UDC 004.9

Y. Stekh, V. Artsibasov
Lviv Polytechnic National University,
CAD Departament

SOME METHODS IN SOFTWARE DEVELOPMENT RECOMMENDATION SYSTEMS

This article analyzes the current state of the models and methods of building recommendation systems. The basic classes of problems that solve the recommendation system are highlighted. The features of the method collaborative filtering are shown. Developed a method for calculating the similarity coefficients, taking into account the sparseness of ratings vectors of goods and people.

Key words: recommender system, data mining, collaborative filtering, coefficients of similarity, user profiles.

Проаналізовано сучасний стан моделей і методів побудови рекомендаційних систем. Виділено основні класи задач, які розв'язують рекомендаційні системи. Показано особливості застосування методу спільної фільтрації. Розроблено метод розрахунку коефіцієнтів подібності, який враховує розрідженість векторів рейтингів товарів і користувачів.

Ключові слова: рекомендаційні системи, інтелектуальний аналіз даних, спільна фільтрація, коефіцієнти подібності, профілі користувачів.

Introduction

Recommendation systems – are systems that operate with a particular type of information filtering system, it is recommended information elements that may be of interest the user. Typical recommendations system receives user input as data aggregates, and sends them to the intended recipients in the form of recommendations. This technology allows users to spend a minimum of time to find the right information on the Internet. Recommendation system compares the data collected from users and create a list of items that are recommended to the user. They are an alternative search algorithm as help users quickly find articles and information that they would not find themselves. Recommendation systems are used mainly to supply the customer in real-time products (films, books, clothing) and services that are likely to be interested in it. Especially, recommendation systems are used in e-commerce. The use of recommendation systems covered recently on a stationary retail trade, information centers, search software, scientific articles, etc. This application is characterized by the provision of advice to users automatically, on the basis of already committed actions (purchases, exposed ratings, visits, etc.) and taking feedback from them (order in shops, referring, etc.). Web recommendation systems (recommendation systems on web pages) are usually implemented on Web servers and use the data obtained from the collection of the revised Web template (explicit data) and user registration information (explicit data). The most famous of recommendation systems include the following : Amazon.com, Inc. - an American company, the largest in the world by turnover among Internet companies that sell products and services online and one of the first online services focused on sales of real goods of mass demand; eBay Inc. - an American company that provides services in the areas of online auctions (main field of activity), online shopping, instant payments, manages the website eBay.com and its local versions in several countries, the company owns PayPal and Ebay Enterprise; MovieLens - recommendation system and virtual community website that recommends movies to its users , recommendations are provided with regard profiles (ratings) of users and use collaborative filtering algorithm; Rozetka.ua ™ - by far the most popular online store electronics and home

appliances in Ukraine, representatives of the company are available in all regions of Ukraine. Recommendation system is one of the most important sections of data mining.

Methods and tools for building recommendation systems

Recommendation system as a separate line began to develop in the last twenty years. So make a classification of methods and tools for building recommendation systems is difficult. We can distinguish the following approaches to building recommendation systems:

- model-based;
- data-based.

In an approach based on models first formed a descriptive model of user preferences, commodities and the relationship between them, and then formed recommendations on the basis of the resulting model. The advantage of this approach is to have a model that gives more insight generated recommendations and relationships in data availability, and the fact that the formation of recommendations is divided into two stages: learning resource model in deferred mode and a fairly simple calculation based on the recommendations of the existing model in real time. However, these models do not support incremental learning (the emergence of new data requires the conversion of the whole model) and mostly show lower prediction accuracy than based on data.

In data-based approach the recommendations are calculated on some similarity degree in all of the accumulated data. These data are a set of vectors of user rating and a set of vectors of item rating. This approach is simpler and showed high accuracy in practice and has the advantage of taking into account new data incremental (new users and new products are added to a database and taken into account when forming forecasts along with available). However, this approach is difficult to calculate in terms of time and memory resources. Also, this approach can not provide a descriptive analysis of existing laws, to give more understanding of the available data and explain the forecast. In modern recommendation systems used in such powerful companies like Amazon.com, Yahoo.com, Google.com, eBay Inc. mainly used the approach based on the data.

In the approach based on the data are the following methods:

- methods that focus on the use of vectors of ratings users (user-centric);
- methods that focus on the use of vectors rankings items (item-centric);
- hybrid methods;
- multicriteria methods.

General block diagram of data-based approach shows in Fig. 1.

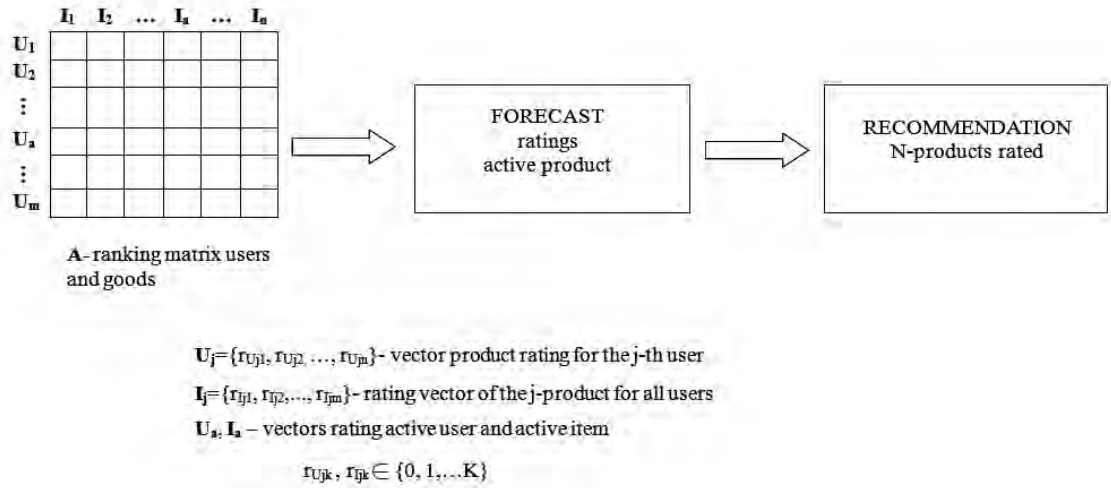

Fig. 1. Overall block diagram of the recommendations search in the data-based approach

Prediction rating in the collaborative filtering techniques

The basic method used in data-based approach is the method of collaborative filtering. The user or item for which is forecasting unknown rating, called the active user or active item, respectively. The task of collaborative filtering can be formulated as follows. Let **U** be a set of *n* users, **I** – a set of *m* items, **R** – a set of $m \times n$ ratings $r_{u,i}$ user $u \in \mathbf{U}$ and product $i \in \mathbf{I}$, $\mathbf{S_u} \in \mathbf{I}$ – a set of products that have already been chosen by the user *u*. The purpose of collaborative filtering is to predict the rating $p_{a,i}$ active user for the item *i*. User *a* is called an active user, if he chose certain items $\mathbf{S_a} \neq \varnothing$. This product, for which is forecast, is not known in advance $i \notin \mathbf{S_a}$. Denoted by $\mathbf{S_v}$ a set of products that the user *v* has selected, $\mathbf{S_u}$ – a set of items that the user *u* has selected. Then $\mathbf{S_{uv}}$ – a set of items that users *u* and *v* have chosen.

$$\mathbf{S}_{uv} = \{i \in \mathbf{S} \,|\, r_{u,i} \neq 0 \wedge r_{v,i} \neq 0\}; \tag{1}$$

$$\mathbf{S}_{uv} = \mathbf{S}_u \cap \mathbf{S}_v; \tag{2}$$

$$m = |\mathbf{S}_{uv}|. \tag{3}$$

Let – $\overline{r_u}, \overline{r_v}$ average rating of the item users *u* and *v*, respectively.

We denote by $\mathbf{T_a}$ a set of users who have jointly selected products with the active user.

The rating forecast to approach focuses on the use of vectors of user ratings is by the following formula

$$r_{a,i} = \overline{r_a} + \frac{\sum_{l \in T_a} (r_{l,i} - \overline{r_l}) w_{a,l}}{\sum_{l \in T_a} |w_{a,l}|}. \tag{4}$$

The rating forecast to approach focuses on the use of vectors of user ratings is by the following formula

$$r_{a,i} = \frac{\sum_{n \in N} r_{u,n} \times w_{i,n}}{\sum_{n \in N} |w_{i,n}|}. \tag{5}$$

The summation is over all selected products $n \in N$ for a user *u*, $w_{i,n}$ - the similarity between the items *i* and *n*.

Accuracy rating forecast is heavily dependent on the accuracy of the calculation of similarity coefficients $w_{i,j}$. Advantageously, similarity coefficient is calculated as follows cosine of the angle between vectors (6) or Pearson correlation formula (7):

$$w_{u,v} = \frac{\sum_{i \in S_{uv}} r_{u,i} \times r_{v,i}}{\sqrt{\sum_{i \in S_{uv}} r_{u,i}^{\,2} \sum_{i \in S_{uv}} r_{v,i}^{\,2}}}; \tag{6}$$

$$w_{u,v} = \frac{\sum_{i \in S_{uv}} (r_{u,i} - \overline{r_u})(r_{v,i} - \overline{r_v})}{\sqrt{\sum_{i \in S_{uv}} (r_{u,i} - \overline{r_u})^2 \sum_{i \in S_{uv}} (r_{v,i} - \overline{r_v})^2}}. \tag{7}$$

Particularity of item rating vectors and user rating vectors users is the fact that they have a large number of zero elements. Each user does not select all items and each item is not selected by all users. Advantageously, the percentage of non-zero elements in these vectors does not exceed 10%. The classic formula for calculating the similarity coefficients do not include this feature and therefore give a significant error in the calculation.

Let R_{max} highest possible rating in the rating scale catalog, R_{min} - the lowest possible rating. Let d(a,b) Euclidean distance between vectors, $d_{max}(a,b)$ – the maximum Euclidean distance to a given set of vectors

$$d_{\max}(a,b) \propto \sqrt{(R_{\max} - R_{\min})^2}. \tag{8}$$

Normalized Euclidean distance between the vectors

$$\sigma_n(a,b) = \frac{d(a,b)}{d_{\max}(a,b)}; \tag{9}$$

$$\sigma_n(a,b) \in (0,1]. \tag{10}$$

Calculated values of the coefficient of similarity for the problem of predicting the rating will take the value converted to normalized Euclidean distance

$$\frac{1}{\sigma_n(a,b)} = \frac{\sqrt{m(R_{\max} - R_{\min})^2}}{\sqrt{\sum_{i \in S_{uv}} (r_{u,i} - r_{v,i})^2}}, \tag{11}$$

where $m = |\mathbf{S}_{uv}|$

The introduction of the coefficient *m* allows to take into account the sparseness of ratings vectors.

Introduction to the calculation of the similarity coefficient Jakard further improves the accuracy of the calculation

$$k_j = \frac{|\mathbf{S}_u \cap \mathbf{S}_v|}{|\mathbf{S}_u \cup \mathbf{S}_v|}. \tag{12}$$

The final form of the expression for the calculation of similarity coefficients following

$$w(u,v) = k_j \times \frac{\sqrt{m(R_{\max} - R_{\min})^2}}{\sqrt{\sum_{i \in S_{uv}} (r_{u,i} - r_{v,i})^2}}, \text{ if } \exists i \in S_{uv}, r_{u,i} \neq r_{v,i}; \tag{13}$$

$$w(u,v) = k_j \times \frac{\sqrt{m(R_{\max} - R_{\min})^2}}{0{,}9 + \sqrt{\sum_{i \in S_{uv}} (r_{u,i} - r_{v,i})^2}}, \text{ if } \forall i \in S_{uv}, r_{u,i} = r_{v,i}. \tag{14}$$

The proposed approach to the calculation of the coefficients of similarity in the problems of collaborative filtering allows you to take into account the considerable sparsity of these vectors and significantly improve the predicted values for the ratings.

Conclusion

This article analyzes the current state of the models and methods of construction of recommendation systems. Highlights the major classes of the problems that solve the recommendation system. Show the features the method of collaborative filtering. Developed a method for calculating the similarity coefficients taking into account the sparseness of ratings vectors of items and users.

1. Agarwal R. C., Aggarwa l C. C., Prasad V. V. V., A Tree Pro jection Algorithm For Generation of Frequent Itemsets // J. Parall. and Distrib. Comput., vol. 61, pp. 350–371, 2001. 2. Aggarwal C. C., Wolf J. L., W u K., Y u P. S., H orting Hatc hes an Egg : A New Graph-Theor etic Ap proach t o Coll aborative Filtering. in Proc. 5th ACMSIGKDD Int. Conf. on Knowledge Discovery and D ata Mining, pp. 201–212, 1999. 3. Agrawa l R., R. Srikant R., Mining S equential Pa tterns, in Pro c. 1 1th In t. Con f. on Dat a Engineering, pp. 3–14, 19 95. 4. Agrawal R., Imie linski T., Swami A., Mi ning Association Rul es be tween Sets of Items in LargeDatabases. in Proc. of the ACM SIGMOD Conf. on Management of Data, pp. 20 7–216, 1993. 5. Al dous D. J. , Reorgan izing Large Web Sites .// Amer. Ma th. Monthly, vo l.108, pp. 16–27, 2001. 6. Anto niou G., v an Harmelen F., A Semantic Web Primer: MIT Pr ess, 2 edi tion, 2008. 7. Baeza-Yates R. A., Rib eiro-Neto B.. Mo dern Information Retrieval.: Add ison-Wesley Longman Pub lishing Co., Inc., 1999. 8. Ba labanovi´c M., Shoham Y., Fab : Content-Based, Col laborative Recommen dation / / Commun.ACM, vol.40 pp.66–72, 1997 9. Ba ldi P., Fr asconi P., Smyth P., Mod eling the Int ernet and the Web: Probabil stic M ethods and Algo rithms: W iley, 2003. 10. Baner jee A., Ghosh J., Clic kstream

Clustering using Weighted Longest Common Subsequences. in Proc. of the Web Mining Workshop at the 1st SIAM Conf. on Data Mining, pp. 33–40, 2001. 11. Berners-Lee T., Hendler J., Lassila O.. The Semantic Web // Scientific American vol. 284 pp.34–43, 2001. 12. Borges J., Levene M., Data Mining of User Navigation Patterns, in Proc. Int. Workshop WEBKDD99 – Web Usage Analysis and User Profiling, pp.31–36, 1999. 13. Burke R. Hybrid Recommender Systems: Survey and Experiments // User Modeling and User-Adapted Interaction, vol.12 pp.331–370, 2002. 14. Cadez D., Heckerman D., Meek C., Smyth P., White S. Model-Based Clustering and Visualization of Navigation Patterns on a Web Site // Data Min. Knowl. Discov., vol.7 pp.399–424, 2003. 15. Chakrabarti S. Data Mining for Hypertext: A Tutorial Survey // ACM SIGKDD Explor. Newsl., vol.1 pp. 1–11, 2000. 16. Cooley R., Mobasher B., J. Srivastava J. Data Preparation for Mining World Wide Web Browsing Patterns // Knowl. and Information Syst., vol.1 pp.5–32, 1999. 17. Cosley D., Lawrence S., Pennock D.M., REFEREE: An Open Framework for Practical Testing of Recommender Systems using Researchindex, in Proc. 28th Int. Conf. on Very Large Data Bases, pp.35–46, 2002. 18. Demir G.N., Uyar S., S., Gündüz-Ögüdücü S.. Multiobjective Evolutionary Clustering of Web User Sessions: A Case Study in Web Page Recommendation // Soft Comput., vol.14 pp.579–597, 2010. 19. Dempster A.P., Laird N.M., Rubin D.B., Maximum Likelihood from Incomplete Data via the EM Algorithm // J. Royal Statistical Society, Series B, vol.39 pp.1–38, 1977. 20. Deshpande M., Karypis G., Item-Based Top-N Recommendation Algorithms // ACM Trans. Information Syst., vol. 22 pp.143–177, 2004. 21. Driga A, Lu P., Schaeffer J., Szafron D., Charter K., Parsons I., FastLSA: A Fast, Linear-Space, Parallel and Sequential Algorithm for Sequence Alignment // Algorithmica, vol.45 pp.337–375, 2006. 22. Adomavicius G., Tuzhilin A., Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions / / IEEE Trans. Knowledge and Data Engineering, vol. 17, pp.734–749, Jun, 2005. 23. Christinsen I. A., Schiaffino S. Entertainment recommender systems for group of users // Expert Systems with Applications, vol. 38, pp.14127–14135, 2011. 24. Tang X., Zeng Q. Keyword clustering for user interest profiling refinement with paper recommender system // Journal of Systems and Softwre, vol. 2, pp.87–101, 2011. 25. Saegusa T. An FPGA implementation of real-time K-means clustering for color images / / Real Time Image Processing, vol. 2, pp. 309–318, 2007. 26. Stekh Y., Lobur M., Faisal M.E. Sardieh, Dombrova M., Artsibasov V. Research and development of methods and algorithms non-hierarchical clustering," in Proc. of the XIth International Conference CADSM, Lviv-Polyana, 2011, pp. 205–207. 27. Lobur M., Stekh Y., Kernytskyy A., Faisal M.E. Sardieh Some trends in knowledge discovery and data mining in Proc. of the IVth International Conference MEMSTECH, Lviv-Polyana, 2008, pp. 205–207.