Tag-Based Social Image Retrieval: An Empirical Evaluation

Aixin Sun, Sourav S. Bhowmick, and Khanh Tran Nam Nguyen
School of Computer Engineering, Nanyang Technological University, Singapore 639798, Singapore.
E-mail: {axsun,assourav}@ntu.edu.sg, namkhanh@pmail.ntu.edu.sg

Ge Bai[1]
School of Computer Science, Fudan University, Shanghai 200433, P. R. China,
E-mail: 07300720208@fudan.edu.cn

Tags associated with social images are valuable information source for superior image search and retrieval experiences. Although various heuristics are valuable to boost tag-based search for images, there is a lack of general framework to study the impact of these heuristics. Specifically, the task of ranking images matching a given tag query based on their associated tags in descending order of *relevance* has not been well studied. In this article, we take the first step to propose a generic, flexible, and extensible framework for this task and exploit it for a systematic and comprehensive empirical evaluation of various methods for ranking images. To this end, we identified five orthogonal dimensions to quantify the matching score between a tagged image and a tag query. These five dimensions are: (i) *tag relatedness* to measure the degree of effectiveness of a tag describing the tagged image; (ii) *tag discrimination* to quantify the degree of discrimination of a tag with respect to the entire tagged image collection; (iii) *tag length normalization* analogous to document length normalization in web search; (iv) *tag-query matching model* for the matching score computation between an image tag and a query tag; and (v) *query model* for tag query rewriting. For each dimension, we identify a few implementations and evaluate their impact on NUS-WIDE dataset, the largest human-annotated dataset consisting of more than 269K tagged images from Flickr**. We evaluated 81 single-tag queries and 443 multi-tag queries over 288 search methods and systematically compare their performances using standard metrics including Precision at top-*K*, Mean Average Precision (MAP), Recall, and Normalized Discounted Cumulative Gain (NDCG).**

Introduction

The prevalence of digital photography devices (e.g., digital cameras, mobile phones) has led to over 200 billion images accessible online and the number is continuously growing (Yahoo!, 2010). Owing to increasing popularity of tagging activities in social media sharing platforms (e.g., Flickr), many of these services enable users to annotate images with tags. The availability of such tags as metadata has given rise to opportunities to build novel and superior tag-based techniques to enhance significantly our ability to understand social images and to retrieve them effectively and efficiently (Goh, Ang, Lee, & Chua, 2011; Jain & Sinha, 2010).

Image retrieval has been widely studied from two paradigms: *content-based* and *annotation-based* image retrieval (Carneiro, Chan, Moreno, & Vasconcelos, 2007; Datta, Joshi, Li, & Wang, 2008). The former requires users to formulate a query using an example image. The retrieval system then returns the set of images that best matches the given example based on visual content, i.e., low-level features such as color and texture. *Annotation-based* image retrieval, on the other hand, enables users to formulate naturally semantic queries using textual keywords. In order to support this retrieval paradigm, many *automatic image annotation* techniques have been proposed, which assign a few relevant keywords to an unannotated image to describe its visual content for image indexing and retrieval (Feng, Manmatha, & Lavrenko, 2004; Guillaumin, Mensink, Verbeek, & Schmid, 2009; Jeon, Lavrenko, & Manmatha, 2003; Makadia, Pavlovic, & Kumar, 2008; Yanai, Shirahatti, Gabbur, & Barnard, 2005). The keywords are often derived from a well-annotated image collection and the number of keywords is often limited to a few hundred.

Compared with these carefully selected keywords in image annotation, social tags are free-form keywords assigned by users for various purposes and not drawn from any controlled vocabulary (detailed in Tagging Motivation and Tag/Query Types). Particularly, tags annotated to an image may not necessarily describe its visual content. For example, tags assigned to images may describe the time (e.g., 2008) and location (e.g., Asia, UK) where the photos are taken, as

Received July 20, 2011; revised August 18, 2011; accepted August 18, 2011

[1] This work was done during Ge Bai's intership at NTU.

© 2011 ASIS&T • Published online 26 September 2011 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/asi.21659

well as camera brands/models (e.g., Canon, 60D). As a result, social tags are noisier than keywords selected for image annotation. For instance, more than 420K distinct tags have been used to annotate 269K images in NUS-WIDE dataset, many of these tags do not describe the visual content of these images (Sun & Bhowmick, 2010). Furthermore, in contrast to annotation-based image retrieval techniques, the number of tags assigned to one social image may differ significantly from another image. Consequently, most state-of-the-art efforts in the area of social image tagging have focused primarily on tag recommendation, disambiguation, and de-noising (Chua et al., 2009; Sigurbjörnsson & van Zwol, 2008; Tang, Yan, Hong, Qi, & Chua, 2009; Weinberger, Slaney, & van Zwol, 2008; Wu, Yang, Yu, & Hua, 2009). Although these studies suggest that recommending (or completing), disambiguating, and de-noising tags lead to superior social image retrieval experiences, the task of ranking images matching a given tag query in descending order of *relevance* has not been systematically studied. We refer to this task as *Tag-based Image Retrieval* (TAGIR).

At first glance, the lack of study may seem to reflect the assessment that TagIR is less challenging. As long as the images are tagged, many techniques developed in the IR community can be easily applied for the TAGIR problem because tags are nothing but textual terms (Manning, Raghavan, & Schtze, 2008; Salton & Buckley, 1988; Zobel & Moffat, 1998). However, a key difference between traditional IR and TAGIR is that a textual document typically has much redundancy of words to convey its semantics, whereas an image is annotated with many fewer words (i.e., tags) with no or minimal redundancy. Moreover, as mentioned above, tags are assigned by different users having different motivations for tagging, different understandings of relatedness between tags and images, or even different interpretations of the meaning of tags arising from knowledge or cultural diversity. Consequently, some of the basic settings commonly accepted in the traditional IR, such as word frequency weighting and document length normalization, demand a revisit.

In this article, we take the first step to propose a *generic*, *flexible*, *and extensible framework for* TAGIR and undertake a *systematic and comprehensive empirical evaluation* of various methods for ranking images using this framework. Our framework consists of five orthogonal dimensions that play pivotal roles in social image tagging, namely, *tag relatedness* for measuring the degree of effectiveness of a tag describing the tagged image, *tag discrimination* for quantifying the degree of discrimination of a tag with respect to the entire tagged image collection, *tag length normalization* analogous to document length normalization in web search, *tag-query matching model* for computing a matching score between an image tag and a query tag, and *query model* for *rewriting* tag queries. Because each dimension may be realized by several alternative formulations, we propose several formulations for each of the five dimensions. For instance, tag relatedness is formulated based on *unit relatedness*, *tag position*, or *user/neighbor voting* (Li, Snoek, & Worring, 2008). Observe that such a framework enables us to describe and compare various formulations associated with each dimension.

We have exhaustively evaluated the impact of these dimensions on NUS-WIDE dataset (Chua et al., 2009), the largest human-annotated dataset consisting of more than 269K images from Flickr, with 81 single-tag queries and 443 multi-tag queries. We evaluated 288 search methods in total and systematically compared their performances using Precision at top-K, Mean Average Precision (MAP), Recall, and Normalized Discounted Cumulative Gain (NDCG). Our experimental results suggest that for single-tag queries, tag relatedness, tag-query matching model, and query model are the most crucial dimensions for superior TAGIR experiences. However, for multi-tag queries, where the information need is much more specifically defined, all these dimensions become significantly less important. The presence of all tags matching a multi-tag query largely guarantees a very good ranking of search results. More complicated formulations over these dimensions typically lead to degradation of search results ranking. In summary, the major contributions of this work are as follows:

- In the third section, we present the TAGIR framework consisting of five orthogonal dimensions and discuss several formulations for each dimension. Particularly, we categorize the works reviewed in Related Work to their corresponding dimensions.
- In the fourth, fifth, and sixth sections, we conduct a systematic and comprehensive empirical evaluation of methods representing different formulations under the aforementioned five dimensions. Specifically, we evaluate 81 single-tag queries on 288 methods and 443 multi-tag queries on 72 methods. We report detailed analysis of the impact of the five dimensions and formulations for answering single-tag and multi-tag queries. The observations made in our experiments serve as a concrete reference to which dimensions need to be further improved for superior TAGIR.

Related Work

Most germane to this work are efforts in the traditional IR environment that rank documents matching a given query in descending order of relevance. Many models and relevance measures have been proposed and studied in the literature (Manning et al., 2008; Salton & Buckley, 1988; Zobel & Moffat, 1998). To explore these relevance measures systematically, Zobel and Moffat (1998) proposed a framework with eight dimensions to describe different components in a relevance measure. Example dimensions are *term weight* (e.g., inverse document frequency), *document-term weight* (e.g., term frequency), and *document length normalization*. Each dimension may be realized by many alternative formulations. For instance, more than eight formulations for document length normalization are enumerated by Zobel and Moffat (1998). Owing to the large number of combinations of alternative formulations, a subset of relevance measures was evaluated on the TREC dataset.

In this work, we share a similar objective to explore and evaluate systematically the relevance measures between tag queries and tagged social images. The major difference between our work and aforementioned efforts is that a textual document contains much redundancy of words to conveys its semantic whereas images are usually associated with only few tags. Furthermore, redundancy of tags is minimal in many social image tagging systems. Particularly, in Flickr, a tag cannot be assigned more than once to the same image. Moreover, the tags are assigned by different users with different motivations and different criteria for determining the degree of relatedness of a tag to an image. All these differences demand systematic investigation of the impact of different formulations on image search ranking.

In the following, we first distinguish TAGIR from automatic image annotation. Next, we address related work from a number of recent research efforts toward understanding image tagging, including: motivations for tagging (Ames & Naaman, 2007; Marlow, Naaman, Boyd, & Davis, 2006; Zollers, 2007), tagging systems (Golder & Huberman, 2006; Marlow et al., 2006), and tag types (Bischoff, Firan, Nejdl, & Paiu, 2008; Overell, Sigurbjörnsson, & van Zwol, 2009; Sigurbjörnsson & van Zwol, 2008, 2010); and tag relatedness (Li, Snoek, & Worring, 2009, 2010; Liu, Hua, Yang, Wang, & Zhang, 2009; Zhu, Yan, & Ma, 2010) and tag representativeness (Lu, Zhang, Tian, & Ma, 2008; Sun & Bhowmick, 2009, 2010).

Image Annotation and Retrieval

Automatic image annotation refers to the task of assigning a few relevant keywords to an unannotated image to describe its visual content; the keywords are then indexed and used to retrieve images (Feng et al., 2004; Guillaumin et al., 2009; Jeon et al., 2003; Makadia et al., 2008; Yanai et al., 2005). These keywords are often derived from a well-annotated image collection, and the latter serves as training examples for automatic image annotation. Jeon et al. (2003) assume that regions in an image can be described using a small vocabulary of blobs. Blobs are generated from low-level image features through clustering. The joint probability distribution of textual keywords and blobs is learned from the annotated image collection to compute the probabilities of keywords associating with a test image. A family of image annotation methods, built on nearest neighbor hypothesis (i.e., visually similar images likely share keywords), are proposed and evaluated by Makadia et al. (2008). Given a query image, the k-nearest neighbors are retrieved and their associated keywords are transferred to the query image. The accuracy of image annotation can be evaluated based on the correctness of the assigned keywords (Makadia et al., 2008) or through image retrieval by using the assigned annotations (Guillaumin et al., 2009).

Although image retrieval is often used to evaluate image annotation methods, the key focus of image annotation is to assign images with keywords. The dimensions in matching a textual query with the keyword-annotated images have not been systematically evaluated. In this work, our focus is to evaluate the TAGIR methods, where annotations in the form of *user-assigned tags* are provided. Moreover, in image annotation research, the keywords are carefully selected and the number of keywords is often very small. For instance, the number of keywords selected in the commonly used image annotation datasets, such as Corel5K, IARP TC12, and ESP game datasets, ranges from 100 to 500 (Guillaumin et al., 2009; Jeon et al., 2003; Makadia et al., 2008; Yanai et al., 2005). The two larger datasets Corel30K and PSU, containing 31K and 60K images, are annotated with 5,587 and 442 keywords, respectively (Carneiro et al., 2007). On the other hand, social tags are keywords assigned by users not from any controlled vocabulary. For the NUS-WIDE dataset used in this work, consisting of 269K tagged images, there are more than 420K distinct tags.

Tagging Motivation and Tag/Query Types

Motivations for tagging is one of the main factors determining the nature and types of tags, which subsequently affect TAGIR experiences. Using data from Delicious,[2] Golder and Huberman (2006) identified seven functions that tags perform, including identifying what or who it is about, what it is, who owns it, refine categories, and self-reference. These functions are largely applicable to Flickr as well (Stvilia & Jörgensen, 2010). Zollers (2007) links the seven functions with possible tagging motivations extrapolated from work by Marlow et al. (2006). These motivations are organizational, attract attention, contribution and sharing, express opinion, and self-presentation. Interviews with participants who annotated their Flickr photos revealed that most participants generally had one or two primary motivations for tagging (Ames & Naaman, 2007). The authors further developed a taxonomy of motivations for tagging consisting of two dimensions, namely, *sociality* (whether the tag's intended usage is for self or others) and *function* (whether the tag's intended use is for organization or communication). Their interviews suggest that organization for the general public (photo pools, search, self-promotion) is the primary motivation for tagging, whereas secondary motivations are self-organization (adding tags for later retrieval) and social communication (adding context for friends, family, and the public).

Design of tagging systems also implicitly affects the resultant tags. Marlow et al. (2006) propose a taxonomy with seven dimensions to describe a tagging system design. Example dimensions include *tagging right* (e.g., self-tagging, permission-based, free-for-all), *tagging support* (e.g., blind, suggested), and *aggregation model* (e.g., bag, set). For instance, bag aggregation model is used in Delicious where the number of times a tag is used to annotate a URL can be used for more objective view of the URL from users. On the other hand, the set model employed in Flickr prevents repetition of tags for a given photo. This leads to a challenging research problem of quantifying the relatedness of a tag to its annotated image, which we review later. Another

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

important dimension is the type of tagging object (i.e., textual and non-textual). Images and videos are examples of non-textual object types; news articles and blog posts are examples of textual object type. For textual data, the query keywords literally appear in the content of the data. Hence, the matching between the query keywords and the content of textual object to be retrieved remains a key role in the retrieval. The tags are considered as additional information over the content (Berendt & Hanser, 2007). However, for non-textual data such as image, the query keywords do not appear in the content of the data to be retrieved. Further, the low-level features extracted from the content and the tags are from two different feature spaces. The match between query keywords and tags, and the relatedness of tags to the tagging objects, are hence expected to affect the tag-based retrieval experience.

Different motivations for tagging naturally lead to *tag type categorization* (Rorissa, 2010). Bischoff et al. (2008) proposed a tag categorization taxonomy with eight categories, including *topic*, *time*, *location*, and *author/owner*, that can be applied for different tagged resources including pictures, web pages, and music. For photos in Flickr, *topic* (e.g., people, flowers) and *location* tags are the most frequently used tag types followed by *time*, *type* (e.g., portrait, landscape), and *opinions/qualities*. Based on AOL query logs, the authors further observed that *topic* and *location* were the most searched tag types where *topic* accounted for about 50% of the searches. Although AOL query logs are for web search, similar observations are made in the *TagExplorer* system, a tool for faceted browsing of Flickr photos (Sigurbjörnsson & van Zwol, 2010). In *TagExplorer*, a slightly simplified taxonomy is adopted to classify tags/queries into three main categories, namely, "where" (locations), "when" (time and activities), and "what" (subjects and names). It is observed that for browsing and searching images, "what" is the most used type (53%), followed by "where" (28%) and "when" (19%). There are also works on classifying tags into tag types automatically and using tags for social data integration (Bischoff, Firan, Kadar, Nejdl, & Paiu, 2009; Ding et al., 2010; Overell et al., 2009).

Tag Relatedness and Representativeness

Categorization of tags and queries paves the way to more effective answering of queries of specific type(s). However, in general social image search, the quantification of *degree of relatedness* of a specific tag to a specific image remains poorly understood. (Li et al. 2008, 2009) proposed a neighbor-voting framework to quantify tag relatedness. It is based on the intuition that if different persons use the same tags to label visually similar images, then these tags are likely to reflect the visual contents of the annotated images. Given an image d, its k-nearest neighbors, denoted by $N_k(d)$, are first obtained based on visual features (e.g., color, edge, texture). The tag relatedness of a tag $t \in d$ is then computed by the probability of t used to annotate the neighborhood images $N_k(d)$ ($P(t|N_k(d))$) offset by its a priori probability of being used in the entire collection $P(t)$, i.e., $P(t|N_k(d)) - P(t)$. Their experiments on 20 queries each with 1,000 labeled examples showed that voting-based tag relatedness can significantly improve image search accuracy (Li et al., 2009). More recently, Li et al. (2010) compared tag relatedness methods using visual similarity defined by multiple types of visual features and concluded that a uniform combination of neighborhood images based on multiple visual features yield results comparable to those based on more complicated supervised or unsupervised combination methods. Their experiments used 20 and 33 queries, respectively, on two datasets. Using neighbor voting as the first step, Liu et al. (2009) applied random walk to further refine tag relatedness by taking pair-wise similarity between tags into consideration. Their main task, however, was to re-rank the tags of a tagged image such that the most relevant tags appear in top positions. Evaluated using 10 popular tags as queries, their re-ranking method achieved better image search results than using original tag positions. Tag relatedness is also related to *tag refinement* where user-assigned tags (e.g., tags for self-reference) may be removed and suggestions for tags that describe image content are provided (Liu, Hua, Wang, & Zhang, 2010; Zhu et al., 2010).

Because both tag relatedness and refinement are based on image visual similarity, they are applicable mainly to "content related" tags, but not tags of other types. However, identifying tags that describe the visual content of images itself is a non-trivial problem. Most existing efforts take a simplistic approach by filtering tags based on frequency or WordNet entries (Chua et al., 2009; Liu et al., 2010; Wu, Hua, Yu, Ma, & Li, 2008; Zhu et al., 2010). Sun and Bhowmick (2009) propose *visual-representativeness* to quantify the effectiveness of a tag in describing the common visual content of its annotated images. For instance, sunset and tiger are visual-representative tags because they describe the visual content of their annotated images effectively. In contrast, most time-related, location-related, or self-reference tags are not visually representative. In their recent work (Sun & Bhowmick, 2010), multiple measures were proposed and evaluated using various visual features of images.

The visual-representative tags can be considered as high-level concepts with small semantic gaps with respect to the image representation in visual space (Lu et al., 2008). To find these tags, Lu et al. derived a confidence score for each image based on the *coherence degree* of its nearest neighbors in both *visual* and *textual* spaces, assuming that each image is surrounded by textual descriptions (e.g., comments). The high-level concepts are then derived through clustering these images with high confidence scores. Similar approach was adopted by Tang et al. (2009).

Discussion

The knowledge of tag types facilitates the search engine's hints generation related to the types of images expected by the searcher based on the types of query tags. An image search system may therefore choose an appropriate ranking method and/or present the matched images in a way to address the search intention effectively. To the best of our knowledge, there is no existing study on detecting all tag types and adjusting the ranking or presentation methods accordingly.

Several recent research have focused on processing location-based tags ("where") and queries (Crandall, Backstrom, Huttenlocher, & Kleinberg, 2009; Kennedy & Naaman, 2008; Serdyukov, Murdock, & van Zwol, 2009). The temporal dimension of the tag ("when") has also been studied recently in the context of event detection in Flickr (Becker, Naaman, & Gravano, 2010; Chen & Roy, 2009). Very recently, Taneva, Kacimi, and Weikum (2010) proposed a weighting and ranking method for searching photos of specific named entities (scientist, politician, building, and mountain), which relates to a very small portion of topic/subject tags ("what"). Recall that topic/subject tags account for more than half of the searches in tag-based image searching and browsing. Hence, effective support for "what"-tag queries is crucial in TAGIR.

Studies on tag relatedness mainly focus on content-related tags with the assumption that relevant tags are assigned to visually similar images; tag visual-representativeness aims at quantifying the effectiveness of a tag on representing the common visual content among its annotated images. Both may therefore improve the search accuracy for content-related tags (Li et al., 2009, 2010; Liu et al., 2009; Zhu et al., 2010). However, the existing studies lack a framework as well as systematic evaluation of a variety of different aspects of TAGIR.

Framework

In this section, we present the TAGIR framework in detail. The notations used in this article are summarized in Table 1.

Figure 1 depicts the architecture of our TAGIR framework. A tagged image consists of two orthogonal components, namely, *visual content* and *a bag of tags*, associated with the image. For simplicity, we model a tagged image d by its tags only, i.e., $d = \langle t_1, t_2, \ldots, t_{|d|} \rangle$, where $|d|$ defines the number of tags[3] associated with d, which is also known as *tag length*. Note that the visual content is not explicitly modeled in our TagIR framework. In particular, the visual content of the images may be used to derive several measures and properties associated with the tags (e.g., tag relatedness, tag visual-representativeness), which can be pre-computed while indexing the images.

TABLE 1. Symbols and semantics.

Symbol	Semantic
$\mathcal{D}$	The tagged image collection
d	A tagged image, $d \in \mathcal{D}$
$\lvert d \rvert$	Number of tags associated with image d
$t \in d$	A tag t associated with image d
D_t	The set of images tagged by tag t
$f(t)$	Number of images tagged by tag t, $f(t) = \lvert D_t \rvert$
$P(t)$	A priori probability of observing t in $\mathcal{D}$, $P(t) = f(t)/\lvert \mathcal{D} \rvert$
$N_k(d)$	The k-nearest neighbors of d based on visual content
$t_q \in Q$	A query tag in query Q consisting of one or more tags

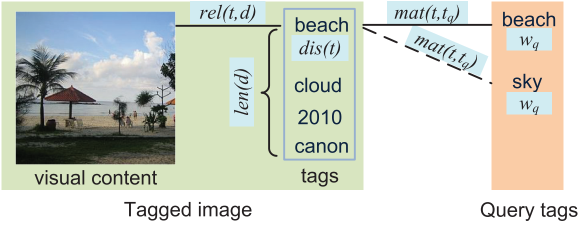

FIG. 1. The TagIR framework.

Inspired by the IR relevance space described in Zobel and Moffat (1998), we identify five dimensions for tag-based image search, namely, *tag relatedness*, *tag discrimination*, *tag length normalization*, *tag-query matching model*, and *query model*. With reference to the default relevance score implemented in Lucene,[4] a widely used Java package for text search, we propose to use Eq. (1) as the scoring function for an image d and a tag query Q, where $\text{rel}(t, d)$, $\text{dis}(t)$, $\text{len}(d)$, and $\text{mat}(t, t_q)$ denote the first four dimensions, respectively, and w_q denotes the weight assigned to a query tag $t_q \in Q$ (see Figure 1 for illustration). Here, a tag query consists of multiple query tags $Q = \langle t_1, t_2, \ldots, t_{|Q|} \rangle$ and their corresponding weights $w_q = \langle w_1, w_2, \ldots, w_{|Q|} \rangle$

$$\text{Score}(d, Q) = \sum_{t_q \in Q, t \in d} w_q \times \text{rel}(t, d) \times \text{dis}(t) \times \text{len}(d) \times \text{mat}(t, t_q) \tag{1}$$

Observe that query rewriting (e.g., query expansion) is not explicitly modeled in the scoring function because it is implicitly represented by the query model dimension in the framework. Also, the aforementioned scoring function has a similar flavor to the *inner product* function in Zobel and Maffat's work (1998) except that the former is computed based on tags associated with an image rather than terms in a textual document. Owing to its closeness to textual document search, the proposed scoring function can be easily implemented on top of an existing textual search system (e.g., Lucene).

Note that the proposed framework returns matched images with their relevance scores. Identifying representative images, diversifying search results, and their presentation are beyond the scope of the proposed framework; as these issues can be handled at the *post-processing phase* after the relevant images are retrieved (Crandall et al., 2009; Kennedy & Naaman, 2008; van Leuken, Garcia, Olivares, &

[3] In Flickr, no duplicate tags can be assigned to the same image. The total number of tags of an image is therefore the same as its number of distinct tags.

[4] http://lucene.apache.org/. The relevance score is known as similarity score in Lucene.

van Zwol, 2009). In the following subsections, we discuss the formulations in each dimension.

Tag Relatedness

Recall that in social image search, there is little common understanding on the issue of quantifying the degree of relatedness between tags and images. Therefore, we evaluate the following three tag relatedness implementations:

$$\mathrm{rel}(t, d) = \begin{cases} 1.0 & \text{Unit relatedness} \\ \dfrac{|d| - \mathrm{pos}(t, d)}{|d|} & \text{Tag position} \\ \alpha + (1-\alpha)\dfrac{\nu(t, d)}{\max\limits_{t' \in d} \nu(t', d)} & \text{Neighbor-voting} \end{cases} \tag{2}$$

Unit Relatedness. A simple approach is to consider all tags equally relevant to their tagged images, i.e., rel(t,d) = 1.0.

Tag Position. Another approach for formulating tag relatedness is to consider the first few tags assigned to an image as a better reflection of the creator/viewer's perception about the image compared with tags assigned later. Tag relatedness is then determined by the *tag position*, shown in Eq. (2), where pos(t, d) denotes the position of a tag t among all tags assigned to image d and $0 \leq \mathrm{pos}(t, d) < |d|$. Here, $|d|$ is the number of tags annotated to the image.

User/Neighbor Voting. Tags voted by more users to a particular resource are believed to be more relevant than tags chosen by very few user (e.g., Delicious). When such a user voting is not available (e.g., Flickr), a tag can be considered as more relevant to an image if the tag is used to annotate other visually similar images (Li et al., 2008; Makadia et al., 2008). Specifically, a *voting score* $\nu(t, d)$ is computed for each tag $t \in d$ using Eq. (3), where $N_k(d)$ is the k-nearest neighbors of d based on visual similarity; $P(t|N_k(d))$ and $P(t)$ are the probabilities of observing tag t among images in $N_k(d)$ and collection $\mathcal{D}$, respectively

$$\nu(t, d) = \max(P(t|N_k(d)) - P(t), 0) \tag{3}$$

The neighbor-voting implementation computes tag relatedness by exploiting visual similarity between images. Hence, content-based tags are likely to receive higher voting scores than tags of other types (e.g., time, location, opinion). In Eq. (2), a parameter $\alpha \in [0,1]$ is used to adjust the contribution of the normalized neighbor voting, which is set to 0.5 in our evaluation. Note that efficient search of k-nearest neighbors of a given image based on visual similarity is beyond the scope of this work.

Tag Discrimination

Analogous to the IDF component in TF×IDF weighting scheme, the tag discrimination dimension, denoted by dis(t), is used to quantify the discriminating power of a tag t with respect to the image collection $\mathcal{D}$. We evaluate the following three tag discrimination implementations:

$$\mathrm{dis}(t) = \begin{cases} 1.0 & \text{discrimination} \\ 1.0 + \log\dfrac{|\mathcal{D}|}{1.0 + f(t)} & \text{IDF discrimination} \\ \mathrm{visual}(t) & \text{Visual-representativeness} \end{cases} \tag{4}$$

Unit Discrimination. In this formulation, dis(t) = 1.0 for any tag appearing in the image collection.

Inverse Document Frequency (*IDF*). As IDF has played a pivotal role in various IR tasks, we also consider a variant of it for our evaluation. Here, a document refers to an image in the tagged image collection. Among many variants of IDF definitions, we evaluate the one defined in Lucene, shown in Eq. (4), where $f(t)$ is the number of images annotated by tag t.

Visual-Representativeness. Recall that about half of tag queries are "what" type, related to visual content of images. Hence, *visual-representativeness* is evaluated as an implementation of tag discrimination, which assigns greater weight to visually representative tags and smaller weight to other tags. A tag is visually representative if its annotated images are visually similar to each other, containing a common visual concept such as an object or a scene. The visual-representativeness is computed using a clarity-based measure (Sun & Bhowmick, 2009, 2010) where the visual content of images is represented using bag of *visual-words*. A *visual-word* is a codeword rather than a unit of language. The clarity measure reflects the difference in the visual-word distributions between the set of images assigned a given tag t against the entire image collection. Specifically, the *clarity* of a tag is the KL-divergence between the *tag language model* $P(w|D_t)$ and the *collection language model* $P(w|\mathcal{D})$ given below and the language models are estimated from visual-word distributions

$$\mathrm{Clarity}(t) = \sum_w P(w|D_t) \log_2 \frac{P(w|D_t)}{P(w|\mathcal{D})} \tag{5}$$

Let $\mu(t')$ and $\sigma(t')$ be the *expected tag clarity score* and standard deviation derived from dummy tag t' randomly assigned to the similar number of images as t (i.e., $f(t) = f(t')$). The visual-representativeness of tag t is given by the zero-mean normalization in Eq. (6). Owing to space constraints, the reader may refer to Sun and Bhowmick (2009, 2010) for the motivation behind zero-mean normalization and the estimations of $P(w|D_t)$ and $P(w|\mathcal{D})$

$$\mathrm{Visual}(t) = \frac{\mathrm{Clarity}(t) - \mu(t')}{\sigma(t')} \tag{6}$$

In our experiments, Visual(t) is further normalized using a sigmoid function such that the most and least visually representative tags have Visual(t) approaching 1.0 and 0, respectively. For a rarely used tag, whose number of annotated images is not large enough to compute visual-representativeness, its Visual(t) is set to 0.5.

Tag Length Normalization

Length normalization is used to reflect the impact of the number of tags assigned to social images. Note that tag length normalization is to normalize the number of tags an image has, but not the length of each individual tag. We evaluate two formulations as shown in Eq. (7) below. The unit normalization represents the baseline formulation where the number of tags is not considered in computing Score(d, Q). The square-root normalization favors images with fewer tags, which is also the default normalization formulation used in Lucene

$$\text{len}(d) = \begin{cases} 1.0 & \text{Unit normalization} \\ \dfrac{1}{\sqrt{|d|}} & \text{Square-root normalization} \end{cases} \tag{7}$$

Tag-Query Matching Model

Tag-query matching enables us to quantify the matching score between a tag $t \in d$ and the query tag t_q. We consider two types of matching detailed below

$$\text{mat}(t, t_q) = \begin{cases} 1.0 & \text{If } t = t_q \\ 0 & \text{Exact match if } t \neq t_q \\ \text{assoc}(t, t_q) & \text{Match-by-association if } t \neq t_q \end{cases} \tag{8}$$

Exact Match. If there is an exact match between t and t_q, then mat$(t, t_q) = 1.0$; otherwise, mat$(t, t_q) = 0$.

Match-by-Association. Given the limited number of tags assigned to most images, exact match is relatively restrictive. Hence, *match-by-association* is introduced to enhance flexibility of matching. In this case, mat$(t, t_q) = 1.0$ if t matches t_q *literally*. Otherwise, the *association score* between t and t_q is used to quantify the matching. For instance, in Figure 1, the dotted line between tags beach and sky indicates match-by-association. The associations between tags can be computed based on tag co-occurrence, *Google distance* (Cilibrasi & Vitanyi, 2007) or *Flickr distance* (Wu et al., 2008). In our experiments, we evaluate three associations, namely, Jaccard coefficient, co-occurrence probability, and interest measure, as defined in Equation (9). In this equation, $f(t \wedge t_q)$ denotes the number of images tagged by both t and t_q; $P(t|t_q)$ is the conditional probability of being tagged by t among the images tagged by t_q

$$\text{assoc}(t, t_q) = \begin{cases} \dfrac{f(t \wedge t_q)}{f(t) + f(t_q) - f(t \wedge t_q)} & \text{Jaccard} \\ P(t|t_q) & \text{Co-occurrence} \\ \max(P(t|t_q) - P(t), 0) & \text{Interest} \end{cases} \tag{9}$$

Observe that in match-by-association an image d may receive a good matching score with a query tag t_q even if t_q does not literally match any of d's tags. The computation of a matching score therefore requires one complete scan of all searchable images, which is not feasible for real-time search when the underlying image collection is large. To reduce the search response time and also benefit from the existing indexing techniques (e.g., inverted indexing), we limit the computation of the matching score to only those images that are tagged by at least one query tag $t_q \in Q$. Obviously, such constraint limits the recall of the search. One possible solution to address this issue is to perform tag query expansion, which we discuss next.

Query Model

The purpose of query model dimension is to rewrite a given query so as to achieve superior search experience. As reported in the recent Multimedia Grand Challenge, people search for images for various reasons and type on average 2.2 tags for each search (Yahoo!, 2010). In fact, the widely adopted tag-cloud depiction of tags makes the problem even more challenging because clicking any tag in a tag-cloud leads to a single-tag query. Tags appearing in a tag-cloud are often extremely popular tags, so each tag leads to a large number of matching images that need to be ranked. Therefore, we focus on query expansion techniques, which have been a long-standing research topic in IR (Xu & Croft, 1996). For clarity, in the following discussion on query expansion we assume that the given query is a single-tag query. Note that the formulations can be easily extended to handle other forms of queries if necessary.

Expansion-by-Association. In this formulation, a single-tag query is expanded by including its top-K most associated tags. In our experiments, we set $K = 5$. We evaluate expansions using the three association measures described in Equation (9). The weights of the expanded tags are their corresponding associations with the given query tag having weight of 1.0.

Concept-Based Expansion. The simple top-K expansion ignores the associations among the expanded tags. For example, the top-5 tags associated with tag rock by Jaccard are: cliff, rocks, concert, music, and band. Clearly, the first two tags are related to *rock stone*, whereas the last three are related to *rock music*. Figure 2 depicts the *tag relationship graph* (TRG) of the query tag rock. Each node in the TRG is labeled with a tag and the label of an edge is the association score of the connected tag pair (e.g., by Jaccard). Observe that the graph structure clearly depicts the two different concepts (*rock stone* and *rock music*) on the left- and right-hand sides of the rock node, respectively. We evaluate *concept-based tag expansion* by exploiting the *concepts* extracted from the TRGS.

The concept detection can be achieved with a community detection or graph cut algorithm. We adopted Modularity Clustering in our experiments (Newman, 2006). In the following, we briefly describe the construction of TRG, followed by concept detection using modularity clustering. A more detailed description is provided by Sun, Bhowmick, and Chong (2011). For a given tag t_q, the TRG is constructed based on an association measure in Equation (9). The top-K most associated tags with t_q are included as the *first-hop* tags in the graph ($K = 8$ in Figure 2). Then, from the top-K most associated tags of each first-hop tag, we select those

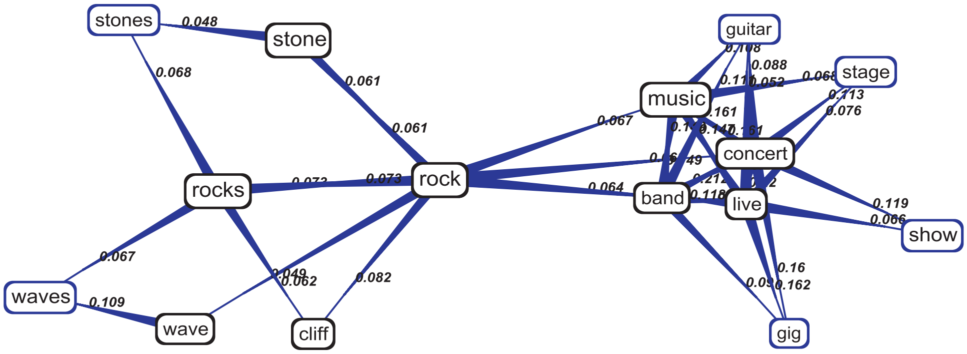

FIG. 2. Tag relationship graph for tag rock.

Rock → rock:1.0 wave:0.049 cliff:0.082 rocks:0.073 stone:0.061
Rock → rock:1.0 concert:0.068 band:0.064 music:0.067 live:0.048
Sunset → sunset:1.0 clouds:0.123 sun:0.148 sky:0.121 silhouette:0.091
Sunset → sunset:1.0 beach:0.088 sea:0.093 ocean:0.078 water:0.082

FIG. 3. Expanded queries for tags rock and sunset, respectively.

tags that are associated with at least two first-hop tags and add them in the TRG as *second-hop* tags. For example, in Figure 2 stones is a second-hop tag (depicted by blue-colored border in the TRG) as it is one of the top-8 most associated tags of stone as well as of rocks (stones is associated with two first-hop tags rocks and stone). We detect concepts by first removing the node representing the query tag t_q from the TRG and then applying modularity clustering to the resultant graph. The removal of t_q is to ensure that t_q does not affect the detection of concepts because it is related to all concepts, but tags in one concept may not be strongly associated with tags in another concept. For instance, removal of the node representing rock leads to two disconnected components in Figure 2 where each component naturally forms a concept. For each component, a concept-based expanded query is generated by adding to t_q the first-hop tags in the component. For example, the expanded queries for tags rock and sunset with $K = 8$ are given in Figure 3, where the values following ":" are Jaccard coefficients. Note that the two expanded queries for tag sunset show two different aspects of the sunset scenes, namely, *landscape* and *seascape*. In our experiments, we set $K = 10$ and on average each query tag leads to 3.1 concept-based expanded queries and each expanded query consists of 4.2 tags. However, it is unreasonable to make assumptions on the specific concept that best matches a user's search intent. Hence, in our evaluation we first compute the score of a matching image with each of the expanded queries of a given query tag and consider only the largest score as the matching score. This ensures that the number of expanded tags considered in the score computation using concept-based expansion is comparable to that of in expansion-by-association with $K = 5$.

Combinations of Formulations

We have identified five dimensions in TAGIR and discussed a few alternative formulations for each dimension. Nevertheless, the formulations discussed above are far from exhaustive and many other formulations can be adopted. For instance, more than eight document length normalization formulations are listed by Zobel and Moffat (1998), which can all be adopted for tag length normalization. However, realistically a complete evaluation of all possible combinations would take years even if each combination takes a few minutes to evaluate. Zobel and Moffat also highlight this practical imitation. Hence, we limit our experimental evaluation on the combinations of the aforementioned formulations. This leads us to *72 methods without query expansion*. A method here refers to a combination of different formulations under each dimension. With query expansion, we add an additional constraint that if the query is expanded based on an association measure in Equation (9) (e.g., Jaccard), then match-by-association must also be based on the same association measure. This leads to *216 methods with query expansion* for evaluation.

Experimental Setup

Ideally, the evaluation should be conducted on a large dataset with queries and their corresponding matching images manually labeled (i.e., ground truth), as are the TREC datasets widely used in IR. Unfortunately, to the best of our knowledge, there does not exist such a benchmark dataset for TAGIR. Hence, we chose NUS-WIDE dataset,[5] containing 269,648 images from Flickr, for our experimental study. Although the number of images seems not very large, this is the largest publicly available web image dataset with manual ground-truth labeling (Chua et al., 2009). In the next section, we give a brief overview of the dataset and justify its use in TAGIR evaluation.

Dataset

All tags provided in the dataset are used in our experiments without filtering. More than 90% of the images in NUS-WIDE dataset are socially tagged with 5–50 tags. The tags are mostly in English and some are in other languages. Each image in the dataset is manually assigned with zero, one, or more

[5] http://lms.comp.nus.edu.sg/research/NUS-WIDE.htm

DOI: 10.1002/asi

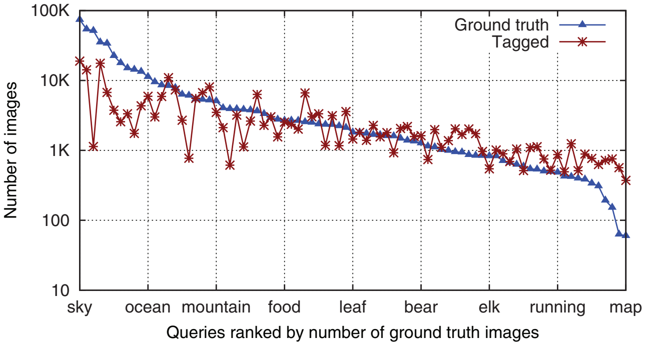

FIG. 4. Distribution of the ground-truth and tagged images.

concepts from a predefined set of 81 concepts. As Chua et al. (2009), stated, the 81 concepts were carefully selected such that they (i) correspond to some tags in Flickr; (ii) cover both general concepts like “animal” and specific ones like “dog” and “flowers”; and (iii) belong to different genres including scene (e.g., airport and beach), object (e.g., tiger, car), event (e.g., earthquake, wedding), program (e.g., sports), people (e.g., police, military), and graphics (e.g., MAP). Among them, scene and object are the two largest genres; each has 33 concepts.

Queries and Ground-Truth. In TAGIR, images are queried by their social tags and each query may consist of one or multiple tags. Naturally, each of the 81 concepts (which also correspond to tags) can be used as a *single-tag query* and the corresponding manually labeled images serve as the ground-truth for TAGIR evaluation. Note that all these 81 queries are “what” queries. The distribution of the number of ground-truth images and the number of images tagged by the concepts is shown in Figure 4 with a few example concepts labeled along the x-axis. Observe that the tag sky has more than 74K ground-truth images and map has 60. Their tagged images are 18.9 K and 372, respectively, representing the most and least popular tags.

Because the dataset does not provide ground-truth labeling for multi-tag queries, we derive *multi-tag queries* by combining the concepts and their ground-truth labelings. For instance, ⟨buildings, garden, sky⟩ form a 3-tag query. To ensure that a derived multi-tag query leads to sufficient number of ground-truth images that fully match all tags in the query, we do not consider those combinations with fewer than 200 images fully matching the query. Further, we do not evaluate a multi-tag query (e.g., ⟨buildings, sky⟩) if there is another more specific query (e.g., ⟨buildings, garden, sky⟩) to be evaluated where the latter contains all query tags from the former. Based on these two conditions, we derived 443 multi-tag queries ranging from 2-tag to 5-tag queries (Table 2).

Low-Level Visual Content Features. The dataset provides six types of low-level features to describe the visual contents of images, including global features such as color, edge, texture, and local feature known as bag of visual-words. We use the 500-D bag of visual-words to compute the visual-representativeness (see Tag Discrimination). Two

TABLE 2. Distribution of the derived multi-tag queries.

Query length	2-Tag	3-Tag	4-Tag	5-Tag	All queries
No. of queries	75	121	91	156	443

similarity definitions as suggested by Li et al. (2010) are used to compute the nearest neighbors for neighbor voting (see Tag Relatedness). Specifically (i) Euclidian distance on three types of global features (64-D color histogram, 73-D edge direction histogram, and 128-D wavelet texture features) are used to obtain 100 nearest neighbors; (ii) cosine similarity on 500-D bag of visual-words is used to obtain another 100 nearest neighbors where the images are processed very much like textual documents and TF×IDF weighting is adopted for 500-D visual-words. Both visual-representativeness and tag relatedness are pre-computed before indexing images by their tags.

Naming of Methods

We use the notations listed in Table 3 to identify uniquely the 288 methods (recall from Combinations of Formulations). For instance, $Q_S R_U D_U L_U M_E$ refers to the method using single-tag query (Q_S) without query expansion, unit relatedness (R_U), unit discrimination (D_U), unit length normalization (L_U), and exact matching (M_E). For clarity, when the same method is evaluated on multi-tag queries, we use $Q_M R_U D_U L_U M_E$ to denote the method. This method is also known as the *baseline* method in our experiments. It assigns the same matching score of 1.0 to any image matching a single-tag query. For multi-tag queries, the method assigns an image a matching score equal to the number of query tags matched by the image.

Evaluation Metrics

For a single-tag query, the ground-truth labels whether an image matches the query. In other words, for a given single-tag query, the ground-truth does not provide a list of images ranked according to their relevance to the query. In the absence of ground-truth on degree of relevance, we adopt MAP, Precision@K, and Recall, in our evaluation. For a multi-tag query, the degree of relevance of an image can be defined based on the number of matching query tags. For instance, an image matching only two query tags of a 3-tag query is considered less relevant than an image matching all the three query tags, but more relevant than an image matching any one of the three tags. We therefore adopt NDCG for evaluating multi-tag queries. In the following, we briefly describe the four metrics.

Mean Average Precision (MAP). For a given query, Average Precision (AP) is the average of the precision values obtained when each relevant image is retrieved for that query; if a relevant image is not retrieved at all, its corresponding precision

TABLE 3. Dimensions and formulation notations.

Dimension	Notations
Relatedness	R_U, R_P, and R_V for unit, position, and voting
Discrimination	D_U, D_F, and D_V for unit, IDF, and visual-rep
Length norm	L_U and L_S for unit and square-root normalization
Matching model	M_E, M_J, M_C, and M_T for exact match, match-by-Jaccard, co-occurrence, and inTerest
Query model	Q for query without expansion. For clarity, single-tag and multi-tag queries are further distinguished by Q_S and Q_M. E_X and C_X, respectively, denote expansion-by-association and concept-based expansion, where $X \in \{J, C, T\}$ for Jaccard, co-occurrence, and in Terest

is 0 (Manning et al., 2008). Hence, AP emphasizes that relevant images be ranked higher and is capped by the recall. MAP is the mean of APS for a set of queries.

Precision@K (or simply *P@K*). It is the ratio of the relevant images among the top-K retrieved images for a given query. In this article, $P@K$ refers to the macro-average of $P@K$ values for the evaluated queries. In our experiments, $K \in \{25, 50, 100, 200, 400\}$.

Recall for a query is the ratio of the retrieved relevant images among all relevant images. Similarly, it refers to the macro-average of recall values of the queries evaluated for each method.

Normalized Discounted Cumulative Gain@K (or NDCG@K) is given in the following equation, where r is the ranking position ($1 \leq r \leq K$); rel(r) is the degree of relevance of the image at rank position r, defined to be the number of matching tags against the given n-tag query ($0 \leq \text{rel}(r) \leq n$). Z is a normalization factor such that a perfect ranking (i.e., ground-truth) gives NDCG@$K = 1$. Similar to $P@K$, in our experiments, $K \in \{25, 50, 100, 200, 400\}$ for NDCG@K

$$\text{NDCG}@K = \frac{1}{Z}\sum_{r=1}^{K}\frac{2^{\text{rel}(r)} - 1}{\log_2(r + 1)} \qquad (10)$$

Among the four metrics, MAP, $P@K$, and NDCG@K evaluate the *ranking* of the retrieved images. However, the images retrieved by the search methods may be scored equally (e.g., two images annotated by the same set of tags). Hence, for each query, the computation of MAP, $P@K$, and NDCG@K involves shuffling and sorting of the retrieved images 50 times using standard Java API. The AP, $P@K$, and NDCG@K for each query is the average over the 50 evaluations in order to minimize the impact of images being equally scored. Note that recall is not affected because it does not rely on the ranking of the retrieved images.

Single-tag Query Evaluation

In this section, we begin by giving a performance overview of the 288 methods evaluated on 81 single-tag queries. Next, we list the best and worst performing methods based on MAP and $P@100$. Lastly, we analyze the impact of the five dimensions and their formulations to the retrieval performance.

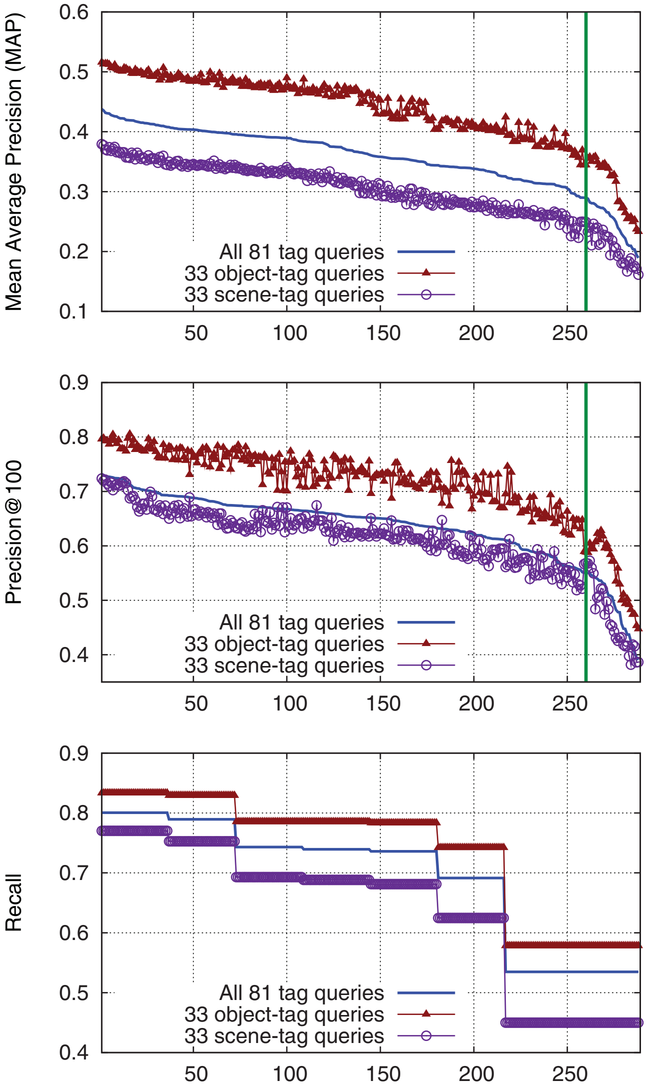

FIG. 5. Performance overview of 288 methods. Methods are ranked by the respective measures on 81 tag queries. The vertical bars in the top two figures indicate the position of the baseline method.

Performance Overview

MAP. Figure 5(a) plots the MAPS against ranks of the 288 methods. They are computed for all 81 single-tag queries, the 33 single-tag queries from object genre, and the 33 from scene genre. The vertical bar in the figure indicates the ranking position of the baseline method. Observe that the best

TABLE 4. Recall of single-tag queries for different query models.

Model	Q_S	E_J	E_C	E_T	C_J	C_C	C_T
Recall	0.535	0.691	0.743	0.739	0.736	0.801	0.790

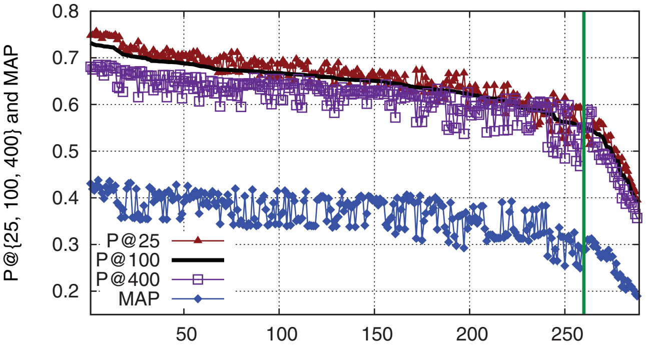

FIG. 6. Method ranked by $P@100$ on 81 queries.

performing method among all 81 queries achieve MAP of 0.438, which is more than 51% of increase over the baseline with MAP of 0.2898 (ranked at 260th position) and an increase of 131% against the worst performing method having MAP equal to 0.1895. It indicates that different methods may lead to significantly different search experiences.

Consider the MAPS of the two types of queries. It is clear that object queries enjoy higher MAPS compared with scene queries. The higher MAPS for object queries can be attributed to both higher precision and higher recall of such queries as depicted in Figures 5(b) and 5(c), respectively. Owing to page constraints, in Figure 5(b) we plot only $P@100$ values. However, the same observation holds for $P@\{25, 50, 200, 400\}$. One possible reason is that users typically tag an image with the objects observable from the image more often and accurately than the scenes in it. Furthermore, observe that if one method performs well on object queries, very likely it also performs well on scene queries, and vice versa. Specifically, Pearson's correlation coefficient for MAP values over all methods for object and scene queries is 0.982. Similar observation holds for $P@100$ with correlation coefficient 0.908. Hence, in the next section, we do not discuss the results of different types of queries separately.

Recall. Recall from Tag-Query Matching Model and Query Model, an image is scored only if it is tagged by at least one query tag. The recall of a method therefore depends solely on the query but not on the ranking model. For the same query, methods with different ranking models achieve the same recall. For instance, the last 72 methods in Figure 5(c) refer to the 72 methods without query expansion. They have the same recall value (0.535) averaged overall 81 single-tag queries. The figure also illustrates that query expansion can improve the average recall over the 81 queries from 0.535 to 0.801 (Table 4).

$P@100$. Figure 6 plots the methods ranked by $P@100$ on all 81 queries along with their MAP, $P@25$, and $P@400$ values. The vertical bar indicates the rank of baseline method. For clarity, we do not plot $P@50$ and $P@200$. Observe that all $P@K$ measures are highly correlated. Pearson's correlation coefficients between $P@100$ and $P@\{25, 50, 200, 400\}$ are 0.959, 0.972, 0.988, and 0.984, respectively, for values over all methods. Owing to this high correlation, we analyze the performances of the methods using $P@100$ only instead of all $P@K$ values. Interestingly, the correlation coefficient between $P@100$ and MAP is 0.85, indicating that some methods may achieve better MAPS but not necessarily better $P@100$s and vice versa. Recall that $P@100$ and MAP reflect the search accuracy of a method based on the top-100 images and all retrieved images, respectively. In the next section, we shall analyze the methods using both $P@100$ and MAP measures.

Best and Worst Performing Methods

72 Methods without Query Expansion. Table 5(a) lists the 10 best and worst performing methods among the 72 methods (without query expansion) ranked by MAP and $P@100$. It is not surprising that the baseline method, indicated by $\prod$ in the table, is among the worst performing methods for both measures. Notice that the best performing methods, $Q_S R_V D_F L_S M_C$ and $Q_S R_V D_U L_S M_C$, achieve an increase of 24% in MAP and 30% in $P@100$ against their corresponding worst performing methods. Interestingly, some methods even delivered poorer $P@100$ than the baseline method, although by a tiny margin and the results are not statistically significant. In summary, we can make the following observations. It is clear that R_V play a critical role in ranking because all the best performing methods use R_V and almost all the worst performing methods engage R_U. Second, M_J seems to be a better tag-query matching formulation than others for MAP. Third, L_S is the dominant length normalization choice for methods that achieve best $P@100$s, but not necessarily for the best MAPS. Lastly, there is no clear pattern related to the tag discrimination dimension.

All 288 Methods. Table 5(b) reports the 10 best and worst performing methods for all 288 methods. It is interesting to observe that for both MAP and $P@100$ the best and worst performing methods involve query expansion. That is, query expansion may not always lead to better search performance and it can perform poorer than the baseline method (ranked at 260th position). On the other hand, the best performing methods achieve 51 and 33% increase in their MAPS and $P@100$ measures, respectively, over the baseline method. In summary, R_V is the dominant tag relatedness formulation for better MAP and $P@100$. Further, most best performing methods based on MAP expanded the corresponding queries using C_J. Most best performing methods also use exact match M_E and all the worst performing methods use M_C coupled with E_C or C_C.

Interestingly, the last point suggests that two-hop matching by co-occurrence probability between a query tag and an

TABLE 5. The 10 best and worst performing methods among the 72 methods without query expansion and all 288 methods, where $\prod$ indicates the baseline method; $^+/^-$ indicates the values are significantly better/worse than baseline by paired *t*-test.

Rank	Method	MAP	Method	$P@100$	Rank	Method	MAP	Method	$P@100$
1	$Q_{\mathrm{S}}R_{\mathrm{V}}D_{\mathrm{F}}L_{\mathrm{S}}M_{\mathrm{C}}$	0.3588^+	$Q_{\mathrm{S}}R_{\mathrm{V}}D_{\mathrm{U}}L_{\mathrm{S}}M_{\mathrm{C}}$	0.7104^+	1	$C_{\mathrm{J}}R_{\mathrm{V}}D_{\mathrm{F}}L_{\mathrm{U}}M_{\mathrm{E}}$	0.4380^+	$C_{\mathrm{T}}R_{\mathrm{V}}D_{\mathrm{F}}L_{\mathrm{U}}M_{\mathrm{E}}$	0.7321^+
2	$Q_{\mathrm{S}}R_{\mathrm{V}}D_{\mathrm{F}}L_{\mathrm{S}}M_{\mathrm{T}}$	0.3579^+	$Q_{\mathrm{S}}R_{\mathrm{V}}D_{\mathrm{F}}L_{\mathrm{S}}M_{\mathrm{C}}$	0.7065^+	2	$C_{\mathrm{J}}R_{\mathrm{V}}D_{\mathrm{U}}L_{\mathrm{U}}M_{\mathrm{E}}$	0.4354^+	$C_{\mathrm{T}}R_{\mathrm{V}}D_{\mathrm{U}}L_{\mathrm{U}}M_{\mathrm{E}}$	0.7298^+
3	$Q_{\mathrm{S}}R_{\mathrm{V}}D_{\mathrm{U}}L_{\mathrm{S}}M_{\mathrm{C}}$	0.3575^+	$Q_{S}R_{\mathrm{V}}D_{\mathrm{F}}L_{\mathrm{S}}M_{\mathrm{T}}$	0.7031^+	3	$C_{\mathrm{J}}R_{\mathrm{V}}D_{\mathrm{F}}L_{S}M_{\mathrm{J}}$	0.4317^+	$E_{\mathrm{C}}R_{\mathrm{V}}D_{\mathrm{F}}L_{\mathrm{U}}M_{\mathrm{E}}$	0.7283^+
4	$Q_{\mathrm{S}}R_{\mathrm{V}}D_{\mathrm{U}}L_{S}M_{\mathrm{T}}$	0.3570^+	$Q_{\mathrm{S}}R_{\mathrm{V}}D_{\mathrm{U}}L_{S}M_{\mathrm{T}}$	0.7024^+	4	$C_{\mathrm{J}}R_{\mathrm{V}}D_{\mathrm{U}}L_{S}M_{\mathrm{J}}$	0.4308^+	$C_{\mathrm{C}}R_{\mathrm{V}}D_{\mathrm{F}}L_{\mathrm{U}}M_{\mathrm{E}}$	0.7269^+
5	$Q_{\mathrm{S}}R_{\mathrm{V}}D_{\mathrm{F}}L_{\mathrm{S}}M_{\mathrm{J}}$	0.3554^+	$Q_{\mathrm{S}}R_{\mathrm{V}}D_{\mathrm{V}}L_{\mathrm{S}}M_{\mathrm{C}}$	0.7009^+	5	$C_{\mathrm{T}}R_{\mathrm{V}}D_{\mathrm{F}}L_{\mathrm{U}}M_{\mathrm{E}}$	0.4305^+	$C_{\mathrm{J}}R_{\mathrm{V}}D_{\mathrm{F}}L_{\mathrm{U}}M_{\mathrm{E}}$	0.7269^+
6	$Q_{\mathrm{S}}R_{\mathrm{V}}D_{\mathrm{V}}L_{\mathrm{U}}M_{\mathrm{J}}$	0.3553^+	$Q_{\mathrm{S}}R_{\mathrm{V}}D_{\mathrm{V}}L_{\mathrm{S}}M_{\mathrm{T}}$	0.6998^+	6	$C_{\mathrm{J}}R_{\mathrm{V}}D_{\mathrm{F}}L_{S}M_{\mathrm{E}}$	0.4279^+	$E_{\mathrm{T}}R_{\mathrm{V}}D_{\mathrm{F}}L_{\mathrm{U}}M_{\mathrm{E}}$	0.7262^+
7	$Q_{\mathrm{S}}R_{\mathrm{V}}D_{\mathrm{F}}L_{\mathrm{U}}M_{\mathrm{J}}$	0.3552^+	$Q_{\mathrm{S}}R_{\mathrm{V}}D_{\mathrm{F}}L_{\mathrm{S}}M_{\mathrm{J}}$	0.6996^+	7	$C_{\mathrm{C}}R_{\mathrm{V}}D_{\mathrm{F}}L_{\mathrm{U}}M_{\mathrm{E}}$	0.4268^+	$C_{\mathrm{C}}R_{\mathrm{V}}D_{\mathrm{U}}L_{\mathrm{U}}M_{\mathrm{E}}$	0.7257^+
8	$Q_{\mathrm{S}}R_{\mathrm{V}}D_{\mathrm{V}}L_{\mathrm{S}}M_{\mathrm{T}}$	0.3551^+	$Q_{\mathrm{S}}R_{\mathrm{V}}D_{\mathrm{U}}L_{\mathrm{S}}M_{\mathrm{J}}$	0.6967^+	8	$C_{\mathrm{J}}R_{\mathrm{V}}D_{\mathrm{V}}L_{\mathrm{U}}M_{\mathrm{E}}$	0.4258^+	$E_{\mathrm{C}}R_{\mathrm{V}}D_{\mathrm{U}}L_{\mathrm{U}}M_{\mathrm{E}}$	0.7245^+
9	$Q_{\mathrm{S}}R_{\mathrm{V}}D_{\mathrm{U}}L_{\mathrm{U}}M_{\mathrm{J}}$	0.3551^+	$Q_{\mathrm{S}}R_{\mathrm{V}}D_{\mathrm{V}}L_{\mathrm{S}}M_{\mathrm{J}}$	0.6922^+	9	$C_{\mathrm{J}}R_{\mathrm{V}}D_{\mathrm{U}}L_{S}M_{\mathrm{E}}$	0.4254^+	$E_{\mathrm{J}}R_{\mathrm{V}}D_{\mathrm{F}}L_{\mathrm{U}}M_{\mathrm{E}}$	0.7242^+
10	$Q_{\mathrm{S}}R_{\mathrm{V}}D_{\mathrm{U}}L_{\mathrm{S}}M_{\mathrm{J}}$	0.3549^+	$Q_{\mathrm{S}}R_{\mathrm{V}}D_{\mathrm{V}}L_{\mathrm{U}}M_{\mathrm{J}}$	0.6890^+	10	$E_{\mathrm{J}}R_{\mathrm{V}}D_{\mathrm{F}}L_{\mathrm{U}}M_{\mathrm{E}}$	0.4235^+	$E_{\mathrm{T}}R_{\mathrm{V}}D_{\mathrm{U}}L_{\mathrm{U}}M_{\mathrm{E}}$	0.7240^+
63	$Q_{S}R_{P}D_{\mathrm{U}}L_{\mathrm{U}}M_{\mathrm{C}}$	0.3132^+	$Q_{\mathrm{S}}R_{\mathrm{U}}D_{\mathrm{U}}L_{\mathrm{U}}M_{\mathrm{T}}$	0.5635	279	$E_{\mathrm{C}}R_{\mathrm{V}}D_{\mathrm{U}}L_{\mathrm{U}}M_{\mathrm{C}}$	0.2295^-	$C_{\mathrm{C}}R_{\mathrm{P}}D_{\mathrm{U}}L_{\mathrm{U}}M_{\mathrm{C}}$	0.4568^-
64	$Q_{\mathrm{S}}R_{\mathrm{U}}D_{\mathrm{F}}L_{\mathrm{S}}M_{\mathrm{E}}$	0.3128^+	$Q_{\mathrm{S}}R_{\mathrm{P}}D_{\mathrm{F}}L_{\mathrm{U}}M_{\mathrm{C}}$	0.5630	280	$E_{\mathrm{C}}R_{P}D_{\mathrm{U}}L_{\mathrm{U}}M_{\mathrm{C}}$	0.2255^-	$C_{\mathrm{C}}R_{\mathrm{V}}D_{\mathrm{V}}L_{\mathrm{U}}M_{\mathrm{C}}$	0.4480^-
65	$Q_{\mathrm{S}}R_{\mathrm{U}}D_{\mathrm{U}}L_{\mathrm{S}}M_{\mathrm{E}}$	0.3128^+	$Q_{\mathrm{S}}R_{\mathrm{P}}D_{\mathrm{U}}L_{\mathrm{U}}M_{\mathrm{C}}$	0.5625	281	$C_{\mathrm{C}}R_{\mathrm{U}}D_{\mathrm{U}}L_{\mathrm{U}}M_{\mathrm{C}}$	0.2217^-	$E_{\mathrm{C}}R_{\mathrm{P}}D_{\mathrm{U}}L_{\mathrm{U}}M_{\mathrm{C}}$	0.4474^-
66	$Q_{\mathrm{S}}R_{\mathrm{U}}D_{\mathrm{V}}L_{\mathrm{S}}M_{\mathrm{E}}$	0.3128^+	$Q_{\mathrm{S}}R_{\mathrm{U}}D_{\mathrm{F}}L_{\mathrm{U}}M_{\mathrm{E}}$	0.5610	282	$E_{\mathrm{C}}R_{\mathrm{U}}D_{\mathrm{U}}L_{\mathrm{U}}M_{\mathrm{C}}$	0.2120^-	$E_{\mathrm{C}}R_{\mathrm{V}}D_{\mathrm{V}}L_{\mathrm{U}}M_{\mathrm{C}}$	0.4372^-
67	$Q_{\mathrm{S}}R_{\mathrm{U}}D_{\mathrm{F}}L_{\mathrm{U}}M_{\mathrm{C}}$	0.3126^+	$Q_{S}R_{\mathrm{U}}D_{\mathrm{V}}L_{\mathrm{U}}M_{\mathrm{E}}$	0.5506	283	$C_{\mathrm{C}}R_{\mathrm{V}}D_{\mathrm{V}}L_{\mathrm{U}}M_{\mathrm{C}}$	0.2097^-	$C_{\mathrm{C}}R_{\mathrm{U}}D_{\mathrm{U}}L_{\mathrm{U}}M_{\mathrm{C}}$	0.4371^-
68	$Q_{\mathrm{S}}R_{\mathrm{U}}D_{\mathrm{V}}L_{\mathrm{U}}M_{\mathrm{C}}$	0.3103^+	$Q_{\mathrm{S}}R_{\mathrm{U}}D_{\mathrm{U}}L_{\mathrm{U}}M_{\mathrm{E}}\prod$	0.5503	284	$C_{\mathrm{C}}R_{P}D_{\mathrm{V}}L_{\mathrm{U}}M_{\mathrm{C}}$	0.2096^-	$E_{\mathrm{C}}R_{\mathrm{U}}D_{\mathrm{U}}L_{\mathrm{U}}M_{\mathrm{C}}$	0.4245^-
69	$Q_{\mathrm{S}}R_{\mathrm{U}}D_{\mathrm{U}}L_{\mathrm{U}}M_{\mathrm{C}}$	0.3100^+	$Q_{\mathrm{S}}R_{\mathrm{U}}D_{\mathrm{F}}L_{\mathrm{U}}M_{\mathrm{E}}$	0.5502	285	$E_{\mathrm{C}}R_{\mathrm{V}}D_{\mathrm{V}}L_{\mathrm{U}}M_{\mathrm{C}}$	0.2040^-	$C_{\mathrm{C}}R_{\mathrm{P}}D_{\mathrm{V}}L_{\mathrm{U}}M_{\mathrm{C}}$	0.4127^-
70	$Q_{\mathrm{S}}R_{\mathrm{U}}D_{\mathrm{V}}L_{\mathrm{U}}M_{\mathrm{E}}$	0.2900	$Q_{\mathrm{S}}R_{\mathrm{U}}D_{\mathrm{V}}L_{\mathrm{U}}M_{\mathrm{C}}$	0.5492	286	$M_{\mathrm{C}}R_{P}D_{\mathrm{U}}L_{\mathrm{U}}M_{\mathrm{C}}$	0.2033^-	$E_{\mathrm{C}}R_{\mathrm{P}}D_{\mathrm{V}}L_{\mathrm{U}}M_{\mathrm{C}}$	0.4049^-
71	$Q_{\mathrm{S}}R_{\mathrm{U}}D_{\mathrm{F}}L_{\mathrm{U}}M_{\mathrm{E}}$	0.2898	$Q_{\mathrm{S}}R_{\mathrm{U}}D_{\mathrm{U}}L_{\mathrm{U}}M_{\mathrm{C}}$	0.5456	287	$C_{\mathrm{C}}R_{\mathrm{U}}D_{\mathrm{V}}L_{\mathrm{U}}M_{\mathrm{C}}$	0.1941^-	$C_{\mathrm{C}}R_{\mathrm{U}}D_{\mathrm{V}}L_{\mathrm{U}}M_{\mathrm{C}}$	0.3945^-
72	$Q_{\mathrm{S}}R_{\mathrm{U}}D_{\mathrm{U}}L_{\mathrm{U}}M_{\mathrm{E}}\prod$	0.2898	$Q_{\mathrm{S}}R_{\mathrm{U}}D_{\mathrm{F}}L_{\mathrm{U}}M_{\mathrm{C}}$	0.5452	288	$E_{\mathrm{C}}R_{\mathrm{U}}D_{\mathrm{V}}L_{\mathrm{U}}M_{\mathrm{C}}$	0.1895^-	$E_{\mathrm{C}}R_{\mathrm{U}}D_{\mathrm{V}}L_{\mathrm{U}}M_{\mathrm{C}}$	0.3910^-

image may significantly hurt the searching results. Suppose t_a is expanded from the original query tag t_q by co-occurrence probability; suppose an image has a tag t_b which is associated with t_a by co-occurrence probability (but may not be associated with t_q). Then the associations between t_q and t_a and between t_a and t_b are both counted in the scoring. That is, it counts the two-hop tag association between t_q and t_b. Consequently, this results in poor performance.

TABLE 6. Distribution of *Good*/*Average*/*Poor* methods.

Method Group	Without query expansion		With query expansion	
	MAP	$P@100$	MAP	$P@100$
Good	12	10	13	19
Average	46	49	167	166
Poor	14	13	36	31

Impact of Dimensions

We now present a set of experiments that analyzes the impact of the dimensions in detail. We partition the methods into three groups, namely, *Good*, *Average*, and *Poor*, and then quantify the discriminative power of each dimension. For instance, consider the 72 methods without query expansion. Let μ_m and σ_m be the mean and standard deviation of the MAP values (resp. $P@100$) of these methods, respectively. The methods whose MAP (resp. $P@100$) values are larger than $\mu_m + \sigma_m$ and smaller than $\mu_m - \sigma_m$ are grouped under *Good* and *Poor*, respectively. The remaining methods are classified as *Average*. Table 6 reports the number of *Good*, *Average*, and *Poor* methods among the 72 methods without query expansion and the 216 methods with query expansion.

We now test the discriminative power of each dimension using the existing feature selection techniques (Sebastiani, 2002). We adopt Information Gain in this work and use Weka[6] to perform the analysis. The results for the 72 methods without query expansion are reported in the left upper part of Table 7(a). Observe that the two most discriminative dimensions based on MAP are tag relatedness and matching model having information gain of 0.41 and 0.31, respectively. In contrast, tag relatedness and length normalization are the most discriminative dimensions for $P@100$ measure. To have a better understanding of which formulations under these dimensions lead to *Good*/*Average*/*Poor* methods, rules are learned from these methods using the ripper algorithm (Cohen, 1995) (Weka's JRip implementation).

The lower part of Table 7(a) highlights the learned rules. In total, 6 (resp. 4) rules are learned from the grouping of 72 methods without query expansion based on MAP (resp. $P@100$). The first rule[7] based on MAP states that: *IF a method uses neighbor voting* R_{V} *for tag relatedness AND square-root* L_{S} *for length normalization, THEN the method achieves good* MAP. The numbers in the parenthesis indicate the number of methods covered by the rule and the number of false positives (e.g., three methods covered by the first rule in fact belong to *Average* group and not the *Good* group). The second rule implies that the neighbor-voting tag relatedness (R_{V}) and Jaccard-based match-by-association (M_{J}) also lead to good MAP. Note that methods that are not covered by the first five rules are addressed by the last rule. Similarly, R_{V} and L_{S} also

[6] http://www.cs.waikato.ac.nz/ml/weka/

[7] We modified the rule presentation format slightly from the Weka output for concise presentation.

TABLE 7. Discriminative power of the dimensions, and rules.

	MAP		P@100	
	Dimension	InfoGain	Dimension	InfoGain
(a) Methods without query expansion				
1	Relatedness	0.4056	Relatedness	0.3849
2	Matching model	0.3073	Length norm	0.2594
3	Length norm	0.0810	Matching model	0.1054
4	Discrimination	0.0019	Discrimination	0.0089
1	$R_{\mathrm{V}} \wedge L_{\mathrm{S}} \Rightarrow \mathit{Good}(12/3)$		$R_{\mathrm{V}} \wedge L_{\mathrm{S}} \Rightarrow \mathit{Good}(12/3)$	
2	$R_{\mathrm{V}} \wedge M_{\mathrm{J}} \Rightarrow \mathit{Good}(3/0)$		$R_{\mathrm{U}} \wedge L_{\mathrm{U}} \Rightarrow \mathit{Poor}(12/2)$	
3	$M_{\mathrm{E}} \wedge R_{\mathrm{U}} \Rightarrow \mathit{Poor}(6/0)$		$M_{\mathrm{C}} \wedge R_{\mathrm{P}} \wedge L_{\mathrm{U}} \Rightarrow \mathit{Poor}(3/0)$	
4	$L_{\mathrm{U}} \wedge M_{\mathrm{C}} \wedge R_{\mathrm{U}} \Rightarrow \mathit{Poor}(3/0)$		$\Rightarrow \mathit{Average}(45/1)$	
5	$M_{\mathrm{E}} \wedge R_{\mathrm{P}} \wedge L_{\mathrm{U}} \Rightarrow \mathit{Poor}(3/0)$			
6	$\Rightarrow \mathit{Average}(45/2)$			
(b) Methods with query expansion				
1	Matching model	0.2042	Matching model	0.1939
2	Query model	0.1861	Length norm	0.1555
3	Relatedness	0.1030	Relatedness	0.1531
4	Length norm	0.0426	Query model	0.1120
5	Discrimination	0.0426	Discrimination	0.0102
1	$R_{\mathrm{V}} \wedge C_{\mathrm{J}} \Rightarrow \mathit{Good}(12/5)$		$\mathrm{R_V} \wedge \mathrm{L_U} \wedge \mathrm{M_E} \Rightarrow \mathit{Good}(18/1)$	
2	$R_{\mathrm{V}} \wedge D_{\mathrm{F}} \wedge M_{\mathrm{E}} \wedge C_I \Rightarrow \mathit{Good}(2/0)$		$\mathrm{L_U} \wedge \mathrm{M_A} \wedge \mathrm{C_C} \Rightarrow \mathit{Good}(9/0)$	
3	$M_{\mathrm{A}} \wedge C_{\mathrm{C}} \wedge L_{\mathrm{U}} \Rightarrow \mathit{Poor}(9/0)$		$\mathrm{E_C} \wedge \mathrm{M_A} \wedge \mathrm{L_U} \Rightarrow \mathit{Poor}(9/0)$	
4	$M_{\mathrm{A}} \wedge E_{\mathrm{C}} \Rightarrow \mathit{Poor}(18/4)$		$\mathrm{L_U} \wedge \mathrm{M_A} \wedge \mathrm{E_I} \Rightarrow \mathit{Poor}(9/3)$	
5	$M_{\mathrm{A}} \wedge C_{\mathrm{C}} \wedge D_{\mathrm{V}} \Rightarrow \mathit{Poor}(3/0)$		$\Rightarrow \mathit{Average}(171/9)$	
6	$M_{\mathrm{A}} \wedge L_{\mathrm{U}} \wedge D_{\mathrm{V}} \wedge E_I \Rightarrow (3/0)$			
7	$M_{\mathrm{A}} \wedge C_I \wedge L_{\mathrm{U}} \wedge D_{\mathrm{V}} \Rightarrow \mathit{Poor}(3/0)$			
8	$\Rightarrow \mathit{Average}(166/8)$			

lead to good P@100. In summary, neighbor-voting-based tag relatedness is one of the key factors to achieve better image search accuracy, whereas tag discrimination seems to have very little impact on the search results.

Table 7(b) reports the discriminative power analysis and the rules learned from the groupings of 216 methods with query expansion. Observe that tag-query matching model and query model are the two most influential dimensions for both MAP and P@100. Recall from Combinations of Formulations, we impose the constraint that if the query is expanded based on an association measure (e.g., Jaccard), then match-by-association must also be based on the same association measure. Hence, we describe the methods with two alternative matching models, M_{E} for exact match and M_{A} for match-by-association. We use six query models, E_{X} and C_{X} for expansion-by-association and concept-based expansion, respectively, where $X \in \{J, C, T\}$ denotes the three association measures. Our results demonstrate that $R_{\mathrm{V}} \wedge C_{\mathrm{J}}$ leads to good MAP regardless of the formulations of other dimensions. For P@100, $R_{\mathrm{V}} \wedge L_{\mathrm{U}} \wedge M_{\mathrm{E}}$ ensures a good precision. However, methods involving match-by-association M_{A} (i.e., rules 2, 3, and 4) all perform poorly.

Impact of Formulations

Tag Relatedness. Consider the 72 methods without query expansion. These methods are placed into three groups such that the 24 methods in one group use the same tag relatedness formulation, i.e., R_{U}, R_{P}, or R_{V}. Table 8 (first row and first column) reports the averaged MAPS and P@100 values for these three groups denoted by R_{U}, R_{P}, and R_{V}, respectively. Clearly, the 24 methods using R_{V} achieve the best MAP and P@100 (highlighted in bold). We also conduct statistical significance tests on the three tag relatedness formulations by considering each method as a sample and the tag relatedness as the variable. The significance test results are summarized in Table 9. The paired t-test shows that $R_{\mathrm{V}} \gg R_{\mathrm{P}} \gg R_{\mathrm{U}}$ based on either MAP or P@100 measure ($\gg$ denotes *significantly better than* with p-value <0.01). The test results are consistent with our earlier observations that all the best performing methods use R_{V} for tag relatedness, whereas most worst performing methods use R_{U}. Similar results are also obtained for the 216 methods with query expansion where each group consists of 72 methods.

Tag Discrimination and Length Normalization. Similarly, the impact of tag discrimination formulations (i.e., D_{U}, D_{F}, and D_{V}) is summarized in Tables 8 and 9. In Table 9, $\approx$ denotes comparable or statistically not significant. D_{F} and D_{V} achieve better MAP and P@100, respectively, for methods without query expansion. On the other hand, D_{F} is a clear winner with relatively large margins of improvement on both MAP and P@100 for methods with query expansion. Lastly, L_{S} is clearly a better choice than L_{U} for tag length normalization.

Tag-Query Matching Model. For methods without query expansion, the four tag-query matching formulations (i.e., exact match M_{E} and match-by-associations M_{J}, M_{C}, and M_{T}), were each used in 18 methods. M_{J} and M_{E} are the best and worst matching models, respectively, for both MAP

TABLE 8. Averaged MAP and P@100 for different formulations.

Method Dimension		Without query expansion MAP	 P@100		With query expansion MAP	 P@100
Relatedness	R_U	0.3218	0.6040	R_U	0.3572	0.6150
	R_P	0.3284	0.6258	R_P	0.3608	0.6284
	R_V	**0.3480**	**0.6737**	R_V	**0.3729**	**0.6606**
Discrimination	D_U	0.3324	0.6322	D_U	0.3650	0.6377
	D_F	**0.3337**	0.6327	D_F	**0.3844**	**0.6513**
	D_V	0.3321	**0.6387**	D_V	0.3415	0.6150
Length norm	L_U	0.3253	0.6062	L_U	0.3528	0.6163
	L_S	**0.3402**	**0.6628**	L_S	**0.3745**	**0.6530**
Matching model	M_E	0.3202	0.6163	M_E	**0.3932**	**0.6688**
	M_J	**0.3407**	**0.6471**	M_A	0.3341	0.6005
	M_C	0.3333	0.6312	E_J	0.3979	**0.6627**
	M_T	0.3368	0.6435	E_C	0.3230	0.6016
Query model	–	–	–	E_T	0.3572	0.6400
				C_J	**0.4072**	0.6595
				C_C	0.3324	0.6054
				C_T	0.3641	0.6388

TABLE 9. Paired t-test for single-tag query evaluation; ">>", ">", and "≈"' indicate p-value <0.01, <0.05, and >0.05, respectively.

Method Dimension	Method without query expansion MAP	 P@100	Method with query expansion MAP	 P@100
Relatedness	$R_V >> R_P >> R_U$	$R_V >> R_P >> R_U$	$R_V >> R_P >> R_U$	$R_V >> R_P >> R_U$
Discrimination	$D_F >> D_V \approx D_U$	$D_V >> D_F \approx D_U$	$D_F >> D_U >> D_V$	$D_F >> D_U >> D_V$
Length norm	$L_S >> L_U$	$L_S >> L_U$	$L_S >> L_U$	$L_S >> L_U$
Matching model	$M_J >> M_T >>$ $M_C >> M_E$	$M_J >> M_T >>$ $M_C >> M_E$	$M_E >> M_A$	$M_E >> M_A$
Query model	–	–	$C_J >> E_J >> E_T >> E_T >> C_C >> E_C$	$E_J >> C_J >> E_T >> C_T >> C_C >> E_C$

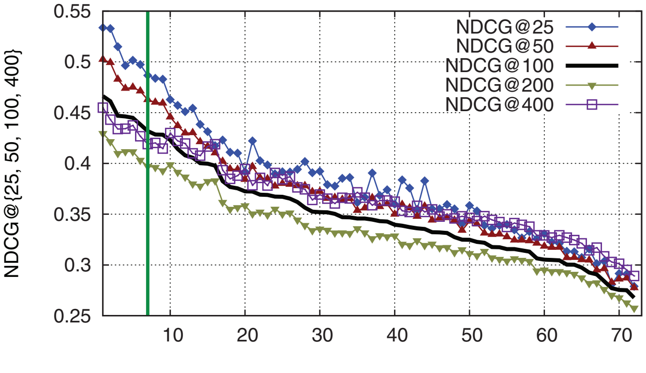

FIG. 7. Methods ranked by NDCG@100 on all queries.

and P@100 (">" denotes *significantly better than* with p-value <0.05). M_T is also a good matching model, which is comparable to M_J for P@100. Methods with M_E outperform methods with M_A by a large margin for both MAP and P@100 for methods with query expansion. This agrees with our earlier observation that two-hop association matching hurts search accuracy.

Query Model. The 216 methods with query expansion are first partitioned into two equal groups representing methods for expansion-by-association (E_X) and concept-based expansion (C_X). The paired t-test shows that: $C_X \gg E_X$ for MAP and $C_X \approx E_X$ for P@100. Next, we partition the 216 methods into six groups based on the association measures in E_X and C_X. As shown in Tables 8 and 9, C_J and E_J are the winners for MAP and P@100, respectively. In general, the Jaccard coefficient is a better association measure than interest and co-occurrence probability.

Multi-tag Query Evaluation

In this section, we report the evaluation using the 443 multi-tag queries. Note that because more than 55% of the 443 queries each contains 4 or 5 tags (Table 2); therefore, we do not further expand these queries. In the next section, we first give a performance overview of these 72 methods without query expansion and then analyze the impact of the dimensions and their formulations to the retrieval performance.

Performance Overview

Figure 7 plots the NDCG@{25, 50, 100, 200, 400} for all the 72 methods evaluated, ranked by NDCG@100. Similar to our earlier findings, all these values are highly correlated. Hence, we conduct our following analysis based on NDCG@100.

DOI: 10.1002/asi

TABLE 10. The best/worst performing methods by NDCG@100, where $\prod$ indicates the baseline method; $^{+}/^{-}$ indicates that the values are significantly better/worse than baseline by paired t-test.

Rank	Method	N@100	Rank	Method	N@100
1	$Q_M R_V D_F L_U M_E$	0.4663^{+}	63	$Q_M R_P D_V L_S M_E$	0.3004^{-}
2	$Q_M R_V D_U L_U M_E$	0.4613^{+}	64	$Q_M R_V D_U L_S M_E$	0.3002^{-}
3	$Q_M R_V D_V L_U M_E$	0.4469^{+}	65	$Q_M R_V D_U L_U M_C$	0.2974^{-}
4	$Q_M R_U D_F L_U M_E$	0.4462^{+}	66	$Q_M R_P D_V L_U M_C$	0.2925^{-}
5	$Q_M R_P D_F L_U M_E$	0.4449^{+}	67	$Q_M R_V D_V L_S M_E$	0.2907^{-}
6	$Q_M R_P D_U L_U M_E$	0.4384^{+}	68	$Q_M R_P D_F L_U M_C$	0.2840^{-}
7	$Q_M R_U D_U L_U M_E \prod$	0.4318	69	$Q_M R_U D_V L_U M_C$	0.2772^{-}
8	$Q_M R_P D_V L_U M_E$	0.4284	70	$Q_M R_P D_U L_U M_C$	0.2754^{-}
9	$Q_M R_U D_V L_U M_E$	0.4283^{-}	71	$Q_M R_U D_F L_U M_C$	0.2751^{-}
10	$Q_M R_U D_F L_S M_J$	0.4230^{-}	72	$Q_M R_U D_U L_U M_C$	0.2677^{-}

TABLE 11. Discriminative power of the dimensions, and rules.

	Dimension	InfoGain	Rules
1	Matching model	0.3542	$M_C \wedge L_U \Rightarrow$ *Poor*(9/2)
2	Length norm	0.0529	$M_E \wedge L_U \Rightarrow$ *Good*(9/0)
3	Discrimination	0.0335	$D_F \wedge R_U \wedge L_S \Rightarrow$ *Good*(4/1)
4	Relatedness	0.0154	$\Rightarrow$ *Average*

TABLE 12. Averaged NDCG@100 for different formulations.

Dimension	Query:	2-Tag	3-Tag	4-Tag	5-Tag	All
Relatedness	R_U	0.3642	**0.3514**	**0.3474**	0.3607	**0.3560**
	R_P	0.3613	0.3477	0.3406	**0.3620**	0.3536
	R_V	**0.3664**	0.3388	0.3329	0.3548	0.3479
Discrimination	D_U	0.3662	0.3456	0.3361	0.3518	0.3493
	D_F	**0.3755**	**0.3654**	**0.3573**	**0.3700**	**0.3671**
	D_V	0.3502	0.3268	0.3276	0.3557	0.3411
Length norm	L_U	0.3526	0.3438	**0.3452**	0.3526	0.3487
	L_S	**0.3754**	**0.3481**	0.3355	**0.3657**	**0.3563**
Matching model	M_E	**0.3906**	**0.3759**	**0.3730**	**0.3930**	**0.3838**
	M_J	0.3683	0.3451	0.3325	0.3535	0.3494
	M_C	0.3356	0.3152	0.3151	0.3363	0.3261
	M_T	0.3614	0.3476	0.3406	0.3538	0.3507

The best and worst 10 performing methods by NDCG@100 for multi-tag queries are listed in Table 10. Observe that the best performing method outperforms the worst by 74%. Among the best performing methods, there is no clear pattern observed from dimensions of tag relatedness and tag discrimination; however, most of these methods use unit length normalization and exact match (i.e., $L_U M_E$), including the baseline method $Q_M R_U D_U L_U M_E$ (ranked at the seventh position). Most worst performing methods use unit normalization and match-by-association with co-occurrence (i.e., $L_U M_C$). In the following, we conduct more detailed analysis on the impact of the dimensions and formulations.

Impact of Dimensions and Formulations

We partition the methods into 13 *Good*, 49 *Average*, and 10 *Poor* methods (see Impact of Dimensions). Table 11 reports the discriminative powers of the dimensions and the rules learned. Clearly, matching model is the most discriminative dimension followed by length normalization. Exact match and unit normalization ($M_E L_U$) lead to 9 out of the 13 good methods. This is consistent with our observations made from Table 10. It also partially explains the seventh ranking of the baseline method.

As in Impact of Formulations, the average NDCG@100 of the methods using the same formulation is reported in Table 12 with the best values highlighted in bold. To understand the methods better, we compute the averages according to the query length and their corresponding statistical significant test (Table 13).

- R_V is a better choice for 2-tag queries and R_U is better for most longer queries for tag relatedness. However, as shown in Table 13, the choices of tag relatedness formulations do not lead to significant differences in performance.
- D_F performs significantly better than either D_U or D_V statistically when tag discrimination is considered. However, the absolute amount of improvement is marginal (Table 12).

TABLE 13. Paired t-test for multi-tag query evaluation; ">>", ">", and "≈" ' indicate p-value <0.01, <0.05, and >0.05, respectively.

Dimension	2-Tag queries	3-Tag queries	4-Tag queries	5-Tag queries	All multi-tag queries
Relatedness	$R_V \approx R_U \approx R_P$	$R_U \approx R_P > R_V$	$R_U > R_P \approx R_V$	$R_U \approx R_P \approx R_V$	$R_U \approx R_P \approx R_V$
Discrimination	$D_F >> D_U >> D_V$	$D_F >> D_U >> D_V$	$D_F >> D_U >> D_V$	$D_F >> D_V \approx D_U$	$D_F >> D_U \approx D_V$
Length norm	$L_S > L_U$	$L_S \approx L_U$	$L_U \approx L_S$	$L_S \approx L_U$	$L_S \approx L_U$
Matching model	$M_E \approx M_J \approx M_T >> M_C$	$M_E \approx M_T \approx M_J >> M_C$	$M_E > M_T > M_J >> M_C$	$M_E > M_T \approx M_J >> M_C$	$M_E > M_T \approx M_J >> M_C$

- Unit length L_U is comparable with L_S except on short queries (i.e., 2-tag), where $L_S > L_U$ is observed.
- Exact match outperforms all other formulations by a large margin for longer queries with 4 or more tags. In fact, considering all multi-tag queries, the improvement of exact match remains statistically significant as shown in the last column in Table 13.

Discussion

Interestingly, compared with the findings from the evaluation of single-tag queries, we observe very different or even orthogonal findings for multi-tag queries as summarized below.

Tag Relatedness. Tag relatedness is the most discriminative dimension for single-tag queries without query expansion and neighbor-voting-based tag relatedness plays a key role in achieving better image search accuracy. However, when evaluated on multi-tag queries for the same 72 methods, tag relatedness became the least discriminative dimension and neighbor voting is comparable to other formulations such as unit relatedness.

Tag-Query Matching. This is the second most discriminative dimension (by MAP) for the 72 methods without query expansion for single-tag queries. Particularly, the results show that match-by-association is superior to the exact match for these methods. Matching model is the most discriminative dimension when evaluated on multi-tag queries (for the same set of methods). However, interestingly exact match significantly outperforms its counterpart. Recall that when evaluating single-tag queries using the 216 methods with query expansion, either match-by-association or query expansion may significantly improve the search accuracy but putting them together adversely affects the result. That is, exact match is a better choice when a query consists of multiple tags, regardless of whether the query is formulated as a multi-tag query or expanded from a single-tag query.

Tag Discrimination. The widely adopted IDF weighting often outperforms other formulations for tag discrimination for both single-tag and multi-tag queries. However, the improvement is marginal and overall tag discrimination seems to have relatively very little impact on the search results for both single-tag and multi-tag queries.

Length Normalization. Methods using square-root normalization achieve best $P@100$s on single-tag queries. However, tag length normalization is not a very discriminative dimension when we consider all evaluations with both single-tag and multi-tag queries.

One possible reason for such different observations made on single-tag and multi-tag queries is the expressive power of the queries. A single-tag query is usually much more general having a large number of potential matching images. Hence, tag relatedness and match-by-association both improve the estimation of the matching score between a tagged image and the query. A multi-tag query, on the other hand, expresses a much more specific information need. The presence of all query tags in a tagged image matching the query largely guarantees a superior match.

Conclusions

In this article, we propose a framework for a systematic empirical evaluation of various methods to rank matched images based on their associated tags in the context of tag-based social image retrieval. Our framework consists of five orthogonal dimensions that play pivotal roles in social image tagging, namely, tag relatedness, tag discrimination, tag length normalization, tag-query matching model, and query model. For each dimension, we discuss several formulation strategies to compute them. A major focus of this work has been on evaluating an array of methods representing various combinations of the proposed dimensions and formulations on NUS-WIDE dataset, the largest human-annotated dataset consisting of more than 269 K images from Flickr.

Our experimental evaluation revealed several interesting findings that play pivotal roles in designing superior social image tagging systems. We observed that single-tag queries and multi-tag queries are best handled separately using *different* relevance measures. Specifically, tag relatedness is the most important dimension for producing superior quality results for single-tag queries but less important for multi-tag queries. Similarly, relatively more advanced tag-query matching models are imperative for answering single-tag queries but not multi-tag queries. Because a large portion of queries are single-tag queries, tag relatedness and tag-query matching model remain important research areas. Our study also showed that neighbor voting is the most effective implementation for tag relatedness. However, it is also a computationally expensive method because of the neighborhood search based on low-level content features. This calls for a more effective and efficient alternative for tag relatedness computation. It is worth mentioning that the computation cost can be reduced if a social tagging system adopts a bag aggregation model and stores the number of times a tag is assigned to an image by multiple users. Our experimental results also demonstrate that either query expansion or match-by-association improves search accuracy for single-tag queries; but the two together result in poorer performance. The issue of which one is better in terms of effectiveness, efficiency, and scalability in large social image retrieval systems remains an open research problem that needs to be explored.

Acknowledgment

This work was supported by Singapore MOE AcRF Tier-1 Grant RG13/10.

References

Ames, M., & Naaman, M. (2007). Why we tag: Motivations for annotation in mobile and online media. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 971–980). New York: ACM Press.

Becker, H., Naaman, M., & Gravano, L. (2010). Learning similarity metrics for event identification in social media. In Proceedings of the Third ACM International Conference on Web Search and Data Mining (WSDM '10) (pp. 291–300). New York: ACM Press.

Berendt, B., & Hanser, C. (2007, March). Tags are not metadata, but "just more content"-to some people. Paper presented at the International Conference on Weblogs and Social Media (ICWSM '07), Boulder, Colorado. Retrieved from http://www.icwsm.org/papers/paper12.html

Bischoff, K., Firan, C.S., Kadar, C., Nejdl, W., & Paiu, R. (2009). Automatically identifying tag types. In Proceedings of the Fifth International Conference on Advanced Data Mining and Applications (ADMA '09) (pp. 31–42). Berlin: Springer.

Bischoff, K., Firan, C.S., Nejdl, W., & Paiu, R. (2008). Can all tags be used for search? In Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM '08) (pp. 193–202). New York: ACM Press.

Carneiro, G., Chan, A.B., Moreno, P.J., & Vasconcelos, N. (2007). Supervised learning of semantic classes for image annotation and retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence, 29, 394–410.

Chen, L., & Roy, A. (2009). Event detection from flickr data through wavelet-based spatial analysis. In Proceedinga of the 18th ACM Conference on Information and Knowledge Management (CIKM '09) (pp. 523–532). New York: ACM Press.

Chua, T.-S., Tang, J., Hong, R., Li, H., Luo, Z., & Zheng, Y. (2009). Nus-wide: A real-world web image database from national university of singapore. In Proceedinga of the ACM International Conference on Image and Video Retrieval (CIVR '09) (pp. 48:1–48:9). New York: ACM Press.

Cilibrasi, R.L., & Vitanyi, P.M.B. (2007). The google similarity distance. IEEE Transactions on Knowledge and Data Engineering, 19, 370–383.

Cohen, W.W. (1995). Fast effective rule induction. In Proceedings of the 12th International Conference on Machine Learning. San Francisco: Morgan Kaufmann.

Crandall, D.J., Backstrom, L., Huttenlocher, D., & Kleinberg, J. (2009). Mapping the world's photos. In Proceedings of the 18th International Conference on World Wide Web (WWW '09) (pp. 761–770). New York: ACM Press.

Datta, R., Joshi, D., Li, J., & Wang, J.Z. (2008). Image retrieval: Ideas, influences, and trends of the new age. ACM Computing Surveys, 40(2), 5:1–5:60.

Ding, Y., Jacob, E.K., Fried, M., Toma, I., Yan, E., Foo, S., & Milojevic, S. (2010). Upper tag ontology for integrating social tagging data. Journal of the American Society for Information Science and Technology, 61(3), 505–521.

Feng, S.L., Manmatha, R., & Lavrenko, V. (2004). Multiple bernoulli relevance models for image and video annotation. In Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision And Pattern Recognition (CVPR '04) (pp. 1002–1009). Washington, DC: IEEE Computer Society.

Goh, D.H.-L., Ang, R.P.-H., Lee, C.S., & Chua, A.Y.-K. (2011). Fight or unite: Investigating game genres for image tagging. Journal of the American Society for Information Science and Technology, 62(7), 1311–1324.

Golder, S.A., & Huberman, B.A. (2006). Usage patterns of collaborative tagging systems. Journal of Information Science, 32, 198–208.

Guillaumin, M., Mensink, T., Verbeek, J., & Schmid, C. (2009). Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In Proceedings of IEEE 12th International Conference on Computer Vision (ICCV '09). Washington, DC: IEEE Press.

Jain, R., & Sinha, P. (2010). Content without context is meaningless. In Proceedings of the International Conference on Multimedia (MM '10) (pp. 1259–1268). New York: ACM Press.

Jeon, J., Lavrenko, V., & Manmatha, R. (2003). Automatic image annotation and retrieval using cross-media relevance models. In Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval (SIGIR '03) (pp. 119–126). New York: ACM Press.

Kennedy, L.S., & Naaman, M. (2008). Generating diverse and representative image search results for landmarks. In Proceedings of the 17th International Conference on World Wide Web (WWW '08) (pp. 297–306). New York: ACM Press.

van Leuken, R.H., Garcia, L., Olivares, X., & van Zwol, R. (2009). Visual diversification of image search results. In Proceedings of the 18th International Conference on World Wide Web (WWW '09) (pp. 341–350). New York: ACM Press.

Li, X., Snoek, C.G.M., & Worring, M. (2008). Learning tag relevance by neighbor voting for social image retrieval. In Proceedings of the First ACM International Conference on Multimedia Information Retrieval (MIR '08) (pp. 180–187). New York: ACM Press.

Li, X., Snoek, C.G.M., & Worring, M. (2009). Learning social tag relevance by neighbor voting. IEEE Transactions on Multimedia , 11(7), 1310–1322.

Li, X., Snoek, C.G.M., & Worring, M. (2010). Unsupervised multi-feature tag relevance learning for social image retrieval. In Proceedings of the ACM International Conference on Image and Video Retrieval (CIVR '10) (pp. 10–17). New York: ACM Press.

Liu, D., Hua, X.-S., Wang, M., & Zhang, H.-J. (2010). Image retagging. In Proceedings of the International Conference on Multimedia (MM '10) (pp. 491–500). New York: ACM Press.

Liu, D., Hua, X.-S., Yang, L., Wang, M., & Zhang, H.-J. (2009). Tag ranking. In Proceedings of the 18th International Conference on World Wide Web (WWW '09) (pp. 351–360). New York: ACM Press.

Lu, Y., Zhang, L., Tian, Q., & Ma, W.-Y. (2008). What are the high-level concepts with small semantic gaps? In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR '08) (pp. 1–8). Washington, DC: IEEE Press.

Makadia, A., Pavlovic, V., & Kumar, S. (2008). A new baseline for image annotation. In Proceedings of European Conference on Computer Vision (ECCV). Berlin: Springer.

Manning, C.D., Raghavan, P., & Schtze, H. (2008). Introduction to information retrieval. Camridge, UK: Cambridge University Press.

Marlow, C., Naaman, M., Boyd, D., & Davis, M. (2006). Ht06, tagging paper, taxonomy, flickr, academic article, to read. In Proceedings of the 17th Conference on Hypertext and Hypermedia (HYPERTEXT '06) (pp. 31–40). New York: ACM Press.

Newman, M.E.J. (2006). Modularity and community structure in networks. Proceedings of the National Academy of Sciences of the United States of America, 103, 8577–8582.

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

Rorissa, A. (2010). A comparative study of flickr tags and index terms in a general image collection. Journal of the American Society for Information Science and Technology, 61(11), 2230–2242.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Information Processing & Management, 24(5), 513–523.

Sebastiani, F. (2002). Machine learning in automated text categorization. ACM Computing Surveys, 34, 1–47.

Serdyukov, P., Murdock, V., & van Zwol, R. (2009). Placing flickr photos on a map. In Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '09) (pp. 484–491). New York: ACM Press.

Sigurbjörnsson, B., & van Zwol, R. (2008). Flickr tag recommendation based on collective knowledge. In Proceedings of the 17th International Conference on World Wide Web (WWW '08) (pp. 327–336). New York: ACM Press.

Sigurbjörnsson, B., & van Zwol, R. (2010). TagExplorer: Faceted browsing of flickr photos (YL-2010-005). Yahoo! Research.

Stvilia, B., & Jörgensen, C. (2010). Member activities and quality of tags in a collection of historical photographs in flickr. Journal of the American Society for Information Science and Technology, 61(12), 2477–2489.

Sun, A., & Bhowmick, S.S. (2009). Image tag clarity: In search of visual-representative tags for social images. In Proceedings of the First SIGMM Workshop on Social Media (WSM '09) (pp. 19–26). New York: ACM Press.

Sun, A., & Bhowmick, S.S. (2010). Quantifying tag representativeness of visual content of social images. In Proceedings of the International Conference on Multimedia (MM '10) (pp. 471–480). New York: ACM Press.

Sun, A., Bhowmick, S.S., & Chong, J.-A. (2011). Social image tag recommendation by concept matching. In Proceedings of the ACM International Conference on Multimedia (MM '11). New York: ACM Press.

Taneva, B., Kacimi, M., & Weikum, G. (2010). Gathering and ranking photos of named entities with high precision, high recall, and diversity. In Proceedings of the Third ACM International Conference on Web Search and Data Mining (WSDM '10) (pp. 431–440). New York: ACM Press.

Tang, J., Yan, S., Hong, R., Qi, G.-J., & Chua, T.-S. (2009). Inferring semantic concepts from community-contributed images and noisy tags. In Proceedings of the 17th ACM International Conference on Multimedia (MM '09) (pp. 223–232). New York: ACM Press.

Weinberger, K.Q., Slaney, M., & van Zwol, R. (2008). Resolving tag ambiguity. In Proceedings of the 16th ACM International Conference on Multimedia (MM '08) (pp. 111–120). New York: ACM Press.

Wu, L., Hua, X.-S., Yu, N., Ma, W.-Y., & Li, S. (2008). Flickr distance. In Proceeding of the 16th ACM International Conference on Multimedia (MM '08) (pp. 31–40). New York: ACM Press.

Wu, L., Yang, L., Yu, N., & Hua, X.-S. (2009). Learning to tag. In Proceedings of the 18th International Conference on World Wide Web (WWW '09) (pp. 361–370). New York: ACM Press.

Xu, J., & Croft, W.B. (1996). Query expansion using local and global document analysis. In Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '96) (pp. 4–11). New York: ACM Press.

Yahoo! (2010). Grand challenge for ACM Multimedia. Retrieved from http://comminfo.rutgers.edu/conferences/mmchallenge/2010/02/10/yahoo-challenge-image/

Yanai, K., Shirahatti, N.V., Gabbur, P., & Barnard, K. (2005). Evaluation strategies for image understanding and retrieval. In Proceedings of the Seventh ACM SIGMM International Workshop on Multimedia Information Retrieval (MIR '05) (pp. 217–226). New York: ACM Press.

Zhu, G., Yan, S., & Ma, Y. (2010). Image tag refinement towards low-rank, content-tag prior and error sparsity. In Proceedings of the International Conference on Multimedia (MM '10) (pp. 461–470). New York: ACM Press.

Zobel, J., & Moffat, A. (1998). Exploring the similarity space. SIGIR Forum, 32(1), 18–34.

Zollers, A. (2007). Emerging motivations for tagging: Expression, performance, and activism. In Proceedings of Tagging and Metadata for Social Information Organization Workshop with www. New York: ACM Press.