Departamento de Computación
UNIVERSIDADE DA CORUÑA

TESIS DOCTORAL CON MENCIÓN INTERNACIONAL

Adquisición y representación del conocimiento mediante procesamiento del lenguaje natural

(Resumen / Abstract)

Doctorando: Milagros FERNÁNDEZ GAVILANES

Directores: Dr. Manuel VILARES FERRO
Dr. Éric VILLEMONTE DE LA CLERGERIE

A Coruña, Octubre 2012

TESIS DOCTORAL:	Adquisición y representación del conocimiento mediante procesamiento del lenguaje natural
AUTOR:	Milagros Fernández Gavilanes
DIRECTORES:	Dr. Manuel Vilares Ferro Dr. Éric Villemonte de la Clergerie
TUTOR:	Dr. Miguel Ángel Alonso Pardo
FECHA:	9 de Octubre de 2012

TRIBUNAL:

PRESIDENTE:

VOCAL 1^{O}:

VOCAL 2^{O}:

VOCAL 3^{O}:

SECRETARIO:

CALIFICACIÓN:

Índice general

Índice de figuras

Resumen

En este contexto, introducimos un marco de recuperación de información que combina procesamiento del lenguaje natural con el conocimiento del dominio. Abordamos la totalidad del proceso de creación, gestión e interrogación de la base de datos documental desde una perspectiva que integra automáticamente conocimientos lingüísticos en un modelo formal de representación semántica, directamente manejable por el sistema. Ello permite la construcción de algoritmos que simplifican las tareas de mantenimiento, proporcionan un acceso más flexible al usuario no especializado, eliminan componentes subjetivos que lleven comportamientos difícilmente predecibles y posibilitan la puesta en marcha de mecanismos para el seguimiento efectivo del funcionamiento del propio sistema.

La adquisición de conocimientos lingüísticos parte de un análisis de dependencias basado en un formalismo gramatical suavemente sensible al contexto. Ello permite conjugar eficacia computacional, potencia expresiva y una excelente capacidad de tratamiento del no-determinismo en programación dinámica. La interpretación matemática de la semántica descansa en la noción de grafo conceptual, que sirve de base tanto para la representación de la colección documental como de las consultas que la interrogan. En el contexto de la Hipótesis de Harris [65], estas representaciones se generan a partir de la información lingüística disponible en los propios textos y constituyen el punto de partida para su indexación.

Las operaciones con grafos se utilizan para el cálculo y la ordenación de las respuestas. Un mecanismo de reconocimiento de patrones aproximado basado en la proyección y generalización de grafos hace posible tener en cuenta la imprecisión intrínseca y el carácter incompleto de la recuperación de información, situando el problema en un marco decidible de cálculo acorde con el *principio de incertidumbre lógica de van Rijsbergen* [150]. Además, el aspecto visual de los grafos permite la construcción de interfaces gráficas de usuario amigables, conciliando precisión e intuición en la gestión de la información.

En relación con anteriores aplicaciones de la teoría de grafos conceptuales al desarrollo de sistemas de recuperación de información, nuestra propuesta resuelve la generación automática de las representaciones semánticas a partir de los textos, evitando la tediosa de la indexación manual. Además, desde nuestro punto de vista, es la primera vez que un marco de pruebas formales se define en este ámbito.

1 | Introducción

La globalización y fiabilidad en el acceso a la información ha justificado la popularización de sistemas de RI, haciendo de su diseño e implementación uno de los mayores retos para la comunidad científica. Tales herramientas se basan en la capacidad que poseen para discernir con respecto de una consulta qué contenidos de la colección documental resultan relevantes de los que no lo son. En concreto, la *relevancia* de un documento viene determinada por la correspondencia entre la representación de su contenido y la de la consulta, por lo que la forma en que estos contenidos se representan es crucial. En este sentido, se requiere de un marco escalable robusto para presentar la información extraída a partir de los documentos, permitiendo su visualización y consulta, lo que nos remite directamente al concepto de ontología en el sentido de la ingeniería del conocimiento y de la *inteligencia artificial* (IA) [35], es decir, al de un marco de referencia para un sistema inteligente en un dominio de conocimiento.

Sin embargo, la mayoría de estos sistemas usan el bien conocido *modelo de espacio vectorial* [129]. Éste se centra principalmente en la coocurrencia de términos, o lo que se suele denominar la recuperación basada en *conjuntos de términos*[1] [64], donde la consideración de estructuras de dependencias sintáctico/semánticas es casi anecdótica. Esto implica que los documentos y las consultas se representen como listas de cadenas que no expresan nada acerca del contexto de la información, ignorando a menudo el aspecto secuencial de las ocurrencias de palabras en los textos[2], a pesar de que el significado de los *lenguajes naturales* (LN's) depende en gran medida de ellas.

Históricamente, este tipo de técnicas se concibió como un medio de almacenamiento de datos más que como un medio para describirlas junto con sus relaciones. Se hace hincapié en cuestiones tales cómo lograr acceder o almacenar dichos datos, pero se ignora en buena medida cual es el sentido que sus autores quieren transmitir [137], lo que se traduce en la falta de consideración de sus significados en un contexto dado. Como consecuencia, aunque estas técnicas han demostrado ser sólidas y eficaces para una gran variedad de textos, en tareas difíciles de recuperación se suele asumir que uno tiene acceso a una gran cantidad de conocimientos previos para mejorar la *precisión* y la *cobertura*. De hecho, la creciente cantidad de datos textuales disponibles electrónicamente ha

[1] en terminología anglosajona, *bag-of-words*.

[2] se ignoran incluso las estructuras más básicas, tales como el orden de los términos en el documento o las fronteras entre frases o párrafos.

incrementado esta necesidad para recuperaciones de alto rendimiento, lo cual sirve de motivación para buscar estrategias que buscan y/o filtran la información basadas en un mayor nivel de comprensión que sólo puede lograrse a través del procesamiento de texto y teniendo en cuenta la semántica. Como punto de partida, las oraciones, y no simplemente los conjuntos de términos, parecen ser la manera natural de mejorar el rendimiento de la recuperación a través de los modelos comunes de documentos [42].

La hipótesis de la que partimos es que con una representación adecuada de los documentos e incorporando conocimientos semánticos limitados, es posible mejorar la eficacia de un sistema de RI. Esto requiere en primer lugar un análisis del texto en profundidad, lo que sitúa de lleno el problema en el marco del *procesamiento del lenguaje natural* (PLN) , aunque con dos características propias. La primera tiene que ver con la cantidad de texto con el que un sistema de RI ha de tratar, y que puede resultar tan grande y heterogéneo que se vuelva poco práctico para llevar a cabo un análisis exhaustivo. La segunda característica viene a suavizar los requerimientos derivados de la primera por cuanto un análisis semántico detallado y preciso no es necesario para las tareas de RI [75], lo que las distingue de otras aún más relacionadas con el PLN como la traducción automática, búsquedas de respuestas o resúmenes automáticos [146]. En este contexto, hemos considerado una estrategia de dos pasos para lidiar con el análisis de textos a nivel de oración. El primero se refiere a la adquisición de conocimiento léxico a nivel de frase, tarea para la que nos hemos inspirado de la arquitectura *Alexina* [120], cuyo núcleo se basa en un analizador de estados finitos. Éste integra un pre-procesador [122] que asume la separación de cadenas de caracteres, la corrección ortográfica y el reconocimiento de entidades nombradas, y que tiene como principal recurso un léxico a gran escala [121]. La salida incluye todas las posibles interpretaciones para cada forma léxica en un *grafo acíclico dirigido* (GAD) que es posteriormente utilizado en una fase de análisis sintáctico. Al respecto, hemos elegido un formalismo *suavemente dependiente del contexto* [153], que proporciona la potencia suficiente para su aplicación sobre LN's, sin por ello renunciar a la eficacia computacional.

En cualquier caso, para rentabilizar las ventajas asociadas al análisis de texto en tareas de RI, también es necesario disponer de una notación formal que sirva como intermediario entre el humano y el ordenador. Concretamente, los *grafos conceptuales* (GC's) [137] poseen la potencialidad necesaria para describir el significado de los datos de acuerdo con la visión del usuario, a la vez que podemos asociarlos con procedimientos que permiten acceder a los datos en la máquina. Estamos así en disposición de evitar el tener que recibir una formación específica para acceder a ellos e interpretar tanto resultados finales como parciales, algo de lo que también adolece la recuperación basada en conjuntos de términos. Por otra parte, la consideración de un mecanismo de inferencia conceptual como el señalado nos permite estimar la *granularidad semántica* de un documento [168], la cual hace referencia al nivel de detalle que conlleva un elemento de información [51]. De esta manera, se abren las puertas para abordar tareas que implican búsqueda de consultas ambiguas, colecciones documentales incompletas y RI aplicada en dominio específico.

Todo ello justifica que nos hayamos decantado por la elección de este tipo de estructura como formalismo de representación semántica.

Formalmente, los GC's obtenidos son derivados de acuerdo con un modelo de dependencias. Concretamente, la colección documental se analiza sintácticamente en un primer momento con el fin de generar un *grafo inicial de dependencias* (GID's) que más tarde será traducido en dependencias *gobernante/gobernado* (GDGG), es decir, relacionando el núcleo de un sintagma con sus modificadores. A partir de aquí y mediante la aplicación de un conjunto de valores iniciales proporcionados por el programador para las clases semánticas (los tipos), marcadores lingüísticos y patrones sintácticos, podemos aproximar y extender de forma fehaciente ambos conjuntos iniciales de dependencias y clases. Una cuidadosa implementación en programación dinámica permite posponer el tratamiento de las ambigüedades tanto de tipo léxico como sintáctico a una posterior fase de definición semántica, donde un protocolo de adquisición de conocimiento iterativo sirve para filtrar interpretaciones irrelevantes con el fin de obtener los GC's. A nuestro conocimiento, no se ha presentado ni documentado una propuesta parecida para la generación automática de representaciones semánticas a partir de textos. Esto es lo que nos va a permitir realizar una formulación simple de la tarea de recuperación. Así, cuando un usuario realiza una pregunta en LN, el sistema la traduce a un GC y luego trata de buscar en la colección documental otros GC's que sean relevantes con respecto al primero. Una vez encontrados, se pueden utilizar para acceder a su información y calcular las respuestas.

Más tarde, necesitaremos incorporar una *función de ordenación*[3] con el fin de clasificar los documentos recuperados en base a su relevancia con respecto a la consulta. El objetivo es evitar que el usuario pierda el tiempo buscando en las listas de resultados obtenidas, entendiendo que en ellas se encuentran numerosos documentos irrelevantes, especialmente cuando sabemos de antemano que quién las revisa rara vez va más allá de la primera página del conjunto recuperado [61], lo cual constituye una causa mayor en la falta de satisfacción asociado a los sistemas de RI [47] y puede llegar a desvirtuar la capacidad real del propio buscador [60]. Para resolver este problema, nos hemos inspirado en trabajos anteriores, donde la función de ordenación se caracteriza mediante una relación de orden parcial sobre el conjunto de transformaciones aplicadas a la consulta con el fin de satisfacer su cometido en la colección documental [56]. La idea consiste en asignar diferentes pesos a estas transformaciones dependiendo de su naturaleza estructural, lo cual nos permitirá centrarnos en criterios de búsqueda lejos de las preferencias personales, descartando los enfoques basados en aprendizaje supervisado debido a su elevado coste en términos humanos.

Sin embargo, existe una preocupación primordial en el campo de la RI que es la evaluación. En este sentido, nuestra propuesta define un marco formal de pruebas que permite la consideración de diferentes técnicas de ordenación para estos sistemas, como son la aplicación o no de *juicios de relevancia* (JREL's), a menudo almacenados en un

[3] también llamada *función de recuperación* por Fuhr y Buckley [53].

fichero denominado *relevancia de la consulta* (CREL), y la selección de un conjunto representativo de *consultas o tópicos* en función de las necesidades de información. De un modo más detallado, en el caso de la tarea de ordenación nuestro punto de partida ha sido el protocolo clásico empleado en la conferencia *Text Retrieval Conference* (TREC) y basado en JREL's [156]. Pero también hemos estimado una simple variación de éstos usando *pseudo*-JREL's (PJREL's), propuestos por Soboroff *et al.* [135] y una alternativa algo diferente, incorporando los JREL's y/o PJREL's pero considerando un criterio algo distinto para la realización de la ordenación. Para ello, hemos retomado una técnica inspirada en la noción de *autoridad del sistema* descrita por Mizzaro *et al.* [101]. En cuanto a las técnicas de ordenación que no tienen en cuenta los JREL's, se optó por evaluar nuestra propuesta mediante un método inspirado en Wu *et al.* [164], que parece ser uno de los más populares en su tipo y que se basa en la idea de comparar la efectividad del motor de búsqueda con los resultados proporcionados por un conjunto de sistemas de RI que sirvan como referencia.

Con respecto a la elección del conjunto de consultas, hemos combinado una serie de trabajos anteriores en torno a dos preguntas complementarias. La primera se refiere a la selección de una consulta para un sistema de RI individual, aplicando el concepto de *precisión media* (PM) [11]. El siguiente consiste en la selección, pero esta vez de un conjunto de tópicos para un determinado sistema [62]. A partir de estas técnicas, y a falta de soluciones definitivas y específicas en el estado del arte, proponemos un método razonado de selección a partir de un conjunto de sistemas de RI, inspirado tanto en la valoración basada en el tipo humano como en la noción de *conectividad del tópico*[4] propuesto por Mizzaro *et al.* [101].

El resto del trabajo se organiza de la siguiente manera. La Sección 2 nos proporciona una vista previa del estado del arte en la RI conceptual, centrándose en las tareas de indexación y de evaluación. En la Sección 3, introduciremos nuestro *corpus* de ejemplo que nos servirá como guía para la discusión. La Sección 4 presenta la teoría básica sobre GC's y sus aplicaciones en RI. Una descripción detallada de nuestra propuesta de adquisición de conocimiento constituirá la Sección 5. La Sección 6 introduce nuestro marco formal de pruebas, centrando la atención en la selección de tópicos y en la ordenación de los sistemas de RI. Más adelante, en la Sección 7 se realiza una serie exhaustiva de pruebas experimentales, mientras que la Sección 8 cierra el documento con las conclusiones del estudio.

2 | Estado del arte

La incorporación del PLN a la RI siempre ha fascinado a los investigadores, en un doble objetivo: la integración de técnicas de interpretación de textos para identificar los términos índice, y la caracterización de su estructura interna. En este punto, el estado del

[4] en terminología anglosajona *topic hubness*.

arte nos sitúa en un marco genérico de trabajo al que se refiere de diferentes formas. Así, algunos autores hablan de *indexación motivada lingüísticamente* [85, 104], mientras que otros consideran como más apropiado el término *indexación semántica* [84, 133]. Algunos trabajos recurren incluso a la expresión de *recuperación inteligente* [38, 57, 134, 146] para subrayar la interacción entre la mente humana y la IA a través de redes y tecnología.

2.1 | Indexación semántica

Tradicionalmente, estas estructuras de indexación pueden ir desde simples palabras hasta unidades multipalabra. Por lo tanto, sobre ellos se suele aplicar un leve análisis lingüístico, utilizando léxicos para lograr una descomposición morfológica simple y la reducción de las palabras a su raíz, eliminando sufijos, afijos y demás de un modo superficial [59, 71, 81]. Pero también se puede aplicar un análisis algo más profundo, que revele la estructura interna de las palabras[5]. Debido a la abundancia de información disponible, estos métodos siguen siendo de los más empleados, y son capaces de hacer frente a algunos fenómenos lingüísticos complejos tales como los pronombres clíticos, contracciones y reconocimiento de nombres propios [2].

Sin embargo, nuestro principal interés se centra en captar la esencia de los documentos mediante la utilización de técnicas de análisis algo más elaboradas, tales como el uso de sintagmas significativos, pero también de frases como condición para la categorización automática de los documentos. Se trata en definitiva de una vieja idea que debiera marcar una mejora sobre el uso de palabras sueltas, aunque en la práctica exista poca evidencia de ello. De hecho, la convicción generalmente aceptada durante mucho tiempo [134, 75] era que sólo las técnicas lingüísticas superficiales podían resultar de interés en el desarrollo de este tipo de aplicaciones [134], aunque, en el mejor de los casos, su efecto positivo sobre la precisión era pequeño [85]. No obstante, la característica que define a estos métodos es que explotan conocimientos léxicos, morfológicos y/o sintácticos, con el fin de detectar relaciones de dependencia lingüística entre palabras, su representación formal y posterior definición de un mecanismo de localización de información en base a ésta.

Podemos diferenciar [85, 171] dos niveles de complejidad en el tratamiento de dependencias en textos. El nivel más bajo se orienta al léxico, lingüísticamente menos sofisticado y representado por un grupo de técnicas conocidas como *modelado de dependencias*. Por lo general, estos sistemas consideran las dependencias existentes entre determinados pares o ternas de palabras [125], a menudo asociadas a un modelo probabilístico [26, 90, 95, 136] con el fin de clasificar las relaciones más plausibles. En este sentido, la mayoría de las estrategias de extracción de términos compuestos se basan en el uso de métodos estadísticos [46] o también en un reconocimiento simple de patrones [78, 132], en lugar de considerar las relaciones estructurales entre los elementos que conforman la oración. Más recientemente, algunos autores propusieron la utilización

[5] por medio de la lematización o de las familias morfológicas sin tener en cuenta la información sintáctica.

de técnicas de análisis superficial para la detección de estos pares [2] y/o ternas [85] de palabras relacionados mediante algún tipo de dependencia sintáctica. Todos estos trabajos muestran la mejora obtenida con respecto al modelo basado en palabras con independencia del idioma[6], en particular cuando se trata de un lenguaje rico en léxico y morfología. Sin embargo, el principal problema radica en la dificultad de integrar la proximidad de términos en el marco descrito. El espacio de parámetros puede volverse muy amplio considerando directamente las dependencias, haciendo la estrategia sensible a la escasa información y al ruido, lo que podría contrarrestar relativamente las pequeñas ventajas que se podrían obtener y sobre las que justificar el interés en modelos de proximidad del lenguaje [171].

Por este motivo, el nivel superior en el tratamiento de dependencias en textos trata de incorporar unidades mayores a las palabras a la hora de afrontar su representación, de modo que las dependencias existentes entre términos pueden ser capturadas indirectamente. Al igual que ocurría en el caso anterior, existen técnicas para la extracción de frases directamente relacionadas con métodos estadísticos [34, 54], con reconocimiento de patrones [116], pero también con técnicas de análisis sintáctico profundo [50, 143]. Sin embargo, aunque no se requiere de un análisis semántico muy detallado y preciso para la realización de tareas de RI [134], con el crecimiento desmesurado de la información, resulta difícil recuperar los documentos relevantes únicamente mediante métodos estadísticos [146]. El origen del problema se sitúa en el excesivo número de términos susceptibles de ser de interés para la descripción de una colección documental, pero a su vez también está relacionada con la dificultad de hacer frente a la escasez de datos en este contexto. En este sentido, las representaciones de textos basadas en grafos etiquetados parecen ser capaces de detectar enlaces no siempre evidentes entre los conceptos [73, 97, 133], independientemente del tamaño del *corpus* considerado. El acercamiento no sólo resulta prometedor, sino que posee el potencial de mejorar el modelo estándar de conjuntos de términos, sobre todo en respuestas a consultas largas [94], una idea en torno a la cuál el consenso es muy amplio [36], siendo varias las estrategias propuestas. Por tanto, aunque hasta hace poco el más conocido de estos acercamientos eran las *redes semánticas* [91], probablemente ninguno de ellos ha sido tan popular en los últimos tiempos como los GC's [137]. En realidad, los GC's son una extensión de las anteriores, introduciendo la noción de dependencia entre nodos. Éstos poseen tres ventajas principales como método de descripción formal. En primer lugar, pueden apoyar una correspondencia directa a partir de una base de datos relacional [32]. En segundo, pueden ser usadas como base semántica para el LN. Finalmente, basándonos en las transformaciones sobre grafos, permiten dar soporte a inferencias automáticas para calcular las relaciones que no son explícitamente mencionadas [57].

Esta aparente versatilidad del modelo basado en grafos debe además dar respuesta

[6] en la práctica, en los entornos de recuperación, normalmente se supone que las palabras asignadas a los documentos de una colección aparecen de manera independiente las unas de las otras [125]. La hipótesis de independencia entre ellas no es realista en muchos de los casos, pero su uso conlleva la utilización de un algoritmo de recuperación simple.

a la búsqueda de aquellos documentos que se encuentran representados de un modo incompleto, incluso a partir de consultas confusas. Este fenómeno, que ha justificado durante bastante tiempo la consideración de estrategias basadas en lógica probabilística, crece ahora de manera exponencial como consecuencia de la imposibilidad de integrar la cantidad total de información disponible en tareas de RI. Se trata en definitiva de formalizar la implementación del *principio de incertidumbre lógica de van Rijsbergen's* [151], según el cual la relevancia es una cuestión de grado y el problema central de la RI radica en como modelarlo y medirlo. Como consecuencia, asumir que dicho proceso puede ser mejorado mediante coincidencias exactas o por medio de la lógica clásica es un intento vano [57]. Este desajuste ha servido en cierto modo de campo propiciatorio para difundir ese sentimiento de que la mejora utilizando frases como índices no parece que sea la alternativa que mejor se ajuste al tratamiento de este tipo de problemas [7] de RI [55].

En este contexto, algunos autores adoptan una posición intermedia, investigando técnicas que hacen uso de conocimientos semánticos limitados, los cuales a su vez pueden ser fácilmente representables a partir del texto usando un formalismo en forma de GC [138]. Esto permite expresar el sentido de la colección documental de una manera lógicamente precisa, humanamente entendible y computacionalmente manejable. Gracias a la correspondencia directa existente entre este tipo de representación y el lenguaje, los GC's desempeñan el papel de lenguaje intermedio para la traducción entre los formalismos orientados a la máquina y el LN. Pero además, este tipo de representación gráfica sirve de lenguaje de especificación y de modelo legible por el usuario, a la vez que formal. Esto justifica que la noción de consulta conceptual date de los primeros tiempos de la investigación en el campo de la RI [139], así como el esfuerzo llevado a cabo en los últimos años con el fin de reemplazar las nociones clásicas probabilísticas por transformaciones formales de grafos [57], o simplemente de completarlas [134, 146].

2.2 | Estrategia de ordenación

Tradicionalmente, la relevancia de los documentos se ha venido estimando usando una variedad de funciones de ordenación basadas en la similitud, que, en la práctica, no dejan de ser simples estrategias empleadas por los motores de búsqueda para ajustar los pesos asociados a los términos de indexación con el fin de optimizar su rendimiento[8]. Más recientemente, las funciones basadas en la popularidad han ganado cierta notoriedad. Estos modelos explotan la existencia de una correlación cercana entre la popularidad y la relevancia, principalmente en el caso de sistemas de RI que gestionan gran cantidad de datos y accesos por parte de los usuarios, como en el ejemplo típico de las búsquedas Web [82, 109]. Sin embargo, aunque en la actualidad los algoritmos encargados de la

[7] en algunos casos, la mejora de la eficacia se logra mientras que para otros, se alcanzan unos resultados marginales o negativos.

[8] algunos autores hablan indistintamente de estrategias de ponderación de términos y de funciones de ordenación [47].

evaluación de la popularidad de los documentos se han vuelto cada vez más sofisticados, es necesario aplicar un esfuerzo específico para evitar algunos problemas inherentes a esta técnica. Nos referimos concretamente al tratamiento de contenidos de nueva incorporación que poseen pocos accesos [9, 18, 23, 41, 45, 86, 105], al hecho de que los documentos más populares tienden a serlo cada vez más [8, 27, 28, 58] o a la eliminación de posibles manipulaciones en las ordenaciones mediante la utilización de enlaces promovidos artificialmente [4, 6, 22, 76, 96, 108, 144].

A pesar de ello, ambos modelos de ordenación, los basados en la similitud y en la popularidad, no parecen ser por sí solos lo suficientemente eficaces como para dar apoyo en la RI aplicada a un dominio general o incluso a uno específico [168]. Este es el motivo por el que se justifica la consideración de propuestas híbridas, ya ampliamente aplicadas [43, 67, 109], incluso cuando las basadas en similitud parecen ser el punto de partida determinante para la obtención de la eficiencia en la recuperación. Con respecto a esto, una alternativa para mejorar su rendimiento consiste en medir directamente la similitud conceptual, la cual puede ser estimada de diferentes maneras. Así, algunos trabajos la calculan mediante el *concepto de menor ancestro común* (CMAC) a partir del contenido de información, algo que parece acercarse a las funciones de ordenación implícitas ejercidas por los humanos [115]. La idea original se debe a Cohen *et al.* [33] que describen un método para calcular la CMAC entre un par de conceptos, el cual nos permite relacionarlos a través de una descripción más específica que integra las respectivas estructuras. De esta manera, podemos inferir relaciones de subconcepto/superconcepto (resp. si un determinado individuo pertenece a un concepto determinado), proporcionando una herramienta para obtener elementos explícitos comunes y derivar conocimiento implícito usando técnicas orientadas a la inclusión en una categoría (resp. instancia) [88]. El estado del arte retoma este estudio con el objetivo de utilizar el contenido de la información para evaluar la similitud semántica en las taxonomías [115], y que más tarde serviría de inspiración para lidiar de diferentes maneras con las tareas de computación en el contexto de la tecnología en RI. Es el caso de algunos autores [102] que se aprovechan directamente de esta técnica para ampliar las medidas clásicas para la comparación de textos, como por ejemplo en el caso del coeficiente Dice [40]. De la misma manera, se consideran otras técnicas diferentes de las que utilizan CMAC, incluyendo a su vez extensiones alternativas a la medida Dice [103], así como relaciones de generalización asociadas a un dominio de conocimiento específico [119]. En cualquier caso, estas propuestas necesitan en primer lugar disponer de una estructura ontológica basada en conocimiento para representar estos conceptos, así como la tecnología estadística basada en *corpus* para generarlos y gestionarlos, situándonos de este modo en el contexto de la RI conceptual [139].

Desde el punto de vista operativo, sea cual sea el criterio de relevancia considerado, una función de ordenación se puede clasificar atendiendo a tres criterios complementarios relacionados con su fase de generación: la capacidad de adaptación al contexto, la naturaleza supervisada y la consideración de un modelo basado en aprendizaje [92].

En relación al primero de ellos, la mayoría de los sistemas de RI utilizan una estrategia fija para apoyar la tarea de clasificación definiendo su contexto de trabajo, independientemente de la heterogeneidad de los usuarios, de las consultas y de las colecciones [47]. Es el llamado *consenso de búsqueda*, en el que la relevancia calculada para toda la población se supone apropiada para todos los individuos y, como consecuencia, todos obtienen los mismos resultados. A pesar de que podríamos interpretar esta uniformidad como una ventaja, debido a que permite la comparación de los resultados de búsqueda entre los diferentes usuarios, lo cierto es que la idea de adecuar las características del proceso de recuperación a nuestras propias preferencias resulta siempre atractiva. Se habla entonces de *búsquedas personalizadas* [110], un enfoque que parece no aplicarse de forma consistente en diferentes contextos [126, 173].

Por otro lado, la RI tradicional se centra principalmente en modelos de ordenación sin supervisión, generalmente basados en el grado de correspondencia entre la consulta y el documento. Es el caso del booleano [150], del vectorial [126], del probabilístico [117], y de los asociados al modelado del lenguaje [111]. Teóricamente resultan sencillos e intuitivos, funcionan razonablemente bien y no requieren de datos etiquetados, una ventaja que no excluye la posibilidad de asociar un número de parámetros de ajuste mediante el uso de alguna técnica de entrenamiento, lo que no es inusual. Sin embargo, como los modelos de ordenación han visto incrementada su sofisticación, el conseguir ajustarlos convenientemente se ha convertido en una cuestión cada vez más difícil [167] y, en la práctica, estos enfoques empíricos sólo disponen de unos pocos parámetros que permitan ser afinados [7].

En contraste con los enfoques no supervisados, los supervisados disfrutan de una mayor precisión y una mejor adaptabilidad, al tiempo que requieren de un esfuerzo humano más importante, lo que durante muchos años limitó el interés práctico en este tipo de estrategias. Sin embargo, la disponibilidad actual de conjuntos etiquetados de evaluación de la relevancia realizados por grupos de expertos ofrecen una alternativa práctica para incorporar técnicas de aprendizaje automático en el diseño de modelos de ordenación. La idea consiste en usar estos recursos etiquetados como medio de entrenamiento para estimar la proximidad semántica entre las consultas y los documentos [168] a través de la minimización de una *función de pérdida* indirectamente relacionada con determinadas medidas de rendimiento de la RI, como el *promedio de la precisión media*[9] (PPM) o la *la ganancia acumulativa reducida normalizada*[10] (GAARN), aunque también existen propuestas que permiten optimizar cualquiera de ellas [166]. En este sentido, se han descrito una gran variedad de estrategias de aprendizaje, tales como las redes neuronales [14, 17], las máquinas soportadas por vectores [16, 68, 69, 74, 152, 169], el «boosting» [52, 93, 166] o la programación genética [37, 39, 48, 148]. En la práctica, aunque estos métodos parecen funcionar mejor que los no supervisados tradicionales [92, 167], se pueden observar algunas diferencias importantes dependiendo

[9] en terminología anglosajona se denomina *mean average precision*.
[10] en terminología anglosajona se denomina *normalized discounted acumulative gain*.

del tipo de instancias utilizados en el aprendizaje. Más en detalle, se han abordado tres modelos diferentes de instanciación: punto a punto, por parejas, y por lista.

En el acercamiento punto a punto [93, 107], cada par de entrenamiento consulta-documento asocia una puntuación de manera independiente, lo que implica que no se consideran las preferencias relativas entre dos documentos recuperados para una misma consulta. Como consecuencia, el método ha demostrado tener un bajo rendimiento durante la fase de aprendizaje de la ordenación, transformando el problema en uno de regresión o de clasificación de un único documento [87]. En cambio, los basados en parejas [14, 16, 52, 69, 74, 87, 149, 168, 170] parecen ser los más populares. Los pares de documentos recuperados dada una consulta, en los que se ha determinado cuál de ellos es el más relevante, constituyen aquí las instancias del conjunto de entrenamiento. Así, el objetivo del proceso de aprendizaje es reducir al mínimo el número medio de inversiones en la ordenación, con el fin de obtener un clasificador binario que pueda indicar qué documento es mejor en un par dado. Esto implica que, dada una consulta, debemos inducir una ordenación total para un conjunto de documentos recuperados a partir de uno parcial entre pares, lo que limita severamente las posibilidades prácticas de este enfoque [10]. Por último, el modelo por lista [10, 15, 17, 89, 112, 165, 166, 169] también ha visto incrementado su popularidad en los últimos años. Considera el conjunto de documentos recuperados para una consulta como instancias en la fase de entrenamiento. Esto debería ser suficiente para superar los problemas anteriormente mencionados en relación con las técnicas punto a punto y por parejas y, de hecho, los resultados prácticos sugieren que éstas poseen cierta superioridad sobre las demás. Sin embargo, la definición de una función de pérdida puede convertirse en una tarea compleja porque la mayoría de las medidas de evaluación en RI no son magnitudes continuas con respecto a los parámetros del modelo de ordenación.

Finalmente, existe un amplio espectro de técnicas básicas de ordenación disponibles. Cada una de ellas tiene su propio conjunto de ventajas que deberíamos tratar de reconciliar mediante propuestas mixtas, y tener claro cuáles son los inconvenientes que se quieren evitar o al menos minimizar. A este respecto, probablemente la combinación de factores óptimos depende de la naturaleza de la tarea de búsqueda con la que queremos tratar. En nuestro caso, se refiere al tratamiento de un dominio específico. La afirmación de la existencia de claros beneficios derivados de la utilización de la similitud basada para la fase de búsqueda nos sitúa directamente en el contexto de algunos trabajos recientes [168], incorporando una dimensión de popularidad cuando el entorno de trabajo puede garantizar un número suficiente de accesos.

2.3 | Evaluación de la recuperación de la información

En este sentido, las técnicas basadas en JREL's y popularizadas por el TREC [156, 157] son consideradas como un estándar *de facto* para la evaluación en RI. Los eventos realizados por el TREC enfocan esta cuestión tomando como fondo común los

100 primeros documentos devueltos por cada sistema participante. Más tarde, estos documentos se revisan por especialistas que juzgan su relevancia, inspirándose en la metodología *Cranfield* [31, 30]. En definitiva, se trata de comparar los sistemas de RI con un conjunto de tópicos o consultas, una serie de documentos referidos a cada uno de ellos, y un conjunto de JREL's por cada documento. Este tipo de experimentación a gran escala ha sido el referente en este campo durante más de veinte años, denominándose *selección profunda*[11]. Sin embargo, el incremento del tamaño, de la complejidad y de la heterogeneidad de las colecciones documentales; así como del conjunto de consultas, lo han hecho inviable.

Por ello, se han propuesto una serie de enfoques alternativos para estimar el rendimiento de los sistemas de RI con recursos limitados de JREL'S, con el fin de reducir el esfuerzo humano en la creación de colecciones de prueba. El primero trata de conseguirlo seleccionando el mejor conjunto de documentos para ser evaluado y teniendo en cuenta medidas de calidad en aquellos casos en los que se pueden realizar pocos juicios. En esta categoría, podemos incluir como primera tentativa las técnicas de *selección*[12] [140], las cuales se centran en aquellos textos que menos probabilidades tienen de ser no relevantes. Sin embargo, trabajos recientes sugieren que el crecimiento en el tamaño de los *corpora* está superando incluso la capacidad de esta técnica para encontrar y juzgar suficientes documentos [13], ya que si se consideraran menos, las estimaciones de las medidas de evaluación tendrían una mayor varianza. En este sentido, algunos autores [21] tratan de reducir el esfuerzo necesario para juzgar a la vez que mantienen un gran número de tópicos, aunque reconocen que analizar los fallos resulta más complejo, por lo que esta vía necesita todavía seguir siendo explorada.

Una segunda alternativa relaja la carga de la valoración de tipo humano de la generación de JREL para introducir la noción de PJREL, los cuales se crean o bien aleatoriamente, seleccionando una correspondencia entre los documentos sobre los tópicos [135], o bien haciendo una lectura rápida de los situados en las posiciones más altas en la ordenación devuelta a partir de un subconjunto de representaciones de tópicos [44].

Por su parte, Mizzaro *et al.* [101] proponen un método de análisis de datos recogidos a partir de recursos de evaluación basados en JREL's o a partir de sistemas de RI similares, como es el caso de los PJREL's. Mediante la introducción de dos versiones normalizadas de PM que los autores usan para construir un grafo bipartito ponderado de motores de búsqueda y tópicos, encontraron que las medidas sobre la autoridad del sistema sirven para medir su rendimiento y que la conectividad revela la sencillez o complejidad de un tópico.

Finalmente, algunas propuestas [164] prescinden del concepto de JREL's, utilizando el solapamiento de los resultados obtenidos. Brevemente, la técnica pasa por interpretar

[11] en terminología anglosajona se denomina *depth pooling*.
[12] en terminología anglosajona se denomina *pooling*.

la relación entre los documentos recuperados a partir de un grupo de sistemas de RI, donde dicha estructura de superposición parece proporcionar un fuerte impacto sobre los resultados. Así, se suele argumentar [142] que este tipo de métodos pueden producir malos resultados en los sistemas con mejor rendimiento cuando éstos se clasifican junto con los de menor rendimiento, a la vez que parece que obtienen peores resultados que el grupo anterior de técnicas.

Otro aspecto a tener en cuenta para definir un marco de pruebas formal en sistemas de RI es la elección adecuada de un conjunto de tópicos o consultas, con el fin de determinar cuáles son los mejores en la predicción de la relevancia real. El trabajo de investigación desarrollado al respecto es escaso y los resultados prácticos se limitan, *de facto*, a algunas ideas relacionadas con la hipótesis del trabajo y propuesta de estrategias de selección cuya validación requiere todavía una seria experimentación. En el apartado de hipótesis ya confirmadas, Mizzaro [100, 101] demuestra formalmente que algunos tópicos son más fáciles que otros y que existen diferencias entre los sistemas a la hora de distinguir entre los fáciles y los difíciles. Sin embargo, aunque podemos decir que no todos ellos son igualmente informativos sobre los sistemas de RI, no tenemos evidencias en cuanto a qué criterio podría ser mejor para calificar esta afirmación.

Estos trabajos en el campo de la evaluación de la RI sugieren de manera reiterada que los tópicos individuales varían enormemente en su capacidad para discriminar entre sistemas, lo cual provoca que se extienda la atención también en la construcción del propio conjunto de tópicos. Se trataría no sólo de discernir cuando un conjunto de este tipo es más útil que otro, siempre con un propósito de evaluación, sino también de seleccionar un número de ellos lo más pequeño posible sin que por ello pierdan esa cualidad. Ello permitiría reducir la carga de trabajo en una metodología cuyo principal problema es el coste, lo que justifica el interés práctico de este tipo de estrategias. Sin embargo, aunque desde hace muchos años ha existido preocupación por esta cuestión, no se han producido contribuciones relevantes hasta hace poco tiempo [62]. Los trabajos anteriores se basan exclusivamente en lo que debe ser la selección profunda, tomando como base metodológica algún tipo de enfoque heurístico [11, 130, 140, 160, 162, 172] que, por desgracia, proporciona para cada caso un resultado diferente. Con respecto a esto, aunque la propuesta de Guiver *et al.* en [62], no intenta conseguir de inmediato una solución completa al problema de la identificación de conjuntos adecuados de tópicos, demuestra formalmente la existencia de fenómenos de complementariedad entre éstos y su influencia en la calidad de la evaluación, desechando la hipótesis de que se trate de un efecto aleatorio. El método se basa en el PPM [63]. Más en detalle, se aplica una búsqueda exhaustiva sobre todos los posibles subconjuntos de tópicos en un intervalo de cardinalidad. Para cada subconjunto, se calcula el correspondiente PPM, así como la correspondiente correlación sobre todos los tópicos con PPM. Los autores argumentan que los mayores valores de correlación (resp. menor) corresponden con los mejores (resp. los peores) conjuntos de tópicos. Sin embargo, el principal obstáculo para la aplicación directa de este método es el complejo análisis combinatorio que requiere, lo que implica

poseer un amplio conjunto de tópicos evaluados y de sistemas asociados ejecutándose. De esta manera, la ganancia de tal reducción, puede ser relativamente pequeña para un esfuerzo importante y es necesario prever algún tipo de estrategia heurística a fin de evitar búsquedas completas en este espacio de trabajo.

3 | El corpus

Con el fin de favorecer la comprensión de nuestra propuesta, vamos a presentarla a partir de un *corpus* botánico que describe la flora del África Occidental. Concretamente, nos hemos centrado en el trabajo *«Flore du Cameroun»*, publicado entre los años 1963 y 2001, el cuál se compone de unos cuarenta volúmenes en francés, donde cada uno consta aproximadamente de 300 páginas organizadas como una secuencia de secciones, cada una dedicada a una especie y siguiendo un esquema estructural sistemático. Así, las secciones incluyen una parte descriptiva enumerando aspectos morfológicos tales como el color, la textura, el tamaño o la forma. Esto implica la presencia de frases nominales, adjetivos y también adverbios para expresar frecuencia e intensidad, pero también el de entidades nombradas para denotar dimensiones.

El texto se organiza taxonómicamente, introduciendo especies (resp. géneros) en capítulos separados (resp. secciones), lo cual resulta ser equivalente a la hiperonimia o a relaciones «`es_un`». Sin embargo, las descripciones incluyen conceptos que están relacionados sin ser taxonómicamente. Podemos distinguir entre las relaciones etiquetadas[13], que pueden ser recuperadas a través de frases nominales que son expresadas de un modo asertivo[14], pueden ser propagadas por estructuras más complejas y que requieren la consideración de técnicas sofisticadas de PLN con el fin de reconocerlas, como en el caso de las enumeraciones y en las definiciones de intervalos[15]. La colección también cuenta con un vocabulario que es compartido por la mayoría de los textos basados en este ámbito, y es de tamaño suficiente para nuestros propósitos. Esto nos va a permitir evaluar nuestra propuesta sobre una variedad de formas verbales y nominales para las que la semántica correcta no es trivial. En particular, debido a la diversidad de las construcciones lingüísticas presentes en el *corpus* y a los diferentes modos de expresarlos, parece una plataforma de pruebas adecuada para estudiar los fenómenos de ambigüedad e integridad gramatical.

El *corpus*[16] ha sido previamente pasado de texto sobre papel a formato electrónico [123], capturando su estructura lógica con el fin de explorarlo, como parte del proyecto BIOTIM [118], una iniciativa de investigación sobre la gestión integral de documentos botánicos que incluye la adquisición conceptual y tareas de minería de texto.

[13] típicamente relacionados con propiedades como `color`, `forma`, `tamaño`, `textura` o `posición`, o bien con entidades como `órgano` o `fruto`.

[14] es el caso, por ejemplo, de las relaciones del tipo «`en forma de`» o «`de color`».

[15] es el caso de construcciones del tipo «`de X a Y`» o «`X e Y`».

[16] proporcionado por el *Instituto Francés de Investigación para el Desarrollo Cooperativo*.

A partir de ahora, vamos a denotar a este *corpus* de ejemplo por $\mathscr{B}$.

4 | Grafos conceptuales y búsquedas

Si nos centramos en [25], el núcleo de nuestra propuesta en RI son los llamados *grafos conceptuales básicos* (GCB's). Se trata de GC's sin negación que describen entidades y relaciones entre éstas, e introducen razonamiento sobre la base de un morfismo de grafos llamado *proyección*. De hecho, se puede demostrar que su comprobación consiste esencialmente en el mismo problema que la satisfacción de restricciones o el de la contención de consultas conjuntivas en las bases de datos [83] y, en particular, sus resultados son sólidos y completos con respecto a la deducción en la *lógica de primer orden* (LPO). En conjunto, esto pone de relieve una cuestión fundamental en RI, a saber la de las respuestas a consultas, cuya finalidad es interesarse por la recuperación de todas las respuestas a una pregunta. La mayor parte de los contenidos de esta sección se toman y/o se inspiran en Chein *et al.* [25] y en Genest *et al.* [57].

4.1 | Grafos conceptuales básicos

El primer paso para la puesta en marcha de la estrategia de interrogación es la definición de un marco de trabajo que permita trazar un mapa cognitivo de los conocimientos ontológicos básicos con los que estamos trabajando. A esta estructura la denominamos *soporte* y compila los principales conceptos, relaciones y el vocabulario que existe en el mundo que estamos tratando de describir.

Definición 1 *Un* soporte *es una tripla* $\mathcal{S} = (\mathcal{T}_\mathcal{C}, \mathcal{T}_\mathcal{R}, \mathcal{I})$ *de conjuntos disjuntos, donde:*

- $\mathcal{T}_\mathcal{C}$ *y* $\mathcal{T}_\mathcal{R}$ *son conjuntos finitos parcialmente ordenados*[17] *de* tipos conceptuales *y* tipos relacionales*, respectivamente, donde el orden que los rige es interpretado como una relación de especialización. Entonces,* $t \leq r$ *se lee como que* r *es una generalización de* t*, o que* t *es una especialización de* r*.*
- *Los tipos de* $\mathcal{T}_\mathcal{C}$ *poseen un* tipo universal *que generaliza a todos los demás, denotado por* $\top$*. Del mismo modo, los tipos de* $\mathcal{T}_\mathcal{R}$ *pueden tener cualquier aridad*[18] *superior o igual a 1, y sólo aquéllos con misma aridad serán comparables.*
- $\mathcal{I}$ *es un conjunto numerable de* referentes individuales*, con un* referente genérico *denotado por* $* \notin \mathcal{I}$*. El conjunto* $\mathcal{I} \cup \{*\}$ *está ordenado parcialmente y sus elementos son dos a dos no comparables entre sí, siendo* $*$ *el más general.*

■

[17] en este caso el conjunto parcialmente ordenado no es más que una jerarquía de tipos.

[18] la aridad de un operador matemático o de una función es el número de argumentos necesarios para que dicho operador o función se pueda calcular.

En definitiva, un soporte consiste en una jerarquía de tipos conceptuales, una jerarquía de tipos relacionales y un conjunto de referentes individuales que pueden ser identificados mediante un diccionario cuyos elementos se asociarán más tarde con tipos conceptuales. En la práctica, este diccionario representa formas léxicas de un tesauro o de un *corpus*, mientras que los tipos conceptuales se referirán a sus clases semánticas, y los relacionales al nexo que las une.

Una vez introducidos los conceptos y las relaciones que formarán parte del mapa general del dominio, podemos enlazarlos entre sí con el fin de describir hechos en los que estamos interesados. En este sentido, un GCB representa la plantilla que se va a rellenar con los conceptos/relaciones tomadas a partir del soporte.

Definición 2 *Formalmente, un* GCB *definidos sobre un soporte* $\mathcal{S} = (\mathcal{T}_\mathcal{C}, \mathcal{T}_\mathcal{R}, \mathcal{I})$*, es una cuádrupla* $\mathcal{G} = (\mathcal{C} \cup \mathcal{R}, \mathcal{A}, \mathcal{E})$ *que satisface las siguientes condiciones:*

- *(*$\mathcal{C} \cup \mathcal{R}, \mathcal{A}$*) es un multigrafo bipartito, no necesariamente conexo, donde* $\mathcal{C}$ *y* $\mathcal{R}$ *son conjuntos disjuntos de* nodos conceptos *y* nodos relaciones*, respectivamente.*
- $\mathcal{A}$ *es el multiconjunto de* aristas*.*
- $\mathcal{E}$ *es una función de etiquetado de nodos y relaciones del grafo* $\mathcal{G}$ *que verifica:*
  - *Un nodo concepto* $c \in \mathcal{C}$ *se etiqueta con un par* $[\text{tipo}(c), \text{ref}(c)] \in \mathcal{T}_\mathcal{C} \times (\mathcal{I} \cup \{*\})$*.*
  - *Un nodo relación* $r \in \mathcal{R}$ *se etiqueta mediante* $\text{tipo}(r) \in \mathcal{T}_\mathcal{R}$*, y su valencia debe ser igual a la aridad de* $\text{tipo}(r)$*.*
  - *Una arista* $a \in \mathcal{A}$*, etiquetada mediante* $i \in \mathbb{N}$*, que conecta un nodo* $r \in \mathcal{R}$ *con un nodo* $c \in \mathcal{C}$*, se denota por* (r, i, c)*. Las aristas* $(r, 1, c_1), \cdots, (r, k, c_k)$ *que inciden sobre* r *son totalmente ordenados y se etiquetan de* 1 *a la aridad de* $\text{tipo}(r)$*. Generalmente, se emplea la notación* $r = \text{tipo}(r)(c_1, \cdots, c_k)$*.*

■

Intuitivamente, un GCB se puede ver como un grafo bipartito que proporciona un conjunto de punteros semánticos sobre dos jerarquías del dominio de conocimiento, uno para conceptos y otro para las relaciones entre estos conceptos. Los conceptos describen a los referentes individuales en el soporte, a los que ahora hemos vinculado con un tipo conceptual. En pocas palabras, tenemos una hoja de ruta que refleja la organización de ese dominio como un sistema de memoria declarativa, facilitando tanto la toma de decisiones como el aprendizaje significativo. En este punto, una vez formalizada la estructura que utilizaremos en la representación del conocimiento, podemos ya introducir la *proyección* como mecanismo básico que permitirá capturar la noción de respuesta a una consulta.

Definición 3 *Sean $\mathcal{G}_1 = (\mathcal{C}_1, \mathcal{R}_1, \mathcal{A}_1, \mathcal{E}_1)$ y $\mathcal{G}_2 = (\mathcal{C}_2, \mathcal{R}_2, \mathcal{A}_2, \mathcal{E}_2)$ dos* GCB*'s definidos sobre un soporte $\mathcal{S} = (\mathcal{T}_\mathcal{C}, \mathcal{T}_\mathcal{R}, \mathcal{I})$. Una* proyección *de $\mathcal{G}_1$ en $\mathcal{G}_2$ es una correspondencia π de $\mathcal{C}_1$ en $\mathcal{C}_2$, y de $\mathcal{R}_1$ en $\mathcal{R}_2$ que verifica:*

$$(r, i, c) \in \mathcal{A}_1 \Rightarrow (\pi(r), i, \pi(c)) \in \mathcal{A}_2 \quad \text{y} \quad x \in \mathcal{C}_1 \cup \mathcal{R}_1 \Rightarrow \mathcal{E}_2(\pi(x)) \leq \mathcal{E}_1(x)$$

donde, si $x \in \mathcal{C}_1$, $\leq$ hace referencia al producto cartesiano del orden en $\mathcal{T}_\mathcal{C}$ y $\mathcal{I} \cup \{*\}$[19]*. En el caso de que $x \in \mathcal{R}_1$, entonces $\leq$ hace referencia al orden de $\mathcal{T}_\mathcal{R}$.*

Del mismo modo, se dice que $\mathcal{G}_1$ es el origen y que $\mathcal{G}_2$ es el destino, pero también se dice que $\mathcal{G}_1$ subsume *a $\mathcal{G}_2$ o que $\mathcal{G}_1$ es* más general *que $\mathcal{G}_2$, usando la notación $\mathcal{G}_1 \succeq \mathcal{G}_2$. El conjunto de proyecciones de $\mathcal{G}_1$ en $\mathcal{G}_2$ se denota por* proy*$(\mathcal{G}_1, \mathcal{G}_2)$.*

■

Intuitivamente, una *proyección* es un *homomorfismo*[20], que permite especializar las etiquetas de los nodos conceptos y de los relacionales. Por lo tanto, la existencia de una proyección de un GCB $\mathcal{Q}$ sobre otro $\mathcal{D}$ significa que el conocimiento representado por $\mathcal{Q}$ está contenido en el conocimiento representado por $\mathcal{D}$.

Teorema 1 *Sea $\mathcal{G}_1 = (\mathcal{C}_1, \mathcal{R}_1, \mathcal{A}_1, \mathcal{E}_1)$ y $\mathcal{G}_2 = (\mathcal{C}_2, \mathcal{R}_2, \mathcal{A}_2, \mathcal{E}_2)$ dos* GCB*'s definidos sobre un $\mathcal{S} = (\mathcal{T}_\mathcal{C}, \mathcal{T}_\mathcal{R}, \mathcal{I})$, entonces $\mathcal{G}_1 \succeq \mathcal{G}_2$ si y sólo si $\exists\pi$, una proyección de $\mathcal{G}_1$ en $\mathcal{G}_2$.*

Demostración: Trivial a partir de la Definición 3.

■

4.2 | El problema de las respuestas a consultas

Estamos ahora en disposición de reescribir el problema de las respuestas a consultas, que toma como entrada una *colección documental $\mathcal{D}$* compuesta de GCB's que representan hechos, un GCB $c \in \mathcal{Q}$ que representa a una consulta integrada dentro de una colección $\mathcal{Q}$, y preguntas para todas las respuestas de $c \in \mathcal{Q}$ en $\mathcal{D}$. Por lo tanto, cada proyección de c con un hecho lleva a una respuesta, o, como veremos, c es deducible a partir de la colección documental $\mathcal{D}$. Como paso preliminar a la formalización de este proceso, estableceremos una correspondencia semántica Φ que asigne una fórmula en LPO $\Phi(\mathcal{G})$ a cada GCB $\mathcal{G}$ [139] definido sobre el soporte $\mathcal{S} = (\mathcal{T}_\mathcal{C}, \mathcal{T}_\mathcal{R}, \mathcal{I})$, donde $\Phi(\mathcal{G})$ es una fórmula positiva, conjuntiva y cerrada existencialmente. En otras palabras, Φ asigna un conjunto de fórmulas $\Phi(\mathcal{S})$ sobre un soporte $\mathcal{S}$, lo que se corresponde con una interpretación de orden parcial de $\mathcal{T}_\mathcal{R}$ y $\mathcal{T}_\mathcal{C}$. Para todo tipo t y t', tal que $t \geq t'$, se tiene la siguiente fórmula:

$$\forall\, C_1, \ldots, C_k,\ t'(C_1, \ldots, C_k) \rightarrow t(C_1, \ldots, C_k)$$

[19] esto es, $[\mathit{tipo}(\pi(x)), \mathit{ref}(\pi(x))] \leq [\mathit{tipo}(x), \mathit{ref}(x)]$ si y sólo si $\mathit{tipo}(\pi(x)) \leq \mathit{tipo}(x)$ y $\mathit{ref}(\pi(x)) \leq \mathit{ref}(x)$.

[20] es un morfismo que preserva las aristas.

donde $k = 1$ para los tipos conceptuales, y en cualquier otro caso k es la aridad de los tipos relacionales. Esto implica que las consultas y los documentos pueden ser interpretados como fórmulas lógicas, y que el proceso de búsqueda se corresponde con un proceso de inferencia lógica. Dicho esto, ya estamos en condiciones de razonar en base a los conocimientos representados mediante grafos en la colección documental y en las consultas.

Teorema 2 *(Suficiencia y completitud) Sean $d \in \mathcal{D}$ y $c \in \mathcal{Q}$ dos* GCB*'s definidos sobre el soporte $\mathcal{S}$, entonces*

$$c \succeq \text{nf}(d) \Leftrightarrow \Phi(\mathcal{S}),\ \Phi(d) \models \Phi(c)$$

donde $\models$ denota la deducción en LPO*; y* nf(d) *es la* forma normal *de d, a saber, aquélla que se obtiene fusionando los nodos concepto con mismo referente individual*[21]*. En definitiva, se trata de aplicar la operación binaria de ligadura externa.*

Demostración: Ver [106].

■

Se puede demostrar que la generación de respuestas a consultas mediante GCB's en el marco descrito es un problema NP-*completo* [24]. En este sentido, el problema de la decisión[22] se puede resolver en un tiempo polinómico [25, 70], dando un sentido computaciontal a nuestro planteamiento.

4.2.1 | Tipos de respuestas

Desde un punto de vista práctico hemos de dotar, además, a las proyecciones de la flexibilidad necesaria para la localización de respuestas cuya estructura no se corresponda exactamente con la proyección de la correspondiente pregunta. En este sentido, será necesario organizar la búsqueda de secuencias de *transformaciones* que permitan a la pregunta o a la colección documental relajar sus estructuras de forma tal que dicha proyección sea posible.

Definición 4 *Sean $d, d' \in \mathcal{D}$ y $c \in \mathcal{Q}$, tres* GCB*'s definidos sobre un soporte $\mathcal{S}$, y ς una correspondencia del conjunto de* GCB*'s definidos sobre $\mathcal{S}$ en él mismo, tal que $\varsigma(d) = d'$. Si $\pi \in$ proy(c, d'), entonces (π, ς) es una* proyección de c en d modulo ς.

■

Intuitivamente, la idea es la de proveer un conjunto de transformaciones que permitan determinar la pertinencia de un documento en relación a una pregunta,

[21] esto es, un GCB está en forma normal si cada referente individual con un tipo conceptual aparece una única vez en él.

[22] esto es, saber si es resoluble, no o simplemente es no decidible.

cuando la información contenida en ambos guarde algún tipo de relación. Formalmente consideraremos tres mechanismos de transformación aplicables a un GCB.

Definición 5 *Sea $\mathcal{G} = (\mathcal{C} \cup \mathcal{R}, \mathcal{A}, \mathcal{E})$ un* GCB *definido sobre un soporte $\mathcal{S} = (\mathcal{T_C}, \mathcal{T_R}, \mathcal{I})$. Una* sustitución *en $\mathcal{G}$ es un par $(t, t') \in (\mathcal{C} \times (\mathcal{T_C} \times (\mathcal{I} \cup \{*\}))) \cup (\mathcal{R} \times \mathcal{T_R})$. Si se puede afirmar que un término concepto (resp. relación) t puede ser sustituido por uno t', se dice que (t, t') son* términos compatibles.

■

Definición 6 *Sea $\mathcal{G} = (\mathcal{C} \cup \mathcal{R}, \mathcal{A}, \mathcal{E})$ un* GCB *definido sobre un soporte $\mathcal{S} = (\mathcal{T_C}, \mathcal{T_R}, \mathcal{I})$. El resultado de aplicar una* unión de los conceptos $c, c' \in \mathcal{T_C}$, *tal que $\mathcal{E}(c) = \mathcal{E}(c')$, es el* GCB *obtenido a partir de $\mathcal{G}$ mediante la identificación de c y c'.*

■

Como una unión puede cambiar sustancialmente la estructura de un GCB, esta transformación se considera que provoca más distanciamiento que las sustituciones.

Definición 7 *Sea $\mathcal{G} = (\mathcal{C} \cup \mathcal{R}, \mathcal{A}, \mathcal{E})$ un* GCB *definido sobre un soporte $\mathcal{S} = (\mathcal{T_C}, \mathcal{T_R}, \mathcal{I})$. El resultado de* agregar un nodo $n \in \mathcal{C} \cup \mathcal{R}$, *tal que $\mathcal{E}(n) = v$, es el nuevo* GCB *$\mathcal{G} + \mathcal{N}$, donde $\mathcal{N}$ es el grafo reducido a n. Si $n \in \mathcal{R}$, entonces es necesario especificar sus aristas vecinas.*

■

Dado que una agregación no sólo varía la estructura del GCB original, sino que además introduce un elemento externo al mismo, esta transformación se considera más compleja que una unión y, en consecuencia, también posee un impacto mayor que el de una sustitución. Por otra parte, y en función de la necesidad o no de combinar las transformaciones definidas, se pueden considerar cuatro posibles tipos de respuestas a una pregunta dada, que introducimos de forma incremental en consideración a la complejidad de su proceso de cálculo. En este sentido, las respuestas más simples serán aquéllas cuyo contenido se refiere de forma exacta a la interrogación planteada.

Definición 8 *Sean $d \in \mathcal{D}$ y $c \in \mathcal{Q}$ dos* GCB*'s definidos sobre un soporte $\mathcal{S}$. Entonces d es una* respuesta exacta de c *si y sólo si* proy$(c, d) \neq \emptyset$.

■

A menudo la ausencia de una respuesta exacta es previsible, bien por la falta de información específica en la base de datos documental, bien por la falta de concretud de la propia pregunta. En el primer caso, hablaremos de *incompletitud documental* y

en el segundo de *ambigüedad de la consulta*. Con el fin de tratar estos casos, primero tenemos que capturar formalmente la noción de respuesta no exacta y situarla en el marco ya definido para los GCB's. A este respecto, en esta tesis adoptamos la estrategia de búsqueda descrita en [57], a su vez inspirada en la implementación de la *segunda forma del Principio de incertidumbre de van Rijsbergen's* [151] propuesto en [79]:

"Sean $d \in \mathcal{D}$ y $c \in \mathcal{Q}$ dos proposiciones, una medida de incertidumbre relativa de $d \rightarrow c$ a una base de conocimiento está determinada por la transformación mínima de d en d', tal que se verifique $d' \rightarrow c$."

donde, en nuestro caso, la transformación de d en d' está basada en las operaciones de grafos, que podrían también ser usadas para transformar una consulta c. En este sentido, cabría preguntarse por qué no transformar c en c' con el fin de conseguir verificar que $d \rightarrow c'$. Con respecto a esto, se puede ver que $d' \rightarrow c$ se verifica si y sólo si $d \rightarrow c'$, donde c' se obtiene a partir de c mediante una transformación dual de una transformación de d en d'. La ventaja usualmente argumentada [57] para modificar la colección documental $\mathcal{D}$ en lugar de las preguntas $\mathcal{Q}$, es que los contenidos de la primera pueden enriquecerse mediante relevancia retroalimentada por el sistema de RI. En cualquier caso, ello permite establecer el marco formal que necesitábamos para flexibilizar el protocolo de interrogación antes introducido en los GCB's. Comenzaremos por describir el caso más simple. Se trata de las *respuestas aproximadas*.

Definición 9 *Sean $d \in \mathcal{D}$ y $c \in \mathcal{Q}$ dos* GCB*'s definidos sobre un soporte $\mathcal{S}$. Entonces d es una* respuesta aproximada de c *si y sólo si existe una secuencia de sustituciones ς, tales que* $\text{proy}(c, \varsigma(d)) \neq \emptyset$.

■

Intuitivamente, para calcular una respuesta aproximada, la estructura del GCB inicial d se ve ligeramente modificada. Dado que las respuestas exactas son un tipo particular de las aproximadas y que constituyen un fenómeno raro sin casi interés práctico, en adelante sólo hablaremos de este tipo de respuestas para referirnos a ambas categorías, exactas y aproximadas. Con el fin de ampliar el grado de flexibilidad asociados a las consultas, aumentaremos el umbral de las transformaciones estructurales permitidas, por ejemplo, incluyendo las uniones. Esto permite definir las *respuestas plausibles*.

Definición 10 *Sean $d \in \mathcal{D}$ y $c \in \mathcal{Q}$ dos* GCB*'s definidos sobre un soporte $\mathcal{S}$. Se dice que una* secuencia ς de sustituciones y uniones es aceptable *si y sólo si ς no contiene demasiadas uniones en relación al número de nodos en c. La proporción de uniones permitidas (μ_u) se establece por el usuario.*

■

Definición 11 *Sean $d \in \mathcal{D}$ y $c \in \mathcal{Q}$ dos* GCB*'s definidos en un soporte $\mathcal{S}$. Se dice que d es una* respuesta plausible *a c si y sólo si existe una secuencia aceptable ς de sustituciones y uniones, tal que* $\text{proy}(c, \varsigma(d)) \neq \emptyset$.

■

Para completar la oferta relacionada con las consultas, incluimos finalmente las agregaciones de nodos. Aunque esto no permite cubrir totalmente el abanico de transformaciones para grafos, sí se centra en aquellas interrogaciones cuyo impacto es menor en lo que a la intención inicial expresada por el usuario se refiere. Se trata de las *respuestas parciales*.

Definición 12 *Sean $d \in \mathcal{D}$ y $c \in \mathcal{Q}$ dos* GCB*'s definidos sobre un soporte $\mathcal{S}$. Se dice que una* secuencia ς de sustituciones, uniones y agregaciones de nodos es aceptable *si y sólo si ς es aceptable para las uniones y no existen demasiados nodos añadidos en relación al número de nodos de c. La proporción de nodos agregados permitidos (μ_a) se establece por el usuario.*

■

Definición 13 *Sean $d \in \mathcal{D}$ y $c \in \mathcal{Q}$ dos* GCB*'s definidos sobre un soporte $\mathcal{S}$. Se dice que d es una* respuesta parcial *a c si y sólo si existe una secuencia aceptable ς de sustituciones, uniones y agregaciones de nodos, tal que* $\text{proy}(c, \varsigma(d)) \neq \emptyset$.

■

Las herramientas formales que venimos de introducir definen un entorno de trabajo basado en GCB's que nos permitirá, por un lado, representar el conocimiento contenido en una colección documental y, por otro, extraer éste en razón de un patrón determinado al que hemos dotado de cierta flexibilidad. En este punto, la puesta en marcha de un sistema de RI requiere, además, de la disponibilidad de un mecanismo efectivo de generación de estos GCB's.

4.3 | La función de ordenación

Una vez formalizado el problema de las respuestas a consultas, necesitamos integrar una estrategia de ordenación como último paso para completar el diseño de nuestra arquitectura de RI conceptual. Con este propósito, la utilización de GCB's como términos de indexación nos permite situar de forma natural la pregunta en el dominio de las funciones basadas en subsunción y en instancias. En este punto, aunque los enfoques basados en CMAC's tienen el potencial suficiente para convertirse en un medio de clasificación poderoso, padecen en la práctica de carencia de eficiencia computacional, debido a su alto coste. Como alternativa, Genest [56] amplía la gama de relaciones

conceptuales para conseguir técnicas más flexibles y menos ambiciosas, buscando un compromiso entre la eficiencia y el poder de discriminación. Por este motivo, el autor introduce las funciones de ordenación como simples órdenes parciales en el conjunto de transformaciones aplicadas a una consulta para alcanzar una proyección sobre la colección documental, es decir, para obtener una respuesta.

Definición 14 *Dado un soporte $\mathcal{S}$, sean $\mathcal{Q}$, $\mathcal{D} = \{d_i\}_{i \in I}$ los* GCB*'s asociados a una consulta y a una colección documental, y sea $\mathcal{R}_{\mathcal{Q}}^{\mathcal{D}}$ la colección de respuestas obtenidas mediante un conjunto $\mathcal{T}_{\mathcal{Q}}^{\mathcal{D}}$ de secuencias de transformaciones sobre grafos aplicadas en $\mathcal{Q}$ para obtener una proyección en algún $d_i,\ i \in I$. Se define una* función de ordenación asociada a $\mathcal{Q}$ y $\mathcal{D}$ *como la ordenación inducida naturalmente en $\mathcal{R}_{\mathcal{Q}}^{\mathcal{D}}$ mediante cualquier orden parcial de $\mathcal{T}_{\mathcal{Q}}^{\mathcal{D}}$.*

■

Este enfoque generaliza a los basados en CMAC's, al tiempo que nos permite flexibilizar las restricciones computacionales. En la práctica, nos centraremos concretamente en el orden parcial introducido por Genest en [56].

Definición 15 *Dado un soporte $\mathcal{S}$, sean $\mathcal{Q}$, $\mathcal{D} = \{d_i\}_{i \in I}$ los* GCB*'s asociados a una consulta y a una colección documental, y sea $\mathcal{R}_{\mathcal{Q}}^{\mathcal{D}}$ la colección de respuestas obtenidas mediante un conjunto $\mathcal{T}_{\mathcal{Q}}^{\mathcal{D}}$ de secuencias de transformaciones sobre grafos aplicadas en $\mathcal{Q}$ para obtener una proyección en algún $d_i,\ i \in I$. Se define el* orden parcial de Genest sobre los elementos $t,\ t' \in \mathcal{T}_{\mathcal{Q}}^{\mathcal{D}}$ *de la siguiente manera:*

$$t <_G t' \text{ si y sólo si } \begin{cases} t' & \text{asocia una respuesta aproximada } \textbf{OR} \\ t & \text{asocia una respuesta parcial } \textbf{OR} \\ t \text{ (resp. } t') & \text{asocia una respuesta parcial (resp. plausible) } \textbf{OR} \\ t,\ t' & \text{asocia el mismo tipo de respuesta } \textbf{AND } |t| > |t'| \end{cases}$$

mientras que

$$t =_G t' \text{ si y sólo si } t \textbf{ AND } t' \text{ asocian el mismo tipo de respuesta, } \textbf{AND } |t| = |t'|$$

■

Intuitivamente esto implica que cualquier respuesta aproximada es considerada más relevante que una plausible, y éstas, a su vez, son consideradas más relevantes que las parciales. Si consideramos un mismo tipo de respuestas, la relevancia es inversamente proporcional al número de transformaciones individuales aplicadas[23]. Desde un punto de vista teórico, esto sigue siendo consistente con respecto a las consideraciones realizadas anteriormente sobre el impacto estructural en los GCB's debido a la aplicación de sustituciones, uniones o agregaciones. A pesar de su simplicidad, esta técnica ha demostrado aparentemente ser superior a las más recientes y sofisticadas [119], lo cual justifica su revisión y consideración formal.

[23] esto es, sustituciones, uniones y agregaciones de nodos.

5 | Adquisición de conocimiento

El objetivo ahora es hilvanar un protocolo de actuación que permita extraer de forma automática el conocimiento atesorado en el texto. En este sentido, nuestra propuesta recurre al encadenamiento de herramientas de análisis léxico, sintáctico y semántico que desemboquen en la generación automática de GCB's directamente a partir de la colección documental, y donde la intervención del usuario se ha reducido a la mínima expresión. Nuestra contribución en este punto se localiza en la novedad de las arquitecturas elegidas para los módulos de análisis léxico y sintáctico y, especialmente, en la originalidad del diseño de la arquitectura del entorno de análisis semántico propuesto.

5.1 | El marco léxico

Aunque nuestra propuesta no requiere de ningún entorno específico de análisis léxico, la elección a efectos de implementación recae en una cadena de procesamiento basada en la arquitectura *Alexina* [120], en esencia una propuesta para el tratamiento léxico digital y su adquisición. Dado que nuestro *corpus* de ejemplo está en francés, hacemos uso de un recurso básico que no es otro que un léxico morfológico y sintáctico a gran escala en ese idioma, denominado LEFFF [121], el cual incluye información originada a partir de diferentes trabajos. Podemos referirnos aquí a la adquisición automática gracias a técnicas estadísticas sobre *corpora* sin tratar, a la adquisición automática de información sintáctica específica o a la corrección manual y la extensión guiada por técnicas automáticas.

Con el fin de proporcionar un tratamiento pre-sintáctico, hemos utilizado una arquitectura denominada SXPIPE [122] que transforma el texto bruto en un GAD capaz de tratar diversos fenómenos que ocurren con una gran frecuencia en *corpora* reales. Esto incluye reconocimiento de varias familias de entidades nombradas, detección de errores ortográficos, tratamiento de ambigüedades durante la separación de cadenas de caracteres en la fase de segmentación de la frase o de la cadena, y tratamiento de ambigüedades léxicas entre palabras que sólo difieren en signos diacríticos o en su capitalización. El analizador léxico utilizado se basa en una morfología de estados finitos que utiliza SXPIPE para preprocesar la entrada, y combina su salida con información léxica extraída a partir del LEFFF.

Independientemente de cual sea el marco léxico que se considere, su salida debe incluir todas las posibles categorías léxicas para la ocurrencia de una forma y que se denota con fines descriptivos de la siguiente manera, introduciendo algunos detalles estructurales adicionales a fin de integrar en adelante datos semánticos.

Definición 16 *Sean* $\{s_i\}_{1\leq i\leq n}$ *la secuencia de frases de un* corpus $\mathscr{C}$ *y* $\Theta_{i,j}$, $1 \leq j \leq |s_i|$ *la ocurrencia de una* forma *en la* j*-ésima posición de la frase* s_i*. Se denota la asociación de una categoría léxica* (a) *y una clase semántica* (b) *con esa forma* $\Theta_{i,j}$*, por* $\Theta_{i,j}^{a,b}$*, y la denominamos* término.

Del mismo modo, se introduce una notación utilizando una variable anónima, $\Theta^{a,\_}_{i,j}$*, denominada* token*, con el fin de designar al conjunto de términos sólo diferenciables por su clase semántica. En ese sentido, también se denota por* $\Theta^{\_,\_}_{i,j}$ *el conjunto de tokens referidos a la misma ocurrencia de una forma, denominada* agrupación*.*

Finalmente, se considera una notación mediante la utilización de variables libres, empleando para ello letras mayúsculas del final del abecedario, con el fin de enumerar rangos de valores. Así, por ejemplo, $\Theta^{a,X}_{i,j}$ *se refiere al conjunto de términos en el token* $\Theta^{a,\_}_{i,j}$*, cuya clase semántica* X *sea aplicable en ese contexto. Además, esta notación puede ser extendida de un modo natural tanto a los tokens como a las agrupaciones.*

■

Con el fin de clarificar estos conceptos, la Fig. 1.1 los ilustra en relación a la frase del *corpus* $\mathscr{B}$*, «feuilles à nervures denticulées»* («hojas con nervaduras dentadas»). Aquí los términos se identifican con triángulos, mientras que los tokens lo hacen con elipses y las agrupaciones con rectángulos. Las clases semánticas asociadas a los términos en esta figura son algunas de las recogidas en la Tabla 3.

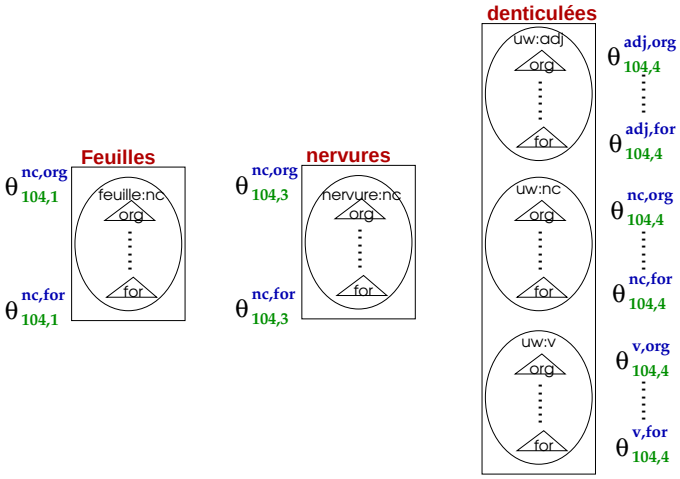

Figura 1.1: Notación léxica

5.2 | El marco sintáctico

Desde un punto de vista descriptivo, nuestra elección recae en las *gramáticas de adjunción de árboles* (GA's) [77], un formalismo suavemente dependiente del contexto que ha ganado gran popularidad en el ámbito del PLN por tres razones fundamentales. La primera, un *dominio de localidad extendido* que permite definir dependencias sintácticas a cualquier nivel. La segunda, la posibilidad de considerar dependencias cruzadas. La tercera, la extensión natural del modelo independiente del contexto clásico, al pasar la unidad básica de reescritura del símbolo al árbol.

5.2.1 | Análisis suavemente dependiente del contexto

Desde el punto de vista computacional, el esquema de análisis sintáctico escogido [3] posee una complejidad temporal (resp. espacial) de orden $\mathcal{O}(n^6)$ (resp. $\mathcal{O}(n^3)$) para un texto de longitud n. Como implementación concreta nos inclinamos por *DyALog* [154]. Ello permite un alto grado de abstracción en el diseño gramatical mediante la consideración del concepto de *metagramática* [155], introduciendo un diseño jerárquico que, además, da cabida a un mecanismo de herencia que simplifica la tarea lingüística. De esta forma la descripción gramatical puede ser refinada progresivamente, facilitanto no sólo su diseño sino también su mantenimiento.

En cuanto al tratamiento de las ambigüedades, una implementación en programación dinámica conlleva la compartición óptima de cálculos y estructuras de representación, derivando en una gestión computacionalmente eficaz del no determinismo. Evitamos así la eliminación de interpretaciones en los procesos de análisis léxico y sintáctico, retrasando la toma de decisiones en este sentido hasta el análisis semántico, cuando dispongamos de toda la información asociada al *corpus* analizado. Hace posible, además, una explotación eficiente del fenómeno de *determinismo local*[24], lo que en la práctica supone que en la medida de lo posible el proceso sea *de facto* determinista y, en consecuencia, con una complejidad espacial y temporal lineal.

5.2.2 | El análisis

Necesitamos que el análisis se resuma en un GDGG que compile las relaciones semánticas iniciales del texto analizado y que han sido capturadas por el analizador sintáctico en un GID. Intuitivamente, se trata de que en este tipo de relaciones el núcleo de un sintagma gobierne a sus modificadores, tal como muestra la Fig. 1.2 mediante líneas discontinuas que conectan los nodos implicados en cada caso. Podemos también observar el impacto que las ambigüedades tanto de tipo léxico como sintáctico generan en el número de posibles dependencias que han de pasar a la posterior fase de análisis semántico. En el primer caso, resulta clara su multiplicación en relación al número de tokens en una misma agrupación, esto es, al número de categorías léxicas asignables a una forma en una posición dada de una frase concreta del *corpus*. En el segundo caso podemos observar un efecto análogo como resultado de la multiplicación de dependencias sobre los modificadores. Es el caso de *«denticulées»* («`dentadas`») como modificador bien de *«feuilles»* («`hojas`») o bien de *«nervures»* («`nervaduras`») en la Fig. 1.2. Se trata en este caso de un conocido fenómeno ligado a la asociación de complementos preposicionales a un sintagma nominal, y que aquí proporciona dos posibles interpretaciones para la frase en este nivel: «`hojas con -nervaduras dentadas-`» o, alternativamente, «`-hojas dentadas- con nervaduras`».

Con respecto a esto, mientras las ambigüedades léxicas sólo dependen de

[24] las ambigüedades de los textos escritos en LN poseen un ámbito de influencia local, disipándose a medida que avanzamos en la lectura. De no ser así, la comunicación entre humanos no sería posible.

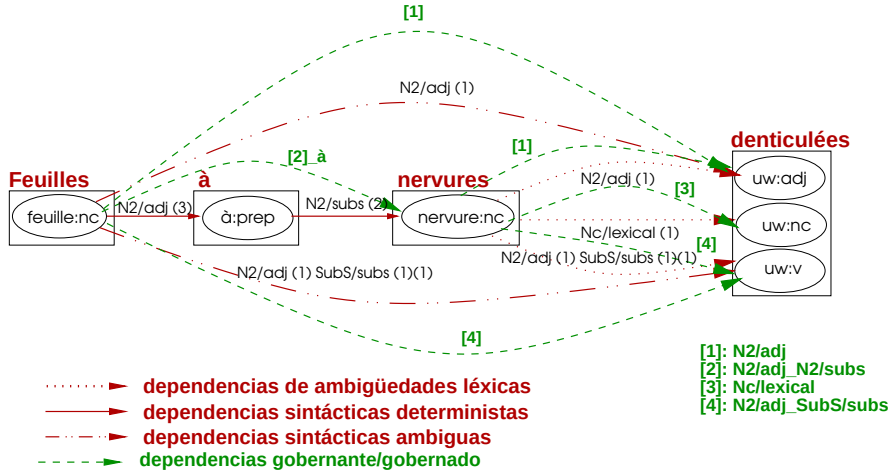

Figura 1.2: Dependencias gobernante/gobernado

la estructura del lenguaje, las sintácticas están fuertemente influenciadas por el formalismo gramatical elegido para describirlo, por la gramática particular considerada y por la falta de una cobertura gramatical completa. Existen incluso no pocas situaciones en las que las ambigüedades han de resolverse forzosamente a nivel semántico, toda vez que su origen puede no ser ni de naturaleza léxica ni sintáctica. Un ejemplo clásico es el uso de estructuras de coordinación relacionando entidades con una lista de adjetivos [118], como en la frase *«des sépales ovales-aigus, glabres ou éparsement hérissés»* (`«sépalos ovalados-agudos, glabros o dispersamente espinosos»`), donde la propiedad *«hérissés»* (`«espinosos»`) se podría unir al adjetivo *«glabres»*(`«glabros»`) o a *«ovales-aigus»* (`«ovalados-agudos»`). En este caso, sólo hay una forma de resolver el problema, y pasa por conocer la naturaleza exacta de los órganos de las plantas, algo que nada tiene que ver ni con la morfología ni con la gramática del lenguaje.

En este sentido, el fenómeno de la ambigüedad puede entenderse como una ilustración de la complejidad del lenguaje en sí mismo [113], siendo éste un problema fundamental a resolver en el PLN. En estas condiciones, es difícil estimar el conjunto de esquemas sintácticos asociados al no determinismo, lo cual podría complicar un acercamiento analítico para resolver el problema. Afortunadamente, existe una condición topológica que resulta ser fácilmente detectable y que lo caracteriza completamente en grafos de dependencias, independientemente de su origen. De un modo más detallado, una ambigüedad se corresponde con una situación donde un token gobernado tiene más de un gobernante. Esto proporciona, a su vez, un mecanismo sencillo para solucionar la cuestión, a saber, se trata de filtrar las dependencias menos plausibles en favor de las que lo son más, asegurando de este modo que un token gobernado tenga únicamente un gobernante.

Sin embargo, la materialización de esta idea no resulta ser tan sencilla. La mayoría de

las ambigüedades pasan inadvertidas, ya que los humanos somos muy hábiles a la hora de resolverlas gracias a un amplio conocimiento del contexto y del mundo, mientras que los sistemas informáticos no tienen plena capacidad en ese terreno. Como consecuencia, a menudo no realizan un buen trabajo de desambiguación y todos los esfuerzos para resolver computacionalmente el problema se centran en explorar los contextos del discurso y explotar los recursos basados en conocimiento [147].

5.3 | El marco semántico

Construido el GDGG, probablemente reflejando toda una variedad de ambigüedades léxicas y sintácticas, ahora lo que queremos es priorizar estas relaciones para extraer de forma efectiva la semántica del texto. Intuitivamente, el proceso consistirá en recopilar información a partir del *corpus* con el objetivo de detectar aquéllas que resulten más plausibles. Técnicamente, la heurística propuesta se organiza en tres niveles de complejidad. Los dos primeros están concebidos para explotar la secuencia de estructuras resultantes de las fases previas de análisis léxico y sintáctico, clasificando en orden de prioridad las ambigüedades correspondientes. El tercer nivel determinará que información semántica está involucrada en cada una de las dependencias.

Para conseguir este objetivo, es necesario introducir una notación específica, ya que deberemos extrapolar nuestras estimaciones desde un contexto local hacia uno global. Así, los datos obtenidos inicialmente de las frases deben ser combinados y evaluados a lo largo de todo el *corpus* con el fin de extraer nuevas conclusiones susceptibles de ser de nuevo aplicadas en cada frase, para luego recomenzar iterativamente el proceso. Deberíamos entonces hablar de *términos*, *tokens* y *agrupaciones plausibles*, nociones que extenderán los conceptos del mismo nombre desde el nivel local a uno de *corpus*.

Definición 17 *Sean* $\{s_i\}_{1\leq i\leq n}$ *la secuencia de frases de un* corpus $\mathscr{C}$ *y* $\Theta_{i,j}$, $1 \leq j \leq |s_i|$ *la ocurrencia de una forma en la* j*-ésima posición de la frase* s_i*. Se denota la asociación de la categoría léxica* (a) *y la clase semántica* (b) *con esa forma* $\Theta_{i,j}$*, por* $\tilde{\Theta}_{i,j}^{a,b}$*, llamado* término plausible*.*

Esta notación puede ser extendida aquí explotando la utilización de las variables anónimas (resp. las variables libres) previamente introducidas para términos, tokens y agrupaciones en la Definición 16.

■

Será necesario igualmente proveernos de la notación necesaria para la gestión de dependencias gobernante/gobernado a nivel de frase (resp. de *corpus*). A este respecto, habremos de referirnos tanto a las transiciones entre tokens (resp. tokens plausibles) que constituyen la salida proporcionada en los GID's por el analizador sintáctico, como a los conjuntos de transiciones entre tokens de dos agrupaciones (resp.

agrupaciones plausibles) diferentes. Finalmente, ya en la fase de categorización semántica consideraremos el tratamiento de transiciones entre términos (resp. términos plausibles).

Definición 18 *Sea* $s_i,\ 1 \leq i \leq n$ *la* i*-ésima frase de un* corpus $\mathscr{C}$ *y* τ *la secuencia de reglas gramaticales necesarias para generar el token* $\Theta^{c,-}_{i,k}$ *a partir del token* $\Theta^{a,-}_{i,j}$ *en el* GDGG. *Se denota* la dependencia entre los tokens $\Theta^{a,-}_{i,j}$ y $\Theta^{c,-}_{i,k}$, etiquetada por τ *como* $\delta^{\theta^{a,-}_{i,j},\tau,\theta^{c,-}_{i,k}}$.

La notación puede extenderse naturalmente a los términos, agrupaciones y estructuras plausibles mediante la utilización de la notación previamente introducida de las variables anónimas. Cuando una dependencia relaciona estructuras plausibles, se habla de dependencias plausibles.

■

5.3.1 | Categorización de los tokens

El objetivo es calcular, para cada agrupación del texto, cual es el token más probable. Es decir, para cada frase del *corpus*, queremos determinar la categoría léxica de cada una de las ocurrencias de las formas que ahí figuren. El proceso, iterativo, se corresponde con las ecuaciones de la Tabla 1, que pasamos a comentar:

$$P(\Theta^{a,-}_{i,j})_{\text{local(0)}} = \frac{1}{|\{\Theta^{X,-}_{i,j}\}|} \tag{1}$$

$$P(\tilde{\Theta}^{a,-}_{i,j})_{\text{global(n+1)}} = \frac{\sum_{\Theta_{k,l}=\Theta_{i,j}} P(\Theta^{a,-}_{k,l})_{\text{local(n)}}}{\sum_{\Theta^{X,-}_{k,l},\Theta_{k,l}=\Theta_{i,j}} P(\Theta^{X,-}_{k,l})_{\text{local(n)}}} \tag{2}$$

$$P(\Theta^{a,-}_{i,j})_{\text{local(n+1)}} = \frac{P(\tilde{\Theta}^{a,-}_{i,j})_{\text{global(n+1)}}}{\sum_{\Theta^{X,-}_{k,l},\Theta_{k,l}=\Theta_{i,j}} P(\tilde{\Theta}^{X,-}_{k,l})_{\text{global(n+1)}}} \tag{3}$$

Tabla 1: Modelo para la categorización de tokens

(1). El proceso se inicia con el cálculo de la probabilidad local a nivel de frase, asociable a un token en una agrupación. Se trata de un simple *ratio* en razón al número de tokens que involucran a dicha agrupación. Obviamente, si sólo existe un token en la agrupación, su probabilidad será de 1.

(2). Define la probabilidad global en el *corpus* de un token plausible, en la iteración $n+1$ del proceso. Se calcula como una proporción de la probabilidad local asociada a tokens con la misma categoría léxica y forma que la del token considerado, en relación a la probabilidad cuando la categoría léxica es libre.

(3). Establece el valor de la probabilidad local asociable a un token en una agrupación, en la iteración $n+1$ del proceso. Para ello, se repercuten las probabilidades calculadas globalmente, distribuyéndolas proporcionalmente entre las globales de los tokens plausibles asociados a la agrupación.

El proceso iterativo continúa hasta la convergencia [123] sobre un punto fijo, o sobre un umbral prefijado de aproximación.

5.3.2 | Categorización de las dependencias entre tokens

Se trata ahora de dar una medida objetiva de la viabilidad de las dependencias sintácticas generadas por el analizador sintáctico, entre los tokens previamente categorizados. Teniendo en cuenta que la caracterización topológica de la ambigüedad sintáctica significa la existencia de varios tokens gobernantes para un mismo gobernado, determinado éste buscaremos definir cual es su gobernante de entre los posibles propuestos por el analizador, con el fin de eliminar dicha ambigüedad. De nuevo consideraremos una estrategia iterativa, en este caso determinada por las ecuaciones de la Tabla 2, que describimos a continuación:

$$W(\delta^{\Theta^{a,\_}_{i,j},\tau,\Theta^{b,\_}_{i,k}}) = \frac{|S \overset{*}{\Rightarrow} \Theta^{a,\_}_{i,j} \overset{\tau}{\Rightarrow} \Theta^{b,\_}_{i,k}|}{\sum_{\delta^{\Theta^{Y,\_}_{i,X},T,\Theta^{Z,\_}_{i,k}}} |S \overset{*}{\Rightarrow} \Theta^{Y,\_}_{i,X} \overset{T}{\Rightarrow} \Theta^{Z,\_}_{i,k}|} \tag{4}$$

$$P(\delta^{\Theta^{a,\_}_{i,j},\tau,\Theta^{b,\_}_{i,k}})_{\text{local(0)}} = \frac{P(\Theta^{a,\_}_{i,j})_{\text{local}} \cdot\ P(\Theta^{b,\_}_{i,k})_{\text{local}} \cdot\ W(\delta^{\Theta^{a,\_}_{i,j},\tau,\Theta^{b,\_}_{i,k}})}{\sum_{\Theta^{Y,\_}_{i,X},\Theta^{Z,\_}_{i,k},\delta^{\Theta^{Y,\_}_{i,X},T,\Theta^{Z,\_}_{i,k}}} P(\Theta^{Y,\_}_{i,X})_{\text{local}} \cdot\ P(\Theta^{Z,\_}_{i,k})_{\text{local}} \cdot\ W(\delta^{\Theta^{Y,\_}_{i,X},T,\Theta^{Z,\_}_{i,k}})} \tag{5}$$

$$P(\delta^{\tilde{\Theta}^{a,\_}_{i,j},\tau,\tilde{\Theta}^{b,\_}_{i,k}})_{\text{global(n+1)}} = \frac{\sum_{\Theta_{l,m}=\Theta_{i,j},\Theta_{l,p}=\Theta_{i,k}} P(\delta^{\Theta^{a,\_}_{l,m},\tau,\Theta^{b,\_}_{l,p}})_{\text{local(n)}}}{\sum_{\delta^{\Theta^{Y,\_}_{l,X},T,\Theta^{Z,\_}_{l,p}},\Theta_{l,p}=\Theta_{i,k}} P(\delta^{\Theta^{Y,\_}_{l,X},T,\Theta^{Z,\_}_{l,p}})_{\text{local(n)}}} \tag{6}$$

$$P(\delta^{\Theta^{a,\_}_{i,j},\tau,\Theta^{b,\_}_{i,k}})_{\text{local(n+1)}} = \frac{P(\delta^{\tilde{\Theta}^{a,\_}_{i,j},\tau,\tilde{\Theta}^{b,\_}_{i,k}})_{\text{global(n+1)}}}{\sum_{\delta^{\tilde{\Theta}^{Y,\_}_{l,X},T,\tilde{\Theta}^{Z,\_}_{l,m}},\Theta_{l,m}=\Theta_{i,k}} P(\delta^{\tilde{\Theta}^{Y,\_}_{l,X},T,\tilde{\Theta}^{Z,\_}_{l,m}})_{\text{global(n+1)}}} \tag{7}$$

Tabla 2: Modelo para la categorización de las dependencias entre tokens

(4). Antes de iniciar el proceso iterativo, calcularemos para cada dependencia sintáctica un peso inicial en función de su etiqueta. Buscamos con ello dar protagonismo a

aquellas dependencias compartidas por un mayor número de análisis, de entre las que comparten un mismo token gobernado.

(5). El proceso iterativo se inicia con el cálculo de la probabilidad local, a nivel de frase, asociable a una dependencia sintáctica. Dado que aquellas se caracterizan por sus tokens gobernante y gobernado, y por su etiqueta, haremos depender esta probabilidad de las locales de dichos tokens; y del peso asignado a la etiqueta asociada. Se calcula como una proporción de los valores citados para la dependencia sintáctica considerada, en relación al conjunto de las asociadas a la agrupación del token gobernado.

(6). Define la probabilidad global en el *corpus* de una dependencia plausible en la iteración $n + 1$ del proceso. Se calcula como una proporción de la probabilidad local asociada a dependencias sintácticas coincidentes con la considerada (salvo en la frase que la localiza), en relación al conjunto de las locales asociadas a tokens gobernados también coincidentes con el considerado (salvo en la agrupación que lo localiza).

(7). Establece el valor de la probabilidad local de una dependencia en la iteración $n + 1$ del proceso. Para ello repercutimos las probabilidades calculadas globalmente, distribuyéndolas proporcionalmente entre las globales de las dependencias sintácticas plausibles asociadas a tokens gobernados coincidentes con el considerado (salvo en la agrupación que lo localiza).

Como en el caso de la categorización léxica el proceso itera hasta la convergencia sobre un punto fijo o la aproximación a un umbral prefijado.

5.3.3 | Categorización de las dependencias entre términos

El objetivo en este nivel es determinar las clases semánticas correctas de los tokens que participan en una misma dependencia sintáctica, con el fin de identificar las que unen términos de dos agrupaciones diferentes. Más exactamente, dado un término gobernado, buscamos definir cual es su gobernante a través de las dependencias sintácticas previamente categorizadas.

Definición 19 *Sea* $s_i,\ 1 \leq i \leq n$ *la* i*-ésima frase de un* corpus $\mathscr{C}$*, y* $\mathscr{T}$ *(resp.* $\mathscr{F}$*) el conjunto de clases semánticas (resp. de formas semánticas) asociadas a* $\mathscr{C}$ *(resp. a* $\mathscr{T}$*) por medio de alguna técnica fiable. Se denota por* $\mathscr{F}(b)$ *al subconjunto de formas asociadas a* $b \in \mathscr{T}$*, y se dice que* $\Theta_{i,j}^{a,b},\ 1 \leq j \leq |s_i|$ *es un* término estable *si y sólo si* $b \in \mathscr{T}$ *y* $\Theta_{i,j} \in \mathscr{F}(b)$*.*

■

Intuitivamente, un término es estable cuando tenemos información fidedigna acerca de la correspondencia entre su categoría semántica y su forma. El origen de ésta puede

Entidades	**Lemas** (en francés)
organe	fleur, staminode, tige, feuille, hypanthe, périanthe, rameau, ...
fruit	fruit, samare, drupe, capsule, akène
Propiedades	**Lemas** (en francés)
couleur	verdâtre, violacé, noirâtre, violet, jaunâtre, orange, roux, rose
forme	obconique, oblancéolé, oblong, bifolié, crateriforme, punctiforme, ...
taille	moyen, petit, double, épais, inégal, entier, longue
texture	hispide, bifide, globuleux, coriace, velutineux, gélatineux, barbu
position	antérieur, dessus, voisin, seul, latéral, transversal

Tabla 3: Conjunto $\mathscr{T}$ de clases semánticas iniciales (tipos) para el ejemplo de funcionamiento

ser el propio usuario o algún método considerado plenamente fiable. Nuestra propuesta considera ambos mecanismos [49]. Por un lado, el usuario define el conjunto de clases semánticas. En nuestro *corpus* de ejemplo botánico $\mathscr{B}$ éstas se organizan en entidades ($\mathscr{E}$) y propiedades ($\mathscr{P}$), de tal manera que dichas propiedades proporcionen información acerca de los atributos aplicables a las entidades; y complementados por un conjunto asociado de formas iniciales, tales como las que se muestran en la Tabla 3.

Marcador(francés)	**Posición**	**Clase**	**Marcador**(francés)	**Posición**	**Clase**
teinté	[2]	Couleur	épaisseur	[1]	Taille
texture	[2]	Texture	atteindre	[1]	Organe/Fruit
taille	[1]	Organe/Fruit	taille	[2]	Taille
teinte	[1]	Organe/Fruit	teinte	[2]	Couleur
couleur	[1]	Organe/Fruit	couleur	[2]	Couleur
texture	[1]	Organe/Fruit	texture	[2]	Texture
forme	[1]	Organe/Fruit	forme	[2]	Forme
position	[1]	Organe/Fruit	position	[2]	Position
altitude	[1]	Organe/Fruit	environ	[2]	Taille
tache	[1]	Organe/Fruit	tache	[2]	Couleur
longueur	[1]	Taille	formé	[2]	Organe/Fruit
composé	[1,2]	Organe/Fruit	dépassant	[2]	Taille
diamètre	[1]	Taille	contour	[2]	Forme/Texture

Tabla 4: Parte del fichero de colocaciones

Por el otro, el sistema saca ventaja de las *colocaciones*, secuencias de palabras que coocurren con más frecuencia de lo esperado y en las cuales conservan su significado original, al contrario de lo que ocurre con las *locuciones*. La idea es filtrar los análisis con el fin de localizar aquéllas que permitan asociar una forma a una clase semántica. Para la ocasión, las representamos como una tripleta de la forma *marcador/posición/clase semántica*. El marcador sirve para identificar la colocación, para la que la forma indicada por la posición pueda ser asociada a la clase semántica, tal y como se muestra en la Tabla 4, en el caso de nuestro *corpus* de ejemplo $\mathscr{B}$. Así, por ejemplo, en la frase *«teintées de rose»* («`teñidas de rosa`»), la presencia del marcador *«teinté»* («`teñida`») pone en evidencia que *«rose»* («`rosa`») es una instancia de la clase semántica *«couleur»*

(«color»). El proceso iterativo se corresponde con las ecuaciones de la Tabla 5 que ahora describimos.

$$W(\Theta_{i,j}^{a,\_}) > \frac{|\{\Theta_{i,j}^{a,X}\}_{X\in\mathscr{T},\Theta_{i,j}\in\mathscr{F}(X)}|}{|\{\Theta_{i,j}^{a,X}\}_{X\in\mathscr{T}}|} \subseteq (0,1] \tag{8}$$

$$W(\Theta_{i,j}^{a,b}) = \begin{cases} \frac{W(\Theta_{i,j}^{a,\_})}{|\{\Theta_{i,j}^{a,X}\}_{X\in\mathscr{T},\Theta_{i,j}\in\mathscr{F}(X)}|} & \text{si } \Theta_{i,j}\in\mathscr{F}(b) \\ \frac{1-W(\Theta_{i,j}^{a,\_})}{|\{\Theta_{i,j}^{a,X}\}_{X\in\mathscr{T},\Theta_{i,j}\notin\mathscr{F}(X)}|} & \text{en otro caso} \end{cases} \tag{9}$$

$$P(\delta^{\Theta_{i,j}^{a,b},\tau,\Theta_{i,k}^{c,d}})_{\text{local(0)}} = \frac{P(\delta^{\Theta_{i,j}^{a,\_},\tau,\Theta_{i,k}^{c,\_}})_{\text{local}}\cdot W(\Theta_{i,j}^{a,b})\cdot W(\Theta_{i,k}^{c,d})}{\sum_{\Theta_{i,X}^{Y,Z},\Theta_{i,k}^{V,W},\delta^{\Theta_{i,X}^{Y,\_},T,\Theta_{i,k}^{V,\_}}} P(\delta^{\Theta_{i,X}^{Y,\_},T,\Theta_{i,k}^{V,\_}})_{\text{local}}\cdot W(\Theta_{i,X}^{Y,Z})\cdot W(\Theta_{i,k}^{V,W})} \tag{10}$$

$$P(\delta^{\tilde{\Theta}_{i,j}^{a,b},\tau,\tilde{\Theta}_{i,k}^{c,d}})_{\text{global(n+1)}} = \frac{\sum_{\Theta_{l,m}=\Theta_{i,j},\Theta_{l,p}=\Theta_{i,k}} P(\delta^{\Theta_{l,m}^{a,b},\tau,\Theta_{l,p}^{c,d}})_{\text{local(n)}}}{\sum_{\delta^{\Theta_{l,X}^{Y,Z},T,\Theta_{l,p}^{V,W}},\Theta_{l,p}=\Theta_{i,k}} P(\delta^{\Theta_{l,X}^{Y,Z},T,\Theta_{l,p}^{V,W}})_{\text{local(n)}}} \tag{11}$$

$$P(\delta^{\Theta_{i,j}^{a,b},\tau,\Theta_{i,k}^{c,d}})_{\text{local(n+1)}} = \frac{P(\delta^{\tilde{\Theta}_{i,j}^{a,b},\tau,\tilde{\Theta}_{i,k}^{c,d}})_{\text{global(n+1)}}}{\sum_{\delta^{\tilde{\Theta}_{l,X}^{Y,Z},T,\tilde{\Theta}_{l,m}^{V,W}},\Theta_{l,m}=\Theta_{i,k}} P(\delta^{\tilde{\Theta}_{l,X}^{Y,Z},T,\tilde{\Theta}_{l,m}^{V,W}})_{\text{global(n+1)}}} \tag{12}$$

Tabla 5: Modelo para la categorización de las dependencias entre términos

(8). Antes de iniciar el proceso, asociaremos a cada token un peso que verifique la condición expuesta, y cuyo valor justificamos a continuación.

(9). Ahora vamos a distribuir equitativamente el peso calculado a partir de la Ecuación 8 entre los términos estables. Esto asegura que el peso que asociamos aquí a un término no estable en dicho token es inferior al asociado a los otros. Tratamos así de dar inicialmente preferencia a los términos estables.

(10). El proceso iterativo se inicia con el cálculo de la probabilidad local, a nivel de frase, asociable a una dependencia semántica. Dado que ésta queda perfectamente caracterizada por sus términos gobernante y gobernado junto con la dependencia sintáctica entre los tokens asociados a éstos, haremos depender este valor de los pesos asociados a dichos términos, así como de la probabilidad local correspondiente a la dependencia sintáctica. Se calcula como una proporción de los valores citados para la dependencia semántica considerada, en relación al conjunto de las asociadas a la agrupación del término gobernado.

(11). Define la probabilidad global en el *corpus* de una dependencia semántica plausible en la iteración $n+1$ del proceso. Se calcula como una proporción de la probabilidad local asociada a dependencias semánticas coincidentes con la considerada (salvo en la frase que la localiza), en relación al conjunto de las locales asociadas a términos gobernados también coincidentes con el considerado (salvo en la agrupación que lo localiza).

(12). Establece el valor de la probabilidad local asociable a una dependencia semántica en la iteración $n+1$ del proceso. Para ello repercutimos las probabilidades calculadas globalmente, distribuyéndolas proporcionalmente entre las globales de las dependencias semánticas plausibles asociadas a términos gobernados coincidentes con el considerado (salvo en la agrupación que lo localiza).

Como en el caso de la categorización de dependencias sintácticas, el proceso itera hasta la convergencia sobre un punto fijo o la aproximación a un umbral prefijado. A la estructura resultante lo denominamos la *semántica del corpus* $\mathscr{C}$ con el que trabajamos.

Definición 20 *Sean* $\{s_i\}_{1\leq i\leq n}$ *una secuencia de frases de un* corpus $\mathscr{C}$*, y* $\mathscr{T}$ *(resp.* $\mathscr{F}$*) el conjunto de clases semánticas (resp. de formas) asociadas a* $\mathscr{C}$ *(resp. a* $\mathscr{T}$*) por medio de alguna técnica fiable. Se define la* semántica del corpus $\mathscr{C}$ *como*

$$\mathscr{S}_{\mathscr{C}} := \{\delta^{\Theta^{a,b}_{i,j},\tau,\Theta^{c,d}_{i,k}},\ P(\delta^{\Theta^{a,b}_{i,j},\tau,\Theta^{c,d}_{i,k}})_{\textit{local}} = \text{máx}\{P(\delta^{\Theta^{X,Y}_{i,j},Z,\Theta^{V,W}_{i,k}})_{\textit{local}}\}\}$$

donde max *es la función maximal en* $\mathbb{N}$*, y* $\delta^{\Theta^{X,Y}_{i,j},Z,\Theta^{V,W}_{i,k}}$ *son las dependencias calculadas como resultado del proceso de adquisición de conocimiento previamente descrito.*

El concepto puede restringirse naturalmente para referirse a la semántica del documento $\mathscr{D}$ en $\mathscr{C}$ *por*

$$\mathscr{S}^{\mathscr{D}}_{\mathscr{C}} := \{\delta^{\Theta^{a,b}_{i,j},\tau,\Theta^{c,d}_{i,k}} \in \mathscr{S}_{\mathscr{C}},\ s_i \in \mathscr{D}\}$$

■

Intuitivamente, definimos la semántica del *corpus* como el conjunto de las dependencias más probables entre sus términos. Esto compila todas las relaciones sintácticas y semánticas consideradas como viables, entre las categorías léxicas en el texto estudiado. La semántica del *corpus* será el punto de partida para la generación de grafos conceptuales que nos sirven como representación del conocimiento formal para propósitos de RI.

5.4 | Generación de grafos conceptuales

Ahora, estamos listos para estructurar los GCB's que vamos a utilizar en nuestras pruebas experimentales. Aunque la propuesta es independiente del ámbito de

conocimiento considerado, es necesario centrar nuestro trabajo en uno concreto, con el fin de modelizar adecuadamente el soporte sobre el que se definirán los grafos. Como ya se ha comentado, nuestra elección recae en un dominio biológico, y más en concreto en una descripción botánica, para la que tomamos como referencia el *corpus* de ejemplo $\mathscr{B}$. La complejidad, extensión y especificidad de este tipo de contenidos hacen difícil que las consultas sean expresadas por un usuario no experto de otra forma que no sea meramente prospectiva. Se trata, por tanto, de una temática especialmente adecuada para la validación de capacidades en el ámbito del tratamiento de información ambigua e incompleta, lo que justifica nuestra decisión.

En este sentido, retomamos el conjunto de clases semánticas (tipos) $\mathscr{T}$ mostrado en la Tabla 3 para el *corpus* $\mathscr{B}$, con el fin de introducir en él un orden parcial en la forma:

$$\forall\, t \in \mathscr{E} = \{\textit{fruit, organe}\},\ t \leq \varepsilon \leq \top$$

$$\forall\, t \in \mathscr{P} = \{\textit{couleur, forme, taille, texture, position}\},\ t \leq \rho \leq \top$$

donde ε (resp. ρ) es el elemento más grande para las entidades $\mathscr{E}$ (resp. propiedades $\mathscr{P}$). De esta manera, introducimos nuestro soporte de ejemplo $\mathcal{S} = (\mathcal{T}_{\mathcal{C}_{\mathscr{B}}}, \mathcal{T}_{\mathcal{R}_{\mathscr{B}}}, \mathcal{I}_{\mathscr{B}})$ definiendo:

$$\mathcal{T}_{\mathcal{C}_{\mathscr{B}}} := \{\varepsilon, \rho\} \cup \mathscr{E} \cup \mathscr{P} \cup \{\top\}$$
$$\mathcal{T}_{\mathcal{R}_{\mathscr{B}}} := \{[b, \tau, d], [b, *, d],\ \exists\, \delta^{\Theta^{a,b}_{i,j}, \tau, \Theta^{c,d}_{i,k}} \in \mathscr{S}_{\mathscr{B}}\} \cup \{[\varepsilon, *, \varepsilon]\} \cup \{[\varepsilon, *, \rho]\} \cup \{[\rho, *, \rho] \cup \{[\top, *, \top]\}$$
$$\mathcal{I}_{\mathscr{B}} := \{\Theta^{a,-}_{i,j},\ \Theta^{c,-}_{i,k}\}_{\delta^{\Theta^{a,-}_{i,j}, -, \Theta^{c,-}_{i,k}}}$$

donde $\mathscr{S}_{\mathscr{B}}$ es la semántica asociada al *corpus* de ejemplo $\mathscr{B}$.

Intuitivamente, consideramos que el conjunto de conceptos $\mathcal{T}_{\mathcal{C}_{\mathscr{B}}}$ que manejaremos para el caso del *corpus* $\mathscr{B}$, se puede clasificar en entidades y propiedades, tal como se describe en la Tabla 3, y no se tiene en cuenta el orden seguido entre elementos similares y/o diferentes. Sólo se define una relación de subsunción entre las entidades individuales (resp. propiedades) y el correspondiente elemento genérico, $*$. Con respecto al conjunto de relaciones $\mathcal{T}_{\mathcal{R}_{\mathscr{B}}}$, se extraen directamente a partir de $\mathscr{S}_{\mathscr{B}}$ a través de la dinámica de transición, resumiéndose desde el punto de vista de las clases semánticas (tipos) de los términos que participan en ella. Como elementos adicionales, se añaden tripletas que representan cualquier posible transición en la semántica que relacione conceptos genéricos. El orden parcial que consideramos en $\mathcal{T}_{\mathcal{C}_{\mathscr{B}}}$ es el inducido naturalmente por el ya definido en $\mathscr{T}$. Finalmente, definimos los referentes individuales $\mathcal{I}_{\mathscr{B}}$ como un conjunto de formas del *corpus* $\mathscr{B}$.

Ahora estamos en disposición de presentar los GCB's que vamos a considerar sobre este soporte. Nuestro punto de partida es la semántica $\mathscr{S}_{\mathscr{D}_m}$ asociada a cada uno de los documentos que constituyen el *corpus*

$$\mathscr{B} = \bigcup_{m \in M} \mathscr{D}_m$$

donde M es el número de estos documentos:

$$\mathcal{C}_{\mathscr{D}_m} := \{\Theta^{a,b}_{i,j}, \Theta^{c,d}_{i,k}\}_{\delta^{\Theta^{a,b}_{i,j},\_,\Theta^{c,d}_{i,k}} \in \mathscr{S}_{\mathscr{D}_m}} \qquad \mathcal{R}_{\mathscr{D}_m} := \{[b,\tau,d],\ \exists\, \delta^{\Theta^{\_,b}_{\_,\_},\tau,\Theta^{\_,d}_{\_,\_}} \in \mathscr{S}_{\mathscr{D}_m}\}$$

$$\mathcal{A}_{\mathscr{D}_m} := \bigcup_{\delta^{\Theta^{a,b}_{i,j},\tau,\Theta^{c,d}_{i,k}} \in \mathscr{S}_{\mathscr{D}_m}} \{([b,\tau,d],1,\Theta^{a,b}_{i,j}),([b,\tau,d],2,\Theta^{c,d}_{i,k})\}$$

$$\mathcal{E}_{\mathscr{D}_m}(X) := \begin{cases} [b,\Theta^{a,\_}_{i,j}] & \text{si} \quad X = \Theta^{a,b}_{i,j} \in \mathcal{C}_{\mathscr{D}_m} \\ X & \text{si} \quad X \in \mathcal{R}_{\mathscr{D}_m} \\ 1 & \text{si} \quad X = (\_,1,\_) \in \mathcal{A}_{\mathscr{D}_m} \\ 2 & \text{si} \quad X = (\_,2,\_) \in \mathcal{A}_{\mathscr{D}_m} \end{cases}$$

Brevemente, un nodo conceptual en $\mathcal{C}_{\mathscr{D}_m}$ es cualquier término involucrado en la semántica $\mathscr{S}_{\mathscr{D}_m}$, mientras que los nodos relaciones en $\mathcal{R}_{\mathscr{D}_m}$ son elementos de $\mathcal{T}_{\mathcal{R}_{\mathscr{B}}}$ asociados a las transiciones en $\mathscr{S}_{\mathscr{D}_m}$. El multiconjunto de aristas $\mathcal{A}_{\mathscr{D}_m}$ contiene en este caso únicamente las relaciones binarias correspondientes a los términos gobernante (resp. gobernado) de la primera tripleta (resp. la segunda).

En cuanto a la función de etiquetado $\mathcal{E}_{\mathscr{D}_m}$, permite recuperar la clase semántica y el token asociado a un término dado representando un concepto, al tiempo que implementa la identidad en las relaciones, ya que en nuestro caso las construimos directamente a partir de la semántica del *corpus*. El valor de esta función sobre las aristas identifica las gobernantes (1) y las gobernadas (2).

6 | El marco de evaluación

El modelo tradicional de evaluación experimental de sistemas de RI [30, 31] implica tres tareas complementarias: la recopilación de una colección documental, la definición de una serie de medidas de confianza para su evaluación y la elección adecuada de un conjunto de tópicos, es decir, de consultas.

A este respecto, es necesario tomar como punto de partida un fondo documental. Con respecto a las otras dos tareas, se trata de minimizar la carga de trabajo asociada a la creación de los JREL's así como a la selección de tópicos. Esto nos permitirá no tener que hacer frente a colecciones de prueba, que incluyen un número arbitrario de documentos en cualquier ámbito del conocimiento, algo difícilmente abordable a escala humana.

Definición 21 *Sean $\mathcal{D} = \{d_i\}_{i\in I}$ una colección documental y $\mathcal{Q} = \{c_j\}_{j\in J}$ un conjunto finito de tópicos (consultas). Se dice que un documento $d_i \in \mathcal{D}$ es* relevante *con respecto a un tópico $c_j \in \mathcal{Q}$ si y sólo si un experto humano considera que dicho documento posee información relativa al mismo. Si no es así, se dice que $d_i \in \mathcal{D}$* no es relevante *a $c_j \in \mathcal{Q}$. Al conjunto de documentos de $\mathcal{D}$ que son relevantes a $c_j \in \mathcal{Q}$, lo denotamos por* $\text{rel}(c_j,\mathcal{D})$*, y por* $\text{nrel}(c_j,\mathcal{D})$ *a los que no lo son.*

■

Nuestro objetivo aquí es tratar de discriminar la eficacia entre diferentes sistemas de RI, detectando cuales resultan ser más sensibles a la hora de identificar documentos relevantes. Por este motivo, en un primer momento será necesario garantizar la estabilidad operativa del propio concepto de relevancia. Sin embargo, lo cierto es que al parecer existen factores que influyen en la concretud de esta definición [131]. Es el caso de las discrepancias entre evaluadores o incluso contradicciones individuales [135] por parte de un mismo evaluador, factores que se ven reforzados por el hecho de que estamos hablando de una magnitud continua que se pretende clasificar mediante una secuencia de valores [141]. Con respecto a esto, asumimos que la influencia de estos factores de desestabilización es mínima, como ya se sugirió en un principio en [66], y que más tarde se corroboró experimentalmente en [158]. Del mismo modo, el desacuerdo en el número de documentos relevantes parece no tener un fuerte impacto en la ordenación de los sistemas [135], probablemente porque tener más documentos relevantes beneficia a la mayoría de los sistemas de manera uniforme.

Si centramos ahora nuestra atención en la tarea de selección de tópicos y en las ordenaciones devueltas por los entornos de RI, se pueden distinguir dos marcos genéricos de acuerdo con el estado del arte. Por un lado, el inspirado en la extensa experiencia acumulada durante décadas en los eventos del TREC y caracterizado esencialmente por el uso preferente de juicios humanos[25], sin tener en cuenta en el proceso de la sencillez o complejidad del tópico. Se habla entonces de un marco *basado en la valoración de tipo humano*. Por el otro, un conjunto de técnicas inspiradas en dos supuestos razonables esbozados en [100] en relación al *«principio de facilidad y/o dificultad»* de determinadas consultas y el *«principio de lo bueno o malo»* que puede resultar ser un sistema de RI. A diferencia de la basada en la valoración de tipo humano, ésta formaliza la sencillez o complejidad de un tópico a partir de medidas basadas en JREL's como un factor importante que impacta en esta tarea. De un modo más detallado, el primer principio establece que deberíamos asignar un peso mayor (resp. menor) tanto si se comete un error en consultas sencillas (resp. difíciles), como si se contesta correctamente en las consideradas difíciles (resp. fáciles). El segundo asume que deberíamos ser capaces de realizar consultas complicadas a los buenos sistemas, mientras que los malos sólo debieran ser capaces de contestar a las sencillas. En adelante, nos referiremos a este marco como el *basado en una valoración tipo máquina*.

Como alternativa, en lo que ocupa exclusivamente la ordenación de sistemas de RI, se ha propuesto una tercera vía que prescinde por completo de uso de recursos basados en JREL's [164]. Se trata en este caso de evaluar el rendimiento de un motor de búsqueda utilizando una medida llamada *contador de referencia*, un tipo específico de puntuación que se calcula mediante el número de ocurrencias de los documentos más relevantes devueltos en los resultados de una colección de otros sistemas de recuperación.

[25] mediante mecanismos de JREL's o similares, como en el caso de PJREL's.

6.1 | Sistemas de RI con ordenación usando JREL's

La utilización de JREL's es la base de la mayoría de las medidas de evaluación de los sistemas de RI, popularizadas entre la comunidad investigadora gracias a las conferencias del TREC. De este modo, podemos distinguir entre dos acercamientos según tengamos en cuenta o no el orden asociado a la clasificación de los resultados devueltos durante la recuperación, lo que actualmente es habitual en los motores de búsqueda.

6.1.1 | Medidas de evaluación basadas en conjuntos

Este tipo de medida estima la calidad de un conjunto no ordenado de documentos recuperados. Se trata de técnicas asociadas a la evaluación de un modelo de RI bidimensional [66]. Esto es, no se considera el orden asociado a las clasificaciones de los contextos y la evaluación sólo se centra en el carácter relevante o no de los documentos recuperados. En este sentido, se introducen una serie de medidas que detallamos a continuación.

Definición 22 *Sean σ un sistema de* RI*, donde $\mathcal{D} = \{d_i\}_{i\in I}$ es una colección documental, y $\mathcal{Q} = \{c_j\}_{j\in J}$ un conjunto finito de tópicos (consultas). Se define la* precisión *(resp. la* cobertura*)* de σ con respecto del tópico c_j para la colección documental $\mathcal{D}$ *como:*

$$\text{P}(\sigma, c_j, \mathcal{D}\ ) := \frac{|\text{rec}(\sigma, c_j, \mathcal{D}) \cap \text{rel}(c_j, \mathcal{D})\,|}{|\text{rec}(\sigma, c_j, \mathcal{D})|} \tag{13}$$

$$(\textit{resp.}\ \text{C}(\sigma, c_j, \mathcal{D}\ ) := \frac{|\text{rec}(\sigma, c_j, \mathcal{D}) \cap \text{rel}(c_j, \mathcal{D})\,|}{|\text{rel}(c_j, \mathcal{D})\,|}) \tag{14}$$

donde $\text{rec}(\sigma, c_j, \mathcal{D})$ *(resp.* $\text{rel}(\sigma, c_j, \mathcal{D})$*) es el conjunto de documentos de $\mathcal{D}$ recuperados por σ (resp. los documentos relevantes) para el tópico $c_j \in \mathcal{Q}$.*

■

Tanto la *precisión* como la *cobertura* fueron introducidas por Cleverton *et al.* en [29]. Intuitivamente, la precisión (resp. la cobertura) representa la proporción entre el número de documentos relevantes recuperados y el número de documentos recuperados en total (resp. documentos relevantes totales), es decir, un valor predictivo positivo de la tarea de búsqueda (resp. la sensibilidad). Por lo tanto, la precisión (resp. la cobertura) evalúa la exactitud (resp. la exhaustividad) de la búsqueda en función de los resultados. En particular, la precisión (resp. la cobertura) no se define cuando no se recuperan documentos (resp. cuando no hay documentos relevantes) en la colección y es mínima (resp. máxima) cuando todos ellos son devueltos por el buscador. En cualquier caso, se trata de conceptos complementarios calculados con respecto a toda la lista de documentos devueltos por el sistema, lo cual plantea algún problema a la hora de estimar la efectividad.

Esto justifica la introducción por von Rijsbergen en [150] de la medida F_β como una manera de estimar la efectividad de la recuperación con respecto al usuario, que concede β veces tanta importancia a la cobertura como a la precisión.

Definición 23 *Sean σ un sistema de* RI*, $\mathcal{D} = \{d_i\}_{i\in I}$ una colección documental y $\mathcal{Q} = \{c_j\}_{j\in J}$ un conjunto finito de tópicos (consultas). Se define, por $\beta \in \mathbb{R}^+ \cup \{0\}$, la medida F_β* de σ con respecto al tópico c_j y la colección documental $\mathcal{D}$ *como:*

$$\mathrm{F}_\beta(\sigma, c_j, \mathcal{D}) := \frac{(1+\beta^2)\ .\ [\mathrm{P}(\sigma, c_j, \mathcal{D})\ .\ \mathrm{C}(\sigma, c_j, \mathcal{D})]}{\beta^2\ .\ \mathrm{P}(\sigma, c_j, \mathcal{D}) + \mathrm{C}(\sigma, c_j, \mathcal{D})} \tag{15}$$

En el caso particular de que $\beta = 1$, se habla de medida F.

■

La medida F_β permite hacer enfásis sobre los pesos asociados a la precisión con respecto a la cobertura, utilizando como valor de control a β. Así, cuando $\beta = 1$, se obtiene la *media armónica* de ambas medidas, que en comparación con la aritmética requiere que los dos valores sean elevados para que a su vez también ella lo sea. En cambio, para valores $\beta < 1$ pesará más la precisión mientras que para valores $\beta > 1$ lo hará la cobertura. Por otro lado, ninguna de estas medidas considera la proporción de documentos no relevantes que se recuperan, situación a la que pretende dar respuesta la introducción del ratio de *fracaso o irrelevancia*[26].

Definición 24 *Sean σ un sistema de* RI*, $\mathcal{D} = \{d_i\}_{i\in I}$ una colección documental y $\mathcal{Q} = \{c_j\}_{j\in J}$ un conjunto finito de tópicos (consultas). Se define el* fracaso de σ con respecto al tópico c_j en la colección documental $\mathcal{D}$ *como:*

$$\mathrm{FR}(\sigma, c_j, \mathcal{D}) := \frac{|\mathrm{rec}(\sigma, c_j, \mathcal{D}) \cap \mathrm{nrel}(c_j, \mathcal{D})\,|}{|\mathrm{nrel}(c_j, \mathcal{D})\,|} \tag{16}$$

donde $\mathrm{nrel}(c_j, \mathcal{D})$ *es el conjunto de documentos de $\mathcal{D}$ que no son relevantes a $c_j \in \mathcal{Q}$.*

■

De esta manera, el fracaso, que fue inicialmente introducido por Salton y McGill [128], se puede interpretar como la probabilidad de que un documento no relevante sea recuperado. Así, este valor devolverá 0 cuando no se recupere ningún documento como respuesta a una consulta.

[26] en terminología anglosajona *fall-out rate*.

6.1.2 | Medidas de evaluación basadas en ordenación

Este tipo de medida considera el orden en el que se presentan los documentos devueltos, una mejora sustancial en relación con las métricas anteriores, ya que estima la precisión en todos los niveles de cobertura. Como consecuencia, se pueden derivar dos mejoras prácticas. La primera hace referencia a la real contribución que implica disponer de información extra sobre el grado de relevancia asociado al sistema de recuperación con respecto a una consulta dada. La segunda permite estimar la eficiencia de un sistema de RI, incluso cuando sólo estamos interesados en calcularlo sobre resultados recuperados en los niveles más bajos. Es el caso típico de la recuperación Web, donde el usuario normalmente se desentiende de las respuestas que no se encuentren en las primeras páginas. Formalmente [124], estas mejoras se traducen en dos aspectos: la *estabilidad*[27] y la *sensibilidad*[28] de la tarea de evaluación.

Una primera aproximación para conseguirlo consiste en determinar la precisión frente a la cobertura de cada uno de los documentos recuperados. Para ello, sincronizaremos ambas medidas sobre la base de los primeros k documentos devueltos.

Definición 25 *Sean σ un sistema de* RI, $\mathcal{D} = \{d_i\}_{i\in I}$ *una colección documental, y $\mathcal{Q} = \{c_j\}_{j\in J}$ un conjunto finito de tópicos (consultas). Se define la* precisión *(resp. la* cobertura*)* de los k documentos devueltos por σ con respecto a los tópicos c_j sobre $\mathcal{D}$*, denotada por* P@$k(\sigma, c_j, \mathcal{D})$ *(resp.* C@$k(\sigma, c_j, \mathcal{D})$*), como:*

$$\text{P@}k(\sigma, c_j, \mathcal{D}) := \frac{|\{\text{reco}(\sigma, c_j, \mathcal{D})_l\}_{l=1}^{k} \cap \text{rel}(c_j, \mathcal{D})\,|}{k} \tag{17}$$

$$(\textit{resp.}\ \text{C@}k(\sigma, c_j, \mathcal{D}) := \frac{|\{\text{reco}(\sigma, c_j, \mathcal{D})_l\}_{l=1}^{k} \cap \text{rel}(c_j, \mathcal{D})\,|}{|\text{rel}(c_j, \mathcal{D})\,|}) \tag{18}$$

donde reco$(\sigma, c_j, \mathcal{D})$ *es la lista, ordenada en base a su relevancia, de los documentos recuperados por σ para el tópico c_j.*

■

Llegados aquí, estamos en disposición de expresar la precisión en función de la cobertura, simplemente calculando ambas medidas en los puntos de sincronización. Como resultado obtenemos un grafo de la precisión/cobertura [98, 114].

[27] la estabilidad de una medida está relacionada con la capacidad que tiene de identificar sistemáticamente las diferencias entre los sistemas a partir de una muestra de tópicos [20].

[28] también llamada *ratio de cobertura*, se refiere a las medidas de evaluación del poder de discriminación de un sistema de RI, sobre una colección de prueba y una serie de ejecuciones realizadas y definidas a partir de la colección [160].

Definición 26 *Sean σ un sistema de* RI*, $\mathcal{D} = \{d_i\}_{i\in I}$ una colección documental, y $\mathcal{Q} = \{c_j\}_{j\in J}$ un conjunto finito de tópicos (consultas). Se puede expresar la* precisión de σ sobre el tópico c_j para la colección documental $\mathcal{D}$ en función de la cobertura *como:*

$$\text{P}_{\text{C}}(\sigma, c_j, \mathcal{D}, c) := \text{P}@k(\sigma, c_j, \mathcal{D}),\ c = \text{C}@k(\sigma, c_j, \mathcal{D}) \tag{19}$$

■

Intuitivamente, la precisión se calcula en el mismo instante que la cobertura, justo en el momento en el que el motor de búsqueda devuelve el documento. Como resultado [99], este tipo de curvas tiene una particularidad y es que presenta la forma de diente de sierra ya que si el $(k+1)$-ésimo documento recuperado no es relevante entonces la cobertura será la misma para los k primeros, pero la precisión experimentará un descenso. Sin embargo, en el caso de que el documento sea relevante, entonces tanto la precisión como la cobertura se incrementarán, y la curva despuntará hacia la derecha. En este sentido, resulta útil eliminar estas sacudidas y la manera estándar de hacerlo es a través de la interpolación.

Definición 27 *Sean σ un sistema de* RI*, $\mathcal{D} = \{d_i\}_{i\in I}$ una colección documental y $\mathcal{Q} = \{c_j\}_{j\in J}$ un conjunto finito de tópicos (consultas). Se define la* precisión interpolada de σ sobre el tópico c_j en función de la cobertura *para la colección documental $\mathcal{D}$, como*

$$\text{PI}_{\text{C}}(\sigma, c_j, \mathcal{D}, c) := \max_{c'\geq c} \text{P}_{\text{C}}(\sigma, c_j, \mathcal{D}, c') \tag{20}$$

■

De esta manera, la medida refiere a la precisión más alta encontrada para la solución del problema planteado. Por el otro lado, aunque hemos utilizado $\text{P}@k$ como primer paso para introducir el grafo de precisión/cobertura, el concepto también posee interés en sí mismo. Así, una de las ventajas que se suele argumentar en su favor es que no requiere de la estimación del conjunto de documentos pertinentes. Sin embargo, por el mismo motivo no calcula correctamente la media y no podemos considerarlo como un criterio estable de evaluación [99]. Una alternativa para aliviar este problema es la *R-precisión* (resp. la *R-cobertura*) [128].

Definición 28 *Sean σ un sistema de* RI*, $\mathcal{D} = \{d_i\}_{i\in I}$ una colección documental y $\mathcal{Q} = \{c_j\}_{j\in J}$ un conjunto finito de tópicos (consultas). Se define la* R-precisión*, denotada por $\text{P}@R(\sigma, c_j, \mathcal{D})$ (resp.* R-cobertura *y denotada por $\text{C}@R(\sigma, c_j, \mathcal{D})$),* de σ sobre el tópico c_j para la colección documental $\mathcal{D}$ *como:*

$$R\text{-P}(\sigma, c_j, \mathcal{D}\ ) := \text{P}@R(\sigma, c_j, \mathcal{D}) \tag{21}$$

$$(\textit{resp. } R\text{-C}(\sigma, c_j, \mathcal{D}\ ) := \text{C}@R(\sigma, c_j, \mathcal{D})) \tag{22}$$

donde $R = |\text{rel}(c_j, \mathcal{D})|$.

■

Intuitivamente, si la colección documental incluye R documentos relevantes para una consulta dada, entonces R-P indicará la cantidad de relevantes una vez que los R mejores resultados hayan sido estudiados por el sistema. En resumen, se refiere a la mejor precisión sobre el grafo P_C, lo que justifica que también sea conocido como el *punto de equilibrio de* P_C, ya que la precisión y la cobertura coinciden en él.

En cualquier caso, ninguna de las métricas de relevancia graduada es tan ampliamente utilizada actualmente como la *precisión media* (PM), que proporciona una interpretación geométrica de los grafos de precisión/cobertura [127]. En efecto, calcula el área bajo la curva P_C, lo que implica estimar el valor medio de la cobertura para el intervalo $[0, 1]$.

Definición 29 *Sean* σ *un sistema de* RI, $\mathcal{D} = \{d_i\}_{i \in I}$ *una colección documental y* $\mathcal{Q} = \{c_j\}_{j \in J}$ *una colección de tópicos (consultas). Se define la* precisión media de σ con respecto al tópico c_j para la colección $\mathcal{D}$ *como:*

$$\text{PM}(\sigma, c_j, \mathcal{D}\ ) := \int_0^1 \text{P}_\text{C}(\sigma, c_j, \mathcal{D})\ d\text{C} \tag{23}$$

En la práctica, este valor se aproxima mediante una suma discreta sobre cada posición de la secuencia ordenada de documentos devueltos, tal como sigue:

$$\text{PM}(\sigma, c_j, \mathcal{D}\ ) := \frac{1}{|\text{rel}(c_j, \mathcal{D})\,|} \sum_{k=1}^{|\text{reco}(\sigma, c_j, \mathcal{D})|} \delta(\text{reco}(\sigma, c_j, \mathcal{D})_k)\ .\ \text{P@}k(\sigma, c_j, \mathcal{D}) \tag{24}$$

donde

$$\delta(\text{reco}(\sigma, c_j, \mathcal{D})_k) := \begin{cases} 1 & \text{si reco}(\sigma, c_j, \mathcal{D})_k \in \text{rel}(c_j, \mathcal{D}\ ) \\ 0 & \text{en cualquier otro caso} \end{cases}$$

■

En la práctica, PM y R-P están altamente correlacionados [145, 161] y muestran una estabilidad similar en términos de comparación de sistemas usando tópicos diferentes [11]. Aunque esto podría parecer algo aparentemente sorprendente[29], se puede demostrar formalmente [5] que si se asume un conjunto razonable de suposiciones, ambas medidas aproximan el área bajo la curva P_C, lo que explica el fenómeno. Además, podemos mejorar la estabilidad calculando el promedio de la PM a través de las consultas [63].

[29] el cómputo de la R-P considera un único punto de precisión mientras que la PM evalúa el área bajo toda la curva P_C.

Definición 30 *Sean σ un sistema de* RI*, $\mathcal{D} = \{d_i\}_{i\in I}$ una colección documental y $\mathcal{Q} = \{c_j\}_{j\in J}$ un conjunto finito de tópicos (consultas). Se define el* promedio de la precisión media de σ sobre el conjunto de tópicos $\mathcal{Q}$ para una colección documental $\mathcal{D}$ *como*

$$\text{PPM}(\sigma, \mathcal{Q}, \mathcal{D}\ ) := \frac{\sum_{j\in J} \text{PM}(\sigma, c_j, \mathcal{D})}{|\mathcal{Q}|} \tag{25}$$

■

Mientras que PM aproxima al área bajo la curva P_C, PPM es aproximadamente el promedio de ese mismo área para un conjunto de consultas. De hecho, PPM es la medida de uso más frecuente en lo que a recuperación con ordenación se refiere, lo que provocó que se convirtiera en un estándar para la comunidad TREC. Considera aspectos orientados tanto a la cobertura como a la precisión, y es sensible a la ordenación devuelta por el sistema, proporcionando una medida de calidad a través de los niveles de cobertura sobre una única figura. Sin embargo, el PPM tiene el efecto de ponderar por igual cada una de las necesidades de información en el resultado final que devuelve, aunque existan muchos documentos relevantes para algunas consultas, mientras que existan muy pocos para otras. Esto significa que un conjunto de prueba debe ser lo suficientemente grande y variado para llegar a ser representativo de la eficacia del sistema sobre las diferentes consultas. Asumiendo estas condiciones, el PPM ha demostrado poseer una especial sensibilidad y estabilidad entre las medidas de evaluación [99]. Por lo demás, es necesaria la utilización de otro tipo de métricas cuando lo que interesa es destacar las mejoras en consultas de bajo rendimiento.

Definición 31 *Sean σ un sistema de* RI*, $\mathcal{D} = \{d_i\}_{i\in I}$ una colección documental y $\mathcal{Q} = \{c_j\}_{j\in J}$ un conjunto finito de tópicos (consultas). Se define el* promedio geométrico de la precisión media de σ sobre el conjunto de tópicos $\mathcal{Q}$ de la colección documental $\mathcal{D}$ *como*

$$\text{PGPM}(\sigma, \mathcal{Q}, \mathcal{D}\ ) := \sqrt[J]{\prod_{j\in J} \text{PM}(\sigma, c_j, \mathcal{D})} \tag{26}$$

■

Tanto el PPM como el PGPM pueden verse como maneras diferentes de alcanzar una medida de calidad a través de la incorporación de diferentes observaciones individuales. Así, mientras la primera es la media aritmética de la PM, considerando un conjunto de tópicos, la segunda es la media geométrica. En este sentido, el PGPM es más representativo de la eficacia a través de un conjunto de consultas, y más robusto frente a situaciones en

las que la presencia de unas pocas interrogaciones con buen rendimiento pueden sesgar la clasificación obtenida mediante el PPM. Concretamente, el PGPM fue introducido por Voorhees en [159].

En este punto, si quisiéramos resumir en una característica común las métricas descritas hasta el momento, tendríamos que decir que todas ellas vienen completamente determinadas por la ordenación de los documentos relevantes en el conjunto resultante. Por lo tanto, no hacen distinción entre los documentos que son explícitamente juzgados como no relevantes y aquéllos que se asume que no son relevantes por no haber sido juzgados, lo cual plantea un problema cuando se sabe que los JREL's proporcionados están lejos de ser completos, haciéndose aconsejable el atenuar esta situación.

Definición 32 *Sean σ un sistema de* RI*, $\mathcal{D} = \{d_i\}_{i\in I}$ una colección documental y $\mathcal{Q} = \{c_j\}_{j\in J}$ un conjunto finito de tópicos (consultas). Se define la* relación de preferencia binaria de σ sobre el tópico c_j en la colección documental $\mathcal{D}$ *como:*

$$\text{PREFB}(\sigma, c_j, \mathcal{D}\ ) := \frac{1}{\text{R}} \sum_{r\in \text{R}} [1 - \frac{|\text{nrel}(c_j, \mathcal{D}) \cap \{\text{reco}(\sigma, c_j, \mathcal{D})\}_{r+1}^{\text{R}}|}{\text{mín}\{\text{R}, |\text{nrel}(c_j, \mathcal{D})|\}}] \qquad (27)$$

donde $\text{R} = |\text{rel}(c_j, \mathcal{D})|$. *Se puede extender de un modo natural esta definición al conjunto finito de tópicos $\mathcal{Q}$.*

■

La medida PREFB, introducida por Buckley *et al.* [12] puede pensarse como la inversa de la fracción de los documentos recuperados que son juzgados como no relevantes y que se sitúan en una posición anterior a los relevantes. De este modo, se calcula una relación de preferencia en función de si los documentos juzgados como relevantes se recuperan antes que los juzgados como irrelevantes, esto es, la medida está basada únicamente en las ordenaciones relativas de los documentos que han sido juzgados previamente. Hablamos de preferencias binarias porque la relación se define a partir de un JREL binario, de tal manera que, dada una consulta, se prefiere cualquier documento relevante frente a los que no lo son. En este sentido, PREFB y PPM están altamente correlacionados cuando se utilizan con JREL's completos. Sin embargo, cuando éstos son incompletos, aunque los sistemas de ordenación mediante la PREFB todavía se correlacionan mucho con los originales, no es el caso de los que ordenan mediante PPM.

Una última propuesta que ha conseguido una aceptación cada vez mayor, especialmente cuando se emplea asociada a sistemas de aprendizaje automático, es la *ganancia acumulativa* (GAA) [99]. Normalmente, la valoración inicial proporcionada por los sistemas de RI posee múltiples grados y, en consecuencia, la mejora debería ser evaluada separadamente en cada nivel de relevancia. En este sentido, los documentos considerados como más relevantes que aparezcan en peores puestos en la lista proporcionada por el sistema debieran ser penalizados, reduciendo el valor de su

relevancia. Sea como sea, las medidas dependientes de la ordenación descritas hasta ahora son calculadas usando unas valoraciones dicotómicas acerca de la relevancia, colapsando éstas en dos para su evaluación.

Definición 33 *Sean σ un sistema de* RI*, $\mathcal{D} = \{d_i\}_{i\in I}$ una colección documental y $\mathcal{Q} = \{c_j\}_{j\in J}$ un conjunto finito de tópicos (consultas). Se define la* ganancia acumulativa reducida de σ sobre el tópico c_j en la colección documental $\mathcal{D}$ en la posición ordenada $r \in [1, R] \cap \mathbb{N}$ *como:*

$$\text{GAAR}(\sigma, c_j, \mathcal{D})_r := \text{G}(\sigma, c_j, \mathcal{D})_1 + \sum_{k=2}^{r} \frac{\text{G}(\sigma, c_j, \mathcal{D})_k}{log_b(k)} \tag{28}$$

donde $R = |\text{rel}(c_j, \mathcal{D})|$ y G *es la secuencia de valores relevantes asociados a la lista* $\text{reco}(\sigma, c_j, \mathcal{D})$*. Se puede extender naturalmente esta definición al conjunto finito de tópicos $\mathcal{Q}$.*

■

En la práctica, la GAAR usa el nivel de relevancia como una medida de valor acumulado en la posición de ordenación asociada al documento, añadiendo esta ganancia progresivamente desde la primera posición a la última. Se asocia una función logarítmica reducida con el fin de aminorar poco a poco el valor del documento al mismo tiempo que se incrementa su posición en la ordenación, pero no demasiado bruscamente. Normalmente, se usa un logaritmo en base dos, esto es, considerando $b = 2$ en la Ecuación 28.

Aunque el conjunto de documentos recuperados puede variar ampliamente entre diferentes sistemas, para comparar sus rendimientos, la versión normalizada de esta medida utiliza el mayor valor posible de GAAR para cada una de las posiciones.

Definición 34 *Sean $\sigma = \{\sigma_i\}_{i\in I}$ una colección de sistemas de* RI*, $\mathcal{D} = \{d_j\}_{j\in J}$ una colección documental, $\mathcal{Q} = \{c_k\}_{k\in K}$ un conjunto finito de tópicos (consultas) y $\{\text{GAAR}(\sigma, c_k, \mathcal{D})_l\}_{l\in L}$ la secuencia (conjunto ordenado) de valores de* GAAR *para el tópico c_k. Se define la* ganancia acumulativa reducida normalizada de σ_i sobre el tópico c_k de la colección documental $\mathcal{D}$ en la posición ordenada $r \in [1, R] \cap \mathbb{N},\ R = |\textit{rel}(c_k, \mathcal{D})|$ *como:*

$$\text{GAARN}(\sigma_i, c_k, \mathcal{D})_r := \frac{\text{GAAR}(\sigma_i, c_k, \mathcal{D})_r}{\text{GARI}(\sigma_i, c_k, \mathcal{D})_r} \tag{29}$$

donde GARI *se denomina la* GAAR ideal*, y se define como el* GAAR *máximo alcanzable en el rango r. Ésta se puede calcular fácilmente a partir de las* GAAR*'s de una lista ordenada que sitúa todos los documentos con mejor clasificación por encima de todos los segundos*

y así sucesivamente. Se puede extender naturalmente esta definición al conjunto finito de tópicos $\mathcal{Q}$.

■

Obviamente, en un algoritmo de ordenación perfecto asociado a un sistema de RI, los valores correspondientes para GAARN serán iguales a 1. Ambas métricas GAAR y GAARN fueron introducidas por Järvelin y Kekäläinen en [72]. Los resultados obtenidos indican una fuerte correlación entre la satisfacción de los usuarios, la GAA y la precisión; una correlación más moderada con la GAAR y una sorprendentemente posible correlación casi despreciable con la GAARN [1].

6.2 | Sistemas de RI con ordenación usando PJREL's

Introducida por Soboroff *et al.* en [135], esta técnica simplemente retoma el proceso oficial de evaluación del TREC [156], cambiando algún aspecto referido a la valoración del entrenamiento basado en asesoramiento humano. Más exactamente, se consideran los siguientes pasos, descritos por los autores:

1. Se selecciona un grupo de 50 consultas siguiendo la propuesta de un grupo de expertos de confianza, normalmente de la organización NIST[30].

2. Se lanzan para su evaluación un número de ejecuciones, asociadas a cada sistema de RI evaluado. Cada una de estas ejecuciones consta (como máximo) de los mejores 1.000'00 documentos recuperados para cada tópico. Por cada participante se crea un subconjunto con estas características que se etiqueta como *ejecución oficial*.

3. El grupo de expertos toma los n primeros documentos devueltos en cada consulta para cada una de las ejecuciones oficiales, eliminando las duplicidades, con el fin de crear un *fondo* para cada una de ellas.

4. Se selecciona aleatoriamente un conjunto de documentos para formar los PJREL's, utilizando un modelo para determinar la relevancia de los documentos que están en ese fondo.

5. A partir del conjunto de PJREL's, se evalúan todas las ejecuciones usando el paquete de evaluación del TREC[31].

Esto es, con respecto al TREC, Soboroff *et al.* tomaron en el tercer paso los valores $n = 10$ ó $n = 100$, mientras que el TREC considera únicamente el caso de que n sea igual a 10. A su vez, en el cuarto paso sustituyeron el papel de los expertos por una elección totalmente

[30] por *National Institute of Standards and Technology*.
[31] consultar `http://trec.nist.gov/trec_eval/`.

aleatoria. Finalmente, en los pasos cuarto y quinto, consideraron los PJREL's en vez de los JREL's.

Obviamente, para estimar este tipo de clasificación podemos considerar todas las medidas previamente descritas para los entornos de evaluación basados en JREL's.

6.3 | Sistemas de RI con ordenación basada en la valoración de la máquina

Descrita por Mizzaro *et al.* [101], esta técnica toma como base la estimación de lo fácil o difícil que puede resultar un tópico, considerando que si el motor de búsqueda quiere tener un alto rendimiento deberá ser suficientemente eficaz en las consultas difíciles. Vamos a bautizar a esta propiedad asociada a un sistema de RI como su *autoridad*, y antes de formalizarla necesitaremos introducir algunos conceptos para la captura de las nociones de facilidad de la consulta y la eficacia del sistema. El punto de partida para esta metodología es la noción de PM, cuyo cálculo puede ser aplicado tanto a JREL's como a PJREL's.

Definición 35 *Sean $\sigma = \{\sigma_i\}_{i\in I}$ un conjunto de sistemas de* RI*, $\mathcal{D}$ una colección documental, y $\mathcal{Q} = \{c_j\}_{j\in J}$ un conjunto finito de tópicos (consultas). Se define la* media de la precisión media del conjunto de sistemas de RI σ aplicado a un tópico c_j para la colección $\mathcal{D}$*, como:*

$$\text{MPM}(\sigma, c_j, \mathcal{D}\ ) := \frac{\sum_{i\in I} \text{PM}(\sigma_i, c_j, \mathcal{D})}{|\sigma|} \tag{30}$$

■

Intuitivamente, la MPM es un indicador de la facilidad asociada a la satisfacción de la consulta, entendiéndola como una magnitud directamente relacionada con el número de sistemas de RI que poseen un buen rendimiento para ese tópico. A partir de la base que ofrece esta medida, Mizzaro *et al.* [101] extienden el concepto de PM con el fin de obtener una directriz fiable para estimar el rendimiento de un sistema de RI sobre las distintas consultas. La idea pasa, en primer lugar, por normalizar la PM con el fin de eliminar cualquier influencia achacable a la facilidad de aquéllas por separado (resp. de la eficacia del sistema de manera individual), con el fin de obtener una medida fiable del rendimiento en un conjunto de sistemas de RI (resp. de lo fácil que resulte ser una consulta).

Definición 36 *Sean $\sigma = \{\sigma_i\}_{i\in I}$ un conjunto de sistemas de* RI*, $\mathcal{D}$ una colección documental, y $\mathcal{Q} = \{c_j\}_{j\in J}$ un conjunto finito de tópicos (consultas). Se define la* precisión media normalizada de σ_i aplicada al tópico c_j de acuerdo con la $\text{MPM}(\sigma, c_j, \mathcal{D})$*, como:*

$$\text{PMN}_{\text{MPM}}(\sigma_i, c_j, \mathcal{D}\ ) := \text{PM}(\sigma_i, c_j, \mathcal{D}) - \text{MPM}(\sigma, c_j, \mathcal{D}\ ) \tag{31}$$

■

De esta manera, la matriz de adyacencia $[\text{PMN}_{\text{MPM}}(\sigma_i, c_j, \mathcal{D})]_{(i,j)\in I\times J}$ puede ser interpretada como un grafo ponderado bipartito, donde el peso de los arcos $c_j \rightarrow \sigma_i$ corresponde a los valores de $\text{PMN}_{\text{MPM}}(\sigma_i, c_j, \mathcal{D})$, lo que refleja el desempeño individual de σ_i sobre el tópico c_j y la eliminación de las desviaciones debido a la facilidad de éste. La medida de PMN_{MPM} fue introducida por Wu y McClean en [163], y Mizzaro [100] calculó su media con el fin de buscar una mejor estabilidad.

Definición 37 *Sean $\sigma = \{\sigma_i\}_{i\in I}$ un conjunto de sistemas de* RI*, $\mathcal{D}$ una colección documental, y $\mathcal{Q} = \{c_j\}_{j\in J}$ un conjunto finito de tópicos (consultas). Se define el* promedio normalizado de la precisión media de σ_i sobre el conjunto de consultas $\mathcal{Q}$ para la colección documental $\mathcal{D}$*, como:*

$$\text{PNPM}(\sigma_i, \mathcal{Q}, \mathcal{D}\,) := \frac{\sum_{j\in J} \text{PMN}_{\text{MPM}}(\sigma_i, c_j, \mathcal{D})}{|\mathcal{Q}|} \tag{32}$$

■

Sorprendentemente, el PNPM muestra un comportamiento algo distinto a los resultados del TREC, proporcionando una clasificación muy diferente en relación con PPM, aunque ambas medidas están relacionadas[32]. En la práctica, lo que generalmente se considera una versión mejorada de un sistema mediante la aplicación de criterios del TREC[33] a menudo resulta no serlo cuando se utiliza PNPM.

Una alternativa para aprovechar la información contenida en la matriz de adyacencia PMN_{MPM} pasa por analizarla sobre la base del algoritmo de HITS de Kleinberg [82] para obtener medidas de evaluación más sofisticadas teniendo en cuenta los conjuntos en su totalidad para ambos, sistemas de RI y consultas. La idea básica propuesta por Mizzaro *et al.* consiste en retomar los indicadores descritos por Kleinberg para la localización de información de alta calidad relacionada con las estructuras de enlace: la *conectividad* y la *autoridad*.

Definición 38 *Sean $\sigma = \{\sigma_i\}_{i\in I}$ un conjunto de sistemas de* RI*, $\mathcal{D}$ una colección documental, y $\mathcal{Q} = \{c_j\}_{j\in J}$ un conjunto finito de tópicos (consultas). Se define la* autoridad de un sistema de RI σ_i sobre el conjunto de consultas $\mathcal{Q}$ *(resp. la* conectividad del tópico c_j en el sistema de RI σ*)* para la colección $\mathcal{D}$*, como:*

$$\text{A}(\sigma_i, \mathcal{Q}, \mathcal{D}\,) := \sum_{j\in J} \text{T}(c_j, \sigma, \mathcal{D}\,) \cdot \text{PMN}_{\text{MPM}}(\sigma_i, c_j, \mathcal{D}\,) \tag{33}$$

$$(\textit{resp.}\ \text{T}(c_j, \sigma, \mathcal{D}\,) := \sum_{i\in I} \text{A}(\sigma_i, \mathcal{Q}, \mathcal{D}\,) \cdot \text{PMN}_{\text{MPM}}(\sigma_i, c_j, \mathcal{D})) \tag{34}$$

■

[32] la correlación tau de Kendall [80] es $0'87$ y la correlación lineal [19] es $0'92$.
[33] es decir, una versión con un mayor PPM.

Intuitivamente, un sistema de RI posee una autoridad alta si es más eficiente sobre los tópicos con una también alta conectividad, es decir, cuando se trata de consultas difíciles. Esto proporciona un criterio de ordenación simple, ya que un sistema que quiere ser eficaz debería presentar unos valores altos en la autoridad asociada.

6.4 | Sistemas de RI con ordenación en base a contadores de referencia ponderados

Descrito por Wu *et al.* en [164], esta propuesta aplica una técnica de fusión de datos que compara los resultados obtenidos para un motor de búsqueda con las tomadas a partir de una colección de otros sistemas de RI distintos. Ello requiere la introducción previa de un cierto número de conceptos.

Definición 39 *Sean* $\sigma = \{\sigma_i\}_{i\in I}$ *un conjunto de sistemas de* RI, $\mathcal{D} = \{d_j\}_{j\in J}$ *una colección documental, y* $\mathcal{Q} = \{c_k\}_{k\in K}$ *un conjunto finito de tópicos (consultas). Denotamos por*

$$\text{CR}(\sigma_i, c_k, \mathcal{D}\,) := \sum_{j_i\in J_i} a(\text{reco}(\sigma_i, c_k, \mathcal{D})_{j_i}) \tag{35}$$

al contador de referencia de σ_i sobre un tópico c_k para la colección documental $\mathcal{D}$, *donde* $a(\text{reco}(\sigma_i, c_k, \mathcal{D})_{j_i})$ *es el número de apariciones de un documento* $\text{reco}(\sigma_i, c_k, \mathcal{D})_{j_i}$ *en la lista* $\{\text{reco}(\sigma_l, c_k, \mathcal{D})\}_{j_l\in J_l, l\neq i}$.

Dado $a(\text{reco}(\sigma_i, c_k, \mathcal{D})_{j_i})$, *bautizamos como* $\text{reco}(\sigma_i, c_k, \mathcal{D})_{j_i}$ *al* documento original *y a sus homólogos en* $\{\text{reco}(\sigma_l, c_k, \mathcal{D})\}_{j_l\in J_l, l\neq i}$ *como los* documentos de referencia *denotados por* $\gamma(\text{reco}(\sigma_i, c_k, \mathcal{D})_{j_i})$.

■

Intuitivamente, dada una consulta y un cierto número de los documentos originales devueltos en las mejores posiciones por un determinado sistema RI en una determinada colección, su CR es la suma de las referencias proporcionadas por los otros sistemas. Esto inspira un método sencillo de ordenación al margen de la consideración de los JREL's y al que Wu *et al.* denominaron *método básico*.

Definición 40 *Sean* $\sigma = \{\sigma_i\}_{i\in I}$ *un conjunto de sistemas de* RI, $\mathcal{D}$ *una colección documental, y* $\mathcal{Q} = \{c_j\}_{j\in J}$ *un conjunto finito de tópicos (consultas). Denotamos por*

$$\text{CRM}(\sigma_i, \mathcal{Q}, \mathcal{D}\,) := \frac{\sum_{j\in J} \text{CR}(\sigma_i, c_j, \mathcal{D})}{|\mathcal{Q}|} \tag{36}$$

al contador de referencia media de σ_i en el conjunto de tópicos $\mathcal{Q}$ para la colección documental $\mathcal{D}$.

■

Intuitivamente, dado un sistema de RI, se calculan sus CRM's como el valor medio de los valores individuales de CR en cada consulta, lo que proporciona una técnica de ordenación fiable para sistemas de RI. Entre las mejoras propuestas por los autores de este método básico se optó por considerar la posición de relevancia de ambos, los documentos originales y los de referencia. Esto hace necesario ampliar la noción de CR con el fin de integrarlos.

Definición 41 *Sean* $\sigma = \{\sigma_i\}_{i\in I}$ *un conjunto de sistemas de* RI, $\mathcal{D}$ *una colección documental,* $\mathcal{Q} = \{c_j\}_{j\in J}$ *un conjunto finito de tópicos (consultas), y* $\{\varrho_{j_i}\}_{j_i\in J_i}$ *las puntuaciones normalizadas*[34] *asociadas a* $\{\text{reco}(\sigma_i, c_j, \mathcal{D})\}_{j_i\in J_i}$. *Sea también* $\forall\ m \in [1, \text{NumDocsMax}],\ k \in [1, 4],\ \text{NumDocsMax}=1.000$*:*

$$\hat{a}(\text{reco}(\sigma_i, c_j, \mathcal{D})_{j_i}) := \sum_{\text{reco}_{(\sigma_k, c_j, \mathcal{D})_{k_l}} \in \gamma(\text{reco}_{(\sigma_i, c_j, \mathcal{D})_{j_i}})} \Delta - l \ \ (\text{resp. } \varrho_{k_l})$$

y

$$\omega_{j_i} := \begin{cases} \zeta(200) - \zeta(m-1)\ , & \text{si} \quad j_i = 5m \\ \omega_{5m} - \frac{1}{m} + \frac{5}{j_i}\ , & \text{si} \quad j_i = 5m - k \end{cases}$$

siendo $\hat{a}(\text{reco}(\sigma_i, c_j, \mathcal{D})_{j_i})$ *y* ω_{j_i} *las funciones de peso asociadas a la relevancia de las posiciones de referencia y a los documentos originales, respectivamente, definiéndose la función auxiliar* ζ *como*

$$\zeta(m) := \begin{cases} 0\ , & \text{si } m = 0 \\ 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{m}\ , & \text{en cualquier otro caso} \end{cases}$$

donde NumDocsMax *es el tamaño máximo de la colección documental* $\mathcal{D}$ *y* Δ *es un valor constante, que los autores establecen empíricamente en sus experimentos a 1.501'00. Denotamos a la expresión*

$$\sum_{j_i\in J_i} \omega_{j_i}.\hat{a}(\text{reco}(\sigma_i, c_j, \mathcal{D})_{j_i}) \tag{37}$$

como $\text{CRP}_{\text{O}}(\sigma_i, c_j, \mathcal{D})$ *(resp.* $\text{CRP}_{\text{P}}(\sigma_i, c_j, \mathcal{D})$*), al* contador de referencia ponderado basado en la ordenación *(resp.* basado en la puntuación*)* de σ_i sobre un tópico c_j para la colección $\mathcal{D}$.

■

Siguiendo el mismo proceso que se aplicó para introducir CRM a partir de los CR, ahora podemos introducir naturalmente la *media de contadores de referencia ponderados,* MCRP_{O} (*resp.* MCRP_{P}) de CRP_{O} (*resp.* CRP_{P}), que ofrece dos medidas adicionales de ordenación.

[34] asumimos, sin pérdida de generalización, que estas puntuaciones están en el intervalo $[0, 1]$.

Sin embargo, algunas de las elecciones en esta propuesta de ordenación son difíciles de justificar, ya que no se han argumentado razones convincentes para presentar la constante Δ, ni los (muy complejos) valores de ω_{j_i}. Como las fórmulas resultantes son poco claras y difíciles de entender, se propone modificar ligeramente el planteamiento original.

Definición 42 *Sean $\{\sigma_i\}_{i\in I}$ un conjunto de sistemas de* RI*, $\mathcal{D}$ una colección documental, $\mathcal{Q} = \{c_j\}_{j\in J}$ un conjunto finito de tópicos (consultas), y $\{\varrho_{j_i}\}_{j_i\in J_i}$ las puntuaciones normalizadas*[35] *asociadas a $\{\text{reco}(\sigma_i, c_j, \mathcal{D})\}_{j_i\in J_i}$. Sea también $\forall\ m, n \in [1, |\mathcal{D}|]$:*

$$\hat{a}(\text{reco}(\sigma_i, c_j, \mathcal{D})_{j_i}) := \sum_{\text{reco}_{(\sigma_k, c_j, \mathcal{D})_{k_l}} \in \gamma(\text{reco}_{(\sigma_i, c_j, \mathcal{D})_{j_i}})} \omega_{k_l} \ \textit{(resp. } \varrho_{k_l}\textit{)},\ \textit{donde}\ \omega_{k_l} := \begin{cases} 1 & \text{si } l = 1 \\ \frac{1}{log_b(l)} & \text{en cualquier otro caso} \end{cases}$$

y

$$\hat{\omega}_{j_i} := \begin{cases} 1 & \text{si } j_i = 1 \\ \frac{1}{log_b(j_i)} & \text{en cualquier otro caso} \end{cases}$$

siendo las funciones de peso asociadas a la relevancia de las posiciones de referencia y a los documentos originales, respectivamente. Denotamos a la expresión

$$\sum_{j_i\in J_i} \hat{\omega}_{j_i}.\hat{a}(\text{reco}(\sigma_i, c_j, \mathcal{D})_{j_i}) \tag{38}$$

como $\text{CRP}_{\text{OL}}(\sigma_i, c_j, \mathcal{D})$ *(resp.* $\text{CRP}_{\text{PL}}(\sigma_i, c_j, \mathcal{D})$*), denominado como el* contador de referencia ponderado basado en la ordenación logarítmica *(resp.* basado en la puntuación logarítmica*)* de σ_i sobre un tópico c_j para la colección $\mathcal{D}$.

■

Siguiendo el mismo proceso que se aplicó para introducir MCRP_{O} (resp. MCRP_{P}), ahora podemos introducir MCRP_{OL} (resp. MCRP_{PL}), lo cual proporciona las medidas de ordenación usando contadores de referencia ponderados que tendremos en cuenta en este trabajo. Tomaremos $b = 2$.

6.5 | Selección del conjunto de tópicos

El objetivo ahora es seleccionar un conjunto de consultas minimal con el fin de evaluar nuestro sistema de RI comparándolo con una colección de las ya existentes, tomando como referencia los diferentes niveles de dificultad en su resolución por parte del usuario. Como visión general, consideramos una técnica de muestreo estratificado para seleccionar *un conjunto inicial de tópicos*, sobre el que más adelante aplicaremos una técnica de minimización para reducir su tamaño sin perder su poder de discriminación.

[35] asumimos, sin pérdida de generalización, que estas puntuaciones están en el intervalo $[0, 1]$.

Esto nos va a permitir simplificar en gran medida la tarea de pruebas que aquí es especialmente compleja por cuanto no sólo pretendemos estimar la eficiencia del sistema de RI, sino también identificar los factores que impactan en términos de imprecisión y de incompletud. Teniendo esto en cuenta, dentro de lo que conocemos, ninguna técnica específica se ha descrito para este fin concreto; por lo que nuestro enfoque tiene el carácter de propuesta.

6.5.1 | El tamaño de la muestra inicial

Una cuestión fundamental consiste en determinar el tamaño del conjunto de consultas que deberíamos utilizar para evaluar la propuesta, para lo que tomamos como referencia la discusión que plantean al respecto Guiver *et al.* [62], a su vez referida a diversos trabajos anteriores. En este sentido, los autores ponen de manifiesto una clara evolución en el estado del arte, atribuyendo las primeras estimaciones a Jones y a van Rijsbergen [140], que llegaron a la conclusión de que usando un número de 75 no era suficiente, 250 eran por lo general aceptable, e incluso 1.000 podían llegar a ser necesarios. Más tarde Zobel [172] apoya la idea de que un conjunto de 25 consultas ya permite realizar un trabajo razonable, mientras que Buckley y Voorhees [11] proporcionan la primera evidencia efectiva de que el número de tópicos necesarios para un buen experimento es de al menos 25, aunque 50 parece ser mejor. Más recientemente, en el contexto de las evaluaciones al estilo TREC, Webber *et al.* [162] afirman que se requieren de unas 150 consultas para distinguir de forma fiable entre sistemas de RI, aunque por lo general sólo se consideran 50 [156]. En nuestro caso, hemos seleccionado en un primer momento una muestra inicial de 150 tópicos.

6.5.2 | El proceso de muestreo

En primer lugar, clasificamos nuestro espacio muestral[36] (población) siguiendo dos criterios independientes, cada uno formando su propia partición, y que creemos puede estar correlacionada con la noción intuitiva de dificultad (durante la resolución) de las consultas. Esta última constituye la variable dependiente deseada para el muestreo, una elección basada en Mizzaro *et al.* [101] que sugiere que es un factor importante en los tópicos para discriminar eficazmente entre sistemas de RI. En la práctica, introducimos de manera concisa estos criterios mediante sus variables asociadas:

- La *especificidad del tópico*, entendiéndola como el nivel de detalle con el que el usuario la expresa. Consideramos tres niveles diferentes: alto, medio y bajo.
- El *tipo de respuesta* devuelto por un motor de búsqueda siguiendo un enfoque conceptual: aproximado, plausible y parcial. Asumimos aquí que una consulta pertenece a un determinado tipo cuando el conjunto de respuestas de esa clase

[36] formado por la totalidad de las posibles consultas a aplicar sobre nuestro *corpus* $\mathscr{B}$.

dentro de las 10 primeras devueltas por el sistema[37] posee un mayor peso estimable que el correspondiente a los demás tipos. Por lo tanto, es necesario fijar la relación μ_u (resp. μ_a) que limita el número de uniones (resp. de agregaciones) asociadas a respuestas plausibles (resp. parciales), así como calcular formalmente dicho peso.

Estos criterios también nos van a permitir combinar ambos puntos de vista, el del usuario y el del sistema de RI. Con el fin de equilibrar la muestra que nos va a servir como conjunto inicial de tópicos, tendremos que minimizar (resp. maximizar) la variabilidad dentro de (resp. entre) las subpoblaciones (estratos) correspondientes a las diferentes particiones. Por lo tanto, distribuimos la muestra entre las tres subpoblaciones introducidas para cada uno de ellas[38], lo que proporciona homogeneidad en todos los niveles de la estratificación. Asimismo, los tópicos de un determinado estrato de una de las particiones se reparten equitativamente entre los estratos de la otra. De este modo, aseguraríamos que la probabilidad de que una de las consultas de la muestra tenga un tipo de respuesta y una especificidad dadas sea aproximadamente la misma, cualquiera que fuera la combinación considerada para estas variables. De esta manera, esperamos mejorar la precisión y la eficiencia de la estimación, sacar conclusiones sobre las subpoblaciones y permitir un mayor equilibrio estadístico en las pruebas sobre las diferencias entre las particiones. Para lograr este objetivo hemos puesto en práctica un cuidadoso proceso de selección.

En relación con la especificidad del tópico, partimos de una colección de tópicos propuestos por expertos y repartida en tres estratos, de tal manera que las consultas de uno se obtienen refinando el contenido de las del estrato anterior. El objetivo es integrar, en número similar, los tópicos con especificidad alta, media y baja. Más en detalle, consideramos una colección inicial de tópicos verificando:

$$\mathcal{Q} := \{\mathcal{Q}_i^{ea}\}_{i\in I} \cup \{\mathcal{Q}_i^{em}\}_{i\in I} \cup \{\mathcal{Q}_i^{eb}\}_{i\in I},\ \mathcal{Q}_i^{ea} \succ \mathcal{Q}_i^{em} \succ \mathcal{Q}_i^{eb},\ \forall i \in I$$

donde $\succeq$ es el orden parcial naturalmente inducido en el espacio muestral por la especificidad detectada por los expertos.

Con respecto al tipo de respuesta, en primer lugar tomamos el valor $\mu_u = 0'34$ (resp. $\mu_a = 0'18$) con el fin de moderar el número de respuestas plausibles devueltas (resp. parciales)[39], lo que equivale a aplicar un muestreo ajustado con la probabilidad adecuada.

Una vez que se ha hecho esto, es necesario introducir algún criterio para medir el peso de un determinado tipo de respuesta en un conjunto finito de éstas, repartiéndolo equilibradamente entre los tipos considerados. Aquí, asumimos que no sólo hemos de

[37] lo que aproximadamente se corresponde con la primera página de resultados devueltos por un motor de búsqueda cualquiera, justo el límite por encima del cual el usuario deja de mostrar interés en la revisión de las respuestas [61].

[38] esto implica que asociamos 50 consultas por estrato, el mismo número considerado por el protocolo clásico del TREC [156] para la evaluación de sistemas de RI.

[39] el número de respuestas plausibles y, especialmente, de parciales pueden incrementar artificialmente su número debido al hecho de que se generan aplicando mecanismos que pueden hacer crecer indefinidamente el tamaño de los GCB's asociados a las consultas, algo que no ocurre con las aproximadas.

tener en cuenta el número de respuestas de determinado caso, sino también la posición de éstas en la ordenación. Por lo tanto, el tipo de respuesta que aparece más abajo en la lista resultante de la búsqueda debería ser penalizado a la vez que se reduce el grado del valor de relevancia. Ello nos sitúa en un contexto equiparable al considerado en la determinación de las medidas de evaluación basadas en ordenación de los sistemas de RI y, más concretamente, en el proceso de construcción de la medida GAARN, que nos servirá ahora de inspiración para introducir la noción de *peso acumulado descontado* asociada a un tipo de respuesta dada.

Definición 43 *Sea σ un sistema de* RI*, $\mathcal{D} = \{d_i\}_{i\in I}$ una colección documental y $\mathcal{Q} = \{c_j\}_{j\in J}$ un conjunto finito de tópicos (consultas). Se define el* peso acumulado descontado *de σ sobre el tópico c_j para un tipo de respuesta ι y una colección documental $\mathcal{D}$ con tamaño de selección $p \in [1, |rec(\sigma, c_j, \mathcal{D})|]$ como:*

$$\text{PAD}(\sigma, \iota, c_j, \mathcal{D})_p := \delta_\iota^{\text{tipo}(\text{reco}_{(\sigma,c_j,\mathcal{D})_1})} + \sum_{k=2}^{p} \frac{\text{tipo}(\text{reco}(\sigma, c_j, \mathcal{D}))_k}{log_b(k)} \tag{39}$$

donde tipo *devuelve el tipo de respuesta que le sirve de argumento, y δ_i^j es la función conocida como* delta de Kronecker*, el cual se define de la siguiente manera:*

$$\delta_i^j := \begin{cases} 1 & \text{si } i = j \\ 0 & \text{en cualquier otro caso} \end{cases} \tag{40}$$

■

En nuestro caso particular, tomamos $p = 10$, $b = 2$ y nuestra propuesta de RI conceptual como σ, lo cual implica que $\iota \in \{aproximada, plausible, parcial\}$. En la práctica, el equipo de expertos emplea la medida PAD para alcanzar una distribución uniforme para la muestra basada en el tipo de respuesta, teniendo en cuenta simultáneamente el criterio de especificidad previamente descrito. Como resultado, se consigue un conjunto inicial de tópicos que verifica todas las restricciones descritas anteriormente a partir de ambos puntos de vista: heterogeneidad entre estratos en las diferentes particiones y homogeneidad en todos los niveles de estratificación. Esto nos coloca en el punto de comienzo de la fase de minimización que introducimos en tres pasos.

6.5.3 | Selección de tópicos individuales para un sistema dado

El primer acercamiento para tratar con la selección de tópicos pasa por fijar una estrategia de estimación de la adecuación de una consulta individual para medir el rendimiento de un proceso de RI. En este sentido, y tomando como fuente de inspiración la experiencia del TREC, la medida PM mide la eficacia de un sistema σ sobre un tópico

individual $c \in \mathcal{Q}$ para una colección documental $\mathcal{D}$, lo que aparentemente podría resolver la cuestión.

Sin embargo, situándonos en el marco de la valoración tipo máquina, no podemos concluir que σ presente un mejor rendimiento para el tópico c que en el tópico $\tilde{c}$ (resp. que σ considera más fácil a c que $\tilde{\sigma}$), en base al dato $\text{PM}(\sigma, c, \mathcal{D}) > \text{PM}(\sigma, \tilde{c}, \mathcal{D})$ (resp. $\text{PM}(\sigma, c, \mathcal{D}) > \text{PM}(\tilde{\sigma}, c, \mathcal{D})$). Simplemente c podría ser un tópico más sencillo[40] y $\tilde{c}$ uno difícil[41] (resp. σ podría ser un buen sistema[42] y $\tilde{\sigma}$ uno malo[43]). Esto nos lleva a volver nuestra atención al concepto de PMN_{MPM} donde, contrariamente a lo que ocurre con PM, la condición $\text{PMN}_{\text{MPM}}(\sigma, c, \mathcal{D}) > \text{PMN}_{\text{MPM}}(\sigma, \tilde{c}, \mathcal{D})$ nos permite inferir que un sistema de RI σ tiene un buen rendimiento en la consulta c y uno malo en $\tilde{c}$.

6.5.4 | Selección de un conjunto de tópicos para un sistema dado

Entre todas las técnicas inspiradas en el TREC y disponibles en el estado del arte para resolver esta cuestión, se optó por trabajar con la de Guiver *et al.* en [62]. El punto de partida es ahora la medida PPM, de hecho un indicador de la eficacia de un sistema de RI que nos orienta sobre su bondad, una vez que el conjunto de consultas ha sido fijado para una colección de documentos dada. La idea consiste en aplicar una búsqueda exhaustiva en todos los posibles subconjuntos de tópicos en una colección determinada. De esta forma, podemos centrarnos en la correlación más alta de estos valores de PPM con el del concepto de la colección, con el fin de estimar la bondad de la predicción sobre un subconjunto de consultas del rendimiento del sistema de RI.

Por otra parte, también podemos retomar aquí un razonamiento similar en el marco de la valoración tipo máquina, usando ahora valores PNPM en lugar de los PPM y teniendo en cuenta que estas dos métricas no siempre coinciden.

6.5.5 | Selección de un conjunto de tópicos para un conjunto de sistemas

A nuestro conocimiento, no se han presentado ni documentado propuestas, hasta ahora, a este respecto en el estado del arte. Nuestra estrategia se apoya tanto en el marco basado en la valoración de tipo humana como en el basado en la valoración tipo máquina, sobre la base de las técnicas presentadas anteriormente, lo mismo para la selección individual que para los conjuntos de consultas en sistemas de RI particulares. Sin embargo, aunque los pasos a aplicar para conseguirlo son los mismos, su naturaleza dependerá en cada momento del tipo de marco de trabajo elegido:

1. El primer paso consiste en generar, a partir de la muestra que sirve de conjunto

[40] esto es, una consulta sobre la cual todos o la mayoría de los sistemas de RI tienen un buen desempeño.

[41] es decir, una consulta sobre la que todos o la mayoría de los sistemas de RI tienen un desempeño deficiente.

[42] es decir, un sistema cuya efectividad se extienda a todos o a la mayoría de las consultas difíciles.

[43] es decir, un sistema cuya efectividad se limita a las consultas fáciles.

inicial de tópicos, una colección de subconjuntos con distintas capacidades para medir el rendimiento del sistema en diferentes niveles, y que denominamos *colección de referencia de tópicos*. En el extremo superior (resp. en el inferior) de esta gradación, situaremos subconjuntos de consultas formadas exclusivamente por aquéllas consideradas difíciles (resp. fáciles) con el poder de discriminación más alto (resp. más bajo). Cualquier tópico no catalogado como difícil o fácil se considerará como medio. El tamaño de cada uno de estos subconjuntos será nuevamente de 50, siguiendo con la propuesta de Webber *et al.* [162].

Se generan dos tipos de colecciones, dependiendo del marco que nos indique la estimación del nivel de sencillez de los tópicos. Por lo tanto, recurrimos a la opinión de un experto en el dominio, en el caso de la estrategia basada en la valoración de tipo humano. Por el contrario, con respecto al criterio basado en máquina, se identifican las consultas difíciles (resp. las fáciles) con la mayor conectividad (resp. la menor conectividad) en el conjunto de sistemas de RI.

2. A continuación, se aplican a cada una de estas colecciones de referencia una estrategia de minimización con el fin de reducir su tamaño sin afectar perceptiblemente su poder de discriminación. El resultado constituirá dos conjuntos de *colecciones finales de tópicos*, uno especialmente orientado a una valoración basada en tipo humano y otro en tipo máquina, distinguiendo cada cual tres niveles de dificultad: alto, medio y bajo. Para calcular el primero, seguimos la técnica propuesta por Guiver *et al.* en [62] sobre la base de la medida de correlación PPM[44]. Dado que tanto la PPM y la PNPM se pueden calcular a partir de JREL's o PJREL's, finalmente obtenemos cuatro colecciones finales de tópicos. Dos de ellos consideran JREL's (resp. PJREL's) como base para calcular la PPM y la PNPM, uno usando una valoración basada en el tipo humano y otra aplicando una basada en el tipo máquina.

   La única cuestión pendiente ahora es determinar la composición de estos subconjuntos finales, un problema para el que los autores no proporcionan un criterio claro. En este sentido, hemos decidido escoger aquéllos candidatos cuya cardinalidad se encuentra en el intervalo $[1, 50)$, mientras alcance un nivel suficientemente alto de correlación PPM (resp. PNPM) con el correspondiente subconjunto de tópicos de referencia.

   En el caso de las consultas basadas en JREL's, tomamos un nivel de correlación PPM (resp. PNPM) con la correspondiente valoración basada en el tipo humano (resp. basada en tipo máquina) orientado a la colección de tópicos de referencia que sea superior o igual a $0'99999932$. Esto supone en una aproximación de tipo humano (resp. valoración tipo máquina) considerar una colección de subconjuntos finales con 12 tópicos (resp. con 10) para dificultades altas, 22 (resp. 15) para dificultades medias y 32 (resp. 8) para las bajas, que denominaremos *colección de tópicos tipo humano sobre* JREL*'s* (resp. *tipo máquina*), o brevemente, CTHJ (resp. CTMJ).

[44] ambos acercamientos se han descrito previamente cuando se introdujo la selección de un conjunto de tópicos para un sistema de RI individual

En el caso de las consultas basadas en PJREL's, tomamos un nivel de correlación PPM (resp. PNPM) con la correspondiente valoración basada en el tipo humano (resp. basada en tipo máquina) orientado a la colección de tópicos de referencia que sea superior o igual a $0'9999990$. Esto supone en una aproximación de tipo humano (resp. valoración tipo máquina) considerar una colección de subconjuntos finales con 30 tópicos (resp. con 2) para dificultades altas, 29 (resp. 22) para dificultades medias y 24 (resp. 48) para las bajas, que denominaremos *colección de tópicos tipo humano sobre* PJREL*'s* (resp. *tipo máquina*), o brevemente, CTHPJ (resp. CTMPJ).

En este contexto, ninguna consulta de la muestra del conjunto inicial posee mayor probabilidad de ser incluido en el conjunto reducido final ni por su tipo de respuesta ni por su especificidad, sino que ello dependerá exclusivamente de su dificultad de resolución, determinada por cualquiera de los dos métodos antes descritos. Esto garantizará la objetividad y la validez de los resultados experimentales que se obtengan usando una muestra reducida. Sin embargo, parece razonable esperar que el protocolo que seguimos para mejorar la selección de consultas proporcione conclusiones sensiblemente diferentes en función del marco específico sobre el que se realice las pruebas. En efecto, los trabajos anteriores [100, 101] muestran que aunque un sistema de RI que quiera ser eficaz en el TREC tendrá que serlo sobre los tópicos fáciles, el sentido común indica que un motor de búsqueda eficaz debe demostrar su verdadero poder en los difíciles.

6.6 | El conjunto de sistemas de RI

Elegimos una muestra de cuatro plataformas de motores de búsqueda bien conocidas con el fin de servir como valores de referencia de comparación para estimar el rendimiento de la eficiencia de nuestra propuesta, que bautizamos como COGIR:

1. ZETTAIR (ver `http://www.seg.rmit.edu.au/zettair/`) es un motor de búsqueda de código abierto desarrollado por el *Search Engine Group* de la Universidad RMIT, desarrollado en C. Fue diseñado buscando simplicidad, así como velocidad y flexibilidad, y su principal característica es su capacidad de manejar grandes cantidades de texto. Este motor de búsqueda admite consultas de tipo booleano y frases.

2. SOLR (ver `lucene.apache.org/solr/`) es una plataforma de búsqueda de código abierto del proyecto Apache Lucene. Sus características principales son que está escrito en JAVA y que se ejecuta como un servidor de búsqueda de texto independiente incluido dentro de un contenedor de servlets como es el caso de TOMCAT. Utiliza la librería de búsqueda de JAVA Lucene en su núcleo para la indexación completa de texto y posterior búsqueda. SOLR proporciona una búsqueda distribuida y la replicación de índices, impulsando la búsqueda y las características de navegación de muchos de los sitios Web más importantes.

3. TERRIER[45] (ver `http://ir.dcs.gla.ac.uk/terrier/`) es motor de búsqueda de código abierto altamente flexible, eficaz, y efectivo, fácilmente desplegable en colecciones de documentos a gran escala y desarrollado en la Universidad de Glasgow. Está escrito en JAVA y proporciona múltiples estrategias de indexación, como el de una sola pasada, de múltiples pasadas y de indexación a gran escala usando algoritmos de MapReduce.

4. INDRI (ver `http://www.lemurproject.org/indri/`) es un motor de búsqueda de código abierto para gran escala, escrito en C++. Fue construido a partir del proyecto LEMUR (ver `http://www.lemurproject.org/`), el cual es un conjunto de herramientas diseñado para la investigación en el modelado de lenguaje y la RI. Este proyecto fue desarrollado gracias al trabajo cooperativo entre las Universidades de Massachusetts y de Carnegie Mellon.

Estos motores de búsqueda proporcionan un abanico representativo de los más populares en la actual oferta de buscadores, incluyendo tanto diferentes lenguajes de implementación como diferentes modelos de búsqueda.

7 | Resultados experimentales

Una vez formalizado el marco de evaluación, ya sólo queda introducir, visualizar e interpretar los resultados, teniendo en cuenta que la manera más sencilla de comparar los diferentes sistemas de RI es ordenándolos mediante valores decrecientes, de acuerdo con las diferentes métricas asociadas al rendimiento. A este respecto, vamos a seguir el mismo orden que el considerado anteriormente a la hora de introducirlas, en función de su tipo.

7.1 | Sistemas de RI con ordenación usando JREL's

En este nivel, hemos considerado el conjunto total de las diferentes métricas de rendimiento presentadas previamente (y en número de catorce) con el fin de experimentar con ellas, lo cual debería ser suficiente para detectar cualquier posible mal funcionamiento en nuestra propuesta, al tiempo que garantizamos la robustez de la evaluación. Así, los tests se realizaron sobre las dos colecciones de conjuntos de tópicos establecidas, CTHJ y CTMJ, buscando adecuar el criterio de la selección de tópicos al enfoque específico de ordenación, ambos basados en JREL's. Esto debería proporcionar fiabilidad al proceso.

[45] de TERabyte RetrIEveR

7.1.1 | Usando una colección de conjuntos de tópicos basada en la valoración tipo humano

Tomamos aquí la CTHJ como colección de tópicos, lo que proporcionará una visión general del comportamiento de nuestra propuesta para hacer frente a la ordenación basada en JREL's sobre tópicos seleccionados mediante una valoración de tipo humano. Esto debería constituir un protocolo de evaluación bien fundado.

Medidas de evaluación basadas en conjuntos

Tratamos aquí con los resultados de las medidas P y C, que se muestran en las Figs. 1.3 y 1.4 respectivamente. Tal y como puede comprobarse, en cualquier caso los resultados indican una mejor precisión del modelo conceptual COGIR sobre los demás, a la vez que una mayor contención de la cobertura.

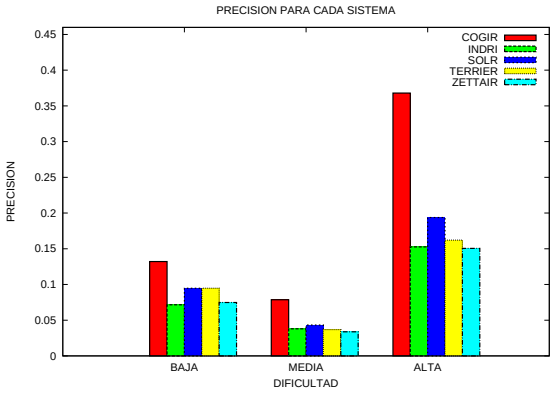

Figura 1.3: P sobre CTHJ usando JREL's

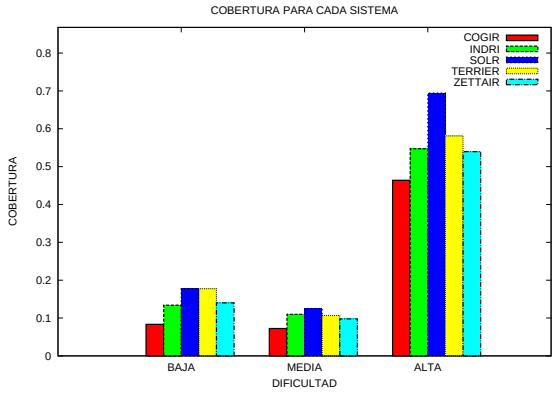

Figura 1.4: C sobre CTHJ usando JREL's

También se incluyen los tests para las métricas F y FR, a fin de tener en cuenta la proporción de documentos no relevantes que son recuperados. Los gráficos asociados se muestran en las Figs. 1.5 y 1.6, respectivamente. En este caso, los valores favorecen claramente al modelo conceptual frente a los otros para el conjunto de tópicos de mayor dificultad, esto es, en aquéllos con mayor poder de discriminación entre sistemas en lo que a evaluación se refiere. Sin embargo, los resultados son menos impactantes para los tópicos con menor poder de discriminación.

Medidas de evaluación basadas en ordenación

Tratamos aquí con los resultados de las medidas P@10 y C@10, que se muestran en las Figs. 1.7 y 1.8, respectivamente. Tal y como puede comprobarse, en cualquier caso los resultados muestran una mejor precisión del modelo conceptual COGIR sobre los demás, a la vez que una mayor contención de la cobertura.

Con el fin de estudiar la posible extensión de los resultados observados en la primera página al conjunto de respuestas obtenidas, calculamos PI_C para niveles 0 (resp. 0'10) de

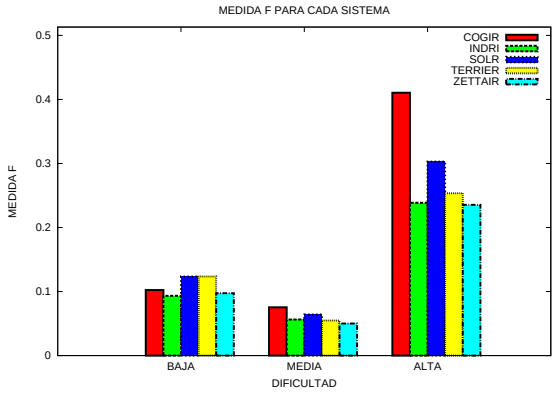

Figura 1.5: F sobre CTHJ usando JREL's

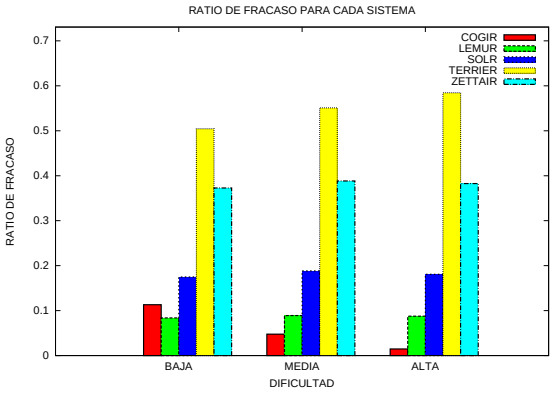

Figura 1.6: FR sobre CTHJ usando JREL's

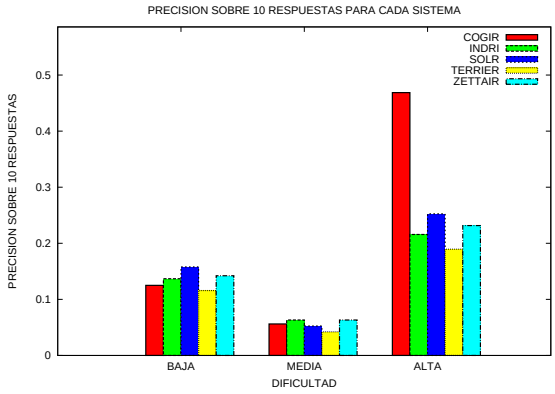

Figura 1.7: P@10 sobre CTHJ usando JREL's

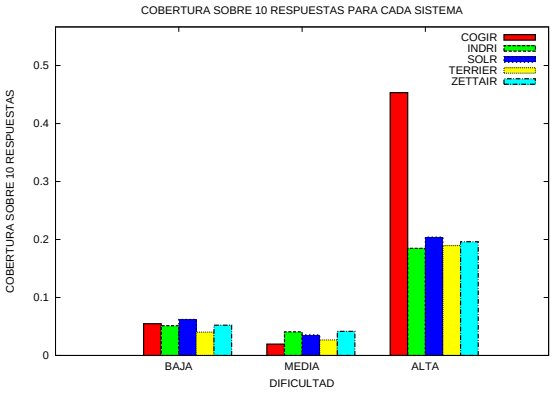

Figura 1.8: C@10 sobre CTHJ usando JREL's

cobertura en la Fig. 1.9 (resp. en la Fig 1.10). De nuevo, como en los casos anteriores, vuelve a quedar patente el mejor comportamiento del modelo conceptual sobre los tópicos con mayor nivel de dificultad.

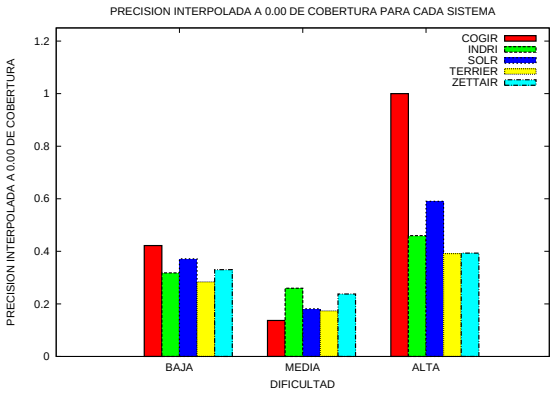

Figura 1.9: $\text{PI}_{C=0'00}$ sobre CTHJ usando JREL's

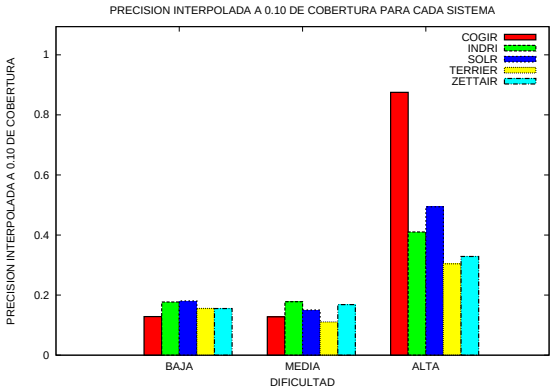

Figura 1.10: $\text{PI}_{C=0'10}$ sobre CTHJ usando JREL's

Por su parte, las Figs. 1.11 y 1.12 vuelven a avalar la robustez del modelo conceptual sobre la base de las medidas R-P y PPM. Al tiempo, estos resultados destacan su rendimiento en el tratamiento de las consultas con mayor dificultad, manteniendo las prestaciones en relación al resto de entornos considerados en otro caso.

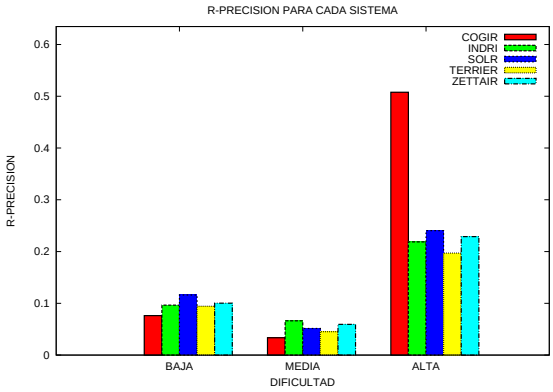

Figura 1.11: R-P sobre CTHJ usando JREL's

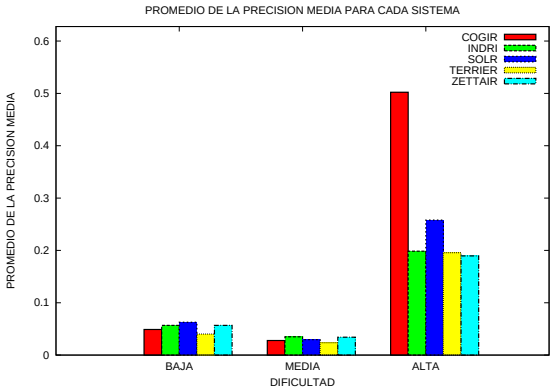

Figura 1.12: PPM sobre CTHJ usando JREL's

En cuanto a los valores obtenidos para PGPM y PREFB, éstos se muestran en las Figs. 1.13 y 1.14. Nuevamente vuelve a repetirse el comportamiento habitual, reflejándose un comportamiento similar en todos los sistemas cuando tratamos consultas con un nivel de dificultad medio o bajo, siendo los resultados mucho mejores en otro caso para el modelo conceptual.

Finalmente, introducimos los valores para GAAR y GAARN en las Figs. 1.15 y 1.16, respectivamente. Al contrario de lo que ocurría en la totalidad de los casos anteriores, aquí los resultados son netamente superiores para el modelo conceptual en el caso de

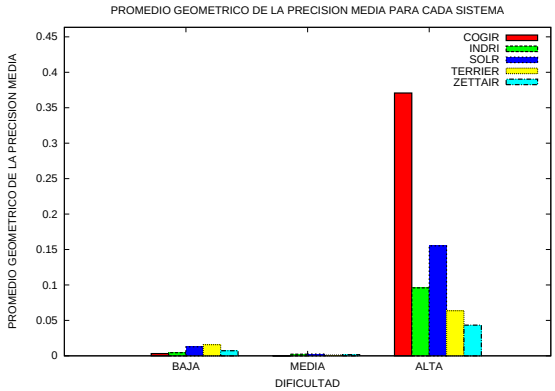

Figura 1.13: PGPM sobre CTHJ usando JREL's

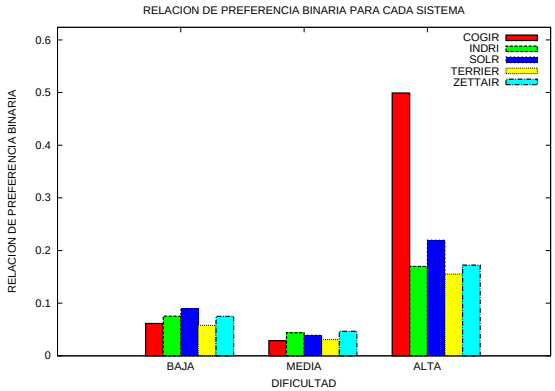

Figura 1.14: PREFB sobre CTHJ usando JREL's

consultas de bajo nivel de dificultad, mientras que en el resto el comportamiento es similar al del conjunto de entornos comparados. Ello no resulta extraño, puesto que ya algunos autores [1] han advertido de los resultados posiblemente sorprendentes en lo que a la correlación con las medidas antes comentadas se refiere.

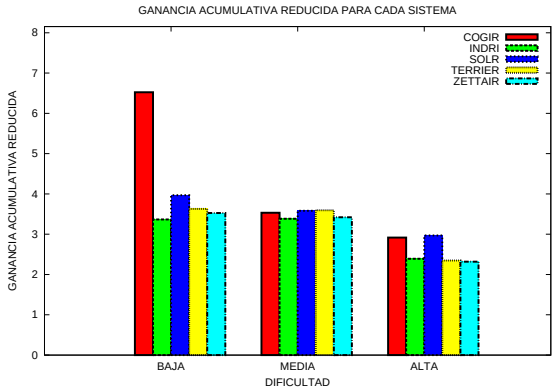

Figura 1.15: GAAR sobre CTHJ usando JREL's

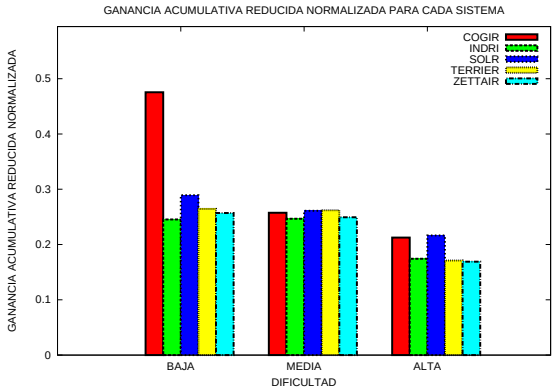

Figura 1.16GAARN sobre CTHJ usando JREL's

7.1.2 | Usando una colección de conjuntos de tópicos basada en la valoración tipo máquina

Ahora aplicamos el mismo conjunto de medidas anteriores sobre el conjunto de tópicos CTMJ. En este caso, el valor del experimento consiste en corroborar las conclusiones ya alcanzadas anteriormente.

Medidas de evaluación basadas en conjuntos

Retomamos aquí el cálculo de las medidas P, C, F y FR, cuyas gráficas son las que se observan en las Figs. 1.17, 1.18, 1.19 y 1.20, respectivamente. En todas ellas, el enfoque conceptual pone de manifiesto un empeoramiento en el funcionamiento sobre el conjunto de tópicos de dificultad alta, en comparación con los de tipo medio y bajo, aunque aún así

logra los mejores resultados para las medidas P y F. Las otras dos medidas resultan estar entre los mejores.

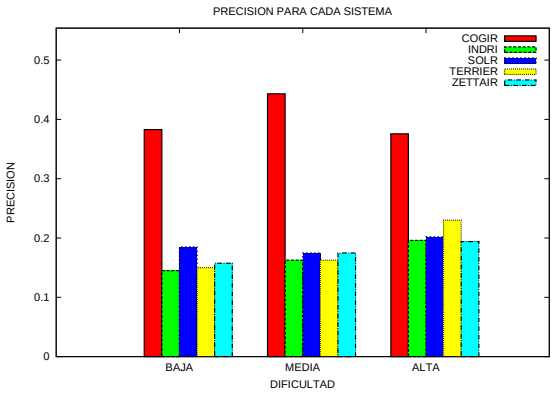

Figura 1.17: P sobre CTMJ usando JREL's

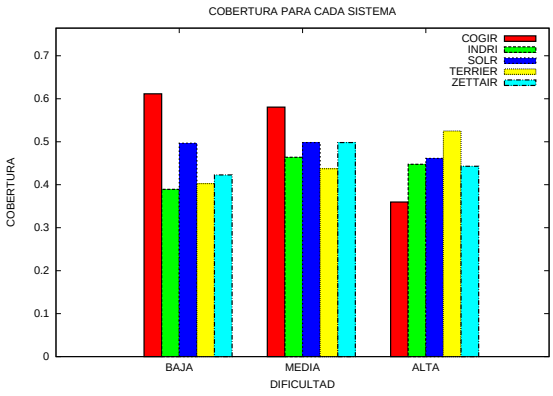

Figura 1.18: C sobre CTMJ usando JREL's

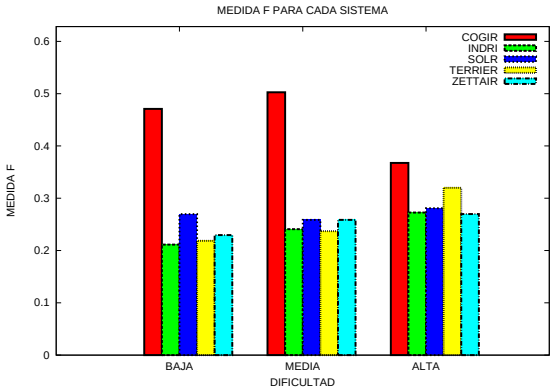

Figura 1.19: F sobre CTMJ usando JREL's

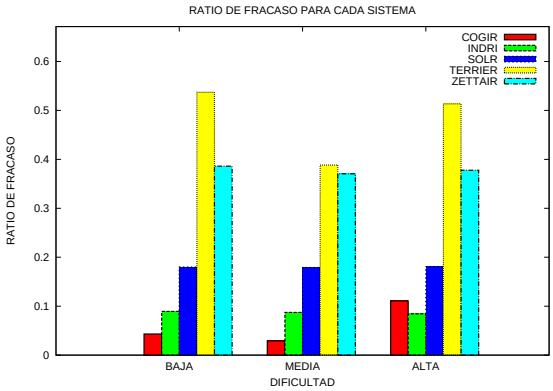

Figura 1.20: FR sobre CTMJ usando JREL's

Medidas de evaluación basadas en ordenación

Recalculamos la P@10, C@10, PI_{C} para los niveles 0 y 0'10 de cobertura, R-P, PPM, PGPM, PREFB, GAAR y GAARN en las Figs. 1.21, 1.22, 1.23, 1.24, 1.25, 1.26, 1.27, 1.28, 1.29 y 1.30, respectivamente. Las figuras muestran como COGIR consigue mejores resultados que los demás sistemas sobre todos los conjuntos de tópicos. Sin embargo, en contraposición a los obtenidos en el caso del CTHJ, proporciona un peor rendimiento sobre los tópicos de dificultad alta en comparación con los de tipo medio y bajo.

7.2 | Sistemas de RI con ordenación usando PJREL's

Seguimos aquí el mismo protocolo aplicado a la ordenación orientada a JREL's, considerando el conjunto total de las diferentes métricas de rendimiento (y en número de catorce) usadas en los anteriores experimentos. La única diferencia es el par de conjuntos de tópicos que usaremos en adelante, remplazando el CTHJ (resp. CTMJ) por CTHPJ (resp.

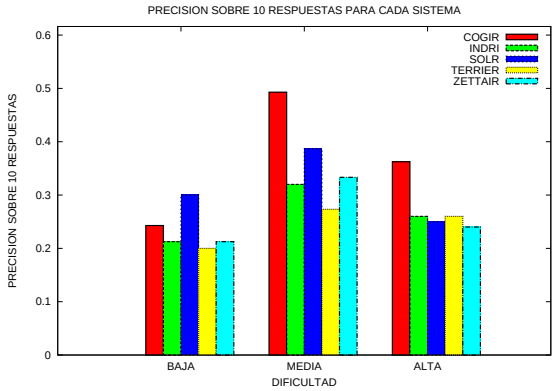

Figura 1.21: P@10 sobre CTMJ usando JREL's

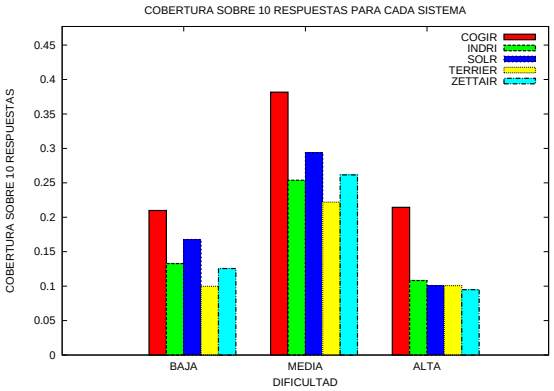

Figura 1.22: C@10 sobre CTMJ usando JREL's

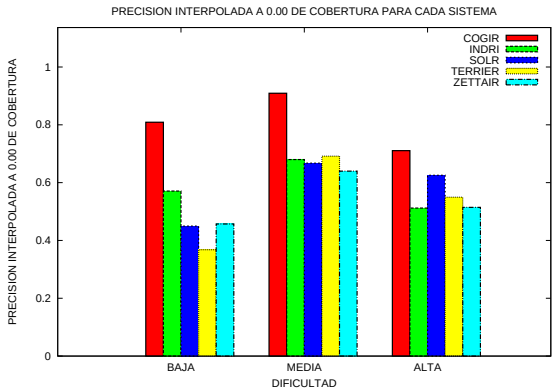

Figura 1.23: $\text{PI}_{C=0'00}$ sobre CTMJ usando JREL's

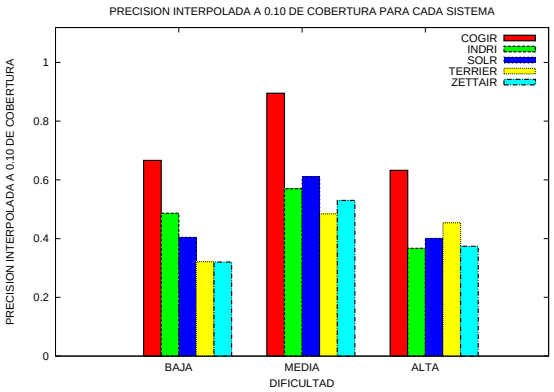

Figura 1.24: $\text{PI}_{C=0'10}$ sobre CTMJ usando JREL's

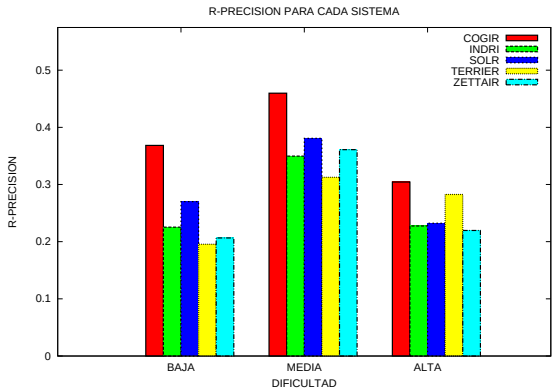

Figura 1.25: R-P sobre CTMJ usando JREL's

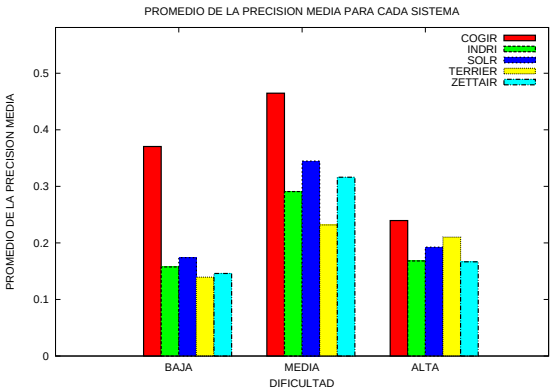

Figura 1.26: PPM sobre CTMJ usando JREL's

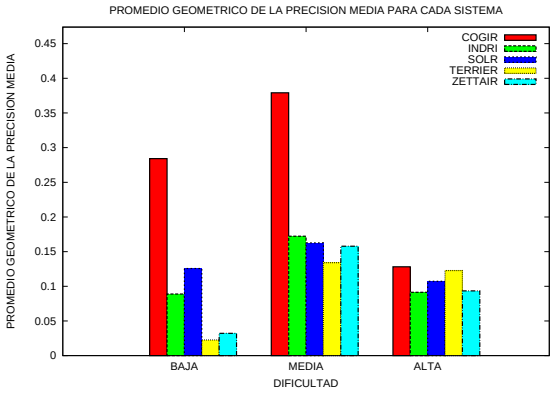

Figura 1.27: PGPM sobre CTMJ usando JREL's

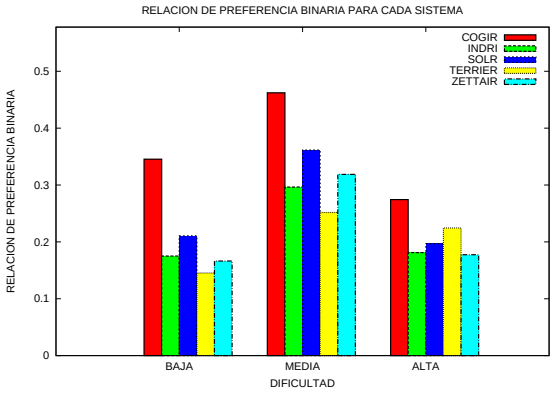

Figura 1.28: PREFB sobre CTMJ usando JREL's

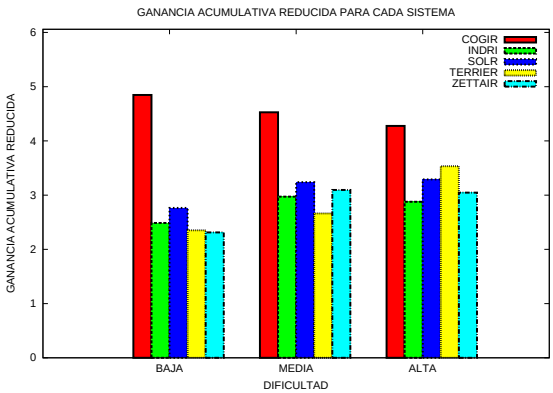

Figura 1.29: GAAR sobre CTMJ usando JREL's

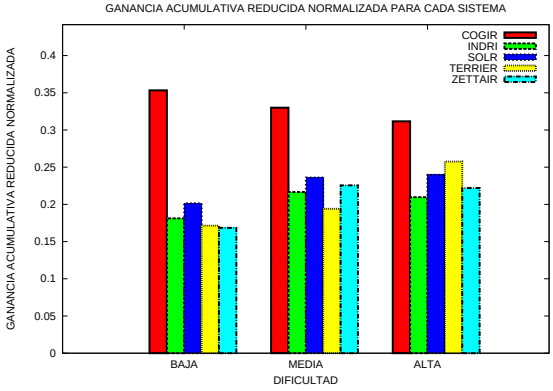

Figura 1.30: GAARN sobre CTMJ usando JREL's

CTMPJ), buscando adecuar el criterio de la selección de tópicos al enfoque específico de ordenación, ambos basados en PJREL's.

7.2.1 | Usando una colección de conjuntos de tópicos basada en la valoración tipo humano

Tomamos aquí CTHPJ como colección de tópicos, lo que nos servirá para proporcionar una visión general de nuestra propuesta para hacer frente a la ordenación basada en PJREL's sobre los tópicos seleccionados usando una valoración de tipo humano.

Medidas de evaluación basadas en conjuntos

Tratamos aquí con los resultados de las medidas P y C, que se muestran en las Figs. 1.31 y 1.32 respectivamente. Los resultados obtenidos constituyen prácticamente un calco de los obtenidos para el caso de los JREL, otorgando nuevamente a COGIR los mejores resultados en cuanto a precisión, manteniendo la contención en la cobertura.

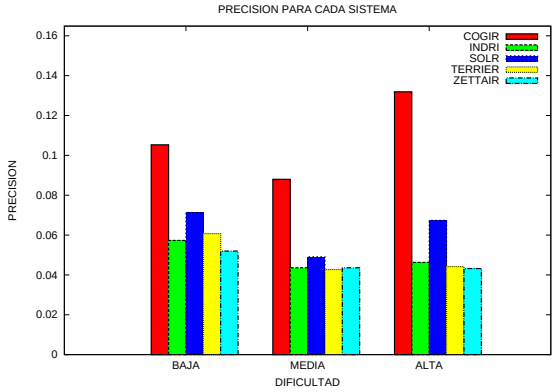

Figura 1.31: P sobre CTHPJ usando PJREL's

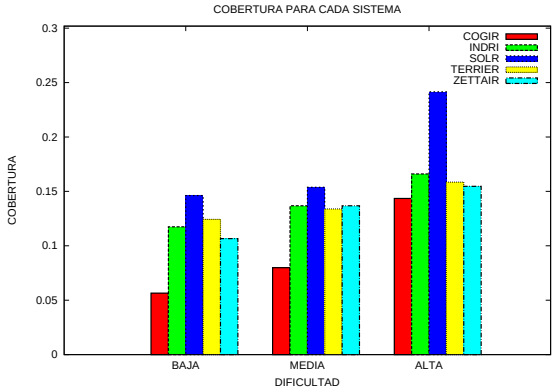

Figura 1.32: C sobre CTHPJ usando PJREL's

Como ya se hizo para los JREL's, también incluimos los resultados de las métricas F y FR en las Figs. 1.33 y 1.34, respectivamente. De nuevo, el modelo conceptual mejora sus resultados sobre el conjunto de tópicos de mayor dificultad, mientras que los resultados son menos impactantes sobre los tópicos con menor poder de discriminación.

Medidas de evaluación basadas en ordenación

Calculamos la P@10, C@10, PI_C para niveles de cobertura 0 y 0'10, R-P, PPM, PGPM, PREFB, GAAR y GAARN en las Figs. 1.35, 1.36, 1.37, 1.38, 1.39, 1.40, 1.41, 1.42, 1.43 y 1.44; respectivamente. Los resultados obtenidos ilustran que COGIR mantiene estable su rendimiento con respecto a los demás motores de búsqueda en el tratamiento de tópicos con mayor dificultad. En este sentido, el uso de PJREL's tiende a favorecer los demás sistemas ya que todos ellos comparten el mismo modelo teórico, lo que provoca listas con resultados similares. Esto repercute en su beneficio, dado que los PJREL's se calculan a partir de dichas listas.

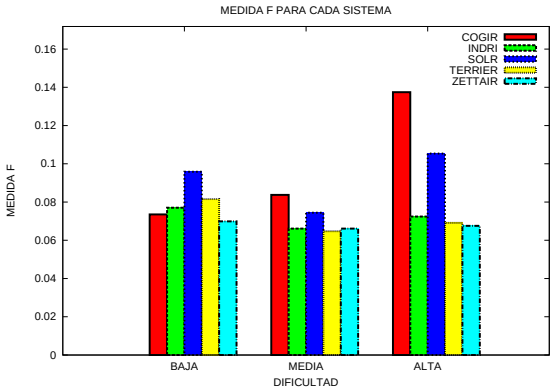

Figura 1.33: F sobre CTHPJ usando PJREL's

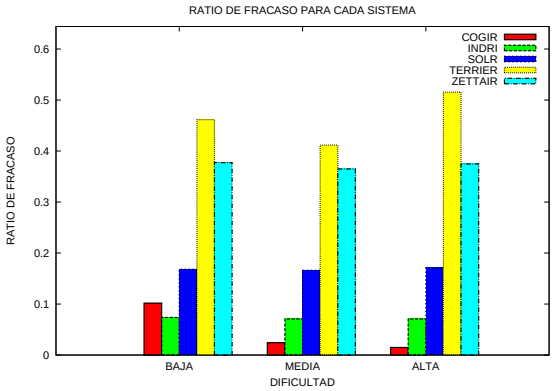

Figura 1.34: FR sobre CTHPJ usando PJREL's

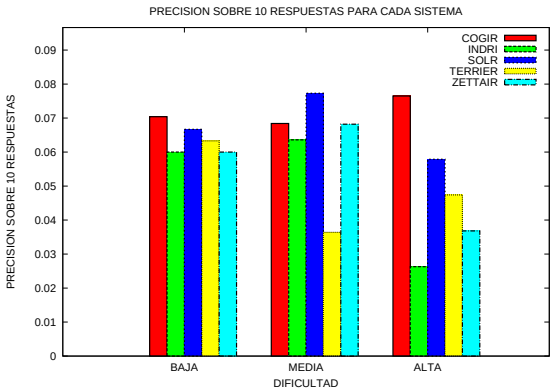

Figura 1.35P@10 sobre CTHPJ usando PJREL's

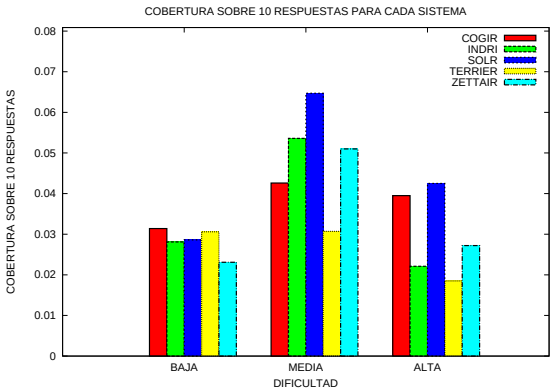

Figura 1.36C@10 sobre CTHPJ usando PJREL's

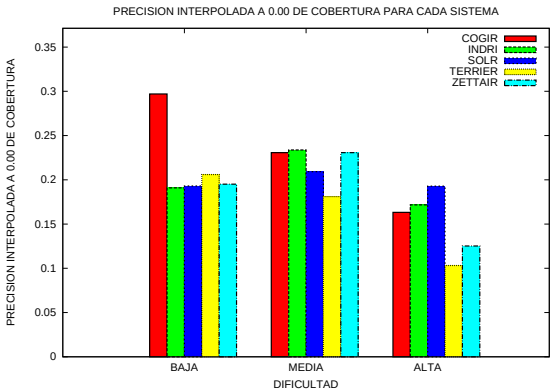

Figura 1.37: $PI_{C=0'00}$ sobre CTHPJ usando PJREL's

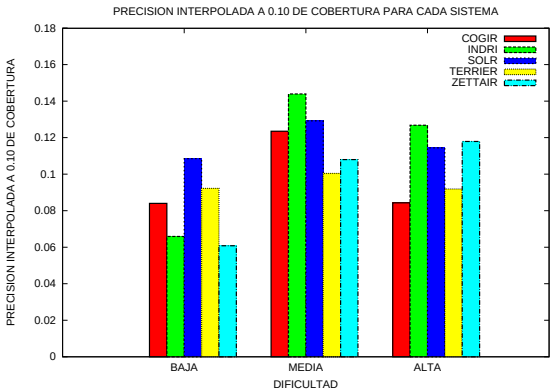

Figura 1.38: $PI_{C=0'10}$ sobre CTHPJ usando PJREL's

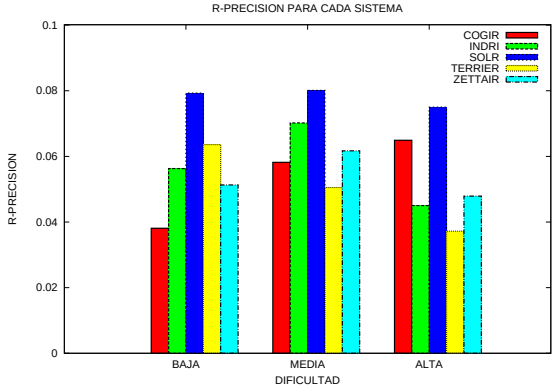

Figura 1.39: R-P sobre CTHPJ usando PJREL's

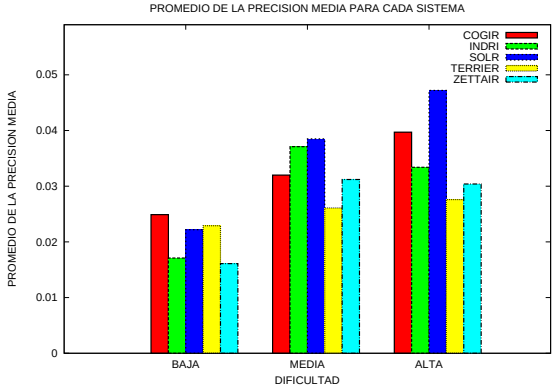

Figura 1.40: PPM sobre CTHPJ usando PJREL's

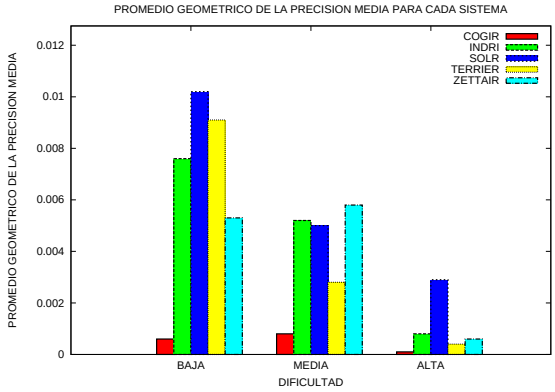

Figura 1.41: PGPM sobre CTHPJ usando PJREL's

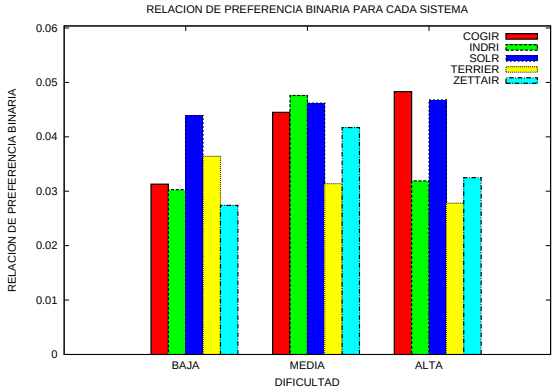

Figura 1.42: PREFB sobre CTHPJ usando PJREL's

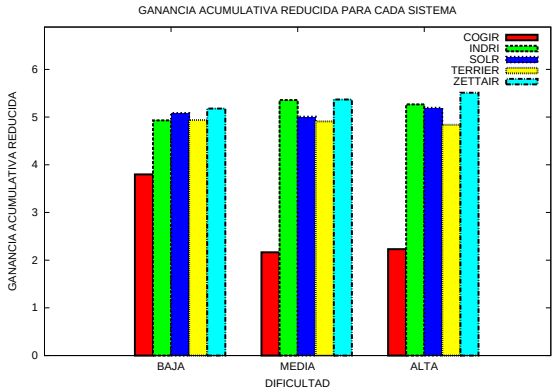

Figura 1.43: GAAR sobre CTHPJ usando PJREL's

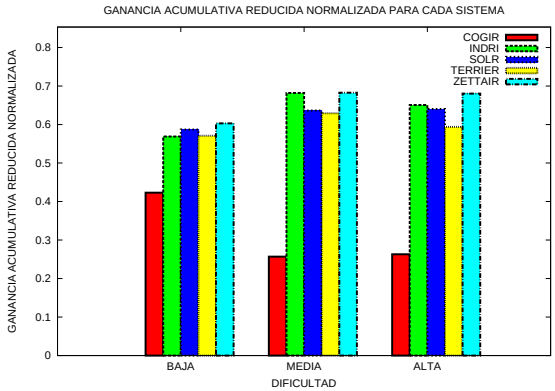

Figura 1.44: GAARN sobre CTHPJ usando PJREL's

7.2.2 | Usando una colección de conjuntos de tópicos basada en la valoración tipo máquina

Ahora calculamos el mismo conjunto de medidas anteriores sobre el conjunto de tópicos CTMPJ. En este caso, el valor del experimento consiste en corroborar las conclusiones alcanzadas con anterioridad.

Medidas de evaluación basadas en conjuntos

Los resultados obtenidos para las medidas P, C, F y FR se muestran en las gráficas de las Figs. 1.45, 1.46, 1.47 y 1.48 respectivamente. Proporcionan valores que claramente favorecen a los demás sistemas con respecto a COGIR sobre los tópicos de dificultad baja y media. Sin embargo, en el caso de los tópicos con mayor poder de discriminación, nuestra propuesta consigue mantener su posición con respecto a los ya comentados para el conjunto de tópicos CTHPJ.

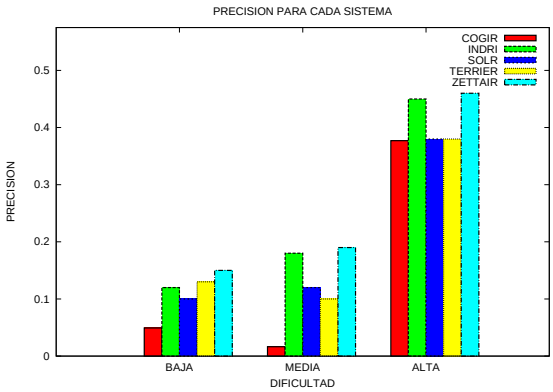

Figura 1.45: P sobre CTMPJ usando PJREL's

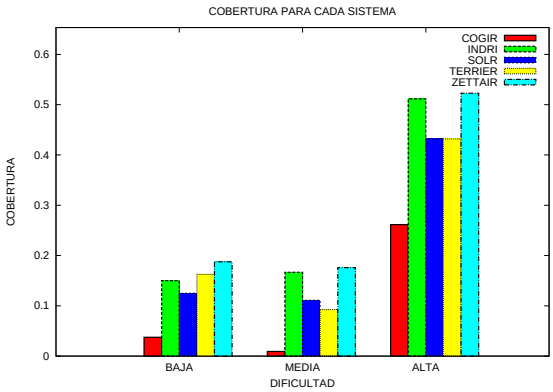

Figura 1.46: C sobre CTMPJ usando PJREL's

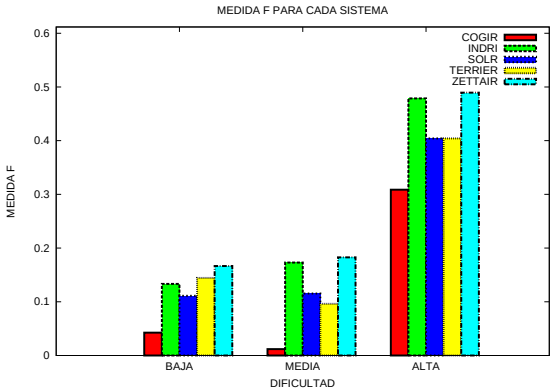

Figura 1.47: F sobre CTMPJ usando PJREL's

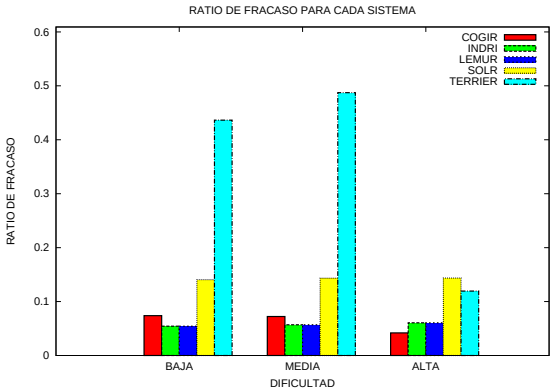

Figura 1.48: FR sobre CTMPJ usando PJREL's

Medidas de evaluación basadas en ordenación

Calculamos la P@10, C@10, PI_{C} para niveles de cobertura 0 y 0'10, R-P, PPM, PGPM, PREFB, GAAR y GAARN en las Figs. 1.49, 1.50, 1.51, 1.52, 1.53, 1.54, 1.55, 1.56, 1.57 y 1.58; respectivamente. Las pruebas sugieren que los resultados obtenidos sobre los tópicos de dificultad baja y media son peores en el caso del motor de búsqueda COGIR. Los obtenidos en el intervalo superior de dificultad se mantienen más o menos en la misma línea que para el conjunto de tópicos CTHPJ, aquí también penalizado por el uso de PJREL. Al igual que para aquel conjunto de tópicos, el enfoque conceptual no supera a sus competidores.

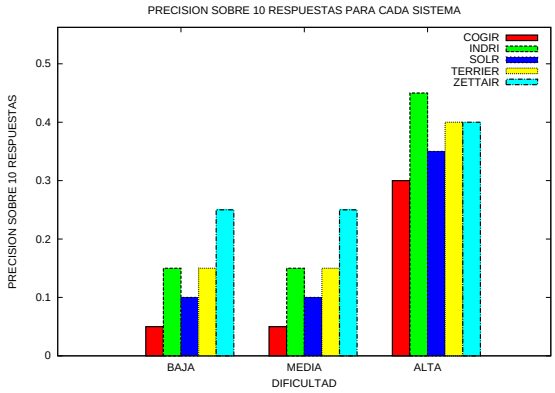

Figura 1.49 P@10 sobre CTMPJ usando PJREL's

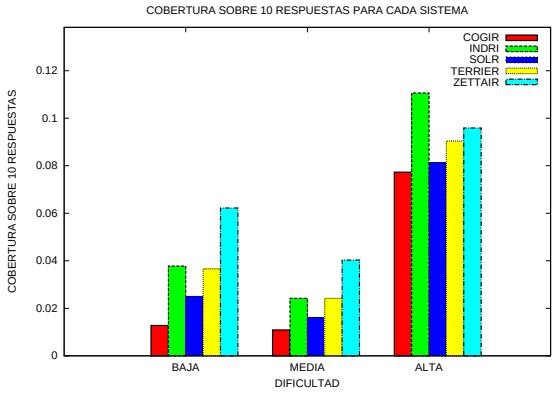

Figura 1.50 C@10 sobre CTMPJ usando PJREL's

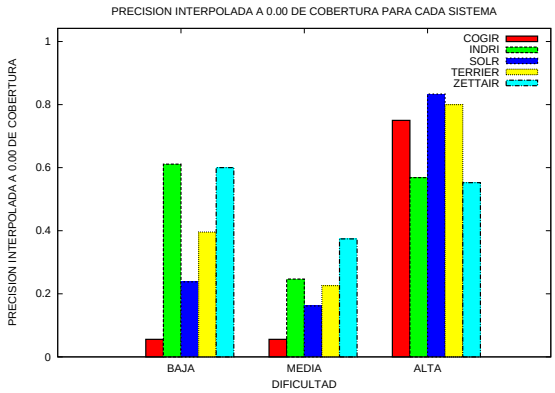

Figura 1.51: $\text{PI}_{\text{C}=0'00}$ sobre CTMPJ usando PJREL's

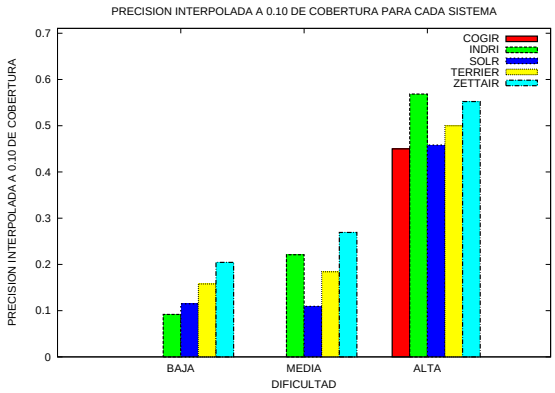

Figura 1.52: $\text{PI}_{\text{C}=0'10}$ sobre CTMPJ usando PJREL's

7.3 | Sistemas de RI con ordenación usando valoración tipo máquina

Como ya se ha dicho, el punto de partida de esta técnica de ordenación [101] es la medida PM, lo que implica que es necesario un cierto número de juicios de relevancia para iniciar el proceso. Teniendo en cuenta que previamente los hemos introducido como

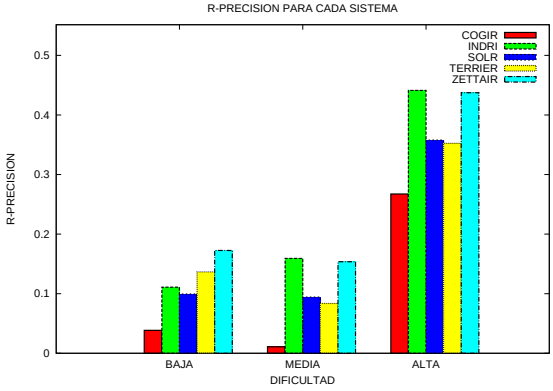

Figura 1.53: R-P sobre CTMPJ usando PJREL's

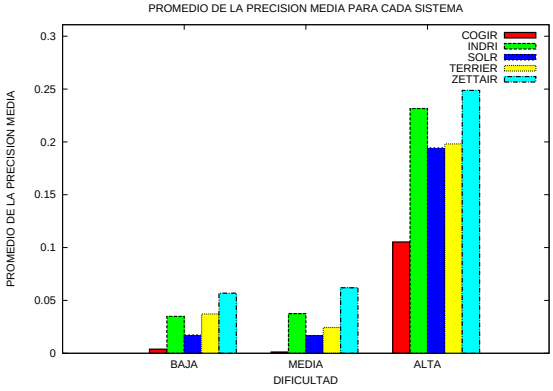

Figura 1.54: PPM sobre CTMPJ usando PJREL's

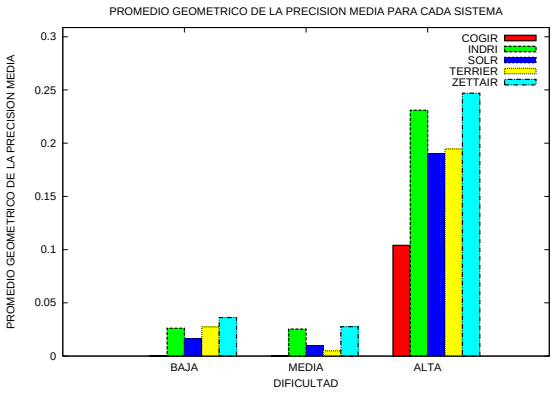

Figura 1.55PGPM sobre CTMPJ usando PJREL's

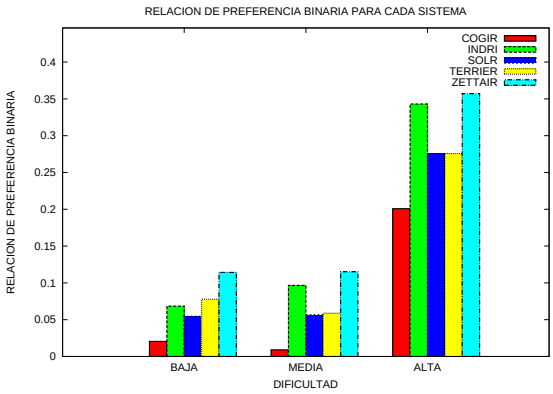

Figura 1.56: PREFB sobre CTMPJ usando PJREL's

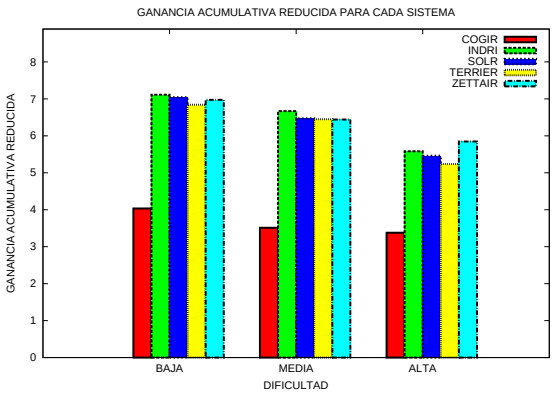

Figura 1.57: GAAR sobre CTMPJ usando PJREL's

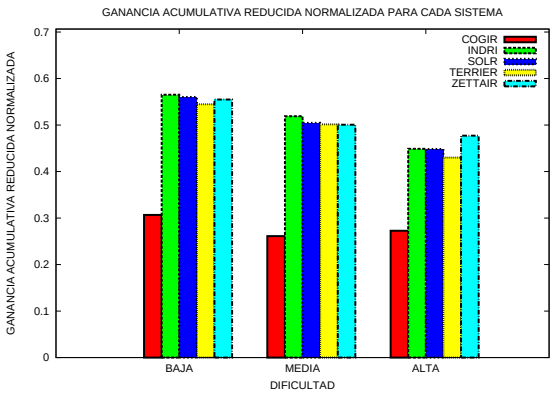

Figura 1.58: GAARN sobre CTMPJ usando PJREL's

estrategias de enjuiciamiento, experimentamos en este nivel tanto con JREL's como con PJREL's.

7.3.1 | Calculando la PM a partir de JREL's

Como ya se había hecho para la clasificación basada en JREL's, en este punto podemos diferenciar dos series de tests, uno por cada conjunto de tópicos construido a partir de JREL's: CTHJ y CTMJ.

Usando una colección de conjuntos de tópicos basada en la valoración tipo humano

En este punto, vamos a probar una ordenación usando una valoración tipo máquina sobre la colección de tópicos tipo humano CTHJ. Los resultados para la medida A se muestran en la Fig. 1.59, dando nuevamente una ventaja al motor de búsqueda conceptual sobre el resto, en especial en el caso de los tópicos con menor y mayor poder de discriminación. De hecho, aunque los peores resultados de COGIR se refieren a los tópicos de dificultad media, aún en este caso su rendimiento mejora el mostrado por cualquiera de los demás sistemas que, en general, muestran un mejor comportamiento justamente sobre ese conjunto de tópicos.

Usando una colección de conjuntos de tópicos basada en la valoración tipo máquina

Probamos ahora una ordenación basada en la valoración tipo máquina sobre la colección de tópicos tipo máquina CTMJ. Los resultados para la medida A se muestran en la Fig. 1.60. Los resultados corroboran el comportamiento previamente observado sobre la colección de tópicos CTHJ.

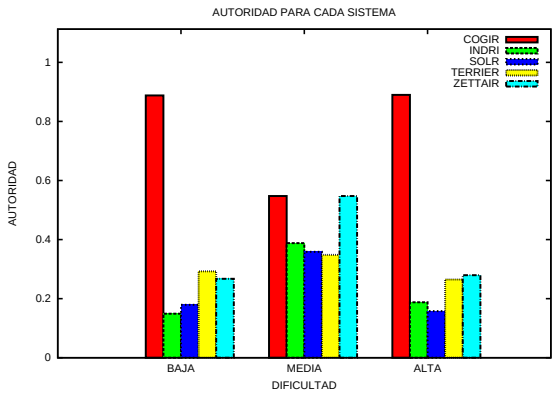

Figura 1.59: A sobre CTHJ usando JREL's

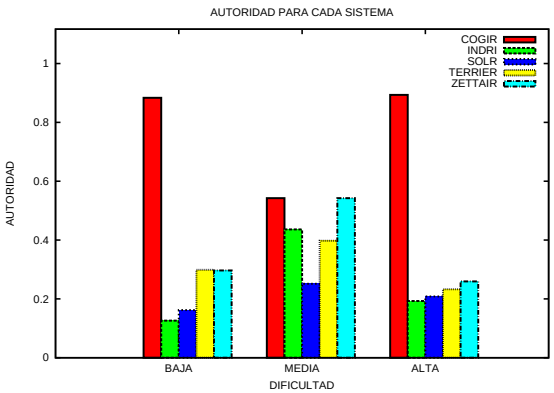

Figura 1.60: A sobre CTMJ usando JREL's

7.3.2 | Calculando la PM a partir de PJREL's

Siguiendo el mismo protocolo descrito para la PM calculada a partir de JREL's, aquí consideramos dos series de pruebas, uno por cada conjunto de tópicos construido a partir

de PJREL's: CTHPJ y CTMPJ.

Usando una colección de conjuntos de tópicos basada en la valoración tipo humano

Probamos ahora una ordenación basada en la valoración tipo máquina usando una colección de conjuntos de tópicos basada en la valoración tipo humano (CTHPJ). Los resultados para la medida A se muestran en la Fig. 1.61. Desde un punto de vista cualitativo, el rendimiento observable en relación a COGIR es análogo al previamente descrito en el caso en el que PM se calculaba a partir de JREL's.

Usando una colección de conjuntos de tópicos basada en la valoración tipo máquina

El turno corresponde ahora a la ordenación basada en la valoración tipo máquina usando una colección de conjuntos de tópicos también basada en el mismo tipo de valoración tipo máquina (CTMPJ). Los resultados para la medida A se muestran en la Fig. 1.62. Aunque el mejor funcionamiento continúa correspondiendo a COGIR, contrariamente a las anteriores gráficas para el caso de la A, en este caso los peores resultados para el modelo conceptual se obtienen en el conjunto de tópicos de mayor dificultad.

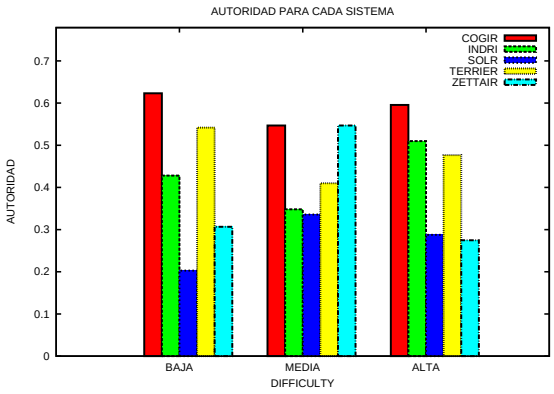

Figura 1.61: A sobre CTHPJ usando PJREL's

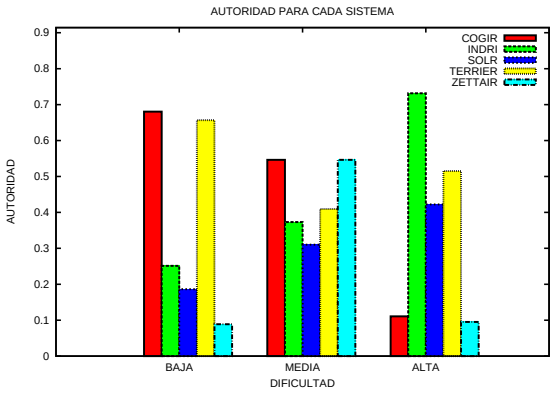

Figura 1.62: A sobre CTMPJ usando PJREL's

7.4 | Sistemas de RI con ordenación usando la media de contadores de referencia ponderados

La última propuesta de ordenación que consideramos fue descrita por Wu *et al.* en [164] y se basa en el concepto de la media de contadores de referencia ponderados. Como ya hemos introducido, se pueden considerar aquí cuatro medidas: $MCRP_O$, $MCRP_P$, $MCRP_{OL}$ y $MCRP_{PL}$.

Dado que en este caso, la estrategia de ordenación no está relacionada con ninguna estrategia de enjuiciamiento en particular, vamos a considerar la gama completa de

conjuntos de tópicos previamente introducidos con el fin de asegurar un procedimiento completo de prueba: CTHJ, CTMJ, CTHPJ y CTMPJ. Esto nos va a permitir considerar tanto la valoración de tipo humano como la de tipo máquina para seleccionar los tópicos, además de las técnicas basadas en JREL's y en PJREL's con el fin de reducir el tamaño de esos conjuntos. De esta manera, pretendemos no favorecer a ninguna estrategia que pudiera ser usada para afinar algunos de los sistemas de RI que se están comparando, un aspecto importante a tener en cuenta cuando se considera un método de ordenación, cuyo punto de partida es el recuento de referencias cruzadas entre el conjunto de documentos devueltos por los motores de búsqueda.

7.4.1 | Usando la reducción de tópicos basados en JREL's

Experimentaremos primero con conjuntos de tópicos obtenidos a partir de técnicas de reducción de tópicos basados en JREL's, que incluyen tanto a las colecciones de conjuntos de tópicos de tipo humano como máquina.

Usando una colección de conjuntos de tópicos basada en la valoración tipo humano

En esta ocasión, los resultados se muestran para las métricas $MCRP_O$, $MCRP_P$, $MCRP_{OL}$ y $MCRP_{PL}$ sobre el conjunto de tópicos CTHJ en las Figs. 1.63, 1.64, 1.65 y 1.66, respectivamente. En estos casos, el enfoque conceptual aparentemente muestra el peor comportamiento posible, especialmente cuando se trata de tópicos de dificultad alta, si bien los resultados son un poco mejores para las medidas $MCRP_O$ y $MCRP_{OL}$. Contrariamente a lo que uno pudiera pensar, tal comportamiento es no sólo congruente con las anteriores medidas sino perfectamente previsible.

En efecto, al aplicar técnicas relativistas, el sistema de RI objeto de test no podría en ningún caso mejorar las prestaciones del conjunto de los que le sirven de referencia comparativa. Es más, este tipo de metodologías puede llevar a situaciones estrepitosamente erróneas cuando el conjunto de esos sistemas referentes muestra un rendimiento común pobre sobre un conjunto de tópicos, mientras que el sistema testeado ofrece una buena precisión. Es justamente el comportamiento que podemos observar en este caso sobre el conjunto de tópicos de mayor dificultad, que hemos visto favorecía al acercamiento conceptual en todas las métricas anteriores y que ahora, por el contrario, parecería mostrar un peor comportamiento.

Usando una colección de conjuntos de tópicos basada en la valoración tipo máquina

Los resultados se muestran ahora para las medidas $MCRP_O$, $MCRP_P$, $MCRP_{OL}$ y $MCRP_{PL}$ sobre el conjunto de tópicos CTMJ, en las Figs. 1.67, 1.68, 1.69 y 1.70 respectivamente. Podemos hacer extensivos exactamente los mismos comentarios previamente realizados con las pruebas sobre el conjunto de tópicos CTHJ, corroborando el razonamiento realizado más allá el tipo de valoración aplicado en la selección de tópicos.

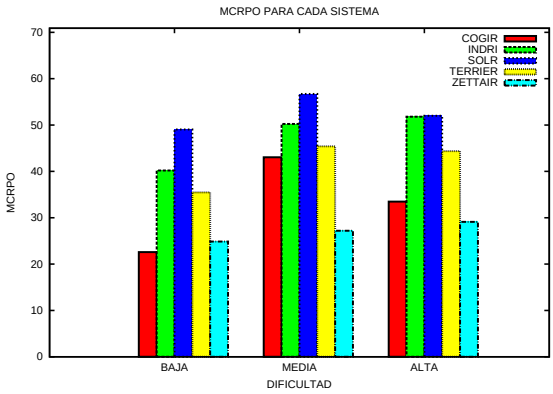

Figura 1.63: $MCRP_O$ sobre CTHJ

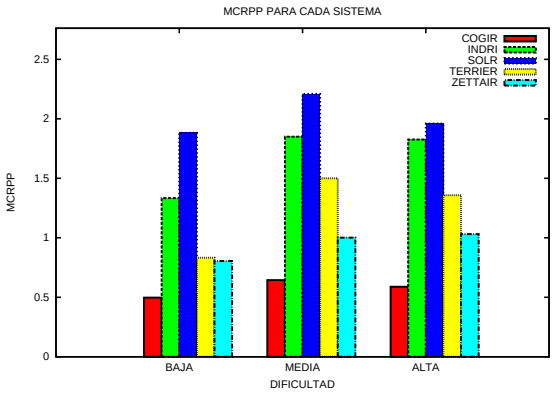

Figura 1.64: $MCRP_P$ sobre CTHJ

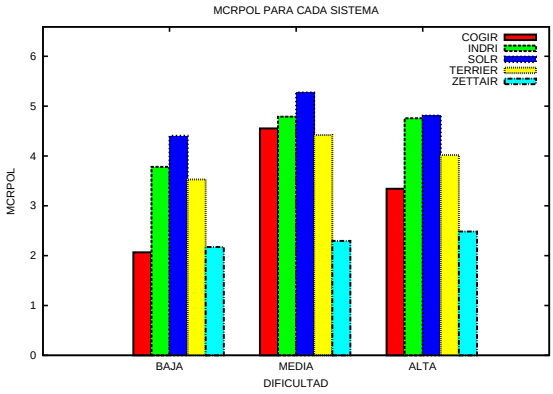

Figura 1.65: $MCRP_{OL}$ sobre CTHJ

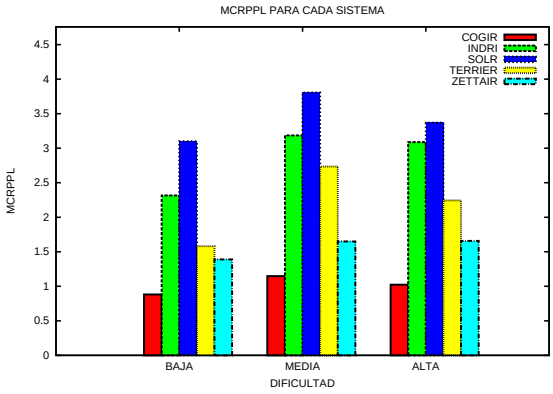

Figura 1.66: $MCRP_{PL}$ sobre CTHJ

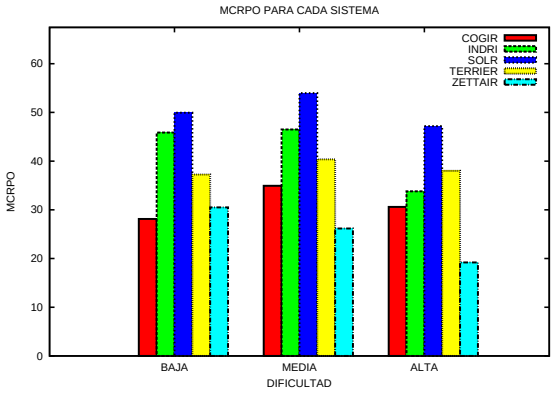

Figura 1.67: $MCRP_O$ sobre CTMJ

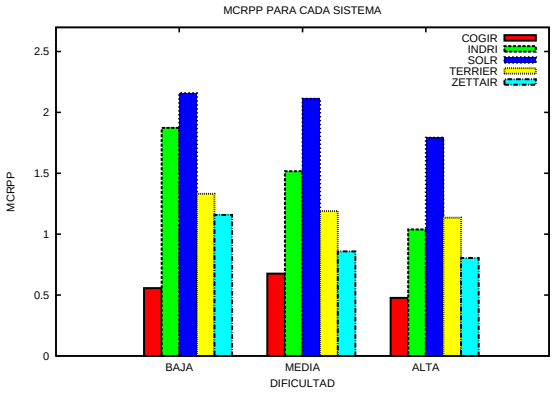

Figura 1.68: $MCRP_P$ sobre CTMJ

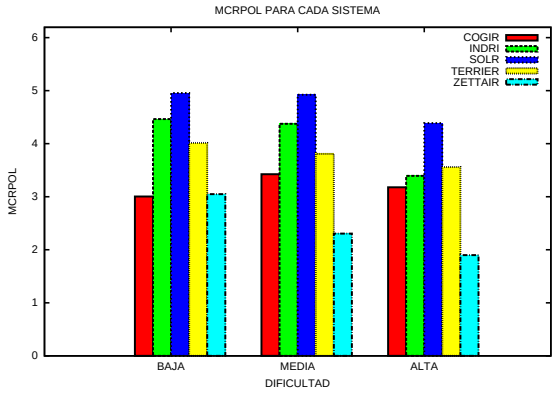

Figura 1.69: MCRP_{OL} sobre CTMJ

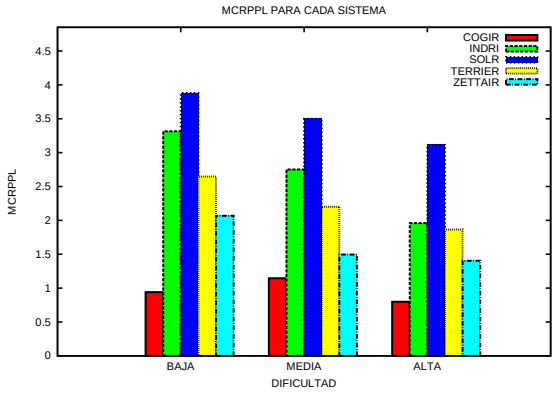

Figura 1.70: MCRP_{PL} sobre CTMJ

7.4.2 | Usando la reducción de tópicos basados en PJREL's

Los experimentos están relacionados ahora con los conjuntos de tópicos obtenidos a partir de PJREL's basados en métodos de reducción, incluyendo colecciones de conjuntos de tópicos tanto de tipo humano como máquina.

Usando una colección de conjuntos de tópicos basada en la valoración tipo humano

Como para el caso anterior de los JREL's, los resultados se muestran para las medidas MCRP_{O}, MCRP_{P}, MCRP_{OL} y MCRP_{PL} sobre el conjunto de tópicos CTHPJ, en las Figs. 1.71, 1.72, 1.73 y 1.74 respectivamente. Los resultados mostrados en las gráficas son cualitativamente equivalentes a los previamente comentados para la reducción de tópicos basados en JREL's, aunque existe una diferencia sustancial. Esto es, el modelo conceptual obtiene los mejores resultados para el conjunto de los tópicos difíciles, cuando para los casos anteriores conseguía los peores.

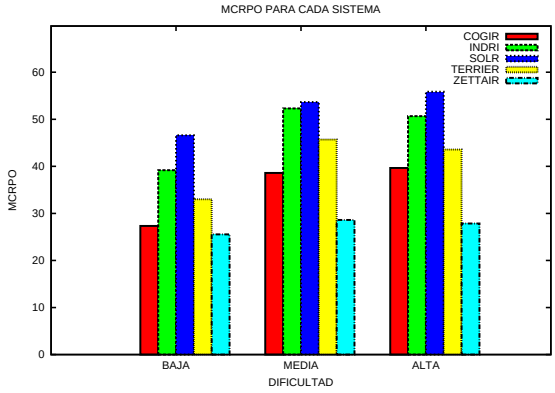

Figura 1.71: MCRP_{O} sobre CTHPJ

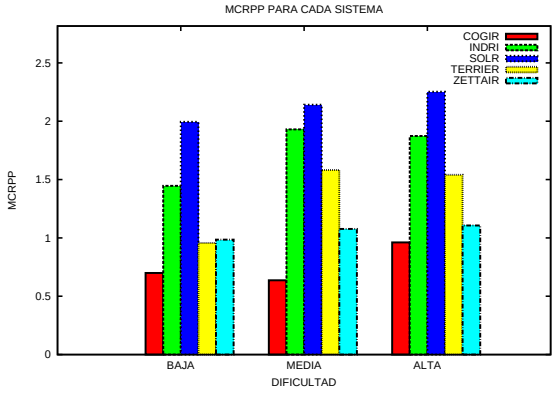

Figura 1.72: MCRP_{P} sobre CTHPJ

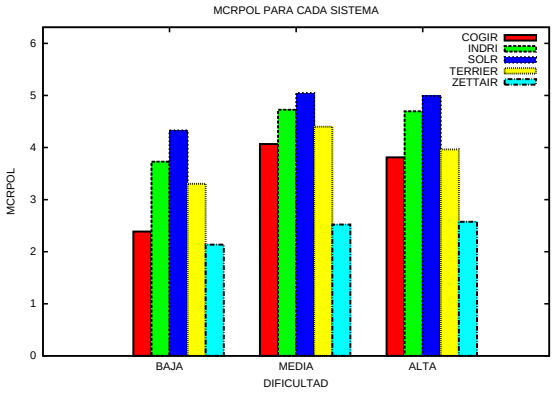

Figura 1.73: MCRP_{OL} sobre CTHPJ

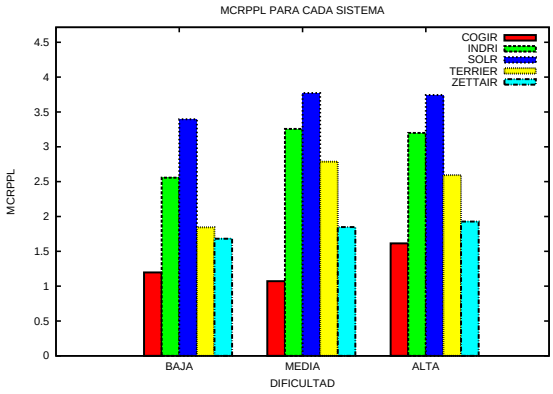

Figura 1.74: MCRP_{PL} sobre CTHPJ

Usando una colección de conjuntos de tópicos basada en la valoración tipo máquina

Los valores se muestran ahora para las medidas MCRP_{O}, MCRP_{P}, MCRP_{OL} y MCRP_{PL} sobre el conjunto de tópicos CTMPJ, en las Figs. 1.75, 1.76, 1.77 y 1.78, respectivamente. Los resultados experimentales son aquí cuantitativamente equivalentes a los comentados anteriormente, aunque sensiblemente diferentes desde un punto de vista cualitativo. En particular, al contrario de las pruebas anteriores, se obtienen los peores resultados para el enfoque conceptual en el caso de las medidas MCRP_{O} y MCRP_{OL}, considerando el conjunto de tópicos de dificultad baja. En relación con las métricas MCRP_{P} y MCRP_{PL}, los resultados son equivalentes a los obtenidos para el caso de la colección de conjuntos de tópicos basada en la valoración humana.

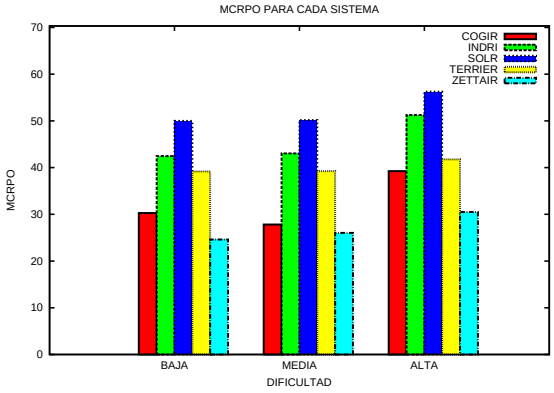

Figura 1.75: MCRP_{O} sobre CTMPJ

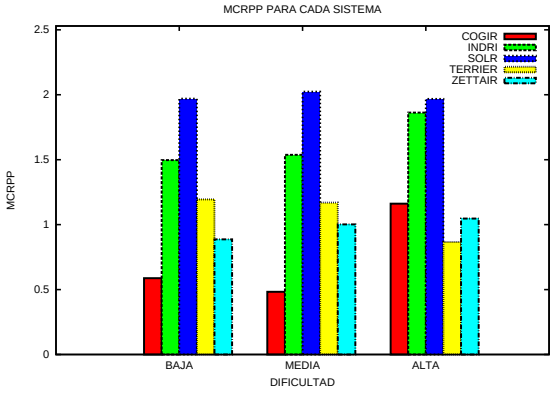

Figura 1.76: MCRP_{P} sobre CTMPJ

8 | Conclusiones

La conveniencia de la inclusión o no de conocimiento lingüístico específico en el diseño de motores de búsqueda, es una discusión que se remonta a los orígenes del propio

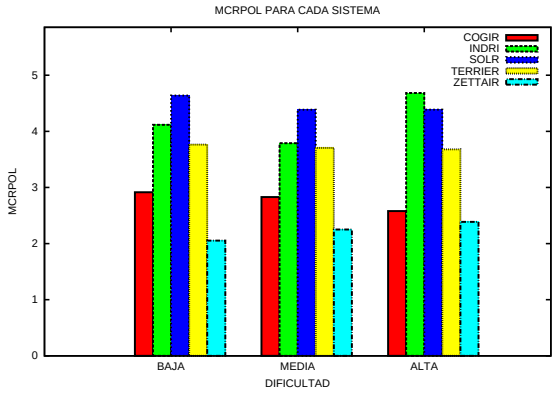

Figura 1.77: MCRP_{OL} sobre CTMPJ

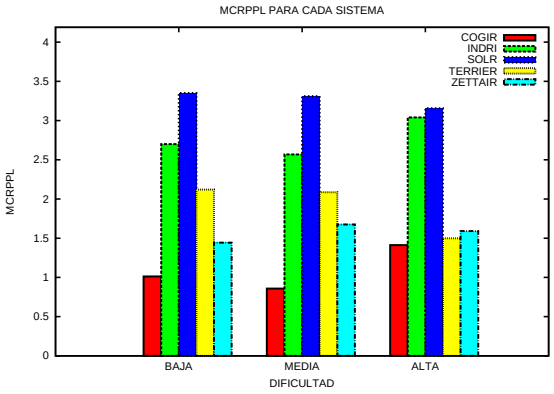

Figura 1.78: MCRP_{PL} sobre CTMPJ

ámbito de la RI. Habitualmente tres han sido las razones argumentadas para obviar esta cuestión: la complejidad algorítmica asociada, la escasez o incluso carencia de recursos lógicos y el aparentemente escaso rendimiento extra asociados a su consideración. Asumida la complejidad técnica de este tipo de estrategias, introducimos una metodología para la adquisición automática de la semántica del texto a partir de la información léxica y sintáctica resumidas en un grafo conceptual que no es sino el reflejo del conjunto de relaciones de dependencia previamente reconocidas. Ello nos permite no sólo disponer de una estructura formal que traslada fielmente el significado de un documento cualquiera, sino que también proporciona una base estructural idónea sobre la que sustentar un algoritmo de correspondencia de patrones aproximado capaz de estimar la proximidad semántica entre dos textos diferentes.

Pretendemos, además, arrojar alguna luz práctica en relación a lo que intuitivamente parece obvio, que una base semántica mejorada en el proceso de recuperación debería tener su reflejo en el rendimiento observado. Para ello, hemos definido un completo entorno de evaluación formal siguiendo lo que, a nuestro conocimiento, constituye una completa muestra de las técnicas actualmente disponibles. Ello nos ha permitido exprimir en profundidad las posibilidades de los acercamientos de RI conceptual, frente a la vocación más genérica de los motores de búsqueda clásicos.

Los resultados obtenidos parecen zanjar definitivamente la discusión por cuanto muestran un rendimiento que, en el peor de los casos, iguala al de los entornos basados en conjuntos de palabras independientemente de cual sea la base de implementación. Como única excepción observada, señalar los resultados de los tests basados en el uso de PJREL's, que favorecen naturalmente a las arquitecturas asociadas a los sistemas de RI que han sido utilizados como referencia para la generación de tales estructuras.

Además, se observa sistemáticamente un salto cualitativo importante cuando se trata de resolver consultas catalogadas como de dificultad creciente en su respuesta y que nosotros asociamos con tópicos con mayor poder de discriminación entre los sistemas comparados.

Intuitivamente este resultado coincide con lo esperable, puesto que la información semántica se revela determinante cuanto más complejo es el significado del texto a analizar, tanto en lo que se refiere a la colección documental a explorar como a la del tópico. Por el contrario, cuando la simplicidad de la interrogación o del propio contenido de los textos estudiados permite prescindir de relaciones semánticas complejas, todos los sistemas objeto de estudio presentan un rendimiento equiparable independientemente del tipo de arquitectura de indexación considerada.

Abstract

We introduce a frame for information retrieval combining natural language processing and knowledge domain. We cover the entire process of creating, managing and querying a document database from a standpoint that automatically integrates linguistic knowledge into a formal model of semantic representation that can be managed directly by the system. This makes it possible to construct algorithms that simplify maintenance tasks, offer non-specialist users greater flexibility of access, eliminate subjective components that lead to behaviors that are hard to predict, and allow for the implementation of effective mechanisms for monitoring system performance.

The linguistic knowledge acquisition process starts from a dependency parse based on a mildly context-sensitive grammatical formalism. We thereby combine computational efficiency, expressive power and an outstanding treatment of non-determinism in dynamic programming. Formal semantic interpretation is based on the notion of a conceptual graph, used both to represent the document collection and the queries used to search it. Within the context of Harris's hypothesis [65], such representations are generated from the linguistic information contained in the texts themselves, and are the starting point for their indexation.

Graph operations are used for computing and ranking answers. An approximate pattern-matching mechanism based on the projection and generalization of graphs makes it possible to take into account the intrinsic vagueness and incompleteness of information retrieval, locating the problem in a decidable framework, in accordance with *van Rijsbergen's logical uncertainty principle*. Furthermore, the visual aspect of graphs enables the construction of friendly graphical user interfaces, reconciling accuracy and intuition in data management.

Compared to previous applications of conceptual graph theory in developing information retrieval systems, our proposal resolves the automatic generation of semantic representations from text, thereby avoiding the tedious task of manual indexing. Also, to the best of our knowledge, it is the first time that a formal testing frame has been defined

in this domain.

1 | Introduction

Globalized access has justified the popularization of *information retrieval (*IR*) systems*, converting their development into a major challenge for the scientific community. Such tools are based on the ability to discern, with respect to a standing query, between the contents of a document database that are relevant and those that are not. More specifically, the *relevance* of a document is determined by a match between the representation of its content and that corresponding to the content of the query, so the way these contents are represented is crucial. In this regard, a robust scalable framework is required to represent the information extracted from documents, enabling its visualization and query, which leads us directly to the concept of ontology in the knowledge engineering and *artificial intelligence* sense [35], i.e., to that of a framework for the domain knowledge of an intelligent system.

The majority of existing IR systems use the *vector space model* (VSM) [129], focusing primarily on word co-occurrence, or what is typically referred to as *bag-of-words* (BOW) [64]. This implies that documents and queries are represented as word lists that do not express anything about context information, often ignoring the sequential aspect of word occurrences in texts[1], although the meaning of natural languages strongly depends on them.

Historically, BOW-based retrieval has been intended as a means of storing data more than a means of describing them together with their relationships. It emphasizes the questions of how data may be stored or accessed, but largely ignores what a collection of documents means to people [137], which implies that important semantic information can be lost, overlooking the meanings and ideas their authors want to convey. As a consequence, although these techniques have proved robust and efficient for a large variety of texts, in difficult retrieval tasks it is commonly assumed that one needs access to a wealth of background knowledge in order to improve *precision and recall*. In fact, the growing amount of textual data available electronically has increased this need for high performance retrieval, which serves as motivation to look for strategies that search for and/or filter information based on some higher level of understanding that can only be achieved through text processing in which semantics are taken into account. As a starting point, sentences, and not simply BOW, seem to be the natural way to enhance retrieval performance over the common document models [42].

The hypothesis is that with an adequate representation of documents, incorporating limited semantic knowledge, it is possible to improve the effectiveness of an IR system. This first requires an in-depth text analysis, thus placing the problem squarely within the

[1] even the most basic structures, such as the order of the terms in the document or the frontiers between sentences or paragraphs, are ignored.

framework of *natural language processing* (NLP), albeit with two specific characteristics. The first of these concerns the amount of text that an IR system has to deal with. It is so huge and heterogeneous that it becomes impractical, given the current state of the art, to exhaustively carry out such analysis. The second characteristic, however, serves to moderate the requirements derived from the first, in that a detailed and accurate semantic analysis is not required for IR tasks [75], thereby differentiating them from other NLP related ones such as machine translation, query answering or text summarizing [146]. In this context, we consider a two-step strategy to deal with text analysis at the sentence level. The first step refers to the acquisition of lexical knowledge, a task for which we take our inspiration from the *Alexina* architecture [120], a frame whose kernel is a finite-state-based lexer that integrates a *pre-processor* [122] providing tokenization, spelling correction and name-entity recognition, taking as primary resource a large-scale lexicon [121]. The output, including all possible interpretations for each lexical form in a *directed acyclic graph*, is ready to be processed in a parsing phase that constitutes the second step. Here, we have chosen a *mildly context-sensitive formalism* [153] which provides enough power to allow for the analysis of natural languages without sacrificing computational efficacy.

Nevertheless, in order to obtain the maximum benefit from text analysis in dealing with IR tasks, we also need to dispose of a formal notation that serves as an intermediary between the human and the computer. Here, *conceptual graphs* (CGs) [137] have the necessary potential to describe the meaning of the data according to the user's view, whilst at the same time enabling us to associate them with procedures that can access the data according to the machine view. We can therefore avoid people having to learn strange conventions in order to access the data and to interpret both final and partial results, an advantage that BOW-based retrieval lacks. On the other hand, conceptual similarity seems to be able to estimate the *semantic granularity* of a document [168], referred to as the level of detail carried by an information item [51], opening the doors to the treatment of searching tasks involving vague queries, incomplete document databases and domain-specific IR. All the above justifies our choice of this kind of structure as semantic representation formalism.

Formally, we derive CGs according to a dependency model. The document collection is first analyzed by the parse in order to generate a *primary graph of syntactic dependencies* that we later translate into *governor/governed ones*, relating the head of a syntagm with its modifiers. From here, and with the addition of a set of initial values provided by the programmer for the semantic classes (types), the consideration of linguistic tags and syntactic patterns allows us to approach and accurately extend both initial sets of dependencies and classes. Careful implementation in dynamic programming makes it possible to postpone the treatment of ambiguities, whether of a lexical or syntactic nature, to a subsequent semantic definition stage, in which an iterative knowledge acquisition protocol serves to filter out irrelevant interpretations in order to obtain the final CG. To the best of our knowledge, such an automatic generation of

semantic representations from texts has never been reported before and it allows us now to achieve a simple formulation of the retrieval task. When the user asks a question in natural language, the system translates it into a CG and then searches in the document database for CGs that are relevant to the query. Once such graphs have been found, we can use them to access the information and compute the answers.

A *ranking function*[2] then comes into operation to classify the documents recovered according to their relevance with respect to the query. The goal is to avoid the user wasting time sifting through hit lists of search results full of irrelevant documents, especially when we know that information seekers rarely review beyond the first page of the result set [61], which constitutes one of the main causes of dissatisfaction with IR systems [47], and can even detract from the real capacity of the search engine [60]. To solve this problem, we have found our inspiration in previous works, characterizing a ranking function as a partial order relation on the set of transformations applied to the query to satisfy its containment in the document database [56]. The idea consists of assigning different weights to these transformations according to their structural nature, which allows us to focus on search criteria that are free from personal preference, ruling out supervised and learning-based approaches due to their high cost in human terms.

However, a primary concern in the IR field is evaluation. In this sense, we define a formal testing frame that includes the consideration of different ranking techniques for IR systems of both types with and without using *relevance judgments*, often stored in a file called *query relevance* (QREL), and the selection of a topic set representative of the information needs. More in detail, for the ranking task our starting point is the classic TREC[3] protocol based on QREL [156], on which we also consider a simple variation using pseudo-QREL (PQREL), as proposed by Soboroff *et al.* [135]. As a further alternative, also incorporating QREL and/or PQREL, but considering a different ranking criterion, we adopt a technique inspired in the notion of *system authority* described by Mizzaro *et al.* [101]. In dealing with ranking techniques without considering QREL, we chose to evaluate our proposal using a method introduced by Wu *et al.* [164], one of the most popular of its kind and based on the idea of estimating the effectiveness of a search engine by comparing its results with those provided by a set of IR systems that act as a reference.

In dealing with the choice of topic set, we combine a series of previous works organized around two complementary questions. The first relates to individual topic selection for an individual IR system and applies the concept of *average precision* (AP) [11]. The next concerns the topic set selection for an individual system [62]. From these techniques, and in the absence of any specific and definitive solution in the state-of-the-art, we propose a method to reasonably provide topic set selection for a set of systems, inspired in both human assessment and the notion of *topic hubness* proposed by Mizzaro *et al.* [101].

[2] also called retrieval function by Fuhr and Buckley [53].

[3] from *Text REtrieval Conference*, an annual benchmarking exercise on IR.

The remainder of the paper is organized as follows. Section 2 gives an overview of the state-of-the-art in conceptual IR, focusing on indexing and evaluation tasks. In Section 3, we introduce the running *corpus* that will serve as a guideline for the discussion. Section 4 is a reminder of the basic theory of CGs and their applications to IR. A detailed description of our proposal for knowledge acquisition constitutes Section 5. Section 6 introduces our formal testing frame, focusing on topic selection and the ranking of IR systems without QREL. An exhaustive series of experimental tests are then discussed in Section 7, while Section 8 closes the paper with the conclusions of the study.

2 | The State-of-the-Art

The incorporation of NLP techniques for IR has always fascinated researchers, with a double intention, namely the integration of text interpretation techniques to identify index terms, and the characterization of their internal structure. In this regard, the state-of-the-art places us within a generic working framework that is referred to in different ways. Thus, some authors talk of *linguistically-motivated indexing* [85, 104], whilst others consider the term *semantic indexing* to be more appropriate [84, 133]. Other works even resort to the expressions *intelligent retrieval* [38, 57, 134, 146] to emphasize the interaction of the human mind and artificial intelligence with networks and technology.

2.1 | Semantic indexing

Often, indexing structures are single words, but they can also be multi-word units. So, linguistic analysis may be applied at sub-word level, for morphological decomposition and stemming [59, 71, 81], but also at word level since conflation of content words via lemmatization or by means of morphological families fails to take syntactic information into account. Nevertheless, indexes of this kind are able to deal with a number of complex linguistic phenomena such as clitic pronouns, contractions, idioms, and proper name recognition [2].

However, our main interest focuses on the use of meaningful string kernels and sentences as terms in the automatic categorization of documents, a longstanding idea that should mark an improvement over the use of single words, yet there is little practical evidence that this is the case. In fact, for a long time the widely held conviction [134, 75] was that only shallow linguistic techniques can be of interest when developing applications of this kind [134], although their positive effect on accuracy is small at best [85]. Formally, the defining characteristic of this kind of techniques is the extraction from text of relations of linguistic dependency between terms, their formal representation and the subsequent definition of an information-locating mechanism based on the latter.

We can distinguish [85, 171] two levels of complexity when dealing with

dependencies in texts. The lowest level is a lexical-oriented one, linguistically less sophisticated and represented by a group of techniques known as *dependency modeling*. Usually, these systems take into account dependencies between certain term pairs or triples [125], often associated to a probabilistic model [26, 90, 95, 136] in order to rank the most plausible relations. In this sense, most techniques for extracting multi-word terms rely on statistics [46] or simple pattern-matching [78, 132], instead of considering the structural relations among the words that form a sentence. More recently, some authors have proposed using *shallow parsing* techniques to extract these pairs [2] and triples [85] of words related by some kind of syntactic dependency. All these works report an improvement on the unigram independence language model[4], in particular when we are dealing with language that has a rich lexis and morphology. However, the main problem lies in the difficulty of integrating term proximity into this framework. The parameter space can become very large by direct dependency consideration, making the strategy sensitive to data sparseness and noise, which may counteract the relatively small benefits one could obtain and justifies interest in proximity language models [171].

The highest level in dealing with dependencies tries to incorporate bigger units than the word, such as the sentence, in text representation. In such a way, dependence between words can be captured indirectly. As in the previous case, techniques for extracting sentences rely on statistics [34, 54], but also on pattern matching [116] and deep parsing techniques [50, 143]. However, although a detailed and accurate semantic analysis is not required for IR tasks [134], with the explosive growth of information it is becoming increasingly difficult to retrieve the relevant documents by statistical means alone [146]. The exceedingly large number of terms that could possibly be of interest in the description of a document collection lies at the root of the problem, related with the well-known difficulty of dealing with data sparseness in this context. In this sense, labeled graph-based text representations seem to be capable to detect unapparent links between concepts across documents [73, 97, 133], regardless of the size of the *corpus* considered. The approach is not only a promising one, but also has the potential to outperform the standard BOW model in response to long queries [94], an idea about which there is an extremely broad consensus [36], with various strategies having been put forward. So, although until recently the most popular of such approaches was the *semantic networks* [91] one, probably none has been as extensive and comprehensive in recent times as CGs [137]. In fact, CGs are an extension of the previous ones by introducing the notion of dependency between nodes, and enjoy three principal advantages as a formal description method. First, they can support direct mapping onto a relational data base as defined by Codd [32]. Second, they can be used as a semantic basis for natural language. Finally, based on graph transformations, they can support automatic inferences to compute relationships that are not explicitly mentioned [57].

[4] in practical retrieval environments the assumption is normally made that the terms assigned to the documents of a collection occur independently of each other [125]. The term independence assumption is unrealistic in many cases, but its use leads to a simple retrieval algorithm.

This apparent versatility of the graph-based model may also be able to provide a solution to the search for incompletely represented documents, even from vague queries. This phenomenon, which has long justified interest in and the use of strategies based on probabilistic logic, is now growing exponentially as a result of the impossibility of integrating the total amount of information available in practical IR tasks. It is a question of formalizing the implementation of *van Rijsbergen's logical uncertainty principle* [151], according to which relevance is a matter of degree and the central problem in IR is how to model and measure it. As a consequence, to assume that retrieval can be performed by exact matching or through classical logic is to miss the point [57], a mismatch which has provided fertile ground for the widespread feeling that the enhancement of using sentences as indexes appears to lack coherence[5] [55].

In this context, some authors take a middle ground by investigating techniques that make use of limited semantic knowledge that can be easily extracted from the text in the form of a CG representation [138]. This makes it possible to express meaning in a form that is logically precise, humanly readable, and computationally tractable. With their direct mapping to language, CGs serve as an intermediate language for translating computer-oriented formalisms to and from natural languages. And with their graphic representation, they serve as a readable but formal design and specification language. This justifies the fact that the notion of conceptual querying dates from the early days of IR research [139], as well as the research effort expended in recent years to finding a way to replace classic probabilistic notions with formal graph transformations [57], or simply to complement them [134, 146].

2.2 | Ranking strategy

Traditionally, document relevance has been estimated using a variety of similarity-based ranking functions that, in practice, are simple strategies to tune the weights associated to indexing terms by a search engine in order to optimize its performance[6]. More recently other models exploiting the close correlation between popularity and relevance have gained notoriety, mainly when dealing with IR systems managing large amounts of both data and user accesses, as is typically the case of Internet-related search engines [82, 109]. However, although the algorithms have nowadays become quite sophisticated in the evaluation of document popularity, a specific effort is required to avoid certain problems inherent to this technique. We can here refer to the treatment of newly incorporated contents with few accesses [9, 18, 23, 41, 45, 86, 105], the fact that the most popular documents tend to become even more popular [8, 27, 28, 58] or the elimination of possible manipulation of the rankings through artificially inflated link popularity [4, 6, 22, 76, 96, 108, 144].

[5] in some cases, an improvement in effectiveness is achieved while in others, marginal or negative results are obtained.

[6] thus, some authors refer indifferently to term weighting strategy or ranking function [47].

In spite of the above, neither similarity-based nor popularity-based ranking models alone seem to be effective enough to support general or even domain-specific IR [168], which justifies the consideration of hybrid proposals that have been so widely applied [43, 67, 109], even when a good similarity-based ranking kernel seems to be the determining starting point for obtaining retrieval efficiency. With regard to this, an alternative for improving performance consists in directly measuring conceptual similarity, which can be estimated in different ways. Thus, some works derive it from the information content of their *least common subsumer* (LCS), which seems to be closer to the implicit ranking functions exercised by humans [115]. The original idea is due to Cohen *et al.* [33], who describe a method to compute the LCS for a pair of concepts, thus enabling us to relate them through the most specific description subsuming the respective structures. In this way, we can inference sub-concept/super-concept relationships (resp. whether a given individual belongs to a certain concept), providing us with a tool to make explicit commonalities and to derive implicit knowledge using subsumption (resp. instance) oriented techniques [88]. The state-of-the-art adopts and continues this study with the aim of using information content to evaluate semantic similarity in taxonomies [115], which has subsequently served as inspiration for different ways of dealing with computation tasks in the context of IR technology. This is the case of some authors [102] who take direct advantage of this technique to extend classic measures for text comparison such as the Dice coefficient [40]. Techniques other than LCS-based proposals have been considered, including alternative extensions to the Dice measure [103], as well as generalization relationships associated to a specific knowledge domain [119]. In any case, such proposals first need to dispose of a knowledge-based ontological structure to represent those concepts as well as *corpus*-based statistical technology to generate and manage them, thus placing us once more within the context of conceptual IR [139].

From an operational point of view, whatever the criterion of relevance considered may be, a ranking function can be classified according to three complementary views related to its generation phase: its adaptability to the context, its supervised nature and its consideration as a learning-based model [92]. With regard to the first of these, most IR systems use a fixed strategy to support the ranking task, regardless of the heterogeneity of users, queries and collections [47] defining their *working context*. This is commonly known as a consensus search, in which the computed relevancy for the entire population is presumed appropriate for individuals and, therefore, everybody gets the same results. Although we could interpret this uniformity as an advantage because it allows for the comparison of search results among different users, it is also true to say that the idea of suiting the features of the retrieval process to our own preferences is always appealing. We would then be talking about *personalized search* [110], an approach that does not seem to work consistently well across different contexts [126, 173].

On the other hand, traditional IR is mostly based on unsupervised ranking facilities, often based on the degree of matching between query and document. This is the case of the Boolean [150], vector space [126], probabilistic [117], and language modeling-based

methods [111]. Theoretically simple and intuitive, they work reasonably well and do not require labeled data, an advantage that does not exclude the possibility of associating a number of tuning parameters by using some kind of training technique, which is not uncommon. However, as ranking models become more sophisticated, parameter tuning becomes an increasingly challenging issue [167] and, in practice, these empirical approaches only have a few parameters to tune [7].

In contrast to unsupervised approaches, supervised ones enjoy higher accuracy and better adaptability at the price of requiring greater human effort, which for a long time limited the practical interest of this kind of strategies. However, the current availability of sets of relevance assessments collected by groups of expert annotators provides a means of incorporating machine learning techniques into the design of ranking models. The idea consists of using these labeled resources as training facilities for estimating the semantic proximity between queries and documents [168] through the minimization of a *loss function* loosely related to some particular IR performance measure, often the *mean average precision* (MAP) or the *normalized discounted cumulative gain* (NDCG), although there are also some proposals that can optimize any such measure used in IR [166]. In this regard, a variety of learning strategies have been proposed, such as neural networks [14, 17], support vector machines [16, 68, 69, 74, 152, 169], boosting [52, 93, 166] or genetic programming [37, 39, 48, 148]. In practice, although these methods seem to work better than the unsupervised traditional ones [92, 167], some major differences can be observed, depending on the type of instances used in learning. More in detail, three different instantiation models have been addressed: pointwise, pairwise, and listwise.

In the pointwise approach [93, 107] each query-document pair in the training data associates a score independently of the other ones, which implies that no relative preferences between two documents retrieved for the same query are considered. Consequently, this method has shown poor performance, transforming the learning-to-rank problem into a regression or classification one of a single document [87]. Pairwise-based proposals [14, 16, 52, 69, 74, 87, 149, 168, 170] seem to be the most popular of the learning-to-rank kind. In this method, document pairs retrieved for a given query in which the most relevant of the two has been determined constitute the instances of the training data set. So, the goal of the learning process is to minimize the average number of inversions in ranking in order to obtain a binary classifier which can tell which document is better in a given pair. This implies that, given a query, we must induce a total ordering for a set of documents retrieved from partial orderings between pairs of them, which severely limits the practical possibilities of this approach [10]. Finally, the listwise model [10, 15, 17, 89, 112, 165, 166, 169] has become increasingly popular in recent years. It considers the entire set of documents retrieved for a query as instances in the training phase. This should make it possible to overcome the problems previously mentioned in relation to the pointwise and pairwise techniques, and, in fact, practical results suggest its superiority over the latter. However, the definition of a listwise

loss function can become a complex task because most IR evaluation measures are not continuous magnitudes with respect to the ranking model's parameters.

Ultimately, there is a wide spectrum of basic ranking techniques currently available. Each has its own set of advantages that we should try to reconcile through hybrid proposals, and drawbacks that we would want to avoid or at least minimize. In this respect, probably the optimum factor combination depends on the nature of the particular search task we want to deal with; in our case, a domain-specific one. This places us directly in the context of some recent works [168] that claim benefits derived from the use of concept-based similarity search, eventually incorporating a dimension of popularity when the working environment can guarantee a sufficient number of accesses.

2.3 | Retrieval evaluation

In this sense, the QREL-based techniques popularized by the TREC events [156, 157] have been considered as a standard in IR evaluation. These workshops address this problem by pooling the top 100 documents retrieved by each participating system, after which an evaluator determines the relevance of each document in the pool. Inspired by *Cranfield's methodology* [31, 30], it compares IR systems over a set of topics, a set of documents for each topic, and a set of QREL for each document. Baptized as *depth pooling*, for over twenty years this large-scale experimentation has been the reference in the field, but the increasing size, complexity and heterogeneity of document collections and query sets has itself made this exercise infeasible.

A number of alternative approaches have been proposed for estimating the performance of IR systems with limited recourse to QREL, in order to reduce the human effort associated with this task. These methods operate on the basis of selecting the best set of documents to evaluate and considering quality measures when few judgments are available. In this category, the first attempts were the well-known pooling techniques [140], which focus on those texts least likely to be non-relevant. However, recent work suggests that the growth in the size of *corpora* is outpacing even the ability of pooling to find and judge sufficient documents [13], since with fewer documents available the variance of the estimates of evaluation measures will be higher. In this sense, some authors [21] try to reduce judging effort whilst maintaining a large number of topics, although they recognize that failure analysis then becomes more difficult and is in need of further exploration.

A second alternative reduces the human assessment load involved in the generation of QREL by introducing the notion of pseudo-QREL, which are generated either by randomly selecting mapping of documents to topics [135], or by skimming the highest-ranked documents returned for an array of topic representations [44].

For their part, Mizzaro *et al.* [101] propose a method of analyzing data gathered from QREL-based or similar IR evaluation resources, such as PQREL. By introducing two

normalized versions for AP that the authors use to construct a weighted bipartite graph of search engines and topics, they find that authority measures systems performance and that hubness reveals topic ease.

Finally, some proposals [164] disregard the QREL concept, using document overlapping instead. Succinctly, this technique involves interpreting the relationship among the documents retrieved by a pool of IR systems, where the overlap structure appears to have a strong impact on the results. Thus, it is often argued [142] that methods of this kind can produce poor results for the best performing systems when they are ranked together with the poorly performing ones, and they seem to perform worse than the previous groups of techniques.

Another point to consider when defining a formal testing frame for IR systems is the choice of an adequate set of topics, i.e., determining those that are best at predicting true effectiveness. Little research has been devoted to this aspect, and the practical results are essentially limited to a series of insights relating to working hypotheses and to the proposal of selection strategies which still require serious experimentation in order to validate them. In the sphere of confirmed hypotheses, Mizzaro [100, 101] proves formally that some topics are easier than others and that there are differences between systems when it comes to distinguishing between easy and difficult ones. However, although we can say that not all topics are equally informative about IR systems, no evidence has emerged as to which criterion might be best to qualify this statement.

These research works on retrieval evaluation also strongly suggest that individual topics vary greatly in their ability to discriminate between systems, which extends our attention to topic set construction. In this regard, it is not merely a question of discerning when one topic set is more useful for evaluation purposes than another, but also of selecting the smallest possible topic set whilst still maintaining this quality. This makes it possible to reduce the workload in a methodology whose major issue is its expense, and thus justifies the practical interest of this kind of strategies. However, although there has been concern about this question for many years, no relevant contributions had been put forward until recently [62]. Previous works do not draw upon any idea to this purpose, concentrating exclusively on what the pool depth should be and using some kind of heuristic approach as a methodological basis [11, 130, 140, 160, 162, 172], the result, unfortunately, being different in each case. In this regard, although the proposal of Guiver *et al.* in [62] does not immediately provide a complete solution to the problem of identifying suitable topic sets, it formally proves the existence of complementarity phenomena between individual sets and their impact on evaluation quality, disproving the hypothesis that it is a random effect. The method is based on the notion of MAP [63] and, more in detail, it applies an exhaustive search on all the possible subsets of topics in an interval of cardinality. For each subset, the corresponding MAP is computed and the corresponding correlation with the MAP on all topics is calculated. Authors argue that the highest (resp. the lowest) correlation values correspond to the best (resp. the worst) subsets. However, a major drawback for the direct application of this method is

the complex combinatorial analysis required, involving a large set of evaluated topics and associated system runs. In such a way, the gain from such condensation may be relatively small for a significant effort and some kind of heuristic strategy needs to be envisaged in order to avoid complete searches in this space.

3 | The running corpus

In order to favor understanding, we introduce our proposal from a botanic *corpus* describing West African flora. We concentrate on the work *«Flore du Cameroun»*, published between 1963 and 2001, which consists of about forty volumes in French, each volume running to about three hundred pages, and organized as a sequence of sections, each one dedicated to a single species and following a systematic structural schema. So, sections include a descriptive part enumerating morphological aspects such as color, texture, size or form. This implies the presence of nominal sentences, adjectives and also adverbs to express frequency and intensity, and named entities to denote dimensions.

The text is organized taxonomically, introducing species (resp. genera) in separate chapters (resp. sections), this being equivalent to hypernymy or «`is_a`» relations. However, the descriptions include concepts that are related non-taxonomically. We can here distinguish between labeled relations[7], which can be retrieved through nominal sentences expressing the former in an assertive way[8] and can be propagated by more complex structures requiring the consideration of sophisticated NLP techniques in order to recognize them, as is the case of enumerations and interval definitions[9]. The collection also possesses a vocabulary that is shared by most texts based on this matter, and is of sufficient size for our purposes. This will allow us to assess our proposal on a variety of verbal and nominal forms for which the correct semantics is not trivial. In particular, due to the diversity of the linguistic constructions present in the *corpus* and the different ways in which they are expressed, it seems a suitable testing platform to deal with ambiguity phenomena and grammatical completeness.

The *corpus*[10] was previously transferred from textual to electronic format [123] and the logical structure of the text captured in order to browse it, within BIOTIM [118], a research initiative on the integral management of botanic documents including conceptual acquisition and text mining tasks. Henceforth, we shall denote this running *corpus* by $\mathscr{B}$.

[7] typically related to properties such as `color`, `form`, `size`, `texture` or `position`; or to entities such as `organ` or `fruit`.

[8] e.g., the case of relations of the type «`in form of`» or «`of color`».

[9] i.e., the case of constructions of the type «`from X to Y`» or «`X and Y`».

[10] supplied by the *French Institute of Research for Cooperative Development*.

4 | Conceptual graphs and searchable bases

Following [25], the kernel of our approach in IR is the notion of *basic conceptual graph* (BG), which is in fact a CG without negation that describes entities and the relationships between them, and also introduces reasoning on the basis of a graph morphism called *projection*. In fact, it can be shown that projection checking is essentially the same problem as those of constraint satisfaction or conjunctive query containment in databases [83] and, in particular, projection proves to be both sound and complete with regard to deduction in *first-order-logic* (FOL). All in all, this highlights a fundamental question in IR, namely query answering, which asks for all answers to a query. Most of the contents in the remainder of this section are taken and/or inspired from Chein *et al.* [25] and Genest *et al.* [57].

4.1 | Basic conceptual graphs

The first step for the implementation of the query strategy is to define a working framework that will make it possible to create a cognitive map of the basic ontological knowledge we are working with. We call this structure *support* and it compiles the main concepts, relations and vocabulary that exist in the world we are trying to describe.

Definition 1 *A* support *is a triple* $\mathcal{S} = (\mathcal{T}_\mathcal{C}, \mathcal{T}_\mathcal{R}, \mathcal{I})$ *of finite sets pairwise disjoint, such that:*

- $\mathcal{T}_\mathcal{C}$ *(resp.* $\mathcal{T}_\mathcal{R}$*) is a partially ordered set of* concept *(resp.* relation*)* types*. These orders are interpreted as specialization relations. So,* $t \leq r$ *is read as* r is a generalization of t *or, also, as* t is subsumed by r*.*
- *Types in* $\mathcal{T}_\mathcal{C}$ *posses a greater element,* $\top$*, called* universal type*. Types in* $\mathcal{T}_\mathcal{R}$ *may be of any arity greater or equal to 1, and only those with same arity are comparable.*
- *The countable set* $\mathcal{I}$ *is a collection of* individual markers*with a* generic marker $* \notin \mathcal{I}$*. The set* $\mathcal{I} \cup \{*\}$ *is partially ordered and its elements pairwise non-comparable, being* $*$ *the greatest one.*

■

In essence, a support consists of a concept type hierarchy, a relation type hierarchy and a set of markers that we can identify with a dictionary whose elements will be later associated to concept types. In practice, this dictionary represents the lexical forms in the *corpus* we are working in, while concepts refer to their semantic categories and relations the relationships between them.

The concepts and relations compiled in this frame can now be linked together in order to describe the facts we are interested in. In this sense, a BG represents the stencil which is going to be filled in with the concept/relations taken from the support.

Definition 2 *A* basic conceptual graph (BG) *defined over a support* $\mathcal{S} = (\mathcal{T}_\mathcal{C}, \mathcal{T}_\mathcal{R}, \mathcal{I})$ *is a 4-tuple* $\mathcal{G} = (\mathcal{C}, \mathcal{R}, \mathcal{E}, \mathcal{L})$*, where:*

- $(\mathcal{C} \cup \mathcal{R}, \mathcal{E})$ *is a bipartite multigraph*[11] *with* $\mathcal{C}$ *and* $\mathcal{R}$ *disjoint sets of* concept *and* relation nodes*, respectively.*
- $\mathcal{E}$ *is the multiset of* edges*.*
- $\mathcal{L}$ *is a labeling function for nodes and edges:*
  - *A node* $c \in \mathcal{C}$ *is labeled by a pair* $[\text{type(c)}, \text{marker(c)}] \in \mathcal{T}_\mathcal{C} \times (\mathcal{I} \cup \{*\})$*. A node* $r \in \mathcal{R}$ *is labeled by* $\text{type(r)} \in \mathcal{T}_\mathcal{R}$ *and the degree*[12] *of* r *must be equal to the arity of* type(r)*.*
  - *An edge in* $\mathcal{E}$*, labeled by* $i \in \mathbb{N}$*, connecting nodes* $r \in \mathcal{R}$ *and* $c \in \mathcal{C}$*, is denoted by* (r, i, c)*. The edges* $(r, 1, c_1), \ldots, (r, k, c_k)$ *incident*[13] *to* $r \in \mathcal{R}$ *are totally ordered and labeled from* 1 *to the degree* k *of* r*. We then shortly denote* $r = \text{type(r)}(c_1, \ldots, c_k)$*.*

■

Intuitively, a BG can be viewed as a bipartite graph that provides an ontology of the application domain. The concepts refer to the markers in the support, to which we have now associated a conceptual type. Briefly, we have a road map that reflects the organization of this domain as a declarative memory system that facilitates both sense-making and meaningful learning. At this point, having formalized the structure we are going to use to represent knowledge, we can now introduce *projection* as the basic mechanism that will make it possible to capture the notion of query answering.

Definition 3 *Let* $\mathcal{G}_1 = (\mathcal{C}_1, \mathcal{R}_1, \mathcal{E}_1, \mathcal{L}_1)$ *and* $\mathcal{G}_2 = (\mathcal{C}_2, \mathcal{R}_2, \mathcal{E}_2, \mathcal{L}_2)$ *be two* BGs *defined on a support* $\mathcal{S} = (\mathcal{T}_\mathcal{C}, \mathcal{T}_\mathcal{R}, \mathcal{I})$*, then a* projection *from* $\mathcal{G}_1$ *to* $\mathcal{G}_2$ *is a mapping* π *from* $\mathcal{C}_1$ *to* $\mathcal{C}_2$*, and from* $\mathcal{R}_1$ *to* $\mathcal{R}_2$ *verifying:*

$$(r, i, c) \in \mathcal{E}_1 \Rightarrow (\pi(r), i, \pi(c)) \in \mathcal{E}_2 \quad \text{and} \quad x \in \mathcal{C}_1 \cup \mathcal{R}_1 \Rightarrow \mathcal{L}_2(\pi(x)) \leq \mathcal{L}_1(x)$$

where, if $x \in \mathcal{C}_1$*,* $\leq$ *refers to the cartesian product of the order on* $\mathcal{T}_\mathcal{C}$ *and on* $\mathcal{I} \cup \{*\}$[14]*. If* $x \in \mathcal{R}_1$*, then* $\leq$ *refers to the order on* $\mathcal{T}_\mathcal{R}$*.*

[11] i.e., there may be several edges between two nodes.
[12] i.e., the number of edges incident to.
[13] where $c_1, \ldots, c_k \in \mathcal{C}$ are not necessarily disjoints.
[14] i.e., $(\mathit{type}(\pi(x)), \mathit{marker}(\pi(x))) \leq (\mathit{type}(x), \mathit{marker}(x))$ iff $\mathit{type}(\pi(x)) \leq \mathit{type}(x)$, and $\mathit{marker}(\pi(x)) \leq \mathit{marker}(x)$.

We call $\mathcal{G}_1$ *the* source *and* $\mathcal{G}_2$ *the* target*, and we say that* $\mathcal{G}_1$ subsumes $\mathcal{G}_2$ *or that* $\mathcal{G}_1$ *is* more general than $\mathcal{G}_2$*, using the notation* $\mathcal{G}_1 \succeq \mathcal{G}_2$*. The set of projections from* $\mathcal{G}_1$ *to* $\mathcal{G}_2$ *is denoted by* $\text{proj}(\mathcal{G}_1, \mathcal{G}_2)$.

■

Intuitively, a *projection* is a *graph homomorphism*[15] that can specialize the labels of concept and relation nodes. So, the existence of a projection from a BG $\mathcal{Q}$ to another one $\mathcal{D}$ means that the knowledge represented by $\mathcal{Q}$ is contained in the knowledge represented by $\mathcal{D}$.

Theorem 1 *Let* $\mathcal{G}_1 = (\mathcal{C}_1, \mathcal{R}_1, \mathcal{E}_1, \mathcal{L}_1)$ *and* $\mathcal{G}_2 = (\mathcal{C}_2, \mathcal{R}_2, \mathcal{E}_2, \mathcal{L}_2)$ *be* BGs *defined on a support* $\mathcal{S}$*, then* $\mathcal{G}_1 \succeq \mathcal{G}_2$ *iff* $\exists\ \pi$*, a projection from* $\mathcal{G}_1$ *to* $\mathcal{G}_2$.

Proof: Trivial from definition 3.

■

4.2 | The query answering problem

We are now ready to rewrite the query answering problem, that here takes as input a knowledge base (KB) $\mathcal{D}$ composed of BGs representing facts and a BG c, which represents a query in a collection $\mathcal{Q}$, and asks for all answers to $c \in \mathcal{Q}$ in $\mathcal{D}$. So, each projection from c to a fact leads to an answer or, as we shall see, c is deducible from the KB $\mathcal{D}$. To attain this goal, we consider a semantic mapping Φ that assign a FOL formula $\Phi(\mathcal{G})$ to each BG $\mathcal{G}$ [139] defined over a support $\mathcal{S} = (\mathcal{T}_\mathcal{C}, \mathcal{T}_\mathcal{R}, \mathcal{I})$, where $\Phi(\mathcal{G})$ is a positive, conjunctive and existentially closed formula. Φ also assigns a set of formulas $\Phi(\mathcal{S})$ to the support $\mathcal{S}$, which corresponds to the interpretation of the partial orderings of $\mathcal{T}_\mathcal{R}$ and $\mathcal{T}_\mathcal{C}$. For all types t and t', such that $t \geq t'$, one has the formula:

$$\forall C_1, \ldots, C_k,\ t'(C_1, \ldots, C_k) \rightarrow t(C_1, \ldots, C_k)$$

where $k = 1$ for concept types, and k is otherwise the arity of the relation. This implies that queries and databases can be interpreted as logical formula, and the search process corresponds to logical inference.

Theorem 2 *(Soundness and completeness) Let* $d \in \mathcal{D}$ *and* $c \in \mathcal{Q}$ *be two* BGs *defined on a support* $\mathcal{S}$*, then:*

$$c \succeq \text{nf}(d) \Leftrightarrow \Phi(\mathcal{S}),\ \Phi(d) \models \Phi(c)$$

[15] i.e., a morphism that preserves edges.

where $\models$ *denotes deduction in* FOL*; and* $\mathrm{nf}(d)$ *is the* normal form *of* d*, i.e., that obtained by merging concept nodes having the same individual marker*[16].

Proof: See [106].

■

It can be proved that the query answering problem for BGs is NP-*complete* [24]. In this sense, any given solution to the decision problem can be verified in polynomial time [25, 70], giving computational meaning to our approach.

4.2.1 | Types of answers

From a practical standpoint we also need to give projections the flexibility required to locate answers whose structure does not correspond exactly with the projection of the corresponding query. In this regard, it will be necessary to organize the search for sequences of *transformations* that will enable the query or the document base to relax its structures in order to make such a projection possible.

Definition 4 *Let* $d, d' \in \mathcal{D}$ *and* $c \in \mathcal{Q}$ *be three* BGs *defined on a support* $\mathcal{S}$*, and* ς *a mapping from the set of* BGs *defined on* $\mathcal{S}$ *onto itself, such that* $\varsigma(d) = d'$*. If* $\pi \in \mathrm{proj}(c, d')$*, then* (π, ς) *is a* projection from c to d modulo ς.

■

Intuitively, the idea is to supply a set of transformations that make it possible to determine the relevance of a document to a query, when there is some kind of relation between the information contained in the two. Formally, we will consider three transformation mechanisms that are applicable to a BG.

Definition 5 *Let* $\mathcal{G} = (\mathcal{C}, \mathcal{R}, \mathcal{E}, \mathcal{L})$ *be a* BG *defined on a support* $\mathcal{S} = (\mathcal{T_C}, \mathcal{T_R}, \mathcal{I})$*, a* substitution on $\mathcal{G}$ *is a pair* $(t, t') \in (\mathcal{C} \times (\mathcal{T_C} \times (\mathcal{I} \cup \{*\}))) \cup (\mathcal{R} \times \mathcal{T_R})$*. When we assert that the concept (resp. relation) term* t *can replace the term* t'*, we say that* (t, t') *are* compatible terms.

■

Definition 6 *Let* $\mathcal{G} = (\mathcal{C}, \mathcal{R}, \mathcal{E}, \mathcal{L})$ *be a* BG *defined on a support* $\mathcal{S} = (\mathcal{T_C}, \mathcal{T_R}, \mathcal{I})$*, the result of the* join of $c, c' \in \mathcal{T_C}$*, such that* $\mathcal{L}(c) = \mathcal{L}(c')$*, is the* BG *obtained from* $\mathcal{G}$ *by identification of* c *and* c'*.*

■

[16] i.e., a BG is in normal form if each individual marker appears at most once in it.

As a join can substantially change the structure of a BG, this transformation is usually considered more distancing than substitutions.

Definition 7 *Let $\mathcal{G} = (\mathcal{C}, \mathcal{R}, \mathcal{E}, \mathcal{L})$ be a* BG *defined on a support $\mathcal{S} = (\mathcal{T}_\mathcal{C}, \mathcal{T}_\mathcal{R}, \mathcal{I})$, the result of* adding a node $n \in \mathcal{C} \cup \mathcal{R}$, *such that $\mathcal{L}(n) = v$, is the new* BG $\mathcal{G} + \mathcal{N}$, *where $\mathcal{N}$ is the graph reduced to n. If $n \in \mathcal{R}$, neighbors must be specified.*

■

Since an addition not only modifies the structure of the original BG, but also introduces an element external to the latter, this transformation is taken to be more complex than a join, and thus also has more impact than a substitution. Furthermore, and according to whether or not it is necessary to combine the transformations defined, we will consider four possible kinds of answer to a given query, which we will introduce incrementally according to in order of complexity of calculation. Thus, the simplest answers will be those whose content refers exactly to the query made.

Definition 8 *Let $d \in \mathcal{D}$ and $c \in \mathcal{Q}$ be two* BG*s defined on a support $\mathcal{S}$. Then d is an* exact answer to c *iff* $\text{proj}(c, d) \neq \emptyset$.

■

The absence of an exact answer is often predictable, either as a result of the lack of specific information in the document database, or of the lack of specificity in the query itself. In the first case, we talk about *document incompleteness* and in the second about *query vagueness*. In order to deal with this situation, we first need to formally capture the notion of non-exact answer and place it within the framework we have already defined for BGs. In this regard, we adopt the search strategy proposed in en [57], which in turn takes its inspiration from the implementation of the *second form of van Rijsbergen's Uncertainty Principle* [151] proposed in [79]:

"Let $d \in \mathcal{D}$ and $c \in \mathcal{Q}$ be two propositions, a measure of the uncertainty of $d \rightarrow c$ relative to a knowledge base is determined by the minimal transformation from d to d', such that $d' \rightarrow c$ holds."

where, in our case, the transformations from d to d' are based on graph operations, which could also be used to transform a query c. Thus, we might also wonder why not transform c to c' in order to obtain that $d \rightarrow c'$ holds. With respect to this, it can be shown that $d' \rightarrow c$ holds if and only if $d \rightarrow c'$ holds, where c' is obtained from c by a dual transformation of a transformation from d to d'. The advantage commonly put forward [57] for modifying database $\mathcal{D}$ instead of queries in $\mathcal{Q}$ is that the content of the former can be enriched through relevance feedback by the IR system itself. Regardless, this makes it possible to

establish the formal framework we need to make the query protocol previous introduced for BGs more flexible. We will begin by describing the simplest case.

Definition 9 *Let $d \in \mathcal{D}$ and $c \in \mathcal{Q}$ be two* BG*s defined on a support $\mathcal{S}$. Then d is an* approximate answer to *c iff there exists a sequence of substitutions ς, such that* $\text{proj}(c, \varsigma(d)) \neq \emptyset$.

■

Intuitively, to compute an approximate answer, the structure of the initial BG d is only slightly modified. Given that exact answers are a particular case of the approximate ones and that they constitute a rare phenomenon of no practical interest, we will henceforth use the term approximate answers to refer both categories, exact and approximate. In order to further increase the degree of flexibility associated to querying, we are obliged to expand the threshold of admissible structural transformations, for example by including joins.

Definition 10 *Let $d \in \mathcal{D}$ be a* BG *defined on a support $\mathcal{S}$. We say that a* sequence ς of substitutions and joins is acceptable *iff ς does not contain too many joins relative to the number of nodes in the* BG *$c \in \mathcal{Q}$. The ratio number (μ_j) of joins can be chosen by the user.*

■

Definition 11 *Let $d \in \mathcal{D}$ and $c \in \mathcal{Q}$ be two* BG*s defined on a support $\mathcal{S}$. We say that d is a* plausible answer *to c iff there is an acceptable sequence ς of substitutions and joins, such that* $\text{proj}(c, \varsigma(d)) \neq \emptyset$.

■

To complete the query offer, we finally include node adds. Although this does not allow us to cover the full range of transformations for graphs, it does focus on those queries whose impact is less than the initial intention expressed by the user when making a query.

Definition 12 *Let $d \in \mathcal{D}$ be a* BG *defined on a support $\mathcal{S}$. We say that a* sequence ς of substitutions, joins and node adds is acceptable *iff ς is acceptable for the joins and there are not too many node adjunctions relative to the number of nodes in the* BG *$c \in \mathcal{Q}$. The ratio number (μ_a) of node added can be chosen by the user.*

■

Definition 13 *Let $\mathcal{D}$ and $\mathcal{Q}$ be two* BG*s defined on a support $\mathcal{S}$. We say that $\mathcal{D}$ is a* partial answer *to $\mathcal{Q}$ iff there is an acceptable sequence ς of substitutions, joins and node adds; such that* $\text{proj}(\mathcal{Q}, \varsigma(\mathcal{D})) \neq \emptyset$.

■

The formal tools we have introduced above define a BG-based working environment that will enable us, on the one hand, to represent the knowledge contained in a collection of documents, and on the other to extract the former using a specific standard model we have endowed with a certain degree of flexibility.

4.3 | The ranking function

Now we have formalized the query answering problem, we need to integrate a ranking strategy as the final step in the design of our concept-based IR architecture. To this end, the use of BGs as indexing terms allows us to naturally place the question in the domain of subsumption and instance-based functions. At this point, even LCS-based proposals have the necessary potential to become a potent ranking facility, although they suffer from a practical lack of efficiency due to their high computational cost. As an alternative, Genest [56] extends the range of conceptual relationships in order to obtain more flexible and less greedy techniques, looking for a compromise between efficiency and discriminative power. For this reason, the author introduces ranking functions as simple partial orders in the set of transformations applied on a query so as to attain a projection on the document database, or in other words, to get an answer.

Definition 14 *Given a support $\mathcal{S}$, let $\mathcal{Q}$, $\mathcal{D} = \{d_i\}_{i\in I}$ be the* BGs *associated to a query and a document database, and let $\mathcal{A}_{\mathcal{Q}}^{\mathcal{D}}$ be the collection of answers obtained through a set $\mathcal{T}_{\mathcal{Q}}^{\mathcal{D}}$ of graph transformation sequences applied on $\mathcal{Q}$ to get a projection on some $d_i,\ i \in I$. We then define a* ranking function associated to $\mathcal{Q}$ and $\mathcal{D}$ *as the ordering naturally induced in $\mathcal{A}_{\mathcal{Q}}^{\mathcal{D}}$ by any partial order on $\mathcal{T}_{\mathcal{Q}}^{\mathcal{D}}$.*

■

This approach generalizes LCS-based ones, while allowing us to relax the computing constraints. In practice, we focus our attention on a particular partial order introduced by Genest in [56].

Definition 15 *Given a support $\mathcal{S}$, let $\mathcal{Q}$, $\mathcal{D} = \{d_i\}_{i\in I}$ be the* BGs *associated to a query and a document database, and let $\mathcal{A}_{\mathcal{Q}}^{\mathcal{D}}$ be the collection of answers obtained through a set $\mathcal{T}_{\mathcal{Q}}^{\mathcal{D}}$ of graph transformation sequences applied on $\mathcal{Q}$ to get a projection on some $d_i,\ i \in I$. We then define the* Genest's partial order on elements $t,\ t' \in \mathcal{T}_{\mathcal{Q}}^{\mathcal{D}}$ *as follows:*

$$t <_G t' \text{ iff } \begin{cases} t' & \text{associates approximate answer } \textbf{OR} \\ t & \text{associates a partial answer } \textbf{OR} \\ t \text{ (resp. } t') & \text{associates a partial (resp. plausible) answer } \textbf{OR} \\ t,\ t' & \text{associate the same type of answer } \textbf{AND} \ \mid t \mid > \mid t' \mid \end{cases}$$

while that

$$t =_G t' \text{ iff } t \textbf{ AND } t' \text{ associate the same type of answer } \textbf{AND } \mid t \mid = \mid t' \mid$$

■

Intuitively this implies that any approximate answer is considered more relevant than a plausible one, and these are more relevant that any partial answer. For answers of the same type, relevance is inversely proportional to the number of individual transformations[17] applied. From a theoretical point of view this is consistent with the previous considerations on structural impact on BGs due to substitutions, joins and adds. In spite of its simplicity, this technique has proved superior to more recent and apparently more sophisticated ones [119], which justifies its formal review and consideration.

5 | Knowledge acquisition

At this point, the implementation of an IR system also requires an effective mechanism for generating the BGs we have adopted as our operating structure. For this purpose we will previously have to devise a protocol for the automatic extraction of the knowledge contained in the texts, and which will then serve as the starting point for the generation process referred to above. In this regard, our proposal contemplates a chain of lexical, syntactic and semantic analysis tools that will lead to the direct automatic generation of BGs from the document database, with minimal user intervention. Our contribution here resides in the novelty of the architectures chosen for the lexical, syntactic and semantic analysis modules, and more particularly in the originality of the design of the semantic analysis environment architecture under consideration.

5.1 | The lexical frame

Although our proposal does not require the use of any specific lexical analysis environment, for our purposes we have chosen a processing chain based on the *Alexina* architecture [120], which is essentially a proposal for digital lexical treatment and acquisition. Given that our running example is a French *corpus*, we consider as our basic resource a large-scale morphological and syntactic lexicon for this language, known as LEFFF [121], which includes information originating from a variety of works. We can here refer to automatic acquisition thanks to the use of statistical techniques on raw *corpora*, the automatic acquisition of specific syntactic information or manual correction and extension guided by automatic techniques.

In order to provide pre-syntactic treatment, we take SXPIPE [122] as a full-featured architecture that produces word lattices from raw text, and is able to handle various phenomena that occur with high frequency in real-life *corpora*. This includes several named-entity families, spelling errors, tokenization ambiguities while detecting sentence and word boundaries, and lexical ambiguities between words differing only by diacritics or capitalization. The lexer is based on a finite state morphology that uses SXPIPE to

[17] i.e., substitutions, joins and node adds.

preprocess the input, and combines its output with lexical information retrieved from LEFFF.

Regardless of the lexical frame considered, its output must include all possible lexical categories for a given occurrence of a form and it will be denoted for description purposes as follows, introducing some additional structural details in order to later integrate semantic data.

Definition 16 *Let* $\{s_i\}_{1 \leq i \leq n}$ *be the sequence of sentences in a* corpus $\mathscr{C}$ *and* $\Theta_{i,j},\ 1 \leq j \leq |\ s_i\ |$ *be the* ocurrence *of a* form *in the i-th sentence,* s_i*. We denote the association of the lexical category* (a) *and semantic class* (b) *to this form, in this sentence, by* $\Theta_{i,j}^{a,b}$ *and we call it* term.

We introduce an anonymous-variable notation, $\Theta_{i,j}^{a,\_}$*, in order to designate the set of terms that can only be differentiated by their semantic class, which we call* token*. In this way, we also denote by* $\Theta_{i,j}^{\_,\_}$ *the set of tokens referring to the same occurrence of a form, which we call* cluster*.*

We also consider a free-variable notation, using capital letters, in order to enumerate a range of values. So, for example, $\Theta_{i,j}^{a,X}$ *refers the sequence of terms in the token* $\Theta_{i,j}^{a,\_}$*, whose semantic class* X *is applicable in that context. We can naturally extend this notation to occurrences of tokens and clusters.*

■

In order to clarify these concepts, Fig. 2.1 illustrates them in relation to the phrase from *corpus* $\mathscr{B}$*, «feuilles à nervures denticulées»* («`leaves with veins dentate`»). Here terms are represented by triangles, tokens by ellipses and clusters by rectangles. The semantic classes associated to terms in this figure are amongst those shown in Table 3.

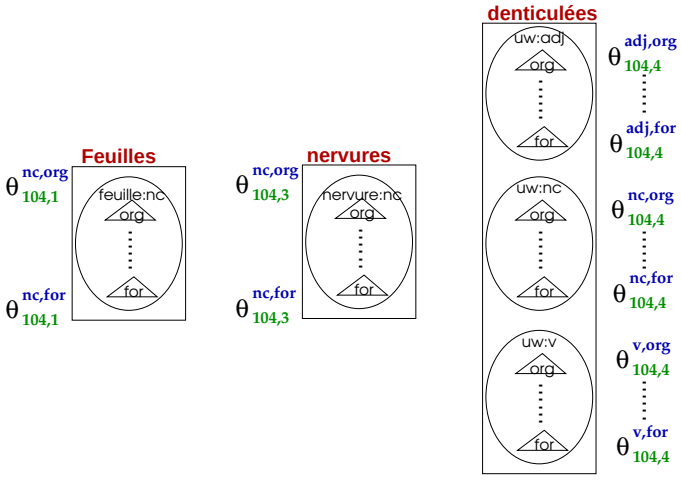

Figure 2.1: Lexical notation

5.2 | The parsing frame

From a descriptive standpoint, our choice fell to *tree adjoining grammars* (TAGs) [77], a mildly context-sensitive formalism that has acquired enormous popularity in the sphere of NLP, for three basic reasons. The first of these is their *extended domain of locality*, enabling them to define syntactic dependencies at any level. The second is the fact that they can consider cross dependencies, whilst the third is that they are a natural extension of the classical context-free model, with the elementary unit of rewriting being the tree rather than the symbol.

5.2.1 | Mildly context-sensitive parsing

From a computational standpoint, the parsing frame chosen [3] has a time (resp. spatial) complexity of the order $\mathcal{O}(n^6)$ (resp. $\mathcal{O}(n^3)$) for a text of length n. As a specific implementation we have opted for*DyALog* [154]. This allows for a high degree of abstraction in the grammatical design through its consideration of the concept of *meta-grammar* [155], introducing a hierarchical design that also includes an inheritance mechanism that simplifies the linguistic task. This enables the grammatical description to be progressively refined, facilitating not only its design but also its maintenance.

With regard to the handling of ambiguities, the use of a dynamic programming tool implies the optimal sharing of calculations and representation structures, leading to a computationally efficient management of non-determinism. We thereby avoid eliminating interpretations in the process of lexical and syntactic analysis, delaying the moment of decision in this regard until the point for semantic analysis is reached, which is when we have available all the information associated with the *corpus* being analyzed. It also allows for the efficient exploitation of the phenomenon of *local determinism*[18], which in practice means that the process is, as far as possible, in fact determinist, and thus possesses linear space and time complexity.

5.2.2 | The parse

he parse has to be summarized in a governor/governed dependency graph that compiles the initial semantic relationships within the text analyzed. Intuitively, in relationships of this kind the nucleus of a syntagma governs its modifiers, as shown in Fig. 2.2 by dotted lines connecting the nodes involved in each case. We can also see the impact that both lexical and syntactic ambiguities have on the number of possible dependencies that will go forward to the subsequent semantic analysis stage. In the first case, it is easy to see how they multiply in relation to the number of tokens in a single cluster, or in other words, to the number of lexical categories that can be assigned to a

[18]the ambiguities in texts written in natural language are subject to a sphere of local influence, and gradually diminish as we read on. If this were not so, communication between human beings would be impossible.

form in a given position in a given phrase from the *corpus*. In the second, we can see an analogue effect resulting from the multiplication of dependencies on the modifiers. An example of this would be *«denticulées»* («`dentate`»), which could be a modifier of either *«feuilles»* («`leaves`») or *«nervures»* («`veins`») in Fig. 2.2. This is a well-known phenomenon linked to the association of prepositional attachments to a nominal syntagma, and which here provides us with two possible interpretations for the sentence at this level: «`leaves with dentate veins`» or, alternatively, «`dentate leaves with veins`».

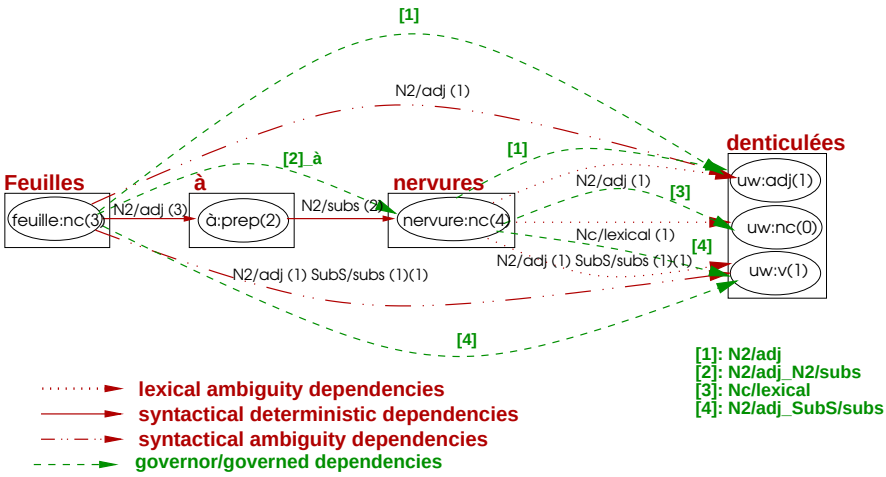

Figure 2.2: Governor/governed dependencies

In this regard, while lexical ambiguity only depends on the structure of the language, syntactic ambiguity is also strongly influenced by the grammatical formalism chosen to describe it, the particular grammar considered and the lack of full grammatical coverage. There are even a good number of situations in which ambiguities will necessarily have to be resolved at semantic level, since they may not originate from either lexical or syntactic causes. A classic example is the use of coordination structures relating entities to a list of adjectives [118], as in *«des sépales ovales-aigus, glabres ou éparsement hérissés»* («`Sepals oval-pointed, smooth or scattered bristly`»), where the property *«hérissés»* («`bristly`») could be attached to the adjectives *«glabres»* («`smooth`») or *«ovales-aigus»* («`oval-pointed`»). In this case, there is only one way to solve the problem, which is to understand the precise nature of the plant organs concerned, something that bears no relationship to the language's morphology or grammar.

Indeed, the phenomenon of ambiguity can be understood as an illustration of the complexity of language itself [113], and is a fundamental problem to be solved in NLP. In these conditions, the set of syntactic schema associated to non-determinism is difficult to estimate, which could complicate an analytic approach to solving the problem. Fortunately, there is a topological condition that is not only easily detectable but also fully characterizes non-determinism in dependency graphs, regardless of its origin. More

in detail, an ambiguity corresponds to a situation where a governed token has more than one governor token. This provides, in turn, a simple mechanism for solving the problem, namely to filter out the less plausible dependencies in favor of the most plausible ones, thereby ensuring that a governed token has only one governor.

However, this idea is not quite so easy to put into practice. Most ambiguities escape our notice because humans are very good at resolving them using context and knowledge of the world, while computer systems do not have full capability in this domain. As a consequence, they often fail to do a good job of disambiguation and all efforts to solve this problem computationally have focused on exploring the context of the discourse and exploiting knowledge-based resources [147].

5.3 | The semantic frame

Once we have constructed the governor/governed dependency graph, which probably reflects a wide range of lexical and syntactic ambiguities, we will now need to prioritize these relationships in order to effectively extract meaning from the text. Intuitively, the process will consist of gathering information from the *corpus* for the purpose of detecting the most plausible relationships. Technically, the proposed heuristic is organized in three levels of complexity, the first two of which are aimed at exploiting the sequence of structures obtained from the previous lexical and syntactic analysis stages, classifying any ambiguities according to their order of priority. The third level will determine what is the semantic information involved in each dependency.

To attain this goal, we first need to introduce some specific notation. Since we are going to extrapolate our estimations from a local context (sentence) to a global one (*corpus*), initial data obtained at sentence level must be then combined and evaluated throughout the whole *corpus* in order to extract new conclusions that can then be applied in each sentence, the process recommencing iteratively. We thus talk about *plausible terms*, *tokens* and *clusters*, notions that will extend the local ones to *corpus* level.

Definition 17 *Let* $\{s_i\}_{1\leq i\leq n}$ *be the sequence of sentences in a* corpus $\mathscr{C}$ *and* $\Theta_{i,j},\ 1 \leq j \leq |\ s_i\ |$ *be the* ocurrence *of a form in the* i*-th sentence,* s_i*. We denote the association of the lexical category (*a*) and semantic class (*b*) to this form, anywhere in* $\mathscr{C}$*, by* $\tilde{\Theta}_{i,j}^{a,b}$ *and we call it* plausible term.

We also naturally extend here the anonymous-variable (resp. free-variable) notation previously introduced for terms, tokens and clusters in Definition 16. ■

We will also need to equip ourselves with the necessary notation for managing governor/governed dependencies at sentence level (resp. *corpus*). In this regard, we will have to refer not only to transitions between tokens (resp. plausible tokens) that constitute the parser output but also the sets of transitions between tokens from two different clusters

(resp. plausible clusters). Finally, and now within the semantic categorization stage, we will have to consider the treatment of transitions between terms (resp. plausible terms).

Definition 18 *Let $s_i,\ 1 \leq i \leq n$ be the i-th sentence in a* corpus $\mathscr{C}$ *and τ be the sequence of the grammar rules necessary to generate the token $\Theta^{c,-}_{i,k}$ from the token $\Theta^{a,-}_{i,j}$ in the governor/governed dependency graph. We denote the* dependency between the tokens $\Theta^{a,-}_{i,j}$ and $\Theta^{c,-}_{i,k}$*, labeled by τ, as $\delta^{\theta^{a,-}_{i,j},\tau,\theta^{c,-}_{i,k}}$.*

The notation can be naturally extended to terms, clusters and plausible structures by exploiting the anonymous-variable notation previously introduced. When a dependency relates plausible structures we talk about a plausible dependency. ■

5.3.1 | Categorization of tokens

The goal is to compute which, for each cluster in the text, is the most probable token. In other words, we want to determine the lexical category for each occurrence of a given form in the position of a sentence in the *corpus*. The process, which is an iterative one, corresponds to the equations in Table 1, which we comment on below:

$$P(\Theta^{a,-}_{i,j})_{\text{local(0)}} = \frac{1}{|\{\Theta^{X,-}_{i,j}\}|} \tag{1}$$

$$P(\tilde{\Theta}^{a,-}_{i,j})_{\text{global(n+1)}} = \frac{\sum\limits_{\Theta_{k,l}=\Theta_{i,j}} P(\Theta^{a,-}_{k,l})_{\text{local(n)}}}{\sum\limits_{\Theta^{X,-}_{k,l},\Theta_{k,l}=\Theta_{i,j}} P(\Theta^{X,-}_{k,l})_{\text{local(n)}}} \tag{2}$$

$$P(\Theta^{a,-}_{i,j})_{\text{local(n+1)}} = \frac{P(\tilde{\Theta}^{a,-}_{i,j})_{\text{global(n+1)}}}{\sum\limits_{\Theta^{X,-}_{k,l},\Theta_{k,l}=\Theta_{i,j}} P(\tilde{\Theta}^{X,-}_{k,l})_{\text{global(n+1)}}} \tag{3}$$

Table 1: Model for categorization of tokens

(1). The process starts by calculating the local probability, at sentence level, that can be associated to a token in a cluster. This is a simple ratio that depends on the number of tokens involved in the said cluster. Obviously, if there is only one token in the cluster, its probability would therefore be 1.

(2). This defines the global probability of a plausible token in the *corpus*, at iteration $n+1$ in the process. It is calculated as a proportion of the local probability associated with tokens of the same lexical category and form as those of the token in question, in relation to the probability when the lexical category is free.

(3). This equation determines the value of the local probability that can be associated with a token in a cluster, at iteration $n+1$ in the process. In order to do so, we allocate the probabilities calculated globally, distributing them proportionally between the global probabilities of the plausible tokens associated with the cluster.

The iterative process continues until convergence [123] at a fixed point, or at a fixed approximation threshold, is achieved.

5.3.2 | Categorization of dependencies between tokens

The objective is to provide an objective measure of the viability of the syntactic dependencies generated by the parser between the previously categorized tokens. Bearing in mind that the topological characterization of syntactic ambiguity is the existence of various governing tokens for a single governed one, we thus seek to define which is the latter's true governor amongst the possibilities put forward by the parser, in order to eliminate such syntactic ambiguity. We once again opt for an iterative strategy, in this case determined by the equations in Table 2, which we now describe:

$$W(\delta^{\Theta^{a,-}_{i,j},\tau,\Theta^{b,-}_{i,k}}) = \frac{|S \overset{*}{\Rightarrow} \Theta^{a,-}_{i,j} \overset{\tau}{\Rightarrow} \Theta^{b,-}_{i,k}|}{\sum_{\delta^{\Theta^{Y,-}_{i,\bar{X}},T,\Theta^{Z,-}_{i,k}}} |S \overset{*}{\Rightarrow} \Theta^{Y,-}_{i,\bar{X}} \overset{T}{\Rightarrow} \Theta^{Z,-}_{i,k}|} \tag{4}$$

$$P(\delta^{\Theta^{a,-}_{i,j},\tau,\Theta^{b,-}_{i,k}})_{\text{local(0)}} = \frac{P(\Theta^{a,-}_{i,j})_{\text{local}} \cdot\ P(\Theta^{b,-}_{i,k})_{\text{local}} \cdot\ W(\delta^{\Theta^{a,-}_{i,j},\tau,\Theta^{b,-}_{i,k}})}{\sum_{\Theta^{Y,-}_{i,\bar{X}},\Theta^{Z,-}_{i,k},\delta^{\Theta^{Y,-}_{i,\bar{X}},T,\Theta^{Z,-}_{i,k}}} P(\Theta^{Y,-}_{i,\bar{X}})_{\text{local}} \cdot\ P(\Theta^{Z,-}_{i,k})_{\text{local}} \cdot\ W(\delta^{\Theta^{Y,-}_{i,\bar{X}},T,\Theta^{Z,-}_{i,k}})} \tag{5}$$

$$P(\delta^{\tilde{\Theta}^{a,-}_{i,j},\tau,\tilde{\Theta}^{b,-}_{i,k}})_{\text{global(n+1)}} = \frac{\sum_{\Theta_{l,m}=\Theta_{i,j},\Theta_{l,p}=\Theta_{i,k}} P(\delta^{\Theta^{a,-}_{l,m},\tau,\Theta^{b,-}_{l,p}})_{\text{local(n)}}}{\sum_{\delta^{\Theta^{Y,-}_{l,\bar{X}},T,\Theta^{Z,-}_{l,p}},\Theta_{l,p}=\Theta_{i,k}} P(\delta^{\Theta^{Y,-}_{l,\bar{X}},T,\Theta^{Z,-}_{l,p}})_{\text{local(n)}}} \tag{6}$$

$$P(\delta^{\Theta^{a,-}_{i,j},\tau,\Theta^{b,-}_{i,k}})_{\text{local(n+1)}} = \frac{P(\delta^{\tilde{\Theta}^{a,-}_{i,j},\tau,\tilde{\Theta}^{b,-}_{i,k}})_{\text{global(n+1)}}}{\sum_{\delta^{\tilde{\Theta}^{Y,-}_{l,\bar{X}},T,\tilde{\Theta}^{Z,-}_{l,m}},\Theta_{l,m}=\Theta_{i,k}} P(\delta^{\tilde{\Theta}^{Y,-}_{l,\bar{X}},T,\tilde{\Theta}^{Z,-}_{l,m}})_{\text{global(n+1)}}} \tag{7}$$

Table 2: Model for categorization of dependencies between tokens

(4). Before beginning the iterative process, we will calculate an initial weight for each syntactic dependency depending on its label. We thereby seek to assign more

importance to the dependencies shared by a greater number of parses, amongst those sharing a single governed cluster.

(5). The iterative process begins by calculating the local probability, at sentence level, that can be associated with a syntactic dependency. Given that such dependencies are characterized by their governor and governed tokens, and by their label, we will make this probability depend on the local probabilities of such tokens, as well as on the weight assigned to the associated label. It is calculated as a proportion of the above-mentioned values for the syntactic dependency in question, in relation to the set of dependencies associated with the governed token cluster.

(6). This equation defines the global probability at *corpus* level of a plausible dependency at iteration $n + 1$ of the process. It is calculated as a proportion of the local probability associated with syntactic dependencies coinciding with the one under consideration (except in the locating sentence), in relation to the set of local dependencies associated with governed tokens that also coincide with the one under consideration (except in the locating cluster).

(7). It establishes the value of the local probability of a dependency in iteration $n+1$ of the process. To this end we allocate the probabilities calculated globally, distributing them proportionally amongst the global probabilities of the plausible syntactic dependencies associated with the governed tokens coinciding with the one under consideration (except in the locating cluster).

As was the case for lexical categorization, the process repeats itself until it converges at a fixed point, or at a fixed approximation threshold.

5.3.3 | Categorization of dependencies between terms

The goal at this level is to attach the semantic classes to the tokens involved in a same syntactic dependency, in order to identify the semantic dependencies between terms in two clusters. To be more precise, given a governed term, we seek to define its governor by means of the syntactic dependencies categorized previously. In order to do so, we need to introduce a certain amount of additional notation beforehand.

Definition 19 *Let $s_i,\ 1 \leq i \leq n$ be the i-th sentence in a* corpus *$\mathscr{C}$, and $\mathscr{T}$ (resp. $\mathscr{F}$) be the set of semantic classes (resp. forms) associated to $\mathscr{C}$ (resp. to $\mathscr{T}$) by means of some reliable technique. We then denote by $\mathscr{F}(b)$ the subset of forms associated to $b \in \mathscr{T}$, and we say that $\Theta_{i,j}^{a,b},\ 1 \leq j \leq \mid s_i \mid$ is an* stable term *iff $b \in \mathscr{T}$ and $\Theta_{i,j} \in \mathscr{F}(b)$.*

■

Intuitively, a term is stable when we have reliable information about the correspondence between its semantic class and its form, obtained either from the user

Entities	**Lemmas** (in French)
organe	fleur, staminode, tige, feuille, hypanthe, périanthe, rameau, ...
fruit	fruit, samare, drupe, capsule, akène
Properties	**Lemmas** (in French)
couleur	verdâtre, violacé, noirâtre, violet, jaunâtre, orange, roux, rose
forme	obconique, oblancéolé, oblong, bifolié, crateriforme, punctiforme, ...
taille	moyen, petit, double, épais, inégal, entier, longue
texture	hispide, bifide, globuleux, coriace, velutineux, gélatineux, barbu
position	antérieur, dessus, voisin, seul, latéral, transversal

Table 3: The set $\mathscr{T}$ of initial semantic classes (types) for the running example

or by means of a method held to be completely trustworthy. Our proposal contemplates the use of both mechanisms [49]. On the one hand, the user defines the set of semantic categories that, in our running *corpus* $\mathscr{B}$, are organized as entities ($\mathscr{E}$) and properties ($\mathscr{P}$), together with a set of initial associated forms such as the one shown in Table 3.

Word(in French)	**Position**	**Class**	**Word**(in French)	**Position**	**Class**
teinté	[2]	Couleur	épaisseur	[1]	Taille
texture	[2]	Texture	atteindre	[1]	Organe/Fruit
taille	[1]	Organe/Fruit	taille	[2]	Taille
teinte	[1]	Organe/Fruit	teinte	[2]	Couleur
couleur	[1]	Organe/Fruit	couleur	[2]	Couleur
texture	[1]	Organe/Fruit	texture	[2]	Texture
forme	[1]	Organe/Fruit	forme	[2]	Forme
position	[1]	Organe/Fruit	position	[2]	Position
altitude	[1]	Organe/Fruit	environ	[2]	Taille
tache	[1]	Organe/Fruit	tache	[2]	Couleur
longueur	[1]	Taille	formé	[2]	Organe/Fruit
composé	[1,2]	Organe/Fruit	dépassant	[2]	Taille
diamètre	[1]	Taille	contour	[2]	Forme/Texture

Table 4: A sample section from the collocations file

On the other, the system makes use of *collocations*, sequences of words that co-occur more often than would be expected by chance and in which they keep their original meaning, in contrast to the case of *locutions*. The idea is to filter out the parse in order to locate collocations that enable a form to be associated with a semantic class. For the occasion, we represent a collocation as a triple of the form *marker/position/semantic class*. The marker serves to identify the collocation for which the form in the indicated position can be associated with the semantic class, as shown in Table 4, once again in the case of the running *corpus* $\mathscr{B}$. So, for example, in the sentence *«teintées de rose»* («`rose-tinted`»), the presence of the marker *«teinté»* («`tinted`») reveals that *«rose»* («`rose`») is an instance of the semantic class «`color`». The iterative process thus corresponds to the equations in Table 5, which we will now describe:

$$W(\Theta^{a,-}_{i,j}) > \frac{|\{\Theta^{a,X}_{i,j}\}_{X\in\mathscr{T},\Theta_{i,j}\in\mathscr{F}(X)}|}{|\{\Theta^{a,X}_{i,j}\}_{X\in\mathscr{T}}|} \subseteq (0,1] \tag{8}$$

$$W(\Theta_{i,j}^{a,b}) = \begin{cases} \frac{W(\Theta_{i,j}^{a,\_})}{|\{\Theta_{i,j}^{a,X}\}_{X\in\mathscr{T},\Theta_{i,j}\in\mathscr{F}(X)}|} & \text{si } \Theta_{i,j} \in \mathscr{F}(b) \\ \frac{1-W(\Theta_{i,j}^{a,\_})}{|\{\Theta_{i,j}^{a,X}\}_{X\in\mathscr{T},\Theta_{i,j}\notin\mathscr{F}(X)}|} & \text{en otro caso} \end{cases} \tag{9}$$

$$P(\delta^{\Theta_{i,j}^{a,b},\tau,\Theta_{i,k}^{c,d}})_{\text{local(0)}} = \frac{P(\delta^{\Theta_{i,j}^{a,\_},\tau,\Theta_{i,k}^{c,\_}})_{\text{local}} \cdot W(\Theta_{i,j}^{a,b}) \cdot W(\Theta_{i,k}^{c,d})}{\sum_{\Theta_{i,X}^{Y,Z},\Theta_{i,k}^{V,W},\delta^{\Theta_{i,X}^{Y,\_},T,\Theta_{i,k}^{V,\_}}} P(\delta^{\Theta_{i,X}^{Y,\_},T,\Theta_{i,k}^{V,\_}})_{\text{local}} \cdot W(\Theta_{i,X}^{Y,Z}) \cdot W(\Theta_{i,k}^{V,W})} \tag{10}$$

$$P(\delta^{\tilde{\Theta}_{i,j}^{a,b},\tau,\tilde{\Theta}_{i,k}^{c,d}})_{\text{global(n+1)}} = \frac{\sum_{\Theta_{l,m}=\Theta_{i,j},\Theta_{l,p}=\Theta_{i,k}} P(\delta^{\Theta_{l,m}^{a,b},\tau,\Theta_{l,p}^{c,d}})_{\text{local(n)}}}{\sum_{\delta^{\Theta_{l,X}^{Y,Z},T,\Theta_{l,p}^{V,W}},\Theta_{l,p}=\Theta_{i,k}} P(\delta^{\Theta_{l,X}^{Y,Z},T,\Theta_{l,p}^{V,W}})_{\text{local(n)}}} \tag{11}$$

$$P(\delta^{\Theta_{i,j}^{a,b},\tau,\Theta_{i,k}^{c,d}})_{\text{local(n+1)}} = \frac{P(\delta^{\tilde{\Theta}_{i,j}^{a,b},\tau,\tilde{\Theta}_{i,k}^{c,d}})_{\text{global(n+1)}}}{\sum_{\delta^{\tilde{\Theta}_{l,X}^{Y,Z},T,\tilde{\Theta}_{l,m}^{V,W}},\Theta_{l,m}=\Theta_{i,k}} P(\delta^{\tilde{\Theta}_{l,X}^{Y,Z},T,\tilde{\Theta}_{l,m}^{V,W}})_{\text{global(n+1)}}} \tag{12}$$

Table 5: Model for categorization of dependencies between terms

(8). Before commencing the iterative process, we will give each token a weight verifying the condition presented, the value of which we justify below.

(9). We shall now distribute equitatively the weight calculated from Equation 8 between the stable terms. This ensures that the weight we here associate with non-stable terms in this token is lower than that associated with the former. We thus aim to give initial preference to the stable terms.

(10). The iterative process commences with the calculation of the local probability, at sentence level, that can be associated with a semantic dependency. Given that the latter is perfectly characterized by its governor and governed terms, together with the syntactic dependency between their associated tokens, we will make this value depend on the weights associated with the said terms, as well as on the local probability corresponding to the syntactic dependency. It is calculated as a proportion of the said values for the semantic dependency in question, in relation to the set of dependencies associated with the governed term cluster.

(11). This equation defines the global probability in the *corpus* of a plausible semantic dependency at iteration $n+1$ of the process. It is calculated as a proportion of the local probability associated with the semantic dependencies that coincide with the

one under consideration (except in the locating sentence), in relation to the set of the local dependencies associated with the governed terms that also coincide with the one under consideration (except in the locating cluster).

(12). It establishes the value of the local probability that can be associated with a semantic dependency at iteration $n + 1$ of the process. To this end we allocate the globally calculated probabilities, distributing them proportionally between the global probabilities of the plausible semantic dependencies associated with governed terms that coincide with the one under consideration (except in the locating cluster).

As in the case of the categorization of syntactic dependencies, the process repeats itself until it convergence at a fixed point or approximation to a predetermined threshold is reached. We call the resulting structure the *semantic of the corpus* $\mathscr{C}$ we are working with.

Definition 20 *Let* $\{s_i\}_{1\leq i\leq n}$ *be the sequence of sentences in a* corpus $\mathscr{C}$*, and* $\mathscr{T}$ *(resp.* $\mathscr{F}$*) be the set of semantic classes (resp. forms) associated to* $\mathscr{C}$ *(resp. to* $\mathscr{T}$*) by means of some reliable technique. We then define the* semantic of the corpus $\mathscr{C}$ *as:*

$$\mathscr{S}_{\mathscr{C}} := \{\delta^{\Theta_{i,j}^{a,b},\tau,\Theta_{i,k}^{c,d}},\ P(\delta^{\Theta_{i,j}^{a,b},\tau,\Theta_{i,k}^{c,d}})_{local} = \max\{P(\delta^{\Theta_{i,j}^{X,Y},Z,\Theta_{i,k}^{V,W}})_{local}\}\}$$

where max *is the maximal function on* $\mathbb{N}$*, and* $\delta^{\Theta_{i,j}^{X,Y},Z,\Theta_{i,k}^{V,W}}$ *are the dependencies computed as result of the knowledge adquisition process previously described.*

We can naturally restrict the concept to refer the semantic of a document $\mathscr{D}$ in $\mathscr{C}$ *by*

$$\mathscr{S}_{\mathscr{C}}^{\mathscr{D}} := \{\delta^{\Theta_{i,j}^{a,b},\tau,\Theta_{i,k}^{c,d}} \in \mathscr{S}_{\mathscr{C}},\ s_i \in \mathscr{D}\}$$

■

Intuitively, we define the semantic of the *corpus* as the set of most probable dependencies between its terms. This comprises all the syntactic and semantic relationships considered as viable, between the lexical categories in the text studied. The semantic of the *corpus* will be the starting point for the generation of the conceptual graphs serving us as formal knowledge representation for IR purposes.

5.4 | Conceptual graph generation

We are now ready to structure the BGs we are going to use in our experimental tests. Although the proposal is independent of the knowledge domain considered, it is necessary to locate our work in a concrete one, in order to suitably model the support serving as a

basis for subsequently defining such graphs. As already commented, our choice has been the biological domain, and more specifically that of botanical description, for which we have taken as our reference the running *corpus* $\mathscr{B}$. The complexity, length and specific nature of this kind of content makes it hard for queries to be expressed by a non-expert user in any way other than a purely prospective one. The chosen topic is thus a particularly suitable one for validating capacities in the sphere of handling vague and incomplete information, and therefore justifies our decision.

In this sense, we retake the set $\mathscr{T}$ of semantic classes (types) shown in Table 3 for the *corpus* $\mathscr{B}$, in order to introduce a partial order on it as follows:

$$\forall\, t \in \mathscr{E} = \{\textit{fruit, organe}\},\ t \leq \varepsilon \leq \top$$

$$\forall\, t \in \mathscr{P} = \{\textit{couleur, forme, taille, texture, position}\},\ t \leq \rho \leq \top$$

where ε (resp. ρ) is the greater element for the entities (resp. properties) $\mathscr{E}$ (resp. properties $\mathscr{P}$). In this way, we introduce our running support $\mathcal{S} = (\mathcal{T}_{\mathcal{C}_{\mathscr{B}}}, \mathcal{T}_{\mathcal{R}_{\mathscr{B}}}, \mathcal{I}_{\mathscr{B}})$ by defining:

$$\mathcal{T}_{\mathcal{C}_{\mathscr{B}}} := \{\varepsilon, \rho\} \cup \mathscr{E} \cup \mathscr{P} \cup \{\top\}$$
$$\mathcal{T}_{\mathcal{R}_{\mathscr{B}}} := \{[b, \tau, d], [b, *, d],\ \exists\, \delta^{\Theta^{a,b}_{i,j}, \tau, \Theta^{c,d}_{i,k}} \in \mathscr{S}_{\mathscr{B}}\} \cup \{[\varepsilon, *, \varepsilon]\} \cup \{[\varepsilon, *, \rho]\} \cup \{[\rho, *, \rho] \cup \{[\top, *, \top]\}$$
$$\mathcal{I}_{\mathscr{B}} := \{\Theta^{a,-}_{i,j},\ \Theta^{c,-}_{i,k}\}_{\delta^{\Theta^{a,-}_{i,j}, -, \Theta^{c,-}_{i,k}}}$$

where $\mathscr{S}_{\mathscr{B}}$ is the semantic associated with the running *corpus* $\mathscr{B}$.

Intuitively, we consider that the set of concepts $\mathcal{T}_{\mathcal{C}_{\mathscr{B}}}$ that we are interested in handling for the *corpus* $\mathscr{B}$ can be classified into entities and properties, as described in Table 3, and ordering is not considered between similar and/or dissimilar elements. A subsumption relation is only defined between individual entities (resp. properties) and the corresponding generic element, $*$. With regard to the set of relations $\mathcal{T}_{\mathcal{R}_{\mathscr{B}}}$, they are directly extracted from $\mathscr{S}_{\mathscr{B}}$ through the transitional dynamic, and summarize a transition between two terms from the point of view of the semantic classes (types) involved. As extra elements, we add triples representing any possible transition in the semantically related generic concepts. The partial order we consider in $\mathcal{T}_{\mathcal{C}_{\mathscr{B}}}$ is the one that is naturally induced by the order previously defined in $\mathscr{T}$. Finally, we define the markers $\mathcal{I}_{\mathscr{B}}$ as the set of forms in the *corpus* $\mathscr{B}$.

We can now introduce the BGs we are considering on this support. Our starting point is the semantic $\mathscr{S}_{\mathscr{D}_m}$ associated with each of the documents constituting the *corpus* $\mathscr{B} = \bigcup_{m \in M} \mathscr{D}_m$, where M is the number of these documents:

$$\mathcal{C}_{\mathscr{D}_m} := \{\Theta^{a,b}_{i,j}, \Theta^{c,d}_{i,k}\}_{\delta^{\Theta^{a,b}_{i,j}, -, \Theta^{c,d}_{i,k}} \in \mathscr{S}_{\mathscr{D}_m}} \qquad \mathcal{R}_{\mathscr{D}_m} := \{[b, \tau, d],\ \exists\, \delta^{\Theta^{-,b}_{-,-}, \tau, \Theta^{-,d}_{-,-}} \in \mathscr{S}_{\mathscr{D}_m}\}$$

$$\mathcal{E}_{\mathscr{D}_m} := \bigcup_{\delta^{\Theta^{a,b}_{i,j}, \tau, \Theta^{c,d}_{i,k}} \in \mathscr{S}_{\mathscr{D}_m}} \{([b, \tau, d], 1, \Theta^{a,b}_{i,j}), ([b, \tau, d], 2, \Theta^{c,d}_{i,k})\}$$

$$\mathcal{L}_{\mathscr{D}_m}(X) := \begin{cases} [b, \Theta^{a,\_}_{i,j}] & \text{si} \quad X = \Theta^{a,b}_{i,j} \in \mathcal{C}_{\mathscr{D}_m} \\ X & \text{si} \quad X \in \mathcal{R}_{\mathscr{D}_m} \\ 1 & \text{si} \quad X = (\_, 1, \_) \in \mathcal{E}_{\mathscr{D}_m} \\ 2 & \text{si} \quad X = (\_, 2, \_) \in \mathcal{E}_{\mathscr{D}_m} \end{cases}$$

Succinctly, a conceptual node in $\mathcal{C}_{\mathscr{D}_m}$ is any term involved in the semantic $\mathscr{S}_{\mathscr{D}_m}$, while relation nodes in $\mathcal{R}_{\mathscr{D}_m}$ are elements of $\mathcal{T}_{\mathcal{R}_{\mathscr{B}}}$ associated to transitions in $\mathscr{S}_{\mathscr{D}_m}$. The multiset of edges $\mathcal{E}_{\mathscr{D}_m}$ contains in this case only binary relations, the governor (resp. governed) term corresponding to the first (resp. second) triple.

With regard to the labeling function $\mathcal{E}_{\mathscr{D}_m}$, it makes it possible to recover the semantic class and the token associated to a given term representing a concept, whilst implementing the identity on the relations, since in our case we build these directly from the semantic of the *corpus*. The value of this function on edges identifies governor (1) and governed edges (2).

6 | The testing frame

The traditional model for experimental evaluation of IR systems [30, 31] implies three complementary tasks: the compilation of a document collection, the definition of a number of trustworthy evaluation measures and the choice of a set of topics (queries). We here take as document collection our running *corpus* $\mathscr{B}$, a complete and real world compendium of botanic contents. With regard to the other two tasks, our goal is to minimize the selecting effort in topics and the judging one in QREL generation. This should allow us to deal with test collections including an arbitrary number of documents on any domain, which would be almost impossible to achieve if done on a human scale.

Whatever the case may be, any IR evaluation task depends on the assumptions that relevance is a meaningful concept and that *relevance* assessments possess the requisite stability on which valid performance can be constructed.

Definition 21 *Let $\mathcal{D} = \{d_i\}_{i\in I}$ be a document collection and $\mathcal{Q} = \{c_j\}_{j\in J}$ be a finite set of topics (queries). We say that a document $d_i \in \mathcal{D}$ is* relevant to a topic $c_j \in \mathcal{Q}$ *iff a human expert considers that the document in question contains relevant information regarding the said query. Otherwise, we say that $d_i \in \mathcal{D}$ is* non-relevant to $c_j \in \mathcal{Q}$*. We denote by* rel$(c_j, \mathcal{D})$ *$\mathcal{D}$ the set of documents in $\mathcal{D}$ which are relevant to $c_j \in \mathcal{Q}$, and by* nrel$(c_j, \mathcal{D})$ *which are non-relevant.*

■

Since our goal here is to discriminate the effectiveness between IR systems, and that this in turn involves the ability to detect which system is more sensitive in the identification of relevant documents, it is first necessary to ensure the operational stability

of relevance. However, we must also mention the apparent existence of factors that impact on the correctness of this definition [131]. This is the case of assessor disagreement or even the individual inconsistencies [135] of a single assessor, factors which are reinforced by the fact that we are talking of a continuous magnitude that we can rank in a sequence of values [141]. With regard to this, we assume that the influence of these destabilizing factors is minimal, as has been primarily suggested in [66], and subsequently experimentally corroborated in [158]. In the same sense, disagreement about the number of relevant documents does not seem to have very much net impact on system rankings at all [135], probably because a greater number of relevant documents benefits most systems uniformly.

Turning our attention now to topic selection and the ranking of IR systems, we can distinguish two common generic frameworks according to the state-of-the-art. The first of these is the one inspired by the experience accumulated in TREC events over the decades, and characterized by the preeminent use of human judging[19], disregarding the influence of topic ease in the process. We talk here about the *human-based assessment frame*. The second, however, is a set of techniques inspired in two reasonable assumptions sketched [100] in the *«easy and difficult principle»* for topics and *«good and bad principle»* for IR systems. Contrary to human-based assessment, this formalizes topic ease from QREL-based measures and as a major impacting factor in this task. More in detail, the former principle establishes that we should attach more (resp. less) weight to both errors on easy (resp. difficult) questions and correct answers on difficult (resp. easy) ones. The latter assumes that we should be able to put difficult questions to good systems, but only easy questions to bad ones. Henceforth, we shall refer to this second frame as the *machine-based assessment one*.

Alternatively, in dealing exclusively with the ranking of IR systems, a third way has been proposed that completely dispenses with the use of QREL-based resources [164]. This is a method for evaluating the performance of a search engine using a measure called *reference count*, a kind of score that computes the number of occurrences for the top documents returned in the results of a collection of other retrieval systems.

6.1 | Ranking IR systems using QRELs

The use of QREL is at the heart of most evaluation measures for IR systems, and has been popularized in the community by TREC events. We shall distinguish two approaches according to whether or not we take into account the order associated with the ranked retrieval results, these currently being standard with search engines.

[19] through QREL or similar mechanisms, such as pseudo-QREL.

6.1.1 | Set-based evaluation measures

Measures of this kind estimate the quality of an unordered set of retrieved documents. Such techniques have been associated, ever since the first retrieval experiments, to a bi-dimensional model [66] of IR evaluation. In other words, no ranked context is taken into account and evaluation only focuses on the relevant or non-relevant character of the documents retrieved.

Definition 22 *Let σ be an* IR *system, $\mathcal{D} = \{d_i\}_{i\in I}$ be a document collection and $\mathcal{Q} = \{c_j\}_{j\in J}$ be a finite set of topics (queries). We then define the* precision *(resp.* recall*))* of σ on the topic c_j for the document collection $\mathcal{D}$ *as:*

$$\mathrm{P}(\sigma, c_j, \mathcal{D}\ ) := \frac{|\mathrm{ret}(\sigma, c_j, \mathcal{D}) \cap \mathrm{rel}(c_j, \mathcal{D})\ |}{|\mathrm{ret}(\sigma, c_j, \mathcal{D})|} \tag{13}$$

$$(\textit{resp.}\ \mathrm{R}(\sigma, c_j, \mathcal{D}\ ) := \frac{|\mathrm{ret}(\sigma, c_j, \mathcal{D}) \cap \mathrm{rel}(c_j, \mathcal{D})\ |}{|\mathrm{rel}(c_j, \mathcal{D})\ |}) \tag{14}$$

where $\mathrm{ret}(\sigma, c_j, \mathcal{D})$ *(resp.* $\mathrm{rel}(\sigma, c_j, \mathcal{D})$*) is the set of documents in $\mathcal{D}$ retrieved by σ (resp. which are relevant) for the topic $c_j \in \mathcal{Q}$.*

■

Both *precision* and *recall* were introduced by Cleverdon *et al.* in [29]. Intuitively, precision (resp. recall) represents the ratio between the number of relevant documents retrieved and the number of retrieved documents (resp. relevant documents), i.e., the positive predictive value (resp. the sensitivity) of the search task. So, precision (resp. recall) assesses the accuracy (resp. the comprehensiveness) of a search. In particular, precision (resp. recall) is undefined when no document is retrieved (resp. when there is no relevant document) in the collection and is minimal (resp. maximal) when all the documents are returned by the search. Be that as it may, these are complementary concepts calculated on the whole list of documents returned by the system, which poses some problems for estimating the latter's effectiveness. This justifies the introduction by von Rijsbergen in [150] of the F_β measure as a way of estimating the effectiveness of retrieval with respect to a user who attaches β times as much importance to recall as precision.

Definition 23 *Let σ be an* IR *system, $\mathcal{D} = \{d_i\}_{i\in I}$ be a document collection and $\mathcal{Q} = \{c_j\}_{j\in J}$ be a finite set of topics (queries). We then define, for $\beta \in \mathbb{R}^+ \cup \{0\}$, the F_β* measure of σ on the topic c_j for the document collection $\mathcal{D}$ *as:*

$$\mathrm{F}_\beta(\sigma, c_j, \mathcal{D}) := \frac{(1+\beta^2)\ .\ [\mathrm{P}(\sigma, c_j, \mathcal{D})\ .\ \mathrm{R}(\sigma, c_j, \mathcal{D})]}{\beta^2\ .\ \mathrm{P}(\sigma, c_j, \mathcal{D}) + \mathrm{R}(\sigma, c_j, \mathcal{D})} \tag{15}$$

In the particular case $\beta = 1$, *we talk about* F-measure..

■

The F_β measure allows us to weight emphasis on precision over recall using β as a control value. When $\beta = 1$, we obtain the *harmonic mean* of precision and recall, which, in contrast to an arithmetic mean, requires both values to be high. For values $\beta < 1$ it weights precision more, whilst for values $\beta > 1$ it weights recall more. On the other hand, none of the above take into account the proportion of non-relevant documents that are retrieved, a situation that the introduction of the *fall-out rate* attempts to remedy.

Definition 24 *Let* σ *be an* IR *system,* $\mathcal{D} = \{d_i\}_{i \in I}$ *be a document collection and* $\mathcal{Q} = \{c_j\}_{j \in J}$ *be a finite set of topics (queries). We then define the* fall-out of σ on the topic c_j for the document collection $\mathcal{D}$ *as:*

$$\text{FO}(\sigma, c_j, \mathcal{D}) := \frac{|\text{ret}(\sigma, c_j, \mathcal{D}) \cap \text{nrel}(c_j, \mathcal{D})\,|}{|\text{nrel}(c_j, \mathcal{D})\,|} \tag{16}$$

where $\text{nrel}(c_j, \mathcal{D})$ *is the set of documents in* $\mathcal{D}$ *which are non-relevant to* $c_j \in \mathcal{Q}$.

■

In such a way, fall-out can be looked at as the probability that a non-relevant document was retrieved. Trivially, this value returns zero when no documents are retrieved in response to any query. The fall-out rate was introduced by Salton and McGill in [128].

6.1.2 | Rank-based evaluation measures

These kinds of measure take into account the order in which the returned documents are presented. As a consequence, two practical improvements can be derived in relation to the above metrics. The first of these is the real contribution implied by being able to dispose of extra information on the degree of relevance associated by the search engine to the documents for a given query. The second is that we can estimate the efficiency of an IR system, even when we are only interested in computing it at fixed low levels of retrieved results. This is typically the case of a web search, where the user commonly pays little or no attention to results not contained in the first few pages. Formally [124], these enhancements translate into both *stability*[20] and *sensitivity*[21] for the evaluation task.

A first approach to obtain this consists of plotting precision against recall after each retrieved document. To do so, we shall first synchronize both precision and recall on the basis of the first k returned documents.

[20] the stability of a measure is related [20] to how consistently it is able to identify differences between systems over a sample of queries or topics.

[21] also called recall rate, it refers to the discrimination power of IR evaluation metrics, given a test collection and a set of runs submitted to the task defined by that collection [160].

Definition 25 *Let σ be an* IR *system, $\mathcal{D} = \{d_i\}_{i\in I}$ be a document collection and $\mathcal{Q} = \{c_j\}_{j\in J}$ be a finite set of topics (queries). We then define the* precision *(resp. la* recall*)* at k returned documents, of σ on the topic c_j for the document collection $\mathcal{D}$, *denoted by* P@$k(\sigma, c_j, \mathcal{D})$ *(resp.* R@$k(\sigma, c_j, \mathcal{D})$*), as:*

$$\text{P@}k(\sigma, c_j, \mathcal{D}) := \frac{|\{\text{rret}(\sigma, c_j, \mathcal{D})_l\}_{l=1}^{k} \cap \text{rel}(c_j, \mathcal{D})\,|}{k} \tag{17}$$

$$(\textit{resp.}\ \text{R@}k(\sigma, c_j, \mathcal{D}) := \frac{|\{\text{rret}(\sigma, c_j, \mathcal{D})_l\}_{l=1}^{k} \cap \text{rel}(c_j, \mathcal{D})\,|}{|\text{rel}(c_j, \mathcal{D})\,|}) \tag{18}$$

where rret$(\sigma, c_j, \mathcal{D})$ *is the list, ranked by relevance, of documents returned by σ for the topic c_j.*

■

We can now express precision as a function of recall, simply by computing both measures at each synchronization point. As a result we obtain a precision/recall graph [98, 114].

Definition 26 *Let σ be an* IR *system, $\mathcal{D} = \{d_i\}_{i\in I}$ be a document collection and $\mathcal{Q} = \{c_j\}_{j\in J}$ be a finite set of topics (queries). We can then express* precision of σ on the topic c_j for the document collection $\mathcal{D}$ as a function of recall *by:*

$$\text{P}_\text{R}(\sigma, c_j, \mathcal{D}, r) := \text{P@}k(\sigma, c_j, \mathcal{D}),\ r = \text{R@}k(\sigma, c_j, \mathcal{D}) \tag{19}$$

■

Intuitively, we are computing the precision in the same instant as recall, just each time that a document is returned by the search engine. As a result [99], curves of this kind have a distinctive saw-tooth shape since if the $(k+1)$th document retrieved is non-relevant then recall is the same as for the top k documents, but precision has dropped. If it is relevant, then both precision and recall increase, and the curve jags up and to the right. In this sense, it is often useful to remove these jiggles and the standard way to do this is through interpolation.

Definition 27 *Let σ be an* IR *system, $\mathcal{D} = \{d_i\}_{i\in I}$ be a document collection and $\mathcal{Q} = \{c_j\}_{j\in J}$ be a finite set of topics (queries). We then define the* interpolated precision of σ on the topic c_j for the document collection $\mathcal{D}$ as a function of recall *by:*

$$\text{IP}_\text{R}(\sigma, c_j, \mathcal{D}, r) := \max_{r'\geq r} \text{P}_\text{R}(\sigma, c_j, \mathcal{D}, r') \tag{20}$$

■

In this manner, the interpolated precision at a certain recall level is defined as the highest precision found for any greater one, solving the problem raised. On the other hand, although we have used P@k as a first step to introducing precision/recall graphs, the concept is also of interest in itself. So, an advantage usually argued in its favor is that it does not require an estimate of the set of relevant documents. However, for this same reason they do not average well, since the set in question strongly impacts precision and cannot be considered as a stable evaluation criterion [99]. An alternative alleviating this problem is *R-precision* (resp. la *R-recall*) [128].

Definition 28 *Let σ be an* IR *system, $\mathcal{D} = \{d_i\}_{i\in I}$ be a document collection and $\mathcal{Q} = \{c_j\}_{j\in J}$ be a finite set of topics (queries). We then define the* R-precision*, denoted by* P@$R(\sigma, c_j, \mathcal{D})$ *(resp.* R-recall*, denoted by* R@$R(\sigma, c_j, \mathcal{D})$*),* of σ on the topic c_j for the document collection $\mathcal{D}$ *as:*

$$R\text{-P}(\sigma, c_j, \mathcal{D}\ ) := \text{P@}R(\sigma, c_j, \mathcal{D}) \tag{21}$$

$$(\textit{resp. } R\text{-R}(\sigma, c_j, \mathcal{D}\ ) := \text{R@}R(\sigma, c_j, \mathcal{D})) \tag{22}$$

where $R = |\text{rel}(c_j, \mathcal{D})|$.

■

Intuitively, if the document collection includes R relevant documents for a query, the R-P provides the number of relevant ones once the top R results have been studied by the system. In short, it refers to the best precision on the P_R graph, which justifies its also being known as the *break-even point of* P_R, since it is where precision and recall coincide.

In any case, none of the graded-relevance metrics are at present as widely used as traditional binary-relevance metrics such as *average precision* (AP), which provides a geometric interpretation for precision/recall graphs [127]. In effect, it calculates the area under the P_R curve, which implies estimating its average value over the interval $[0, 1]$ for recall.

Definition 29 *Let σ be an* IR *system, $\mathcal{D} = \{d_i\}_{i\in I}$ be a document collection and $\mathcal{Q} = \{c_j\}_{j\in J}$ be a collection of topics (queries). We then define the* average precision of σ on a topic c_j for the collection $\mathcal{D}$ *as:*

$$\text{AP}(\sigma, c_j, \mathcal{D}\ ) := \int_0^1 \text{P}_\text{R}(\sigma, c_j, \mathcal{D})\ d\text{R} \tag{23}$$

In practice, this value is approximated by a discrete sum over every position in the ranked sequence of documents, as follows:

$$\text{AP}(\sigma, c_j, \mathcal{D}\ ) := \frac{1}{|\text{rel}(c_j, \mathcal{D})\,|} \sum_{k=1}^{|\text{rret}_{(\sigma,c_j,\mathcal{D})}|} \delta(\text{rret}(\sigma, c_j, \mathcal{D})_k)\ .\ \text{P@}k(\sigma, c_j, \mathcal{D}) \tag{24}$$

where

$$\delta(\text{rret}(\sigma, c_j, \mathcal{D})_k) := \begin{cases} 1 & \text{if rret}(\sigma, c_j, \mathcal{D})_k \in \text{rel}(c_j, \mathcal{D}) \\ 0 & \text{otherwise} \end{cases}$$

■

In practice, AP and R-P are highly correlated [145, 161] and show similar stability in terms of comparing systems using different queries [11]. Even though this could apparently seem surprising[22], it can be formally proved [5] that, assuming a reasonable set of assumptions, both measures approximate the area under the P_R curve, thus explaining the phenomenon. Also, we can improve stability by averaging AP across queries [63].

Definition 30 *Let σ be an* IR *system, $\mathcal{D} = \{d_i\}_{i \in I}$ be a document collection and $\mathcal{Q} = \{c_j\}_{j \in J}$ be a finite set of topics (queries). We then define the* mean average precision of σ on the topic set $\mathcal{Q}$ for the document collection $\mathcal{D}$ *as:*

$$\text{MAP}(\sigma, \mathcal{Q}, \mathcal{D}) := \frac{\sum_{j \in J} \text{AP}(\sigma, c_j, \mathcal{D})}{|\mathcal{Q}|} \tag{25}$$

■

While AP approximates the area under the P_R curve, MAP is roughly the average area under it for a set of queries. In fact, MAP is the most frequently used summary measure of a ranked retrieval run and it became standard in the TREC community. It contains both recall and precision oriented aspects; and it is also sensitive to entire ranking, providing a single-figure measure of quality across recall levels. However, MAP has the effect of attributing an equal weight to each piece of information included in the final reported number, even if many documents are relevant to some queries whereas very few are relevant to other queries. This means that a set of test information must necessarily be large and diverse enough to be representative of system effectiveness across different queries. Assuming these conditions, MAP has been shown to have especially good sensitivity and stability among evaluation measures [99]. Otherwise, when we are interested in highlighting improvements for low-performing topics, a different kind of metric becomes necessary.

Definition 31 *Let σ be an* IR *system, $\mathcal{D} = \{d_i\}_{i \in I}$ be a document collection and $\mathcal{Q} = \{c_j\}_{j \in J}$ be a finite set of topics (queries). We then define the* geometric mean average precision of σ on the topic set $\mathcal{Q}$ for the document collection $\mathcal{D}$ *as:*

$$\text{GMAP}(\sigma, \mathcal{Q}, \mathcal{D}) := \sqrt[J]{\prod_{j \in J} \text{AP}(\sigma, c_j, \mathcal{D})} \tag{26}$$

■

[22] computation of R-P considers only a single precision point while AP evaluates the area under the entire P_R curve.

Both MAP and GMAP, can be seen as different ways to reach a single quality measure through different ways of aggregating individual observations. So, while the former is the arithmetic mean of the AP in a set of queries, the latter is the geometric one. In this sense, GMAP is more indicative of effectiveness across an entire set of queries, and more robust when faced with situations in which the presence of a few well-performing queries can skew the ranking obtained by MAP. The GMAP measure was introduced by Voorhees in [159].

At this point, if we summarize the metrics described so far in terms of a single common characteristic, we would have to say that they are completely determined by the ranks of the relevant documents in the result set. They therefore make no distinction between documents that are explicitly judged as non-relevant and those that are only assumed to be non-relevant because they are unjudged, which raises a problem when QREL are known to be far from complete, it becoming advisable to mitigate this situation.

Definition 32 *Let σ be an* IR *system, $\mathcal{D} = \{d_i\}_{i\in I}$ be a document collection and $\mathcal{Q} = \{c_j\}_{j\in J}$ be a finite set of topics (queries). We then define the* binary preference relation of σ on the topic c_j for the document collection $\mathcal{D}$ *as:*

$$\text{BPREF}(\sigma, c_j, \mathcal{D}\,) := \frac{1}{\text{R}} \sum_{r\in \text{R}} [1 - \frac{|\text{nrel}(c_j, \mathcal{D}) \cap \{\text{rret}(\sigma, c_i, \mathcal{D})\}_{r+1}^{\text{R}}|}{\text{mín}\{\text{R}, |\text{nrel}(c_j, \mathcal{D})|\}}] \tag{27}$$

where $\text{R} = |\text{rel}(c_j, \mathcal{D})|$. *We can naturally extend this definition to the finite set of topics $\mathcal{Q}$.*

■

The BPREF measure, due to Buckley *et al.* [12], can be intuitively thought of as the inverse of the fraction of judged irrelevant documents that are retrieved before relevant ones. So, it computes a preference relation of whether judged relevant documents are retrieved ahead of judged irrelevant ones, i.e., it is based on the relative ranks of judged documents only. We talk about binary preference because the relation is defined from binary QREL in such a way that, for a given topic, any relevant document is preferred over any non-relevant one. In this sense, BPREF and MAP are very highly correlated when used with complete QREL. However, when these are incomplete, rankings of systems by BPREF still correlate highly to the original ones, whereas rankings by MAP do not.

A final approach that has increasingly been adopted, especially when employed with machine learning, is *cumulative gain* (CG) [99]. Normally, the original assessments provided by IR systems have multiple degrees and, in consequence, their performance should be evaluated separately at each relevance level. In this sense, highly relevant documents appearing lower in a search result list should be penalized as the graded

relevance value is reduced. Nevertheless, the ranking-dependence measures previously described are computed by using dichotomous relevance assessments, collapsing these into two for the purposes of evaluation.

Definition 33 *Let σ be an* IR *system, $\mathcal{D} = \{d_i\}_{i \in I}$ be a document collection and $\mathcal{Q} = \{c_j\}_{j \in J}$ be a finite set of topics (queries). We then define the* discounted cumulative gain *of σ on the topic c_j for the document collection $\mathcal{D}$ at a rank position $r \in [1, R] \cap \mathbb{N}$ as:*

$$\text{DCG}(\sigma, c_j, \mathcal{D})_r := \text{G}(\sigma, c_j, \mathcal{D})_1 + \sum_{k=2}^{r} \frac{\text{G}(\sigma, c_j, \mathcal{D})_k}{log_b(k)} \tag{28}$$

where $R = |\text{rel}(c_j, \mathcal{D})|$ and G *is the sequence of relevance values associated to the list* $\text{rret}(\sigma, c_j, \mathcal{D})$. *We can naturally extend this definition to the finite set of topics $\mathcal{Q}$.*

■

In practice, DCG uses the relevance level as a gained value measure for the ranked position associated to a document, by adding this gain progressively from the first position to the last one. A logarithmic discounting function is associated in order to progressively reduce the document value as its rank increases, but not too steeply. Usually, a logarithm to base two is used, i.e., taking $b = 2$ in Equation 28.

Since the set of retrieved documents may vary widely between different systems, in order to compare performance the normalized version of this measure uses the highest possible DCG value at each position.

Definition 34 *Let $\sigma = \{\sigma_i\}_{i \in I}$ be a collection of* IR *systems, $\mathcal{D} = \{d_j\}_{j \in J}$ be a document collection, $\mathcal{Q} = \{c_k\}_{k \in K}$ be a finite set of topics (queries) and $\{\text{DCG}(\sigma, c_k, \mathcal{D})_l\}_{l \in L}$ be the sequence (ordered set) of* DCG *values on the topic c_k. We then define the* normalized discounted cumulative gain *of σ_i on the topic c_k for the document collection $\mathcal{D}$ at a rank position $r \in [1, R] \cap \mathbb{N},\ R = |rel(c_k, \mathcal{D})|$ as:*

$$\text{NDCG}(\sigma_i, c_k, \mathcal{D})_r := \frac{\text{DCG}(\sigma_i, c_k, \mathcal{D})_r}{\text{IDCG}(\sigma_i, c_k, \mathcal{D})_r} \tag{29}$$

where IDCG *refers to the* ideal DCG *defined as the maximum achievable* DCG *value at rank r. It is easily found by calculating the* DCG *of a ranked list that places all the highest-graded documents above all the second ones and so on. We can naturally extend this definition to the finite set of topics $\mathcal{Q}$.*

■

Obviously, in a perfect ranking algorithm associated to an IR system, the corresponding values for NDCG will be equal to 1. The DCG and NDCG metrics were

both introduced by Järvelin and Kekäläinen in [72]. Results indicate strong correlation between user satisfaction, CG and precision; moderate correlation with DCG; and a possible surprisingly negligible correlation with NDCG [1].

6.2 | Ranking IR systems using PQREL

Introduced by Soboroff *et al.* in [135], this technique simply retakes the official TREC evaluation process [156], modifying certain aspects concerning trained human assessment. More exactly, we consider the following steps, as described by the authors:

1. A group of 50 topics are selected following the proposal of a trustworthy group of experts, usually the NIST[23] organization.

2. A number of runs, associated to each IR system evaluated, is submitted for evaluation. Each run consists of the top (at most) 1.000,00 retrieved documents for each topic. A subset of the runs from each participant are labeled *official runs*.

3. The group of experts takes the top n-documents per topic from each official run to form the *pool* for that topic. Duplicates are removed from the pool.

4. Using a model for how relevant documents occur in the pool, select a set of documents randomly to form PQREL.

5. From the set of PQREL, evaluate all the runs using the `trec_eval` package[24].

That is, with respect to TREC, Soboroff *et al.* take in step 3 $n = 10$ or $n = 100$, while TREC considers only $n = 10$. They also replace in step 4 the role of experts by a random choice. Finally, in steps 4 and 5, they consider PQREL instead of QREL. Obviously, we can consider all the measures previously described for the previous QREL-based one when estimating this kind of ranking.

6.3 | Ranking IR systems using machine-based assessment

Due to Mizzaro *et al.* [101], this technique takes as its basis the estimation of topic ease in order to consider that if a search engine seeks to perform well it should be effective on difficult topics. We shall baptize this property associated with an IR system as its *authority*, and before formalizing it we need to introduce some concepts capturing and formalizing topic ease and system effectiveness. The starting point for this methodology is the notion of AP, the computation of which can be done from both QREL and PQREL.

[23] from *National Institute of Standards and Technology*.
[24] see `http://trec.nist.gov/trec eval/`.

Definition 35 *Let $\sigma = \{\sigma_i\}_{i\in I}$ be a set of* IR *systems, $\mathcal{D}$ be a document collection and $\mathcal{Q} = \{c_j\}_{j\in J}$ be a topic (query). We then define the* average average precision of the set of IR systems σ on the topic c_j for the document collection $\mathcal{D}$*, as:*

$$\text{AAP}(\sigma, c_j, \mathcal{D}) := \frac{\sum_{i\in I} \text{AP}(\sigma_i, c_j, \mathcal{D})}{|\sigma|} \tag{30}$$

■

Intuitively, AAP is an indicator of the ease associated with topic satisfaction, understood as a magnitude directly related with the number of IR systems having good performance on that topic. From the basis offered by the AAP measure, Mizzaro *et al.* [101] extend the AP concept in order to obtain a reliable guideline for estimating the performance of an IR system on individual topics. The idea consists, initially, of normalizing AP in order to remove the influence of single topic ease (resp. single system effectiveness) on AP to obtain a reliable measure of performance on a set of IR systems (resp. on topic ease).

Definition 36 *Let $\sigma = \{\sigma_i\}_{i\in I}$ be a set of* IR *systems, $\mathcal{D}$ be a document collection and $\mathcal{Q} = \{c_j\}_{j\in J}$ be a set of topics (queries). We then define the* normalized average precision of σ_i on the topic c_j according to $\text{AAP}(\sigma, c_j, \mathcal{D})$*, as:*

$$\text{NAP}_{\text{AAP}}(\sigma_i, c_j, \mathcal{D}) := \text{AP}(\sigma_i, c_j, \mathcal{D}) - \text{AAP}(\sigma, c_j, \mathcal{D}) \tag{31}$$

■

In this way, the adjacency matrix $[\text{NAP}_{\text{AAP}}(\sigma_i, c_j, \mathcal{D})]_{(i,j)\in I\times J}$ can be interpreted as a weighted bipartite graph, where the weight on arcs $c_j \to \sigma_i$ corresponds to the values for $\text{NAP}_{\text{AAP}}(\sigma_i, c_j, \mathcal{D})$, reflecting now the individual performance of σ_i on the topic c_j and eliminating the deviations due to topic ease. The NAP_{AAP} measure was introduced by Wu and McClean in [163], and is averaged by Mizzaro in [100] across the queries in pursuit of greater stability.

Definition 37 *Let $\sigma = \{\sigma_i\}_{i\in I}$ be a set of* IR *systems, $\mathcal{D}$ be a document collection and $\mathcal{Q} = \{c_j\}_{j\in J}$ be a finite set of topics (queries). We then define the* normalized mean average precision of σ_i on the topic set $\mathcal{Q}$ for the document collection $\mathcal{D}$*, as:*

$$\text{NMAP}(\sigma_i, \mathcal{Q}, \mathcal{D}) := \frac{\sum_{j\in J} \text{NAP}_{\text{AAP}}(\sigma_i, \mathcal{Q}_j, \mathcal{D})}{|\mathcal{Q}|} \tag{32}$$

■

Surprisingly, NMAP shows a behavior that differs somewhat from TREC results, and also provides quite a different ranking in relation to MAP, although both measures are related[25]. In practice, what is generally considered an improved version of a system using TREC criteria[26] would often turn out not to be an improvement at all when using NMAP.

An alternative that can be used to take advantage of the information in the NAP_{AAP} adjacency matrix consists of analyzing it on the basis of the Kleinberg's *hits algorithm* [82] in order to obtain more sophisticated evaluation measures, taking into account the whole sets considered for both IR systems and topics. The basic idea proposed by Mizzaro *et al.* consists of using the indicators described by Kleinberg for locating high-quality information related to link structures: *hubness* and *authority*.

Definition 38 *Let $\sigma = \{\sigma_i\}_{i\in I}$ be a set of* IR *systems, $\mathcal{D}$ be a document collection and $\mathcal{Q} = \{c_j\}_{j\in J}$ be a set of topics (queries). We then define the* authority of the IR system σ_i on topics $\mathcal{Q}$ *(resp. the* hubness of the topic c_j on IR systems σ*)* for the document collection $\mathcal{D}$*, as:*

$$\text{A}(\sigma_i, \mathcal{Q}, \mathcal{D}\,) := \sum_{j\in J} \text{H}(c_j, \sigma, \mathcal{D}\,) \cdot \text{NAP}_{\text{AAP}}(\sigma_i, c_j, \mathcal{D}\,) \tag{33}$$

$$(\textit{resp.}\ \text{H}(c_j, \sigma, \mathcal{D}\,) := \sum_{i\in I} \text{A}(\sigma_i, \mathcal{Q}, \mathcal{D}\,) \cdot \text{NAP}_{\text{AAP}}(\sigma_i, c_j, \mathcal{D})) \tag{34}$$

■

Intuitively, an IR system has high authority if it is more effective on topics with high hubness, that is, on difficult topics. This provides a simple ranking criterion, since a system that wants to be effective should then show high associated authority values.

6.4 | Ranking IR systems using weighted reference counts

Described by Wu *et al.* in [164], this proposal applies a data fusion technique that compares the results obtained from a given search engine with those taken from a collection of other different IR systems. This means having previously to introduce a certain number of concepts.

Definition 39 *Let $\sigma = \{\sigma_i\}_{i\in I}$ be a set of* IR *systems, $\mathcal{D} = \{d_j\}_{j\in J}$ be a document collection, and $\mathcal{Q} = \{c_k\}_{k\in K}$ be a set of topics (queries). We denote by*

$$\text{RC}(\sigma_i, c_k, \mathcal{D}\,) := \sum_{j_i\in J_i} o(\text{rret}(\sigma_i, c_k, \mathcal{D})_{j_i}) \tag{35}$$

[25] the Kendall's tau correlation [80] is $0,87$ and the linear correlation [19] is $0,92$.
[26] i.e., a version with a higher MAP.

the reference count of σ_i on the topic c_k for the document collection $\mathcal{D}$, *where* $o(\text{rret}(\sigma_i, c_k, \mathcal{D})_{j_i})$ *is the number of occurrences of a document* $\text{rret}(\sigma_i, c_k, \mathcal{D})_{j_i}$ *in the list* $\{\text{rret}(\sigma_l, c_k, \mathcal{D})\}_{j_l \in J_l, l \neq i}$.

Given $o(\text{rret}(\sigma_i, c_k, \mathcal{D})_{j_i})$, *we baptize* $\text{rret}(\sigma_i, c_k, \mathcal{D})_{j_i}$ *as the* original document *and the counterparts in* $\{\text{rret}(\sigma_l, c_k, \mathcal{D})\}_{j_l \in J_l, l \neq i}$ *as its* reference documents *denoted by* $\gamma(\text{rret}(\sigma_i, c_k, \mathcal{D})_{j_i})$.

■

Intuitively, given a query and a certain number of top original documents returned by a particular IR system in a given collection, its RC is the addition of the reference ones provided by the other systems. This inspires a simple ranking method independently of the consideration of QREL, which Wu *et al.* call *Basic*.

Definition 40 *Let* $\sigma = \{\sigma_i\}_{i \in I}$ *be a set of* IR *systems,* $\mathcal{D}$ *be a document collection, and* $\mathcal{Q} = \{c_j\}_{j \in J}$ *be a set of topics (queries). We denote by*

$$\text{ARC}(\sigma_i, \mathcal{Q}, \mathcal{D}) := \frac{\sum_{j \in J} \text{CR}(\sigma_i, c_j, \mathcal{D})}{|\mathcal{Q}|} \tag{36}$$

the average reference count of σ_i on the topic set $\mathcal{Q}$ for the document collection $\mathcal{D}$.

■

Intuitively, given an IR system, we compute its ARC as the average value of the individual values for RC on each topic; which provides a reliable ranking technique for IR systems. Amongst the refinements proposed by the authors to this Basic method, we chose to take into account the relevance position of both original and reference documents. This makes it necessary to extend the RC notion in order to integrate them.

Definition 41 *Let* $\sigma = \{\sigma_i\}_{i \in I}$ *be a set of* IR *systems,* $\mathcal{D}$ *be a document collection,* $\mathcal{Q} = \{c_j\}_{j \in J}$ *be a set of topics (queries), and* $\{\varrho_{j_i}\}_{j_i \in J_i}$ *be the normalized scores*[27] *associated to* $\{\text{rret}(\sigma_i, c_j, \mathcal{D})\}_{j_i \in J_i}$. *Let also* $\forall \ m \in [1, \text{MaxNumDocs}], \ k \in [1, 4],$ MaxNumDocs=1.000*:*

$$\hat{o}(\text{rret}(\sigma_i, c_j, \mathcal{D})_{j_i}) := \sum_{\text{rret}(\sigma_k, c_j, \mathcal{D})_{k_l} \in \gamma(\text{rret}(\sigma_i, c_j, \mathcal{D})_{j_i})} \Delta - l \ \ (\text{resp. } \varrho_{k_l})$$

and

$$\omega_{j_i} := \begin{cases} \zeta(200) - \zeta(m-1) \ , & \text{if} \quad j_i = 5m \\ \omega_{5m} - \frac{1}{m} + \frac{5}{j_i} \ , & \text{if} \quad j_i = 5m - k \end{cases}$$

[27] we assume, without loss of generalization, that these scores are in the interval $[0, 1]$.

being $\hat{o}(\text{rret}(\sigma_i, c_j, \mathcal{D})_{j_i})$ *and* ω_{j_i} *the weight functions associated to the relevance positions of reference and original documents, respectively, with the auxiliary function* ζ *defined by*

$$\zeta(m) := \begin{cases} 0\,, & \text{if } m = 0 \\ 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{m}\,, & \text{otherwise} \end{cases}$$

where MaxNumDocs *is the maximum size of the document collection* $\mathcal{D}$ *and* Δ *is a constant value, which the authors set empirically to 1.501,00 in their experiments. We denote the expression*

$$\sum_{j_i \in J_i} \omega_{j_i}.\hat{o}(\text{rret}(\sigma_i, c_j, \mathcal{D})_{j_i}) \tag{37}$$

by $\text{WRC}_\text{o}(\sigma_i, c_j, \mathcal{D})$ *(resp. by* $\text{WRC}_\text{s}(\sigma_i, c_j, \mathcal{D})$*), baptizing it the* ordering-based *(resp.* scoring-based*)* weighted reference count of σ_i on the topic c_j for the document collection $\mathcal{D}$.

■

Following the same process applied to introduce ARC from RC, we can now naturally introduce the *average weighted reference count*, AWRC_o (resp. AWRC_s) from WRC_o (resp. WRC_s), which provides two additional ranking measures.

However, it is difficult to justify some choices in this ranking proposal since no convincing reasons are argumented to introduce the constant Δ nor (very complex) values ω_{j_i}. As result formulae are also unclear and difficult to understand. In this sense, we propose to lightly modify the original proposal.

Definition 42 *Let* $\{\sigma_i\}_{i \in I}$ *be a set of* IR *systems,* $\mathcal{D}$ *be a document collection,* $\mathcal{Q} = \{c_j\}_{j \in J}$ *be a set of topics (queries), and* $\{\varrho_{j_i}\}_{j_i \in J_i}$ *be the normalized scores*[28] *associated to* $\{\text{rret}(\sigma_i, c_j, \mathcal{D})\}_{j_i \in J_i}$*. Let also* $\forall\ m, n \in [1, |\mathcal{D}|]$*:*

$$\hat{o}(\text{rret}(\sigma_i, c_j, \mathcal{D})_{j_i}) := \sum_{\text{rret}(\sigma_k, c_j, \mathcal{D})_{k_l} \in \gamma(\text{rret}(\sigma_i, c_j, \mathcal{D})_{j_i})} \omega_{k_l} \ \textit{(resp. } \varrho_{k_l}\textit{)},\ \textit{where}\ \omega_{k_l} := \begin{cases} 1 & \text{if } l = 1 \\ \frac{1}{log_b(l)} & \text{otherwise} \end{cases}$$

and

$$\hat{\omega}_{j_i} := \begin{cases} 1 & \text{if } j_i = 1 \\ \frac{1}{log_b(j_i)} & \text{otherwise} \end{cases}$$

be the weigth functions associated to the relevance positions of reference and original documents, respectively. We denote the expression

$$\sum_{j_i \in J_i} \hat{\omega}_{j_i}.\hat{o}(\text{rret}(\sigma_i, c_j, \mathcal{D})_{j_i}) \tag{38}$$

[28] we assume, without loss of generalization, that these scores are in the interval $[0, 1]$.

by $\text{WRC}_{\text{LO}}(\sigma_i, c_j, \mathcal{D})$ *(resp. by* $\text{WRC}_{\text{LS}}(\sigma_i, c_j, \mathcal{D})$*), baptizing it* logaritmic ordering-based *(resp. the* scoring-based*)* weighted reference count of σ_i on the topic c_j for the collection $\mathcal{D}$.

■

Following the same process applied to introduce AWRC_{O} (resp. AWRC_{S}), we can now define AWRC_{LO} (resp. AWRC_{LS}), which provides the ranking measures using weighted reference counts we shall definitively consider in this work. We take $b = 2$.

6.5 | Selecting the topic set

The goal now is to select a minimal set of topics in order to evaluate our IR system in comparison with a collection of existing ones, taking as reference different levels of difficulty in the solving of user queries. As a general overview, we consider a stratified sampling technique to select an *initial set of topics*, on which we shall later apply a minimization technique in order to reduce its size without loss of discrimination power. This will allow us to significantly simplify the testing task that here is especially complex, since we not only aim to estimate the efficiency of the IR system, but also to identify those factors impacting it in terms of vagueness and incompleteness. Given that, to the best of our knowledge, no specific techniques have been described for this concrete purpose, our approach is in the nature of a proposal.

6.5.1 | The size of the initial sample

One fundamental question consists in determining the size of the topic set we shall use to evaluate the proposal, for which purpose we take as reference the discussion put forward in this sense by Guiver *et al.* [62], which in turn derives from previous works. In this regard, the authors reveal an evident evolution of the state-of-the-art, attributing the first estimations to Jones and van Rijsbergen [140], who concluded that 75 were not enough, 250 were usually acceptable, and even 1.000 could be needed. They later refer to Zobel's work [172] supporting the idea that a set of 25 topics did a reasonable job, while Buckley and Voorhees [11] give the first effective evidence that the number of queries needed for a good experiment is at least 25, although 50 seems to be better. More recently, in the context of TREC-like evaluations Webber *et al.* [162] claim that 150 topics are required to make a reliable distinction between IR systems, although usually only 50 are considered [156]. In our case, we shall first select an initial sample with 150 topics.

6.5.2 | The sampling process

To begin with, we classify our sample space[29] (population) following two independent criteria, each one conforming its own partition, and which we believe can be correlated

[29] formed by the complete set of possible topics to be applied on our running *corpus* $\mathscr{B}$.

with the intuitive notion of difficulty in solving topics. This latter constitutes our desired dependent variable for sampling, a choice based on Mizzaro *et al.* [101] that suggests it as a major factor in topics when seeking to efficiently discriminate between IR systems. In practice, we succinctly introduce these criteria by their associated variables:

- The *specificity of the query*, understood as the level of detail with which the user expresses it. We consider three different levels: high, medium and low.
- The *type of answer* returned by a search engine following a conceptual approach: approximate, plausible and partial. We assume here that a topic has a given type when the set of answers of that class within the first 10 returned by the system[30] has a greater estimable weight than that corresponding to the other types. Thus, we need to fix the ratio μ_j (resp. μ_a) limiting the number of joins (resp. aggregations) associated to plausible (resp. partial) answers, as well as formally compute such a weight.

These criteria also allow us to combine points of view from both the user and the IR system. In order to balance the sample that will serve as an initial set of topics, we should minimize (resp. maximize) variability within (resp. between) subpopulations (strata) corresponding to different partitions. So, we distribute the sample between the three sub-populations introduced for each one of these[31], which provides homogeneity in all levels of that stratification. Also, topics in a given stratum of a partition are equitably shared between the strata of the other one. We thereby ensure that the probability of one of the topics in the sample having a given type of answer and specificity will be approximately the same, regardless of the combination considered for the variables in question. In this way, we expect to improve the accuracy and efficiency of estimation, draw inferences about those sub-populations and permit a greater balancing of statistical power of tests of differences between partitions by sampling equal numbers from them, varying widely in size. To attain this goal we employ a careful selection process.

In regard to query specificity, we take as our starting point a collection of topics proposed by human experts and distributed over three strata, in such a way that the queries in each stratum are obtained by refining the content of those from the previous one. The objective is to integrate, in similar number, topics with high, medium and low specificity. More in detail, we consider an initial collection of topics

$$\mathcal{Q} := \{\mathcal{Q}_i^{hs}\}_{i\in I} \cup \{\mathcal{Q}_i^{ms}\}_{i\in I} \cup \{\mathcal{Q}_i^{ls}\}_{i\in I},\ \mathcal{Q}_i^{hs} \succ \mathcal{Q}_i^{ms} \succ \mathcal{Q}_i^{ls},\ \forall i \in I$$

where $\succeq$ is the partial order naturally induced in the sample space by the specificity detected by the experts.

[30] which corresponds approximately to the first page of results returned by any search engine, precisely the threshold above which the user ceases to show interest in looking at the answers [61].

[31] this implies that we associate 50 topics per stratum, the same number considered by the classic TREC protocol [156] for evaluating IR systems.

In regard to the type of answer, we first take value $\mu_j = 0,34$ (resp. $\mu_a = 0,18$) in order to moderate the number of plausible (resp. partial) answers returned[32], which is equivalent to sampling them with an appropriate probability.

Once this has been done, we need to introduce a criterion which will measure the weight of a given type of answers in a finite set, balancing the distribution between the types under consideration. Here, we assume that we not only have to take the number of answers of a given case into account, but also their position in the ranking. So, the type of answers appearing lower in the search result list should be penalized as the degree of relevance value is reduced. This sets us in a context comparable to the one considered when determining ranked-based evaluation measures for IR systems and, more specifically, in the process for constructing the NDCG metric, which we now take as our inspiration for introducing the notion of *discounted cumulative weight associated to a given type of answer*.

Definition 43 *Let σ be a* IR *system, $\mathcal{D} = \{d_i\}_{i \in I}$ be a document collection and $\mathcal{Q} = \{c_j\}_{j \in J}$ be a finite set of topics (queries). We then define the* discounted cumulative weight of σ on the topic c_j for a type ι of answer and document collection $\mathcal{D}$ with pool size $p \in [1, |ret(\sigma, c_j, \mathcal{D})|]$ *as:*

$$\text{DCW}(\sigma, \iota, c_j, \mathcal{D})_p := \delta_\iota^{\text{type}(\text{rret}_{(\sigma, c_j, \mathcal{D})_1})} + \sum_{k=2}^{p} \frac{\text{type}(\text{rret}(\sigma, c_j, \mathcal{D}))_k}{log_b(k)} \tag{39}$$

where type *returns the type of the answer that serves as argument, and δ_i^j is the function known as* Kronecker's delta, *which is defined as follows:*

$$\delta_i^j := \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases} \tag{40}$$

■

In our particular case, we take $p = 10$, $b = 2$ and our conceptual-based IR proposal as σ, which implies that $\iota \in \{approximate, plausible, partial\}$. In practice, the team of human experts uses DCW to attain a uniform distribution for the sample based on the type of the answer, simultaneously taking into account the criterion for specificity previously described. As result, we obtain our initial set of topics verifying all the constraints previously described from the point of view of both heterogeneity between strata in different partitions and homogeneity in all levels of each stratification. This places us at the starting point to begin with the minimization phase, which we now introduce in three steps.

[32] the number of plausible, and in particular partial, answers can artificially expand due to the fact that they are generated by applying mechanisms that can indefinitely increase the size of the BGs associated to the topics, a situation that does not occur in the case of approximate answers.

6.5.3 | Individual topic selection for an individual system

Our first approach to deal with topic selection consists of determining a strategy for estimating the suitability of an individual topic for measuring the performance of an individual IR system. In this sense, and taking the TREC experience as our source of inspiration, AP measures the effectiveness of an individual IR system σ on a similarly individual topic $c \in \mathcal{Q}$ for a document collection $\mathcal{D}$, which would appear to solve the question.

However, placing ourselves in the machine-based assessment frame, we cannot conclude that an IR system σ presents a better performance on the topic c than on topic $\tilde{c}$ (resp. that c is considered easier by σ than by $\tilde{\sigma}$), on the basis of $\text{AP}(\sigma, c, \mathcal{D}) > \text{AP}(\sigma, \tilde{c}, \mathcal{D})$ (resp. $\text{AP}(\sigma, c, \mathcal{D}) > \text{AP}(\tilde{\sigma}, c, \mathcal{D})$). It may simply be that c might be an easy topic[33] and $\tilde{c}$ a difficult one[34] (resp. σ might be a good system[35] and $\tilde{\sigma}$ a bad one[36]). This means that we have to turn our attention to the concept of NAP_{AAP} where, unlike the situation that occurs with AP, the condition $\text{NAP}_{\text{AAP}}(\sigma, c, \mathcal{D}) > \text{NAP}_{\text{AAP}}(\sigma, \tilde{c}, \mathcal{D})$ allows us to infer that an IR system σ has a good performance on the topic c and a bad one on $\tilde{c}$.

6.5.4 | Topic set selection for an individual system

Among all the TREC inspired techniques available in the state-of-the-art to solve this question, we chose to work with the one proposed by Guiver *et al.* in [62]. The starting point is now the MAP measure, in fact an indicator of the effectiveness of the IR system orienting us about its goodness, once the set of topics has been fixed for a given document collection. The idea consists of applying an exhaustive search on all possible subsets of topics within a given collection. In this way, we can focus on the highest correlation of these MAP values with that for the whole collection, in order to estimate how well a subset of topics predicts the performance of the IR system.

Alternatively, we can also here retake a similar reasoning on the machine-based assessment frame, but now using NMAP values instead of MAP ones and taking into account the fact that these two metrics do not always agree.

6.5.5 | Topic set selection for a set of systems

To the best of our knowledge, no proposals in this regard have been documented in the state-of-the-art. We place this question on both the human-based and the machine-based assessment frames, on the basis of the techniques previously introduced for individual and topic set selection for individual IR systems. However, although the steps to be applied in order to achieve this will be the same, their nature will depend at any given moment on

[33] i.e., a topic on which all or most IR systems have a good performance.
[34] i.e., a topic on which all or most IR systems have a poor performance.
[35] i.e., a system whose performance extends to all or most difficult topics.
[36] i.e., a system whose performance is close to easy topics.

the type of working frame chosen:

1. The first step consists of generating, from the sample serving as our initial set of topics, a collection of subsets with different capabilities to measure system performance in different degrees, which we baptize the *baseline topic collection*. At the top (resp. bottom) end of this scale, we consider subsets exclusively formed by difficult topics (resp. easy topics) with the highest (resp. lowest) discrimination power. Any topic not catalogued as difficult or easy is considered as a medium difficulty one. The size of each of these subsets will once again be 50, following the proposal of Webber *et al.* [162].

   Two type of collections are generated, according to the frame used to estimate the ease level of the topics. So, in the case of the human-based assessment strategy, we resort to the opinion of an expert in the domain,. On the other hand, when dealing with machine-based assessment criteria, we identify difficult topics (resp. easy topics) with the highest hubness (resp. lowest hubness) on the set of IR systems.

2. We then apply to each of these two baseline topic collections a minimization technique in order to reduce their size without appreciably affecting their discrimination power. The result will be two sets of *final topic collections*, one particularly oriented towards human-based assessment and the other to machine-based assessment, distinguishing between three levels of difficulty (high, medium and low) in each collection. To compute the former we follow the technique proposed by Guiver *et al.* in [62] on the basis of MAP measure correlation, while for the second the process will be analogous, although based on NMAP correlation[37]. Given that both MAP and NMAP can be computed from QREL or PQREL, we finally obtain four final topic collections. Two of them consider QREL (resp. pseudo-QREL) as basis for computing both MAP and NMAP, one using human-based assessment and another one applying the machine-based variety.

   The only matter still pending is to determine the composition of these final subsets, a problem for which the authors provide no clear criterion. We here choose to select those candidates whose cardinality falls within the interval $[1, 50)$, while achieving a high enough level of MAP (resp. NMAP) correlation with the associated baseline topic subset.

   In the case of the QREL-based topics, we take a level of MAP (resp. NMAP) correlation with the corresponding human-based (resp. machine-based) assessment oriented baseline topic collection that is higher than or equal to $0,99999932$. This means, in the case of the human-based (resp. machine-based) approach, considering a collection of final subsets with 12 (resp. 10 topics for the higher difficulty, 22 (resp. 15) for the medium difficulty and 32 (resp. 8) for the lower one, which we

[37] both approaches were previously described when we introduced topic set selection for individual IR systems.

baptize the *human-based (resp. machine-based) on* QREL *topic set collection* or, briefly, HBQTC (resp. MBQTC).

In the case of PQREL-based topics, we take a level of MAP (resp. NMAP) correlation with the corresponding human-based (resp. machine-based) assessment oriented baseline topic collection that is higher than or equal to $0,999990$. This means, in the case of the human-based (resp. machine-based) approach, considering a collection of final subsets with 30 (resp. 2) topics for the higher difficulty, 29 (resp. 22) for the medium difficulty and 24 (resp. 48) for the lower one, which we baptize the *human-based (resp. machine-based) on pseudo-*QREL *topic collection* or, briefly, HBPQTC (resp. MBPQTC).

Given that no topic from the initial sampling set has a greater probability of being included in the reduced final set, either on the basis of its kind of answer or its specificity, this will depend exclusively on its level of difficulty, determined by one of the methods described above. This should guarantee the objectivity and validity of the experimental results we shall later obtain using this reduced sample. However, it also seems reasonable to try to ensure that the protocol we follow to perform topic selection will provide sensitively different conclusions depending on the specific frame in which the tests are be located. In effect, previous works [100, 101] show that even an IR system that seeks to be effective in TREC needs to be effective on easy topics, with common sense indicating that a high-performance search engine should prove its real power on difficult ones.

6.6 | The set of systems

We chose a sample of four well known search engine platforms in order to serve as comparison reference values to estimate the performance efficiency of our proposal, which we have baptized COGIR:

1. ZETTAIR (see `http://www.seg.rmit.edu.au/zettair/`) is an open source search engine developed by the *Search Engine Group* at RMIT University, written in C. It has been designed for simplicity as well as speed and flexibility, and its primary feature is the ability to handle large amounts of text. This search engine supports ranked, simple (non-nested) Boolean, and phrase queries.

2. SOLR (see `lucene.apache.org/solr/`) is an open source enterprise search platform from the Apache Lucene project. As major features, it is written in JAVA and runs as a standalone full-text search server within a servlet container such as TOMCAT. It uses the Lucene JAVA search library at its core for full-text indexing and search. SOLR provides distributed search and index replication, powering the search and navigation features of many of the world's largest internet sites.

3. TERRIER[38] (see `http://ir.dcs.gla.ac.uk/terrier/`) is a highly flexible, efficient, and effective open source search engine, readily deployable on large-scale

[38] from TERabyte RetrIEveR

collections of documents and developed at the University of Glasgow. It is written in JAVA and provides multiple indexing strategies, such as multi-pass, single-pass and large-scale MapReduce indexing.

4. INDRI (see `http://www.lemurproject.org/indri/`) is an open source search engine for large-scale search, written in C++. It is built on top of the LEMUR project (see http://www.lemurproject.org/), which is a toolkit designed for research in language modeling and information retrieval. This project was developed as a cooperative work between the Universities of Massachusetts and Carnegie Mellon.

These search engines provide a fan-shaped coverage of some of the most popular current search engines on offer, including both different implementation languages and different search models.

7 | Experimental results

Once the testing frame has been formalized, we only have to input, visualize and interpret the results, taking into account that the simplest way to compare different IR systems is to sort them by decreasing values, according to the different effectiveness metrics. In this regard, we now follow the same order previously considered to introduce those metrics, according to their type.

7.1 | Ranking IR systems using QREL

We consider the complete set of different effectiveness metrics (fourteen in total) previously introduced for experimental purposes at this level, which should be sufficient to detect any possible malfunction in our proposal, whilst also guaranteeing the robustness of the evaluation. The tests are performed on two topic set collections, HBQTC and MBQTC, seeking to fit the criterion for topic selection to the specific ranking approach, both of which are based on QREL. This should lend reliability to the process.

7.1.1 | Using the human-based assessment topic sets collection

We take here HBQTC as our topic collection, which will provide a general view of the behavior of our proposal in dealing with QREL-based ranking on topics selected using human assessment. This should constitute a well-founded evaluation protocol.

Set-based evaluation measures

We deal here with results for P and R measures, shown in Figs. 2.3 and 2.4 respectively. As can be seen, in both cases the results show greater precision from the conceptual model (COGIR) than from the others, as well as a greater control of coverage.

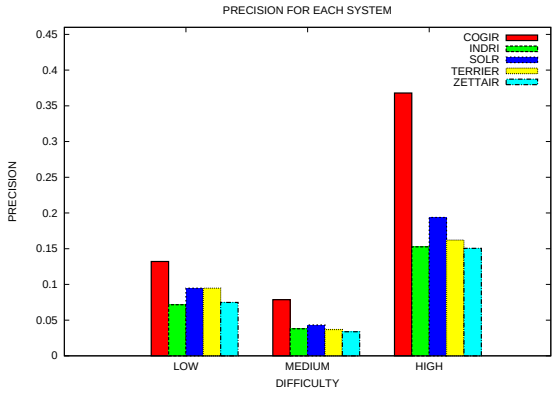

Figure 2.3: P on HBQTC using QREL

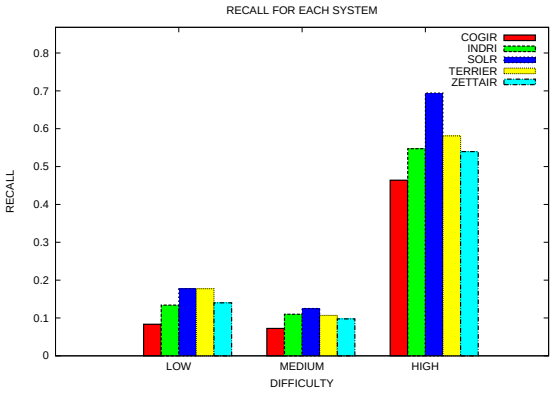

Figure 2.4: R on HBQTC using QREL

We also include tests for F and FO metrics, in order to take into account the proportion of non-relevant documents recovered. The associated plots are shown in Figs. 2.5 and 2.6, respectively. The values clearly favor the conceptual model over the rest in the set of topics of the highest difficulty, i.e., in those with the greatest capacity to discriminate between systems as far as evaluation is concerned. The results are less striking in the case of topics with a lesser discriminatory capacity.

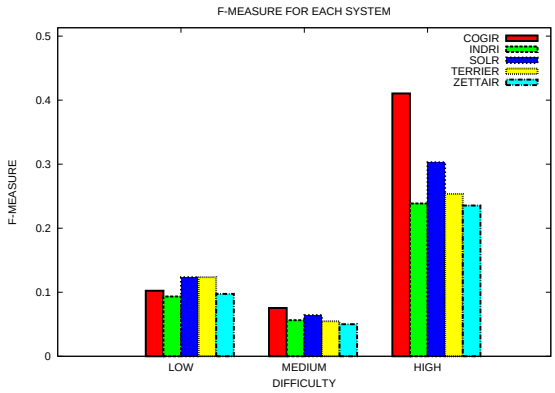

Figure 2.5: F on HBQTC using QREL

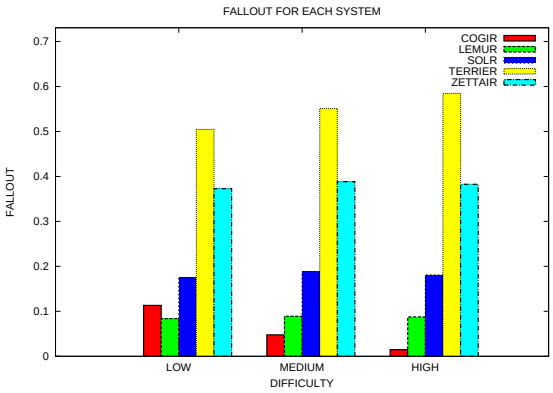

Figure 2.6: FO on HBQTC using QREL

Rank-based evaluation measures

We deal here with results for P@10 and R@10 measures, shown in Figs. 2.7 and 2.8 respectively. As can be seen, in both cases the results show greater precision from the conceptual model (COGIR) than from the others, as well as keeping recall within reasonable bounds.

In order to study the possible extension of the first page of results to the complete set obtained, we calculated IP_R at recall level 0 (resp. 0,10) in Fig. 2.9 (resp. in Fig 2.10). Once again, as in the previous cases, the conceptual model clearly showed a better performance in the case of high difficulty topics.

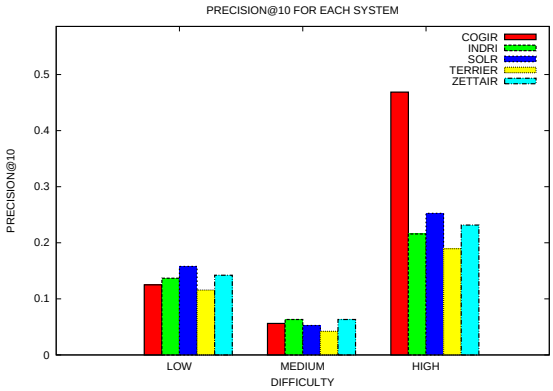

Figure 2.7: P@10 on HBQTC using QREL

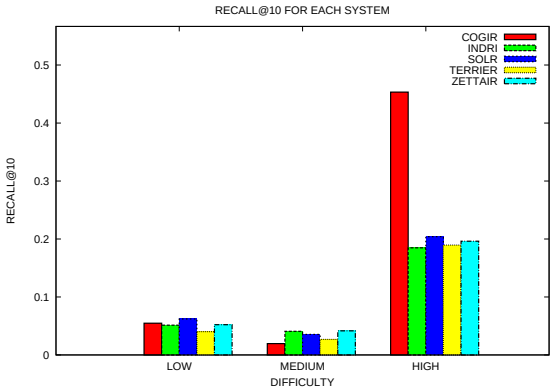

Figure 2.8: R@10 on HBQTC using QREL

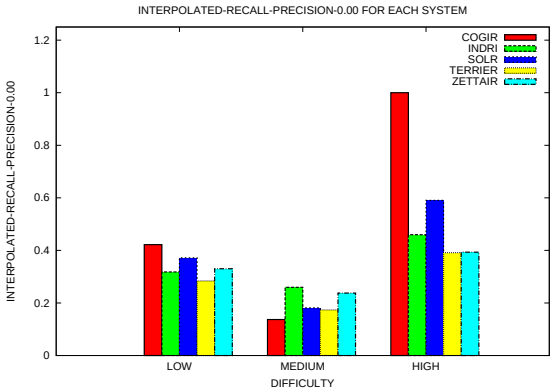

Figure 2.9: $IP_{R=0,00}$ on HBQTC using QREL

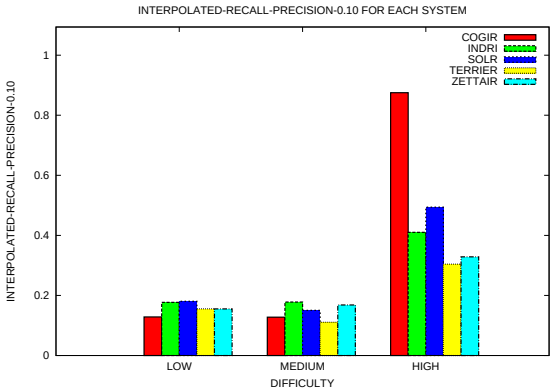

Figure 2.10: $IP_{R=0,10}$ on HBQTC using QREL

For their part, Figs. 2.11 and 2.12 again support the robustness of the conceptual model, this time on the basis of the R-P and MAP metrics. At the same time, these results highlight its performance in dealing with queries of the highest level of difficulty, whilst maintaining a comparable performance to the rest of the systems evaluated when dealing with the other levels of difficulty.

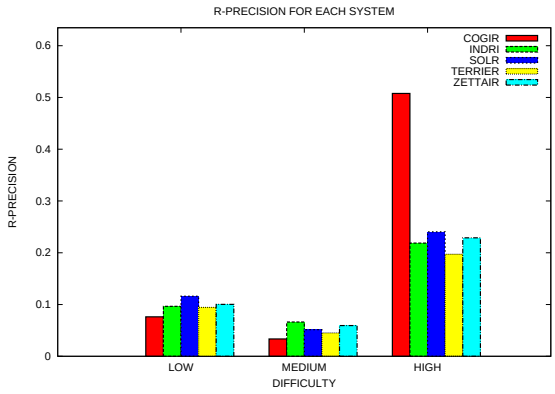

Figure 2.11: R-P on HBQTC using QREL

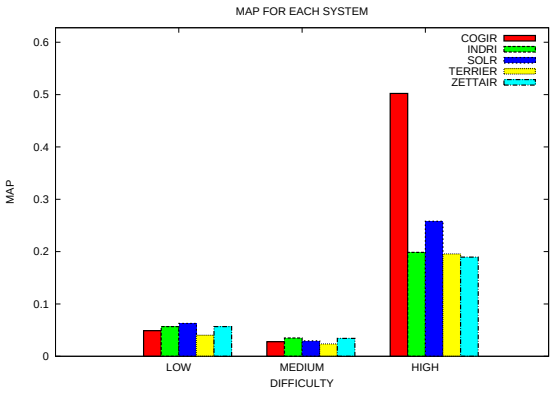

Figure 2.12: MAP on HBQTC using QREL

The values obtained for GMAP and BPREF are shown in Figs. 2.13 and 2.14. The same behavior as before is once again repeated, as shown by the similar level of performance obtained by all the systems considered when dealing with queries of low or medium difficulty, whilst the results of the conceptual model were much better when these were of high difficulty.

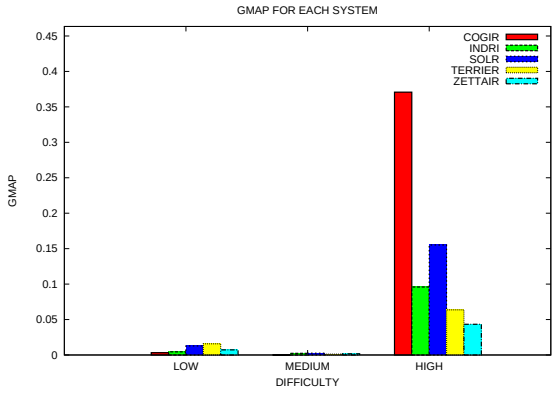

Figure 2.13: GMAP on HBQTC using QREL

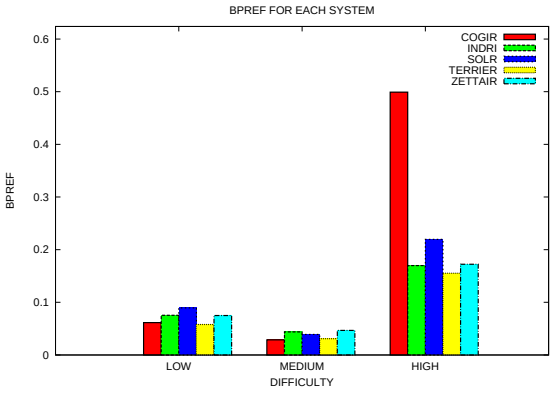

Figure 2.14: BPREF on HBQTC using QREL

Finally, we present the values for DCG and NDCG in Figs. 2.15 and 2.16, respectively. Unlike all the preceding cases, here the results are clearly better for the conceptual model when dealing with low difficulty queries, whilst in the others its performance is comparable to that of the other systems. This comes as no surprise, since various authors [1] have warned of the potentially surprising results as far as correlation with the above-mentioned metrics is concerned.

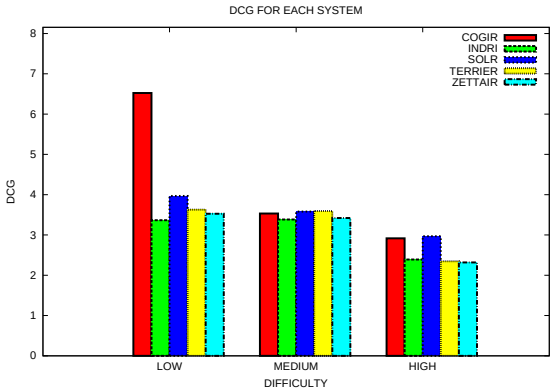

Figure 2.15: DCG on HBQTC using QREL

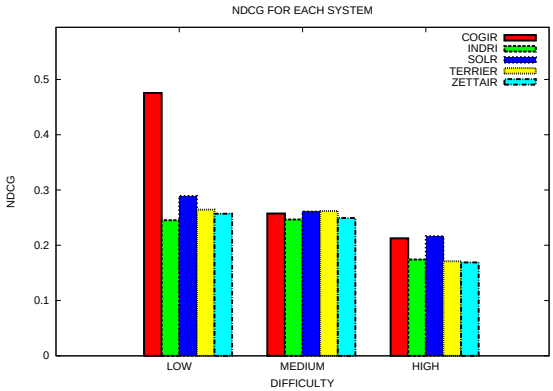

Figure 2.16: NDCG on HBQTC using QREL

7.1.2 | Using the machine-based assessment topic sets collection

We now compute the same set of previous measures on the topic set MBQTC. In this case, the value of the experiment consists of corroborating the conclusions previously attained.

Set-based evaluation measures

We once again compute for P, R, F and FO measures, whose plots are shown in Figs. 2.17, 2.18, 2.19 and 2.20, respectively. In all these cases, the conceptual approach performs worse for the set of high difficulty topics in comparison to the medium and low difficulty ones, although it still achieves the best results for P and F, while performing quite well for the other two measures.

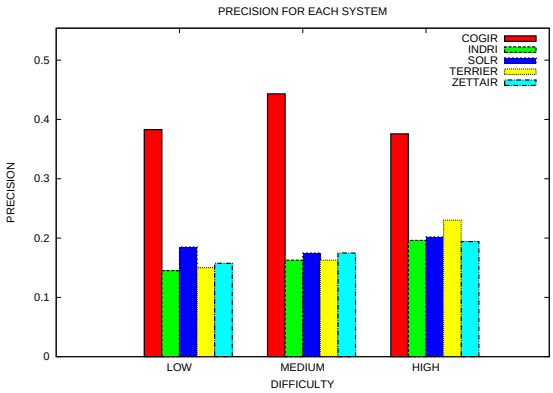

Figure 2.17: P on MBQTC using QREL

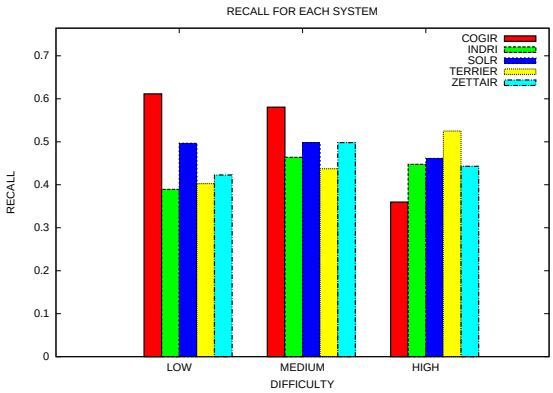

Figure 2.18: R on MBQTC using QREL

Rank-based evaluation measures

We now re-calculate the P@10 and R@10, IP_R for recall levels 0 and 0,10, R-P, MAP, GMAP, BPREF, DCG and NDCG in Figs. 2.21, 2.22, 2.23, 2.24, 2.25, 2.26, 2.27, 2.28, 2.29

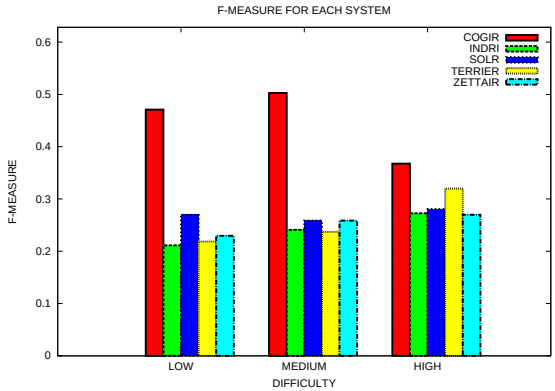

Figure 2.19: F on MBQTC using QREL

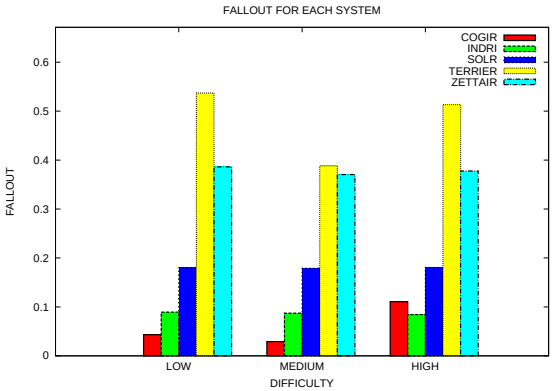

Figure 2.20: FO on MBQTC using QREL

and 2.30, respectively. The figures show that COGIR achieves better results than the other IR systems for all sets of topics. However, in contrast to those obtained in the case of HBQTC, it performs worse for the set of high difficulty topics than for the medium and low difficulty ones.

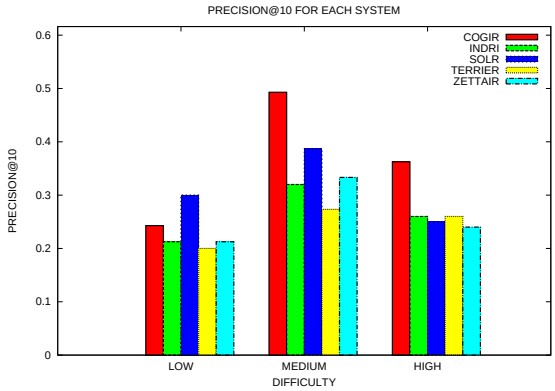

Figure 2.21: P@10 on MBQTC using QREL

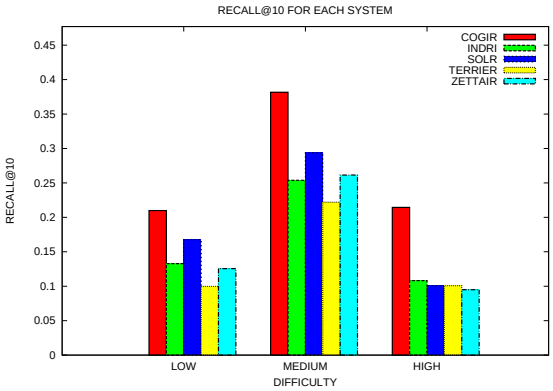

Figure 2.22: R@10 on MBQTC using QREL

7.2 | Ranking IR systems with PQREL

We follow here the same protocol applied to the QREL-oriented ranking, considering the complete set of difference effectiveness metrics (fourteen in total) used in the previous experiments. The only difference is the pair of topic sets we now use, replacing HBQTC (resp. MBQTC) by HBPQTC (resp. MBPQTC), looking to fit the criterion for topic selection to the specific ranking approach, in both cases based on PQREL.

7.2.1 | Using the human-based assessment topic sets collection

We take here HBPQTC as topic collection, which should serve us to provide a general view of our proposal in dealing with PQREL-based ranking on topics selected using human assessment.

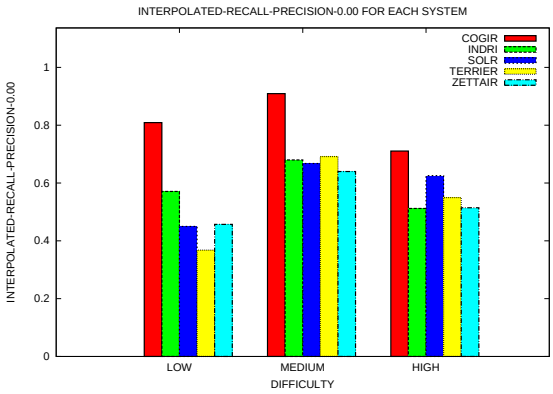

Figure 2.23: $\text{IP}_{\text{R}=0,00}$ on MBQTC using QREL

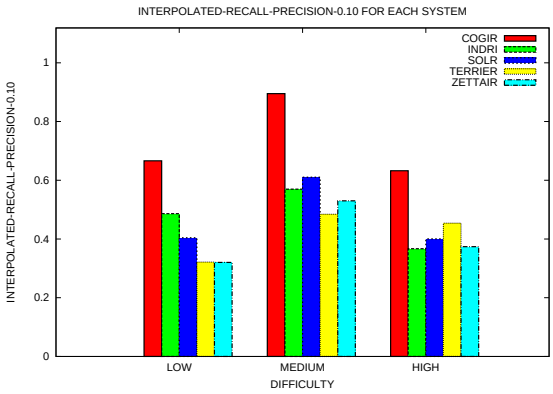

Figure 2.24: $\text{IP}_{\text{R}=0,10}$ on MBQTC using QREL

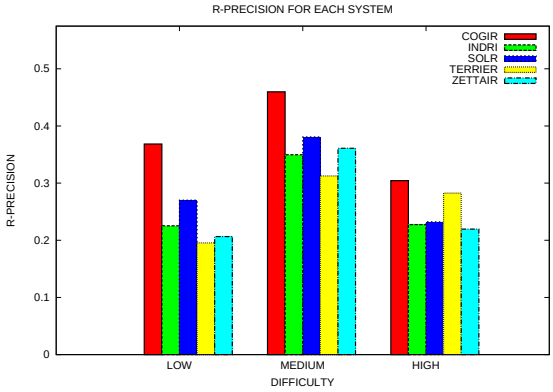

Figure 2.25: R-P on MBQTC using QREL

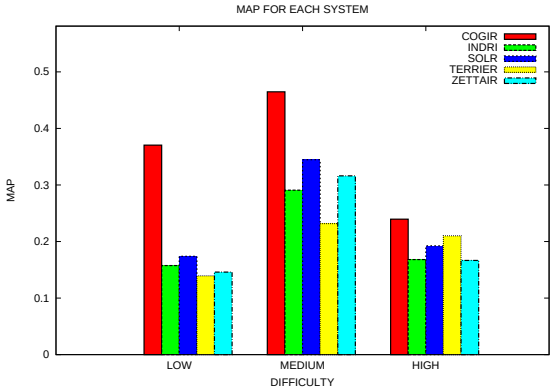

Figure 2.26: MAP on MBQTC using QREL

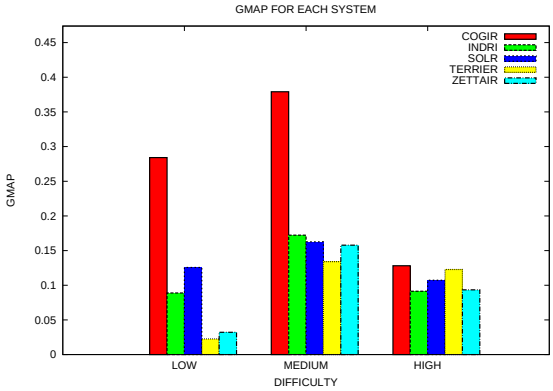

Figure 2.27: GMAP on MBQTC using QREL

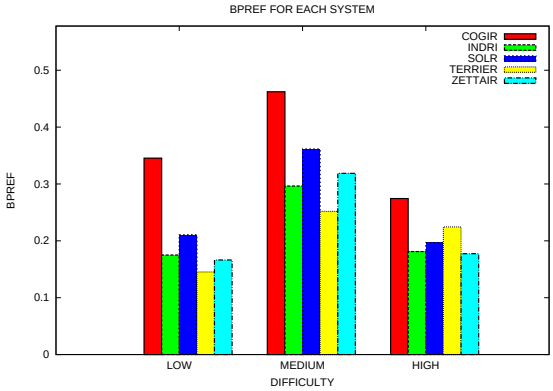

Figure 2.28: BPREF on MBQTC using QREL

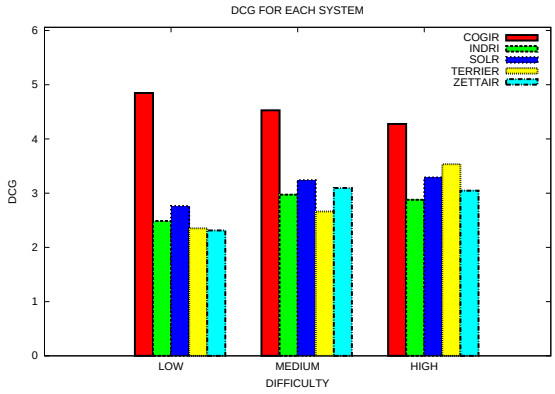

Figure 2.29: DCG on MBQTC using QREL

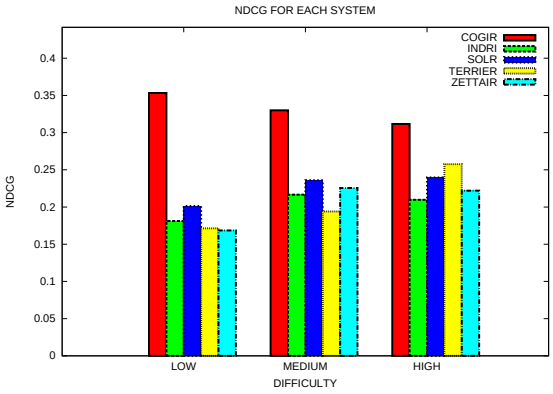

Figure 2.30: NDCG on MBQTC using QREL

Set-based evaluation measures

We deal here with results for P and R measures, shown in Figs. 2.31 and 2.32 respectively. The results obtained are practically a carbon copy of those obtained in the case of QREL, once again revealing that COGIR gives the best performance with regard to precision, whilst keeping recall within reasonable bounds.

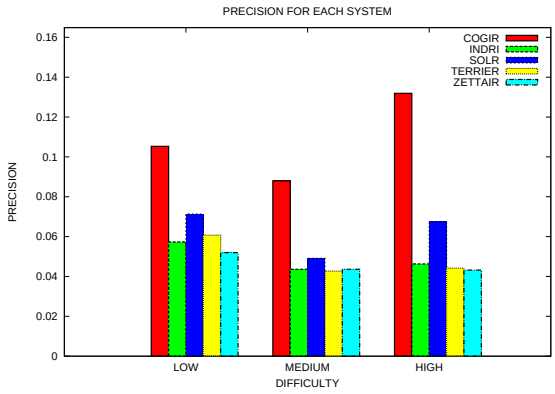

Figure 2.31: P on HBPQTC using PQREL

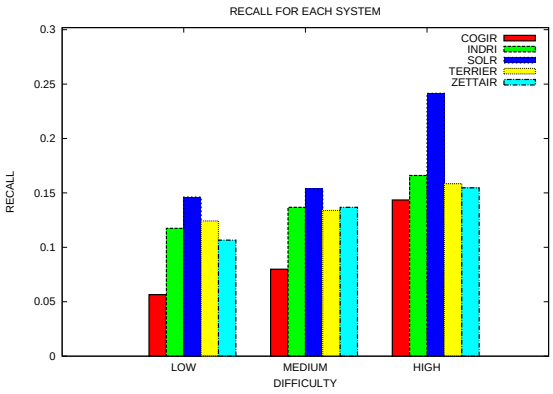

Figure 2.32: R on HBPQTC using PQREL

As previously for QREL, we also include the results for F and FO metrics in Figs. 2.33 and 2.34, respectively. Once again, the conceptual model gives improved results with the set of high difficulty topics, but these are less spectacular in the case of topics with less discriminatory capacity.

Rank-based evaluation measures

We calculate the P@10, R@10, IP_R for recall levels 0 and 0,10, R-P, MAP, GMAP, BPREF, DCG and NDCG in Figs. 2.35, 2.36, 2.37, 2.38, 2.39, 2.40, 2.41, 2.42, 2.43 and 2.44, respectively. The results obtained illustrate COGIR's stable performance, when compared to the other search engines, in handling topics of the highest degree of difficulty.

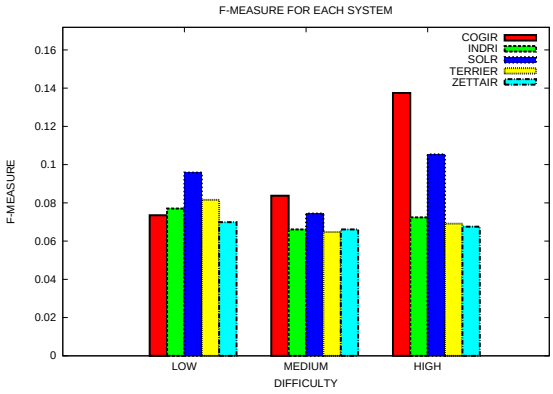

Figure 2.33: F on HBPQTC using PQREL

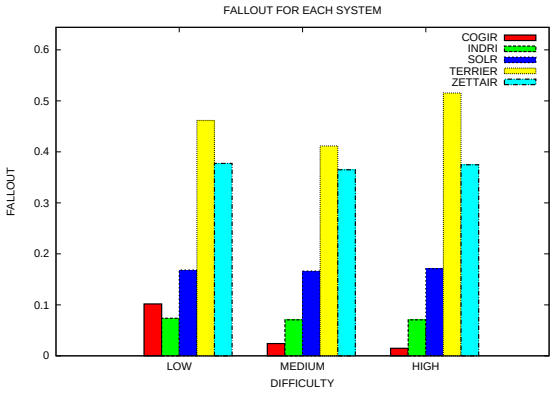

Figure 2.34: FO on HBPQTC using PQREL

In this sense, the use of PQREL tends to favor the other systems because they all of them share the same theoretical model, and thus produce similar result lists. This is to their benefit, given that PQREL are calculated from such result lists.

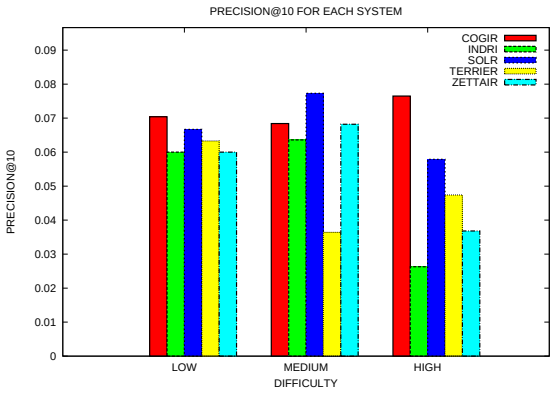

Figure 2.35: P@10 on HBPQTC using PQREL

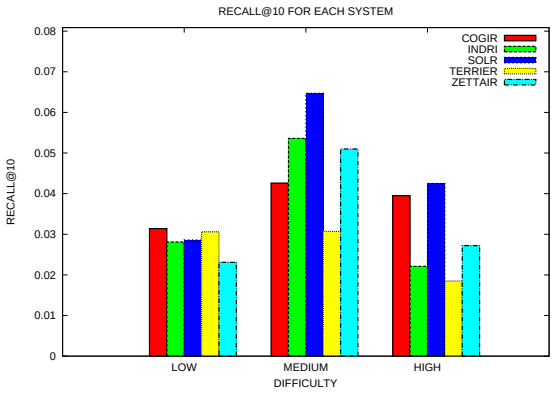

Figure 2.36: R@10 on HBPQTC using PQREL

7.2.2 | Using the machine-based assessment topic sets collection

We now compute the same set of previous measures on the topic set MBPQTC. In this case, the value of the experiment consists of corroborating the conclusions previously reached.

Set-based evaluation measures

The results for the P, R, F and FO metrics, the plots for which are shown in Figs. 2.45, 2.46, 2.47 and 2.48, respectively, show values that clearly favor the other search engines over COGIR in the sets of topics of low and medium difficulty. However in the case of topics of the highest degree of difficulty, our proposal successfully maintains its position relative to those commented on for the topic set HBPQTC.

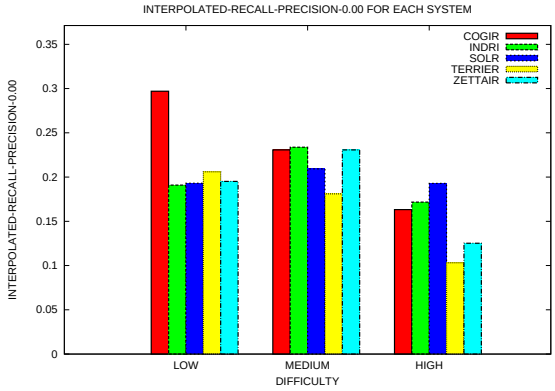

Figure 2.37: $\text{IP}_{\text{R}=0,00}$ on HBPQTC using PQREL

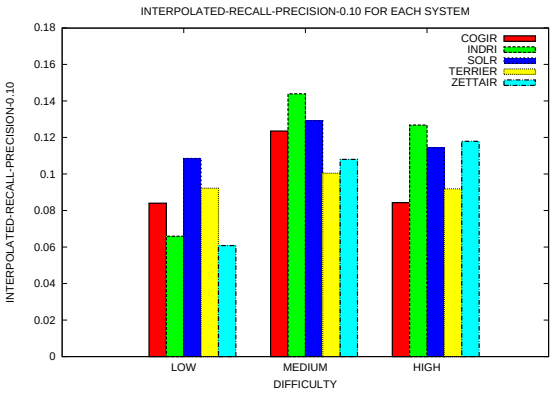

Figure 2.38: $\text{IP}_{\text{R}=0,10}$ on HBPQTC using PQREL

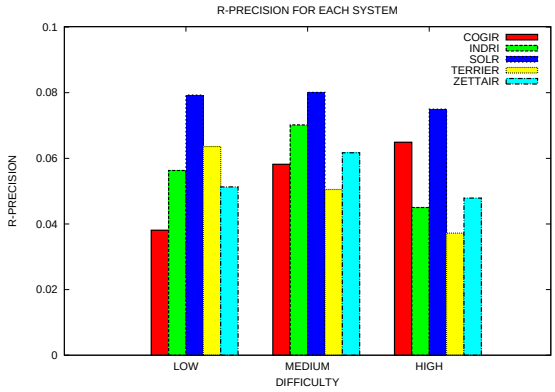

Figure 2.39: R-P on HBPQTC using PQREL

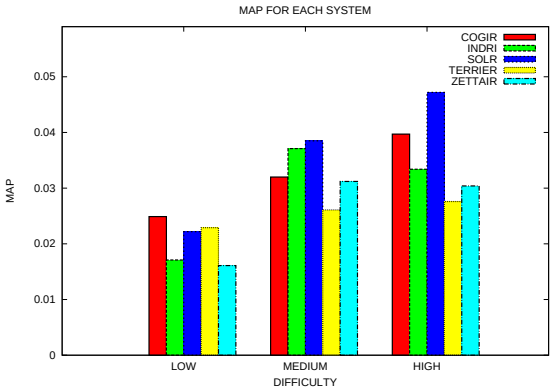

Figure 2.40: MAP on HBPQTC using PQREL

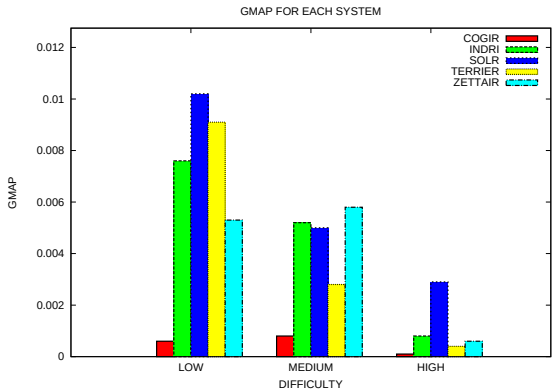

Figure 2.41: GMAP on HBPQTC using PQREL

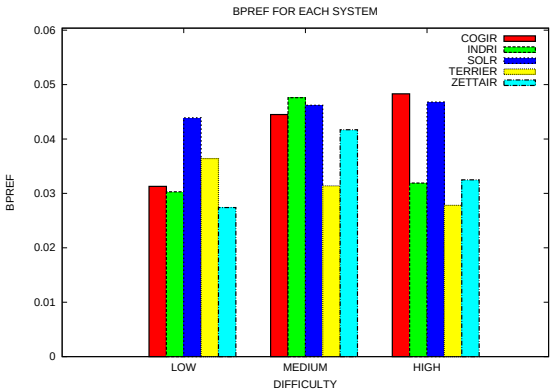

Figure 2.42: BPREF on HBPQTC using PQREL

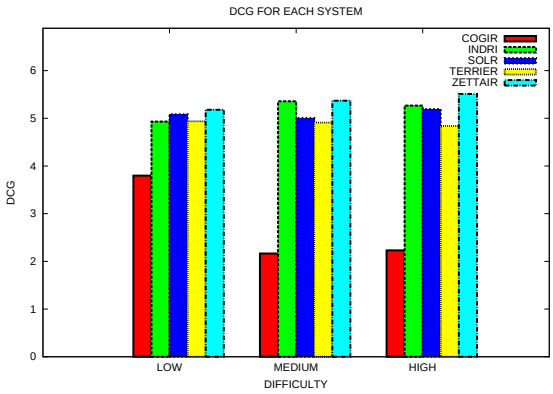

Figure 2.43: DCG on HBPQTC using PQREL

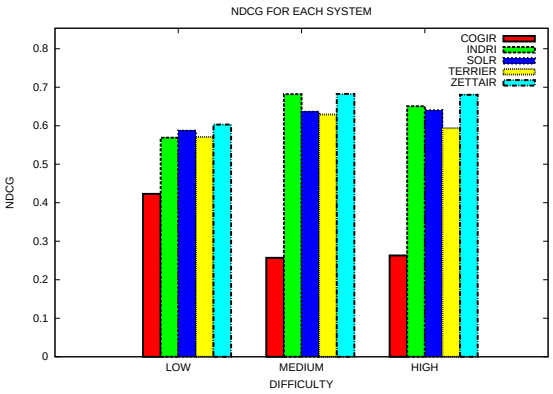

Figure 2.44: NDCG on HBPQTC using PQREL

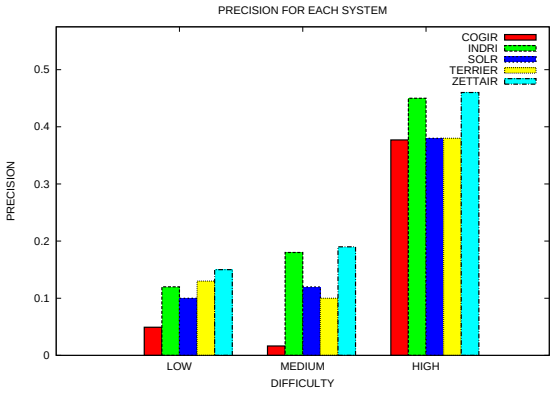

Figure 2.45: P on MBPQTC using PQREL

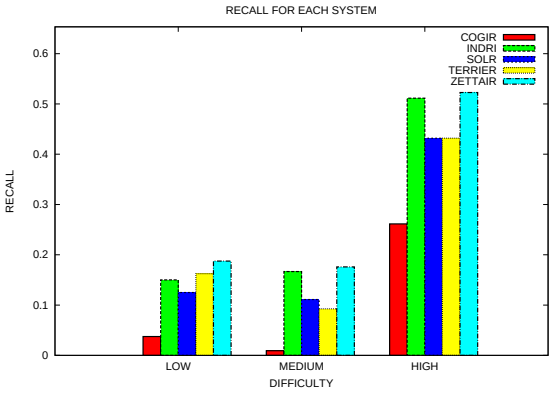

Figure 2.46: R on MBPQTC using PQREL

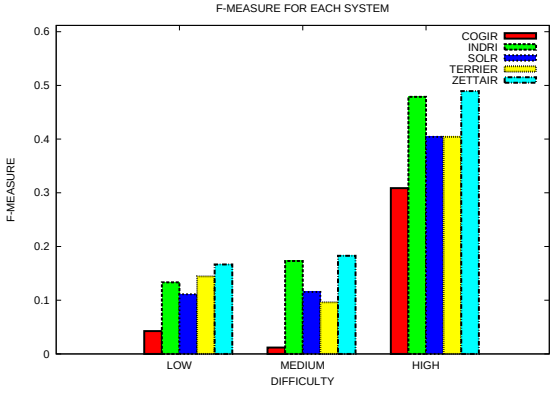

Figure 2.47: F on MBPQTC using PQREL

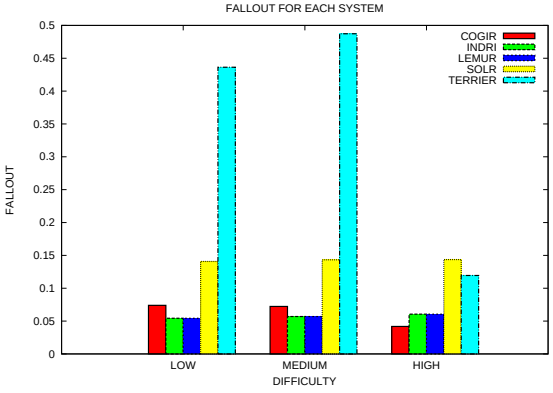

Figure 2.48: FO on MBPQTC using PQREL

Rank-based evaluation measures

We calculate the P@10 and R@10, IP_R for recall levels 0 and 0,10, R-P, MAP, GMAP, BPREF, DCG and NDCG in Figs. 2.49, 2.50, 2.51, 2.52, 2.53, 2.54, 2.55, 2.56, 2.57 and 2.58, respectively. The tests suggest that the COGIR search engine produces the poorest results when handling topics in the low and medium difficulty ranges. The results in the top range of difficulty are more or less on the same lines as those for the topic set HBPQTC, here also penalized by the use of PQREL. As in the case of the latter topic set, the conceptual approach fails to surpass its competitors.

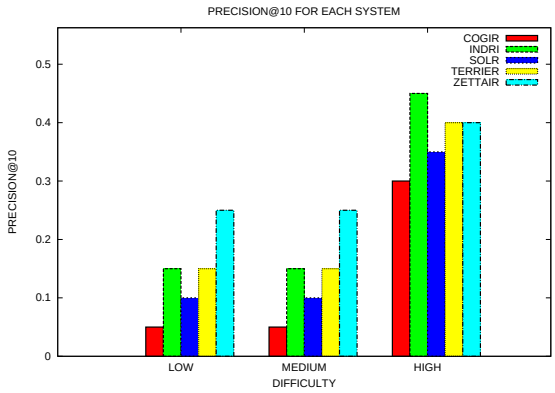

Figure 2.49: P@10 on MBPQTC using PQREL

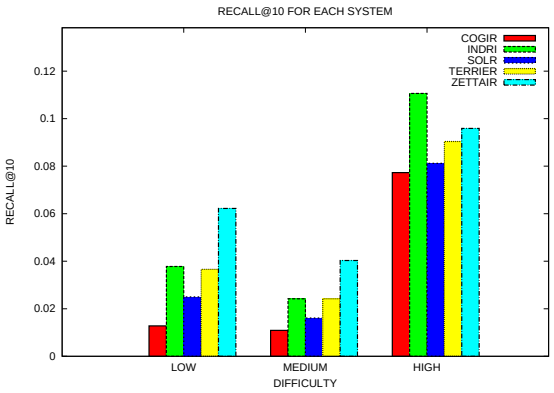

Figure 2.50: R@10 on MBPQTC using PQREL

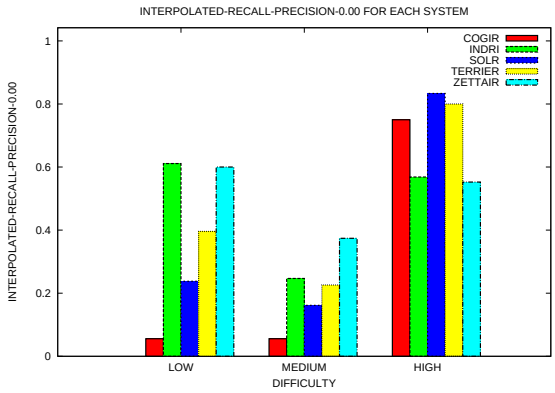

Figure 2.51: $IP_{R=0,00}$ on MBPQTC using PQREL

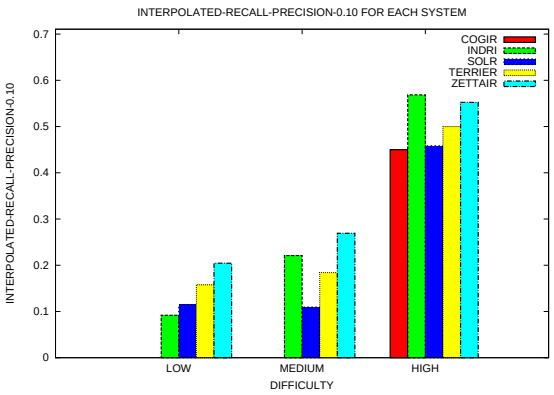

Figure 2.52: $IP_{R=0,10}$ on MBPQTC using PQREL

7.3 | Ranking IR systems using machine-based assessment

As has already been stated, the starting point for this ranking technique [101] is the AP measure, which implies that we need a certain number of relevance judgments in order to initiate the process. Given that they have both previously been introduced as judging strategies, we experiment at this level with QREL and PQREL.

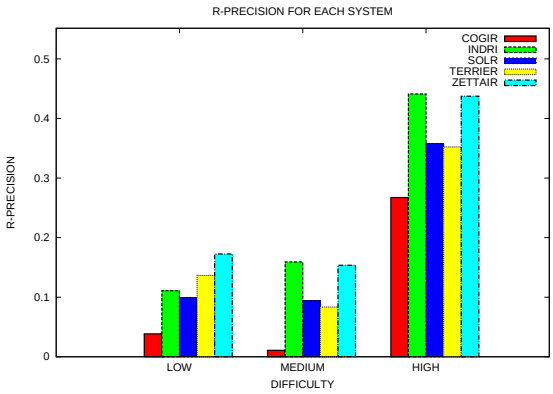

Figure 2.53: R-P on MBPQTC using PQREL

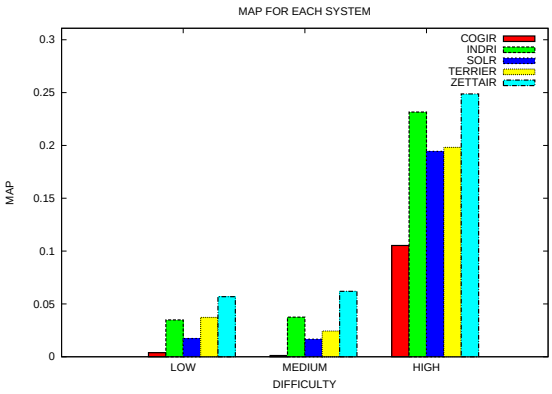

Figure 2.54: MAP on MBPQTC using PQREL

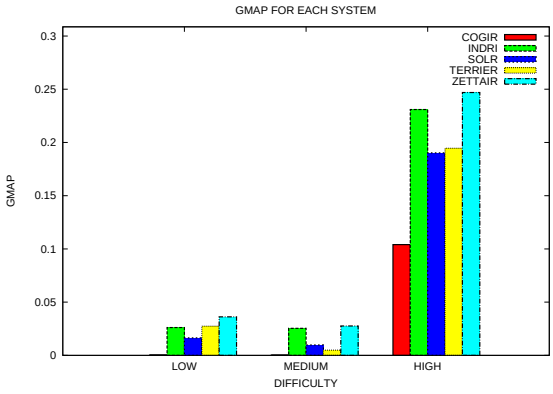

Figure 2.55: GMAP on MBPQTC using PQREL

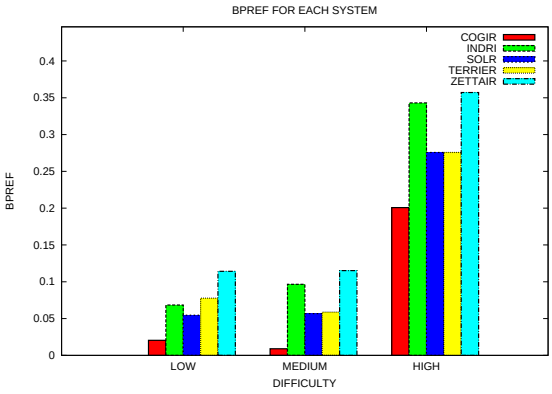

Figure 2.56: BPREF on MBPQTC using PQREL

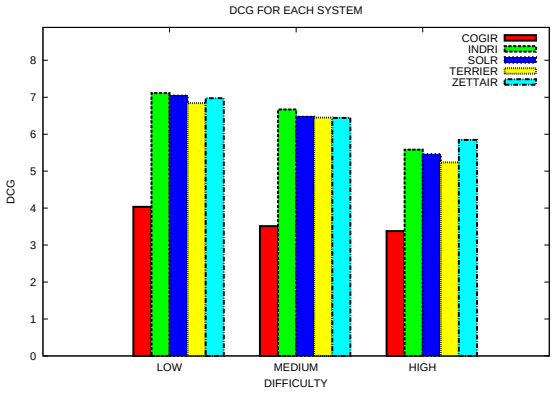

Figure 2.57: DCG on MBPQTC using PQREL

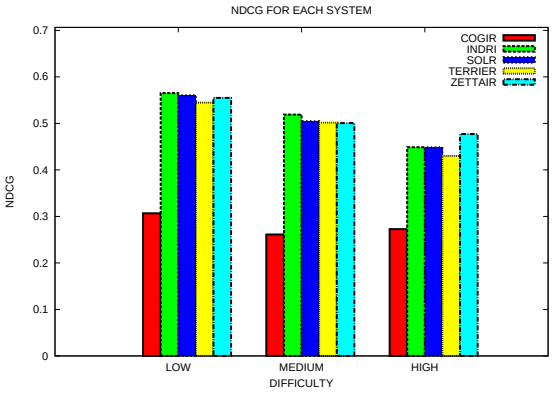

Figure 2.58: NDCG on MBPQTC using PQREL

7.3.1 | Computing AP from QREL

As previously done for ranking based on QREL, we differentiate between two series of tests at this point, one for each topic set built from QREL: HBQTC and MBQTC.

Using the human-based assessment topic sets collection

We here test a machine-based assessment ranking on the human-based assessment topic set collection (HBQTC). The results for the A measure are shown in Fig. 2.59, and once again give the conceptual search engine the advantage over the rest, particularly in the case of topics at the bottom and top ends of the difficulty range. In fact, even though COGIR registers its worst results with the medium difficulty topics, it still manages to outperform any of the other systems, which precisely perform best with this topic set.

Using the machine-based assessment topic sets collection

We now test a machine-based assessment ranking on the machine-based assessment topic set collection (MBQTC). The results for the A measure are shown in Fig. 2.60, and corroborate the behavior observed previously for the HBQTC topic set.

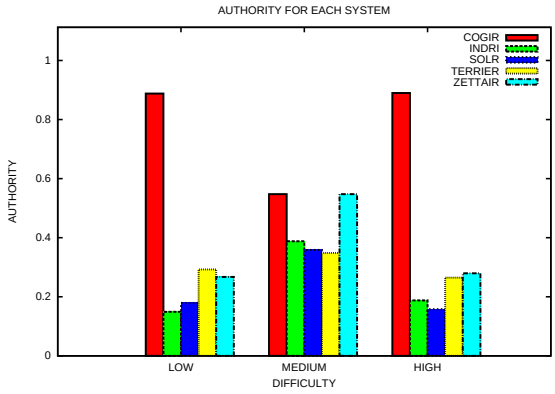

Figure 2.59: A on HBQTC using QREL

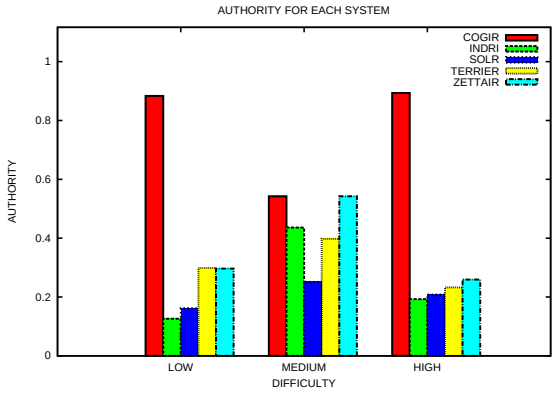

Figure 2.60: A on MBQTC using QREL

7.3.2 | Computing AP from PQREL

Following the same protocol described in dealing with AP computed from QREL, we here consider two series of tests, one for each topic set built from PQREL: HBPQTC and MBPQTC.

Using the human-based assessment topic sets collection

We now test a machine-based assessment ranking on the human-based assessment topic set collection (HBPQTC). The results for the A measure are shown in Fig. 2.61.

From a qualitative point of view, the performance observed for COGIR is analogous to that previously described when AP was calculated from QREL.

Using the machine-based assessment topic sets collection

We now turn our attention to machine-based assessment ranking performed on the machine-based assessment topic set collection (MBPQTC). The results for the A measure are shown in Fig. 2.62. Although the best performance continues to correspond to COGIR, in contrast to the previous plots for A, the worst results for the conceptual model are in this case obtained on the set of high difficulty topics.

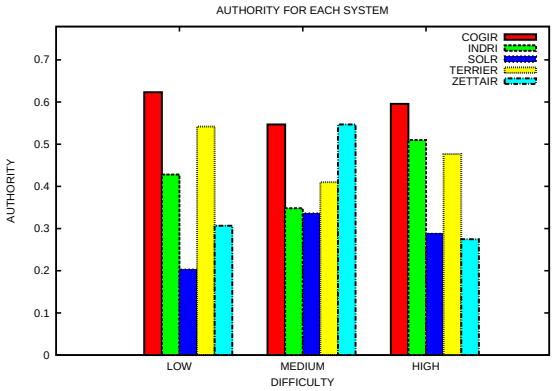

Figure 2.61: A on HBPQTC using PQREL

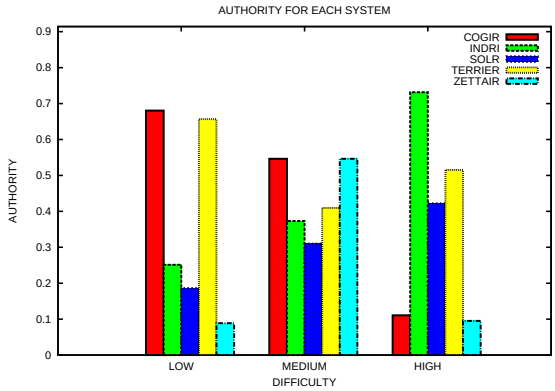

Figure 2.62: A on MBPQTC using PQREL

7.4 | Ranking IR systems using average weighted reference counts

The last ranking proposal we consider was described by Wu *et al.* in [164] and is based on the concept of average weighted reference count. As we have already mentioned above, four ranking measures can be here considered: $AWRC_O$, $AWRC_S$, $AWRC_{LO}$ and $AWRC_{LS}$.

Given that in this case the ranking strategy is not related with any judging strategy in particular, we shall consider the complete set of topic sets previously introduced in order to ensure a complete testing procedure: HBQTC, MBQTC, HBPQTC and MBPQTC. This allows us to consider both human and machine-based assessment when selecting the topic, as well as QREL and PQREL-based techniques in order to reduce the size of the topic sets. In this way, we do not favor any strategy that could be used to refine some of the IR systems we are comparing, an important point to take into account when we consider a ranking method whose starting point is the counting of cross-references between the set of documents returned by the search engines.

7.4.1 | Using QREL-based topic reduction

We first experiment with topic sets obtained from QREL-based topic reduction techniques, which includes both human and machine-based topic sets collections.

Using the human-based assessment topic sets collection

On this occasion the results are shown for $AWRC_O$, $AWRC_S$, $AWRC_{LO}$ and $AWRC_{LS}$ metrics on the topic set HBQTC in Figs. 2.63, 2.64, 2.65 and 2.66 respectively. In these cases, the conceptual approach apparently displays the worst performance possible, especially when dealing with high difficulty topics, although the results are slightly better for $AWRC_O$ and $AWRC_{LO}$ measures. Contrary to what one may think, such behavior is not only congruent with the previous metrics, but also perfectly foreseeable.

When relativist techniques are applied, the IR system being tested will never be able to improve on the performance of the set of those that act as a comparative reference. What is more, methodologies of this kind can lead to disastrously erroneous situations when the set of such reference systems shows a universally poor performance on a given set of topics, whilst the system being tested offers a good level of precision. This is exactly the behavior observed here on the set of topics at the highest level of difficulty, for which, as we have seen, the conceptual approach comes off best in all the previous metrics: in this case, however, it appears to come off worst.

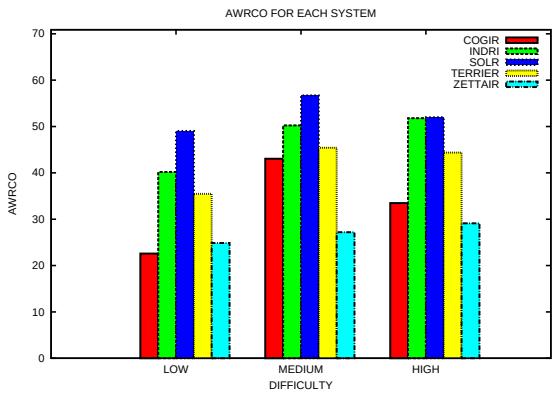

Figure 2.63: $AWRC_O$ on HBQTC

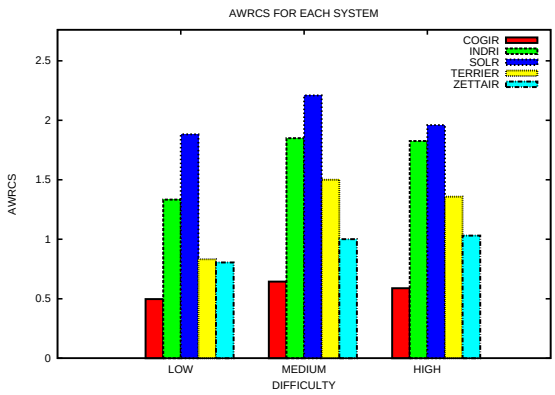

Figure 2.64: $AWRC_S$ on HBQTC

Using the machine-based assessment topic sets collection

The results are now shown for $AWRC_O$, $AWRC_S$, $AWRC_{LO}$ and $AWRC_{LS}$ measures on the topic set MBQTC, in Figs. 2.67, 2.68, 2.69 and 2.70 respectively. We can apply exactly the same comments made above for the tests on the set of HBQTC topics, corroborating the reasoning given regarding the kind of assessment applied to topic selection.

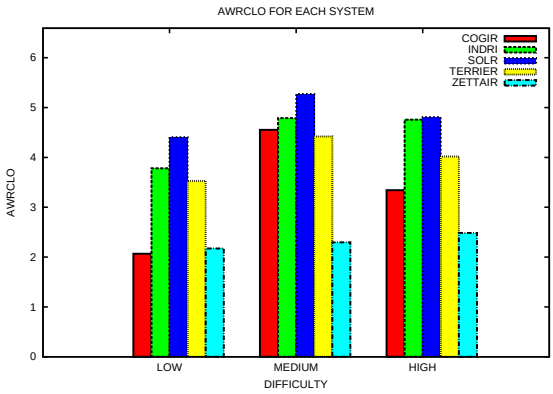

Figure 2.65: $AWRC_{LO}$ on HBQTC

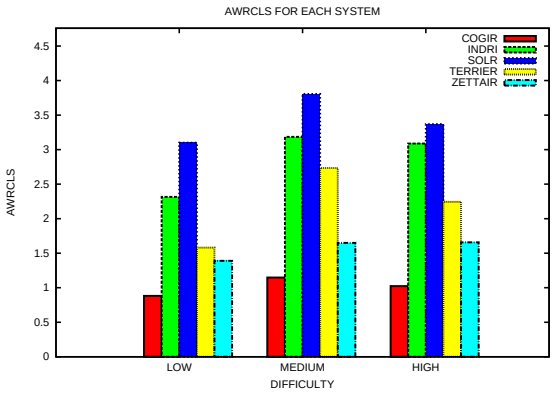

Figure 2.66: $AWRC_{LS}$ on HBQTC

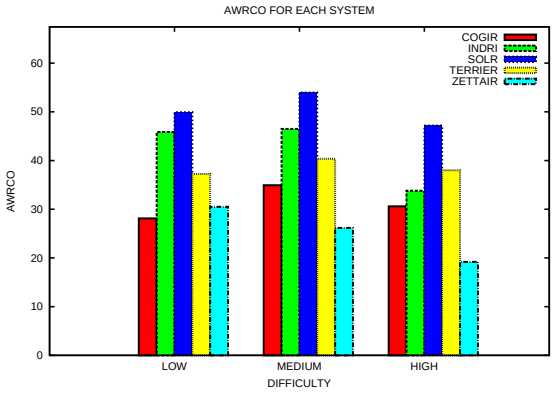

Figure 2.67: $AWRC_{O}$ on MBQTC

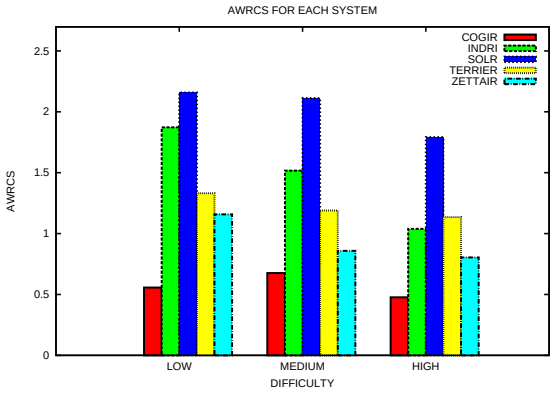

Figure 2.68: $AWRC_{S}$ on MBQTC

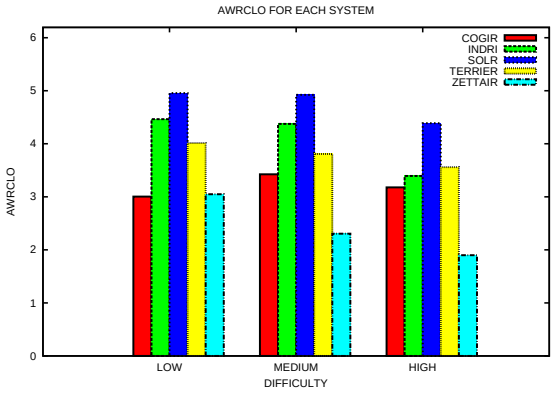

Figure 2.69: $AWRC_{LO}$ on MBQTC

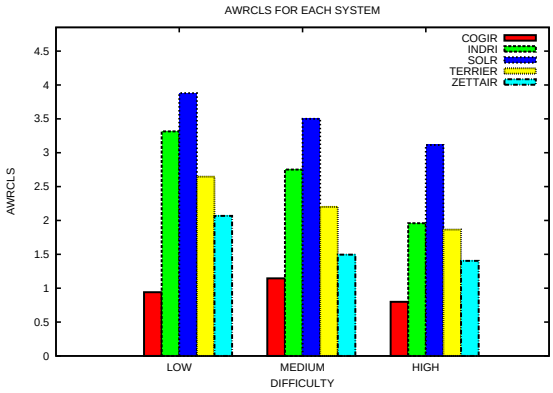

Figure 2.70: $AWRC_{LS}$ on MBQTC

7.4.2 | Using PQREL-based topic reduction

The experiments now concern topic sets obtained from PQREL-based reduction methods, including both human and machine-based topic set collections.

Using the human-based assessment topic sets collection

As previously for the case of QREL, the results are shown for the $AWRC_O$, $AWRC_S$, $AWRC_{LO}$ and $AWRC_{LS}$ metrics on the topic set HBPQTC are shown in Figs. 2.71, 2.72, 2.73 and 2.74 respectively. The results shown in the plots are qualitatively equivalent to those previously commented on for QREL-based topic reduction, but with one important difference, namely that the conceptual model achieves the best results for the set of difficult topics, when in all previous cases it obtained the worst ones.

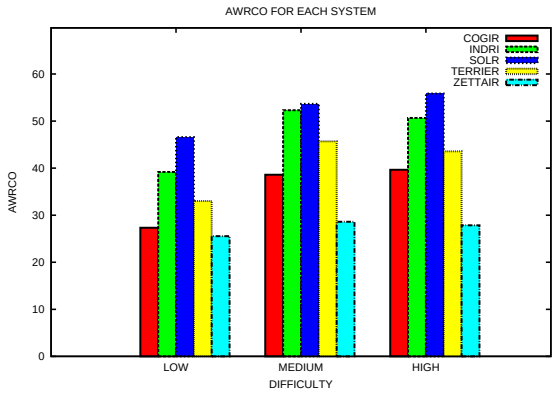

Figure 2.71: $AWRC_O$ on HBPQTC

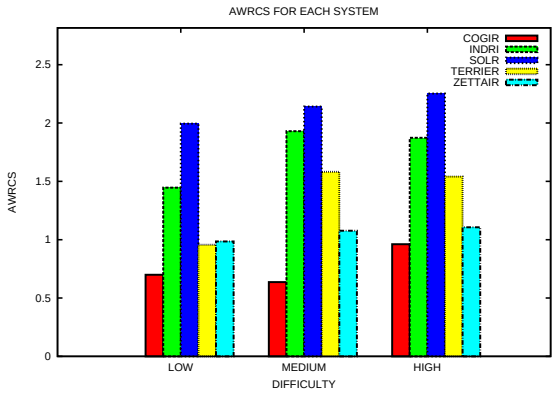

Figure 2.72: $AWRC_S$ on HBPQTC

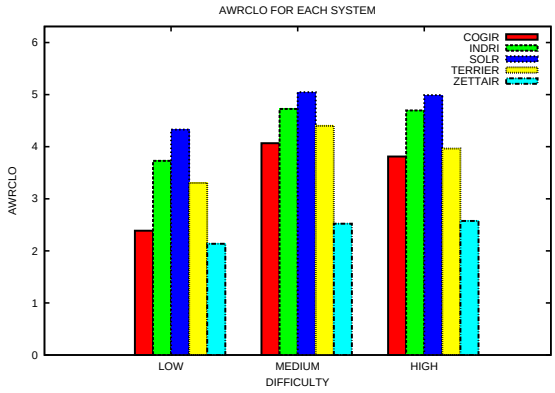

Figure 2.73: $AWRC_{LO}$ on HBPQTC

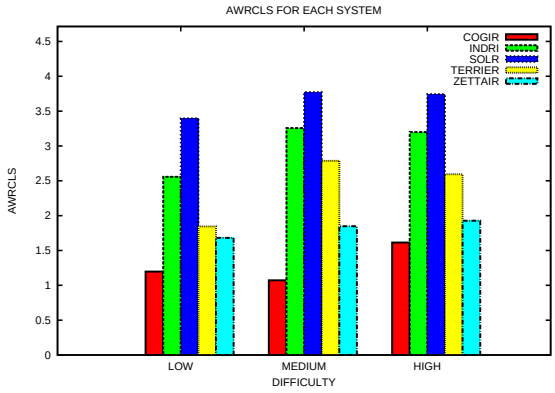

Figure 2.74: $AWRC_{LS}$ on HBPQTC

Using the machine-based assessment topic sets collection

The results are now shown for $AWRC_O$, $AWRC_S$, $AWRC_{LO}$ and $AWRC_{LS}$ metrics on the topic set MBPQTC, in Figs. 2.75, 2.76, 2.77 and 2.78 respectively. The experimental results

are here quantitatively equivalent to those commented on above, although sensitively different from a qualitative point of view. In particular, in contrast to the previous tests, the worst results for the conceptual approach in the case of AWRC_{O} and AWRC_{LO} are obtained for the set of easy topics. With regard to the AWRC_{S} and AWRC_{LS} metric the results are equivalent to those obtained for the human-based assessment topic sets collection.

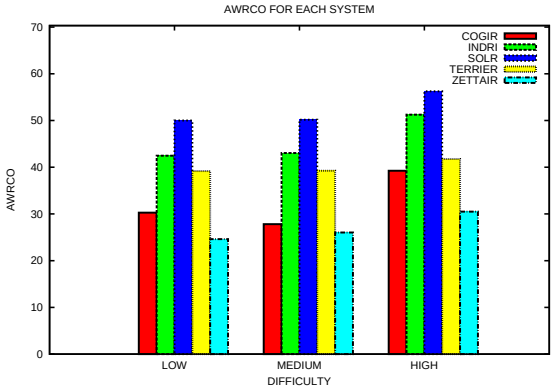

Figure 2.75: AWRC_{O} on MBPQTC

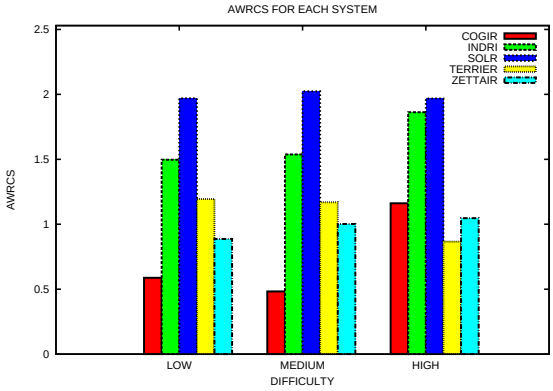

Figure 2.76: AWRC_{S} on MBPQTC

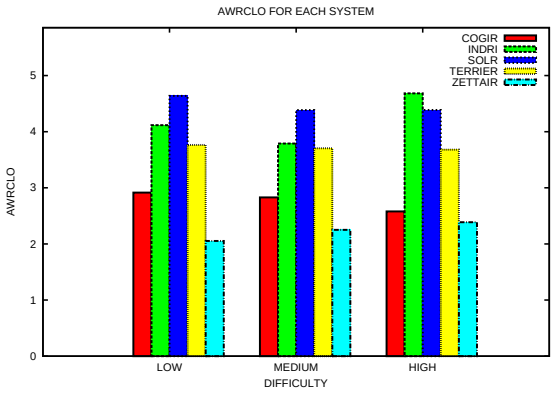

Figure 2.77: AWRC_{LO} on MBPQTC

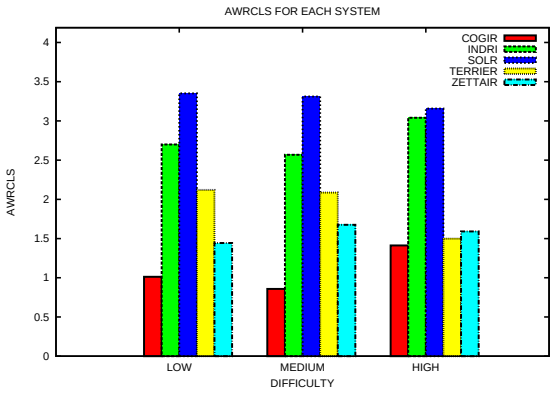

Figure 2.78: AWRC_{LS} on MBPQTC

8 | Conclusions

The convenience of whether to include specific linguistic knowledge in search engine design or not is a debate that goes back to the very origins of IR itself. Traditionally, three arguments have been put forward against doing so: the algorithmic complexity it entails, the scarcity or even lack of logical resources, and the apparently meager improvement in performance obtained. Having assumed the technical complexity inherent to strategies of this kind, we introduce a methodology for the machine acquisition of the text semantics based on the lexical and syntactic information, summarized in a conceptual graph that is nothing less than a reflection of the set of dependency relations recognized earlier. This

enables us not only to make use of a formal structure that faithfully transmits the meaning of any document, but also provides an ideal structural foundation upon which to place an approximate pattern-matching algorithm able to calculate the semantic proximity between two different documents.

A further aim is to throw some practical light on what seems obvious from an intuitive standpoint, namely that a semantic foundation that is improved in the retrieval process should be mirrored in the performance of the system. To this end we defined a full formal assessment environment that to the best of our knowledge constitutes a complete sample of currently available techniques. This enabled us to exploit to the full the possibilities of the conceptual approach to IR, as compared to the more generic nature of traditional search engines.

The results obtained would appear to put an end to the debate for once and for all, since they reveal a performance that, even in the worst case, is equal to that of BOW-based systems. The only exception we have observed to the above are the results of the PQREL's-based tests, which naturally favor architectures associated with the IR systems used as a reference for generating such structures.

Additionally, we systematically observed a significant qualitative leap forward when attempting to resolve queries classified as being of increasing difficulty, and which we associate with topics having a greater capacity for discrimination between the different systems compared.

Intuitively, this result coincides with what one could expect, since the importance of semantic information grows in line with the increasing complexity of the meaning of the text to be analyzed, whether in terms of the document collection we are exploring or of the topic itself. On the other hand, when the simplicity of the query or of the content of the texts in question means that we do not have to deal with complex semantic relationships, all the systems evaluated show similar performance, regardless of the nature of the tagging architecture employed.

Índice alfabético

Alphabetical index

Bibliografía

[1] A. Al-Maskari, M. Sanderson, and P. Clough. The relationship between IR effectiveness measures and user satisfaction. In *Proc. of the 30th Int. Conf. on Research and Development in Information Retrieval*, SIGIR'07, pages 773–774, New York, NY, USA, 2007. ACM.

[2] Miguel A. Alonso, Jesús Vilares Ferro, and Victor M. Darriba. On the usefulness of extracting syntactic dependencies for text indexing. In *Proceedings of the 13th Irish International Conference on Artificial Intelligence and Cognitive Science*, AICS '02, pages 3–11, London, UK, 2002. Springer-Verlag.

[3] Miguel A. Alonso Pardo, David Cabrero Souto, Manuel Vilares, and Éric Villemonte de La Clergerie. Tabular algorithms for TAG parsing. In *Proc. of EACL'99*, 1999.

[4] G. Amati and C.J. Van Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Transanctions on Information Systems*, 20(4):357–389, October 2002.

[5] J.A. Aslam, E. Yilmaz, and V Pavlu. A geometric interpretation of R-precision and its correlation with average precision. In *Proc. of the 28th Int. Conf. on Research and Development in Information Retrieval*, SIGIR'05, pages 573–574, New York, NY, USA, 2005. ACM.

[6] J. Attenberg and T. Suel. Cleaning search results using term distance features. In *Proc. of the 4th Int. Workshop on Adversarial Information Retrieval on the Web*, AIRWeb '08, pages 21–24, New York, NY, USA, 2008. ACM.

[7] T. Galen Ault and Y. Yang. Information filtering in trec-9 and tdt-3: A comparative analysis. *Information Retrieval*, 5:159–187, April 2002.

[8] R. Baeza-Yates, F. Saint-Jean, and C. Castillo. Web structure, dynamics and page quality. In *Proc. of 9th Int. Symposium on String Processing and Information Retrieval*, volume 2476 of *SPIRE'02*, pages 117–132, Lisbon, Portugal, 2002. Springer.

[9] S. Bani-Ahmad and G. Ozsoyoglu. On popularity quality: growth and decay phases of publication popularities. In *Proc. of the 6th Int. Conf. on Innovations in Information Technology*, IIT'09, pages 231–235, Piscataway, NJ, USA, 2009. IEEE Press.

[10] D. Bollegala, N. Noman, and H. Iba. Rankde: learning a ranking function for information retrieval using differential evolution. In *Proc. of the 13th Annual Conf. on Genetic and Evolutionary Computation*, GECCO'11, pages 1771–1778, New York, NY, USA, 2011. ACM.

[11] C. Buckley and E.M. Voorhees. Evaluating evaluation measure stability. In *Proc. of the 23rd Int. Conf. on Research and Development in Information Retrieval*, SIGIR'00, pages 33–40, New York, NY, USA, 2000. ACM.

[12] C. Buckley and E.M. Voorhees. Retrieval evaluation with incomplete information. In *Proc. of the 27th Int. Conf. on Research and Development in Information Retrieval*, SIGIR'04, pages 25–32, New York, NY, USA, 2004. ACM.

[13] Chris Buckley, Darrin Dimmick, Ian Soboroff, and Ellen M. Voorhees. Bias and the limits of pooling. In *In* Proc. of the 29th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, SIGIR'06*, pages 619–620, 2006.

[14] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender. Learning to rank using gradient descent. In *Proc. of the 22nd Int. Conf. on Machine learning*, ICML'05, pages 89–96, New York, NY, USA, 2005. ACM.

[15] C.J.C. Burges, R. Ragno, and Q. Viet Le. Learning to rank with nonsmooth cost functions. In B. Schölkopf, J.C. Platt, and T. Hoffman, editors, *Proc. of the 20th Annual Conf. on Neural Information Processing Systems*, volume 19, pages 193–200. MIT Press, 2006.

[16] Y. Cao, J. Xu, T.-Y. Liu, H. Li, Y. Huang, and H.-W. Hon. Adapting ranking SVM to document retrieval. In *Proc. of the 29th Annual Int. Conf. on Research and Development in Information Retrieval*, SIGIR'06, pages 186–193, New York, NY, USA, 2006. ACM.

[17] Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai, and H. Li. Learning to rank: from pairwise approach to listwise approach. In *Proc. of the 24th Int. Conf. on Machine learning*, ICML'07, pages 129–136, New York, NY, USA, 2007. ACM.

[18] D. Carmel, H. Roitman, and E. Yom-Tov. On the relationship between novelty and popularity of user-generated content. In *Proc. of the 19th Int. Conf. on Information and knowledge Management*, CIKM'10, pages 1509–1512, New York, NY, USA, 2010. ACM.

[19] D. Carmel and E. Yom-Tov. *Estimating the Query Difficulty for Information Retrieval*. Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan & Claypool Publishers, 2010.

[20] B. Carterette and P.N. Bennett. Evaluation measures for preference judgments. In *Proc. of the 31st Int. Conf. on Research and Development in Information Retrieval*, SIGIR'08, pages 685–686, New York, NY, USA, 2008. ACM.

[21] B. Carterette, V. Pavlu, E. Kanoulas, J.A. Aslam, and J. Allan. Evaluation over thousands of queries. In *Proc. of the 31st Int. Conf. on Research and Development in Information Retrieval*, SIGIR'08, pages 651–658, New York, NY, USA, 2008. ACM.

[22] C. Castillo and B.D. Davison. Adversarial web search. *Foundations and Trends in Information Retrieval*, 4(5):377–486, May 2011.

[23] C. Castillo, D. Donato, and A. Gionis. Estimating number of citations using author reputation. In *Proc. of 14th Int. Symposium on String Processing and Information Retrieval*, SPIRE'07, pages 107–117, Berlin, Heidelberg, 2007. Springer-Verlag.

[24] Michel Chein and Marie laure Mugnier. Conceptual graphs: fundamental notions. *Revue d'Intelligence Artificielle*, 6:365–406, 1992.

[25] Michel Chein and Marie-Laure Mugnier. *Graph-based Knowledge Representation: Computational Foundations of Conceptual Graphs*. Springer, London, 2008.

[26] Bong-Hyun Cho, Changki Lee, and Gary Geunbae Lee. Exploring term dependences in probabilistic information retrieval model. *Inf. Process. Manage.*, 39:505–519, July 2003.

[27] J. Cho and S. Roy. Impact of search engines on page popularity. In *Proc. of the 13th Int. Conf. on World Wide Web*, WWW'04, pages 20–29, New York, NY, USA, 2004. ACM.

[28] J. Cho, S. Roy, and R.E. Adams. Page quality: in search of an unbiased web ranking. In *Proc. of the 24th Int. Conf. on Management of Data*, SIGMOD'05, pages 551–562, New York, NY, USA, 2005. ACM.

[29] C. Cleverdon, J. Mills, and E.M. Keen. An inquiry in testing of information retrieval systems. 1966.

[30] C.W. Cleverdon. The significance of the Cranfield tests on index languages. In *Proc. of the 14th Int. Conf. on Research and Development in Information Retrieval*, SIGIR'91, pages 3–12, New York, NY, USA, 1991. ACM.

[31] Cyril Cleverdon. *The Cranfield tests on index language devices*, pages 47–59. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1997.

[32] E.F. Codd. A relational model of data for large shared data banks. *Communications of the ACM*, 26:64–69, January 1983.

[33] W.W. Cohen, A. Borgida, and H. Hirsh. Computing least common subsumers in description logics. In *Proc. of the Tenth Int. Conf. on Artificial intelligence*, AAAI'92, pages 754–760. AAAI Press, 1992.

[34] W.W. Cohen and Y. Singer. Context-sensitive learning methods for text categorization. *ACM Transactions on Information Systems*, 17(2):141–173, 1999.

[35] D. Corbett. Graph-based representation and reasoning for ontologies. In John Fulcher and L. Jain, editors, *Computational Intelligence: A Compendium*, volume 115 of *Studies in Computational Intelligence*, pages 351–379. Springer Berlin / Heidelberg, 2008.

[36] Olivier Corby. Web, graphs and semantics. In *Proceedings of the 16th international conference on Conceptual Structures: Knowledge Visualization and Reasoning*, ICCS '08, pages 43–61, Berlin, Heidelberg, 2008. Springer-Verlag.

[37] R. Cummins and C. O'Riordan. Term-weighting in information retrieval using genetic programming: A three stage process. In *Proc. of the 17th European Conf. on Artificial Intelligence*, ECAI'06, pages 793–794, Amsterdam, The Netherlands, The Netherlands, 2006. IOS Press.

[38] K. Curran, C. Murphy, and S. Annesley. Intelligent information retrieval. *Int. Journal of Advanced Media and Communication*, 1(2):139–147, 2006.

[39] H.M. de Almeida, M.A. Gonçalves, M. Cristo, and P. Calado. A combined component approach for finding collection-adapted ranking functions based on genetic programming. In *Proc. of the 30th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, SIGIR'07, pages 399–406, New York, NY, USA, 2007. ACM.

[40] L.R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, July 1945.

[41] A. Dong, Y. Chang, Z. Zheng, G. Mishne, J. Bai, R. Zhang, K. Buchner, C. Liao, and F. Diaz. Towards recency ranking in web search. In *Proc. of the Third ACM Int. Conf. on Web Search and Data Mining*, WSDM'10, pages 11–20, New York, NY, USA, 2010. ACM.

[42] A. Doucet and H. Ahonen-Myka. Non-contiguous word sequences for information retrieval. In *Proc. of the Workshop on Multiword Expressions*, MWE'04, pages 88–95, Morristown, NJ, USA, 2004. Association for Computational Linguistics.

[43] Y. Duan, L. Jiang, T. Qin, M. Zhou, and H.-Y. Shum. An empirical study on learning to rank of tweets. In *Proc. of the 23rd Int. Conf. on Computational Linguistics*, COLING'10, pages 295–303, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[44] Miles Efron. Using multiple query aspects to build test collections without human relevance judgments. In *ECIR '09: Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval*, pages 276–287. Springer-Verlag, 2009.

[45] J.L. Elsas and S.T. Dumais. Leveraging temporal dynamics of document content in relevance ranking. In *Proc. of the Third ACM Int. Conf. on Web Search and Data Mining*, WSDM '10, pages 1–10, New York, NY, USA, 2010. ACM.

[46] J.L. Fagan. Automatic phrase indexing for document retrieval: An examination of syntactic and non-syntactic methods. In *Proc. of the 10th Int. Conf. on Research and Development in Information Retrieval*, SIGIR'87, pages 91–101. ACM, 1987.

[47] W. Fan, M.D. Gordon, and P. Pathak. A generic ranking function discovery framework by genetic programming for information retrieval. *Information Processing & Management*, 40:587–602, May 2004.

[48] W. Fan, M.D. Gordon, and P. Pathak. Genetic programming-based discovery of ranking functions for effective web search. *Journal of Management Information Systems*, 21:37–56, April 2005.

[49] D. Faure and C. Nedellec. A corpus-based conceptual clustering method for verb frames and ontology. In P. Velardi, editor, *Proceedings of the LREC Workshop on Adapting lexical and corpus resources to sublanguages and applications*, pages 5–12, 1998.

[50] Milagros Fernández, Eric Villemonte de la Clergerie, and Manuel Vilares Ferro. Mining conceptual graphs for knowledge acquisition. In Fotis Lazarinis, Efthimis N. Efthimiadis, Jesús Vilares, and John Tait, editors, *CIKM-iNEWS*, pages 25–32. ACM, 2008.

[51] F. Fonseca, M. Egenhofer, C. Davis, and G. Câmara. Semantic granularity in ontology-driven geographic information systems. *Annals of Mathematics and Artificial Intelligence*, 36:121–151, September 2002.

[52] Y. Freund, R. Iyer, R.E. Schapire, and Y. Singer. An efficient boosting algorithm for combining preferences. *The Journal of Machine Learning Research*, 4:933–969, December 2003.

[53] N. Fuhr and C. Buckley. A probabilistic learning approach for document indexing. *ACM Transanctions on Information Systems*, 9:223–248, July 1991.

[54] J. Gao and J.-Y. Nie. A study of statistical models for query translation: finding a good unit of translation. In *Proc. of the 29th Int. Conf. on Research and Development in Information Retrieval*, SIGIR'06, pages 194–201, New York, NY, USA, 2006. ACM.

[55] Jianfeng Gao, Jian-Yun Nie, Guangyuan Wu, and Guihong Cao. Dependence language model for information retrieval. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '04, pages 170–177, New York, NY, USA, 2004. ACM.

[56] David Genest. *Extension du modèle des graphes conceptuels pour la recherche d'informations*. PhD thesis, Université Montpellier II, 2000.

[57] David Genest and Michel Chein. A content-search information retrieval process based on conceptual graphs. *Knowl. Inf. Syst*, 8(3):292–309, 2005.

[58] P. Ghodsnia, A.M.Z. Bidoki, and N. Yazdani. A punishment/reward based approach to ranking. In *Proc. of the 2nd Int. Conf. on Scalable information systems*, InfoScale'07, pages 58:1–58:4, ICST, Brussels, Belgium, Belgium, 2007. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering).

[59] J.A. Goldsmith, D. Higgins, and S. Soglasnova. Automatic language-specific stemming in information retrieval. In *Revised Papers from the Workshop of Cross-Language Evaluation Forum on Cross-Language Information Retrieval and Evaluation*, CLEF'00, pages 273–284, London, UK, 2001. Springer-Verlag.

[60] M. Gordon and P. Pathak. Finding information on the world wide web: the retrieval effectiveness of search engines. *Information Processing & Management*, 35:141–180, March 1999.

[61] L.A. Granka, T. Joachims, and G. Gay. Eye-tracking analysis of user behavior in WWW search. In *Proc. of the 27th Int. Conf. on Research and Development in Information Retrieval*, SIGIR'04, pages 478–479, New York, NY, USA, 2004. ACM.

[62] J. Guiver, S. Mizzaro, and S. Robertson. A few good topics: Experiments in topic set reduction for retrieval evaluation. *ACM Transactions on Information Systems*, 27:21:1–21:26, November 2009.

[63] D. Harman. Overview of the second text retrieval conference (trec-2). In *Proc. of the workshop on Human Language Technology*, HLT'94, pages 351–357, Stroudsburg, PA, USA, 1994. Association for Computational Linguistics.

[64] Zellig Harris. Distributional structure. *Word*, 10(23):146–162, 1954.

[65] Zellig Harris. *Mathematical Structures of Language*. John Wiley and Son, New York, 1968.

[66] S.P. Harter. Variations in relevance assessments and the measurement of retrieval effectiveness. *Journal of the American Society for Information Science*, 47:37–49, January 1996.

[67] Taher H. Haveliwala. Topic-sensitive pagerank. In *Proc. of the 11th Int. Conf. on World Wide Web*, WWW'02, pages 517–526, New York, NY, USA, 2002. ACM.

[68] R. Herbrich, T. Graepel, and K. Obermayer. Support vector learning for ordinal regression. In *In Proc. of Int. Conf. on Artificial Neural Networks*, ICANN'99, pages 97–102, 1999.

[69] R. Herbrich, T. Graepel, and K. Obermayer. Large margin rank boundaries for ordinal regression. In P.J. Bartlett, B. Schölkopf, D. Schuurmans, and A.J. Smola, editors, *Advances in Large Margin Classifiers*, pages 115–132. MIT Press, 2000.

[70] John E. Hopcroft and Jeffrey D. Ullman. *Introduction to Automata Theory, Languages and Computation*. Addison-Wesley, 1979.

[71] Christian Jacquemin, Judith Klavans, and Evelyne Tzoukermann. Expansion of multi-word terms for indexing and retrieval using morphology and syntax. In *ACL*, pages 24–31, 1997.

[72] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. volume 20, pages 422–446, New York, NY, USA, October 2002. ACM.

[73] W. Jin and R.K. Srihari. Graph-based text representation and knowledge discovery. In *Proc. of the Symposium on Applied Computing*, SAC'07, pages 807–811, New York, NY, USA, 2007. ACM.

[74] T. Joachims. Optimizing search engines using clickthrough data. In *Proc. of the Eighth ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, KDD'02, pages 133–142, New York, NY, USA, 2002. ACM.

[75] K.S. Jones. *What is the role of* NLP *in text retrieval ?*, pages 1–24. Text, Speech and Language Technology Book Series. Kluwer Academic Publishers, 1999.

[76] T. Jones, D. Hawking, P. Thomas, and R. Sankaranarayana. Relative effect of spam and irrelevant documents on user interaction with search engines. In *Proc. of the 20th Int. Conf. on Information and Knowledge Management*, CIKM'11, pages 2113–2116, New York, NY, USA, 2011. ACM.

[77] Aravind K Joshi. An introduction to tree adjoining grammar. In A Manaster-Ramer, editor, *Mathematics of Language*. John Benjamins, Amsterdam, 1987.

[78] J.S. Justeson and S.M. Katz. Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1:9–27, 1995.

[79] F. Lepage J.Y. Nie. *Toward a broader model for information retrieval*, chapter nformation Retrieval, Uncertainty and Logics, pages 17–38. eds. M. Lalmas, F. Crestani, C.J. van Rijsbergen, Kluwer Academic Publishers, 1998.

[80] M.G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.

[81] Kimmo Kettunen, Eija Airio, and Kalervo Järvelin. Restricted inflectional form generation in management of morphological keyword variation. *Inf. Retr.*, 10:415–444, October 2007.

[82] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46:604–632, September 1999.

[83] Phokion G. Kolaitis and Moshe Y. Vardi. Conjunctive-query containment and constraint satisfaction. In *Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*, PODS '98, pages 205–213, New York, NY, USA, 1998. ACM.

[84] T.G. Kolda and D.P. O'Leary. A semidiscrete matrix decomposition for latent semantic indexing information retrieval. *ACM Transactions on Information Systems*, 16(4):322–346, 1998.

[85] C.H.A. Koster and J.G. Beney. Phrase-based document categorization revisited. In *Proc. of the 2nd Int. Workshop on Patent Information Retrieval*, PaIR'09, pages 49–56, New York, NY, USA, 2009. ACM.

[86] A. Kulkarni, J. Teevan, K.M. Svore, and S.T. Dumais. Understanding temporal query dynamics. In *Proc. of the Fourth ACM Int. Conf. on Web Search and Data Mining*, WSDM '11, pages 167–176, New York, NY, USA, 2011. ACM.

[87] J.-W. Kuo, P.-J. Cheng, and H.-M. Wang. Learning to rank from bayesian decision inference. In *Proc. of the 18th Int. Conf. on Information and Knowledge Management*, CIKM'09, pages 827–836, New York, NY, USA, 2009. ACM.

[88] R. Küsters and R. Molitor. Structural Subsumption and Least Common Subsumers in a Description Logic with Existential and Number Restrictions. *Studia Logica*, 81:227–259, 2005.

[89] Y. Lan, T.-Y. Liu, Z. Ma, and H. Li. Generalization analysis of listwise learning-to-rank algorithms. In *Proc. of the 26th Annual Int. Conf. on Machine Learning*, ICML'09, pages 577–584, New York, NY, USA, 2009. ACM.

[90] C. Lee and G.G. Lee. Probabilistic information retrieval model for a dependency structured indexing system. *Information Processing & Management*, 41(2):161–175, 2005.

[91] Fritz Lehmann. Semantic networks. *Computers & Mathematics with Applications*, 23(2-5):1 – 50, 1992.

[92] M. Li, H. Li, and Z.-H. Zhou. Semi-supervised document retrieval. *Information Processing & Management*, 45:341–355, May 2009.

[93] P. Li, C.J.C. Burges, and Q. Wu. Mcrank: Learning to rank using multiple classification and gradient boosting. In J.C. Platt, D. Koller, Y. Singer, and S.T. Roweis, editors, *Proc. of Advances in Neural Information Processing Systems*, volume 20 of *NIPS'07*, pages 897–904. MIT Press, 2007.

[94] C. Liu, H. Wang, S. Mc Clean, J. Liu, and S. Wu. Syntactic information retrieval. In *Proc. of the Int. Conf. on Granular Computing*, GRC'07, page 703, Washington, DC, USA, 2007. IEEE Computer Society.

[95] L. Maisonnasse, E. Gaussier, and J.-P. Chevallet. Revisiting the dependence language model for information retrieval. In *Proc. of the 30th Int. Conf. on Research and Development in Information Retrieval*, SIGIR'07, pages 695–696, New York, NY, USA, 2007. ACM.

[96] S. Maiti, D.P. Mandal, and P. Mitra. Tackling content spamming with a term weighting scheme. In *Proc. of the Joint Workshop on Multilingual OCR and Analytics for Noisy Unstructured Text Data*, MOCR-AND'11, pages 6:1–6:5, New York, NY, USA, 2011. ACM.

[97] D. Manjula, G. Aghila, and T. V. Geetha. Document knowledge representation using description logics for information extraction and querying. In *Proc. of the Int. Conf. on Information Technology: Computers and Communications*, ITCC'03, page 189, Washington, DC, USA, 2003. IEEE Computer Society.

[98] C.D. Manning and H. Schütze. *Foundations of statistical natural language processing*. MIT Press, Cambridge, MA, USA, 1999.

[99] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schtze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.

[100] S. Mizzaro. The good, the bad, the difficult, and the easy: something wrong with information retrieval evaluation ? In *Proc. of the 30th European Conf. on Information Retrieval*, ECIR'08, pages 642–646, Berlin, Heidelberg, 2008. Springer-Verlag.

[101] S. Mizzaro and S. Robertson. Hits hits TREC: exploring IR evaluation results with network analysis. In *Proc. of the 30th Int. Conf. on Research and Development in Information Retrieval*, SIGIR '07, pages 479–486, New York, NY, USA, 2007. ACM.

[102] M. Montes y Gómez. *Minería de texto empleando la semejanza entre estructuras semánticas*. PhD thesis, Instituto Politécnico Nacional, México D.F., México, 2005.

[103] M. Montes y Gómez, A. López-López, and A. Gelbukh. Information retrieval with conceptual graph matching. In *Proc. of 11th Int. Conf. on Database and Expert Systems Applications*, number 1873 in Lecture Notes in Computer Science, pages 312–321. Springer-Verlag, 2000.

[104] J. Mothe and L. Tanguy. Linguistic analysis of users'queries: Towards an adaptive information retrieval system. In *Proc. of the Third Int. Conf. on Signal-Image Technologies and Internet-Based System*, SITIS'07, pages 77–84, Washington, DC, USA, 2007. IEEE Computer Society.

[105] A. Mowshowitz and A. Kawaguchi. Bias on the web. *Communications of the ACM*, 45:56–60, September 2002.

[106] Marie-Laure Mugnier and Michel Leclère. On querying simple conceptual graphs with negation. *Data Knowl. Eng.*, 60(3):468–493, 2007.

[107] R. Nallapati. Discriminative models for information retrieval. In *Proc. of the 27th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, SIGIR'04, pages 64–71, New York, NY, USA, 2004. ACM.

[108] A. Ntoulas, M. Najork, M. Manasse, and D. Fetterly. Detecting spam web pages through content analysis. In *Proc. of the 15th Int. Conf. on World Wide Web*, WWW'06, pages 83–92, New York, NY, USA, 2006. ACM.

[109] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999. Previous number = SIDL-WP-1999-0120.

[110] J. Pitkow, H. Schütze, T. Cass, R. Cooley, D. Turnbull, A. Edmonds, E. Adar, and T. Breuel. Personalized search. *Communications of the ACM*, 45:50–55, September 2002.

[111] J.M.. Ponte and W.B. Croft. A language modeling approach to information retrieval. In *Proc. of the 21st Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, SIGIR'98, pages 275–281, New York, NY, USA, 1998. ACM.

[112] T. Qin, X.-D. Zhang, M.-F. Tsai, D.-S. Wang, T.-Y. Liu, and H. Li. Query-level loss functions for information retrieval. *Information Processing & Management*, 44:838–855, March 2008.

[113] C. Quiroga-Clare. Language ambiguity: A curse and a blessing. *Translation Journal*, 7(1), 2003.

[114] V. Raghavan, P. Bollmann, and G.S. Jung. A critical investigation of recall and precision as measures of retrieval system performance. *ACM Transactions on Information Systems*, 7:205–229, July 1989.

[115] Philip Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *IJCAI'95: Proceedings of the 14th international joint conference on Artificial intelligence*, pages 448–453, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc.

[116] Francisco J. Ribadas, Manuel Vilares Ferro, and Jesús Vilares Ferro. Semantic similarity between sentences through approximate tree matching. In *IbPRIA (2)*, pages 638–646, 2005.

[117] S. E. Robertson and Sparck K. Jones. Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27(3):129–146, 1976.

[118] Guillaume Rousse and Éric Villemonte de La Clergerie. Analyse automatique de documents botaniques: le projet Biotim. In *proc. of TIA'05*, pages 95–104, Rouen, France, April 2005.

[119] Catherine Roussey. *Une méthode d'indexation sémantique adaptée aux corpus multilingues*. Thèse de doctorat en informatique, INSA de Lyon, December 2001.

[120] B. Sagot. *Analyse automatique du français: lexiques, formalismes, analyseurs*. PhD thesis, Université Paris VII, Paris, France, 2006.

[121] B. Sagot. The lefff, a freely available and large-coverage morphological and syntactic lexicon for french. In *Proceedings of LREC'10*, Valetta, Malta, 2010.

[122] B. Sagot and P. Boullier. Sxpipe 2: architecture pour le traitement pré-syntaxique de corpus bruts. *Traitement Automatique des Langues*, 2(49):155–188, 2008.

[123] Benoît Sagot and Éric Villemonte de La Clergerie. Error mining in parsing results. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 329–336, Sydney, Australia, July 2006. Association for Computational Linguistics.

[124] T. Sakai. On the reliability of information retrieval metrics based on graded relevance. *Information Processing & Management*, 43:531–548, March 2007.

[125] G. Salton, C. Buckley, and C.T. Yu. An evaluation of term dependence models in information retrieval. In *Proc. of the 5th Int. Conf. on Research and Development in Information Retrieval*, SIGIR'82, pages 151–173, New York, NY, USA, 1982. Springer-Verlag New York, Inc.

[126] Gerard Salton. *Automatic text processing: the transformation, analysis, and retrieval of information by computer*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1989.

[127] Gerard Salton. Developments in automatic text retrieval. *Science*, 253:974–979, 1991.

[128] Gerard Salton and Michael J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, NY, USA, 1986.

[129] Gerard Salton, Anita Wong, and Chung-Shu Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975. The paper where vector space model for IR was introduced.

[130] Mark Sanderson and Justin Zobel. Information retrieval system evaluation: effort, sensitivity, and reliability. In *SIGIR*, pages 162–169, 2005.

[131] T. Saracevic. Relevance: A review of the literature and a framework for thinking on the notion in information science. part II: nature and manifestations of relevance. *Journal of the American Society for Information Science and Technology*, 58:1915–1933, November 2007.

[132] J. Seo and J. Jeon. High precision retrieval using relevance-flow graph. In *Proc. of the 32nd Int. Conf. on Research and Development in Information Retrieval*, SIGIR'09, pages 694–695, New York, NY, USA, 2009. ACM.

[133] K. Shaban. *A semantic graph model for text representation and matching in document mining*. PhD thesis, Waterloo, Ont., Canada, 2006.

[134] T.J. Siddiqui. Intelligent techniques for effective information retrieval: a conceptual graph based approach. *SIGIR Forum*, 40(2):73–74, 2006.

[135] I. Soboroff, C. Nicholas, and P. Cahan. Ranking retrieval systems without relevance judgments. In *Proc. of the 24th Int. Conf. on Research and Development in Information Retrieval*, SIGIR'01, pages 66–73, New York, NY, USA, 2001. ACM.

[136] F. Song and W.B. Croft. A general language model for information retrieval. In *Proc. of the 8th Int. Conf. on Information and Knowledge Management*, CIKM'99, pages 316–321, New York, NY, USA, 1999. ACM.

[137] John F. Sowa. Conceptual graphs for a data base interface. *IBM Journal of Research and Development*, 20:336–357, July 1976.

[138] John F. Sowa. Semantics of conceptual graphs. In *Proceedings of the 17th annual meeting on Association for Computational Linguistics*, ACL '79, pages 39–44, Stroudsburg, PA, USA, 1979. Association for Computational Linguistics.

[139] John F. Sowa. *Conceptual Structures: Information Processing in Mind and Machine*. Systems Programming Series. Addison-Wesley, July 1983.

[140] Karen Sparck Jones and C J Van Rijsbergen. Information retrieval test collections. *Journal of Documentation*, 32(1):59–75, 1976.

[141] A. Spink and H. Greisdorf. Regions and levels: measuring and mapping users'relevance judgments. *Journal of the American Society for Information Science and Technology*, 52:161–173, January 2001.

[142] Anselm Spoerri. Using the structure of overlap between search results to rank retrieval systems without relevance judgments. *Inf. Process. Manage.*, 43(4):1059–1070, 2007.

[143] T. Strzalkowski. Natural language information retrieval. *Information Processing & Management*, 31(3):397–417, 1995.

[144] A.-J. Su, Y.C. Hu, A. Kuzmanovic, and C.-K. Koh. How to improve your google ranking: Myths and reality. In *Proc. of the Int. Conf. on Web Intelligence and Intelligent Agent Technology*, volume 1 of *WI-IAT'10*, pages 50–57, Washington, DC, USA, 2010. IEEE Computer Society.

[145] J. Tague-Sutcliffe and J. Blustein. A statistical analysis of the TREC-3 data. In *Overview of the Third Text REtrieval Conference*, TREC-3, pages 385–398, 1994.

[146] U. S. Tiwary and Tanveer Siddiqui. *Natural Language Processing and Information Retrieval*. Oxford University Press, Inc., New York, NY, USA, 2008.

[147] E.G. Traugott. *The Ubiquity of metaphor: metaphor in language and thought*, chapter Conventional and dead metaphors revisited, pages 17–56. Amsterdam studies in the theory and history of linguistic science: Current issues in linguistic theory. J. Benjamins, 1985.

[148] A. Trotman. Learning to rank. *Information Retrieval*, 8:359–381, May 2005.

[149] M.-F. Tsai, T.-Y. Liu, T. Qin, H.-H. Chen, and W.-Y. Ma. Frank: a ranking method with fidelity loss. In *Proc. of the 30th Annual Int. Conf. on Research and Development in Information Retrieval*, SIGIR'07, pages 383–390, New York, NY, USA, 2007. ACM.

[150] C. J. van Rijsbergen. *Information Retrieval*. Butterworth, 1979.

[151] C. J. van Rijsbergen. Another look at the logical uncertainty principle. *Inf. Retr.*, 2:17–26, February 2000.

[152] V.N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1998.

[153] K. Vijay-Shanker and David J. Weir. The equivalence of four extensions of context-free grammars. *Mathematical Systems Theory*, 27(6):511–546, 1994.

[154] É. Villemonte de La Clergerie. DyALog: a tabular logic programming based environment for NLP. In *Proceedings of 2nd International Workshop on Constraint Solving and Language Processing (CSLP'05)*, Barcelona, Spain, October 2005.

[155] É. Villemonte de La Clergerie. From metagrammars to factorized TAG/TIG parsers. In *Proceedings of IWPT'05 (poster)*, pages 190–191, Vancouver, Canada, October 2005.

[156] Ellen M. Voorhees. Trec: Continuing information retrieval's tradition of experimentation. *Commun. ACM*, 50:51–54, November 2007.

[157] Ellen M. Voorhees and Donna Harman. Overview of the sixth text retrieval conference (trec-6). In *TREC*, pages 1–24, 1997.

[158] E.M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing & Management*, 36:697–716, September 2000.

[159] E.M. Voorhees. Overview of the TREC 2004 robust retrieval track. In *In Proc. of the Thirteenth Text REtrieval Conference*, TREC-13, page 13, 2004.

[160] E.M. Voorhees and C. Buckley. The effect of topic set size on retrieval experiment error. In *Proc. of the 25th Int. Conf. on Research and Development in Information Retrieval*, SIGIR'02, pages 316–323, New York, NY, USA, 2002. ACM.

[161] E.M. Voorhees and D. Harman. Overview of the seventh text retrieval conference trec-7. In *Proc. of the Seventh Text REtrieval Conference (TREC-7*, pages 1–24, 1998.

[162] W. Webber, A. Moffat, and J. Zobel. Statistical power in retrieval experimentation. In *Proc. of the 17th Int. Conf. on Information and Knowledge Management*, CIKM '08, pages 571–580, New York, NY, USA, 2008. ACM.

[163] S. Wu and S.I. McClean. Evaluation of system measures for incomplete relevance judgment in IR. In *Proc. of the 7th Int. Conf. on Flexible Query Answering Systems*, pages 245–256, 2006.

[164] Shengli Wu and Fabio Crestani. Methods for ranking information retrieval systems without relevance judgments. In *Proc. of the 2003 ACM Symposium on Applied computing*, SAC'03, pages 811–816, New York, NY, USA, 2003. ACM.

[165] F. Xia, T.-Y. Liu, J. Wang, W. Zhang, and H. Li. Listwise approach to learning to rank: theory and algorithm. In *Proc. of the 25th Int. Conf. on Machine learning*, ICML'08, pages 1192–1199, New York, NY, USA, 2008. ACM.

[166] J. Xu and H. Li. Adarank: a boosting algorithm for information retrieval. In *Proc. of the 30th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, SIGIR'07, pages 391–398, New York, NY, USA, 2007. ACM.

[167] J. Xu, T.-Y. Liu, M. Lu, H. Li, and W.-Y. Ma. Directly optimizing evaluation measures in learning to rank. In *Proc. of the 31st Annual Int. Conf. on Research and Development in Information Retrieval*, SIGIR'08, pages 107–114, New York, NY, USA, 2008. ACM.

[168] X. Yan, R.Y.K. Lau, D. Song, X. Li, and J. Ma. Toward a semantic granularity model for domain-specific information retrieval. *ACM Transanctions on Information Systems*, 29:15:1–15:46, July 2011.

[169] Y. Yue, T. Finley, F. Radlinski, and T. Joachims. A support vector method for optimizing average precision. In *Proc. of the 30th Annual Int. Conf. on Research and Development in Information Retrieval*, SIGIR'07, pages 271–278, New York, NY, USA, 2007. ACM.

[170] Zhaohui Z., Hongyuan Z., Tong Z., Olivier C., Keke C., and Gordon S. A general boosting method and its application to learning ranking functions for web search. In J.C. Platt, D. Koller, Y. Singer, and S.T. Roweis, editors, *Proc. of Advances in Neural Information Processing Systems*, volume 20 of *NIPS'07*, pages 1697–1704. MIT Press, 2007.

[171] Jinglei Zhao and Yeogirl Yun. A proximity language model for information retrieval. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '09, pages 291–298, New York, NY, USA, 2009. ACM.

[172] J. Zobel. How reliable are the results of large-scale information retrieval experiments? In *Proc. of the 21st Int. Conf. on Research and Development in Information Retrieval*, SIGIR'98, pages 307–314, New York, NY, USA, 1998. ACM.

[173] J. Zobel and A. Moffat. Exploring the similarity space. *SIGIR Forum*, 32:18–34, April 1998.