DOI: 10.3745/KIPSTB.2010.17B.2.177

주성분 분석과 퍼지 연관을 이용한 문서군집 방법

박 선[†] · 안 동 언[††]

요 약

본 논문은 주성분 분석과 퍼지 연관을 이용한 새로운 문서군집 방법을 제안한다. 제안된 방법은 주성분 분석의 의미특징을 이용하여 군집 레이블과 군집의 대표 용어들을 선택하기 때문에 문서군집의 내부구조를 더 잘 표현할 수 있다. 또한 퍼지연관 값을 이용한 군집은 문서군집에 유사하지 않은 문서를 더 잘 구분함으로써 문서군집의 성능을 높일 수 있다. 실험결과 제안방법을 적용한 문서군집방법이 다른 문서군집 방법에 비하여 좋은 성능을 보인다.

키워드 : 문서군집, 주성분분석, 의미 특징, 퍼지 연관

Document Clustering Method using PCA and Fuzzy Association

Sun Park[†] · Dong Un An[††]

ABSTRACT

This paper proposes a new document clustering method using PCA and fuzzy association. The proposed method can represent an inherent structure of document clusters better since it select the cluster label and terms of representing cluster by semantic features based on PCA. Also it can improve the quality of document clustering because the clustered documents by using fuzzy association values distinguish well dissimilar documents in clusters. The experimental results demonstrate that the proposed method achieves better performance than other document clustering methods.

Keywords : Document Clustering, Principal Componet Analysis, Semantic Features, Fuzzy Association

1. 서 론

근래의 정보 검색 분야에는 사용자의 요구사항을 만족시키기 위하여 다양한 정보를 효율적으로 처리할 수 있는 문서의 범주화에 대해서 많은 연구가 있다. 문서의 범주화는 대량의 문서들을 각각의 문서의 특성 및 주제에 맞게 분류하는 것으로, 사전에 학습이 필요한 지도학습방법인 문서분류와 학습이 필요 없는 비지도학습 방법의 문서군집으로 구분할 수 있다[4].

전통적인 군집방법은 분할기반 방법, 계층적 기반 방법, 밀도기반 방법, 격자 기반 방법으로 분류 할 수 있다. 이들 대부분의 방법들은 거리 기반의 목적 함수를 사용하기 때문에 고차원의 객체들을 군집하는 것에는 비효율적이다. 이중에서 대표적인 군집방법으로는, 군집을 생성하는 방법에 따라서 k개의 군집을 임의로 정하여 군집을 확장해가는 비계층적 방법인 Kmeans와 군집간의 결합 방법에 의한 계층적 군집방법인 직접 군집방법이 있다[4, 8].

문서군집은 군집 알고리즘을 사용하여서 문서집합으로부터 유사한 특성을 가진 문서들의 그룹을 발견하는 것이다. 문서군집은 자료를 분석하는 중요한 기술로 자료의 조직화, 웹 검색결과의 브라우징, 다중문서 요약 등 다양한 정보검색 응용분야에 활용되는 중요한 방법[4, 10]으로, 정보통신 및 개인화 단말기의 발전으로 중요성이 더욱 부각되고 있다. 그러나 문서군집 방법의 근본적인 문제는 자료 집합의 분포나 내부구조, 사용자가 원하는 군집 형태 등이 군집결과에 중요한 영향을 미친다는 것이다[6]. 또한 용어 집합으로 구성된 문서와 이러한 문서들의 집합, 즉 이러한 고차원의 객체를 효율적으로 군집할 수 있는 방법의 요구가 증가하고 있다.

본 논문은 주성분분석과 퍼지연관을 이용하여 문서를 군집하는 새로운 문서군집 방법을 제안한다. 주성분분석(PCA, Principal Component Analysis)은 다차원적인 변수들을 축소, 요약하는 차원의 단순화와 더불어 일반적으로 서로 상관되어 있는 반응변수들 간의 복잡한 구조를 분석하는데 주

† 정 회 원 : 전북대학교 전기전자정보인력양성사업단 박사후과정
†† 종신회원 : 전북대학교 전기전자컴퓨터공학부 교수(교신저자)
논문접수 : 2009년 12월 7일
수 정 일 : 1차 2010년 1월 6일
심사완료 : 2010년 1월 7일

로 이용되는 방법이다[1, 7]. 퍼지 연관(Fuzzy Association)은 퍼지집합 이론을 사용하여 정보검색 과정의 모호성을 정형화하는 방법으로, 문서집합의 용어들이나 다른 색인어들 간의 관계를 인식할 수 있다[9, 16]. 제안 방법은 주성분분석을 이용하여 군집의 레이블과 군집의 대표 용어들을 선택하고, 선택한 대표 용어들과 문서에 포함된 용어의 퍼지 연관 관계를 이용하여 문서를 군집한다.

제안된 방법은 다음과 같은 장점을 갖는다. 첫째, 주성분 분석을 사용하여 군집을 대표할 수 있는 몇 개의 대표 용어들을 이용함으로서 고차원의 특징을 갖는 문서군집에 효율적이다. 둘째, 대표 용어와 문서내의 용어들 간의 퍼지 연관 관계를 사용하고, 이것은 군집에 더욱 관련 있는 용어를 포함한 문서들로 군집함으로써 문서군집의 정확도를 높일 수 있다. 마지막으로, 군집을 대표할 수 있는 군집 레이블을 추출함으로써 사용자는 쉽게 군집에 포함된 문서 집합의 특성을 파악할 수 있다.

본 논문의 구성은 다음과 같다. 제2장은 관련연구로 기존 문서군집방법, 주성분분석과 퍼지연관을, 제3장은 제안한 문서군집방법을, 제4장에서는 실험 및 평가에 대해 기술한다. 마지막으로 제5장에서는 결론을 맺는다.

2. 관련연구

2.1 문서군집

본장에서는 제안방법과 유사한 의미특징이나 군집의 레이블을 이용한 문서군집에 대한 기존연구에 대하여 알아본다. *Ji*외 저자들이 제안한 문서 군집방법으로 문서 군집 분석에 군집의 구성원에 대한 사전지식을 통합한 준지도 문서 군집 모델을 제안하였다. 이들의 방법은 사용자가 분류를 원하는 클러스터를 사전 지식으로 지정하고, 사전 지식을 군집의 비용 함수에 적용하여 문서를 군집한다[6]. *Basu*외 저자들은 준지도 Kmeans방법을 이용한 문서군집 방법을 제안하였다. 이들의 방법은 분류표시가 된 자료를 이용하여 초기 시드 클러스터를 생성하고, 분류표시가 된 자료로부터 제약사항을 생성하여 군집한다[3]. Li 이외 저자들은 문서군집을 위하여 각각의 군집과 관련된 군집의 하위 공간 구조의 명시적 모델링 방법을 이용한 ASI(Adaptive Subspace Iteration) 알고리즘을 제안하였다[11]. Wang과 Zhang은 문서군집을 위하여 지역 레이블의 예측과 전역 레이블의 조직화 방법을 이용한 CLGR(Clustering with Local and Global Regularization) 알고리즘을 제안하였다[14]. Xu 이외 저자들은 비음수 행렬 분해(NMF, Non-negative Matrix Factorization)의 의미특징을 이용하여 문서를 군집하는 방법을 제안하였다[15]. 본 논문의 저자들은 이전에 비음수 행렬 분해와 군집의 정제방법을 이용한 문서군집 방법을 제안하였다. 이 방법은 비음수 행렬 분해의 유사한 문서집합을 구분 하지 못하는 문제를 해결하기 위하여서 군집 후, 군집내의 유사도를 이용하여 재 군집하는 방법을 제안하였다[12].

2.2 주성분 분석

주성분 분석(PCA, Principal Component Analysis)이란 전체 자료의 공분산 행렬의 구조를 파악하여 고유값이 큰 고유벡터들의 축으로 자료의 축을 변환하여 주성분을 구하는 분석으로 다음과 같다[1, 7].

p개의 확률특징 $X_1, X_2, ..., X_p$를 원소로 하는 확률특징벡터 X가 평균 벡터 $\overline{x}$와 공분산 행렬 $S(p \times p)$를 갖고 있을 때 다음과 같다.

$$X = \begin{vmatrix} X_1 \\ X_2 \\ \dots \\ X_p \end{vmatrix}, \overline{x} = \begin{vmatrix} \overline{x_1} \\ \overline{x_2} \\ \dots \\ \overline{x_p} \end{vmatrix}, S = \begin{vmatrix} s_{11}\, s_{12} \dots s_{1p} \\ s_{21}\, s_{22} \dots s_{2p} \\ \dots \quad \dots \quad \dots \quad \dots \\ s_{p1}\, s_{p2} \dots s_{pp} \end{vmatrix}$$

여기서, $s_{ik} = \frac{1}{n}\sum_{j=1}^{n}(x_{ij} - \overline{x_i})(X_{kj} - \overline{x_k}) = s_{ki}$, X_i와 X_k의 공분산, $\overline{x_i} = \frac{1}{n}\sum_{j=1}^{n} X_{ij}$, X_i의 산술평균이다.

주성분 분석[1, 7]은 원래 특징벡터 X를 적절히 선형 변환 시켜 원본자료의 성질을 유지하면서 자료를 축소하고 해석하는데 사용한다. 이 선형변환은 X의 원소들 간의 상관구조 관계를 나타내는 S를 분석대상으로 하며, S는 $\overline{x}$ 의 값의 변화에 의한 영향을 받지 않는다. 우선 S의 p개의 고유값(eigen value) λ_j 들을 크기 순으로 배열하고 각각의 고유값에 대응되는 고유벡터(eigen vector) e_j의 짝들을 $(\lambda_1 \cdot e_1), ..., (\lambda_j \cdot e_j)$라 하고 λ_j들의 크기순으로 배열하면, $Se_j = \lambda_j \cdot e_j, j = 1,2,...,p$, $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_j$와 같은 관계가 있으며, 이를 다음과 같은 행렬 기호로 정의 할 수 있다.

$$SP = P\Lambda,\ S = P\Lambda P' \tag{1}$$

여기서 P는 p개의 고유벡터 e_i들로 구성된 크기 $p \times p$ 직교행렬이고, Λ는 λ_i를 i번째 대각원소, 그리고 모든 비대각 원소가 0인 크기 $p \times p$의 대각행렬, 그리고 P'는 P의 전치 행렬이다. 즉, $P = (e_1, e_2, ..., e_p)$, $\Lambda = Diagonal(\lambda_1, \lambda_2, ..., \lambda_p)$ 이와 같은 P를 이용하여 다음과 같은 X의 직교변환을 할 때, $\Phi' = P'X$ 이 변화에 의해 새로이 창조되는 벡터 $\Phi' = (\phi_1, \phi_2, ..., \phi_p)$를 X의 주성분이라 정의한다. 이때 j번째 고유값 λ_i에 대응하는 고유벡터 e_j의 원소들을 X와의 선형결합에서 가중계수로 사용한다. 즉, Φ'의 j번째 원소 ϕ_j를 X의 j번째 주성분이라고 하고, $e'_j = (e_{1j}, e_{2j}, ..., e_{pj}), j = 1,2,...,p$ 일때, 다음 식(2)와 같다.

$$\Phi_j = e_j X = e_{1j}X_1 + e_{2j}X_2 + ... + e_{pj}X_p = \sum_{i=1}^{p} e_{ij}X_j \tag{2}$$

2.3 퍼지 연관

이 장에서는 문서 군집에 사용되는 퍼지 연관 이론에 대하여 알아본다. 퍼지 이론은 다음과 같이 정의 된다[9, 16].

[정의 1] 두 유한 집합 $X = \{x_1, ..., x_u\}$와 $Y = \{ y_1, ..., y_v\}$ 사이의 퍼지 연관은 이진 퍼지 관계 $f:X\times Y\rightarrow[1,0]$으로 정의된다. 여기서 u와 v는 X와 Y 각각의 원소의 수이다.

[정의 2] 용어 색인 집합 $T = \{t_1, ..., t_u\}$와 문서 집합 $D = \{d_1, ..., d_v\}$가 주어질 때, t_i는 문서들의 퍼지 집합 $h(t_i)$에 의해 표현된다. 즉, $h(t_i) = \{F(t_i, d_j) \mid \forall\ d_i \in D\}$이다. 여기서 $F(t_i, d_j)$는 문서 d_j에서 t_i의 중요도의 정도이다.

[정의 3] 퍼지 연관 용어 관계 (fuzzy associated terms relation)는 문서집합 D에서 용어 t_i와 t_j가 동시 나타남을 기반으로 하여서 다음 식과 같이 정의 된다.

$$RT(t_i,t_j)=\frac{\sum_k \min(F(t_i,d_k),F(t_j,d_k))}{\sum_k \max(F(t_i,d_k),F(t_j,d_k))} \quad (3)$$

퍼지 연관 용어 관계는 동시에 존재하는 용어들에 기반하여 다음과 같이 단순화된다.

$$r_{i,j}=\frac{n_{i,j}}{n_i+n_j-n_{i,j}} \quad (4)$$

여기서, $r_{i,j}$는 용어 i와 j 사이의 퍼지 연관 용어 관계이다. $n_{i,j}$는 i번째 용어와 j번째 용어를 동시에 포함하는 문서들의 개수이며, n_i는 i번째 문서를 포함하는 문서의 개수이고, n_j는 j번째 문서를 포함하는 문서의 개수이다.

3. 주성분 분석과 퍼지 연관을 이용한 문서군집

이 논문에서 제안한 문서군집 과정은 다음 (그림 1)과 같이 전처리, 군집 대표용어 추출, 문서군집으로 구성된다. 전처리단계에서는 문서집합을 전처리하여 용어-문서 빈도행렬을 구성한다. 군집 대표용어 추출 단계에서는 주성분 분석을 이용하여 군집의 대표 용어와 군집 레이블을 추출한다. 문서군집단계에서는 군집 대표 용어와 퍼지연관을 이용하여 문서를 군집한다.

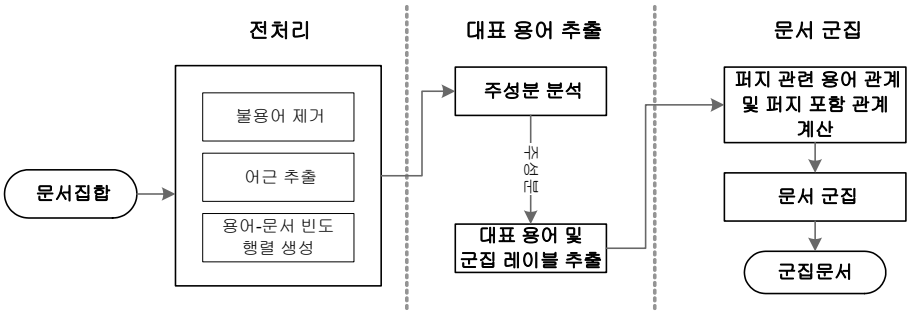
3.1 전처리

전처리 단계는 주어진 문서집합으로부터 불용어 제거, 어근추출, 용어빈도 벡터를 생성한다[5, 9]. 불용어 제거는 Rijsbergen의 불용어 목록[5]을 이용하고, 어근추출은 Porter의 어근추출 알고리즘[5]을 이용한다. 용어빈도 벡터 생성에 사용되는 벡터 $T_j = [\ t_{1j}, t_{2j}, \cdots , t_{nj}\ ]^T$는 j번째 문서의 용어빈도이다. 여기서 요소 t_{ij}는 j번째 문서에서 출현한 i번째 용어의 빈도이다.

3.2 주성분 분석을 이용한 군집의 대표 용어 추출

주성분 분석은 다차원 자료를 선형 변환시켜서 서로 상관되지 않는 새로운 인공 자료들인 주성분을 유도한다. 이 때 각 주성분이 보유하는 변이의 크기를 기준으로 그 중요도 순서를 생각 할 수 있다[1, 7]. 즉, 주성분 분석을 이용하면 군집내의 문서를 표현하기 위해서 문서내의 모든 용어들을 사용하는 대신에, 정보 손실을 최소화하면서 소수의 몇 개의 대표 용어로 문서를 포함하는 군집을 표현 할 수 있다. 즉, 가장 큰 주성분을 가지는 대표 용어는 군집을 구성하는 문서들의 특성을 잘 표현 할 수 있는 군집 레이블로 사용할 수 있고, 높은 값을 가지는 주성분과 일치하는 몇 개의 용어들로 문서군집의 특성을 나타낼 수 있다. 다음 예1은 주성분 분석에 의해 8개의 문서를 3개의 문서군집으로 만들기 위해서, 3개의 군집 레이블 및 군집의 대표 용어를 추출하는 예이다.

예1) 다음 <표 1>은 8개의 문서와 10개의 용어로 구성된 용어-문서 빈도 행렬이다. <표 2>는 <표 1>의 행렬에 주성분 분석을 수행하여 결과를 보여준다. 군집의 레이블은 각각의 주성분 열벡터에서 가장 큰 절대 값을 가지는 용어를 군집 레이블로 상위의 값을 가진 용어들을 대표 용어로 선택한다. <표 3>은 <표 2>로부터 3개의 군집 레이블 및 각각 2개의 대표 용어를 선택한 것이다. 즉, 첫 번째 주성분 열벡터에서 t_4(-0.7520), 두 번째 주성분 열벡터에서는 t_9(-0.6698),



(그림 1) 의미 특징과 군집연관을 이용한 문서군집

〈표 1〉 용어-문서 빈도 행렬

용어 \ 문서	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8
t_1	2	0	1	0	0	1	0	0
t_2	0	1	1	0	0	0	0	0
t_3	3	2	1	0	0	0	0	1
t_4	5	4	6	0	0	0	0	0
t_5	4	2	3	1	2	0	0	0
t_6	1	0	0	4	3	0	0	0
t_7	0	0	0	4	4	0	1	0
t_8	0	1	0	0	0	1	2	1
t_9	0	0	0	0	1	5	4	3
t_{10}	0	0	0	0	0	1	1	1

〈표 2〉 용어-문서 빈도행렬의 주성분 분석결과

용어 \ 주성분	p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8	p_9	p_{10}
t_1	-0.1715	-0.0089	-0.0840	0.5105	0.3939	0.1489	0.4404	-0.2472	-0.5005	0.1411
t_2	-0.0903	0.0349	-0.0246	-0.3819	-0.0223	0.1426	-0.4522	-0.5169	-0.3349	0.4893
t_3	-0.3044	0.0508	-0.0882	0.4271	-0.7309	0.0892	-0.2399	-0.0410	-0.2493	-0.2259
t_4	**-0.7520**	0.2460	-0.1291	-0.3789	0.1870	0.2807	0.0825	0.1702	0.0969	-0.2506
t_5	**-0.5069**	-0.0354	0.1275	0.2373	0.0750	-0.6598	-0.1010	-0.0318	0.2878	0.3664
t_6	-0.1570	-0.3931	***0.4836***	0.1826	-0.0968	0.5822	0.0058	-0.0945	0.4134	0.1829
t_7	-0.1201	**-0.5227**	**0.4931**	-0.2759	-0.0109	-0.2618	0.0502	0.1051	-0.4853	-0.2745
t_8	-0.0556	-0.1904	-0.2297	-0.3149	-0.4773	-0.0762	0.6823	-0.2086	0.1104	0.2321
t_9	-0.0703	**-0.6698**	-0.6280	0.0506	0.1855	0.0017	-0.2377	-0.0951	0.1256	-0.1835
t_{10}	-0.0135	-0.1470	-0.1656	0.0049	-0.0768	0.1490	-0.0490	0.7533	-0.2378	0.5442

〈표 3〉 군집의 대표 용어 및 군집 레이블

군집	군집 레이블	군집 대표 용어
C_1	t_4	t_4, t_5
C_2	t_9	t_9, t_7
C_3	t_7	t_7, t_6

세 번째 주성분 열벡터에서는 t_7(0.4931)을 각각 군집의 레이블로 선택한다. 이 것은 <표 2>의 주성분 열벡터의 순서는 고유값의 크기 순서로 되어 있고, 이는 고유값이 클수록 군집 내의 문서들을 더욱 잘 설명 할 수 있기 때문이다. 즉, 선택된 용어들이 주성분 분석의 특성상 보다 많은 문서에서 같이 출현하는 용어들을 군집의 레이블로 추출한다는 의미를 갖고 있다.

3.3 퍼지연관을 이용한 문서 군집

퍼지연관을 이용한 문서 군집 방법은 다음과 같다. 용어-문서 빈도행렬에 식(4)의 퍼지 연관 용어 관계를 이용하여 퍼지 연관 용어 관계 상관 행렬을 계산한다. 식(5)를 이용하여 퍼지 연관 용어 상관 행렬과 대표 용어들로부터 퍼지 포함 관계를 계산하다. 계산된 퍼지 포함관계를 이용하여 문서를 군집한다. 즉, 퍼지 포함관계 $\mu_{i,j}$ 가 최고값을 가지면, d_i 문서를 C_j 군집에 할당한다. 예2는 퍼지연관을 이용하여 문서를 군집하는 예이다.

각각의 문서들이 각각의 군집 집합에 포함되는 정도인 퍼지 포함관계[9, 16]는 다음과 같이 정의 된다.

$$\mu_{i,j} = \max_{\forall t_a \in d_i} \left[1 - \prod_{\forall t_b \in CT_j} (1 - r_{a,b}) \right] \quad (5)$$

여기서, $\mu_{i,j}$ 는 j번째 군집 C_j에 i번째 문서 d_i가 속하는 정도이며, $r_{a,b}$는 용어 $t_a \in d_i$와 용어 $t_b \in CT_j$ 사이의 퍼지관계이고, CT_j는 주성분 분석을 이용하여 선택한 j번째 군집의 대표용어 집합이다.

예2) 다음 <표 4>는 <표 1>에 식(4)를 이용하여 퍼지 연관 용어관계의 상관 행렬을 계산한 결과이다. <표 5>는 주성분 분석에 의해 선택된 <표 3>의 군집의 대표 용어와 <표 4>의 퍼지 연관 용어관계 상관행렬에 식(5)를 이용하여 문서를 군집한 결과이다.

〈표 4〉 〈표 1〉의 퍼지 연관 용어 관계의 상관 행렬

용어	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}
t_1	1	0.25	0.4	0.5	0.33	0.2	0	0	0.167	0.2
t_2	0.25	1	0.5	0.667	0.4	0	0	0.2	0	0
t_3	0.4	0.5	1	0.75	0.5	0.167	0	0.333	0.148	1.167
t_4	0.5	0.667	0.75	1	0.6	0.2	0	0.167	0	0
t_5	0.333	0.4	0.5	0.6	1	0.6	0	0.125	0.125	0
t_6	0.2	0	0.167	0.2	0.6	1	0.5	0	0.167	0
t_7	0	0	0	0	0	0.5	1	0.167	0.4	0.2
t_8	0	0.2	0.333	0.167	0.125	0	0.167	1	0.6	0.75
t_9	0.167	0	0.148	0	0.125	0.167	0.4	0.6	1	0.75
t_{10}	0.2	0	0.167	0	0	0	0.2	0.75	0.75	1

〈표 5〉 〈표 1〉의 문서군집 결과

군집	포함문서
$C_1(t_4)$	d_1, d_2, d_3
$C_2(t_9)$	d_6, d_7, d_8
$C_3(t_7)$	d_4, d_5

3.4 제안 알고리즘

다음은 본 논문에서 제안한 주성분 분석과 퍼지 연관을 이용한 문서군집 알고리즘이다. 1행에서는 문서집합을 전처리하며, 2행에서 9행까지는 전처리된 용어-문서 빈도 행렬에 주성분 분석을 수행하여 군집의 레이블과 군집의 대표 용어들을 선택한다. 10행에서 13행은 군집의 대표 용어들과 문서에 포함된 용어들과의 퍼지 연관관계를 이용하여서 문서를 군집한다.

Algorithm. 주성분 분석과 퍼지 연관을 이용한 문서 군집
Input: 용어-문서 빈도 행렬 A, 군집 개수 k, 대표 용어의 개수 r, k번째 군집의 대표 용어 집합 T^k, 전체문서의 개수 n,
Output: k개의 군집문서집합 C^k

1. 전처리 수행
2. 주성분 분석 수행
3. *Repeat*
4. 왼쪽 주성분 행렬 P로부터 c개의 주성분 열벡터 선택
5. *Repeat*
6. 상위 f개의 주성분 값에 일치하는 대표 용어 추출하여 T^C에 저장
7. 가장 큰 주성분 값에 일치하는 용어를 군집 레이블로 선택
8. *Until* f = 1, ..., r
9. Until c = 1, ..., k
10. A의 퍼지 연관 용어 관계 계산
11. *Repeat*
12. Select $d_j = \arg\max_{1 \le n \le j} \mu_{i,j}$ then include d_j in C^i
13. *Until* i = 1, ..., k

4. 실험 및 평가

제안방법에 대한 실험은 문서군집의 표준 성능평가 자료인 20 Newsgroups 문서자료[2] 중 일부를 무작위로 추출하여 실험하였다. 20 Newsgroups 평가자료는 뉴스 그룹이 20개가 있으며, 20개의 뉴스 그룹에는 총 20000 개의 문서를 포함하고 있다. 뉴스그룹은 컴퓨터 그래픽, 운영체제 윈도우, 컴퓨터 하드웨어, 종교, 의학, 정치 등 20개의 다양한 주제로 구성되어 있으며, 각 주제에 포함된 기사의 수는 같다. 다음 <표 6>은 실험에 사용된 평가자료의 특성표이다.

본 논문의 성능평가는 문서군집의 표준 평가척도 중 하나인 식(7)의 NMI(normalize mutual information)를 사용한다[11, 14, 15]. NMI의 상호정보이득은 두 개의 문서군집 C와 C'가 주어질 때 이들 간의 상호정보 $MI(C,C')$로 다음 식(6)과 같이 정의된다.

$$MI(C,C') = \sum_{c_i \in C, c'_j \in C'} p(c_i, c'_j) \cdot \log_2 \frac{p(c_i, c'_j)}{p(c_i) \cdot p(c'_j)} \quad (6)$$

$$NMI(C,C') = \frac{MI(C,C')}{\max(H(C), H(C'))} \quad (7)$$

〈표 6〉 20 Newsgroups 문서집합의 특성

문서집합의 속성	20 Newsgroups
총 문서 갯수	20000
사용문서 갯수	5400
클러스터 갯수	20
사용 클러스터 갯수	10
최대 클러스터의 문서 갯수	1000
최소 클러스터의 문서 갯수	100
중간 클러스터의 문서 갯수	500
평균 클러스터의 문서 갯수	540

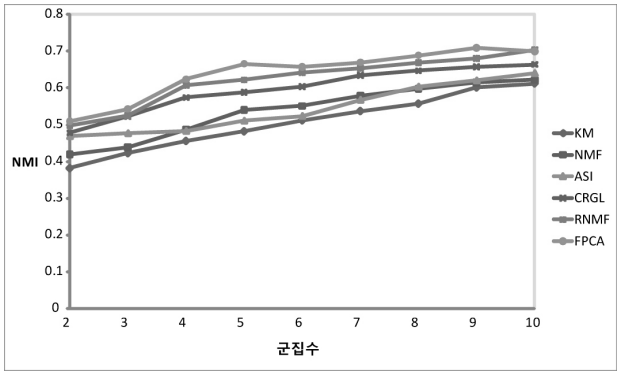
〈표 7〉 NMF를 이용한 제안방법의 비교결과

군집방법 \ k	2	3	4	5	6	7	8	9	10
KM	0.382	0.423	0.456	0.482	0.512	0.537	0.557	0.601	0.612
NMF	0.42	0.439	0.487	0.541	0.551	0.579	0.598	0.615	0.622
ASI	0.469	0.477	0.483	0.511	0.524	0.568	0.604	0.621	0.641
CRGL	0.478	0.523	0.574	0.589	0.604	0.635	0.647	0.658	0.664
RNMF	0.498	0.525	0.607	0.622	0.642	0.654	0.669	0.681	0.704
FPCA	0.51	0.543	0.624	0.666	0.657	0.67	0.689	0.71	0.7

여기서, $p(c_i)$와 $p(c'_j)$ 는 각각 군집 c_i 와 c'_j 에 문서집합의 문서가 포함될 확률이고, $p(c_i, c'_j)$는 문서집합의 문서가 동시에 군집 c_i 와 c'_j 에 포함될 확률이다. $H(C)$와 $H(C')$는 C와 C'의 엔트로피이다.

본 논문의 실험은 서로 다른 다섯 가지 문서군집방법과 제안방법간의 성능을 비교하였다. 성능 비교에서는 문서군집의 성능평가에 주로 사용되는 방법으로 군집의 개수를 2에서 10까지 증가하면서 NMI를 비교 하였다. <표 7>은 각각의 문서군집 방법 간의 비교 실험의 평균 NMI 결과이고, (그림 2)는 <표 7>을 도식화한 결과이다. 여기서, FPCA는 제안방법으로 주성분 분석과 퍼지연관을 이용한 문서군집방법이며, KM은 표준 Kmeans 군집을 이용한 문서군집방법[1, 8]이고, NMF는 비음수 행렬분해의 의미특징을 이용한 Xu의 문서군집방법이다[15]. 또한, ASI는 Li가 제안한 문서군집방법으로 반복 적응형 군집의 하위 공간 구조를 이용하고[11], CLRG는 Wang이 제안한 방법으로 군집의 지역과 전역의 정규화 속성을 이용하며[14], RNMF는 저자들의 이전 제안 방법으로 비음수 행렬분해와 군집의 정제방법을 이용한다[12].

(그림 2)에서 NMF군집방법이 KM군집방법보다 성능이 좋은 것은 KM에서의 단순한 유사도를 이용한 군집보다 NMF를 이용하여 자료의 내부구조를 반영하여 군집하는 것이 더 정확도에 영향을 미치는 것을 알 수 있다. 또한 군집의 하위 공간 구조의 속성을 사용하는 ASI나 군집의 전역 및 지역적 정규화 특성을 사용하는 CRGL보다는 군집의 내부 구조와 군집간의 유사도를 사용하는 RNMF가 좋은 군집 결과를 나타냄을 알 수 있다. 특히, FPCA는 군집의 각각의 특



(그림 2) 평균 NMI 비교결과

성을 나타내는 대표용어와 군집에 포함되는 문서의 용어들 간의 연관관계를 고려함으로써 가장 좋은 성능을 보인 것으로 생각된다. 즉, 군집의 내부특성과 이러한 내부특성을 대표하는 용어들에 가장 적합한 연관 관계를 가진 용어들을 포함한 문서들로 군집을 구성함을 알 수 있다.

5. 결 론

본 논문은 주성분 분석과 퍼지 연관을 이용하여 문서를 군집하는 새로운 문서군집방법을 제안하였다. 제안 방법은 주성분 분석을 사용하여 군집을 대표할 수 있는 몇 개의 대표 용어들로 선택함으로서 군집의 고차원적인 특성으로부터 몇몇 의미 특징을 갖는 용어로 저차원화함으로서 군집을 효율적으로 표현하였으며, 군집의 대표 용어와 가장 높은 연관관계를 갖는 용어를 포함하는 문서들로 군집함으로써 문서군집의 정확도를 높였다. 또한, 군집을 대표할 수 있는 군집 레이블을 추출함으로써 사용자는 쉽게 문서군집의 특성을 파악할 수 있다. 실험결과 제안방법의 FPCA의 평균 NMI가 KM군집 방법에 비하여서는 26.46%가, NMF군집 방법보다는 18.89%가, ASI군집 방법보다는 17.78%가, CRGL 군집 방법보다는 7.39%가, RNMF군집 방법보다는 2.98%가 각각 높음으로서 다른 문서군집 방법에 비하해서 더 좋은 성능을 나타냄을 알 수 있다.

앞으로 제안 방법의 성능 향상을 위하여 용어에 대한 가중치를 계산할 수 있는 다양한 정책과 비음수 행렬 분해를 이용한 군집방법에 적용할 수 있는 방법에 대하여 연구가 진행 되어야 할 것이다.

참 고 문 헌

[1] 이창범, 김민수, 이기호, 이귀상, 박혁로. “주성분 분석을 이용한 문서 주제어 추출”, 정보과학회논문지 : 소프트웨어 및 응용 제 29권 제 10호, 2002.

[2] The 20 newsgroups data set. http://people.csail.mit.edu/jrennie/20Newsgroups/, 2009.

[3] S. Basu, A.Banerjee, R. Mooney, “Semi-supervised Clustering by Seeding,” Proceeding of International Conference on Machine Learning (ICML), pp.19-26, 2002.

[4] S. Chakrabarti, “mining the web: Discovering Knowledge from Hypertext Data,” Morgan Kaufmann Publishers, 2003.

[5] W. B. Frankes, B. Y. Ricardo, “Information Retrieval : Data Structure & Algorithms,” Prentice-Hall, 1992.

[6] X. Ji, W. Xu, S. Zhu, “Document Clustering with Prior Knowledge”, Proceeding of Special Interest Group on Information Retrieval (SIGIR), pp.405-412, 2006.

[7] R. A. Johnson, D. W. Wichern, Applied Multivariate Statistical Analysis 5th ed., Prentice hall, 2007.

[8] J. Han, M. Kamber, “Second Edition Data Mining Concepts and Techniques,” Morgan Kaufman, 2006.

[9] C. Haruechaiyasak, M. L. Shyu, S. C. Chen, "Web Document Classification Based on Fuzzy Association," In proceedings of the 25th Annual International Computer Software and Applications Conference (COMPSAC'02), 2002.

[10] Y. Huang, T. M. Mitchell, “Text Clustering with Extended User Feedback”, Proceeding of Special Interest Group on Information Retrieval (SIGIR), pp.413-420, 2006.

[11] T. Li, S. Ma, M. Ogihara, "Document Clustering via Adaptive Subspace Iteration," In proceeding of SIGIR'04, pp.218-225, 2004.

[12] S. Park, D. U. An, B. R. Char, C. W. Kim, "Document Clustering with Cluster Refinement and Non-negative Matrix Factorization," In proceeding of ICONIP'09, pp.281-288, 2009.

[13] B. Y. Ricardo, R. N. Berthier, “Moden Information Retrieval,” ACM Press, 1999.

[14] F. Wang, C. Zhang, "Regularized Clustering for Documents," In proceeding of ACM SIGIR'07, 95-102, 2007.

[15] W. Xu, X. Liu, Y. Gon, “Document Clustering Based On Non-negative Matrix Factorization,” Proceeding of Special Interest Group on Information Retrieval (SIGIR), pp.267-274, 2003.

[16] L. A. Zadeh, "Fuzzy Sets, in Dubois, D., Prade, H. and Yager, R. R. editiors, Readings in Fuzzy Sets for Intelligent Systems," Morgan Kaufmann Publiishers, 1993.

[17] H. J. Zeng, Q. C. He, Z. Chen, W. Y. Ma, J. Ma, “Learning to Cluster Web Search Results,” Proceeding of Special Interest Group on Information Retrieval (SIGIR), 210-217, 2004.

박 선

e-mail : sunbak@jbnu.ac.kr

1996년 전주대학교 전자계산학과(이학사)

2001년 한남대학교 정보산업대학원 정보통신학과(공학석사)

2007년 인하대학교 컴퓨터정보공학과(공학박사)

2008년~2009년 호남대학교 컴퓨터공학과 전임강사

2009년~현 재 전북대학교 전기전자정보인력양성사업단 박사후과정

관심분야 : 정보검색, 데이터마이닝, 데이터베이스

안 동 언

e-mail : duan@jbnu.ac.kr

1981년 한양대학교 전자공학과(공학사)

1987년 KAIST 컴퓨터공학과(공학석사)

1995년 KAIST 컴퓨터공학과(공학박사)

1995년~현 재 전북대학교 전기전자컴퓨터공학부 교수

관심분야 : 자연어처리, 정보검색, 기계번역