CWI at ImageCLEF 2008

Theodora Tsikrika and Henning Rode and Arjen P. de Vries
CWI, Amsterdam, The Netherlands
{Theodora.Tsikrika, Henning.Rode, Arjen.de.Vries}@cwi.nl

Abstract

CWI used PF/Tijah, a flexible XML retrieval system, to evaluate image retrieval based on textual evidence in the context of the wikipediaMM task at ImageCLEF 2008. We employed a language modelling framework and found that the text associated with the Wikipedia images is a good source of evidence. We also investigated a length prior and found that biasing towards images with longer descriptions than the ones retrieved by our language modelling approach is not beneficial.

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; H.2.3 [**Database Managment**]: Languages—*Query Languages*

General Terms

Measurement, Performance, Experimentation

Keywords

Image retrieval, language models, textual evidence, length prior

1 Introduction

CWI participated in the wikipediaMM task at ImageCLEF 2008, where we examined the value of textual evidence for the retrieval of Wikipedia images associated with sparse and noisy English text formatted in XML. Our main objective was to establish a strong text-based baseline against which we will be able to compare the cross-media retrieval approaches currently being developed in the context of our research activities in the VITALAS[1] project. To this end, we employed a state-of-the-art language modelling approach that is implemented by PF/Tijah [5], a flexible XML retrieval system.

The remainder of this paper is organised as follows. Sections 2 and 3 provide background by introducing PF/Tijah and presenting the employed language modelling framework, respectively. Section 4 discusses our participation in the wikipediaMM task. Section 5 concludes this paper.

2 The PF/Tijah System

PF/Tijah, a research project run by the University of Twente, aims at creating a flexible environment for setting up search systems. It achieves that by including out-of-the-box solutions for common retrieval tasks, such as index creation (that also supports stemming and stopword

[1] http://vitalas.ercim.org/

removal) and retrieval in response to structured queries (where the ranking can be generated according to any of several retrieval models). Moreover, it maintains its versatility by being open to adaptations and extensions.

PF/Tijah is part of the open source release of MonetDB/XQuery[2], which is being developed in cooperation with CWI, Amsterdam and the University of München. PF/Tijah combines database and Information Retrieval (IR) technologies by integrating the PathFinder (PF) XQuery[3] compiler [1] with the Tijah XML IR system [7]. This provides PF/Tijah with a number of unique features that distinguish it from most other open source IR systems:

- It supports retrieval of arbitrary parts of XML documents, so that a query can simply ask for any XML tag-name as the unit of retrieval.
- It allows complex scoring and ranking of the retrieved results by directly supporting the NEXI query language [12, 13].
- It embeds NEXI queries as functions in the XQuery language, leading to ad hoc result presentation by means of its query language.
- It supports text search combined with traditional database querying.

Information on PF/Tijah can be found at: `http://dbappl.cs.utwente.nl/pftijah/`.

3 Language models

Generative language models, also known as statistical language models and commonly referred to as *language models*, estimate the probability distribution over all possible linguistic units of word sequences by applying statistical estimation techniques. In the language modelling approach to IR [10, 4, 9], a language model φ_D is inferred for each document D in the collection, i.e., the parameters of this language model are estimated from the document's text. Then, this language model is used for estimating probabilities of samples such as a query. Given a query Q, the ranking of the documents in the collection is produced by estimating the likelihood of the query (i.e., the probability of *generating* the query) $P(Q|\varphi_D)$, given φ_D the estimated language model for each document. For simplicity, we denote the query likelihood as $P(Q|D)$.

Here, we consider that the text associated with each image corresponds to a textual document and that queries consist only of a textual part. Documents are modelled using a multinomial distribution, and queries and documents are represented as sequences of terms [4]. With a query represented as a sequence of k random variables each corresponding to a term, its likelihood is:

$$P(\mathbf{q}|D) = P(q_1, q_2, \ldots, q_k|D) = \prod_{i=1}^{k} P(q_i|D) \tag{1}$$

assuming that each q_i is generated independently from the previous ones given the document model. The language model is thus reduced to modelling the distribution of each single term.

The simplest estimation strategy for an individual term probability is the *maximum likelihood estimate* (*mle*). This corresponds to the relative frequency of a term t_i in a specific document d $P_{mle}(t_i|d) = \frac{tf_{i,d}}{\sum_t tf_{t,d}}$, where $tf_{i,d}$, the term frequency of term t_i in document d, is normalised by the document's length (the sum of the term frequencies of all of its terms). However, this estimate is not suitable for IR, since it will assign zero query likelihood probabilities to documents missing even a single query term. This sparse estimation problem is addressed by *smoothing* techniques, e.g., the Jelinek-Mercer smoothing, which is a *mixture model* (a linear interpolation) of the document model with a background model (the collection model in this case):

$$P(\mathbf{q}|D) = \prod_{i=1}^{k} (1-\lambda) P_{mle}(q_i|D) + \lambda P_{mle}(q_i|C) \tag{2}$$

[2] `http://www.sourceforge.net/projects/monetdb/`
[3] `http://www.w3.org/TR/xquery/`

where $P_{mle}(t_i|C) = \frac{df_i}{\sum_t df_t}$, df_i the document frequency of term t_i in the collection.

Such language models can also be used to rank the documents in the collection by the posterior probability of a document being relevant to a query:

$$P(D|Q) \propto P(D)P(Q|D) \tag{3}$$

where $P(D)$ is the prior probability of the document being relevant, typically estimated given the document's query-independent features. To estimate these priors, two approaches can be distinguished [6]: (i) direct estimation on some training data, and (ii) definition based on some general modelling assumptions. For direct estimation of the probability of relevance of a document given its feature $\mathbf{f}$, we can use the distribution of the feature in the relevant set and in the collection:

$$P(R|D_f) = \frac{P(D_f|R)P(R)}{P(D_f)} \propto \frac{P(D_f|R)}{P(f)} = \frac{\#(rel, f)}{\#(f)} \tag{4}$$

where $\#(rel, f)$ is defined as the number of relevant documents with feature $\mathbf{f}$ and $\#(f)$ as the total number of documents with that feature [18]. This estimate can then be used directly in Equation 3. Alternatively, one could make the general modelling assumption (with or without training data) that the a-priori probability of relevance is taken to be a function of that feature, e.g., a linear function when the feature under consideration is the document's length:

$$P(R|D_f) \propto C \times doclength(D) \tag{5}$$

where C a constant that can be ignored in the ranking. In previous research, CWI has investigated the application of such generative probabilistic models both to image and to video retrieval [16, 17, 18, 8, 14].

4 WikipediaMM task

We participated in the wikipediaMM task where we used PF/Tijah for indexing and retrieval. Each image was represented by its accompanying textual description in English. Stopword removal and stemming were applied.

4.1 Runs

By using the topics' title only field, we submitted two runs based on the language model in Equation 2, with $\lambda = 0.8$ in both cases (a standard value for many retrieval tasks [14]). Our first run **cwi_lm_txt** simply applied this smoothed language model. Our second run **cwi_lm_lprior_txt** used Equation 3 to produce the ranking by incorporating a prior based on a linear function of length, so as to bias retrieval towards images with richer descriptions. Length was defined as the number of terms in the image description. Length priors have played an important role in IR, with previous research investigating the influence of this basic query-independent feature in the retrieval of textual documents [11, 4, 6].

4.2 Results

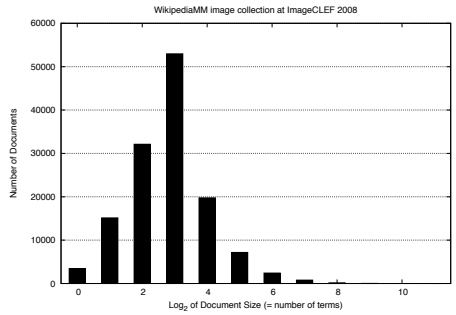
Table 1 presents the results of our official submissions. Both these text-based runs performed satisfactorily. Given though that the use of length priors has previously shown to be beneficial in the context of the INEX MM task [14, 15] (which used the same collection as the wikipediaMM task), we decided to analyse the distribution of the length (size) of the descriptions in relevant and non-relevant images.

The cwi_lm_lprior_txt run is based on the assumption that the distribution of size is different for relevant and non-relevant images. We performed a retrospective analysis of the distribution of length in the wikipediaMM collection (Figure (1a)), and the relevant images for the 2008 topics

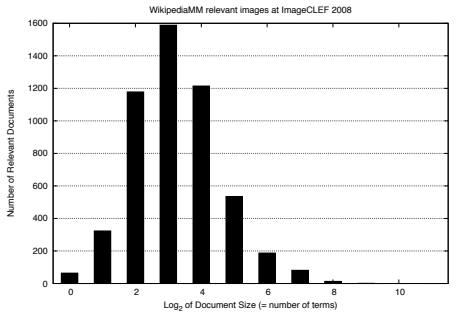
Table 1: Results for the CWI official submissions to the wikipediaMM task at ImageCLEF 2008.

runID	MAP	P@5	P@10	P@20	R-prec.	Bpref
cwi_lm_txt	0.2528	0.3840	0.3427	0.2833	0.3080	0.2673
cwi_lm_lprior_txt	0.2493	0.4293	0.3467	0.2787	0.2965	0.2622

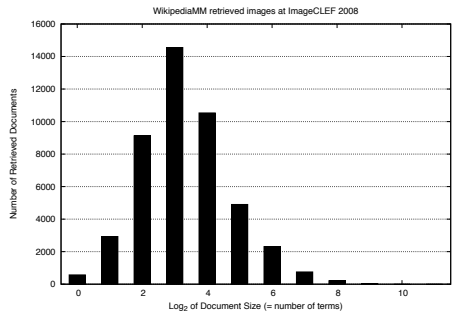
(Figure (1b)). While the collection mostly contains images with shorter descriptions, the relevant ones appear to be associated with slightly longer descriptions. If we would not pay attention to document length and just use a retrieval model that does not have a bias for documents of any size, we would retrieve shorter documents than the relevant ones. Simply giving a bias towards longer documents would have the potential of improving the retrieval results.



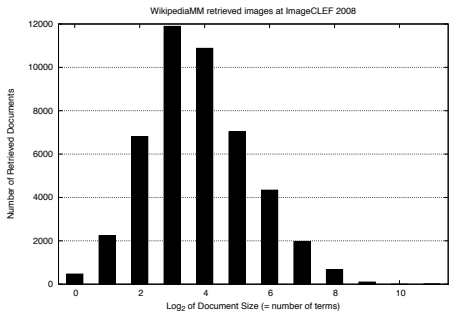
(a) wikipediaMM 2008 image collection



(b) wikipediaMM 2008 relevant images



(c) wikipediaMM 2008 retrieved images (cwi_lm_txt)



(d) wikipediaMM 2008 retrieved images (cwi_lm_lprior_txt)

Figure 1: Size distribution of images in the wikipediaMM collection, of images relevant to the wikipediaMM 2008 topics, and of images retrieved by the smoothed language modelling approach.

However, in reality, a retrieval model does not retrieve documents of all sizes uniformly. For example, the language model we use interpolates foreground and background probabilities in a standard manner and computes the foreground probability based on the relative frequency of query terms in documents. This has the effect that short documents containing query terms get a high score. Figure (1c) shows the distribution of documents that we retrieve using this language modeling approach if we do not compensate for document length (i.e., those retrieved by cwi_lm_txt). This analysis indicates that this approach already retrieves documents with sizes similarly distributed to those of the relevant ones. As a result, further biasing towards longer documents (Figure (1d)) is not beneficial. Further experiments are needed to find appropriate priors and to analyse their impact on retrieval effectiveness.

5 Conclusions

CWI used PF/Tijah to participate in the wikipediaMM task at ImageCLEF 2008. We employed a language modelling approach that considers only textual evidence and produced satisfactory results. We also examined the incorporation of length priors. Our analysis indicated that the textual descriptions of the relevant images tend to be of equal length to the ones our approach retrieves, thus biasing towards images with richer descriptions is not beneficial.

6 Acknowledgements

This material is based on work supported by the European Union via the European Commission project VITALAS (contract no. 045389).

References

[1] P. Boncz, T. Grust, M. van Keulen, S. Manegold, J. Rittinger, and J. Teubner. MonetDB/XQuery: a fast XQuery processor powered by a relational engine. In S. Chaudhuri, V. Hristidis, and N. Polyzotis, editors, *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 479–490. ACM Press, June 2006.

[2] N. Fuhr, M. Lalmas, S. Malik, and G. Kazai, editors. *Advances in XML Information Retrieval and Evaluation, Proceedings of the 4th International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2005), Revised Selected Papers*, volume 3977 of *Lecture Notes in Computer Science*. Springer, March 2006.

[3] N. Fuhr, M. Lalmas, A. Trotman, and J. Kamps, editors. *Focused access to XML documents, Proceedings of the 6th International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2007)*, volume 4862 of *Lecture Notes in Computer Science*. Springer, December 2008.

[4] D. Hiemstra. A linguistically motivated probabilistic model of information retrieval. In C. Nikolaou and C. Stephanidis, editors, *Proceedings of the 2nd European Conference on Research and Advanced Technology for Digital Libraries (ECDL 1998)*, volume 1513 of *Lecture Notes in Computer Science*, pages 569–584. Springer, September 1998.

[5] D. Hiemstra, H. Rode, R. van Os, and J. Flokstra. PF/Tijah: text search in an XML database system. In M. Beigbeder, W. Buntine, and W. G. Yee, editors, *Proceedings of the 2nd International Workshop on Open Source Information Retrieval (OSIR) (held in conjunction with SIGIR 2006)*, pages 12–17, August 2006.

[6] W. Kraaij, T. Westerveld, and D. Hiemstra. The importance of prior probabilities for entry page search. In K. Järvelin, M. Beaulie, R. Baeza-Yates, and S. H. Myaeng, editors, *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 27–34. ACM Press, August 2002.

[7] J. List, V.Mihajlović, G.Ramírez, A.P. de Vries, D. Hiemstra, and H.E. Blok. TIJAH: Embracing IR Methods in XML Databases. *Information Retrieval*, 8(4):547–570, 2005.

[8] V. Mihajlović, G. Ramírez, T. Westerveld, D. Hiemstra, H. E. Blok, and A. P. de Vries. TIJAH scratches INEX 2005: Vague element selection, image search, overlap, and relevance feedback. In Fuhr et al. [2], pages 72–87.

[9] D. R. H. Miller, T. Leek, and R. M. Schwartz. A hidden markov model information retrieval system. In F. Gey, M. Hearst, and R. Tong, editors, *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 214–221. ACM Press, August 1999.

[10] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, and J. Zobel, editors, *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 275–281. ACM Press, August 1998.

[11] A. Singhal, C. Buckley, and M. Mitra. Pivoted document length normalization. In H. P. Frei, D. Harman, P. Schaüble, and R. Wilkinson, editors, *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 21–29. ACM Press, August 1996.

[12] A. Trotman and B. Sigurbjörnsson. Narrowed Extended XPath I (NEXI). In Fuhr et al. [2], pages 16–40.

[13] A. Trotman and B. Sigurbjörnsson. Nexi, now and next. In Fuhr et al. [2], pages 41–53.

[14] T. Tsikrika, P. Serdyukov, H. Rode, T. Westerveld, R. Aly, D. Hiemstra, and A. P. de Vries. Structured document retrieval, multimedia retrieval, and entity ranking using PF/Tijah. In Fuhr et al. [3], pages 306–320.

[15] T. Tsikrika and T. Westerveld. The INEX 2007 Multimedia Track. In Fuhr et al. [3].

[16] T. Westerveld. *Using generative probabilistic models for multimedia retrieval*. PhD thesis, Centre for Telematics and Information Technology, University of Twente, 2004.

[17] T. Westerveld, R. Cornacchia, A. P. de Vries, J. C. van Gemert, and D. Hiemstra. An integrated approach to text and image retrieval- the Lowlands team at TRECVID 2005. In *TREC Video Retrieval Evaluation Online Proceedings*, 2005.

[18] T. Westerveld, A. P. de Vries, and G. Ramírez. Surface features in video retrieval. In M. Detyniecki, J. M. Jose, A. Nürnberger, and C. J. van Rijsbergen, editors, *Adaptive Multimedia Retrieval: User, Context, and Feedback, Third International Workshop (AMR 2005) Revised Selected Papers*, volume 3877 of *Lecture Notes in Computer Science*, pages 180–190. Springer, July 2006.