문서구조를 이용한 질의응답문서 클러스터링에 관한 연구

A Study on Clustering Query-answer Documents with Structural Features

최 상 희(Sanghee Choi)*

목 차

초 록

이용자가 직접 참여하여 질의를 제공하고 응답을 하면서 공동으로 지식을 생산해내는 형태의 정보서비스는 응답된 내용이 축적되어 가면서 새로운 대용량 정보검색 분야로 성장하고 있다. 이 연구에서는 질의와 응답이 결합되어 있는 질의응답문서의 구조적 특성을 반영하여 질의응답문서의 효율적인 이용 방안을 도모하고자, 문서 구성요소인 질의제목, 질의설명, 응답내용을 기반으로 클러스터를 자동 생성하여 수작업 주제 범주와 비교한 후 각 구성요소의 주제 표현 성능을 평가하였다. 실험 결과 응답내용 클러스터링 결과가 수작업 결과와 가장 유사한 것으로 나타나 응답내용이 문서의 주제를 표현하는데 효과적인 것으로 분석되었다.

ABSTRACT

As the number of users who ask and give answers in the query-answer documents retrieval system is growing exponentially, the query-answer document become a crucial information resource, as a new type of information retrieval service. A query-answer document consists of three structural parts: a query, explanation on query, and answers chosen by users who asked the query. To identify the role of each structural part in representing the topics of documents, the three structural parts were clustered automatically and the results of several clustering tests were compared in this study.

키워드: 클러스터링, 질의응답문서, 질의 클러스터링, 문서 클러스터링
Clustering, Query-answer Documents, Query Clustering, Document Clustering

* 연세대학교 문헌정보학과 강사(tudultudul@naver.com)
논문접수일자 2005년 11월 15일
게재확정일자 2005년 12월 15일

1. 서 론

많은 사람들이 정보를 교환하는 네트워크 정보환경이 발전해가면서 다양한 지식이 여러 가지 형태로 축적되고 있다. 특히, 인터넷과 같이 특별히 정보생산자가 제한되어있지 않은 환경에서 다양한 이용자들이 스스로 정보요구를 표현하고 그에 적합한 정보를 다른 이용자에게서 구하는 형태의 정보생산체제가 늘어나고 있다. 각종 웹사이트에서 제공하는 FAQ나 Naver의 지식iN과 같이 이용자가 직접 참여하여 질의를 제공하고 응답을 하여 공동으로 지식을 생산해 내는 질의응답형 지식검색 서비스가 이에 해당하는 정보생산체제라고 할 수 있으며 주요 정보원으로 점점 더 많은 주목을 받고 있다. 그러나 이와 같은 질의응답 서비스에 많은 질의와 답변이 축적되기 시작하면서 이용자가 원하는 내용을 찾기가 점점 어려워지고 있다.

웹 검색이 아닌 질의응답형 지식검색서비스 내에서 검색된 결과만 해도 수천 건이 넘게 나오는 경우가 종종 있으며 결과적으로 유사한 질의와 답변도 같이 늘어나게 된다. 즉, 한 질의에 여러 답변이 중복되어 제공되는 사례도 많이 발생하지만 질의 자체도 중복되는 경우가 많아 응답정보가 중복되는 사례가 발생하게 된다.

질의응답 문서는 일반문서와 구조적으로 다른 특성을 나타내고 있다. 문서의 제목이 질의에 해당하며 문서내용에는 질의내용을 보충한 질의설명과 그 질의에 대한 정보인 응답내용이 혼합되어 있기 때문이다.

이 연구에서는 질의와 응답이 결합되어 있는 문서환경에서 문서의 구조적 특성을 반영하여 질의응답문서의 효율적인 이용 방안을 도모하고자 한다. 이 연구에서는 계층적 클러스터링 기법을 적용한 문서 클러스터링을 적용하여 질의응답문서의 구조적 특성을 분석하였다. 질의응답문서를 수작업으로 주제 적합성을 판정하여 범주화 한 후 문서 구성요소인 질의제목, 질의설명, 응답내용을 기반으로 클러스터를 자동생성하고 수작업 주제 범주와 비교하여 각 구성요소의 주제 표현 성능을 평가하였다.

2. 연구배경

질의와 관련된 정보검색 연구들은 대부분 질의에 초점을 맞추고 있고 검색효율을 높이기 위한 시도라고 할 수 있다(Gopal and Ramesh 1995; Zhang and Liu 2004). 검색질의의 개념을 파악하여 질의간 의미관계를 파악하여 이용자가 찾고자 하는 정보의 핵심을 파악하려는 연구의 결과로 ConQuer II는 이미 상용화되어 있다. ConQuer II는 객체와 개체간의 의미간 관계를 파악하여 구축된 개념체계에서 질의를 생성하는 방식을 적용하고 있다(Bloesch and Halpin 1997). 이와 비슷한 방식으로 Vizla라는 시스템은 개체와 속성 관계를 분석하여 질의를 단계별로 생성할 수 있도록 지원하고 있다(Berztiss 1993). 이용자가 입력하는 자연어 질의를 개념적으로 처리하려고 시도한 오웨이는 질의가 불완전한 정보로 구성되어 있어 검색효율을 떨어트리는 요인이 된다고 하였다(Owei 2002). 따라서 검색 질의의 개념간 의미관계를 파악하여 질의를 재생성함으로써 질의의 불완전성을 해소하려 하였다.

웹에서 사용되는 이용자의 질의를 처음부터 용도별로 분류하여 검색되는 과정을 질의 유형에 맞추어 검색효율을 높이려는 시도를 한 연구도 있다(Kang and Kim 2003). 이 연구에서는 이용자의 질의를 정보검색, 네비게이션, 트랜잭션 등 세 가지 유형으로 나누었다. 첫 번째 유형은 주제정보를 찾아 적합성을 판정하는 것이 핵심이고, 두 번째 유형은 홈페이지를 찾는 것이 주요 목적이며 트랜잭션은 서비스를 찾는 것이 목적이다. 즉, 찾고자 하는 유형에 따라 질의를 처리하는 방식을 달리하여 검색결과의 정확도를 향상시키고자 하는 것이다.

최근 클러스터링 기법은 질의를 처리하는데도 적용되고 있어 질의 클러스터링은 클러스터링 연구 분야로 주목을 받기 시작했다. 질의 클러스터링은 이용자의 질의를 군집화하는 과정으로 톰브로스는 질의를 기반으로 하여 검색결과를 클러스터링 하는 연구를 수행하였다(Tombros, Villa, and Rijsbergen 2001). 그러나 이런 연구 시도들은 모두 검색 질의 중심으로 이루어진 것으로 질의응답문서를 검색하는 환경에서의 질의에 해당하는 것이라 보기 힘들다. 질의응답문서에서 질의를 클러스터링 한다는 것은 오히려 문서의 제목으로 클러스터링 하는 방식과 비교하는 것이 더 적절할 것이다.

웬의 연구에서는 질의를 클러스터링 하는데 이용자 로그를 이용하고 있다. 이 연구는 이용자가 어떤 문서를 보았는지 추적하여 질의의 유사성을 파악하고자 하는 것이었다(Wen, Nie, and Zhang 2001). 즉, 두 이용자가 유사한 문서를 결과로 보았다면 두 이용자의 질의가 유사한 것으로 여기는 방식이다. 이 연구에서는 결과로 본 문서를 기반으로 질의의 주제성을 파악하고 하는 측면에서 기존의 연구와 다르다고 할 수 있다.

질의응답문서는 일반적으로 질의어로 검색되는 문서와는 다른 환경에 속한다고 할 수 있다. 일반 문서검색에서 질의어로 검색하면 검색된 문서가 문서제목과 문서내용으로 구성되어 있는 것과는 달리 질의응답문서 검색환경에서 질의어로 검색된 문서에는 기존에 등록되어 있는 질의, 질의설명, 응답내용이라는 요소가 결합되어 있다. 즉답형 참고질의(ready reference) 또는 FAQ, 지식검색결과와 같은 형태로 나타나고 있는 질의응답문서(query-answer documents)는 일반적으로 질의제목, 질의내용를 표현한 질의설명, 그 질의에 해당하는 답변으로 구성되어 있다. 〈예 1〉을 보면 질의제목과 구체적인 질문 사항을 정리한 질의내용, 그리고 질의자가 자신이 한 질의에 적합하다고 판단한 답변내용이라고 선택한 답변으로 구성되어 있다. 따라서 일반 문서검색에 적

〈예 1 질의응답문서 구조〉

질의제목: 집에서의 무선인터넷 사용

질의설명: 데스크톱 1대, 그리고 노트북 1대가 있습니다. 노트북은 현재 무선 랜카드 사용하고 있구요, 집에서 인터넷을 두대 다 사용 하려고 하는데 넷스팟은 너무 비싸서, 지식인을 대충 보니 공유기를 쓰면 된다고 하시는 것 같은데 어떤 공유기를 구입해야 하는지(저렴한 놈으로다가.) 설정 복잡하면 진짜 난감하거든요. ~~~~~~

답 변 : 보통 공유기는 선을 꼽아 신호를 보내어 컴퓨터와 무선으로 통신하는 건데요 공유기에 케이블을 꼽고 그에 맞는 랜카드를 넣으면 되여… 그때 공유기는 외부에서 하면 안되여… 전마야 약해지니까요. ~~~~~~~~~~

용되는 질의 연구나 문서 클러스터링 연구와는 차별화되어야 한다. 특히 지식검색과 같은 질의응답문서 검색이 점점 늘어나고 있기 때문에 이러한 구조적인 특성을 기반으로 질의응답문서 검색의 효율성을 도모하는 연구가 수행되어야 할 필요성이 대두되었다.

3. 연구대상과 방법

3. 1 실험집단 및 클러스터링 실험 단계

실험대상이 된 문서집단을 추출한 대상은 네이버 지식iN으로 주제분야는 컴퓨터, 통신분야이다. 네이버 지식iN에서 검색되는 문서는 질의제목, 질의설명, 질의자가 답변으로 선택한 응답내용으로 구성되어 있어 질의응답문서 구조를 나타내고 있다. 실험집단의 주제 범위는 컴퓨터, 통신 분야로 제한을 하였다. 대주제분야가 크게 다르게 실험집단 문서를 추출할 경우 문서 내 중복되는 용어가 거의 없게 되므로 결과적으로 클러스터링의 성능이 높게 나타날 것이 예측되기 때문이다. 예를 들어 어린이 건강을 주제로 다루는 문서와 컴퓨터에 윈도우 설치를 다루는 문서가 같은 클러스터에 배정될 가능성은 낮을 것이다. 또한 실제 이용자가 이 두 주제의 문서를 한 검색의 결과에서 보게 될 가능성도 매우 낮으므로 클러스터링 연구 결과를 적용하기에도 적절하지 않다. 따라서 같은 용어를 사용하면서 이용자가 검색을 하였을 때 검색결과에 뒤섞여 나올 가능성이 많은 문서집단을 설정하고자 대주제분야를 한정하여 그 주제분야 내에서 세부주제를 선정, 실험집단을 구축하였다.

네이버 지식iN에서도 컴퓨터, 통신분야의 하위 주제 카테고리를 제시하고 있지만 하위 주제카테고리는 실제 이용자가 선정하는 것으로 같은 주제의 질문이 여러 카테고리에 분산되어 나타나는 경우가 많았다. 예를 들어 프린터에 문제 해결을 문의하는 질의가 주변기기 카테고리가 아닌 하드웨어나 운영체제 등의 다른 카테고리에서 검색되기도 한다. 따라서 기존의 카테고리를 반영하기는 하였으나 〈표 1〉에서 기술한 주제별로 제시한 키워드로 네이버 지식iN에서 다시 검색을 하여 검색결과에서 실험집단 주제에 맞는 문서를 선정하였다.

〈표 1〉 실험집단 주제 범주

주제범주번호	주제	검색키워드
M1	프린터 문제 해결	프린터, 잉크젯, 레이저젯, 문제, 오류
M2	인터넷 파일공유	인터넷, 네트워크, 파일, 공유
M3	이메일 문제 해결	이메일, 전송
M4	이동통신, 핸드폰 기기	이동통신, 핸드폰, 휴대전화, 신상품
M5	윈도우 설치	윈도우, 설치, 문제
M6	웹페이지 제작	웹, 웹페이지, 태그, 제작, 작성
M7	보안 프로그램 설치 및 바이러스 해결	보안, 바이러스, 백신, 설치
M8	무선네트워크 연결	무선네트워크, 무선랜, 연결, 접속
M9	동영상 재생 문제	동영상, 재생, 플레이어,
M10	노트북 추천	노트북, 신상품, 비교, 추천

실험집단에 맞는 문서 선정은 총 2인이 〈표 1〉 주제범주에 맞추어 주제를 평가한 결과에 따라 생성하였다. 그 과정을 살펴보면 먼저 주제범주별 검색 키워드로 검색된 결과에서 실험집단으로 선정할 30건의 문서 2배수인 60건의 문서를 추출하였다. 각 범주 주제별로 추출된 문서를 실험자와 일반 이용자 1인이 내용을 읽어보고 〈표 1〉의 각 주제범주에 해당한다고 판단된 문서를 표시한 다음 두 사람이 특정 주제범주에 속한다고 일치하게 평가한 문서를 순차적으로 30건을 선정하였다. 그 결과 각 주제범주별로 30건씩 총 300건의 문서집단이 생성되었다. 이 문서집단은 구축과정에서 수작업으로 분류되었으므로 자동클러스터링 결과 생성된 문서집단과 비교하는데 사용되었다.

실험집단으로 선정된 질의응답문서 집단을 문서 내 구조로 분할하여 크기를 분석한 결과 〈표 2〉와 같다. 그 중 질의설명의 경우는 평균 395자이나 20자 미만의 질의설명을 입력한 경우도 있었다. 또한 응답내용보다 많은 질의설명이 있는 경우도 발생하여 800자가 넘는 등 질의제목이나 응답내용의 문자 수와 비교하였을 때 상대적으로 글자 수의 편차가 많은 편이다.

문서 구조별 평균 문자 수

〈표 2〉 실험집단 구성요소별 평균 문자 수

	질의제목	질의설명	응답내용
평균 문자 수	88자	395자	818자

클러스터링 실험을 위해 질의응답문서는 질의제목, 질의설명, 응답내용으로 분할하였다. 각 분할된 문서구성요소는 다섯 가지 방식으로 적용, 조합하여 클러스터링 대상으로 설정하였다. 구성요소를 적용한 방식별로 각각 클러스터링을 한 후 결과를 비교하여 보았다.

- 클러스터링 대상
  - 질의제목
  - 질의설명
  - 응답내용
  - 질의제목 + 질의설명
  - 질의제목 + 응답내용

3. 2 클러스터링 기법 및 평가 방법

클러스터링하기 위해 문서에서 추출한 용어에 적용한 용어 가중치는 단어빈도 × 역문헌빈도이다. 역문헌빈도는 스파크존스의 역문헌빈도를(Spark Jones 1972)를 적용하였다. 클러스터링 기법으로 적용한 것은 워드 기법인데 이 기법은 되도록이면 같은 크기로 클러스터를 생성시키려는 성향을 보이고 있고 클러스터에 할당된 멤버 수의 편차가 상대적으로 적다는 연구결과가 제시되어 있다(최상희 2004; Millgan, Soon, and Sokol 1983). 실험집단은 이미 수작업으로 각 범주별로 30건의 문서가 고르게 할당되어 있기 때문에 수잡업 분류와 유사하게 클러스터 결과를 생성시키려면 되도록이면 고르게 클러스터를 생성시키는 기법이 적절하다고 분석되었다.

클러스터링 생성 시 대상 간 유사도비교를 할 때 적용되는 유사도 공식은 코사인계수를 적용하였다. 코사인계수는 유사도를 측정하는데 가장 일반적으로 적용되는 계수로서, 문장간 유사도를 측정하는 경우 두 문장간의 중복되는 값을 전체 문장길이로 정규화해주는 역할을 한다.

생성된 클러스터링을 평가하는 척도는 문헌 단위의 평가척도인 클러스터링 정확률, 재현율, F 척도와 비편향적 단일척도인 WACS (Weighted Average Cluster Similarity)를 적용하였다. 클러스터링 정확률과 재현율은 정보검색에서의 정확률과 재현율을 응용한 것으로, 특정 문헌을 질의로 적용한다. 그 다음 특정 문헌이 속하여 있는 수작업 주제 범주를 적합문헌 집단으로 규정하고 자동으로 생성된 클러스터 중 그 문헌이 포함된 것을 검색된 문헌집단으로 하여 각각 정확률과 재현율을 계산하는 것이다. 문헌 D_k가 배정된 수작업 범주가 M_i이고 자동생성된 클러스터가 C_j 일때 문헌별 정확률과 재현율을 계산하면 아래와 같다(정영미, 이재윤 2002).

$$\text{정확률}(D_k) = \frac{|M_i \cap C_j|}{|C_j|}$$

$$\text{재현율}(D_k) = \frac{|M_i \cap C_j|}{|M_j|}$$

클러스터링 정확률과 재현율은 각 문헌에 대해 구한 정확률과 재현율의 평균이 된다.

$$\text{클러스터링 정확률(p)} = \frac{1}{D}\sum_{j=1}^{n}\sum_{i=1}^{m}\frac{|M_i \cap C_j|^2}{|C_j|}$$

$$\text{클러스터링 재현율(r)} = \frac{1}{D}\sum_{j=1}^{n}\sum_{i=1}^{m}\frac{|M_i \cap C_j|^2}{|M_j|}$$

d : 전체문헌수
n : 자동생성 클러스터링의 수
m : 수작업 범주의 수

F척도는 문헌별 정확률과 재편율의 평균을 반영한 복합척도로서 클러스터 크기의 편차에 크게 영향받지 않는다는 특성이 있다.

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

비편향적 단일척도 WACS는 정영미, 이재윤(2002)이 클러스터링 평가를 위하여 고안한 척도로서 비교 대상인 수작업 분류 범주와 자동생성 클러스터를 각각 소속 문헌의 벡터로 표현한 다음 유사한 정보를 벡터 유사도 공식으로 계산한 것이다. 즉, 범주와 클러스터간 유사도를 계산하여 각각의 크기를 고려한 가중편균을 산출하면 전체의 유사도가 산출된다. 즉, 자동생성된 특정 클러스터 C_j에 배정된 문헌이 하나 이상 속한 수작업 분류 범주를 검색하여 각각의 유사도를 계산한 후 클러스터와 수작업 분류 범주와 중복되는 문헌 수를 반영하여 가중 평균을 산출한다. 이 공식에서 적용된 유사도 계수는 다이수 계수이다.

$$\text{유사도 Sim}(M_i, C_j) = \frac{2|M_i \cap C_j|}{|M_i| + |C_j|}$$

클러스터 C_j에 대한 WACS 공식과 클러스터링 전체 성능에 대한 WACS 공식은 아래와 같다.

$$\text{WACS}(C_j) = \frac{1}{|C_j|}\sum_{i=1}^{m}\frac{2|M_i \cap C_j|^2}{|M_i| + |C_j|}$$

전체 성능에 대한 WACS 공식은 각 클러스

터별로 산출된 값을 기반으로 하여 가중 평균을 구한 것이다.

$$WACS(C) = \frac{1}{D}\sum_{j=1}^{n}\sum_{i=1}^{m}\frac{2|M_i \cap C_j|^2}{|M_i|+|C_j|}$$

4. 문서구조를 이용한 질의응답문서 클러스터링 분석결과

1) 질의제목

〈표 3〉을 보면 질의제목을 기반으로 클러스터링 한 결과 웹페이지 제작(m6)에 관련된 클러스터만 변별되게 다른 클러스터에 할당되는 현상이 나타났다. 이동통신(m4)와 보안바이러스(m7)는 멤버의 대부분이 C1으로 배정되었고 파일공유(m2) 범주는 3개의 클러스터에 분할되었다. 나머지 주제 범주의 경우 대부분 C1과 나머지 클러스터에 이원화되는 결과를 나타내었다. 따라서 전체의 55%의 질의제목들이 한 클러스터로 할당되는 현상이 드러났다.

2) 질의설명

질의설명을 기반으로 클러스터링 한 결과, 질의제목에서 독립적으로 한 클러스터로 할당되었던 웹페이지 제작(m6) 범주가 다시 여러 주제가 섞인 가장 큰 클러스터에 혼합되는 현상이 나타났고 무선네트워크(m8)은 두 번째 큰 클러스터를 형성하면서 노트북(m10) 범주와 같은 클러스터에 할당되었다. 〈표 4〉를 참

〈표 3〉 질의제목 클러스터링 결과

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
m1	19	0	0	1	0	0	0	0	10	0
m2	11	0	0	1	8	0	9	1	0	0
m3	17	0	0	1	0	0	0	0	0	12
m4	27	1	0	1	1	0	0	0	0	0
m7	29	0	0	0	1	0	0	0	0	0
m8	18	3	0	1	0	8	0	0	0	0
m10	10	20	0	0	0	0	0	0	0	0
m9	13	0	17	0	0	0	0	0	0	0
m5	14	0	0	16	0	0	0	0	0	0
m6	6	0	0	0	0	0	0	24	0	0

〈표 4〉 질의 설명 클러스터링 결과

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
m2	1	29	0	0	0	0	0	0	0	0
m4	5	17	1	0	0	0	7	0	0	0
m5	0	19	0	0	0	0	0	8	0	3
m6	2	23	0	0	0	0	0	0	5	0
m7	2	28	0	0	0	0	0	0	0	0
m9	0	30	0	0	0	0	0	0	0	0
m8	0	4	26	0	0	0	0	0	0	0
m10	1	12	17	0	0	0	0	0	0	0
m3	0	10	0	20	0	0	0	0	0	0
m1	0	2	0	0	10	18	0	0	0	0

조하면 노트북(m10), 이메일(m3), 프린터(m1) 주제 범주는 범주내 질의 설명이 1/3과 2/3씩 두 클러스터에 양분되는 현상이 나타났다.

질의설명은 평균문자 수가 395자로 질의제목(88자)보다 평균문자 수가 거의 4배정도 늘어난 것으로 클러스터링 대상이 되는 자질 수가 커지므로 클러스터링의 성능이 향상될 것을 기대했었다. 또한 주제를 나타내지 못하는 질의(예 2 참조)의 경우 질의 주제를 나타내는데 더 효과적인 것으로도 예측되었다.

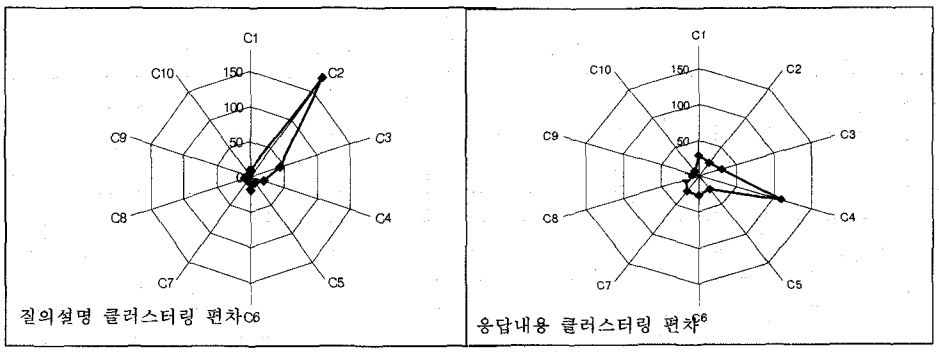
〈예 2〉 질의설명의 질의제목 주제 설명 보충 사례

질의제목: 가르쳐 주세요...
질의설명: 네스팟을 사용하려고 합니다… 제가 거주하는 원룸에서… 친구의 노트북(무선랜내장)을 사용하면 네스팟이 연결되고… 저의 노트북(무선랜을 사서 장착)은 접속이 되질 않습니다… 신호가 상당히 약합니다… 다들 무선랜이 좋지 않다고 하는데요~~ 혹시 AP신호를 증폭할 수 있는 그런 장비도 있는지요~~ 자주 옮겨다녀야하는 직업으로서 무조건 무선랜을 교환하는 것 보다… AP신호를 증폭할 수 있다면… 다음에 다른 원룸으로 옮길때에도 많은 도움이 될 것 같아서 그러는데… 아시는 분들 계시면… 꼭 알려주세요…

질의설명이 질의제목을 보충하여 주제를 표현하는데 효과적일 수 있다는 예측과는 달리 클러스터링 결과 전체 질의설명의 58%가 C2에 할당되는 현상이 나타나 질의설명은 질의제목보다 클러스터를 분할하지 못하는 결과를 보였다. 이는 질의에 대한 보충 설명을 하는 과정에서 추가된 용어들이 주제성을 띄지 않는 일반적인 용어이고 질문을 설명하기 위해 주변 상황에 대해 포괄적으로 기술하는 과정에서 다른 주제와 연관된 설명을 추가하는 경우가 발생한 것으로 분석될 수 있다. 질의설명에서 클러스터링 자질은 증가되었으나 증가된 자질이 주제를 변별하는 기능이 없는 용어들이어서 클러스터링 성능향상에는 효과적이지 못하였다.

3) 응답내용

응답내용을 클러스터링을 한 결과 〈그림 1〉을 보면 질의제목과 질의설명에서 한 클러스터에 전체 문서집단의 문서 1/2 이상이 할당되었던 현상이 완화되고 상대적으로 클러스터 크기 편차가 줄어드는 성향을 보였다. 응답내용은 질의설명의 2배로 자질이 늘어났기 때문에 클



〈그림 1〉 질의설명과 응답내용 클러스터링 편차 비교

러스터를 분할하는 성능이 향상된 것으로 분석될 수 있으나 질의설명이 질의제목보다 자질이 4배 늘어났어도 클러스터를 분할하지 못한 결과를 나타냈으므로 자질 수의 단순 증가가 주요 요인이라고 할 수 없다. 즉, 질의설명은 질의제목보다 자질이 늘어났지만 주제 변별력이 없는 일반적인 내용이 추가된 결과를 보여주었지만 응답내용은 자질 수가 늘어났을 때 클러스터 분할에 영향을 미치는 주제를 표현하는 용어들이 증가된 경우라고 할 수 있다. 응답내용을 클러스터링 한 결과, 수작업 범주와 대응되는 정도가 질의제목과 질의설명보다 향상된 것으로 나타났다.

파일공유(m2), 보안바이러스(m7), 노트북(m10) 3개 범주에 해당하는 문서는 가장 큰 클러스터에 대부분 할당되었지만 나머지 범주는 한 범주의 2/3가 넘는 멤버가 소속되어 대응되는 클러스터가 생성된 것으로 나타났다. 예를 들어 〈표 5〉를 보면 자동생성된 클러스터 C1은 주제범주 무선네트워크(m8)에 대응한다고 할 수 있다. 문서의 응답내용을 클러스터링 하면서부터 자동생성된 각 클러스터에 주제 범주가 연결되기 시작하였다.

4) 질의제목과 질의설명

질의응답문서 구조에서 제목과 질의부분에 해당하는 질의제목과 질의설명을 결합하여 클러스터링을 한 결과 질의제목과 질의설명에서

〈표 5〉 응답내용 클러스터링 결과

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
m8	27	0	1	2	0	0	0	0	0	0
m9	0	22	0	8	0	0	0	0	0	0
m1	0	0	28	2	0	0	0	0	0	0
m2	1	0	0	28	0	1	0	0	0	0
m7	0	0	0	21	0	0	0	0	0	9
m10	1	0	0	29	0	0	0	0	0	0
m6	0	0	0	6	23	0	1	0	0	0
m3	0	0	1	0	0	26	3	0	0	0
m4	0	0	0	0	0	0	22	0	8	0
m5	0	0	0	13	0	0	0	17	0	0

〈표 6〉 질의제목+질의설명 클러스터링 결과

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
m2	1	19	0	0	0	10	0	0	0	0
m4	5	17	1	0	0	0	7	0	0	0
m5	0	17	0	0	0	0	0	13	0	0
m7	2	28	0	0	0	0	0	0	0	0
m8	0	4	26	0	0	0	0	0	0	0
m10	1	11	18	0	0	0	0	0	0	0
m3	0	10	0	20	0	0	0	0	0	0
m1	0	2	0	0	10	18	0	0	0	0
m9	0	13	0	0	0	0	0	0	17	0
m6	2	4	0	0	0	0	0	0	0	24

크기가 큰 클러스터가 다른 소수멤버의 클러스터로 다소 분할되는 현상을 보였다. 전체 50%가 넘는 멤버가 몰렸던 클러스터가 전체 41% 수준으로 감소되었다. 그러나 〈표 6〉에서 분할된 현상을 보면 주제범주 프린터(m1)와 동영상문제(C9), 이메일(M3)의 경우 2개의 클러스터에 범주내 멤버의 1/3과 2/3가 양분되는 결과가 나타났다. 또한 자동생성된 클러스터 C3과 C6의 경우 두 개 범주가 같이 할당된 현상이 드러났다. 따라서 질의제목과 질의설명을 결합하여 클러스터링을 하였어도 클러스터를 수작업 주제범주에 대응되도록 생성시키지 못한 것으로 나타났다.

5) 질의제목과 응답내용

질의응답문서 구조에서 제목부분에 해당하는 질의제목과 응답부분에 속하는 응답내용을 결합하여 클러스터링을 한 결과, 응답내용을 클러스터링 했을 때 3개의 주제 범주가 한 클러스터에 묶였던 것을 2개 범주로 나타나 전체 실험집단의 36%에 해당하는 문서가 한 클러스터에 할당되었던 현상이 전체 27% 수준으로 감소되었다. 또한 자동생성된 각 클러스터와 주제범주를 대응하여 비교하였을 때(표 7에서 회색으로 표시된 셀 부분 참조) 자동생성된 클러스터 C4와 C10을 제외하고는 대부분 대응되는 수작업 주제 범주가 있는 것으로 나타났다. 응답내용만을 클러스터링 했을 때는 파일공유(m2), 보안바이러스(m7), 노트북(m10) 3개 범주가 한 클러스터에 할당되었었는데, 질의제목과 결합한 결과에서는 파일공유(m2)가 새로운 클러스터로 분할되어 클러스터 분할 성능이 질의제목에 의해서 향상된 것으로 나타났다.

6) 문서구성 요소별 클러스터링 성능 비교

질의응답문서 구조에 따라 분할한 질의제목, 질의설명, 응답내용을 기반으로 각각 클러스터링 한 결과와 질의제목과 질의설명, 응답내용을 조합하여 클러스터링한 결과를 비교하여 성능 평가한 것은 〈표 8〉과 같다.

단일 구성요소로는 응답내용이 클러스터링 정확률(precision)과 클러스터링 재현율(recall), F척도, WACS 모든 평가 척도에서 가장 성능이 좋은 것으로 나타났다. 응답내용은 질문을 한 이용자가 자신이 문의한 사항에 적절한 답을 제공하였다는 것을 선정한 것으로 실제 질의에 대한 적합성 피드백 정보에 해당한다고 할 수 있다. 따라서 문의하는 사람이 자신도

〈표 7〉 질의제목+응답내용 클러스터링 결과

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
m8	27	0	1	2	0	0	0	0	0	0
m9	0	22	0	8	0	0	0	0	0	0
m1	0	0	28	2	0	0	0	0	0	0
m7	0	0	0	21	0	0	0	0	0	9
m10	1	0	0	29	0	0	0	0	0	0
m6	0	0	0	6	23	0	1	0	0	0
m3	0	0	1	0	0	26	3	0	0	0
m4	0	0	0	8	0	0	22	0	0	0
m5	0	0	0	4	0	0	0	26	0	0
m2	1	0	0	0	2	1	0	0	26	0

〈표 8〉 문서구성 요소별 클러스터링 성능 비교

	PRECISION	RECALL	F	WACS
질의제목	0.45	0.57	0.50	0.38
질의설명	0.40	0.66	0.50	0.36
응답내용	0.65	0.72	0.68	0.61
질의제목 + 질의설명	0.40	0.58	0.53	0.45
질의제목 + 응답내용	0.73	0.73	0.73	0.70

표현하지 못했던 내용을 응답자가 오히려 적확하게 표현했을 가능성이 있다. 즉, 잘 모르고 있는 내용을 기반으로 질의설명을 기술하였지만 제시된 답변을 선택할 때는 모호했던 질의설명이 구체화된 응답내용을 맞는 답변이라고 선택하게 된다. 그 결과 응답질의문서의 주제가 응답내용에서 더 정확하게 표현될 가능성이 있게 되는 것이다. 클러스터링 성능 평가 결과도 이와 같은 가능성을 반영한 것으로 응답내용만 클러스터링 한 결과와 질의제목 + 응답내용을 조합하여 클러스터링한 결과가 질의부분만 이용하여 클러스터링 한 세 결과 보다 수작업 주제범주와 대응한 성능평가에서 모두 높은 것으로 나타났다. 따라서 질의응답문서의 경우 문서의 주제를 파악하는데 있어 응답부분이 가장 중요한 역할을 한다고 분석되었다.

질의제목과 질의설명은 상대적으로 특히 클러스터링 정확률과 WACS 척도에서 낮은 성능을 보여 질의설명의 경우 응답내용의 WACS 척도 값의 절반에 해당하는 성능을 보이고 있었다. 질의제목과 질의설명을 비교하면 질의설명은 질의제목에 비해 클러스터링 재현율은 크게 향상되었지만 반면 클러스터링 정확률은 오히려 떨어지는 것을 볼 수 있다. 클러스터링 분할 결과에서도 나타났듯이 질의설명은 질의응답문서를 주제범주별로 분할하기 보다는 한 클러스터로 모으는 성향을 보였다. 즉, 질의설명에는 다른 주제범주의 문서에서도 출현하는 비주제성 일반 용어들이 많이 포함되어 있어 그런 용어들을 기반으로 클러스터링을 하기 때문에 문서들을 한 클러스터로 모으게 되는 것이다. 그 결과 클러스터링 정확률을 저하시키고 클러스터링 재현율을 높이는 현상을 보이게 된 것이다.

질의응답문서의 제목에 해당하는 질의제목은 질의설명과 응답내용에 결합했을 때 모두 클러스터링 성능 향상이 있는 것으로 나타났다. 질의설명의 F척도 값이 0.5, WACS 값이 0.36에 해당하였는데 질의제목과 결합한 결과 각각 0.53과 0.45로 나타나 두 척도면에서 모두 향상되었다. 또한 응답내용의 F척도 값은 0.68, WACS 값이 0.61에 해당하였는데 질의제목이 적용되자 각각 0.73과 0.70으로 나타나 역시 향상된 결과를 보여주었다. 따라서 질의제목은 질의하고자 하는 내용의 주제를 표현하는 주제어 기능이 있는 것으로 분석된다.

5. 결 론

질의응답문서의 구조를 분할하여 클러스터링한 결과, 자동 생성된 클러스터와 수작업 범

주의 대응도와 클러스터링 정확률, 클러스터링 재현율, F척도, WACS 등 전반적으로 모든 측면에서 응답내용이 가장 수작업 범주에 가까운 성능을 보이는 것으로 나타났다. 즉, 응답내용이 질의응답문서 내에서 주제를 표현하는 역할을 가장 효율적으로 하고 있는 것이다. 질의제목은 전체 문서의 주제를 표현하는데 있어 자질 수가 충분하지 못해 주제별로 클러스터를 분할하는데 효과적이지 못한 반면 질의설명은 자질측면에서 질의제목보다 월등히 많음에도 불구하고 질의주제를 나타내고 있지 못한 것으로 나타났다.

이용자가 질의를 한다는 것은 자신이 물어보는 내용에 대해 정확히 알고 있지 못하기 때문에 문의를 하는 것이므로 질의설명을 질의제목보다 길게 설명하다고 해서 질의 주제를 표현하는 내용을 효과적으로 추가하지 못하는 것으로 분석되었다. 따라서 질의응답문서 내에서는 질의자의 질의하는 의도가 자신이 기술한 부분보다는 다른 사람의 응답내용 속에서 더 잘 표현되어 있는 것이다. 이러한 분석 결과는 질의응답문서를 이용하여 향후 연구를 할 때 다양한 분야에 적용될 수 있다.

첫째, 질의응답문서를 검색하는 지식검색 서비스에서 검색된 결과를 클러스터링하여 중복된 정보를 처리해 줄 때 응답내용에서 추출된 자질을 중심으로 처리할 수 있다. 질의제목에서 질의설명과 응답내용으로 자질이 늘어났을 때 증가된 클러스터링 자질 중 응답정보의 자질이 클러스터링 성능 향상에 더 많은 영향을 미치는 것으로 나타났으므로 클러스터링 자질 선정에 있어서 주목할 부분에 해당한다. 또한 질의응답문서 클러스터링 결과로 산출된 클러스터는 질의응답 지식검색 서비스를 위한 분류체계 지원방법으로 활용될 수 있다. 이용자가 직접 질의를 생성하고 분류하는 체제에서 이용자가 기존의 분류체계를 숙지하고 유사한 정보가 분산되지 않게 질의를 분류한다는 것은 거의 불가능하다. 응답내용을 중심으로 문서 클러스터가 주제별로 효과적으로 생성된다면 기존의 분류체계를 보완하는 방법이 될 수 있다.

둘째, 질의응답문서를 검색하는 지식검색 서비스에서 질의를 개선시키고자 할 때 응답내용을 중심으로 처리하는 것을 고려할 수 있다. 질의를 처리하는 과정을 통해 핵심 개념으로 질의를 재생성한 후 다시 검색을 하거나 질의를 확장하고자 할 때, 검색된 질의응답문서의 질의제목과 질의설명이 아닌 응답내용을 적용하였을 경우 더 효과적일 수 있다는 가능성이 제시되었다.

셋째, 질의응답문서에서 질문자의 적합성 피드백을 파악하는 목적으로 응답내용을 적용할 수 있다. 질문자가 자신이 질문한 내용에 가장 적절하다고 생각되는 내용을 응답내용으로 선택한 것이고 실험결과, 응답내용이 질의응답문서의 구성요소 중 질의 주제를 가장 효과적으로 표현하였으므로 적합성 피드백 과정으로 인식할 수 있다.

이 연구는 컴퓨터, 인터넷 분야에서의 질의응답문서를 추출하여 클러스터링 한 결과를 중심으로 수행된 것이므로 향후 다양한 주제 분야로 확장하여 주제별 특성을 고찰해 볼 필요가 있다. 또한 질의응답문서 검색 환경도 주제별 FAQ나 일반인이 참여하는 지식검색 등 점차 세분화되고 있으므로 이에 따른 분야별 질의응답문서 검색환경에 대한 연구도 수행되어야 한다고 제언하는 바이다.

참 고 문 헌

노정순. 2004. OPAC에서 탐색결과의 클러스터링에 관한 연구. 『한국문헌정보학회지』, 38(1): 36-50.

정영미, 이재윤. 2001. 클러스터링 성능 평가를 위한 비편향적 척도의 개발. 『제8회 한국정보관리학회 학술대회 논문집』, 167-172.

정영미, 최상희. 2001. "문장 클러스터링에 기반한 자동요약 모형." 『정보관리학회지』, 18(3): 159-177.

최상희. 2004. 질의응답을 위한 복수문서 요약에 관한 연구. 연세대학교, 박사학위 논문.

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

Bloesch, A. C., and T. A. Halpin. 1997. "Conceptual Queries Using Conquer II." *Proceedings of the ER'97: 16th International Conference on Conceptual Modeling*, (Los Angeles). 112-126.

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

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

Milligan, G. W., S. C. Soon, and L. M. Sokol. 1983. "The Effect of Cluster Size, Dimensionality, and the Number of Cluster on Recovery of True Cluster Structure." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5(1): 40-47.

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

Roussinov, D., and H. Chen. 2001. "Information Navigation on the Web by Clustering and Summarizing Query Results." *Information Processing & Management*, 37(4): 789-816.

Wen, Ji-Ron, Jian-Yun Nie, and Hong-Jiang

Zhang. 2001. "Clustering User Queris of a Serach Engine." *In Proceedings of WWW10*, pp.162-168.

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