한국항행학회 논문지 제14권 제1호 2010년 2월

비음수 행렬 분해와 퍼지 관계를 이용한 문서군집

Document Clustering using Non-negative Matrix Factorization and Fuzzy Relationship

박선*, 김경준**

Sun Park*, Kyung-Jun Kim **

요 약

본 논문은 비음수 행렬 분해와 퍼지 관계를 이용한 새로운 문서군집 방법을 제안한다. 제안된 방법은 비음수 행렬 분해된 의미특징을 이용하여 군집 레이블과 군집의 대표 용어들을 선택함으로서 문서군집의 내부구조를 더 잘 표현할 수 있으며, 퍼지 관계 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.

Abstract

This paper proposes a new document clustering method using NMF and fuzzy relationship. The proposed method can improve the quality of document clustering because the clustered documents by using fuzzy relation values between semantic features and terms to distinguish well dissimilar documents in clusters, the selected cluster label terms by using semantic features with NMF, which is used in document clustering, can represent an inherent structure of document set better. The experimental results demonstrate that the proposed method achieves better performance than other document clustering methods.

Key words : Document Clustering, Non-negative Matrix Factorization, Fuzzy Relationship, Semantic Features

I. 서 론

인터넷의 발전은 온라인 뉴스, 블러그, 이메일, 게시판 등 다양한 종류의 문자 정보를 증가 시키고 있다. 이러한 대량의 문자정보의 증가는 문서조직화, 자동 문서요약, 주제 추출, 정보 필터링 등 다양한 정보검색 방법의 기반 기술로 효율적인 문서군집 방법을 필요로 하고 있다[1, 2]. 이 때문에 사용자의 요구사항을 만족시키기 위하여 다양한 정보를 효율적으로 처리할 수 있는 문서의 범주화에 대한 연구를 많이 진행 하고 있다.

문서의 범주화는 학습과 평가가 필요한 문서분류와 학습이 필요 없는 문서군집으로 구분할 수 있다[3].

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 문서군집과 같은 고차원의 객

* BK21-전북 전자정보 고급인력양성사업단(CJEIT, Chonbuk National University)
** 한국과학기술원 전산학과 연구조교수
· 제1저자 (First Author) : 박선
· 투고일자 : 2009년 12월 28일
· 심사(수정)일자 : 2009년 12월 28일 (수정일자 : 2010년 4월 1일)
· 게재일자 : 2010년 4월 30일

체들을 군집하는 것에는 비효율적이다[4].

문서군집은 문서집합으로부터 유사한 특성을 가진 문서들의 그룹을 발견하는 것이다. 문서군집은 다양한 정보검색 응용분야에 활용되는 중요한 방법[2]으로, 정보화 기술의 발전으로 중요성이 더욱 부각되고 있다. 그러나 문서군집 방법의 근본적인 문제는 자료 집합의 분포나 내부구조, 사용자가 원하는 군집 형태 등이 군집결과에 중요한 영향을 미친다는 것이다[5]. 또한 점차 용량이 증가하는 문서들의 고차원 객체를 효율적으로 군집할 수 있는 기술의 필요성이 증가 하고 있다.

본 논문은 비음수 행렬분해와 퍼지관계를 이용하여 문서를 군집하는 새로운 문서군집 방법을 제안한다. 비음수 행렬 분해는(NMF, non-negative matrix factorization) Lee와 Seung이 제안한 방법으로 인간이 객체를 인식할 때 객체의 부분정보의 조합으로 인식하는 것에 착안하여, 객체정보를 기초특징(base feature)과 부호특징(encoding feature)으로 나누어 부분정보(part-base)로 표현한다[6]. 퍼지 관계(Fuzzy Relationship)는 퍼지집합 이론을 사용하여 정보검색 과정의 모호성을 정형화하는 방법으로, 문서집합의 용어들이나 다른 색인어들 간의 관계를 인식할 수 있다[7]. 제안 방법은 비음수행렬분해를 이용하여 군집의 레이블과 군집의 대표 용어들을 선택하고, 선택한 대표 용어들과 문서에 포함된 용어의 퍼지 관계를 이용하여 문서를 군집한다.

제안된 방법은 다음과 같은 장점을 갖는다. 첫째, 비음수 행렬분해를 이용하여 군집을 대표할 수 있는 몇 개의 대표 용어들을 추출함으로서 고차원의 특징인 문서군집에 효율적이다. 둘째, 대표 용어와 문서 내의 용어들 간의 퍼지 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, 비음수 행렬분해와 퍼지관계를, 제3장은 제안한 문서군집방법을, 제4장에서는 실험 및 평가에 대해 기술한다. 마지막으로 제5장에서는 결론을 맺는다.

Ⅱ. 관련연구

2-1 문서군집

본장에서는 제안방법과 유사한 의미특징이나 군집의 레이블을 이용한 문서군집 대한 기존연구에 대하여 알아본다.

Xu 이외 저자들은 비음수 행렬 분해(NMF, Non-negative Matrix Factorization)의 의미특징을 이용하여 문서를 군집하는 방법을 제안하였다[8]. Xu의 문서군집방법은 단순히 의미 특징의 가중치인 의미변수의 크기에 따라서 문서를 군집함으로써, 원본 문서의 구조에 큰 영향을 받는다. 즉, 원본 문서가 극단적으로 다른 내용을 포함하던지 비슷한 구조를 가지면서 다른 내용의 문서들로 구성되면, 이들을 분류하지 못하던지, 아니면 하나의 군집으로 문서를 분류하는 문제점을 가지고 있다.

Ji외 저자들은 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[5].

Basu외 저자들은 준지도 Kmeans방법을 이용한 문서군집 방법을 제안하였다. 이들의 방법은 분류표시가 된 자료를 이용하여 초기 시드 클러스터를 생성하고, 분류표시가 된 자료로부터 제약사항을 생성하여 군집한다[9].

Li 이외 저자들은 문서군집을 위하여 각각의 군집과 관련된 군집의 하위 공간 구조의 명시적 모델링 방법을 이용한 ASI(Adaptive Subspace Iteration) 알고리즘을 제안하였다[10].

Wang과 Zhang은 문서군집을 위하여 지역 레이블의 예측과 전역 레이블의 조직화 방법을 이용한 CLGR(Clustering with Local and Global Regularization) 알고리즘을 제안하였다[11].

본 논문의 저자들은 이전에 비음수 행렬 분해와 군집의 정제방법을 이용한 문서군집 방법을 제안하

였다. 이 방법은 비음수 행렬 분해의 유사한 문서집합을 구분 하지 못하는 문제를 해결하기 위하여서 군집 후, 군집내의 유사도를 이용하여 재 군집하는 방법을 제안하였다[12]. 또한 저자들은 주성분 분석과 퍼지연관을 이용한 문서군집 방법을 제안하였다[13].

2-2 비음수 행렬 분해

비음수 행렬 분해는 주어진 비음수 행렬로부터 비음수의 인수를 찾는 행렬분해 알고리즘이다[6]. 비음수 행렬 분해 알고리즘은 식(1)의 목표함수 J 가 0에 가깝게 수렴 할 때까지 식(2)과 식(3)을 이용하여 행렬 W와 H의 값을 동시에 갱신한다.

$$J = \| A - WH \|^2 \quad (1)$$

식(1)의 목적은 행렬 A를 비음수 m×r 행렬 W와 비음수 r×n 행렬 H로 분해하는 것이다. 여기서, A는 m개의 용어와 n개의 문장으로 이루어진 m×n 행렬이고, r은 의미특징의 개수이다.

$$H_{rj} \leftarrow H_{rj} \frac{(W^T V)_{rj}}{(W^T WH)_{rj}} \quad (2)$$

$$W_{ir} \leftarrow W_{ir} \frac{(VH^T)_{ir}}{(WHH^T)_{ir}} \quad (3)$$

본 논문에서 행렬 X의 j번째 열벡터는 X*j로, i번째 행벡터는 Xi*로, i번째 행과 j번째 열의 원소는 Xij표시한다.

행렬 A의 j번째 열벡터 A*j는 행렬 W의 l번째 열벡터 W*l와 행렬 H의 요소 Hkj가 선형조합을 이루며 식(4)과 같다.

$$A_{*j} = \sum_{l=1}^{r} H_{kj} W_{*l} \quad (4)$$

예1) 다음은 식(1)을 이용하여 A행렬을 W 와 H 행렬로 분해 한 예이다. r = 2, 수렴할 반복수는 50 이고, 수렴 허용오차가 0.001이다. W 와 H 행렬의 초기값은 각각 0.5이다.

$$\underset{A}{\begin{bmatrix}1&2&3\\4&5&6\\7&8&9\\10&11&12\end{bmatrix}} \approx \underset{W}{\begin{bmatrix}0.15&1.60\\0.66&0.97\\1.15&0.57\\1.61&0.41\end{bmatrix}} \times \underset{H}{\begin{bmatrix}6.11&6.68&7.18\\0.09&0.60&1.22\end{bmatrix}} = \underset{\tilde{A}}{\begin{bmatrix}1.05&1.94&3.03\\4.12&4.98&5.93\\7.07&8.01&8.95\\9.90&11.01&12.08\end{bmatrix}}$$

예2) 다음은 예제(1)의 열벡터 A*3 를 식(4)와 같이 의미특징벡터와 의미변수의 선형조합으로 나타낸 예이다.

$$\underset{A_{*3}}{\begin{bmatrix}3\\6\\9\\12\end{bmatrix}} \approx 7.2 \times \underset{H_{13}}{\begin{bmatrix}0.15\\0.66\\1.15\\1.61\end{bmatrix}} + 1.2 \times \underset{H_{23}}{\begin{bmatrix}1.60\\0.97\\0.57\\0.41\end{bmatrix}} = \underset{\tilde{A}_{*3}}{\begin{bmatrix}3.03\\5.93\\8.95\\12.08\end{bmatrix}}$$

W_{*1} W_{*2}

2-3 퍼지 관계

이 장에서는 문서 군집에 사용되는 퍼지 관계 이론에 대하여 알아본다. 퍼지 이론은 다음과 같이 정의 된다[7].

(정의 1) 두 유한 집합 X = {x1, ..., xu}와 Y = {y1, ..., yv} 사이의 퍼지 관계는 이진 퍼지 관계 f:X×Y →[1,0]으로 정의된다. 여기서 u와 v는 X와 Y 각각의 원소의 수이다.

(정의 2) 용어 색인 집합 T = {t1, ..., tu}와 문서 집합 D = {d1, ..., dv}가 주어질 때, ti는 문서들의 퍼지 집합 h(ti)에 의해 표현된다. 즉, h(ti) = {F(ti, dj) | ∀ di ∈D}이다. 여기서 F(ti, dj)는 문서 dj에서 ti의 중요도의 정도이다.

(정의 3) 퍼지 관련 용어 관계 (fuzzy related terms relation)는 문서집합 D에서 용어 ti와 tj가 동시 나타남을 기반으로 하여서 다음 식과 같이 정의 된다.

$$RT(t_i,t_j)=\frac{\sum_k \min(F(t_i,d_k),F(t_j,d_k))}{\sum_k \max(F(t_i,d_k),F(t_j,d_k))} \quad (5)$$

퍼지 관련 용어 관계는 동시에 존재하는 용어들에 이용하여 다음과 같이 단순화된다.

$$r_{i,j}=\frac{n_{i,j}}{n_i+n_j-n_{i,j}} \quad (6)$$

여기서, ri,j는 용어 i와 j 사이의 퍼지 관련 용어 관계이다. ni,j는 i번째 용어와 j번째 용어를 동시에 포함하는 문서들의 개수이며, ni는 i번째 문서를 포함하는 문서의 개수이고, nj는 j번째 문서를 포함하는 문서의 개수이다.

Ⅲ. 제안방법

본 논문에서 제안한 문서군집 과정은 다음 그림1과 같이 전처리, 군집 대표용어 추출, 문서군집으로 구성된다. 전처리단계에서는 문서집합을 전처리하여서 용어-문서 빈도행렬을 구성한다. 군집 대표용어 추출 단계에서는 비음수 행렬분해를 이용하여 군집의 대표 용어와 군집 레이블을 추출한다. 문서군집단계에서는 군집 대표 용어와 퍼지관계를 이용하여 문서를 군집한다.

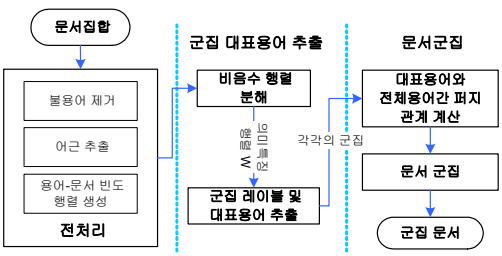

그림 1. 퍼지관계와 의미특징을 이용한 문서군집
Fig. 1. Document clustering using semantic features and fuzzy relationship.

3-1 전처리

전처리 단계는 주어진 문서집합으로부터 불용어 제거, 어근추출, 용어빈도 벡터를 생성한다[14]. 불용어 제거는 Rijsbergen의 불용어 목록[14]을 이용하고, 어근추출은 Porter의 어근추출 알고리즘[14]을 이용한다. 용어빈도 벡터 생성에 사용되는 벡터 Ti = [t1i, t2i, … , tni]T는 i번째 문장의 용어빈도이다. 여기서 요소 tij는 j번째 문서에서 출현한 i번째 용어의 빈도이다.

3-2 군집의 대표 용어 추출

비음수 행렬분해에 의한 의미특징은 원본자료의 부분정보를 나타낼 수 있고, 다시 이러한 부분정보의 조합으로 원본자료를 표시할 수 있다. 즉, 이러한 특성을 갖는 의미특징을 이용하면 군집을 구성하는 문서집합의 특성을 몇몇 특정 의미특징과 대응되는 용어들로 나타낼 수 있다. 이는 정보 손실을 최소화하면서 소수의 몇 개의 대표 용어로 문서 군집을 표현할 수 있다. 특히 가장 높은 값을 가지는 의미특징과 대응되는 용어는 군집을 구성하는 문서들의 특성을 잘 표현 할 수 있는 군집 레이블로 사용할 수 있다.

다음은 비음수 행렬분해와 퍼지 관계를 이용하여 7개의 문서를 3개의 군집으로 분류하는 예이다. 표1은 7개의 문서를 전처리하여서 6개의 용어와 7개의 문서로 구성된 용어-문서 빈도 행렬이다. 표2는 표1을 비음수 행렬분해 하여서 얻어진 의미특징 행렬이다. 표3은 표2를 이용하여 추출한 각각의 군집에 대한 대표 용어이다. 표3은 표2의 각 군집 r 별로 상위 2개의 의미특징 값을 가지는 의미특징과 대응되는 용어들을 추출하였다.

표 1. 용어-문서 빈도행렬
Table 1. Term-document frequency matrix.

문서 / 용어	*d1*	*d2*	*d3*	*d4*	*d5*	*d6*	*d7*
t1	2	1	0	0	0	0	0
t2	1	2	0	0	0	0	1
t3	3	1	0	0	1	1	0
t4	0	0	1	2	1	1	1
t5	0	0	1	1	1	1	1
t6	0	0	1	1	0	0	0

표 2. 의미 특징 행렬 W
Table 2. Semantic feature matrix W.

	r1	*r2*	*r3*
t1	0	**1.8455**	**0.4791**
t2	0	0	**2.4913**
t3	0.2884	**2.6364**	0
t4	**2.8135**	0	0.0048
t5	**2.1483**	0.0834	0.0341
t6	1.1197	0	0

표 3. 군집의 대표용어
Table 3. Cluster label terms.

	cluster label terms
C1	*t4, t5*
C2	*t3, t1*
C3	*t2, t1*

3-3 퍼지관계를 이용한 문서군집

퍼지관계를 이용한 문서 군집 방법은 다음과 같다. 용어-문서 빈도행렬에 식(6)의 퍼지 관련 용어 관계를 이용하여 퍼지 관련 용어 관계 상관 행렬을 계산한다. 식(7)을 이용하여 퍼지 관련 용어 상관 행렬과 대표 용어들로부터 퍼지 포함 관계를 계산하다. 계산된 퍼지 포함관계를 이용하여 문서를 군집한다. 즉, 퍼지 포함관계 μi,j가 최댓값을 가지면, di 문서를 Cj 군집에 할당한다.

각각의 문서들이 각각의 군집 집합에 포함되는 정도인 퍼지 포함관계[7]는 다음과 같이 정의 된다.

$$\mu_{i,j} = \max_{\forall t_a \in d_i} \left[1 - \prod_{\forall t_b \in CT_j} (1 - r_{a,b}) \right] \qquad (7)$$

여기서, μi,j 는 j번째 군집 Cj에 i번째 문서 di가 속하는 정도이며, ra,b는 용어 tj∈dj와 용어 tb∈CTj 사이의 퍼지관계이고, CTj는 비음수 행렬 분해를 이용하여 선택한 j번째 군집의 대표용어 집합이다.

다음 표4와 표6은 비음수 행렬 분해를 이용하여 추출한 군집의 대표용어와 퍼지관계를 이용하여서 문서를 군집하는 예를 보여준다. 표4는 표1의 용어-문서 빈도행렬과 식(6)의 퍼지 관련 용어 관계식을 이용하여서 계산한 용어의 상관행렬이다. 표5는 표4의 용어의 상광행렬에 식(7)의 퍼지 포함관계를 이용하여 문서를 군집한 결과이다.

표 4. 용어의 상관행렬
Table 4. Term correlation matrix.

용어	*t1*	*t2*	*t3*	*t4*	*t5*	*t6*
t1	1	0.5	0.5	0	0	0
t2	0.5	1	0.4	0.17	0.14	0
t3	0.5	0.4	1	0.29	0.29	0
t4	0	0.17	0.29	1	1	0.4
t5	0	0.14	0.29	1	1	0.4
t6	0	0	0	0.4	0.4	1

표 5. 문서 군집 결과
Table 5. Result of document clustering.

군집	documents
C1	*d4, d5, d6, d7*
C2	*d1*
C3	*d2*

3-4 제안 알고리즘

다음은 비음수 행렬분해와 퍼지관계를 이용하여 문서를 군집하는 제안 알고리즘이다.

Algorithm. 의미특징과 퍼지 관계를 이용한 문서 군집
Input: 용어빈도행렬 A, 군집의 개수r, 전체 문서의 개수 n, 전체 용어의 개수 k,군집 대표 용어의 개수 t, 군집 대표 용어 집합 CT
Output: 군집 문서 집합 C^k

1. 전처리 수행.
2. 비음수 행렬 분해.
3. repeat
4. repeat
5. Select $p = \underset{1 \le n \le k}{\arg max}\{W_{ic}\}$
6. Put p와 일치하는 군집 대표 용어 into CT
7. until p=1, ..., t
8. until c=1, ..., r
9. 퍼지 관련용어 관계 계산
10. repeat
11. Select $d_j = \underset{1 \le n \le j}{\arg max}\{\mu_{i,j}\}$
then include d_j in cluster C^i
12. until i=1, ..., k

위 알고리즘의 2번에서 8번까지는 비음수 행렬 분해의 의미특징을 이용하여 군집의 대표 용어를 추출한다. 알고리즘의 9에서 12번에서는 퍼지 관계와 군집의 대표 용어를 이용하여 문서를 군집한다.

Ⅳ. 실험 및 평가

제안방법에 대한 실험은 문서군집의 표준 성능평가 자료인 20 Newsgroups 문서자료[15] 중 일부를 무작위로 추출하여 실험하였다. 20 Newsgroups 평가 자료는 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 표6은 실험에 사용된 평가 자료의 특성표이다.

표 6. 20 Newsgroups 문서집합의 특성
Table 6. Property of 20 Newsgroups document set.

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

본 논문의 성능평가는 문서군집의 표준 평가척도 중 하나인 식(9)의 NMI(normalize mutual information)를 사용한다[16]. NMI의 상호정보이득은 두 개의 문서군집 C와 C'가 주어질 때 이들 간의 상호정보 MI(C,C')로 다음 식(8)과 같이 정의된다.

$$MI(C,C') = \sum_{c_i \in C, c'_j \in C'} p(c_i, c'_j) \cdot \log_2 \frac{p(c_i, c'_j)}{p(c_i) \cdot p(c'_j)} \quad (8)$$

$$NMI(C,C') = \frac{MI(C,C')}{\max(H(C), H(C'))} \quad (9)$$

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

본 논문의 실험은 서로 다른 여섯 가지 문서군집 방법과 제안방법간의 NMI를 군집의 개수를 2에서 10까지 증가하면서 비교 하였다. 그림2는 각각의 문서군집 방법 간의 비교 실험의 평균 NMI 결과이다. 여기서, FNMF는 제안방법으로 비음수 행렬 분해와 퍼지 관계를 이용한 문서군집방법이다. KM은 표준 Kmeans 군집을 이용한 문서군집방법[3]이고, NMF는 비음수 행렬분해의 의미특징을 이용한 Xu의 문서군집방법이다[8]. 또한, ASI는 Li가 제안한 문서군집방법으로 반복 적응형 군집의 하위 공간 구조를 이용하

고[10], CLRG는 Wang이 제안한 방법으로 군집의 지역과 전역의 정규화 속성을 이용하며[11], RNMF와 FLSA는 저자들의 이전 제안 방법으로 각각 비음수 행렬분해와 군집의 정제방법[12]과 주성분 분석과 퍼지연관을 이용한다[13].

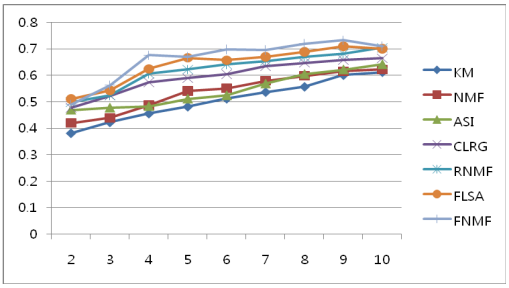

그림 2. 평균 NMF 비교 결과
Fig. 2. Result of comparison of average NMI

그림2의 실험결과 제안방법의 FNMF의 평균 NMI가 KM군집 방법에 비하여서는 30.53%가, NMF군집 방법보다는 22.73%가, ASI군집 방법보다는 22.58%가, CRGL군집 방법보다는 10.85%가, RNMF군집 방법보다는 6.30%가, FLSA군집 방법보다는 3.22%가 각각 높음으로서 다른 문서군집 방법에 비해서 더 좋은 성능을 나타냄을 알 수 있다.

그림2의 실험결과를 분석하면 NMF군집방법이 KM군집방법 보다 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 RNMF와 군집의 내부 구조와 문서의 용어들 간의 퍼지 연관을 이용한 FLSA가 좋은 군집 결과를 나타냄을 알 수 있다. 특히, FNMF 군집의 각각의 특성을 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.

V. 결 론

본 논문은 비음수 행렬 분해와 퍼지 관계를 이용하여 문서를 군집하는 새로운 문서군집방법을 제안하였다. 제안 방법은 비음수 행렬 분해를 사용하여 군집을 대표할 수 있는 몇 개의 대표 용어들로 선택함으로서 군집의 고차원적인 특성으로 부터 몇몇 의미 특징을 갖는 용어로 저차원화 함으로서 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.

참 고 문 헌

[1] Hu, T., Xiong, H., Zhou, W., Sung, S. Y., Luo, H.: Hypergraph Partitioning for Document Clustering: A Unified Clique Perspective. *In proceeding of SIGIR'08*, 871-872 (2008)

[2] Ricardo, B. Y., Berthier, R. N.: Moden Information Retrieval, ACM Press (1999)

[3] Chakrabarti, S.: Mining the Web : Discovering Knowledge from Hypertext Data. Morgan Kaufmann (2003)

[4] Han. J., Kamber., M.: Data Mining Concepts and Techniques Second Edition. Morgan Kaufmann (2006)

[5] Ji, X., Xu, W., Zhu, S.: Document Clustering with Prior Knowledge. *In proceeding of SIGIR'06*, 405-412 (2006)

[6] Lee, D. D., Seung, H. S.: Learning the parts of objects by non-negative matrix factorization. Nature, 401:788-791 (1999)

[8] Xu, W., Liu, X. and Gong, Y.: Document clustering based on non-negative matrix factorization. In proceeding of ACM SIGIR'03 (2003)

[7] Haruechaiyasak, C., Shyu, M. L., Chen, S. C.: Web Document Classification Based on Fuzzy Association. In proceedings of the 25th Annual International Computer Software and Applications

Conference (COMPSAC'02) (2002)

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

[10] Li, T., Ma, S., Ogihara, M.: Document Clustering via Adaptive Subspace Iteration. *In proceeding of SIGIR'04*, 218-225 (2004)

[11] Wang, F., Zhang, C.: Regularized Clustering for Documents. *In proceeding of ACM SIGIR'07*, 95-102 (2007)

[12] Park, S., An, D. U., Char, B. R., Kim, C. W.: Document Clustering with Cluster Refinement and Non-negative Matrix Factorizaion. *In proceeding of ICONIP'09*, (2009)

[13] Park, S., An, D. U., Cha, B. R., Kim, C. W.: Document Clustering with Semantic Feature and Fuzzy Association. *In proceeding of ICISTM'10*, (2010)

[14] Frankes, W. B. Ricardo, B. Y.: Information Retrieval, Data Structure & Algorithms. Prentice-Hall (1992)

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

[16] Xu, W., Gong, Y.: Document Clustering by Concept Factorization. *In proceeding of SIGIR'04*, 202-209 (2004)

박 선 (朴仙)

1996년 2월 : 전주대학교 전자계산학과(이학사)
2001년 8월 : 한남대학교 정보산업대학원 정보통신학과(공학석사)
2007년 8월 : 인하대학교 컴퓨터정보공학과(공학박사)
2008~2009년 8월 : 호남대학교 컴퓨터공학과 전임강사,
2009년 9월~현재 : 전북대학교 전기전자정보인력양성사업단 박사후과정
관심분야 : 정보검색, 데이터마이닝, 데이터베이스

김 경 준 (金京浚)

1996년 2월 : 경일대학교 컴퓨터공학과 (공학사)
1999년 8월 : 경북대학교 컴퓨터공학전공 (공학석사)
2005년 2월 : 경북대학교 정보통신학과 (공학박사)
2005 3월 : 경북대학교 컴퓨터공학과 PostDoc. 연구원
2005년 9월 ~ 2006년8월 : 대구대학교 정보통신공학부 초빙교수
2006년 9월 ~ 2009년 5월 : 호남대학교 전파이동통신공학과 전임강사
2009년 12월 ~ 현재 : 한국과학기술원 전산학과 연구조교수
관심분야 : 무선채널할당기법, 토폴로지제어, 지능형전송시스템(ITS), 무선MAC 프로토콜, 네트워크보안