Query Reconstruction for Searching QA Documents by Utilizing Structural Components*

Sang-Hee Choi, Eun-Gyoung Seo*****

Abstract

This study aims to suggest an effective way to enhance question-answer(QA) document retrieval performance by reconstructing queries based on the structural features in the QA documents. QA documents are a structured document which consists of three components: question from a questioner, short description on the question, answers chosen by the questioner. The study proposes the methods to reconstruct a new query using by two major structural parts, question and answer, and examines which component of a QA document could contribute to improve query performance. The major finding in this study is that to use answer document set is the most effective for reconstructing a new query. That is, queries reconstructed based on terms appeared on the answer document set provide the most relevant search results with reducing redundancy of retrieved documents.

초록

질의응답문서는 이용자가 입력한 질의, 질의설명, 답을 아는 다른 이용자가 제시한 응답으로 구성된 구조화된 문서로서, 최근 웹 문서처럼 검색이 일반적으로 일어나고 있는 정보원이다. 이 연구에서는 질의응답문서의 구조적 특성을 기반으로 질의를 재생성하여 질의응답문서의 검색효율을 향상시키고자 하였다. 질의재생성 실험에서 성능이 비교된 문서구조는 질의와 응답내용이다. 질의를 기반으로 질의를 재생성하는 방식에서는 질의응답검색 시스템에 입력되어 있는 유사질의를 활용하여 클러스터링하는 기법이 적용되었다. 응답정보를 기반으로 질의를 재생성하는 방식에서는 가장 유사한 기존 질의에 대해 응답된 내용에서 단락검색으로 적합한 문장들을 선정하여 활용하는 기법이 적용되었다. 실험 결과 응답정보를 활용하여 질의를 재생성하는 방식이 정확률은 유지하면서 더 다양한 검색결과를 제공하는 것으로 나타났다.

Keywords: Query-Answer Documents, Query Reconstruction, Query Clustering, Query Performance, Clustering, Passage Retrieval

질의응답문서, 질의재생성, 질의클러스터링, 클러스터링, 단락검색

---

* This research was supported by Post Doc program of Korea Research Foundation(H00007)

** Lecturer of Library & Information Science Dept. at Yonsei Univ. (tudultudul@naver.com)

*** Professor of Division of Knowledge & Information at Hansung Univ. (egseo@hansung.ac.kr)

1. Introduction

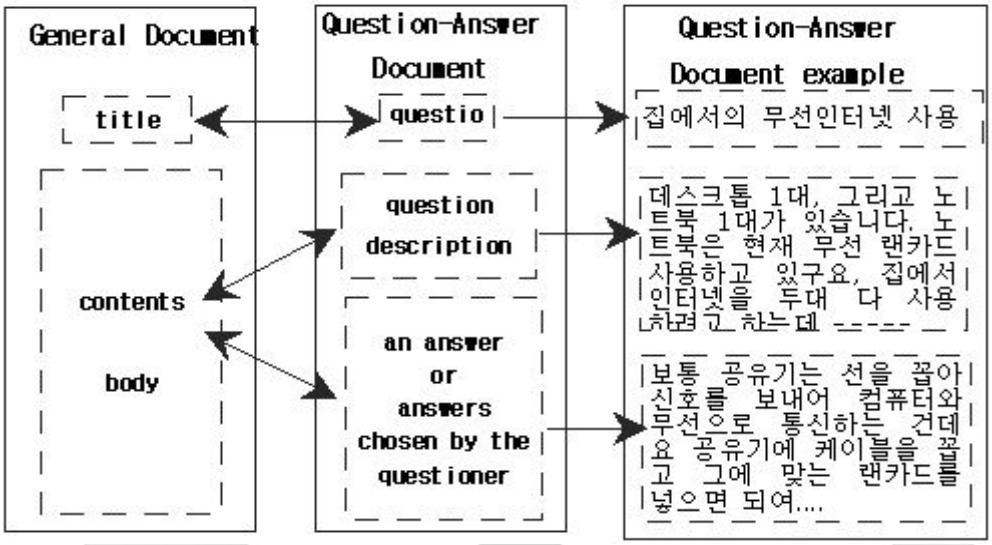
The Internet enables larger numbers of people to be brought into a dialog and to share and combine the information set into a larger, multifaceted knowledge base. The Internet as a new medium for information exchange has to do with the quality and quantity of information available. When internet users ask carefully about the information they want to know, other users can give them the variety of levels of answers. For example, once a user posts a question on a web site that provides the question-answering(QA) service, people who know the answers put answers on the site. As soon as the certain amount of answers are delivered to the questioner, he or she picks the right answer among them and attaches to the original question. Most of web sites provide QA services which allow users to be able to post questions or to answer the questions.

The Internet provides a means for information users to participate fully in the exchange of information. By making their questions and answers available to others, they build a new type of critical information resources as the output of question-answering activities. As this type of information resources has been increased and accumulated, the web allows users to search QA resources. Now, searching a data set of questions and answers accumulated through QA service becomes one of the popular information seeking approaches. Furthermore, users don't need to put always their questions on the web site. Users, first, browse the questions and then identify whether similar questions which they want to ask are posted. If the question is posted, users just read the answers and if not, they ask the question. In a certain way, searching cumulated questions and answers could be regarded as reusing the results of query answering process.

Documents generated as the result of QA process have unique components, comparing with the general documents. The general documents have two components such as title and contents, while the documents generated by QA process have three components such as question, question descriptions and answers. The documents with theses three components can be called as the question-answer documents(QA documents). The <Figure 1> shows the difference between general documents and QA documents in terms of the structure of components. The question statement of a user could be matched to the title of a document. The body of general document can be divided into two sections: question description section and answer section in the QA document.

Because QA services on the web allow users to ask any kinds of questions including even same questions and to give similar answers, all of user's questions and their answers are filed up so that the number of retrieved QA documents is getting higher. Therefore, searchers of QA

documents have to compare redundant documents in order to select more relevant information. It makes poor time for users to browse or/and find the relevant information among the large number of search results. Environment of searching QA documents becomes more like general web document retrieval environment.



<Figure 1> Comparison of structures in general document and QA document

However, the retrieved results of QA documents could be better if the query statement be more built elaborately. It has been always difficult for end-users to compose a query statement which can fulfill their own information needs. That is, users can't represent appropriate query terms which describe their information needs, or they don’t know relevant terms because they are not familiar with the search topics. Therefore the retrieval performance could be improved if the retrieval interface system helps users to construct a query statement or automatically reconstruct queries.

In the QA documents, the search topic is represented by two kinds of people, questioners and answerers. The questioner has a primary responsibility to express his own search interest in a question and the answerer expresses answers of the question, which contributes to describe the same topic with his own terms. So, QA documents have two distinctive parts to represent the search topic.

This study aims to suggest an effective way to enhance QA document retrieval performance by reconstructing queries based on the structural features in the QA document. Specifically the study examines which component of a QA document could contribute to reconstruct queries and to improve query performance. In general, terms in queries have been considered to be the

most effective to describe the topics of queries, and comparatively answers are not regarded so important as the queries to represent questioner's intention. It is necessary to redefine the role of answers to grasp the major concepts of queries. In QA document retrieval, answers could contribute themselves to represent the topics of queries and clarify the meaning of queries as much as the queries for because answerers could know the topics better than the questioners. On this assumption, the roles of the query part and the answer part in the QA document were examined in terms of representing the topic of the queries effectively for query reconstruction. For this, the study evaluates the result of QA document searched by the new queries.

2. Literature Reviews

2.1 Query Processing

To improve retrieval performance, there are some remarkable researches in the field of query processing. Owei(2002) mentioned that a query was made up of incomplete information from user's limited knowledge base and it ended up with insufficient search results. He proposed to reconstruct a query by identifying concepts in user's natural queries.

Bloesch and Halpin(1997) investigated semantic relations among queries to grasp what users really try to search and then, built the commercialized system, ConQuer II, which automatically constructed a new query using conceptual scheme based on the semantic relations between object and entity. In the same way, Vizla system developed by Berztiss in 1993 supported to complete a query statement step by step as identifying conceptual relations between an entity and attributes of user's first query input.

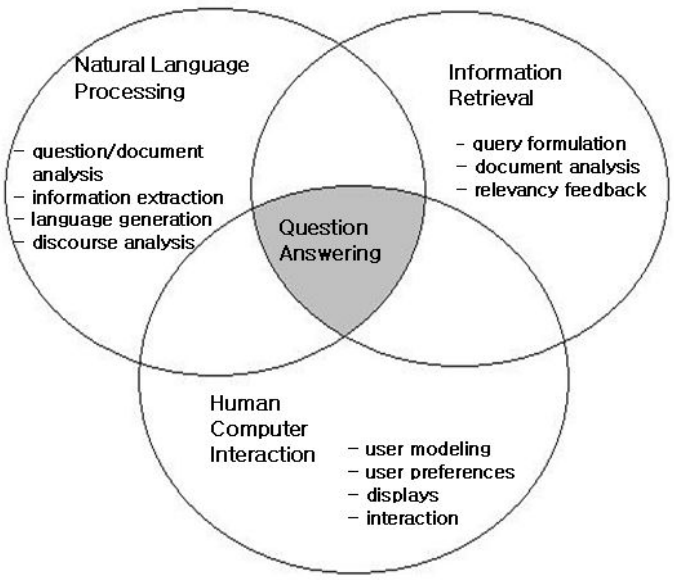
Kang and Kim(2003) investigated characteristics of web search queries, 'search, 'navigation', and 'transaction'. The first type of queries, 'search', was to retrieve relevant information directly. The second type of query, 'navigation', was to search homepage or site and the last type of query, 'transaction', was to find relevant services. They aimed to improve the precision of research results by using the different query processing depending on the retrieval activities and purposes.

Clustering methods have been applied to congregate relevant objects in many areas. Recently, clustering has been used to compare and identify meaning of queries. So, query clustering is a comparatively new research area considering term clustering or document clustering, but is noted as one of major clustering research areas. Many studies identified similar queries, add query terms, and reconstructed a new query using clustering methods(Gopal and Ramesh 1995, Wen, Nie, and Zhang 2001, Zhang and Liu 2004). Especially in Wen's research, they clustered queries using by users' log transaction data. If two

users viewed same pages with different queries, these used queries might be about the same topic. Also, they traced the similarity of queries using the retrieved results. On the contrary, Tombros and his colleagues clustered retrieved documents based on the queries in order to improve the representation of query terms(Tombros, Villa, and Rijsbergen2001). Tombros's study investigated the role of documents in presenting query's topic and it turned out to be effective to identify the meaning of queries. We considered the answer could be effective like document corresponding to queries and investigated the role of answers to intensify the query terms.

2.2 Question Answering

Maybury(2004) defines question-answering activity(QA activity) is an interactive human computer process that encompasses understanding a user information need, typically expressed in a natural language query; retrieving relevant documents, data, or knowledge from selected sources; extracting, qualifying and prioritizing available answers from these sources; and presenting and explaining responses in an effective manner. From this point of view, query answering is an elaborate process which is associated with various information processing methods of the core information science disciplines such as natural language processing, information retrieval, or HCI(see <Figure 2>).



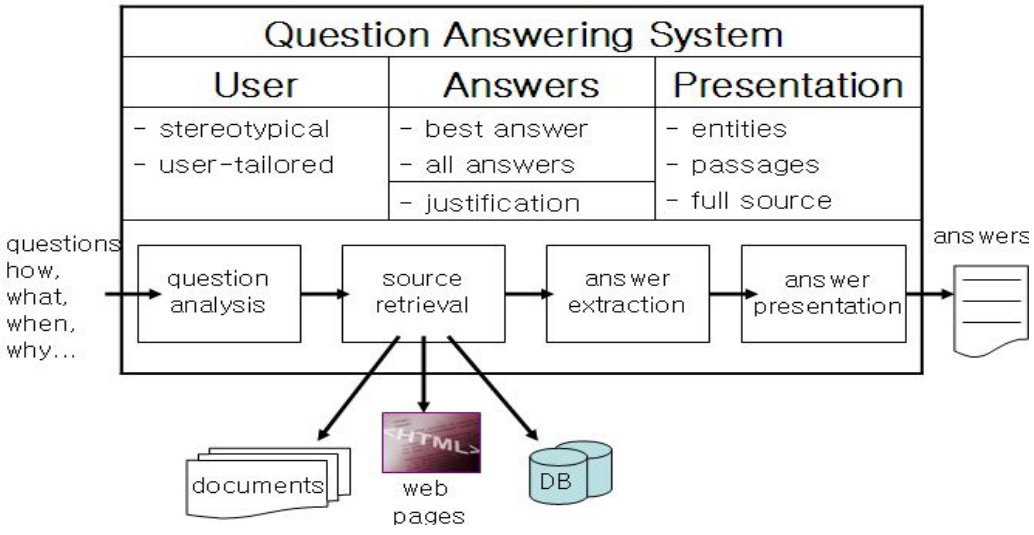
< Figure 2 > Relationship of Question Answering to Other Disciplines

The significant QA activities have been introduced in TREC(Text REtrieval Evaluation Conference) QA track and the ARDA Advanced Question Answering for Intelligence(AQUAINT) Program. The first web-based QA system, START, was developed by MIT in 1993 and in 1996 commercial QA system, Ask Jeeves started to process natural language query. In 1999, TREC-8 initiated question answering track with 20 participants.

The QA researches, in general, can be divided and descried according to the content of answers to questions such as the following:

1) Temporal question answering: processing question with temporal elements such as relative times, points, or durations.
2) Spatial question answering: processing question associating spatial elements such as location, regions or size.
3) Definitional and descriptive question answering: creating answers that define of describe terms, objects, or concepts.
4) Biographical question answering: creating or extracting significant events or facts in the life span of a person, or organization.
5) Opinioned question answering: detecting opinions and responses to questions regarding subjective content.

General QA process architecture can be addressed in <Figure 3>.



<Figure 3> Question Answering Architecture (Maybury 2004, 9)

This study focused on question analysis and source retrieval stages in whole question-answering(QA) process. Clustering QA documents components was applied to reconstruct queries for question analysis and the retrieved documents by enhanced queries were evaluated for the stage of source retrieval. Well-constructed queries can retrieve more relevant and various information source for question answering system. We expected reconstructing queries could contribute to improve not only general performance of QA documents retrieval but also basic functionality of QA system.

3. Query Reconstruction Methods

This study conducted two kinds of query reconstruction experiments using the different components of QA documents. For constructing new query, one used only query clusters and the other used the query-answer clusters.

3.1 Query reconstruction methods with QA document's components

Users who search QA documents primarily don't know well about the search topic or sometimes have wrong information on it. Choi(2005) found that the users who search QA documents often fail to express their information needs in appropriate query terms. Therefore this study was conducted for the purpose of selecting new relevant query terms which represent users' information needs more appropriately. Two approaches for new query reconstruction were examined in this paper; one was based on query clusters and the other was to use answers attached to the questions by the questioner.

QA document clusters were formed by three different QA document components, that is, 'question', 'question description', and 'answer'. According to the results of analysis of the clustering performance of each QA document component, it is found that users could effectively find relevant documents from answer-document cluster. That is, answer-document cluster could make to gather relevant documents most effectively comparing the prejudged clusters. On the contrary, the terms in query-description-documents were not effective to represent query's topic (see <Table 1>), even though the size of question-description-documents was almost four times as much as that of query-documents(see <Table 2>). Simple increment of the clustering features occurred in question-description-documents didn't affect the performance of QA document clustering. It proved also that the retrieval performance of answer-documents representing query's topic didn't depend on their size but their contents.

< Table 1 >Comparison of QA Document Component Clustering Results

	PRECISION	RECALL	F	WACS
Question	0.45	0.57	0.50	0.38
Question description	0.40	0.66	0.50	0.36
Answer	0.65	0.72	0.68	0.61

Precision = $\frac{1}{D}\sum_{j=1}^{n}\sum_{i=1}^{m}\frac{|M_i \cap C_j|^2}{|C_j|}$

Recall = $\frac{1}{D}\sum_{j=1}^{n}\sum_{i=1}^{m}\frac{|M_i \cap C_j|^2}{|M_j|}$

F(p, r) = $\frac{2pr}{p+r}$

WACS(c) = $\frac{1}{D}\sum_{j=1}^{n}\sum_{i=1}^{m}\frac{2|M_i \cap C_j|^2}{|M_i|+|C_j|}$ (Chung and Lee, 2001)

d: number of all documents

n: number of clusters generated by clustering

c: clusters made by person

m: clusters generated automatically

<Table 1> also shows that answer-documents contributed most to represent the topic of QA documents in terms of four evaluation criteria. Questioners could clarify what they really looked for while they were selecting the right answer written in appropriate manner. Accordingly, answer-documents could include more pertinent terms to describe questioner's intention because answerer could know more and precisely about the query's topic than the questioner.

<

<Table 2 > Average Number of Letters in QA Documents Components of Dataset

	Question	Question Description	Answer
average # of letters	88 letters	395 letters	818 letters

Based on the previous QA document clustering results, the two approaches to reconstruct queries were examined and compared in the experiment. The components of QA documents used in the experiment were query and answer. Each procedure of the approaches associating two components is presented in <Figure 4>.

<Figure 4> The procedure of two experiments for reconstructing a query

1) Query reconstruction by clustering queries

Users search QA documents to check out if there are similar questions answered previously. Especially, when the users aren't familiar with what they were asking, terms in the similar questions could help them to refine their queries. Query clustering was applied to identify relevant terms in similar queries to regenerate a query.

By using query centroid in the query cluster, a new query could be generated. The procedure for regenerating a query is described as below:

① Clustering queries in QA document data set

② Identifying the most relevant query cluster by comparing the similarity between query and query cluster centroid

③ Generating new query with the query centroid of the most relevant query cluster

④ Adapting cosine coefficient to identify similarity in queries and clusters.

Cosine Coefficient(x, y) = $$\frac{\sum_{i=1}^{t} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{t} (x_i)^2 \cdot \sum_{i=1}^{t} (y_i)^2}}$$

2) Query reconstruction by selecting sentences in the answer of the most similar query

In order to reconstruct a query, the second method used the answer-documents which were attached to the most relevant query to the original test query. The relevant answers were chosen by a questioner. From several candidate answers, that is, the questioner selected only few answers contained the most appropriate information about in terms of his question. Answer-documents could be the outcome of users' relevance evaluation and these procedures could also be treated as a user's feedback activity. Therefore, the answer-documents supplied more appropriate query terms for a new user who searches QA documents with similar information need.

In this study, the relationship between a query and an answer in QA documents was utilized to select pertinent terms. The procedure of utilizing the relationship and reconstructing a new query is as follows:

① Retrieving the most relevant query by a test query.

② Extracting the answer attached to the retrieved query. The extracted answer was previously chosen by the questioner.

③ Selecting the most relevant three sentences in the extracted answer. Cosine coefficient was applied to retrieve sentence vectors.

④ Generating a sentence centroid among top three relevant sentences.

⑤ Reconstructing a new query with a sentence centroid.

Tombros and Sanderson(1998) addressed that the relevant sentences to the query in the retrieved documents play a great role to effectively represent information needs implicated in the query. Therefore, they, first, selected some sentences related to the query and second, generated a summary with those relevant sentences. Under the perception of this previous finding about relationship between a query and relevant sentences, this study used sentences as information source to generate a new query.

3.2 Test Dataset and Limitation

The dataset of this study was extracted from the question-answering service provided by Naver, commercial web portal. Naver's QA documents in question-answering service consisted of a question, question description, and answers chosen by the questioner. Naver has provided question-answering services in the 11 subject fields, but the subject fields of test dataset in this study are the computer and network subject. The dataset was restricted in these fields to make the condition of QA document retrieval system similar to real QA document retrieval system as possible with limited number of documents. If the test dataset is built with diverse subjects, only few terms in query are occurred in other queries so that the recall ratio of experimental results appears very low unlike that of real world. In the other way, the retrieved results could have the high precision ratio if the number of QA documents in the dataset be limited and the subject fields of QA documents be too diverse. Therefore, if the experiment can not be conducted with huge size of dataset like real circumstance, it would be as well to build test dataset limiting a subject field to increase possibilities of problematic situation that could be caused by large number of redundant or similar information in test dataset. <Table 3> shows 10 topics of test dataset in the field of computer and network.

< Table 3 > Topics of Test Dataset

Number of Topics	Topics
M1	printer problem solution
M2	file sharing on the internet
M3	e-mail program setting problem
M4	mobile communication, mobile phone
M5	window setting
M6	coding web pages
M7	setting virus program
M8	connecting wireless LAN
M9	playing movie files
M10	notebook recommendation

On this limitation, the test dataset was constructed with 300 QA documents related in 10 topics shown in <Table 3>. For collecting test QA documents, two subject experts, first, searched QA documents with predefined query terms in Naver QA service and then chose 60

relevant documents from the result of search on each 10 topic categories. By each category, two subject experts critically read the documents and evaluated their relevancy to the topics of categories. After evaluating process, they selected the most relevant 30 documents for each test topic category. And test query set was constructed as several keywords occurred in the title of each selected QA test document. As a result, 300 QA documents and 300 new queries were built as the test dataset for the study. The predefined relevancy evaluation of the query to the topic categories was also applied to evaluate query clustering performance. This study was conducted in the experimental data set and predefined topics so it is necessary to expand the topic areas and the number of dataset to generalize the findings of this study.

4. Analysis of the Results

4.1 Evaluation method

The performance of reconstructed queries was evaluated by two methods. One is to use precision at ten, which was suggested in TREC. This evaluation method investigated how precisely the reconstructed queries could retrieve relevant documents on the topics. The other method is to evaluate how much various information is included in search results. The entropy formula, in this study, was used as an evaluation criterion to examine the performance of reconstructed queries, that is, the ability of reducing the redundancy of information in the search results. Han(1998) used the entropy formula in order to determine the topic coherence within the clusters. Han in his research proved that the cluster had many members on the same topic if the entropy value of a cluster was close to 0. It means entropy formula could be used to measure the diversity in a group. If the entropy value is high, we could interpret that there are various members in a cluster. The entropy formula used Kim and Lee's study is also used for the study(Kim and Lee 2000).

‘

Entropy (C) = $\sum_{i=1}^{n} - p_i \log_2 p_i$

p_i: rate of members of category i in cluster c

In this study, n in entropy formula is the number of topic categories in a search result by a query.

4.2 Analysis on the performance of reconstructed query

To evaluate the performance of reconstructed queries, QA test documents were retrieved by three types of queries: i) original query; ii) reconstructed query by query clustering; and iii) reconstructed query by selecting sentences in the answer previously chosen. From the search results, top ten documents were extracted and evaluated. Top ten documents of each query were treated as a cluster for entropy measurement. New 300 search results were produced by the new queries and evaluated by precision-at-ten and entropy methods.

According to the evaluation described in <Table 4>, the queries reconstructed by answer-documents showed the highest performance at the precision criterion. Overall queries reconstructed by two approaches had better performance than original queries even though the precision rate was not remarkably improved.

<Table 4> The evaluation on the performance of reconstructed queries

	Original query	Reconstructed query by query clustering	Reconstructed query by answer previously chosen
Precision-at-ten	0.70	0.72	0.74
Entropy	0.91	0.96	1.13

In terms of entropy, the tendency of reconstructed queries appeared in different manner. The performance of new queries reconstructed by query clustering is almost as same as that of original queries in entropy evaluation. The new query terms selected by query clustering were not effective to present various aspects of information needs implicated in the original query. On the contrary, new queries reconstructed by answer previously chosen have the highest entropy value. It could be interpreted that the method based on the answer-documents be able to expand topic diversity without decreasing precision. The entropy value of query by answer was the highest as well as that of precision-at-ten.

5. Conclusion

This study explored the nature of QA documents and proposed to utilize the structural components in the QA document to reconstruct original query. To identify the characteristics of structural components, each document set of the QA documents such as ‘question’, ‘question

description', 'answer' was clustered and evaluated in the first stage of the study. And then, the study conducted to evaluate structural components' performance on the query's topic.

Among the components of QA documents, the answer-document set showed consistently the best performance in two experiments: one is the QA document clustering experiment and the other the query reconstructing experiment. Even though the answer-document was not originated from the primary user who had information need, it was most effective to present his information need.

In the query reconstructing experiment, queries formed by selecting sentences in answers provided the most relevant and various documents in search results. As more relevant documents were retrieved, information redundancy was also increased. It is not easily achieved to reduce redundancy in search results without decreasing precision. In this aspect, answer-document set should be more weighted than other sets because it performed effectively at two criteria, precision rate and redundancy rate measured by entropy.

QA documents should also be treated with understanding their own structure. In particular, the body of QA documents consists of two different components written by two different users' groups. Question-description set is written out by a questioner who might not be able to describe his or her information need effectively and answer set is written out by other user who might be able to express questioner's information need better than the questioner. For that reason, the structural characteristics of QA documents should be considered before attempting to improve retrieval performance. Answer set in the QA documents could also play a role like user's feedback. Answers in the QA documents are reviewed by the questioner so that it could be considered as the result of primary user's relevance evaluation.

This study was, however, conducted in few topic areas and limited number of documents. On the account of this limitation, various attempts in other topic areas with larger number of dataset are suggested. In addition, the query reconstruction method suggested in this study could require the certain condition of QA document retrieval system because the similar queries are necessary to reconstruct a new query. This query reconstruction method, therefore, is not suitable for the QA retrieval system which has not accumulated sufficient queries. However, this method is recommendable for large QA document retrieval system to enhance the retrieval precision performance and reduce the redundancy in the search results.

Reference

Berztiss, A. T. 1993. "The Query Language Vizla". IEEE TKDE 5(5): 813 - 825.

Bloesch, A. C. and Halpin, T. A. 1997. “Conceptual Queries Using Conquer II”. *Proceedings of the ER'9*7: 16th International Conference on Conceptual Modeling(Los Angeles). 112-126.

Choi, Sang Hee. 2005. "A Study On Clustering Query-answer Documents with Structural Features". *Journal of the Korean Society for Library and Information Science.* 39(4): 105-118.

Chung, Young Mee and Lee, Jae Yun. 2001. "Development of a Clustering Model for Automatic Knowledge Classification" *Journal of the Korean Society for information Management*.18(2): 203-230.

Gopal, Ram D. and Ramesh, R. 1995. "The Query Clustering Problem: A Set Partitioning Approach," *IEEE Transactions on Knowledge and Data Engineering,* 7(6): 885-899

Han, Eui-hong. 1998. "WebACE: a web agent for document categorization and exploration" *In Proceeding of the 2nd International Conference on Autonomous Agents.*

Kang, In-Ho and Kim, GilChang. 2003. "Query Type Classfication or Web Document Retrieval" In *Proceedings of the 26th Annual International ACM SIGIR Conference,.* July 28 - August 1, 2003,Toronto, Canada. p 64-71.

Kim, Jung Ha, and Lee, Jae Yun. 2000. "A Comparative Study on Performance Evaluation of Document Clustering Results" *7th Proceedings of the Korean Society for Information Management Conference.* p45-50.

Maybury, Mark. T.2004. *New Directions in Question Answering*. Menlo Park: AAAI prss.

Owei,Vesper. 2002. "An Intelligent Approach to Handling Imperfect Information in Concep-Based Natural Language Queries". *ACM Transaction on Information Systems*, 20(3): 291-328.

Tombros, Anastasios and Sanderson, Mark. 1998. "Advantages of Query Biased Summaries in Information Retrieval" In *Proceedings of the 21st Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval,* 2-10.

Tombros, Anastasios, Villa, Robert and Rijsbergen, C. J. Van. 2002. "The Effectiveness of Query-Specific Hierarchic Clustering in Information Retrieval," *Information Processing & Management,*. 38(4): 559-582

Wen, Ji-Ron, Nie, Jian-Yun, and Zhang, Hong-Jiang. 2001. "Clustering User Queris of a Serach Engine," In *Proceedings of WWW10,* 162-168

Zhang, Ya-Jun and Liu, Zhi-Qiang. 2004. "Refining Web Search Engine Results Using Incremental Clustering," *International Journal of Intelligent Systems*, 19(2): 191-199