Integrating semantic relatedness in a collaborative filtering system

Felice Ferrara, Carlo Tasso

Artificial Intelligence Laboratory, Department of Mathematics and Computer Science, University of Udine, Udine, Italy

Abstract

Collaborative Filtering (CF) recommender systems use opinions of people for filtering relevant information. The accuracy of these applications depends on the mechanism used to filter and combine the opinions (the feedback) provided by users. In this paper 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 also opens opportunities for providing explanations on the obtained recommendations.

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 mechanism 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 (i) automatically finding users (technically named *neighbors*) who can provide suggestions to a given information need and (ii) combining the feedback provided by the neighbors for generating the prediction (Koren & Bell, 2011). For this reason, the accuracy of a user-based CF recommender system depends on the capability of the system to identify the set of people who share knowledge, tastes and preferences with the active user and combining the feedback of the neighbors for generating useful predictions.

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 behaviour in the past will probably agree also in the future. However, the social process executed by humans also uses other contextual information: according to the current information need, people ask 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 idea of predicting a rating by taking into account the characteristics of the target item: the opinions expressed for the resources more related to the target item are 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. By integrating the semantic relatedness in the computation we also open some interesting perspectives for facing the task of supporting the user with explanations on the recommendations (Tintarev & Masthoff, 2011). 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 describes related work; the metrics introduced in this paper for computing the semantic relatedness among the concepts of Wikipedia are presented in Section 3; Section 4 focuses on the methods for integrating the semantic relatedness for weighting the opinions of neighbors and producing the predictions of the ratings; the evaluation of the proposed CF approach is illustrated in Section 5; a final discussion concludes the paper in Section 6.

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, to improve the accuracy of the predictions. The idea of selecting the neighbors in an adaptive way has been also proposed in the BIPO framework (Baltrunas & Ricci, 2008) where only the most *predictive items* for a given movie (i.e. the movies more correlated to the target item) are used to compute recommendations and the other movies are treated as noise. In order to identify the predictive items two kinds of approaches are used in (Baltrunas & Ricci, 2008):

- Statistical approach. The correlation between two items is computed by taking into account the feedback of the users by discovering latent relationships among the opinions of the users.
- A genre-based approach. The correlation among the movies is given by the number of shared genres.

We also followed the idea of adaptively filtering the feedback of the users in order to provide recommendations in social tagging (Ferrara & Tasso, 2011). In particular, we utilized tags for grouping resources associated to each topic of interest of the active user and then a specific set of resources was used to select the neighbors for a given topic. However we noticed that by using only a subset of the ratings provided by the active user we increase significantly the sparsity of the matrix containing the feedback of the users. For this reason, in this work, we do not discard the ratings in the computation of the recommendations but we weight more the opinions expressed for the other movies semantically related to the target item. In this way, the opinions expressed for the movies semantically related to the target item are considered as more relevant and no opinions are discarded.

Wikipedia has been used also in other work to compute the semantic relatedness among concepts. The textual contents uploaded by the users of Wikipedia for describing concepts have been used in (Gabrilovich & Markovitch, 2007) and (Strube & Ponzetto, 2006) for computing the semantic relatedness among them. The link structure of Wikipedia was utilized for computing the semantic relatedness in (Milne, 2007) by using two metrics (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 concepts, the COUT metric computes the semantic relatedness between two concepts as the cosine similarity between the two corresponding vectors. The GDIN metric, on the other hand, slightly modifies the Google Distance measure (Cilibrasi, 2007) in order to compute the semantic relatedness among concepts of Wikipedia. In this case the distance is computed as

$$GDIN(a,b) = 1 - \frac{\log\left(\max(|A|,|B|) - \log(|A \cap B|)\right)}{\log(|W|) - \log\left(\min(|A|,|B|)\right)}$$

where A is the set of pages with a link to concept a, B is the set of pages with a link to concept b, and W is the set of pages available in Wikipedia.

In this work we will also provide some variations of the COUT and the GIN metrics which are described in Section 3.

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 shared genres. However, by taking into account only the genre of the movies we do not consider other possible significant relations. For this reason, we follow the idea of inferring a semantic relatedness measure from Wikipedia. In fact, two movies may be

related since they share one of more actors, the director, the subject, and/or various other aspects of their content information usually uploaded in Wikipedia.

We decided to use the variations of the COUT and the GIN metrics described in Section 2 for inferring the semantic relatedness among the movies present in of Wikipedia. However, we also implemented some variations of these metrics in order to find settings able to improve the accuracy of the recommendations. In particular, we called are going to describe two variations of the COUT and the GDIN metrics (described above) which will be referred as CIN and GDOUT. The CIN metric is a variation of the COUT measure. This metric still computes the semantic relatedness between two concepts as the cosine similarity among two weighted vectors of Wikipedia pages. However, in our case, the weighted set of pages used to represent a concept is constituted by the Wikipedia pages W which have a link to the concept. Moreover, the weight of a page in the vector is equal to $\log\left(\frac{|W|}{|T|}\right)$ where W is the set of pages in Wikipedia and T is the number of articles linked by the page. We also modified the GIN metric by defining the GOUT measure which takes into account the pages linked by the concept. In particular the GOUT is computed as

$$GDOUT(a,b) = 1 - \frac{\log\left(\max(|A|,|B|) - \log(|A \cap B|)\right)}{\log(|W|) - \log(\min(|A|,|B|))}$$

where *A* is the set of pages linked by the concept *a*, *B* is the set of pages linked by the concept *b*, and W is still the set of pages available in Wikipedia.

We integrated the semantic relatedness metrics described in this paper in the user-based CF recommender system described in the following section.

4 Predicting the ratings by weighting the opinions

In order to compute the rating for a target item *j*, a baseline mechanism can compute the similarity among two users by means of the Pearson Correlation

$$BasePC(u,v,j) = \frac{\sum_i (r_{ui} - \overline{r_u})(r_{vi} - \overline{r_v})}{\sqrt{\sum_i (r_{ui} - \overline{r_u})^2 \sum_i (r_{vi} - \overline{r_v})^2}}$$

where: r_{ui} and r_{vi} are the ratings provided respectively by the user *u* and *v* for the item *i*, $\overline{r_u}$ and $\overline{r_v}$ are the means of the ratings returned by the user *u* and *v* and the sum runs over all the items *i* that both the users *u* and *v* rated. By using this formula all the available ratings provided by the users are used in the computation regardless of the given target item j.

On the other hand, in the BIPO framework, the correlation among the target item *j* and the other movies is integrated as follows:

$$BipoPC(u,v,j) = \frac{\sum_i w_{ij}(r_{ui} - \overline{r_u})(r_{vi} - \overline{r_v})}{\sqrt{\sum_i w_{ij}(r_{ui} - \overline{r_u})^2 \sum_i w_{ij}(r_{vi} - \overline{r_v})^2}}$$

where w_{ij} is the weight in $[0,1]$ used to compute the correlation between the item *i* and *j*. In this way the BIPO framework assigns higher relevance to the items more correlated to the target item, but the items which are not related to the target item are discarded from the computation. In our work the correlation between the item *i* and *j* is the semantic relatedness inferred from Wikipedia according to the metrics described above.

Moreover we propose a different approach for integrating the semantic distances, since we follow the idea that by discarding ratings we increase the sparsity of the matrix which contains the feedback of the users. More technically, we integrated the semantic relatedness as follows:

$$WikiPC(u,v,j) = \frac{\sum_i (\frac{1+w_{ij}}{2})(r_{ui} - \bar{r_u})(r_{vi} - \bar{r_v})}{\sqrt{\sum_i (\frac{1+w_{ij}}{2})(r_{ui} - \bar{r_u})^2 \sum_i (\frac{1+w_{ij}}{2})(r_{vi} - \bar{r_v})^2}}$$

In our approach, given a target item, we use the opinions expressed for all the items in the dataset since, also when the weight for the opinion is equal to zero, we still take into account the feedback. On the other hand, 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, so we are not able to model other characteristics such as:

- semantic relations not reported in Wikipedia. The users of Wikipedia could provide a not accurate description of the movie;
- latent 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 the ratings assigned to the related items have a stronger impact. 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 the metrics described above, we can compute the set N of the top *n* neighbors. Then, their weighted opinions can be used to predict the rating. In particular, the opinions of the neighbors are finally combined by predicting the rating r_{ui} of the user *u* for the item *i* as follows:

$$r_{ui} = \bar{r_u} + \frac{\sum_{v \in N} PC(u,v,i) \times (r_{vi} - \bar{r_v})}{\sum_{v \in N} |PC(u,v,i)|}$$

By substituting PC(u,v,j) with *BasePC*, *BipoPC* and *WikiPC* respectively, we obtain three different ways of generating predictions that we exploited in the experimental evaluation illustrated in Section 4.

5 Evaluation

In order to evaluate our approach we exploited an off-line analysis by using the MovieLens dataset (MovieLens Data Sets) which is composed by ratings provided by *943* users who rated *1682* movies (each user in the dataset rated at least *20* items).

We pre-processed the dataset for associating each movie in the dataset to the corresponding pages of Wikipedia. Since some movies are not described by a Wikipedia page we had to remove these movies (*68* items) and the corresponding opinions from the dataset. We used a 5-fold cross validation technique for evaluating the results returned when:

- the opinions are weighted according to the metrics described in Section 2 (the GDIN and COUT metrics) and the metrics that we proposed in Section 3 (the GDOUT and the CIN metrics);
- the opinions are not weighted, i.e. the *BasePC* is used;
- the opinions are weighted by taking into account only the genres as shown in [Baltrunas, 2008] (i.e. the semantic relatedness is equal to the number of genres shared between the two considered movies divided by the total number of genres).

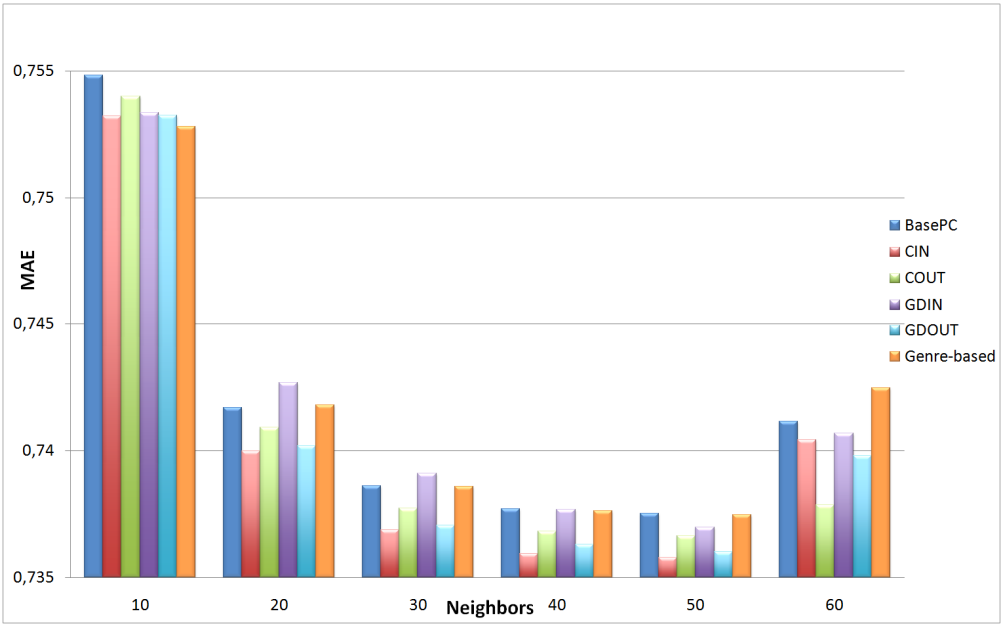

Figure 1: The accuracy of the weighting approaches

We evaluated the accuracy of the recommendations by computing the *Mean Absolute Value* (*MAE*)

$$MAE = \frac{\sum |r_{ui} - f_{ui}|}{M}$$

where r_{ui} is the rating predicted by the recommender system, f_{ui} is the rating provided by the user u for the item i and M is the number of predictions. This metric runs over the list of computed predictions and lower MAE values correspond to more accurate results. The results of the evaluation executed by integrating the metrics by means of the *WikiPC* are reported in Figure 1. The figure shows that all the approaches which weight the ratings outperform the results of the *BasePC* approach (which does not weight the opinions). In particular, the CIN measure that we introduced in this work provides the best results (*MAE=0.735*) by using the feedback of *50* neighbors. Moreover, the genre-based approach returns the worst results when it is compared to the Wikipedia based metrics. This shows that by using only the genre we lose a lot of significant information about the items. We certified the statistic significance of the results by means of the Wilcoxon test ($p < 0.05$).

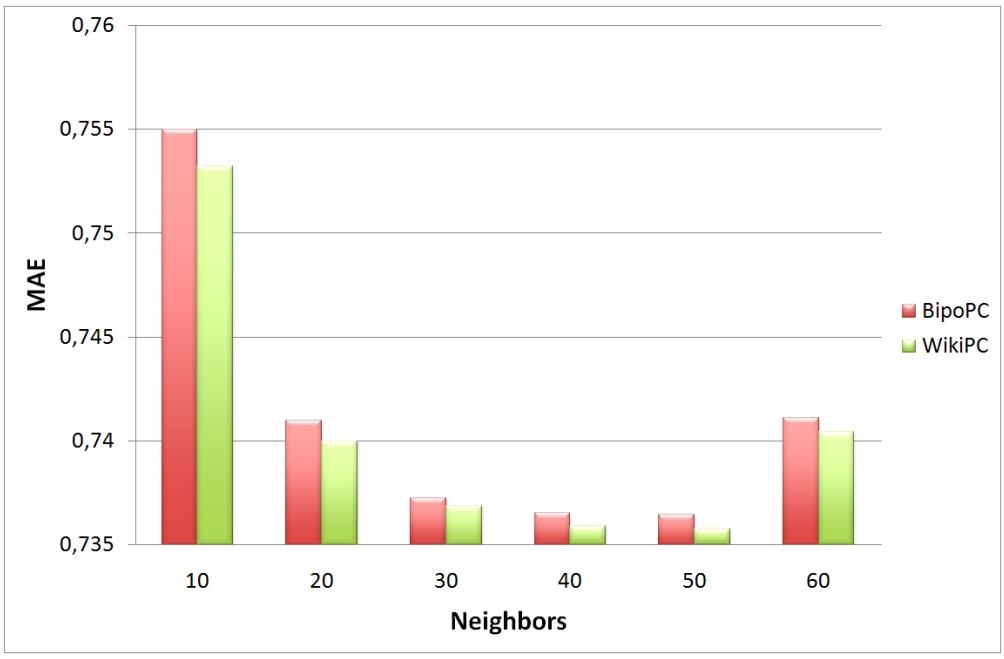

Figure 2: Comparing the BipoPC approach to the WikiPC approach

We also compared our approach for integrating the semantic relatedness with the *BipoPC* mechanism (by using the CIN measure to compute the correlation among the movies). According to the results shown in Figure 2 the *WikiPC* outperforms the *BipoPC*, but the difference between the approaches is quite limited. This is probably due to the increasing sparsity generated by the *BipoPC* approach when semantic information is used to identify the most predictive items.

6 Conclusions and Future Work

In this paper we proposed a new approach for integrating the information available in Wikipedia into a CF recommender system able to provide accurate recommendations. The improvements are actually not very meaningful but we also claim that the integration of semantic relatedness also opens interesting opportunities. The semantic relations used to adaptively identify the neighbourhood could be used to implement an explanation mechanism where the system shows the semantic features (such as the actors, the genre, or the director) used to compute the recommendations. The explanation could be proposed to the active user by a sentence like *'we recommend this movie since you liked: horror movies, Francis Ford Coppola'*. In turn the user could refine the recommendations be deleting or adding other meaningful semantic features. The semantic features could also be used for clustering into groups of semantically related movies. By clustering movies we could support the diversity in the recommendations by suggesting movies in clusters which have not been considered by the active user but are still interesting to users who shared a similar feedback. Future works will consider also the idea of extracting semantic relations from Linked Data cloud such as (Linked Movie Database).

Bibliography

Koren, Y., & Bell, R. (2011). Advances in Collaborative Filtering. In Ricci, F., & Shapira, R., & Kantor, P. (Eds.) Recommender Systems Handbook (pp. 145-186) Springer

Tintarev, N., & Masthoff, J. (2011). Designing and Evaluating Explanations for Recommender Systems. In Ricci, F., & Shapira, R., & Kantor, P. (Eds.) Recommender Systems Handbook (pp. 479-510) Springer

Baltrunas, L., & Ricci, F. (2008). Locally Adaptive Neighborhood Selection for Collaborative Filtering Recommendations. In Proceedings of UMAP 2008 Conference (pp. 22-31). Hannover, Germany

Ferrara F., & Tasso, C. (2011). Extracting and Exploiting Topics of Interests from Social Tagging Systems. In Proceedings of ICAIS 2011Conference (pp. 285-296) Klagenfurt, Austria

Strube, M., & Ponzetto, S. (2006). WikiRelate! Computing Semantic Relatedness Using Wikipedia. In Proceedings of the AAAI 2006 Conference (pp. 1419-1424). Boston, Massachusetts, USA

Gabrilovich, E., & Markovitch, S. (2007). Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis. In proceedings of IJCAI 2007 Conference (pp. 1606-1611). Hyderabad, India

Milne, D. (2007). Computing Semantic Relatedness using Wikipedia Link Structure. In proceedings of the New Zealand Computer Science Research Student Conference. Hamilton, New Zealand

Cilibrasi, R., & Vitányi, P. (2007). The Google Similarity Distance. IEEE Transaction on Knowledge Data Engeneering 19(3): 370-383

MovieLens Data Sets. Available from http://www.grouplens.org/node/73/

Linked Movie Database. Available from http://www.linkedmdb.org/