Associating Relevant Photos to Georeferenced Textual Documents through Rank Aggregation

Rui Candeias and Bruno Martins
rui.candeias,bruno.g.martins@ist.utl.pt

Instituto Superior Técnico, INESC-ID
Av. Professor Cavaco Silva, 2744-016 Porto Salvo, Portugal

Abstract. The automatic association of illustrative photos to paragraphs of text is a challenging cross-media retrieval problem with many practical applications. In this paper we propose novel methods to associate photos to textual documents. The proposed methods are based on the recognition and disambiguation of location names in the texts, using them to query Flickr for candidate photos. The best photos are selected with basis on their popularity, on their proximity, on temporal cohesion and on the similarity between the photo's textual descriptions and the text of the document. We specifically tested different rank aggregation approaches to select the most relevant photos. A method that uses the *CombMNZ* algorithm to combine textual similarity, geographic proximity and temporal cohesion obtained the best results.

1 Introduction

The automatic association of illustrative photos to paragraphs of text is a challenging cross-media retrieval problem with many practical applications. For instance the Zemanta[1] blog enrichment extension is a commercial application capable of suggesting photos from Flickr to blog posts. Another example concerns with textual documents describing travel experiences, usually called *travelogues*, which can give interesting information in the context of planning a trip. Today, there are several websites where these documents are shared and the use of web information for travel planning has also increased. However, the use of the travelogues by themselves is very restrictive. It is our conviction that the visualization of photos associated with specific parts from the travelogue, like common scenarios and points of interest, may lead to a better usage of travelogues.

Despite the huge number of high quality photos in websites like Flickr[2], these photos are currently not being properly explored in cross-media retrieval applications. In this paper, we propose methods to automatically associate photos, published on Flickr, to textual documents. These methods are based on mining geographic information from textual documents, using a free web service to

This work was partially supported by the Fundação para a Ciência e a Tecnologia (FCT), through project grant PTDC/EIA-EIA/109840/2009 (SInteliGIS)

[1] http://www.zemanta.com/

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

recognize and disambiguate location names and points of interest mentioned in the documents. The places recognized in the documents are then used to query Flickr for related photos. Finally, the best photos are selected with basis on their popularity and on the similarity between their information (e.g., textual, geographical and temporal metadata) and the information from the document (e.g., textual contents, recognized places and temporal metadata).

The rest of this paper is organized as follows: Section 2 presents the main concepts and related works. Section 3 describes the proposed methods, detailing the mining of geographic information contained in texts and the selection of the best photos, based on their popularity and similarity. Section 4 describes how a system, containing the proposed methods, was implemented. It also presents the results of an initial evaluation experiment. Finally, Section 5 presents our conclusions and points guidelines towards future work.

2 Related Work

Problems related with the treatment of geographic references in textual documents have been widely studied in *Geographic Information Retrieval* [1,11,15,16]. Using this information requires the recognition of place names in the texts (i.e., delimiting the text tokens referencing locations) and the disambiguation of those place names in order to know their real location in the surface of the Earth (i.e., give unique identifiers, typically geospatial coordinates, to the location names that were found). The main challenges in both tasks are related with the ambiguity of natural language. Amitay et al. characterized those ambiguity problems according to two types, namely geo/non-geo and geo/geo [1]. Geo/non-geo ambiguity occurs when location names have a non-geographic meaning (e.g., Turkey, the country or the bird). Geo/geo ambiguity refers to distinct locations with the same name (e.g. London in England and London in Ontario).

Leidner studied different approaches for the recognition and disambiguation of geographic references in documents [11]. Most of the studied methods resolve places references by matching expressions from the texts against dictionaries of location names, and use disambiguation heuristics like default senses (e.g., the most important referenced location is chosen, estimated by the population size) or the spatial minimality (e.g., the disambiguation must minimize the polygon that covers all the geographic references contained in the document). Recently, Martins et al. studied the usage of machine learning approaches in the recognition and disambiguation of geographic references, using Hidden Markov Models in the recognition task, and regression models with features corresponding to the heuristics surveyed by Leidner, in the disambiguation task [15]. Other recent works focused on recognition and disambiguation problems that are particularly complex, involving the processing of texts where geographic references are very ambiguous and with a low granularity (e.g., mountaineering texts mention tracks and specific regions in mountains), and where it is important to distinguish between the location names pertinent to route descriptions and those that are pertinent to the description of panoramas [16].

Currently, there are many commercial products for recognizing and disambiguating place references in text. An example is the Yahoo! Placemaker[3] web service, which was used in this work and is better described in Section 3.1.

Previous works have also studied the usage of Flickr as a *Geographic Information Retrieval* information source [4]. The information stored in this service revealed itself to be useful for many applications, due to the direct links between geospatial coordinates (i.e., the coordinates of the places where the photos were taken, either given by cameras with GPS capabilities or by the authors), dates (i.e., the moments when the photos were taken) and text descriptions that are semantically rich (i.e., descriptions and *tags* associated to photos).

In particular, Lu et al. addressed the automatic association of photos, published on Flickr, to Chinese travelogues [14], with basis on a probabilistic topic model detailed on a previous work [8], which is an extension of the Probabilistic Latent Semantic Indexing (pLSA) method [9]. The main idea in the work by Lu et al. is similar to the basis of our work, as the authors tested different methods for the selection of photos, obtained by querying Flickr's search engine with the location names recognized in the texts. The probabilistic topic model is used by the authors to avoid the gap between the vocabulary used in the documents and the textual descriptions used in photos, modeling photos and/or documents as probabilistic distributions over words. The authors tested four different approaches for the selection of relevant photos, namely (i) a baseline approach based on a simple word-to-word matching with the words from the travelogue texts and the tags that represent the photos (ii) a mechanism based on a probabilistic model created with the travelogue texts (iii) a mechanism based on a probabilistic model created with tags that represent the photos, and (iv) a mechanism based on a probabilistic model using the texts and the tags, which obtained the best results. In our work, we approached the problem in a slightly different way, by querying Flickr with the geospatial information associated with the places recognized in the documents.

In terms of previous works related to the area of cross-media retrieval, Deschacht and Moens presented an approach that tries to find the best picture of a person or an object, stored in a database of photos, using the captions associated to each picture [5]. The authors built appearance models (i.e., language models that represent the text captions from images), to capture persons or objects that are featured in an image. Two types of entity-based appearance models were tested, namely an appearance model based on the visualness (i.e., the degree to which an entity is perceived visually), and another appearance model based on the salience (i.e., the importance of an entity in a text). As baseline approaches, the authors built two simpler appearance models, namely (i) a bag-of-words (BOW) model based on the words of the image captions, and (ii) a bag-of-nouns (BON) model based on the nouns and proper nouns contained in the image captions. From a dataset composed of several image-caption pairs, the authors created two different sets of images annotated with the entities, namely (i) an easy dataset composed of images with one entity, and (ii)

[3] `http://developer.yahoo.com/geo/placemaker/`

a difficult dataset composed of images with three or more entities. The results showed that when the dataset was queried with only one entity, the method using the appearance model based on the visualness achieved the best results. On the other hand, when the query was composed of two entities, the method using the bag-of-words had better results.

3 Automatic Association of Photos to Texts

The proposed method for the automatic association of photos to textual documents is essentially based on a pipeline of three stages, which involves (i) recognizing and disambiguating location names and points of interest referenced in documents, (ii) collecting candidate photos through Flickr's API[4], and (iii) selecting the best photos with basis on their importance and on their similarity (e.g., textual, geographical and temporal) towards the document. In this section we describe the three steps in detail.

3.1 Mining Geographic Information in Documents

In this work, we used the Yahoo! Placemaker web service in order to extract locations and specific points of interest from texts. Placemaker can identify and disambiguate places mentioned in textual documents. The service takes as input a textual document with the information to be processed, and returns an XML document that lists the referenced locations. For each location found in the input document, the service returns also its position in the text, the complete expression that was recognized as the location, the type of location (e.g., country, city, suburb, point of interest, etc.), an unique identifier in the locations database used by the service (i.e., the Where On Earth Identifier - WOEID - used by Yahoo! GeoPlanet[5]), and the coordinates of the centroid that is associated to the location (i.e., the gravity center of the minimum rectangle that covers its geographic area). Also, for each document taken as input, the service returns the bounding box corresponding to the document (i.e., the minimum rectangle that covers all its geographic locations).

3.2 Collecting and Selecting Relevant Photos

The main challenge in collecting and selecting photos relevant to a segment of text is related to the semantic gap between the photo metadata and the text, as well as the noise present in the documents and in the descriptions of the photos. For instance, in the case of travelogues, and despite the fact that these documents have a uniform structure, their authors frequently mention information related to transportation and accommodation, and not only descriptions of the most interesting locations. For example, if the text of a travelogue mentions an airport

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

[5] http://developer.yahoo.com/geo/geoplanet/

or the city where the trip ends, while describing the arrival, one can select photos related to these locations, which are not important for illustrating the most interesting contents of the document. We have that travelogues frequently mention locations that are only slightly relevant, and so it is very important to distinguish between relevant and irrelevant locations.

Other challenges in collecting and selecting relevant photos are related with the fact that photos published in Flickr are frequently associated to tags or textual descriptions irrelevant to their visual contents (e.g., tags are usually identical among different photos uploaded by the same person, at the same time), and also the vocabulary used in Flickr can be very different from the vocabulary used in textual documents.

Having these limitations in mind, we tested different approaches for the selection of relevant photos, combining different sources of evidence for estimating the relevance of the photos. These approaches are as follows:

T1: **Selection based on textual similarity:** We compute the textual similarity between the tags plus the title of the photos, and the text of the document. Specifically, we compute the cosine measure between the textual descriptions of the photos (i.e., joining tags and title) and the textual document, using the Term Frequency × Inverse Document Frequency (TF-IDF) method to weight terms in the feature vectors. The idea behind this method is that, if a photo has textual descriptions more similar to the text of a document, then it can be considered as a good photo to be associated to the document.

T2: **Selection based on textual similarity and geographical proximity:** We combined the textual similarity from T1 with the similarity, based on the geospatial coordinates, between the locations recognized in the document and the locations where photos were taken. The geographical similarity is computed according to the formula $\frac{1}{(1+d)}$, where d is the great-circle distance between the two locations. Because multiple locations can be recognized in the document, we computed the maximum and the average similarity towards each photo. The idea behind this method is that a photo that was taken near a location recognized in the document can be considered as a good photo to be associated to the document.

T3: **Selection based on textual similarity, geographical proximity and temporal cohesion:** We combine the method from T2 with the temporal distance, in semesters, between the publication date of the document and the moment when a photo was taken. Similarly to what is done in method T2, the temporal similarity is computed according to the formula $\frac{1}{(1+t)}$, where t id the number of semesters separating the photo from the document. The idea behind this method is that a photo taken in a moment close to the date when the document was written can often be considered as a good photo to be associated to the document.

T4: **Selection based on textual similarity, geographical proximity, temporal cohesion and photo interestingness:** We combine the method T3 with other information related to the interestingness of the photos (e.g., the number of comments and the number of times other users considered the photo as a favorite). In this case, if a photo was taken in a location inside the bounding box of the document (i.e., the bounding box that contains all locations), then the number of comments and the number of times a photo was marked as favorite are considered as features, and otherwise these features assume the value of minus one. The idea behind this method is that a photo that was taken near the locations recognized in the document, and that is considered an interesting photo due to the number of comments and the number of times users marked it as a favorite, can be considered a good photo to be associated to the document.

The above combination approaches were based on the usage of rank aggregation schemes to combine the multiple features. Specifically, two approaches were considered, namely the *CombSUM* and the *CombMNZ* methods originally proposed by Fox and Shaw [7]. Both *CombSUM* and *CombMNZ* use normalized sums when combining the different features. To perform the normalization, we applied the min-max normalization procedure to the scores of the individual features, which is given by Equation 1.

$$V_{normalized} = \frac{V - min}{max - min} \tag{1}$$

The *CombSUM* score of a photo p, for a given document D, is the the sum of the normalized scores received by the photo in each of the k individual rankings, and is given by Equation 2.

$$CombSUM(p, D) = \sum_{j=1}^{k} score_j(p, D) \tag{2}$$

Similarly, the *CombMNZ* score of a photo p for a given document D is defined by Equation 3, where r_e is the number of non-zero similarities.

$$CombMNZ(t, P) = CombSUM(t, P) \times r_e \tag{3}$$

For measuring the similarity between the textual description of the photos and the text of the document, in all the above methods, stopwords were first removed. To calculate the cosine measure between the photos textual descriptions and the document, using the Term Frequency × Inverse Document Frequency (TF-IDF) method, we considered tags to be more important to describe the photo, followed by the title. Thus, we applied different weights for the different types of textual descriptions, weighting the tags as twice more important.

4 Validation Experiments

We implemented a prototype system based on the techniques described in the previous section, using the Qizx[6] XQuery engine as an execution environment.

[6] http://www.xmlmind.com/qizx/

This XQuery engine supports the latest version of the standard, together with the XQuery Full Text extension to perform full-text search with the cosine measure and TF/IDF vectors, in collections of XML documents.

In order to validate the proposed methods, we created a corpus of 450 photos downloaded from Flickr, with geographical information and a sufficiently large textual description (i.e., more then 100 words and containing location names or points of interest). We used expressions frequently used in travelogues, such as *monument*, *vacation*, *trip* or *castle* to filter the photos collected from Flickr. The collected photos were taken in a point contained in the bounding box corresponding to the geospatial footprint of one of the world's most visited cities[7]. Also, the considered photos were taken in a date from 2000-01-01 to 2010-05-01. For each photo, the number of comments and the number of times it was considered as favorite by other users were also collected.

In order to conduct the experiments, we needed a collection of documents with relevance judgments for photos, i.e., a correct relevant photo associated to the document. This collection was not already available and creating a collection of travelogue documents, illustrated with Flickr photos that had been manually selected by human experts, would be extremely time consuming, also implying some knowledge about the locations described in the documents. This collection is not already available, and creating a collection of photos from Flickr selected and associated by experts to travelogues would be extremely time consuming, and would imply a certain knowledge of the city to where the travel was made.

The photo descriptions from Flickr, with the above characteristics, are fairly good examples of documents with relevance judgments, because the owner considered the photo as a relevant example to be associated to the large textual description. So, for the purpose of our experiments, we considered the textual descriptions as representations of textual documents having the same characteristics as travelogues, and the photos from which the textual descriptions were taken as the relevant photos that should be automatically associated.

The prototype system, implementing different configurations for the proposed method, was then used to process the documents, associating them to relevant photos. The configurations used are described in Section 3.2.

With the results for each document, and considering all four possible configurations with the two voting schemes, we used the `trec_eval` evaluation tool to evaluate the matchings between photos and documents. Figure 1 presents the results obtained in terms of Precision at position 1 (Precision@1), and in terms of the Reciprocal Rank, in the all the considered cities. The horizontal lines represent the mean value of Reciprocal Rank, in red, and the mean value of Precision@1, in blue, for all the considered cities and when using the best configuration. In all the charts, the bar in red, full colored, represents the value of Reciprocal Rank, and the bar in blue, with a shaded color, represents the value for the metric of Precision@1.

The graphics show that method T3 using the *CombMNZ* approach (i.e., T3-MNZ) outperforms method T1 in all the cities. These results suggest that the

[7] `http://en.wikipedia.org/wiki/Tourism`

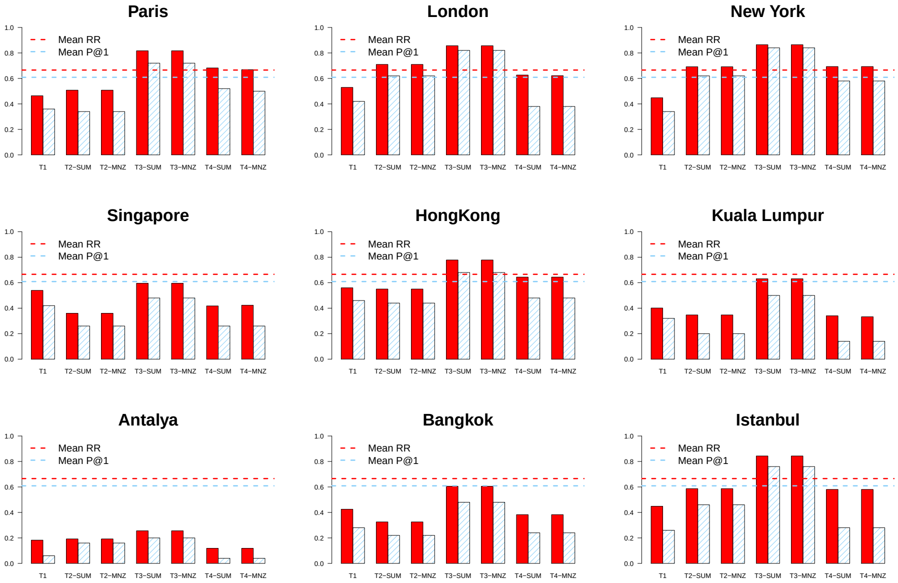


Fig. 1. Reciprocal Rank and Precision@1 for each method and city.

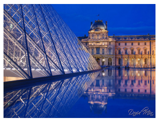
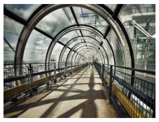
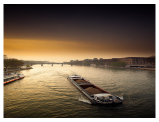
usage of multiple features (e.g., geographical proximity and temporal cohesion) combined with the textual similarity is better then the usage of the textual similarity alone. Also, methods using *CombMNZ* as the rank aggregation approach have similar results to the methods using *CombSUM*.

It is also interesting to notice that the values in the cities of Paris, London and New York are higher, although the dataset contained an equal number of photos for each city (i.e., 50 photos). In these cities, all the combination methods using *CombMNZ* outperform method T1. These results suggest a higher precision of Placemaker in the recognition and disambiguation of the location names mentioned in the descriptions for those cities, although it should be noticed that textual similarity alone also presents good results in these cities.

Figure 2 illustrates the obtained results for two example textual descriptions, presenting the top-3 most relevant photos as returned by the best performing method, together with their tags in Flickr.

Figure 3 presents the number of documents, in the collection, containing each possible number of words, and the number of documents mentioning different numbers of places. In the collection, there is a higher number of documents with 100 to 200 words. Also, the number of recognized places is frequently low, with most of the documents containing 1 to 5 places.

Figure 4 illustrates the relationships existing between the values of Precision@1 and Reciprocal Rank, with the number of words and the number of places, when considering the combination method that had the best results, i.e.,

The **Louvre Pyramid** is a large glass and metal **pyramid**, surrounded by three smaller pyramids, in the main courtyard of the **Louvre Palace** in **Paris**. The large **pyramid** serves as the main entrance to the **Louvre Museum**. Completed in 1989, it has become a landmark for the city of **Paris**. The construction of the **pyramid** triggered considerable controversy because many people felt that the futuristic edifice looked quite out of place in front of the **Louvre Museum** with its classical architecture.			
	id=3756841917	**id=3784201917**	**id=3418929006**
	louvre **paris** **pyramid** **palace** **europe** **vacation** **night** **reflections** **summer** **museum**	**paris** **france** **pompidou** **frontpage** **longexposure** **centrepompidou** **longexposure** **nikon** **guidomusch** **parijs**	**paris** **france** **colour** **seine** **boat** **filters** **lee** **leefilters** **sunset** **light**

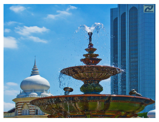
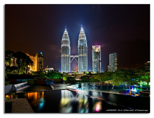
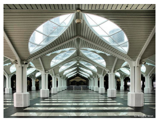
Water, without which, we would not be on Earth. This was captured to show that **Kuala Lumpur** is a blend of old **buildings** as well as new. This **fountain** sits on one end of the famous Selangor Club field used for Merdaka Day Celebrations. Reggie Wan of Singapore and I were here on this bright and hot day. Couldn't get out fast enough away from this **tourist spot**! Actually, all the photos showcased here were pretty nice and I really couldn't decide which one to be the main picture. I chose this one because it was more artistic (the **fountain** is dark, the tall **building** medium grayish and the minaret is white surrounded by blue)!			
	id=294945309	**id=4194402499**	**id=327607369**
	water **fountain** **buildings** **sky** **tourist** **aqua** **cityscape** **reflections** **jalanraja** **fujifilm**	**kuala** **lumpur** **petronas** **explore** **digital** **blending** **dynamic** **nikon** **range** **malaysia**	**building** **kualalumpur** **malaysia** **soe** **southeastasia** **reggiewan** **asia** **top20travelpix** **klcc** **mosque**

Fig. 2. The top three most relevant photos returned for two example documents.

T3 using *CombMNZ*. These results suggest that a higher number of words does not improve the results, neither in terms of Precision@1 or Reciprocal Rank. The higher value of Reciprocal Rank and Precision@1 in documents with 1200 to 1300 words can be explained by the corresponding small number of documents (i.e., only 2 documents). It is also interesting to notice that the values for Precision@1 and for the Reciprocal Rank seem to improve when more than one place is referenced in the document.

5 Conclusions and Future Work

In this paper, we have described novel methods for the automatic association of photos to textual documents. The described methods are based on a pipeline of three steps, in which geographic references are first extracted from documents,

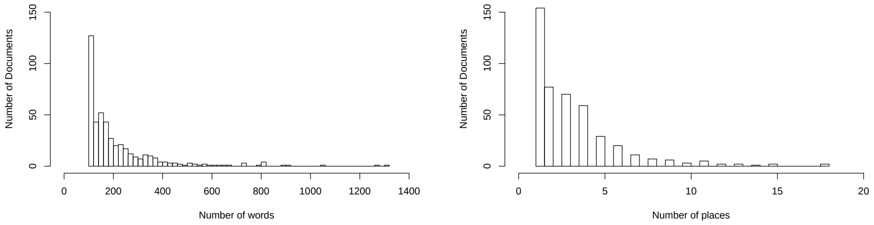


Fig. 3. Histograms with the number of words and the number of places.

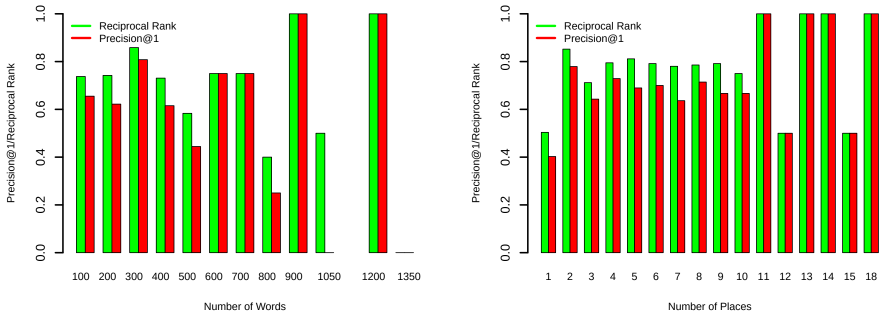


Fig. 4. Variations in the values of Precision@1 and Reciprocal Rank, in terms of the number of words and the number of places referenced in the documents.

then photos matching the geographic references are collected, using Flickr's API, and finally the best photos are selected with basis on their similarity and relevance. Different methods to select relevant photos were compared and a method based on the combination of textual similarity, geographic proximity and temporal cohesion, using the *CombMNZ* rank aggregation method for performing the combination, obtained the best results.

Despite the good results from our initial experiments, there are also many challenges to future work. From our point of view, the major challenge lies in improving the evaluation protocol. The validation of the proposed methods should be made through a collection of static photos, with relevance judgments clearly established by humans. The Content-based Photo Image Retrieval (CoPhIR) collection, described in [3] and built from 106 million photos from Flickr, could be a starting point for building such a test collection. Another idea is to experiment the proposed methods in a collection not related to the domain of travelogues. For instance, the dataset with news texts from BBC which was described by Feng and Lapata [6], containing approximately 3400 entries and where each entry is composed by a news document illustrated with a image that contains a textual caption, could also be used to as a starting point to build a better test collection to evaluate our method. This corpus contains near 3400 entries, where each entry is composed by a news document, a news image related with the document and its caption. Also, besides the usage of the cosine similarity to measure the textual similarity between photos and documents, it would be interesting to use different methods, for instance based on probabilistic topic models such as the Latent Dirichlet Allocation (LDA) model [2].

It would also be interesting to experiment with supervised learning methods for combining the different relevance estimators. Several supervised learning to rank methods [13,12], recently proposed in the information retrieval community to address the problem of ranking search engine results, could be used to develop models that can sort photos based on their relevance, considering different sources of evidence (i.e., several similarity and importance metrics). Recent works in the area of information retrieval have also described several advanced unsupervised learning to rank methods, capable of outperforming the *CombSUM* and *CombMNZ* approaches. This is currently a very hot topic of research and, for future work, we would for instance like to experiment with the ULARA algorithm, which was recently proposed by Klementiev et al. [10].

References

1. E. Amitay, N. Har'El, R. Sivan, and A. Soffer. Web-a-where: geotagging web content. In *Proceedings of the 27th Annual international ACM SIGIR Conference on Research and Development in information Retrieval*, 2004.
2. D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3, 2003.
3. P. Bolettieri, A. Esuli, F. Falchi, C. Lucchese, R. Perego, T. Piccioli, and F. Rabitti. Cophir : A test collection for content-based image retrieval. Technical report, Institute of Information Science and Tecnologies, National Reasearch, Pisa, Italy, 2009.
4. D. J. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg. Mapping the world's photos. In *Proceedings of the 18th international conference on World Wide Web*, 2009.
5. K. Deschacht and M. Moens. Finding the best picture: Cross-media retrieval of content. In *Proceedings of the 30th European Conference on Information Retrieval*, 2008.
6. Y. Feng and M. Lapata. Automatic image annotation using auxiliary text information. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2008.
7. E. A. Fox and J. A. Shaw. Combination of multiple searches. In *Proceedings of the 2nd Text Retrieval Conference*, 1994.
8. Q. Hao, R. Cai, X. Wang, J. Yang, Y. Pang, and L. Zhang. Generating location overviews with images and tags by mining user-generated travelogues. In *Proceedings of the 17th ACM international Conference on Multimedia*, 2009.

9. T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd Annual International ACM SIGIR conference on Research and development in information retrieval*, 1999.
10. A. Klementiev, D. Roth, K. Small, and I. Titov. Unsupervised rank aggregation with domain-specific expertise. In *Proceedings of the 21st International Joint Conference on Artifical intelligence*, 2009.
11. J. Leidner. *Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding of Place Names*. PhD thesis, Institute for Communicating and Collaborative Systems, School of Informatics, University of Edinburgh, 2007.
12. H. Li. *Learning to Rank for Information Retrieval and Natural Language Processing*. Morgan & Claypool Publishers, 2011.
13. T. Liu. Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, 3(3), 2009.
14. X. Lu, Y. Pang, Q. Hao, and L. Zhang. Visualizing textual travelogue with location-relevant images. In *Proceedings of the 2009 international Workshop on Location Based Social Networks*, 2009.
15. B. Martins, I. Anastácio, and P. Calado. A machine learning approach for resolving place references in text. In *Proceedings of the 13th AGILE International Conference on Geographic Information Science*, 2010.
16. M. Piotrowski, S. Liubli, and M. Volk. Towards mapping of alpine route descriptions. In *Proceedings of the 6th ACM Workshop on Geographic information Retrieval*, 2010.