검색과 분류가 동시에 가능한 JULSE 시스템의 설계 및 구현

장정효, 손주성, 김도연, 이상곤,[O] 이원휘, 안동언[‖]
전주대학교 일반대학원 컴퓨터공학전공 언어과학실[O]
전북대학교 일반대학원 컴퓨터공학과 지능정보공학실[‖]

{fsfsharp, w24e, kdy, samuel}@jj.ac.kr[O] and {wony, duan}@moak.chonbuk.ac.kr[‖]

Design and Implementation of Field Classification and Information Retrieval Engine: JULSE

Jeong-Hyo Jang, Ju-Sung Son, Do-Yun Kim, Samuel Sangkon Lee,[0]
Won-Hee Lee, and Dong-un Ahn[‖]
Language Science Lab., Dept. of Computer Science & Engineering, Jeonju University,[0]
Intelligence Engineering Lab., Dept. of Computer Engineering, Chonbuk National University[‖]

요 약

기존의 정보검색 엔진은 문서의 분야에 상관없이 본문 전체의 내용을 보여주므로 사용자가 적합한 내용인지를 파악하기 위해서는 본문 전체를 읽어 보아야 그 적절성 여부를 알 수 있다. 본 논문에서 제안하는 방법은 질의어가 지시하는 분야를 분야연상어를 이용하여 자동으로 파악하고, 사용자가 원하는 분야에서의 검색이 이루어지도록 하는 검색과 분류가 동시에 가능한 엔진을 설계하여 검색결과의 성능을 향상하고자 한다. 이와 함께 적당한 분야연상어가 다수 출현한 단락을 사용자에게 제공하여 본문 전체를 보지 않아도 질의어에 적당한 문서인지를 빠르게 파악하도록 설계하여 구현하였다.

1. 서론

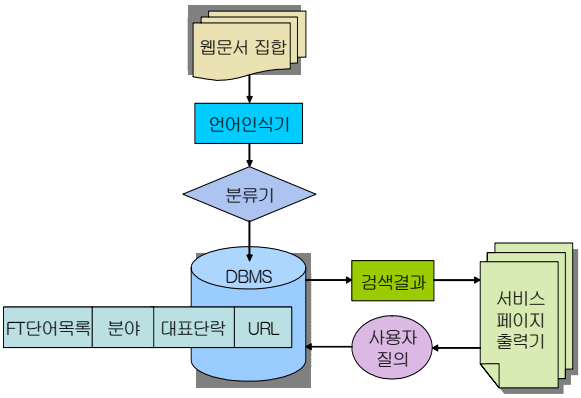
기존의 검색엔진은 인터넷에 등록된 문서를 검색할 때 몇 개의 키워드만을 이용하여 검색을 수행하므로 검색된 문서의 정확도를 확신할 수 없는 실정이며, 인터넷에 등록되지 않는 문서는 검색할 수 없는 한계를 보이고 있다. 이러한 텍스트 기반의 정보검색 엔진이 갖는 성능하락의 문제점을 해결하기 위해서는 질의어를 잘 분석하여 사용자가 찾고자 하는 분야에 한정된 검색이 필요하다.

본 논문에서는 문서의 해당 분야를 자동으로 결정할 수 있는 분야연상어(어느 문서에서나 존재)[5]를 이용하여 분야를 결정하고, 질의어와 동일한 분야에 해당하는 문서만을 검색하여 검색의 정확도를 향상하고자 한다. 또한 인터넷의 모든 문서를 자동으로 수집하여 대량의 문서 정보를 필터링하고, 분야별 분야연상어를 추출하고 이를 이용하여 사용자의 검색에 적합한 분야 내에서 문서를 검색한다. 분류결과가 안정적으로 이용되도록 완전/준완전/중간/다 분야의 분야연상어(FT; Field-associated Terms)를 추출하여 문서검색의 정밀도를 획기적으로 높이고자 한다.

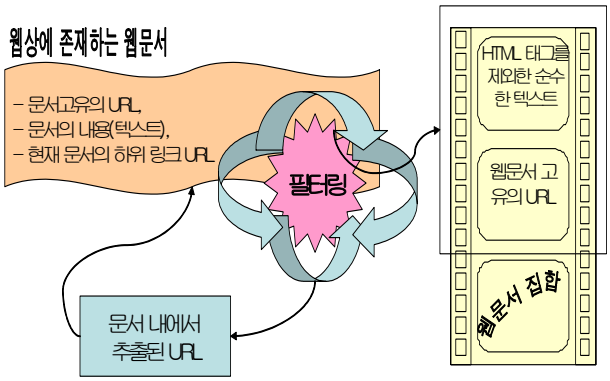
2. 분류방법

본 장에서는 인간두뇌의 연상지식에 기반한 정보검색/분류엔진(JULSE; Jeonju University Language Search Engine)을 설계하였다. 먼저 검색/분류엔진 시스템의 전체 구조는 다음의 (그림 1)과 같다. 본 논문에서 제시하는 시스템은 크게 네 가지의 핵심 모듈 (1) 웹 문서 자동 수집기(AWDC; Agent for Web Document Collector), (2) 분류기(LS-Classifier), (3) 데이터베이스 관리기(DBMS), (4) 서비스 페이지 출력기(Service Provider) 등으로 구성하였다.

웹에서 자동으로 수집한 문서를 분야연상어 사전을 통해 분류하는 기능과 분류된 정보를 데이터베이스에 체계적으로 지식베이스화하여 저장한 후, 구축된 지식을 바탕으로 사용자의 질의에서 추정된 분야와 동일한 분야의 문서를 검색하여 그 결과를 웹 페이지에 다시 되돌려 주는 기능을 포함한다. 다음 절에서는 위에서 언급한 4가지 모듈에 대하여 차례대로 설명한다.



(그림 1) 전체 시스템의 구조



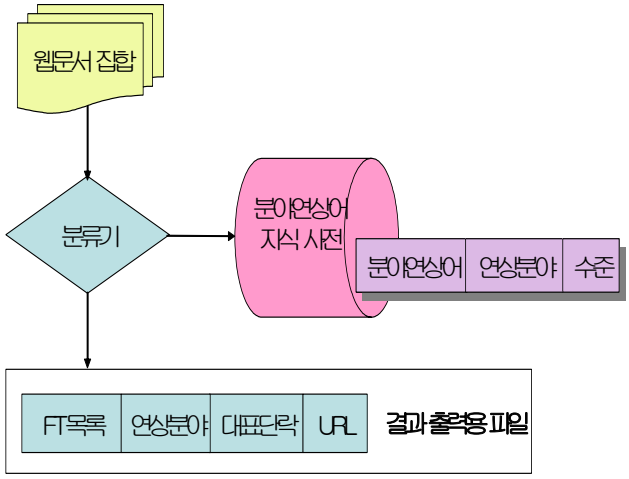
(그림 2) 웹 문서 자동 수집기의 개요

2.1 웹문서 자동 수집기

웹문서 수집기는 인터넷의 문서를 자동으로 수집하기 위한 모듈로서 (그림 2)에서 표시된 것과 같이 인터넷상에 존재하는 웹 문서를 텍스트(Text)부분과 주소(URL)로 분할하고, 분할된 정보는 웹문서 집합(set)으로 구축된다. 웹문서 내에서 추출된 주소정보는 이전과 동일한 방법으로 웹문서 집합을 수집한다. 이와 같이 인터넷상의 모든 웹문서를 자동 수집한다. 필요한 정보(예를 들면, HTML 태그 등)를 필터링하여 제거한 후 텍스트만을 추출한다. 구현된 알고리즘을 (그림 3)에 기술하였다. 순수하게 텍스트만 추출된 문서는 다음의 분류기(LS-Classifier)로 보내지게 된다.

```
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
                                     불필요한 정보를 제거
  String ProcessedLinkText = getLinkData(TextMine);
            // 링크정보의 추출과 태그, 특수문자 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); // 분류하기 위한
                                      LS-classifier 시동
  String NextURL = getURL_from_DB(); // DB에서 접속할 URL 추출
  if (NextURL.equals(null)) { // 다음 접속할 URL이 없으면 종료
  System.out.println("프로그램 종료~! WrWn 확인 요망!");
  } else { AWDC_mainExecute(NextURL); }
}
```

(그림 3) 웹 문서 자동 수집기(AWDC) 알고리즘



(그림 4) 분류기의 개요도와 분야연상어의 사전 구조

2.2 분류기

위의 (그림 4)에 제시한 바와 같이 수집된 웹문서 집합에서 구축된 분야연상어 사전[3]을 이용하면 분류기의 작업을 거쳐 단어목록, 분야, 대표단락의 선정[1, 6], 주소정보 등으로 분류하여 출력파일에 분야연상어 사전이 구축된다. 분야연상어 사전은 분야연상어와 그 단어가 연상하는 분야 및 수준 정보로 구성되어 다섯 가지의 수준별로 분류된 완전/준완전/중간/다 분야의 분야연상어[4] 등을 이용하기 때문에 분류작업이 가능하게 된다. 분야연상어 사전의 구축 방법은 참조문헌 [3, 4]에서 제시된 방법을 사용하였고, 구현된 알고리즘은 (그림 5)에 설명하였다.

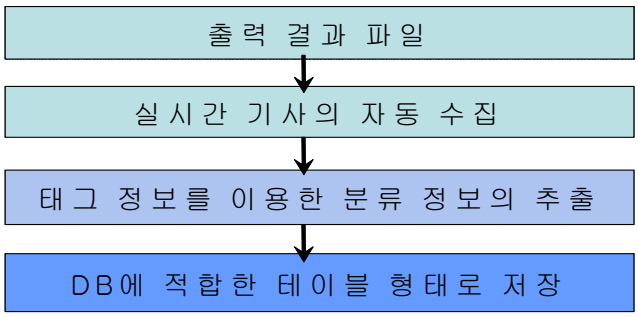
No	FT_Lists	Associated Fields	Paragraph	URL
1	야구, 홈런, 김응룡감독, 방어율, 이닝, 투수,	</야구>, ...	P32, P76, P127, ...	http://www.chosun.com/~sport/942345.html...
...	...	...	...	...

<표 1> 데이터베이스 테이블

2.3 데이터베이스 관리기

분류기에서 추출된 출력파일은 실시간 자동 문서수집 모듈로 수집되어 출력파일의 태그정보로 사용되는데 분류기가 분석한 내용을 분류하여 추출한 정보를 데이터베이스에 저장한다. 데이터베이스의 테이블 구조의 생성 알고리즘을 아래에 표시하고 생성된 결과를 위의 <표 1>에 정리하였다.

(그림 5) 분류기의 순서도



(그림 6) 데이터베이스 관리기

입력 : 순수한 텍스트 파일
출력 : <표 1>의 데이터베이스 테이블

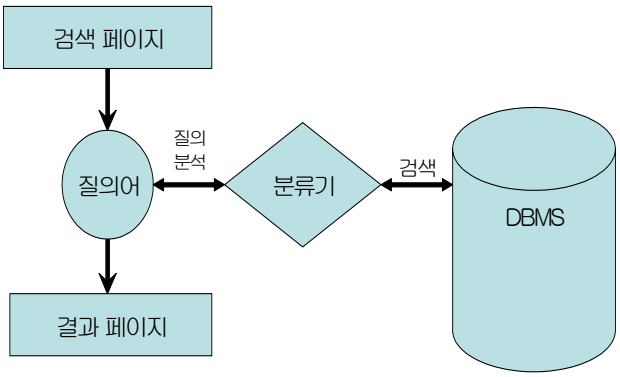
```
Data_Base_Management_System()
{
  DirInfo SourceDir;  // 순수 텍스트 파일이 저장된 폴더
  DirInfo BackupDir;  // 순수 텍스트 파일이 백업되는 폴더
        // 백업한 폴더 이름이 오늘 날짜와 동일한지 비교
  if (BackupDir.Name != Date.Now) {
     BackupDir = Dir.Create(Date.now); }

  Timer_Tick  // 일정 시간 마다 동작
  {
  FileInfo[] TextFileName = SourceDir.GetFiles();
                    // SourceDir안의 파일을 배열로 생성
  for (int i = 0 ; i < TextFileName.Count ; i++) {
        // 파일에서 분야를 추출
    string Field = TextFileName[i].Read(Field);
    string Word_List = TextFileName[i].Read(WordList);
    string Paragraph = TextFileName[i].Read(Paragraph);
    string URL = TextFileName[i].Read(URL);
  if (!DB.tablename.Exists(Field))  // DB Table 중 분야와
                                    일치하는지 검사
  {
    Create.table(Field); // Table 생성
  }
     DB.Insert(WordList, Paragraph, URL); // DB구조에 맞게 저
장
     TextFileName[i].MoveTo(BackupDir); } // 순수 텍스트 백업
  }
}
```

(그림 7) DBMS의 알고리즘

2.4 서비스 페이지 출력기

서비스 페이지 출력기(그림 8 참고)는 사용자의 요구조건을 받아들이고 그 결과를 출력하는 기능을 담당한다. 검색페이지에서 사용자의 질의를 입력받아 질의를 분석하는 과정을 거친다. 사용자의 질의를 분류기에서 웹문서 집합을 분류하는 것과 같은 작업을 거쳐 사용자의 질의가 어떤 분야의 문서를 검색하고자 하는가를 알아내어 사용자의 질의에 적합한 분야를 결정하고 이와 동일한 분야에서의 검색이 이루어진다. 이는 데이터베이스에서 검색과 질의가 병렬적으로 이루어지며 검색결과는 결과 페이지를 통해 사용자에게 인터페이스 된다.



(그림 8) Service Provider의 구조

사용자의 검색 질의에 해당하는 분야를 중심으로 검색되기 때문에 검색대상의 문서를 축소할 수 있으며 동시에 검색시간도 줄일 수 있다. 이것은 기존의 검색엔진 보다 빠른 검색이 가능하고, 사용자가 원하는 정확도가 높은 문서를 출력할 수 있는 장점이 있다.

2.5 구현환경

본 논문에서 제시한 모듈의 구현환경은 아래와 같다.

- CPU : Pentium IV 3.0 GHz,
- Memory : 512 MB,
- 개발언어 : Java, C#.Net, JSP, XML,
- 데이터베이스 : MS-SQL

3. 결론

본 논문에서는 분야연상어를 이용하여 질의어와 검색되는 문서를 동일한 분야(영역) 내에서 검색하는 방법을 제시하였다. 분야연상어 사전의 자동구축과 동시에 분류가 가능한 새로운 개념의 인터넷 정보 검색/분류 엔진을 구현하였다.

단락검색 방법을 통해 대표단락을 먼저 제시함으로서 사용자의 질의에 보다 정확한 결과를 제공할 수 있게 되며, 검색과 관련이 적은 정보는 빠르게 차단할 수 있다. 분야연상어 구축은 분야체계를 미리 정의해야 하지만 분류체계가 변경되어도 손쉬운 적용이 가능할 것으로 기대된다[5].

분야연상어는 인간의 연상지식을 사용하기 때문에 컴퓨터가 인간 두뇌의 인지작용과 유사하게 웹문서를 읽어 어떤 분야에 속하는지를 자동으로 빠르게 판단하도록 한다. 또한 웹검색 엔진에 등록되지 않은 웹문서도 수집되므로 기존의 정보 검색 엔진보다 폭 넓은 검색 결과를 제공할 수 있다.

현재는 단락검색에서 대표적인 단락(대표 단락)만을 제공하고 있지만, 향후에는 화제를 추적하여 웹문서 내에 있는 분야별 단락을 추출하고 제공한다면 자동 요약기술에 응용할 수 있으리라 기대된다.

감사의 글

이 논문은 2003년도 한국학술진흥재단의 지원에 의하여 연구되었음(KRF-2003-003-D00415). 재단의 연구비 지원에 감사 드립니다.

참고 문헌

[1] 이상곤, "분야연상어를 이용한 화제분야의 계산 방법과 단락검색", 정보처리학회논문지(B), 제 12권, 제 1호, pp. 57-68, 2005.

[2] 이상곤, "한글 문서분류용으로 이용할 복합어로 구성된 분야연상어의 추출법", 한국 정보과학회 논문지: 소프트웨어 및 응용, 제 32권, 제 7호, pp. 636-649, 2005.

[3] 이원휘, 김도연, 이상곤, "그래픽컬한 분야인식기의 설계 및 구현", 한국정보과학회 가을 학술발표 논문집, 제 31권, 제 2호, pp. 769-771, 2004.

[4] 이원휘, 최현, 이상곤, "분야연상어 추출방법의 설계와 구현", 한국정보처리학회 2004년도 춘계학술발표 논문집, 제 11권, 제 1호, pp. 651-654, 2004.

[5] 이상곤, 이완권, "분야연상어의 수집과 추출 알고리즘", 정보처리학회 논문지(B), 제 10권, 제 3호, pp. 347-358, 2003.

[6] 이상곤, "분야연상어를 이용한 화제의 계속성과 전환성을 추적하는 단락분할방법", 정보처리학회 논문지(B), 제 10권, 제 1호, pp. 57-66, 2003.