2011

DUTH at ImageCLEF 2011 Wikipedia Retrieval

Arampatzis, Avi

ImageClef, Wikipedia Retrieval Task

DUTH at ImageCLEF 2011 Wikipedia Retrieval

Avi Arampatzis, Konstantinos Zagoris, and Savvas A. Chatzichristofis

Department of Electrical and Computer Engineering,
Democritus University of Thrace, Xanthi 67100, Greece.
{avi,kzagoris,schatzic}@ee.duth.gr

1 Introduction

As digital information is increasingly becoming multimodal, the days of single-language text-only retrieval are numbered. Take as an example Wikipedia where a single topic may be covered in several languages and include non-textual media such as image, audio, and video. Moreover, non-textual media may be annotated with text in several languages in a variety of metadata fields such as object caption, description, comment, and filename. Current search engines usually focus on limited numbers of modalities at a time, e.g. English text queries on English text or maybe on textual annotations of other media as well, not making use of all information available. Final rankings are usually results of fusion of individual modalities, a task which is tricky at best especially when noisy modalities are involved.

In this paper we present the experiments performed by Democritus University of Thrace (DUTH), Greece, in the context of our participation to the ImageCLEF 2011 Wikipedia Retrieval task.[1] The ImageCLEF 2011 Wikipedia collection is the same as in 2010. It has image as its primary medium, consisting of $237,434$ items, associated with noisy and incomplete user-supplied textual annotations and the Wikipedia articles containing the images. Associated annotations are written in any combination of English, German, French, or any other unidentified language. This year there are 50 new test topics, each one consisting of a textual and a visual part: three title fields (one per language—English, German, French), and 4 or 5 example images. The exact details of the setting of the task, e.g., research objectives, collection etc., are provided at the task's webpage.

We kept building upon and improving the experimental multimodal search engine we introduced last year, `www.mmretrieval.net` (Fig.1). The engine allows multiple image and multilingual queries in a single search and makes use of the total available information in a multimodal collection. All modalities are indexed separately and searched in parallel, and results can be fused with different methods. The engine demonstrates the feasibility of the proposed architecture and methods, and furthermore enables a visual inspection of the results beyond the standard TREC-style evaluation. Using the engine, we experimented with different score normalization and combination methods for fusing results. We eliminated the least effective methods based on our last year's participation to ImageCLEF [1] and improved upon whatever worked best.

[1] `http://www.imageclef.org/2011/Wikipedia`

The rest of the paper is organized as follows. In Section 2 we describe the MMretrieval engine, give the details on how the Wikipedia collection is indexed and a brief overview of the search methods that the engine provides. In Section 3 we describe in more detail the fusion methods we experimented with and justify their use. A comparative evaluation of the methods is provided in Section 4; we used the 2010 topics for tuning. Experiments with the 2011 topics are summarized in Section 5. Conclusions are drawn in Section 6.

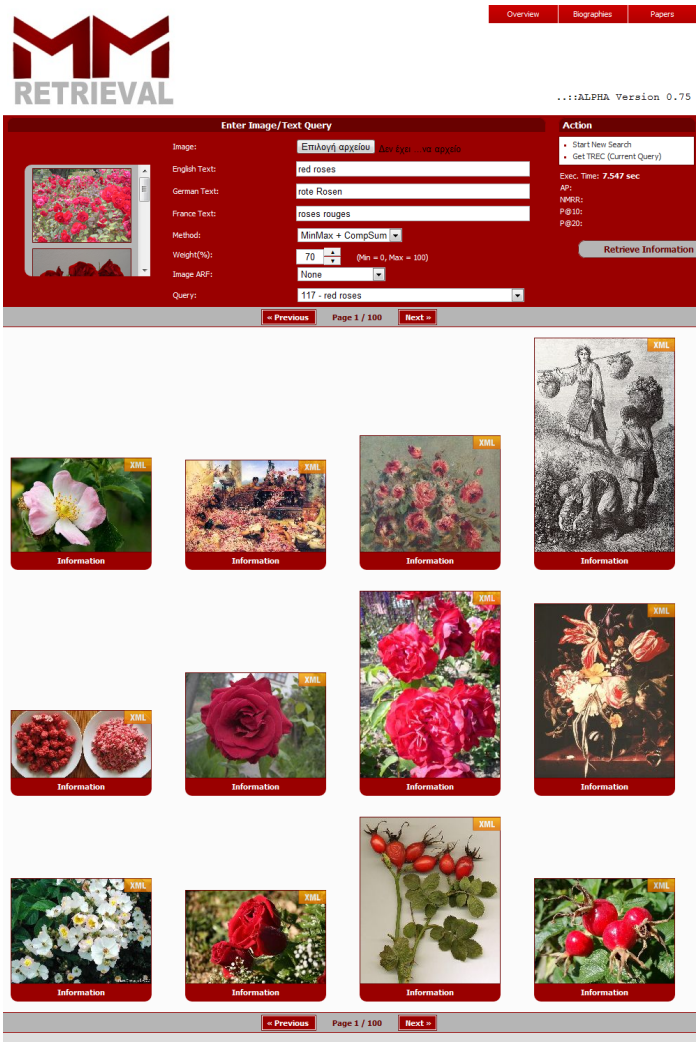

Fig. 1. The www.MMRetrieval.net search engine.

2 www.MMRetrieval.net: A Multimodal Search Engine

During last year's ImageCLEF Wikipedia Retrieval, we introduced an experimental search engine for multilingual and multimedia information, employing a holistic web interface and enabling the use of highly distributed indices [8]. Modalities are searched in parallel, and results can be fused via several selectable methods. This year, we built upon the same engine eliminating the least effective methods and trying to improve whatever worked best last year.

2.1 Indexing

To index images, we employ the family of descriptors known as Compact Composite Descriptors (CCDs). CCDs consist of more than one visual features in a compact vector, and each descriptor is intended for a specific type of image. We index with two descriptors from the family, which we consider them as capturing orthogonal information content, i.e., the Joint Composite Descriptor (JCD) [3] and the recently proposed Spatial Color Distribution (SpCD) [4]. JCD is developed for color natural images, while SpCD is considered suitable for colored graphics and artifficially generated images. Thus, we have 2 image indices.

The collection of images at hand, i.e. the ImageCLEF 2010/2011 Wikipedia collection, comes with XML metadata consisting of a description, a comment, and multiple captions, per language (English, German, and French). Each caption is linked to the wikipedia article where the image appears in. Additionally, a raw comment is supplied which may contain some of the per-language comments and any other comment in an unidentified language. Any of the above fields may be empty or noisy. Furthermore, a name field is supplied per image containing its filename. We do not use the supplied `<license>` field.

For text indexing and retrieval, we employ the Lemur Toolkit V4.11 and Indri V2.11 with the tf.idf retrieval model.[2] In order to have clean global (DF) and local statistics (TF, document length), we split the metadata and articles per language and index them separately. Thus, we have 4 indices: one per language which includes metadata and articles together but allows limiting searches in either of them, plus one for the unidentified language metadata including the name field (which can be in any language). For English text, we enable Krovetz stemming; no stemming is done for other languages in the current version of the system. We also Krovetz-stem the unidentified language metadata, assuming that most of it is probably English.

2.2 Searching

The web application is developed in the C#/.NET Framework 4.0 and requires a fairly modern browser as the underlying technologies which are employed for the interface are HTML, CSS and JavaScript (AJAX). Fig.2 illustrates an overview of the architecture. The user provides image and text queries through the web interface which are dispatched in parallel to the associated databases. Retrieval results are obtained from each of the databases, fused into a single listing, and presented to the user.

[2] `http://www.lemurproject.org`

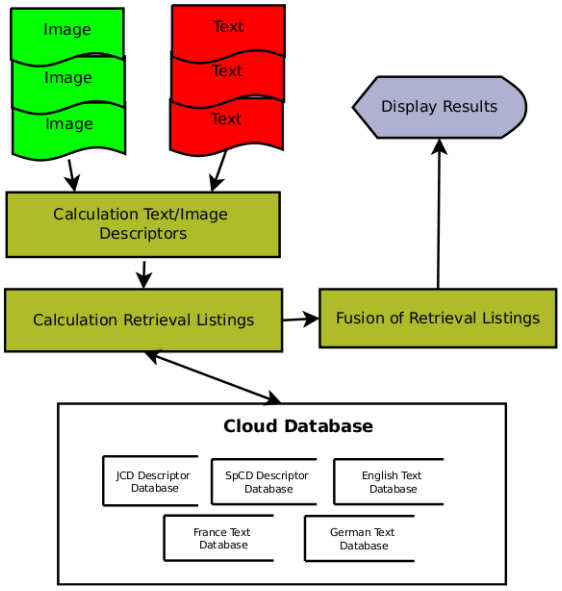

Fig. 2. System's architecture.

Users can supply no, single, or multiple query images in a single search, resulting in $2 * i$ active image modalities, where i is the number of query images. Similarly, users can supply no text query or queries in any combination of the 3 languages, resulting in $3 * l$ active text modalities, where l is the number query languages. Each supplied query results to 3 modalities: it is run against the corresponding language metadata, articles, as well as, the unidentified language metadata. The current alpha version assumes that the user provides multilingual queries for a single search, while operationally query translation may be done automatically.

The results from each modality are fused by one of the supported methods. Fusion consists of two components: score normalization and combination. In CombSUM, the user may select a weigh factor $W \in [0, 100]$ which determines the percentage contribution of the image modalities against the textual ones.

For efficiency reasons, only the top-2500 results are retrieved from each modality. If a modality returns less than 2500 items, all non-returned items are assigned zero scores for the modality. When a modality returns 2500 items, all non-occurring items in the top-2500 are assigned half the score of the 2500th item.

3 Fusion

Let $i = 1, 2, \ldots$ be the index running over example images, and j running over the visual descriptors (only two in our setup), i.e. $j \in \{1, 2\}$. Let DESC_{ji} be the score of a collection image against the ith example image for the jth descriptor.

Let $l \in \{1, 2, 3\}$ be the index running over provided natural languages (or example text queries, i.e. three in our setup), and $m \in \{1, 2, 3\}$ running over the textual data streams per language (we consider three: metadata, articles, and undefined language metadata). Let TEXT_{ml} be the score of a collection item against the text query in the lth language for the mth text stream.

Fusion consists of two successive steps: score normalization and score combination.

3.1 Score Combination

CombSUM

$$s = (1 - w)\frac{1}{ji}\sum_{j,i}\mathrm{DESC}_{ji} + w\frac{1}{ml}\sum_{m,l}\mathrm{TEXT}_{ml} \tag{1}$$

The parameter w controls the relative contribution of the two media; for $w = 1$ retrieval is based only on text while for $w = 0$ is based only on image.

CombDUTH

Image Modalities Assuming that the descriptors capture orthogonal information, we add their scores per example image. Then, to take into account all example images, the natural combination is to assign to each collection image the maximum similarity seen from its comparisons to all example images; this can be interpreted as looking for images similar to *any* of the example images. Summarizing, the score s for a collection image against the topic is defined as:

$$s = \max_{i}\left(\sum_{j}\mathrm{DESC}_{ji}\right) \tag{2}$$

Text Modalities Assuming that the text streams capture orthogonal information, we add their scores per language. Then, to take into account all the languages, the natural combination is to assign to each collection item the maximum similarity seen from its comparisons to all text queries; this can be interpreted as looking for items in *any* of the languages. Summarizing, the score s for a collection image against the topic is defined as:

$$s = \max_{l}\left(\sum_{m}\mathrm{TEXT}_{ml}\right) \tag{3}$$

Combining Media Incorporating text, again as an orthogonal modality, we add its contribution. Summarizing, the score s for a collection image against the topic is defined as:

$$s = (1 - w)\max_{i}\left(\frac{1}{j}\sum_{j}\mathrm{DESC}_{ji}\right) + w\max_{l}\left(\frac{1}{m}\sum_{m}\mathrm{TEXT}_{ml}\right) \tag{4}$$

3.2 Score Normalization

MinMax For the text modalities, we apply MinMax in different 'flavours':

- Per Modality. This is the standard MinMax taking the maximum score seen per ranked-list.
- Per Modality Type. We take the maximum score seen across ranked-lists of the same modality type. For example, to MinMax a ranked-list coming from English metadata, we take the maximum score seen across the ranked-lists of English, French, and German metadata, produced by the queries in the corresponding languages.
- Per Index Language. We take the maximum score seen across all ranked-lists from modalities coming from the same index. For example, to MinMax a ranked-list coming from English metadata, we take the maximum score seen across the ranked-lists of English metadata and English articles.
- Per Query Language. We take the maximum score seen across all ranked-lists produced by the same query language. For example, to MinMax a ranked-list coming from English metadata, we take the maximum score seen across the ranked-lists produced by English metadata, English articles, and undefined language metadata, using the same English query.

The minimum score is always 0 for tf.idf.

Given that image modalities produce scores in $[0, 100]$ (using the Tanimoto coefficient for similarity matching), we do not apply any MinMax normalization to image scores.

Query Difficulty Inverse document frequency (IDF) is a widely used and robust term weighting function capturing *term specificity* [7]. Analogously, *query specificity* (QS) or query IDF can be seen as a measure of the discriminative power of a query over a collection of documents. A query's IDF is a log estimate of the inverse probability that a random document from a collection of N documents would contain all query terms, assuming that terms occur independently. QS is a good pre-retrieval predictor for query performance [6]. For a query with k terms $1, \dots k$, QS is defined as

$$\mathrm{QS}_k = \log \left(\prod_{i=1}^{k} \frac{N}{\mathrm{df}_i} \right) = \sum_{i=1}^{k} \log \frac{N}{\mathrm{df}_i} \tag{5}$$

where df_i is the document frequency (DF), i.e. the number of collection documents in which the term i occurs.

In the Query Difficulty (QD) normalization, we divide all scores per modality by QS, using the df statistics corresponding to the modality. This will promote the scores of 'easy' modalities and demote the scores of 'difficult' modalities for the query.

For image modalities, we do a similar normalization as defined in the above equation, except that the k terms are replaced by each descriptor's bins.

w	MAP	P10	P20	bpref
0.5	0.1712	0.5329	0.4757	0.2273
0.6	0.2283	**0.5771**	0.5164	0.2825
0.7	0.2741	0.5743	**0.5221**	0.3258
0.8	**0.3004**	0.5543	0.4971	**0.3442**
0.9	0.2940	0.5186	0.4629	0.3335

Table 1. MinMax per modality + CompSUM

w	MAP	P10	P20	bpref
0.6	0.1837	0.5300	0.4807	0.2411
0.7	0.2360	**0.5700**	**0.5100**	0.2935
0.8	0.2712	0.5586	0.4879	**0.3228**
0.9	**0.2773**	0.5257	0.4557	0.3187
1.0	0.2461	0.4871	0.4121	0.2871

Table 2. MinMax per mod. type + CompSUM

w	MAP	P10	P20	bpref
0.6	0.2072	0.5571	0.5043	0.2652
0.7	0.2592	**0.5871**	**0.5264**	0.3140
0.8	**0.2882**	0.5686	0.5071	**0.3362**
0.9	0.2857	0.5400	0.4786	0.3283
1.0	0.2561	0.4986	0.4329	0.2975

Table 3. MinMax per index lang. + CompSUM

w	MAP	P10	P20	bpref
0.6	0.1653	0.4986	0.4500	0.2171
0.7	0.2177	**0.5686**	0.5021	0.2775
0.8	0.2675	**0.5686**	**0.5236**	0.3227
0.9	**0.2860**	0.5229	0.4814	**0.3348**
1.0	0.2506	0.4771	0.4214	0.2982

Table 4. MinMax per query lang. + CompSUM

4 Experiments with the 2010 Topics

4.1 MinMax+CompSUM

Tables 1, 2, 3, and 4 summarize the MinMax+CompSUM results. Best early precision is achieved by per-index-language MinMax at $w = 0.7$, while the best effectiveness in terms of MAP and all other measures is achieved by per-modality MinMax at $w = 0.8$. Per-modality-type is the weakest MinMax normalization, while per-query-language is competitive.

4.2 MinMax+CompDUTH

w	MAP	P10	P20	bpref
0.4	0.1781	0.5157	0.4636	0.2322
0.5	0.2427	**0.5471**	**0.4900**	0.2921
0.6	0.2789	0.5457	0.4864	0.3207
0.7	**0.2937**	0.5329	0.4679	**0.3304**
0.8	0.2903	0.5029	0.4421	0.3238

Table 5. MinMax per modality + CompDUTH

w	MAP	P10	P20	bpref
0.5	0.1920	0.5129	0.4486	0.2465
0.6	0.2382	**0.5171**	**0.4593**	0.2875
0.7	0.2622	0.4886	0.4521	**0.3085**
0.8	**0.2646**	0.4543	0.4136	0.3061
0.9	0.2520	0.4271	0.3879	0.2900

Table 6. MinMax per mod. type + CompDUTH

Tables 5, 6, 7, and 8 summarize the MinMax+CompDUTH results. Per-modality-type is the weakest MinMax normalization, followed by per-query-language. Best early precision is achieved by per-modality (best P10) at $w = 0.5$ and per-index-language (best P20) at $w = 0.6$. Per-modality at $w = 0.7$ achieves the best MAP, while per-index-language achieves the best bpref at $w = 0.7$. Although per-index-language has lower MAP than per-modality, its MAP comparable to per-modality; moreover, per-index-language achieves a higher bpref which signals that we may be retrieving unjudged relevant items. All in all, we conclude that per-index-language is the strongest MinMax normalization.

w	MAP	P10	P20	bpref
0.5	0.2237	0.5414	0.4764	0.2738
0.6	0.2724	**0.5429**	**0.4950**	0.3167
0.7	**0.2922**	0.5343	0.4750	**0.3325**
0.8	0.2911	0.5186	0.4593	0.3259
0.9	0.2772	0.4900	0.4371	0.3113

Table 7. MinMax per index lang. + CompDUTH

w	MAP	P10	P20	bpref
0.5	0.1702	0.4971	0.4429	0.2161
0.6	0.2283	**0.5400**	**0.4736**	0.2783
0.7	0.2624	0.5143	0.4657	0.3122
0.8	**0.2711**	0.5029	0.4514	**0.3177**
0.9	0.2596	0.4614	0.4193	0.3012

Table 8. MinMax per query lang. + CompDUTH

4.3 Overall Comparison of CompSUM, CompDUTH, and MinMax Types

Overall, best early precision is achieved by per-index-language MinMax with CompSUM at $w = 0.7$, and all other measures are optimized by per-modality MinMax with CompSUM at $w = 0.8$. However, since the 2011 topic set consists of 4 or 5 example images per topic, CompDUTH may show larger effectiveness differences than these on the 2010 topic set; consequently, we will retain CompDUTH runs with 2011 topic set, using per-index-language MinMax and $w = 0.6, 0.7, 0.8$. All these will result to 5 runs in total.

4.4 QD Normalization

w	MAP	P10	P20	bpref
0.3	0.2090	0.5557	0.4986	0.2598
0.4	0.2562	**0.5914**	**0.5243**	0.3087
0.5	0.2807	0.5657	0.5129	0.3296
0.6	**0.2928**	0.5471	0.4971	**0.3383**
0.7	0.2907	0.5286	0.4714	0.3344

Table 9. QD + CompSUM

w	MAP	P10	P20	bpref
0.2	0.1913	0.5200	0.4507	0.2427
0.3	0.2595	**0.5671**	**0.4957**	0.3034
0.4	0.2851	0.5586	0.4779	0.3252
0.5	**0.2901**	0.5371	0.4707	**0.3287**
0.6	0.2864	0.5100	0.4507	0.3236

Table 10. QD + CompDUTH

Tables 9 and 10 summarize the QD normalization results with both combination methods. In early precision, the QD normalization works much better with CompSUM than with CompDUTH. The best CompSUM results are achieved for $w = 0.4$; this run has also the best P10 we have reported so far. In all other measures, although CompSUM is slightly better than CompDUTH, their effectiveness is comparable.

In comparison to the MinMax normalizations, the QD normalization achieves the best initial precision results (when CompSUM is used for combination), and comparable effectiveness to the best MinMax normalization in all other measures.

In summary, we will retain QD+CompSUM at $w = 0.4$ and QD+CompDUTH at $w = 0.3$ and 0.5; thus, we will have 3 QD runs in total.

4.5 Summary

While we have experimented with radically different normalization and combination methods, our results have not shown a large variance. This suggests that we are 'pushing' at the effectiveness ceiling of the 2010 dataset. It is worth noting that most of

the runs reported so far have a better MAP and bpref than last year's best automatic run submitted to ImageCLEF, and a slightly lower but comparable initial precision.[3] Nevertheless, a visual inspection of our results reveals that with CompDUTH we are retrieving un-judged items which are sometimes relevant, a fact that most of the times does not seem to get picked up by bpref.

5 Experiments with the 2011 Topics

Run	w	Details	MAP	P10	P20	bpref
QD + CompSUM	0.6		**0.2886**	0.4860	0.3870	0.2905
QD + CompDUTH	0.5		0.2871	0.4620	0.3870	0.2885
QD + CompSUM	0.4		0.2866	**0.5120**	**0.4190**	**0.3014**
MinMax + CompDUTH	0.7	PerIndexLang	0.2840	0.4580	0.3990	0.2775
MinMax + CompSUM	0.7	PerIndexLang	0.2818	0.4840	0.3990	0.2945
MinMax + CompDUTH	0.6	PerIndexLang	0.2786	0.4640	0.4110	0.2815
MinMax + CompDUTH	0.8	PerIndexLang	0.2751	0.4360	0.3730	0.2677
MinMax + CompSUM	0.8	PerModality	0.2717	0.4380	0.3740	0.2728
QD + CompDUTH	0.3		0.2605	0.4840	0.4090	0.2768

Table 11. Results with the 2011 topics, sorted on MAP.

Table 11 summarizes a selection of our official runs with the 2011 topics. Note that we submitted more runs employing pseudo-relevance feedback methods for the image modalities which we do not include or analyse here; their performance was comparable to the ones included in the table.

In score combination, the simplest method of linearly combining evidence (CompSUM) is once more found to be robust, irrespective of the normalization method. However, CompDUTH is very competitive with similar performance. In score normalization, query difficulty normalization (QD) gives the best effectiveness in both MAP and initial precision when scores are combined with CompSUM.

The current experiment points to that the choice of normalization method is more important than the combination method. MinMax achieves the best results for $w = 0.6$ or 0.7, i.e. retrieval based on 60-70% text, while with QD the contribution of text can be reduced to 40-60% improving overall effectiveness. It seems that with a better score normalization across modalities or media, we can use more of content-based image retrieval in a multimodal mix.

[3] Last year's best MAP, P10, P20, and bpref were 0.2765, 0.6114, 0.5407, and 0.3137, respectively; they were all achieved by the XRCE group [5].

6 Conclusions

We reported our experiences and research conducted in the context of our participation to the controlled experiment of the ImageCLEF 2010 Wikipedia Retrieval task. As second-time participants, we improved upon and extended our experimental search engine, `http://www.mmretrieval.net`, which combines multilingual and multi-image search via a holistic web interface and employs highly distributed indices. Modalities are search in parallel, and results can be fused via several methods.

All in all, we are modestly satisfied with our results. Although our best MAP run ranked our system as the second-best among the other participants' systems (excluding all relevance feedback and query expansions runs), we believe that the content-based image retrieval part of the problem has a large room for improvement. A promising direction may be using new image modalities such as those based on the bag-of-visual-words paradigm and other similar approaches. Furthermore, we consider score normalization and combination important problems; while effective methods exist in traditional text retrieval, those problems are not trivial in multimedia setups.

References

1. Arampatzis, A., Chatzichristofis, S.A., Zagoris, K.: Multimedia search with noisy modalities: Fusion and multistage retrieval. In: Braschler et al. [2]
2. Braschler, M., Harman, D., Pianta, E. (eds.): CLEF 2010 LABs and Workshops, Notebook Papers, 22-23 September 2010, Padua, Italy (2010)
3. Chatzichristofis, S.A., Boutalis, Y.S., Lux, M.: Selection of the proper compact composite descriptor for improving content-based image retrieval. In: SPPRA. pp. 134–140 (2009)
4. Chatzichristofis, S.A., Boutalis, Y.S., Lux, M.: SpCD - Spatial Color Distribution Descriptor - A fuzzy rule-based compact composite descriptor appropriate for hand drawn color sketches retrieval. In: Proceedings ICAART. pp. 58–63. INSTICC Press (2010)
5. Clinchant, S., Csurka, G., Ah-Pine, J., Jacquet, G., Perronnin, F., Sánchez, J., Minoukadeh, K.: Xrce's participation in wikipedia retrieval, medical image modality classification and ad-hoc retrieval tasks of imageclef 2010. In: Braschler et al. [2]
6. Cronen-Townsend, S., Zhou, Y., Croft, W.B.: Predicting query performance. In: SIGIR. pp. 299–306. ACM (2002)
7. Spärck Jones, K.: A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation 28, 11–21 (1972), `http://www.soi.city.ac.uk/~ser/idf.html`
8. Zagoris, K., Arampatzis, A., Chatzichristofis, S.A.: www.mmretrieval.net: a multimodal search engine. In: SISAP. pp. 117–118. ACM (2010)