Tag and Neighbour Based Recommender System for Medical Events

Karunakar Reddy Bayyapu and Peter Dolog

IWIS — Intelligent Web and Information Systems,
Aalborg University, Computer Science Department
Selma Lagerlöfs Vej 300 DK-9220 Aalborg, Denmark
E-mail: {kreddy, dolog}@cs.aau.dk

Abstract. This paper presents an extension of a multifactor recommendation approach based on user tagging with term neighbours. Neighbours of words in tag vectors and documents provide for hitting larger set of documents and not only those matching with direct tag vectors or content of the documents. Tag popularity, tag representativeness and tag similarity are applied similarly as in the original approach but also to neighbours. By doing so, we treat the documents which have been added to the result set by considering word neighbours in the same way as the others. This provides an advantage in the situations where the quality of tags is lower. We discuss the approach on the examples from the existing Medworm system to indicate the usefulness of the approach.

1 Introduction

Search systems typically return a ranked list of web pages on different aspects of the same topic in the returned list in a response to a users request. Recently, social activities such as tagging have emerged mostly to help people to organize resources of their personal interest on the web. The tagging information has been applied to help information retrieval and recommender systems. Although some successful applications have been developed (see, for instance,[4]), implementing and extending a hybrid tag-based recommender system with personalization for social bookmarking systems is still a challenge. Many applications would benefit from tag based analysis, which is sufficiently general to be useful in a wide range of applications, is already performed.

Tags in Medical bookmarking systems such as $Medworm$ are usually assigned to organize and share resources on the Web. Tag clouds are weighted lists of tags. The relative importance of a tag is visualized with bigger font size, bolder letters, and is measured as a count of the popularity of the tag i.e. how many times users have used it to describe a resource. Tags are generally submitted by any user in bookmarking services such as for example $http://medworm.com$. The data source tags provide useful information with sufficient background and context, even though the set of tags are quite limited to provide an accurate degree of relatedness between tags and neighbours.

Furthermore, the tags themselves as they appear in the*Medworn* system, for example, represent multiple domains. Therefore, it is not directly applicable to consider only tag measures. In our approach, we focus our efforts on neighbour objects of the tags for search and retrieval systems. Our approach combines basic similarity calculus with external factors such as a tag popularity, tag representativeness and closest neighbours semantic similarity, document score, semantic similarity of tags as described in [4]. The proposed contribution of this paper is:

- The extended hybrid tag based recommender system which bases the computation on a user query,
- Finding the closest neighbour vector of the tag from medical data source.

The rest of the paper is structured as follows. Section 2 discusses the problem and proposed solution on an example. Section 3 defines our approach to multi-factor recommendation with word neighbours. Section 4 discusses related work and positions our work in this context. Section 5 discusses experimental evaluations and outcome results. Section 6 explains analysis of experiment and conclusions .Section 7 discusses future work.

2 Working example and Motivating scenario

Let's consider the following scenario. If the user submits a query to the system, the system evaluates which documents are relevant to the query and returns a rank ordered list of documents to the users. Normally, the system considers content of the documents or collaborative user activities as factors to judge the relevance. Recently, tag-based approaches have been coined in the literature as well but only in the general situations. The domain specific searches by experts usually do not hit the most relevant documents in the first top n results. For example, the medical *Medworm* system gives almost 14470 records just based on swine flu tag. However, the system does not know what exactly user is looking for, or user doesn't know the proper words to describe what it is that he wants. Then the returned results are often unsatisfactory.

However, we could resolve this problem by proposed extension of tag and neighbour based recommender systems for medical events. This approach searches for the most trusted information. The traditional approach based on simple tag based recommendation factors are not efficient enough for domain specific systems such as *Medworm* due to lower relevance of user generated tags used. Therefore, we also apply neighbours to consider wider set of documents.

Figure 1 shows which kind of information *Medworm* can offer to improve the search to find information with respect to a certain aspect of a document. One just needs to refer to its associated tags and predicted neighbours in the corresponding documents. There are two columns depicted in the figure which represent the positions of neighbours in the text or the tag vector. By considering one side or both sides of the neighbourhood, we can target wider space of documents than with original query. This allows for expansion of the document set hit by the query. Therefore, we ensure that the user will not miss important

medical events documented in a document or a blog even when it does not match exactly the query. By applying personalization factors as in original query to extracted neighbours, we achieve similar ranking and therefore provide a means to access the most relevant documents.

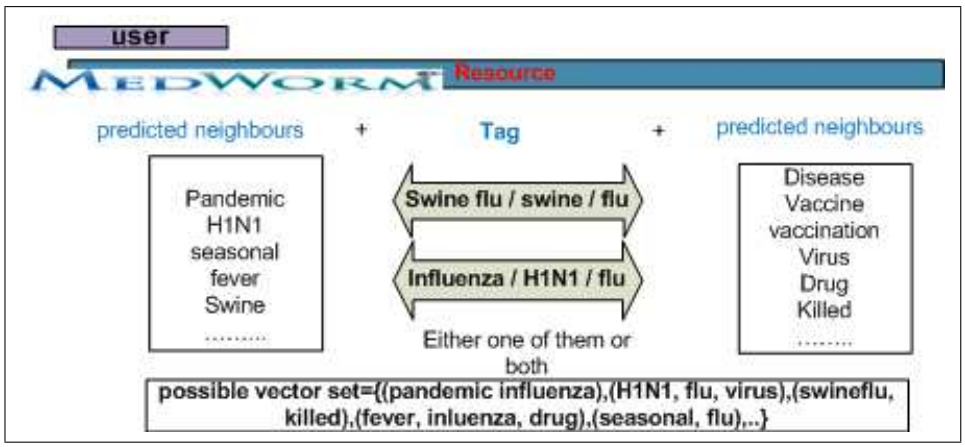

Fig. 1. Relation between predicted tags and their neighbours of user?s interest

3 Recommender System

3.1 Concept

The main concept is to achieve recommender system in medical events. The proposed recommender system combines associated neighbours with different aspects of similarity and tags. Consequently, a hybrid recommender system obtained integrate many independent recommendations by applying tag popularity, tag representativeness, tag similarity and neighbours. It exploits neighbours that are dynamically re-calculated according to the effectiveness of the recommendation. The system also uses the semantic similarity between the neighbours to calculate neighbours set. These results influence the final recommendation list to re-order the rank. The process of the calculation looks as follows. First, a users click on a tag or he submits a query. Then, the recommendation algorithm is applied to produce a set of recommended resources. Second, this set is then sorted by taking the user interest and tag neighbours into account and re-ranks the results accordingly.

3.2 Multi-Factor Recommendation with Word Neighbours

The tag and neighbour based recommender approach based on [4] is calculated as:

The extended hybrid similarity score (HS),

$$HS_{(D_i,D_{ii})} = [(Ds_{D_i} \times Ds_{D_{ii}}) \times (TS_{(D_i,D_{ii})})] \times NS_{(D_i,D_{ii})} \ ,$$

where D_i and D_{ii} are a particular documents from a set of documents D. Ds is the document score, TS is a function for measuring the tag similarity and NS is the closest neighbour vector space. We define document score as [4]:

$$Ds = \sum_{i=1}^{n} Popularity(Tag_i) \times \sum_{i=1}^{n} Representativeness(Tag_i),$$

where n is the total number of existing tags in the repository and for the definitions see in [4]. Informally, each one of the factors in the above formulas is calculated as follows:

Tag Popularity. The tag popularity is calculated as a count of occurrences of one tag per total of resources available [4]. We rely on the fact that the most popular tags are like anchors to the most confident resources. As a consequence, it decreases the chance of dissatisfaction by the receivers of the recommendations.

Tag Representativeness. [4] It measures how much a tag can represent a document it belongs to. It is believed that those tags which most appear in the document can better represent it. The tag representativeness is measured by the term frequency.

Tag Similarity (TS). It combines the classical cosine similarity (CosSim) from user query and information retrieval field with a semantic similarity (SemSim) which is defined in [4].

Cosine similarity (CosSim). The cosine similarity in our approach is a measure between, a user query Q transformed into a tag vectore and a set of tags or words (represented as a vector) of particular web page ($\bar{W}$). Each word, w(ti), in each dimension corresponds to the importance of a particular tag ti.

The Cosine similarity (Q, wi) is calculated for every tag word resource $wi \in \bar{W}$. As an output, this stage of the algorithm will produce a subset of resources W', that have some similarity to the query tag and similarity scores for each [15].

Let us assume that the user interacts with the system by selecting a query tag and expects to receive resource recommendations. Therefore, a query is a unit vector consisting of a single tag, and the equation is (adapted from [12]):

$$CosSim(Q_{D_i,wD_{ii}}) = \frac{\bar{W}_{(Q_{D_i},w_{D_{ii}})}}{\sqrt{\sum_{t \in T} \bar{W}_{(Q_{Di},w_{D_{ii}})}}}$$

where T is the set of tags, the similarity of the selected tag to each resource and recommends the top n., $\bar{W}$ is the set of words of that particular visited web page, D_i and D_{ii} are a particular documents from a set of web pages/documents.

Semantic Similarity (SemSim). The semantic relation between two tags is defined as follows:

$$SemSim(s,t) = MDSim(s,t) \times OntoSim(s,t), \forall s,t \in T$$

Where s and t are particular tags of set of tags T. MD Sim (s,t) is the Medical Dictionary similarity score and OntoSim (s,t) is the similarity score achieved from ontologies.

Neighbour Semantic Similarity (NS). Calculates closest neighbour influence for personalized recommendation by vector space models [13]. In order to find the nearest neighbours of the tag word, it must measure the similarity of the tag words [20], and select several words that have the highest similarity as the nearest neighbours of the tag word. We adopt cosine similarity algorithm to measure the similarity between word $w(ti)$ and $w(tj)$. If the user does not rate the words, we can assume the rating is zero. Assuming the rating of the n-dimensional word space of word $w(ti)$ and $w(tj)$ is respectively vector $w(\bar{ti})$ and $w(\bar{tj})$.

The similarity between word $w(ti)$ and $w(tj)$ is $sim(w(ti), w(tj))$

$$sim(w(ti), w(tj)) = \cos w(\bar{ti}), w(\bar{tj}) = \frac{w(\bar{ti}) * w(\bar{tj})}{\|w(\bar{ti})\| * \|w(\bar{tj})\|}$$

In this view, the solution to be addressed includes how to represent the tags and their closest neighbours and how to use it to influence the activation of user preferences. The approach is to predicted neighbours, the current visited context is represented as (is approximated by) a set of words concepts from the domain ontology. Ultimately, the perceived effect of neighbours extended hybrid approach is that user interests that are in focus for a current context, and those that are in the semantic scope of the ongoing user activity are considered for personalization [3].

4 Related Work

Tags have been recently studied in the context of recommender systems due to various reasons. Tags are signals or labels that particular resource was interesting for a user, and he bookmarked it as well as tagged it with a specific tag relevant for a particular situation a user was encountered in. Recommendations of relevant events should be based on the sufficient occurrences for similar signals expressed by tags. Therefore, different similarity measures need to be studied in this context for effectiveness and efficiency. [10] argues for a solution where tagging from social bookmarking provides a context for recommender systems in terms of context clues from tags as well as connectivity among users to improve the collaborative recommender system. [11] constructed a web recommender based on large amount of public bookmark data on Social Bookmarking systems. For means of personalization, [11] utilizes folksonomy tags to classify web pages and to express user's preferences. By clustering folksonomy tags, they

can adjust the abstraction level of user's preferences to the appropriate level. [11] experiment did not measure the efficiency of the recommendations in terms of user satisfaction what could give us a parameter for comparison. [17] extends a content based recommender system by deriving current and general personal interests of users from different tags according to different time intervals. However, the similarity of the tags is given by two Na?ve Bayes classifiers trained over different timeframes: one classifier predicts the user's current interest, whereas the other classifier predicts the user's general interest in a bookmark. The two classifiers are trained with a subset of the bookmarks created by a user. The tags of each bookmark, converted into a "bag of words", are used as training features. The bookmarks are recommended in the case of both two classifiers predicting a bookmark as interesting. The effectiveness of the recommendations, however, is totally dependent on the quality of the subset of bookmarks used for training the classifiers.

[19] proposes a collaborative filtering approach TBCF (Tag-based Collaborative Filtering) based on the semantic distance among tags assigned by different users to improve the effectiveness of neighbour selection. That is, two users could be considered similar not only if they rated the items similarly, but also if they have similar understanding over these items. To calculate the semantic similarity, the WordNet dictionary is being accessed to find the shortest path connecting a tag and its synonym in the graph synsets. The semantic distance based calculation, which might be difficult depending on the context of users. Special vocabularies hardly are found in general purpose dictionaries such as WordNet. Furthermore, the WordNet lacks much data useful to support proper name disambiguation, and it is not collaboratively edited [8]. [7] develops a page rank based algorithm for recommendations of resources based on preference vectors in folksonomy systems. [5] shows the benefits of using tag based profiles for personalized recommendations of music on Last.fm. The purpose of tags varies as well as tagging itself may be influenced by different factors. For example, [14] studies a model for tagging evolution based on the community influence and personal tendency. It shows how 4 different options to display tags affect user's tagging behavior. [1] studies how the tags are used for search purposes. It confirms that the tags can represent a different purpose such as topic, self reference, and so on and that the distribution of usage between the purposes varies across the domains. It compares the purposes with another literature (such as [6, 18, 14]) where these are called differently.

Other works such as [16] and [9] coined the term emergent semantics as the semantics which emerge in communities as social agreement on tag's meaning that the semantics is derived from its frequent use instead of the contract given by ontologies from ontology engineering point of view. However, the approaches based on emergent semantics are characterized by the power law which gives a long tail of the tags of which semantics have not emerged yet. Therefore, [2] looks at grounding of the tag relatedness with a help of WordNet.

5 Evaluation and Results

We have conducted an experiment to preliminary assess the performance of the recommender approach proposed in this paper. The nature of the experiment was based on a simulation a mix of scenarios regarding the amount of pages, tags and their neighbours. The proposed scenarios were created aiming at simulating realistic usage of $Medworm$. The variables addressed by each scenario are:

- *Amount of Pages:* each page has a set of tags that are compared for processing the recommendations. Therefore, the more pages exist then more time will be spent to calculate the similarity between the pages.
- *Amount of Tags:* the similarity of the pages is given by their tags. The whole set of tags of each page must be compared to verify which ones are similar.
- *Set of neighbours:* set of neighbours are depending on similarity of the tags. The whole set of neighbours of each page must be compared with semantic meaning of the tag.

These variables were chosen because we are using them for calculating the recommendations. This process is time consuming and invariably affects the system performance. However, it does not mean that other factors such as page size should not be considered[4].

We found that the choice of $tf * idf$ played an important role to find tag representativeness. In our evaluation, $tfidf$ have identical trends, but $tfidf$ always provides superior results, so we have reported only results found based on those weights. We were able to extract tag neighbours. Some samples were taken from each dataset of neighbours to find neighbours semantic similarity using cosine similarity. The validation was performed to measure the improvement in recommendation. We used MedicineNet[1] online free medical dictionary to measure semantic similarity.

Let's assume that given semantic similarity is $\varphi(s)$, where S is the semantic similarity. We also have to define cosine similarity between the user's query and tags ti, tj, tk, tl, etc. These tags could be in vector node of current webpage: $\bar{W} = \{\ldots, ti, \ldots, tj, \ldots, tk, \ldots, tl \ldots\ldots\}$, where $\varphi(s) \subseteq \bar{w}$ and $\varphi(s) \cap \bar{w} = \emptyset$

The neighbour similarity depends on the similarity of the tag words. So, neighbour semantic similarity (NS) is subset of cosine similarity. NS can be computed as follows:
$\bar{\varphi}(NS) = \{\ldots, ti-1, ti+1, \ldots, tj-1, tj+1, \ldots, tk-1, tk+1, \ldots, tl-1, tl+1, \ldots\ldots\}$, $\bar{\varphi}(NS) \subseteq \bar{w}$ and $\bar{\varphi}(NS) \sim \varphi(s)$.

In order to test the effectiveness of the algorithm, we compute the factors with a collection of documents from $Medworm$ data source about 90 pages. The documents were encoded with xml format, so we decided to make 466 tags manually. Content of the pages and some tags were extracted from web sites on the Internet. Similarly, we utilized manually generated neighbours to assign tags to $Medworm$ pages tagged by a particular user. Due to certain constraints,

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

we had to limit the number of user queries. We needed just adequate number of satisfactorily different pages and sufficiently different assignment of tags to them. Each test case consists of tag, neighbours, semantic factors and resource. We consider this resource as the target results, since we know that the user is interested in it.

Here, we have one user interest query "avian flu pandemic". A simple way to start out is by eliminating documents that do not contain all three tags "avian", "flu", "pandemic", but in general it hits many documents. Our algorithm distinguished relevant and irrelevant documents and tags like "avian", "flu", "pandemic" that occur rarely and good keywords. After performing a recommendation using both the tag and neighbours, the rank of the target resource in the recommendation set was recorded and shown in table1.

Table 1. First five documents with highest HS returned by our Hybrid Recommendation approach for a query= *avian flu pandemic*. The top document entitled *Birds in the news*, is intuitively relevant to the query

ReturnPos	Document#	Document score(DS)	Recommedation Score(HS)
1	34	19.12	14.531
2	12	14.06	8.857
3	56	12.03	4.879
4	8	10.05	3.601
5	44	10.79	2.421

As seen at Table 1 our extended hybrid based algorithm accepts a user interest, a set of neighbours, and a selected tag. The recommendation and document scores are from the interval between 0 and 20. As we can see, the top recommendation score is 14.53 (72.65% accuracy). It means that the particular document is the most relevant to users query. Result positions 2, 3, 4 are also relevant to user query but not most relevant when comparing to position 1. Position 5 document got mixed results and it is partially related to the query. This result was not considered excellent but satisfactory since our recommendations relied basically on syntax similarity of the tags.

6 Discussions and Conclusions

Categorized *Medworm* is the application of medical RSS feeds [2] to better target the delivery of health care, facilitate the discovery of new products, and helps to determine a person's predisposition to a particular disease or condition through web. In this recommender systems, the extended hybrid algorithm

[2] http://www.medworm.com/rss/aboutmedworm.php

can perform tasks such as discovering documents (much like the web robots), ranking documents, filtering them, and automatically routing useful and interesting information to users and it has learning and adaptation capabilities. In fact, the extended hybrid algorithm perfectly suits information discovery and retrieval in the web. For example, information discovery and ranking can be handled by document score which depends on the tag popularity and tag representativeness. Another tag similarity can specialize in indexing, yet another like neighbours similarity can implement an information retrieval, and so forth. Applying these techniques to the web pages/documents, retrieved by a search tool could substantially weed out unrelated documents and improve the ranking quality of the remaining page/documents [15].

As per our experiment, the user has tagged several medical object web resources with the tag "flu". If a user selects that tag, the system should recommend resources concerning the number of records about the "flu". Certainly in addition, another "flu" related documents may have been tagged with alternative tags: disease, swine, H1N1, avian, bird flu, influenza, flu attacks, respiratory problems, pandemic, symptoms, lung infection, etc. These resources may have been tagged with "flu" and may not have been "flu" but they should still be made available to a user. The recommendation strategies must also be adapted to deal with the neighbours. Typically, recommender systems have dealt with two dimensions: users query and object neighbours semantic similarity. Consider again the "flu" related topics. After selecting "flu" the system may recommend resources only related by semantic similarities of neighbours. User may notice as of his query "avian flu pandemic", the "avian", "flu", and "pandemic" are related tags. These are generated by the closest neighbour semantic objects.

Finally, user may notice resources in *Medworm's* profile dealing with the medical neighbour objects, and view one of those resources as his priority is most trusted information in the top position.

7 Future Work

The evaluation of the system provided some supplementary conclusions, namely, a recommendation performed with association neighbours appeared to be the most useful but only positive influence neighbours. Future work will focus on positive and negative neighbours and tag similarity i.e. sentiment analysis. It would benefit the system to utilize such opinions and to lower the score of bad results, even if other strategies show them as recommendable.

Acknowledgments

This work is partially supported by the European Union under the IST project M-Eco (http://www.meco-project.eu).

References

1. Kerstin Bischoff, Claudiu S. Firan, Wolfgang Nejdl, and Raluca Paiu. Can all tags be used for search? In James G. Shanahan et al., editor, *Proc. of the 17th ACM Conference on Information and Knowledge Management, CIKM 2008*, pages 193–202, Napa Valley, California, USA, October 2008. ACM.
2. Ciro Cattuto, Dominik Benz, Andreas Hotho, and Gerd Stumme. Semantic grounding of tag relatedness in social bookmarking systems. In Amit P. Sheth et al., editor, *The Semantic Web - ISWC 2008, Proc. of the 7th Intl. Semantic Web Conference, ISWC 2008*, volume 5318 of *Lecture Notes in Computer Science*, pages 615–631, Karlsruhe, Germany, October 2008. Springer.
3. C. Dolbear, P. Hobson, D. Vallet, M. Fernndez, I. Cantador, and P. Castells. Personalised multimedia summaries. In *Semantic Multimedia and Ontologies Part III*. 2008.
4. Frederico Duarao and Peter Dolog. Extending a hybrid tag-based recommender system with personalization. In *Accepted for ACM Symposium on Applied Computing 2010*, Lausane, Switzerland, 2010. ACM Press. Accepted for publication.
5. Claudiu S. Firan, Wolfgang Nejdl, and Raluca Paiu. The benefit of using tag-based profiles. In Virgílio A. F. Almeida and Ricardo A. Baeza-Yates, editors, *Fifth Latin American Web Congress (LA-Web 2007)*, pages 32–41, Santiago de Chile, November 2007. IEEE.
6. Scott A. Golder and Bernardo A. Huberman. Usage patterns of collaborative tagging systems. *Journal of Information Science*, 32(2):198–208, 2006.
7. Andreas Hotho, Robert Jäschke, Christoph Schmitz, and Gerd Stumme. Information retrieval in folksonomies: Search and ranking. In York Sure and John Domingue, editors, *The Semantic Web: Research and Applications, Proc. of the 3rd European Semantic Web Conference, ESWC 2006*, volume 4011 of *Lecture Notes in Computer Science*, pages 411–426, Budva, Montenegro, June 2006. Springer.
8. Jay J. Jiang and David W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. Technical report, 1997.
9. Peter Mika. Ontologies are us: A unified model of social networks and semantics. *Web Semant.*, 5(1):5–15, 2007.
10. Reyn Nakamoto, Shinsuke Nakajima, Jun Miyazaki, and Shunsuke Uemura. Tag-based contextual collaborative filtering. *IAENG Intl. Journal of Computer Science*, 34(2), 2007.
11. Satoshi Niwa, Takuo Doi, and Shinichi Honiden. Web page recommender system based on folksonomy mining for itng '06 submissions. In *ITNG '06: Proc. of the Third Intl. Conference on Information Technology: New Generations*, pages 388–393, Washington, DC, USA, 2006. IEEE.
12. C. Papakonstantinou, I. Panagiotou, and F. Verbeek. Tag based meta-search for browsing the web: The tictag application. In *Proceedings 13th Computer-Human Interaction Netherlands Conference*, Leiden, The Netherlans, 2009.
13. G. Salton, A.Wong, and C. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.
14. Shilad Sen, Shyong K. Lam, Al Mamunur Rashid, Dan Cosley, Dan Frankowski, Jeremy Osterhouse, F. Maxwell Harper, and John Riedl. tagging, communities, vocabulary, evolution. In Pamela J. Hinds and David Martin, editors, *Proc. of the 2006 ACM Conference on Computer Supported Cooperative Work, CSCW 2006*, pages 181–190, Banff, Canada, November 2006. ACM.

15. A. Shepitsen, J. Gemmell, B. Mobasher, and R. Burke. Personalized recommendation in social tagging systems using hierarchical clustering. In *Proceedings of the 2008 ACM conference on Recommender systems*, pages 259 – 266, 2008.
16. Steffen Staab. Emergent semantics. *IEEE Intelligent Systems*, 17(1):78–86, 2002.
17. Pavan Kumar Vatturi, Werner Geyer, Casey Dugan, Michael Muller, and Beth Brownholtz. Tag-based filtering for personalized bookmark recommendations. In *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge mining*, pages 1395–1396, New York, NY, USA, 2008. ACM.
18. Zhichen Xu, Yun Fu, Jianchang Mao, and Difu Su. Towards the semantic web: Collaborative tag suggestions. In *WWW2006: Proc. of the Collaborative Web Tagging Workshop*, Edinburgh, Scotland, 2006.
19. Shiwan Zhao, Nan Du, Andreas Nauerz, Xiatian Zhang, Quan Yuan, and Rongyao Fu. Improved recommendation based on collaborative tagging behaviors. In *IUI '08: Proc. of the 13th Intl. conference on Intelligent user interfaces*, pages 413–416, New York, NY, USA, 2008. ACM.
20. L. Zheng, Y. Wang, J. Qi, and D. Liu. Research and improvement of personalized recommendation algorithm based on collaborative filtering. *IJCSNS International Journal of Computer Science and Network Security*, 7(7), jul 2007.