HAL
archives-ouvertes.fr

Movie Recommendation Using OLAP and Multidimensional Data Model

Worapot Jakkhupan, Supasit Kajkamhaeng

▶ **To cite this version:**

Worapot Jakkhupan, Supasit Kajkamhaeng. Movie Recommendation Using OLAP and Multidimensional Data Model. Khalid Saeed; Václav Snášel. 13th IFIP International Conference on Computer Information Systems and Industrial Management (CISIM), Nov 2014, Ho Chi Minh City, Vietnam. Springer, Lecture Notes in Computer Science, LNCS-8838, pp.209-218, 2014, Computer Information Systems and Industrial Management. <10.1007/978-3-662-45237-0_21>. <hal-01405583>

HAL Id: hal-01405583
https://hal.inria.fr/hal-01405583

Submitted on 30 Nov 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution| 4.0 International License

Movie Recommendation using OLAP and Multidimensional Data Model

Worapot Jakkhupan and Supasit Kajkamhaeng

Information and Communication Technology Programme, Faculty of Science,
Prince of Songkla University, Hat Yai Campus, Songkla 90112, Thailand
{worapot.j,supasit.k}@psu.ac.th

Abstract. This research proposes an adoption of data warehousing concepts to create a movie recommender system. The data warehouse is generated using ETL process in a desired star schema. The profiles of users and movies are created using multidimensional data model. The data are analyzed using OLAP, and the reports are generated using data mining and analysis tools. The recommended movies are selected using multi-criteria candidate selection. The movies which present the genres that match individual preference are recommended to the particular user. The multidimensional data model and OLAP provide high performance to discover the new knowledge in the big data.

Keywords: Movie recommender system, Data warehousing, MDX language, OLAP, Multidimensional data model

1 Introduction

In the information overload era, it is very difficult for users to search for the interested information from the large amount of data. To facilitate users eliminating the useless data, recommender system (RS) has been developed. RS is a system that predicts the preferences of user from their previous interests, and offers them the pleasurable items that might meet user's satisfaction [1]. There are many successful examples of adopting the RS in the websites, for example, Amazon[1], Last.fm[2], and Movielens[3].

Nowadays, RS has been increasingly implemented in diverse areas [1, 2]. RS provides two important advantages. First, RS helps user to deal with big data presented in the internet by eliminating the information that user may not interest, and gives users the information that meets their interests. Second, RS helps business to gain more profits by increase an opportunity to offer their customers the items that related to the customers' preferences. RS uses various techniques, which suitable for the diverse purposes [3]. The well-known techniques are content-based RS [4], collaborative filtering RS [5], and hybrid RS [6].

1 http://www.amazon.com
2 http://www.last.fm
3 http://www.movielens.org

adfa, p. 1, 2011.
© Springer-Verlag Berlin Heidelberg 2011

In the real world, movie RS should recommend users the new releases and top hits movies. Moreover, user prefers to see the movies based on personal taste, especially on the movie genres, thus, movie RS is a content-based [7]. Movie RS has to know which genres of movie that users prefer to see, called user profile, and the genre of each movie, called movie profile. If movie RS has sufficient user profile, it is possible to offer the movies that meet user preferences. Anyway, user behavior on rating the movies is unpredictable. Some users always rate the movies when they like, call positive rating, on the other hand, some users always rate the movies when they don't like, called negative rating. If user always obviously give positive rating, RS can recommend the items that similar with the previous items. On the other hand, if user always give negative rating, RS should select the opposite items with the previous items. The traditional content-based RS, if there is inadequate information, RS cannot find enough movies that similar to the target movie, on the other hand, if there are large amount of similar movies, RS requires high computational performance. Another solution for the content-based profiling is to use the data warehouse and multidimensional data model, which has been applied not only in the movie RS [8] but also in the various types of RS such as website [9, 10] and book [11].

This study proposes the adoption of data warehouse concepts to create the movie RS. The propose method was experimented on HetRec2011 MovieLens dataset[4]. The dataset is prepared using ETL to create data warehouse. OLAP aggregation is used to create user and movie profiles in the multidimensional data models. The recommended movies are selected using multi-criteria candidate selection. The details of the movies are gathered from IMDb movie database via data sharing API[5]. Finally, the result is shown to user using the web based mobile browser.

The rest of this paper is organized as follows. Sect. 2 introduces the overview of the recommender system. Sect. 3 describes the proposed method. Sect. 4 reveals the experimental results. Finally, discussion and conclusion are described in Sect. 5.

2 Background of Recommender System

2.1 Recommender System

Recommender system (RS) is the system that predict the preferences of user in which user would give to the items user had not yet considered. RS uses various techniques which suitable for the various types of data and purposes [3]. RS recommends the items to user using model built from the characteristics of an item, called content-based RS [4], or the social relationship between users, called collaborative filtering RS [5], or mix both techniques, called hybrid RS [6].

RS predicts user's preferences following three steps; extract user preferences from the data source, compute recommendation using appropriate techniques, and present the recommendation candidates to users. In the content-based and collaborative RS, system requires the similar matrix among users and items. The basic requirement of RS is shown in Fig. 1.

---

4 http://www.grouplens.org/system/files/HetRec2011-movielens-2k.zip

5 http://www.omdbapi.com/

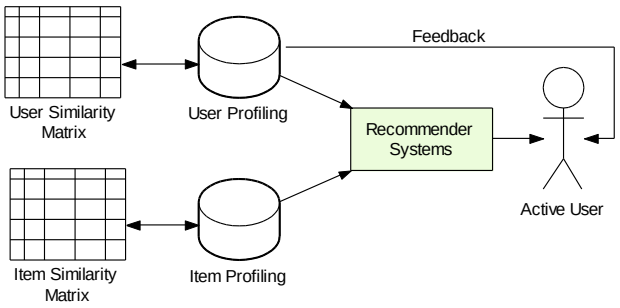


Fig. 1. Basic requirements of a recommender system

2.2 Content-based Recommender System

The content-based RS predicts rating that user may give to the undesired items. There are two steps of content-based recommendation: finding the similar items using similarity function, after that, calculating the rating using rating prediction function.

Similarity Function. The similarity function is necessary in the collaborative filtering and content-based recommender system. There are two well-known and widely used similarity functions; Pearson Coefficient Correlation (PCC), and cosine similarity. The PCC is widely used to calculate the similarity between users in a collaborative filtering approach, and the cosine similarity is widely used to calculate the similarity between items in a content-based filtering approach. The cosine similarity equation is shown in Eq. 1.

$$\text{sim}(a,b) = \frac{\sum_{i=1}^{n} w_{a,i}, w_{b,i}}{\sqrt{\sum_{i=1}^{n} w_{a,i}^2}\sqrt{\sum_{i=1}^{n} w_{b,i}^2}} \tag{1}$$

where $\text{sim}(a,b)$ is the cosine similarity between item *a* and item *b*, $w_{a,i}$ is the weight of attribute *a* on item *i*, and $w_{b,i}$ is the weight of attribute *b* on item *i*.

Rating Prediction. Content-based recommender system recommends users the items that similar or related to their preferences in the past. This approach itself requires data of individual user, and the attribute of the item. Using this approach, there is a chance for new items to get recommended, and there is no population bias. The content-based recommender system estimates user's rating using Eq. 2.

$$PR_{u,i} = \frac{\sum_{j=1}^{n} \left(\text{sim}_{i,s_j} \times r_{u,s_j}\right)}{\sum_{j=1}^{n} \text{sim}_{u,s_j}} \tag{2}$$

where $PR_{u,i}$ denotes the predicted rating of user *u* on item *i*, $\text{sim}_{i,sj}$ is the weight of attribute similarity between active item *i* and the selected item *s*, $r_{u,sj}$ is the rating of user

u on selected item *s*, $sim_{u,sj}$ is the rating of user *u* on the selected item *s*, and *n* is the number of selected items that similar with item *i*.

2.3 Data Warehouse and Multidimensional Recommender System

Data warehouse provides extreme performance to manage and analyze big data in terms of explicitly finding the new knowledge from the large amount of massive data, including RS [12]. The OLAP and multidimensional data model are used to implement the RS in many applications such as web sites [11], movies [8], and books [10]. Multidimensional data model allows user to view the data in multi aspects. Moreover, data warehouse can be used to handle the cold-start problem [7].

3 Experimental Methodology

3.1 Architecture of the proposed system

This research applies data warehouse into a movie recommender system. The architecture of the movie recommender system proposed in this research is drawn in Fig. 2. Firstly, the raw HetRec2011 MovieLens dataset are transformed to the desired data warehouse using ETL process. The data in data warehouse are analyzed using OLAP features, and the multidimensional data are created. The candidate movies are selected using multi-criteria selection method, and subsequently are presented to user.

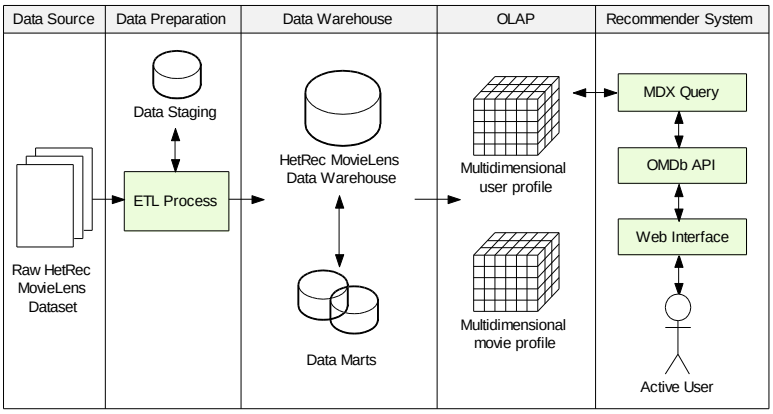


Fig. 2. Architecture of the proposed system

3.2 Data Preparation using Extract-Transform-Load (ETL)

The HetRec2011 MovieLens dataset has collected the ratings of users on movies. In the HetRec2011-movielens-2k.zip, three data files were selected to generate the multidimensional data model. The movie_genres.dat, assigns 20,809 movies with 20 genres. The user_ratedmovies.dat provides 855,598 ratings of 2,113 users on 10,197

movies, which means 8,684 movies have never been rated. The allowed rating scores are 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, and 5, respectively. The movie.dat file gives the details of each movie, including imdbID, which is used to gather the details of movies from OMDb API. To integrate the data from these files and to generate the desired multidimensional data cube, this research used Pentaho Data Integration (PDI)[6] to operate the Extract-Transfer-Load (ETL) process. The multidimensional data cube is represented in ROLAP star schema stored in MySQL database as shown in Fig. 3.

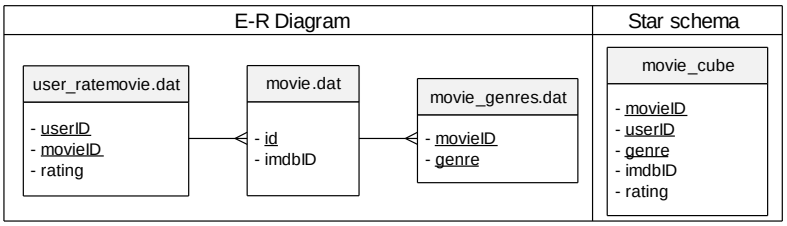
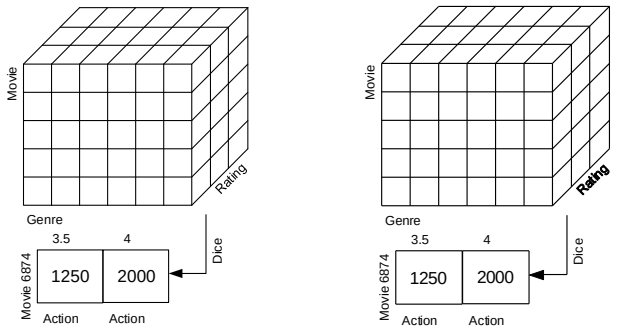


Fig. 3. ETL process using Pentaho Data Integration (PDI)

3.3 Multidimensional Data Analysis using OLAP

After generating the data warehouse using ETL process, the data in data warehouse are transformed into the desired multidimensional data cube. There are two data cube generated from OLAP; user profile and movie profile. The facts (or measurements) are acquired by count aggregation feature in OLAP. As shown in Fig. 4, the user profile consists of three dimensions: user, genre, and rating. Likewise, the movie profile consists of 3 dimensions: movie, genre, and rating. The advantage of multidimensional data is the data can be represented in multi aspect using slice or dice method. For example, R_{user}: {75, Action, 3.5} = 8 means user 75 has rated action genre with score 3.5 for 8 times. R_{movie}: {6874, {Action, Crime, Thriller}, 5} = 1,765 means movie 6874, contains 3 genres; action, crime and thriller, has been rated with score 5 for 1,750 times.

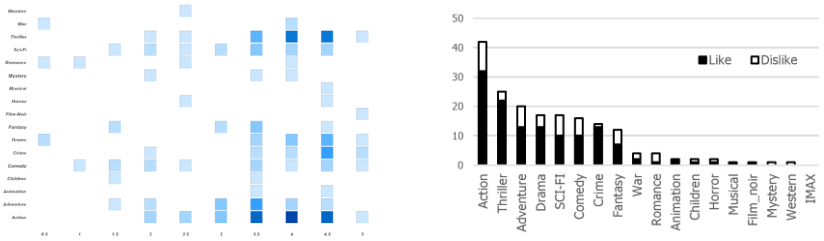


(a) Profile of user 75 (b) Profile of movie 1820

Fig. 4. Multidimensional data cube generated from movie data warehouse

[6] http://www.pentaho.com/explore/pentaho-data-integration

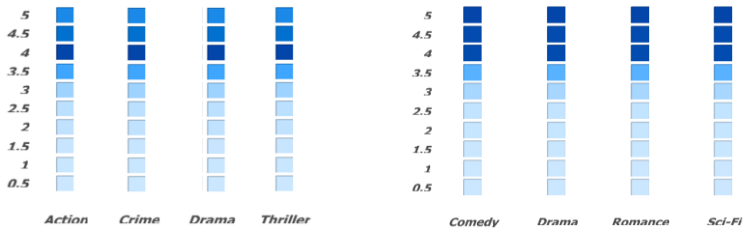
Multidimensional User Profile. The average rating is used as a baseline score to define the behavior of users. For HetRec2011 dataset, the mean rating score is 2.5, therefore, if the average rating of user is higher than 2.5, it can infer that this user has positive rating behavior. For example, the average rating of user 75 is 3.46, therefore, user 75 prefers to give positive rating on the movies. To define the genres that user like or dislike, the frequency and the ratio between ratings higher and lower than average are compared. For example, user 75 has rated 56 movies, which contains action genre 41 times, thriller 26 times, and adventure 20 times, respectively. For the action genres, there are 30 times of rating higher than 3.46 and 11 times lower than 3.46, thus, the ratio is 30:11. On the other hand, user 75 has rated romance genre 4 times, 3 times lower than 3.46, 1 time higher than 3.46, thus, the ratio is 3:1. That means, obviously, user 75 likes to see action movies but doesn't like to see romance movies. The profile of user 75 is analyzed and represented using heat grid graph and bar chart as shown in Fig. 5.



(a) Distribution of rating (positive behavior) (b) Frequency and ratio of like and dislike

Fig. 5. Profile of user 75 using multidimensional data cube slice and dice

Multidimensional Movie Profile. There are three information used to create the movie profile; genres, rating score, and frequent of rating. Movie genres are explicitly defined in movie_genres.dat file. The global rating score and the frequent of rating are acquired from ratings.dat file. The multidimensional movie profile is created and is diced into a heat grid report as an example shown in Fig. 6. According to the profile of user 75, movie 6874 should be recommended to user 75 since it has high score on action and thriller, and movie 7361 should not be recommended.



(a) Profile of movie 6874 (b) Profile of movie 7361

Fig. 6. Movies profiling using OLAP dice method

3.4 Multi-criteria Candidates selection

To select the candidates for recommendation, the user profile and movie profile are considered. Firstly, RS selects the movies which user has never rated before. The selected movies are filtered by matching with the genres that match user preferred. Subsequently, the top-k movies are ordered by the frequent of global users view and rating. Finally, the candidate movies are ordered by newest release and are recommended to user.

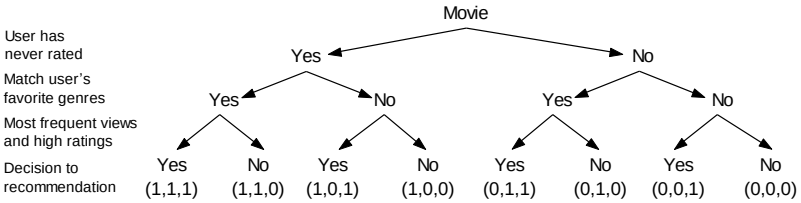
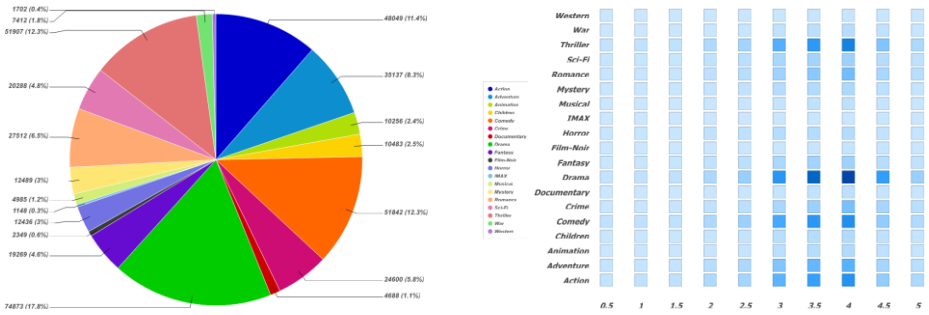


Fig. 7. Multi-criteria candidate selection for the movie recommendation

4 Experimental Results

4.1 Social Recommendation using frequent of rating and distribution of score

The frequent rating is calculated using count aggregation feature of OLAP presented in pie graph, which reflects the interest of all users on each genre, but not reveals the positive or negative opinion. If the genre has high frequent of rating, it means users like to see the movies which present that genre. Likewise, the distribution of score reveals the trend of users given to each genre presented in heat grid graph. The high frequent and high score rating should be recommended to every user, especially new or anonymous user. Data warehouse and OLAP dice the large amount of data and reveal the report in multi-aspect as shown in Fig. 8.



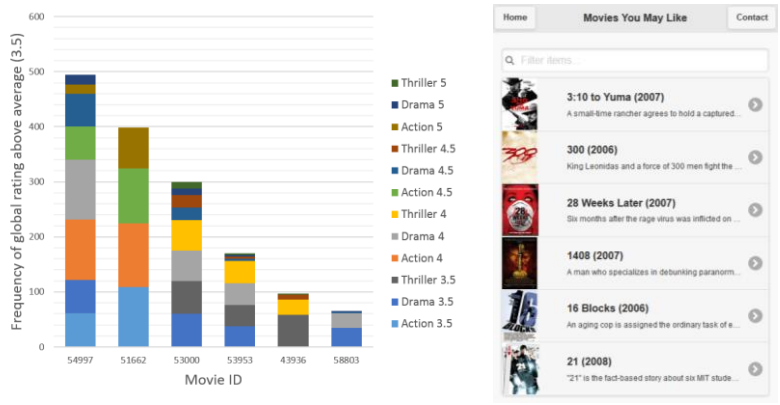
(a) Frequent of rating each genre (b) Distribution of rating score grouped by genres

Fig. 8. Global high frequent rating and high score rating generated from OLAP

The top-5 most frequent rated genres are drama, comedy, thriller, action, and adventure, respectively. This information reveals the concentrate of the genres users like to give rating. Likewise, the top-4 highest rating genres are drama, action, thriller, and comedy, respectively. This information reveals the genres of the movies that users prefer to see. Thus, According to the OLAP data analysis results, the movies that contain drama, action, thriller, and comedy should be recommended first for the new or anonymous user because those genres are most hits and most frequent view. The data aggregation feature provided by OLAP helps data analyst to generate the information from big data, which can be drilled down to see the details, or rolled up to see the summarization.

4.2 Personalized Recommendation

Addition from the global recommendation, the personalized preferences on the movie genres as stated in the multidimensional user profile are considered. The system generates the profile of every user, as an example, we have selected user 75 as a case study because user 75 has high frequent on movie rating, which adequate for the data analysis using OLAP. For example, according to Fig. 6a, user 75 has given high frequent and high rating on 3 genres; action, thriller, and drama. Thus, there are 7 sets (calculated from 2^n-1) of genres should be recommend to user 75 are M_{set} = [{action}, {action, thriller}, {action, drama}, {action, thriller, drama}, {thriller}, {thriller, drama}, {drama}]. According to the multidimensional movie profile, Fig. 9a reveals the list of movies that user 75 has never rated which have high score on action, thriller and drama extracted from movie profile using OLAP analyzer tool. The candidate movies are selected using multi-criteria and are presented to user 75 via web interface. The information of the movies are collected from IMDb using JSON API. The example screenshot of the movie recommendation prototype on mobile for user 75 is shown in Fig. 9b.



(a) Most frequent and hits extracted from OLAP (b) Prototype of RS

Fig. 9. The movie recommendation for user 75 using OLAP analyzer

5 Discussion and Conclusion

In the content-based recommender system using cosine similarity and rating prediction, there are many factors affected the accuracy of the rating prediction. For example, since the average rating of user is high due to user's positive rating behavior, it has a high opportunity that the predicted rating might lower than average if only one genre that user does not like appears in the movie. Moreover, the calculation might not reliable if there is a lack of similarity items, and the calculation requires high computational performance if the data contain large amount of similar items. Therefore, this research adopts the data warehouse concept as an alternative technique to develop the movie recommender system. The proposed concept was experimented on the HetRec2011 MovieLens dataset. The raw data are transformed into star schema using ETL process, and subsequently the movie data warehouse is created. The multidimensional user and movie profiles are generated using OLAP aggregation tool. The data cube is slice and dice in order to generate the desired reports represented in the appropriate report types.

The multidimensional user profile reveals variety aspects of user, such as user rating behavior, and the distribution of like or dislike genres. The multidimensional movie profile with OLAP aggregation tool reveals the distribution of global rating, like or dislike from all users, and the frequent of rating on each genre. The most top hits and high frequent rating movies should be recommended to all users, especially new or anonymous user who the system has no previous record. Moreover, the personalized user preferences on movie genres are matched with the genres of each movie to select the movies using multi-criteria candidate selection method. Finally, after generating the movie recommendation list, the details of the recommended movies are acquired from IMDb web service API using JSON. The prototype of movie recommender system represent the results in web interface. The multidimensional data model and OLAP data analysis provide high performance to analyze the big amount of data, and to discover the new knowledge, which can be applied in the movie recommender system.

References

1. Bobadilla, J., Ortega, F., Hernando, A., Gutiérrez, A.: (2013) Recommender systems survey. Knowledge-Based Systems. 46, 109-132 (2013)
2. Park, D.H., Kim, H.K., Choi, I.Y., Kim, J.K.: A literature review and classification of recommender systems research. Expert. Syst. Appl. 39, 10059-10072 (2012)
3. Adomavicius, G., Tuzhilin, A.: Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. IEEE T. Knowl. Data En. 17(6), 734-749 (2005)
4. Meteren, R., Someren, M.: Using content-based filtering for recommendation. In: Machine Learning in the New Information Age: MLnet/ECML 2000 Workshop, pp. 47-56. (2000)
5. Candillier, L., Meyer, F., Boullé, M.: Comparing state-of-the-art collaborative filtering systems. In: Perner, P. (ed.) Machine Learning and Data Mining in Pattern Recognition. LNCS, vol. 4571, pp. 548-562. Springer, Heidelberg (2007)
6. Antonopoulus, N., Salter, J.: Cinema screen recommender agent: combining collaborative and content-based filtering. IEEE Intell. Syst. 21(1), 35-41 (2006)

7. Elsa, N., Franck, R., Oliver, T., Ronan, T.: Cold-Start Recommender System Problem Within a Multidimensional Data Warehouse. In: IEEE Seventh International Conference on Research Challenges in Information Science, pp. 1-8, IEEE press, New York (2013)
8. Miller, B.N., Albert, I., Lam, S.K., Konstan, J.A., Riedl, J.: MovieLens unplugged: experiences with an occasionally connected recommender system. In: 8th international conference on Intelligent user interfaces, pp. 263-266. ACM, New York (2003)
9. Gediminas, A., Alexander, T.: Extending Recommender Systems: A Multidimensional Approach. In: IJCAI Workshop on Intelligent Techniques for Web Personalization, pp. 1-5. Seattle, WA, USA (2001)
10. Tiwari, R.G., Husain, M., Gupta, B., Agrawal, A.: Amalgamating Contextual Information into Recommender System. In: 3rd International Conference on Emerging Trends in Engineering and Technology, pp. 15-20. IEEE Press, New York (2010)
11. Thor, A., Rahm, E.: AWESOME – A Data Warehouse-based System for Adaptive Website Recommendations. In: 30th International Conference on Very large data bases, pp. 384-395. VBLC Endowment (2004)
12. Krohn-Grimberghe, A., Nanopoulos, A., Schmidt-Thieme, L.: A Novel Multidimensional Framework for Evaluating Recommender Systems. In: ACM RecSys Workshop on User-Centric Evaluation of Recommender Systems and Their Interfaces, pp. 34-41. CEUR-WS.org (2010)