Developing a Test Collection to Support Diversity Analysis

Monica Lestari Paramita, Mark Sanderson and Paul Clough
Department of Information Studies, University of Sheffield
Regent Court, 211 Portobello Street
Sheffield S1 4DP, UK
{m.paramita, m.sanderson, p.d.clough}@shef.ac.uk

ABSTRACT

Promoting diversity in search has gained much recent interest, however there exists a lack of sufficient benchmarks for evaluation. In this paper, we describe the creation of a new test collection for evaluating diversity in image search. To ensure the creation of a realistic resource, query variations from an image search log were used as the basis for identifying topics that might require a search system to produce diverse output. We describe the development of our benchmark being used in the ImageCLEFPhoto 2009 image retrieval task.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search

General Terms

Experimentation, Human Factors, Verification.

Keywords

Diversity, image test collection, image retrieval, building test collection

1. INTRODUCTION

For a topic that has long been recognized as important to information retrieval [1][2], it is perhaps surprising that publicly available test collections for measuring *diversity* in search are so rare. Diversity can take many forms, from queries that are ambiguous with multiple distinct meanings (e.g. Java the island; Java the drink; Java the programming language), to queries referring to broad topics that have multiple relevant aspects associated with them (London weather; London tourist information; London history, etc). Up until 2008, researchers working on promoting diversity in search could only evaluate their approaches using a small test collection from the TREC interactive track composed of 20 topics, or they could construct small test sets of their own, or be creative in their use of test collection data so as to simulate test collections with support for diversity [2].

The dearth of test collections was highlighted in 2008 by Clarke et al [3] and by Sanderson [4] who each created test collections with support for diversity by repurposing existing test sets: one originally built to test Question Answering (QA); the other for image search. The question answering collection, from TREC, was composed of "list type" questions for which QA systems were expected to find multiple nuggets of information. Clarke et al viewed the nuggets as representing different aspects of a query. Arni et al. [5] also realized that relevant images for topics from the ImageCLEF 2007 test collection addressed multiple aspects. Consequently, Arni et al were able to adapt the relevance judgments of 39 topics from the 2006/7 collection to create a suitable (albeit limited) test collection. This was used in a large-scale system evaluation that involved >20 research groups, the results of which are published in [6].

In both cases, existing test collections intended for exploring other research questions were adapted to allow the measurement of search diversity and it is unclear whether topics and relevance judgments in the collections reflected typical diverse user needs. Consequently, it was decided to create a new test collection with a more principled approach to topic selection and definition of diversity.

This paper describes the development of this new collection starting with a brief summary of past relevant work in diversity (section 2), followed by our motivation in developing this collection (section 3). Next, the data set used in this work and the process employed to build the test collection is described with a particular focus on the topics of the collection (section 4). The results of query development and its statistics are described (section 5) and an analysis was being made to examine these diverse queries more thoroughly (section 6).

2. LITERATURE REVIEW

A brief review of literature in the area of diversity focuses on two aspects, general research on the lack of specificity in user queries (section 2.1) and more specifically the development of test collections supporting diversity (section 2.2).

2.1 Dealing with Ill-Specified Queries

It has long been recognized that user's search queries can be ill-specified: Fairthorne [7] stated that a user's query can be "*an exceedingly ambiguous phrase*". More recent work on web search continued to observe this trend. Having analyzed the length of queries from nine different search engine logs, Jansen & Spink [8] reported that around 25%-30% of queries used only one term and the average query length was 2.2 terms which, as Sanderson showed [4], are often ambiguous. More recently Clough et al [9] showed that non-ambiguous queries can also need diverse search results.

One approach to dealing with ambiguous queries is to ask the user to provide some form of clarification to their query. An alternative approach is to provide diversity in the search results in the expectation that some results will contain information from at least one interpretation of the query. Consequently, the probability that users find relevant information, regardless of their intent, is increased.

SIGIR 2009 Workshop: Redundancy, Diversity, and Interdependent Document Relevance, July 23rd, Boston, Massachusetts, USA. Copyright 2009 ACM 978-1-60558-483-6/09/07

2.2 Testing Diversity in Image Retrieval

One of the first test collections to examine diversity in search was created in 2008 by the organizers of ImageCLEFPhoto. Arni et al. [5] adapted an existing ad-hoc image test collection to support diversity. The collection contained 20,000 captioned photos on a range of subjects. Its 60 topics were chosen to provide a good coverage of the collection [10]. From the 60, Arni et al. [5] identified 39 in which there were obvious sub-sets (or *clusters*) of images relevant to the query. For example, relevant results for the query "destinations in Venezuela" could be clustered into photos from different locations within the country of Venezuela.

Arni et al. [5] took the existing relevant images of the 39 topics and clustered them. The groupings were mainly based on location, but other semantically related set of images were also found. The diversity of a retrieval system was measured using *cluster recall*, a measure that calculated the proportion of retrieved clusters to all available clusters for a particular topic. The collection was used in the ImageCLEF 2008 evaluation where results showed that participant's systems could produce diverse outputs. In a subsequent study it was also found that users preferred more diversified output over a less diverse one [11].

3. DEVELOPING A TEST COLLECTION

Given the importance of promoting diversity in search, it is important for researchers to have access to standardized evaluation resources. As successful as ImageCLEFPhoto 2008 was, it was clear that a number of aspects of the test collection needed improvement including:

(1) **A larger document collection**: the need for diversity is more likely to arise in larger more heterogeneous collections;

(2) **More realistic topics**: the topics and diversity clusters were constructed by assessors based on a retroactive examination of relevant images in the collection. In other words, the topics and diversity clusters were entirely constructed from the collection and not from user needs and consequently, it could be argued that they were unrealistic;

(3) **Topic specification**: the manner in which a topic is specified in a test collection follows a well recognized structure of listing the text of the topic (usually placed in the title) followed generally by a detailed outlining of the user need (usually placed in the description part of the topic). By contrast, there is no convention of how diversity should be specified in a test collection topic. In ImageCLEFPhoto 2008, a new "cluster" tag was added which detailed the type of diversity required for each topic. It was felt that this approach to specifying diversity was not the best, as in an operational setting, diversity was unlikely to be specified in such a way. Therefore a different approach was needed;

(4) **Relevance assessments**: the manner in which relevance will be assessed in this new collection also needs to be examined. In ImageCLEFPhoto 2008, relevance was determined in previous years when the collection was first formed. Re-arranging the relevant documents into the diversity clusters was conducted by two assessors. However, if diversity is at least in part a reflection of different users interpreting the same query in different ways, it may be necessary to ensure that relevance judgments are made by a broad diversity of assessors.

3.1 Larger Document Collection

Our initial focus was to provide a larger data set upon which the test collection could be based. We were able to locate a collection from Belga[1], a Belgian press agency, which contained just under half a million images.

3.2 More Realistic Topics

It is important to develop a test collection which is realistic; otherwise, IR systems optimized to work well on the collection might not work as well in practice. In order to ensure that the collection was as accurate as possible, we derived queries based on actual users needs expressed in a query log. It was possible to exploit a preliminary analysis of Belga query logs in 2008 by Tsikrika [12]. Even though query reformulations were not able to be extracted from this data, an examination of the query lists revealed the possibility of exploiting query variations in finding diverse queries. Each variety could be seen as a different aspect of the diversity in a test collection topic.

This data was used as a basis for constructing all the queries in this collection; not only to help identify the most submitted queries, but also to recognize common queries that appeared to need diverse search output. Clusters needed by users for a particular query were also determined based on this list. We therefore ensured that this test collection is as realistic as possible.

3.3 Topic Specification

The first two issues guided collection development to be as representative as possible of a real life situation. It was consequently also important, to decide the most sensible way to define the cluster. The topics of ImageCLEFPhoto 2008 specified a cluster type using a tagged keyword. However, it is not common for a search engine to have the result clusters of a diverse topic defined in such a way. It might also not be possible to categorize the clusters into one particular type. We therefore decided to seek an alternative and less artificial approach to specifying diversity in the topics.

4. DATA SETS

This section describes the data sets we are using in developing the collection. Data from Belga contains 498,039 captioned images, where each caption contains information about the contents of the image, such as people shown in the photo, the event taking place, and the location where the image was captured.

In order to develop realistic queries, we were provided with a list of queries submitted to the Belga search engine from 1 January 2008 to 31 December 2008. There were 1,402,990 queries, of which, 51% were submitted by registered users; the rest were submitted anonymously. Comparing the number of distinct users, Tsikrika [12] found that the number of registered users was significantly smaller than the anonymous users: 429 compared to 285,184, respectively. Since anonymous users were identified by their session ID every time they used the system, it was possible that the same user was identified by more than one ID. However, since no further information was available to identify this issue, it was assumed that different IDs referred to different users throughout the analysis.

Tsikrika [12] deleted empty queries and removed symbols from query text. Duplicate queries submitted by the same users were

[1] Belga News Agency. www.belga.be

also deleted. A list showing 225,291 queries submitted by all users was produced from this research. Another list was also produced to contain only queries which were submitted by registered users. Each of these lists contained the query text and the number of different users who submitted that text. These two lists were used in identifying diverse information needs.

4.1 Identifying Diversity

We chose to identify a set of diverse needs by finding queries issued by more than 5 users, which might be the reformulation of another query. However, having only the lists of popular queries described above, it was not possible to detect the reformulation performed by the users. We discovered from previous research that reformulation of queries is a common indicator of a need for diversity by users. We therefore tried to simulate such reformulation by using the *variations* of queries, as they are the closest approximation of reformulations. Seeking a query Q and from it searching for all queries which were supersets of Q, which could be seen as its variations. We analyzed the 500 most common queries in our two lists and found the number of diverse queries as shown in Table 1.

Table 1. Number of Diverse Queries in the Top 500 Queries

All Users	Logged In Users
89 (17.8%)	111 (22.2%)

We are aware that using this method in identifying clusters will not base the collection on a uniform sample of the search workload. However, considering that the collection set for diversity evaluation is very rare, we believe that it is important to see how IR systems deal not only with general queries, but also with those requiring diversity. We therefore deliberately chose only the top 500 queries as the base of creating the queries in this set to focus the collection in the most popular diverse queries only.

4.2 Identification of Clusters

Having found the most common diverse queries and their variations, it was necessary to identify which of the variations would become part of the list of clusters in a diverse topic in the test collection. This cluster identification was performed by studying the frequency distribution of query variations in the lists and establishing a threshold, which would determine those variations that were part of a cluster topic and those which were not. Three different types of query distribution were revealed and each required a different cluster identification process. We therefore adopted three manually applied heuristics to help locate commonly used diverse topics. Across all three heuristics, never more than 10 clusters were selected per topic. This limit was set as it was assumed that a search engine would only display the first 10 results of a search, therefore only 10 clusters would be visible on the screen.

4.2.1 Gap Method

To decide the correct cut-off point in identifying the *major clusters*, we calculated the gap score for each pair of query variations.

$$gap\_score_i = \frac{freq(Q_i)}{freq(Q_{i+1})} \qquad (1)$$

with Q_i represents query variation at rank i. This gap score computes the ratio between the frequencies of one query variation and the next.

Since a big gap score illustrates that there exists a drastic decrease of the need of next query variation, we chose cluster with the greatest gap score as the threshold. Only query variations from the first rank up to the threshold rank are chosen to be part of the topics. An example of the query handled by this method is shown in Table 2.

The query “cruz beckham” has the largest gap score compared to the others. Therefore we chose all variations up to that rank as the major clusters, which are “david beckham”, “victoria beckham”, “romeo beckham”, “brooklyn beckham” and “cruz beckham”, which are shown in bold.

Table 2. “Beckham” Query and its Variations

Queries		Freq	Gap Score
Initial Query	beckham	3,688	-
Query Variations	**david beckham**	1,394	3.05
	victoria beckham	456	3.14
	romeo beckham	145	1.73
	brooklyn beckham	84	1.29
	cruz beckham	65	**10.83**
	david beckham 1999	6	1.2
	david beckham 1998	5	1
	sandra beckham	5	-

In the above example, other query variations which fall below the threshold were not included in the clusters. This decision was made because the variations’ frequencies were very small compared to the initial query’s frequency. Consequently, the needs of including them in the clusters were not strongly justified. However, there are some other cases which show relatively large frequencies for the rest of the variations. An example of this topic is shown in Table 3.

Table 3. “Prince” Query and its Variations

Queries		Freq	Gap Score
Initial Query	prince	3,688	-
Query Variations	**prince albert**	2,334	1.99
	prince william	1,169	1.88
	prince philippe	622	1.09
	prince felipe	573	**2.01**
	prince charles	285	1.52
	prince frederik	188	1.03
	prince carl philip	182	1.3
	prince laurent	140	1.01
	prince amadeo	139	...
	...	...	...

The first four variations were chosen as major clusters based on the largest gap score. In this example, however, the frequencies of the rest of the variations are relatively big, and the total frequencies of these variations easily exceeded the frequency of "prince felipe" cluster, which was the last major cluster in the query. This shows that there exists a need of the rest of the variations.

To accommodate this situation, we created an *"other" cluster* which contained all images which were relevant to the query but not included in the initial four clusters. This query could be thought of as the initial query with all of the clusters negated. In the above example, the "other" cluster will be "prince –albert –william –philippe –felipe".

4.2.2 Upper Bound Other Method

In some other cases, the frequencies of the queries did not decrease significantly which results in rather similar gap scores between each pair. In this case, we decided to use a different approach rather than simply choosing the clusters based on the largest gap. An example of the queries in this situation is shown in the Table 4.

Table 4. "Brussels" Query and its Variations

Queries		Freq	Gap Score
Initial Query	brussels	81	-
Query Variations	**brussels airport**	50	1.47
	brussels airlines	34	2.27
	police brussels	15	1
	fc brussels	15	1.25
	metro brussels	12	1
	demonstration brussels	12	1.09
	stock exchange brussels	11	1
	brussels parliament	11	1
	brussels grand place	11	1.1
	ring brussels	10	1
	bourse brussels	10	1.25
	tunnel brussels	8	1
	school brussels	8	1
	grand place brussels	8	1
	euronext brussels	8	1
	brussels stock exchange	8	...
	...	...	...

As shown in Table 4, the gap scores and frequencies of each query variations are relatively similar to one another. Therefore, instead of setting a threshold based on the largest gap, we chose the top nine query variations as major clusters which are shown in bold. Since there are other variations outside of these clusters which have similar frequencies to the ones included, we also created a final "other" cluster. The final cluster for this query could be thought of as: "brussels –airport –airline –police –fc –metro –demonstration –stock –exchange –parliament –grand –place". In this example, nine major clusters and one "other" cluster are chosen to suit the upper bound of 10 clusters per query.

4.2.3 Exception

Note in the previous examples, the variations chosen completely defined the focus of a topic's information request. However, in some cases, the initial query was taken to be one of the clusters as illustrated in Table 5.

Table 5. "Brad Pitt" Query and its Variations

Queries		Freq
Initial Query	**brad pitt**	2,204
Variations	**angelina jolie brad pitt**	36
	brad pitt berlin	19

In this case, the frequency of occurrence of the "brad pitt" query was significantly larger than its variants. We therefore decided that it would not be reasonable to define the information need by the two variants alone. Therefore, the information need of the query was composed of three clusters: "brad pitt", "angelina jolie brad pitt" and "brad pitt berlin".

4.3 Checking Document Availability and Adding Example Images

After listing all the diverse queries and their clusters, we ran each cluster in a search engine to find one example image per cluster. In the context of this test collection, it was assumed that the example image was akin to an image commonly chosen by users after issuing the query of the cluster.

By way of testing if clusters were well represented in the collection, the cluster was eliminated if it was found that a cluster did not have a relevant image. Some elimination occurred because the image collection seemed to cover a somewhat different timeframe from the query log. This caused some popular queries and clusters being eliminated due to the unavailability of relevant images. After this process, any topics that were left with one cluster were removed.

The resulting ImageCLEFPhoto 2009 test collection contained 50 topics: 26 queries were constructed from the logged in list of users; 24 were developed from the list of all users.

4.4 Relevance Assessments

The relevance judgment is performed by using DIRECT system, which is a distributed IR evaluation campaign tool [13]. We limited the pool by selecting only the top 100 images found in the queries by each participant. There were around 100,000 images in total which were assessed by 23 assessors. The judgment process is divided into two phases. The first phase, the topic judgment, assesses whether or not the images are relevant to the query title. After all images are judged, the second phase will be started, which is the cluster judgment. In this phase, each of the relevant images will be assessed toward the cluster title. By the time this document was being written, the first judgment phase has been completed, and we are about to start the cluster judgment.

5. PRESENTING DIVERSE QUERIES

The next question to be dealt with was how to present the diverse topics to participants in the 2009 evaluation exercise. As care had been taken to ensure that the queries and their clusters were well chosen, it was also important to ensure that the way the topics were represented was realistic.

The question asked was how might queries look to a system developer seeking to adapt their retrieval system to cope with diverse requests. If the developer examined a query log, they might locate the same query variations identified in the Belga query log. Therefore, the developer could use past query and click data to train a system on the forms of diversity likely for a particular topic. However, it was also possible that a developer might wish to have their searching system diversify search outputs for previously unseen queries.

Therefore, it was decided to split the 50 test collection topics into two types, one to address each of the situations described above. Topics in the first type would include information about the cluster variations, a description (used by relevance assessors) and example images from each cluster. One example of the first type of query is shown in Table 6.

Table 6. First Type Query

Query Number	**12**
Query Title	**clinton**
Cluster Title	**hillary clinton**
Cluster Desc	Relevant images show photographs of Hillary Clinton. Images of Hillary with other people are relevant if she is shown in the foreground. Images of her in the background are irrelevant.
Image	belga26/05859430.jpg
Cluster Title	**obama clinton**
Cluster Desc	Relevant images show photographs of Obama and Clinton. Images of those two with other people are relevant if they are shown in the foreground. Images of them in the background are irrelevant.
Image	belga28/06019914.jpg
Cluster Title	**bill clinton**
Cluster Desc	Relevant images show photographs of Bill Clinton. Images of Bill with other people are relevant if he is shown in the foreground. Images of him in the background are irrelevant.
Image	belga44/00085275.jpg

In this type of topic the expected clusters were well-defined. Participants therefore know how broad or how diverse the results should be.

Although all topics in the test collection have full tests of cluster titles, descriptions and example images, the second half of the test collection topics were released to participants with all cluster information removed including information about the number of clusters. For these topics participants were required to decide on how broad the results should be and what form diversity should take. An example of the second type queries is shown in Table 7.

Table 7. Second Type Query

Query Number	**26**
Query Title	**obama**
Image	belga30/06098170.jpg
Image	belga28/06019914.jpg
Image	belga30/06107499.jpg

In the first type of topics, the number of images was the same as the clusters; in the second type, a consistent number of example images were given throughout in order to not hint to participants on how many clusters were expected. By using this type of topic, we encourage the participants to freely interpret the diversity need in the search results. The statistics of the queries are shown in Table 8 while the distribution of the clusters is shown in Figure 1. This information is not distributed to the participants of the conference.

Table 8. Queries Statistics

Number of Queries	50
Percentage of Queries with "Other" Cluster	58%
Average Number of Major Clusters per Topic	3.54
Average Number of Clusters (Major + "Other") per Topic	4.12
Range of Clusters	2 to 10
Average Initial Query Length	1.12
Average Major Clusters Length	2.18

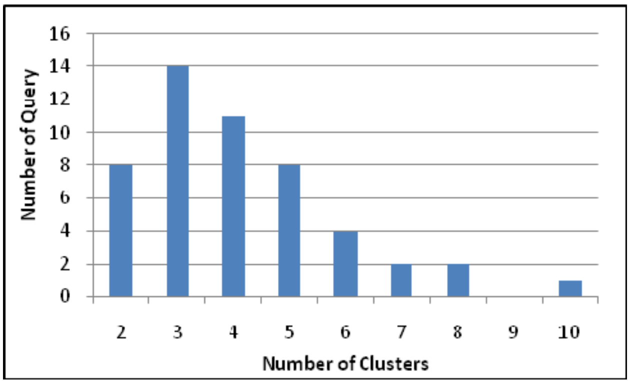

Figure 1. Distribution of Clusters

6. ANALYSIS OF QUERIES

Whilst developing the queries, we detected occasionally clusters within topics that themselves could represent a number of information needs. For example, the query "tom boonen", which is a cluster in the "boonen" topic, also has potential clusters on its own: "tom boonen 2007" and "tom boonen press". This shows that the need of diversity does not stop in one level. Addressing this multi-level diversity in collection building is a planned focus for next year's work.

The methodology we developed helped us enormously in identifying the major clusters of topics. However, we do realize some issues with this approach. Assuming that long queries are variations from the short queries might cause some problems when the initial query has different meanings. For example, query about "queen" does not necessarily refer to the royal family, as it might represent the rock band. However, since users do not normally reformulate their query to "queen rock band", use of query variations to spot clusters might not detect this variation. It is also possible for a cluster to not be a subset of the initial queries. For example, the query "new york" might not be required as one of the results when "york" is submitted to the search engine as the two of them are completely different.

Nevertheless, using the query variation approach we think is a promising start to a principled approach to creating a diverse test collection.

7. EVALUATION

To measure the effectiveness of the runs submitted to the system, two methods will be used in evaluating the results. Precision at rank 10 (Prec@10) will be used to evaluate the fraction of retrieved documents. Diversity will be evaluated by using cluster recall at rank 10 (CR@10). The latter measure was used in last year's ImageCLEFPhoto 2008 [5] to assess diversity in participants' submissions.

CR@10 evaluates the diversity factor by presenting the fraction of retrieved clusters of all the existing clusters in that query. Given the 'beckham' query shown in Table 2, results showing only images of "david beckham" will have CR score of 0.2, representing one retrieved cluster from the five clusters. Meanwhile, images showing people from all the clusters will have CR score of 1, representing fully retrieved clusters in the result.

It is realistic to expect a bigger proportion of documents from the most popular cluster, and fewer documents from the less popular ones. However, this factor cannot be detected by using the original cluster recall measure, as results with different proportions will have the exact same score if the numbers of retrieved clusters are the same. We are now developing a new measure to evaluate this factor. We intend to give higher score to IR systems which consider the clusters' popularity in finding the images. This algorithm is under development by the time the paper is being written.

According to the information in Table 8, there are 42% queries without the "other" cluster. It is therefore possible that participants managed to find relevant images which are not included in any of these clusters. Since these clusters were justified against the query logs, unidentified clusters represent a weak need of the images. It is not realistic to punish IR systems which do not find images from these unidentified clusters. Therefore, we ignored these images in the cluster recall evaluation.

8. CONCLUSIONS

In order to analyze and evaluate diversity appropriately, a realistic test collection is required, however very few suitable resources exist. A previous test collection created for ImageCLEFPhoto 2008 was constructed with a focus on promoting diversity in image search results. However, the queries were collection driven, and therefore it was hard to justify whether or not the need of diversity was correctly represented. Moreover, the collection used was relatively small.

Exploiting a new dataset and query logs from the Belga news agency, we created a larger and more representative diverse image test collection. Containing around half a million images and fifty diverse topics, this benchmark will form the basis of the test collection being used in the ImageCLEFPhoto 2009 image retrieval task. Since these topics are based on query logs associated with the document collection, they provide more realistic examples of diversity than those used in past test collections.

There was a concern about exactly how to present diversity to participating groups in the test collection evaluation campaign. Therefore, in this collection, we defined the clusters by their titles and one relevant image for each cluster. Fifty percent of the queries were also released without any cluster information at all, so that we can analyze how participants deal with diversity with no additional information. Having two different types of queries enables us to compare the participants' runs and analyze how they process diversity without knowing the clusters.

9. FUTURE WORK

Our analysis also indicates that multi-level diversity exists in user requests. Up to now, this issue is very little studied and no evaluation was available to assess this feature. Consequently, further research should be conducted to study this issue more thoroughly.

Using query variations as the base of this collection development misses some factor where diversity is not represented in keywords, such as visual diversity. Due to the data limitation, we were not able to cover this aspect in this collection development. We plan to focus on this factor more in the future as visual diversity is an important part of image retrieval.

10. ACKNOWLEDGMENTS

We would like to thank Belga Press Agency for providing us the collection and query logs and Theodora Tsikrika for the pre-processed queries which we used as the basis for this research. We would also like to thank the reviewers for their valuable comments and inputs about this work.

The work reported has been partially supported by the TrebleCLEF Coordination Action, within FP7 of the European Commission, Theme ICT-1-4-1 Digital Libraries and Technology Enhanced Learning (Contract 215231).

11. REFERENCES

[1] Carbonell, J. and Goldstein, J. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries, Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, p.335-336, August 24-28, 1998, Melbourne, Australia

[2] Chen, H. and Karger, D. R. 2006. Less is more: probabilistic models for retrieving fewer relevant documents. In Proc. ACM SIGIR, 429-436.

[3] Clarke, C. L. A., Kolla, M., Cormack, G.V., Vechtomova, O., Ashkan, A., Buttcher, S., and MacKinnon, I. 2008. Novelty and Diversity in Information Retrieval Evaluation. In Proceedings of the 31st ACM SIGIR Conference on

Research and Development in Information Retrieval (SIGIR 2008), Singapore, July 2008.

[4] M. Sanderson. 2008. Ambiguous queries: test collections need more sense. In Proceedings SIGIR'08, 2008, pp. 499-506.

[5] Arni, T., Tang, J., Sanderson, M. and Clough, P. 2008. Creating a Test Collection to Evaluate Diversity in Image Retrieval.

[6] Arni, T., Clough, P., Sanderson, M., Grubinger, M. 2008. Overview of the ImageCLEFPhoto 2008 Photographic Retrieval Task, in CLEF 2008 Working Notes

[7] Fairthorne, R. A. 1963. Towards Information Retrieval. Journal of the Operational Research Society (1963) 14, 215–216.

[8] Jansen, B. and Spink, A. 2006. How are we searching the World Wide Web? A comparison of nine search engine transaction logs. Information Processing and Management, 42 (1), 248-263.

[9] Clough, P., Sanderson, M., Abuammoh, M., Navarro, S., and Paramita, M. 2009. Multiple Approaches to Analysing Query Diversity. To appear in SIGIR 2009.

[10] Grubinger, M. and Clough, P. 2007. On the Creation of Query Topics for ImageCLEFPhoto, In Proceedings of the third MUSCLE / ImageCLEF workshop on image and video retrieval evaluation, Budapest, Hungary, 19-21 September 2007.

[11] M. Sanderson, J. Tang, T. Arni, P. Clough. 2009. What else is there? Search Diversity Examined. In proceedings of ECIR 2009. Toulouse, April 2009.

[12] Tsikrika, T. 2009. Queries Submitted by Belga Users in 2008.

[13] Di Nunzio, G.M., Ferro, N. 2005. DIRECT: a Distributed Tool for Information Retrieval Evaluation Campaigns. In: Proc. 8th International Workshop of the DELOS Network of Excellence on Digital Libraries on Future Digital Library Management Systems (System Architecture & Information Access).