동의어와 유의어 개념에 기반 한 키워드 추출기의 설계 및 구현

박 은 석, 박 현 진, 이 상 곤
전주대학교 컴퓨터공학과 언어과학실
pes81sp@nate.com, {eungkah, samuel}@jj.ac.kr

Design and Implementation of Keyword Extractor based on Synonyms and Related Terms

Eunsuk Park, Hyunjin Park and Samuel Sangkon Lee
Language Science Lab.,
Dept. of Computer Science & Engineering,
Jeonju University

요 약

인간은 문서를 읽고 그 내용을 머릿속에서 개념적으로 정리하여 몇 개의 명사를 이용하여 키워드로 인지한다. 본 논문은 이러한 점에 착안하여 문서를 대표하는 키워드를 추출하는 시스템을 설계하고 구현하였다. 본 논문에서는 단어의 개별적인 개념 정보를 동의어와 유의어 사전을 통해 주요 개념어를 추출하고, 추출된 개념어들 사이의 공기 관계를 계산하여 키워드로써의 중요도를 계산하고자 한다. 이를 통해 문서를 대표할 수 있는 키워드 후보를 생성하는 생성 규칙을 자동화하고 문서를 잘 대표할 수 있는 키워드 추출기를 제안하였다.

1. 서 론

사람이 문서를 읽고 그 내용을 개념상으로 정리하는 일련의 프로세스를 살펴보면, 문서에서 출현하는 주요 명사를 이용하여 문서를 대표할 수 있는 복합 단어로 구성된 키워드를 생성[3]하고 이를 기억한다. 그러나 문서 내에 적절한 키워드(주제어)가 존재하는 경우는 추출이 가능하지만, 존재하지 않는 경우에는 적절한 추출이 불가능하다. 본 논문에서는 문서에서 적당한 키워드가 출현하지 않은 경우에도 적당한 추출이 가능하도록 단어의 개념에 기반 하여 복합어로 구성된 키워드를 추출[4]할 수 있도록 하는 기술을 확장하여 개발하고자 한다. 다음 장에서는 본 연구의 중요 과제인 사전 구조에 대해 설명하고, 3장에서는 키워드 추출 방법에 대하여 기술한다. 마지막으로는 결론과 향후의 연구 과제에 대하여 논의한다.

2. 지식 사전의 구조

본 시스템의 지식베이스가 되는 지식 사전의 자료 구조는 B+ Tree[7]로 구축하였다. <그림 1>은 "가가호호"의 경우가 구성된 예를 나타낸다. 이 구조는 크게 인덱스(Index), 구성된 트리의 노드(Tree Node), 스트링 데이터(Data) 등 세 가지 층으로 구분하여 구성하였다. 다음에 각각의 층에 대해 설명한다.

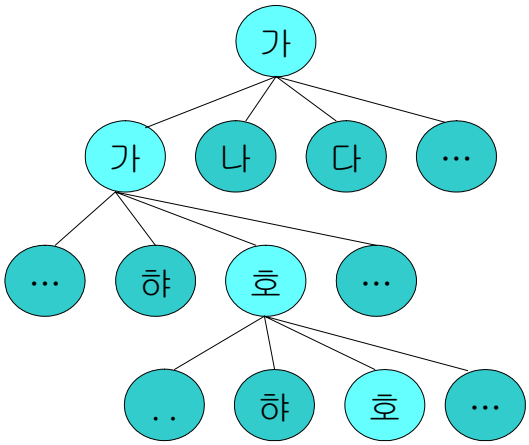
● 인덱스(Index)

인덱스 층은 하부에 구성된 B+ Tree의 개수가 기록되어 있으며, 다음 노드에 대한 트리 진입점의 위치와 이에 해당하는 음절 정보가 저장되어 있다. 예를 들어, 어떤 단어

<표 1> 인덱스 층의 구성 예

100	1	600	갸	1500	거	3500	겨	5600	…

가 <표 1>과 같은 트리가 구성된 경우, 숫자 100은 전체 트리의 수를 나타내며, 다음의 숫자 1은 음절 정보, 600은



<그림 1> 지식 사전의 구조도

트리의 진입점 위치이다. 그 다음의 데이터 스트링 "갸, 거, 겨" 등은 음절 정보를 의미하고, 각각 2바이트를 차지한다. 이 때, 숫자 1은 1 바이트이지만 실제로 데이터 파일을 기록할 때는 모든 1 바이트의 문자 앞에 널(NULL, ' ') 값을 삽입하여 음절 정보와 동일한 크기인 2 바이트로 수정하여 저장한다. 이는 바이너리 탐색을 가능하도록 하여 트리 진입점을 동일한 자료구조로 검색하기 위한 조치이다.

● 트리 노드(Tree Node)

트리의 각 노드는 최소 정보로 구성하고, 모든 노드의 크기를 일정하게 구성하였다. 각 음절 별로 세 개의 서로 다른 노드가 생성되고, 이 노드들은 부모와 자식 관계를 갖도록 하였다. 노드의 정보는 ① 음절 정보, ② 단어의 생성 가능 여부, ③ 자식의 수, ④ 첫 번째 자식의 위치, ⑤ 데이터 층의 위치, ⑥ 부모 노드의 위치 등을 기록하고 있다.

<표 2> 트리 노드의 구성 예

한	0	12	14500	98600	0
국	0	5	24300	105480	14200
어	1	2	32300	140490	20440

예를 들어 "한국어"의 트리 구성 예가 <표 2>와 같이 구성되었다면, 노드 "한"의 경우, "한"에 해당하는 한 글자만으로는 의미 있는 단어가 될 수 없으므로 의미 있는 단어의 생성이 가능한가를 판가름하는 여부(이 경우 '0')를 기록하고, 자식의 수(12)와 첫 번째 자식의 위치(14500)를 이용하여 자식 노드를 탐색한다. 파일 기록 시, 한 노드의 모든 자식 노드는 정렬되어 순차적으로 기록되며, 각 노드의 크기가 동일하기 때문에 자식 수와 첫 번째 자식의 위치만을 알고 있으면 빠른 탐색이 가능하다. 따라서 이진 탐색을 이용해 모든 자식 노드를 빠르게 탐색할 수 있다. 노드 "어"의 경우 루트에서 탐색을 시작하여 하향 순회하면 "한국어"가 되어 의미 있는 단어로 생성될 수 있기 때문에 "단어 생성 가능 여부"를 나타내는 플래그에 1이 기록되어 있고, "한국어"의 한문 정보는 다음에서 설명하는 데이터 층의 위치를 기록하였다.

● 데이터(Data)

데이터 층에는 이전에 설명한 두 가지 층에 기록되지 않은 부수적인 정보들을 모두 기록한다. 각 개별 사전마다 데이터 층의 정보가 서로 상이하기 때문에 데이터 층을 별도로 구성하였다.

2.1 사전별 특징

본 연구의 지식 베이스로 구축한 사전은ⓐ 동의어와 유의어 사전, ⓑ 명사 사전, ⓒ 전문 분야별 사전 등 3가지의 사전을 구축하였다. 동/유의어 사전은 부록 I에 샘플을 제시하였다. 문서에서는 동일한 의미로 사용되었지만 다른 형태의 형태소가 출현한 명사들의 동의어와 유의어 정보를 저장하고 있다. 명사 사전은 개념어 사전을 통해 얻어진 개념어 정보 중 명사만을 추출하는데 이용하고, 부록 II에 제시한 바와 같이 전문 용어 사전은 기술적인 문서의 해당 분야를 판정하고 분야에 맞는 용어를 키워드로 추출하기 위해 구축하였다.

● 명사 사전

명사 사전의 구성 엔트리는 국어사전[5, 6]에 등록된 103,425개로 구성하여 Tree Node 층만을 유지한다. 따라서 명사 사전의 Data 층은 존재하지 않는다. 한 어절에서의 명사의 검색 방법은 조사를 제거하는 방식을 사용하지 않고, 명사 사전과의 완전 일치를 우선으로 하여 검색하며, 완전한 일치(perfect match)가 이루어지지 않은 경우에는 최장 일치 검색 기술을 통해 명사를 추출하였다.

때로는 동의어 및 유의어 사전에 등록되어 있지 않아 중요한 명사의 추출이 탈락하는 경우가 발생한다. 이를 방지하기 위해 일반 명사 사전은 반드시 필요하다.

● 동의어 및 유의어 사전

동의어/유의어 사전의 데이터 층 정보는 ㉮ 동/유의어, ㉯ 결합 시 우선 순위 여부, ㉰ 개념어의 수, ㉱ 개념어 노드의 위치 정보, ㉲ 한문이나 ㉳ 영문 정보로 구성하였다. 한 단어에 대한 개념어가 두 개 이상 존재 할 수 있기 때문에 개념어의 개수를 계속 유지하고 있어야 한다. 이 개

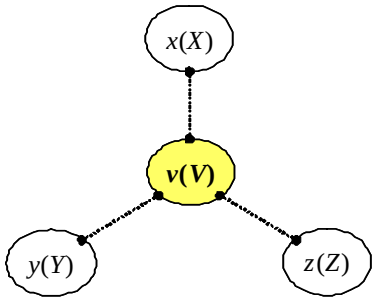
넘어의 수만큼 개념어 노드의 위치 정보를 모두 읽는다

● 전문 분야별 사전

전문 분야 사전은 인덱스 층 이전에 분야 수와 분야 명이 구분되어 기록되어 있다. 노드 정보는 위의 사전들과 동일하지만, 사전으로 생성되는 내용을 하나 이상의 어절로 구성하였다. 예를 들어, 추출된 키워드 후보가 '식별 코드, 신경 세포 모델, 신경망 사전'의 경우, 키워드 생성 시 띄어쓰기에 의해 사용된 공백까지 모두 단어 사전에 포함하여 여러 어절로 구성된 전문 분야 용어도 검색할 수 있게 구성하였다. 사전 검색 시, 명사 사전에 검색된 명사만을 가지고 검색하면 문서에 "정보의 검색"라는 단어가 존재하여 "정보 검색"라는 전문 용어도 아울러 검색하고 이 단어에 대한 분야 정보도 추출한다. 이 경우, 키워드 추출 시 "정보"와 "검색"과 같이 두 개의 단어로 분리되어 추출하는 것보다 복합어 "정보 검색" 혹은 "정보검색"이 키워드로 적당하기 때문에 여러 개의 어절을 하나의 복합어로 병합한다. 또한 데이터(Data) 층에는 분야 명 배열의 인덱스들이 기록되어 문서에서 전문 용어가 검색되면 인덱스 숫자를 누적하여 해당 분야를 추정[2]하는데 사용한다.

3. 키워드 추출 방법

키워드 추출기의 정확도를 향상하기 위해서는 중요도 계산 방법이 필요하며 중요도 값의 계산[1]은 참고 문헌 [1]이 제시하고 있는 단어의 개념간 거리(CD; Conceptual Distance)가 필요하다.



(그림 2) 공기 관계의 수
(참고 문헌 [1]에서 인용)

3.1 공기 관계

중요도의 지표로 개념어의 각 요소를 포함하고 있는 문장간의 거리를 이용할 수 있지만, 이 거리만으로는 개념어들 사이의 관련성을 정확히 파악할 수 없다[1]. 따라서 개념어의 공기 관계에 주목하여 개념 간의 거리를 이용한다. 어떤 단어 V, X, Y, Z가 출현 한 경우, X와 Z에 공통으로 공기하는 개념어 V를 사이에 두고 X에서 Z의 개념어 간 거리를 계산한다. CD는 공통 개념어의 수가 1이다. 각 개념어간 거리는 VX는 같은 문장 내에서 공기하고 있으므로 1이고, XZ 또한 마찬가지로 구할 수 있다. 그러므로 XZ의 개념어간 거리는 2가 된다.

3.2 공기 관계의 수

주제를 대신하는 개념어 혹은 개념어의 각 구성 요소는 문서 중에 자주 출현한다는 가정을 이용하면 이들 개념어는 다른 단어들과의 많은 수의 공기 관계를 갖는다. 따라서 다른 단어와 많은 수의 공기 관계를 갖는 개념어가 문서의 주제를 잘 대표한다고 사료 되고 이러한 단어를 문서의 적절한 키워드로 추출한다.

예를 들어, 어떤 문서에서 i번째의 복합어 w의 공기 관계 수를 w_i라 하면 개념어가 V가 갖는 공기 관계 수는 3(=1+1+1)이 된다. 또한 X, Y, Z는 모두 1이 된다. 따라서 V는 X, Y, Z보다 중요도가 높다. 특정한 명사에는 의미는 같지만 문서에는 적합하지 않은 개념어 후보들이 수 없이 생성될 수 있다.

<표 3> 개념어의 후보들

출현 단어	개념어 후보
산출	해산, 셈, 제작, 제조, 산출, ...
관리	간수, 사관, 관구, 서규, 담당, ...

개념어 사전을 통해 얻어진 많은 개념어 후보를 문서의 성격에 적당한 개념어 후보로 요약할 필요가 있다. 많은 개념어들 간의 공기관계 수를 모두 계산하는 비효율을 줄이고, 문서에 적합한 개념어 후보를 생성하기 위해서 이 작업은 반드시 필요하다. 문서에 출현한 어절과 개념어 후보들을 빠르게 비교하여 하나의 개념어로 압축하는 기술이 필요하다.

3.3 중요도의 계산

개념어간 거리와 개념어의 공기 관계 수를 고려한 키워드 후보의 중요도를 계산하는 식은 주석 1)에서 제시된 방법을 이용하였다. 위의 식은 개념어간의 거리가 작을수록, 공기관계 수가 많을수록, 개념 요소에 대한 동의어 및 유의어의 빈도가 높을수록, 후보어의 중요성은 높아진다. 중요도 계산은 다음과 같이 두 가지로 나누어 설명한다. 첫째, 개념어 후보를 중심어로 선택하고, 중심어와 공기하는 단어들을 대상으로 계산한다. 둘째, 문서에서는 중요하지

1) 중요도(I) 계산식은 다음과 같다. 자세한 내용은 참고 문헌을 참고하여야 한다.

$$I = \left[\frac{1}{n \times \alpha d}\right] \times \sum_{i=1}^{n} [(\{(S(W_i) \times \alpha) + (R(W_i) \times \beta)\})]$$

만 개념어 사전에 존재하지 않는 명사를 중심어로 하고, 이 단어와 공기하는 단어들과의 중요도를 계산한다. 중요 명사를 추출하는 것은 문서의 공기관계 수가 가장 많은 명사를 주요 구성 요소로 이용하였다.

3.4 키워드의 생성

키워드는 중요도 계산 값을 이용하여 생성한다. 중요도 계산에 이용한 임계치(threshold)는 문서의 길이, 키워드 후보의 출현 빈도 등에 의해 많이 변경된다. 이것은 향후에 계속 연구하여야 하지만, 이번 연구에서는 수 차례의 실험 후 얻은 임계치 0.5를 설정하였고, 사용자의 개별적인 설정에 따라 0.3~1.5까지의 임계치 변경을 통해 후속에 랭크되는 다른 후보어를 제시할 수 있도록 설계하였다. 또한 복합 명사로 구성된 키워드는 각 구성 요소들의 순서가 문제가 되는데, "-하다", "-되다", "-시키다"가 붙을 수 있는 명사를 별도로 추출하였다. "하다"가 붙을 수 있는 단어는 총 수는 12,767개, "되다"는 519개, "시키다"는 163개 등 총 14,131개가 조사되었다. 따라서 이러한 명사들은 복합어 구성 시, 맨 뒤에 부착하여 복합 키워드를 생성하였다. 다시 말하면, 위의 단어 리스트에 포함되는 단어는 복합어 생성 시, 맨 뒤로 붙여 생성하여 생성된 복합어가 의미가 잘 통하도록 고려하였다.

4. 결론

실험에 사용한 문서 집합은 각 분야별 논문의 초록을 이용하였다. 논문의 저자가 자신의 논문에 정의한 키워드를 그대로 이용하여 시스템이 추출한 키워드와 비교하였다. 참고 문헌 [8]에서 제시한 정보과학회 2000년 춘계 학술대회 논문 초록 441개와 2001년 춘계학술대회 512개 등 총 953개의 문서 집합으로 구성되어 있다. 각 초록 문서는 최대 10개의 문장으로 구성되어 있다.

본 시스템은 가장 중요도가 큰 값을 갖는 키워드를 인간에게 알려주어, 사용자가 그 문서를 읽을 것인가를 빠르게 판단하도록 제시하는 장점이 있다. 전문 분야 사전을 이용하여 분야를 추정하고 복합어로 구성된 키워드를 추출하여 사용자의 판단에 도움을 준다. 개념어들 간의 공기관계, 개념어간 거리, 중요도 계산을 적용함으로서 추출의 정확도를 향상하였다. 향후에는 사전에 등록되어 있지 않은 미등록어의 중요성 측정과 지시대명사가 지시하는 명사를 파악하고, 지시어까지 고려한 공기 관계 계산 방법을 연구하여야 한다.

참고 문헌

[1] 이 상 곤, 이 태 헌, "개념기반 복합 키워드 추출방법", 컴퓨터교육학회논문지 제 6권, 제 2호, pp. 23-31, 2003.

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

[3] Nagata, M. et al., "A Newspaper Keyword Generation Method Based on Key-Concept Extraction.", 제 37회 정보처리 전국대회 논문집 pp. 1030-1031. 1988. (in japanese)

[4] 김 양 선, 이 상 곤, "단어 개념에 기반한 한국어 복합 키워드의 추출", 정보처리학회 논문집, 제 10권, 제 2호, pp. 477-480, 2003.

[5] 남 영 신, 혼+ 국어사전, 성안당, 1997.

[6] 남 영 신, 새로운 우리말 분류 대사전, 성안당, 1994.

[7] Rudolf Bayer et al., "Prefix B-Trees," ACM Transactions on Database Systems, Vol. 2, No. 1, pp. 11-26, 1977.

[8] 컴퓨터 연구 정보센터 http://cseric.cau.ac.kr/new_cseric/main_frame.htm

부 록

I. 동의어와 유의어 사전의 구성 예

Concept(계절(季節)) = [<철, 계후(季候), 절후(節候), 절기(節期), 기절(期節), 사철(四철), 사시(四時), 시절(時節), 사시사철, 시즌(season), 때>, {시기}]...

Concept(삭제(削除)) = [<삭말(削抹), 말소(抹消), 말살(抹殺), 소거(消去), 절제(切除), 할거(割去), 제거(除去)>, {}]...

II. 전문 분야별 사전의 구축

과학기술, 정보통신, 공업용어, 전자공학, 전자산업, 전파통신, 컴퓨터공학(인터넷 용어 포함), 금속공업, 기계공학, 건설 용어, 교량 용어, 항공 용어, 자동차 용어, 철강 금속 용어, 비철 금속 용어,

물리학 용어, 화학 용어, 수학 용어, 생물학 용어, 통계학 용어, 자연 지리학, 해부학 용어, 법률 용어, 방사선 용어, 해양 용어, 환경공학 용어, 플라스틱 용어,

영양학 사전, 패션-비즈니스, 패션머천다이징 및 패션브랜드, 한국어류, 건강/영양식품, 조리용어, 가족 상담 심리 용어 등 34가지의 전문 분야 사전을 구축하였다.