지능정보연구 제17권 제1호 2011년 3월(pp.111~125)

중복을 허용한 계층적 클러스터링에 의한 복합 개념 탐지 방법

홍수정
LG유플러스
(crystaland@lguplus.co.kr)

최중민
한양대학교 컴퓨터공학과
(jmchoi@hanyang.ac.kr)

클러스터링(Clustering)은 유사한 문서나 데이터를 묶어 군집화해주는 프로세스이다. 클러스터링은 문서들을 대표하는 개념별로 그룹화함으로써 사용자가 자신이 원하는 주제의 문서를 찾기 위해 모든 문서를 검사할 필요가 없도록 도와준다. 이를 위해 유사한 문서를 찾아 그룹화하고, 이 그룹의 대표되는 개념을 도출하여 표현해주는 기법이 요구된다. 이 상황에서 문제점으로 대두되는 것이 복합 개념(Complex Concept)의 탐지이다. 복합 개념은 서로 다른 개념의 여러 클러스터에 속하는 중복 개념이다. 기존의 클러스터링 방법으로는 문서를 클러스터링할 때 동일한 레벨에 있는 서로 다른 개념의 클러스터에 속하는 중복된 복합 개념의 클러스터를 찾아서 표현할 수가 없었고, 또한 복합 개념과 각 단순 개념(Simple Concept) 사이의 의미적 계층 관계를 제대로 검증하기가 어려웠다.

본 논문에서는 기존 클러스터링 방법의 문제점을 해결하여 복합 개념을 쉽게 찾아 표현하는 방법을 제안한다. 기존의 계층적 클러스터링 알고리즘을 변형하여 동일 레벨에서 중복을 허용하는 계층적 클러스터링(Hierarchical Overlapping Clustering, HOC) 알고리즘을 개발하였다. HOC 알고리즘은 문서를 클러스터링하여 그 결과를 트리가 아닌 개념 중복이 가능한 Lattice 계층 구조로 표현함으로써 이를 통해 여러 개념이 중복된 복합 개념을 탐지할 수 있었다. HOC 알고리즘을 이용해 생성된 각 클러스터의 개념이 제대로 된 의미적인 계층 관계로 표현되었는지는 특징 선택(Feature Selection) 방법을 적용하여 검증하였다.

논문접수일 : 2010년 12월 30일 논문수정일 : 2011년 01월 25일 게재확정일 : 2011년 02월 18일
투고유형 : 국문일반 교신저자 : 최중민

1. 서 론

클러스터링(Clustering)은 유사한 문서나 데이터를 묶어 군집화해주는 프로세스이다. 클러스터링은 문서들을 대표하는 개념별로 그룹화함으로써 사용자가 자신이 원하는 주제의 문서를 찾기 위해 모든 문서를 검사할 필요가 없도록 도와준다. 클러스터링의 원칙은 동일한 클러스터(Cluster)에 속하는 문서들 간에는 상당한 유사성이 존재하지만, 서로 다른 클러스터에 속하는 문서들 간에는 이질성을 갖도록 클러스터를 구성하는 것이다. 클러스터링은 데이터 집합으로부터 지식을 발견하기에 유용한 기법이다(Lavine, 2000). 이를 위해서는 문서를 클러스터링하여 각 클러스터의 주요 의미를 나타내는 개념을 도출하여 표현해주는 기법이 요구된다.

이 상황에서 문제점으로 대두되는 것이 복합 개념(Complex Concept)의 탐지이다. 복합 개념은 서로 다른 개념의 여러 클러스터에 속하는 중복 개념이다. 기존의 클러스터링 방법으로는 문서를 클러스터링할 때 동일한 레벨에 속하는 서로 다른

개념의 클러스터에 속하는 중복된 복합 개념의 클러스터를 찾아서 표현할 수가 없었고, 또한 복합 개념과 각 단순 개념(Simple Concept) 사이의 의미적 계층 관계를 제대로 검증하기가 어려웠다.

본 논문에서는 기존 클러스터링 방법의 문제점을 해결하여 복합 개념을 쉽게 찾아 표현하는 방법을 제안한다. 기존의 계층적 클러스터링 알고리즘을 변형하여 동일 레벨에서 중복을 허용하는 계층적 클러스터링(Hierarchical Overlapping Clustering, HOC) 알고리즘을 개발하였다. HOC 알고리즘은 문서를 클러스터링하여 그 결과를 트리가 아닌 개념 중복이 가능한 Lattice 계층 구조로 표현하고 이를 통해 여러 개념이 중복된 복합 개념을 탐지할 수 있었다. HOC 알고리즘을 이용해 생성된 각 클러스터의 개념이 제대로 된 의미적인 계층 관계로 표현되었는지는 특징 선택(Feature Selection) 방법을 적용하여 검증하였다.

본 논문의 구성은 다음과 같다. 제 2장에서는 기존의 클러스터링 방법과 핵심 기술에 대해 살펴보고, 제 3장에서는 본 논문에서 제안하는 복합 개념 탐지를 위한 HOC 클러스터링 알고리즘을 설명하고 기존의 계층적 클러스터링 알고리즘과 비교한다. 제 4장에서는 제안한 HOC 알고리즘을 이용한 전체 시스템 구조를 예제와 함께 설명하고, 제 5장에서는 실제 데이터 예제에 적용한 실험을 통해 복합 개념 탐지와 개념 관계 검증을 평가한다. 마지막으로 제 6장에서는 결론을 맺고 향후 과제를 제시한다.

2. 관련 연구

클러스터링 방법은 비계층적 방법(Non-hierarchical Clustering)과 계층적 방법(Hierarchical Clustering)으로 구분한다. 이 분류 방법은 클러스터링을 수행하여 얻어진 결과 클러스터의 구조 형태를 기준으로 한다. 즉, 클러스터링 결과가 클러스터 객체간의 계층을 나타내는 트리 형태의 모습으로 나타나면 계층적 방법, 그렇지 않으면 비계층적 방법으로 분류한다.

비계층적 클러스터링 방법에는 K-means 알고리즘과 같은 분할 클러스터링(Partitional Clustering)(Jain et al., 1999; Likas et al., 2003; Liu, 2007), Fuzzy K-means(Gath and Geva, 1989)나 Overlapping K-means(Cleuziou, 2008)와 같은 중첩 클러스터링(Overlapping Clustering), 그리고 확률 기반 클러스터링(Probabilistic Clustering) 등이 있다.

계층적 클러스터링은 각각 하나의 문서를 클러스터로 보고 가장 유사도가 높은 두 개의 클러스터를 병합하거나 분리하면서 계층적으로 군집을 형성해 가는 방법이다. 클러스터링의 결과를 시각적으로 보여주기 위해 덴드로그램(dendrogram)이라고 불리는 트리 형태를 이용한다. 계층적 클러스터링 방법은 상향식(bottom-up)과 하향식(top-down)의 두 가지 기본 접근 방식이 있다. 병합 계층적 클러스터링(Agglomerative Hierarchical Clustering)은 상향식 방식으로, 두 개의 작은 클러스터를 병합하여 최종으로 하나의 클러스터가 될 때까지 가장 유사한 클러스터를 찾아 병합해 가는 방식이다. 분리 계층적 클러스터링(Divisive Hierarchical Clustering)은 하향식 방식으로, 하나의 큰 클러스터를 두 개의 작은 클러스터로 레벨을 형성해가면서 가장 유사한 그룹으로 분리하는 방식이다(Liu, 2007). 보편적으로 병합 계층적 클러스터링을 더 많이 사용하며, 본 연구에서도 이 클러스터링 방법을 변형하여 새로운 알고리즘을 개발하였다.

클러스터링을 이용하여 생성된 클러스터를 분석하여 숨겨진 개념을 찾아내는 방법은 여러 분야에서 응용될 수 있기 때문에 지속적인 연구가 이

루어지고 있다. 그 중에서도 계층적 클러스터링 방법을 이용하여 개념을 탐지하는 연구는 각 클러스터가 표현하는 개념의 연관 관계를 표현해줄 수 있다는 이득이 있다. 다음은 계층적 클러스터링을 이용하여 개념을 탐지하고 표현하는 기법을 다루고 있는 연구들이다.

(Jonyer, 2001)은 그래프를 기반으로 하는 개념적 계층 클러스터링 방법을 다루고 있다. 그래프 기반의 구조 표현은 개념을 쉽게 발견할 수 있게 한다. 이 방법은 원래의 데이터를 압축하고 데이터 안의 개념을 구조적으로 표현하는 서브구조를 탐지하여 발견하게 해 준다. 이 방법은 클러스터가 가지는 개념 간의 많은 연관 관계를 발견하여 표현할 수 있다.

기존의 개념 계층 형성에 관한 연구에서는 대부분 계층적 클러스터링을 이용해서 비슷한 용어(term)를 가진 그룹으로 나누었지만 결과 계층은 그다지 만족스럽지 못하였다. (Yeh and Sie, 2006)은 이를 개선하기 위해 specific knowledge 네트워크의 결과와 자동적으로 형성된 개념 계층을 합친다. 이 방법에서는 클러스터링 단계에서 voting과정을 기본으로 하여 기존의 병합 계층적 클러스터링 방법에 voting 정보를 혼합하여 클러스터링을 실행한다.

(Chuang and Chien, 2004)는 텍스트 세그멘테이션(text segmentation)의 계층적 토픽 구조를 만드는데 클러스터링을 적용한다. 문서에서 텍스트를 세그멘테이션하고 나눠진 문장에서 중심이 되는 단어를 추출한다. 추출된 단어를 기반으로 세그멘테이션으로 나눠진 각 문장에 병합 계층적 클러스터링과 분할 계층적 클러스터링을 적용하여 개념 계층을 생성한다.

개념 계층의 동일 레벨에 존재하는 서로 다른 개념에 속하는 복합 개념을 탐지하여 표현하기 위해서는 클러스터 사이의 연관된 상하 관계 뿐만 아니라 클러스터의 중복(overlapping)을 표현할 수 있어야 한다. 하지만 위에서 기술한 기존의 클러스터링 방법은 복합 개념을 탐지하는데 적절하지 않다. 본 논문은 이를 해결하는데 초점을 맞춘다.

3. 중복 허용 계층적 클러스터링(HOC) 알고리즘

본 논문에서 제안하는 중복 허용 계층적 클러스터링(Hierarchical Overlapping Clustering, HOC) 알고리즘은 기존의 병합 계층적 클러스터링 알고리즘을 변형, 발전시킨 것이다. 본 장에서는 기존 알고리즘과 HOC 알고리즘을 pseudo-code 형태로 제시하고, 같은 예제에 대해서 두 알고리즘을 수행한 결과를 비교함으로써 HOC 알고리즘의 효과를 보여주고자 한다.

<알고리즘 1>은 기존의 계층적 클러스터링 알고리즘에서 가장 널리 쓰이는 병합 계층적 클러스터링 알고리즘을 나타낸다.

Algorithm : Agglomerative_Hierarchical_Clustering(*D*)

1. 데이터 집합 *D*의 각 데이터 포인트를 하나의 클러스터로 간주한다.
2. *D*에 속한 모든 데이터 포인트들을 각 한 쌍씩 거리를 계산한다.
3. **repeat**
4. 서로 간의 거리가 가장 가까운 두 클러스터를 찾는다.
5. 두 클러스터를 합쳐서 새로운 클러스터 *C*를 만든다.
6. *C*와 다른 모든 클러스터들 간의 거리를 계산한다.
7. **until** 하나의 클러스터가 남을 때까지.

<알고리즘 1> 병합 계층적 클러스터링 알고리즘

다음과 같이 데이터 포인트 a, b, c, d, e, f로 구성된 예제에 병합 계층적 클러스터링을 적용하면 어떤 결과가 나오는지 살펴보자. 데이터간의 유사도에 따라 <그림 1>과 같이 초기 클러스터링 된 예제에 이 알고리즘을 적용한 결과는 <그림 2>와 같이 표현된다. 이를 덴드로그램으로 표현하면 <그림 3>과 같이 트리의 형태로 나타난다.

Algorithm : Hierarchical_Overlapping_Clustering(D)

1. 데이터 집합 D의 각 데이터 포인트를 하나의 클러스터로 간주한다.
2. **repeat**
3. D에 속한 데이터 포인트(클러스터 C) 간의 거리를 계산한다.
4. **for** $i \leftarrow 1$ **to** 클러스터 수 **do**
5. i번째 클러스터에서 서로 간의 거리가 가장 가까운 클러스터를 찾는다.
6. i번째 클러스터와 찾은 클러스터를 합쳐서 새로운 클러스터 C를 만든다.
7. **end for**
8. **until** 하나의 클러스터가 남을 때까지.

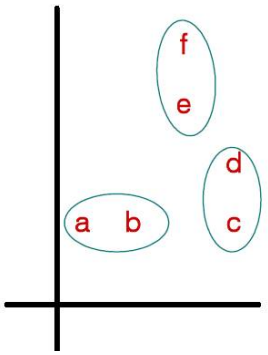
<알고리즘 2> 중복 허용 계층 클러스터링(HOC) 알고리즘

<알고리즘 2>는 본 논문에서 제안하는 HOC 알고리즘으로 기존의 병합 계층적 클러스터링 알고리즘을 변형하여 클러스터링된 결과가 트리의 형태가 아닌 Lattice 형태로 표현될 수 있게 하였다.

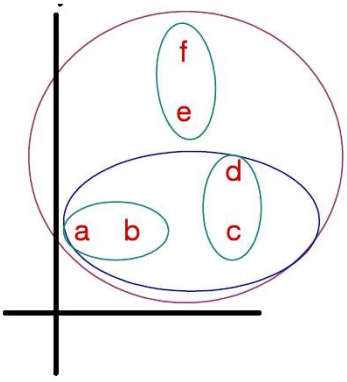
데이터 포인트 a, b, c, d, e, f로 구성된 이전과 동일한 예제(<그림 1>)에 제안하는 HOC 알고리즘을 적용한 결과는 <그림 4>로 표현되고 이를 덴드로그램으로 표현하면 <그림 5>와 같다.

<그림 3>과 <그림 5>를 비교해보면 기존 병합 계층적 클러스터링 알고리즘을 수행했을 때와 제안하는 HOC 알고리즘을 수행했을 때 각각의 클러스터링 결과가 다르다는 것을 알 수 있다. <그림 5>에서 보듯이 HOC 알고리즘은 개념의 의미적인 상하관계를 표현하면서도 기존 계층적 클러스터링에서 하지 못했던 개념 계층구조의 동일 레벨에 존재하는 서로 다른 개념에 중복으로 속하는 복합 개념을 탐지하여 표현할 수 있다. 또한, <그림 6>과 같이 각 클러스터를 대표하는 개념을 찾아 이를 표현할 수 있다면 차이가 명확히 보인다.

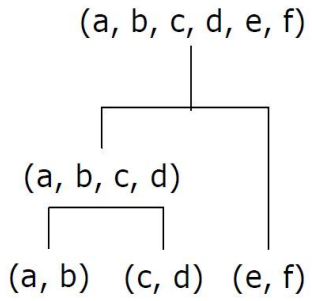
<그림 6>에서 (a)는 병합 계층적 알고리즘의 클러스터링 결과 덴드로그램이고 (b)는 제안하는 HOC 알고리즘의 클러스터링 결과 덴드로그램이다. <그림 6>의 (b)에서 “사람” 개념의 클러스터



<그림 1> 예제 데이터포인트



<그림 2> 병합 계층적 클러스터링 알고리즘 적용 결과



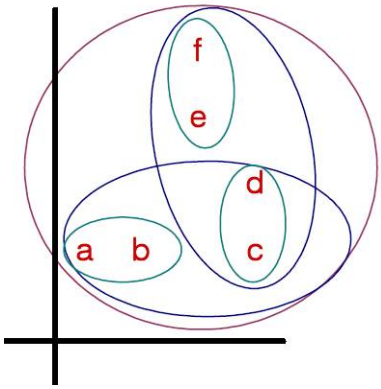
<그림 3> 병합 계층적 클러스터링 결과 덴드로그램

(c, d)가 상위의 같은 레벨에 있는 서로 다른 클러스터인 “초식” (a, b, c, d)와 “육식” (c, d, e, f)에 중복으로 속하는 것을 볼 수 있다. 이러한 중복된 연결(edge)이 존재하는 것이 복합 개념이다.

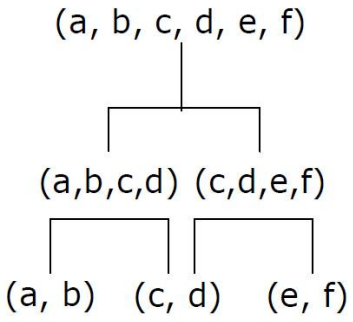
기존의 계층적 알고리즘으로는 “사람”과 “사자”의 데이터로 “육식”이라는 개념을 표현하지 못했고 또한 “사람”이 “초식”과 “육식”에 중복으로 속하는 복합 개념이라는 것을 알 수 없었다. 그러나 제안하는 HOC 알고리즘은 “육식” 개념을 찾아서 표현해줄 뿐만 아니라 “사람”이 복합 개념이라는 것도 표현한다.

4. HOC 알고리즘을 이용한 시스템 구조

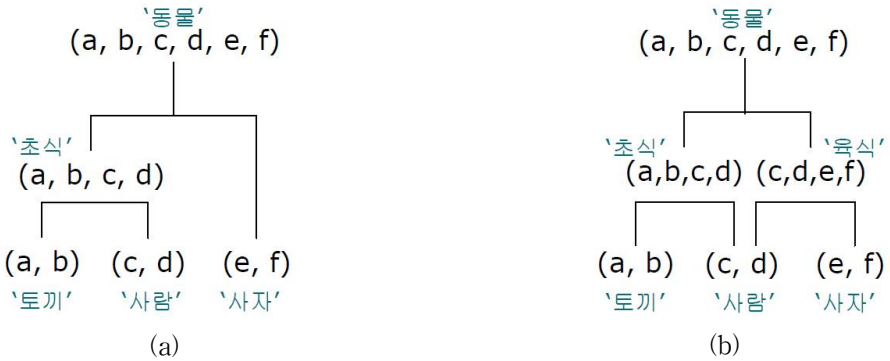
이 장에서는 본 논문에서 제안하는 HOC 알고리즘을 사용하여 복합 개념을 탐지하는 전반적인 시스템 구조를 설명한다. 이 시스템 구조는 1) 문서의 전처리 과정, 2) HOC 알고리즘을 이용한 클러스터링 과정, 3) 특징 선택을 이용하여 Lattice 상에 표현된 개념들의 의미적 계층 관계를 검증하는 과정 등의 세 부분으로 구성되어 있다. <그림 7>은 HOC를 사용하는 전반적인 시스템 구조이다.



<그림 4> HOC 알고리즘 적용 결과



<그림 5> HOC 알고리즘 클러스터링 결과 덴드로그램



<그림 6> 기존 계층적 알고리즘과 HOC 알고리즘의 클러스터링 결과 덴드로그램 비교

4.1 문서의 전처리 과정

클러스터링을 위해서는 알고리즘을 적용하기 전에 문서들을 객관적인 수치인 좌표 값으로 변형하여 좌표평면 또는 공간에 표현해주어야 한다. 이러한 과정을 문서 전처리 과정이라 한다.

우선 문서에서 단어를 추출하고 정제된 단어를 얻기 위해 불필요한 단어를 제거한다. 이를 위해 불용어 제거(stopword removal)와 어근처리(stemming)를 한다. 이러한 과정을 거치면 각 문서를 대표하는 색인어(index term)가 추출된다. 이 색인어들을 식 (1)을 적용한 TF-IDF 값에 따라 최종적인 좌표 값을 부여하고(Baeza-Yates and Ribeiro-Neto, 1999), 그 값에 따라 문서들을 좌표평면 또는 공간에 표현한다.

$$TF-IDF_{ik} = tf_{ik} \times \log \frac{N}{df_k} \quad (1)$$

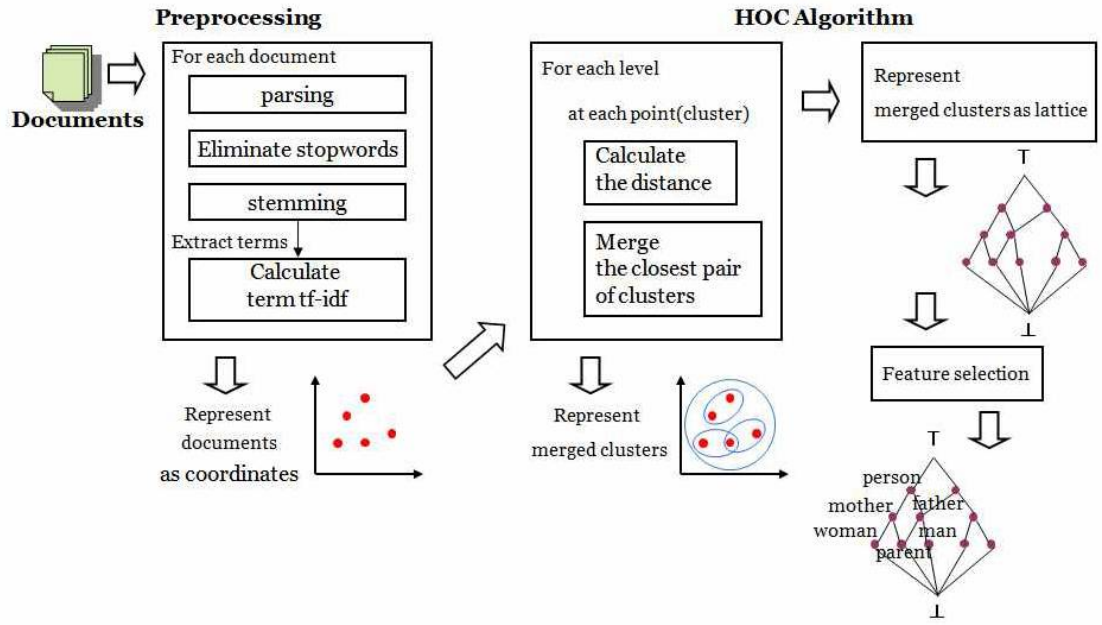
df_k : k번째 단어가 출현하는 클래스의 개수
tf_{ik} : i번째 클래스의 k번째 단어의 출현빈도수
N : 클래스의 전체 개수

<그림 8>은 샘플 문서의 좌표값 테이블과 이를 좌표 평면에 표현한 예제를 보여준다.

4.2 HOC를 적용한 클러스터링 과정

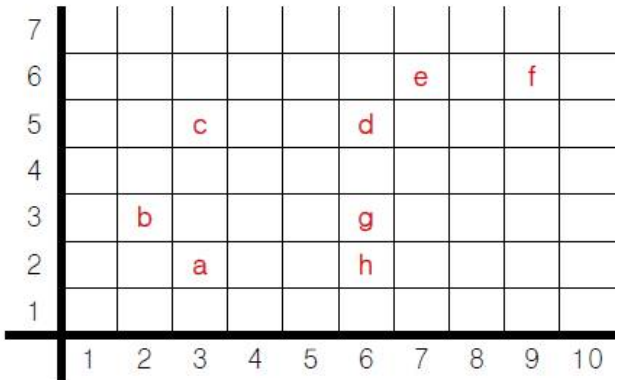
두 문서의 유사도는 문서간의 거리 값이 작을수록 높다는 전제 하에, 본 논문에서는 유클리디안 거리(Euclidean distance) 측정 방식을 사용하여 문서간의 거리를 계산하여 유사도를 판별하였다.

문서 전처리 과정 예제에서 다룬 <그림 8>의 문서간의 거리 값을 계산하면 <그림 9>(a)와 같이 나타낼 수 있고, 이 값을 이용하여 각 문서마다 거리가 가장 가까운 문서를 구하면 <그림 9>(b)와



<그림 7> HOC 알고리즘을 사용하는 전체 시스템 구조

	x	y
a	3	2
b	2	3
c	3	5
d	6	5
e	7	6
f	9	6
g	6	3
h	6	2



<그림 8> 문서들의 좌표 값 테이블과 좌표평면에 표현된 예제 문서들

	a	b	c	d	e	f	g	h
a		1.41	3	4.24	5.66	7.21	3.16	2
b			2.24	4.47	5.83	7.62	4	4.12
c				3	4.12	6.08	3.61	4.24
d					1.41	3.16	2	3
e						2	3.16	4.12
f							4.24	5
g								1
h								

(a) 문서간의 거리 값

	Closest doc.
a	b
b	a
c	b
d	e
e	f
f	e
g	h
h	g

(b) 문서별 가장 가까운 거리의 문서 테이블

<그림 9> 문서별 가장 가까운 문서 구하기

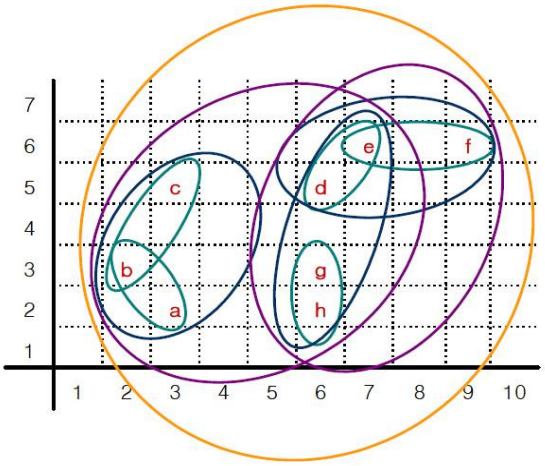
같이 정리될 수 있다.

이제 이 테이블 데이터를 이용하여 문서들을 서로 묶어 합친다. 이렇게 하면 좌표 평면 상에서는 <그림 10>과 같이 중첩된 그룹으로 표현되고, 이를 트리 형태로 표현하면 <그림 11>과 같다.

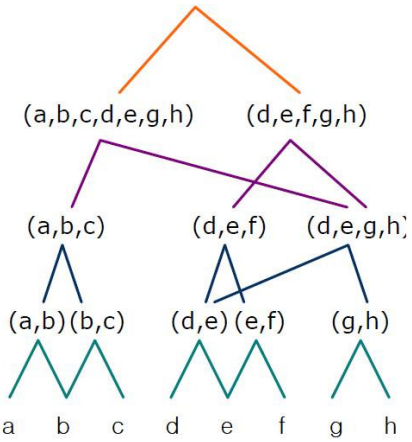
<그림 11>에서 클러스터 (d, e)는 상위의 동일 레벨에 있는 클러스터 (d, e, f)와 클러스터 (d, e, g, h)에 중복으로 속해있고, 클러스터 (d, e, g, h)는 클러스터 (a, b, c, d, e, g, h)와 클러스터 (d, e, f, g, h)에 중복으로 속해있다. 이를 통해 클러스터 (d, e)와 클러스터 (d, e, g, h)는 복합 개념이라는 것을 알 수 있다.

4.3 특징 선택을 이용한 개념 관계 검증 과정

HOC 알고리즘에 의해 Lattice 형태로 표현된 클러스터들의 개념 관계(concept relationship)를 알아보기 위해 본 논문에서는 특징 선택을 이용하여 클러스터를 대표할 수 있는 단어들을 특징으로 추출한다. 그리고 추출된 단어들을 바탕으로 클러



<그림 10> HOC 알고리즘에 의한 클러스터링 결과



<그림 11> HOC 알고리즘에 의한 클러스터링 결과 트리

스터의 대표 개념을 추정하여 의미적 계층 관계를 검증한다. 특징 선택을 하기 위해 용어의 기여도를 측정하는 수식을 이용하여 각 클러스터에 대한 용어의 정보량을 계산한 후 전체 클러스터 문서에서의 용어 정보량을 계산한다.

학습에 이용될 중요한 속성들을 추출하는 과정에서 신뢰성을 향상시키기 위해 해당 문서의 공통적인 특징을 가려내어 이를 기준으로 각 속성마다 가중치를 차별적으로 두는 중요 속성을 추출하는 방법을 특징 선택(Feature selection)이라 한다. 즉, 특징 선택은 학습 자료의 중요 속성들을 구분된 클래스별로 중요도를 정의하는 특징 추출 가중치 설정 기법이다. 여기에서의 특징(feature)은 클래스를 대표하는 단어 집합으로서 이러한 단어들에 높은 가중치가 설정된다. 특징을 선택하기 위해서는 각 단어가 자신이 속해 있는 클래스에 얼마만큼 기여하고 있는가를 알아야 하며 그러한 기여도를 측정하는 데에는 여러 가지 방법이 있는데, 앞 절의 식 (1)에서 기술한 TF-IDF와 아래 식 (2)에서 나타낸 χ^2(chi-square) 통계량을 주로 이용한다.

본 논문에서는 χ^2 통계량을 이용하여 특징 선택을 하였다. χ^2 통계량은 용어 t와 클래스 c의 의존성(dependency)을 측정하는 것으로서 이를 이용하여 클래스 c와 용어 t사이에 존재하는 관계를 수치적으로 나타낼 수 있다. 용어 t와 클래스 c가 완전히 독립적이면 0의 값을 가지게 된다. 수식은 식 (2)와 같다.

$$x^2(t,c) = \frac{N\times(A\times D - C\times B)^2}{(A+C)\times(B+D)\times(A+B)\times(C+D)} \quad (2)$$

$$x^2_{\max}(t) = \max_{t=1}^{m} x^2(t,c_i)$$

A : c에 속한 문서 중 t를 포함하는 문서 수
B : c 아닌 클래스에 속한 문서 중 t를 포함하는 문서 수
C : c에 속한 문서 중 t를 포함하지 않는 문서 수
D : c 아닌 클래스에 속한 문서 중 t를 포함하지 않는 문서 수
N : 전체 학습 문서 수

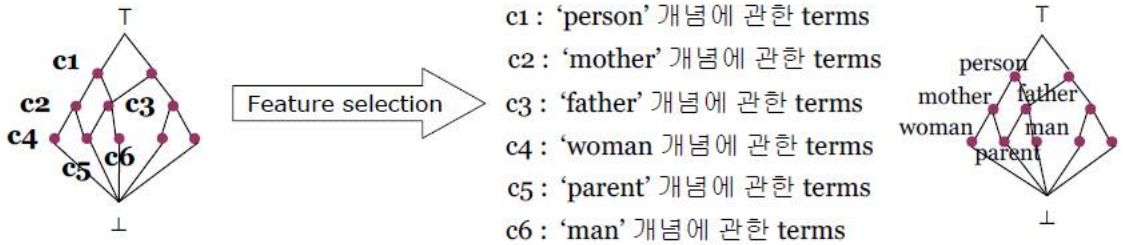
예를 들어 <그림 12>와 같이 진행한 특징 선택

과정의 결과는 좌표 평면 상에서 <그림 13>과 같이 표현된다. 특징 선택을 통해 추출된 클러스터간의 단어를 보고 각 클러스터의 개념을 추정하여 트리에 적용해 보면 <그림 14>와 같이 의미적 계층 관계가 성립하는 개념들의 관계를 확인할 수 있다.

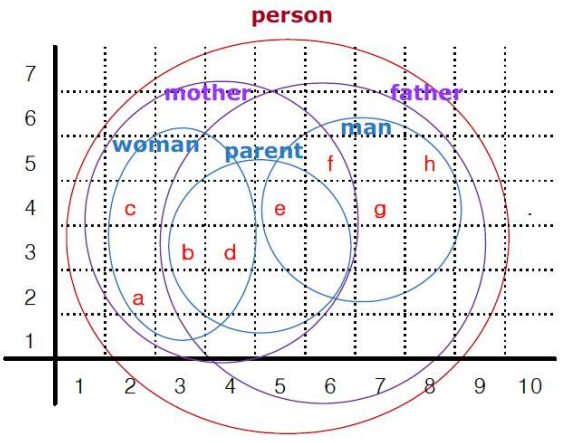
a, b, c, d 문서 집합(클러스터)은 특징 선택을 통해 추출된 단어를 통해 woman이라는 대표 개념을 연상할 수 있다. 추정된 개념들은 <그림 14>와 같은 구조로 표현될 수 있다. <그림 14>를 통해 parent라는 대표 개념으로 표현된 클러스터는 상위의 동일 레벨에 있는 mother와 father로 표현된 서로 다른 개념 클러스터에 중복으로 속하는 복합 개념임을 알 수 있다. 또한, 기존의 계층적 클러스터링 알고리즘에서 발견하지 못했던 father라는 클러스터를 새로이 탐지하여 표현하였음을 알 수 있다.

5. 실험 및 평가

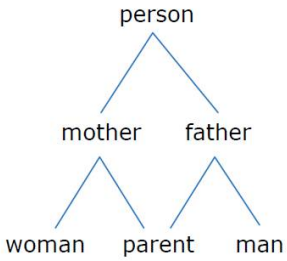
이 장에서는 실험 데이터(예제 문서 집합)를 이용하여 본 논문에서 제안한 HOC 알고리즘으로 클러스터링을 수행하고, 그 결과에 특징 선택을 적용하여 클러스터별로 대표되는 개념을 추정하고, 개념들의 의미적 계층 관계를 검증하는 과정을 보여준다. 이를 바탕으로 문서 집합의 내용을 분석하고



<그림 12> 특징 선택 흐름도



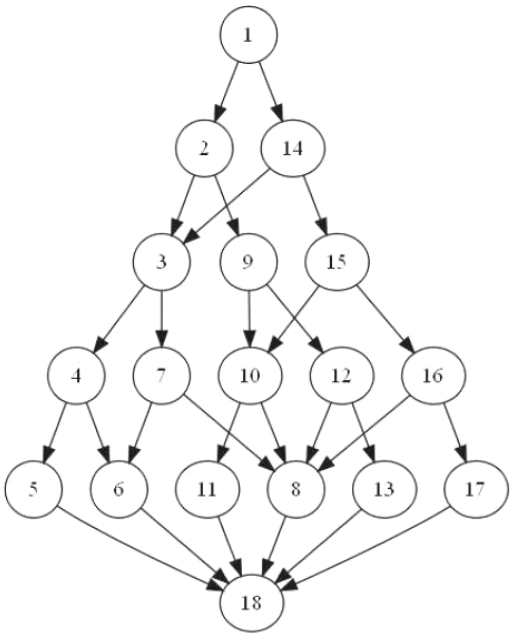
<그림 13> 특징 선택의 좌표 평면 상의 결과



<그림 14> 개념의 의미적 계층

평가하였다.

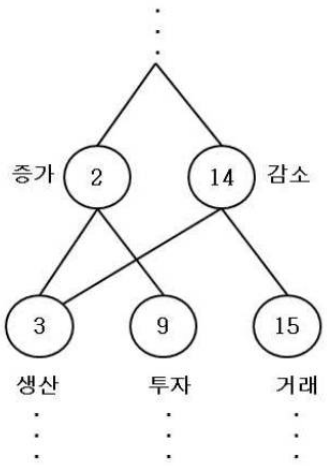
실험에 사용된 데이터는 로이터 뉴스 기사를 모은 Reuters-21578자료 집합이다(Lewis, 2004). Reuters-21578자료 집합은 119개의 클래스와 21578개의 문서를 가지고 있다. 본 논문에서는 이 중에서 일반적으로 많이 쓰이는 클래스인 “wheat”와 “corn” 클래스에서 각각 3개씩 문서를 선정해 총 6개의 문서로 실험하였다. 실험 데이터에 HOC 알고리즘으로 클러스터링을 하여 얻어진 구현 결과는 <그림 15>와 같이 Lattice로 표현되었다.



<그림 15> 예제 문서집합에 대한 클러스터링 결과 Lattice

<그림 15>에서 클러스터 노드 3, 8, 10은 동일한 레벨의 서로 다른 클러스터들에 각각 중복으로 속하므로 복합 개념으로 간주할 수 있다(클러스터 3은 클러스터 2, 14에, 클러스터 8은 클러스터 7, 10, 12, 16에, 클러스터 10은 클러스터 9, 15에 각각 속하는 것을 Lattice 구조를 통해 알 수 있다). 특징 선택을 이용해 각 클러스터 노드의 주요 단어들을 구하여 각 클러스터를 대표하는 개념을 추정하였다. 특징 선택을 위한 수식으로는 앞 절의 식 (2)에서 기술한 max를 고려하는 χ^2 통계량을 사용하였다.

<그림 15>의 클러스터 2, 3, 9, 14, 15에 특징 선택 과정을 적용한 결과 <그림 16>과 같이 표현되었다. 즉, 클러스터 2에서 특징 선택을 사용해 추출된 단어를 통해 2를 대표하는 개념을 “증가”라고 추정하였고, 나머지 클러스터 3, 9, 14, 15도 대표하는 개념을 각각 “생산”, “투자”, “감소”, “거래”라고 추정할 수 있었다.



<그림 16> 실험결과 Lattice에 속한 일부 개념의 추정과 계층 표현

<그림 16>은 클러스터 노드 2, 3, 9, 14, 15의 추정된 대표 개념과 각 개념 간의 계층 관계를 보여준다. <그림 16>을 통해 이 실험 데이터 문서 집합에서 투자는 증가하고 거래는 감소한다는 것을 알 수 있다. 클러스터 노드 3의 대표 개념인 생산은 복합 개념으로서 실험 데이터 집합 내에서는

증가되기도 하고 감소되기도 하며 양쪽에서 공존한다는 정보도 알 수 있다. 이런 과정을 거쳐 알고자 하는 문서들을 일일이 살펴보지 않아도 문서들의 주요 내용을 쉽게 파악할 수 있다.

<그림 15>에서 클러스터 노드 8은 다른 어떤 노드보다도 가장 많은 클러스터에 중복으로 속하는 복합 개념이다. 클러스터 노드 8은 실제 실험에서 위의 문서 집합을 표현하는 데 가장 구체적인(specific) 단어들이 포함되어 있다. 이 단어들은 8번 노드에서 특징 선택의 기여도 수식 값이 가장 컸고 위의 노드로 올라갈수록 값이 작게 나왔다. 실제 예제 문서를 대상으로 한 실험에서 클러스터 8에서 특징 선택을 통해 추출된 단어로는 "crop", "wheat", "corn" 등이 있고 이는 클러스터링 된 상위 노드의 어느 문서 집합에서도 공통적으로 나오는 가장 흔한 단어들이다. 이 단어들은 상위 노드의 다른 클러스터에서는 그 클러스터를 대표할 만큼 특징적이지 않다고 판단하여 특징 선택의 수식 값이 작다.

6. 결 론

본 논문에서는 기존의 계층적 클러스터링 알고리즘을 변형하여 동일 레벨에서 중복을 허용하는 계층적 클러스터링(HOC) 알고리즘을 개발하였고, 이를 이용하여 개념 간에 의미적인 계층관계를 가지면서도 여러 개념에 중복으로 속하는 복합 개념을 탐지하여 문서 집합의 내용을 쉽게 이해할 수 있는 방법을 제안하였다.

본 논문에서는 클러스터링 후 특징 선택을 이용하여 클러스터별 대표 단어를 추출하여 개념을 추측하고 각 개념들의 의미적 계층 관계를 검증하였다. 이를 통해 클러스터 노드별 고유정보를 표현하는 것이 가능해졌다. 제안한 HOC 알고리즘을 실제 데이터에 대해 실험한 결과 Lattice 형태의 클러스터 계층 구조를 얻을 수 있었고, 전체 실험 데이터문서를 일일이 검토하지 않고도 문서 집합인 클러스터 노드들이 각각 무슨 의미의 개념을 가지며 각 개념의 계층 관계가 어떠한지를 판단할 수 있었다.

본 논문에서 제안한 HOC 알고리즘의 문제점 중의 하나는 수행시 동일한 레벨의 서로 다른 클러스터에 중복으로 속하는 클러스터를 탐지하여 새롭게 추가하기 때문에 클러스터 노드의 의미적 계층 관계를 검증할 때 검증해야 할 노드의 수가 증가하여 분석해야 할 데이터가 너무 많아진다는 것이다. 따라서 향후 과제로 이러한 상황에서 발생하는 추가 비용를 줄일 필요가 있다. 이는 복합 개념으로 추정되는 노드만을 남기고 나머지 노드를 가지치기를 하면 어느 정도 해결되겠지만 향후 개념의 의미적 상하관계를 분석하는 일을 자동으로 하여 개념을 추론하고 레이블을 달 수 있게 한다면 지금보다 더 좋은 성능을 보이리라 예상된다.

참고문헌

Baeza-Yates, R. and B. Ribeiro-Neto, *Modern Information Retrieval*, Addison Wesley, 1999.

Chuang, S. and L. Chien, "A Practical Web-based Approach to Generating Topic Hierarchy for Text Segments, *Proc. 13th ACM Intl. Conf. on Information and Knowledge Management (CIKM'04)* (2004), 127~136.

Cleuziou, G., "An Extended Version of the *K*-means Method for Overlapping Clustering", *Proc. 19th Intl. Conf. on Pattern Recognition (ICPR 2008)*, (2008), 1~4.

Gath, I. and A. B. Geva, "Unsupervised Optimal Fuzzy Clustering", *IEEE Transactions on*

Pattern Analysis and Machine Intelligence, Vol.11, No.7(1989), 773~780.

Jain, A. K., M. N. Murty and P. J. Flynn, "Data Clustering : A Review", *ACM Computing Surveys*, Vol.31, No.3(1999), 264~323.

Jonyer, I., D. J. Cook and L. B. Holder, "Graph-Based Hierarchical Conceptual Clustering", *Journal of Machine Learning Research*, Vol.2(2002), 19~43.

Lavine, B. K., "Clustering and Classification of Analytical Data", in R. A. Meyers (ed.), *Encyclopedia of Analytical Chemistry*, (2000), 1~21.

Lewis, D., "Reuters-21578 Text Categorization Test Collection", 2004.

http://www.daviddlewis.com/resources/testcollections/reuters21578/.

Likas, A., N. Viassis and J. J. Verbeek, "The Global *K*-means Clustering Algorithm", *Pattern Recognition*, Vol.36, No.2(2003), 451~461.

Liu, B., *Web Data Mining : Exploring Hyperlinks, Contents, and Usage Data*, Springer, 2007.

Yeh, J. and S. Sie, "Towards Automatic Concept Hierarchy Generation for Specific Knowledge Network", *Lecture Notes in Artificial Intelligence(LNAI)*, Vol.4031 (2006), 982~989.

Abstract

Hierarchical Overlapping Clustering to Detect Complex Concepts

Sujeong Hong* · Joongmin Choi**

Clustering is a process of grouping similar or relevant documents into a cluster and assigning a meaningful concept to the cluster. By this process, clustering facilitates fast and correct search for the relevant documents by narrowing down the range of searching only to the collection of documents belonging to related clusters. For effective clustering, techniques are required for identifying similar documents and grouping them into a cluster, and discovering a concept that is most relevant to the cluster. One of the problems often appearing in this context is the detection of a complex concept that overlaps with several simple concepts at the same hierarchical level. Previous clustering methods were unable to identify and represent a complex concept that belongs to several different clusters at the same level in the concept hierarchy, and also could not validate the semantic hierarchical relationship between a complex concept and each of simple concepts.

In order to solve these problems, this paper proposes a new clustering method that identifies and represents complex concepts efficiently. We developed the Hierarchical Overlapping Clustering (HOC) algorithm that modified the traditional Agglomerative Hierarchical Clustering algorithm to allow overlapped clusters at the same level in the concept hierarchy. The HOC algorithm represents the clustering result not by a tree but by a lattice to detect complex concepts.

We developed a system that employs the HOC algorithm to carry out the goal of complex concept detection. This system operates in three phases; 1) the preprocessing of documents, 2) the clustering using the HOC algorithm, and 3) the validation of semantic hierarchical relationships among the concepts in the lattice obtained as a result of clustering.

The preprocessing phase represents the documents as x-y coordinate values in a 2-dimensional space by considering the weights of terms appearing in the documents. First, it goes through some refinement process by applying stopwords removal and stemming to extract index terms. Then, each index term is assigned a TF-IDF weight value and the x-y coordinate value for each document is determined by combining the TF-IDF values of the terms in it.

The clustering phase uses the HOC algorithm in which the similarity between the documents is calculated by applying the Euclidean distance method. Initially, a cluster is generated for each

* LG U+
** Department of Computer Science and Engineering, Hanyang University

document by grouping those documents that are closest to it. Then, the distance between any two clusters is measured, grouping the closest clusters as a new cluster. This process is repeated until the root cluster is generated.

In the validation phase, the feature selection method is applied to validate the appropriateness of the cluster concepts built by the HOC algorithm to see if they have meaningful hierarchical relationships. Feature selection is a method of extracting key features from a document by identifying and assigning weight values to important and representative terms in the document. In order to correctly select key features, a method is needed to determine how each term contributes to the class of the document. Among several methods achieving this goal, this paper adopted the χ^2 statistics, which measures the dependency degree of a term t to a class c, and represents the relationship between t and c by a numerical value.

To demonstrate the effectiveness of the HOC algorithm, a series of performance evaluation is carried out by using a well-known Reuter-21578 news collection. The result of performance evaluation showed that the HOC algorithm greatly contributes to detecting and producing complex concepts by generating the concept hierarchy in a lattice structure.

Key Words : Hierarchical Overlapping Clustering, Complex Concept Detection, Feature Selection, Concept Labeling

저 자 소 개

홍수정

한양대학교 전자컴퓨터공학부를 졸업하였고, 2009년에 한양대학교 대학원 컴퓨터공학과에서 석사학위를 취득하였다. 현재 LG유플러스(구, LG텔레콤)에 재직 중이다. 관심분야는 데이터마이닝, 정보검색/추출, 데이터베이스, 이중화, DR구축, 클라우드 등이다.

최중민

서울대학교 컴퓨터공학과를 졸업하였고, 1986년에 서울대학교 대학원 컴퓨터공학과에서 석사학위를, 1993년에 미국 State University of New York at Buffalo에서 컴퓨터학 박사 학위를 각각 취득하였다. 1993년부터 1995년까지 한국전자통신연구원(ETRI)에서 선임연구원으로 재직하였으며, 현재 한양대학교 컴퓨터공학과 교수로 재직 중이다. 한국 정보과학회, 정보처리학회, 지능정보시스템학회, 인터넷 정보학회, 미국 IEEE, ACM 등의 정회원이며, 관심분야는 웹지능, 텍스트마이닝, 정보검색/정보추출, 인공지능, 지능형 모바일 정보시스템 등이다.