ISSN No. 0976-5697

Volume 5, No. 1, Jan-Feb 2014

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Web-News Document Clustering Based on Incremental Conceptual Clustering

R.Vidya*[1], Mrs.S. Akila Devi[2], Prof. Mrs.A.Askarunisa[3]
PG Scholar*[1], Assistant Prof. M.E[2], M.E. Ph.D[3]
Department of computer Science and Engineering
Vickram College of Engineering, Enathi. India. PIN 630 561
rvidyakanna@gmail.com

Abstract: Internet provides drastic access to the news articles from different information sources around the world. The main approach is used to find out the users preference from both news content and user information. Incremental clustering is done on the web news document in order to group the documents for recommendation. The idea of conceptual clustering is used. It finds the similarity between them which is called as correlation measures. Here the data is collected from data set through various web sites of news group.

Keywords: Web-News, clustering, Information, filtering

I. INTRODUCTION

Recommender systems a specific type of information filtering (IF) technique that attempts to present information items (news, articles, music, books, news, images, web pages, etc.) that are likely of interest to the user. Web-based news reading services, like Google News and Yahoo! News, have become increasingly prevalent as the Internet provides fast access to news articles from various information sources around the world. With the gigantic amount of news articles, a key issue of online news services is how to help users find interesting articles that match the users' preference as much as possible, by making use of both news content and user information. This is the problem of personalized news recommendation. News systems in particular, benefit from recommendations given the fact that online news systems are exploratory by nature. People browse through the list of daily news usually driven by personal interest, curiosity, or both. Due to the amount of news items available, online news services deploy recommender systems that help the users find potentially interesting news Clustering is a division of data into groups of similar objects. Each group, called cluster, consists of objects that are similar between themselves and dissimilar to objects of other groups.

In other words, the goal of a good document clustering scheme is to minimize intra-cluster distances between documents, while maximizing inter-cluster distances (using an appropriate distance measure between documents). News recommender system is one of the challenges we face in this research. The most general setting in which recommender systems are studied is presented in Figure 1. Known user preferences are represented as a matrix of *n* users and *m* items, where each cell are *u* and *i* corresponds to the rating given to item *i* by the user *u* is user ratings matrix is typically sparse, as most users do not rate most items.

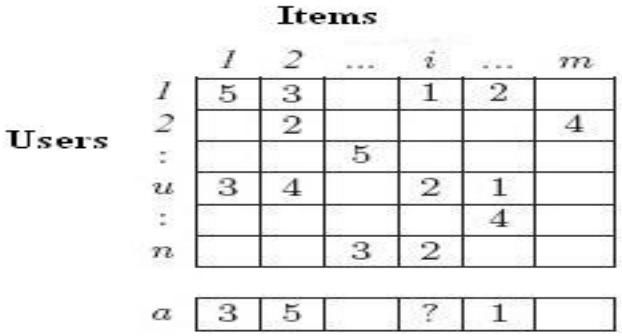

Figure1: General setting in recommender systems.

Notably, these explanations do not aim at increasing the click-through ratio of recommendations. Rather, they try to help users to realize whether a news item can be of her interest or not, by providing her with additional information on the recommendations being proposed. News Recommendation system is inevitable in the news website to recommend the updated news to the interested users. However, the system consumes high resources (computation and memory) for recording user's news navigation history, extracting creating profiles from history and maintaining user interest matrix. Our proposed hybrid recommendation model presents accurate news page to user with low Computational process and storage requirements.

II. LITERATURE SURVEY

The two basic entities which appear in any Recommender System are the user (sometimes also referred to as customer) and the item (also referred to as product in the bibliography). A user is a person who utilizes the recommender system providing his opinion about various items and receives recommendations about new items from the system. The goal of Recommender Systems is to generate suggestions about new items or to predict the utility of a specific item for a particular user. In both cases the process is based on the input provided, which is related to the preferences of that user.

Hao Wen et al. [1] reported a hybrid method for personalized recommendation of Web news to users has been presented. They proposed an approach which classifies

Web pages by calculating the respective weights of terms. A user's interest and preference models are generated by analyzing the user's navigational history. Kim et al. [2] proposed collaborative filtering principle for a network consists of group of customers. His study discuss about users contribution by uploading multimedia content, writing wiki pages, and posting blog articles.

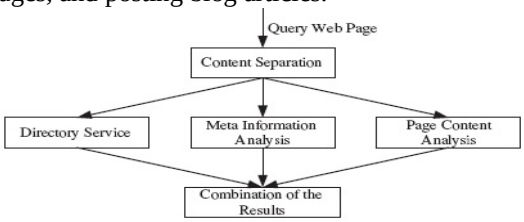

Figure 2. Web page classification by Hao Wen et al. [1].

Adomavicius [3] in his work he explored a number of item ranking techniques that can generate recommendations that have substantially higher aggregate diversity across all users while maintaining comparable levels of recommendation accuracy. In Hao Ma et al. [4] paper, they aimed at providing a general framework on mining Web graphs for recommendations, they first propose a novel diffusion method which propagates similarities between different nodes and generates recommendations; then we illustrate how to generalize different recommendation problems into our graph diffusion framework. McFee et al. [5] in their paper, they propose a method for optimizing content-based similarity by learning from a sample of collaborative filter data. Oliveira et al. [6] presented an online recommendation system that eases the matching of a user with the most relevant products and services. Their paper discuss the results gathered on experimental data analysis and the statistical hypotheses tests that were performed, which allowed concluding in which circumstances trust-based recommendation is advantageous.

Cheng et al [7] proposes an adaptive recommendation mechanism that rests on a congestion-aware scheduling method for multi-group travelers on multidestination travels. Tao et al. [8] reported a personalized ontology model for web information gathering. As a model for knowledge description and formalization, ontologies are widely used to represent user profiles in personalized web information gathering. Zorzo et al. [9] reported an adaptive automaton in recommendation systems. The recommendation systems look for to offer customized products to their users. An accurate user profile can greatly improve a search engine's performance by identifying the information needs for individual users Kenneth Wai-Ting Leung and Dik Lun Lee [10]. In this paper, they proposed and evaluated several user profiling strategies. Dimitrios Pierrakos and Georgios Paliouras presented [11] a knowledge discovery framework for the construction of Community Web Directories, a concept that they introduced in their recent work, applying personalization to Web directories. In this context, the Web directory was viewed as a thematic hierarchy and personalization was realized by constructing user community models on the basis of usage data.

A. Problem Definition:

News recommender system is one of the challenges we face in this research. Notably, these explanations do not aim at increasing the click-through ratio of recommendations. Rather, they try to help users to realize whether a news item can be of her interest or not, by providing her with additional information on the recommendations being proposed. News Recommendation system is inevitable in the news website to recommend the updated news to the interested users. However, the system consumes high resources (computation and memory) for recording user's news navigation history, extracting creating profiles from history and maintaining user interest matrix. Our proposed hybrid recommendation model presents accurate news page to user with low Computational process and storage requirements.

Web news recommendation remains challenging for at least three reasons. First, the scalability of most news recommendation services needs more research for fast and real-time processing; Second, news articles are not independent in most scenarios, i.e., browsing one news item may affect the subsequent news reading; Third, the popularity and regency of news articles change dramatically over time, which differentiates news items from other web objects, such as products and movies, rendering traditional recommendation methods ineffective.

III. METHODOLOGY

The architecture of the exiting hybrid recommender system for news recommendation on the Web is shown in Figure 3. In the system, a Web user is distinguished by identifying his or her interest and preference models. A user's navigational data is monitored and analyzed to conduct user modeling. An automatic classification method is utilized to categorize the Web contents browsed by a user. In the existing system, the user modeling method consists of two steps: determining the content of a Web page using the Web page classification method; and utilizing the Nave Bayes model for updating the user's interest and preference models. In the Web page classification method, the terms are determined by the ontology base WordNet (Miller 2009), and the weights of terms are calculated by the tf-idf (term frequency-inverse document frequency) method.

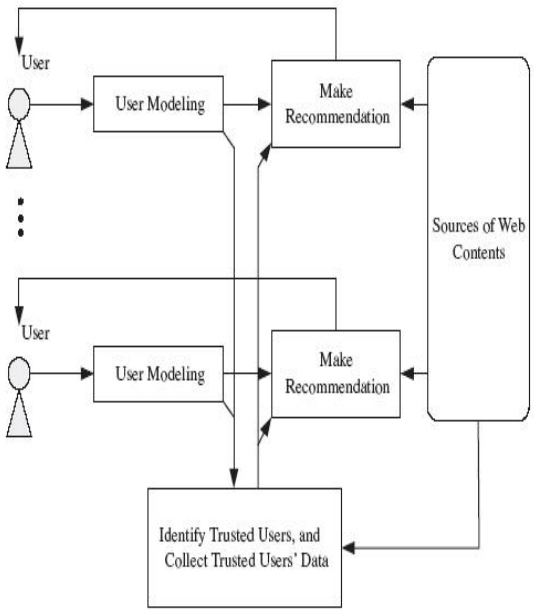

Figure 3: Existing NEWS recommending system by Hao Wen et al.[1].

In this work, a user's preference model scores a Web site based on the degree to which the user prefers to retrieve information from that Web site. The recommendation rating process of the proposed system can be divided into two steps. First, a content-based algorithm is utilized to determine the probability of recommending Web content to a user, considering the factors of the user's interest and preference models, the Web content, and the time limitation. Second, the method of collaborative filtering is used to modify the probability of recommending Web content. The system will distribute some test Web content, which has been well classified and identified by users. The users who send back positive responses are considered as the trusted users.

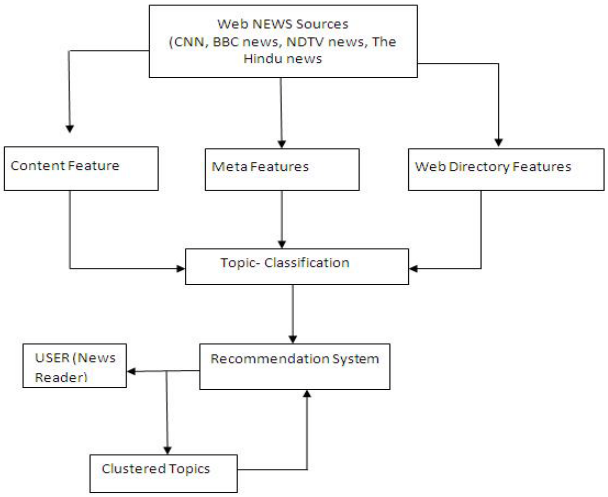

Figure 4. Block diagram

The web sources provide news documents for recommendation system with following features. Features In the proposed Web classification method, three components of a new document page are separated: hyperlink address, meta information, and effective content area. The schematic representation of the classification solution used in the system is illustrated in Figure 4. The classification method includes three steps *viz.*: content separation, parallel feature classification, and combination of the results. Initially, a Web page's hyperlink information, meta information, and content information respectively are analyzed for classification. Then, the final classification result is generated by a fusion algorithm. In this approach, in order to classify text information such as a Web page's meta and content information, weights of terms in categories of topics are used for computing the cumulative weights of terms of a target text.

A. Recommendation System:

In the proposed system, the user modelling method consists of two steps:

Determining the content of a Web page using the Web page classification method; and utilizing the Nave Bayes model for updating the user's interest and preference models. In this work, a user's preference model scores a Web site based on the degree to which the user prefers to retrieve information from that Web site. The recommendation rating process of the proposed system can be divided into two steps. First, a content-based algorithm is utilized to determine the probability of recommending Web content to a user, considering the factors of the user's interest and preference models, the Web content, and the time limitation. Second, the method of collaborative filtering is used to modify the probability of recommending Web content. The system will distribute some test Web content, which has been well classified and identified by users. The users who send back positive responses are considered as the trusted users. Additionally, the Web content browsed by more trusted users will obtain higher scores in the recommending process.

B. Topic Clustering:

Cluster analysis or clustering is the task of assigning a set of objects into groups (called clusters) so that the objects in the same cluster are more similar (in some sense or another) to each other than to those in other clusters. In proposed system, we include incremental clustering techniques for clustering user's history of news pages.

C. Techniques:

TF-IDF TF - IDF, term frequency-inverse document frequency, is a numerical statistic which reflects how important a word is to a document in a collection or corpus. It is often used as a weighting factor in information retrieval and text mining. The tf-idf value increases proportionally to the number of times a word appears in the document, but is offset by the frequency of the word in the corpus, which helps to control for the fact that some words are generally more common than others. Variations of the TF - IDF weighting scheme are often used by search engines as a central tool in scoring and ranking a document's relevance given a user query. TF - IDF can be successfully used for stop-words filtering in various subject fields including text summarization and classification. TF - IDF is the product of two statistics, term frequency and inverse document frequency. Various ways for determining the exact values of both statistics exist. In the case of the term frequency tf(t,d), the simplest choice is to simply use the raw frequency of a term in a document, i.e. the number of times that term t occurs in document *d*. TF - IDF is calculated as *TFID(T,d,D)=TF(t,d)XIDF(t,D)*

A high weight in TF - IDF is reached by a high term frequency (in the given document) and a low document frequency of the term in the whole collection of documents; the weights hence tend to filter out common terms. Since the ratio inside the IDF's log function is always greater than or equal to 1, the value of IDF (and TF - IDF) is greater than or equal to 0. As a term appears in more documents, the ratio inside the logarithm approaches 1, bringing the IDF and TF - IDF closer to 0.

D. Stemmer:

Affix removal conflation techniques are referred to as stemming algorithms and can be implemented in a variety of different methods. All remove suffices and/or prefixes in an attempt to reduce a word to its stem.. The algorithms that are discussed in the following sections, and those that will be implemented in this project, are all suffix removal stemmers. During the development of a stemmer the issues of iteration and context awareness must be addressed. Suffices that are concatenated to words are often done so in a certain order, such that a set of order-classes will exist among suffices. An iterative stemming algorithm will remove suffices one at a time, starting at the end of the word and working towards the beginning. An issue also exists

about whether a stemmer should be context-free or context-sensitive. A context-sensitive algorithm involves a number of qualitative contextual restrictions that are developed to prevent the removal of endings that, in certain situations, can lead to erroneous stems being produced. A context free algorithm removes endings with no restrictions placed on the circumstances of the removal.

Bayesian inference is a method of inference in which Bayes' rule is used to update the probability estimate for a hypothesis as additional evidence is learned. The usual form of Bayes' theorem used in the present work is given by:

$$P(H_i|D,I) = \frac{P(H_i|I)\ P(D|H_i,I)}{P(D|I)}$$

Equation shows how the prior probability of a hypothesis Hi is updated to a posterior probability p(D=Hi) which includes all the information provided by the data D . The updating factor is the ratio of two terms and only the likelihood function (or sampling distribution)), p(Hi= D,I) depends explicitly on Hi , the denominator) p(D=I) , called the prior predictive probability or the global likelihood, being independent of Hi.

IV. RESULTS

A. Incremental Conceptual Clustering:

Clustering can be considered the most important unsupervised learning problem; so, as every other problem of this kind, it deals with finding a structure in a collection of unlabelled data. A loose definition of clustering could be "the process of organizing objects into groups whose members are similar in some way". A cluster is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters.
Conceptual clustering is a machine learning paradigm for unsupervised classification developed mainly during the 1980s. It is distinguished from ordinary data clustering by generating a concept description for each generated class. Most conceptual clustering methods are capable of generating hierarchical category structures; see Categorization for more information on hierarchy. The results are shown by various screen shots from the figure 5-11.

Figure 5. Screen shot 1- displays the text from the document.

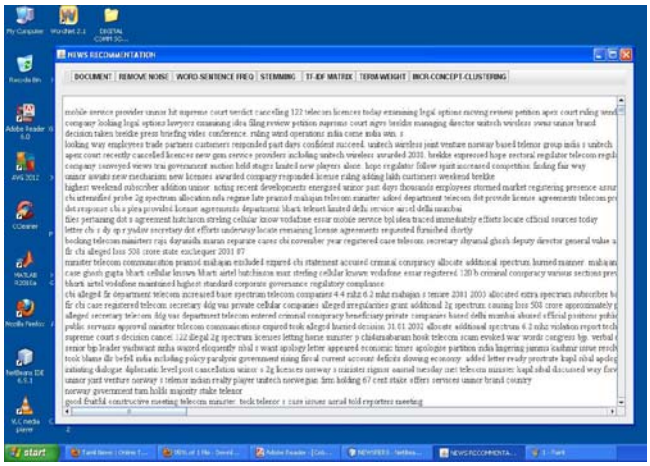

Figure 6: Screen shot 2 - removes the noise words (conjunction and prepositions).

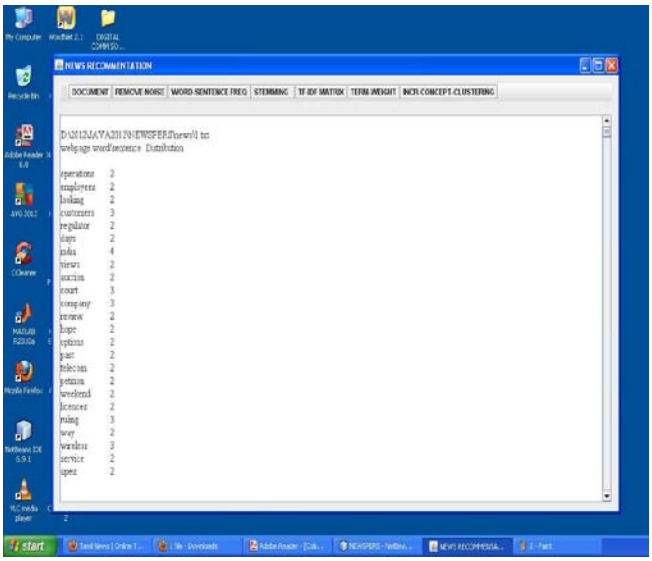

Figure 7. Screen shot 3-number of occurrences of words in the sentence.

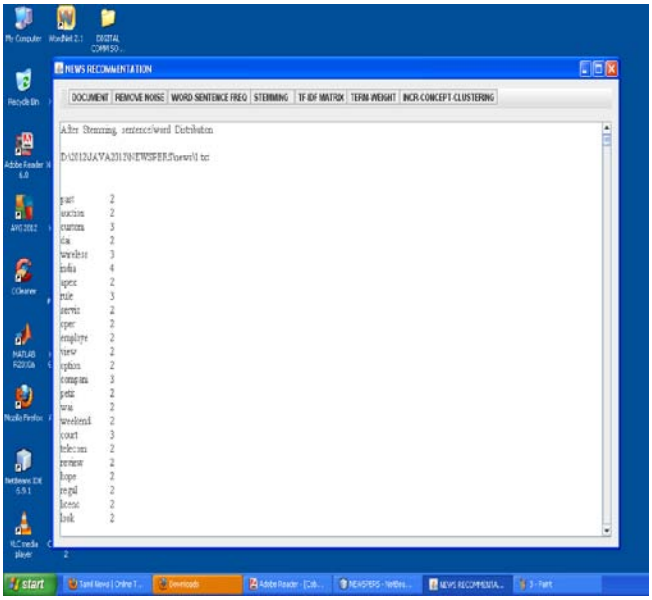

Figure 8: Screen shot 4-removes plural and ing-words.

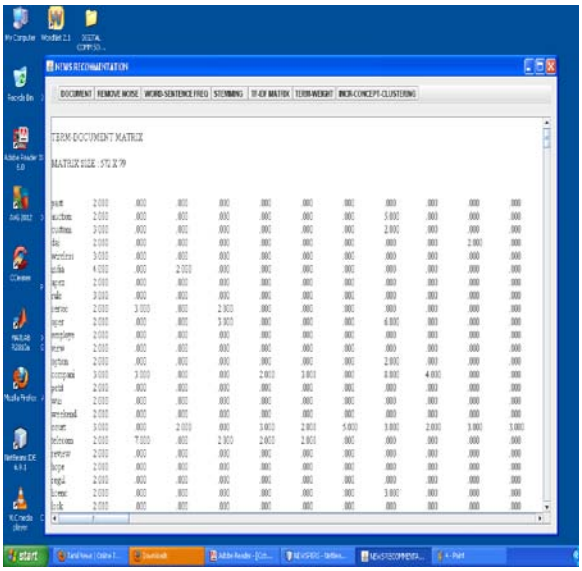

Figure 9: Screen shot 5 - number of occurrences of the term in matrix Representation.

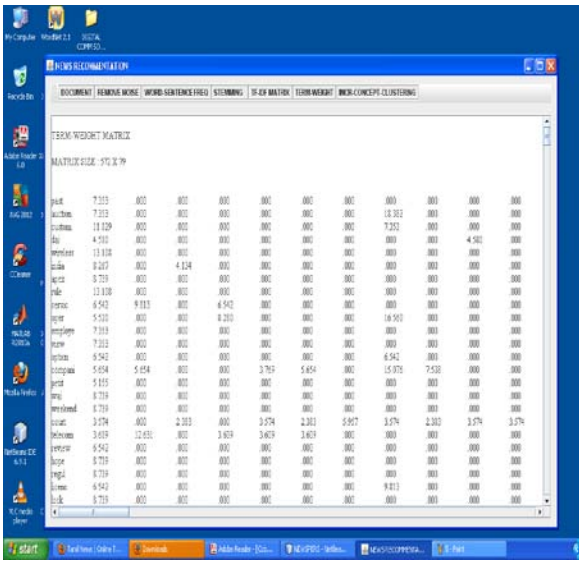

Figure 10: Screen shot 6 – term weight calculation.

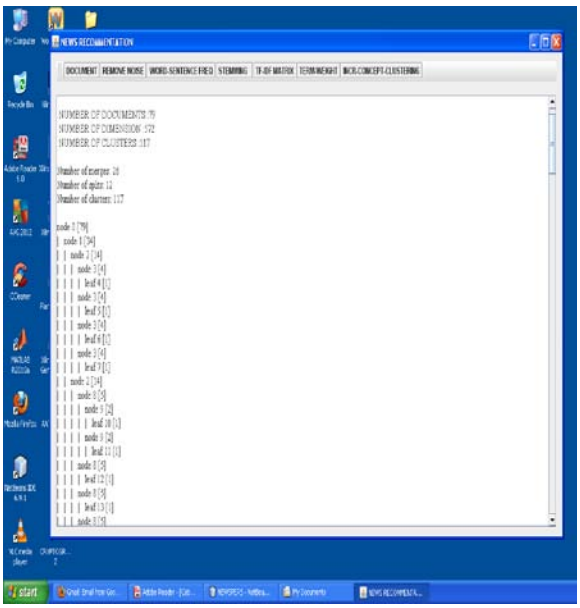

Figure 11: Screen shot 7 -correlation measurements

V. CONCLUSION

In this work, we propose a user's topic preference based recommendation system to web news recommendation. Web news collected from various news sites and explores the intra relations among news articles, along with different characteristics of news items, including news content, similar access patterns and named entities preferred by users. User's reading web news content and Meta data are tracked and analyzed through a Web page automatic topic classifying process, which is used to construct and update the user's preference topic model using Bayesian algorithm. Web news collected from various news sites is classified by the Web page classification method. Our system supports efficient incremental conceptual clustering on newly published news articles, as well as high quality of recommendation results.

VI. REFERENCES

[1]. H. Wen, L. Fang, L. Guan, A hybrid approach for personalized recommendation of news on the web, Expert Systems with Applications 39, (2012) 5806–5814.

[2]. H. K. Kim, Y. U. Ryu, Y. Cho, J. K. Kim, Customer-driven content recommendation over a network of customers, IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans 42 (2012) 48–56.

[3]. Adomavicius, Improving aggregate recommendation diversity using ranking-based techniques, IEEE Transactions on Knowledge and Data Engineering 24 (2012) 1–15.

[4]. H. Ma, I. King, M. R.-T. Lyu, Mining web graphs for recommendations, IEEE Transactions on Knowledge and Data Engineering 24 (2012) 1051–1064.

[5]. B. McFee, L. Barrington, G. Lanckriet, Learning content similarity for music recommendation, IEEE Transactions on Audio, Speech, and Language Processing 20 (2012) 2207–2218.

[6]. A. D. R. Oliveira, L. N. Bessa, T. R. Andrade, L. V. L. Filgueirase, J. S. Sichman, Trust-based recommendation for the social web, IEEE (Revisit IEEE America Latina Latin America Transactions 10 (2012) 1661 – 1666.

[7]. S.-T. Cheng, G.-J. Horng, C.-L. Chou, The adaptive recommendation mechanism for distributed group in mobile environments, IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews 42 (2012) 1081–1092.

[8]. X. Tao, Y. Li, N. Zhong, A personalized ontology model for web information gathering, IEEE Transactions on Knowledge and Data Engineering 23 (2011) 496–511.

[9]. S.D. Zorzo, P.R.M. Cereda, R.A. Gotardo, Adaptive automata in recommendation systems,IEEE (Revista IEEE America Latina) Latin America Transactions 9 (2011)152-59.

[10]. K. W.-T. Leung, D. L. Lee, Deriving concept-based user profiles from search engine logs, IEEE Transactions on Knowledge and Data Engineering 22 (2010) 969–982.

[11]. D. Georgios and P. Paliouras, Personalizing Web Directories with the Aid of Web Usage Data, IEEE Transactions on Knowledge and Data Engineering, (2010), vol. 22 no. 9) pp. 1331-1344.