Well-argued recommendation: adaptive models based on words in recommender systems

Julien Gaillard
University of Avignon
INRIA Sophia Antipolis
Avignon, France
julien.gaillard[1,2]

Marc El-Beze
University of Avignon
Agorantic
Avignon, France
marc.elbeze[1]

Eitan Altman
INRIA Sophia Antipolis
Maestro
Sophia-Antipolis, France
eitan.altman[2]

Emmanuel Ethis
University of Avignon
Nobert Elias Center
Avignon, France
emmanuel.ethis[1]

[1] :@univ-avignon.fr
[2] :@inria.fr

Abstract

Recommendation systems (RS) take advantage of products and users information in order to propose items to consumers. Collaborative, content-based and a few hybrid RS have been developed in the past. In contrast, we propose a new domain-independent semantic RS. By providing textually well-argued recommendations, we aim to give more responsibility to the end user in his decision. The system includes a new similarity measure keeping up both the accuracy of rating predictions and coverage. We propose an innovative way to apply a fast adaptation scheme at a semantic level, providing recommendations and arguments in phase with the very recent past. We have performed several experiments on films data, providing textually well-argued recommendations.

1 Introduction

Recommender systems aim at suggesting appropriate items to users from a large catalog of products. Those systems are individually adapted by using a specific profile for each user and item, derived from the analysis of past ratings. The last decade has shown a historical change in the way we consume products. People are getting used to receive recommendations. Nevertheless, after a few bad recommendations, users will not be convinced anymore by the RS. Moreover, if these suggestions come without explanations, why people should trust it? Numbers and figures cannot talk to people.

To answer these key issues, we have designed a new semantic recommender sytem (SRS) including at least two innovative features:

- Argumentation: each recommendation relies on and comes along with a textual argumentation, providing the reasons that led to that recommendation.
- Fast adaptation: the system is updated in a continuous way, as each new review is posted.

In doing so, the system will be perceived as less intrusive thanks to well-chosen words and its failures will be smoothed over. It is therefore necessary to design a new generation of RS providing textually well-argued recommendations. This way, the end user will have more elements to make a well-informed choice. Moreover, the system parameters have to be dynamically and continuously updated, in order to provide recommendations and arguments in phase with the very recent past. To do so, we have adapted the algorithms we described in Gaillard (Gaillard et al., 2013), by including a semantic level, i.e words, terms and phrases as they are naturally expressed in reviews.

This paper is structured as follows. In the next section, we present the state of the art in recommendation systems and introduce some of the improvements we have made. Then, we present our approach and define the associated methods in section 3. We describe the evaluation protocol and how we have performed some experiments in section 4. Finally we report results including a comparison to a baseline in section 5.

2 Related work and choice of a baseline

We present here some methods used in the literature. Collaborative Filtering (CF) systems use logs

of users, generally user ratings on items (Burke, 2007; Sarwar et al., 1998). In these systems, the following assumption is made: if user a and user b rate n items similarly, they will rate other items in the same way (Deshpande and Karypis., 2004). This technique has many well-known issues such as the "cold start" problem, i.e when new items or users appear, it is impossible to make a recommendation, due to the absence of rating data (Schein et al., 2002). Other limitations of RS are sparsity, scalability, overspecialization and domain-dependency problems.
In Content Based Filtering (CBF) systems, users are supposed to be independent (Mehta et al., 2008). Hence for a given user, recommendations rely only on items he previously rated.

Some RS incorporate semantic knowledge to improve quality. Generally, they apply a concept-based approach to enhance the user modeling stage and employ standard vocabularies and ontology resources. For instance, ePaper (scientific-paper recommender), computes the matching between the concepts constituting user interests and the concepts describing an item by using hierarchical relationships of domain concepts (Maidel et al., 2008). Codina and Ceccaroni (2010) propose to take advantage of semantics by using an interest-prediction method based on user ratings and browsing events.
However, none of them are actually based on the user opinion as it is expressed in natural language.

2.1 Similarity measures

Similarity measures are the keystone of RS (Herlocker et al., 2005). Resnick (1997) was one of the first to introduce the Pearson correlation coefficient to derive a similarity measure between two entities. Other similarity measures such as Jaccard and Cosine have been proposed (Meyer, 2012). Let S_u be the set of items rated by u, T_i the set of users who have rated item i, $r_{u,i}$ the rating of user u on item i and $\overline{r_x}$ the mean of x (user or item). PEA(i,j) stands for the Pearson similarity between items i and j and is computed as follows:

$$\frac{\sum_{u \in T_i \cap T_j}(r_{u,i} - \overline{r_i})(r_{u,j} - \overline{r_j})}{\sqrt{\sum_{u \in T_i \cap T_j}(r_{u,i} - \overline{r_i})^2 \sum_{u \in T_i \cap T_j}(r_{u,j} - \overline{r_j})^2}} \quad (1)$$

In the remainder, the Pearson similarity measure will be used as a baseline. The Manhattan Weighted and Corrected similarity (MWC), that we introduced in (Gaillard et al., 2013), will be used as a point of comparison as well[1]. Again, for none of them, textual content is taken into account.

2.2 Rating prediction

Let i be a given item and u a given user. We suppose the pair (u, i) is unique. Indeed, most of social networks do not allow multiple ratings by the same user for one item. In this framework, two rating prediction methods have to be defined: one user oriented and the other item oriented. *Sim* stands for some similarity function in the following formula.

$$rating(u,i) = \frac{\sum_{v \in T_i} Sim(u,v) \times r_{v,i}}{\sum_{v \in T_i} |Sim(u,v)|} \quad (2)$$

A symmetrical formula for items $rating(i,u)$ is derived from and combined with (2).

$$\hat{r}_{u,i} = \beta \times rating(u,i) + (1-\beta) \times rating(i,u) \quad (3)$$

3 Methods

In this section, we describe the methods we have used and propose some of the enhancements we have elaborated in our system. In formula (2), Sim can be replaced by several similarity such as Pearson, Cosine or MWC similarity (Tan et al., 2005). All these methods provide a measurement of the likeness between two objects. We then conclude if two users (or items) are "alike" or not. One has to define what "alike" should mean in this case. If two users rate the same movies with equals ratings, then these similarities will be maximal. However, they may have rated identically but for completely different reasons, making them not alike at all. Moreover, none of these similarity measures can express why two users or items are similar. This is due to the fact that they rely on ratings only.

3.1 New similarity based on words

We propose a new similarity method, taking into account words used by users in their past reviews about items. In the remainder, we call it the *Word Based Similarity* (WBS). Each user x (or item) has a vocabulary set V_x and each word w in it is associated

[1] Details on MWC can be found in supplementary material.

with a set of ratings $\mathbf{R}_{w,x}$ and an average usage rating $\overline{r_w}$. In order to balance the contribution of each word, we define a weight function F_w, mixing the well-known Inverse Document Frequency $IDF(w)$ with the variance σ_w^2. Common words and words w associated with very heterogenous ratings $\mathbf{R}_{w,x}$ (i.e a high variance) will have a smaller weight in the similarity. N_w is the number of items in which the word w appears. N_{tot} is the total number of items. D is the maximum difference between two ratings. Note that F_w has to be updated at each iteration.

$$F_w = -log\left(\frac{N_w}{N_{tot}}\right) \times \frac{1}{\sigma_w^2} \tag{4}$$

$$WBS(x,y) = \frac{\sum_{w \in V_x \cap V_y}(D - |\overline{r_{w,x}} - \overline{r_{w,y}}|)F_w}{D \times |V_x \cap V_y| \sum_{w \in V_x \cap V_y} F_w} \tag{5}$$

3.2 Adaptation

An adaptive framework proposed in (Gaillard et al., 2013) allows the system to have a dynamic adaptation along time, overcoming most of the drawbacks due to the cold-start. The authors have designed a dynamic process following the principle that every update (u, i) needs to be instantly taken into account by the system. Consequently, we have to update the σ_w^2 and IDF(w) at each iteration, for every word. Paying attention to avoid a whole re-estimation of these two variables, we derived an iterative relation for the two of them[2]. We thus reduced the complexity by one degree, keeping our system very well-fitted to dynamic adaptation.

3.3 Textual recommendation

The main innovative feature of our proposal is to predict what a user is going to write on an item we recommend. More precisely, we can tell the user why he is expected to like or dislike the recommended item. This is possible thanks to the new similarity measure we have introduced (WBS). Let us consider a user u and an item i. To keep it simple, the system takes into account what u has written on other items in the past and what other users have written on item i, by using WBS. The idea consists in extracting what elements of i have been liked or disliked by other users, and what u generally likes. At the intersection of these two pieces of information, we extract a set of matching words that we sort by relevance using F_w. Then, by taking into account the ratings associated with each word, we define two sub-sets P_w and N_w. P_w contains what user u is probably going to like in i and N_w what u may dislike. Finally, we provide the most relevant arguments contained in both P_w and N_w, and each of them is given in the context they have been used for item i. As an example, some outputs are shown in section 5.2.

[2]More details can be found in the supplementary material.

4 Evaluation criteria

We present here the evaluation protocol we designed. It should be noted that we are not able to make online experiments. Therefore, we can not measure the feedback on our recommendations. However, the cornerstone of recommender system is the accuracy of rating predictions (Herlocker et al., 2004). From this point of view, one could argue that the quality of a recommender engine could be assessed by its capacity to predict ratings. It is thus possible to evaluate our system comparing the prediction $\hat{r}_{u,i}$ for a given pair (u, i), with the actual real rating $r_{u,i}$.

The classical metrics[3] (Bell et al., 2007) *Root Mean Square Error* (RMSE) and *Mean Absolute Error* (MAE) will be used to evaluate our RS.

Last but not least, we make the following assumption: if WBS results are as good as MWC's, the words presented by the system to users as arguments are likely to be relevant.

5 Experiments

This work has been carried out in partnership with the website Vodkaster [4], a Cinema social network. Researchers have used other datasets such as the famous Netflix. Unfortunately, the latter does not include textual reviews. It is therefore strictly impossible to experiment a SRS on such a dataset.

5.1 Corpus

The corpus has been extracted from Vodkaster's database. Users post *micro-reviews* (MR) to express their opinion on a movie and rate it, within a

[3]Details on metrics are given in the supplementary material.

[4]www.vodkaster.com

140 characters Twitter-like length limit. We divided the corpus into three parts, chronologically sorted: training (Tr), development (D) and test (T). Note that in our experiments, the date is taken into account since we also work on dynamic adaptation.

	Tr	D	Tr+D	T
Size	55486	9892	65378	9729
Nb of Films	8414	3184	9130	3877
Nb of Users	1627	675	1855	706

Table 1: Statistics on the corpus

5.2 Results

Figure 1 compares four different methods: the classical Pearson (PEA) method that does not allow quick adaptation, the MWC method with and without quick adaptation MNA and ours (WBS). Within

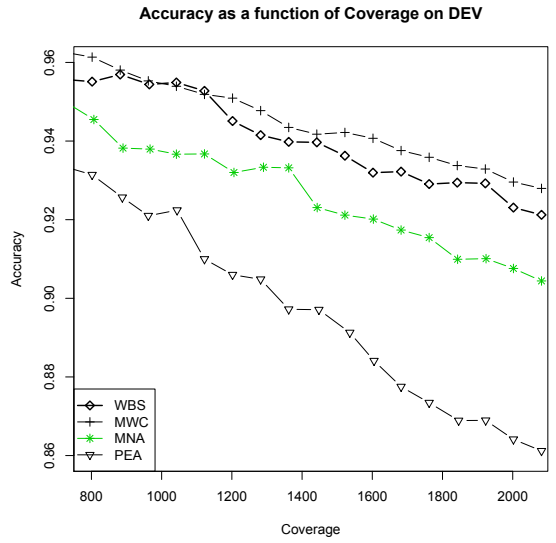

Figure 1: Evolution of accuracy as a function of coverage for PEA, MWC and WBS methods on D corpus.

the confidence interval, in terms of accuracy, the same perfomances are obtained by MWC and WBS. Both outperform[5] PEA and MNA. Our word based approach is thus able to offer the arguments feature without any loss of perfomances with respect to any others RS methods that we know of.

In Table 2, we set a constant coverage (2000 predictions) in order to be able to compare results obtained with different methods.

Corp.	Met.	RMSE	MAE	%Acc.	CI
D	PEA	0.99	0.76	86.41	1.49
E	MNA	0.93	0.72	90.75	1.26
V	MWC	0.89	0.69	92.95	1.12
	WBS	0.89	0.70	92.45	1.16
T	PEA	1.01	0.78	86.02	1.51
E	MNA	0.98	0.75	90.04	1.30
S	MWC	0.92	0.71	91.46	1.22
T	WBS	0.94	0.72	91.15	1.24

Table 2: Results with Pearson (PEA), MWC, MWC without Adaptation (MNA), WBS. CI is the radius confidence interval estimated in % on accuracy (Acc.).

MNA (MWC without adaptation) being better and more easily updated than Pearson (PEA), we have decided to use the adaptive framework only for MWC. Moreover, for Pearson dynamic adaptation, the updating algorithm complexity is increased by one degree.
We want to point out that the results are the same for both MWC and WBS methods, within a confidence interval (CI) radius of 1.16%. From a qualitative point of view, these results can be seen as an assessment of our approach based on words.

Example of outputs: The movie *Apocalypse Now* is recommended to user Theo6 with a rating prediction equal to 4.3. Why he might like: *some brillant moments* (0.99), *among the major masterpiece* (0.91), *Vietnam's hell* (0.8); dislike: *did not understand everything but...* (0.71).

The data we have does not contain the information on the reaction of the user to the recommendation. In particular, we do not know if the textual argumentation would have been sufficient for convincing Theo6 to see the film. But we know that after seeing it, he put a good rating (4.5/5) on this movie.

[5]Note that the key point here is the comparison of results obtained with the baseline and with the method we propose. Both of them have been evaluated with the same protocol: RMSE is computed with respect to rating predictions above some empirical threshold as done in (Gaillard et al., 2013).

6 Conclusion and perspectives

We have presented an innovative proposal for designing a domain-independent SRS relying on a word based similarity function (WBS), providing textually well-argued recommendations to users. Moreover, this system has been developed in a dynamic and adaptive framework. This might be the first step really made towards an anthromorphic and evolutive recommender. As future work, we plan to evaluate how the quality is impacted by the time dimension (adaptation delay, cache reset,etc.).

Acknowledgment

The authors would like to thank Vodkaster for providing the data.
This work has been partly supported by the European Commission within the framework of the CONGAS Project (FP7- ICT-2011-8-317672), see www.congas-project.eu.
We also would like to thank Agorantic for their support (http://blogs.univ-avignon.fr/sfr-agorantic).

References

R. Bell, Y. Koren and C. Volinsky. 2007. *The BellKor 2008 Solution to the Netflix Prize.* The Netflix Prize.

R. Burke. 2007. *Hybrid Web Recommender Systems.* The Adaptive Web, 377–408.

V. Codina and Luigi Ceccaroni. 2010. *Taking Advantage of Semantics in Recommendation Systems.* Proceedings of the 13th International Conference of the Catalan Association for A.I.,163–172

M. Deshpande and G. Karypis. 2004. *Item based top-N recommendation algorithms.* ACM Transactions on Information and System Security.

J. Gaillard, M. El-Beze, E. Altman and E. Ethis. 2013. *Flash reactivity: adaptive models in recommender systems.* International Conference on Data Mining (DMIN), WORLDCOMP.

J. Herlocker, J.A Konstan, L. Terveen and J. Riedl. 2004. *Evaluating collaborative filtering recommender systems.* ACM Transactions on Information Systems (TOIS).

V. Maidel, P. Shoval, B. Shapira, M. Taieb-Maimon. 2008. *Evaluation of an ontology-content based filtering method for a personalized newspaper.* RecSys'08: Proceedings, 91–98.

B. Mehta, T. Hofmann, and W. Nejdl. 2008. *Robust collaborative filtering.* In RecSys

F. Meyer. 2012. *Recommender systems in industrial contexts.* PhD thesis, University of Grenoble, France.

P. Resnick and R. Varian Hal. 1997. *Recommender systems (introduction to special section.)* Communications of the ACM

B.M Sarwar, J.A Konstan, A. Borchers,J. Herlocker, B. Miller, J. Riedl 1998. *Using filtering agents to improve prediction quality in the groupLens research collaborative filtering system.* Proceedings of the ACM Conference on Computer Supported Cooperative Work

A.I Schein, A. Popescul and L.H Ungar. 2002. *Methods and metrics for cold-start recommendations.* ACM SIGIR Conference on Research and Development in Information Retrieval.

P. Tan, M. Steinbach and V. Kumar. 2005 *Introduction to Data Mining.* Addison-Wesley, 500–524.

C. Ziegler, S.M McNee, J.A Konstan and G. Lausen. 2005. *Improving recommendation lists through topic diversification.* Fourteenth International World Wide Web Conference