Distrib Parallel Databases (2015) 33:69–93
DOI 10.1007/s10619-014-7164-8

Quality based dynamic incentive tagging

Haoran Xu · Dandan Zhou · Yuqing Sun · Haiqi Sun

Published online: 5 November 2014

Abstract Social tags take an important role in exploratory search. In collaborative tagging systems, users are allowed to annotate resources with tags. The significant challenges in such systems are the uncertainty of tag quality and the incomplete annotation on a large number of resources. Based on the observation that these problems can be statistically negligible after receiving sufficient tags, we propose a novel incentive mechanism to reward taggers according to the quality of their bookmarks, called the *Quality-based dynamic Incentive Mechanism* (*QIM*). To well evaluate the quality of bookmarks, we design some quantitative evaluation methods. The reward allocation function is proposed to allocate the budget to different taggers based on their bookmark quality and the tagging states of annotated resources. We perform experiments to evaluate our method on three public datasets collected from real tagging systems. Comparing with previous works, the adopted principle of "*high quality deserves high price*" in this paper can encourage users to annotate seriously. The experimental results show that our method gets higher tagging quality of resources under a fixed budget. Moreover, it requires less time and less money to achieve the stable tagging state of a system.

Keywords Collaborative tagging · Incentive mechanism · Bookmark quality

H. Xu · D. Zhou · Y. Sun (✉) · H. Sun
School of Computer Science and Technology, Shandong University, Jinan, China
e-mail: sun_yuqing@sdu.edu.cn

H. Xu
e-mail: hr_xu1990@163.com

D. Zhou
e-mail: zdd_water1989@163.com

H. Sun
e-mail: shqonline@yeah.net

1 Introduction

Crowdsourcing is the practice of obtaining needed services, ideas, or contents by soliciting contributions from a large group of people, and especially from an online community, rather than from traditional employees or suppliers. Collaborative tagging is a typical crowdsourcing application. It allows users to annotate web resources like URLs or photos with tags. These tags can be used to categorize and summarize these online resources, which provides a convenient way to manage web contents for searching, mining and recommendation [1–4]. A well known tagging system is *Del.icio.us*, which allows users to store, share, and discover bookmarks of web resources. In the tagging process, users choose tags according to their cognitive understanding of resource content, which may result in quite a few casual or unrelated tags. For example, the statistics show that there are about 60 % low-frequency used tags (less than 50 times) out of the total 1,300 million tags in *Flickr* database [5], which may be caused by misspelling, synonym, polysemy etc. [6–8]. Most of these low-frequency tags are not related to resource contents, which influence the utility of tag-based applications [9], such as resource categorization and information retrieval [10].

The current works solve this problem from two perspectives. One is from the semantic perspective. For example, by the tag recommendation method, a collaborative system provides some related tags as suggestions when a user annotates a resource. Taggers can either choose some tags from the suggested list or submit new tags. Although this method reduces the ambiguity of tags to some extent, it restricts user creativity and may influence the collective intelligence on topic extraction from Web resources [4]. Another typical solution allows users to describe a resource with some semantic assertions instead of tags, which describe some properties of a resource [7]. However, this method highly increases the workload of user tagging and its implementation is complicated.

Another representative is from the quantity aspect, also named as the sufficient tagging method. This is based on the observation that having enough bookmarks, the relative tag frequencies of a resource can naturally reflect the significance of different aspects of its content [11]. In another word, the noisy tags are statistically negligible due to their low number of occurrences. However, in real collaborative tagging systems, only a small portion of resources receive enough bookmarks. Research shows that the tags of the under-tagged resources will affect the correctness of tag based retrieval results and recommendation [12]. Considering this point, the incentive-based tagging solution is proposed to encourage users to annotate under-tagged resources [13]. It improves the tagging quality of resources by rewarding a uniform amount of money to each bookmark. Although this method accelerates to some extent the process of a collaborative tagging system approaching its stable tagging state, it may cost more money than necessary since rewarding a tagger does not consider bookmark quality. There may exist the case that a tagger, for the reward purpose, casually submits tags that are not related with the content of the annotated resource. This deviates from the initial purpose of the incentives.

To encourage taggers seriously working, we take into account bookmark quality as the criterion of rewards, and present a novel incentive mechanism, called *Quality-based dynamic Incentive Mechanism* (*QIM* for short). In this mechanism, we propose some methods to quantitatively evaluate a bookmark quality based on "*crowd proof*". That is to say whether a tag is useful depends on the collective results. A series of experiments on real datasets are performed to verify these methods. We first carefully analyze when the collection of tag assignments are appropriate to evaluate posterior bookmarks. Then we adopt the *backtracking* idea to testify the effectiveness of these methods by comparing the quality for the same bookmark against the present tag set and the stable tag set. Besides, the efficiency and the incentive results are verified carefully. In this mechanism, we also employ a reward allocation function to decide how much a system rewards a tagger. To accelerate the improvement of tagging quality, the reward is designed relevant to the tagging state of resource. We perform some experiments to evaluate this mechanism and the results show three benefits comparing with pervious methods. For a fixed budget, our mechanism makes a higher tagging quality of a system, while for an expected tagging quality state of a system, it pays less. And, it requires less time for a collaborative tagging system to achieve its stable tagging state.

The rest of the paper is organized as follows. Related works are given in Sect. 2, and preliminaries are given in Sect. 3. We present an overview of our incentive mechanism and analyze tagger behaviors under different incentive mechanisms in Sect. 4. In Sect. 5, we propose the evaluation methods of bookmark quality and a dynamic reward allocation strategy. Experiments are performed on real datasets in Sect. 6. And future works are discussed in Sect. 7.

2 Related works

2.1 Tag usage in collaborative tagging systems

To solve the tagging quality problem, it is important to have a deep understanding of tag usage in social tagging systems. In social tagging systems, there exists a phenomenon that certain tags gradually become much more popular than others, named as tag convergence [14,15]. This is mostly caused by the frequent selection of certain tags by a large number of taggers. Lin et al. [16] study the tag convergence phenomenon and find that the aggregated frequency of the top 30 % tags account for 70 % of the total tags. Sood et al. [17] indicate that suggestion of relevant tags for users improves the probability for tag convergence. Besides, Li et al. [18] study the tags of all URLs in *Del.icio.us* and enhance tag convergence by removing noisy tags with low frequency. These studies are helpful for many tag based applications, such as information retrieval [16,19], classification [4,20] and recommendation [21,22]. They utilize tag information to improve the correctness of results. The straightforward usage of tags is as key words based on tag frequencies. High frequency tags are regarded as the crowd cognitive consensus on resource content in these works. The above works are highly related with the representative tags, which motivate us to explore popular tags for bookmark quality evaluation.

2.2 Recommendation based tagging quality improvement

Some methods solve the low-quality tag problem by recommending a set of tags when a tagger annotates a resource. Heymann et al. [23] study the use of machine learning techniques to enrich the information of resources by tag-based association rules. By collaborative filtering, the authors in [24] analyze how users tag photos and what information is contained in tags. The Waking And Sleeping (*WAS*) is another recommendation algorithm proposed by Wen-Hau et al. [25]. The authors in [26–28] also propose some personalized and interactive tag recommendation methods. The above tag recommendation methods to some extent avoid casual tagging and reduce the quantity of noisy tags. But there are some limitations. First, tag recommendation algorithms often require complex computation which reduces the efficiency. Second, recommended tags restrict user creativity, which deviate from the natural of crowdsourcing.

2.3 Semantics based tagging quality improvement

Another kind of methods introduce semantics in folksonomy so as to reduce tag ambiguity. Marchetti et al. [7] propose a semantic tagging algorithm that utilizes external knowledge, such as WordNet and Wikipedia, to discover the semantic relationships and create tag hierarchies so as to to reduce tag ambiguity. Majid et al. [29] compare different methods on fuzzy semantic problems of tags. They point out that knowledge based approaches have higher precision in disambigution than statistical approaches. Recently, Daniela et al. [30] take advantage of semantic similarity to enhance recommendation. However, this kind of methods often need to build a comprehensive ontology, which requires the participation of experts in various fields. It is time-consuming to update the ontology and thus it is not suitable for dynamic tagging process.

2.4 Incentive based tagging quality improvement

The most related work is the quantity based incentive tagging. This is based on the finding that tag frequencies of a resource gradually get stable as the number of received bookmarks increases [6,11]. The noisy tagging problem then can be naturally solved after having sufficient bookmarks. By this observation, the incentive-based mechanism is proposed to encourage taggers to submit bookmarks on under-tagged resources [13,31]. It rewards a user for annotating an under-tagged resource so as to improve the tagging quality of the resource. Since this method is easy to apply to the existing applications, Lei et al. [32] present an incentive-based tagging system on traditional crowdsourcing systems. However, it focuses only on the number of bookmarks without consideration of bookmark quality, which can not avoid casual or irrelevant bookmarks. Besides, Weng et al. [33] design a tagging game and scoring mechanism to encourage users to annotate resources. It rewards a higher score if the chosen bookmarks by a tagger overlap with previous bookmarks on the same resource. Although this method provides an enjoyable way to increase the tagging quality, it restricts user choices on tagging resources and the evaluation of novel tags is not appropriate.

To overcome these shortcomings, we propose a quality-based dynamic incentive mechanism, which rewards a tagger based on both bookmark quality and resource tagging state. We also provide a compensation mechanism to reward popular tag originators for their innovation. Comparing to previous incentives, our method can encourage users to provide high quality annotations and it also accelerates the process of a system approaching tagging stable.

3 Preliminaries

In this section, we present some terms used in this paper, some of which have been previously defined by [13].

Let $\mathcal{R} = \{res_1, res_2, \ldots, res_n\}(n \in N^+)$ be a set of n resources and $\mathcal{T} = \{t_1, t_2, \ldots, t_m\}(m \in N^+)$ denote the set of all tags. A *bookmark* is a non-empty set of tags annotated to a resource by a tagger in one tagging operation. The kth bookmark received by resource res_i is denoted as $b_i(k) = \{t_1, t_2, \ldots\} \subset \mathcal{T}, k \geq 1$. All the bookmarks of resource res_i are arranged in a chronological order. In the following discussion, we adopt the $\pi_i^k\ point$ to denote the time point of resource res_i receiving its kth bookmark.

Definition 1 The $tag\ set$ of a resource res_i, denoted by $\mathcal{T}_i(k)$, is the union set of previous k times bookmarks received by res_i. Initially it is empty, and is iteratively updated by

$$\mathcal{T}_i(k) = \mathcal{T}_i(k-1) \cup b_i(k) \tag{1}$$

Definition 2 At the $\pi_i^k\ point$, the $frequency$ of tag t for res_i is the number of bookmarks that contain tag t in res_i's preceded $[1, \ldots, k]$ bookmarks, denoted as $h_i(t, k)$[13] and given by

$$h_i(t, k) = |\{b_i(j) | 1 \leq j \leq k, t \in b_i(j)\}| \tag{2}$$

Then the $relative\ frequency$ of tag t for res_i at the $\pi_i^k\ point$ is the frequency $h_i(t, k)$ normalized by the number of tags among the res_i's first k bookmarks, denoted as $f_i(t, k)$, given by

$$f_i(t, k) = \frac{h_i(t, k)}{\sum_{t' \in \mathcal{T}_i(k)} h_i(t', k)}, \quad k > 0 \tag{3}$$

At the $\pi_i^k\ point$, the *relative tag frequency distribution* (rfd) of res_i is a vector $\vec{F}_i(k)$, whose jth component is the relative tag frequency of tag t_j for res_i, given by

$$\vec{F}_i(k)[j] = f_i(t_j, k) \tag{4}$$

For a resource res_i, the $tag\ set$ $\mathcal{T}_i(k)$ is used to record all the historical tags received by res_i at the $\pi_i^k\ point$, while the rfd of res_i indicates the relative frequency distribution of tags in $\mathcal{T}_i(k)$.

Definition 3 Given a parameter $\omega \geq 2$, a resource res_i's *Moving Average score* (*MA*) at the π_i^k *point* ($k \geq \omega$) is denoted by $m_i(k, \omega)$, given by

$$m_i(k, \omega) = \frac{1}{\omega - 1} \sum_{j=k-\omega+2}^{k} sim\left(\vec{F}_i(j-1), \vec{F}_i(j)\right) \tag{5}$$

where *sim* is a metric to quantify the similarity of two adjacent rfds for resource res_i, and ω is the window size.

The *MA* score is the average of $\omega - 1$ adjacent similarity scores and it is defined after receiving ω or more bookmarks. In this paper, *MA* score reflects the changes of tag rfd for a resource. The range of the *MA* score is [0,1] and it generally increases with the number of bookmarks received by a resource. We introduce a threshold τ to quantify whether rfd is relatively stable.

Definition 4 Given parameters $\omega \geq 2$ and $\tau \in (0, \ldots, 1)$, an rfd is *practically-stable* when satisfying $m_i(k, \omega) \geq \tau$, denoted as $\hat{\varphi}_i(\omega, \tau)$.

The semantics of the *practically-stable* state of a resource is that its rfd remains almost the same even more bookmarks are received, namely the rfd can be regarded as the appropriate description of this resource [11]. So, *MA* score is adopted in this paper as the indicator of the *tagging state* of a resource. The number of bookmarks where $m_i(k, \omega) \geq \tau$ firstly holds is defined as the ***Stable Point***, denoted as $sp \in N^+$, and the tagging state is called *stable*. After res_i's tagging state is stable, its tag set is denoted as $\widetilde{\mathcal{T}}_i$. The frequency of tag $t \in \widetilde{\mathcal{T}}_i$ is denoted as $h_i(t)$.

4 Quality based incentive mechanism for collaborative tagging

In this section, we will present the overview of the proposed *Quality Based Incentive Mechanism* (*QIM* for short) and analyze tagger behavior under this mechanism.

4.1 Mechanism design

We consider the situation where there are many under-tagged resources in a collaborative tagging system. The purpose of our mechanism is to optimize the process of making a tagging system reach stable under a given budget, which resides on two sides. One is to encourage taggers seriously working on tagging action so that the reward benefits those effective works. Another is to accelerate the process of a system approaching its stable tagging state in an efficient way.

Considering the first problem of quantitatively evaluating a bookmark quality, there are two key points: *what* can be as the evidence to evaluate a bookmark and *how* to evaluate it. According to the spirit of crowdsourcing, the intelligence should be based on user online activities without any *specialist* interaction. Theoretically, only when a tag set is stable, it can be used as the evidence for evaluating a bookmark. However, it is impossible to make an instant reward in practice after a user tagging according to

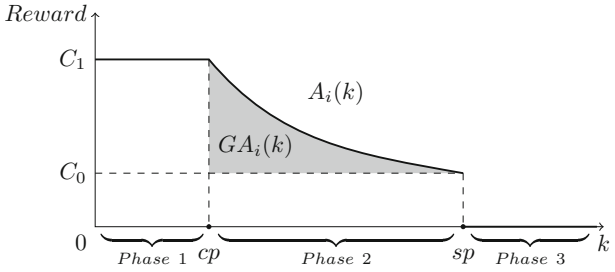


Fig. 1 The dynamic incentive function. $A_i(k)$ denotes the max reward for resource res_i at the π_i^k $point$, and $GA_i(k)$ is the actual reward a tagger can get based on the quality of one's bookmark $b_i(k)$

its stable tag rfd. The evidence of an instant evaluation should be the collection of previous bookmarks. So the upcoming problem becomes how many bookmarks that a resource receives are enough to be used for evaluating the posterior posts. We perform a series of experiments to analyze this problem and have an average point based on the statistic results. That point is called ***Critical Point*** afterwards and is denoted as $cp \in N^+$. By analyzing multiple aspects of tagging records, we find that for any resource that has received the number cp of posts, the main characteristics of the tag distribution are quite similar with its stable tagging state. The concrete evaluation methods will be given in Sect. 6.2.

To accelerate the process of a system approaching its stable tagging state, we need to balance rewards among different resources. Theoretically, the ultimate goal of an incentive mechanism should make all resources in a tagging system reach stable. However, under a given budget, it is difficult to achieve this goal in a limited period. A policy must be made on either maximizing the number of stable resources or averagely improving the tagging quality of all resources. Considering the tag usage in tag based applications, such as information retrieval [21] or recommender system [22], a nearly stable tag set for a resource, is more helpful than a far unstable one, although it is worse than a final stable tag set. Hence, many related works choose the policy of improving average quality [13]. Since the trend of the tagging quality function likes a logarithm function for any resource, receiving the same number of bookmarks, the tagging quality of a nearly stable resource increases less than for a far unstable resource [13].

Under this consideration, we propose two principles of designing the incentives. The basic principle is to reward tagging under-tagged resources, and the reward amount decreases with the stable state of a resource. This principle encourages users to annotate the most unstable resources. Another principle takes the precedence on the early tagging stage of a resource so as to accelerate a resource receiving cp number of bookmarks. It encourage users to bookmark the resources with fewest tags. So the reward function scales negatively with the number of bookmarks that a resource received.

To sum up the above arguments, we present a dynamic incentive function, as shown in Fig. 1. The whole tagging process is divided into three phases. In the first phase, namely before the $Critical\ Point(cp \in N^+)$, it is inappropriate to evaluate the quality of a bookmark. So a constant reward is granted to each user annotation. In

Table 1 Payoff matrix of incentive based tagging mechanism

Strategy	User	
	Casual-Tagging	*Serious-Tagging*
UIM	$(Q - C_1, \underline{C_1 - C_0})$	$(\widehat{Q} - C_1, C_1 - \widehat{C}_0)$
QIM		
Phase 1	$(Q - C_1, \underline{C_1 - C_0})$	$(\widehat{Q} - C_1, C_1 - \widehat{C}_0)$
Phase 2	$(\underline{Q - C_0}, C_0 - C_0)$	$(\underline{\widehat{Q} - A}, \underline{A - \widehat{C}_0})$

the second phase, namely between cp and the $Stable\ Point(sp \in N^+)$, a bookmark quality can be evaluated. The reward is set inversely proportional to a resource tagging state determined by $MA\ score$. To encourage user seriously bookmarking, the exact reward a user gets in practice is determined by the evaluation of his/her bookmarks (given in next section). The higher the quality of a bookmark, the more reward. A minimum reward is set so as not to detour user annotation. So, the reward a tagger gets is flexible within a range, shown as the shadow part in Fig. 1. After the stable point sp is the third phase and there is no reward.

We will discuss the following key points of the proposed incentive mechanism in the next section: how to evaluate the quality of a bookmark, how to balance the budget to different resources and how to set the rewards for different periods of a resource.

4.2 Mechanism comparison

In this subsection, we compare our mechanism with the existing incentive mechanisms from the perspective of game theory. The typical incentive mechanisms is presented by Yang et al. [13], which rewards a uniform reward for a user annotating an under-tagged resource. We denote it as *Uniform Incentive Mechanism* (*UIM* for short) in the following discussion. The purpose of this comparison is to theoretically understand how rational users behave under different incentive mechanisms. The payoff matrix is shown as Table 1.

The rows of this table indicate the strategies of a tagging system, *UIM* or *QIM*. While the columns show user strategies, *Casual-Tagging* or *Serious-Tagging*. Each cell contains both players' payoff under their strategy combination. The former value in each cell corresponds to the system payoff, and the latter is for the user. The objective of the game is to maximize their payoffs.

In the matrix, C_0 and $\widehat{C}_0$ denote different user costs in *Casual-Tagging* and *Serious-Tagging*, respectively. Obviously, $C_0 \leq \widehat{C}_0$ holds since a user needs less time and effort when tagging at discretion. C_1 represents the uniform reward in both UIM and the first phase of *QIM*, A represents the dynamic reward during the second phase of *QIM*. Then we have $A \leq C_1$ according to the *QIM* defined in previous subsection. It is reasonable to assume $A \geq \widehat{C}_0$, since a rational user needs a reward for annotation. Specially in the second phase, a basic reward should be assigned to a tagger so as to encourage annotation. For example, it can be set the estimate of an average user annotation cost C_0. The gain of a system is the improvement of resource tagging quality, denoted by $\widehat{Q}$ and Q under user *Serious-Tagging* and *Casual-Tagging*, respectively. Theoretically, $\widehat{Q} > Q$ holds since a user seriously tagging brings better results for system.

Since this is a *Stackelberg Game* [34], where a tagging system acts first and users take subsequent actions, we firstly study how a system chooses an incentive mechanism. We mark a player's higher payoff as an underline in the table. From the table, we can see there is no difference between *U I M* and the first phase of *QIM*. When a system adopts *U I M*, rational users prefer *Casual-Tagging* for a higher payoff. When a system chooses *QIM*, we assume that for the same user, a casual tagging user could not get any more reward besides the basic part C_0. So, in the second phase rational users prefer *Serious-Tagging* so as to get a higher payoff under the condition $A > \widehat{C}_0$. So a system should choose *QIM*. Now, we analyze user behavior under *QIM*. Since user payoff in *Phase* 1 is always larger than in *Phase* 2, rational users prefer to annotate a resource as early as possible. That is to say, rational taggers prefer tagging the most unstable resource so as to get a higher payoff. According to the game analysis [35], *QIM* is a dominant strategy for the system, and (*QIM*, *Serious-Tagging*) is the Nash Equilibrium. This illustrates why *QIM* can accelerate the process of a system approaching its stable tagging state.

5 Bookmark quality evaluation and reward

In this section, we first propose four methods to evaluate a bookmark quality. Then we present a dynamic incentive function to allocate reward for a bookmark. Next, we discuss the compensation on novel and popular tags, which are ignored in the instant evaluation of a bookmark quality. Finally, we discuss how to allocate a given budget among under-tagged resources.

5.1 The quantitative evaluation of bookmark quality

In this subsection, we propose four different quantitative methods for evaluating bookmark quality. The main idea of these methods is based on the "*crowd proof*" that testifies a bookmark against previous bookmarks from two perspectives: what tags in it and how popular they are. For a resource r_i, the quality of a bookmark $b_i(k)$ at the π_i^k *point* is denoted as $bq_i(k)$ in the following discussion, where $k \geq 1$.

1. **Tag hitting** ($\mathcal{TH}$) The tag hitting method evaluates a bookmark quality according to the intersection of $b_i(k)$ and the previous tag set $\mathcal{T}_i(k-1)$ for r_i at π_i^k *point*. It computes the ratio of intersection tags in $b_i(k)$. This ratio shows which assigned tags are consistent with previous bookmarks. Hence, the larger this ratio, the higher the bookmark quality. Formally,

$$bq_i(k) = \frac{|b_i(k) \cap \mathcal{T}_i(k-1)|}{|b_i(k)|} \tag{6}$$

2. **High-frequency hitting** ($\mathcal{HF}$) The $\mathcal{TH}$ method only takes into account the consisted tags in a bookmark without considering tag frequency, which loses the meaningful information on different importance of tags. The high frequency hitting method evaluates a bookmark quality by counting the tag frequencies in the intersection of a bookmark and its previous tag set. If a bookmark hits more top-

frequency tags, its quality is higher. Formally, given the relative tag frequency distribution of res_i at π_i^{k-1} $point$, i.e. $\vec{F}_i(k-1) = \{f_i(t_1, k-1), f_i(t_2, k-1), \ldots\}$, the $\mathcal{HF}$ method is computed as:

$$bq_i(k) = \sum_{t_j \in b_i(k) \cap \mathcal{T}_i(k-1)} f_i(t_j, k-1) \tag{7}$$

3. **Least effort** ($\mathcal{LE}$) In $\mathcal{HF}$ evaluation, we can not avoid the malicious $full\ cover$ assignments. That is to say, malicious tag assignments may include as much tags as one can so as to acquire a high bookmark quality. To avoid this, the evaluation of a bookmark quality should consider the size of a bookmark. According to the least effort criterion, the number of tags for identifying a resource should be minimized [28]. Hence, we introduce the least effort method which determines a bookmark quality by the average frequency of tags in the hitting set. Formally,

$$bq_i(k) = \frac{1}{|b_i(k)|} \sum_{t_j \in b_i(k) \cap \mathcal{T}_i(k-1)} f_i(t_j, k-1) \tag{8}$$

4. **Stability improvement** ($\mathcal{SI}$) Since the above methods only consider the tag intersection with previous bookmarks, they does not take a view of the whole tag set to evaluate how much a bookmark contributes to the stability of a resource tagging quality. So, we introduce the stability improvement method to evaluate a bookmark quality by the promotion of a resource tagging quality. Formally,

$$bq_i(k) = \begin{cases} 1 & m_i(k, \omega) - m_i(k-1, \omega) \geq \delta \\ 0 & otherwise \end{cases} \tag{9}$$

where $m_i(k, \omega)$ denotes the MA score of resource res_i and $\delta > 0 \in R^+$ is a threshold.

In summary, the above evaluation of a bookmark quality ranges from 0 to 1. The higher a bookmark quality, the larger the value. We will compare these methods in details in Sect. 6.3 and discuss their adaptation in practice.

5.2 Dynamic incentive reward

Having the quantitative evaluation of a bookmark quality, the subsequent question is how to allocate the reward. The core of the incentive is to accelerate the process of a system approaching the stable tagging state and to encourage taggers to annotate. According to the mechanism in Sect. 4.1, in $Phase$ 1, a constant C_1 money is rewarded to each bookmark. In $Phase$ 2, a dynamic incentive function is designed based on both bookmark quality and the tagging state of a resource. The less stable a resource's tagging state, the more reward for tagging. For any resource res_i at π_i^{k-1} $point$, a system computes an announced reward for the next bookmark, denoted as $A_i(k)$. It consists two part: a constant part C_0 as the basic encouragement for user tagging

and a variant part expressing how far a resource tagging quality need improve for $practically\text{-}stable$. Formally,

$$A_i(k) = \begin{cases} C_1 & k \leq cp\ (phase\ 1) \\ C_0 + (C_1 - C_0) \cdot \frac{\tau - m_i(k-1,\omega)}{\tau - m_i(cp,\omega)} & cp < k \leq sp\ (phase\ 2) \end{cases} \tag{10}$$

where ω is the window size in computing MA score, τ is a threshold for determining whether the tagging state is $stable$.

Based on $A_i(k)$, a tagger's actual reward is computed based on the quality $bq_i(k)$ of the assigned bookmark. For the k^{th} bookmark on resource res_i, we adopt $GA_i(k)$ to denote the actual reward a tagger can get, which is defined as follows:

$$GA_i(k) = A_i(k) \cdot bq_i(k) \tag{11}$$

5.3 Compensation for popular tag originator

Revisit the above evaluation methods, we notice two shortcomings. One is in the first phase that a tagger may casually annotate a resource due to a constant reward. Another is in the second phase that an originator of popular tag does not get any reward on this tag. Since all the evaluation methods are based on previous tag set, any novel tag is not taken into account for reward. To solicit novel and popular tags, we propose a compensation mechanism to remedy the above shortcomings, which resides on three points $which$, $when$ and how.

The first key point is $which$ new generated tags should be rewarded. Obviously, the compensation should be applied to useful tags rather than noisy ones. Justifying whether a tag is meaningful or not requires either $specialist$ verification or $crowd\ proof$. According to the spirit of crowdsourcing, we adopt the $crowd\ proof$ idea to verify the usefulness of a tag, namely by the popularity of tags. For example, the top ranked tags or high frequency tags in the final tag set can represent the popularity. The second key point is $when$ to reward. The $backtracking$ method is introduced and the evaluation is performed after a resource has been tagging stable. The third key point is how to reward. It should be the same with the adopted method in evaluating a bookmark quality so as to remain consistent along the whole incentive process. We adopt the same method to evaluate a bookmark against the stable tag set.

For example, for resource res_i, if we take the top ranked tag set $\widetilde{\mathcal{T}}_i^r$ as the popularity criterion and the $least\ effort$ $\mathcal{LE}$ method for bookmark quality evaluation, the compensation for a tag $t \in (b_i(k) \cap \widetilde{\mathcal{T}}_i^r) \wedge t \notin \mathcal{T}_i(k-1)$ is calculated as

$$cps_i(t) = C_2 * f_i(t) \tag{12}$$

where $f_i(t)$ is the relative frequency of t in $\hat{\varphi}_i(\omega, \tau)$, C_2 is a basic reward for compensation (e.g. $C_2 = C_1 - C_0$).

Another candidate method of compensation can be calculated as the difference of evaluating the same bookmark $b_i(k)$ at π_i^k $point$ and at final stable state. Formally,

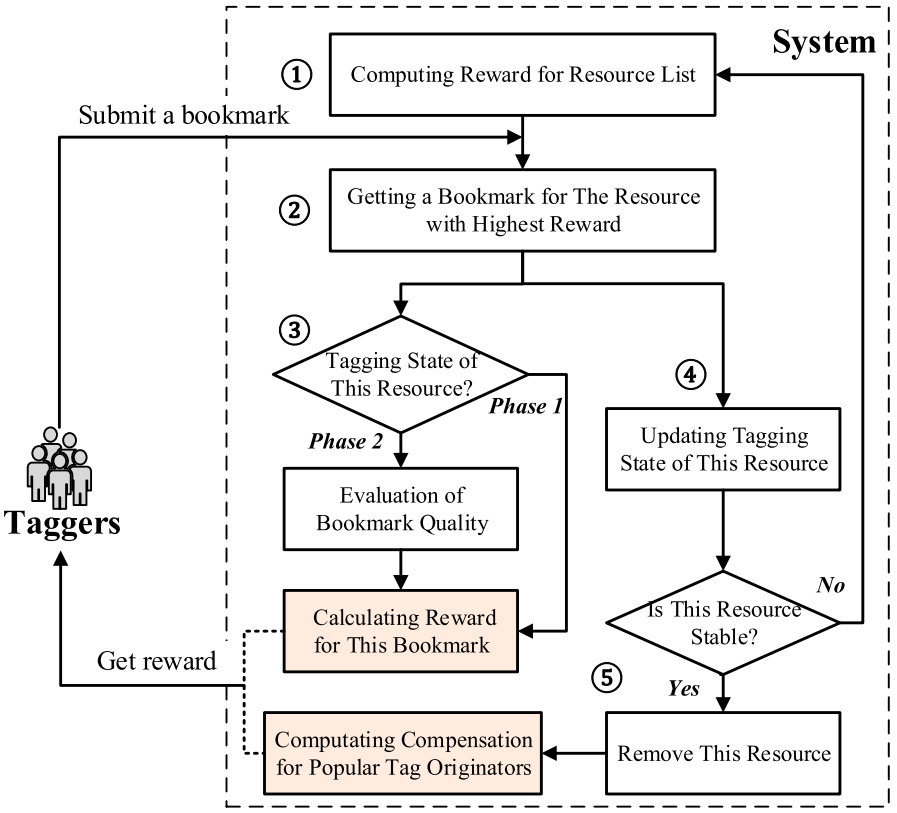


Fig. 2 Highest reward strategy

$$final_i(k) = \frac{1}{|b_i(k)|} \sum_{t_j \in b_i(k) \cap \widetilde{T}_i^r} f_i(t_j) \tag{13}$$

where $f_i(t_j)$ denotes the relative frequency of t_j in $\hat{\varphi}_i(\omega, \tau)$.

$$cps_i(t) = \begin{cases} C_1 * new_i(k) & Phase1 \\ max\{A_i(k) * (final_i(k) - bq_i(k)), 0\} & Phase2 \end{cases} \tag{14}$$

where $A_i(k)$ denotes the announced reward at π_i^k $point$ and $bq_i(k)$ is the quality of bookmark $b_i(k)$ at π_i^k $point$.

5.4 Incentive allocation strategy

According to the proposed incentive mechanism in Sect. 4.1, the amount of reward reflects the requirements of a system for the stable tagging purpose. So we propose the Highest Reward strategy (*HR*) for incentive tagging, shown in Fig. 2. It announces

Algorithm 1 The Bookmark-Quality Based Reward Allocation

Require: Under-tagged Resources $\mathcal{R} = \{res_1, \ldots, res_n\}$, Budget B,
1: Compute the reward $A_i(k_i)$ for each resource $res_i \in \mathcal{R}$
2: **while** ($B > 0$ & $\mathcal{R} \neq \emptyset$) **do**
3: Select $res_h \in \mathcal{R}$ with the highest reward
4: A tagger submits bookmark $b_h(k_h)$ for res_h
5: Compute the reward $GA_h(k_h)$ according to the Equation 11.
6: $B \leftarrow B - GA_h(k_h)$
7: Update tagging state of res_h
8: **if** tagging state of res_h is stable **then**
9: Compute compensation $cps_h(t)$ for popular tag originators
10: $B \leftarrow B - cps_h(t)$
11: $\mathcal{R} = \mathcal{R} - \{res_h\}$
12: **else**
13: Compute the reward $A_h(k_h + 1)$ for next bookmark $b_h(k_h + 1)$ on res_h
14: **end if**
15: **end while**

higher rewards on less stable resources and assigns a user resources in descending order on reward. When there are multiple resources with the same reward, it associates the precedence to the resource with least bookmarks. It contains the following steps and details are given in Algorithm 1.

Step 1 The system computes the provided reward $A_i(k)$ for each under-tagged resource res_i according to Eq. 10 and displays them in descending order of rewards.

Step 2 A tagger annotates the resource res_h with the highest reward with the bookmark $b_h(k)$.

Step 3 The system calculates the actual reward $GA_h(k)$ for $b_h(k)$ according to the Eq. 11.

Step 4 The system updates the tagging state of the annotated resource.

Step 5 If a resource gets stable tagging state, the system computes compensation for the popular tag originators.

6 Experimental study

6.1 Datasets and experiment settings

There are quite a few datasets about collaborative tagging systems, such as Flickr,[1] Bibsonomy[2] and Goodreads[3] etc. However, most of them lack the detailed process of tagging, such as the tagger and the time stamp of each bookmark. So they do not satisfy the testification requirements of incentive tagging. We select three publicly available sets of bookmarks in the experiments, which are $Last.FM$, $Del.icio.us$-2004 and $Del.icio.us$-2007 respectively. These datasets provide full information involved in

[1] http://www.flickr.com

[2] http://www.bibsonomy.org

[3] http://www.goodreads.com

Table 2 Dataset statistics

	Last.FM	*Del.icio.us*-2004	*Del.icio.us*-2007
Number of tags	5,519	62,326	87,489
Number of resources	41	759	5,000
Number of bookmarks	8,319	134,594	562,048
Tags per res.	135	82	18
Bookmarks per res.	203	177	112
Tags per bookmark	2	2	4

each bookmark, i.e. the time stamp, resource ID and the tags etc. *Last.FM* is a website that allows users to annotate their music collection. The *Last.FM* dataset is given by the HetRec 2011 workshop,[4] which contains 186,480 bookmarks from August 1, 2005 to May 9, 2011. *Del.icio.us* provides bookmarking web service for URLs. The *Del.icio.us*-2004 dataset includes 3,352,035 bookmarks, which is crawled by authors in [36] in Year 2004. Since these datasets initially contain many under-tagged resources, we select the stable resources for the purpose of testifying the effectiveness of our incentives on tagging quality improvement. The final used data includes 41 stable resources from the *Last.FM* dataset and 759 stable resources from the *Del.icio.us*-2004 dataset. Recently some websites introduce tag recommendation into collaborative tagging, in which users are suggested some previous used tags when annotating a resource. For example, the *Del.icio.us* system introduced tag recommendation in June 2005 [37]. This method to some extent restricts user choices. To justify whether and how much the tag recommendation influences the evaluation of tagging quality, we adopt another *Del.icio.us* dataset, which is collected from *Del.icio.us* in Year 2007 and provided by the authors of [13]. It contains a set of 5,000 stable resources, 87,489 tags and 562,048 bookmarks. This *Del.icio.us*-2007 dataset initially contains much more resources, from which the authors select 5,000 stable resources for the purpose of evaluating stable tagging state by setting $\omega = 5$ and $\tau = 0.9999$ [13]. The statistics of these filtered datasets are shown in Table 2.

All the programs are implemented in C++ and experiments are conducted on a computer with Intel Core i5 CPU (3.10 GHz), 4 GB memory and Linux system. In the following subsections, we first discuss how to choose a *Critical Point cp*. Then we compare the effectiveness, efficiency and adaptability of the proposed methods on bookmark quality evaluation. Finally, we compare the effectiveness of our mechanism with other existing incentive mechanisms.

6.2 Critical point selection

In our proposed *QIM* mechanism, the qualitative evaluation of bookmark quality is based on the collection of previous bookmarks. The following discussion will determine when the accumulated bookmarks are enough as an evidence for evaluating the

[4] http://ir.ii.uam.es/hetrec2011/datasets.html

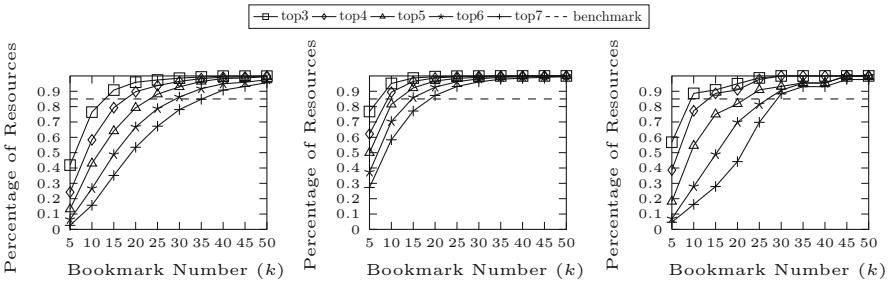


Fig. 3 Ratio of *top-r set* covered resources

posterior bookmarks. We adopt the *tag set* (see Definition 1) as the collection of bookmarks on a resource. By analyzing collaborative tagging systems, the authors in [28] find there are two typical properties of the tag set of each well tagged resource: the popularity of tags and multiple facets that tags can represent. Based on this observation and the discussion on tag usage in Sect. 3, we consider two representative aspects top-ranked tags and high frequency tags to evaluate a *tag set*. Since the *Del.icio.us* system introduced tag recommendation in June 2005 [37], we adopt *Del.icio.us*-2004 and *Del.icio.us*-2007 so as to make comparison on natural tagging and tag recommendation.

6.2.1 Top ranked tags

The stable tag set of resource res_i is denoted as $\widetilde{\mathcal{T}}_i$ (defined in Sect. 3). For a given integer $r \in N^+$, the *Top-r Set* of resource res_i (denoted as $\widetilde{\mathcal{T}}_i^r$) is defined as the subset of $\widetilde{\mathcal{T}}_i$, in which the frequencies of tags rank top r in $\widetilde{\mathcal{T}}_i$.

A *Top-r Set* actually is one of the representative characteristic of a stable tagging resource. Considering the purpose of tag based applications, only a part of tags are used [11]. That means only some top ranked tags are adopted to represent a resource. So, in the process of collaborative tagging, if the tag set of a resource res_i covers $\widetilde{\mathcal{T}}_i^r$, namely for some $k \in N^+$, $\mathcal{T}_i(k) \supseteq \widetilde{\mathcal{T}}_i^r$ holds, the tag set $\mathcal{T}_i(k)$ represent some main characteristic of res_i. That is to say after receiving k bookmarks, the tag set $\mathcal{T}_i(k)$ can be regarded as the evidence of evaluating the posterior bookmarks. This k setting is the semantically same with the concept *Critical Point* (cp) in Sect. 4.1.

To determine an appropriate cp, we evaluate for a fixed integer r how many bookmarks a resource need to receive so that its tag set covers *Top-r Set*, namely to determine the k setting such that $\mathcal{T}_i(k) \supseteq \widetilde{\mathcal{T}}_i^r$. Having the observation that a resource is often well described by about five top tags in My Web 2.0 [28], we set r range from 3 to 7 in the experiment. Then we compute the percentage of the *coverage resources* satisfying $\mathcal{T}_i(k) \supseteq \widetilde{\mathcal{T}}_i^r$. The results are shown in Fig. 3. The x-axis is the number of bookmarks (k) submitted to a resource and the y-axis indicates the percentage of *coverage resources* when they receive k bookmarks. For the convenience of comparison, we adopt 85 % as the benchmark to indicate that most resources satisfy the *Top-r Set* coverage requirement.

From Fig. 3, we observe that the ratio increases quickly with k and approaches 100 % before $k = 50$ on all three datasets. This is consist with the trend of a resource tagging quality. Considering different k, we notice that the smaller the value of r, the faster the percentage increases. For example, in Fig. 3a, when $r = 5$, the percentage is 96.43 % at $k = 35$ and increases to 98.68 % at $k = 45$. For $r = 7$, there are over 84 % resources cover their $Top\text{-}r\ Sets$ at $k = 35$ and this percentage is over 93 % at $k = 45$. This indicates for most resources, the tag set can be used as an evidence for bookmark evaluation when k ranges from 35 to 45. There are similar phenomena in Fig. 3b, c. There is more fluctuation in $Last.FM$ because it has much smaller number of stable resources than another two datasets. Considering the influence of tag recommendation, we compare Fig. 3a, b, and find that tag recommendation brings a quick convergence on tags than natural tagging. It requires a smaller k to cover the popularity characteristic of a tag set. For example, for the same $r = 7$, $Del.icio.us$-2007 requires an average $k = 30$ bookmarks to cover $\widetilde{\mathcal{T}}_i^r$, while $Del.icio.us$-2004 requires $k = 45$. This is because tag recommendation was introduced in the $Del.icio.us$-2007 dataset, which causes that users tend to annotate a resource with the most popular tags due to system recommendations.

6.2.2 High frequency tags

Tag frequency describes how much this tag is relevant to the resource on people consensus. The larger a tag frequency, the more likely the tag being used to identify the tagged resource [28]. For a given integer f, a $High\text{-}f\ Set$ for a resource is defined as the subset of its stable tag set $\widetilde{\mathcal{T}}_i$, denoted as $\widetilde{\mathcal{T}}_i^f$, in which the frequency of every tag is larger than f. Formally,

$$\widetilde{\mathcal{T}}_i^f = \{t | t \in \widetilde{\mathcal{T}}_i \cap h_i(t) > f\} \tag{15}$$

If the tag set of a resource covers its $High\text{-}f\ Set$, it could be as the evidence to evaluate other bookmarks, namely it covers the most important tags.

Figure 4 illustrates the relationship between the percentage of $High\text{-}f\ Set$ covered resources and the bookmark number. The x-axis is the number of bookmarks (k) submitted to a resource and the y-axis describes the percentage of resources satisfying

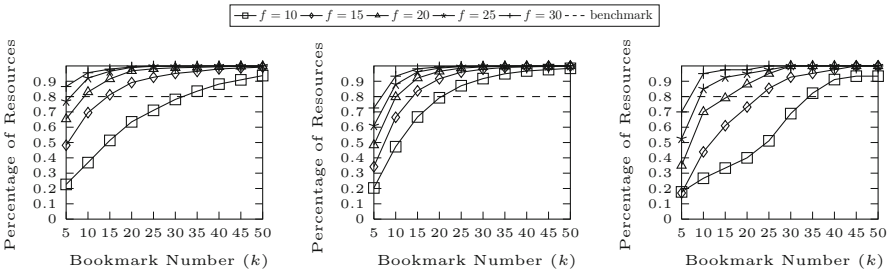


Fig. 4 Ratio of $high\text{-}f\ set$ covered resources

$\mathcal{T}_i(k) \supseteq \widetilde{\mathcal{T}}_i^f$. We adopt 80 % as the benchmark to indicate that most resources' tag sets have covered the $High\text{-}f\ Set$.

From this figure, we find that the coverage ratio increases with k and reaches 100 % soon on all three datasets. For instance in Fig. 4a, the ratio is 98.80 % at $k = 35$ and reaches 99.73 % at $k = 45$ when $f = 20$. Considering different k, a larger f brings faster ratio increase. Considering the influence of tag recommendation, as shown in Fig. 4a, b, we can see that the natural tagging in $Del.icio.us$-2004 requires more bookmarks than $Del.icio.us$-2007 to achieve the same ratio. For example, for the same $f = 30$, $Del.icio.us$-2004 requires an average $k = 45$ bookmarks to cover $\widetilde{\mathcal{T}}_i^f$, while $Del.icio.us$-2007 only requires $k = 30$.

Overall considering the above findings, we conclude that after receiving a number k of bookmarks, the tag sets of a resource can cover the main characteristics of its final stable tag set. This number k is called the $Critical\ Point(cp)$ in our mechanism. Considering different tagging modes, say natural tagging and tag recommendation based tagging, the $Critical\ Point$ is different. Generally, cp is lager in nature tagging than in rag recommendation. Statistically, an optional setting can be $cp \in [35, 45]$ for a natural tagging system and $cp \in [20, 30]$ for a tag recommendation based tagging system.

6.3 Comparison of the proposed bookmark quality evaluation methods

6.3.1 Comparison of evaluation effectiveness

The first experiment testifies the effectiveness of proposed evaluation methods by comparing the quality for the same bookmark $b_i(k)$ against the tag set at $\pi_i^k\ point$ and at final stable state. The purpose of this evaluation is verifying the suitability of our quantitative evaluation method. Theoretically, whether a bookmark is helpful for a resource should be justified by the final stable tag set, which is the core of $crowd\ proof$. However, it is impossible for an instant evaluation during the tagging process before a resource reaching stable. In practice, since we need to make an instant reward, the quality of a bookmark should be computed at the tagging point. To avoid the influence of tag recommendation, we only perform this experiment on the $Last.FM$ and $Del.icio.us$-2004 dataset. The results are shown in Fig. 5, where the x-axis represents the bookmark number, and the y-axis means the average of the kth bookmark quality for all the resources. The solid lines give the actual bookmark quality computed at $\pi_i^k\ point$, while the dashed lines show the quality computed by the same evaluation method against the final stable tag set, dented as $standard$.

Figure 5a, d are the results of the Tag Hitting method ($\mathcal{TH}$) on two datasets, respectively. Since the final stable tag set collects all tags a resource received and $\mathcal{TH}$ computes a bookmark quality as the intersection with the final tag set, the $standard$ quality is always 1. It is also easy to understand that the actual bookmark quality increases positively with the bookmark number k and the difference between the actual bookmark quality and $standard$ gradually becomes smaller. This reflects that the amount of collected tags increase with k. The comparison on the High-Frequency Hitting method ($\mathcal{HF}$) is shown in Fig. 5b, e. From these figures, we can see there is the same trend

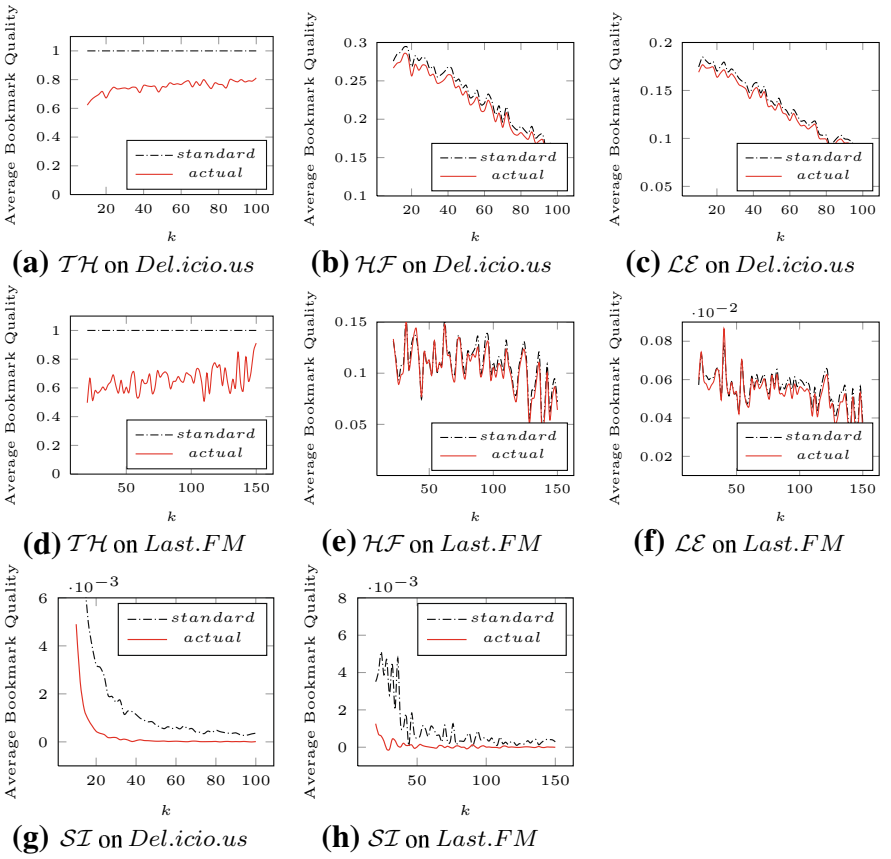


Fig. 5 Comparison results of effectiveness. k denotes the number of bookmarks

for both the actual bookmark quality and the *standard* and their difference becomes smaller with the increasing k. The results on the Least Effort method ($\mathcal{LE}$) are shown in Fig. 5c, f. Since $\mathcal{LE}$ evaluation is similar with $\mathcal{HF}$, the phenomenon is similar too. The difference is that its value is overall smaller than in $\mathcal{HF}$ because $\mathcal{LE}$ computes the average quality and $\mathcal{HF}$ only computes the sum.

Figure 5g, h depict the results on the Similarity Improvement method ($\mathcal{SI}$). Different with the above three methods, $\mathcal{SI}$ outputs a boolean value (refer to Eq. 9). To clearly reflect a bookmark quality, we adopt the difference of *MA* score between two adjacent bookmarks as the actual quality, i.e. $m_i(k, \omega) - m_i(k-1, \omega)$. This is consistent with the idea on quality improvement with $\mathcal{SI}$ definition. Likewise, the *standard* quality is computed as the difference of *tagging quality* between two adjacent bookmarks against the stable tag set. Formally,

$$q_i(k) - q_i(k-1) = sim\left(\vec{F}_i(k), \hat{\varphi}_i(\omega, \tau)\right) - sim\left(\vec{F}_i(k-1), \hat{\varphi}_i(\omega, \tau)\right) \quad (16)$$

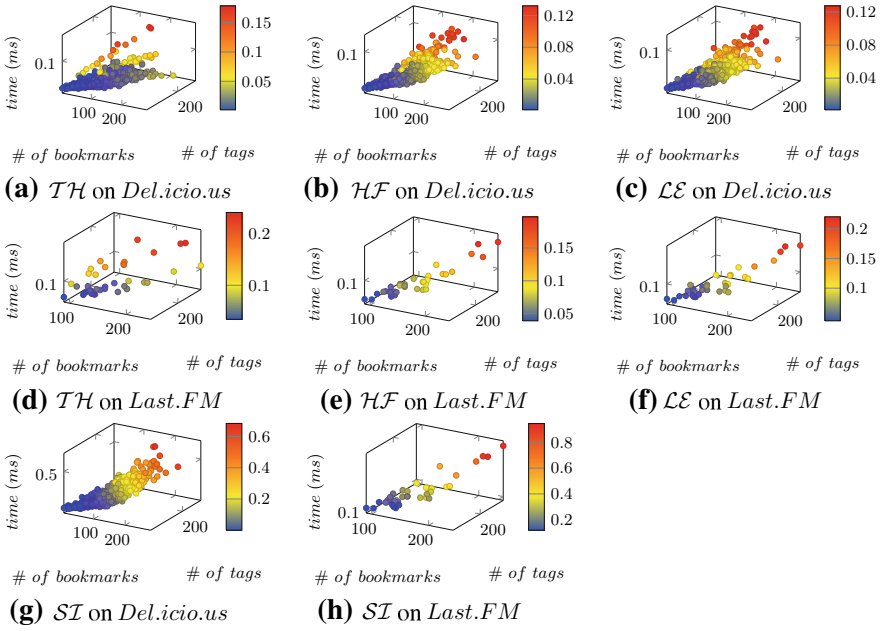


Fig. 6 Comparison results of efficiency

From these figures, we can also find that the difference between actual bookmark evaluation and *standard* gets smaller as resources receives more bookmarks. This is because that the trend of *MA* score is more and more similar with that of *Tagging quality* when the bookmark number increases.

In the above figures, there is common phenomenon that the results on $Last.FM$ fluctuate larger than in $Del.icio.us$-2004. This is because we adopt the average of bookmark quality. There are more resources in $Del.icio.us$-2004 than in $Last.FM$.

Overall, the small difference between the actual and standard quality in all cases well testify the effectiveness of the proposed evaluation methods.

6.3.2 Comparison of efficiency

This section evaluates the efficiency of these bookmark evaluation methods on the $Last.FM$ and $Del.icio.us$-2004 dataset. Figure 6 gives the experimental results, the x-axis is the number of bookmarks, the y-axis is the number of tags and the z-axis is the runtime. Each point in these figures denotes the overall runtime to evaluate all the bookmarks of a resource with a certain number of bookmarks and tags.

The results in figures show that all the evaluation methods are efficient and the runtime increases with the number of bookmarks or tags. For the same resources, the performance of four methods are different. The fastest method is the $\mathcal{TH}$ method and the slowest method is the $\mathcal{SI}$ method. This is because the computation of *MA* score

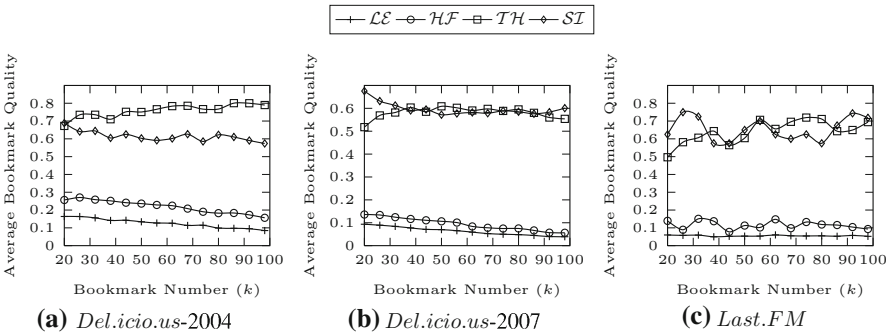


Fig. 7 Comparison of bookmark quality evaluation methods

in $\mathcal{SI}$ is complicated than others. The efficiency of both the $\mathcal{HF}$ method and the $\mathcal{LE}$ method are almost the same. There are similar phenomenons on both datasets.

6.3.3 Adaptation analysis

Generally, when choosing an evaluation method, we need to take into account three aspects: effectiveness, efficiency and value range. In the above subsections, we have analyze the first two aspects of the four methods. Here we make a further comparison of their value domains. The results are shown in Fig. 7. The x-axis is the number of bookmarks (k), which varies from 20 to 100, and the y-axis is the average bookmark quality for the kth bookmark for all the resources.

From this figure, we find that the value domains of these methods are different when evaluating the same bookmarks. The domains of $\mathcal{SI}$ and $\mathcal{TH}$ are higher than $\mathcal{HF}$ and $\mathcal{LE}$. For example, in Fig. 7a, the values for $\mathcal{TH}$ and $\mathcal{SI}$ ranges from 0.5 to 0.8, while the value interval for $\mathcal{HF}$ and $\mathcal{LE}$ is [0, 0.3]. Since our reward is based on the bookmark quality, the actual value should map for a reasonable interval, such as to apply a standardization to a unified range. Considering the influence of tag recommendation, there is less fluctuation of quality in $Del.icio.us$-2007. There is more fluctuation in $Last.FM$ because it has much smaller number of stable resources than another two datasets.

To take into account the effectiveness and efficiency of these methods, $\mathcal{HF}$ and $\mathcal{LE}$ are much more efficient than $\mathcal{SI}$ while it has a relatively equal effectiveness. So, it is not necessary to combine them together. Comparing with $\mathcal{TH}$, $\mathcal{HF}$ and $\mathcal{LE}$ have much better results. So, we conclude $\mathcal{HF}$ and $\mathcal{LE}$ are appropriate for most cases. Since $\mathcal{HF}$ and $\mathcal{LE}$ are the evolution of $\mathcal{TH}$, there is no need to combine them together.

6.4 Incentive effects

In order to compare with the most related work [13], we adopt the same dataset, i.e., $Del.icio.us$-2007. We choose two strategies with best results from their work: the Fewest First strategy (*FP*) and the Hybrid strategy (*FP-MU*). The *FP* strategy chooses the resource that has received the fewest bookmarks and allocates to taggers

as the next bookmarking task. The *FP-MU* strategy is a combination of *FP* strategy and the Most Unstable First strategy (*MU*). It uses *FP* strategy to allocate bookmarking tasks if resources have not received at least ω bookmarks. Otherwise, it chooses the resource with the smallest *MA* score.

Our incentive allocation strategy is the Highest Reward strategy (*HR*), as given in Sect. 5.4. In the following experiments, we set $\omega = 5$ and $\tau = 0.9999$ to determine whether the tagging state of a resource is stable. According to the discussion in Sect. 6.2, we set the $Critical\ Point\ cp = 25$. Furthermore, we adopt the $\mathcal{HF}$ to evaluate the bookmark quality and set $C_1 = 1, C_0 = 0.4$.

To evaluate the $tagging\ quality$ of a resource res_i at $\pi_i^k\ point$, we adopt rfd similarity between $\pi_i^k\ point$ and $Stable\ Point\ sp$. Formally,

$$q_i(k) = sim\left(\vec{F}_i(k), \hat{\varphi}_i(\omega, \tau)\right) \tag{17}$$

where $\hat{\varphi}_i(\omega, \tau)$ denotes the rfd at sp. The $tagging\ quality$ of a system is defined as the average tagging quality of all the resources $\mathcal{R}$ in the system, denoted as following.

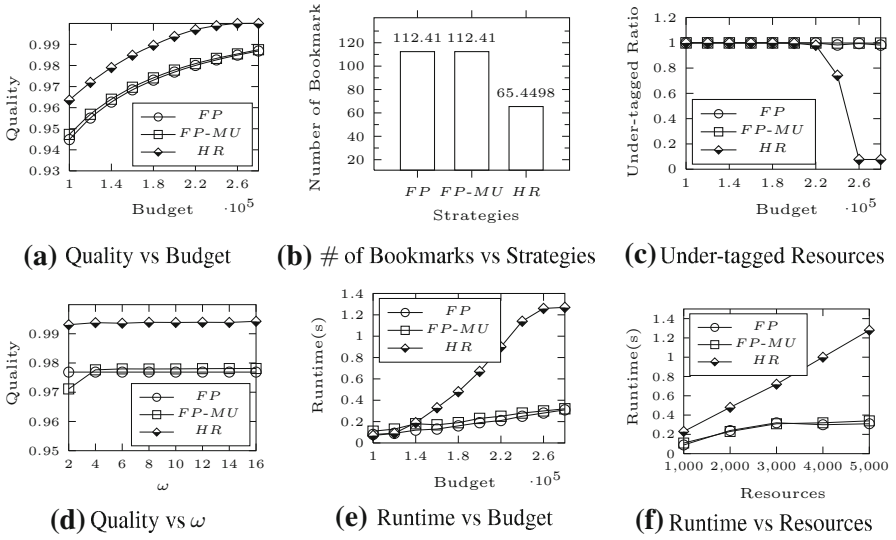
$$Q\left(\mathcal{R}, \vec{k}\right) = \frac{1}{n}\sum_{i=1}^{n} q_i(k_i) \tag{18}$$

6.4.1 Tagging Quality versus Budget

First, we evaluate how different strategies affect the system tagging status under a fixed budget. The budget scales from 100,000 to 280,000 in the experiment. The comparison results are shown in Fig. 8a. In this figure, the x-axis denotes the budget that we can used to reward users when they annotate on resources, and the y-axis is the tagging quality of the system (refer to Eq. 18). Our strategy *HR* is represented with diamond mark, while *FP* with circle mark and *FP-MU* with square mark. From this figure, we see that all the three strategies improve the tagging quality as the budget increases and eventually achieve stability. Moreover, *HR* always edges over *FP* and *FP-MU* significantly. For example, when the budget $B = 200{,}000$, *FP* and *FP-MU* improve the quality to 98.69 and 97.79 % while *HR* improves the quality to 99.38 %, which is 0.69 % more. This shows that our strategy further improve the tagging quality compared with other incentive strategies under a fixed budget. Besides, our strategy requires a smaller budget for achieving an expected tagging stability. For instance, *HR* costs 140,000 to improve the tagging quality to 97.986 % while *FP* and *FP-MU* have to spend 220,000 to achieve the same quality.

6.4.2 Tagging quality versus bookmark number

We perform the experiment to evaluate the impact of bookmark number on tagging quality. Figure 8b shows the average number of bookmarks when the tagging state of resources achieve stable. We observe that for both *FP* and *FP-MU*, the average bookmark number is 112, while *HR* requires only 64 bookmarks. This illustrates that the reward improves the received bookmark quality. Or we can say that the system only



(a) Quality vs Budget **(b)** # of Bookmarks vs Strategies **(c)** Under-tagged Resources

(d) Quality vs ω **(e)** Runtime vs Budget **(f)** Runtime vs Resources

Fig. 8 Experiment results

rewards the good-quality bookmarks. Since less bookmarks are needed for a resource to get stable tagging state under *HR* strategy, it means that our strategy can speed the process for a system to achieve tagging stable.

6.4.3 Under-tagging

We further evaluate how the budget affects the under-tagged resources. Figure 8c shows the percentage of resources that are under-tagged after performing a certain budget of bookmarking tasks. Overall, the percentage of under-tagged resources drops as the budget increases. *HR* sharply reduces the under-tagged percentage to 7.7 % when B is 200,000–260,000, while the percentage only drops to 1.78 and 0.22 % when B is 280,000 under *FP* and *FP-MU*, respectively. This phenomenon is consistent with our goal of improving the average tagging quality of all the resources. The tagging quality of a resource increasing with bookmarks is similar to a logarithm function [13]. When allocating a fixed amount of bookmarking tasks, a resource receiving a small number of bookmarks will get a greater quality improvement than other resources with lots of bookmarks. Hence the *HR* always chooses the resource with highest reward instead of selecting resources that almost get stable tagging state, which results in the nearly synchronous enhancement of tagging quality for each resource. Once *HR* has improved the tagging state of all under-tagged resources to be almost stable, any additional bookmarking tasks will carry these under-tagged resources over their stable points and sharply drop in the under-tagged resource percentage. Theoretically, having enough money, the under-tagged percentage should drop to 0. For *FP* and *FP-MU*, the percentage of under-tagged resources is still high even when the budget $B = 260{,}000$. This illustrates that these two strategies need more budget for the resources achieving stable tagging state.

6.4.4 Effect of ω

Both *FP-MU* and *HR* use ω to control the moving average window size. Figure 8d shows the effect of ω on *FP*, *FP-MU* and *HR* with budget $B = 200{,}000$. The results show that our method has better effectiveness for all values of ω. For *FP-MU* strategy, it adopts *FP* strategy before each resource receives ω bookmarks. Since each resource can only receive 8 bookmarks under budget $B = 200{,}000$ for 5,000 resources, the *FP-MU* always operates as *FP* when $\omega > 8$. For our strategy, the tagging quality is not affected by ω since in $Phase\ 1$ we do not need to compute tagging quality and the ω setting is always smaller than cp.

6.4.5 Efficiency

We evaluate the computational performance required by each strategy. Figure 8e shows the performance of these strategies for 5,000 resources under different budgets. In these experiments, *HR*'s running time increases with budget and remains stable when $B = 260{,}000$, while *FP*'s and *FP-MU*'s running time keep increasing. Since *HR* has to calculate the bookmark quality and update the maximum reward for the resource in $Phase\ 2$, its running time is a little longer than *FP* and *FP-MU*. However, when the tagging states of all resources are in $Phase\ 2$, the running time tends to stability. Figure 8f shows the performance scales with the number of resources with $B = 280{,}000$. Overall, the running time of *HR* is acceptable for all cases.

6.4.6 Summary

Based on the above experimental results, we conclude that our strategy *HR* is more effective than *FP* and *FP-MU* although it requires a little more time. For a fixed budget, *HR* makes a higher tagging quality of a system. For an expected state of tagging quality, it requires a lower budget. At the same time, *HR* accelerates the process of a system approaching stable than previous methods. Overall, our method is applicable for real tagging systems.

7 Conclusion

In collaborative tagging systems, tags can be used to categorize and manage online resources. The tagging quality of resources are fundamental to the availability and validity of these systems. In this paper, to encourage taggers to seriously annotate under-tagged resources, we propose a quality-based dynamic incentive mechanism. Four quantitative methods are proposed to evaluate a bookmark quality and the compensation is discussed so as to solicit popular and novel tags. Accordingly, a budget allocation strategy is proposed to balance a given budget among different resources in a system which accelerates the process of the system approaching tagging stable. Both the theoretical analysis by game theory and experiments on real datasets indicate that our method is more effective than previous works. For a fixed budget, our method makes a higher tagging quality of resources in the whole system, while for an expected

state of system tagging quality, it costs less. Furthermore, it requires less time for a collaborative tagging system to achieve its stable tagging status. In future, we will design personalized quantitative evaluation methods for bookmark quality according to different system requirements. Another future work is to analyze user behavior in tagging systems for better tag selection.

Acknowledgments The authors would like to thank the anonymous reviewers for their valuable comments and suggestions to improve the quality of this paper. We are also grateful to Dr. Reynold Cheng, from Department of Computer Science in the University of Hong Kong, for his comments and providing datasets. This work is supported by the National Natural Science Foundation of China (61173140), the National Science & Technology Pillar Program (2012BAF10B03-3), Special Program on Independent Innovation & Achievements Transformation of Shandong Province (2014ZZCX03301) and Science & Technology Development Program of Shandong Province (2014GGX101046).

References

1. Kang, R., Fu, W.T., Kannampallil, T.G. : Exploiting knowledge-in-the-head and knowledge-in-the-social-web: effects of domain expertise on exploratory search in individual and social search environments. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 393–402 (2010)
2. Sen, S., Harper, F.M., LaPitz, A., Riedl, J.: The quest for quality tags. In: Proceedings of the 2007 International ACM Conference on Supporting Group Work, pp. 361–370. ACM (2007)
3. Sen, S., Vig, J., Riedl, J.: Tagommenders: Connecting users to items through tags. In: Proceedings of the 18th International Conference on World Wide Web, pp. 671–680. ACM (2009)
4. Zubiaga, A., Fresno, V., Martinez, R., García-Plaza, A.P.: Harnessing folksonomies to produce a social classification of resources. IEEE Trans. Knowl. Data Eng. **25**(8), 1801–1813 (2013)
5. Wu, L., Yang, L., Yu, N., Hua, X.S.: Learning to tag. In: Proceedings of the 18th International Conference on World Wide Web, pp. 361–370. ACM (2009)
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 (2007)
7. Marchetti, A., Tesconi, M., Ronzano, F., Rosella, M., Minutoli, S: Semkey: a semantic collaborative tagging system. In: Proceedings of the Workshop on Tagging and Metadata for Social Information Organization at WWW, vol. 7, pp. 8–12 (2007)
8. Wetzker, R., Zimmermann, C., Bauckhage, C.: Analyzing social bookmarking systems: a del.icio.us cookbook. In: Proceedings of the ECAI 2008 Mining Social Data Workshop, pp. 26–30 (2008)
9. 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)
10. Zubiaga, A., Körner, C., Strohmaier, M.: Tags vs shelves: from social tagging to social classification. In: Proceedings of the 22nd ACM Conference on Hypertext and Hypermedia, pp. 93–102. ACM (2011)
11. Golder, S.A., Huberman, B.A.: Usage patterns of collaborative tagging systems. J. Inform. Sci. **32**(2), 198–208 (2006)
12. Körner, C., Kern, R., Grahsl, H.-P., Strohmaier, M.: Of categorizers and describers: An evaluation of quantitative measures for tagging motivation. In: Proceedings of the 21st ACM Conference on Hypertext and Hypermedia, pp. 157–166. ACM (2010)
13. Yang, X.S., Cheng, R., Mo, L., Kao, B., Cheung, D.W.: On incentive-based tagging. In: Proceedings of the 2013 IEEE International Conference on Data Engineering Data Engineering (ICDE). pp. 685–696. IEEE (2013)
14. Kipp, M.E.I.: Convergence and divergence in tagging systems: an examination of tagging practices over a four year period. In: Proceedings of the American Society for Information Science and Technology, vol. 47(1), pp. 1–2 (2010)
15. Hope, G., Wang, T.G., Barkataki, S.: Convergence of web 2.0 and semantic web: a semantic tagging and searching system for creating and searching blogs. In: Proceedings of the International Conference on Semantic Computing, pp. 201–208. IEEE Computer Society (2007)
16. Lin, X., Beaudoin, J.E., Bui, Y., Desai, K.: Exploring characteristics of social classification. Adv. Classif. Res. Online **17**(1), 1–19 (2006)

17. Sood, S., Owsley, S., Hammond, K.J, Birnbaum, L.: TagAssist: automatic tag suggestion for blog posts. In: Proceedings of the International Conference on Weblogs and Social Media (2007)
18. Li, X., Guo, L., Zhao, Y.E.: Tag-based social interest discovery. In: Proceedings of the 17th international conference on World Wide Web, pp. 675–684. ACM (2008)
19. 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 IEEE 27th International Conference on Data Engineering (ICDE), pp. 27–38. IEEE (2011)
20. Wu, H., Zubair, M., Maly, K.: Harvesting social knowledge from folksonomies. In: Proceedings of the Seventeenth Conference on Hypertext and Hypermedia, pp. 111–114. ACM (2006)
21. Lamere, P.: Social tagging and music information retrieval. J. New Music Res. **37**(2), 101–114 (2008)
22. Zeng, D., Li, H.: How useful are tags? an empirical analysis of collaborative tagging for web page recommendation. In Intelligence and Security Informatics, pp. 320–330. Springer, Berlin (2008)
23. Heymann, P., Ramage, D., Garcia-Molina, H.: Social tag prediction. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 531–538. ACM (2008)
24. Sigurbjörnsson, B., Van Zwol, R.: Flickr tag recommendation based on collective knowledge. In: Proceedings of the 17th International Conference on World Wide Web, pp. 327–336. ACM (2008)
25. Du, W.H., Rau, J.W., Huang, J.W., Chen, Y.S.: Improving the quality of tags using state transition on progressive image search and recommendation system. In: Proceedings of IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 3233–3238. IEEE (2012)
26. Chen, H.-M., Chang, M.-H., Chang, P.-C., Tien, M.-C., Hsu, W.H., Wu, J.-Li.: SheepDog: group and tag recommendation for flickr photos by automatic search-based learning. In: Proceedings of the 16th ACM International Conference on Multimedia, pp. 737–740. ACM (2008)
27. Krestel, R., Fankhauser, P.: Personalized topic-based tag recommendation. Neurocomputing **76**(1), 61–70 (2012)
28. Xu, Z., Fu, Y., Mao, J., Su, D.: Towards the semantic web: collaborative tag suggestions. In: Proceedings of Collaborative Web Tagging Workshop at WWW (2006)
29. Majid, A., Khusro, S., Rauf, A.: Semantics in social tagging systems: a review. In: Proceedings of the International Conference on Computer Networks and Information Technology (ICCNIT), 191–203. IEEE (2011)
30. Godoy, D., Rodriguez, G., Scavuzzo, F.: Leveraging semantic similarity for folksonomy-based recommendation. IEEE Internet Comput. **18**(1), 1 (2013)
31. Mo, L., Cheng, R., Kao, B., Yang, X.S., Ren, C., Lei, S., Cheung, D.W., Lo, E.: Optimizing plurality for human intelligence tasks. In: Proceedings of the 22nd ACM International Conference on Information and Knowledge Management, pp. 1929–1938. ACM (2013)
32. Lei, S., Yang, X.S., Mo, L., Maniu, S., Cheng, R.: iTag: incentive-based tagging. In: Proceedings of the IEEE 30th International Conference on Data Engineering (ICDE), pp. 1186–1189. IEEE (2014)
33. Weng, L., Menczer, F.: GiveALink tagging game: an incentive for social annotation. In: Proceedings of the acm sigkdd Workshop on Human Computation, pp. 26–29. ACM (2010)
34. Shokri, R., Theodorakopoulos, G., Troncoso, C., Hubaux, J.-P., Le Boudec, J.-Y.: Protecting location privacy: optimal strategy against localization attacks. In: Proceedings of the 2012 ACM Conference on Computer and Communications Security, pp. 617–627. ACM (2012)
35. Squicciarini, A.C., Griffin, C., Sundareswaran, S.: Towards a game theoretical model for identity validation in social network sites. In: Proceedings of the International Conference on Social Computing (Socialcom), pp. 1081–1088. IEEE (2011)
36. Görlitz, O., Sizov, S., Staab, S.: PINTS: peer-to-peer Infrastructure for Tagging Systems. In: Proceedings of the 7th International Conference on Peer-to-peer Systems, p. 19 (2008)
37. Jäschke, R., Marinho, L., Hotho, A., Schmidt-Thieme, L., Stumme, G.: Tag recommendations in social bookmarking systems. Ai Commun. **21**(4), 231–247 (2008)