ON THE IMPORTANCE OF BEING DIVERSE:

ANALYSING SIMILARITY AND DIVERSITY IN WEB SEARCH

Maurice Coyle and Barry Smyth*
Smart Media Institute
University College Dublin, Dublin, Ireland
firstname.lastname@ucd.ie

Abstract We argue that the emphasis normally placed on query-similarity in Web search limits search precision. We draw on related work in case-based reasoning (CBR) and recommender systems research, which shows how enhancing diversity can improve the quality of retrieved cases and recommendations. We investigate the use of related diversity-enhancing retrieval techniques in Web search, showing that similar benefits are available, i.e. that result diversity can be significantly enhanced without compromising query similarity or result precision and recall.

Keywords: Web search, diversity, relevance, topic coverage

Introduction

Web search engines are the primary tool for online information discovery and significant strides have been made to build upon their information retrieval (IR) origins in order to address the specific needs of Web users. Nevertheless, search engines frequently fail to deliver the right results at the right time.

It has been shown that users have a tendency to formulate under-specified queries consisting of between 2 and 3 search terms [15]. This coupled with the fact that most commercial search engines index over 1 billion documents leads to large result-lists with poor precision characteristics. Most search engines rank search results according to their similarity to the query terms and this can lead to result-lists with low diversity and poor topic coverage.

As an example, the first 200 Google results for the intentionally vague query 'lisp' all refer to the Lisp programming language with only a few references to other meanings, none of which refer to speech impediments. With a predominance of computer-related information on the Web, it's not hard to see why this

*The support of the Informatics Research Initiative of Enterprise Ireland is gratefully acknowledged

is the case, but with an increasing number of online users from non-computing backgrounds, there is no longer a corresponding bias among Web searchers.

The point is that given vague queries, result diversity in modern search engines is poor which will inevitably lead to search failures – a speech therapist will not be served well by a typical search engine for the term 'lisp' and will be left with no choice but to refine their query. Thus, researchers have focused their efforts on a number of different possible approaches. Recently, ranking metrics have been developed using factors other than query-page similarity [3].

In this paper we focus on diversity among search engine results for vague queries. Research in the areas of CBR and recommender systems has begun to question the similarity assumption, arguing that in many scenarios query-similarity can be sacrificed in favour of improved result diversity in order to maximise the coverage of the retrieved cases. A successful solution has involved a ranking metric incorporating diversity as well as similarity, rather than attempting to elaborate the queries or change the result presentation paradigm.

We adapt this diversity-enhanced approach for use in Web search and evaluate its performance on a range of test data (Sections 3 and 4). We show that the technique introduces result diversity without compromising overall query-page similarity or precision and recall characteristics of the result-lists (see Section 4). First, we will review a range of related research, covering context-sensitive search methods, result-clustering, and diversity-enhancing techniques.

1. Background

1.1 Related Work

In related work, *search context* was introduced to elaborate vague queries and focus search [10] - this encompasses *explicit context manipulation* ([6, 11]) and *implicit context inference* ([4, 7]). The context-sensitive technique yielded promising results indicating that Web search can benefit from its use ([16, 17]).

1.2 Results Clustering

The IR community uses clustering both as a pre-retrieval process to speed up search performance [19] and as a post-retrieval document browsing technique for handling vague queries [5, 12]; it is the second paradigm that concerns us.

[21] and [22] are examples of early work on result clustering. A technique called *suffix tree clustering* (STC) is introduced which shows potential as a means of generating meaningful clusters. A fuzzy similarity metric is proposed in [8] as part of a relational fuzzy clustering algorithm that is $O(n^2)$ (STC is $O(n)$), apparently capable of producing more focused clusters than STC.

It is also worth mentioning [20] for their approach to clustering using connectivity information rather than textual content.

1.3 Towards Diversity-Enhanced Retrieval

The above strategies help users to find information following a vague query. However, they place different obligations on both search engine and searcher and move away from the accepted *ranked list presentation paradigm.*

A number of case retrieval systems have concentrated on improving the diversity of a single set of recommendations while preserving the query-similarity of these recommendations to a lesser or greater extent. [14] introduces a system focused on diversity, however although recommendations are maximally diverse from each other, query-similarity is compromised. Thus, the candidate cases must be sufficiently similar to the query to begin with.

[13] introduces *similarity layers* and *similarity intervals.* The former preserve case-query similarity while enhancing diversity and the latter achieve greater diversity by relaxing the constraint that query similarity must be preserved. It is worth noting that a retrieval technique may enhance diversity as a side-effect. *Order-based retrieval* is an example of such a technique [2], exhibiting an inherent ability to enhance the diversity of a set of retrieval results.

The above techniques are designed for use in case retrieval scenarios and as such it is not clear how they may be adapted for Web search. However, one of the earliest proposals for diversity-enhanced retrieval ([1],[18]) is sufficiently general for it to be directly applied to Web search. This technique is described in detail below and serves as the focus for the remainder of this paper.

2. The Case for Diversity in Web Search

The average Web search is unlikely to result in a focused list of relevant results [9] and Web users are unlikely to venture beyond the first results page [15]. Thus, search engines must maximise the probability that a relevant result will be presented within the first page. Furthermore without any assessment of user preferences or search context, it is valuable to ensure that the first k search results reflect a representative sample of as many relevant results as possible.

In the next section we describe the Bounded Greedy Selection technique first introduced by [18]. We will argue that it provides a reasonable balance between similarity and diversity with only a small extra computational cost.

3. Similarity vs. Diversity

We assume a standard similarity function for computing the similarity between a search query, q, and a page p_i, $Sim(q, p_i)$. Further, we assume that this function can also measure the similarity between two pages, $Sim(p_i, p_j)$.

$$Div(p_1, ..., p_n) = \frac{\sum_{i=1..n} \sum_{j=i..n} (1 - Sim(p_i, p_j))}{\binom{n}{2} * (n-1)} \quad (1)$$

We define the diversity of a set of pages, $p_1, ..., p_n$ to be the average dissimilarity between all pairs of pages in this set (see Equation 1). Standard search engines will tend to display a diversity profile which increases and a similarity profile which decreases, as result-list size increases (see Section 4.2). Thus the trade-off between query-similarity and result-diversity is a simple one: for small result-lists, high query-similarity means low diversity. We aim to optimise this trade-off, delivering result-lists that are diverse and that thus offer greater coverage of the result-space, without compromising their similarity to the query or their relevance to the end-user.

Table 1a. Greedy Algorithm.

q: target query, P: set of pages matching q, k: # results

```
1. define GreedySelection(q,P,k)
2. begin
3. R:={}
4. For i := 1 to k
5. Sort P by Qual(q, p, R) ∀ p in P
6. R := R + First(P)
7. P := P - First(P)
8. EndFor
9. return R
10. end
```

Table 1b. Bounded Greedy Algorithm.

q, P, k: as in Table 1a, b: bound

```
1. define BoundedGreedySelection(q,P,k,b)
2. begin
3. P':=bk pages in P most similar to q
4. R:={}
5. For i := 1 to k
6. Sort P' by Qual(q, p, R) ∀ p in P'
7. R := R + First(P')
8. P' := P' - First(P')
9. EndFor
10. return R
11. end
```

3.1 Greedy Selection

A novel approach to improving diversity, while at the same time maintaining similarity, is to explicitly consider both diversity and similarity during retrieval [18]. The *greedy selection* algorithm (Table 1a) achieves this by incrementally building a final result-list, R. During each step the remaining pages are ordered according to their *quality* with the highest quality page added to R.

The quality (see Equation 2) of a page p is proportional to the similarity between p and the current query q, and to the diversity of p *relative* to those pages so far selected, $R = \{r_1, ..., r_m\}$ (see Equation 3). The first page to be selected is always the one with the highest similarity to the query. During each iteration, the page with the highest quality value is selected.

$$Qual(q, p, R) = Sim(q, p) * RelDiv(p, R) \tag{2}$$

$$\begin{aligned} RelDiv(p, R) &= 1 \quad if\ R = \{\}; \\ &= \frac{\sum_{i=1..m}(1 - Sim(p, r_i))}{m} \quad , otherwise \end{aligned} \tag{3}$$

However, this algorithm is expensive. For an initial result-list of n pages, during each of the k iterations we must calculate the diversity of each remaining page relative to those so far selected. This means an average of $\frac{n-k}{2}$ relative diversity calculations, each one consisting of an average of $\frac{k}{2}$ similarity calculations. This gives an average total cost of $k * \frac{n-k}{2} * \frac{k}{2}$ similarity computations per retrieval. For example, for an initial result-list of 1000 pages, retrieving the top 3 pages can mean approximately 2250 similarity computations.

3.2 Bounded Greedy Selection

To reduce the complexity of the greedy selection algorithm we implement a bounded version adapted from that found in [18]. The *bounded greedy selection* algorithm (Table 1b) selects the best bk pages using their query-similarity (line 3) and then applies the greedy selection method to these (lines 4 - 9).

This algorithm has a greatly reduced cost since k pages are selected from bk pages instead of from n pages and $bk \ll n$ for typical low values of b and k. This only means a total of $k * \frac{k(b-1)}{2} * \frac{k}{2}$ extra similarity computations on top of the normal retrieval cost. For example, for a 1000 page initial result-list, retrieving the 3 best pages with $b = 2$ will now require about 7 extra similarity computations on top of the standard similarity-based retrieval cost.

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.

[18] shows that the bounded greedy algorithm offers the best combination of diversity and efficiency, at least in CBR systems. 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.

4. Evaluation

In this section we describe a recent evaluation to investigate this diversity-conscious ranking strategy. We compare the similarity-based and diversity-based methods and focus on their diversity and similarity characteristics, the degree of re-ordering that takes place as a result of introducing diversity and the effects of this on precision and recall characteristics.

4.1 Set-up

We produced 760 separate queries taken from 5 distinct topical domains (mammals, programming languages, researchers, computer science and travel). We also produced two search engines based on the Jakarta Lucene search engine. The *SIM* version of Lucene used standard similarity-based retrieval with TF*IDF term weighting, and corresponds to a standard Web search engine.

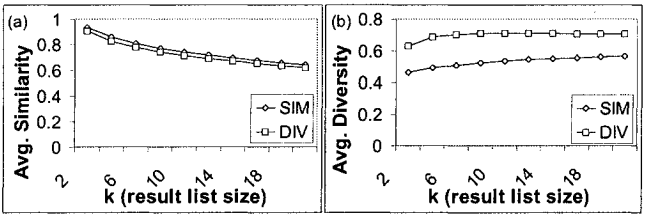

Figure 1. (a) Query-similarity profile, (b) Avg. diversity profile for SIM and DIV at various result list sizes

The *DIV* version was set up to incorporate the diversity-enhancing retrieval technique on top of the TF*IDF functions. Thus, for each query we were able to generate and compare result-lists of varying sizes for SIM and DIV.

Next we needed to populate our test search engines with a collection of Web pages and we also needed to establish a set of relevant pages for each query. To do this we adopted a similar approach to that reported by [16, 17]. Specifically, a basic or *non-contextualised* (e.g. 'java') and a *contextualised* (e.g. 'programming language java') version of each query was submitted to the HotBot search engine and the top 1000 results retrieved. To determine which results for the basic query were relevant, the intersection between the 2 lists for each query was taken. Thus, we had a list of relevant results for each query which was used to assess the precision and recall of the result-lists produced for the basic queries by the SIM and DIV search engines.

Finally, an index was created from the candidate result-lists produced for each of the queries, producing an index of approximately 250,000 pages.

4.2 The Similarity-Diversity Tradeoff

In this first experiment we evaluate the similarity-diversity trade-off – the degree to which query-similarity is compromised as we introduce diversity. We do this by submitting each query to the SIM and DIV (with $b = 4$) search engines to produce results-lists of various sizes, for $k = 2...20$. For each result-list produced by each search engine, we compute its average similarity (i.e., the average similarity between its results and the current query) and its average diversity (i.e., the average pairwise dissimilarity between its results).

The results are shown in Figure 1(a&b) as graphs of average similarity and diversity vs. result-list size. As expected, the diversity-enhanced technique used by DIV leads to a drop in query-similarity when compared to SIM. For example, the average similarity for DIV drops from 0.905 at $k = 2$ to 0.617 at $k = 20$ whereas for SIM it starts at 0.93 at $k = 2$ and falls to 0.639 at $k = 20$. So for different values of k there is only around a 3% drop in average query-similarity for the result-lists produced by DIV compared to those of SIM.

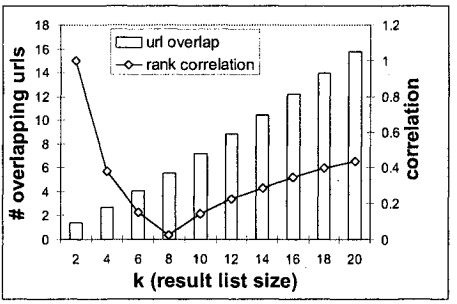

Figure 2. Rank correlation and result overlap characteristics

Also as expected, in Figure 1(b) the advantage goes to DIV, which offers result-lists with significant diversity increases compared to those offered by SIM. For instance, the average diversity for DIV remains stable at approximately 0.7 for k>=4. This is in contrast with the average diversity of the SIM result-lists, which starts at 0.46 at $k = 2$ and grows to only 0.56 at $k = 20$.

The thing to note here is the difference between the scale of the drop in similarity versus the increase in diversity. A minor drop in query-similarity experienced by DIV is accompanied by a significant increase in result diversity.

4.3 A Comparison Of Rankings

Comparing the result-list produced by SIM, for a given k, to that produced by DIV for the same k should illustrate two things. First, DIV will have dropped some of SIM's results in favour of new, more diverse results from outside of SIM's top k results. Second, the results they have in common should be ordered differently to reflect their different quality contributions. Here we evaluate the extent to which this is happening by comparing result-lists from SIM and DIV and measuring the number of results that they have in common and their rank correlation (Spearman's rank correlation is used).

The results are shown in Figure 2 as graphs of overlap and rank correlation against result-list size. As expected, the number of shared results between SIM and DIV increases with k. To begin with, at $k = 4$ SIM and DIV share, on average, about 2.7 results and this increases to nearly 16 results at $k = 20$. Interestingly, this indicates that the percentage overlap grows slowly across the values of k, from an percentage overlap of about 66% at $k = 4$ to 79% at $k = 20$. Thus, on average the diversity-enhancing technique (at $b = 4$) tends to drop approximately 20% to 35% of SIM's top k results in producing its own top k diverse results. This percentage is at a suitably low level that we are not relying too heavily on diversity and not enough on query-page similarity.

The rank correlation results are also interesting. The rank correlation is seen to drop rapidly as k increases initially but then begins to increase slowly again

beyond $k = 8$. For example, at $k = 4$ the rank correlation is 0.37 and this falls to near-zero at $k = 8$ before rising again to 0.43 at $k = 20$. The higher correlation values at low values of k are probably a reflection of the small result-set sizes which will limit the reordering possibilities. Nevertheless, the low correlation values noted across the different values of k indicate that there is a considerable order difference between the shared results in SIM and DIV.

4.4 Precision vs. Recall

We have shown that the benefits of more diverse result lists can be enjoyed without overly compromising the query-similarity of the selected results. However, if increasing diversity in the hope of improving result coverage reduces the precision and recall characteristics of the result-lists (where precision is the proportion of retrieved results that are relevant and recall is the proportion of relevant results that have been retrieved) then our approach is unlikely to bear fruit in practice. Here we consider this issue directly by estimating the accuracy of the SIM and DIV result-lists, in terms of precision and recall estimates on the generated result lists, using the relevant results identified earlier.

The precision and recall results, graphed against k (result-list size), are presented in Figure 3 for the mammals and travel domains. Each data-point represents the mean precision or recall results for either SIM or DIV ($b = 4$) calculated across all queries for the specific domain. The obvious point about these results is that they indicate an improvement in both precision and recall for DIV when compared to SIM. For example, in Figure 3(c) we see that SIM achieves an average precision score of 0.25 at $k = 2$ and that this grows to 0.31 at $k = 20$. In contrast, the same graph indicates that the DIV method achieves an average precision of just under 0.28 at $k = 2$, growing to around 0.35 at $k = 20$. For all result-list sizes we find that the precision characteristics of DIV represent improvements of between 12% and 23% over SIM.

In Figure 3(d) we find that DIV enjoys a similar benefit when it comes to recall. At $k = 2$ both SIM and DIV offer recall of just over 0.01 (actually 0.13 for SIM and 0.16 for DIV) but by $k = 20$ DIV's recall has grown to just under 0.20 whilst SIM has achieved only 0.17. For all values of k this means that DIV benefits from an improvement in recall over SIM by between 12% and 24%. Similar results can be seen for the mammals domain in Figure 3(a&b).

The significance of these results is based on the fact that DIV does not result in a drop in precision and recall – this was always a danger given that there is a reduction in query-similarity.

5. Conclusions

Most search engines rely mainly on query-similarity when it comes to selecting and ordering search results. This often leads to a lack of diversity within

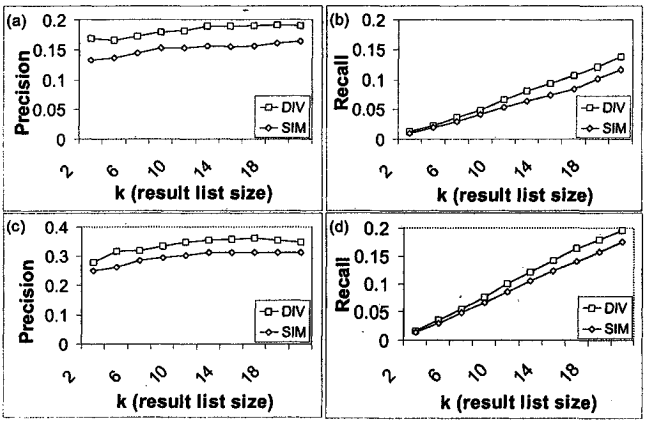

Figure 3. (a) Precision results for the mammals domain, (b) Recall results for the mammals domain, (c) Precision results for the travel domain, (d) Recall results for the travel domain

result-lists where the top-scoring documents may be very similar to the query but very similar to each other as well. A user looking for information on a different topic may need to sift through many similar but irrelevant results.

We have proposed a solution to this problem that employs the standard ranked-list presentation paradigm of today's search engines and that is general enough to work with all search engines that rank results according to a well-defined similarity metric. It calls for the introduction of diversity when it comes to selecting and ranking search results. This diversity-enhancing algorithm is efficient and effective, leading to significant increases in diversity and relatively minor compromises in query similarity. It reorders search results to maximise result diversity as well as query similarity and initial experiments indicate that it does not compromise precision and recall characteristics.

References

[1] K. Bradley and B. Smyth "Improving Recommendation Diversity", Proceedings of the 12th National Conference in Artificial Intelligence and Cognitive Science (AICS-01), pp. 75-84, Maynooth, Ireland, 2001.

[2] D. Bridge and A. Ferguson "Diverse Product Recommendations using an Expressive Language for Case Retrieval", Proceedings of the 16th European Conference on Case-Based Reasoning, pp. 43-57, 2002.

[3] S. Brin and L. Page "The Anatomy of A Large-Scale Hypertextual Web Search Engine" , Proceedings of the 7th International World-Wide Web Conference, 2001.

[4] J. Budzik and K. Hammond "User Interactions with Everyday Applications as Context for Just-in-time Information Access", Proceedings of the International Conference on Intelligent User Interfaces, pp. 44-51, ACM Press, 2000.

[5] D. R. Cutting and D. R. Karger and J. O. Pedersen and J. W. Tukey "Scatter Gather: a cluster-based approach to browsing large document collections", Proceedings of

the 15h International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 318-29, ACM Press, 1992.

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

[7] T. H. Haveliwala "Topic-Sensitive PageRank", Proceedings of the 11th World-Wide Web Conference, ACM Press, 2002.

[8] Z. Jiang and A. Joshi and R. Krishnapuram and L. Yi "Retriever: Improving Web Search Engine Results Using Clustering", Managing Business with Electronic Commerce: Issues and Trends, Idea Press, 2001.

[9] R. Krovetz and W. B. Croft "Lexical Ambiguity and Information Retrieval", Information Systems, 10(2), pp. 115-141, 1992.

[10] S. Lawrence "Context in Web Search", IEEE Data Engineering Bulletin, 23(3), pp. 25-32, 2000.

[11] S. Lawrence and C. Lee Giles "Searching the Web: General and Scientific Information Access", IEEE Communications 37(1), pp. 116-122, 1999.

[12] A. Leouski and W. Croft "An Evaluation of Techniques for Clustering Search Results", Technical Report IR-76, Department of Computer Science, University of Massachusetts, Amherst, 1996.

[13] D. McSherry "Diversity-Conscious Retrieval", Proceedings of the 6th European Conference on Case-Based Reasoning, pp. 219-233, Aberdeen, Scotland, 2002.

[14] H. Shimazu "ExpertClerk: Navigating Shoppers' Buying Process with the Combination of Asking and Proposing", Proceedings of the 17th International Joint Conference on Artificial Intelligence, pp. 1443-1448, Seattle, Washington, USA, 2001.

[15] C. Silverstein and M. Henzinger and H. Marais and M. Moricz "Analysis of a Very Large AltaVista Query Log", Technical Report 1998-014, Digital SRC Technical Notes http://gatekeeper.dec.com/pub/DEC/SRC/technical-notes/abstracts/src-tn-1998-014.html, 1998.

[16] B. Smyth and E. Balfe and P. Briggs and M. Coyle and J. Freyne "Collaborative Web Search", Proceedings of the 18th International Joint Conference on Artificial Intelligence, pp. 1417-1419, Acapulco, Mexico, 2003.

[17] B. Smyth and E. Balfe and P. Briggs and M. Coyle and J. Freyne "I-SPY - Anonymous, Community-based Personalization by Collaborative Meta-search", Proceedings of the 23rd SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence, Cambridge, UK, 2003.

[18] B. Smyth and P. McClave "Similarity vs. Diversity", Proceedings of the 4th International Conference on Case-Based Reasoning, Vancouver, Canada, 2001.

[19] C. J. van Rijsbergen "Information Retrieval, 2nd Edition", Department of Computer Science, University of Glasgow, 1979.

[20] Y. Wang and M. Kitsuregawa "Link-based Clustering of Web Search Results", Lecture Notes in Computer Science, Advances in Web-Age Information Management, Second International Conference, 2118, pp. 225-237, WAIM 2001.

[21] O. Zamir and O. Etzioni "Web Document Clustering: A Feasibility Demonstration", Research and Development in Information Retrieval, pp. 46-54, 1998.

[22] O. Zamir and O. Etzioni "Grouper: A Dynamic Clustering Interface to Web Search Results", Computer Networks, 31(11-16), pp. 1361-1374, Amsterdam, Netherlands, 1999.