클러스터링 기법을 이용한 키워드 유사도 순위화 알고리즘에 따른 사용자 질의 확장

이상훈° 김기태
중앙대학교 컴퓨터공학과

User Query Expansion Through Keyword Similarity Ranking Algorithm Using Clustering Methods

Sang-Hoon Lee, Ki-Tae Kim
Dept. of Computer Science & Engineering, Chung-Ang University

요 약

본 논문에서는 여러 가지 클러스터링 기법들을 사용하여 키워드 유사도를 순위화하여 사용자의 질의를 확장하는 기법을 제안한다. 클러스터링 기법에는 연관(Association) 클러스터링, 메트릭(Metric) 클러스터링, 스칼라(Scalar) 클러스터링 기법을 사용하고, 이들간의 가중치를 적절히 조절하여 검색 시스템을 만든다. 사용자의 질의가 주어졌을 때, 질의 키워드와 연관된 키워드들을 순위화 하여 사용자에게 보여주고, 사용자의 추가입력을 받아서 질의를 확장한다. 사용자가 적당한 질의어로 판단하여 확장된 질의로 검색을 수행할 때까지 이 과정을 반복한다. 실험에서 사용한 문헌집합은 Korea Herald의 2003년 1월과 2월의 경제 관련 기사들을 수집하여 사용하였고, 실험을 거쳐서 질의를 확장한 결과 만족할 만한 결과가 도출되었다.

1. 서론

일반적으로 사용자가 검색을 할 때에는 자연어 질의를 하기보다는 단일 키워드 혹은 여러 키워드를 혼합해서 질의로 입력하고, 검색된 문서들에서 자신이 원하는 것을 찾아낸다. 사용자가 검색환경이나 자신이 찾고자 하는 문헌에 대한 정확한 정보를 가지고 있다면 이러한 검색 방법이 효과적 일 수 있지만, 정확한 정보를 가지고 있지 않다면, 원하지 않는 문헌들이 검색될 수 있다[2].

본 논문에서는 사용자의 질의를 보다 효과적으로 확장하여 사용자가 원하는 문헌을 찾을 수 있도록 도와주는 기법을 제안하고자 한다. 클러스터링의 개념을 이용해서 키워드 간의 유사도를 순위화 하여 저장하고, 사용자의 질의가 들어왔을 경우, 질의 키워드와 연관된 키워드들을 순위에 따라 사용자에게 보여주고, 사용자의 선택에 따라 사용자의 질의를 확장하여 사용자의 검색을 도와준다.

질의 확장을 위해서 클러스터링 기법을 사용하는 것은 정보검색 초기부터 시도된 기본적인 방법으로, 본 논문에서는 문헌 내에서 두 키워드가 공기 하는 정도에 따라 연관성을 계산하는 연관(Association) 클러스터링 기법과 두 용어간의 문헌 내에서의 거리에 따라 연관성을 계산하는 메트릭(Metric) 클러스터링 기법, 두 용어의 이웃 하는 용어에 따라 연관성을 계산하는 스칼라(Scalar) 클러스터링 기법을 적절히 조절하여 사용한다.

본 논문의 구성은 다음과 같다. 2장에서는 질의 확장과 본 논문에서 사용한 클러스터링 기법에 대해서 자세히 알아보고, 3장에서는 이러한 클러스터링 기법을 이용해서 구현한 검색 시스템에 대해서 설명하고, 여기에 사용된 문헌집합에 대해서 알아본다. 4장에서는 검색 시스템의 실험 결과를 보이고, 5장에서는 실험 결과에 대한 분석과 추후 연구과제를 기술한다.

2. 관련 연구

정보 검색에 있어서 사용자의 질의 확장에 대한 연구는 많이 있어왔다. 여기에는 사용자 연관 피드백을 이용하는 방법과 처음 검색된 문헌 집합에서 추출된 정보를 이용하는 방법, 문헌 집합 전체에서 추출된 정보를 이용하는 방법이 있다[2, 3].

본 논문에서는 전체 문헌 집합에서 작성된 키워드 유사도 순위를 이용해서 사용자의 연관 피드백을 받아 질의를 확장한다. 연관 피드백 방법은 가장 널리 사용되는 질의 확장의 방법으로[2, 4], 초기에 검색된 문헌 집합에서 사용자가 연관 문헌들을 선택하면 그 문헌 내의 키워드들의 가중치를 재계산하여 질의를 재작성 하는 방법으로, 본 논문에서는 이것을 연관된 키워드를 순위화 하여 보여주고 사용자의 키워드 추가 선택에 의해 질의를 확장하는데 사용한다.

문헌 집합에서 키워드의 연관성을 판단하는 방법으로는 전통적으로 문헌 내에서의 공기 관계를 이용하는 연관 클러스터링 방법이 많이 쓰여 왔다[2]. 그 정의는 다음과 같다.

> 전체 문헌 집합 D 내의 문헌 d_j 내에서의 어떤 키워드 k_i의 빈도를 $f_{k_i,j}$로 표시할 때, 문헌 d_j 내에서의 두 키워드 k_u, k_v간의 유사도 $S_{u,v}$는 다음과 같이 구할 수 있다.
>
> $$S_{u,v} = \sum_{d_j \in D_i} f_{s_u,j} \times f_{s_v,j}$$

연관 클러스터링 기법은 문헌 내에서 어떤 키워드 쌍이 공기 하는 빈도만을 고려하고, 그 키워드들의 출현 위치는 고려하지 않는다. 하지만, 문헌 내에서 가까이에 위치하고 있는 키워드가 멀리 떨어져 있는 키워드 보다 더 연관성이 크다고 볼 수 있기 때문에, 키워드 사이의 거리를 고려하는 것도 의미가 있다. 메트릭 클러스터링 기법은 이와 같은 아이디어에 기반하고 있다[2, 5].

> 두 키워드 k_i와 k_j 사이의 거리 $r(k_i,k_j)$를 한 문헌 내에서의 두 키워드 사이에 출현하는 키워드의 수로 정의하면, 하나의 문헌 내에서의 두 키워드 k_u, k_v간의 유사도 $S_{u,v}$는 다음과 같이 구할 수 있다.
>
> $$S_{u,v} = \sum_{k_i \in V} \sum_{k_j \in V} \frac{1}{r(k_i,k_j)}$$

위의 식에서 V 는 문헌 내의 전체 키워드 집합을 의미한다. 다른 여러 논문들에서 $\frac{1}{r^2(k_i,k_j)}$ 과 같은 여러 가지 다른 변형들이 제안되었으나 차이가 별로 없음이 판명되었다[2].

스칼라 클러스터링은 두 키워드 간의 연관성을 측정하는 또 다른 방법으로 두 키워드가 비슷한 이웃을 가지면 두 키워드가 연관되어 있다고 판단하는 것이다. 이것을 측정하는 방법으로는 벡터 $\vec{k}_u$ 의 모든 연관 계수 $k_{u,i}$ 를 나열하고 다른 벡터 $\vec{k}_v$ 의 모든 연관 계수 $k_{v,i}$ 를 나열하여 두 벡터의 스칼라 척도를 이용하는 것이다.

$\vec{k}_u = (k_{u,1}, k_{u,2}, \cdots, k_{u,n})$, $\vec{k}_v = (s_{v,1}, s_{v,2}, \cdots, s_{v,n})$ 을 키워드 k_u 와 k_v 의 연관 계수 벡터라고 하면, 두 키워드 간의 유사도 $S_{u,v}$ 는 다음과 같이 구할 수 있다.

$$S_{u,v} = \frac{\vec{s}_u \cdot \vec{s}_v}{|\vec{s}_u| \times |\vec{s}_v|}$$

3. 질의 확장 시스템

본 논문에서 사용한 시스템의 개요는 아래의 그림 1과 같다. 이 시스템은 전체 문헌 집합에서 키워드간의 유사도를 분석해서 저장한 다음 사용자의 질의가 입력되면 질의의 키워드와 유사한 키워드를 순위화해서 사용자에게 보여주고, 사용자의 선택에 따라 추가적인 질의 확장과 검색이 이루어지는 시스템이다.

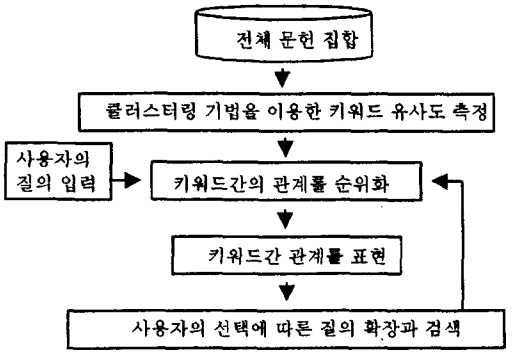

그림 1. 질의 확장 시스템 개요

3.1 문헌 집합

실험에서 사용한 문헌 집합은 Korea Herald의 2003년 1월과 2월의 경제 관련 기사 500개를 사용하였다. 여러 분야의 문헌들이 혼합되어 있는 문헌 집합 보다는 같은 분야의 문헌들 사이에서 키워드들간의 연관성을 분석하는 것이 더 용이 하기 때문에 경제 분야의 문헌들 만을 문헌 집합으로 선정하였다.

3.2 키워드 추출

Html 형식으로 기술되어 있는 문헌들을 하나씩 읽어 들여 html 태그들은 제거하고, 내용만을 읽어 들인다. 공백을 기준으로 키워드를 추출하고, Francis와 Kucera가 1982년에 제안한 영어 불용어(a, to, in, and 등) 425개를 제거하여 키워드를 선정한다[8].

3.3 키워드간의 유사도 측정

2장에서 언급한 세 가지의 클러스터링 기법을 이용해서 전체 문헌 집합 내에서의 키워드간 유사도를 측정한다. 문헌 집합의 문헌에 각각 문헌 번호를 부여하고, 각각의 문헌 내에서의 키워드 간의 유사도를 측정한다. 하나의 문헌 내에서의 유사도는 0~1 사이의 값으로 측정하는데, 세 가지의 클러스터링 기법 중 가장 좋은 성능을 내는 메트릭 클러스터링에 많은 가중치를 두고 측정하고, 하나의 문헌에서 측정된 키워드간 유사도를 모두 더해 전체 문헌 집합에서의 키워드간 유사도를 측정한다.

새로운 문헌이 문헌 집합에 추가 되면 추가된 문헌의 키워드간 유사도를 측정해서 전체 유사도에 더해주면 된다.

3.4 사용자 연관 피드백을 통한 질의 확장과 검색

사용자의 질의가 입력되면 질의를 키워드로 분리하여 순위화된 유사한 키워드를 상위 10개까지 사용자에게 보여주고, 상위 10개 이후의 키워드들은 사용자가 원할 경우, 추가적으로 보여준다. 사용자가 원하는 키워드를 선택하면, 사용자가 선택한 키워드로 질의를 확장하여 그 질의와 유사한 새로운 키워드를 순위화해 보여준다. 여러 개의 키워드로 구성된 질의는 각각의 키워드와의 유사도가 아니라 전체 키워드와의 유사도를 측정한다.

사용자가 적당한 질의로 판단하여 검색을 선택하면 그 질의와 연관된 문헌들을 검색해 준다. 연관된 문헌을 판단하는 기준은 질의내에 포함된 각 키워드들간의 유사도의 합이 가장 큰 문헌을 순위화에 검색해준다.

4. 실험 및 평가

논문에서 제안한 질의 확장 시스템에 임의의 질의 키워드를 넣고, 질의를 확장하는 실험을 수행하였다. 첫번째 실험에서는 질의어로 Samsung을 입력하였고, 그 결과는 아래의 그림 2, 3과 같다. 두번째 실험에서는 질의어로 Hyundai를 입력하였고, 그 결과는 그림 4, 5와 같다.

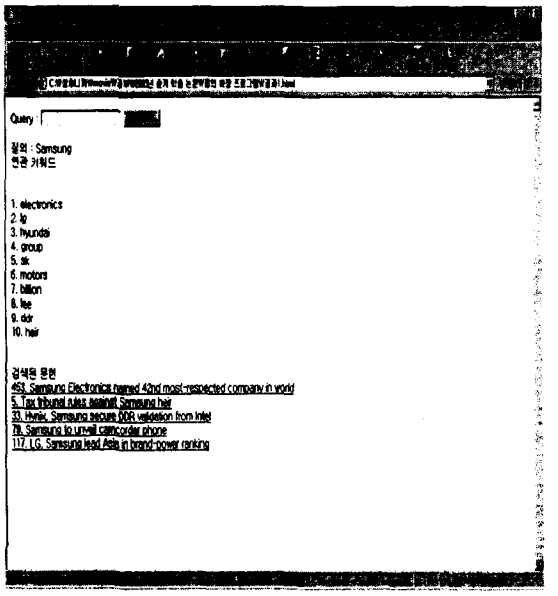

그림 2. 질의어로 Samsung을 입력한 경우

위의 결과에서 heir를 추가 키워드로 선택한 경우의 결과는 다음의 그림3과 같다.

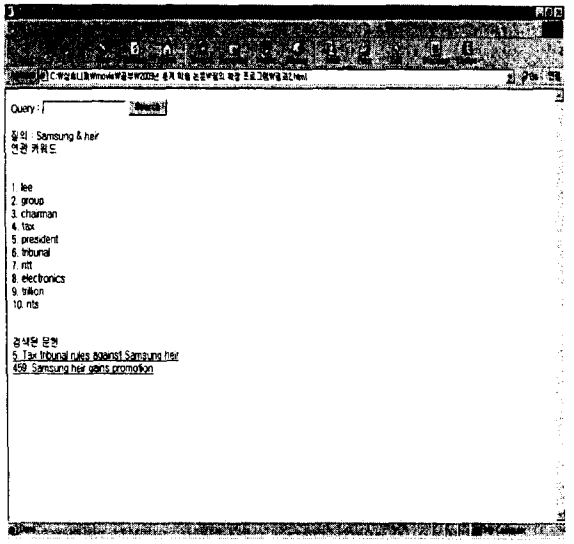

그림 3. 질의 확장을 위해 추가로 heir을 선택한 경우

질의 확장을 통한 위의 실험에서 삼성그룹의 후계자에 관련된 문헌을 찾기를 원했다면, 결과에서 보듯이 충분히 만족할 만한 결과가 나왔다는 것을 볼 수 있다.

아래의 그림 4는 질의어로 Hyundai를 입력한 실험 결과이다.

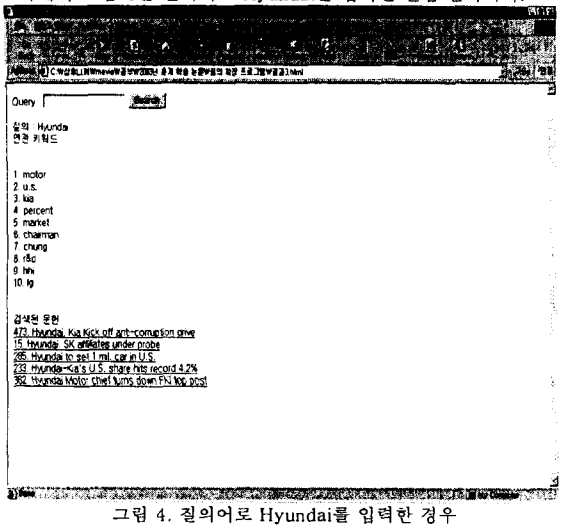

그림 4. 질의어로 Hyundai를 입력한 경우

위의 결과에서 R&D를 추가 키워드로 선택한 경우의 결과는 다음의 그림 5와 같다. 현대의 R&D 관련 문헌을 찾기를 원했다면, 충분히 만족할 만한 결과가 나왔다는 것을 볼 수 있다. 하지만, 현대의 R&D에 대한 문헌이 충분치 않아서 그다지 많은 문헌이 검색되지는 않았다.

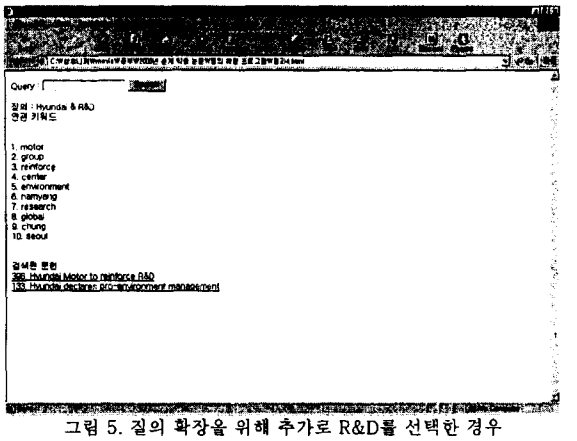

그림 5. 질의 확장을 위해 추가로 R&D를 선택한 경우

5. 결론 및 추후과제

질의 확장에 대한 결과를 판단하는 것은 사용자의 주관적인 판단이기 때문에 정확한 성능 평가를 할 수는 없지만, 실험의 결과에서 보듯이 만족할만한 결과가 도출되었으므로, 클러스터링 기법을 이용한 키워드 유사도 측정이 질의 확장을 위한 유사 키워드 순위화에 유용하다는 것이 증명되었다.

하지만, 경제 분야만의 소형 문헌집합으로는 시스템에 대한 신뢰성을 보장할 수 없기 때문에, 좀 더 많은 문헌과 여러 분야의 문헌들에 대한 실험이 필요하다. 그리고, 여러 분야에서 동음이의어로 쓰이는 키워드에 대한 처리와 스테밍 알고리즘의 적용이 필요하고, 실제 웹 환경에서 일반 사용자들을 통한 성능 평가 실험이 필요하다.

참고문헌

1. R. Attar and A. S. Fraenkel, **Local Feedback in Full-Text Retrieval Systems,** Journal of the ACM 1977 pp. 397-417.
2. Ricardo Baeza-Yates and Berthier Ribeiro-Neto, **Modern Information Retrieval,** 1999 by the ACM press pp. 117-139.
3. Jinxi Xu and W.Bruce Croft, **Query Expansion Using Local and Global Document Analysis.**
4. Chia-Hui Chang and Ching-Chi Hsu, **Integrating query expansion and conceptual relevance feedback for personalized Web information retrieval.**
5. Chavez, E. and Navarro, G. **An effective clustering algorithm to index high dimensional metric spaces.** String Processing and Information Retrieval, 2000. SPIRE 2000. Proceedings. Seventh International Symposium on, 2000. pp. 75-86.
6. Mandar Mitra and Amit Singhal and Chris Buckley. **Improving Automatic Query Expansion.**
7. Chris Buckley and Mandar Mitra and Janet Walz and Claire Cardie. **Using Clustering and SuperConcepts Within SMART : TREC6.**
8. W. Francis and H. Kucera. **Frequency Analysis of English Usage.** Houghton Mifflin, New York, 1982.
9. S. Robertson. **On term selection for query expansion.** Journal of Documentation. 46. pp 359-364. 1990.
10. Ryen W. White and Ian Ruthven and Joemon M. Jose. **Finding Relevant Documents using Top Ranking Sentences : An Evaluation of Two Alternative Schemes.**