의미 특징과 퍼지를 이용한 문서군집

박선*, 김철원**, 안동언***
*전북대학교 전기전자정보인력양성사업단-BK21
**호남대학교 컴퓨터공학과
***전북대학교 전자정보공학부
e-mail:sunbak@jbnu.ac.kr, cwkim@honam.ac.kr, duan@jbnu.ac.kr

Document Clustering using Semantic Features and Fuzzy

Sun Park*, Chul Won Kim**, Dong Un An***
*Advanced Graduate Education Center of Jeonbuk for Electronics and Information Technology-BK21
**Dept of Computer Engineering, Honam University
***Division of Electronic & Information Engineering, Chonbuk National University

요 약

본 논문은 문서의 의미특징과 퍼지를 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,퍼지를 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.

1. 서론

인터넷의 발전은 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[4, 5, 7, 8, 13]. 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].

문서군집은 문서집합으로부터 유사한 특성을 가진 문서들의 그룹을 발견하는 것이다. 문서군집은 다양한 정보검색 응용분야에 활용되는 중요한 방법[3, 4, 6, 13]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[8]. 또한 점차 용량이 증가하는 문서들의 고차원 객체를 효율적으로 군집할 수 있는 기술의 필요성이 증가 하고 있다.

의미특징이나 군집의 레이블을 이용한 문서군집 방법에 대한 기존연구는 다음과 같다. Xu이외의 저자들은 비음수 행렬 분해(NMF, Non-negative Matrix Factorization)의 의미특징을 이용하여 문서를 군집하는 방법을 제안하였다[16]. *Ji*외 저자들은 문서 군집 분석에 군집의 구성원에 대한 사전지식을 통합한 준지도 문서 군집 모델을 제안하였다[8]. *Basu*외 저자들은 준지도 Kmeans방법을 이용한 문서군집 방법을 제안하였다한다[2]. 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[16]. 본 논문의 저자들은 이전에 비음수 행렬 분해와 군집의 정제방법을 이용한 문서군집 방법을 제안하였다[13]. 또한 저자들은 주성분 분석과 퍼지연관을 이용한 문서군집 방법을 제안하였다[12].

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)로 나누어 부분정보(part-base)로 표현한다[9]. 퍼지 관계(Fuzzy Relationship)는 퍼지집합 이론을 사용하여 정보검색 과정의 모호성을 정형화하는 방법으로, 문서집합의 용어들이나 다른 색인어들 간의 관계를 인식할 수 있다[5, 17]. 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 문서내의 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.

2. 관련연구

2.1 비음수 행렬 분해

비음수 행렬 분해는 주어진 비음수 행렬로부터 비음수의 인수를 찾는 행렬분해 알고리즘이다[9]. 비음수 행렬분해 알고리즘은 식(1)의 목표함수 J 가 0에 가깝게 수렴할 때까지 식(2)를 이용하여 행렬 W와 H의 값을 동시에 갱신한다.

$$J = \|A - WH\|^2 \tag{1}$$

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 의미특징의 개수이다.

$$H_{\alpha\mu} \leftarrow H_{\alpha\mu} \frac{(W^T V)_{\alpha\mu}}{(W^T W H)_{\alpha\mu}}, \quad W_{i\alpha} \leftarrow W_{i\alpha} \frac{(V H^T)_{i\alpha}}{(W H H^T)_{i\alpha}} \tag{2}$$

2.2 퍼지이론

퍼지 이론은 다음과 같이 정의 된다[5, 17].

(정의 1) 두 유한 집합 $X = \{x_1, ..., x_u\}$와 $Y = \{ y_1, ..., y_v\}$ 사이의 퍼지 관계는 이진 퍼지 관계 $f: X \times Y \rightarrow [1,0]$으로 정의된다. 여기서 u와 v는 X와 Y 각각의 원소의 수이다.

(정의 2) 용어 색인 집합 $T = \{t_1, ..., t_u\}$와 문서 집합 $D = \{d_1, ..., d_v\}$가 주어질 때, t_i는 문서들의 퍼지 집합 $h(t_i)$에 의해 표현된다. 즉, $h(t_i) = \{F(t_i, d_j) \mid \forall\ d_j \in D\}$이다. 여기서 $F(t_i, d_j)$는 문서 d_j에서 t_i의 중요도의 정도이다.

(정의 3) 퍼지 관련 용어 관계는 동시에 존재하는 용어들에 이용하여 다음과 같다.

$$r_{i,j} = \frac{n_{i,j}}{n_i + n_j - n_{i,j}} \tag{3}$$

여기서, $r_{i,j}$는 용어 i와 j 사이의 퍼지 관련 용어 관계이다. $n_{i,j}$는 i번째 용어와 j번째 용어를 동시에 포함하는 문서들의 개수이며, n_i는 i번째 문서를 포함하는 문서의 개수이고, n_j는 j번째 문서를 포함하는 문서의 개수이다.

3. 제안방법

본 논문에서 제안한 문서군집 과정은 다음과 같이 전처리, 군집 대표용어 추출, 문서군집으로 구성된다.

전처리 단계는 주어진 문서집합으로부터 불용어 제거, 어근추출, 용어빈도 벡터를 생성한다[4, 13]. 불용어 제거는 Rijsbergen의 불용어 목록[4]을 이용하고, 어근추출은 Porter의 어근추출 알고리즘[4]을 이용한다. 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.

군집 대표용어 추출 단계에서는 비음수 행렬분해를 이용하여 군집의 대표 용어와 군집 레이블을 추출한다. 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 표현 할 수 있다. 특히 가장 높은 값을 가지는 의미특징과 대응되는 용어는 군집을 구성하는 문서들의 특성을 잘 표현 할 수 있는 군집 레이블로 사용할 수 있다.

문서군집단계에서는 군집 대표 용어와 퍼지관계를 이용하여 문서를 군집한다. 퍼지관계를 이용한 문서 군집 방법은 다음과 같다. 용어-문서 빈도행렬에 식(3)의 퍼지 관련 용어 관계를 이용하여 퍼지 관련 용어 관계 상관 행렬을 계산한다. 식(4)를 이용하여 퍼지 관련 용어 상관 행렬과 대표 용어들로부터 퍼지 포함 관계를 계산하다. 계산된 퍼지 포함관계를 이용하여 문서를 군집한다. 즉, 퍼지 포함관계 $\mu_{i,j}$가 최고 값을 갖으면, d_i 문서를 C_j 군집에 할당한다.

각각의 문서들이 각각의 군집 집합에 포함되는 정도인 퍼지 포함관계[5, 17]는 다음과 같이 정의 된다.

$$\mu_{i,j} = \max_{\forall\ t_a \in d_i} \left[1 - \prod_{\forall\ t_b \in CT_j} (1 - r_{a,b}) \right] \tag{4}$$

여기서, $\mu_{i,j}$ 는 j번째 군집 C_j에 i번째 문서 d_i가 속하는 정도이며, $r_{a,b}$는 용어 $t_j \in d_j$와 용어 $t_b \in CT_j$ 사이의 퍼지관계이고, CT_j는 비음수 행렬 분해를 이용하여 선택한 j번째

군집의 대표용어 집합이다.

4. 실험 및 평가

제안방법에 대한 실험은 문서군집의 표준 성능평가 자료인 20 Newsgroups 문서자료[1] 중 일부를 무작위로 추출하여 실험하였다. 20 Newsgroups 평가 자료는 뉴스그룹이 20개가 있으며, 20개의 뉴스 그룹에는 총 20000개의 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 성능평가는 문서군집의 표준 평가척도 중 하나인 NMI(normalize mutual information)를 사용한다[10, 15, 16]. NMI의 상호정보이득은 두 개의 문서군집 C와 C'가 주어질 때 이들 간의 상호정보로 정의된다.

본 논문의 실험은 서로 다른 여섯 가지 문서군집방법과 제안방법간의 NMI를 군집의 개수를 2에서 10까지 증가하면서 비교 하였다. 비교방법으로는 FNMF, KM, NMF, ASI, CLRG, RNMF, FLSA를 비교하였다. 여기서, FNMF는 제안방법으로 비음수 행렬 분해와 퍼지 관계를 이용한 문서군집방법이다. KM은 표준 Kmeans 군집을 이용한 문서군집방법[3, 6]이고, NMF는 비음수 행렬분해의 의미특징을 이용한 Xu의 문서군집방법이다[16]. 또한, ASI는 Li가 제안한 문서군집방법으로 반복 적응형 군집의 하위공간 구조를 이용하고[10], CLRG는 Wang이 제안한 방법으로 군집의 지역과 전역의 정규화 속성을 이용하며[14], RNMF와 FLSA는 저자들의 이전 제안 방법으로 각각 비음수 행렬분해와 군집의 정제방법[11]과 주성분 분석과 퍼지연관을 이용한 다[12].

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 비하해서 더 좋은 성능을 나타냄을 알 수 있다.

5. 결론

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.

참고문헌

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

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

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

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

[5] Haruechaiyasak, C., Shyu, M. L., Chen, S. C.: Web Document Classification Based on Fuzzy Association. In proceedings of the 25thAnnualInternationalComputerSoftwareandApplicationsConference(COMPSAC'02) (2002)

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

[7] 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)

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

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

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

[11] 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)

[12] 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)

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

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

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

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

[17] Zadeh, L. A., Fuzzy Sets, in Dubois, D., Prade, H. and Yager, R. R. editiors, Readings in Fuzzy Sets for Intelligent Systems, Morgan Kaufmann Publiishers, (1993)