Folksonomy-Based Term Extraction for Word Cloud Generation

DAVID CARMEL, EREL UZIEL, IDO GUY, YOSI MASS, and HAGGAI ROITMAN,
IBM Research - Haifa Lab

In this work we study the task of term extraction for word cloud generation in sparsely tagged domains, in which manual tags are scarce. We present a folksonomy-based term extraction method, called *tag-boost*, which boosts terms that are frequently used by the public to tag content. Our experiments with tag-boost based term extraction over different domains demonstrate tremendous improvement in word cloud quality, as reflected by the agreement between manual tags of the testing items and the cloud's terms extracted from the items' content. Moreover, our results demonstrate the high robustness of this approach, as compared to alternative cloud generation methods that exhibit a high sensitivity to data sparseness. Additionally, we show that tag-boost can be effectively applied even in nontagged domains, by using an external rich folksonomy borrowed from a well-tagged domain.

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

General Terms: Algorithms, Experimentation

Additional Key Words and Phrases: Tag-cloud generation, keyword extraction, tag-boost

ACM Reference Format:
Carmel D., Uziel E., Guy I., Mass Y., and Roitman H. 2012. Folksonomy-based term extraction for word cloud generation. ACM Trans. Intell. Syst. Technol. 3, 4, Article 60 (September 2012), 20 pages.
DOI = 10.1145/2337542.2337545 http://doi.acm.org/10.1145/2337542.2337545

1. INTRODUCTION

The emergence of social media applications in recent years has encouraged people to be actively involved in content creation and classification. Users share content such as photos and videos and annotate public content through comments, ratings, and recommendations. Collaborative bookmarking systems such as Delicious[1] and Dogear for the enterprise [Millen et al. 2006], as well as many other content sharing sites (such as Flickr[2], Last.fm[3], and YouTube[4]), encourage users to tag available content for their own use as well as for the public. Sites providing blogging services, for example, encourage their bloggers to tag their own content to improve the disclosure and findability of their posts.

[1]delicious.com
[2]www.flickr.com
[3]www.last.fm
[4]www.youtube.com

Portions of the work reported here were previously presented in a short conference paper [Carmel et al. 2011].
Authors' address: D. Carmel, E. Uziel, I. Guy, Y. Mass, H. Roitman, IBM Research - Haifa Lab, Haifa 31905, Israel, email: carmel@il.ibm.com.

DOI 10.1145/2337542.2337545 http://doi.acm.org/10.1145/2337542.2337545

Tags annotated by users form a taxonomy of the tagged items, commonly referred to as a *folksonomy*. The value of a folksonomy is derived from the unique vocabulary and explicit meanings added by the people who tag the items. For example, new personal aspects may derive from a person's inferred understanding of an item's value. Folksonomies have been found to be extremely useful for many information retrieval (IR) applications, including tag cloud representation of social media items [Hassan-Montero and Herrero-Solana 2006; Kaser and Lemire 2007], query refinement [Wang and Davison 2008], and search and browsing enhancement [Bischoff et al. 2008; Heymann et al. 2008a; Hotho et al. 2006; Xu et al. 2008].

The quality of a folksonomy heavily depends on users' engagement through tagging activity. Users are expected to provide accurate tags that truthfully represent the item in order to bring value to the folksonomy-based applications. In contrast, tagging can be detrimental if nonappropriate tags are associated with the items, either on purpose (e.g., by spamming [Heymann et al. 2007]), or by using terminology that has no relevance beyond the tagger's limited context [Carmel et al. 2009a]. However, most user-provided tags are indeed related to the annotated content. Heymann et al. [2008a] analyzed the Delicious folksonomy and found that tags appear in over 50% of the pages with which they are associated, and in only 20% of the cases they do not appear in the page text, backlink page text, or forward link page text. Similarly, Bischoff et al. [2008] showed that a large number of tags are accurate and reliable over several domains. For example, more than 73% of the tags used to annotate items in a music domain also appear in online music reviews.

Folksonomy-based applications are mostly useful in well-tagged domains. While many such domains with extensively tagged resources do exist (e.g., Web pages in Delicious or photos in Flickr), many other domains are sparsely tagged. This occurs either because that site's tagging policy does not encourage users to tag content of other providers (e.g., blog services typically allow users to tag only their own blog entries), or because the content is mostly exposed through other vendors, such as search engines or feed readers that often do not expose the tagging functionality to the users.

1.1. Word Cloud Generation

A word cloud is a visual depiction, typically used to provide a visual summary or a semantic view of an item or a cluster of items that have something in common (e.g., the search results for a specific query). Terms in the cloud are normally listed alphabetically and the importance of a term is represented using font size or color. Thus, users can easily find a term either alphabetically or by importance. A term in the cloud usually links to all items that are associated with it.

Clouds are usually generated using the tags assigned to an item (referred to as *tag clouds*), or important terms extracted from the item's description (referred to as *word clouds*). Clearly, meaningful, high-quality tag clouds can be generated in well-tagged domains where the resources are widely tagged. An item can be successfully represented by a tag cloud that is based on its own tags, or on tags associated with similar items. However, when manual (user-provided) tags are not available, feature selection techniques can be used to extract meaningful terms from the item's content or from other textual resources that are related to the item, such as anchor text or the item's metadata [Heymann et al. 2008b]. These extracted terms can be used to create a word cloud [Brooks and Montanez 2006; Liu et al. 2009; Lu et al. 2009a; Song et al. 2008]. Word clouds, however, are usually inferior to tag clouds, since significant terms, from a statistical perspective, do not necessarily serve as good labels for the content from which they were extracted [Carmel et al. 2009b].

Fig. 1. Word cloud of tweets related to *Apple*, for August 28, 2011, generated by the TweetCloud application[5].

Recently, word clouds have been popularized by leading social media sites such as Twitter and Facebook, as an alternative to tag clouds due to the sparseness of manual tags in these domains. Figure 1 presents a word cloud of tweets related to Apple, generated by the TweetCloud application[5] in August 28, 2011, after the noticeable announcement of the resignation of Steve Jobs, Apple's famous chairman.

1.2. Tag-Boost for Word Cloud Generation

Our study focuses on term extraction methods for word cloud generation and is specifically tailored for sparsely tagged domains where manual tags are scarce.[6] We propose a novel approach that enhances term selection methods for word cloud generation. Our approach, termed *tag-boost*, promotes terms in the item's description that the public frequently uses as tags. Our method selects terms from the item description, according to statistical selection criteria, and according to their relative frequency in the tag-based folksonomy. Thus, terms that people frequently use to tag content are boosted, in comparison to terms that are not frequently used as tags. The folksonomy used for tag-boosting can be imported from any external domain and is not limited to the domain-based folksonomy we address, which might be poor or noisy. Thus, this method can be applied to any content, including nontagged or sparsely tagged domains. However, as we note in the following sections, the efficiency of the tag boost approach depends on the quality of the folksonomy used and its suitability for the content being represented.

The tag-boost approach for term extraction is motivated by traditional *named entity recognition* techniques that boost important terms according to their linguistic features (part-of-speech, for example, [Wu et al. 2002]). In our case, terms that are frequently used as tags, even in other domains, are believed to serve as better tags than (statistically) significant terms that have never been used to tag content by the public.

For example, looking at the terms appearing in the "Apple" word cloud in Figure 1, many terms, which represent different aspects of the concept, enable the users to drill down and focus on one such aspect (e.g., "iPad," "iPhone," "iTune"), or different aspects

[5]http://tweetcloud.com

[6]Domains could be sparsely tagged either when only few items have been tagged or that there are few tags in the folksonomy. In this work we measure domain sparseness according to the number of nontagged items in the collection.

such as "juice," "drinking," etc. However, quite a few terms (e.g., "best," "click," "email," "join") do not provide much value to the user. These terms are significant from a statistical perspective, yet they do not contribute to the essence of the word cloud, given that such terms would rarely be used to describe content. The tag-boost approach can identify and filter out such terms, giving preference to terms that are suitable for tagging, and hence are more likely to be useful.

In order to evaluate the quality of a generated word cloud, we used a benchmark of well-tagged items for testing. Given a testing item associated with manual tags, the quality of the generated word cloud can be estimated by measuring how many of those manual tags are retrieved, as well as their relative ranking in the ranked list of terms. We evaluated the word cloud quality for an item, or for a cluster of items such as the set of results retrieved for a specific query, or the set of items tagged by an individual person. Our results clearly show that our tag-boost technique significantly improves term extraction methods for word cloud generation, and significantly outperforms alternative methods in sparsely tagged domains. Additionally, we show that tag-boost can be effectively applied even in nontagged domains, by using an external rich folksonomy borrowed from a well-tagged domain.

The main contributions of our work are as follow:

— tag-boost: a novel, effective, and efficient technique for enhancing term extraction for word cloud generation, based on boosting terms that are frequently used by the public to tag content;
— an automatic evaluation methodology (without human intervention) that measures the quality of a word cloud according to the agreement between manual tags of the testing items and the cloud's terms extracted from the items' content;
— Extensive experimentation over various domains that demonstrates the valuable contribution of using tag-boost to extract terms for word cloud generation.

The rest of this article is organized as follows. Section 2 addresses related work on word cloud generation and tag recommendation. Section 3 details the tag-boost enhancement approach and how it can be used to enhance term extraction from an item's description. Section 4 describes our evaluation methodology, the experiments we conducted in both enterprise and nonenterprise domains, and the results obtained. Finally, Section 5 summarizes and suggests directions for further research.

2. RELATED WORK

In this section we review related work on tag cloud generation methods that are based on manual tags associated with items, and word cloud generation methods that are based on internal terms extracted from the items' content.

2.1. Usability

Word cloud usability literature is mostly focused on evaluation methodologies that estimate different cloud visualization methods for the user [Halvey and Keane 2007; Hassan-Montero and Herrero-Solana 2006; Kaser and Lemire 2007; Rivadeneira et al. 2007; Sinclair and Cardew-Hall 2008]. The terms in the cloud are displayed according to their relevance score using various font sizes, colors schemes, etc. [Kaser and Lemire 2007]. This results in a higher level of content representation [Hassan-Montero and Herrero-Solana 2006].

Venetis et al. [2011] propose several evaluation criteria for tag cloud quality. They focused on tag cloud representation of a ranked list of items such as the search results for a query. For example, the coverage metric measures which part of the set

is accessible from the cloud, that is, how many items are covered by at least one of the terms in the cloud. Another metric measures the overlap of the cloud, that is, the average level of intersection between two sets of items that are associated with two different terms in the cloud. These metrics are focused on evaluating the selection(and ranking) of a subset of tags from a fixed set of associated manual tags.

As opposed to generating a cloud from manual tags, the word cloud generation process is based on term extraction from the items' content. Extracting terms to be used as a word cloud for a given set of documents is strongly related to *cluster labeling*. A common approach to cluster labeling is to extract important terms from the cluster's content using statistical extraction methods [Manning et al. 2008], or alternatively, from external sources such as Wikipedia [Carmel et al. 2009b]. Our work focuses on the term extraction process; therefore, in our experiments, we focused on evaluating different term extraction methods. Our evaluation is based on measuring the agreement between the manual tags of the testing items and the cloud's terms we extracted from the items' content.

2.2. Tag Recommendation

The requirement for a high-quality folksonomy has led to the development of many tag recommendation methods for suggesting appropriate tags in social media applications [Jäschke et al. 2008; Krestel et al. 2009; Lu et al. 2009b; Mishne 2006; Rendle et al. 2009; Song et al. 2008; Symeonidis et al. 2008]. Tag recommendation has much in common with tag cloud generation; both generate a ranked list of tags from the same resources. However, several discrepancies exist between the two tasks. Tag recommendation is evaluated according to user satisfaction, as measured by the tags actually selected for annotation from the list of recommended tags. Therefore, many popular tag recommendation approaches recommend tags that have already been assigned to the item, since those tags are more likely to be selected for annotation. In contrast, tag cloud generation is evaluated according to the quality of representation [Venetis et al. 2011], independently of the tags' likelihood of being selected by users.

State-of-the-art tag recommendation methods can be roughly classified into three main types: usage-based, similarity-based, and content-based. Usage-based tag recommendation methods usually work by recommending the most popular tags already been used. In this manner, Sigurbjörnsson and van Zwol [Börkur and van Zwol 2008] considered terms that frequently co-occurred with tags previously given by the user. Xu et al. [2006] proposed a tag recommendation method aimed at providing tags with high coverage (i.e., that have a minimum overlap of concepts among the suggested tags), high popularity (i.e., many users use them), and low discovery effort (i.e., cooccur with other suggested tags).

The similarity-based tag recommendation approach applies item-based collaborative filtering techniques to recommend tags associated with similar items [Carmel et al. 2009a; Chirita et al. 2007; Krestel et al. 2009; Lu et al. 2009b; Mishne 2006; Rendle et al. 2009; Symeonidis et al. 2008]. Thus, a nontagged item can be recommended with suitable tags, given that several similar items have already been tagged appropriately. Based on this line of reasoning, Mishne's AutoTag system annotated blogs using tags that were extracted from similar blogs [Mishne 2006]. Chirita et al. [2007] suggested implementing a personalized tag recommendation system by looking at similar documents that were located on the user's desktop. Work by Krestel et al. [2009] and Si and Sun [2009] learned a topic-tag distribution from a training set of well-annotated documents, for which new documents can be mapped and tagged based on topic similarity. Lu et al. [2009b] further recommended tags based on a combination of content-based similarity and tag-based similarity. In Section 4 we experiment

with several variants of collaborative filtering for tag cloud generation as a baseline, to compare with our term extraction approach.

Content-based tag recommendation methods, which are the most relevant to our work, extract keywords from the text of items [Brooks and Montanez 2006; Givon and Lavrenko 2009; Heymann et al. 2008b; Song et al. 2008; Zhang et al. 2009]. Brooks and Montanez Brooks and Montanez [2006] analyzed the effectiveness of extracted terms for classifying blog entries. Their approach recommends terms with highly *tf-idf* scores as tags. Heymann et al. [2008b] predicted tags based on page text, anchor text, surrounding hosts, and other tags applied to the Web page. They found that tag-based association rules can produce very high-precision predictions and also provide deep insight into the relationships among tags.

2.3. Mutual Relations between Content and Tags

Our tag-boost-based term extraction approach utilizes relations between manual tags and the items' content. Several content-based tag recommendation methods have also utilized relationship between content and tags for tag recommendation. Liu et al. [2009] developed an approach to recommend tags for weblogs. They separate the recommended tags into keywords that appear in the content of the weblogs and external tags that do not appear in the content. Their method recommends external tags by mining the cooccurrences of tags and keywords in a training dataset and retrieving the tags that have the highest cooccurrence with the extracted keywords. Lee and Chun [2007] use a neural network for recommending tags to blog entries. In their approach, the network is trained by positive examples to find mutual relationships between internal keywords and tags, and then used these results to recommend tags for new blog entries according to their content. Zhang et al. [2009] enhance document representation using available document tags. They measure the mutual reinforcement relationship between the document's keywords and its tags to determine the best sets of representative keywords and tags for that document. Song et al. [2008] propose a two-stage classification framework that also utilizes tag-keyword associations and provides a highly-automated solution for real-time tag recommendation. Ramage et al. [2009] studied *Labeled LDA*, a topic model that defines a one-to-one correspondence between the content's latent topics and the user tags. This allows *Labeled LDA* to directly learn word-tag correspondences that can be used by several applications, including term extraction, for word cloud generation. Givon and Lavrenko [2009] annotated books with tags, using an estimation of the joint tag and keyword probability distribution. The joint distribution provides an estimation that a book will be annotated with certain tags, given a background collection of annotated books. Their method retrieves tags that maximize the joint distribution and recommends them for annotation. This approach has much in common with collaborative filtering-based tag recommendation, as it can be shown that recommended tags to a given book are those assigned to similar books.

All of these works appear to learn how to tag an unseen item from existing manually tagged items. Therefore, their effectiveness is strictly affected by the amount and quality of the training data, i.e., the existence of sufficient tagged items and the richness of the folksonomy. Apparently, the effectiveness of such methods in nontagged, or sparsely tagged, domains is expected to drop due the lack of training data. This is especially true for nontagged items, or for those with similar items that are sparsely tagged. In the following we will show that the effectiveness of CF based methods is highly sensitive to data sparseness. Tag-boost, on the other hand, is found to be highly robust to data sparseness and therefore provides an effective alternative method for term extraction in such domains.

3. WORD CLOUD GENERATION

Word cloud generation techniques are based on extracting internal terms from the items' content, and then generating a cloud using these terms. In the following, we describe the cloud generation method we used and the term extraction method we propose.

3.1. Cloud Generation

In this section, we describe the technique we used for cloud generation, originally described in work by Amitay et al. [2009]. This method proved to perform properly when applied on manual tags of well-tagged items and therefore provides a basis for our work.

Most cloud generation techniques that are based on existing manual tags rank the tags associated with the items in order to retrieve the top-ranked tags for a cloud. In many cases, tags are ranked according to the number of times they were used to tag an item. Let e be an item tagged by k manual tags, $(t_1, \ldots, t_k)$, then a tag cloud (a ranked list of representative tags) can be generated by ranking the tags according to the tag score, $s(t, e)$, defined by the following weighting formula:

$$s(t, e) = tf(t, e) \cdot ief(t);$$

$tf(t, e) = \log(freq(t, e) + 1)$ monotonically increases with $freq(t, e)$, the number of times e was tagged by t, and $ief(t) = \log \frac{N}{N_t}$ is the *inverse entity frequency* of the tag t, where N is the total number of items, and N_t is the number of items tagged by t. Such a weighting scheme is analogous to the popular vector-space *tf-idf* weighting approach. Thus, a highly used tag (high tf), which is assigned to only a few items in the collection (high ief), is ranked higher in the cloud.

A cloud for a cluster of tagged items is generated by ranking the tags assigned to all items in the cluster. Let $S = \{e_1 \ldots, e_n\}$ be a list of tagged items, ranked according to an arbitrary score function, $Score(e)$. Each item e_i is associated with a list of tags $(t^i_1, \ldots, t^i_{k_i})$, $k_i \geq 0$ (for a nontagged item e_j, $k_j = 0$). The aggregated score of a tag, with respect to S, is determined as follows:

$$s(t, S) = \sum_{e \in S} Score(e) \cdot s(t, e). \quad (1)$$

Thus, tags that are frequently assigned to many highly scored items are ranked higher in the tag-cloud.

The same formula can be used for word cloud generation. Given a set of items, each associated with a set of extracted terms, then a word cloud can be generated using Equation (1), where the term score per item, $s(t, e)$, is determined by the the term extraction method.

3.2. Term Extraction

Sparsely tagged domains may not contain enough manual tags to use for a cloud representation. However, important terms can be extracted from an item's related content to be used for word cloud generation. In this work we assume that each item is associated with a textual description, from which significant terms can be extracted.

Term extraction is strongly related to feature selection, which is the process of selecting a set of terms for text representation. Common approaches for feature selection evaluate terms according to their ability to distinguish the given text from the text of the whole collection. In our case, we aim to find a set of terms that best distinguishes an item, or a cluster of items, from the entire collection.

We experimented with four fundamental extraction techniques [Manning et al. 2008]:

(1) *tf-idf*, which selects terms from the item textual description with maximum *tf-idf* weights;
(2) *Mutual Information* (MI), which measures how much information the presence/absence of a term contributes to the item description;
(3) χ^2, which measures the statistical independence of the occurrence of the term in the item description and its occurrence in the collection; and
(4) *Kullback-Leibler divergence* (KL), which looks for a set of terms that maximize the KL divergence between the language model of the item's content and the language model of the entire collection [Berger and Lafferty 1999].

If the items were annotated with only a few tags, which were not sufficient for cloud representation, then the tags could be supplemented with extracted terms. The tags and terms combination policy we applied is based on placing the manual tags at the top of the ranked list, and then completing the list with the top scoring extracted terms. This policy is based on the assumption that manual tags are superior to internal extracted terms.

We can also consider more sophisticated combination policies between tags and terms, like measuring the mutual relationships between them [Givon and Lavrenko 2009; Song et al. 2008; Zhang et al. 2009]. However, existing methods assume the existence of sufficient associated tags. Advanced integration approaches of terms and tags in sparsely tagged domains have not yet been studied, to the best of our knowledge, and we suggest this as a direction for future work.

3.3. Tag-Boost Enhancement

Internal extracted terms are very useful for IR applications such as text clustering and categorization [Manning et al. 2008]. However, in general, extracted terms are not always optimal for tagging. Carmel et al. [2009b] showed that in many cases, even when manually associated tags appear in the text, pure statistical methods have difficulty in identifying them as good descriptors. An important term, as determined by common statistical criteria, is not always considered a good label by humans.

Therefore, our main hypothesis is that terms considered as good labels by humans have specific characteristics that are not always exposed by standard statistical extraction approaches. We attempt to measure the likelihood of a term being considered by humans as a good tag.

Let C be a collection of items, and let w be a term. Let $C_w \subseteq C$, and $T_w \subseteq C$ be the set of items containing w, and the set of items tagged by w, respectively. We mark $w \in e$, if term w appears in e's description. We mark $tag(w, e)$, when w is used to tag e. We mark $tag(w)$ if w is used to tag any item in C.

The probability of term w to tag an item e when appearing in its description can be approximated by *maximum likelihood estimation* (MLE):

$$Pr_C(tag(w, e)|w \in e) \stackrel{def}{=} \frac{|C_w \cap T_w|}{|C_w|}.$$

$Pr_C(tag(w, e)|w \in e)$ estimates the probability that a term w found in the item's description will also be used to tag that item. Thus, terms with high values should be biased for word cloud generation.

The second measure we apply approximates the global likelihood of a term to be used as a tag. Our approximation is based on the assumption that a term w, used to

tag many items in the collection, is more likely to be used as a tag for any item in the collection, compared to terms that are rarely used as tags.

$$Pr_C(tag(w)) \stackrel{def}{=} \frac{|T_w|}{|C|}.$$

When estimating probabilities based on a limited amount of data, smoothing is a fundamental approach to adjust the maximum likelihood estimator and thereby correct the inaccuracy due to data sparseness [Zhai and Lafferty 2001]. For example, a term that has never been used to tag the items to which it belongs ($Pr_C(tag(w,e)|w \in e) = 0$), should still be considered a good candidate for tagging when it is frequently used to tag other items in the collection (with high $Pr_C(tag(w))$). Therefore, the tag-boost probability we apply is based on the *Jelinek-Mercer* smoothing of the two estimators:

$$Pr_C^{smooth}(tag(w,e)|w \in e) \stackrel{def}{=} \lambda \cdot Pr_C(tag(w,e)|w \in e) + (1-\lambda) \cdot Pr_C(tag(w)). \quad (2)$$

The smoothing coefficient λ, can be optimally tuned for each individual collection. In our experiments, described in Section 4, we set λ to 0.9 as it seems to perform well for all collections with which we experimented.

Finally, we boosted each term extracted from the item's description, by any statistical term extraction technique, by multiplying its (statistical) weight $s(w,e)$, by the estimated tag-boost probability. We then selected terms with the maximum boosted score for word cloud representation:

$$s_{boost}(w,e) = s(w,e) \cdot Pr_C^{smooth}(tag(w,e)|w \in e). \quad (3)$$

When both estimators of term w are zero, that is, w is never used to tag an item, then its boosted score is zeroed, and it will not be selected by the tag-boost approach, regardless of its statistical score. This observation is especially important from a practical point of view, as only terms with positive tag-boost probability ($Pr_C(tag(w) > 0$) should be considered by the term extraction method. In this way, the efficiency of term extraction is significantly improved as many terms in the text can be filtered out in advance, including short phrases for which statistics extraction is usually an expensive task. This strict approach might be relaxed by an alternative flexible combination of the statistical score with tag-boost score, for example by a linear combination; however, all terms should be analyzed in this case by the extraction process.

We also note that both estimators of the term-tagging characteristics can be inferred from any collection of tagged items. Thus, as we will show, it is possible to estimate those probabilities from a well-tagged collection. These probabilities can then be used to boost terms in sparsely tagged collections that suffer from insufficient statistics.

3.4. Collaborative Filtering

Collaborative filtering (CF) is a popular technique for recommending items that are related to similar users. For example, an online book store can recommend books to its users, based on books bought by others with similar buying patterns.

In our case, item-based CF can be used to enrich an item's cloud with tags used for annotating similar items [Carmel et al. 2009a; Chirita et al. 2007; Krestel et al. 2009; Lu et al. 2009b; Mishne 2006; Rendle et al. 2009; Symeonidis et al. 2008]. The principal idea is that a manual tag associated with an item is also expected to be suitable for a similar item. The main advantage of CF methods over term extraction-based methods is that external (manual) tags, even ones that are not used to tag the item, may be superior to internal terms. On the other hand, a CF-based approach is sensitive to

the way similarity is measured among items, and appropriate tags for a given item do not always fit its similar items. Moreover, CF assumes the existence of good tags for similar item, an assumption that often does not hold in sparsely tagged domains.

Nevertheless, CF methods are very effective recommendation techniques. We therefore applied them for cloud generation as a strong baseline approach. Given an item e, we first find a set of similar items, $S(e) = (e_1, \ldots e_n)$, scored according to their similarity to e, and then create a cloud for this set using Equation (1). In our experiments, described in Section 4, we measure similarity among items using the Lucene open source search engine[7]. That is, each item's description is indexed as a document by Lucene, and the most similar items are retrieved for a query based on the most significant terms extracted from the item's description. We experimented with different term extraction methods for similarity measurement.

The CF method can also be integrated with term extraction methods. If no tags are assigned to a similar item, or only a few are assigned, then the cloud of that item can be supplemented with internal terms extracted from the item description. Subsequently, the CF-based cloud of both tags and terms can be aggregated from all similar items.

4. EVALUATION

In this section, we summarize the experiments we conducted with term extraction methods for word cloud generation in sparsely tagged domains. We describe the evaluation methodology for cloud quality estimation, the datasets we used, the methods with which we experimented, and the results obtained.

4.1. Evaluation Methodology

The main assumption behind our evaluation approach is that manual tags assigned to an item are good labels for summarizing its main characteristics. Therefore, a set of well-tagged items can be used to evaluate word cloud quality by measuring the agreement between the manual tags and the extracted terms. Given a testing item associated with good manual tags, the quality of a word cloud generated for that item can be estimated by measuring the number of manual tags retrieved, as well as their relative position in the ranked list of extracted terms. The average precision (AP) measure estimates the quality of a ranked list of terms by measuring the number and rank of manual tags in the list. MAP, the mean AP over a set of testing items, is the main evaluation measure we used for this work. We also used P@10, the relative frequency of manual tags in the top ten retrieved terms, as a complementary measurement.

The assumption that manual tags are always relevant for representing an item does not always hold, as some of the tags are private, noisy, or incorrect [Carmel et al. 2009a; Heymann et al. 2007]. Moreover, the set of manual tags is incomplete and does not cover all good terms that can be used for labeling. Nevertheless, different generation methods are still comparable according to their ability to retrieve those manual tags, and to agree with the crowd on the right terms to use for item representation. Moreover, the evaluation is fully automatic; it can be applied to a very large set of testing items, and therefore it is fairly robust to different testing sets. The manual tags associated with the testing items are used for testing only and are hidden from the word cloud generation methods. A similar evaluation approach was recently suggested to measure the quality of personalized search [Xu et al. 2008].

The term extraction methods we experimented with include several statistical extraction methods from the items' descriptions, CF-based methods that retrieve manual tags of similar items, and hybrid methods that combine the two approaches.

[7]www.lucene.apache.org

Table I. Social Media Dataset

	#items	#unique tags	avg. tags/item	median tags/item
Delicious	144.5K	67K	12.8	11
CiteULike	235K	54K	3.5	3
Dogear	198K	47K	3.8	3
BlogCentral	119K	24K	2.7	1

To evaluate the word cloud generated for a cluster of items, we experimented with two types of item sets. The first one is a ranked list of results for a text query. A testing set of 100 queries was randomly selected from the query-log of the search system of our organization. For each testing query, a manual tag cloud was generated for the top-10 ranked items, based on their manual tags, using Equation (1).

Then, for each item, we extracted 30 terms and created a word cloud for the set using Equation (1). We measured the AP of the word cloud using the top-10 tags in the manual tag cloud, considered as relevant tags for that query[8].

The second type is a set of items tagged by a specific user. We selected all items tagged by a user, and generated a word cloud for this set by extracting 30 terms per item. This word cloud summarizes the user interests and preferences, and can be used for personalization tasks [Guy et al. 2010]. To test the quality of such a word cloud, we randomly selected 1000 users who were tagged by others, with at least 5 tags each, in an enterprise tagging application that allows employees to annotate each other with descriptive tags [Farrell et al. 2007]. These manual tags, assigned by others to describe the user's role, activities, and preferences, were used to measure the AP of the word cloud of the user's related items. Manual tags used by others to annotate a person were shown to be strongly correlated with his actual preferences [Guy et al. 2010].

In our final experiment, we examined whether an external folksonomy can be used for tag-boost-based term extraction. This is especially important for sparsely tagged domains, in which the local folksonomy is too poor or too noisy to provide sufficiently reliable statistics. Thus, a rich external folksonomy from a close domain may provide a suitable alternative. We used the Twitter collection, which has a poor and noisy folksonomy, to examine whether the richer Delicious or *CiteULike* folksonomies can be used effectively for tag-boost-based term extraction in this domain.

4.2. Data

We experimented with term extraction methods over four social media datasets, described in Table I.

Delicious T140. Delicious T140 is a dataset made up of 144.5K unique Web pages, all of them with their corresponding social tags retrieved from *Delicious* during June 2008.[9] This is a well-tagged dataset, as all documents (Web pages) are assigned at least one tag. The median number of tags per page is 11, and the average is 12.8[10]. As is true for many other social bookmarking systems [Hotho et al. 2006; Jäschke et al. 2008; Mishne 2006], the number of tags given to a document in this set is power-law distributed (N^k), with a power factor of $k = -0.09$ ($R^2 = 0.41$).

[8]We also experimented with NDCG@10 while considering the rank of the manual tags in the cloud as different relevance levels. We omit reporting this evaluation measure due to its high agreement with MAP results.

[9]http://nlp.uned.es/social-tagging/delicioust140/

[10]These numbers relate to the total number of tags assigned to an item. We do not report the statistics of unique tags assigned to an item as almost all tags are unique.

CiteUlike. CiteULike is an online bookmarking service that allows users to bookmark academic articles. For our experiments, we used a random sample of 235K documents, with a total of 209K tags. The median number of tags given to a page is 3 and the average is 3.5. The power law for tags given to a document is $k = -0.28$ ($R^2 = 0.9$).

Dogear. Dogear is an enterprise social bookmarking system [Millen et al. 2006], popularly used by thousands of employees, to organize their bookmarks of both intranet and Internet documents. The Dogear dataset contains 198K Web pages, bookmarked with a total of 743K tags. The median number of tags assigned to a page is 3 and the average is 3.8. The power law for tags to a document is $k = -0.35$ ($R^2 = 0.91$).

BlogCentral. BlogCentral is an enterprise blog service allowing employees to publish personal blogs within the intranet and to comment on other people blogs [Huh et al. 2007]. The dataset contains 119K blog posts, tagged with a total of 350K tags. The median number of tags given to a page is 1 and the average is 2.65. The relatively low average number of tags per item in this domain is due to the fact that in contrast to the other collections, only the author is allowed to tag his/her own blog posts. Therefore, we use the *BlogCentral* dataset as an example of sparsely tagged domains. The power law for tags to a document is $k = -0.27$ ($R^2 = 0.89$).

4.3. Word Cloud Generation for a Single Item

In our first experiment, we evaluated several term extraction methods for a single item, with and without tag-boost over the four domains. For each domain, we randomly selected 1000 items, each assigned with at least 5 unique tags for testing. We selected web pages from Delicious and Dogear, blog posts concatenated with their associated comments from BlogCentral, and the abstracts of scientific articles from CiteULike.

For each item, we generated a word cloud by extracting the most important terms from the item's content. We measured its quality by AP and P@10 according to the manual tags of that item. The (statistical) term extraction methods select the most informative terms from the item's content; we then boosted each term by the tag-boost probability, according to Equation (3). Table II represents the MAP and P@10 achieved by the different term extraction methods over the four domains.

The results reveal no significant difference in performance among the different statistical methods, in terms of the agreement between the generated word cloud and the manual tags assigned to an item, as measured by MAP and P@10. However, tag-boost improved that agreement significantly for all methods, particularly for Delicious (80% on average for MAP). The improvement was less impressive for the other domains but still significant (65% for BlogCentral, 23% for Dogear and 41% for CiteULike). In all domains, the improvement achieved by tag-boost was statistically significant (paired t-test $p < 0.001$). Interestingly, χ^2, which is the inferior method over all collections, outperformed all other methods (albeit insignificantly) when combined with tag-boost.

Since no significant difference in performance was measured among the term extraction methods, we focused the following experiments on the KL extraction method as a representative for statistical term extraction.

4.4. Word Cloud Generation for a Cluster of Items

Clouds are usually created for a set of items, rather than for an individual item, due to the sparseness of manual tags per item, even in well-tagged domains. In this section, we describe two experiments that evaluate term extraction methods for word cloud generation for a cluster of items.

In the first experiment, we measured the quality of a word cloud created for the result list of a query, by measuring the agreement between the word cloud generated

Table II.

Word cloud quality of several term extraction methods, with and without tag-boost, as measured by MAP and P@10.

Collection	Term Ext. Method	MAP		P@10	
		no-boost	tag-boost	no-boost	tag-boost
Delicious	*tf-idf*	0.16	0.25 (+56%)	0.25	0.35 (+40%)
	χ^2	0.11	**0.27** (+145%)	0.20	**0.37** (+87%)
	MI	0.16	0.26 (+62%)	0.25	0.36 (+44%)
	KL	0.16	0.25 (+56%)	0.25	0.36 (+44%)
CiteULike	*tf-idf*	0.17	0.21 (+24%)	0.16	0.20 (25%)
	χ^2	0.14	**0.24** (+71%)	0.15	**0.22** (+47%)
	MI	0.17	0.23 (+35%)	0.16	0.22 (+38%)
	KL	0.17	0.23 (+35%)	0.17	0.21 (+24%)
Dogear	*tf-idf*	0.19	0.21 (+11%)	0.19	0.20 (+5%)
	χ^2	0.16	**0.25** (+56%)	0.17	**0.23** (+35%)
	MI	0.19	0.23 (+21%)	0.19	0.21 (+11%)
	KL	0.19	0.22 (+16%)	0.19	0.21 (+11%)
BlogCentral	*tf-idf*	0.15	0.26 (+73%)	0.15	0.23 (+53%)
	χ^2	0.14	**0.27** (+93%)	0.14	**0.24** (+71%)
	MI	0.18	0.27 (+50%)	0.18	0.23 (+28%)
	KL	0.18	0.26 (+44%)	0.18	0.23 (+28%)

Table III.

Quality of word clouds generated for the result lists of queries, as measured by MAP and P@10.

Collection	Term Extr. method	MAP	P@10
Delicious	KL	0.14	0.41
	KL+TB	0.44 (+214%)	0.87 (+112%)
CiteULike	KL	0.13	0.40
	KL+TB	0.29 (+123%)	0.69 (+73%)
Dogear	KL	0.13	0.43
	KL+TB	0.23 (+77%)	0.60 (+40%)
BlogCentral	KL	0.15	0.42
	KL+TB	0.27 (+80%)	0.56 (+33%)

from extracted terms and the tag cloud generated from manual tags. For each domain, we ran 100 text queries, randomly selected from a query log, and created two clouds for the top-10 results per query.

We created the manual tag-cloud from the tags of retrieved items according to Equation (1). We created the term-based word cloud by the same formula, using the 30 terms per item extracted from its content. We used the top 10 tags of the manual tag cloud to measure the precision of the word cloud. Table III shows the quality of the word cloud, while using KL and tag-boost-based KL (KL+TB) for term extraction, over the four domains. Tag-boost significantly improved the word cloud quality (by more than 200% on Delicious!), to better agree with manual based tag clouds.

In the second experiment, we measured the quality of word clouds generated for a cluster of items tagged by a person in the Dogear and the BlogCentral domains. We

Table IV.

Quality of word clouds generated for clusters of items; a cluster is the set of items all tagged by the same person.

Collection	Term Extr. method	MAP	P@10
Dogear	KL	0.06	0.12
	KL+TB	0.12 (+100%)	0.23 (+92%)
	Manual Tags	0.15	0.30
BlogCentral	KL	0.15	0.30
	KL+TB	0.24 (+60%)	0.46 (+20%)
	Manual Tags	0.25	0.50

selected 1000 people for testing, each associated with at least 5 in-tags (terms used to tag that user by others).[11] Those in-tags were used to measure the quality of the word clouds, created from the 30 terms extracted from each item's content. This evaluation is highly reliable, since a person's in-tags, provided by other users, are independent of the items the user tagged and their corresponding content. Therefore, the higher the agreement of the user's word cloud with those in-tags, the higher its quality.

In addition, we created a manual cloud for each cluster, based on the manual tags assigned to the cluster's items, and measured its agreement with the given in-tags.

Table IV shows the quality of the different term extraction methods used for word cloud generation. The last row shows the quality of the manual-based tag cloud of those items, which is given for reference.

While the KL-based extraction method fails to agree with the users' in-tags, using tag-boost improves the quality by more than 100% on Dogear, and by more than 60% on BlogCentral, performing almost as well as manual tags. In this experiment, we also observed the dominance of all extraction methods in the BlogCentral domain compared to the Dogear domain. This is probably due to the closer agreement between in-tags and personal blog posts as both in-tags and blog content are highly correlated with the author's topics of interest, in contrast to user personal bookmarks in Dogear, which are much more diversified.

4.5. Collaborative Filtering

Internal terms extracted from the items' content provide an alternative to manual tags for word cloud generation, especially when manual tags are scarce. Alternatively, tag-recommendation based approaches may be used for recommending tags for word cloud representation. Collaborative filtering is a highly effective approach for recommending tags assigned to similar items, and actually many of the tag recommendation methods, covered in Section 2, are based on collaborative filtering approaches.

In this experiment, we studied CF-based methods compared to tag-boost-based term extraction methods for cloud generation. Given an item, the CF method extracts 10 terms from the item's content to represent the item as a query, then selects the 30 nearest-neighbors (the top-30 search results) and retrieves the manual tag cloud of those similar items. We experimented with two term-extraction methods for similarity measurement: one based on KL, marked by *CF(KL)*, and the other based on KL with tag-boost, marked by *CF(KL+TB)*.

Additionally, we experimented with a combination of CF and term extraction. After retrieving the 30 nearest neighbors of the item, we completed the set of tags for each

[11]We could not run this experiment on *Delicious* and *CiteULike* since we do not have the in-tags for users in these domains.

Table V.

Quality of the tag clouds generated by collaborative filtering variants, as measured by MAP and P@10; the word cloud quality of the term extraction method (KL+TB) is given for reference.

Collection	CF method.	MAP	P@10
Delicious	KL+TB	0.25	0.36
	CF(KL)	0.43	0.49
	CF(KL+TB)	0.37	0.44
	Combine(CF,KL)	0.42	0.48
CiteULike	KL+TB	0.23	0.21
	CF(KL)	0.29	0.24
	CF(KL+TB)	0.29	0.25
	Combine(CF,KL)	0.27	0.23
Dogear	KL+TB	0.22	0.21
	CF(KL)	0.23	0.21
	CF(KL+TB)	0.21	0.20
	Combine(CF,KL)	0.24	0.22
BlogCentral	KL+TB	0.28	0.25
	CF(KL)	0.22	0.20
	CF(KL+TB)	0.22	0.20
	Combine(CF,KL)	0.23	0.20

neighbor with its top-scored extracted terms. Thus, we assigned each neighbor with a ranked list of 30 terms. These terms are a combination of the item's manual tags at the top of the list, augmented with its most important terms that are ranked below the manual tags. The cloud was then constructed by aggregating the combined lists. We mark the method with *Combine(CF,KL)*.

The tag cloud quality of all methods was measured by MAP and P@10 over the four collections, using the same 1000 testing items used in the first experiment. Table V shows the performance of the CF-based methods over the four domains. The performance of the term-extraction method, based on KL plus tag-boost, is given for reference.

As expected, the CF-based methods perform very well on the well-tagged domains, and significantly outperforms the *KL+TB* method over Delicious and CiteULike. This is not surprising, as the nearest-neighbors in the Delicious collection are associated with many manual tags, and the term-extraction methods are limited on CiteULike due to the limited available content (only paper abstracts). However, on a sparsely-tagged domain such as BlogCentral, KL+TB outperforms CF. The effectiveness of the CF methods drops significantly in these domains, since similar items cannot provide sufficient tags for cloud generation. Augmenting the manual tags of similar items with extracted terms did not significantly improve the cloud quality over all collections. Furthermore, the effect of tag-boost on CF while being used for term extraction for similarity measurement was insignificant.

The fact that CF is highly effective in well-tagged domains, compared to its inferiority in sparsely tagged domains, raises the question: how rich should the folksonomy be to still be effective for CF? To answer this question, we measured the MAP of CF(KL+TB) and KL+TB, using only a fragment of the folksonomy. We diluted the folksonomy by randomly selecting a fraction of the documents from the collection and removing their associated tags.

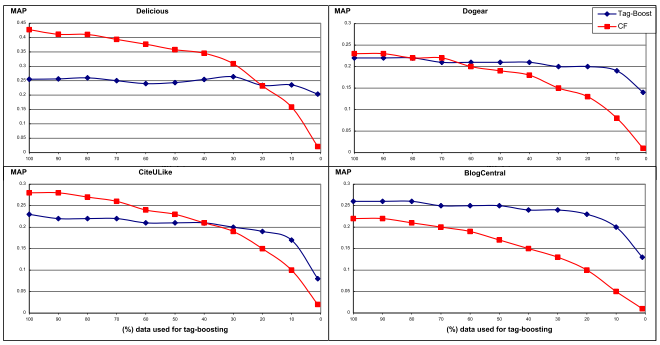

Fig. 2. Word cloud quality (MAP) as a function of data sparseness; applying only a fraction of the folksonomy for CF and for tag-boost-based term extraction.

Figure 2 shows the effect of data sparseness on CF(KL+TB) as well as the KL+TB effectiveness over the four domains. The results clearly show, over all domains, the high sensitivity of the CF method to data sparseness compared to the high robustness of tag-boosting. The drop in effectiveness of KL+TB is quite moderate in cases of data sparseness; the tag-boost effectiveness is significantly reduced only when 70% or more of the tags are dropped.

The results indicate that when data is becoming sparse, tag-boost outperforms CF. For the richest domain, Delicious, tag-boost's effectiveness becomes equal to that of CF only after 80% of the data is removed. For CiteULike, the intersection occurs when 60% of the data is removed. For Dogear, both CF and tag-boost are comparably effective until CF drops when about 50% of the data is removed. For the sparsest domain, BlogCentral, tag-boost outperforms CF even with the original data, and the performance gap increases as more data is being removed. These results consistently show the superiority of tag-boost over CF for sparsely tagged data.

4.6. Using External Folksonomy for Tag-Boosting

In the final experiment, we examined whether an external folksonomy can be used for word cloud generation. The data we used was taken from Twitter. Users can tag their tweets by hashtags; however, less than 30% of the tweets are tagged, and many of the hashtags are used for marking a thread of discussion rather than for labeling [Kwak et al. 2010]. Consequently, manual tags in Twitter are useless for cloud generation. Moreover, its poor folksonomy cannot even be used for tag-boosting. Therefore, Twitter is an excellent example for a sparsely tagged domain.

We used the Twitter domain to examine whether an external folksonomy can be effectively used to replace the useless folksonomy for tag-boosting. Existing hashtags were treated as regular terms while the Delicious and CiteUlike folksonomies were used for tag-boosting. In order to evaluate the word cloud quality in this case, we manually tagged a random sample of 1000 tweets, extracted from a collection of 100K tweets that we collected during one week in November 2010. Tweets were tagged according to their content while considering internal terms, hashtags, and urls (when they existed). The average number of manual tags assigned per tweet was 2.5 and about 20% of the tweets were not tagged due to difficulties in understanding their underlying topics. However, we believe that in general, the manual terms that we used to tag the tweets are more precise and complete than the tags used for evaluation in the other domains as the tagging process in this experiment is fully controlled.

Table VI.

Quality of the word clouds generated for Twitter items, using the Delicious (Del.) and the CiteULike (CUL) folksonomies for tag-boosting.

	Term Extr. Method	MAP	P@5
Single items	KL	0.39	0.26
	KL+TB(Del.)	0.54(+38%)	0.44(+69%)
	KL+TB(CUL)	0.47(+21%)	0.36(+38%)
Cluster items	KL	0.11	0.36
	KL+TB(Del.)	0.27(+145%)	0.62(+72%)
	KL+TB(CUL)	0.21(+91%)	0.58(+61%)

The word-cloud quality was measured in two cases. First, we extracted 5 terms from each tweet's content,[12] using KL with and without tag-boost, and measured the agreement between them with the tweet's manual tags. Second, we ran 100 randomly selected queries and measured the quality of the word cloud generated for the top-10 results for each query, based on these extracted terms. Table VI shows the effectiveness of tag-boost for the Twitter domain, first using the Delicious folksonomy, and then using the CiteULike folksonomy. Despite the fact that the two folksonomies were borrowed from different domains, both significantly contributed to term extraction, as both are rich enough to filter out noisy terms that are unsuitable to be used as tags. The Delicious folksonomy was superior to the CiteULike folksonomy, presumably since it is more general and better fits the Twitter domain.

5. SUMMARY

In this work we studied the task of term extraction for word cloud generation in sparsely tagged domains, in which manual tags are scarce. We presented the tag-boost enhancement approach for term-extraction methods, which boosts terms that are frequently used by the public to tag content. Our experiments with tag-boost enhancement, over several enterprise and nonenterprise domains, demonstrated a tremendous improvement in word cloud quality, as reflected by the agreement between the generated word clouds and the corresponding manual tags of the testing items. In addition, our experimental results demonstrated the high effectiveness of tag-boost for word cloud generation, especially in sparsely tagged domains, compared to the high sensitivity of alternative CF-based methods to data sparseness. Moreover, we showed that tag-boost can even be applied in nontagged domains, by using an external folksonomy borrowed from a well-tagged domain.

The high robustness of tag-boost based term extraction to data sparseness is probably due to its simplicity and to the global computation of the tag-boost score, which considers the term distribution over the whole folksonomy, in addition to its distribution within the item's content. Thus, even items with nonassigned tags can still benefit from an effective tag-boost-based term extraction. This is even true in nontagged domains, where an external folksonomy can provide a tag-boost score for internal terms. In contrast, alternative tag recommendation methods that consider only the item's own tags, or the tags assigned to similar items, are much more sensitive to data sparseness. In addition, such approaches are less efficient as they should analyze the external content for measuring item similarity, while tag-boost only analyzes the external folksonomy. In our experiments, we showed the high robustness of tag-boost

[12]Tweets are often short and contain less than 5 terms. In this case we extracted all the tweet's terms.

to data sparseness compared to CF based methods. In the future we intend to further explore the behavior of alternative tag-based term extraction approaches, and the potential of aggregating different methods, in sparsely tagged domains.

We found that in our experiments, CF was clearly shown to be a highly effective method for cloud generation in well-tagged domains, performing significantly better than term extraction methods. The existence of many well-tagged items in such a collection enables CF to provide a high-quality cloud that is superior to any word cloud that is based on internal terms. However, CF may still have difficulties in cloud generation for specific items in such domains, where similar items are not associated with sufficient relevant tags. For such items, the manual cloud can be enhanced with terms extracted by tag-boost. Our initial attempts to enhance CF with tag-boost were not productive. Future research may further examine how to integrate tag-boost into CF to enhance its performance in well-tagged domains.

Tag-boost enhancement for term extraction is effective, efficient, and easily applied by any extraction method that looks for internal terms for labeling or for text representation and summarization. Thus, it can be used by many applications that utilize term extraction, in addition to word cloud generation. For future work we intend to investigate the contribution of tag-boost-based term extraction for other applications. We also intend to explore mutual relations between tags and terms in sparsely tagged domains, in which manual tags (if they exist) and internal terms can be mixed to provide a better cloud representation. Additionally, we intend to investigate the types of terms that are boosted by the tag-boost methods, to gain a better understanding of the characteristics of important terms, from a statistical perspective, that are also considered by the public as appropriate terms for labeling.

REFERENCES

AMITAY, E., CARMEL, D., HAR'EL, N., OFEK-KOIFMAN, S., SOFFER, A., YOGEV, S., AND GOLBANDI, N. 2009. Social search and discovery using a unified approach. In *Proceedings of the ACM Conference on Hypertext and Hypermedia*. ACM, 199–208.

BERGER, A. AND LAFFERTY, J. 1999. Information retrieval as statistical translation. In *Proceedings of the Annual Interntaional ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 222–229.

BISCHOFF, K., FIRAN, C. S., NEJDL, W., AND PAIU, R. 2008. Can all tags be used for search? In *Proceeding of the International Conference on Information and Knowledge Management*. ACM, 193–202.

BÖRKUR, S. AND VAN ZWOL, R. 2008. Flickr tag recommendation based on collective knowledge. In *Proceedings of the International World Wide Web Conference*. ACM, 327–336.

BROOKS, C. H. AND MONTANEZ, N. 2006. Improved annotation of the blogosphere via autotagging and hierarchical clustering. In *Proceedings of the International World Wide Web Conference*. ACM, 625–632.

CARMEL, D., ROITMAN, H., AND YOM-TOV, E. 2009a. Who tags the tags?: A framework for bookmark weighting. In *Proceedings of the International Conference on Information and Knowledge Management*. ACM, 1577–1580.

CARMEL, D., ROITMAN, H., AND ZWERDLING, N. 2009b. Enhancing cluster labeling using wikipedia. In *Proceedings of the Annual Interntaional ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 139–146.

CARMEL, D., UZIEL, E., GUY, I., MASS, Y., AND ROITMAN, H. 2011. Folksonomy-based term extraction for word cloud generation. In *Proceedings of the International Conference on Information and Knowledge Management*. ACM, 2437–2440.

CHIRITA, P. A., COSTACHE, S., NEJDL, W., AND HANDSCHUH, S. 2007. P-tag: large scale automatic generation of personalized annotation tags for the web. In *Proceedings of the International World Wide Web Conference*. ACM, 845–854.

FARRELL, S., LAU, T., NUSSER, S., WILCOX, E., AND MULLER, M. 2007. Socially augmenting employee profiles with people-tagging. In *Proceedings of the Annual ACM Symposium on User Interface Software and Technology*. ACM, 91–100.

GIVON, S. AND LAVRENKO, V. 2009. Large scale book annotation with social tags. In *Third International AAAI Conference on Weblogs and Social Media*.

GUY, I., ZWERDLING, N., RONEN, I., CARMEL, D., AND UZIEL, E. 2010. Social media recommendation based on people and tags. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 194–201.

HALVEY, M. J. AND KEANE, M. T. 2007. An assessment of tag presentation techniques. In *Proceedings of the International World Wide Web Conference*. ACM, 1313–1314.

HASSAN-MONTERO, Y. AND HERRERO-SOLANA, V. 2006. Improving tag-clouds as visual information retrieval interfaces. In *Proceedings of the International Conference on Multidisciplinary Information Sciences and Technologies*.

HEYMANN, P., KOUTRIKA, G., AND GARCIA-MOLINA, H. 2007. Fighting spam on social web sites: A survey of approaches and future challenges. *IEEE Internet Computing 11,* 6, 36–45.

HEYMANN, P., KOUTRIKA, G., AND GARCIA-MOLINA, H. 2008a. Can social bookmarking improve web search? In *Proceedings of the ACM International Conference on Web Search and Data Mining*. ACM, 195–206.

HEYMANN, P., RAMAGE, D., AND GARCIA-MOLINA, H. 2008b. Social tag prediction. In *Proceedings of the Annual Interntaional ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 531–538.

HOTHO, A., JÄSCHKE, R., SCHMITZ, C., AND STUMME, G. 2006. Information retrieval in folksonomies: Search and ranking. In *Proceedings of the 3rd European Semantic Web Conference (ESWC'06)*. 411–426.

HUH, J., JONES, L., ERICKSON, T., KELLOGG, W. A., BELLAMY, R. K. E., AND THOMAS, J. C. 2007. Blogcentral: The role of internal blogs at work. In *Proceedings of the Conference on Human Factors in Computing Systems*. ACM, 2447–2452.

JÄSCHKE, R., MARINHO, L., HOTHO, A., SCHMIDT-THIEME, L., AND STUMME, G. 2008. Tag recommendations in social bookmarking systems. *AI Comm. 21,* 4, 231–247.

KASER, O. AND LEMIRE, D. 2007. Tag-cloud drawing: Algorithms for cloud visualization. CoRR abs/cs/0703109.

KRESTEL, R., FANKHAUSER, P., AND NEJDL, W. 2009. Latent dirichlet allocation for tag recommendation. In *Proceedings of the ACM International Conference on Recommender Systems*. ACM, 61–68.

KWAK, H., LEE, C., PARK, H., AND MOON, S. 2010. What is twitter, a social network or a news media? In *Proceedings of the International World Wide Web Conference (WWW'10)*. ACM, 591–600.

LEE, S. O. K. AND CHUN, A. H. W. 2007. Automatic tag recommendation for the web 2.0 blogosphere using collaborative tagging and hybrid ann semantic structures. In *Proceedings of the WSEAS International Conference on Applied Computer Science*. World Scientific and Engineering Academy and Society (WSEAS), 88–93.

LIU, Y., LIU, M., CHEN, X., XIANG, L., AND YANG, Q. 2009. Automatic tag recommendation for weblogs. In *Proceedings of the Conference on Innovations in Theoretical Computer Science*. IEEE Computer Society, 546–549.

LU, Y.-T., YU, S.-I., CHANG, T.-C., AND HSU, J. Y.-J. 2009a. A content-based method to enhance tag recommendation. In *Proceedings of the International Joint Conference on Artificial Intelligence*. Morgan Kaufmann Publishers Inc., 2064–2069.

LU, Y.-T., YU, S.-I., CHANG, T.-C., AND HSU, J. Y.-J. 2009b. A content-based method to enhance tag recommendation. In *Proceedings of the International Joint Conference on Artificial Intelligence*. Morgan Kaufmann Publishers Inc., 2064–2069.

MANNING, C. D., RAGHAVAN, P., AND SCHUTZE, H. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

MILLEN, D. R., FEINBERG, J., AND KERR, B. 2006. Dogear: Social bookmarking in the enterprise. In *Proceedings of the Conference on Human Factors in Computing Systems*. ACM, 111–120.

MISHNE, G. 2006. Autotag: A collaborative approach to automated tag assignment for weblog posts. In *Proceedings of the International World Wide Web Conference*. ACM, 953–954.

RAMAGE, D., HALL, D., NALLAPATI, R., AND MANNING, C. D. 2009. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the Conference on Empirical Methods on Natural Language Processing and Computational Natural Language Learning*. ACL, 248–256.

RENDLE, S., BALBY MARINHO, L., NANOPOULOS, A., AND SCHMIDT-THIEME, L. 2009. Learning optimal ranking with tensor factorization for tag recommendation. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 727–736.

RIVADENEIRA, A. W., GRUEN, D. M., MULLER, M. J., AND MILLEN, D. R. 2007. Getting our head in the clouds: Toward evaluation studies of tagclouds. In *Proceedings of Conference on Human Factors in Computing Systems*. ACM, 995–998.

SI, X. AND SUN, M. 2009. Tag-lda for scalable real-time tag recommendation. *J. Computa. Info. Syst*.

SINCLAIR, J. AND CARDEW-HALL, M. 2008. The folksonomy tag cloud: when is it useful? *J. Info. Sci. 34,* 1, 15–29.

SONG, Y., ZHUANG, Z., LI, H., ZHAO, Q., LI, J., LEE, W.-C., AND GILES, C. L. 2008. Real-time automatic tag recommendation. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 515–522.

SYMEONIDIS, P., NANOPOULOS, A., AND MANOLOPOULOS, Y. 2008. Tag recommendations based on tensor dimensionality reduction. In *Proceedings of the ACM International Conference on Recommender Systems*. ACM, 43–50.

VENETIS, P., KOUTRIKA, G., AND GARCIA-MOLINA, H. 2011. On the selection of tags for tag clouds. In *Proceedings of the ACM International Conference on Web Search and Data Mining*. ACM, 835–844.

WANG, J. AND DAVISON, B. D. 2008. Explorations in tag suggestion and query expansion. In *Proceedings of the Workshop on Search and Social Media*. ACM, 43–50.

WU, D., NGAI, G., CARPUAT, M., LARSEN, J., AND YANG, Y. 2002. Boosting for named entity recognition. In *Proceedings of the International Conference on Computational Linguistics*. Association for Computational Linguistics, 1–4.

XU, S., BAO, S., FEI, B., SU, Z., AND YU, Y. 2008. Exploring folksonomy for personalized search. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 155–162.

XU, Z., FU, Y., MAO, J., AND SU, D. 2006. Towards the semantic web: Collaborative tag suggestions. In *Proceedings of the International World Wide Web Conference Workshop on Collaborative Web Tagging*.

ZHAI, C. AND LAFFERTY, J. 2001. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 334–342.

ZHANG, X., YANG, L., WU, X., GUO, H., GUO, Z., BAO, S., YU, Y., AND SU, Z. 2009. sdoc: Exploring social wisdom for document enhancement in web mining. In *Proceedings of the International Conference on Information and Knowledge Management*. ACM, 395–404.

Received August 2011; revised November 2011; accepted January 2012