군집과 위키피디아를 이용한 문서군집

박선* · 이성호** · 박희만** · 김원주** · 김동진** · 산드라 아벨** · 이성로***

*, **, ***목포대학교

Document Clustering using Clustering and Wikipedi

Sun Park* · Seong Ho Lee** · Hee Man Park** · Won Ju Kim** · Dong Jin Kim** · Abel Chandra** · Seong Ro Lee***

*, **, ***Mokpo National University

E-mail : *sunpark@mokpo.ac.kr, ***srlee@mokpo.ac.kr

요 약

본 논문은 군집과 위키피디아(Wikipedia)를 이용하여 문서를 군집하는 새로운 방법을 제안한다. 제안된 방법은 비음수행렬분해를 이용하여 군집을 대표할 수 있는 군집 주제(topic)의 개념을 잘 표현할 수 있으며, 위키피디아의 동음이의어를 사용함으로써 문서와 군집 간의 의미관계를 고려하지 않는 용어집합(bag-of-words) 문제를 해결할 수 있다. 실험결과 제안방법을 적용한 문서군집방법이 다른 문서군집 방법에 비하여 좋은 성능을 보인다.

ABSTRACT

This paper proposes a new document clustering method using clustering and Wikipedia. The proposed method can well represent the concept of cluster topics by means 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. The experimental results demonstrate that the proposed method achieves better performance than other document clustering methods.

키워드

문서군집(document clustering), 비음수행렬분해(NMF, non-negative matrix factorization), 의미 특징(semantic features), 위키피디아(wikipedia)

Ⅰ. 서 론

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, 문서집합에서 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[2, 3, 4, 5, 6].

본 논문에서는 의미특징과 위키피디아 기반 방법의 제한 사항을 극복하는 의미특징과 위키피디아를 이용한 새로운 문서군집방법을 제안한다.

II. 본 론

제안 방법은 다음과 같다. 첫 번째는 초기군집 단계로 kmeans 군집방법을 이용하여서 설정된 k개로 문서로 군집한다.

kmeans은 n개의 자료를 주어진 k개의 군집으로 묶는 알고리즘이다[7]. 본 논문에서는 문서를 초기 군집하기 위하여 식(1)의 코사인 유사도를 이용한 거리 척도를 사용한다.

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

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

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.

두 번째는 군집의 중요한 용어들을 추출하는 단계로, 각각의 군집에 비음수 행렬의 의미특징을 이용하여 군집의 주제를 나타내는 중요도가 높은 용어들을 추출한다.

군집의 대표 용어를 추출하는 식은 다음과 같다.

$$IT^p \leftarrow T_{ij} \text{ if } p = \underset{1 \le l \le r}{\arg max}\, W_{il} \text{ and } W_{il} \geq cv^l \quad (3)$$

여기서, IT^p는 p번째 군집을 대표하는 용어집합이고, Tij는 j번째 열벡터(군집)에 속하는 i번째 행의 의미특징에 대응되는 용어이다. cv^l는 l번째 열벡터에 포함된 의미특징의 평균값이다.

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.

III. 결론

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.

Acknowledgement

이 논문은 2011년도 정부(교육과학기술부)의 재원으로 한국연구재단의 대학중점연구소 지원사업으로 수행된 연구임(2011-0022980), 본 연구는 지식경제부 및 정보통신산업진흥원의 대학 IT연구센터 지원사업의 연구결과로 수행되었음 (NIPA-2012-H0301-12-2005)

참고문헌

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