Improving Query Expansion for Image Retrieval via Saliency and Picturability

Chee Wee Leong[1], Samer Hassan[1], Miguel Enrique Ruiz[2], and Rada Mihalcea[1]

[1] Department of Computer Science, University of North Texas
cheeweeleong@my.unt.edu, samer@unt.edu, rada@cs.unt.edu
[2] School of Library and Information Management, Emporia State University
mruiz2@emporia.edu

Abstract. In this paper, we present a Wikipedia-based approach to query expansion for the task of image retrieval, by combining salient encyclopaedic concepts with the picturability of words. Our model generates the expanded query terms in a definite two-stage process instead of multiple iterative passes, requires no manual feedback, and is completely unsupervised. Preliminary results show that our proposed model is effective in a comparative study on the ImageCLEF 2010 Wikipedia dataset.

1 Introduction and Motivation

The growth of the Internet has encompassed an enormous increase in the amount of data available in different modalities (e.g., texts, images, symbols, sound and video clips), formats (semi-structured vs unstructured), topics (e.g., politics, sports, entertainment) and languages (English is by far the most common). While different views of web data continued to emerge, the manner in which users specify queries remains largely unchanged. Typically, a user would enter one or more words in a natural language of choice, indicate the search domain (e.g., image search), and proceed to submit the query. Regardless of the type of web data requested, the length of the query is usually short, placing the onus on Information Retrieval (IR) systems to extrapolate beyond the surface forms of the query to extract its underlying semantics. Consequently, a good IR system must be intelligent enough to infer the true intentions of the user using further semantic analysis.

In this paper, we introduce a corpus-based approach of utilizing salient encyclopaedic concepts to expand queries for retrieving images. Our method is unsupervised, requires no manual feedback and involves a definite two-stage process to generate additional query terms that are semantically similar to the original query. We hypothesize that images are better represented using *picturable* words, and model this dimension of picturability in the expansion process using Flickr as a knowledge-base.

The paper is organized as follows. We briefly introduce how two resources, Wikipedia and Flickr, can be used to model the meaning and picturability of a word using *salient* concepts and corpus evidence respectively. After that, we proceed to construct our expansion model, and provide empirical results on a dataset showing its effectiveness.

2 Salient Semantic Analysis

We model the meaning of a word using its associated salient concepts that are linked in articles containing the word, using an approach termed Salient Semantic Analysis (SSA) [2]. The links within Wikipedia articles are regarded as clues or salient features (concepts) within the text that help define and disambiguate its context. By measuring the semantic association between words and salient concepts found in its neighborhood using co-occurrence statistics, SSA creates semantic profiles for words featuring the top concepts associated (co-occurring) with these words in a given window. Let us consider the following snippet extracted from a Wikipedia article:

> *An automobile, motor car or car is a wheeled motor vehicle used for transporting passengers, which also carries its own engine or motor.*

All the underlined words and phrases represent linked concepts, which are disambiguated and connected to the correct Wikipedia article. SSA semantically interpret each term in this example as a vector of its neighboring concepts (instead of words, as in other corpus-based measures). For example the word *motor* can be represented as a weighted vector of the salient concepts *automobile*, *motor car*, *car*, *wheel*, *motor vehicle*, *transport*, and *passenger*.

Formally, given a corpus C with m tokens, vocabulary size N, and concept size W (number of unique Wikipedia concepts), a $N \times W$ matrix (P) is generated representing the pairwise mutual information between each of the corpus terms with respect to its context concepts. The elements of P are defined as follows:

$$P_{ij} = log_2 \frac{f^k(w_i, c_j) \times m}{f^C(w_i) \times f^C(c_j)} \tag{1}$$

where f^k is the number of times the terms w_i and concept c_j co-occur together within a window of k words in the entire corpus.

To calculate the semantic relatedness between two words/texts, A and B, given the constructed matrix, we have:

$$Sim(A,B) = \begin{cases} 1 & Score_{cos}(A,B) > \lambda \\ Score_{cos}(A,B)/\lambda & Score_{cos}(A,B) \leq \lambda \end{cases} \tag{2}$$

where

$$Score_{cos}(A,B) = \frac{\sum_{y=1}^{N}(P_{iy} * P_{jy})^{\gamma}}{\sqrt{\sum_{y=1}^{N} P_{iy}^{2\gamma} * \sum_{y=1}^{N} P_{jy}^{2\gamma}}}, \tag{3}$$

The γ parameter allows the control of weight bias and λ is a normalization factor which help closing the semantic gap between perfect synonyms (tiger-tiger) and near-synonyms (tiger-feline).[3]

[3] In a sense, λ represents a synonymy threshold, a value above which the semantic relationship is considered synonymous.

3 Flickr Picturability

We hypothesize that some words are more *picturable* than others (e.g. banana vs paradigm), i.e., it is easier to find an image to visually describe concepts invoked by the more picturable word. In our work, we attempt to model the picturability of a word using Flickr[4] as a resource.

Following [3], we can model the picturability of a word alone, or in association with other words in text. The latter is of particular interest since we wish to discriminate words based on their picturability in order to observe the corresponding effects on the performance of query expansion models. The algorithm proceeds as follows: given a word in a free text, we use the $getRelatedTags$[5] API to retrieve the most frequent Flickr tags associated with the word, and use them as corpus evidence to compute its picturability score. We disregard stopwords and any word less than three characters long or not found in Flickr tag repository. Next, any retrieved tag appearing as surface forms in the text is rewarded proportionally to its term frequency in the text, as words having a high frequency and featuring more in the Flickr tags are relatively more picturable in the text. Additionally, we score all words that return a non-empty related tags set with a discounted weight (β=0.5) of its term frequency to promote outstanding, picturable words in the rewarding phase. As an illustration, consider the following snippet:

> *On the Origin of Species, published by Charles Darwin in 1859, is considered to be the foundation of evolutionary biology.*

For each of the content words w_i (i.e. $origin$, $species$, $published$, $charles$, $darwin$, $foundation$, $evolutionary$, and $biology$), we query Flickr and obtain their related tag set R_i. The words $origin$, $published$, and $foundation$ return an empty set of related tags and hence are not scored and also removed from our set of consideration, leaving $species$, $charles$, $darwin$, $evolutionary$, and $biology$ with the initial score of 0.5. Next, we score each w_i based on the number of votes it receives from the remaining w_j (Figure 1). Each vote represents an occurrence of the candidate tag w_i in the related tag set R_j of the candidate tag w_j. For example, if $darwin$ appeared in the Flickr related tags for $charles$, $evolutionary$ and $biology$, then it would have a weight of 0.5+1+1+1=3.5.

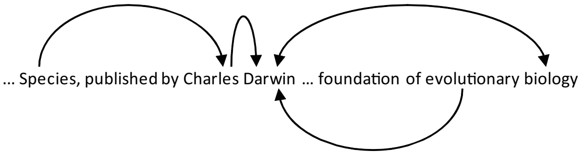
... Species, published by Charles Darwin ... foundation of evolutionary biology

Fig. 1. Flickr Picturability Labels

[4] http://www.flickr.com

[5] http://www.flickr.com/services/api/flickr.tags.getRelated.html

4 Automatic Query Expansion via Saliency and Picturability

In this work, we focus on the task of query expansion for retrieving images. Our model entails a two-stage process, namely, candidates generation and candidates selection, described below.

4.1 Candidate Generation

Given a set of initial query terms Q={q_1,q_2,...,q_k}, we retrieve a list of the top m Wikipedia articles most relevant to Q, computed by summing over individual concept vectors of all query terms in Q using SSA, as described in section 2. A pre-processing step is first applied to remove stopwords from the title of each article. We consider each of the remaining words w_i a candidate for query expansion, weighted using a simple fusion rule:

$$Weight(w_i) = tf(w_i) * 1/rank(w_i) * flickr(w_i) \tag{4}$$

where $tf(w_i)$ is the term frequency of w_i appearing as a word across all m Wikipedia titles, $rank(w_i)$ is the highest rank of the title (reverse sorted) containing w_i, and $flickr(w_i)$ is picturability score provided by corpus evidence using the Flickr picturability method in section 3.

4.2 Candidate Selection

From the generation phase, we select as working set the top ranked W words according to (4) which can be potentially applied to expand Q. We next adopt a bootstrapping procedure to expand Q as follows: for each word w_j traversed in W (reverse sorted), if $Sim(Q, w) \geq \alpha$, we update Q to include w_j, where $Sim(Q, w)$ is provided by SSA in section 2. Our expansion model focuses on extracting picturable terms associated with salient concepts that are semantically similar to the initial query text Q. The weighting scheme in the candidate generation phase ensures that terms strongly associated with these concepts are first used to expand Q, and the expansion is performed incrementally, adding the most relevant terms each time while preserving the overall semantic consistency.

5 Empirical Evaluation

For evaluating our expansion model, we use the data from the ImageCLEF 2010 Wikipedia [6]. This collection includes 237,434 images with associated texts in English, French and German. Approximately 10% of these images are annotated in all three languages, 24% with annotations in two languages, and 62% with annotations in one language. For each image used in the evaluation, we translated the French and German texts in the captions into English using the Bing Translation service. The collection also contains 70 topics written in all three languages. For a fair comparative study, we build our retrieval system using the same specifications provided by the best performing system using exclusively monolingual features (untaTxEn) [7]. Following their approach, we first index all data in the collection using the Indri/Lemur information retrieval system. Next, we build a unigram model with Dirichlet smoothing, Krovetz stemming and a list of English stopwords [5]. During retrieval, each retrieved document, D, is scored as follows:

$$P(Q|D) = \prod_i P(q_i|D)^{\frac{1}{|Q|}} \tag{5}$$

where Q is the query in question and q_i is a query term in Q. Prior to retrieval, we perform query expansion for each query Q using the expansion model explained in section 4, with m=1000 to ensure adequate coverage of concepts, W=50 and α = 1. When measuring similarity, our SSA model is set to γ = 1.2 and λ = 0.02 which are derived from experiments using three manually constructed queries. The topics can be classified into three different tiers of difficulty based on results from teams participating in ImageCLEF. Here, we show examples of automatically expanded query for each tier. {*cockpit of an airplane*} yields {*cockpit airplane compartment aircraft flight plane airplane passenger carrier boeing crash*} (easy), {*dna helix*} yields {*dna helix strand fold protein molecular alpha binding*} (medium), while {*building site*} yields {*building site construction structures*} (hard).

6 Discussion

Table 1 shows the results from our experiments. The performance of each system is reported using a collection of metrics typically used in IR. As observed, our query monolingual expansion model SSA (flickr + ENG) generally yield improvements over the top-performing monolingual querying system (untaTxEn) in ImageCLEF 2010, scoring better in number of relevant images retrieved, MAP and bpref, and is competitive (0.2916 vs 0.3025) on the Rprec metric. The difference in retrieval performance between SSA (flickr+EN) and SSA (flickr+EN+DE+FR) represents the advantage of querying in a multilingual setting, which records improvements in all metrics except for number of relevant images retrieved. Similarly, the difference between SSA (EN+DE+FR) and SSA (flickr+EN+DE+FR) indicates the role played by the flickr picturability component in equation 4. When omitted, the flickr picturability causes a drop in retrieval performance recorded in all metrics. Overall, with the exception of Precision at 20, our expansion models (monolingual or multilingual) scores significantly better performance over the average system participating in CLEF 2010.

Table 1. Retrieval Performance

System	#Relevant	Map	Rprec	bpref	P@20
SSA (flickr + EN)	**8176**	0.2277	0.2916	0.2654	0.3314
SSA (EN+DE+FR)	8057	0.2240	0.2900	0.2655	0.3871
SSA (flickr + EN+DE+FR)	8162	**0.2293**	0.2971	**0.2685**	**0.4114**
untaTxEn	7840	0.2251	**0.3025**	0.2617	0.4057
Average in CLEF2010	5789	0.1611	0.2323	0.2032	0.3582

7 Related Work

Several expansion models based on Wikipedia have sprung into existence recently [1, 4, 8]. The work most closely related to ours is [1], where each query is run using a dependence model to retrieve a list of ranked Wikipedia documents, of which anchor words in lower-ranked documents referencing higher-ranked ones are utilized to expand the query. In contrast, our system is different in a number of aspects. First, we are adopting a lightweight approach that performs word mining from the surface forms of articles titles, rather than analysing entire documents for anchor words. Second, we employ an additional level of semantic relatedness check using SSA to ensure the words in the expanded query are semantically consistent. Finally, our system preferentially selects words that are not only semantically similar to the original query, but also picturable ones. As future work, we plan to address the optimization of system parameters.

References

1. Arguello, J., Elsas, J.L., Callan, J., Carbonell, J.G.: Document representation and query expansion models for blog recommendation. In: Proceedings of the Second International Conference on Weblogs and Social Media (2008)
2. Hassan, S., Mihalcea, R.: Semantic relatedness using salient semantic analysis. In: Proceedings of AAAI Conference on Artificial Intelligence (2011)
3. Leong, C.W., Mihalcea, R., Hassan, S.: Text mining for automatic image tagging. In: Proceedings of the International Conference on Computational Linguistics (2010)
4. Li, Y., Luk, W.P.R., Ho, K.S.E., Chung, F.L.K.: Improving weak ad-hoc queries using wikipedia as external corpus. In: Proceedings of ACM SIGIR conference on Research and development in information retrieval (2007)
5. Ogilvie, P., Callan, J.: Experiments using the lemur toolkit. In: Proceedings of 2001 Text REtrieval Conference (TREC 2001), special publication. National Institute of Standards and Technology (2001)
6. Popescu, A., Tsikrika, T., Kludas, J.: Overview of the wikipedia retrieval task at imageclef 2010. In: CLEF 2010 Labs and Workshops, Notebook Papers. Padua, Italy (2010)
7. Ruiz, M., Chen, J., Pasupathy, K., Chin, P., Knudson, R.: Unt at imageclef 2010: Clir for wikipedia images. In: CLEF 2010 Labs and Workshops, Notebook Papers. Padua, Italy (2010)
8. Xu, Y., Jones, G.J., Wang, B.: Query dependent pseudo-relevance feedback based on wikipedia. In: Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval (2009)