지식추출엔진 및 특허출원엔진의 개발을 위한 모형 연구

A Study on the Development Model for Knowledge Portal Site and Automated Patent Application Engine

노 동 조(Dong-Jo Noh)*

초 록

기술정보와 지적재산권정보의 효과적인 입수와 이용은 연구자들의 연구개발 과정에서 발생하는 시행착오나 중복연구를 방지할 뿐만 아니라 산업재산권의 침해를 사전에 예방할 수 있기 때문에 보다 효율적인 연구를 추구하는 사람들에게 있어서는 필수적인 요소이다. 하지만 정보가 폭증하는 지식정보사회에서 연구자들이 정보활동에 쏟는 시간의 과다는 연구시간의 단축을 의미하며 이는 연구의 생산성을 저하시키는 결과를 초래한다. 본 연구는 연구자들의 연구생산성 향상을 위한 하나의 도구로써 연구자들을 대신하여 전자문서의 내용을 자동적으로 분석하여 핵심적인 내용들을 추려서 문제와 해결방안의 형태로 된 지식 데이터베이스를 구축하고 이와 연계된 자동화된 특허출원엔진의 개발에 초점을 맞추었다. 전술한 두 시스템의 개발 가능성과 타당성에 대하여 논의하고 시스템 구축에 필요한 요소와 기술적인 문제들을 모형 개발을 통하여 제시하였다.

ABSTRACT

The purpose of this study is to achieve two goals. One is to construct knowledge database which could read and analyse electronic documents in place of researchers for the purpose of improvement of research productivity, and the other is to develop automated patent application engine which is connected to the knowledge database. This study discusses the possibilities and appropriateness of two systems mentioned earlier, and provides elements necessary to system developments and technical problems through model development.

키워드: 검색엔진, 지식추출엔진, 지적재산권, 특허, 특허출원, 기술정보, 지적재산권정보, 산업재산권, 지식 데이터베이스, 모형개발
knowledge database, patent application, intellectual property right

---

* 혜천대학 정보시스템계열 교수(djnoh@hcc.ac.kr)
논문접수일자 2002년 5월 9일
게재확정일자 2002년 5월 24일

1. 서 론

1.1 연구의 필요성

원저적 이론을 창출하는 과학자나 신기술, 신제품 등을 개발하는 기술자는 모두가 새로운 아이디어를 추구하는 사람들이며, 보다 윤택한 인간의 삶을 위해 필요한 원천정보를 제공하는 사람들이기도 하다. 이들은 특정 연구와 기술개발에 관련된 정보를 얻기 위해서 방대한 양의 기술정보와 지적재산권정보로부터 해당 지식과 기술을 파악해야 할 뿐만 아니라 연구 및 기술개발 활동의 결과로 양질의 정보를 생산해 냄으로써 지식정보사회에서 선도적 역할을 담당하고 있다.

하지만 이들조차도 폭증하는 정보들로 인해서 정보의 선별과 검색, 입수에 많은 어려움을 호소하고 있다. 연구자들이 정보활동에 쏟는 시간의 과다는 상대적으로 연구와 실험, 기술개발에 투여하는 시간의 단축을 의미하며 이는 곧 연구의 생산성을 저하시키는 결과를 초래한다.

전술한 문제들에 대한 해결책중의 하나는 연구자들이 본연의 업무인 연구에 전념할 수 있도록 사회적 장치를 마련하여 주는 것이다. 연구자들을 대신하여 누군가가 그들이 필요로 하는 각종 지식과 정보원들을 점검하고 그곳에 수록된 정보와 중요 내용들을 분석하여 이용하기 편리한 형식으로 가공하여 제공해 주는 것이다.

문제(Problem)와 해결방안(Solution)의 형식으로 정리된 지식 데이터베이스를 통해서 연구자들이 특정 연구와 프로젝트에서 요구되는 지식과 정보를 손쉽게 추출할 수 있으며, 연구활동의 결과로 얻은 지적재산정보에 대한 권리부여는 지식 데이터베이스와 연계된 자동화된 특허출원 소프트웨어를 통하여 이루어 질 수 있다면 가능할 것이다.

본 연구는 이러한 연구자들의 기대와 요구를 하나의 시스템 안에서 해결해 보고자 하는 하나의 시도이다. 즉, 지식추출엔진과 특허출원엔진의 개발을 통하여 전술한 문제들에 대한 현실적 가능성과 타당성에 대하여 예증하고 시스템 구성에 필요한 요소와 기술적인 문제들에 대하여 미리 점검하여 보는 것이다.

본 연구를 통하여 효율적인 지식추출이 가능해 지고 지적재산권정보 및 기술정보의 검색이 보다 원활해지며 나아가 특허출원까지 자동으로 이루어지게 된다면 연구자들의 연구개발 과정에서 발생하는 시행착오나 중복연구를 방지하고 연구시간과 비용을 절약할 수 있으며 나아가 산업재산권의 침해를 사전에 예방할 수 있는 등의 여러 가지 효과를 기대할 수 있을 것이다.

1.2 연구의 목적

본 연구는 현실적으로 제기되는 문제들에 대한 답을 얻기 위한 연구로서 문제해결에 적합한 모형을 개발하는 것이다. 따라서 본 연구의 목적은 모형 개발을 통하여 문제와 해결방안의 형식으로 정리된 시스템의 구축 가능성과 타당성에 대하여 지각하고 시스템의 개발 및 구축과정에서 발생하는 여러 요인과 기술적인 문제들에 대하여 선지하는 것이다. 즉, 연구자들을 대신하여 전자문서의 내용을 자동적으로 읽고 분석하여 그 중에서 핵심적인 내용들을 추려서 문제와 해결방안의 형태로 된 지식 데이터베이스를 구축하고 이와 연계된 자동화된 특허출원 프로

그램을 개발을 통하여 그와 같은 사실을 확인하는 것이다.

1.3 연구의 방법 및 제한점

본 연구의 목적을 달성하기 위하여 동원된 연구방법은 다음과 같다. 첫째, 문헌조사와 데이터베이스 검색을 통하여 관련분야의 정보를 수집하고 시스템 개발의 주요 대상이 되는 기술정보와 지적재산권정보에 대한 현황과 특성을 파악하였다. 이 과정을 통하여 기술정보와 지적재산권정보의 특성상, 새로운 개념들이 자주 등장하며 그에 따른 미등록어의 문제, 생략된 요소들의 복원 및 이들간의 관계 설정이 지식 데이터베이스를 구축하는데 필요하다는 사실을 인지하고 그러한 문제점들이 시스템의 설계에 반영될 수 있는 이론적 근거를 마련하였다. 둘째, 시스템 설계분야의 전문가들과의 브레인스토밍(Brain Storming)을 통하여 시스템의 구축 가능성과 타당성에 대하여 검토하였다. 그 결과, 시스템의 개발에는 개발의 대상이 되는 정보자료의 주제와 형태적 특성 및 시스템을 이용하는 이용자들의 특성들이 반영되어야 한다는 결론을 도출하였다. 셋째, 시스템 개발 및 구축을 전문으로 하는 벤처기업의 협력을 받아 일정 기간동안 특정 주제분야(의학분야)의 기술정보와 특허정보를 대상으로 한 지식 데이터베이스를 구축하였다. 구축된 데이터베이스의 성능을 검토하는 과정에서 시스템 개발에 필요한 요소와 기술적인 문제들이 밝혀졌다.

앞서 1.2 연구의 목적에서 밝힌 바와 같이 이 연구는 현실적으로 제기되는 문제들에 대한 적절한 솔루션을 제공하기 위하여 마련된 모형연구이다. 따라서 모형연구의 특성상, 개발된 모형이 모든 실체를 다 표현해 주지 못할 뿐만 아니라 본 연구를 통하여 제시된 모형의 효과와 실효성에 대해서도 신뢰할 수 없는 부분이 있다. 다만 본 연구를 통하여 기술정보 및 지적재산권정보의 특성을 고려한 시스템 개발의 필요성이 제고되고 그러한 특성들이 반영된 시스템 개발의 타당성과 가능성들에 대한 논의가 진전되며 시스템 구축에 필요한 요소와 기술적인 문제들이 보다 명확하게 밝혀져서 앞으로 발생할 시행착오를 줄일 수 있기를 기대한다.

2. 지식추출엔진 및 특허출원엔진을 위한 모형 개발

2.1 모형 개발의 목적

연구자 및 기술자들의 연구생산성 향상을 위한 하나의 도구로서 특정주제와 관계된 지식과 기술을 분석하여 특정 문제의 해결에 필요한 핵심적인 내용을 추려서 문제와 해결방안의 형태로 지식 데이터베이스를 구축하고 이 데이터베이스와 연계된 특허출원 프로그램을 개발할 수 있다면 기술개발을 위한 사전 정보검색 및 연구결과에 대한 권리부여까지 한 시스템에서 해결할 수 있게 된다.

따라서 모형 개발은 다음의 두가지 부분에 초점을 맞추고 있다. 하나는 전자화된 한국어 문서들에 대하여 지식 추출이 가능한 검색엔진을 개발하는 것이고 다른 하나는 그렇게 해서 개발된 검색엔진을 활용하여 자동화된 특허출원 소프트웨어를 개발하는 것이다. 즉, 특정 전문분야에

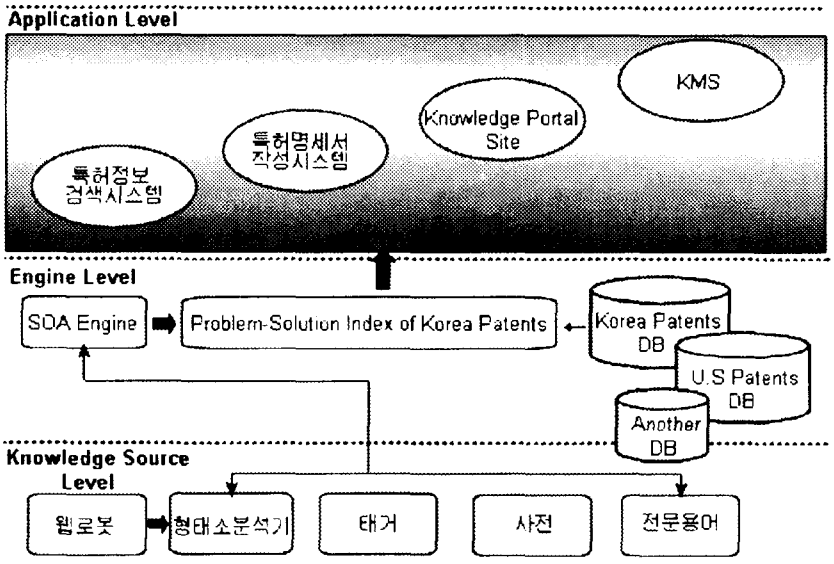
해당되는 전자화된 한국어 문서들을 분석하여 인과관계를 추출하는 기본엔진과 추출된 인과관계를 바탕으로 검색과 브라우징(Browsing)을 할 수 있는 인터페이스를 제공하고 사용자에게 최적화된 특허명세서 작성흐름도(Workflow)를 제공하기 위한 도구(Tool)를 개발하는 것이다(그림 1 참조).

2.2 지식추출의 기본 개념 및 과정

본 시스템에서의 지식이란 전자문서를 통하여 얻을 수 있는 객체간의 관계(규칙)를 말한다. 구체적으로 지식은 한 문장내에서 서술어를 중심으로 하는 주어와 목적어의 관계(한글의 경우는 주어 + 목적어 + 서술어의 순서이며, 영어의 경우는 주어 + 동사 + 목적어의 순서이다)나 ~에 의한, ~를 통한, ~를 사용한 등과 같은 특정 어구를 중심으로 연결되는 두 개의 구 사이의 관계들이 된다.

'문제(Problem)'를 문장간의 또는 한 문장내에서의 목적어 + 서술어에 해당되는 어절이라고 정의할 때, 그 문제에 대한 '해결책(Solution)'은 문제와 인과관계에 있는 주어에 해당되는 단어나 어구들이라고 할 수 있다. 따라서 추출된 지식(주어 + 목적어 + 서술어 관계)에서 어의 색인(Semantic Indexing)을 통하여 문제에 대한 해결책을 제시할 수 있게 되는 것이다.

지식의 추출과 시스템의 구축 과정을 단계적으로 기술하면 다음과 같다. 제 1단계에서는 전자화된 문서를 어절 단위로 분리하여 위치 정보를 구하고 각 문장에 대하여 형태소분석을 하여 주어와 서술어를 구별하고 파싱을 통해서 분석



* KMS(Knowledge Management System)
* SOA Engine(Subject-Object-Action Engine)

그림 1. 전체 시스템 구성도

정보를 구한다. 제 2단계에서는 앞선 단계에서 얻어진 어절의 위치 정보를 숫자 형태의 포인터로 변환하여 저장한다. 제 3단계에서는 상기 어절의 위치 정보와 파싱 정보를 이용하여 어절의 형태소분석 정보의 중의성을 해소하는 단계이며 제 4단계에서는 문장에서 주어(Subject) - 목적어(Object) - 서술어(Action)의 형태로 의미있는 지식을 분석하여 추출한다. 제 5단계에서는 추출된 단어에 대한 목적어, 서술어와 주어간에 존재하는 원인(문제)과 결과(해결) 관계를 바탕으로 하는 데이터베이스를 구축하는 단계이며, 마지막 6단계에서는 단어들간의 원인(문제)와 결과(해결)의 관계를 분석하여 문서의 요약문을 작성한다(그림 2 참조).

2.3 시스템의 구성요소 및 기술

본 시스템을 개발하는 궁극적인 목표가 한국어 문서들에서의 인과관계 추출과 이를 기반으로 한 데이터베이스 구축과 특허출원 자동 소프트웨어 개발이므로 전술한 내용들을 실제로 구현하려면 다음과 같은 요소와 기술들이 수반되어야 한다.

(1) 형태소분석기

효용성이 검증된 형태소분석기(예를 들면, 한국과학기술원 전문용어센터(KORTERM)에서 개발한 형태소분석기인 '한나눔')를 사용하여 전문분야의 문서들을 대상으로 한 형태소 해석기 사전의 적용과 미등록어 추정 같은 부분들에 대한 튜닝(Tuning)이 반드시 필요하다. 특허분야의 문서는 그 성격상 필연적으로 새로운 아이디어들을 포함하게 됨에 따라 제품명이나 세부기술에 있어서 새로운 용어들이 자주 등장하기 때문에 분석에 실패할 확률이 높다. 따라서 특허분야 문서들을 대상으로 전문용어사전

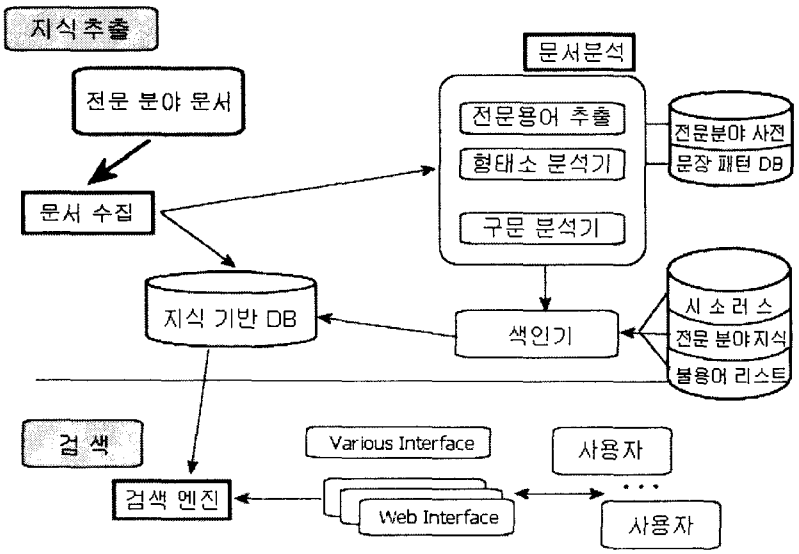


그림 2. 세부 컴포넌트 구성도

을 구축하는 일과 새로운 문서들을 분석할때 자동적으로 전문용어를 추출하고 이를 효과적으로 제어할 수 있어야 한다. 이를 위해서는 첫째, 전문분야의 사전 구축 둘째, 전문용어 추출기와 형태소분석기의 결합 셋째, 특허분야 문서에 필요한 미등록어 추정 알고리즘 개발 등이 필요하다.

(2) 구문분석기

효용성이 검증된 구문분석기를 기반으로 한 코퍼스(Corpus)에 대한 튜닝이 필요하다. 실용적인 수준의 구문분석기는 비문법적인 문장들이 입력되어도 적절한 구조를 출력해 줄 수 있어야 하며 한 문장내에서 반복으로 인해 생략된 요소들에 대한 복원과 생략된 요소들간의 관계를 밝혀서 문장내에서 찾아 줄 수 있어야 한다. 이를 위해서는 첫째, 주어 - 목적어 - 서술어 관계의 추출용 구문분석기의 튜닝 둘째, 분석이 힘든 표현들에 대한 패턴처리기의 개발 셋째, 문장 탐색을 통한 생략 복원용 프로그램의 개발 등이 전제되어야 한다.

(3) 지식기반 인덱싱

앞서 형태소분석과 구문분석의 결과를 토대로 색인의 대상이 되는 지식들을 추출하여야 한다. 일차적으로 주어 - 목적어 - 서술어 관계를 추출한 다음, 문서에서 수집된 패턴을 바탕으로 특정어구(~에 의한, ~를 통한, ~를 사용한 등)를 중심으로 연결되는 어구들을 추출하여 문장내에서 '문제 - 해답'의 형태로 관계를 얻어내야 한다. 이를 위해서는 첫째, 해당 문서의 수집 및 전처리기의 개발 둘째, 지식관계의 추출 및 인덱싱 셋째, 불용어처리 리스트의 구축 등이 전제되어야 한다.

(4) 검색 및 지원도구

검색엔진과 전문용어사전, 불용어 리스트,

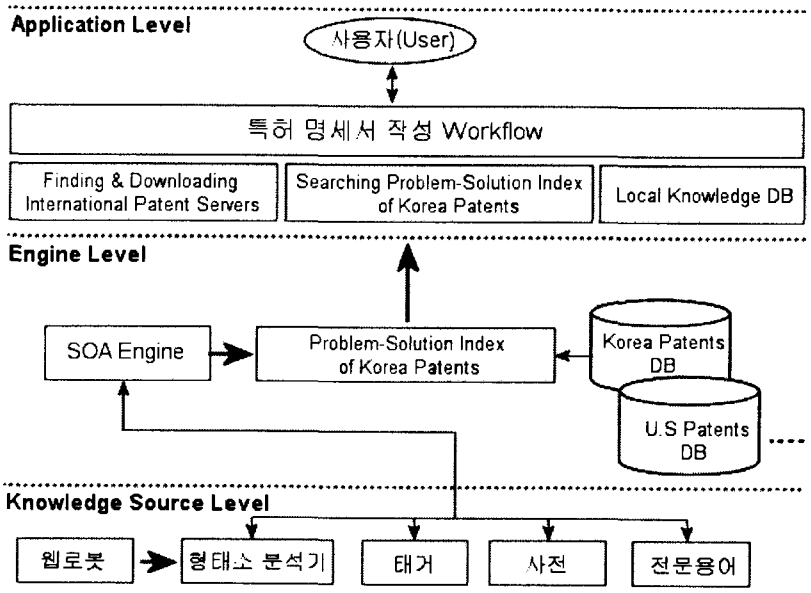


그림 3. 특허명세서 작성 시스템 구성도

어휘 패턴들을 관리하기 위한 도구들이 마련되어야 한다. 국내외 특허전문 검색사이트를 참고하여 인터페이스를 설계하고 한국어 특성에 맞는 검색 환경과 브라우징을 제공한다. 이를 위해서는 첫째, 자연언어 질의 처리기 둘째, 자원관리도구의 제작 셋째, 검색 인터페이스의 제작 등이 전제되어야 한다.

(5) 특허출원 명세서 작성 흐름도

특허출원명세서를 작성할 수 있는 흐름도를 제공한다. 이를 위해서는 특허출원명세서 작성 소프트웨어가 제작되어야 한다(그림 3 참조).

결론적으로 전자화된 문서들에 대한 지식추출엔진과 특허출원엔진을 개발하기 위해서는 시스템 구성과 개발 단계에 따른 다음과 같은 구성 요소와 기술들이 필요하다. 첫째, 문서의 수집 및 분석단계에서는 형태소분석기, 구문분석기, 자원관리기, 전문분야의 사전, 어휘지식패턴, 시소러스, 전문분야의 지식, 문서의 출처 및 전체 문서의 리스트가 필요하다. 둘째, 색인기에서는 색인어 추출기, 상위개념의 색인기, 검색용 데이터베이스 구축기, 지식기반 인덱싱 데이터베이스, 색인용 구문패턴기, 불용어리스트가 필요하다. 셋째, 검색기에서는 자연어 질의처리기, 사용자 인터페이스, 원문전체 및 출처 정보가 필요하다(그림 4 참조).

2.4 향후 과제

정보시스템의 개발에는 다음의 두가지 요인들이 고려되어야 한다. 하나는 개발의 대상이 되는 정보자료의 주제와 형태적 특성이 반영되어야 하며 다른 하나는 시스템을 이용하는 이용자들의 특성이 반영되어야 한다. 하지만 본 연구에서는 기술정보 및 지적재산권 정보의 특성을 고려한 시스템의 개발에 초점을 맞추었기 때문에 기술정보와 지적재산권 정보를 이용하는 이용자들의 정보요구나 이용특성 등은 시스템의

한국어문서의 지식추출엔진	문서수집 및 분석	형태소 분석기 구문분석기 자원 관리기
		전문분야 사전 어휘 지식 패턴 시소러스 분야지식 문서 출처 및 전체 문서 리스트
	색인기	색인어 추출기 상위개념 색인기 검색용 DB 구축기
		지식 기반 인덱싱 DB 색인용 구문 패턴 불용어 리스트
	검색기	자연어 질의 처리기 사용자 인터페이스
		원문 전체 및 출처 정보

그림 4. 지식추출엔진의 구성 요소 및 기술

개발에 반영되지 못하였다. 또, 일정 기간동안 특정 주제분야의 지식과 정보만을 대상으로 한 결과이기 때문에 본 연구를 통하여 밝혀진 사실이나 제시된 모형도 이론상으로는 완전할지 모르나 운용에 있어서는 기술적인 문제들이 엄연히 존재할 것이다.

앞으로의 과제는 본 연구에서는 고려하지 못한 기술정보와 지적재산권정보 이용자들의 특성까지도 반영한 시스템을 구축하는 것이며 그 과정에서 발생하는 기술적인 문제들과 고려 요인들을 명확히 밝혀내는 것이다. 시스템의 적용도 단계별 접근이 바람직하다. 특정 주제분야나 특정 자료의 형태로 제한하여 시스템을 시험개발하고 그 효용성이 입증되면 점차 그 범위를 넓혀 나가는 것이 바람직한 수순일 것이다. 실용화는 시스템의 성능이 객관적으로 인정할 수 있는 수준에 도달했을 때의 논의사항이다.

본 연구를 통하여 지식기반사회에서 기술정보 및 지적재산권정보의 중요성을 새롭게 인식하는 계기가 되고 후속연구를 통하여 기술정보 및 지적재산권 정보의 특성과 이용자들의 특성이 모두 반영된 시스템의 설계가 이루어지고 그 과정에서 발생하는 문제와 기술적인 요인들이 보다 정밀하게 밝혀지기를 기대해 본다.

3. 결론 및 제언

기술정보와 지적재산권정보의 효과적인 입수와 처리, 이용은 연구자들의 연구개발과정에서 발생하는 시행착오나 중복연구를 방지할 뿐만 아니라 연구시간과 비용을 절약하며 나아가 산업재산권의 침해를 사전에 예방할 수 있기 때문에 보다 효율적인 연구를 추구하는 연구자들에게 있어서는 필수적인 요소이다. 연구자들이 본연의 연구활동에 전념하기 위해서는 누군가가 이들을 대신하여 전자문서의 내용을 자동적으로 읽고 분석하여 핵심적인 내용들을 추려서 문제와 해결방안의 형태로 구성된 지식 데이터베이스를 구축하여 제공하는 것이 바람직하다. 본 연구는 이러한 연구자들의 기대에 대한 하나의 부응으로서 지식추출엔진과 이와 연계된 자동화된 특허출원엔진의 개발 가능성과 타당성에 대하여 논의하고 프로그램 개발에 필요한 요소와 기술적인 문제들을 모형 개발을 통하여 제시하였다.

본 연구를 통해서 얻어진 기본 개념들을 정리하면 다음과 같다. 첫째, 지식은 객체간의 관계이며 한 문장내에서 서술어를 중심으로 하는 주어와 목적어의 관계이거나 ~에 의한, ~를 통한, ~를 사용한 등과 같은 특정어구를 중심으로 연결되는 두 개의 구 사이의 관계가 된다. 둘째, 문제(Problem)를 문장간 또는 한 문장내에서 목적어 + 서술어에 해당되는 어절이라고 정의하면 그 문제에 대한 해결책(Solution)은 문제와 인과관계에 있는 주어에 해당되는 단어나 어구가 된다. 셋째, 주어 + 목적어 + 서술어 관계에서 추출된 지식은 어의 색인(Semantic Indexing)을 통해서 문제에 대한 해결책을 제시할 수 있다.

전자화된 문서들에 대한 지식추출엔진과 특허출원 프로그램을 개발하기 위해서는 다음과 같은 요소와 기술들이 필요하다. 첫째, 문서의 수집 및 분석단계에서는 형태소분석기, 구문분석기, 자원관리기, 전문분야의 사전, 어휘지식패턴, 시소러스, 전문분야의 지식, 문서의 출처 및 전체문서의 리스트가 필요하다. 둘째, 색인

기에서는 색인어 추출기, 상위개념의 색인기, 검색용 데이터베이스 구축기, 지식기반 인덱싱 데이터베이스, 색인용 구문패턴기, 불용어리스트가 필요하다. 셋째, 검색기에서는 자연어 질의처리기, 사용자 인터페이스, 원문전체 및 출처 정보가 필요하다.

지식추출엔진과 특허출원엔진의 효율성을 제고하기 위해서는 특허문서의 특성들이 시스템의 개발 및 구축과정에서 반드시 반영되어야 한다. 즉, 새로운 개념들이 빈번히 등장함에 따른 미등록어의 문제, 생략된 요소들의 복원 및 관계 설정 등에 대한 조정작업이 필요하다.

향후 연구과제는 본 연구를 통하여 밝혀진 사실과 제안된 요소들을 바탕으로 실용가능한 지식 데이터베이스 및 특허출원엔진을 구축한 다음, 시스템의 성능을 검토하고 그 과정에서 발생하는 문제점들을 최소화하여 실제로 적용가능한 시스템을 완성하는 일이다. 이 과정에서 기술정보와 지적재산권정보를 이용하는 이용자들의 특성과 그들의 정보요구, 이용행태 등이 반영되어야 한다. 전술한 요소들이 모두 반영된 시스템이 완성되면 기술정보 및 지적재산권정보의 입수와 활용이 보다 용이해 질 뿐만 아니라 연구자들이 개발한 이론과 기술에 대한 권리부여까지도 이 시스템 안에서 모두 해결되므로 연구자들의 연구 생산성을 향상시키고 연구의욕을 고취시켜 국가경쟁력의 확보에도 일익을 담당하게 될 것이다.

참 고 문 헌

노동조. 2002. 지식추출 엔진을 위한 모형 개발 연구. 『한국비블리아 발표논집』, 6:89-97.

노동조, 이상렬. 2002. 『지식정보시대의 디지털 정보검색론』. 서울: 에듀컨텐츠.

노동조, 이상렬. 2002. 『ASP 웹 데이터베이스 프로그래밍』. 서울: 한올출판사.

송도규. 1996. 『인지언어학과 자연언어 자동처리』. 서울: 홍릉과학출판사.

이종언, 윤영희, 장인정. 2001. 『산업기술정보-검색과 활용』. 서울: 한국산업기술진흥협회.

정교민. 2000. 『특허분석과 기술가치』. 서울: 한울출판사.

최기선 외. 1996. 『전문용어연구1』. 서울: 홍릉과학출판사.

황도삼 외. 1996. 『자연어처리』. 서울: 홍릉과학출판사.

한국어 정보처리연구소. 1999. 『C로 구현한 한글코드시스템 프로그래밍 가이드』. 서울: 골드.

한국어 정보처리연구소. 1999. 『C로 구현한 인터넷 데이터베이스시스템』. 서울: 골드.

한국어 정보처리연구소. 1999. 『C로 구현한 인터넷 정보검색시스템』. 서울: 골드.

http://kita.technet.or.kr

http://www.kipo.go.kr

http://www.kipris.or.kr

http://www.invention-machine.com