Preface

This volume contains the refereed papers presented at DDR 2011: 1st International Workshop on Diversity in Document Retrieval held on 18th April 2011 in Dublin, Ireland, as part of the 33rd European Conference on Information Retrieval (ECIR 2011).

When an ambiguous query is received, a sensible approach is for the information retrieval (IR) system to diversify the results retrieved for this query, in the hope that at least one of the interpretations of the query intent will satisfy the user. Diversity is an increasingly important topic, of interest to both academic researchers (such as participants in the TREC Web and Blog track diversity tasks), as well as to search engines professionals. In this workshop, we solicited submissions both on approaches and models for diversity, the evaluation of diverse search results, and on applications of diverse search results.

As diversity is, in general, an emerging topic, these is no consensus on various aspects of the topic. The primary aim of this workshop is to foster an interactive, in-depth environment with papers and attendees representing and discussing one of three workshop themes:

- Modelling - e.g. "What are the key components of diversification models?"
- Evaluation - e.g. "How can a better evaluation experiment for diversification be structured?"
- Applications - e.g. "What are the key applications for diversity in commercial search?"

We called for both position and technical papers. Each received submission was reviewed by three program committee members. The committee decided to accept ten papers, each within the context of one of the three workshop themes. The program also includes two invited talks in the Evaluation and Application themes.

We would like to thank the program committee members for the great efforts in reviewing all the submissions, the ECIR 2011 organising committee for their support, and all the authors for their contributions.

March 28, 2011

Craig Macdonald
Jun Wang
Charles Clarke

Table of Contents

Session 3

Applications

Program Committee

Ben Carterette	University of Delaware
Olivier Chapelle	Yahoo! Research
Charles Clarke	University of Waterloo
Nick Craswell	Microsoft
Ben He	Graduate University of Chinese Academy of Sciences
Jaap Kamps	University of Amsterdam
Craig Macdonald	University of Glasgow
Jian-Yun Nie	Universit de Montral
Iadh Ounis	University of Glasgow
Filip Radlinski	Microsoft Research
Tetsuya Sakai	Microsoft Research Asia
Rodrygo Santos	University of Glasgow
Jun Wang	University College London
Ryen White	Microsoft Research
Jianhan Zhu	University College London

Program

09:00-09:10 *Welcome and Opening*

09:10-10:30 **Session 1: Evaluation**

09:10-09:40 *Invited Talk: Challenges in Diversity Evaluation*
Tetusya Sakai

09:40-09:50 *Position Paper: Towards the Foundations of Diversity-Aware Node Summarisation on Knowledge Graphs*
Marcin Sydow

09:50-10:05 *Technical Paper: Analysis of various evaluation measures for diversity*
Praveen Chandar and Ben Carterette

10:05-10:20 *Technical Paper: Novelty and Diversity Metrics for Recommender Systems: Choice, Discovery and Relevance*
Pablo Castells, Saúl Vargas and Jun Wang

10:15-10:30 Discussion

10:30-11:00 **BREAK**

11:00-12:00 **Session 2: Modelling**

11:00-11:10 *Position Paper: Diversifying for Multiple Information Needs*
Rodrygo Santos and Iadh Ounis

11:10-11:20 *Position Paper: A Search Architecture Enabling Efficient Diversification of Search Results*
Gabriele Capannini, Franco Maria Nardini, Raffaele Perego and Fabrizio Silvestri

11:20-11:35 *Technical Paper: Diversification of search results as a fuzzy satisfiability problem*
Steven Schockaert and Martine De Cock

11:35-11:50 *Technical Paper: A Comparative Study of Search Result Diversification Methods*
Wei Zheng and Hui Fang

11:50-12:00 Discussion

12:00-13:00 **Poster Session**

13:00-14:00 **LUNCH**

14:00-15:00 **Session 3: Applications**

14:00-14:30 *Invited Talk: Analysis of Document Diversity through Sentence-Level Opinion and Relation Extraction*
Alessandro Moschitti

14:30-14:40 *Position Paper: Diversity in Expert Search*
Vassilis Plachouras

14:40-14:50 *Position Paper: SOPHIA: bridging the gap between thematic modelling to interactive diverse search*
Niall Rooney, David Patterson and Vladimir Dobrynin

14:50-15:05 *Technical Paper: Explicit Query Diversification for Geographical Information Retrieval*
Davide Buscaldi and Paolo Rosso

15:05-15:25 Discussion

15:25-16:00 **BREAK**

16:00-18:00 **Session 4: Breakout Groups**

16:00-17:00 *Breakout groups: Evaluation, Modeling, Applications*

17:00-18:00 *Wrapup from breakout groups*

Challenges in Diversity Evaluation (Keynote)

Tetsuya Sakai

Microsoft Research Asia

Abstract. In this paper, I first survey existing approaches to evaluating diversified search results very briefly. Then I list up some open problems in this area that might initiate discussions at this workshop. Finally, I report on the ongoing efforts at NTCIR that are related to diversity evaluation.

1 Introduction

Given an ambiguous or underspecifid query and no knowledge of the user, presenting a diversified search result to the user is probably a sensible approach to accomodating different user needs, and this research area has received a lot of attention lately. For example, TREC[1] started the diversity task within the Web track in 2009, and there are ongoing related tasks at NTCIR[2].

Between 2008 and 2010, some new evaluation metrics have been proposed that are designed specifically for diversity evaluation. Unlike *subtopic recall* which simply looks at the number of subtopics (or *intents*) covered by a search output [18], these new metrics consider both relevance and diversity in ranked document retrieval. However, there remain several open problems in diversity evaluation, both within the traditional document retrieval paradigm and beyond.

2 Current Status

There are at least three different approaches to incorporating relevance and diversity in ranked retrieval evaluation. Below, I will briefly discuss their characteristics. Mathematical details can be found elsewhere [1, 8, 9, 14].

2.1 α-nDCG

α-nDCG [8, 9] regards information needs (or *intents*) and documents as sets of *nuggets*. This metric defines graded relevance as the number of different nuggets covered by each document. That is, a document that covers many intents is a highly relevant document. α-nDCG first discounts the value of each retrieved relevant document based on "nuggets already seen," and then further discounts

[1] `http://trec.nist.gov/`
[2] `http://research.nii.ac.jp/ntcir/`

it based on the document rank. For example, if two documents relevant to a particular intent is retrieved, the second relevant document is considered rather *redundant* and receives very little credit. The key idea behind α-nDCG is to encourage diversity by means of discouraging redundancy. A similar metric called NRBP has also been proposed [8, 9].

One of the weaknesses of these nugget-based metrics is that they are difficult to normalise and may not range fully between 0 and 1 [5, 14]. A few other issues will be discussed below.

2.2 Intent Aware Metrics

Given the query "apple," suppose that we somehow know that 80% of the user population seek information on a certain company, while the other 20% seek information on a fruit. The diversifed search engine result page (SERP) probably should reflect this uneven distribution over intents somehow. For example, the SERP can allocate more space to information relevant to the *company* intent. Thus, it seems sensible to incorporate *intent likelihood* in diversity evaluation.

Moreover, search engine companies routinely label URLs using a graded relevance scale for evaluation with metrics such as nDCG [10]. Hence, if we can identify multiple intents per query in advance, it would probably make sense to extend this practice, so that we can obtain *per-intent graded relevance* assessments.

Neither intent likelihood nor per-intent graded relevance is taken into account in the original α-nDCG (although incorporating intent likelihood into this metric has been discussed [8]). In contrast, the *Intent-Aware* (IA) metrics [1] utilise both of them in a straightforward manner. Given intent probabilities estimated in some way [1, 16] as well as per-intent graded relevance data, *nDCG-IA*, for example, can be computed as follows: for each intent, define an ideal ranked list and compute nDCG based on the per-intent graded relevance data; then take an expectation using the intent probabilities. Thus nDCG-IA is designed to satisfy the "average" user.

IA metrics also have weaknesses. One of them is that it does not range fully between 0 and 1: it is usually impossible for a single system output to be ideal for every intent at the same time. More importantly, it has been shown that IA metrics do not necessarily reward diversity: a system that returns many documents relevant to a highly likely intent can receive a very high score without diversifying [9, 14].

2.3 D♯-measures ("Dee Sharp")

D-measures and *D♯-measures*, originally called "div-measures" and "Idiv-measures" [14], have been designed to solve the problems that apply to α-nDCG, NRBP and IA metrics. These new metrics utilise both intent likelihood and per-intent graded relevance data, *and* range fully between 0 and 1. The assumptions behind D-measures are simple: (1) the intents for a given query are exclusive; and (2) the

graded relevance level of a document for an intent is proportional to the relevance probability of that document to the intent. Intuitively, D-measures prefer systems that return documents that are highly relevant to major intents before those that are marginally relevant to minor intents. Moreover, D♯-measures, which are simply linear combinations of *intent recall* (i.e. subtopic recall) and *D-measures*, have demonstrated high *discriminative power* [9, 12, 14] as well as high intuitiveness.

Just like α-nDCG and IA metrics, however, D♯-measures can only handle diversified ranked retrieval with a pre-defined set of possible intents for a given query. As I shall illustrate in the next section, the existing approaches are clearly not sufficient for handling a wider range of search result diversification tasks.

3 Challenges

3.1 Balancing Relevance and Diversity

I have closely examined some pairs of ranked lists from the TREC 2009 diversity task [9, 14] where α-nDCG, nDCG-IA and D♯-nDCG disagreed with one another, to see which metric is more *intuitive* than others. (The original TREC 2009 diversity test collection provides neither intent probabilities nor per-intent graded relevance data, but my colleague Ruihua Song and I have enriched the collection in order to satisfy these two requirements.)

As was mentioned in Section 2.2, IA metrics can be counterintuitive as they do not necessarily encourage diversification. On the other hand, α-nDCG can be counterintuitive as it does not know the difference between a major and a minor intent or that between a highly relevant and a marginally relevant document for an intent. It can also be counterintuitive for *informational* intents, because returning multiple relevant documents for these intents may actually be a good thing, even though α-nDCG regards these documents as redundant. However, there are also cases where it is difficult to say which metric is better. The bottom line is, you have a ranked list with high diversity and low relevance, and one with low diversity and high relevance: which one should you prefer?

Some subquestions: how should we balance diversity and relevance for nagivational and for informational intents? The original nDCG is inherently suitable for informational queries and intents, as the main idea is to *accumulate* pieces of information. For navigational queries and intents, graded-relevance extensions of Reciprocal Rank such as *Expected Reciproal Rank* (ERR) [6] or P^+*-measure* [12] may be more suitable. (ERR assumes that the user is dissatisfied with documents within top $r-1$ and is finally satisfied with a relevant document at rank r, and that graded relevance reflects the satisfaction probabilities. P^+ assumes that the user may examine documents down to rank r_p, where r_p is the rank of the first "most relevant" document within the SERP. Both metrics can be regarded as an instance of the Normalised Cumulative Utility (NCU) metrics family [13].) Moreover, as the TREC diversity task has demonstrated, a query may contain both informational and navigational intents. Can we get the diversity/relevance

balance right for navigational intents and for informational intents, for example by using clickthrough data, and then seamlessly integrate the results in computing evaluation scores? Will the resultant metric match up with user ratings [15]?

3.2 Listing Up Intents

Suppose a SERP is designed to show 10 URLs initially. And suppose we have somehow mined over 100 "intents" for a query from some query log. Then no matter how a system diversifies, it will not get a high intent recall score with its first SERP. Moreover, considering too many "tail" intents may unnecessarily complicate the relevance assessment and evaluation procedures. What is the right *granularity* of intents that we should consider? (Given the query "Web browser," should we consider "Internet Explorer 8" and "Internet Explorer 9" as two distinct intents or just consider "Internet Explorer" along with "Firefox" etc?) Another issue that actually complicates the enumeration of possible intents is *orthogonality*: for example, given the query "apple," suppose we have obtained "Apple the Steve Jobs company" and "apple the fruit" as the two most likely intents. Now, a third intent candidate comes along: "apple products." What does *this* represent? (Is it about cider or about iPad?) Considering these matters, what is the appropriate, systematic way to list up intents that are "good" to include in diversity evaluation?

3.3 Estimating Intent Likelihoods

A few methods exist for estimating intent probabilities given a query [1, 16]. What is the level of accuracy required here for conducting diversity evaluation reliably? For example, do we want the actual intent probabilities, or would it suffice to just rank the intents and thereby define *relative* importance?

3.4 Evaluating Structured URL Lists

Diversity evaluation metrics such as α-nDCG, IA metrics and D♯-measures are for ranked retrieval. However, a flat ranked list may not necessarily be a good presentation format especially when multiple orthogonal dimensions such as relevance and diversity come into play. Clustering, categorisation and dynamic presentation [4] may be useful. If indeed they are, how can we evaluate them properly?

3.5 Evaluating beyond URL Lists

Instead of presenting a list of URLs to the user, search engines can try to satisfy the user immediately after a click on the search button, by presenting (say) a direct answer to the user's information need (e.g. [7]) or a collection of information gathered from different media and sources (e.g. [2]). Diversity is probably important not only in ranked document retrieval but also for these new search

tasks. How can we quantify the system's diversity/relevance performances in such cases? Evaluating *whole page relevance* [3] is a step towards this direction, but can we design more quantitative, repeatable evaluation methods?

Moreover, while satisfying the user immediately is important, we may also have to consider diversity at the other end of the spectrum: diversity in *session-based* IR evaluation (as opposed to query-based) [11]. This is about evaluating interactive and exploratory IR quantitatively, and we probably need a plausible user model that incorporates the user's post-query navigation (e.g. [17]).

Can we build a unified evaluation framework that can handle diversity and relevance within individual ranked lists, across aggregated verticals, and across queries within the same session? Can we keep the new evaluation metric almost as simple and elegant as nDCG? (I believe that an evaluation metric should be easy to compute and easy to interpret.) For example, instead of using the document rank as the basic unit for defining the ideal search scenario (as nDCG and their extensions do), can we use "atomic user action" as the basic unit, and define an ideal user action sequence for a given information need (not query)? Here, a user action could be a click (on a URL, query suggestion, next button etc.), a scroll, a character input via a QWERTY keyboard, or even a short time interval between explicit actions (assumed to represent the user's reading action etc.). Whether our system is an aggregated search system or an exploratory one, we want to give the user as much useful information as possible while minimising the burden on the user's side. A diversified SERP evaluation is but one small step towards this goal. How can we make a giant leap?

Can we build a test collection for the unified evaluation framework?

4 What's Happening at NTCIR-9

At NTCIR-9, two new tasks (well, one task and a "pilot subtask") related to diversity are underway, and the NTCIR-9 final workshop meeting will take place in December 2011 in Tokyo. I am hoping that some of the (smaller) problems discussed above will be tackled along the way.

NTCIR-9 INTENT is similar to the TREC Web Track Diversity Task. The main differences are: (a) participants themselves mine possible intents for each query; (b) per-intent graded relevance assessments are used for evaluating selectively diversified search results; and (c) we consider Chinese and Japanese queries (but not English queries). Thus, in the *subtopic mining* subtask, participants submit a ranked list of *possible intents* for each query. Then, in the *document ranking* subtask, some of these intents are used for computing metrics such as intent recall and D♯-measures. 33 and 32 teams have signed up for the two subtasks, respectively. Details can be found at `http://www.thuir.org/intent/ntcir9/`.

NTCIR-9 One Click Access is a subtask under INTENT, but is more like an independent task. Systems are expected to return a fixed-length textual output (either 500 or 140 Japanese characters) in response to a given query. It is called "1CLICK" as the idea is that the user can access the right information immediately after clicking on the search button. Systems that return important

nuggets first and minimise the amount of text the user has to read will be rewarded. We are experimenting with a new nugget-based evaluation framework, and our query set contains some ambiguous queries. 25 teams have signed up for the task. Details can be found at `http://research.microsoft.com/en-us/people/tesakai/1click.aspx`.

5 Summary

I have briefly discussed the current status on diversity evaluation (within the traditional ranked retrieval paradigm), and then listed up some open questions both for diversified ranked retrieval and for diversification in more advanced systems. Finally I touched upon relevant ongoing efforts at NTCIR. I hope that some of my points will be useful for discussion at the Diversity in Document Retrieval (DDR) workshop and that I will get useful feedback for NTCIR from the DDR attendees.

Acknowledgments

I would like to thank the organisers and reviewers of the DDR workshop at ECIR2011 for giving me the opportunity to write this paper and for their valuable feedback. I also thank the NTCIR-9 INTENT/1CLICK organisers and participants for their courage and effort.

References

1. Agrawal, R., Gollapudi, S., Halverson, A. and Leong, S.: Diversifying Search Results, Proceedings of ACM WSDM 2009 (2009) 5–14
2. Arguello, J., Diaz, F., Callan, J. and Carterette, B.: A Methodology for Evaluating Aggregated Search Results, ECIR 2011, to appear (2011)
3. Bailey, P., Craswell, N., White, R. W., Chen, L., Satyanarayana, A. and Tahaghoghi, S. M. M.: Evaluating Search Systems Using Result Page Context, Proceedings of IIiX 2010 (2010)
4. Brandt, C., Joachims, T., Yue, Y. and Bank, J.: Dynamic Ranked Retrieval, Proceedings of ACM WSDM 2011 (2009) 247–256
5. Carterette, B.: An Analysis of NP-Completeness in Novelty and Diversity Ranking, Advances in Information Retrieval Theory (ICTIR 2009), LNCS 5766 (2009) 200-211
6. Chapelle, O., Metzler, D., Zhang, Y. and Grinspan, P.: Expected Reciprocal Rank for Graded Relevance, Proceedings of ACM CIKM 2009 (2009) 621–630
7. Chilton, L. B. and Teevan, J.: Addressing People's Information Needs Directly in a Web Search Result Page, Proceedings of WWW 2011 (2011)
8. Clarke, C. L. A., Kolla, M. and Vechtomova, O.: An Effectiveness Measure for Ambiguous and Underspecified Queries, Advances in Information Retrieval Theory (ICTIR 2009), LNCS 5766 (2009) 188–199
9. Clarke, C. L. A., Craswell, N., Soboroff, I. and Ashkan, A.: A Comparative Analysis of Cascade Measures for Novelty and Diversity, Proceedings of ACM WSDM 2011 (2011)

10. Järvelin, K. and Kekäläinen, J.: Cumulated Gain-Based Evaluation of IR Techniques, ACM Transactions on Information Systems, 20(4) (2002) 422–446
11. Järvelin, K., Price, S. L., Delcambre, L. M. L. and Nielsen, M. L.: Discounted Cumulated Gain Based Evaluation of Multiple-Query IR Sessions, ECIR 2008, LNCS 4056 (2008) 4–15
12. Sakai, T.: Bootstrap-Based Comparisons of IR Metrics for Finding One Relevant Document, AIRS 2006, LNCS 4182 (2006) 374–389
13. Sakai, T. and Robertson, S.: Modelling A User Population for Designing Information Retrieval Metrics, Proceedings of EVIA 2008 (2008) 30–41
14. Sakai, T., Craswell, N., Song, R., Robertson, S., Dou, Z. and Lin, C.-Y.: Simple Evaluation Metrics for Diversified Search Results, Proceedings of EVIA 2010 (2010) 42–50
15. Sanderson, M., Paramita, M. L., Clough, P. and Kanoulas, E.: Do User Preferences and Evaluation Measures Line Up? Proceedings of ACM SIGIR 2010 (2010) 555–562
16. Song, R., Qi, D., Liu, H., Sakai, T., Nie, J.-Y., Hon, H.-W. and Yu, Y.: Constructing a Test Collection with Multi-Intent Queries, Proceedings of EVIA 2010 (2010) 51–59
17. Yilmaz, E., Shokouhi, M., Craswell, N. and Robertson, S.: Incorporating User Behavior Information in IR Evaluation, SIGIR 2009 Workshop on Understanding the User - Logging and Interpreting User Interactions in Information Search and Retrieval (UIIR 2009) (2009)
18. Zhai, C., Cohen, W. W. and Lafferty, J.: Beyond Independent Relevance: Methods and Evaluation Metrics for Subtopic Retrieval, Proceedings of ACM SIGIR 2003 (2003) 10–17

Analysis of Document Diversity through Sentence-Level Opinion and Relation Extraction

Alessandro Moschitti

Department of Computer Science and Information Engineering
University of Trento
Via Sommarive 14, 38100 POVO (TN) - Italy
moschitti@disi.unitn.it

Abstract. Diversity in document retrieval has been mainly approached as a classical statistical problem, where the typical optimization function aims at diversifying the retrieval items represented by means of language models. Although this is an essential step for the development of effective approaches to capture diversity, it is clearly not sufficient. The effort in Novelty Detection has shown that sentence-level analysis is a promising research direction. However, models and theory are needed for understanding the difference in content of the target sentences.
In this paper, an argument for using current state-of-the-art in Relation and Opinion Extraction at the sentence level is made. After presenting some ideas for the use of the above technology for document retrieval, advanced extraction models are briefly described.

Keywords: Relation Extraction; Opinion Mining; Diversity in Retrieval

1 Introduction

Diversity in document retrieval has been mainly approached as a classical statistical problem, where the typical optimization function aims at diversifying the retrieval items represented by means of language models, see for example the novelty detection track [2]. Although, this is an essential step for the development of effective approaches to diversity in retrieval, it is not sufficient. Indeed, while for standard document retrieval, frequency counts and the related weighting schemes help in defining the most probable user information needs, they play an adversary role in capturing diversity.

For example, when retrieving documents related to the entity *Michael Jordan*, a huge amount of text will be related to the basket player; perhaps other items will be related to the Jordan, statisticians and professor, but very few of them, e.g., will be devoted to the *Michael Jordan* accounting employee for Rolfe, Benson LLP. The occurrences of the latter in Web documents will be so small that no powerful language model will be able to effectively exploit them, considering the ocean of the basket player related information. In other words, there will not be enough statistical evidence to build a language model for such

employee, consequently the related context, e.g. words, can be confused with the one of other documents unrelated to *Michael Jordan*.

The solution of this problem requires the use of techniques for fine grained analysis of document semantics. In a statistical framework this means that we need to extract features semantically related[1] to the object about which the users expressed their information needs. Such features cannot be just constituted by simple context words as the frequency problem highlighted above would prevent them to be effective. In contrast, textual relations between entities like those defined in ACE [8] provide an interesting level of characterization of the target entity. For example, the sole relation *Is_employed_at* can easily diversifies the three *Michael Jordan* above. A search engine aiming at providing diversity in retrieval will need to integrate such technology in the classical language model.

Another interesting dimension of document diversity is the opinion expressed in text. Documents can be 99% similar according to scalar product based on weighting schemes (especially if traditional stoplists are applied) but express a completely different viewpoint. This is manly due to the fact that documents reporting different opinions on some events describe them by manly only changing adjectives, adverbs and syntactic constructions. Typical opinion polarity classifiers can help to separate diverse retrieved documents but, when several events are described, the opinion analysis at the document level is ineffective. In contrast, by extracting topics, opinion holders and opinion expressions would make it possible to retrieve documents that are diverse with respect to events and opinion on them. In this perspective, one main goal of the LivingKnowledge project[2] is to reveal and analyze the diversity of the information in the Web, as well as the potential bias existing on the related sources.

In the reminder of this paper, Section 2 will report on latest results of sentence-level Relation Extraction, Section 3 will describe our approach to opinion mining in LivingKnowledge and finally, Section 4 will derive the conclusions.

2 Sentence-Level Relaton Extraction

The extraction of relational data, e.g. relational facts, or world knowledge from text, e.g. from the Web [26], has drawn its popularity from its potential applications in a broad range of tasks. The Relation Extraction (RE) is defined in ACE as the task of finding relevant semantic relations between pairs of entities in texts. Figure 1 shows part of a document from ACE 2004 corpus, a collection of news articles.

In the text, the relation between *president* and *NBC's entertainment division* describes the relationship between the first entity (person) and the second (organization) where the person holds a managerial position.

To identify such semantic relations using machine learning, three settings have been applied, namely supervised methods, e.g. [27, 7, 12, 30], semi-supervised methods, e.g. [4, 1], and unsupervised methods, e.g. [9, 3]. Work on supervised

[1] At a higher level than the simple lexical co-occurences.

[2] `http://livingknowledge-project.eu/`

Jeff Zucker, the longtime executive producer of NBC's "Today" program, will be named Friday as the new **president** of **NBC's entertainment division**, replacing Garth Ancier, NBC executives said.

Fig. 1. A document from ACE 2004 with all entity mentions in bold.

Relation Extraction has mostly employed kernel-based approaches, e.g. [27, 7, 5, 28, 6, 21, 29]. However, such approaches can be applied to few relation types thus distant supervised learning [14] was introduced to tackle such problem. Another solution proposed in [23] was to adapt models trained in one domain to other text domains.

Although, the supervised models are far more accurate than unsupervised approaches, they require labeled data and tend to be domain-dependent as different domains involve different relations. This is a clear limitation for the purpose of improving diversity retrieval since document aspects like entities and events are typically very diverse and thus require different sources of annotated data.

The drawback above can be alleviated by applying a form of weakly supervision, specifically named distant supervision (DS), using Wikipedia data [3, 14, 10]. The main idea is to exploit (i) relation repositories, e.g. the *Infobox*, x, of Wikipedia to define a set of relation types $RT(x)$ and (ii) the text of the page associated with x to produce the training sentences, which are supposed to express instances of $RT(x)$.

Previous work has applied DS to RE at *corpus level*, e.g., [3, 14]: relation extractors are (i) learned using such not completely accurate data and (ii) applied to extract relation instances from the whole corpus. The multiple pieces of evidence for each relation instance are then exploited to recover from errors of the automatic extractors. Additionally, a recent approach, i.e., [10], has shown that DS can be also applied at level of Wikipedia article: given a target *Infobox* template, all its attributes[3] can be extracted from a given document matching such template.

In contrast, sentence-level RE (SLRE) has been only modeled with the traditional supervised approach, e.g., using the data manually annotated in ACE [7, 12, 30, 5, 28, 29, 6, 21]. The resulting extractors are very valuable as they find rare relation instances that might be expressed in only one document. For example, the relation *President(Barrack Obama, United States*) can be extracted from thousands of documents thus there is a large chance of acquiring it. In contrast, *President(Eneko Agirre, SIGLEX)* is probably expressed in very few documents (if not just one sentence), increasing the complexity for obtaining it.

[3] This is a simpler tasks as one of the two entity is fixed.

2.1 Automated Extraction of General Purpose Relationships

We have proposed a substantial enhancements of SLRE: first, the use of DS, where the relation providers are external repositories, e.g., YAGO [24], and the training instances are gathered from Freebase [13]. These allow for potentially obtaining larger training data and many more relations, defined in different sources.

Second, we have adapted state-of-the-art models for ACE RE, based on Support Vector Machines (SVMs) and kernel methods (KM), to Wikipedia. We used tree and sequence kernels that can exploit structural information and interdependencies among possible labels. The comparative experiments show that our models are flexible and robust to Web documents as we achieve the interesting F1 of 74.29% on 52 YAGO relations. To give a very rough idea of the importance of the results, the document-level attribute extraction based on DS showed an F1 of 61% [10].

Third, we have verified the quality of our SLRE, by manually mapping relations from YAGO to ACE based on their descriptions. We designed a joint RE model combining DS and ACE data and tested it on ACE annotations (thus according to expert linguistic annotators). The improvement of 2.29 percent points (76.23%-73.94%) shows that our DS data is consistent and valuable.

Finally, since our aim is to produce RE for real-world applications, we have experimented with end-to-end systems. For this purpose, we also exploit Freebase for creating training data for our robust Named Entity Recognizer (NER). Consequently, our RE system is applicable to any document/sentence. The satisfactory F1 of 67% for the 52 YAGO relations suggests that our technology can be applied to real scenarios. This is an important piece of evidence that the use of general purpose RE technology for achieving diversity in retrieval is a viable research direction.

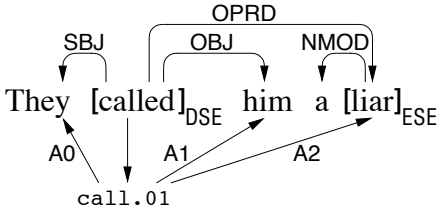

Fig. 2. Syntactic and shallow semantic structure.

3 Sentence-Level Opinion Extraction

Judgements, assessments and opinions play a crucial role in many areas of our societies, including politics and economics. They reflect knowledge diversity in

perspective and goals. The vision inspiring LivingKnowledge (LK) is to consider diversity as an asset and to make it traceable, understandable and exploitable, with the goal to improve navigation and search in very large multimodal datasets (e.g., the Web itself).

To design systems that are capable of automatically analyzing opinions in *free text*, it is necessary to consider syntactic/semantic structures of natural language expressed in the target documents. Although several sources of information and knowledge are considered in LK, we here illustrate an example only focused on text. Given a natural language sentence like for example:

They called him a liar.

the opinion analysis requires to determine: (i) the opinion holder, i.e. *They*, (ii) the direct subjective expressions (DSEs), which are explicit mentions of opinion, i.e. *called*, and (iii) the expressive subjective elements (ESEs), which signal the attitude of the speakers by means of the words they choose, i.e. *liar*.

In order to automatically extract such data, the overall sentence semantics must be considered. In turn, this can be derived by representing the syntactic and shallow semantic dependencies between sentence words. Figure 2 shows a graph representation, which can be automatically generated by off-the-shelf syntactic/semantic parsers, e.g. [11], [15]. The oriented arcs, above the sentences, represent syntactic dependencies whereas the arcs below are shallow semantic (or semantic role) annotations. For example, the predicate *called*, which is an instance of the PropBank [22] frame `call.01`, has three semantic arguments: the Agent (A0), the Theme (A1), and a second predicate (A2), which are realized on the surface-syntactic level as a subject, a direct object, and an object predicative complement, respectively.

Once the richer representation above is available, we need to encode it in the learning algorithm, which will be applied to learn the functionality (subjective expression segmentation and recognition) of the target system module, i.e. the opinion recognizer. Since such graphs are essentially trees, we exploit the ability of tree kernels [16, 20, 17, 19, 18] to represent them in terms of subtrees, i.e. each subtree will be generated as an individual feature of the huge space of substructures.

Regarding practical design, kernels for structures such us trees, sequences and sets of them are available in the SVM-Light-TK toolkit (`http://disi.unitn.it/moschitti/Tree-Kernel.htm`). This encodes several structural kernels in Support Vector Machines, which is one of the most accurate learning algorithm [25].

Our initial test on the LivingKnowledge tasks suggests that kernel methods and machine learning are an effective approach to model the complex semantic phenomena of natural language.

4 Conclusion

In this paper, we have described some limits of only using language models for diversity in document retrieval. As shown by previous work in novelty detection,

an analysis of document at sentence level should be carried out. In this respect, we have shown state-of-the-art natural language processing techniques for Relation Extraction and Opinion Mining, where for the former innovative approaches based on distant supervision allow for training many general purpose relation extractors.

Once accurate sentence analysis is available, several scenarios in the field of Information Retrieval open up:

- Search engines for people retrieval: the availability of automatically derived relations allows for an accurate entity disambiguation;
- Retrieval based on diversity of events: relations along with temporal information constitute basic events and are building blocks of more complex ones;
- Retrieval based on diversity in opinion: retrieval of review fragments targeting a special product or its subpart.

The FET (future emerging technology) project, LivingKnowledge, is studying such innovative approaches to diversity, although the rapid development of the above-mentioned technology suggests that such futuristic approaches are already our present.

Acknowledgements

This research has been supported by the EC project, EternalS – "Trustworthy Eternal Systems via Evolving Software, Data and Knowledge" (project number FP7 247758) and by the EC Project, LivingKnowledge – "Facts, Opinions and Bias" in Time (project number FP7 231126).

References

1. Agichtein, E., Gravano, L.: Snowball: Extracting relations from large plain-text collections. In: Proceedings of the 5th ACM International Conference on Digital Libraries. pp. 85–94 (2000)
2. Allan, J., Wade, C., Bolivar, A.: Retrieval and novelty detection at the sentence level. In: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval. pp. 314–321. SIGIR '03, ACM, New York, NY, USA (2003)
3. Banko, M., Cafarella, M.J., Soderland, S., Broadhead, M., Etzioni, O.: Open information extraction from the web. In: Proceedings of IJCAI. pp. 2670–2676 (2007)
4. Brin, S.: Extracting patterns and relations from world wide web. In: Proceedings of WebDB Workshop at 6th International Conference on Extending Database Technology. pp. 172–183 (1998)
5. Bunescu, R., Mooney, R.: A shortest path dependency kernel for relation extraction. In: Proceedings of HLT and EMNLP. pp. 724–731. Vancouver, British Columbia, Canada (October 2005)
6. Bunescu, R.C.: Learning to extract relations from the web using minimal supervision. In: Proceedings of ACL (2007)
7. Culotta, A., Sorensen, J.: Dependency tree kernels for relation extraction. In: Proceedings of ACL. pp. 423–429. Barcelona, Spain (July 2004)

8. Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S., Weischedel, R.: The automatic content extraction (ace) program?tasks, data, and evaluation. In: Proceedings of LREC. pp. 837–840. Barcelona, Spain (2004)
9. Hasegawa, T., Sekine, S., Grishman, R.: Discovering relations among named entities from large corpora. In: Proceedings of ACL. pp. 415–422. Barcelona, Spain (July 2004)
10. Hoffmann, R., Zhang, C., Weld, D.S.: Learning 5000 relational extractors. In: Proceedings of ACL. pp. 286–295. Uppsala, Sweden (July 2010)
11. Johansson, R., Nugues, P.: Dependency-based syntactic–semantic analysis with PropBank and NomBank. In: Proceedings of the Shared Task Session of CoNLL-2008 (2008)
12. Kambhatla, N.: Combining lexical, syntactic, and semantic features with maximum entropy models for information extraction. In: The Companion Volume to the Proceedings of ACL. pp. 178–181. Barcelona, Spain (July 2004)
13. Metaweb Technologies: Freebase wikipedia extraction (wex) (March 2010), `http://download.freebase.com/wex/`
14. Mintz, M., Bills, S., Snow, R., Jurafsky, D.: Distant supervision for relation extraction without labeled data. In: Proceedings of ACL-AFNLP. pp. 1003–1011. Suntec, Singapore (August 2009)
15. Moschitti, A., Coppola, B., Giuglea, A., Basili, R.: Hierarchical semantic role labeling. In: CoNLL 2005 shared task (2005)
16. Moschitti, A.: Efficient convolution kernels for dependency and constituent syntactic trees. In: Proceedings of ECML'06. pp. 318–329 (2006)
17. Moschitti, A.: Making tree kernels practical for natural language learning. In: Proccedings of EACL'06 (2006)
18. Moschitti, A.: Kernel methods, syntax and semantics for relational text categorization. In: Proceeding of CIKM 2008 (2008)
19. Moschitti, A., Quarteroni, S., Basili, R., Manandhar, S.: Exploiting syntactic and shallow semantic kernels for question/answer classification. In: Proceedings of ACL'07 (2007)
20. Moschitti, A., Zanzotto, F.M.: Fast and effective kernels for relational learning from texts. In: ICML'07 (2007)
21. Nguyen, T.V.T., Moschitti, A., Riccardi, G.: Convolution kernels on constituent, dependency and sequential structures for relation extraction. In: Proceedings of EMNLP. pp. 1378–1387. Singapore (August 2009)
22. Palmer, M., Gildea, D., Kingsbury, P.: The proposition bank: An annotated corpus of semantic roles. Comput. Linguist. 31(1), 71–106 (2005)
23. Riedel, S., Yao, L., McCallum, A.: Modeling relations and their mentions without labeled text. In: Machine Learning and Knowledge Discovery in Databases. Lecture Notes in Computer Science, vol. 6323, pp. 148–163. Springer Berlin / Heidelberg (2010)
24. Suchanek, F.M., Kasneci, G., Weikum, G.: Yago - a core of semantic knowledge. In: 16th international World Wide Web conference. pp. 697–706 (2007)
25. Vapnik, V.N.: Statistical Learning Theory. Wiley-Interscience (1998)
26. Yates, A.: Extracting world knowledge from the web. IEEE Computer 42(6), 94–97 (June 2009)
27. Zelenko, D., Aone, C., Richardella, A.: Kernel methods for relation extraction. In: Proceedings of EMNLP-ACL. pp. 181–201 (2002)
28. Zhang, M., Su, J., Wang, D., Zhou, G., Tan, C.L.: Discovering relations between named entities from a large raw corpus using tree similarity-based clustering. In:

Proceedings of IJCNLP'2005, Lecture Notes in Computer Science (LNCS 3651). pp. 378–389. Jeju Island, South Korea (2005)
29. Zhang, M., Zhang, J., Su, J., , Zhou, G.: A composite kernel to extract relations between entities with both flat and structured features. In: Proceedings of COLING-ACL 2006. pp. 825–832 (2006)
30. Zhou, G., Su, J., Zhang, J., , Zhang, M.: Exploring various knowledge in relation extraction. In: Proceedings of ACL. pp. 427–434. Ann Arbor, USA (June 2005)

Towards the Foundations of Diversity-Aware Node Summarisation on Knowledge Graphs

Marcin Sydow

Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland
Polish-Japanese Institute of Information Technology, Warsaw, Poland,
msyd@poljap.edu.pl

Abstract. This paper aims at initiating a discussion of the foundations of the notion of diversity in a novel problem of computing graphical node summarisations in knowledge multi-graphs (equivalently viewed as RDF-graphs). As it reports an ongoing work, it proposes a general framework of basic concepts and adaptations of two diversity-aware evaluation measures previously studied in the context of information retrieval to the studied problem and briefly discusses them.

Keywords: diversity, node summarisation, evaluation measures, axioms

1 Introduction

Consider a large knowledge base in the form of a directed multi-graph where nodes represent entities from some domain and directed arcs represent some binary relations between the entities. For example, in the movie domain, a directed arc labeled as "acted in" could point from the node labeled as "Woody Allen" to the node labeled as "Zelig" (the title of a movie). Assume that arcs have associated numerical weights that represent some notion of "strength" of the particular relation instance.

Now, imagine the task of local, graphical *summarisation* of a particular node q in such a graph G, i.e. the task of extracting a connected subgraph of G surrounding q, that conveys as complete information about q as possible but has a very limited size.

In this paper, we propose a view on this problem that is based on an analogy with information retrieval as follows. The node q can be viewed as a "query" and the elements of the graph G (e.g. nodes and arcs) as pieces of information ("quasi-documents") to be included in the summary based on their "relevance" to the summarised node.

In this context, the simplest approach to construct a summary seems to greedily select the elements of the surrounding graph in the order of their relevance until the size limit is reached. Such approach has a clear analogy with the PRP principle in IR [11].

Actually, [13] has recently presented a greedy algorithm for computing arc-number-limited entity summarisation on RDF-graphs that works in this way, by selecting edges based on their weighted distance from the summarised node. However, the experimental results in [13] revealed that this approach has the problem of high risk of *redundancy of the information* in the summary (such as the dominance of a single relation name), a problem that is inherent to any greedy, PRP-based approach of this kind known in IR. Due to this, [12] proposed "DIVERSUM", a *diversity-aware* variant of the problem

and appropriate novel algorithm that is diversity-aware, in a simple and intuitive way, by explicitly avoiding edge-label repetition when selecting the edges based on their relevance (proximity) and importance (multiplicity). Furthermore, the recent user evaluation study [14] demonstrated that diversity-awareness introduced in the novel algorithm has been clearly appreciated by the users and resulted in higher-quality summaries.

Thus, as explained above, the diversification approach has *a very natural novel application to the problem of node summarisation in knowledge multi-graphs*, an application whose foundations have not been deeply discussed yet, up to the author's best knowledge. Hence, the focus of this paper is to *initiate the discussion of the theoretical foundations of the concept of diversity in node summarisation*. A desired long-term goal of such a discussion is to propose diversity-aware evaluation measures that would make it possible to design diversity-aware node-summarisation algorithms in a more principled way than was done previously (e.g. in [13, 12]).

Contributions: building on [13, 12, 14], we propose an IR analogy to graphical node summarisation, define the general framework of the basic concepts, consider three diversity axioms and propose two (implicit and explicit) diversity-aware evaluation measures for the problem adapted from measures known in IR and briefly discuss them.

Related Work: [9] proposes a random-walk-based summarisation of an information network (not a single node) that is diversity-aware. [10] studies summarisation of tree-structured XML documents within a constrained budget.

Text summarisation is a mature field, see the [8] survey, for example. The issue of diversity has recently gained much interest in IR community. The fact that the relevance of each retrieved document should be evaluated *dependently* on the other retrieved documents was noticed quite early [6]. An early practical diversity-aware re-ranking algorithm, MMR, utilising so called *implicit* (similarity-based) approach to diversification was proposed in [2] and then became a basis for many followers. The problem of diversity naturally appears in the context of *ambiguous* search queries. [3] studied a related problem of providing at least one document relevant to an ambiguous user query. [5] is an example of an *explicit* approach, that directly models various *aspects* of an *under-specified* query, by means of *information nuggets* and proposes a diversity-aware evaluation measure α-nDCG (later combined with other measures in [4]). [1] proposed a category-based model, "intent-aware" evaluation measure and a greedy algorithm approximately optimising it.

2 Generic Specification of the Problem

The underlying knowledge base is a directed multi-graph G with *unique* labels on nodes (representing entities) and (non-uniquely) labelled arcs (representing binary relations) and rational, non-negative weights on arcs reflecting their "strength".

INPUT: a node q to be summarised and $k \in N$, the limit on the summary's S size defined by the function $l(S) \in N$. We consider the size constraint for practical reasons, due to the limited user comprehension capacity or/and the limited display space (especially important in the context of potential applications on small, mobile devices).

OUTPUT: a (weakly) connected subgraph S of G containing q that *satisfies the size constraint:* $l(S) \leq k$ and *maximises* the properly defined *evaluation measure* $f(S)$. In this paper, we discuss the desired properties and propose specific choices for the evaluation measure f.

Considering the definition of the size function l of the summary S, the potential most natural choices are: number of arcs in S (l_a), number of nodes in S (l_n), sum of the two numbers (l_s). Another, more complex choice for the limit function could be the total length l_t of the textual labels of nodes and arcs in the summary.

We introduce a helper operational notion of *information piece*, inspired by the notion of information nuggets in [4], that represents the unit of information contained in S and view S via the notion of D_S i.e. the *collection of its information pieces*. There is some choice on what to consider a single information piece of S: only nodes, only arcs, unique arc labels (relation names), nodes and arcs, nodes and relation names, etc.

Another important concept related to our problem that should be specified is the notion of "relevance" of information pieces in D_S to q. The simplest approach is to define a function $w : D_S \to [0, 1]$ that represents relevance to q, and the total relevance of S is aggregated over the elements of D_S. The relevance function $w(d)$ can have two components: dynamic (i.e. query-dependent) and static (query-independent). Considering the dynamic component, it should take into account: 1) proximity of information piece d to q in terms of the structure of the underlying graph G (e.g. minimum weighted path length from q to d); and 2) similarity between d and q (e.g. based on textual similarity of the labels or other, more sophisticated notions of similarity, based on some ontology, for example). Considering the static component of w, it can be based on some global properties of the graph G such as centrality or prestige measures, etc. known in the field of social network analysis. Due to space limitations we leave a detailed discussion on how to compute $w(d)$ in multi-graphs for future extension of this work.

Notice our assumption of the connectedness of the resulting summary S that literally means that only those information pieces that are connected to q by a path in G could be considered potentially relevant to q.

Specification of the size function $l(S)$, relevance and similarity functions, and decision of what is to be considered a single information piece, seems to be necessary to start a general discussion of evaluation measures for the node summarisation problem.

3 "Axioms" of Diversity-Aware Evaluation Measures

We propose adaptations of 3 "axioms" out of 8 discussed in [7] (for the context of document retrieval) that could be considered in the context of graphical node summarisation.
1) *monotonicity*: $f(S) \leq f(S')$ for any summaries S, S' such that $D_S \subseteq D_{S'}$ (i.e. adding a piece of information to a summary cannot make it worse);
2) *consistency*: the optimal summary S (according to f) does not change if we make its information pieces D_S more relevant to q and less similar to each other and/or other pieces (outside of D_S) less relevant to q and more similar to each other (this definition is valid only if the relevance $w : D_S \to [0, 1]$ and similarity $s : D_S^2 \to [0, 1]$ functions are defined, we will call it *relevance-consistency* if only w is defined).
3) *stability*: $S \subseteq S'$ for any optimal (according to f) summaries S, S' such that S' has larger size than S (i.e. $l(S) \leq l(S')$). Actually, this property is quite strong and it is not clear if it is really desired for a good evaluation measure. It is possible to imagine reasonable examples when an optimum summary containing 2 information pieces is *not* contained in an optimum summary containing 3 information pieces, etc. On the other hand, this property makes it possible that a greedy algorithm that iteratively selects information pieces to add into the summary can find a global optimum.

4 Diversity-Aware Evaluation Measures

We propose two evaluation measures for the node summarisation problem.

DIVERSUM-Based Evaluation Measure: an "implicit" measure that aims at generalising the approach taken in [12] in the form of a convex combination that directly balances the total relevance and maximum allowed redundancy among information pieces:

$$f(S) = \lambda \sum_{d \in D_S} w(d) - (1 - \lambda)|D_S| \max_{d,d' \in D_S, d \neq d'} s(d, d')$$

$\lambda \in [0, 1]$ is a parameter that controls the balance, $w(d)$ is relevance of d to q and s is a pairwise similarity function among information pieces of D_S. $|D_S|$ coefficient stands for balancing the number of terms in the sum. Alternatively, other aggregation functions can be used instead of sum or max. For example, instead of max, it seems reasonable to use sum (with another normalising coefficient: $\frac{2}{|D_S|-1}$ to account for the number of unordered pairs compared to single elements in D_S, in this case we assume $|D_S| > 1$). Using max instead of the first sum is not a good idea since it would not prevent against the "topic drift", i.e. adding irrelevant pieces to the summary. We conjecture that both variants of the measure satisfy consistency but are not monotonic or stable.

Category-Aware Evaluation Measure: (adapted from [1] to our context) it explicitly focuses on ambiguity of the user information need by modeling the distribution of *categories* of information that a user wishing to summarise a node q may be interested in. Let C denote the set of possible categories (or interpretations of the query). The distribution of query interpretations over categories is modeled by $P(c|q)$ (with $\sum_{c \in C} P(c|q) = 1$). Similarly, the relevance function $w(d|c)$ is category-aware. The measure can be viewed as the expected (over all possible interpretations c) value of the chance of satisfying the user with *at least one* information piece of the summary that is relevant to q in the context of the *actual* interpretation of their unknown interest:

$$f(S) = \sum_{c \in C} P(c|q)(1 - \Pi_{d \in D_S}(1 - w(d|c)))$$

For computational tractability, the measure implicitly assumes independent relevance $w(d|c)$ of information pieces *conditioned* on the actual category (due to product) but does not assume independent relevance that would obviously be counter-diversity-aware. We conjecture that the measure is monotonic and relevance-consistent but not stable. Now, we briefly present some novel ideas on how $P(c|q)$ or $w(d|c)$ can be computed for graphs. $P(c|q)$ can be pre-computed once for each node as a soft membership measure obtained from applying any soft clustering method to the nodes of the knowledge base G that takes proximity and textual labels into account. Our preliminary idea for computing $w(d|c)$ is to use probability of getting to d with a k-limited random walk starting at q where the transition probability is skewed towards arcs and nodes that are more relevant to the particular category c. Due to space limitations, more detailed discussion is left for the extended version of this paper. There exists an efficient greedy approximation algorithm optimising this measure (what is claimed to be NP-hard [1]).

Future Work: 1) Deeper discussion and analysis of proposed measures and their variants; 2) practical ways of computing all their ingredients and effective algorithms. Since some of the above measures lead to non-trivial combinatorial search problems, one can consider brute force (for small size of the summary) or some sub-optimal optimisation heuristics (such as simulated annealing, for example); 3) Extensive experimental evaluation of the proposed methods on real data; 4) Further discussion of diversity "axioms".

Acknowledgements. The author is supported by N N516 481940 and N N516 443038 grants of Polish Ministry of Science and Higher Education.

References

1. Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, and Samuel Ieong. Diversifying search results. In *WSDM*, pages 5–14, 2009.
2. Jaime Carbonell and Jade 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. Harr Chen and David R. Karger. Less is more: probabilistic models for retrieving fewer relevant documents. In *SIGIR 2006: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, Washington, USA, August 6-11, 2006*, pages 429–436. ACM, 2006.
4. Charles Clarke, Maheedhar Kolla, and Olga Vechtomova. An effectiveness measure for ambiguous and underspecified queries. In Leif Azzopardi, Gabriella Kazai, Stephen Robertson, Stefan Rger, Milad Shokouhi, Dawei Song, and Emine Yilmaz, editors, *Advances in Information Retrieval Theory, Proceedings of ICTIR 2009*, volume 5766 of *Lecture Notes in Computer Science*, pages 188–199. Springer Berlin / Heidelberg, 2009.
5. 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 *SIGIR*, pages 659–666, 2008.
6. W. Goffman. A searching procedure for information retrieval. *Information Storage and Retrieval*, 2(2):73–78, 1964.
7. Sreenivas Gollapudi and Aneesh 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. Karen Sparck Jones. Automatic summarising: The state of the art. *Information Processing & Management*, 43(6):1449 – 1481, 2007. Text Summarization.
9. Qiaozhu Mei, Jian Guo, and Dragomir Radev. Divrank: the interplay of prestige and diversity in information networks. In *Proc. of the 16th ACM KDD Conference*, KDD '10, pages 1009–1018, New York, NY, USA, 2010. ACM.
10. Maya Ramanath, Kondreddi Sarath Kumar, and Georgiana Ifrim. Generating concise and readable summaries of xml documents. *CoRR*, abs/0910.2405, 2009.
11. S. Robertson. The probability ranking principle. *J. of Documentation*, 33(4):294–304, 1977.
12. Marcin Sydow, Mariusz Pikuła, and Ralf Schenkel. DIVERSUM: Towards diversified summarisation of entities in knowledge graphs. In *Proceedings of Data Engineering Workshops (ICDEW) at IEEE 26th ICDE Conference*, pages 221–226. IEEE, 2010.
13. Marcin Sydow, Mariusz Pikuła, and Ralf Schenkel. Entity summarization with limited edge budget on undirected and directed knowledge graphs. *Investigationes Linguisticae*, 21:76–89, 2010. `http://www.staff.amu.edu.pl/~inveling/index.php?direct=227`.
14. Marcin Sydow, Mariusz Pikuła, and Ralf Schenkel. To diversify or not to diversify entity summaries on RDF knowledge graphs? In *(to appear in) The Proceedings of the ISMIS 2011 Conference*, Lecture Notes in Artificial Intelligence. Springer Verlag, 2011.

Analysis of Various Evaluation Measures for Diversity

Praveen Chandar and Ben Carterette

Dept of Computer Science, Univ. of Delaware, Newark, DE 19716 USA
{pcr,carteret}@udel.edu

Abstract. Evaluation measures play a vital role in analyzing the performance of a system, comparing two or more systems, and optimizing systems to perform some task. In this paper, we analyze and highlight the strengths and weaknesses of commonly used measures for evaluating the diversity in search results. We compare MAP-IA, α-nDCG, and ERR-IA using data from TREC'09 web track diversity runs and simulated data. We describe a class of test sets that could be used to compare evaluation measure and systems used for diversifying search results.

1 Introduction

IR researchers have long been interested in optimizing systems to provide results for different users with different information needs when they happen to use the same query. This is known as "novelty and diversity"; the goal of the current research program is to be able to optimize and evaluate retrieval systems by:

1. ability to find relevant material;
2. ability to rank relevant material;
3. ability to satisfy diverse needs;
4. ability to rank documents to satisfy diverse needs.

We evaluate 1 and 3 using simple measures like precision/recall or generalizations to "subtopics" like subtopic precision and subtopic recall [12]. These are set-based measures that, taken alone, do not capture anything about the quality of the ranking. Measures like MAP, DCG, ERR are affected by the ranking as well as the relevance of the documents; generalizations like MAP-IA, α-nDCG, and ERR-IA capture diversity and ranking.

Rank-based diversity measures like MAP-IA, α-DCG, and ERR-IA conflate relevance, diversity, and ranking. They are necessary to have a single value for which to optimize system effectiveness, but the more properties we are evaluating with a single measure, the more likely it is that we mistakenly ascribe an improvement in effectiveness to the wrong cause. Our goal in this work is to investigate the degree to which each of relevance, diversity, and ranking influence the outcome of a measurement of MAP-IA, α-DCG, and ERR-IA.

Amongst the diversity runs submitted to the TREC'09 web track [6] we observed that systems with high relevance nearly always had high diversity scores,

while systems with lower relevance were able to achieve higher diversity. This encouraged us to investigate the sensitivity of MAP-IA, α-DCG, and ERR-IA to relevance, diversity, and ranking. We look at real data (runs submitted to the TREC'09 Web track [6]) as well as simulated data covering more possible cases.

2 Analysis of Evaluation Measures

2.1 Evaluation Measures

As discussed above, evaluation measures for diversity account for both relevance and diversity in the ranking. The degree to which a particular measure is dependent on relevance rather than diversity could potentially have a big impact on system design and optimization. In this section, we briefly discuss commonly used evaluation measures for diversity. We use the values of these measures reported by the `ndeval` utility developed for the TREC Web track.

α-nDCG α-nDCG, an extension of DCG [9], uses a position-based user model [8]. The measure takes into account the position at which a document is ranked along with the subtopics contained in the documents. α-nDCG scores a ranking by rewarding newly-found subtopics and penalizing redundant subtopics geometrically, discounting all rewards with a log-harmonic discount function of rank. α is a parameter controlling the severity of redundancy penalization; we use $\alpha = 0.5$ as done for TREC evaluation.

MAP-IA Mean average precision (MAP) is a very well-known evaluation measure for ad hoc retrieval. The "intent-aware" version computes the MAP for each subtopic separately (assuming the documents relevant to that subtopic are the full set of relevant documents; each subtopic is treated as a distinct interpretation for a given query). MAP-IA is then a weighted average over the subtopics [1].

ERR-IA Chapelle et al. proposed an evaluation measure that is based on interdependent ranking [5]. According to this measure, the contribution of each document is based on the relevance of documents ranked above it. The discount function is therefore not just dependent on the rank but also on the relevance of previously ranked documents. Like MAP-IA, ERR-IA is computed by calculating ERR for each subtopic, then computing a weighted average over subtopics.

2.2 Methods of Analysis

Our primary motivation behind this analysis was to find the relative degree of influence of relevance, diversity, and document ranking on each of α-nDCG, ERR-IA, and MAP-IA. In this section, we describe our determination of categories and the methods used to generate data. Since we use the `ndeval` utility developed for the TREC Web track, the parameters (such as $\alpha = 0.5$ for α-nDCG) for the evaluation measures are the same as used in TREC Web track.

Real Systems In order to observe the levels of relevance and diversity on the current systems, we first looked at the 48 runs submitted to the diversity task

	relevace		
diversity	low	medium	high
high	0	4	12
medium	0	14	2
low	15	1	0

Table 1. Classification of TREC 2009 Web diversity runs into a 3-by-3 table of increasing ability to find relevant documents and increasing ability to find diverse documents.

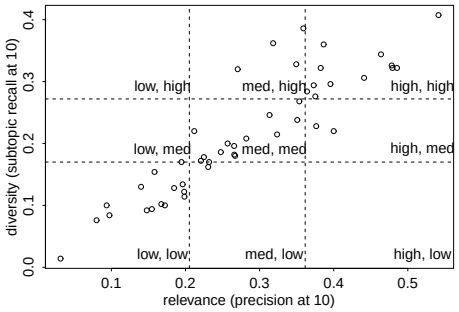

Fig. 1. Precision@10 vs s-recall@10 for 48 systems submitted to the TREC 09 Web track's diversity task. The dashed lines shows relevance and diversity class boundaries.

of the TREC 2009 Web track [6]. We categorized these systems into three levels of relevance based on precision at rank 10 (with a document judged relevant to any subtopic considered relevant for precision@10) and three levels of diversity based on subtopic recall (S-recall) at rank 10 [12] (which is the ratio of unique subtopics retrieved in the top 10 to total unique subtopics). With three levels of each factor, there were nine categories in total. Table 1 gives the number of systems observed in each category. Figure 1 plots S-recall@10 vs precision@10 for these 48 systems to show the breakdown of categories in more detail.

Note that relevance and diversity among these systems are highly correlated. None of the Web track systems have high relevance and low diversity, nor low relevance and high diversity, though both situations are theoretically possible—high relevance/low diversity could be achieved by a system finding many redundant documents, while low relevance/high diversity could be achieved by a system that finds a few relevant documents covering many subtopics. The fact that few systems fall off the diagonal in the Figure 1 suggests that current systems confound relevance and diversity in their ranking approach and therefore may not be good for analyzing general properties of measures.

Simulated Systems Since the real systems do not account for all possible scenarios that we may want to investigate using our measures, we generate several systems in each category using simulations. Since the dependent variable here is the MAP-IA, α-DCG, and ERR-IA scores, the simulated data must be obtained by varying independent variables such as relevance, diversity, document order-

ing, and subtopic distribution. We generated two kinds of simulated systems to study the effect of independent variables on the evaluation measures.

Rel+Div: First, we randomly sample documents from the full Web 2009 *qrels* to create random rankings that satisfy one of our nine experimental conditions: low/medium/high precision@10 and low/medium/high S-recall@10, with labels corresponding to values between $0-0.3$ for low, $0.3-0.6$ for medium, and $0.6-1$ for high. We sampled until we had 10 random rankings in each condition.

Rel+Ord: Next, we controlled diversity ranking in the following way: ten different rankings in each of the same nine relevance/diversity conditions were carefully chosen by varying the minimum rank at which maximum S-recall is obtained. In each category we generate ten rankings in which the documents are re-ordered such that maximum S-recall is obtained only at rank i, where i ranges from 1 to 10. The first ranking (ranking 1, i.e $i = 1$) would attain maximum S-recall at rank 1, the second (ranking 2, i.e $i = 2$) attains max S-recall at rank 2, and so on. In this way we model degrading ability of a system to rank documents.

2.3 Re-ranking Methods

A common way to achieve diversity in a ranking is to first rank by relevance, then re-rank those documents to achieve greater diversity. We briefly describe two re-ranking approaches that we will investigate in this work.

Maximal Marginal Relevance linearly combines a typical bag-of-words relevance score of a document with the amount of "novelty" the document adds to the ranking [2]. The degree of novelty in ranking can be controlled as MMR is a linear combination of relevance and novelty scores. The algorithm prefers documents relevant to the query and least similar to previously ranked documents.

Similarity Pruning is a greedy approach that diversifies the result set by iterating through the initial ranking and removing similar documents [4]. The algorithm iterates over an initial ranking sorted by relevance and prunes documents with similarity scores above a threshold θ.

3 ANOVA

Our goal is to decompose the variance in an evaluation measure into components:

1. variance due to changes in the system's ability to find relevant documents;
2. variance due to changes in the ability of a system to satisfy diverse needs;
3. variance due to changes in the system's ability to rank relevant and diverse documents;
4. variance due to interactions among the above;
5. variance due to topics;
6. variance due to other attributes of a system or other factors.

	SSE in measure (and %age)		
component	ERR-IA	α-nDCG	MAP-IA
relevance	819.0 (22%)	639.9 (16%)	386.2 (11%)
diversity	1075.7 (29%)	1979.3 (52%)	648.8 (20%)
interaction	48.7 (1%)	75.6 (2%)	19.6 (1%)
topic	482.7 (13%)	567.5 (15%)	1362.8 (42%)
residual	1282.5 (35%)	561.5 (15%)	822.1 (25%)

Table 2. Variance decomposition for components affecting the value of each measure. The first three are independent variables we control. The "topic" component is a random effect due to topic sample. The "residual" component comprises everything about the measure that cannot be explained by the independent variables. Percentages sum to 100 (modulo rounding error) for each measure. All effects are significant with $p < 0.01$.

Multi-way analysis of variance (ANOVA) is the statistical tool that we will use. In each of our experiments we have at least two independent factors from numbers 1–3 above, as well as one random effect (TREC 2009 topics) for which we have repeated measures on every independent factor. We will not go into details on computing ANOVA, since they can be found in standard statistics textbooks. The numbers we report are derived from the ANOVA procedures in the statistical programming environment R [10]; they are meant to provide intuition about how much we can distinguish between systems that are different on one factor when the rest are held constant.

There are many ways to evaluate evaluation measures; this is one way, but others include detailed examination of single-topic rankings [11], examination of mathematical properties of measures [3], or other data analytic approaches [7].

4 Results

4.1 Varying relevance and diversity

As described above, our first set of simulated data uses two independent factors—relevance as measured by precision@10 and diversity as measured by S-recall@10—with three levels each. We have 47 topics (after dropping those with two or fewer subtopics) and 10 random rankings at each pair of levels. Thus we have $3 \cdot 3 \cdot 45 \cdot 10 = 4050$ total data points for our ANOVA.

Table 2 shows ANOVA variance decomposition for our three measures of interest. From this table we conclude the following:

1. α-nDCG does a much better job at distinguishing between systems that provide different levels of diversity, with 52% of its variance being explained by diversity level as compared to 29% for ERR-IA and 20% for for MAP-IA.
2. MAP-IA is dominated by random variance due to topic sample. This is because the range of achievable MAP-IAs for a given topic depends heavily on the number and distribution of subtopics in documents [3].

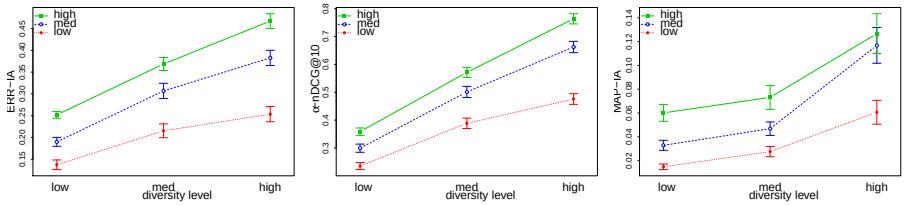

Fig. 2. Effect of increasing diversity and relevance independently on ERR-IA, α-nDCG, and MAP-IA and their standard error over a topic sample.

	SSE in measure (and %age)		
component	ERR-IA	α-nDCG	MAP-IA
relevance	682.3 (16%)	586.0 (16%)	386.2 (9%)
diversity	891.7 (22%)	1174.6 (47%)	648.8 (14%)
ranking alg	1174.6 (29%)	593.7 (16%)	19.6 (3%)
interactions	477.5 (12%)	298.3 (8%)	152.9 (3%)
topic	347.9 (9%)	497.2 (13%)	1362.8 (35%)
residual	375.0 (12%)	288.1 (7%)	822.1 (35%)

Table 3. Variance decomposition for components affecting the value of each measure. The first four are the independent variables we control (interactions between the first three are aggregated together). The "topic" component is a random effect due to topic sample. The "residual" component comprises everything about the measure that cannot be explained by the independent variables or the random effect. Percentages sum to 100 (modulo rounding error) for each measure. All effects are significant with $p < 0.01$.

3. ERR-IA is more strongly affected by unmodeled factors captured in residual error than the other two measures. This may imply that ERR-IA is more sensitive to the ranking of documents than α-nDCG or MAP-IA.
4. Interaction between relevance and diversity plays relatively little role in any of the three measures (though these effects are significant). Our classification of Web track runs suggests interaction effects play a much bigger role in system optimization, however.

Figure 2 shows the mean value of each measure increasing with diversity level for each relevance level, with standard error bars showing randomness due to topic sample. This shows that each measure can distinguish between both different levels of relevance and diversity (as ANOVA analysis suggests). Interestingly, standard error tends to increase with diversity and relevance; this suggests that other factors are affecting the measures more when the systems are better.

4.2 Varying relevance, diversity, and ranking algorithm

The fact that there was so much residual error in the previous results suggests that the ranking algorithm may play a role in determining the measure value (which is not surprising considering that all three use information about ranks).

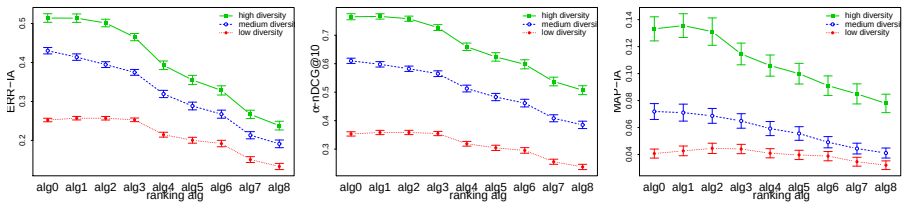

Fig. 3. Effect of degrading a ranking algorithm at independent diversity levels on ERR-IA, α-nDCG, and MAP-IA and their standard error over a topic sample.

To investigate that, we used our data simulating different ranking algorithms; our 10 random rankings above are now non-random levels of a "ranking" factor. Table 3 summarizes the ANOVA analysis; we see the same trends as before regarding diversity, relevance, and topic effects, but now we see ranking accounts for a large amount of variance in the measure. Residual variance decreases, except in MAP-IA; this suggests that MAP-IA is dominated by undesirable factors.

Figure 3 shows the effect of degrading the simulated ranking algorithm on measure value at different diversity levels (averaged over all relevance levels). Note that the maximum ERR-IA values here are much higher than those shown in Figure 2; this is because the ranking of documents is much more important to ERR-IA than either relevance or diversity alone.

4.3 Effect of reranking algorithm

Finally, we looked at whether the initial level of relevance and diversity affect the efficacy of the reranking-for-diversity approaches we describe above. We reranked results for the random systems using the approaches, then looked at the effect of each of our components on variance in the difference in a measure from the initial ranking to the re-ranked results.

Figure 4 shows that MMR and SimPrune work best when there's high relevance and medium diversity in the initial ranking, and worst when there is already high diversity in the initial ranking, likely because both tend to exclude documents from the original ranking. The wide range in the error bars shows that in general relevance is not a strong factor, only being significant at $p < 0.1$.

5 Conclusions

In this paper, we perform a thorough analysis on various evaluation measures for diversity. We observe that ERR-IA is more sensitive to document ranking and α-nDCG is more sensitive to the diversity among documents retrieved. Further, it is interesting to note that MAP-IA is more sensitive to the topic sample and other factors, which is not desirable in any evaluation measure. The re-ranking approaches were found to be influenced more by diversity in the initial ranking than relevance, with only a medium level of diversity being conducive to improving results after re-ranking.

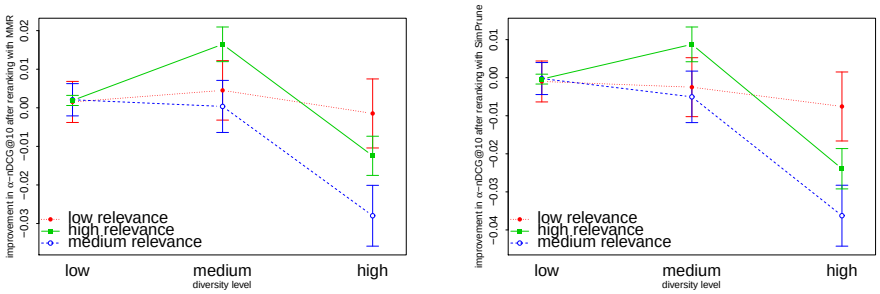

Fig. 4. Effect on α-nDCG@10 of reranking an initial set of results with the given relevance and diversity levels using MMR or SimPrune.

References

1. Agrawal, R., Gollapudi, S., Halverson, A., Ieong, S.: Diversifying search results. In: Proceedings of WSDM '09. pp. 5–14 (2009)
2. Carbonell, J., Goldstein, J.: The use of mmr, diversity-based reranking for reordering documents and producing summaries. In: Proceedings of SIGIR '98. pp. 335–336 (1998)
3. Carterette, B.: An analysis of NP-completeness in novelty and diversity ranking. In: Proc. ICTIR. pp. 200–211 (2009)
4. Carterette, B., Chandar, P.: Probabilistic models of ranking novel documents for faceted topic retrieval. In: Proceeding of CIKM'09. pp. 1287–1296 (2009)
5. Chapelle, O., Metlzer, D., Zhang, Y., Grinspan, P.: Expected reciprocal rank for graded relevance. In: Proceeding of CIKM '09. pp. 621–630 (2009)
6. Clarke, C.L., Craswell, N., Soboroff, I.: Overview of the TREC 2009 web track. In: Proceedings of TREC (2009)
7. Clarke, C.L., Craswell, N., Soboroff, I., Ashkan, A.: A comparative analysis of cascade measures for novelty and diversity. In: Proc. WSDM. pp. 75–84 (2011)
8. 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. pp. 659–666 (2008)
9. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of ir techniques. ACM Trans. Inf. Syst. 20, 422–446 (October 2002)
10. R Development Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2010), http://www.R-project.org, ISBN 3-900051-07-0
11. Sakai, T., Craswell, N., Song, R., Robertson, S., Dou, Z., Lin, C.Y.: Simple evaluation metrics for diversified search results. In: Proc. EVIA (2010)
12. Zhai, C.X., Cohen, W.W., Lafferty, J.: Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In: Proceedings of SIGIR '03. pp. 10–17 (2003)

Novelty and Diversity Metrics for Recommender Systems: Choice, Discovery and Relevance

Pablo Castells[1], Saúl Vargas[1], and Jun Wang[2]

[1] Departamento de Informática, Universidad Autónoma de Madrid, 28049 Spain
[2] Department of Computer Science, University College London, WC1E 6BT, UK

{pablo.castells,saul.vargas}@uam.es, wang.jun@acm.org

Abstract. There is an increasing realization in the Recommender Systems (RS) field that novelty and diversity are fundamental qualities of recommendation effectiveness and added-value. We identify however a gap in the formalization of novelty and diversity metrics –and a consensus around them– comparable to the recent proposals in IR diversity. We study a formal characterization of different angles that RS novelty and diversity may take from the end-user viewpoint, aiming to contribute to a formal definition and understanding of different views and meanings of these magnitudes under common groundings. Building upon this, we derive metric schemes that take item position and relevance into account, two aspects not generally addressed in the novelty and diversity metrics reported in the RS literature.

Keywords: novelty, diversity, metrics, evaluation, recommender systems

1 Introduction

Several approaches to assess novelty and diversity in search results have been proposed in the last few years [1,5,7,8,10]. Datasets have been released at evaluation campaigns such as TREC 2009/10, fostering convergence and sharing of common benchmarks. Metrics such as α-nDCG, nDCG-IA, MAP-IA, ERR-IA, NRBP have been used in the diversity task of the TREC Web track [6]. Further diversity-related metrics have been proposed outside this, such as subtopic precision and recall [5] or k-call [10]. Even though diversity and novelty are a largely open research topic in IR today, one can see a fair extent of convergence and reuse of metrics and methodologies in the community.

In contrast, studies of comparable depth on measuring novelty and diversity, and/or an array of well-understood metrics, are still missing in the Recommender Systems (RS) area. In fact, the range of metrics described in the literature is considerably scant. For instance, to the best of our knowledge, a measure that takes into account the order of recommended items is completely missing –except for the obvious application of diversity metrics at different top-n cutoffs. Yet novelty and diversity play an arguably even more central role in the recommendation context, where the practical value and gain from recommendation are closely linked to the notion of discovery in most scenarios. Moreover, the ambiguity in user needs is considerably higher than in ad-hoc IR, and intrinsic to the task, since there is no explicit expression of such needs. Despite a significant and growing stream of research and interest in diversity and novelty in the field [11,12,13], there would seem to be a gap in the definition and systematic study of metrics.

Following this motivation, we discuss the definition of suitable metrics for the needs and specifics of RS regarding novelty and diversity. Our study considers two main ground concepts in recommendation novelty, namely item *similarity* and user-item *interaction*, upon which different recommendation novelty and diversity models unfold. User-item interaction is in turn modeled upon three core conditions: *choice*, *dis-*

covery, and *relevance*. As a specific result, we find modular means to introduce rank and relevance sensitiveness in the metrics, two properties currently not present in the diversity and novelty metrics reported in the RS literature.

2 Novelty and Diversity in Recommender Systems

Novelty and diversity are different though related notions. The novelty of a piece of information generally refers to how different it is with respect to "what has been previously seen", by a specific user, or by a community as a whole. Diversity generally applies to a set of items, and is related to how different the items are with respect to each other. This is related to novelty in that when a set is diverse, each item is "novel" with respect to the rest of the set. Moreover, a system that promotes novel results tends to generate global diversity over time in the user experience; and also enhances the global "diversity of sales" from the system perspective. Another fundamental take on diversity is defined in ad-hoc IR in terms of query interpretations or aspects. The adaptation of this perspective to a recommendation task certainly deserves investigation. For reason of available space, we leave it aside in the present paper, and we only discuss here –besides novelty itself– notions of diversity that result from a novelty model, as we shall see. Moreover, we focus on novelty and diversity as perceived by the end-user, i.e. we do not cover here the system or the business perspective. Finally, we assume an application scenario where the items that the user has already chosen in the past are not recommended again –leaving out scenarios such as recommendation for grocery shopping (where the same products are bought periodically), or personalized music playlist generation (where it is generally ok to recommend known music tracks).

We distinguish two main notions upon which recommendation novelty and diversity can be defined: item popularity and similarity. Recommendation novelty and diversity can be modeled upon the novelty and dissimilarity of recommended items, which in turn we formalize in terms of user-item interaction models, and distance functions. Novelty and diversity can be measured generically, that is, irrespective of the user they are delivered to, or they can actually take into account the target user. In a generic approach, the diversity in a list of items can be measured, for instance, in terms of the objective variety of items in the list (e.g. as pairwise dissimilarity), and novelty can be defined in terms of how many users are familiar with the items. In a user-relative approach, novelty can take into account what the specific target user has already seen, and diversity can consider the variety of interests within his individual user profile. Metrics may just analyze the composition of recommended lists, or they may also take into account that the top positions have a higher impact on the effective diversity and novelty value of the list. A metric may strictly focus on novelty, leaving relevance for a complementary metric to capture it, or actually require items to be relevant for their novelty to be counted in. Which among all such variants is more appropriate depends on the evaluation goals and requirements, the specifics of the recommendation task and/or the application domain.

3 Item Novelty Models

We consider two models of item novelty, one oriented to popularity, and one based on inter-item distance. We consider two variants for each formulation: generic and relative to an item set. In general we will consider two items sets in relative novelty: the profile of the target user, and (in distance-based models) a list of recommended items. Each will induce different recommendation metrics, as we shall see later on in Section 5.

Popularity-based Item Novelty. The novelty of an item can be defined relative to a set of observed events on the set of all items. A common way to formalize the *generic novelty* of an item is by the amount of information its observation conveys [12], in terms of some distribution involving the item. This is expressed in Information Theory as:

$$novelty(i) = I(i) = -\log_2 p(i) \quad (1)$$

where $p(i)$ represents the probability that i is observed, and $I(i)$ is commonly called self-information or surprisal. In this model, we propose to interpret the i random variable as an event of user choice, that is, "i is picked" by a random user. This reflects a factor of item popularity, whereby $novelty(i)$ corresponds to the log of the *inverse popularity*. High novelty values correspond to long-tail items in the density function, that few users have chosen or interacted with, and low novelty values correspond to popular head items. This scheme measures generic novelty as far as it is the same for all users. A *user-relative novelty* variant can be defined by simply taking $p(i|u)$ in equation 1, which amounts to restricting our observations to the target user:

$$novelty(i|u) = -\log_2 p(i|u) \quad (2)$$

An alternative, discovery-based popularity model is to consider the probability $p(K|i)$ that an item is known or is familiar to (rather than chosen by) a random user. In this case, we define generic and user-relative novelty respectively as:

$$novelty(i) = 1 - p(K|i) \qquad novelty(i|u) = 1 - p(K|i,u) \quad (3)$$

In order to emphasize the effect of highly novel items (favoring few very novel items vs. many moderately novel), one may also consider the logarithm of the inverse probability:

$$novelty(i) = -\log_2 p(K|i) \qquad novelty(i|u) = -\log_2 p(K|i,u) \quad (4)$$

Distance-based Item Novelty. Relative novelty can also be modeled with respect to a set of items on a Euclidean view. This can be defined as the *average* or *minimum distance* between the item at hand, and the items in the set:

$$novelty(i|S) = \sum_{j \in S} p(j|S) d(i,j) \quad \text{or} \quad novelty(i|S) = \min_{j \in S} d(i,j) \quad (5)$$

where d is some distance measure. The distance can be defined e.g. as $d(i,j) = 1 - sim(i,j)$ for some similarity measure (cosine-based, Pearson correlation, etc., normalized to [0,1]) in terms of the item features (content-based view) or their user interaction patterns (collaborative view). If we take S as the set of items a user has interacted with (i.e. the items in his profile), we get a user-relative novelty version of equation 5.

In the next section we discuss several estimation approaches for the distributions that have come up so far. $p(i|S)$ will be discussed only in the case where S is a user profile. In section 5 we discuss the case where $S = R$, a list of recommended items.

4 Ground Models

The models upon which novelty is defined in the previous section can use different estimation approaches, depending on the availability and type of observation data, the choice of random variables and any additional restriction on the observed events upon which the distributions are estimated. We broadly distinguish three main categories of user-item relationships:

- *Choice*: an item is used, picked, selected, accessed, browsed, bought, etc. It is common to have a frequency associated to this event, though the relation can also be binary (e.g. one-time purchase).

- *Discovery*: an item has/has not been seen before. This is understood as a binary fact, independently from the frequency of interaction, or the degree of enjoyment / dislike.
- *Relevance*: in the context of RS, relevance can be related to notions of preference, i.e. how much a user likes or enjoys an item, or how useful the item is.

Choice and discovery aspects in the interaction between users and items gives rise to different novelty and diversity model variants, which can be implemented in different ways depending on the available data, as we discuss next. We do not see relevance as playing a role in the popularity models discussed in the previous section, but we also discuss it here, as another ground aspect of user-item interaction modeling, which we shall use later on in recommendation metrics. Choice models are most naturally associated to observations in the form of usage data, whereas relevance models are best estimated upon explicit user ratings, and both types of observation suit discovery modeling well. We nonetheless discuss estimation approaches for choice in terms of ratings as well, and relevance in terms of usage. As a general rule, choice and discovery model estimates should use training data only –preferably separate from the training data used by the recommender–, whereas relevance estimates should use test data.

Item Choice. As a simple abstraction for observed usage, let us assume the observed data consists of a set Λ of user/item/timestamp records, reflecting item access by users (e.g. in an online music site). Taking $p(i)$ as the probability that i is used, a maximum likelihood item prior estimate would be:

$$p(i) \sim \frac{|\{(u,i,t) \in \Lambda\}|}{|\Lambda|} \tag{6}$$

Under this formulation, novelty as defined in equation 1 is the so-called *inverse collection frequency* ICF of the item. The posterior $p(i|u)$ for user-relative novelty (equation 2) is trickier as far as we should assume no observation of u accessing i in the past (as stated in the introduction, we otherwise assume i would not be recommended to u). We can take an indirect estimate based on other items the user has accessed:

$$p(i|u) \sim \sum_{j \in \mathbf{u}} p(i|j)p(j|u), \quad p(j|u) \sim \frac{|\{(u,j,t) \in \Lambda\}|}{|\{(u,k,t) \in \Lambda\}|} \text{ if } j \in \mathbf{u}, \quad p(i|j) \sim \frac{|\mathbf{i} \cap \mathbf{j}|}{|\mathbf{j}|} \tag{7}$$

where **u** denotes the set of items in the user's profile, and **i** the set of users who have accessed i. User ratings are sparse observations to support choice models, but can still enable an acceptable rough estimation if enough data are available. This can be done by equating a positive rating (i.e. above some threshold τ) to a one time access observation (as in e.g. a purchase), which fits as a model for equations 6 and 7 above (see Table 1).

Item Discovery. The prior that a random user knows about i can be estimated as:

$$p(K|i) \sim \frac{|\mathbf{i}|}{|\mathrm{Y}|} = \frac{|\{u \in \mathrm{Y} | \exists t \in \mathrm{T} : (u,i,t) \in \Lambda\}|}{|\mathrm{Y}|} \tag{8}$$

where Y is the set of all users, and T is the timestamp data type. In this formulation, item novelty in equation 4 becomes the *inverse user frequency* IUF [2]. When the observed data consists of item ratings by users, this becomes:

$$p(K|i) \sim \frac{|\mathbf{i}|}{|\mathrm{Y}|} = \frac{|\{u \in \mathrm{Y} | r(u,i) \neq \emptyset\}|}{|\mathrm{Y}|} \tag{9}$$

where $r(u,i) \neq \emptyset$ means the rating of u for i is known. If the interaction between users and items is binary (e.g. one-time purchase), then equations 8 and 9 are the same. To model $p(K|i,u)$ in user-relative novelty, assuming again no past interaction between u and i, we can take an indirect estimate: $p(K|i,u) \sim \sum_{j \in \mathbf{u}} p(K|j)p(i|j)\, p(j|u)/p(i|u)$.

Item Relevance. Relevance in RS can be equated to the user interest for items. How relevance can be modeled depends again on the nature of available observations. For usage logs, a correspondence can be fairly established between item usage counts and user interest, in such a way that probability estimates of an item being used –$p(i|u)$– can be (properly scaled and) taken as a reasonable proxy for the probability of the item being liked (i.e. relevant). Under this view, the approaches discussed above for item choice (eq. 6 and 7) would apply here. If instead the available input consists of explicit user ratings, the probability of items being liked can be modeled by some heuristic mapping between rating values and probability of relevance. For instance, drawing from the ERR metric scheme [4]:

$$p(rel|i,u) \sim \frac{2^{g(u,i)} - 1}{2^{g_{max}}}$$

where g is a utility function to be derived from ratings, e.g. $g(u,i) = \max(0, r(u,i) - \tau)$, where τ represents the "indifference" rating value, as proposed by Breese et al [2].

The estimation approaches described here thus provide complete means to instantiate and compute the novelty models defined in the preceding subsection, in terms of item/user frequencies and/or rating data. Table 1 below summarizes some of the combinations that result from the alternatives discussed so far.

Table 1. Summary of item novelty models (smoothing to be applied in the estimates as appropriate), where I denotes the set of all items.

			Model estimation	
	Novelty model	Approach	Usage data	Rating data
Generic	Item choice (ICF)	$-\log_2 p(i)$	$p(i) \sim \frac{\lvert\{(u,i,t) \in \Lambda\}\rvert}{\lvert\Lambda\rvert}$	$p(i) \sim \frac{\lvert\{u \in \mathrm{Y} \lvert r(u,i) > \tau\}\rvert}{\lvert\{(u,j) \in \mathrm{Y} \times \mathrm{I} \lvert r(u,j) > \tau\}\rvert}$
	Item discovery (IUF)	$1 - p(K\|i)$ $-\log_2 p(K\|i)$	$p(K\|i) \sim \frac{\lvert\{u \in \mathrm{Y} \lvert \exists t \in \mathrm{T} : (u,i,t) \in \Lambda\}\rvert}{\lvert\mathrm{Y}\rvert}$	$p(K\|i) \sim \frac{\lvert\{u \in \mathrm{Y} \lvert r(u,i) \neq \emptyset\}\rvert}{\lvert\mathrm{Y}\rvert}$
Relative	Item choice	$-\log_2 p(i\|u)$	$p(i\|u) \sim \sum_{j \in \mathbf{u}} \frac{\lvert\mathbf{i} \cap \mathbf{j}\rvert}{\lvert\mathbf{j}\rvert} \frac{\lvert\{(u,j,t) \in \Lambda\}\rvert}{\lvert\{(u,k,t) \in \Lambda\}\rvert}$	$p(i\|u) \sim \sum_{j \in \mathbf{u}} \frac{\lvert\mathbf{i}_\tau \cap \mathbf{j}_\tau\rvert}{\lvert\mathbf{j}_\tau\rvert\lvert\mathbf{u}_\tau\rvert} \Big\lvert \begin{matrix} \mathbf{u}_\tau = \{i \in \mathrm{I} \lvert r(u,i) > \tau\} \\ \mathbf{i}_\tau = \{u \in \mathrm{Y} \lvert r(u,i) > \tau\} \end{matrix}$
	Item discovery	$1 - p(K\|i,u)$ $-\log_2 p(K\|i,u)$	$p(K\|i,u) \sim \sum_{j \in \mathbf{u}} p(K\|j) \frac{\lvert\mathbf{i} \cap \mathbf{j}\rvert}{\lvert\mathbf{j}\rvert} p(j\|u)/p(i\|u)$	
	Avg. user distance	$\frac{1}{\lvert u \rvert} \sum_{j \in \mathbf{u}} d(i,j)$	–	–
	Min. user distance	$\min_{j \in \mathbf{u}} d(i,j)$	–	–

5 Recommendation Novelty and Diversity Metrics

As a general scheme, we define metrics on recommender systems' output as the expected novelty of the recommended items:

$$m(R) = \sum_{i \in R} p(i|R) novelty(i) \tag{10}$$

where R is the list of recommended items. An interesting user-relative derivation of this formulation consist in modeling $p(i|R,u)$ by considering that a user u picks item i if a) the user browses as far as the position of i in the ranking, and b) he decides to pick i because he is interested in it (relevance). If we assume that both facts are independent, based on a generic user model, who at each position k in the ranking continues browsing down to the next position with some probability $p(k)$ (as modeled in [9]), we get:

$$m(R|u) = \sum_n \left(\prod_{k<n} p(k) \right) p(rel|i_n, u) novelty(i_n|u)$$

where i_n is the item at position n in R and $p(rel|i_n,u)$ is the probability that u finds i_n

relevant. The term $p(rel|i_n, u)$ thus introduces a condition of relevance: the potential novelty of i_n shall be counted in the overall novelty assessment as much as the user possibly likes the item. This is in contrast with the metrics in the RS literature, which focus on novelty or diversity only, and require a complementary accuracy metric to capture relevance. Furthermore, $p(k)$ introduces a component that makes the metric rank-sensitive. For instance, similar to the RBP scheme [9], we may consider a constant $p(k) = p$ for all k, thus getting an exponential rank discount: $(R|u) = (1-p)\sum_n p^{n-1} p(rel|i_n, u)$ $novelty(i_n|u)$. Other heuristic discount schemes can be considered as well, such as a logarithmic discount (as in nDCG), a linear discount, etc. The general form would thus be:

$$m(R|u) = \sum_n disc(n) p(rel|i_n, u) novelty(i_n|u) \tag{11}$$

where $disc(n)$ is the discount function. The discount, the relevance term, and the user-dependencies can be included or excluded as best fits the evaluation requirements.

The schemes discussed so far apply to all the item novelty models defined in Section 3. In general, popularity and user-relative distance give rise to recommendation novelty metrics, whereas distance-based set-relative novelty with respect to R results in recommendation diversity metrics, as we discuss next.

Recommendation Novelty Metrics. Introducing the simple self-information (choice-oriented) item novelty model defined in equation 1 into equation 10, we get:

$$novelty(R) = -\sum_{i \in R} p(i|R) \log_2 p(i)$$

which gives a measure of overall recommendation novelty. Under an item choice model, this can be read as the expected ICF of the recommended items. Further, if we make the approximation $p(i|R) \sim p(i)\,|\mathrm{I}|/|R|$ for $i \in R$ (I being the set of all items), it can be seen that this turns out to be $novelty(R) \sim H(R) + C$, which is (except for a constant $C = \log_2 \frac{|\mathrm{I}|}{|R|}$) the entropy of R under the $p(\cdot\,|R)$ distribution, a common RS novelty metric [12]. Alternatively, the rank-sensitive, relevance-aware development, and the user-specific variants discussed above (equation 11) may apply, whereby we get enhanced alternatives to the plain entropy metric.

Using the discovery-oriented popularity model defined by equation 4, we get:

$$novelty(R) = -\sum_{i \in R} p(i|R) \log_2 p(K|i)$$

which corresponds to the expected IUF of the recommended items. Using equation 3 instead of 4, we get $novelty(R) = \sum_{i \in R} p(i|R)\big(1 - p(K|i)\big)$, the expected probability that an item in the recommended list is not known by the user. This can also be read as the expected number of unknown items in the recommendation, a natural and direct measure of novelty. Again, relevance and rank bias can be introduced by refining $p(i|R, u)$ into $disc(n)p(rel|i_n, u)$ in the user-relative discovery-oriented formulations.

If we take a distance-based user-relative novelty model (equation 5 with $S = \mathbf{u}$), starting from equation 11, we get an alternative novelty measure consisting of the expected distance between the recommended items and the items in the user profile:

$$novelty(R|u) = \sum_{n, j \in \mathbf{u}} disc(n) p(rel|i_n, u) p(j|u) d(i, j)$$

where $p(j|u)$ can be e.g. simplified to a uniform distribution, or be understood as an additional relevance factor, equated to $p(rel|j, u)$, in which case item relevance would be twice accounted for in the metric.

Novelty-based Diversity Metrics. Taking on from equation 11, and instantiating set-relative distance-based novelty models (eq. 5) with $S = R$, we get a measure of diversity:

$$diversity(R|u) = \sum_{n,k} disc(n)p(rel|i_n, u)p(i_k|R)d(i_n, i_k)$$
$$= 2\sum_{k<n} disc(n)disc(k)p(rel|i_n, u)d(i_n, i_k) \quad (12)$$

This general form provides a rank-sensitive and doubly rank-aware *expected intra-list diversity* metric (where assuming d is symmetric and since $d(i, i) = 0$, it is enough to sum for $k < n$). Equation 12 generalizes the average intra-list distance –used in several works on recommendation diversity [11,13]– with the introduction of rank-sensitivity and relevance. Again, the discount and relevance factors can be included or excluded as best fits the evaluation requirements. In particular, if we simplify the discount factors to uniform priors at each raking position (no discount is applied), and relevance is not considered in the model, equation 12 reduces to plain average intra-list diversity: $diversity(R|u) = \frac{2}{|R|(|R|-1)}\sum_{k<n} d(i_n, i_k)$, as used in the RS literature.

6 Experimental Results

Table 2 shows the value of particular instantiations of the above metric schemes in an experiment with MovieLens 100K data. The metrics apply to a common state of the art kNN collaborative filtering recommender (user-based with 50 neighbors), and three diversification algorithms on the baseline output, which rerank the top 500 items based on three diversification algorithms: an adaptation of IA-Select [1], two MMR schemes [3] (with diversity components based on Genre similarity and IUF, respectively, both tuned towards high diversity with λ = 0.6), and a random reranking.

Table 2. Sample results for three representative metric schemes (generic novelty, user-relative novelty, generic diversity), in four configurations: rank and relevance insensitive (None), rank-sensitive (Rank), relevance-aware (Rel), and Both. Values better than random are in bold, italics indicate above the kNN baseline, and the best value for each metric is underlined. All differences are statistically significant (Wilcoxon $p < 0.01$) except when in parenthesis (w.r.t. random) and brackets (kNN).

	Expected IUF (EIUF@50)				Expected profile dist. (EPD@50)				Expected ILD (EILD@50)			
	None	Rank	Rel	Both	None	Rank	Rel	Both	None	Rank	Rel	Both
kNN	3,3815	(3,4149)	**0,2108**	**0,2178**	0,8289	0,8303	**0,0529**	**0,0541**	0,7944	0,4812	**0,0507**	**0,0323**
IA-Select	3,1983	3,2753	**0,1814**	**0,1918**	***0,8707***	***0,8630***	**0,0510**	**0,0519**	***0,8836***	***0,5379***	***0,0516***	***0,0331***
MMR-dist	3,3598	(*3,4487*)	**0,2065**	**0,2162**	***0,8666***	***0,8733***	***0,0545***	***0,0559***	***0,8900***	***0,5453***	***0,0562***	***0,0360***
MMR-iuf	***4,5297***	***4,8009***	***0,2478***	***0,2649***	*0,8319*	(*0,8344*)	**0,0467**	**0,0472**	[0,7917]	[0,4795]	**0,0450**	**0,0282**
Random	*3,4326*	[*3,4396*]	0,1726	0,1729	*0,8371*	*0,8370*	0,0436	0,0436	*0,8268*	*0,5004*	0,0439	0,0275

It can be seen how the different metrics capture different aspects of recommendations. Random reranking beats the baseline in all of the relevance-unaware metrics (and is even second best on EIUF), an effect that is consistently reversed with the introduction of relevance. Relevance also reveals an above-random performance by IA-Select and MMR-dist on EIUF (unnoticed by the relevance-unaware variant). Note to this respect that EIUF and EILD in the "None" variant correspond with approaches reported in the RS literature. MMR-dist is best at relevance except for EIUF, where MMR-iuf is best, as one would expect, as it greedily targets IUF. It can also be seen that rank sensitivity uncovers a better performance by MMR-dist over the baseline and random reranking on EIUF without relevance –which is not perceived when disregarding the ranking. It also shows that while IA-

Select is slightly better than MMR-dist at pure EPD regardless of item order, MMR-dist ranks the novel items better. The overall effect of rank is less significant than relevance in this experiment though. This is because the diversifiers rerank the top 500 items, while the metrics take a fairly shorter top 50 cutoff, thereby capturing rank improvement to some extent even in the rank-unaware variants. Experiments with different baselines and configurations (which we omit here for lack of space) confirm and extend our observations.

7 Conclusion

The presented study aims to contribute to the understanding of the different perspectives on novelty –and derived diversity– in RS, laying out the different views, alternatives, variants, and means of estimation, upon a common, formalized ground. Our effort aims to cover and generalize the metrics reported in the RS literature [11,12,13], and derive new ones. Two novel features in novelty and diversity measurement arise from our study: ranking sensitivity, and relevance-awareness. Both aspects are introduced in a generalized way by easy to configure terms in any metric supported by our scheme. Preliminary experiments confirm our hypotheses and provide initial observations on the behavior of the different metric configurations. Room remains for deeper examination, and additional empirical studies in specific tasks and scenarios to provide further insights on the qualities of the metrics for different purposes.

Acknowledgments. This work was supported by the Spanish Ministry of Science and Innovation (TIN2008-06566-C04-02) and the Government of Madrid (S2009/TIC-1542).

References

1. R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. WSDM 2009, Barcelona, Spain, pp. 5-14.
2. J. S. Breese, D. Heckerman, and C. Kadie. Empirical analysis of predictive algorithms for collaborative filtering. UAI 1998, Madison, WI, USA, pp. 43-52.
3. J. G. Carbonell, J. Goldstein. The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. SIGIR 1998, Melbourne, Australia, pp. 335-33.
4. O. Chapelle, D. Metzler, Y. Zhang, and P. Grinspan. Expected Reciprocal Rank for Graded Relevance. CIKM 2009, Hong Kong, China, pp. 621-630.
5. H. Chen and D. R. Karger. Less is More. SIGIR 2006, Seattle, WA, USA, pp. 429-436.
6. C. L. A. Clarke, N. Craswell, and I. Soboroff. Overview of the TREC 2009 Web Track. TREC 2009, Gaithersburg, MD, USA.
7. C. L. A. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkann, S. Büttcher, and I. MacKinnon. Novelty and diversity in information retrieval evaluation. SIGIR 2008, Singapore, pp. 659-666.
8. C. L. A. Clarke, M. Kolla, and O. Vechtomova. An Effectiveness Measure for Ambiguous and Underspecified Queries. ICTIR 2009, Cambridge, UK, pp. 188-199.
9. A. Moffat and J. Zobel. Rank-biased precision for measurement of retrieval effectiveness. ACM Transactions on Information Systems 27(1), December 2008.
10. C. Zhai, W. W. Cohen, and J. Lafferty. Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval. SIGIR 2003, Toronto, Canada, pp. 10-17.
11. M. Zhang and N. Hurley. Avoiding Monotony: Improving the Diversity of Recommendation Lists. RecSys 2008, Lausanne, Switzerland, pp. 123-130.
12. T. Zhou, Z. Kuscsik, J-G. Liu, M. Medo, J. R. Wakeling, and Y-C. Zhang. Solving the apparent diversity-accuracy dilemma of recommender systems. Proceedings of the National Academy of Sciences of the United States of America 107(10), 2010, pp. 4511-4515.
13. C-N. Ziegler, S. M. McNee, J. A. Konstan, G. Lausen. Improving recommendation lists through topic diversification. WWW 2005, Chiba, Japan, pp. 22-32.

Diversifying for Multiple Information Needs

Rodrygo L.T. Santos and Iadh Ounis

School of Computing Science
University of Glasgow
G12 8QQ, Glasgow, UK
{rodrygo,ounis}@dcs.gla.ac.uk

Abstract. Several approaches have been proposed in recent years to diversify the search results for an ambiguous or underspecified query. In common, most of these approaches are driven by intrinsic characteristics of the search results, such as their content or their coverage of a particular taxonomic scheme. In this position paper, we argue that a true diversification should be driven by the perspective of the search users as opposed to the perspective of the search results. In particular, we claim that an ambiguous query should be regarded as representing multiple possible information needs. The effectiveness of diversifying for multiple information needs is supported by our recent empirical results.

1 Introduction

Query ambiguity is a problem for information retrieval (IR) systems in general, and for web search engines in particular [18]. While an ambiguous query (e.g., 'jaguar') is open to multiple *interpretations* (e.g., 'animal', 'car', 'guitar'), a query with a clearly defined interpretation (e.g., 'jaguar car') may still be underspecified, in that it is open to multiple *aspects* of this interpretation (e.g., 'dealers', 'rental', 'insurance', 'tuning', 'maintenance', 'parts') [9]. An effective approach to tackle query ambiguity is to diversify the search results. By doing so, the chance that different users posing the same query will find at least one relevant result to their particular information need is maximised [6].

Current approaches in the literature seek a diverse ranking by promoting search results that cover multiple aspects[1] of the query or results that cover aspects not well covered by the other results. In common, most of these approaches exploit characteristics of the search results themselves—e.g., their textual content [4] or their coverage of a taxonomy of categories [1]—as surrogates for the actual query aspects. In this position paper, we argue that such an aspect representation only loosely caters for the possible information needs that might have led different users to pose the same query. Instead, we claim that a representation that explicitly aims to encompass multiple information needs is more effective.

In the rest of this paper, Section 2 discusses the limitations of the results-driven diversification performed by existing approaches. Our view of search result diversification as a process driven by users and their multiple possible information needs is detailed in Section 3. We conclude this paper in Section 4.

[1] Unless otherwise noted, we will refer to query interpretations and aspects indistinctly.

2 Opposing Views: Users' vs. Search Results' Diversity

Most diversification approaches in the literature attempt to promote diversity from the perspective of the search results themselves. As illustrated in Figure 1, these approaches derive some representation of the aspects underlying the query from the search results as opposed to the query itself. For instance, novelty-based diversification approaches directly compare the search results to one another without explicitly representing the aspects underlying the query—e.g., based on the search results' textual dissimilarity [4], the divergence of their language models [21], or the correlation of their relevance scores with respect to the initial query [13, 20]. In contrast, coverage-based approaches seek to maximise the search results' coverage of some explicit representation of the aspects underlying the query—e.g., categories from an existing taxonomy [1], or topic models estimated from the search results themselves [5]. In both cases, there is no attempt to account for the multiple possible information needs underlying the query.

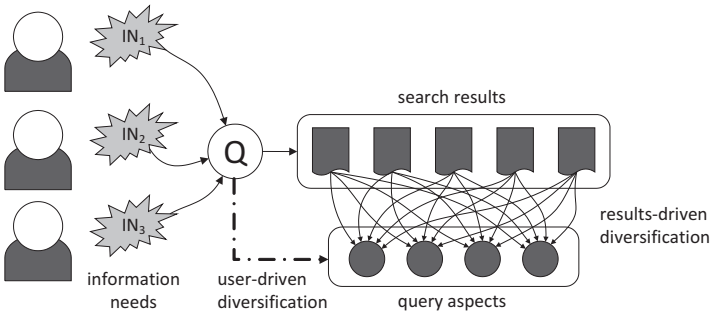

Fig. 1. User- vs. results-driven diversification.

We argue that a results-driven diversification has two key limitations. Firstly, the final ranking can be only as diverse as the aspects identified from the results retrieved for the initial query, which may be biased [12]. As a result, important query aspects (from the user population perspective) may be overlooked simply because they are not well represented among the initial results; conversely, marginally important aspects may be overemphasised. Secondly, the query aspects identified solely based on the search results are a loose surrogate for the actual information needs that may have motivated different users to issue the query in the first place. For instance, search results that cover different topics or categories—or results that are just dissimilar from each other—can feasibly meet the same information need, in which case they would be deemed redundant.

In contrast to promoting diversity from the perspective of the search results, we claim that a user-driven diversification is more effective, as corroborated by our recent empirical results [11, 14–17]. In the next section, we further detail our view of search result diversification in light of multiple possible information needs, and highlight the key areas of investigation involved in this view.

3 Diversifying for Multiple Information Needs

In this section, we detail our view of search result diversification as the problem of satisfying the multiple possible information needs underlying an ambiguous or underspecified query. Although this view is supported by our own successful experiences [11, 14–17], we focus on the principles underlying these experiences rather than on our particular solutions. In Sections 3.1 and 3.2, we describe the building blocks for a general and effective framework for diversifying the search results with respect to multiple information needs, as described in Section 3.3.

3.1 Representing Multiple Information Needs

Inspired by Spärck-Jones et al. [19], we argue that an ambiguous query should be seen as representing an ensemble of possible information needs. The problem then lies in uncovering this ensemble of information needs for a given query. For instance, in a web search scenario, the most natural approach for identifying the possible information needs underlying an ambiguous query is to analyse what previous users that issued the same query were after. Using external resources, such as a query log, one could mine queries related to the initial query, by analysing patterns of query reformulations [2, 14]. On the other hand, the search results themselves could still be leveraged as a resource for a user-driven diversification. In fact, there might be cases when the search results are the most appropriate (or maybe the only available) resource. For instance, in a blog search scenario, multiple information needs could reflect different facets (e.g., left-wing, opinionated, local) of the topic of the query [10], which in turn could be better inferred from the search results for this query. Generally speaking, the suitability of a particular resource for uncovering the possible information needs underlying a query depends on the nature of the diversification task—and hence, of the information needs themselves—at hand.

3.2 Satisfying Individual Information Needs

In order to diversify the search results with respect to the identified information needs, we first need to be able to estimate how well each search result meets every one of the identified information needs. A natural and effective approach is to deploy a ranking model to perform such estimations. As a result, the key step for diversifying the search results for a query becomes to estimate the relevance of each of these results to multiple information needs. The more refined these estimations, the more effective the attained diversification performance. For instance, we have achieved considerable success by leveraging ranking models of various calibres, from traditional document weighting models to learned models based on several features [11, 14–17]. Another important consideration regarding our view of user-driven diversification is that the identified information needs may be rather different from one another, in terms of the underlying intent of the user [3]. As such, these needs may benefit from different features. For instance, while an informational need might benefit from query expansion, a navigational need is more likely to benefit from query analysis features.

3.3 Satisfying Multiple Information Needs

Sections 3.1 and 3.2 described our view for representing the multiple possible information needs underlying an ambiguous or underspecified query, and for satisfying each of the represented information needs individually. The next step for producing a diverse ranking is to integrate these ideas into a unified diversification framework. In particular, such a framework should account for the overall coverage of each search result with respect to the identified information needs, so as to rank highly diverse documents first. Moreover, it should account for how well each information need is covered by the other search results, so as to avoid promoting redundant results [14, 17]. Additionally, another crucial feature of an effective diversification framework is the ability to infer how much emphasis should be placed on each of the identified information needs. For instance, there may be dozens of possible information needs underlying the query. If our goal is to satisfy most users in the first page of results, a bias towards the most important information needs for the user population should be enforced [14, 17]. Finally, an effective diversification framework should also cater for the ambiguity levels of different queries. In particular, not all queries are equally ambiguous. For instance, the query 'jaguar' is arguably more ambiguous than 'jaguar uk dealer locator'. To deal with the specificities of different queries, a diversification framework should be able to automatically decide not only whether, but also how much to diversify the search results on a per-query basis [15].

Altogether, the aforementioned requirements can be naturally mapped into components of a framework for diversifying for multiple information needs. In particular, our xQuAD (Explicit Query Aspect Diversification) framework [11, 14–17] fulfils all these requirements in order to provide a general and effective approach to search result diversification. As a matter of fact, building upon these ideas, xQuAD attained the top performance in the category B of the diversity task of the TREC 2009 and 2010 Web tracks [7, 8].

4 Conclusions

In this paper, we have questioned the effectiveness of search results-driven diversification approaches, and argued for a user-driven diversification instead. In particular, we have detailed our position towards diversifying the search results for multiple information needs, which naturally led to a general framework for search result diversification. Our recent results [11, 14–17] support the stated position, with the described framework attaining a state-of-the-art performance.

Our view of diversification as a user-driven process could be further extended towards satisfying multiple possible information needs across multiple search scenarios (e.g., web, image, news, blogs). In particular, this would open up research directions on several fronts, including the estimation of query ambiguity, the identification and estimation of the likelihood of different information needs, and the estimation of appropriate models for satisfying information needs in different scenarios. As a result, this extended view could form the basis for a holistic approach to search result diversification in an aggregated search scenario.

References

1. Agrawal, R., Gollapudi, S., Halverson, A., Ieong, S.: Diversifying search results. In: Proceedings of WSDM. pp. 5–14 (2009)
2. Baeza-Yates, R.A., Hurtado, C.A., Mendoza, M.: Query recommendation using query logs in search engines. In: EDBT Workshops. pp. 588–596 (2004)
3. Broder, A.: A taxonomy of Web search. SIGIR Forum 36(2), 3–10 (2002)
4. Carbonell, J., Goldstein, J.: The use of MMR, diversity-based reranking for reordering documents and producing summaries. In: Proceedings of SIGIR. pp. 335–336 (1998)
5. Carterette, B., Chandar, P.: Probabilistic models of ranking novel documents for faceted topic retrieval. In: Proceedings of CIKM. pp. 1287–1296 (2009)
6. Chen, H., Karger, D.R.: Less is more: probabilistic models for retrieving fewer relevant documents. In: Proceedings of SIGIR. pp. 429–436 (2006)
7. Clarke, C.L.A., Craswell, N., Soboroff, I.: Overview of the TREC 2009 Web track. In: Proceedings of TREC (2009)
8. Clarke, C.L.A., Craswell, N., Soboroff, I., Cormack, G.V.: Preliminary overview of the TREC 2010 Web track. In: Proceedings of TREC (2010)
9. Clarke, C.L.A., Kolla, M., Cormack, G.V., Vechtomova, O., Ashkan, A., Büttcher, S., MacKinnon, I.: Novelty and diversity in information retrieval evaluation. In: Proceedings of SIGIR. pp. 659–666 (2008)
10. Macdonald, C., Ounis, I., Soboroff, I.: Overview of the TREC 2009 Blog track. In: Proceedings of TREC (2009)
11. McCreadie, R., Macdonald, C., Ounis, I., Peng, J., Santos, R.L.T.: University of Glasgow at TREC 2009: Experiments with Terrier—Blog, Entity, Million Query, Relevance Feedback, and Web tracks. In: Proceedings of TREC (2009)
12. Mowshowitz, A., Kawaguchi, A.: Assessing bias in search engines. Inf. Process. Manage. 38(1), 141–156 (2002)
13. Rafiei, D., Bharat, K., Shukla, A.: Diversifying Web search results. In: Proceedings of WWW. pp. 781–790 (2010)
14. Santos, R.L.T., Macdonald, C., Ounis, I.: Exploiting query reformulations for Web search result diversification. In: Proceedings of WWW. pp. 881–890 (2010)
15. Santos, R.L.T., Macdonald, C., Ounis, I.: Selectively diversifying Web search results. In: Proceedings of CIKM. pp. 1179–1188 (2010)
16. Santos, R.L.T., McCreadie, R., Macdonald, C., Ounis, I.: University of Glasgow at TREC 2010: Experiments with Terrier in Blog and Web tracks. In: Proceedings of TREC (2010)
17. Santos, R.L.T., Peng, J., Macdonald, C., Ounis, I.: Explicit search result diversification through sub-queries. In: Proceedings of ECIR. pp. 87–99 (2010)
18. Song, R., Luo, Z., Nie, J.Y., Yu, Y., Hon, H.W.: Identification of ambiguous queries in Web search. Inf. Process. Manage. 45(2), 216–229 (2009)
19. Spärck-Jones, K., Robertson, S.E., Sanderson, M.: Ambiguous requests: implications for retrieval tests, systems and theories. SIGIR Forum 41(2), 8–17 (2007)
20. Wang, J., Zhu, J.: Portfolio theory of information retrieval. In: Proceedings of SIGIR. pp. 115–122 (2009)
21. Zhai, C., Cohen, W.W., Lafferty, J.: Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In: Proceedings of SIGIR. pp. 10–17 (2003)

A Search Architecture Enabling Efficient Diversification of Search Results

Gabriele Capannini, Franco Maria Nardini,
Raffaele Perego, and Fabrizio Silvestri

ISTI–CNR, Pisa, Italy
{name.surname}@isti.cnr.it

Abstract. In this paper, we deal with efficiency of the diversification of results returned by Web Search Engines (WSEs). We extend a search architecture based on additive Machine Learned Ranking (MLR) systems with a new module computing the diversity score of each retrieved document. Our proposed solution is designed to be used with other techniques, (e.g. early termination of rank computation, etc.). Furthermore, we use an efficient state-of-the-art diversification approach based on knowledge extracted from query logs, and prove that it can efficiently works in a additive machine learned ranking system, and we study its feasibility.

1 Introduction

Diversification of Web search results is a hot research topic. The majority of research efforts have been spent on studying effective diversification methods able to satisfy web users. In this paper we take a different turn and we consider the problem from the efficiency perspective. As Google's co-founder Larry Page declares[1]: *"Speed is a major search priority, which is why in general we do not turn on new features if they will slow our services down"*. In [5], we define a methodology for detecting when, and how, query results need to be diversified. We rely on the well-known concept of query refinement to estimate the probability of a query to be ambiguous. In the same paper, we show how to derive the most likely refinements, and how to use them to diversify the list of results. Then, we propose an original algorithm allowing the diversification task to be accomplished effectively and efficiently.

In this paper, our focus is on plugging efficient diversification in additive MLR systems. In modern WSE query response time constraints are satisfied employing a two-phase scoring. The first phase inaccurately selects a small subset of potentially relevant documents from the entire collection (e.g. a BM25 variant). In the second phase, resulting candidate documents are scored again by a complex and accurate MLR architecture. The final rank is usually determined by additive ensembles (e.g. boosted decision trees [6]), where many scorers are executed sequentially in a chain and the results of the scorers are added to compute the final document score.

[1] http://www.google.com/corporate/tech.html

The main contribution of this paper is to propose an architectural solution that can be integrated in an additive MLR pipeline so as to efficiently perform the diversification process at query-processing time. We provide a formal analysis of the problem and we study its feasibility in the existing MLR architectures.

The paper is organized as follows: Section 2 presents our formalization of the diversification problem. Section 3 describes the search architecture aiming at enabling the efficient diversification of search results. In Section 4, we present our conclusions and we outline possible future work.

2 Diversification using Query Logs

Users query WSEs by submitting sequences of requests that are recorded in query logs. Let Q be a query log. Let q and q' be two queries submitted by the same user during the same logical session recorded in Q. We adopt the terminology proposed in [1], and we say that a query q' is a "specialization" of q if the user information need is stated more precisely in q' than in q. Let us call S_q the set of specializations of an ambiguous/faceted query q mined from the query log. Given the popularity function that computes the frequency of a query topic in Q, and a query recommendation algorithm trained with query log Q, an algorithm that exploits the query log sessions to provide users with suggestions concerning related queries, can be adapted for devising specializations S_q at query-processing time.

Now, let us give some additional assumptions and notations. $\mathcal{D}$ is the collection of documents indexed by the WSE which returns, for any given query q, an ordered list of documents $R_q \subseteq \mathcal{D}$. The rank of document $d \in \mathcal{D}$ within R_q is indicated with $rank(d, R_q)$. A distance function $\delta : \mathcal{D} \times \mathcal{D} \rightarrow [0,1]$, having non-negative and symmetric properties is defined as $\delta(d_1, d_2) = 1 - cosine(d_1, d_2)$, where $cosine(\,)$ denotes the cosine similarity function.

The utility function defined in Equation (1) denotes how good $d \in R_q$ is for satisfying a user intent that is better represented by specialization q'.

$$U(d|R_{q'}) = \sum_{d' \in R_{q'}} \frac{1 - \delta(d, d')}{rank(d', R_{q'})} \tag{1}$$

The intuition for U is that a result $d \in R_q$ is more useful for specialization q' if it is very similar to a highly ranked item contained in the results list $R_{q'}$.

Using the above definitions of distance (δ) and utility (U), we are able to define a query-log-based approach to diversification.

MaxUtility(k): *Given: query q, the set R_q of results for q, two probability distributions $P(d|q)$ and $P(q'|q)\,\forall q' \in S_q$ measuring, respectively, the likelihood of document d being observed given q, and the likelihood of having q' as a specialization of q, the utilities $U(d|R_{q'})$ of documents, a mixing parameter $\lambda \in [0, 1]$, and an integer k. Find a set of documents $S \subseteq R_q$ with $|S| = k$ that maximizes*

$$U(S|q) = \sum_{d \in S} \sum_{q' \in S_q} (1 - \lambda) P(d|q) + \lambda P(q'|q)\, U(d|R_{q'})$$

with the constraints that every specialization is covered proportionally to its probability. Formally, let $R_q \bowtie q' = \{d \in R_q | U(d|R_{q'}) > 0\}$*. We require that for each* $q' \in S_q$, $|R_q \bowtie q'| \geq \lfloor k \cdot P(q'|q) \rfloor$.

Our technique aims at selecting from R_q, the k results that maximize the overall utility of the results list.

MaxUtility(k) aims to maximize directly the overall utility. The problem can thus be simplified and solved in a very simple and efficient way [5]. In the same paper we propose *OptSelect*, an algorithm that aims to maximize $U(S|q)$ by simply computing for each $d \in R_q$ the utility of d for specializations $q' \in S_q$ and, then, to *select* the top-k highest ranked documents. Obviously, we have to carefully select results to be included in the final list in order to avoid choosing results that are relevant only for a single specialization. To select results, we use a set of $|S_q|$ min-heaps each of those keeps the $\lfloor k \cdot P(q'|q) \rfloor + 1$ most useful documents for that specialization. *OptSelect* returns the set S maximizing the objective function of MaxUtiliy(k) in linear time. Moreover, the running time of *OptSelect* is linear in the size of document considered. Indeed, all the heap operations are carried out on data structures having a constant size $\leq k$.

3 Proposed Architecture

In the previous section we have sketched the *OptSelect* algorithm as an efficient solution for the diversification task. Here, we show how such a solution needs to be adapted in order to be plugged in a modern MLR system having a pipelined architecture. Let us assume that, given a query q, MLR algorithms are used to rank a set $D = \{d_1, \ldots, d_m\}$ of documents according to their relevance to q. Then the k documents with the highest score are returned. To this end, additive ensembles are used to compute the final score $s(d_i)$ of a document d_i as a sum over many, simple scorers, i.e. $s(d_i) = \sum_{j=1}^{n} f_j(d_i)$, where f_j is a scorer that belongs to a set of n scorers executed in a sequence. Moreover, the set of scorers is expected to be sorted by decreasing order of importance. This because, as argued in [4], if we can estimate the likelihood that d_i will end up within the top-k documents, we can early exit the $s(d_i)$ computation at any position $t < n$, computing a partial final score using only the first t scorers. For these reasons, it is important to define a solution that is fully integrable with the existing systems. Another important aspect to consider is the cost of each f_j that must be sustainable w.r.t. the others scorers. In particular, we assume that the cost c of computing $f_j(d_i)$ is constant and the total cost of scoring all documents in D is, thus $\mathcal{C}(D) = c \cdot m \cdot n$. For tasks with tight constraints on execution time, this cost is not sustainable if both m and n are high (e.g. $m > 10^5$ and $n > 10^3$ as shown in [4]).

To achieve the previously specified goal, WSE needs some additional modules in order to enable the diversification stage, see Figure 1. Briefly, our idea is the following. Given a query q, perform simultaneously both the selection of the documents potentially relevant for q from the entire collection (module `BM25`) and the retrieve of the specializations for q (module `SS`). Assuming that `SS`

performs faster than both DR and BM25, the module $f_{\texttt{DVR}}$ can be placed in any position of the MLR pipeline, i.e. $f_1 \rightarrow_{\ldots} f_n$. The target of $f_{\texttt{DVR}}$ is, then, to exploit Equation (1) for properly increasing the rank of the incoming documents as the other pipelined scorers do. Note that in this case, that is different from *OptSelect* running context, the final extraction of top-k documents is left to the MLR pipeline that already performs this operation automatically. In the following, we give more detail on our approach.

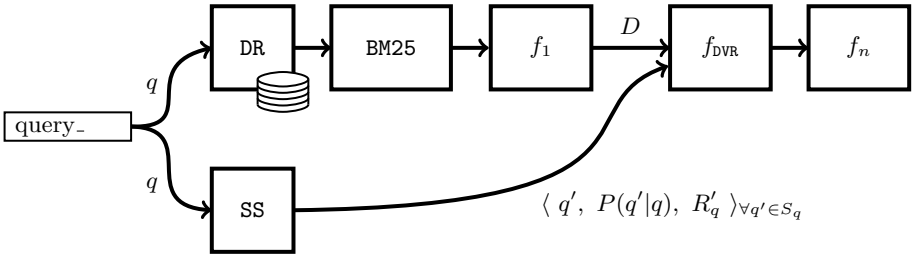

Fig. 1. A sketch of the WSE architecture enabling diversification.

For any given query q submitted to the engine, we dispatch q to the document retriever DR that processes the query on the inverted index, and to the module SS that generates the specializations S_q for q. SS processes q on a specific inverted index structure derived from query logs: the same proposed in [2]. SS returns a set of specializations S_q, a distribution of probability $P\left(q'|q\right) \forall q' \in S_q$, and a set $R_{q'} \forall q' \in S_q$ of sketches representing the most relevant documents for each specialization. Concerning the feasibility in space of the inverted index in SS, note that each set $R_{q'}$ related to a specialization $q' \in S_q$ is very small compared to the set of whole documents R_q to re-rank, i.e. $|R_{q'}| \ll |R_q|$. Furthermore, using *shingles* [3], only a sketch of a few hundred bytes, and not the whole documents, can be used to represent a document without significant loss in the precision of our method[2]. Resuming, let ℓ be the average size in bytes of a shingle representing a document and let h be the average space needed to store the set S_q of specializations for a query q by using the related inverted index, we need at most $(\ N \cdot |S_{\widehat{q}}| \cdot |R_{\widehat{q}'}| \cdot \ell\ +\ N \cdot h\ )$ bytes for storing N ambiguous query along with the data needed to assess the similarity among results lists. For example, considering a number of ambiguous queries of order of hundreds of thousands, tens of specializations per query, and hundreds of documents per specialization, we need an inverted index for SS of about 10 GB.

Now, let us focus on $f_{\texttt{DVR}}$. As the other modules belonging to the MLR pipeline, also $f_{\texttt{DVR}}$ receives a set of documents D as a stream from its preceding module, scores the elements, then release the updated set. However, contrarily to other diversifying methods analyzed in [5], $f_{\texttt{DVR}}$ is able to compute on the fly the

[2] note that shingles are already maintained by the WSE for near duplicate document detection.

diversity-score for each document d. In fact, exploiting the knowledge retrieved from the query log, our approach does not require to know in advance the composition of D to diversify the query result because SS provides the proper mix of different means related to q. In particular, we firstly compute for each $d \in D$ the related shingle. As stated in [3], the related sketch can be efficiently computed (in time linear in the size of the document d) and, given two sketches, the similarity $1-\delta(d,d')$ of the corresponding documents (i.e. $d \in D$ and each document d' returned by SS, i.e. $d' \in R_{q'}\ \forall q' \in S_q$) can be computed in time linear in the size of the sketches. The resulting similarity thus concurs to compute $U(d|R_{q'})$, i.e. the variation of final score of the document d.

4 Conclusions

We studied the problem of plugging the WSE results diversification step in a additive MLR system. In order to do that, we exploited a diversification technique suitable for working in this ranking system and thus able to compute at query-processing time the diversity score of each document. By exploiting this approach, the selection of the relevant results to return to the user can be done by simply selecting the top-k documents with the highest score. Our proposed solution is designed to be used with other techniques, (e.g. early termination of rank computation, etc.). We sketched the resulting MLR search architecture, and we outline a first preliminary study on the feasibility in space of the technique.

References

1. Boldi, P., Bonchi, F., Castillo, C., Vigna, S.: From 'dango' to 'cakes': Query reformulation models and patterns. In: Proc. WI'09. IEEE CS Press (2009)
2. Broccolo, D., Marcon, L., Nardini, F.M., Perego, R., Silvestri, F.: Generating suggestions for queries in the long tail with an inverted index. submitted for review
3. Broder, A.Z., Glassman, S.C., Manasse, M.S., Zweig, G.: Syntactic clustering of the web. Comput. Netw. ISDN Syst. 29, 1157–1166 (September 1997)
4. Cambazoglu, B., Zaragoza, H., Chapelle, O., Chen, J., Liao, C., Zheng, Z., Degenhardt, J.: Early exit optimizations for additive machine learned ranking systems. In: Proceedings of the third ACM international conference on Web search and data mining. pp. 411–420. ACM (2010)
5. Capannini, G., Nardini, F.M., Perego, R., Silvestri, F.: Efficient diversification of web search results. Proceedings of the VLDB, Vol. 4, No. 7 (April 2011)
6. Zheng, Z., Zha, H., Zhang, T., Chapelle, O., Chen, K., Sun, G.: A general boosting method and its application to learning ranking functions for web search. Advances in Neural Information Processing Systems 19 (2007)

Diversification of search results as a fuzzy satisfiability problem

Steven Schockaert and Martine De Cock

Dept. of Applied Mathematics and Computer Science, Ghent University, Belgium
{steven.schockaert,martine.decock}@ugent.be

Abstract. In various information retrieval settings, it is of interest to the user to receive search results that are not only relevant, but also diverse. As the precise goals and the underlying understanding of diversity differs considerably from application to application, there is a need for a language in which diversification strategies can be encoded and modified in an intuitive way, yet which is sufficiently rich to capture all of the subtleties that may arise. In this paper, we propose a language based on ideas from fuzzy logics, and illustrate its flexibility and ease-of-use by providing several examples of diversification strategies. Through a number of use-case scenarios, we also point out how some of the weaknesses of existing methods can be avoided in our framework.

1 Introduction

The results returned by a search engine are useful to a user only insofar that they are relevant to her information need, are up-to-date, and arise from an authoritative source, among others. In addition to considering these qualities of individual documents, however, it is also important to ensure that the list of results is sufficiently diverse [1–3], for at least two reasons. First, when the query issued by a user is ambiguous, it makes sense to display at least one search result related to each possible understanding of the query. For example, when sending the query *apple* to google, all results on the first page are related to apple computers[1], which is cumbersome for users who are looking for information about fruit. Second, the list of search results should preferably not contain redundant results: when a given document is relevant, but very similar to a higher ranked document, it may be of little added value to the user. For example, in an image retrieval setting where the query is *Paris*, it does not make much sense to present the user with 20 photos of the Eiffel tower.

One way to deal with the problem of diversifying search results is to treat it as a combinatorial optimization problem. Starting from a given set of documents D, relevance estimates for these documents, and pairwise (dis)similarities, the primary task is then to find an optimal subset $S \subseteq D$ of k documents, which are as relevant as possible, while being as different from each other as possible. In

[1] Verified on February 14, 2011.

[3], S is selected as the set of k documents which maximizes some optimization criterion f. A first possibility is [3]:

$$f(S) = (k-1) \cdot \sum_{d \in S} \mathit{rel}(d) + 2\lambda \sum_{d_i, d_j \in S} \mathit{dist}(d_i, d_j) \tag{1}$$

where $|S| = k$, $\lambda > 0$, $\mathit{rel}(d)$ is the relevance estimate of document d and dist is a measure of dissimilarity. Two other possibilities are [3]:

$$f(S) = \min_{d \in S} \mathit{rel}(d) + \lambda \min_{d_i, d_j \in S} \mathit{dist}(d_i, d_j) \tag{2}$$

$$f(S) = \sum_{d \in S} \Big(\mathit{rel}(d) + \frac{\lambda}{|D| - 1} \sum_{d' \in D} \mathit{dist}(d, d') \Big) \tag{3}$$

Note that the latter sum ranges over the set of all documents, rather than those in S alone. As shown in [3], each of the alternatives (1)–(3) satisfies different properties, corresponding to different aspects of diversity. In general, it is not straightforward to translate a given intuition about diversification to an actual optimization criterion, which always yields the most intuitive result. Moreover, in different settings, different factors may have to be taken into account. When retrieving product reviews, for instance, it may be of interest to the user to know whether most reviews are positive or negative. In this sense, when 90% of the reviews are negative, it would not be a good idea to display 5 positive reviews and 5 negative reviews, even though this choice may maximize diversity and all 10 reviews might be relevant.

As different settings thus require different diversification mechanisms, there is a need for a flexible framework in which the intuitions underlying a particular setting can easily be translated to declarative specifications. In this paper, we propose an approach based on fuzzy logics. As in classical logic, formulas in fuzzy logics are built from constants, variables and logical connectives. In contrast to classical logic, however, formulas may take an arbitrary truth value from the unit interval $[0, 1]$ instead of only 0 (false) and 1 (true). On one hand, the resemblance with classical logic allows us to encode relations between graded properties, such as relevance or similarity, in a logical fashion. This leads to intuitive, declarative specifications, whose qualitative behavior can readily be seen from the syntactic structure of the formulas. On the other hand, Boolean connectives can be generalized to fuzzy logic connectives in different ways, which provides a form of parameterization in fuzzy logic models. The exact behavior of the resulting systems is therefore a combination of the syntactic structure of the underlying formulas, and an appropriate choice for each of the logical connectives.

The structure of this paper is as follows. In the next section, we introduce a language based on fuzzy logics, which we will use to encode diversification mechanisms. Subsequently, Section 3 presents a number of use case scenarios to illustrate some issues with existing methods such as (1)–(3). Section 4 then proposes various encodings of diversification strategies, as constraints on fuzzy logic formulas. Finally, a discussion with some concluding remarks is provided.

2 Constraints on fuzzy logic formulas

Let $D = \{d_1, ..., d_n\}$ be the set of document under consideration, with corresponding relevance scores $rel(d_i)$ and pairwise similarities $sim(d_i, d_j)$. Both relevance scores and similarities are assumed to be in $[0, 1]$. When encoding diversification mechanisms, we will also consider a number of additional predicates. If p is an m-ary predicate, then an expression of the form $p(d_{i_1}, ..., d_{i_m})$ is called a term. From these terms, formulas are constructed as follows:

- Constants in [0,1], such as $rel(d_i)$ and $sim(d_i, d_j)$, are formulas.
- Each term is a formula.
- If α and β are formulas, then $\alpha \otimes \beta$, $\alpha \oplus \beta$, $\alpha \rightarrow \beta$, $\alpha \wedge \beta$, $\alpha \vee \beta$ and $\neg\alpha$ are formulas.
- If α is a formula and $\lambda \in [0, 1]$, then $\lambda \cdot \alpha$ is a formula.
- If each of $\alpha_1, ..., \alpha_m$ are formulas, then also $avg\{\alpha_1, ..., \alpha_m\}$, $\forall\{\alpha_1, ..., \alpha_m\}$ and $\exists\{\alpha_1, ..., \alpha_m\}$ are formulas.

For the ease of presentation, we also write e.g. $\forall d \in D \,.\, p(d)$ for $\forall\{p(d) \,|\, d \in D\}$, or even expressions such as $avg_{d \in D}(\forall d' \in D \,.\, p(d, d'))$. There are three important differences with classical logic. First, arbitrary values from $[0, 1]$ may appear in formulas as constants. Second, there are a number of connectives that have no counterpart in classical logic, namely the scaling operator $\cdot$ and the averaging operator *avg*, which are tied to the numerical interpretation of truth degrees. Finally, there are two types of conjunction ($\wedge$ and $\otimes$) and two types of disjunction ($\vee$ and $\oplus$), which are defined as follows:

$$\alpha \wedge \beta = \min(\alpha, \beta) \qquad \alpha \vee \beta = \max(\alpha, \beta)$$
$$\alpha \otimes \beta = \max(\alpha + \beta - 1, 0) \qquad \alpha \oplus \beta = \min(\alpha + \beta, 1)$$

Note that $\wedge$ and $\otimes$ indeed correspond to logical conjunction when their arguments are restricted to the classical truth values 0 and 1, and that $\vee$ and $\oplus$ correspond to logical disjunction. The operators $\otimes$ and $\oplus$ are the connectives from Łukasiewicz logic, and provide a truth degree which is a bounded linear combination of their arguments. The operators $\rightarrow$ and $\neg$ are the implication and negation from Łukasiewicz logic, defined as

$$\alpha \rightarrow \beta = \min(1, 1 - \alpha + \beta) \qquad \neg\alpha = 1 - \alpha$$

The scaling operator $\cdot$ is simply interpreted as multiplication. Finally, *avg*, $\forall$ and $\exists$ are defined as

$$avg\{\alpha_1, ..., \alpha_m\} = \frac{\alpha_1 + ... + \alpha_m}{m}$$
$$\forall\{\alpha_1, ..., \alpha_m\} = \min(\alpha_1, ..., \alpha_m)$$
$$\exists\{\alpha_1, ..., \alpha_m\} = \max(\alpha_1, ..., \alpha_m)$$

In the following, we will consider sets E of equalities of the form $\alpha = \beta$, with α and β formulas. Such a set of equalities will be called a theory. These equalities

are seen as constraints on the possible truth values of terms, which are treated as variables. An assignment ω from terms to $[0, 1]$ is called a model of a set of equalities E if substituting every term t by its value $\omega(t)$ causes all equalities in E to be satisfied. We will consider sets of equalities E such that every model of E corresponds to a solution of the diversification problem, i.e. a reasonable choice of k documents among those in D.

By construction, all formulas can be written as the combination of a number of linear expressions using the minimum and maximum operators. Seeing terms as variables, it is therefore possible to translate a set of equalities E to a mixed integer program P, such that there is a one-on-one correspondence between the models of E and the solutions of P, using a straightforward extension of the procedure proposed in [4]. This means that models of E can be found using fast mixed integer programming solvers such as CBC[2]. Under some conditions, models can also be found using finite constraint satisfaction methods [6]. Alternatively, approximate models can be found using heuristic search techniques.

3 Motivating examples

Before illustrating how equalities of fuzzy logic formulas may be used to specify diversification mechanisms, we point out some weaknesses of existing methods using a number of scenarios:

> **Scenario A** Suppose that the set D contains two duplicates (or near-duplicates) d_1 and d_2 which are highly relevant. Ideally, only one of d_1 and d_2 should appear in the set S, no matter how relevant these documents are.

If the set S is selected based on (1) or (3), both of d_1 and d_2 may appear when these documents are sufficiently relevant and/or sufficiently different from the documents in $S \setminus \{d_1, d_2\}$ (when using (1)) or $D \setminus \{d_1, d_2\}$ (when using (3)).

Next, we consider the scenario where a query term is ambiguous, and all documents that correspond to the same understanding of the query are very similar:

> **Scenario B** Suppose that D can be partitioned in $D_1 \cup ... \cup D_m$ such that documents from the same partition are highly similar, and documents from different partitions are highly dissimilar.

Let us first assume that $m < k$. When using (1), S will then more or less be balanced, in the sense that approximately the same number of documents are chosen from each partition block D_i. However, as at least two highly similar documents will be contained in S, criterion (2) trivializes, causing many different sets S to be considered as optimal, not all of which may also be intuitively satisfactory. Finally, criterion (3) will lead to the unintuitive behavior of choosing only documents from the partition blocks with the fewest documents.

[2] `http://www.coin-or.org/projects/Cbc.xml`

Now assume that $m \geq k$. Then (1)–(2) will select one document from k different partition blocks, mainly chosen based on their relevance, while (3) would still lead to choose documents from the smallest partition blocks.

> **Scenario C** Suppose that D contains one document d_1 which is not among the k most relevant documents, but which is highly dissimilar from all other documents in D. Assume furthermore that the documents in $D \setminus \{d_1\}$ are all somewhat similar to each other.

Using (1) and (3), d_1 would typically be included in S, with the remaining documents to a large extent being chosen based on their relevance. Using (2), however, depending on the value of λ, either d_1 would not be included in S or relevance would not be taken much into account for selecting the other $k-1$ documents.

4 Encoding diversification strategies

As the previous section illustrates, it is difficult to specify global optimization criteria that always lead to those results that are intuitively most desirable. In this section, we present an alternative, in which equalities between fuzzy logic formulas encode in a declarative fashion whether a given choice of S is optimal. In particular, we introduce a predicate *imp*, such that for each document d, $imp(d)$ represents the degree to which it is important to include d in S. By construction, the set S then contains the k most important documents w.r.t. this predicate:

$$in(d) = (in_1(d) \vee ... \vee in_k(d)) \tag{4}$$

where we use $in(d)$ to denote that d is included in S and $in_i(d)$ to denote that d is the i^{th} ranked document. The terms $in_i(d)$ and $in(d)$ are assumed to be Boolean, and the right-hand side of (4) should accordingly be regarded as a Boolean expression. The following formulas encode that $in_i(d)$ should be the i^{th} most important document:

$$in_1(d) = (\forall d' \neq d \,.\, imp(d) > imp(d') \vee (imp(d) = imp(d') \wedge \neg in_1(d'))) \tag{5}$$

$$in_2(d) = (\forall d' \neq d \,.\, imp(d) > imp(d') \vee in_1(d') \vee (imp(d) = imp(d') \wedge \neg in_2(d'))) \tag{6}$$

...

$$in_k(d) = (\forall d' \neq d \,.\, imp(d) > imp(d') \vee in_1(d') \vee ... \vee in_{k-1}(d') \vee (imp(d) = imp(d') \wedge \neg in_k(d'))) \tag{7}$$

Intuitively, d should be the i^{th} ranked document if all documents which are more important are ranked higher, i.e. for every other document d' we should either have one of $in_1(d'), ..., in_{i-1}(d')$ (in which case d' is indeed ranked higher), or $imp(d) \geq imp(d')$ (in which case d is at least as important as d'). Due to the last disjunct in (5)–(7), ties are broken arbitrarily.

To complete the specification of a diversification strategy, we introduce a number of equalities to define the predicate *imp*, which together with the equalities (4)–(7) form a theory E, whose models define the optimal choices for S. As a first strategy, we may define *imp* as follows:

$$\mathit{redundant}(d) = (\exists d' \neq d \,.\, \mathit{in}(d') \wedge \mathit{sim}(d, d')) \tag{8}$$

$$\mathit{imp}(d) = \mathit{rel}(d) \otimes \neg \mathit{redundant}(d) \tag{9}$$

Note that (9) clearly reveals the intuition of the underlying diversification mechanism: it is important to include d in the set S if (i) d is relevant and (ii) no other document in S is similar to it. To conjunctively combine both aspects, the Łukasiewicz conjunction is used, which, together with the use of negation boils down to a bounded difference, i.e. $\mathit{imp}(d) = \max(0, \mathit{rel}(d) - \mathit{redundant}(d))$. The linear combination of relevance scores with redundancy scores presupposes some form of commensurability. In practice, this means that the relevance scores and similarity scores we have at our disposal may have to be manipulated somehow. Such a manipulation would moreover allow us to tweak the trade-off between relevance and similarity. Also note that (9) specifies a cyclic definition: the value of the predicate *imp* depends on the predicate *in*, which in turn depends on *imp*. The models of E thus correspond to some form of equilibria or fixpoints of these equations, an observation which can be made more explicit via the theory of fuzzy answer set programming [5]. The underlying intuition is also reminiscent of Nash equilibria, in the sense that S is defined as a set (cfr. a global strategy) which cannot be improved by replacing a single document (cfr. in which no player can improve his utility without cooperation).

Let us now reconsider the three scenarios from Section 3. In Scenario A, (9) ensures that d_1 and d_2 cannot both be included in S, as then both $\mathit{imp}(d_1)$ and $\mathit{imp}(d_2)$ would be (close to) 0. In Scenario B, assuming $m < k$, we find that at least one document from each partition block will be included in S, although the remaining documents may be chosen somewhat arbitrarily. Finally, in Scenario C, we find that d_1 would typically be included in S. Hence, in all three scenarios, more or less desirable results are found. The main problem seems to be that when selecting two highly similar documents is unavoidable, as in Scenario B, some of the remaining documents may not be selected in an optimal way. This is due to the fact that the value of $\mathit{redundant}(d)$ depends on the occurrence of a single document d' in S. In this respect, using (9) resembles the optimization criterion (2). As an alternative to (8)–(9), we may consider

$$\mathit{disparate}(d) = \mathit{avg}\{\neg \mathit{sim}(d, d') \mid d \neq d', \mathit{in}(d')\} \tag{10}$$

$$\mathit{imp}(d) = \mathit{rel}(d) \otimes \mathit{disparate}(d) \tag{11}$$

which encodes the intuition that a document d is important if, on average, the other documents in S are dissimilar to it. Using (10)–(11) in Scenario B, approximately the same number of documents will be selected from each partition block, similar as when using (1) or (3). However, in contrast to (8)–(9), using (10)–(11) does not always lead to the desired result in Scenario A. One way to

ensure optimal behavior both in Scenarios A and B would be to combine the intuitions of (9) and (11) as follows ($\lambda \in [0,1]$):

$$imp(d) = rel(d) \otimes \big(\lambda \cdot (\neg redundant(d)) \oplus (1-\lambda) \cdot disparate(d)\big) \qquad (12)$$

where we assume that $\cdot$ takes priority over $\oplus$. If λ is sufficiently high, using (12) will avoid that both d_1 and d_2 are included in S in Scenario A. Moreover, in Scenario B, typically $redundant(d)$ will be close to 1 for all documents, in which case (12) behaves qualitatively similar to (11).

As already mentioned in the introduction, when ranking reviews or opinions, it is important that the set S accurately reflects whether most reviews are positive or negative, and even which type of complaints most people have (e.g. about a given product). This means that it may be beneficial to include several documents in S which express the same opinion, and are in this sense similar. To some extent, this requirement is at odds with the idea of diversifying search results, or at least, it can be seen as a tempering factor. This latter intuition of adding a tempering factor can be translated as follows:

$$prevalent(d) = avg_{d' \in D}\, sim(d,d') \qquad (13)$$

$$imp(d) = rel(d) \otimes \big(\lambda \cdot (\neg redundant(d)) \oplus (1-\lambda) \cdot prevalent(d)\big) \qquad (14)$$

which translates the intuition that d should be included if it is relevant, and it is either different from the other documents in S or it conveys a prevalent opinion. For large values of λ, (13)–(14) behave similarly as (9), while for small values of λ, diversity will only play a minimal role. To the best of our knowledge, such a trade-off has not yet been considered in existing methods.

5 Discussion

The language that was introduced in Section 2 offers the flexibility to encode a wide array of diversification strategies. In addition to the illustrations that were provided in Section 4, it is also possible to simulate existing strategies such as (1)–(3), as well as various greedy algorithms that decide which documents to add one at a time (e.g. [1]). One of the main advantages of our approach is that degrees between 0 and 1 can be treated both as numerical values (when using averaging or scaling operations), or as logical truth degrees (when using generalizations of logical connectives), which allows us to encode diversification strategies in such a way that the syntactic structure of the formulas immediately reveals the underlying intuitions.

The examples that were given correspond to basic mechanisms for diversifying search results. In practice, more structured information may be available, in which case the flexibility offered by our framework would play an even bigger role. For instance, our strategy for diversifying product reviews, i.e. (14), may be further refined when information is available about which ratings have been given by the users, or classification information about the type of complaints that are conveyed. Similarly, we may think of diversification mechanisms that

take user profiles into account, ensuring that reviews are displayed from a diverse set of users (e.g. regarding age or geographic location).

Given the observation that different applications involve different subtleties, we advocate a declarative approach, in the sense that the specification of a particular strategy should be decoupled from its implementation. One possibility for implementing the strategies encoded as constraints on fuzzy logic formulas is to translate these constraints to mixed integer programs, for which various highly efficient solvers exist. This approach has the advantage that an additional global (linear) optimization criterion can be specified to make an informed decision when there are multiple solutions. Another implementation method would be to use more heuristic techniques, e.g. taking advantage of the cyclic nature of the examples in Section 4. One idea would be to guess an arbitrary set S, i.e. a particular solution to (4)–(7), and then incrementally improve this guess by repeatedly evaluating the values of $imp(d)$, and adapting the set S accordingly.

Acknowledgments

Steven Schockaert was funded as a postdoctoral fellow of the Research Foundation – Flanders.

References

1. J. Carbonell and J. Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Cconference on Research and Development in Information Retrieval*, pages 335–336, 1998.
2. 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*, pages 659–666, 2008.
3. S. Gollapudi and A. Sharma. An axiomatic approach for result diversification. In *Proceedings of the 18th International Conference on World Wide Web*, pages 381–390, 2009.
4. R. Hähnle. Many-valued logic and mixed integer programming. *Annals of Mathematics and Artificial Intelligence*, 12:231–264, 1994.
5. J. Janssen, S. Schockaert, D. Vermeir, and M. De Cock. General fuzzy answer set programs. In *Proceedings of the 8th International Workshop on Fuzzy Logic and Applications (WILF)*, pages 352–359, 2009.
6. S. Schockaert, J. Janssen, D. Vermeir, and M. De Cock. Finite satisfiability in infinite-valued Łukasiewicz logic. In *Proceedings of the Third International Conference on Scalable Uncertainty Management*, volume 5785 of *Lecture Notes in Computer Science*, pages 240–254. 2009.

A Comparative Study of Search Result Diversification Methods

Wei Zheng and Hui Fang

University of Delaware, Newark DE 19716, USA
zwei@udel.edu, hfang@ece.udel.edu

Abstract. Top-ranked documents returned by traditional retrieval functions may cover the same piece of relevant information and cannot satisfy different user needs. Search result diversification solves this problem by diversifying results to cover more information needs, i.e., query subtopics, in top-ranked documents. Many diversification methods have been proposed and studied, and most of them re-rank original retrieved documents according to both relevance and diversity functions in a probabilistic framework. Although official TREC results make it possible to compare the effectiveness of different diversification systems, it remains unclear whether the better performance of a system comes from better diversification methods or component estimation methods. In this paper, we conduct a systematic study on comparing three representative diversification methods which can be implemented using probabilistic methods. We not only analytically compare the methods but also conduct empirical studies and evaluate the effectiveness of these methods in a controlled manner.

1 Introduction

Traditional retrieval functions ignore the relations among returned documents. As a result, top ranked documents may contain relevant yet redundant information. In order to maximize the satisfaction of different search users, it is necessary to diversify search results.

Many diversification methods have been proposed. For example, Carbonell and Goldstein [2] proposed the maximal marginal relevance (**MMR**) ranking strategy to balance the relevance and the redundancy among the returned documents. Yin et. al. [7] derived a diversification method using the language modeling approach, i.e., **WUME**. Santos et. al. [6] proposed a probabilistic framework, i.e., **xQuAD**, that estimates the diversity based on the relevance of documents to query subtopics and the importance of query subtopics. The first method is a classical method and has been widely cited, but none of the top-ranked diversity systems from TREC used this method. The last two methods were implemented in the systems participating in TREC 2009 Web track. Although the evaluation results of these two methods are quite different according to the official TREC results [3], it is unclear whether the performance differences are

caused by the underlying diversification methods or the ways of estimating the component functions.

In this paper, we conduct a systematic study to compare the above three representative diversification methods both analytically and empirically. Specifically, we first analyze the methods and summarize their commonalities and differences. These methods mainly differ in *diversity modeling*, i.e., whether the diversity is implicitly modeled through document similarities or explicitly modeled through the coverage of query subtopics, and *document dependency*, i.e., whether the diversity score of a document is related to other documents or not. To make a more meaningful empirical comparison, we modify the three methods under the same framework and use the variants of the three methods in this paper. All of the variants of the methods re-rank the original retrieved documents based on a linear combination of relevance and diversity scores. The variants also use the same methods to estimate components in their functions. This would allow us to focus on the differences in the diversification methods. Moreover, following the idea of diagnostic evaluation [5], we conduct four sets of experiments using simulated collections. Our goal is to not only compare different diversification methods but also study how the performance of a diversification method can be impacted by different factors, i.e., the quality of relevant functions, the tradeoff between relevance and diversity, and the number of query subtopics.

Experiment results show that *diversity modeling* has a large impact on the effectiveness of a diversification method. Explicitly modeling the diversity with query subtopics is more effective than implicitly modeling the diversity through document similarities. As an example, MMR performs worse than the other two methods consistently. Moreover, *document dependency* has a smaller impact on the diversity performance. Although computing the diversity score of a document based on other documents is intuitively desirable, the empirical performance gain is small. Finally, we can also make the following interesting observations.

- The effectiveness of a diversification method is closely related to the effectiveness of its relevance function. In particular, the performance improvement of the diversification method decreases as the performance of the relevance function increases.
- The number of query subtopics affects the diversity performance of the methods that explicitly model the diversity based on subtopics. However, they may still achieve reasonably good performance when the quality of subtopics is good and the number of missed subtopics is small.

2 Analytical Comparisons of Diversification Methods

Most of existing diversification methods first retrieve a set of documents based on only their relevance scores, and then re-rank the documents so that the top-ranked documents are diversified to cover more query subtopics [2–4, 6, 7]. Since the problem of finding an optimum set of diversified documents is NP-hard [1], a greedy algorithm is often used to iteratively select the diversified document.

In this paper, we focus on three representative diversification methods discussed in the previous section.

- MMR [2]: It maximizes the margin relevance of the documents and iteratively select the document that is not only relevant to the query but also dissimilar to the previously selected documents.
- $WUME$ [7]: It maximizes the probability that the document meets the user needs. Its diversification function iteratively selects the document that covers both the query and the important subtopics of the query.
- $xQuAD$ [6]: It uses the probability model to maximize the combination of the likelihood of a document is observed given the query and the likelihood of the document while not the previously selected documents is observed given the query. It iteratively selects the document that is not only relevant to the query but also covers the subtopics that have not be well covered by previously selected documents.

All these three methods iteratively select the document that is not only relevant to the query but also diversified to cover more query subtopics, explicitly or implicitly. Therefore, all of them fit into a general framework that iteratively selects the document with the highest relevance and diversity scores [2, 6, 1]:

$$d^* = \arg \max_{d \in D \setminus D'} (\lambda \times (Rel(d, q) + (1 - \lambda) \times Div(d, q, D'))) \tag{1}$$

where D is a set of documents that need to be re-ranked, D' is the set of previously selected documents, λ is a parameter that balances the relevance score of the document i.e., $Rel(d, q)$, and the diversity score $Div(d, q, D')$.

We then implement the variants of these methods under the framework and they are referred to as MMR^*, $WUME^*$ and $xQuAD^*$:

1. Maximal marginal relevance (MMR) variant method [2]:

$$Div_{MMR^*}(d, q, D') = - \max_{d' \in D'} p(d|d') \tag{2}$$

2. $WUME$ variant method [7]:

$$Div_{WUME^*}(d, q, D') = \sum_{s \in S(q)} p(s|q)p(d|s) \tag{3}$$

3. Explicit query aspect diversification ($xQuAD$) variant method [6]:

$$Div_{xQuAD^*}(d, q, D') = \sum_{s \in S(q)} p(s|q)p(d|s) \prod_{d' \in D'} (1 - p(d'|s)) \tag{4}$$

$S(q)$ is the subtopic set of query q. $p(d|d')$ measures the similarity between current document and selected document, $p(d|s)$ measures the similarity between the document and the subtopic, $p(s|q)$ measures the importance of the subtopic in the query and $\prod_{d' \in D'} (1 - p(d'|s))$ is the subtopic importance penalization

component that penalizes the importance of the subtopic that has been covered in previously selected documents.

In $WUME^*$, we split the probability of the document given both the query and subtopics existing in $WUME$, in order to make it comparable with the other methods. We consider the probability of the document given the query in $Rel(d, q)$ and the probability of the document given the subtopics in $Div_{WUME^*}(d, q, D')$. Other main differences between the original diversification methods and these variants are how to estimate the component functions in the methods and how to find query subtopics. Since we focus on comparing different diversity functions, we use query subtopics given in the judgment file as $S(q)$ and use the same method to estimate the components.

Comparing these three diversity functions, we can see that they mainly differ in two aspects. The first aspect is the *diversity modeling*. MMR^* implicitly models the diversity through document similarities and ignores the information about query subtopics. On the contrary, the other two methods explicitly model the diversity through the coverage of query subtopics. The second aspect is the *document dependency*. $WUME^*$ assumes that the diversity score of a document is independent of other documents while the other two methods assume that the diversity score depends on the previously selected documents.

Intuitively, it is more reasonable to explicit use subtopics to model diversity and assume that the documents are dependent of each other. Therefore, $xQuAD^*$ should perform best and the performance of $WUME^*$ would be the second best. However, it is unclear whether both explicit subtopics and document dependence have big effects on the diversification results, and whether the difference between the diversification methods is significantly. We will compare their performances in the following section.

3 Experiments

In our experiment, we use the TREC09 and TREC10 collections [3], each of which has 50 queries, and the Category B of ClueWeb09 collection that contains 428 million documents. We use α-nDCG@100, together with α-nDCG@20 used in TREC, as the measures to evaluate the diversification results. The reason is that we want to observe the performance of a longer document ranking list. We use the Dirichlet retrieval function [9] to retrieve the original results and compute the probabilities in Equation (1)-(4). We use the real subtopics given in the judgment file for diversification in explicit subtopic based methods. We then design the experiments to study the following questions: (1) the optimum performances of diversification methods; (2) the impact of retrieval performance of the original ranking on diversification results; (3) the impact of parameters, i.e., tradeoff between diversity and relevance, and number of subtopics.

3.1 Comparison of diversification methods

In this section, we test whether using explicit subtopics and document dependence can significantly perform better. Table 1 shows the optimum performances

	TREC09 result		TREC10 result	
	α-nDCG@20	α-nDCG@100	α-nDCG@20	α-nDCG@100
MMR^*	0.365	0.427	0.344	0.415
$WUME^*$	0.479	0.546	0.579	0.630
$xQuAD^*$	0.482	0.550	0.588	0.636

Table 1. The performances of diversification methods when using all real subtopics for diversification

of the diversification methods both on the original TREC09 and TREC10 collections. All the parameters in each method are set to the optimum values. We can see that both $xQuAD^*$ and $WUME^*$ perform significantly better than MMR^*. It shows that using explicit subtopics in diversification is better than implicit subtopics, which is consistent with the observation in [6]. However, the performances of $xQuAD^*$ and $WUME^*$ are not significantly different. It tells that the component of subtopic importance penalization in Equation 4 of $xQuAD^*$ needs to be modified to further improve the performance. We leave this study for our future work.

3.2 Impact of original retrieval result quality

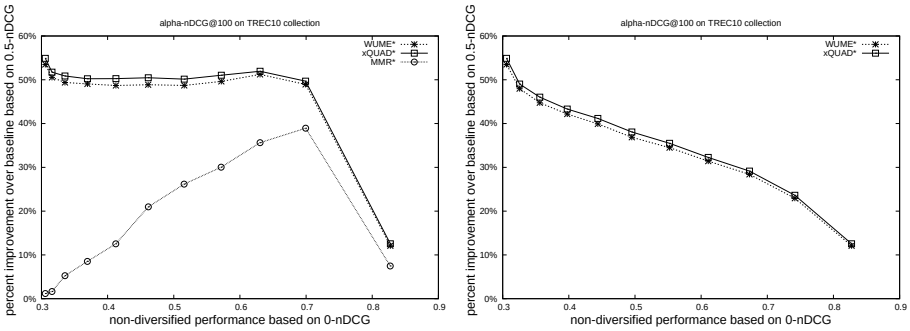

Fig. 1. The percentage improvements of diversification methods over non-diversified methods that combined all relevant documents with non-relevant documents selected from the top results (left) or random selected (right) from the original retrieval result.

We now test the impact of original retrieval result quality on diversification results. Due to the space limitation, we only show results of TREC2010 while ignore TREC2009 that has similar trend in the following experiments. We simulate the original retrieval results with different relevance qualities, evaluated by $0-nDCG$. We combine all the relevant documents in the judgment file with N non-relevant documents selected from the top documents in the original retrieval result in each query. We then re-compute the relevance scores of all these documents given the query. The simulated retrieval result only contains the relevant documents when N is 0 and is the same as the original retrieval result when N

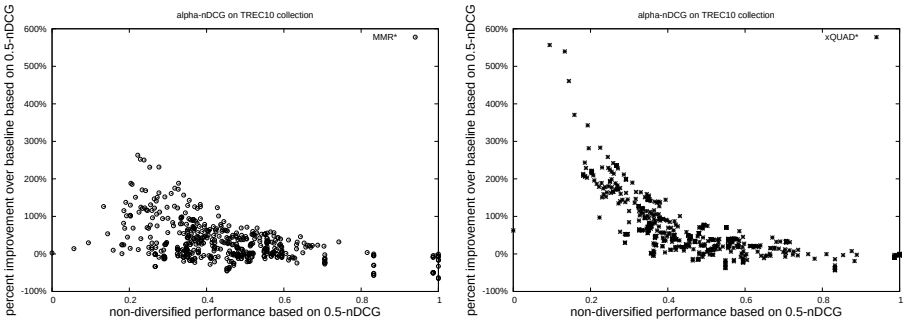

Fig. 2. The improvements of diversity methods in each query

is 100. The left plot of Figure 1 shows the performance improvements of different diversification methods in diversifying these simulated retrieval results. The values of N corresponding to points from left to right on each line are 100, 90, ..., 10 and 0, respectively.

There are three interesting observations in the plot. (1) $xQuAD^*$ and $WUME^*$ can consistently outperform MMR^*. What's more, the difference between MMR^* and the other two methods is bigger when the original retrieval result is worse. MMR^* aggressively selects the document that are most different from the previously selected document. This helps diversify the relevant documents but also selects more non-relevant documents when the original result is worse. The reason is that many non-relevant documents are less similar to relevant documents [8] and the non-relevant documents themselves are also different. (2) The performance differences between $WUME^*$ and $xQuAD^*$ are always small. We also use the other method to select the N non-relevant documents and compare these two methods on the new stimulate retrieval results. We randomly select these non-relevant documents 10 times from the original retrieval result for each value of N. We then diversify each 10 results corresponding to the same value of N and use their average relevance performance to represent the performance of that value of N. The right plot of Figure 1 shows the performances of $WUME^*$ and $xQuAD^*$. Their performances are still similar. It again shows that a new method to penalize the subtopic importance is needed to further improve the performance. (3) The worse the non-diversified method performs, the larger the improvement of diversification is. The reason is that these methods can use the subtopics to not only diversify relevant documents but also rank non-relevant documents lower when the quality of non-diversified result is poor. It is also interesting to study the improvement trend of diversification methods with different diversity performances of the original retrieval method, evaluated by $0.5 - nDCG$. Figure 2 shows the improvements of $xQuAD^*$ and MMR^* over baseline in each query with simulated retrieval results. We can also see that $xQuAD^*$ performance has larger gain when the diversity quality of the query is worse.

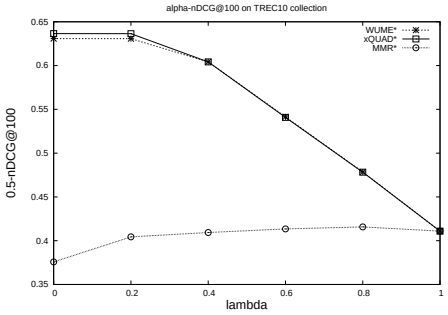

Fig. 3. Impact of λ on the diversification performance

3.3 Impact of parameters in diversification

There are two parameters in the diversification methods. One is λ that balances the relevance score and diversity score in Equation 1. The other is the number of subtopics in explicit subtopic based methods. We use the original retrieval result and the real subtopics in the judgement to tune these parameters. Figure 3 shows the impact of λ on different methods when using all real subtopics. The smaller the value of λ is, the more the methods are focusing on diversity. The optimum value of λ in methods based on explicit subtopics is 0. The subtopics used in these methods are the real subtopics in judgment file and therefore they can achieve optimum performance without considering the document relevance with the original query. However, the optimum value of λ may not be 0 if they do not use real subtopics and use other methods to extract subtopics from the collection.

In the above experiment, we use all real subtopics in diversification. However, the extracted subtopics in methods based on explicit subtopics may be incomplete. Therefore, we study the impact of the number of subtopics while using the optimum value of λ in each method. We randomly select $n\%$ of real subtopics for diversification in each query. We extract each possible combination of real subtopics for each value of n. For each value of n, we evaluate the diversification performance of using its subtopic sets and use the average performance to represent the diversification performance corresponding to that value of n.

Figure 4 shows the diversification performance using the incomplete subtopic set. The improvements of $WUME^*$ and $xQuAD^*$ decrease when the percentage of missed real subtopics decreases, but they can still outperform MMR^*. What's more, their performance decrease is not significant when the percentage of missed real subtopics is small, i.e., 20%. The right plot in Figure 4 shows the percentage of queries in different categories when comparing the diversification using $n\%$ of real subtopics and that using all real subtopics. When n is greater or equal to 80, the result of using these incomplete subtopics is very close to the result using all real subtopics, which shows that the explicit subtopic modeling methods are robust to the quality of subtopics and can still achieve reasonably good performance when their extracted subtopics do not contain all real subtopics.

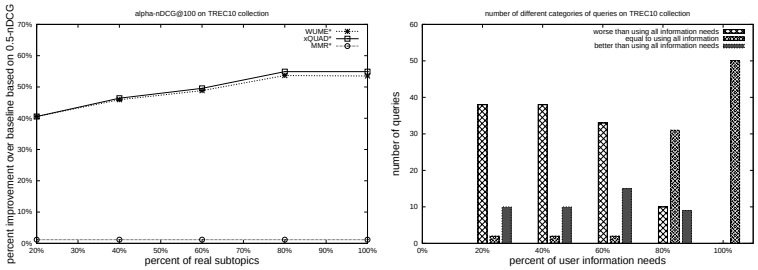

Fig. 4. Performance improvement (left) and query comparison (right) when using $n\%$ of real subtopics for diversification

4 Conclusion

In this paper, we revisited the existing diversification methods based on the language model and systematically compared their diversity functions. We compared the diversity modeling and document dependency strategies used in diversification functions. The experiment result shows that the explicit subtopic modeling and subtopic importance penalization strategies perform better but the effect of the penalization is small. It is also interesting to find that the explicit subtopic based methods are robust to the number of subtopics and can still achieve reasonable good performance when missing a small number of real subtopics.

References

1. R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In *Proceedings of WSDM'09*, 2009.
2. J. Carbonell and J. Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of SIGIR'98*, pages 335–336, 1998.
3. C. L. A. Clarke, N. Craswell, and I. Soboroff. Overview of the trec 2009 web track. In *Proceedings of TREC'09*, 2009.
4. M. Drosou and E. Pitoura. Search result diversification. In *Proceedings of SIGMOD'2010*, 2010.
5. H. Fang, T. Tao, and C. Zhai. Diagnotic evaluation of information retrieval models. *ACM Transactions of Information Systems*, To Appear.
6. R. L. T. Santos, C. Macdonald, and I. Ounis. Exploiting query reformulations for web search result diversification. In *Proceedings of WWW'10*, 2010.
7. D. Yin, Z. Xue, X. Qi, and B. D. Davison. Diversifying search results with popular subtopics. In *Proceedings of TREC'09*, 2009.
8. C. Zhai, W. Cohen, and J. Lafferty. Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In *Proceedings of SIGIR'03*, 2003.
9. C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of SIGIR'01*, 2001.

Diversity in Expert Search

Position Paper

Vassilis Plachouras

PRESANS
XTEC, École Polytechnique, 91128 Palaiseau Cedex, France
vassilis.plachouras@presans.com

Abstract. Diversity in retrieval results has been mainly studied in the context of Web search, where queries may be broad or ambiguous. Web search, however, is not the only area where diversity may be applied. Diversity may be beneficial in expert search, where users are looking for people or organizations with relevant expertise. In this position paper, we define a measure to quantify the topical ambiguity or broadness of a query, and we demonstrate it on a sample of queries from an operating expert search engine. We then discuss the aspects of diversity which are specific to expert search, and note the implications of diversification at different stages of ranking experts.

1 Introduction

Web search engines deal with a wide range of broad or ambiguous queries, commonly expressed with only few words [3, 9]. One approach to deal with such queries is to diversify the retrieved results, so that different interpretations or subtopics are promoted in the results for ambiguous or broad queries, respectively. The diversification of results, hence, requires the definition of a new objective function with a trade-off between relevance and diversity, by reranking documents according to how dissimilar they are, or how well they cover the query subtopics [5, 8]. Research on diversifying search results has been facilitated by the corresponding task in TREC 2009 Web track [2], where documents are relevant to a topic but also to subtopics.

The diversity of search results so far has been studied in the context of Web search. In this position paper, we argue that diversity is important for expert search systems, where users do not search for documents about a topic, but for people or organizations with expertise relevant to the query topic. Expert search has been introduced as a task in the Enterprise track of TREC 2005 [4], and various techniques based on voting models [6], or language models [1, 7] have been proposed. The issue of diversity, however, has not been raised in the context of expert search.

In expert search, queries may correspond to a description of scientific or technological needs. The results are references to people or organizations with relevant expertise. When developing a cross-disciplinary expert search system, or one that covers many geographic regions, diversity of results is important for

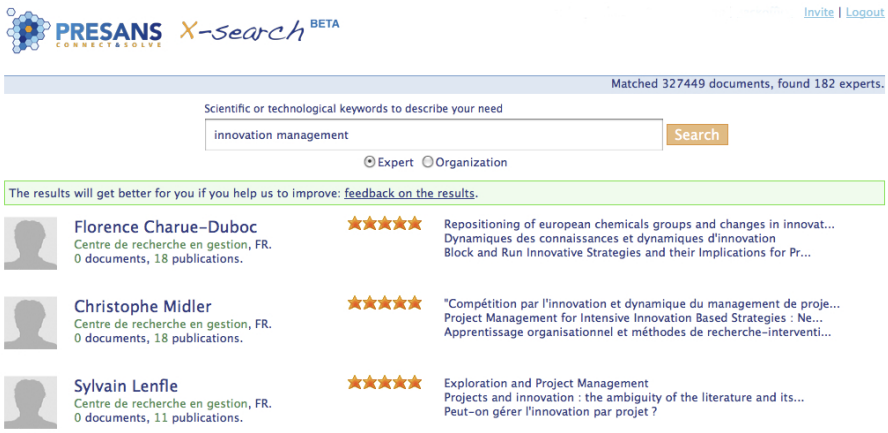

Fig. 1. Screenshot of X-Search and top results for the query *innovation management*.

broad or ambiguous queries. In this position paper, we first define a measure of topical diversity based on the distribution of matched documents and experts over a taxonomy of scientific disciplines. We demonstrate the topical diversity measure on a sample of queries from X-Search[1], an expert search engine developed by PRESANS, a startup company connecting business needs to world-wide expertise, in collaboration with LIX at École Polytechnique (France) and DB-NET laboratory at AUEB (Greece). Figure 1 shows a screenshot from X-Search. We also discuss aspects of search results diversity for expert search, and the implications of diversification at different stages of an expert ranking model.

2 Quantifying Topical Diversity for expert search queries

We introduce a measure for quantifying the topical diversity of queries based on the tree distance between nodes in a taxonomy of scientific disciplines. Intuitively, a query is more broad or ambiguous when the retrieved documents relate to scientific disciplines, which are far apart in the taxonomy. We demonstrate the defined measure with a dataset of publication titles and abstracts, harvested using the protocol OAI-PMH from a repository of publications. Most of the publications are annotated by the users who import the data with relevant scientific disciplines from a taxonomy.

For a query q, we rank the top-50 publications, using conjunctive matching and BM25 scores based on terms and bigrams. We form the set $\mathcal{S}_q = \{s : p_i \text{ annotated with } s\}$, where p_i is the publication ranked at position i, and s is a scientific discipline. We rank experts using CombSUM [6], where each publication is associated to each one of the authors. Each expert is annotated with all the topics related to all his publications. We define the set

[1] Available at http://www.presans.com/x-search

$\mathcal{S}'_q = \{s : e_i \text{ annotated with } s\}$, where e_i is the expert ranked at position i. Similar to [5], we define the distance between two nodes s and t as follows:

$$dist(s,t) = \sum_{i=l(lca)}^{l(s)} \frac{1}{2^{i-1}} + \sum_{i=l(lca)}^{l(t)} \frac{1}{2^{i-1}}$$

where $l(s)$ is the depth of s in the taxonomy (the root of the taxonomy has a depth equal to 1) and lca is the lowest common ancestor of s and t.

The weight $w(s)$ of a node s in the taxonomy is equal to the sum of scores of all p_i annotated with the corresponding discipline. The weights $w(s)$ are normalized so that $\sum_{s \in \mathcal{S}_q} w(s) = 1$. The topical diversity of documents $td_{doc}(q)$ and experts $td_{exp}(q)$ retrieved for query q is defined as the average distance between pairs of nodes in $\mathcal{S}_q$ and $\mathcal{S}'_q$, respectively:

$$td_{doc}(q) = \frac{\sum_{s,t \in \mathcal{S}_q, s \neq t} dist(s,t)}{\binom{|\mathcal{S}_q|}{2}} \quad td_{exp}(q) = \frac{\sum_{s,t \in \mathcal{S}'_q, s \neq t} dist(s,t)}{\binom{|\mathcal{S}'_q|}{2}}$$

Table 1 shows the values of $td_{doc}(q)$ and $td_{exp}(q)$ for a sample of queries submitted to X-Search. The queries are ordered according to td_{doc}. We observe that $td_{exp}(q) > td_{doc}(q)$ for 7 out of 10 queries, because we consider all the disciplines associated with an expert instead of the ones corresponding to the matched documents only. As a consequence, td_{exp} may be useful in returning more semantically associated research disciplines from the taxonomy when processing very specific queries for which there are few experts.

Table 1. Topical diversity values obtained from the ranking of documents and experts, respectively, for a sample of queries submitted to X-Search.

No.	Query	td_{doc}	td_{exp}	No.	Query	td_{doc}	td_{exp}
1	calcium dimer	1.5250	2.0057	6	complex systems	2.0314	2.0994
2	ultracapacitor	1.6667	2.0929	7	turf	2.1045	2.2487
3	data mining	1.7009	1.9492	8	digestion	2.1588	2.0120
4	innovation	1.8172	1.7823	9	biomass	2.2542	2.3709
5	voice recognition	1.9942	2.0715	10	nanotechnologies	2.2702	2.2426

Table 2 shows the top-5 disciplines associated to the queries *data mining* and *biomass*, respectively. We see that the first query is more specific to Computer Science. The second query is associated to three different top-level scientific disciplines, and hence, it is more broad as shown by the values $td_{doc}(q)$ and $td_{exp}(q)$ in Table 1.

3 Aspects of Diversity in Expert Search

The concept of diversity in expert search has several aspects. Some aspects also apply in Web search, while others are specific to expert search.

Table 2. Top 5 taxonomy nodes associated to the queries *data mining* and *biomass*, respectively, and the corresponding weights $w(s)$.

Scientific disciplines for query: data mining	$w(s)$
Computer Science/Databases	0.3945
Computer Science/Artificial Intelligence	0.1020
Computer Science/Learning	0.0550
Computer Science/Information Retrieval	0.0519
Computer Science/Other	0.0425
Scientific disciplines for query: biomass	**$w(s)$**
Sciences of the Universe/Ocean, Atmosphere	0.2502
Sciences of the Universe/Continental interfaces, environment	0.0732
Life Sciences/Ecology, environment	0.0509
Sciences of the Universe/Astrophysics	0.0465
Physics/Astrophysics	0.0465

One aspect of diversity is *topical diversity*. A query matching a broad field of science, such as physics, can potentially match any expert on physics, while the intention of a user might be to find experts in quantum physics. In both Web search and expert search, we can employ a taxonomy, such as the DMOZ directory, to estimate the distribution of topics in the results. A key point which is specific to expert search, however, is the possibility to estimate the distribution of topics at the level of documents, or at the level of experts (see Section 4).

Another aspect of diversity is *geographical diversity*. An expert search engine matches experts affiliated to institutions. Hence, one objective of an expert search engine can be to optimize the coverage of a geographic region in the matched results. For example, the intention of a user, who is forming a consortium for a European Union project proposal, may be to find experts from at least a minimum number of European countries, rather than from only one country.

Typically, expert search engines explain the ranking of an expert by offering a list of supporting documents, which may be for example Web pages, publications, patents, posts in forums or blogs. Each type of document carries different weight in supporting expertise, and *supporting document diversity* can be important when the objective of the expert search engine is to match experts having both publications and patents, or who actively participate in online forums.

4 Diversity at different stages of ranking experts

Expert ranking is typically a three-stage process. In the first stage, documents matching the query are retrieved and ranked. In the second stage, experts are associated to the ranked documents. The association of documents to experts can take place offline at indexing time, when the association is based on the occurrence of the expert's name in the title or the anchor text of a document, or when the expert is the author of a publication. The strength of the association

can also be computed at query time, when it depends on features such as the distance between the occurrence of an expert's name and the query terms in documents. In the third stage, a score is computed for each expert based on the associated documents, and a ranked list of experts is produced.

Methods to promote diversity of results can be applied at each of the three stages with different implications about the final ranking. When diversifying the initial document ranking, the final ranking favors people with cross-field expertise. When directly diversifying the matched experts, then the final ranking will favor experts specialized in distinct subtopics. Finally, diversifying the supporting documents of each ranked expertise expected to boost in the final ranking experts with a variety of supporting documents. For example, an expert who has both published articles and filed patents on a topic, or someone who has been active in different countries, would obtain more associated documents.

5 Concluding remarks

While diversity has been mostly studied in the context of Web search so far, in this paper, we argue for the need to apply diversity techniques in the case of expert search. We have defined a measure to quantify the ambiguity or broadness of queries and demonstrated it on a sample of queries collected from an expert search engine, X-Search. We have also discussed how different objectives of an expert search engine relate to the various aspects of diversity in results, as well as the implications of applying diversification at different stages in ranking experts.

References

1. Balog, K., Azzopardi, L., de Rijke, M.: Formal models for expert finding in enterprise corpora. In: Proceedings of the 29th ACM SIGIR. pp. 43–50 (2006)
2. Clarke, C., Craswell, N., Soboroff, I.: Overview of the TREC 2009 Web track. In: Proceedings of the 18th Text REtrieval Conference (TREC 2009) (2009)
3. Clough, P., Sanderson, M., Abouammoh, M., Navarro, S., Paramita, M.: Multiple approaches to analysing query diversity. In: Proceedings of the 32nd ACM SIGIR. pp. 734–735 (2009)
4. Craswell, N., de Vries, A., Soboroff, I.: Overview of the trec 2005 enterprise track. In: Proceedings of the 14th Text REtrieval Conference (2005)
5. Gollapudi, S., Sharma, A.: An axiomatic approach for result diversification. In: Proceedings of the 18th WWW Conference. pp. 381–390 (2009)
6. Macdonald, C., Ounis, I.: Searching for expertise. Comput. J. 52, 729–748 (2009)
7. Petkova, D., Croft, W.B.: Hierarchical language models for expert finding in enterprise corpora. In: Proceedings of the 18th IEEE International Conference on Tools with Artificial Intelligence. pp. 599–608 (2006)
8. Skoutas, D., Minack, E., Nejdl, W.: Increasing Diversity in Web Search Results. In: Proceedings of WebSci10 (2010)
9. Song, R., Luo, Z., Wen, J.R., Yu, Y., Hon, H.W.: Identifying ambiguous queries in web search. In: Proceedings of the 16th WWW Conference. pp. 1169–1170 (2007)

SOPHIA: bridging the gap between thematic modelling to interactive diverse search

Niall Rooney[1], Vladimir Dobrynin[2], David Patterson[2], Mykola Galushka[2]

[1] School of Computing & Mathematics, University of Ulster,
University Of Ulster at Jordanstown
Newtownabbey, BT37 OQB, United Kingdom
nf.rooney@ulster.ac.uk
[2] Vladimir Dobrynin, David Patterson, Mykola Galushka
{v.dobrynin, d.patterson, m.galushka}@sophiasearch.com

Abstract. Sophia provides a unique approach to structure a document corpus based on identification of key and distinct intrinsic themes that form the basis to a partitional clustering. We advanced this information theoretic model to describe the various facilities to provide diverse search for knowledge workers in various domains of usage.

Keywords: clustering, interactivity, visualization

1 Introduction

Our motivation, even at an early stage of our research, was to provide an information retrieval model that would provide the foundation for a rich and structured mechanism to facilitate exploratory search, where a user's information needs are not clear [2]. In this manner, a user would not rely entirely on the formulation of precise queries (with all the difficulties this entails), but would hopefully be "guided" by the system to improve recall. The information retrieval model was based on an information theoretic approach to document clustering referred to as contextual document clustering or **Sophi**sticated **I**nformation **A**nalysis (SOPHIA) in the literature [1,4,5]. SOPHIA is an efficient and easily parallelizable mechanism to cluster a whole corpus of documents based on the identification of the primary and diverse themes or contexts within a corpus. With each document assigned to its best matching cluster according to a similarity metric between a document and a cluster, graph-based mechanisms are the basis to organize the documents within a cluster to form regions of document similarity at different levels of granularity. We refer to this process as **SOPHIA-indexing** which was based on a relational data model. There are many challenges in this area such as how themes are best represented, how best to choose the set of diverse themes, and how to ensure a relatively high theme coverage of documents. We do not have space in this paper to describe in detail the SOPHIA model or how we addressed the latter issues. The interested reader is directed to our publications. The research into SOPHIA lead to the formation of a startup company SOPHIASEARCH (www.sophiasearch.com) as a collaboration between researchers

at the University of Ulster and St Petersburg State University in 2007. In the commercial arena, there was a strong focus on being able to provide robust and efficient query visualization mechanisms to a number of proprietary document collections and publicly available document stores such as the New York (NY) Times annotated corpus[6], PUBMED/MEDINE abstracts [3] and US patents abstracts [7]. As such, two components were added to the original base system, **SOPHIA-query** concerned with visualization of search results and **SOPHIA-bridge** which provided the construction and access of additional relational structures to enhance **SOPHIA-indexing** and allow a higher level of functionality within **SOPHIA_query.** A constraint on the level of functionality was based on a consideration of efficiency of end usage. The system is currently multithreaded with work ongoing to make it distributed for cloud computing but we also wanted to minimize the footprint of the databases produced as a result of building the index.
Recently we have been able to refine, through direct user feedback, how best to enhance our query tool and where necessary the underlying component functionality. This iterative process is shown in Figure 1 and has been in effect, a form of AGILE development.

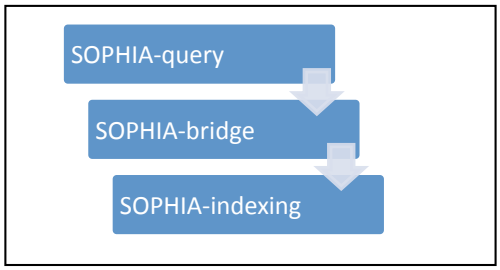

Fig. 1. The components of SOPHIA and their dependencies

Here we describe in a subjective fashion, our experiences and challenges in developing such a system.

2 Interactive Search

One obvious approach to allow users to execute richer queries than key words and phrases was to provide a query by example. This allows a sample of text to be uploaded as a query. These queries were treated in a similar fashion to the handling of original documents by **SOPHIA_indexing** in the construction of a database. As such we were able to identify the most relevant clusters, and within each cluster the most relevant documents. Keyword queries also provided the most relevant clusters and documents within a cluster but their formulation was akin to traditional keyword retrieval mechanisms formed at a cluster rather than at a corpus level. In either case the user is presented with a flat list of clusters, each consisting of a list of relevant documents. In the interests of brevity, the rest of the description relates to key word queries only. This raised two challenges a) how best to convey the contextual nature of a cluster given the query and b) how best to provide a ranking of relevant

documents within a cluster where relevancy was simply a binary measure – a document is relevant if it contains one or more of the query terms. The first issue proved problematic as a context is a probabilistic distribution in a similar manner that a topic in a topic model is a distribution of terms. A context is thematically more broad ranging than the relevant documents returned per cluster. A naïve approach of providing context descriptions was insufficiently specific as we were not paying attention to the actual query content. Mechanisms that tried to capture the cluster graphically proved unwieldy as for large corpora, many hundreds of documents may be relevant. It was necessary to focus not only on the context but the most relevant documents to the query, where the previous definition of relevancy is insufficient. This required us to develop a new integral measure of relevancy which considered not only the relevancy of a retrieved document but also that of similar documents. This was facilitated through **SOPHIA_bridge**, which derived not only a graph for the cluster but also sub-graphs or neighbourhoods for each document as areas of "closeness" within the tree. We were able to provide neighbourhood key words for each neighbour and provide key snippets or extracts from each document each document based on neighbourhood key words. In addition, key phrasal tags were derived for each neighbourhood. By necessity these neighbourhoods were non-mutually exclusive. By forming this improved ranking of documents, we were able to better solve the first issue but also address the second issue, by gleaning information from the top ranked documents as well as the cluster as a whole. We wanted to use such information to best provide a distinct description of each retrieved thematic cluster, in light of the query, the cluster's context and the top ranked retrieved documents.

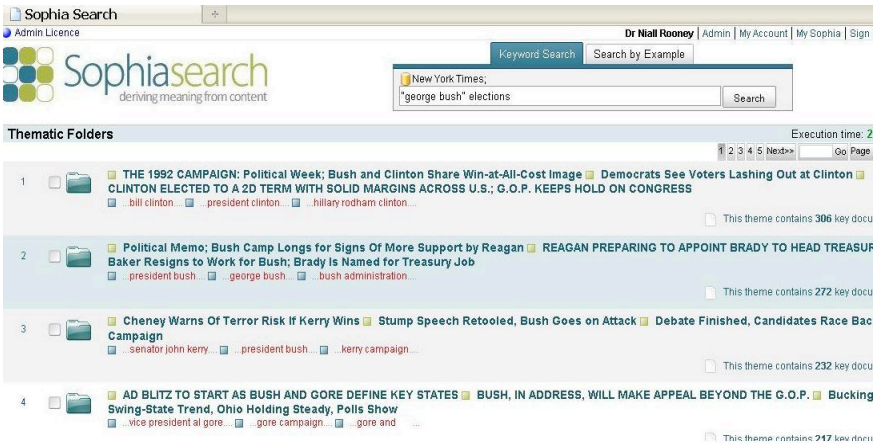

Fig. 2. "George Bush" elections query against the NYTimes corpus

We considered many mechanisms to balance the need to provide distinct descriptors against the need to prevent either information underload (descriptions *are too terse and the user could overlook or misinterpret an important thematic cluster*) and overload (*the user is required to interpret too much information to make a choice as to which theme to explore*). All approaches focused on the use of titles, snippets for top ranked documents and most frequently occurring phrasal tags for the cluster as a

whole. For the referenced corpora, we found that titles and phrase tags provided the best approach.

However in general there is no optimal solution and per corpus, we provide configuration mechanisms to allow for different approaches to be chosen. For example, Wikipedia documents have very short titles and the use of snippets is preferable. Figure 2 shows a screenshot for the query consisting of the phrase "**George Bush**" and keyword **elections** in the Sophia user interface for the NY Times corpus. A cluster and its set of relevant documents are referred to as a thematic folder, as the user has no requirement to understand how clusters are formed. This, in light of the nature of the given corpus is not a very specific query and as such for each retrieved folder, there are many relevant documents. However it is clear that the first four retrieved folders relate to different US elections, that either George Bush Senior or Junior competed in. If the user had a particular election in mind but not conveyed, he/she can quickly eliminate irrelevant thematic folders and discover only other lesser ranked thematic folders also related to this election. Of necessity, the existence of the latter is likely to be the case as SOPHIA context/themes are often more general in scope than simple topics, and the distinction between relevant thematic folders to this query will be based on different and often subtle aspects of this election e.g. the 2nd thematic folder is concerned with support from Ronald Reagan for the 1988 election. It may be that the user requires high recall (or at the very least diversity in the retrieved documents), and SOPHIA provided mechanisms for users to retain per query session and documents that he/she deemed relevant.

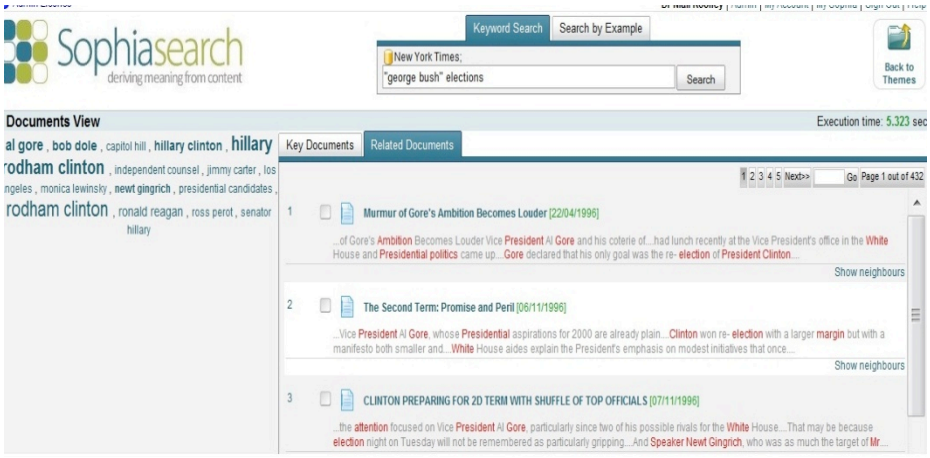

Fig. 3. Document level View within a folder

The next level of exploration is based within the thematic folder. The presentation of documents at first glance may seem similar to standard presentation mechanisms, but we allowed for two features to aid the user. Firstly we were able not only to provide lists of ranked documents as previously described but also list of ranked related documents which did not contain the query term. These related documents were based on the ranking of relevant documents and their pertinent neighbourhoods. The second

important mechanism was to allow these lists to be filtered. Each cluster and hence thematic folder, independent of the query has a set of key phrase tags and these provide a mechanism to filter documents. An example of this process is shown in Figure 3, where the first returned thematic cluster has been selected to view documents. The left hand side panel shows the list of cluster phrasal tags. If the user were to select the tag "al gore" then only documents that contain that term would be displayed in the currently highlighted list (in this case, the list of Related documents). The right hand panels show lists of Key and Related documents. Note it is possible also to explore the neighbourhood of any document as well through the *Show Neighbour* links. Each document in a Key and Related list has a document summary beneath its title based on document snippets and the document summaries have communal key words highlighted in red.

3 Conclusions

In this paper, we outlined the strategies to provide exploratory search based on SOPHIA's ability to structure a corpus related to implicit themes. We demonstrate the necessity of having flexible browsing based mechanisms to maximize the potential for users to find relevant documents. We believe many of the approaches we have adopted, may also be of relevancy to other mechanisms of structuring corpora based on intrinsic linguistic constructs such as topic modeling.

References

[1] Dobrynin, V., Patterson, D. and Rooney, N., Contextual Document Clustering. In *Proceedings of the 26th European conference on Information Retrieval Research*, LNCS 2997, pp. 167-180, Springer, 2004.

[2] Marchionini, G. (2006). Exploratory search: From finding to understanding. *Communications of the ACM, 49*(4), 41-46.

[3] PubMed, U.S. National Library of Medicines National Institutes of Health, http://www.ncbi.nlm.nih.gov/pubmed/.

[4] Rooney, N. Patterson, D., Galushka, M., Dobrynin, V.: A scaleable document clustering approach for large document corpora. Information Processing and Management, 42(5): 1163-1175, 2006.

[5] Rooney, N.,Patterson, D., Galushka, M., Dobrynin, V., Smirnova, E.: An investigation into the stability of contextual document clustering. JASIST 59(2): 256-266, 2008.

[6] The New York Times Annotated Corpus New York Times corpus, LDC2008T19, Linguistic Data Consortium, http://www.ldc.upenn.edu/.

[7] US Patent data abstracts, United States Patents and Trademark Office, http://www.uspto.gov/.

Explicit Query Diversification for Geographical Information Retrieval

Davide Buscaldi[1] and Paolo Rosso[2]

[1] LIFO, Laboratoire d'Informatique Fondamentale d'Orléans,
Université d'Orléans, 45100 Orléans, France
davide.buscaldi@univ-orleans.fr
[2] Natural Language Engineering Lab,
ELiRF Research Group,
Dpto. de Sistemas Informáticos y Computación,
Universidad Politécnica de Valencia, 46022 Valencia, Spain
prosso@dsic.upv.es

Abstract. In this paper we make a first attempt to evaluate the potential of diversity in the Geographical Information Retrieval task. This task represents an opportunity to take advantage of diversity, given that documents are relevant not only from a thematic point of view, but also spatially. A user of a GIR system may be interested in results that are geographically distributed and equally relevant. We attempted to diversify results explicitly, reformulating queries using the meronyms of the places contained in the original queries, with the help of a geographical ontology. The obtained results show that a theoretical improvement is possible, but this approach may be effective only in the case that the relevant documents do not contain enough geographical data.

1 Introduction

Diversity search is an Information Retrieval (IR) paradigm that is somehow opposed to the classic IR vision of "similarity search", in which documents are ranked accordingly to their similarity to the query. In the case of diversity search, similarity to the query is not the only criterion to determine relevant results: they should be different one from each other under some aspect, in order to satisfy the user information needs from different points of view which may be known to the user or not. For instance, if the user submit an ambiguous query, it is possible that he is not aware of its ambiguity, and the system should return a mixture of documents which may provide a complete picture of all interpretations, allowing the user to take a further step and decide which aspect of the query is more relevant to him/her. However, ambiguity is not the only source of diversity. Information is often temporally and/or geographically constrained, such that the results of a given query may be diversified in the temporal or spatial dimension, in order to provide the user with a picture of the evolution of a topic in time, or to give him/her an idea of how the topic may be relevant to a specific sub-region of a broader region named in the query. For instance, the temporal diversification

of the query "Countries of the European Union" may result in a list of documents where each document describes the countries entering into the Union in a specific year (the complete set of relevant documents show the history of the adhesions); the geographical diversification of the same query may return documents where the perspective is switched to the membership of a single country (the complete set of relevant documents provides a full coverage, from a geographical point of view, of the topic).

Until now, the objective of most diversity-related research works has been to provide multiple distinct interpretations for ambiguous queries [1,2,3]; less works have dealt with the representation of sub-topics within search results for queries with broad thematic scope [4]. Spatial diversity has been successfully applied to image search in [5]; Tang and Sanderson [6] showed that spatial diversity is appreciated by users. Clough et al. [7] analysed query logs and found that in the case of place names ambiguities users tend to reformulate queries more often.

The objective of this paper is to determine the potential of geographical diversity in the context of Geographical Information Retrieval (GIR). In GIR, queries are geographically constrained: therefore, it is possible, with the help of a geographical ontology, determine the sub-topics directly from the query (for instance: Europe is diversified in all its component countries) and build a set of reformulated queries, one for each subtopic. With the help of GeoWorSE, a GIR-enabled search engine, and the evaluation framework (queries and documents) of GeoCLEF[3], we attempted to determine the effects of the diversified sub-queries on the retrieval results.

The remainder of this paper is structured as follows: in Section 2 we describe the retrieval framework, in Section 3 we describe the collection used and the experiments carried out; in Section 4 we present the results of the experiments and an analysis of these results; finally in Section 5 we draw some conclusions and set the path for future works.

2 The GeoWorSE Retrieval System

In our experiments we used the GeoWorSE retrieval system [8]. This system is built around the Lucene search engine and a geographical ontology based on Geonames[4] and WordNet [9]. It is based on the enrichment of the index with terms that are not contained in the examined document but which can be inferred from the geographical entities in the document text.

During the indexing phase, the documents are examined in order to find location names (*toponyms*) by means of the Stanford conditional random fields-based NER system. When a toponym is found, in the case it has more than one referent according to the geographical ontology, the correct reference for the toponym is selected using a density-based disambiguation algorithm [10], with a context composed by the other toponyms contained in the document. Then, holonyms and synonyms of the toponym are extracted from the ontology and

[3] http://ir.shef.ac.uk/geoclef/
[4] http://www.geonames.org

added to an *expanded* index, together with the original toponym. For instance, consider the following text from document GH950630-000000 in the GeoCLEF collection:

> ...The *British* captain may be seen only once more here, at next month's world championship trials in **Birmingham**, where all athletes must compete to win selection for *Gothenburg*...

Let us suppose that in the ontology there are two possible referents for "Birmingham": "Birmingham/Alabama", and "Birmingham/England". "Gothenburg" is found only once but with synonyms *Goteborg* (the original Swedish name) and the alternate spelling "Goetenborg". Let us suppose that the disambiguator correctly identifies "Birmingham" with the English referent, then its holonyms are *England*, *United Kingdom*, *Europe*. In the case of "Gothenburg" we obtain *Sweden* and *Europe* as holonyms, "Goetenborg" and "Goteborg" as synonyms. Therefore, the words added to the expanded index for the above paragraph are: *Birmingham*, *England*, *United Kingdom*, *Europe*,*Gothenburg*, *Goteborg*, *Goeteborg*, *Sweden*.

The *geo* index contains the geographical coordinates associated to the above toponyms. All document terms are stored in the *text* index. The *text* and expanded indices are used during the search phase; the *geo* index was not used for search in this work, but only in the analysis of the results. In Figure 1 we show the architecture of the indexing module.

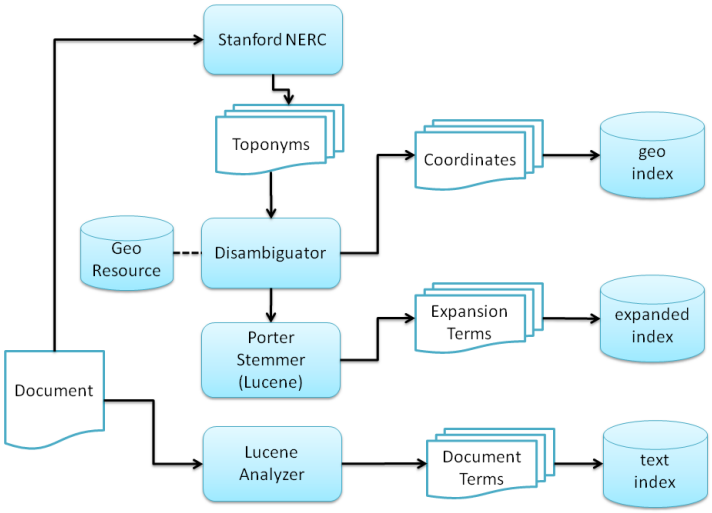

Fig. 1. Diagram of the Indexing module

The topic text is searched by Lucene in the text index. All the toponyms are extracted by the NER system and searched for by Lucene in the *expanded* index The result of the search is a list of documents ranked using the $tf \cdot idf$ weighting scheme, as implemented in Lucene.

2.1 Query Diversification

The query terms $t_0 \dots t_n \in Q$ are grouped into two subsets, a *content* set C_q, containing the words which represent the "focus" or thematic part of the query, and a *geographic footprint* set G_q, which contains the place names P identifying the geographical constraint associated to the query. For every toponym $t \in G_q$, we search the geographical ontology for *meronyms* $M_t = \{m_0, \dots m_k\}$ (i.e., places contained in the place represented by toponym t). The diversified queries are assembled by taking the terms in C_q together with a meronym $m \in M_t$, for every t. Therefore, the number or queries built in such a way is $\sum_{t \in G_q} |M_t|$. For instance, the query "golf tournaments in Europe" would be diversified into: "golf tournaments Spain", "golf tournaments Italy", "golf tournaments UK", "golf tournaments France", etc.

Among all the produced queries, we selected the ν most promising queries as the ones having the highest mutual information (MI) between the content terms and the terms in the geographic footprint:

$$I(C_q; G_q) = p(C_q \cap G_q) \log \frac{p(C_q \cap G_q)}{p(C_q)p(G_q)} \quad (1)$$

Where probabilities are calculated as the number of hits (obtained with the baseline ranking and the indicated set of terms) divided the number of documents in the collection. If there are less than ν possible reformulations, all reformulated queries are taken into account. This selection process has the objective of identifying the relative importance of the geographical aspects underlying the original query.

3 Experimental Setup

Our experiments were conducted over the GeoCLEF 2005-2008 test collection, including a total of 100 topics with the relative relevance judgements. The document collection consists of 169,477 documents and is composed of stories from the British newspaper "The Glasgow Herald", year 1995, and the American newspaper "The Los Angeles Times", year 1994. We run the experiments using only the topic title. It was possible to build a reformulation for 45 of the 100 topics. This means that 55 topics did not include a place name or the included place names did not have meronyms in the geographical ontology (this may happen if the place can be approximated to a point or a line, such as cities or rivers). We used two baselines: the first baseline is constituted by the result obtained with the original query, without reformulation. The second baseline is made of the merged results of reformulations, using the cmbMNZ fusion algorithm [11].

The potential of diversification was examined using an oracle, that is, a system which returns the results obtained by the best (among the ν) reformulation. The "best reformulation" is the one which obtains the highest score according to the selected metric. Since we are still at the beginning of our work on diversity search, we implemented a naïve round robin technique (subsequently indicated as "RR") for the fusion of the results of the reformulated queries, consisting in building a list by taking one document in turn from each individual list and alternating them in order to construct the final merged output. This is how a user would behave while examining different sets of results (examining the top ones from each set, then the second best results, and so on). Duplicate results are removed. In this way, the merged result set can be compared with the ones obtained with the baselines.

The metrics used in the evaluation are: Mean Average Precision (MAP), Mean Relevance Rank (MRR), Precision at 5 (P@5), and Normalized Cumulative Discounted Gain (NDCG). NDCG ability to handle degrees of relevance was not exploited since the relevance judgements in GeoCLEF are binary judgements. It should be also noted that existing diversity metrics cannot be easily deployed in this task, since there are no relevance assessments at the query aspect level.

4 Experimental Evaluation

We carried out two evaluations, one with $\nu = 5$ (Table 1) and another with $\nu = 10$ (Table 2). In all measures, the baseline is better than the fused results, while the oracle is always better than the baseline. It is interesting to note that the round-robin technique allowed to obtain better results than CombMNZ with MRR and NDCG (although the difference in NDCG is not statistically relevant). NDCG has been observed in [12] to be the measure that most effectively models user preferences.

Table 1. Results with $\nu = 5$

	base	CombMNZ	RR	Oracle
MAP	0.2074	0.1935	0.1818	0.2543
MRR	0.5301	0.4923	0.5185	0.7131
NDCG	0.4710	0.4605	0.4644	0.5401
P@5	0.3435	0.3087	0.2696	0.4304

We analysed the data and found some queries that obtained always a significant improvement over the baseline with the *RR* fusion, and others for which the oracle was not able to obtain a result better than the baseline. These "critical" topics are shown in Table 3.

We examined the distributions of places in the set of relevant documents in order to understand whether geographical diversity is supported by the data contained in the test collection or not. For each query q we carried out a k-means clustering, with $k = \nu$, of the points contained in the set R_q of relevant

Table 2. Results with $\nu = 10$

	base	CombMNZ	RR	Oracle
MAP	0.2074	0.1862	0.1777	0.2612
MRR	0.5301	0.4323	0.4948	0.7512
NDCG	0.4710	0.4555	0.4616	0.5510
P@5	0.3435	0.3217	0.2783	0.4522

Table 3. "Critical" topics

Topics mostly benefitted by query reformulation (group 1)	
10.2452/GC-001	Shark Attacks off Australia and California
10.2452/GC-006	Oil Accidents and Birds in Europe
10.2452/GC-008	Milk Consumption in Europe
10.2452/80-GC	Politicians in exile in Germany
Topics negatively affected by query reformulation (group 2)	
10.2452/GC-048	Fishing in Newfoundland and Greenland
10.2452/GC-013	Visits of the American president to Germany
10.2452/GC-010	Flooding in Holland and Germany
10.2452/51-GC	Oil and gas extraction found between the UK and the Continent

documents. The desired behaviour was to obtain clusters centered on geographic areas corresponding to the places identified in the query diversification process.

We found that reformulation of queries in group 1 was effective because actually the centroids *did not* match the diversified places, while for queries in group 2, the data showed clusters centered mostly on relevant areas (we plotted these clusters in Figure 2 and Figure 3 for topics $10.2452/GC-006$ and $10.2452/GC-010$, respectively). It can also be observed that many places are distributed accordingly to the sources of the news (Glasgow and Los Angeles). Therefore, the diversification of the queries based on the geographical ontology seems to be effective only when the data do not offer enough clues to group results from a geographical viewpoint.

Fig. 2. Distribution of places in documents judged relevant for topic $10.2452/GC-006$. Cluster centroids indicated with star-shaped markers. Places are sparsely distributed and do not reflect the geographic footprint of the query.

Fig. 3. Distribution of places in documents judged relevant for topic $10.2452/GC-010$. Cluster centroids indicated with star-shaped markers. Data mostly reflect the geographic footprint of the query.

5 Conclusions and Further Work

We developed a simple method to geographically diversify GIR queries, based on the meronyms extracted from a geographical ontology. With this method, if the original query contains the name of a region which includes n places, the $\nu \leq n$ most significant places (according to the mutual information between the query content and the geographical constraint) are selected, and ν queries are submitted to the search engine. We evaluated this method over the GeoCLEF test set. The results showed that an oracle capable of selecting the best reformulation always obtains better results than the baseline for all metrics, indicating that a theoretical improvement is possible; however, the tested fusion methods are not able to capture this potential. The error analysis showed that apriori diversification of the query was useful when the geographical data are sparse, and therefore it is necessary to "drive" the query towards possible relevant results. If the geographical data in the relevant documents are dense enough to support the diversification of results, then diversity can be inferred from data and query reformulation adds noise.

In order to validate these conclusions, we will have to carry out more experiments. We will have to design a data-driven diversification algorithm (or use an existing one, such as the one proposed by [1]) and verify that in this way it is possible to exploit the geographical diversity contained in the data to improve the results in GIR. We should also evaluate the results using metrics specifically aimed to diversity.

Acknowledgments

We would like to thank the MICINN TEXT-ENTERPRISE 2.0 TIN2009-13391-C04-03 (Plan I+D+i) research project.

References

1. Santos, R.L.T., Peng, J., Macdonald, C., Ounis, I.: Explicit Search Result Diversification through Sub-queries. In Gurrin, C., He, Y., Kazai, G., Kruschwitz, U., Little, S., Roelleke, T., Rüger, S.M., van Rijsbergen, K., eds.: ECIR. Volume 5993 of Lecture Notes in Computer Science., Springer (2010) 87–99
2. Agrawal, R., Gollapudi, S., Halverson, A., Ieong, S.: Diversifying search results. In: WSDM '09: Proceedings of the Second ACM International Conference on Web Search and Data Mining, New York, NY, USA, ACM (2009) 5–14
3. 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 the 31st annual international ACM SIGIR conference on Research and development in information retrieval. SIGIR '08, New York, NY, USA, ACM (2008) 659–666
4. Vee, E., Srivastava, U., Shanmugasundaram, J., Bhat, P., Yahia, S.A.: Efficient computation of diverse query results. In: Proceedings of the 2008 IEEE 24th International Conference on Data Engineering, Washington, DC, USA, IEEE Computer Society (2008) 228–236
5. Paramita, M.L., Tang, J., Sanderson, M.: Generic and Spatial Approaches to Image Search Results Diversification. In: ECIR '09: Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval, Berlin, Heidelberg, Springer (2009) 603–610
6. Tang, J., Sanderson, M.: Spatial Diversity, Do Users Appreciate It? In: GIR10 Workshop. (2010)
7. Clough, P., Sanderson, M., Abouammoh, M., Navarro, S., Paramita, M.: Multiple Approaches to Analysing Query Diversity. In: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval. SIGIR '09, New York, NY, USA, ACM (2009) 734–735
8. Buscaldi, D., Rosso, P.: Using geowordnet for geographical information retrieval. In: Evaluating Systems for Multilingual and Multimodal Information Access, CLEF 2008 revised selected papers. (2008) 863–866
9. Miller, G.A.: Wordnet: A lexical database for english. Communications of the ACM **38**(11) (1995) 39–41
10. Buscaldi, D., Rosso, P.: A conceptual density-based approach for the disambiguation of toponyms. International Journal of Geographical Information Systems **22**(3) (2008) 301–313
11. Fox, E.A., Shaw, J.A.: Combination of Multiple Searches. In: Proceedings of the 2nd Text REtrieval Conference (TREC-2). (1994) 243–249
12. Sanderson, M., Paramita, M.L., Clough, P., Kanoulas, E.: Do user preferences and evaluation measures line up? In: Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval. SIGIR '10, New York, NY, USA, ACM (2010) 555–562

Author Index