군집 주제의 유의어와 유사도를 이용한 문서군집 향상 방법

박선*, 김철원**
*목포대학교 정보산업연구소
**호남대학교 컴퓨터공학과
e-mail:sunpark@mokpo.ac.kr

Enhancing Document Clustering Method using Synonym of Cluster Topic and Similarity

Sun Park*, Chul-won Kim**
*Research Faculty Institute of Information Science and Engineering Research, Mokpo National University
**Dept of Computer Engineering, Honam University

요 약

본 논문은 군집 주제의 유의어와 유사도를 이용하여 문서군집의 성능을 향상시키는 방법을 제안한다. 제안된 방법은 비음수행렬분해의 의미특징을 이용하여 군집 주제(topic)의 용어들을 선택함으로서 문서 군집 집합의 내부구조를 잘 표현할 수 있으며, 군집 주제의 용어들에 워드넷의 유의어를 사용하여서 확장함으로써 문서를 용어집합(bag-of-words)으로 표현하는 문제를 해결할 수 있다. 또한 확장된 군집 주제의 용어와 문서집합에 코사인 유사도를 이용하여서 군집의 주제에 적합한 문서를 잘 군집하여서 성능을 높일 수 있다. 실험결과 제안방법을 적용한 문서군집방법이 다른 문서군집 방법에 비하여 좋은 성능을 보인다.

1. 서론

문서군집 방법을 정의하면 군집 알고리즘을 사용하여서 유사한 특성의 문서들을 집합으로 묶는 기술이다[1, 2]. 일반적인 문서군집 방법들에서는 문서를 용어의 집합(BOW, bag-of-words)으로 표현하는 방법을 주로 사용한다. 그러나 이러한 문서를 용어의 집합으로 표현하는 방법은 문서집합에 포함된 용어들의 의미적 관계를 전혀 고려하지 않고, 단지 용어들이 문서에 출현된 빈도만을 이용한다. 이 때문에 용어 빈도 집합을 사용한 문서군집 방법은 문서집합 내에 포함된 문서의 특성이 군집의 결과에 많은 영향을 미친다. 즉, 문서들의 분포나 문서집합의 내부구조, 사용자가 요구하는 군집형태 등에 따라서 군집의 결과가 달라진다. 또한 문서 집합을 군집 할 때에 문서간의 거리를 측정하여서 군집하는 거리기반의 목적함수를 사용함으로써 두 문서간의 실제 거리를 잘 반영할 수 없다[3]. 이러한 문제를 해결하기 위해서 요즘은 문서군집에서는 온톨로지(ontology, 공유된 개념화)나 의미특징(semantic feature)을 이용한 방법을 많이 사용하고 있다.

온톨로지에 기반한 문서군집 방법은 워드넷이나 위키피디아 등의 외부 지식으로부터 용어 온톨로지를 구축하여서 문서군집의 성능을 향상시킨다. 그러나 문서집합에서 사용되는 용어에 대해서 포괄적인 개념을 찾아 온톨로지를 구축하는 것이 어렵다. 또한 온톨로지 구축비용이 많이 들고, 온톨로지를 구축하더라도 정확한 범위를 적용대상에 적용하는 것도 힘들다. 또한 이러한 특성 때문에 때로는 정보손실 문제가 발생한다[3]. 최근의 온톨로지를 이용한 문서군집에 대한 연구로는 다음과 같다. Hu외 저자들은 문서군집을 위하여 위키피디아의 외부 지식을 이용하여 온톨로지를 구축하였다[3]. Trappey외 저자들은 특허문서를 군집하기 위하여 퍼지에 기반한 온톨로지 방법을 제안하였다[4].

의미특징에 기반한 문서군집 방법은 문서집합 내부의 특성을 나타내는 의미특징을 이용한 방법으로서 쉽게 군집의 특성을 나타내는 주제들을 추출할 수 있다. 또한 명확한 의미의 군집주제를 나타내는 의미특징을 이용하여 좋은 군집결과를 얻을 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[5]. 의미특징에 기반한 문서군집의 최근 연구는 다음과 같다. Li 이외 저자들은 문서군집과 관련된 군집의 하위공간구조의 특징을 이용한 ASI(Adaptive Subspace Iteration) 알고리즘을 제안하였다[6]. 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[7]. 본 논문의 저자들은 이전에 문서군집을 위한 세 가지 방법을 제안하였다. 제안방법으로는 의미특징과 군집의 응집도를 이용한 방법[8, 9], 의미특징과 퍼지관계를 이용한 방법[11], 마지막으로 주성분 분석과 퍼지연관을 이용한 방법[3]이 있다. 이들 방법은 의미특징에 기반을 두고 있기 때문에 구성 문서의 특성이 극단적으로 유사하거나 다르면 군집의 성능이 좋지 않을 수 있는 문제를 가지고 있다.

본 논문에서는 의미특징 방법의 제한 사항을 극복하기 위하여

서 군집 주제의 유의어와 코사인 유사도를 이용한 문서군집의 성능 향상 방법을 제안한다. 제안 방법은 다음과 같다. 첫 번째 단계로 비음수 행렬의 의미특징을 이용하여서 군집의 주제를 나타낼 수 있는 중요도가 높은 용어들을 추출한다. 이렇게 추출된 용어들은 군집의 내부 특성을 잘 반영할 수 있는 군집의 주제를 요약된 형태로 잘 표현할 수 있다. 두 번째 단계에서는 추출된 용어들을 워드넷을 이용하여서 유의어로 확장한다. 확장된 군집 주제의 용어들은 의미특징이 원본 문서집합의 문서구성에 제한받는 문제를 극복할 수 있다. 마지막으로 확장된 군집 주제의 용어들과 원본 문서들 간에 코사인 유사도를 이용하여서 문서를 군집한다. 군집 주제를 잘 반영 할 수 있는 문서들을 코사인 유사도를 이용하여서 군집함으로써 군집의 성능을 향상 시킬 수 있다.

2. 비음수행렬분해

이번 장에서는 비음수행렬분해의 개념과 알고리즘에 대하여 알아보고, 다음 장에서 비음수행렬분해를 이용하여서 군집 주제의 중요 용어를 추출하는 제안방법에 대하여 알아본다.

비음수행렬분해는 대량의 객체정보로부터 비음수로 된 부분정보를 추출하고, 이들의 선형 조합으로 객체를 표현할 수 있도록 하는 방법이다. 비음수행렬분해 알고리즘은 비음수 자료로 구성된 원본 자료를 두 개의 비음수로된 행렬로 분해한다[5]. 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.

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

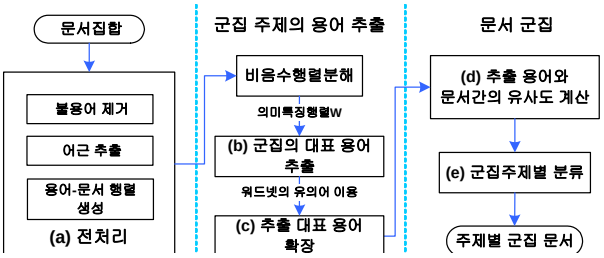
식(1)의 목적은 행렬 A를 비음수 $m \times r$ 행렬 W와 비음수 $r \times n$ 행렬 H로 분해하는 것이다. 여기서, A는 m개의 용어와 n개의 문장으로 이루어진 $m \times n$ 행렬이고, r은 의미특징행렬의 크기를 결정할 수 있는 의미특징의 개수이다. 또한 두 개의 비음수 의미 특징 행렬을 구별하기 위하여서, 비음수행렬분해 알고리즘을 제안한 Lee와 Seung은 두 행렬 W와 H를 의미특징 행렬 W와 의미변수 행렬 H로 각각 이름을 정의하였다[5].

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

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

3. 제안 문서군집 방법

본 논문에서 제안한 문서군집 과정은 다음 그림1과 같이 전처리, 군집 주제의 용어 추출, 문서군집으로 구성된다. 전처리단계에서는 문서집합을 전처리하여서 용어-문서 빈도행렬을 구성한다. 군집 주제의 용어 추출 단계에서는 비음수행렬분해를 이용하여 군집의 주제를 요약하여서 설명할 수 있는 중요 용어들을 추출하고, 워드넷을 이용하여 추출된 용어들을 확장한다. 문서군집 단계에서는 추출된 군집주제의 용어 집합과 문서들 간의 유사도를 계산하여 문서를 군집한다.



(그림 1) 의미 특징과 유사도를 이용한 문서군집

그림1(a)의 전처리 단계는 주어진 문서집합으로부터 불용어 제거, 어근추출, 용어빈도 벡터를 생성한다[1, 2]. 불용어 제거는 Rijsbergen의 불용어 목록[2]을 이용하여서 목록에서 정의하고 있는 무의미한 용어들을 제거한다. 어근추출은 Porter의 어근추출 알고리즘[8]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 문장의 용어빈도이다. 여기서 요소 t_{ij}는 j번째 문서에서 출현한 i번째 용어의 빈도이다[7, 8, 15].

비음수행렬분해를 이용하여 그림1(b)와 같이 군집 주제를 나타내는 군집의 대표 용어들을 추출하고, 그림1(c)와 같이 추출된 군집의 대표 용어를 워드넷의 유의어를 이용하여 확장한다.

그림1(b)와 같이 군집의 주제를 잘 요약하여 설명 할 수 도록 군집의 대표 용어를 추출하는 방법은 다음과 같다. 문서집합을 전처리하여서 용어-문서 빈도행렬 A를 생성하고, 추출하고자 하는 군집의 개수(의미특징 r)를 설정한다. 설정된 군집의 개수를 이용하여서 비음수행렬분해 한다. 행렬분해 된 의미특징행렬 W를 이용하여 군집의 주제를 잘 설명할 수 있는 용어들을 추출한다. 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.

$$R^p \leftarrow A_{ij} \text{ if } p = \underset{1 \le j \le r}{\arg max}\, W_{ij} \text{ and } W_{ij} \ge cv^j \quad (4)$$

여기서, R^p는 p번째 군집을 대표하는 용어집합이고, A_{ij}는 j번째 열벡터(군집)에 속하는 i번째 행의 의미특징에 대응되는 용어이다. cv^j는 j번째 열벡터에 포함된 의미특징의 평균값으로 식(5)과 같다.

$$cv^j = \frac{\sum_{i=1}^{n} W_{ij}}{n} \quad (5)$$

여기서, m은 i행의 개수이다. 즉, 용어(의미특징)의 개수이다.

군집의 대표 용어를 이용하여 문서를 군집할 때, 대표용어와 일치하는 용어들로 구성된 문서들은 잘 군집 되나, 대표 용어가 나타내는 군집의 주제를 포함하고 있으면서 다른 용어들로 구성된 문서들은 좋은 군집 결과가 나오지 않는 문제를 가지고 있다. 이러한 문제를 해결하기 위하여 본 논문에서는 워드넷을 이용하여 대표 용어들을 유의어 집단으로 확장한다. 확장방법은 대표 용어를 워드넷을 이용하여서 명사에 대한 유의어를 검색하고, 이 유의어 집합을 대표용어에 추가하여서 확장된 군집주제의 용어 집합 ER^p를 구성한다. 여기서 ER^p는 p번째 군집에서 확장된 군집의 대표 용어 집합 ER이다.

유사도를 이용한 문서 군집 방법은 다음과 같다. 식(6)을 이용하여서 각각의 문서와 각각의 군집의 대표 용어집합 R간의 유사도를 계산한다. 군집의 대표 용어집합과 가장 높은 유사도를 갖는 문서를 대표 용어집합의 군집에 포함시킨다. 그러나 일반 적으로 문서집합에 구성된 문서의 특성들을 보면 군집에 나타내는 주제에 일치하면서 동음이의어(homonym)나 이음동의어(유의어,synonym)으로 구성되어 있어서 유사도를 이용하여 구별할 수 없는 경우가 있다. 이러한 이유에서 본 논문에서는 군집의 대표 용어 집합과의 유사도가 0인 문서가 있다면 확장된 군집의 대표 용어집합 ER과 유사도를 계산하여 군집한다. 다음은 본 논문에서 유사도 계산에 사용되는 코사인 유사도 $csin()$이다[15].

$$csim(A_{*a},A_{*b}) = \frac{\sum_{i=1}^{m} A_{ia} \times A_{ib}}{\sqrt{\sum_{i=1}^{m} A_{ia}^2} \times \sqrt{\sum_{i=1}^{m} A_{ib}^2}} \qquad (6)$$

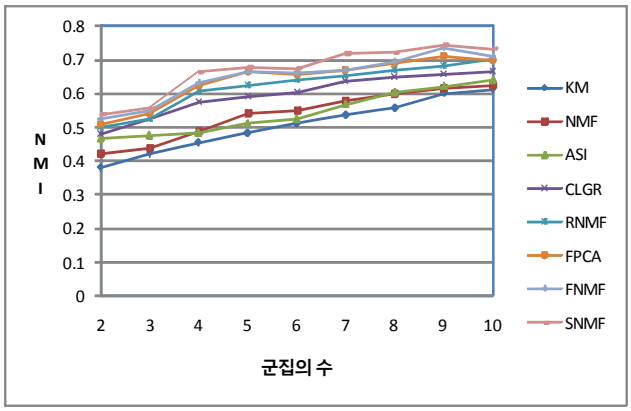
여기서, A_{*a}와 A_{*b}는 행렬 A의 a번째와 b번째 열벡터이다. 이것 들은 비음수 값을 가지므로 $0 \leqq csim() \leqq 1$ 이다.

4. 실험 및 평가

본 논문의 평가자료는 20 Newsgroups문서자료[5]를 이용하였다. 20 Newsgroups문서자료는 문서군집 및 분류의 표준 성능평가 자료로 많이 사용하는 자료이다. 20 Newsgroups는 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.

본 논문의 실험은 서로 다른 일곱 가지 문서군집방법과 제안 방법간의 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.

평가에 사용된 비교방법들은 직접 구현하였으며, 다음 그림2와 같이 KM[1], NMF[7], ASI[6], CLGR[9], RNMF[8], FPCA[10], FNMF[11], SNMF등의 문서군집방법을 비교 평가 하였다. 여기서 KM은 전통적인 분할기반의 군집방법으로 Kmeans를 이용한 방법이다. 본 논문에서는 기존 방법과의 비교 기준을 세우기 위하여서 사용되었다. 나머지 NMF, ASI, CLGR, RNMF, FPCA, FNMF, SNMF등은 의미특징을 이용한 방법으로 RNMF, FPCA, FNMF는 이전에 저자들이 제안한 방법이다. 여기서, SNMF는 본 논문에서 제안한 방법이며, FNMF와 FPCA는 비음수행렬분해와 주성분 분석에 각각 퍼지연관을 이용한 문서군집방법들이고, RNMF는 비음수행렬분해와 군집의 정제방법을 이용한 군집방법이다. 또한, NMF는 비음수 행렬분해의 의미특징을 이용한 Xu의 문서군집방법이며, ASI는 Li가 제안한 문서군집방법으로 반복 적응형 군집의 하위 공간 구조를 이용하였으며, CLGR는 Wang이 제안한 방법으로 군집의 지역과 전역의 정규화 속성을 이용하여서 문서를 군집하는 방법이다. 성능평가한 결과, 제안방법인 SNMF의 평균 NMI가 KM군집 방법에 비하여서는 16.29%가, NMF군집 방법보다는 13.10%가, ASI군집 방법보다는 12.56%가, CRGL군집 방법보다는 7.29%가, RNMF군집 방법보다는 4..73%가, FPCA군집 방법보다는 2.88%가, FNMF군집 방법보다는 2.06%가 각각 높음으로서 다른 문서군집 방법에 비하해서 더 좋은 성능을 나타냄을 알 수 있다.



(그림 2) 문서군집방법들 간 평균 NMI 비교결과

5. 결론

본 논문은 군집 주제의 유의어와 유사도를 이용하여서 문서군집의 결과를 향상시키는 방법을 제안하였다. 제안 방법은 비음수 행렬분해를 이용하여서 문서집합의 주제를 잘 표현 할 수 있는 군집 주제의 용어들을 추출하였으며, 비음수행렬의 의미특징이 문서집합의 내부 구조만을 반영하여서 특정 자료 집합에 군집이 제한되는 것을 극복하기 위하여, 워드넷의 유의어를 사용하여 군집 주제의 용어집합을 확장하였다. 또한, 군집 주제 용어와 확장된 용어 집합에 유사도를 이용하여서 문서집합으로부터 군집의 주제를 잘 반영한 문서를 분류하였다.

참고문헌

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

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

1992.
[3] X. Hu, X. Zhang, C. Lu, E. K. Park, X. Zhou, "Exploiting Wikipedia as External Knowledge for Document Clustering," In proceeding of 15th ACM SIGKDD Conference On Knowledge Discover and Data Mining (KDD'09), Paris, Fance, Jun. 2009. pp. 389-396
[4] 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. 2009, pp.806-814
[5] D. D. Lee, H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," Nature, 401, pp. 788-791, Oct. 1999.
[6] T. Li, S. Ma, M. Ogihara, "Document Clustering via Adaptive Subspace Iteration", In proceeding of SIGIR'04, pp. 218-225 (2004)
[7] 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.
[8] 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.
[9] F. Wang, C. Zhang, "Regularized Clustering for Documents", In proceeding of ACM SIGIR'07, pp. 95-102, 2007.
[11] 박선, 김경준, "비음수 행렬 분해와 퍼지 관계를 이용한 문서군집", 한국항행학회 논문지, 제14권 제2호, pp. 239-246, 2010.
[10] 박선, 안동언, "주성분 분석과 퍼지 연관을 이용한 문서군집 방법", 한국정보처리학회 논문지, 제17-B권, 제2호, pp. 177-182, 2010.