

# Application and Evaluation of Multi-Dimensional Diversity

Teerapong Leelanupab, Martin Halvey, Joemon M. Jose  
University of Glasgow, Glasgow, G12 8RZ, United Kingdom  
{kimm,halvey,jj}@dcs.gla.ac.uk

## Abstract

Traditional information retrieval (IR) systems mostly focus on finding documents relevant to queries without considering other documents in the search results. This approach works quite well in general cases; however, this also means that the set of returned documents in a result list can be very similar to each other. This can be an undesired system property from a user's perspective. The creation of IR systems that support the search result diversification present many challenges, indeed current evaluation measures and methodologies are still unclear with regards to specific search domains and dimensions of diversity. In this paper, we highlight various issues in relation to image search diversification for the ImageClef 2009 collection and tasks. Furthermore, we discuss the problem of defining clusters/subtopics by mixing diversity dimensions regardless of which dimension is important in relation to information need or circumstances. We also introduce possible applications and evaluation metrics for diversity based retrieval.

## Categories and Subject Descriptors

H.3 [Information Storage and Retrieval]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software

## General Terms

Algorithms, Evaluation Measures

## Keywords

Multi-Dimensional Diversity, Relevance, Novelty

## 1 Introduction

With the recent explosion of digital information available in multiple forms e.g. text, image, or video etc., it has become increasingly important for IR systems to improve retrieval algorithms in order to allow users to search information more effectively. Promoting diversity in search result sets has been recognised as an important process in information retrieval for decades [2, 8]. Also, several research works have empirically motivated the need for diversity in a result [4, 11]. However, it can harm the effectiveness of the systems with respect to query relevance. Conventional IR systems employs an independent ranking approach to rank and assimilate documents in order of relevance with respect to user queries. This approach ignores the content of documents ranked in the search results. The IR systems implementing this approach are mostly appropriate when the relevant documents are very few and high-recall is required. An example of this situation is

the topic distillation on web search, where a typical surfer wishes to find a very few key relevant web sites rather than every relevant web page [14]. Nevertheless, the pure relevance ranking is unsuitable in some situations. First, there are often an enormous number of potentially relevant documents containing highly similar contents, resulted in partially or nearly duplicate information within documents in the ranking. Secondly, in a large number of cases users pose a query, for which the result set contains very *broad* topics related to multiple search aspects, or has multiple distinct meanings. For instance, the query “Brussels” represents an example of a broad query that might refer to “Brussels airport”, “Brussels parliament”, and “Brussels weather” etc. The query “Chelsea” represents an example of an equivocal query that might be “Chelsea Clinton”, “Chelsea football club”, or “Chelsea area in London” etc.

Clarke et al. [3] identify the precise distinction of the aforementioned issues between *novelty* and *diversity* in the IR domain: novelty is the need to avoid *redundancy* in search results, and diversity is the need to resolve *ambiguity* of search queries. They also point out that the data set from the TREC Question and Answering track can be used as a test collection with support for diversity rather than that of the TREC Interactive track [18] or the TREC Novelty track [15]. In the area of multimedia, Paramita et al. [9] and Sanderson [12] have created test collections, “ImageClef 2008/9”, for image search diversification. A popular approach to dealing with the redundancy problem is to provide diversity in the set of search results using explicit re-ranking functions with a user-tunable parameter, referred to as MMR diversity ranking [2] or a Harmonic measure [13] etc. as combining functions of similarity and novelty. To cope with poorly specified or ambiguous queries, a traditional approach also relies on promoting diversity in an expectation that some results containing information from a different query interpretation are presented in the search results, maximizing the chance to retrieve relevant results to the users information need. Therefore, diversity in search results can overcome these problems.

In this paper, we discuss many issues in diversity ranking within the ImageClefPhoto 2009 and describe the possible solution in dealing with the problems of defining clusters across heterogeneous dimensions in promoting diversity to reorder initial results and produce a summary. Some studies raised the importance [6, 1, 16] of specific dimensions of diversity – temporal, spatial, and visual diversification etc. Figure 1 shows various dimensions of diversity. A multi-dimensional diversity based retrieval system should consider the dimensions in search results individually according to different applications, user’s preferences and circumstances. For example, in a product search, a user, who is searching for a new laptop, queries his desired specification and features of the laptop. The shopping search system should deliver, rather than the same products, many laptops from different brands sharing the same specification and feature. The system could also alternatively recommend various accessories related to the products. Another example in a different circumstance might be a magazine editor, who has to write an entertainment column about a bibliography of “Kylie Minogue”. She wants to search images of Kylie, appearing at different times, locations and with other singers. The expected results that might satisfy her should be diversified in those dimensions that she is interested in.

The rest of this paper is organised as follows. In Section 2, we provide a literature survey of related work. Section 3 proposes a possible application of promoting diversity, followed by an evaluation measure for multi-dimensional diversity in Section 4. Finally, we discuss and conclude our proposal in Section 5.

## 2 Background

### 2.1 Multi-Dimensional Diversity

The number of top document results retrieved by conventional IR systems may not be relevant to users if nearly duplicate results are closely ranked. A very broad and poorly specified query posed by users can also cause a similar problem in finding relevant documents. This realisation leads to the need to promote diversity against redundancy and to increase the possibility of finding relevant documents in retrieval ranking. Despite the fact that promoting diversity is of

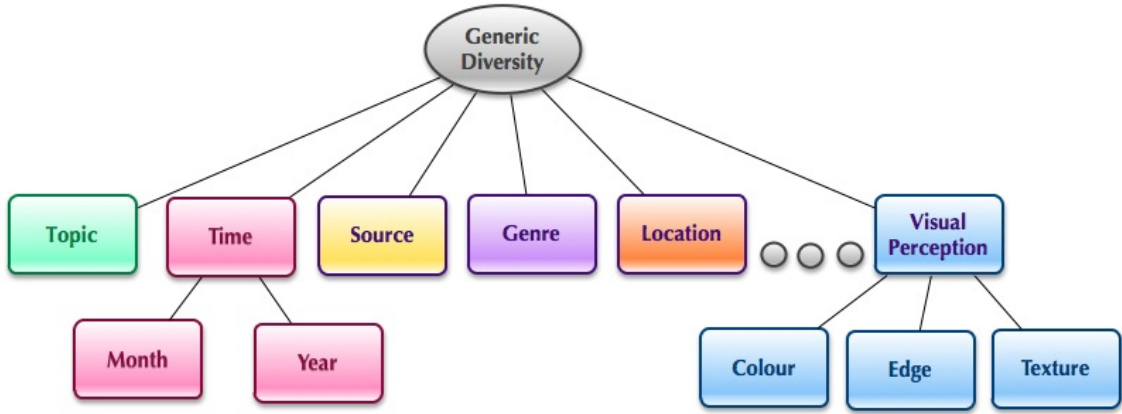


Figure 1: Dimensions of Diversity

importance, we recognize that diversity is still a broad and under-specified concept. Diversity of search results can be represented in various dimensions, such as topic [19], time [1], source (supplier in commercial search), genre (e.g. products and their accessories), location [10], visual presentation in multimedia retrieval [5, 7] etc. From our viewpoint it should be assumed that these dimensions are independent so as to simplify the difficult task of defining diversified clusters. Although defining clusters based on analysing the distribution of query variations assists in more accurately specifying diversity based on user information needs [9], the mixture of different dimensions in defining clusters casts doubt on how to effectively develop diversity algorithms and to evaluate their results from the combination of varied dimensions in diversity. For example, the clusters of topic “Beckham”, defined according to ImageClef 2009 [9], may be “David Beckham”, “Victoria Beckham”, “Beckham fragrances”, “Beckham at AC Milan”, “David Beckham 2009”, and “David Beckham and Tom Cruise”. The sample images from different dimensions related to this topic are shown in Figure 2. As we notice, this search topic is at least composed of four dimensions (i.e. anchor person (considered as topic), genre, location, and time). It was felt that defining clusters across different dimensions is inappropriate where some documents may fall into a cluster which overlaps two dimensions e.g. “David Beckham 2009”. So, there is a need to find an effective method to deal with multi-dimensional diversification.

## 2.2 Diversity Evaluation Measures

As well as devising IR models or approaches tailored for exploiting diversification within result sets, new metrics for evaluating search results must also be developed. The evaluation measures widely used for measuring the diversity of results in IR systems, such as S-recall, WS-precision [18],  $\alpha$ -nDCG [3], and subMRR [17] etc., account for subtopics or clusters of documents in a single dimension of diversity. Therefore by using clusters that are related to different dimensions, it becomes more difficult to evaluate diversity effectiveness. The commonly used assessment of the performance of the systems aimed at promoting diversity is mainly measured by a subtopic/cluster recall [18], which measures the number of subtopics/content-clusters in the investigating position. This measure is defined as follow:

$$CR@k = \frac{|\cup_{i=1}^k subtopics(d_i)|}{n_Q} \quad (1)$$

where the function  $subtopics(d_i)$  returns the subtopics or facets that are covered by document  $d_i$ ,  $n_Q$  is the total number of subtopics for the given topic/query and  $k$  is a rank position. The cluster recall is able to be then combined with standard precision for quantitative measure using



Figure 2: Sample Images in Different Dimensions Related to “Beckham” topic

a  $F$ -measure. However, this measure might not be suitable to benchmark the results when the results are mixed together from different algorithms intended to diversify the results from different aspects. The question is how can we measure the performance of diversity algorithms. If an algorithm performs better or worse, from which dimension do we get the gain. Ideally, evaluation measures, i.e. S-recall, should either investigate the clusters within the same dimension or we should find a new method to evaluate diversity over several dimensions.

### 3 Application of Diversity

Existing approaches for result diversification are system-centred where algorithms attempt to balance relevant and diverse documents containing different aspects. On the contrary, diversity algorithms should depend on the context and the information need of user. Ideally, they should separately promote diversity based on pre-defined dimensions. The retrieval systems can then visually present the results, separated into different viewpoints according to dimensions. The systems then allow for fast navigation in order to either find more similar images or narrow down search domains in a specific dimension. Otherwise, the retrieval systems can fuse the results by considering which dimension is important to users and when these dimensions are important. In this paper, we focus on the latter, which finally presents the results in a single ranking. The weight of each dimension can be initially defined according to different search domains. This weight is then adapted to the user context and information need, or manually adjusted by the users. Additionally, the ideal algorithm should rank documents by covering as many dimensions as possible.

In the Question and Answering task in TREC [3], it is suggested that relevant documents that can answer two or more questions are more important than ones answering one question. Similarly, documents which fit into multiple subtopics from different dimensions or which fall into a particular dimension where no other or few documents exist, should be ranked higher. These concepts relatively refer to the concept of *information nuggets* suggested by Clarke et al. [3], where user’s information, causing a user to formulate a query, is modeled as a set of nuggets

$\{N = n_1, \dots, n_m\}$  that the result list should contain. In the other word, this is a set of all possible individual subtopics in all dimensions  $\{C = C_{\alpha_1,1}, \dots, C_{\alpha_A,B}\}$ . Moreover, we propose to add the weight specifying the importance of dimensions with respect to the user context. An example of this concept is presented in Table 1.

<i>Documents</i>	$\alpha_1$ -Topic (2)		$\alpha_2$ -Location (3)			$\alpha_3$ -Genre (1)		<i>Total Scores</i>
	$C_{\alpha_1,1}$	$C_{\alpha_1,2}$	$C_{\alpha_2,1}$	$C_{\alpha_2,2}$	$C_{\alpha_2,3}$	$C_{\alpha_3,1}$	$C_{\alpha_3,2}$	
$x_1$	X				X		X	6
$x_2$		X		X				5
$x_3$			X					3
$x_4$					X	X		1
$x_5$	X							0

Table 1: Top Five Retrieved Documents ( $d$ ) on Three Dimension Coverage ( $\alpha$ ) and Subtopics/Clusters ( $C$ ) for Each Dimension (Weight of Dimensions for This User Context)

Let  $d$  be a document in the ranking,  $\alpha$  be a dimension given a query, and  $C$  be a cluster/subtopic in a dimension. The table shows top five documents retrieved in the ranking and the clusters that are fulfilled by the documents, according to the relevance judgement. For the purpose of this example, we define the weight of three dimensions: “Topic”, “Location”, and “Genre”, with 2, 3, and 1 respectively regarding a search domain. The last column gives the total scores of documents assigned by the sum of weights assigned when the document covers that dimension. In this example, we assume that the relevance to a query and dissimilarity of contents in the ranking are not considered. We could therefore view the coverage of dimensions as reasonable representatives and treat the total scores given as a relevance value. The ideal “ranking” for the documents would be  $x_1 - x_2 - x_3 - x_4 - x_5$ , with those documents covering more dimensions placed before those covering less. For the document  $x_4$  and  $x_5$ , the given scores are lower than what they actually gain, since we consider the novelty, the clusters  $C_{\alpha_1,1}$  and  $C_{\alpha_3,1}$  have already been covered by the document  $x_1$  ordered ahead of them. Furthermore, the documents  $x_3$  is placed at the third position since only document  $x_3$  can cover the cluster  $C_{\alpha_2,1}$ . In a real application when relevance to queries and dissimilarity of content amongst documents in the ranking are needed to be considered, this dimension coverage score can be treated as a graded diversity value that will be added to a dissimilarity value. We then obtain a novelty score, used in a combining function such as MMR [2]. The adjusted MMR can be defined as the following equation:

$$MMR_{J+1} \equiv \operatorname{argmax}_{x_i \in I \setminus J} [\lambda S(x_i; q) + (1 - \lambda)(D(x_i; (x_1, \dots, x_J)) + \alpha(x_i; q))] \quad (2)$$

where  $I$  is the set of initial results retrieved by the IR system;  $J$  is a set of re-ranked results at iteration  $J$ ;  $q$  is a query;  $x_i$  is a candidate document in  $I \setminus J$ , which is the set of documents that have not been ranked yet; and  $x_J$  is a document in  $J$ , i.e. the set of documents that have been already ranked. Function  $S(x_i; q)$  is a normalised similarity metric used for document retrieval, such as Okapi BM25, while  $D(x_i; (x_1, \dots, x_J))$  is a dissimilarity metric which, for instance, is the opposite of the cosine similarity between document vectors. The function  $\alpha(x_i; q)$  is a normalised diversity coverage score in an ideal ranking. The parameter  $\lambda > 0.5$  means that similarity to the query is more important than novelty, while  $\lambda < 0.5$  represents situations where novelty is more important than relevance to the query.

## 4 Evaluation of Diversity

Common evaluation metrics are inappropriate for retrieval tasks as discussed in Section 2. Therefore, we propose to evaluate systems which do attempt to promote diversity by separately considering clusters from the same dimension. Similar to a conventional S-recall measure, our proposed measure for evaluating the recall of subtopics within a dimension is defined as follows.

$$CR_{\alpha_a; q}@k = \frac{|\cup_{i=1}^k \text{subtopics}(d_i)|}{n_Q} \quad (3)$$

S-recall of dimension  $\alpha_a$  at  $K$  represents percentage of retrieved subtopics in the top  $K$  documents in the dimension  $\alpha_a$ . We can use this measure to evaluate multiple ranking lists presented as a group of diversity dimensions.

In addition to measure S-recall of individual dimensions, we can evaluate the overall performance of diversification approaches by the average sum of S-recall from possible dimensions related to a query. Here, we can include the weight specified by search domains or user context. The following is S-recall for total performance:

$$CR_{total; q}@k = \frac{1}{A} \sum_{a=1}^A w_{\alpha_a} \cdot (CR_{\alpha_a; q}@k) \quad (4)$$

where  $CR_{total; q}@k$  is the average S-recall at  $K$  to the query  $q$ ;  $A$  is the number of diversity dimensions;  $w_{\alpha_a}$  is the weight of dimensions; and  $A = \{\alpha_1, \dots, \alpha_a\}$  is the set of potential relevant dimensions. This proposed S-recall measure can be employed to evaluate algorithms for promoting multi-dimensional diversity.

## 5 Conclusion

In this paper we have presented a re-ranking method and evaluation measure based on multi-dimensional diversity. It is clear that current methods for promoting diversity ranking based on simple re-ranking are unlikely to be optimal for a realistic situations that show a variety of dimensions of diversity. We suggest that result diversification should be processed and evaluated separately for specific dimensions. The results from this independent diversification can be either presented in different views, or combined into a single ranking by taking into consideration search domains and user contexts. The weight of dimension can be tuned to user's information need. The overall measure can evaluate the effectiveness of the combined ranking. Moreover, to satisfy a diverse population of searchers, there are many research challenges for new methodologies to personalise diversity dimensions or identify the level of granularity of diversity, which is desirable. We believe that future alternative approaches could include investigating user query formulation in interactive search with respect to user intention, or employing an ontology to identify possible meanings of a query and a semantic subspace to specify subtopics of documents. This also leads to new challenges related to the development of these new approaches for promoting and evaluating diversity in IR.

## 6 Acknowledgements

This research was supported by the Royal Thai Government and the European Commission under contract FP6-027122-SALERO.

## References

- [1] Omar Alonso, Michael Gertz, and Ricardo Baeza-Yates. On the value of temporal information in information retrieval. *SIGIR Forum*, 41(2):35–41, 2007.
- [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, Melbourne, Australia, 1998.

- [3] 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 '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 659–666, Singapore, Singapore, 2008.
- [4] M. Eisenberg and C. Barry. Order effects: A study of the possible influence of presentation order on user judgments of document relevance. *Journal of the American Society for Information Science and Technology (JASIST)*, 39(5):293–300, 1988.
- [5] Martin Halvey, P. Punitha, David Hannah, Robert Villa, Frank Hopfgartner, Anuj Goyal, and Joemon M. Jose. Diversity, assortment, dissimilarity, variety: A study of diversity measures using low level features for video retrieval. In *ECIR '09: Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval*, pages 126–137, Toulouse, France, 2009.
- [6] Christopher B. Jones, Ross S. Purves, Anne Ruas, Mark Sanderson, Monika Sester, Marc J. van Kreveland, and Robert Weibel. Spatial information retrieval and geographical ontologies an overview of the spirit project. In *SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 387–388, Tampere, Finland, 2002.
- [7] Teerapong Leelanupab, Frank Hopfgartner, and Joemon M. Jose. User centred evaluation of a recommendation based image browsing system. In *IICAI '09: Proc. 4th Indian Int. Conf. on AI*, December 2009. to appear.
- [8] M. E. Maron and J. L. Kuhns. On relevance, probabilistic indexing and information retrieval. *J. ACM*, 7(3):216–244, 1960.
- [9] Monica Lestari Paramita, Mark Sanderson, and Paul Clough. Developing a test collection to support diversity analysis. In *SIGIR '09: Proceedings of the 32nd annual international ACM SIGIR conference on Diversity, and Interdependent Document Relevance workshop*, Boston, USA, 2009.
- [10] Monica Lestari Paramita, Jiayu Tang, and Mark Sanderson. 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*, pages 603–610, Toulouse, France, 2009. Springer-Verlag.
- [11] Flip Radlinski and Susan Dumais. Improving personalized web search using result diversification. In *SIGIR '06: Proc. 29th Int. ACM SIGIR Conf. on Research and Development in IR*, pages 691–692, Seattle, USA, 2006.
- [12] Mark Sanderson. Ambiguous queries: test collections need more sense. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 499–506, Singapore, Singapore, 2008.
- [13] Barry Smyth and Paul McClave. Similarity vs. diversity. In *ICCBR '01: Proceedings of the 4th International Conference on Case-Based Reasoning*, pages 347–361, London, UK, 2001.
- [14] Ian Soboroff. On evaluating web search with very few relevant documents. In *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 530–531, Sheffield, United Kingdom, 2004.
- [15] Ian Soboroff. Overview of the trec 2004 novelty track. In *The Thirteenth Text Retrieval Conference (TREC 2004)*, 2004.

- [16] Reinier H. van Leuken, Lluís Garcia, Ximena Olivares, and Roelof van Zwol. Visual diversification of image search results. In *18th International World Wide Web Conference*, pages 341–341, 2009.
- [17] Jun Wang and Jianhan Zhu. Portfolio theory of information retrieval. In *SIGIR '09: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 115–122, Boston, USA, 2009.
- [18] Cheng Xiang Zhai, William W. Cohen, and John Lafferty. Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 10–17, Toronto, Canada, 2003.
- [19] Guido Zuccon, Teerapong Leelanupab, Anuj Goyal, Martin Halvey, P. Punitha, and Joe-mon M. Jose. The university of glasgow at imageclefphoto 2009. In *CLEF Workshop*, Corfu, Greece, 2009. to appear.