(제 10회 한글 및 한국어 정보처리 학술대회)

문서 클러스터를 이용한 재순위화 모델

이경순 박영찬* 최기선

한국과학기술원 전산학과
대전시 유성구 구성동 373-1, 우:305-701
{kslee,kschoi}@world.kaist.ac.kr

*한국전자통신연구원 자연어정보처리부
대전시 유성구 가정동 161, 우:305-350
parkyc@etri.re.kr

Document Reranking Model Using Clusters

Kyung-Soon Lee Young-Chan Park* Key-Sun Choi

Dept. of Computer Science
KAIST

*Dept. of Natural Language Information Processing
ETRI

요약

본 연구[1]에서는 정보검색시스템의 모델로 문서 클러스터를 이용한 재순위화 모델을 제시한다. 이 방법은 검색단계와 분석단계로 이루어지는데, 검색단계에서는 역화일기법을 이용해서 질의어를 포함하는 문서들을 검색하여 질의어-문서 유사도에 따라 순위를 결정한다. 분석단계에서는 이미 구축된 문서 클러스터를 이용해서 검색되어진 문서들의 분석을 통해 질의어-클러스터 유사도를 계산한다. 질의어-문서 유사도와 질의어-클러스터 유사도를 결합하고, 이 유사도에 기반해서 문서들을 재순위화한다. 이때 이용하는 클러스터는 정적 클러스터이고, 질의어에 따라 서로 다른 클러스터를 생성하는 동적인 뷰를 제공한다. 재순위화 모델은 역화일 기법과 클러스터 분석 기법이 가지는 장점을 결합하여 질의어 뿐만 아니라 문서에 포함된 모든 단어들을 분석함으로써 문서의 문맥을 고려할 수 있다. 제안하는 모델은 역화일 기법을 이용한 검색 결과에 비해서 우수한 성능 향상을 나타내고 있다.

1 서론

정보검색의 대표적인 방법으로 역화일 기법(inverted file method)은 질의어에 나타난 키워드들이 문서에서 어느 정도의 가중치를 가지고 존재하느냐를 기준으로 문서들을 순서화 한다. 역화일 기법은 사용자의 질의어에 나타난 단어들이 문서내에 존재하는지 존재하지 않는지를 단순히 검색할 뿐 문서에 포함된 다른 단어들 즉, 문맥을 고려하지 않기 때문에 관련 있는 문서를 찾아내는 능력에 있어 한계가 있다.

정보검색을 위한 또다른 방법인 클러스터 분석은 질의어에 적합한 문서를 검색하기 보다는 문서를 어떠한 범주에 할당하거나 검색 결과를 보다 효과적으로 제공하기 위한 방법으로 많이 이용되고 있다. 정보검색에서의 질의어는 통계적으로 의미있는 빈도 벡터를 얻기에는 너무 적은 양의 단어들로 구성되기 때문에 클러스터 분석기법은 질의어-문서의 유사도를 계산하기에는 부적절하지만[10], 문서-문서의 유사도를 측정하기에는 적절한 방법이다.

일반적으로 사람이 질의어와 문서와의 관련도를 측정할 때는 어떠한 문서가 사용자의 질의어에 포함된 단어들을 적게 포함하더라도 질의어와 관련된 단어들이 문서에 많이 포함되어 있다면, 그 문서는 사용자의 요구에 적합한 문서가 된다. 정보검색기법에서도 질의어확장, 은닉의미색인, 상호정보를 이용한 검색 등과 같이 질의어 뿐만 아니라 질의어가 내포하고 있는 의미를 고려하고자 하는 연구들이 많이 진행되고 있다.

질의어 확장(query expansion)은 사용자가 제시한 질의어에 이와 관련된 단어들을 추가해서 문서를 검색함으로써 보다 연관된 문서들을 검색하고자 하는 것이다. 시소러스를 이용해서 질의어를 확장하거나 코퍼스에 나타나는 단어들의 행태에 따라 단어들의 상호관계를 분석해서 질의어를 확장하는 방법 등이 있다. 자동적인 질의어 확장 방법은 재현율은 높일 수 있으나 높은 순위의 문서들에서 정확율은 일반적으로 낮아지기 때문에 실용적이지 않다[8]. 이에대한 보완기법으로 사용자의 적합성 피드백을 이용하여 새로운 질의어를

1) 본고는 과학기술부의 지원을 받아 수행중인 '대용량 국어정보 심층처리 및 품질관리 기술개발'의 일환으로 이루어졌다.

형성하는 방법은 검색된 문서를 기반으로 해서 사용자가 직접 적합 문서와 비적합 문서를 판단해야 하고 사용자의 판단의 질에 매우 종속적이다. 최근에는 사용자가 제시한 질의어에 대해 검색된 문서들을 분석해서 질의어를 자동으로 확장하는 방법 등이 연구되고 있다[2].

은닉의미색인(*latent semantic indexing*)은 벡터공간 검색기법의 확장으로, 개념기반 검색 기법이다. 단어들 사이의 의존 관계가 문서와 질의의 표현에서 고려되고, 검색에서 이를 활용하기 때문에, 문서가 질의어에 포함된 단어를 포함하지 않더라도 관련있는 문서로 검색될 수 있다[13].

최근 연구에서는 단어들 사이의 관계를 나타내는 척도인 상호 정보(mutual information)를 이용해서 검색 효율을 향상시킬수 있음을 보이고 있다. 상호정보를 이용한 2단계 문서순위 결정 기법[1]에서는 일차로 벡터공간모델을 사용하여 질의어-문서간 유사도 값에 따라 문서 순위를 결정한다. 이차로 자동으로 구축된 상호 정보를 통해 일차 검색된 문서 순위를 재조정한다.

본 연구에서는 질의어 뿐만 아니라 문서의 문맥을 반영하여 검색을 수행하기 위한 방법으로 문서 클러스터를 이용한 재순위화 모델을 제시한다. 이 모델은 먼저, 역화일 기법으로 검색을 수행하는 SMART시스템을 이용하여 질의어를 포함하는 문서들을 검색한 후, 이 검색된 문서들에 대해 클러스터 분석을 통해서 문서들을 재평가한다. 질의어를 포함하는 문서들이 형성하는 클러스터를 통해서 질의어 뿐만 아니라 문서에 포함된 모든 단어들이 문서와의 유사도에 반영될 수 있다.

2절에서는 재순위화 모델에서 사용하고 있는 문서 클러스터의 구축에 대해 설명하고, 3절에서는 재순위화 모델에서의 정보검색 과정을 설명한다. 그리고 4절에서는 실험을 통해서 제안하는 모델의 성능을 평가한다.

2 문서 클러스터의 구축

다차원 자료에 대한 클러스터링은 많은 분야에서 필요로 한다. 클러스터링 방법에는 계층적 클러스터링과 비계층적 클러스터링이 있는데, 클러스터 분석에 관한 많은 연구에서 계층적 클러스터링 방법을 이용하고 있다.

계층적 클러스터링은 각 문서들을 각 하나의 클러스터로 하여 시작한다. 유사도가 높은 두개의 클러스터를 하나의 클러스터로 만드는 과정을 반복하여 하나의 클러스터가 남을 때까지 반복한다. 계층적 클러스터의 구조는 그림1과 같이 단순히 하나의 트리로 각 단계에서 클러스터들이 결합되는 것을 볼 수 있다.

계층적 클러스터링 방법에는 유사도가 높은 두개의 클러스터를 선택하는 기준에 따라 단순 링크 기법, 복합 링크 기법, 그룹 평균 링크 기법, Ward 기법 등이 있는데, 이들 중 Ward기법에 의해 형성된 클러스터는 균등하고 대칭적인 특성을 갖는 경향이 있다[14].

제안하는 모델에서는 문서클러스터링 방법으로 Ward 기법을 사용한다. Ward 기법[5]은 일반적인 계층적 클러스터링 기법을 따르고 있는데, 두 클러스터가 결합될 때 클러스터 중심에 대한 거리계산에서 전체 그룹내의 분산이 최소로 유지되는 두 클러스터를 각 단계에서 결합한다. 이때 클러스터간의 관련도는 비유사도 값을 기준으로 비유사도 값이 작을수록 두 클러스터의 관련도가 높은 것이 된다. 그래서, Ward기법을 최소분산기법이라고도 한다.

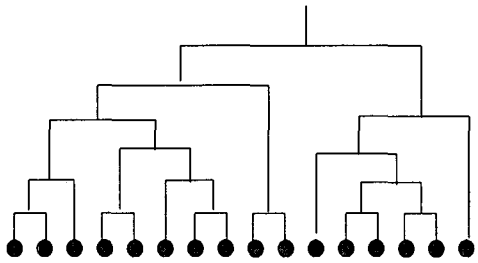

그림 1: 계층적 클러스터의 구조

본 연구에서는 문서 클러스터를 구축하기 위해서 먼저, 각 문서를 벡터형태로 표현한 후, 문서들 사이의 유사도를 계산한다. 이 유사도에 따라 문서들을 클러스터링한다.

각 문서는 단어와 그 단어가 가지는 가중치의 쌍으로 표현한다. 문서 표현을 위한 과정은 형태소 해석후 명사 추출을 하고, 단어 빈도수와 문서 빈도수를 계산하여 각 단어에 가중치를 계산한다.

단어에 가중치를 부여하기 위한 방법으로 정보검색 문헌에 이들 요소들을 이용한 단어 가중치를 계산하는 공식에 따른 검색효율에 대한 비교연구는 [6]에 있다. 단어의 가중치 계산은 질의어에 대한 단어 가중치 계산과 문서에 대한 단어 가중치 계산에 각기 어떤 것을 적용하느냐에 따라 여러 방법으로 계산을 할 수 있다. 여러 단어 가중치 계산 공식들 중 비교적 높은 검색 효율을 나타내는 문서 · 질의어의 가중치 기법으로 atc · atc기법이 있다.

$$a=0.5+0.5\cdot\frac{tf}{\max tf}\ t=\ln\frac{N}{n}\ c=\frac{1}{\sqrt{\sum w_k^2}},$$

여기서 tf는 단어빈도수, maxtf는 최대값을 갖는 단어빈도수, N은 전체 문서의 수, n은 어떤 단어를 포함하고 있는 문서의 수를 나타낸다. 클러스터링을 위한 문서의 표현에서는 이 방법에 따라 단어들의 가중치를 부여하였다. 문서-문서의 유사도는 두 벡터사이의 대응되는 색인 단어의 가중치에 의존하므로, 클러스터링에서 단어에 대한 가중치 계산 기법은 문서 클러스터링의 효율에 중요한 영향을 미치는 요소이다.

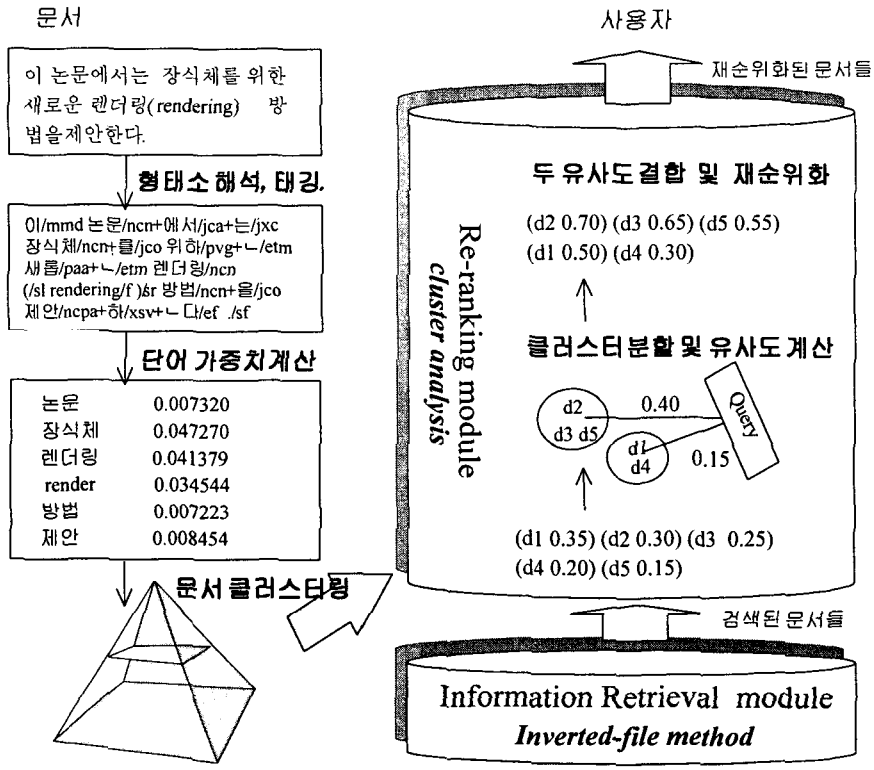

그림 2: 문서 클러스터를 이용한 재순위화 모델

문서와 클러스터 사이의 유사도 계산을 위한 방법으로는 코사인 계수(cosine coefficient)를 이용하였다.

문서 클러스터링은 N개의 문서에 대해서 2N-1개의 클러스터를 형성하고, 각 클러스터에 대한 클러스터 대표를 생성한다. 이 클러스터 중심(cluster centroid) 또는 클러스터 대표는 클러스터에 포함되어 있는 문서들의 특성을 표현하기 위해 사용되는데, 이는 단어와 단어의 가중치의 쌍으로 이루어진 벡터로 표현된다.

재순위화 과정에서 이미 구축된 클러스터 계층 구조를 이용해서 검색된 문서들에 대한 동적인 클러스터를 결정하고, 클러스터 중심과 질의어 사이의 유사도를 계산한다.

3 문서클러스터를 이용한 재순위화모델

문서 클러스터를 이용한 재순위화 모델에 대한 전체적인 구조는 그림2와 같다.

우선, 문서들 사이의 유사도에 따라 문서들에 대한 계층적 클러스터를 구축한다. 정보 검색시 역화일 기법과 클러스터 분석 기법을 결합하고 있는데, 1차는 검색단계이고, 2차는 분석단계로 이루어진다. 검색단계에서는 질의어와 문서 사이의 유사도를 계산하여 순위를 결정한다. 분석단계에서는 1차적으로 검색된 문서들이 클러스터에 분포하는 행태에 따라 클러스터를 분할한 후, 각 클러스터의 중심과 사용자 질의어와의 유사도를 계산한다. 질의어-문서의 유사도와 질의어-클러스터의 유사도의 값을 결합해서 새로운 유사도를 계산한다. 이 유사도에 따라 문서들을 다시 순서화하여 사용자에게 제시한다.

3.1 역화일 기법에 기반한 검색 단계

검색 단계에서는 질의어를 포함하는 문서들을 검색한다. 이를 위해 한국어 검색을 위해 n-그램을 사용한 SMART시스템[6]을 이용한다.

SMART시스템[4]은Havard와 Cornell 대학에서 30여년에 걸쳐 개발되어온 벡터공간모델의 정보검색 시스템이다. 이 시스템은 다양한 가중치기법을 적용할 수 있는 실험환경을 제공하고 있다.

역화일 기법에 의한 검색을 통해 질의어-문서 사이의 유사도에 따라 문서들을 내림차순으로 순서화한다. 두 번째 단계에서는 질의어-문서의 유사도가 0이상인 N개의 문서를 대상으로 하여 재순위화를 하게 된다.

본 연구에서는 질의어와 문서에 대해 다양한 가중치 기법을 적용했을 때 재순위화 모델의 검색 효율을 평가하였다.

3.2 클러스터 분석에 기반한 문서의 재순위화 단계

문서에 포함되어 있는 단어들의 부합 정도가 높을수록 두 문서는 보다 유사하다고 할 수 있다. 따라서, 문서 클러스터링을 할때도 유사도가 높은 문서들을 우선적으로 하나의 클러스터를 형성한다. 그러므로, 관련된 문서는 비관련된 문서보다 서로 유사한 경향이 있다는 *클러스터 가설*[3]에 따라 서로 관련있는 문서들은 같은 클러스터에 속하게 된다.

클러스터의 중심이 가지는 단어와 그 단어의 가중치는 클러스터에 속해있는 문서들이 가지는 단어와 단어의 가중치에 의해 결정된다. 그러므로, 하나의 클러스터에 속해있는 문서들은 클러스터 중심을 형성하는 과정에서 자신들의 특성을 반영하게 된다. 즉, 두 문서 또는 두 클러스터 C_i와 C_j가 하나의 클러스터를 형성할 때, 클러스터 중심을 계산하는 과정은 다음과 같다.

$$C_{ij} = \frac{m_i C_i + m_j C_j}{m_i + m_j},$$

여기서 m은 클러스터의 크기 즉, 클러스터에 포함되어 있는 문서의 수를 나타낸다.

클러스터 분석을 통한 재순위화 단계에서는 문서를 하나의 단위로 다루는 것이 아니라, 문서들의 그룹인 클러스터를 하나의 단위로 보고, 질의어-클러스터의 유사도를 계산해서 클러스터에 속하는 각 문서들에 동일하게 적용시킨다. 클러스터에 속해있는 문서들은 클러스터 중심을 통해서 서로 다른 문서들에 영향을 주게된다. 이때, 질의어 뿐만 아니라 문서내에 포함된 다른 단어들도 작용을 하게 되므로 문맥을 고려한 검색을 가능하게 하는 효과를 얻을 수 있다.

3.2.1 클러스터 분할

클러스터 분석 분야에서는 검색 결과를 향상시키기위해 클러스터링을 어떻게 이용할 것인가에 관한 많은 연구가 있어 왔다. 대부분의 기존 연구에서는 전체 문서집합에 대해서 정적 클러스터링(static clustering)을 하고 나서, 질의어와 클러스터의 부합정도를 계산했다[11]. 대조적으로, Scatter/Gather 시스템[9]은 클러스터 기반 문서 브라우징기법으로, 동적 클러스터링(dynamic clustering) 기법을 적용하였다. 이는 질의어에 대해 검색된 문서들을 대상으로 클러스터링을 수행한다.

정적 클러스터링과 동적 클러스터링의 장단점은 [12]에 의하면 다음과 같다. 실행시 동적 클러스터링을 하면 분할이 질의어에 대해 보다 정교하게 되고, 적은 수의 클러스터로도 전체 결과 집합을 포함하는 것이 보장된다. 정적 클러스터링은 결과집합이 많은 클러스터로 분할될 수도 있고, 이들 클러스터 중 일부는 결과 집합에서 하나 또는 둘 정도의 적은 수로 구성될수도 있다. 반면에, 미리 클러스터링을 하면, 문서집합의 내용을 고려하면서 질의에 따라 분할할 수 있고 다양하고 유용한 범주를 만들수 있는 보다 견고한 클러스터링 알고리즘을 채택할 수 있는 기회를 갖게 된다.

본 연구에서 제안하는 모델은 *정적 클러스터링 기법에 동적 뷰(dynamic view)를 이용한다.* 즉, 먼저 전체 문서집합에 대해 정적 클러스터링을 수행하고, 검색된 문서들에 따라 재순위화 단계에서 클러스터를 동적으로 분할한다.

계층 클러스터링은 N개의 문서에 대해 2N-1개의 클러스터와 클러스터 중심을 생성하는데, 이러한 클러스터들 중에서 현재의 질의어에 대해 어떠한 클러스터들을 선택할 것인지를 결정해야 한다. 이를 위해서, 첫단계에서 검색된 문서들이 클러스터에 어떻게 분포하고 있는지를 분석하여 질의어의 관점에서 클러스터를 동적으로 분할한다.

그림3에 나타난 것처럼, 각 클러스터에 대해서

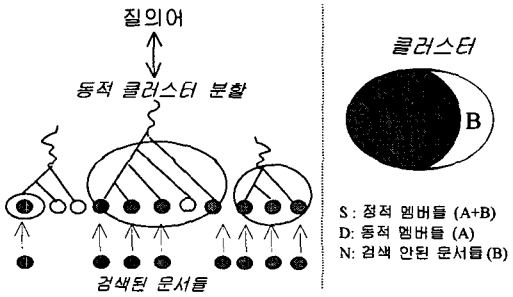

그림 3: 동적 클러스터 분할

정적 클러스터링에서의 정적 멤버들(S)과 질의어에 대해 검색된 문서들의 분포에 따른 동적 멤버들(D)의 비율에 따라 클러스터의 계층구조 상에서 상위 클러스터를 택할 것인지 하위 클러스터를 택할 것인지를 결정한다. 이 기준은 임계치(threshold)에 따라 선택된다. 이 단계에서 선택되는 클러스터들은 질의어에 대해 검색된 문서 집합이 달라진다면 결과 클러스터들도 달라진다. 따라서, 질의어의 특성에 맞는 클러스터들을 생성할 수 있다.

3.2.2 질의어-클러스터 유사도 계산

클러스터의 중심은 단어와 그 단어의 가중치 쌍으로 이루어진 벡터로 표현되는데, 검색 집합 A_a에 대한 클러스터 중심 $C(A_a)$은 다음과 같이 정의된다.

$$C(A_a) = \frac{\sum_{a \in A_a} \alpha}{|A_a|},$$

여기서 $|A_a|$는 A_a의 크기, α는 문서a에 대한 벡터를 나타낸다. 클러스터 중심을 형성하는 과정을 통해서 문서들은 서로 이웃한 문서들에게 영향을

미치게 된다.

클러스터 분할의 결과는 첫단계에서 검색된 문서가 아닌 것들도 포함할 수 있다. 이러한 문서들은 질의어에 대한 클러스터 중심에 부정적인 효과를 줄 수 있으로, 이 단계에서는 검색 되지 않은 문서의 부정적인 효과를 최소화시키기 위해서 클러스터 중심의 값을 조정한다.

질의어에 포함된 각 단어 w_i에 대해서 다음과 같이 클러스터 중심이 갖는 값을 다음과 같이 조정한다.

$$C_{w_i}' = C_{w_i} \cdot (1 + \frac{S-D}{S}),$$

여기서 S는 클러스터의 정적 멤버의 수, D는 클러스터의 동적인 멤버의 수를 나타낸다. 그리고, C_{wi}는 현재 클러스터 중심의 가중치, C_{wi}'는 조정된 클러스터 중심의 가중치를 나타낸다.

새롭게 조정된 클러스터 중심을 이용하여 질의어-클러스터 유사도를 계산한다. 계산된 클러스터 유사도는 같은 클러스터에 속하는 모든 문서들에 대해서 동일한 효과를 가진다. 즉, 같은 클러스터에 속한 문서들은 첫단계 질의어-문서의 유사도는 서로 다른 값을 갖더라도 두번째 단계 질의어-클러스터의 유사도는 서로 동일한 값을 갖게 된다.

사용자의 요구와 관련이 있는 클러스터는 높은 유사도를 갖고, 사용자의 요구와 거리가 먼 클러스터는 낮은 유사도를 가진다. 그러므로, 사용자의 요구와 관련된 클러스터에 높은 우선권을 주는 실마리를 갖게 된다.

3.2.3 두 유사도의 결합

이제까지 역화일 기법을 이용하여 질의어-문서 유사도를 계산하고, 클러스터 분석을 통해서 질의어-클러스터 유사도를 계산하였다. 검색 단계에서는 각 문서들에 중점을 두었고, 분석 단계에서는 문서들의 집합에 중점을 두었다.

여기서는 검색단계와 분석단계로부터 계산된 두가지 유사도 값을 다음과 같이 결합한다.

$$Sim_{combined} = \alpha \cdot Sim_{inverted} + \beta \cdot Sim_{cluster}$$

, 여기서 α, β는 각 단계에서 계산된 유사도 값에 중요도를 부여하는 의미와 이용된 가중치 기법의 차이에 따른 유사도 값의 차이를 조정하는 의미를 가진다. 이렇게 결합된 유사도에 따라 문서들을 재순위화 한 후, 사용자에게 결과를 제시한다.

어떤 문서가 검색 단계에서 문서에 대한 유사도는 낮은 값을 가졌더라도 분석 단계에서 클러스터에 대한 유사도가 높은 값을 가질 경우에 사용자에게 높은 우선순위로 제시될 수 있다. 반대로, 검색 단계의 문서 유사도는 높은 값을 가졌더라도 분석 단계의 클러스터 유사도가 낮은 값을 가진다면 사용자에게는 낮은 우선순위로 제시될 수도 있다.

문서클러스터를 이용한 재순위화 모델에서는 질의어를 포함하는 정도에 따라 계산된 각 문서의 유사도에 질의어 관점에 따라 동적으로 분할된 클러스터의 유사도를 결합함으로써 문서의 순위를 재결정하였다.

4. 실험 및 평가

제안된 모델의 성능을 평가하기 위해 ETRI-KEMONG 실험 집합을 이용하였다. 표1은 실험 집합에 대한 통계 정보이다.

문서 수	23113
질의어 수	45
평균 문서 길이	56 words
평균 질의어 길이	3 words
평균 적합문서 수	9

표 1: 실험집합의 통계 정보

다음은 비교 실험을 위한 두 검색 기법에 대한 설명이다.

- SMART 시스템 : 한국어 검색을 위한 n-그램 방식의 SMART시스템은 역화일 기법의 검색 시스템으로, 검색한 문서를 질의어-문서의 유사도에 따라 내림차순으로 순위를 부여한다.

- 문서클러스터를 이용한 재순위화 모델: 이 모델은 역화일 기법을 이용하여 질의어를 포함하는 문서를 검색하고, 클러스터 분석을 통해서 문서들의 문맥을 반영하여 문서들의 순위를 재평가한다.

제안모델의 성능 효율을 실험하기 위해서 먼저, SMART 시스템의 실험 환경을 bi-그램으로 설정하였다. 문서와 질의어에 대한 여러 가지 가중치 계산방법들 중에서 [6,7]의 연구결과에서 검색 효율이 비교적 높은 lnc · ltc, atn · ntc, ltn · ntc, atc · atc 4개의 문서 · 질의어 가중치 기법과 단순히 단어빈도수만을 고려한 nnn · nnn 가중치 기법 각각에 대해서 SMART시스템의 성능과 재순위화 모델에서의 성능 차이를 비교 실험하였다.

재순위화 단계에서 클러스터를 동적으로 분할할 때 사용하는 임계치의 변화에 따라서도 성능의 차이를 비교하였고, 또한 역화일 기법에 의한 유사도와 클러스터 분석에서의 유사도를 결합할 때 두 유사도의 결합 중요도를 다르게 했을 때의 성능의 차이를 비교하였다.

표2는 1차 검색에서 사용된 SMART 시스템의 각 가중치 기법에 대해 제안한 모델의 성능 향상률을 재현율(recall)과 정확률(precision)에서 상세하게 보여주고 있다. 11pt avg는 11포인터 재현율에서의 질의어 집합의 평균 정확률을 적용함으로써 평가되었고, change%는 SMART시스템의

precision / recall	smart nnn-nnn	rerank1 (0.8,2:3)	smart atc-atc	rerank2 (0.8,1:1)	smart lnc-ltc	rerank3 (0.5,5:1)	smart atn-ntc	rerank4 (0.9,2:1)	smart ltn-ntc	rerank5 (0.5,1:1)
0.0	0.6397	0.9110	0.6004	0.6775	0.6142	0.6962	0.7776	0.8422	0.8097	0.9107
0.1	0.6209	0.8999	0.5942	0.6775	0.6142	0.6907	0.7631	0.8390	0.7937	0.8996
0.2	0.5559	0.8663	0.5767	0.6498	0.5956	0.6647	0.7389	0.8174	0.7626	0.8591
0.3	0.5039	0.8225	0.5446	0.6387	0.5860	0.6470	0.6896	0.7694	0.7284	0.8334
0.4	0.3774	0.7152	0.4836	0.5970	0.5103	0.5905	0.6112	0.7125	0.6508	0.7350
0.5	0.3638	0.6983	0.4731	0.5885	0.4761	0.5700	0.6102	0.6999	0.6278	0.7136
0.6	0.3112	0.6017	0.4228	0.4753	0.3986	0.4759	0.5579	0.5955	0.5515	0.6331
0.7	0.2815	0.5305	0.3289	0.4160	0.3183	0.4229	0.4984	0.5215	0.5029	0.5683
0.8	0.2632	0.4634	0.2919	0.3648	0.2795	0.3844	0.4477	0.4450	0.4522	0.4879
0.9	0.2524	0.4351	0.2542	0.3320	0.2364	0.3535	0.4155	0.4261	0.4166	0.4715
1.0	0.2383	0.4199	0.2478	0.3195	0.2228	0.3389	0.4024	0.4114	0.4011	0.4535
11pt avg	0.4008	0.6694	0.4380	0.5215	0.4411	0.5304	0.5920	0.6436	0.6088	0.6878
change %		67.16%		19.06%		20.24%		8.72%		12.98%

표 2: SMART시스템과 재순위화 모델의 비교 실험

각 가중치 기법에 의한 검색 결과에 대한 재순위화 모델에서의 성능 향상률을 나타낸다. 표2의 smart atc · atc와 rarank2(0.8 1:1)에서 atc · atc는 SMART시스템의 가중치 기법을 뜻하고, rarank2는 SMART시스템의 atc · atc 가중치기법으로 1차 검색후 클러스터 분석을 거친 재순위화 실험 결과를 나타낸다. 괄호안의 숫자에서 0.8은 임계치로, 계층적 클러스터를 분할할때 정적 클러스터에 대한 동적 클러스터의 비율이 0.8이상이면서 최소값을 갖는 클러스터를 선택한 것이다. 그리고, 1:1은 SMART 시스템의 문서 유사도와 클러스터 유사도의 결합 비중을 각각 1:1로 계산하여 결합한 것을 의미한다. 결합에서의 비중의 차이는 가중치 계산 기법이 달라지면 계산된 유사도의 분포가 영향을 받기 때문에 이러한 현상을 최소화하기 위한 것이다.

제안한 기법에서는 SMART시스템의 nnn · nnn가중치기법에 대해서는 67.16%, atc · atc 가중치기법에 대해서는 19.06%, lnc · ltc 가중치 기법에 대해서는 20.24%, atn · ntc 가중치 기법에 대해서는 8.72%, ltn · ntc 가중치 기법에 대해서는 12.98% 의 향상을 보이고 있다.

문서클러스터를 이용한 재순위화 모델은 클러스터 분석을 통해서 암시적으로 문서들의 문맥을 반영함으로써 전체 검색에서 성능을 향상시킬 수 있었다.

5. 결 론

본 연구에서는 정보검색시스템의 모델로 문서 클러스터를 이용한 재순위화 기법을 제안하였다. 이 기법에서는 처음 검색된 문서들의 행태에 따라서 클러스터를 동적으로 분할함으로써 준동적(semi-dynamic) 형태의 클러스터를 이용한다. 따라서, 제안된 모델에서는 질의어의 특성에 맞는 클러스터를 생성할 수 있다. 또한, 사용자의 요구에 대해 문서의 검색과 클러스터 분석을 결합함으로써 문서의 문맥이 반영될 수 있었다. 이때 클러스터 문맥은 질의어를 정교하게 하는 논리적인 촛점의 집합으로서의 역할을 하였다.

실험을 통해서 역화일 기법에 의한 질의-문서 유사도에 의한 검색에서 보다 문서 클러스터를 이용한 재순위화 모델이 아주 우수한 성능 향상을 나타냄을 보였다.

향후 연구로, 제안된 모델에서는 재순위화 단계에 사용자 프로파일 관리 시스템을 추가함으로써, 쉽게 사용자 특성을 반영한 검색이 가능하도록 확장할 수 있다. 동일한 질의어에 대해서 검색된 문서의 적합성 여부는 사용자의 관심에 따라 달라질 수 있는데, 문서 클러스터는 사용자 프로파일과 비교하기에 적합한 구조이다. 재순위화 단계에서 사용자 프로파일과 클러스터 중심을 비교함으로써 사용자의 관심이 반영될 수 있다. 사용자 프로파일이 질의어의 확장 형태로 볼 때, 사용자 프로파일을 사용함으로써 보다 더 성능을 향상시킬 수 있음을 기대할 수 있을 것이다.

참고문헌

[1] 강현규. 1997. 자연언어 정보검색에서 상호정보를 이용한 2단계 문서 순위 결정 방법. 한국과학기술원 박사학위논문.

[2] C. Buckley and G. Salton and J. Allan. 1994. The effect of adding relevance information in a relevance feedback environment. In *Proc. 17'th ACM SIGIR International Conference on Research and Development in Information Retrieval*, pages 292-298.

[3] C. J. van Rijsbergen. 1979. *Information Retrieval.* Butterworths, 2nd edition.

[4] Gerard Salton. 1989. *Automatic Text Processing: The transformation, analysis, and retrieval of information by computer.* Addison-Wesley.

[5] J. H. Ward. 1963. Hierarchical Grouping to Optimize an Objective function. *Journal of the American Statistical Association*, 58(301):235-244.

[6] Joon Ho Lee and Jeong Soo Ahn. 1996. Using n-Grams for Korean Text Retrieval. In *Proc. 19'th ACM SIGIR International Conference on Research and Development in Information Retrieval*, pages 216-224.

[7] Joon Ho Lee. 1995. Combining Multiple Evidence from Different Properties of Weighting Schemes. In Proc. 18'th ACM SIGIR International Conference on Research and Development in Information Retrieval. pages 180-188.

[8] Larry Fitzpatrick and Mei Dent. 1997. Automatic Feedback Using Past Queries: Social Searching? In *Proc. 20'th ACM SIGIR International Conference on Research and Development in Information Retrieval*, pages 306-313.

[9] Marti A. Hearst and Jan O. Pedersen, Reexamining the Cluster Hypothesis: Scatter/Gather on Retrieval Results. In Proc. 19'th ACM SIGIR International Conference on Research and Development in Information Retrieval, pages 76-84.

[10] Michael L. Mauldin and Jaime G. Carbonell. 1991. *Conceptual Information Retrieval: A Case Study in Adaptive Partial Parsing*, Kluwer Academic Publishers.

[11] P. Willett. 1988. Recent Trends in Hierarchical Document Clustering: a Critical Review. *Information Processing and Management*, 24(5):577-597.

[12] Peter G. Anick and Shivakumar Vaithyanathan. 1997. Exploiting Clustering and Phrases for Context-Based Information Retrieval. In *Proc. 20'th ACM SIGIR International Conference on Research and Development in Information Retrieval*, pages 314-323.

[13] Scott Deerwester and Susan T. Dumais and Richard Harshman. 1990. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41(6):391-407.

[14] William B. Frakes and Ricardo Baeza-Yates. 1992. *Information Retrieval : Data Structures & Algorithms*. Prentice Hall, pages 435-436.