Multidimensional Search Result Diversification: Diverse Search Results for Diverse Users

Sumit Bhatia
Computer Science and Engineering
The Pennsylvania State University
University Park, PA-16802, USA
sumit@cse.psu.edu

1. INTRODUCTION

One major assumption underlying most of the document retrieval models is of **independent document relevance**, i.e., relevance of a document given a query is independent of the (ir)relevance of any other document in the corpus. This assumption, however, is an oversimplification and leads to many problems in IR. Since documents are analyzed independently for computing relevance, many times two highly similar and relevant documents are both presented to the user in top 10 search results. This is not desirable as no new information is being provided to the user by the duplicate documents. Thus, there is a need to diversify search results so as to maximize novel information and minimize redundancy in top-k documents.

In addition to the need of minimizing the redundant results from the search result list presented to the user, there is also a need to cater to the vast and diverse user population. Hundreds of millions of people today rely on Web based Search Engines (WSEs) to satisfy their information needs. Many of the users issuing the same query to the search engine may have varying backgrounds, information needs, context etc. yet they use the same terms to indicate their information needs to the search engine. In order to meet the expectations of this vast and diverse user population, the search engine should present a list of results such that the probability of satisfying the *average user* is maximized [1]. This leads us to the problem of **Search Result Diversification**. Given a user submitted query, the search engine should include results that are relevant to the user query and at the same time, *diverse* enough to meet the expectations of diverse user populations. However, *it is not clear in what respect the results should be diversified.* Much of the current work in diversity [1, 3, 4, 11, 15] focuses on *ambiguous and underspecified queries* and tries to include results corresponding to diverse interpretations of the ambiguous query. However, this is not always sufficient. In my analysis of a commercial web search engine's logs (Section 3), I found that even for well-specified informational queries, click entropy is very high indicating that different users prefer different types of documents. Very recently, a diversification algorithm fine-tuned for such informational queries has been proposed [18]. Further, high click entropies were also observed for a large fraction of transactional queries. Current state-of-the-art diversification algorithms that are optimized for sub-topic coverage are not suited for this scenario as the user goal here is to *perform a transaction.*

One major goal of my PhD thesis will then be to identify the various possible dimensions along which the search results can be diversified. Further, we require techniques that given a user entered query can identify what diversity dimensions are best suited for the query. Given such explicit diversity requirements for the query, appropriate algorithms can then be used to diversify search results for the query. The remainder of this paper discusses these issues in further detail (Sections 3 and 4).

2. RELATED WORK

Maximum Marginal Relevance (MMR) [2] introduced by Carbonell and Goldstein represents one of the earliest attempts for search result diversification. For a given user query MMR selects documents that are relevant to the user query as well as provide novel information when compared to previously selected documents. Chen and Karger [4] argue that the strategy of returning as many relevant results as possible (the *Probability Ranking Principle (PRP)*) is not always optimal. Hence they put forward the idea of returning a set of documents that maximizes the probability of finding a relevant document in top-k documents. Agrawal et al. [1] study the problem of diversifying search results of ambiguous web queries. They assume the availability of a taxonomy of information and that both queries and documents may belong to one or more categories in this taxonomy. The problem is formulated as an optimization problem that aims to maximize the probability of satisfying the average user. However, it turns out that it is NP-hard to optimize the resulting objective function. They describe a greedy algorithm to select a set of diverse documents that is a $(1 - 1/e)$ approximation algorithm for the problem. Gollapudi and Sharma [7] describe an axiomatic framework that can be used for designing and characterizing diversification mechanisms. Santos et al. [11] proposed the xQuAD (explicit Query Aspect Diversification) framework that takes into account various *aspects* of an underspecified query. In the proposed framework, the different aspects of a given query are represented in terms of *sub-queries* and the documents are ranked based on their relevance to each sub-query. Santos et al. [12] propose a supervised selective diversification

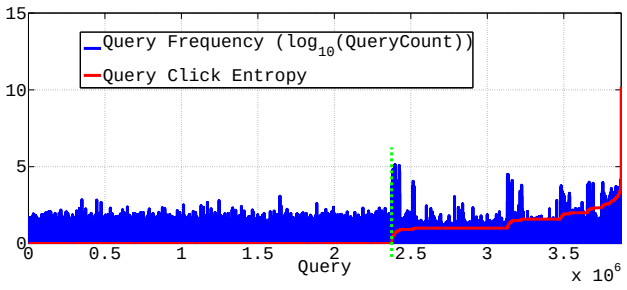

Figure 1: Queries sorted by their click entropy values. The green dashed line indicates beginning of queries with non-zero click entropies. Roughly 1.2M queries have non-zero click entropy values.

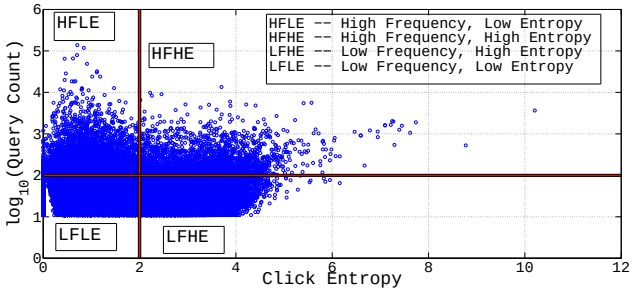

Figure 2: Scatter plot showing query frequency and associated click entropy as observed in the WS Logs.

approach that trades off relevance and diversity on a per query basis. As discussed before, the focus of these works is on creating search results lists so as to maximize the coverage of different query interpretations and does not take into account various other diversity dimensions that we discuss in following sections.

3. MOTIVATION

In order to test my hypothesis that there is a need to diversify search results along multiple dimensions, I analyzed search query log data from a commercial web search engine manually by selecting a small subset of queries. There are a total of 14.9 million query transactions in the logs out of which $\approx 7M$ are unique queries. Out of these unique queries, click-through information is available for $\approx 3.87M$ queries with $\approx 4.97M$ unique URLs.

The queries selected for analyses are not sampled randomly but are selected so as to ensure that the queries analyzed manually are representative of the queries that can benefit from diversity. I use *click entropy* [13] to identify such queries. It has been used previously to identify ambiguous queries [16] and queries that can potentially benefit from personalization [14] and diversification [6].

Click entropy (CE) for a query q is defined as follows.

$$CE(q) = \sum_{d \in D_q} -P(d|q) \log_2 P(d|q) \quad (1)$$

Here, D_q is the set of documents selected by various users for query q.

A higher click entropy indicates that users selected different documents for the given query indicating that the query was used by users looking for different information and hence, indicates a potential for diversification. The idea here is to identify queries with high click entropies and observe the reasons for users clicking different URLs for the query.

Figure 1 shows a plot of click entropy values and frequency (on log scale) for each of the 3.87M unique queries with click-through information in the WSE logs. From the figure we note that roughly two-thirds of all the queries have a zero click entropy value. Most of these queries appeared only once in the logs and thus only had a single clicked document that results in zero click entropy. A large fraction of these queries corresponds to navigational or transactional queries (e.g. `google`, `yahoo`, `pnc bank` etc.). Roughly one-thirds of all the queries have non-zero click entropies.

Next, I considered only those queries that appeared in the logs more than ten times. That resulted in a total of 80,756 unique queries that appeared for a total of 5,104,536 times in the query logs. Figure 2 shows a scatter plot between query frequency and query click entropies for this set of queries. Each point on the plot represents a query with its frequency (log scale) on y-axis and its click entropy on x-axis. Then I divided the queries into four classes based on their frequency and entropy values. Table 1 summarizes the criterion for classification and also lists some other statistics about queries in each of the four classes. I then selected a random sample of 50 queries from each class for analysis. Observing the queries in each class and their associated clicked documents, following observations were drawn.

- Queries in the LFLE class account for 55.29% of all the unique queries and appear roughly one million times in the query logs (21.32%). A large faction of queries in this class are generally *long-tail* queries where the user is generally looking for a specific piece of information. E.g. `ohio department of corrections`, `mutual savings credit union` etc. Many of the queries in this class are specific website names

- Queries belonging to LFHE class are also generally quite specific. The reason for the high entropy values is due to the fact these queries are generally "literatue survey" type queries – user is looking for various aspects of the query or a single document is not able to provide the complete information. E.g. `peru facts`, `katie morgan` etc.

- Queries in HFLE class are mostly navigational or transactional queries where the user is looking for a specific website (e.g. `pogo`, `askjeeves.com` etc.) or answers to some common questions (e.g. `calories in strawberry` etc.).

- From Figure 2 we note that there are a number of queries that have high frequency as well as high click entropies. These are the queries that have high potential for diversification and are most appropriate for our purposes. Even though the number of unique queries in this class is small (2.80% of all the unique queries), the fact that these queries have high frequencies indicate that these queries are issued repeatedly by a

Query Class	Condition	Number of Unique Queries	Number of Times Query Issued
Low-Frequency–Low-Entropy (LFLE)	Frequency $\leq$ 100, Entropy $\leq$ 2	44,653 (55.29%)	1,088,466 (21.32%)
Low-Frequency–High-Entropy (LFHE)	Frequency $\leq$ 100, Entropy $>$ 2	29,947 (37.08%)	759,565 (14.88%)
High-Frequency–Low-Entropy (HFLE)	Frequency $>$ 100, Entropy $\leq$ 2	3,898 (4.83%)	2,465,363 (48.30%)
High-Frequency–High-Entropy (HFHE)	Frequency $>$ 100, Entropy $>$ 2	2,258 (2.80%)	791,142 (15.50%)

Table 1: Four classes of queries based on frequency and click entropy values.

considerable fraction of user population (15.50% of all the queries). Thus, improving search results for these queries is extremely crucial. We also note that these queries are underspecified and their average length is smaller than those of in the remaining three classes.

- For many queries, *medium diversity* is an important requirement. For example, for the query `ferrari`, many of the user clicks were for images of the sports-car and some clicks were for the ferrari website.
- For many queries, *source diversity* is crucial. For example, for the query `bausch & lamb` while some users clicked the company's website, many users were looking for news about bausch & lamb's new products and thus clicked results from many news websites.
- For transactional queries, people may click results from different websites that offer that particular product/service.

Above analysis, even though small scale, provides some evidence to the hypothesis that indeed, there is a need to diversify search results along different dimensions. Further, search result diversification can benefit other types of queries as well in addition to ambiguous queries that have been the focus of current research. In light of these observation, I now describe the specific problems I want to address in my research.

4. PROBLEMS I WANT TO ADDRESS

4.1 What are different diversity dimensions?

Diversity requirements in information retrieval can generally be classified as either *extrinsic* or *intrinsic* [10]. Building upon this classification and based on my initial observations drawn from the above analysis, I describe here an initial hierarchy of diversity requirements.

Intrinsic Diversity: Intrinsic Diversity refers to the inherent need of avoiding *redundancy* as a part of the information need itself. Important dimensions to consider here are:

- **Novelty and Redundancy considerations:** It is desired that the successive documents returned by the search engine should provide new information that is not covered by the previous documents. For example, while doing a literature survey on `svm`, the user would like to see results that cover different aspects of SVMs (theory, implementation algorithms, variations etc.). Thus, for such queries we want to *minimize redundancy* and *maximize novelty* in top-k documents.
- **View points/Opinions/Sentiments:** For many queries, such as `abortion`, the user might be interested in finding various positive and negative opinions about the topic. Similar arguments apply for product reviews etc. For such queries, the search engine should present results that cover different view points regarding the query topic.
- **Medium Diversity:** For many queries *medium diversity* is an important requirement. This was particularly found to be true for product names, celebrity names, song and movie titles etc. For example, for the query `ferrari`, many of the user clicks were for images of the sports-car and some clicks were for the Ferrari website.

Extrinsic diversity: Extrinsic Diversity arises due to the uncertainty about the information need itself and can be attributed to the ambiguity in query (e.g. `java, jaguar` etc.) or lack of information about the user. Different users might prefer different types of documents for the same query. Given the lack of knowledge about the user's query intent and his background and preferences, the search engine can diversify the search results so as to maximize coverage over the user base. Some typical dimensions to consider can be:

- **User Expertise Level:** Consider the query `pagerank`. A computer scientist would prefer technical documents for this query that provide algorithmic details and mathematical principles behind the algorithm. A layman, on the other hand, would like to minimize the technical jargon and is perhaps looking for a much simpler explanation. Given no information about the user's expertise level, it is desired to include documents that would cover both types of users. Further, aggregate statistics from query logs can be obtained about users' gender, age etc. [8, 17] and results can be diversified accordingly. Such techniques have very useful applications for sponsored search. For example, for the query `shoes`, using information about user gender distribution, results for both male and female shoes can be included.

- **Document Source and Style/Readability:** In the above example (*pagerank*), we can include documents from different sources like CiteSeerX, ACM Digital Library, Wikipedia or may be some tech blogs that discuss about pagerank. Writing style in the documents is another aspect with respect to which diversification can be performed. Such a system can be very useful for teachers, students and non-native speakers of a language [9]. Consider two documents such that each of them contains all the information a user is looking for. One document is very easy to read and uses simple, non-technical language while the other one uses a more formal and technical language. Some users might prefer a document written in a simpler language while some might be looking for formally written documents. Google has recently provided an option to users to sort their results by document reading levels[1]. This, however, does not solve the problem in general because other aspects of style, viz. locale, discourse, genre, etc, have not been taken into account. We need a principled approach for balancing relevance and these readability measures.

Having such an hierarchy of diversification requirements will enhance our understanding about the *expectations of an average user from the search engine*. The above observations are based on a small scale, manual analysis of limited query log data available to me. By utilizing aggregate statistics about queries, users and their interaction with the search engine for different queries, more concrete evidences about diverse user preferences as well as relative importance of different diversity dimensions can be derived.

4.2 Classifying Queries Based on Diversity Requirements

Once we know different diversity dimensions, the next natural question I would like to address is: **Given a query, how can we determine the diversification requirement best suited for the query?** For some queries subtopic coverage may be more important while for others diversification with respect to document source or stylistics might be important. This problem is related to the problem of selective diversification [12] where the goal is to identify queries for which diversification techniques should be used. However, in addition, we are also interested in identifying different diversity classes a given query belongs to? Further, for some queries it may be required to diversify along multiple diversity dimensions. In such cases, it is also important **to determine the relative importance of different diversity dimensions for the given query**. By utilizing past user interaction data, query level features (like query clarity, entropy, lexical features etc.) and document level features (e.g. popularity, content quality, previous click history etc.), classifiers for diversification requirements can be developed.

4.3 Diversification Framework

Given a user query, once we know the type of diversity requirements for the user, an appropriate diversification technique is required. **I would like to study the problem of simultaneously diversifying search results along multiple dimensions**, as discussed above. One possible way here could be to build upon the *nugget* based framework introduced by Clarke et al. [5] where we represent each document as a set of nuggets, each nugget corresponding to a diversity dimension. Further, we represent the average user also as a set of nuggets. A search result list could then be created by including documents that satisfy maximum user nuggets.

[1] http://www.google.com/support/websearch/bin/answer.py?hl=en&answer=1095407

5. EXPECTED CONTRIBUTIONS OF PROPOSED RESEARCH

1. **A hierarchy of diversity dimensions along which a search query can be diversified.** Such an hierarchy will provide deeper insights about the importance of diversity in information retrieval and will enable the development of algorithms optimized and fine-tuned for specific diversity requirements.

2. **Algorithms for classifying queries based on their diversity requirements.** Having information about the exact diversification requirements for a query, the search engine can compute optimized results for the query. Further, due to the long tail nature of web queries, results for a (relatively) small number of high frequency queries that cater to a large fraction of users can be pre-computed for a faster response time.

3. **Frameworks for search result diversification along multiple dimensions.** As opposed to the current work, we advocate a two stage approach to search result diversification. Given a query, the first task should be to identify various diversity dimensions and their relative importance for the query followed by an appropriate diversification strategy. Such an approach will help offer search results to the user, optimized for the query under consideration.

6. REFERENCES

[1] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In *WSDM '09: Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 5–14. ACM, 2009.

[2] J. Carbonell and J. Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336, New York, NY, USA, 1998. ACM.

[3] B. Carterette and P. Chandar. Probabilistic models of ranking novel documents for faceted topic retrieval. In *CIKM '09: Proceeding of the 18th ACM conference on Information and knowledge management*, pages 1287–1296, New York, NY, USA, 2009. ACM.

[4] H. Chen and D. R. Karger. Less is more: probabilistic models for retrieving fewer relevant documents. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, pages 429–436, New York, NY, USA, 2006. ACM.

[5] C. L. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon. Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 659–666, New York, NY, USA, 2008. ACM.

[6] P. Clough, M. Sanderson, M. Abouammoh, S. Navarro, and M. Paramita. Multiple approaches to analysing query diversity. In *SIGIR '09: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 734–735, New York, NY, USA, 2009. ACM.

[7] S. Gollapudi and A. Sharma. An axiomatic approach for result diversification. In *WWW '09: Proceedings of the 18th international conference on World wide web*, pages 381–390, New York, NY, USA, 2009. ACM.

[8] B. J. Jansen and L. Solomon. Gender demographic targeting in sponsored search. In *Proceedings of the 28th international conference on Human factors in computing systems*, CHI '10, pages 831–840, New York, NY, USA, 2010. ACM.

[9] N. Ott and D. Meurers. Information retrieval for education: Making search engines language aware. *Themes in Science and Technology Education, Special issue on Computer-aided language analysis, teaching and learning: approaches, perspectives and applications*, 3(1-2):9–30, 2010.

[10] F. Radlinski, P. N. Bennett, B. Carterette, and T. Joachims. Redundancy, diversity and interdependent document relevance. *SIGIR Forum*, 43(2):46–52, 2009.

[11] R. L. Santos, C. Macdonald, and I. Ounis. Exploiting query reformulations for web search result diversification. In *WWW '10: Proceedings of the 19th international conference on World wide web*, pages 881–890, New York, NY, USA, 2010. ACM.

[12] R. L. Santos, C. Macdonald, and I. Ounis. Selectively diversifying web search results. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10, pages 1179–1188. ACM, 2010.

[13] F. Silvestri. Mining query logs: Turning search usage data into knowledge. *Foundations and Trends in Information Retrieval*, 4(1–2):1–174, 2010.

[14] J. Teevan, S. T. Dumais, and D. J. Liebling. To personalize or not to personalize: modeling queries with variation in user intent. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 163–170, New York, NY, USA, 2008. ACM.

[15] J. Wang and J. Zhu. Portfolio theory of information retrieval. In *SIGIR '09: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 115–122, New York, NY, USA, 2009. ACM.

[16] Y. Wang and E. Agichtein. Query ambiguity revisited: clickthrough measures for distinguishing informational and ambiguous queries. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 361–364, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[17] I. Weber and A. Jaimes. Who uses web search for what: and how. In *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM '11, pages 15–24, New York, NY, USA, 2011. ACM.

[18] M. J. Welch, J. Cho, and C. Olston. Search result diversity for informational queries. In *WWW '11: Proceedings of the 20th international conference on World wide web*, 2011.