The Xtrieval Framework at CLEF 2008: ImageCLEF Wikipedia MM task

Thomas Wilhelm, Jens Kürsten, and Maximilian Eibl
Chemnitz University of Technology
Faculty of Computer Science, Dept. Computer Science and Media
09107 Chemnitz, Germany
[thomas.wilhelm | jens.kuersten | maximilian.eibl] at cs.tu-chemnitz.de

Abstract. This paper describes our participation at the ImageCLEF Wikipedia MM task. We used our Xtrieval framework for the preparation and execution of the experiments. We submitted 4 experiments in total. The results of these experiments were mixed. The text-only experiment scored second best with a mean average precision (MAP) of 0.2166. In combination with image based features the MAP dropped to 0.2138. With the addition of our thesaurus based query expansion it scored best with a MAP of 0.2195. Without query expansion and with the inclusion of the provided concepts the lowest MAP of 0.2048 was achieved, but there were 23 more relevant documents retrieved than in all 3 other experiments. Furthermore, the retrieval speed and comparison operations for vectors could be speeded up by implementing an interface to the PostgreSQL database.

Categories and Subject Descriptors

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

Keywords

Evaluation, Cross-Language Information Retrieval, Content-based Image Retrieval, Query Expansion, Experimentation

1 Introduction

For the ImageCLEF 2008: Wikipedia MM task a large set of data is supplied. Additional to the images and texts a 120 dimensional feature vector and concepts were available. This encouraged us to find a solution for vector storage, retrieval and comparison. Therefore we implemented an interface to index and search data into a PostgreSQL database. Additional to indexing the supplied feature vector we also indexed the MPEG-7 descriptors computed by Caliph & Emir.

2 Experiment Setup

We used our base system from the last year (see [1], [2] and [3]) with the following setup: Apache Lucene, customized analyzer with positional stopword[1] removal and Snowball stemmer[2]. For the content-based image retrieval we used Caliph & Emir as described below.

For query expansion a thesaurus was used. The parameters of the last year were further tuned to reduce unfitting synonyms. As source for the thesauri we still use OpenOffice.org[3].

[1] http://members.unine.ch/jacques.savoy/clef/index.html
[2] http://snowball.tartarus.org/
[3] http://wiki.services.openoffice.org/wiki/Dictionaries

The MPEG-7 features were calculated by Caliph & Emir (see [4]). Contrary to our experiments of the last year the MPEG-7 descriptors were not stored as text representations in Lucene[4], but as vectors in a PostgreSQL[5] database. PostgreSQL was chosen because it supports arrays as data types. In fact it is not necessary to know the actual size of the arrays at design time. This approach is expected to achieve a much higher retrieval speed and it is possible to use descriptors of Caliph & Emir which has no string representation implemented (e.g. the dominant color descriptor). The supplied 120 dimensional vectors are stored in the PostgreSQL database as well.

The computations of the distance measures were externalized into the PostgreSQL database by implementing the algorithms as stored procedures in PL/pgSQL[6]. PL/pgSQL is an internal programming language of PostgreSQL which adds support for additional logic to SQL such as control structures. The following algorithms are implemented so far: cosine similarity, Dice coefficient, Euclidean metric, intersection, Jaccard similarity coefficient. The main advantage is the reduction of extra round trips between our application and the database server. On the other hand the speed could be reduced by the fact that PL/pgSQL is an interpreted language.

All topics were preprocessed ad-hoc to retrieve all needed resources to perform the experiments. Especially the example images were retrieved and analyzed in advance.

3 Results

The results in table 1 show that the text-only results can only be marginally improved by additional data. Just experiment “cut-mix-concepts” retrieved more relevant documents in total, but with a lower MAP than all other experiments.

Table 1. Retrieval results

run	type	modality	feedback/ expansion	MAP	retrieved documents	relevant retrieved documents
cut-txt-a	auto	txt	nofb	0.2166	52623	3111
cut-mix	auto	txtimg	nofb	0.2138	52623	3111
cut-mix-qe	auto	txtimg	qe	0.2195	52623	3111
cut-mix-concepts	auto	txtimgcon	nofb	0.2048	70803	3134

Our baseline is “cut-txt-a”, which retrieved a total of 3111 relevant documents and reached a mean average precision (MAP) of 0.2166. By adding the content-based image features, which consist of the 120 dimensional feature vector and four MPEG-7 descriptors (scalable color, edge histogram, color layout and dominant color descriptor), the MAP decreased to 0.2138. This is a hint towards the low visual similarity between relevant pictures, which attract our attention during the relevance assessment process. After the preprocessing of the topics with our query expansion our highest MAP of 0.2195 was achieved. The inclusion of the concepts scored the worst MAP of 0.2048, but retrieved 23 more relevant documents than any other of our experiments.

4 Future Work

The PostgreSQL database support many other programming languages to implement stored procedures, i.e. PL/Tcl, PL/Perl, PL/Python and PL/Java[7]. Because the retrieval system itself is written in Java it would be suitable to use PL/Java as programming language.

[4] http://lucene.apache.org
[5] http://www.postgresql.org
[6] http://www.postgresql.org/docs/current/static/plpgsql.html
[7] http://pgfoundry.org/projects/pljava/

We also intend to implement our own content-based image retrieval algorithms on the basis of vectors stored in a PostgreSQL database.

References

[1] T. Wilhelm, J. Kürsten, and M. Eibl, “Experiments for the ImageCLEF 2007 Photographic Retrieval Task”; http://clef-campaign.org/2007/working_notes/wilhelmCLEF2007.pdf.

[2] J. Kürsten, T. Wilhelm, and M. Eibl, “The xtrieval framework at clef 2007: Domain-specific track,” *LNCS - Advances in Multilingual and Multimodal Information Retrieval*, C. Peters et al., ed., Berlin: Springer Verlag, 2008.

[3] T. Wilhelm, J. Kürsten, and M. Eibl, “Extensible retrieval and evaluation framework: Xtrieval,” *LWA 2008: Lernen - Wissen - Adaption*, Würzburg: 2008.

[4] M. Lux, W. Klieber, and M. Granitzer, “Caliph & Emir: Semantics in Multimedia Retrieval and Annotation,” *19th International CODATA Conference*, 2004.