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 geographic information retrieval methods to automatically associate photos to textual documents. The proposed methods are based on the recognition and disambiguation of location names in the texts, using the corresponding geospatial coordinates to query Flickr for candidate photos. The photos are also clustered with basis on visual similarity, and we use relevance estimators obtained from both individual images and image clusters. 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 documents. We tested different learning to rank approaches to combine a rich set of features capturing photo relevance. A method that uses the Coordinate Ascent algorithm to combine term-level textual similarity, topical 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\textsuperscript{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 informations in the context of planning a trip. Today, there are several websites where these documents are shared and the use of web information for planning travels 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 available in websites like Flickr\textsuperscript{2}, these photos are currently not being properly explored in cross-media retrieval applications. In this paper, we propose novel geographic information retrieval 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 service to recognize and disambiguate location names and points of interest from the mentioned documents. The places recognized in the documents are then used to query Flickr for related photos. The photos are clustered together with basis on their visual similarity, and we use relevance estimators obtained from both the individual photos and from the visual image clusters, measuring the association between the photos and the textual documents. 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, from individual images and from image clusters) and the information from the document (e.g., textual contents, recognized places and temporal metadata). To combine the multiple relevance estimates, we experimented with the usage of state-of-the-art learning to rank methods. Experimental results attest for the adequacy of the proposed approaches, showing that a method that uses the Coordinate Ascent algorithm to combine term-level textual similarity, topical similarity, geographic proximity and temporal cohesion can lead to values of 0.73 and .0.83, respectively in terms of the Precision@1 and Reciprocal Rank metrics.

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 the validation methodology and the obtained results. Finally, Section 5 presents our conclusions and points directions for future work.

2. RELATED WORK
Typical problems related with the treatment of geographic references in textual documents have been widely studied in the area of Geographic Information Retrieval [1, 16, 21, 24]. Using this information requires the recognition of place names in texts (i.e., delimiting the text tokens referencing locations) and the disambiguation of those place names in

\textsuperscript{1}http://www.zemanta.com/

\textsuperscript{2}http://www.flickr.com
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. Anitay 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., the word Turkey can refer to either the country or the bird). Geo/geo ambiguity refers to distinct locations with the same name (e.g., London, a city in England or in Ontario).

Leidner studied different approaches for the recognition and disambiguation of geographic references in text [16]. 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 correct disambiguation should 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 SVM regression models with features corresponding to the heuristics surveyed by Leidner, in the disambiguation task [21]. Other recent works focused on place reference 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 mentioning 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 [24].

Currently, there are also many commercial products for recognizing and disambiguating place references in text. An example is the Yahoo! Placemakerootnote{http://developer.yahoo.com/geo/placemaker/} 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 [6]. The information stored in this service revealed itself to be very useful for many different 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 manually set 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 [20], with basis on a probabilistic topic model detailed on a previous work [10], which was an extension of the Probabilistic Latent Semantic Indexing (pLSA) method [11]. 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 topics, which in turn can be seen 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, starting by querying Flickr with the geospatial information associated with 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 [7]. The authors built appearance models (i.e., language models that represent the image textual captions), 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) 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.

Coelho and Ribeiro approached the task of finding suitable images to illustrate text, from specific news stories to more generic blog entries, through a content-based multimedia retrieval technique based on a three-stage process involving (i) textual search, (ii) score filtering and (iii) visual clustering [5]. The system was tested on the SAPO-Labs media collection, containing photos for known personalities, and on the MIRFlickr-25000 collection, with photos and user tags collected from Flickr. Visual content was described by the Joint Composite Descriptor and indexed by a Permutation-PreFix Index. The obtained MIRFlickr results correspond to precision values above 70%, while the SAPO-Labs personality search fared worse, with precision values below 40%.

Previous works in information retrieval, and also in geographic information retrieval [22], have addressed the usage of supervised machine learning for developing search engine ranking formulas, combining multiple estimators for docu-
ment relevance in an optimal way. Both Tie-Yan Liu [19] and Hang Li [17] presented good surveys on the subject of learning to rank for information retrieval (L2R4IR), categorizing the existing algorithms into three groups, according to their input representation and optimization objectives:

- **Pointwise approach** - The L2R4IR task is seen as either a regression or a classification problem. Given feature vectors of each single resource (e.g., a photo) from the data for the input space, the relevance degree of each of those individual resources is predicted with scoring functions. Through these scores we can sort resources and produce the final ranked list. Several pointwise methods have been proposed, including Multi-class Classification for Ranking (McRank) [18].

- **Pairwise approach** - The L2R4IR task is seen as a binary classification problem for the pairs of resources to be ranked, since the relevance degree can be regarded as a binary value telling which ordering is better for a given pair of resources. Given feature vectors of pairs of resources from the data for the input space, the relevance degree of each resource can be predicted with scoring functions which try to minimize the average number of misclassified resource pairs. Several different pairwise methods have been proposed, including SVMrank [14], RankNet [4] or RankBoost [9].

- **Listwise approach** - The L2R4IR task is addressed in a way that takes into account an entire set of resources, associated with a query, as instances. These methods train a ranking function through the minimization of a listwise loss function defined on the predicted list and the ground truth list. Given feature vectors of a list of resources from the data for the input space, the relevance degree of each of those documents can be predicted with scoring functions which try to directly optimize the value of a particular information retrieval evaluation metric averaged over all queries in the training data [19]. Several different listwise methods have also been proposed, including SVMmap [28], AdaRank [26] or Coordinate Ascent [23].

In this paper, we made experiments with the application of state-of-the-art learning to rank algorithms available through the RankLib\(^4\) open-source Java package, namely, RankBoost and Coordinate Ascent. We modeled the task of selecting relevant photos to associate to textual documents as a learning to rank problem where we represent the associations between documents and photos as feature vectors containing multiple relevance estimators (e.g., the textual similarity between the document and the tags from the photo), and use training data to learn a ranking model capable of sorting photos in relation to a given document in a way that optimizes the Reciprocal Rank evaluation metric.

### 3. ASSOCIATING PHOTOS TO TEXTS

The proposed method for the automatic association of photos to textual documents is essentially based on a pipeline of four stages, which involves (i) recognizing and disambiguating location names and points of interest referenced in documents, (ii) collecting candidate photos through Flickr’s API\(^5\), (iii) clustering photos with a basis on visual image similarity, and (iv) selecting the best photos with a basis on their importance and on their similarity (e.g., textual, geographical and temporal similarity, computed from the individual photo and from the corresponding image cluster) towards the document. This section details the previous four steps.

#### 3.1 Mining geographic data 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 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\(^6\)), 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 Finding relevant photos

The main challenges in collecting and selecting photos relevant to a segment of text are related to the semantic gap between the photo metadata and the text, as well as with 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 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 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, titles or textual descriptions that are irrelevant to their actual contents (e.g., titles and 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, attempting to combine multiple estimators of relevance in order to improve results. These approaches are as follows:

\(^4\)http://www.cs.umass.edu/~vdang/ranklib.html

\(^5\)http://www.flickr.com/services/api/

\(^6\)http://developer.yahoo.com/geo/geoplanet/

T1: Selection based on a baseline textual similarity metric: 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, in terms of having many common words, then it can be considered as a good photo to be associated to the document.

T2: Selection based on rich textual similarity features: We compute the textual similarity between the tags plus the title of the photos, and the text of the document, using a rich set of textual similarity features that goes beyond the baseline TF-IDF metric from method T1. We represent the textual contents of the documents and the textual metadata for the photos as either term vectors or as probability distributions over a set of topics, afterwards measuring similarity with basis on these representations. In the case of the topic-based representations, they are obtained through the application of the Latent Dirichlet Allocation (LDA) topic modeling technique [2] as implemented in the LingPipe package.

We build the LDA model in an offline pre-processing step, afterwards computing Bayesian point estimates for the topic distribution of documents and photo descriptions through Gibbs sampling. The complete set of textual features is as follows:

- Two features capturing (i) the number of terms in the textual descriptions of the photos (i.e., joining tags and title), and (ii) the number of terms in the textual document.
- The number of terms shared between the descriptions of the photos (i.e., joining tags and title) and in the textual document, normalized according to the number of terms in the textual descriptions of the photo (i.e., a term frequency feature).
- The cosine similarity measure between the term vectors corresponding to the descriptions of the photos and the textual document, using the Term Frequency × Inverse Document Frequency (TF-IDF) method to weight individual terms.
- The cosine measure between the topic vectors obtained through LDA from the descriptions of the photos and from the textual document. Each of the 25 dimensions on these vectors contains a score representing the probability of the photo or document belonging to that particular topic.
- A symmetrized form of the Kullback-Leibler divergence computed between the probability distributions for the topics present in the descriptions of the photos and in the textual document.

A binary feature indicating if the most probable topic present in the description of the photo is the same as the most probable topic obtained for the textual document.

By representing photos and documents in terms of the underlying topics, instead of just tags and words, we hope to avoid the semantic gap between the vocabulary used in the textual documents and the vocabulary used to tag photos, similarly to Lu et al. [20].

T3: Selection based on rich textual similarity features and sentimental polarity match: We combined the rich textual similarity from T2 with features based on the sentimental mood associated to both the documents and the photos. A simple opinion mining method, based on counting the number of words from the document appearing in a subjectivity lexicon, was used to estimate the polarity of the textual document, assigning it to a value ranging from minus one (i.e., negative sentimental mood) to plus one (i.e., positive mood). A similar value is computed from the visual contents of the candidate image, assigning it to a score ranging from minus one (i.e., cooler images, were the blue component dominates over the red component) to plus one (i.e., warmer image, where the red component dominates over the blue component). Under the assumption that cooler images should be assigned to documents having a negative sentimental polarity, and warmer images should be assigned to documents having a positive sentimental polarity, we compute the absolute difference between the values computed from the document d and the image i. The actual formula is shown below.

\[
\text{match}(d, i) = \| \frac{|\text{words}(d, +)| - |\text{words}(d, -)|}{|\text{words}(d)|} - \frac{\sum_{p \in r} p \cdot (r - p)}{\sum_{p \in B}} \|
\]

In the formula, the function words() returns the number of words from the textual document matching the polarity lexicon, and \( p \) and \( B \) denote the RGB representation of a pixel from image \( i \), from where we take the \( R \) and \( B \) components.

T4: Selection based on rich textual similarity features, sentimental polarity and geographical proximity: We combined the textual and polarity similarity features from T3 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 geodesic 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.

T5: Selection based on rich textual similarity, sentimental polarity, geographical proximity and temporal cohesion: We combine method T4 with the temporal distance, in terms of semesters, between the publication date of the document and the moment
when a photo was taken. Similarly to what is done in method T3, we use a temporal similarity metric computed according to the formula \(\frac{1}{1+t^2}\), 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 be considered as a good photo to be associated to the document.

**T6: Selection based on rich textual similarity, sentimental polarity, geographic proximity, temporal cohesion and photo interestingness:** We combine method T5 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, 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 high number of comments and the high number of times users marked it as a favorite, can be considered a good candidate photo to be associated to the document.

**T7: Selection combining individual photo similarity, polarity and interestingness, together with features computed from visual image clusters:** We group together images with similar visual characteristics (i.e., with similar colors in the same regions) and compute the similarity between these groups of photos and the textual document, combining these similarity scores with the features from method T6. Specifically, we use the \(K\)-means clustering algorithm together with a simple visual similarity metric based on local color features to group the photos taken in a location inside the bounding box of the document (i.e., the bounding box that contains all locations) into \(K\) clusters. The algorithm for computing the visual similarity starts by normalizing image dimensions to 300x300 pixels, afterwards computing 25 \(\times\) 3 features corresponding to the average of the RGB values on 25 regions with 30x30 pixels each, uniformly sampled from the image. To compute the similarity between two images \(A\) and \(B\), we take the 25 regions from each image, compute the Euclidean distance between the regions, and accumulate the values, afterwards normalizing the aggregated distance \(d\) to a similarity score according to the formula \(\frac{1}{1+d^2}\). For photos coming from a location outside the bounding box of the document, these features assume the value of minus one. The number of clusters \(K\) is determined through the method described in [25], which involves choosing a value that maximizes inter-cluster similarity and minimizes intra-cluster similarity. The photo is assigned to the most similar cluster and the following set of features is then computed:

- The total number of terms in the set of all textual descriptions from the photos belonging to the most similar cluster.
- The number of terms shared between the descriptions of all the photos belonging to the most similar cluster, and the textual document, normalized according to the number of terms in the full set of textual descriptions for the photos.
- The cosine similarity measure between the term vectors corresponding to the descriptions of all the photos belonging to the most similar cluster, and the textual document, using the Term Frequency \(\times\) Inverse Document Frequency (TF-IDF) method to weight individual terms.
- The average similarity, based on the geospatial coordinates, between the locations recognized in the document and the centroid location for the most similar cluster towards the photo. The geographical similarity is computed according to the formula \(\frac{1}{1+d^2}\), where \(d\) is the average geodesic distance between the two locations.
- The maximum similarity, based on the geospatial coordinates, between the locations recognized in the document and the closest location associated to one of the photos from the most similar cluster. Similarity here is also computed according to the formula \(\frac{1}{1+d^2}\), where \(d\) is the minimum geodesic distance between the two locations.
- The average temporal distance, in semesters, between the publication date of the document and the moments when the photos from the most similar cluster were taken. We use the average of the temporal similarities computed according to the formula \(\frac{1}{1+t^2}\), where \(t\) is the number of semesters separating photos from the document.
- The maximum temporal similarity, computed in terms of semesters as in the case of the previous feature, between the publication date of the document and the moments when the photos from the most similar cluster were taken.

The above features computed from the most similar cluster are used together with those from method T6, with the intuition that visual clusters can be used to propagate relevance information across similar images, thus avoiding the data sparseness issues commonly associated to Flickr tags. If we can group together images depicting the same monuments and the same points of interest, then we can propagate tags and other metadata information between these visual clusters of images, hopefully improving the results.

The above combination approaches were based on the usage of learning to rank algorithms to combine the multiple features. The following subsection details the considered learning to rank algorithms.

One should also refer that when processing the textual descriptions of the photos and the text of the documents, stopwords were first removed, followed by stemming process and...
tags were considered to be more important than titles to describe the photos. Thus, we applied different weights for the different types of textual descriptions, weighting the tags as twice more important (i.e., representing the photos through virtual text documents where the text from the tags is repeated twice, and the text from the title appears only once).

### 3.3 Learning to rank photos

As previously stated, we experimented with the application of state-of-the-art learning to rank algorithms available through the RankLib open-source Java package, namely:

- **RankBoost** - This is a pairwise boosting technique for ranking [9]. Training proceeds in rounds, starting with all the pairs of photos being assigned to an equal weight. At each round, the learner selects the weak ranker that achieves the smallest pairwise loss on the training data, with respect to the current weight distribution. Pairs that are correctly ranked have their weight decreased and those that are incorrectly ranked have their weight increased, so that the learner will focus more on the hard samples in the next round. The final model is essentially a linear combination of weak rankers. Weak rankers can theoretically be of any type, but they are most commonly chosen as a binary function with a single feature and a threshold. This function assigns a score of 1 to a document of which feature value exceeds the threshold, and 0 otherwise. In our experiments, the number of threshold candidates was set to 50 and the number of rounds to train was set to 1000.

- **Coordinate Ascent** - This is a state-of-the-art list-wise method, originally proposed by Metzler and Croft, which uses coordinate ascent to optimize the ranking model parameters [23]. Coordinate ascent is a well-known technique for unconstrained optimization, which optimizes multivariate objective functions by sequentially doing optimization in one dimension at a time. The method cycles through each parameter and optimizes over it, while fixing all the other parameters.

We modeled the task of selecting relevant photos to associate to textual documents as a learning to rank problem where we represent the associations between documents and photos as feature vectors containing multiple relevance estimators (e.g., the textual similarity between the paragraph and the tags plus the title from the photo), and used training data to learn a ranking model capable of sorting photos in relation to a given document in a way that optimizes the reciprocal rank (RR) evaluation metric.

Previous authors have argued that some evaluation metrics are more informative than others, and thus the target metrics used in the optimization of learning to rank models may act as bottlenecks that summarize the training data. Yilmaz and Robertson challenged the assumption that retrieval systems should be designed to directly optimize for a metric that is assumed to evaluate user satisfaction, showing that even if user satisfaction can be measured by a metric X, optimizing the model for a more informative metric Y may result in a better test performance according to X [27]. This is particularly true in cases when the number of features is large and when we only have access to a limited amount of training data. Experiments confirmed this hypothesis, although we do not report those experiments here for constraints in terms of space.

### 4. Validation Experiments

In order to validate the proposed methods, we created a collection with 900 photos downloaded from Flickr, containing geographical information and a sufficiently large textual description (i.e., more than 100 words and containing location names or points of interest). We used expressions frequently used in travelogues, such as monument or vacations 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 18 most visited cities[^10] (i.e., a set of 50 photos was collected from each of the 18 cities, all containing large textual descriptions and associated to geospatial coordinates). Also, the considered photos were taken in a date ranging from 2000-01-01 to 2010-06-01. For each photo, the number of comments, the number of times it was considered as favorite by other users and the number number of times other users accessed the photo were also collected.

In order to conduct the experiments, we needed a collection of documents with relevance judgments for photos, i.e., a correct photo associated to each document. 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 photo 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 with basis on their titles, tags, geospatial coordinates and popularity information derived from comments and visits.

A prototype photo association system, implementing different configurations for the feature vectors and also using one of the considered learning to rank algorithms, was then used to to process the documents, associating them to relevant photos. The configurations for the feature vectors are described in Section 3.2, and Section 3.3 detailed the considered learning to rank algorithms.

The 50 textual descriptions associated to each of the 18 cities were randomly split into two groups, and we used $25 \times 18$ textual descriptions for building a training dataset for the learning to rank algorithms, and the other $25 \times 18$ textual descriptions for building a test dataset. In the case of the training dataset, and in an attempt to build a balanced collection with both relevant and non-relevant photos, we associated each textual description to a maximum of ten photos, namely the relevant photo to be associated to the text, a non-relevant photo containing the highest textual similarity according to the TF-IDF metric, a non-relevant photo containing the highest textual similarity according to the topic distributions metric, a non-relevant photo taken in a place

With the results for each textual description, obtained by taking in the same semester, a non-relevant photo whose image cluster date is the highest cluster textual similarity according to the TF-IDF metric, a non-relevant photo whose image cluster has the highest sentimental polarity, and finally a photo selected randomly.

With the results for each textual description, obtained by...
considering all seven possible feature vector configurations and the two different learning to rank algorithms, we used the `trec_eval` tool to evaluate the matchings between photos and documents. Figure 1 presents the results obtained in terms of Precision at cut-off position 1 (Precision@1) and in terms of the Mean Reciprocal Rank, separately for each of 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 cities and with the best configuration. In all the charts, the bar in red that is fully colored represents the value of Mean Reciprocal Rank, while the bar in blue with a shaded color, represents the value of Precision@1.

The graphics show that method T5 using either the Coordinate Ascent or the RankBoost algorithm, outperforms the other methods in all the cities. These results suggest that the 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.

Table 1 summarizes the obtained results averaged across all the considered cities, showing a higher performance of the CA algorithm in method T5, either in terms of Reciprocal Rank or Precision@1.

Figure 4 illustrates the obtained results for two example textual descriptions, presenting the top-3 most relevant photos as returned by the best performing method (i.e., T5 with the CA algorithm), together with their tags from Flickr. Notice that for an text referencing *France*, the top-3 most relevant photos returned are indeed related to France.

Figure 2 presents the number of documents, in the entire collection of test and training documents, for 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 3 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. These results suggest that either a higher number of words or a higher
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.

number of places do 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 900 and 1000 words can be explained with the small number of documents with these number of words. Also, some of the higher value of Reciprocal Rank and Precision@1 in documents with more than 10 places can be explained with the small number of documents with these number of places.

<table>
<thead>
<tr>
<th>Combination Method</th>
<th>Algorithm</th>
<th>Precision@1</th>
<th>Reciprocal Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>RB</td>
<td>0.44</td>
<td>0.54</td>
</tr>
<tr>
<td>T2</td>
<td>CA</td>
<td>0.50</td>
<td>0.62</td>
</tr>
<tr>
<td>T3</td>
<td>RB</td>
<td>0.44</td>
<td>0.59</td>
</tr>
<tr>
<td>T4</td>
<td>CA</td>
<td>0.50</td>
<td>0.64</td>
</tr>
<tr>
<td>T5</td>
<td>RB</td>
<td>0.80</td>
<td>0.86</td>
</tr>
<tr>
<td>T6</td>
<td>CA</td>
<td>0.76</td>
<td>0.87</td>
</tr>
<tr>
<td>T7</td>
<td>RB</td>
<td>0.70</td>
<td>0.83</td>
</tr>
<tr>
<td></td>
<td>CA</td>
<td>0.68</td>
<td>0.80</td>
</tr>
</tbody>
</table>

Table 1: The obtained results.

5. CONCLUSIONS AND FUTURE WORK
In this paper, we described methods for the automatic association of photos to textual documents. The described methods are based on a pipeline of four steps, in which geographic references are first extracted from documents, then photos matching the geographic references are collected, using Flickr’s API, afterwards photos are clustered with basis on visual similarity, and finally the best photos are selected with basis on their similarity and interestingness. Different methods to select relevant photos were compared and a method based on the combination of term-level textual similarity, topical similarity, geographic proximity and temporal cohesion, using the Coordinate Ascent learning to rank approach 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 is to improve the evaluation mechanism. The evaluation should be done with 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, as well as the MIRFlickr collections for visual concept detection [12, 13], could be important baselines for the development of a more adequate test collection for our work.

A particularly interesting idea for future work concerns with using spatial outlier detection approaches, or even place reference classification methods such as those proposed by Piotrowski at al. for mountaineering accounts [24], in order to filter from the textual documents place references that are only marginally relevant. Besides filtering place references, it would also be interesting to experiment with other features (e.g., features derived from image classifiers trained for detecting specific objects or concepts) or refined versions of the features that were already considered (e.g., better methods for estimating the sentimental polarity, or better methods for measuring image similarity when clustering images).

Finally, we also believe that it would be interesting to experiment with unsupervised learning to rank methods for combining the different relevance estimators. Recent works in the area of information retrieval have described several advanced unsupervised learning to rank methods, capable of outperforming the popular CombSUM and CombMNZ approaches [8]. In IR, rank aggregation 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. [15].

References


