특허 및 기술정보의 연계 검색에 관한 연구

Research for Interlink Retrieval of Patent and Technical Information

송종철(J.C. Song) 정보유통연구팀 연구원
이성용(S.Y. Lee) 정보체계연구팀 연구원
홍기채(G.C. Hong) 정보유통연구팀 책임기술원
강윤희(Y.H. Kang) 천안대학교 정보통신학부

정보통신을 비롯한 다양한 분야에서 새로운 기술과 아이디어를 이용한 기술개발이 활발하게 이루어짐에 따라 창의적 연구결과에 따른 특허 출원도 급격히 증가하고 있다. 본 고에서는 사용자의 특허검색 과정에서 특허와 관련된 기술 문서를 동시에 검색이 용이하도록 지원하는 시스템의 개발에 대하여 논하고자 한다. 특허 및 기술정보 연계 시스템은 신규 특허 문서에 대한 분류를 위해 주제별 주요 용어를 추출하고 특허 문서와 유사한 기술 문서를 코사인 유사도 기법을 사용하여 유사도에 따라 기술 문서를 사용자에게 제공할 수 있도록 설계하였다.

I. 서 론

최근 정보통신을 비롯한 다양한 분야에서 새로운 기술과 아이디어를 이용한 기술개발이 활발하게 이루어 짐에 따라 창의적 연구결과에 따른 특허 출원도 급격히 증가하고 있다. 본 논문에서 설계한 특허 및 기술정보 연계 시스템은 특허 정보 사용자에게 특허와 관련된 기술문헌을 동시에 제공함으로써 특허 정보에 대한 이해를 높여 특허 정보의 활용도를 향상시키는 동시에 보다 명확한 특허출원이 가능하도록 지원하는 것을 목적으로 한다.

또한, 본 연구에서는 사용자의 특허검색 과정에서 특허와 관련된 기술 문서를 동시에 검색이 용이하도록 지원하는 방식을 개발하고자 하는 것이다. 이를 위해 특허 관련 정보를 사전에 분석하고, 관련된 기술 정보를 클러스터링 함으로써 두 유형의 문서간 상호 연계를 통한 통합검색이 가능하도록 하기 위한 지능형 정보검색 방식의 개발을 목적으로 한다.

본 논문에서는 지능형 정보 검색을 위한 주요 구성요소인 특허 정보와 기술정보에 대한 연계 시스템 구조, 유사도를 기반으로 한 특허계층 구조 내의 특허 문서를 위치시키기 위한 분류 기법 및 특허 문서와 기술 문서의 연계방법을 기술한다.

본 논문의 제 II장에서는 특허 및 기술정보 연계를 위한 내용 기반 라우팅 및 클러스터링 관련연구를 기술하며 제 III장에서는 두 개의 상이한 도메인의 정보에 대한 연계 검색 지원을 위한 시스템 구성, 연구 수행을 위한 기술정보와 특허 정보 문서의 클러스터링 기법의 흐름 및 단계별 수행내용을 기술한다. 제 IV장에서는 클러스터링 처리를 위한 주요용어추출 방법 및 문서벡터 구성을 위한 벡터공간 구성 방식, 기술정보 문서와의 유사도 계산 방법 등에

대하여 기술하고, V장에서는 결론으로 설계된 시스템의 확장성 및 활용성 등에 대하여 기술한다.

II. 관련 연구

사용자의 특허검색 과정에서 특허와 관련된 기술문서를 동시에 검색이 용이하도록 지원하기 위해서는 연관검색을 위한 메커니즘을 제공하여야 한다. 분산 환경 하에서 정보 서버는 연관 접근 제공이 어려우며, 이것으로 인하여 정보공간 확장성이 제약된다. 폭발적 정보공간의 확대는 단일 전역 색인을 기반으로 한 색인 계획을 불가능하게 하며 질의의 분산처리를 가로막는 요인으로 작용한다. 이러한 이유로 효율적인 정보접근을 위해서 내용 기반 라우팅과 질의정재가 요청된다. 내용 기반 라우팅은 사용자 질의를 적절한 서버로 지시 또는 전달하는 과정이다. 내용 기반 질의를 응용하여 구현될 수 있는 질의 서비스는 다음과 같다.

- 점진적인 네트워크 내용의 발견 및 질의 구성 시 변별력을 높이기 위한 지도로서 균일한 질의 인터페이스를 지원한다.
- 정보 서버의 계층구조를 통한 서버 내용의 기술을 전파한다.
- 질의 서버에서 기대되는 관련 정도에 따라 가용 서버에 개별 질의들을 라우팅한다.

내용 레이블(content label)은 분산 환경 하에서 연관 접근의 의미를 기반으로 하여 조직되며, 분산된 정보 서버들이 상호 접근할 수 있는 서비스를 제공한다. 내용 레이블을 기반으로 한 내용 라우터(content router)는 분산 연관 접근 구현의 기본으로 작동한다[1].

사용자의 관점에서 바라본 정보의 브라우징과 검색은 개별 데이터 객체의 위치에 관계 없이 수행되어야 한다. 내용 레이블은 개별 서버에 대한 기술로서 자동적인 내용 레이블의 구성을 수행하며, 또한 정보 서버에 질의를 라우팅 한다. 내용 라우팅 시스템 아키텍처의 장점은 다음과 같다.

- 활용성: 내용 레이블은 사용자에게 피드백을 제공받아 가용자원에 대한 학습과 질의를 구성한다. 내용 레이블을 이용하여 검색 공간을 항해하며, 또한 정보 검색을 수행할 수 있다.
- 확장성: 내용 라우팅은 내용을 기반으로 한 정보 서버의 방대한 네트워크에 대한 균일한 정보 접근을 제공한다.
- 효율성: 사용자가 내용 라우팅 시스템을 이용하여 검색할 경우, 내용 라우팅 시스템은 서버 네트워크를 효율적으로 지원하며, 내용 레이블은 정보 서버의 내용을 사용자 검색에 적합한 검색 결과로 정재한다.

정보 서버는 사용자의 중첩 검색(interleaved searching) 및 브라우징 행위를 지원함으로써 특정 문서에 대한 정보 가용성을 이해할 수 있다. 또한, 정보 서버는 사용자 질의에 대한 정재 및 자동 질의 라우팅을 위한 메타 정보로서 내용 레이블을 사용한다.

본 연구에서는 특허 및 기술정보의 정보량 증가에 따라 단일정보 서버 내에 모든 정보를 유지하지 않고 문서를 분류한 후 다수의 정보 서버에 정보를 분산하여 유지하는 방법을 적용할 수 있도록 시스템 설계 시 고려한다. 다수 정보 서버의 활용을 위해서는 내용 기반 라우팅과 같은 분산 정보시스템이 사용되며 효율적인 검색을 지원하기 위해서는 관련 문서들에 대한 주제별 및 특성별 클러스터링이 이루어져야 한다[1],[2].

클러스터링이란 주어진 데이터 셋을 서로 유사성을 가지는 몇 개의 클러스터로 분할해 나가는 과정으로, 하나의 클러스터에 속하는 데이터 점들 간에는 서로 다른 클러스터 내의 점들과는 구분되는 유사성을 갖게 된다[3]. 클러스터링은 각각의 정보들을 관련된 항목의 클러스터로 조직하고, 조직된 클러스터는 광역 정보공간 내에 속하는 사용자와 시스템을 지원한다. 클러스터 추상화는 내용의 상세함에

관계 없이 광역 정보공간을 단일 단위로써 취급한다. 정보공간의 영역을 탐색하는 사용자에게 관련된 클러스터를 식별할 수 있도록 함으로써 정보필요 및 정보공간의 복잡성에 대하여 사용자가 명시할 수 있도록 한다.

클러스터는 시스템의 분산 구성요소 간의 작업의 분리와 자원 할당을 위한 편리한 단위를 제공한다. 클러스터는 용어 표현 및 문서로 구성된 메타정보 기술이다. 클러스터링은 데이터 마이닝 방법의 일부분으로서 자동적인 질의 형성과 유사문서 검색을 위해 관련된 문서의 특징을 추출하는 것을 목적으로 한다.

클러스터링 방법은 크게 파티션 접근(Partitioning approach)과 계층적 접근(Hierarchical approach)으로 나눌 수 있다. 파티션 접근 방법은 어떠한 범주 함수를 최적화 시키는 K개의 파티션을 결정해 나가는 방법으로, 유클리언 거리(Euclidean distance) 측정법에 기반한다. 파티션 접근 방법은 클러스터의 무게중심점을 대표 값으로 분할해 나가는 K-means 방법과, 클러스터 내의 중심과 가장 근접한 객체로 대표점을 찾아가는 K-medoid 방법이 있으며, 분할을 위한 초기 값 선정방식이나 대표 값 선정방식에 따라 다양한 형태로 변형될 수 있다[4].

계층적 접근은 처음에 각각의 데이터 점을 하나의 클러스터로 설정한 후 이들 쌍(pair)간의 거리를 기반으로 하여 분할/합병을 수행하는 상향식 방식으로 모든 점들이 하나의 단위 클러스터에 속하게 될 때까지 그 이력정보를 유지해 나가게 된다. 쌍 간의 거리를 어떻게 측정하느냐에 따라 단일 링크(single linkage), 완성 링크(complete linkage) 또는 센트로이드 링크(centroid linkage) 등을 이용하는 다양한 방법이 존재한다[5].

클러스터링 알고리즘은 클러스터링을 위해 거리 기법과 확률적 기법을 사용한다. 거리와 유사도를 정의하거나 확률을 사용하여 문서 구조에 대한 선험적 지식(a priori)을 사용한다. <표 1>은 문서 클러스터링을 위한 기법 중 거리기반 방법론과 확률 모델에 대하여 기술한 것이다.

<표 1> 문서 클러스터링 방법

방법론	접근 방법	기법
거리기반 방법론	• 클러스터링을 위한 특성(feature)으로서 서로 다른 문서 내에 출현하는 선택된 단어 집합을 사용 • 각 문서는 특성 벡터로서 표현 • 문서집합은 다차원 공간으로 표현	• *k*-means 분석 • 계층적 클러스터링 • 근접 이웃 클러스터링 (최단 근접)
확률 모델	• 문서 *j*가 다수의 속성 값을 갖는 클래스 *Ci*에 포함될 확률 값으로서 문서 분류 수행 • 속성을 표현하는 변수는 독립적임 • 문서를 가장 높은 확률을 갖는 클래스에 포함	• Bayesian 분류 (AutoClass)

거리기반 방법론은 공간 내의 거리 평가 기준을 정의하기가 쉽지 않고 문서 내의 구성 단어 수가 적으므로 희소 벡터공간을 유지하는 단점을 갖고 있다. 확률 모델은 특성공간이 너무 커지거나 벡터공간이 내재된 특성에 의존하는 문제점을 갖는다[6]. 이러한 문제점을 해결하기 위해서는 문서 클러스링을 수행하는 도메인의 특성에 따라 효과적인 클러스터링 기법이 사용되어야 한다. 정보기술 및 특허 문서와 같은 다차원 공간(high dimensional space)의 클러스터링을 수행하는 경우 문서를 구성하는 도메인에 대한 분석 및 기존 클러스링 방법의 개선이 필요하다.

III. 특허 및 기술정보 연계 시스템

본 절에서는 특허 및 기술정보 연계 시스템의 구성 과정에서 사용하는 텍스트 선택기법을 기술하고 특허 및 기술정보 연계를 위한 시스템 구성의 단계별 수행내용을 기술한다. 구성될 특허 및 기술정보 연계 시스템에서 사용자는 특허 정보에 대한 질의 후 얻어진 결과문서에 대해 필요로 하는 특허 문서를 선택한다. 기술정보 검색 시스템에서는 위에서 선택되어진 특허문서와 유사성이 있는 기술 문서를 검색하여 결과 집합으로 출력하며, 제시된 결과 집합에 대하여 문서 클러스터링을 수행한 후 유사도가

높은 상위 기술 문서를 사용자에게 제공할 수 있도록 설계한다.

1. 텍스트 선택 기법

특허정보 연계 검색 시스템은 텍스트 선택기법을 사용하여 특허 정보와 유사한 기술 문서를 연계한다. 이를 위해 다음 4가지 기본 구성요소를 갖는다.

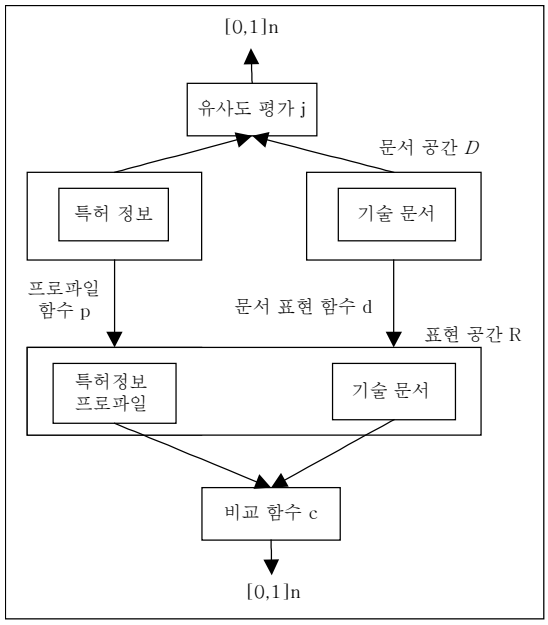
- 문서 표현을 위한 기법
- 정보 필요인 프로파일을 표현하기 위한 기법
- 프로파일과 문서 표현 비교 방법
- 결과 비교를 사용하는 방법

텍스트 선택의 목적은 정보 요구와 문서 표현 간의 유사도를 비교 연산 등을 통하여 자동으로 계산함으로써 실제로 필요한 정보를 찾기 위해 사람이 문서들을 비교하여 필요 정보들을 구성하고 생성하는 것과 유사한 결과를 도출하는 것이다. 4가지 구성요소 중 결과 비교는 사용자에게 결과를 출력하는 모듈로서 선택 모듈과 출력 모듈로 구성된다. 선택 모듈은 각 문서에 한 개 이상의 값을 할당하고 출력 모듈은 사용자의 선택에 대한 검색결과를 구성한다.

프로파일 획득 함수 p의 도메인은 I이고 가능한 정보의 집합과 프로파일 영역은 R이다. 비교 함수 c의 도메인은 $R\,X\,R$이며 영역은 $[0,1]^n$이다. 즉 문서와 프로파일은 0과 1 사이의 값을 갖는 n 개의 튜플이다. 이상적인 정보 필터링 시스템은 식 (1)과 같다.

$$c(p(IN),d(doc)) = j(IN, doc), \forall IN \in I, \forall doc \in D,$$
$$j : I \times D \rightarrow [0,1]^n \qquad (1)$$

식 (1)에서 IN은 사용자의 관심분야를 j 는 사용자의 관심분야와 문서 간의 관계 판단, 그리고 n은 차원을 표현한다. 본 연구에서는 텍스트 선택에서 적용되는 텍스트 필터링 기법을 특허 정보에 대한 정보 기술 문서 선택에 적용한다. (그림 1)은 특허 정보와 유사한 기술정보 텍스트 문서를 선택함으로써 전자통신연구원 내에 이미 구축된 기술정보와 새롭게 구축될 특허관련 서지 및 전문을 연결할 수 있음을 보인 것이다.



(그림 1) 특허 정보와 기술 문서를 위한 텍스트 필터링

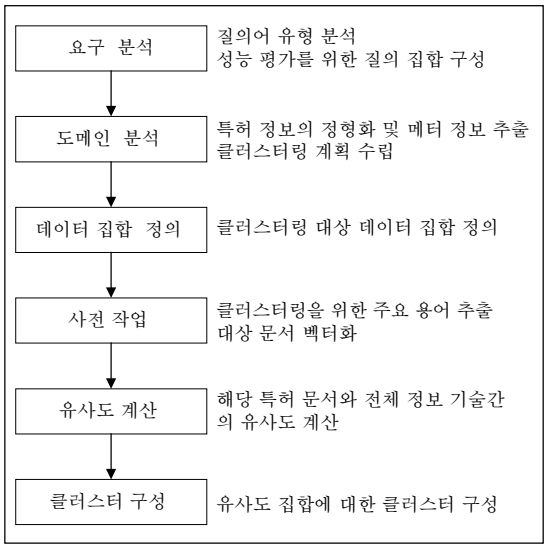
2. 특허 문서 분류 및 기술 문서 연계

특허 문서와 기술 문서간의 연계 검색을 위한 클러스터링은 (그림 2)의 과정을 통해 수행한다.

요구분석 과정에서는 클러스터링을 위해서는 사용 시스템에 대한 검색 유형을 정의한다. 본 연구에서 모든 검색의 시작은 특허 정보 검색으로부터 시작된다고 가정한다. 즉, 특허 정보 검색 서버에 대한 사용자 검색에 대해 관련 특허 정보 및 정보기술 관련 문서제공이 이루어진다.

데이터 집합 정의 과정에서는 특허 문서에 대한 파싱과 색인 구성을 위해 문서에 대한 태그 선정 및 중요도를 평가한다. 클러스터링을 위한 특허 문서는 특허 그룹 식별자, 특허 번호, 특허명, 특허 출현 연도 및 특허 관련 키워드 등의 메타 정보를 갖는다고 가정한다. 특허 정보의 검색 및 색인을 위해 특허 정보는 태그를 갖도록 구조화함으로써 문서에 대한 내용 기반 검색 및 색인 구성 시 태그에 대한 가중치를

적용할 수 있다. 특허 정보에서는 특허출원일과 특허 청구 범위 등이 검색에 중요하게 사용되는 필드이기 때문에 문서의 중요도를 결정하는 요소로도 활용될 수 있다.



(그림 2) 특허 문서 클러스터링 흐름도

메타 정보는 특허 정보를 구성하는 서지 및 본문의 분석과정에서 정형화된다. 클러스터링은 해당 문서 집합의 특징에 따라 적합한 기법이 활용되므로 문서 도메인의 특성 분석이 선행되어야 한다. 특허 정보의 구성은 특허 정보에 대한 문서 표현을 위해 서지 정보와 전문에 대해 벡터공간을 구성한다.

<표 2>의 특허 검색 질의 유형은 두 종류의 특허 정보에 대한 검색을 지원하기 위해 두 분야의 서지 및 전문을 연계할 수 있도록 검색을 위한 색인 구조를 설계한다.

<표 2> 특허검색 질의 유형

검색 유형	내용	비고
특허의 분야만 알고 있는 경우	특허 DB 내의 자신이 원하는 특허에 어떠한 것이 있는지 검색하는 경우	유사 특허 검색
관련 기술 문서를 검색하는 경우	정확한 특허의 내용을 알고 있으며 기술 문서 내에서 특허에 관련된 문서가 어떠한 것이 있는지 검색하는 경우	특정 특허 관련 기술 문서 검색

IV. 특허 정보의 클러스터링 기법

본 논문에서는 문서 집합의 클러스터링을 위해 용어 관계를 사용한다. 용어는 문서간의 관계를 표현하기 위한 기반으로 용어집합에 따라 문서가 클러스터될 수 있다. 클러스터 분석 방법은 N개 항목들의 데이터 집합을 M개 클러스터로 나누는 비계층적 방법과 항목들이나 클러스터 쌍이 성공적으로 연계된 근접한 데이터 집합을 생성하는 계층적 방법이 있다. 클러스터를 통해 형성된 동일 집단의 구성요소들은 높은 연관성을 갖고, 다른 집단의 요소들과는 낮은 연관성을 가져야 한다. 문서들의 클러스터 구성은 다음의 방법으로 이루어진다.

- 문서는 그 속에 포함되는 용어를 기반으로 클러스터링 된다.
- 문서는 여러 분야에 대해 클러스터링 된다.
- 용어는 지식을 축적하거나 질의의 정재 등을 위해 동시에 발생되는 문서를 기반으로 클러스터링될 수 있다.

특허 문서 클러스터링의 사전 작업은 문서 분류를 위한 분류체계의 구성, 클러스터링을 위한 주요 용어 추출 및 특허 문서를 읽은 후 각 문서에 대한 가중치 용어 벡터를 구성하는 단계 등 3단계로 이루어진다. 본 절에서는 (그림 2)의 과정 중 주요 용어 추출 및 용어 벡터 구성, 문서간 유사도 계산에 대하여 기술한다.

1. 주요 용어 추출

어떤 특정 분야에 속하는 용어들이 문서에 많이 나타나는 경우, 그 문서는 용어들이 속한 분야의 문서일 가능성이 높다. 정보추출을 통해 용어가 어떠한 개체를 나타내는지 인식하여 문서를 표현함으로써 문서가 내포하는 의미를 보다 정확히 반영할 수 있다. 주요 용어 추출과정을 통해 추출된 주요 용어는 반입된 특허 문서의 분류를 위해 사용된다.

질의와 관련된 용어를 선택하기 위해 질의 정재 알고리즘은 다른 용어와 동시에 출현하는 다른 용어

와의 조건부 확률을 사용한다. 즉, 특정 용어와 일치하는 문서가 주어질 때 다른 용어가 이들 문서 내에 존재할 수 있다고 가정하며 문서 내에서 동시에 출현하는 용어는 높은 관련성을 갖는다. 용어 T가 문서 그룹 G 내에 발생할 확률은 식 (2)와 같다.

$$P(G|T) = \frac{P(T|G)}{\sum_g P(T|G)} \quad (2)$$

베이지언 확률을 이용한 문서 내의 단어 T1, T2, T3, … Tn이 출현할 확률은 다음에 의해 계산된다.

$$P(G_g|T_{1,}T_{2,\ldots,}T_n) = P(G_g)\times P(T_1|G_g)\times P(T_2|G_g)\ldots\times P(T_n|G_g)$$

$p(G_g)$ 와 $P(T_g|T_g)$ 는 문서 집합 내의 용어의 출현빈도를 기반으로 근사치를 계산한다[8].

$P(T_g|G_g)$ 의 값은 용어 *count(Ti)/count*(G_g의 용어빈도) 에 의해 값을 구하며 신규 특허 문서 반입 시에 문서의 가중치를 기반으로 상위 용어들에 대해 소속 문서 그룹을 결정한다.

본 연구에서의 주요 용어 추출은 특허 문서에 대한 분류 과정에서 활용되어지는 용어집합을 생성하는 과정으로 앞으로 정보통신 관련 시소러스 구축에 활용하거나 구축된 시소러스에 추가될 수 있다.

- 특허 관련 문서로부터 용어를 추출한다.
- 특허와 관련된 그룹 식별자를 가정하며 가정된 특허 그룹 ID 내에 포함되는 특허 서지로부터 주요한 용어를 추출한다.
- 주요 용어의 추출은 문서 내의 용어 가중치를 고려하여 전체 문서 N개 중 최소 N/2에 발생하는 용어를 대상으로 하며 용어의 발생 위치(특허명, 특허범위, 기술의 특징)를 고려한다.

구성된 가중치 용어 벡터에 대하여 동시출현 관계의 용어 집합을 생성한다. 선정된 용어집합은 개념 용어 벡터로서 질의 확장 및 정재에 적용한다. 관련 용어 집합을 구성하기 위해 트리거 쌍을 사용할 수 있으나 트리거 쌍을 구성하기 위해서는 말뭉치로부터 통계 작업을 수행해야 하므로 나중에 고려한다. 용어 추출 과정을 위해 주제를 포함하는 문서로부터 관련 용어 집합을 구성하며 동시 출현 용어와 가중치 값이 0.05 이상의 값을 갖는 용어 집합을 추출한 후, 추출된 용어집합을 대상으로 용어가 문서 집합 내에 포함된 확률을 구한다. 용어 집합 구성 과정에서 적용한 용어선정 규칙을 기술하면 다음과 같다.

1) 문서 출현빈도 기반: *N/2*의 문서 내에서 공기 관계를 갖는 용어를 찾아 용어집합을 구성한다.
2) 전이 클로저(transitive closure) 기반: 용어 a와 b가 관련되고 b가 c와 관련될 때 a는 c와 관련된다. 연산자 "<"가 전이 연산자이며 a <b 이고 b <c 이면 a<c이다. 새로운 용어 집합 [a c]를 문서 출현빈도 집합에 추가하여 얻어진 전이집합을 용어집합에 추가한다.
3) 시소러스 내의 해당 용어가 표제어로 출현하는 경우 그 용어와 관련된 용어를 용어집합에 추가한다.

2. 문서 벡터 구성

수집된 문서에 대한 색인 구축은 문서 내에 발생된 용어의 빈도수를 기반으로 하여 가중치를 설정한다. 본 연구에서는 식 (3)을 사용하여 문서의 벡터공간을 구성한다. 벡터공간은 희소 행렬의 형태로 표현된다[6],[7].

$$w_{ik} = tf_{ik} \times \log(N/n_k)$$

T_k = D_i내의 k번째 용어
tf_{ik} = D_i내에서의 T_k의 출현빈도
idf_k = 문서집합 C에서 T_k의 역문서빈도
N = 문서집합 C의 문서수
n_k = 문서집합 C에서 T_k를 포함수는 문서수
$idf_k = \log(N/n_k)$

다음은 벡터공간 모델에서의 기본 전제조건을 나타낸다.

- 질의와 문서는 용어에 의해 식별된다.
- 문서 D는 m 차원의 벡터로써 표현된다.
- $D = <w_1, w_2, \cdots w_m>$, 이때 W_i는 용어 t_i의 가중치

다음은 구성된 문서 벡터의 예로, 용어 번호 3643과 문서번호 94의 가중치는 0.0385이다.

(3643,94)	0.0385
(3744,94)	0.0385
(3745,94)	0.0385
(3792,94)	0.0385
(3801,94)	0.0385

3. 정보기술 문서와의 유사도 계산

정보기술 문서와의 유사도 계산은 벡터 유사도 계산방법을 사용한다. 이를 위해 해당 특허 정보와 기술 문서에 대한 벡터를 구성한 후 두 벡터 사이의 유사 정도를 비교하는 방법을 사용한다. 본 방법은 두 벡터 사이의 각도를 나타내는 코사인 계수를 사용한다. 본 연구에서는 특허 문서와 정보 기술 문서와의 유사도를 측정하기 위해 다음의 순서에 따라 유사문서를 찾는다.

1) 모든 기술 문서에 대한 색인을 수행한다.
2) 신규 반입된 특허 문서에 대해 용어벡터를 구성한다.
3) 코사인 유사도를 계산하여 신규 반입 문서와 높은 연관성을 갖는 기술 문서를 얻는다.
4) 상위 N개에 대해 관련 링크를 구성한다.

구성된 용어집합 벡터와 기존에 구축된 정보기술 문서와의 유사도를 계산한 후, 유사도 값에 따라 유사문서를 연계한다. 본 연구에서는 문서에 대해 다차원의 벡터공간을 구성한 후 문서 간의 유사도를 계산하였다. 차원의 수는 문서를 구성하는 용어의 수에 의해 결정된다. 벡터 사이의 코사인 값을 측정하는 데 쓰이는 코사인 상관 계수를 바탕으로 하는 벡터 부합 연산을 사용하여 문헌과 질의 사이의 유사성을 계산할 수 있고 이 유사도를 사용하여 문헌에 대한 순위를 결정한다. 식 (4)에서 D_i 정보 기술 문서를, Q는 반입된 특허 문서의 벡터를 표현한다.

$$sim(Q, D_i) = \frac{\sum_{j=1}^{t} w_{qj} \times w_{d_{ij}}}{\sqrt{\sum_{j=1}^{t} (w_{qj})^2 \times \sum_{j=1}^{t} (w_{d_{ij}})^2}} \quad (4)$$

$$D_i = w_{d_{i1}}, w_{d_{i2}}, \ldots, w_{d_{it}}$$

$$Q = w_{q1}, w_{q2}, \ldots, w_{qt}$$

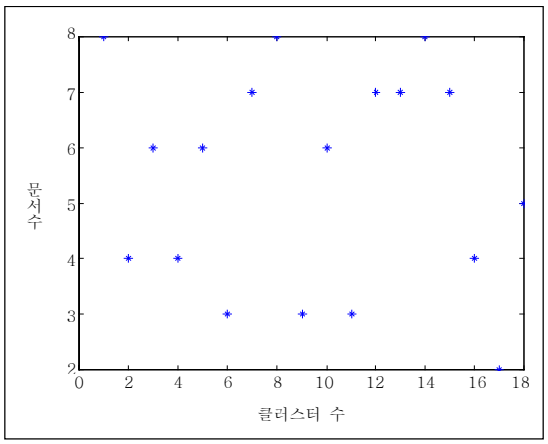
W = 문서 내의 용어의 가중치

본 모델은 순위부여 실험의 기본모델로 사용되었다. 문서 내의 용어빈도에 대한 코사인 상관계수를 용어 가중치로 사용함으로써 코사인 유사도 함수에 문헌 길이를 자동으로 정규화 할 수 있다. <표 3>은 실험을 위해 사용된 문서 리스트를 보인 것이다. 실험대상 문서는 검색 후 수동 분류된 분야별로 구분된 문서리스트를 대상으로 함으로써 클러스터링 후에 유사문서들의 리스트에 대한 유사 문서 집합이 정확도를 평가한다. 본 실험은 98개 문서를 대상으로 수행하여 18개의 클러스터를 얻었다.

<표 3> 문서 리스트

문서번호	문서구분	문서
58	AGENT	http://logic.stanford.edu/sharing/knowledge.html
59	AGENT	http://www-ksl.stanford.edu/owledge-haring/ndex.html
60	AGENT	http://www.cs.umbc.edu/kqml/index.html
61	AGENT	http://www.cl.cam.ac.uk/users/rwab1/agents.html
62	AGENT	http://www.cs.bham.ac.uk/~amw/agents/links/index.html
63	MULTI-MEDIA	http://www.dpc.or.kr/whitepaper/wp98/625.html
64	MULTI-MEDIA	http://etlars.etri.re.kr/EtlarsHome/lecture/multimedia/Multimedia.html
65	MULTI-MEDIA	http://www.tta.or.kr/StdInfo/WhiteBook/htm/2-3.htm

(그림 3)은 코사인 유사도 함수의 수행을 통해 얻어진 클러스터 결과를 보인 것이다. 초기 문서 집합은 주제별로 5개의 문서집합으로 분류하였으며 분류한 클러스터의 크기 편차를 사용하여 클러스터 구성의 정확도를 얻을 수 있다.



(그림 3) 클러스터링 결과

식 (5)는 연구 수행과정에서 유사도에 따라 구성된 문서 집합의 평가를 위해 사용된다. 정확도(c)는 클러스터 c의 정확도를 표현한다.

$$\text{정확도}(c) = \frac{c\text{ 내의 바르게 분류된 문서수}}{c\text{ 내의 분류대상 문서수}} \quad (5)$$

V. 결 론

본 연구에서는 정보량이 확대되고 있는 기술정보와 특허관련 서지 및 전문을 연계하여 검색할 수 있도록 클러스터링을 이용한 연계 검색 기법을 설계하였다. 앞으로는 클러스터링을 통하여 분류된 문서들을 분산 정보 서버상에 유지하며, 분산 환경에 적합한 내용 기반 라우팅 기법 등에도 활용할 계획이다.

본 연구는 사용자의 특허 정보 검색에서 관련된 기술 문서를 연결하는 것을 목적한다. 연구방법으로는 특허 정보와 기술 문서에 대한 상호 연계 검색을 위해 기술 문서에 대한 유사도를 기반으로 클러스터링을 구축한 후 특허 정보와 링크함으로써 사용자의 통합 검색을 지원한다. 이를 위해 특허 정보와 유사한 텍스트 문서를 선택함으로써 전자통신연구원 내의 기존에 구축된 기술정보와 신규로 구축되는 특허 관련 서지 및 전문을 대상으로 연결할 수 있다. 또한 문서에 포함된 주제어나 핵심어에 따라 문서를 다차원의 벡터공간에 배열시킨 후, 문서가 배열된 벡터 공간 상에서 문서 간의 유사도를 계산하고 정량화한다. 코사인 상관계수를 바탕으로 벡터 부합 연산을 사용하여 문서와 질의 사이의 유사성을 계산하며 질의에 대한 문서 순위를 결정함으로써 관련된 기술 문서들을 특허 정보와 연결한다.

참 고 문 헌

[1] M.A. Sheldon “Content Routing: A Scalable Architecture for Network-Based Information Discovery,” *PhD thesis*, MIT, Dec. 1995.

[2] J. Cowie and W. Lehnert, “Information Extraction,” *Communications of the ACM*, Jan. 1996, Vol. 39, No. 1, pp. 80 – 91.

[3] Alexander Hinneburg and Daniel A. Keim, “Clustering Method for Large Data Sets,” SIGMOD99, 1999.

[4] J. Michael, A. Berry, and Gordon Linoff, “Data Mining Techniques for Marketing, Sales, and Customer Support,” *Wisley computer publishing*, 1997.

[5] G. Salton, Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer, Addison-Wesley, Reading, MA, 1988.

[6] Michael W. Berry, Zlatko Drmac, Elizabeth R. Jessup, Matrices, Vector Spaces, and Information Retrieval, *SIAM Review*, Vol. 41, No. 2, pp. 335 – 362.

[7] Daphne Koller and Mehran Sahami, “Hierarchically Classifying Documents Using Very Few Words,” *In Proc. ICML-97*, 1997, pp. 170 – 176.

[8] N.Fuhn, “Models for Retrieval with Probabilistic Indexing,” *Information Processing and Management*, Vol. 25, No. 1, 1989.