Granular Semantic User Similarity in the Presence of Sparse Data

Granular semantic user similarity in the presence of sparse data

Francesco Osborne, Silvia Likavec and Federica Cena

Università di Torino, Dipartimento di Informatica, Torino, Italy
{osborne,likavec,cena}@di.unito.it

Abstract. Finding similar users in social communities is often challenging, especially in the presence of sparse data or when working with heterogeneous or specialized domains. When computing semantic similarity among users it is desirable to have a measure which allows to compare users w.r.t. any concept in the domain. We propose such a technique which reduces the problems caused by data sparsity, especially in the cold start phase, and enables granular and context-based adaptive suggestions. It allows referring to a certain set of most similar users in relation to a particular concept when a user needs suggestions about a certain topic (e.g. cultural events) and to a possibly completely different set when the user is interested in another topic (e.g. sport events). Our approach first uses a variation of the spreading activation technique to propagate the users' interests on their corresponding ontology-based user models, and then computes the concept-biased cosine similarity (CBC similarity), a variation of the cosine similarity designed for privileging a particular concept in an ontology. CBC similarity can be used in many adaptation techniques to improve suggestions to users. We include an empirical evaluation on a collaborative filtering algorithm, showing that the CBC similarity works better than the cosine similarity when dealing with sparse data.

Keywords: Recommender systems, ontology, propagation of interests, user similarity, data sparsity, cosine similarity

1 Introduction

One of the most basic assumptions in social communities is that people who are known to share specific interests are likely to have additional related interests. This assumption is used by many approaches, especially in collaborative-filtering recommender systems, which recommend items that people with tastes and preferences similar to the target user liked in the past [13]. User profiles can be compared using a range of metrics, such as cosine similarity [16], the Pearson Correlation Coefficient [1], Jaccard's index [6], to name just a few.

However, these traditional methods for computing similarity do not come without limitations. First, similarity is difficult to calculate in the presence of data sparsity, a problem which arises when the user ratings are spread over items that seldom overlap. This especially happens in two situations: at the beginning of the interaction, when we

do not know much about the user (the *cold start* problem [12]) and when the domain is huge and we do not have the user interest values for all the concepts in the domain.

Second, these similarity measures usually take into account all the concepts in the domain. This solution may not always be flexible enough, especially in a non-specialized and heterogeneous domain, i.e. the domain covering a range of very different topics and subtopics. *In this case, rarely similar people show to have the same tastes for all the different concepts.* Even when the domain is not so heterogeneous, it frequently happens that people have the same tastes for some portions of the domain, but different ones for others. There are many cases in which it is very important to be able to be more granular and to establish a different set of similar users for different concepts. For example, two users can have a very different tastes for many elements of the domain and still be very similar w.r.t. a certain topic or subculture, such as `Rock_Music` or `Wine`. Computing the user similarity over all the elements of the domain does not consider this aspect, wasting useful information, which is a shame especially in the presence of sparse and poor data.

To address these two issues (sparsity problem and computing similarity over the whole domain), we introduce *the concept-biased cosine similarity* (CBC similarity), a novel similarity metric designed to privilege certain concept in an ontology, according to the recommendation context, and overcome data sparsity.

The prerequisites for our approach are the following:

- *semantic representation of the domain knowledge* using an OWL ontology[1] where domain concepts are taxonomically organized;
- *user model defined as an overlay over the domain ontology*;
- a *strategy for propagation of the user interest values in the user model*.
- a methodology to calculate for each user a *set of similar users w.r.t. a certain topic*.

Our approach consists of three steps. First, we use a variation of the spreading activation technique, introduced in [3, 4], to propagate the users interests on their ontology based user models (Sect. 2). Then, we pre-compute a matrix in which every pair of users is assigned a similarity score using cosine similarity for each concept in the ontology for which there is enough feedback from both users. Finally, we calculate the concept-biased cosine similarity (CBC similarity), a variation of the cosine similarity designed for privileging a particular concept in an ontology (Sect. 3).

The main contribution of the paper is this novel method for calculating fine grained user similarity for specific sub-portions of the domain (CBC similarity) even in the presence of sparse data. We tested the CBC similarity in a collaborative filtering algorithm and showed that it outperforms the standard cosine similarity when dealing with sparse data. The method is general and applicable in wide variety of domains, provided that the domain is described using an ontology.

The remainder of the paper is organized as follows. In Sect. 2 we provide some details of our ontology-based user model, how to determine the user interest in domain concepts and how to subsequently propagate it to similar concepts in the domain ontology. We describe how to calculate the similarity between users for different specific sub-portions of the domain ontology in Sect. 3 . We present the results of the evaluation

[1] `http://www.w3.org/TR/owl-features/`

of our approach in Sect. 4, followed by related work in Sect. 5. Finally, we conclude and give some directions for future work in Sect. 6.

2 User interest

Our technique for finding similar users in particular contexts as the first step includes the determination and propagation of user interests in order to populate and update the user model. User interests are propagated to similar concepts, using various approaches, in order to reduce data sparsity and reach distant concepts. We start this section with the definition of the user model in Sect. 2.1, followed by a brief overview of how to determine the user interest in Sect. 2.2 and conclude with some possible propagation techniques in Sect. 2.3.

2.1 User model definition

In order to satisfy the previously listed requirements, our domain is represented using an OWL ontology[2]. As will be seen later, the way in which this ontology is formulated, influences the choice of propagation technique.In our approach we employ an *ontology-based user model*, defined as an overlay over the domain ontology. More precisely, each ontological user profile is an instance of the domain ontology, and each node in the ontology has an interest value associated to it. This means that each node N in the domain ontology can be seen as a pair $\langle N, \mathcal{I}(N)\rangle$, where $\mathcal{I}(N)$ is the interest value associated to node N.

2.2 Determining user interest

We describe here the initial step of our approach which considers determining the user interest in domain concepts and is based on our previous work described in [3, 4]. When the user provides feedback for a certain concept in the domain ontology[3], this feedback is implicitly recorded by the system so that it can be further used to calculate user interest values for other domain concepts.

For each concept N in the domain ontology we distinguish two types of interest values: *sensed interest* and *propagated interest*. Given the user feedback for the concept N, we first calculate the user *sensed interest* as follows:

$$\mathcal{I}_{\mathrm{S}}(N) = \frac{l(N)}{\mathrm{MAX}(1 + e^{-\mathsf{f}(N)})}$$

where $l(N)$ is the level of the node receiving the feedback, MAX is the level of the deepest node in the ontology and $\mathsf{f}(N)$ is the user feedback for the node N. The sensed interest depends on the user feedback for the node and the position of the node in the ontology, since the nodes lower down in the ontology represent specific concepts, and as such signal more precise interest than the nodes represented by upper classes in the ontology, expressing more general concepts. Subsequently, the sensed interest value is used in *propagation phase* (see Sect. 2.3) to propagate the user interest to similar objects.

[2] `http://www.w3.org/TR/owl-ref/`

[3] How the feedback is obtained is out of the scope of this paper.

These two interest values, $\mathcal{I}_S$ and $\mathcal{I}_P$ are kept separated for each concept, and the total interest for each concepts in the ontology is calculated as:

$$\mathcal{I}(N) = \phi\mathcal{I}_O(N) + \sigma\mathcal{I}_S + \Sigma_{i=1}^{n}\pi_i\mathcal{I}_P(N_i, N)$$

where $\phi, \sigma, \pi_1, \ldots, \pi_n \in \mathbb{R}$ and $\phi+\sigma+\Sigma_{i=1}^{n}\pi = 1$, $\mathcal{I}_O$ is the *old interest value* (initially set to zero), $\mathcal{I}_S(N)$ is the sensed interest for the node N and $\mathcal{I}_P(N_i, N)$ is the total propagated interest from the node N_i to the node N, n being the number of nodes which propagate their interest values to the node N. By varying the constants ϕ, σ and π_i it is possible to assign different level of importance to either sensed or propagated interest values.

2.3 Propagating user interest values

Depending on how the ontology is designed, the concepts to which to propagate the user interest can be determined in different ways. Here we describe two different propagation techniques:

- distance-based propagation,
- property-based propagation.

Distance-based propagation One of the simplest ways to measure the similarity of two concepts in the ontology is to use the ontology graph structure and calculate the distance between nodes by using the number of edges or the number of nodes that need to be traversed in order to reach one node from the other (see [11]).

Using the distance between nodes, the user interest in a certain domain object can be propagated *vertically* in the ontology, upward to the ancestors and downward to the descendants of a given node. In this case the propagated interest is calculated by modulating the sensed interest of the node N that receives the feedback by the exponential factor which describes the attenuation with each step up or down and a weight inversely correlated with the amount of feedback already received by the node as follows:

$$\mathcal{I}_P(N, M) = \frac{e^{-kd(N,M)}}{1 + \log(1 + n(M))}\mathcal{I}_S(N)$$

where $d(N, M)$ is the distance between the node N receiving the feedback and the node M receiving the propagated interest, $n(M)$ is the number of actions performed in the past on the node M and $k \in \mathbb{R}$ is a constant.

Further improvement of distance-based propagation can be obtained with *conceptual distance* where the main idea is to modify the lengths of the edges in the ontology graph, as initially proposed in [5]. First, the set of relevant concepts is determined. Then for the relevant concepts, the notion of conceptual specificity is introduced, in order to specify their relevance in the given context. Using the conceptual specificity, the edge lengths are modified so that they exponentially decrease as the levels of the nodes increase. Then the propagated interest is calculated using these exponentially decreasing edge lengths to calculate the distances between concepts.

Property-based propagation If the ontology has explicit specification of the concepts' properties, we can use these properties to calculate the similarity between the concepts, the distance between them. To do this, we can adopting the approach described in [4],

and decide to which elements to propagate the user interest. This propagation does not have any particular direction and permits propagation of user interests to various nodes in the ontology. The propagated interest value is calculated using the hyperbolic tangent function as:

$$\mathcal{I}_{\mathrm{P}}(N, M) = \frac{e^{2\mathrm{SIM}(N,M)} - 1}{e^{2\mathrm{SIM}(N,M)} + 1} \mathcal{I}_{\mathrm{S}}(N)$$

where $\mathcal{I}_{\mathrm{S}}(N)$ is the sensed interest for the node N which received the feedback and $\textsc{sim}(N, M)$ is the *property-based similarity* between the node N which received the feedback and the node M which receives the propagated interest.

Property-based similarity calculates the similarity of classes, starting from Tversky's feature-based model of similarity [17], where similarity between objects is a function of both their common and distinctive characteristics. For two domain elements N_1 and N_2, for each property p, we calculate $\textsc{cf}_p$, $\textsc{df}_p^1$ and $\textsc{df}_p^2$, which denote how much p contributes to common features of N_1 and N_2, distinctive features of N_1 and distinctive features of N_2, respectively, depending on how p is defined in N_1 and N_2. We calculate the similarity between N_1 and N_2 as follows:

$$\textsc{sim}(N_1, N_2) = \frac{\textsc{cf}(N_1, N_2)}{\textsc{df}(N_1) + \textsc{df}(N_2) + \textsc{cf}(N_1, N_2)}.$$

The property-based similarity of equivalent classes is equal to 1, whereas for instances the values-property pairs declared for each instance are compared.

The choice of one of these propagation methods depends on the ontology structure. If the ontology has a deep taxonomy, (conceptual) distance-based propagation can be used. Our experiments showed that conceptual distance performs better than the standard one. If the ontology does not have a deep hierarchy and the classes are defined by means of properties, property-based propagation is the most suitable one.

3 The concept-biased cosine similarity

In this section we propose a novel technique for computing granular semantic user similarity w.r.t. any concept formalized in the ontology, called *concept-biased cosine similarity* or *CBC similarity*. The CBC similarity measure introduces a bias in the standard cosine similarity measure, in order to privilege a certain concept in the domain ontology. It is particularly useful when the user is interested in a specific part of the domain and there is a need to find the most similar users concerning that part of the domain. For example, we can refer to a set of the most similar users when a user needs suggestions about one topic (e.g. cultural events) and to a possibly different set when the user is interested in another topic (e.g. sport events). This similarity measure can be used by many kinds of recommendation techniques which build on similarity measures between users, allowing a more granular prospective on the domain. However, the investigation of recommendation strategies is out of the scope of this paper.

The algorithm to compute the CBC similarity exploits a pre-computed tridimensional matrix which contains for each pair of users the similarity values obtained using the standard cosine similarity w.r.t. the significant concepts of the domain ontology. The significant classes may be pre-labeled or can be defined autonomically by setting a threshold on the level or the number of their instances. Table 1 displays a part of this

structure using an ontology describing social events. The meaning is intuitive: Ann is similar to Bill w.r.t. some concepts, such as gastronomical events and cooking courses, but different w.r.t. cultural events. Hence, it makes sense to use the information from Bill's user model when recommending gastronomical events to Ann.

To compute this matrix we scan the ontology top-down and for each significant class X and each pair of users we compute the cosine similarity, using as input the interest values for the subclasses and instances of X. To save time and space, it is advisable to ignore all the subclasses of a given class which for a specific pair of users have cosine similarity below a certain threshold (0.5 in our approach) or do not include enough common feedback values to be significant. For example, if Ann and Bill have few feedback values w.r.t. `Sport_Event`, their similarity w.r.t. the significant classes which are subclasses of `Sport_Event` (e.g. `Race`, `Open_Day` etc.) are not computed.

Similarity with Ann	**Bill**	**Cindy**	**Damian**
`Event`	0.79	**0.82**	0.75
`Gastronomical_Event`	**0.89**	0.84	0.65
`Cultural_Event`	0.54	0.81	**0.94**
`Sport_Event`	0.72	**0.92**	0.6
`Cooking_Course`	**0.96**	0.84	0.69
`Concert`	0.72	0.74	**0.92**
`Tasting`	**0.83**	0.71	0.63
`Soccer_Match`	0.56	**0.97**	0.67
...	...	...	...

Table 1. Ann's similarity with Bill, Cindy and Damian w.r.t. different concepts. In bold the highest value for each row, representing the user the most similar to Ann, w.r.t. a certain class.

A naive way to use the matrix given in Table 1 would be to take directly the cosine similarity scores computed for each concept when a suggestion for that concept is needed. This approach, however, may miss the big picture. In fact, while most of the information expressing similarity w.r.t. a certain concept can be derived from the feedback values for the items classified under this concept, by not considering the other feedback values we may miss precious and more subtle information about the users' common preferences. For example, we might compute the cosine similarity using only the interest values for the subconcepts of the class `Sport_Event` and use this similarity to find similar users with the same interest for sport events. However, if we also consider the user feedback about some `Book_Presentation` which talks about `Yoga`, we would reinforce our knowledge about user interest in yoga related sport events.

We address this issue with: (i) *interest propagation* and (ii) *super-class weighing*. Interest propagation, as discussed in Sect. 2, is not only appropriate for reducing sparsity, but it also forces the interest value assigned to each class to take into consideration also the feedback values for any semantically related concepts. However, interest propagation effectiveness depends on how well the ontology is designed. For a well-designed domain ontology which describes all the explicit and subtle relationships among the domain elements, interest propagation would be enough to solve the problem of including the information contained in different classes. In a realistic scenario, instead, the

ontology is often far from perfect. For this reason, when computing the similarity w.r.t. concept X it is better not to exclude completely all the other concepts, but rather weigh them differently. This is where CBC similarity comes into picture.

Our approach includes in the computation of the CBC similarity w.r.t. X the similarity scores of the super-classes of X. For example, when the concept `Race` is taken into consideration we want a similarity metric that does not only award people similar w.r.t. `Race` but also gives a bonus to people similar with respect to `Sport_Event` and `Event`. We compute the CBC similarity for concept X using the following formula:

$$CBC(X, a, b) = w \cdot \cos(v^a(X), v^b(X)) + (1 - w)\frac{\sum_{i \in S} \frac{s_i}{T} \cos(v^a(X_i), v^b(X_i))}{\sum_{i \in S} \frac{s_i}{T}} \quad (1)$$

where

- a, b are the users we want to compare;
- X is the starting class for which we calculate the CBC similarity;
- $\cos(x, y)$ is the cosine similarity for vectors x and y;
- i refers to the element X_i of the set S composed of super classes of X;
- $v^a(X)$ and $v^b(X)$ are the interest value vectors associated to the subclasses/instances of the class X according to the user models of users a and b;
- $v^a(X_i)$ and $v^b(X_i)$ are the interest value vectors associated to the sub-classes/instances of the class X_i according to the user models of users a and b;
- s_i is the number of subclasses of X_i for which both users provided feedback;
- T is the total number of classes in the ontology;
- $0 < w < 1$ weighs the relationship between a class and its super classes.

We believe that w should depend on the sparsity degree of the data for a class: if it is high we will need more support from the super classes, whereas if it is low, it can take care of itself. Thus, in our system we estimate w for concept X as the average of the ratios between the subclasses which received a direct feedback and the total number of subclasses for each ontology-based user model which received at least 10 feedbacks.

It should be noticed that in order for CBC to be meaningful for a user, she/he should have a decent amount of feedback values for class X for which we calculate the CBC similarity. If this is not true, it is better to compute the CBC on the first super class of X that has enough feedback. For example, if the context is `Race` and the user did not provide enough feedback on `Race` concept we can check if she/he has the interest values for `Sport_Event` or for `Event`, relying to the first usable super class.

We now give an example of the last step of CBC computation, using the values in Table 1. Assume that we need to suggest to Ann a concert for tomorrow night. A standard way would be to exploit the feedback values of the most similar users according to different techniques (e.g. memory-based collaborative filtering). If we adopt the standard cosine similarity we will conclude that Cindy is the most similar user to Ann, followed by Bill and Damian. While this is true taking into consideration the whole domain, the situation may be different if we consider only the items related to the `Concert` class. Thus, taking into consideration also the context in which the suggestion is needed we can compute CBC relative to class `Concert`. In the domain ontology the super-classes of `Concert` are `Cultural_Event` and `Event`. For simplicity we assume that `Cultural_Event` subsumes one third of the items and $w = 0.5$.

$$CBC(\texttt{Concert}, Ann, Bill) = 0.72 \cdot 0.5 + \frac{0.54 \cdot 1/3 + 0.79}{1/3+1} \cdot 0.5 = 0.36 + 0.36 = 0.72$$
$$CBC(\texttt{Concert}, Ann, Cindy) = 0.74 \cdot 0.5 + \frac{0.81 \cdot 1/3 + 0.82}{1/3+1} \cdot 0.5 = 0.37 + 0.41 = 0.78$$
$$CBC(\texttt{Concert}, Ann, Damian) = 0.92 \cdot 0.5 + \frac{0.54 \cdot 1/3 + 0.79}{1/3+1} \cdot 0.5 = 0.46 + 0.40 = 0.86$$

We can see that using the CBC similarity Damian is the user most similar to Ann w.r.t. the class `Concert`, followed by Cindy and Bill.

4 Evaluation

To evaluate the CBC similarity we implemented a memory-based collaborative filtering recommender that computes the suggested rating for an item as the weighed sum of the feedback values given by the most similar users [2].

Our assumption was that using the CBC similarity would yield more accurate results than adopting the standard cosine similarity, especially in the presence of sparse data. Furthermore, we wanted to compare the conceptual distance-based propagation and the property-based propagation for the CBC similarity computation. We decided to only test conceptual distance-based propagation since it provides better recommendation accuracy than standard distance-based propagation. To this aim we compared three similarity measures:

- Cosine similarity (COS);
- CBC similarity after conceptual distance-based propagation (CBC-C);
- CBC similarity after property-based propagation (CBC-P).

We used as input concept for the CBC metrics the most direct super-class of any item for which the rating had to be suggested.

4.1 Experiment Setup and Sample

We employed a domain ontology describing social events, focusing on gastronomic events (e.g. tastings, food fairs), cultural events (e.g. concerts, movies) and sport events (e.g. races, open days). The ontology was designed for recommending purposes and it includes 16 main classes at 3 levels. In order to collect users preferences regarding the domain items, we used a questionnaire in which 44 events had to be graded on a scale from 1 to 10. The sample was composed of 231 subjects, 19-38 years old, recruited according to an availability sampling strategy. We obtained a total of 10,164 ratings.

To test the accuracy of CBC similarity combined with different propagation techniques we generated an ordered list of rated items starting from a portion of each user's ratings and compared this list with the remaining part of the user ratings. We ran different tests varying the initial percentage of rated items (5%, 10%, 20%), to simulate different degrees of sparsity in the input data.

4.2 Measures

The difference between the estimated ratings and the real ones was computed using the mean absolute error (MAE). The correlation between the list generated by the tested

algorithm and the original user's list was estimated using Spearman's rank correlation coefficient ρ, which provides a non-parametric measure of statistical dependence between two ordinal variables and gives an estimate of the relationship between two variables using a monotonic function. In the absence of repeated data values, a perfect Spearman correlation of $\rho = +1$ or $\rho = -1$ is obtained when each of the variables is a prefect monotone function of the other. The performance of the different techniques was compared with the χ-square test.

4.3 Results

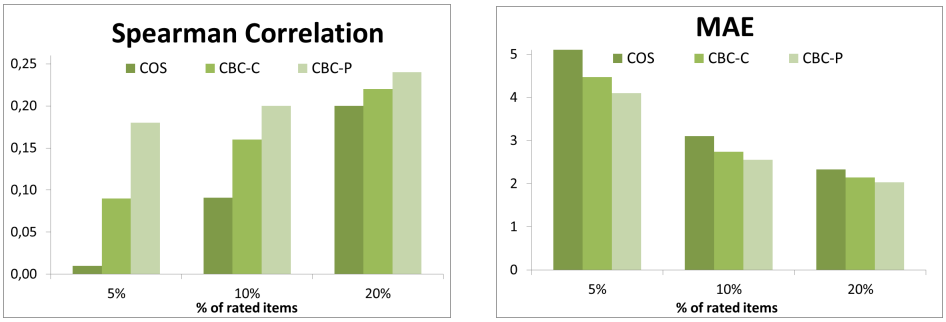

Fig. 1. Left panel: Spearman correlation between the suggested and real preference list in case of 5%, 10% or 20% rated items. Right panel: MAE of the suggested and real rating in case of 5%, 10% or 20% rated items.

The left panel in Fig. 1 shows that the lists produced with the CBC similarity have a higher degree of association ρ with the real ratings than the ones produced by using the standard cosine similarity. CBC similarity is particularly efficient in case of data sparsity, since it allows to use domain knowledge to compensate for the lack of data. Hence, as the percentage of rated items grows the difference between the two approaches becomes less prominent. In all tests the CBC-P approach was more effective than the CBC-C approach confirming the results presented in [4]. The right panel of Fig. 1 displays the mean absolute error of the suggested ratings w.r.t. the real ratings. Also in this case CBC similarity performed better then the cosine similarity for all the tests and CBC-P yielded a slightly better result than CBC-C.

To understand better the aforementioned results we studied the frequency distribution of the Spearman Coefficient for a percentage of rated items equal to 5% and 10%, respectively (Fig. 2). The standard cosine similarity was unable to produce any case with a good association ($\rho > 0.5$) in the 5% test, whereas CBC-P did so for 10% of users. On the 10% test COS obtained a good association in 2% of the cases as opposed to the 15% of CBC-P. This trend continues also for the 20% test yielding 20% for COS and 34% for CBC-P.

The difference between the cosine similarity and the CBC similarity is statistically significant according to the χ-square test for the 5% and 10% test, but not for the 20% one. More precisely, in the 5% test the χ-square between COS and CBC-C yielded $p = 6 \cdot 10^{-8}$, between COS and CBC-P it yielded $p = 2 \cdot 10^{-100}$ and between CBC-C and CBC-P it yielded $p = 6 \cdot 10^{-19}$. For the 10% it yielded $p = 5 \cdot 10^{-9}$ for COS and

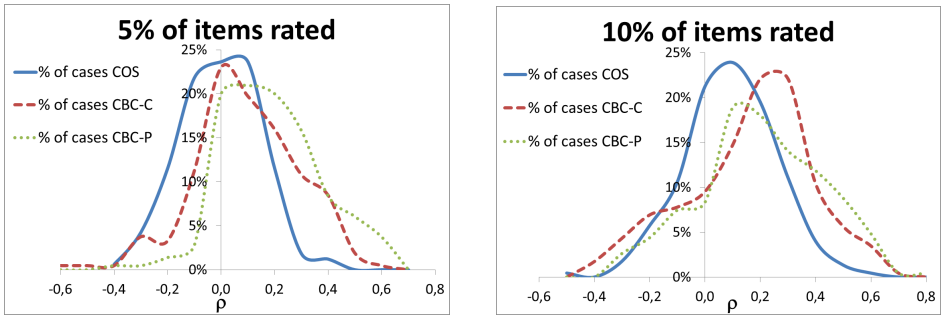

Fig. 2. Distribution of cases for various values of ρ for the 5% (left panel) and 10% (right panel) of rated items.

CBC-C, $p = 10^{-11}$ for COS and CBC-P and $p = 0.6$ (thus not significant) for CBC-C and CBC-P.

Hence, we can conclude that the CBC similarity approach works significantly better then the standard cosine similarity and can be effective in alleviating the data sparsity and the cold start problem.

5 Related work

In this section we briefly describe the works in the literature which exhibit similarities with our approach. In particular, we present works that share the cornerstones of our approach: usage of ontology-based user model, and propagation of user interests in an ontology. Then, we present other modalities of calculation of similarity among users.

The most similar works are the ones of Sieg et at. [14, 15] and Middleton et al. [8], since they all employ ontology-based user models and propagate user interests in the ontology. Sieg et al. use ontological user profiles to improve personalized Web search in [14] and in collaborative-filtering recommender in [15]. User interaction with the system is used to update the interest values in domain ontology by using spreading activation. While Sieg et al. annotate instances of a reference ontology with user interest values, we annotate also the classes, not only the instances. Middleton et al. [8] present a hybrid recommender system in which the user feedback is mapped onto the domain ontology. They propagate user interest to superclasses of specific topics using IS-A similarity. Their propagation is bottom-up and always propagates 50% of the value to the super class.

A review and comparison of different similarity measures and algorithms can be found in Cacheda et al. [2]. Our alternative similarity metric makes use of domain ontological structure to calculate similarity among users, and in particular uses the interest values of these users' corresponding user models. In a similar fashion to us, other works follow this approach.

For example, Mobasher et al. [9] calculate user similarities based on interests scores across ontology concepts, instead of their ratings on individual items like traditional collaborative filtering. A main difference w.r.t. our approach is the method they use

to calculate similarity. They first turn the ontological user profiles into flat vectors of interest scores over the space of concepts, and then compare the user profiles to figure out how distant each users profile is from all other users profiles. The distance between the target user and a neighbor user is calculated using the Euclidean distance.

In Yuan et al. [18] a structured cosine similarity is proposed for supporting text clustering by considering the structure of the documents. Similarly to us they built on the cosine similarity and flatten the values assigned to an ontology classes in a vector to feed to the cosine similarity. However, they simply pass the lower class value to the super-class and do not weigh the relations between different classes.

In Thiagarajan et al. [16] the authors represent user profiles as bags-of-words (BOW) where a weight is assigned to each term describing user interests to create extended user profiles. They then use a spreading activation technique to find and include additional terms in user profiles. The similarity measure is obtained by combining cosine similarity (for overlapping parts) with bipartite graph approaches (for remaining profile terms).

The idea that not all domain concepts should be treated the same is elaborated in [7] which shows that some of the ratings carry more discriminative information than others. They argue that less common ratings for a specific item tend to provide more discriminative information than the most common ratings. Thus, they propose to divide user similarity into two parts: local user similarity (a vector similarity among users) and global user similarity (considers the number of similar neighbors).

6 Conclusions and future work

We proposed the concept-biased cosine similarity (CBC similarity), a novel approach to measure the user similarity relative to a specific concept, which is able to alleviate data sparsity problem. Our approach can be adopted for several purposes. For example, it can be used to enhance collaborative filtering recommender in a semantic direction [15], for alleviating data sparsity and for improving of recommendation results. It can also be exploited by social applications in order to suggest new connections to users, based on shared interests.

An empirical evaluation showed that CBC similarity outperforms the cosine similarity in supporting collaborative filtering, especially in the presence of sparse data. In the tests with less then 20% of the items rated the difference between the two techniques was statistically significant.

We are working in several directions to exploit interest propagation and refine ontology-based similarity metrics. We plan on allowing bias in the similarity metric not only for preferring a specific topic, but also for other dimensions such as context and expertise. A more ambitious work would be to even allow this multiple dimensions to influence the structure of the ontology as suggested in [10]. For this reason, we are also working on an approach to compute the CBC similarity across user profiles represented as overlays over different ontologies or personal ontology views. This avenue of work may also be useful to compare user profiles from different systems, making cross-system personalization easier.

References

1. P. Ahlgren, B. Jarneving, and R. Rousseau. Requirements for a cocitation similarity measure, with special reference to Pearson's correlation coefficient. *Journal of the American Society for Information Science and Technology*, 54(6):550–560, 2003.
2. F. Cacheda, V. Carneiro, D. Fernandez, and V. Formoso. Comparison of collaborative filtering algorithms: Limitations of current techniques and proposals for scalable, high-performance recommender systems. *ACM Transactions on the Web*, pages 2–2, 2011.
3. F. Cena, S. Likavec, and F. Osborne. Propagating user interests in ontology-based user model. In *12th International Conference of the Italian Association for Artificial Intelligence, AI*IA '11*, volume 6934 of *LNCS*, pages 299–311. Springer, 2011.
4. F. Cena, S. Likavec, and F. Osborne. Property-based interest propagation in ontology-based user model. In *20th Conference on User Modeling, Adaptation, and Personalization, UMAP 2012*, volume 7379 of *LNCS*, pages 38–50. Springer, 2012.
5. E. Chiabrando, S. Likavec, I. Lombardi, C. Picardi, and D. Theseider Dupré. Semantic similarity in heterogeneous ontologies. In *22nd ACM Conference on Hypertext and Hypermedia, Hypertext '11*, pages 153–160. ACM, 2011.
6. L. Hamers, Y. Hemeryck, G. Herweyers, M. Janssen, H. Keters, R. Rousseau, and A. Vanhoutte. Similarity measures in scientometric research: The Jaccard index versus Salton's cosine formula. *Information Processing and Management*, 25(3):315–318, 1989.
7. H. Luo, C. Niu, R. Shen, and C. Ullrich. A collaborative filtering framework based on both local user similarity and global user similarity. *Machine Learning*, 72(3):231–245, 2008.
8. S. E. Middleton, N. R. Shadbolt, and D. C. De Roure. Ontological user profiling in recommender systems. *ACM Transactions on Information Systems*, 22:54–88, 2004.
9. B. Mobasher, X. Jin, and Y. Zhou. Semantically enhanced collaborative filtering on the web. In *Web Mining: From Web to Semantic Web, First European Web Mining Forum, EMWF '03*, volume 3209 of *LNCS*, pages 57–76. Springer, 2003.
10. F. Osborne and A. Ruggeri. A prismatic cognitive layout for adapting ontologies. In *21st International Conference on User Modeling, Adaptation, and Personalization, UMAP '13*, volume 7899 of *LNCS*, pages 359–362. Springer, 2013.
11. R. Rada, H. Mili, E. Bicknell, and M. Blettner. Development and application of a metric on semantic nets. *IEEE Trans. on Systems Management and Cybernetics*, 19(1):17–30, 1989.
12. G. Salton and M. McGill. *Introduction to Modern Information Retrieval.* McGraw-Hill Book Company, 1984.
13. J. B. Schafer, D. Frankowski, J. L. Herlocker, and S. Sen. Collaborative filtering recommender systems. In *The Adaptive Web, Methods and Strategies of Web Personalization*, volume 4321 of *LNCS*, pages 291–324. Springer, 2007.
14. A. Sieg, B. Mobasher, and R. Burke. Web search personalization with ontological user profiles. In *16th ACM Conference on Information and Knowledge Management*, CIKM '07, pages 525–534. ACM, 2007.
15. A. Sieg, B. Mobasher, and R. Burke. Improving the effectiveness of collaborative recommendation with ontology-based user profiles. In *1st International Workshop on Information Heterogeneity and Fusion in Recommender Systems*, HetRec '10, pages 39–46. ACM, 2010.
16. R. Thiagarajan, G. Manjunath, and M. Stumptner. Finding experts by semantic matching of user profiles. In *3rd Expert Finder Workshop on Personal Identification and Collaborations: Knowledge Mediation and Extraction, PICKME '08*, 2008.
17. A. Tversky. Features of similarity. *Psychological Review*, 84(4):327–352, 1977.
18. S.-T. Yuan and J. Sun. Ontology-based structured cosine similarity in document summarization: with applications to mobile audio-based knowledge management. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 35(5):1028–1040, 2005.