계층적 클러스터링을 이용한 텍스트 정보 검색 모델

신동호[1], 장병탁[1,2]
1.서울대학교 인지과학 협동과정, 2.서울대학교 컴퓨터공학과
Email: {dhshin, btzhang}@scai.snu.ac.kr

A Retrieval Model Using Cluster Information

Dong-Ho Shin, Byoung-Tak Zhang
1.Interdisciplinary Program in Cognitive Science. Seoul National University
2. Department of Computer Engineering. Seoul National University

요 약

일반적인 정보검색 시스템은 백만 건 이상의 문서 집합을 미리 색인하여 저장해 놓고, 사용자로부터 몇 단어 안팎의 질의어를 입력받아 둘 사이의 유사도(query-document similarity)를 계산하여 우선 순위별로 뽑아 나열해 준다. 한편, 문서 클러스터링은 문서집합에서 문서들간의 상대적인 유사도(document-document similarity)를 계산하여 분류한다. 유사도 계산은 단어 벡터로 표현(represent)된 문서와 질의어 사이에 내적이나 코싸인 각도로 계산되는데, 질의어는 몇 단어 안팎에 지나지 않기 때문에 질의어와 문서와의 유사도 계산으로 산출된 검색 결과와 문서와 문서간의 유사도 계산으로 산출된 클러스터링 정보는 그 성격이 다르다. 본 논문에서는 이와 같이 상이한 두 유사도 정보를 효율적으로 결합하여 검색 성능을 높이는 방법에 대해 다룬다. 실험에서는 TREC의 데이터 집합과 평가 방법을 따랐다.

1. 개 요

정보검색 시스템은 비교적 변화가 없는 백만건 이상의 문서들을 미리 저장 및 색인해 놓고 사용자의 다양한 질의어가 입력될 때마다 질의어와 문서와의 유사도를 계산하여 적절성 여부를 판별해 결과로 보여준다. 문서를 색인하는 과정에서는 가능한 모든 색인어들을 추출하여 역화일 구조로 저장하기 때문에 한 문서는 여러개의 색인어를 가지고 따라서 서로 다른 여러 형태의 질의어에 의해 검색(retrieval)될 수 있다.

한편, 사용자 질의어는 몇 단어 안팎의 짧은 문장으로 되어 있다. 질의어의 단어들은 사용자가 신중하게 직접 고른 것이기 때문에 사용자의 정보 요구를 잘 표현한다고 볼 수 있겠으나 단어 수가 워낙 적기 때문에 자연언어가 가지는 중의성(ambiguity)의 문제가 내재해 있다. 일례로 웹에서 브라우저의 정보를 서버측에 전달하기 위해 설정하는 것을 'cookie'라 하는데 검색 엔진을 통해 검색해 보면, 작고 납작한 케이크를 뜻하는 쿠키와 관련된 사이트도 검색되어 나온다. 검색자는 분명 한 영역에서 사용하는 특징적인 의미로 질의를 한 것이지만, 이런 경우는 질의어 자체가 중의성을 가지고 있기 때문에 검색 결과가 잘못 되었다고 할 수는 없을 것이다.

질의어의 형태는 검색 결과에 큰 영향을 미치기 때문에, TREC에서는 질의어에 해당하는 topic의 형식을 title, desc, narr로 구분하고 이 중 어느 항목들을 선택하여 질의어를 만드는 것이 효과적인지를 분석하고 있다. TREC은 미국의 NIST에서 주관하는 세계 최대 규모의 정보검색 대회로 정보검색의 표준 평가 방법 개발등을 목적으로 하고 있다. 매년 2기가 바이트 분량의 텍스트 데이터로부터 미리 정해진 50개의 질의어 각각에 대해 1,000 개씩 뽑아 제출토록 규정되어 있다 [1].

데이터베이스 시스템에서는 검색자가 자신의 정보 요구를 정확히 파악하고 있고 또한 그 요구를 정확히 질의어로 만들 수 있다고 가정하는데 반해 정보검색 시스템에서는 검색자도 자신의 정보 요구를 잘 알지도 못하고 정확히 표현하지도 못한다고 가정한다. 그러므로 정보검색 시스템에서는 이 문제를 해결하기 위해 적합도 피드백(relevance feedback), 질의어 확장(query expansion)의 기법을 사용하여 보충한다.

검색이 비교적 많은 수의 단어로 이루어진 문서와 비교적 적은 수

의 단어로 이루어진 질의어간의 유사도를 비교하여 이루어지는데 반해 문서 클러스터링은 비교적 많은 수의 단어들을 가지고 있는 문서들 간에 상대적 유사도를 계산하여 이루어진다.

분류(classification)는 검색과는 달리 미리 정의된 주제에 따라 문서를 나누는 방식이기 때문에 설계자가 원하는 대로 주제를 학습시킬 수 있는데 반해, 클러스터링(clustering)은 대표적인 무감독학습(unsupervised learning) 기법으로 최종 클러스터링 된 결과가 과연 설계자가 원하는 대로 된 것인지에 대해서는 확신할 수 없다. 이러한 근본적인 어려움에도 불구하고 클러스터링은 나름대로의 의미를 가지고 있는데, 첫째는 문서간에 전반적인 비교가 이루어진다는 것이고 둘째는 설계자가 미처 예상치 못한 주제도 찾아 낼 수 있다는 것이다.

본 논문에서는 질의어-문서 유사도, 문서-문서 유사도라는 상이한 성격의 두 정보를 통합하여 검색 효율을 개선하는 방법을 다룬다.

2. 클러스터링과 검색

i번째 문서 d_i는 서로 독립이라고 가정된 n차원의 단어 벡터로 표현된다.

$$d_i = < t_1 \ldots t_n >$$

그런데 문서내의 단어들에 따라 그 문서를 표현하는 정도가 다르기 때문에 일반적으로 아래와 같은 가중치식 $tf \cdot idf$ 을 사용하여 나타낸다 [7]:

$$w_{ij} = tf_{ij} \cdot \log(\frac{N}{df_j}),$$

여기에서 w_{ij} 는 i 번째 문서의 j번째 단어의 가중치를 나타내고, tf_{ij} 는 i번째 문서의 j번째 단어의 빈도수를 나타내며, N은 전체 문서 수, df_j는 j번째 단어가 나타난 문서수를 나타낸다. 문서 클러스터링은 문서들간의 유사도에 따라 결정되는데, 두 문서 d_i, d_j의 유사도는 코싸인 유사도로 계산한다.

$$S(d_i, d_j) = \cos\theta = \frac{d_i \cdot d_j}{|d_i||d_j|} = \frac{\sum_{k=1}^{n} d_{ik} \cdot d_{jk}}{\sqrt{\sum_{k=1}^{n} d_{ik}^2}\sqrt{\sum_{k=1}^{n} d_{jk}^2}}$$

문서 클러스터링은 한 문서와 다른 모든 문서들과의 상대적 유사도에 의해 결정되므로 대규모 문서 집합에서는 클러스터링을 하기가 쉽지 않다. 본 실험에서는 회귀적인 방법으로 계층적 클러스터링(hierarchical clustering)을 하였다 (그림1).

클러스터링시 어느 정도 유사한 것들끼리 묶어 한 클러스터를 형성할 것인가를 결정하는 기준은 크게 두 가지 방식이 있는데, 하나는 특정 임계치(threshold)를 두어 거리가 그 값 이하인 것들을 하나로 묶는 방법이 있고 다른 하나는 클러스터의 수를 결정해 놓고 그 수에 맞도록 임계치를 자동 조절하여 클러스터를 만들어 가는 방법도 있다. 본 실험에서는 후자의 방법을 따랐다.

문서 집합 D는 계층도의 깊이 1에서 M개의 클러스터로 나누어

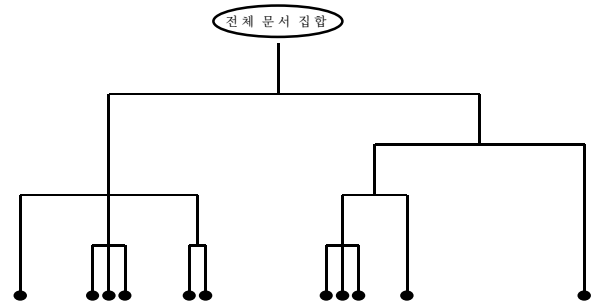

그림 1: 전체 문서 집합의 계층적 클러스터링

지고, 다음 단계에서는 M개의 클러스터들이 다시 각각 M개의 클러스터로 나누어진다. 따라서 각 단계 l 에서의 클러스터 개수는 M^l 이다.

D의 부분집합(한 클러스터) D'을 M개의 클러스터로 나누기 위해 임계치(threshold)를 자동 조절하는 방법을 사용하였다. 각 클러스터는 문서 리스트와 프로토타입 문서 벡터를 가지고 있다. 새로 들어온 문서와 프로토타입 문서 벡터와의 거리가 임계치 이하면 그 클러스터에 추가하고 프로토타입 벡터를 업데이트한다. 만약 모든 클러스터로부터의 거리가 임계치 이상이 되면 서로 최소 거리에 있는 클러스터들을 합병(merge)하여 새로운 문서 리스트와 프로토타입 문서 벡터를 만들고 임계치를 합병된 클러스터간의 거리로 상향조정한다. 전체적인 도식은 그림 2와 같다.

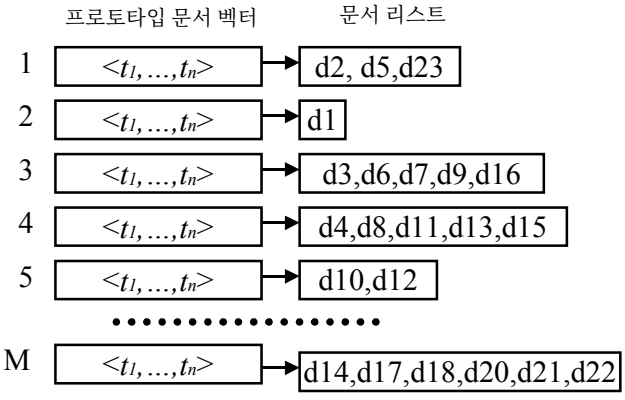

그림 2: 부분 문서집합 D'을 M개의 클러스터로 나누기

주어진 문서집합에서 입력된 질의어에 따라 검색하는 방법은 문서-문서 유사도 계산과 동일한 방법을 따랐다. 문서 클러스터링의 결과는 문서들간의 전반적인 비교의 결과로 생성된 것인데 반해 질의어는 사용자가 입력하는 그 때마다 새로운 주제(정보 요구)가 되며 문서와의 부분 비교에 의한 결과이다. 그림 3은 전체 문서 집합에서 문서들간의 클러스터 구조와 검색의 관련성을 보여준다.

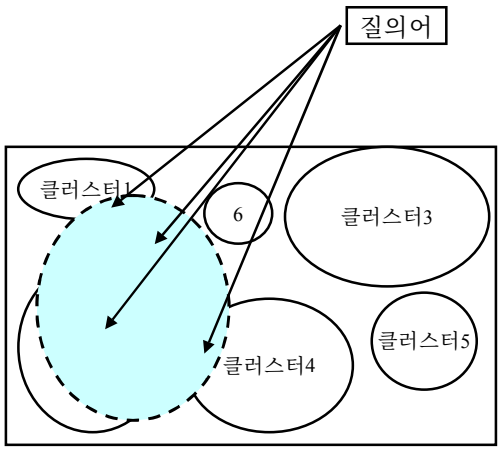

그림 3: 문서 클러스터링과 검색

3. 실험결과

3.1 TREC-7 Ad Hoc 실험 분석

실험은 TREC-7 Ad Hoc 과제 형식에 따라 진행하였다. TREC Ad Hoc 과제는 여러 종류의 신문, 잡지들로부터 얻은 대략 52만개 (2GB)의 문서집합으로부터 주어진 50개의 주제에 대해 각각 적절하다고 판단되는 1000개의 문서를 뽑아 검색율을 측정한다. 방대한 양의 자료이기 때문에 상용 정보검색 시스템의 평가 기준으로도 손색이 없고 표준적인 평가방법이기 때문에 신뢰할만 하다.

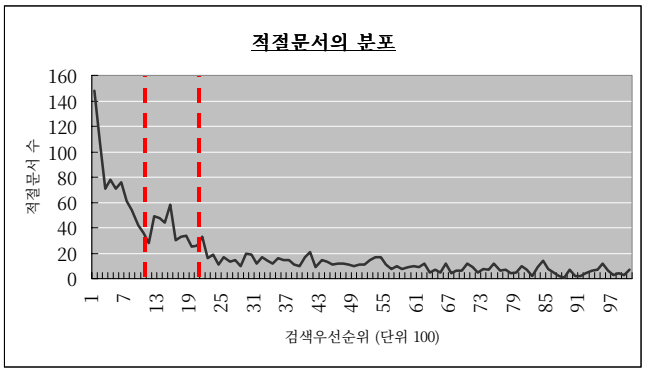

그림 4: 검색 결과 적절 문서의 분포도

그림 4는 단순한 벡터 공간 모델을 이용하여 질의어-문서간의 유사도에 따라 우선순위별로 나열했을 때, 적절 문서들이 나타나는 분포를 보여주고 있다. 두 개의 점선은 각각 1000번째와 2000번째를 가리킨다. 유사도가 높을수록 적절 문서일 확률이 높기 때문에 상당부분은 1000번째 안쪽에 위치하게 되는데 여기까지가 TREC 평가 방법에서 규정한 검색 문서수이다. 1000번째에서 2000번째 사이에도 많은 적절 문서들이 존재하는데, 이 문서들은 검색에 실패한 문서들이다.

이와 같이 단순한 벡터 공간 모델을 이용해서 TREC7의 데이터를 분석해 본 결과 재현율(recall)은 77.79%이다. 따라서 어근화(stemming)나 질의어 확장을 하지 않았는데도 적절 문서의 상당 부분이 검색되어 나온 것이다. 단지 유사도 계산에서 상위 1000등 안에 못 드는 것이 정확률(precision)을 떨어지게 하는 주 요인인 것이다. 본 논문에서는 어떤식으로든 검색되어 나온 문서들의 리스트에 그 문서집합으로부터 구한 클러스터 정보를 적용, 유사도를 재계산하여 정확률을 올리는 방법을 다룬다.

클러스터 수	평균 적절 클러스터 비율
10	43.6 %
100	9.7 %
1000	0.017%

표 1: 적절 클러스터 비율

표 1은 전체 문서집합을 계층적으로 10개, 100개, 1000개로 클러스터링 했을 경우, 질의어에 의한 검색을 통해 나오는 적절문서들이 존재하는 클러스터의 비율을 나타내고 있다. (앞으로 적절문서들이 나오는 클러스터를 '적절 클러스터'라 하겠다.) 10개의 경우에도 50%를 넘지 못하는 것으로 볼 때, 그림 3에서 직관적으로 제시한 검색과 클러스터링의 관계가 옳다는 것을 강력하게 시사한다.

3.2 클러스터 정보의 반영

이제 남은 문제는 적절 클러스터를 찾아내어 효율적으로 반영하는 방법이다. 클러스터링은 대표적인 무감독(unsupervised) 학습으로 비교적 어려운 문제인데다가 52만개의 문서집합 전체에 대해 이루어진 클러스터링의 결과를 어느 정도 신뢰할 수 있을지 알기 어렵다. 한 가지 분명한 사실은 클러스터링이 주제별로 잘 이루어질수록 검색결과에 좋은 영향을 줄 수 있다는 것이다.

클러스터의 적절성을 판별하는데 유일하게 주어진 자료는 질의어-문서 유사도로 주어진 문서 리스트이다. 그림 4)에서 보듯이 상위 순위에 들수록 좀 더 적절한 문서일 가능성이 커진다. 따라서 어느 클러스터가 상위의 문서들을 많이 가지고 있는냐에 따라 적절 여부를 계산하는 것은 타당해 보인다.

클러스터의 적절성을 판별하는데 사용할 문서의 수를 N, 개별 문서들의 유사도 점수를 y_i라고 하자. 그러면 상위 N개 문서들의 평균 $\overline{Y}$와 j 번째 클러스터의 평균 $\overline{c_j}$는 다음과 같다.

$$\overline{Y} = \sum_{i=1}^{N} y_i$$

$$\overline{c_j} = \frac{\sum_{y_i \in c_j} y_i}{n_j}$$

위에서 n_j는 j번째 클러스터에 포함되는 문서수이다.

각 클러스터는 $\frac{\overline{c_j}}{\overline{Y}}$ 만큼의 가중치를 얻게 된다. 즉, 상위 N개의 문서 중 높은 점수의 문서를 많이 가지고 있는 클러스터는 나머지 문서들도 그에 상응하는 가중치를 갖게 된다.

$$s'_{jk} = s_{jk}\frac{\overline{c_j}}{\overline{Y}}$$

최종 결과는 재조정된 유사도값 s'_{jk} 으로 평가한다.

s'_{jk}의 값에 따라 각 질의어에 대해 1,000개의 문서를 뽑아 제출하면 TREC의 방식에 따라 평가를 하게 된다. 일반적으로 검색결과를 평가하는 방법에는 여러 가지가 있는데, TREC에서는 아래 그림 5,6에서와 같은 두 가지 방법으로 평가한다. 그림 5는 서로 반비례 관계에 있는 재현율과 정확률을 함께 측정하는 것으로 각각의 재현율에 대한 정확률이 표현되어 있다. 그림 6은 문서를 몇 개 뽑는냐에 따라 정확률을 나타내고 있다. 대체적으로 많이 뽑을수록 정확률은 떨어진다.

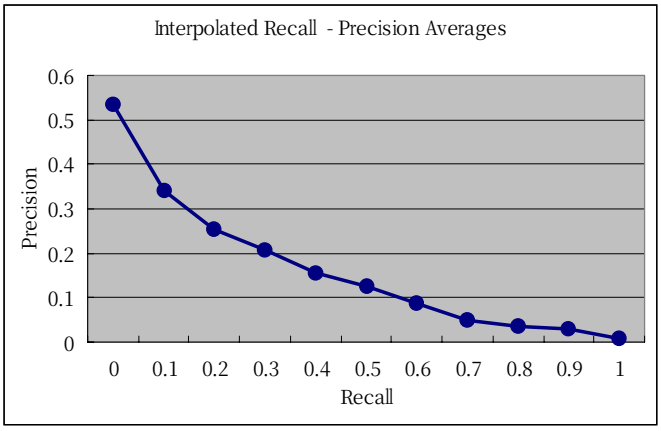

그림 5: TREC-7 재현율-정확률 곡선

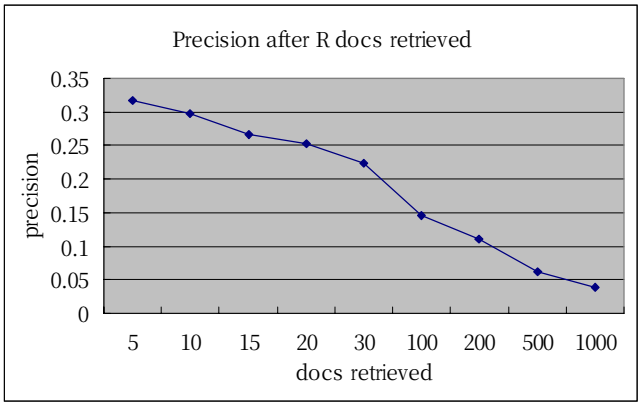

그림 6: 검색된 문서가 R개일 때의 정확률

4.결론

본 연구에서는 질의어-문서 유사도로 이루어지는 TREC-7 검색 환경에 문서-문서 유사도에 의한 클러스터링 정보를 효과적으로 반영하는 방법에 대해 고찰해 보았다. 데이터에 대한 여러 가지 분석들은 본 방법이 타당함을 보여준다. 단, 검색시 클러스터 정보를 반영하기 위해서는 정확한 클러스터링 작업이 사전에 이루어져야 하는데 지금까지 제안된 많은 방법들에도 불구하고 여전히 어려운 문제로 남아 있다. 좀 더 정확한 클러스터링 정보를 가지고 있다면 본 논문에서 제안한 기본 통계학적인 방법 이상의 것들도 적용시켜 볼 수 있을 것이다. 또한, 본 연구는 인터넷상에서 검색된 문서들을 자동으로 분류해 주는 기술에도 쉽게 접목될 수 있으리라 생각된다.

감사의 글 : 본 연구는 정보통신부에서 시행한 대학기초 연구기술 지원사업(98-199)에 의해 지원을 받았음

참고문헌

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

[2] Korfhage, Robert R., *Information Storage And Retrieval*, John Wiley & Sons, 1997.

[3] Lee, J.H., Analyses of Multiple Evidence Combination, *SIGIR-97*, pp.267-276, 1997.

[4] Miller, G.A., Five papers on WordNet, *International Journal of Lexicology*, vol.3, no.4, 1990.

[5] Robertson, S.E. and Sparck-Jones, K., Relevance weighting of search terms. *Journal of the American Society for Information Science* 27:-1976.

[6] Rocchio, J., Relevance feedback information retrieval, In G. Salton, editor, *The Smart Retrieval System - Experiments in Automatic Document Processing*, Prentice-Hall, pp.313-323, 1971.

[7] Salton, G., *Automatic Text Processing*, Addison-Wesley, 1989.

[8] Silverstein, C., Perdersen, J. O., Almost-Constant-Time Clustering of Arbitrary Corpus Subsets, *SIGIR-97*, pp.60-66, 1997.

[9] Yang, Y., Noise Reduction in a Statistical Approach to Text Categorization, *SIGIR-95*, pp.256-263, 1995.