MDS를 이용한 개별문서의 계층적 지식구조 브라우징 인터페이스 설계

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

한 승 희*, 이 재 윤**
Seung-Hee Han, Jae-Yun Lee

차 례

초 록

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

키 워 드

용어 클러스터링, 클러스터 대표어, 계층적 브라우징 인터페이스, 다차원 척도법

* 일본 게이오대학교 도서관 · 정보학과 방문연구원
(Visiting Researcher, Dept. of Library and Information Science, Keio Univ., libinfo@yonsei.ac.kr),
** 경기대학교 문헌정보학과 전임강사
(Full-time lecturer, Dept. of Library and Information Science, Kyonggi University, memexlee@kyonggi.ac.kr)
• 논문접수일자 : 2004년 8월 30일
• 게재확정일자 : 2004년 9월 15일

ABSTRACT

The purpose of this study is to propose a hierarchical user interfaces for browsing the knowledge structure of a single document. To generate the hierarchical knowledge structure, hierarchical term clustering and cluster representative term selection were performed with a single thesis in information science field, and the result was applied to design the interfaces which browse a single document hierarchically using multidimensional scaling. The interfaces can be applied to develop the user-friendly information retrieval system.

KEYWORDS

term clustering, cluster representative term, hierarchical browsing interface, multidimensional scaling

1. 서론

인터넷의 대중화로 이용자들은 온라인 정보원을 이용하여 많은 양의 정보에 접근할 수 있게 되었다. 정보 과부하(information overload)와 같이 양적인 접근만을 강조하는 현재의 정보 환경에서 이용자는 많은 양의 정보 중에서 정보의 질을 판단하고 자신이 원하는 정보를 선택할 수 있는 정보 관리 능력이 필요하게 되었다. 특히 디지털 도서관의 개념이 대두되면서, 이용자들은 효율적으로 정보를 탐색하고 이용할 수 있는 새로운 정보검색 환경을 요구하게 되었다.

실제 현 정보검색 기법과 활용도구를 이용하는 검색환경은 대규모 정보처리 작업과 탐색결과 출력의 효율성 관점에서 다음과 같은 문제점을 갖는다(서은경 2002).

첫째, 검색 대상이 되는 대량의 데이터의 전체 구조와 찾고자하는 정보간의 상호관련성에 대해서 전혀 알 수 없다.

둘째, 대용량의 데이터를 대상으로 주제 구분 없이 단순 용어일치 검색이 이루어지므로 입력된 탐색어 또는 키워드에 대한 결과가 너무 많거나 부정확한 경우가 많다.

셋째, 이용자의 요구와 일치하는 정확한 질의를 표현하기가 쉽지 않아 기존 검색시스템에서의 질의어 선정은 이용자에게 큰 부담이 된다.

현재의 정보검색 환경에서 이용자들은 일반적으로 텍스트 형태의 질의문을 입력하고 텍스트 형태의 리스트로 검색결과를 제공받기 때문에 많은 검색결과로부터 이용자에게 적합하다고 판정되는 정보를 일일이 확인해야 하는 불편함이 있는데, 이 때 이용자의 적합성 판정은

검색결과의 표제나 저자, 목차, 요약문 등에 주로 의존하고 있다. 그러나 이러한 원문 대용물(surrogate)은 이용자가 원문의 내용을 이해하는 데 결정적인 역할을 하지 못한다는 연구 결과들이 보고되었다(Maloney 1974, 이태영 1990).

이러한 기존의 정보검색 환경의 문제를 해결하는 한 방법으로, 검색결과로 나타나는 개별문서의 지식구조를 시각화하여 이용자가 원하는 정보에 대해 편리하게 접근하고 이를 효과적으로 이용할 수 있도록 하는 개별문서의 브라우징 인터페이스를 설계할 수 있다.

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 척도법(MDS: Multidimensional Scaling)을 이용하였다.

기존의 정보시각화 연구가 검색결과로 나타나는 복수의 문헌집단이 갖는 특성을 표현하는 방법을 중심으로 연구되었다면, 이 연구는 검색결과로 나타나는 개별문서의 주제적 특성을 표현하고자 하는 것으로, 기존의 정보시각화 연구와는 근본적인 차이가 있다.

2. 개별문서의 지식구조 생성

이 연구에서는 개별문서의 지식구조를 자동으로 생성하기 위해 정보학 분야의 학위논문 1편을 대상으로 용어 클러스터링 실험과 클러스터 대표어 선정 실험을 수행하였다. 개별문서의 지식구조 자동 생성과정은 <그림 1>과 같다. 이 중에서 문서 단락분할 과정, 자동색인 과정, 용어 연관성 측정 과정은 용어 클러스터를 생성하고 클러스터 대표어를 선정하기 위한 텍스트 전처리 과정에 해당한다.

2. 1 텍스트 전처리

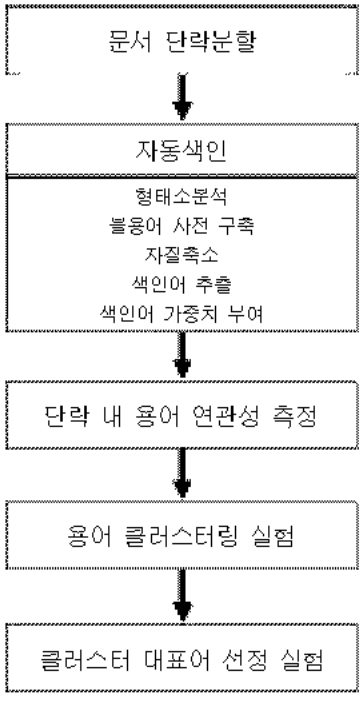
2.1.1 문서의 단락분할

문서를 단락으로 구분하는 방법 중 가장 손쉬운 것은 저자가 나눈 장, 절이나 문단을 그대로 이용하는 것이다. 그러나 일반적으로 모든 단락의 길이가 유사하지 않기 때문에 단락의 길이를 정규화해야 한다.

단락검색과 지역적 질의확장 분야를 중심으로 단락길이의 정규화 방법에 대해서 많은 연구가 있었다. 이와 관련된 선행연구에서 일반적으로 특정 크기의 단어 창(word window)을 만들어 문헌을 고정길이의 단락으로 구분하는 것이 저자에 의한 장, 절, 문단 단위로 나눈 것보다 성능이 우수한 것으로 나타났기 때문에 이 실험에서도 단락의 길이를 문단으로 고정하여 분할하는 방식을 채택하였다.

단락의 길이를 결정하기 위해 실험 문서의

문단별 평균 단어 수를 계산한 결과 한 문단이 평균 29.7개의 단어를 포함하는 것으로 나타났다. 이를 근거로 이 실험에서는 30개의 단어를 한 문단으로 하는 고정길이 단락분할 기법을 적용하였다.



〈그림 1〉 개별문서의 지식구조 자동 생성 과정

2.1.2 자동색인

단락 수준이 결정된 후에는 단락을 기준으로 형태소 분석을 이용하여 색인어를 추출하였다. 색인 과정에서 한 글자짜리 색인어는 단어를 지나치게 세분하여 오분석된 경우가 많고 클러스터링에 잡음으로 추가되므로 제외하였고 복합명사는 분할하지 않았다. 실험 문서의 성격상 외국어 표기가 많았는데, 한국어와 병기된 형태의 동의어, 논문에서 인용한 저자의 이름, 고유명사 등과 같이 외국어 표기의 대부분이 원문에서 핵심이 되는 정보였으므로 색인 대상에 포함시켰다.

문서 내에 분포하는 고빈도어와 저빈도어는 시스템의 성능 향상에 긍정적인 영향을 주지 않는 것이 일반적이기 때문에 이 실험에서는 색인어 집합을 축소하기 위해 두 가지 방법을 이용

하였다. 첫번째는 불용어 사전을 구축하여 고빈도 불용어를 제거하는 방법이고, 두번째는 저빈도어 집합을 제거하는 방법이다.

우선 용어 클러스터링의 성능 향상을 위해 자동색인 결과에 대해서 불용어 사전을 구축하였다. 색인어를 분석한 결과, 실험 문서 집단의 특성을 반영하는 용어들이 자주 출현하는 것을 알 수 있었다. 예를 들면, '연구', '논문', '방법', '사용', '결과', '비교' 등과 같이 논문을 작성할 때 일반적으로 많이 쓰이는 용어들은 대부분 고빈도어이므로 용어간의 연관성 측정에 긍정적인 영향을 미치지 못하였다. 그러므로 이러한 용어들을 중심으로 불용어 사전을 구축하였다.

불용어 사전을 구축하여 불용어를 제거한 후에는 *tf*가 2 이하인 저빈도어를 제거하는 자질축소 단계를 거쳐 최종 색인어 집합을 구축하였고, 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 1> 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

2.1.3 용어간 연관성 측정

텍스트 전처리 과정을 거쳐 추출된 149개의 색인어를 대상으로 문서 내에 출현한 용어간의 연관성을 측정하였다. 용어간의 연관성은 용어 쌍간의 동시출현빈도를 나타내는 단락 용어 행렬에 유사계수를 이용하여 측정되었다. 여러 종류의 유사계수가 있으나 이 실험에서는 용어 클러스터링에 적합한 것으로 알려진 코사인 유사계수(cosine coefficient)를 적용하였다(한승희 2004). 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).

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

2.2 용어 클러스터링

다음 단계로 코사인 유사계수를 이용하여 생성된 단락 용어 행렬을 대상으로 용어 클러스터링(term clustering) 실험을 수행하였다.

용어 클러스터링은 서로 관련 있는 용어들을 일정한 기준에 따라 모아서 여러 개의 용어 클래스를 형성하는 것으로, 텍스트를 대상으로 용어간의 통계적 연관성(term association) 분석

에 기초한다. 용어간의 연관성은 주로 문헌 집단 내 용어들의 동시출현빈도를 이용하여 측정하는데, 두 개의 용어가 많은 수의 문헌에 함께 출현하였다면 이 두 용어는 서로 관련이 있다고 보고 같은 클래스에 포함시킨다(정영미 1993). 용어 자동분류의 접근 방법은 비계층적 방법과 계층적 방법이 있으나, 일반적으로 용어간의 계층 관계를 표현하기 위해 주로 계층적 클러스터링 기법을 적용한다. 그러므로 이 연구에서도 개별문서의 지식구조를 계층적으로 표현하기 위해 계층적 기법을 선택하였다.

용어 클러스터링에 적용되는 계층적 클러스터링 알고리즘으로는 완전연결 기법(complete linkage), 단일연결 기법(single linkage), 그룹평균연결 기법(group average linkage), 워드 기법(Ward's method) 등 다양한 클러스터링 알고리즘이 있으나, 이 연구에서는 용어 클러스터링에 일반적으로 이용되는 워드 기법을 이용하였다.

워드 기법은 클러스터를 구성하는 객체간의 유클리드 거리의 제곱오차를 최소화하는 방식으로 클러스터를 통합하는데(Ward 1963), 이 기법은 다른 계층적 기법에 비해 클러스터의 크기를 작고 균일하게 분류해주는 경향이 있기 때문에 용어나 개념의 자동분류에 적합하다고 알려져 있다(Ding et al. 2001; Nedaniĉ, Spasiĉ, and Ananiadou 2002; 이미경 2002; 유영준 2003). Nedaniĉ, Spasiĉ, and Ananiadou (2002)은 2082개의 Medline 검색결과 중에서 추출한 상위 174개의 용어를 가지고 워드 기법과 단일연결 기법을 이용하여 용어 클러스터링 실험을 하였는데, 실험결과 단일연결 기법에 비해 워드 기법이 동일한 조건에서 더 많은 군집을 생성하는 것으로 나타났으며, 클러스터의 분류 정확률에 있어서도 단일연결 기법에 비해 나은 성능을 보였다.

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

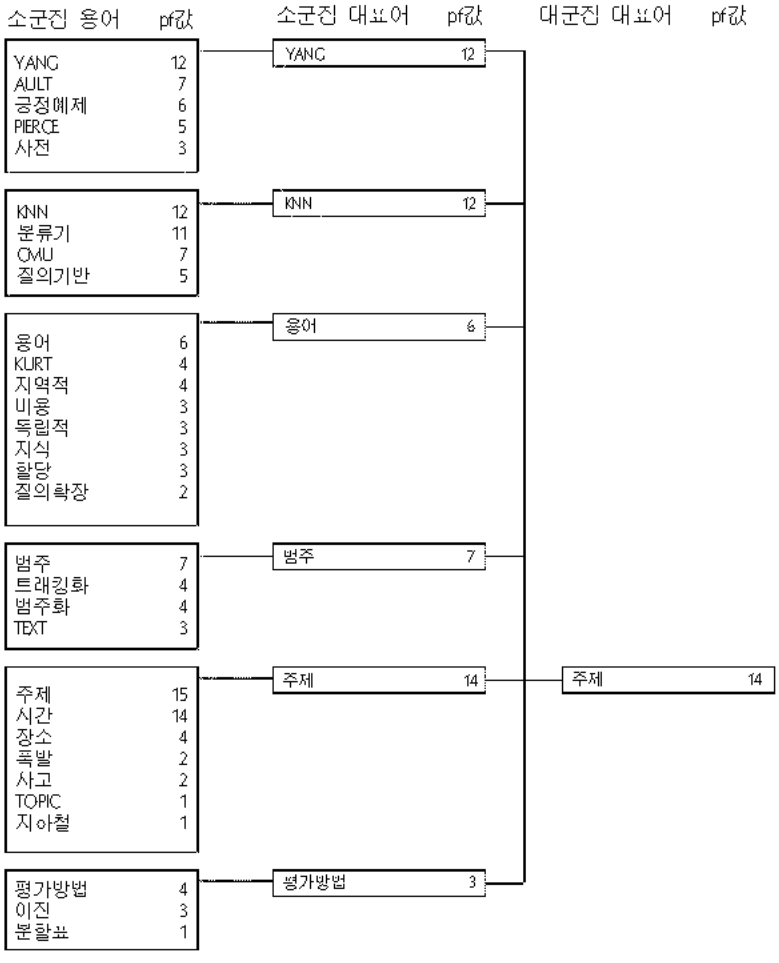
2.3 용어 클러스터 대표어 선정

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.

〈그림 2〉는 소군집과 대군집 수준에서의 용어 클러스터링 결과에 대한 클러스터 대표어 선정 결과의 일부를 나타낸다. 군집 수준의 변화에 따라 대표어를 선정할 때에는 승자 노드를 채택하는 방식을 이용한다. 〈그림 2〉에서 보는

바와 같이 클러스터를 처음에 소군집으로 생성하였을 때 각 클러스터에서 단락빈도가 가장 높은 'YANG', 'KNN', '용어', '범주', '주제', '평가방법'이 소군집에서의 대표어가 된다. 이 여섯 개의 군집이 대군집에서의 통합 과정을 거치면 여섯 개의 대표어 중에서 단락빈도가 큰 용어가 있는 클러스터로 단락빈도가 작은 용어가 있는 나머지의 클러스터가 통합되어 새로운 클러스터를 형성하면서 단락빈도가 큰 '주제'가 승자 노드로 채택되어 대군집에서의 대표어가 된다. 이러한 과정을 거쳐 지식구조가 자동으로 생성되는데, 소군집에서 대군집으로 통합되면서 그 결과는 계층적인 형태를 이룬다.

3. 개별문서의 계층적 지식구조 브라우징 인터페이스 설계



〈그림 2〉 계층적 구조로 자동 생성된 지식구조의 일부

자동으로 생성된 개별문서의 지식구조를 이용하여 이용자에게 개별문서의 지식구조에 대한 이해와 접근을 돕기 위한 계층적 형태의 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.

3.1 개별문서의 주제 개요 브라우징

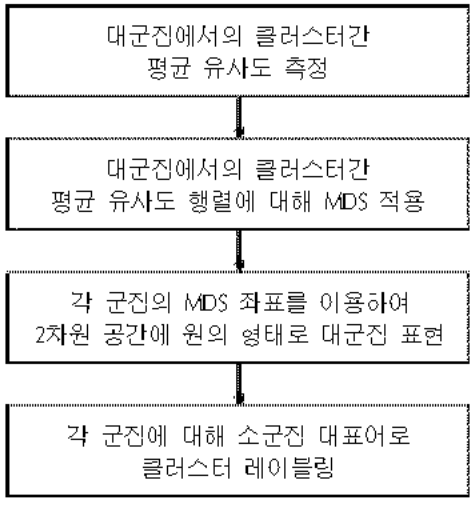
개별문서를 구성하는 대표 주제(개념)들을 2차원 공간에 표현하여 이용자에게 제공함으로써, 이용자들은 특정 문서에 어떠한 주요 개념들이 출현하였으며, 그 개념과 관련된 개념들에는 어떠한 것이 있는지를 한 눈에 파악할 수 있으며, 이러한 유형의 도구는 검색결과에 대한 이용자의 적합성 판정 보조 도구로 이용될 수 있다.

이용자가 개별문서의 내용을 이해하기 쉽도록 그 주제 개요를 제공하기 위해 다음의 〈그림 3〉과 같은 과정으로 인터페이스를 생성하였다.

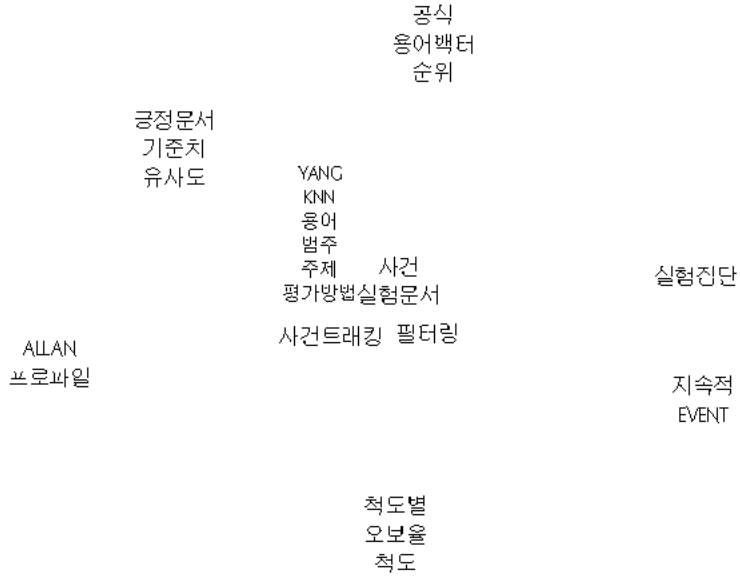
첫번째 과정에서는 대군집 수준에서 클러스터간의 평균 유사도를 측정하기 위해 코사인 유사계수를 이용하였으며, 용어 클러스터를 2차원 공간에 표현하기 위해 다차원 척도법을 이용하였다. 다차원 척도법은 일차원의 개념으로 측정할 수 없는 개념을 측정할 때 사용하는 기법으로 대상들간의 유사성을 평가하게 하여 평가자가 대상을 평가하는 데 내재하고 있는 평가기준을 발견하고, 각 기준에 따라 평가대상들이 갖는 측정치를 찾는 데 목적이 있다. 이러한 원리에 의하여 포지셔닝 맵(positioning map)을 작성하는 데 다차원 척도법을 주로 이용하고 있다.

다차원 척도법에 의해 얻어진 각 군집에 대한 좌표값을 이용하여 2차원 공간에 대군집을 표현한 후, 각 클러스터의 특성을 표현하고 이용자들이 개별문서의 주제적인 특성을 식별하게 하기 위해 각 클러스터에 대해 소군집 대표어로 레이블링을 하였다. 이 때 대군집의 클러스터를 소군집의 클러스터로 레이블링 한 이유는 대군집을 대표하는 10개의 대표어보다는 소군집을 대표하는 25개의 대표어를 이용자에게 제공하는 것이 이용자의 원문에 대한 주제적 이해를 보다 용이하게 하기 위해서이다. 〈그림 3〉과 같은 과정을 거치면 〈그림 4〉와 같은 인터페이스 화면이 완성된다.

〈그림 4〉에서와 같이 원문의 내용을 나타내는 대군집 수준의 10개의 용어 클러스터가 2차원 공간에서 원의 형태로 표시되며, 각 원은 원문의 내용을 구성하는 주제를 나타낸다. 이 때 원 안에 표시된 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.



〈그림 3〉 개별문서의 주제 개요를 제공하기 위한 인터페이스 생성 과정



〈그림 4〉 개별문서의 주제 개요 브라우징 인터페이스

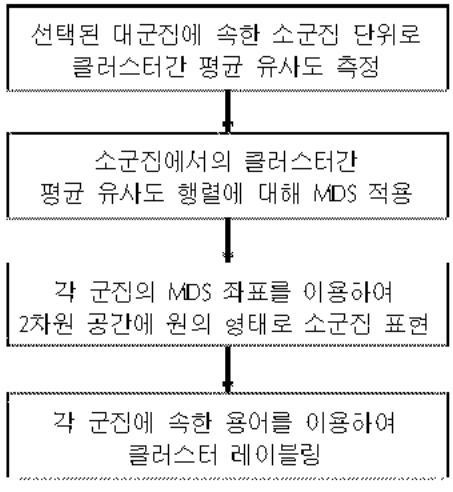
3.2 개별문서의 세부주제 접근 인터페이스

개별문서의 주제 개요에 대한 브라우징을 통해 관심이 있는 클러스터를 선택하고 나면, 다음 단계에서는 선택한 대군집을 구성하는 소군집 용어 클러스터를 이용자에게 제공함으로써 개별문서를 구성하는 구체적인 개념들에 대해 보다 쉽게 이해할 수 있도록 한다. 예를 들어, <그림 4>에서 이용자가 'YANG, KNN, 용어, 범주, 주제, 평가방법'으로 표시된 군집을 선택하였다면, 이 단계에서는 이 대군집을 구성하는 소군집의 개념을 <그림 5>의 과정을 거쳐 <그림 6>과 같이 제공하게 된다.

개별문서의 주제 개요 브라우징 단계에서와 마찬가지로, 첫번째 과정에서 클러스터간의 평균 유사도를 측정하기 위해 코사인 유사계수를 이용하였다. 이 때 클러스터의 레이블링은 소군집에 속한 용어들을 그대로 이용하였다. 이 인터페이스의 형태적 특성 역시 앞 단계에서와 마찬가지로 2차원 공간에서 원의 형태로 표시되는데, 공간에서의 원의 크기와 거리 특성은 앞 단계에서와 같다.

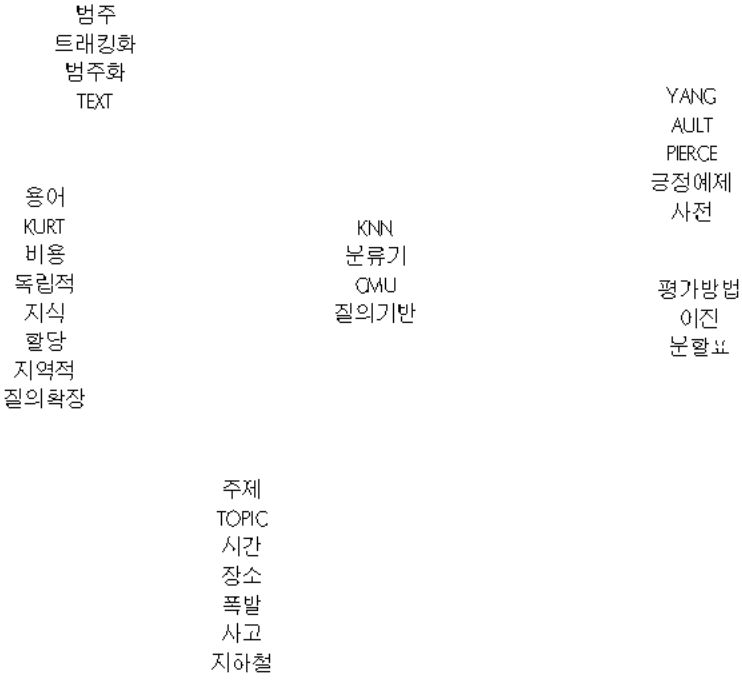
이러한 인터페이스의 장점은 원문을 구성하는 세부 개념들을 파악할 수 있다는 것에 있다. 예를 들어, <그림 6>에서 보는 바와 같이, 원문에서 자주 출현하는 저자 집단('YANG', 'AULT', 'PIERCE' 등)을 파악함으로써, 특정 주제 분야에서 연구 전선(research front)에 있는 연구자들을 파악할 수 있고, 이에 따라 그 주제 분야와 관련된 다른 연구들을 이용자로 하여금 찾아볼 수 있게 도와준다.

앞에서 생성된 <그림 2>의 지식구조가 여섯 개의 소군집에서 한 개의 대군집으로 통합되는 계층적 구조를 나타낸다고 할 때, <그림 4>에서는 대군집을 표현함으로써 계층구조의 상위계층을 표현하고 있으며, <그림 6>에서는 소군집



<그림 5> 개별문서의 세부주제 접근을 제공하기 위한 인터페이스 생성 과정

을 표현함으로써 계층구조의 하위계층을 표현하고 있는 것을 알 수 있다. 그러므로 〈그림 4〉의 인터페이스를 통해 이용자는 개별문서가 갖는 주제 개요(상위 개념)를 이해할 수 있고, 〈그림 6〉의 인터페이스를 통해서는 개별문서를 구성하는 세부주제(하위 개념)에 접근할 수 있다.



〈그림 6〉 개별문서의 세부주제 접근 인터페이스

3.3 단락검색 결과의 제공

이용자가 첫 번째 단계에서 개별문서의 전체적인 주제를 이해하고 두 번째 단계에서 개별문서를 구성하는 세부주제에 접근하고 나면, 그 다음 단계에서는 이용자가 선택한 세부주제와 개별문서의 본문을 연결함으로써 이용자가 개별문서를 구성하는 세부 주제가 나타난 본문을 단락형태로 제공받게 된다.

이용자가 두 번째 단계에서 특정 세부주제를 나타내는 용어 클러스터를 선택하면, 시스템에서는 단락검색을 통해 그 클러스터에 속한 용어들을 포함하고 있는 본문의 단락을 이용자에게 제공한다. 예를 들어, 이용자가 〈그림 6〉에서 'KNN, 분류기, CMU, 질의기반' 이라는 클러

스터를 선택하였다면, 그 클러스터에 속한 용어들을 모두 포함하거나 일부를 포함하는 단락을 이용자에게 제공하게 되며, 그 형태는 〈그림 7〉과 같다.

질의어 : 'KNN' '분류기' 'CMU' '질의기반'

단락검색결과 : 19건

순위	단락 내용	본문 페이지 연결
1	...사건트래킹과 **질의기반** 정보필터링을 동시에 실시하여 그 성능을 비교 및 분석하는 것을 그 내용으로 한다. 사건트래킹의 기법으로는 TDT의 연구기관 중 하나인 **CMU**에서 고안한 번영된 **kNN 분류기**를 사용하였으며, 정보필터링 기법에서는 Carpineto(2001)가 제안한 KLD(Kullback-Leible Divergence) 색심어 추출 공식을 사용해 질의를 생성하고...	p.3
2	예제기반 범주화 기법인 **kNN 분류기**는 성능이 우수하고 트래킹 작업에서 그에 필요한 용어나 사건의 처리, 결정값의 최적화등에 대한 별도의 가정없이 **kNN 분류기**의 기법을 그대로 사건트래킹에 적용할 수 있다. 따라서 **CMU**에서는 이를 사용해 사건트래킹을 수행하였다(Yang 2000). **kNN 분류기**는 하나의 입력문서에...	p.9

〈그림 7〉 개별문서의 세부주제 접근을 통한 단락검색 결과 제공

〈그림 7〉은 이용자가 선택한 특정 용어 클러스터에 대한 단락검색 결과를 임의로 표현한 것으로, 이용자에게 특정 용어를 모두 혹은 일부 포함하고 있는 단락을 제공하며, 그 내용을 보다 자세하게 확인하고 싶은 경우에는 원문의 페이지와 연결해주는 구조를 갖는다.

이러한 세 단계의 계층적 브라우징 과정을 거쳐 이용자는 검색결과로 얻은 개별문서에 대해 그 문서의 전체적인 주제를 이해하는 동시에 그 문서가 이용자에게 적합한 것인지를 판정할 수 있다. 또한 적합하다고 판정한 후에는 원문을 구성하는 관심 있는 세부주제에 대해 접근할 수 있으며, 결과적으로는 이용자가 탐색하기를 원하는 개별문서의 본문에까지도 접근할 수 있게 된다.

4. 결론

이 연구에서는 정보검색 환경에서 개별문서의 계층적 지식구조를 자동으로 생성하고 이를

대상으로 개별문서에 대한 이용자의 이해와 접근을 용이하게 하는 개별문서의 지식구조 브라우징 인터페이스를 제안하였다.

인터페이스를 생성하기 위한 첫 번째 과정으로 개별문서에 출현한 용어를 가지고 클러스터링과 클러스터 대표어 선정 과정을 거쳐 개별문서의 계층적 지식구조를 생성하였다. 두 번째 과정에서는 계층적 브라우징 인터페이스를 설계하였는데, 첫 번째 과정에서 생성된 용어 클러스터간의 유사도 측정 데이터에 다차원 척도법을 적용해 용어 클러스터를 2차원 공간에 배치함으로써 이용자가 개별문서에 대한 주제적인 이해나 접근을 쉽게 하고, 더불어 관련된 원문의 단락을 탐색할 수 있도록 이용자 인터페이스를 설계하였다.

많은 연구자들이 검색시스템에서의 정보시각화 인터페이스는 이용자에게 의미 있는 정보를 제공함으로써 검색시스템의 효율성을 높여줄 것이며 이용자가 정보시각화 인터페이스를 더 만족할 것이라 믿고 있다(서은경 2002). 그러므로 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 다양한 종류의 문헌에 적용해 볼 필요가 있으며, 이용자 평가를 수행하여 보다 이용자 친화적인 개별문서 브라우징 인터페이스를 제안할 필요가 있다.

참고문헌

서은경. 2002. 정보시각화에 대한 스킴모형별 비교 분석. 『한국문헌정보학회지』, 36(4): 175 205.

유영준. 2003. 『문헌정보학의 지식 구조에 관한 연구』. 박사학위논문, 연세대학교 대학원, 문헌정보학과.

이미경. 2002. 『동시출현 단어 분석을 통한 지식구조 파악에 관한 연구: 인공지능 분야를 대상으로』. 석사학위논문, 연세대학교 대학원, 문헌정보학과.

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

정영미. 1993. 『정보검색론』. 개정판. 서울: 구미무역㈜출판부.

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

Anderberg, Michael R. 1973. *Cluster Analysis for Applications*. New York: Academic Press.

Ding, Ying, Gobinda G. Ghowdhury, and Schubert Foo. 2001. "Bibliometric Cartography of Information Retrieval Research by using Co word Analysis". *Information Processing & Management*, 37:

817 842.

Doyle, Lauren B. 1961. "Semantic Road Maps for Literature Searchers". *Journal of the ACM*, 8(4): 553 578.

Lin, Xia. 1997. "Map Displays for Information Retrieval". *Journal of the American Society for Information Science*, 48(1): 40 54.

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

Nenadiĉ, G., Spasiĉ, I., and Ananiadou, S. 2002. "Term Clustering using a Corpus Based Similarity Measure". in Sojka, P., Ivan Kopecek, and Karel Pala Eds. *Text, Speech and Dialogue(TSD 2002)*, Berlin: Springer, 151 154.

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

Tombros, Anastasios, Robert Villa, and C. J. Van Rijbergen. 2001. "The Effectiveness of Query specific Hierarchic Clustering in Information Retrieval". *Information Processing and Management*, 38: 559 582.

Tombros, Anastasios. 2002. *The Effectiveness of Query based Hierarchical Clustering of Documents for Information Retrieval*. Ph.D. diss., Department of Computing Science, University of Glasgow.

van Rijbergen, C. J. 1973. *Information Retrieval*. 2nd ed. London: Butterworths.

Ward, J. H. 1963. "Hierarchical Grouping to Minimize an Object Function". *Journal of the American Statistical Association*, 58: 236 244.