Advanced Personalized Research Paper Recommendation System 
Based on Expanded UserProfile through Semantic Analysis

1 Kwanghee Hong, 2 Hocheol Jeon, 3 Changho Jeon 
1Department of Computer Science & Engineering, Hanyang University, Korea, 
hkhong@cse.hanyang.ac.kr 
2 Agency for Defense Development, Seoul, Korea, hochuls@chollian.net 
3 Department of Computer Science &Engineering, Hanyang University, Korea, 
chj5193@hanyang.ac.kr

Abstract

This paper proposed APRPRS (Advanced Personalized Research Paper Recommendation System) based on UserProfile which is applied keyword expansion through semantic analysis. An algorithm for semantic keyword expansion is designed and implemented. Whenever collected research papers by topic are selected, a renewal of UserProfile increases the frequency of each domain, topic and keyword. Each ratio of occurrence is recalculated and reflected on UserProfile.

Semantic keyword expansion extracts the semantically similar keywords that were used in similar papers among including research paper in same domain and adds the extracted keywords to UserProfile.

We measured satisfaction and accuracy for each system recommended paper to test and evaluate performances of the suggested keyword expansion method and system. Finally, there are performance improvements about 9% when semantic keyword expansion is applied. As a result of experiment, the suggested system represents high level of satisfaction and accuracy.

Keywords: Semantic Keyword Expansion, Personalization, Recommendation System

1. Introduction

Many researchers collect a lot of similar theme research papers to write high quality papers using search engines such as Google. They emphasize that their basic idea or method is what kinds of differences have and how this system is efficient and effective compared to other previous announced methods by comparing and analyzing similar research papers.

Especially, the process of collecting and analyzing the similar title research papers requires much efforts and time to researchers.

In this paper, we proposed research paper recommendation system based on semantic keyword expansion that will help researchers to write research papers by reducing the time and efforts.

When the user clicks the recommended research paper, UserProfile has stored keywords of that research paper. It improved accuracy and satisfaction of recommended papers which given to user by expanding semantically similar keywords in the same domain with taking advantage of WordNet.

2. Related work

2.1. Personalized recommendation based on UserProfile

The relevance of information is related to the user’s preferences, which are commonly referred to as the user profile. UserProfiles known as user modeling are a very active field of research in information
retrieval [25]. UserProfiles are generally represented as sets of weighted keywords, semantic networks, or weighted concepts, or association rules. UserProfiles are constructed from information sources using a variety of construction techniques based on machine learning or information retrieval [26].

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 [20]. 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 [22].

Recommendation system is an ‘Information Filtering Technology’ which assists to find the research paper and items what the user wants quickly and accurately [11], [18]. ‘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 [14]. 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 [19].

2.2. Semantic keyword expansion

Query expansion is to search the research paper which has high relevance by adding the words related to proposed query from user and searching. Recently, the research about query expansion has two methods. One is using personal or group profile, and the other is using the feedback of user.

If the user enters the query, the expanding method that makes profile refers the previous written profile with keyword related to user interest area. Then it expands the query by referring them. In the other expanding method by feedback of user, after the users see the searching results, it revises the question and find repeatedly.

We construct the Researcher Profile based on evaluation about paper that researcher read previously. Also we construct the semantic expansion of word that the semantic relationship is applied to [17]. We recommend some papers to user by measuring the similarity between the researcher’s profile applied the constructed semantic expansion and recommending candidate research paper.

3. System architecture

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

![Figure 1. System architecture of APRPRS](image)

---

68
Advanced PRPRS is consisted of UserInterface, Semantic Analysis Manager, Crawler, Converting Manager, Extraction Manager, KeyFind Manager, Filtering Manager and UserProfile Manager.

First of all, UserInterface interacts with users directly, analyses the meaning utilizing syntax analysis and WordNet through Semantic Analysis Manager by obtaining topics from users. Above all, APRPRS uses this information to find similar topics and to extract candidate keywords.

Crawler collects research papers by using extracted similar topic as a question of Google scholar from Semantic Analysis Manager as well as entered topic from users.

A Converting Manager is given some collected research paper, also each collected PDF document will be transferred to text that is only accessible type within system through PDFToTXT module. Because the transferred text files don’t cover any images, but only include headers, footers and annotation, they are required to filter.

The unnecessary parts of document interrupt recognizing and extracting the exact sentence as well as the parts in text. That causes degradation of performances in terms of accuracy and efficiency of system. Thus, this system needs process to delete the unnecessary parts like annotation. Extraction Manager will perform this process.

Another function of Extraction Manager is partially extracting and recognizing the title, author, abstract, keyword, body and reference, etc. By this algorithm, it is possible to produce the index information about relevant paper automatically.

KeyFind Manager does function as extracting keyword applicable to UserProfile. If it cannot extract any keyword in main context, it utilizes analogized keyword in title and used keyword in resembling papers.

Database Manager performs managing function of index information about collected research papers. UserProfile Manager handles UserProfile which stores preference information about user’s research paper topic. Then, whenever users click each suggested research paper as a result of research, it will be renewed continuously.

Filtering Manager performs by using UserProfile to provide adequate papers among saved research papers for user selected titles, also give the refined list of papers to user via UserInterface.

4. Personalized research paper recommendation

4.1. Keyword extraction and expansion

In this paper, APRPRS extracts or infers keywords whether ‘keywords’ section is present or not. That can be confirmed through extraction algorithm used in related research [3].

If there is any ‘keywords’ section on that paper, keyword will be extracted with the utilization of previously defined rules and extraction algorithm in related research [3]. Otherwise, keyword would be selected better suitable word among words used in Title, and then uses it as keyword.

That is, TF(Term Frequency) in the body of this paper uses 2 words as keyword for each used word in the paper. At this point, words include the compound noun and select occurring words in the body which coincide with title. Every research paper’s keyword represents that article very properly, so this paper only considers semantic based keyword expansion.

In the beginning, keyword expansion starts with searching of similar research paper and measuring similarity between papers in sequence. Similarity between papers is measured by Cosine similarity, which used to be performed on information retrieval.

For this measure, Cosine similarity performed the basic indexing process about each paper and constructed inverted index file on DB.

\[
\text{Sim}(d_i, d_j) = \frac{d_i \cdot d_j}{\|d_i\| \cdot \|d_j\|} = \frac{1}{\sqrt{w_{d_i}} \sqrt{w_{d_j}}} \sum_{t} w_{d_i, t} w_{d_j, t}
\]

\[
= \frac{1}{w_{d_i} w_{d_j}} \sum_{t \in d_i} g_{d_t} \cdot \left( \log \frac{N}{df_t} \right)^2
\]  

(1)
When the similarity between papers represents more than a critical value \((\alpha = 0.64)\), we extract the keyword from paper, then utilize that for keyword expansion. Critical value is decided by experiment that determines the correlation of similarity between papers and similarity between the keywords which are used in two papers as shown in Figure 3. The similarity between the used keywords was calculated by equation (2), (3).

\[
\text{KeySim}(D_i, D_j) = \frac{\sum \max(\text{WordMatch}(d_i, k_n, d_j, k_m))}{D_N_i}
\]  

(2)

\[
\text{WordMatch}(d_i, k_n, d_j, k_m) = \sum_{\rho} \text{WordNetSim}(w_{\rho}, w_{\alpha})
\]  

(3)

At this point, \(D_N_i\) means the number of keywords in paper \(i\), also \(d_i, k_m\) stands for \(m\)-th keyword of paper \(i\). \(D_j, k_n\) implies \(n\)-th keyword of paper \(j\). Moreover, \(\text{WordMatch}\) signifies the similarity between two words by using WordNet.

If a length of keyword is 1 character, WordNet can be applied directly. However when a length of keyword is more than 2 characters, WordNet will be applied for each word in keyword, the sum of keywords will be used as a value of \(\text{WordMatch}\) because of difficulty to apply directly.

We check keyword similarity between keyword \(d_i, k_m\) and all keywords in paper \(d_j\), and then take the biggest value in measured values, as a denominator of equation (2).

![Figure 2. Keyword similarity according to the similarities between papers](image)

Because keywords compose the words that express the paper representatively and implicitly in research paper, applying keyword to other paper beside relevant paper would be inappropriate.

Considering these keyword’s characteristics, we take the used keyword in similar paper as a target of keyword expansion. We judge the availability whether to use for keyword expansion by measuring semantic similarity of each keyword in pseudo paper.

The semantic similarity between keywords is measured by Figure 3. When the value is more than 0.68, we judge that the keyword has semantic similarity to use as keyword expansion. If the keyword are not registered in WordNet, so can’t measure any similarity, 1 character word cannot be used to keyword expansion. In case of parts of more than 2 characters keyword are not registered, we treat the relevant similarity value as zero.

The following algorithm 1 explains performing process of keyword expansion.
Algorithm 1: Keyword Expansion

1: function KeywordExpansion(keywords[])
2:     SimPapers[] ← getSimilarPapers(); // extract similar research papers with the given research paper
3:     while i < length of keywords do // repeat until i is greater than or equal to length of keywords
4:         while j < length of SimPapers do // repeat until j is greater than or equal to length of SimPapers
5:             sim_keywords[] ← SimPapers[j].getKeywords(); // extract keywords from a similar research paper
6:             while k < length of sim_keywords do // repeat until k is greater than or equal to length of sim_keywords
7:                 word_match_value ← getWMV (keywords[i], sim_keywords[k]); // calculate keyword similarity
8:                 if word_match_value > 0.68 then
9:                     ExKeywords[m] ← sim_keywords[j]; // add conditionally accepted keyword to expanded keyword list
10:                    m++; // increase index m
11:                 end if
12:             end while
13:         end while
14:     end while
15:     return ExKeywords;
16: end function

Keyword expansion starts with receiving keywords list of relevant paper as a factor. Algorithm 1 measures Cosine Similarity for given paper and other papers which in same domain. It extracts research papers showing greater than 0.7 similarities (Line2). After extracting conditionally accepted keyword (Line 3-4), it checks the similarity between each keywords which are transferred as factor (Line 5-6). If that similarity is greater than word match value, add the keyword to expanded keyword list (Line 7-8).

Moreover, it makes available to keyword expansion of relevant paper as well as similar papers by adding transferred keyword as factors to similar papers. These mutual save considers whole situation when keyword expansion may not be possible due to nonexistence of similar paper.

4.2. Expanded UserProfile-based personalization

The similar keywords that are obtained from keyword expansion process would be applied to UserProfile together. UserProfile that applied to related research is utilized [3]. The structure of UserProfile is as below Figure 3.
At this point, $WD_{i,j}$ is a weighted value of $j$-th Topic in $i$-th Domain. $WT_{j,k}$ means a weighted value of $k$-th keyword of $j$-th Topic. Each $EK$ is the added keyword through keyword expansion, it will be treated equally. When users click the collected paper by topic with UserInterface, UserProfile will be renewed.

5. Experiment

5.1. Environment

We had been performed an experiment to test and evaluate the expansion of UserProfile through keyword expansion proposed this paper and performance of research paper recommendation method based on proposed system with 5 volunteer participants for about a month.

Each user chooses the domain related them, produces topics wanted to collect, collects research papers for each topic through UserInterface from time to time and has been recommended refined papers based on UserProfile. The topics which are generated from each user are 5.9 on the average. The collected paper is 5,696 the whole. The paper which is compiled by topic is about 781 on the average.

Users recorded the possibility of satisfaction as ‘true or false’ about each paper collected by that system. They measured satisfaction and accuracy as recording how they are satisfied with every 50 papers recommended by the system about his selected topics. To confirm the performance of proposed method in this paper, we compared this paper with previous research and measured satisfaction and accuracy by utilizing the same method used in previous research.

Figure 4, 5 represent UserInterface that contains more various and concrete information regarding each UserInterface and research papers.
5.2. Evaluation

We defined the accuracy of paper collection, the satisfaction of recommendation and accuracy to measuring the performance of proposed method on this paper.

\[
SAT = \frac{\# Satisfied \ Research \ Paper}{\# \ Recommended \ Research \ Paper} \tag{4}
\]

\[
ACC = \frac{\# \ Recommended \ Sat + \ # \ Not \ Recommended \ UnSat}{Total \ Saved \ Research \ Paper \ for \ a \ given \ Topic} \tag{5}
\]

SAT means in a ratio of the number of recommended research paper for users to the number of satisfied research paper for a given topic, and ACC means in a ratio of total number of saved research paper to recommended satisfied paper plus not-recommended paper unsatisfied paper among the collected research papers for a given topic. Then, for measuring SAT, we restrict recommended paper...
numbers maximum as fifty. But for measuring the accuracy, we don’t have any limits how many papers recommended to users; all of the research papers are to be subject of recommendation.

Figure 6, 7 represent the average satisfaction and accuracy of research papers recommended to users by week with (APRPRS) and without Advance Personalized Research Paper Recommendation System (PRPRS). As shown in Figure, it represents rather high level of Satisfaction.

We can see that the adaptive speed of satisfaction almost represented similarly, while the accuracy adapted a little rapidly as shown the steep slope, when keyword expansion is used.

![Figure 6](image_url)

**Figure 6.** The average of recommended satisfaction for user weekly

![Figure 7](image_url)

**Figure 7.** The average of recommended accuracy for user weekly

Figure 8 represents the entire average of satisfaction and accuracy to all users on last week (fourth week) of experiment with and without using keywords expansion. As shown in Figure 8, when it was
used the keywords expansion, satisfaction and accuracy represent very high level for each 0.894, 0.851. This represented the upgrade of performance with about 9.3% and about 9.5% than without using keywords expansion.

![Average satisfaction and accuracy of user](image)

**Figure 8.** Average satisfaction and accuracy of user

6. Conclusion and future work

In this paper, we have proposed the semantic similar keyword expansion method through the similar paper, measured the performance of proposed method through our experiments.

In case of recommendation of research papers, expansion of UserProfile through keyword expansion can be represented very high satisfaction and accuracy through various experiments. However, in case of keywords expansion, because it doesn’t have any limits of the numbers, it has the possibility of limitless expansion. Accordingly, in the future, we will research the methods that prevent the performance degradation of system and manage efficiently.

7. References


