XRCE and CEA LIST's Participation at Wikipedia Retrieval of ImageCLEF 2011

Gabriela Csurka[1] and Stéphane Clinchant[1,2] and Adrian Popescu[3]

[1] Xerox Research Centre Europe, 6 chemin de Maupertuis 38240, Meylan France
firstname.lastname@xrce.xerox.com
[2] LIG, Univ. Grenoble I, BP 53 - 38041 Grenoble cedex 9, Grenoble France
[3] CEA, LIST, Vision & Content Engineering Laboratory, 92263 Fontenay aux Roses, France
adrian.popescu@cea.fr

Abstract.
In this document, we first recall briefly our baseline methods both for text and image retrieval and describe our information fusion strategy, before giving specific details concerning our submitted runs.

As text retrieval, XRCE used either and Information-Based IR model [4] or a Lexical Entailment IR model based on statistical translation IR model [5]. Alternatively, we also used an approach from CEA List that models the queries using on one hand socially related Flickr tags and on the other hand Wikipedia concepts introduced in [13]. The combination of these runs have shown that the approaches were rather complementary.

As image representation, we used spatial pyramid of Fisher Vectors built on local orientation histograms and local RGB statistics. The dot product was used to define the similarity between two images and to combine the color and texture based ranking we used simple score averaging.

Finally, to combine visual and textual information, we used a so called the Late Semantic Combination (LSC) method [3], where first the text expert is used to retrieved semantically relevant documents, and than the visual and textual scores are averaged to rank these documents. This strategy allowed us to significantly improve over mono-modal retrieval performances. Using the late fusion of the best text expert from XRCE and from CEA and combining with our Fisher Vector based image run with LSC leaded to a MAP of 37% (best score obtained in the Challenge).

Keywords

Multi-modal Information Retrieval, Wikipedia Retrieval, Fisher Vector, Lexical Entailment

1 Introduction

The Wikipedia Retrieval task consisted of multilingual and multimedia retrieval [14]. The collection contains images with their captions extracted from Wikipedia in different languages namely French, English and German. In addition, participants were provided with the original Wikipedia pages in wikitext format. The aim was to retrieve as many relevant images as possible from the aforementioned collection, given a textual query translated in the three different languages and one or several query images.

Each team submitted different types of runs (see Table 1) : mono-media runs (text) and multimedia (mixed) runs with different fusion approaches. As the table shows, we also submitted common runs for which we applied the semantic late filtering or late combination of XRCE textual runs with the textual from CEA List.

As the results show, these textual runs were complemantary to our runs that used no external resources and hence their combination allowed for further boosting the performance of mono and multi-modal retrieval.

Our image representation based on Fisher Vectors is briefly recalled in section 4. To combine visual and textual information, we used a semantic filtered late combination method described in section 5. Finally, we give specific details about the submitted run in section 6 and conclude in 7

2 XRCE Text Based IR Models

We start from a traditional bag-of-words representation of pre-processed texts where pre-processing includes tokenization, lemmatization, and standard stopword removal. However, in some cases lemmatization might lead to a loss of information. Therefore before building the bag-of-words representation we concatenated a lemmatized version of the document with the original document. We build an index for the image captions and one for the surrounding paragraph containing the images (as last year). For all runs, we average the score obtained on the captions and paragraph index.

We used basically two textual models, the Smoothed Power Law (SPL) Information-Based Model [4] and the Lexical Entailment (AX) IR Model [5].

2.1 Information Based IR Model (SPL)

Information models draw their inspiration from a long-standing hypothesis in IR, namely the fact that the *difference in the behaviors of a word at the document and collection levels brings information on the significance* of the word for the document. This hypothesis has been exploited in the 2-Poisson mixture model, in the notion of eliteness in BM25, and more recently in DFR models. In particular, several researchers, Harter [6] being one of the first ones, have observed that the distribution of significant, ”specialty” words in a document deviates from the distribution of ”functional” words. The more the distribution of a word in a document deviates from its average distribution in the collection, the more likely is this word significant for the document considered. This can be easily captured in terms of information:

$$\text{Info}(x) = -\log P(X = x|\lambda) = \text{Informative Content} \tag{1}$$

If a word behaves in the document as expected in the collection, then it has a high probability $P(X = x|\lambda)$ of occurrence in the document, according to the collection distribution, and the information it brings to the document, $-\log P(X = x|\lambda)$, is small. On the contrary, if it has a low probability of occurrence in the document, according to the collection distribution, then the amount of information it conveys is greater. In a nutshell, information can be understood as a *deviation from an average behavior.*

Overall, the general idea of the information-based family is the following:

1. Due to different document length, discrete term frequencies (x_w^d) are renormalized into continuous values $t_w^d = t(x_w^d, l_d)$
2. For each term w , we assume that the renormalized values t_w^d follow a probability distribution P on the corpus. Formally, $T_w \sim P(.|\lambda_w)$.
3. Queries and documents are compared through a measure of surprise, or a mean of information of the form
$$RSV(q, d) = \sum_{w \in q} -q_w \log P(T_w > t_w^d|\lambda_w)$$

So, information models are specified by *two main components*: a function which normalizes term frequencies across documents, and a probability distribution modeling the normalized term frequencies. Information is the key ingredient of such models since information measures the significance of a word in a document.

We choosed for our runs the Smoothed Power model proposed in [4]. This model is specified in two steps: the Divergence from Randomness (DRF) normalization of terms frequencies and the Smooth Power Law (SPL) distributions:

- DFR Normalization with parameter c: $t_w^d = x_w^d \log(1 + c\frac{avg_l}{l_d})$
- $Tf_w \sim \mathrm{SPL}(\lambda_w = \frac{N_w}{N})$

where avg_l is the mean document length, l_d the document length, c a parameter N is the number of documents in the collection and N_w the number of document containing word w. The retrieval function is then:

$$RSV(q,d) = \sum_{w \in q \cap d} -x_w^q \log(\frac{\lambda_w^{\frac{t_w^d}{t_w^d+1}} - \lambda_w}{1 - \lambda_w}) \qquad (2)$$

2.2 Lexical Entailment based IR Models (AX)

Berger and Lafferty [2] addressed the problem of information retrieval as a statistical translation problem with the well-known noisy channel model. This model can be viewed as a probabilistic version of the generalized vector space model. The analogy with the noisy channel is the following one: To generate a query word, a word is first generated from a document and this word then gets "corrupted" into a query word. The key mechanism of this model is the probability $P(v|u)$ that term u is "translated" by term v. These probabilities enable us to address a vocabulary mismatch, and some kinds of semantic enrichments. The problem now lies in the estimation of such probability models.

We refer here to a previous work [5] on lexical entailment models to estimate the probability that one term entails another. It can be understood as a probabilistic term similarity or as a unigram language model associated to a word (rather than to a document or a query). Let u be a term in the corpus, then lexical entailment models compute a probability distribution over terms v of the corpus $P(v|u)$. These probabilities can be used in information retrieval models to enrich queries and/or documents and to give a similar effect to use of a semantic thesaurus. However, lexical entailment is purely automatic, as statistical relationships are only extracted from the considered corpus. In practice, a sparse representation of $P(v|u)$ is adopted, where we restrict v to be one of the N_{max} terms that are the closest to u using an Information Gain metric. We computed the probabilistic term similarity on the paragraph collections because we believed paragraphs could capture a better semantical context as opposed to captions. and retain the top 10 words for each word in the collection.

3 CEA LIST Text and Visual Prototype Based Retrieval

The main idea behind in CEA LIST's approach is to model the queries using on one hand socially related Flickr tags and on the other hand Wikipedia concepts and to combine them. Here we provide only a quick description of Flickr and Wikipedia models which are similar to those introduced in [13]. The main novelty introduced this year is the creation of a visual topic prototype with visual concepts. Each result image is then characterized in the visual prototype space and the results which match the prototype are ranked higher using a late fusion approach.

3.1 Flickr Query Modeling

Term relations extracted from Flickr are defined within a photographic tagging language. We use this data source to define an adaptation of the TF-IDF model to the social space. Given a query Q, we define its social relatedness to term T using:

$$SocRel(T|Q) = users(Q,T) * 1/log(pre(T)) \ (1)$$

where $users(Q,T)$ is the number of distinct users which associate tag T to query Q among the top 20,000 results returned by the Flickr API for the query Q; and where $pre(T)$ is the number of distinct users from a prefetched subset of 30,000 Flickr users that have tagged photos with tag T.

In this new social weighting scheme, term frequency and document counts from the classical IR formulas are replaced with user counts, which prevents the final relatedness score from being biased by heavy contributions from a reduced number of users. The computation of $users(Q,T)$ from 20,000 top results is destined to keep computation time low, while accounting for different query relevant contexts. $pre(T)$ is precomputed from all the tags submitted by a random subset of 120,000 Flickr users.

Related terms are computed from preprocessed queries. This processing involves removing photographic terms (which can have a negative impact on queries with few results) as well as prepositions and articles from the queries. Both prepositions and photographic terms are kept in precompiled lists, extracted from Wikipedia: the "List of English prepositions" page, lists of photographic terms exist on the "Category:Film techniques" and "Category:Photographic processes" pages. In the following sections, we will use the preprocessed forms of the queries. For instance, *skeleton of dinosaur* becomes *skeleton dinosaur*. For this query, the most related Flickr terms are: *bones*, *museum*, *trex*, *fossil*, *natural history museum* and *tyrannosaurus rex*.

Before using Wikipedia, we create an enriched version of the query (Q_E) by selectively stemming orginal query terms. A Flickr related term is retained as a variant if its edit distance to one stemmed term from the initial query is smaller than three or if the stemmed form is a prefix of the related term, so *skeleton dinosaur* becomes *(skeleton:skeletons) (dinosaur:dinosaurs)*. Words in the initial query and the top related Flickr term have weights of 1 while terms starting from x = 2 have weights which are normalized with $SocRel(T|Q_1)$. This weighting expresses the fact that the importance of a term in the Flickr model decreases with its rank among the socially related terms. . We create a Flickr query model for each language in an identical manner.

3.2 Wikipedia Query Modeling

When the terms in a query (or a part of a query) are categorical in nature, results that correspond to their subtypes or to other semantically related concepts are ignored in absence of query expansion. For instance, *tyrannosaurus* or *allosaurus* are valid subtypes of *dinosaur*, which is a part of the topic *skeleton of dinosaur*. Images tagged with these related concepts are potentially relevant for *skeleton of dinosaur* but they would not be returned when querying with the initial terms. Query expansion is particularly useful in cases when initial queries return a small number of results or for languages that are seldom used in the annotations of the images. Since image queries cover a broad range of concepts and are expressed in different languages, a generic, detailed and multilingual data source is needed to enable an efficient expansion and we consider Wikipedia to be an appropriate data source for the extraction of semantically related concepts.

We express the semantic relatedness between a query and a Wikipedia article as a combination of two scores. We first measure the overlap between the words in the query and the words in the category section and in the first sentence of encyclopedic articles and then compute the dot product between the query and a vectorial representation of the entire article content. Priority is given to the first score because categorical and definitional information have a priviledged role in defining semantic relatedness. For English, queries are run through WordNet and synonyms are added to the query terms that are unambiguous in WordNet, to avoid introducing noise from polysemous word senses. Overlap is normalized by the number of words in a query and its values vary from 0 (no terms in common between the query and the article's categories) and 1 (all terms in common). Due to the fact that queries usually contain a small number of words, the overlap scores offer a coarse expression of semantic relatedness and a large number of articles will share the same scores. In our system, given the query *golf player on court*, an article categorized under with *golf* and *player* is always better ranked than an article categorized only under *player*.

3.3 Text Retrieval

The collection is preprocessed in order to represent textual annotations using a TF-IDF formalism. Wikipedia and Flickr models have different roles in the retrieval system. Related concepts from the encyclopedia are used for semantic query expansion, whereas the set of related Flickr tags is exploited for result ranking. We first retrieve all documents in the collection that are annotated with at least one word from the initial query or with one of the related Wikipedia concepts and give these documents a coarse ranking score which is based on the number of terms from the initial query the document is related to. To break the many ties that result from the use of the coarse score, we compute a fine-grained score as the cosine similarity between the document representation and the Flickr query model.

3.4 Visual Prototype Based Retrieval

Each 2011 Wikipedia Retrieval topic is provided with four or five example images. We extract visual concepts from these positive examples in order to extract a visual prototype of the query. Extracted visual concepts include one or several of the following: presence of a face, indoor/outdoor scene, black & white vs. color image; photograph/clipart/map content. For instance, the prototype of *portrait of Che Guevara* includes the *face* and *photograph* while a the prototype of *golf player on green* will include *outdoor*, *photograph*, *color*. Each image in the collection is preprocessed in order to extract the visual concepts described above. We then compare each image in the textual ranking to the query prototype and increase its visual score each time for each matching visual concept. We take a simple approach a give the same score to each different visual concept we extracted.

4 XRCE Fisher Vector based Image Representation

As for the image representation, we used an improved version [12, 11] of the Fisher Vector [10]. The Fisher Vector can be understood as an extension of the bag-of-visual-words (BOV) representation. Instead of characterizing an image with the number of occurrences of each visual word, it characterizes the image with the gradient vector derived from a generative probabilistic model. The gradient of the log-likelihood describes the contribution of the parameters to the generation process.

Assuming that the local descriptors $I = \{x_t, x_t \in \mathbb{R}^D, t = 1 \dots T\}$ of an image I are generated independently by Gaussian mixture model (GMM) $u_\lambda(x) = \sum_{i=1}^{M} w_i \mathcal{N}(x|mu_i, \Sigma_i)$, I can be described by the following gradient vector (see also [7, 10]):

$$G_\lambda^I = \frac{1}{T} \sum_{t=1}^{T} \nabla_\lambda \log u_\lambda(x_t) \tag{3}$$

where $\lambda = \{w_i, \mu_i, \Sigma_i, i = 1 \dots M\}$ are the parameters of the GMM. A natural kernel on these gradients is the Fisher Kernel [7]:

$$K(I, J) = G_\lambda^{I\prime} F_\lambda^{-1} G_\lambda^J, \qquad F_\lambda = E_{x \sim u_\lambda} \left[\nabla_\lambda \log u_\lambda(x) \nabla_\lambda \log u_\lambda(x)' \right]. \tag{4}$$

where F_λ is the Fisher information matrix. As it is symmetric and positive definite, F_λ^{-1} has a Cholesky decomposition $F_\lambda^{-1} = L_\lambda' L_\lambda$ and $K(I, J)$ can be rewritten as a dot-product between normalized vectors $\mathcal{G}_\lambda$ with: $\mathcal{G}_\lambda^I = L_\lambda G_\lambda^I$. We will refer to $\mathcal{G}_\lambda^I$ as the *Fisher Vector* (FV) of the image I.

In the case of diagonal covariance matrices Σ_i (we denote by σ_i^2 the corresponding variance vectors), closed form formulas can be derived for $\mathcal{G}_{w_i^d}^I$, $\mathcal{G}_{\mu_i^d}^I$, $\mathcal{G}_{\sigma_i^d}^I$, for $i = 1 \dots M$, $d = 1 \dots D$ (see details in [11]). As we do not consider $\mathcal{G}_{w_i^d}^I$ (the derivatives according to the weights), $\mathcal{G}_\lambda^I$ is the concatenation of the derivatives $\mathcal{G}_{\mu_i^d}^I$ and $\mathcal{G}_{\sigma_i^d}^I$ and is therefore $N = 2MD$-dimensional.

The Fisher Vector is further normalized with Power ($\alpha = 0.5$) and L2 normalization as suggested in [11] and the dot product is used as similarity between the Fisher Vectors. We also used in some cases the spatial pyramid [8] to take into account the rough geometry of a scene. The main idea is to repeatedly subdivide the image and represent each layout as a concatenation of the representations (in our case Fisher Vectors) of individual sub-images. As we used three spatial layouts (1×1, 2×2, and 1×3), we obtained 3 image representations of respectively N, $4N$ and $3N$ dimensions.

As low level features we used our usual (see for example [1]) SIFT-like Orientation Histograms (ORH) and local color statistics (COL), *i.e.* local color means and standard deviations in the R,G and B channels, both extracted on regular multi-scale grids and reduced to 50 or 64 dimensional with Principal Component Analysis (PCA).

5 The Late Semantic Combination Fusion Method

There has been many research works addressing text/image information fusion. The method we mainly used in our runs was the one we described in [3]. The intuition behind this technique is that since different media (here text and image) are semantically expressed at different levels, we should not combine them independently as most of information fusion techniques so far do, but on the contrary, we should consider the underlying complementarities that exist between these media. In the case of text/image data fusion, as the results in most ImageClef Task shows [9], text based search is more efficient than visual based since it is more difficult to extract the semantics of an image compared to a text. However, basic late fusion approaches showed that combination of visual and textual information can outperform the only text based search. This shows that the two media are often complementary to each other despite the differences between monomedia performances.

In [3], we have shown that the late fusion can be improved by simply adding a semantic filtering step before score combination. This filtering step has the role of enforcing the visual system to search among the set of retrieved objects by the text expert. In this way we impose that images visually similar to the query images share a common semantic (given by the textual query). While this filtering step is the basis of the image reranking methods, we have shown in [3] that remaining at this level is unsufficient. Indeed, when the visual system has low performance, the image reranking allows to significantly out-perform it however its performance is generally poorer than using only the text alone. Therefore, what we proposed was to *combine image reranking with late fusion* in order to overcome their respective weaknesses. Note that the strength of image reranking is to realign the visual system to search in a relevant subset with respect to the semantic viewpoint, while the strength of late fusion relies on a well performing text expert.

Hence, the *Late Semantic Combination (LSC)* combination works as follows. First, we define a semantic filter for the image scores according to the textual expert:

$$SF(q,d) = \begin{cases} 1 & \text{if } d \in \mathrm{KNN}_t(q) \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

where $\mathrm{KNN}_t(q)$ denotes the set of the K most similar objects to q according to the textual similarities. Hence, this will give us a reduced list (K) for which we need to compute the image similarities.

After normalization (all scores are transformed to have values between 0 and 1), the semantically filtered image scores are combined with the text ones:

$$s_{LSC}(q,d) = \alpha_t \mathbf{N}(s_t(q,d)) + \alpha_v(\mathbf{N}(SF(q,d)s_v(q,d))) \tag{6}$$

where $\mathbf{N}$ is an operator normalizing score between 0 and 1, $\alpha_t = \alpha$ and $\alpha_v = 1 - \alpha$ are positive weights that sum to 1. Note that the similarity for all documents d that are not in $\mathrm{KNN}_t(q)$ are set to 0.

6 Descriptions of our runs

Similarly to 2010, we evaluated different mono-media and multimedia runs on the Wikipedia corpus using the new set of queries. The MAP and P10 performances of these runs are shown in Table 1.

Some of the runs were submitted separately by each of the two participants while others were the result of a combination of text experts from XRCE and CEA LIST. The latter runs were obtained either using the late semantic filtering where CEA LIST Runs were used as text expert or first we combined their run with our text expert before using the LSC method.

In addition, in Table 1 we show also the performance of our visual run (not submitted). The performance of these visual runs were again poor, as averaging the Fisher Vectors of the set of image queries was not sufficient to capture the underlying semantics of the query. However, as in 2010, we can observe that in spite of the low performance of the visual expert, the Semantic Filtered Late was able to take advantage of the complementarity of the media types to improve over the pure text based approach.

Table 1. Wikipedia retrieval: overview of the performances of our different runs. Top table deals with monomedia runs and bottom table with multimedia runs.

ID	RUN	MAP	P@10
T6	XRCE_CEA_TXT_RUN_SPLAX_ENFRDE	0.3141	0.516
T5	XRCE_CEA_TXT_RUN_AX_ENFRDE	0.3130	0.530
T1	XRCE_TXT_RUN_AX	0.2780	0.470
T4	XRCE_TXT_RUN_SPLAX	0.2769	0.464
T3	CEA_enfrde_all	0.2591	0.466
T2	XRCE_TXT_RUN_SPL	0.2432	0.422
I1	XRCE_VIS_FV	0.0271	0.0860

Txt Run	ID	RUN	MAP	P@10	Rel. Improv.
	M1	XRCE_CEA_MULTI_RUN_SFL_AX_viscon	0.3880	0.632	
T5	M2	XRCE_CEA_MULTI_RUN_AX_ENFRDE_FV_SFL	0.3869	0.624	+23.6 %
T6	M3	XRCE_CEA_MULTI_RUN_SPLAX_ENFRDE_FV_SFL	0.3848	0.620	+22.5 %
T1	M4	XRCE_MULTI_RUN_AX_FV_SFL	0.3557	0.594	+27.9 %
T4	M5	XRCE_MULTI_RUN_SPLAX_FV_SFL	0.3556	0.578	+28.4 %
	M6	XRCE_CEA_MULTI_RUN_SPLAX_VISCON	0.3471	0.574	
T3	M7	CEA_XRCE_RUN_ENFRDE_FV_SFL	0.3075	0.54	
	M8	CEA_viscon_1.07	0.2703	0.480	

In which follows, we give details on each run individually.

6.1 *Text based runs:*

- **T1:** XRCE Text based retrieval with the Lexical Entailment (AX) based IR Models (section 2.2).
- **T2:** XRCE Text based retrieval with the the Smoothed Power Law (SPL) Information-Based Model (section 2.1).
- **T3:** CEA LIST multilingual textual run (section 3.3).
- **T4:** Late fusion between T1 and T2.
- **T5:** Late fusion between T1 and T3.
- **T6:** Late fusion between T4 and T3.

We can see from the Table 1 that on this dataset AX works better than SPL. While their combination does not seem to help, when we further combine them with T3 (T6) we obtain better performance than just combining T1 with T3 (T5). In all cases, combining XRCE runs with the CEA run (T3) leaded to an absolute improvement of 4% in MAP.

6.2 *Image based run.*

Our image run (I1) was based on similarity between Fisher Vector based signatures as described in section 4. We built 4 image signatures for each image corresponding to the two different low level features (ORH and COL) and for each using either a global FV or a spatial pyramid (1x1_2x2_1x3). The 4 FVs were used independently to rank the Wikipedia images using the dot product as similarity measure and the 4 scores were simply averaged.

6.3 *Text and image based runs.*

The combination of visual and textual runs were done using the Late Semantic Combination combination described in section 5. The image scores were all the same, correponding to I1 described above, so only the text expert changed from one run to other. Hence:

- **M2:** used T5, a late fusion between AX and the XRCE text run from CEA.
- **M3:** used T6, a late fusion between AX, SPL and the text run from CEA.
- **M4:** used T1 corresponding to the AX model.
- **M5:** used T4 corresponding to the late fusion between AX and SPL.
- **M7:** used T3, CEA multilingual textual run.

The three remaining runs were based on M8, a linear combination of textual results and visual prototype based results from CEA (see details in section 3.4). Hence, we had:

- **M6:** a late fusion between T4 and M8.
- **M1:** a late semantic combination where we used M6 as “the semantic filter” and also we averages the visual scores with the normalized scores of the run M7. Hence, while M7 was not purely text based, it had the role of the “text expert” in the LCS approach.

As a conclusion, we can again see that the CEA and XRCE runs were complementary with a gain of absolute improvement of 3% in MAP. While adding the visual characterization of the topics to pure text retrieval helps (T2 compared to M6), when we added the visual expert the gain was almost negligible (M1 compared with M2). This is interesting as it reinforces some of our observations we made in [1] experimenting with the IAPR TC-12 photographic collection.

7 Conclusion

This year XRCE participated again with success in the Wikipedia Retrieval Task showing again that despite the fact that pure visual based retrieval led to poor results, when we appropriately combined them with our text ranking we are able to outperform the mono-modal ones. This was shown with various text expert, including those from the CEA LIST.

Acknowledgments

We would like also to acknowledge Florent Perronnin and Jorge Sánchez for the efficient implementation of the Fisher Vectors we used in our experiments.

Adrian Popescu was supported by the French ANR (Agence Nationale de la Recherche) via the Periplus project.

References

1. J. Ah-Pine, S. Clinchant, G. Csurka, F. Perronnin, and J-M. Renders. *Leveraging image, text and cross-media similarities for diversity-focused multimedia retrieval*, chapter 3.4. Volume The Information Retrieval Series of etal [9], 2010. ISBN 978-3-642-15180-4.

2. Adam Berger and John Lafferty. Information retrieval as statistical translation. In *In Proceedings of the 1999 ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 222–229, 1999.
3. Stéphane Clinchant, Julien Ah-Pine, and Gabriela Csurka. Semantic combination of textual and visual information in multimedia retrieval. In *ACM International Conference on Multimedia Retrieval (ICMR)*, 2011.
4. Stéphane Clinchant and Eric Gaussier. Information-based models for ad hoc ir. In *SIGIR '10: Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 234–241, New York, NY, USA, 2010. ACM.
5. Stéphane Clinchant, Cyril Goutte, and Éric Gaussier. Lexical entailment for information retrieval. In *Advances in Information Retrieval, 28th European Conference on IR Research, ECIR 2006, London, UK, April 10-12*, pages 217–228, 2006.
6. S. P. Harter. A probabilistic approach to automatic keyword indexing. *Journal of the American Society for Information Science*, 26, 1975.
7. T. Jaakkola and D. Haussler. Exploiting generative models in discriminative classifiers. In *Advances in Neural Information Processing Systems 11*, 1999.
8. S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.
9. H. Müller, P. Clough, Th. Deselaers, and B. Caputo, editors. *ImageCLEF- Experimental Evaluation in Visual Information Retrieval*, volume The Information Retrieval Series. Springer, 2010. ISBN 978-3-642-15180-4.
10. F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In *CVPR*, 2007.
11. F. Perronnin, J. Sánchez, and Y. Liu. Large-scale image categorization with explicit data embedding. In *CVPR*, 2010.
12. Florent Perronnin, Yan Liu, Jorge Sánchez, and Hervé Poirier. Large-scale image retrieval with compressed fisher vectors. In *CVPR*, 2010.
13. Adrian Popescu and Gregory Grefenstette. Social media driven image retrieval. In *ACM International Conference on Multimedia Retrieval (ICMR)*, 2011.
14. Theodora Tsikrika, Adrian Popescu, and Jana Kludas. Overview of the wikipedia retrieval task at imageclef 2011. In *Working Notes of CLEF 2011, Amsterdam, The Netherlands*, 2011.