Integrating semantic relatedness in a collaborative filtering system

Felice Ferrara and Carlo Tasso

University of Udine, Via delle Scienze 206, 33100 Udine, Italy
{felice.ferrara, carlo.tasso}@uniud.it

Abstract. Collaborative Filtering (CF) recommender systems use opinions of people for filtering relevant information. The accuracy of these applications depends on the mechanisms used to filter and combine the opinions (the feedback) provided by the users. Here, we propose a mechanism aimed at using semantic relations extracted from Wikipedia in order to adaptively filter and combine the feedback of people. The semantic relatedness among the concepts/pages of Wikipedia is used to identify the opinions which are more significant for predicting a rating for an item. We show that our approach improves the accuracy of the predictions and it opens, on the other hand, opportunities for providing explanations of the computed recommendations by means of the semantic relations.

1 Introduction

Collaborative Filtering (CF) recommender systems use opinions of people for filtering relevant information. These tools face the information overload problem by simulating the word-of-mouth adopted by people who ask suggestions to friends or experts when they need to take a decision. In the area of CF systems, user-based CF mechanisms predict the relevance of a resource (referred also as target item) for a target user (referred also as *active user*) by: automatically finding people (technically named *neighbors*) who can provide the response to the given information need; combining the feedback of the neighbors for generating the prediction. For this reason, the accuracy of a user-based CF recommender system depends on the ability of the system of: identifying the set of people who share knowledge, tastes and preferences with the active user; combining the feedback of the neighbors for producing a useful prediction.

In order to identify the set of neighbors, a user-based recommender system compares the feedback provided by the active user with the feedback provided by the other users following the idea that people who showed a similar behavior in the past will probably agree also in the future. However, the social process executed by humans uses also other contextual information: according to the current information need, people ask the suggestions to a specific set of people since they can provide more authoritative opinions.

In this work we propose a mechanism aimed at getting closer to this social mechanism. In particular, we follow the the idea of predicting a rating by taking

into account the characteristics of the target item: the opinions expressed for the resources more strictly related to the target item are considered as more relevant for identifying the neighbors and for computing the final prediction. We apply this idea to the movie domain and, more specifically, we use Wikipedia for inferring the relatedness among the movies. The semantic relatedness among the movies is used in order to weight the opinions of the users: the opinions/ratings provided for the movies more related to the target item are considered as more relevant for computing the user-to-user similarity and for predicting the rating.

By integrating the semantic relatedness in the computations we also open some interesting perspectives for facing the task of supporting the user with explanations of the recommendations. The semantic features (such as the actors, the director or other meaningful characteristics) used to identify the most authoritative opinions can be presented to the active user for showing how the system works.

The paper has the following structure: Section ....

2 Related Work

In this paper we propose to weight the opinions provided by the users in order to identify an authoritative set of neighbors and consequently improve the accuracy of the prediction of the rating for an item. The idea of selecting the neighbors in an adaptive way has been proposed in the BIPO framework [**?**]. In the BIPO framework different approaches are used to identify the most predictive items for a given movie, where the predictive items are the movies more correlated to the given movie. In BIPO, only the most predictive items are used to compute the recommendations since the other movies are treated as noise. In order to reach this goal two kinds of approaches are used:

- Statistical approaches. The correlation among two items is computed by taking into account the feedback of the users: the correlation among the movies is computed by discovering latent relations among the opinions of the users.
- A genre-based approach. The relevance of a movie for predicting the rating of the target item grows up according to the number of genres shared between the target item and the specific movie.

We also followed the idea of adaptively filtering the feedback of the users in order to provide recommendations in social tagging systems. In particular, in [**?**] we proposed a user-based collaborative filtering where tags with a shared meaning where used to identify the resources in each topic of interest, the users interested in each topic of interest and to produce different list of recommendation for each interest of the active user.

On the other hand, in this work we exploit semantic similarities extracted from Wikipedia in order to weight the opinions without filtering the feedback of people since we follow the idea that by using only the opinions for the most

predictive items we increase significantly the sparsity of the matrix containing the feedback of the users.

The link structure of Wikipedia has been used also in other works to compute the semantic relatedness among concepts. In [], the authors proposed two metrics for computing the semantic distance among concepts of Wikipedia which will be referred in this paper as COUT and GDIN. The COUT metric describes each concept as a weighted vector of Wikipedia pages: given a concept of Wikipedia, the pages linked by the concept describe it and the weight of each page is equal to $log\left(\frac{|W|}{|T|}\right)$ where W is the set of pages in Wikipedia and T is the number of pages which have a link to the page. Given such representation of the concepts, the COUT metric compute the distance among two concepts as the cosine similarity between the corresponding vectors. The GDIN metric, on the other hand, slightly modifies the Google Distance measure for computing the semantic relatedness among concepts of Wikipedia. More technically, the distance between the concepts a and b, is computed as

where A is the set of pages which have a link to the page a, B is the set of pages which have a link to b and W is the set of all the pages available in Wikipedia.

3 Computing Semantic Relatedness in Wikipedia

In order to assign a weight to each opinion provided by the users we compute the semantic relatedness between the target item and the item evaluated by the specific opinion. In the BIPO framework the only semantic approach used for identifying the predictive items was based on the number of genres shared by the movies. However, by taking into account just the genre of the movies we do not consider other possible significant relations and, for this reason, we follow the idea of inferring a semantic relatedness measure from Wikipedia. In fact two movies may be related since they may share one of more actors, the director, or the subject, which are information usually uploaded in Wikipedia.

We decided to use the methods described in literature (the COUT and the GIN metrics) for inferring the semantic relatedness among the concepts of Wikipedia. However, we also implemented some variation of these metrics since the semantic relatedness metrics described in literature have never been used to compute the correlation among two movies and we can find different settings able to improve the accuracy of the recommendations. In particular, here we use two variations of the metrics reported above which will be referred as CIN and GDOUT.

The CIN metric is a variation of the COUT measure. This metric still compute the semantic relatedness between two concepts as the cosine similarity among two weighted vectors of Wikipedia pages but, in this case, the weighted set of pages used to represent a concept is defined by the articles which has a link to it. In this case the weight of a page in the vector is equal to

where T is the set of pages linked by the page.

On the other hand we also modified the GIN metric by defining the GOUT measure which is defined by taking into account the pages linked by the concept. In particular the GOUT is computed as

where

We decided to propose these other metrics since we were interested in checking if there were some ways to exploit the link structure of Wikipedia to have better results.

4 Computing the Semantic Relatedness among Concepts of Wikipedia

In this work we use the link structure of Wikipedia in order to compute the semantic relatedness among the movies. In particular in our approaches each movie is represented by taking into account the corresponding Wikipedia page and, more specifically, the pages linking this page and linked by this page. More technically, each movie is described by a vector of weighted links where the weight of a link has to

Both the proposed approaches are based on the idea of using the Normalized Google Distance (Cilibrasi and Vitanyi, 2007) for computing the similarities among two Wikipedia pages. The Normalized Google Distance was originally defined to compute the similarities among Web pages by taking into account the number of links shared among two documents. The two approaches described in this section, on the other hand, slightly modify this idea since the similarity among two Wikipedia pages is computed according to the number of links the resources share.

We inherit the first approach from [], where each Wikipedia page is described according the incoming links, i.e. the set of

We inherit the first metric from []

The first one is the method defined in [] which basically uses the Normalized Google Distance (Cilibrasi and Vitanyi, 2007), in order to compute the semantic relatedness. In particular, we will refer this metric as GoogleLinkIN measure since it takes into account the links entering in the page in order to infer the semantic similarity among the two concepts. The idea of this measure is based on the assumption that two concepts A and B are more strongly related if many other concepts link both A and B. Technically, given A and

5 Integrating semantic relatedness in user-based CF system

The semantic relatedness is used in pour approach to weight the opinions of people following the idea that some opinions are more relevant when we predict the rating for a given resource.

In fact a baseline mechanism for identifying the neighbors can use the Pearson similarity

person

where .... In this scenario all the available ratings provided by the users are used without regards of other contextual information. On this other hand, in the BIPO framework, the target item is integrated in this formula as follows:

where

In this way the BIPO framework assigns an higher relevance to the items more correlated to the target item. In fact, the items which are not related to the target item are discarded from the computation. According to the idea that such approach can increase the sparsity of the matrix containing the feedback of the users, in this work we propose a slight different approach for integrating the semantic distances without removing a significant part of the feedback. More technically, we integrated the semantic relatedness as follows:

where ...

In our approach, given a target item we use the opinions expressed for all the items in the dataset since when the weight for the opinion is equal to zero we still take it into account, where the ratings for the movies which are more related to the target item are considered as more relevant.

This choice mainly depends on the fact that we compute the similarity among the items by using semantic features. By using semantic features we are not able to model other latent characteristics such as:

- semantic relations not reported in Wikipedia. The users of Wikipedia could provide a not accurate description of the movie.
- relations which cannot be expressed in a semantic way.

From a conceptual point of view this means that we are taking into account all the feedback of the users but we have more attention to the items which appear more predictive. For example if we are going to predict the rating given by the user Bob for the animation movie 'Dumbo' than we will consider movie semantically related to it (such as other animation movies, other Disney movies, etc.) as more significant than others.

By using this metric the top N neighbors can be selected and, consequently their weighted opinions are used to predict the rating. In fact, the opinions of the neighbors are finally combined by computing the rating as follows:

formula

6 Results

In order to evaluate our idea we exploited an off-line analysis by using the MovieLens dataset which is composed by ... movies where each movie is rated by at least .. users and each user provided at least ... ratings.

We had to execute a preprocessing phase for associating each movie in the dataset to the corresponding concepts of Wikipedia. Since some movies are not described by a Wikipedia page we had to remove these movies and the corresponding opinions (.. items) from the dataset.

We computed the recommendations and we computed the accuracy of the recommendation by computing the MAE measure which ...

The results of the generated recommendations are reported in Figure.

As reader can notice all the approaches which weight the ratings improve the results obtained when the PCC is used.

7 Conclusions

In this paper we proposed a user-based CF recommender system where the opinions of people are used by taking into account the characteristics of the target item. In particular, the

References