한국 퍼지 및 지능시스템 학회 2002, Vol. 12, No. 3, pp. 219-226

웹마이닝을 위한 퍼지 클러스터링 알고리즘

Fuzzy Clustering Algorithm for Web-mining

임영희* · 송지영** · 박대희**
Younghee Im*, Jiyoung Song**, and Daihee Park**

*대전대학교 컴퓨터정보통신공학부
**고려대학교 컴퓨터정보학과

요 약

웹 검색 엔진의 검색 결과를 클러스터링하는 후처리 클러스터링 알고리즘은 그 특성상 일반적인 클러스터링 알고리즘과는 다른 요구조건을 갖는다. 본 논문에서는 이러한 후처리 클러스터링 알고리즘의 요구조건들을 최대한 만족하는 새로운 클러스터링 알고리즘을 제안하고자 한다. 제안된 Fuzzy Concept ART는 문서 클러스터링에 있어 여러 가지 장점을 갖는 개념벡터와 실시간 클러스터링 알고리즘으로 알려진 Fuzzy ART를 퍼지이론에 기반하여 결합한 형태로써, 후처리 클러스터링 뿐 아니라 범용의 클러스터링 알고리즘으로도 응용이 가능하다.

Abstract

The post-clustering algorithms, which cluster the results of Web search engine, have some different requirements from conventional clustering algorithms. In this paper, we propose the new post-clustering algorithm satisfying those of requirements as many as possible. The proposed Fuzzy Concept ART is the form of combining the concept vector having several advantages in document clustering with Fuzzy ART known as real time clustering algorithms on the basis of fuzzy set theory. Moreover we show that it can be applicable to general-purpose clustering as well as post-clustering.

Key Words : clustering algorithm, fuzzy ART, concept vector, fuzzy concept ART

1. 서 론

최근 컴퓨터와 인터넷 환경의 발달로 인하여 비구조화된 텍스트 문서의 양은 폭발적으로 증가하고 있다. 따라서 텍스트 문서의 집합으로부터 잠재적 개념(latent concept)을 발견하여 그룹화하는 문서 클러스터링은 웹마이닝이나 정보 검색과 같은 실제적인 영역에서 매우 유용하게 사용될 수 있다. 전통적인 정보 검색 엔진의 경우, 사용자의 질의에 대한 검색 결과를 질의와의 관련 정도에 따라 순위가 매겨진 매우 긴 문서 목록의 형태로 사용자에게 제공한다. 이는 초기 인터넷상의 정보 부재로 인하여 검색 엔진의 목표가 되도록 많은 검색 결과를 보여주는데 있었기 때문이다. 그러나 이제는 넘쳐나는 정보들로 인하여, 오히려 검색 결과의 양이 너무 많아 단순한 문서 목록 형태의 검색 결과만으로는 사용자들이 원하는 정보를 찾기가 더욱 힘들어지고 있다. 따라서 검색 엔진에 의해 제공된 일차 검색 결과를 공통된 주제끼리 그룹화하여 사용자에게 가공된 정보의 형태로 제공하고자 하는 연구들이 최근 들어 활발히 진행되고 있으며[1-5], 이는 자연어 검색 기법과 함께 차세대 정보 검색 서비스의 대안으로 여겨지고 있다[6].

문서 클러스터링 기법은 정보 검색 시스템이나 지식 관리 시스템의 전체 문서 코퍼스(corpus)를 오프라인 상에서 미리 클러스터링하여, 질의 요청 시 해당 질의와 가장 유사한 클러스터에 대해서만 검색을 수행하는 "전처리 클러스터링 기법"[8,9]과 질의에 대한 검색 결과를 온라인 상에서 즉각적으로 클러스터링하는 "후처리 클러스터링 기법"으로 나눌 수 있다. 본 연구에서는 문서 클러스터링 기법 중 후처리 클러스터링 기법에 초점을 맞추고자 한다.

후처리 클러스터링 기법은 그 성격상 전처리 클러스터링 기법 혹은 범용의 클러스터링 기법과는 매우 다르므로, 자체 기법의 성격을 잘 반영한 새로운 평가기준이 요구된다. 관련 연구에 의하면, 후처리 클러스터링 기법의 유효성을 평가하는데 사용되는 평가기준(즉, 기본 요건)들은 아래와 같이 정리된다[1]:

1. 관련성(relevance): 검색된 문서들을 사용자의 질의와 관련된 것과 그렇지 않은 것으로 클러스터링할 수 있어야 한다.
2. 요약(summaries): 사용자는 클러스터의 내용이 관심이 있는 것인지 없는 것인지 한 눈에 판단할 수 있어야 한다. 따라서 해당 클러스터에 대한 간결하고 정확한 요약을 제공할 수 있어야 한다.
3. 중복(overlap): 일반적으로 문서들은 하나 이상의 주제를 내포하고 있을 수 있으므로 하나의 문서가 여

접수일자 : 2001년 10월 6일
완료일자 : 2002년 5월 2일
이 논문은 2000년도 한국학술진흥재단의 지원에 의하여 연구되었음(KRF-2000--041-E00273)

러 개의 클러스터에 속할 수 있어야 한다.

4. 발췌문-허용오차(snippet-tolerance): 웹 문서 전체를 입력으로 수행하는 방법 대신, 검색 엔진에 의하여 리턴된 문서의 발췌문만을 가지고 클러스터링을 수행하여도 좋은 결과를 보여야 한다.
5. 속도(speed): 실제 정보 검색에 응용되기 위해서는 클러스터링의 수행 속도가 빨라야 한다.
6. 점증성(incrementality): 클러스터링 수행시간을 줄이기 위하여 모든 검색 결과가 전송될 때까지 기다리는 것이 아니라, 검색된 문서들이 도착하는 즉시 클러스터링을 수행할 수 있어야 한다.

관련 연구 [1]에서 제시한 위의 평가기준 외에도, 다음의 평가기준들을 추가적으로 고려해야 한다.

7. 사전지식(priori knowledge): 클러스터링에 적용되는 문서 집합들이 질의에 따라 각기 다른 특성을 나타내므로 클러스터의 개수와 같은 사전 지식을 요구하지 않는 방법이어야 한다.
8. 메모리 복잡도(memory complexity): 상업용 검색 엔진의 경우, 하루에도 수 백만 건의 질의를 처리해야 하므로 검색 엔진 서버의 과부하 문제 및 메모리 복잡도 문제 등이 고려되어야 한다.

일반적으로, 인공지능 분야에서 개발된 기존의 클러스터링 알고리즘들은 고차원의 대규모 데이터 집합에 적용하기가 어려우며, 후처리 클러스터링 기법을 위하여 개발된 최근의 알고리즘들도 위의 여러 가지 기준들 중 일부 항목에만 초점을 맞추는 방식을 취하고 있다. 따라서 본 논문에서는 후처리 클러스터링 기법이 갖추어야 할 기본 요건들을 최대한 충족시킬 수 있는 새로운 형태의 후처리 클러스터링 알고리즘을 제안하고자 한다. 제안된 후처리 클러스터링 알고리즘인 Fuzzy Concept ART는 문서 클러스터링에 있어 여러 가지 장점을 갖는 개념 벡터와 실시간 클러스터링 알고리즘으로 알려진 Fuzzy ART[13] 신경망을 퍼지이론에 기반하여 결합한 형태로써, 기본 요건 중, 1, 2, 3, 4, 5, 7, 8번째 요건을 모두 만족한다. 또한 Fuzzy Concept ART를 신경망 관점에서 바라보면, Fuzzy ART가 갖는 단점들을 극복한, 보다 강건하고 효율적인 새로운 ART 모델의 개발이라는 데 그 의의가 있다.

본 논문의 구성은 다음과 같다. 2장에서는 문서 클러스터링에 대한 관련 연구들을 간략히 살펴보고, 3장에서는 클러스터링 알고리즘의 입력으로 사용될 문서들의 벡터화에 대해 설명한다. 4장에서는 Fuzzy Concept ART의 근간을 이루는 개념 벡터와 Fuzzy ART, 그리고 본 논문에서 새롭게 제안된 Fuzzy Concept ART에 대해 기술한다. 5장에서는 실험결과 및 분석을 기술한다. 마지막으로 6장에서는 결론 및 향후 연구과제에 대해 논한다.

2. 관련 연구

Dhillon[7]은 문서들이 포함하고 있는 용어(term)들을 특징량(feature)으로 하여, 전체 문서 집합에 대한 벡터 공간을 모델링하였다. 이렇게 모델링된 문서 벡터들은 매우 고차원적이며, sparse한 특성을 나타내므로, 저차원의 dense한 데이터를 마이닝할 때와는 다른 특성을 갖는 Spherical K-Means 클러스터링 알고리즘을 제안하였다. Spherical K-Means 알고리즘에 의해 생성된 각 클러스터는 단위 유클리디안(Euclidean) 노름(norm)을 갖도록 정규화된 중점 벡터(centroid vector), 즉 개념 벡터(concept vector)에 의해 대표된다. Spherical K-Means 알고리즘은 각 문서 벡터가 자신을 포함하는 클러스터의 개념 벡터와 가장 큰 코사인 유사도(cosine similarity)를 갖도록 문서 벡터 공간을 분할한다. 상당히 큰 규모의 문서 집합에 대한 실험 결과, Spherical K-Means 알고리즘은 상당히 좋은 클러스터링 결과를 보이며, 각 클러스터는 개념 벡터에 의해 잘 요약된 클러스터 레이블(label)을 갖는다. 그러나 Spherical K-Means 알고리즘은 클러스터의 개수를 미리 지정해줘야 하는 문제점이 있다. 실제로 검색 결과는 질의에 따라 매우 동적인 특성을 보이므로, 검색 결과가 몇 개의 클러스터로 구성되어 있는지를 사전에 알기란 불가능하다. 따라서 클러스터의 개수를 미리 지정해줘야 하는 제약조건은 Spherical K-Means 알고리즘의 실제적인 응용을 어렵게 만드는 요인이 된다. 또한 K-Means 알고리즘과 마찬가지로 초기 클러스터가 클러스터링 성능을 좌우한다는 문제점을 안고 있다.

한편, 신경망의 ART(Adaptive Resonance Theory)가 가지고 있는 여러 가지 장점에도 불구하고, 현재까지 문서 클러스터링에 ART를 적용한 연구는 변형된 ART2를 이용하여 웹 페이지를 분류하는 Vlajic[8,9]의 연구를 제외하고는 발견되지 않고 있다. Vlajic[8]은 문서 집합을 벡터 모델로 변환한 다음, ART2의 변형을 이용하여 문서 벡터들을 분류하였다. 변형된 ART2는 각 클러스터의 가중치 벡터가 해당 클러스터에 소속된 문서 벡터들의 중점값을 갖도록 학습되며, 경계 변수 대신 가중치 벡터와 문서 벡터사이의 불일치 정도에 대한 허용오차(tolerance) 변수를 두어 분류 강도를 조정하였다. 또한, Vlajic[9]에서는 문서상의 용어뿐 아니라, 웹 문서가 갖는 12개의 하이퍼링크 정보도 문서 표현에 함께 사용하였다. 이들의 연구가 단지 25개의 웹 문서만을 대상으로 수행되었고, 전처리 클러스터링에 초점을 맞추고는 있지만, ART가 문서 클러스터링에 적용될 수 있음을 보였다는데 그 의미를 찾을 수 있다.

3. 문서의 표현형태

본 논문에서는 후처리 클러스터링을 위한 문서의 표현형태로써 벡터공간 모델을 사용하고자 한다. 따라서 각 문서들은 가중치가 부여된 용어 빈도수의 벡터로써 표현된다. 이때 앞서 언급했듯이, 검색 결과의 클러스터링은 온라인 상에서 실시간으로 수행되어야 하므로 성능 평가시 속도가 매우 중요한 평가기준이 된다. 따라서 문서 전체에 대해 문서 벡터를 구성하는 대신 문서의 일부만으로 문서 벡터를 구성하는 방법 등이 제안되었다[1,2,5]. 관련 연구결과들을 종합해 볼 때, 검색엔진에 의해 제공되는 발췌문만을 가지고 문서 벡터를 구성하는 방법론은 클러스터링의 정확도를 어느 정도 유지하면서 클러스터링의 탐색공간을 크게 줄일 수 있다. 또한 문서 제목의 경우, 해당 문서의 전체 내용을 대표하는 특성을 가지므로 문서의 제목 역시 문서 벡터 구성 시 매우 중요한 요소가 된다. 따라서, 본 논문에서는 발췌문과 해당 문서의 제목만을 입력하여 문서 벡터를 구성함으로써 클러스터링의 속도를 향상하고자 한다(평가기준 4: 발췌문-허용오차). 또한 발췌문과 제목만을 문서 표현으로 이

용할 경우, 해당 클러스터링 툴은 사용자의 클라이언트 머신(client machine)에 탑재되어 실행될 수 있으며, 이는 검색 엔진 서버의 과부하를 줄일 수 있다는 또 다른 장점을 갖는다(평가기준 5: 속도 및 8: 메모리 복잡도).

다음은 문서 벡터로의 변환을 위한 일반적인 전처리 과정이다[10].

1. 전체 문서 집합으로부터 모든 용어을 추출한다.
2. 의미 없는 용어, 즉 "stop list"에 등록된 단어들을 제거한다.
3. 각 문서에 대하여 용어의 빈도수를 계산한다.
4. 휴리스틱에 의해, 매우 높은 빈도수(high frequency)를 갖는 용어와 매우 낮은 빈도수(low-frequency)를 갖는 단어는 기능어(function word)로 간주하여 제거한다.
5. 위 과정의 수행 후, d 개의 용어가 남았다고 가정하면, 각 단어에 1부터 d까지의 인덱스를 할당한다. 마찬가지로 각 문서에 1부터 n까지의 인덱스를 할당하면, 전체 문서 집합에 대한 벡터공간 모델인 $d \times n$ 차원의 행렬 D가 완성된다. 이때 $D_{i,j}$는 j번째 문서에서의 i번째 용어에 대한 tf/idf(term frequency/inverse document frequency) 값이다

단, 각 문서가 검색 엔진으로부터 얻은 웹 문서라면 단계 1의 수행 전에 HTML 태그를 제거해주는 추가 작업이 필요하다.

본 논문에서는 문서의 벡터화를 위해 MC[11] 프로그램을 사용한다. MC는 대규모의 문서 집합으로부터 아주 빠르게 문서 벡터를 생성해주며, 생성된 문서 벡터는 nonzero 값만을 저장하는 CCS(Compressed Column Storage)[12] 포맷으로 저장된다. 따라서 저장 공간을 절약할 수 있으며, Fuzzy Concept ART에서의 코사인 유사도 계산량을 크게 줄일 수 있다.

4. Fuzzy Concept ART

4.1 개념 벡터

클러스터링 알고리즘에서 두 문서간의 유사도를 결정하는 문제는 클러스터링 알고리즘의 선택 못지 않게 중요한 문제이다. 본 연구에서는 코사인 유사도로 두 문서간의 유사도를 측정하고자 한다. 코사인 유사도는 이해하기 쉽고, sparse 벡터에 대해 계산이 단순하기 때문에 정보 검색이나 텍스트 마이닝에서 널리 사용되는 유사도이다[10].

본 논문에서는 각 문서 벡터 $X_1, X_2, \cdots, X_n$가 단위(unit) L_2 노름을 갖도록 정규화한다[7]. 이러한 정규화는 문서 벡터들의 방향성(direction)만을 유지하게 해주므로, 문서의 길이가 다르더라도 같은 주제를 다루는 문서들(즉, 유사한 용어들로 구성된 문서들)을 유사한 문서 벡터로 변환해 주는 효과가 있다. 또한, 두 문서 벡터 X_i와 X_j사이의 코사인 유사도는 다음과 같이 두 벡터 사이의 내적(inner product)으로 간단히 구할 수 있다.

$$S(X_i, X_j) = X_i^T X_j = \|X_i\| \|X_j\| \cos(\theta(X_i, X_j)) = \cos(\theta(X_i, X_j)) \quad (4\text{-}1)$$

여기서, 두 벡터사이의 각(angle)은 $0 \le \theta(X_i, X_j) \le \pi/2$이다.

n개의 문서 벡터들이 c개의 클러스터 $\pi_1, \pi_2, \cdots, \pi_c$로 나누어진다고 가정하자. 각 클러스터 π_j의 대표 벡터(representative vector)로 가장 널리 사용되는 평균(mean) 벡터 또는 중점 벡터는 다음과 같이 정의된다.

$$M_j = \frac{1}{n_j} \sum_{X \in \pi_j} X \quad (4\text{-}2)$$

여기서, n_j는 클러스터 π_j에 속해있는 문서의 수이다. 이때 만약 중점 벡터 M_j를 다음과 같이 단위 노름을 갖도록 정규화하면, 중점 벡터의 방향성만을 갖는 개념 벡터 C_j를 정의할 수 있다[7].

$$C_j = \frac{M_j}{\|M_j\|} \quad (4\text{-}3)$$

위와 같이 정의된 개념 벡터 C_j는 다음과 같은 특성을 갖는다. $R^d_{\ge 0}$상의 임의의 단위 벡터 Z에 대해, 다음의 Cauchy Schwarz 부등식을 유도할 수 있다.

$$\sum_{X \in \pi_j} X^T Z \le \sum_{X \in \pi_j} X^T C_j \quad (4\text{-}4)$$

위의 식 (4-4)에 의해 개념 벡터 C_j는 클러스터 π_j에 속해 있는 모든 문서 벡터에 대해 가장 근접한 코사인 유사도를 갖는 벡터임을 알 수 있다. 이러한 개념 벡터는 문서 벡터의 sparse한 특성을 그대로 유지하므로, 간단한 계산만으로 코사인 유사도나 클러스터의 응집력 등을 계산할 수 있다. 따라서 수행 속도에 상당히 민감한 후처리 클러스터링의 계산 복잡도를 크게 줄일 수 있다. 또한 개념 벡터들이 각 클러스터에 대해 지역화되므로[7], 본 연구에서 충족시키고자 하는 후처리 클러스터링의 요약 조건을 만족시킬 수 있는 좋은 아이디어를 제공한다. 즉, 클러스터링 수행 결과를 사용자에게 어떤 형태로 서비스 할 것인가(visualization)는 클러스터링 과정 못지 않게 매우 중요한 과제이다. 따라서 클러스터 내 문서들의 내용을 보다 쉽게 이해할 수 있는 레이블을 제공한다면, 사용자는 클러스터에 대한 레이블만을 보고 자신이 원하는 정보들이 포함되어 있는 클러스터를 선택할 수 있을 것이다.

n개의 문서 벡터들이 c개의 클러스터 $\pi_1, \pi_2, \cdots, \pi_c$로 나누어진다고 가정하면, 문서 클러스터 π_j의 키워드(keyword)는 용어 클러스터 $Word_j$로 표현할 수 있다. 클러스터 π_j의 대표 벡터인 개념 벡터 C_j의 각 용어 중 다른 개념 벡터에서의 가중치보다 큰 가중치를 갖는 용어는 용어 클러스터 $Word_j$에 속하게 된다[7].

$$Word_j = \{ k\text{th word} : 1 \le k \le d,\ C_{k,j} \ge C_{k,m},\ 1 \le m \le c,\ m \ne j \} \quad (4\text{-}5)$$

이때, d는 전체 용어의 개수이다. 이렇게 구성된 각 클러스터에 대한 용어 클러스터 $Word_j$는 개념 벡터가 해당 클러스터에 대해 지역화되는 특성을 가지므로, 비교적 좋은 키워드들을 제공해준다. 또한, 각 클러스터에 소속된 문서들 중 해당 클러스터의 개념 벡터와 가장 큰 코사인 유사도를 갖는 문서를 요약(summary)으로 제공함으로써, 단순한 용어의 나열에서는 얻기 어려운 각 클러스터에 대한 직관적인 이해가 가능하다.

$$Summary_j = \arg \max_{x \in \pi_j} \{\cos(\theta(X, C_j)\} \quad (4\text{-}6)$$

4.2. Fuzzy ART

신경망 ART는 기존에 학습되었던 것이 새로운 학습에 의해 지워지지 않도록 새로운 지식을 전체 데이터베이스에 일관성 있는(self-consistent) 방법으로 통합한다. 실시간 클러스터링 알고리즘으로 알려진 ART는 다른 신경망에 비해 다음과 같은 장점을 갖는다. 첫째, ART는 비교사 학습에 의해 입력 패턴을 클러스터링하므로, 사전에 학습 데이터를 통한 훈련없이 새로운 입력 패턴을 학습할 수 있다. 둘째, ART는 기존 신경망들의 딜레마인 "stability-plasticity" 문제를 해결한다. 신경망에서 stability란 이전에 학습한 패턴들에 대한 기억을 안정적으로 유지하는 능력을 말하며, plasticity란 이전에 학습한 적이 없는 새로운 패턴을 처리할 수 있는 능력을 말한다. ART는 입력 패턴과 학습된 클러스터간의 비교를 통해, 이미 학습된 클러스터에 영향을 미치지 않으면서 학습을 수행할 수 있는 reset 메커니즘을 사용하여 이 딜레마를 해결하였다. 셋째, ART는 경계 변수(vigilance parameter) 값에 따라 클러스터링의 분류 결과를 조정할 수 있다. 즉, 경계 변수의 값을 크게 주면, 좀 더 세분화되고 구체적인 클러스터들을 얻을 수 있다. 또한, 학습이 완료된 클러스터에 대한 가중치 값들은 해당 클러스터에 속해 있는 패턴들에 대한 대표 벡터로 해석될 수 있다.

ART에는 이진 벡터를 클러스터링하는 ART1과 아날로그 벡터를 클러스터링하는 ART2가 있으며, Fuzzy ART는 ART1의 교집합(intersection) 연산을 퍼지 집합 이론의 min 연산으로 대치함으로써 ART1이 아날로그 벡터에 대하여 학습할 수 있도록 하였다. Fuzzy ART 알고리즘의 세부적인 수행 과정은 다음과 같다[13]. 먼저, 입력 벡터 $\boldsymbol{X}_i=(X_{1,i}, X_{2,i}, \cdots, X_{d,i})$, $i=1,\cdots,n$는 각 콤포넌트가 [0, 1]의 값을 가지며, $\boldsymbol{W}_j^{(t)}=(W_{1,j}^{(t)}, W_{2,j}^{(t)}, \cdots, W_{d,j}^{(t)})$, $j=1,\cdots,c$는 카테고리 j에 대한 가중치 벡터라 하자. 이때 n은 입력 패턴의 수이고, d는 입력 패턴의 차원, c는 카테고리의 개수이다. 또한 Fuzzy ART의 동적 특성을 결정하는 변수들은 선택 변수 α (>0), 학습 변수 β ($\in[0,1]$), 경계 변수 ρ ($\in[0,1]$)등이 있다.

<u>초기화</u>. 카테고리 개수 c와 변수 α, β, ρ값을 초기화하고, 초기 가중치 벡터는 첫 번째 다음과 같이 초기화한다:

$$W_{1,j} = W_{2,j} = \cdots = W_{d,j} = 1, \quad j=1,\cdots,c \quad (4\text{-}7)$$

<u>활성화 함수(Activation Fuction: AF)</u>. 입력 패턴과 가중치 벡터 사이의 매칭 정도를 측정하는 활성화 함수는 퍼지 집합 이론의 min 연산을 이용하여 다음과 같이 계산한다.

$$AF(W_j^{(t)}, X_i) = \frac{\sum_{k=1}^{d} \min(X_{k,i}, W_{k,j}^{(t)})}{\alpha + \sum_{k=1}^{d} W_{k,j}^{(t)}}, \quad j=1,\cdots,c \quad (4\text{-}8)$$

<u>클러스터의 선택</u>. ART는 경쟁 학습의 winner-take-all(WTA) 전략을 통해 입력 패턴들을 학습한다. 따라서 Fuzzy ART는 다음과 같이 각 클러스터에 대한 활성화 함수 값에 의해 가장 유사한 클러스터 유닛 $j*$을 선택한다.

$$j* = \arg \max_{j=1,\cdots,c} \{AF(W_j^{(t)}, X_i)\}. \quad (4\text{-}9)$$

<u>Resonance 유닛의 선택</u>. 식 (4-9)에 의해 선택된 클러스터에 대해 다음의 매칭 함수(Matching Function: MF)에 경계 변수 조건을 적용하여 resonance 유닛을 선택한다.

$$MF(W_{j*}^{(t)}, X_i) = \frac{\sum_{k=1}^{d} \min(X_{k,i}, W_{k,j*}^{(t)})}{\sum_{k=1}^{d} X_{k,i}} \geq \rho \quad (4\text{-}10)$$

만일 클러스터 $j*$가 위의 경계 변수 조건을 만족하면, 해당 입력 패턴을 학습할 수 있도록 가중치 갱신이 발생하고, 그렇지 않으면 해당 클러스터의 활성화 함수를 reset하며(즉, $AF(W_{j*}^{(t)}, X_i^{(t)}) = -1$), 조건을 만족하는 클러스터를 찾을 때까지 식 (4-9)에 의해 새로운 $j*$를 탐색한다.

<u>가중치 갱신(또는 학습)</u>. 선택된 카테고리 $j*$에 대해 조건 (4-10)이 만족되면, 다음의 식에 의해 가중치 벡터를 조정한다.

$$W_{j*}^{(t+1)} = \beta \cdot \min(W_{j*}^{(t)}, X_i) + (1-\beta) \cdot W_{j*}^{(t)} \quad (4\text{-}11)$$

Fuzzy ART는 아날로그 입력을 처리할 수 있도록 ART1을 일반화한 것이므로, ART1의 구조적 문제점을 그대로 안고 있다[14]. 즉 입력 패턴의 적용 순서에 매우 민감하며, resonance 유닛을 찾기 위해 최악의 경우 모든 활성화 함수 값을 정렬해야 한다. 또한 진정한 의미의 퍼지 클러스터링을 수행하지 않는다. 따라서 본 논문에서는 위와 같은 Fuzzy ART의 문제점을 해결하는 동시에 정보 검색 결과를 클러스터링하는 문제에 적용이 가능한 새로운 형태의 Fuzzy Concept ART를 제안하고자 한다.

4.3. Fuzzy Concept ART

Fuzzy Concept ART의 기본 아이디어는 각 클러스터 유닛의 가중치 벡터가 해당 클러스터의 개념 벡터가 되도록 하는 것이다. 이는 Dhillon[7]의 Spherical K-Means 알고리즘과 같이 각 클러스터의 개념 벡터와 입력 패턴의 코사인 유사도를 계산하여 가장 유사한 클러스터로 해당 입력 패턴을 할당하는 작업이지만, Spherical K-Means 알고리즘과는 달리 클러스터의 개수와 같은 사전 지식이 필요 없는 비교사 학습이며, 새로운 입력 패턴이 제시되었을 때 전체 시스템을 재학습할 필요 없이 점증적 학습(incremental learning)이 가능하다는 장점을 갖는다. 또한, 퍼지 이론을 적용함으로써 진정한 의미의 퍼지 클러스터링을 수행한다. 즉, 상대적 퍼지 소속값에 의해 입력 패턴이 각 클러스터에 소속될 정도를 표현하고(context-sensitive), 절대적 퍼지 소속 값으로 해당 입력 패턴이 잡음 데이터나 outlier인지를 결정한다(context-insensitive). 또한 입력 패턴과 가장 유사한 클

러스터의 가중치만을 갱신하는 것(WTA 전략)이 아니라, 입력 패턴이 각 클러스터에 소속될 정도에 따라 각 클러스터의 가중치를 학습함으로써 soft-competitive learning을 수행하게 된다. 따라서 Fuzzy Concept ART의 학습이 완료되면 입력 패턴들은 각 클러스터에 대한 상대적 퍼지 소속 값을 가지므로, 여러 주제를 내포하는 문서의 경우 여러 클러스터에 중복되어 소속될 수 있다.

초기화. 카테고리 개수 c는 1로 초기화하고, 입력 패턴들을 단위 L_2 노름을 갖도록 정규화한다. 또한, 초기 가중치 벡터는 첫 번째 입력 패턴으로 초기화한다:

$$W_1^{(0)} = X_1 \tag{4-12}$$

Fuzzy Concept ART에서 입력 패턴과 가중치 벡터 사이의 매칭 정도는 코사인 유사도에 의해 측정되므로, 식 (4-12)에 의해 첫 번째 입력 패턴과 초기 클러스터 유닛 사이의 코사인 값은 항상 1이 된다. 따라서 사용자가 어떤 값의 경계 변수($\rho \in [0,1]$)를 주더라도 첫 번째 패턴이 항상 첫 번째 카테고리에 할당됨을 보장할 수 있다.

활성화 함수. Fuzzy Concept ART의 활성화 함수는 다음과 같이 상대적 퍼지 소속 값으로 정의한다.

$$AF(W_j, X_i) = R_{ij} = \frac{A_{ij}}{\sum_{h=1}^{c} A_{ih}} \tag{4-13}$$

이때 절대적 퍼지 소속 값 A_{ij}는 다음과 같이 입력 패턴과 가중치 벡터 사이의 코사인 유사도로 정의한다.

$$A_{ij} = \cos(\theta(\mathrm{W_j}, \mathrm{X_i})) = \mathrm{X_i} \cdot \frac{\mathrm{W_j}}{\|\mathrm{W_j}\|} \tag{4-14}$$

매칭 함수. 경계 변수 조건의 만족 여부를 결정하는 매칭 함수는 위의 식 (4-14)에 의해 코사인 유사도 값을 갖는 절대적 퍼지 소속 값으로 정의된다.

$$MF(W_j, X_i) = A_{ij} \tag{4-15}$$

이렇게 정의된 활성화 함수와 매칭 함수는 다음의 조건 (4-16)를 만족하므로,

$$\begin{aligned} &MF(W_1, X_i) > MF(W_2, X_i) \\ &\Leftrightarrow AF(W_1, X_i) > AF(W_2, X_i) \end{aligned} \tag{4-16}$$

Concept ART[15][16]와 마찬가지로 최대 활성화 함수 값을 갖는 카테고리에 대해 매칭 함수가 경계 변수 조건을 만족하지 않으면 별도의 탐색 과정 없이 곧바로 새로운 카테고리를 생성하고, 해당 입력 패턴을 가중치 벡터로 할당한다.

Resonance 유닛의 선택. Resonance 유닛의 선택을 위한 경계 변수 테스트는 다음과 같다.

$$MF(W_{j*}, X_i) \geq \rho \tag{4-17}$$

이때, $j* = \arg \max_{j=1,\cdots,c} \{AF(W_j, X_i)\}$이다.

즉, Fuzzy Concept ART에서 활성화 함수 값은 현재 입력 패턴에 대한 해당 클러스터의 신뢰(credit) 정도를 나타내며, 매칭 함수 값은 입력 패턴이 해당 카테고리의 outlier인지 아닌지를 결정하는 척도가 된다.

가중치 갱신(또는 학습). Fuzzy Concept ART는 입력 패턴들이 식 (4-13)에 의해 각 클러스터에 대한 소속 정도 값을 가진다. 따라서 가중치 갱신시, 입력 패턴이 각 클러스터의 상대적 소속정도에 따라 해당 클러스터의 개념 벡터에 영향을 미쳐야 한다.

$$W_j^{(new)} = W_j^{(old)} + (R_{ij})^m \cdot X_i, \quad j = 1, \cdots, c \tag{4-18}$$

위의 식 (4-18)에 의해 Fuzzy Concept ART에서의 가중치 갱신은 최대값을 갖는 카테고리뿐만 아니라 기존의 모든 클러스터에 대해 수행되며, 이때 $m \in (1, \infty)$은 소속 정도에 대한 가중치 지수(weighting exponent)이다.

다음의 그림 1은 Fuzzy Concept ART 알고리즘을 요약 정리한 것이다.

Step0. Normalize input pattern with L2 norm.
Initialize Weights:
$W_1 = X_1$
Step1. While Stopping Condition is false, do Step 2-7
Step2. For each training input, do Step 3-6
Step3. Set activation to zero
Step4. Compute Activation Function:

$$AF(W_j, X_i) = \frac{A_{ij}}{\sum_{h=1}^{c} A_{ih}}, \quad j = 1, \cdots, c$$

where $A_{ij} = \cos(\theta(\mathrm{W_j}, \mathrm{X_i})) = \mathrm{X_i} \cdot \frac{\mathrm{W_j}}{\|\mathrm{W_j}\|}$

Step5. Find $j*$ with max activation
Step6. Test for reset:

If, $MF(W_{j*}, X_i) = A_{ij*} = X_i \cdot \frac{W_{j*}}{\|W_{j*}\|} \geq \rho$ then

$W_j^{(new)} = W_j^{(old)} + (R_{ij})^m \cdot X_i, \quad j = 1, \cdots, c$

else new processing element allocation:

$c = c + 1$

$W_c^{(new)} = X_i$

Step7. Test for stopping condition

그림 1. Fuzzy Concept ART 알고리즘
Fig. 1. Fuzzy Concept ART algorithm

5. 실험 결과 및 분석

본 실험은 질의어 "guinea"에 대해, 검색 엔진 Google이 리턴해준 상위 185개 문서들을 문서 제목과 발췌문만을 가지고 클러스터링하는 실험이다. 먼저 HTML 태그를 모두 제거한 다음, 불용어와 두 개 이하의 문서에서 발생한 용어는 제거한다[3]. 그 결과 문서 집합 GUINEA의 경우, 각 문서 벡터는 159차원의 sparse 벡터(96% sparsity)가 된다. 이렇게 구성된 문서 벡터 집합을 Fuzzy Concept ART를 사용하여 클러스터링하였다. 질의어 "guinea"외에도 다른 질의어에 대한 실험도 비슷한 결과를 관찰할 수 있었다[15]. 지면 관계상 이는 생략하

고자 한다.

데이터 집합 GUINEA를 Fuzzy Concept ART를 이용하여 퍼지 클러스터링한 결과는 다음의 표 1과 같다. 이때 경계 변수 ρ = 0.01, 가중치 지수 $m=2$로 설정하였으며, 각 클러스터에 대한 키워드와 요약, 그리고 개념 벡터와 가장 유사한 상위 5개의 문서를 표시하였다. 또한 표 1에서 각 문서의 제목 뒤에 표시된 수치는 해당 클러스터에 대한 상대적 소속 값을 의미한다.

표 1. Fuzzy Concept ART를 이용한 GUINEA의 클러스터링 결과(ρ = 0.01)

Table 1. Clustering result of Fuzzy Concept ART(ρ = 0.01)

	ρ = 0.01, Cluster 1, size = **47**
keywords	**pig, pigs, page, cavy, cavies, fowl**
summary	**Seagull's Guinea Pig Compendium (0.91)**

The Guinea Pig page (0.80)
Alyssa Buecker's Guinea Pig Tales (0.93)
Guinea Pig page(cavy, cavies) (0.83)
Guinea pig care: learn to tend to these docile... (0.91)
George's Guinea Pig (0.77)
⋮

	ρ = 0.01, Cluster 2, size = **58**
keywords	**papua, png, online, government, national, conversation**
summary	**Pupua New Guinea (0.56)**

About Pupua New Guinea (0.62)
Pupua New Guinea (0.51)
Papua New Guinea - research level book's... (0.67)
Pupua New Guinea (0.47)
Milne Bay Airlines(MBA)-Airlines of Papua New Guinea (0.56)
⋮

	ρ = 0.01, Cluster 3, size = **23**
keywords	**bissau, africa, map, index, guin, top**
summary	**Guinea-Bissau (0.78)**

Political Resources on the Net - Guinea-Bissau (0.71)
Guinea-Bissau vacation guide (0.67)
Guinea-Bissau: Peace Agreements Degital ... (0.70)
Government on WWW: Guinea-Bissau (0.69)
Guinea-Bissau Page (0.58)
⋮

	ρ = 0.01, Cluster 4, size = **23**
keywords	**equatorial, republic, recent, data, human, version**
summary	**Equatorial Guinea (0.81)**

Equatorial Guinea (0.87)
Adminet - Equatorial Guinea (0.67)
Equatorial Guinea (0.82)
NSRC Equatorial Guinea (0.68)
Ethnologue: Equatorial Guinea (0.77)
⋮

	ρ = 0.01, Cluster 5, size = **17**
keywords	**conakry, adventure, excite, september, travel, destination**
summary	**Travels in Guinea Conakry, African Adventure Trips (0.76)**

The Anglican Diocese of Guinea - Conakry (0.94)
Guineanews - News about Guinea Conakry (0.65)
Guinea (0.67)
Welcome to Paradise Live (0.88)
Excite Travel (0.73)
⋮

	ρ = 0.01, Cluster 6, size = **14**
keywords	**african, information, country, network, related, west**
summary	**Guinea - Consular Information Sheet (0.70)**

Guinea (0.94)
Guinea Economic Information & Indicator (0.68)
Rare Dog Bred - New Guinea Singing Dog (0.62)
Pupua New Guinea - News, Information, Foods, Travels... (0.37)
The USAID Leland Initiative and Guinea (0.72)
⋮

	ρ = 0.01, Cluster 7, size = **3**
keywords	**species, located, http, location**
summary	**Insect Ecology in New Guinea (0.99)**

Crocodilian Species - New Guinea Crocodile... (0.99)
Papua New Guinea OrchidNews: Species Photos (0.48)

실험 결과, Fuzzy Concept ART는 후처리 클러스터링 알고리즘으로써의 평가 기준 1, 2, 3, 5, 7, 8을 만족하며, 추가적으로 평가 기준 3의 중복 문제를 해결했음을 의미한다. 예를 들어, 표 1의 Cluster 7에 소속되어 있는 문서 "Pupua New Guinea Orchid News"의 경우, 생물 종

(species)에 관한 문서를 담고 있는 Cluster 7뿐만 아니라 Pupua New Guinea에 관한 문서들로 구성된 Cluster 2에도 속할 수 있다. Fuzzy Concept ART에 의한 학습 결과, 이 문서가 각 클러스터에 소속될 상대 퍼지 값은 다음의 표 2와 같다.

표 2. 문서 "Papua New Guinea OrchidNews"에 대한 상대 퍼지 소속값

Table 2. Fuzzy membership for document "Papua New Guinea OrchidNews"

Papua New Guinea Orchid News						
Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
0.03	0.23	0.06	0.03	0.12	0.05	0.48

또 다른 예로 Cluster 6의 문서 "Pupua New Guinea - News, Information, Foods, Travels..."를 살펴보자. Papua New Guinea에 대한 전반적인 정보를 모두 담고 있는 이 문서는 "information"이 키워드인 Cluster 6뿐 아니라, Pupua New Guinea에 관한 Cluster 2와 여행에 관한 Cluster 5에 대해서도 비교적 높은 소속 정도를 갖는다(표 3).

표 3. 문서 "Papua New Guinea - News, Information, Foods, Travels..."에 대한 상대 퍼지 소속값

Table 3. Fuzzy menmbership for document "Papua New Guinea - News, Information, Foods, Travels..."

Papua New Guinea - News, Information, Foods, Travels...						
Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
0.03	0.20	0.07	0.06	0.24	0.37	0.02

따라서 Fuzzy Concept ART에 의해 학습된 문서는 여러 주제를 내포할 경우, 여러 클러스터에 중복되어 소속될 수 있다(평가 기준 3: 중복).

6. 결론 및 향후 연구과제

본 논문에서 제안된 Fuzzy Concept ART는 서론에서 제시된 후처리 클러스터링 알고리즘의 여러 가지 평가 기준 중, 관련성, 요약, 발췌문-허용오차, 속도, 사전 지식, 메모리 복잡도, 중복 등 대부분의 요구조건들을 만족한다. 점증성은 검색 엔진에 의해 리턴된 문서가 도착하는 즉시 클러스터링을 수행함을 의미하는데, 이런 의미의 점증성은 문서 데이터를 벡터 공간으로 모델링하는 경우, 불가능한 요구조건이다. 왜냐하면 각 문서들이 갖는 특징량, 즉 용어들이 전체 문서 집합으로부터 추출된 것이므로 벡터 모델의 경우, 모든 문서들이 모두 도착한 후에야 문서 벡터를 구성할 수 있다. 그러나 일반적으로 언급되는 점증적 갱신의 경우, Fuzzy Concept ART는 전체 시스템을 재구성할 필요 없이 새로운 입력 패턴을 학습할 수 있으므로 점증적 학습이 가능하다. 또한 실험에서도 살펴보았듯이, Fuzzy Concept ART는 구조적 특성상, 후처리 클러스터링뿐만 아니라 범용의 클러스터링 알고리즘으로도 그 응용이 가능하다[15]. 향후 후처리 클러스터링 알고리즘의 좀 더 정확한 성능 평가를 위해 클러스터링 결과의 정확도에 대한 정량적인 평가방법의 개발이 요구된다. 또한 Fuzzy Concept ART의 실제적인 적용을 위해서는 보다 지능적이며, 사용자와 상호작용이 가능한 클러스터링 결과의 가시화에 대한 연구가 필요할 것이며, 경계 변수의 민감성 문제에 대한 추가 연구가 요구된다.

참 고 문 헌

[1] O. Zamir and O. Etzioni, "Web Document Clustering: A Feasibility Demonstration", Proceedings of the 19th International ACM SIGIR Conference on Research and Development in Information Retrieval(SIGIR '98), pp. 46-54, 1998.

[2] A. Leouski and W. B. Croft, "An Evaluation of Techniques for Clustering Search Results", Technical Report IR-76, University of Massachusetts at Amherst, 1996.

[3] D. S. Modha and W. S. Spangler, "Clustering Hypertext With Applications To Web Searching", Proceedings of ACM Hypertext Conference, 2000.

[4] M. A. Hearst and J. O. Pedersen, "Reexamining the Cluster Hypothesis: Scatter/Gather on Retrieval Results", Proceedings of ACM SIGIR '96, pp. 76-84, 1996.

[5] O. Zamir and O. Etzioni, "Grouper: A Dynamic Clustering Interface to Web Search Results", available at http://www.cs.washington.edu/zamir/papers/www8.ps.gz

[6] 박민우, "검색엔진의 과거와 현재 그리고 미래", *마이크로소프트웨어*, 2000년 3월호, pp. 220-235, 2000.

[7] I. S. Dhillon and D. S. Modha, "Concept Decomposition for Large Sparse Text Data using Clustering", Technical Report RJ 10147(9502), IBM Almaden Research Center, 1999.

[8] N. Vlajic and H. C. Card, "Categorizing Web Pages using Modified ART", *IEEE Canadian Conference*, Vol. 1, pp. 313-316, 1998.

[9] N. Vlajic and H. C. Card, "An Adaptive Neural Network Approach to Hypertext Clustering", *IEEE-INNS-ENNS International Joint Conference on Neural Networks*, Vol. 6, pp.3772-3726, 1999.

[10] W. B. Frakes and R. Baeza-Yates, "*Information Retrieval: Data Structures and Algorithms*", Prentice Hall, Englewood Cliffs, New Jersey, 1992.

[11] J. J. Fan, "MC: A Fast Sparse Matrix Generator For Large Text Collections", available at http://www.cs.utexas.edu/users/jfan/dm/

[12] Available at http://www.cs.utexas.edu/users/inderjit

/Resources/sparse_matrices.
[13] G. A. Carpenter, S. Grossburg, and D. B. Rosen, "Fuzzy ART: An Adaptive Resonance Algorithm for Rapid, Stable Classification of Analog Patterns", *Proceedings of 1991 International Conference Neural Networks*, Vol. II, pp. 411-416, 1991.
[14] A. Baraldi and E. Alpaydin, "Simplified ART: A New Class of ART Algorithms", International Computer Science Institute, TR 98-004, 1998.
[15] 임영희, "Fuzzy Concept ART: 웹 정보 검색을 위한 후처리 클러스터링 알고리즘", *고려대학교 박사학위 논문*, 2001.
[16] 임영희, "후처리 웹 문서 클러스터링 알고리즘", *정보처리학회논문지B*, 제 9-B권 제 1호, pp. 7-16, 2002.

저 자 소 개

임영희(Young-Hee Im)
1994년 : 고려대학교 전산학과(학사)
1996년 : 고려대학교 전산학과(석사)
2001년 : 고려대학교 전산학과(박사)
2001년~현재 : 대전대학교 컴퓨터정보통신공학부 강의전담교수

관심분야 : 인공지능, 정보 검색, 텍스트 마이닝, 데이터 마이닝, 데이터베이스 보안
E-mail : yheeim@dju.ac.kr

송지영(Ji-Young Song)
1996년 : 고려대학교 전산학과(학사)
1999년 : 고려대학교 전산학과(석사)
1999년~현재 : 고려대학교 전산학과 박사과정

관심분야 : 데이터 마이닝, 인공지능, 인공신경망, 생체인식, SVM
E-mail : songjy@korea.ac.kr

박대희(Dai-Hee Park)
1982년 : 고려대학교 수학과(학사)
1984년 : 고려대학교 수학과(석사)
1989년 : 플로리다 주립대학 전산학과(석사)
1992년 : 플로리다 주립대학 전산학과(박사)
1993년~현재 : 고려대학교 컴퓨터정보학과 교수

관심분야 : 인공지능, 지능 데이터베이스, 데이터마이닝, 인공신경망, 퍼지 이론
E-mail : dhpark@korea.ac.kr