클러스터링 기법을 이용한 개별문서의 지식구조 자동 생성에 관한 연구*

Automatic Generation of the Local Level Knowledge Structure of a Single Document Using Clustering Methods

한 승 희(Seung-Hee Han)**
정 영 미(Young-Mee Chung)***

초 록

이 연구에서는 전통적인 인쇄매체 환경에서 지식에 대해 지역적인 접근법을 제공하는 권말색인과 목차의 기능에 착안하여 용어 클러스터링 실험과 클러스터 대표어 선정 실험을 통해 개별문서의 지식구조 자동 생성 기법을 제안하였고, 자동 생성된 지식구조가 갖는 기능성을 평가하여 정보검색 환경에서의 적용 가능성을 확인하였다. 용어 클러스터링 실험에서는 워드 기법의 성능이 중복 분류를 허용하는 퍼지 K-means 클러스터링 기법에 비해 높았으며, 클러스터 대표어 선정 기법으로는 단락빈도를 이용한 경우가 가장 좋은 성능을 나타냈다. 또한, 이용자 태스크를 기반으로 하여 최종적으로 생성된 지식구조의 기능성을 평가한 결과, 이 연구에서 자동 생성된 지식구조가 인쇄매체 환경에서의 권말색인과 목차가 갖는 기능을 어느 정도 수행한다는 것을 입증하였다.

ABSTRACT

The purpose of this study is to generate the local level knowledge structure of a single document, similar to end-of-the-book indexes and table of contents of printed material, through the use of term clustering and cluster representative term selection. Furthermore, it aims to analyze the functionalities of the knowledge structure, and to confirm the applicability of these methods in user-friendly information services. The results of the term clustering experiment showed that the performance of the Ward's method was superior to that of the fuzzy K-means clustering method. In the cluster representative term selection experiment, using the highest passage frequency term as the representative yielded the best performance. Finally, the result of user task-based functionality tests illustrate that the automatically generated knowledge structure in this study functions similarly to the local level knowledge structure presented in printed material.

키워드: 용어 클러스터링, 클러스터 대표어, 지역적 지식구조, 워드 기법, 퍼지 K-means 클러스터링 기법, term clustering, cluster representative term, local level knowledge structure, Ward's method, fuzzy K-means clustering method

---

* 본 연구는 연세대학교 대학원 박사학위논문의 일부를 요약한 것임.

** 日本 慶應義塾大學(Keio University) 圖書館・情報學科 訪問研究員(libinfo@yonsei.ac.kr)

*** 연세대학교 문헌정보학과 교수(ymchung@yonsei.ac.kr)

■ 논문접수일자 : 2004년 8월 17일

■ 게재확정일자 : 2004년 9월 10일

1. 서 론

전통적인 인쇄매체 환경에서 정보 이용자들은 전체 학문 분야나 특정 주제 분야에 구조적으로 접근하기 위해서 분류표나, 시소러스, 주제명 표목을 이용하며, 특정 저작 내에서 발생한 지식에 구조적으로 접근하기 위해서 목차와 권말색인을 이용하여 왔다. 전자를 전역적 수준의 지식구조(global level knowledge structure), 후자를 지역적 수준의 지식구조(local level knowledge structure)라고 할 수 있는데, 이러한 지식구조는 유형에 관계없이 근본적으로 지식에 대한 탐색을 지원하는 일종의 탐색 도구(search aids)로서의 기능을 수행한다.

전통적인 인쇄매체 환경에서와 마찬가지로 온라인 정보 환경에서도 이용자의 효율적이고 효과적인 정보 이용을 위해 정보에 대한 구조적인 접근 도구가 필요하다. 이러한 관점에 비추어 온라인 정보 환경에서 정보에 대한 구조적 접근에 관한 연구 패턴을 분석하면 크게 두 가지로 나누어 볼 수 있는데, 첫째는 동시인용분석(co-citation analysis)이나 동시출현 단어분석(co-word analysis)과 같은 방법을 이용해서 특정 주제 분야의 지식구조를 파악하는 연구(White and McCain 1998; Ingwersen, Larsen and Noyons 2001; Ding, Chowdhury, and Foo 2001)가 있고, 둘째는 'Scatter-Gather'와 같이 복수의 문서집단을 대상으로 하여 하이퍼텍스트 기반 브라우징 도구를 개발하는 것에 초점을 둔 연구가 있다(Hearst and Pedersen 1996).

이 두 가지 연구 패턴을 살펴보면 학문 분야나 특정 주제 분야 전체에 대해 구조적으로 접근하는 전역적 수준의 지식구조와 유사하다. 반면에 온라인 환경에서 개별문서 단위의 지식에 대한 구조적인 접근을 시도하거나 이를 바탕으로 개별문서에 대한 탐색 도구를 개발한 연구는 찾아보기 어렵다.

일반적으로 전통적인 인쇄매체 환경에서 정보 이용자는 개별문서의 지식에 접근하기 위해 두 가지 방식의 접근법을 이용하는데, 문서의 전체적인 구조를 이해하기 위해서는 목차를 이용하고, 문서에 출현한 특정 용어나 개념을 기준으로 문서의 주제나 관련 개념간의 관계를 파악하기 위해서는 권말색인을 이용한다. 만약 목차와 권말색인의 기능이 통합된 새로운 유형의 지식구조가 개발된다면 정보검색 환경에서 이용자는 검색 결과로 얻은 개별문서의 전체 구조 및 그 문서를 구성하는 세부 주제와 개념을 쉽게 파악할 수 있고, 정보검색 시스템 개발자나 정보 제공자의 입장에서는 많은 비용과 시간과 노력을 들이지 않고 이용자에게 효과적인 서비스를 제공할 수 있다.

이러한 관점에서, 이 연구는 인쇄매체 환경에서 주로 이용되는 개별저작 단위의 지식구조인 목차와 권말색인의 특성과 기능에 착안하여 이 두 가지 접근법의 기능을 통합한 형태로 개별문서에 대한 구조적 접근을 제공하는 지역적 지식구조를 자동 생성하는 것을 목적으로 한다. 이러한 유형의 지식구조는 정보검색 환경에서 개별문서에 대한 탐색 도구와 원문에 대한 새로운 유형의 대용물(surrogate)로서의 기능을 할 수 있다.

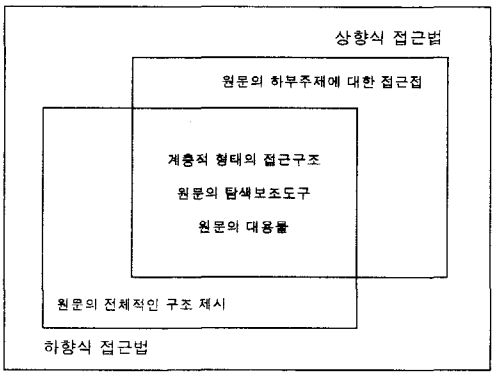
이 연구에서는 용어 클러스터링 실험과 대표어 선정 실험을 통해 원문 전체의 구조적인

특성과 원문을 구성하는 세부 주제의 특성을 동시에 표현할 수 있는 개별문서 단위의 지식구조를 자동 생성하였고, 저자와 이용자 집단을 대상으로 자동 생성된 개별문서의 지식구조가 갖는 기능성을 평가하였다.

2. 이론적 배경

2. 1 지역적 수준의 지식구조 기능 분석

유안나(1992)와 김효열(1995)의 연구를 참조하여 목차와 권말색인의 기능 분석을 통해 온라인 정보 환경에 유용한 지역적 수준의 지식구조가 갖추어야 할 요소를 제시하였으며, 그 내용은 다음과 같다.



〈그림 1〉 지역적 지식구조의 기능적 요소

첫째, 원문에 대한 전체적인 구조를 제시하는 동시에 원문을 구성하는 세부 주제에 대해서도 접근점을 제공해야 한다.

둘째, 원문에 대한 효과적인 접근을 위해서 지식구조는 계층적인 형태로 표현되어야 한다.

셋째, 원문의 내용 탐색을 용이하게 하는 탐색 도구로서의 기능을 해야 한다.

넷째, 원문의 내용을 대신할 수 있는 원문의 대용물로서의 기능을 해야 한다.

이 연구에서는 지역적 지식구조의 기능적 요소 분석을 통해 지역적 지식구조의 자동 생성 실험 결과를 평가하였고, 그 내용은 제 4장에서 언급하였다.

2. 2 지식구조 생성을 위한 용어 클러스터링 기법

2. 2. 1 계층적 워드 기법

정보검색 분야에서 주로 적용되고 있는 계층적 응집 클러스터링 기법(Hierarchical Agglomerative Clustering Method)은 객체간의 유사성을 어떻게 측정하는가에 따라 단일연결 기법(single linkage), 완전연결 기법(complete linkage), 그룹평균연결 기법(group average linkage), 워드 기법(Ward's method) 등으로 구분된다.

이 중, 워드 기법은 클러스터를 구성하는 객체간의 유클리드 거리의 제곱오차를 최소화하는 방식으로 클러스터를 통합하는데(Ward 1963), 이 기법은 클러스터를 모두 비슷한 크기로 생성하는 경향이 있다(Milligan, Soon, and Sokol 1983). 워드 기법은 다른 계층적 기법에 비해 클러스터의 크기를 작고 균일하게 분류해주는 경향이 있기 때문에 용어나 개념의 자동분류에 적합하다고 알려져 있다(Ding, Chowdhury, and Foo 2001; Nedanić, Spasić, and Ananiadou 2002).

2. 2. 2 퍼지 K-means 클러스터링 기법

퍼지 이론(fuzzy theory)이란 인간이 사용하는 애매한 표현을 이해하기 위해 주관성이 개입되는 애매성을 정량적으로 취급하여 정보의 손실을 최소화하고, 컴퓨터가 인간과 비슷한 판단 및 결정을 하도록 도와주는 방법을 말하는 것으로, Zadeh가 처음 소개하였다(이광형, 오길록 1991). 용어의 자동분류에 이러한 퍼지 개념을 도입함으로써 본질적으로 여러 클래스에 속할 수 있는 용어, 즉 의미모호성을 가진 용어를 정확하게 분류할 수 있다.

퍼지 이론에 근거한 퍼지 클러스터링(fuzzy clustering) 기법은 소프트 클러스터링(soft clustering)이라고도 하는데, 흑백논리에 입각한 단순 클러스터링 기법에 비해 의미모호성을 갖는 용어의 자동분류에서 그 효과가 더욱 클 수 있다.

퍼지 K-means 클러스터링 기법은 퍼지 클러스터링 기법의 일종으로 단순 K-means 클러스터링 기법과 마찬가지로 k개의 센트로이드를 기준으로 클러스터를 생성한다. 그러나 단순 K-means 클러스터링과 퍼지 K-means 클러스터링의 차이점은 데이터의 중복 분류 허용 여부에 있다. 퍼지 K-means 클러스터링에서는 퍼지 클러스터링의 원리와 같이 모든 데이터가 k개의 클러스터에 대한 소속함수값을 갖는다. 소속함수값의 기준치에 따라 클러스터의 중복도가 결정되기 때문에 소속함수값의 기준치가 높아질수록 클러스터의 중복도는 낮아지고 클러스터링 결과는 단순 K-means 클러스터링 결과와 유사하게 된다.

3. 개별문서의 지식구조 자동 생성 실험

3. 1 실험개요

대부분의 정보검색 실험에서는 실험 집단에 포함된 복수의 문서를 대상으로 다양한 기법들을 적용한다. 그러나 이 연구의 목적이 개별문서의 지식구조를 자동으로 생성하는 것이므로 실험 대상의 특성이 기존의 정보검색 실험과는 다르다. 이 실험에서는 효과적으로 개별문서의 지식구조를 생성하기 위해 국내에서 2000년에서 2003년 사이에 출판된 정보학 분야의 학위논문 중 연구 방법으로 실험 방법을 채택한 10편을 실험 대상으로 선정하였다.

개별문서의 지식구조를 자동으로 생성하기 위해 개별문서를 단락 단위로 분할하여 색인어를 추출하고, 단락 내 용어간의 연관성을 측정하는 텍스트 전처리 과정을 거쳐 단락-용어 행렬을 생성한 후, 용어 클러스터링 실험을 수행하였으며, 가장 성능이 우수한 용어 클러스터링 실험 결과를 대상으로 클러스터 대표어 선정 실험 및 평가를 거쳐 개별문서의 지식구조를 최종적으로 생성하였다.

3. 2 평가방법

실험의 결과로 제시되는 개별문서의 지식구조를 평가하기 위해서는 앞의 2.1에서 언급한 지역적 지식구조의 기능적 요소 분석 내용을 토대로 다음과 같은 측면을 고려하여야 한다.

첫째, 클러스터링 기법이 주제적으로 연관성 있는 용어들을 효과적으로 군집화해주는가?

둘째, 자동 생성된 지식구조가 저자의 의도

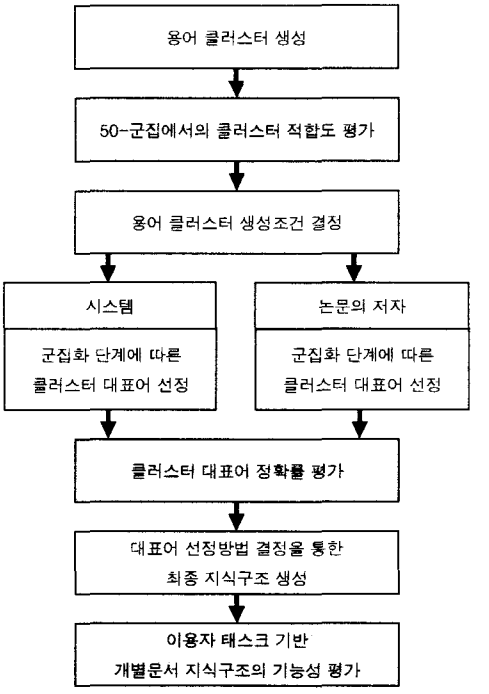
를 반영하는가?

셋째, 자동 생성된 지식구조를 이용하여 원문의 세부 주제에 대한 이해가 가능한가?

넷째, 자동 생성된 지식구조가 원문에 대한 전체적인 구조를 제시하는 원문의 대용물의 기능을 하는가?

위의 내용에서 첫번째 측면과 두번째 측면은 자동으로 지식구조를 생성하기 위해 거쳐야 하는 다양한 환경의 용어 클러스터링 실험과 클러스터 대표어 선정 실험 중에서 가장 적합한 환경을 선택하기 위한 것이다. 첫번째 측면을 평가하기 위해 다양한 용어 클러스터 생성 조건을 대상으로 용어 클러스터 적합도를 측정하여 최적의 생성 조건을 결정하였고, 이 클러스터링 결과를 토대로 대표어 선정 실험을 하여 생성된 지식구조가 원문에 나타난 저자의 의도를 반영하는가를 평가하기 위해 클러스터 대표어 정확률을 측정하여 최적의 대표어 선정 방법을 결정하였다.

세번째 측면과 네번째 측면은 최종적으로 자동 생성된 지식구조가 갖는 기능성을 평가하기 위한 것으로 이용자 태스크를 기반으로 평가가 이루어졌다. 세번째 측면은 자동 생성된 지식구조가 전통적인 인쇄매체 환경에서의 권말색인의 기능을 하는가를 평가하기 위한 것으로 이를 평가하기 이용자의 세부 주제 이해도를 평가하였고, 네번째 측면은 목차의 기능성을 평가하는 동시에 생성된 지식구조가 원문의 대용물로서의 기능을 할 수 있는가를 평가하기 위한 것으로 지식구조의 통보성을 측정하였다. 이용자의 세부 주제 이해도 평가와 지식구조의 통보성 평가는 제 4장에서 다루었다. 실험의 평가 과정은 <그림 2>와 같다.



〈그림 2〉 자동 생성된 지식구조의 평가 과정

(1) 클러스터 적합도

이 실험에서는 주제적으로 연관성 있는 용어들을 군집화하는 클러스터링 기법의 효과를 측정하기 위해서 다양한 조건에서 생성된 클러스터링 결과를 대상으로 각 논문의 저자에 의한 클러스터 적합성 판정 결과를 이용하여 클러스터 적합도를 측정하였으며, 그 공식은 다음과 같다.

$$\text{클러스터 적합도} = \frac{\text{클러스터 대표주제에 적합한 용어 수}}{\text{클러스터에 속한 용어 수}}$$

(2) 클러스터 대표어 정확률

클러스터 적합도를 측정하여 가장 우수한 성능을 나타내는 클러스터링 생성 조건을 선택한 후에는 세 가지의 클러스터 대표어 선정 방법을

적용하여 개별문서의 지식구조를 생성하였다.

시스템이 선정한 클러스터 대표어가 원문에 나타난 논문 저자의 의도를 얼마나 잘 반영하였는가를 평가하기 위하여 논문의 저자를 대상으로 클러스터 대표어 정확률을 측정하였다. 논문의 저자가 선정한 대표어를 기준으로 시스템이 선정한 대표어의 정확률을 계산하면 어떠한 대표어 선정 방식이 원문에 나타난 저자의 의도를 얼마나 효과적으로 나타내었는가를 평가할 수 있다. 클러스터 대표어 정확률의 공식은 다음과 같다.

$$\text{클러스터 대표어 정확률} = \frac{\text{논문저자와 시스템이 동일하게 선정한 대표어 수}}{\text{클러스터 대표어 총 수}}$$

3. 3 텍스트 전처리

3. 3. 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 있었는데, 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 10개 문서의 문단별 평균 단어 수를 계산한 결과 한 문단이 평균 29.62개의 단어를 포함하는 것으로 나타났다. 이를 근거로 이 실험에서는 30개의 단어를 한 문단으로 하는 고정길이 단락분할 기법을 적용하였다. 이 실험에서는 단락의 중복은 허용하지 않았다. 개별문서의 고정길이 단락 수는 <표 1>과 같다.

3. 3. 2 자동색인

단락 수준이 결정된 후에는 단락을 기준으로 형태소 분석을 이용하여 색인어를 추출하였다. 문서 내에 분포하는 고빈도어와 저빈도어는 시스템의 성능 향상에 긍정적인 영향을 주지 않는 것이 일반적이기 때문에 이 실험에서는 색인어 집합을 축소하기 위해 불용어 사전을 구축하여 고빈도 불용어를 제거하였으며, 단어빈도(tf)가 2 이하인 저빈도어를 제거하는 자질 축소 단계를 거쳐 최종 색인어 집합을 구축하였다. 그 결과는 <표 1>과 같다.

자동색인의 마지막 단계로 색인어에 가중치를 부여하였다. 이 실험에서는 가중치 공식으로 단락 내 단어빈도(tfp), 단락 내 이진 단어빈도($btfp$), 그리고 단락 내 단어빈도와 역단락빈도의 곱($tfp \cdot ipf$)을 이용하였다. N이 총 단락 수이고 n이 특정 용어 i를 포함한 단락 수일 때, 역단락빈도(ipf) 공식은 다음과

〈표 1〉 개별문서의 고정길이 단락 수와 고유 색인어 수

논문	A	B	C	D	E	F	G	H	I	J	평균
단락 수	173	161	217	188	160	188	168	176	217	141	178.9
고유 색인어 수	149	128	178	186	167	135	134	113	161	129	148.0

같다(박지연 2001).

$$ipf = \log_2 \frac{N}{n}$$

3. 3. 3 용어간 연관성 측정

색인어에 가중치 공식을 적용한 후 용어쌍간의 동시출현빈도를 나타내는 단락 - 용어 행렬에 유사계수를 적용하여 용어간의 연관성을 측정하였다. 이 실험에서는 코사인 유사계수(cosine coefficient)와 함께 동시인용 분석이나 동시출현 단어빈도 분석에 기초한 전역적 수준의 지식구조 분석 연구에서 많이 사용하고 있는 피어슨 상관계수(Pearson's coefficient)를 이용하였다. 용어 x와 용어 y에 대해 x_i는 단락 i에 출현한 용어 x의 가중치이며, y_i는 단락 i에 출현한 용어 y의 가중치일 때, 코사인 유사계수와 피어슨 상관계수(r) 공식은 다음과 같다(Sneath and Sokal 1973).

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

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

3. 4 용어 클러스터링 실험

3. 4. 1 클러스터 생성 조건

텍스트 전처리 단계를 거쳐 생성된 단락-용어간의 연관성 행렬을 가지고 다양한 조건의 용어 클러스터링 실험을 하였다. 이 실험에서의 클러스터 생성 조건은 <표 2>와 같다. 클러스터링 기법을 적용할 때에는 클러스터의 수를 고려해야 하는데, 이 실험에서는 휴리스틱을 적용하여 개별문서마다 50-군집, 25-군집, 10-군집의 단위로 클러스터 계층을 형성하였으며, 퍼지 K-means 클러스터링에서 군집이 정확하게 50, 25, 10개로 생성되지 않을 때에는 그와 유사한 수로 클러스터를 생성하도록 하였다.

tfp와 $btfp$, 그리고 $tfp \cdot ipf$를 용어가중치로 하여 워드 기법을 적용한 결과 tfp를 가중치로 이용했을 때와 $tfp \cdot ipf$를 가중치로

〈표 2〉 클러스터 생성 조건

클러스터 생성 조건	설 명		
	용어 가중치	유사계수	클러스터링 알고리즘
tfp_wc	단락 내 단어빈도	코사인 유사계수	워드 기법
tfp_wp	단락 내 단어빈도	피어슨 상관계수	워드 기법
tfpipf_wc	단락 내 단어빈도 × 역단락빈도	코사인 유사계수	워드 기법
tfp_ipf_wp	단락 내 단어빈도 × 역단락빈도	피어슨 상관계수	워드 기법
btfp_wc	단락 내 이진 단어빈도	코사인 유사계수	워드 기법
btfp_wp	단락 내 이진 단어빈도	피어슨 상관계수	워드 기법
tfp_fuzzyk	단락 내 단어빈도	피어슨 상관계수(※ 클러스터링 소프트웨어에서 결정된 사항임)	퍼지 K-means 기법
btfp_fuzzyk	단락내 이진 단어빈도		퍼지 K-means 기법

이용했을 때의 클러스터링 결과가 일치하여 나타났다. 이것은 용어 군집화 과정에서 *idf*가 아무런 영향력을 주지 못했다는 것을 의미하는데, 이러한 현상이 발생한 데에 대해서는 텍스트 전처리 과정에서 불용어 사전을 통해 고빈도어와 저빈도어 집합을 제거한 상태에서 *idf* 용어 가중치를 부여했기 때문에 고빈도어와 저빈도어의 편차를 정규화해주는 *idf*가 제 기능을 하지 못한 것으로 해석할 수 있다.

3. 4. 2 용어 클러스터 특성 분석

<표 3>은 워드 기법과 퍼지 K-means 기법으로 생성된 50-군집 클러스터링 결과를 대상으로 한 군집에 속한 평균 용어 수를 나타낸 것이다. 워드 기법에서는 평균적으로 2.96개의 용어가 한 군집에 소속된 반면, 퍼지 K-means 기법에서는 평균적으로 6.68개의 용어가 한 군집에 소속된 것으로 나타났다. 이것은 퍼지 K-means 클러스터링 기법이 하나의 용어를 복수의 군집에 중복 분류하기 때문인 것으로 해석할 수 있다. 또한 논문 C와 논문 D와 같이 워드 기법에서 한 클러스터에 속하는 용어의 수가 많은 경우에는 퍼지 K-means 기법에서도 마찬가지로 한 클러스터에 속하는 용어의 수가 많은 것으로 나타났다.

퍼지 K-means 기법에서 용어의 분류 중복 정도를 살펴보기 위해 한 용어가 평균적으로 몇 개의 군집에 포함되는지를 살펴보았다. <표 4>와 같이 퍼지 K-means 클러스터링 결과에서 용어의 평균 소속 군집 수를 분석한 결과, 평균적으로 한 용어가 약 2.2개의 군집에 속하는 것으로 나타났다.

3. 4. 3 클러스터 적합도 평가

다양한 용어 클러스터 생성 조건 중에서 최적의 조건을 선정하기 위해 논문의 저자의 적합성 판정 데이터를 이용하여 용어 클러스터의 적합도를 평가하였으며, 그 결과는 <표 5>와 같다.

논문 C의 성능이 특별히 다른 논문들의 성능에 비해 낮은 이유는 논문 C의 고정 단락의 수에 비해 고유 색인어의 수가 상대적으로 적어서 용어간 유사도가 다른 논문에 비해 제대

〈표 3〉 클러스터링 기법에 따른 한 군집 당 평균 소속 용어 수(50-군집)

조건 \ 논문		A	B	C	D	E	F	G	H	I	J	평균
워드 기법		2.98	2.56	3.56	3.72	3.34	2.70	2.68	2.26	3.22	2.58	2.96
퍼지 K-means 기법	tfp	6.35 (k=48)	5.02 (k=52)	10.08 (k=51)	6.98 (k=51)	9.14 (k=50)	6.29 (k=49)	6.18 (k=51)	4.08 (k=50)	6.70 (k=50)	6.02 (k=50)	6.68
	btfp	6.27 (k=51)	5.73 (k=49)	8.80 (k=49)	10.81 (k=48)	6.31 (k=49)	6.57 (k=49)	5.32 (k=50)	4.10 (k=50)	7.56 (k=48)	5.34 (k=50)	6.68

〈표 4〉 퍼지 K-means 기법에서 용어의 평균 소속 군집 수(50-군집)

가중치 \ 논문	A	B	C	D	E	F	G	H	I	J	평균
tfp	2.03	2.04	2.89	1.92	2.74	2.28	2.35	1.81	2.08	2.33	2.25
btfp	2.15	2.20	2.42	2.80	1.85	2.39	1.99	1.81	2.25	2.07	2.19

〈표 5〉 다양한 클러스터 생성 조건에서의 용어 클러스터링 성능(50-군집)

생성 조건 \ 논문	A	B	C	D	E	F	G	H	I	J	평균
tfp_wc	0.928	0.789	0.670	0.804	0.808	0.733	0.828	0.850	0.807	0.876	0.809
tfp_wp	0.938	0.792	0.607	0.798	0.790	0.756	0.821	0.850	0.807	0.868	0.803
btfp_wc	0.942	0.812	0.645	0.828	0.802	0.741	0.843	0.867	0.820	0.868	0.817
btfp_wp	0.925	0.798	0.596	0.811	0.814	0.770	0.828	0.858	0.814	0.829	0.804
tfp_fuzzyk	0.888	0.646	0.298	0.672	0.532	0.568	0.460	0.721	0.537	0.654	0.598
btfp_fuzzyk	0.851	0.620	0.352	0.601	0.542	0.559	0.511	0.722	0.496	0.787	0.604

로 측정되지 않았기 때문인 것으로 추정할 수 있다.

<표 5>에서 보는 바와 같이 가장 좋은 성능을 나타내는 클러스터 생성 조건은 용어 가중치로 단락 내 이진 단어빈도를, 유사계수로 코사인 유사계수를, 그리고 워드 기법을 사용한 경우(*btfp_wc*) 인 것으로 나타났다. 그러나 용어 가중치나 유사계수 같은 요인에 따라서는 클러스터링의 성능이 큰 차이가 없다는 것을 알 수 있다. 그러므로 이러한 조건의 변화가 클러스터링의 성능에 미치는 영향은 유의할만한 수준이 아니기 때문에 클러스터를 생성할 때 적절한 용어 가중치와 유사계수를 이용하면 된다.

그러나 클러스터링 알고리즘에 따라서는 클러스터링 성능의 차이가 큰 것을 알 수 있다. 알고리즘을 기준으로 클러스터링 성능을 비교해보면, 중복을 허용하지 않는 워드 기법이 중복을 허용하는 퍼지 K-means 기법보다 더 나은 성능을 보이는 것으로 나타났다. 이러한 현상은 결국 한 용어가 평균적으로 2개 이상의 군집에 중복되어 분류되면서 한 클러스터에서 복수의 주제가 함께 나타남으로써 논문의 저자가 클러스터 내 용어가 클러스터 대표 주제에 적합한가를 판정하는 데에 좋지 않은 영향을 준 것으로 해석할 수 있다.

또한 워드 기법의 클러스터 적합도가 낮을수록 일반적으로 퍼지 K-means 기법의 클러스터 적합도도 낮은데, 특히 용어 당 평균 소속 군집 수가 상대적으로 컸던 논문 C는 퍼지 클러스터링 기법을 적용한 경우 그 성능이 다른 논문에 비해 상당히 낮게 측정되었다. 이러한 점 역시 용어의 지나친 중복 분류가 오히려 클러스터링 성능의 저하 요인이 된다는 사실을 입증한다.

3. 5 클러스터 대표어 선정 및 지역적 지식구조의 생성

3. 5. 1 클러스터 대표어 선정 방법

용어 클러스터링 실험에서 가장 높은 클러스터링 성능을 나타낸 클러스터 생성 조건인 '*btfp_wc*' 기법으로 생성된 50-군집, 25-군집, 10-군집의 클러스터를 계층적으로 표현하기 위해 클러스터 대표어 선정 실험을 하였다. 이 실험에서 클러스터 대표어 선정을 위해 사용된 방법들은 다음과 같다.

첫번째 대표어 선정 방법은 단락빈도(*p f*)를 이용하는 것이다. 단락 단위의 문서분할 환경에서 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.

두번째 대표어 선정 방법은 단어빈도와 역단락빈도의 곱(tw)을 이용하는 것이다. 정영미와 이재윤(2001)은 문헌 클러스터링 연구에서 자질 축소를 위해 여러 문헌에 고르게 출현하는 단어보다는 일부의 문헌에 집중적으로 출현하는 단어가 문헌 클러스터의 식별에 도움이 될 것이라고 가정하고 다음과 같은 공식을 적용하였다.

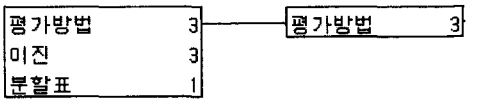
$$cw = ln(cf) \times \frac{cf}{df}$$

위와 같은 가정은 용어 클러스터의 대표어 선정에도 적용이 될 수 있다. 왜냐하면 여러 클러스터에 고르게 출현하는 용어보다는 특정 클러스터에 집중적으로 출현하는 용어가 클러스터의 주제 식별에 도움이 될 수 있기 때문이다. 이러한 이유로 이 공식을 용어 클러스터의 대표어 선정 방법 중 하나로 적용하였다. 이 공식을 단락 단위 문서분할 환경에 적합하도록 변형하면 다음과 같다.

$$tw = ln(tf) \times \frac{tf}{pf}$$

세번째 대표어 선정 방법은 클러스터 센트로이드와의 유사도($s\ cen$)를 이용하는 것이다. 10-군집을 기준으로 하여 각 군집의 클러스터 센트로이드를 계산하고, 그 센트로이드와의 코사인 유사도가 제일 높은 용어가 각 군집 수준에서 클러스터의 대표어가 된다. 어떠한 대표어 선정 방법을 사용하든지 각 클러스터에서 가장 큰 값을 갖는 용어를 대표어로 선정한다.

클러스터 대표어 선정을 위한 사전 실험에서 두 가지 문제점이 발생하였는데, 첫 번째는 복수의 용어가 최고 대표값을 갖는 경우에 어떠한 용어를 대표어로 선정할 것인가를 결정하는 문제이고, 다른 하나는 최고 대표값을 갖는 용어가 대표어로 선정되었으나 클러스터의 내용을 대표하지 못하는 경우에 생기는 문제이다.



〈그림 3〉 클러스터 대표어의 대표값이 같은 경우의 해결 방안

첫 번째 문제는 <그림 3>에서의 '평가방법'과 '이진'이라는 용어에서 찾아볼 수 있다. 이 두 용어는 동일한 대표값을 가지므로 그 클러스터의 대표어 후보가 된다. 이러한 경우를 해결하기 위해 이 실험에서는 길이가 긴 용어를 대표어로 선정하였는데, 정보검색 환경에서 일반적으로 문헌이나 문장의 길이가 길수록 그렇지 않은 문헌이나 문장에 비해 정보량이 크기 때문에 이와 같은 원칙을 용어에도 적용하였다.

두 번째 문제는 <그림 4>에서 찾아볼 수 있다. 그림에서 최고 대표값을 갖는 용어 '지속적'이 대표어로 선정되었으나 이 용어는 클러스터의 주제적인 특성을 반영하지 못한다. 이러한 문제점의 대안으로 두 개의 대표어를 선정하는 방법을 제안하였다. 최상위어 '지속적'과 함께 두 번째 상위어 제 2 대표어로 '뉴스'를 이용자에게 함께 보여주는 것이다. 그렇게 되면 이용자는 '지속적'과 '뉴스'를 결합하여 '지속적 뉴스'라는 개념으로 이 클러스터의 주

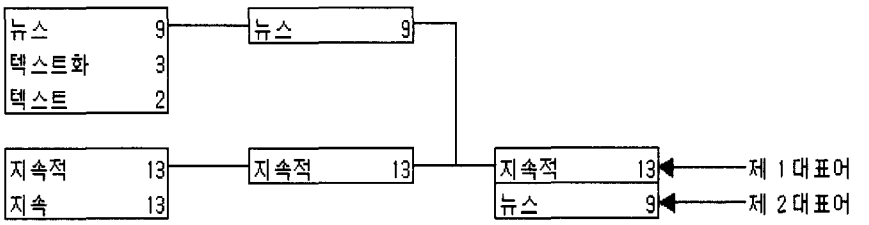
제를 이해하게 된다.

3. 5. 2 대표어 선정을 통한 지역적 지식구조의 생성

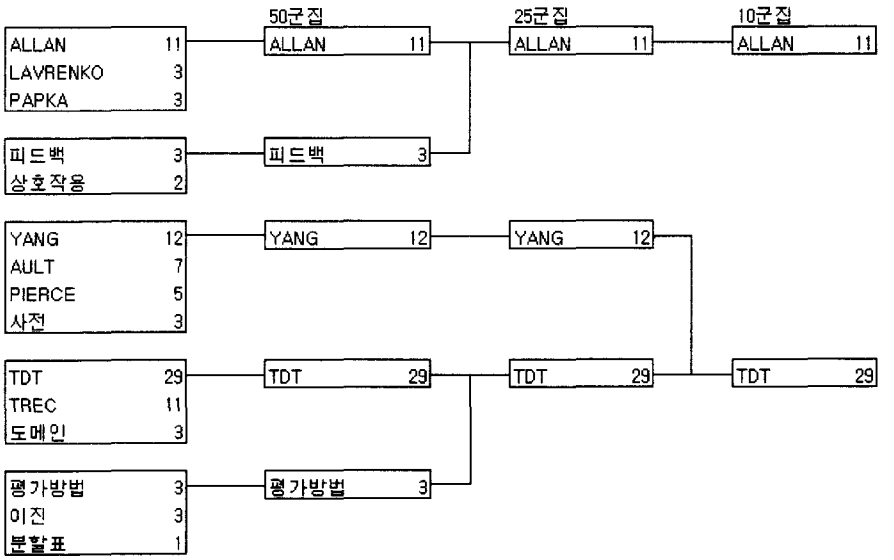
군집 수준에 따라 계층적으로 대표어를 선정하면 개별문서의 지식구조가 완성된다. <그림 5>는 논문 A를 대상으로 자동 생성된 지식구조 중 일부를 나타내고 있다. 지식구조의 기본 표현 단위는 용어와 그 용어의 pf이며, pf의 크기에 따라 군집 수준의 변화에 따른 클러스터 결합 구조를 표현한다. 군집 수준의 변화에 따라 대표어를 선정할 때에는 승자 노드를 채택하는 방식을 이용한다.

3. 5. 3 클러스터 대표어 정확률 평가

세 가지 클러스터 대표어 선정 방법 중 원문의 저자의 의도를 가장 잘 반영하는 최적의 방법을 결정하기 위해 클러스터 대표어 정확률을 측정하였다. <표 6>에서 보는 바와 같이 각 군집 수준에서 세 가지 대표어 선정 방법 중 pf를 이용한 방법이 50-군집에서 0.758, 25-군집에서 0.608, 10-군집에서 0.570의 정확률



〈그림 4〉 클러스터 대표어의 정보량이 낮은 경우의 해결 방안



〈그림 5〉 워드 기법에 단락빈도 대표어 선정 방법을 적용하여 생성한 논문 A의 지식구조의 일부

〈표 6〉 각 군집 수준에서의 대표어 선정 방법별 클러스터 대표어 정확률

군집 \ 방법 \ 논문		A	B	C	D	E	F	G	H	I	J	평균
50	pf	0.760	0.700	0.800	0.760	0.740	0.640	0.820	0.760	0.820	0.780	0.758
	tw	0.500	0.580	0.580	0.580	0.660	0.500	0.560	0.760	0.600	0.600	0.592
	s_cen	0.460	0.580	0.660	0.600	0.580	0.500	0.740	0.720	0.560	0.680	0.608
25	pf	0.520	0.440	0.760	0.680	0.640	0.400	0.760	0.600	0.800	0.480	0.608
	tw	0.360	0.320	0.440	0.320	0.600	0.400	0.440	0.520	0.640	0.440	0.448
	s_cen	0.400	0.200	0.560	0.440	0.400	0.320	0.520	0.400	0.320	0.480	0.404
10	pf	0.700	0.400	0.800	0.600	0.600	0.300	0.600	0.500	0.700	0.500	0.570
	tw	0.300	0.300	0.400	0.100	0.600	0.200	0.500	0.400	0.700	0.100	0.360
	s_cen	0.300	0.200	0.400	0.200	0.200	0.300	0.400	0.200	0.300	0.300	0.280

을 나타내어 가장 성능이 우수한 것으로 나타났다.

이 결과에서 한 가지 주목할 사실은 50-군집에서 10-군집으로 군집의 수를 줄여갈수록 클러스터 대표어의 정확률이 낮아진다는 것인데, 이것은 50-군집에서 결정된 승자 노드가 결국 25-군집과 10-군집에서의 대표어로 결정되는 방식 때문인 것으로 해석할 수 있다.

4. 자동 생성된 지역적 지식구조의 기능성 평가

4. 1 평가방법

클러스터 대표어 정확률 평가를 통해 클러스터 대표어 선정 방법이 결정되면 개별문서에 대한 최종의 지식구조가 생성된다. 이 지식구조가 전통적인 인쇄매체 환경에서의 목차나 권말색인의 기능을 하고 있는가를 평가하기 위해 이용자 태스크를 기반으로 하여 자동 생성된 지식구조의 기능성을 평가하였다.

전통적인 인쇄매체 환경에서 주제에 대해 상향식으로 접근하는 방식을 취하는 권말색인은 이용자가 특정 용어에서 시작하여 원문을 구성하는 용어에 대한 저자의 의도를 파악하고 그와 관련된 개념이나 상위의 개념들을 이해하도록 돕는다. 반면 목차는 이용자가 원문을 구성하는 상위의 개념이나 주제에서 시작하여 그와 관련된 하위의 개념이나 주제에 접근함으로써 원문의 구조적인 흐름을 이해하도록 돕는다.

이 실험에서는 이용자를 대상으로 세부 주제 이해도를 측정하여 지식구조가 갖는 상향식 기능성을 평가하였으며, 지식구조의 통보성을 측정하여 하향식 기능성을 평가하였다. 평가에 참여한 이용자 집단의 구성은 현재 정보학에 관심 있는 문헌정보학 전공 대학원생 10명으로 하였다.

(1) 원문의 세부 주제 이해도

원문의 세부 주제 이해도를 평가하기 위해서 이용자에게 50-군집에서부터 10-군집까지 군집화 수준에 따라 각 군집을 대표하는 용어를 선택하도록 하는 방식으로 클러스터 대표어 일치율을 측정하였다. 클러스터 대표어 일치율 공식은 다음과 같다.

$$\text{클러스터 대표어 일치율} = \frac{\text{이용자와 시스템이 동일하게 선정한 대표어 수}}{\text{클러스터 대표어 총 수}}$$

(2) 지식구조의 통보성

지식구조에 대한 하향식 기능성을 가지고 이용자가 원하는 정보에 접근할 수 있는가를 평가하기 위해 지식구조의 통보성을 평가하였다. 지식구조의 통보성 평가는 원문에서 발췌한 내용에 대한 이해를 묻는 질문-응답 방식으로 이루어졌으며, 다음의 공식과 같이 질문에 대한 응답 정확률을 계산하여 측정하였다.

$$\text{응답 정확률} = \frac{\text{이용자의 정답 수}}{\text{질문 총 수}}$$

4. 2 평가결과

(1) 원문의 세부 주제 이해도 평가

<표 7>에서 보는 바와 같이 50-군집에서의 평균 클러스터 대표어 일치율은 0.577, 25-군집에서는 0.751, 10-군집에서는 0.802로, 군집의 수가 작아질수록, 즉 군집화의 수준이 높아질수록 이용자-시스템과의 클러스터 대표어의 일치율이 높아지는 것을 알 수 있다.

이것은 결국 자동으로 생성된 지식구조가 제공하는 상향식 기능성을 통해 이용자는 전통적 인쇄매체 환경에서의 권말색인과 같이 원문을 구성하는 특정 개념간의 관계와 그에 내재된 저자의 의도를 파악함으로써 원문의 세부 주제에 대한 이해가 가능하다는 것으로 해석할 수 있다.

(2) 지식구조의 통보성 평가

<표 8>에서 보는 바와 같이 이용자는 전체 질문 중 약 70%를 정답으로 응답하였다. 이러한 결과를 통해 이용자는 상위 계층에서 시작하여 하위 계층을 탐색하면서 원문의 내용에 대한 구조적인 흐름을 파악하고 원문의 내용을 이해하였다고 할 수 있다. 결국 이 실험에서 생성된 개별문서의 지식구조를 통해 이루어진 상위 개념에서 하위 개념으로의 하향식 탐색 과정이 이용자의 정보 문제를 해결하는데 긍정적인 영향을 주었다고 해석할 수 있다. 그러므로 이 실험에서 생성된 개별문서의 지식구조는 하향식 기능성을 갖고 있다고 해석할 수 있다.

〈표 7〉 군집화 수준에 따른 이용자-시스템간의 클러스터 대표어 일치율

군집수준 \ 논문	A	B	C	D	E	F	G	H	I	J	평균
50-군집 평균	0.632	0.688	0.488	0.536	0.532	0.524	0.540	0.640	0.576	0.616	0.577
25-군집 평균	0.672	0.704	0.856	0.856	0.760	0.768	0.776	0.704	0.696	0.720	0.751
10-군집 평균	0.720	0.840	0.740	0.920	0.820	0.860	0.920	0.660	0.720	0.820	0.802

〈표 8〉 개별문서별 지식구조의 통보성 평가 결과

내용 \ 논문	A	B	C	D	E	F	G	H	I	J	평균
질문 수	6	4	5	4	4	5	5	5	6	4	4.80
평균 정답응답 수	5.20	2.80	2.80	2.80	3.40	2.60	3.80	4.00	3.20	3.40	3.40
평균 응답정확률	0.87	0.70	0.56	0.70	0.85	0.52	0.76	0.80	0.53	0.85	0.71

5. 결론 및 제언

이 연구에서는 전통적인 인쇄매체 환경에서 지식에 대한 지역적 접근법을 제공하는 권말색인과 목차의 특성과 기능에 착안하여 용어 클러스터링 실험과 클러스터 대표어 선정 실험을 통해 개별문서의 지식구조를 자동으로 생성하였고 그 기능성을 평가함으로써, 정보검색 환경에서의 적용 가능성을 확인하였다. 그러므로 이 연구는 정보검색 환경에서 이용자를 중심으로 한 새로운 유형의 서비스에 대한 기반 연구로서의 성격을 갖는다고 할 수 있다.

이 연구에서는 효과적으로 개별문서의 지식구조를 생성하기 위해 계층적 클러스터링 기법 중 워드 기법과 중복을 허용하는 퍼지 K-means 클러스터링 기법을 적용해 보았다. 워드 기법은 군집을 작고 균일하게 나누는 경향성 때문에 용어 클러스터링에서 많이 이용되고 있으며, 퍼지 K-means 알고리즘을 적용하여 용어 자동분류 환경에서 문제가 되는 의미모호성 문제를 해결하고자 하였다. 용어 클러스터링의 성능은 50-군집에서의 원문의 저자에 의한 클러스터 적합도로 측정하였다.

용어 클러스터링 결과 퍼지 K-means 클러스터링 기법이 워드 기법에 비해 낮은 클러스터 적합도를 나타냈는데, 이러한 현상은 한 용어가 평균적으로 2개 이상의 군집에 과도하게 중복되어 분류되었기 때문인 것으로 해석할 수 있다. 용어 클러스터링 성능이 클러스터링 기법간에는 큰 차이를 보인 반면 용어 가중치나 유사계수와 같은 기타의 생성 조건간에는 별 차이가 없었다. 결국 용어 가중치와 유사계수의 차이는 클러스터링 성능을 결정하는데 크게 영향을 미치지 않는 것으로 나타났다.

클러스터 적합도 평가를 통해 단락 내 이진 단어빈도 가중치와 코사인 유사계수를 이용하여 워드 기법으로 클러스터를 생성(*btp wc*)한 결과가 0.817로 가장 성능이 우수한 것으로 나타났다.

최적의 성능을 나타내는 용어 클러스터 생성 조건을 결정한 후에는 이를 대상으로 50-군집, 25-군집, 10-군집에서의 클러스터 결합 과정을 반영하여 단락빈도(*pf*), 단어빈도×역단락빈도(*tw*), 클러스터 센트로이드와의 유사도(*s cen*)와 같은 세 가지 클러스터 대표어 선정 방법을 적용하였다. 그리고 세 가지 방법 중 논문의 저자의 의도를 가장 유사하게 표현한 대표어 선정 방법을 결정하기 위해 클러스터 대표어 정확률을 측정한 결과 단락빈도(*pf*)가 50-군집에서 0.758, 25-군집에서 0.608, 그리고 10-군집에서 0.570의 값으로 가장 효과적인 대표어 선정 방법이라는 것을 밝혀냈다.

최적의 클러스터 대표어 선정 방법을 결정하여 개별문서에 대한 최종의 지식구조가 생성된 후에는 이용자 태스크를 기반으로 하여 자동 생성된 지식구조가 갖는 기능성을 평가하였는데,. 권말색인과 목차의 기능성 분석을 기초로 하여 이용자의 세부 주제 이해도와 지식구조의 통보성을 평가하였다.

이용자의 세부 주제 이해도 평가 결과 50-군집에서의 클러스터 대표어 일치율(0.577)보다 10-군집에서의 클러스터 대표율(0.802)이 높게 나타났고, 지식구조의 통보성 평가 결과 평균적으로 약 0.7의 응답정확률을 나타났다. 이것은 결국, 이 연구에서 제안한 지식구조가

전통적인 인쇄매체 환경에서의 목차와 권말색인의 기능을 어느 정도 수행한다는 것이 증명한다. 그러므로 이 연구에서 제시한 지역적 지식구조의 생성 기법은 온라인 환경에서 목차나 권말색인의 기능을 대신하거나 적어도 보완할 수 있는 새로운 유형의 원문에 대한 대용물을 생성하기 위해 적용될 수 있다.

이 연구의 향후 과제는 다음과 같다.

첫째, 이 연구에서 사용된 기법들을 다른 성격의 문서집합에 적용해 볼 필요가 있다. 이 실험에서 사용한 학위논문은 원래 그 구조가 명확하기 때문에 용어간의 연관성 측정에 유리하다. 그러므로 이 연구의 결과를 일반화하기 위해서 신문이나 웹 문서 등과 같이 다양한 성격의 문서집합에 대해 지식구조를 자동으로 생성하는 실험을 해 볼 필요가 있다.

둘째, 퍼지 K-means 클러스터링 기법을 이용한 용어의 중복 분류와 관련하여 용어 클러스터링에 적합한 용어의 중복 분류 수준에 대해 연구할 필요가 있다. 이 실험결과에서는 퍼지 K-means 클러스터링 기법이 과도한 중복을 허용한 나머지 오히려 클러스터링 성능이 중복을 허용하지 않는 클러스터링 기법에 비해 낮게 나타났는데, 이러한 결과는 용어의 중복 분류에 대한 허용 수준에 따라 달라질 것으로 보인다. 그러므로 다양한 크기의 용어 집단을 대상으로 퍼지 클러스터링에 대한 심도 있는 연구가 필요하다.

셋째, 용어 클러스터링에 이용되는 용어의 수를 효과적으로 축소하는 기법에 대해 연구해야 할 필요가 있다. 이 연구에서는 불용어 사전을 구축하고 저빈도어를 제거하는 등 텍스트 전처리 과정에 많은 시간이 소요되었다. 따라서 용어 클러스터링에 적합한 용어 자질 축소 기법과 축소 수준에 대한 연구가 필요하다.

참 고 문 헌

김효열. 1995. 『도서권말색인의 작성지침과 자동생성에 관한 연구』. 석사학위논문, 연세대학교, 문헌정보학과.

박지연. 2001. 『질의확장에 의한 단락검색의 성능 향상에 관한 연구』. 석사학위논문, 연세대학교 대학원 문헌정보학과.

서은경. 1984. 『용어의 자동분류에 관한 연구』. 석사학위논문, 연세대학교 대학원 도서관학과.

유안나. 1992. 『원문대표정보의 비교평가에 관한 연구』. 석사학위논문, 연세대학교, 문헌정보학과.

이광형, 오길록. 1991. 『퍼지 이론 및 응용: 1권 이론』. 서울: 홍릉과학출판사.

정영미, 이재윤. 2001. 지식분류의 자동화를 위한 클러스터링 모형 연구. 『정보관리학회지』, 18(2): 203-230.

한승희. 2003. 『용어 자동분류를 위한 퍼지 클러스터링 기법 분석』. 제10회 한국정보관리학회 학술대회 논문집, 2003년 8월 22일-23일. [서울: 이화여자대학교 포스코관]. 95-103.

Bezdek, James C. 1981. *Pattern Recognition with Fuzzy Objective Function Algorithms*. New York: Plenum Press.

Cao, Guihong, Dawei Song, and Peter Bruza. 2004. "Fuzzy K-means Clustering on a High Dimensional Semantic Space". *Advanced Web Technology and Applications: 6th Asia-Pacific Web Conference (APWeb 2004)*.

Ding, Ying, Gobinda G. Chowdhury, and Schubert Foo. 2001. "Bibliometric Cartography of Information Retrieval Research by Using Co-word Analysis". *Information Processing & Management*, 37: 817-842.

Gaush, Audrey P., and Michael B. Eisen. 2002. "Exploring the Conditional Coregulation of Yeast Gene Expression through Fuzzy K-means Clustering". *Genome Biology*, 3(11): 1-22.

Hearst, Marti, A., and Jan O. Pedersen. 1996. "Reexamining the Cluster Hypothesis: Scatter/Gather on Retrieval Results". *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval(SIGIR '96)*, 76-84.

Ingwersen, P., B. Larsen, and E. Noyons. 2001. "Mapping National Research Profiles in Social Science". *Journal of Documentation*, 57(6): 715-740.

Jain, A. K., M. N. Murty, and P. J. Flynn. 1999. "Data Clustering: A Review". *ACM Computing Surveys*, 31(3): 264-323.

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.

Mima, H., S. Ananiadou, and G. Nenadiĉ. 2001. "ATRACT Workbench: An Automatic Term Recognition and Clustering of Terms". In Matoušek, V. et al. Eds. *Text, Speech and Dialog(TSD 2001)*. Berlin: Springer, 126-133.

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.

Song Dawei, Guihong Cao, and Peter Bruza. 2003. "Fuzzy K-means Clustering in Information Retrieval". [pdf file]. *Distributed Systems Technology Centre Technical Report*. <http://www.dstc.edu.au/Research/Projects/Infoeco/publications/tech-report-K-means.pdf>.

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

White, Howard D., and Katherine W. McCain. 1998. "Visualizing a Discipline: An Author Co-Citation Analysis of Information Science, 1972-1995." *Journal of the American Society for Information Science*, 49(4): 327-355.