Using Tags for Measuring the Semantic Similarity of Users in Enhancing Collaborative Filtering Recommender Systems

Editor’s Note: In this title, “to Enhance” would be better than “in Enhancing.”

Abstract Recent years have seen a significant growth in social tagging systems. Social tagging systems which allow users to use their own generated tags to organize, categorize, describe and search digital content on social media. The growing popularity of tagging systems is leading to an increasing need for automatic generation of recommended items for users. Much previous research focuses on incorporating recommender systems techniques in social tagging systems to support the suggestion of suitable tags for annotating related items. Collaborative filtering is one of these recommender system such techniques technique. The most critical task in collaborative filtering is finding related users with similar preferences, “liked-minded” users. Despite the popularity of collaborative filtering, it still suffers from some certain limitations for example in relation to “cold-start” users, for example, where it is often the case that often there are insufficient preferences to make recommendations. In addition, there is the data-sparsity problem, where there is limited user feedback data to identify similarities in users’ interests because there is no intersection between users’ transactional data, a situation which also leads to degraded recommendation quality. Therefore, in this paper report we present a new collaborative filtering approach based on users’ semantic annotations (tags), which calculates the similarity between users by discovering the semantic spaces in their posted tags. We believe that this approach better reflects the semantic similarity between users according to their tagging perspectives and consequently improves recommendations through the identification of semantically related items for each user. Our experiment on a real-life dataset shows demonstrates that our approach outperforms the traditional user-based collaborative filtering approach in terms of recommendation improving the quality of recommendations.

Keywords: Collaborative Filtering, Folksonomies, Semantic Neighborhood, Social Tagging System

Editor’s Note: All Keywords should appear in the Abstract. However, "Folksonomies" and "Semantic Neighborhood" do not occur here.
1 Introduction

The World Wide Web (WWW), has undergone exponential growth over the past two decades; the first generation enabled Internet users to have direct access to a large diversity of available knowledge items. The second generation of the WWW, usually denoted as “Web2.0” and also referred to as “participatory Web2.0,” has led to a significant change in the way in which people interact with and through the Web. Web2.0 can be characterized as a paradigm that facilitates communication, interoperability and information sharing and collaboration on the Web [1]. Web2.0 allows users to easily annotate any item (sites, pages, media, etc.) that someone else has authored. These annotations (tags) take many forms such as editing, rating, organizing and classification. These annotations enable users to easily retrieve, search or filter these items in the future. Moreover, the social phenomenon of collaborative tagging (also known as “folksonomies” or “social tags”) is a big shift from earlier local and solitary to global and collaborative Internet activity. This shift has enabled users to be information producers, rather than just information browsers. However, rich information is increasing exponentially in social web systems, and a huge substantial amount of new information is being produced every day. Since this phenomenon is already exceeding human processing capabilities, and it is becoming difficult for users to find the needed information quickly because they face the problem of information overload [2, 3]. Consequently, recommender systems have emerged in response to the information-overload challenge of information overload, providing users with recommendations of items that are relevant and likely fit their needs [4]. Collaborative tagging systems such as Delicious¹, YouTube², Flickr³ and Twitter⁴ are allowing the creator or visitors of such content to assign freely chosen keywords or tags to such content.

Many researchers have recently focused on using recommender systems with social collaborative tagging [2, 6-8] to mitigate limitations such as “cold start” and sparsity [3], which are present in the traditional recommender systems limitations such as “cold start” and sparsity [3]. However, without considering the semantics of user tags in the recommendation process, the recommender system cannot distinguish and interoperate the user’s interests for the same tags. Furthermore, almost all of the studies incorporate recommender systems in social tagging to deal with tag suggestions and recommendations and help users annotate a related item.

We believe that semantic tags can tackle the limitations inherent in traditional collaborative filtering and improve the quality of collaborative filtering by capturing users’ semantic preferences based on the user tags. In traditional user-based collaborative filtering, two users are similar if they co-rate a particular item with similar score values. Consider two users, u₁ and u₂, both of whom rate the movie “Avatar” with a similar high score. In traditional user-based approaches, u₁ and u₂ are considered similar and “like-minded” users. However, traditional methods discard the semantic perspective of the respective users, where u₁ gives awards a high score because he likes “science fiction” movies, whereas the other user u₂ likes “Avatar” because she likes “adventure movies.”

To elaborate the problem further, in state-of-the-art social tagging and collaborative filtering, two or more users are considered similar if they both annotate a particular item with similar tags. For example, let user u₁ post tags (java, tour) on an item, and u₂ post tags (java, XML) on another item. In traditional methods, u₁ and u₂ are similar because both of them have tagged (java); however, there is no similarity between u₂ and u₃. Unfortunately, the similarity is incorrect in this scenario because traditional methods do not distinguish between “Java,” the island, for u₁ and “java,” the programming language, for u₂. To solve this problem, our approach determines the semantic similarity

¹) http://www.delicious.com
²) http://www.youtube.com
³) http://www.flickr.com
⁴) http://www.twitter.com
between users such so that \( u_2 \) and \( u_3 \) are identified as being more semantically similar according to their semantic tags, wherein the tags “java” and “XML” of \( u_2 \) are more similar to the tag “RDF” of \( u_3 \) than to the tags “java” and “tour” of \( u_1 \).

This paper report presents an approach to measure the similarity between two users based on the basis of the occurrences co-occurring features of tags as exhibited by in other research, we extend those tags into their respective semantics by exploiting available open-semantic [notice inserted apostrophe] lexical resources. The remainder of this paper report is organized as follows: Section 2 presents an overview of the traditional collaborative filtering approach and related work; Section 3 describes the preliminaries related to the proposed approach; Section 4 describes introduces the proposed approach; Section 5 presents describes the experiments and results; and finally, Section 6 offers a conclusion and our directions recommendations for future work are detailed in section 6.

* Editor’s Note: Use “semantics” as a noun but “semantic” as an adjective. See http://www.merriam-webster.com/dictionary/semantic and http://www.merriam-webster.com/dictionary/semantics, noticing “noun plural but singular or plural in construction.”

2 Background and Related work

In a common formulation, the recommender system task is reduced to the problem of discovering related items that have not been seen by the user [3]. Collaborative filtering (hereafter CF hereafter) is considered to be the best recommendation technique that automates the process of the “word-of-mouth” paradigm [9]. Collaborative filtering CF compares a user’s basis on the basis of the similarity of their preferences and those of other users [10, 11]. The main two main approaches in CF are the item-based approach [12, 13] and the user-based approach [14]. Usually, the recommendation process in both of these CF approaches depends on finding a similar pattern for the target user (the term used in the user-based approach) and other users holding similar preferences to form a “neighborhood,” and these the preferences from these like-minded neighbors which are called the most similar users (or similar items is the term used in the item-based approach). Many computing methods have been used to measure the similarity between users in CF, such as the Pearson correlation coefficient and cosine similarity [4]. The most critical task in a CF recommender system is the formation of a similar neighborhood, because differences in these like-minded neighbors leads within result to in different recommendations, which influences thereby influencing the accuracy of the recommendation process. The similarity between two users, \( u \) and \( v \), is calculated as the cosine angle between users the corresponding feature vectors, [insert comma] as follows:

\[
similarity(u, v) = \cos(\theta) = \frac{u \cdot v}{||u|| \cdot ||v||}
\]

When these neighbors have been found, the next step is the process of estimating the predicted value of items which have not been seen or rated as yet unseen and unrated by the target user. The greater the number of similar users found in the recommendation environment for the active user, the more influences this user has on the prediction-estimation process of estimating the predicted value for unseen items. The last step is the recommendation of the top \( M \) items with the highest predicted values to the target user [15].

As a result of the growing popularity of social media sites in recent years, many researchers have investigated the recommender system domain under the social-tagging area of research, where tags have
been considered as an additional information resource for designing effective recommendation systems. Social tagging systems or folksonomies allow users to assign content with a freely chosen keyword or tag [5], which can reflect the users’ cognitive preferences on for the content. Hence, the tag co-occurrence properties might express similarity between users or items to build a user community and item clusters, which can be employed to estimate the likely items for targeted individuals. Therefore, tags in social tagging provide a promising way to tackle some of the limitations in recommender systems, such as the cold-start [notice inserted hyphen] and sparsity problems [16]. The phenomenon of social tagging has resulted in two areas of research in recommender systems: i) tag recommendations and suggestions and ii) resource filtering and recommendations.

In When making tag recommendations and suggestions, the main idea is to provide assistance to assist users by recommending appropriate tags for users to annotate given items. The proposed approach presented in this paper report, however, falls into the latter category, i.e., resource filtering and recommendation. Therefore, it is beyond the scope of this paper report to discuss various approaches in this area to tagging. Further information is available in [17-22].

The area of resource filtering and recommendations has attracted many researchers to who have proposed novel methods for improving current recommender systems. Tso-Sutter et al. [17] integrate tags into CF by reducing the *three-dimensional (user, item, tag) relationship into to three two-dimensional relationships, as (user, tag), (item, tag) and (user, item). The idea behind their approach is to consider tags as items in a two-dimensional relationship; for example, in the (user, tag) relation, these tags <user, tag> are should be considered as a single item in the user–item rating matrix. To integrate tags with CF, the authors researchers then apply a fusion method to re-associate these relationships. However, generally, the process of reducing the three-dimensional dimensions relation into flat two-dimensional relation generally leads to the discarding of potentially useful information for the recommender system.

De Gemmis et al. [18] propose a strategy that enables a content-based recommender to infer user interests. Machine-learning techniques are applied both on both official content descriptions of items (static content) and on the tagging data (dynamic content) to build user profiles and learn user interests. Static and dynamic content are preventively analyzed in order to capture the semantic preferences of the users behind the keywords to increase the prediction accuracy of the recommender system. De Gemmis et al. use the content-based approach in their proposed strategy because they believe that the content-based this approach can provide accurate recommendations [16]. However, this approach is only effective if the recommended items contain rich content information, such as books, articles and bookmarks, that can be easily extracted.

Bao et al. [19] propose two algorithms to incorporate social tagging and web searching into CF, called SocialSimRank (SSR), which is used to calculate the similarity between tags and web queries, and SocialPageRank (SPR), which is used to capture tabulate the popularity of a web page by considering the count number of a web page’s annotations on the page. Bao et al. attempt to improve web searching by incorporating social tags into user query expansion. Liang et al. [23] incorporate tags in CF to generate a tag-based similarity to improve the traditional CF by clustering users based on their tagging behavior instead of their similar-rating behavior. However, the number of clusters has to must be defined because, in the case of a huge large tagging space, clustering can be a very expensive computation. Sen et al. [21] propose a tag-based recommendation algorithm called “tagommenders,” the underlying idea of their algorithm is being that it this algorithm can be used to predict user preferences for items based on the basis of their inferred tag preferences data. Au et al. [22] suppose assuming that it is still possible that users can influence one another in the process of item adoption through various implicit mechanisms,
Our approach differs from the aforementioned studies in that our aim is to explore the tagging of the semantic space of users. In other words, we consider semantic tags to discover like-minded users to recommend so that semantically relevant items can be recommended to a particular user. We expect this to be more useful not only in improving the recommendation quality of recommendations, but also in realizing better user perception of relevant items. *Editor’s Note: Sometimes the difference between “relation(s)” and “relationship(s)” can be very subtle. The sound of these revisions in green intuitively seems more appropriate to me as a highly educated native speaker of English. Compare http://www.merriam-webster.com/dictionary/relation and http://www.merriam-webster.com/dictionary/relationship.*

3 Semantic similarity for tags

In our approach the semantic similarity between tags is obtained by exploiting the well-known WordNet lexical database for English. WordNet is a large conceptual model database of nouns, verbs, adjectives and adverbs that are grouped into sets of cognitive synonyms (synsets) [24]. In WordNet there is a conceptual-semantic interlinkage and lexical relationship between synsets. The terms which hold the same meaning are referred to as synonyms, and those which belong to the same concept and are placed in the same synset. These hierarchical concepts are hierarchical and can quantify how much the extent to which concept A is similar to concept B. For example, these hierarchical concepts and relationships might show indicate that an automobile is more similar to a boat than to a tree; [delete semi-colon] this is because “boat” and “automobile” share “vehicle” [notice 3 pairs of inserted quotation marks] as an a common ancestor in the WordNet structure [25].

Our proposed approach enables any user to tag any item in the collaborative tagging environment, as well as duplicate a tag for the same item as a different user. We based our approach on the triple \(<user, item, tag>\) representation which is widely adopted in the collaborative tagging community [26]. A folksonomy is a set of triples. Each triple represents a user's annotation of an item with a tag. More technically speaking, if there is a list of users \(U = \{u_1, u_2, u_3, \ldots, u_m\}\), a list of items \(I = \{m_1, m_2, m_3, \ldots, m_k\}\), and a list of tags \(T = \{t_1, t_2, t_3, \ldots, t_n\}\), the folksonomy \(F = (U, I, T, Y)\), where \(Y\) is the user tag assigned for an item [27].

This designation differs from traditional CF where each user assigns a tag to an item, rather with triples; [delete semi-colon & insert period]. There is the existence of a scope of real numbers. [delete period This leads us to consider the user feature as a vector of tags posted by the user. For example, in Fig.1 (see section 4.1), the user \(u_2\) posts \(t_1, t_3\) for \(item_1\) which can be represented as \((u_2, m_1, (t_1, t_3))\).

To provide a semantic grounding for our folksonomies, we use WordNet as the external semantic space for measuring the semantic similarity between tags. Calculating the semantic similarity in WordNet can be done by measuring the distance between nodes related to the associated concepts. When the links between these nodes are considered in terms of distance, then the that distance between nodes indicates how similar the concepts are. We measure the similarity between tags by using Lin's semantic similarity [28], which uses information content in calculating the semantic similarity for calculation. Lin’s semantic measures relate the information content (IC) of the most informative common ancestor (MICA) to the IC of the associated concepts thus:

\[
Sim_{Lin}(c_1, c_2) = \frac{2 \times IC(C_{MICA})}{IC(c_1) + IC(c_2)}
\]
Lin's similarity ranges from 0 for tags without similarity to 1 for tags with maximum similarity. Budanitsky et al. [29] point out that similarity can be considered as a special case of relatedness because both are semantic notations. When measuring the semantic similarity for of social tags, we need to map the tags to an existing lexicon or thesaurus such as the WordNet [24]. However, tags are by nature free keywords which can include many community-specific terms that do not exist in any lexicon. Therefore, we propose the use of co-occurrence distribution to identify the semantic similarity for such tags.

Christian et. al [30] propose an approach to define the co-occurrence relationship between tags in social tagging systems as follows: Let \( n(m,t) \) be the number of occurrences of tag \( t \) on item \( m \). Let \( n(t) = \sum_m n(m,t) \) be the number of tag occurrences for all items in \( I \), and let \( N(t) = \sum_m n(m,t) \) be the number of tag occurrences for item \( m \), where \( m \) is an instance of \( I \). In [30], the similarity between two tags \( t_x \) and \( t_y \) is calculated as the weighted average of the tag distributions for the items, which denotes the co-occurrence distribution between tags for such an item. The co-occurrence distribution of a tag for all items in a social tagging system is calculated by Eq. (3):

\[
 p_{t_y}(t_x) = \sum_{m \in I} q(t_x|m)Q(m|t_y) (3)
\]

where,

\[
 Q(m|t_y) = \frac{n(m,t_y)}{n(t_y)} \text{ on } I.
\]

\[
 q(t_x|m) = \frac{n(m,t_x)}{N(m)} \text{ on } m \in I.
\]

### 4 The Proposed approach – exploiting semantic similarity of tags for collaborative filtering

Fig.1 demonstrates the overall process in our proposed approach with utilizing the two main CF steps: i) generating a neighborhood and ii) recommending relevant items.
The basic idea behind our study assumes that active users are interested in items that have been tagged by like-minded users, and that these tags are similar to the tags used by the same active users. The first step in our approach is to look for a set of similar users who have tagged a target item. Then we compute the semantic similarity between that similar user set and the active user. Based on the basis of this similarity, the semantic ranking of the item is computed to decide whether to recommend the target item or not. In the last step, the top items are recommended to the target user based on the basis of the semantic similarity with like-minded users. This process as is illustrated in Fig.1. 

**Editor’s Note:** The denotation of “whether” implies the existence of alternatives or possibilities; therefore, “or not” is unnecessary or redundant.

4.1 Generation of a semantic-based neighborhood
As mentioned in section 2, one of the critical tasks of a user-based CF recommender system is the generation of a set of like-minded or nearest-neighbor users to the target user. Consider two users, \( u \) and \( v \), where \( u, v \in U \). First, we obtain items \( m_{uv} \in I \) which are sharable in terms of tagging behavior between the related users. For each item \( m \in m_{uv} \), we present each user with the tags that have been posted by users \( u \) and \( v \). The tags of users \( u \) and \( v \) for item \( m \) are presented as \( (u, m, (t_{u1}, ..., t_{un})) \) and \( (v, m, (t_{v1}, ..., t_{vn})) \), respectively, where both \( t_{un} \) and \( t_{vn} \subseteq T \). For each tag \( t_u \in (t_{u1}, ..., t_{un}) \) and \( t_v \in (t_{v1}, ..., t_{vn}) \), we calculate the semantic-similarity-of-tag (STSIm) value. STSIm\((t_u, t_v)\) can be calculated by using Eq. (2) if both \( t_u \) and \( t_v \) exist in the WordNet lexicon; otherwise, the value can be calculated by using Eq. (3) if one of the tags does not exist in WordNet. Based on the basis of the STSIm value for the tags given by both user \( u \) and user \( v \) to \( m_{uv} \), we can determine the semantic similarity between the two users by Eq. (4).

\[
SUsim(u, v) = \sum_{m \in m_{uv}} \sum_{t_u, t_v} \text{STSIm}(t_u, t_v) \quad (4)
\]

In Eq. (4), \( t_u \) is denotes the tags posted by user \( u \) on item \( m \), and \( t_v \) is denotes user \( v \)'s tags on item \( m \), where \( m \in m_{uv} \). The higher the \( SUsim \) value between the two users \( m_{uv} \), the more similar they are.

Finally, for a given user \( u \in U \) we determine the top \( N \) users with the highest \( SUsim \) for user \( u \). We denote this set as a set of semantically similar users (SSU), and it is defined as follows:

\[
\text{SSU}_N(u) = \arg\max_{v \in U - \{u\}} (SUsim(u, v)) \quad (5)
\]

4.2 Item recommendation

When the set of \( N \) semantically similar users has been identified, the last step consists of the actual prediction for each item and the generation of the top \( M \) list of recommended items. In our approach, the basic idea of estimating relevant unseen items for the active user starts from the assumption that users prefer items that have been tagged by like-minded users. We describe this assumption as a semantic social rank from the set of SSU, and it is defined as follows:

\[
\text{SSR} (u, r) = \sum_{v \in \text{SSU}_N(u)} SUsim(u, v) \times \text{social rank}(v) \quad (6)
\]
In Eq. (6), \( r \in I - m_u \) where \( r \) denotes the items that have not been seen by user \( u \). \( m_u \) is the items tagged by user \( u \); \( I \) is the set of all items. The social rank is equal to 1 if the item has been tagged by semantically similar users; otherwise, \( I \) obtains a value equal to 0. Finally, a set of top \( M \) ranked items that obtained higher SSR scores is recommended to user \( u \).

### 5 Evaluation

#### 5.1 Dataset

The dataset used in our experiments is the hetrec2011-movielens-2k dataset dated May 2011 which has been made available to the public by Cantador et al.\(^3\). It is based on the original MovieLens10M dataset, published by the GroupLens research group. The dataset has been used in previous studies such as [32, 33]. One of the major issues when dealing with tagging data is the quality of the tags because tags are words or combinations of words that are freely assigned by users. In order to ensure the quality of our experiment and the findings, it was necessary to remove meaningless data by filtering the dataset. Since our proposed approach depends on co-occurrence distribution, we followed the filtering steps employed by previous research in this area [21, 30, 34]. We removed meaningless tags, i.e., those that had not been posted by at least two users. Also, in line with previous studies such as [35], we also removed tags that had not been assigned to at least five items because such assignment would lead to a low co-occurrence score with other tags. According to the high sparsity of the tagging dataset, we considered items that had at least 15 tags. The final pruned abridged dataset used in our study consisted of the following: 2,013 users, 500 items, 14,800 tagging records and 1,400 tags.

#### 5.2 Tag cleaning

The main problem when trying to map tags in the MovieLens dataset to WordNet is that not all the tags in the dataset are recognized by the lexicon. Specifically, 51% of the tags in the dataset were not in WordNet. Therefore, we tried to increase this percentage by streaming the original tags in the dataset. Then we used Edit Distance-based Word Similarity (Levenshtein distance) \(^3\), which is one of the well-known edit distance functions. The Levenshtein distance is defined as the number of deletions, insertions, or substitutions of characters required to transform string \( s_1 \) into another string, \( s_2 \); the lower the distance between the strings, the more similar the two strings. MovieLens tags can be mapped to WordNet lemmas if the edit distance ratio between the test tag and the WordNet lemma is greater than 85%. Undertaking this cleaning process resulted in 58.8% of the tags in the dataset being mapped to the WordNet lexicon.

#### 5.3 Evaluation matrices

\(^3\) http://www.grouplens.org
To evaluate the performance of our proposed approach, we adapted the two famous matrices from information retrieval, namely, precision and recall which judge how relevant the recommended items are to the target users [38]. Precision measures the ratio of the number of items in a list of recommendations to those that were also contained in the test set. Recall measures the ratio of the number of relevant items retrieved to the total number of relevant items in the test set. In our experiment we withheld the items that had been previously tagged by the active user, and then we calculated the precision and recall for user \( u \) as follows:

\[
\text{Precision}(u) = \frac{|\text{Test}(u) \cap \text{TopM}(u)|}{|\text{TopM}(u)|}
\]

(7)

\[
\text{Recall}(u) = \frac{|\text{Test}(u) \cap \text{TopM}(u)|}{|\text{Test}(u)|}
\]

(8)

To make our result realistic, we considered that some users would have large tagging records, whereas others would make just a few tags for items. Next, we calculated the precision and recall for each user in the user space. Then the Average Precision@M (AP) and the Average Recall@M (AR) was calculated by the following equations:

\[
\text{AP}(U) = \frac{1}{U} \sum_{u \in U} \text{Precision}(u)
\]

(9)

\[
\text{AR}(U) = \frac{1}{U} \sum_{u \in U} \text{Recall}(u)
\]

(10)

However, according to the number of recommended items, the values of precision and recall conflict with each other. Generally, an increment in the number of items recommended tends to increase recall but decreases precision [38]. Therefore, we also considered the F1 measure, which combines both recall and precision with an equal weight in a single value [38]. The F1 measure is denoted by the following formula:

\[
\text{F1} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}
\]

(11)

In order to compare the performance of our proposed approach, we compared our approach with a popular tagging approach [39, 40] based on classical cosine similarity and presented as (cosine_CF), which depends on users’ tagging histories. The results of this comparison are discussed in the next section.

5.4 Results and discussion

In this section we present the results of our experiment with respect to the quality of the items recommended. We compared the performance of the top M recommendations for our proposed approach with a popular tagging approach based on the cosine similarity measure [39, 40]. The experiment was done with a variant number of recommended items, we by considered considering M from 5 to 100 with an increment of 5. Furthermore, we also considered the number of K similar users.

The results shown in Fig.2 and Fig.3 yield an interesting finding. We can observe that precision gradually decreases while recall increases with the increment in the top M recommended items.
One possible explanation for this result is that with the increment of M recommended items, more false positives are likely to be returned in the recommendation, thereby resulting in low precision; whereas, more true positives are likely to be returned for the increment of M recommended items that obtain higher recall values. This pattern of findings is very common in information retrieval research. However, our proposed approach (denoted semantic_CF in the Fig.2 and Fig.3) outperforms traditional CF (denoted cosine_CF) in terms of precision and recall, as shown in Fig.2 and Fig.3.

Fig.4 and Fig.5 also show that the proposed approach outperforms the cosine-based approach in terms of F1 with regard to number of recommended items and number of neighbors. The M = 10 shows the highest F1 measure, indicating that higher values of M will result in more ‘junk’ recommendations.
We also examined the F1 measure with various numbers of top K similar users. The neighborhood selection of our proposed approach was found to differ from that of the classical cosine-based similarity. We presumed that this is because the proposed approach is based on users’ semantic perspectives on tags, whereas, the cosine-based approach depends solely on the co-occurrence tagging between users.

Our approach outperforms the cosine-based in all variants of top K users. The superiority of our approach, however, decreases when the number of top K users ranges between 50 and 70, due to the effect of lower precision; whereas, the cosine-based is hardly affected by the increment in the neighbor size.

### 6 Conclusion and Future Work

This paper has presented an approach for deriving semantic similarity between users (i.e., neighborhood) by exploiting user tags. The idea stems from the belief that, ‘similar’ tags provided by different users might indicate their relatedness and potential input for recommender systems. However, most tagging activities are subject to little-to-no control in terms of terms and vocabulary used. Therefore, different tags can be semantically mean the
same equivalent. Therefore, in order to overcome such a situation, we used the WordNet to assist in measuring the semantic relatedness between tags. In the case of non-existence words not existing in the WordNet database, the co-occurrence distribution measure was used.

Evaluation on of the MovieLens dataset has shown interesting and promising results. The proposed approach outperforms the conventional tag-driven user-based CF based on the basis of the cosine-based similarity in terms of precision, recall and harmonic-means (F1) measures. Hence, it shows demonstrates that representing simple semantic information is capable of enhancing the performance of recommendation systems. However, there is no doubt that the complexity and extra processing required to perform implement the semantic processing analysis might be the setback a disadvantage of this approach.

Our future works projects include evaluating the approach on a different or bigger larger dataset and for further compare comparison with other state-of-the-art approaches. With the emergence of the Semantic Web and particularly Linked Open Data (LOD) in particular, expansion of tags to such open data is another potential work in this area.

References

12. Mahapatra S, Tareen A, Yang Y. A cold start recommendation system using item correlation and user similarity. ACM Transactions on Information Systems (TOIS), 2011(CSE 635 ’11 Buffalo, NY USA)

14. Carl kadie J S B D H. Empirical Analysis of Predictive Algorithms for Collaborative Filtering. Microsoft Research Microsoft Corporation One Microsoft Way Redmond, WA, 1998, 98052 Editor’s Note: Check the author’s name with the source. Is there only one author? Does he really have 5 initials in addition to (what appears to be) the personal name Carl? Remember to place (what appears to be) the surname [family name] Kadie first.


22. Au Yeung C, Iwata T. Capturing implicit user influence in online social sharing. Proceedings of the 21st ACM conference on Hypertext and hypermedia. Toronto, Ontario, Canada: ACM, 2010. 245-254 Editor’s Note: Verify the 1st author’s name. I think Yeung is the surname, in which case the syntax should probably be Yeung Au C; however, the correct syntax might be Au-Yeung C—[notice inserted hyphen] if the author has a compound surname.


