Introduction to the special issue on Non-English Web Retrieval

Fotis Lazarinis[1], Jesus Vilares[2], John Tait[3], Efthimis N. Efthimiadis[4]

[1]Technological Educational Institute, 30200 Mesolonghi, Greece
lazarinf@teimes.gr

[2]Department of Computer Sciences, University of A Coruna
Facultad de Informatica, Campus de Elvina s/n, 15071 - A Coruna (SPAIN)
jvilares@udc.es

[3]Information Retrieval Facility
Palais Eschenbach, Eschenbachgasse 11/3. Stk., 1010 Vienna, Austria
john.tait@ir-facility.org

[4]The Information School, University of Washington
Box 352840, Seattle, WA 98195, USA
efthimis@u.washington.edu

For more than a decade the growth of the level of non-English activity on the web has been noted by many authors (Spink et al 2002; Global Reach 2004; Yang 2005; Kwok 2006; Chung 2008; Miniwatts 2009a; Miniwatts 2009b) and there is no reason to expect the pace of this change to slacken. The pace is likely to increase especially in continents that currently have low Internet penetration. The Web has become a dominant global multicultural and multilingual pool of data. Although in recent years search engines have improved their handling of non-English queries, studies show that many problems still exist and are worthy of further research.

This special issue aims at addressing the challenges and directions in Non-English Web retrieval by providing insights into the existing problems and presenting specific solutions. The call for papers for this special issue was released on February 2008. Twenty-nine papers were received by June 2008 and each was reviewed by three

independent reviewers. After the review process nine papers were accepted for inclusion in the special issue. These studies address various aspects of the special issue topics and concern many non-English languages such as Arabic, Polish, Spanish, Greek, Japanese, Amharic, and a few other European languages.

In the first paper, the special issue editors, Lazarinis, Vilares, Tait and Efthimiadis, provide an overview of the research on non-English search through an extensive literature review. The research issues discussed in these studies are categorized in order to identify the research questions and solutions proposed. Further research is proposed at the end of each section.

Eguchi and Croft use a structured query approach using word-based units to capture compound words, as well as more general phrases, in a query. The paper discusses problems, such as compound words and segmentation that appear in Japanese information retrieval and some research efforts to address these problems.

Knowledge-poor methods for tackling person name matching and lemmatization in Polish, a highly inflectioned language with a complex personal name declension paradigm is discussed in Piskorski, Wieloch and Sydow.

Hammo presents a framework to enhance the retrieval effectiveness of search engines to search for diacritic and diacritic-less Arabic text through query expansion techniques. Query expansion for searching Arabic text is promising, according to the results of the study.

The effect of multilingual queries for homepage finding is studied in Blanco and Lioma, where the aim of their retrieval system is to return a specific homepage. The study reports that Latinized versions of the queries and the local adaptations of the search engines produce better results in many cases.

Efthimiadis, Malevris, Kousaridas, Lepeniotou and Loutas conducted an evaluation using Greek and Latinized homepage finding queries for known Greek organizations. The analysis showed that the global search engines ignore the characteristics of the Greek language, hence treating semantically similar Greek queries differently.

The information-seeking behaviour of non-English Web users is studied in Berendt and Kralisch. The study established that content and link creation behaviour leads to an under-representation of non-English languages in the Web. It also provides evidence that link-following behaviour leads to an under-utilization of non-English content.

Guzman, Montes-y-Gómez, Rosso and Villaseñor-Pineda study the use of the Web as a Spanish linguistic resource for text classification. They retrieved their initial data using Google and they were able to develop a self-training method, which makes use of the Web as a lexical support resource.

Classification of Amharic texts compiled from the Web is discussed in Asker, Argaw, Gambäck, Asfeha and Habte The effect of operations like stemming or part-of-speech tagging on text classification was also investigated. The experiments indicated that stemming plays a less important role than expected for text classification for a highly inflected language like Amharic.

The main conclusion from the special issue papers is that there are still many open research issues for non-English Web search. The papers highlight the need for more research.

Acknowledgements

We would like to thank the editors of the Journal of Information Retrieval W. Hersh, J. Mothe, and J. Zobel, for all their help, as well as, the special issue reviewers (alphabetically):

Mustafa Abusalah, Miguel A. Alonso, Einat Amitay, Ricardo Baeza-Yates, Guillermo Barrutieta, Richard (Rui) Cai, Fazli Can, Raman (Chandra) Chandrasekar, Keh-Jiann Chen, Kuang-hua Chen, Zheng Chen, Theodore Dalamagas, Arjen P. de Vries, Nicola Ferro, Atsushi Fujii, Sumio Fujita, Ayse Goker, Julio Gonzalo, Kalervo Jarvelin, Gareth F. Jones, Ghassan Kanaan, Noriko Kando, Murat Karamuftuoglu, Nikitas Karanikolas, A. Kumaran, Mark Levene, Jimmy J. Lin, Thomas Mandl, Mandar Mitra, Alexandros Ntoulas, Michael Oakes, Iadh Ounis, Gabriel Pereira, Carol Peters, Vassilis Plachouras, Adam Przepiorkowski, Owen Rambow, Satoshi Sekine, Nasredine Semmar, Aya Soffer, Sofia Stamou, Richard F. E. Sutcliffe, Seyed M.M. Tahaghoghi, Vasudeva Varma, Manuel Vilares, Ryen W. White, Fadi Yamout.

References

Chung W. (2008) Web Searching in a Multilingual World. *Communications of the ACM*, 51(5), 32-40.

Global Reach (2004)*, Global Internet Statistics (by Language)* (2004). Available at: www.global-reach.biz/globstats (accessed 31 July 2006).

Kwok, S. H. (2006) P2P searching trends: 2002–2004, *Information Processing & Management*, 42(1), 237-247.

Miniwatts International. (2009) *Internet Usage Statistics: The Internet Big Picture.*

Available at: www.internetworldstats.com/stats.htm (accessed January 6, 2009).

Spink, A. Jansen, B.J. Wolfram, D. Saracevic, T. (2002) From e-sex to e-commerce: Web search changes, *IEEE Computer*, 35(3), 107-109.

Yang, C. C. (2005) Changes in queries in Gnutella peer-to-peer networks, *Journal of Information Science,* 31(2) 124-135.