Int J Multimed Info Retr (2013) 2:273–288
DOI 10.1007/s13735-013-0039-3

REGULAR PAPER

Event-related image retrieval: exploring geographical and temporal distribution of user tags

Massimiliano Ruocco · Heri Ramampiaro

Received: 27 February 2013 / Revised: 19 June 2013 / Accepted: 21 June 2013 / Published online: 25 August 2013

Abstract Providing effective tools to retrieve event-related pictures within media-sharing applications, such as Flickr, is an important but challenging task. One interesting aspect is to search pictures related to a specific event with a given annotated image. Most existing methods have focused on doing this by extracting visual features from the pictures. However, pictures in media-sharing applications increasingly come with location information, such as geotags. Therefore, we stress the importance of exploring the possibility to leverage on the geographical and temporal distribution of terms in a tag-based search process, within event-related image retrieval. Specifically, we propose extended query expansion models that exploit the information about the temporal neighborhoods among pictures in a collection, and leverage on the geo-temporal distribution of the candidate expansion terms to reweight and expand the initial query. To evaluate our approach, we conduct extensive experiments on a dataset consisting of pictures from Flickr. The results from these experiments demonstrate the effectiveness of our method with respect to retrieval performance.

Keywords Tag-based image search · Image retrieval · Spatio-temporal approach · Query expansion

This paper is an extended and revised version of ACM HT2013 [41]. This work is supported by the Research Council of Norway, under the VERDIKT program Grant number 176858.

M. Ruocco (✉) · H. Ramampiaro
Department of Computer and Information Science, Norwegian University of Science and Technology, 7491 Trondheim, Norway
e-mail: ruocco@idi.ntnu.no

H. Ramampiaro
e-mail: heri@idi.ntnu.no

1 Introduction

The explosion of photos shared on the web has not only opened many possibilities but also resulted in new needs, and hence new challenges. Although recent developments and technological advances have helped the user to access public photos on the web, e.g., through media-sharing applications, the amount of available information makes the access to these photos still a less straightforward task. To partly address this challenge, the development of event-related image retrieval systems has been proposed [28]. An event-related image retrieval system is optimized to retrieve all pictures related to a specific event. Here, an event has a specific semantic meaning. Focusing on media-sharing applications, an event can be "something happening in a certain place at a certain time and tagged with a certain term" [40]. So in an event-retrieval system, the intent of a user might be to retrieve resources related to a particular event, or to use a given tagged photo representing an event to retrieve other photos related to any similar events from a large image collection. Our main focus is on the latter.

Due to their characteristics, pictures in photo-sharing applications such as *Flickr*[1] and *Panoramio*[2] are here particularly interesting. Pictures in such applications are accompanied by contextual metadata, containing heterogeneous fields, such as camera-specific data, e.g., the Exchangeable image file format (EXIF)[3] data, *Title*, *Tags*, *Description*, temporal information, i.e., capture and upload time, and geolocation, i.e., geotags. In this work, we study how we can exploit the above metadata to retrieve event-related pictures. In doing

[1] See http://www.flickr.com/.

[2] See http://www.panoramio.com/.

[3] EXIF is a standard of the Camera and Imaging Products Association (CIPA). See also http://www.cipa.jp/english.

so, we aim at addressing the following challenges. First, not all pictures may contain reliable tags, description or title. They may either be missing or have no relation to the content of the picture. As a result, they may not contribute much in a retrieval task. Second, Tags are unstructured, subjective and full of noise, thus worsening the retrieval performance. Third, many of the queries are short, i.e., pictures containing only few tags. Dealing with short queries is in itself a challenge. Fourth, a complete collection of images from photo-sharing applications is inherently large, and handling large datasets is in itself an important challenge.

In summary, the main goal of this work is to deal with the above challenges, focusing on situations where a user searches for pictures related to a specific event, each of which is represented by an image with a possibly small number of tags. We believe this area is still not mature, and that only few approaches are available, e.g., [24,28,46]. Further, within information retrieval (IR), existing work has mainly been focused on applying temporal information in the retrieval models [17]. At the same time, the most related approaches, such as [46], are promising with respect to retrieval performance, but seems to be mainly based on using the visual features. We also considered using visual features as part of our approach. However, we early learned that performance (speed) could be a challenge with large datasets. Further, due to the characteristics of event-based pictures, also pointed out by Brenner and Izquierdo [7], we decided to mainly focus on using the metadata. As shown in this paper, we manage to get good retrieval performance, even without using visual features.

That is, we show that by mining and extracting the geo-profiles of terms from textual tags, we can further improve the retrieval performance. To our best knowledge, this has not been explored in depth before. Existing work has mainly been concerned with point-of-interests (POI) extraction [34,39] and trajectory mining [50]. With the constantly increasing number of geotagged pictures in, e.g., Flickr[4], exploring this dimension is important. To this end, our main contributions are as follows. First, we conduct a study comparing the effectiveness of different retrieval models when using only the textual metadata in event-related image retrieval. As part of this, we thoroughly analyze how different combinations of textual fields affect the retrieval effectiveness, depending on the adopted retrieval model. Second, we propose a new weighting model for a query expansion step-based temporal proximity in combination with existing term weighting and similarity models. Third, we develop a new extended model that also includes the mined spatial profile for terms in the textual tags. Our extensive evaluation shows that using both of our new models yields better retrieval performance than the baseline models, especially with short queries, i.e., pictures with only one to three tags.

This paper is organized as follows. Section 2 gives an overview of the related work. Section 3 outlines some preliminary theory that our approach is based on, and defines the problems addressed in this paper. Section 4 elaborates on our new weighting model for query expansion, accompanied by the query expansion models that we use as baselines for our experiments in Sect. 5. Section 6 presents the result from these experiments. Finally, in Sect. 7 we conclude the paper and outline our future work.

2 Related work

Extracting pictures related to real-life events is an active research field [14,38,46], and in the past decades, detection of events from textual document streams and databases has been extensively treated in the literature [1,6]. Still, despite being active, we believe that event-related image retrieval and matching for photo-sharing applications is not a fully mature field. To put our research into perspective, in the following we briefly discuss some approaches that are related to ours.

2.1 Event retrieval and matching

Most related approaches within event retrieval and matching have been aimed at extracting events from different kinds of datasets. To our best knowledge, only few works have addressed the problems of retrieving events related to media sharing. Most of these approaches were presented in the social event detection (SED) task at MediaEval 2011 [28], where the main objective was to develop event retrieval systems for Flickr pictures. Most interesting is the work by Trad et al. [46]. Similar to our approach, the authors proposed methods to match a given (query) picture representing an event to pictures representing the same events in a picture collection. The query image is provided with both temporal and spatial information, and the matching algorithm is based first on visual similarity, followed by a reranking step based on geo-temporal coherence. To handle the scalability, they use Map Reduce in the content analysis and indexing process, and conducted their experiments on a set of around 1 million of pictures, from the LastFM-Flickr dataset [47]. Our work differs from this work in that rather than applying visual features, we only use textual data. This allows us to work on a much larger dataset, i.e., a dataset consisting of around 88 million Flickr pictures.

2.2 Query expansion

Query expansion (QE) has been proven to increase retrieval effectiveness, where an often applied approach is so-called

[4] In 2009, more than 3.3 % (approx. 100 million pictures) were geotagged. See also http://goo.gl/fvjPg.

pseudo-relevance feedback [26,37,51]. The use of temporal information in information retrieval has been previously widely investigated both for ranking models [22] and query expansion [11]. Approaches incorporating geographical context in query expansion has, on the other hand, been mainly proposed in the field of geographical information retrieval (GIR) systems [8,13,30,32]. In particular in [13], the authors propose an expansion process by deriving the spatial query footprint from SPIRIT[5] ontologies, while in [8], the term suggestion is supported by Wordnet.[6] The main difference to our work is that in both [8,13], the query expansion process uses similarity derived from ontologies, whereas, in our work we measure the geographical co-occurrence using the dynamic context of social media resources such as Flickr. This also allows us to use free text search rather than relying on queries with specific query structure, such as the triplet <*what, relation, where*>, often used in GIR [8]. Note that tag suggestion can be related to the topic of query expansion. For example in the work in [27,45], the authors proposed tag recommendation methods for Flickr pictures. The limitation of these approaches is that the input query image must be necessarily geotagged. However, with our approach we only require the tag-based search to be performed with timestamped textual queries.

2.3 Events in photo-sharing applications

Concerning image-sharing applications such as Flickr in general, there are several approaches that are worth discussing. Most notable is the approaches presented in [29,33]. In both of these approaches, the authors proposed methods for detecting groups of events and landmark pictures from community photo collections, by applying a clustering step and followed by a classification step. The main difference is the way the clustering step is carried out. While the former used an agglomerative clustering algorithm, the latter is based on community detection clustering algorithm. Nevertheless, at a first glance, these approaches seem to be related to ours in that we can apply the event detection part in the event retrieval process. However, the focus is different in that we are most interested in directly retrieving event-related pictures without having to cluster and classify them first. Also, to our best knowledge, both of the approaches are based on visual features. In addition to these two approaches, the work by Becker et al. [4] is another approach on extracting events from community photo collections. Here, the authors mainly focused on event clustering. For this reason, the focus is different from ours.

[5] http://www.geo-spirit.org/.

[6] http://wordnet.princeton.edu/.

2.4 Event detection and extraction from Microblogs

Due to the advance of Internet-based social community, much effort has been put on developing approaches to identify and extract events from different social community resources such as Microblogs [5,10,25,42,48], and image-sharing applications [4,29,33,38]. Focusing on Microblogs,such as twitter[7], the user contributes to the social media by posting text messages that are generally short and tagged with a temporal tag. The most important differences of this type of text compared with textual documents are the average length of the textual messages and the noises in which such messages contain. Works on event detection within this domain have tried to tackle the above two characteristics in different ways. For example, Long et al. [25] propose a language-independent approach for detecting, summarizing and tracking events from tweeter posts. Further, Chakrabarti and Punera [10] suggested a real-time approach to summarize the tweeter posts as events, using a modified variant of the Hidden Markov Model to model the hidden state representation of an event. Other examples of real-time approaches were presented in [5,42,48]. In [48], the goal was to detect events from tweet posts by leveraging on their geographical and temporal tags. In [5], the authors presented a method composed by a clustering step, followed by a classification step to group tweets and separate event clusters from non-event clusters, respectively. Finally, Sakaki et al. [42] investigated the possibility to detect events such as earthquake using the real-time stream of tweet posts as sensors. For this, the authors proposed a specific spatio-temporal model based on Kalman filter to detect such a kind of event.

3 Problem definition

The main focus of our work is on a tag-based search of event-related pictures from a photo collection. Here, we assume a query to be a set of tags from a picture, e.g., a Flickr picture tagged with textual information, including Title, Tag and Description, as well as a timestamp specifying when the picture was taken.

So, consider a collection of Flickr images as our target dataset $\mathscr{D}$, where each image $\mathscr{I}$ comes with metadata consisting of information about when the picture was taken and the textual annotations. Then, each image $\mathscr{I} \in \mathscr{D}$ can be represented as $\mathscr{I} = \{\mathscr{T}, d_t\}$, where $\mathscr{T} = \{Title, Description, Tags\}$ denotes the set of textual annotations for $\mathscr{I}$, and d_t is the timestamp for when the photo was taken. With the aforementioned challenges in mind, we want to investigate approaches to deal with the situation in which *a user wants to retrieve a set of pictures*

[7] See http://www.twitter.com/.

representing a specific event, given a picture representing the same event. Formally, if we let $\hat{\mathscr{S}}$ be a set of pictures representing the target event related to the user query intention, and $\mathscr{I}_q \in \hat{\mathscr{S}}$ denote our query image, then our problem is to retrieve all $\mathscr{I} \in \hat{\mathscr{S}}$ representing the same event as $\mathscr{I}_q$.

As part of the solutions to our problem, we will answer the following research questions: First, how do different tag fields of a picture from a media-sharing application such as Flickr affect the retrieval effectiveness? Second, can a query expansion step be useful in retrieving event-related pictures, if we have a query consisting only of the metadata for a single picture? Third, which temporal and spatial features can be useful to improve the search effectiveness in retrieving event-related pictures? Forth, can we still improve the retrieval effectiveness when applying queries with small number of tags?

As we explain in Sect. 4, we aim particularly at exploring the temporal proximity between term distributions and considering the spatial profile of tag terms in retrieving event-related pictures. Further, to partly answer the above questions, in our evaluation we will first perform a set of baseline experiments in which we explore the effectiveness of different retrieval and query expansion models. Then, we will evaluate the retrieval effectiveness of our query expansion model based on temporal and spatio-temporal reranking of the retrieved list.

4 Query expansion for event retrieval

Query expansion is a post-processing step in retrieval systems, aiming at ensuring good retrieval performance when the query is too short, poor and does not contain all the terms, and therefore does not sufficiently reflect the user's search intent [3]. The effectiveness of QE has been proved in many works [15,26,49]. One of the most used QE approaches is pseudo-relevance feedback [3]. The main idea is to assume that a top-k ranked list of retrieved documents are relevant to a specific query. Then, we perform QE by extracting terms from these documents, and use them to reweight and extend the terms in the original query. Depending on the method being used, the choice of the terms can be done by comparing the distributions of terms in the retrieved (or feedback) documents and the entire collection. Note that since we want to tackle the challenges connected to searching event-related pictures using metadata—assuming timestamped pictures with small number of tags, it is necessary to improve and adapt existing query expansion techniques. In the following, we elaborate on how we do this after giving an overview of the baseline QE approaches.

4.1 Baseline query expansion approaches

Generally speaking, a query expansion approach is a two step approach consisting of (1) choosing the terms to be used in the expansion, and (2) assigning the weight to the chosen terms. Focusing on (1), there are several approaches that have been suggested. Among these, we have specifically considered two methods that have been proven to be very effective: the Kullback–Liebler (KL) divergence-based approach [9] and the divergence from randomness (DFR) model [2]. With the KL divergence approach, the idea is to analyze the term distributions, and maximize the divergence between the distribution of terms from the top-k retrieved documents and the distribution of terms over the entire collection [9]. The terms chosen for the query expansion are those contributing to the highest divergence, i.e., the highest KL score [9]. This means that expansion terms with low probability in the entire collection and high probability on the retrieved top-k documents are given more weights than the other terms. The following equation is used to calculate the KL score for a given term t in the feedback (top-k) documents [9]:

$$\mathrm{KL}(t) = P_{\mathrm{Rel}}(t) \log \left[\frac{P_{\mathrm{Rel}}(t)}{P_{\mathrm{Coll}}(t)} \right], \tag{1}$$

where $P_{\mathrm{Rel}}(t)$ and $P_{\mathrm{Coll}}(t)$ are the probability that t appears in the top-k documents and in the collection, respectively. Here, $P_{\mathrm{Rel}}(t)$ can be estimated by the normalized term frequency of t in the top-k documents, whereas $P_{\mathrm{Coll}}(t)$ can be computed as the normalized frequency of t in the entire collection. With the DFR model, on the other hand, the idea is to weight the expansion terms by calculating the divergence between the distribution of terms in the feedback documents (the top-k documents) and a random distribution [2]. In our work, we have chosen to implement this method based on the Bose–Einstein statistics (Bo1), which has been shown to be one of the most effective approaches. Bo1 is computed as follows [2]:

$$\mathrm{Bo1}(t) = tf_{\mathrm{feedback}} \log \left[\frac{1 + P_n(t)}{P_n(t)} \right] + \log \left[1 + P_n(t) \right], \tag{2}$$

where tf_{feedback} is the frequency of term t in the feedback documents, and $P_n(t) = F/N$ is the ratio between the frequency F of t in the entire collection and N the size of the data set. After the expansion terms have been selected using one of the approaches above, we can proceed to step (2), i.e, reweighting the terms in the query. One of the classical approach to reweight query terms is the Rocchio's algorithm [37]. In particular, we use the Rocchio's Beta equation [31] as follows:

$$\hat{w}(t_q) = \frac{tf_{q_{tq}}}{\max tf_q} + \beta \frac{w(t_q)}{\max w}, \tag{3}$$

where $\hat{w}(t_q)$ is the new weight of a term t_q of the query, $w(t_q)$ is the weight from the expansion model, i.e., $\mathrm{KL}(t_q)$ or $Bo1(t_q)$, max w is the maximum weight from the expanded weight model, max tf_q is the maximum term frequency in the query and $tf_{q_{tq}}$ is the frequency of the term in the query.

4.2 Extended query expansion models for event retrieval

In this section, we propose a set of methods to extend the above baseline models. Our main goal is to exploit the temporal and geographical information encapsulated in the picture tags. Previous approaches have focused on investigating the application of the temporal information in pseudo-relevance feedback approaches. For example, the approaches by Efron and Golovchinsky [12] and Keikha et al. [18] proposed methods to incorporate time into the relevance model by Lavrenko and Croft [20]. In contrast to this, our objective is to use the characteristics of an event, in combination with the temporal proximity of the term distribution as features in the term selection process for a query expansion framework. We assume that all pictures in our collection contain a temporal annotation identifying when the picture was taken, i.e., a timestamp. Further, we hypothesize that pictures related to the same event have some temporal proximity or temporal closeness. This means that the more temporally close to the query the retrieved pictures are, the more likely that they are related to the same event. Such a property is useful in a query expansion framework, since we can use the temporal information to decide the term weights. For example, we can give higher weights to terms having higher probability to appear in a document and being temporally close to the query. With this in mind, in the following we propose a query expansion model to improve the retrieval of events.

4.2.1 Temporal-proximity reranking

As a first improvement, we explore the effectiveness of using a ranking function that considers both the textual similarity and the temporal proximity of the document, in the query expansion process. The idea is to push documents with higher temporal proximity up in the top-k feedback documents. Note, however, that the temporal similarity and the textual similarity are not two unified measures. Therefore, the scores assigned by performing two queries, one with textual query and another with the temporal data, are not straightforward to merge by a score-based ranked list fusion. For this reason, we merge the two ranked lists by adopting *rCombMNZ* [21], which is the ranked-based version of *CombMNZ* [43], given by

$$\mathrm{score}^{R_i} = h(d, \mathbf{R}) \sum_{R_i \in [R_1, R_2]} g^{R_i}(d), \tag{4}$$

where d is a document of a ranked list, R_1 and R_2 are the two ranked lists and $h(d, \mathbf{R})$ is the rank hits representing the number of ranking lists in which the document d is present. Further, $g^{R_i}(d)$ denotes the normalized ranking score of the document d in the ranked list R_i.

4.2.2 Temporal-proximity-aware KL divergence

As a second improvement, we actively use the assumption about temporal proximity, mentioned before. In both of the presented baseline query expansion models, the core premise is that a query expansion word should be more common in the feedback documents and less common in the whole collection. This means that we have a high divergence between the distribution of the candidate term expansion in the feedback document set, and the distribution of the same term in the whole collection. Hence, our intuition is the following: the distribution of a good candidate expansion term should commonly co-occur as much as possible in documents that are temporally close to the query picture and less common in the whole collection. This is the same as having a high divergence between the distribution of the co-occurrence of the candidate expansion terms and the query terms in the set of temporal neighbors pictures, and the distribution in the whole collection. The idea is that in addition to the original KL-divergence computation, our weighting process also considers the divergence of the term distributions within a time slice $\mathscr{L}$, centered in the timestamp of the query image, and the co-occurrence with the query terms within the same time slice. Now, let $\theta^{\mathscr{L}}_{[t,t_i]}$ be the distribution of the co-occurrence between the candidate expansion term t and the query terms $t_i \in Q$ within the set of temporal neighbors, and $\theta^{\mathrm{Coll}}_{[t,t_i]}$ denote the distribution of the co-occurrence terms in the whole collection. Then, our temporal-aware KL score can be computed as follows:

$$\mathrm{KL}^{\mathscr{L}}(Q, t) = \sum_{t_i \in Q} \mathrm{KL}(\theta^{\mathscr{L}}_{[t,t_i]} \| \theta^{\mathrm{Coll}}_{[t,t_i]}) \tag{5}$$

$$= \sum_{t_i \in Q} P_{\mathscr{L}}([t|t_i]) \log \left[\frac{P_{\mathscr{L}}(t|t_i)}{P_{\mathrm{Coll}}(t|t_i)} \right]. \tag{6}$$

In this reweighting process, the new weight of a candidate expansion term t is the sum of the divergence between $\theta^{\mathscr{L}}_{[t,t_i]}$ and $\theta^{\mathrm{Coll}}_{[t,t_i]}$, for all the $t_i \in Q$. In other words, a candidate expansion term gets a higher weight if the divergence between these two distributions $\theta^{\mathscr{L}}_{[t,t_i]}$ and $\theta^{\mathrm{Coll}}_{[t,t_i]}$ is high. Further, $P_{\mathscr{L}}(t|t_i)$ is the co-occurrence probability of the terms t and t_i within a time interval $\mathscr{L}$, and $P_{\mathrm{Coll}}(t|t_i)$ is the co-occurrence probability of the terms t and t_i within the whole collection. We evaluate the co-occurrence probability as proposed in [44] by adding a normalization factor:

Table 1 The scores of the query expansion terms after baseline KL divergence (KL), and temporal KL divergence ($KL^{\mathscr{L}}$)

KL (t)		$KL^{\mathscr{L}}(Q,t)$	
atmedia	1.400	london	1.316
london	1.168	atmedia	1.266
ajax	1.108	ajax	1.135
atmedia2009	**0.270**	media	0.400
atmediaajax	0.089	atmediaajax	0.244
javascript	0.077	**event**	**0.182**
atmedia09	0.055	**conference**	**0.146**
media	0.050	**web**	**0.130**
web	**0.047**	javascript	0.097
conference	**0.040**	**presentation**	**0.048**
atmedia2008	**0.030**	**session**	**0.046**
event	**0.025**	abbey	0.033
presentation	**0.020**	pub	0.031
brendaneich	0.015	bar	0.021
session	**0.014**	screen	0.021
johnresig	0.013	brendaneich	0.015
christianheilmann	0.012	lectern	0.014
patrickgriffiths	0.012	christianheilmann	0.009

Boldunderline values indicates increasing position in the ranked list

$$P_{\mathscr{L}}(t|t_i) = \frac{\left[\frac{n_d^{\mathscr{L}}(t,t_i)}{n_d^{\mathscr{L}}(t)+n_d^{\mathscr{L}}(t_i)}\right]}{|\mathscr{D}_{\mathscr{L}}|}, \quad \text{and} \tag{7}$$

$$P_{\text{Coll}}(t|t_i) = \frac{\left[\frac{n_d^{\text{Coll}}(t,t_i)}{n_d^{\text{Coll}}(t)+n_d^{\text{Coll}}(t_i)}\right]}{|\mathscr{D}|}, \tag{8}$$

where $\mathscr{D}$ is the whole dataset and $D_{\mathscr{L}} \subset D$ is a subset of D composed by documents having timestamp within the time interval $\mathscr{L}$. This means that $n_d^{\mathscr{L}}(t,t_i)$ and $n_d^{\text{Coll}}(t,t_i)$ are the number of documents in the set $D_{\mathscr{L}}$ and D, respectively, in which the terms t and t_i co-occur. Similarly, $n_d^{\mathscr{L}}(t)$ and $n_d^{\text{Coll}}(t)$ are the number of documents in the set $D_{\mathscr{L}}$ and D, respectively, that are tagged with the term t.

Example To explain the motivation behind Eq. 6, consider the tag scores in Table 1. This table shows the results of two reweighting processes: (1) using the baseline KL divergence in Sect. 4.1, and (2) using the temporal KL in Eq. 6. Here, our query was a picture with the tags {atmedia, london, ajax} and the timestamp (27.09.2007), referring to a periodic conference event, "atmedia", in 2007. As shown in Table 1, our dataset at least contains pictures from the 2008 and 2009 conferences.

In this example, we can make the following interesting observations. First, with the baseline approach, although several tags may refer to the same periodic event, e.g, the tag atmedia2008 and atmedia2009, different times may lead to different scores. Second, using our temporal KL divergence approach, generic event- related terms in the user query Q, such as event, conference and session, get higher scores than with the baseline approach. This is because the distribution of the co-occurrences of such terms with the query terms have a higher divergence in the set of temporal neighbors, compared to the divergence of the same distribution in the whole collection. To further illustrate the usefulness of applying the temporal information, Fig. 1 shows how the temporal distributions of two tags conference and atmedia, and their co-occurrences look like within a given time interval. Also, the results of our experiments in Sect. 5 demonstrate that our observation also apply to most cases.

4.2.3 Combining the KL divergences

To include the influences of both scores in the calculation of the final expansion weight, the last two models can be mixed in a linear combination, given by

$$\text{KLT}(Q,t) = \gamma \text{KL}(t) + (1-\gamma)\text{KL}^{\mathscr{L}}(Q,t), \tag{9}$$

where γ is factor used to determine the amount of influence each score has on the final weight. In our experiment, we will analyze the retrieval effectiveness as function of the values of γ on the weighting step, in the query expansion process. This gives us also the possibility to evaluate the impact of the proposed temporal weighting model.

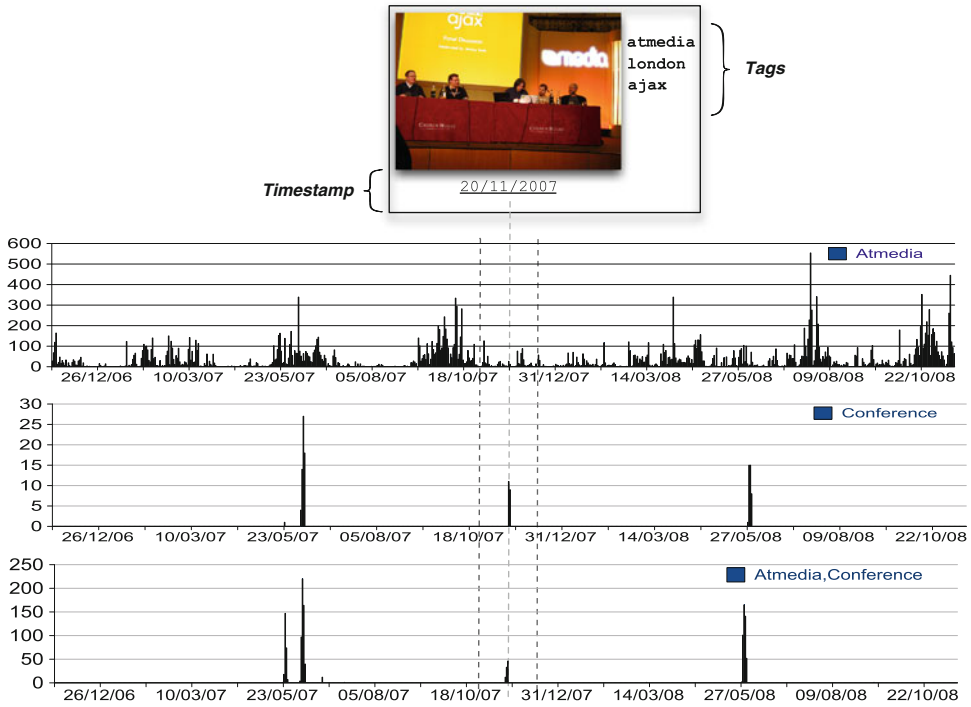


Fig. 1 Temporal distribution of the tags `conference` and `atmedia` and their co-occurrence

4.3 Exploring term spatial distribution

As explained in our hypothesis, pictures related to the same event tend to appear in a limited geographical area. In this approach, we mainly consider query pictures that are not geotagged. There are two main reasons for this. First, we believe the problem would be less challenging to solve when having both the temporal and geographical information available. Second, we aim at making our approach as generic as possible, and thus enabling it for media-sharing applications and social media in general. For example, in the Flickr dataset only 3.3 % of the pictures are geotagged. As a conclusion, although the probability to have a geotagged picture is low, the portion of pictures that are geotags can still be useful to extract geographical profile of the terms.

With this in mind, we propose a method to find a good expansion term t, given a set of query terms $Q = \{t_i\}_i$. Including the geographical dimension, a good expansion term is a term related to the same event of the query picture. In particular it is a term that commonly co-occur in documents that are temporally close to the query picture and in a geographic delimited area, and less common in the whole temporal timeline in the same delimited area. To define a realistic problem, the query picture is not geotagged.

The method presented is based on the discretization of the world map. We first divide the world map in *M tiles* $\Theta = \{\mathscr{T}_k\}_{k=1\ldots M}$ of size one degree as proposed by [52]. This means that the tiles does not have the same size. This is because, on the world map, the size corresponding to one degree varies depending on the latitude values; spanning from 0 Km on the poles, to 100 Km close to the equator. This approximation is suitable to use since most of the highly populated areas are closer to the equator than the poles.

To include the spatial dimension in the candidate expansion term score, we use a similar hypothesis to the one proposed in Sect. 4.2.2 as a starting point. This means that a good expansion term t is the one for which there is a high divergence between the distribution of the pictures tagged with the query term and the expansion term in a temporal time slice $\mathscr{L}$ and a tile $\mathscr{T}_k$, and the distribution of the terms in the same geographical tile $\mathscr{T}_k$ but covering the whole timeline.

Formally the new divergence is computed using KL-divergence as follow:

$$\mathrm{KL}^{\mathscr{L}}_{\mathscr{T}_k}(Q,t) = \sum_{t_i \in Q} \mathrm{KL}\left(\theta^{\mathscr{L}}_{[t,t_i,\mathscr{T}_k]} \| \theta^{\mathrm{Coll}}_{[t,t_i,\mathscr{T}_k]}\right) \tag{10}$$

$$= \sum_{t_i \in Q} P_{\mathscr{L}}(t|t_i,\mathscr{T}_k) \log\left[\frac{P_{\mathscr{L}}(t|t_i,\mathscr{T}_k)}{P_{\mathrm{Coll}}(t|t_i,\mathscr{T}_k)}\right]. \tag{11}$$

Here, $P_{\mathscr{L}}(t|t_i,\mathscr{T}_k)$ is the co-occurrence probability of the query term t_i and expansion term t, within a time interval $\mathscr{L}$ and a geographical tile $\mathscr{T}_k$. Similarly, $P_{\mathrm{Coll}}(t|t_i,\mathscr{T}_k)$, is the same probability without the temporal restriction. We approximate these probability as follows:

$$P^{G}_{\mathrm{Coll}}(t|t_i,\mathscr{T}_k) = \frac{\left[\frac{n^{\mathrm{Coll}}_d(t,t_i|\mathscr{T}_k)}{n^{\mathrm{Coll}}_d(t|\mathscr{T}_k)+n^{\mathrm{Coll}}_d(t_i|\mathscr{T}_k)}\right]}{|\mathscr{T}_k|} \tag{12}$$

$$P^G_{\mathcal{L}}(t|t_i, \mathcal{T}_k^{\mathcal{L}}) = \frac{\left[\frac{n_d^{\mathcal{L}}(t,t_i|\mathcal{T}_k)}{n_d^{\mathcal{L}}(t|\mathcal{T}_k)+n_d^{\mathcal{L}}(t_i|\mathcal{T}_k)}\right]}{|\mathcal{T}_k^{\mathcal{L}}|}. \quad (13)$$

We calculate the pair of probabilities $P^G_{\text{Coll}}(t|t_i, \mathcal{T}_k)$ and $P^G_{\mathcal{L}}(t|t_i, \mathcal{T}_k^{\mathcal{L}})$ for each tile $\mathcal{T}_k \in \Theta$. We calculate the divergence between the two distribution values, tile by tile. We consider the maximum as the final score. To include the influence of KLT, we mixed the models in a linear combination, given by

$$\text{KLST}(Q,t) = \sigma\,\text{KLT}(Q,t) + (1-\sigma)\max\{\text{KL}^{\mathcal{L}}_{\mathcal{T}_k}(Q,t)\}_{\mathcal{T}_k}. \quad (14)$$

4.4 Scalability of the method

Recall that the purpose of our work is to improve the tag-based search effectiveness of event related resources, such as Flickr pictures, by improving the keyword-based ranking models in IR. In our approach, the images are indexed based on their textual metadata (the tags), using inverted index structure. It is a data structure that efficiently store and retrieve textual resources, and has been proven scalable [23].

As for our framework, the temporal and spatial dimensions are included in the ranking model, and our query expansion method does not need extra data structure. Thus, the only bottleneck might be the increased size of queries. However, as we mentioned before, we assume that the query size is normally small. Therefore, this would not be an issue.

Nevertheless, our expansion algorithm is depicted in Algorithm 1. To further understand the scalability of our approach, let us analyze the computation cost of this algorithm. As can be observed, to compute the final score, the algorithm requires $N = |\mathscr{E}| * |Q| * |\Theta|$ steps, where $|\mathscr{E}|$ is the number of expansion terms, $|Q|$ is the size of the query and $|\Theta|$ denotes the number of tiles. Since $|\Theta|$ is a finite number and that not all tiles contain images, plus $|Q|$ is normally small, it is safe to assume that our algorithm has a complexity of $\mathscr{O}(|\mathcal{E}|)$.

In general situations where the above sizes are unlimited, we can parallelize the core of the algorithm, i.e., Step 6 to Step 8 in Algorithm 1. Moreover, computing Eq. 10 is done with a query limited in a spatial area (the tile). During the computation, this area is fixed for any queries. In such a case, scalability would not cause any problem.

As a final note, to perform our experiments, we indexed and run our queries using Terrier[8] for the text search and Solr[9] for the spatial search, both providing features for searching and storing web-scale indexes. Further, we defined three random test queries with one keyword, two keywords and three keywords. Then, we measured the response time as function of the size of the dataset, i.e., the number of indexed documents. We performed the experiments on an Intel i7-950 Processor, with 24 Gb RAM and 1 Tb Hard Disk. Figure 2 summarizes the results of our experiments, showing the average response time of the baseline QE method and the average response time of our QE approach.

As can be derived from these results, even though the size grew, the execution times did not follow the increase of the dataset size. Note that the code written to perform the experiments was not optimized, and thus this lack of optimization might affect the response time, in general. More specifically, we did not perform any parallelization of the queries in Steps 6–8 in Algorithm 1. We did not optimize Solr neither, but used standard tuning values. Finally, we did not warm up the cache of the Solr system before each experiment, i.e., the cache was empty at each query processing.

Algorithm 1 Pseudo code of the QE procedure that incorporates geo-temporal dimensions.

1: $\mathcal{L} \leftarrow$ time interval centred in the query timestamp
2: **Query Q** by using ranking model r and get the **D** set of top-N relevant docs
3: **Extract** unique tags from **D** and get the candidate expansion term set $\mathcal{E}$
4: **for** e_j in $\mathcal{E}$ **do**
5: **for** t_i in **Q do**
6: **for** $\mathcal{T}_k$ in Θ **do**
7: **Calculate** $KL^{\mathcal{L}}_{\mathcal{T}_k}(t_i, e_j)$
8: **end for**
9: **Calculate** $KLT(t_i, e_j)$
10: **Calculate** $KLST(t_i, e_j)$
11: **end for**
12: **Calculate** $\sum_{t_i \in Q} KLST(t_i, e_j)$
13: **end for**
14: **Rank** $e_j \in \mathcal{E}$ terms according to $KLST(Q, e_j) \rightarrow \mathcal{E}_{Rank}$
15: **Re-build Q** with the top-k terms from $\mathcal{E}_{Rank} \rightarrow \hat{\mathbf{Q}}$
16: **Query** $\hat{\mathbf{Q}}$ by using ranking model r

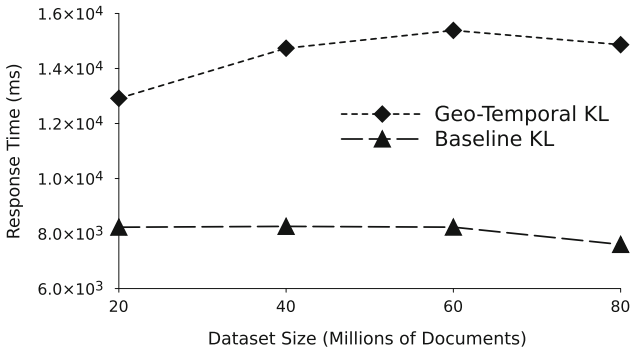


Fig. 2 Response time for our QE executions with the random test queries as function of the dataset size

[8] See http://www.terrier.org.

[9] See http://lucene.apache.org/solr.

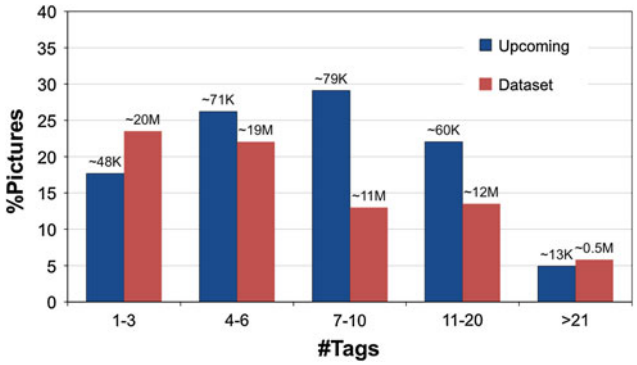


Fig. 3 Distribution of number of tags per pictures

5 Experimental setup

5.1 Dataset

To evaluate our method, we use the *Upcoming dataset* [4] as the ground-truth for our experiments. This dataset consists of 270.425 pictures from Flickr, taken between 1 January 2006 and 31 December 2008, each of which belongs to a specific event from the Upcoming event database.[10] The unique number of events are 9.515. Each event is composed by a variable number of images, varying from 1 to 2.398 pictures. Further, since the size of this dataset is relatively small for our purpose, and due to the lack of other datasets that are very large, we decided to build an additional dataset by merging the Upcoming dataset with other pictures gathered from Flickr[11] covering a time period from 01.01.2006 to 31.12.2010 and without spatial restrictions. Our final dataset now consists of 88.257.485 pictures, of which 18.861.585 pictures are without any tags and around 23.5 % with 1–3 tags (see also Fig. 3 for more information about the distribution of the number of tags). This further illustrates the necessity of supporting short queries, as mentioned in Sect. 1. Also, this shows that both the ground-truth and the final dataset contain sufficiently enough portions of short queries.

Before performing our experiments, we first indexed all image tags using Terrier. As part of the dataset preparation we run a preprocessing step consisting of tokenization, i.e., UTFTokenizion based on whitespace and punctuation marks, and English stopword removal. Then, we randomly selected set of pictures from each event cluster in the Upcoming dataset and use these as queries.

[10] See http://upcoming.yahoo.com/.

[11] We used Flickr API to do this. See also http://www.flickr.com/services/api/.

5.2 Evaluation methodology

To assess the effectiveness of our approach, we compare our models with existing models, which also serve as baseline for our evaluation. Our baseline models are the Vector Space Model (TFIDF) [3], Okapi BM25 (BM25) [35], Hiemstra Language Modelling (LM) weighting model [16] and KL divergence retrieval model (KLDM) [19]. For both BM25 and LM, we use the default parameter values, i.e., for BM25 we set $k_1 = 1.2$, $k_3 = 8$ and $b = 0.75$, and for LM is $c = 0.15$.

To evaluate the retrieval performance, we use standard in information retrieval evaluation metrics, including the mean average precision (MAP) and R-Precision (RP) [3]. To make sure that any improvements are statistically significant, we perform paired two-sample one-tailed t tests at $p < 0.05$ or 95 % confidence interval. Therefore, any stated improvements in this paper are all statistically significant, unless otherwise specified.

5.3 Considerations related to query expansion

Studying our dataset, we observed that more than one picture related to the same event have been annotated with the same set of tags by the same user. This is because many users in Flickr often copy and paste the same set of tags for pictures related to the same events or same group of pictures. To illustrate this, Fig. 4 shows the difference between the number of picture retrieved and the number of unique pictures in the retrieved set, using our query set presented above and with a BM25 retrieval model.

This histogram shows that a set of retrieved documents generally contains a high percentage (around 80 % in all the cases) of pictures with duplicated set of tags. This observation is useful when performing a query expansion on the type of dataset as ours. Further, when extracting candidate expansion terms from the top-K retrieved documents, it can happen that a high number of duplicates of tag sets are in the documents (pictures) within the top-K positions. This would reduce the space of candidate expansion terms. To avoid this problem,

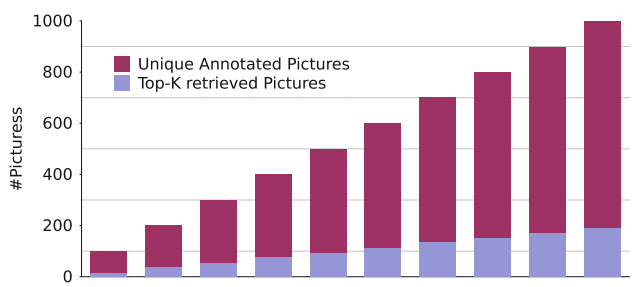


Fig. 4 Comparison between the number of pictures retrieved and the uniqueness of their TagSet

we decide to remove the duplicates from the retrieved document set during the process of selecting the top-K retrieved pictures. So in our experiments, the number of pictures in the top-k retrieved set used to select the candidate expansion terms is the number after removing the duplicates.

Finally, to handle noisy and non-informative tags, we first filter candidate expansion terms from the whole dataset that do not comply with $tf < 100$. Then, we remove candidate terms that do not match $tf_{\mathcal{T}_k} < 50$, where $tf_{\mathcal{T}_k}$ is the term frequency of a term extracted from images taken within the geographical tile $\mathcal{T}_k$.

6 Results

Aiming at answering our research questions in Sect. 3, we analyze the effectiveness of each textual field in the pictures to find out which of the fields contributes to the best retrieval performance. Thereafter, we perform different sets of experiments to study the effectiveness of our proposed query expansion model with respect to different parameter values.

6.1 Field effectiveness

Our first experiment aims at exploring the effectiveness of using Flickr images as queries. To assess this effectiveness and analyze the role of the fields in the metadata, we use different combinations of the textual metadata as queries and document representation. Specifically, we evaluate how Title, Tag and their combination affect the retrieval effectiveness. To do this, we first use Title only as a document, then Tag only, and finally Description only. Thereafter, we test different combinations of these fields as follows: Title and Tag; and Title, Tag and Description.

Note that the efforts of the TREC community on retrieval of structured and unstructured documents, i.e., the INEX benchmarking for XML information retrieval, and the field-based retrieval models such as BM25F [36] can seem to be related to this part of our work. However, because the focuses of these are more on full text contents, they are beyond the scope of this work.

The set of queries is formed by randomly selecting one picture from each event cluster in the Upcoming Dataset. Here, we only consider event clusters containing more than 500 pictures from a total of 50 clusters. Thus, the total number of queries is 50 for each sample. This random sampling is repeated five times to obtain five sets of 50 queries, which means that the total number of queries submitted are 250.

Tables 2, 3 and 4 summarize the results from the experiments for our retrieval effectiveness analyses. Here, TAG_{TAG} means that we use the tag field in both the indexing and the

Table 2 MAP and RP by querying using the Title field

Bold indicates statistically significant highest RP and MAP value (then for each column)

Comb	TFIDF		BM25		LM		KLDM	
	MAP	RP	MAP	RP	MAP	RP	MAP	RP
TIT_{TAG}	0.498	0.502	**0.500**^{234}	**0.506**^{234}	**0.484**^{234}	**0.492**^{234}	**0.503**^{234}	**0.510**^{234}
TIT_{TIT}	0.350	0.358	0.324	0.332	0.357	0.364	0.353	0.360
TIT_{DES}	**0.550**^{124}	**0.559**^{124}	0.459	0.467	0.460	0.468	0.460	0.468
TIT_{TT}	0.113	0.129	0.106	0.124	0.127	0.140	0.130	0.147

Table 3 MAP and RP by querying using the Tag field

Bold indicates statistically significant highest RP and MAP value (then for each column)

Comb	TFIDF		BM25		LM		KLDM	
	MAP	RP	MAP	RP	MAP	RP	MAP	RP
TAG_{TAG}	0.685	0.695	**0.687**^{234}	**0.697**^{234}	0.691	0.704	**0.6925**^{23}	**0.7043**^{23}
TAG_{TIT}	0.064	0.082	0.085	0.105	0.067	0.083	0.064	0.081
TAG_{DES}	0.281	0.290	0.281	0.287	0.434	0.448	0.281	0.288
TAG_{TT}	**0.695**^{123}	**0.707**^{123}	0.530	0.540	**0.696**^{123}	**0.708**^{123}	**0.691**^{23}	**0.704**^{23}

Table 4 MAP and RP by querying using both the Tag and Title field

Bold indicates statistically significant highest RP and MAP value (then for each column)

Comb	TFIDF		BM25		LM		KLDM	
	MAP	RP	MAP	RP	MAP	RP	MAP	RP
TT_{TAG}	0.663	0.680	**0.669**^{23}	**0.686**^{23}	0.468	0.484	**0.667**^{23}	**0.682**^{23}
TT_{TIT}	0.117	0.139	0.108	0.129	0.120	0.144	0.129	0.154
TT_{DES}	0.369	0.376	0.287	0.295	0.288	0.297	0.289	0.297
TT_{TT}	**0.673**^{123}	**0.690**^{123}	**0.665**^{23}	**0.683**^{23}	**0.693**^{123}	**0.705**^{123}	**0.670**^{23}	**0.686**^{23}

query, whereas TAG_{TIT} means we apply tag (TAG) in the indexing but title (TIT) in the query, and so on. TT stands for tags and title combination, while DES is the description field. The numbers 1, 2, 3 and 4 in superscript in the tables indicate the statistical significance improvements on the dataset indexed with TAG field, TIT field, DES field and TT fields, respectively.

With the results in these three tables, we can make the following observations. First as shown in Table 2, querying using the title resulted in the lowest MAP and R-precision values compared to querying with the title and the tags. Further, looking at the best results in each table, for each retrieval model, the most representative field for each picture was the Tag field, with which the MAP and the RP values were the highest.

Finally, in Tables 3 and 4, we can see that in all the cases, the highest MAP and RP values were obtained when the same fields were used both to represent the documents/images and to generate the set of queries. In summary, since these results are conclusive, we can safely base our experiments to test our query expansion step using the combination TAG_{TAG}.

6.2 Short queries versus long queries

In this section, we compare the retrieval effectiveness of using query pictures with less than three tags and query pictures with more than four tags. To do this, we randomly select 40 query pictures with less than three tags and 40 query pictures with more than four tags. To make the experiment more realistic, we consider only event clusters containing more than 100 pictures. This is because a small number of users normally contribute to small clusters. Thus, there would be a high probability that a high percentage of the pictures would be annotated with the same tags.

To perform this experiment, as well as executing the standard models, we also applied the query expansion models described in the previous section. Specifically, we used the Rocchio's framework weighting model, with both the Kullback–Leibler divergence model (KL), and the Bose–Einstein weighting scheme (Bo1) to choose the expansion terms. For each QE run, we used the default values, i.e., setting $\beta = 0.4$ and choosing the first n terms of the top-K documents for the Rocchio's Beta weighting model. The values of K, i.e., the number of pseudo-relevant documents, were chosen from {30, 60, 90}, and n, i.e., the number of selected terms, from {8, 18}.

Figure 5 and Table 5 present the results of our comparisons of the effects of using short and long queries. Specifically, in Fig. 5 we focus on comparing the effects of short and long queries on the retrieval effectiveness when using only standard IR models. In Table 5, on the other hand, we compare the impacts of the query lengths when applying the two different query expansion models, Bo1 and KL. Here, we summarize the percentage improvements from standard IR models to applying the query expansion models.

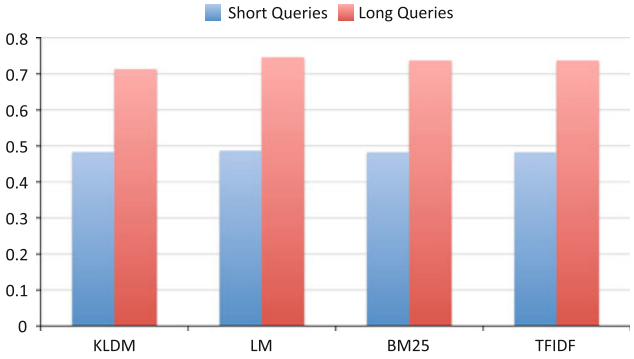


Fig. 5 MAP values with respect to using different retrieval models

Table 5 Short versus long queries: percentage improvements using the query expansion model compared to the standard retrieval model, in terms of MAP values

	Bo1		KL	
	Long	Short	Long	Short
TFIDF (%)	2.75	5.79	2.75	5.78
BM25 (%)	2.62	5.71	2.54	5.83
LM (%)	−0.33	3.98	−0.32	3.98

As Fig. 5 shows, with all standard IR models, we obtained the highest MAP and R-precision values with long queries. In contrast to this, as shown in Table 5, when applying the query expansion step, we generally get the best results with the short queries. More specifically, apart from the Language Model (LM), where long queries resulted in decreased MAP values, applying the query expansion step yielded two times higher improvements with short queries than using long queries.

As a conclusion, if we only use standard retrieval models, we get the best results with long queries. The reasons for this is that (1) the use of a higher number of tags make the query more effective, and (2) many users usually annotate groups of pictures with high number of tags. Since we extract the expansion terms from a list of top-K documents, thus making most of the query terms either an excess or more important, short queries with the expansion steps give the best results. For this reason, we focus on improving the query expansion models based on short queries.

6.3 Evaluating the extended QE models

In this experiment, we evaluated the approaches proposed in Sect. 4.2.1. As in the previous experiment, we first randomly selected 100 queries from the event clusters, containing more than 100 pictures. Then we selected pictures with less than three tags.

In the first set of experiments, we compared the results obtained by performing the retrieval process followed first by the standard KL divergence for query expansion (KL), and thereafter by the proposed proximity-based temporal KL (KLT). In the second set of experiments, we tried the combination of the two proposed methods, i.e., selecting the expansion query terms by considering the pseudo-relevant top-K documents and weighting the terms extracted applying KL and KLT, using the linear combination in Eq. 9. In the third experiment, we compared the previous models with the one based on spatial distribution of terms (KLST). In the fourth experiment set, to assess the effectiveness of KLT, we compared the effectiveness of KL and KLT, when doing a reranking step as explained in Sect. 4.2.1, with Eq. 4 being either applied or not applied. Here, the QE was performed on pseudo-relevant pictures, still using Bo1 and KL in the *Rocchio's Beta* framework (RB), with the same default values of β.

Now, to perform a complete set of experiments, we considered different values of the following parameters. First, as query expansion parameters, we varied the value of K such that $K \in \{30, 60, 90, 120\}$ and the values of n such that $n \in \{8, 18\}$. Second, as a parameter for KLT, we varied the time slice $\mathscr{L}$ in the following set: {1 day, 3 days, 7 days}. Third, for the reranking step, we varied the R values, i.e., number of top-R documents to rerank, in the set {1,000, 2,000, 3,000, 4,000}.

In addition to the above models, we also implemented the Mixture Model [51] and the Relevance Model [20]. However, the results were comparable to the KL and Bo1 query expansion models. Thus, due to the space constraints, we did not include them in this paper.

6.3.1 The impact of γ on mixed KL

With this set of experiments, we tested the impact of the parameter γ in Eq. 9 used to linearly combine the KLT and the standard KL divergences. We varied its values from 0 to 1 such that $\gamma \in \{0, 0.25, 0.5, 0.75, 1\}$, where 0 means that we only have the contribution of KLT and to 1 we only have the contribution of KL. We repeated this experiment for the six combinations of the number of query expansion terms n, and the number of top-K documents considered in the query expansion process.

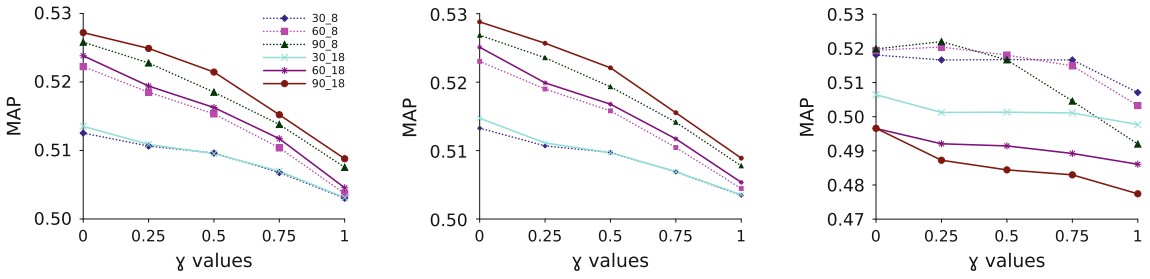
Figure 6 shows the impact of varying the values of γ on the MAP values. As can be seen in this figure, for all the combinations of K and n values, the MAP values decreased when we increased the γ value. This means that mixing both of the contributions was not very effective with respect to retrieval performance, but the most important contribution came from our KLT divergence.

6.3.2 KL versus KLT

To further assess the performance of our KLT approach, we compared it with the baseline approach, using the linear combination in Eq. 9, with $\gamma = 0$.

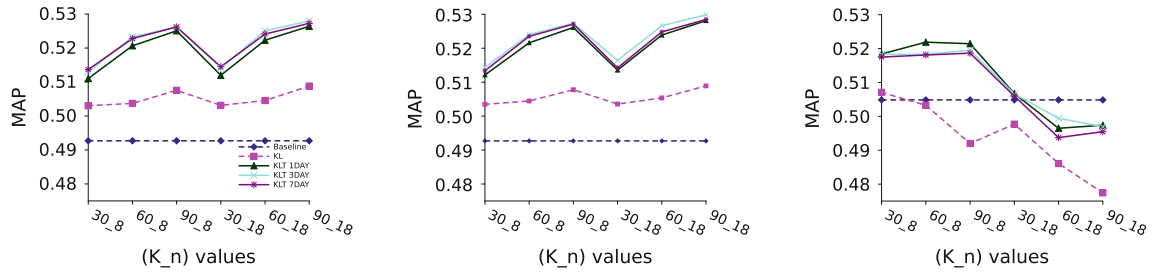
First, we compared KL and KLT without any reranking step. The result from this experiment is summarized in Fig. 7, showing comparison between the retrieval effectivenesses of our QE models and the baseline models. As can be observed, using BM25 and TFIDF retrieval models in the initial retrieval step, our KLT outperforms KL, with all combinations of K and n. With LM and $n = 30$, the KLT also outperforms the baseline model. With $n = 60$, the KLT still outperforms KL but in this case the query expansion process is not very effective. Overall, we can conclude that our query expansion models are better than the baseline QE model, and that all presented improvements are statistically significant at 95 % confidence interval.

We carried out our next experiment to assess the effectiveness of our KLT compared to the baseline query expansion, including the reranking step. Specifically, we evaluate the impact of R, i.e., the number documents reranked with respect to the temporal proximity. Here, we performed the retrieval process, first by reranking and then applying the KL divergence for query expansion (RERANKING+KL), and second by reranking and then applying our proposed



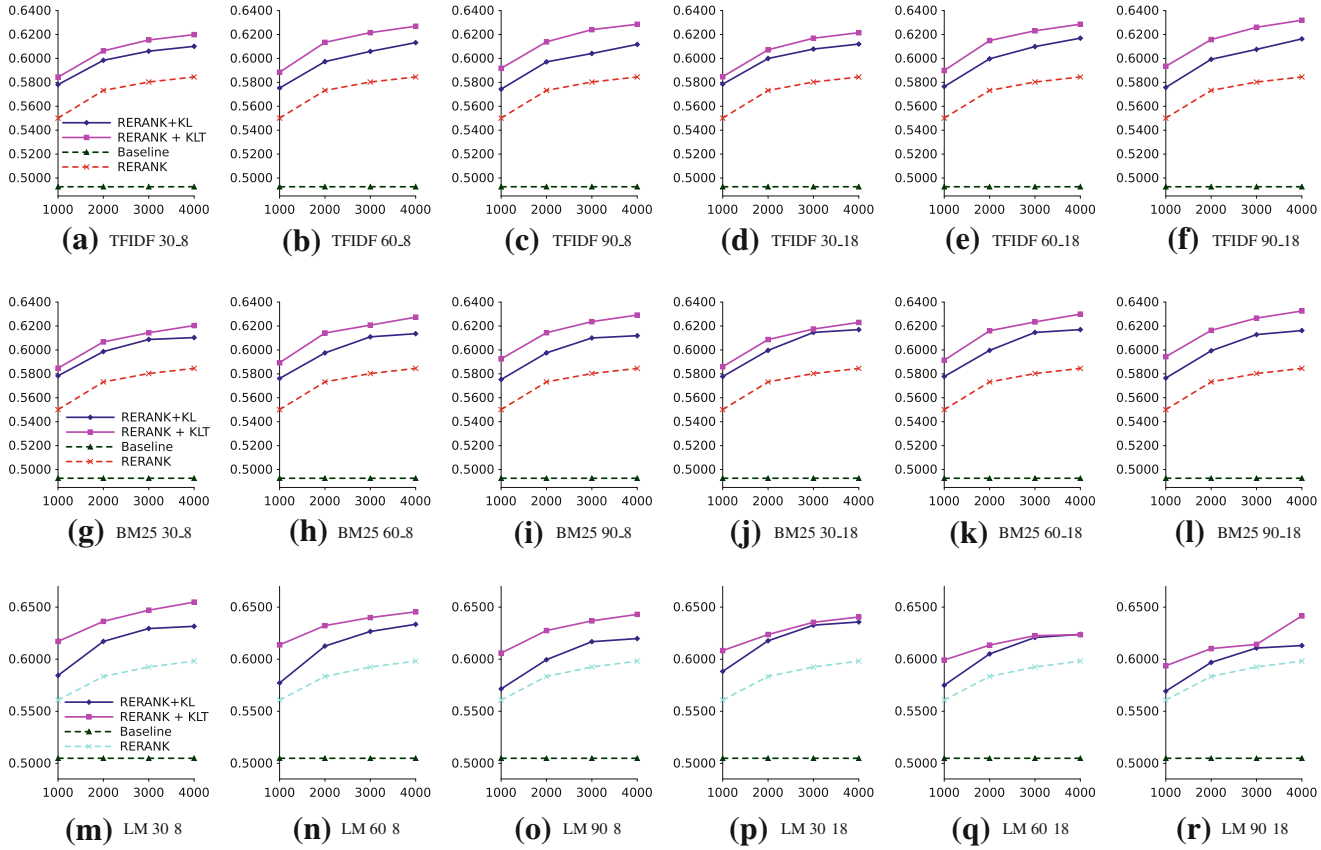
(a) TFIDF as base model **(b)** BM25 as base model **(c)** LM as base model

Fig. 6 The MAP values as function of the value of γ for three different standard retrieval models as base for the query expansion models. In each figure from **a**–**c**, each graph represents a combination of K and n values, expressed as $\{K\}\_\{n\}$



(a) TFIDF as base model **(b)** BM25 as base model **(c)** LM as base model

Fig. 7 The MAP values as function of the values of K and n, expressed as $\{K\}\_\{n\}$, with the three different retrieval models as bases for the different query expansion models



(a) TFIDF 30_8 **(b)** TFIDF 60_8 **(c)** TFIDF 90_8 **(d)** TFIDF 30_18 **(e)** TFIDF 60_18 **(f)** TFIDF 90_18

(g) BM25 30_8 **(h)** BM25 60_8 **(i)** BM25 90_8 **(j)** BM25 30_18 **(k)** BM25 60_18 **(l)** BM25 90_18

(m) LM 30 8 **(n)** LM 60 8 **(o)** LM 90 8 **(p)** LM 30 18 **(q)** LM 60 18 **(r)** LM 90 18

Fig. 8 The MAP values as function of the of number of documents to rerank, different retrieval models and different values of K and n

temporal KL (RERANKING+KLT). As before, we varied the values of n and K.

Figure 8 presents the results from this experiment. It depicts several graphs comparing the retrieval performance of the above approaches, using different combinations of the size of the feedback document set K and the number of candidate query expansion terms.

So Fig. 8a shows the results from running QE with TFIDF as a base retrieval model[12], and with $K = 30$ and $n = 8$, and so on. As we can observe in this figure, in all our tests, our proposed KLT with the reranking outperforms the standard KL. Moreover, we can see that in all the cases, we obtained the highest MAP values with R=4,000. And, as before, all the improvements of KLT are statistically significant at 95 % confidence interval.

6.3.3 KLT versus KLST

In this subsection, we compare the temporal-aware query expansion model with the spatio-temporal-aware query

[12] A base model is the retrieval model we run prior to a QE step.

Table 6 Percentage of improvement of MAP and RP using different reweighting model on BM25

	$\Delta MAP(\%)$			$\Delta RP(\%)$		
	KL	KLT	KLST	KL	KLT	KLST
30_8	2.97	8.20	**10.84**^{12}	3.74	8.29	**10.59**^{12}
60_8	4.57	10.66	**15.06**^{123}	5.46	9.79	**16.66**^{123}
90_8	2.91	13.19	**16.77**^{123}	3.02	11.30	**17.42**^{123}
120_8	2.61	13.34	**16.94**^{123}	2.47	11.01	**17.63**^{123}
30_18	3.62	8.58	**11.44**^{123}	4.35	8.61	**12.30**^{123}
60_18	3.62	12.25	**15.55**^{123}	4.35	11.30	**16.63**^{123}
90_18	2.95	14.33	**16.85**^{123}	3.19	13.38	**18.30**^{123}
120_18	2.84	14.52	**17.25**^{123}	3.14	13.29	**19.05**^{123}

The numbers 1, 2 and 3 in superscript in the table indicate the statistical significance improvements on the baseline, KL and KLT reweighting models, respectively

Bold indicates statistically significant highest values of ΔMAP and ΔRP at different values of feedback documents (*K*) and selected terms (*n*) (then for each half row)

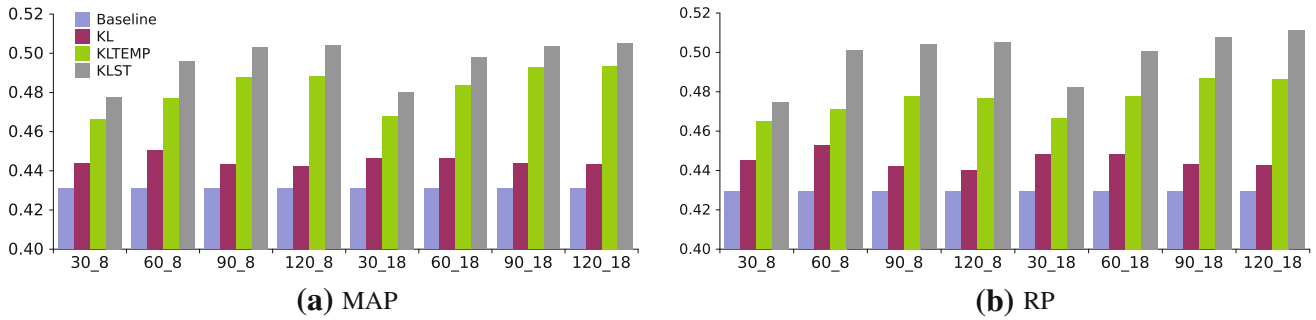


Fig. 9 Comparison of MAP and RP values of KL_ST against other query expansion models, as function of the values of K and n [expressed as $\{K\}\_\{n\}$], using BM25

expansion model KLST. We use the values of $\gamma = 0$ for the linear combination between KL and KLT, which has been shown to yield the best result. Further, we set $\delta = 0.5$ to compute KLST(t) as given by Eq. 14. Due to the space limitation, we do not present any tuning process for the δ value.

Table 6 summarizes our comparison experiments. In this table, we show how much our proposed query expansion models, KLT and KLST, improve the retrieval performance, i.e., the MAP and RP values, as compared to the base IR model BM25 and the baseline KL. The temporal window used is 3 days. For both the MAP and RP values, the first columns of the table is the percentage improvement from BM25 to KL; the second column is the percentage improvement from BM25 to KLT; while the third column is the percentage improvement from BM25 to KLST. Note that as mentioned previously, we have omitted the results from applying the Bo1-based QE model because they were not significantly different from the KL results. Further, we chose to include BM25 as the base IR model here since it was the model that gave us the overall best results, compared to TFIDF and LM (see also Sect. 6.3.2).

Analyzing the results in Table 6, we can observe that for all combinations of the number of feedback documents K and number of expansion terms n, our geo-spatial and temporal-based QE model, KLST, outperforms both the baseline KL and our temporal-based, KLT, reweighting model. Specifically, KLST is from 10.6 to 19 % better than the baseline method. Moreover, with the best MAP and RP values, KLST is six times better than the baseline KL and around 50 % better than the KLT model. To give a better overview of our comparison, Fig. 9 depicts the differences between the four models with respect to the MAP and RP values. As this figure shows, the effectiveness of our KLST model is noticeably better than the three other models, which also specifically answers our third research question in Sect. 3.

7 Conclusions

Photo-sharing applications, such as Flickr, contain many pictures related to real life events, and many of them are annotated with time and location information. The main goal of this work has been to improve existing retrieval models

by exploiting this information within event-related image retrieval. Our main idea has been to use picture metadata to emulate a query-by-example analogy. To achieve this goal, we have proposed an extended query expansion model that exploits the temporal information of pictures and the spatial distribution of terms. We thoroughly evaluated our approach by first analyzing the retrieval effectiveness with respect to different combinations of metadata fields, and using different standard retrieval models. Then, we conducted several experiments to assess the effectiveness of our two proposed query expansion models; one based on temporal proximity of tag terms and the other based on spatial distribution of tag terms. We compared both methods with existing baseline approaches. The results of these experiments have shown that our approach outperforms the state-of-the-art query expansion models, and that the improvements were statistically significant at a $p < 0.05\,\%$ level. In particular, we demonstrated that our method is effective even when the amount of information surrounding a picture is small. Finally, by testing our approach on a large dataset, and still getting good results, we can conclude that our approach can handle large-scale data.

Nevertheless, there are still interesting results and aspects of this work that we have omitted, but will be part of our future research. More specifically, we are currently investigating the effects of including semantic similarities among terms using and linking to knowledge bases, such as *Wikipedia*, in term reweighting. We are also investigating the possibility of integrating features from (web-based social) user interactions to further improve our retrieval performance.

Acknowledgments We acknowledge the anonymous reviewers' comments and suggestions, which have been valuable in improving the quality of our manuscript.

References

1. Allan J, Papka R, Lavrenko V (1998) On-line new event detection and tracking. In: Proceedings of ACM SIGIR 1998. ACM, New York, pp 37–45
2. Amati G, Joost C, Rijsbergen V (2002) Probabilistic models for information retrieval based on divergence from randomness. ACM Trans Inf Syst 20(4):357–389
3. Baeza-Yates R, Ribeiro-Neto B (2011) Modern information retrieval: the concepts and technology behind search. Addison-Wesley, New York
4. Becker H, Naaman M, Gravano L (2010) Learning similarity metrics for event identification in social media. In: Proceedings of WSDM 2010, pp 291–300
5. Becker H, Naaman M, Gravano L (2011) Beyond trending topics: real-world event identification on twitter. In: Proceedings of ICWSM 2011
6. Brants T, Chen F, Farahat A (2003) A system for new event detection. In: Proceedings of ACM SIGIR 2003. ACM, New York, pp 330–337
7. Brenner M, Izquierdo E (2012) Social event detection and retrieval in collaborative photo collections. In: Proceedings of ACM ICMR 2012. ACM, New York
8. Buscaldi D, Rosso P, Sanchis E (2006) A wordnet-based indexing technique for geographical information retrieval. In: CLEF, pp 954–957
9. Carpineto C, de Mori R, Romano G, Bigi B (2001) An information-theoretic approach to automatic query expansion. ACM Trans Inf Syst 19:1–27
10. Chakrabarti D, Punera K (2011) Event summarization using tweets. In: Proceedings of ICWSM 2011
11. Choi J, Croft WB (2012) Temporal models for microblogs. In: Proceedings of the 21st ACM international conference on information and knowledge management, CIKM '12. ACM, New York, pp 2491–2494. ISBN 978-1-4503-1156-4
12. Efron M, Golovchinsky G (2011) Estimation methods for ranking recent information. In: Proceedings of ACM SIGIR 2011. ACM, New York, pp 495–504
13. Fu G, Jones CB, Abdelmoty AI (2005) Ontology-based spatial query expansion in information retrieval. In: Proceedings of the 2005 OTM confederated international conference on "On the move to meaningful internet systems: CoopIS, COA, and ODBASE"-volume part II, OTM'05. Springer, Berlin
14. Gkalelis N, Mezaris V, Kompatsiaris I (2010) A joint content-event model for event-centric multimedia indexing. In: Proceedings of IEEE ICSC '10. IEEE Computer Society, New York, pp 79–84
15. He B, Ounis I (2009) Finding good feedback documents. In: Proceedings of ACM CIKM 2009. ACM, New York, pp 2011–2014
16. Hiemstra D (2001) Using language models for information retrieval. University of Twente, Netherlands. ISBN 978-90-75296-05-1
17. Jones R, Diaz F (2007) Temporal profiles of queries. ACM Trans Inf Syst 25(3). ISSN 1046–8188
18. Keikha M, Gerani S, Crestani F (2011) Time-based relevance models. In: Proceedings of ACM SIGIR 2011. ACM, New York, pp 1087–1088. ISBN 978-1-4503-0757-4
19. Lafferty J, Zhai C (2001) Document language models, query models, and risk minimization for information retrieval. In: Proceedings of ACM SIGIR 2001. ACM, New York, pp 111–119
20. Lavrenko V, Croft WB (2001) Relevance based language models. In: Proceedings of ACM SIGIR 2001. ACM, New York, pp 120–127
21. Lee JH (1997) Analyses of multiple evidence combination. In: Proceedings of ACM SIGIR 1997. ACM, New York, pp 267–276
22. Li X, Croft WB (2003) Time-based language models. In: Proceedings of the twelfth international conference on information and knowledge management, CIKM '03. ACM, New York, pp 469–475. ISBN 1-58113-723-0
23. Lin J, Dyer C (2010) Data-intensive text processing with MapReduce. In: Synthesis lectures on human language technologies. Morgan and Claypool Publishers, San Rafael
24. Liu X, Troncy R, Huet B (2011) Finding media illustrating events. In: Proceedings of the ACM ICMR 2011. ACM, New York, pp 58:1–58:8
25. Long R, Wang H, Chen Y, Jin O, Yu Y (2011) Towards effective event detection, tracking and summarization on microblog data. In: Proceedings of WAIM 2011. Springer, Berlin, pp 652–663
26. Lv Y, Zhai C (2010) Positional relevance model for pseudo-relevance feedback. In: Proceedings of ACM SIGIR 2010. ACM, New York, pp 579–586
27. Moxley E, Kleban J, Manjunath BS (2008) Spirittagger: a geo-aware tag suggestion tool mined from flickr. In: Proceedings of the 1st ACM international conference on multimedia information retrieval, MIR '08. ACM, New York, pp 24–30. ISBN 978-1-60558-312-9

28. Papadopoulos S, Troncy R, Mezaris V, Huet B, Kompatsiaris I (2011) Social event detection at MediaEval 2011: challenges, dataset and evaluation. In: MediaEval 2011 Workshop, Pisa, Italy, September 1–2
29. Papadopoulos S, Zigkolis C, Kompatsiaris Y, Vakali A (2011) Cluster-based landmark and event detection for tagged photo collections. IEEE Multimed 18(1):52–63
30. Perea-Ortega JM, Ure na López LA (2012) Geographic expansion of queries to improve the geographic information retrieval task. In: Proceedings of the 17th international conference on applications of natural language processing and information systems, NLDB'12. Springer, Berlin. ISBN 978-3-642-31177-2
31. Pérez-Agüera J, Araujo L (2008) Comparing and combining methods for automatic query expansion. Adv Nat Lang Process Appl Res Comput Sci 33:177–188
32. Pu Q, He D, Li Q (2009) Query expansion for effective geographic information retrieval. In: Proceedings of the 9th cross-language evaluation forum conference on evaluating systems for multilingual and multimodal information access, CLEF'08. Springer, Berlin. ISBN 3-642-04446-8, 978-3-642-04446-5
33. Quack T, Leibe B, Van Gool L (2008) World-scale mining of objects and events from community photo collections. In: Proceedings of the 2008 international conference on content-based image and video retrieval. ACM, New York, pp 47–56
34. Rae A, Murdock V, Popescu A, Bouchard H (2012) Mining the web for points of interest. In: Proceedings of ACM SIGIR 2012. ACM, New York, pp 711–720
35. Robertson SE, Walker S (1994) Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In: Proceedings of the ACM SIGIR 1994. Springer, New York Inc., pp 232–241
36. Robertson S, Zaragoza H, Taylor M (2004) Simple BM25 extension to multiple weighted fields. In: Proceedings of ACM CIKM 2004. ACM, New York, pp 42–49
37. Rocchio JJ (1971) Relevance feedback in information retrieval. In: Salton G (ed) The SMART retrieval system — experiments in automatic document processing. Prentice-Hall, Englewood Cliffs, NJ
38. Ruocco M, Ramampiaro H (2010) Event clusters detection on flickr images using a suffix-tree structure. In: Proceedings of the 2010 IEEE international symposium on multimedia, ISM'10. IEEE Computer Society, Washington, DC, USA, pp 41–48. doi:10.1109/ISM.2010.16
39. Ruocco M, Ramampiaro H (2012) Heterogeneous tag-point patterns for ranking and extracting hot-spot related tags. In: Proceedings of the 4th ACM SIGSPATIAL international workshop on location-based social networks, LBSN '12. ACM, New York
40. Ruocco M, Ramampiaro H (2013) A scalable algorithm for extraction and clustering of event-related pictures. Multimed Tools Appl:1–34. ISSN 1380–7501
41. Ruocco M, Ramampiaro H (2013) Exploring temporal proximity and spatial distribution of terms in web-based search of event-related images. In: Proceedings of the 24th ACM HT 2013. ACM, New York
42. Sakaki T, Okazaki M, Matsuo Y (2010) Earthquake shakes twitter users: real-time event detection by social sensors. In: Proceedings of the WWW 2010. ACM, New York, pp 851–860
43. Shaw JA, Fox EA (1994) Combination of multiple searches. In: The second text REtrieval conference (TREC-2), pp 243–252
44. Sigurbjörnsson B, van Zwol R (2008) Flickr tag recommendation based on collective knowledge. In: Proceedings of the WWW 2008. ACM, New York, pp 327–336
45. Silva A, Martins B (2011) Tag recommendation for georeferenced photos. In: Proceedings of the 3rd ACM SIGSPATIAL international workshop on location-based social networks, LBSN '11. ACM, New York, pp 57–64. ISBN 978-1-4503-1033-8
46. Trad MR, Joly A, Boujemaa N (2011) Large scale visual-based event matching. In: Proceedings of the 1st ACM ICMR 2011. ACM, New York, pp 53:1–53:7
47. Troncy R, Malocha B, Fialho ATS (2010) Linking events with media. In: Proceedings of I-SEMANTICS
48. Watanabe K, Ochi M, Okabe M, Onai R (2011) Jasmine: a real-time local-event detection system based on geolocation information propagated to microblogs. In: Proceedings of CIKM 2011. ACM, New York, pp 2541–2544
49. Xu J, Croft WB (1996) Query expansion using local and global document analysis. In: Proceedings of ACM SIGIR 1996. ACM, New York, pp 4–11
50. Yin Z, Cao L, Han J, Luo J, Huang TS (2011) Diversified trajectory pattern ranking in geo-tagged social media. In: SDM, pp 980–991
51. Zhai C, Lafferty J (2001) Model-based feedback in the language modeling approach to information retrieval. In: Proceedings of CIKM 2001. ACM, New York, pp 403–410
52. Zhang H, Korayem M, You E, Crandall DJ (2012) Beyond co-occurrence: discovering and visualizing tag relationships from geo-spatial and temporal similarities. In: Proceedings of WSDM 2012. ACM, New York, pp 33–42