논문 2012-49SP-2-6

군집의 중요 용어와 위키피디아를 이용한 문서군집 향상

(Enhancing Document Clustering using Important Term of Cluster and Wikipedia)

박 선*, 이 연 우**, 정 민 아***, 이 성 로****

(Sun Park, Yeonwoo Lee, Min-A Jeong, and Seong Ro Lee)

요 약

본 논문은 군집 중요 용어들과 위키피디아(Wikipedia)의 동음이의어를 이용하여 문서군집의 성능을 향상시키는 새로운 방법을 제안한다. 제안된 방법은 비음수행렬분해의 의미특징을 이용하여 군집 중요 용어들을 선택함으로서 군집을 대표할 수 있는 군집 주제(topic)의 개념을 잘 표현할 수 있으며, 군집의 중요 용어에 위키피디아의 동음이의어를 사용하여 확장함으로써 문서와 군집 간의 의미관계를 고려하지 않는 용어집합(bag-of-words) 문제를 해결할 수 있다. 또한 확장된 군집의 중요 용어를 이용하여 문서집합을 재 군집하여 초기 군집을 정제함으로써 군집방법의 성능을 높일 수 있다. 실험결과 제안방법을 적용한 문서군집방법이 다른 문서군집 방법에 비하여 좋은 성능을 보인다.

Abstract

This paper proposes a new enhancing document clustering method using the important terms of cluster and the wikipedia. The proposed method can well represent the concept of cluster topics by means of selecting the important terms in cluster by the semantic features of NMF. It can solve the problem of “bags of words” to be not considered the meaningful relationships between documents and clusters, which expands the important terms of cluster by using of the synonyms of wikipedia. Also, it can improve the quality of document clustering which uses the expanded cluster important terms to refine the initial cluster by re-clustering. The experimental results demonstrate that the proposed method achieves better performance than other document clustering methods.

Keywords : 문서군집(document clustering), 비음수행렬분해(NMF, non-negative matrix factorization), 의미 특징(semantic features), 위키피디아(wikipedia), 동음이의어(synonym), 중요 용어(important term).

* 정회원-교신저자, 목포대학교 정보산업연구소
(Institute Research of Information Science and Engineering, Mokpo National University)
** 정회원, 목포대학교 정보통신공학과
(Department of Information Communication Engineering, Mokpo National University)
*** 정회원, 목포대학교 컴퓨터공학과
(Department of Computer Engineering, Mokpo National University)
**** 정회원, 목포대학교 정보전자공학과
(Department of Information Electronic Engineering, Mokpo National University)
※ 이 논문은 2010년도 정부(교육과학기술부)의 재원으로 한국연구재단의 대학중점연구소 지원사업으로 수행된 연구임(2010-0028295), 본 논문은 한국통신학회 지원에 의하여 연구되었음.
접수일자: 2011년11월16일, 수정완료일: 2011년12월27일

Ⅰ. 서 론

문서군집은 군집 알고리즘을 사용하여 문서집합으로부터 유사한 특징을 가진 문서들의 집합으로 그룹화 하는 방법이다. 문서군집은 정보검색의 중요한 기술로 문서분류, 요약, 주제추출, 정보 필터링 등의 효율적인 기반 기술로 많이 사용하고 있다. 특히, 트위터, 페이스북, 블로그, RSS(really simple syndication) 등과 같이 개인화된 서비스가 증가하면서 각각의 서비스 특성에 맞도

록 요구하는 문서 분류에 대한 필요성이 증가하고 있다. 이로 인하여 문서군집에 대한 관심이 점차 높아지고 있다.

전통적인 문서군집 알고리즘은 대부분 문서를 단어의 집합(BOW, bag-of-words)으로 표현하는 방법을 주로 사용하고 있다. 그러나 이러한 방법은 문서 집합에 포함된 용어(term; 단어)들의 의미적 관계를 전혀 고려하지 않고, 단지 용어들이 문서에 출현된 빈도만을 이용하고 있다[1].

용어의 빈도를 기반으로 한 문서군집 방법은 크게 두 가지 요인에 따라서 군집 결과에 많은 영향을 받는다. 첫 번째 요인으로 문서 집합의 자체 특성이다. 즉, 문서 집합에서 문서의 분포나 내부구조, 사용자가 요구하는 군집 개수 등에 따라서 군집의 결과가 달라진다. 두 번째 요인은 군집 알고리즘에서 사용되는 목적함수들이다. 문서군집 알고리즘에 많이 사용하는 거리기반의 목적함수는 두 문서 간의 실제 거리를 잘 반영할 수 없는 문제를 가지고 있다[1]. 이러한 문제를 해결하기 위해서 최근 연구에서는 외부지식인 온톨로지(ontology, 공유된 개념화) 및 위키피디아(wikipedia)를 이용하거나, 문서집합의 내부구조를 나타내는 의미특징(semantic feature)을 많이 사용하고 있다.

외부지식 기반은 주로 워드넷이나 전문가의 수작업으로 온톨로지를 구축하여서 문서군집의 성능을 향상시키나, 군집을 위한 용어의 포괄적인 개념을 찾아 온톨로지를 구축하는 것이 어렵고 또한 구축비용이 많이 든다. 이외에도 온톨로지를 구축하더라도 정확한 범위를 적용대상에 적용하는 것도 힘들어서 정보손실 문제가 발생한다[1]. 온톨로지 기반의 문제점을 해결하기 위해 최근에는 위키피디아를 활용한 방법이 많이 연구[1~4]되고 있다. 위키피디아 기반의 방법들은 온톨로지 기반의 방법에 비해서 좋은 성능을 보이나, 위키피디아의 모든 정보를 전처리하여 필요한 개념으로 모델화 할 때 많은 비용이 드는 단점과 학습이 필요한 경우가 있다.

위키피디아는 웹 기반의 다국적 언어의 자유로운 컨텐츠를 지향하는 온라인 백과사전으로 사용자들이 정보의 생산자로 참여하여 광범위한 지식 정보를 제공하고 있다. 위키피디아는 실시간적 이면서도 지속적으로 문서가 생성 또는 가공되기 때문에 워드넷과는 다르게 시간경과에 따른 정보의 유효성에 대한 제약을 받지 않는다. 또한 전문적인 사전이나 서적, 기사, 연구문헌 등을 기반으로 참고하여 기재되기 때문에 문서의 개념과 내용에 대한 신뢰성을 가지고 있다[5].

의미특징에 기반한 문서군집 방법은 문서집합 내부구조의 특성을 이용하여서 의미 있는 군집의 주제들을 추출하고, 추출된 주제들과 관련된 문서들의 집합으로 쉽게 군집할 수 있다. 그러나 문서집합의 구성 문서들이 유사한 특성을 같거나, 극단적으로 다른 특성을 갖고 있으면 추출된 의미특징들의 문서집합의 내부 구조를 충분하게 반영할 수 없으므로 좋은 군집 결과를 얻기 힘들다[6].

본 논문에서는 의미특징과 위키피디아 기반 방법의 제한 사항을 극복하는 의미특징과 위키피디아를 이용한 새로운 문서군집방법을 제안한다. 제안 방법은 다음과 같다. 첫 번째는 초기군집 단계로 *k*means 군집방법을 이용하여서 설정된 *k*개로 문서로 군집한다. 두 번째는 군집의 중요한 용어들을 추출하는 단계로, 각각의 군집에 비음수 행렬의 의미특징을 이용하여 군집의 주제를 나타내는 중요도가 높은 용어들을 추출한다. 세 번째는 중요 용어들의 확장단계로, 군집의 중요 용어와 위키피디아의 동음이의어 문서 목록을 이용하여 중요 용어들을 확장한다. 마지막은 군집의 정제단계로, 확장된 군집의 중요 용어와 군집 간의 유사도를 이용하여 문서집합을 재 군집한다.

제안된 방법은 다음과 같은 장점을 갖는다. 첫째, 의미특징을 이용하여 추출된 군집의 중요 용어들은 군집의 내부 특성을 잘 반영할 수 있는 군집의 주제를 요약된 형태로 잘 표현할 수 있다. 둘째, 확장된 군집 주제의 용어들은 의미특징이 원본 문서집합의 문서구성에 제한받는 문제를 극복할 수 있으며, 학습이 필요 없으며 중요 용어들만을 이용하여 용어를 확장하기 때문에 위키피디아 전체 내용을 전처리하는 비용 부담을 덜 수 있다. 마지막으로, 확장된 군집 주제의 중요 용어들을 이용하여 문서를 재 군집하여 초기군집을 정제함으로써 군집의 성능을 향상 시킬 수 있다.

본 논문의 구성은 다음과 같다. 제Ⅱ장은 관련연구로 문서군집에 대한 최신 연구를, 제Ⅲ장은 제안된 문서군집 방법을, 제Ⅳ장에서는 실험 및 평가에 대해 기술한다. 마지막으로 제Ⅴ장에서는 결론을 맺는다.

Ⅱ. 관련 연구

최근의 온톨로지를 이용한 문서군집에 대한 연구로는 다음과 같다. Trappey외 저자들은 특허문서를 군집

하기 위하여 퍼지에 기반한 온톨로지 방법을 제안하였다[7].

위키피디아 기반의 최근 문서군집 연구는 다음과 같다. Hu외 저자들은 외부 지식인 위키피디아를 이용한 문서군집 방법을 제안하였다. 이들의 방법은 위키피디아의 개념과 문서집합을 연결하여 위키피디아의 분류(category)를 문서군집에 이용할 수 있도록 제안하였다. 즉 문서와 위키피디아의 분류와의 관계를 유도하였다. 이들의 방법은 위키피디아의 분류 및 개념으로부터 문서간의 관계를 유도하기 위해 모든 위키피디아의 정보를 전처리하여서 계산 비용이 높은 문제를 가지고 있다[1]. Huang외 저자들은 위키피디아를 기반으로 활성학습(active learning)하여 문서를 군집하는 방법을 제안하였다. 이들의 방법은 학습을 통하여 위키피디아로 부터 위키피디아의 주제에 관련된 개념기반의 문서를 생성하고, 생성된 의미적 관계의 문서를 이용하여 문서를 군집한다. 그러나 이들의 방법은 군집하기 이전에 사전작업으로 위키피디아를 학습해야한다[2]. Huang외 저자들은 위키피디아 기반의 개념의 표현을 이용하여 문서를 군집하는 방법을 제안하였다. 이 방법 역시 문서군집에 관련된 위키피디아의 개념을 추출하기위해서는 학습해야 한다[3]. Kiran외 저자들은 연관규칙과 위키피디아를 이용한 계층적 문서군집방법을 제안하였다. 이들의 방법은 문서군집을 위해 생성되는 연관규칙이 일반화 근접 빈도 항목(generalized closed frequent itemsets)을 이용하여 일반적인 연관규칙에 비해서 계산 양을 줄였더라도 역시 많은 계산을 필요로 하는 문제를 가지고 있다[4].

의미특징에 기반한 문서군집의 최근 연구는 다음과 같다. Li 이외 저자들은 문서군집과 관련된 군집의 하위 공간구조의 특징을 이용한 ASI(Adaptive Subspace Iteration) 알고리즘을 제안하였다[8]. Wang과 Zhang은 문서군집을 위하여 지역 레이블과 전역 레이블의 특징을 이용한 CLGR(Clustering with Local and Global Regularization) 알고리즘을 제안하였다[9]. Xu이외 저자들은 비음수 행렬 분해(NMF, Non-negative Matrix Factorization)의 의미특징을 이용하여 문서를 군집하는 방법을 제안하였다[10]. 본 논문의 저자들은 이전에 의미특징에 기반 하여서 문서군집을 위한 세 가지 방법을 제안하였다. 제안방법으로는 의미특징과 군집의 응집도를 이용한 방법[11~12], 의미특징과 퍼지관계를 이용한 방법[13], 마지막으로 주성분 분석과 퍼지연관을 이용한 방법[14]이 있다. 이들 방법은 의미특징에 기반을 두고 있기 때문에 구성 문서의 특성이 극단적으로 유사하거나 다르면 군집의 성능이 좋지 않을 수 있는 문제를 가지고 있다. 이전 저자들이 제안한방법의 문제점을 해결하기 위해서 비음수행렬분해와 워드넷에 기반한 군집 주제의 유의어와 유사도를 이용한 문서군집 향상 방법을 제안하였다[15]. 그러나 이 방법 역시 워드넷에 기반하고 있기 때문에 시간경과에 따른 용어의 의미 변화를 군집에 반영하지 못하는 문제를 가지고 있다.

Ⅲ. 제안 문서군집 방법

본 논문에서 제안한 방법은 전처리, 초기 군집, 군집의 중요 용어 추출, 군집의 중요 용어 확장, 군집의 정제 단계로 구성된다. 전처리단계에서는 문서집합을 전처리하여서 용어-문서 빈도 행렬을 구성한다. 초기 군집단계 에서는 *k*means 군집 방법을 이용하여 문서를 군집한다. 군집의 중요 용어 추출 단계에서는 비음수행렬분해를 이용하여 군집의 주제(topic)를 설명할 수 있는 중요 용어들을 추출한다. 군집의 중요 용어 확장 단계에서는 위키피디아의 동음이의어 문서 목록과 유사도를 이용하여 중요 용어를 확장한다. 군집의 정제 단계에서는 확장된 군집주의 중요 용어와 군집에 포함된 문서 간의 유사도를 계산하여 문서를 재 군집한다. 다음 장에서 제안방법의 각 단계에 대하여 자세히 설명한다.

3.1 전처리

현재 1위인 영문 위키피디아의 경우 항목의 개수는 3,795,912개 이며 한글 위키피디아는 20위 순위로 181,054개의 항목으로 구성되어 있다[16]. 영문 위키피디아의 경우 전 분야에 고르게 항목들의 분포되어 있으며, 일반 백과사전 이상의 내용과 전문성을 가지고 있다. 그에 비하여 한글 위키피디아의 경우 특정 분야에 편중되어 있고, 항목들이 느리게 생성되고 있으며 아직은 많은 분야가 미흡한 실정이다. 이러한 이유에서 본 논문 전처리 방법은 영문 문서를 기준으로 설명한다. 한글 문서에 본 논문의 제안방법을 적용하려면, 전처리 단계에 한글 형태소분석 도구[17]로 용어를 추출하여 용어-문서 빈도행렬을 구성하면 된다.

전처리 단계는 주어진 문서집합으로부터 불용어 제거, 어근추출, 용어빈도 벡터를 생성한다[18]. 불용어 제거는 Rijsbergen의 불용어 목록[18]을 이용하여서 목록에

서 정의하고 있는 무의미한 용어들을 제거한다. 어근추출은 Porter의 어근추출 알고리즘[18]을 이용하여서 영어의 파생어들을 가장 중심이 되는 용어인 어근으로 변환한다. 용어-문서 빈도 행렬의 용어빈도 벡터 생성에 사용되는 벡터 $T_i = [\ t_{1i}, t_{2i}, \cdots, t_{ni}\ ]^T$는 i번째 문서의 용어빈도이다. 여기서 요소 t_{ij}는 j번째 문서에서 출현한 i번째 용어의 빈도이다[8, 18].

3.2 초기 군집

군집의 주제를 나타내는 중요 용어를 추출하기 위해서는 먼저 문서집합을 초기 군집해야 한다. 초기 군집 단계를 위해서 본 논문에서는 일반적으로 군집방법에 가장 많이 사용하는 *k*means 알고리즘을 이용하여 전처리된 용어-문서 빈도행렬을 초기 군집한다.

*k*means은 n개의 자료를 주어진 k개의 군집으로 묶는 알고리즘이다[19]. 본 논문에서는 문서를 초기 군집하기 위하여 식(1)의 코사인 유사도를 이용한 거리 척도를 사용한다.

$$d(T_{*a}, T_{*b}) = 1 - csim(T_{*a}, T_{*b}) \tag{1}$$

$$csim(T_{*a}, T_{*b}) = \frac{\sum_{i=1}^{m} T_{ia} \times T_{ib}}{\sqrt{\sum_{i=1}^{m} T_{ia}^2} \times \sqrt{\sum_{i=1}^{m} T_{ib}^2}} \tag{2}$$

여기서, T_{*a}와 T_{*b}는 문서행렬 T의 a번째와 b번째 열벡터이다. 이 것 들은 비음수 값을 가지므로 $0 \leqq csim() \leqq 1$이고, 따라서 $0 \leqq d() \leqq 1$이다.

3.3 군집의 중요 용어 추출

군집의 중요 용어 추출 단계에서는 비음수행렬분해를 이용하여 군집 주제를 설명할 수 있는 중요 용어들을 초기 군집으로 부터 추출한다.

비음수행렬분해는 비음수 자료로 구성된 원본 자료를 두 개의 비음수로 구성된 행렬로 분해한다[6]. 비음수행렬 분해 알고리즘은 식(3)의 목표함수 J 가 0에 가깝게 수렴 할 때까지 식(4)를 이용하여 행렬 W와 H의 값을 동시에 갱신하여 원본 자료를 두개의 비음수 행렬로 분해한다.

$$J = \| T - WH \|^2 \tag{3}$$

식(1)의 목적은 행렬 T를 비음수 $m \times r$ 행렬 W와 비음수 $r \times n$ 행렬 H로 분해하는 것이다. 여기서, T는 m개의 용어와 n개의 문서로 구성된 $m \times n$ 행렬이고, r은 의미특징행렬의 크기를 결정할 수 있는 의미특징의 개수이다. 또한 두 개의 비음수 의미 특징 행렬을 구별하기 위하여서, 비음수행렬분해 알고리즘을 제안한 Lee와 Seung은 두 행렬 W와 H를 의미특징 행렬 W와 의미변수 행렬 H로 각각 이름을 정의하였다[6].

$$H_{rj} \leftarrow H_{rj} \frac{(W^T V)_{rj}}{(W^T WH)_{rj}},\ W_{ir} \leftarrow W_{ir} \frac{(VH^T)_{ir}}{(WHH^T)_{ir}} \tag{4}$$

본 논문에서는 비음수행렬분해를 이용하여 군집의 중요 용어를 추출하는 방법으로 저자들이 이전에 제안한 방법[15]을 수정하여서 이용한다. 이전에 제안 방법의 수정된 방법은 다음과 같다. 추출을 원하는 용어의 개수만큼 의미특징의 개수(r)를 설정하고, 초기 군집의 각각의 군집을 전처리한 후에 비음수해렬분해 한다. 행렬분해 된 의미특징행렬 W에 식(5)를 이용하여 군집의 주제를 잘 설명할 수 있는 중요 용어들을 추출한다. 추출된 중요 용어들 중에서 중복되는 용어들은 제거한다. 즉, 행렬 W의 열벡터는 군집의 주제에 대응되며, 행벡터는 군집을 구성하는 문서들의 용어에 대응된다. 이러한 이유에서 열벡터에 포함된 높은 값의 의미특징은 그 열벡터에 대응되는 군집에 중요한 용어가 된다. 군집의 대표 용어를 추출하는 식은 다음과 같다.

$$IT^p \leftarrow T_{ij} \text{ if } p = \underset{1 \leq l \leq r}{\arg max}\ W_{il} \text{ and } W_{il} \geq cv^l \tag{5}$$

여기서, IT^p는 p번째 군집을 대표하는 용어집합이고, T_{ij}는 j번째 열벡터(군집)에 속하는 i번째 행의 의미특징에 대응되는 용어이다. cv^l는 l번째 열벡터에 포함된 의미특징의 평균값으로 식(6)과 같다.

$$cv^l = \frac{\sum_{i=1}^{n} W_{il}}{n} \tag{6}$$

여기서, n은 i행의 개수이다. 즉, 용어(의미특징)의 개수이다.

본 논문에서 군집 주제의 중요 용어를 추출할 때 비음수행렬분해를 사용하는 이유는, 비음수행렬분해에 의해 생성되는 의미특징들은 원본 문서집합의 내부특징을 잘 표현할 수 있다. 즉, 일반적으로 문서집합은 다양한 주제를 갖는 문서들로 구성되어 있고, 각각의 주제를

포함 하는 문서들을 모아서 군집을 구성할 수 있다. 이 때문에 비음수행렬분해 된 의미특징들은 문서들이 포함하고 있는 중요한 주제들을 쉽게 의미 있는 특징들로 그룹화하여서 나타낼 수 있다. 그러나 의미특징들은 문서집합의 내부의 구조특성만을 이용하기 때문에 실제로 문서들이 같은 주제를 포함하고 있으면서 다른 형태로 표현하는 경우 잘 구분할 수 없는 문제를 가지고 있다. 이러한 문제를 해결하기 위하여서 본 논문에서는 군집의 중요 용어를 확장한다.

3.4 군집의 중요 용어 확장

군집의 중요 용어만을 이용하여 문서를 군집할 때, 중요 용어와 일치하는 용어들로 구성된 문서들은 잘 군집 되나, 중요 용어가 나타내는 군집의 주제를 포함하고 있으면서 다른 용어 집합으로 구성된 문서들은 좋은 군집 결과가 나오지 않는 문제를 가지고 있다. 이러한 문제를 해결하기 위하여 본 논문에서는 영어 위키피디아의 동음이의어 문서목록을 이용하여서 군집의 중요 용어를 확장한다.

군집의 중요 용어들 중에는 위키피디아의 항목에 일치하지 않는 용어들이 존재한다. 이는 중요 용어들이 주제를 표현할 만큼 의미가 있는 못하기 때문이다. 이러한 중요 용어들은 확장하지 않고, 위키피디아의 항목에 일치되는 용어만을 이용하여 다음과 같이 확장한다.

확장방법은 중요 용어를 위키피디아의 동음이의어 문서 목록 검색하고, 동음이의어 목록에 포함된 동음이의어 항목 및 설명과 군집 간의 유사도를 식(2)를 이용하여 계산한다. 상위 순위에 있는 동음이의어 항목만을 군집의 중요 용어 집합에 추가하여서 확장된 군집 중요 용어 집합 EIT^p를 구성한다. 여기서 EIT^p는 p번째 군집에서 확장된 군집의 대표 용어 집합 EIT이다.

3.5 군집의 정제

군집의 정제를 위한 재 군집 방법은 다음과 같다. 유식(2)를 이용하여서 각각의 문서와 각각의 군집의 확장된 중요 용어 집합 EIT 간의 유사도를 계산한다. 확장된 중요 용어 집합과 가장 높은 유사도를 갖는 문서를 확장 중요 용어 집합의 군집에 포함시킨다. 만약 확장된 중요 용어의 유사도가 0이어서 분류할 수없는 문서가 있는 경우 $k$means 알고리즘을 이용하여 분류된 원래의 군집에 이 문서를 배정한다.

3.6 제안방법의 알고리즘

다음은 본 논문에서 제안한 문서군집방법의 알고리즘이다. 제안 알고리즘의 1행에서는 문서집합 D를 전처리하여 용어-문서 빈도행렬 T를 생성한다. 2행에서는 $k$means 군집방법을 이용하여 T를 초기 군집한다. 3행에서는 전처리된 초기 군집 행렬 C를 비음수행렬분해 하여서 비음수 의미특징행렬 W와 비음수 의미변수 행렬 H를 계산한다. 4행에서는 의미특징 행렬 W와 식(6)을 이용하여서 의미특징 열벡터의 평균 cv를 계산한다. 5행에서 7행까지는 군집의 중요 용어집합을 추출한다. 8행에서는 군집의 중요 용어집합과 위키피디아의 동음이의어 목록을 이용하여서 중요 용어집합을 확장한다. 9행에서 15행까지는 식(2)의 유사도를 이용하여 문서를 재 군집한다. 이중 10행에서는 확장된 군집의 중요 용어집합과 문서간의 유사도를 이용하여 문서를 재 군집하며, 12행과 같이 유사도가 0인 경우 $k$means 알고리즘을 이용하여 분류된 원래의 군집에 이 문서를 분류한다.

Algorithm: ***WikiNMF***(D,r);

Input: 문서집합 D, 군집의 개수 k.

Output: 용어(m)-문서(n) 빈도행렬 T, 비음수 행렬 W and H, 군집된 문서집합 C, 재 군집된 문서집합 RC

1: $T \leftarrow$ 전처리(D);
2: $C \leftarrow kmeans(T, k)$;
3: $[W, H] \leftarrow$ 비음수행렬분해(전처리(C));
4: $cv \leftarrow CV(W)$;
5: for $p \leftarrow 1$ to k do
6: $IT^p \leftarrow T_{ij}$ if $p = \underset{1 \le l \le r}{\arg max} W_{il}$ and $W_{il} \ge cv^l$;
7: end
8: $EIT \leftarrow$중요용어확장(IT);
9: for $j \leftarrow 1$ to n do
10: if $csin(T_{*j}, EIT^k) \neq 0$
11: $RC^p \leftarrow T_{*j}$ if $p = \underset{1 \le j \le n, 1 \le k \le r}{\arg max} csin(T_{*j}, EIT^k)$;
12: else
13: $RC^j \leftarrow T_{*j}$;
14: endif
15: end

Ⅳ. 실험 및 평가

본 논문에서는 문서군집 및 분류의 표준 성능평가 자료로 많이 사용하는 20 Newsgroups 문서자료[20]를 이

용하여 제안방법의 성능을 비교평가 하였다. 20 Newsgroups의 구성으로는 뉴스 그룹이 20개가 있으며, 20개의 뉴스 그룹에는 총 20000 개의 문서가 있다. 뉴스그룹은 컴퓨터 그래픽, 운영체제 윈도우, 컴퓨터 하드웨어, 종교, 의학, 정치 등 20개의 다양한 주제에 같은 수의 기사를 포함하고 있다.

본 논문에서는 문서군집의 성능평가를 위하여 20 Newsgroups문서자료 중 일부를 무작위로 추출하여 사용하였다. 다음 표 1은 평가에 사용된 평가 자료의 특성표이다. 본 논문에서는 성능평가 방법의 척도(measure)로 문서군집에서 주로 사용하는 식(8)의 NMI(normalize mutual information)를 이용한다[8, 15]. NMI는 두 군집간의 정보이득을 계산하여서 성능을 평가하는 방법이다. NMI의 상호정보이득은 두 개의 문서군집 C와 C'가 주어질 때 이들 간의 상호정보 MI(C, C')로 다음 식(7)과 같다.

$$MI(C,C') = \sum_{c_i \in C, c'_j \in C'} p(c_i, c'_j) \cdot \log_2 \frac{p(c_i, c'_j)}{p(c_i) \cdot p(c'_j)} \quad (7)$$

$$NMI(C,C') = \frac{MI(C,C')}{\max(H(C), H(C'))} \quad (8)$$

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 문서군집방법과 제안방법간의 성능을 비교 평가 하였다. 평가방법은 20 Newsgroups 문자서료로 부터 임의로 추출된 10개의 군집문서를 이용하여서 군집하고, 군집결과를 실제 20 Newgroups에 분류되어 있는 문서와 NMI를 비교하였다. 비교방법으로는 군집의 개수를 2에서 10까지 증가시키며 각각 50번 반복하여서 각각의 군집에 평균을 계산하여서 평가하였다.

표 1. 평가에 사용된 문서집합의 특성
Table 1. The property of document set with respect to evaluation.

문서집합의 속성	20 Newsgroups
총 문서 갯수	20000
사용문서 갯수	5000
클러스터 갯수	20
사용 클러스터 갯수	2~10
최대 클러스터의 문서 갯수	500
최소 클러스터의 문서 갯수	20
중간 클러스터의 문서 갯수	300
평균 클러스터의 문서 갯수	280

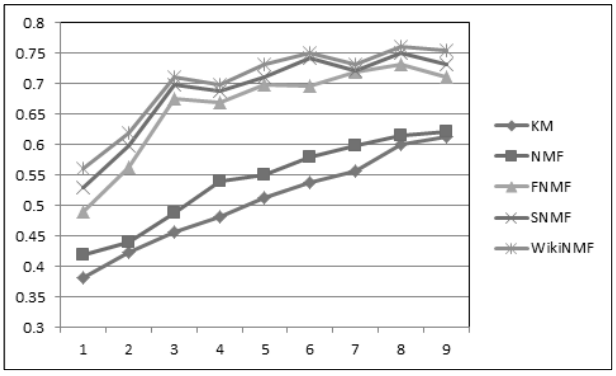


그림 1. 문서군집방법들 간 평균 NMI 비교결과
Fig. 1. The result of comparison of average NMF in document clustering methods.

평가에 사용된 비교방법들은 다음 그림1과 같이 KM[19], NMF[10], FNMF[13], SNMF[15], WikiNMF 등의 문서군집방법을 구현하여 비교평가 하였다. 여기서 KM은 전통적인 분할기반의 군집방법으로 *k*means를 이용한 방법이다. NMF와 FNMF는 의미특징을 이용한 방법이며 SNMF는 의미특징과 온톨로지를 사용한 방법이다. 이들 중에서 FNMF와 SNMF는 이전에 저자들이 제안한 방법이다. WikiNMF는 본 논문에서 제안한 방법이다. FNMF는 비음수행렬분해와 퍼지연관을 이용한 문서군집방법이고, SNMF는 의미특징과 워드넷을 이용하여 군집 주제의 유의어와 유사도를 이용하여 문서를 군집하는 방법이다.

그림 1에서 NMF군집방법이 KM군집방법 보다 좋은 성능을 보인다. 이는 KM에서의 단순히 두 문서간의 거리 척도를 사용하는 것보다는 NMF의 의미특징들을 이용하여 자료의 내부구조를 반영하는 것이 군집결과에 더 영향을 미치는 것을 알 수 있다[6, 10]. 또한 KM이나 NMF 보다는 군집의 각각의 특성을 나태 내는 대표용어와 군집에 포함되는 문서의 용어들 간의 연관관계를 고려한 FNMF가 좋은 성능을 보인다. 또한 문서집합의 내부 특징을 이용하는 FNMF 보다는 군집 대표 용어를 확장하는 SNMF이 더 좋은 성능을 보인다. 특히 제안된 WikiNMF가 가장 좋은 성능을 보이는데, 이것은 군집의 주제를 잘 표현하는 중요 용어집합과 문서집합의 내부 자료 특성을 고려하여 외부 지식인 위키피디아의 동음이의어를 이용하여서 중요 용어들을 확장하는 것이 군집결과에 더 좋은 영향을 미치는 것으로 보인다.

V. 결 론

본 논문은 군집 주제의 중요 용어를 확장한 후에, 초기 군집을 재 군집하여 정제함으로써 문서군집의 결과를 향상시키는 방법을 제안하였다. 제안 방법은 비음수 행렬분해를 이용하여서 문서집합의 주제를 잘 표현 할 수 있는 군집의 중요 용어들을 추출하였으며, 문서집합의 내부 구조만을 반영하는 의미특징이 특정 자료 집합에 군집이 제한되는 것을 극복하기 위하여 위키피디아의 동음이의어 문서목록을 사용하여 중요 용어를 확장하였다. 또한, 확장된 중요 용어 집합으로 문서집합을 재 군집하여 군집을 정제함으로써 군집의 성능을 향상시켰다. 성능평가 결과 제안방법인 WikiNMF의 평균 NMI가 KM 방법에 비하여서는 27.78%가, NMF 방법보다는 23.19%가, FNMF 방법보다는 5.73%가, SNMF 방법 보다는 2.29%가 각각 높음을 보였다.

참 고 문 헌

[1] X. Hu, X. Zhang, C. Lu, E. K. Park, X. Zhou, “Exploiting Wikipedia as External Knowledge for Document Clustering”, Proceeding of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 397-406, 2009.

[2] A. Huang, D. Milne, E. Frank, I. H. Witten, “Clustering Document with Active Learning using Wikipedia”, Proceeding of the 8th IEEE International Conference on Data Mining (ICDM'08), pp. 839-844, 2008.

[3] A. Huang, D. Milne, E. Frank, I. H. Witten, “Clustering Document using a Wikipedia-based Concept Representation”, Proceeding of Advances in Knowledge discovery and data mining, LNCS 5476, pp.628-636, 2009.

[4] G. V. R. Kiran, K. Ravi Shankar, V. Pudi, “Frequent Itemset based Hierarchical Document Clustering using Wikipedia as External Knowledge”, Technical Report No: IIT/TR/2010/33, Wales, UK, 2010.

[5] wikipedia, “http://www.wikpedia.com/”, 2011.

[6] D. D. Lee, H. S. Seung, “Learning the parts of objects by non-negative matrix factorization,” Nature, 401, pp. 788-791, Oct. 1999.

[7] A. J. C. Trappey, C. V. Trappey, F. C. Hsu, and D. W. Hsiao, “A Fuzzy Ontological Knowledge Docment Clustering Methodolgoy, The Journal of IEEE Transcation On System, Man and Cypernetics,” vol. 39, no. 3, Jun. pp.806-814, 2009.

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

[9] F. Wang, C. Zhang, “Regularized Clustering for Documents”, In proceeding of ACM SIGIR'07, pp. 95-102, 2007.

[10] 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.

[11] 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.

[12] 박선, 김철원, “비음수 행렬 분해와 군집의 응집도를 이용한 문서군집”, 한국해양정보통신학회 논문지, 제13권 제12호, pp. 2603-2608, 2009.

[13] 박선, 김경준, “비음수 행렬 분해와 퍼지 관계를 이용한 문서군집”, 한국항행학회 논문지, 제14권 제2호, pp. 239-246, 2010.

[14] 박선, 안동언, “주성분 분석과 퍼지 연관을 이용한 문서군집 방법”, 한국정보처리학회 논문지, 제17-B권, 제2호, pp. 177-182, 2010.

[15] 박선, 김경준, 이진석, 이성로, “군집 주제의 유의어와 유사도를 이용한 문서군집 향상 방법”, 전자공학회논문지 제48권 SP편 제5호, pp. 30-38, 2011.

[16] List of Wikipedias, “http://meta.wikimedia.org/wiki/List_of_Wikipedias”, 11월, 2011.

[17] 한경한, 남경완, “한국어 정보 처리 입문 : 컴퓨터가 우리말을 이해하려면”, 커뮤니케이션북스, 2007.

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

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

[20] The 20 newsgroups data set. http://people.csail.mit.edu/jrennie/20Newsgroups/, 2011.

저 자 소 개

박 선(정회원)-교신저자

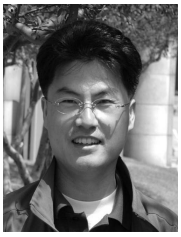
1996년 전주대학교 전자계산학과 학사 졸업.
2001년 한남대학교 정보산업 대학원 정보통신학과 석사 졸업.
2007년 인하대학교 컴퓨터정보 공학과 박사 졸업.
2008년～2009년 호남대학교 컴퓨터공학과 전임강사.
2010년 전북대학교 전기전자정보인력양성사업단 박사후과정.
2011년～현재 목포대학교 정보산업연구소 연구교수.

<주관심분야 : 정보검색, 데이터마이닝, 데이터베이스, 해양생물 IT정보융합>

정 민 아(정회원)

1994년 2월 전남대학교 전산 통계학과 석사
2002년 2월 전남대학교 전산통계학과 박사
2005년 3월～현재 목포대학교 컴퓨터공학과 부교수

<주관심분야 : 데이터베이스/데이터마이닝, 생체인식시스템, 무선통신응용분야(RFID, USN, 텔레메틱스), 임베디드시스템>



이 연 우(정회원)

1994년 2월 고려대학교 전자공학과 석사
2000년 2월 고려대학교 전자공학과 박사
2000년 10월～2003년 12월 영국 Edinburgh 대학교 Research Fellow
2004년 1월～2005년 8월 삼성종합기술원
2005년 9월～현재 국립목포대학교 정보공학부 정보통신공학전공, 부교수

<주관심분야 : 해상무선통신, e-Navigation, Cognitive Radio, 4G 이동통신>

이 성 로(정회원)

1987년 고려대학교 전자공학과 졸업
1990년 한국과학기술원 전기및 전자공학과 석사
1996년 한국과학기술원 전기및 전자공학과 박사
1997년 9월～현재 목포대학교 공과대학 정보전자공학과 교수

<주관심분야 : 디지털통신시스템, 이동 및 위성통신시스템, USN/텔레미틱스응용분야, 임베디드시스템>