대규모 데이터베이스에서의 지식정보 추출을 위한 클러스터링 기법 연구동향

Research Trends of Clustering Methods for Extracting Knowledge in Large Database

문병주(B.J. Moon) 정보유통연구팀 선임연구원
정현수(H.S. Jung) 정보유통연구팀 책임연구원, 팀장
이동일(D.I. Lee) 기술정보센터, 센터장

정보검색시스템에서는 방대한 양의 데이터에서 보다 효율적이고, 보다 정확한 데이터를 어떻게 추출할 것인가가 항상 가장 중요한 문제로 인식되어 왔다. 특히, 앞으로 데이터베이스는 지식정보를 담는 대규모 데이터베이스가 되므로 이러한 문제를 해결하기 위한 방법은 갈수록 복잡해 질 것이다. 현재 이의 해법으로 데이터마이닝에 대한 연구가 활발하게 진전되고 있으며, 특히 문서의 연관관계를 정의해 주는 클러스터링은 향후 지식발견의 가장 중요한 요소가 될 것으로 보인다. 따라서, 본 논문은 대규모 데이터베이스에서 지식정보 발견에 관한 기법에 대한 최근의 연구동향을 소개한다. 즉, 이용자 질의에 대한 검색 결과를 개선하기 위한 방편인 데이터마이닝 기법의 기본개념과 데이터마이닝 기법 중에서도 최근 가장 빠르게 실용화가 이루어지고 있는 클러스터링에 대한 최근의 동향을 살펴본다.

I. 서 론

최근 데이터베이스 분야에서 가장 주목받고 있는 기술로 데이터마이닝(Data Mining)을 꼽을 수 있다. 특히, 21세기 지식기반사회는 지식정보의 DB화 정도에 크게 의존하게 될 것이며, 이 경우 데이터베이스는 엄청난 양의 지식정보를 담게 되는 대규모 데이터베이스가 될 가능성이 크다. 따라서 이러한 대규모 데이터베이스에서 얼마만큼 유용하고 정확한 정보를 추출할 것인가가 가장 중요한 문제가 될 것이며, 최근 이의 해법으로 데이터마이닝이 등장하게 된 것이다.

데이터마이닝이란 방대한 양의 데이터에서 의미있는 패턴(pattern)이나 룰(rule)을 발견하여 이를 추출하고 분석하는 기법으로 클러스터링(clustering), 분류(classification) 등의 기법이 이용되고 있다.

이중 클러스터링에는 단어 클러스터링(term clustering)과 문서 클러스터링(document clustering)의 두 가지 방법이 있다. 단어 클러스터링이란 인접한 단어 정보로부터 각 단어에 대한 특성벡터를 추출하고, 모든 단어쌍(term pair)에 대하여 특성벡터를 이용한 유사도를 계산하여, 가장 유사도가 높은 단어들을 클러스터링하는 기법이다. 문서 클러스터링이란 문서를 구성하는 색인어들로 문서에 대한 특성벡터를 추출하고, 모든 문서쌍(document pair)에 대하여 특성벡터를 이용한 유사도를 측정하여, 가장 유사도가 높은 문서쌍을 클러스터링하는 기법이다.

특히, 클러스터링은 분류를 위한 사전 데이터를 이용하여야 하며, 각 분류체계에 해당하는 전문용어 사전 구축과 각 문서의 특성벡터(색인어)와 분류체계의 사전을 비교하여 가장 부합되는 분류에 할당하

는 것이 현재 가장 널리 이용되고 있는 클러스터링 기법이다.

본 고에서는 이러한 데이터마이닝의 기본개념과 클러스터링 관련 연구동향을 살펴보기로 한다.

II. 데이터마이닝

데이터베이스의 발전단계를 살펴보면, 지금까지 크게 4단계로 발전되어 왔다고 볼 수 있다[1]. <표 1>에서 보는 바와 같이 제1단계는 1960년대로 데이터베이스의 개념이 처음 도입되기 시작하여 데이터 수집(data collection)과 저장을 목적으로 발전되어 왔다. 그리고, 데이터가 어느 정도 저장되기 시작하자 1980년대 들어 데이터에 보다 효율적이고 손쉽게 접속할 수 있는 방법들이 모색되기 시작하였다. 이를 제2단계라고 볼 수 있으며, RDBMS나 SQL, ODBC 등의 개념들이 등장하기 시작하였다. 1990년대에는 데이터의 통합적이고도 대용량으로 제공할 수 있는 데이터 웨어하우스(data warehouse)가 기업을 중심으로 급속히 전개되어 왔다. 데이터 웨어하우스는 기존 데이터를 최종 이용자가 편리하게 접근할 수 있도록 동적으로 제공하게 된다.

하지만, 데이터의 양이 보다 광범위해지고, 특히 분산된 환경에서의 데이터 통합이 이루어짐에 따라 기존 데이터 처리방식의 한계가 나타나게 되었다. 즉, 이용자들이 기존의 광범위한 데이터에서 자신이 필요한 데이터만을 추출하는 데 있어 기존 기술로는 한계에 달하게 된 것이었다. 이에 이와 같이 방대한 양의 데이터에서 의미있는 패턴이나 룰을 발견하여 이를 추출하고 분석하여 체계화되고 유용한 데이터를 사용자에게 제공하는 데이터마이닝 개념이 등장하게 된 것이다.

1. 데이터마이닝의 개념

데이터마이닝, 혹은 데이터베이스에서의 지식발견(Knowledge Discovery in Databases: KDD)은 방대한 양의 데이터 집합에서 자동 혹은 반자동으로 의미 있는 패턴이나 룰을 발견하여 이를 추출하고

<표 1> 데이터베이스 발전단계

단 계	특 징	관련기술	주요업체
1단계 (1960년대)	• 데이터 수집 • 보관개념 • 정적 데이터 제공	컴퓨터, 테이프, 디스크	IBM, CDC
2단계 (1980년대)	• 데이터 접근 • 보관개념 • 단일정보에 대한 동적 데이터 제공	RDBMS, SQL, ODBC	Oracle, Sybase, Informix, IBM, Microsoft
3단계 (1990년대)	• 의사 결정 지원 • 데이터 웨어하우징 • 다중정보에 대한 동적 데이터 제공	OLAP, 다차원 DB, 데이터 웨어하우스	Pilot, Comshare, Arbor, Cognos, Microstrategy 등
4단계 (1990년대 후반)	• 데이터마이닝 • 예측/선행 정보 제공	병렬컴퓨터, massive database	Pilot, Lockheed, IBM, SGI 등

분석하는 기법이다[2]. 발견된 지식은 또한 데이터의 특성을 묘사하거나 패턴을 만들거나 혹은 데이터베이스에서의 객체의 클러스터링 등에 대한 규칙이 될 수 있다. 따라서 데이터마이닝에서는 이용자가 지식을 발견하고 자기 용도에 부합되도록 이를 적용할 수 있는 분석적 데이터 조작 툴이 필요하다.

데이터마이닝의 목적은 의사 결정 또는 마케팅에 적용할 수 있는 중요한 정보를 캐내는 것이다. 데이터마이닝의 기본 개념은 새로운 것이 아니라 인공지능 분야의 기계학습(machine learning) 이론에 그 뿌리를 두고 있다. 즉, 현실 세계에서 데이터베이스가 발달하여 수많은 데이터가 쌓여가고 있으므로 이로부터 감춰진 유용한 정보를 캐내고자 하는 욕구가 데이터베이스 종사자들에게 일어나게 되어 기계학습에서 사용된 기법을 데이터베이스에 응용하기에 이르렀다.

일반적으로 데이터마이닝에서 얻고자 하는 지식은 연관(association), 분류(classification), 클러스터링(clustering), 순서(sequence) 등에 관한 지식이다[3, 4, 10, 14]. 이들 요소에 대해 간략히 기술하면 다음과 같다.

가. 연관

연관은 데이터베이스 각 항목간에 존재하는 연관규칙(association rule)을 찾아내는 것이다. 연관규

칙이란 데이터베이스의 레코드 셋에 대하여 아이템의 집합 중에 존재하는 친화도나 패턴을 찾아내는 규칙이다. 예를 들어, 슈퍼마켓에서 운영하는 판매 데이터베이스의 경우 판매된 항목 중 서로 같이 팔리는 연관성이 높은 항목들을 알 수 있다면 상품진열, 상품주문 등 마케팅에 큰 도움이 될 것이다.

나. 분류

분류란 데이터베이스 내의 객체의 셋에 대하여 그 안에 내재하는 공통 특성을 추출하여 이 객체들을 서로 다른 클래스로 그룹핑하는 것이다. 즉, 데이터베이스에 있는 데이터들을 서로 중첩되지 않는 그룹으로 쪼개는 규칙을 찾는 것이다.

다. 클러스터링

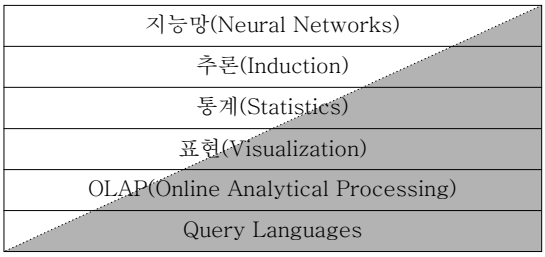
클러스터링이란 물리적 혹은 추상적 객체를 비슷한 객체군으로 그룹핑하는 것이다. 이 경우 유사성 때문에 함께 모여진 객체의 셋을 클러스터(cluster)라 한다. 클러스터링은 먼저 필수 객체들이 셋으로 모여지고 이로부터 일련의 규칙이 유도된다.

라. 순서

순서란 데이터를 순서화(ordering)하는 것이다. 이를 위해선 데이터의 특징 등을 식별해 내기 위해 일정기간 동안 레코드를 분석하여 순서 패턴을 찾아낸다. 이러한 지식 추출 작업에 자주 사용되는 기법에는 신경망(neural network), 결정트리(decision tree) 등이 많이 쓰이고 있다.

2. 데이터마이닝 기반기술

Gartner Group에 따르면, 향후 2~3년 내에 산업계에 폭 넓게 영향을 미칠 5가지 기술 중 데이터마이닝과 인공지능을 꼽고 있다. 또한, 향후 5년간 산업계에서 중점적으로 투자하게 될 10가지 기술 중 병렬 구조와 더불어 데이터마이닝을 들고 있다. 이와 같이 데이터마이닝은 21세기 초에 가장 중점적으로 연구되고 활용될 기술 중 하나로 주목 받고 있는 것이다[3]. 여기선 이러한 데이터마이닝을 가능하게 하는 기반기술에 대해 살펴보기로 한다.



(그림 1) 데이터마이닝의 기계/인적 관계도

데이터마이닝이란 대용량 데이터베이스에서 보다 정확한 지식을 추출하는 것으로 학습(learning)과 추론(induction)을 바탕으로 한다. 이러한 학습과 추론을 위한 절차로 Gartner Group에서는 지능망, 추론기법(induction), 통계기법(statistics), 가시화기법(visualization), OLAP(On-line Analytical Processing), 그리고 질의언어를 꼽고 있다.

(그림 1)의 대각선은 기계 대 인적 작업의 처리비율을 표시한 것이다. 즉, 대각선의 상위는 기계에 의한 처리부분이며, 하위는 인적 요소에 의한 처리부분을 나타낸다. 즉, 지능망은 기계적 처리가 거의 대부분 발생하는 작업이며, 질의언어는 인적 요소에 의한 처리가 대부분인 작업이 된다.

더불어 데이터마이닝에서 가장 공통적으로 사용되는 기술로는 인공지능망(artificial neural networks), 결정트리(decision trees), 유전자 알고리즘(genetic algorithms), 인접이론(nearest neighbor method), 규칙 추론(rule induction) 등을 들 수 있다.

가. 인공지능망

트레이닝을 통하여 학습하는 비선형 예측모델(non-linear predictive model)로 구조상 생물학적 신경계와 유사하다.

나. 결정트리

결정트리는 결정 집합(set of decision)을 묘사하는 트리 모양의 구조이다. 특히, 이러한 결정은 데이

터 집합의 분류에 필요한 규칙을 생성한다. 이러한 결정트리 기법으로는 Classification Regression Trees(CART) 그리고 CHAID(Chi Square Automatic Interaction Detection) 등이 있다.

다. 유전자 알고리즘

유전 조합(genetic combination), 변이(mutation) 등과 같은 프로세스를 사용하는 최적화 기법이다.

라. 인접이론

과거의 데이터 집합에서 가장 유사한 k 레코드의 클래스 조합을 근간으로 한 데이터 집합에서 각 레코드를 분류하는 기술이다.

마. 규칙 추론

통계적 중요도를 근간으로 한 데이터에서 유용한 if-then 규칙을 추출해 내는 기법이다.

Ⅲ. 클러스터링 연구동향

클러스터링은 정보검색에서 유사한 객체, 즉 문서(documents)나 단어(terms)를 그룹핑하는 데 이용되는 알고리즘이다. 클러스터링에 대한 연구는 정보검색 이론의 초창기인 1960년대부터 이루어져 왔다. 하드웨어적 및 기술적 한계로 인하여 1980년대까지 크게 활성화되지 못하였으나, 1990년대부터 브라우징 측면에서 연구가 활발하게 진전되었다.

클러스터링이 실질적으로 정형화된 기술로 적용되기 시작한 것은 Van Rijsbergen의 Single-link Hierarchies 이론에서 기인한다. 그후 Sparck Jones가 단어 클러스터링에 대한 이론을 정립하면서 클러스터링에 대한 이론적 바탕이 이루어지게 되었다.

1. 클러스터링 관련 알고리즘

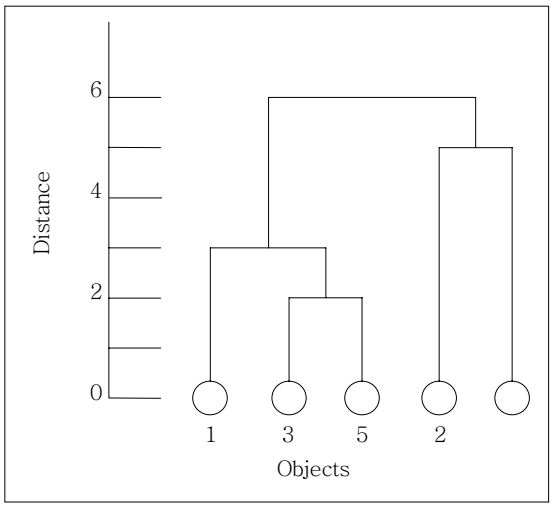
클러스터링 알고리즘은 크게 4가지 유형으로 나누어진다[5, 11]. 첫 번째는 특성(단어, 기능)과 클래스(클러스터)간 관계에 대한 알고리즘이며, Monothetic 알고리즘과 Polythetic 알고리즘으로 분류된다. 두 번째는 객체와 클래스간의 관계를 정의하는 알고리즘이며, Exclusive와 Overlapping 알고리즘으로 분류된다. 그리고, 클래스와 클래스간 관계를 정의하는 알고리즘이 있으며, Ordered(혹은 Hierarchic) 알고리즘과 Unordered(혹은 Simple Partition) 알고리즘으로 나누어진다. 마지막으로 클러스터를 최적화하기 위해, 미리 정의해 둔 기준이나 함수에 의해 하나의 클러스터를 생성하는 최적화 기법에서 이용되는 유전자 알고리즘이 있다. 유전자 알고리즘은 생체적 진화론을 적용하여 최적화 문제를 해결해보려는 기법이다.

하지만, 이러한 클러스터링 알고리즘은 일반적으로 문서간 비교나 클러스터와 문서 비교 등과 같은 유사한 방식을 이용하여 객체를 비교하는 Pairwise 알고리즘을 기반으로 한다. Pairwise 알고리즘의 기본개념은 비교 대상 객체의 비율이나 수가 증가하면, 유사도도 증가한다는 개념이다. Pairwise 알고리즘은 최소 0와 최대 1의 값으로 정형화되며, 비교 객체가 동일한 경우 최대값을 지니게 된다. 즉, S(X, X) = 1이다. 특히, Pairwise 알고리즘은 S(X, Y) = S(Y, X)라는 대칭성을 지닌다.

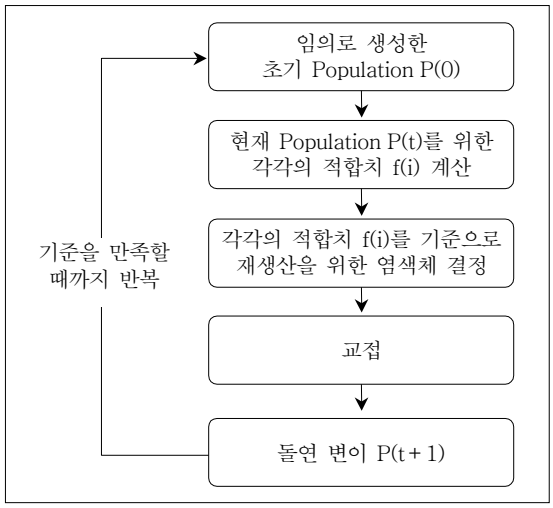
Pairwise 알고리즘을 바탕으로 매칭계수 $|X \cap Y|$ 나 고유의 가중치를 이용하는 알고리즘들이 발표되고 있다. 매칭계수에는 Dice의 계수 $2|X \cap Y| / (|X| + |Y|)$, Jaccard의 계수 $|X \cap Y| / |X \cup Y|$ 등이 있다. 하지만, 일반적으로 고유의 가중치를 이용한 알고리즘들이 보다 유용한 것으로 알려져 있다.

특히, 위의 알고리즘에서 가장 널리 이용되는 방법은 클래스간 관계를 정의하는 계층적 알고리즘(hierarchic algorithms)과 생체적 진화론을 적용한 유전자 알고리즘이다. 계층적 알고리즘은 일련의 연속적인 병합이나 분할에 의해 수행되어지는 기법을 말한다. 그 수행 결과는 트리 구조 또는 계층적 구조로서 (그림 2)와 같이 Dendogram으로 알려진 도식으로 표현된다.

응집계층기법은 분리된 클러스터를 각각 관찰하



(그림 2) 계층적 클러스터링을 표현한 Dendogram



(그림 3) 유전자 알고리즘의 개요

며 그들의 유사도에 따라 병합하는 일련의 과정을 하나의 클러스터만 남을 때까지 계속한다.

분할계층기법은 반대로 모든 객체가 포함된 하나의 클러스터로부터 비유사성을 기준으로 각 객체가 자신의 그룹을 가질 때까지 서브그룹으로 분할한다.

계층적 기법의 성능은 데이터 유형(data type)에 따라 변화되므로 모든 환경에 최적의 기법은 존재하지 않는다. 계층적 클러스터링의 몇 가지 단점을 살펴보면, 다음과 같다.

1) 유사성 행렬을 저장하기 위해 제한된 작은 데이터 집합 사용
2) 앞 단계에서 잘못 그룹핑된 객체에 대한 재설정 과정이 없다.
3) 알고리즘에 내재된 구조적 형태에 의해 자료의 등급이 결과에 반영된다.

계층적 클러스터링 기법과는 달리 최적화하기 위해 미리 정의해 둔 기준이나 함수에 의해 하나의 클러스터를 생성하는 기법으로 최적화 기법이 있다. 최적화 기법에는 생체적 진화론을 적용하여 최적화 문제를 해결해보려는 유전자 알고리즘이 널리 이용되고 있다.

간단히 말해 유전자 알고리즘이란 객체 집합에 선택, 교접, 돌연변이와 같은 일련의 유전적 연산을 계속적으로 적용하는 것을 말한다. 염색체(chromosomes)라 불리는 요소는 가능한 문제 해결책으로 표현된다. 초기의 염색체는 해결책 집합으로부터 임의로 선택된다.

유전적 연산은 객체의 유전적 정보와 결합하여 새로운 유전체를 생성한다. 이러한 과정을 재생산(reproduction)이라 한다. 각 염색체는 문제 해결의 정도를 수치화한 적합치(fitness value)를 가지며 보다 최적의 해를 가지는 염색체는 보다 높은 적합치를 가진다. 즉, 적자생존의 법칙을 적용하여 높은 적합치를 가지는 염색체를 선택하는 기법이다.

유전자 알고리즘은 (그림 3)과 같이 풀고자 하는 문제의 변수 값을 이진 스트링으로 표현한다. 이 코딩 방법은 문제의 변수가 이진값이거나 다른 이산치를 갖는 경우에 특히 적합한 표현법이다.

2. 클러스터링 방법

클러스터링 방법으로는 문서를 구성하는 색인어들을 이용하는 문서 클러스터링과 인접단어의 특성을 이용하는 단어 클러스터링이 있다[5, 7, 12].

두 방식은 우선 클러스터링 단위(단어, 문서)가 다르므로 클러스터링에 이용되는 특성벡터 계산과정이 다르다는 차이점은 있으나 같은 클러스터링 문

제이므로 일단 특성벡터가 구해지면 유사한 클러스터링 알고리즘을 적용하게 된다.

가. 문서 클러스터링

문서 클러스터링은 문서를 구성하는 색인어들로 문서에 대한 특성벡터를 추출하여 모든 문서쌍에 대하여 특성벡터를 이용하여 유사도를 측정함으로써 가장 유사도가 높은 문서쌍을 클러스터링하는 기법으로 클러스터링된 결과를 문서사이의 유사도 계산에 반영하게 된다. 문서 클러스터링 기법으로는 Graph Theoretic Methods, Fast Partition Methods, Nearest Neighbor Clusters 등이 이용되고 있다.

Graph Theoretic Methods는 어떤 한계치(threshold) 이상의 유사도를 지니는 객체를 그래프 형식으로 표현하는 기법이다. 여기에는 서브그래프간에 하나의 링크만을 지니는 Single linkage cluster(혹은 Connected component)와 서브그래프간에 복합적인 링크를 지니는 Complete linkage cluster(혹은 Maximal complete subgraph), 그리고 최근에는 클러스터간 평균값을 계산하는 Average linkage cluster와 객체가 거리의 제곱의 합으로 표현되는 Ward's minimum variance 기법도 이용되고 있다. Graph Theoretic Methods의 각 기법별 특성은 다음과 같다.

- Single linkage

Nearest Neighbor Clustering이라고도 하며 두 클러스터 간의 거리를 각각의 클러스터 내의 객체 가운데 가장 가까운 객체간의 거리로 설정하는 기법으로 클러스터간의 관계가 명확하지 않을 때 하나의 긴 체인 형태로 클러스터링 되어지는 문제가 발생된다.

- Complete linkage

Furtherest neighbor clustering이라고도 하며 두 클러스터 간의 거리를 각각의 클러스터 내의 객체 가운데 가장 멀리 떨어진 객체간의 거리로 설정하는 기법으로 Single linkage 클러스터링 기법의 chaing 현상을 제거할 수 있다.

- Average linkage

Group average clustering이라고도 하며 두 클러스터 간의 거리를 각각의 클러스터 내의 모든 객체로부터 다른 클러스터 내의 모든 객체로의 거리의 평균값으로 설정하는 기법으로 작은 변화에 적응하여 클러스터링 할 수 있는 기법으로 사용된다.

- Ward's minimum variance 기법

위에서 열거한 기법과는 달리 객체를 클러스터 중심과 각 객체간의 거리의 제곱의 합을 이용한 통계적 최적화 기법을 말한다. 클러스터 병합 알고리즘의 각 스텝에 의해 이 통계치의 증가율은 감소할 것이다. 이 기법은 작은 수의 객체를 클러스터링하고, 거의 같은 크기의 클러스터를 생성한다.

Fast Partition Methods는 문서간 클러스터링의 속도를 높이기 위한 기법으로 Single Pass methods, K-means methods가 있다. Single Pass Methods는 동일한 클러스터(C1) 내의 특정 문서(D1)를 대표문서화 함으로써 각 클러스터간에 대표문서(Di)를 비교하여 유사도 Si를 계산하는 방식이다. 이 때 Si가 어떤 한계치 St보다 큰 경우, 해당 문서를 대응하는 클러스터에 추가하고, 클러스터의 대표문서를 다시 계산하게 된다. 이러한 작업은 문서가 모두 클러스터링 될 때까지 계속 이루어지게 된다. K-means 혹은 Reallocation methods는 특정 클러스터의 대표문서를 추출하고, 클러스터링 하고자 하는 문서를 가장 유사한 대표문서가 있는 클러스터에 포함시키게 된다.

Nearest Neighbor Clusters는 가장 근접한 문서들을 클러스터링하는 기법이다. 이때 K라는 근접도를 주게 되며, 클러스터간에 오버래핑(중복부분)이 발생하는 특징이 있다. Nearest Neighbor Clusters로는 Sparck Jones의 Star cluster가 계층적 클러스터(hierarchic cluster)를 생성하는 데 이용되고 있다.

나. 단어 클러스터링

단어 클러스터링은 인접한 단어 정보로부터 각 단어에 대한 특성벡터를 추출하여 모든 단어쌍에 대하여 특성벡터를 이용하여 유사도를 계산함으로써

가장 유사도가 높은 단어들을 클러스터링 하는 기법이다. 특히, 클러스터링된 결과를 단어사이의 유사도 계산에 반영하게 된다.

초창기의 단어 클러스터링은 시소러스 사전을 이용하여 단어들을 클러스터링하는 기법이 이용되었다. 특히, Van Rijsbergen의 확률 검색모델(Probabilistic retrieval model) 등이 단어 클러스터링을 기반으로 하고 있다. 더불어 LSI 등은 단어 클러스터를 생성하는 문서 클러스터를 이용하고 있다. 최근의 정보검색에서 이용되고 있는 질의 확장 기술은 문맥을 기반으로 한 단어 클러스터링을 이용하고 있으며, 최근의 클러스터링에 관한 연구는 단어 클러스터링을 확장하는 방향으로 이루어지고 있다.

Ⅳ. 결론 및 향후 연구방향

지금까지 정보검색시스템에서는 방대한 양의 데이터에서 보다 효율적이고, 보다 정확한 데이터를 어떻게 추출할 것인가가 항상 가장 중요한 문제로 인식되어 왔다. 특히, 데이터량이 갈수록 방대해지고 있어 이러한 문제를 해결하기 위한 방법은 갈수록 복잡해질 것이다.

현재 이의 해법으로 데이터마이닝에 대한 연구가 활발하게 진전되고 있다. 특히, 문서의 연관관계를 정의해 주는 클러스터링은 향후 지식발견의 가장 중요한 요소가 될 것으로 보인다.

따라서, 본 고에서는 데이터마이닝과 클러스터링에 대한 개념과 현황을 살펴보았다.

참 고 문 헌

[1] Pilot, “An Introduction to Data Mining?,” http://www.pilotsw.com /dmpaper/dmindex.htm, 1999. 1.

[2] Heikki Mannila, “Methods and Problems in Data Mining,” http://www.csd.uch.gr/~chrysos/uni/dmining/index.html, 1997. 7.

[3] Heikki Mannila, “Database Mining: A Performance Perspective?,” 1996. 10.

[4] 나민영, “대규모 지식데이터베이스에서 유용한 지식 추출하는 기법,” http://www.dpc.or.kr/dbworld/document/ 9709/spec.html, 1997. 9.

[5] Bruce Croft, “Document and Term Clustering,” http://ciir.cs.umass.edu/cmpsci646/ir9/index.html, 1998. 4.

[6] Moses Charikar *et al.*, “Incremental Clustering and Dynamic Information Retrieval,” *Proc. of the 29th Annual ACM Symp. on Theory of Computing*, 1997. 5. pp. 626 – 635.

[7] Heikki Mannila, “Data Mining and Hierarchical Models,” http:// www.cs.helsinki.fi/~mannila/, 1997. 11.

[8] Helena Ahonen *et al.*, “Applying Data Mining Techniques in Text Analysis,” 1998. 6.

[9] John Davies *et al.*, “Using Clustering in a WWW Information Agent,” http://www.labs.bt.com/jasper/html/jasclus.html.

[10] Heikki Mannila, “Data Mining: Machine Learning, Statistics, and Databases,” http://www.cs.helsinki.fi/~mannila/.

[11] Rakesh Agrawal *et al.*, “Database Mining: A Performance Perspective,” *IEEE Trans. on Knowledge and Data Engineering*, Vol. 5, No. 6, 1993. 12. pp. 914 – 925.

[12] Oskari Heinonen *et al.*, “Attribute-Oriented Induction and Conceptual Clustering,” 1996. 9.

[13] Doug Fisher, “Iterative Optimization and Simplification of Hierarchical Clustering,” “http://www.cs.washington.edu/research/jair/volume4/fisher96a -html/html -final.html,” 1996. 3.

[14] Hannu Toivonen, “Discovery of Frequent Patterns in Large Data Collections,” 1996. 5.