Télécom Bretagne at ImageCLEF WikipediaMM 2010

Adrian Popescu

Institut Télécom/Télécom Bretagne, Département Informatique
adrian.popescu@telecom-bretagne.eu

Abstract. In this paper, I describe the approach proposed by Télécom Bretagne for the WikipediaMM 2010 evaluation campaign [6]. One of the main challenges in large scale image retrieval is the mismatch between query terms and image textual descriptions from the database. This mismatch can be reduced using query expansion and here I present a Wikipedia based query expansion approach. In order to boost results' accuracy, the expansion is followed by a reranking step which uses query models extracted from Flickr.

1 Introduction

Retrieving images from heterogeneous and noisy databases was thoroughly studied but remains an interesting research topic. Some open questions include:

- how to deal with difficult queries?
- how to perform query expansion in a manner that improves both precision and recall?
- what resources to use in order to model query content?

In this paper, I present techniques which represent possible answers to these questions. A query expansion technique is adapted from previous work, and is augmented a query modeling module based on Flickr tags. The categorical structure of Wikipedia is exploited in order to find and rank concepts from the encyclopedia which are semantically similar to the initial query. The experiments are focused on English queries but the same techniques are applicable to other languages.

With over 3,000,000 articles in its English version, Wikipedia is a rich resource and is used in a variety of research tasks, such as: sense disambiguation, ontology extraction or semantic relatedness. The last problem can be formulated as follows: given an input (a concept or a longer text), find the concepts which are most closely related to the input. Wikipedia based techniques to find semantic relatedness include WikiRelate! [5], Explicit Semantic Analysis (ESA) [1] and Wikipedia Link-based Measure (WLM) [2]. WikiRelate! modifies techniques previously applied to WordNet in order to suit Wikipedia's structure. The authors of [1] map queries to Wikipedia concepts representation in order to find related concepts. ESA is interesting because it finds related concepts for any

given query and not only for mono-conceptual queries and is thus suited for use in Web information retrieval. WLM exploits only Wikipedia links to find related concepts.A comparison of the three techniques [2] shows that ESA achieves the best performances, followed by WLM and WikiRelate!. Modeling Flickr content is another very interesting area of research. Related to the present paper are techniques that analyze Flickr tags to find frequent topics and their correlations [4] and [8]. Correlations are then used in order to suggest new tags based on supplied tags. The authors of [4] also map tags into WordNet to extract the distribution of Flickr tags in different conceptual domains. They report that main tag categories include artifacts and places. Wu et al. [8] analyze image content in order to improve tag choice using textual correlations.

2 Query modeling

The proposed retrieval model has two main components: query modeling with Flickr, respectively with Wikipedia. These two components are described in the following subsections subsections.

2.1 Query modeling with Flickr

Flickr is a photo sharing site that contains over 4.8 billion photos as of August 2010. A part of these photos are tagged and these tags can be used to build query models. The 2010 WikipediaMM textual topics, and – more generally – image queries, contain visual cues (terms specific to photography such as close up, black and white), which are not useful during the query modeling stage. A list of photographic terms is extracted from `http://en.wikipedia.org/wiki/Category:Film_techniques` and `http://en.wikipedia.org/wiki/Category:Photographic_processes` (respectively the corresponding categories for French and German) and these terms are removed from the queries. Prepositions are also stripped from the queries and the remaining words are of each topic are used to query the Flickr API and download metadata for up to top 20,000 images associated to the query. The relatedness is defined by counting the photos that are annotated with the respective term. In table 1, we present the top 10 related terms for *fractals*, *tennis player on court* and *cactus in desert*. Most of the terms presented in the table are closely related to the initial query and they constitute an acceptable model of its content. They range from generic terms such as abstract, digital, nature for *fractals* to specific terms such as wimbledon, federer or centre court for *tennis player on court*.

One known problem in information retrieval is that the word form in the queries is not the same as their form in the database and this mismatch hurts recall. One common solution to this problem is to use stemming but this solution has its drawbacks since the stemmed forms of the words can match completely different words, especially when dealing with multilingual datasets. Stemming was only used for the words in the query and was followed by a look-up in the query models in order to find word variants. Related words that have an edit

Topic ID	Topic text	Top related terms
1	fractals	fractal, apophysis, abstract, mandelbrot, romanesco, art, green, digital, cauliflower, nature
8	tennis player on court	wimbledon, tennis court, racket, players, sport, federer, ball, wta, atp, centre court, tournament
39	cactus in desert	arizona, cacti, saguaro, tucson, cholla, sonoran, phoenix, california, barrel cactus, az

Table 1. Flickr query models for English. Top 10 related terms are presented.

distance smaller than 3 with respect to words in the query or terms that begin with the stem of the terms in the query were considered relevant. For instance, the topic *cactus in desert* becomes *cactus:cacti desert* and each word variant will be used in order to search relevant results. Although the discussion here is focused on English queries, the same procedure was applied to French and German versions of the topics.

2.2 Query modeling with Wikipedia

Words in the topic do not cover the entire semantic field of the underlying concepts and many potentially relevant results are ignored. One particularly useful relation for improving the conceptual coverage of the topic is the conceptual inheritance (X *isA* Y). For instance, *Elena Dementieva* or *Rafael Nadal* are *tennis players* and images annotated with their names are potentially relevant the topic *tennis player on court*. To discover semantically similar concepts, I rely on own previous work [3], which exploits the categorical structure of Wikipedia.

The main exploited resource is Wikipedia, which provides its dumps for free use. I downloaded the English version of March 2010, split them into individual files and search for articles which are related to the words in the WikipediaMM topics. Prior to that, topics were preprocessed in order to remove photographic terms and prepositions and to find term variants in the query models (see previous subsection). Also, topic words were run trough WordNet and synonyms were extracted for unambiguous ones (words that have a single WordNet entry). This extraction was limited to unambiguous terms because noise can be introduced by secondary senses of polysemous words. The enriched versions of topics were compared to Wikipedia categories in order to find articles categorized with words in the topic. The number of common terms between the topic and an article's categories represents a rough measure of their semantic similarity. It is defined here as a score between 0 (no terms in common) and 1 (all terms in common) and used to propose a first ranking of Wikipedia articles (first column in table 2). Since many articles share the same coarse grained similarity score, ties are broke by introducing a second score which is directly dependent of the number and frequency of terms from the Flickr query model that appear in the target article and inversely proportional to the log of the length of the article.

This second score was determined empirically and results would be probably improved if a more principled similarity distance was used.

Topic ID	Topic text	Top related concepts (with coarse and fine grained similarity scores)
1	fractals	Fractal antenna (1, 0.125); Fractal cosmology (1, 0.106); Mandelbrot set (1, 0.092); Fractal dimension (1, 0.076); Barnsley fern (1; 0.066); Fractal landscape (1, 0.065); Iterated function system (1, 0.06); Newton fractal (1, 0.06); Fractal art (1, 0.06); Lyapunov fractal (1, 0.051)
8	tennis player on court	Roger Federer (0.667, 1.771); Rafael Nadal (0.667, 1.423); Billie Jean King (0.667, 0.9501); Serena Williams (0.667, 0.95); Juan Martin del Potro (0.667, 0.936); Venus Williams (0.667; 0.859); Jimmy Connors (0.667, 0.859); Ken Rosewall (0.667, 0.795); Maria Sharapova (0.667, 0.777); Pete Sampras (0.667, 0.776); Justine Henin (0.667, 0.734)
39	cactus in desert	Saguaro (1, 0.143); Stenocereus thurberi (1, 0.099); Opuntia ficus-indica (1, 0.091); Pachycereus pringlei (1, 0.08); Cylindropuntia bigelovii (1, 0.068); Opuntia basilaris (1, 0.048); Cylindropuntia fulgida (1, 0.044); Opuntia engelmannii (1, 0.04); Hylocereus undatus (1, 0.022); Wharram Percy (0.5, 4.58)

Table 2. Wikipedia query models obtained obtained using category look-up and Flickr query models. Top 10 related terms are and their similarity scores are presented.

The examples in 2 show that extracted Wikipedia concepts are generally closely related to the initial topics. Also, the example for topic 39 illustrates well the importance of the coarse similarity score because the top 9 concepts (similarity = 1) are related to both terms in the query; whereas the last term, Wharram Percy, a deserted village in England, is only related to desert. The Wikipedia translation graph was used to also get similar Wikipedia concepts for French and German.

Images in the WikipediaMM collection come with brief textual descriptions. In this setting, query expansion is an appealing way to improve recall and, if performed in a judicious way, to also improve results precision.

3 Retrieval experiments

Both textual and multimodal runs were submitted this year. Multimodal runs involved a k-NN inspired visual reranking of textual results and actually degraded the final quality of results. Also, runs were submitted for English only queries and for all three versions of each topic. The multilingual scores of images were formed by adding up scores in individual languages. As with multimodal runs, the overall performance was degraded with respect to English only runs. Therefore, in this section we focus on textual English only runs. Unlike past years'

WikipediaMM campaigns, the texts of the articles that contained collection images were provided to participants. This thorough textual context of the images facilitates the retrieval task. However, since the main purpose of Télécom was to test a retrieval approach in a database with short and noisy textual descriptions, submitted runs were based on the metadata files only and not on article texts.

Given a query, image results from the Wikipedia collection are retrieved by searching for images which are described either by terms in the initial query or by related concepts from Wikipedia. We assume that the relatedness of an image to a query is proportional to the number of words shared by its description in the collection and the query (in its extended form). A score between 0 and 1 is attributed to the image, in function of the number of topic words found. For instance, if the initial query contained three words (tennis player court) and two of them were retrieved (*tennis and player* or *tennis and court*), the image gets a score of 0.667. If a Wikipedia related concept is retrieved in the image's metadata, the corresponding coarse similarity score is attributed to the image, with a small penalization to account for the rank of the Wikipedia concept. For instance, an image annotated with *Roger Federer* gets a score of 0.667 while another image annotated with *Venus Williams* gets a score of 0.666. Relevant images are found by launching queries with the initial terms and the expanded queries in the following order:

- all terms in the initial query and a related concept
- the initial query
- parts of the initial query (starting with largest subparts - and favoring rare terms) and a related concept
- related concept or parts of the initial query

One effect of this type of scoring is that many images have the same ranking score and they need to be separated. To break ties, Flickr query models are used to search related terms in the image description. For two images are annotated with *Venus Williams*, which both have an initial score of 0.666, an image also annotated with *wimbledon* will be ranked higher than an image annotated with *wta* because the first word is more closely related to *Venus Williams* than the second.

To study the influence of the use of Wikipedia related concepts, two runs were submitted:

- telecom_en_flickr_wiki – involves searches with up to top 1000 Wikipedia related concepts, regardless of their coarse similarity score.
- telecom_en_flickr_wiki_maj – is limited to search with only those concepts among the top 1000 Wikipedia related concepts that have a similarity score higher than 0.5. For instance, in the case of *cactus in desert*, only the top nine terms from table 2 are used. This limitation was imposed in order to study the effect of semantic similarity on the retrieval results.

The MAP and precision (@10 and @20) results for the two analyzed runs are presented in table 3. The run that uses only Wikipedia concepts with a high

Overall rank	Run ID	MAP	P@10	P@20
14	telecom_en_flickr_wiki	0.2052	0.4786	0.4236
18	telecom_en_flickr_wiki_maj	0.2227	0.4829	0.4407

Table 3. Wikipedia query models obtained obtained using category look-up and Flickr query models. Top 10 related terms are and their similarity scores are presented.

similarity to the initial query has better scores for both types of measures. The relative improvement of the mean average precision when limiting the usage of related concepts is of 8.5%. This difference supports the hypothesis that query expansion should be performed only with terms that are closely related to the initial query.

4 Conclusion and future work

Here a query modeling technique that involves the use of Wikipedia related concepts but also of Flickr tags is introduced. The role of the Wikipedia concepts is to improve recall when working with scarce annotations. Flickr models are used to enrich initial queries with words that derivations of the words in the initial query, to propose a fine grained ranking of Wikipedia concepts and to break ties between initial scores provided after searching the database with terms in the enriched version of the initial query and Wikipedia concepts.

One interesting future work direction is to replace the ad-hoc concept ranking schemes used in the experiments with more principled methods. Also necessary for the validation of the approach is the comparison to standard retrieval models which are already implemented in freely available indexing frameworks. Third, it would be interesting to evaluate the usage of richer textual descriptions (i.e Wikipedia articles that contain the images or textual windows that surround the images in these articles). As noted, only image metadata were used and, as underlined in [6], these metadata provide only a scarce textual representation of the images in the collection, which probably penalizes the quality of the final results.

5 Aknowledgement

This research is part of the French National Agency for Research (ANR) project Georama (ANR-08-CORD- 009).

References

1. E. Gabrilovich, S. Markovich. Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. In Proc. of IJCAI, 2007
2. D. Milne, I. H. Witten. An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In Proc. of WIKIAI, 2008

3. A. Popescu, H. Le Borgne, P.-A. Moëllic, Conceptual Image retrieval over a Large Scale Database. In Evaluating Systems for Multilingual and Multimodal Information Access, In Proc. of the 9th Workshop of the Cross-Language Evaluation Forum, Lecture Notes in Computer Science, 2009
4. Sigurbjornsson, B., van Zwol, R.. Flickr Tag Recommendation based on Collective Knowledge. In Proc. of WWW 2008 (Beijing, China).
5. M. Strube, S. P. Ponzetto. WikiRelate! Computing Semantic Relatedness Using Wikipedia. In Proc. of AAAI, 2006
6. A. Popescu, T. Tsikrika, J. Kludas. Overview of the WikipediaMM task at ImageCLEF 2010. CLEF working notes, 2010
7. R. H. Van Leuken, L. Garcia, X. Olivares, R. van Zwol. Visual Diversification of Image Search Results. In Proc. of WWW, 2009
8. Wu, L, Hua, X.-S., Yu, N., Ma, W.-Y. and Li, S. 2008. Flickr Distance. Proc. of ACM Multimedia 2008, Vancouver, Canada