용어 자동분류를 위한 퍼지 클러스터링 기법 분석

Analytical Study of Fuzzy Clustering Technique for Automatic Term Classification

한승희, 연세대학교 대학원 문헌정보학과

Han, Seung-Hee
Dept. of Library and Information Science, Graduate School of Yonsei University

목차 및 권말색인과 같은 인쇄형태의 정보내용에 대한 구조화된 접근방식에서 착안하여 전자 문서의 내용에 대한 새로운 형태의 접근방식을 개발할 수 있는데, 이를 위한 방안으로 용어 자동분류 기법이 있다. 본 연구에서는 용어의 의미모호성 문제를 해결하는 동시에 용어간의 계층관계 표현이 가능한 자동분류 기법으로 퍼지 클러스터링 기법을 제안하고, 대표적인 퍼지 클러스터링 알고리즘인 퍼지 c-means 기법에 대해 분석하고자 한다.

1 서 론

인터넷의 대중화로 정보 이용자들은 온라인과 오프라인 정보원을 통해 많은 양의 정보를 이용하고 있다. 그 결과, 정보를 이용하는데 드는 것보다 더 많은 시간과 노력을 정보의 조직이나 관리에 할애하고 있다. 정보 과부하(information overload) 문제가 심각해지면서 수집한 정보에서 양질의 정보를 선택해야 하는 이용자의 역할이 커지고 있어, 이용자 중심의 체계적인 정보 관리 방법이 절실하게 필요하다.

한편, 이용자들은 전통적인 인쇄매체에서 정보 내용에 접근하기 위해 정보 구조를 표현하고 있는 목차나 권말색인 등을 이용하고 있다. 전자 문서는 하이퍼텍스트 기술을 적용하면서 정보 자체에 대한 접근이 용이해졌으나, 과다한 정보의 양과 하이퍼텍스트 기술이 갖는 단순한 링크 연결과 같은 낮은 구조화 수준으로 인해 정보 내용의 접근이 더욱 어려운 것이 사실이다.

전통적인 인쇄매체의 접근방식을 전자문서에 적용한다면, 새로운 방식으로 많은 양의 정보를 관리할 수 있다. 대부분의 이용자들은 텍스트의 제목과 같은 일부 제한된 요소들을 가지고 검색결과의 적합성 판정을 내린다. 그러나 온라인 정보 검색 환경에 적합하도록 정보 내용의 구조를 자동으로 생성한다면, 하이퍼텍스트 기술에만 의존했던 기존의 정보 접근방식보다 효율적으로 정보 내용에 접근할 수 있다. 수 천건의 검색 문헌을 대상으로 개별 문헌의

주요 개념을 식별하고, 그 개념을 표현하는 구조를 자동으로 생성하여 이용자에게 제공한다면, 이용자는 대부분의 검색엔진에서 현재 제공하고 있는 순위 기반 검색 서비스보다 쉽게 정보를 브라우징하거나 접근할 수 있다.

이에, 본 연구에서는 전자 문서를 자동으로 분석하고 그 구조를 생성하기 위한 방안으로 용어 자동분류 기법을 제시하고자 한다. 구체적으로는, 용어 자동분류의 개념과 연구동향을 살피고, 대표적인 자동분류 기법들을 대상으로 정보 구조 생성 도구로써의 적용가능성과 문제점을 분석한다. 그리고 분석 내용을 바탕으로 용어의 자동분류에 적합한 기법으로 퍼지 클러스터링 기법을 제안함으로써 앞으로의 용어 자동분류 연구를 위한 이론적인 기초를 마련하고자 한다.

2 용어 자동분류 연구의 전개와 논의

2.1 용어 자동분류의 개념

용어 자동분류(automatic term classification) 또는 용어 클러스터링(term clustering)이란 용어의 문맥적인 성질을 근거로 하여 용어의 의미를 결정, 동의어 및 관계어군을 자동으로 만드는 방법이다(서은경 1984). 즉, 컴퓨터에 입력된 용어들 가운데 서로 관련 있는 용어들을 일정한 기준에 따라 모아서 여러 개의 용어 클래스를 형성하는 것을 말한다(정영미 1993).

전문가가 수작업으로 용어를 분류하여 시소러스나 주제명 표목을 구축하는 경우에는 시간과 노력이 많이 들고, 객관적으로 일관성 있는 분류결과를 얻기 어렵다. 또한 수작업에 의한 시소러스나 주제명 표목은 일반적이고 전역적인(global) 수준의 개념체계로서, 일반적인 개념간의 관계를 표현하는데 효과적인 반면 최신의 학문분야를 표현하는 개념이나 구체적인 개념을 즉각적으로 반영하기가 어렵다.

이러한 문제점을 해결하기 위해 많은 연구자들이 자동 용어 분류에 관심을 갖기 시작했다. 특히 1960년대 후반에 들어 지식의 자동분류를 위해 클러스터링의 개념이 도입되고 컴퓨팅 기술의 발전으로 인해 그 처리 속도가 향상되면서, 1980년대부터 용어뿐만 아니라 문헌을 대상으로 한 자동분류 연구가 증가했다.

현재 용어 자동분류는 탐색용 시소러스를 이용한 질의확장뿐만 아니라 정보조직 및 접근 도구와 데이터 마이닝 등 그 적용범위가 넓다.

2.2 탐색용 시소러스와 질의 확장

일반적으로 용어 자동분류의 목적은 자동 구축된 탐색용 시소러스를 이용하여 탐색자의 초기 질의와 관련된 용어를 새로운 질의에 자동으로 추가함으로써 부적절한 질의어로 인해 발생하는 검색 성능의 저하를 막고 정보검색의 효율을 높이는데 있다. 초기의 자동분류 연구 역시 이러한 목적에

기초를 두고, 탐색용 시소러스를 작성하여 질의 확장에 필요한 관련 용어군을 수집하는 것에 초점을 두었다.

탐색용 시소러스의 구축은 텍스트를 대상으로 한 용어간의 연관성(term association) 분석에 기초한다. 용어 자동분류에서 용어간의 연관성은 주로 문헌집단 내 용어들의 동시출현빈도를 이용하여 측정하는데, 두 개의 용어가 많은 수의 문헌 속에 함께 출현하였다면, 이 두 용어는 서로 관련이 있다고 보고 같은 클래스에 포함시킨다(정영미 1993). 실제로, 단어의 동시출현에 기초하여 문서에서 직접 수집된 용어들이 수작업 시소러스보다 질의 확장에 더욱 효율적인 경우도 있는 것으로 나타났다(Ekmekcioglu, Robertson, and Willet 1992).

2.3 용어 자동분류와 의미모호성 문제

용어의 자동분류에 관한 연구는 정보 검색에서의 단어 불일치 문제를 해결하기 위해 시작되었다고도 할 수 있다. 여기서 단어 불일치 문제란 동음이의어나 이음동의어와 같이, 정보 검색 환경에서 동일한 개념에 대해 문헌의 저자와 정보탐색자의 표현이 일치하지 않거나, 반대로 다른 개념에 대해 저자와 탐색자의 표현이 일치하는 것을 의미한다.

정보 검색에서 단어 불일치 문제는 결국 검색의 정확률을 감소시킨다. 이러한 문제를 해결하기 위해 많은 연구자들이 지역적 문맥 분석이나 질의 확장과 같은 해결책을 개발해왔다. 그러나 용어의 자동분류와 관련된 기존 연구들을 살펴보면, 주로 용어의 연관성을 찾아내는 기법에 초점을 두어왔을 뿐, 용어의 의미모호성(word-sense ambiguity) 문제에 대해서는 거의 고려하지 않고 있다. 물론 문서집합의 성격에 따라 용어의 의미모호성 고려여부가 용어 자동분류의 성능에 영향을 줄 수는 있다. 예를 들면, 전문용어의 출현빈도가 높은 주제특정적인 영역의 컬렉션에서는 용어의 의미모호성을 발견하기 어렵다. 반면, 신문과 같은 비주제영역의 문헌 집합에서 의미모호성을 갖는 용어를 찾아보는 것은 어려운 일이 아니다. 그러므로 효율적인 용어분류 시스템을 구축하기 위해서는 용어의 의미모호성 문제를 고려해야 할 필요가 있다.

3 용어 자동분류 기법

Wu(2001)의 연구에 의하면, 효율적인 용어 자동분류 시스템을 구축하기 위해 다음과 같은 네 가지 사항을 고려해야 한다.

① 용어 자동분류 시스템의 자동분류 결과는 일반적인 분류체계나 시소러스에서 개념을 표현하고 조직하는 방식과 유사하게 계층적인 구조로 표현되어야 한다.
② 개별 용어는 다른 용어들과 함께 적어도 하나 이상의 계층 관계를 가져야 한다.
③ 의미모호성을 갖는 용어를 인정해야 한다. 즉, 특정 용어가 하나 이상의 의미를 갖는 경우는 하나 이상의 계층 관계로 표현될 수 있어야 한다.
④ 용어 자동분류의 결과는 현재 지식과

학문의 경향을 표현하기 위해 개별문서나 문서집합에 의존적(collection-dependent)이어야 한다.

용어 자동분류에 관한 기존의 연구들을 살펴보면, 텍스트에서 연관된 용어군을 찾기 위해 베이즈 기법, 클러스터링, 신경망, 지역적 문맥 분석과 같은 기법들을 적용해 왔다. 본고에서는 이 중에서 가장 많이 이용되는 통계기반 용어 클러스터링 기법과 신경망 기반 자동분류기법의 특징을 살펴보고, 위에서 언급한 고려사항과 비교하여 이들이 용어 자동분류에 적용되었을 때 갖는 장·단점을 분석하였다.

3.1 통계기반 용어 클러스터링

클러스터 분석(cluster analysis)의 목적은 데이터 집합이 갖고 있는 구조를 발견(structure- seeking)하는 것이다. 즉, 데이터 집합을 구성하는 객체간의 통계적 유사성에 기초하여 객체를 군집화함으로써 벡터 공간 상에서 가까운 거리에 있는 객체들은 같은 클러스터에, 그렇지 않은 객체들은 다른 클러스터에 포함시키는 일련의 분류작업을 거쳐 데이터 집합의 구조를 탐색하는 것이다. 일반적으로, 한 데이터 집합에는 서로 다른 클러스터들이 함께 존재하게 된다. 일반적으로 클러스터링 기법은 크게 비계층적 기법(non-hierarchical clustering)과 계층적 기법(hierarchical clustering)으로 나뉜다.

비계층적 클러스터 분석 기법 중 하나인 k-means 기법은 미리 선택된 k개의 센트로이드를 중심으로 센트로이드와 객체와의 거리를 최소화할 때까지 n개의 데이터를 k개의 클러스터로 나누는 방식이다. 그러므로 이러한 비계층적 기법은 재배치 기법이라고도 하는데, 일반적으로 시간과 비용을 최소화할 수 있다는 장점을 가지고 있다. 그러나, 클러스터링 결과가 k개의 센트로이드 선택에 따라 많은 영향을 받는다.

비계층적 기법을 용어 클러스터링에 적용할 때 발생하는 첫 번째 문제점은, 용어간의 계층관계를 표현할 수 없다는 것이다. 또한, 비계층적 클러스터링 알고리즘을 적용하면 개별 용어는 오직 하나의 클러스터에만 속할 수 있는데, 이것은 하나 이상의 의미를 포함하고 있는 의미모호성을 갖는 용어들은 동시에 다른 클러스터에 속할 수 없다는 것을 뜻한다.

계층적 기법은 유사도가 강한 데이터를 포함하는 작은 클러스터를 상대적으로 유사도가 낮은 데이터를 포함하고 있는 좀 더 큰 클러스터에 포함함으로써 트리 구조로 데이터를 분류하는 기법을 말한다. 이 기법에서 보여주는 트리 형태의 데이터 구조를 덴드로그램(dendrogram)이라 한다. 계층적 기법은 일반적으로 상향식의 응집적(agglomerative) 기법과 하향식의 분할(divisive) 기법으로 나눌 수 있으며, 특히 계층적 응집 알고리즘에는 단일연결(single linkage), 완전연결(complete linkage), 그룹평균연결(group average linkage), 워드 기법(Ward's method)이 있다. 일반적으로

계층적 기법이 비계층적 기법에 비해 클러스터링 성능이 우수하나, 처리 시간 및 속도가 오래 걸린다.

계층적 기법을 용어 자동분류에 적용하면 덴드로그램을 통해 용어간의 계층관계 및 개념상 서로 가까운 용어를 쉽게 식별할 수 있다는 장점이 있으나, 트리 구조의 가장 하위에 있는 개별 가지는 고유한 개별 용어를 나타내기 때문에, 용어들은 오직 하나의 계층관계만을 갖게 된다. 즉, 계층적 클러스터링 기법 역시 비계층적 기법과 마찬가지로 용어의 의미모호성을 고려하지 않는다.

3.2 신경망 기반 용어 자동분류

용어나 개념을 분류하거나 시소러스를 구축하기 위해 Kohonen의 자기조직화 신경망(Self-Organizing Map, K-SOM)이나 홉필드망(Hopfield net)과 같은 신경망 기반 알고리즘을 이용할 수 있다.

K-SOM은 입력층과 출력층으로 구성되는 자율학습 신경망 알고리즘으로, 입력층의 노드들은 임의의 연결강도를 가지는 모든 출력층의 노드들과 연결되어 있으며, 가장 작은 유클리드 거리를 가지는 출력노드가 승자노드로 선택되어 출력노드에 입력노드가 매핑되는 방식으로 클러스터링하는 방식을 의미한다(Orwig, Chen, and Nunamaker 1997).

SOM 알고리즘의 장점 중 하나는 데이터의 분류결과를 지도 형태로 보여줄 수 있다는 것이다. 이러한 지도 형태의 디스플레이는 계층적 클러스터링 기법에서 보여주는 트리 형태 구조보다 용어간의 관계를 쉽게 표현할 수 있다. 지도상 영역의 크기는 데이터의 상대적인 중요도를, 인접영역은 특정 개념에 대한 관련 개념을 의미한다. 개념 분류를 위해 신경망 알고리즘을 적용한 연구 중에는 웹 문서의 자동 분류에 SOM을 적용한 WEBSOM이 있다.

그러나 용어 자동분류에 SOM 알고리즘을 적용하는 경우에도, 앞에서 언급한 클러스터링 기법과 마찬가지로 용어의 의미모호성 문제를 고려하기 어렵다. 또한 신경망 기반 자동분류는 처리 속도가 느려서 시간이 오래 걸린다는 단점이 있다. SOM을 이용한 자동분류의 성능과 관련해서는, 개념의 자동분류에 있어 K-SOM 기법이 워드 클러스터링 기법보다 정확하지 않다는 연구 결과가 있다(Roussinov and Chen 1999).

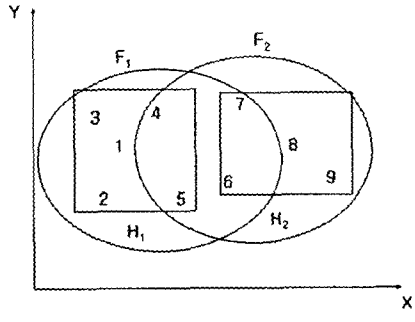
4 퍼지이론과 용어 자동분류

4.1 용어분류에서의 퍼지이론의 의의

단순 용어 클러스터링(hard clustering) 연구에서는 특정 클러스터에 용어를 포함하거나 불포함하는 이진논리를 적용해왔다. 즉, 하나의 용어는 하나의 클래스에만 속하도록 하는 것이 일반적이었다. 그러나 클래스 간의 경계가 모호하거나, 용어 자체의 의미가 모호하여 다의성을 띄는 경우에는 엄격한 용어분류가 불가능한 경우가 있다.

이러한 문제는 경계가 모호한 단어들을 하나의 클래스에 지정하는 것이 아니라 여러 클래스에 지정하는 방식으로 해결할 수 있다.

퍼지이론(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 가지고 있는 용어를 정확하게 분류할 수 있다.



〈그림 1〉 단순 클러스터와 퍼지 클러스터
(Jain, Murty, and Flynn 1999)

4.2 퍼지 클러스터링의 개념

퍼지이론에 근거한 퍼지 클러스터링 기법(fuzzy clustering, soft clustering)은 흑백논리에 입각한 단순 클러스터링 기법에 비해 의미모호성을 갖는 용어의 자동분류에서 그 효과가 더욱 크다.

단순 클러스터링과 퍼지 클러스터링의 차이는 〈그림 1〉과 같이 설명할 수 있다(Jain, Murty, and Flynn 1999). 데이터 $X=1,2,3,4,5,6,7,8,9$에 대해 사각형으로 둘러싸인 H_1과 H_2는 일반 클러스터를, 타원으로 둘러싸인 F_1과 F_2는 퍼지 클러스터를 나타낸다. 퍼지 클러스터링에서는 모든 데이터가 각 클러스터에 대한 [0,1]의 소속함수값(membership function value)을 갖게 되며, 각각의 클러스터는 모든 데이터에 대한 퍼지 집합이 된다. 예를 들어, 퍼지 클러스터 F_1과 F_2는 아래와 같은 소속함수값으로 표현할 수 있다. 순서쌍 (x_j, u_{ij})은 데이터(x_j)와 클러스터 i에 대한 데이터 x_j의 소속함수값(u_{ij})을 나타낸다. 소속함수값이 크면 클수록 클러스터에 대한 데이터 할당 신뢰도가 높다.

F_1=(1,0.9),(2,0.8),(3,0.7),(4,0.6),(5,0.55),(6,0.2),
(7,0.2),(8,0.0),(9,0.0)

F_2=(1,0.0),(2,0.0),(3,0.0),(4,0.1),(5,0.15),(6,0.4),
(7,0.35),(8,1.0),(9,0.9)

한편, 퍼지 클러스터링 기법에 계층적 클러스터 분석기법을 적용하기 위한 연구에서는 다양한 유사도 측정 기법을 이용하여 기존의 퍼지 클러스터링의 단점으로 지적되고 있는 계산 복잡도의 문제를 해결하는 동시에 기존의 퍼지 클러스터링 방법에 비해 향상된 클러스터링 성능을 보여주었다(정혜천 1999).

4.3 퍼지 c-means 기법

정확한 퍼지 클러스터링 결과를 얻기 위해 가장 중요한 것은 최적의 소속함수값으로 구성된 행렬 $U=u_{ij}$을 얻는 것이다. 행렬 U는 일반적으로 다음과 같은 조건을 만족해야 한다.

조건 1: 모든 i에 대해, $0<\sum_{j}u_{ij}<1$이다.

조건 2: 모든 j에 대해, $\sum_{i}u_{ij}=1$이다.

퍼지 클러스터링 기법 중 가장 일반적으로 쓰이는 퍼지 c-means 기법(Bezdek 1981)은 최적의 행렬 U를 계산하기 위해 데이터와 클러스터 센트로이드와의 제곱 오차합을 최소화하는 방식으로 아래의 〈공식 1〉과 같은 목적함수 J를 취한다. 즉, 입출력 공간상의 데이터 x_j에 대한 미지의 센트로이드 v_i를 구하기 위해 퍼지 논리를 적용하여 각각의 데이터에 대한 퍼지 소속함수값 u_{ij}를 계산하고, 이를 통해 목적함수 J를 최소화함으로써 효율적으로 데이터를 클러스터링하는 기법이다.

$$J(u_{ij},v_k,x_j)=\sum_{i=1}^{c}\sum_{j=1}^{n}(u_{ij}^{m})\,|x_j-v_i|^2\,,\quad m>1$$

〈공식 1〉

이 때 m은 퍼지 소속함수 u_{ij}의 퍼지 정도(fuzziness)를 나타내는 지수형 계수이다. 목적함수의 최소화 문제를 해결하기 위한 미지의 센트로이드 v_i와 소속도 u_{ij}는 아래의 〈공식 2〉, 〈공식 3〉과 같이 계산한다.

$$v_i=\frac{\sum_{j=1}^{n}(u_{ij})^m x_j}{\sum_{j=1}^{n}(u_{ij})^m}\,,\qquad i=1,2,\ldots,c$$

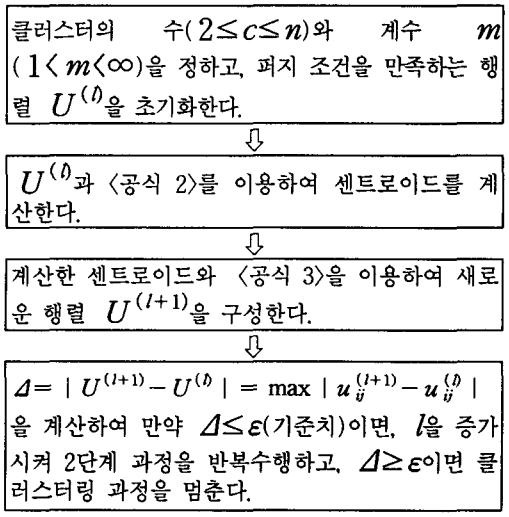
〈공식 2〉

$$u_{ij}=\frac{\left(\frac{1}{|x_j-v_i|^2}\right)^{\frac{1}{m-1}}}{\sum_{k=1}^{c}\left(\frac{1}{|x_j-v_k|^2}\right)^{\frac{1}{m-1}}}\,,$$

$$i=1,2,\ldots,c\qquad j=1,2,\ldots,n$$

〈공식 3〉

퍼지 c-means 클러스터링 알고리즘은 다음과 같이 4단계의 과정을 거쳐 이루어진다.



〈그림 2〉 퍼지 c-means 알고리즘

5 결론 및 제언

용어의 자동분류에 대한 연구는 탐색용 시소러스를 구축하여 질의를 확장함으로써 정보검색의 효율성을 높이기 위해 시작했다. 그러나 기존의 용어 자동분류에 대한 실험적인 연구를 살펴보면, 용어의 의미모호성 해소나 용어간의 관계 표현에 대해서는 연구가 아직 미흡하다.

본고에서는 용어의 의미모호성을 해소하는 동시에 용어간의 관계표현이 가능한 용어 자동분류 기법의 대안으로 퍼지 클러스터링 기법을 제안했다. 하나의 용어를 하나의 클래스에만 포함하는 기존의 분류기법에 반해, 퍼지 클러스터링 기법은 하나의 용어를 하나 이상의 클래스에 포함함으로써, 용어의 의미모호성을 해소할 수 있으며, 이러한 퍼지 클러스터링 기법에 계층형 알고리즘을 적용함으로써 용어간의 계층관계를 표현할 수 있다.

온라인 정보 환경에서 용어의 자동분류 기법은 탐색용 시소러스를 이용한 질의 확장 이외에, 개별 문서단위의 개념체계를 구축하기 위한 기법에도 적용할 수 있다.

개별 문서를 대상으로 생성되는 지역적(local) 수준의 개념체계는 전역적 개념체계에 대응하는 개념으로, 특정 문서나 문서집합에 의존하여 개별 문서나 문서집합 단위로 생성한 개념 체계를 의미한다. 지역적 개념체계는 특정 문서의 의미 지도(semantic road map) 역할을 하며, 전체적인 문서의 내용 이해를 돕는다. 이러한 개별 문서 단위의 접근법은 기존의 정보검색 시스템에서 검색결과에 대한 이용자의 접근과 브라우징, 그리고 적합성 판정에 대한 새로운 대안을 제시할 수 있다. 특히, 단행본의 전통적인 접근수단이었던 표제, 저자, 목차, 권말색인 정보 이외에, 하이퍼텍스트 기반의 전자 도서에 대한 접근을 새롭게 할 수 있으며, 웹 사이트의 경우에는 사이트맵 자동 구축이 가능하다.

질의확장을 통한 정보 검색의 효율성 향상뿐만 아니라 개별 문서 및 컬렉션 단위의 개념체계 자동 구축을 위해서는 앞에서 언급한 용어의 의미모호성 문제나 관계 표현 문제 등이 해결되어야 한다. 이러한 문제 해결책으로 제안된 퍼지 클러스터링 기법의 타당성을 확인하기 위해서는 후속연구로 전자도서나 전자적인 형태의 학위논문을 대상으로 한 용어 자동분류 실험이 필요하다.

참 고 문 헌

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

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

정영미. 1993. **정보검색론**. 서울: 구미무역(주)출판부.

정혜천. 1999. **유사도 측정에 의한 계층적 퍼지 클러스터링**. 석사학위논문, 영남대학교 대학원, 전기공학과 제어

및 시스템 전공.

Bezdek, J. C. 1981. *Pattern Recognition with Fuzzy Objective Function Algorithms*. NY: Plenum Press.

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

Ekmekcioglu, F. C., Robertson, A. M., and Willet, P. 1992. Effectiveness of Query Expansion in Ranked-Output Document Retrieval Systems. *Journal of Information Science*, 18(2): 139-147.

Orwig, R. E., Chen, H., and Nunamaker, J. F. J. 1997. A Graphical. Self-Organizing Approach to Classifying Electronic Meeting Output. *Journal of American Society for Information Science*, 48(2): 157-170.

Roussinov, D., and Chen, H. 1999. Document Clustering for Electronic Meetings: An Experimental Comparison of Two Techniques. *Decision Support Systems*, 27: 67-79.

WEBSOM. http://websom.hut.fi/websom

Wu, Y. B. 2001. *Automatic Concept Organization: Organizing Concepts from Text through Probability of Co-occurrence Analysis*. Ph. D. diss., University of Albany, State University of New York.