Dynamic FAQ Retrieval with Rough Set Theory

***Deng-Yiv Chiu* [†], *Pei-Shin Chen* [††], *and Ya-Chen Pan* [†††]**

[†]*Faculty of Information Management, Chung Hua University, HsinChu, Taiwan 300, R.O.C.*

Summary

We explore FAQ (frequently asked questions) retrieval by applying hierarchical agglomerative clustering method and rough set theory. The clustering method and FAQ collection are used to construct a FAQ clustering concept hierarchy. Then, we use lower/upper approximations in rough set theory to classify users' queries. The rough set theory can help solve uncertain problem well. In experiments, the data is collected from accounting systems and requisition systems of non-profit organizations in Taiwan. The empirical results show that the proposed classification approach is valuable.

Key words:
Frequently asked questions, Hierarchical agglomerative clustering method, Rough set theory, Concept hierarchy

1. Introduction

Nowadays the enterprise has regarded the customer relationship as managing core to satisfy customer needs and maintain customer relationship well. There are many channels of communication between enterprise and customers, such as internet service, e-mail, fax, customer service center, local retailer, etc [1]. To assist customers to find the solution quickly, many enterprises began to design FAQ systems. FAQ systems not only help customers solve questions, but assist the customer service to answer customer questions [2].

However, the majority of the FAQ systems are established with enterprise views. Enterprises often set up FAQ collection and the appropriate answers in advance. But it cannot satisfy customers' needs well via such kind of FAQ systems.

In FAQ retrieval systems, users may typically encounter some difficulties [3]. The matched FAQs are often not satisfactory when keywords are used to search for relevant FAQs. The reason is that mostly information providers try to answer in advance questions that customers may inquery. Another probability is that the extracted representative words from a FAQ can not represent the FAQ sufficiently since the length of the FAQ is brief. Some researches use similarity degree among questions to classify customers' questions. The measurement criteria of similarity degree include Dice coefficients, Jaccard coefficients and the overlap coefficients [4]. Also, some researches apply query log clustering method in which a new similarity measurement using a machine readable dictionary is adopted to improve the deficiencies of FAQ retrieval [5].

The famous FAQ retrieval systems include FAQ Finder [6], Auto-FAQ [7], Sneiders' system[3], and Ask Jeeves[8]. In FAQ Finder system, vector-space model (VSM) is used to calculate similarity degree and WordNet is used to perform concept matching. In Auto-FAQ system, techniques in natural language processing are adopted to improve the performance of keyword comparison. In Sneiders' system, to match users' queries to FAQ collection, keywords are classified into required keywords, optional keywords, and irrelevant keywords. In Ask Jeeves system, the FAQ collection is classified into 11classes. Then keywords of user queries are used to search for relevant FAQs. In some researches, case-based reasoning (CBR) method is adopted to find a set of rules, and user queries will be added to FAQ collection incrementally to propose a dynamic retrieval method [9]. But, it can not process uncertain classification efficiently.

In this research, to improve the performance of FAQ retrieval, we first apply hierarchical clustering method to construct the clustering concept hierarchy for the FAQs used as training data. Then, we use lower/upper approximations in rough set theory to classify user queries. The empirical results show that the proposed classification approach is valuable.

This paper is organized as follows: Chapter 2 covers theoretical considerations. Section 2.1 introduces the details of hierarchical agglomerative clustering method. Section 2.2 introduces the concept of rough set theory. Section 2.3, 2.4, 2.5 and 2.6 introduce the structure and the details of the proposed FAQ retrieval approach. Chapter 3 introduces experiments and compares the proposed method with other methods. The last chapter introduces the conclusion, findings, and future works of this study.

Manuscript received August 5, 2007
Manuscript revised August 20, 2007

2. Theoretical Consideration

2.1 Hierarchical agglomerative clustering method

Hierarchical agglomerative clustering method is used to cluster data items repeatedly with the hierarchical structure from bottom to up. At the beginning, each individual item forms a cluster in its own. Then, clusters with nearest distance are merged until all items belong to one cluster (or to a predefined number of clusters). The hierarchical agglomerative clustering algorithm is presented as follows.

(i) Each individual item forms a cluster in its own $C_1,\ldots,C_n$.
(ii) Find the pair of C_i and C_j with the nearest distance.
(iii) Merge C_i and C_j to form a new cluster.
(iv) If the number of remaining clusters is equal to one (or predefined number), the process is terminated; otherwise, repeat (ii) and (iii).

2.2 Rough set theory

Rough set theory was proposed by Pawlak [10]. It is based on the rules of data mining and artificial intelligent algorithms. It is suitable to discover uncertain and incomplete implied knowledge. In rough set theory, a database consists of four components $S=\{U,Q,V,f\}$, where U is the universe consisting of a finite set of objects, Q is a finite set of attributes, V is a set of values $=U_{q\in Q}V_q$ where V_q is a value of the attribute q, and $f:U\times Q\rightarrow V$ is a function such that $f(x,q)\in V_q$ is the function between record x and attribute q [11].

Rough set theory is based on the approximation concept and lower/upper approximations of a set. Each target set is defined by lower/upper approximations. If X is a subset of U, A is a subset of Q , the lower/upper approximations of A to X can be represented by $\underline{A}X$ and $\overline{A}X$. All of objects in $\underline{A}X$ (lower approximation) must belong to X. All of objects in $\overline{A}X$ (upper approximation) could belong to X. A set X is said to be rough if there exists an element in upper approximation but not in lower approximation. The formulas are defined as below.

$$\underline{A}X=\{x\in U:[x]_A\subseteq X\} \quad (1)$$

$$\overline{A}X=\{x\in U:[x]_A\cap X\neq\varnothing\} \quad (2)$$

Where

$[x]_A$ is the A-elementary set.

2.3 The structure of the proposed FAQ retrieval approach

The structure of the proposed FAQ retrieval approach is shown in Fig. 1. Hierarchical clustering method is applied to construct the clustering concept hierarchy for the training data. Then rough set theory is used to classify users' queries and to generate relevant FAQs. The detailed steps are as follows.

(i) Mining features of FAQs in FAQ collection. First, we extract representative keywords from FAQs in FAQ collection. Keywords are used as mining features to represent FAQs.
(ii) Constructing clustering concept hierarchy. In this step, we apply hierarchical clustering method to keywords extracted from FAQ collection to construct a clustering concept hierarchy. FAQs with high similarity degree are clustered together.
(iii) Mining features of the user query. In order to classify the user query, here we extract representative keywords for user query.
(iv) Applying rough set theory to classify user query. In this step, we apply lower/upper approximations in rough set theory to classify the user query by using keywords of user query. Then, the user query is added into FAQ collection.
(v) Generating relevant FAQs. Finally, the relevant FAQs for the user query are generated. The relevant FAQs are those in the cluster to which the user query is assigned.

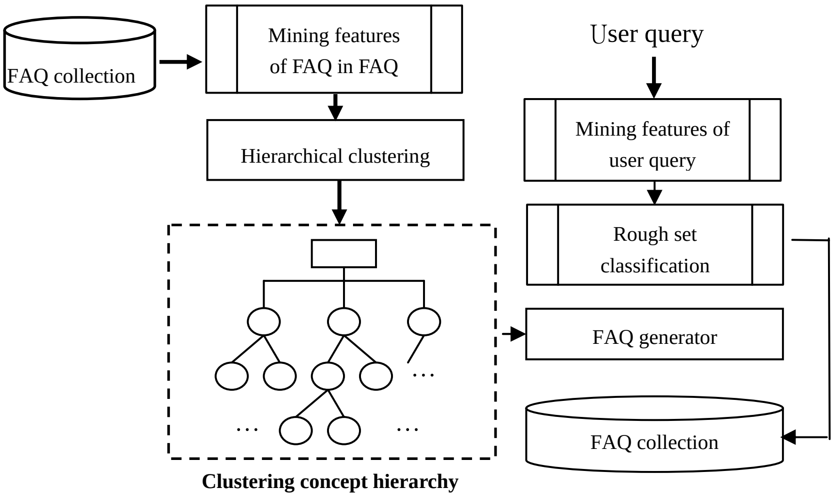

Fig. 1 The structure of proposed FAQ retrieval approach

2.4 Mining features of FAQs in FAQ collection

In order to extract representative keywords of FAQs, we adopt CKIP system (Chinese Knowledge Information Processing Group) to segment questions into separated terms. Then, with the assistance of domain experts, we recombine domain terms that are wrongly considered as several separated ordinary terms in previous step, such as domain terms, 請購 (requisition) in accounting is segmented into two separated ordinary terms, 請(please) and 購(purchase) in pervious step. The steps are as below and an example is shown in Table 1.

(i) Delete all of the punctuation marks, such as ，,。，、，；etc.

(ii) Delete all of the numeral character, such as 0,1,2…etc.

(iii) Recombine domain terms.

(iv) Delete the stop words in traditional Chinese, such as "的"(of), "在"(at).

(v) The remaining words labeled as domain terms, noun, and verb are regarded as representative keywords and are stored to database.

Table 1: An example of FAQ feature mining

Process step	Question：新增請購案，要在何處新增 Question in English：Where to add a requisitions								
	*feature*1	*feature*2	*feature*3	*feature*4	*feature*5	*feature*6	*feature*7	*feature*8	*feature*9
Segment a FAQ with CKIP system	新增 add	請 please	購 purchase	案 case	， ,	要 needing	在 at	何處 where	新增 add
Delete punctuations	新增 add	請 please	購 purchase	案 case		要 needing	在 at	何處 where	新增 add
Recombine domain terms	新增 add	請購案 requisitions				要 needing	在 at	何處 where	新增 add
Delete stop words	新增 add	請購案 requisitions						何處 where	新增 add

2.5 FAQ clustering concept hierarchy

Here we apply hierarchical agglomerative clustering analysis method [12] to cluster FAQs in FAQ collection. First, we compute distance among FAQs. And, each FAQ is regarded as a single cluster at the beginning. Then we construct the concept hierarchical structure of the FAQs. Finally, we define each cluster with the assistance of domain experts.

Here we illustrate the clustering procedure. Table 2 lists 5 FAQs and their extracted keywords.

To calculate the similarity degree, the formula used to calculate similarity degree is shown in Eq.3. For example, the similarity degree between FAQ 4 and FAQ 5 is 0.375 (3/4*3/6) since there are 3 extracted keywords appearing in both of FAQ 4 and FAQ 5, there are 4 keywords in FAQ 4, and there are 6 keywords in FAQ 5.

$$similiarity\ (A,B) = \frac{(number\ of\ same\ keywords\ in\ both\ A\ and\ B)^2}{number\ of\ keywords\ in\ A*number\ of\ keywords\ in\ B} \quad (3)$$

The similarity degrees are used in clustering process as shown in Fig. 2.

Table 2: An example of FAQs and their extracted keywords

FAQ2	請購案沖銷出現”核銷數大於決標數” The write-off of requisitions shows approved amount greater than amount of contract					
Keywords	請購案 requisitions	沖銷 write-off	核銷數 approved amount	決標數 amount of contract		
FAQ2	開傳票的會計科目欲存檔時會出現請輸入最低層級 The accounts of voucher shows ‘please input lowest layer’ when you want to save to file					
Keywords	傳票 voucher	會計科目 accounts	存檔 save to file	輸入 input	最低 lowest	層級 layer
FAQ3	傳票作業登打會計科目，出現請輸入最低層級 When you key in accounts in accounts operation, it shows ‘please input lowest layer’					
Keywords	傳票 voucher	作業 operation	會計科目 accounts	輸入 input	最低 lowest	層級 layer
FAQ4	沖銷暫付款出現”核銷數大於決標數” write-off of temporary payment shows ‘approved amount greater than amount of contract					
Keywords	沖銷 write-off	暫付款 temporary payment	核銷數 approved amount	決標數 amount of contract		
FAQ5	某幾筆暫付款在開傳票時出現“ 核銷數大於決標數” 不能存檔的問題 Some temporary payment shows the problem of ‘approved amount greater than amount of contract so that it cannot save to file when you write voucher					
Keywords	暫付款 temporary payment	傳票 voucher	核銷數 approved amount	決標數 amount of contract	存檔 save to file	問題 problem

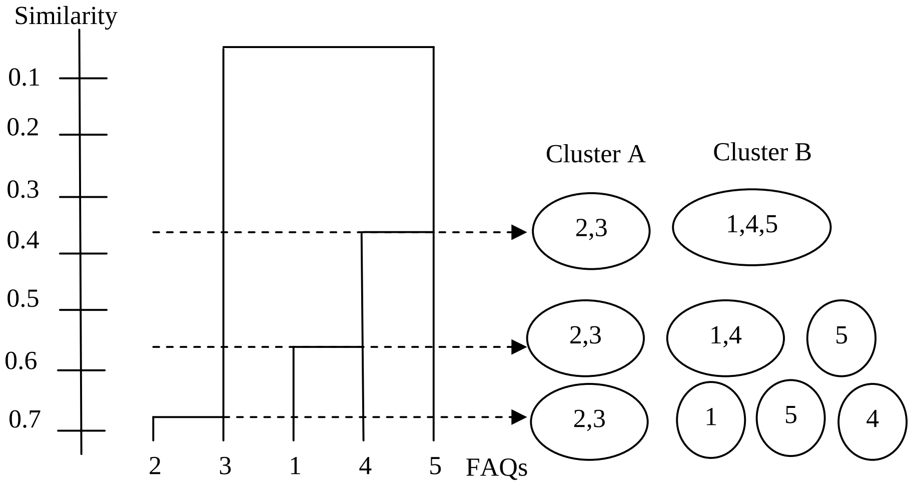

Fig. 2 An example of clustering concept hierarchy

Finally, we obtain two clusters, respectively called cluster A including FAQ 2 and FAQ 3, and cluster B including FAQ 1, FAQ 4, and FAQ 5. With this method, we can cluster our training data of FAQ collection and construct the clustering concept hierarchy.

2.6 Classifying user query with rough set theory

We use rough set theory to determine the cluster to which the user query belongs as below.

Algorithm: rough set classification

```
Input:
    new $F_1$,…new $F_i$          //features of user query
    $C_1,...,C_j$                 //clusters in FAQ collection
    $minSupport$                  //minimum support value
    $Q_1,...Q_m$                  //FAQs in FAQ collection
    $F_{1n},...,F_{qt}$           //features of FAQ $Q_t$ in FAQ
                                    collection
Step：
(i)   Find out the lower approximation, $\underline{A}$, and upper
      approximation, $\overline{A}$, of the user query to cluster, $C_k$,
      in FAQ collection.
(ii)  If (new $F_1$ ,…new $F_i$ ) $\in \underline{A}$ , then
      (new $F_1$,…new $F_i$) $\in C_k$ .
(iii) If (new $F_1$ ,…new $F_i$ ) $\in \overline{A}$ and minimum support
      value is greater than $minSupport$ , then
      (new $F_1$,…new $F_i$) $\in C_k$
```

3. Experiments

3.1 The data and clustering concept hierarchy

The data is collected from FAQ collections of accounting systems and requisition application systems used by over 50 national universities and juridical person organizations in Taiwan. Those organizations use Organization Resource Plane (ORP) information system developed by AIFU Co.* to integrate their accounting systems and requisition application systems. The data collection consists of 1241 FAQ data items recorded from 2002/03/29 to 2004/03/17.

We use following steps to preprocess noisy data.

(i) Delete records not written in traditional Chinese.
(ii) Delete records containing content exceeding five sentences or 60 words.
(iii) Delete records related to specific organizations, such as 'what is the date to sign the accounting system contract with XX university?'.
(iv) Delete records related to specific date, such as requisition application dated June 17 was not found.

*AIFU company develops Organization Resource Planning (ORP) information system, which includes human resource management, assets management, and financial control management and so on. The company develops information management system for non-profit organizations in Taiwan.

There are 809 data items remained. We use 521 data items (from 2002/3/29 to 2003/11/5) as training data and the rest 288 data items (from 2003/12/7 to 2004/3/17) as testing data. Then we apply the hierarchical agglomerative clustering method to training data and construct a FAQ clustering concept hierarchy. In the hierarchy, there are 3 levels in which first level includes 2 clusters, second level includes 8 clusters, and third level includes 28 clusters, as shown in Table 3.

Table 3: The FAQ clustering concept hierarchy in the study

First level	Second level	Third level	
Cluster name	Cluster name	Cluster name	
Accounting system (357)*	Voucher (128)	Voucher preparation (35)	Write-off (23)
		Expenditure purpose (29)	Posting (16)
		Print-out of vouchers (18)	Voucher number (7)
	Purchase request (83)	Temporary payments (36)	Verification (31)
		Purchase order number (16)	
	Detail (53)	Details of receipts and expenditure (23)	Details of expenditure (5)
		Details (19)	Chart of accounts (6)
	Error message (25)	Negative request purchase’s changes (6)	Negative temporary payment’s changes (9)
		Verification numbers greater than mark number (10)	
	Budget report (21)		
	Other categories (47)	Accounts (14)	Lowest level (6)
		Department budget (13)	Adding accounts (8)
		Bank reconciliation (6)	
Requisition system (164)	Purchase request (111)	Adding purchase request (32)	Purchase order number (21)
		Expenditure categories (17)	Purchase order query (41)
	Other questions (53)	Web page display (25)	Print out (17)
		Disordered code (11)	

*The number in the parentheses beside cluster name is the total number of data items belonging to the cluster.

3.2 User query classification and evaluation

Here we apply rough set theory to the 288 testing data items to classify users’ queries into appropriate clusters in the clustering concept hierarchy. The minimum support value is set to 0.64 with trial and error approach. To evaluate the performance of the proposed method, we use true positive fraction (TP), false position fraction (FP), and false negative fraction (FN). They are defined as below. The evaluation of performance is shown in Table 4.

TP: a user query is classified into the cluster to which it belongs.

FP: a user query is classified into the cluster to which it does not belong.

FN: a user query is not classified into the cluster to which it belongs.

Table 4: Performance of user query classification

Cluster Name	Total of data items belonging to the cluster	True Positive fraction(TP)		False Positive fraction(FP)		False negative fraction(FN)	
		Total	Rate	Total	Rate	Total	Rate
Voucher	83	82	98.80%	1	0.49%	1	1.20%
Purchase request	64	60	93.75%	3	1.34%	4	6.25%
Detail	33	30	90.91%	4	1.57%	3	9.09%
Error message	14	14	100.00%	0	0.00%	0	0.00%
Budget report	11	10	90.91%	1	0.36%	1	9.09%
Other categories	22	19	86.36%	2	0.75%	3	13.64%
Purchase request	47	44	93.62%	3	1.24%	3	6.38%
Other questions	14	14	100.00%	0	0.00%	0	0.00%
	Average rate	94.79%		0.69%		5.21%	

The overall true positive fraction is 94.79%, the false positive fraction is 0.69%, and the false negative fraction is 5.21%. Also, false negative of clusters ‘Detail’, ’Budget report’, and ‘Other categories’ are high. It seems that we need more training data items of those clusters to extract representative keywords for those clusters.

Then, we generate relevant FAQs for each user query and use precision, recall, and F-measure to evaluate the retrieval performance. Precision is the ability to retrieve only relevant items and recall is the ability to retrieve all of relevant items existing in FAQ collection. The formulas are shown as follows. The distribution of precision and recall in a certain range is shown in Table 5. The average precision, recall, and F-measure are shown in Table 6 in which we also show performance of some other research [13].

$$\text{Precision} = \frac{\text{numberof relevant FAQs retrieved}}{\text{total number of FAQs retrieved}} \quad (4)$$

$$\text{Recall} = \frac{\text{number of relevant FAQs retrieved}}{\text{total number of relevent FAQs in FAQ collection}} \quad (5)$$

$$\text{F-measre} = \frac{2*\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}} \quad (6)$$

Table 5: Distribution of precision and recall in a certain range

Range	0.00~0.19	0.20~0.39	0.40~0.59	0.60~0.79	0.80~1.00
Precision	2	5	12	29	240
Recall	1	4	5	19	259

The performance under 60% for the precision occurs 19 of 288 times. The performance under 60% for recall occurs 10 times of 288 times. Those are mainly due to not sufficient training data. Also, the recall is better than 80% for 259 times since the hierarchical clustering method is applied to training data such that many relevant FAQs are clustered together in the hierarchy.

The performance of the proposed approach is significant. The reasons are that the hierarchical clustering method is utilized to set up an appropriate structure for the FAQ collection, rough set theory is used to help solve uncertain problems well, and domain experts provide assistance in extracting enterprise terms and preprocessing noisy data.

Table 6: Compare with the other methods

Method	Precision	Recall	F-measure
The proposed approach	93.31%	90.64%	91.96%
Prioritized Keyword Matching	88.00%	85.00%	86.47%

4. Conclusions

In the research, we explore dynamic FAQ retrieval approach. The clustering method and training data are used to construct a FAQ clustering concept hierarchy. Then we use lower/upper approximation in rough set theory to classify users' queries.

We conclude that appropriate usage of enterprise terms does help. Also, rough set theory can be used to improve the performance of user query classification significantly since it can help process uncertain classification well.

In the future, we expect to continue the works as below.

(i) We may apply methods in natural language to analyze the structure of semasiology and sentence to extract appropriate and representative keywords for FAQs,

(ii) Instead of gaining assistance from domain experts, we plan to use rule-based knowledge database to provide rules to preprocess noisy data and to extract enterprise terms automatically. It can promote research value although the accuracy rate may decrease.

Acknowledgments

The authors were supported in part by National Science Foundation and AIFU Co., Grant No. NSC 92-2622-E-216-015-CC3

References

[1] O. Walton, Jr., : "Customer relationship management in an e-business environment," Proceeding of the 2001 IEEE International Conference, pp.311-316 , 2001.

[2] F. Yeh, "The study of supporting online FAQ generation," Institute of Human Resource Management, National Sun Yat-sen University, Master Thesis, 2003.

[3] E. Sneiders, "Automated FAQ answering: continued experience with shallow language understanding," AAAI Fall Symposium, pp.97–107, 1999.

[4] C. J. Van Rijsbergen, "Information retrieval (Second ed.)", London: Butterworths, 1979.

[5] H. Kim and J. Seo, "High-performance FAQ retrieval using an automatic clustering method of query logs," Information Processing and Management, vol. 42, pp.650–661, 2006.

[6] K. Hammond, R. Burke, C. Martin and S. Lytinen, "FAQ Finder: A case-based approach to knowledge navigation," Proceedings of the 11th Conference on Artificial Intelligence for Applications, pp. 80–86, 1995.

[7] S. D. Whitehead, "Auto-FAQ: An experiment in Cyberspace leveraging," Computer Networks and ISDN Systems, vol. 28(1–2), pp.137–146, 1995.

[8] Ask Jeeves, http://www.ask.com.

[9] Y. Fu and R. Shen, **"**GA based CBR approach in Q&A system," Expert Systems with Applications, vol. 26, Issue 2. pp.167-170, 2004.

[10] Y. Yang and T. C. Chiam, "Rule discovery based on rough set theory," Proceedings of the ISIF Conference, pp.11-16, 2000.

[11] Z. Pawlak, "Rough sets," International Journal of Computer and Information Science 11, pp.341-356, 1982.

[12] W. Hui, I. Duntsch and G. Gedigal ," Classificatory filtering in decision systems," International Journal of Approximate Reasoning, pp. 111-136, 2000.

[13] E. Sneiders, " Automated FAQ answering: continued experience with shallow language understanding, " AAAI Fall Symposium, pp.97–107, 1999.

Deng-Yiv Chiu received the B.A. from Averett College, Virginia, USA in 1988, M.S. from University of Maryland, USA in 1990. He received the PhD in Computer Since from Illinois Institute of Technology, USA in 1994. After working as an assistant professor in the Dept. of Math and Computer Science, Chicago State University, USA, he has been an associate professor at Chung Hua University, HsinChu, Taiwan since 1996. His research interest includes machine learning, information retrieval, and their applications to knowledge management and finance.