# Context-Aware Image Semantic Extraction in the Social Web

Massimiliano Ruocco
*Supervised by Prof. Heri Ramampiaro*
Norwegian University of Science and Technology
Trondheim
Norway
ruocco@idi.ntnu.no

## ABSTRACT

Media sharing applications such as Panoramio and Flickr contain a huge amount of pictures that need to be organized to facilitate browsing and retrieval. Such pictures are often surrounded by a set of metadata or image tags, constituting the image context. With the advent of the paradigm of Web 2.0 especially the past five years, the concept of image context has further evolved, allowing users to tag their own and other people's pictures. Focusing on tagging, we distinguish between static and dynamic features. The set of static features include textual and visual features, as well as the contextual information. Further, we may identify other features belonging to the social context as a result of the usage within the media sharing applications. Due to their dynamic nature, we call these the dynamic set of features. In this work, we assume that every media uploaded contains both static and dynamic features. In addition, a user may be linked with other users with whom he/she shares common interests. This has resulted in a new series of challenges within the research field of semantic understanding. One of the main goals of this work is to address these challenges.

## Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information Storage and Retrieval—*Information Search and Retrieval*

## Keywords

Image Retrieval, Tag Recommendation, Flickr, Social Web

## 1. INTRODUCTION

The advancement of digital technologies and the widespread of Web 2.0 paradigm have made media sharing communities such as *Flickr* or *Panoramio*, a common place where pictures are freely uploaded and tagged. Many people own mobile phones with camera and it is a common practice to take pictures and upload these pictures into a media sharing community. Often these pictures are accompanied by two kinds of metadata: A set of annotations, added by the user that are generally named with the term *tag*, and camera-specific metadata, i.e. the EXIF data. Temporal data and locational data are often included in the EXIF data. We call this set of information related to the picture, *contextual data*. Media sharing applications and *tagging systems*[12] normally allow users to upload pictures with their personal tags. These are free text, short and unstructured and pictures are stored in a social context. From this point of view, a picture belongs to a specific user, generally defined by a unique id. Further each user may be linked with other users with whom he/she shares common interests and some resources. As a result to these considerations, we must consider a pictures to be defined by a set of *static features* such as visual, temporal, spatial and textual information and stored in a social context. In content-based image retrieval systems[6], images were indexed based on their visual features. With the advent of the web, text based image retrieval systems index and retrieve the images based on the surrounding text, on a web scale environment. Now, because of the above characteristics there is a need to define different features model and a need to make this new features useful in order to obtain more improvement knowledge extraction and retrieval algorithms. In particular, the main contributions of this research are the follows: 1) Analyse the power of the contextual and social information in order to extract tag and picture semantics. 2) Most of the work until now in tag semantics, categorize tag in *event-related* or *locational-related*. Our work will try to improve the state of the art by using knowledge databases such as Dbpedia, in particular linking tags to dbpedia entities. 3) We will use the work of point 1 and 2 to improve image retrieval performance.

This paper is organized as follows. The State of the Art is presented in Section 2. In Section 3 we present the research questions of our research. Finally conclusion will be presented in Section 5.


*WWW 2012 Companion*, April 16–20, 2012, Lyon, France.
ACM 978-1-4503-1230-1/12/04.

## 2. STATE OF THE ART

Since the tag associated to a pictures are freely-chosen, subjective, short and unstructured, there is a need to give some kind of meaning and semantic to the tags. The purposes of this are generally to improve image search, support the tag suggestion process and facilitating the browsing of pictures. Extracting semantics from tags is a field treated in different works in the literature[16, 2]. In these works semantics means to categorize tags in *event-related* or *location-related tags*.

In [13], starting from a geo-referenced collection , their approach learn tag semantic by mining the dataset. The tags are categorized as place (i.e. `Paris`), time event (i.e.

2008), landmark (i.e. `Tour Eiffel`) and visual description (`sunset`, `sky`). For this purpose the features used to mine the dataset was time, location, visual information and tags co-occurrence. For place extraction they extend the SSI idea, using quad-tree for space scale definition and using Jaccard measure for co-occurrence. In [10] an event/activity detection framework is presented. It aims to infer generic activities to a picture from Flickr by using a probabilistic approach that incorporates geographical information. In practice, a classifier is trained for each kind of activity, such as *Golf*, *Hiking*, *Lake* according to the visual information. Late fusion from results of visual and geo-classifier show improvement in precision. In this work *Geonames*[1] is used for reverse geo-coding.

# 3. RESEARCH QUESTIONS

Our research starts with the following research question:
**RQ:** *How can context information be used to improve image retrieval?*
Our research will focus primarily around this research question. In order to catch all the challenges related to this question we need to decompose and split the principal research question in the following sub-questions:
**RQ 1.1:** *How can semantic be extracted from the textual information associated to a picture?*
**RQ 1.2:** *May contextual information be useful to extract semantic from the textual information associated to a picture?*
**RQ 1.3:** *How can other open content resources or knowledge database be useful for tag semantic extraction?*
**RQ 1.4:** *May the extracted semantic be useful for automatic annotation and retrieval purpose?*
So far we already got some results useful to answer RQ1.1 and RQ1.2 and we are currently working on RQ1.3.

# 4. EXPECTED CONTRIBUTIONS

The contributions of this research will be in the area of extraction of tag semantics by first considering contextual and social information and then using the same features to associate a tag to an entity in knowledge databases. In addition, we will analyse the impact of the tag semantics association in retrieval systems and in automatic tag annotation.

To evaluate our work, our experiments will be conducted on dataset consisting of Flickr pictures. Different challenges have been encountered using this dataset. Lack of contextual information (not all the pictures contain spatial tag) and short, subjective and not always reliable tags are examples of challenges we have to deal with.

In order to extract tags and picture semantics, we will explore three different but related approaches. First, we will propose a scalable algorithm to group event-related pictures, by employing spatial and temporal metadata. Second, we will exploit the possibility to employ spatio-temporal statistic algorithms to catch the spatial and spatio-temporal aggregation capabilities of the tag point-patterns to categorize tags in event-tag or locational-tag to improve the existing co-occurrency measure for tag recommendation[20]. Third and finally we will explore the possibility to enrich tag semantics by linking them with entities of existing knowledge databases (such as *Dbpedia*). In the following subsection we will explain our preliminary results, motivation and future works in these three directions.

[1]http://www.geonames.org

## 4.1 A scalable approach for clustering and Extraction of Event Related Pictures

We have proposed a mining algorithm to extract and cluster group of images representing events (i.e. *wedding*, *football matches*, *parties*). Most work in event detection has mainly been done on text documents, where event were defined as *something happening in a certain place at a certain time* [1]. However, focusing on our context, this definition is too limited. In this work we see an event as *something happening in a certain place at a certain time and tagged with a certain term* in the context of social media such as Flickr or Panoramio where pictures are normally surrounded by a set of metadata (textual annotation, temporal and spatial information).

The proposed approach[18] is based on the well-known *Suffix Tree Clustering Algorithm*[22], previously used only on textual documents. In contrast to the original method, to extract our base clusters, we do not need to consider all the nodes of the suffix tree that group at least two documents. We first prune the set of nodes according to our definition of event. All the node labelled with a temporal label $V$, a geographical label $G$ and a tag label $t$ are possible events. We detect these nodes, and we will call them *candidate event clusters* $S_{VGt}$ by traversing the Suffix Tree for all the nodes achievable by a sequence of temporal branch, positional branch and tag branch. Each such a cluster will then be tagged with the sequence of label of the branches passed. This means that each candidate event cluster is a collection of pictures grouped by a certain time slice, a certain geographical area, and tagged with a certain term. According to the definition in [18] we must compare the $S_{VGt}$ set with the collection of images grouped by the same geographical area $G$ and the same tag $t$, $S_{Gt}$. This hypothesis always holds because a tag representing an event can only belong to a single combination of date/time and a geographical area. It still holds even if we have a situation where an object (image object) or a place appear in several images taken over a long period of time. Although the tags for these images will have the same geo-tags but different time tags, their combinations are still unique.

To capture this, in our tree structure, the images in the set $S_{VGt}$, ideally, must be the same images of the set $S_{Gt}$. Thus all the candidate event clusters $S_{VGt}$ labelled with a temporal label, a geographical label and a tag label will be compared with the set $S_{Gt}$. In practice, considering noise in the tags, and in the temporal and spatial information associated to the picture, we can relax this assumption. The assumption , for a given $S_{VGt}$, to be a an *event cluster* becomes the following:

$$\frac{|S_{VGt} \cap S_{Gt}|}{\max(S_{VGt}, S_{Gt})} \geq K \quad (1)$$

Using $K = 1$ means that $S_{Gt}$ is equal to $S_{VGt}$. Using a smaller $K$ we can control the assumption according to the noise in the dataset and the spatio/temporal granularity used for the extension of the tags.

This work has been extended by implementing a new algorithm that were enriched with a more sophisticated step for merging group of images representing the same events. The intuition behind this last refinement block is inspired from

the well know DBSCAN[5] clustering algorithm. Extended analysis on the dataset are also provided. This hypothesis always holds because a tag representing an event can only belong to a single combination of date/time and a geographical area. It still holds even if we have a situation where an object (image object) or a place appear in several images taken over a long period of time. Although the tags for these images will have the same geo-tags but different time tags, their combinations are still unique. We performed analysis on a dataset of around 240K pictures. We analysed the behaviour of the parameter $K$ introduced in the equation 1 to relax the hypothesis of event cluster. We will perform the algorithm on value $K = 1$ and $K = 0.8$ (see Figure 1)

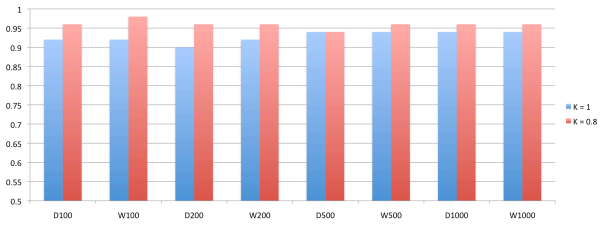


**Figure 1: Precision at different values of $K$ over different granularities in space and time**

To summarize, our principal contributions are the following: Use of spatial and temporal information in order to extract picture semantics, employing a scalable algorithm feasible for large-scale dataset, extension of a clustering algorithm used previously only in dataset of textual documents and extension of the previous definition of event, in the context of media sharing application.

## 4.2 Spatio Temporal Analysis of tag point pattern

The datasets considered consist of pictures from photo-sharing tools. These images are associated with metadata. In particular we consider, for each picture, the locational information represented by a pair of real number $\mathbf{g} = (lat, lon)$ representing latitude and longitude, the temporal information $t$ represented by the timestamp, and the set of tags, the collection of term $T$. Hence, these photos are characterized to be placed in a spatial(-temporal) domain. Each picture considered has been taken in a certain time and in a certain place.

Formally if we consider each term $t_i$ in the Vocabulary $V$ constructed from the dataset, we can consider to have a set of $M$ points $s_{1,t_i}, ..., s_{M,t_i}$ representing the term distribution in the spatio-temporal domain. The presence of a term $t$ in a certain point in the spatial(-temporal) domain, is represented by the picture tagged with $t$. We call this distributions *tag-point pattern* and each of these distributions may be modelled by a random variable.

The assumption is that event-related tags are grouped in spatio-temporal and temporal space, while locational-tags are grouped in the spatial domain. This is not a trivial problem since the underlying picture distribution is not homogeneous (not equally distributed) in both the spaces (*heterogeneous point pattern*, see Figures 2 and 3).

In order to capture the tag semantics we will explore and analyse the existing spatio-temporal and spatial clustering algorithm in order to evaluate the *clusterability* of each point

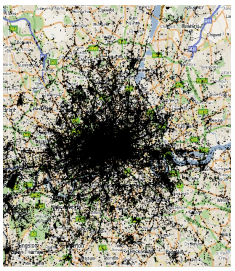

**Figure 2: Spatial distribution of the pictures collected**

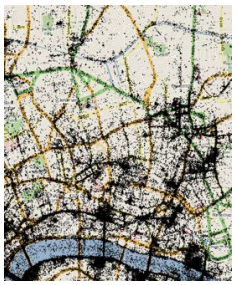

**Figure 3: A spatial zoom of the pictures dataset**

pattern. As part of this work, we will investigate the use of statistical methods to analyse the spatial and spatio-temporal regularity of tag-point patterns. As part of this we have already performed some preliminary analysis applying global methods such as *Complete Spatial Randomness* (CSR) Test[4]. Existing method for testing the CSR hypothesis may be divided in *quadrat method*, *nearest-neighbour method* and *method of K-functions* [3]. The first method divide the sampling windows $W$ in $n$ quadrants (square regions of equals area). The randomness can be tested by using a *index of dispersion test*. The drawback of this procedure is that it is strictly related to the size of the selected square. Different types of index of dispersion try to overcome this problem ([3] and [7]). The second approach is based on observing the distances from each point $m_i$ to all its nearest neighbour in $W$. Different statistics exist to test CSR Hypothesis based on nearest neighbours. The drawback of this method is that since they use only the closest events, it only catches the smallest scale of the pattern and then only the smallest-scale clustering tendency. The third method, which we will base our work on is the usage of *K-function*. This point pattern analysis takes into account the scale effect and permit the exhibition of different structures at different scales. The *K-function method* or *reduced second moment order* was originally proposed in [17] for homogeneous and isotropic spatial point process over the whole space and consider randomly sampled cells of different sizes. The definition is:

$$\lambda K(h) = E(\#(\text{evts within dist. } h \text{ of an evt})) \qquad (2)$$

where $\lambda$ is assumed constant throghout $\mathbf{R}$.

As part of this work we have already done preliminary studies on some tags of pictures of a dataset gathered from Flickr of 1.5M of pictures, by performing the inhomogeneous version of the *Ripley's K-function*. For an inhomogeneous Poisson process with $\lambda(x)$, as known first-order intensity the estimator of $K(r)$ is:

$$\hat{K}(r, \lambda) = \frac{1}{a(A)} \sum_{i=1}^{n} \sum_{j \neq i} \frac{w_{ij} I(d_{ij} \leq r)}{\lambda(x_i)\lambda(x_j)} \qquad (3)$$

where $I(d_{ij} \leq h)$ is the same indicator of the equation 2, $w_{ij}$ is the edge corrector. In general, the intensity function $\lambda(x)$ is unknown and need to be estimated. The CSR test may be performed by comparing the empirical $\hat{K}(r)$ with $K(r)$. In particular, $\hat{K}(r) > \pi r^2$ indicates some degree of clustering at scale $r$ and $\hat{K}(r) < \pi r^2$ indicates some degree

of dispersion at scale $r$. To standardize the $K$-function it is possible to define the $L$-function:

$$L(r) = \sqrt{\frac{K(r)}{\pi}} \quad (4)$$

Under CSR $L$-function is $r$. Moreover $L(r) > r$ indicates clustering at scale $r$ and $L(r) < r$ indicates dispersion at scale $r$. These studies, by far show the effectiveness of the

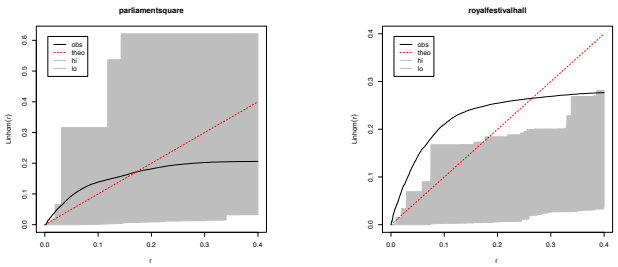


**Figure 4: $L$ function for `royalfestivalhall` and `parliamentsquare` tag-point pattern**

use of these statistic. We evaluated the function L for inhomogeneous tag-point patterns in order to evaluate the CSR, for a subsample (around 1000 points) of the tag-point pattern. In Figure 4.2 we show our preliminary results. We can see that both the tag point pattern show a *clusterability* until reaching a diameter of around 200 meters for `parliamentsquare` and around 300 meters for `royalfestivalhall`. The problems to be solved are first the computational time cost of this global method and second to find metrics to compare the *clusterability* between two tag-point patterns. Although this method has generally been used in epidemiology, to our best knowledge, the application and comparative analysis of Spatio-Temporal statistics[21] for analysing spatio and temporal distribution of the pictures-related tags are still lacking.

To summarize, the motivations behind this study are the following: First give semantics to the tags of a picture and analyse the power of existing spatio-temporal statistics (global and local[21]) and second, enhance the existing co-occurrence measures[20] for automatic annotation purposes including spatial and spatio-temporal co-occurrence distance.

## 4.3 Linking Flickr Tag to Dbpedia entity

As already mentioned, tags are free, subjective and can contain noise. For these reasons they can be unreliable . Existing approaches try to tackle this problem by performing *tag re-ranking*([11, 23]). In these works the tags related to the pictures are ranked according to some relevance measures. We will study the possibility to associate a tag to a *Dbpedia* entity page. According to [14], the use of Dbpedia can enrich the tag semantics and is more reliable than Wordnet. To our best knowledge, however existing works have been mainly focused on entity extraction from textual documents[8] or tweets. The challenges in linking Flickr Tag to *Dbpedia* entity is the lack of context considering a single picture. The tags are short text, and that disambiguation problem is here difficult to solve by applying the approaches normally used for textual documents. Our idea to solve the disambiguation problem is to leverage neighbours pictures according to different features as mentioned before.

## 4.4 Application: Event Retrieval

So far the pictures are considered surrounded with spatial and temporal metadata. In the real case there is a lack of contextual metadata. In Flickr, only 10% of pictures are geotagged. An event-related image retrieval system must consider this lack of information and deal with it.
In this direction we perform some preliminary experiments in a proposed event retrieval system[19]. In particular, we proposed a system to solve two specific event extraction challenges. In the first challenge, the main goal was to retrieve all soccer events in Rome and Barcelona, while in the second challenge, we were asked to retrieve all events from two specified venues in Amsterdam (NL) and Barcelona (ES) within a certain temporal range. The results of the queries were presented as groups of images - i.e, one group per event. More details on the proposed approach and results can be founded in [15].

The system presented is composed by different blocks. First a query expansion is performed by using two knowledge databases such as *LastFm*[2] and *Dbpedia*, to get the venue names in different languages and their location. A set of textual and spatial queries were submitted to a search engine. Then the results were categorized in the following block to filter these results. Next, a temporal clustering algorithm has been performed on the retrieved and filtered list. Here we use the *quality threshold*[9] clustering algorithm. Finally a refinement and a semantic merge step on the resulting clusters were employed.

For each query different runs were presented. For the first query we presented 2 runs: 1) categorization performed by considering only Tag and 2) categorization performed by using all textual metadata (`Title`, `Description` and `Tag`). For the second challenge 3 runs were presented: 1) No Refinement step 2) Refinement with top-100 tags, 3) Refinement with entity names. The results can be summarized in the following points: 1) Tag metadata were more descriptive, 2) better performance were obtained using entity names in the refinement block, 3) The refinement block was useful to increase recall. Figures 5 and 6 also summarize the results with respect to Precision, Recall, NMI, F-Score metrics.

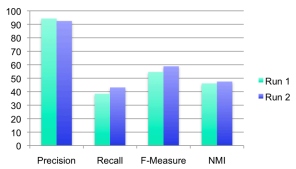


**Figure 5: First query**

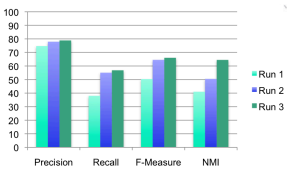


**Figure 6: Second query**

Our future work here will go in the direction of improving the refinement step by merging the ranked lists coming from pseudo relevance feedback employed on spatial, temporal and visual features.

# 5. CONCLUSIONS

Contextual data represent an extra and useful source of information related to resources in the media sharing applications. In order to take advantage of these contextual information we propose two mining algorithm to extract semantic from the tags leveraging on spatial and temporal metadata

[2]http://www.last.fm

with promising results. We believe also that it is necessary to compare these results with a second method such as linking the tag with the entity of an existing knowledge database (i.e. *Dbpedia*). To deal with the challenges caused by the short nature of the textual tags, we need some clues coming from contextual neighbours, in large dataset. Possible applications from the tag extraction are event-retrieval systems and automatic annotation systems. We propose a baseline event-retrieval system to evaluate preliminary results by successfully employing knowledge databases in the query expansion block. Further, we propose cluster refinement block based on entity extraction to improve the quality of extracted clusters.

## 6. ACKNOWLEDGEMENTS

This work is supported by the Research Council of Norway, grant number 176858 under the VERDIKT program.

## 7. REFERENCES

[1] Allan, et al. Topic detection and tracking pilot study final report. In *In Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop* (1998), pp. 194–218.

[2] Chen, L., and Roy, A. Event detection from flickr data through wavelet-based spatial analysis. *Proceeding of the 18th ACM conference on Information and knowledge management CIKM 09 09* (2009), 523.

[3] Cressie, N. A. C. *Statistics for Spatial Data*, rev sub ed. Wiley-Interscience, Jan. 1993.

[4] Diggle, P. J. *Statistical Analysis of Spatial Point Patterns*. Hodder Arnold Publishers, London, 2003.

[5] Ester, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proc. of 2nd International Conference on Knowledge Discovery and* (1996), pp. 226–231.

[6] Flickner, M., et al. Query by image and video content: The qbic system. *Computer 28* (1995), 23–32.

[7] Greig-Smith, P. *Quantitative plant ecology*, vol. 9. Butterworths, 1983.

[8] He, J., de Rijke, et al. Automatic link generation with wikipedia: A case study in annotating radiology reports. In *20th ACM Conference on Information and Knowledge Management (CIKM 2011)* (Glasgow, 2011), ACM, ACM.

[9] Heyer, et al. Exploring Expression Data: Identification and Analysis of Coexpressed Genes. *Genome Research 9*, 11 (Nov. 1999), 1106–1115.

[10] Joshi, D., and Luo, J. Inferring generic activities and events from image content and bags of geo-tags. *Proceedings of the 2008 international conference on Contentbased image and video retrieval CIVR 08* (2008), 37.

[11] Liu, D., et al. Tag ranking. In *Proceedings of the 18th international conference on World wide web* (New York, NY, USA, 2009), WWW '09, ACM, pp. 351–360.

[12] Marlow, et al. Ht06, tagging paper, taxonomy, flickr, academic article, to read. In *Proceedings of the seventeenth conference on Hypertext and hypermedia* (New York, NY, USA, 2006), HYPERTEXT '06, ACM, pp. 31–40.

[13] Moxley, et al. Not all tags are created equal: learning flickr tag semantics for global annotation. In *Proceedings of the 2009 IEEE international conference on Multimedia and Expo* (Piscataway, NJ, USA, 2009), ICME'09, IEEE Press, pp. 1452–1455.

[14] Overell, S., Sigurbjörnsson, B., and van Zwol, R. Classifying tags using open content resources. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining* (New York, NY, USA, 2009), WSDM '09, ACM, pp. 64–73.

[15] Papadopoulos, S., et al. Social Event Detection at MediaEval 2011: Challenges, Dataset and Evaluation. In *MediaEval 2011 Workshop* (Pisa, Italy, September 1-2 2011).

[16] Rattenbury, et al. Towards automatic extraction of event and place semantics from flickr tags. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval* (New York, NY, USA, 2007), ACM, pp. 103–110.

[17] Ripley, B. D. The second-order analysis of stationary point processes. *Journal of Applied Probability 13* (1976), 255–266.

[18] Ruocco, M., and Ramampiaro, H. Event clusters detection on flickr images using a suffix-tree structure. In *Proceedings of the 2010 IEEE International Symposium on Multimedia* (Washington, DC, USA, 2010), ISM '10, IEEE Computer Society, pp. 41–48.

[19] Ruocco, M., and Ramampiaro, H. Ntnu @ mediaeval2011: Social event detection task (sed). *Working Notes Proceedings of the MediaEval 2011 Workshop* (2011).

[20] Sigurbjörnsson, B., and van Zwol, R. Flickr tag recommendation based on collective knowledge. In *Proceeding of the 17th international conference on World Wide Web* (New York, NY, USA, 2008), WWW '08, ACM, pp. 327–336.

[21] Tango, T. *Statistical methods for disease clustering.* New York, NY: Springer, 2010.

[22] Zamir, O., and Etzioni, O. Web document clustering: a feasibility demonstration. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval* (New York, NY, USA, 1998), SIGIR '98, ACM, pp. 46–54.

[23] Zhuang, J., and Hoi, S. C. A two-view learning approach for image tag ranking. In *Proceedings of the fourth ACM international conference on Web search and data mining* (New York, NY, USA, 2011), WSDM '11, ACM, pp. 625–634.