

문서의 의미론적 분석에 기반한 키워드 추출에 관한 연구
(A Study on Keywords Extraction based on Semantic Analysis of Document)

송민규^a, 배일주^b, 이수홍^c, 박지형^d

^{a,b,c} 연세대학교 기계공학부
서울 서대문구 신촌동 134, 120-749
Tel: +82-2-2123-2823, Fax: +82-2-312-2159, E-mail: blacklism@cadcam.yonsei.ac.kr,
iljusado@cadcam.yonsei.ac.kr, shlee@yonsei.ac.kr

^d 한국과학기술연구원 지능인터랙션 연구센터
서울 성북구 월송길 5 하월곡동 39-1, 136-791
Tel: +82-2-958-5631, Fax: +82-2-958-5649, E-mail: jhpark@kist.re.kr

Min-Kyu Song^a, Il-Ju Bae^b, Soo-Hong Lee^c, Ji-Hyung Park^d

^{a,b,c} School of Mechanical Engineering, Yonsei University
134, Shinchon-Dong, Seodaemun-Gu, Seoul, 120-749, Korea
Tel: +82-2-2123-2823, Fax: +82-2-312-2159, E-mail: blacklism@cadcam.yonsei.ac.kr,
iljusado@cadcam.yonsei.ac.kr, shlee@yonsei.ac.kr

^d Intelligence and Interaction Research Center, Korea Institute of Science and Technology
39-1, Hawolgok-Dong, Wolsong-Gil 5, Seongbuk-Gu, Seoul, 136-791, Korea
Tel: +82-2-958-5631, Fax: +82-2-958-5649, E-mail: jhpark@kist.re.kr

Abstract

지식 관리 시스템, 정보 검색 시스템, 그리고 전자 도서관 시스템 등의 문서를 다루는 시스템에서는 문서의 구조화 및 문서의 저장이 필요하다. 문서에 담겨있는 정보를 추출하기 위해 가장 우선시되어야 하는 것은 키워드의 선별이다. 기존 연구에서 가장 널리 사용된 알고리즘은 단어의 사용 빈도를 체크하는 TF(Term Frequency)와 IDF(Inverted Document Frequency)를 활용하는 TF-IDF 방법이다. 그러나 TF-IDF 방법은 문서의 의미를 반영하지 못하는 한계가 존재한다. 이를 보완하기 위하여 본 연구에서는 세 가지 방법을 활용한다. 첫 번째는 문헌 속에서의 단어의 위치 및 서론, 결론 등의 특정 부분에 사용된 단어의 활용도를 체크하는 문헌구조적 기법이고, 두 번째는 강조 표현, 비교 표현 등의 특정 사용 문구를 통제 어휘로 지정하여 활용하는 방법이다. 마지막으로 어휘의 사전적 의미를 분석하여 이를 메타데이터로 활용하는 방법인 언어학적 기법이 해당된다. 이를 통하여 키워드 추출 과정에서 문서의 의미 분석도 수행하여

키워드 추출의 효율을 높일 수 있다.

Keywords:

Keyword Extraction (키워드 추출), Analysis Lexical Meaning (의미론적 분석)

1. 서론

지능형 회의 공간을 실현하기 위해서는 컴퓨터로 구현되는 사이버 공간과 인간의 실제 생활 공간을 구별되지 않게 융합시켜, 인간과 인간, 인간과 사물, 인간과 정보 환경과의 커뮤니케이션의 한계를 극복할 수 있는 요소 기술의 개발을 필요로 한다.

이를 위하여 초기 연구에서는 회의 도메인을 분석하고, 도메인 내에 존재하는 개념을 규명하였으며, 이를 상위, 하위 온톨로지와 각 온톨로지의 클래스와 속성, 관계를 정의하였다. 그러나 이 온톨로지는 정적인 상태로 존재하며 변하는 상황에 대응하기에는 부족하다. 인간과 컴퓨터 간의 커뮤니케이션이 활발하게 이루어지기

위해서는 변하는 주변 환경과 정보에 따라 동적인 변화가 가능해야 하므로 온톨로지가 주변 상황과 지속적인 상호 응답을 하면서 보완되고 확장될 수 있어야 한다. 이를 위해 반자동적인 온톨로지 수정, 확장에 대한 알고리즘과 모듈을 개발하였다. [1]

구축된 온톨로지를 바탕으로 ‘회의 문서 처리’와 ‘실시간 회의 기록 처리’ 서비스를 제공하는 것이 지능형 회의 공간 구현의 목적이라고 할 수 있다. 그러나 회의 문서를 처리하는 서비스 구현을 위해서는 문서 처리를 통하여 온톨로지를 자동으로 생성할 수 있는 기능이 필요한데, 구조화되어 있지 않은 문서를 바탕으로 온톨로지를 자동으로 생성하는 것은 쉽지 않다. 온톨로지 구축 과정을 자동화하는 것은 그 대상이 되는 문서가 모두 일정한 틀에 의하여 구조화되어 있어야 한다. 그러나 모든 문서를 구조화하기에는 어려움이 있기 때문에, 현실적으로 완전 자동화는 불가능하다. 따라서 사람이 직접 모델링 작업을 수행하고, 이를 학습 알고리즘을 통하여 반자동화하는 연구가 수행되고 있다. 반자동화를 위해서는 문서가 구조화되거나, 또는 준구조화 되면 된다.

본 연구에서는 반자동화를 위하여 문서의 준구조화를 수행하고, 이를 바탕으로 키워드를 추출하고자 한다. 문서로부터 정보 추출의 자동화 및 온톨로지를 자동으로 생성하기 위해서는 키워드 추출이 필수적이다. 이를 위하여 통계적 기법, 문헌구조적 기법, 언어학적 기법, 통계 어휘 활용 방법을 병행하여 사용하는 방법론을 제안하고, 이를 수행할 수 있는 시스템을 구현하고자 한다.

2. 관련 연구

2.1 키워드 추출

키워드 추출 프로세스는 크게 네 가지 단계로 구분할 수 있다. 첫 번째 단계는 키워드를 추출하기에 앞서 문서를 처리하는 선처리 단계이다. 이 단계에서는 문서를 지정된 포맷으로 준구조화하거나, 또는 형태소 분석기를 통하여 구성성분을 분해하는 작업을 수행한다. 두 번째 단계로는 후보군을 선정하는 단계이다. 앞선 단계에서 구분된 각 단어들 중 키워드가 될 가능성이 있는 단어들만 추려내는 작업이 이에 해당된다. 숫자 및 특수 문자, 조사 등은 키워드가 될 가능성이 극히 낮기 때문에 제외되고, 보통 명사 형태의 단어가 후보군으로 선정된다. 또한, 키워드 추출 알고리즘에서 가장 많이 사용되는 TF(Term Frequency), IDF(Inverted Document Frequency) 등도 이 단계에서 활용된다. TF는 단어의 출현 빈도를 나타내는 지표로, 문서에서 많이 출현하는 단어일수록 그 중요성이 높다고 사료되어 키워드일

가능성이 높기 때문에 측정되는 지표이다. 그러나 TF 지표에도 한계가 있다. 조사나 일반 동사 등과 같이 거의 모든 문서에서 일반적으로 사용되는 단어들은 그 출현 빈도가 높을 수 밖에 없기 때문이다. 그러나 출현 빈도에 비해서 해당 문서에서 그러한 단어들의 중요성이 높다고 할 수 없다. 따라서 이를 보완할 수 있는 지표가 IDF이다. IDF는 단어가 다른 문서에서도 얼마만큼 폭넓게 사용되고 있는가를 측정하는 지표이다. 즉, 일반적으로 널리 사용되는 단어가 아닌, 특정 문서에서만 나타나는 단어가 해당 문서의 키워드일 가능성이 높다는 점에서 측정하는 지표이다. 이러한 지표들을 측정하여 후보군을 선정하게 된다. 세 번째 단계는 선정된 후보군에 대하여 키워드 지표를 측정하는 알고리즘을 적용하는 단계이다. 앞서 사용한 TF와 IDF를 사용하기도 하며, 그 외에 각 연구에서 새롭게 제안하는 방법을 사용하기도 한다. 마지막으로 이 단계를 거쳐 키워드 지표를 적용하고, 키워드를 추출하는 단계가 있다.

Yasin Uzun은 핵심 알고리즘으로 특정 분류에 속할 확률을 산출하는 방법인 Naïve Bayes를 활용하였다. 그는 TF와 IDF를 산출한 TF x IDF Score 외에도 이전에 함께 사용되었던 단어와의 거리, 전체적인 텍스트 내에서의 단어의 위치, 문장 내에서의 단어의 위치 등을 지표로 하여 Naïve Bayes 방법을 적용하였다. [2]

Y. Matsuo 등은 단일 문서 내에서 키워드 추출이 가능한 방법을 제안하였다. 기존의 연구에서는 기준 문서를 선정하고, 이 기준 문서와의 비교를 통하여 IDF 지표를 측정하였는데 반해, 이 연구는 기준 문서가 필요하지 않다는 장점을 가지고 있다. 기본적으로 단어들의 TF 값을 측정하여 후보군을 선정하고, 해당 단어들이 함께 사용되는 횟수를 측정하게 된다. 이를 통하여 단어로서뿐만 아니라 구문으로 추출할 수 있다는 장점도 가지고 있다. [3]

마지막으로 Taeho Jo 등은 신경망 모델을 활용하여 키워드를 추출하였다. 이 연구에서는 TF, IDF와 함께 ITF(Inverted Term Frequency), T(Title), FS(First Sentence), LS>Last Sentence)를 지표로 하여 신경망 모델을 적용하였다. TF, IDF, ITF는 그 수치로 표현하였고, T, FS, 그리고 LS는 해당되는지의 여부를 0과 1로써 표현하였다. [4]

3. 의미분석 기반의 키워드 추출

현재까지 이루어지고 있는 연구들은 단어의 사용 빈도 및 위치 등의 형태에 의존한 방법만을 사용하고 있기 때문에, 문서의 의미를 반영하는 점은 미흡하다고 할 수 있다. 이를 보완하여 기존의 형태 분석 기반의 장점은 살리고, 의미 분석 기반의

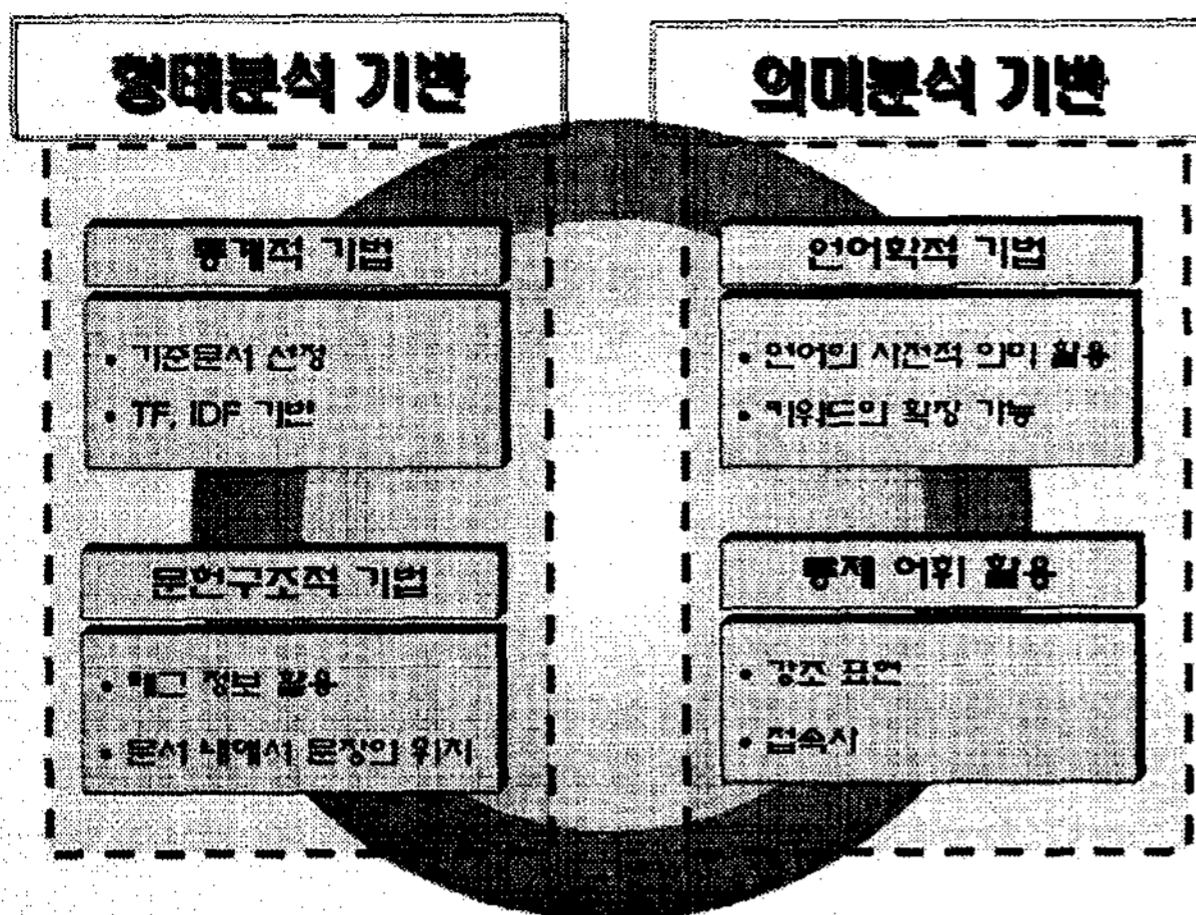


Figure 1 - 방법론

방법을 모색하고자 하였다.

이를 위해서는 키워드일 가능성이 높은 단어의 특징을 정리하는 것이 선행되어야 할 것이다. Kageura와 Umino는 이를 다섯 가지로 요약하였다. 첫 번째 특징은 “문서에서 나타나는 단어”이고, 두 번째 특징은 “문서에서 빈번하게 나타나는 단어”이다. 세 번째 특징은 “제한된 수의 문서에서만 나타나는 단어”이고, 네 번째 특징은 “전체 문서에서보다 문서에서 상대적으로 빈번하게 사용되는 단어”이다. 마지막으로 “데이터베이스에서 특정 분포 특성을 나타내는 단어”이다. [5] 이 특징은 이전 연구와 마찬가지로 문서 및 단어의 형태 분석에 기반하였다. 본 연구는 문서의 의미 분석을 기반으로 하므로 “구문에 부가적인 의미를 부여하는 특정 단어와 함께 사용되는 단어”를 여섯 번째 특성으로 추가하였다. 이는 강조 표현, 접속사, 부정 표현 등 특정 표현과 함께 사용되는 단어는 그렇지 않은 단어보다 중요성이 더해진다는 점에 기인하였다. 이러한 점을 바탕으로 Figure 1과 같이 네 가지 방법을 혼용한 방법론을 제안한다.

3.1 통계적 기법

통계적 기법은 크게 세 가지 단계로 나눌 수 있다. 첫 번째 단계는 선처리 단계이다. 이 단계에서는 다양한 분야와 주제를 포괄할 수 있는 기준 문서를 선정한다. 또한, 문서에서 숫자 및 특수 문자를 제거한다. 그리고 경계 기호 및 Stop-Words로 구분하는 작업도 이 단계에서 수행된다. 이러한 작업은 형태소 분석기를 활용하여 수행되는데, 이 연구에서는 포항공대 지능형 소프트웨어 연구실에서 개발한 형태소 분석기 POSTAG/Sejong 모듈을 활용한다. [6]

두 번째 단계는 알고리즘을 적용하는 단계이다. 이 단계에서는 기존 연구에서 수행하던 TF와 IDF를

측정한다. 이를 바탕으로 TF x IDF Score 지표를 산출한다. 이는 단어의 사용 빈도가 높을수록, 그리고 특정 문서에서만 사용될수록 키워드일 가능성이 높다는 특성을 이용한다.

마지막 단계로는 보정하는 단계이다. 문서의 길이가 길거나, 문장의 길이가 길면 사용하는 단어의 수가 많아진다. 이는 자연스레 단어의 TF 값이 커지는 결과를 낳게 된다. 또한, 문서의 길이에 따라서 IDF 값 역시 달라질 수 밖에 없다. 이러한 점을 보완하기 위하여 문서의 길이에 따른 보정값을 적용할 필요가 있다. 통계적 기법에서는 최종적으로 보정된 TF x IDF Score 지표를 산출한다.

3.2 문헌구조적 기법

문헌구조적 기법에서는 기본적으로 문서의 태그 정보를 활용한다. 문서의 형태에 따라 다르지만 문서가 제목, 소제목, 본문 등으로 구성되어 있다면, 본문에서 등장하는 단어보다 제목, 소제목에서 사용된 단어가 키워드일 가능성이 높다. 따라서, 태그 정보를 활용하여 해당 단어가 제목에서 쓰였는지, 소제목에서 쓰였는지를 판단하면 키워드로써 결정하기 쉬워진다. 이를 위해서는 먼저 문서를 준구조화할 필요가 있다. Figure 2와 같이 문서를 XML 포맷으로 변환하여 사용한다면 문헌구조적 기법을 쉽게 적용할 수가 있다.

또한, 문헌구조적 기법에서는 문단 및 문장 내에서의 단어의 위치 정보도 활용한다. 글을 쓸 때 따로 단락을 구분 짓지 않더라도 보통 서론, 본론, 결론의 형태로 글을 작성하게 된다. 이러한 점에서 문서, 또는 문장의 처음이나 마지막에 쓰일수록 키워드일 가능성이 높다. 따라서, 위치 정보를 활용하여 해당 위치에 놓였을 때 가중치를 부여함으로써 문서의 특징을 반영할 수가 있다. 이

```
<?xml version="1.0" encoding="UTF-8" standalone="no" ?>
- <HWPML Style="export" SubVersion="7.0.0.0" Version="2.7">
- <BODY>
- <SECTION Id="0">
- <P ColumnBreak="false" PageBreak="false" ParaShape="
- <TEXT CharShape="4">
  <CHAR>시각장애인 재활도우구</CHAR>
</TEXT>
</P>
<P ParaShape="0" Style="0" />
- <P ParaShape="0" Style="0">
- <TEXT CharShape="5">
  <CHAR>(1) 원지팡이</CHAR>
</TEXT>
<TEXT CharShape="5" />
</P>
- <P ParaShape="0" Style="0">
- <TEXT CharShape="0">
  <CHAR>시각장애인이 외출할 때 길을 찾거나 장애물을
  것으로 긴 지팡이로 안내나 지팡이로 점미식 지팡이로 전
  있다.</CHAR>
</TEXT>
```

Figure 2 - XML 포맷의 예

방법 역시 통계적 기법과 마찬가지로 문서 및 문장의 길이에 따라 보정할 필요가 있다.

3.3 언어학적 기법

3.1절과 3.2절에서 언급한 방법은 기존의 연구와 크게 다르지 않은 형태분석 기반의 방법이다. 이를 보완할 수 있는 방법이 이 절에서 제시할 언어학적 기법이다. 언어학적 기법에서는 문서의 의미를 반영하고자 하였다.

예를 들어, 한 대상에 대하여 설명을 하는 문서가 있다고 가정해보자. 문서에서는 대상에 대한 언급보다는 대상의 의미를 나타내는 단어, 대상과 관련이 있는 단어, 대상의 하위 개념을 나타내는 단어들에 주를 이룰 것이다. 이 경우 대상에 대한 설명임에도 불구하고, 대상을 나타내는 단어는 키워드로 선정되지 않을 수가 있다. 이런 경우는 문서의 의미를 파악하지 않고, 형태에만 국한하여 판단하였기 때문이다.

이를 보완할 수 있는 방법으로 단어의 사전적 의미를 반영하는 방법과 시소러스를 활용하는 방법을 생각하였다. 그러나, 현재 활용할 수 있는 시소러스가 존재하지 않고, 이 연구를 위해 시소러스를 새롭게 구축하는 것은 불가능에 가까워, 이 연구에서는 단어의 사전적 의미만을 활용하고자 한다. 언어학적 기법을 적용하기 위해서는 선처리 단계로 사전 정보를 구축하여야 한다. 사전에서 단어를 검색하여, 검색된 의미에서 키워드를 추출한다. 그 후, 도출된 키워드들을 해당 단어의 메타데이터로 저장한다. 추후 언어학적 기법을 적용할 때 이 메타데이터를 활용하게 된다.

문서에서 키워드를 추출한 후 추출된 키워드들과 메타데이터를 매칭하여 비교한다. 그 후 기준치보다 높은 매칭률을 보였을 때는 해당 키워드를 추가하게 된다. 이 기법은 새롭게 키워드를 추출하는 작업이 아니라 다른 방법을 적용한 후 보정 및 확장 단계에서 이루어지는 방법이다.

3.4 통계 어휘 활용

논문 등과 같이 서론, 본론, 결론으로 구분할 수 있는 문서에서는 결론 부분에 해당 문서의 핵심적인 내용을 작성하게 된다. 이처럼 특정 단어 뒤에 오는 단어들은 그 중요성이 다른 단어보다 크다고 할 수 있다. 그 중요성을 결정지어줄 수 있는 단어를 이 연구에서는 통계 어휘라 지칭한다.

통계 어휘에는 해당 문장에 중요도를 부여하는 강조 표현, 접속사, 부정 표현 등이 포함된다. 강조 표현에는 다양한 형태가 있다. 특정 단어만을

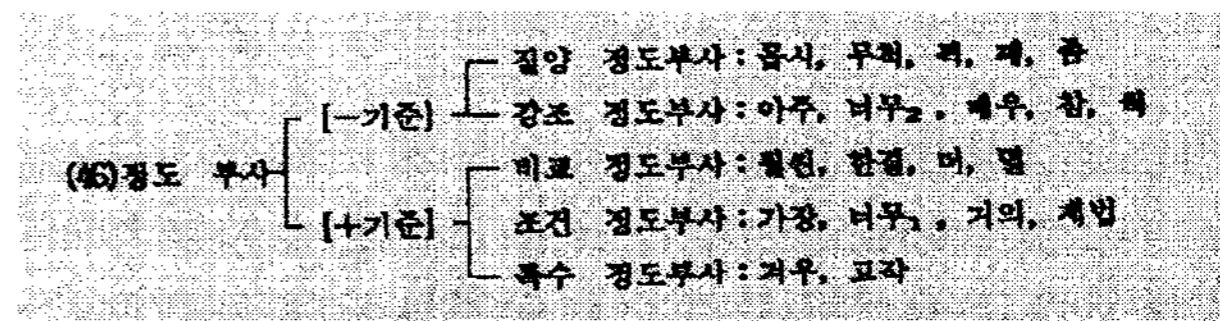


Figure 3- 강조 표현을 하는 정도 부사 [7]

강조하는 표현이 있고, 문장 전체를 강조하는 표현이 있다. 또한, '무척', '매우' 등과 같이 극대의 의미를 가지는 강조 표현이 있고, '겨우', '고작'과 같이 극소의 의미를 가지는 강조 표현이 있다.

강조 표현의 종류도 다양하다. 월 부사, 화식 부사, 술부 부사, 정도 부사 등에 의한 강조 표현이 있고, 접두사나 접미사에 의한 강조 표현의 형태가 있다. 이 연구에서는 문학 작품을 대상으로 하는 것이 아니라, 기술 문서를 주 대상으로 하고 있다. 따라서, 기술문서에서 쉽게 찾아볼 수 없는 접두사나 접미사에 의한 강조 표현 등은 제외시키고자 한다. Figure 3에 나타나 있는 정도 부사에 의한 강조 표현 정도를 반영하고자 한다.

4. 시스템 구현

본 연구에서 구현 중인 시스템은 Figure 4와 같은 아키텍처를 갖는다. 메인 시스템은 지식 관리 시스템이다. 이 시스템에서는 형태소 분석기와 문서 처리기와의 연동을 통하여 후보군을 추출하는 역할을 수행한다. 또한, 사전 분석기와 연동을 통하여 사전 정보를 활용하고 키워드를 추출하는 역할을 수행한다. 이 때 적용하는 알고리즘은 알고리즘 라이브러리 모듈과의 연동을 통하여 수행한다.

최초 키워드를 추출할 문서는 준구조화 작업을 통하여 구성 성분으로 분해하고, 또한 형태소 분석기를 통하여 형태소 분석을 수행한다. 문서의

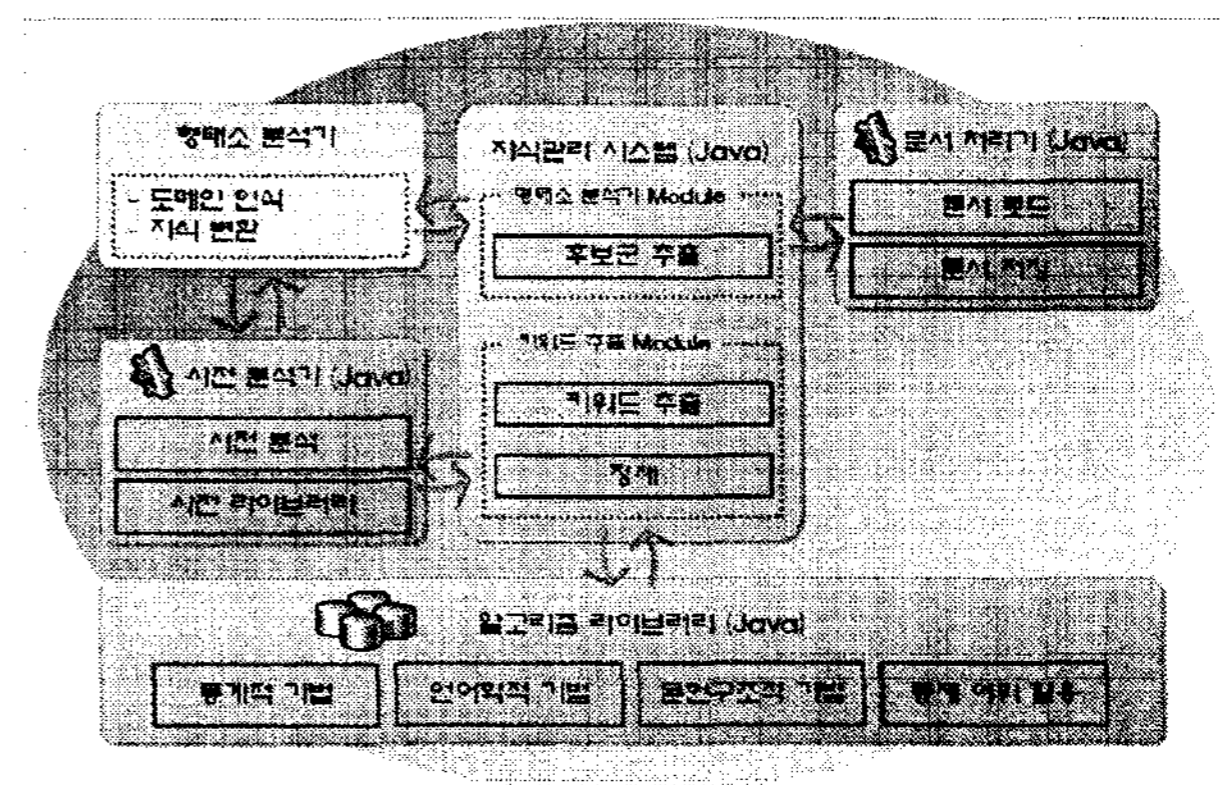


Figure 4 - 시스템 아키텍처

준구조화 작업은 문서 처리기에서 담당한다. 이 모듈은 문서를 읽어 이를 XML 포맷으로 변환하는 역할을 수행하는데, 구현된 문서 처리기는 Figure 5와 같다. Figure 6은 이 문서 처리기를 통하여 문서를 XML 포맷으로 변환한 예이다.

또한, 형태소 분석기는 앞서 언급했듯이 기존 연구에서 개발된 모듈을 사용하며, 그 구동 예는 Figure 7과 같다. 형태소 분석기는 지식 관리 시스템, 그리고 사전 분석기와 연동을 하여 역할을 수행한다.

형태소 분석이 완료되면 그 다음으로는 핵심어 후보군을 선정하게 된다. 핵심어 후보군은 TF의 값에 따라 선출되며, 이는 지식 관리 시스템에서 담당하게 된다. 지식 관리 시스템에서는 이 외에도

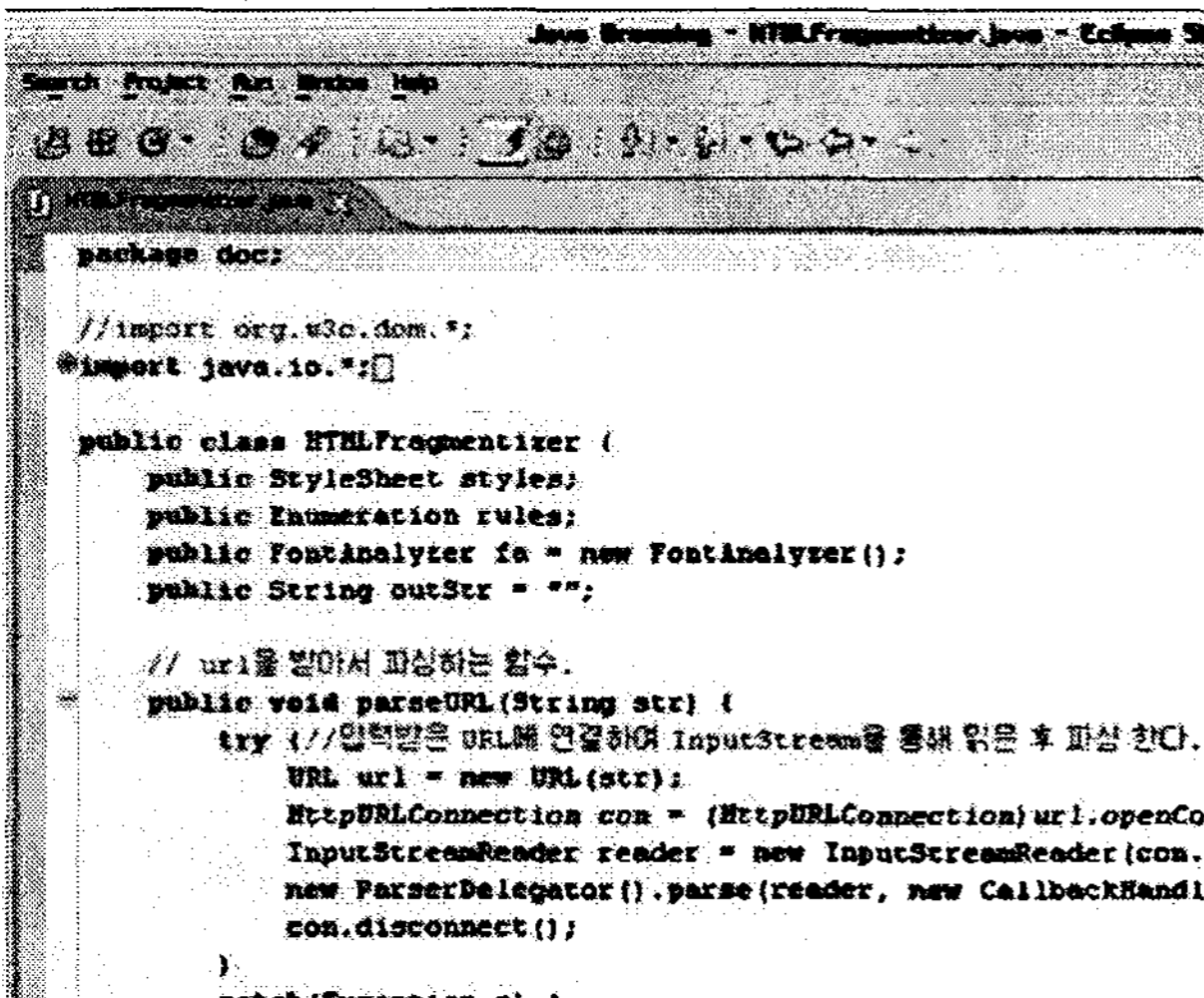


Figure 5 - 문서 처리기

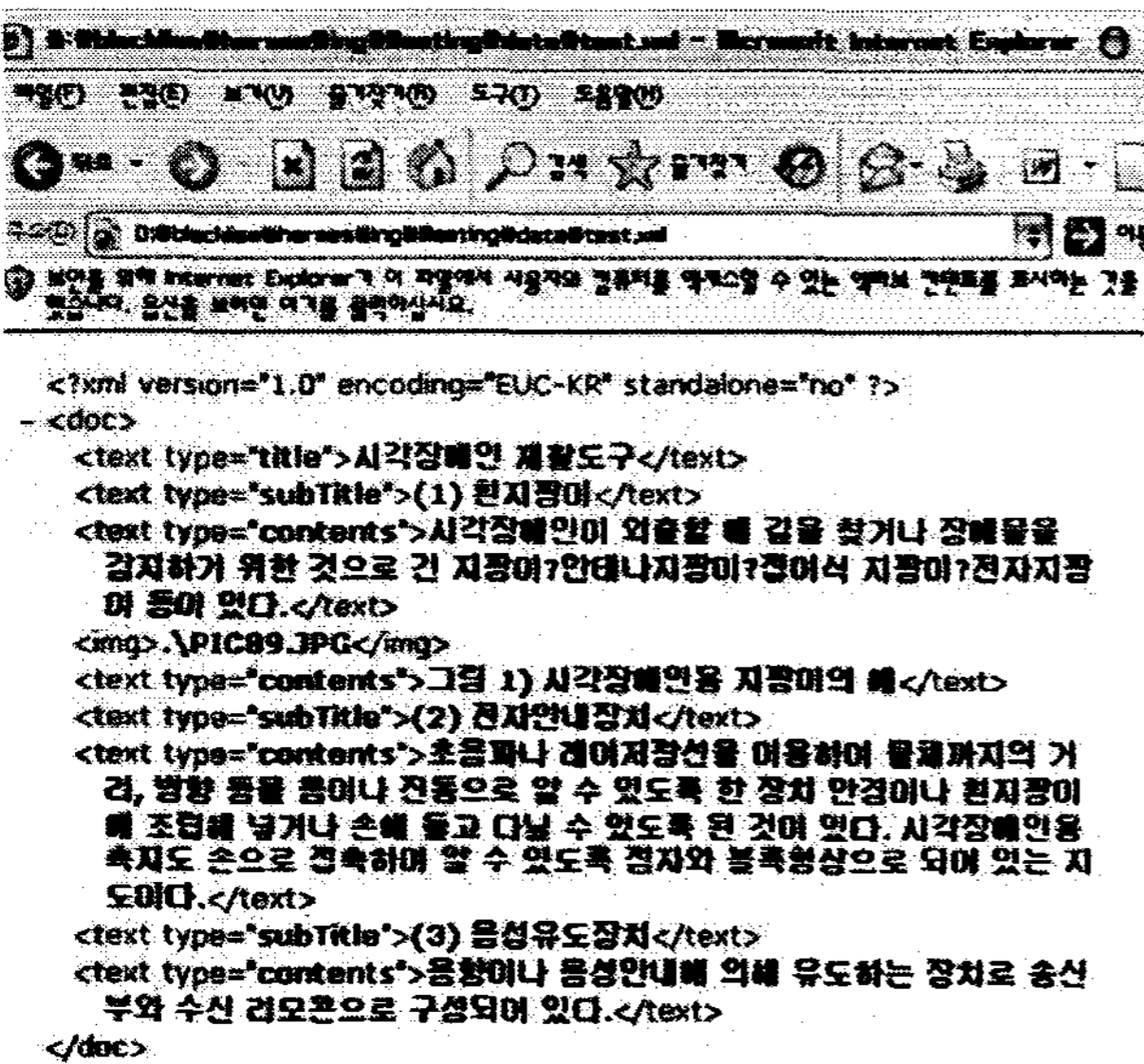


Figure 6 - XML 변환의 예

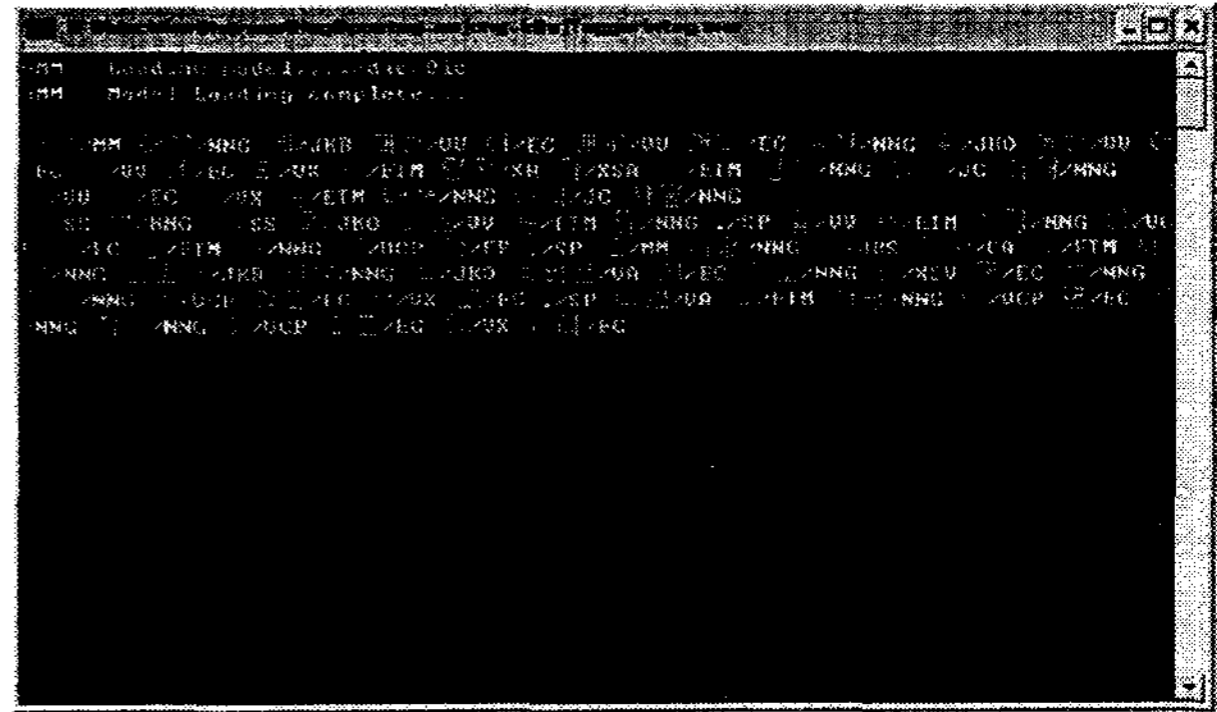


Figure 7 - 형태소 분석기의 구동 예 [6]

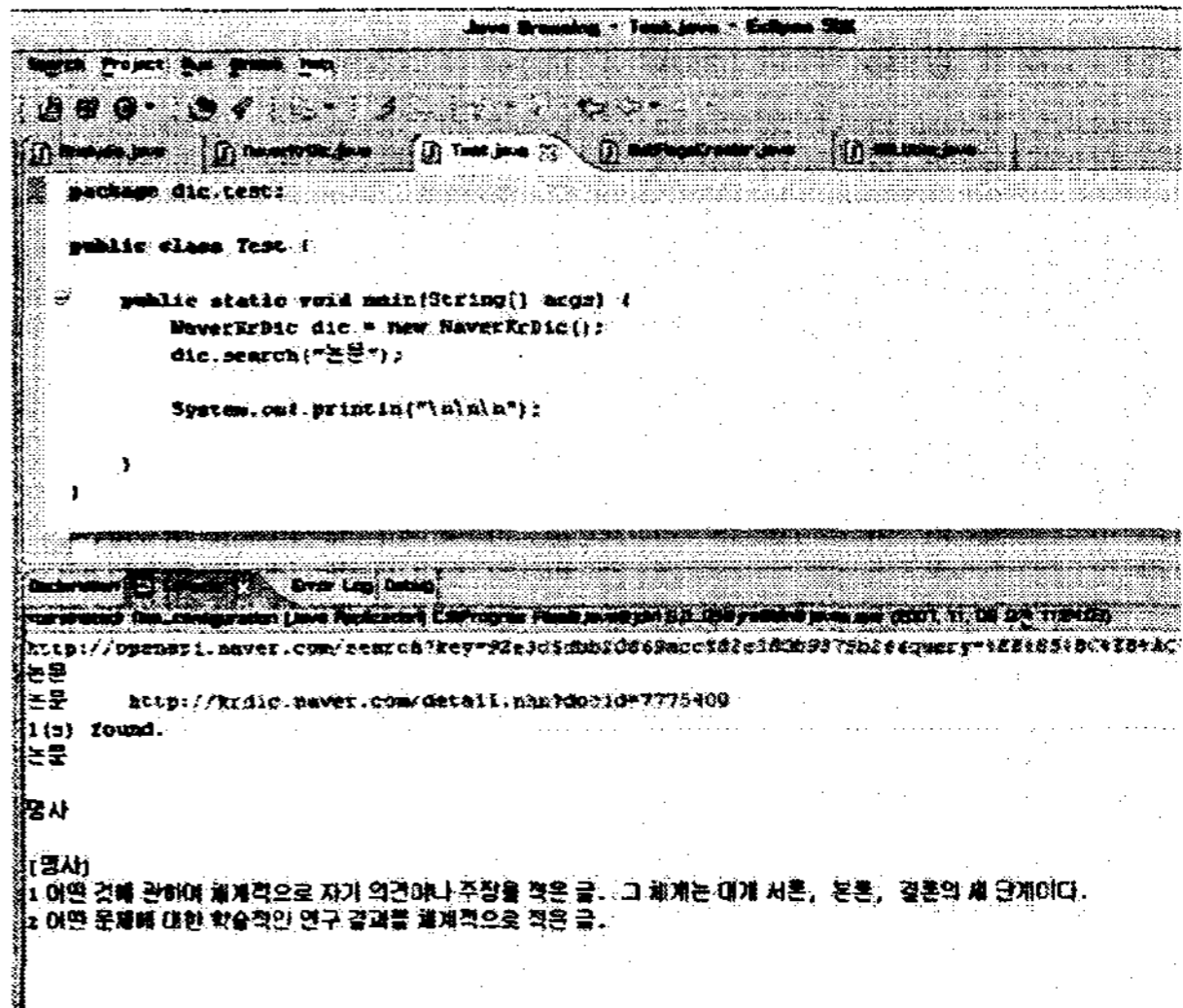


Figure 8 - 사전 분석기

통계적 기법, 문헌구조적 기법, 통계 어휘를 적용하며, 최종적으로 키워드를 선출하는 역할도 수행한다. 현재 지식 관리 시스템은 구축 단계에 있다.

키워드 추출에 앞서 마지막으로 수행하는 단계는 언어학적 기법 적용 단계이다. 이를 위해서는 사전 분석기를 구현하여야 한다. 이 연구에서는 단어의 사전적 의미를 검색할 사전으로 인터넷 포털사이트인 네이버의 국어 사전을 활용하였다. 이를 시스템화하기 위하여 네이버에서 제공하는 Open API를 활용하여 시스템을 구현하였다. [8] 해당 모듈은 Naver Open API를 활용하여 자바와 연동하였고, 검색에 따른 결과값 중 단어의 의미만을 활용하기 위하여 파싱을 수행하였다. 구현된 시스템은 Figure 8과 같다.

5. 결론

본 연구에서는 문서의 의미론적 분석을 위하여

통계적 기법, 문헌구조적 기법, 언어학적 기법, 통제 어휘 활용 방법을 병행하여 사용하는 방법론을 제안하였다. 또한, 위의 방법을 적용할 시스템의 일부를 구현하고, 현재 시스템의 보완 및 확장을 진행하고 있다. 방법을 적용하기에 앞서 동음이의어의 처리 문제와 각 알고리즘을 보완하는 작업이 필요할 것이다. 하지만, 이 작업을 수행하고, 시스템 구현이 완성된다면 기존의 연구가 형태 분석에만 국한되어 있다는 점을 보완하여 문서의 의미를 어느 정도 반영한 키워드 추출을 할 수 있을 것으로 기대한다. 또한, 지능형 회의 공간에서의 ‘회의 문서 처리’와 ‘실시간 회의 기록 처리’ 서비스에 활용할 수 있다는 점에서 효율적인 도구가 될 것으로 기대된다. 추후, 활용 가능한 시소러스가 있어 시소러스에 의한 분석 방법도 추가한다면 더욱 다양한 분야에 적용할 수 있을 것으로 기대된다.

참고문헌

- [1] 배일주, 송민규, 이수홍, 박지형. (2007). “회의 이벤트 및 온톨로지 내부 규칙 처리를 이용한 회의 온톨로지의 동적 거동에 관한 연구”, 2007 한국 CAD/CAM 학회 학술발표회 논문집, pp. 785-791.
- [2] Yasin Uzun. (2005). “Keyword Extraction using Naïve Bayes”, <http://www.cs.bilkent.edu.tr/~guvenir/courses/CS550/Workshop/>
- [3] Matsuo. (2004). “Keyword Extraction from a Single Document Using Word Co-Occurrence Statistical Information”, International Journal on Artificial Intelligence Tools, Vol. 13, No. 1, pp. 157-169.
- [4] Taeho Jo, Malrey Lee, and Thomas M Gatton. (2006). “Keyword Extraction from Documents using a Neural Network Model”, 2006 International Conference on Hybrid Information Technology, Vol. 2, pp. 194-197.
- [5] K. Kaguera and B. Umino. (1996), “Methods of Automatic Term Recognition”, Terminology, Vol. 3, No. 2, pp. 259-289.
- [6] <http://nlp.postech.ac.kr/>
- [7] 임성규. (1989), “현대 국어의 강조법 연구”, 박사학위논문.
- [8] <http://openapi.naver.com/index.nhn>