자동 키워드 제작기 시스템 설계

이창열, 강현규, 장호욱, 박세영
한국전자통신연구소 언어정보연구실

A Design of the Automatic Keyword Maker

ChangYeol Lee, Hyunkyu Kang, Howook Jang, SeYoung Park
Language Information Sec. ETRI, lcy@modu.etri.re.kr

요약

본 논문에서는 대규모 텍스트 데이타 베이스를 구축하거나 전자 도서를 구축할 때 중요한 정보에 관한 화일 구축과 정보 검색시 필요한 자동 키워드 제작기의 설계에 대하여 논하였다.

자동 키워드 제작기는 명사 사전과 조사 사전의 도움을 받아서 명사 및 복합 명사를 추출하고 중요한 키워드를 자동으로 색인하는 과정을 설계하였으며 이들 검색에 필요한 속도 및 정확도 향상에 중점을 두었다.

I. 서론

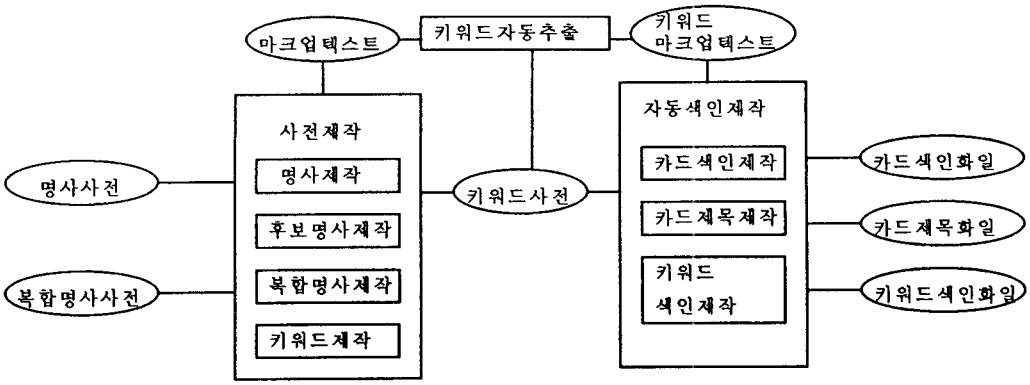
본 논문은 전자도서나 대규모 텍스트 데이타 베이스를 구축할 때 많은 정보의 효율적 관리 및 검색시 필요한 자동 키워드 제작기 설계에 관하여 기술한 것이다. 검색의 매개체인 키워드를 구축하는 과정은 다음과 같다.

우선 대상이 되는 소오스에 대한 명사와 그 빈도를 구축하고 복합 명사를 구축한다. 이들 명사나 복합 명사로 부터 중요한 단어를 키워드로 선정한뒤 선정된 키워드가 포함되어 있는 카드(사용자의 질의시 보여주는 최소 단위. 논리적으로 볼때 마크업 구조의 화일에서 마크업의 최소단위와 일치한다.)에 대한 역 화일을 구축하는 것이다.

이렇게 선정된 키워드는 사용자가 질의시 검색의 대상이 되고 최종적으로 원하는 정보를 제공하는 매개체가 되는 것이다.

본 논문에서는 '세계를간다'[1] 책자에 대하여 본 시스템을 적용하여 설계한 것으로 II장에서는 키워드 제작기의 구조를 III장에서는 전체 설계 과정으로서 명사, 복합 명사, 키워드 구축에 관하여 기술하고 그리고 IV장에서는 결론을 기술한다.

II. 키워드 제작기 구조



<그림 1.> 키워드 제작기의 구조

키워드 제작기는 사전 제작 모듈과 자동색인 제작 모듈, 그리고 사용자 인터페이스로 구성된다. 사전 제작 모듈에서는 마크업 텍스트에 대하여 해당 명사, 복합명사를 추출한 뒤 키워드 선택 알고리즘을 이용하여 키워드를 선정한다. 선정된 키워드는 다시 마크업 텍스트에 마크를 하여 줌으로써 나중에 마크된 키워드를 밝게 보여 주거나, 사용자가 키워드를 선택할 경우 관련된 내용을 검색(네비게이션)할 수 있게 한다.

자동색인 제작 모듈에서는 마크된 텍스트에 대하여 그의 계층상의 구조에 따른 카드들을 생성한다. 또한 정보 검색의 효율을 높이기 위해 계층 구조에 따른 카드들의 제목을 생성하며, 정보 검색을 위한 키워드 색인으로써 역색인 화일을 생성 한다[5].

III. 자동 키워드 제작기 설계

3.1. 명사 구축

명사만들기는 제공되는 소오스에 있는 명사 리스트를 작성하는 것이다. 기존에 존재하는 전역(Global) 명사 사전을 사용하여 명사 화일('명사 + 빈도'구조로 됨)을 만들고, 명사 사전에 등록되지 않은 단어 중에서 조사 사전을 이용하여 후보 명사 화일('후보명사 + 단어 + 라인' 구조로 됨)을 만들수 있다. 후보 명사는 추후 수동적 검증을 통한후 정식 명사로 등록된다.

"프랑스 50"은 명사 화일의 내용 예제이며,
"하조대 하조대에서는 강원도에 있는 하조대에서는 많은"은
후보 명사 화일의 예제 내용이다.

3.2. 복합명사 구축

복합명사는 명사사전에 등록된 단어들의 의미있는 나열이다. 복합 명사 처리의 목적은 정보 검색에 있어서 정확도를 높이기 위하여 필요하며 명사사전에

있는 단어의 나열에서 복합명사를 찾기 위하여는 구문적, 의미적 처리를 필요로 하나, 여기서는 주로 구문적 처리에 중점을 둔다. 왜냐하면 의미적 분석은 완전한 품사 사전이 필요로 하며, 많은 사람의 수동적인 도움없이 해결하기가 요원하기 때문이다.

3.2.1. 패턴 분류

복합 명사는 단어 사이의 순서를 간직하고 있으며 전이적이지 않다. 복합 명사의 구성 패턴은 다음과 같이 분류될 수 있다[2].

명사 2개(2개) :	명사 + 명사 (타입 1), 명사 + '의' + 명사(타입 2)
명사 3개(4개) :	명사 + 명사 + 명사(타입 3),
	명사 + '의' + 명사 + 명사(타입4),
	명사 + 명사 + '의' + 명사(타입5),
	명사 + '의' + 명사 + '의' + 명사(타입 6)
명사 4개(8개) :	명사 + 명사 + 명사 + 명사(타입 7),
	명사 + '의' + 명사 + 명사 + 명사(타입 8), ...
	명사 + '의' + 명사 + '의' + 명사 + '의 + 명사(타입 14)
명사 5개(5개) :	명사 + 명사+ 명사 + 명사 + 명사(타입 15), ...
	명사 + 명사 + 명사 + 명사 + '의 ' + 명사(타입 19)
기타 (타입 20).	

위와 같은 분류에 해당되는 타입을 전부 분류한뒤 <표 1>과 같은 의미 분석을 하여 통계적 자료로 사용한다. 명사 사이에 있는 '의' 대신에 '을/를'을 이용하여 분석하여도 가능하다.

```
void    compnoun dict_n index  source dict_c
        입력 : 명사 사전, 명사 사전 인덱스, 소오스, 복합명사 화일 이름
        결과 : 복합명사 화일

{       초기화;
        do_loop(소오스로 부터 한라인씩 읽는다.) {
                화일의 끝이면 빠져나간다.
                /* 명사사전을 검색하고 출력 결과의 타입을 결정한다 */
                type = look(dict_n, index, word);
                if(type == 0)  if(noun > 2) 복합 명사 발견 ;
                if(type == 1)  저장; /* "조사 없는 명사" 형태 */
                if(type == 2)  저장 ; /* "의"가 포함된 형태 */
                if(type  == 3) if(noun > 2) 복합명사 발견; /* 조사가 포함된 타입 */
                if(type >= 1) noun_++; else noun = 0;
        } end_of_do_loop
}
```

3.2.2 패턴 의미 분석

'타입'은 명사 분류상 타입을 말하며, '선택'은 복잡한 형태의 복합 명사가 의미를 가지는 또다른 복합 명사를 내포할 수 있으므로 더 세분화된 분류인 것이다. '제거'는 복합 명사로써 의미를 상실한 단어이며, '기타'는 명사의 가치를 상실한 단어를 말한다.

샘플	타입	선택	제거	기타
프랑스 혁명	1			
원자로 등 첨단 무기	7			등
겨울의 파리	2			
프랑스 물건의 유명 마크	9	프랑스 물건 유명 마크		
대한 정보	1		제거	대한

<표 1.> 복합 명사 추출 과정표

3.2.3. 패턴 통계 분석

가. 타입 분석

전체 7,465개의 복합 명사 후보에서 99.9%에 해당되는 7,456개의 복합 명사 후보를 각 타입별로 분류하는 기능이 필요하다.

타입	타입1	타입2	타입3	타입4	타입5
갯수	4,363	1,684	525	229	384
퍼센트	58.4	22.6	7.0	3.1	5.1
타입	타입6	타입7	타입8	타입9	타입10
갯수	21	64	40	48	41
퍼센트	0.3	0.9	0.5	0.7	0.5
타입	타입11	타입12	타입13	타입14	타입15
갯수	0	6	7	1	7
퍼센트	0.0	0.1	0.1	0.0	0.0
타입	타입16	타입17	타입18	타입19	타입20
갯수	6	4	13	13	0
퍼센트	0.1	0.08	0.2	0.2	0.12

<표 3> 복합명사의 타입 분석

타입 1, 2, 3, 4, 5, 6, 7, 8,9, 10가 전체 99.1%를 차지한다. 명사가 2개, 3개인 경우가 96.5%, 4개 인것을 포함할때 99.3%를 차지한다.

나. 길이 분석

'세계를 간다'[1]에 대하여 위의 방법을 시도하여 얻은 복합명사 후보는 7,465개이며 그중 위의 패턴에서 명사가 포함된 갯수에 따라 분류하면 <표 2>와 같이

나온다. 이 단어에는 현재 의미가 없거나, 명사 사전의 잘못된 엔트리로 인하여 잘못 선정된 명사, 동음이의어 문제에 대하여 처리를 하고 있으나(각 타입 별로 구별한뒤 수동으로 오류 패턴을 찾아서 통계 처리) 결과에 대하여는 많은 수동적 작업으로 인하여 요원하다.

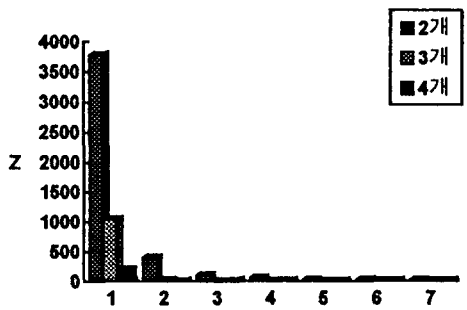
명사갯수	2	3	4	5	6	7	총계
갯수	6,047	1,159	207	43	8	1	7,465
퍼센트	81	15.5	2.8	0.58	0.11	0.01	100

<표 2> 복합명사에 포함되는 명사 갯수와 비율

명사가 4개 이하로 복합된 복합명사는 전체 명사의 99.3%, 5개 이하는 99.9%를 차지 한다.

다. 빈도 분석

빈도분석은 '의'를 제외한 상태에서 한다. 즉 모든 예비 복합 명사를 기본형 (타입 1, 타입 3, 타입 7, 타입 15)으로 변경한다. 그뒤 각 복합 명사에 대하여 빈도 표 <표 4>를 구성하고 빈도 분석을 통한 복합 명사 후보를 선정한다.



<표 4> 복합 명사의 빈도 분석

명사가 2개로 구성된 복합 명사의 빈도는 1회-3,781, 2회-391, 3회-97, 4회-48, 5회-26, 6회-20, 7회-11, 8회-6, 9회-7, 10회-4, 11회-3, 12회-1, 13회-4, 14회-1, 15회-2, 16회-1, 17회-0, 18회-1, 19회-1, 20회-4, 24회-1, 26회-1, 36회-1, 38회-1, 39회1, 63회-2이다.

명사가 3개로 구성된 복합명사의 빈도는 1회-1,042, 2회-26, 3회-7, 4회-2, 5회-1, 8회-1, 11회-1, 12회-1이다.

명사가 4개로 구성된 복합 명사의 빈도는 1회-199, 2회-4개이다. 그외 명사가 5개인 복합 명사는 1회가 43, 6개인 것은 1회가 8, 7개인것은 1회가 1개이다.

3.3. 키워드 만들기

키워드는 명사나 복합명사 중에서 구한다. 구하는 방식은 다음과 같다.

첫째 빈도수로 구한다. 전체 소오스에서 명사와 복합 명사의 출현 빈도를 사용하여 키워드로 정한다. 출현 빈도는 단어의 가치를 나타낼수 있기 때문이다. 너무 자주 출현하거나 1번 정도 출현하면 그의 중요도는 감소된다. 그러므로 여기서는 2회이상 20회 이하를 만족하는 명사와 복합 명사를 선정한다. 5,750개의 유일한 복합명사중에서 669개가 그 조건을 만족하는 복합명사이다.

둘째 각 카드별로 키워드의 최소, 최대 갯수를 정한다. 이 이유는 한카드에 키워드가 하나도 없는 경우도 나타날수 있으며 너무 많은 경우도 있기 때문이다. 그러므로 적절한 키워드 갯수를 유지하여야 한다.

3.4. 키워드 추출

키워드 추출이란 입력 텍스트에서 키워드 사전의 키워드와 일치하는 단어에 대해서 마크를 해 주는것을 말한다. 마크된 단어는 최종적으로 노드를 밝게 보이게 하여 사용자에게 또다른 연산(예를들어 네비게이션)을 제공하는 역할을 한다. 키워드 추출 알고리즘 수행 결과[3] 샘플은 다음과 같다.

입력화일 : 러시아의 곡창 지대는 러시아 평원이다.
출력화일 : \러시아\의 \곡창 지대\는 \러시아 평원\이다.

동음이의어 문제 : \이\ 변화,\등\을,\한\ \가지\

3.5. 카드 색인

카드 색인은 시스템에서 카드를 빠르게 찾아내고 처리하도록하기 위하여 카드의 상하 구조, 제목 및 카드의 색인 정보를 포함한 화일을 구성하는 것이다. 텍스트의 계층구조에 따른 마크업이 되어 있으며, 이 마크업에 따라 카드의 마크업 텍스트에서 시작주소, 크기, 계층구조에 따른 제목등을 기록하게 한다[4].

void **parse** *source* ; 카드의 계층적인 구조를 분석한다.
입력 : 마크업 텍스트
결과 : 카드의 계층적 구조를 분석한 정보

void **append_cardstruc** *cardf cardidx*
; 카드 계층구조에 따른 제목 및 카드번호 주소를 만든다.
입력 : 카드의 계층적인 구조 정보
결과 : 카드의 구조 정보 및 인덱스

3.6. 카드 제목 만들기

정보 검색의 효율을 높이기 위하여, 상위 카드들의 제목 정보가 하위 카드에 상속되도록 한다. 이는 비록 특정 하나의 카드에는 정보가 들어있지 않지만 상위 카드의 제목에 이러한 정보가 존재하는 경우 이를 이용할 수 있다.

마크업 텍스트로 부터 상위 카드의 계층을 분석한 후 이 계층 구조에 따른 상위 제목들을 자동 생성한다. 자동 생성된 카드 계층 구조에 따른 제목들은 마크업

텍스트 화일의 계층에 따라 제목들을 삽입한다. 나중에 키워드 색인시에 이러한 제목 정보는 함께 고려되어 키워드 색인이 될 수 있다.

3.7. 키워드 색인

정보를 검색하는 경우에 어떠한 키워드에 대하여 어떻게 검색을 할 수 있을지를 정하는 것으로 색인의 알고리즘에 따라 시스템의 효율에 영향을 미친다. 특정 키워드의 중요도를 나타내는 무게 Wij는 임의의 카드 i에서 키워드 j의 출현 빈도로 나타낼 수 있다.

중요도가 결정된후 모든 키워드에 대하여 "키워드 + 총 나타난 카드 갯수(n) + 카드 번호(첫번째) + 중요도 + 카드번호(두번째) + ... + 카드번호(n번째) + 중요도"의 역 색인 화일을 작성한다.

IV. 결론

키워드 제작기는 온라인 도서에 대하여 전자도서를 만들어 주는 시스템이다. 즉 책에 나타나는 단어에 대하여 키워드를 선택하고 각 키워드가 있는 위치에 대한 정보를 간직하고 있다가 사용자가 원할시 해당 키워드가 있는 곳을 보여줄 수 있다.

키워드 제작기를 구축하기 위하여는 사전 제작 도구, 자동 색인 제작 도구가 필요하며 사용자 인테페이스 구축을 필요로 한다. 키워드 제작기는 온라인 도서를 전자화 하는데 필수적 과정으로 복합 명사의 자동 선택 루틴, 후보 키워드 설정 방법, 키워드 중요도 선정이 시스템의 정확도에 많은 영향을 미치며 키워드 추출은 사용자 인터페이스에 도움을 주는 부분으로 이의 개선에 중점을 두었다. 앞으로 마이크로 소프트 윈도우스상에서 구현될 예정이다.

참고 문헌

[1] 일본 다이아몬드 Big사. 세계를 간다 :유럽 14개국, 중앙일보사. 1991.

[2] 이창열, "복합명사 추출 알고리즘에 관한 연구", TM93-6330-44. ETRI, 1993.

[3] 장호욱, 강현규, "키워드 추출 알고리즘의 개선", TM93-6330-48, ETRI, 1993.

[4] 강현규, "관광정보 검색 시스템(TIRS)에서의 자동 색인 실험 결과 보고서", TM92-6330-62, ETRI,1992.

[5] 강현규외 6인, "전자도서 제작 시스템 기능 규격 정의서", BLK-KB032-1.0, ETRI, 1998.