JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지(J. Korea Inst. Inf. Commun. Eng.) Vol. 18, No. 3 : 625~630 Mar. 2014

주성분 분석과 k 평균 알고리즘을 이용한 문서군집 방법

김우생[1*] · 김수영[2]

Document Clustering Technique by K-means Algorithm and PCA

Woosaeng Kim[1*] · Sooyoung Kim[2]

[1*]Department of Computer Software, Kwangwoon University, Seoul 139-701, Korea
[2]Department of Computer Engineering, Handong Global University, Pohang 791-708, Korea

요 약

컴퓨터의 발전과 인터넷의 급속한 발전으로 정보의 양이 폭발적으로 증가하게 되었고 이러한 방대한 양의 정보들은 대부분 문서 형태로 관리되기 때문에, 이들을 효과적으로 검색하고 처리하는 방법의 연구가 필요하다. 문서 군집은 문서간의 유사도를 바탕으로 서로 연관된 문서들을 군집화하여 대용량의 문서들을 자동으로 분류하고 검색하고 처리하는데 효율과 정확성을 증대시킨다. 본 논문은 특징 벡터 공간 상의 벡터들로 표현되는 문서들을 K 평균 알고리즘으로 군집화할 때, 주성분 분석을 사용하여 초기 시드점들을 선정함으로써 군집의 효율을 높이는 방법을 제안한다. 실험 결과를 통하여 제안하는 기법이 기존의 K 평균 알고리즘보다 좋은 결과를 얻을 수 있음을 보였다.

ABSTRACT

The amount of information is increasing rapidly with the development of the internet and the computer. Since these enormous information is managed by the document forms, it is necessary to search and process them efficiently. The document clustering technique which clusters the related documents through the similarity between the documents help to classify, search, and process the large amount of documents automatically. This paper proposes a method to find the initial seed points through principal component analysis when the documents represented by vectors in the feature vector space are clustered by K-means algorithm in order to increase clustering performance. The experiment shows that our method has a better performance than the traditional K-means algorithm.

키워드 : 문서 군집화, K 평균 알고리즘, 주성분 분석

Key word : Document Clustering, K-means algorithm, PCA

접수일자 : 2013. 09. 03 심사완료일자 : 2013. 10. 14 게재확정일자 : 2013. 10. 31

*** Corresponding Author** Woosaeng Kim(E-mail:kwsrain@kw.ac.kr, Tel:+82-2-940-5217)
Department of Computer Software, Kwangwoon University, Seoul 139-701, Korea

Open Access http://dx.doi.org/10.6109/jkiice.2014.18.3.625 print ISSN: 2234-4772 online ISSN: 2288-4165

Ⅰ. 서 론

컴퓨터의 발전과 인터넷의 급속한 발전으로 정보의 양이 폭발적으로 증가하게 되었고 이러한 방대한 양의 정보들은 대부분 문서 형태로 관리되어 이들을 효과적으로 검색하고 처리하기 위한 많은 연구가 진행되고 있다. 이 중에서도 문서 군집화(Document Clustering)는 정보 검색 시스템에서 방대한 양의 문서들을 구조화하는데 중요한 역할을 담당하고 있다. 문서들에 대한 군집화는 유사한 문서들의 그룹을 만들어 특정한 카테고리 안에서 검색과 처리를 용이하게 하고, 체계적인 문서 관리와 문서 저장을 위해서도 효율적이다. 또한 군집화 된 데이터들은 데이터들 간에 일종의 경향 또는 규칙성을 보이고 심지어 주목할 가치가 있는 관련 지식을 보여 주기까지 한다.

군집 방법은 크게 확률 모형에 기초한 방법과 확률 모형을 가정하지 않은 방법으로 나눌 수 있다[1,2]. 확률모형에 기초한 방법론 중에서 대표적인 방법은 가우스 혼합모형이며, 확률 모형을 가정하지 않은 방법론 중 대표적인 방법은 계층 군집법과 K 평균 군집법이다. K 평균 군집법은 계층적 군집법에 비하여 계산량이 적고 대용량 데이터를 빠르게 처리할 수 있는 장점이 있으나, 군집의 수를 사전에 알아야 하며 무작위로 초기 시드점을 선택하는 문제점이 있다.

문서는 문서 집합의 단어들을 차원으로 하는 특징 벡터 공간 상에서 하나의 벡터로 표현될 수 있다. 주성분 분석(PCA: Principal Component Analysis)은 다차원적인 변수들을 축소 하는 차원의 단순화와 더불어 일반적으로 서로 상관되어 있는 반응 변수들 간의 복잡한 구조를 분석하는데 주로 이용된다[2,3]. 따라서 본 논문은 특징 벡터 공간 상의 문서들을 K 평균 알고리즘으로 군집화할 때, 주성분 분석을 사용하여 적절한 시드점을 구하는 방법을 제안한다. 실험 결과를 통하여 제안하는 방법이 기존 K 평균 기법보다 더 좋은 결과를 얻음을 보였다.

본 논문의 구성은 다음과 같다. 2장은 관련 연구이고, 3장은 문서-단어 행렬로 표현된 문서들에 주성분을 사용해 적절한 시드점을 구하는 방법에 대하여 설명한다. 4장은 실제 데이터로 실험을 하여 제안한 방법이 효율적인지를 검증하고, 5장에서 결론을 낸다.

Ⅱ. 본 론

2.1. 문서 군집 및 처리

본 장에서는 제안하는 방법과 관련된 문서 군집이나 처리 방법에 관하여 알아본다. [4]는 주성분 분석과 퍼지 연관을 이용하여 문서를 군집화 하였다. 주성분 분석을 사용하여 군집을 대표할 수 있는 몇 개의 대표 용어들을 선택함으로써 군집의 고차원적인 특성으로부터 몇몇 의미 특징을 갖는 용어들로 군집을 효율적으로 표현하였으며, 군집의 대표 용어와 가장 높은 연관 관계를 갖는 용어를 포함하는 문서들로 군집함으로써 문서 군집의 정확도를 높였다. [5]는 문서의 내용을 대표할 수 있는 주제어를 추출하기 위해 주성분 분석을 사용하였다. 주성분 분석의 고유값과 고유벡터를 이용하여 문서 자체 내의 단어의 흐름을 파악한 후 주제어를 추출하는 방법을 사용하였다. [6]는 주성분 분석과 비정칙치 분해를 이용해 문서를 요약하는 방법을 제안하였다. 주성분 분석과 비정칙 분해를 사용해 문장 벡터와 주제어 벡터를 획득한 후, 추출된 주제어와 문장 간의 거리가 가장 짧은 문장들을 중요 문장으로 추출하여 문서 요약으로 사용하였다. [7]은 K 평균과 비음수 행렬 분해를 이용하여 주제 기반의 다중 문서를 요약하는 방법을 제안하였다. 비음수 행렬 분해를 이용해 가중치가 부여된 용어-문장 행렬을 의미 특징 행렬과 비음수 변수 행렬로 분해함으로써 직관적으로 이해할 수 있는 의미적 특징을 추출하고, 주제와 의미 특징간의 유사도에 가중치를 부여해 유사도는 높으나 실제 의미 없는 문장이 추출되는 것을 막았다. 또한 K 평균 군집을 이용해 문장에 포함된 잡음을 제거하여 문서의 의미가 요약에 편향되게 반영하는 것을 피했다. [8]은 비음수 행렬분해와 군집의 정제 방법을 이용한 문서 군집을 제안하였다. 이 방법에서는 비음수 행렬 분해의 유사한 문서 집합을 구분하지 못하는 문제를 해결하기 위해 군집 후, 군집 내의 유사도를 이용하여 재 군집하는 방법을 제안하였다. [9]는 SVD (Singular Value Decomposition)를 사용하여 단어와 문서에서 의미가 있는 레이블을 추출하여 군집화하는 방법을 제안하였다. 그러나 이 방법은 데이터의 양이 많은 경우 시간이 오래 걸리는 단점이 있다.

2.1.1. 주성분 분석

주성분 분석은 고차원 입력 벡터를 저차원의 벡터로

표현하여 몇 개의 주성분 값으로 나타내는 다변량 데이터 처리 방법 중의 하나이다[2,3]. *n*차원의 벡터 x=[x1 x2 ... xn]T가 있을 때 식 (1)과 식(2)를 적용해 나온 평균 벡터와 공분산 행렬을 통해 고유벡터를 구한 뒤에 대응되는 고유값의 크기에 따라 고유벡터를 정렬하여 새로운 행렬 A를 만든다.

$$m_x = \frac{1}{M}\sum_{k=1}^{M} x_k \quad (1)$$

$$Cx = \frac{1}{M}\sum_{k=1}^{M} x_k x_k^T - m_k m_k^T \quad (2)$$

이 새로운 행렬 A를 변환 행렬로 사용해 식 (3)과 같이 벡터 x를 벡터 y로 변환하면, y의 열에 있는 새 변수들 y1 y2 ... y*n*은 비상관성을 가지며 단조 감소 분산 순서로 배열되어 분산 값이 큰 주성분들로 차원을 줄일 수 있다.

$$y = A(x - m_x) \quad (3)$$

Ⅲ. 주성분 분석과 K 평균 알고리즘을 통한 문서 군집화

본 논문에서 제안한 문서 군집 과정은 그림 1과 같이 전처리, 시드점 추출, 문서 군집으로 구성된다. 전처리 단계에서는 문서 집합을 전처리하여 문서-단어 행렬을 구성한다. 문서-단어 행렬에 대응하는 특징 벡터들에 주성분 분석을 통한 시드점들을 찾은 후 K 평균 알고리즘을 적용하여 문서들을 군집화한다.

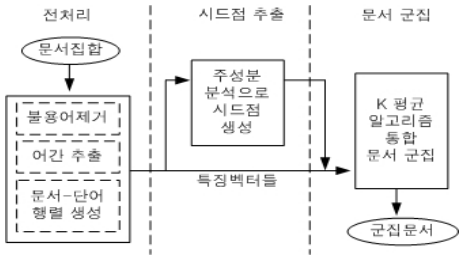

그림 1. 문서 군집 과정
Fig. 1 Process of Document Clustering

3.1. 전처리

전처리 단계는 주어진 문서 집합으로부터 불용어 제거, 어간 추출, 단어의 가중치 등을 생성한다. 어간 추출은 Porter의 어간 추출 알고리즘을 이용한다[10]. 문서 군집화에서 단어의 가중치는 많은 변형이 존재하지만 대부분 문서 내 단어 빈도수(TF: Term Frequency)와 역문헌 빈도수(IDF: Inverse Document Frequency)의 조합에 기반한다[11]. 다음 표 1은 8개의 문서와 10개의 단어로 구성된 문서-단어 행렬로 단어 가중치로는 단어 출현수를 나타낸다. 여기서 (A1, A2, A3), (B1, B2), (C1, C2, C3)는 각각 유사한 문서들로 군집의 대상이 된다.

표 1. 문서-단어 행렬
Table. 1 Document-Word Matrix

	t1	t2	t3	t4	t5	t6	t7	t8	t9	t10
A1	2	0	3	5	4	1	0	0	0	0
A2	0	1	2	4	2	0	0	1	0	0
A3	1	1	1	6	3	0	0	0	0	0
B1	0	0	0	0	1	4	4	0	0	0
B2	0	0	0	0	2	3	4	0	1	0
C1	1	0	0	0	0	0	0	1	5	1
C2	0	0	0	0	0	0	1	2	4	1
C3	0	0	1	0	0	0	0	1	3	1

3.2. 주성분 분석을 이용한 시드점 추출

문서들을 문서-단어 행렬로 표현할 때 단어들을 특징 벡터 공간상의 기준 축들로 하면, 문서는 특징 벡터 공간 상의 벡터 x=[x1 x2 ... x*n*]T 로 표현된다. 예를 들어, 표 1의 A1 문서는 (2, 0, 3, ..., 0, 0, 0)의 벡터로, A2 문서는 (0, 1, 2, ..., 1, 0, 0)의 벡터로 표현된다. 따라서 본 논문에서는 이러한 특징 벡터 공간 상의 문서들을 K 평균 알고리즘을 사용하여 군집화한다. 일반적인 K 평균 알고리즘에서는 시드점이라고 불리는 초기의 K 개 패턴 즉 벡터들이 무작위로 선택되지만, 군집 구조에 관한 어떤 지식을 사용해 적절한 시드점을 구할 수 있다면 더 나은 성능을 얻을 수 있다[2,3]. 표 1의 문서-단어 행렬에서, (A1, A2, A3), (B1, B2), (C1, C2, C3)의 각 그룹 문서들은 특징 벡터 공간상에서 서로 다른 군집을 형성함을 알 수 있다. 이처럼 각 그룹이 서로 다른 군집을 형성할 때 가장 적절한 시드점들은 각 그룹에서 한 개씩의 시드점을 선택하는 것이다. 예를 들어, 표 1의 경우 A2, B1, C2 문서를 3개의 시드점으로 선택하면

무작위 시드점으로 K 평균 알고리즘을 수행하는 것보다 좋은 결과를 얻을 수 있다. 본 연구에서는 주성분 분석을 통해 이러한 적절한 시드점을 구하는 방법을 제안한다.

첫 번째는 문서들의 특징 벡터 공간 특성을 잘 표현할 수 있는 몇 개의 대표 문서들을 선택해 시드점으로 사용하는 방법이다. 주성분 분석은 다차원 자료를 선형 변환시켜서 서로 상관되지 않는 새로운 인공 자료들인 주성분을 유도한다. 이 때 소수 몇 개의 주성분이 원래 자료에 내재하는 전체 변이 중 가능한 많은 부분을 보유하도록 변환시킴으로써 정보의 손실을 최소화하며 차원의 축약을 기할 수 있다. 결국, 주성분 분석을 이용한다면 정보의 손실을 최소화하면서 소수의 몇 개 문서로 문서 집합의 단어들을 표현할 수 있다. 즉 높은 값을 가지는 주성분과 일치하는 대표 문서들은 손실을 최소화하면서 대부분의 단어들을 표현할 수 있기에, 이들을 군집 시드로 사용할 수 있다. 대표 문서들을 찾기 위해 표 1에서 문서들을 기준 축으로 하는 특징 벡터 공간 상에서 주성분 분석을 적용한 결과는 표 2와 같다. 표 2의 경우 3개 군집화가 요구되므로 3개의 시드점을 구해야 한다. 표 2의 주성분 열(P1~P8)은 고유값이 적은 즉, 분산 성분이 줄어드는 순서로 정렬 되어 있으므로, P1 열에서 가장 절대값이 큰 -0.572 즉 A3 문서, P2 열에서 가장 절대값이 큰 0.626 즉 B1 문서, P3 열에서 가장 절대값이 큰 0.490 즉 C1 문서를 각각 대표 문서 즉, 시드점으로 선택한다. 주성분 분석을 통하여 A, B, C의 각 그룹에서 한 문서씩이 시드점으로 선택되었음을 알 수 있다. 이것은 각 그룹에서 하나씩의 문서가 선택될 때, 이 문서들이 대부분의 단어 들을 포함할 수 있기 때문이다.

표 2. 표 1의 주성분 분석 결과
Table. 2 Table 1's PCA Result

	P1	P2	P3	P4	P5	P6	P7	P8
A1	-0.560	0.006	0.290	0.694	0.185	-0.069	0.284	0.022
A2	-0.386	-0.090	0.121	-0.414	0.581	0.103	-0.184	-0.526
A3	-0.572	-0.074	0.324	-0.429	-0.504	0.073	-0.090	0.332
B1	0.147	0.626	0.311	-0.061	0.049	0.655	0.234	0.006
B2	0.140	0.501	0.474	0.001	0.021	-0.601	-0.376	-0.044
C1	0.266	-0.425	0.490	0.224	-0.370	0.230	-0.180	-0.486
C2	0.270	-0.292	0.420	-0.312	0.217	-0.262	0.655	0.152
C3	0.158	-0.281	0.242	0.112	0.433	0.258	-0.469	0.592

두 번째는 특징 벡터 공간상에서 각 그룹의 크기가 비슷할 때, 서로 간에 거리가 먼 벡터들을 시드점으로 사용하는 방법이다. 그러나 n 차원 특징 벡터 공간상에서 서로 간에 거리가 먼 m개의 벡터들을 찾기 위해서는 모든 벡터들 간의 거리를 조사해야 하는 문제점이 있다. 주성분 분석은 고차원 입력 벡터를 저차원의 벡터로 표현할 수 있다. 따라서 본 논문에서는 주성분 분석을 통하여 n 차원 상의 벡터들을 1차원 상의 점들로 변환한 후, 서로 간에 거리가 떨어진 m개의 점들을 찾는 방법을 사용한다. 1차원 상의 점들 가운데 서로 간에 거리가 떨어진 m개의 점들을 찾기 위해, 본 연구에서는 모든 점들을 정렬한 후 m 균등 분할하여 각 분할의 중심이 되는 m개의 점을 구한다. 예를 들어, 표 1의 10차원 특징 벡터 공간 상의 8개 문서 벡터들을 주성분 분석을 통해 1차원 상의 8개 점들로 변환하면 -4.924, -3.017, -4.806, 1.909, 1.864, 3.318, 3.321, 2.336이 된다. 이 점들을 정렬한 후 3 균등 분할하여 각 분할의 중심점을 구하면 -4.806, 1.909, 3.321이 되며, 이 점들은 각각 A2, B2, C3로 각 그룹에서 한 개씩의 문서가 시드점으로 선택됨을 알 수 있다.

Ⅳ. 실험 및 평가

본 연구에서는 제안하는 방법의 군집 성공률을 측정하기 위해 20 Newsgroups 문서 자료[12] 중 일부를 사용하였다. 20 Newsgroups은 20개의 다양한 주제로 구성되어 있으나, 이 들 중 종교, 컴퓨터 윈도우, 우주 과학, 오토바이, 스포츠 야구, 자동차, 의학, 총기, 정치의 9개 뉴스 그룹을 선택하고, 각 뉴스 그룹의 100개의 문서 들 중에서 무작위로 50개씩의 문서를 사용하였다. 문서 전처리를 통한 문서 단어 행렬에서 단어의 가중치로는 단어-역문헌 빈도수(TF-IDF)를 사용하였다. 표 3은 형성되는 군집의 수를 다르게 했을 때 주성분 분석을 통해 적절한 시드점을 찾는 비율과 K 평균 알고리즘의 수행 결과를 보여 준다. 표 3에서 적절한 시드점이란 모든 군집에서 한 문서씩을 시드점으로 선택하는 경우를 의미한다. 예를 들어, 4개의 군집에서 4개의 시드점을 찾을 때 3개의 군집에서만 시드점을 찾는다면 적절한 시드점을 찾는 비율은 75%가 된다. 표 3에서 보듯이 방법 1의 경우 군집 수가 많아 질수록 적절한 시드점을

찾는 비율이 낮아지는 것을 알 수 있다. 즉 군집 수가 많아 질수록 일부 군집에서만 시드점들이 선택되는 경우가 발생한다. 이것은 군집 수가 많아 질수록 다양한 문서들이 사용되어 같은 단어를 공유하는 문서들도 많아지기 때문에, 일부 군집의 문서들이 시드점으로 선택되더라도 문서 집단에 존재하는 대부분의 단어들을 포함할 수 있기 때문이다. 반면 방법 2의 경우 좀 더 적절한 시드점을 찾는 것을 알 수 있다. 이것은 각 군집의 크기가 비슷할 때 주성분을 적용하여 문서 벡터들을 1차원 상의 점들을 변환하여 균등 분할하면, 주제와 상관없는 문서 데이터들이 있는 경우를 제외하고는 각 그룹의 시드점을 제대로 찾을 수 있기 때문이다. 군집 성공 비율은 주성분 분석 기법의 시드점들을 사용한 K 평균 알고리즘이 무작위 시드점들을 사용한 K 평균 알고리즘보다 좋은 것을 알 수 있으며, 방법 2는 방법 1보다 좀 더 적절한 시드점을 찾기 때문에 군집 성공률도 보다 좋은 것을 알 수 있다.

표 3. 군집수에 따른 실험 결과
Table. 3 Experimental Result by Number of Groups

군집 수	무작위 군집 성공률	방법 1 적절한 시드점 비율	방법 1 군집 성공률	방법 2 적절한 시드점 비율	방법 2 군집 성공률
2	75	90	98	100	100
3	78	88	96	100	98
4	68	83	75	100	98
5	57	78	72	96	86
6	56	75	67	92	82
7	62	69	66	92	82
8	60	71	70	92	84
9	56	63	65	90	77

Ⅴ. 결 론

문서 군집은 문서간의 유사도를 바탕으로 서로 연관된 문서들을 군집화하여 대용량의 문서들을 자동으로 분류하고 검색하고 처리하는데 효율과 정확성을 증대시키는 방법이다. 단어들을 포함하는 문서는 특징 벡터 공간 상의 벡터로 표현할 수 있기에 K 평균 알고리즘을 사용하여 문서들을 군집화할 수 있다. 기존의 K 평균 알고리즘은 무작위로 시드점을 선택하기 때문에, 잘못된 시드점을 선택하는 경우 성능이 좋지 않았다. 따라서 본 연구에서는 주성분 분석을 적용하여 적절한 시드점을 찾아 효율적으로 군집화하는 방법을 제안하였다. 실험을 통해 무작위 K 평균 알고리즘보다 주성분 분석으로 적절한 시드점을 통한 K 평균 알고리즘이 더 좋은 군집화 성능을 얻음을 보였다. 추후 과제로는 군집 크기의 변이가 큰 경우에도 적절한 시드점을 찾는 연구가 필요하다.

감사의 글

본 연구는 2014년도 광운대학교 교내 연구비 지원에 의하여 이루어진 연구로서, 관계부처에 감사드립니다.

REFERENCES

[1] C. Park, Y. Kim, J. Kim, J. Song, and H. Choi, *Data Mining using R,* Kyowoosa, 2011.

[2] H. Park, and K. Lee, *Pattern Recognition and Machine Learning from Basic to Application,* Leehan Pub., 2011.

[3] L. Oh, *Pattern Recognition,* Kyobo Book Centre, 2010.

[4] S. Park, D. An, "Document Clustering Method using PCA and Fuzzy Association," *Journal of Korea Information Processing Society B,* 2010.

[5] C. Lee, M. Kim, K. Lee, G. Lee, H. Park, "Document Thematic words Extraction using Principal Component Analysis," *Journal of the Korea Society of Computer and Information B,* 2002.

[6] C. Lee, M. Kim, J. Paik, H. Park, "Text Summarization using PCA and SVD," *Journal of Korea Information Processing Society B,* 2003.

[7] S. Park, J. Lee, "Topic-basied Multi-document Summarization Using Non-negative Matrix Factorization and K-means," *Journal of the Korea Society of Computer and Information B,* 2008.

[8] S. Park, D. U. An, B. R. Char, and C. W. Kim, "Document Clustering with Cluster Refinement and Non-negative Matrix Factorization," In *Proceeding of ICONIP'09,* 2009.

[9] S. Osinski and D. Weiss, "Conceptua Clustering using lingo algorithm: Evaluation on open directory project data,"

in *Proc. IIPWM04,* 2004.
[10] The Porter Stemming Algorithm. Available: http://tartarus.org/~martin/PorterStemmer/
[11] B. Lee, *Information Retrieval,* Green Pub. 2012.
[12] http://qwone.com/~jason/ 20Newsgroups/

김우생(Woosaeng Kim)

1985년 텍사스 주립대학 전산학과 졸업 (학사)
1991년 미네소타 주립대학 전산학과 졸업(박사)
1987년 ~ 1988년 현대전자
2001년 UC 버클리 대학 교환 교수
1992년 ~ 현재 광운대학교 컴퓨터소프트웨어 학과 교수
※관심분야 : 데이터베이스, 멀티미디어

김수영(Sooyoung Kim)

한동대학교 전산전자공학부 재학 중
※관심분야 : 데이터베이스