검색결과의 브라우징을 위한 계층적 클러스터링

윤보현, 김현기, 노대식, 강현규
지식정보연구부, 한국전자통신연구원
{ybh, hkk, rosik, hkkang}@etri.re.kr

A Hierarchical Clustering for Browsing Retrieval Results

Bo-Hyun Yun, Hyun-Ki Kim, Dae-Sik Roh, Hyun-Kyu Kang
Knowledge Information Department,
Electronic Telecommunications and Researches Institute

요약

대부분 웹 검색엔진들의 검색결과로 수십 혹은 수백만건의 문서가 제시되어 사용자가 원하는 문서를 찾는데 어려움이 있다. 이러한 문제를 해결하기 위해 본 논문에서는 검색 결과의 브라우징을 위한 검색 결과 문서에 대한 자동 클러스터링 방법을 제안한다. 문서간 유사도를 계산하기 위해 공통 키워드 빈도를 이용하고, 클러스터링 방법은 계층적 클러스터링을 사용하고, 각 클러스터에 대한 디스크립터를 추출하기 위해 빈도를 이용한다. 실험 결과, 완전 연결 방법이 가장 나은 정확도를 보였지만 계산시간이 많이 소요되어 동적 환경에 부적합하다는 것을 보였다. 아울러 집단 평균 연결이 정확도나 계산 시간 측면에서 우수함을 알 수 있었다.

1. 서론

기존의 정보 검색 시스템은 사용자가 적합한 문서를 꼼꼼하게 찾아야하는 긴 검색결과 리스트를 제시한다. 이러한 시스템에서 문서에 포함된 정보는 분산되며, 비슷한 색인어를 갖는 문서들은 근접해 있지 않아 검색에 어려움이 있다. 따라서 검색결과 문서들에 대해 클러스터링을 수행하여 클러스터를 브라우징할 수 있다면, 임의의 클러스터에 제한하여 원하는 문서들을 찾을 수 있을 것이다.

문서 클러스터링은 문서간의 유사도에 근거하여 문서를 그룹화함으로써 하나의 주제에 관련된 문서들은 하나의 클러스터에 속하도록 하는 것이다. 문서 클러스터링은 문서 집합에 대해 미리 수행될 수 있으나 검색결과 클러스터링이 보다 더 나은 성능을 보여 왔다. 그 이유는 검색결과 클러스터링이 색인된 문서집합에 대해 클러스터가 계산되지만 검색된 클러스터링은 부적합한 문서들이 클러스터 정보에 영향을 미치기 때문이다[2,5,6].

사용자의 편리한 검색을 위해 본 논문에서는 검색결과에 대해 클러스터링을 수행하여 클러스터에 대해 브라우징이 가능하도록 하는 클러스터링 방법을 제안한다. 제안한 방법에서는 계층적 클러스터링 방법을 클러스터링 방법으로 이용하고, 문서간 유사도를 계산하기 위해 공통키워드의 빈도를 사용하고, 빈도에 기반하여 디스크립터를 추출한다.

2. 관련연구

클러스터링 기법은 비계층적 클러스터링 방법과 계층적 클러스터링 방법으로 나뉘어 진다. 비계층적인 클러스터링 방법은 임의로 선택된 초기의 클러스터로부터 문서를 클러스터로 재배치하는 작업을 반복하여 최종의 클러스터를 형성하는 방법이다. 이 방법은 클러스터링 시간이 빠르나 대부분 검색 효율이 떨어지고 문서의 입력 순서에 따라 클러스터링 결과가 변화하는 문제점을 갖는다. 선형시간 클러스터링 알고리즘으로 K-means 알고리즘[3], 단일 패스(Single Pass) 알고리즘[1], Buckshot 과 Fractionation[2], 그리고 STC(Suffix Tree Clustering) 알고리즘[7]이 있다.

계층적 클러스터링 방법[1,4]은 문서간의 유사도를 계산하여 유사도 행렬을 작성하고 클러스터링 알고리즘에 의해 계층적 클러스터를 구축하는 방법이다. 이 방법은 클러스터링 시간이 비계층적인 방법에 비해 느리지만 상대적으로 보다 정확한 클러스터링을 수행한다.

3. 공통 키워드 빈도를 이용한 계층적 클러스터링

3.1 시스템 구성도

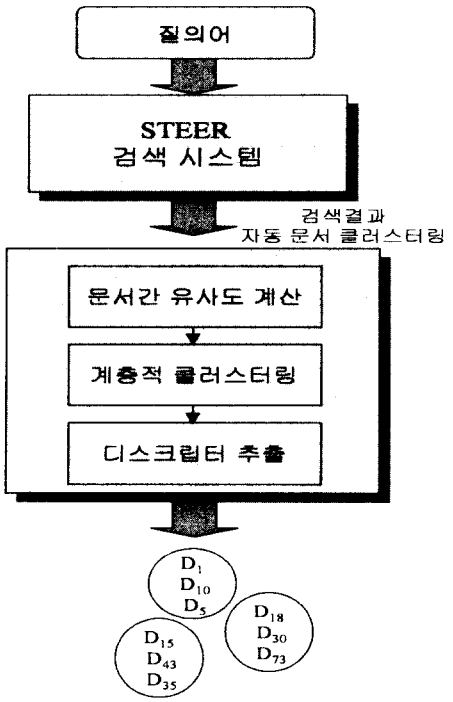

그림 1. 시스템 구성도

그림 1에는 제안하는 검색결과 브라우징을 위한 클러스터링의 시스템 구성도이다. 검색시스템으로는 본 연구실에서 기개발된 STEER 검색시스템[8]을 이용한다. STEER 검색시스템으로 검색된 문서들에 대해 공통 키워드 빈도를 이용하여 문서간 유사도를 계산한다. 계산된 유사도를 기반으로 계층적 클러스터링을 수행한다. 마지막으로 각 클러스터내의 문서들에서 빈도를 이용하여 디스크립터를 추출하는 과정을 수행한다.

3.2 공통 키워드의 빈도를 이용한 문사간 유사도 계산

문서사이의 유사도를 계산하기 위해 공통키워드의 빈도를 이용하는 방법은 "두 문서에 같은 형태소가 많이 나타날 수록, 문서의 내용이 더 유사하다"는 가정을 기반으로 한다. 즉, 두 문서가 완전히 같은 키워드로 이루어져 있으면 그 내용도 완전히 동일하고, 두 문서에 같은 키워드가 하나도 나타나지 않으면 두 문서의 내용은 전혀 연관이 없다고 보는 방법이다. 그러나 문서에 키워드가 많을수록 같은 키워드가 나타날 가능성이 많으므로, 문서사이의 유사도가 문장의 길이에 영향을 받지 않게 하기 위해서, 두 문서에 나타나는 같은 키워드의 개수를 문서의 키워드의 개수의 평균으로 나누어준다.

같은 키워드가 나타나는 빈도를 이용하여 문서들 사이의 유사도를 계산하는 식은 식(1)과 같다.

$$Sim(D_i, D_j) = \frac{2|D_i \cap D_j|}{|D_i| + |D_j|} \quad (1)$$

식 (1)에서 $|D_i|$는 문서 D_i에 있는 키워드의 개수이고, $|D_i \cap D_j|$는 두 문서 D_i와 D_j에 동시에 나타나는 문서의 개수이다.

3.3 계층적 클러스터링

가장 일반적으로 사용하는 계층적 클러스터링 알고리즘은 다음 세 가지가 있다[1,4].

• 단일연결(SL, Single Link) 방법

각각의 단계에서 같은 클러스터에 있지 않은 객체의 가장 유사한 쌍을 결합한다. 이 방법은 비교적 효율적으로 구현할 수 있어 널리 사용한다. 그러나 클러스터를 길게 나열하는 경향이 있어서 타원형의 클러스터를 표현하기에는 적합하지만 분리형 클러스터를 표현하기에는 부적합하다.

• 완전연결(CL, Complete Link) 방법

클러스터 사이의 유사도를 결정하기 위하여 두 클러스터에서 가장 유사도가 작은 쌍을 이용한다. 즉, 클러스터에 있는 모든 객체가 가장 작은 유사성으로 서로 연결되기 때문에 완전연결이라 칭한다. 작고 단단하게 묶인 클러스터가 이 방법의 특징이다.

• 집단평균연결(GAL, Group Average Link) 방법

유사성을 결정하기 위해 클러스터 내에서 연결한 쌍의 평균값을 이용한다. 모든 객체들은 클러스터간의 유사성에 기여하기 때문에 느슨하게 묶인 단일연결 클러스터와 단단하게 묶인 완전연결 클러스터 사이의 중간적인 구조를 나타낸다.

3.4 디스크립터(Descriptor) 추출

클러스터링이 수행되고 난 후 필요한 작업은 디스크립터 추출이다. 디스크립터 추출은 각 그룹내의 문서들로부터 디스크립터를 추출하여 사용자가 해당 클러스터를 쉽게 브라우징할 수 있도록 하는 역할을 한다. 그러나 이 작업은 일반적으로 계산시간이 많이 소요되므로 계산시간을 최소화하면서 효율적인 방법이 필요하다.

본 논문에서는 디스크립터를 추출하기 위해 빈도를 이용한다. 각 클러스터내의 문서들에서 색인시 추출했던 키워드를 색인어 화일에서 추출한다. 이중에서 1음절과 2음절 키워드는 제거한다. 그 이유는 1음절과 2음절의 키워드는 그 빈도가 매우 높아 불용어일 가능성이 높기

때문이다. 아울러 3음절이상의 키워드가 대부분 복합명사이기 때문에 변별력이 있는 디스크립터로 사용할 수 있다. 추출한 키워드를 빈도순의 정렬하여 빈도가 높은 다섯 개의 키워드를 디스크립터를 추출한다.

4. 실험 및 평가

실험 자료는 정보과학회 논문을 대상으로 하고, 사용한 질의는 "정보", "시스템", 그리고 "평가" 세 가지를 사용하였다. 실험 결과에서 검색 결과의 클러스터링의 평가 기준은 문서의 내용이 문서가 속해 있는 클러스터의 내용과 유사할 때 정확하게 평가된 것으로 간주하였다. 클러스터의 의미는 클러스터 내에 있는 문서들의 의미중 가장 많이 사용된 의미로 정의하였다. 클러스터링 정확도는 식(2)에 의해 계산할 수 있다.

$$\text{정확도} = \frac{\text{정확하게 분류된 문서의 개수}}{\text{분류하고자하는 문서의 개수}} \quad (2)$$

문서 사이의 유사도를 계산하는 방법들의 정확도를 비교하기 위해서는, 분류 결과의 클러스터의 수를 같게 해야한다. N개의 문서를 N개의 클러스터로 분류한다면 어떤 방법을 사용하던지 정확도가 100%가 될 것이고, N개의 문서를 1개의 클러스터로 분류하라고 하면 정확도가 가장 많이 사용된 의미의 비율로 나올 것이다. 그러므로 정확도 평가를 위해서는 클러스터의 개수를 같게 해 주어야 한다. 그러나 분류할 문서의 개수에 비해서 클러스터의 개수가 너무 많게 되면 모든 방법이 어느 정도 이상의 정확도를 보이기 때문에 정확한 비교를 할 수가 없게 된다. 이러한 이유로 클러스터의 개수를 적정한 수준으로 결정해야 한다.

표1에서 ()의 의미는 해당 질의를 사용했을 때 검색되는 문서 개수이다. 완전 연결 방법이 전반적으로 가장 좋게 나타났지만, 클러스터의 개수가 다른 방법보다 훨씬 많고 계산시간이 많이 소요되므로 전체적으로 성능이 좋다고 말할 수 없다. 완전 연결 방법의 클러스터 개수가 많은 이유는 클러스터 사이의 유사도를 측정하는데 두 클러스터에서 가장 유사도가 작은 쌍을 사용하므로, 두 클러스터에 유사도가 0인 문서쌍이 하나라도 있으면 두 클러스터는 클러스터링이 되지 않기 때문에 단일 연결과 집단 평균 연결에 비해 클러스터의 수가 많게 나타났다.본 논문에서는 유사도가 0.2이고 집단 평균 연결을 사용했을 때 정확도가 비교적 높은 편이었기에 이 방법을 사용한다.

5. 결론

본 논문에서는 검색 결과의 브라우징을 위한 검색 결과 문서에 대한 자동 클러스터링 방법을 제안하였다. 문서간 유사도를 계산하기 위해 공통 키워드 빈도를 이용하였고, 클러스터링 방법은 집단 평균 연결 방법을 사용하였고, 빈도에 의해 디스크립터를 추출하는 방법을 이용하였다. 실험 결과, 완전 연결 방법이 가장 나은 정확도를 보였지만 계산시간이 많이 소요되어 동적환경에 부적합하다는 것을 보였다. 아울러 집단 평균 연결이 정확도나 계산시간 측면에서 우수함을 알 수 있었다.

표 1. 실험 결과

유사도	정보(95)			시스템(83)			평가(98)		
	방법	클러스터 개수	정확도	방법	클러스터 개수	정확도	방법	클러스터 개수	정확도
0.1	SL	1	45.1	SL	1	45.1	SL	1	45.1
	CL	17	82.4	CL	20	84.8	CL	16	83.7
	GAL	1	45.1	GAL	1	45.1	GAL	1	45.1
0.2	SL	1	45.1	SL	1	45.1	SL	1	45.1
	CL	27	86.5	CL	31	87.2	CL	28	88.6
	GAL	9	83.3	GAL	11	83.2	GAL	10	82.5
0.3	SL	2	63.8	SL	2	61.5	SL	2	66.2
	CL	36	82.7	CL	42	80.6	CL	34	82.5
	GAL	20	80.3	GAL	25	79.7	GAL	17	81.2
0.4	SL	7	78.5	SL	10	76.3	SL	9	77.2
	CL	67	85.2	CL	73	86.4	CL	61	84.6
	GAL	19	83.6	GAL	22	84.1	GAL	17	82.4
0.5	SL	13	76.4	SL	18	73.9	SL	15	70.6
	CL	78	83.8	CL	82	85.3	CL	76	83.2
	GAL	28	81.6	GAL	34	80.8	GAL	24	79.1

참고문헌

[1] Frakes, W.B., Baeza-Yates, R., Information Retrieval, Prentice Hall, 1992.

[2] Cutting, D.R., Karger, D.R., Perderson, J.O., Tukey, J.W., "Scatter/Gather: a cluster-based approach to browsing large document collections," SIGIR'92, pp. 318-329, 1992.

[3] Rocchio, J.J., Document Retrieval Systems - optimization and evaluation, Ph.D. Thesis, Havard University, 1966.

[4] Salton, G., Automatic Text Processing, Addison-Welsley Publishing Company, 1989.

[5] Schütz, H., Sulverson, C., "Projections for efficient document clustering," SIGIR'97, pp. 74-81, 1997.

[6] Silverson, C., Perderson, J.O., "Almost-constant time clustering of arbitarily corpus subsets," SIGIR'97, pp. 60-66, 1997.

[7] Zamir, O., Etzioni, O., "Web Document Clustering: A Feasibility Demonstration," SIGIR'98, pp. 46-54, 1998.

[8] 박영찬, 김문석, 김남일, 주종철, "SGML/XML 정보검색 시스템의 구성과 구현 방법론 사례연구 : STEER-SGML/XML" 제10회 한글 및 한국어 정보처리 학술대회, pp. 105-110, 1998.