Current Approaches to Search Result Diversification

Enrico Minack, Gianluca Demartini, and Wolfgang Nejdl

L3S Research Center, Leibniz Universität Hannover,
30167 Hannover, Germany, {lastname}@L3S.de

Abstract With the growth of the Web and the variety of search engine users, Web search effectiveness and user satisfaction can be improved by diversification. This paper surveys recent approaches to search result diversification in both full-text and structured content search. We identify commonalities in the proposed methods describing an overall framework for result diversification. We discuss different diversity dimensions and measures as well as possible ways of considering the relevance / diversity trade-off. We also summarise existing efforts evaluating diversity in search. Moreover, for each of these steps, we point out aspects which are missing in current approaches as possible directions for future work.

1 Introduction

In the last years, the Web has become the largest and most consulted public source of information, and Web search emerged as the primary technique for finding relevant information on the Web. Search engines usually provide a long list of results that contains thousands of entries, where the most relevant results tend to be quite similar [1]. In particular for informational queries [2], users reading through a list of relevant but redundant pages quickly stop as they do not expect to learn more. The phenomenon of *saturated user satisfaction* is a well-understood and extensively studied field in economics called "*law of diminishing marginal returns*" [3].

The amount of data on the Web is growing exponentially, and so does the amount of relevant results for a query. Given that most search engine users only look at the first page of available results, to improve user satisfaction, this search result list should be optimised to contain both *relevant* and *diverse* results [4], fairly representing the thousands of relevant results. This task is also known as *search result diversification*.

For an ambiguous query like "Jaguar", a search result list should contain results about the *car*, the *animal*, the *operating system* and other senses. In case of an unambiguous query like "nuclear power plant", the list should be diverse in the contained information: objective and opinionated sites, supportive and opposing thoughts, related topics and subtopics. It is easy to see how this can be a computational expensive process that is difficult to run at query time.

The goal of this paper is to survey recent approaches in this area, identifying commonalities and differences between these works. We also present possible

open questions not yet addressed by state-of-the-art techniques. Here, we focus on the field of search result diversification, however, we want to point to other fields where similar problems have been addressed and solutions might be adaptable. For example, recommender systems provide a list of items which are interesting (*i. e.*, relevant) and *novel* (*i. e.*, diverse from the ones the user already knows) [5]. Another example is image or video search where near-duplicate results are removed [6], or multiple senses of ambiguous queries are covered [7]. Dynamic clustering algorithms on image features are used in [8] to provide visually diverse result sets. In general, clustering algorithms may provide adaptable (dis)similarity measures that are used to create sets of items with high intra-set and low inter-set similarity [9].

In this paper, we compare current work in search result diversification. To the best of our knowledge, there is no such recent comparison. First, we identify common aspects and different notions of diversity in all proposed approaches. We show how the trade-off between relevance and diversity is solved, which is an NP-hard optimisation problem. As last step, search effectiveness is evaluated not only in terms of relevance but also of diversity. Finally, we point out open problems and areas which can be improved.

The rest of the paper is structured as follows. In Section 2, we define the problem of search result diversification. Section 3 presents dimensions and types of diversity, and how approaches measure them. Further in Section 4, we show the strategies and algorithms of balancing between relevance and diversity, efficiently. The evaluations of the effectiveness of current approaches are described in Section 5. We conclude by discussing open research questions in Section 6.

2 Search Result Diversification: Problem Definition

Search result diversification is an optimisation problem aiming to find k items which are the subset of all relevant results that contains both most *relevant* and most *diverse* results. Usually, increasing the diversity in the subset leads to a decrease in relevance; therefore, the optimal trade-off between relevance and diversity needs to be found. Looking at previous work on search result diversification, it is possible to notice that, in order to achieve the optimisation goal, three components are usually adopted. Here, we follow the notion and structure of a general result diversification approach presented in [10]:

Relevance Measure: It provides a relevance score for each results which creates an initial ranking of the items.

Diversity Measure: This measure reflects the dissimilarity between two given items, or the overall dissimilarity of a set of results.

Diversification Objective: The objective defines the way both measures are merged into a single score that has to be maximised.

The first step of result diversification is to rank the items by a relevance score as a normal retrieval task. In Information Retrieval (IR), several models and relevance measures have been developed. In result diversifying systems, such

standard techniques have been used to rank items by their relevance. For example, [11] uses a vector space model to represent items and queries, while [12] exploits language models and KL-divergence as relevance functions.

The second and actually diversifying component is the measure of diversity. Such a measure provides means to represent the dissimilarity of two results – or the dissimilarity within a whole set of results – with a single value. Different types of diversity and proposed diversity measures will be described in Section 3.

The third component, the diversification objective, formalises the strategy to find a trade-off between the two measures in order to diversifying a result set. This optimisation is known to be *NP-hard* [3,10], so there is a need to develop efficient algorithms. In Section 4, we will see what diversification objectives and algorithms current approaches employ to efficiently diversify search results.

Finally, the quality of the result set has to be evaluated using standardised metrics, repeatable experiments and publicly available datasets. In Section 5, we give detailed information about the evaluation efforts of the reviewed works.

3 Notions of Diversity

We first introduce to some properties of diversity and take a look at the various kinds of diversity known to exist in information sources. We then review notions of diversity considered in recent work.

3.1 Dimensions of Diversity

Considering Web search, two levels of diversity can be found [13]: (1) query terms may be ambiguous, which is word sense diversity, and (2) for a specific word sense, the available information sources may be diverse. Different *causes* of diversity in such information sources are known to be, *e. g.*, educational, cultural, spatio-temporal [14], or simply the goal of communication. These become manifest in an orthogonal dimension, the *type* of diversity: *e. g.*, conflicting information [15], opposing opinions and sentiment [16], ideological perspectives [17], or text genre [18]. Further, as the usage of the term "diversity" is itself diverse, diversity is studied from different perspectives in fields like ecology, geography, psychology, linguistics, sociology, economics, and communication [19].

This diversity in information sources should not be ignored or avoided. Instead, it should be seen as a rich feature that, handled explicitly and being exploited, could lead to better ways to deal with diverse information sources [20].

3.2 Measures of Diversity

We saw that there are many dimensions of diversity that can be considered for diversification. We will now investigate which notions of diversity current approaches consider and how they are measured. Note that the term similarity can be used interchangeably to denote the same concept as of dissimilarity:

$dissimilarity = 1 - similarity$, where $similarity \in [0, 1]$.

Semantic Distance. Gollapudi *et al.* [10] reuse the known *min-hashing scheme* sketching algorithm, which produces sketches similar to random term samples using a number of different hashing functions. They use the *Jaccard similarity* between those sketches as the dissimilarity measure, *i. e.*, one minus the fraction of the cardinality of the intersection and the union of the two sketches. This dissimilarity measure diversifies based on content dissimilarity.

Categorical Distance. Additionally, [10] presents a *categorical distance* where dissimilarity is based on the distance between the category of the results within a taxonomy. As a distance measure, the *weighted tree distance* measure is used. In case of multiple categories being assigned, the shortest distance from each category of one result to the categories of the other result is added up after weighting with the minimal probability that any of the respective two categories is assigned. This measure emphasises word senses diversification.

Agrawal *et al.* [3] also use categories, derived from query click logs. However, they abstain from using an inter-result dissimilarity measure. They directly use the information about the categories in their diversification objective.

Vee *et al.* [21] introduce a *diversity order* for relational databases being an order among attributes (*e. g.*, for cars: $Make \prec Model \prec Colour \prec \ldots$). This order expresses that certain attributes have higher priority to be diversified than others (*e. g.*, first $Make$ is diversified, then $Model$). They show how result tuples can be seen as paths in a tree of values, where the paths satisfy the diversity order. Tuples that have a longer path from the root in common are more similar than others. Therefore, this measure is similar to a tree distance measure.

Novel Information. In [12], *unigram language models* are used to represent results. The authors define functions that quantify novel information a new result conveys additionally to an (the) existing result(s) using the *KL-divergence*. This measure diversifies in a general sense regarding content dissimilarity.

Conclusion. The diversity measure used by a system defines the kind of diversity the system can handle. However, none of the presented works focus on their diversity measure. The measures are mentioned very briefly without motivation.

Looking at these diversity measures, two groups can be observed. One group measures dissimilarity based on content similarity, whereas the other group uses metadata about the content (*e. g.*, the categories), which are not extracted from the content but taken from additional information sources (*e. g.*, user click logs). Still, no measure exploits intrinsic properties of the results, *e. g.*, the genre (blog post, a news article, a manual) or the sentiment regarding the query topic. Therefore, these kinds of diversity are not yet exploited explicitly for search result diversification.

4 The Relevance / Diversity Optimisation Problem

The relevance and diversity of a search result set can be maximised using various strategies. The main challenge for all these strategies is to select those results that add more diversity to the set, probably at the cost of relevance. Finding a good compromise is the primary goal.

4.1 Diversification Objectives

Gollapudi *et al.* [10] combine the relevance measure and the dissimilarity in three different ways: *max-sum*, *max-min*, and an *average dissimilarity* like measure. These set selection functions are to be maximised.

Max-sum Diversification. The first objective in [10] combines the sums of the relevance and diversity measure as a weighted sum.

Max-min Diversification. The second objective targets at maximising the sum of the minimum relevance and minimum dissimilarity within the set.

Average Dissimilarity Diversification. Their third objective adds the original relevance for a result with the average dissimilarity regarding all other results in the set. The sum over the whole set is to be maximised.

Max-sum of max-score Diversification. Similarly to max-sum diversification, [21] maximises the sum of dissimilarity of the result set, but it only produces sets that have the maximal relevance sum. Therefore, it does not find sets with higher diversity scores but slightly lower relevance sum.

Max-product Diversification. Based on the already chosen results, Zhai *et al.* [12] select the next result by maximising the parameterised product of the relevance of the next result and its dissimilarity to the chosen results.

Categorical Diversification. Agrawal *et al.* [3] use a relevance measure that considers the categories of a document and query. The result set is diversified so that its results cover all categories, weighted by their probability to occur.

4.2 Diversification Algorithms

The problem of search result diversification is NP-hard [3,10]. Therefore, approximation algorithms have to exploit inherent structural properties of the solution space to achieve adequate system response times. IR systems based on inverted lists are proven to be unable to directly provide diverse results [21]. In the following, we present algorithms used to efficiently find top-k diverse search results.

Gollapudi *et al.* [10] show that their max-sum and max-min diversification objectives can be casted to a facility dispersion problem for which approximation algorithms exist. Agrawal *et al.* [3] use a Greedy algorithm that starts with an empty list of results and select the next result with the highest *marginal utility* until k results are selected. The marginal utility measures the probability that the result satisfies a category the current result set does not yet satisfy. Similarly, Zhai *et al.* [12] uses the same Greedy algorithm, but with their function that represents the novel information being introduced by the next document. Vee *et al.* [21] cluster results into buckets based on their diversity order and selects results from those buckets in order to retrieve balanced diverse results.

Conclusion. Apparently, most approaches find a solution for the diversification problem using Greedy approximation algorithms. All optimisation algorithms work online on the relevant results provided by the retrieval phase. Therefore, the presented works do not investigate the applicability of offline pre-computation or special data structures that could improve online performance.

5 Evaluating Diversity in Search

This section presents methods for evaluating diversity-aware search techniques. We describe datasets used and evaluation metrics designed for this purpose.

5.1 Datasets for Diversity-aware Search

In previous works, different types of datasets have been used. Gollapudi *et al.* [10] use Wikipedia disambiguation pages as ground truth for the word senses. They also use a structured dataset in the context of product disambiguation evaluating the goodness of a measure based on a product taxonomy. In [3], the authors use 10,000 queries and top 50 retrieved results from a commercial search engine, judgements obtained with the Amazon Mechanical Turk[1], and the Open Directory Project (ODP)[2] taxonomy to classify results. Zhai *et al.* [12] use topics from the Text REtrieval Conference (TREC) Interactive Track where assessors identify a list of subtopics for each topics and mark the relevance of retrieved results with respect to each subtopic. Vee *et al.* [21] have based their experiments on a structured dataset using Yahoo! Autos. They perform experiments generating keyword and structured queries measuring response times for different cases. Real and synthetic structured data are used in [11]. They create feature vectors they want to retrieve back as a set of diverse results.

As we have seen, previous work use different and non-standard datasets. In order to create a benchmark for diversity in search, in the Web Track at TREC 2009 the new "Diversity Task" started. We notice that the notion of diversity used is rather a topical diversity. This leaves open the aspect of evaluating other dimensions as, *e. g.*, diversity of opinions (see Section 3.1).

Conclusion. As we can see, in most cases two main types of datasets have been used: classical textual documents to be ranked (*i. e.*, TREC-like tasks) and structured datasets (*i. e.*, for Database-like search task). In both cases, the goal is to provide the user with a smaller set of relevant and diverse results. While we have also seen that standard benchmarks are being created, there is still need for creating benchmarks for specific diversification tasks.

5.2 Diversity-aware Evaluation Measures

In order to evaluate the effectiveness of proposed diversity-aware search approaches, new metrics need to be designed. In most cases, adaptation from already existing metrics have been done.

In [4], an evaluation framework for novelty and diversity is proposed. They see information needs and results as sets of "information nuggets", and relevance is defined as a function of the nuggets contained in the user's need and previous results. Moreover, as graded relevance seems a reasonable assumption for

[1] Amazon Mechanical Turk: `http://www.mturk.com/`
[2] ODP – Open Directory Project: `http://www.dmoz.org/`

such task, they propose α-NDCG: an adaptation of the well-known NDCG metric proposed in [22]. They experiment on past TREC collections showing the feasibility of the proposed approach.

In [12] *S-Recall at k* is defined as the percentage of subtopics covered by one of the first k results. Values of S-Recall at k cannot be directly compared among topics having a different number of subtopics, that is, this metric does not account the difficulty of a certain topic. For this reason they define, *S-Precision at recall r* which is the ratio between the minimal rank at which the system has S-recall r and such minimal rank obtained by an optimal system. Additionally, for penalising redundancy (*i. e.*, low diversity) in the ranking, they define *weighted S-precision at recall r* taking into account the cost of presenting a result to the user as well as the cost of processing a subtopic in a result.

In [3] the authors propose an adaptation of common metrics taking into account the user intent. They consider ambiguous queries to belong to different categories (*i. e.*, senses) and relevance to be rated differently for different categories. They take into account the "popularity" of each query's category (*e. g.*, for the query "Jaguar" the *car* sense might be more prominent than the *animal* sense) computing a distribution on the categories for a query.

In the database query scenario, the evaluation is usually based on comparing the approximation done by the system against the "optimal" result (see, *e. g.*, [11]) which can be computed (but this computation is NP-hard).

6 Discussion and Conclusion

In this paper, we surveyed recent advances in search result diversification. We found that all approaches fit well in the notation and structure of a general diversification system as given in [10]. Quite a number of diversity measures and diversification objectives are already available. However, the reviewed notions of diversity are still limited to content or category similarity, though a range of more specific diversity types exists. Further, no new (dis)similarity measures were developed, but rather existing metrics (*e. g.*, Sketching, KL-divergence) were reused. Here we see potential for further advances.

Moreover, it would be interesting to design ranking functions that directly focus on diversity rather then to see diversification as a re-ranking step. Even if Vee *et al.* [21] show that no inverted list based system can produce a relevant and diverse ranking of results, we still believe that the retrieval of diverse and relevant results may benefit from an integrated retrieval phase, as well as data structures supporting result diversification.

Finally, regarding the evaluation metrics, there have been adaptations of widely used and well understood metrics such as NDCG. Standard benchmarks created for other purposes or proprietary datasets are used, but no dataset for diversity in search is available yet. We believe that different dataset for different notions of diversity (*e. g.*, opinions, topics, or genre) should be constructed.

Acknowledgment. This work was supported by the European Seventh Framework Programme FP7 (Grant 231126, Project LivingKnowledge).

References

1. Carbonell, J., Goldstein, J.: The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. In: Proceedings of SIGIR '98, ACM (1998) 335–336
2. Broder, A.Z.: A Taxonomy of Web Search. SIGIR Forum **36**(2) (2002) 3–10
3. Agrawal, R., Gollapudi, S., Halverson, A., Ieong, S.: Diversifying Search Results. In: Proceedings of WSDM '09, ACM (2009) 5–14
4. Clarke, C.L., Kolla, M., Cormack, G.V., Vechtomova, O., Ashkan, A., Büttcher, S., MacKinnon, I.: Novelty and Diversity in Information Retrieval Evaluation. In: Proceedings of SIGIR '08, ACM (2008) 659–666
5. Adomavicius, G., Tuzhilin, A.: Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. IEEE Transactions on Knowledge and Data Engineering **17**(6) (2005) 734–749
6. Wu, X., Hauptmann, A.G., Ngo, C.W.: Practical Elimination of Near-Duplicates from Web Video Search. In: Proceedings of MULTIMEDIA '07, ACM 218–227
7. Weinberger, K.Q., Slaney, M., Van Zwol, R.: Resolving Tag Ambiguity. In: Proceeding of MM '08, ACM (2008) 111–120
8. van Leuken, R.H., Pueyo, L.G., Olivares, X., van Zwol, R.: Visual Diversification of Image Search Results. In: Proceedings of WWW '09. (2009) 341–350
9. Jain, A., Murty, M., Flynn, P.: Data Clustering: a Review. ACM Computing Surveys **31**(3) (1999)
10. Gollapudi, S., Sharma, A.: An Axiomatic Approach for Result Diversification. In: Proceedings of WWW '09, ACM (2009) 381–390
11. Jain, A., Sarda, P., Haritsa, J.R.: Providing Diversity in K-Nearest Neighbor Query Results. In: Proceedings of PAKDD '04. (May 26–28 2004) 404–413
12. Zhai, C.X., Cohen, W.W., Lafferty, J.: Beyond Independent Relevance: Methods and Evaluation Metrics for Subtopic Retrieval. In: Proceedings of SIGIR '03, ACM
13. Clough, P., Sanderson, M., Abouammoh, M., Navarro, S., Paramita, M.: Multiple Approaches to Analysing Query Diversity. In: Proceedings of SIGIR '09, ACM
14. Giunchiglia, F., Maltese, V., Madalli, D., Baldry, A., Wallner, C., Lewis, P., Denecke, K., Skoutas, D., Marenzi, I.: Foundations for the Representation of Diversity, Evolution, Opinion and Bias. Report D1.1, Living Knowledge European Project (*to appear in* 2009)
15. Etzioni, O., Banko, M., Soderland, S., Weld, D.S.: Open Information Extraction from the Web. Communications of the ACM **51**(12) (2008) 68–74
16. Pang, B., Lee, L.: Opinion Mining and Sentiment Analysis. In: Foundations and Trends in Information Retrieval. Volume 2. (2008) 1–135
17. Lin, W.H.: Identifying Ideological Perspectives in Text and Video. PhD thesis, Language Tech. Inst., School of Comp. Sci., Carnegie Mellon University (Oct 2008)
18. Biber, D.: The Multi-Dimensional Approach to Linguistic Analyses of Genre Variation: An Overview of Methodology and Findings. Computers and the Humanities **26** (1993) 331–345
19. McDonald, D.G., Dimmick, J.: The Conceptualization and Measurement of Diversity. Communication Research **30**(1) (2003) 60–79
20. Giunchiglia, F.: Managing Diversity in Knowledge. In Ali, M., Dapoigny, R., eds.: IEA/AIE 2006, LNAI 4031, Springer-Verlag Berlin Heidelberg (2006) 1
21. Vee, E., Srivastava, U., Shanmugasundaram, J., Bhat, P., Yahia, S.A.: Efficient Computation of Diverse Query Results. In: Proceedings of ICDE '08. 228–236
22. Järvelin, K., Kekäläinen, J.: Cumulated Gain-Based Evaluation of IR Technique. ACM Transactions on Information Systems (TOIS) **20**(4) (2002) 422–446