ISSN 2319-2720

Volume 3, No.1, January – March 2014
International Journal of Computing, Communications and Networking
Available Online at http://warse.org/pdfs/2014/ijccn023120143.pdf

Tag Recommendations Combining Content based and Collaborative Techniques

Rohit Jha[1] Ajay Goyal[2]
[1]Dept. of I.T, SATI, Vidisha (MP), India, rohitjha@live.in
[2]Dept. of I.T, SATI, Vidisha (MP), India, ajay_champ_sati@yahoo.co.in

ABSTRACT

Tagging has become a powerful tool discover and search, it directly allow users to freely create and choose the categories that best describe a piece of information. However, It becomes difficult to distinguish similar interests between customers because the sparsity problem is caused by the insufficient number of the transactions. Various Different solutions have been proposed but they are effective only on heavily used system. In this paper, we proposed a system that utilizes strengths of various tag sources and relations between concepts captured in tag co-occurrence graphs mined from collaborative actions of users. The architecture of the proposed system is based on a text indexing engine, which allows the system to deal with large datasets in real time, while constantly adapting its models to newly added posts. We defined the importance of the utilization of a feedback loop in the tag recommendation process.

1. INTRODUCTION

Tag Recommendation is an extremely active area of research in software engineering in recent years. Several varieties of tag recommendation algorithms have been proposed by researchers all over the world. These have been divided into three categories namely (i) Graph based, (ii) Content based and (iii) Hybrid approaches. Graph based systems use collaborative filtering and utilizes the relationship between the tags, the users and the resource represented in a folksonomy graph. Content-based systems are based solely on the textual metadata related to the resource. Hybrid systems combine these two types of input.
Emphasizing on practicality of tag recommendation systems a wide range of approaches have been reported. The main objective of tag recommendation system is to predict tag that users would like to user for their resource. Therefore, while designing tag recommendation system it becomes essential that best suitable experimental studies are undertaken. The researchers in the past have clearly defined the overview of tagging models in order to explain tagging. Information about folksonomy data structure has been also provided. The existing tag recommendation solutions have been categorized into approaches that analyze the folksonomy in order to come up with recommendations, and content-based approaches where the textual content and/or meta-data of documents is considered.

Tagging is a popular means of annotating objects on the web. Tagging is becoming an increasingly important tool to help people organize their information in huge item collections. A detailed account of different types of tag recommender systems can be found in [Golder and Huberman[3 and 4]. For example, it allows people to bookmark the items they are interested in and to organize them into various topic sets by adding tags. Once sufficiently many items are tagged, the tags can also be used to search for items on a specific topic. Since tags are associated to both items and users, tags also can be used for generating personalized recommendations. However, unlike keywords or subject headings assigned by information professionals, tags usually lack any form of explicit organization and normalization.

Thus, search and recommendation need to be adapted to the characteristics of tagging systems.Automatic content recommendation has already become a mature field of academic study. A number of standard algorithms have evolved.Most of which are based on implicit or explicit feedback from users on items, usually in the form of item ratings.The aim of tagging is to group and organize objects and make it easier to find a particular object in the collection. In contrast to a hierarchical organization, tags are usually not organized in a fixed taxonomy. This unstructured form makes it easier for users to select tags for objects without having to worry about the location of the tags they use in a hierarchy.

1.1 The Tag Recommendation System

With the advent of affordable domestic high-speed communication facilities, in-expensive digitization devices, and the open access nature of the Web, a new and exciting family of Web applications known as Web 2.0 has been born. The underlying idea is to decentralize and cheapen content creation, thus leading the Web into a more open, connected, and democratic environment. In this chapter we focus on a particular family of Web 2.0 applications known as Social Tagging Systems.

1.2Content based recommendation

In content-based approaches, the textual content of the documents is used for tag extraction and expansion [15, 17], word-tag co-occurrence [17], or with document classification techniques [19]. Important aspects of content based approaches are the content source and the document representation used. Experimentations have shown that the most informative words generally appear

in the title and URL [20], and the document text (21). For structured text documents such as HTML, further sources such as anchors, links and paragraphs are available.

1.3 Hybrid Systems

Hybrid tag recommendation systems try to combine the advantages of resource content and folksonomy graphs. As they usually start the processing with the resource content, they are often classified as content-based methods. Graph and content based systems usually tailor a well-known machine learning or information retrieval approach to the tag recommendation problem. In comparison, hybrid systems try to utilize specific strengths of several information sources in folksonomies. Such approach allows them to be more efficient and process a wider variety of posts, hence it makes them more practical.

2.LITERATURE REVIEW

Heymann et al. [22] carried out experiments on HTML pages comparing the value of page text, anchor text and text of surrounding hosts for tag prediction. They concluded that out of the three, the document text was most informative and anchor text was more informative than surrounding hosts. The document representation in content-based approaches is usually a bag-of-words. There are many different methods of determining the importance score of each word to the document, most of which include a Tf-Idf score in the calculations. The content based tags are linearly combined with tags from resource and user profiles. The system retrieves resources, which textual content is related to the posted resource title, and builds the recommendation based on prominent tags from their profiles. Specific attention is given to resources posted previously by the author of the current post —their tags are weighted higher when tags from all relevant resources are combined.

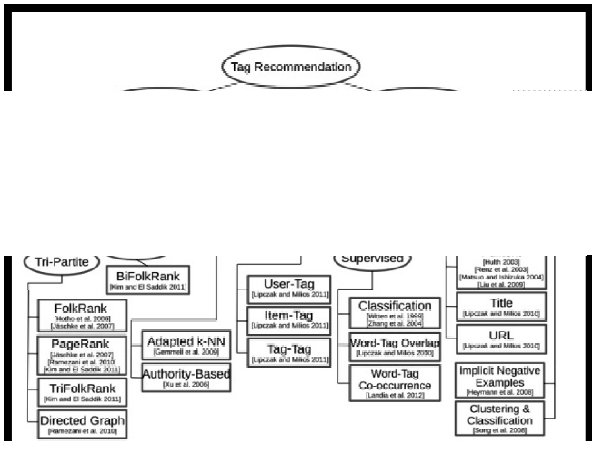

Figure 1: Overview of Tag Recommendation Approaches

Lipczaket al.[16]presented their hybrid tag recommender which won the content-based tag recommendation task of the ECML PKDD Discovery Challenge. The part of the hybrid tag recommender was reported to be closely comparable to the content-based approaches, for which Lipczak et al.[16] gave individual results which is a combination of two tag recommendation sets: past tags of the query user and tags related to the content of the query document. The only source of content data in their approach is the document title. In order to generate the content-based tag recommendations, the words in the title of the query document which have been used as tags before in the training data are first extracted (word-tag overlap). The tag recommendation set was finally expanded depending upon the tag-tag co-occurrence. Due to the initial filtering, the content words only included words which also appeared as tags in the training data. The differences between their approaches for including content data and the content-related part of the hybrid were the content sources and the document representation used. Therefore finally they considered and evaluated two different content sources namely; document title and full text content, in their approaches.

Tatu and D'Silva[29] proposed a system based on tags extracted from resource and user profiles. The set of tags is extended using NLP techniques and later merged with content based tags. A system by Ju and Hwang [30] scans the content of previously tagged documents to evaluate the likelihood of a content word being used as a tag. The likelihood is later used as a score for words that occur in the content of currently posted resource. The content based tags are linearly combined with tags from resource and user profiles.

Musto et al.[29] based their system on a search engine. The system retrieves resources, which textual content is related to the posted resource title, and builds the recommendation based on prominent tags from their profiles. Specific attention is given to resources posted previously by the author of the current post — their tags are weighted higher when tags from all relevant resources are combined.

2.1 Tag Cloud

A tag cloud (word cloud or weighted list in visual design) is a visual representation for text data, typically used to depict keyword on websites, or to visualize free form text. Tags are usually single words, and the importance of each tag is shown with font size or colour. This format is useful for quickly perceiving the most prominent terms and for locating a term alphabetically to determine its relative prominence. When used as website navigation aids, the terms are hyperlinked to items associated with the tag.

A text cloud or word cloud is a visualization of word frequency in a given text as a weighted list. The technique has recently been popularly used to visualize the topical content of political speeches.

Tag clouds have been subject of investigation in several usability studies. The following summary is based on an overview of research results given by Lohmann et al:
Tag size: Large tags attract more user attention than small tags (effect influenced by further properties, e.g., number of characters, position and neighbouring tags).
Scanning: Users scan rather than read tag clouds.
Cantering: Tags in the middle of the cloud attract more user attention than tags near the borders (effect influenced by layout).
Position: The upper left quadrant receives more user attention than the others (Western reading habits).
Exploration: Tag clouds provide suboptimal support when searching for specific tags (if these do not have a very large font size).

3. PROPOSED METHODOLOGY

The challenges in generating successful tag recommendations include the personalization aspect, the dimensionality and sparsity of tagging data, and the new document problem. These are all general problems of traditional recommender systems, however, they are highly pronounced in the social tagging domain and the tag recommendation task. Due to the mentioned differences in the tagging behavior of individual users (motivation and expertise), tag recommendations have to be personalized to the preferences of the query user in order to be successfully accepted.

The information contained in social tagging data has a high dimensionality in types of objects. In contrast to traditional recommender systems that deal with ratings given to documents by users (such as a user giving a 5-star rating to a book on Amazon), tag recommendation models have to be learnt on data that contains the added dimension of tags. While numerical ratings of the same document by two different users can be compared directly, different tags assigned to the same document cannot be directly compared. The tags are an additional dimension that has to be considered by the models. The additional tag dimension combined with the differences in the tagging behaviour of users leads to a high data sparsity. Furthermore, the query always includes two objects, the user and document, and the task is to recommend tags that are appropriate for these two objects in combination. Similar challenges exist in item recommender systems which consider contextual information [Adomavicius and Tuzhilin, 2011], and where additional data dimensions such as author names or related places are analyzed. Another challenge is the new document problem, akin to the new item and cold-start issues in recommender systems [Jannach et al., 2010]. A large proportion of documents in social tagging systems is only tagged by one user [Wetzker et al., 2008], and thus many query documents for which the tag recommender is asked to make predictions have no previous tagging information associated with them. This is an important issue to address since recommending tags for new query documents based solely on the overall tagging profile of the query user is likely to result in a low success rate.

The hybrid tag recommender is conceptualized to be composed of five basic recommenders in the proposed system. This modular structure allows the system to utilize various tag sources and properties of folksonomy data structure created by taggers collaboratively.

The three basic tag sources are as follows:

- Content of the tagged resource
- Resource profile that are the tags used for the same resource by other users.
- The user profile, tags previously used. To extend and refine the set of tags extracted from resource content the system uses two graph-based recommenders which run a spreading activation algorithm using content-to-tag or tag-to-tag co-occurrence graphs.

The main idea behind the design of the recommendation process has been represented in Figure. 4.1. The figure shows different stages of processing where the tags from five basic recommenders have been merged. It is to utilize the specific advantages of each source of tags and combine the results produced by each of them to produce the final recommendation. Since the proposed system is a hybrid tag recommender, there is a large space for the possible combinations of basic system components.

Title-to-tag Recommender
Title-to-tag recommender runs the spreading activation algorithm on a directed co-occurrence graph of terms, which were used as title words or tags.

Tag-to-tag Recommender
An analogous approach can be applied to a tag-to-tag graph. The graph captures the relations between tags that frequently co-occur in the same posts.

Resource Profile Recommender
The set of tags related to the resource content is extended by the tags extracted from the resource profile. The combined efforts of users make the resource profile a very precise source of tags, pushing the best tags to the top of frequency-ranked list.

User Profile Recommender
Tags frequently used in the past are not necessary a good current recommendation. The user profile is a very rich source of potential tag recommendations. It is likely to contain tags representing different user's interests and activities, which change dynamically. The user profile recommender uses an additional scoring scheme, complementing the frequency-based scheme, as in the

resource profile recommender to adjust to this fact. The identical sets of tags with different scores are produced with the two schemes. The sets are later merged, so the final score is a linear combination of the scores proposed by both schemes.

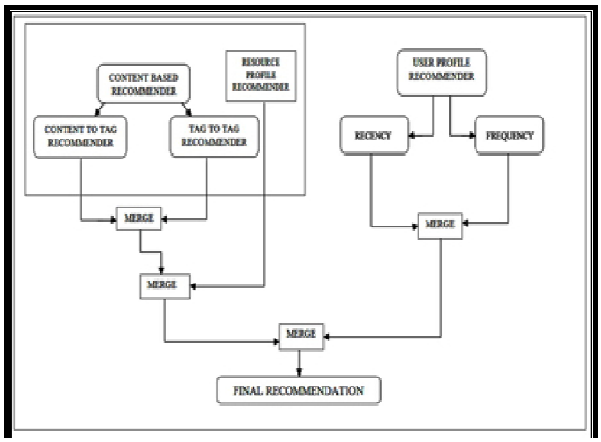

Figure: Scheme for tag recommendation system

3.1 Recommendation

For performing a recommendation task the system needs to extract two tag profiles (for the resource and the user) and a series of references to the co-occurrence graphs. The number of these references is limited by the size of the content based recommendation set. To simplify the problem, the co-occurrence graph lookup can be reduced to the tag profile lookup task. A tag profile for a term represents all tags that co-occurred with it in any of the posts, while the frequency of co-occurrences can be used to calculate the weight of the connection. To extract a tag profile for a post element (i.e., user, resource, tag or content word) the system uses a text indexing engine, which stores all previously processed posts. By accessing the Lucene index directly, the system is able to quickly retrieve a list of posts that contain a given element. As the extraction of posts is a much more time consuming task, we decided to limit the number of posts, based on which the profile is built, to the 1000 most recent posts that contain the element.

Each element type possess a separate tag profile cache. To reduce the number of references to the index, the system contains a layer of caches (Fig. 4.2). If the system hits the profile in the cache, it does not have to refer to the index. In case of a miss, the profile is built based on the information extracted from the index. If the element was used in more than 20 posts, its profile is added to the cache replacing the profile with the lowest value of replacement function. We experimented with two basic replacement policies: In the system it was decided to use a combination of recency and frequency factors described in the following equation (4.2) which in most cases is able to match or outperform the better of the two basic policies.

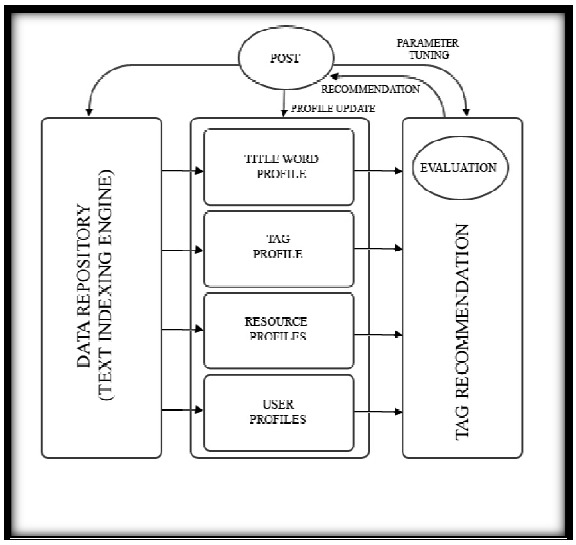

Figure: System architecture. The cache layer improves system efficiency.

$$rf(item) = \frac{\text{frequency}}{\{\text{Current Time} - \text{Last Time Used}\}} \qquad \text{(Eq. 4.2)}$$

4. RESULT & ANALYSIS

We evaluated the proposed tag recommendation system using datasets from six collaborative tagging systems. The datasets represent a wide variety of tagging systems in terms of type of folksonomy, its size, time-span of posts and character of posted resources. The system was evaluated from the perspectives of its electiveness and efficiency. The electiveness evaluation included the experiments, which tested the system's ability to tune its parameters to the characteristics of a specific collaborative tagging system and the quality of recommendations produced by the system and its processing stages.

Despite the large number of publications on tag recommendation problem, little has been done on the unification of the evaluation methods. In fact, most of the systems are evaluated in a unique way proposed by their authors. In some cases the evaluation methodology follows the specific application of the system and it is unlikely that we can find a "one-for-all" evaluation approach for all tag recommendation systems.

To observe the impact of online content adaptation on the results and provide a base- line for the system we ran a series of experiments in which this feature was turned off. The parameters of the system were re-trained to tune it to the new conditions. The adaptation improves the results of the recommendations for all tested datasets (Table 6.1). The statistical significance of the difference was confirmed by a Wilcoxon signed-rank test ($P< 0.001$). For the three broad folksonomies, online content adaptation has a clear impact on the relative importance of different

tag sources. We present the plots of recall and precision for each stage of the recommendation process, without and with adaptation, to show how they contributed to the final result (Figure6.2). For all datasets the largest improvement is noticed for the user related tags.

Adaptation allows the system to extend the repository of user related tags by the tags that describe user's recent interests. The system is also able to gather in- formation about new coming users, from the moment they start to use the system. It is especially important for the BibSonomy and CiteULike datasets, for which we observed a large number of users who started to use the system in the test period. For these two datasets user related tags become the richest and most accurate source of tags. This is not the case for the Delicious dataset where the improvement of user related tags is comparable to resource related tags. It seems that the availability of a large number of newly added posts allows resource profiles to overcome the problem of cold start — the noisiness of profiles of infrequently posted resources.

Finally, the adaptation seems to have little or no impact on the content related tags extracted from the co-occurrence graphs. The associations between tags are well established at the time of the evaluation and they are not changed by the adapted content. In this case the adaptation is likely to be useful in the early stage of folksonomy formulation only.

Datasets	With Adaptation	Without Adaptation	Per cent Increase
BibSonomy	0.380	0.237	60.34
CiteULike	0.433	0.273	58.61
Delicious	0.449	0.344	30.52

Table: Adaption results for Broad Folksnomies datasets

Datasets	With Adaptation	Without Adaptation	Per cent Increase
Stack Overflow	0.550	0.499	10.22
BlogSpot	0.384	0.356	7.86
WordPress	0.465	0.430	8.14

Table: Adaption results for Narrow Folksnomies datasets

REFERENCES

1. Nikolas Landia, Stephan Doerfel, Robert Jäschke, Sarabjot Singh Anand, Andreas Hotho, and Nathan *Griffiths.2013.* **Deeper Into the Folksonomy Graph: FolkRank Adaptations and Extensions for Improved Tag Recommendations***in CoRR (2013).*

2. Symeonidis, P., Nanopoulos, A. and Manolopoulos, Y. 2008. **Tag recommendations based on tensor dimensionality reduction**. In *Proceedings of the 2008 ACM Conference on RecommenderSystems (RecSys '08).* ACM, New York, NY, USA, 43–50. DOI:http://dx.doi.org/10.1145/1454008.1454017.

3. Rendle, S., Leandro BalbyMarinho, AlexandrosNanopoulos, and Lars Schmidt-Thieme. 2009**. Learning optimal ranking with tensor factorization for tag recommendation.** In *Proceedings of the 15th ACMSIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '09)*. New York, NY,USA, 727–736. DOI:http://dx.doi.org/10.1145/1557019.1557100.

4. Jaschke, R., Balby, L., Hotho, M., Andreas Schmidt-Thieme, Lars and Stumme, Gerd 2007. **Tag Recommendations in Folksonomies.** In *Knowledge Discovery in Databases: PKDD 2007, 11th European Conference on Principles and Practice of Knowledge Discovery in Databases*. 506–514.

5. Ramezani, M., Gemmell, J., Schimoler, Th and BamshadMobasher. 2010. **Improving Link Analysis for Tag Recommendation in Folksonomies**. In *Proceedings of the 2nd Recommender Systemsand the Social Web Workshop at RecSys '10*. 39–45.

6. Gemmell, J., Schimoler, Th., Ramezani, M., Christiansen, L. and Mobasher. B. 2008. **Improving folkrank with item-based collaborative filtering**. In *ACM RecSys'09 Workshop on Recommender Systems and the Social Web*, pages 17–24.

7. Xu, Zhichen; Yun Fu, Jianchang Mao, and Difu Su. 2006. **Towards the Semantic Web: Collaborative Tag Suggestions.** In *Proceedings of the Collaborative Web Tagging Workshop at WWW 2006*. Edinburgh, Scotland.

8. Lipczak, M. and Milios. E. 2011. **Efficient Tag Recommendation for Real-Life Data.***ACM Trans.Intell. Syst. Technol.* 3, 1, Article 2 (Oct. 2011), 21 pages.DOI:http://dx.doi.org/10.1145/2036264.2036266.

9. Gemmell, J., Schimoler, Th.,Ramezani, M.and MobasherB.2009. **Adapting K-Nearest Neighbour for Tag Recommendation in Folksonomies**. In *Proceedings of the 7th Workshop onIntelligent Techniques for Web Personalization and Recommender Systems*.

10. Lipczak, Marek, Yeming Hu, Yael Kollet, and EvangelosMilios. 2009. **Tag Sources for Recommendation in Collaborative Tagging Systems**. In *Proceedings of the ECML/PKDD 2009 Discovery Challenge Work-shop*. 157–172.

11. Landia, Nikolas; Sarabjot Singh Anand, Andreas Hotho, Robert Jaschke,¨ Stephan Doerfel, and FolkeMit-zlaff. 2012. **Extending FolkRank with content data**. In Proceedings of the *4th ACM RecSys work-shop on Recommender Systems and the Social Web (RSWeb '12).*

ACM, New York, NY, USA, 1–8. DOI:http://dx.doi.org/10.1145/2365934.2365936

12. Lipczak, M. 2008. **Tag recommendation for folksonomies oriented towards individual users**. In *Proc. the ECML/PKDD 2008 Discovery Challenge Workshop*, pp. 84–95.

13. Lipczak, M. and Evangelos M. 2010. **The impact of resource title on tags in collaborative tagging systems**. In *HT '10: Proceedings of the 21st ACM Conference on Hypertext and Hypermedia*. ACM, New York, NY, USA, 179–188. DOI:http://dx.doi.org/10.1145/1810617.1810648.

14. Tatu, M. and D'Silva, 2008. M. Srikanth M. Tatu and T. D'Silva. RSDC'08: **Tag recommendations using bookmark content.** In Hotho*et al.* 2008 pp. 96–107.

15. Ju, S. and Hwang, Kyu-Baek. 2009. **A weighting scheme for tag recommendation in social bookmarking systems**. In *Proc. the ECML/PKDD 2009 Discovery Challenge Workshop*, pages 109–118.

16. Rendle, Steffen and Schmidt-Thieme, Lars. **Pairwise interaction tensor factorization for personalized tag recommendation.** In *Proceedings of the third ACM* 114 I*nternational conference on Web search and data mining*, WSDM '10, pages 81–90. ACM, 2010.

17. Halpin, Harry Valentin Robu, and Hana Shepherd. **The complex dynamics of collaborative tagging.** In *WWW '07: Proc. the 16th International Conference on World Wide Web*, pages 211–220. ACM, 2007.

18. Ricci, F., Rokach, L., Shapira, B. and Paul B. Kantor, editors. **Recommender Systems Handbook**. Springer, 2011.

19. Cheng, W. and H¨ullermeier, Eyke 2009**. Combining instance-based learning and logistic regression for multilevel classification**. *Machine Learning*, 76(2):211–225, 2009.

20. Hotho, A., Jaschke, R., Schmitz, C. and Stumme. G. 2006. **Information Retrieval in Folksonomies: Search and Ranking.** In *The Semantic Web: Research and Applications (Lecture Notes inComputer Science)*, Vol. 4011. Springer, 411–426.DOI:http://dx.doi.org/10.1007/11762256.

21. Wetzker, R., Zimmermann, C., Bauckhage, C., and Albayrak S. 2010. **I tag, you tag: Translating tags for advanced user models**. In *WSDM '10: Proc. the Third ACM International Conference on Web Search and Data Mining*, pages 71–80, New York, NY, USA, 2010. ACM.

22. Eisterlehner, F., Hotho, Andreas and Robert Jaschke, editors. **ECML PKDD Discovery Challenge** *2009 (DC09)*, volume 497 of *CEUR-WS.org*, 2009.

23. Sood, S.C. Hammond, K.J.. Owsley, S.H and Birnbaum. **TagAssist: Automatic tag suggestion for blog posts**. In *Proc. the International* L. 2007. *Conference on Weblogs and Social Media (ICWSM 2007)*, 2007.