Journal of Fundamental and Applied Sciences

Research Article

ISSN 1112-9867

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

A CONTEXTUAL INFORMATION BASED SCHOLARY PAPER RECOMMENDER SYSTEM USING BIG DATA PLATFORM

N.Jokar[1], A.R.Honarvar[1], Kh.Esfandiari[1]

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

Published online: 18 June 2016

ABSTRACT

Recommender systems for research papers have been increasingly popular.In the past 14 years more than 170 research papers,patents and webpageshave been published in this field. Scientific papers recommender systemsare trying to provide some recommendations to each user which are consistent with the users' personal interests based on performance, personal tastes and users behaviors.Since the volume of papers are growing day after day and the recommender systemshave not the ability for covering these huge volumes ofprocessing papers according to the users' preferences it is necessary to use parallel processing (mapping – reducing programming) for covering and fast processing of these volumes of papers. The suggested system for this research constitutes a profile for each paper which contains context information and the scope of paper. Then, the system will advise some papers to the user according to the user work domain and the papers domain. For implementing the system it has been used hadoop bed and the parallel programming because the volume of data was a part of a big data and the time was also an important factor. The performance of the suggested system was measured by the criteria such as user satisfaction and the accuracy and the results have been satisfactory.

Keywords: Recommender systems; big data; Hadoop; contextual information.

Author Correspondence, e-mail: nasrin.jokar@gmail.com

doi: http://dx.doi.org/10.4314/jfas.v8i2s.144

1. INTRODUTION

The general application of the Web as a global information system has led to the users deal with huge volumes of data. The final users are confusing with choosing the desired information in facing with these huge volumes of data from tens of thousands of sources.Today, the organization of this data is essential.Web personalization gadgets have facilitated the access to this data.In other words these gadgets guarantee the delivery of the right data to the people at the right time. The recommender system is the most successful example of personalization tools on the Web which in recent years has been of interest to the researchers. The recommender systems are systems which are trying to offer suggestions to each user based on personal taste and user behaviors and also to help the users in decision making process. Designing and implementing such systems will be undeniable by growing business on the Web,E-learning, increasing the users' communication with each other and emerging of social networks. For this purpose, several algorithms have been used and most of them are based on collaborative filtering algorithm and content-based algorithm. Thecontext aware recommender systems are systems which are trying to offer suggestions to users so that they are consistent with their personal desires based on performance, personal tastes and the user's behaviors and depending on the context in which they are used. In recent years, the recommender systemshave used the description of the status and the situation of the users' information such as the location, time and work for more relevant and personalized recommendations.For example, it is given a different recommendation to a user who is an undergraduate student and follow a phased approach to deliver in the classroom than a user who is a postgraduate student and is writing research paper[1]. One of the most cited definitions from computer science perspectivehas been proposed by Dey and Abowd[2].They said,any information that can be used to identify an entity is a context or text. They have classified the context to fourdimensions:location, identity, time and activity.There are two levels of contexts in this definition:Primary contexts or origin contexts which have four undermentioned dimensionsand secondary contexts which are derived from primary contexts (Figure 1). For example, many issues related to the information such as phone number,

address, email address, date of birth, etc.can be obtained from the entity.Such information obtained from the primary contexts is considered as secondary contexts.

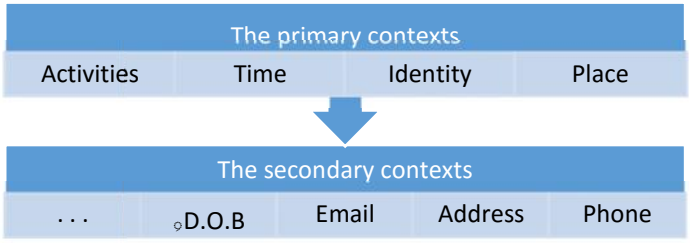

Fig.1. classification of context

The context-aware recommender systems and the conventional recommender systems have the same goals and provide services and information for the users. The difference between these two systems is that context-aware systems work based on the users' context and the recommender systems work based on users'interests and preferences.It must be said thatthese two systems are not competing togetherbut are complementary to each other[2]. The recommender systems have been increasingly popularfor research papers.In the past 14 years, it has been published more than 170 research papers,Patents and web pagesin this field. The recommender systems for research papers areuseful programs. For example, they help researchers who pursue their research area[3].Text-based paper recommender systemsand referral -based paper recommender systems are papers which will return the results according to user query.Text-based recommendations are usually done by using specific keywords which are given to search engines like Citeseer and ScholarGoogle[8]. In referral-based recommendations the purpose of recommendation is referringhelp to the authorsin selecting the most relevant papers for referring the number of potential sources[9].The brackets are used as contexts for referring. The extracted contextual for digital libraries information have classified into three main groups; user documents and environment[1].Users contextual information include user profile, user type, the purpose, the activity, the work, prior knowledge,skill, social networking and information of the main pages,logs,the behavior of the information and the level of information need. For example, the meaning of user type,student and faculty member is the researcher or the meaning of the level of information need is that

the user needs that level urgent or non-urgent.Each paper has certain characteristics that are different than other papers.These features can be considered as contextual information of the papers. The contextual information include bibliographic information, referring between papers and the popularity of the papers. The contextualenvironment information include the place of items in the libraryor geographic location, time and the type of services. The bibliographic information of the papers include title, number, abstract, keywords, author, publisher, date of publication, paper classification, The original paper, paper format, language and thepaper status in terms of ranking[1].

Hadoop isa set of software and libraries that provide the mechanisms of processing huge amounts ofdistributed data. In fact, Hadoop can be likened to an operating system which is designed toprocess and manage huge volumes of dataon different machines.Map Reduce is a parallel data processing method which is used for implementing the logic of data processing.This method helps users produce programs that they would be able to process data in parallel on a large clusterof ordinary computers and have a high tolerance for errors[11].The aim of this study is recommending the most relevant papers to a user according to the paper domain and the user working area.These recommendations are carried out through content-based method and using contextual information such aspaper title, abstract, keywords (including keywords from the author of that paper, IEEE terms, controlled indicators in INSPECT,and non-controlled indicators in INSPECT). Since the volumes of papersis growing day by day and the recommender systems have not the ability of covering these massive volumes for processing papers according to user preferences thus, for fast covering and processing of these volumes of datawe need to use parallel processing (Map ReduceProgramming)which is implemented in the context of Hadoop.This study includes the following hypotheses. It can be submitted useful answerto these questions after the implementation of the system.

Is it possible to offer more accurate recommendations using contextual information and limiting the domain?

Is it possible to satisfy the user by recommending the most relevant papers?

Related works

In this study [1] it is essential tounderstand and exploituser requirements for predicting the exact recommendations to users on a specific domainlike DLS.The context-aware recommender systems authors studied contextual information to build a recommender system and to identify the methods for adopting contextual information and to identify the communication between the contextual systems in the recommender systems and in an academic domain. Finally, the results showed that the contextual systems extracted for the recommender systems in the academic DLS are classified into three main groups: user, documents and the contextual systems environment.

In [4,5] their studies have suggested a personalized recommender system that has increased the accuracy of finding research papers. This system also can be possible and applicable for developing the recommender system of a specific domain. The proposed algorithms to extract the keywords is that at first the unnecessary characters such as or, off, is are removed and then keywords in the title and preface, keywords for paper and keywords on the first page of paper are extracted.If the keyword does not exist a few words from the title can be selected and formula 1 shows the reflection between the title and keywords.

$$\text{Reflex} = \frac{\#\,\text{Keyword Title}}{\#\,\text{Title Term}} \qquad \text{Formula 1}$$

The user profile is collected by each user click on a research paper. Whenever a paper was selected domain frequency, subject, and profile keywords will be increased.To calculate the similarity it was used the cosine similarity and then the rate of each event and user profile reflector recalculated.To calculate the similarity it was used the cosine similarity.

In [6] study the important concepts and prepositions or some interesting phrases was extracted and then these interesting phrases was classified through using a simple Bayesian classification.The domain of scientific paper main subjects is a branch of an academic field. For example, computer context is domains such as networking, parallel computing, theory of computation, data analysis.Each domain can have different subdomains. In such case the second domain is considered. For example,the second domain for data analysis domain includes NLP and pattern recognition.The method of this study extracts the root or context of each paper in the form of paper domain. The interesting phrases based on their place beside of

the specific prepositions are derived from the results of disambiguation of prepositions. The authors of this study have used integration and keywords for extracting the interesting phrases because the titles of the papers tend to have unique and on the other hand, phrases or keywords are usually used more widely for tagging the domain. Then the common interesting phrases in the title and also the keywords as a derivative are maintained for retaining.In order to classify the interesting phrases and the keywords few input data is tagging and each derivative is classified as the "domain" or "No domain" using Bayesian algorithm.

In [10] 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.

In [7] has focused on local recommendations.Citation context [C] is defined as a sequence of words which seem around specific citation. Usually citation context contains words which describe or summarize the referred papers. Directly, the Ontology of the cited documents should be close to the citation context. Authors by a neural network model, estimate the percent of possibility of citation to paper, according to the citation context. This method ensures that keywords used to refer similar documents have a high semantic similarity. To assess this model, a general test is performed on data collection (citeseer) and has better operation than State Of Art.

Proposed Method

The recommender system of this study is that at first instance, for each paper a profile is considered using contextual information such as title, abstract, keywords (include keywords that the author of the paper has considered for that paper, terms from IEEE,controlled indicators in INSPECT and non-controlled INSPECT).Then, using the cosine similarity between the user work domain and the papers domain recommends the papers which have more similarity to the user work domain. The user work domain can be registered by the user in his or her profile or the system offers the papers from each domain randomly when the domain was not determined by the user.When the user selects one of these papers, system

saves the paper domain for the user work domain and it will recommend papers using it. The data in this study was a set of controlled data from IEEE conferences and journalsfrom 2013 to 2015[12] that using a reptile has received URI of each paperas input and has saved the contextual information in the form of an xml form. Preprocessing operations such as removing the word inhibitor,wordrooting out and word weighing are applying on data based on the parameters tf / idf in the form of bigram.Alsowords that have listed in the abstract for more than twiceare added tokeywords and domaindeterminant words (papers that have not keywords and domains are filtered).The domain system for 80% of trained papers and20% of the remaining papers can be recognized using cosine similarity between domain vectors of each paper. Since the size of data was huge and the response time was important it was used Map Reduce programming for implementing and processing of data in the context of Hadoop. In this study it was used user satisfaction and accuracy criteria which were presented by [5] for evaluation.

$$SAT= \frac{the\ number\ of\ correct\ papers}{the\ number\ of\ recommended\ papers} \qquad \text{Formula 2}$$

$$ACC= \frac{recommended\ by\ unsat\ and\ sat}{the\ number\ of\ saved\ papers}$$

According to the above formula SAT equals to the number of correct papers on the number of recommended papers and ACC equals to the number of correct and non-correct papers on the total number of saved research papers for a specific subject. To measure user satisfaction we limitthe maximum number of recommended papers to 20 but for accuracywe have no limit.

2. EXPERIMENTS AND RESULTS

To evaluate the system it was asked 20 undergraduate and master's candidates and 5 PhD candidates to use this systemfor 3 weeks for each of computer, electricity and mechanic domain. Each user selected the work domain. According to that domain system recommended papers to them. The users' recorded true or false possibility of the domain for each recommended papers and then each user declares the average of his or her accuracy and satisfaction.According to the above criteria and the average of each users' accuracy and

satisfaction the results show the system function in figures 2, 3 and 4.The system performance will reduce by increasing the level of education because of recommending papers with more specific domains.

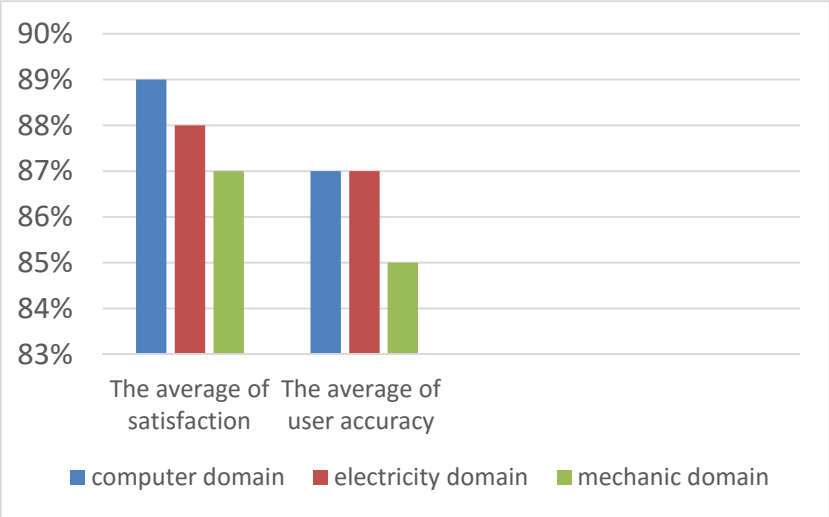

Fig.2. The undergraduate candidates' results

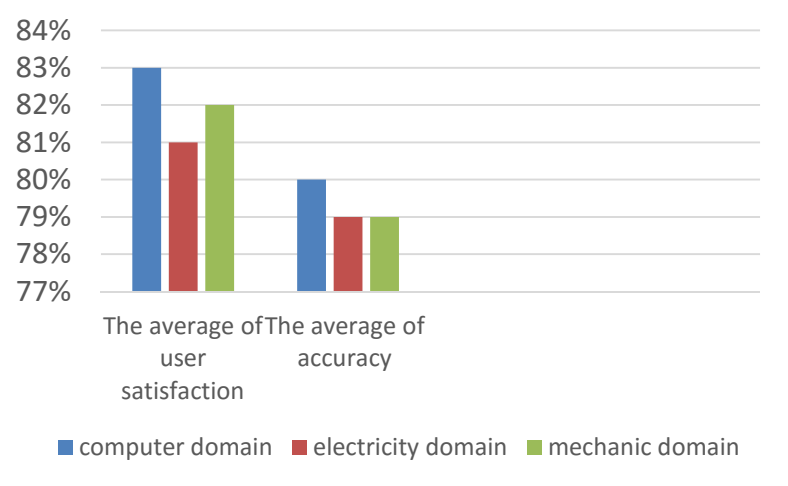

Fig.3. The graduate candidates' results

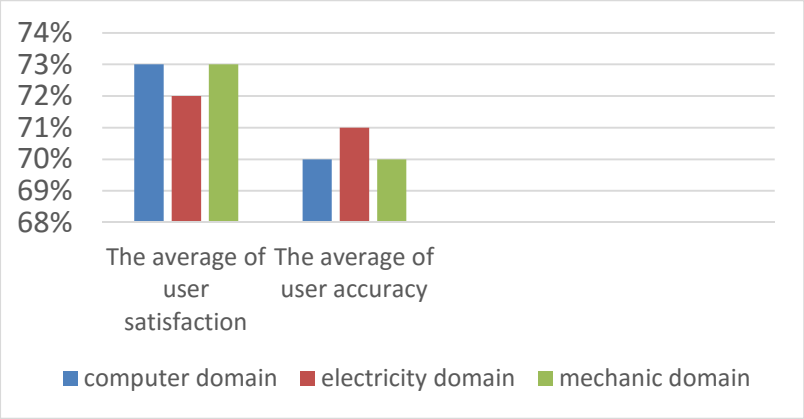

Fig.4. The PhD candidates' results

3. CONCLUSION

In this study was created a profile for each paper including contextual information such as title, abstract, keywords andthe words that identify the domain of each paper.For each userwas created a profile including the user work domain.Finally, the papers will recommend to the users which have more similarity to the user work domain using cosine similarity between the papers domains and the user work domain. In (Hong et al, 2012) study it has been used the keywords to identify the domain but in this study in addition to the keyword it has been used the domain detector words and the words which are listed more than two times in the abstract that cause to recognize the paper domain more accurate and in order to evaluate the system it has been used 5 candidates whose their educational level has not been mentioned.According to the conducted assessments on this system the system performance was satisfactory. Since the total data has been papers from digital IEEE library, for future studies it will be used the IEEE library published books and Amazon website. Moreover, foridentifying the domain it will be used the interesting phrases in the title in addition to the keywords.

4. REFERENCES

[1] Champiri, Z. D., Shahamiri, S. R., & Salim, S. S. B. A systematic review of scholar context-aware recommender systems. *Expert Systems with Applications*, 2015, *42*(3), 1743-1758

[2] Abowd, G. D., Dey, A. K., Brown, P. J., Davies, N., Smith, M., &Steggles, P. (1999, January). Towards a better understanding of context and context-awareness. In *Handheld and ubiquitous computing* (pp. 304-307). Springer Berlin Heidelberg.

[3] Beel, J., Langer, S., Genzmehr, M., Gipp, B., Breitinger, C., &Nürnberger, A. (2013, October). Research paper recommender system evaluation: a quantitative literature survey. In *Proceedings of the International Workshop on Reproducibility and Replication in Recommender Systems Evaluation* (pp. 15-22). ACM

[4] Hong, K., Jeon, H., &Jeon, C. (2012, August). UserProfile-based personalized research paper recommendation system. In *Computing and Networking Technology (ICCNT), 2012 8th International Conference on* (pp. 134-138), IEEE.

[5] Hong, K., Jeon, H., &Jeon, C. Personalized Research Paper Recommendation System using Keyword Extraction Based on User Profile. *Journal of Convergence Information Technology*, 2013, *8*(16), 106-116.

[6] Lakhanpal, S., Gupta, A., & Agrawal, R. (2015, July). Discover trending domains using fusion of supervised machine learning with natural language processing. In *Information Fusion (Fusion), 2015 18th International Conference on* (pp. 893-900), IEEE.

[7] Huang, W., Wu, Z., Liang, C., Mitra, P., & Giles, C. L. (2015, February). A Neural Probabilistic Model for Context Based Citation Recommendation. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*.

[8] Caragea, C., Silvescu, A., Mitra, P., & Giles, C. L. (2013, July). Can't see the forest for the trees?. a citation recommendation system. In *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries* (pp. 111-114). ACM.

[9] Rokach, L., Mitra, P., Kataria, S., Huang, W., & Giles, L. (2013). A supervised learning method for context-aware citation recommendation in a large corpus. *INVITED SPEAKER: Analyzing the Performance of Top-K Retrieval Algorithms*, 2013.

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

[11] http://hadoop.apache.org

[12] http://hadoop.apache.org

How to cite this article:
Jokar N, Honarvar AR, Esfandiari Kh. A contextual information based scholary paper recommender system using big data platform. J. Fundam. Appl. Sci., 2016, *8(2S), 914-924.*