# INEX 2005 Multimedia Track

Roelof van Zwol[1], Gabriella Kazai[2], and Mounia Lalmas[2]

[1] Utrecht University, Department of Computer Science, Center for Content and Knowledge Engineering, Utrecht, the Netherlands
roelof@cs.uu.nl
[2] Department of Computer Science, Queen Mary, University of London, London, United Kingdom
{gabs, mounia}@dcs.qmul.ac.uk

**Abstract.** This paper reports on the activities of the INEX 2005 Multimedia track. The track was successful in realizing its objective to provide a pilot evaluation platform for the evaluation of retrieval strategies for XML-based multimedia documents. In this first exploratory year the focus of the evaluation experiment was to test approaches for the retrieval of XML fragments using a combination of content-based text and image retrieval techniques. The track is set to continue at INEX 2006.

## 1 Challenge and Objectives

The main objective of the INEX 2005 multimedia track was to provide a pilot evaluation platform for structured document retrieval systems which are not limited to textual content, but combine multiple media types. Many, real-life structured document collections today contain a range of media such as text, image, speech, and video. Incorporating different media types within the retrieval process and producing meaningful rankings of multimedia documents and components is far from trivial. The goal of the multimedia track at INEX 2005 was to investigate this problem from a new perspective, using the structure of documents as the semantic/logical backbone for the retrieval of multimedia document fragments. In this first year, the track resulted in the construction of a pilot evaluation platform for the retrieval of multimedia structured document fragments. The methodology used in the construction of the test collection was built on established methods used at TREC [1].

The task set for the multimedia track was to retrieve relevant document fragments based on an information need with a structured multimedia character. The challenge for a structured document retrieval system in this case is then to combine the relevance of the different media types into a single (meaningful) ranking that can be presented to the user. The INEX multimedia track differs from other approaches in multimedia information retrieval, like TRECVID [2] and ImageCLEF [3], in the sense that it focuses on the use of document structure to estimate, relate, and combine the relevance of different multimedia fragments. The focus for 2005 was on the combination of text and image retrieval, where a strict interpretation of the structural constraints within the specified information need was adopted.

## 1.1 Participants

Eight groups participated in the first year of this track. These are summarised in Table 1.

**Table 1.** Participants of the multimedia track

| ID | Organisation | Created topics | Assessed topics | Submitted runs |
|---|---|---|---|---|
| utwente | Cirquid Project (CWI and U. of Twente) | 6 | 3 | 5 |
| qmul | Queen Mary University of London | 5 | 3 | 1 |
| utrecht | Utrecht University | 4 | 3 | 5 |
| rmit | RMIT University | 3 | 3 | 5 |
| qutau | Queensland University of Technology | 3 | 3 | 5 |
| ugrenoble | University of Grenoble | 0 | 2 | 0 |
| uberkely | University of Berkely | 0 | 2 | 0 |

## 1.2 Setting up the multimedia track

A step-by-step outline of the activities involved in setting up the multimedia track is given below.

- **Acquisition of the Lonely Planet WorldGuide XML collection** [4]. One of the first tasks was to acquire a suitable XML collection that was easily accessible and contained integrated multimedia objects. Such a collection was donated by the Lonely Planet organization in the form of the WorldGuide XML collection.
- **Extension of the NEXI query language**. The original NEXI query language [5] supported only text-based information access. For the multimedia track, it was necessary to allow for searching images as well as text. Therefore, a small extension to the NEXI query language was defined.
- **Baseline system for topic formulation**. Both a text-based and an image-based retrieval system was provided for participants to support the topic creation process.
- **Topic creation procedure and selection**. A topic creation procedure was set up, similar to that used by the INEX ad-hoc track, but extended to address the additional requirements of the multimedia track. For example, in addition to searching for relevant text fragments to a candidate topic, participants were asked to carry out preliminary searches for relevant images. Topic creation resulted in a pool of 25 topics, which now form part of the INEX multimedia test collection. These topics were then used to evaluate the participants' retrieval systems.
- **Assessment procedure**. Binary relevancy judgments and the yellow-marker design for obtaining relevance assessments were employed in the track. Assessments were collected using the XRAI assessment tool [6], which was

adapted by Benjamin Piwowarski to fit the track's assessment procedure (which is different from that employed by the ad-hoc track). As a result of the employed binary relevance, the two-step assessment procedure used in the ad-hoc track was replaced with a single step: Assessors were asked to mark those multimedia fragments relevant that satisfied all requirements of the information need. Based on the SSCAS interpretation of a topic's structural constraints, both support elements, i.e. where to look, and target elements, i.e. what to return, had to be strictly matched by relevant fragments. All eight participants took part in the assessments. Two topics, topic 18 and 20, remained un-assessed.

– **Evaluation of results**. In total, twenty-five runs were submitted by five participants (qmul, qutau, rmit, utrecht and utwente). Using the TREC evaluation tool, trec_eval [1], the results of several standard measures used in TREC were reported as indications of retrieval performance.

We detail each of the above steps in separate sections in the remainder of this paper.

## 2 Lonely Planet WorldGuide XML Document Collection

The corpus used for the INEX 2005 multimedia track was based on the Lonely Planet WorldGuide collection, made available by the Lonely Planet organization. The collection consists of 462 XML documents, each providing information about holiday or travel destinations. The most likely users of the collection are hence travelers who are interested in researching the locations of their next holiday or business trip. The collection can be viewed online at: `http://www.lonelyplanet.com/worldguide/`. The collection contains information about countries, regions and major cities. For each destination an introduction is given, complemented with information about transport, culture, major events, facts, and an image gallery that provides an impression of the local scenery. In Table 2 some additional statistics for the LonelyPlanet document collection are given.

**Table 2.** Lonely Planet WorldGuide collection statistics

| | |
|---|---|
| Total number of XML documents | 462 |
| Total number of images | 2633 |
| Average number of images per file | 6.7 |
| Average depth of XML structure | 4.73 |
| Maximum depth of XML structure | 8 |
| Average number of XML nodes per document | 440 |

## 3 Topic Creation

We first provide two example topics to introduce the requirements and the chosen approach for combining text and image components within a user request. The main consideration is the inclusion of content-based image retrieval into the specification of an information request using the NEXI query language [5][3].

In its first year, the multimedia track focused on the use of Content And Structure (CAS) topics as these allow for the explicit representation of the multimedia character of an information request. We refer to these as NEXI-CAS queries.

### 3.1 Examples

*Example 1.*

> **Information need**: *Find images depicting scuba diving for destinations with a tropical climate and with activities that discuss exploring the beautiful underwater nature by diving.*
> **Information request**:

```
//destination[ about(.//weather,tropical climate)
     and about(.//activities, beautiful "underwater nature" diving)]
  //images//image[about(., scuba diving)]
```

The information need of Example 1 contains both textual and image components. E.g. `about(.//weather,tropical climate)` specifies the condition requesting information about a tropical climate that is to be found within the XML element `weather`, which is a descendant of a `destination` element. Furthermore, requested `image` elements should depict scuba diving scenes.

Although the target elements of the above example are images, so far, simple textual retrieval approaches may have been sufficient to produce the required output, e.g. by searching on image captions. However, a combination of text and image retrieval techniques was encouraged within the track with the aim that these may in fact produce better results. An example of a query that combines both these aspects is shown in Example 2.

*Example 2.*

> **Information need**: *Find images depicting scuba diving, like in BN5970_6.jpg, for destinations with a tropical climate and with activities that discuss exploring the beautiful underwater nature by diving.*
> **Information request**:

```
//destination[about(.//weather,tropical climate)
     and about(.//activities, beautiful "underwater nature" diving)]
  //images//image[about(., scuba diving src:/image/BN5970_6.jpg)]
```

[3] NEXI is the XML query language used in INEX, which has been specifically developed to emphasize the content-oriented access to XML documents.

The extension to the information need in Example 2, where a sample image is given by the user, enforces the inclusion of content-based image retrieval techniques into the retrieval process. To specify the corresponding information request in NEXI, a small extension to the query language was required, which is discussed next.

### 3.2 Multimedia Extension to NEXI

Two possible options were open for extending NEXI with image querying capabilities: 1.) introduce a new `depict` clause that takes an example image as one of its parameters, or 2.) to extend the already existing `about` clause for image as well as text querying. The latter approach was chosen. By expressing both the content and image components of the information need within the same `about` clause of NEXI, we are effectively overloading its meaning, leaving it to the retrieval system to decide if a text or image search (or both) is required. The reason for doing so was to emphasize the multimedia nature of the track. By adopting this overloaded `about` clause, we can specify query constraints for a document fragment (which may be pure text, image, or a combination of the two media types) using textual descriptions (e.g. `about(//image, scuba diving)` or `about(//destination, scuba diving)`) or using example images (e.g. `about(//image, src:/image/BN5970_6.jpg)` or `about(//destination, src:/image/BN5970_6.jpg)`) or any combinations of the above, e.g. as shown in Example 2.

### 3.3 Topic Format and Topic Development Procedure

*Topic Format.* The topic format for the multimedia track consists of the following fields: a description, castitle, and narrative. The following information is contained in each of these fields:

- <**description**> A brief description of the information need, specifying any structural, textual, and visual requirements/composition on the content. The description must be precise, concise, and informative. It must contain the same terms and the same structural requirements that appear in the castitle, albeit expressed in natural language.
- <**castitle**> A valid NEXI expression based on the Lonely Planet document collection that contains at least one `about` clause containing at least one image component. The expression is of the form //A[B] or //A[B]//C[D].
- <**narrative**> A detailed explanation of the information need and the description of what makes and element relevant or not. The narrative should explain not only what information is being sought, but also the context and motivation of the information need, i.e. why the information is being sought and what purpose it may serve.

*Topic Development Guidelines.* Each participating group was requested to submit 3 CAS topics following these steps.

**Step 1: Initial topic statement**. Creation of a one or two sentence description of the information being sought. This should be a simple description of the information need without regard to retrieval system capabilities or document collection peculiarities. The context and motivation of the information need, i.e. why the information is being sought, also had to be recorded.
**Step 2: Exploration phase**. In this step the initial topic statement is used to explore the collection and obtain an estimate of the number of relevant elements. This is necessary to evaluate whether a topic can be judged consistently and whether enough but not too many relevant answers exist for it within the collection. For this purpose two search engines were made available to participants: a text and an image retrieval system.
**Step 2a: Assess the top 25 text fragments**. Participants were asked to judge the relevance of the top 25 retrieved text fragments. Each result had to be judged on its own merits. A search was to be abandoned if there were fewer than 2 or more than 20 relevant text fragments in the top 25 results.
**Step 2b: Assess top 25 images**. Since most participants did not have an off-the-shelf system available for the multimedia track, we have chosen to carry out a separate scan for the relevance of the image component. Here, participants had to assess the top 25 returned images (using binary relevance only). As with text, each result had to be judged on its own merits. A minimum of 2 and a maximum of 20 relevant images were required within the top 25 results.
**Step 2c: Inspect document matching**. To assure that the document collection had a reasonable chance of completely fulfilling the text and image-based constraints of the information need a check at document level was needed. Participants were asked to count the number of documents that satisfied both the textual and image conditions. A candidate topic was to be rejected if less than 2 documents were found in the top 25 results for both components.
**Step 3: Topic formulation**. During this step, participants finalised the topic description, CAS title, and narrative.
**Step 4: Topic submission**. Topics were submitted using the on-line Candidate Topic Submission Form on the INEX 2005 website.

*Topic Pool.* Five of the participating groups submitted a total of 25 topics. Example 3 shows one of these submitted topics. The target elements of this query are `destination` elements, where a relevant fragment needs to fulfill the various conditions formulated using both textual and visual requirements.

*Example 3.*

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<inex_topic topic_id="13" inex_track="MM">
<castitle>//destination[about(., city river)]//images//image[
```

```
            about(., river city) and about(., src:/images/BN6288_22.jpg)]
</castitle>
<description>Find images depicting a city with a river, like in
 BN6288_22.jpg</description>
<narrative>Brisbane is a beautiful city as there is a river flowing
through the city. We want to find other cities that also have a river
so that we can create a tour of river cities. </narrative>
</inex_topic>
```

## 4 Relevance Assessments

The definition of relevance used for the assessments was based on the definition employed in the INEX ad-hoc track with the exception that exhaustivity was measured only on a binary scale. In addition, reflecting the SSCAS task, an XML element was only considered relevant if it strictly matched the structural conditions specified within the query, i.e. only target elements could be relevant and only if they were contained in an XML document that satisfied all the query's containment constraints.

Therefore, a given multimedia fragment was said to be relevant if it discussed (or depicted) the topic of request to any degree and if it strictly adhered to the structural conditions requested by the user. Similarly to the ad-hoc track, the assessment procedure followed the highlighting approach. However, given the binary nature of relevance (exhaustivity), the assessment procedure for the multimedia track consisted only of a single pass. During this single assessment phase, assessors were requested to highlight multimedia fragments that contained only relevant content, i.e. contained no (or only minimal) non-relevant information. In the case of textual content, only relevant text fragments, e.g. words or sentences, were to be highlighted. In the case of images, since it was not possible to highlight only a part of an image, the whole image was highlighted if it contained relevant content (regardless of how much of the image may have been non-relevant).

Relevance assessments were carried out using the XRAI assessment tool, developed by Benjamin Piwowarski. Figure 1 shows a snapshot of the interface, where fragments from the Lonely Planet collection are marked relevant (highlighted).

In total, eight groups took part in the assessments. As a result, 23 of the 25 topics were assessed. These then provided the basis for the evaluation in this first exploratory year of the multimedia track. Table 3 provides a summary of the obtained assessment results. Topics *8*, *21*, *22*, and *24* are shown in italic, because the assessments did not match the castitle, and the topic description. These topics were therefore removed from the official evaluation.

In this report we only present results based on the official topic and assessment pools for the original submissions from participants. The official evaluation is therefore based on a pool of 19 topics, whereas for the extended evaluation 23 topics are available.

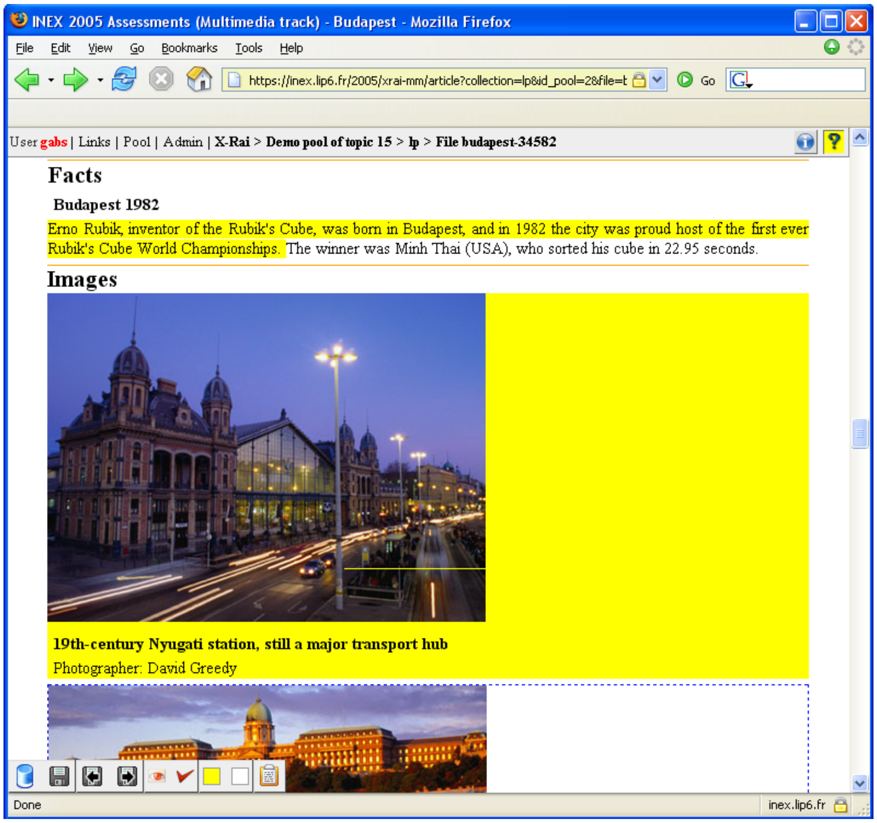


**Fig. 1.** Snapshot of the XRAI interface

## 5 Evaluation of Retrieval Performance

In this section, we discuss the results of the evaluation of the submitted runs. We report interpolated recall-precision averages and precision at fixed element cutoff levels for the best scoring run of each participant. In addition, mean average precision (MAP) scores and results of the binary preference (bpref) measure are given as overall performance indicator. All results were calculated using the TREC evaluation scripts, implemented in trec_eval version 7.3.

*MAP, Precision@10 and bpref.* Table 4 shows results for the best performing run of each participating group (selected based on MAP score). In total 19 topics were used for the official evaluation, however utrecht did not produce a ranking for two topics.

When comparing the number of retrieved elements (Ret.) with the number of relevant elements (Rel.) and the number of relevant and retrieved elements (Rel. & ret.), rmit was particularly successful in retrieving 202 relevant elements out

**Table 3.** Details of the obtained assessments per topic

| Topic | Relevant Elements | Topic | Relevant Elements |
|---|---|---|---|
| 1 | 29 | 14 | 44 |
| 2 | 75 | 15 | 18 |
| 3 | 13 | 16 | 40 |
| 4 | 13 | 17 | 10 |
| 5 | 4 | 18 | - |
| 6 | 8 | 19 | 20 |
| 7 | 10 | 20 | - |
| *8* | *5* | *21* | *25* |
| 9 | 31 | *22* | *21* |
| 10 | 50 | 23 | 4 |
| 11 | 2 | *24* | *77* |
| 12 | 11 | 25 | 2 |
| 13 | 64 | | |

of only 784 retrieved elements. However, the most number of relevant elements were retrieved by qutau: 303 out of a total of 448 relevant elements.

**Table 4.** MAP, Precision@10 and bpref

| **Participant** | **qmul** | **rmit** | **qutau** | **utrecht** | **utwente** |
|---|---|---|---|---|---|
| **Run** | **vsm aggr 06** | **alpha-0.3** | **text only** | **annotation** | **automatic noimg lm 05** |
| **Topics** | 19 | 19 | 19 | 17 | 19 |
| **Ret.** | 4750 | **784** | 3366 | 1112 | 4750 |
| **Rel.** | 448 | 448 | 448 | 390 | 448 |
| **Rel. & ret.** | 83 | 202 | **303** | 216 | 282 |
| **MAP** | 0.0412 | **0.2779** | 0.2711 | 0.2392 | 0.2751 |
| **Precision@10** | 0.0368 | **0.3105** | 0.2842 | 0.2706 | 0.2789 |
| **bpref** | 0.2388 | 0.4455 | **0.6516** | 0.5113 | 0.6272 |

MAP reports the average of the precision values after each retrieved relevant element. Based on the statistics of Table 4, it can be concluded that MAP favors rmit, closely followed by utwente and qutau.

Precision@10 reflects on the ability of a system to retrieve relevant elements in the top 10 positions of the ranking. For this measure, rmit has best performance, followed by qutau and utwente. This suggests that although qutau shows better performance at the top of the ranking, utwente achieves better overall performance.

Bpref has been shown to be a highly stable measure when relevance judgments are sparse [7]. This is therefore especially useful here given the size of the assessment pool. Intuitively, bpref measures the average number of times non-relevant material appears before relevant material. Here the differences between the runs are substantial: qutau and utwente are the only ones with a score of over 60%, while the others follow at some distance. Interestingly, however, bpref leads to a much improved performance score for qmul, compared with MAP and Precision@10.

More detailed information on the overall performances for all other runs submitted by participants can be found in Appendix A, or on the website [8].

*Interpolated Recall-Precision Averages.* Figure 2 shows the interpolated precision scores calculated at 11 standard recall points for the best performing run of each participating group (selected based on MAP score). Most curves are closely correlated. Two observations may be of interest: 1.) The best rmit run outperforms all others for mid-recall levels (0.1 – 0.6), and 2.) the best run for utrecht achieves the highest precision of all for low recall levels (<0.1).

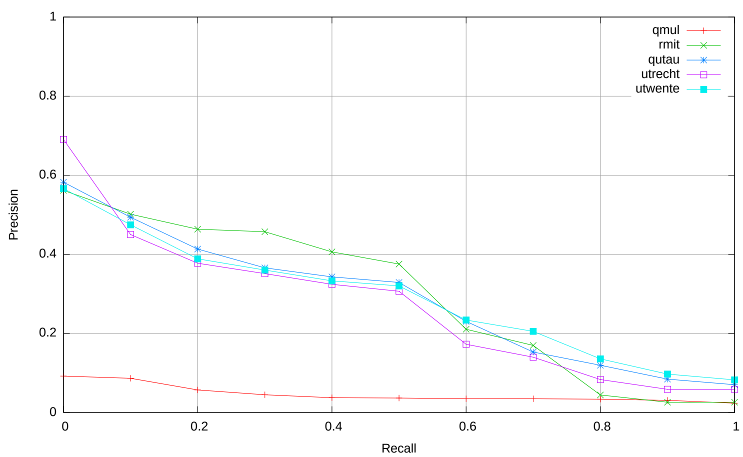


**Fig. 2.** Interpolated precision-recall averages @ 11 standard recall levels.

*Precision at Element Cutoff Levels.* Figure 3 plots scores for precision at various element cutoff levels in the range of 1 – 250. Precision at element cutoff is sensitive of the order in which relevant elements are returned. With the exception of the qmul run, all runs show an almost equal performance.

*Mean Average Precision per Topic.* The analysis presented here allows us to investigate the performance of the different strategies per topic. In addition, it reveals if there are topics, which could not be answered, or which require specific approaches.

In Figure 4, MAP per topic is plotted for the 19 topics. It shows that retrieval performance is poor for all strategies on topics 1, 9, 12, 15, and 27. Looking at Table 3, it becomes clear that the assessment pools for these topics are relatively

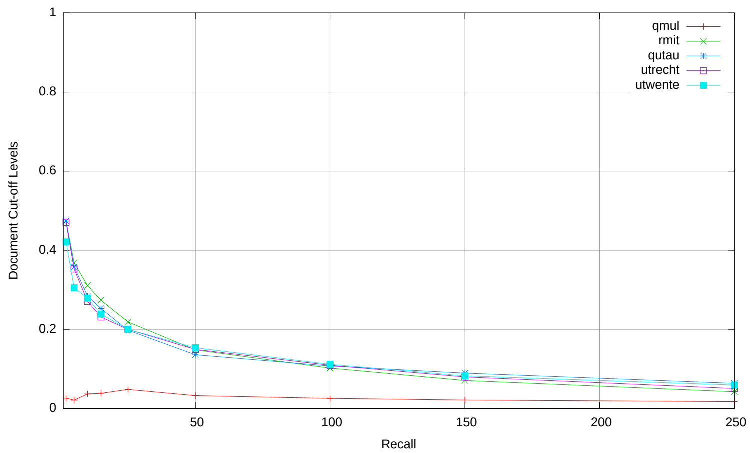


**Fig. 3.** Precision at element cutoff levels

small. For example, in the case of topic 25 none of the systems were able to find any of the 2 relevant elements. On the other hand, most systems were more successful on topic 11, which also has only 2 relevant elements. Obviously, topics with sparse relevant elements will have a larger impact on the overall evaluation. It would hence make sense to withdraw such topics from the evaluation.

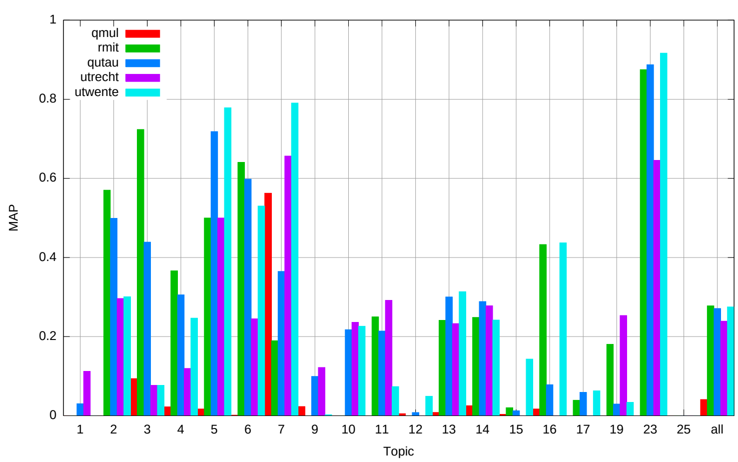


**Fig. 4.** Mean average precision per topic

Interesting to note that each participant seems to stand out on a number of topics, while performance on the other topics is about average. Furthermore, overall performance for most systems on topics 2 – 7 and 23 seems to be significantly higher than for topics 9 –19. A more detailed analysis on the characteristics of the topics is needed to examine what causes this difference in performance.

## 6 Discussions at Dagstuhl

At the workshop in Dagstuhl a lively discussion was held, during which the issue was raised that more expertise was needed on image retrieval approaches. It was a common problem that most of the participants did not have an 'off-the-shelf' image retrieval system available and ended up with a poor performing image retrieval strategy. Consequently, their multimedia information retrieval approach suffered, which explains why at current the multimedia-based approaches did not do convincingly better than a plain text-based search.

One of objectives for next year, is to arrange a state-of-the-art image retrieval system that can be used by participants. Ideally, image segmentation and object recognition should be supported by this system.

Another initiative, forthcoming from this discussion is to either extend the Lonely Planet collection with a larger collection of images[4], or to switch to the Wikipedia collection, which is also enriched with a large collection of images. This collection will also be used by the INEX 2006 ad-hoc track [9].

## 7 Conclusions and Future Work

A detailed analysis of the results for the multimedia track remains to be done. However, at this point we can conclude that, despite the exploratory nature of the first year, many achievements have been realized. We have successfully acquired and exploited the Lonely Planet WorldGuide, which proved to be a very useful starting point to flash out initial problems both in the retrieval approaches and the evaluation methodology. With a minimum extension, we successfully adopted the NEXI query language for *multimedia* structured document retrieval. A topic pool of 23 topics has been created and assessed. Five participating groups succeeded in building a multimedia retrieval system for structured documents and submitted a total of 25 runs for the evaluation.

A solid basis has been created to run the multimedia track again next year. We will have to reconsider many of the choices made, such as for instance the topic creation procedure and the evaluation metrics, which are currently based on the standard TREC methodology.


### Acknowledgments

We would like to thank all participants for their efforts in making the first year for the multimedia track a success. In particular, we are grateful for the support


[4] In this case, we will try to obtain access to the stock photography library of the Lonely Planet, which contains a large amount of images associated with a destination. See for more information: `http://www.lonelyplanetimages.com/`.

we received from Benjamin Piwowarski, who provided the relevance assessments tool. We would like to thank Andrew Trotman, among others, who have helped with the implementation of the NEXI extension. Finally, we would like to express our gratitude towards the Lonely Planet organization for providing us with this interesting XML collection.

## References


1. Voorhees, E., Harman, D., eds.: TREC - Experiment and Evaluation in Information Retrieval, MIT Press (2005)
2. Smeaton, A., Kraaij, W., Over, P.: The TREC Video Retrieval Evaluation (TRECVID): A case study and status report. In: Coupling approaches, coupling media and coupling languages for information retrieval - RIAO 2004, Vaucluse, France (2004) 25 – 37
3. Clough, P., Mueller, H., Sanderson, M.: The CLEF Cross Language Image Retrieval Track (ImageCLEF) 2004. In: Fifth Workshop of the Cross-Language Evaluation Forum (CLEF 2004). Lecture Notes in Computer Science, Springer (2005)
4. Planet, L.: (World guide - `http://www.lonelyplanet.com/worldguide/`)
5. Trotman, A., Sigurbjornsson, B.: Narrowed Extended XPath I (NEXI). In: Advances in XML Information Retrieval. Volume 3493. Springer, Schloss Dagstuhl, Germany (2005) 16 – 40
6. Lalmas, M., Piwowarski, B.: INEX 2005 relevance assessment guide. In: INEX 2005 Workshop Pre-proceedings. (2005) 391 – 401
7. Buckley, C., Voorhees, E.M.: Retrieval evaluation with incomplete information. In: SIGIR 2004: Proceedings of the 27th annual international ACM SIGIR conference on Research and Development in Information Retrieval, New York, NY, USA, ACM Press (2004) 25 – 32
8. van Zwol, R.: (INEX 2005 Multimedia Track Evaluation Results - `http;//contentlab.cs.uu.nl/~roelof/mmtrack/`)
9. INEX: (INitiative for the evaluation of XML Retrieval 2006 - `http://inex.is.informatik.uni-duisburg.de/2006/`)


## A Performance results

In this appendix the performance of the additional runs that were submitted by participants are reported. Performance results for best runs per group were reported in Section 5.

**Table 5.** Performance results for rmit

| Run | alpha-0.0 | alpha-0.1 | alpha-0.3 | alpha-0.5 | alpha-0.9 | alpha-1.0 |
|---|---|---|---|---|---|---|
| **Topics** | 19 | 19 | 19 | 19 | 19 | 19 |
| **Ret.** | 784 | 784 | 784 | 784 | 784 | 784 |
| **Rel.** | 448 | 448 | 448 | 448 | 448 | 448 |
| **Rel. & ret.** | 202 | 202 | 202 | 202 | 202 | 202 |
| **MAP** | 0.2759 | 0.2771 | **0.2779** | 0.2764 | 0.2664 | 0.2244 |
| **Precision@10** | 0.3053 | 0.3053 | **0.3105** | 0.3053 | 0.2579 | 0.2105 |
| **bpref** | 0.4455 | 0.4455 | 0.4455 | 0.4455 | 0.4455 | 0.4455 |

**Table 6.** Performance results for qutau

| Run | all Features 5 | all Features 15 | Text Only | all Features 15 10 | all Features 3 100 | global Features 10 100 |
|---|---|---|---|---|---|---|
| **Topics** | 19 | 19 | 19 | 19 | 19 | 19 |
| **Ret.** | 3767 | 3793 | 3366 | 3882 | 4009 | 4132 |
| **Rel.** | 448 | 448 | 448 | 448 | 448 | 448 |
| **Rel. & ret.** | 297 | 297 | **303** | 300 | 285 | 266 |
| **MAP** | 0.1995 | 0.2064 | **0.2711** | 0.1844 | 0.2037 | 0.2066 |
| **Precision@10** | 0.1947 | 0.2053 | **0.2842** | 0.1895 | 0.2105 | 0.2053 |
| **bpref** | 0.6507 | 0.6518 | 0.6516 | **0.6647** | 0.6501 | 0.6319 |

**Table 7.** Performance results for utrecht

| Run | Text | Annotation | PCA | PCA-cross |
|---|---|---|---|---|
| **Topics** | 13 | 17 | 17 | 17 |
| **Ret.** | 601 | 1112 | 1550 | 1291 |
| **Rel.** | 188 | 390 | 390 | 390 |
| **Rel. & ret.** | 88 | 216 | 216 | **220** |
| **MAP** | 0.2329 | **0.24** | 0.1769 | 0.2324 |
| **Precision@10** | 0.2615 | 0.27 | 0.2235 | **0.2824** |
| **bpref** | 0.4467 | 0.51 | 0.5041 | **0.5145** |

**Table 8.** Performance results for utwente

| Run | autom. noimg LM 05 | manual LM 05 | autom. noimg GPX 05 | autom. noimg Okapi 15 075 | manual GPX 05 | manual Okapi 15 075 |
|---|---|---|---|---|---|---|
| **Topics** | 19 | 19 | 18 | 19 | 18 | 19 |
| **Ret.** | 4750 | 4750 | 3142 | 4750 | 3826 | 4750 |
| **Rel.** | 448 | 448 | 408 | 448 | 408 | 448 |
| **Rel. & ret.** | **282** | 278 | 233 | 216 | 239 | 206 |
| **MAP** | **0.2751** | 0.26 | 0.2567 | 0.211 | 0.2627 | 0.2133 |
| **Precision@10** | 0.2789 | 0.2579 | 0.2667 | 0.2263 | **0.2833** | 0.2263 |
| **bpref** | **0.6272** | 0.6244 | 0.5532 | 0.475 | 0.5909 | 0.4853 |