그래픽컬한 분야인식기의 설계 및 구현

이원휘[1], 김도연[2], 이상곤[1]
전주대학교 정보기술공학부 언어과학실[1,2]
{wony[1], samuel[1]}@jj.ac.kr & neoholland@hanmail.net[2]

Design and Implementation of Graphical Field Recognizer

Won-Hee Lee[1], Do-Yun Kim[2], and Samuel Sangkon Lee[1]
Language Science Lab., School of Information Technology & Engineering, Jeonju University

요 약

사람은 문서를 읽을 때 문서 전체를 읽지 않더라도 대표적인 단어를 보는 것만으로 정치, 경제, 교육, 스포츠 등의 분야를 정확히 인지한다. 이러한 단어를 분야연상어로 정의하고, 빈도수 조사를 통해 전자사전에 자동으로 구축한다. 이러한 분야연상어는 문서의 초기인식 시 어느 분야인지 분명하지 않거나 애매한 경우에도 문서를 읽어가면서 분야를 인지할 수 있도록 도와준다. 본 논문에서는 이러한 특성을 가지고 있는 분야연상어를 이용하여 시스템에 새로운 문서가 주어질 때 해당 문서의 분야정보를 추출하고, 분야연상어의 분포정보를 인간에게 그래픽컬 하게 보여줄 수 있도록 분야인식기를 설계하고 구현한다.

1. 서론

대량의 전자문서에서 사용자의 검색요구에 맞는 문서를 검색하는 종래의 검색방법은 문서전체를 하나의 객체로 생각하여 검색요구에 적당한 문서를 검색하여 왔다. 그러나 실제문서에서는 복수의 화제가 혼합되어 있기 때문에 문서전체를 검색대상으로 하지 않고, 검색요구에 정확히 일치하는 텍스트 단편만을 검색하는 단락검색(PR; Passage Retrieval) 기술이 미국이나 일본 등지에서 주목을 받고 있으나, 한국에서의 연구는 활발하지 못한 실정이다[2].

본 논문에서는 분야연상어를 이용하여 검색요구에 일치하는 문서를 인식하고, 검색된 문서에서 일치되는 단락의 추출을 목적으로 한다. 단락이 문서의 특정화제에 대해 쓰여진 것인지를 판별하기 위해 단락의 범위를 결정하고, 그 단락의 분야를 결정하는 방법을 제안한다.

이하 제 2장에서는 본 논문의 이론적 배경이 되는 분야연상어에 대하여 살펴보고, 3장에서는 주어진 문서에서의 분야와 단락의 결정에 대하여 설명한다. 4장에서는 설계 시스템의 개요를 설명하고, 구현한다.

2. 분야연상어

사람은 문서를 읽을 때 문서 전체를 읽지 않더라도 대표적인 단어를 보는 것만으로 정치, 경제, 교육, 스포츠 등의 분야를 정확히 인지한다. 이러한 단어를 분야연상어로 정의한다. 이 분야연상어는 특정한 분야를 정확하게 연상할 수 있는 단어로서 잘 분류된 문서 컬렉션에서 구축할 수 있다. 분야연상어를 이용하면 문서를 관련된 분야별로 추출할 수 있다.

2.1 분야체계

분야체계란 각 분야의 상위·하위관계를 트리구조로 표현한 분야별 체계를 말한다. 트리분야에 따라 문서 데이터(문서 컬렉션)를 미리 분류하고, 각 문서 내에 존재하는 분야연상어를 추출한다. 각 분야에 속하는 문서 데이터 내에 출현하는 분야연상어의 집중률을 계산한다. 여기서 구해진 분야연상어는 형태소사전에 등록되어 있는 단어(단일어 혹은 단위어)이며, 복합어에 대한 분야연상어는 단일어의 분야계승에 기초하여 반자동적으로 구축할 수 있다. 그러나 본 논문에서는 형태소 분석을 거치지 않아 많은 노이즈가 있음에도 분야연상어를 추출한 참조논문 [5]에서 구현한다. 분야연상어 추출방법을 이용하여 분야연상어 사전을 이용한다.

2.2 분야연상어의 수준

문서 컬렉션에서 추출한 분야연상어는 연상되는 분야의 넓이에 차이가 있다. 이 단어는 유일한 종단분야나 중간분야를 지시하는 단어 혹은 복수의 종단분야나 중간분야를 지시하므로, 각 분야연상어 w의 수준을 완전분야연상어(수준1), 준완전분야연상어(수준2), 중간분야연상어(수준3), 다분야연상어(수준4) 등으로 나누어 정의하여 사용한다[1]. 분야연상어의 자세한 내용은 참조[1, 3-4]를 참조하면 좋을 것이다.

3. 분야와 단락의 결정

본 논문에서는 문서의 각 문단마다 처리를 진행해 단락 별 분야와 문서 전체의 분야를 추출한다. 이하 설명에서 사용되는 각각의 변수를 정의한다. 먼저, 처리대상 문서 $d_i=\{p_{i,1}, p_{i,2}, p_{i,3}, \ldots, p_{i,j}, \ldots p_{i,m}\}$이다. 단, $p_{i,j}$는 문서 d_i내의 j번째 문단을 표시한다. 문단 $p_{i,j}=\{w_1, w_2, w_3, \ldots, w_k, \ldots, w_m\}$이다. w_k는 문장을 이루는 어절로서 분야연상어 후보가 된다. F는 분야트리 전체집합을 의미하며, $\{F_1, F_2, F_3, \ldots, F_l, F_m\}$으로 구성된다. $Frequency(p_{i,j}, F_l)$는 문서 d_i의 한 문단 p_j내에 존재하는 분야 F_k의 분야연상어의 정수(분야연상어의 빈도수를 의미)이다.

본 방법은 문서 내에 존재하는 분야연상어를 각 문단에서 추출한다. 참조[4]에서 미리 작성된 분야연상어 사전을 이용하여 각 문장에 존재하는 모든 분야연상어를 추출하고, 추출된 분야연상어는 각 수준별·분야별로 집계된다. 이렇게 집계된 수준별·분야별 집계내용을 토대로 각 문단별 분야를 결정하게 된다. 또한 최종적으로 각 문단별로 취합된 수준별·분야별 집계내용은 다시 상위폴더에 취합되어 문서 전체의 분야를 결정하는데 사용된다. 이로서 문서전체의 분야와 각 문단별 분야를 결정할 수 있다. 하나의 문서가 여러 분야를 포함하고 있는 문서라 할지라도 문서내의 문단별로 분야의 분포를 분석할 수 있다.

4. 분야인식기의 설계 및 구현

본 장에서는 앞장에서 논의된 알고리즘을 기반으로 실제 분야인식기를 설계하고 구현한다.

4.1 구현환경

분야인식기의 구현환경은 다음과 같다. 메모리 256MB와 CPU Pentium IV 2.6 GHz 속도를 가진 시스템에서 마이크로시스템사에서 개발한 자바(JAVA) 컴파일러를 이용하였다. 데이터 구조는 디렉토리 구조이며, 분야연상어 사전(FT.dic)이며, *.txt은 분석의 대상이 되는 모든 문서파일이다.

4.2 구현방법

분야인식기의 주요 기능은 주어진 문서를 제공되는 분야연상어 사전을 이용하여 분석한 후, 각 수준별로 출력하는 기능과 이를 토대로 각 문단별 분야정보와 문서전체에 걸쳐 분야정보를 그래프로 표현한다. 수준별 문서의 출력은 분야연상어를 수준별로 인식하고, 인식된 분야연상어는 배경색을 다르게 표현하며, 표현 그래프는 막대그래프와 꺾은선그래프를 사용하였다.

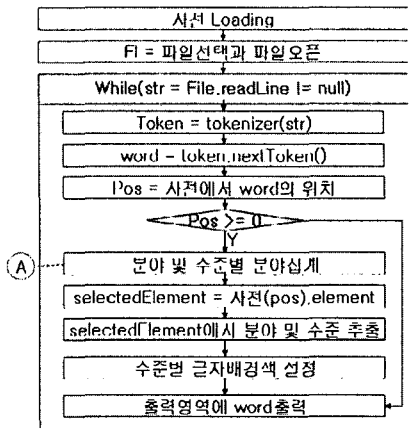
4.3 텍스트 분석

텍스트의 분석은 사전 로딩 과정을 거쳐 분석하고자 하는 파일에서 한 문단씩 읽어들여 어절단위로 분리하고, [그림 1]에서와 같이 분야연상어 사전에서 분리된 어절을 찾는 메서드를 실행한다. 이 메서드는 찾고자 하는 분야연상어가 사전에 있을 경우에는 사전에서 해당 분야연상어의 위치를 반환하며, 실패한 경우 -1을 반환한다. 따라서 반환된 위치정보가 0보다 크거나 같을 경우 사전에서 해당 위치에 있는 분야 및 수준을 추출할 수 있다. 추출된 수준에 따라 출력영역에 출력할 분야연상어의 배경색을 설정한다. 본 논문의 방법은 각 분야별로도 배경색을 설정할 수 있으나, 오히려 분석자로 하여금 혼란을 초래할 수 있어 단순히 수준별로 4가지 색만 허용한다. 이 과정을 읽어들인 문단이 NULL이 될 때까지 반복한다. 사전탐색은 이진탐색 알고리즘을 이용하였다.

분야연상어 사전의 구조와 텍스트 분석 흐름도는 다음 (그림 1)과 (그림 2)와 같다

분야연상어	연상분야	수준

(그림 1) 분야연상어사전의 구조



(그림 2) 텍스트분석을 위한 개요도

위 (그림 2)에서 Ⓐ 부분의 "분야 및 수준별 분야집계"는 해쉬 테이블을 이용하며, 필드명을 키 값으로, 빈도수를 값으로 하는 데이터 쌍을 각 수준별로 저장하고, 각 문단별 벡터를 생성하여 관리한다. 다음 (그림 3)과 (그림 4)는 문단별 수준과 분야 정보를 저장하기 위한 해쉬테이블과 벡터의 구조이며, (그림 5)는 문서 전체의 수준과 분야정보를 저장하기 위한 해쉬 테이블이다.

Sub_Level 1	Field_1	Field_2	...	Field_n
	Value_1	Value_2	...	Value_n
Sub_Level 2	Field_1	Field_2	...	Field_n
	Value_1	Value_2	...	Value_n
Sub_Level 3	Field_1	Field_2	...	Field_n
	Value_1	Value_2	...	Value_n
Sub_Level 4	Field_1	Field_2	...	Field_n
	Value_1	Value_2	...	Value_n

(그림 3) 문단별-수준별 분야 및 빈도수를 저장하기 위한 해쉬테이블의 구조

Vector vSubHash

Sub_Level 1	Sub_Level 1	Sub_Level 1	...
Sub_Level 2	Sub_Level 2	Sub_Level 2	
Sub_Level 3	Sub_Level 3	Sub_Level 3	
Sub_Level 4	Sub_Level 4	Sub_Level 4	

(그림 4) 문단별로 취합된 수준별 분야 및 빈도수 해쉬테이블을 저장하는 벡터의 구조

Total_Level 1	Field_1	Field_2	...	Field_n
	Value_1	Value_2	...	Value_n
Total_Level 2	Field_1	Field_2	...	Field_n
	Value_1	Value_2	...	Value_n
Total_Level 3	Field_1	Field_2	...	Field_n
	Value_1	Value_2	...	Value_n
Total_Level 4	Field_1	Field_2	...	Field_n
	Value_1	Value_2	...	Value_n

(그림 5) 문서내의 수준별 분야 및 빈도수를 저장하기 위한 해쉬테이블의 구조

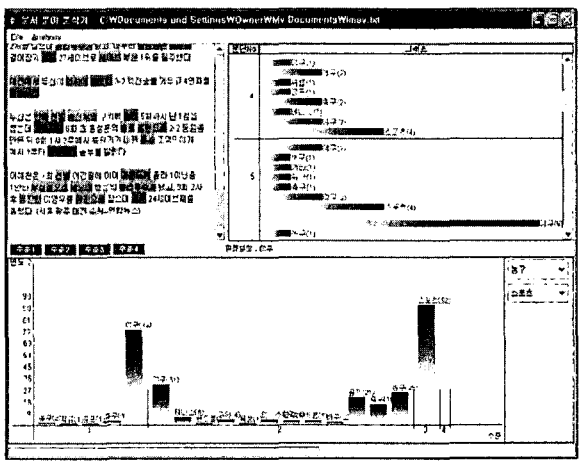
4.4 그래프 표현

그래프는 위 텍스트 분석단계에서 얻어진 해쉬테이블을 이용한다. 먼저, 문단별 그래프는 벡터로부터 해쉬테이블 묶음을 읽어와 수준 1부터 수준 4까지의 해쉬테이블 내용을 차례로 읽어들여 막대그래프로 표현된다. 이 과정에 앞서, 그래프 영역은 한정되어 있고, 막대그래프의 크기는 유동적이므로 우선 그래프의 막대 중 가장 긴 막대의 크기(빈도)를 이용하여 막대의 1단위 크기를 미리 계산한다. 아래 <표 1>은 문단별 그래프 출력을 위한 알고리즘 예이다.

<표 1> 문단별 막대그래프 출력 알고리즘 예

```
stick_scale = 그래프영역 너비/가장 큰 빈도수
while(vSubhash has More){
    for(int i=1;i<=4;i++){
        Hashtable tmp = 수준 i번째 해쉬테이블;
        변수 fild = tmp.nextKey;
        변수 val = tmp.nextValue;
        stick_length = val * stick_scale;
        막대그래프 도시;
        그래프 영역 + 막대 위치 + 10;
        그래프영역 높이 재지정
    }
}
```

전체문서에 걸쳐 수준별 분야의 그래프는 <표 1>의 문단별 막대그래프 출력 흐름도의 for문 부분을 차용하여 변형한 후 출력한다. 아래의 (그림 6)은 텍스트 분석과 막대그래프 출력 화면이다.

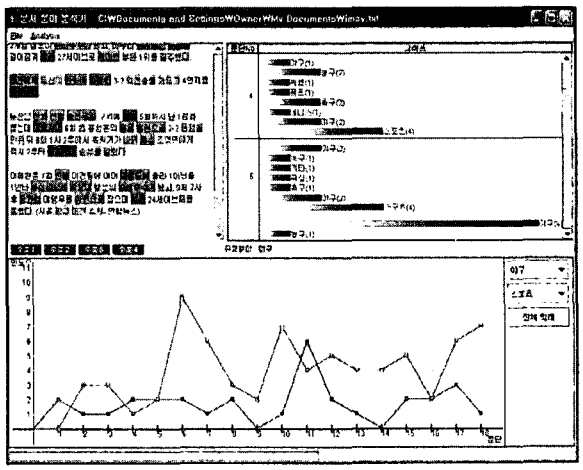


(그림 6) 텍스트 분석과 막대그래프의 출력 화면

4.5 꺾은선 그래프를 이용한 분야의 분포분석

꺾은선 그래프는 분석대상 문서에서 특정 분야의 분포도를 분석하기 위해 제공된다. 수준 1과 수준 2가 종단분야이므로 현재 문서에 포함된 종단분야의 리스트를 콤보박스에 넣고 분석하고자 하는 분야를 선택하여 해당분야의 분포 상황을 한눈에 볼 수 있다.

구현은 단원별 그래프 출력에 사용되었던 벡터와 sub_Level 해쉬 테이블을 활용하여, 분석자가 선택한 필드만을 추출하여 꺾은선 그래프로 출력한다. 흐름도는 위의 막대그래프와 유사하므로 생략한다. 아래의 (그림 7)은 막대그래프 출력 후 분석자가 전체 막대그래프의 우측에 위치한 분야 선택에서 "야구" 분야를 선택했을 때 결과를 출력한 화면이다.



(그림 7) 꺾은선 그래프 출력 화면

5 결론

단락검색은 사용자가 작성한 질의어에 대해 정확하고 빠르게 동작하며 동시에 검색과 무관한 정보는 신속하게 차단할 수 있다. 또한 단락검색은 사용자가 원하는 정보의 존재여부를 빠르게 지시한다. 본 논문은 텍스트의 특정 화제분야를 대표하는 실마리로서 분야연상어를 이용하였기 때문에 인간의 두뇌 혹은 인지작용과 유사하게 컴퓨터가 텍스트를 읽어감에 따라 텍스트가 어느 분야에 속하는지 빠르게 판단한다. 또한 본 연구는 단락검색 시 화제의 전환성과 계속성을 고려하여 복수의 분야에 대해 언급된 문서상에서 분야를 손쉽게 분리할 수 있고, 반대로 동일분야의 텍스트가 분리되는 현상을 방지하여 복수 분야에 속하는 텍스트의 중복을 제거하는 한국어 문서의 단락검색에 기초 연구자료로 이용될 수 있다.

참고문헌

[1] 김숙영, 최창원, 이상곤, "한글문서분류용 분야연상어의추출 알고리즘", 한국정보과학회 2003 가을 학술발표 논문집(Ⅰ), 제 30권, 제 2호, pp. 544-546, 2003.

[2] 이상곤, "분야연상어를 이용한 화제의 계속성과 전환성을 추적하는 단락분할 방법", 정보처리학회 논문지B, 제 10권, 제 1호, pp. 57-66, 2003.

[3] 이상곤, 이완권, "분야연상어의 수집과 추출 알고리즘", 정보처리학회 논문지B, 제 10권, 제 3호, pp. 347-358, 2003.

[4] 이원휘, 최현, 이상곤, "분야연상어 추출 방법의 설계 및 구현", 한국정보처리학회 춘계학술발표 논문집 제 11권 제1호, pp. 651-654. 2004.

[5] 홍성옥, 이상곤, "연상정보를 이용한 단락분할 방법", 2003년도 한국정보처리학회 춘계 학술발표 논문집(상), 제 10권, 제 1호, pp. 497-500, 2003.