Journal of Fundamental and Applied Sciences

Research Article

ISSN 1112-9867

Available online at http://www.jfas.info

A PAPER RECOMMENDER SYSTEM BASED ON USER'S PROFILE IN BIG DATA SCHOLARLY

Sh. Aghamirzad[1], A.R. Honarvar[1] and N. Jokar[1]

[1]Department of Electrical and Computer Engineering, Safashahr Branch, Islamic Azad University, Safashahr, Iran

Published online: 18 June 2016

ABSTRACT

Users encounter a huge volume of papers in digital libraries and paper search engines such as IEEE Explore, ACM Digital library, Google scholar and etc. these high number of papers make some difficulties for researchers for finding proper information and items. Recommender systems contain successful tools for knowledge of users and identification of their priorities. These systems present a personalized proposal to users who seek to find a special kind of relevant data or their priorities through the big number of data. Recommendersystem based on personalization uses the user profile and in view of the fact that the user profile encompass information pertaining to the user priorities; so it is a very active district in data recovery. Recommendersystem is an attitude that presented in order to encounter difficulties caused by abundant data and it helps users to attain their goals quickly through huge number of data. In this paper, we have presented an approach that received entire of available information in a paper, and formed a profile for each user by short and long inquiries; this profile is personalized for user and then recommends the closest paperto the

Author Correspondence, e-mail:aghamirzadeh.sh@gmail.com

doi: http://dx.doi.org/10.4314/jfas.v8i2s.150

user by the user profile. Findings indicate that suggested approach outperformsthe similar approaches.

Keywords: recommender system; bigdata; user profile; content-based recommender system;hadoop

1. INTRODUCTION

The considerable increase of data in digital library and increase of information have caused that human being is encountered with some challenges to select information. Recommender systems are applied to find the best information. Recommender systems are used to facilitate data processing [1]. In these systems, based on personalpreference, performance and behavior of users, some recommendations are given to the users consistent with their personal preference and he can decide well.

Recommender systems are classified into three groups in terms of generation method: Content-based, collaborative filtering, knowledge-based. Recommender system (information filtering technology) is useful to find the research papers and the items a user needs exactly and rapidly [2]. Text-based filter is based on the similarity between the texts of a candidate paper and target paper and each item indicates a content model with item features. Collaborative filtering is based on a technique in which a system asks some questions regarding demographic data and identifiable data and the required data are retrieved as recommendation to the user. Knowledge-based filtering presents knowledge extraction about users and items of a system and inference of links between users and items and the user requirements are fulfilled via the items and some recommendations are created.

Recommender systems have received much attention in researcher papers considerably. In the past 14 years, more than 170 papers, patents, web pages have been published in this regard. Recommender systems are useful in research papers [3]. For example, it helps the researchers in research field.

The goal of a recommender system is increasing precision and satisfaction of users of required information [3]. A user can be interested in relevant research papers but other users are interested in the publication of a paper in special field or the relevant fields of their

research. Mostly, after reading academic papers, users try to find the association between the papers by which they can alleviate the problems of papers [4].

By existing data and analysis of behavior and properties of users, recommender system presents recommendations. This smart system is defined to facilitate transactions with high information overload and information problems. Numerous information on web and internet is called information overload making decision making as challenging. Recommender system approach overcomes the challenges of information overload and these systems help the users to achieve their goal earlier. Indeed, recommender systems act as a filter, a filter showing what is required of the user. This is called information personalization. Personalization is one of the most important approaches of recommender system. In these systems, sensitivity and reaction to profiles of user are created. The user profile in recommender systems has a structure showing the information directly or indirectly and preferences of user and his working fields. A user profile is a description of the user preferences.

The user profile as a set of key weighted words shows the semantic network or semantic concepts or forum rules. The user profile is made of the information resources by different types of learning machine and information retrieving techniques [6].

Most of the recommender systems of existing papers are based on the paper profile [6].The profile-based system requires that the users register their profile in the system and recommend the papers based on similarity between the existing paper and their profile. This method has some limitations for the researchers their profile is not registered and new researchers and the user profile is not available always.

In recent years, many recommender systems are designed and implemented to identify the users, prediction of their preference and recommendations. In each of systems, based on the work field and goals, a set of techniques are updated and data are extracted. The main component in all these systems is the user profile. The method of making profile and data source in recommendation system and updating profile data and extraction techniques and application of profile data are the factors with great importance in design of a recommender system. To develop a recommendation system of scientific paper in which keywords and abstract are used in the filter via long and short queries to extract the main idea of paper, the

scientific recommendation system is based on the text filter.

The goal of building profile for users and presenting the most relevant paper is based on the user queries. At first, via the study title, the subject is defined (one's working field) and then by the abstract of candidate paper showing the main idea of paper, the paper of the user is presented to the user via search engines as google scholar.

This study attempts to extract the paper idea by short queries, work field of user (author) and long queries and by building the profile for each user, better recommendations are presented to the user. Later, the studies regarding recommender system and user profile are explained. Then, the study method (recommended method) is defined and finally findings (results) are presented.

2. RELATED WORKS

The general evaluation is divided into three groups: At first the studies regarding recommender systems and extraction of work field (subject) and then building a profile for user.

The review of literature was conducted during 2001-2010. This study was regarding the classification of recommender system based on their applied file and recommendation method. Based on this study, most papers were regarding film applied plan and the film dedicated first rank and second rank was dedicated to recommender system for purchase. The third rank for book and research was increased in recommender system regarding book and references since 2007[16].

In a survey of recommender services regarding digital library by [17], they mentioned some recommender systems. Scientific organizations as ACM, IEEE didn't present recommender systems. They stated that although recommender systems had considerable benefits for students and scientists, in digital library as present productivity of recommender system is not implemented to the social space and communities [17].

Evaluated morethan 80 methods for academic recommendations including 170 research papers and patent, weblogs. They applied three main evaluation methods (user-online evaluation and off-line evaluation). They found that users don't score the papers in

recommender systems. They apply implicit scores as reference and download to eliminate this problem [15].

Forscientific paper's topic extractionLatent Dirichlet allocation (LDA) has been considered in previous methods to extract topics from text documents. The basic idea behind LDA is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words [8].

[9] usedLDA to find scientific topics from abstracts of papers published in the proceedings of the national academy of sciences. but the main disadvantages of this model are that the topics are distribution over single words and thus the semantics is lost; and this method needs a pre-defined number of latent topics (people can choose a different number of topics, thus producing different results) and manual topic labeling, which is usually difficult for people.

Proposed[5] a method that uses closed frequent keyword-set of the titles' phrases to form topics. In their proposed approach, they form closed frequent keyword-sets by top-down dissociation of keywords from the phrases present in the paper's titles on a user-defined minimum support.

Proposed a collaborative topic regression model which combines ideas from CF and content analysis based on probabilistic topic modeling. They developed an algorithm for recommending scientific articles to users of online archives where each user has a library of articles that he /she is interested in, and their goal was to match each user to articles of interest that are not in his/her library. They used the abstract and title of the paper to model a user and characterize candidate papers to recommend, which occasionally results in irrelevant recommendations. The abstract and title are not good enough to help know the content of the paper as some abstracts are not well written due to the expertness of the author or the abstract length limit(may be 100 words) suggested by journal or conference format[13].

In [12] the problem of paper recommendation in Big Data scholarly was addressed. In this work an approximate approach for recommending papers to researchers based on local sensitive hashing proposed by converting the citations of papers to signatures and comparing these signatures against each other to detect similar papers according to their citations. A parallel and distributed aspects of the proposal is also discussed.

Introduced a source independent framework for research paper recommendation. Their method requires as input only a single research paper and generates several potential queries by using terms in that paper, which are then submitted to existing web information sources that hold research papers**.** They consider title, abstract and body as target section for queries generation and used title and abstract section for candidate paper generation, stating that title and abstract are only publicly available section for researcher. Their approach generates a 2-gram word and noun-phrases extracted using part of speech tagging as queries. However, we feel that such a small span of text does not effectively represent a user's interest of the candidate paper [7].

Proposed an unsupervised learning method for research papers organization, this method extracted topics based on the relationship between the paper's title, frequent sentences and most similar references to the paper's title. It means that only frequent sentences and cited references most related to the paper's title are only considered in topics extraction; and based only on this, some topics are not extracted. A better approach is to extract all paper's topics, it is proposed in this paper, because paper's title features of top frequent sentences relationship, keywords relationship and references relationship are not enough to get all topics that the paper is addressing [11].

Conductedsome studies on ranks each user gives to items. Rating methods were used to build the user pro file instead of user customization and to recommend a new item, collaborative filtering-based system is applied and recommender system to recommend each new item should have a history in system of the item. To build user profile, feature vector is used and here is 5 vectors and each victor shows an aspect of features (or a dimension of user preferences) and a profile vector including some attributes (value). Another method to build the profile is using fuzzy logic. At first we put a summary of each film favorite of the user inside a candidate set. Then, we eliminate the stop words and by porter algorithm, the stem of words is found and this set is used as fuzzy inference input [14].

3. PROPSED METHOD AND EXPRIMENTS

To build profile for user (based on data), at first working filed and the paper idea should be

achieved and they are important factors to create profile for the user and are applied with other mentioned items in the previous section to produce the profile. The method of extraction of idea of paper and working field as called long and short queries in this study are explained in this study.

In this study, the set of data is collected from the papers in 2014 by digital library Cite Seer[x]and each paper is a file with format XML with the name of author, DOI (each paper has special number), title, keywords, abstract and each part of paper with reference, there is the text of reference and there are the year of publish and place for the author [18]. Place feature is an important factor to recommend and assesse the papers but it is not based on our goal and we ignore this factor.

3.1 Extraction of long queries

In this study, to achieve the main idea, two methods are used. First method, in which abstract as the main part of paper is available to all researchers and the second method is using the paper text (the text of references inside the paper) and this method is called the extraction of long queries.

First method: Using the abstract of paper

At first, some sentences as Cue word are shown in Table 1 and are found inside the abstract and these terms include keywords and key terms. These sentences are selected via candidate queries and sometimes, they include unnecessary terms are labeled via Part of speech tagging (POST).

Table1. An example ofcue words

in this paper	approach	method	contribution	framework	study	algorithm	solution	model

After finding the terms with cue words and labelling these terms, the sequences of words consisting of two or some nouns or adjectives are considered as the idea of paper.

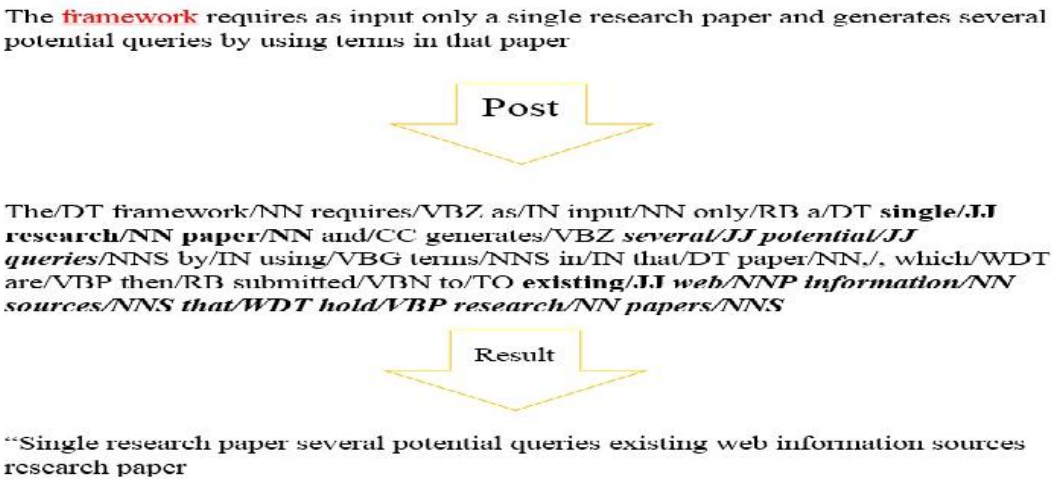

Fig.1. An example of extraction of paper idea

An example [7] of idea extraction is shown in Figure 1. The first part considers a sentence of abstract with cue word (framework) and then via labeling and eliminating extra words, the residual words are considered as the idea of paper.

Second method: Body of paper

Here, the paper body is used (the text of existing references) and the applied sentences are similar to the title and the title has key terms and the sentences with these terms are considered as they consist of important terms. At first, the sentences similar to title are found, then via TF, the terms with highest frequency (important terms of title) are achieved and are adapted with the title terms and the terms with highest similarity with title are extracted.

At first, we consider the similarity of each sentence of text as title.

$$\text{Sim}(S_i, T) = \frac{\vec{s_i} \cdot \vec{T}}{max_{s_j \in P}(\vec{s_j} \cdot \vec{T})} \quad (1)$$

Title and any sentence of target paper are presented as a vector of content words. $\vec{T}$ denotes title vector and $\underset{S_i}{\rightarrow}$ denotes sentence vector in target paper p.

This method depends upon the title and at first we should measure the important values of terms and then we add up these important values with the sentence. To do this, each S_i should have definite frequency and via TF, frequency of each sentence is achieved:

$$\text{Score}(S_i) = \frac{\sum_{t \in P} tf(t)}{max\{\sum_{t \in p} tf(t)\}} \quad (2)$$

Tf indicates frequency of terms of term t

Combining both formulas is used to achieve value (size) of Si sentence.

Total Score(S_i)= Sim(S_i,T)+ Score(S_i) (3)

Finally, a sentence with highest frequency and much similarity to the title is extracted from the body of paper and to eliminate the unnecessary terms, post is applied [4].

3.2Short queries extraction

A scientific paper deals with some subjects and the words in the paper reflect a set of features of the subject. The subjects of scientific paper define the main issue of author. They show a short summary of the paper content and can perform the recommendation of papers by retrieving or via similar subjects. In this study, short queries extraction is called defining the paper subject.

To do this, two methods are used: a) Using the title and reference of paper, b) Using paper abstract

a) Using the title and reference of paper

At first, we consider the title term and define stop words. Then, the title term is decomposed to some definite terms based on the number of stop words (from where stop word exist). For example, if there is one stop word in the title, the title is decomposed to two sub-terms, then the words of each term is stemmed by porter algorithm and finally are weighted by tf of decomposed terms and the words higher in terms of frequency in the text are considered as the subject.

In references beside what we performed in the title, we define punctuation. At first, reference terms are decomposed by stop words and they are also decomposed by punctuations and then their frequency is computed in the paper and the terms with highest frequency are considered as the paper subject.

b) Using the abstract

In this method, statistical data are used, then the terms with the highest frequency are extracted. These terms are considered with the order of the words in title. This method is as

each term of abstract is considered and the words with the highest frequency are defined, we consider the terms with frequency above 2. We believe that if a term has the highest frequency in a paper, some neighbor terms can be occurred at the same time (this idea is considered in the paper). After extraction, repetitive issues are excluded and sometimes, overlapping is occurred in the extraction and these terms are the sub-set of long terms and repetitive terms are used as the subject. For example, if the terms “research”, “recommendation” are repeated frequently and neighbor of the term (paper), then “paper recommendation” and “recommendation of research paper” are extracted and based on the method “recommendation of research paper” compared to the other one, the paper subject is expressed better.

3.3 Recommendation method

As it was said, in the previous section, for each user, a profile is formed, then for each attribute, a vector is considered and these vectors show the user interest. The system checks each vector of target paper in the user profile with the candidate paper feature vector via cosines similarity and in case of highest similarity, the paper is recommended to him.

There are two states for the user entering the system for the first time a) It is possible the user enters a word to show his interest, as he enters just a word and it doesn’t have all the features (user profile), it recommends to the user all the papers with having the required term, b) In the second case, it is possible the user enters the system and no word is entered to show his interest and in this case, the system recommends some papers with different issues randomly to identify the interest field and recommends the relevant paper based on selection of one of the subjects by the user.

3.4 Implementations and Experiments

Due to high volume of data in this study, Mahoutprocessoras the set of Hadoop project is applied. Hadoop is an open software framework enabling the distributed processing of big data on some clusters of servers. This framework is written in Java and is designed to do distribute processing on thousands of machines with high error toleration. Mahout processor as the inseparable component of Hadoop project is a space to create machine learning algorithms as distributed. This sub-project includes various algorithms and libraries for data mining. Mahout provides math calculation libraries according to linear algebra and statistics

for JAVA. Mahout is recognized as one of the subj-projects of Hadoop but it doesn't mean it is dependent upon Hadoop. Mahout can be used without Hadoop and on a single node, even non-Hadoop cluster. Mahout is applied in this project due to having prepared functions and more libraries and increased speed of processing.

In this study, the set of data is collected from the papers in 2014 by digital library Cite Seer[x]and each paper is a file with format XML with the name of author, DOI (each paper has special number), title, keywords, abstract and each part of paper with reference, there is the text of reference and there are the year of publish and place for the author [12]. Place feature is an important factor to recommend and assesse the papers but it is not based on our goal and we ignore this factor.

3.5 EvaluationCriteria

As recommender systems are evolved in recent years, we can hardly state about their assessment criteria. The main challenges is selection of a suitable metric covering high diversity of published criteria for quantitative assessment of recommender systems and include all of them.

Here, the proposed method is considered to recognize its effectiveness. There are two assessment techniques to measure the effect of proposed method as influence assessment of proposed method to extract the subjects and idea of paper and second case is performance of assessment of applied metrics to recommend the study paper in this study.

3.6 The findings of assessment of extraction of subject and idea of papers

To evaluate the method, the authors and co-authors of papers are used as a group consisting of 20 researchers publishing papers in computer sciences. They gave their published papers to us and they were asked regarding the support or reject of idea and subject. Many ideas and subjects were supported by researchers and a few of them were rejected. Generally, the papers presented by 20 researchers are 1142 subjects, of which 927 were supported by the researchers as 81.2. Of 541 extracted ideas, 424 terms and about 79.5 of extracted terms were supported by the subjects.

In this study, the two recall evaluation criteria are used and to evaluate precision, Discounted Cumulative Gain is used to measure the quality of recommendation rating.

Recall: A fraction of positive items evaluated by queries is considered and as an input document of a query is assumed, the relevant papers are called positive and irrelevant papers are called negative and recall precision is computed as follows.

$$\text{Recall}=\frac{tp}{tp+fn} \quad (4)$$

tp[1] is the number of papers preferred by the user and is recommended truly to the user, fn[2] is the number of papers preferred by the user but is not recommended to the user.

The results of recall metric are shown in Table 2 and assessment of each of methods is defined separately.

Table 2. As a result the extraction queries with a metric recall

Fields	Source (cite seer ^{x})
Title and references	%42
Body	%39
Both fields for short queries	**%81.2**
Abstract	%41.5
Body (top and important sentences)	%38
Both fields for long queries	**%79.5**

DCG[3]: It is measurement of rating quality computed as follows.

$$CG_p= rel_1 + \sum_{i=2}^{p} \frac{rel\ i}{\log_2 i} \quad (5)$$

P is situation for calculation in DCG and $rel_i$is the association value of items in situation I. The association value is as: Highly relevant items score 2, relevant items score 1 and irrelevant score zero.

The assessment of recommended recommendation method focuses on retrieving papers and presenting the most relevant paper. DCG compares and rates strategies by assessment metric. The results are shown in Table 3.

Table 3. The results proposed method

Approach	Source (cite seer ^{x})
The proposed method(Both short and long queries)	%85

[1]True positive
[2]False Negative
[3]Discounted Cumulative Gain

As shown in Table 3, the proposed method can present papers with the highest rate to the researchers. As the considered fields play important role in the main content of paper as abstract as exact summary of paper, introduction and text showing the details regarding the paper and problems and solution in this section and the best search is created and the most relevant paper is presented to the researcher.

4. DISCUSSION AND CONCLUSION

As the scientific papers are increasing in digital libraries, presenting a relevant paper to the author is a challenging issue and users ask for the personalization of recommendations. This systematic paper is recommended and it is an improvement method of recommendation of scientific papers by user profile (author) and by the papers written by the author in the past, the work field and the main idea of paper are defined and long and short queries are used to retrieve the applied papers. By the existing features in dataset as title and abstract and publish year for each user, a profile is built. This method is text based in which the content of both target and candidate papers is used to recommend the best paper to the user. Some methods are recommended to generate the queries and these queries are used for dataset of the online information to recommend the most relevant paper. Finally, by cosines similarity between the fields and selected papers to be presented for paper can be computed. Compared to the proposed other methods, the results of tests show the best performance of proposed method of this study as in the proposed method, the entire paper is used and no part of paper is eliminated like the previous methods. By eliminating each of parts of paper, exact recommendation is less possible. In the proposed method of this paper, all parts are considered and Mahut processor is used as we are encountered with considerable number of papers. The comparison of processor is shown in Figure 2. The proposed method as query is shown in Figure.

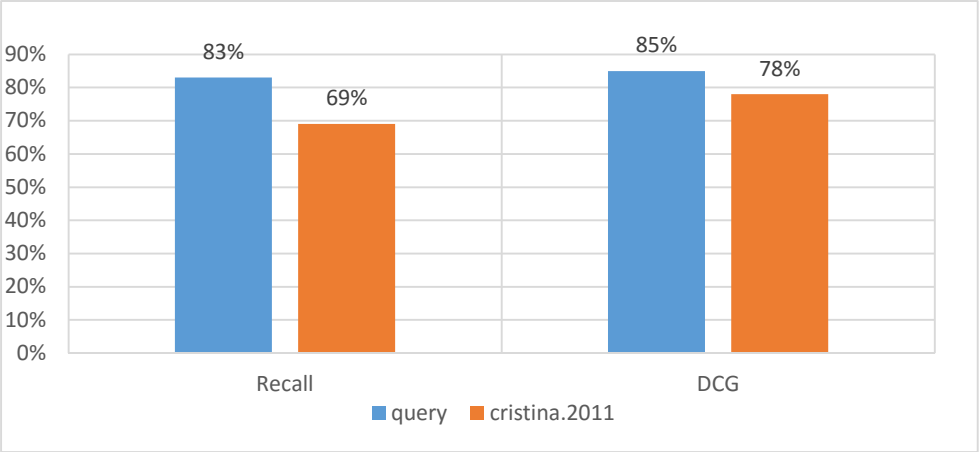

Fig.2. The comparison of the proposed method with the similar methods

The recommendations for further studies are using labeling of total data and all useful words can be selected among many other words and this increases the speed of queries processing and indexing the attributes can improve the system efficiency.

5. REFERENCES

[1] Joonseok L., Kisung L., Jennifer G. Kim. (2010), Personalized Academic Research Paper Recommendation System.

[2] Dehghani Champiri, Z. et al. (2015). A systematic review of scholar context-aware recommender systems.elsevier.

[3] Joeranbeel, S. L. A Comparison of Offline Evaluations, Online Evaluations, and User Studies in the Context of Research Paper Recommender Systems

[4] Damien H, Liao B,Vincent H., Dennis N. Faustin K., (2015), An Effective Academic Research Papers Recommendation for Non-profild Users

[5] Shubankar K., Singh A.and Pudi V. (2011),” A Frequent Keyword-Set Based Algorithm for Topic Modeling and Clustering of Research Papers”, In the proceeding of the 2011 third Conference on Data Mining and Optimization (DMO) , IEEE, June 28-29, pp. 96-102, Selangor, Malaysia

[6] Kwanghee H., Hocheol J., Changho J. (2013), Personalized Research Paper Recommendation System using Keyword Extraction Based on User Profile

[7] Nascimento C., Laender A. H. F., da Silva A. S. and Gonçalves M. A., (2011) A Source Independent Framework for Research Paper Recommendation”. ACMJune 13–17, Ottawa,

Ontario, Canada

[8] Blei D. M., Ng A. Y. and Jordan M. I., "Latent Dirichlet Allocation", Journal of Machine Learning Research, 2003, 3, 993–1022

[9] T. Griffiths and M. Steyvers, "Finding scientific topics. In Proceedings of the National Academy of Sciences", (2004), 5228–5235

[10] He Q., Kifer D., Pei J., Mitra P. and Gilee C. L., (2011) "Citation recommendation without Author Supervision", In the Proceedings of the fourth ACM international conference on Web search and data miningFebruary 9–12, Hong Kong, China

[11] Hanyurwimfura D., Bo L., Njangi D. and Dukuzumuremyi J. P., "A Centroid and Relationship based Clustering for Organizing Research Papers", International Journal of Multimedia and Ubiquitous Engineering, 2014, 9(3), 219-234

[12]SiroosKeshavarz, Ali Reza Honarvar, (2015), A Parallel Paper recommender system in Big Data Scholarly, International Conference on Electrical Engineering and Computer

[13] Wang C.and Blei M.D., "Collaborative Topic Modeling for Recommending Scientific Articles", In Proc. of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, (2011), 448–456

[14] Qing L. and Byeong M. K. (2004). Constructin User Profiles for Collaborative Recommender System.

[15] Joeranbeel, stafanlanger,marcelgenzmehr,Belagipp .(2013)Research paper recommender system Evaluation a Quantitative literanture survey.

[16] Park, D. H., Kim, H. K., Choi, I. Y., & Kim, J. K. A literature review and classification of recommender systems research. Expert Systems with Applications, 2012, 39, 10059–10072

[17] Franke, M., Geyer-Schulz, A., & Neumann, A. W. (2008). Recommender services in scientific digital libraries. In Multimedia services in intelligent environments.Springer, pp. 377–417

[18] http://citeseer^{x}.ist.psu.edu

How to cite this article:
Aghamirzad Sh, Honarvar AR, Jokar N. A paper recommender system based on user's profile in big data scholarly. J. Fundam. Appl. Sci., 2016, *8(2S), 941-955.*