Improving the Accuracy and Efficiency of Tag Recommendation System by Applying Hybrid Methods

Ali Kohi
Department of Software Engineering
Mashhad Branch - Islamic Azad University
Mashhad, Iran
Ali.kohi@gmail.com

Seyed Javad Ebrahimi
Department of Mathematical Sciences
Mashhad Branch - Islamic Azad University
Mashhad, Iran
Sj.ebrahimi@gmail.com

Mehrdad Jalali
Department of Software Engineering
Mashhad Branch - Islamic Azad University
Mashhad, Iran
mehrdadjalali@ieee.org

Abstract— Recently applications of social tagging systems have increased. These systems allow users to organize, manage and search the required resource freely, thus by combination and integration of recommendation systems in social software, assisting users to appropriately assign tag to resources and try to improve annotation among users. The challenges of recommendation systems are large-scale data, inconsistence data, usage of time-consuming machine learning algorithms, long and unreasonable time of recommendation and not being scalable to the demands of real world applications. Recently more efforts have been conducted to solve these problems. In this paper we proposed a tag recommendation system that is able to work with large-scale data and being applied in real world. The proposed system’s evaluation performed on a dataset collected from Delicious.com. The results demonstrated the efficiency and accuracy of proposed system.

Keywords: collaborative-based; collaborative tagging system; folksonomies; recommendation system; social tagging system; tag recommender

I. INTRODUCTION

Daily increasing of information in web and being unclassified caused difficulties in accessing to information. Therefore web directories were built to classify information but due to predefined structure of web directory, it forced users to follow this structure to classify information and made structure to lose its importance. By using new web technology called web 2.0, a new method of organizing and classifying web content called social tagging systems was made. Since social tagging systems allow users to organize, manage and search their required resource freely its applications have increased.

Due to being able to assign tag to resources freely, the popularity of these systems has increased and one of the advantages of applying these systems is resource classification according to users’ knowledge without any cost. Arbitrary assigning tag by users produces a large amount of data and being disintegrated in tag assignment. In order to appropriately assign tag to a resource to correctly accessing a resource the needs of a recommendation system was felt. Delicious.com is a popular social tagging system in which users able to annotate, search and share web addresses. In recent years researcher remarkably focused on tag properties which produced by users such as ambiguous and synonym words and analyzed them to discover the interrelations [1].

Recently many works were done in order to discover knowledge based on tags which result in creating user model in [2, 3, 4], tag prediction [5, 6], semantic web search [7] and extracting social networks [8]. Furthermore in these years tag recommendation systems made much progress [9, 10]. The challenges of recommendation systems are large-scale data, inconsistence data, usage of time-consuming machine learning algorithms, long and unreasonable time of recommendation and not being scalable to the demands of real world applications. In this paper the proposed tag recommendation system assists user to properly assign tag to a resource and includes the following steps: extracting useful information of web content, finding related words, discovering the relation between pages and users and afterwards combining the result in order to integrate the recommendation outcome. Experimental results demonstrated accuracy, efficiency and faster respond of the proposed system.

This paper organized as following: section 2 analyzes related work, while section 3 introduces data model. In section 4 introduces the proposed system and its modules. In section 5 evaluation of information retrieval metrics present, then dataset and experimental results discuss in section 6 and finally in section 7 we present conclusion.

II. RELATED WORK

Usually Tag recommendation systems divided into three classes: collaborative-based, content-based and graph-based approaches [11]. Methods of tag recommendation based on content, extracted useful words from resources using information retrieval techniques [12, 13]. For instance in [14] using TF/IDF technique to exploit top three terms for automated tag recommendation. Another method used Bayesian classification and exploit appropriate terms from content [15]. Some systems apply ontology to suggest tags. In [16] words extracted from documents then applying ontology to suggest abstract and conceptual tags.

Collaborative methods for tag recommendation are similar to collaborative filtering [17]. These systems found a community behavior to recommend tags. Collaborative methods have 3 steps: a tool to find similar resources and extracted its tags, a method to combine tags then ranked
them and afterwards suggesting top tags to user. This model was employed and implemented for automated tag recommendation systems in [18]. In this system, according to a sent post, tags associated with similar posts extracted and recommended. An adaptation of K-nearest neighbor method was employed to compose user information, tags and resources to suggest tag to user [19]. In this method users which assigned the same tags to similar resources considered as a neighbor and extracted tags according to relation between users and resources result in better outcome.

Graph-based methods apply PageRank algorithm. In [20, 21] FolkRank algorithm which is based on PageRank algorithm used to recommend tag. In FolkRank algorithm social tagging system is modeled as a triple graph including documents, users and tags. Each graph node presented a user and its weight computed according to its relation with other nodes. FolkRank algorithm has an acceptable output [22, 23] but since this algorithm has complicated computation and being time consuming, it could not utilize large scale data [24]. Since data in social tagging systems has relations among its features including users, documents and tags, Fig. 1, tensor factorization employed to transform data. In [25, 26] for decomposing information and transforming three dimensional space into three features and a tensor core, tensor factorization was used. It has less computation than graph based methods in recommendation step but the process of building model is not scalable and cannot be utilized in real world [24]. In [27, 22] pair-wise interaction tensor factorization presented which gave a rank to each tag and due to existence of user-resource pair in training data has an acceptable run time in both model building and recommendation. In other papers several methods composed to reduce their drawbacks and improve their quality. In [28] the proposed system suggested tags according to a combination of extracted information from web page content (i.e. title and URL), similar resources’ tags with respect to specific resource and relevant tags to user profile. According to extracted information from resource content (i.e. title, URL, Meta keyword and Meta description), weighted them and composing this information with assigned user’s tags and its neighbors, system recommended tags [11]. To recommend a set of tags this system sequentially used content-based, user profile and collaborative-based sets i.e. next set is used if the previous set will be empty. In [24] a combination of six methods based on different models in social tagging system was used. In this method because of a compound technique the quality of results improved but it is time consuming and suffer from cold-start problem thereby it cannot be utilized in real world.

In brief most of the mentioned methods and techniques try to improve the result and solving the problems. Unfortunately because of being time consuming and having complicated computation, they are not able to cover all dimensions (user, resource and tag) of social tagging system as well as not being able to be utilized in real world.

In this paper we proposed a new recommendation system which can be used in real world and covering all dimensions of social tagging systems.

III. DATA MODEL

At first data model and problem assumptions in real world presented.

The foundation of a social tagging system is the annotation. An annotation contains a user, resource and all tags the user applied to the resource. A collection of annotations results in a complex network of interrelated users, resources and tags (Fig. 1) [29].

Supposed that the social tagging system in real world using a structure the same as Fig. 2 to store information which presents relational tables in database.

The numeric values are assigned to Users and pages and stored them in this structure. The middle table which is the most important part of this structure appropriately stored the relation among users, pages, tags and their information. According to this relationship among tables variant report having any interrelation among users, resources and tags can be extracted simply, thus accessing to information is a lot faster and has less complication. Furthermore no information sacrifices because of three dimensional space decomposition [25, 26].

IV. THE PROPOSED SYSTEM (TAG RECOMMENDATION SYSTEM BY APPLYING HYBRID METHODS – TRS-HM)

The architecture of TRS-HM depicted in Fig. 3. The proposed system includes the following modules: Web Page Crawler, Related Word, Extracted Tag, User Popular Tags and Web Page(s), Related Web Page(s) Tags, Recommendation Engine. Due to combination and integration methods of content-based, collaborative-based, related words and user profile, the proposed system does not suffer from cold-start problem. This problem solved using Web Page Crawler and Related Word modules. Also all information that was used only extracted from employed database. Each module discuss in the following.

A. Web Page Crawler

Web Page Crawler objective is solving the cold-start problem using all useful extracted words from content of a specific page.

Figure 1. Structure of social tagging system

Figure 2. Database Structure to store the information of users, resources and tags
Figure 3. The architecture of the proposed system

User interface sent web page address to Crawler module then it downloaded the specified page and analyzed its textual information by applying given weight in Table I. The order of different part of an html page in this table is according to [32]. By applying the given weights to each part according to their importance in search engines and normalizing them to range [0..1] their values obtained and presented in Table I.

In order to speed up only textual information extracted. In addition to Table I, Crawler module has a list of inappropriate words which called black list. This list includes words such as ‘the’, ‘too’, ‘either’ and etc. If this module encounters these words it would reject and prevent them as output. The weight of extracted word is defined as:

\[ W = \frac{N_t \cdot C_t}{TN_t} \]  

(1)

Which \( W \) is the obtained weight, \( N_t \) is the number of word occurrence in text content, \( C_t \) is the given weight of word according to Table I and \( TN_t \) is the total number of extracted words from text content.

After extracting words, Crawler applies the following filtering:
1) Eliminating of words which their weight according to (1) is less than threshold \( \alpha \).
2) Eliminating of words which have less character than \( \beta \).
3) Words which are only digit eliminated.
4) Descending sort due to their weight then top \( \mu \) words selected.

According to conducted experiments on parameters \( \alpha \), \( \beta \) and \( \mu \), the most proper values for these parameters calculated and presented in Table II. Since most of the words in Black List have less than three characters, the value of \( \beta \) is three.

B. Related Word

This module objective is finding the related words according to the output of the Web Page Crawler module. For instance Web Page Crawler module extracted a set of words such as ‘sale’, ‘buy’, ‘material’ and ‘cost’ from the content of a store web page. Related Word module accordingly performed and finds words such as ‘store’, ‘shop’, ‘salesroom’ and etc. A set of extracted words from this module assists collaborative based methods to find a better relation among tags. Related words in this module were extracted using a composition of gained information from three databases. In this paper we used WordNet, GoRank.com and WordsLike.net databases which are relevant to ontology or related words.

The following filtering has been applies on extracted words:
1) Removing Words that have less character than threshold \( \beta \).
2) Removing Words which are only digit.
3) If a word from Web Page Crawler module has less related words than threshold \( \delta \), it is removed.

This filtering has been performed on all extracted words from previous module. In this module the value of parameter \( \beta \) is the same as the previous module but According to conducted experiments on parameter \( \delta \) the most proper value for this parameter computed and presented in Table III. This module causes false-positive extracted words from Web Page Crawler module eliminated which result in more accurate final output.

C. Extracted Tag

This module combines extracted words from last two modules. The extracted set called Combinations of Web Page Crawler and Related Word Modules (CWRM).

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \alpha )</td>
<td>4</td>
</tr>
<tr>
<td>( \beta )</td>
<td>3</td>
</tr>
<tr>
<td>( \mu )</td>
<td>10</td>
</tr>
</tbody>
</table>

TABLE I. THE OBTAINED WEIGHT FOR DIFFERENT PART OF AN HTML WEB PAGE

<table>
<thead>
<tr>
<th>Item</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>0.36</td>
</tr>
<tr>
<td>H1 Tag</td>
<td>0.22</td>
</tr>
<tr>
<td>Text(Underline , Italic , Bold)</td>
<td>0.11</td>
</tr>
<tr>
<td>Body</td>
<td>0.09</td>
</tr>
<tr>
<td>Anchor Tag</td>
<td>0.08</td>
</tr>
<tr>
<td>Image Text</td>
<td>0.06</td>
</tr>
<tr>
<td>URL</td>
<td>0.04</td>
</tr>
<tr>
<td>Heading</td>
<td>0.03</td>
</tr>
<tr>
<td>Meta Keyword</td>
<td>0.01</td>
</tr>
</tbody>
</table>

TABLE II. WEB PAGE CRAWLER MODULE PARAMETERS
The aim of this module is selecting a set of tags which are related to specific page. This set assists the collaborative-based methods to improve the performance of searching and also has the added benefits of solving the cold-start problem.

D. User Popular Tags and Web Page(s)

This module using user ID and CWRM set and performing the following steps that divided into several parts:

1) a) pages from user pages set are selected whose assigned tags existed (at least one) in CWRM set.
   
   2) b) X1 tags which are relevant to extracted pages in previous part from all users whose have most occurrences are selected and called TR1.

2) a) Selecting X2 top pages from a descending list of user pages which is sorted according to the number of assigned tags to page.
   
   3) b) users whose set of pages existed in extracted set of previous part are selected.

3) a) X3 tags which have most occurrences from assigned tags to pages in previous part were selected and called TR2.
   
   4) b) X4 pages from user pages set whose assigned tags existed (at least one) in CWRM set and have most assigned tags were selected.
   
   5) b) X5 top tags from set of pages’ tags in previous part from all users were selected and called TR3.

4) a) pages from user pages set are selected whose assigned tags existed in CWRM set.
   
   5) b) X6 top tags from set of pages’ tags in previous part were selected and called TR4.

In this module related information collected using user profile and other users profile. If a user enters the system for the first time then TR1 . . . TR4 sets are empty sets since extracting information from other users profile have a direct relation with user profile. This problem solved by Extracted Tag and Related Web Page(s) Tags modules.

E. Related Web Page(s) Tags

This module uses user ID and CWRM set and performing the following steps that divided into several parts:

1) a) all pages whose assigned tags existed in CWRM set were selected.

2) b) X7 pages which have most occurrences of assigned tags from previous part were selected.

3) c) X8 top tags from set of pages’ tags in previous part which have most occurrences were selected and called TR5.

F. Recommendation Engine

In proposed system the Recommendation Engine module presented the final results. The TR1 . . . TR6 sets was designed to be able to present the best final result using a combination of these sets and then suggesting most appropriate tags whose have most occurrences to user. The number of tags suggested by system is according to parameter λ.

V. SYSTEM EVALUATION

In this section evaluation metrics presented.

A. Evaluation of metrics

For efficiency and accuracy of proposed system, information retrieval metrics such as recall, precision and f measure was used [26, 31].

1) Recall: it is a common metric for evaluating the utility of recommendation algorithms and defined as:

   \[ Recall = \frac{T_h \cap T_r}{T_h} \]  

   where:
   
   - \( T_h \): Set of pages which have most occurrences of assigned tags.
   - \( T_r \): Set of pages which have most occurrences of assigned tags in previous part.

   TABLE IV. THE PARAMETERS OF USER POPULAR TAGS AND WEB PAGE(S) AND RELATED WEB PAGE(S) TAGS MODULES

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>X₁</td>
<td>25</td>
</tr>
<tr>
<td>X₂</td>
<td>5</td>
</tr>
<tr>
<td>X₃</td>
<td>15</td>
</tr>
<tr>
<td>X₄</td>
<td>50</td>
</tr>
<tr>
<td>X₅</td>
<td>70</td>
</tr>
<tr>
<td>X₆</td>
<td>35</td>
</tr>
<tr>
<td>X₇</td>
<td>4</td>
</tr>
<tr>
<td>X₈</td>
<td>5</td>
</tr>
<tr>
<td>X₉</td>
<td>20</td>
</tr>
</tbody>
</table>
Which $T_h$ is a set of tags assigned by user; $T_r$ is a set of recommended tags by system.

2) Precision: it is common metric for measuring the usefulness of recommendation algorithms and defined as:

$$\text{Precision} = \frac{|T_h \cap T_r|}{|T_r|}$$

(3)

3) F-measure: it is another common metric of evaluation in which recall and precision parameters used as following:

$$F = \frac{2PR}{P+R} = \frac{2}{\frac{P}{R} + \frac{R}{P}}$$

(4)

This metric presents the relation between precision and recall.

VI. EXPERIMENTAL EVALUATION

In this section the description of dataset presented. Furthermore system requirement for evaluating the proposed schema and the effect of parameters $\alpha$ and $\mu$ on the quality of results presented and afterwards comparison of proposed system with other relevant models and algorithms presented. The dataset which used in this paper is gathered from a popular annotation web site called Delicious.com that was collected from 1 Dec 2007 to 31 Dec 2007 by DAI-labor, Technische university of Berlin; Germany [30] includes 362,866 users, 3,311,815 pages, 555,506 tags and 20,280,277 annotations. The specifications of system that was used for evaluation are Processor Intel Core™ Duo 2.66 GHz, RAM 3.0 GB, Hard Disk 500 GB, Operating System Windows 7 Ultimate and for implementation using C# .Net and SQL Server 2008.

As mentioned before in Web Page Crawler module the most proper values are $\alpha = 4$ and $\mu = 10$. For instance suppose the value of $\alpha$ be one of $\{3,4,5\}$ and the value of $\mu$ be one of $\{8,10,12\}$. The experiments in which F-measure metric used for evaluation shown, Fig. 4, that if $\alpha = 5$ and $\mu = 8$ then the search domain became smaller then it contains less appropriate words thus the quality of result reduced and if $\alpha = 3$ and $\mu = 12$ then the search domain became bigger then it contains more inappropriate words so again the quality of result decreased and also the search time increased.

In order to evaluate the proposed system which strictly depends on pages content, several pages selected uniformly random and before experiments began we made sure of the availability of the selected pages. For experiment step 150 pages selected and parameter $\lambda$ is 20 for all experiments. The proposed system compared to STaR and Hybrid tag recommendation methods [24, 11], since these methods have high quality output in social tagging system. For comparison these methods simulated due to applied dataset. The results of comparing the proposed system with STaR method depicted in Table V.

According to Table V, since the proposed system using related words, new methods of searching neighbors and similar information it is superior to STaR method.

In comparison to applied method in [24], since it used 20-core [21], it was applied on database. Users, resources and tags are removed from the database in order to produce a residual database that guarantees each user, resource and tag occur in at least 20 annotations thus it contains 372 users, 25,912 pages, 48,644 tags and totally 2,061,300 annotations. The result of comparison depicted in Figure 5.

Since the proposed system’s modules strictly depend on user and submitted page, the results improved in comparison to method applied in [24]. The advantages of this system are being competitive with the state-of-the-art techniques in terms of accuracy, the benefits of being scalable to large real world applications, faster respond, not suffering from cold-start problem and because the proposed system has no offline phase, all modules can be employed online.

E. Comparison of Hybrid tag recommendation and the proposed system

TABLE V. THE COMPARISON OF STaR AND THE PROPOSED SYSTEM

<table>
<thead>
<tr>
<th>Method</th>
<th>Recall</th>
<th>Precision</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>STaR</td>
<td>15.93</td>
<td>25.94</td>
<td>19.73</td>
</tr>
<tr>
<td>Proposed System</td>
<td>36.58</td>
<td>59.56</td>
<td>45.32</td>
</tr>
</tbody>
</table>

![Figure 5. The comparison of Hybrid tag recommendation and the proposed system](image-url)
VII. CONCLUSIONS AND FUTURE WORKS

In this paper the main challenges of social tagging systems discussed. Experiments show that systems whose employed combination methods have better output than others. Most of the applied methods have extensive and complicated computation thus they cannot be applied for large-scaled deployments. In this paper a new method presented in order to solve several problems of social tagging systems. The proposed system is competitive with the state-of-the-art techniques in terms of accuracy, also has the benefit of being scalable to large real world applications, simple implementation, faster respond, not suffering from cold-start problem and since the proposed system has no offline phase, all modules can be employed online.

In order to improve the quality of recommended results, semantic relation among users can be utilized as well as discovering social networks among users.

ACKNOWLEDGMENT

The authors would like to express their appreciation to Professors Bamshad Mobasher, Jonathan Gemmell and Mr Cataldo Musto for their help and Mr Alan Said for providing Delicious dataset.

REFERENCES

Learning and Principles and Practice of Knowledge Discovery in Databases, pp. 235-243, 2009.


