Fast and Semantic Measurements on Collaborative Tagging Quality

Yuqing Sun[1,2(✉)], Haiqi Sun[3], and Reynold Cheng[4]

[1] School of Computer Science and Technology, Shandong University, Jinan, China
sun_yuqing@sdu.edu.cn
[2] Engineering Research Center of Digital Media Technology, MOE, Jinan, China
[3] Software College, Shandong University, Jinan, China
shqonline@yeah.net
[4] Department of Computer Science, The University of Hong Kong, Hong Kong, China
ckcheng@cs.hku.hk

Abstract. This paper focuses on the problem of tagging quality evaluation in collaborative tagging systems. By investigating the dynamics of tagging process, we find that high frequency tags almost cover the main aspects of a resource content and can be determined stable much earlier than a whole tag set. Motivated by this finding, we design the swapping index and smart moving index on tagging quality. We also study the correlations in tag usage and propose the semantic measurement on tagging quality. The proposed methods are evaluated against real datasets and the results show that they are more efficient than previous methods, which are appropriate for a large number of web resources. The effectiveness is justified by the results in tag based applications. The light weight metrics bring a little loss on the performance, while the semantic metric is better than current methods.

1 Introduction

Collaborative tagging, also known as crowdsourcing or folksonomy system, is widely adopted by web social applications, such as *Del.icio.us*, *Flickr* and *MovieLens*. It encourages users to annotate resources, such as URLs, images or movies, with bookmarks according to their understanding. Each bookmark contains a set of tags. After receiving a number of bookmarks, the tag frequency distribution of a resource would remain stable [14]. These tags and frequencies are regarded as the meta data of resources and are used for recommendation or information retrieval. This method provides an easy way to organize a large quantity of web resources, which is more efficient than traditional classification by specialists. Besides, collaborative tagging plays an important role in tag based applications, such as recommendation [4], web clustering [10], and search [1]. For example, tag-based retrieval is more efficient and effective than traditional full-text search [2].

This work is supported by NSF China (61173140), SAICT Experts Program, Independent Innovation & Achievements Transformation Program (2014ZZCX03301) and Science & Technology Development Program of Shandong Province (2014GGX101046).

J. Bailey et al. (Eds.): PAKDD 2016, Part II, LNAI 9652, pp. 363–375, 2016.
DOI: 10.1007/978-3-319-31750-2_29

Since collaborative tagging is a kind of user subjective action, users often choose their interested resources for tagging. So resources have different numbers of bookmarks. In practice only a small proportion of resources receive enough bookmarks and their tagging states reach stable. That is to say, the tag frequency distributions remain almost the same even if they continuously receive new bookmarks. Considering a large number of resources with few bookmark, their tag frequency distributions change with a new coming bookmark. The states of these resources are called *under-tagged* [16]. According to the findings in [13], a stable tag set is helpful for tag-based applications, such as retrieval or recommendation. But the tags of *under-tagged* resources often affect the correctness of results.

To verify whether a resource reaches tagging stable, the notion of tagging quality is introduced. One widely adopted criterion is the similarity score, which computes the tag distribution similarity in several consecutive tagging points [9,15]. Relative entropy is another way to gauge the stability, which computes the distance between tag frequency distributions on two tag sets [6]. Although these methods provide objective evaluation on tagging state, there are two shortcomings. One is lack of internal link evaluation on tag usage. For example, the phrases *iOS* and *Apple phone* together appear in many resources, which indicates their closeness in semantics and similar usages. However, these correlations have been overlooked in the current methods. Another is time consuming. It takes much computation on measuring two tag frequency distribution, which is not appropriate for a large quantity of web resources. So, it is necessary to take these characteristics into consideration when evaluating the tagging state.

By analyzing the tagging process, we find that noisy tags are the main cause on influencing a tag set stable. According to the previous study [7], most noisy tags are not related to a resource content, which may be caused by user misoperation or misunderstanding. Since the purpose of collaborative tagging is to find the semantics of a resource, the representative tags are often desired to describe a resource and the noisy tags can be neglected. By exploiting the evolution of tag set in collaborative tagging process, we find that some high frequency tags almost cover the main aspects of a resource content and can be determined stable much earlier than the whole tag set. This motivates us to design novel measurements, the swapping index and the smart moving index, to evaluate a resource tagging quality against these representative tags. We also study the inner links between tags and find that some tags often together appear in bookmarks. The notion *concept* is introduced to represent their relationships and semantics. Then a tag set can be reorganized semantically. Based on *concept*, we propose a semantic measurement on tagging quality. We perform a series of experiments on real datasets to justify our methods. The results show that the proposed metrics are more efficient than previous methods. We also adopt some tag based applications to verify their results against stable tag sets under different metrics. The results show the effectiveness of our methods.

The remainder of this paper is organized as follows: Sect. 2 presents the related works and basic notions. In Sect. 3 we describes the datasets and present our findings. Sections 4 and 5 introduce the efficient metrics and semantic

measurement on tagging quality, respectively, as well as the experiments. Finally, conclusions and future works are discussed.

2 Related Works

Dynamics on Collaborative Tagging. Collaborative tagging systems have attracted much attention in recent years. Golder et al. analyze the characteristics of tagging systems and find that the tag frequency distribution of a resource remains almost unchanged after receiving enough bookmarks, named as the stable state [5]. Halpin et al. analyze several aspects of the dynamics in collaborative tagging, including why tag distribution follows the *Power Law*, the patterns of tag distribution for stable resources, and tag correlation or completeness etc. [6]. Trushkowsky et al. estimate the completeness of answers for enumerative questions in crowdsourced database [12]. Different from these works, we study the influence of tag correlations and representatives in tagging quality.

Tag Based Application. Bischoff et al. evaluate the effectiveness of tags obtained in collaborative tagging systems in information retrieval [2]. They find that not every tag well describes the content of a resource. There are vocabulary problems such as tag polysemy and tag synonymy in collaborative systems. Heymann et al. compare the tag based search to the full text based search so as to improve retrieval results [7]. Chi et al. also investigate the efficiency of collaborative tagging systems in information retrieval by use of *information theory* and propose some methods to improve the tag-based search [3]. These motivate us to evaluate the quality of a stable tag set against tag-based applications.

Incentive System and Measurement on Tagging Quality. In collaborative tagging systems, only a few resources can get enough tags such that their tagging quality is good enough to describe resource contents [16]. To promote the tagging quality of under-tagged resources, Yang et al. propose an incentive-based tagging mechanism which rewards users on tagging unstable resources. They propose the Moving Average (MA) score to measure the tagging quality of resources [16]. Halpin et al. also evaluate the tagging quality by the relative entropy (KL divergence) between tag frequency distributions at several consecutive tagging steps [6]. Since these measurements highly rely on the whole tag set, the computation is time consuming and the results are affected by noisy tags.

Basic Notions. The basic notions are cited from the related work [16]. Let $\mathcal{R} = \{r_1, r_2, \ldots, r_n\}$ and $\mathcal{T} = \{t_1, t_2, \ldots, t_m\}, n, m \in N^+$, be the resource set and tag set in a collaborative tagging system, respectively. A bookmark is a finite nonempty set of tags annotated to a resource by a user in one tagging operation. The j^{th} bookmark received by resource r_i is denoted as $b_i(j) = \{t_1, t_2, \ldots, t_l\} \subset \mathcal{T}, j \geq 1, l \in N^+$. Let π_i^n denote the time point of r_i receiving its n^{th} bookmark. For r_i, the tag set at π_i^n is denoted by $T_i(n) = \bigcup_{1 \leq j \leq n} b_i(j)$. The frequency of tag t for r_i at π_i^n is the number of bookmarks containing t that r_i has received at π_i^n, denoted by $h_i(t, n) = |\{b_i(j) | 1 \leq j \leq n, t \in b_i(j)\}|$. The relative frequency is the

normalized frequency $f_i(t,n) = \frac{h_i(t,n)}{\sum_{t' \in T_i(n)} h_i(t',n)}$. The **relative tag frequency distribution (rfd)** of r_i at π_i^n is a vector $\boldsymbol{F_i}(n)$, where $\boldsymbol{F_i}(n)[j] = f_i(t_j, n)$.

When a resource receives enough bookmarks , its *rfd* changes less and reaches stable, called **stable rfd** and denoted by $\varphi_i = \lim_{n\to\infty} \boldsymbol{F_i}(n)$. The **tagging quality** of r_i at π_i^n is the similarity between its current rfd $\boldsymbol{F_i}(n)$ and its stable rfd φ_i, denoted by $q_i(n) = sim(\boldsymbol{F_i}(n), \varphi_i)$. Actually, the ideal stable state is impossible to get to since collaborative tagging is an infinite process. So a practical evaluation on tagging quality **Moving Average score(MA)** is used to quantify the changes of tag frequency distribution in several consecutive steps. Given a parameter $\omega \geq 2$, the MA score of r_i at $\pi_i^n (n \geq \omega)$ is $m_i(n,\omega) = \frac{1}{\omega-1} \sum_{j=n-\omega+2}^{n} sim(\boldsymbol{F_i}(j-1), \boldsymbol{F_i}(j))$. For a given parameter τ (close to 1) as the threshold, r_i is defined tagging stable when $m_i(n,\omega) \geq \tau$.

3 Dataset and the Dynamics

In this section, we introduce the datasets adopted in this paper and investigate their dynamics from two aspect: which are the representative tags of a resources content and their evolution in a collaborative tagging process.

We adopt three real datasets. The *del.icio.us*- 2007 dataset contains the resources from web application *del.icio.us*. There are 5000 stable resources, 562,048 bookmarks, and 2,027,747 tags. On average each resource received 112 bookmarks, which contain 83 distinct tags. The second dataset *Last.fm* is about a music website, which contains the data from August 2005 to May 2011. There are 33 stable resources and each receives 178 bookmarks on average. The third dataset *MovieLens* contains the rating data for movies from Dec 2005 to March 2015 selected from website *MovieLens*. We select 256 resources, where each receives 413 bookmarks on average. The resources we select have reached their stable states against the MA score with $\tau = 0.9999$ and $\omega = 20$.

To understand the tag set of a resource, we analyze tag distribution and representatives tag. We firstly compute all tag frequency distributions, which follow the *Power Law* as the previous work [6]. To further evaluate a high frequency tag, we count the ratio of users on each tag, namely $\frac{f_i(t,n)}{n}, t \in \mathcal{T}$, and find that the *top-1* tag is used by approximately 62 % users and even the 10^{th} popular tag is used by more than 10 % users. High frequency tags indicate the consensus of a large population on resource content, which are helpful for tag-based applications. On the contrary, a low frequency tag is considered as someone's personal understanding and is often regarded noisy. This motivates us to employ tag frequency as the criterion on representative tags.

Then we study when representative tags can be determined in a tagging process. For each resource, we select four time points of its tagging process. The first point N_s records the number of bookmarks when a resource reaches stable. The average N_s of resources in datasets *delicious*, *Last.fm* and *MovieLens* are 84, 130 and 246, respectively. The other three points are selected $\frac{1}{4}N_s$, $\frac{2}{4}N_s$ and $\frac{3}{4}N_s$, respectively. For each point, we count the frequencies of popular tags of $r_i \in \mathcal{R}$, shown in Fig. 1(a), where x axis gives a tag rank (top 1 to 60), y lists

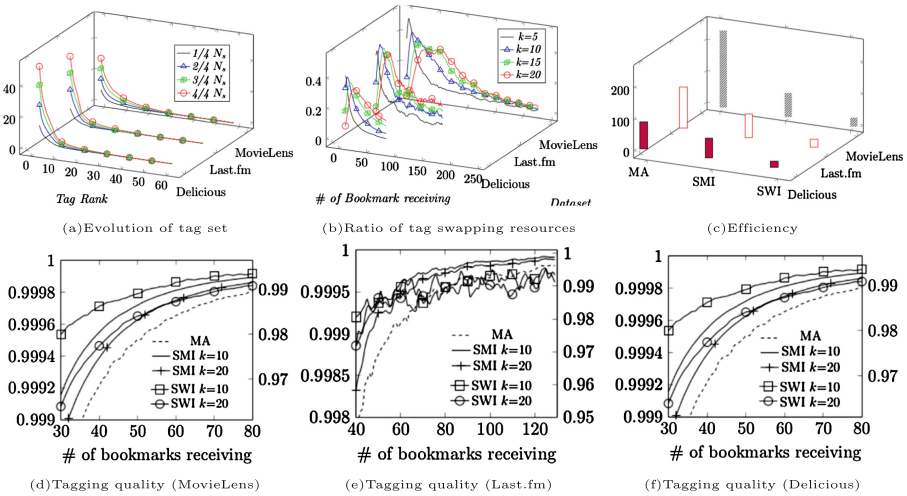


Fig. 1. The dynamics of collaborative tagging

the datasets and z gives the average tag frequency of resources in a dataset. The results show that all tag frequency distributions follow the *Power Law* on different time points. The more bookmarks, the more obvious the trend. For example, the top-1 tag in *Dilicious* is covered by approximate 50 bookmarks at N_s, while this number is about 24 at $1/2N_s$.

A coming question is whether the representative tags of a resource remain unchanged in a tagging process. We consider the top k tags and introduce the concept *tag swapping* to denote the change of top k tags at two consecutive points. For a fix k, we compute the ratio of tag-swapping resources at each point, shown in Fig. 1(b), where x gives time points of tagging process, y lists the datasets and z gives the ratio. We can see as the increase of bookmarks, the ratio decreases, namely the stability of a tag set gets better. Comparatively, there are more *tag swapping* under a larger k. This is caused by low frequency tags included in a top-k tag set, whose frequencies are very close and a new bookmark may change their rankings. But if k is set too small, only a few aspects of a resource can be captured. So, a decent k is desired.

4 The Light Weight Metrics on Tagging Quality

To reduce the influence of noisy tags and improve the efficiency of measuring tagging quality, we introduce two novel measurements *Swapping Index* and *Smart Moving Index*, which are *light weight* in computation.

4.1 Swapping Index and Smart Moving Index

Definition 1. *Given parameters* $\omega \geq 2$, $k \geq 1$ *and* $T_i^k(n) = \{t|rank(t) \leq k \cap t \in T_i(n)\}$, *the* ***Swapping Index (SWI)*** *of* r_i *at* π_i^n $(n \geq \omega)$ *is given by*

$$swi_i(n, \omega, k) = \frac{\sum_{j=n-\omega+2}^{n} |T_i^k(j) \cap T_i^k(j-1)|}{k(\omega - 1)} \tag{1}$$

SWI measures the tagging state by computing the intersection between two consecutive *top-k* tag sets under window ω. To determine whether a *SWI* is good enough, we adopt a parameter $\tau > 0$ as the indicator of stable state. r_i is tagging stable when $swi_i(n, \omega, k) > \tau$. Notice that, only if a resource receives more than ω bookmarks and its tag set contains not less than k distinct tags, the *SWI* can be calculated. This requirement is easy to satisfy in practice. Different from the current *MA* score, *SWI* only considers the representative tags.

Definition 2. *Given parameters* $\omega \geq 2$ *and* $k \geq 1$, *the* ***Smart Moving Index (SMI)*** *of resource* r_i *at* π_i^n $(n \geq \omega)$ *is given in Eq. 2, where* $\boldsymbol{F_i}^k(n)$ *is the top-k relative tag frequency distribution whose members* $\boldsymbol{F_i}^k(n)[j] = f_i(t_j, n), t_j \in T_i^k(n)$, *and* $\boldsymbol{F_i}(j-1)_{T_i^k(j)}[m] = f_i(t_m, j-1)$, $t_m \in T_i^k(j-1) \cap T_i^k(j)$. *sim is a metric to quantify the similarity between two adjacent tag vectors.*

$$smi_i(n, \omega, k) = \frac{1}{\omega - 1} \sum_{j=n-\omega+2}^{n} sim(\boldsymbol{F_i}^k(j-1)_{T_i^k(j)}, \boldsymbol{F_i}^k(j)) \tag{2}$$

SMI evaluates a tag set by the similarity between adjacent tag frequency distributions. Given a parameter $\tau > 0$ as a threshold, r_i is called tagging stable when $smi_i(n, \omega, k) > \tau$. Notice that only if a resource has received more than ω bookmarks and the size of its tag set is not less than k, *SMI* can be defined.

4.2 Experimental Study

The programs in this paper are written in C++ and experiments are executed on a machine with 4G memory, Intel i3 CPU, installed with 32-bit Linux system. As comparison, we adopt the widely adopted moving score (*MA*).

Tagging State Evaluation. We first verify the trend of each index in a tagging process. The results for three datasets are shown in Fig. 1(d),(e) and (f), where x axis gives the number of bookmarks a resource receives, y gives the score of *MA*, *SMI* and *SWI*. Since *MA* and *SMI* have the same domain, they follow the left vertical axis, while *SWI* is against the right axis. The results are the average value of all resources. The parameters here are $\omega = 5, k = 10, 20$. The results show that there are similar trends for three methods. With the increase of bookmarks, the tagging quality of resources get better.

Table 1. Number of Bookmarks Required vs k

	SMI k=5	*SMI* k=10	*SMI* k=15	*SMI* k=20	*SWI* k=5	*SWI* k=10	*SWI* k=15	*SWI* k=20
MovieLens	38.7188	78.695	113.234	137.87	22.2812	29.5156	37.2031	45.8125
Lastfm	49.0909	75.8182	83.9394	89.5758	15.3636	24.4242	30.9091	37.4242
Delicious	51.0962	62.7162	67.5676	70.7408	13.9108	19.1994	24.52	30.306

Efficiency Measurement. The efficiency is evaluated on two sides, where in experiments $\omega = 5$, $\tau = 0.9999$, and $k = 10$. One is to compute the required number of bookmarks for a resource to reach stable under different metrics, shown in Fig. 1(c), where x-axis lists the three stability indexes, y lists the datasets and the z-axis gives the average number of required bookmarks. The results show that *SWI* requires the least bookmarks for a resource to be stable and *MA* requires the most bookmarks. We further compare the results under different k. From the results in Table 1, we can see that the required number of bookmarks scales with the increasing k for both *SMI* and *SWI*. This is because a larger k takes into account more low frequency tags, whose rankings are less stable than high frequency tags.

Another is to evaluate the runtime. The first experiment tests how the performance scales with the number of resources, which helps us understand a system workload. The average runtime on 5000 resources in *delicious* dataset are shown in Fig. 2(a), where x-axis gives the quantity of resources and y-axis gives the runtime. The results show that our methods are more efficient than *MA*, which illustrates that they are appropriate for a large scale system. Comparatively, *SWI* is faster than *SMI*. We further analyze the effect of k setting on runtime. The results in Fig. 2(b) show that *SWI* and *SMI* are much more efficient than *MA*, since *MA* computes the similarity for all tag frequencies, which is time-consuming. Besides, *MA* stays relative stable since it is not affected by k. The runtime for both *SMI* and *SWI* gets longer as the increasing k. The reason is that a larger k means more tags involved in measurement. *SWI* is always more efficient than *SMI*.

Comparison on Stable Tag Set Usage. In order to understand how our proposed methods work in practical applications, we investigate the usage of a stable tag set in tag related applications. Specially, we consider the tag-based recommendation, which is one of the most web popular applications. Three representative recommendation algorithms are adopted in our experiments, namely *Item-based Hierarchical Clustering*, *Item-based k-Means Clustering* etc. The benchmark is the recommendation results returned by the final tag set, denoted by Rec_{final}. When resources are evaluated stable by *SMI*, the recommendation results are computed against the current tag sets, denoted as Rec_{SMI}. Similarly, the recommendation results for *SWI* and *MA* are denoted as Rec_{SWI} and Rec_{MA}, respectively.

We consider two criteria on the accuracy of recommendation results. One is the similarity between two recommended sets, denoted as $Accuracy_{set} = sim(Rec^*, Rec_{final})$, where $sim()$ is the similarity evaluation function and Rec^*

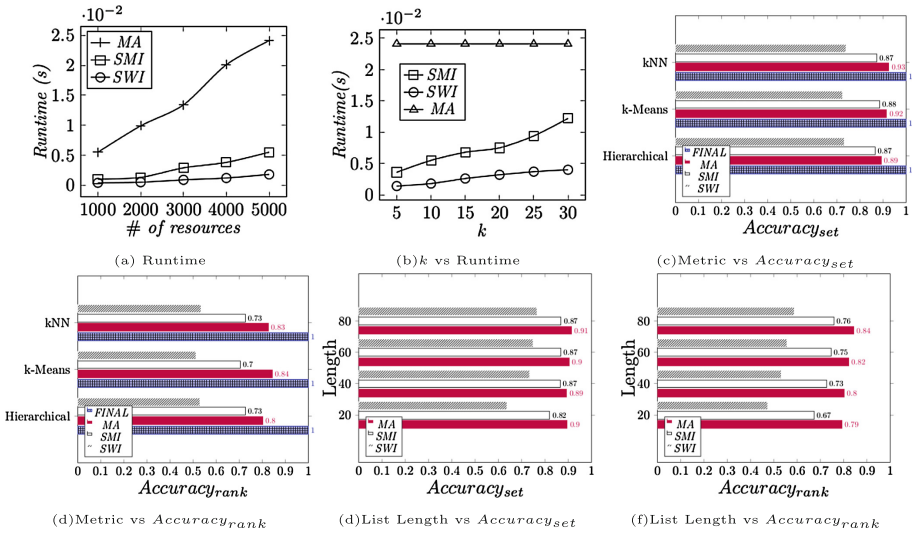


Fig. 2. Tag-based Recommendation Accuracy

indicates an alternative recommendation result based on different metric, such as Rec_{SMI}, Rec_{SWI} or Rec_{MA}. Another criterion evaluates the ranking difference between two recommended lists. We adopt the *Kendall tau distance*, which is widely used to evaluate the number of pairs in opposite order in two rankings [8]. Let $K(Rec^*, Rec_{final})$ denote the *Kendall tau distance*. This accuracy is defined as $Accuracy_{rank} = \frac{|Rec^* \cap Rec_{final}|}{|Rec^*|} * (1 - K(Rec^*, Rec_{final}))$.

For each recommendation algorithm, we compare the results on different stable tag sets against *MA*, *SWI* and *SMI*, respectively. This experiment is performed on the *delicious* dataset and the parameters are $\omega = 5$, $\tau = 0.9999$ and $k = 10$. Each value is computed as the average recommendation accuracy for all resources, i.e. the 5000 resources in the dataset. The results under different algorithms are shown as different colored bars. We first investigate the accuracy on different metrics and show the results in Fig. 2(c) and (d), where x gives the accuracy and y lists the algorithms. Here we choose the top 20 elements in the recommendation results. The result of *SWI* is lower than others. This is because *SWI* only considers the components of representative tags and ignores their frequency. The accuracy of *SMI* is close to *MA*. Comparing two accuracy evaluation methods, $Accuracy_{set}$ is higher than $Accuracy_{rank}$. This is because the $Accuracy_{set}$ metric only focuses on the members in a recommendation result, while the *Kendall tau distance* method considers the ranking of each element. Then we investigate how accuracy scales with the recommendation list length, shown in Fig. 2(e) and (f), where y gives the length of recommendation results. The accuracy increases with the length.

Summary. Taking into account the above evaluations, we can see that *SWI* and *SMI* are more efficient than the current method, which resides on two sides: the direct computation time and the required period for a resource becoming stable. For an expected stable tagging state, fewer bookmarks are required against our proposed metrics. For example, *SWI* needs about 36 % bookmarks against *MA*. This is because they compute the representative tags, which reflect the consent by a large population and can reach stable much earlier than some unpopular tags. Consiering the usage of stable tag set, *SMI* is much closer to *MA*. As shown in Fig. 2(b), when k is set 25, the result of *SMI* is very close to *MA*. But it saves about 60 % computation time than *MA*. So the proposed metrics on tagging quality are appropriate for a large scale system.

5 Semantic Measurement on Tagging Quality

Tags reflect user understanding on a resource content. Different tags reflect different facets of a resource, while two users may adopt different tags to describe the same aspect of a resource. So, multiple tags may share the identical sense and reflect the similar facet of resource content. For example, the word *iphone* is highly related to *Apple* than *tomato*. This motivates us to investigate the intrinsic associations between tags and take into account tag semantics in tagging quality evaluation.

5.1 The Semantic Metric on Tagging Quality

In this paper, we adopt the notion *concept* to model a resource content, which can be hierarchically organized. A *concept* is defined as a set of tags associated with the semantics in a system. Considering a collaborative tagging system, a concept is calculated by the tags together used in a bookmark. Formally, a tag vector is defined as $\boldsymbol{t} =< f_1(t), f_2(t), \ldots, f_{|R|}(t) >, t \in \mathcal{T}$, where $f_i(t)$ is the relative frequency of tag t for resource r_i. For two tags t_i and t_j, its similarity *sim*(t_i, t_j) can be calculated against the tag vector. Given a threshold τ and a distance increment δ, concepts can be iteratively clustered as a hierarchical tree. Details on the hierarchical clustering can be got in [11]. Each concept maps to a node in the hierarchies. Let h denote the height of the tree, and $\eta \in [1, h]$ denote a specific level of tree, which reflects the semantic closeness of tags. Each concept of level 1 (leaf node) maps to a concrete tag. A concept in a higher level has a broader semantics, which is iteratively generated by combining the semantically close concepts. Formally,

Definition 3. *A hierarchical tree consists of* ***concepts****. The concept on level* $\eta \in [1, h]$ *with* m *children is given below, where* $c_j^{\eta-1}$ *is the* j^{th} *child of* c^{η}.

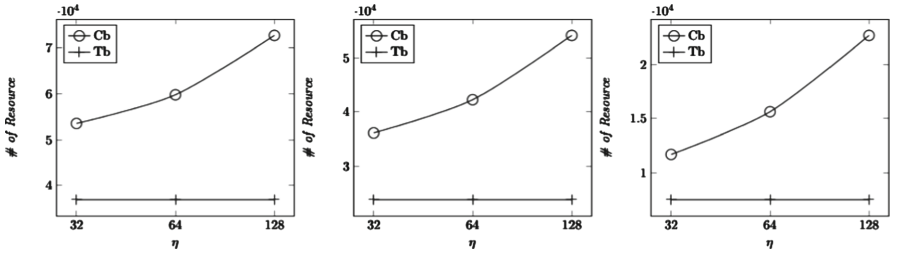
$$c^{\eta} = \begin{cases} \{t\} \subset \mathcal{T} & \eta = 1 \\ \bigcup_{1 \le j \le m} c_j^{\eta-1} & \eta > 1 \end{cases} \tag{3}$$

Actually, the hierarchies reflect the crowd intelligence. Compared to the plain folksonmy, it takes the advantage of the taxonomy systems to manage tags in a systematic way. Besides, it can locate the related resources quickly when processing a query [5]. Base on the notion *concept*, we introduce the semantic measurement of tagging quality.

Definition 4. *Give a parameter* η*, the* ***resource concept distribution function (rcf)*** *of resource* r_i *at* π_i^k *is a vector* $\boldsymbol{F_i}^\eta(k)$*, s.t. the* j^{th} *component is the frequency of concept* c_j^η *for* r_i *at* π_i^k*,* $\boldsymbol{F_i}^\eta(k)[j] = f_i(c_j^\eta, k) = \frac{\sum_{t \in T_i(k) \cap c_j^\eta} h_i(t,k)}{\sum_{t \in T_i(k)} h_i(t,k)}$.

Definition 5. *Given parameters* $\omega \geq 2$*,* η *and* τ*, the* ***Semantics Index(SI)*** *of* $r_i \in \mathcal{T}$ *at* $\pi_i^k(k \geq \omega)$ *is given by Eq. 4.* r_i *is called stable if* $sem_i^\eta(k, \omega) > \tau$.

$$sem_i^\eta(k, \omega) = \frac{1}{\omega - 1} \sum_{j=k-\omega+2}^{k} sim(\boldsymbol{F_i}^\eta(j-1), \boldsymbol{F_i}^\eta(j)) \tag{4}$$



(a)$\delta = 0.03$ (b)$\delta = 0.06$ (b)$\delta = 0.12$

Fig. 3. Comparison on Semantic Index and Tag based Metrics

5.2 Experiment Analysis

The Tag Based Application. To evaluate the effectiveness of the semantic index, we select the tag based retrieval, one of the most popular applications, as the evaluation tool, which answers a tag based query with related resources. We justify the tagging quality metrics by the output of tag based retrieval against each tag set that is evaluated stable against a metric. The experiments are performed on the *delicious* dataset. In a collaborative tagging system, a tag frequency reflects how much people have a consensus on a resource by the tag, which is also admitted by the tag based retrieval. Let parameter $\delta \in [0..1]$ denote the relatedness threshold between a tag and a resource. For once query with tag t, resource r_i is selected as the result if $f_i(t, n) > \delta$. In practice, this parameter can be learned from user multiple retrievals. Similarly, the concept based evaluation is $f_i(c, n) > \delta$ or $\frac{f_i(c^\eta, n)}{|\{\hat{t} | \hat{t} \in c^\eta \cap T_i(n)\}|} > \delta$, where $t \in c^\eta$.

Effectiveness Comparison. We compare the concept-based measurement (*Cb* for short) and Semantic Index (SI) with the *Tag-based* measurement (*Tb* for short) MA. The parameters in this experiment are $\omega = 5, \tau = 0.9999$. Figure 3 shows the results for the retrieval application under different constraints, where x axis gives η settings and y gives the average number of returned resources on tag based queries. Overall, *Cb* is better than *Tb* since it can find more related resources which are not in an obvious mode. A larger δ means a stricter relatedness on tag and resource and results in fewer satisfied resources. A concept in a higher level η contains more semantics such that more resources satisfy a query.

We further evaluate the returned results on retrieval application against the final tag set in a collaborative tagging system. Let tp denote the number of resources in the *true positive* case, i.e. the resources appear in the retrieval results by both the SI stable tag set and the final tag set. Similarly, tn, fp and fn denote the cases *true negative*, *false positive*, and *false negative*, respectively. Thus, $precision = \frac{tp}{tp+fp}$, and $recall = \frac{tp}{tp+fn}$. The parameters are $\omega = 5$, $\tau = 0.9999$ and $\delta = 0.05$ for SI and MA. The comparison results are listed in Table 2, which show that the *precision* and *recall* are very close under SI and MA. A higher η brings a slight lower *precision* and *recall* since more general content are extracted and fewer bookmarks are desired for a resource being tagging stable. In practice, the parameters is set by an administrator at first so as to solve the *cool start* problem. Then it can be learned in the process of tag based applications.

Efficiency Measurement. It is necessary to consider both the required bookmarks for a resource be stable and the runtime for computing metrics. Comparing SI and MA, the complexity on each metric computation is the same. So, their difference depends on the number of required bookmarks for each resource reaching its stable state. From the results in Table 2, we can see the required number of bookmarks for a resource to be tagging stable is smaller under SI than MA. With an increasing η, fewer bookmarks are desired for a resource being tagging stable. At the first glance, the reduced number of bookmarks is not very large, such as the average number of saved bookmarks is about 20 when $\eta = 128$. However, if we take into account the whole tagging process, each resource receives 100 bookmarks on average around a whole year. It is easy to conclude that 20 bookmarks are expected for approximately one month. So, the improvement on evaluation efficiency is meaningful in practice.

Table 2. Comparison of SI and MA on Retrieval Results

	SI η=32	MA η=32	SI η=64	MA η=64	SI η=128	MA η=128
Precision	0.962848	0.967474	0.960242	0.968447	0.956816	0.97073
Recall	0.958032	0.963424	0.952392	0.963933	0.945782	0.96517
Required Bookmark	77.247	83.03	70.949	83.03	63.1314	83.03

6 Conclusion

In this paper, we propose several metrics on collaborative tagging quality evaluation, which are light weight on computation and effective than the current method. The SWI and SMI metrics take the representative tags for measurement to get rid of the influence of noisy tags in making a resource tagging stable. The semantic measurement SI takes tag intrinsic associations into consideration, which makes tagging quality evaluation more effective. A series of experiments are performed against several real datasets and results show the efficiency and effectiveness of our methods. This illustrates that the proposed methods are especially appropriate for a large quantity of web resources. In the future, we will take user personal preferences into account for quality study. Another direction is to investigate how to apply our methods into other crowdsourcing applications.

References

1. Bi, B., Lee, S.D., Kao, B., Cheng, R.: Cubelsi: an effective and efficient method for searching resources in social tagging systems. In: Proceedings of the 27th IEEE International Conference on Data Engineering (ICDE), pp. 27–38. IEEE (2011)
2. Bischoff, K., Firan, C.S., Nejdl, W., Paiu, R.: Can all tags be used for search? In: Proceedings of the 17th ACM Conference on Information and knowledge Management, pp. 193–202. ACM (2008)
3. Chi, E.H., Mytkowicz, T.: Understanding the efficiency of social tagging systems using information theory. In: Proceedings of the 19th ACM Conference on Hypertext and Hypermedia, pp. 81–88. ACM (2008)
4. Durao, F., Dolog, P.: A personalized tag-based recommendation in social web systems. In: Adaptation and Personalization for Web 2.0, p. 40 (2009)
5. Golder, S.A., Huberman, B.A.: Usage patterns of collaborative tagging systems. J. Inf. Sci. **32**(2), 198–208 (2006)
6. Halpin, H., Robu, V., Shepherd, H.: The complex dynamics of collaborative tagging. In: Proceedings of the 16th International Conference on World Wide Web, pp. 211–220. ACM (2007)
7. Heymann, P., Koutrika, G., Garcia-Molina, H.: Can social bookmarking improve web search? In: Proceedings of the 2008 International Conference on Web Search and Data Mining, pp. 195–206. ACM (2008)
8. Kendall, M.G.: Rank Correlation Methods. Griffin, London (1948)
9. Lei, S., Yang, X.S., Mo, L., Maniu, S., Cheng, R.: Itag: incentive-based tagging. In: Proceedings of 30th IEEE International Conference on Data Engineering (ICDE), pp. 1186–1189. IEEE (2014)
10. Ramage, D., Heymann, P., Manning, C.D., Garcia-Molina, H.: Clustering the tagged web. In: Proceedings of the Second ACM International Conference on Web Search and Data Mining, pp. 54–63. ACM (2009)
11. Shepitsen, A., Gemmell, J., Mobasher, B., Burke, R.: Personalized recommendation in social tagging systems using hierarchical clustering. In: Proceedings of the 2008 ACM Conference on Recommender Systems, pp. 259–266. ACM (2008)
12. Trushkowsky, B., Kraska, T., Franklin, M.J., Sarkar, P.: Crowdsourced enumeration queries. In: Proceedings of the 29th IEEE International Conference on Data Engineering (ICDE), pp. 673–684. IEEE (2013)

13. Van Damme, C., Hepp, M., Coenen, T.: Quality metrics for tags of broad folksonomies. In: Proceedings of International Conference on Semantic Systems (I-SEMANTICS), pp. 118–125 (2008)
14. Wagner, C., Singer, P., Strohmaier, M., Huberman, B.A.: Semantic stability in social tagging streams. In: Proceedings of the 23rd International Conference on World Wide Web, pp. 735–746. ACM (2014)
15. Xu, H., Zhou, D., Sun, Y., Sun, H.: Quality based dynamic incentive tagging. Distrib. Parallel Databases **33**(1), 69–93 (2015)
16. Yang, X.S., Cheng, R., Mo, L., Kao, B., Cheung, D.W.-l.: On incentive-based tagging. In: Proceedings of the 29th IEEE International Conference on Data Engineering (ICDE), pp. 685–696. IEEE (2013)