제 18 회 한국정보처리학회 추계학술발표대회 논문집 제 9 권 제 2 호 (2002.11)

클러스터 중심 결정 방법을 개선한 K-Means Algorithm의 구현

조시성, 김호영, 오형진, 이신원*, 안동언, 정성종
*정인대학 컴퓨터정보학과

e-mail : santted@duan.chonbuk.ac.kr

An Implementation of K-Means Algorithm improving cluster centroids decision methodologies

Si-Sung Cho, Ho-Young Kim, Hyung-Jin Oh, Shin-Won Lee*, Dong-Un An, Sung-Jong Chung
*Dept. of Computer Information, ChongIn College
Dept. of Computer Engineering, Chonbuk National University

요 약

K-Means 알고리즘은 재배치 기법의 일종으로 K 개의 초기 클러스터중심(centroid)를 중심으로 K 개의 클러스터가 될 때까지 클러스터링을 반복하는 것이다. K-Means 알고리즘은 특성상 초기 클러스터 중심과 새롭게 생성된 클러스터 중심에 따라 클러스터링 결과가 달라진다. 본 논문에서는 K-Means Algorithm 의 초기 클러스터중심 선택 방법과 새로운 클러스터 중심 결정 방법을 개선한 변형 K-Means Algorithm 을 제안한다. SMART 시스템에서 제안한 16 가지 가중치 계산 방식에 의하여 두 알고리즘의 성능을 평가한 결과 제안한 변형 알고리즘이 재현률과 F-Measure 에서 20%이상 향상된 결과를 얻을 수 있었으며 특정 주제 아래 문서가 할당되는 클러스터링 성능이 우수하였다.

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 클러스터링하여 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 문서 클러스터링 기법 중 후처리 클러스터링 기법에 대하여 논의하고자 한다.

본 논문의 구성은 다음과 같다. 2 장에서는 관련연구를 살펴보고 3 장에서는 기존의 K-Means 알고리즘과 살펴보고 K-Means 알고리즘의 초기 클러스터 선택을 변형한 변형 K-Means 알고리즘을 제안하며 4 장에서는 실험결과와 분석을 기술한다.마지막으로 5 장에서는 결론 및 향후 연구과제에 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. K-Means Algorithm과 변형 K-Means Algorithm

K-Means 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

3.1 K-Means Algorithm

K-Means Algorith 은 다음과 같다..

1.클러스터 개수 K 를 선택한다

2.K 개의 초기 중심을 구한다.

3.각 문서(d)들과 K 개의 중심(c)와의 거리를 구한다..

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

4.문서를 가장 짧은 거리의 중심에 할당한다.

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

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

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

5.클러스터 중심을 재계산한다

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

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

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

$$else\ goto\ 3$$

K-Means Algorithm 은 특성상 생성된 클러스터 중심)에 따라 클러스터링 결과가 달라진다[6][8]. 특히 초기 클러스터중심을 어떻게 선택하는가에 따라 빠른 시간에 최적의 클러스터링 결과가 나오는 경우와 그렇지 않은 경우가 존재한다. 기존의 K-Means 알고리즘에서는 특정 문헌 집단에 속하는 임의의 문서를 초기 클러스터 중심벡터로 선택하였다.

클러스터링에 영향을 미치는 또 다른 요소는 클러스터링 과정에서 발생하는 새로운 클러스터 중심(cluster Centroid)를 결정하는 것이다. K-Means Algorithm 에서는 클러스터에 속하는 문서들의 색인어와 가중치만을 단순히 하나의 클러스터 벡터로 병합하였다.

3.2 Modified K-Means Algorithm

변형 K-Means Algorithm 은 특정 문헌 집합에 속하는 임의의 한 개의 문서를 선택하는 대신 색인어와 가중치로 표현되는 문서를 3 개로 선택하여 중복된 색인어를 제외하고 병합한 후 초기 클러스터 중심 벡터로 설정하였다. 변형한 초기 클러스터 중심은 식(1)와 같다.

$$c_i^{initial} = \sum_{j}^{m} d_j \qquad \text{식(1)}$$

$c_i^{initial} : i$ 번째 클러스터 벡터

$d_j : j$ 번째 문서 벡터, j=rand()%100

변형 K-Means Algorithm 에서 새로운 클러스터 중심 벡터는 클러스터에 포함된 모든 문서들이 갖는 색인어의 가중치의 평균으로 계산한다. 클러스터 중심 c_i 와 문서 d_j 가 병합되어서 생성된 클러스터 중심은 식(2)과 같이 계산한다.

$$c_i^{new} = \frac{m_i c_i + m_{ij} d_{ij}}{m_i + m_{ij}} \qquad \text{식(2)}$$

c_i : i 번째 클러스터 벡터

d_j : j 번째 클러스터에 할당된 j 번째 문서 벡터

m_i : 클러스터의 크기

$m_{ij} : i$ 번째 클러스터에 할당된 j 번째 문서의 크기

c_i^{new} : i 번째 새로운 클러스터 중심 벡터

새롭게 생성된 클러스터 중심은 클러스터에 속하는 문서들이 클러스터 중심을 형성하는 과정에서 문서에 표현되어 있는 색인어들의 가중치들로 자신들의 특성을 반영하며 서로 이웃한 문서들에게 영향을 미치게 되어 문서간의 문맥을 고려한 클러스터링 효과를 얻을 수 있게 된다.

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 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 [6]

4. 실험 및 결과

본 논문에서 구현한 전체 시스템은 크게 자동문서 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 데이터

본 논문에서 사용한 실험 문서는 Reuter21578 newswire 이다[15]. 실험 문서는 모두 TOPIC 태그를 가지고 있으며 TOPIC 태그는 문서의 내용을 파악하여 해당 주제에 속하는 문서임을 나타낸다. 선택한 실험 문서는 Reuter21578 문서에서 가장 많이 존재하는 TOPIC 10 개를 선정하였으며 각 TOPIC 당 문서 10 개씩, 총 100 개의 문서를 실험 문서로 선택하였다.[2][2]

4.2. 실험 결과

자동 문서 요약기의 출력을 10 Passage 로 하였을 때 K-Means 알고리즘(그림 1, 가중치 ANN)과 변형 K-Means Algorithm(그림 2, 가중치 ANN)을 적용하여 가중치 기법에 대한 실험 결과이다. 각 그림에서 cid 1 은 실험 데이터인 REUTER2157 newswire 의 TOPIC 1 번인 EARN, cid 2 는 TOPIC 2 번인 ACQ, cid 10 은 TOPIC 10 번인 SHIP 을 대표하는 주제이며 그림에서는 할당된 문서의 수와 문서 번호를 함께 나타내고 있다.

4.3 두 알고리즘의 성능 비교

본 장에서는 K-Means Algorithm 과 제안하는 Modified K-Means 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 K-Means 알고리즘과 변형 K-Means 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.

그림 1. K-ANN

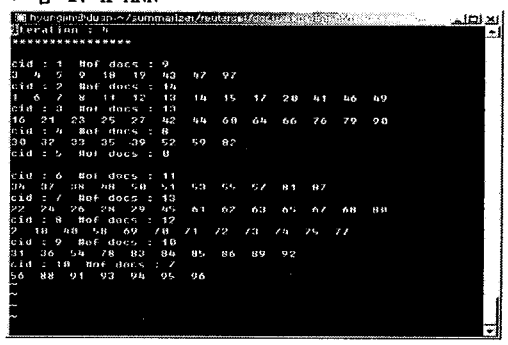

그림 2. M-ANN

. 각 클러스터의 정확률(P)은 식 5 와 같다.

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

. 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 K 는 K-Means Algorithm 을, M-K 는 Modified K-Means Algorithm 을, F 는 F-measure 를 나타낸다.

가중치 조합	K -AP	k- AR	K-F	M-AP	M-AR	M-F
BNN	72.2	40.3	28	52.6	47.9	44
BNC	73.8	40.6	28	54.3	49.2	45
BTN	72.2	40.4	28	51.0	46.1	42

BTC	83.6	34.8	22	84.0	53.3	39
NNN	62.7	40.6	30	57.7	55.3	53
NNC	65.3	47.0	37	58.9	55.8	53
NTN	72.2	40.4	28	49.1	44.1	40
NTC	75.6	31.6	20	81.6	46.0	32
ANN	71.5	45.1	33	62.3	62.1	62
ANC	71.5	45.1	33	62.3	62.2	62
ATN	74.8	36.3	24	63.0	52.5	45
ATC	55.7	35.4	26	61.6	53.3	47
LNN	57.6	45.1	37	56.7	55.3	54
LNC	67.5	44.3	33	54.0	50.8	48
LTN	76.1	32.9	21	70.3	57.7	49
LTC	68.6	38.7	27	63.5	56.0	50

표 1. Passages 5 평균 정확률, 평균 재현률, 단위 %

가중치 조합	K-AP	K-AR	K-F	M-AP	M-AR	M-F
BNN	72.8	23	35	60.1	56	58
BNC	39.7	19	25.7	50.7	53	51.8
BTN	69.4	18	28.6	60.1	56	58.0
BTC	86.0	10	17.9	75.2	44	55.5
NNN	70.1	17	27.4	62.6	57	59.7
NNC	77.8	13	22.3	46.0	46	46
NTN	77.8	23	35.5	60.0	56	57.9
NTC	80.8	18	29.4	82.5	26	39.5
ANN	65.9	16	25.7	42.8	45	43.9
ANC	66.0	16	25.8	42.8	45	43.9
ATN	61.9	23	33.5	62.8	49	55.0
ATC	74.7	21	32.8	55.2	39	45.7
LNN	67.4	29	40.6	61.3	57	59.1
LNC	81.2	12	20.9	50.8	46	48.3
LTN	81.1	20	32.1	63.8	53	57.9
LTC	78.8	16	26.6	49.6	37	42.4

표 2 Passages 10 평균 정확률, 평균 재현률, 단위 %

가중치 조합	K-AP	K-AR	K-F	M-AP	M-AR	M-F
BNN	80.5	15	25.3	50.6	48	49.2
BNC	80.5	14	23.9	49.1	48	48.5
BTN	80.5	15	25.3	50.6	48	49.2
BTC	85.5	13	22.6	80.6	31	44.8
NNN	77.5	15	25.1	43.1	28	33.9
NNC	80.5	16	26.7	45.3	36	40.1
NTN	80.5	15	25.3	50.6	48	49.3
NTC	85.5	14	24.1	81.9	30	43.9
ANN	80.5	15	25.3	57.03	50	53.3
ANC	80.5	15	25.3	57	41	47.7
ATN	80.5	14	23.9	56.1	45	49.9
ATC	85.5	14	24.1	59.3	42	49.2
LNN	80.5	16	25.3	47.2	44	45.6
LNC	80.5	15	25.3	44.9	46	45.4
LTN	80.5	15	25.3	57.1	46	51.0
LTC	85.5	15	25.5	46.5	38	41.8

표 3. Passages 5 평균 정확률, 평균 재현률, 단위 %

표 1~3 는 16 가지 가중치 기법을 적용하였을 때 클러스터링 결과로서 평균 정확률 측면에서는 K-Means 알고리즘이 보다 높게 나타났지만, 평균 재현률과 F-measure 에서는 변형 K-Means 알고리즘이 20%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.

5. 결론

본 논문에서는 문서 클러스터링 기법을 소개하고 재배치 기법의 일종인 K-Means 알고리즘의 초기 클러스터 선택과 새로운 클러스터 중심 결정을 변경한 변형 K-Means 알고리즘을 제안하였다. 제안한 변형 알고리즘은 초기 클러스터를 3 개의 임의의 문서로 선택하여 색인어를 병합하였으며 새로운 중심은 각 클러스터에 속하는 문서의 색인어 가중치를 평균한 것이다. 실험 결과 요약문의 크기에 상관없이 평균 정확률 측면에서는 K-Means 알고리즘이 높게 나타났지만 평균 재현률과 F-measure 측면에서는 변형 K-Menas 알고리즘이 20%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, 요약문서 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] 김경순, " 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] khaled Alsabti, 1998, Sanjay Ranka, Vineet Singh, An Efficient K-Means Clustering Algorithm, IIPS 11th International Parallel Processing Symposium.
[7] Prabhakar Raghavans Lecture Notes of Principles of Information Retrieval.
[8] Qin He, 'A Review of Clustering Algorithms as Applied in IR", UIUCLIS--1999/6+IRG.
[9] Ray R. Larsons Lecture Notes of Principles of Information Retrieval.
[10] Tapas Kanung, 2000, The Analysis of a Simple k-Means Clustering Algorithm. Proc. of ACM Symposium on Computational Geometry Hong Kong, June 12-14.