연상 지식을 이용한 문서 분류 엔진의 구현

장 정 효, 손 주 성, 이 상 곤*, 안 동 언‖
전주대학교 일반대학원 컴퓨터공학과 언어과학실*
전북대학교 일반대학원 컴퓨터공학과 지능정보공학실‖
{fsfsharp, w24e, samuel}@jj.ac.kr*, duan@moak.chonbuk.ac.kr‖

Implementation of Document Classification Engine by Using Associative Knowledge

Jung-Hyo Jang, Ju-Sung Son, Samuel Sangkon Lee*, Dong-Un Ahn‖
Language Science Lab., Dept. of Comp. Sci. & Eng., Jeonju University*
Intelligence Engineering Lab., Dept. of Comp. Eng., Chonbuk National University‖

요 약

인간은 문서 내용의 적절성을 파악하기 위해서는 문서 전체를 읽어 보아야 그 적절성 여부를 알 수 있다. 그러나 문서의 양이 많은 경우나 문서 내에 여러 화제가 산재되어 있으면 문서의 분야를 파악하기 위해 많은 시간과 노력이 필요하게 된다. 따라서 본 논문에서 제안하는 방법은 이러한 비용을 절감하기 위해 카테고리의 트리 정보와 문서의 내용에서 추출한 분야연상어를 지식사전으로 구축하고 이를 이용하는 분류기를 설계하여 수집과 분류에 소요되는 비용을 절감하는 자동 분류기를 구현하였다.

1. 서론

최근의 오프라인 미디어는 물론 인터넷의 이용 증가로 인한 전자문서가 급증하여 수많은 정보를 다양한 경로를 통해 습득하고 있다. 특히, 전자 텍스트 정보의 양이 기하급수적으로 늘어나는 현대에는 정보의 효과적인 관리 및 빠른 검색의 필요성이 대두되고 있다. 인간이 문서 내용의 적절성을 파악하기 위해서는 전체 내용을 읽어 그 적절성 여부를 판단해야 하지만 모든 문서를 읽고 내용의 적절성을 판단하기 위해서는 대단한 양의 비용이 소모되게 된다.

본 논문에서는 대용량의 신문 카테고리 정보를 이용하여 자동으로 분야체계를 구성하고, 어느 문서에나 존재하는 분야연상어를 추출하면 어떤 단어가 어떠한 분야에 대해 어느 정도를 연상하는지에 대한 정보를 구축할 수 있다. 이러한 단어를 연상지식으로 이용하여 사전을 구축할 수 있다[2, 4]. 연상정보를 바탕으로 임의의 문서에 대한 분야를 자동으로 파악하고 분류하는 문서 자동 분류기를 설계하여 작업자의 작업량을 크게 감소시키고, 문서 수집과 분류 비용의 절감 그리고 사용자의 효과적인 정보 관리 및 검색 욕구를 충족시켜 효율적이고 체계적인 정보 관리가 가능하도록 분류엔진을 구현하였다.

2. 관련 연구

2.1 분야연상어

사람이 문서를 읽을 때 문서 전체의 내용을 읽지 않더라도 대표적인 단어를 읽는 것만으로도 스포츠, 정치, 경제 등의 분야를 정확히 인지한다. 이러한 문서의 분야를 대표하는 단어를 분야연상어(FT; Field-associated Terms)라 정의한 참고 문헌 [1, 3, 6, 7]의 연구 방법을 이용하여 분야별 텍스트를 추출할 수 있다.

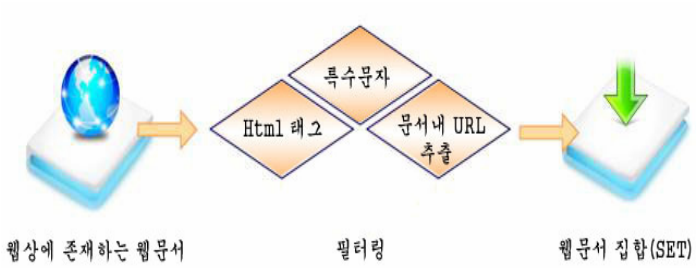
2.2 분야 체계

각 분야에 대한 상/하위 관계의 정보를 표현한 것을 분야 체계라 한다. 분야 체계[7]는 신문 카테고리의 정보를 바탕으로 트리 구조의 형태로 각 분야의 관계 정보를 표현하여 구성하고, 분야 체계의 지정은 <Path>로 기술한다. 예를 들어 <P>=</스포츠/축구>로 표시하며, 모순이 생기지 않는 한 전체 분야는 생략한다.

2.3 분야연상어의 수준

분야연상어가 분야를 결정할 때 연상되는 범위나 개수가 다르기 때문에 분야 체계 내에서 연상되는 분야의 범위에 제약을 부여하고, 각 분야연상어의 수준을 완전 분야연상어(수준 1), 준완전 분야연상어(수준 2), 중간 분야연상어(수준 3), 다분야 연상어(수준 4), 비 연상어(수준 5) 등으로 정의한 것을 이용하였다[3, 6].

3. 웹 문서 수집기

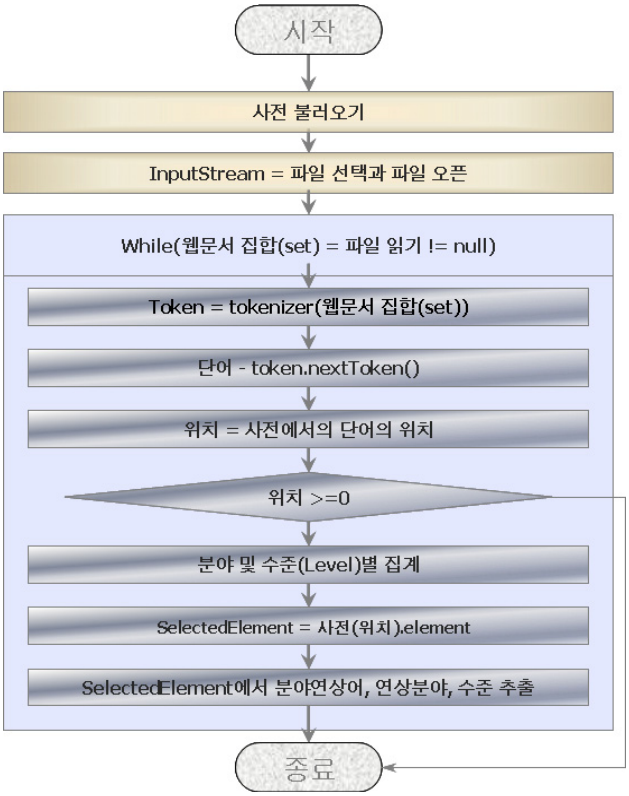


(그림 1) 웹 문서 수집기의 개요

웹 문서 수집기는 2005년에 제시된 연구결과인 참고문헌 [2]의 알고리즘을 그대로 이용하여 인터넷의 문서를 자동으로 수집하기 위한 모듈로서 위의 (그림 1)에서 표시된 것과 같이 인터넷상에 존재하는 웹 문서를 텍스트(Text) 부분과 주소(URL)로 분할하고, 분할된 정보는 웹문서 집합(set)으로 구축한다. 웹 문서 내에서 추출된 주소 정보는 이전과 동일한 방법으로 웹 문서 집합을 수집한다. 이와 같이 인터넷상의 모든 웹 문서에서 불필요한 정보(예를 들면, HTML 태그 등)를 제거한 후 텍스트만을 추출한다. 구현된 알고리즘을 아래의 (그림 2)에 기술하였다[2].

```
입력 : 웹상에 존재하는 문서 집합
출력 : 순수한 텍스트 문서 집합
AWDC_main-Execute(String Url) {  // 시작 메서드
  String TextMine = null; // 호스트에서 온 HTML(링크 & 원문) 문서
  TextMine = ConnectionHost(Url);  // 호스트에 접속한다.
  TextMine = removeScript(TextMine); // 스크립트와 스타일 등의
                                     불 필요한 정보를 제거
  String ProcessedLinkText = getLinkData(TextMine);
              // 링크 정보의 추출과 태그, 특수 문자 등의 제거
  saveURL_to_DB(ProcessedLinkText); // 정제된 Link를 DB에 저장
  String saveFileLocation = Save_to_File(ProcessedLinkText);
                              // 본문과 현재 URL을 파일에 저장
  Parse ps = new Parse(); // 분야 분류 클래스 인스턴스
  ps.Ls-classifier(saveFileLocation); // 자동 분류를 위한
                                      LS-classifier 가동
  String NextURL = getURL_from_DB(); // DB에서 접속할 URL 추출
  if (NextURL.equals(null)) { // 다음에 접속할 URL이 없으면 종료
  System.out.println("프로그램의 종료! WrWn 재확인하세요!");
  } else { AWDC_mainExecute(NextURL); }
}
```

(그림 2) 웹 문서 자동 수집기(AWDC) 알고리즘



(그림 3) 분야연상어 사전 구축 처리 흐름도

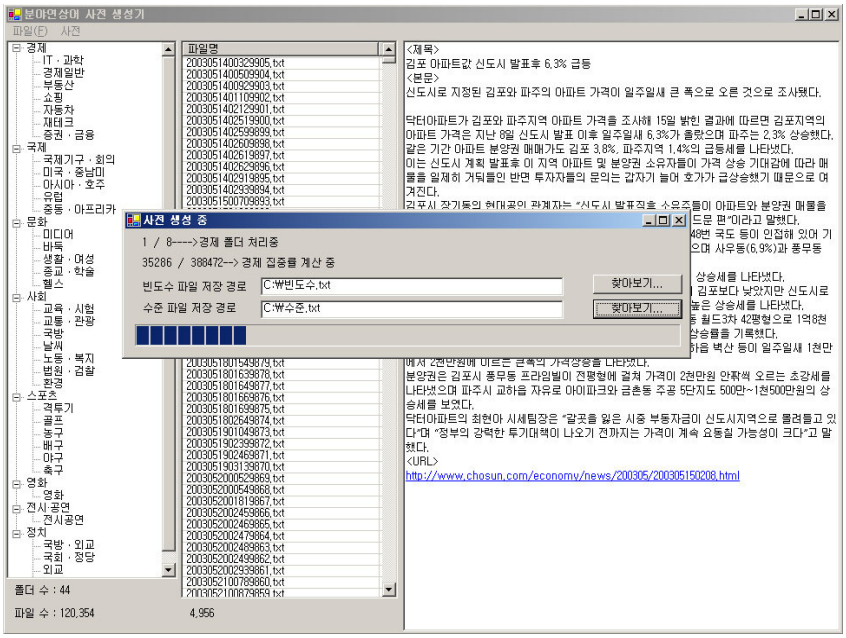
4. 분야연상어 사전 생성기

웹 문서 수집기로 수집한 웹 문서 집합(set)을 분야연상어 사전 생성기는 특수 문자 제거 작업 후, 명사로 구성된 분야연상어 등을 사용하여 형식 형태소를 인식할 수 있는 사전과 반복적이고 의미 없는 불필요한 단어를 제외하는 사전으로 명사를 추정하는 분석 단계를 거쳐 분야연상어 사전을 생성한다. (그림 3)은 사전 생성기가 웹 문서 집합을 분석해 빈도수의 계산과 빈도수 정규화 및 집중률[6]을 계산하여 분야연상어 사전을 구축하는 간략한 처리 흐름도이다.

<표 1> 분야연상어 사전의 구성

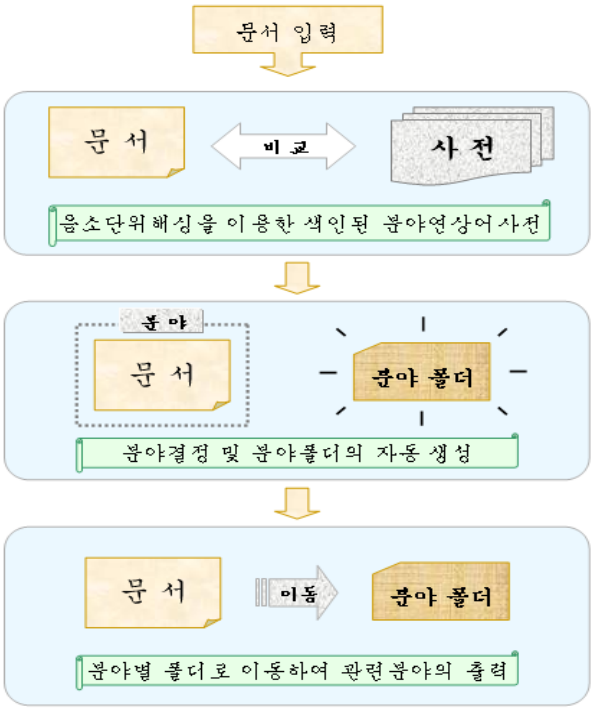
분야연상어	수준	연상분야

위의 <표 1>은 사전 생성기에 의해 구축된 분야연상어 사전의 구조이다[3]. (그림 4)는 웹 문서 자동 수집기가 수집한 웹 문서 집합 파일을 사전 생성기가 읽어 트리 구조를 형성하고, 웹 문서 집합에 대해 빈도수 및 집중률을 계산하여 분야연상어 사전을 구축하는 모습이며, 웹문서 집합 파일에 대한 확인 및 수정이 가능하고, 또한 형식 형태소나 제외해야 할 불필요한 단어의 제외 사전의 업데이트 및 수정을 할 수 있다.



(그림 4) 분야연상어 사전 생성기의 실행 화면

5. 분류기



(그림 5) 분류기 시스템의 구조도

분류기는 임의의 문서를 읽어 들여 자동으로 분야를 판별한다. 사전 생성기에 의해 구축된 분야연상어 사전을 이용하여 분야에 대한 수준별 가중치를 계산하며, 분야별 합계로 문서의 분야를 결정한다. 문서의 분야는 완전 / 준완전 / 중간 분야의 분야연상어의 가중치가 높은 문서의 분야를 판별하여 안정적인 분류 결과를 얻도록 하였다. 다 분야의 분류는 본 논문의 분류기의 목적과 상이하여 이번 실험에서는 제외하였다.

대량의 문서를 읽어 들여 (그림 5)와 같이 음소단위 해시로 색인된 사전의 정보와 비교하여 문서의 분야를 결정하고, 분야 폴더를 생성하여 문서를 분야 폴더로 이동하고 결과를 사용자에게 보여주게 된다. 이러한 분야 폴더 생성과 이동으로 분류와 관리를 용이하게 하였다. (그림 6)은 문서 분류기가 대량의 문서를 읽어 문서의 분야를 자동으로 파악하고, 분류된 문서를 분야 폴더로 생성하여 이동시킨 후 트리 구조의 형태로 사용자에게 보여주며 분류된 통계 정보를 제공한다. 파일의 확인 및 수정 역시 가능하다.

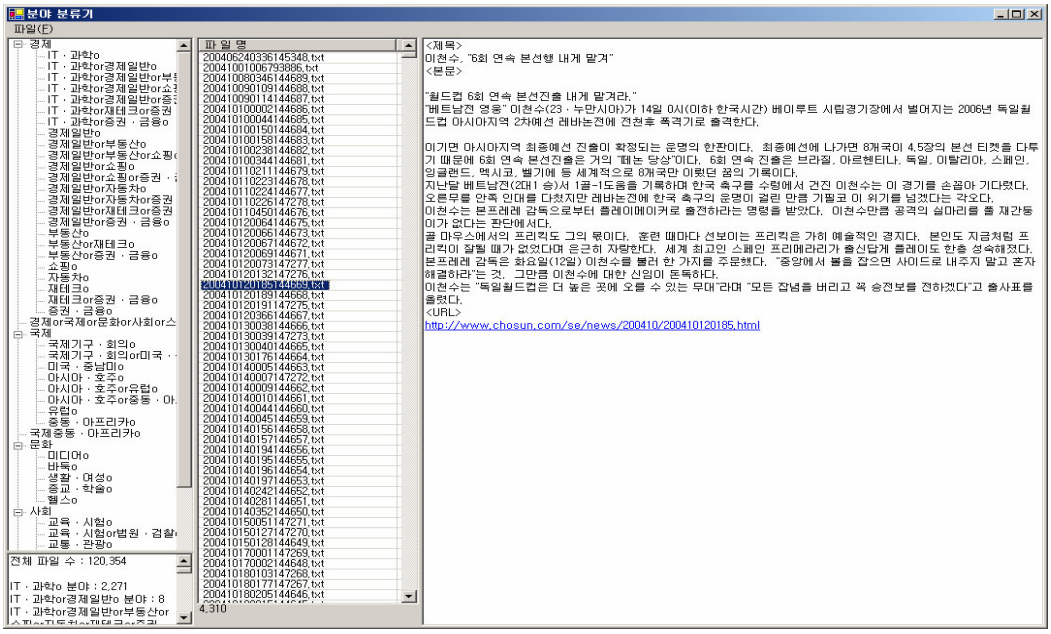
본 논문의 문서 자동 분야 분류기의 구현 환경은 다음과 같다.

- CPU : Pentium IV 3.0 GHz,
- Memory : 1 GB,
- 개발 언어 : Java, C#. net,
- 데이터 베이스 : MS-SQL

<표 2> 수준별 분야연상어의 수

수 준	분야연상어의 수
1	532,100
2	23,492
3	7,420
4	159,355

웹 문서 자동 수집기에 의해 수집된 120,350건의 조선일보 신문 기사와 신문 카테고리의 44개의 분야를 이용하였다. 수준별 분야연상어는 전체 722,367개 이며, <표 2>는 각 수준 1—4까지의 집계를 보여주고 있다.



(그림 6) 문서 분류기의 실행 화면

6. 결론

본 논문에서는 신문 카테고리의 트리 정보와 연상지식을 이용하여 전자문서의 분야를 자동으로 분류하는 방법을 제시하였다. 각 분야를 빠르게 판별할 수 있는 대표 단어인 분야연상어의 사전을 인간의 개입 없이 자동으로 구축하였고, 분야를 자동으로 파악하여 분류하는 문서 분류기를 구현하였다. 대량의 문서를 인간이 분류할 때 걸리는 시간과 비용보다 단축시킬 수 있었으며, 자동으로 분류된 내용을 분리시켜 필요한 자료의 검색 및 사용 등 효과적인 정보의 관리가 가능하도록 하였다. 분야연상어는 인간의 연상 지식을 사용하기 때문에 컴퓨터가 인간의 인지작용과 흡사하게 문서를 읽어 문서 분야를 빠르게 판단하도록 한다. 분야연상어의 구축은 분야체계를 미리 정의해야 하지만 분류 분야가 변경되어도 쉽게 분야를 확장할 수 있을 것으로 기대된다.

향후에는 문서에 화제가 산재하여 있는 경우에는 다 분야로 판단하기 때문에 화제의 추적 기술과 한국어의 단락 검색에 적당한 방법을 이용하면 좀 더 성능의 개선이 생각된다. 또한 현재는 분야 체계를 신문 카테고리를 이용하여 제한적 이었지만 혼합형 트리 구조의 형태를 이용하면 더욱 폭 넓은 분야를 분류할 수 있을 것이다[1, 6, 7].

감사의 글

본 과제(결과물)는 교육인적자원부 · 산업자원부 · 노동부의 출연금으로 수행한 산학협력중심대학 육성사업의 연구결과입니다.

참고 문헌

[1] 이 상 곤, "분야연상어를 이용한 화제분야의 계산방법과 단락검색", 정보처리학회논문지(B), 제 12권, 제 1호, pp. 57-68, 2005.

[2] 장 정 효, 손 주 성, 김 도 연, 이 상 곤, 이 원 휘, 안 동 언, "검색과 분류가 동시에 가능한 JULSE 시스템의 설계 및 구현", 정보처리학회 2005년도 추계 학술발표 논문집(상), 제 12권, 제 2호, pp. 673-676, 2005.

[3] 이 상 곤, "한글 문서분류용으로 이용할 복합어로 구성된 분야연상어의 추출법", 정보과학회논문지: 소프트웨어 및 응용, 제 32권, 제 7호, pp. 636-649, 2005.

[4] 이 원 휘, 김 도 연, 이 상 곤, "그래픽컬한 분야인식기의 설계 및 구현", 정보과학회 가을 학술발표 논문집, 제 31권, 제 2호, pp. 769-771, 2004.

[5] 이 원 휘, 최 현, 이 상 곤, "분야연상어 추출방법의 설계와 구현", 정보처리학회 2004년도 춘계 학술발표 논문집, 제 11권, 제 1호, pp. 651-654, 2004.

[6] 이 상 곤, 이 완 권, "분야연상어의 수집과 추출 알고리즘", 정보처리학회논문지(B), 제 10권, 제 3호, pp. 347-358, 2003.

[7] 이 상 곤, "분야연상어를 이용한 화제의 계속성과 전환성을 추적하는 단락분할방법", 정보처리학회논문지(B), 제 10권, 제 1호, pp. 57-66, 2003.