Redundancy, Diversity and Interdependent Document Relevance

Filip Radlinski
Microsoft Research
filiprad@microsoft.com

Paul N. Bennett
Microsoft Research
paul.n.bennett@microsoft.com

Ben Carterette
University of Delaware
carteret@cis.udel.edu

Thorsten Joachims
Cornell University
tj@cs.cornell.edu

1 Introduction

The goal of the Redundancy, Diversity, and Interdependent Document Relevance workshop was to explore how ranking, performance assessment and learning to rank can move beyond the assumption that the relevance of a document is independent of other documents. In particular, the workshop focussed on three themes: the effect of redundancy on information retrieval utility (for example, minimizing the wasted effort of users who must skip redundant information), the role of diversity (for example, for mitigating the risk of misinterpreting ambiguous queries), and algorithms for set-level optimization (where the quality of a set of retrieved documents is not simply the sum of its parts).

This workshop built directly upon the Beyond Binary Relevance: Preferences, Diversity and Set-Level Judgments workshop at SIGIR 2008 [3], shifting focus to address the questions left open by the discussions and results from that workshop. As such, it was the first workshop to explicitly focus on the related research challenges of redundancy, diversity, and interdependent relevance – all of which require novel performance measures, learning methods, and evaluation techniques. The workshop program committee consisted of 15 researchers from academia and industry, with experience in IR evaluation, machine learning, and IR algorithmic design. Over 40 people attended the workshop.

This report aims to summarize the workshop, and also to systematize common themes and key concepts so as to encourage research in the three workshop themes. It contains our attempt to summarize and organize the topics that came up in presentations as well as in discussions, pulling out common elements. Many audience members contributed, yet due to the free-flowing discussion, attributing all the observations to particular audience members is unfortunately impossible. Not all audience members would necessarily agree with the views presented, but we do attempt to present a consensus view as far as possible.

2 Workshop Structure

The workshop's structure, which was designed to promote discussion, consisted of 6 invited talks and 4 paper presentations. The schedule was split into four sessions, namely *Axioms and Evaluation*, *Algorithms*, *Evaluation* and *Collection Building*. Each session was organized to maximize the amount of discussion time.

The axioms and evaluation session included an invited talk by Sreenivas Gollapudi from Microsoft Research titled *A Framework for Result Diversification*, an invited talk by Khalid El-Arini from Carnegie Mellon University on *Turning Down the Noise in the Blogosphere* and a talk by Scott B. Huffman from Google presenting a paper with Elizabeth DeVaul Tucker on *Evaluating Diversity for a Vertical Slice of the Query Stream*. Following this, the second session on Algorithms included an invited talk on *Diversified Retrieval as Structured Prediction* by Yisong Yue from Cornell University and an invited talk on *Improving Diversity in Ranking Using Absorbing Random Walks* by Andrew B. Goldberg from the University of Wisconsin at Madison.

The third session consisted of an invited talk by ChengXiang Zhai from the University of Illinois at Urbana-Champaign on *Modeling Diversity in Information Retrieval* and an invited talk by Charles Clarke from the University of Waterloo titled *Evaluating Novelty and Diversity*. The final session on Collection Building included three paper presentations – *Evaluation of Redundant Information from Distillation Systems Using Nuggets and Fuzzy Sets* presented by Olga Babko-Malaya from BAE Systems (co-authored with James White, Daniel Hunter, Connie Fournelle and Michael Schneider), *C-Test: Supporting Novelty and Diversity in Testfiles for Search Tuning* presented by David Hawking from Funnelback Pty Ltd (co-authored with Tom Rowlands and Paul Thomas), and *Development of a Collection to Support Diversity Analysis* presented by Monica Lestari Paramita from the University of Sheffield (co-authored with Mark Sanderson and Paul Clough).

3 Views on Diversity

Many of the talks addressed similar themes, while focussing on different aspects of research problems. This resulted in many related questions, suggestions and discussions being offered in response to different talks. We therefore chose to present this summary grouped into themes, rather than chronologically.

3.1 What is Diversity?

The most common theme from the workshop is best summed up as a question: What is diversity? At a high level, Charles Clarke and ChengXiang Zhai presented a similar breakdown of the topic of diversity into two distinct classes: Either a dynamic property determined by the users of a system, or an inherent static property of the data being presented to the users. While some queries may require both types of diversity, it seems clear that there is some separation; how to clearly describe it was discussed at length by the workshop participants, and the structure we present here came from rereading our notes of the discussion and attempting to fit the different types of questions and comments brought up.

3.1.1 Extrinsic Diversity: Diversity as uncertainty about the information need

One type of diversity aims to address uncertainty about the information need given a query. A canonical example in this class is the query "jaguar". One would wish for diverse results for this query since we cannot know if the user was interested in the animal, the car or another meaning of this query. However, this type of diversity can be broken down further into two subgroups: The uncertainty can come from (1) ambiguity in the entity the query refers to or (2) uncertainty about the user. An illustration of the former is the query "jaguar", while an illustration of the latter is the query "swine flu" where doctors and patients may be interested in different aspects of the same topic.

A recurring question brought up by attendees and speakers regarded the challenge of trading off between different groups of users when optimizing this type of utility. Should we optimize the mean utility (for instance presenting many results for the dominant information needs of queries), or should we aim to satisfy the largest fraction of users at least a little (effectively allowing minority information needs to be over-represented)? For instance, Khalid El-Arini presented an approach to providing a search over blog posts that maximizes the coverage over topics [6], and Yisong Yue presented an algorithm for covering maximally many distinct words in a diverse ranking (while learning the importance of different words) [11]. Similarly, Andrew Goldberg and ChengXiang Zhai both presented algorithms that trade-off between these loss functions ([14] and [12] respectively).

3.1.2 Intrinsic Diversity: Diversity as part of the information need

In contrast to diversity in response to uncertainty in the information need, diversification can also be seen as a question of avoiding redundancy and thus presenting a novel and useful set of results to a single well defined information need. Charles Clarke talked about this type of diversity as optimizing for *novelty* in results, where the goal is finding results that cover different *aspects* of an information need. This can also be seen in terms of the ideal outcome for a search: While for some searches the ideal outcome is a single result, such as for navigational queries, in many cases the ideal outcome is a set of results. For example, intrinsic diversity is required when (1) there is no single result that fully answers the information need, (2) the user desires different views – for example, a variety of reviews about a product, or a variety of opinions about a political issue, (3) the user desires a selection of options to choose between – for example, different prices and options for a service as described in the case of commercial queries by Scott Huffman, (4) the information need is to get an overview of a topic, (5) different results from different sources are needed to build confidence in the correctness of the answer to an information need. One interesting example of this type of information need was brought up by a workshop attendee in the legal domain, where a searcher would often be interested in finding every distinct legal case that is relevant to some current one.

Both Charles Clarke and Olga Babko-Malaya discussed the question of how to evaluate the presence of such different aspects of an intent, arguing for the use of *nuggets* that describe any binary property of a document. While Clarke suggested dealing with possible complex interactions between arbitrarily crafted nuggets using simplifying assumptions, Babko-Malaya described a set of rules that can be used to create nuggets based on a document collection and then aggregated. With her co-authors, she proposed *nugs* as clusters of semantically equivalent nuggets, while allowing nuggets to partially be assigned to different nugs. However, in addition to the presence

or absence of documents that cover a nugget, a final evaluation must ascribe an importance to each. In his presentation, David Hawking described a file format that could be used to encode a range of information about a ranked list of documents, including the presence or absence of nuggets and their relative importance.

3.2 Algorithms

Sreenivas Gollapudi set the stage for algorithms with his opening talk about different properties you may want for an algorithm that picks a diverse set of results given a query. He presented an impossibility result – that all axioms we may consider desirable cannot be satisfied [7] – and showed how different goals, represented by different axioms, correspond to different algorithms. One of the axioms that provoked a number of questions was that of stability: if an algorithm selects k diverse results, is it desirable that the same algorithm picks a strict superset of these results if asked to select $k + 1$ diverse results?

Khalid El-Arini presented an algorithm for selecting blog posts to show users, optimizing for the coverage over topics. However, he also argued that given the large number of topics that exist, a personalized diversification algorithm is preferable and showed how to learn the importance of covering topics on a per-user basis [6]. In a related manner, the algorithm discussed by Yisong Yue optimized the diversity of results in general search, learning the importance of individual words and then selecting the optimal set of results that cover the largest number of words [11]. Both of these speakers noted that such an approach to diversity is a special case of the NP-hard maximum set cover problem, and thus only admits an approximate polynomial time algorithm for such a coverage-based metric. Andrew B. Goldberg described an algorithm that covers the space of documents instead, relying on a known document/document similarity function. The random walk algorithm he presented sequentially selects documents that have the largest number of expected visits by a random walk that is absorbed by nodes representing previously selected documents [14]. One of the algorithms described by ChengXiang Zhai on the other hand took a risk-minimization approach, estimating the utility of presenting users specific results and optimizing so as to minimize the risk from presenting a result set [8, 13]. He also described an active diversification technique, where a ranking system can diversify aggressively initially, to help disambiguate users' interests [9]. One particularly relevant question in designing diversification algorithms came from an audience member during ChengXiang Zhai's talk, asking whether all queries *should* be diversified the same way – perhaps different types of queries should be diversified using different approaches.

3.3 Test Collections and Evaluation

During the presentations, evaluation challenges were brought up often. Traditional IR metrics, such as Average Precision and Discounted Cumulative Gain, clearly suffer when search results are diversified, because near duplicates of relevant documents are removed. Sources for information about diversity discussed included Wikipedia disambiguation pages, TREC interactive track topics labeled for subtopics, CLEF image data, metrics of wasted user effort due to duplicates, coverage of Open Directory Project classes, search engine query and click distributions, and average utility – all of which have various benefits and drawbacks. For instance the approach presented by Khalid El-Arini assessed blog posts for topicality and redundancy using a small user study; Yisong Yue

evaluated his approach using the coverage of TREC subtopics for a small set of queries; Scott Huffman's approach was evaluated on a few hundred randomly sampled commercial queries that were labeled for diversity among results, which were categorized as online stores, formal reviews, blog posts or manufacturer pages; Andrew B. Goldberg assessed document summaries produced by finding diverse sentences in documents.

There is no evaluation metric that seems to be universally accepted as the best for measuring the performance of algorithms that aim to obtain diverse rankings, perhaps in part because, as presented in the previous section, there is a wide diversity in what diversity means. For instance, Charles Clarke, Sreenivas Gollapudi, ChengXiang Zhai and Olga Babko-Malaya presented a number of precision and recall based metrics for diversity [4, 5, 7, 12]. Clarke in particular spoke at length about the importance of evaluation, presenting a measure that uses nugget-based relevance judgments and discounts nuggets appearing in ranked documents by both the rank k at which they appear, and their redundancy with nuggets in documents at ranks $1 \ldots k-1$. The measure sits in a framework that can be used to evaluate both extrinsic and intrinsic diversity, and incorporate a model of nugget or document importance to users.

There is also no standard dataset that currently appears to be in wide use for studying diversity. However, of particular note, the release of a large dataset for diversity evaluation on image data as part of ImageCLEF 2009 may be of interest to practitioners [1]. Presented by Monica Lestari Paramita, this dataset includes a large number of real user queries, images from the Belga News Agency, and subtopic or "cluster" judgments indicating which of several aspects an image relates to. In her talk, Paramita also described how the ImageCLEF queries were labeled for extrinsic diversity based on the frequency with which other words co-occur with a given query in the logs of the image search engine.

New types of evaluations were also brought up during discussions. For instance, one audience member proposed an evaluation based on triplets in the form (query, document 1, document 2). A judge may be told that given the query, document 1 is known to be relevant. The judge is then asked to judge if document 2 is relevant and/or redundant given this information. The evaluation of entire sets of results were mentioned in relation to David Hawking's C-Test system and work by Hawking and colleagues on direct comparison of result sets [10]. In addition, the question of how diverse are the results currently returned by state of the art information retrieval systems provoked some discussion. Such diversity could in principle be measured using any of the above metrics, demonstrating a need for new research in diversification algorithms in general information retrieval settings.

Finally, it seems that the different types of diversity, extrinsic and intrinsic, likely require different evaluation metrics to more closely estimate utility to users. For example, the intent-aware metrics highlighted by Sreenivas Gollapudi [2] seem more appropriate to extrinisic diversity where the expectation over intents are taken with respect to the uncertainty regarding the need (that is, likely intent by population, property of the user, and so forth). However, this seems to likely be a mismatch for intrinsic diversity where the need itself requires some notion of recall as mentioned above. Here, performance metrics such as Zhai et. al's [12] subtopic recall and precision seem a priori to be a more likely fit to utility.

3.4 Applications

One of the most interesting outcomes of the workshop was the discussion of the large range of different applications where diversity is important. In addition to standard web search retrieval tasks, some of the applications where some form of optimizing for diversity is important include:

- Image search
- News and blog aggregation
- Web queries with commercial intent .
- Summarizing documents
- Finding related movie stars
- Finding experts in enterprise search
- Searching for legal precedents and patents

In particular, each application provides use cases with a different balance over the types of diversity that are most important.

4 Conclusion

The SIGIR 2009 workshop on Redundancy, Diversity, and Interdependent Document Relevance provided a forum for researchers to explore how ranking algorithms and evaluation techniques can move beyond the assumption that relevance of documents is independent of other documents presented. In the ten talks of the workshop, and over two and a half hours of discussion, a number of themes recurred. In particular the different meanings of diversity (as well as the lack of a standardized vocabulary to describe them) became very clear, and challenges in evaluation and creating datasets promoted much discussion. The applications and algorithms described showed the breadth of research already undertaken but also highlighted the challenges remaining.

We finish by noting that in this report, we have attempted to reflect both the general sentiment of the discussions and key themes brought up in questions from the workshop attendees as well as by the speakers. As in any discussion, it should not be assumed that these opinions were unanimous. We hope they will prove useful to the reader in understanding the different aspects of diversity, and the many open research directions in this area. Other resources, such as talk slide decks, have been made available on the workshop website, `http://ir.cis.udel.edu/IDR-workshop/`.

5 Acknowledgments

We would like to thank ACM SIGIR for its sponsorship and the SIGIR 2009 committee for its support. We are particularly thankful for the feedback received from the workshop chair (Diane Kelly) and general chairs (James Allan and Javed Aslam), and their quick responses to questions. We would also like to thank contributors, invited speakers and program committee members for their effort in supporting this workshop.

References

[1] ImageCLEF 2009 photo retrieval task. http://www.imageclef.org/2009/photo.

[2] Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, and Samuel Ieong. Diversifying search results. In *Proceedings of WSDM*, pages 5–14, 2009.

[3] Paul N. Bennett, Ben Carterette, Olivier Chapelle, and Thorsten Joachims. Beyond binary relevance: Preferences, diversity, and set-level judgments. *SIGIR Forum*, 42(2):53–58, December 2008.

[4] Charles L. A. Clarke, Maheedhar Kolla, Gordon V. Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. Novelty and diversity in information retrieval evaluation. In *Proceedings of SIGIR*, pages 659–666, 2008.

[5] Charles L.A. Clarke, Maheedhar Kolla, and Olga Vechtomova. An effectiveness measure for ambiguous and underspecified queries. In *Proceedings of ICTIR*, pages 188–199, 2009.

[6] Khalid El-Arini, Gaurav Veda, Dafna Shahaf, and Carlos Guestrin. Turning down the noise in the blogosphere. In *Proceedings of KDD*, pages 289–298, 2009.

[7] Sreenivas Gollapudi and Aneesh Sharma. An axiomatic approach to result diversification. In *Proceedings of WWW*, pages 381–390, 2009.

[8] John Lafferty and ChengXiang Zhai. Document language models, query models, and risk minimization for information retrieval. In *Proceedings of SIGIR*, pages 111–119, 2001.

[9] Xuehua Shen and ChengXiang Zhai. Active feedback in ad hoc information retrieval. In *Proceedings of SIGIR*, pages 59–66, 2005.

[10] Paul Thomas and David Hawking. Evaluation by comparing result sets in context. In *Proceedings of CIKM*, pages 94–101, 2006.

[11] Yisong Yue and Thorsten Joachims. Predicting diverse subsets using structural svms. In *Proceedings of ICML*, pages 1224–1231, 2008.

[12] ChengXiang Zhai, William Cohen, and John Lafferty. Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval. In *Proceedings of SIGIR*, pages 10–17, 2003.

[13] ChengXiang Zhai and John Lafferty. A risk minimization framework for information retrieval. *Information Processing and Management*, 42(1):31–55, Jan 2006.

[14] Xiaojin Zhu, Andrew B. Goldberg, Jurgen Van Gael, and David Andrzejewski. Improving diversity in ranking using absorbing random walks. In *Proceedings of NAACL HTL*, pages 97–104, 2007.