# Social Media Retrieval using Image Features and Structured Text

D.N.F. Awang Iskandar, Jovan Pehcevski, James A. Thom, and S. M. M. Tahaghoghi

School of Computer Science and Information Technology, RMIT University
Melbourne, Australia
{dayang, jovanp, jat, saied}@cs.rmit.edu.au

**Abstract.** Use of XML offers a structured approach for representing information while maintaining separation of form and content. XML information retrieval is different from standard text retrieval in two aspects: the XML structure may be of interest as part of the query; and the information does not have to be text. In this paper, we describe an investigation of approaches to retrieve text and images from a large collection of XML documents, performed in the course of our participation in the INEX 2006 Ad Hoc and Multimedia tracks. We evaluate three information retrieval similarity measures: Pivoted Cosine, Okapi BM25 and Dirichlet. We show that on the INEX 2006 Ad Hoc queries Okapi BM25 is the most effective among the three similarity measures used for retrieving text only, while Dirichlet is more suitable when retrieving heterogeneous (text and image) data.

**Key words:** Content-based image retrieval, text-based information retrieval, social media, linear combination of evidence

## 1 Introduction

A structured document could contain text, images, audios and videos. Retrieving the desired information from an eXtensible Markup Language (XML) document involves retrieval of XML elements. This is not a trivial task as it may involve retrieving text and other multimedia elements.

The INitiative for the Evaluation of XML Retrieval (INEX) provides a platform for participants to evaluate the effectiveness of their XML retrieval techniques using uniform scoring procedures, and a forum to compare results. Of the nine tracks at INEX 2006, this paper presents the RMIT university group's participation in two tracks: the Ad Hoc track, where we investigate the effects of using different information retrieval (IR) similarity measures; and the Multimedia (MM) track, where we combine retrieval techniques based on text and image similarity.

There are four XML retrieval tasks within the INEX 2006 Ad Hoc track: *Thorough*, *Focused*, *All In Context (AIC)* and *Best In Context (BIC)*. Using three IR similarity measures — Pivoted Cosine, Okapi BM25, and Dirichlet — in this paper we focus on the results obtained under *Thorough* and *AIC* tasks. Since the system we used is a full-text IR system which only does retrieval at document level, we only expected it to perform well on article retrieval in the *AIC* task.

The objective of the INEX 2006 MM track is to exploit the XML structure that provides a logical level at which multimedia objects are connected and to improve the retrieval performance of an XML-driven multimedia information retrieval system.[1] Existing research on multimedia information retrieval from XML document collections is shown to be challenging [3, 12, 13]. For the *Multimedia Images (MMImages)* and *Multimedia Fragments (MMFragments)* tasks of the INEX 2006 MM track, we explore and analyse methods for combining evidence from content-based image retrieval (CBIR) with full-text IR. We describe a fusion system that combines evidence and ranks the query results based on text and image similarity. The fusion system consists of two subsystems: the GNU Image Finding Tool (GIFT), and the full-text IR system (Zettair). A technique for linear combination of evidence is used to merge the relevance scores from the two subsystems.

The retrieval strategy has been evaluated using Wikipedia, a social media collection that is an online encyclopedia. Social media describes the online tools and platforms that people use to share opinions, insights, experiences, and perspectives with each other. Social media can take many different forms, including text, images, audio, and video. Popular social mediums include blogs, message boards, podcasts, wikis, and vlogs.[2]

The remainder of this paper is organised as follows. Section 2 describes the text retrieval approach used for the Ad Hoc and MM tracks followed by the performance results obtained on the *Thorough* and *AIC* tasks of the Ad Hoc track. In Section 3 we present the INEX 2006 multimedia topics and their corresponding relevance judgements. In Section 4 we describe our approach to retrieve XML articles and the associated images based on the multimedia topics used in the MM track. In Section 5 we present results obtained from our experiments on the two tasks of the INEX 2006 MM track. We conclude in Section 6 with a discussion of our findings and outline directions for future work.

## 2 Full-Text Information Retrieval

In this section, we describe the three similarity measures implemented in Zettair, and show performance results on the *Thorough* and *AIC* tasks of the INEX 2006 Ad Hoc track.

### 2.1 The Zettair Search Engine

Zettair is a compact and fast text search engine developed by the Search Engine Group at RMIT University.[3] Zettair supports on-the-fly indexing and retrieval of large textual document collections. To process the queries for the INEX 2006 Ad Hoc and MM tracks, we first obtained the document content by extracting the plain document text (and by completely removing all the XML tags). We then indexed these documents using fast and efficient inverted index structure as implemented in many modern search

[1] INEX 2006 Multimedia Track Guidelines

[2] `http://en.wikipedia.org/wiki/Wiki`

[3] `http://www.seg.rmit.edu.au/zettair`

engines [14]. A similarity measure is used to rank documents by likely relevance to the query; in this work, we report on experiments using three different similarity measures implemented in Zettair, which respectively follow the three major models to information retrieval: the vector-space model, the probabilistic model, and the language model.

### 2.2 Similarity Measures

The *similarity* of a document to a query, denoted as $S_{q,d}$, indicates how closely the content of the document matches the query.

To calculate the query-document similarity, statistical information about the distribution of the query terms (within both the document and the collection as a whole) is often necessary. These term statistics are subsequently utilised by the similarity measure. Following the notation and definitions of Zobel and Moffat [16], we define the basic term statistics as:

- $q$, a query;
- $t$, a query term;
- $d$, a document;
- $N_{\mathscr{D}}$, the number of all the documents in the collection;
- For each term $t$:
  - $f_{d,t}$, the frequency of $t$ in the document $d$;
  - $N_{\mathscr{D}_t}$, the number of documents containing the term $t$; and
  - $f_{q,t}$, the frequency of $t$ in query $q$.
- For each document $d$:
  - $f_d = |d|$, the document length approximation.
- For the query $q$:
  - $f_q = |q|$, the query length.

We also denote the following sets:

- $\mathscr{D}$, the set of all the documents in the collection;
- $\mathscr{D}_t$, the set of documents containing term $t$;
- $\mathscr{T}_d$, the set of distinct terms in the document $d$;
- $\mathscr{T}_q$, the set of distinct terms in the query, and $\mathscr{T}_{q,d} = \mathscr{T}_q \cap \mathscr{T}_d$.

**Vector-Space Model** In this model, both the document and the query are representations of $n$-dimensional vectors, where $n$ is the number of distinct terms observed in the document collection. The best-known technique for computing similarity under the vector-space model is the cosine measure, where the similarity between a document and the query is computed as the cosine of the angle between their vectors.

Zettair uses pivoted cosine document length normalisation [8] to compute the query-document similarity under the vector-space model:

$$S_{q,d} = \frac{1}{W_D \times W_q} \times \sum_{t \in \mathscr{T}_{q,d}} \left(1 + \log_e f_{d,t}\right) \times \log_e \left(1 + \frac{N_{\mathscr{D}}}{N_{\mathscr{D}_t}}\right) \tag{1}$$

In Equation (1), $W_D = \left((1.0 - s) + s \times \frac{W_d}{W_{AL}}\right)$ represents the pivoted document length normalisation, and $W_q$ is the query length representation. The parameter $s$ represents the *slope*, whereas $W_d$ and $W_{AL}$ represent the document length (usually taken as $f_d$) and the average document length (over all documents in $\mathscr{D}$), respectively. We use the standard value of 0.2 for the slope, which is shown to work well in traditional IR experiments [8].

**Probabilistic Model** In IR, the probabilistic models are based on the principle that documents should be ranked by decreasing probability of their relevance to the expressed information need. Zettair uses the Okapi BM25 probabilistic model developed by Sparck Jones et al. [10]:

$$S_{q,d} = \sum_{t \in \mathscr{T}_{q,d}} w_t \times \frac{(k_1+1)\, f_{d,t}}{K + f_{d,t}} \times \frac{(k_3+1)\, f_{q,t}}{k_3 + f_{q,t}} \tag{2}$$

where $w_t = \log_e\left(\frac{N_{\mathscr{D}} - N_{\mathscr{D}_t} + 0.5}{N_{\mathscr{D}_t} + 0.5}\right)$ is a representation of inverse document frequency, $K = k_1 \times \left[(1-b) + \frac{b \cdot W_d}{W_{AL}}\right]$, and $k_1$, $b$ and $k_3$ are constants, in the range 1.2 to 1.5 (we use 1.2), 0.6 to 0.75 (we use 0.75), and 1 000 000 (effectively infinite), respectively. The chosen values for $k_1$, $b$ and $k_3$ are shown to work well with the TREC Collection experiments [10]. $W_d$ and $W_{AL}$ represent the document length and the average document length.

**Language Model** Language models are probability distributions that aim to capture the statistical regularities of natural language use. Language modelling in IR involves estimating the likelihood that both the document and the query could have been generated by the same language model. Zettair uses a query likelihood approach with Dirichlet smoothing [15]:

$$S_{q,d} = f_q \times \log \lambda_d + \sum_{t \in \mathscr{T}_{q,d}} \log\left(\frac{N_{\mathscr{D}} \times f_{d,t}}{\mu \times N_{\mathscr{D}_t}} + 1\right) \tag{3}$$

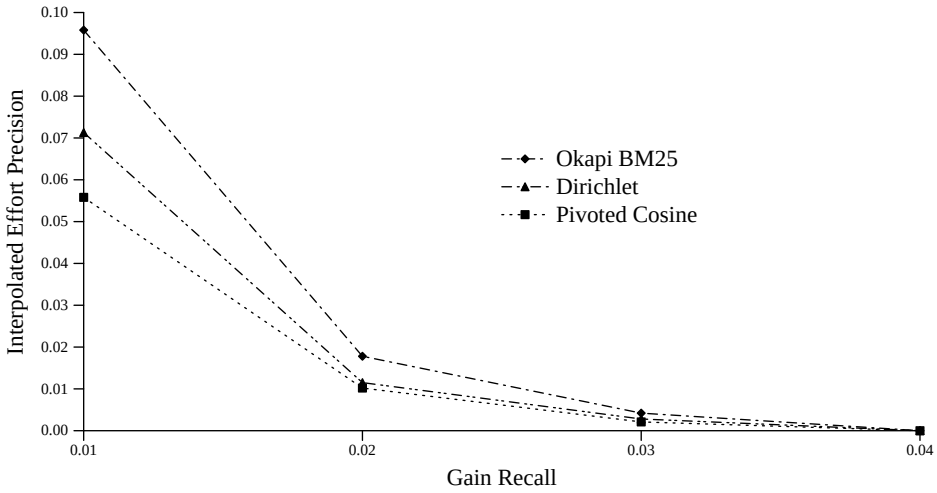
where $\mu$ is a smoothing parameter, while $\lambda_d$ is calculated as: $\lambda_d = \mu / (\mu + f_d)$. We use the value of 2 000 for $\mu$ as according to Zhai and Lafferty [15] it is the optimal value used in most IR experiments.

### 2.3 Performance Results

We now compare the performance of the three similarity measures implemented in Zettair for the *Thorough* and *AIC* tasks of the INEX 2006 Ad Hoc track.[4] We used the information in the `title` element of the topic as the query for Zettair.

The official measures of retrieval effectiveness for the INEX 2006 *Thorough* task are `ep/gr` and `MAep` of the XCG metrics family [4]. The `ep/gr` graphs provide a detailed

[4] Similar relative performance differences between the three similarity measures were also observed on the Focused task of the INEX 2006 Ad Hoc track.




| Similarity Measure | MAeP |
|---|---|
| Okapi BM25 | **0.0058** |
| Dirichlet | 0.0052 |
| Pivoted Cosine | 0.0047 |

**Fig. 1.** Retrieval performance of the three similarity measures implemented in Zettair on the *Thorough* task of the INEX 2006 Ad hoc track.

view of the run's performance at various gain-recall levels. The MAep measure provides a single-valued score for the overall run performance. This evaluation measure was also used for the *MMFragments* task evaluation of the INEX 2006 MM track. The measures make use of the *Specificity* relevance dimension, which is measured automatically on a continuous scale with values in the interval [0, 1]. A relevance value of 1 represents a fully specific component (that contains only relevant information), whereas a relevance value of 0 represents a non-relevant component (that contains no relevant information). Values of *Specificity* were derived on the basis of the ratio of relevant to both relevant and non-relevant text, as highlighted by the assessor.

Figure 1 shows the performance results obtained for the three similarity measures using both the MAep scores and the ep/gr graphs. We observe that Okapi BM25 produced the best MAep score among the three similarity measures, substantially outperforming the other two similarity measures. This performance difference is especially reflected on the ep/gp graphs. Of the other two measures, Dirichlet seems to perform better than the Pivoted Cosine measure. Interestingly, the ep/gp graphs generated on the *article-level* Fetch and Browse retrieval task of the INEX 2005 Ad hoc track show similar relative performance differences between the three similarity measures, even though a different XML document collection (IEEE instead of Wikipedia) was used as part of the evaluation testbed [6]. However, the Pivoted Cosine similarity measure out-

**Table 1.** Retrieval performance of the three similarity measures implemented in Zettair on the *AIC* task of the INEX 2006 Ad hoc track

| Similarity Measure | MAgP | gP[5] | gP[10] | gP[25] | gP[50] |
|---|---|---|---|---|---|
| Okapi BM25 | **0.1751** | **0.3766** | **0.3049** | **0.2220** | **0.1566** |
| Dirichlet | 0.1655 | 0.3266 | 0.2559 | 0.1844 | 0.1372 |
| Pivoted Cosine | 0.1489 | 0.3236 | 0.2611 | 0.1830 | 0.1301 |

performed the other two measures on the *element-level* Fetch and Browse retrieval task of the INEX 2005 Ad Hoc track. Compared to runs submitted by other participants in the INEX 2006 *Thorough* task, all three measures performed relatively poor as our system only returned whole articles (our run using the Okapi BM25 measure was ranked as 82 out of 106 submitted runs).

Table 1 shows that, when using the official evaluation measures for the INEX 2006 *AIC* task, Okapi BM25 again outperforms the other two similarity measures. With the MAgP measure, our run using the Okapi BM25 measure was ranked as fourth out of 56 submitted runs in the INEX 2006 *AIC* task. With the measures at rank cutoffs, this run was consistently ranked among the top five best performing runs in the INEX 2006 *AIC* task.

In the next section we describe our research activities carried out for the INEX 2006 MM track. We start with a description of the INEX 2006 MM tasks, along with their associated topics and their corresponding relevance judgements.

## 3 Multimedia Tasks, Topics and Relevance Judgements

The INEX 2006 MM topics were organised differently compared to the INEX 2005 MM topics. The INEX 2005 MM topics were only based on the *MMFragments* task, whereas the *MMImages* task was additionally introduced in the INEX 2006 MM track.

Since there are two tasks, the Wikipedia collection has been divided into two subcollections: Wikipedia Ad Hoc XML collection (Wikipedia), which contains XML documents as well as images; and Wikipedia image collection (Wikipedia_IMG), which contains 170 000 royalty free images. The *MMFragments* task utilises the Wikipedia collection and the Wikipedia_IMG is used for the *MMImages* task.

### 3.1 Multimedia Images Task

In the *MMImages* task, the participants were required to find relevant images in the articles based on the topic query. Hence, this task is basically using image retrieval techniques. Even though the target element is an image, the XML structure in the documents could be exploited to get to the relevant images. An example of an *MMImages* topic is depicted in Figure 2.

Each XML document in the Wikipedia_IMG collection contains an image. Therefore, the *MMImages* task essentially represents a document retrieval task, as the only results allowed were full documents (articles) from the XML image collection. The path of each of the resulting answers for this task were in the form of /article[1], so no document fragments are retrieved.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!DOCTYPE inex_topic SYSTEM "topic.dtd">
<inex_mm_topic topic_id="8"  ct_no="18"
task="MMimages">
<title>Images of bees with flowers.</title>
<castitle>//article[about(., src:60248)
and about(., bee)
and about (.,concept:animal)]
</castitle>
<description> Find images depicting a bee or
bees with flowers. </description>
<narrative>
Bees play an important role in pollinating flowering plants,
and are the major type of pollinators in ecosystems that contain
flowering plants. The flower's nectar is the primary source for energy,
and the pollen is primarily for protein and other nutrients. We are
looking for pictures depicting a bee or bees with flowers. We are not
interested in pictures of a basketball coach "Clair Bee" and album cover
for the Bee Gee's Stayin' Alive.
</narrative>
</inex_mm_topic>
```

**Fig. 2.** Example of a *MMImages* query with image ID 60248, a bee and a flower (original in colour)

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!DOCTYPE inex_topic SYSTEM "topic.dtd">
<inex_mm_topic topic_id="16"  ct_no="12" task="MMFragments">
<title>Kiwi shoe polish</title>
<castitle>
//article[about(.//history,kiwi shoe polish)]//image[about(., kiwi)]
</castitle>
<description>
Find images related to the Kiwi shoe polish product.
</description>
<narrative>Kiwi is the brand name of a shoe polish, first made in Australia
in 1906 and as of 2005 sold in almost 180 countries. Owned by the Sara Lee
Corporation since 1984, it is the dominant shoe polish in some countries,
including the United Kingdom and the United States, where it has about
two-thirds of the market. Find images related to the Kiwi shoe polish
product. We are not interested in the kiwi fruit.</narrative>
</inex_mm_topic>
```

**Fig. 3.** Example of a *MMFragments* query.

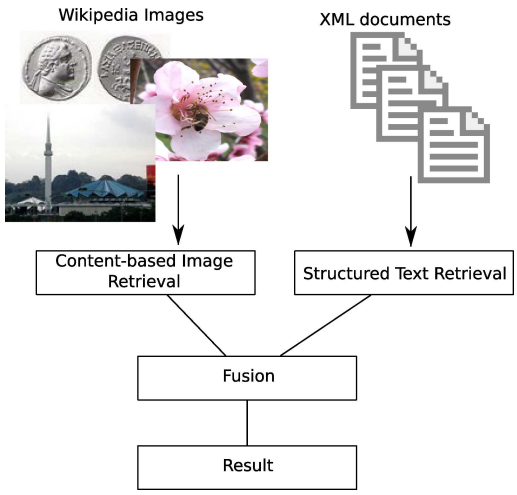



**Fig. 4.** Conceptual representation of the system (original in colour)

### 3.2 Multimedia Fragments Task

The objective of the *MMFragments* task is to find relevant XML fragments given an multimedia information need. Figure 3 illustrates a *MMFragments* topic. The target elements are ranked in relevance order and element overlapping is permitted.

## 4 Our Approach

In this section, we describe our approach adopted for the INEX 2006 MM track. We used two systems and fused the results from these systems to obtain the results for the multimedia queries. The overall structure of the system is depicted in Figure 4. Since the XML document structure serves as a semantic backbone for retrieval of the multimedia fragments, we use Zettair to retrieve the relevant articles. The GNU Image Finding Tool (GIFT),[5] a content-based image retrieval system, is used to retrieve the results based on the visual features of the images.

For INEX 2006 MM track, we adopted similar approach as the one we used in the INEX 2005 MM track [3]. The only difference is that we now use Zettair instead of the hybrid XML retrieval approach. With Zettair, our officially submitted runs used the Pivoted Cosine similarity measure as it performed best among the three similarity measures in the INEX 2005 Ad Hoc track (using the IEEE document collection) [6]. However, we also performed additional runs to examine the effect of using Okapi BM25 and Dirichlet in the two INEX 2006 MM tasks.

[5] `http://www.gnu.org/software/gift`

### 4.1 Content-Based Image Retrieval

The GNU Image Finding Tool was used to retrieve relevant images. The image features from the `Wikipedia_IMG` collection were extracted and indexed using an inverted file data structure.

Two main image features (colour and texture) were extracted during the indexing process. GIFT uses the HSV (Hue-Saturation-Value) colour space for local and global colour features [11]. For extracting the image texture, a bank of circularly symmetric Gabor filters is used. GIFT evaluates and calculates the query image and the target image feature similarity based on the data from the inverted file. The results of a query are presented to the user in the form of a ranked list.

For the multimedia topics, we used the image references listed in the source (`src`) element of the multimedia CAS query as the query image to GIFT. We used the default Classical IDF algorithm and set the search pruning option to 100%. This allows us to perform a complete feature evaluation for the query image, even though the query processing time is longer. For each query, we retrieved and ranked all the images in the `Wikipedia_IMG` collection. Referring to the multimedia topic presented earlier, the query image of Figure 2 is provided to GIFT.

### 4.2 Fusing and Ranking The Image and Text Retrieval

To fuse the two retrieval status value (RSV) lists into a single ranked result list $R$ for the multimedia queries, we use a simple linear combination of evidence [1] that is also a form of polyrepresentation [5]:

$$R = \begin{cases} \alpha \cdot S_I + (1-\alpha) \cdot S_T & \text{if the query contains image;} \\ S_T & \text{otherwise.} \end{cases}$$

Here, $\alpha$ is a weighting parameter (determines the weight of GIFT versus Zettair retrieval), $S_I$ represents the image RSV obtained from GIFT, and $S_T$ is the RSV of the same image obtained from the Zettair.

To investigate the effect of giving certain biases to a system, we vary the $\alpha$ values between 0 to 1. When the value of $\alpha$ is set to 1, only the RSVs from GIFT are used. On the other hand, only the Zettair's RSVs are used when the value of $\alpha$ is set to 0. If there was no image in the query then only the Zettair's RSVs are used, irrespective of the value of $\alpha$.

For the INEX 2006 MM track official runs, we submitted six runs with the $\alpha$ value set to 0.0, 0.5 and 1.0. We then conducted additional runs with the $\alpha$ values ranging between 0.0 to 0.5 to further investigate which $\alpha$ value produces the best retrieval performance. The fusion RSVs of the image and structured text retrieval are then ranked in a descending order of similarity.

## 5 Experiments and Results

The experiment for the runs was conducted by varying the $\alpha$ values and investigating the retrieval effectiveness of the three similarity measures in Zettair. For each multi-

media task, our runs were categorised into two types depending on which INEX 2006 multimedia topic elements were automatically translated as an input query to Zettair:

1. Title runs, which utilise the content of the `title` element; and
2. Extended runs, which utilise the content of the `title`, `castitle`, and `description` elements from each multimedia query.

### 5.1 Evaluation Metrics

The TREC evaluation metric was adopted to evaluate the *MMImages* task and the evaluation is based on the standard precision and recall retrieval performance measures:

- Mean Average Precision (`MAP`): The mean of the average precisions calculated for each topic. Average precision represents the average of the precisions calculated at each natural recall level.
- `bpref`: It computes a preference relation of whether judged relevant documents are retrieved ahead of judged irrelevant documents. Thus, it is based on the relative ranks of judged documents only.
- Average interpolated precision at 11 standard recall levels (0%-100%).

For the *MMFragments* task, the EvalJ evaluation software [6] was utilised. We used EvalJ with the following parameters: metrics (ep-gr), overlap (off), quantisation (gen), topic (ALL). The following evaluation measures were used:

- The effort-precision/gain-recall (`ep/gr`) graphs, which provide a detailed view of the run's performance at various gain-recall levels.
- Non-interpolated mean average effort-precision (`MAep`), which provides a single-valued score for the overall run performance. `MAep` is calculated as the average of effort-precision values measured at natural gain-recall levels.

### 5.2 Multimedia Images Task

For the *MMImages* task we conducted seven Title runs and three Extended runs using each of the IR similarity measures. We varied the $\alpha$ values between 0.0 and 1.0. As the results of the Title runs were promising, we applied a finer variation for the $\alpha$ values between the interval 0.0 and 0.5 (with the step of 0.1) to investigate the best $\alpha$ value for each similarity measure.

In Table 2, we observe that using the `title` element as the query produces better `MAP` and `bpref` performances compared to using the extended query for the *MMImages* task. Among the similarity measures, Dirichlet performed best, and this can be seen in Figure 5 that depicts the interpolated recall/precision averages for all the best runs for each similarity measure.

In the Title runs, having the $\alpha$ values between 0.1 and 0.4 yielded the best `MAP` and `bpref` performance. Using the text retrieval system alone produces better retrieval

[6] `http://evalj.sourceforge.net`

**Table 2.** Retrieval performance for the *MMImages* task: mean average precision (MAP) and bpref. *Italic* values – best performance runs using the various $\alpha$ values for each similarity measure **Bold** values – best overall performance among all runs

| Similarity Measure | $\alpha$ value | MAP | bpref |
|---|---|---|---|
| Title runs | | | |
| Pivoted Cosine | 0.0 | 0.3054 | 0.2861 |
| | 0.1 | *0.3153* | *0.2957* |
| | 0.2 | *0.3153* | *0.2957* |
| | 0.3 | 0.3152 | 0.2957 |
| | 0.4 | 0.3150 | 0.2956 |
| | 0.5 | 0.3071 | 0.2880 |
| | 1.0 | 0.2149 | 0.2033 |
| Okapi BM25 | 0.0 | 0.2679 | 0.2605 |
| | 0.1 | 0.2686 | 0.2622 |
| | 0.2 | *0.2700* | *0.2643* |
| | 0.3 | 0.2674 | 0.2599 |
| | 0.4 | 0.2660 | 0.2572 |
| | 0.5 | 0.2664 | 0.2592 |
| | 1.0 | 0.1909 | 0.1814 |
| Dirichlet | 0.0 | 0.3130 | 0.2973 |
| | 0.1 | 0.3175 | 0.3014 |
| | 0.2 | 0.3175 | 0.3014 |
| | 0.3 | **0.3203** | **0.3034** |
| | 0.4 | **0.3203** | **0.3034** |
| | 0.5 | 0.3202 | 0.3032 |
| | 1.0 | 0.2158 | 0.2080 |
| Extended runs | | | |
| Pivoted Cosine | 0.0 | 0.2608 | 0.2307 |
| | 0.5 | *0.2642* | *0.2366* |
| | 1.0 | 0.2071 | 0.1926 |
| Okapi BM25 | 0.0 | *0.2674* | 0.2369 |
| | 0.5 | *0.2674* | *0.2464* |
| | 1.0 | 0.2087 | 0.2002 |
| Dirichlet | 0.0 | 0.2988 | 0.2787 |
| | 0.5 | *0.3094* | *0.2805* |
| | 1.0 | 0.2147 | 0.1987 |

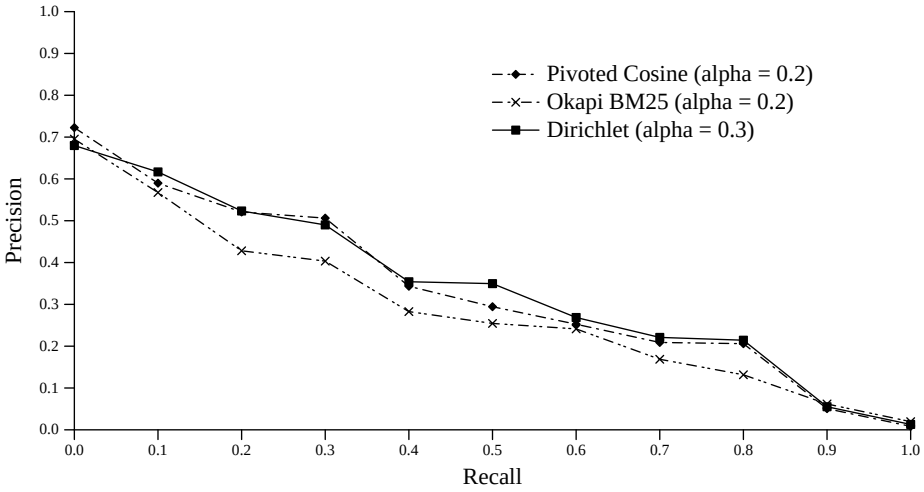



**Fig. 5.** Interpolated precision averages at eleven standard recall levels for the Title runs of the *MMImages* task.

performance compared to using only the content-based image retrieval system; however the best performance is found when combining evidence and weighting the text retrieval as more important than the content based image retrieval. Comparing the performances between INEX 2005 and INEX 2006 MM track, we observed a similar trend in the $\alpha$ values, where the best $\alpha$ values were in the same range.

Overall, using Dirichlet as the similarity measure produces the best retrieval performance compared to Pivoted Cosine and Okapi BM25 for the *MMImages* task. We also found that the Extended runs performed worse in most cases when compared to the Title runs.

### 5.3 Multimedia Fragments Task

For the *MMFragments* task, we conducted six runs using the Pivoted Cosine, Okapi BM25 and Dirichlet similarity measures. We used the default value of $\alpha = 0.0$ for all the runs (since for this task we only used the text retrieval system).

We observe an opposite behaviour for the Title and Extended runs for this task. As reflected from the `ep/gr` graphs in Figure 6, the Extended runs performed better than the Title runs. This shows that the presence of the `title`, `castitle` and `description` from the query improves the retrieval performance when compared to only using the `title` element of the MM queries in the *MMFragments* task. This result also reflects the nature of the task, where XML fragments need to be returned as the retrieved answers.

When comparing the retrieval performance of the three IR similarity measures, we observe that Dirichlet once again outperformed Pivoted Cosine and Okapi BM25. This can also be seen in Figure 6 and from the overall `MAep` scores presented in Table 3.

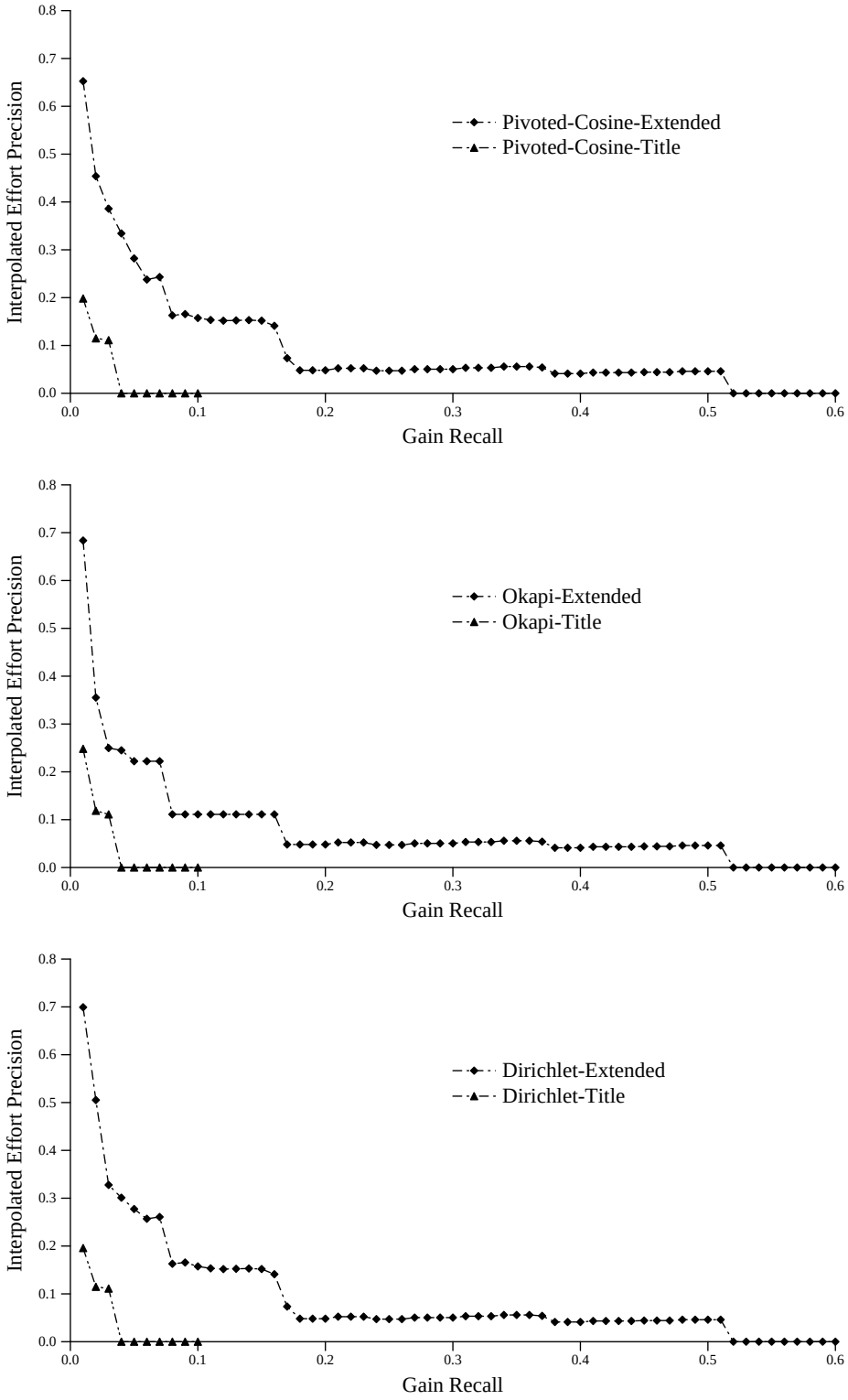



**Fig. 6.** Interpolated effort-precision averages at standard gain-recall levels for the Title and Extended runs of the *MMFragments* task, using Pivoted Cosine (top), Okapi BM25 (middle), and Dirichlet (bottom) similarity measures in Zettair.

**Table 3.** Retrieval performance of the Extended runs on the *MMFragments* task

| **Similarity Measure** | MAep |
|---|---|
| Pivoted Cosine | 0.0655 |
| Okapi BM25 | 0.0586 |
| Dirichlet | **0.0663** |

To investigate whether combining evidence from the CBIR system improves the retrieval performance for this task, we conducted several preliminary runs that fuse the RSVs from the CBIR system and Zettair. This resulted in a minor performance improvement. However, without better fragment retrieval system, we cannot conclude whether combining text and image RSVs will improve retrieval performance for the *MMFragments* task.

## 6 Conclusions and Future Work

In this paper we have reported on our participation in the Ad Hoc and MM tracks of INEX 2006. We utilised a full-text information retrieval system for both tracks to retrieve the XML documents and combined this with a content-based image retrieval system for the MM track.

For the Ad Hoc track, Okapi BM25 similarity measure produced the best retrieval performance for the *Thorough* and *AIC* tasks.

For the two XML multimedia retrieval tasks, we officially submitted six runs using the Pivoted Cosine similarity measure. We also conducted additional runs to investigate the effectiveness of the Okapi BM25 and Dirichlet similarity measures. The runs for the *MMImages* task reflect the various relative weights of 0.0 to 1.0 for the $\alpha$ values. We found that Dirichlet was the best similarity measure for the *MMImages* task, and that $\alpha$ values between 0.1 and 0.4 produced the best retrieval performance. For the *MMFragments* task, the official runs were only based on the text retrieval system. We executed four additional runs using Okapi BM25 and Dirichlet similarity measures. As for the *MMImages* task, Dirichlet was also found to be the best among the three similarity measures used in the *MMFragments* task.

We have used the linear combination of evidence to merge the RSVs from two retrieval subsystems for retrieving multimedia information. We conclude that a text retrieval system benefits by using some evidence from a CBIR system. More specifically, giving more weight to text retrieval system RSVs in the fusion function yields better performance than when the two subsystems are used on their own.

This work can be extended in two ways. First, to cater for the *MMFragments* task more effectively, the hybrid XML retrieval approach [7] can be used as the content-oriented XML retrieval system. Second, it would also be interesting to fuse the RSVs from CBIR and text systems with the 101 image concepts such as those provided by the University of Amsterdam [9].

**Acknowledgments** This research was undertaken using facilities supported by the Australian Research Council, an RMIT VRII grant, and a scholarship provided by the Malaysian Ministry of Higher Education.

## References

1. Y. A. Aslandogan and C. T. Yu. Evaluating strategies and systems for content-based indexing of person images on the web. In *MULTIMEDIA 2000: Proceedings of the Eighth ACM International Conference on Multimedia*, pages 313–321, New York, NY, USA, 2000. ACM Press.
2. N. Fuhr, M. Lalmas, S. Malik and G. Kazai (editors). *Advances in XML Information Retrieval and Evaluation, 4th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2005, Dagstuhl Castle, Germany, November 28-30, 2005, Revised Selected Papers*, Volume 3977 of *Lecture Notes in Computer Science*. Springer, 2006.
3. D. N. F. Awang Iskandar, J. Pehcevski, J. A. Thom and S. M. M. Tahaghoghi. Combining image and structured text retrieval. In Fuhr et al. [2], pages 525–539.
4. G. Kazai and M. Lalmas. INEX 2005 evaluation measures. In Fuhr et al. [2], pages 16–29.
5. B. Larsen, P. Ingwersen and J. Kekäläinen. The polyrepresentation continuum in IR. In *IIiX: Proceedings of the 1st international conference on Information interaction in context*, pages 88–96, New York, NY, USA, 2006. ACM Press.
6. J. Pehcevski, J. A. Thom and S.M. M. Tahaghoghi. RMIT University at INEX 2005: Ad Hoc Track. In Fuhr et al. [2], pages 306–320.
7. J. Pehcevski, J. A. Thom and A-M. Vercoustre. Hybrid XML retrieval: Combining information retrieval and a native XML database. *Information Retrieval*, Volume 8, Number 4, pages 571–600, 2005.
8. A. Singhal, C. Buckley and M. Mitra. Pivoted document length normalization. In *Proceedings of the ACM-SIGIR International Conference on Research and Development in Information Retrieval*, pages 21–29, Zurich, Switzerland, 1996. ACM Press.
9. C. G. M. Snoek, M. Worring, J. C. V. Gemert, J. Geusebroek and A. W. M. Smeulders. The challenge problem for automated detection of 101 semantic concepts in multimedia. In *MULTIMEDIA '06: Proceedings of the 14th annual ACM international conference on Multimedia*, pages 421–430, New York, NY, USA, 2006. ACM Press.
10. K. Sparck Jones, S. Walker and S. E. Robertson. A probabilistic model of information retrieval: Development and comparative experiments. Parts 1 and 2. *Information Processing and Management*, Volume 36, Number 6, pages 779–840, 2000.
11. D. M. Squire, W. Müller, H. Müller and T. Pun. Content-based query of image databases: Inspirations from text retrieval. *Pattern Recognition Letters*, Volume 21, Number 13–14, pages 1193–1198, 2000. (special edition for SCIA'99).
12. D. Tjondronegoro, J. Zhang, J. Gu, A. Nguyen and S. Geva. Integrating text retrieval and image retrieval in XML document searching. In Fuhr et al. [2], pages 511–524.
13. R. van Zwol. Multimedia strategies for $b^3$-sdr, based on principal component analysis. In Fuhr et al. [2], pages 540–553.
14. I. H. Witten, A. Moffat and T. C. Bell. *Managing Gigabytes: Compressing and Indexing Documents and Images, Second Edition*. Morgan Kaufmann Publishers, 1999.
15. C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems*, Volume 22, Number 2, pages 179–214, 2004.
16. J. Zobel and A. Moffat. Exploring the similarity space. *ACM SIGIR Forum*, Volume 32, Number 1, pages 18–34, 1998.