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,a0080003,u0906797}@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 strategize 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 changes in perceptions of the user. Given the fairly mature approaches in this field for retrieving relevant documents, recent research has shifted focus on more sophisticated aspects such as addressing the implicit information needs of the user by extrapolating from the explicit information in the search query. Standard retrieval methods have been extended to draw additional information from a user's geographic location (Lu et al., 2010) and personal profiles of the user (Teevan et al., 2005) collected by encouraging user registration and personalization. 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 there is lack of information specific enough to facilitate retrieval that accurately satisfies the users' information needs. 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 (Bhatia et al., 2012). Inclusion of diversity as a parameter to retrieve and rank search results warrants additional considerations. The traditional probabilistic ranking principle purely based on relevance would be sub-optimal(Gollapudi and Sharma, 2009). More often than not, diversity in search results require a trade-off against relevance.

1.1 Types of Diversity

Although important works in search result diversification such as (Agrawal et al., 2009), (Santos et al., 2010c) agree that diversity stems from lack of enough information, it is important to realise that variety can be introduced into the query results by considering different basis for diversification . (Gollapudi and Sharma, 2009) provide a comprehensive set of axioms, only a proper subset of which any diversification algorithm would satisfy. Providing search results corresponding to the unambiguous forms of an ambiguous query gives semantic diversity. For example, an user issuing a query, 'Jaguar' could intend to find information on 'Jaguar, the wild cat' or 'Jaguar, the sports car' or 'Jaguar the operating system'. Query results on different topics of a search query provides topical diversity. For example, an user with the query 'Barrack Obama', assuming the reference is to the 'Barrack Hussain Obama', President of the United States of America, could be provided with topical diversity through 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.

We note that these various aspects may depend mainly on the interpretations of the named entities in the query. Interpretations are governed by ambiguities and facets of the named entities. Ambiguities and facets of a named entity depends on the level of detail encoded in the query. This leads us to the thought that not all queries need diversification and the quantum of diversity would depend on the query and the entities in the query.

1.2 Named Entity Recognition in Queries(NERQ)

(Guo et al., 2009) (Pasca, 2007) claim that a significant portion of web search queries 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. Dbpedia (Auer et al., 2007) is a resource that harvests structured content in Wikipedia and provides it in machine-readable form. In our methods we extensively use dbpedia to leverage on the crowd-sourced content from Wikipedia to recognise named entities and find diversified versions of the original search query.

2 Related works

(Agrawal et al., 2009) describe an algorithm to provide diverse search results for ambiguous queries. They assume a taxonomy to build their solution. (Santos et al., 2010a) present a framework to identify unspecified information needs of underspecified queries to provide a diverse list of search results. (Cucerzan, 2007) uses Wikipedia (wik, b) to do large scale entity disambiguation. (Hoffart et al., 2011) uses dbpedia among other knowledge bases to disambiguate named entities. We propose to build on these works by identifying unspecified relations among named entities in a query to diversify the search results. In the remainder of this paper we describe our methods in Section 3.

3 Method

As mentioned earlier in Section 1.1, we start with analysing the user search- query form ambiguity. We consider the search query as a whole and check for Wikipedia Disambiguation pages (wik, c) through dbpedia's SPARQL end-point service. If we find the query considered whole is ambiguous, the reformulations of the query are the different disambiguation pages that Wikipedia points to. This part handles the semantic ambiguity of the query.

If the query is not ambiguous, we check if it is a category under wikipedia category hierarchy (wik, a). If it is a category, then we retrieve all the subcategories of the category as possible reformulations. We interpret that these subtopics represent the different sub-topics of the query. This part handles the topical diversity of the search query.

Our method attempts to maximize the utility from the semantics of the complete query before analysing its parts. We aim to include elements of generalization to the overall search results since we assume that the user is not completely aware of his information need and is rather exploring his way through the search. If we fail to detect a match in disambiguation and the category pages, we clean the query of prepositions,infinitives and other stopwords and break them into many bigrams. We again attempt to match the bigrams to the category hierarchy. If there is a match, the reformulation would be to substitute the matched bigram with the *sub-categories* identified category. For example, the query "Obama family tree" would have the bigrams, "Obama family", "family tree", "Obama tree". of these, "Obama family" would match the category page with the same title. We retrieve the following sub-categories,

- dbpedia:Category:Barack_Obama
- dbpedia:Category:Non-free_images_of_Obama_family

The corresponding reformulations would be to clean the strings and append the token tree with each of these category titles. Although the second category title seems less meaningful, we do not filter it at this stage and allow the reformulation.

Until now, we have not considered the named entities in the search query. Our consideration of named entities is limited to place names and person names. Queries that fail to match any of

the above three methods would now be checked for name and place entities. We use the dbpedia lookup service (dbp,) that looks up mirror pages of Wikipedia on dbpedia that stores all the structured content of its peer on Wikipedia. Again we start by checking the query as a whole for entity. Pages in Wikipedia and their corresponding mirror on dbpedia provide class attribute that enables entity class checking. A null value for the class indicates that the page does not represent an entity. We filter only those with the value `http://dbpedia.org/ontology/Place` and `http://dbpedia.org/ontology/Person`. We then generate bigrams of the search query to check for parts of the query that might qualify as an entity. This time the bigrams are limited to those that conserve the original word order in the search query.

We do not attempt to reformulate queries that fail to retrieve reformulations in any of the above methods. We assume such queries are specific enough to not qualify for diversification. (Santos et al., 2010b) argues that diversification is not suitable for every search query and one should abstain from it when not appropriate.

3.1 Ranking reformulations

We then rank the reformulated search queries using the method prescribed by (Santos et al., 2010a) in equations Eq.8 and Eq.9. They use the metric to rank the sub-queries they generate based their relative importance. Since our method also generative similar reformulations we reuse this metric. The metric helps bubble up those reformulations that are meaningful, that is, those reformulations that share some of it top 10 search results with the original search query result would be ranked at the top. We then select the top n reformulations to pick search results to append with the original search query result.

4 Conclusions

We introduce a novel method to use semantic web data sources with a taxonomy of things and identify web search queries or its parts in the taxonomy. We then use the retrieved information from the taxonomy to reformulate the web search queries to direct the information retrieval in multiple directions. These search results from a variety of sub-queries can then be combined to produce a diverse search result.

Acknowledgments

We thank Dr. NG, Hwee Tou 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.

Dbpedia lookup.

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.

Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust disambiguation of named entities in text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 782–792. Association for Computational Linguistics.

Yumao Lu, Fuchun Peng, Xing Wei, and Benoit Dumoulin. 2010. Personalize web search results with user's location. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 763–764. 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.

Jaime Teevan, Susan T Dumais, and Eric Horvitz. 2005. Personalizing search via automated analysis of interests and activities. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 449–456. ACM.

Wikipedia category hierarchy.

Wikipedia, the free encyclopedia.

Wikipedia:disambiguation.