클러스터 중심 결정 방법에 따른 문서 클러스터링 성능 분석

오형진, 변동률, 이신원* 박순철, 정성종, 안동언
전북대학교 컴퓨터공학과, *정인대학
전화 : 063-270-2416 / 핸드폰 : 016-9860-2489

Analysis of Document Clustering Varing Cluster Centroid Decisions

Hyung Jin Oh, Dong Ryul Byun, Shin Won Lee*, Soon Chul Park, Sung Jong Chung, Dong Un An.
Dept of Electronics and Information Engineering of Conbuk National University, Chongin College
E-mail : hyungjin@duan.chonbuk.ac.kr

Abstract

K-means clustering algorithm is a very popular clustering technique, which is used in the field of information retrieval. In this paper, We deal with the problem of K-means Algorithm from the view of creating the centroids and suggest a method reflecting document feature and considering the context of each document to determine the new centroids during the process of forming new centroids. For experiment, We used the automatic document summarizer to summarize the Reuter21578 newslire test dataset and achieved 20% improved results to the recall metrics.

I. 서론

정보검색 시스템에서 문서 클러스터링 기법은 사용자 질의에 대하여 검색된 문서를 문서간의 유사도에 따라 클러스터로 구성하고 검색결과로 사용자에게 제시하여 주는 것이다.

기존의 정보검색 시스템은 사용자가 적합한 문서를 꼼꼼하게 찾아야 하는 긴 검색결과 리스트를 제시한다. 따라서 사용자는 찾고자 하는 문서를 빠르게 찾아주는 방법외에 검색결과로 나온 문서 집합의 의미를 쉽게 파악할 수 있는 방법에 대한 연구가 이루어져야 한다. 왜냐하면 다양한 주제에 관련된 텍스트를 검색하고 조직화 하는 것은 많은 시간과 노력을 필요로 하는 작업이기 때문이다.[1][2][3][4]

문서 클러스터링에 영향을 가장 많이 미치는 요소는 초기 클러스터 선택, 클러스터링 과정에서 발생하는 클러스터 중심값 결정, 문서를 대표하는 색인어에 가중치 부여 문제등이 있다. [1][3]

본 논문에서는 클러스터의 중심을 결정하는 문제에 대하여 기존의 k-means 알고리즘과 제안하는 방법을 사용하며, 자동 문서 요약기를 사용하여 전문을 요약한 요약 문서를 클러스터링의 결과를 논의 한다.

본 논문의 구성은 다음과 같다. 2장에서는 관련연구에 대하여 살펴보고, 3장에서는 기존의 k-means 알고리즘을 설명하겠다. 4장에서는 제안하는 클러스터 중심을 계산하는 방법을 논의 한다. 5장에서는 자동문서 요약기의 출력, passage(문장내 명사 수, 5, 10, 15) 수를 변화시키면서 수행한 클러스터링 결과를 비교 분석하고, 끝으로 6장에서는 결론을 맺는다.

II. 관련연구

문서 클러스터링은 정보 검색의 효율성과 유효성을 증대시키기 위한 목적으로 사용한다. 대표적인 문서 클러스터링의 방법론은 클러스터링의 결과로 생성되는 그룹의 구조에 따라서 계층적 클러스터링(hierarchical clustering method)과 비계층적 클러스터링(non-hierarchical clustering method)으로 나눌 수 있는데 각각의 방법론에 따라 여러 가지 구현 알고리즘이 있다.[2]

비계층적 클러스터링은 입력되는 문서의 순서에 따라 클러스터링 결과가 달라지는 단일 처리 방법(single pass method)과 이의 단점을 보완한 재배치 방법(reallocation method)이 있다. 계층적 클러스터링은 문서간의 유사도 정보를 토대로 단계적으로 계층적인 클러스터를 형성하는 방법으로 응집 알고리즘(agglomerative method)과 분할 알고리즘(divisive method)이 있다. 계층적 응집 알고리즘에는 단일 링크 방법(single link method), 완전 링크 방법(complete link method), 그룹 평균 연결 방법(group average link method)등이 있다.[6]

III. 클러스터 중심 결정 방법에 따른 문서 클러스터링

3.1 k-means 알고리즘

문서 클러스터링의 성능은 어떤 클러스터링 기법을 사용하여 클러스터링을 하였는지에 따라 다른 결과를 나타낸다. 본 논문에서는 비계층적 클러스터링 방법에서 널리 사용하는 K-means 알고리즘을 사용한다.

1. K값 클러스터 개수를 구한다.
2. K개의 초기 중심값(proto-centroids)을 구한다.
3. 각 문서(d)들과 중심값(c) 사이의 거리를 구한다.

$$dist(\overline{d_i}, \overline{c_j}) = \sqrt{\sum_{k=1} (d_{ki} - c_{kj})^2},$$

$i = 1, 2, \ldots n$ n: 전체문서개수
$j = 1, 2, \ldots K$ k: centroid의개수

4. 문서를 가장 짧은 거리의 중심값에 할당한다.

$$\arg\min dist(\overline{d_i}, \overline{c_j})$$
$$i = 1, \ldots n \ , \ j = 1, \ldots k$$

$d_i \in G_c$, if $dist(\overline{d_i}, \overline{c_j}) < dist(\overline{d_i}, \overline{c_l})$
(for all $l = 1, 2, \ldots, k \ \ l \neq j$)

5. 새로운 클러스터 중심값을 재계산 한다..

$$\overline{c_j} = \frac{1}{|c_j|} \sum_{l=1}^{|c_j|} \overline{d_l}$$

6. 이전의 중심값과 새로운 중심값을 비교하여 벡터간 차이가 거의 없을 때까지 반복한다.

If $\max \delta(\overline{c_j^{old}}, \overline{c_j^{new}}) < \theta$ *then return*
else goto 3

K-means 알고리즘

클러스터 중심은 클러스터에 포함되어 있는 문서들의 특성을 나타내기 위하여 사용하는데, 단어와 가중치의 쌍으로 이루어진 벡터로 표현한다. k-means 알고리즘에서의 클러스터 중심은 식 1)과 같다.[1]

$$\overline{c_j} = \frac{1}{|c_j|} \sum_{l=1}^{|c_j|} \overline{d_l} \qquad (1)$$

클러스터 중심을 L-차원의 공간에서 벡터 $(x_{i1}, x_{i2}, \ldots, x_{iL})$로 표현하였을 때 클러스터에 속하는 문서를 대표하는 색인어와 가중치만을 단순히 하나의 클러스터 벡터로 머지(merge)하였다.

3.2 제안하는 클러스터 중심값 계산 방법

본 논문에서 제안하는 클러스터 중심 계산 방법은 클러스터에 포함된 모든 문서들이 갖는 단어의 가중치의 평균으로 계산한다. 클러스터 중심 Ci와 문서 dj가 병합되어서 생성된 클러스터 중심은 식 2)와 같이 계산한다.

$$c_i^{new} = \frac{m_i \cdot C_i + m_j \cdot d_j}{m_i + m_j} \qquad (2)$$

i : cluster number
j : number of allocated documents to the current cluster.

새롭게 생성된 클러스터 중심은 클러스터에 속하는 문서들은 클러스터 중심을 형성하는 과정에서 문서에 나타난 단어들의 가중치로 자신들의 특성을 반영한 것이다. 따라서 클러스터에 속해있는 문서들은 클러스터 중심을 통해서 서로 다른 문서들에 영향을 주게 되어 문맥을 고려한 클러스터링 효과를 얻을 수 있다.

Ⅳ. 실험 및 결과 분석

본 논문의 실험을 위해서 자동 문서 요약기를 거쳐 출력한 데이터인, passage(문장내 명사) 수(5, 10, 15)를 변화시키면서 요약된 문서를 사용한다.

4.1 실험 데이타

본 논문에서 제안하는 클러스터 중심값 계산 방법을 평가하기 위하여 사용한 실험 문서는 Reuter21578 newswire이다. 클러스터링 효과를 평가하기 위하여 정확률 척도를 사용하는데 각 Reuter21578 newswire문서는 각 문서가 할당되어야 하는 TOPIC 태그를 가지고 있다.[8] 자주 등장하는 TOPIC 10개와 각 TOPIC당 문서 10개씩, 총 100개의 문서를 실험 문서로 선택하였으며, TOPIC 1번에 해당하는 문서번호는 1~10, TOPIC 2번에 해당하는 문서는 11~20, TOPIC 10번에 해당하는 문서는 91~100번 문서이다. 실험 문서는 자동 문서 요약기를 사용하여 문서에서 중요한 문장 5라인씩 패시지(passage) 수를 달리하면서 (5, 10, 15passages) 기존의 k-means 방법론과 제안하는 모델의 성능을 평가하였다.[8]

표 1. Reuter21578 newswire에서 자주 등장하는 TOPIC

TOPIC1	EARN	TOPIC6	TRADE
TOPIC2	ACQ	TOPIC7	INTEREST
TOPIC3	MONEY-FX	TOPIC8	GNP
TOPIC4	GRAIN	TOPIC9	WHEAT
TOPIC5	CRUDE	TOPIC10	SHIP

4.2 실험 데이터의 특성

자동 문서 요약기를 통한 Reuter21578 문서가 가지는 특성은 다음과 같다.

- 문서를 대표하는 평균단어수가 5, 10, 15passage당 각각 30, 54, 65이다.
- 한 문서에서 단어의 출현 빈도는 아주 작다
- 같은 토픽내의 문서에서 출현하는 단어는 유사하다.

문서 클러스터링을 할 때에 문서 전문을 이용하지 않고 자동 문서 요약기를 사용한 이유는 클러스터링의 속도 뿐만 아니라 요약문서의 위와 같은 특성 때문에 각 클러스터에 해당하는 TOPIC 문서 문서들을 대표하는 단어들이 유사한 의미를 갖추고 있고 같은 토픽내의 문서는 의미상 분포 형태가 밀접하다고 할 수 있다.[3]

4.3 결과 및 분석

그림 1과 2는 요약문의 출력 결과를 10개의 passage를 사용하였을 때 기존의 k-means 방법론에서와 제안한 방법론을 실험한 결과이다.

클러스터 수는 10개(k=10)를 선택하였으며 cid1은 TOPIC 1번인 EARN, cid2는 TOPIC 2번인 ACQ, cid10은 TOPIC 10번인 SHIP을 대표하며 해당 클러스

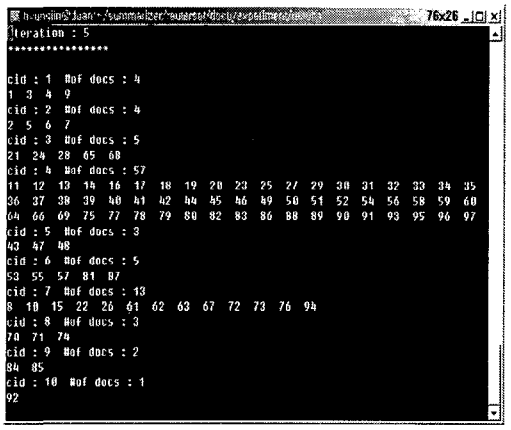

그림 3 k-means 알고리즘

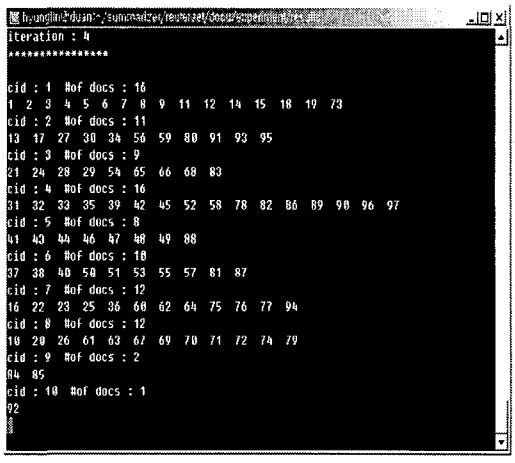

그림 4 제안한 방법론의 클러스터중심 결정 - 10passage

터에 할당된 문서수와 할당된 문서번호를 나타낸다.

그림 1과 그림 2를 비교해보면 기존의 k-means 알고리즘에서는 발생하는 한 클러스터에 할당된 문서가 많은 클러스터가 존재하는 사실을 발견할 수 있으며, 제안하는 클러스터 중심을 결정하는 방법이 기존의 k-means 방법보다 클러스터에 할당된 문서수에서 균등하게 할당되었으며 재현률 측면에서 더 나은 결과를 나타남을 알 수 있다.

표 2. passage 5인 경우(%)

kmeans	정확률	재현률	제안 방법	정확률	재현률
cid 1	100	10	cid 1	60	60
cid 2	75	30	cid 2	75	60
cid 3	18	30	cid 3	15	20
cid 4	67	20	cid 4	100	30
cid 5	18	70	cid 5	100	70
cid 6	40	80	cid 6	78	70
cid 7	25	20	cid 7	31	40
cid 8	100	10	cid 8	50	70
cid 9	60	30	cid 9	40	60
cid 10	100	20	cid 10	75	60

표 3. passage 10인 경우(%)

kmeans	정확률	재현률	제안 방법	정확률	재현률
cid 1	40	40	cid 1	56	90
cid 2	0	0	cid 2	18	20
cid 3	60	30	cid 3	44	40
cid 4	18	100	cid 4	31	50
cid 5	100	30	cid 5	87	70
cid 6	60	30	cid 6	40	40
cid 7	30	40	cid 7	29	20
cid 8	100	30	cid 8	33	40
cid 9	100	20	cid 9	100	20
cid 10	100	10	cid 10	100	10

표 4. passage 15인 경우(%)

kmeans	정확률	재현률	제안 방법	정확률	재현률
cid 1	100	20	cid 1	62	80
cid 2	100	20	cid 2	83	50
cid 3	50	10	cid 3	36	50
cid 4	14	10	cid 4	100	30
cid 5	100	10	cid 5	56	90
cid 6	100	10	cid 6	80	80
cid 7	12	90	cid 7	17	10
cid 8	100	10	cid 8	50	40
cid 9	100	10	cid 9	70	90
cid 10	100	10	cid 10	100	40

표 5에서 볼 수 있듯이 평균 정확률 측면에서는 기존의 k-means 알고리즘이 본 논문에서 제안한 기법보다 다소 높게 나타났지만 각 클러스터에 할당된 문서수가 적기 때문이며 평균 재현률 측면에서는 제안한 기법이 20%이상 좋은 성능을 보이고 있다. 즉, 클러스터링의 효과가 우수함을 알 수 있다.

표 5. passage수 변화에 대한 평균 정확률, 재현률(%)

kmeans	평균정확률	평균재현률	제안 방법	평균정확률	평균재현률
5passage	60.3	32	5passage	65.4	54
10passage	60.8	33	10passage	53.8	40
15passage	77.6	20	15passage	65.4	56

V. 결론

본 논문에서는 k-means 알고리즘에서의 클러스터 중심 생성 방법과 클러스터에 속하는 모든 문서들의 가중치를 평균하는 방법을 비교 하였고 제안하는 기법이 평균 정확률 측면보다는 평균 재현률 측면에서 20%이상 좋은 성능을 보이고 있음을 알 수 있는데 그 이유는 클러스터에 속해있는 문서들은 클러스터 중심을 통해서 서로 다른 문서들에 영향을 주게 되어 문맥을 고려한 클러스터링 효과를 얻을 수 있기 때문이다.

k-means 알고리즘의 대표적인 문제는 초기 중심값 선택에 따라 클러스터링에 영향을 미치는 결과가 매우 다름을 실험을 통하여 발견하였으며 이를 개선하는 연구가 필요하다.

참고 문헌

[1] Tapas Kanung, "The Analysis of a Simple k-Means Clustering Algorithm".

[2] Qin He, 'A Review of Clustering Algorithms as Applied in IR", UIUCLIS--1999/6+IRG.

[3] 오형진 외 ,"요약 문서 기반 문서 클러스터링", 전북대학교 컴퓨터공학과, 정보처리학회지 제 9권 pp.589-592, 2002.4.

[4] 고지현 외, "LSI를 이용한 가중치 변화에 따른 클러스터링 결과 분석", 전북대학교 정보통신공학과, 정보처리학회지 제 9권, pp. 1009-1012, 2002.4.

[5] 김금영 외, "질의기반 자동문서 요약", 전북대학교 컴퓨터공학과, 정보처리학회지 제 9권, pp.593-596. 2002.4.

[6] 이문기 외, '웹 디렉토리 서비스를 위한 문서 클러스터링', 포항공과대학교 컴퓨터공학과, pp.351-351.

[7] 김명철 외, "최신 정보 검색론", 홍릉과학출판사. 2001.1

[8] http://www.research.att.com/~lewis/reuters21578.html