

Evaluation and user acceptance issues of a Bayesian classifier based TV Recommendation System

Benedikt Engelbert
University of Applied Sciences
Osnabrück
Artilleriestr. 46
49076 Osnabrück
+49 541 969 3262

Karsten Morisse
University of Applied Sciences
Osnabrück
Barbarastr. 16
49076 Osnabrück
+49 541 969 3615

Kai-Christoph Hamborg
University of Osnabrück
Seminarstr. 20
49069 Osnabrück
+49 541 969 4703

b.engelbert@hs-osnabrueck.de k.morisse@hs-osnabrueck.de k.hamborg@uni-osnabrueck.de

ABSTRACT

Nowadays there is a variety of TV channels and programs. This seems to be an advantage for the TV user, but in most cases the user is overwhelmed and not able to choose the most appropriate content though. Assistive systems are needed to support the user in selecting the most appropriate content regarding the user's interests. The research group Next Generation PVR faced the task to develop a user supporting Personal Video Recorder (PVR) in the form of a Bayesian classifier based recommendation system. The work on the prototype of the system is almost done. This paper focuses on the evaluation of the given system. We are presenting two types of evaluation scenarios as well as an approach for measuring user acceptance of a TV recommendation system. Within the evaluation, the acceptance will be questioned. In addition, the results of both scenarios and of the user acceptance survey are presented and discussed.

Categories and Subject Descriptors

I.2.m [Artificial Intelligence]: Miscellaneous

General Terms

Algorithms, Experimentation, Human Factors

Keywords

Recommendation System, Television, Evaluation, Bayesian classifier, User Acceptance

1. INTRODUCTION

The way of consuming media clearly changed in recent years. Especially in times of high bandwidth and loads of appropriate media delivering services, the Internet plays an important role in cases of consuming audio/video content. However, live television is still the most popular media. In Germany 97% of all households possess a TV set with an average use of 220 minutes a day [1]. Therefore it is evident that the television market is still interesting for broadcasters. The satellite operator ASTRA holds up to 1700 TV channels just for the region of Germany. Regardless of the encrypted and shopping program, 53 is a realistic number of receivable TV channels [2]. For the TV user it is quite difficult to handle the enormous offer of content. In most cases extensive TV guides list just a limited number of TV channels and often only popular ones. The user will invest time to get an overview of all the available content. Due to too much effort most of the users focus on favored or popular TV channels and the most interesting

content regarding the user's interests remains unnoticed. Because of this, a user supporting Personal Video Recorder based on a Bayesian classifier has been presented in [3] to generate personalized TV recommendations and to counteract the problems in the given TV landscape from a user perspective. For the sake of missing evaluation results, it was not possible to give a statement about the quality of recommendations. In this paper we present two evaluation scenarios to measure the quality of the developed Bayesian classifier based TV recommendation system. The quality within a recommendation system is obviously one of the most important facts to be determined. Nevertheless in a user supporting system the question about user acceptance is just as important. For this reason this paper describes also an approach of measuring user acceptance for recommendation systems and the associated results for the given system. It will be shown, that the quality of recommendations and the user acceptance are strongly related. To present our work, the paper is organized as follows: section 2 gives an overview of related work on recommendation system with the focus on multimedia and TV systems. Within section 3, a short overview about the evaluated system is presented. We just explain the parts of the system, which are necessary for the work of the evaluation. Section 4 is divided into a part where both of the evaluation scenarios are explained and where our approach of measuring user acceptance is presented. The results of the evaluation are described and discussed in section 5. The paper ends with a conclusion and future work.

2. RELATED WORK

Recommendation systems (RS) exist in several application areas and for different types of content. One of the most common examples is Amazon.com. Within the online shop, Amazon recommends items like Books or CDs on the basis of already purchased items. For this, Amazon is using an item-to-item based approach, where relations between items get calculated [4]. Also in the area of multimedia applications RS are common. YouTube for instance is a popular online video broadcaster with million queries every day. The YouTube RS generates recommendations for related video content by analyzing user activities on the portal website [5]. With the intention to counteract the enormous offer of video content on YouTube, the work in [6] presents a mobile RS based on an extended *Bayesian classifier*. Also in other fields of RS the Bayesian classifier has been proofed as an efficient and proper working method. In 1996 Billsus and Pazzani presented their work on classifying web sites using the Bayes classifier [7]. The main goal of the work was to classify web sites automatically into the classes *like* or *dislike*. This is done by extracting information from the given HTML tags within the source code. The accuracy measurement of the classification process gained up

to 81% of properly classified web sites. The work points out, that an increasing database does not yield in an increasing accuracy of the classification process. A similar effect for a Bayesian classifier approach has been reported in [8]. The work describes a system for classifying Arabic text documents. For an evaluation a data set with increasing number of words has been created. The results have shown that the accuracy of correct classified documents is the highest in a number of 800 words with 74%. With an increasing number of words the accuracy continuously decreases. This information needs to be carefully respected in our evaluation framework described in section 4. Furthermore, in the context of television, recommendation systems are also available. Already in 1998 [9] discusses fundamental ideas and methods for RS in the *TV application area*. Das and Herman characterize the use of two different user profiles, which are influenced by the implicit and explicit behavior of a TV user. Those user profiles are also considered in our work. A more concrete scenario is described in the work of Gutta et. al. [10]. The paper describes an intelligent TV guide, which generates personalized TV recommendations. Those recommendations are generated on an adaptive user profile, where search requests for TV content of a user are captured. For the recommendation process a Bayesian classifier is also used. In [11] a personalized Electronic Program Guide (EPG) is described. Through user interactions with a Set-Top-Box, a user profile can be derived to generate personalized recommendation. Users are assigned to a user group on which basis the recommendations depend. An evaluation showed that the accuracy for this approach runs against 70% at maximum. A popular example in the TV area is the Set-Top-Box System TiVo [12]. The TiVo approach is similar to the Amazon item-to-item recommendation process described in [4], where the TiVo user base rates TV content. With the use of those ratings, the system tries to find related and similar TV items. Problematic is the state in the beginning period of the system. An item-to-item approach needs a certain number of ratings before the system is able to generate accurate recommendations. So the TiVo system is using a context aware Bayesian classifier to counteract those circumstances. The work in [12] speaks of accurate, but internal evaluation results. That's why it isn't possible to name any results at this point.

3. SYSTEM OVERVIEW

The following section describes the system architecture of the developed system. In section 2 several approaches have been presented, where some ideas were also considered for the following system architecture. In Figure 1 the components for the recommendation process within the system are shown. The main functionality of the Bayesian classifier is to classify given content objects into the classes *like* or *dislike*. So the classifier is based on a two-class decision model, where the conditional probability that a certain content fits into one of the classes like or dislike gets calculated. Within the calculation the classifier compares object attributes and counts how often they occur in one of the pre defined classes. For a detailed description of the classifier calculation we refer to our early work [3]. The object attributes are derived from a given EPG data set the system uses. More specifically the system uses generic *DomainObjects*, which are described for each application by metadata. In the given application, the metadata is derived from the EPG data with the following attributes: *TV Channel*, *Title*, *Subtitle*, *Category* (e.g. *Movie*, *show*), *Genre* (e.g. *Comedy*), *Actors*, *Description*, *Year*. With the generic data model it is possible to use the classifier also in other application areas.

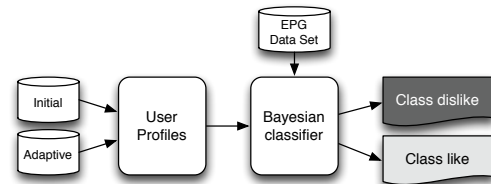


Figure 1. System Architecture

For the classification process a database is needed on which the calculation depends. The system provides for this purpose *User Profiles*. There are two types of user profiles: *initial* and *adaptive* user profile. The initial profile is important in the beginning of the system usage. In section 2 the cold start problem has already been discussed. The user can fill the initial profile with TV content he or she likes to get recommendation also in the beginning period of the system. For this purpose, the system provides a web-based TV guide. The initial profile is just created once, whereas the adaptive profiles gets build up continuously during the system use. So it is planned that the adaptive profile grows over the time and the influence as a database for the recommendation process increases. This means that, conversely, the influence of the initial profile decreases. There is the assumption that the higher the adaptive profile the better the generated recommendations. It remains an open question how the adaptive profile is created. We are dividing between implicit and explicit feedback, which is based on the work of Das and Herman in [9]. With an explicit feedback the user consciously expresses whether he or she likes a certain TV content or not (e.g. rating a TV content). The implicit feedback bases on the viewing behavior of the user. For example, if a user watched a whole TV content the system assumes, that the user likes the content. Otherwise the system would register the content as disliked. In this context the work of e.g. Hu et. al. in [19] should be mentioned. The paper deals with the profound analysis of implicit feedback, however, it describes the use of a much more complex model. The improvement of our model could be future work. It is important to say, that the stored *DomainObjects* within the user profiles are assigned to a class like or dislike. This is necessary for the calculation mentioned at the beginning of this section.

4. EVALUATION

In section 3 the main components with associated functionality of the system architecture has been described. We just focused on those parts of the system, which are important for the evaluation process in the following section. First we explain two evaluation scenarios. The first one is an online scenario to reach a high number of potential volunteers. The second one is a more realistic scenario, where the volunteers write down their TV habits in a TV diary. In section 4.2 we describe an approach to measure user acceptance for the given system. The approach is based on the Technology Acceptance Model (TAM), which characterizes the relationship between information systems and user acceptance.

4.1 Scenarios

4.1.1 Online Scenario

An important question at the beginning of the scenario design is at what point of the recommendation process the quality can be measured. As pointed out before, the system uses a Bayesian classifier, which classifies TV content in one of the classes like or dislike. In the long term, the main database for generating recommendations is the adaptive user profile, which is gathered by the feedback of the user during the use of the system. The idea

for the online scenario is to simulate the creation of an adaptive user profile on which basis the user gets personalized recommendations. We consider at least five steps where the user fills the adaptive profile and the system generates recommendations. After every step the user needs to rate each of the given recommendations. Every step represents one week of a user's watching behavior, which means that the system generates recommendations from one week of EPG data. The rating of the recommendations is more a verification of the classification the system did. The system presents 20 TV items where ten of these items belong to the class like and ten to the class dislike. The items are displayed to the user without class affiliation. The user needs to perform a classification on his/her own. In the back of the process the system compares both classifications and stores the classification of the user within the adaptive profile for the next step. At the end of the scenario there is the sixth final classification step, where the system calculates the quality of the classification. To measure the quality, we are using the F1-score, which can be interpreted as the weighted average of precision and recall values. Precision and recall is a common metric for assessing the accuracy of classification systems. To meet the requirements of the evaluation, the relationship between temporal development of the adaptive user profile and the quality of the system classification needs to be respected. For this reason, every evaluation user needs to do overall six steps, so that all users need the same length for the scenario and have the same requirements. Internally, every volunteer gets randomly another number of steps. Thus, the completion of the adaptive profile varies between one and five weeks, so that just the data for that random count of steps gets stored. At the end of the evaluation we can compare the calculated quality between the varying lengths of the scenario. At the beginning of the scenario an initial profile by the user with ten objects will be created.

4.1.2 TV diary

The measurement of quality is similar to the scenario described in section 4.1.1. Significant is the development of the adaptive user profile. Where the Online Scenario was more a simulation of data creation, whereas the TV diary deals with real watched TV content. The participants watch television as usual. They need to write down all the watched TV content in a TV diary with the name of a series, date and time so that the system can find the watched content within the given EPG data. For the participants it is also mandatory to write down TV content they began to watch, but switched the channel in case of dissatisfaction. Zapping behavior isn't respected due to the fact that the system wouldn't do either. The participants write down their viewing behavior for five weeks. After that, the evaluation supervisor collects all the diaries and transfers the diary data to the system. On the basis of the data, the system generates recommendations. Each participant needs to rate the system recommendations by dividing the displayed and unordered TV content into the classes like and dislike. The classification of the system and the participant are compared and on this basis the quality gets calculated. This is equally done within the sixth step in the Online Scenario (cp. section 4.1.1.). An initial profile won't be created.

4.2 User Acceptance

Measuring user acceptance for a software system is related to the field of psychology. It is necessary to figure out the conditions for a software system, which result in an actual use of the system. Already in 1989 Davis presented the Technology Acceptance Model (TAM). The TAM is an information-theoretic model, which defines the Variables Perceived Usefulness (PU) and

Perceived Ease of Use (PEOU). These two variables are defined as the main factors influencing the use of a system. Davis defines the variables as follows:

Perceived Usefulness „the degree to which a person believes that using a particular system would enhance his or her job performance“ cp. [13] p. 320

Perceived Ease of Use „the degree to which a person believes that using a particular system would be free of effort“ cp. [13] p. 320.

There is a lot of research on User Acceptance Issues in recommender systems cp. [14][15]. Especially the work of Jones and Pu in [14] about User Acceptance Issues in Music recommender systems is interesting for the work presented in this paper due to similar conditions. Jones and Pu adapted the TAM for the use in music recommendation systems by defining the variables more Application-specific. They pointed out, that 1) *the entertainment due to given recommendations*, 2) *the correct adaption of the user feedback* and 3) *the entirety of the given data base* is important for Perceived Usefulness. The factors for Perceived Ease of Use Jones and Pu name *Usability* and *Effort* until the system works properly. Caused by the fact, that we are using a simulation within the evaluation process, the point of Usability has been almost discarded. Only the Usability of the creation of the initial user profile can be questioned. We presented an adapted questionnaire from the given work in [14] with 20 items at the end of the evaluation. The user needed to apply a five step Likert scale from 1 (totally disagree) to 5 (totally agree). The items were formulated as positive or negative statements so that the evaluation user could answer with the given Likert scale.

F1	I liked the TV content which has been recommended to me
F2	The TV content which has been recommended to me tailored my taste
F3	The TV content which has been recommended were new to me
F4	I liked the TV content I already knew
F5	In general I was satisfied with the recommendations
F6	The recommendations were as good as the recommendations from my friends
F7	Many recommendations were too similar to each other
F8	The system has a huge selection of possible content
F9	I like that the system identifies my taste
F10	I can influence the quality of the recommendations with my feedback
F11	The system recognizes my taste
F12	I know how the system generates recommendations after I've used it
F13	The time the system needs to generate recommendations is appropriate
F14	For the creation of the initial profile I needed to spend too much time
F15	The creation of the initial Profile was easy and comfortable
F16	I would create the initial profile again to get proper recommendations fast
F17	The system asks me for my television watching too much
F18	If there is another technology which recommends other things to me (e.g. books), I would use it
F19	I think the system is useful and I would use it again
F20	I think the system is useful choosing interesting TV content

Table 1 Questionnaire User Acceptance

5. RESULTS

In the following section we present the evaluation results. The section is again divided into two subsections. Within section 5.1 we present the results for the evaluation scenarios; in section 5.2 the evaluated user acceptance questionnaire is presented. The results will be discussed in section 6.

5.1 Evaluation Results

5.1.1 Online Scenario

For the Online Scenario we had 51 participants, where nine of them didn't finish the evaluation. So the number can be reduced to 42 with 34 male and eight female participants. A large part of the group can be classified as educated (higher education level or university degree) with an age between 18 and 39. Table 2 presents the results for the Online Scenario. The data were tested for independence using the chi square (χ^2) test with a significance level of 5% ($\alpha=0.05$). We've proved the calculated chi square value against the quantile of the chi square distribution $p=3.84$ respecting the significance level of $\alpha=0.05$ and degree of freedom $df=1$.

Iterat.	Participants	Objects	F1 Score	χ^2 -Value
1	8	30	0.61345	16.799
2	9	50	0.65	27.023
3	8	70	0.81319	69.237
4	8	90	0.74534	55.844
5	9	110	0.73143	41.216

Table 2 Results Online Scenario

The table shows the iteration number and the associated number of participants to those the iteration number was randomly assigned. The column number Objects presents the size of the adaptive user profile e.g. for the scenario with one iteration the adaptive user profile stored 30 TV objects like films, series, etc.. In the fourth column the quality in form of the calculated F1 score is shown. As you can see at first the quality increases steadily. The system achieved a maximum quality with a database of 70 objects. With a more increasing data volume the quality decreases.

5.1.2 TV diary

For the TV diary eight people attended, mostly between 18 and 39 years and with a high educational level. The number of participants decreases from ten to eight caused by incomplete TV diaries. Table 3 presents the results for the TV diary equally to the results in section 5.1.1. Due to a limited number of participants/data records we used Fisher's exact test proving independence of the data set. We have no quality data for the 5th iteration because of insignificant test results. In contrast to the Online Scenario the quality increases continuously. The highest value in iteration four is similar to the highest quality in the Online Scenario.

Iterat.	Participants	Objects	F1 Score	p-Value
1	2	12.5	0.73913	0.0069
2	2	11	0.75472	0.0083
3	1	32	0.80	0.0325
4	2	56	0.83333	0.0108
5	1	47	-	0.0573

Table 3 Results TV diary

5.2 User Acceptance Results

The following section presents the results of the proposed user acceptance approach. In Figure 2 the median of the 20 items are shown. Negative formulated questions are marked with a dot on the top of the bar.

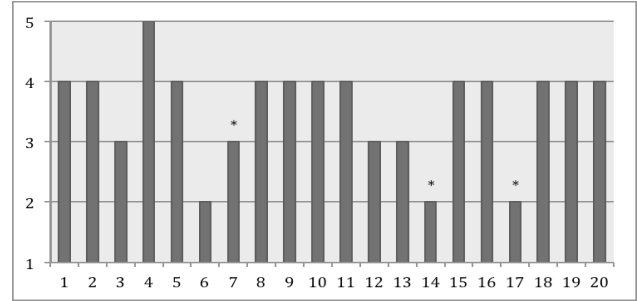


Figure 2 Acceptance Items/Median

Figure 2 shows high results within the questionnaire, which is an indication for general user acceptance. Negated questions were overall answered with a low valence, nevertheless the results tend into the right direction. For a closer look we refer to Table 1 including the question overview. The questions are divided into three sections asking for Quality, Effort and Acceptance as already mentioned in section 4.2. To analyze the results, we interpret every subarea separately. Items one till twelve concern the quality of the given recommendations. The questioned quality within the questionnaire is a rather subjective interpretation of each user, than an objective measurement. Especially the items 1,2,4 and 5 question the perceived quality by the user. All of the 4 items were rated with a minimum value of 4, which tends to be a positive feedback. Item 6 questioned if the quality of the recommendations correspond to the quality of TV recommendations from friends. The middle value of 3 is a little low, but to reach human interpretation level for the system wasn't a compound goal at all. Item 12 questioned the transparency of the system, which is with a value of 3 also a little low. However items 8 to 11 were with a value of 4 approvingly answered. On the basis of the given results with agreeing (4) and fully agreeing (5) feedback we can say that a certain quality of the system recommendations is given, which influences the user acceptance positively. For the subarea of effort we questioned the items 13 to 17. Those items questioned the fact of needed effort for using the system properly. We got a positive feedback for the items 14 to 17 with a value of 4 (agree) and 2 (not agree) for negative formed questions. These items were questioned for the construction of the initial feedback. The creation of the initial feedback is the only direct interaction between user and system needed to get recommendations presented. Just the time for creating new recommendations by the system was rated with a neutral value (3). So we can say that the effort for using the system is limited, which is also a positive influence for user acceptance. Within the subarea of acceptance for recommender systems in general and for the tested system all questions were approvingly answered (4). These results also tend to a positive user acceptance. These results have proven that a general user acceptance for the system is given. We have some particular shortcomings in comparison to recommendations by friends. In this case the system could be advanced by a connection to one of the big social networks like Facebook or Twitter. A concrete approach for this was already presented in one of our previous works cp. [16], but not respected within the evaluation.

6. DISCUSSION

In the following section the results from the evaluation presented in section 5 are discussed. We proposed two different approaches for evaluating a Bayesian classifier based recommendation system. Both approaches differ in the way data for the generating process was collected. Within the Online Simulation data was collected in a continuous manner. The adaptive profiles within the TV diary were created more qualitatively with a real user behavior. However, both scenarios were designed to examine the quality of given recommendations and to examine if there is connection between increasing database (adaptive user profile) and generated TV recommendations. First of all we can say that the system classifies objects with a quality between 81% and 83% with a database between 56 and 70 already classified objects. Because of this, it is clear that a sufficiently large database is needed to reach a proper quality. Both result tables show that there is an increasing quality with an increasing number of classified data objects. Table 2 shows a decreasing quality at a high number of data objects though. This effect has been observed in other application areas, where also a Bayesian classifier has been used for classification. In [17] the classification of Arabic text documents is described. Also in this research a large database causes a decreasing quality. The work of Billsus and Pazzani on classifying spam mails described a decreasing quality at one point of the evaluation [18]. It can be assumed that the effect just occurs within the online scenario. The Set-Top-Box prototype provides a mechanism, so that older data objects are deleted. That means the database needs to be updated and shouldn't exceeds a certain threshold. A number between 56 and 70 data objects has been determined within our evaluation. The evaluation has also shown that the quality is at a proper level, but can be increased. One approach proposed in section 5.2 is the integration of social networking to satisfy the user with recommendations done by friends. As mentioned the approach is already implemented, but not evaluated at all. A resulting increase in quality is thus just a hypothesis. It would also be possible to increase the quality by interpreting the attributes e.g. a series name synonymously. For this it is possible to implement a thesaurus or ontology.

7. CONCLUSION

In this paper we presented an evaluation for a Bayesian classifier based recommendation system for generating TV content on the basis of user behavior. In addition to two evaluation scenarios, which measure the quality of recommendations, we presented results of user acceptance evaluation based on the work of [14]. With a correct classification rate up to 83% the system reached a good quality and fulfills the demanded requirements. The results of the user acceptance gave a good feedback for the acceptance and the actual use of the system. Even though we achieved a good quality, there is still space for improvements (cp. section 6). Our future work concentrates on the evaluation of the social recommendation approach.

8. REFERENCES

- [1] B. Engel, C.-M. Ridder, "Massenkommunikation 2010", Press Conference of ARD, ZDF, Sep. 2010.
- [2] SES Astra. Available at <http://www.astra.de/2117/de>
- [3] Engelbert, B., Blanken, M. B., Kruthoff-Brüwer, R., and Morisse, K. 2011. A user supporting Personal Video Recorder by implementing a generic Bayesian classifier based recommendation system. In *Proceedings of 7th IEEE International Workshop on Pervasive Learning, Life and Leisure* (Seattle, WA, USA, March 21 – 25, 2011).
- [4] Linden, G., Smith, B., and York, J. 2003. *Amazon.com Recommendations*. IEEE Press.
- [5] Davidson, J., Liebal, B., Liu, J., Nandy, P., and Vleet, T.V. 2010. The YouTube Video Recommendation System. In *ACM Conference Proceedings on Recommender Systems* (Barcelona, Spain, September 2010).
- [6] De Pessemier, T., Deryckere, T., and Martens, L. 2010. Extending the Bayesian Classifier to a Context-Aware Recommender System for Mobile Devices. In *Fifth International Conference on Internet and Web Applications and Services* (Barcelona, Spain, 2010).
- [7] Billsus, D., and Pazzani, M. 1996. Revising User Profiles: The Search for Interesting Web Sites. In *Proceedings of 3rd International Workshop on Multistrategy Learning*.
- [8] Thabtah, F., Eljinini, M. A., Zamzeer, M., and Hadi, W. M. 2009. Naive Bayesian Based on Chi Square to Categorize Arabic Data. *Communications of the IBIMA* (10).
- [9] Das, D., and Herman, t. H. 1998. *Recommender Systems for TV*. Eindhoven: American Association for Artificial Intelligence.
- [10] Gutta, S., Kurapati, K., Lee, K., Martino, J., Milanski, J., Schaffer, D., et al. 2000. TV Content Recommender System. *Briarcliff Manor: American Association of Artificial Intelligence*.
- [11] Ardissono, L., Gena, C., Torasso, P., Bellifemine, F., Chiarotto, A., Difino, A., et. Al. 2003. Personalized Recommendation of TV Programs. In *8th Advanced in Artificial Intelligence Conference*.
- [12] Kamal, A., & van Stam, W. 2004. TiVo: Making Show Recommendations Using a Distributed Collaborative Filtering Architecture. *ACM Knowledge Discovery and Data Mining*. Seattle, Washington: ACM.
- [13] Davis, F. D. 1989. Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly Vol. 13*, pp. 319-339.
- [14] Jones, N., & Pu, P. 2008. User Acceptance Issues in Music Recommender Systems. *EPFL Technical Report*. Lausanne.
- [15] Hu, R., & Pu, P. 2009. Acceptance Issues of Personality-based Recommender Systems. *3rd ACM Conference on Recommender Systems* (pp. 221-224). New York: ACM.
- [16] Engelbert, B., Blanken, M., Kruthoff-Brüwer, R., & Morisse, K. 2011. A User Supporting Personal Video Recorder Based on a Generic Bayesian Classifier and Social Network Recommendations. *J.J. Park, L.T. Yang, C.Lee (Eds.): Future Tech 2011, Proceedings, Part II. Communications in Computer and Information Science, Vol 185*, pp. 1-8.
- [17] Thabtah, F., Eljinini, M. A., Zamzeer, M., & Hadi, W. M. 2009. Naive Bayesian Based on Chi Square to Categorize Arabic Data. *Communications of the IBIMA* (10).
- [18] Billsus, D., & Pazzani, M. 1996. Revising User Profiles: The Search for Interesting Web Sites. *Proceedings of 3rd International Workshop on Multistrategy Learning*.
- [19] Hu, Y., Koren, Y., & Volinsky, C. 2008. Collaborative Filtering for Implicit Feedback Datasets. *Proceedings of the 2008 8th IEEE International Conference on Data Mining*. Washington, D.C., USA.