# 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.[←delete period & insert comma→], ~~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, [←insert comma] ~~or~~ i.e., [←insert comma] "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, [←insert comma] where ~~it is often the case that~~ often there are insufficient preferences to make recommendations. ~~In addition~~ Moreover, there is the data-sparsity [←notice inserted hyphen] 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, [←delete space between previous letter and comma] 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 ~~improvement~~ of recommendations.

**Keywords:** [←insert colon] 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. ~~It~~ 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, [←insert comma] ~~and~~ a ~~huge~~ substantial amount of new information ~~is~~ being produced every day. Since this phenomenon is already exceeding human processing capabilities, [←insert comma] ~~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~~ for items that are relevant and likely fit their needs [4], [←delete comma & insert period]. ~~and~~ Furthermore, [←notice inserted comma] collaborative tagging systems such as Delicious[1], YouTube[2], Flicker[3] and Twitter[4] ~~are allowing~~ allow the creator or visitors ~~of such content~~ to assign freely chosen keywords or tags [5] 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 [←notice inserted apostrophe] 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, [←insert comma] two users are similar if they co-rate a particular item with similar score values. Consider two users, $u_1$ and $u_2$, both of whom rate the movie "Avatar" with a similar high score. In traditional user-based approaches, $u_1$ and $u_2$ are considered similar and "like-minded" users. However, traditional methods discard the semantic perspective of the respective users, where $u_1$ ~~gives~~ awards a high score because he likes "science fiction" movies, whereas ~~the other user~~ $u_2$ 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 $u_1$ post tags (java, tour) on an item, $u_2$ post tags (java, XML) on an item, and $u_3$ post a tag (RDF) on another item. In traditional methods, $u_1$ and $u_2$ are similar because both ~~of them have~~ tagged (java); [←insert semi-colon] ~~and~~ however, [←notice inserted comma] there is no similarity between $u_2$ and $u_3$. Unfortunately, the similarity is incorrect in this scenario because traditional methods do not distinguish between "Java," the island, for $u_1$ and "java," the programming language, [←notice 4 inserted commas] for $u_2$. 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 *semantics of the annotation or tags ~~given~~ attached by both users. Therefore, instead of considering the ~~co-occurrence~~ 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 ~~rest~~ remainder of this ~~paper~~ report is organized as follows: Section 2 presents an overview of the traditional collaborative filtering approach and related work; [←delete period & insert semi-colon] Section 3 describes the preliminaries related to the proposed approach; [←delete period & insert semi-colon] ~~and~~ Section 4 ~~describes~~ introduces the proposed approach; [←delete period & insert semi-colon] Section 5 ~~presents~~ describes the experiments and results; [←delete period & insert semi-colon] and finally, Section 6 offers a ~~our~~ 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 (~~hereinafter~~ CF hereafter) is considered to be the best recommendation technique that automates the process of the "word-of-mouth" paradigm [9], [←delete comma] in estimating the utilization of unseen items ~~by a user~~. ~~Collaborative filtering~~ CF compares ~~a~~ users ~~based~~ 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 ~~forming~~ formation of a similar neighborhood, [←delete comma] 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, $\mathrm{u}$ and $\mathrm{v}$, is calculated as the cosine angle between ~~users'~~ the corresponding feature vectors, [←insert comma] as follows:

$$\mathrm{similarity}(u,v) = \cos(\theta) = \frac{\vec{u} \cdot \vec{v}}{\| \vec{u} \| \cdot \| \vec{v} \|} \quad (1)$$

When these neighbors have been found, [←insert comma] the next step is the process of ~~estimation 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 ~~the~~ this user has on ~~the~~ this 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].} [←Move this {bracketed} sentence to the beginning of the next paragraph.]

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 [←notice inserted hyphen] 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~~ annotating given items. The proposed approach presented in this ~~paper~~ report, however, falls into the latter category, i.e., [←insert 2 commas] 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, [←insert comma] 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, [←insert comma] 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 [←notice inserted hyphen] 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), [←insert comma] ~~which is~~ used to calculate the similarity between tags and web queries), [←insert comma] 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 [←notice inserted hyphen] 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," [←delete period & insert comma] 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., [←insert comma] [22] ~~suppose~~ assuming that it is still possible that users can influence one another in the process of item adoption through various implicit mechanisms, [←delete period & insert comma] ~~Au et al.~~ capture the influence preferences among the users in a social system, ~~which they~~ considering the preferences to be related to the tagging behavior between users for certain items.

Our approach differs from the aforementioned studies in that ~~our~~ we 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 these~~ 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 ~~that~~ 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 ~~users~~ 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 $U = \{u_1, u_2, u_3, \ldots\ldots, u_m\}$, [←insert comma] ~~and~~ a list of items $I = \{m_1, m_2, m_3, \ldots\ldots, m_k\}$, [←insert comma] and a list of tags $T = \{t_1, t_2, t_3, \ldots\ldots, t_n\}$, the folksonomy $F = \langle U, I, T, Y \rangle$, 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 $u_1$ posts $t_1, t_2$ for $item_1$, which can be represented as $(u_1, m_1, (t_1, t_2))$.

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:

$$Sim_{Lin}(c_1, c_2) = \frac{2 \times IC(C_{MICA})}{IC(c_1) + IC(c_2)} \quad (2)$$

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 $m$, [←delete comma & insert semi-colon→]; ~~Let~~ $n(t) = \sum_m n(m,t)$, [←insert comma] ~~be~~ the number of tag occurrences for all items in $I$, [←delete comma & insert semi-colon→]; and ~~let~~ $N(t) = \sum_t n(m,t)$, [←insert comma] ~~be~~ the number of tag occurrences for item m, where $m$ is an instance of $I$, [←delete comma] and m ∈ $I$. 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). [←delete period & insert colon→]:

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

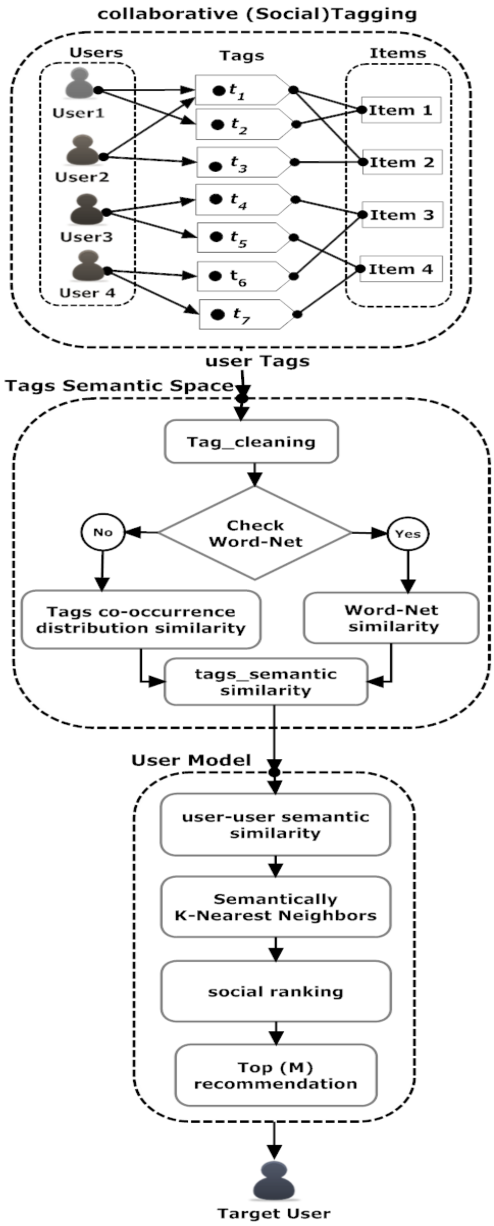
where,

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

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

## 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.




Revisions to text in image: Collaborative | Semantic Tag Space | Tags'…distribution…| tags'_semantic…| Semantically nearest K-neighbors or Nearest semantic K-neighbors?

**Fig.1** System overview of collaborative filtering based on semantic tags.

The basic idea behind our study assumes that active users are interested in items that have been tagged by like-minded users, [←delete comma] 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 [←notice inserted hyphen] users ~~who have~~ having ~~similar~~ tastes similar to the target user's. Consider two users, $u$ and $v$, where $u, v \in U$. First, we obtain items

$m_{u,v} \in I$, which are sharable in terms of tagging behavior between the related users. For each item $m \in m_{u,v}$, 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 $(u, m, (t_{u1}, \ldots, t_{un}))$ and $(v, m, (t_{v1}, \ldots, t_{vn}))$, respectively, where both $t_{un}$ and $t_{vn} \subseteq T$. For each tag $t_u \in (t_{u1}, \ldots, t_{un})$ and $t_v \in (t_{v1}, \ldots, t_{vn})$, we calculate the semantic-similarity-of-tag [←notice 3 inserted hyphens] (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, [←delete comma & insert semi-colon→]; otherwise, [←insert comma] ~~it~~ 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 $m_{u,v}$, [←insert comma] we can determine the semantic similarity between the two users by Eq. (4). [←delete period & insert colon→]:

$$SUsim(u,v) = \sum_{m \ni m_{u,v}} \sum_{t_u, t_v} STSim(t_u, t_v) \quad (4)$$

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 $m \in m_{u,v}$. The higher the SUsim value between the two users $m_{u,v}$, [←insert comma] the ~~more similar they are~~ greater their similarity.

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) , [←insert comma] ~~and~~ defined ~~it~~ as follows:

$$SSU_N(u) = \mathrm{argmax}_{v \ni U-\{u\}} (SUsim(u,v)) \quad (5)$$

4.2 Item recommendation

When the set of N semantically similar users has been identified, [←insert comma] 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, [←insert comma] ~~and it is~~ defined as follows:

$$SSR(u, r) = \sum_{v \ni SSU_N(u)} SUsim(u,v) \times social\ rank(v) \quad (6)$$

In Eq. (6), $r \in I - m_u$ where $r$ ~~is~~ denotes the items that have not been seen by user $u$,[←delete comma & insert semi-colon→]; $m_u$ , [←insert comma] ~~is~~ the items tagged by user $u$ ; [←insert semi-colon] and $I$, [←insert comma] ~~is~~ the set of all items. The social rank is equal to 1 if the item has been tagged by semantically similar users; otherwise, [←insert comma] it 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.[31]. It is based on the original MovieLens10M dataset, published by the Group Lens5 research group. ~~The~~ This 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, [←insert comma] 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. , [←insert comma] those that had not been posted by at least two users. Also, in line with previous studies such as [35] , [←insert comma] we ~~also~~ removed tags that had not been assigned to at least five items, [←delete comma] because ~~this~~ such assignment would lead to a low co-occurrence score with other tags. According to the high sparsity of the tagging dataset, [←insert comma] we considered items that had at least 15 tags [36]. 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; [←delete semi-colon & insert period→]. Specifically, [←notice inserted comma] 51% of the tags in the dataset were not in WordNet. Therefore, [←insert comma] we tried to increase this percentage by ~~steaming~~ streaming [?] the original tags in the dataset. Then we used Edit Distance-based Word Similarity (Levenshtein distance) [37], ~~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 $s1$ into another string, $s2$; the ~~lower~~ shorter the distance between the strings, [←insert comma] the ~~more similar the two strings~~ greater the similarity. 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

[5] 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, [←delete comma & insert semi-colon→]; ~~and~~ then we calculated the precision and recall for user u as follows:

$$\mathrm{Precision}(u) = \frac{|\mathrm{Test}(u) \cap \mathrm{TopM}(u)|}{|\mathrm{TopM}(u)|} \quad (7)$$

$$\mathrm{Recall}(u) = \frac{|\mathrm{Test}(u) \cap \mathrm{TopM}(u)|}{|\mathrm{Test}(u)|} \quad (8)$$

To make our result realistic, we considered that some users would have large tagging records, [←delete comma & insert semi-colon→]; ~~while~~ whereas, [←notice inserted comma] others ~~users~~ would make ~~just~~ only a few tags for items, [←delete comma & insert period→]. ~~and~~ Next, we calculated the precision and recall for each user in the user space. [←insert period] ~~and~~ Then the $\mathrm{Average\ Precision@M}$ (AP) and the $\mathrm{AverageRecall@M}$ (AR) was calculated by the following equations:

$$\mathrm{AP}(U) = \frac{1}{U}\sum_{u \ni U}^{U} \mathrm{Precision}(u) \quad (9)$$

$$\mathrm{AR}(U) = \frac{1}{U}\sum_{u \ni U}^{U} \mathrm{recall}(u) \quad (10)$$

However, according to the number of recommended items, the values of precision and recall conflict with each other; [←delete semi-colon & insert period→]. Generally, an increment in the number of items recommended tends to increase recall but ~~it~~ decreases precision [38]. Therefore, we also considered the F1 measure, [←insert comma] which combines both recall and precision with an equal weight in a single value [38]. The F1 measure is denoted by the following formula:

$$F1 = \frac{2 \times \mathrm{recall} \times \mathrm{precision}}{\mathrm{recall} + \mathrm{precision}} \quad (11)$$

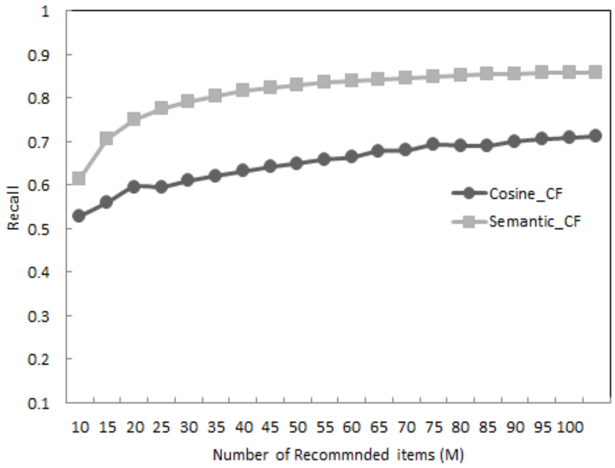
In order to compare the performance of our proposed approach, we compared our approach with a popular tagging approach [39, 40] based on classical cosine similarity and presented as (cosine_CF), which depends on users' tagging histories. [←insert period] ~~and~~ The results of this comparison are discussed 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 ~~of~~ on top M recommended items ~~is~~ was done with a variant number of recommended items, [←delete comma] ~~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.

**Fig.2** Precision with increment of top M recommended items

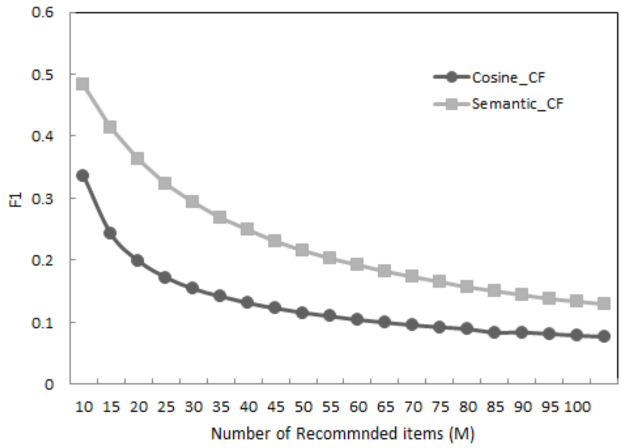



Revision to text in image: Recommended

**Fig.3 R**ecall with increment of 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, [←delete comma & insert semi-colon→]; whereas, [←insert comma] 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 ~~shows~~ ~~show~~ indicate that the proposed approach outperforms the cosine-based approach in terms of F1 ~~in terms of~~ with regard to number of recommended items and number of neighbors. The M = 10 shows the highest F1 measure, [←insert comma] indicating that higher values of M will result in more 'junk' recommendations.




Revision to text in image: Recommended

**Fig.4** F1 ~~measures~~ measure values with increment of Top M recommended items

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~~ the difference occurred because the proposed approach is based on users' semantic perspectives on tags, [←delete comma & insert semi-colon→]; whereas, [←insert comma] the cosine-based approach depends ~~purely~~ solely on the co-occurrence tagging between users.

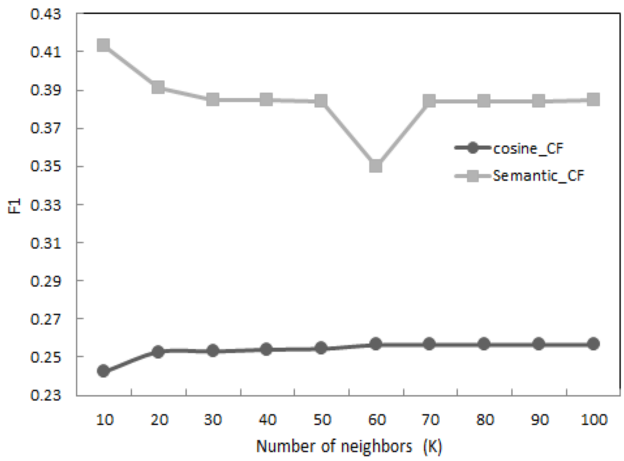



**Fig.5** F1 measures with variant size of top K similar users

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 ~~is~~ ranges between 50 and 70, [←insert comma] due to the effect of lower precision, [←delete comma & insert semi-colon→]; whereas, [←insert comma] the cosine-based is hardly affected by the increment in the neighbor size.

## 6 Conclusion and Future Work

This ~~paper~~ report has presented an approach for deriving semantic similarity between users (i.e., neighborhood) [←notice inserted comma] by exploiting user tags. The idea stems from the ~~believe~~ belief that, [←delete comma] 'similar' tags provided by different users might ~~indicates~~ indicate their relatedness and potential input for recommender systems. However, most tagging activities are ~~with~~ subject to little-to-no [←notice 2 inserted hyphens] control ~~in terms of~~ with regard to terms and vocabulary used. Therefore, different tags can be semantically ~~mean the~~

~~same~~ equivalent. ~~Therefore~~ Thus, 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 ~~is~~ 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) [41] in particular, expansion of tags to such open data is another potential work in this area.

## References


1. ~~Sharma P.~~ Core Sharma P. Characteristics of web 2.0 services. Online at http://www.techpluto.com/web-20-services, 2008
2. Siersdorfer S, Sizov S. Social recommender systems for web 2.0 folksonomies. Proceedings of the 20th ACM Conference on Hypertext and Hypermedia, 2009: 261-270
3. Adomavicius G, Tuzhilin A. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE Trans. on Knowl. and Data Eng., 2005, 17(6): 734-749 Editor's Note: The stylesheet says: "Journal names should be spelled out in full. Examples are given in [1–4]."
4. Su X, Khoshgoftaar T M. A survey of collaborative filtering techniques. Advances in Artificial Intelligence, 2009, 2009: 4
5. Golder S A, Huberman B A. Usage patterns of collaborative tagging systems. Journal of Information Science 2006, 32(2): 198-208
6. Zhang Z-K, Zhou T, Zhang Y-C. Personalized recommendation via integrated diffusion on user–item–tag tripartite graphs. Physica A: Statistical Mechanics and its Applications, 2010, 389(1): 179-186
7. Han H, Cai Y, Shao Y, Li Q. Improving recommendation based on features' co-occurrence effects in collaborative tagging systems, web technologies and applications. In: Sheng Q, Wang G, Jensen C, Xu G, eds: Springer Berlin / Heidelberg, 2012, 652-659 Editor's Note: It appears that you omitted the name of the larger collection/compilation which contains this article.
8. Lee D H, Brusilovsky P. Improving recommendations using WatchingNetworks in a social tagging system. Proceedings of the 2011 iConference, 2011: 33-39
9. Schafer J B, Konstan J A, Riedl J. E-commerce recommendation applications. Data mining and knowledge discovery, 2001, 5(1): 115-153
10. Krishnan V, Narayanashetty P K, Nathan M, Davies R T, Konstan J A. Who predicts better?: Results from an online study comparing humans and an online recommender system. ACM, 2008. 211-218
11. Lam X N, Vu T, Le T D, Duong A D. Addressing cold-start problem in recommendation systems. Proceedings of the 2nd International Conference on Ubiquitous Information Management and Communication, 2008: 208-211
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)

13. Linden G, Smith B, York J. Amazon. com recommendations: Item-to-item collaborative filtering. Internet Computing, IEEE, 2003, 7(1): 76-80
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.
15. Su X, Khoshgoftaar T M. A survey of collaborative filtering techniques. ~~Journal ,~~Advances in Artificial Intelligence, 2009, 2009: 2-2
16. Hamouda S, Wanas N. PUT-Tag: personalized user-centric tag recommendation for social bookmarking systems. Social Network Analysis and Mining, 2011, 1(4): 377-385
17. Tso-Sutter K H L, Marinho L B, Schmidt-Thieme L. Tag-aware recommender systems by fusion of collaborative filtering algorithms. Proceedings of the 2008 ACM Symposium on Applied Computing, 2008: 1995-1999
18. Gemmis M D, Lops P, Semeraro G, Basile P. Integrating tags in a semantic content-based recommender. Proceedings of the 2008 ACM Conference on Recommender Systems, 2008: 163-170
19. Bao S, Xue G, Wu X, Yu Y, Fei B, Su Z. Optimizing web search using social annotations. Proceedings of the 16th International Conference on World Wide Web, 2007: 501-510
20. Liang H, Xu Y, Li Y, Nayak R. Tag Based Collaborative Filtering for Recommender Systems. Proceedings of the 4th International Conference on Rough Sets and Knowledge Technology, 2009: 666-673
21. Sen S, Vig J, Riedl J. Tagommenders: connecting users to items through tags. Proceedings of the 18th International Conference on the World Wide Web, 2009: 671-680
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.
23. Liang H, Xu Y, Li Y, Nayak R. Tag-based collaborative filtering for recommender systems. In: Wen P, Li Y, Polkowski L, Yao Y, Tsumoto S, Wang G, eds. Rough sets and knowledge technology: Springer Berlin Heidelberg, 2009, 666-673
24. Fellbaum C. WordNet: An Electronic Lexical Database. MIT Press, 1998 Editor's Note: This book appears to have more than one author. See http://www.amazon.com/s/ref=nb_sb_noss?url=search-alias%3Daps&field-keywords=24.%20Fellbaum%20C.%20WordNet%3A%20An%20electronic%20lexical%20database.%20MIT%20.

25. Pedersen T, Patwardhan S, Michelizzi J. WordNet:: Similarity: measuring the relatedness of concepts. Demonstration Papers at HLT-NAACL 2004, 2004: 38-41[delete duplicate colon in above line]
25. Hotho A, J R, #228, schke, Schmitz C, Stumme G. Information retrieval in folksonomies: search and ranking. Proceedings of the 3rd European conference on The Semantic Web: research and applications, 2006: 411-426 Editor's Note: The assorted characters highlighted in purple do not make good sense in this context. Consult the source and make the appropriate correction.
26. Hu J, Wang B, Liu Y, Li D Y. Personalized tag recommendation using social influence. Journal of Computer Science and Technology, 2012, 27(3): 527-540
27. Lin D. An information-theoretic definition of similarity. Proceedings of the Fifteenth International Conference on Machine Learning, 1998: 296-304
28. Budanitsky A, Hirst G. Evaluating WordNet-based measures of lexical semantic relatedness. Comput. Linguist., 2006, 32(1): 13-47 Editor's Note: The stylesheet says: "Journal names should be spelled out in full. Examples are given in [1–4]."

29. Wartena C, Brussee R, Wibbels M. Using tag co-occurrence for recommendation. Proceedings of the 2009 Ninth International Conference on Intelligent Systems Design and Applications, 2009: 273-278
30. Cantador N, Konstas I, Jose J M. Categorising social tags to improve folksonomy-based recommendations. Web Semantics, 2011, 9(1): 1-15
31. Jones C, Ghosh J, Sharma A. Learning multiple models for exploiting predictive heterogeneity in recommender systems. Proceedings of the 2nd International Workshop on Information Heterogeneity and Fusion in Recommender Systems, 2011: 17-24
32. Said A, Luca E W D, Kille B, Jain B, Micus I, Albayrak S. KMulE: A framework for user-based comparison of recommender algorithms. Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces, 2012: 323-324
33. Zhang Z-K, Zhou T, Zhang Y-C. Tag-aware recommender systems: A state-of-the-art survey. Journal of Computer Science and Technology, 2011, 26(5): 767-777
34. Vig J, Sen S, Riedl J. Tagsplanations: Explaining recommendations using tags. Proceedings of the 14th International Conference on Intelligent User Interfaces, 2009: 47-56
35. Kim H-N, Alkhaldi A, El Saddik A, Jo G-S. Collaborative user modeling with user-generated tags for social recommender systems. Expert Systems with Applications, 2011, 38(7): 8488-8496
36. Levenshtein V I. Binary codes capable of correcting deletions, insertions and reversals. Soviet Physics Doklady, 1966, 10: 707-710 Editor's Note: Verify accuracy of word highlighted in purple.
37. Herlocker J L, Konstan J A, Terveen L G, Riedl J T. Evaluating collaborative filtering recommender systems. ACM Transactions on Information Systems (TOIS), 2004, 22(1): 5-53
38. Durao F, Dolog P. Extending a hybrid tag-based recommender system with personalization. Proceedings of the 2010 ACM Symposium on Applied Computing, 2010: 1723-1727
39. Jäschke R, Marinho L, Hotho A, Schmidt-Thieme L, Stumme G. Tag recommendations in social bookmarking systems. AI Communications 2008, 21(4): 231-247
40. Heath T, Bizer C. Linked Data: Evolving the web into a global data space. Synthesis Lectures on the Semantic Web: Theory and Technology, 2011, 1(1): 1-136