

# Scalable search-based image annotation

Changhu Wang · Feng Jing · Lei Zhang ·  
Hong-Jiang Zhang

Received: 18 October 2007 / Accepted: 15 May 2008 / Published online: 14 June 2008  
© Springer-Verlag 2008

**Abstract** With the popularity of digital cameras, more and more people have accumulated considerable digital images on their personal devices. As a result, there are increasing needs to effectively search these personal images. Automatic image annotation may serve the goal, for the annotated keywords could facilitate the search processes. Although many image annotation methods have been proposed in recent years, their effectiveness on arbitrary personal images is constrained by their limited scalability, i.e. limited lexicon of small-scale training set. To be scalable, we propose a search-based image annotation algorithm that is analogous to information retrieval. First, content-based image retrieval technology is used to retrieve a set of visually similar images from a large-scale Web image set. Second, a text-based keyword search technique is used to obtain a ranked list of can-

didate annotations for each retrieved image. Third, a fusion algorithm is used to combine the ranked lists into a final candidate annotation list. Finally, the candidate annotations are re-ranked using Random Walk with Restarts and only the top ones are reserved as the final annotations. The application of both efficient search techniques and Web-scale image set guarantees the scalability of the proposed algorithm. Moreover, we provide an annotation rejection scheme to point out the images that our annotation system cannot handle well. Experimental results on U. Washington dataset show not only the effectiveness and efficiency of the proposed algorithm but also the advantage of image retrieval using annotation results over that using visual features.

---

Communicated by E. Chang.

C. Wang (✉)  
Department of Electronic Engineering and Information Science,  
University of Science and Technology of China,  
230027 Hefei, China  
e-mail: wch@ustc.edu

F. Jing  
Beijing New Sunlight Technologies Co. Ltd, Room 1104,  
D Building of the place No. 9 of Guanghua Rd,  
Chaoyang District, 100020 Beijing, China  
e-mail: scenery.jf@gmail.com

L. Zhang  
Microsoft Research Asia, 49 Zhichun Road,  
100190 Beijing, China  
e-mail: leizhang@microsoft.com

H.-J. Zhang  
Microsoft Advanced Technology Center,  
49 Zhichun Road, 100190 Beijing, China  
e-mail: hjzhang@microsoft.com

## 1 Introduction

With the popularity of digital cameras, more and more people have considerable digital images on their personal devices. How to effectively index and search these personal images emerges as a crucial issue. Unlike Web images that have rich metadata such as filename, ALT text, URL and surrounding text for indexing and searching, personal images have little textual information. A possible solution is to index and search images with visual information, i.e. content-based image retrieval (CBIR) [26]. Although CBIR has been extensively studied for more than a decade, it has three limitations that limit its practicability. First, due to the so-called semantic gap between low level visual features and high level semantic concepts, the precision of CBIR is usually unsatisfactory. Second, due to the high dimensionality of visual features and the curse of dimensionality, the efficiency and scalability of CBIR are usually low. Finally, the query form of CBIR is unnatural for personal image search. On the one hand, for query by example (QBE), the example image is often absent.

On the other hand, query by sketch (QBS) is too complex and adds too much burden to normal users.

To resolve the aforementioned issues, a straightforward solution is to manually annotate the images and search them with annotated keywords. However, manually annotating large quantity of images is too tedious and time-consuming for common users. To solve this problem, many automatic image annotation methods have been proposed in recent years. Most of the existing image annotation approaches can be classified into two categories, i.e. classification-based methods and probabilistic modeling-based methods.

The classification-based methods use image classifiers to represent annotation keywords (concepts). The classifier of a keyword (concept) is trained to separate the training images with the keyword (concept) from other keywords (concepts). For a new image, the outputs of the classifiers will be used to annotate it. Many representative classifiers have been used, such as the two-dimensional multi-resolution hidden Markov models (2D MHMMs) [19], support vector machine (SVM) [9, 14, 33], Bayes Point Machine [8], and Mixture Hierarchical Model [6, 7]. Since each annotation keyword (concept) has a classifier and all the classifiers should be tested to annotate an image, the classification-based methods are unsuitable for image dataset with unlimited lexicon, e.g. personal image sets.

The probabilistic modeling-based methods attempt to infer the correlations or joint probabilities between images and annotation keywords. As the pioneer work, Mori et al. [23] proposed a method for annotating image grids using co-occurrences in 1999. Although the model is simple, the annotation performance is relatively low. Another way of capturing co-occurrence information is to introduce latent variables that link image features with keywords. Duygulu et al. [10] proposed a novel approach that treated image annotation as a machine translation problem. A statistic machine translation model was used to “translate” the keywords of an image to the blob tokens obtained by clustering. The use of EM algorithm in the model constrains its scalability to large image collections. Other representative work includes the Gaussian Mixture Model, the Latent Dirichlet Allocation Model (LDA) and the correspondence LDA [4]. In spite of the profound statistic model and strict deduction, the parametric probabilities used in the model are too simple to model real distributions and the parameter estimation process is too complex to be scalable. Inspired by the relevance language models, several relevance models have been proposed recently, such as, the Cross-Media Relevance Model (CMRM) [15], the Continuous Relevance Model (CRM) [16, 18], and the Multiple Bernoulli Relevance Model (MBRM) [12]. Although these models are shown to be more effective than the previous ones, their dependency on small-scale high quality training set restricts their effectiveness on arbitrary personal images. That is, only images that are consistent with

the training images could be annotated with keywords in a limited vocabulary.

With the prosperity of the Web, it has become a huge deposit of almost all kinds of data. Several problems that were believed to be “unsolvable” have been successfully solved by leveraging the rich information of the Web [11, 34]. Compared with the limited number of concepts that can be modeled using a relatively small-scale image training set, the potentially unlimited vocabulary of a Web-scale image database can be utilized to annotate images. Motivated by Web search technologies in many commercial systems, we have proposed the AnnoSearch system [31]. Assuming that an accurate keyword of the image to be annotated is available, the keyword is used to retrieve several semantically relevant images. Then, the resulting images are further re-ranked based on visual features. Finally, a search result clustering (SRC) technique [35] is used to mine the final annotations from the top images. Although the initial keyword might speed up the search process and enhance the relevance of the retrieved images with the image to be annotated, it is not always available, especially for the personal images. Moreover, since SRC was originally designed for general Web search, it may not be an optimal solution to annotation mining.

In this paper, we focus on annotation of large personal image collections. A scalable search-based image annotation (SBIA) algorithm is proposed, which is analogous to information retrieval. The algorithm basically has four steps. First, CBIR technique is used to retrieve a set of visually similar images from a large-scale Web image dataset. Second, a text-based keyword search technique is used to obtain a ranked list of candidate annotations for each retrieved image. Third, a fusion algorithm is used to combine the ranked lists into a final candidate annotation list. Finally, the candidate annotations are re-ranked using Random Walk with Restarts (RWR) and only the top ones are reserved as the final annotations. The application of both efficient search technologies and Web-scale image set guarantees the scalability of the proposed algorithm. Similar to [31], the proposed approach enables annotating with almost unlimited vocabulary, which is impossible for most existing methods. To be capable of annotating personal images with little textual information, we remove the assumption of an initial keyword and employ an efficient indexing technique to speed up the search. Furthermore, instead of mining annotations with SRC, we consider this process as a text-based keyword search problem. Instead of treating the retrieved images equally as in [31], the images are treated differently according to their visual similarities to the query image. Moreover, we also provide an annotation rejection scheme to point out the images that our annotation system cannot handle well.

The scalability of the proposed framework is reflected in the following aspects. First, the system leverages a Web-scale training set (millions of training images), which could

not be efficiently handled by traditional learning based annotation methods. Second, the Web-scale training set provides an unlimited vocabulary, which is very different from the small scale lexicon used by traditional methods. Third, by leveraging the Web-scale image and keyword set, the system has the potential to annotate arbitrary query image. Fourth, the system could annotate images in real time. Based on the proposed framework, an online image annotation service has been deployed. A desktop image search system could benefit from such a service. When indexing desktop images, the system can submit images (or features) one by one to the annotation service and receive automated annotation results in the background. All the annotation processes are conducted on the server side and do not add any burden to the desktop system. Then the annotated images are indexed according to their annotations, and users can easily search images on desktop by keywords.

The rest of the paper is organized as follows. Section 2 presents our motivation and the “search” view of image annotation problem. The proposed image annotation algorithm is described in Sect. 3. Experimental results are shown in Sect. 4. We conclude the paper and discuss future work in Sect. 5.

The main results in this paper were first presented in [30].

## 2 Image annotation in the “search” view

### 2.1 Motivation

Large scale personal image collections have several special properties, which make image management and annotation even challenging. First, unlike Web images, personal images usually have little text information and therefore little training data is available to learn numerous image concepts. Second, the visual contents of personal images in one collection are too diverse, making it difficult to model various concepts. Third, since there are usually a large number of images in one’s collection, to efficiently manage these images, all of them need to be automatically annotated first. The first two properties make traditional annotation methods not appropriate in this special environment, and motivate us to seek for other solutions to annotate this kind of images. The third property necessitates the efficiency of the algorithm.

Due to the lack of training data in personal image collections, we need to leverage training data outside personal collections. And to be capable of annotating diverse personal images, the training set has to be diverse enough, which means that a large scale training set is crucial. Therefore, leveraging large scale Web images may be a good solution, and they can be considered as a low quality and diverse training set.

Motivated by text-based search technologies in many commercial systems, we formulate image annotation as a search

problem. That is, given a query image, how to search for accurate annotation keywords from a Web-scale training set. From this perspective, we further find that search-based image annotation and information retrieval can be treated as dual problems, and therefore we can systematically leverage those mature information retrieval technologies. In the following section we will discuss the correspondences between the two problems.

## 2.2 Correspondences

### 2.2.1 Information retrieval

A typical information retrieval (IR) system can be simplified as Fig. 1. Given a text query “*Bayesian introduction*”, the system first locates the inverted lists for each query term, and then finds relevant documents that contain all the query terms. If the user is not satisfied with the results, he/she may change the query, e.g. “*Bayesian tutorial*”, and search again to get more results. Or, he/she can modify his/her query to “*Bayesian~introduction*”<sup>1</sup> to find more results in one search session. However, in most cases, latter searches are ignored, because search results for the original query, e.g. “*Bayesian introduction*”, are usually sufficient.

Let it be assumed that the retrieval system always performs query synonym expansion as a general case. The remaining problem is how to rank the found relevant documents.

Let  $Dist(t)$  denote the distance between the original query term and the synonym term  $t$ , which can be measured based on word similarity as in WordNet [22]. For each query term  $t$  (or synonym term) in a document  $doc\_a$ , a term frequency  $\times$  inverse document frequency (tf-idf) [3] score  $Score_{tfidf}(doc\_a, t)$  is used to measure the relevance of the document to the term. In a simple case, the relevance score of the document can be obtained by summing up all terms’ tf-idf scores related to that document, weighted by a function of  $Dist(t)$ . The relevance score of document  $doc\_a$  can be calculated as follows:

$$Score_{relevance}(doc\_a) = \sum_{t \in S} f(Dist(t)) \times Score_{tfidf}(doc\_a, t), \quad (1)$$

where  $S$  is the set of query terms and synonym terms.

Although we could rank the documents according to their relevance scores and return top ones as the search results, the search results could be further refined using pseudo-relevance feedback technology [32]. The basic idea of pseudo-relevance feedback is to extract expansion terms from the top-ranked documents to formulate a new query. Through

<sup>1</sup> ~ is the synonym operator used by Google. “~introduction” is similar to “introduction OR tutorial OR FAQ ...”.

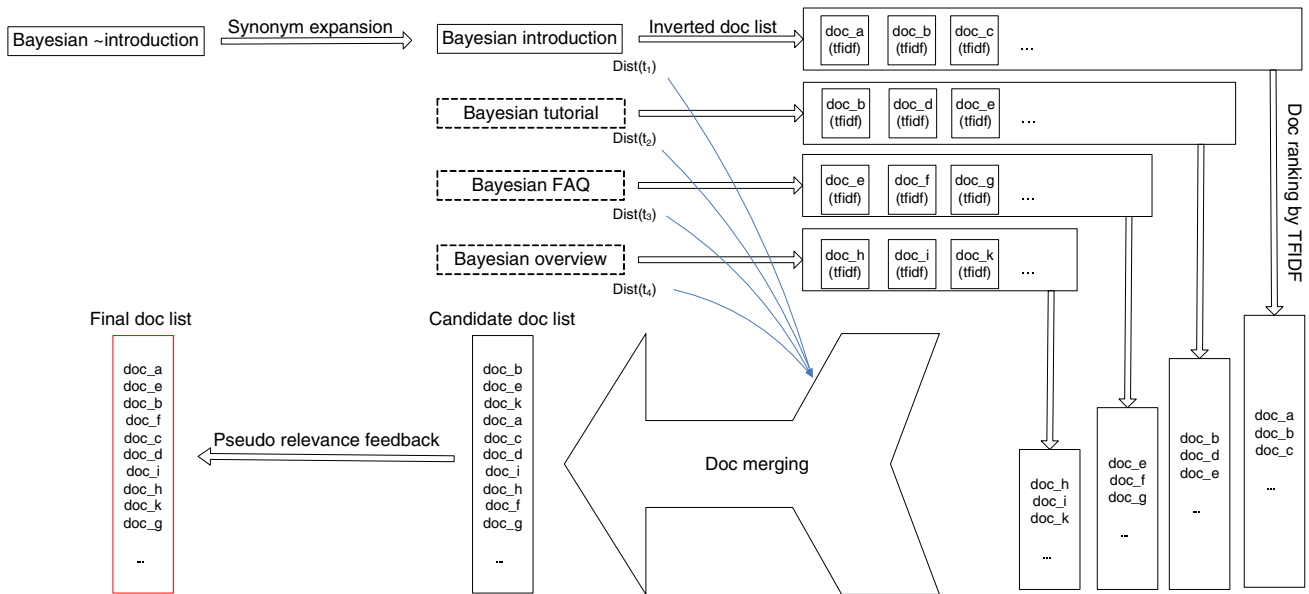


Fig. 1 Framework of information retrieval

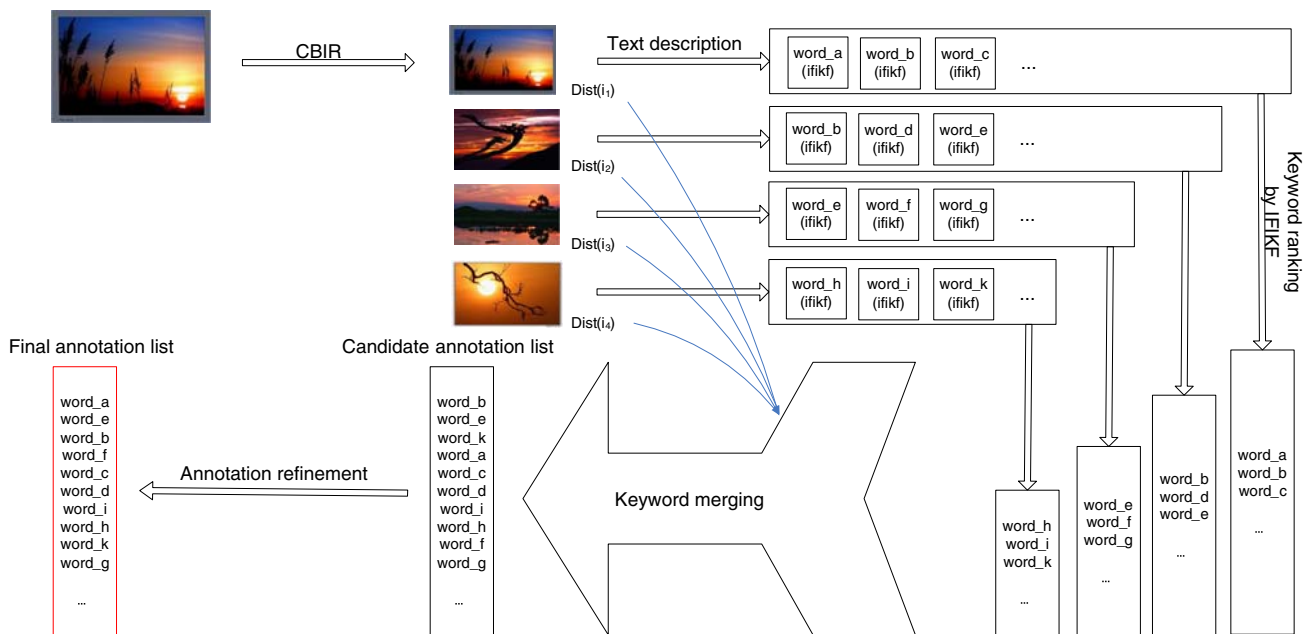


Fig. 2 Framework of search-based image annotation

query expansion, some relevant documents missed in the initial round can possibly be retrieved and more relevant documents could be ranked higher. Pseudo-relevance feedback can be considered as a score refinement process as follows:

$$Score(doc_a) = Refine(Score_{relevance}(doc_a)). \quad (2)$$

Based on the refined scores for each relevant document, we can rank all the relevant documents and return top ones as the search results to users.

### 2.2.2 Search-based image annotation

We re-formulate image annotation as a search process analogous to the aforementioned information retrieval. The search-based image annotation (SBIA) framework is shown in Fig. 2. Comparing Figs. 2 with 1, we can see that the images in image annotation correspond to the textual terms in information retrieval, while the annotation keywords in image annotation correspond to the documents in information retrieval. Given a target image as the query, our aim is to find the most

related keywords. Similar to the synonyms finding in information retrieval, CBIR technique is used to retrieve the most similar images to the target image from a large-scale Web image set. Assume that there are  $s$  similar images  $\{I_i\}_{i=1,\dots,s}$ . The distance between the target image and a similar image  $I_i$  is denoted as  $Dist(i)$ . Since all the similar images are Web images with textual descriptions, these descriptions could be considered as inverted files. For a keyword appearing in the description of an image, an image frequency  $\times$  inverse keyword frequency (if-ikf) measure similar to the tf-idf measure could also be defined to reflect the relevance between the image and keyword. We will give more details in Sect. 3.2. Then, the relevance score of a keyword to the target image is a weighted sum of the relevance scores corresponding to the similar images. More specifically, the relevance score of a keyword  $word\_a$  can be calculated as follows:

$$Score_{relevance}(word\_a) = \sum_{I_i \in S} f(Dist(i)) \times Score_{ifikf}(word\_a, I_i), \tag{3}$$

where  $S$  is the set of similar images of the target image.

Analogous to the pseudo-relevance feedback process of information retrieval, a novel approach to automatically refine the candidate annotations of images is proposed. We reformulate the image annotation refinement process as a graph ranking problem and solve it with the Random Walk with Restarts (RWR) algorithm. We will give more details in Sect. 3.4.

Similar to Eq. 2, the refined score of keyword  $word\_a$  could be calculated as follows:

$$Score(word\_a) = Refine(Score_{relevance}(word\_a)) \tag{4}$$

### 3 Search-based image annotation

There are four main components of the proposed SBIA algorithm: a CBIR stage to retrieve visually similar images, a text-based keyword search stage to obtain a ranked list of candidate annotations for each retrieved image, a fusion stage to combine the ranked lists into the final candidate annotation list, and an annotation refinement stage to automatically refine the candidate annotations. We also introduce the proposed annotation rejection method at the end of this part.

#### 3.1 Content-based image retrieval

About 2.4 million images were collected from several photo forum sites, e.g. photosig [2]. Images of such sites have rich textual information, such as title and photographer’s description. The textual information reflects the content of corresponding image to some extent. For details of the Web-scale dataset please refer to [36].

For a target image  $I_t$ , a typical CBIR technique is used to retrieve a set of visually similar images denoted by  $S$ . To represent an image, a 64-dimensional feature [37] was extracted. It is a combination of three features: 6 dimensional color moments, 44 dimensional banded auto-correlogram and 14 dimensional color texture moments. For color moments, the first two moments from each channel of CIE-LUV color space were extracted. For correlogram, the HSV color space with inhomogeneous quantization into 44 colors is adopted. Note that all existing global or local features and the corresponding distance measures could be used by the algorithm. To speed up the similarity search process, a K-means-based indexing algorithm is used [13]. The rationale of clustering-based high-dimensional indexing techniques is to reduce the search space by first finding a few cluster blocks whose centers are nearest to the query point, and then searching through the data points residing in these blocks. Since the data points visited are much less compared to the whole database, the search procedure can thus significantly speed up. In the implementation, 2.4 million images are indexed into 2,000 clusters. Given a query image, the most nearest cluster blocks are retrieved to insure that there are at least 10,000 images in the retrieved blocks. Several hundreds of images with largest similarities will be kept for further use in annotation mining step. With the indexing technique, less than 0.03 s is needed to retrieve 500 most similar images from 2.4 million images. For each retrieved image  $I_i \in S$ , the distance between  $I_i$  and the target image  $I_t$  is denoted as  $Dist(i)$ . In this implementation, the Euclidean distance is adopted.

#### 3.2 Text-based keyword search

For each image  $I_i \in S$ , a text-based keyword search process is used to rank all the related keywords. The related keywords are the ones that appear in title or description of  $I_i$  except the stop words and some inappropriate words such as “image” and “photo”. More specifically, for  $I_i$ , the set of related keywords are denoted as  $K_i$ . Denote  $K$  as the combination of all  $K_i$ , i.e.  $K = \cup_{I_i \in S} K_i$ . For each keyword  $K_j \in K$ , its relevance score to  $I_i$  is denoted as  $Score_{relevance}(i, j)$ .

Two strategies could be used to calculate  $Score_{relevance}(i, j)$ . One is to use the prominence score. Prominence score reflects the prominence of a keyword to annotate an image. The prominence score of a keyword  $K_j$  to  $I_i$  is defined as follows:

$$Score_{prominence}(i, j) = \begin{cases} \frac{occurrence(i, j)}{\sum_{K_k \in K_i} occurrence(i, k)} & K_j \in K_i \\ 0 & K_j \notin K_i \end{cases} \tag{5}$$

where  $occurrence(i, j)$  denotes the number of  $K_j$  in title or description of  $I_i$ .

Recall that the images in image annotation correspond to the textual terms in information retrieval, while the annotation keywords in image annotation correspond to the documents in information retrieval. Thus, in image annotation problem, a keyword could be considered as a document with several related images being keywords. An image is deemed as related to a keyword if the keyword appears in the title or description of the image. Then the aforementioned occurrence number could be considered as image frequency (IF) that is analogous to term frequency (TF). Similar to document frequency (DF), the keyword frequency (KF) of an image could be defined as the number of related keywords of the image. Enlightened by the tf-idf weighted scheme, a new scoring scheme based on if-ikf is proposed as follows:

$$Score_{ifikf}(i, j) = \begin{cases} \frac{occurrence(i, j)}{\log_2(sizeof(K_i)+1)} & otherwise \\ 0 & K_i = \phi \text{ or } K_j \notin K_i. \end{cases} \quad (6)$$

For efficiency, we index textual descriptions of images in the database. The textual description of each image  $I_i$  in the entire database is represented by a list of keyword-score pairs:  $(ID\_of\_K_j, Score_{ifikf}(i, j))$ . Here  $K_j \in K_i$ . For an image  $I_i$  in the database, a text engine could directly return the list of keyword-score pairs to the annotation system.

### 3.3 Merging stage

The ranked lists of all similar images in  $S$  are further merged to get the final candidate annotation keywords of the target image. Considering that the similar images have different similarities with the target image and more similar image should have more impact on the final annotation results, we use the following formula to score and rank the keywords  $K_j \in K$ :

$$Score_{relevance}(K_j) = \sum_{I_i \in S} f(Dist(i)) \times Score_{relevance}(i, j), \quad (7)$$

where  $f(*)$  is a function that transforms distance to similarity.  $f(*)$  is defined as follows:

$$f(d) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{-d^2}{2\sigma^2}\right\} \quad (8)$$

Once the relevance score of each keyword in  $K$  is obtained, we can pick up the top  $N$  as the final candidate annotations  $\{w_i\}_{i=1, \dots, N}$  with corresponding confidence score  $\{Score_{relevance}(w_i)\}_{i=1, \dots, N}$  to be refined in the next annotation refinement stage.

### 3.4 Annotation refinement with RWR

In the candidate annotation set  $\{w_i\}_{i=1, \dots, N}$ , there may be some irrelevant annotations. We can explore the relationship among candidate annotations to try to discard the noisy ones.

Jin et al. [17] have done pioneer work on annotation refinement using a generic knowledge-based WordNet. From the small candidate annotation set obtained by an annotation method, the irrelevant annotations will be pruned using WordNet [22]. The basic assumption is that highly correlated annotations should be reserved and non-correlated annotations should be removed. However, in [28, 29] we have testified that there are two main limitations using WordNet. One is that the similarity defined using WordNet is sometimes not appropriate for the annotation refinement problem. For example, “mountain” and “sky” usually appear in a scenery photo together, while “tree” and “flag” seldom simultaneously appear in an image. However, with the JCN measure in WordNet, the similarities of the above two pairs of words are 0.061 and 0.148, respectively, which is unreasonable. With the proposed similarity measure, the two similarities will be 0.430 and 0.095, respectively, which is more reasonable. The other limitation is that it cannot deal with the annotations that do not exist in the lexicon of WordNet. For example, there are 49 out of 374 words of the Corel dataset<sup>2</sup> which either do not exist in WordNet lexicon or have zero similarity with all other words using the JCN measure [22], while all of them exist in our unlimited lexicon.

In order to fully utilize the confidence scores of the candidate annotations obtained by former stages and the Web-scale image set as aforementioned in Sect. 3.1, we reformulate the image annotation refinement process as a graph ranking problem and solve it with the RWR algorithm. The basic assumption of the algorithm is that highly correlated keywords should be ranked higher. Thus, by modeling the candidate keywords as vertices of a graph and defining certain similarity as the edges to measure the correlation between keywords, the scores could be propagated among correlated keywords. By formulating the image annotation refinement process as a graph ranking problem, two disconnected keywords could also influence each other iteratively, and finally their scores converge to a static state.

#### 3.4.1 Graph construction

Each candidate annotation  $w_i$  is considered as a vertex of a graph  $G$ . All vertices of  $G$  are fully connected with proper weights. The weight of an edge is defined based on the “co-occurrence” similarity as below.

We have built an image search engine named as Enjoy-Photo [36] based on the Web-scale image set as aforementioned

<sup>2</sup> <http://www.photosig.com>

tioned in Sect. 3.1. Each word  $w_i$  will be used as a query to query EnjoyPhoto. The number of search results is denoted as  $num(i)$ . For two different word  $w_i$  and  $w_j$ , “ $w_i w_j$ ” will be used as the query. The number of search results is denoted as  $num(i, j)$ . The weight of the edge between  $w_i$  and  $w_j$  is then calculated by the following formula:

$$sim(w_i, w_j) = \begin{cases} \frac{num(w_i, w_j)}{\min(num(w_i), num(w_j))} + \epsilon & num(w_i, w_j) > 0 \\ \epsilon & num(w_i, w_j) \leq 0 \end{cases} \quad (9)$$

where  $\epsilon$  is a constant satisfying  $0 < \epsilon \ll 1$ .

### 3.4.2 The RWR algorithm

The RWR algorithm [24] performs as follows. Assume that there is a random walker that starts from node  $w_i$  with a certain probability. At each time-tick, the walker has two choices. One is to randomly choose an available edge to follow. The other choice is to jump to  $w_j$  with probability  $c \times v(j)$ , where  $v$  is the restart vector and  $c$  is the probability of restarting the random walk [24].

### 3.4.3 Image annotation refinement with RWR

Assume that  $G$  is a graph with  $N$  vertices  $\{w_i\}_{i=1\dots N}$  constructed as in Sect. 3.4.3. Let  $A$  be the adjacency matrix of  $G$ .  $A$  is column-normalized to ensure that the sum of each column in  $A$  is one. The original confidence scores of candidate annotations are considered as the restart vector  $v$ .  $v$  is normalized to ensure the sum of all elements in  $v$  is one. The aim is to estimate the steady-state probability of all vertices, which is denoted by  $u$ . Let  $c$  be the probability of restarting the random walk. Then the  $N$ -by-1 steady state probability vector  $u$  satisfies the following equation:

$$u = (1 - c)Au + cv \quad (10)$$

The iteration of Eq. 10 can be guaranteed to converge if the transition matrix  $A$  is stochastic and primitive, which can be satisfied by our definition to the edge of the graph in Eq. 9. Thus, we have

$$u = c(I - (1 - c)A)^{-1}v \quad (11)$$

where  $I$  is the  $N \times N$  identity matrix.

The  $i$ th element  $u(i)$  of the steady-state vector  $u$  is the probability that  $w_i$  can be the final annotation.

We can choose the top  $m$  annotations with highest probabilities as the final annotations.

### 3.5 Image annotation rejection

In our system, we use a 64-dimensional global feature to represent an image. It is worth noting that global features are

helpful for image-level concept annotation, but are not very powerful for object-level annotation. It is thus necessary to develop a scheme to estimate the confidence of our system to annotate the query images, and thereby reject the ones we cannot handle well. We first construct a consistency map for the Web-scale image collection, on which each image has a consistency value to represent the local semantic consistency between this image and its visually similar images. Then, for a query image, we use the consistency map to calculate the annotation confidence of the query image.

Following the notations in Sects. 3.1 and 3.2, we use the following formula to calculate the consistency value of each image  $I_i$  in the collection:

$$Score_{consistency}(I_i) = \sum_{I_j \in S_i} Sim_{visual}(I_i, I_j) \times Sim_{semantic}(I_i, I_j), \quad (12)$$

where  $S_i$  is the similar image set of image  $I_i$ .  $Sim_{visual}(I_i, I_j)$  represents the visually similarity between image  $I_i$  and  $I_j$ , which can be calculated by Eq. 8.  $Sim_{semantic}(I_i, I_j)$  represents the semantic consistency between image  $I_i$  and  $I_j$ . Since in the Web-scale image collection, each image  $I_i$  has related keyword descriptions denoted as  $K_i$  (see Sect.3.2), we can utilize the cosine similarity [27] between  $K_i$  and  $K_j$  to calculate  $Sim_{semantic}(I_i, I_j)$ .

After the whole consistency map of the Web-scale image collection is constructed, we utilize this map to calculate the annotation confidence of the target image  $I_t$ :

$$Score_{confidence}(I_t) = \sum_{I_j \in S_t} Sim_{visual}(I_t, I_j) \times Score_{consistency}(I_j). \quad (13)$$

Therefore, for a query image, besides the annotation list, our system can offer a confidence score of the annotation list for the query image. We can inform users that we cannot handle the query image if this score is below certain threshold.

### 3.6 Comparison with AnnoSearch

In AnnoSearch [31], given a target image  $I_t$  and a related keyword  $w_t$ , the goal is to annotate  $I_t$  with more related keywords. The annotation process could be roughly separated into two stages. First, the most similar images  $S$  of a large-scale Web image dataset are retrieved using both  $w_t$  and  $I_t$ . Then, the search result clustering (SRC) technology [35] is used to mine the annotations  $\mathbf{w}^*$  from text descriptions of  $S$ . The two stages can be expressed as the following formula:

$$p(\mathbf{w}^*|I_t, w_t) = p(\mathbf{w}^*|S)p(S|I_t, w_t). \quad (14)$$

With SRC, the most representative topics  $\mathbf{t}^*$  could be detected based on which  $\mathbf{w}^*$  could be learned. Therefore,

Eq. 14 can be rewritten as follows:

$$p(\mathbf{w}^*|I_t, w_t) \approx p(\mathbf{w}^*|\mathbf{t}^*)p(\mathbf{t}^*|S)p(S|I_t, w_t). \quad (15)$$

As the first work that leverages the Web-scale training corpus, both the scalability and performance of AnnoSearch are satisfactory. However, it still has spaces for improvement. The main limitation is the requirement of an initial accurate keyword which is not always available, especially for personal images. Moreover, since SRC was initially designed for general Web search, directly using it in annotation mining may not be an optimal solution. Furthermore, the similar images are treated equally ignoring their similarities to the target image. The proposed four-stage SBIA algorithm could be expressed as the following formula:

$$p(\mathbf{w}^*|I_t) = RWR \left( \sum_{I_i \in S} p(\mathbf{w}^*|I_i)p(I_i|I_t) \right). \quad (16)$$

First, the most similar images  $S$  are retrieved based on CBIR techniques. Meanwhile, the similarity between target image  $I_t$  and an image  $I_i \in S$  is reserved as  $p(I_i|I_t)$ . Second, for each image  $I_i \in S$ , the relevance score of the keywords  $\mathbf{w}^*$  is obtained by text-based keyword search technology. It is denoted by  $p(\mathbf{w}^*|I_i)$ . Third, all relevance scores from different images in  $S$  are merged for each keyword in  $\mathbf{w}^*$ . Finally, the top candidate keywords in  $\mathbf{w}^*$  are refined by the RWR algorithm, and the top ones after refined are returned to users as the final annotations of  $I_t$ . Note that no initial keyword is needed in advance and the similar images are used according to their similarities to the target image.

## 4 Experimental results

A series of experiments were conducted on U. Washington dataset [1] to evaluate the proposed SBIA algorithm. First, we use the Web-scale dataset as the training set to test the U. Washington images in the image annotation task. Then, to show the effectiveness of the annotations on image retrieval, the retrieval performance of query by keyword (QBK) using SBIA results was compared with that of query by example (QBE) using visual features.

### 4.1 Image annotation

The aim of this work is to automatically annotate arbitrary images with unlimited vocabulary by leveraging a Web-scale image set. Thus, in this section, we show detailed experimental results of this part.

As mentioned in Sect. 3.1, we use the 2.4 million Web images associated with meaningful descriptions as the Web-scale training set, which is consistent with the implementation in [31] and [21]. We use U. Washington dataset (UW) as

the testing set. UW is a CBIR database, which can be downloaded from the University of Washington. There are about 5 manually labeled ground truth annotations for each image in UW. Totally, there are 1,109 images and more than 350 unique words. Although, for some images, not all objects in them are annotated, we strictly use the annotations of UW as the ground truth annotations and the synonyms and non-appearing correct annotations are assumed incorrect, if there is no other explanation.

First, several variations of SBIA algorithm were evaluated. Then SBIA was compared with a modified version of AnnoSearch [31] that removed the dependency of initial keywords and LiAnnoSearch [21].

#### 4.1.1 Experimental design

In AnnoSearch, there is a strong constraint that an accurate query keyword should be provided. The UW folder names are used as the initial query keyword [31]. Although the initial keyword might speed up the search process and result in more relevant images to the query image, it is not always available, especially for the personal images. Due to the unavailability of accurate query keywords for personal images, we ignore the query keyword when we implement AnnoSearch algorithm, and only use the query image. More specifically, the first stage of AnnoSearch is replaced with the first stage of SBIA. Moreover, the average member image score criterion [31] is used in the modified version of AnnoSearch (MANnoSearch) due to its good performance in [31].

Two strategies are used to evaluate the annotation performance: phrase-level strategy and term-level strategy. The original ground truth annotations include both words and phrases. In the phrase-level strategy, an annotation is considered to be correct if and only if it is a ground truth annotation of the target image. In the term-level strategy, both the ground truth annotation phrases and the result annotation phrases are divided into separate words. If there is more than one same word in the annotations of an image, only one is reserved. An annotated word is considered to be correct if and only if it appears in the ground truth annotation of the target image. The precision and recall for the two strategies are defined as follows:

$$\begin{aligned} precision_{phrase} &= \frac{1}{n} \sum_{k=1}^n \frac{correct_p(k)}{automatic_p(k)} \\ recall_{phrase} &= \frac{1}{n} \sum_{k=1}^n \frac{correct_p(k)}{groundtruth_p(k)} \\ precision_{term} &= \frac{1}{n} \sum_{k=1}^n \frac{correct_t(k)}{automatic_t(k)} \\ recall_{term} &= \frac{1}{n} \sum_{k=1}^n \frac{correct_t(k)}{groundtruth_t(k)}, \end{aligned} \quad (17)$$

where  $correct_p(correct_t)$  is the number of correctly annotated phrases (terms) of the testing image  $I_k$ .  $automatic_p$



( $automatic_t$ ) is the number of automatically annotated phrases (terms) in  $I_k.groundtruth_p$  ( $groundtruth_t$ ) is the number of ground truth phrases (terms) in  $I_k.n$  is total number of testing images.

Although the proposed algorithm is able to produce phrase level annotations by using n-gram model [5], currently only term-level annotations are considered for simplicity. As a result, the phrase-level evaluation strategy will be disadvantageous for the proposed SBIA algorithm. However, the experimental results in the following sections will show that the annotation results of SBIA are still satisfactory even gauged by the phrase-level strategy.

Due to the use of almost unlimited vocabulary, there will be a long list of annotations for both MAnnoSearch and SBIA. Therefore, the number of annotation results is restricted to be no more than 10. With  $m$  result annotations, the precision and recall are denoted by  $precision@m$  and  $recall@m$ .

First, two variations i.e. size of similar image set and scoring strategy were evaluated, without the refinement process (SBIA-N). After the variations were fixed, SBIA-N was compared with MAnnoSearch. Second, we evaluated the proposed image annotation refinement process in SBIA. After fixing the restart parameter, we compared the SBIA and MAnnoSearch with the refinement process (MAnnoSearch-Y), followed by the comparison between SBIA and LiAnnoSearch. Then, the experiment of annotation rejection was shown. Finally, since our algorithm can predict annotations outside the ground truth, we also provided some results using manual evaluation.

#### 4.1.2 Size of similar image set

The size of visually similar image set of the first CBIR stage is a common and crucial parameter in both MAnnoSearch and SBIA. Let's denote it by  $s$ . To facilitate the further evaluations,  $s$  is first decided for both algorithms by comparing the annotation performance with different  $s$ . The number of annotation results  $m$  is fixed to 5 and both prominence-based and if-ikf-based scoring strategies are considered. The precision and recall are shown in Fig. 3a and b with  $s$  changed from 20 to 2,500. Five conclusions could be drawn from Fig. 3a and b. First, the changing trends of precision and recall are similar. Second, the absolute values of term-level evaluation are consistently better than that of phrase-level evaluation, which coincides with our intuition. Third, the performance of MAnnoSearch is best when  $s$  is 200. Therefore,  $s$  is set to be 200 for MAnnoSearch in the following evaluations. Fourth, the changing trends of both scoring strategies are similar. Last, the performance of SBIA is similar when  $s$  is equal or larger than 500. Considering that the smaller the value of  $s$ , the more efficient the annotation process will be, we set  $s$  to be 500 for SBIA.

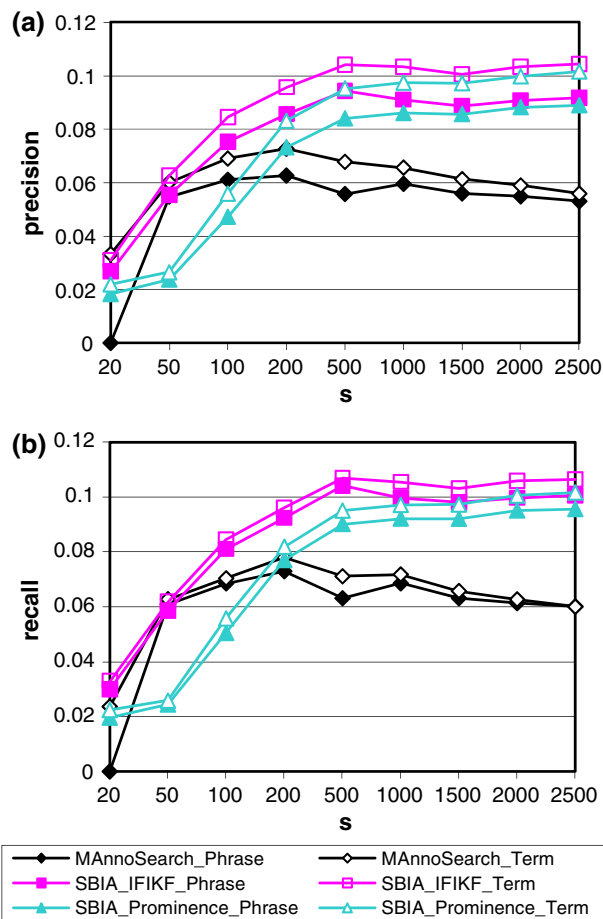


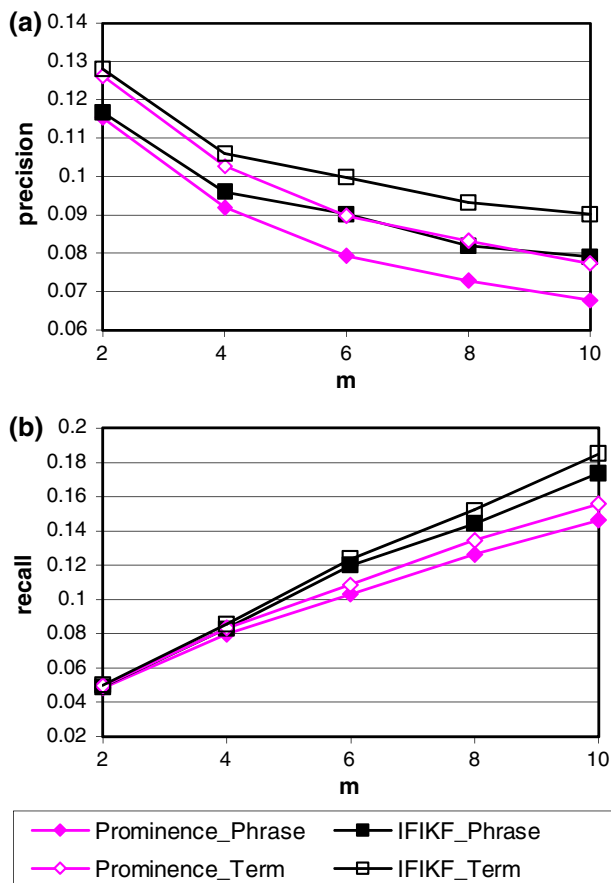
Fig. 3 a, b Annotation precision and recall of different sizes of similar image set

#### 4.1.3 Scoring strategy

The two scoring strategies of SBIA as discussed in Sect. 3.2 were evaluated and compared. From Fig. 3a and b we can see that the if-ikf strategy is consistently better than prominence strategy when  $m$  is 5. To further compare the two strategies,  $m$  was varied from 2 to 10. The comparison results of if-ikf strategy and prominence strategy are shown in Fig. 4a and b. Although the performance of if-ikf strategy is similar to that of prominence strategy when  $m$  is less than 4, both precision and recall of if-ikf strategy outperform those of prominence strategy when  $m$  is larger than 4. Therefore, in the following experiments, the if-ikf strategy is used for SBIA.

#### 4.1.4 MAnnoSearch versus SBIA-N

Based on the aforementioned preparations, SBIA without refinement process (SBIA-N) was compared with MAnnoSearch using the according parameters. Figure 5a and b show the comparison results. SBIA-N consistently outperforms MAnnoSearch while  $m$  is ranging from 2 to 10.

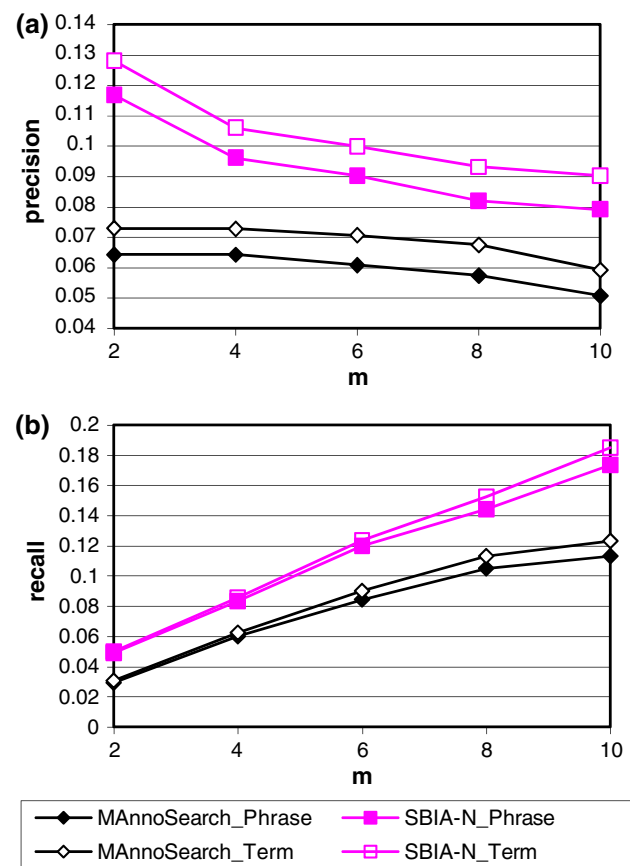


**Fig. 4** a, b Annotation precision and recall of different relevance scores

#### 4.1.5 Image annotation refinement

In this section, we evaluate the image annotation refinement process of SBIA. The number of candidate annotations is set to be 10. For each query image, we refine the ten candidate annotations obtained by SBIA-N using the proposed RWR algorithm. The restart parameter  $c$  was varied from 0 to 1. Recall that SBIA degenerates to be SBIA-N when  $c$  is 1.

The experimental results of annotation refinement based on SBIA-N are shown in Figs. 6a, b and 7a, b. Both results of term-level and those of phrase-level are shown. Four conclusions could be drawn from Fig. 7a and b. First, the changing trends of term-level evaluation and phrase-level evaluation are similar. Second, both precision and recall rates reach to the lowest value when  $c$  is 1 with  $m$  fixed, which indicates that the refinement process did improve the image annotation performance. Third, the improvement of precision is more significant than that of recall. Fourth, since the refinement process is based on the same candidate annotation list, the performances with different  $c$  tend to be consistent while  $m$  is approaching 10.



**Fig. 5** a, b Annotation precision and recall comparison of MAnnoSearch and SBIA-N

The refinement process is orthogonal to the method of producing candidate annotations. We have conducted the same evaluation of annotation refinement for MAnnoSearch. All the aforementioned conclusions still stand. Since MAnnoSearch with refinement (MAnnoSearch-Y) outperforms MAnnoSearch, we compared SBIA with MAnnoSearch-Y in the next sub-section. We set  $c$  to be 0.3 for both SBIA and MAnnoSearch-Y in the following evaluations, for it is the best parameter for both of them.

#### 4.1.6 MAnnoSearch-Y versus SBIA

Based on the aforementioned preparations, SBIA was compared with MAnnoSearch-Y using the according parameters. Figure 8a and b show the comparison results. SBIA consistently outperforms MAnnoSearch-Y while  $m$  is ranging from 2 to 10.

#### 4.1.7 LiAnnoSearch versus SBIA & MAnnoSearch

In [21], Li et al. presented a modified version of AnnoSearch (LiAnnoSearch) by removing the dependency of the initial accurate keyword  $w_l$ . As the same as AnnoSearch,

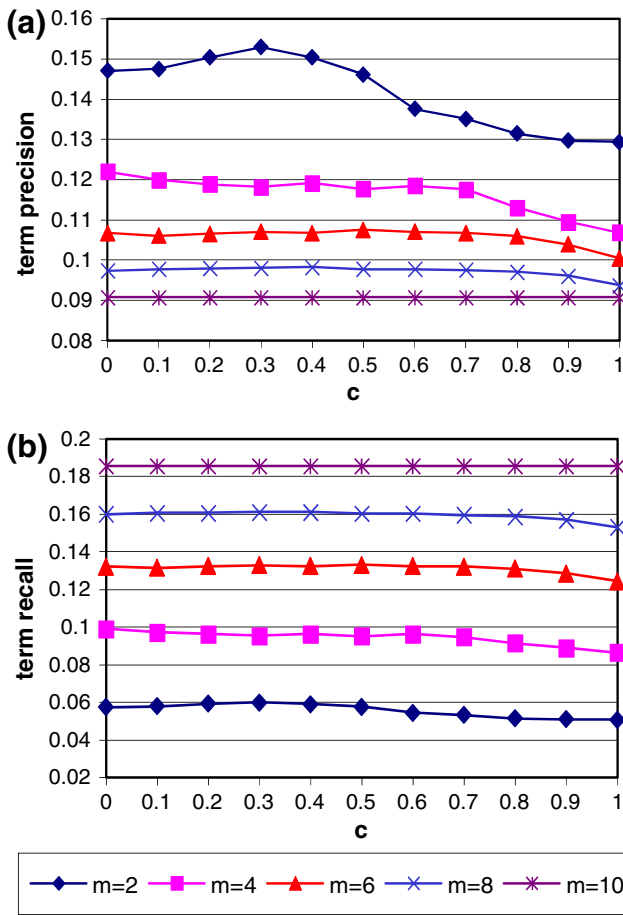


Fig. 6 a, b Annotation precision and recall of different restart parameters using term-level evaluation

LiAnnoSearch solved image annotation problem in a search and mining process. Besides the independency of  $w_t$ , the main difference between AnnoSearch and LiAnnoSearch in search stage is that, a novel high-dimensional indexing algorithm is proposed to index high-dimensional visual features in LiAnnoSearch.

Table 1 shows the comparisons of SBIA, LiAnnoSearch and MAnnoSearch under term-level evaluation strategy, which is the only evaluation strategy in LiAnnoSearch. The results of LiAnnoSearch are directly referred to [21], in which the number of final annotations  $m$  is dynamically determined. To get a fair evaluation, we directly fix  $m$  to be 10 in MAnnoSearch and SBIA. In Table 1 we can see that LiAnnoSearch is comparable with MAnnoSearch. That is reasonable since both of them can be considered as the modified versions of AnnoSearch by removing the dependency of initial keywords, and except the dependency of initial accurate keyword, all other problems in AnnoSearch could also be suffered in both of MAnnoSearch and LiAnnoSearch. Moreover, the results also show that the high-dimensional indexing algorithm proposed in LiAnnoSearch cannot remarkably

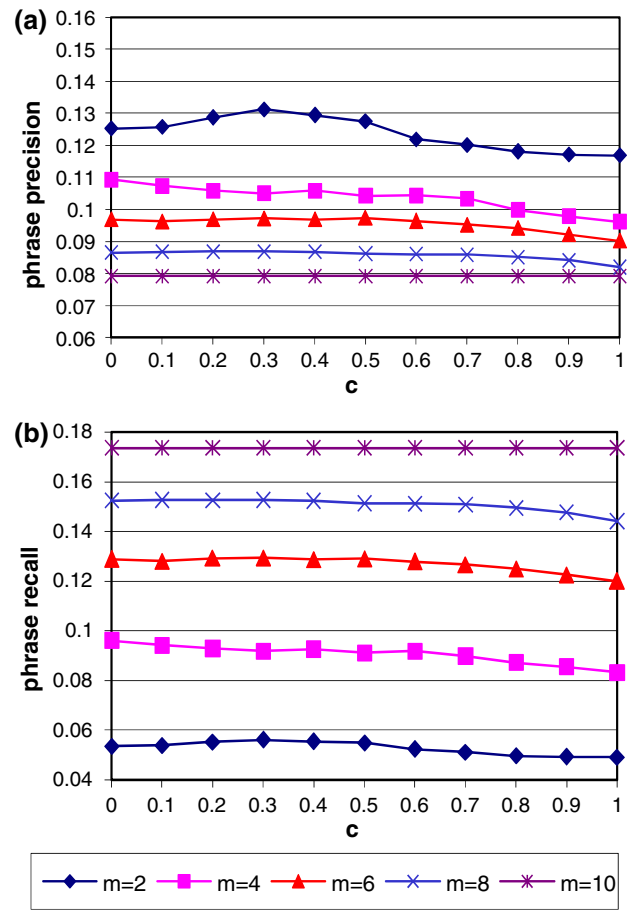


Fig. 7 a, b Annotation precision and recall of different restart parameters using phrase-level evaluation

improve annotation performance. Different from AnnoSearch and LiAnnoSearch, we formulate image annotation as a search problem. From this perspective, we further find that search-based image annotation and information retrieval can be treated as dual problems, and therefore we can systematically leverage those mature information retrieval technologies. Therefore, SBIA outperforms both of LiAnnoSearch and MAnnoSearch.

It took less than 0.23 s (less than 0.03 s for image search process and less than 0.2 s for annotation mining process) and about 800 M memory (300 M for visual feature and 500 M for textual feature) on average to annotate one image on a Web server with two 2.0 GHz CPUs and 2 GB memory, which is comparable with LiAnnoSearch and MAnnoSearch. The rather high efficiency of SBIA guarantees its application to large personal image collections.

#### 4.1.8 Image annotation rejection

In Sect. 3.5, we have developed a method to calculate the consistency score to represent the local semantic consistency

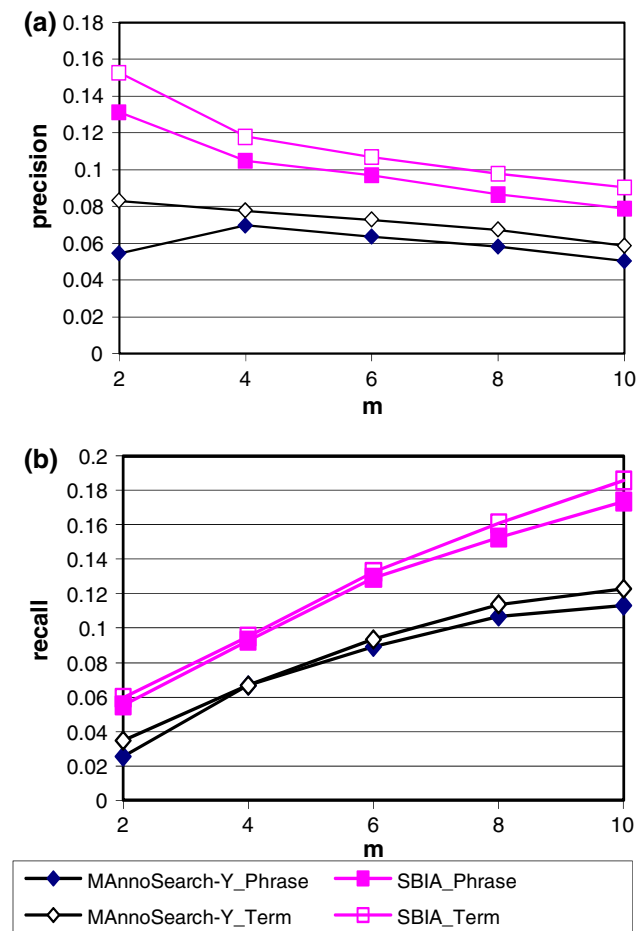


Fig. 8 a, b Annotation precision and recall comparison of MAnnoSearch-Y and SBIA

Table 1 Comparison of SBIA, LiAnnoSearch, and MAnnoSearch

Models	MAnnoSearch	LiAnnoSearch	SBIA
Precision	0.06	0.07	0.09
Recall	0.12	0.11	0.18

between the query image and its visually similar images in the database. This score could be used to estimate the confidence of our system to annotate the query image, and thereby reject the ones we cannot handle well. To show the effectiveness of the rejection scheme, we remove the images of UW with lowest confidence scores, and evaluate other confident images. Figure 9a and b shows the average precision and recall performance of remaining query images after rejecting certain images with lowest consistency scores (e.g. 20% of the query image set), using term-level evaluation. The larger the rejection rate is, the more confident the left query images are. The constant improvement by increasing rejection rate shows that the rejection scheme could improve the annotation performance by automatically rejecting the ones we cannot handle well.

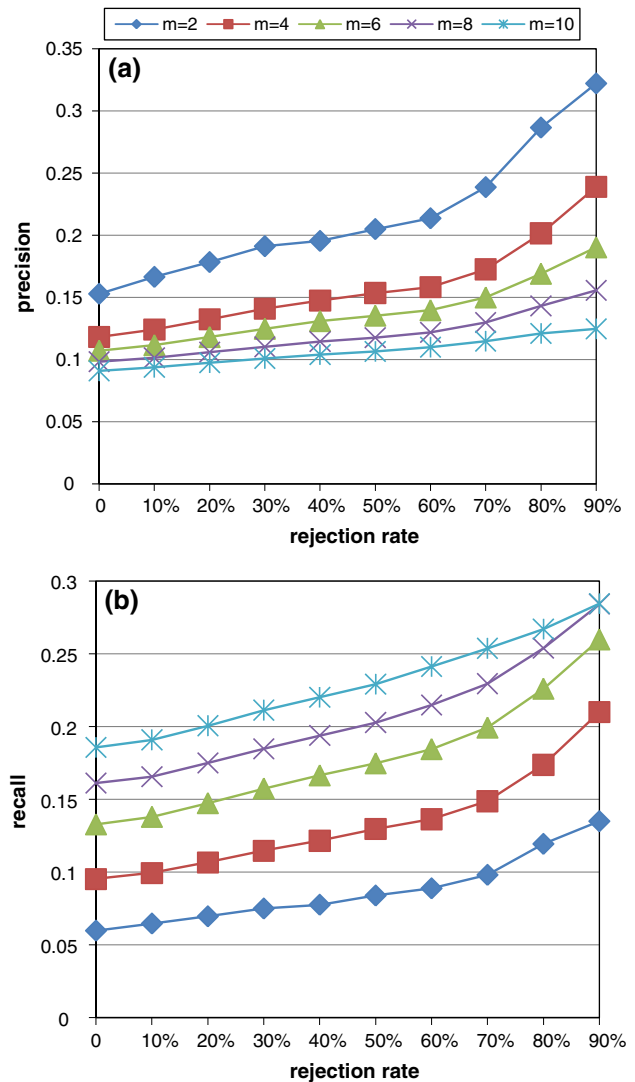
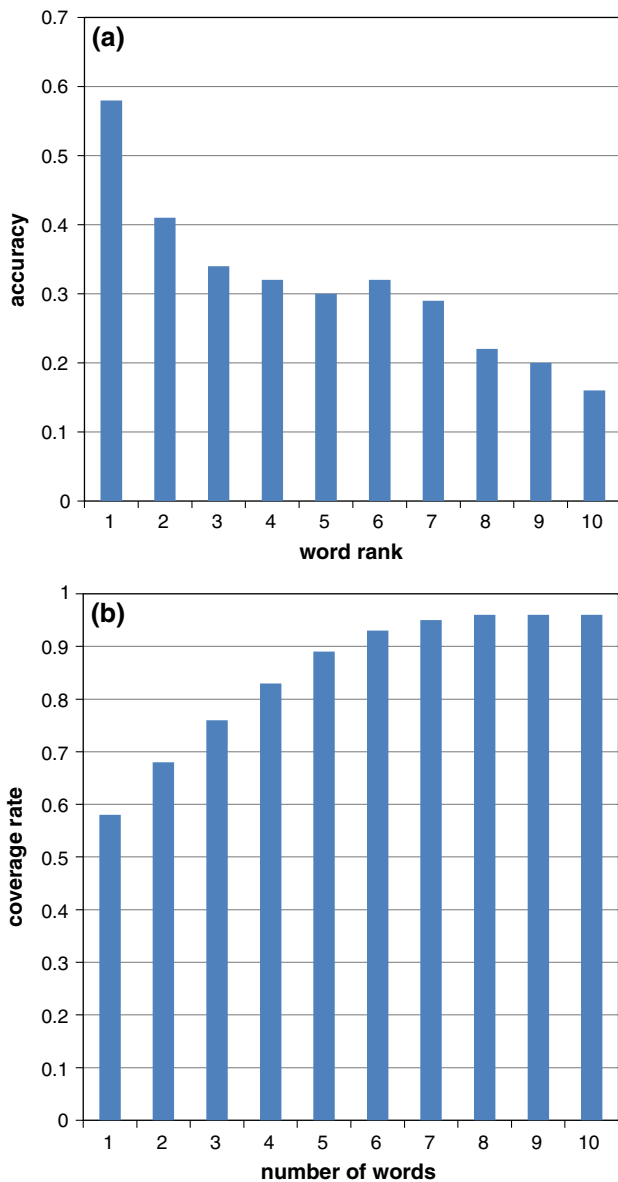


Fig. 9 a, b Annotation precision and recall with different rejection rate

#### 4.1.9 Results using manual evaluation

In the previous evaluations, we strictly used the annotations of UW as the ground truth annotations, and thus the synonyms and non-appearing correct annotations were assumed incorrect. Although this evaluation method can provide a fair and easy comparison between different annotation works, it may shrink the annotation performance, for the proposed SBIA annotation system can predict annotations outside of the ground truth. Since one of the advantages of our method is to annotate images using unlimited vocabulary, we also show some results using manual evaluation instead of strictly using the ground truth of UW dataset.

Similar to the evaluation system of [20], each image in 100 random selected images of UW dataset is shown together with 10 annotation words assigned by SBIA algorithm. A trained person, who did not participate in the development



**Fig. 10** Annotation performance based on manual evaluation of 100 random selected images of UW dataset. **a** Percentages of images correctly annotated by the  $m$ th word. **b** Percentages of images correctly annotated by at least one word among the top  $m$  words

of the system, examines every word against the image and checks a word if it is judged as correct. Annotation performance is reported from two aspects in Fig. 10. Figure 10a shows accuracies, that is, the percentages of images correctly annotated by the  $m$ th word. The first word achieves an accuracy of 58%. The accuracy decreases gradually with  $m$  except for minor fluctuation with the sixth and seventh words. This reflects that the ranking of the words by the SBIA system is on average consistent with the true level of accuracy. Figure 10b shows the percentages of images correctly annotated by at least on word among the top  $m$  words. Using only six auto-

matic annotations produced by the system, we can achieve 90% coverage. Some examples in Fig. 11 also show the ability of the proposed annotation system. From Fig. 11 we can see that although some auto-annotations do not appear at UW ground truth, they are still reasonable to be annotations of the corresponding image.

#### 4.2 Image retrieval

To show the effectiveness of the annotations of SBIA on image retrieval, the retrieval performance of query by keyword (QBK) using SBIA results was compared with that of query by example (QBE) using visual features. Ten terms were selected from the UW dataset as the query concepts. The terms are chosen based on both popularity and diversity. The ten terms with corresponding occurrence number in both ground truth and the results of SBIA are listed in Table 2.

##### 4.2.1 Evaluation measures

Precision and recall are used as the basic measures. Since the queries of QBK and QBE are different, we re-define the precision and recall accordingly.

For QBK, the precision and recall of a term  $t$  are defined as follows:

$$\begin{aligned}
 precision@m &= \frac{correct\_t(m)}{m} \\
 recall@m &= \frac{correct\_t(m)}{groundtruth\_t},
 \end{aligned}
 \tag{18}$$

where  $correct\_t(m)$  is the number of correctly retrieved images in the first  $m$  images and  $groundtruth\_t$  is the number of images annotated by term  $t$  in ground truth.

For QBE, denote the  $m$  most similar images to an image  $I_i$  by  $S_i$ . Denote the images in  $S_i$  that are annotated by a term  $t$  in ground truth by  $S_i^t$ . The precision and recall of the term  $t$  are defined as follows:



$$\begin{aligned}
 precision@m &= \frac{1}{n_t} \sum_{I_i \in S^t} \frac{correct_i^t(m)}{m} \\
 recall@m &= \frac{1}{n_t} \sum_{I_i \in S^t} \frac{correct_i^t(m)}{groundtruth\_t},
 \end{aligned}
 \tag{19}$$

where  $S^t$  is the set of images annotated by  $t$  in ground truth.  $n_t$  is the size of  $S^t$ .  $correct_i^t(m)$  is the number of images in  $S_i^t$ .

##### 4.2.2 Retrieval comparison

The experimental results of image retrieval for each selected term using either QBK or QBE are shown in Figs. 12a, b.  $m$  is set to be 10. Although for people, snow and rock, QBE is slightly better than QBK, QBK is apparently superior to QBE

**Fig. 11** Annotation results for several UW images. The *second row* is the auto-annotation produced by SBIA system (top 5 results). The *third row* is the ground truth annotations of UW images set

			
<b>Auto-annotation:</b> blue, sky, landscape, water, sea	<b>Auto-annotation:</b> people, street, children, girl, friend	<b>Auto-annotation:</b> landscape, sky, water, white, black	<b>Auto-annotation:</b> sunset, sunrise, sun, people, tree
<b>UW Ground truth:</b> partly cloudy sky, hills, trees, small sailboat, water	<b>UW Ground truth:</b> husky alumni band, cheerleaders, hec ed pavilion, cobblestones, windows, 2 rounded entryways, drum, horn instruments	<b>UW Ground truth:</b> clouds, frozen, lake, mountain, sky	<b>UW Ground truth:</b> boats, sailboats, sky, sunrise/sunset

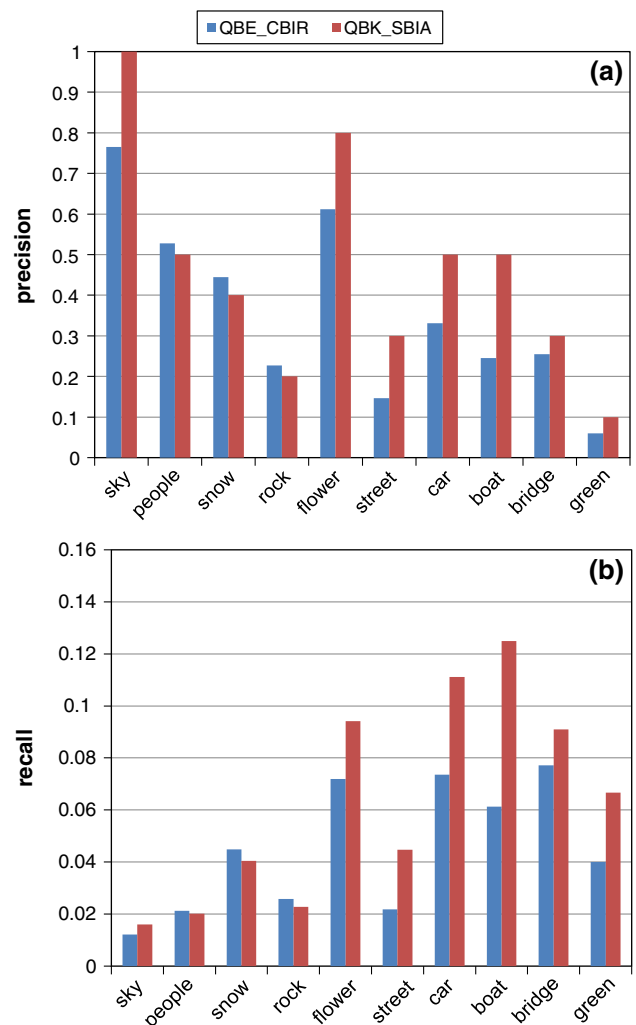
**Table 2** Selected terms and their corresponding occurrence number in ground truth and our annotation results

Term:	Sky	People	Snow	Rock	Flower
Num (GT)	627	248	99	88	85
Num (My)	150	558	88	116	112
Term:	Street	Car	Boat	Bridge	Green
Num (GT)	67	45	40	32	15
Num (My)	390	46	179	239	13

for the other 7 concepts. The average retrieval performances with different  $m$  are shown in Figs. 13a and b. Both precision and recall of QBK are consistently higher than those of QBE.

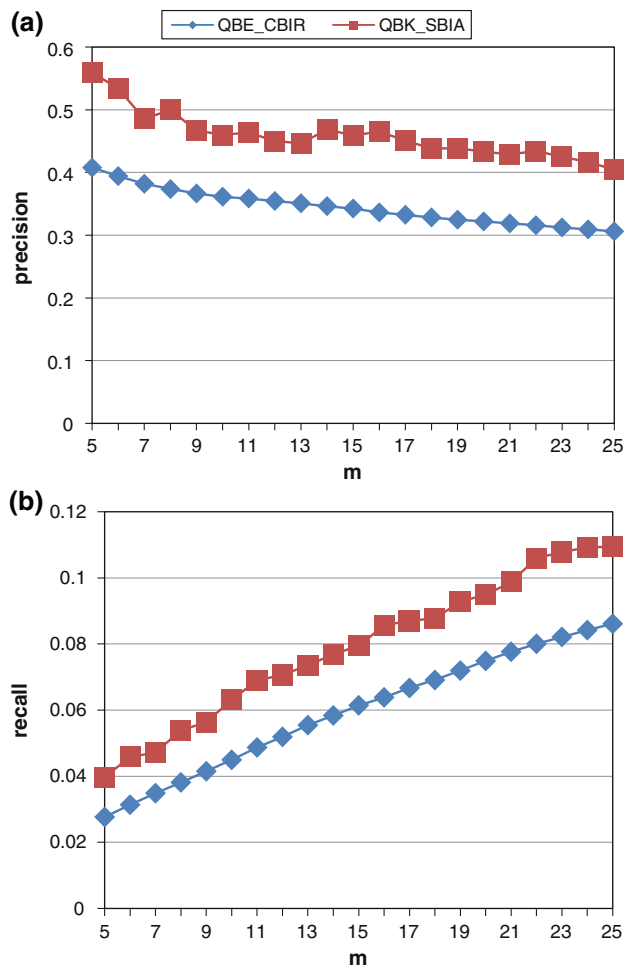
## 5 Conclusions and future work

In this paper, by formulating image annotation as a search problem, we have presented a novel search-based image annotation algorithm that is analogous to information retrieval. First, CBIR technology is used to retrieve a set of visually similar images from a large-scale Web image set. Second, a text-based keyword search technique is used to obtain a ranked list of candidate annotations for each retrieved image. Third, a fusion algorithm is used to combine the ranked lists into the final candidate annotation list. Finally, the candidate annotations are re-ranked using RWR and only the top ones are reserved as the final annotations. The application of both efficient search technologies and Web-scale image set guarantees the scalability of the proposed algorithm. Experimental results on U. Washington datasets show not only the effectiveness and efficiency of the proposed algorithm but also the advantage of image retrieval using annotation results over that using visual features.



**Fig. 12** a, b Retrieval precision and recall of different terms

By re-formulating image annotation as a search process, several effective information retrieval techniques could be adaptively used and deeper insight into the annotation



**Fig. 13** a, b Average retrieval precision and recall

problem could be gained. For example, more effective ranking functions, e.g. BM25 [25] could be possibly used with proper search techniques in image annotation environment in the future.

## References

- <http://www.cs.washington.edu/research/imagedatabase/groundtruth/>
- <http://www.photosig.com>
- Baeza-Yates, R.A., Ribeiro-Neto, B.: Modern Information Retrieval. Addison-Wesley Longman Publishing Co., Inc., Reading (1999)
- Blei, D.M., Jordan, M.I.: Modeling annotated data. In Proceedings of SIGIR, 2003
- Brown, P.F., deSouza, P.V., Mercer, R.L., Pietra, V.J., Lai, J.C.: Class-based  $n$ -gram models of natural language. *Comput. Linguist.* **18**(4), 467–479 (1992)
- Carneiro, G., Vasconcelos, N.: Formulating semantic image annotation as a supervised learning problem. In: Proceedings of CVPR (2005)
- Carneiro, G., Vasconcelos, N.: A database centric view of semantic image annotation. In: Proceedings of SIGIR (2005)
- Chang, E., Kingshy, G., Sychay, G., Wu, G.: CBSA: content-based soft annotation for multimodal image retrieval using Bayes point machines. *IEEE Trans. CSVT* **13**(1), 26–38 (2003)
- Cusano, C., Ciocca, G., Schettini, R.: Image annotation using SVM. In: Proceedings of Internet Imaging, vol. IV SPIE (2004)
- Duygulu, P., Barnard, K.: Object recognition as machine translation: learning a lexicon for a fixed image vocabulary. In: Proceedings of the 7th European Conference on Computer Vision, vol. 4, pp. 97–112 (2002)
- Fan, X., Xie, X., Li, Z., Li, M., Ma, W.Y.: Photo-to-Search: using multimodal queries to search the web from mobile devices. In: Proceedings of the 7th ACM SIGMM Workshop on MIR (2005)
- Feng, S.L., Manmatha, R., Lavrenko, V.: Multiple bernoulli relevance models for image and video annotation. In: Proceedings of CVPR (2004)
- Ferhatsmanoglu, H., Tuncel, E., Agrawal, D., Abbadi, A.E.: Approximate nearest neighbor searching in multimedia databases. In Proceedings of the 17th IEEE International Conference on Data Engineering, Heidelberg, pp. 503–511 (2001)
- Gao, Y., Fan, J., Luo, H., Xue, X., Jain, R.: Automatic image annotation by incorporating feature hierarchy and boosting to scale up SVM classifiers. In: Proceedings of ACM Multimedia, Santa Barbara (2006)
- Jeon, J., Lavrenko, V., Manmatha, R.: Automatic image annotation and retrieval using cross-media relevance models. In: Proceedings of ACM SIGIR (2003)
- Jeon, J., Manmatha, R.: Automatic image annotation of news images with large vocabularies and low quality training data. In: Proceedings of ACM Multimedia (2004)
- Jin, Y., Khan, L., Wang, L., Awad, M.: Image annotations by combining multiple Evidence & Wordnet. In: Proceedings of ACM Multimedia, Singapore (2005)
- Lavrenko, V., Manmatha, R., Jeon, J.: A model for learning the semantics of pictures. In: Proceedings of the 17th Annual Conference on Neural Information Processing Systems (2003)
- Li, J., Wang, J.Z.: Automatic linguistic indexing of pictures by a statistical modeling approach. *IEEE Trans. Pattern Anal. Mach. Intell.* **25**(9), 1075–1088 (2003)
- Li, J., Wang, J.Z.: Real-time computerized annotation of pictures. In: Proceedings of the 14th Annual ACM international Conference on Multimedia, Santa Barbara (2006)
- Li, X., Chen, L., Zhang, L., Lin, F., Ma, W.Y.: Image annotation by large-scale content-based image retrieval. In: Proceedings of ACM Multimedia, Santa Barbara (2006)
- Miller, G.A.: WordNet: a lexical database for English. *Commun. ACM* **38**(11), 39–41 (1995)
- Mori, Y., Takahashi, H., Oka, R.: Image-to-word transformation based on dividing and vector quantizing images with words. In: MISRM'99 (1999)
- Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: bringing order to the Web, technical report. Stanford University, Stanford (1998)
- Robertson, S.E., Walker, S.: Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. In: Proceedings of SIGIR, pp. 345–354. Springer, Heidelberg (1994)
- Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern Anal. Mach. Intell.* **22**(12), 1349–1380 (2000)
- Steinbach, M., Karypis, G., Kumar, V.: A comparison of document clustering techniques. In: KDD Workshop on Text Mining (2000)
- Wang, C., Jing, F., Zhang, L., Zhang, H.: Content-based image annotation refinement. In: Proceedings of the 2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR. IEEE Computer Society, Minneapolis (2007)

29. Wang, C., Jing, F., Zhang, L., Zhang, H.J.: Image annotation refinement using random walk with restarts. In: *Proceedings of ACM Multimedia* (2006)
30. Wang, C., Jing, F., Zhang, L., Zhang, H.J.: Scalable search-based image annotation of personal images. In: *ACM SIGMM International Workshop on Multimedia Information Retrieval*, Santa Barbara (2006)
31. Wang, X.J., Zhang, L., Jing, F., Ma, W.Y.: AnnoSearch: image auto-annotation by search. In: *International Conference on Computer Vision and Pattern Recognition*, New York (2006)
32. Xu, J., Croft, W.B.: Querying expansion using local and global document analysis. In: *Proceedings of the 19th International Conference on Research and Development in Information Retrieval* (1996)
33. Yang, C., Dong, M., Hua, J.: Image annotation using asymmetrical support vector machine-based multiple-instance learning. In: *Proceedings of CVPR* (2006)
34. Yeh, T., Tollmar, K., Darrell, T.: Searching the Web with mobile images for location recognition. In: *International Conference on Computer Vision and Pattern Recognition* (2004)
35. Zeng, H., He, Q., Chen, Z., Ma, W., Ma, J.: Learning to cluster web search results. In: *Proceedings of SIGIR*, New York, pp. 210–217 (2004)
36. Zhang, L., Chen, L., Jing, F., Deng, K.F., Ma, W.Y.: EnjoyPhoto—a vertical image search engine for enjoying high-quality photos. In: *Proceedings of ACM multimedia* (2006)
37. Zhang, L., Hu, Y., Li, M., Ma, W., Zhang, H.: Efficient propagation for face annotation in family albums. In: *Proceedings of ACM multimedia*, New York, pp. 716–723 (2004)