Diversifying for Multiple Information Needs

Rodrygo L.T. Santos and Iadh Ounis

School of Computing Science
University of Glasgow
G12 8QQ, Glasgow, UK
{rodrygo,ounis}@dcs.gla.ac.uk

Abstract. Several approaches have been proposed in recent years to diversify the search results for an ambiguous or underspecified query. In common, most of these approaches are driven by intrinsic characteristics of the search results, such as their content or their coverage of a particular taxonomic scheme. In this position paper, we argue that a true diversification should be driven by the perspective of the search users as opposed to the perspective of the search results. In particular, we claim that an ambiguous query should be regarded as representing multiple possible information needs. The effectiveness of diversifying for multiple information needs is supported by our recent empirical results.

1 Introduction

Query ambiguity is a problem for information retrieval (IR) systems in general, and for web search engines in particular [18]. While an ambiguous query (e.g., ‘jaguar’) is open to multiple *interpretations* (e.g., ‘animal’, ‘car’, ‘guitar’), a query with a clearly defined interpretation (e.g., ‘jaguar car’) may still be underspecified, in that it is open to multiple *aspects* of this interpretation (e.g., ‘dealers’, ‘rental’, ‘insurance’, ‘tuning’, ‘maintenance’, ‘parts’) [9]. An effective approach to tackle query ambiguity is to diversify the search results. By doing so, the chance that different users posing the same query will find at least one relevant result to their particular information need is maximised [6].

Current approaches in the literature seek a diverse ranking by promoting search results that cover multiple aspects[1] of the query or results that cover aspects not well covered by the other results. In common, most of these approaches exploit characteristics of the search results themselves—e.g., their textual content [4] or their coverage of a taxonomy of categories [1]—as surrogates for the actual query aspects. In this position paper, we argue that such an aspect representation only loosely caters for the possible information needs that might have led different users to pose the same query. Instead, we claim that a representation that explicitly aims to encompass multiple information needs is more effective.

In the rest of this paper, Section 2 discusses the limitations of the results-driven diversification performed by existing approaches. Our view of search result diversification as a process driven by users and their multiple possible information needs is detailed in Section 3. We conclude this paper in Section 4.

[1] Unless otherwise noted, we will refer to query interpretations and aspects indistinctly.

2 Opposing Views: Users' vs. Search Results' Diversity

Most diversification approaches in the literature attempt to promote diversity from the perspective of the search results themselves. As illustrated in Figure 1, these approaches derive some representation of the aspects underlying the query from the search results as opposed to the query itself. For instance, novelty-based diversification approaches directly compare the search results to one another without explicitly representing the aspects underlying the query—e.g., based on the search results' textual dissimilarity [4], the divergence of their language models [21], or the correlation of their relevance scores with respect to the initial query [13, 20]. In contrast, coverage-based approaches seek to maximise the search results' coverage of some explicit representation of the aspects underlying the query—e.g., categories from an existing taxonomy [1], or topic models estimated from the search results themselves [5]. In both cases, there is no attempt to account for the multiple possible information needs underlying the query.

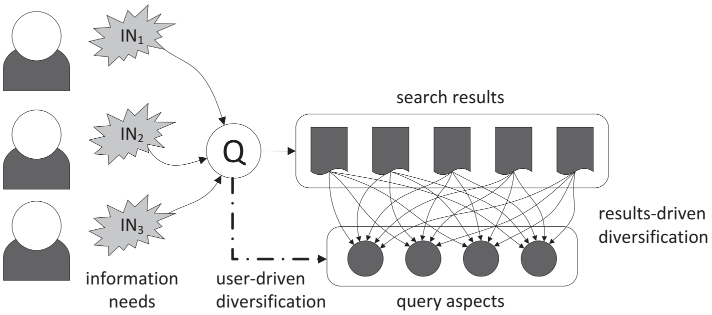

Fig. 1. User- vs. results-driven diversification.

We argue that a results-driven diversification has two key limitations. Firstly, the final ranking can be only as diverse as the aspects identified from the results retrieved for the initial query, which may be biased [12]. As a result, important query aspects (from the user population perspective) may be overlooked simply because they are not well represented among the initial results; conversely, marginally important aspects may be overemphasised. Secondly, the query aspects identified solely based on the search results are a loose surrogate for the actual information needs that may have motivated different users to issue the query in the first place. For instance, search results that cover different topics or categories—or results that are just dissimilar from each other—can feasibly meet the same information need, in which case they would be deemed redundant.

In contrast to promoting diversity from the perspective of the search results, we claim that a user-driven diversification is more effective, as corroborated by our recent empirical results [11, 14–17]. In the next section, we further detail our view of search result diversification in light of multiple possible information needs, and highlight the key areas of investigation involved in this view.

3 Diversifying for Multiple Information Needs

In this section, we detail our view of search result diversification as the problem of satisfying the multiple possible information needs underlying an ambiguous or underspecified query. Although this view is supported by our own successful experiences [11, 14–17], we focus on the principles underlying these experiences rather than on our particular solutions. In Sections 3.1 and 3.2, we describe the building blocks for a general and effective framework for diversifying the search results with respect to multiple information needs, as described in Section 3.3.

3.1 Representing Multiple Information Needs

Inspired by Spärck-Jones et al. [19], we argue that an ambiguous query should be seen as representing an ensemble of possible information needs. The problem then lies in uncovering this ensemble of information needs for a given query. For instance, in a web search scenario, the most natural approach for identifying the possible information needs underlying an ambiguous query is to analyse what previous users that issued the same query were after. Using external resources, such as a query log, one could mine queries related to the initial query, by analysing patterns of query reformulations [2, 14]. On the other hand, the search results themselves could still be leveraged as a resource for a user-driven diversification. In fact, there might be cases when the search results are the most appropriate (or maybe the only available) resource. For instance, in a blog search scenario, multiple information needs could reflect different facets (e.g., left-wing, opinionated, local) of the topic of the query [10], which in turn could be better inferred from the search results for this query. Generally speaking, the suitability of a particular resource for uncovering the possible information needs underlying a query depends on the nature of the diversification task—and hence, of the information needs themselves—at hand.

3.2 Satisfying Individual Information Needs

In order to diversify the search results with respect to the identified information needs, we first need to be able to estimate how well each search result meets every one of the identified information needs. A natural and effective approach is to deploy a ranking model to perform such estimations. As a result, the key step for diversifying the search results for a query becomes to estimate the relevance of each of these results to multiple information needs. The more refined these estimations, the more effective the attained diversification performance. For instance, we have achieved considerable success by leveraging ranking models of various calibres, from traditional document weighting models to learned models based on several features [11, 14–17]. Another important consideration regarding our view of user-driven diversification is that the identified information needs may be rather different from one another, in terms of the underlying intent of the user [3]. As such, these needs may benefit from different features. For instance, while an informational need might benefit from query expansion, a navigational need is more likely to benefit from query analysis features.

3.3 Satisfying Multiple Information Needs

Sections 3.1 and 3.2 described our view for representing the multiple possible information needs underlying an ambiguous or underspecified query, and for satisfying each of the represented information needs individually. The next step for producing a diverse ranking is to integrate these ideas into a unified diversification framework. In particular, such a framework should account for the overall coverage of each search result with respect to the identified information needs, so as to rank highly diverse documents first. Moreover, it should account for how well each information need is covered by the other search results, so as to avoid promoting redundant results [14, 17]. Additionally, another crucial feature of an effective diversification framework is the ability to infer how much emphasis should be placed on each of the identified information needs. For instance, there may be dozens of possible information needs underlying the query. If our goal is to satisfy most users in the first page of results, a bias towards the most important information needs for the user population should be enforced [14, 17]. Finally, an effective diversification framework should also cater for the ambiguity levels of different queries. In particular, not all queries are equally ambiguous. For instance, the query 'jaguar' is arguably more ambiguous than 'jaguar uk dealer locator'. To deal with the specificities of different queries, a diversification framework should be able to automatically decide not only whether, but also how much to diversify the search results on a per-query basis [15].

Altogether, the aforementioned requirements can be naturally mapped into components of a framework for diversifying for multiple information needs. In particular, our xQuAD (Explicit Query Aspect Diversification) framework [11, 14–17] fulfils all these requirements in order to provide a general and effective approach to search result diversification. As a matter of fact, building upon these ideas, xQuAD attained the top performance in the category B of the diversity task of the TREC 2009 and 2010 Web tracks [7, 8].

4 Conclusions

In this paper, we have questioned the effectiveness of search results-driven diversification approaches, and argued for a user-driven diversification instead. In particular, we have detailed our position towards diversifying the search results for multiple information needs, which naturally led to a general framework for search result diversification. Our recent results [11, 14–17] support the stated position, with the described framework attaining a state-of-the-art performance.

Our view of diversification as a user-driven process could be further extended towards satisfying multiple possible information needs across multiple search scenarios (e.g., web, image, news, blogs). In particular, this would open up research directions on several fronts, including the estimation of query ambiguity, the identification and estimation of the likelihood of different information needs, and the estimation of appropriate models for satisfying information needs in different scenarios. As a result, this extended view could form the basis for a holistic approach to search result diversification in an aggregated search scenario.

References

1. Agrawal, R., Gollapudi, S., Halverson, A., Ieong, S.: Diversifying search results. In: Proceedings of WSDM. pp. 5–14 (2009)
2. Baeza-Yates, R.A., Hurtado, C.A., Mendoza, M.: Query recommendation using query logs in search engines. In: EDBT Workshops. pp. 588–596 (2004)
3. Broder, A.: A taxonomy of Web search. SIGIR Forum 36(2), 3–10 (2002)
4. Carbonell, J., Goldstein, J.: The use of MMR, diversity-based reranking for reordering documents and producing summaries. In: Proceedings of SIGIR. pp. 335–336 (1998)
5. Carterette, B., Chandar, P.: Probabilistic models of ranking novel documents for faceted topic retrieval. In: Proceedings of CIKM. pp. 1287–1296 (2009)
6. Chen, H., Karger, D.R.: Less is more: probabilistic models for retrieving fewer relevant documents. In: Proceedings of SIGIR. pp. 429–436 (2006)
7. Clarke, C.L.A., Craswell, N., Soboroff, I.: Overview of the TREC 2009 Web track. In: Proceedings of TREC (2009)
8. Clarke, C.L.A., Craswell, N., Soboroff, I., Cormack, G.V.: Preliminary overview of the TREC 2010 Web track. In: Proceedings of TREC (2010)
9. Clarke, C.L.A., Kolla, M., Cormack, G.V., Vechtomova, O., Ashkan, A., Büttcher, S., MacKinnon, I.: Novelty and diversity in information retrieval evaluation. In: Proceedings of SIGIR. pp. 659–666 (2008)
10. Macdonald, C., Ounis, I., Soboroff, I.: Overview of the TREC 2009 Blog track. In: Proceedings of TREC (2009)
11. McCreadie, R., Macdonald, C., Ounis, I., Peng, J., Santos, R.L.T.: University of Glasgow at TREC 2009: Experiments with Terrier—Blog, Entity, Million Query, Relevance Feedback, and Web tracks. In: Proceedings of TREC (2009)
12. Mowshowitz, A., Kawaguchi, A.: Assessing bias in search engines. Inf. Process. Manage. 38(1), 141–156 (2002)
13. Rafiei, D., Bharat, K., Shukla, A.: Diversifying Web search results. In: Proceedings of WWW. pp. 781–790 (2010)
14. Santos, R.L.T., Macdonald, C., Ounis, I.: Exploiting query reformulations for Web search result diversification. In: Proceedings of WWW. pp. 881–890 (2010)
15. Santos, R.L.T., Macdonald, C., Ounis, I.: Selectively diversifying Web search results. In: Proceedings of CIKM. pp. 1179–1188 (2010)
16. Santos, R.L.T., McCreadie, R., Macdonald, C., Ounis, I.: University of Glasgow at TREC 2010: Experiments with Terrier in Blog and Web tracks. In: Proceedings of TREC (2010)
17. Santos, R.L.T., Peng, J., Macdonald, C., Ounis, I.: Explicit search result diversification through sub-queries. In: Proceedings of ECIR. pp. 87–99 (2010)
18. Song, R., Luo, Z., Nie, J.Y., Yu, Y., Hon, H.W.: Identification of ambiguous queries in Web search. Inf. Process. Manage. 45(2), 216–229 (2009)
19. Spärck-Jones, K., Robertson, S.E., Sanderson, M.: Ambiguous requests: implications for retrieval tests, systems and theories. SIGIR Forum 41(2), 8–17 (2007)
20. Wang, J., Zhu, J.: Portfolio theory of information retrieval. In: Proceedings of SIGIR. pp. 115–122 (2009)
21. Zhai, C., Cohen, W.W., Lafferty, J.: Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In: Proceedings of SIGIR. pp. 10–17 (2003)