사전 정보에 기반한 효율적인 자동색인기 설계

진정환*, 김태완**
*인제대학교 전산학과
**인제대학교 정보컴퓨터공학부
jjhjia@cs.inje.ac.kr

A Design of Efficient Automatic Indexing based on Dictionary Information

Joung-Hwan Jin*, Tae-Wan Kim**
*Dept of Computer Science, Inje University
**Dept of Infomation & Computer Engineering, Inje University

요약

웹상에 공유되어진 문서의 내용을 대표하는 색인어 추출은 정보 검색 시스템의 질을 좌우한다. 한국어의 자유로운 복합명사나 띄어쓰기 규약, 사전 미등록 어휘 등으로 색인어 추출시 질의어와 색인어 사이의 형태상의 불일치(Syntactic Term Mismatch)가 발생하여 검색성능을 저하시키는 경우가 많다.

따라서 본 논문에서는 사전을 통한 형태소 해석을 통해 단위명사(Unit Noun)로 색인어를 추출하고 사전 미등록어는 N-gram 기반 색인 방법을 이용하여 질의어와 색인어 사이의 부분 일치된 문서도 추출될 수 있는 방법을 제안하였으며, 색인어와 질의어 사이의 유사도 계산을 통해 문서의 우선순위를 정함으로써 색인기의 성능을 높이는 방법을 제안한다.

1. 서론

인터넷의 발달로 인해 많은 웹사이트가 생겨나고 다양한 정보가 문서화된 형태로 공유되어지고 있으며 그 양도 급격히 증가하고 있다. 따라서 공유된 수많은 문서들 중에서 원하는 정보를 효과적으로 검색하기 위한 정보 검색 시스템에 대한 연구가 활발히 진행되어지고 있다.

정보 검색 시스템(Information Retrieval System)은 사용자가 필요로 하는 정보를 수집하여 내용을 분석하고 찾기 쉬운 형태로 조직화하여서 정보에 대한 요구가 발생하였을 때 해당 정보를 찾아 제공하는 시스템을 말한다[6].

정보 검색시 저장된 문서 중에 사용자에게 유용한 정보를 포함하고 있는 문서들을 신속하게 찾아내기 위해 문서를 미리 분석하여 주요어를 추출한 후 찾기 편리한 형태로 저장해 둔다.

어떤 문헌에 대하여 그 문헌의 전체적인 내용을 나타내거나 그 문헌을 다른 문서들로부터 구별할 수 있도록 그 문서의 선택 단서가 되는 단어 또는 단어구 등을 추출하는 것을 색인이라고 한다. 정보검색 시스템은 이 색인이 얼마나 잘 수행되느냐에 따라 그 성능이 좌우된다.

하나의 문서를 표현하는데 있어 주로 사용되어지는 요소는 그 문서를 구성하고 있는 단어들이고, 명사는 문서를 가장 잘 나타낼 수 있는 어휘요소이므로 본 논문에서는 효율적인 명사추출에 중점을 두고 연구를 시도하였다.

자연언어 문장에서 많은 명사들은 여러 가지 형태의 중의성을 가지므로 정확하게 명사를 추출하기 위해서는 복잡한 과정을 필요로 한다. 그러나 정보검색이나 정보요약과 같은 분야에서는 짧은 시간내에 방대한 양의 문서를 처리하고, 특정 장르에 구분없이 처리해야 하는 경우에는 의미 분석과 같은 복잡한 과정을 이용하는 것은 적절한 방법이 아니다[11].

따라서 본 논문에서는 사전을 이용한 빠른 색인어 추출을 목적으로 한다.

사전탐색에 실패한 어절은 신조어나 고유명사, 철

자오류로 인식하고 색인어로 가정한다. 미등록어에 대하여 질의어와 n-gram방식의 색인을 수행하고 질의어와 색인어 사이의 부분일치를 고려하여 유사도를 계산한다. 색인어의 빈도수가 높은 문서에 대하여 우선적으로 검색될 수 있도록 우선순위를 줌으로 사용자는 필요한 정보를 얻는데 소모되는 시간을 최소화 할 수 있다.

2. 관련연구

2.1. 명사추출

명사 추출방법은 크게 세 가지로 나눌 수 있는데 첫 번째로 형태소 분석기를 이용하는 방법, 둘째로 형태소 분석기와 품사 태거를 이용하는 방법, 셋째로 언어 분석 도구를 사용하지 않는 방법이 있다.

첫 번째 방법은 가장 보편적인 방법으로 형태소 분석 결과로 가능한 모든 명사를 추출하는 방법으로 한국어에서는 형식 형태소가 실질 형태소 보다 발달되어 있으며 그 수가 상대적으로 적은 성질을 이용하여 우좌 분석을 수행한다. 이 방법은 재현율은 높은 반면 체언 유형의 분석 중의성 문제로 부정확한 결과가 포함 될 수 있다.

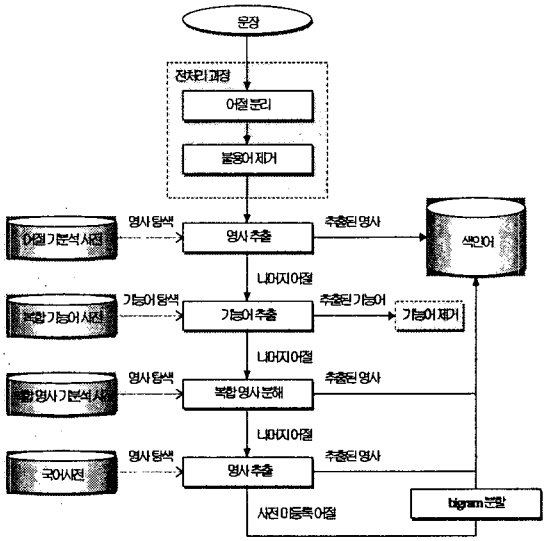
두 번째 방법은 품사태깅을 수행하여 명사로 결정된 단어만을 추출함으로써 분석 중의성 문제를 해결한다. 그러나 품사태거 구축에 많은 시간과 노력이 필요하며 형태소 분석과 태깅단계를 거치는 이중 단계로 분석시간이 오래 걸리는 단점이 있다.

세 번째 방법은 사전에 저장된 어휘 정보만을 사용하여 명사를 추출한다[11]. 타 방법에 비하여 시스템이 단순하고 구현이 쉬우며 분석속도 또한 빠른 장점이 있는 반면 언어 분석도구를 사용하는 방법에 비하여 정확율이 낮은 단점이 있다.

본 논문은 공유된 문서에 대하여 비교적 빠르고 정확한 문서의 색인어를 추출하는 것을 목적으로 하고 있으므로 실험실에 보유하고있는 어절 기분석사전, 복합기능어사전, 복합명사 기분석 사전, 국어사전을 이용하여 명사를 색인어로 추출하고 사전내에 등록되지 않은 미등록어에 대하여 n-gram기반의 색인을 수행하여 검색 효율을 높이는 방법에 대하여 연구한다.

3. 색인어 추출 방법

색인어 추출은 사전을 이용한 형태소 해석을 기반으로, 한 어절에서 기능 형태소를 제외한 실질 형태소가 명사임이 밝혀지면 이를 하나의 색인어로 가정한다. 본 논문에서 제시하는 사전 정보를 기반으로 색인어를 추출하는 과정은 [그림 1]과 같다.



[그림 1] 사전탐색을 통한 색인어 추출 시스템

3.1. 전처리

입력된 문장에 대하여 공백, 마침표, newline 문자, 쉼표, 따옴표를 기준으로 각각의 어절을 token으로 분리한다. 색인어로 선정되기 어려운 수식언(부사, 관형사, 감탄사)이나 수사, 대명사등의 불용어를 제거함으로 사전 탐색시간을 줄인다.

3.2 사전 탐색

제안한 시스템은 네 번의 사전탐색을 통해 사전에 등록되어있는 명사를 색인어로 추출하고 미등록 어절에 대해서는 n-Gram으로 분할하여 색인어로 추출한다.

1 단계 : 어절 기분석 사전을 이용한 명사 추출

말뭉치에서 상위 빈도를 나타내는 어절에 대하여 미리 분석해 놓은 어절 기분석 사전을 이용하여 입력 어절로부터 명사로 태깅된 단어를 색인어로 선정한다. [표1]은 어절 기분석 사전을 이용한 색인어 선정의 과정을 보여준다.

[표 2] 어절 기분석 사전을 이용한 색인어 추출

입력어절	고대, 사회의, 문화, 수준이란, 군사력과, 경제력, 국력을, 의미한다
어절 기분석사전	고대/NN, 사회/NN+의/J, 문화/NN, 수준/NN+이란/J, 군사력/NN+과/J, 경제력/NN, 국력/NN+을/J, 의미/NN+하/SV +ㄴ다/EF
추출색인어	고대, 사회, 문화, 수준, 군사력, 경제력, 국력, 의미

2 단계 : 기능어 분리

색인어로 사용될 가능성이 높은 체언 뒤에는 주로 형식형태소(조사, 어미, 접미사 등)들이 붙어 하나의 어절을 형성하는데 명사를 정확하게 찾기 위해서는 명사와 형식형태소인 기능어들을 분리해야 한다. 어절 기분석 사전에 등록되어 있지 않은 어절에 대하여 복합기능어 사전을 이용하여 기능어를 찾는다.

이를 위해 우좌(right-left) 역방향 분해 알고리즘을 통한 최장일치의 원리(principle of the longest match)로 기능어를 찾고 입력 어절에서 발견된 기능어를 체언과 분리한다. [표2]는 복합명사 기능어 사전을 이용하여 기능어를 분리하는 과정을 보여준다. [표3]은 복합기능어 사전에 수록되어있는 "보다"의 활용형인 복합기능어의 예이다.

[표 3] 복합기능어 사전을 이용한 기능어 분리

입력어절	부모님으로부터나마, 인공지능에서로라도 사람한테서조차도
기능어인식	부모님+(으로부터 나마) 인공지능+(에서 로 라도) 사람+(한테서 조차 도)

[표 4] 복합기능어의 예

보다[보다]	보다가 [보다가]
보다가는[보다가 는]	보다간 [보다가 는]
보다는[보다 는]	보다는야 [보다 는 야]
보다는커녕[보다 는커녕]	보다도 [보다 도]
보다라도[보다 라도]	보다만 [보다 만]
보다만도[보다 만 도]	보다만은 [보다 만 은]
보다만의[보다 만 의]	보다만이 [보다 만 이]
보다만이라도[보다 만 이라도]	보다야 [보다 야]
보단 [보다 는]	보단야 [보다 는 야]
보단커녕 [보다 는 커녕]	

3단계 : 복합명사의 분해

기능어가 분리된 문자열을 입력으로 받아서 복합명사 기분석 사전을 이용하여 복합명사를 단위 명사로 분리하고 각각의 단위명사를 취하여 색인어로 선정한다.

[표 5] 복합명사 사전을 이용한 색인어 추출

입력어절	총득표, 정보화기술,고신뢰전송서비스
복합명사 기분석 사전	총/PF 득표/NN 정보/NN 화/SN 기술/NN 고/PF 신뢰/NN 전송/NN 서비스/NN
추출 색인어	득표, 정보, 기술, 신뢰, 전송, 서비스

4 단계: 한국어 사전을 이용한 색인어추출

기능어가 분리된 문자열로써 복합명사 기분석 사전을 이용하여도 단위 명사로 분리되지 않은 문자열들에 대해서는 한국어 사전을 이용하여 사전에 수록된 명사를 추출하고 색인어로 선정한다.

5 단계 : 사전 미등록어 처리

4단계까지 진행하여도 처리되지 못한 문자열들은 고유명사, 신조어, 외래어, 전문용어, 사용자 표기오류 등 사전 미등록어로 가정하고 색인어를 추출한다. 이때 사용자의 질의에서 추출된 질의어 또한 사전에 등록되지 않은 어절일 가능성도 배제할 수 없으므로 질의어와 색인어 사이의 형태상 불일치(Syntactic Term Mismatch)로 인하여 문서의 내용을 나타내는 색인어가 누락되는 경우가 발생할 수 있다. 이를 해결하기 위해 미등록 색인어나 질의어에 대하여 n-gram방식의 색인 방법을 적용함으로써 부분적으로 일치하는 문자열도 색인어로 추출될 수 있도록 한다. 질의와의 유사도 계산에 있어 추출된 색인어 중 완전일치된 색인어와 부분일치된 색인어에 대한 고려를 함으로써 검색시 정확률을 향상시킨다[8]. n-gram이란 인접한 n개의 음절을 말하는데 사전에 수록되어있지 않은 어절에 대하여 음절 단위로 분할할 때, 한국어에서 나타나는 명사어절들의 출현 비율은 2음절 명사어절이 가장 많이 출현하는 성질을 고려하여 bigram으로 분리하여 저장한다.

[표 6] 사전 미등록어의 bigram 분할

입력어절	오차곡선, 한국기계연구소, 평생교육
bigram 분할에 의한 추출색인어	오차, 차곡, 곡선 한국, 국기, 기계, 계연, 연구, 구소 평생, 생교, 교육

4. 부분일치를 고려한 유사도 계산

질의와 문서 사이의 유사도 계산을 통하여 사용자

가 필요로하는 정보를 포함하는 문서에 순위를 부여하고 사용자들이 우선순위가 높은 문서를 먼저 검색함으로써 필요로하는 정보를 찾는데 필요한 시간을 최소화 할 수 있다. 문서의 순위 결정은 사전에서 추출된 질의어와, 문서가 포함하는 색인어의 출현빈도(term frequency), 역문헌빈도(inverse document frequency)를 이용한 불리언 연산을 통해 문서 리스트를 획득하고, 벡터공간모델(vector space model)을 이용하여 획득한 문서와 사용자 질의문과의 유사도를 계산한다[4]. 사전에 미등록된 질의어에 대해서는 부분일치를 통한 유사도를 계산하는데, 질의어가 색인어와 완전히 일치하지 않더라도 복합 명사를 이루는 단위명사나 bigram에 일치하면 의미적으로 연관성이 있다고 가정한다. 이때 사전에 존재하지 않는 질의어에 대한, 색인어의 부분일치 가중치는 (1)의 식으로 계산하여 완전 일치 색인어로 계산한 가중치보다 낮게 둔다.

$$Weight_{Part} = \alpha \times \mathrm{Sim}(D, Q_{Part}) \quad (1)$$

part : 부분일치

α : 부분일치 색인어 가중치 상수 ($0 < \alpha < 1$)

본 논문에서는 질의어와 문서와의 관련정도를 계산할때 색인어와 완전일치된 유사도와 부분일치된 색인어의 유사도를 더해서 문서의 우선순위를 계산한다. 또한 작은 크기의 문서들이 우선순위 계산에 있어 불공정하게 취급되는 것을 막기 위하여 모든 문서 벡터들의 길이를 정규화한다.

5. 결론 및 향후 발전방향

본 논문에서는 공유된 대량의 문서들 중에서 사용자가 필요로 하는 문서의 효율적인 검색을 위하여 구축된 사전정보를 이용한 색인어 추출 방법을 제시한다. 복합명사의 띄어쓰기 자유로움, 일관성없는 외래어표기, 신조어, 전문용어 등 사전 미등록 어휘로 인한 질의어와 색인어의 형태상 불일치를 해결하기 위하여 bigram 기반의 색인을 사용한다. 또한 검색된 문서에 대한 유사도를 계산하고 이를 이용하여 검색문서에 순위(ranking)를 결정한다.

향후 연구과제로 문서 내에서 색인어의 출현빈도와 함께 검색의 정확도를 높이는 중요한 요소로 작용할 수 있는 색인어의 위치정보를 고려한 문서의 순위결정방법에 대한 연구가 필요하며, 사전탐색시 발생하는 사전 매칭 실패 원인들을 개선시키는 방안과 명사 단위의 색인어에 대한 중요도계산 방법을 연구하여 문서간의 분별력있는 키워드를 추출해 내는 연구가 필요하다. 또한 의미없는 n-gram의 생성으로 분리된 문자열로 추출되는 색인어 수가 많아 짐에 따라 저장공간이 늘어날 수 있는데, 이를 해결하기위해 색인어의 효과적인 저장방법에 대한 연구와 분리된 문자열로 인하여 부적합한 문서들이 검색될(false match)가능성을 줄일 수 있는 방법에 대한 연구가 필요하다.

참고문헌

[1] Baeza-Yates,Ribeiro-Neto, "Mordern Information Retrieval", Addison Wesley, 1999

[2] Roger S. Pressman, "Software Engineering A Practltiners' Approach" 3rd Ed. McGraw Hill

[3] Gerard Salton, "Automatic Text Processing", Addison Wesley, 1989

[4] Gerald Salton, Chris Buckly. "Term-Weighting approaches in automatic retrieval", Information processing Management, 24(5), pp. 513-523, 1988

[5] 강승식, 장병탁, "음절 특성을 이용한 범용 한국어 형태소 분석기 및 맞춤법 검사기", 정보과학회논문지, 제23권, 5호, pp. 530-539, 1996

[6] 김영택 "자연 언어 처리", 교학사, 1994

[7] 김판구, 조유근, "상호 정보에 기반한 한국어 텍스트의 복합어 자동색인", 한국정보과학회논문지, 제 21권, 7호, pp. 1333-1340, 1994

[8] 이준호, 안정수, 박현주, 김명호, "한국 문서의 효과적인 검색을 위한 n-Gram 기반의 색인 방법", 정보관리학회, 제13권, 1호, pp. 47-63, 1996

[9] 이현민, 박혁로, "복합명사의 역방향 분해 알고리즘", 한글 및 한국어 정보처리, 학술대회 논문집, pp.56-59, 2000

[10] 임해창, 윤보현, 강승식, "한국학 서지정보와 전자 텍스트를 위한 자동색인 및 검색시스템 개발 연구", 한국어전산학회, 제2집, pp.273-286, 1998

[11] 장동현, 맹성현, "학습데이터를 이용하여 생성한 규칙과 사전을 이용한 명사 추출기", 제1회 형태소분석기 및 품사태거 평가 워크숍 발표논문집, pp. 151-156, 1999