제21회 한국정보처리학회 춘계학술발표대회 논문집 제11권 제1호 (2004. 5)

계층적 클러스터링에서 분류 대표어 선정에 관한 연구*

이상선, 이신원, 안동언, 정성종
전북대학교 컴퓨터공학과
e-mail : wish@duan.chonbuk.ac.kr

A Study on Cluster Topic Selection in Hierarchical Clustering

Sang-Seon Yi, Shin-Won Lee, Dong-Un An, Sung-Jong Chung
Dept. of Computer Engineering, Chonbuk National University

요 약

정보의 양이 많아지면서 정보 검색 시스템에 검색 결과를 자동으로 구조화하는 계층적 클러스터링을 적용하는 시도가 늘고 있다. 계층적 클러스터링은 문서 간의 유사도를 통해 클러스터를 계층 구조로 만들어 검색 성능을 높이고 결과를 사용자에게 이해하기 쉽게 보여준다. 계층 구조는 검색 결과를 요약하는 것이기 때문에 클러스터의 내용을 효과적으로 함축할 수 있는 대표어의 선정이 중요하다. 각 클러스터의 대표어를 선정하기 위해 대표어에 명사인 단어만 추출하고 상위 클러스터 대표어에 사용된 단어는 하위 클러스터에 사용하지 않는 방법을 적용하여 대표어의 질을 높였다.

1. 서론

기존의 정보 검색 시스템은 질의어를 입력하면 긴 검색 결과 리스트를 제시하여 사용자가 세밀하게 적합한 문서를 찾아야 한다. 다양한 주제에 관련된 문서를 검색하고 조직화하는 일은 많은 시간과 노력을 필요로 하기 때문에 시스템에서 자동으로 문서 집합을 조직화하는 방법에 대한 연구가 필요하다.

문서 클러스터링은 대용량의 문서 집합을 주제에 따라 분류하는 것으로 정보 검색 분야에서 문헌 구조를 분석하거나, 검색 효과와 성능을 높이기 위해 이용되고 있다. 그 중 계층적 클러스터링은 문서 간의 유사도를 통해 단계적으로 계층 구조로 만들어 저장한 것으로 검색 엔진의 계층적 구조를 하향 탐색하여 검색 성능을 높일 수 있다. 또, 사용자에게 검색된 문서 구조를 계층적으로 보여주어 평탄한 구조보다 검색 결과를 직관적으로 이해하기 쉽게 보여준다. 그 중 Vivisimo[1]는 문서 자동 분류를 이용한 대표적인 정보 검색 시스템으로 검색 결과를 클러스터링하여 계층적인 구조로 보여준다.

검색 결과를 계층적으로 보여줄 때 결과를 정확하게 분류하는 것도 중요하지만 분류한 결과를 요약해서 보여주는 대표어 선정 또한 중요하다. 몇 단어만으로 클러스터의 내용을 함축해야 하기 때문에 정제된 단어의 선택이 필수적이다. 본 논문에서는 대표어의 질을 높이기 위해 사전을 이용하여 명사만 추출하는 방법과 상위 클러스터 대표어에 사용된 단어를 하위 클러스터에 반복해서 사용하지 않는 방법을 제시한다.

본 논문의 구성은 다음과 같다. 3 장에서는 정보 검색 시스템인 Condor 시스템의 구성과 계층적 클러스터링 모듈, 대표어 선정 방법에 대하여 기술을 하며, 4 장에서는 실험 결과를 보여주고 끝으로 5 장에서는 결론을 맺겠다.

2. 관련 연구

문서 클러스터링 기법은 클러스터를 구성해나가는 방법에 따라 계층적 클러스터링과 비계층적 클러스터

* 본 연구는 한국과학재단 목적기초연구 R01-2003-000-11588-0 지원으로 수행되었음.

링으로 나누어 볼 수 있다. 계층적 클러스터링은 비계층적인 방법에 비해 처리 속도가 느리지만 정확한 클러스터링이 수행된다는 장점을 갖는다. 두 개의 유사도가 높은 클러스터를 선택하는 방법에 따라 단순 링크 기법, 복합 링크 기법, 그룹 평균 링크 기법, Ward 기법 등이 있다. [2]

비계층적 클러스터링은 임의로 선택된 초기 클러스터로부터 문서를 클러스터에 재배치하는 작업을 반복적으로 수행하여 최종 클러스터를 형성하는 방법으로 Single pass 방법, K-means 알고리즘이 대표적이다. [3] 계층적 클러스터에 비해 클러스터링 성능이 떨어지고 문서의 입력 순서에 따라 클러스터링 결과가 달라지는 등의 단점을 갖지만 클러스터링 처리 속도가 빨라서 문서의 양이 매우 많거나 실시간으로 클러스터링 해야 할 때 자주 사용된다.[4][5][6]

Vivisimo 는 여러 개의 검색 엔진을 통합 검색하는 메타검색 기능과 문서 클러스터링 기술에 의해 실시간 자동 분류하여 계층적으로 보여주는 기능을 제공한다. 클러스터마다 한두 개의 단어를 대표어로 선정하는데 대표어 선정에 대체적으로 좋은 성능을 보이지만 한국어의 경우 동사나 형용사 등 명사 이외의 품사도 대표어로 선택하여 영어에 비해 좋지 못한 성능을 보이고 있다.

3. 시스템

실험은 전북대학교 지능공학연구실과 CMU, LNI, ㈜서치라인[7]이 공동 개발한 Condor 시스템을 기반으로 이루어졌다.

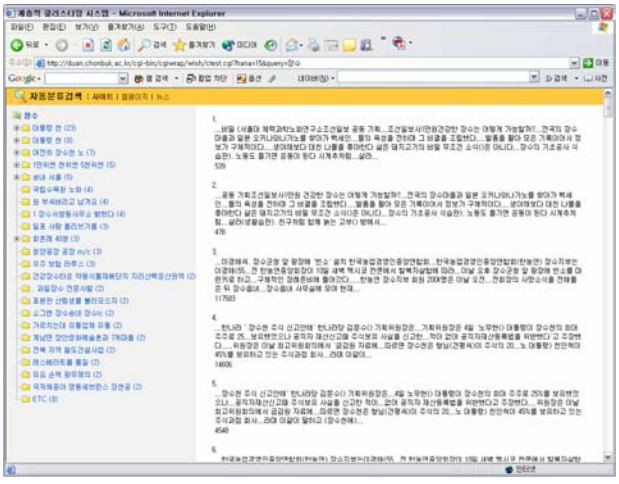

그림 1. Condor 시스템의 검색 결과

Condor 는 대용량의 문서를 검색 및 색인 하는 정보 검색 엔진으로 크게 Data, 검색 엔진, 사용자 인터페이스 부분 등 세 부분으로 나눠져 있다. Condor 시스템에서 계층적 클러스터링은 Data 를 색인, 질의 처리, 요약 등을 처리하는 엔진 컴포넌트와 API, 사용자 인터페이스 사이에 위치하여 색인이 끝난 뒤 전처리 과정을 담당하게 된다.

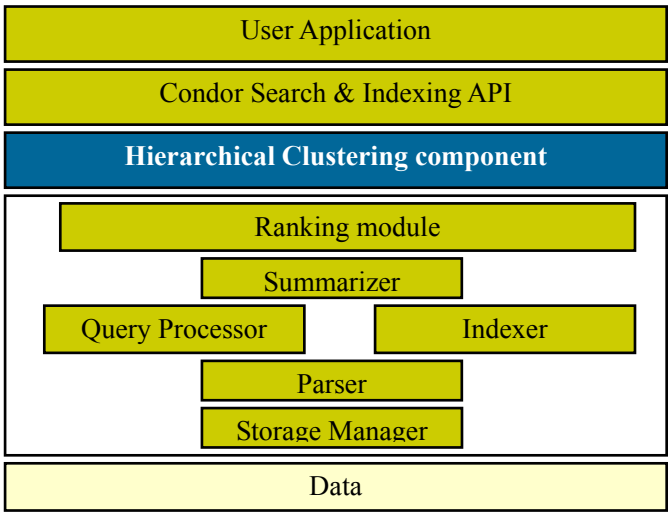

그림 2. Condor 시스템의 구조

Condor 시스템은 K-means 알고리즘을 이용하여 계층적 클러스터링을 구현했는데 그 이유는 K-means 알고리즘이 다른 계층적 클러스터링 알고리즘에 비해 정확성은 떨어지지만 구현이 간단하고 처리 속도가 매우 빨라 사용자의 질의에 따라 실시간으로 많은 양의 문서를 클러스터링해야 하는 웹 검색엔진에 어울리기 때문이다. K-means 알고리즘을 이용하여 계층적 클러스터링을 구현한 알고리즘은 다음과 같다.

① 클러스터의 초기 중심값을 구한다.

② 각 문서(D)들과 중심값(C) 사이의 거리를 구한다.

$$dist(D_i, C_j) = \frac{\sum_{k=1}^{n}(D_{i,k} \times C_{j,k})}{\sum_{k=1}^{n}(D_i,k)^2 \times \sum_{k=1}^{n}(C_j,k)^2}$$

i : 1, 2, …, n n : 전체 문서 수
j : 1, 2, …, K k : centroid 의 수

③ 문서를 가장 짧은 거리의 중심값에 할당한다.

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

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

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

④ 새로운 클러스터 중심값을 재계산한다.

$$\vec{C}_j = \frac{1}{|C_j|}\sum_{l=1}^{|C_i|}\vec{d}_l$$

⑤ 이전의 중심값과 새로운 중심값을 비교하여 벡터 간 차이가 거의 없을 때까지 반복한다.

$$if\ \max \delta(\overline{C}_j^{old}, \overline{C}_j^{new}) < \theta\ then\ return$$

$$else\ goto\ ③$$

⑥ 클러스터 내 문서의 유사도가 한계치보다 적으면 클러스터 안에서 다시 클러스터링한다.

⑦ 클러스터를 트리에 저장한다.

총 클러스터의 계층은 문서 집합 상황에 따라 최고 3 단계까지 세부 분류를 한다. 그렇지만 소속 문서들의 상황에 따라 어떤 클러스터는 2,3 단계까지 세부 분류가 될 수도 있고 어떤 클러스터는 1 단계에서 분류가 멈출 수도 있다.

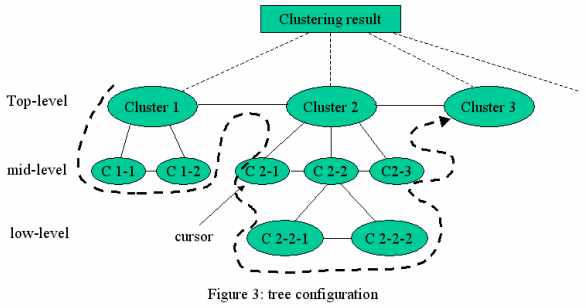

Figure 3: tree configuration

그림 3. 계층적 클러스터링

문서의 세부 분류를 마치면 각 클러스터의 대표어를 선정한다. 클러스터의 단어 중 가중치가 높은 3 단어를 선정하는데 단어의 가중치 계산 방법은 식 (1)과 같다.

$$weight = \frac{tf}{tf + 2.0} \times \frac{df + 2.0}{df} \qquad (1)$$

선정되는 단어의 질을 높이기 위해 다음과 같은 처리를 추가한다.

① 사전을 이용하여 명사만 추출한다.

$$if\ (term_{j/C_i} \in Noun)\ then\ Terms_{C_i}$$

$Terms_{C_i}$: i번째 $Cluster$의 단어 집합

$term_{j/C_i}$: $Cluster_i$의 j번째 단어

i : $Cluster$의수

j : $Cluster$의단어수

② 이미 상위 클러스터에서 사용한 단어는 중복을 배제한다.

$$if\ (Terms_{C_{i+1}} \notin Terms_{C_i})\ then\ Terms_{C_{i+1}}$$

③ 클러스터의 단어를 가중치가 높은 순으로 3 개 추출하여 대표어로 선정한다.

4. 실험 및 결과 분석

실험 데이터는 2003 년 5 월부터 11 월까지의 조선일보, 중앙일보, 동아일보 등의 뉴스와 10 월부터 11 월의 한겨레신문, 한국일보, 문화일보 등의 뉴스를 매 10 분 간격으로 수집한 후 색인하여 실험에 사용하였다.

그림 4. 수정하기 전 Condor 시스템

그림 5. 수정한 후 Condor 시스템

논문에서 제안한 명사 추출 방법과 중복 제거 방법을 적용한 콘도르 시스템과 적용하지 않은 콘도르 시스템을 비교하였다. 또, 여러 클러스터링 상황에 따른 성능을 비교하기 위해 실행시킬 검색어를 넓은 의미를 갖는 것, 좁은 의미를 갖는 것, 중의성을 띄는 것으로 나눠 비교했다.

표 1. 실험 검색어

중의성을 띄는 단어	Q1	유산	Q2	장수
넓은 의미를 갖는 단어	Q3	여성	Q4	정치
좁은 의미를 갖는 단어	Q5	게임	Q6	금연

검색 시스템에 질의를 실행시켜 클러스터의 대표어 선정이 정확한 것인지를 평가한다. 어떻게 클러스터링이 되었는지 보다는 클러스터링 되어 모아진 문서 들의 대표어가 잘 선정되었는지를 평가하기 위해 대표어로 선정된 3 개의 단어 중 클러스터의 내용에 적합한 단어의 수를 세어 정확율을 계산하였다. 상위 클러스터의 대표어와 중복되거나 대표성을 띠지 않는 의미를 알 수 없는 단어는 틀린 것으로 했다.

표 2. 기존 클러스터링 결과

	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6	합계
총 대표어	89	114	90	90	99	99	581
맞은 대표어	51	63	34	39	47	37	271
정확율	.57	.55	.38	.43	.47	.37	.47

표 3. 명사만 클러스터링 결과

	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6	합계
총 대표어	95	105	89	85	94	97	565
맞은 대표어	50	69	53	57	62	50	341
정확율	.52	.66	.60	.67	.66	.52	.60

표 4. 주제어 중복 제거 결과

	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6	합계
총 대표어	84	94	76	73	71	83	481
맞은 대표어	50	70	51	60	55	66	352
정확율	.60	.74	.67	.82	.71	.80	.73

표 5. 명사 추출 + 중복 제거 결과

	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6	합계
총 대표어	95	108	89	85	94	97	568
맞은 대표어	61	81	59	72	71	70	414
정확율	.64	.75	.66	.84	.76	.72	.73

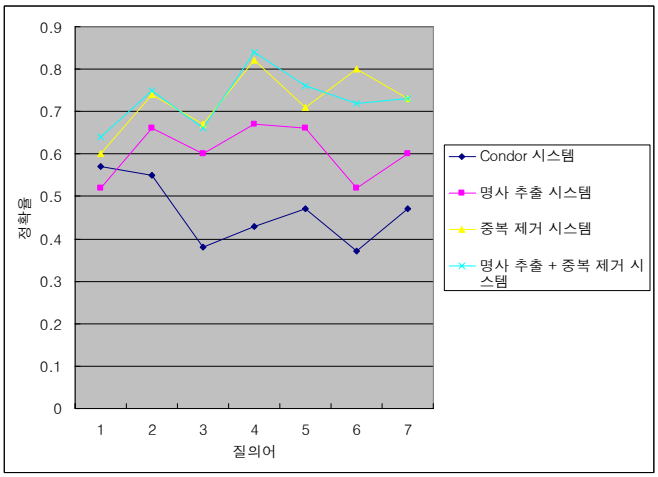

그림 6. 실험 결과

실험 결과(표 2, 표 3, 표 4, 표 5, 그림 4)에 의하면 명사 추출 및 중복 제거를 한 시스템이 적용하지 않은 시스템에 비해 더 나은 성능을 보이는데 필요 없는 정보나 중복되는 단어를 제거했기 때문이다. 대표어에 명사만 추출하여 선정한 경우 성능이 좋아지는 이유는 동사나 부사, 형용사 같은 품사는 단어에 특정한 의미를 지니지 않아 클러스터의 내용을 표현하는데 어려움이 있기 때문이다. 상위 클러스터의 대표어와 중복되는 단어를 제거한 경우에도 많은 정보를 사용자에게 보여주고 각각의 클러스터를 효과적으로 구분 지을 수 있기 때문에 좋은 성능을 보였고 대표어에 명사만 추출하고 중복도 제거한 시스템이 가장 좋은 성능을 보였다. 평가는 사람에 의하여 판정하였다.

5. 결론

본 논문에서는 정보 검색 시스템에서 검색 결과를 계층적 클러스터링하여 대표어를 선정할 때 명사만 추출하거나 상위 클러스터의 대표어와의 중복을 제거하는 방법으로 대표어의 질을 높였다. 시스템의 성능을 비교한 결과 원래 시스템보다 명사만 추출한 방법이나 중복을 제거한 방법이 더 높은 결과를 보였는데 대표어에 많은 의미를 주고 각각의 클러스터의 구분이 쉬워지기 때문이다.

향후 연구로는 계층적 클러스터링을 한 K-means 알고리즘은 일반적으로 비계층적 클러스터링에 사용되는 알고리즘이기 때문에 성능이 좋지 않으므로 다른 계층적 클러스터링 알고리즘을 사용하여 클러스터링 성능을 높일 필요가 있다. 또, 가중치가 높은 단어를 우선으로 대표어로 선정하기 때문에 각 단어의 조합이 어울리지 않아 의미가 모호한 대표어가 있는데 이를 개선하는 연구가 필요하다.

참고문헌

[1] Vivisimo http://www.vivisimo.com/
[2] Qin He, “A Review of Clustering Algorithms as Applied in IR,” UIUCLIS—1999/6+IRG
[3] Tapas Kanung, “The Analysis of a Simple k-Means Clustering Algorithms” in Proceedings of the sixteenth annual symposium on Computational geometry, 2000
[4] 김명철 외, “최신 정보 검색론”, 홍릉과학출판사
[5] 오형진, 변동률, 이신원, 박순철, 정성종, 안동언, “클러스터 중심 결정 방법에 따른 문서 클러스터링 성능 분석”, 대한전자공학회 2002 년 하계학술대회
[6] khaled Alsabti, 1998, Sanjay Ranka, Vineet Singh, An Efficient K-Means Clustering Algorithm, IIPS 11th International Parallel Processing Symposium.
[7] 박순철, 안동언, “콘도르 정보 검색 시스템”, 한국산업정보학회 논문지, 제 8 권 제 4 호, pp. 31- 37, 2003.