Different Strokes for Different Folks: an analysis of similarity and diversity in Web search[1]

Maurice Coyle and Barry Smyth[2]

Abstract. Relying purely on query-page similarity when ranking Web search results limits the scope of the result set to the detriment of search performance. In this paper we propose that introducing diversity into the ranking metric can increase topic coverage without adversely affecting result relevance in the face of vague queries.

1 Introduction

The failure of large-scale commercial search engines to satisfy a user's query at the first time of asking - if at all - and the user frustration that occurs as a consequence is well documented. When faced with a vague or under-specified query Web search engines, and most other types of information retrieval (IR) system, are renowned for returning imprecise and otherwise unsatisfactory result lists [4].

Part of the blame must rest with the users themselves; they rarely look beyond the first page of results and have a tendency to formulate under-specified queries consisting of between 2 and 3 search terms [8]. Coupled with the fact that most commercial search engines index over 1 billion documents, this leads inevitably to large result-lists with poor precision characteristics.

Generally, search engines rank results according to their similarity to the query terms and this, combined with the *ranked-list presentation paradigm* that has been almost universally adopted, can lead to result-lists with low diversity and poor coverage of the information space. For example, for the query 'lisp', the top 60 Google results relate to the programming language Lisp and only a handful of pages in the top 200 relate to other valid topics, none of which contain information on speech impediments. One might argue that this is to be expected, that there is inevitably a predominance of computer related content on the Web. This is undoubtedly true, but it is no longer matched by a corresponding bias among Web searchers and the lack of diversity among these results will frustrate searchers who are interested in the less well represented interpretations of 'lisp'.

The above bias is endemic in modern Web search engines because their ranking metrics take only a local view of result relevance; in short, relevance is computed on the basis of similarity to the query without consideration of any other candidate results that may have already been selected for retrieval.

The problem of how to handle vague queries has been addressed before, where factors other than query-page similarity have been considered when ranking search results [1]. The introduction of a *search context* has been used as a method for disambiguating queries and shows promise as a means for focussing search when faced with vague queries (see [3, 5, 9] for more). An alternative solution involves the clustering of search results ([11, 12]), presenting them as an organised collection of documents rather than a flat list of results. This method, although representing a move away from the widely accepted ranked-list presentation paradigm, has shown that it is capable of producing well-defined clusters of Web documents which help to focus the searcher on the topic at hand.

The case-based reasoning (CBR) community has recently begun to question the similarity assumption inherent in related IR applications such as recommender systems. Their argument is that in many scenarios query-similarity can be sacrificed in favour of improved result diversity to maximise the coverage of the retrieved cases ([6, 7, 10]).

In this paper we investigate the ranking of search results by considering both query-page similarity and result diversity. This technique is adapted from one used in the field of recommender systems [10] and we show that it preserves query similarity while increasing overall result diversity, and that it can even improve precision and recall.

2 Similarity vs. Diversity

We use a ranking metric that uses a measure of similarity and diversity to compute result quality (see Equation 1), assuming a standard query-page similarity metric, $Sim(q, p)$ and also assuming that $Sim(p_i, p_j)$ gives the similarity between pages p_i and p_j. The quality of a page p for a query q is calculated relative to the set of pages already selected, $R = r_1, r_2, .., r_m$ using the equations below:

$$Qual(q, p, R) = Sim(q, p) * RelDiv(p, R) \quad (1)$$

$$\begin{aligned} RelDiv(p, R) &= 1 \ if \ R = \{\}; \quad (2) \\ &= \frac{\sum_{i=1..m}(1 - Sim(p, r_i))}{m}, otherwise \end{aligned}$$

2.1 Bounded Greedy Selection

We implement a bounded greedy selection algorithm similar to that in [10]. This algorithm selects the best bk (where b is a bound used to limit algorithm complexity and k is our desired result-list size) pages according to their query-similarity. Pages are then iteratively selected by choosing the page with the highest *quality* (see Equation 1) on each iteration. The first selected page is always the most similar to the query but the remainder depend on query similarity and their similarity to previous selections. The result is a set of pages that are similar to the query but different in terms of content from each other.

This algorithm has a small selection cost since k pages are selected from bk pages instead of from n (where n is the size of the initial result list) pages and $bk \ll n$ for typical low values of b and k.

[1] The support of the informatics initiative of Enterprise Ireland is gratefully acknowledged.

[2] Smart Media Institute, University College Dublin, Dublin, Ireland, email: firstname.lastname@ucd.ie

Since we do not examine all pages, we may miss a page with a marginally lower similarity value than the best bk pages but a significantly better diversity value. However, the likelihood of this decreases with page similarity, so for suitable values of b it is unlikely.

[10] shows that the bounded greedy selection algorithm offers a good combination of diversity and computational efficiency, at least in CBR systems. Of course here we are interested in Web search and in our evaluation we investigate whether the advantages of this diversity preserving technique transfer into the Web search context.

3 Evaluation

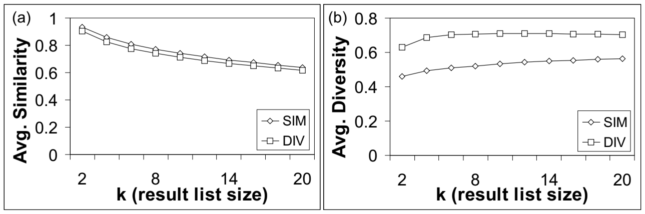

Figure 1. (a) Query-similarity profile, (b) Avg. diversity profile for SIM and DIV at various result list sizes

For our evaluation, we implemented one search engine which relied on pure similarity-based ranking (SIM) and another which used the diversity-enhancing technique (DIV). These engines were used to rank the results for around 670 queries from a variety of domains ([2]). For each query, a set of relevant results was obtained as follows. The query was submitted to the HotBot search engine and the top 1000 results were retrieved. The top 1000 results for a contextualised version of each query (e.g. 'jaguar mammal' for the query 'jaguar') were also retrieved. The set of relevant results for the non-contextualised query was taken to be the intersection of these two result lists. Using these sets of relevant results, the number of relevant results returned for different list sizes was calculated and precision and recall characteristics were measured for each query.

The similarity and diversity profiles for each engine's result lists can be seen in Figure 1. As expected, increasing the diversity of a result-list leads to a drop in query-page similarity. The thing to note here is the difference between the magnitude of the drop in similarity versus the increase in diversity. The minor drop in similarity experienced by DIV is accompanied by a large increase in result diversity.

To investigate whether the enhanced-diversity result lists have an effect on the overall relevance of the result lists, we calculated precision and recall values for each result list using the pre-computed set of relevant results mentioned above. The results of this analysis can be seen in Figure 2, for the mammals and travel domains. The graphs for the other 3 domains are qualitatively similar. The significance of these results is that where a drop in result list relevance might be expected for DIV, no such drop is experienced. The fact that DIV's precision and recall characteristics are better than those of SIM for these experiments is an added bonus but may not be reliable.

4 Conclusions

We conclude from the above evaluation that the diversity-enhancing ranking strategy provides us with the benefits of diverse search results, such as increased topic coverage, with minor sacrifices in query-page similarity. Furthermore, the DIV engine provided precision and recall characteristics that were better than those provided by the SIM engine where the same or slightly worse performance would be acceptable, although this may not be statistically significant.

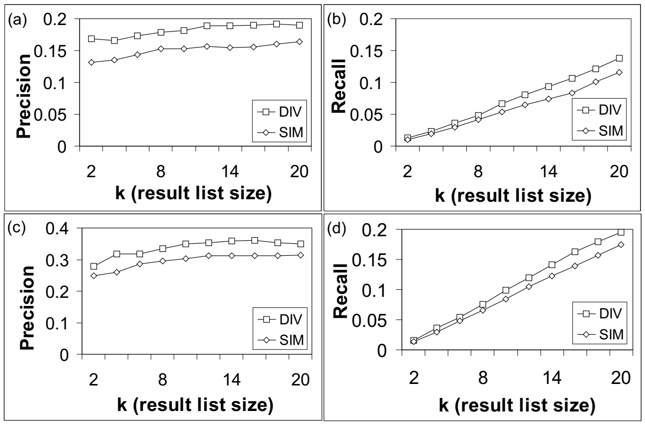

Figure 2. (a) Precision results for the mammals domain, (b) Recall results for the mammals domain, (c) Precision results for the travel domain, (d) Recall results for the travel domain

REFERENCES

[1] S. Brin and L. Page, 'The Anatomy of A Large-Scale Hypertextual Web Search Engine.', in *Proceedings of the Seventh International World-Wide Web Conference*, (2001).

[2] J. Freyne, B. Smyth, M. Coyle, P. Briggs, and E. Balfe, 'Further Experiments in Collaborative Ranking in Community-Based Web Search.', *AI Review: An International Science and Engineering Journal (In Press)*, (2004).

[3] E. Glover, S. Lawrence, M. D. Gordon, W. P. Birmingham, and C. Lee Giles, 'Web Search - Your Way', *Communications of the ACM*, **44**(12), 97–102, (2000).

[4] Robert Krovetz and W. Bruce Croft, 'Lexical ambiguity and information retrieval', *Information Systems*, **10**(2), 115–141, (1992).

[5] S. Lawrence, 'Context in Web Search', *IEEE Data Engineering Bulletin*, **23(3)**, 25–32, (2000).

[6] D. McSherry, 'Diversity-Conscious Retrieval.', in *Proceedings of the Sixth European Conference on Case-Based Reasoning (ECCBR 2002)*, ed., Susan Craw, pp. 219–233. Springer, (2002). Aberdeen, Scotland.

[7] H. Shimazu, 'ExpertClerk : Navigating Shoppers' Buying Process with the Combination of Asking and Proposing.', in *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI-2001)*, ed., Bernhard Nebel, pp. pages 1443–1448. Morgan Kaufmann, (2001). Seattle, Washington, USA.

[8] Craig Silverstein, Monika Henzinger, Hannes Marais, and Michael Moricz, 'Analysis of a Very Large AltaVista Query Log', Technical Report 1998-014, Digital SRC, (1998). http://gatekeeper.dec.com/pub/DEC/SRC/technical-notes/abstracts/src-tn-1998-014.html.

[9] B. Smyth, E. Balfe, P. Briggs, M. Coyle, and J. Freyne, 'I-SPY - Anonymous, Community-Based Personalization by Collaborative Meta-Search', in *Proceedings of the 23rd SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence*. Springer-Verlag, (2003). Cambridge, UK. accepted for publication.

[10] B. Smyth and P. McClave, 'Similarity vs. Diversity.', in *Proceedings of the International Conference on Case-Based Reasoning*, eds., D. Aha and I. Watson, pp. 347–361. Springer, (2001).

[11] Yitong Wang and Masaru Kitsuregawa, 'Link based clustering of Web search results', in *Advances in Web-Age Information Management, Second International Conference, WAIM 2001*, volume 2118, pp. 225–236, (2001). Xi'an, China.

[12] Oren Zamir and Oren Etzioni, 'Grouper: a dynamic clustering interface to Web search results', *Computer Networks (Amsterdam, Netherlands: 1999)*, **31**(11–16), 1361–1374, (1999).