자동 문서 클러스터링을 위한 디스크립터 추출 방안

윤보현*, 강현규*, 고형대**
한국전자통신연구원 언어공학연구부*
목포대학교 정보공학부**
e-mail:ybh@etri.re.kr

A Method of Descriptor Extraction for Automatic Document Clustering

Bo-Hyun Yun, Hyun-Kyu Kang, Hyung-Dae Ko
Dept. of Language Engineering, ETRI*,
Dept. of Information Engineering, Mokpo National University**

요약

기존의 검색엔진은 검색결과를 적합도 순서로 나열하여 사용자가 원하는 문서를 찾는데 어려움이 있다. 이러한 문제의 해결책으로 검색결과 문서에 대해 자동 클러스터링을 수행하여 문서 내용이 유사한 문서가 하나의 클러스터내에 존재하도록 한다. 본 논문에서는 검색 결과 문서의 클러스터링에서 필요한 디스크립터 추출 방안을 제안한다. 각 클러스터 내에서 디스크립터를 추출하기 위해 정보검색의 색인과정에서 사용하는 용어 가중치 계산 방법을 이용한다.

1. 서론

검색 결과 문서 클러스터링과 같은 자동 문서 처리 기술은 대량의 검색 결과 문서에서 사용자가 원하는 문서를 쉽게 찾도록 도와준다[2,3,7]. 이러한 검색 결과 문서 클러스터링에 있어서 클러스터에 속하는 문서들에서 디스크립터를 추출하는 작업은 사용자가 검색결과 문서에 대해 브라우징이 가능하도록 하기 때문에 반드시 필요한 작업이다.

디스크립터 추출은 하나의 클러스터를 대표하는 키워드들을 추출하는 작업을 말한다. 이러한 디스크립터 추출은 문서 요약 시스템[10], 정보 추출 시스템[9], 그리고 주제 인식 시스템[4]에도 이용될 수 있다. 문서 요약 시스템의 첫 과정으로서 문서들의 주제를 인식하는 주제인식 과정이 수행된다. 정보 추출 시스템에서도 자연어 문서로부터 원하는 정보만을 찾는 과정에 응용될 수 있으며, 주제 인식 시스템에서 문서의 주제를 찾는 과정에 적용될 수 있다.

기존의 디스크립터 추출의 연구에서는 불용어 사전을 구축하여 필터링하는 과정이 필요하거나 의미없는 용어를 디스크립터로 선정하는 문제가 있다.

본 논문에서는 검색결과 클러스터링에 필요한 용어 가중치 기반 자동 디스크립터 추출 방안을 제시하고자 한다.

2. 관련연구

기존에 디스크립터를 추출하는 방법[5]에는 Odds 기반 기법과 Centroid 기반 방법이 있다.

Odds 기반 기법은 각 문서 그룹 c_j안의 한 문서에서 나타나는 용어 t_i의 확률적 odds에 대한 다른 그룹의 한 문서에서 나타나는 용어의 비인 $O_j(t_i)$는 식 (1)과 같다.

$$O_j(t_j)=\frac{P(t_i|\,c_j)}{\sum_{c_k \neq c_j} P(t_i|\,c_k)} \quad (1)$$

문서집합 c_j에 대한 디스크립터로서 가장 높은 O_j의 값을 갖는 k개의 용어를 선택한다.

Centroid 기반 방법은 각 그룹 c_j에 속한 모든 문서들에 대해 Euclidean centroid를 계산한다. Centroid vector에서 가장 높은 값을 가지는 k개의 용어를 그 그룹의 디스크립터로 선택한다.

Centroid 기반 방법이 Odds 기반 기법보다 더 나은 성능을 보인다. 그러나 Centroid 기반 방법의 성능은 불용어들이 디스크립터 리스트에 나타나지 않도록 미리 제거하는 불용어 제거의 효율에 의존적이다. 대조적으로 Odds 기반 방법은 다른 문서 집합에서는 자주 나타나지만 한 문서집합에서는 드물게 나타나는 용어를 선호하는 경향이 있다.

3. 용어 가중치 기법에 의한 디스크립터 추출

3.1 시스템 구성도

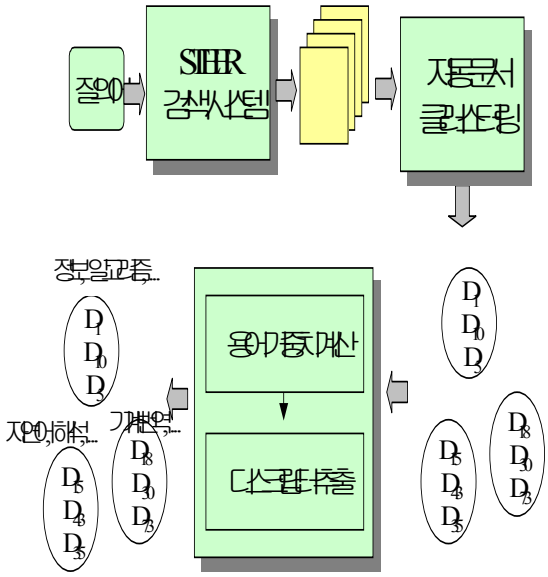

그림 1. 시스템 구성도

그림 1에는 제안하는 디스크립터 추출의 시스템 구성도이다. 검색시스템으로는 본 연구실에서 기개발된 STEER 검색시스템을 이용한다. STEER 검색시스템[8]으로 검색된 문서들에 대해 공통 키워드 빈도를 이용하여 문서간 유사도를 계산한다. 계산된 유사도를 기반으로 계층적 클러스터링을 수행한다. 마지막으로 용어 가중치 계산 과정을 거쳐 각 클러스터내의 문서들에서 디스크립터를 추출하는 과정을 수행한다.

3.2 검색 결과 클러스터링

검색 결과 문서 클러스터링은 문서간 유사도 계산 단계와 계층적 문서 클러스터링 단계로 이루어 진다.

첫째, 문서사이의 유사도를 계산하기 위해 공통키워드의 빈도를 이용한다. 이 방법은 "두 문서에 같은 형태소가 많이 나타날 수로, 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(2)와 같다.

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

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

둘째, 계층적 클러스터링 방법으로는 집단평균연결(Group Average Link) 방법을 이용한다. 이 방법은 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.

3.3 용어 가중치 계산

용어 가중치를 계산하기 위해 정보검색의 색인과정에 색인어의 가중치를 계산하기 위해 전통적으로 사용해온 $tf \times idf$ 계산식을 이용한다[1,6]. 용어 가중치 계산식은 식(3)과 같다.

$$w = tf * \log(N/n) + 1 \quad (3)$$

여기에서 tf는 문서 D에서 색인어 t_i의 발생 회수이고, $\log(N/n)$는 색인어 t_i가 나타나는 문서 D의 개

수의 역이고, N은 전체 문서에서의 문서의 수를 의미한다. 위의 가중치 계산식의 왼쪽 부분은 한 문서에서의 용어의 중요성을 의미하고, 오른쪽 부분은 전체 문서들 중에서 용어의 분별력을 표현한다.

이러한 *tf×idf* 기반 키워드 추출은 특별한 문서에 속하는 중요한 키워드를 인식하는데 유용하다. 예를 들어 표1에서는 Wall Street Journal의 항공 범주에서 선정한 한 문서와 *tf×idf*에 의해 추출된 상위 10개의 키워드를 보여준다.

표 3. 용어가중치에 의해 추출한 상위 10개 키워드

Rank	Term	Weight
1	lorenzo	19.90
2	holder	9.66
3	voting	9.05
4	proposal	8.03
5	50.7%	7.61
6	eastern	7.54
7	class	7.26
8	authorize	7.08
9	labor	6.42
10	airline	6.23

표1에서 보면 상위 3개의 용어 “lorenzo", "holder", 그리고 ”voting“은 이 문서를 대표하는 디스크립터로 볼 수 있다.

3.4 디스크립터 추출

가중치가 계산되어 내림차순으로 정렬된 키워드 중에서 1음절과 2음절 키워드를 삭제한다. 이것은 1음절과 2음절 키워드는 대부분이 의미가 포괄적인 단어들이기 때문이다. 또한 3음절이상인 단일 명사나 복합명사를 디스크립터로 추출하기 위함이다. 3음절이상의 키워드중에서 상위 5개를 각 클러스터의 디스크립터로 추출한다.

그림2에는 정보과학계열의 논문을 대상으로 검색한 결과를 클러스터링하고 디스크립터를 추출한 결과를 보여준다. 각 클러스터마다 의미있는 디스크립터들이 추출되었음을 알 수 있다.

5. 결론

본 논문에서는 검색 결과 문서의 클러스터링에서 필요한 디스크립터 추출 방안을 제안하였다. 디스크립터를 추출하기 위해 정보검색의 색인과정에서 사용하는 용어 가중치 계산 방법을 이용하였다. 또한 변별력이 있는 3음절 이상의 복합명사를 디스크립터로 추출하였다. 이 방법은 불용어를 제거과정이 불필요한 효율적인 방법이다.

향후에는 클러스터내의 문서들을 대상으로 디스크립터를 추출하기 위해 공기 패턴(co-occurrence pattern)이나 명사구를 추출하는 방법을 고려할 것이다.

그림 2. 디스크립터 추출의 결과화면

참고문헌

[1] Frakes, W.B., Baeza-Yates, R., Information

Retrieval, Prentice Hall, 1992.

[2] Hearst, M. A., Pederson, J.O., "Reexamining the cluster hypothesis: Scatter/Gather on retrieval results," SIGR'96, pp. 76-84, 1996.

[3] Leouski, A.V., Croft, W.B., "An evaluation of techniques for clustering search results," Technical Report IR-76, Dept. of Computer Science, University of Massachusettes, 1996.

[4] Lin, C.-Y., "Robust Automated Topic Identification," Ph.D. Dissertation, Dept. of Electrical Engineering-Sytem, University of Southern Califonia, Califonia, 1997.

[5] Sahami, M., Yusufali, S., Baldonado, M.Q.W., "SONIA: A Service for Organizing Networked Information Autonomously," Digital Libraries'98, pp. 200-209, 1998.

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

[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” 제 10회 한글 및 한국어 정보처리 학술대회, pp. 105-110, 1998.

[9] 오효정, 임정묵, 이만호, 맹성현, “유한 오토마타를 이용한 정보 추출 시스템의 구현 및 분석”, 제 10회 한글 및 한국어 정보처리 학술대회, pp. 97-104, 1998.

[10] 장동현, 맹성현, “문서 구조 정보를 이용한 확률 모델 기반 자동요약 시스템”, 제 9회 한글 및 한국어 정보처리 학술대회, pp. 15-22, 1997.

부록 1. Wall Street Journal의 항공범주 샘플문서

@WSJ870521-0028

TEXAS AIR CORP. holders approved a proposal that will increase Chairman Frank Lorenzo's voting power. The proposal doubled the voting power of each Class A common share to 10 votes. Mr. Lorenzo holds 50.7% of that class, which elects three-quarters of the Houston airline holding_company's directors. Holders also cleared an increase in authorized common to 200 million shares from 75 million. An amendment to lift authorized preferred_shares to 50 million from 10 million was withdrawn, however, because the company received less than 50% of the preferred proxies required to vote on the proposal. Mr. Lorenzo told holders that the company still is committed to shrinking labor costs at its Eastern Airlines, despite the unit's break-even first quarter. "We wouldn't expect this type of performance in a bad year because Eastern labor costs are just too high," he said.