계층구조를 이용한 문서 클러스터 제목의 자동생성

김태현[0] 맹성현

충 남 대 학 교 컴 퓨 터 과 학 과
{heemang, shmyaeng}@cs.cnu.ac.kr

Automatic Naming of Document Clusters by Using their Hierarchical Structure

Tae-Hyun Kim[0] Sung Hyon Myaeng

Dept. of Computer Science, Chungnam National University

요 약

웹에서 정보를 찾고자 하는 사용자들을 돕기 위해서는 조직화된 방법으로 검색 결과들을 제시하는 것이 바람직하다. 이러한 목적을 위해, 문서 클러스터링 기법들이 제안되었다. 문서 클러스터링은 사용자들이 관심의 대상이 되는 문서들을 더욱 쉽게 배치할 수 있게 하고, 검색된 문서 집합에 대한 개관을 손쉽게 얻을 수 있게 한다. 클러스터링 결과로 주어지는 각 클러스터의 주제를 사용자들이 빠르게 파악할 수 있게 하려면 클러스터 제목을 표현하는 문제가 중요시 된다. 본 연구에서는, 웹 디렉토리의 계층적 구조를 사용하여 자동으로 클러스터 제목을 생성하는 방법을 제안한다. 이 방법은 대상이 되는 클러스터에 있는 문서들의 내용과 부합되는 계층상의 노드를 계층구조 상에서 찾아내어, 계층구조의 루트로부터 그 노드에 이르는 경로명을 클러스터의 제목으로 사용자에게 제시하도록 한다. 본 연구에서 제안한 모델은 '야후' 디렉토리를 사용하여 실험되었다. 실험 결과, 실험대상 클러스터의 본래 제목과 정확하게 일치하는 제목을 찾을 수 있는 경우의 정확률이 57.5%, 의미적으로 본래 제목에 부합되는 제목을 찾을 수 있는 경우의 정확률이 대략 90%에 이른다는 것을 알 수 있었다.

1. 서론

현재 정보검색 서비스들은 주로 대량의 웹 문서들 중에서 사용자 요구에 가장 적합한 문서들을 가능한 빠른 시간 내에 찾아주는 것을 목적으로 검색 서비스를 제공하고 있다. 사용자들에게 주어지는 정보검색 결과는 대부분의 경우 단순히 단어빈도에 기초한 적합도에 따라 나열식으로 제시된다. 이 경우, 사용자가 생각하는 적합도와는 다른 순서로 그 결과가 주어지는 경우도 있고, 전혀 다른 의미를 갖는 문서들이 결과집합에 포함되는 경우도 있다. 따라서, 사용자들은 자신이 진정으로 원하는 문서를 찾기 위해, 이러한 검색결과 리스트를 순차적으로 훑어 보아야 한다는 문제를 안고 있다. 그러므로, 정보검색에서는 사용자에게 질의에 적합하다고 판단되는 문서들을 빠르게 찾아주는 방법 이외에 사용자들이 검색결과로 나온 문서 집합들의 의미를 빠르게 파악할 수 있게 하는 방법에 대한 연구가 이루어져야 한다.

검색 결과들을 사용자들이 파악하기 쉬운 형태로 제공하기 위한 노력으로 다양한 연구가 이루어지고 있다. 디렉토리 형태로 분류된 문서집합을 대상으로 검색을 수행하도록 하여 그 결과를 제시함으로써 사용자들이 검색되어져 나온 문서가 어떠한 분류에 속하는지 쉽게 파악할 수 있도록 하는 방법이 그 중 하나이고, 또 다른 하나로 문서 클러스터링(Document clustering)을 이용하여 결과문서집합을 작은 그룹으로 나누어 제시하는 방법이 있다. 전자는 검색의 대상이 제한적이며, 검색대상이 되는 문서들을 분류하는데 많은 노력이 들뿐만 아니라, 검색되어 나온 결과를 제시하는 방법이 기존의 나열식 검색결과 제시 방법과 크게 다르

지 않다는 단점을 갖는다. 후자의 경우는 사용자가 검색결과를 하나의 큰 집합이 아닌 서로 다른 특성을 가지는 여러 개의 작은 집합으로 보면서 접근할 수 있다는 점에서 사용자가 전체적인 검색결과를 쉽게 파악할 수 있다는 장점을 제공하고 있다. 그러나 이 역시 클러스터링 결과로 만들어진 각 클러스터(Cluster)가 어떠한 주제 하에 묶였는지를 적절히 표현하지 못할 경우 사용자가 각 클러스터를 이해하는데 많은 노력이 필요하게 된다는 문제점을 안고 있다.

본 연구에서는 위의 두 가지 방법의 단점을 최소화하고 장점을 살릴 수 있는 검색결과제시 방법을 제안하고자 한다. 즉 정보검색의 결과로 만들어진 문서집합을 클러스터링 기법을 이용해 유사한 문서들의 집합으로 그룹화하고, 그 결과 만들어진 각 클러스터에 대해 사용자가 클러스터의 내용을 쉽게 파악할 수 있게 하는 적절한 제목을 붙여주기 위해 디렉토리 정보를 이용하는 것이다.

웹 문서를 각 문서가 갖는 특징적인 의미에 따라 수작업으로 분류한 기존의 디렉토리 정보는 자동적으로 분류한 다른 어떤 문서집합보다 의미적으로 유용한 정보집합이다. 따라서, 이러한 디렉토리 정보의 계층적인 특성을 클러스터 제목을 생성하는데 이용하자는 것이 본 논문의 기본 아이디어이다. 이를 위해 본 연구에서는 웹 상에 이미 존재하는 디렉토리 구조를 이용하여 웹 문서들의 계층적인 분류 정보를 구성하였다. 즉, 계층구조의 각 노드에 소속된 단어집합의 통계적인 정보와 문서 클러스터링의 결과로 수집된 각 클러스터의 단어통계정보를 비교하여 이들간의 유사도를 측정, 가장 적합한 계층구조상의 위치를 찾아내고, 이러한 계층정보를 문서 클러스터의 제목으로 활용하도록 하였다.

2. 관련 연구

2.1 문서 클러스터링 기법

문서 클러스터링이란, 특정 문서집합 내에 있는 각 문서들간의 유사도를 측정하여 유사한 문서들을 그룹지어 주는 것을 말한다. 문서들간의 유사도는 각 문서가 갖는 특징들을 비교하여 계산하게 되는데 일반적으로 문서에 포함되어 있는 단어의 빈도수를 그 특징으로 사용한다. 문서 클러스터링 기법은 클러스터를 구성해나가는 방법에 따라 계층적 클러스터링과 비계층적 클러스터링으로 나누어 볼 수 있다.[3,4,5]

계층적 클러스터링은 비계층적인 방법에 비해 비교적 클러스터링 시간이 느리지만 보다 정확한 클러스터 링이 수행된다는 장점을 갖는다. 또한 계층적 클러스터링은 이진트리(binary tree)와 유사한 형태의 클러스터 구조를 생성해내는 방법을 이용하므로, 클러스터 결과에 대한 서로 다른 레벨의 해상도(resolution)를 제공할 수 있다는 장점을 갖는다.[2,3,5]

비계층적 클러스터링은 임의로 선택된 초기 클러스터로부터 문서를 클러스터에 재배치하는 작업을 반복적으로 수행하여 최종 클러스터를 형성하는 방법으로, 계층적 클러스터링에 비해 클러스터링 시간은 빠르지만 검색효율이 떨어지고 문서의 입력 순서에 따라 클러스터링 결과가 달라진다는 단점을 갖는다.[3]

비계층적 클러스터링 기법의 일종인 어미 트리 클러스터링(STC)은 클러스터링의 대상이 되는 문서들에 포함되어 있는 구(phrase)들을 이용하여 공통 구(common phrase)를 포함하고 있는 문서들을 동일 클러스터에 할당한다. STC에서는 필요한 클러스터의 수를 명시할 필요가 없고, 대신 기초 클러스터(base cluster)간의 유사도를 결정하기 위한 임계점(threshold)을 명시해 주어야 한다. 그러나 계층적 클러스터링 알고리즘의 경우와 달리 클러스터링 결과가 이 임계점에 의해 크게 좌우되지는 않는다는 장점을 갖는다. 또한 STC는 시간 복잡도가 대상 문서집합의 크기에 비례하는 선형시간 알고리즘이다.[2,3]

2.2 클러스터 제목 생성

클러스터의 제목을 자동으로 생성하기 위해, 일반적으로 클러스터 내에 포함된 문서들이 공통적으로 갖는 정보들을 이용한다. 본 절에서는 이러한 정보를 이용한 기존 연구들 중 대표적인 두 가지 방법에 대해 살펴보기로 한다.

2.2.1 클러스터 내 문서들에서의 용어 가중치 이용 방법

클러스터를 대표하는 제목을 추출하기 위해, 클러스터에 포함되어 있는 문서들에 나타난 용어들의 가중치를 이용하는 방안으로, 용어 가중치는 정보검색에서 일반적으로 사용되는 용어빈도수(tf: term frequency)와 역문서빈도수(idf: inverse document frequency)를 이용하여 다음의 식과 같이 계산된다.

$$w_i = tf_i \times [\log(N/n) + 1]$$

위 식에서 tf_i는 문서 D에서 색인어 t_i의 출현 횟수이고, $\log(N/n)$은 클러스터 내에서 색인어 t_i가 나타나는 문서 개수의 역이다. 여기에서 tf는 한 문서에서 용어의 중요성을 나타내고, idf는 전 체 문서들에서 용어의 분별력을 나타낸다. 클러스 터 내의 용어들을 위 식을 이용해 구한 가중치에 따라 내림차순으로 정렬하고 정렬된 키워드 중에서 상위에 존재하는 키워드를 선택함으로써 클러스터에 대한 제목을 만든다. 그러나, 이러한 방법으로 선택되는 클러스터 제목은 단순히 키워드들의 나열 형태로 나타나므로 클러스터의

의미를 총체적으로 나타내는 추상화 된 제목을 표현해 줄 수 없다는 단점을 갖는다.[5]

2.2.2 클러스터 내 문서들의 공유 구 이용 방법

STC를 이용한 클러스터링의 경우, 구(phrase)를 이용해 클러스터를 만들기 때문에 이때 이용되었던 구를 클러스터의 제목으로 사용할 수 있다. 이는 용어 가중치를 이용한 단순한 방법에 비해 클러스터가 갖는 의미를 보다 효과적으로 표현해 줄 수 있다. 그러나, 클러스터를 만드는 데 사용된 모든 구를 클러스터의 제목으로 사용할 경우 불필요한 구들이 포함되어 오히려 사용자의 클러스터 이해를 저해하는 요인이 될 수 있다. 따라서, 클러스터에서 중요한 몇 개의 구들만을 선택하여 제목으로 사용하는 것이 바람직하다.

클러스터 내 문서들이 공유하는 구들 중에서 중요한 구들을 선택하기 위해서 주로 경험적 규칙(heuristic)이 이용된다. STC를 이용하여 생성한 클러스터에 대해 공유 구를 제목으로 붙여주는 실험적인 시스템인 Grouper에서 사용된 경험적 규칙들로는 단어 중복 규칙, 부분 또는 포함 구 관계 규칙, 낮은 적용범위를 갖는 가장 일반적인 구 제외 규칙과 같은 것들이 있다.

Grouper에서는 경험적 규칙에 의해 만들어진 구들을 클러스터의 제목으로 선택하고, 이 구들이 문서 내에 나타나는 빈도수에 따라 클러스터 내에 있는 문서들을 순서화(ordering)하여 사용자에게 제시한다. 즉, 사용자는 클러스터 내 문서들이 공유하는 구들에 의해 클러스터의 내용을 파악할 수 있을 뿐만 아니라 클러스터 내 문서들 중 어떤 문서가 클러스터 제목과 가장 관련이 깊은 것인지를 알 수 있다. 그러나, 이 방법도 클러스터 제목을 생성하는데 사용되는 구가 클러스터 내의 문서에서 발생하는 구에 한정되어 있어 추상화된 제목을 제시하지 못한다는 단점을 갖는다.[3]

3. 클러스터 제목 자동 생성

본 연구는 크게 세 단계로 나뉜다. 그 첫째는 클러스터 제목 생성에 사용되는 디렉토리 정보를 구성하기 위한 디렉토리 구조정보 구성 단계이고, 둘째는 주어진 문서집합을 클러스터링하는 문서 클러스터링 및 클 러스터 정보표현 단계이며, 마지막 단계는 이전의 두 단계에서 형성된 정보들을 바탕으로 클러스터의 제목을 자동으로 생성하는 클러스터 제목 생성 단계이다. 본 장에서는 각 단계에 대해 상세히 설명하기로 한다.

3.1 디렉토리 구조정보 구성

3.3.1 디렉토리 구조

본 연구에서는 문서 클러스터의 제목을 붙여주기 위해서 사용하는 정보구조로 '한글 야후 디렉토리 서비스' 사이트에 있는 디렉토리 구조를 이용하였다. 이는 일반적인 컴퓨터 파일 시스템에서 사용하는 디렉토리 구조와 유사하게 계층을 이루고 있고, 디렉토리 마다 각기 이름이 주어진다. 그러나, 다른 점은 디렉토리의 상위와 하위 계층간의 관계가 단순한 위치적 포함관계가 아닌 개념적인 포함관계로 이루어져 있다는 것이다. 또한 각 디렉토리 마다 해당 디렉토리를 대표하는 정보를 포함하는 HTML 파일이 존재한다. 각 디렉토리에 주어지는 HTML 파일에는 해당 디렉토리의 이름과 루트 디렉토리로부터 현재 디렉토리까지에 이르는 경로, 현재 디렉토리와 관련이 있는 웹 사이트로의 링크들과 그에 대한 간략한 설명, 하위 개념을 갖는 디렉토리로의 링크 등에 대한 정보가 포함되어 있다.[1]

Part		Number
Health		506
Education		156
News_and_Media		177
Recreation		782
Business_and_Economy		2950
Social_Science		248
Society_and_Culture		641
Entertainment		957
Arts		567
Science		585
Government		255
Regional	Countries	843
	South_Korea	11364
	Regions	59
Reference		61
Computers_and_Internet		584
total		20229

[표 1] 야후 디렉토리 구조의 하위 디렉토리 수

'야후 디렉토리 서비스' 사이트에서 얻은 디렉토리 계층 구조는 14개의 최상위 디렉토리를 시작 으로 하여 개념적인 구조가 점차적으로 세분화되면서 계층이 깊어진다. 이러한 디렉토리 구조는 전체적으로 트리 형태로 볼 수 있다. 이 계층에서 트리의 단말노드(terminating node)에 해당하는 디 렉토리 수는 13323개이고, 이들의 평균 깊이는 6.46이고, 최대 깊이는 12이다. 또한 트리의 중간노드(intermediate node)에 해당되는 디렉토리의 수는 7415개 이고, 이들이 갖는 하위 디렉토리 개수의 평균 값은 2.78이고, 최대 값은 163이다. 이러한 디렉토리 구조의 최상위 14개 디렉토리에 주어지는 하위 디렉토리의 총 개수는 [표 1]과 같다.

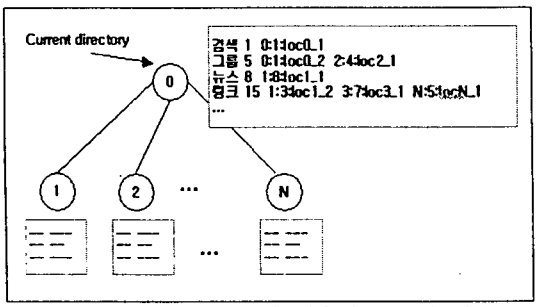
3.3.2 디렉토리 계층정보 구성

디렉토리 구조는 계층적인 정보를 가지고 있기 때문에, 이러한 정보를 문서 클러스터의 제목을 붙여

주는데 이용하기 위해서는 계층구조 내의 각 디렉토리가 가지는 정보를 적절히 표현하여야 한다. 이를 위해 본 연구에서는 각 디렉토리가 가지는 이름을 이용하여 계층정보를 구성하였다.

연구의 초기 단계에서는 각 디렉토리의 이름뿐만 아니라 이들이 가지는 HTML 파일에 있는 정보들까지도 이용하여 계층정보를 구성하고 이를 이용하고자 하였다. 그러나, 이러한 경우 개념적인 계층구조를 저해하는 정보요소가 증가하여 이후에 계층정보와 클러스터 정보를 비교하여 클러스터 제목을 생성하고자 하는데 있어 좋지 않은 영향을 미치게 되었다. 따라서, 순수 디렉토리 계층정보만을 구성하고자 하는 목적으로 각 디렉토리에 주어지는 이름들만을 이용한 것이다.

디렉토리 계층정보는 상향식 기법으로 구성된다. 즉, 디렉토리 계층구조 상에서 각 디렉토리는 자신의 하위 디렉토리가 가지는 모든 색인(index) 정보와 자신의 디렉토리 이름에 대한 정보를 통합한 색인 정보를 갖는다. 이는 다음의 [그림 1]과 같이 각 디렉토리에 대해 만들어진다.



[그림 1] 디렉토리 계층정보 구성 예

이러한 색인 정보는 용어를 단위로 하여 색인 파일에 기록되며, 각 용어에 대해 주어지는 정보는 [그림 1]에서 나타난 바와 같이 현재 디렉토리 를 기준으로 하여 그 이하에 해당 용어가 나타난 총 회수와 이 용어가 현재 디렉토리의 몇 번째 서브 디렉토리에서, 몇 번 나타나고, 서브 디렉토리 색인 파일의 어느 곳에 나타나는지에 대한 정보를 포함한다. 이 때 서브 디렉토리 색인 파일의 어느 곳에 해당 용어가 나타나는지에 대한 정보는 실제로 서브 디렉토리 색인 파일 내에서 해당 용어가 나타나는 물리적인 변위(offset)를 기록한다. 이는 최상위 디렉토리의 색인 파일에서 필요한 용어정보를 찾은 이후에는, 그 하위에 있는 디렉토리들에 대해서는 색인 파일을 순차적으로 검색하여 해당 용어를 찾지 않고 변위를 이용해 직접 찾을 수 있도록 하기 위해 사용되는 정보이다.

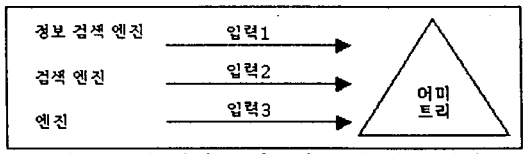
3.2 문서 클러스터링 및 클러스터 정보표현

3.2.1 웹 검색결과에 대한 클러스터링

웹 검색결과에 대한 클러스터링 방법으로 앞선 관련연구에서 설명한 어미 트리 클러스터링을 이용하였다. 문서 클러스터를 생성하기 위한 과정으로 '검색결과 문서 수집 및 처리', '어미 트리 구성', '클러스터 생성'의 세 단계를 거치게 된다.[3]

'검색결과 문서 수집 및 처리' 단계는 사용자가 검색엔진에 질의를 던져 그 결과로 받은 웹 문서로부터 실질적인 검색결과 문서들을 추출하는 단계이다. 이 단계에서 웹 문서에 대한 파싱(parsing)이 이루어져 각 검색결과 문서에 대한 제목, URL, 요약정보가 추출된다. 추출된 정보는 색인기(indexer)에 의해 재처리되어 다음 단계인 '어미 트리 구성'에서 이용할 있는 명사 리스트의 형태로 변환된다.

'어미 트리 구성' 단계에서는 앞 단계에서 만들어진 검색결과 문서 정보에 대한 명사 리스트를 어미 트리의 입력으로 주어, 검색결과로 얻은 모든 문서의 모든 어미를 갖는 어미 트리를 구성한다. 다음의 [그림 2]는 검색결과 문서를 대표하는 명사 리스트가 '정보 검색 엔진'이라 할 때, 이에 대해 어미 트리로 주어지는 입력을 나타낸다.



[그림 2] 어미 트리로의 문서 정보 입력

'클러스터 생성' 단계에서는 최종적으로 만들어진 어미 트리의 중간노드(intermediate node)들에 주 어지는 문서집합을 각각 기초 클러스터(base cluster)로 간주하고, 이를 대상으로 클러스터 병합 알고리즘을 적용하여 최종적인 클러스터들을 얻는다.

3.2.2 클러스터 정보표현

어미 트리 클러스터링 기법을 이용하여 최종적으로 만들어진 문서 클러스터에 대한 제목을 붙여주기 위해서는 각 클러스터를 대표할 수 있는 정보를 구성해야 한다. 이를 위해 클러스터가 포함하고 있는 다음과 같은 기본 정보들을 수집한다.

N : 클러스터 내의 용어 개수
D : 클러스터 내 문서의 총 개수
tf_i : i번째 용어의 클러스터 내 출현빈도
df_i : 클러스터 내에서 i번째 용어가 나타나는 문서의 수

수집된 기본 정보를 이용하여 하나의 문서 클러스터 내에 포함되어 있는 각 용어들의 중요도를 계산하고, 이를 클러스터를 대표하는 정보로 이용한다. 용어 중요도를 계산하기 위한 수식은 다음과 같다. 이는 '클러스터 내에서 출현빈도가 높으면서, 클러스터 내의

문서들에 고르게 나타나는 용어가 더 중요한 용어이다.'라는 가정을 기초로 구성한 수식이다.

$$w_i = \frac{tf_i}{\max tf_i} \times \frac{df_i}{D} : \text{i번째 용어의 중요도}$$

위의 수식을 이용하여 클러스터 내 용어의 중요도를 모두 계산한 후에는 이들 중에서 그 중요도 값이 큰 상위 30개의 용어들을 찾아내어 이를 클러스터를 대표하는 정보로 이용한다. 클러스터를 대표하는 정보에서 중요도가 낮은 용어들을 배제하는 이유는, 문서 클러스터 내에서 용어 중요도가 낮은 용어는 해당 클러스터를 대표하는데 기여하는 바가 낮고 오히려 클러스터 정보를 표현하는데 있어 잡음(noise)으로 작용할 소지가 있기 때문이다.

3.3 클러스터 제목 생성

3.3.1 디렉토리 경로 선택

검색결과 문서집합에 대해 클러스터링을 수행하여 만들어진 각 문서 클러스터에 대한 제목을 붙여주기 위해서는, 앞 절에서 설명한 바와 같이 각 클러스터를 대표할 수 있는 정보들을 우선 구성해 야 한다. 클러스터 제목의 생성은 이렇게 구성된 클러스터 정보가 디렉토리 계층구조 상의 어떤 디 렉토리와 유사한지를 찾아내는 문제라 할 수 있다. 즉, 클러스터 정보와 가장 유사한 디렉토리에 대한 계층구조 상의 경로가 해당 클러스터의 제목으로 사용될 수 있는 것이다. 이를 위해 클러스터 정보와 디렉토리 계층구조 상에 있는 색인 정보를 루트에서부터 하향식 기법으로 비교해 나가면서, 그 유사도가 높은 경로를 선택한다.

디렉토리 계층구조 상에서 클러스터의 제목으로 사용될 경로명을 선택하기 위해, 루트로부터 각 계층에서 유사도가 높은 두 개의 서브 디렉토리를 반복적으로 선택하여 나가는 방식을 이용한다. 즉, 최상위 14개 디렉토리 중 유사도가 높은 두 개의 디렉토리를 선택하는 것을 시작으로 하여 선택된 디렉토리의 서브 디렉토리에서 다시 클러스터 정보와의 유사도를 계산, 그 값이 큰 두 개의 디렉토리를 선택하는 것이다. 이러한 디렉토리 경로 선택은 더 이상 서브 디렉토리가 존재하지 않거나 클러스터 정보표현에 사용된 단어가 서브 디렉토리에 나타나지 않을 경우에 도달할 때까지 반복적으로 수행된다. [그림 3]은 클러스터 제목을 찾기 위해 사용된 알고리즘을 의사 코드로 나타낸 것이다. 이 알고리즘에서 사용되는 클러스터와 각 디렉토리와의 유사도를 구하는 수식은 다음과 같다.

클러스터 대표 용어의 수 (30) : N
클러스터 내 i번째 대표 용어의 중요도 : w_i
클러스터 용어 i의 디렉토리 내 출현빈도 : tf_i

클러스터 C와 디렉토리 D의 유사도 :

$$Sim(C, D) = \sum_{i=0}^{N} tf_i \times w_i$$

이는 클러스터 정보표현에 사용된 30개의 용어가 비교 대상이 되는 디렉토리 내에서 나타난 빈도수에 클러스터 내에서 해당 용어가 갖는 중요도를 반영시킴으로써 최종적으로 클러스터와 디렉토 리와의 유사도를 구하는 것이다.

```
search the cluster terms' info in the root index file;
make subdirectory's cluster term index info using above info;

while(there exist any cluster term)
{
    calculate subdirectories' similarities;
    find the two subdirectories with the high similarity;
    goto selected subdirectories;
    make subdirectory's cluster term index info using current directory
info;
    calculate subdirectories' similarities;
}
```

[그림 3] 클러스터 제목 생성 알고리즘

위의 방법을 이용하면 최종적으로 여러 개의 디렉토리 경로명이 클러스터 제목 후보로 선택된다. 따라서, 이들 클러스터 제목 후보 중에서 클러스터 제목으로 가장 적합한 대상을 선택해야 한다는 2차적인 문제가 발생한다. 이 문제를 해결하기 위해서는 선택된 클러스터 제목 후보들에 대해 순위를 부여해야 한다.

3.3.2 선택된 제목에 대한 순위 부여

디렉토리 경로명을 선택하는데 사용된 유사도 계산 방법은 현재 디렉토리를 기준으로 그 하위의 모든 디렉토리가 포함하고 있는 정보를 하나의 비교 대상으로 삼고 있다. 디렉토리 경로명을 선택하기 위해 사용된 이러한 계산 방법을 그대로 이용하여 최종적으로 선택된 디렉토리 경로명들에 대한 순위를 부여하는 것은 바람직하지 않다. 왜냐하면, 디렉토리 경로 선택 단계에서 사용되는 계산 방법은 선택된 디렉토리 경로명 자체에 대한 정보뿐만 아니라 해당 디렉토리의 형제 디렉토리(sibling directory)에 대한 정보까지도 비교대상에 포함시키고 있고, 계산이 수행되고 있는 계층에서 동일 계층에 있는 다른 디렉토리와의 비교를 위해 사용되는 것이므로, 최종적으로 선택된 서로 다른 계층을 갖는 디렉토리 경로 자체에 대한 순위를 부여하는데 이용하기에는 부적절하기 때문이다. 따라서, 디렉토리 경로명에 포함되어 있는 용어정보만을 하나의 비교 대상으로 삼아 이를 클러스터 정보표현과 비교하는 방식으로 선택된 각 디렉토리 경로명에 대한 순위를 결정해야 한다. 이를 위해 다음의 수식을 이용하여 선택된 각 디렉토리 경로명에 대한 적합성을 계산하였다.

T : 디렉토리 경로명에서 클러스터 대표 용어의 출현 수
L : 디렉토리 경로명이 갖는 레벨

$\text{Relevance}(DP) = \frac{T}{L}$: 디렉토리 경로명 DP의 적합도

디렉토리 경로 선택 단계에서 선택된 모든 클러스터 제목 후보에 대해 위의 수식을 적용한 계산을 수행하고, 그 결과로 얻은 값에 의해 내림차순으로 정렬하여 클러스터 제목에 대한 순위를 얻는다. 이렇게 얻은 순위에 따라 상위 5위에 포함되는 클러스터 제목을 선택하여 이를 최종적으로 클러스터를 대표하는 제목으로 이용하도록 하였다.

이상으로 본 연구에서 수행한 검색결과 문서집합에 대한 클러스터링과 이를 이용해 만들어진 문서 클러스터에 대해 디렉토리 계층정보를 이용하여 제목을 붙이는 방법에 대해 살펴보았다. 다음에서는 실제로 문서 클러스터에 제목을 붙여주는 실험을 통해 본 연구에서 제시한 제목 생성 방법에 대한 평가를 수행한다.

4. 실험 및 평가

본 논문에서 제안한 클러스터 제목 생성 기법의 성능을 검증하기 위해 다음 두 가지 실험을 수행하였다.

[실험 1] 검색결과 문서집합에 대해 어미 트리 클러스터링을 수행하여 만들어진 문서 클러스터에 대한 클러스터 제목 생성
[실험 2] '야후 디렉토리 사이트' 내에 분류되어 있는 문서집합에 대한 클러스터 제목 생성

4.1 실험 1

[실험 1]은 검색결과 문서집합에 대해 어미 트리 클러스터링을 수행한 결과로 만들어진 문서 클러스터에 대해 클러스터 제목을 생성한 것이다. 이 실험을 위해서 4개 질의를 검색엔진 '엠파스'에 던져 그 결과로 받은 각 100개의 문서를 대상으로 어미 트리 클러스터링을 수행하였다.

[표 2]는 어미 트리 클러스터링의 결과로 만들어진 클러스터의 개수와 문서집합의 클러스터 할당정도를 나타낸다. 문서 클러스터링을 수행한 결과로 만들어진 각 클러스터는 중복 허용이라는 클러스터링 알고리즘의 특성 때문에 하나의 문서가 대략 두 개의 클러스터에 할당되는 현상을 보인다.

이러한 클러스터들을 대상으로 클러스터 제목 생성 기법을 적용해 본 결과, 하나의 질의어에 대해 만들어진 클러스터들이 유사한 클러스터 제목을 갖게 됨을 알 수 있었다. [그림 4]는 이런 예를 나타낸다.

질의어	생성된 클러스터의 개수	클러스터에 할당된 문서 수	클러스터에 할당되지 않은 문서 수	클러스터에 할당된 문서의 중복을 포함한 총합
TV	6	64	36	115
검색엔진	6	74	26	135
인공지능	7	88	12	190
정보검색	6	45	55	65
	6.25	67.75	32.25	126.25

[표 2] 어미 트리 클러스터링 결과

Cluster [0]
뉴스와미디어:텔레비전:전국방송국:케이블TV:아리랑TV
뉴스와미디어:텔레비전:케이블TV
뉴스와미디어:텔레비전:전국방송국
비즈니스와경제:회사:컴퓨터:통신,네트워킹:한국에이아이소프트(주):ILLUSTRA
비즈니스와경제:회사:컴퓨터:하드웨어:워크스테이션:한국썬.마이크로시스템즈(주)

Cluster [1]
컴퓨터와인터넷:인터넷:WWW:웹엔터테인먼트:인터넷드라마,시리즈
뉴스와미디어:텔레비전:전국방송국:케이블TV:아리랑TV
컴퓨터와인터넷:인터넷:기관,단체:한국전산원
뉴스와미디어:라디오:전국방송
뉴스와미디어:텔레비전:케이블TV

[그림 4] 질의 'TV'에 대한 클러스터 제목 생성 예

[실험 1]을 통해 본 논문에서 제안한 클러스터 제목 자동생성 기법이 전체적으로 볼 경우에는, 본래의 질의어에 부합되는 적절한 클러스터 제목을 생성함을 알 수 있었다. 그러나, 하나의 질의어에 대해 만들어진 각 클러스터를 뚜렷이 구분하여 주지는 못함을 알 수 있었다. 이러한 결과는 검색결과 문서집합에 대해 클러스터링을 수행하여 얻은 각 클러스터가 가지는 문서의 중복성이 원인이 되어 나타나는 결과이다. 이는 STC 기법의 특성으로 클러스터간 유사성을 줄이고, 클러스터를 특징지을 수 있는 정보를 추출해야 한다는 문제를 안고 있다. [실험 1]의 방법으로는 본 연구에서 제안한 문서 클러스터 제목의 자동생성에 대한 성능을 독립적으로 평가하는데 한계가 있기 때문에 다음의 [실험 2]와 같은 방법을 이용하여 평가를 수행하였다.

4.2 실험 2

[실험 2]는 본 논문에서 제안하는 디렉토리 계층정보를 이용한 클러스터 제목의 자동생성에 대한 성능만을 평가하기 위해 수행되었다. 이는 문서 클러스터를 신뢰하는 상태에서 클러스터 제목 생성에 대한 성능을 평가하기 위한 실험 방법이다. 신뢰할 만한 문서 클러스터를 얻기 위해 '야후 디렉토리 사이트'에 이미 수작업으로 분류되어 있는 각 디렉토리 문서집합을 하나의 클러스터로 간주하여 문서 클러스터를 만들었다. 디렉토리 계층 상에서 문서집합을 선택하는데 있어 트리의 중간노드에 해당되는 디렉토리와 단말노드에 해당되는 디렉토리를 구분하여 각 20개의 디렉토리를 임의로 선택하였다. 그리고, 이에 대해 각 10개

의 문서를 다시 임의로 선택하여 실험의 대상이 되는 문서 클러스터를 구성하였다.[1]

선택된 실험 대상 클러스터들에 대해 각각 계층정보를 이용한 클러스터 제목 생성을 수행하였다. 클러스터의 제목으로 선택되는 디렉토리 경로명은 그 일치정도에 따라 순위가 주어지고, 상위 5위 안에 포함되는 경로명들이 클러스터의 제목으로 선택된다. 다음의 [그림 5]는 클러스터 2와 17에 대해 계층정보를 이용해 제목을 생성한 결과이다.

```
[2] 건강과 의학:체중조절:다이어트
비즈니스와 경제:회사:식품, 음식:조식, 아침식사 [0.8]
건강과 의학:체중조절:다이어트 [0.75]
건강과 의학:식품과 영양 [0.75]
비즈니스와 경제:회사:건강, 의학:체중조절, 다이어트 [0.67]
비즈니스와 경제:회사:식품, 음식:냉동식품 [0.6]

[17] 레크리에이션과 스포츠:자동차
레크리에이션과 스포츠:자동차 [0.5]
레크리에이션과 스포츠:자동차:운전:자동차보험 [0.4]
레크리에이션과 스포츠:자동차:제조사와 모델명 [0.4]
비즈니스와 경제:회사:자동차:중개업:제조사 [0.29]
비즈니스와 경제:회사:동물:개,강아지:서비스:애견관리 [0.14]
```

[그림 5] 클러스터 제목 생성 예

클러스터 2의 경우, 클러스터 제목 자동 생성을 통해 얻은 5개의 클러스터 제목 중 2위에 해당하는 제목에 본래의 클러스터가 갖는 경로와 동일한 제목이 주어졌으며, 클러스터 17의 경우 1위에 동일한 제목이 주어졌다. 또한, 나머지 순위에 주어지는 제목들도 본래의 클러스터 제목과 유사한 정보를 포함하고 있음을 알 수 있다. 다음의 [표 4]는 각 클러스터의 제목으로 선택된 디렉토리 경로명들이 본래의 디렉토리 이름과 일치하는 정도를 분석한 결과를 나타낸다.

클러스터 ID	일치하는 제목의 순위	정성적 제목 평가의 순위	일치 레벨 /전체 디렉토리 레벨
1	없음	3	2 / 3
2	2	2	3 / 3
3	없음	2	2 / 3
4	2	2	3 / 3
5	1	1	7 / 7
6	2	2	3 / 3
7	2	2	5 / 5
...	...	...	...
38	1	1	3 / 3
39	1	1	3 / 3
40	없음	3	0 / 3
평균	57.5%	90%	3 / 3.975

[표 4] 클러스터에 대한 제목 생성 결과 분석

위의 표에서 사용된 순위는 최종적으로 클러스터의 제목으로 선택된 5개 제목 중에서 순위 몇 위에 해당되는 제목이 클러스터 제목으로 가장 적합한 것인가를 선택한 것이며, '없음'의 경우는 자동 생성된 5개의 클러스터 제목 내에 클러스터를 대표하는 제목이 나타나지 않았다는 것을 의미한다.

[표 4]에서 '일치하는 제목의 순위'는 클러스터 제목 자동 생성 결과에서 본래의 클러스터 제목과 정확하게 일치하는 제목이 나타난 순위를 의미한다. 정확하게 일치하는 제목이 나타나는 경우는 평균 57.5%로 다소 낮은 수치를 보이고 있다. '정성적 제목 평가의 순위'는 생성된 제목이 본래의 클러스터 제목과 개념적으로 일치하는 경우, 이를 클러스터의 제목으로 적합하다고 판단하는 방법으로 결과를 평가한 것이다. 이 경우 본래의 클러스터 제목과 의미적으로 유사한 제목이 90%정도 나타남을 알 수 있었다. 이 방법에서 클러스터 제목으로 적합하다고 평가된 것의 순위가 4위 혹은 5위인 것을 제외하고도 유사 제목이 80%정도 생성되었음을 감안할 때, 본 연구에서 제안한 클러스터 제목 자동생성 기법이 클러스터의 의미를 잘 대표할 수 있는 제목을 생성해냄을 알 수 있다.

[표 4]에서 '일치 레벨/전체 디렉토리 레벨'은 본래 디렉토리 이름이 갖는 디렉토리 구조 상의 계층과 클러스터 제목으로 선택된 디렉토리 경로의 계층이 일치하는 정도를 나타낸 것이다. 예를 들어, 클러스터로 사용되는 본래 디렉토리의 경로명이 '엔터테인먼트:영화:개봉영화안내'이고, 제목 자동생성의 결과로 선택된 디렉토리 경로명이 '엔터테인먼트:영화:작품'일 경우, 전자가 갖는 레벨이 3이고, 선택된 디렉토리 경로명에서 이와 일치하는 레벨은 '엔터테인먼트:영화'의 2 레벨이므로 '일치 레벨/전체 디렉토리 레벨'에 주어지는 값은 '2/3'가 된다. 이러한 레벨 평가의 결과로, 선택된 클러스터 제목의 레벨이 기준이 되는 디렉토리 이름의 레벨에 대해 평균 75.5% 정도 일치함을 알 수 있었다.

클러스터 제목 생성의 결과로 본래의 디렉토리 이름과 동일한 제목이 주어지지 않는 경우는 일반적으로 문서집합을 나타내는 디렉토리의 이름이 해당 문서집합에 포함된 포괄적인 정보를 추상적으로 표현한 경우에 해당되었다. [그림 6]은 이러한 경우에 해당되는 클러스터 제목 생성 결과의 예이다. 클러스터 3의 경우, 클러스터에 포함된 문서들에 일반적으로 나타나는 정보는 상품광고를 위한 각종 상품명과 가격, 상품판매처등에 대한 정보들이였다. 클러스터 15의 경우는 취업정보와 관련하여 각종 회사에 대한 정보를 포함하고 있어 그 결과 특정 회사와 관련된 라벨이 추출되었다. 이러한 문제는 본 연구에서 클러스터 제목 을 생성함에 있어 이용한 정보는 단순 단어출현빈도와 이들에 대한 계층적인 정보뿐이기 때문에 발생한다. 즉, 의미적인 정보를 고려하지 않기 때문에 발생하는 근본적인 문제이다.

클러스터를 대표하는 용어로 선택된 단어가 본래의 디렉토리가 아닌 다른 디렉토리에서 더 많은 비중을 가지면서 사용되는 경우에도, 디렉토리 계층정보의 편중으로 인해 잘못된 클러스터 제목이 선택됨을 알 수 있었다. [그림 7]은 이러한 경우에 해당되는 예이

다. 예에서 클러스터 13에 포함된 문서들에 가장 빈번하게 나타나는 단어들은 '노동' 과 '조합' 이다. 그렇기 때문에 이러한 단어가 많이 사용되고 있는 다른 디렉토리들이 선택되었음을 결과를 통해 알 수 있다.

[3] 비즈니스와 경제:회사:소매:온라인 쇼핑:디렉토리
예술과 인문:디자인 아트:가구 디자인:이벤트 [0.4]
비즈니스와 경제:생활광고 [0.33]
비즈니스와 경제:회사:컴퓨터:통신, 네트워킹 [0.33]
비즈니스와 경제:회사:컴퓨터:컨설팅 [0.33]
비즈니스와 경제:회사:컴퓨터:통신, 네트워킹:소프트웨어 [0.29]

[15] 비즈니스와 경제:취업,채용:구인정보
비즈니스와 경제:회사:컴퓨터:통신,네트워킹:LG정보통신 [0.83]
건강과 의학:교육:연구소,기관:한국 학교보건교육 연구회 [0.6]
비즈니스와경제:회사:컴퓨터:통신,네트워킹:소프트웨어:전자통신[0.5]
비즈니스와경제:회사:컴퓨터:컨설팅:시스템통합:KCC정보통신[0.5]
비즈니스와경제:회사:건축:기업간거래(B2B):취업,채용 [0.5]

[그림 6] 잘못된 클러스터 제목 생성 예 1

[13] 정부:법:노동법
비즈니스와 경제:노동:노동조합:서울여성노동조합 [1.0]
교육:기관,단체:전문단체:노동조합:전국교직원노동조합 [1.0]
교육:기관,단체:전문단체:노동조합:한국교원노동조합 [0.83]
비즈니스와 경제:회사:예술과 공예:기관,단체:노동조합 [0.5]
비즈니스와 경제:생활광고 [0.25]

[그림 7] 잘못된 클러스터 제목 생성 예 2

5. 결론 및 향후 연구

본 논문은 문서 클러스터에 대한 기존 연구에 기반하여, 문서 클러스터에 대한 적절한 제목을 생성하기 위한 모델을 제안하였다. 디렉토리 계층정보를 이용한 클러스터 제목 생성 기법은 '야후 디렉토리 사이트'의 디렉토리 정보를 구조화하여 이를 이용하였다.

본 연구에서 제안한 문서 클러스터에 대한 제목 생성 방법을 평가하기 위해서, '야후'의 각 디렉토리에 있는 문서집합을 문서 클러스터로 간주하여, 이를 대상으로 실험을 수행하였다. 그 결과 문서 클러스터를 구성한 본래의 디렉토리 이름과 일치하는 클러스터 제목을 얻은 경우가 57.5% 정도이고, 정성적인 평가에 의한 평가의 경우 의미적으로 적합한 클러스터 제목이 생성되는 경우가 90%정도 이였다. 따라서, 본 연구에서 제안한 문서 클러스터 제목의 자동생성 기법이 클러스터의 의미를 잘 대표할 수 있는 제목을 생성해내는 유용한 방법임을 알 수 있었다.

문서 클러스터에 대해 적절한 제목을 부여하는데 있어 전제되어야 하는 사항은 클러스터의 응집도가 높아야 하고 각 클러스터가 뚜렷이 구분되어야 한다는 것이다. 이러한 클러스터를 대상으로 클러스터 제목 자동생성을 수행해야만 변별력 있는 클러스터 제목을 생성해낼 수 있기 때문이다. 따라서, 향후 연구에서는 클러스터의 응집성과 각 클러스터 간의 변별력을 높이기 위한 방법에 대한 연구가 수행되어야 한다. 또한, 클러스터 제목 생성에 이용되는 디렉토리 계층 구조와 관련하여서는 디렉토리 계층 구조 내의 각 계층이 갖는 특성에 대한 연구와 디렉토리 계층정보의 불균형을 해소하기 위한 방법과 관련한 연구가 이루어져야 한다.

6. 참고 문헌

[1] Yahoo Korea "http://kr.dir.yahoo.com"

[2] Oren Zamir, "Fast and Intuitive Clustering of Web Documents", Qual's Paper, University of Washington.

[3] Oren Zamir & Oren Etzioni, "Grouper: A Dynamic Clustering Interface to Web Search Results", WWW8.

[4] Ricardo Baeza-Yates & Berthier Ribeiro-Neto, "Modern Information Retrieval", Addison-Wesley, 1999.

[5] 윤보현, 강현규, 고형대, "자동 문서 클러스터링을 위한 디스크립터 추출 방안", 제13회 한국정보처리학회 봄 학술발표논문집, 2000.