韓國 컴퓨터情報學會 論文誌
第13卷 第4號, 2008. 7.

2008-13-4-1-2

NMF 기반의 용어 가중치 재산정을 이용한 문서군집

이 주 홍*, 박 선**

Document Clustering using Term reweighting based on NMF

Ju-hong Lee*, Sun Park**

요 약

문서군집은 정보검색의 많은 응용분야에 사용되는 중요한 문서 분석 방법이다. 본 논문은 비음수 행렬 분해(NMF, non-negative matrix factorization)를 기반한 용어 가중치 재산정 방법을 이용하여서 사용자의 요구에 적합한 군집결과를 얻도록 하는 새로운 군집모델을 제안한다. 제안된 모델은 군집형태에 대한 사용자 요구와 기계에 의한 군집 형태의 차이를 최소화하기 위하여 사용자 피드백에 의한 가중치가 재계산된 용어를 이용한다. 또한 제안방법은 용어의 가중치 재계산과 문서군집에 문서집합의 내부구조를 나타내는 의미특징행렬과 의미변수행렬 이용하여 문서군집의 성능을 높일 수 있다. 실험결과 제안방법을 적용한 문서군집방법이 적용하지 않은 문서군 방법에 비하여 좋은 성능을 보인다.

Abstract

Document clustering is an important method for document analysis and is used in many different information retrieval applications. This paper proposes a new document clustering model using the re-weighted term based NMF(non-negative matrix factorization) to cluster documents relevant to a user's requirement. The proposed model uses the re-weighted term by using user feedback to reduce the gap between the user's requirement for document classification and the document clusters by means of machine. The proposed method can improve the quality of document clustering because the re-weighted terms, the semantic feature matrix and the semantic variable matrix, which is used in document clustering, can represent an inherent structure of document set more well. The experimental results demonstrate appling the proposed method to document clustering methods achieves better performance than documents clustering methods.

▸ Keyword : 문서군집(Document Clustering), 사용자 피드백(User Feedback), 비음수 행렬 분해(Non-negative Matrix Factorization), 용어의 재가중치(Re-weighted Term)

• 제1저자 : 이주홍
• 접수일 : 2008. 4. 1, 심사일 : 2008. 6. 5, 심사완료일 : 2008. 7. 25.
* 인하대학교 컴퓨터정보공학과교수 **호남대학교 컴퓨터공학과교수

Ⅰ. 서 론

문서군집은 군집 알고리즘에 의하여서 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 방법이다[2, 6]. 그러나 문서군집 방법의 근본적인 문제는 자료 집합의 분포나 내부구조, 사용자가 원하는 군집 형태 등이 군집결과에 중요한 영향을 미친다는 것이다[4].

최근에는 이러한 문제를 해결하기 위하여 사용자에게 다양한 유형의 사용자 피드백을 적용하여 매개변수(parameters)를 조정하여서 군집의 질을 향상시키고 있다[4, 12].

본 논문은 NMF에 기반한 용어의 가중치를 지도 학습방법으로 계산하고, 계산된 용어의 가중치를 군집할 문서에 적용하여 문서를 군집하는 새로운 문서군집 모델을 제안한다. 본 논문에서 사용되는 사용자 피드백은 일반적인 질의를 확장하여 용어를 재작성하는 연관피드백[9]과는 달리, 사용자가 원하는 결과의 군집에 포함되는 문서를 추출하여 정확한 군집이 될 수 있도록 용어에 대한 가중치를 재계산하는 새로운 방법이다. NMF(non-negative matrix factorization, 비음수 행렬 분해)는 Lee와 Seung이 제안한 방법으로 인간이 객체를 인식할 때 객체의 부분정보의 조합으로 인식하는 것에 착안하여, 객체정보를 기초특징(base feature)과 부호특징(encoding feature)로 나누어 부분정보(part-base)로 표현한다. 이러한 부분정보의 조합으로 전체 객체를 표현하는 방법은 대량의 정보를 효율적으로 표현 할 수 있는 방법이다[7, 8]. Xu등은 NMF를 이용하여 문서를 군집하는 방법을 제안하였다[11].

제안된 모델은 다음과 같다. 문서집합으로부터 학습에 사용할 문서를 사용자가 추출하고 군집을 결정한다. 추출문서집합을 전처리 하여서 벡터모델로 표시하고[9], NMF를 이용하여 추출문서집합을 비음수 의미특징 행렬과 비음수 의미변수 행렬로 분해한다[7]. 추출문서집합 내의 문서 벡터들은 의미특징벡터에 가중치인 의미변수를 곱한 값의 선형합으로 표시된다. 의미특징 벡터는 문서의 내부특징을 나타내며, 의미변수는 문서 내에서 의미특징의 중요도를 나타낸다. 비음수 의미변수 행렬에 제안된 가중치 재계산 방법을 적용하여 용어의 가중치를 재계산한다. 계산된 용어 가중치를 문서집합 전체에 적용한 후 전통적인 문서군집 방법을 이용하여 문서를 군집한다.

제안된 모델은 다음과 같은 장점을 갖는다. 첫째, 의미특징과 의미변수를 사용하여 군집의 내부구조와 의미특징의 분포를 쉽게 파악할 수 있고, 이를 이용하여 용어의 가중치를 쉽게 재계산함으로써 문서군집의 정확도를 높일 수 있다. 둘째, 제안된 모델의 용어의 가중치 계산은 NMF에 기반한 문서군집방법뿐만 아니라 단수값 분해(SVD; singular value decomposition)도 적용하여 계산 할 수 있다. 마지막으로 계산된 용어의 가중치는 Kmeans와 같은 전통적인 문서군집 방법에 적용하여 군집의 성능을 높일 수 있다.

본 논문의 구성은 다음과 같다. 제2장은 관련연구를, 제3장은 제안한 문서군집방법을, 제4장에서는 실험 및 평가에 대해 기술한다. 마지막으로 제5장에서는 결론을 맺는다.

Ⅱ. 관련연구

문서군집 및 분류에대한 기존연구는 다음과 같다. (X. Ji, et al.) [4]의 논문에서 문서 군집 분석에 군집의 구성원에 대한 사전지식을 통합한 준지도 문서 군집 모델(semi-supervised document clustering model)을 제안하였다. 이들의 방법은 사용자가 분류를 원하는 클러스터를 사전 지식으로 지정하고, 사전 지식을 군집의 비용 함수(cost function)에 적용하여 문서를 군집한다.

(S. Basu, et al.) [1]의 논문에서는 준지도 K-means방법을 이용한 문서군집 방법을 제안하였다. 이들의 방법은 분류표시가 된 자료를 이용하여 초기 시드 클러스터를 생성하고, 분류표시가 된 자료로부터 제약사항을 생성하여 군집한다.

(H. J. Zeng, et al.) [12]의 논문에서는 비지도 학습방법인 웹 검색 결과의 군집을 지도학습방법으로 변환하는 방법을 제안하였다. 이들의 방법은 주어진 질의와 검색 결과의 순위 리스트로부터 여러 개의 속성을 계산하고, 이 속성과 학습자료를 이용하여 회귀 모델 학습에 적용하여 검색 결과에 대한 성능을 향상시켰다.

SpeClustering 모델을 (Y. Huang, et al.) [6]의 논문에서 제안하였다. 이들의 방법은 군집의 이름과 관련이 없는 일반적인 특질로부터 군집에 필요한 문서의 특질을 분류하고, 제안 모델에 다양한 유형의 사용자 피드백을 적용하여 매개변수(parameters)를 조정할 수 있는 방법을 제공하였다.

(김천식외 저자) [13]의 논문에서는 등은 신경망알고리즘과 C4.5알고리즘을 이용하여 문서를 분류하는 방법을 제안하였다.

(송재원외 저자) [14]의 논문에서는 영역 기반의 이미지 검색 시스템을 위하여 공간 위치 정보를 적합성 피드백을 위한 가중치를 사용하는 방법을 제안하였다.

III. 용어 재가중치에 의한 문서군집

제안 모델은 용어의 가중치 재계산 단계와 문서군집 단계로 이루어진다. 다음 그림1은 제안된 문서군집 모델의 개요이다. 제안 모델의 가중치 재계산 방법은 SVD과 NMF와 같이 원본데이터를 분해하여 군집하는 모든 방법에 적용할 수 있다. 본 논문에서는 군집방법으로 Xu[11]의 NMF를 이용한 군집방법과 K-Means[5] 군집방법을 이용한다.

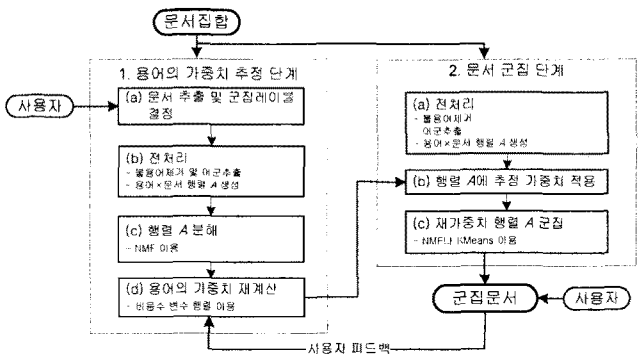

그림 1. 제안된 문서군집 방법
Fig. 1 The proposed document clustering method

3.1 NMF를 이용한 문서군집 방법

본 논문에서는 용어의 가중치 추정과 문서군집을 위하여 Xu[11]의 NMF를 이용한 문서군집방법을 이용한다. 본 논문에서 행렬 X의 j번째 열벡터는 X_{*j}로, i번째 행벡터는 X_{i*}로, i번째 행과 j번째 열의 원소는 X_{ij}로 표시한다.

NMF는 주어진 양의 행렬로부터 양의 인수를 찾아내는 행렬분해 알고리즘이다[7, 8, 11]. 문서집합이 k개의 군집으로 구성된다고 가정할 때, 행렬 X를 식(2)의 목적 함수가 최소값을 갖도록 식(1)과 같이 $m \times k$ 비음수 의미특징 행렬(NSFM) W와 $k \times n$ 비음수 의미변수 행렬(NSFM) H로 분해한다.

$$X \approx WH \quad \cdots\cdots (1)$$

$$J = \frac{1}{2}\|X - WH\| \quad \cdots\cdots (2)$$

여기서 $W = [w_{ij}]$이고 $H = [h_{ij}]$이며 $W = [W_1, W_2, \cdots, W_k]$이다. W와 H의 원소 값을 갱신하기 위하여 목적함수 J값이 수렴 허용오차 보다 작아지거나 지정한 반복횟수를 초과할 때까지 식(3)을 반복한다[7, 8, 11].

$$w_{ij} \leftarrow w_{ij}\frac{(XH^T)_{ij}}{(WHH^T)_{ij}}, h_{ij} \leftarrow h_{ij}\frac{(W^TX)_{ij}}{(W^TWH)_{ij}} \quad \cdots\cdots (3)$$

여기서 H^T는 H의 전치행렬이고, W^T는 W의 전치행렬이다.

Xu등이 제안한 NMF를 이용한 문서군집 방법은 다음과 같다[11]. 먼저 주어진 문서 집합에서 j열로 가중치가 부여된 용어빈도 벡터 문서를 가지는 용어문서 행렬 A를 구성한다. 행렬 A에 식(3)으로 NMF를 수행하여 비음수 행렬 W와 H를 얻는다.

식(4)를 이용하여 행렬 W와 H를 정규화 한다. 행렬 W를 이용하여 각 문서의 군집 레이블을 결정한다. 예를 들어, 만약 $x = \arg max_j h_{ji}$이면 문서 d_i를 군집 x에 할당한다.

$$w_{ij} \leftarrow \frac{w_{ij}}{\sqrt{\sum_i w_{ij}^2}}, \quad h_{ij} \leftarrow h_{ij}\sqrt{\sum_i w_{ij}^2} \quad \cdots\cdots (4)$$

행렬 A의 j번째 열벡터 A_{*j}는 행렬 W의 l번째 열벡터 W_{*l}와 행렬 H의 요소 H_{kj}^T가 선형조합을 이루며 식(5)과 같다.

$$A_{*j} = \sum_{l=1}^{k} H_{kj}^T W_{*l} \quad \cdots\cdots (5)$$

3.2 용어의 가중치 추정 단계

용어의 가중치 추정 단계는 용어의 가중치를 재계산하는 단계로 문서군집 단계에서 적용하거나, 사용자가 만족하지 않을시 다시 용어의 가중치를 재계산하는데 이용된다. 용어의 가중치 추정 단계는 문서추출 및 군집결정, 전처리, 비음수 행렬분해, 용어의 가중치 재계산으로 이루어진다.

그림1의 1.(a)는 문서 추출 및 군집결정 단계로 사용자에 의하여 문서집합으로부터 용어의 재가중치 계산을 위한 문서를 추출하고, 군집 레이블에 맞도록 문서를 분류한다. 즉, 일부 샘플 문서들을 사용자가 원하는 군집으로 분류한다.

그림1의 1.(b)는 전처리 단계로서, 그림1의 1.(a)에서 사용자가 군집한 문서를 불용어 제거, 어근추출을 하여서 벡터모델[3, 9]로 생성하는 단계이다. 여기서 벡터모델은 문서를 전처리하여 총 m개의 용어와 n개의 문서로 이루어진 $m \times n$ 행렬 A이다. 행렬 A는 $[A_{\cdot 1}, A_{\cdot 2}, \cdots, A_{\cdot n}]$로 나타내며, 각 행 벡터 $A_{\cdot i}$는 i번째 문서의 용어빈도 벡터이다.

그림1의 1.(c)는 비음수 행렬 분해 단계로서 행렬 A에 식(3)을 이용한 NMF알고리즘을 적용하여 비음수 의미 특징 행렬 W와 비음수 의미변수 행렬 H로 분해한다. 행렬 W의 의미특징 벡터 W_j는 군집의 내부 특징을 나타내며, 행렬 H의 원소인 의미변수 H_{ij}는 군집 내에서의 의미특징의 중요도를 나타낸다. 즉, 군집에 포함된 문서의 의미변수가 높다는 것은 군집에서 중요한 문서라는 것을 의미하며, 비슷한 의미변수의 값을 갖는 문서들은 비슷한 유형의 문서라는 것을 의미할 수 있다. Xu 등[11]은 비음수 행렬분해에서의 이와 같은 의미변수의 특징을 이용하여 문서군집방법을 제안하였다.

그림1의 1.(d)는 용어 재가중치 계산 단계이다. 용어 재가중치 계산단계의 최종 목적은 식(10)과 같이 a번째 용어의 새로운 가중치를 계산하는 것이다. 그러나, 직접 a번째 용어의 새로운 가중치를 계산 할 수 없다. 이러한 이유 때문에 본 논문에서는 다음 식(9)와 같이 a행의 전체 원소에 대한 평균 가중치의 변화량을 계산하고, 계산된 평균 가중치의 변화량에 이전의 가중치를 더하여 식(10)과 같이 새로운 가중치를 계산하다. 다음 식(9)는 식(6), 식(7), 식(8)에 의해서 유도 할 수 있다.

$$\tilde{A}_{ai} = g_a A_{ai} \quad \cdots\cdots (6)$$

여기서 g_a는 a번째 용어의 가중치 값, A_{ai}는 a번째 용어와 i번째 문서의 용어의 빈도이다.

$$\Delta g_a^i A_{ai} = \sum_{k \in I_i} \Delta H_{ki} W_{ak} \quad \cdots\cdots (7)$$

여기서, Δg_a는 i번째 문서의 a번째 용어 가중치의 변화량, I_i는 의미변수행렬 H에서 i번째 문서의 의미변수벡터 $H_{\cdot i}$에서 $\Delta H_{ki} \neq 0$이 아닌 k들의 집합이다.

$$H_{ki}^{new} = H_{ki}^{old} + \Delta H_{ki} \quad \cdots\cdots (8)$$

여기서, 의미변수 H_{ki}^{old}는 i번째 문서의 k번째 용어에 대한 의미변수의 원래값이다. 즉, 사용자에 의하여 분류되기 이전 군집문서의 의미변수 값이다. 의미변수 H_{ki}^{new}는 i번째 문서의 k번째 용어에 대한 의미변수 값이 사용자의 분류에 의해서 수정된 의미변수의 값이다. H_{ki}^{new}는 다음 식(15)에 의하여 계산할 수 있고, 다음과 같이 식(11), 식(12), 식(13), 식(14)에 의해서 유도 할 수 있다.

$$\Delta g_a = E(\Delta g_a^i) = \frac{1}{n}\sum_{i=1}^{n} \Delta g_a^i = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{A_{ai}} \sum_{k \in I_i} \Delta H_{ki} W_{ak} \quad \cdots\cdots (9)$$

여기서 Δg_a는 a행의 전체 원소에 대한 평균 가중치의 변화량이고, n은 전체문서의 개수이다.

$$g_a^{new} = g_a^{old} + \Delta g_a \quad \cdots\cdots (10)$$

여기서 Δg_a^{old}의 초기값은 1이다.

$$\tilde{H}_{ji}^{old} = \frac{H_{ji}^{old}}{\sum_{k=1}^{r} H_{ki}^{old}} \quad \cdots\cdots (11)$$

여기서 $\tilde{H}_{ji}^{old}$는 정규화된 H_{ji}^{old}이다.

$$b_j^c = \frac{1}{f^c}\sum_{k=1}^{f^c} \tilde{H}_{ji_k^c}^{old}, \quad \sum_{j=1}^{r} b_j^c = 1 \quad \cdots\cdots (12)$$

여기서 c는 군집의 일련번호로 c = 1, 2, … , e이고 e는 군집의 개수, D_c는 c번째 군집이고, c 군집에 포함된 문서의 개수는 f_c이며 $d_{i_k} \in D_c$로 k =1, 2, … , f_c이다.

$$l_{ji_k^c} = \left| b_j^c - \tilde{H}_{ji_k^c}^{old} \right| \quad \cdots\cdots (13)$$

여기서 l_{ji_k}는 c번째 군집 내에서 의미특징의 중요도의 차이를 나타낸다.

$$\Delta\tilde{H}_{ji_k} = \begin{cases} if\ b_j^c > \tilde{H}_{ji_k^c}^{old}, \Delta\tilde{H}_{ji_k^c} = +\xi \cdot l_{ji_k^c} \\ if\ b_j^c < \tilde{H}_{ji_k^c}^{old}, \Delta\tilde{H}_{ji_k^c} = -\xi \cdot l_{ji_k^c} \end{cases},$$

$$\tilde{H}_{ji_k^c}^{new} = \tilde{H}_{ji_k^c}^{old} + \Delta\tilde{H}_{ji_k^c} \quad \cdots\cdots (14)$$

여기서 ξ는 조절 상수로 0.5를 갖는다.

$$H_{ji_k}^{new} = \tilde{\tilde{H}}_{ji_k^c}^{new} \times \sum_{p=1}^{r} H_{pi_k^c}^{old}, \quad \tilde{\tilde{H}}_{ji_k^c}^{new} = \frac{\tilde{H}_{ji_k^c}^{new}}{\sum_{p=1}^{r} \tilde{H}_{pi_k^c}^{new}} \quad \cdots\cdots (15)$$

다음 식(16)는 식(10)에 의하여 재계산된 용어의 가중치 행렬이다.

$$G = \begin{pmatrix} g_1 & 0 & \cdots & 0 \\ 0 & g_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & g_n \end{pmatrix} \quad \cdots\cdots (16)$$

여기서 가중치 행렬 G의 원소값은 g_a^{new} 의 a번째 용어와 일치하는 $A_{a\cdot}$의 용어가 존재하는 경우 g_a^{new} 의 원소값을 가지며, 그렇지 않으면 1의 값을 가진다.

3.3 문서군집단계

문서의 군집단계는 전체문서 집합에 대하여 전처리 단계, 문서집합 행렬에 계산된 용어 가중치 적용단계, 군집단계로 이루어진다.

그림1의 2.(a)는 전처리 단계로서 전체문서를 대상으로 불용어 제거, 어근추출, 벡터모델생성으로 이루어진다[2, 3, 10]. 문서를 전처리 하여 총 m개의 용어와 n개의 문서로 이루어진 $m \times n$행렬 A는 $A = [A_{\cdot 1}, A_{\cdot 2}, \cdots, A_{\cdot n}]$로 나타내며, 각 행 벡터 $A_{i\cdot}$는 i번째 문서의 용어빈도 벡터이다.

그림1의 2.(b)에서 행렬 A에 계산된 용어 가중치 적용은 식(16)을 이용하여 식(17)과 같은 행렬 $\tilde{A}$로 적용한다.

$$\tilde{A} = GA \quad \cdots\cdots (17)$$

마지막으로 그림1의 2.(c)와 같이 재계산된 가중치가 적용된 행렬 $\tilde{A}$에 NMF를 이용한 Xu[11]의 방법이나 K-Means방법[5]을 이용하여 군집한다. 군집결과 사용자가 만족하면 종료하고, 만족하지 않으면 그림1의 1.(d) 용어 재가중치 계산 단계로 가서 가중치를 다시 계산한다.

Ⅳ. 실험 및 평가

제안방법에 대한 실험은 20 Newsgroups 문서자료 [10] 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 표1은 실험에 사용된 평가자료의 특성표이다.

표 1 20 Newsgroups 문서집합의 특성
Table. 1 Property of 20 Newsgroups document set

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

본 논문의 성능평가는 문서군집의 표준 평가척도 중 하나

인 식(19)의 NMI(normalize mutual information)를 사용한다[4, 6, 13]. NMI의 상호정보이득은 두 개의 문서군집 C와 C'가 주어질 때 이들간의 상호정보 MI(C, C')로 다음 식(18)과 같이 정의된다.

$$MI(C,C') = \sum_{c_i \in C, c'_j \in C'} p(c_i, c'_j) \cdot \log_2 \frac{p(c_i, c'_j)}{p(c_i) \cdot p(c'_j)} \quad \cdots\cdots (18)$$

$$NMI(C,C') = \frac{MI(C,C')}{\max(H(C), H(C'))} \quad \cdots\cdots (19)$$

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

실험은 서로 다른 두 가지 군집방법의 NMI를 군집의 개수를 2에서 10까지 증가하면서 비교 하였다. 그림2는 K-Means 문서군집방법과 K-Means 문서군집방법에 제안방법으로 가중치를 재계산하여 적용한 방법간의 비교 결과 이다. 여기서 KMeans는 표준 K-Means를 이용한 문서 군집방법이고[5], NMF는 비음수 행렬 분해를 이용한 문서를 군집한 방법이다[11]. WeightKmeans과 WeightNMF는 각각 Kmeans과 NMF 문서군집방법에 추정된 용어의 가중치를 적용한 방법이다.

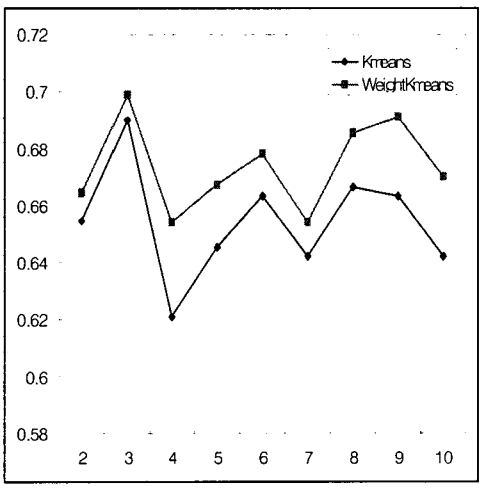

그림 2 20 Newsgroups에 대한 Kmeans과 WeightKmeans 문서군집 간의 비교 결과
Fig. 2 Comparison of evaluatoin results between Kmeans document clustering and WeightKmeans document clustering by using 20 Newsgroups

그림3은 NMF문서군집방법과 NMF문서군집방법에 제안방법으로 가중치를 재계산하여 적용한 방법 간의 비교 결과이다.

그림2와 그림3에서와 같이 기존의 군집방법에 비하여 재계산된 가중치를 적용한 방법이 더 좋은 결과를 나타낸다.

그림4는 20 Newsgroups에 대한 4가지 방법간의 평균 NMI이다. 평가결과 KMeans 군집방법에 비하여는 Weight KMeans의 평균 NMI가 2.9%, NMF 군집방법에 비하여 WeightNMF 군집방법이 평균 NMI가 5.2% 성능이 더 높다. 그림4에서 가중치를 재계산하지 않은 KMeans이 최하의 성능을 나타낸다. 여기서 NMF가 KMeans보다 성능이 좋은 것은 KMeans와 같이 단순한 유사도를 이용한 군집보다 NMF를 이용하여 자료의 내부구조를 반영하여 군집하는 것이 더 정확도에 영향을 미치는 것을 알 수 있다. 또한 WeightNMF가 가장 좋은 성능을 보이는 것은 자료의 내부구조를 군집에 반영하면서, 사용자가 부여한 학습을 통하여 용어의 가중치를 재계산하여서 사용자가 원하는 군집결과로 유도하기 때문이다.

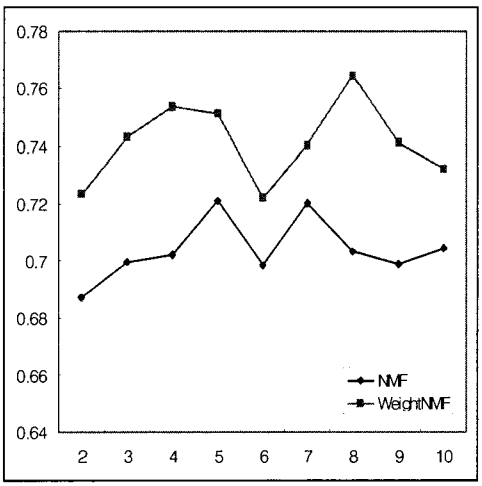

그림 3 20 Newsgroups에 대한 NMF와 WeightNMF 문서군집 간의 비교 결과
Fig. 3 Comparison of evaluation results between NMF document clustering and WeightNMF document clustering by using 20 Newsgroups

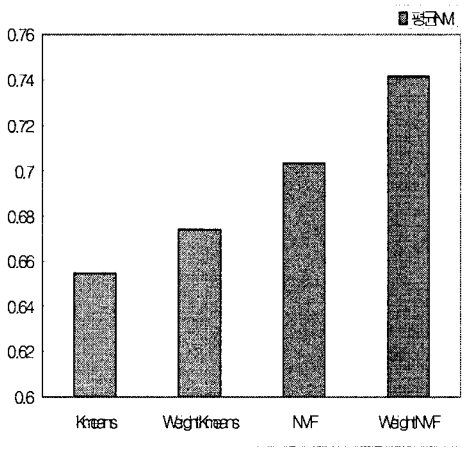

그림 4 20 Newsgroups에 대한 평균 NMI 비교결과
Fig. 4 Comparison of average NMF of 20 Newsgroups

V. 결론

본 논문은 비음수 행렬 분해를 이용하여 용어의 가중치를 재계산하는 사용자 중심의 문서를 군집하는 새로운 모델을 제안하였다. 제안모델은 사용자가 분류한 문서를 이용하여 용어의 가중치를 재계산하여 문서를 군집함으로써 사용자의 요구와 군집결과 사이의 의미적 차이를 최소화 시켰으며, 의미 특징과 의미변수를 사용하여 문서의 내부구조를 군집에 반영함으로써 군집의 정확도를 향상 시켰다. 또한 가중치를 재계산할 수 있는 새로운 방법을 개발하였으며, 재계산된 가중치를 전통적인 문서군집에 적용하여 군집의 효율을 향상시켰다. 실험 결과 가중치를 재계산한 방법이 가중치를 계산하지 않은 방법에 비해서 더 좋은 성능을 나타냄을 알 수 있다.

앞으로 제안 모델의 성능 향상을 위하여 용어에 대한 재가중치를 계산할 수 있는 다양한 정책과 SVD를 이용한 군집방법에 적용할 수 있는 용어 가중치 재계산 방법에 대하여 연구가 진행 되어야 할 것이다.

참고문헌

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

[2] S. Chakrabarti, "mining the web: Discovering Knowledge from Hypertext Data", Morgan Kaufmann Publishers, 2003.

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

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

[5] J. Han, M. Kamber, "Second Edition Data Mining Concepts and Techniques", Morgan Kaufman, 2006.

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

[7] D. D. Lee, H. S. Seung, "Learning the parts of objects by non-negative matrix factorization', Nature, vol.401, 788-791, 1999.

[8] D. D. Lee, H. S. Seung, "Algorithms for non-negative matrix factorization", In Advances in Neural Information Processing Systems, vol.13, 556-562, 2001.

[9] B. Y. Ricardo, R. N. Berthier, "Moden Information Retrieval", ACM Press, 1999.

[10] The 20 newsgroups data set. http://people.csail.mit.edu/jrennie/20Newsgroups/, 2007.

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

[12] 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.

[13] 김천식, 홍유식, "텍스트 마이닝을 이용한 XML 문서 분류 기술", 한국컴퓨터정보학회 논문지 11권2호, 2006.5.

[14] 송재원, 김덕환, 이주홍, "공간 위치 정보를 적합성 피드백을 위한 가중치로 사용하는 영역 기반 이미지 검색 시스템", 한국컴퓨터정보학회 논문지 11권4호, 2006.9.

저 자 소 개

이주홍 (Ju-Hong Lee)

1983년 서울대학교 컴퓨터공학과 졸업(학사)

1985년 서울대학교 대학원 컴퓨터공학과 졸업(석사)

2001년 한국과학기술원 컴퓨터공학과 졸업(박사)

2001년~현재 인하대학교 컴퓨터정보공학과 부교수

관심분야 : 데이터마이닝, 데이터베이스, 정보검색, 신경망, 기계학습

E-mail : juhong@inha.ac.kr

박 선(Sun Park)

1996년 전주대학교 전자계산학과 졸업(학사)

2001년 한남대학교 정보산업대학원 정보통신학과 졸업(석사)

2007년 인하대학교 컴퓨터정보공학과 졸업(박사)

2008~현재 호남대학교 컴퓨터공학과 전임강사

관심분야 : 정보검색, 데이터마이닝, 데이터베이스

E-mail : sunpark@honam.ac.kr