# IIIT-H at ImageCLEF Wikipedia MM 2009

Srinivasarao Vundavalli
International Institute of Information Technology,
Hyderabad-500032,
Andhra Pradesh, India
srinivasarao@research.iiit.ac.in

**Abstract**

In this paper, we describe the IIITH retrieval system used for the ImageCLEF Wikipedia MM task. The system automatically ranks the most similar images to a given textual query. The system preprocesses the data set in order to remove the non-informative terms. For each query, the system finds a ranked list of its most similar images using the textual information only.

### Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.3.7 Digital Libraries

### Keywords

TF-IDF, Information Retrieval, Image Retrieval, Vector Space Model, Boolean Model

## 1 Introduction

ImageCLEFs WikipediaMM uses image collection created and employed by the INEX competition [1]. This image collection contains approximately 150,000 images that cover diverse topics of interest. These images are associated with unstructured and noisy textual annotations in English. There are 75 topics as the query to be searched on the collection. Each topic consists of textual data (and/or sample images and/or concepts) describing a users (multimedia) information need. These data are given as the XML files in each one there is image textual information such as image title, image concept, image location path and a short text which describes the desired results, i.e it explains what could be the relevant result and what is the irrelevant results. The aim of this task is to try to find as many relevant images as possible from the Wikipedia image collection for a given textual-and/or-visual query.

Section 2 describes the models we used in our experiments, in section 3 we show the results we obtained and finally section 4 concludes the notes.

## 2 Models

Our system uses a combination of the Vector Space Model (VSM) of Information Retrieval and the Boolean model to determine how relevant a given Document is to a User's query. In general, the idea behind the VSM is the more times a query term appears in a document relative to the number of times the term appears in all the documents in the collection, the more relevant that document is to the query. It uses the Boolean model to first narrow down the documents that need to be scored based on the use of boolean logic in the Query specification.

The score is calculated based on TF-IDF model[2].

### 2.1 Term Frequency-Inverse Document Frequency Model

In the TF-IDF model [2], each document in the collection and a query are represented by their associated vector of the length of the vocabulary.

$tf(t \text{ in } d)$ correlates to the term's frequency, defined as the number of times term $t$ appears in the currently scored document $d$. Documents that have more occurrences of a given term receive a higher score. The default computation of $tf(t \text{ in } d)$ in our system is $sq.rt(frequency)$.

$idf(t)$ stands for Inverse Document Frequency. This value correlates to the inverse of $docFreq$ (the number of documents in which the term $t$ appears). This means rarer terms give higher contribution to the total score. The default computation of $idf(t)$ in our system is $1+log(numDocs/(docFreq+1))$

### 2.2 Vector Space Model(VSM)

In a Vector Space Model[3,4], a document is represented as a vector. Each dimension corresponds to a separate term. If a term occurs in the document, its value in the vector is non-zero. Several different ways of computing these values, also known as (term) weights, have been developed. One of the best known schemes is tf-idf weighting[2].

The score of query q for document d correlates to the cosine-distance or dot-product between document and query vectors. A document whose vector is closer to the query vector in that model is scored higher. The scoring we used was the lucene[5] scoring.

## 3 Runs and Results

By using the image's filename and the text associated with the image, we submitted one run based on the models described in the previous section.

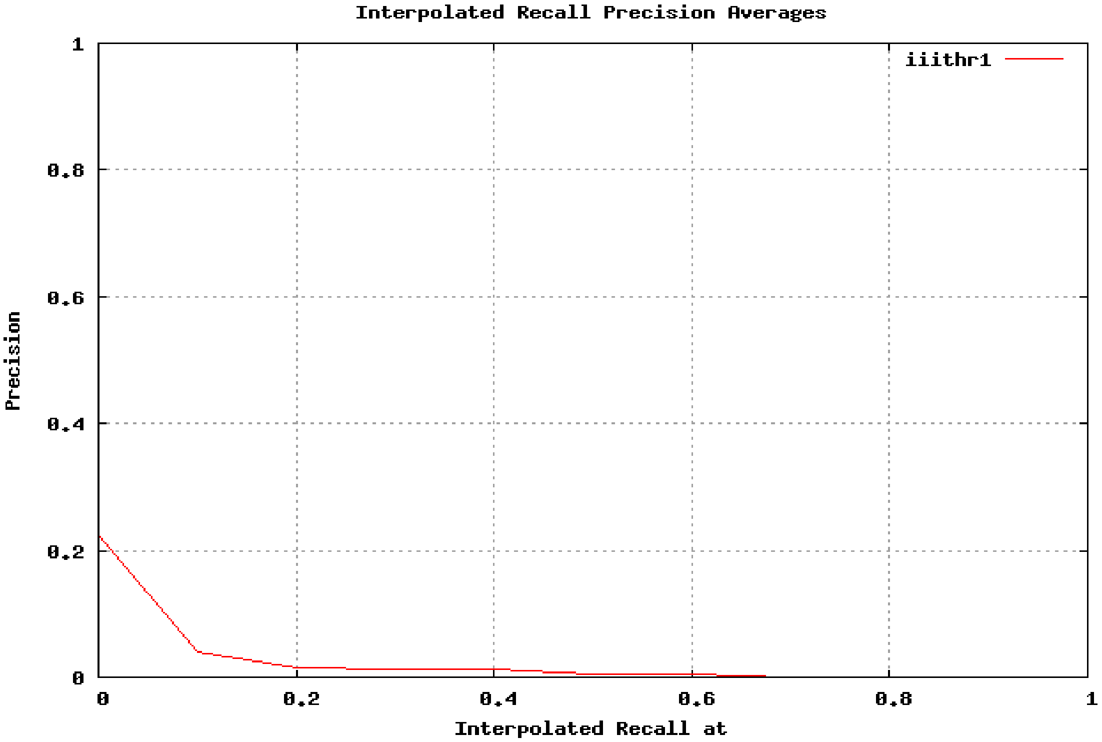
The results for the wikipediaMM task have been computed with the trec_eval tool (version 8.1). The submitted runs have been corrected (where necessary) so as to correpond to valid runs in the correct TREC format. The following corrections have been made:

- The runs comply with the TREC fomat as specified in the submission guidelines for the task
- When a topic contains an image example that is part of the wikipediaMM collection, this image is removed from the retrieval results, i.e., we are seeking relevant images that the users are not familiar with (as they are with the images they provided as examples).
- When an image is retrieved more than once for a given topic, only its highest ranking for that topic is kept and the rest are removed (and the ranks in the retrieval results are appropriately fixed).
- Ensure that each of the submitted runs has a unique name.

The interpolated recall precision averages are shown in Figure 1, and the summary statistics for the runs sorted by MAP are shown in Table 1.

| Run | Modality | Topic Fields | FB/QE | MAP | P@10 | P@20 | R-prec. | Bpref | Number of retrieved documents | Number of relevat retrieved documents |
|---|---|---|---|---|---|---|---|---|---|---|
| iiithr1 | TXT | TITLE | NOFB | 0.0186 | 0.0689 | 0.0389 | 0.0246 | 0.022 | 618 | 98 |

Table 1: Retrieval Results



## 4 Conclusion

In this working note, we presented a retrieval system which automatically ranks the most similar images to a given textual query. The system preprocesses the data set in order to remove the non-informative terms. For each query, the system finds a ranked list of its most similar images using the textual information only. Even though our system did not perform well, we are happy that we participated in the ImageCLEF WikipediaMM task and are looking forward to ImageCLEF2009.

## 5 References

1. http://inex.is.informatik.uni-duisburg.de/2007/mmtrack.html

2. Salton, Gerard and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Information Processing and Management 24 (5): 513-523.

3. G. Salton, A. Wong, and C. S. Yang (1975), "A Vector Space Model for Automatic Indexing". Communications of the ACM, vol. 18, nr. 11, pages 613620.

4. F.Song and W.B.Croft (1999). A General Language Model for Information Retrieval. Research and Development in Information Retrieval: 279-280.

5. http://lucene.apache.org