개별문서의 지식구조 브라우징 인터페이스에 관한 연구

Designing User Interface Model for Browsing the Knowledge Structure of a Single Document

한승희, 연세대학교 문헌정보학과, libinfo@yonsei.ac.kr
이재윤, 경기대학교 문헌정보학과, memexlee@kyonggi.ac.kr

Seung-Hee Han, Dept. of Library and Information Science, Yonsei University
Jae-Yun Lee, Dept. of Library and Information Science, Kyonggi University

이 연구에서는 현재의 정보검색 환경에서 이용자 친화적인 검색 시스템을 개발하기 위한 한 방안으로 개별문서의 지식구조 브라우징 인터페이스를 제안하였다. 개별문서에 대한 지식구조를 자동 생성하기 위해 개별문서에 출현한 용어를 이용하여 용어 클러스터링과 클러스터 대표어 선정 작업을 수행하였고, 이를 대상으로 다차원 축척법을 이용하여 2차원 공간에 개별문서의 지식구조를 표현함으로써 이용자가 개별문서에 대해 보다 용이하게 접근할 수 있는 브라우징 인터페이스를 마련하였다.

1. 서론

현재의 정보검색 환경에서 이용자들은 일반적으로 원문의 표제나 저자, 요약문 등에만 의존하여 검색결과로 얻은 개별문서가 적합한가의 여부를 판단하고 그것을 이용하게 된다. 그러나 이러한 원문 대용물(surrogate)은 이용자가 원문의 내용을 이해하는 데 결정적인 역할을 하지 못한다는 연구 결과들이 보고되었다(Maloney 1974, 이태영 1990).

이 연구에서는 개별문서에 출현한 용어간의 연관성 분석을 기초로 개별문서의 지식구조를 자동으로 생성하고, 이를 정보검색 환경에서 실제로 적용할 수 있도록 개별문서의 지식구조를 브라우징하기 위한 인터페이스를 제안하였다. 개별문서의 지식구조 자동 생성은 용어 클러스터링과 클러스터 대표어 선정을 통해 수행되었고, 자동 생성된 지식구조를 탐색하기 위한 인터페이스를 생성하기 위해서는 다차원 축척법(MDS: Multidimensional Scaling)을 이용하였다.

2. 개별문서의 지식구조 생성

이 연구에서는 개별문서의 지식구조를 자동으로 생성하기 위해 정보학 분야의 학위논문 1편을 대상으로 용어 클러스터링 실험과 클러스터 대표어 선정 실험을 수행하였다.

실험을 수행하기 전에 텍스트 전처리 과정에서는 각 논문을 30개의 색인어를 기준으로 단락 단위로 분할한 후 형태소 분석을 통한 자동색인 결과로부터 불용어 제거 및 자질 축소 단계를 거쳐 색인어를 추출하고 이에 대해

다양한 색인어 가중치를 부여하였다. 이 때 사용한 색인어 가중치 공식은 단락 내 이진 단어빈도(binary term frequency in passage, btfp)이다. 실험 대상의 통계적 특성은 <표 1>과 같다.

<표 22> 실험 대상의 통계적 특성

고정길이 단락 수	불용어 제거 전 색인어 수	불용어 제거 후 색인어 수	불용어 제거+tf=2 이하 자질축소 후 색인어 수
173	727	577	149

2.1 용어 클러스터링

텍스트 전처리 과정을 거쳐 추출된 149개의 색인어를 대상으로 문서 내에 출현한 용어간의 연관성 측정을 기초로 용어 클러스터링 실험을 수행하였다. 용어간의 연관성 측정을 통해 단락-용어 행렬을 작성하기 위해 코사인 유사계수(cosine coefficient)를 사용하였다. 용어 x와 용어 y에 대해 x_i는 단락 i에 출현한 용어 x의 가중치이며, y_i는 단락 i에 출현한 용어 y의 가중치일 때, 그 공식은 다음과 같다(Sneath and Sokal 1973).

$$cosine(x,y) = \frac{\sum_i (x_i y_i)}{\sqrt{(\sum_i x_i^2)(\sum_i y_i^2)}}$$

다음 단계로 코사인 유사계수를 이용하여 생성된 단락-용어 행렬을 대상으로 용어 클러스터링 실험을 수행하였다. 다양한 클러스터링 알고리즘이 있으나, 이 연구에서는 용어 클러스터링에 일반적으로 이용되는 워드 기법(Ward's method)을 이용하였다. 워드 기법은 다른 계층적 기법에 비해 클러스터의 크기를 작고 균일하게 분류해주는 경향이 있기 때문에 용어나 개념의 자동분류에 적합하다고 알려져 있다(Ding 2001; Nedaniĉ, Spasiĉ, and Ananiadou 2002).

클러스터링 기법을 적용할 때에는 클러스터의 수를 고려해야 한다. 이 실험에서는 워드 기법에 휴리스틱을 적용하여 25-군집의 대군집과 10-군집의 소군집으로 클러스터 계층을 형성하였다.

2.2 용어 클러스터 대표어 선정

용어 클러스터링 실험에서 생성된 소군집의 클러스터를 대상으로 각 클러스터를 대표하는 25개의 클러스터 대표어를 선정하였다. 이 연구에서 클러스터 대표어를 선정하기 위해 단락빈도(passage frequency)를 이용하였다. 특정 용어의 단락빈도가 높다는 것은 그 용어가 문헌 전체에서 골고루 출현했다는 것을 의미한다. 그러므로 특정 클러스터에 속한 용어들 중에서 단락빈도가 가장 높은 용어를 문헌 전체에서 주제적으로 의미있는 것으로 보고, 그 클러스터의 대표어로 선정하였다.

3 개별문서의 지식구조 브라우징 인터페이스

자동으로 생성된 개별문서의 지식구조를 이용하여 이용자에게 개별문서의 지식구조에 대한 이해와 접근을 돕기 위한 브라우징 인터페이스를 설계하였다. 이용자가 원하는 주제에 대해 얻어낸 검색 결과 중에서 특정 문서를 선택하면 시스템은 이용자에게 다음과 같은 세 단계로 개별문서에 대한 브라우징 인터페이스를 제공하게 된다.

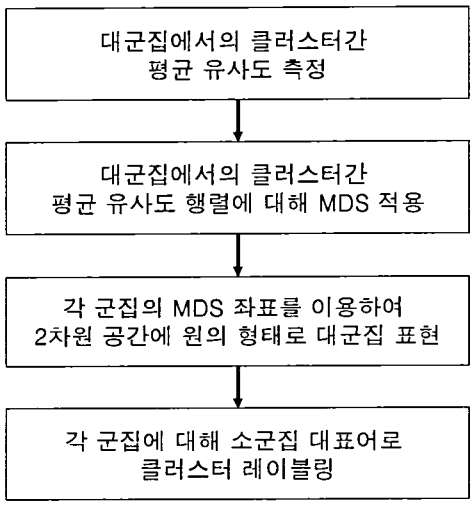
(1) 개별문서의 주제 개요 브라우징

개별문서를 구성하는 대표 주제(개념)들을 2차원 공간에 표현하여 이용자에게 제공함으로써, 이용자들은 특정 문서에 어떠한 주요 개념들이 출현하였으며, 그 개념과 관련된 개념

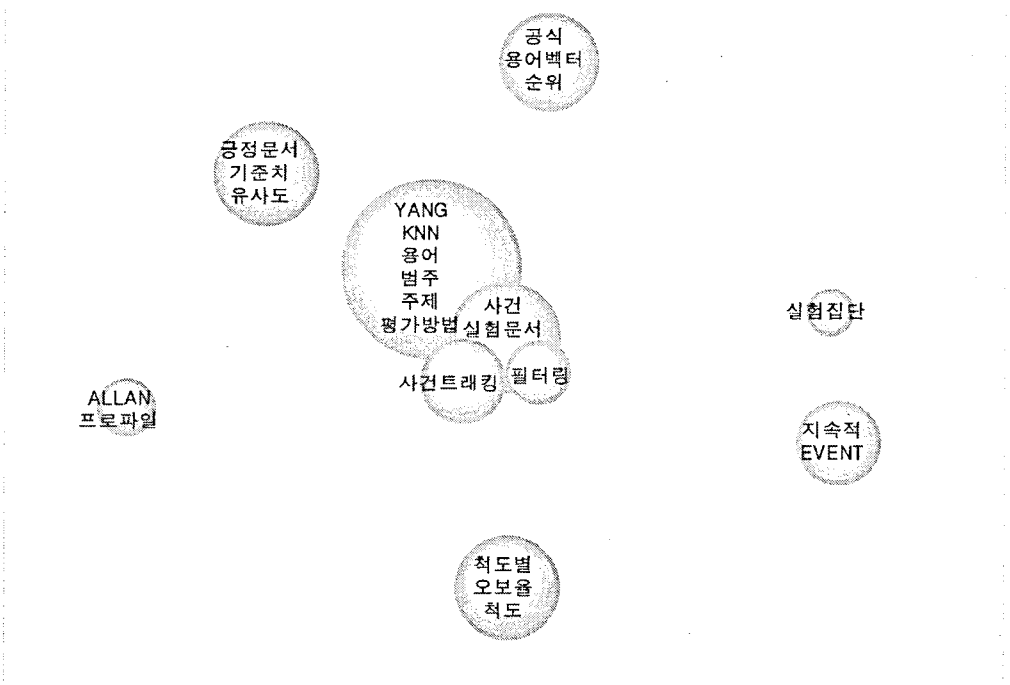
들에는 어떠한 것이 있는지를 한 눈에 파악할 수 있다.

이용자가 이해하기 쉽도록 개별문서의 주제 개요를 제공하기 위해 다음의 <그림 1>과 같은 과정으로 인터페이스를 생성하였다.

첫번째 과정에서 클러스터간의 평균 유사도를 측정하기 위해 코사인 유사계수를 이용하였으며, 용어 클러스터를 2차원 공간에 표현하기 위해 다차원 축척법을 이용하였다. 대군집의 클러스터를 소군집의 클러스터로 레이블링한 이유는 대군집을 대표하는 10개의 대표어보다는 소군집을 대표하는 25개의 대표어를 이용자에게 제공하는 것이 이용자의 원문에 대한 주제적 이해를 보다 용이하게 하기 위해서이다.



<그림 1> 개별문서의 주제 개요를 제공하기 위한 인터페이스 생성 과정



<그림 2> 개별문서의 주제 개요 브라우징 인터페이스

그림 1과 같은 과정을 거치면 그림 2와 같은 인터페이스 화면이 완성된다.

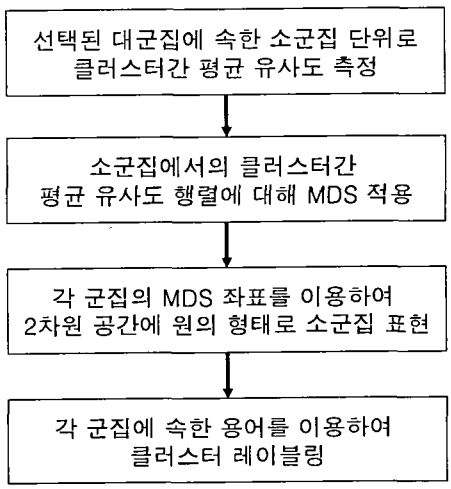
그림 2에서와 같이 원문의 내용을 나타내는 용어 클러스터는 2차원 공간에서 원의 형태로 표시되며, 원 안의 용어는 소군집에서의 클러스터 대표어를 나타낸다. 원의 크기는 용어 클러스터의 상대적인 크기를 나타내는데, 그림에서 보는 바와 같이, 대표어가 많은 용어 클러스터일수록 원의 크기가 큰 것을 알 수 있다. 또한 원과 원 사이의 거리는 주제간의 상대적 유사도를 나타낸다. 원 사이의 거리가 가까울수록 두 용어 클러스터는 유사한 주제를 표현하며, 거리가 멀수록 주제적으로 거리가 멀다는 것을 의미한다.

(2) 개별문서의 세부주제 접근

단계 1에서 개별문서의 주제 개요를 살핀 후 관심이 있는 클러스터를 선택하고 나면, 단계 2에서는 선택한 대군집을 구성하는 소군집을 이용자에게 제공함으로써 개별문서를 구성하는 구체적인 개념들에 대해 보다 쉽게 이해할 수 있도록 한다. 예를 들어, 그림 2에서 이용자가 'YANG, KNN, 용어, 범주, 주제, 평가방법'으로 표시된 군집을 선택하였다면, 단계 2에서는 이 대군집을 구성하는 소군집의 개념을 그림 3의 과정을 거쳐 그림 4와 같이 제공하게 된다.

단계 1에서와 마찬가지로, 첫번째 과정에서 클러스터간의 평균 유사도를 측정하기 위해 코사인 유사도를 이용하였다. 클러스터의 레이블링은 소군집에 속한 용어들을 그대로 이용하였다.

인터페이스의 형태가 갖는 특성 역시 단계 1에서와 마찬가지로 2차원 공간에서 원의 형태로 표시되는데, 공간에서의 원의 크기와 거리 특성은 단계 1에서와 같다.

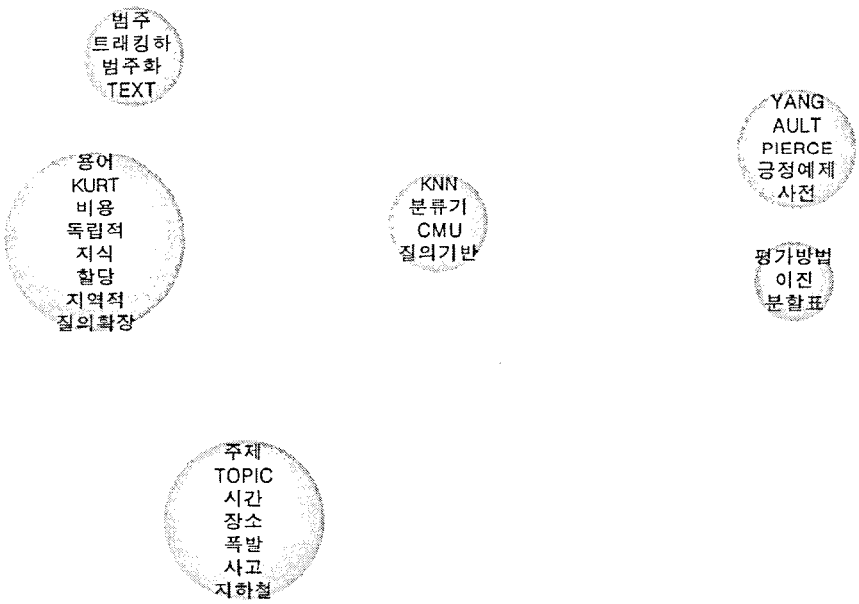


<그림 3> 개별문서의 세부주제 접근을 제공하기 위한 인터페이스 생성 과정

(3) 단락검색 결과의 제공

이용자가 단계 1에서 개별문서의 전체적인 주제를 이해하고 단계 2에서 개별문서를 구성하는 세부주제에 접근하고 나면, 단계 3에서는 이용자가 선택한 세부주제와 개별문서의 본문을 연결함으로써 이용자가 개별문서를 구성하는 세부 주제가 나타난 본문을 단락형태로 제공받게 된다.

이용자가 단계 2에서 특정 세부주제를 나타내는 용어 클러스터를 선택하면, 시스템에서는 단락검색을 통해 그 클러스터에 속한 용어들을 포함하고 있는 본문의 단락을 이용자에게 제공한다. 예를 들어, 이용자가 그림 4에서 'KNN, 분류기, CMU, 질의기반'이라는 클러스터를 선택하였다면, 그림 5와 같이 그 클러스터에 속한 용어들을 모두 포함하거나 일부를 포함하는 단락을 이용자에게 제공하게 된다.



<그림 11> 개별문서의 세부주제 접근 인터페이스

...사건트래킹과 **질의기반** 정보필터링을 동시에 실시하여 그 성능을 비교 및 분석하는 것을 그 내용으로 한다.
사건트래킹의 기법으로는 TDT의 연구기관 중 하나인 CMU에서 고안한 변형된 KNN 분류기를 사용하였으며, 정보필터링 기법에서는 Carpineto(2001)가 제안한 KLD(Kullback-Leible Divergence) 핵심어 추출 공식을 사용해 질의를 생성하고...

<그림 5> 개별문서의 세부주제 접근을 통한 단락검색 결과의 제공

이러한 세 단계의 브라우징 과정을 거쳐 이용자는 검색결과로 얻은 개별문서에 대해 그 문서의 전체적인 주제를 이해하고, 그 안에서 관심이 있는 세부주제에 대해 접근할 수 있으며, 결과적으로는 이용자가 탐색하기를 원하는 개별문서의 본문에까지도 접근할 수 있게 된다.

4 결론

이 연구에서는 정보검색 환경에서 개별문서의 지식구조를 생성하고 이를 대상으로 개별문서에 대한 이용자의 이해와 접근을 용이하게 하는 브라우징 인터페이스를 제안하였다. 정보검색 환경에서 기존의 이용자 인터페이스 연구가 복수의 문서집단의 주제적 표현에 초점을 맞추어 진행되었다면, 이 연구는 개별문서의 지식구조를 표현하기 위한 수단으로 브라우징 인터페이스를 적용하였다는 데에 의의

를 둘 수 있다.

인터페이스를 생성하기 위해 용어 클러스터링과 클러스터 대표어 선정 과정을 거쳐 클러스터간의 유사도 측정 데이터를 이용하여 다차원 축척법을 적용하였고, 이를 통해 2차원 공간에 용어 클러스터를 배치함으로써 이용자가 개별문서에 대한 주제적인 이해나 접근을 쉽게 하고, 더불어 관련된 원문의 단락을 탐색할 수 있도록 이용자 인터페이스를 설계하였다.

이러한 유형의 인터페이스는 실제 정보검색 환경에서 기존의 원문 대용물과 함께 검색결과의 이용이나 적합성 판정에 효과적으로 응용되어 이용자 친화적인 정보검색 시스템의 구현에 도움이 될 수 있다.

이 연구의 결과를 일반화하기 위해 이 연구에서 제안한 인터페이스 모형을 대상으로 이용자 평가를 수행하여 보다 이용자 친화적인 개별문서 브라우징 인터페이스를 제안할 필요가 있다.

참고문헌

이태영. 1990. "한국어 초록문의 문장과 내용에 관한 연구". **情報管理研究**, 21(1): 1-33.

한승희. 2004. **클러스터링 기법을 이용한 개별문서의 지식구조 자동 생성에 관한 연구**. 박사학위논문, 연세대학교 대학원 문헌정보학과.

Maloney, R. K. 1974. "Title versus Title/Abstract Text Searching SDI System". *Journal of the American Society for Information Science*, 25(6): 370-373.

Sneath, Peter, H. A., and Robert R. Sokal. 1973. *Numerical Taxonomy: The Princeiples and Practice of Numerical Classification.* San Francisco: W. H. Freeman and Company.