문서 수에 따른 가중치를 적용한 K-means 문서 클러스터링

조시성, 안동언, 정성종, 이신원*
*정인대학 컴퓨터정보학과
전북대학교 컴퓨터공학과
e-mail : santted@duan.chonbuk.ac.kr

K-means Clustering Method according to Documentation Numbers

Cea-Sung Cho, Dong-Un An, Sung-Jong Jeong, Shin-Won Lee*
*Dept. of Computer Information, ChongIn College
Dept. of Computer Engineering, Chonbuk National University

요 약

본 논문에서는 이 문서 클러스터링 방법 중 계층적 방법인 Kmeans 클러스터링 알고리즘을 이용하여 문서를 클러스터링 하고자 한다. 기존의 Kmeans 클러스터링 알고리즘은 문서의 수가 많을 경우 하나의 클러스터링에 너무 많은 문서들이 할당되는 문제점이 있다. 이 치우침을 완화하고자 각 클러스터링에 할당된 문서 수에 따라서 문서에 가중치를 부여한 후 다시 클러스터링을 하는 방법을 제안하였다. 실험 결과는 정확률, 재현율을 결합한 조화 평균(F-measure)를 사용하여 평가하였으며 기존 알고리즘보다 9%이상의 성능 향상을 나타냈다.

1 서론

문서 클러스터링은 다량의 문서를 특정 주제 아래 자동 분류하는 것으로써 사용자가 특정 정보에 대한 검색 요구를 하였을 때 모든 문서를 검색하는 대신 사용자의 요구와 가장 가까운 주제의 클러스터 내의 문서만을 검색함으로써 탐색 시간을 절약할 수 있고 검색의 효율을 향상시킬 수 있다. 문서 클러스터링 기법은 정보 검색 시스템의 전체 문서 집합을 오프라인에서 미리 클러스터링 하여 질의 요청시 해당 질의와 가장 유사한 클러스터에 대해서만 검색을 수행하는" 전처리 클러스터링 기법" 과 질의 검색 결과를 온라인 상에서 즉시 수행하는 " 후처리 클러스터링" 으로 나눌 수 있다.[5] 본 논문에서는 후처리 기법 중 성능이 좋은 알고리즘과 이 알고리즘의 단점을 보완한 알고리즘을 실험하여 비교 평가하고자 한다.

본 논문의 구성은 다음과 같다. 2 장에서는 관련 연구를 살펴보고 3 장에서는 비계층적 클러스터링 알고리즘인 M-Kmeans 알고리즘과[3] 성능을 개선한 R-Kmeans 알고리즘에 대해서 살펴보고, 4 장에서는 실험 데이터 및 실험 결과와 분석을 기술한다. 마지막으로 5 장에서는 결론 및 향후 연구 과제에 대하여 논한다.

2 관련 연구

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.

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.[4]

3 M-Kmeans Algorithm 과 R-Kmeans Algorithm

3.1 M-Kmeans Algorithm (Modified Kmeans)

M-KMeans 알고리즘은 비계층적 클러스터링 기법으로 문서와 클러스터의 중심값을 나타내는 centroid 와의 유사도를 측정하여 문서를 적합한 클러스터에 재배치하는 방법이다. 여기에서 centroid 는 클러스터에 속하는 문서들의 평균 벡터값을 이용한다

K-Means Algorithm 은 다음과 같다..

1. 클러스터 개수 K 를 선택한다
2. K 개의 초기 중심을 구한다.

$$c_i^{initial} = \sum_{j}^{m=3} \overrightarrow{d_l}$$

3. 각 문서(d)들과 K 개의 중심(c)와의 거리를 구한다..

$$\arg\min_{\substack{i=1,n \\ j=1,k}} dist(\overrightarrow{d_i}, \overrightarrow{c_j})$$

$$C_i^{new} = \frac{(m_i c_i + m_{ij} d_{ij})}{m_i + m_{ij}}$$

c_i : i 번째 클러스터 벡터
d_{ij} : j 번째 클러스터에 할당된 j 번째 문서 벡터
m_i : i 번째 클러스터의 크기
m_{ij} : i 번째 클러스터에 할당된 j 번째 문서의 크기
C_i^{new} : i 번째 새로운 클러스터 중심 벡터

4. 문서를 가장 짧은 거리의 중심에 할당한다.

$$\arg\min dist(\overline{d}_i, \overline{c}_i), i = 1...n, j = 1...k$$

$$d_i \in_{i=1,n} G_{c_j} \quad if\ dist(d_i, c_j) < dist(d_i, c_l)$$

$$for\ all\ l = 1, 2, \cdots, k \quad l \neq j$$

5. 클러스터 중심을 재계산한다

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

6. 새로 생성된 클러스터 중심과 이전에 생성된 클러스터 중심과의 거리가 임의의 값 이상이면 3 으로 가서 반복한다

$$if\ \max \delta(\overrightarrow{c_j^{old}}, \overrightarrow{c_j^{new}}) < \theta\ then\ return$$

$$else\ goto\ 3$$

M-KMeans Algorithm 은 특성상 생성된 클러스터 중심에 따라 클러스터링 결과가 달라진다[8][10]. 특히 초기 클러스터중심을 어떻게 선택하는가에 따라 빠른 시간에 최적의 클러스터링 결과가 나오는 경우와 그렇지 않은 경우가 존재한다. M-KMeans 알고리즘에서는 특정 문헌 집합에 속하는 임의의 한 개의 문서를 선택하는 대신 색인어와 가중치로 표현되는 문서를 3 개(m=3)로 선택하여 중복된 색인어를 제외하고 병합한 후 초기 클러스터 중심 벡터로 설정하였다. 식은 알고리즘 3.1 의 2 번 식과 같다.[3][4]

클러스터링에 영향을 미치는 또 다른 요소는 클러스터링 과정에서 발생하는 새로운 클러스터 중심(cluster Centroid)를 결정하는 것이다. M-KMeans Algorithm 에서의 새로운 클러스터 중심 벡터는 포함된 모든 문서들이 갖는 색인어의 가중치의 평균으로 계산한다. 클러스터 중심 Ci 와 문서 di 가 병합되어서 생성된 클러스터 중심은 알고리즘 3.1 의 3 번이다.

기존의 Kmeans Algorithm 보다 좋은 결과를 보였지만,[3] M-Kmeans 알고리즘 역시 하나의 클러스터에 너무 많은 문서가 할당되었다.

3.2 R-Kmeans Algorithm (Revised Kmeans)

이 알고리즘은 기존 M-Kmeans 알고리즘 단점의 해법을 제안하였다. 3.1 의 알고리즘에 의해 얻어진 각 클러스터로부터 문서 수를 구한 후, 문서 수에 따른 임의의 가중치를 부여하여 클러스터링을 하는 방법이다.

R-Kmeans Algorithm 은 다음과 같다.

1. 3.1 에 의해 얻어진 클러스터로부터 각 문서 수를 구한다.
2. 문서 수에 따른 가중치를 부여한다.

```
If (cmcount[] > 20) r_dist[] = 0.15
Else if (cmcount[] > 10 ) r_dist[] = 0.3
Else r_dist[] = 0.5
```

cmcount : 각 클러스터링에 할당된 문서 수
r_dist : 임의의 가중치

3. 3.1 의 알고리즘에 의해 재 클러스터링 한다. (거리 계산시 문서의 거리에 위의 가중치를 결합한다.)

3.3 색인어의 가중치 계산 방법

색인어의 가중치 부여는 문서와 문서를 비교하기 위해서 분류자질, 즉 단어에 적절한 가중치를 부여하는 방법이다. 문서 내용을 설명하는데 같이 사용된 단어라 할지라도 다양한 비중을 가지고 있으며, 단어는 문서 안에서 중요성에 대한 척도로서 문서의 각 단어에 가중치를 부여 해야 한다.

본 논문에서는 색인어에 부여한 가중치는 SMART 시스템에서 사용하고 있는 다양한 가중치 계산 방법을 사용하였다.

SMART 시스템에서 가중치는 세가지 요소에 대한 조합으로 BNN, BNC, BTN, BTC, NNN, NNC, NTN, NTC, ANN, ANC, ATN, ATC, LNN, LNC, LTC 조합이 가능하다 [8]

4. 실험 및 결과

본 논문에서 구현한 전체 시스템은 크게 자동 문서 요약 모듈과 문서 클러스터링 모듈로 구성되어 있다[3], 실험을 위하여 요약문의 색인어에 대하여 SMART 시스템에서 제안한 가중치를 계산하는 방법 16 가지(BNN, BNC, BTN, BTC, NNN, NNC, NTN, NTC, ANN, ANC, ATN, ATC, LNN, LNC, LTN, LTC)를 적용하여 클러스터링을 하였다.

4.1 실험 문서

본 논문에서 사용한 실험 문서는
http://kinds.or.kr
에서 주제별 검색에 의해 수집된 문서이다.

실험 문서는 모두 TOPIC 태그를 가지고 있으며 TOPIC 태그는 문서의 내용을 파악하여 해당 주제에 속하는 문서임을 나타낸다. 선택한 실험 문서는 http://kinds.or.kr에서 가장 많이 존재하는 TOPIC 10 개를 선정하였으며 각 TOPIC 당 문서 10 개씩, 총 100 개의 문서를 실험 문서로 선택하였다.

Topic 1	부동산	Topic 6	보험
Topic 2	증권	Topic 7	선거
Topic 3	노동	Topic 8	역사
Topic 4	패션	Topic 9	요리
Topic 5	물가	Topic 10	유학

표 1 http://kinds.or.kr Topic10

4.2. 실험 결과

아래 그림은 M-Kmeans Algorithm(그림 1, 가중치 LTN)과 문서 가중치를 적용한 R-Kmeans Algorithm(그림 2, 가중치 M-LTN)을 적용하여 가중치 기법에 대한 실험 결과이다. 각 그림에서 cid 1 은 실험 데이터인 http://kinds.or.kr 의 주제별 검색에서 수집한 TOPIC 1 번인 부동산, cid 2 는 TOPIC 2 번인 증권, cid 10 은 TOPIC 10 번인 유학을 대표하는 주제이며 그림에서는 할당된 문서의 수와 문서 번호를 함께 나타내고 있다.

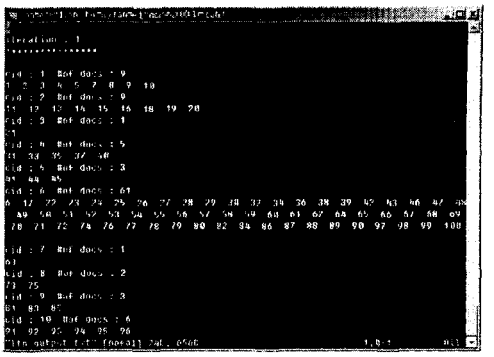

그림 1 M-Kmeans(ltn)

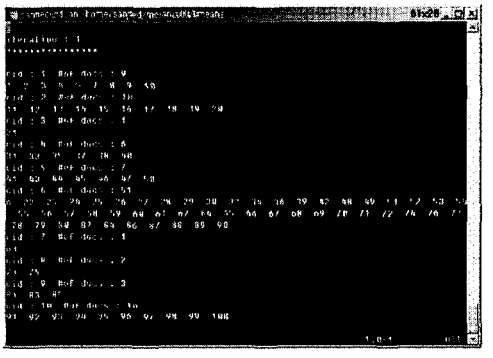

그림 2 R-Kmeans(ltn)

4.3 두 알고리즘의 성능 비교

본 장에서는 M-Kmeans Algorithm 과 제안하는 R-Kmeans Algorithm 의 성능을 비교 분석한다.클러스터링 성능 평가 척도는 클러스터링의 경우에는 생성된 클러스터가 어느 범주에 해당하는지, 또는 특정 문헌이 어느 범주로 자동 분류되었는지를 판정하기가 어렵지만 클러스터의 수인 K 를 10 개로 고정시켜 동일한 환경에서 기존의 M-Kmeans 알고리즘과 R-Kmeans 알고리즘의 성능을 상대적으로 평가 하였다.

실험 평가 정확률과 재현률을 각각의 클러스터에 대하여 적용하였고, 전체적인 시스템의 성능을 평가하기 위하여 각 실험마다 평균 정확률과 평균 재현률을 정의한다. 또한 평균 재현률과 평균 정확률을 결합한 조화 평균(F-measure)을 정의하여 재현률과 정확률을 하나의 척도로 나타내어 두 알고리즘의 성능을 그래프로 나타내어 본다.

. 각 클러스터의 정확률(P)은 식 5 와 같다.

$$p = \frac{\text{해당클러스터에할당된관련문서수}}{\text{해당클러스터에할당된총문서수}} \quad \text{식(5)}$$

. 각 클러스터의 재현률(R)은 식 6 과 같다.

$$R = \frac{\text{해당클러스터에할당된관련문서수}}{\text{해당클러스터에관련된문서수}} \quad \text{식(6)}$$

. 클러스터링의 평균 정확률(AP)은 식 7 과 같다.

$$AP = \frac{1}{K}\sum_{k=1}^{K} P_k, k = 10 \quad \text{식(7)}$$

. 클러스터링의 평균 재현률은 식 (8)과 같다.

$$AR = \frac{1}{K}\sum_{k=1}^{K} R_k, k = 10 \quad \text{식(8)}$$

. F-Measure(AP 와 AR 의조화평균)은 식 (9)와 같다.

$$F = \frac{2 \bullet AP \bullet AR}{AP + AR} \quad \text{식(9)}$$

표에서 M 은 M-Kmeans Algorithm 을, R 은 R-Kmeans Algorithm 을, F 는 F-measure 를 나타낸다.

가중치 조합	M- AP	M- AR	M- F	R- AP	R- AR	R- F
ATC	81.12	27.0	40.51	81.14	28.0	41.63
ATN	91.44	41.0	56.61	91.43	40.0	55.65
LTC	83.00	24.0	37.23	91.35	36.0	51.65
LTN	91.64	48.0	63.00	90.96	58.0	70.83
NTC	91.26	31.0	46.28	91.35	36.0	51.65
NNN	54.65	45.0	49.35	49.27	46.0	47.58

표 2. 평균 정확률, 평균 재현률, F-measure 단위 %

표 2 는 16 가지 가중치 기법 중에서 6 가지만을 적용하였을 때 클러스터링 결과로서 평균 정확률 측면에서는 두 알고리즘이 별로 차이 없지만, 평균 재현률과 F-measure 에서는 R-Kmeans 알고리즘이 9%이상 좋은 성능을 보이고 있다. 재현률이 높다는 것은 특정한 주제 아래에 해당하는 문서가 제대로 할당되며 특정한 주제 아래 문서가 할당되는 클러스터링 성능이 우수함을 알 수 있다.

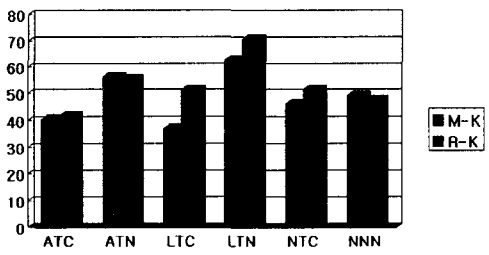

그림 3 F-measure

그림 3 은 표 2 를 F-measure 로 나타낸 도표이다. M-K 는 기존의 M-Kmeans Algorithm 의 결과 값을 나타내고, R-K 는 본 논문에서 새로 제안한 R-Kmeans 의 결과를 나타낸 것이다.

5. 결론

본 논문에서는 문서 클러스터링 기법을 소개하고 재배치 기법의 일종인 M-Kmeans 알고리즘의 성능과 문서 가중치 알고리즘을 제안하였다. 제안한 변형 알고리즘은 클러스터링 해서 얻어진 각 클러스터에 할당된 문서 수에 따른 가중치 부여 후 재클러스터링 하는 알고리즘이다. 실험 결과 평균 정확률 측면에서는 기존 M-Kmeans 알고리즘과 차이가 별로 없었지만 평균 재현률과 F-measure 측면에서는 R-Kmeans 알고리즘이 9%이상 좋은 성능을 보이고 있다.
향후 연구에서는 문서 전문과 요약문을 대상으로 하여 제안하는 알고리즘의 성능을 평가해보고자 한다.

참고문헌

[1] 오형진, 고지현, 안동언, 정성종, 2002.4, 요약 문서 기반 문서 클러스터링, 한국정보처리학회, 춘계학술발표논문집, pp.589-592.
[2] 오형진, 변동률, 이신원, 박순철, 안동언, 정성종, 2002.6. 클러스터 중심 결정 방법에 따른 문서 클러스터링 성능 분석, 대한전자공학회,하계학술대회.
[3] 오형진, " 클러스터 중심 결정 방법을 개선한 변형 K-Means 알고리즘의 구현" , 2002.8. 석사학위 논문, 전북대학교
[4] 이경순, " 정보검색에서 벡터공간 검색과 클러스터 분석을 통한 문서 순위 결정 모델" , 2001.5. 박사학위 논문, 한국과학기술원
[5] 임영희, " 후처리 웹문서 클러스터링 알고리즘" , 2002.2, 한국정보처리학회, 정보처리학회 논문집.
[6] 정영미, " 정보검색론"
[7] Baeza-Yates, Ribeiro_Neto Modern information Retrieval"
[8] khaled Alsabti, 1998, Sanjay Ranka, Vineet Singh, An Efficient K-Means Clustering Algorithm, IIPS 11th International Parallel Processing Symposium.
[9] Prabhakar Raghavans Lecture Notes of Principles of Information Retrieval.
[10] Qin He, 'A Review of Clustering Algorithms as Applied in IR", UIUCLIS--1999/6+IRG.
[11] Ray R. Larsons Lecture Notes of Principles of Information Retrieval.
[12] Tapas Kanung, 2000, The Analysis of a Simple k-Means Clustering Algorithm. Proc. of ACM Symposium on Computational Geometry Hong Kong, June 12-14.