Diversification of search results by exploiting entity relationships in web search queries

Muthu Kumar C.[1] **Senthil Kumar Chandramohan**[1] **Lee Jia Wei Shaun**[1]
[1] National University of Singapore
{a0092669,,}@nus.edu.sg

Abstract

1 Introduction

Web informational retrieval by commercial search engines focus on satisfying information needs of a majority of its users. So, they focus on retrieving documents relevant to the most popular interpretation(s) of web search queries. They rely on their dynamic query logs to constantly tune their search results to adapt to the changes in perceptions of the user. Given the fairly mature approaches in this field for retrieving relevant documents recent research has focussed on more sophisticated aspects such as addressing the implicit information needs of the user by extrapolating from the explicit information in the search query. Drawing additional information from a user's geographic location and personal profiles of the user collected by encouraging user registration personalization being the most popular extensions to the standard retrieval methods. Although finding the broad geographic location of a user is fairly easy, profiling a user based on his search queries requires logging a considerable number of his queries over longer time periods.

To present a diverse list of search results is one alternative when faced with lack of information specific enough to facilitate accurate retrieval. Presenting a diverse list of documents as the search result of a query enables web search engines to address the information needs of a wider audience. The quantum of diversity in the search results may depend mainly on the interpretations of the named entities in the query. Such interpretations are governed by their ambiguities and facets.

1.1 Types of Diversity

Although ... agree diversity stems from lack of enough information, variety can be introduced into the query results by diversification along different lines. Providing search results corresponding to the unambiguous forms of an ambiguous query gives semantic diversity (Gollapudi and Sharma, 2009). For example, an user issuing a query, 'Jaguar' could intend to find information on 'Jaguar, the wild cat' or 'Jaguar, the sports car'. Query results on different topics of a search query provides topical diversity. For example, an user with a query on 'Barrack Obama', assuming the reference is to the 'Barrack Hussain Obama', President of the United States of America, topical diversity could be provided by search results on 'Barrack Obama on being a Nobel Peace prize winner' or 'Barrack Obama, an American political leader with African lineage'. In our method, we prioritize semantic diversity through disambiguation over topical diversity through categorization.

(Santos et al., 2010c) (Santos et al., 2010b) (Bhatia et al., 2012)

2 Motivation and Related work

(Agrawal et al., 2009) describe an algorithm to provide diverse search results for ambiguous queries. (Santos et al., 2010a) present a framework to identify unspecified information needs of underspecified queries to provide a diverse list of search results. We propose to build on these works by identifying unspecified relations among named entities in a query to diversify the search results. The remainder of this paper is organized as follows.

2.1 Named Entity Recognition in Queries(NERQ)

(Guo et al., 2009) (Pasca, 2007) claim a significant portion of web search queries to contain named entities. (Guo et al., 2009) found that 70% of queries from the query logs of a commercial search engine to contain named entities. Thus recognising named entities becomes valuable in

interpreting the semantics of a web search query. Given such value in recognising named entities, to devise a method to recognise them in web search queries becomes important. To recognise named entity in search queries is different from the traditional task of Named Entity Recognition(NER) in Natural Language Processing. Traditional NER extracts and tags entities to a pre-defined set of entity classes from long coherent textual discourses in documents. The task of Named Entity Recognition in Queries(NERQ) is more complicated. (Guo et al., 2009) found that less than 1% of their sample of web search queries had two or more entities and the average length of a query was 2–3 words. While a standard NER today such as (Finkel et al., 2005) and (Ratinov and Roth, 2009) rely on automatic sequence labelling using Conditional Random Fields(CRF) it fails miserably on the much shorter web search queries. (Bunescu and Pasca, 2006) show a method to extract named entities with the help of encyclopaedia text such as Wikipedia pages using heuristics on the full text content of the pages.

3 Method

Search is verified for ambiguous interpretations (Cucerzan, 2007) through

(Pasca, 2007)

(Auer et al., 2007)

3.1 Ranking reformulations

4 Evaluation

5

5.1 Fonts

Type of Text	Font Size	Style
paper title	15 pt	bold
author names	12 pt	bold
author affiliation	12 pt	
the word “Abstract”	12 pt	bold
section titles	12 pt	bold
document text	11 pt	
captions	11 pt	
abstract text	10 pt	
bibliography	10 pt	
footnotes	9 pt	

Table 1: Font guide.

5.2 The First Page

5.3 Sections

5.4 Footnotes

5.5 Graphics

6 Translation of non-English Terms

7 Length of Submission

8 Other Issues

Acknowledgments

We thank Dr. Ng,Hwee Thou for helping us in choosing our project topic and further advise in defining our problem.

References

Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, and Samuel Ieong. 2009. Diversifying search results. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 5–14. ACM.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer.

Sumit Bhatia, Cliff Brunk, and Prasenjit Mitra. 2012. Analysis and automatic classification of web search queries for diversification requirements. *Proceedings of the American Society for Information Science and Technology*, 49(1):1–10.

Razvan Bunescu and Marius Pasca. 2006. Using encyclopedic knowledge for named entity disambiguation. In *Proceedings of EACL*, volume 6, pages 9–16.

Silviu Cucerzan. 2007. Large-scale named entity disambiguation based on wikipedia data. In *Proceedings of EMNLP-CoNLL*, volume 6, pages 708–716.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 363–370. Association for Computational Linguistics.

Sreenivas Gollapudi and Aneesh Sharma. 2009. An axiomatic approach for result diversification. In *Proceedings of the 18th international conference on World wide web*, pages 381–390. ACM.

Jiafeng Guo, Gu Xu, Xueqi Cheng, and Hang Li. 2009. Named entity recognition in query. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 267–274. ACM.

Marius Pasca. 2007. Weakly-supervised discovery of named entities using web search queries. In *CIKM*, pages 683–690.

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pages 147–155. Association for Computational Linguistics.

Rodrygo LT Santos, Craig Macdonald, and Iadh Ounis. 2010a. Exploiting query reformulations for web search result diversification. In *Proceedings of the 19th international conference on World wide web*, pages 881–890. ACM.

Rodrygo LT Santos, Craig Macdonald, and Iadh Ounis. 2010b. Selectively diversifying web search results. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1179–1188. ACM.

Rodrygo LT Santos, Jie Peng, Craig Macdonald, and Iadh Ounis. 2010c. Explicit search result diversification through sub-queries. In *Advances in information retrieval*, pages 87–99. Springer.