Association for Information Systems
AIS Electronic Library (AISeL)

PACIS 2004 Proceedings

Pacific Asia Conference on Information Systems (PACIS)

December 2004

# CBR-based Recommender Systems for Research Topic Finding

Deng Liu
*Fudan University*

Limin Lin
*Fudan University*

Jie Lu
*University of Technology*

# CBR-based Recommender Systems for Research Topic Finding

Deng Liu
Fudan University, Shanghai,
P.R.China, 200433
022025086@fudan.edu.cn

Limin Lin
Fudan University, Shanghai,
P.R.China, 200433
lmlin@fuan.edu.cn

Jie Lu
Faculty of IT, University of Technology,
Sydney, Australia
Jielu@it.uts.edu.au

**Abstract**

M*ore and more universities and research institutes are taking scientific research as one of the main objectives. Finding a valuable research topic is an important activity that scientific researchers are more concerned about. But the existing Recommender Systems for Research Topic Finding that only provide simply document mining don't satisfy the need of the researchers. This paper explores a new type of Recommender Systems for Research Topic Finding based on the CBR technology. We propose a framework of CBR-based Recommender Systems for Research Topic Finding (CRSRTF). Furthermore, the edit distance is used to solve a basic problem — knowledge representation and rule mining of the senior researchers' experience.*

**Keywords:** Research topic finding, Case-based reasoning (CBR), the edit distance, Recommender systems

## 1. Introduction

A good research topic is very important for both researchers and research organizations, since it implies the possibility of quality and quantity of research products, as well as distinguished leaderships in academic society. Providing supports to researchers on finding good research topics is currently one of the main tasks taken by research organizations. The support is also shifting from an 'indirect' way, which provides researchers with help on funds and hardwires, to a 'direct' way, which provides researchers with helps on their research topic finding process. For example, most of the past supports focused on financial and substantial aspect, including various organizational or governmental research grants, and electronic or online library systems. These kinds of supports help researchers 'indirectly' without going to the point of how good research topics being developed. Nowadays they are shifting to a more 'direct' way to 'the inside of' the topic finding process. For example Research and Development Office's Grants Development Teams are founded in many research organizations targeted at providing assistance, advice and feedback to applicants for the whole process of application writing. In particular, as a 'direct' support, more and more efforts are put on collecting and making full use of the experience of senior researchers who have been successful in previous grants applications.

Studies on research topics finding have started since 2001 (Schwartz 2001, Ramamonjisoa 2001, Ramamonjisoa 2003). However, these studies were conducted as a sub topic of 'Topic Detection and Tracking' (TDT) and therefore took a similar research methodology as that of TDT. TDT is a research project sponsored by The Defense Advanced Research Projects

Agency (DARPA) since 1997. It aims to 'identify event-based topics and follow them across multilingual incoming streams of broadcast news and newswire documents'. The methods related to TDT research can be viewed as specializations of the broader concept for data mining, which deal with the extraction of knowledge from text documents, and are involved with natural language processing and information retrieval techniques in addition to the standard data mining and machine learning methods. In the paper by Ramamonjisoa(2003), he calculates the frequency of the keywords of the known topics and the unknown topics and applies a data mining method to estimating maximum relevance between them to identify a possible new research topic. From this example, we can see that for an unknown topic to be a possible research topic, it must have been discussed explicitly in existing documents and must be in a clear relationship with certain existent topic(s).

However, finding research topics should be more than 'mining' topics from existing documents by analyzing the associate relationships between keywords. Furthermore, research topics have different features from news topics. In particular, a new research topic means that it should be mostly new and seldom discussed in existing documents. However, the existing mining method could only identify the relationships among topics already discussed in existing documents. It is not enough for new topics recommendation.

Therefore, recommender systems for research topic finding should have the intelligence of learning. The research topic finding process is a complicated process of human mind. Data should not only refer to documents, but also to the existing research topics, the requirements from industry, the new research results from adjacent research fields and the ideas from discussions in conference meetings. The various rules used for data processing should not be confined to data mining, besides it should contain the experience of senior researchers, even their intuition, which should be paid more attention to and learned. Moreover both rules and data are subject to continuous changing as long as researchers keep on developing new topics with new experience. Recommender systems should be able to support the extension and adjustment of both rules and data.

Recommender systems with learning ability have been researched by Rashid, et al. (2002). However, Rashid's research only focuses on the learning of 'new user preference' that is derived from the mining of the existing user data. It is not about the learning of various rules. Researches on adding new rules into Recommender systems have not been found in current literature.

Since there are not any proper Recommender systems to find research topics, or good mining methods of representing the researchers' experience. This paper focuses on developing the Recommender Systems that can provide researchers with excellence topics. The CBR technology is used to support the Recommender Systems. The systems not only have the learning ability of various rules, such as experience and professional intuition of senior researchers, but also deal with various kinds of data, including documents, conference records, professional knowledge and so on. For this purpose, we firstly propose the framework of the CBR-based Recommender Systems for Research Topic Finding (CRSRTF), which has two significant features, one is the learning abilities of new cases and rules, and the other is the rules describing all various kinds of data processing. Furthermore, a method of the edit distance is applied in knowledge representation and rule mining. It not only can extract the cases excellently from the data, but can find out rules that recall the connotative thinking process of senior researchers. It plays a very important role in initiating a new research topic.

The paper is organized as follows: in section 2 we briefly introduce the framework of CRSRTF. In section 3 we describe the edit distance. In section 4 we describe how to represent knowledge and reason rule by the edit distance. The paper ends with a discussion of the results, and the planned future work.

## 2. CBR- based Recommender Systems for Research Topic Finding (CRSRTF).

CBR is firstly proposed in 1977 by Schank who claimes that 'a memory system that fails to learn from its experience is unlikely to be very usefully' (Riesbeck and Schank 1989). The learning ability of CBR can be shown by it's '4Rs' cycle reasoning process (Fig.1). In the first 'R' stage, when a user inputs a new decision problem, CBR **retrieves** the most similar case from the case base. In the second 'R' stage, the solution of the retrieved case is **reused**. In the third 'R' stage, the solution is **revised** to suit the new problem. And in the fourth 'R' stage, the problem and the revised solution are **retained** for the future reuse. The recommender systems for research topic finding could be enabled with learning ability by

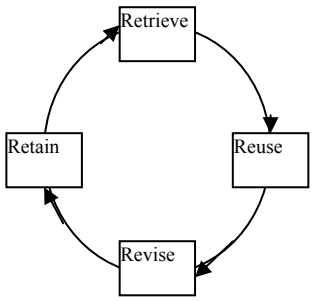


Fig. 1 CBR reasoning process: '4Rs'cycle

applying CBR technique.

Usually the representation of cases in CBR has two parts, the problem space and the solution space. The problem space is composed of a set of features of the problem description part, and the solution space is composed of a set of features of the solution description part. During the retrieve stage, CBR compares the feature values of the problem space between source cases and a new case in order to find the most similar case from the stored cases. However, for some applications where case features can not be clearly described in quantity, the fuzzy method can handle both case representation and case retrieval.

The framework of CRSRTF has been designed as shown in Fig. 2. In the framework, there are three bases and four modules. The three bases are data base, rule base, and case base. The case base contains two parts: one is the representation of the feature of the problem space; the other is that of the solution space. In the system, the features of the problem space and the solution space can be divided into several parts according to the nature of the case, such as mentioned above, the background knowledge of specialty, the newest research achievement about this field，the hot current situation of the correlative fields and so on. When the attributes are drawn out from the case, the fuzzy method is applied to describe the feature of the case because of abstract and non-structural features of the case itself. The fuzzy method is easy to describe the feature of the case, and it is useful to case retrieval. The rule that represents the optimal reflection from the problem space to the solution space is put in the rule base. The rule and the feature attribute of problem space and solution space together describe the process that researchers think over the research topics. The data base is composed of some information which is used in the process of finding research topics. The

data in date base are corresponding with the feature of problem space and rules. They may be the survey about the papers in this field, as well as the reports about industry, and some articles of related field. The four modules are data base maintaining module, rule base maintaining module, case base maintaining module, and research topic finding module. Data base maintaining module, rule base maintaining module and case base maintaining module work as interfaces to the data base, the rule base and the case base respectively. Research topic finding module works on three bases and the other three modules when implementing CBR reasoning process to produce topics recommendation to users. The reasoning process and the data flow have been described in Fig 3.

**The Retrieval Step** The features of the new problem are compared with the features of the existing cases in the case base using the fuzzy retrieval technology in order to find the most similar cases. It can be achieved by following steps. At first, the domain of the new topic is compared with the domain of existing topics in the case base. If they are similar, they are reserved. For each part feature the new case is compared with preserving cases according to the features of each part. Finally, the set of similar cases is ranked according to the degrees of similarity. The output of the retrieval step is a collection of candidate cases for each part of the feature description.

**The Reuse Step** Firstly rule is gotten in keeping with cases retrieved in the former step. For each part of the features the rule of the most similar case should be chose. Then the rule is applied to the corresponding date in date base. At last a new topic is presented as a preliminary recommendation topic.

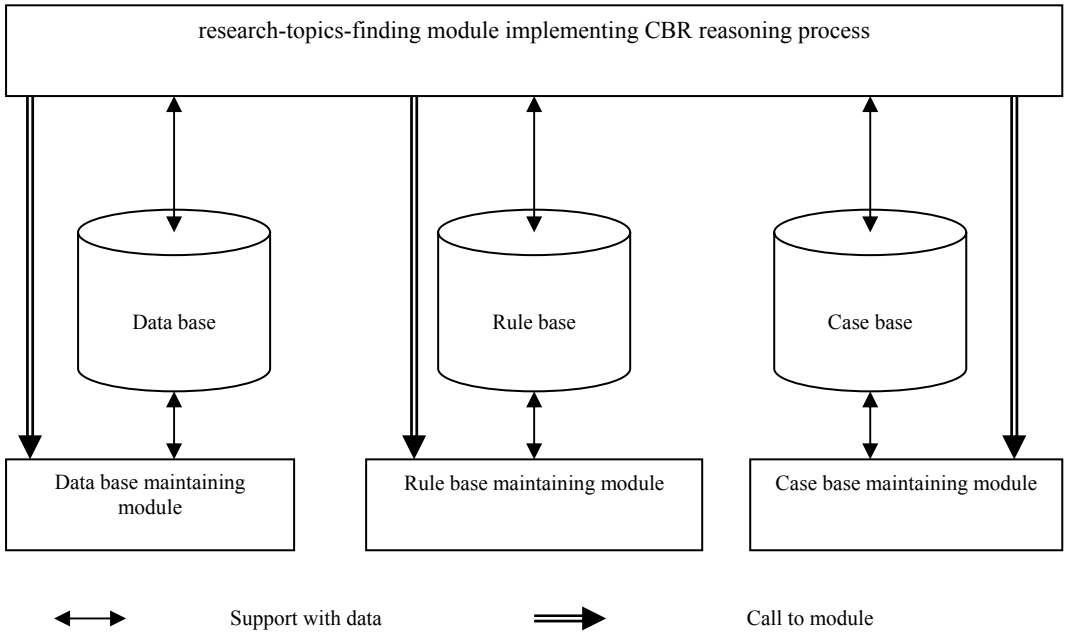


Fig. 2 CBR-based recommender system for research topic finding (CRSRTF) framework

**The Revise Step** In this stage the proposed solution which comes from the former step is adjusted to fit case representation structure of rule better, according to the needs of the users.

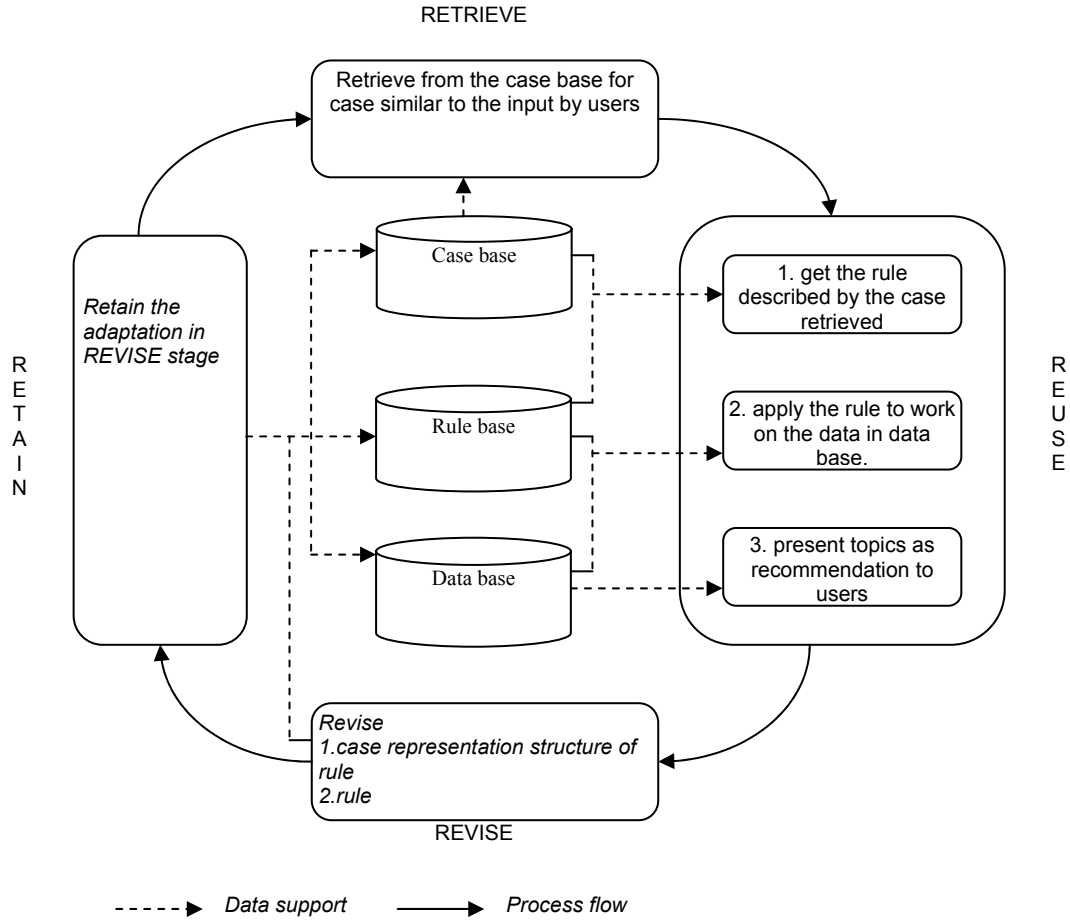


*Fig. 3 the CBR reasoning process for the CRSRTF*

**The Retain Step** After an excellent research topic is found, it should be added to the memory of cases. Meanwhile the rule base and the date base should be updated accordingly. The newly added case will be available for the reasoning in future problems.

By applying CBR technology, the Recommender systems could learn from experience of senior researchers during the reasoning process. The rule set could be extended to include more than just mining topics from documents, and the data base could be extended to include other kinds of data besides documents.

## 3. the Edit Distance Enabling Knowledge Representation

After the framework of CRSRTF is proposed, it is crucial to find a proper way to construct the cases and reflect the experience of the senior researchers. The edit distance approach is used in knowledge representation and rule mining.

### *3.1 the Edit Distance Approach*

The concept of the edit distance has a well established history dating back to the 60s and 70s (Gilbert and Schroeder 2000). It has played an important role in a wide array of applications due to its representational efficacy and computational efficiency (Wei 2004), and is still widely used in bioinformatics to comparing genomic (Gilbert and Schroeder 2000), parsing theory (Pighizzini 2001), molecular biology (Karp 1993), image and signal processing, error correction, information retrieval, and pattern recognition and pattern matching in large databases (Arslan and Egecioglu 1999). The edit distance between two sequence, which is also called Levenshtein distance, can be defined as the minimum total cost of converting one

sequence (the source sequence) into the other (the target sequence), given a set of allowed edit operations and a cost function giving out the cost of each edit operation. The sequences under comparison are not restricted to quantitative date, and they do not even have to be of the same nature, since the edit operations and their costs can be designed to handle any kind of data (Arcos and Grachten 2003).

The edit distance generally contains three operations: insertion, deletion, and substitution. Each operation has a non-negative real value representing the cost of this operation. Insertion is the addition of an element at some point in the target sequence; deletion refers to the operation that removes an element from the source sequence; substitution means replacing an element in the source sequence by one in the target sequence. Although the effect of the substitution is equivalent to that of deleting an element in the source sequence then adding an element in the target sequence, the substitution operation emphasizes on that the source and target elements somehow correspond to each other. Furthermore, it has lower cost than the sum of the cost of deleting the source element and inserting the target element, when a source element and a target element are thought to be relevant. Certainly, many other operations can be added, according to the nature of the sequence. In the paper, in order to suit researchers' thought process of topic finding, the set of operations is composed of following four operations: insertion, deletion, substitution and combination. Combination refers to combining several elements in the source sequence into an element of the target sequence.

After the operations are decided, the following thing is to define of the cost for each edit operation. Different problem could have different cost functions depending on the nature of the problem to be solved. In this paper, the costs of the edit operations are defined as the function about similarity between the corresponding elements of two sequences. The special meaning about them is explicated in the following section.

### *3.2 Computing the Distance*

The edit distance can be computing easily by using the dynamic programming (Pighizzini 2001, Gilbert and Schroeder 2000). We firstly suppose two sequence $A<a_1, a_2, a_3 \cdots a_m>$ as the source sequence and $B<b_1, b_2, b_3 \cdots b_n>$ as the target sequence. The edit distance $d_{i,j}$ is defined as follows:

$$d_{i,j}=\min \begin{cases} d_{i,j-1}+w(b_j) & \text{(insertion)} \\ d_{i-1,j}+w(a_i) & \text{(deletion)} \\ d_{i-1},j-1+w(a_i,b_j) & \text{(substitution)} \\ d_{i-k,j-1}+w(a_{i-k+1},\cdots,a_i, b_j) & \text{(combination)} \end{cases}$$

For all $0 \leqslant i \leqslant m$, $2 \leqslant k \leqslant i$ and $0 \leqslant j \leqslant n$ where m and n are the length of the source sequence and the target sequence respectively. In addition, the initial conditions are:

$$\begin{aligned} &d_{0,j} = d_{0,j-1} + w(b_j) && \text{(insertion)} \\ &d_{i,0} = d_{i-1,0} + w(a_i) && \text{(deletion)} \\ &d_{0,0} = 0 \end{aligned}$$

In the formula, $d_{i,j}$ is the edit distance between $A_i < a_1,\cdots,a_i >$ and $B_j < b_1,\cdots,b_j >$. Ws are the cost functions representing the cost of the corresponding operation. $W(b_j)$ corresponds to the cost resulted from an insertion of $b_j$, $w(a_i)$ is the cost incurred by a deletion of $a_i$, $w(a_i,b_j)$ is the cost associated with the change form $a_i$ to $b_j$, and $w(a_{i-k+1},\cdots,a_i, b_j)$ returns the cost of combining $a_{i-k+1},\cdots,a_i$ into $b_j$.

### *3.3 Optimal Alignments*

After calculating the distance value between two sequences, the optimal alignment is found out, which makes the source sequence transform into the target sequence and also yields this value. It contains some operations mentioned before. In order to get the optimal alignments, it is necessary to store every $d_{i,j}(0\leqslant i\leqslant m, 0\leqslant j\leqslant n)$, because $d_{i,j}$ is the minimum cost by the last operation that was performed to arrive at $d_{i,j}$. From this operation it can be inferred what the last step is, which makes the source sequence $A_i$ transform the target sequence $B_j$. For example, perhaps $b_j$ is added, or $b_j$ is substituted for $a_i$. It is the matrix of m rows and n lines. When tracing back the matrix in this way from $d_{m,n}$ to $d_{0,0}$, the alignment of operations can be found that matches this value and completes this transformation. A corresponding relation is set up between the elements in the source sequence and those in the target sequence. The alignment is the rule that is been looking for. It indicates the thought process of the researchers' looking for a good topic and the tacit knowledge in the researchers' brain, which is essential to the research topic finding. It is crucial to the new topic finding in the Reuse step.

An example of the alignment is shown in figure 4. Some geometric figures are used to replace the elements in order to simplify the problem. The alignment is presented between two sequences. The letters stand for insertion (I), deletion (D), substitution (S) and combination(C), respectively. For instance in figure 4, the first element of the source sequence is the same as the first element of the target sequence, so an operation of substitution is to execute it. The fourth and the fifth elements in the source sequence are two right-angled triangles, but the relevant third element in the target sequence is the triangle that is made up of them, therefore the operation is combination. Using this method, the optimal alignment could be found: {SDSCI} . It would make the least cost in the whole, which transforms the source sequence to the target sequence.

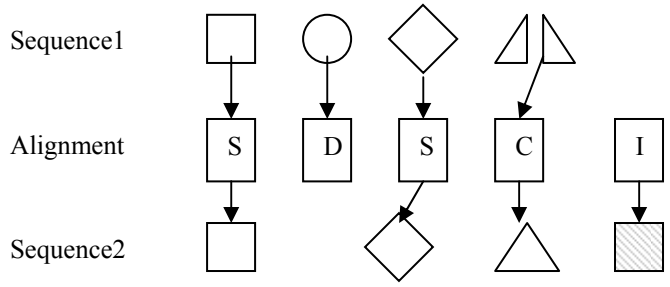


Figure 4. A possible alignment between two sequences

## 4. Case Representation and Rule Mining

Before researchers choose a topic to study, something based on the current condition must be considered. How is the topic related to researchers' professional background? What is the hot

problem in current research field? What can be provided by the university or the institute? In addition, it is necessary to review the papers nowadays in the field. All kinds of factors must be thought over and over again. Finally, a perfect research topic is gained. The process of looking for a research topic is so valuable, that we hope to use the CBR technology and the edit distance to express it to help research topic finding in the future.

For this purpose, the conformation of the case is very important. It contains two sides. One is the conformation of the question space (the source sequence); the other is the conformation of the solution space (the target sequence). The conformation of the question space could be set up according to the factors mentioned before, using the CBR technology and the fuzzy method. Firstly, the problem space is divided into the following several parts: the knowledge of the professional field, the current hot research, the background of applications, the review of studies, and the technology and support provided. Then every part is decomposed into several attributes. The feature value of each attribute is drawn out from the collected resources by fuzzy technology. The conformation of the solution space could be planed by the same method. The solution space is separated into the following fields: the keywords of the research topic, the main study content, the technology used, the study method, and the theoretical base. The feature value of each attribute is confirmed by the records of the project. Certainly, the parts of the problem space and the solution space could be increased and changed freely according to the nature of the problem, only keeping small relativity among every part and covering the whole study content. The number of attributes is not limited yet.

If an excellent rule is able to be found, which could represent the researchers' performance, a crucial problem is to identify which element(s) of the source sequence corresponds to each element of the target sequence. Especially during the process of the topic finding, the performance of researchers is profound and complicated, which is too difficult to describe it with an easy rule. Researchers deviate from the existing technology to solve the same problem. Hence, it is normally not assumed that the problem space contains all parts of the solution space. neither every element of the problem space can be found in the solution space. From this perspective, the edit distance, as described in the previous section, will be very useful.

### *4.1 the Edit Operation*

All edit operations could be classified to make clear the features of their behavior. One kind is the Reference operations in that the elements operated appear in only sequence, such as Insertion and Deletion. The other kind is the Correspondence operations which refer to the operated elements from both sequence, such as substitution and combination. In the case, we are not primarily interested in the order of the element of the sequence, but in the changes between the source sequence and the target sequence.

In the attribute set of the solution space, some attributes from the problem space still stay there. This is enough only by the substitution operation. Besides these attributes, it also contains the extra attributes, or lacks some attributes comparing with the attributes of the problem space. This implies that besides the substitution operation, the insertion operation and deletion operation are also wanted. Furthermore, it is common that combining several attributes in the problem space together becomes a new attribute in the solution space. For instance, researchers combine the fuzzy technology with CBR together and gain a new CBR technology—fuzzy-CBR. Therefore, the set of operations mentioned in section 3 is basically sufficient for the transformation from the source sequence to the target sequence in research topic finding.

### *4.2 the Cost Value*

Once the set of operations of the edit distance is determined, the following work to decide a proper cost value for each one. Ideally, the cost value will be such that resulting optimal alignment reflects researchers' performance completely. In the paper, the cost function is defined as a function related to similarity. Equations 1, 2, 3, 4 define the costs of Insertion, Deletion, Substitution and combination, respectively.

$$w(b_j)=1-S(b_j) \quad (1)$$

$$w(a_i)=S(a_i) \quad (2)$$

$$w(a_i,b_j)=1-S(a_i,b_j) \quad (3)$$

$$w(a_{i-k+1},\cdots,a_i,b_j)=\sum(1-S(a_{i-h+1},bj))/k \quad (1\leqslant h\leqslant k,\ k\in N,h\in N) \quad (4)$$

Where $S(a_i)$ or $S(b_j)$ is the similarity of the element $a_i$ or $b_j$ and the research topic, and $S(a_i,b_j)$ represents the similarity of the element $a_i$ and the element $b_j$. For insertion and substitution, the cost function is defined as subtracting similarity from 1. Deletion can only use the similarity as its cost function value. But combinations are more complex. The similarity of each element in source sequence and the corresponding element in target sequence should be calculated firstly, and then the average value of them should be computed as the cost of combination. From the equations it can be observed that the cost value of transformation will be zero if the source sequence and the target sequence are the same completely. In this condition, all of operations are simply substitutions.

### *4.3 the Rule Application*

When the edit distance is acted on two sequences drawn from the problem space and the solution space respectively, the optimal alignment could be found. This alignment reflects the nature of researchers' thought in the process of research topic finding. To mine a new topic finding, firstly the feature of the problem space about the new problem is represented. So in the retrieval phase the feature of the new problem is compared with the feature of the existing cases in the case base according to the attributes of different parts using fuzzy retrieval technology. A set of candidate cases for each part feature of the new problem could be gained. The corresponding alignment for each case could be gotten in the rule base. Then the best rule is applied to corresponding data of the new problem in the data base, and gain the new recommender research topic.

## 5. Conclusions and Future Work

This paper explores a new type of recommender systems based on the CBR technology for research topic finding. The framework of it is presented, and the edit distance is applied to solve two very important problems in the recommender systems, knowledge representation and rule mining.

To improve academic research level, finding a new valuable research topic is the first step. Therefore helping researchers to find research topics is one of the most important tasks. The proposed recommender system can be effectively applied to universities and research institutes helping researchers directly on the research topic finding process. The recommender systems as tools for research topic finding will play an important role in the research process.

Furthermore, the CBR technology is firstly applied to the recommender systems to facilitate research topic finding. The edit distance is used to represent the cases, which perfectly reflects the experience of the senior researchers' finding a new research topic. The existing

rules are used to mine a new research topic for the future research work. This paper will contribute to the development of CBR research by improving the CBR structural model to have a new feature in the dynamic structural change of case representation and rule extending, which will enable the recommender systems to possess a strong learning ability.

But this paper only proposes a conceptual framework of CRSRTF. The only preliminary work is done to carry out knowledge representation and rule mining. More research is required to further consummate the theory and develop a prototype system of CRSRTF, including determining interfaces, software/hardware environments, designing three bases and four modules and so on. At the same time, the edit distance should be improved to find the valuable research topics more efficiently. For example, the cost function could be specified in further detail, not only a simple function of the similarity.

## Acknowledgement

The research in this paper has been funded by the NSFC No.70201009.

## References

Arcos, J.L., Grachten, M., and Mantaras, R.L. "Extracting Performers' Behaviors to Annotate Cases in a CBR System for Musical Tempo Transformations," *the 5th International Conference on Case-based Reasoning*, Trondheim, Norway, 2003

Arslan, A.N., and Egecioglu, O. "An Efficient Uniform-Cost Normalized Edit Distance Algorithm**,"** *International Workshop on Groupware* (22:24), 1999, pp. 8 - 15

Franz, M., Ward, T., McCarley, J., and Zhu, W. "Unsupervised and supervised clustering for topic tracking," *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, 2001, pp. 310-317

Gilbert, D., and Schroeder, M. "FURY: Fuzzy Unification and Resolution Based on Edit Distance," *Proceedings of the IEEE International Symposium on Bio-Informatics and Biomedical Engineering*, 2000

Karp, R. "Mapping the genome: Some combinatorial problems arising in molecular biology," *Proc. 25th ACM Symposium on Theory of Computing,* 1993, pp. 278–285

Nunamaker, J. F., JR., M. C., and Titus, P. D. M. "Systems development in information systems research," *Journal of management information systems* (7:3), 1990-1991, pp. 88-106

Pighizzini, G. "How hard is computing the edit distance?" *Information and Computation* (165:1), 2001, pp.1-13

Ramamonjisoa, D. et al. "Research topic Discovery from WWW by Keywords Association Rules," *Second International Conference, RSCTC2000, Revised papers*, LNAI 2005, Ziarko and Yao Eds., 2001, pp.412--419

Ramamonjisoa, D. "Towards Automated Research topic Discovery on Scientific Domain by Agents System," *http://www.ssgrr.it/en/ssgrr2003w/papers/157.pdf*, *http://citeseer.nj.nec.com/555050.html*, 2003

Rashid, A.M., Albert, I., Cosley, D., Lam, S.K., McNee, S., Konstan, J.A., and Riedl, J. "Getting to Know You: Learning New User Preferences in Recommend Systems," In *Proceedings of the 2002 International Conference on Intelligent User Interfaces*, *Sa*n Francisco, CA, 2002, pp.127-134

Resnick, P., and Varian, H.R. **"**Recommender Systems," *Communication of ACM* (40:3), 1997, pp. 56-58.

Riesbeck, C.K., and Schank, R.C. *Inside Case-based Reasoning*, Lawrence Erlbaum Associates, 1989

Schwartz, R. et al. "Unsupervised Topic Discovery," in *Proceedings of Workshop on Language Modeling and Information Retrieval*, 2001, pp.72-77

Watson, I. *Applying Case-Based Reasoning: Techniques for Enterprise Systems*, Morgan Kaufmann Publishers, Inc. San Francisco, California, 1997

Wayne, C. L. "Topic detection and tracking in English and Chinese," *Proceedings of the fifth international workshop on Information retrieval with Asian languages*, 2000, pp.165-172

Wei, J. "Markov edit distance," *IEEE Transactions on Pattern Analysis and Machine Intelligence* (26:3), 2004, pp.311-322