Recommendation System using Keyword Extraction Based on User Profile

[1]Name of 1st Author, [2]Name of 2nd Author, [3]Name of 3rd Author
Radha Sali, Apurva Bhavsar , Mamta Patel

[1]Assistant Professor, [2] Assistant Professor, [3] Assistant Professor,
[1,2,3] School of Computer Science and Engineering,
[1] Sandip University, Nashik, India.

Abstrac : In this paper, we proposed PRPRS (Personalized Research Paper Recommendation System) that designed expansively and implemented a User Profile-based algorithm for extracting keyword by keyword extraction and keyword inference. If the papers don't have keyword section, we consider the title and text as an argument of keyword and execute the algorithm. Then, we create the possible combination from each word of title. We extract the combinations presented in the main text among the longest word combinations which include the same words. If the number of extracted combinations is more than the standard number, we used that combination as keyword. Otherwise, we refer the main text and extract combination as much as standard in order of high Term-Frequency. Whenever collected research papers by topic are selected, a renewal of User Profile increases the frequency of each Domain, Topic and keyword. Each ratio of occurrence is recalculated and reflected on User Profile. PRPRS calculates the similarity between given topic and collected papers by using Cosine Similarity which is used to recommend initial paper for each topic in Information retrieval. We measured satisfaction and accuracy for each system-recommended paper to test and evaluated performances of the suggested system. Finally PRPRS represents high level of satisfaction and accuracy.

IndexTerms - **Recommendation System, Personalization, User Profile.**

I. INTRODUCTION

Due to the development of technology of Internet, Web Programming and Web environment in recent years, the huge amount of data extremely increases in the Web, then following new exceed information search engines are developed and overload problems occur. So, newly high technology made to solve these problems and to provide user-wanted information quickly and accurately. Content retrieval model of Web data which adds Metadata has been studied for a way to search the information effectively. User can quickly and accurately approach the information what they want through this model. Academics and researchers search information, store and share them, finally organize them . The Collaborative filtering-based service is to provide the user-preferred catalog documents which have similar preference by using the profile information instead of using query. In this research paper recommendation system, recent interests focus on increasing accuracy of recommendation . In other words, providing personalized of customized information is wanted and a necessity of new marketing strategy, such as web personalization, one-to-one marketing, and customer relationship management is needed in social practice working area and academic research . An interest in web-based recommendation system is very high because personalized services are important factors to Internet shopping malls and Web service providers . The researches that analyze web using information of users are very useful for recommendation of Web-page as a based technology. The Web-page recommendation technology assesses the web pages that user visited, reflects the assessment results to confidence and uses them for web searching recommendation. The another way is to analyze the keyword that user used with interest and the behavior information with using mouse and keyboard, then screen and recommend the web page to users. Therefore, researches for algorithm of recommendation system and for Information Filtering had been studied, also recommendation system has been used as competition tool in E-commerce business such as Amazon, CD now, and yes24. We believe that suggested PRPRS recommendation system improves accuracy to find the research papers, as well as it can be possibly applied to development of the recommendation system of practical area . This paper makes up for deficits of existing systems which cannot respond to user's profile information sensitively. Then experiments confirm the high level of satisfaction and accuracy with reflecting user's profile information.

II. RELATED WORK

2.1 Recommendation system

Recommendation system is an ‘Information Filtering Technology’ which assists to find the research paper and items what the user wants quickly and accurately. 'Extracts' extracts user's favorite data automatically among a lot of data and processes data. The way to extract information is recorded through the extraction rules. This way stores extraction data in pre-defined templates for processing . A set of data in information extraction appears in the form of the document. Document is an environment for learning the rules of information extraction and extracting wanted information. The method of learning that rules is different depending on the type of the document , also a type of documents is divided into three kinds, as ‘Unstructured documents’, ‘Structured documents’ and ‘Semi-structured documents’

2.3 Personalization

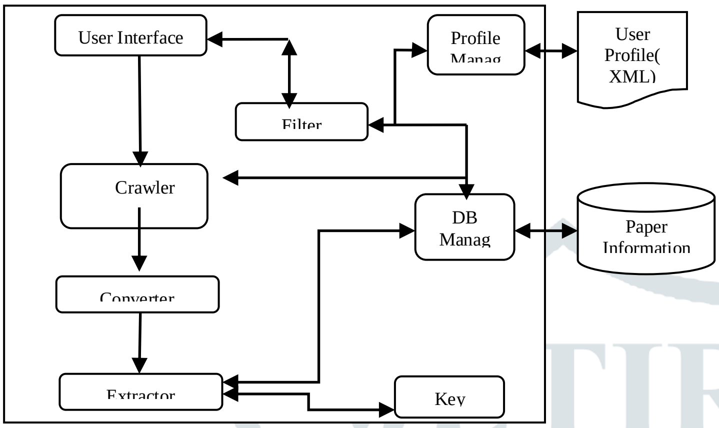
There are Personalization techniques for recommendation system, as ‘Rule-Based Filtering’, ‘Collaborative Filtering’ and ‘Learning agent Filtering’. Rule-Base Filtering is a technique that system asks the question to users about demographic information and personally identifiable information, and then recommends something based on the rule for matched to answer. Collaborative Filtering uses other customer's preferences similar to user's favorite patterns and recommends related services to users. Learning agent keeps track of user's properties, habits, and personal preferences through the analysis of log file such as website history, frequency, access location, time.

2.4 UserProfile

The relevance of information is related to the user's preferences, which are commonly referred to as the user profile. UserProfile known as user modeling is very active field of research in information retrieval. UserProfiles are generally represented as a sets of weighted keywords, semantic networks, or weighted concepts, or association rules. User profiles are constructed from information sources using a variety of construction techniques based on machine learning or information retrieval.

III. SYSTEM ARCHITECTURE

This Figure 1 shows system architecture which is proposed in this paper.



PRPRS(Personalized Research Paper Recommendation System) is composed to UserInterface, Crawler, Converter, Extractor, Filter, KeyFinder, Profile Manager and DB Manager.

As UserInterface interacts with users, makes users input topics and outputs each topic to Crawler, it collects research papers which are adequate to their topics.

Also, PRPRS presents papers collected by topic on topic selected by the user for refining list throughout Filter, provides some information, such as a title, keywords, abstract and so on for paper selected by the user. It reflects the signal information by providing Keyword of the relevant paper to Filter.

Crawler collects huge amount of various research papers related given topic from user, obtains a list of research paper by using Google Scholar(http://scholar.google.com) and stores them in the 'Local' for each paper's URL. In this time, Crawler lets DB Manager to store title, URL, metadata of each paper in the Database, which is for preventing redundancy storage.

Converter transforms each research paper's 'PDF files' which are stored in Local into 'Text form' that the system can handle.

Also, Extractor extracts titles, keywords, abstracts and body of each research paper research changed into 'Text form'. At this moment, if the research paper didn't have any keywords, it could be extracted proper keywords in titles via KeyFinder.

Filter performs filtering process to provide proper articles to users among the stored paper on user's chosen topic via UserProfile which contains personal preference information with XML forms, and provides list of refining papers to users.

UserProfile has been stored the preference information by topic for user's research paper. When users select some article with UserInterface, they reflect user's preference information by weight value of upgrade.

IV. PERSONALIZED RESEARCH PAPER RECOMMENDATION

4.1 Keyword extraction

In this paper, we have designed and implemented algorithm in order to extract keywords section, and Pseudo code is the same as Algorithm 1. Algorithm 1 is to be defined the function which receives an argument. Algorithm starts the preprocessing as remove the tab character or overlap vacancy character to the content of a paper transfer text by Converter and then extracts 'Keywords' section of the paper (line3).

In order to get keywords section from content, we assume as follows:

1) Keywords section exists between introduction and topic.
2) Keywords section begins with 'Keywords', ends '\n' or ''.
3) Keywords section exists in the first page of the paper.

We can get the string composed Keyword by removing unnecessary characters such as ':' or '-' used with reserved word 'Keywords' in 'Keyword' section (line 5). Each Keyword is identified using separator ':' or ',' from the string of Keywords (line 7-11).

Algorithm 1. Keyword extraction

```
1: function KeywordExtraction(content)
2: // preprocessing the content of a paper to get 'Keywords' section of the paper
3: key_content · getKeywordsSection(content);
4: // removing reserved keyword 'keywords' from keyword section
5: key_content · removeKeyword(key_content);
6: // Tokenizing by ';' or ','
```

```
    while exist token do // repeat until there are no more tokens
        key · getToken(key_content); // tokenize from key_content
        Keywords[i] · key; // add key to Keywords at i
        i++; // increase index i
    end while
    return Keywords;
end function
```

There are cases that are not included Keywords like an academic paper or some of conference paper in the research paper. In this case, because we cannot be extracted Keywords, we use them as Keywords after we locate most appropriately representative words. Paper uses the assumption that what is the most appropriately expression of the paper is topic and keywords. That is, when the paper doesn't include keywords, we select the appropriate words among the words which are composed of topic. These assumptions have got through the experiment which performs to the paper with keywords. That is, we have examined the capability of substitution of keywords to the title by measuring the degree of reflection between the title and the keywords to the paper with keywords.

$$\text{Reflex}=\frac{KeywordTitle}{Titleterm} \qquad (4.1)$$

With experiment result, Reflex is shown average value of 0.645; this means that words shown the title can be useful of Keywords. Also, the number of words used the title is an average 3.04, 1.76 words of these number is included the Keywords. Therefore, we used two words for the Keywords on title. If there isn't any keywords section, it utilizes 'Algorithm 2'(keyword selection) to extract the keywords with using the title and contents. Title and contents will be used as factors; therefore algorithm makes possible keyword candidates according to relative distances from each words on title (Line 3). Afterward, It extracts the matched combination showed in the main text among the matched maximum length keywords including same words (line 4). If the size of extracted keywords is lower than 2, use that keyword as it is (line 12). Otherwise, it extracts two keywords which have max term frequency referred to the main body (line 6-10).

Algorithm 2. Keyword selection

```
1: function KeywordSelection(content, title)
2:     // preprocessing the content and title of a paper to get keywords
3:     key_title · getKeywords(title);
       // making keyword candidates depending on the distance from each word
4:     key_title · findMaxMatchedKeywords(key_title);
       // finding the matched max length keywords
5:     if size of key_title > 2 then
6:         while size of Keywords < 2 do // repeat until there are no more tokens
7:             key · findMaxTF(key_title) // find a keyword having max term-frequency
8:             Keywords[j] · key; // add key to Keywords at j
9              j++; // increase index j
10:        end while
11:    else
12:        Keywords · key_title;
13:    end if
14:    return Keywords;
15: end function
```

4.2 Profile-based personalization

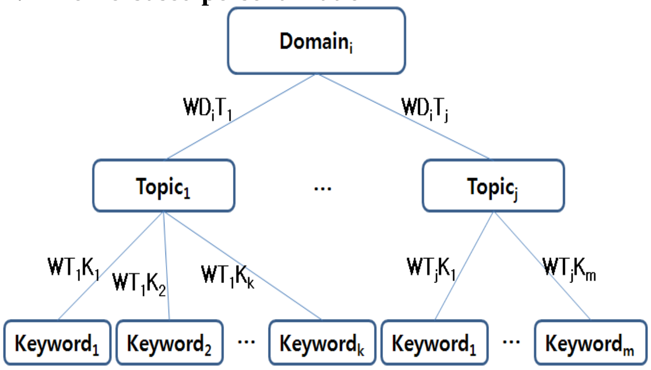
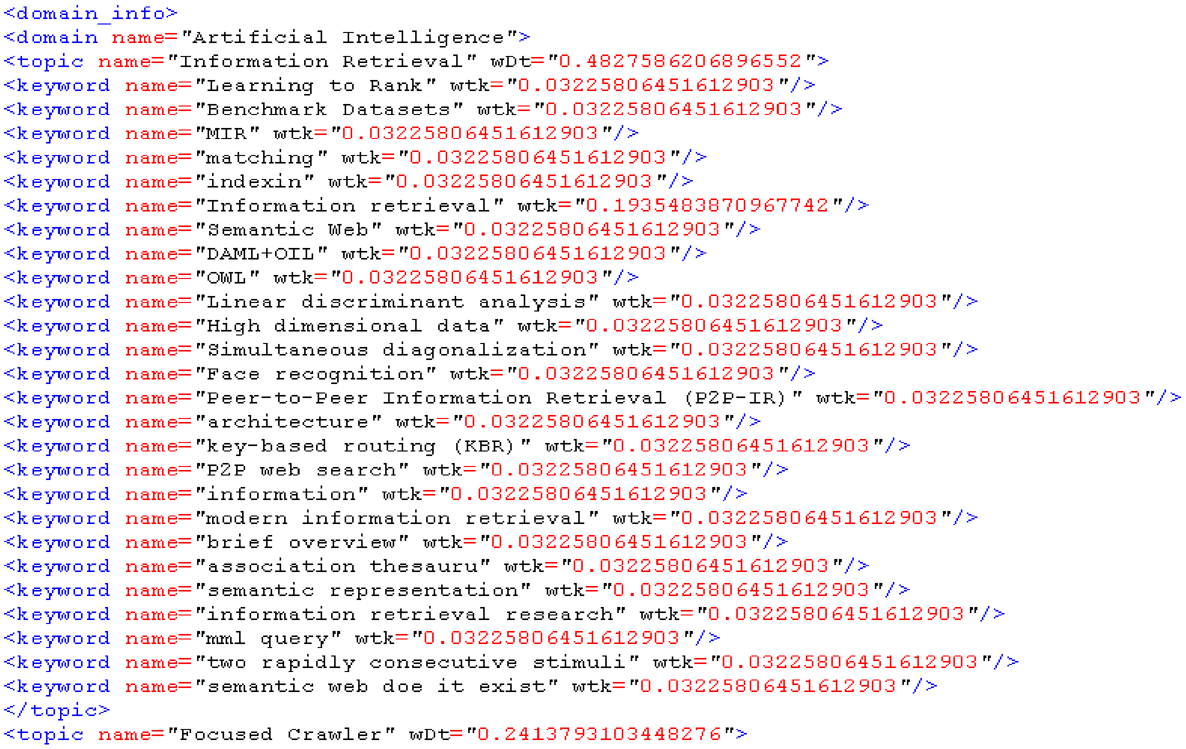


Figure 2. Architecture of User Profile

Where WD_iT_j means the weight of Domain i to the Topic j, and WT_jK_k means the weight of Topic jk to the Keyword.



```
<domain_info>
<domain name="Artificial Intelligence">
<topic name="Information Retrieval" wDt="0.4827586206896552">
<keyword name="Learning to Rank" wtk="0.03225806451612903"/>
<keyword name="Benchmark Datasets" wtk="0.03225806451612903"/>
<keyword name="MIR" wtk="0.03225806451612903"/>
<keyword name="matching" wtk="0.03225806451612903"/>
<keyword name="indexin" wtk="0.03225806451612903"/>
<keyword name="Information retrieval" wtk="0.1935483870967742"/>
<keyword name="Semantic Web" wtk="0.03225806451612903"/>
<keyword name="DAML+OIL" wtk="0.03225806451612903"/>
<keyword name="OWL" wtk="0.03225806451612903"/>
<keyword name="Linear discriminant analysis" wtk="0.03225806451612903"/>
<keyword name="High dimensional data" wtk="0.03225806451612903"/>
<keyword name="Simultaneous diagonalization" wtk="0.03225806451612903"/>
<keyword name="Face recognition" wtk="0.03225806451612903"/>
<keyword name="Peer-to-Peer Information Retrieval (P2P-IR)" wtk="0.03225806451612903"/>
<keyword name="architecture" wtk="0.03225806451612903"/>
<keyword name="key-based routing (KBR)" wtk="0.03225806451612903"/>
<keyword name="P2P web search" wtk="0.03225806451612903"/>
<keyword name="information" wtk="0.03225806451612903"/>
<keyword name="modern information retrieval" wtk="0.03225806451612903"/>
<keyword name="brief overview" wtk="0.03225806451612903"/>
<keyword name="association thesauru" wtk="0.03225806451612903"/>
<keyword name="semantic representation" wtk="0.03225806451612903"/>
<keyword name="information retrieval research" wtk="0.03225806451612903"/>
<keyword name="mml query" wtk="0.03225806451612903"/>
<keyword name="two rapidly consecutive stimuli" wtk="0.03225806451612903"/>
<keyword name="semantic web doe it exist" wtk="0.03225806451612903"/>
</topic>
<topic name="Focused Crawler" wDt="0.2413793103448276">
```

Figure 3. An example of UserProfile

Figure 3 is an example of UserProfile which is used in PRPRS. As shown in Figure, UserProfile is

stored a form of XML, formed hierarchical architecture. It can be traced easily preference information by Domain and Topic of the users, adapt easily with a variation of signal information by forming hierarchical architecture. The update of UserProfile is made for every click of the research paper collected by topic.

Whenever a research paper is selected, it makes increasing of a frequency of Domain, Topic, and

Keyword then, recalculates a rate of each occurrence and reflects to UserProfile. There isn't any addition or delete of the Domain as well as delete of the topic. The addition process will be performed when Domain doesn't have the same topic which users input.

At this point, All WDT would be recalculated. Also, there isn't any delete of Keywords. The addition of Keywords will be performed when each keyword doesn't have the same domain and topic.

All WTK within the same topic would be recalculated according to the addition of Keywords.

PRPRS let upper paper of 50 to recommend by calculating degree of similarity between given topic and collecting paper for utilizing Cosine Similarity generally to be used in order to solve 'Cold Start Problem' which UserProfile-based recommend systems hold[3].

This following formula (1) can get the similarity between the title of 100 top priority articles in research result and given topic to extract the 50 top priority articles which are applied to in this paper.

In this case of having the same similarity, they randomly select.

$$\mathrm{Sim(D,T)}= \sum_{i=1}^{n} w\ di\ \times w\ ti$$

$$\mathrm{Wdi = tfdi}\ \times idf\ \mathrm{di}\ ,\ \mathrm{wti\ = tfti}\ \times \mathrm{idfti}$$

V. EXPERIMENT

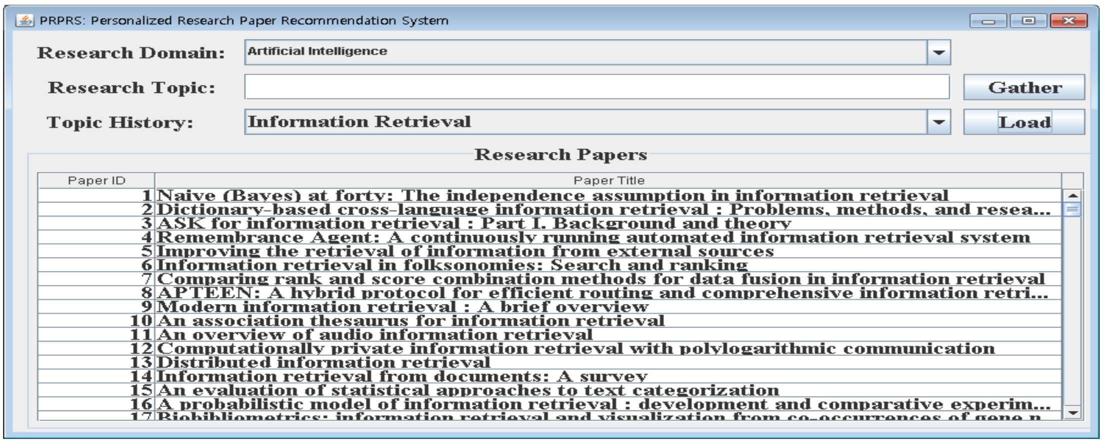
5.1 Environment

We had been performed an experiment to test and evaluate the proposed system's performance with 5 volunteer participants for about a month. Each user chooses the domain related them, produces topics, collects proper research papers for each topic and has recommended refining papers throughout this system. The topics which are generated from each user are 6.8 on the average. We excluded nonresearch papers or results which are written in languages other than English among the result of research throughout Google Scholar. Consequently, the collected paper is 7,236 the whole. The paper which is complied by topic is 658 on the average.

Users recorded possibility of relevance as 'true or false' depending on topical relevance about each paper collected by that system. Accordingly, they estimated accuracy of collecting process. They

measured satisfaction and accuracy about recommendation, as recording how satisfied with every 50 papers recommended by the system about his selected topics.

Figure 4, 5 represent user interface of system, and figure 5 gives us more detail information than figure 4, such as downloadable URL, extracted keywords, abstract as well as title at a glance regarding user's choice paper among papers suggested to users.



Research Domain: Artificial Intelligence

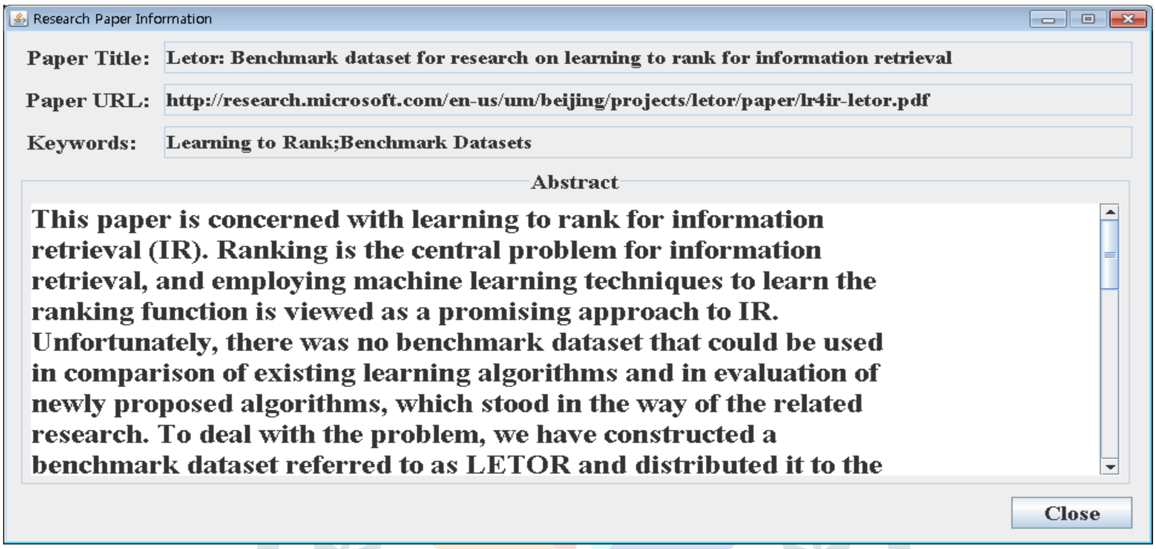
Research Topic:

Topic History: Information Retrieval

Research Papers

Paper ID	Paper Title
1	Naive (Bayes) at forty: The independence assumption in information retrieval
2	Dictionary-based cross-language information retrieval : Problems, methods, and resea...
3	ASK for information retrieval : Part I. Background and theory
4	Remembrance Agent: A continuously running automated information retrieval system
5	Improving the retrieval of information from external sources
6	Information retrieval in folksonomies: Search and ranking
7	Comparing rank and score combination methods for data fusion in information retrieval
8	APTEEN: A hybrid protocol for efficient routing and comprehensive information retri...
9	Modern information retrieval : A brief overview
10	An association thesaurus for information retrieval
11	An overview of audio information retrieval
12	Computationally private information retrieval with polylogarithmic communication
13	Distributed information retrieval
14	Information retrieval from documents: A survey
15	An evaluation of statistical approaches to text categorization
16	A probabilistic model of information retrieval : development and comparative experim...
17	Biobibliometrics: information retrieval and visualization from co-occurrences of gene n

Figure 4. User Interface



Paper Title: Letor: Benchmark dataset for research on learning to rank for information retrieval

Paper URL: http://research.microsoft.com/en-us/um/beijing/projects/letor/paper/lr4ir-letor.pdf

Keywords: Learning to Rank;Benchmark Datasets

Abstract

This paper is concerned with learning to rank for information retrieval (IR). Ranking is the central problem for information retrieval, and employing machine learning techniques to learn the ranking function is viewed as a promising approach to IR. Unfortunately, there was no benchmark dataset that could be used in comparison of existing learning algorithms and in evaluation of newly proposed algorithms, which stood in the way of the related research. To deal with the problem, we have constructed a benchmark dataset referred to as LETOR and distributed it to the

Figure 5. User Interface for more detailed paper information

VI. CONCLUSION AND FUTURE WORK

When academic and researchers wrote their research paper, they had to spend a lot of time and various efforts to find the latest research paper and materials. Therefore, active research paper search and re-search are needed related to specific topic. This paper makes up for deficits of existing system which cannot respond to user's profile information sensitively. Then experiments verify that PRPRS has the high level of satisfaction and accuracy with reflecting user's profile information rapidly.Future work will be implemented about grouping of research papers related to specific subject and active recognition of research trends continuously.

VII. REFERENCES

[1] Jing, J., Helal, A.S., Elamagarmid, A., "Client-Server computing in mobile environments", ACM comp. Surveys, vol.31, no.4, pp.117-157, 2012.

[2] Rey-Long Liu, "Dynamic category filtering profiling for text filtering and classification", ELSEVIER, Information Processing and Management, pp.154-158, 2012.

[3] Kwanghee Hong, Hocheol Jeon, Changho Jeon, "UserProfile-based personalized research paper recommendation system", 2012 8th International Conference On Computing and Networking Technology (ICCNT 2012), vol.8, No.1, pp.134-138, 2012.

[4] Bill Schilit, Norman Adams, and Roy Wand, "Context-aware computing Applications", In Proc, of IEEE Workshop on Mobile Computing Systems and Applications, pp.235-239, 2012.

[5] Manabu Ohta, Toshihiro Hachiki, Atsuhiro Takasu, "Related Paper Recommendation to Support Online-Browsing of Research Papers", IEEE, 2011 Fourth International Conference on the Applications of Digital Information and Web Technologies (ICADIWT), pp.130-136, 2011.