문법형태소 네트워크를 이용한 자동색인 시스템의 설계

안 성 현, 장 재 우
전북대학교 컴퓨터공학과

Design of Automatic Indexing System Using Korean Morpheme Network

Sung-Hyun Ahn and Jae-Woo Chang
Chonbuk National Univ. Dept. of Computer Engineering

요 약

본 논문은 한국어 특성을 적용하여 키워드를 자동으로 추출하는 기법을 제시한다. 기존에 제안된 명사 추출 시스템인 문법형태소 네트워크를 확장하여 단일 명사 뿐만 아니라 복합 명사를 색인어로 추출한다. 복합 명사는 단일 명사에 비해 보다 한정적 개념을 가지므로, 색인어로 추출될 때 문헌의 식별력을 높일 수 있다. 복합 명사를 구성하는 각각의 단일 명사를 인식함으로써 복합 명사를 분해하고, 간단한 구단위 구문분석을 수행하는 명사 결합 규칙에 따라 단일 명사들을 복합 명사로 합성하는 방법을 제시한다. 마지막으로 이와 같이 추출된 복합 명사에, 복합 명사를 구성하는 단일 명사 간의 연관성을 고려하여 보다 정확한 가중치를 부여할 수 있는 새로운 가중치 부여 방안을 제시한다.

1. 서론

자동색인이란 문헌으로부터 그 문헌을 대표할 수 있는 대표어구를 컴퓨터를 이용하여 자동으로 찾아내는 것을 말한다[1]. 자동색인을 위한 여러 가지 방법들이 제안되고 있는데, 이러한 방법들은 크게 어구의 출현 빈도를 고려하는 통계적 방법과 형태소분석, 구문분석, 의미분석 등 언어 정보를 이용하는 언어학적 방법으로 분류할 수 있다. 그러나 최근의 연구에 의하면 형태소분석을 이용한 방법이 다른 방법들에 비해 정확한 색인어의 추출에는 어려움이 있지만, 구현이 비교적 간단하고 한국어에 적용하기 용이하기 때문에 가장 현실적인 방법으로 인식되고 있다[2,3]. 따라서 본 논문에서는 음절 단위 분석법[4]에 근거하여 형태소분석을 수행한 기존의 "문법형태소 네트워크를 이용한 한글 문헌의 자동 키워드 추출[2]"의 명사 추출 시스템을 확장하여 단일 명사 뿐만 아니라 복합 명사를 색인어로 추출하는 자동색인 시스템을 설계한다. 아울러 추출된 색인어에 대해 중요도를 나타낼 수 있는 가중치를 부여하여 사용자의 질의에 가장 적합한 문서들을 검색할 수 있는 정보를 제공한다. 2 장에서는 전체적인 시스템의 구성을 언급하고, 3 장에서는 복합 명사를 처리하는 과정을 살펴 보고, 4 장에서는 추출된 색인어에 가중치를 부여하는 방법을 제시하고 마지막으로 결론을 맺는다.

2. 시스템 구성

본 시스템에서는 주어진 대상 문서에서 색인어를 추출하기 위하여 기존에 제안된 명사 추출 시스템인 문법형태소 네트워크를 확장하여 색인 작업을 수행한다.

2.1 문법형태소 네트워크

문법형태소 네트워크(GMN : Grammatical Morpheme Network)는 문법형태소가 한 어절 내에서 보이는 결합 패턴과 어절 간의 특정한 연결 구조를 네트워크로 표현하여 형태소분석과 모호성을 해결을 수행하며 이를 통해 보다 정확한 명사 추출을 수행한다. 즉, 한국어에서 문법형태소로 사용되는 음절 및 음소 151 개를 노드로 갖고 어절 내에서 혹은 어절 간에서의 문법형태소 연결 패턴을 링크로 표현한다. 따라서 분석 과정 중에 필요로 하는 조사나 어미 사전과 같은 어휘 사전들을 사용하지 않음으로 분석 시간을 줄일 수 있다.

[정의] GMN = (N, IL, EL)

N(Node) : 문법형태소 노드

IL(Internal Link) : 어절 내에서의 문법형태소 연결 패턴을 나타내는 내부링크

EL(External Link) : 어절 간의 문법형태소 연결 패턴을 나타내는 외부링크

그러나 색인어로써 단일 명사만 추출하고 명사 사전을 사용하지 않기 때문에 "색인"과 같이 조사가 생략되고 단독으로 쓰인 명사를 "색이 + ㄴ"으로 오분석하는 문제점이 있다. 그리고 명사 사전이 사용하지 않았기 때문에 복합 명사를 처리하는데 어려움이 있다.

2.2 문법형태소 네트워크의 확장

명사 사전을 사용하지 않음으로 인한 오분석을 방지하고 복합 명사를 구성하는 각각의 단일 명사를 인식하기 위하여 약 8 만여개의 명사를 포함하는 명사 사전을 사용하여 분석을 수행한다. 복합 명사가 단일 명사에 비해 보다 한정적인 의미를 가지므로 색인어로 추출함으로써 문헌의 식별력을 높일 수 있다. 복합 명사를 구성하는 각각의 단일 명사를 인식함으로써 복합 명사를 분해하고, 간단한 구단위 구문분석을 수행하여 명사 결합 규칙에 따라 단일 명사들을 복합 명사로 합성하게 된다. 이렇게 추출된 색인어에 중요도를 표현할 수 있는 가중치를 부여함으로써 사용자의 질의에 가장 적합한 문서를 검색할 수 있도록 한다. 가중치 부여는 단일 명사 색인어에 대해서는 색인어의 문헌내 출현 빈도와 역문헌 빈도의 곱을 사용하고, 복합 명사의 경우에는 단일 명사와 같은 방법으로 구한 가중치와 구성 단일 명사 간의 연관성을 함께 고려하여 보다 정확한 가중치를 부여할 수 있도록 한다. 전체적인 시스템 구성은 아래 그림 1 과 같다.

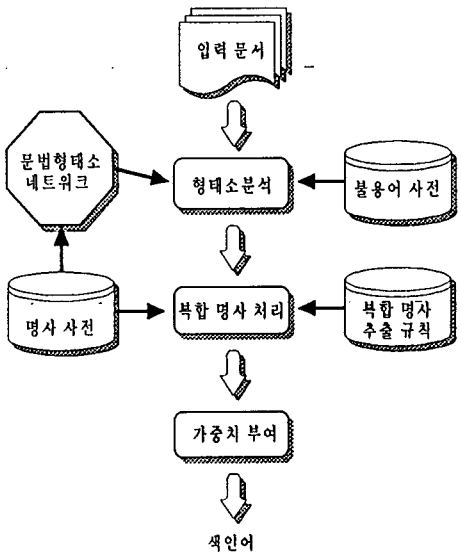

그림 1. 시스템 구성

3. 복합 명사 처리 단계

최근의 정보검색 시스템에서는 단일 명사 뿐만 아니라 복합 명사를 색인어로 사용하여 문서의 내용을 표현하고 있다. 복합 명사는 단일 명사보다 단어의 의미가 한정되므로, 색인어로 사용될 때 문헌의 식별력을 높일 수 있다. 이러한 복합 명사를 처리하는 기법 중의 하나는 복합 명사 사전을 구성하여 복합 명사를 처리하는 것이다. 그러나 한국어에서 복합 명사는 그 수가 매우 많고, 또한 필요에 따라 쉽게 생성되기 때문에 모든 복합 명사를 사전에 수록하여 처리하는 것은 매우 비효율적이다. 따라서 이미 구성된 어휘 사전을 이용하여 복합 명사를 추정하는 방법이 주로 사용된다[6]. 본 논문에서는 단일 명사로 구성된 명사 사전과 복합 명사 추출 규칙을 이용하여 복합 명사를 색인어로 추

출한다.

3.1 복합 명사의 분류

본 논문에서는 복합 명사를 3가지 타입으로 분류하여 처리하였다. 단일 명사를 복합 명사로 합성하여 만들어지는 타입 1과 단일 명사로 분해될 수 있는 복합 명사를 타입 2로 설정하였다. 그리고 타입 1과 타입 2가 혼합된 타입 3이 있다.

타입 1 : 정보의 검색 -> 정보//검색
타입 2 : 정보검색 -> 정보/검색
타입 3 : 정보검색 시스템 -> 정보/검색//시스템

3.2 단일 명사의 복합 명사화

한 문장에서 단일 명사들을 복합 명사로 합성하기 위해서는 일반적으로 간단한 구문분석을 수행한 후에 복합 명사 추출 규칙을 적용하여 복합 명사를 구성하게 된다. 복합 명사 추출 규칙을 사용하면 시소러스와 같은 의미 정보 구축 없이 문장 내에서의 명사 위치만을 이용하여 합성하므로 비교적 단순하다는 장점이 있다. 따라서 본 논문에서는 기존에 제안된 복합 명사 추출 규칙[5]을 적용하여 복합 명사를 추출한다. 복합 명사 추출 규칙은 다음과 같다.

1) 띄어 씌여진 명사군
   : 명사 + 명사
2) 명사의 수식을 받는 명사
   : 명사 + 관형격 조사 + 명사
3) '-적'류의 접미사에 의한 파생어의 수식을 받는 명사
   : 명사 + -적 + 명사
4) '-하다'류의 접미사에 의한 파생 동사와 동사의 필수격
   : 명사 + 격조사(동사에 따라 결정) + 명사 + 하다
5) 관형형 어미를 가지는 조사 기능의 상용어구 중심으로 앞뒤의 명사
   : 명사 + '-에 대한' + 명사

위와 같은 규칙에 의해 복합 명사로 합성되면 복합 명사를 형성하는 각각의 단일 명사 또한 색인어로 추출하여 문헌의 내용을 대표할 수 있는 단일 명사가 색인어로 추출되지 않는 문제점이 없도록 하였다.

3.3 복합 명사의 단일 명사화

복합 명사를 단일 명사로 분해하는 방법에 대한 기존 연구는 복합 명사의 길이를 4음절에서 7음절로 제한하고, 자주 발생하는 패턴에 근거하여 복합 명사를 추정해내는 방법을 사용한다[6]. 이 방법은 분석 시간을 줄일 수 있다는 장점은 있지만, 음절 수가 7음절로 제한되어 있어 7음절 이상의 복합 명사는 처리할 수 없고, 또한 통계적인 패턴을 적용하므로, 부적절한 분석 결과를 유발할 수 있다는 단점이 있다. 따라서 본 논문에서는 최장일치법을 적용하여 복합 명사를 구성하는 단일 명사들을 인식하는 방법을 사용한다. 이를 통하여 글자 수의 제한 없이 분석이 가능하며 단일 명사 사전에 없는 고유 명사도 추정이 가능하다. 그리고 1음절 명사는 정확한 의미 표현이 어렵고 분석시 모호성을 유발하므로 2음절 이상의 단일 명사만 색인어로 추출하고, 4음절 이상의 복합 명사만을 분석 대상으로 한다.

복합 명사를 구성하는 단일 명사를 명사 사전에서 검색하는 방법에는 두가지 방법이 있다. 첫번째는 복합 명사를 뒤에서부터 검색하는 방법이고, 두번째는 앞에서부터 검색하는 방법이다. 첫번째 방법은 알고리즘이 단순하고 빠르지만 "대학생선교회" 같은 복합 명사는 "대학/생선/교회"와 같은 오분석을 유발할 수 있다. 두번째 방법에서는 "김영삼대통령"과 같이 미등록어가 포함된 복합 명사는 제대로 분석하지 못한다. 따라서 본 논문에서는 두가지 방법을 결합한 최장일치법으로 분석하는 방법을 제시한다.

복합 명사는 2음절 이상의 단일 명사들로 이루어진다고 가정하였으므로 "정보검색시스템"이란 복합 명사에서 앞의 2음절을 제외하고 "검색시스템"이란 단어를 단일 명사 사전에서 찾는다. 만약에 존재하지 않는다면 그 다음으로 "색시스템"을 찾게 된다. 이러한 방법으로 사전을 검색하면 "시스템"이라는 단일 명사를 인식하고 분리할 수 있다. 그리고 나머지 부분인 "정보검색"에서도 위와 같은 방법을 적용하면 최종적으로 "정보/검색/시스템"과 같이 각각의 단일 명사로 분리할 수 있다. 이와 같이 하면 뒤에서부터 분석함으로써 유발되는 오분석을 방지할 수 있고, 미등록어가 포함된 복합 명사도 처리할 수 있다. 그러나 이렇게 최장일치법으로 처리하면 많은 사전 검색이 필요하므로 사전의 구조를 효율적으로 구성해야 한다. 본 논문에서는 Trie를 기본 구조로 하여 사전을 구성하였다. Trie는 전체 사전 엔트리를 주기억 장치에 저장하므로 B-Tree 등과 같이 보조기억 장치를 사용하는 사전 구조에 비해 검색 속도가 빠르다[7].

4. 가중치 부여 단계

문서에 부여되는 색인어는 다음의 두가지 요구 사항을 만족해야 한다.

1) 문서의 내용을 잘 표현할 것.
2) 문서를 다른 문서와 잘 구분 지울 것.

첫번째 사항은 색인어의 표현력이라 하고, 두번째 사항은 식별력이라 한다. 색인어의 표현력은 문서와 색인어와 관계를 나타내는 것으로 이 색인어가 얼마나 문서를 잘 표현할 수 있는 지를 나타내게 된다. 그리고 식별력은 이 색인어가 색인어로써 얼마나 가치가 있는 지를 나타내게 된다. 만약에 어떤 색인어가 모든 문서에 존재한다면 이 색인어를 통한 문서 검색은 의미가 없게 된다. 따라서 식별력이 높은 색인어는 특정 문서에만 집중적으로 나타나게 된다. 좋은 색인어는 표현력과 식별력이 모두 좋아야 하므로 색인어의 중요도는 일반적으로 표현력과 식별력의 곱으로 나타내게 된다[8].

본 논문에서 사용한 가중치 부여 방법은 색인어의 문헌 내 출현 빈도(Term Frequency, TF)와 역문헌 빈도 (Inverse Document Frequency , IDF)의 곱을 사용한다. 색인어에 부여되는 역문헌 빈도 산출은 Sparck Jones 가 제시한 공식을 이용하였다[9].

$$IDF_i = \log_2(N) - \log_2(DF_i) + 1$$

N : 전체 문헌의 수

DF_i : 단어 i 의 문헌 빈도 (Document Frequency)

색인어 i 가 문서 j 에서 가지는 중요도는 다음과 같이 계산된다.

$$W_{ij} = TF_{ij} * IDF_{ij}$$

TF_{ij} : 문서 j 내에서의 색인어 i 의 출현 빈도

위의 식에서 TF 는 색인어의 표현력을 나타내고, IDF 는 색인어의 식별력을 나타낸다.

4.1 복합 명사의 가중치 부여

복합 명사에 가중치를 부여하는 방법에는 복합 명사를 하나의 색인어로 간주하고 빈도수에 의해 가중치를 부여하거나, 복합 명사를 이루는 각각의 단일 명사에 대해 가중치를 부여하는 방법이 있다. 그러나 위와 같은 방법들은 복합 명사가 가지는 중요도를 정확히 표현하지 못한다는 단점이 있다. 따라서 본 논문에서는 복합 명사에 가중치를 부여할 때, 복합 명사 자체의 가중치와 복합 명사를 구성하는 각각의 단일 명사 간의 연관도를 고려하여 보다 정확한 가중치를 부여할 수 있도록 한다.

가중치를 계산하는 식은 다음과 같다.

$$W(K|D) = w(K|D) * [(n-1) * Sim(K|C)]$$

- K : 단일 명사로 이루어진 복합 명사
  ($K = k_1 + \ldots + k_n$, k_i : 단일 명사)
- D : 복합 명사 K 가 출현한 문서
- C : 전체 문헌들의 집합
- n : 복합 명사 K 를 구성하는 단일 명사의 개수
- W(K|D) : 복합 명사 K 의 가중치
- w(K|D) : 복합 명사 K 자체의 가중치
  (TF * IDF 로 얻어진 가중치)
- Sim(K|C) : 복합 명사 K 를 구성하는 단일 명사들 간의 관련도

위의 식에서 w(K|D)는 복합 명사를 하나의 색인어로 간주하고 얻어진 가중치를 가리킨다. 일반적으로 복합 명사를 구성하는 단일 명사의 개수가 많아질수록 그 의미는 더욱 한정적이고 보다 많은 정보를 가지고 있다고 할 수 있다. 따라서 n 값이 클수록 보다 높은 가중치를 부여한다. Sim(K|C)는 전체 문헌에서 복합 명사 K 를 구성하는 각각의 단일 명사들이 얼마나 많은 관련성을 가지고 결합하고 있는 정도를 나타낸다. 단어 간의 관련도를 계산하기 위해서는 시소러스 등과 같은 의미 정보가 구축되어 있어야 한다. 그러나 현재 기술로는 완벽한 시소러스의 구축이 불가능하기 때문에 본 논문에서는 그러한 의미 정보없이 복합 명사와 단일 명사의 출현 빈도에 의한 간단한 계산으로 단일 명사 간의 관련도를 측정한다[10].

$$Sim(K|C) = \frac{\alpha|A(K)| + \beta|B(K)| + \gamma|C(K)|}{|k1| + \ldots + |kn|}$$

$|A(K)|$: 타입 1 의 복합 명사 K 의 출현 빈도

$|B(K)|$: 타입 2 와 3 의 복합 명사 K 의 출현 빈도

$|C(K)|$: 한 문헌 내에서 복합 명사 K 를 구성하는 각각의 단일 명사가 모두 독립적으로 출현한 문헌의 개수

$|k_i|$: 단일 명사 k_i 의 출현 빈도

α, β, γ : 복합 명사 타입을 고려한 가중치

위의 식에서 α, β, γ는 복합 명사 타입 간에 가중치를 나타내는 인자이다. 일반적으로 $\alpha \geq \beta \geq \gamma$ 이다. 본 실험에서는 실험의 간략화를 위해 $\alpha = 3, \beta = 2, \gamma = 1$로 가정하고 수행한다. 그러나 보다 정확한 가중치를 얻기 위해서는 많은 실험을 통해 최적의 α, β, γ 값을 구해야 한다.

5. 결론 및 향후 연구 방향

본 논문에서는 기존에 제안된 명사 추출 시스템인 문법 형태소 네트워크를 확장하여 단일 명사 뿐만 아니라 복합 명사를 색인어로 추출하는 자동색인 시스템을 설계하였다. 단일 명사 뿐만 아니라 복합 명사를 색인어로 추출하고, 복합 명사를 구성하는 각각의 단일 명사의 연관성을 고려하여 가중치를 부여함으로써 사용자의 질의에 가장 적합한 문서를 검색할 수 있는 정보를 제공할 수 있다.

향후 연구 방향으로는 보다 다양한 복합 명사 추출 규칙을 찾아내고, 복합 명사를 구성하는 각각의 단일 명사 간의 관련성을 보다 정확히 계산할 수 있도록 α, β, γ의 값을 최적화하는 연구가 필요하다.

참고 문헌

[1] 정영미, 정보검색론, 구미 무역 출판부, 1993.

[2] 김철완, "문법형태소 네트워크를 이용한 한글 문헌의 자동 키워드 추출," 전북대학교 석사학위 논문, 1995.

[3] 강승식 외, "한국어 자동색인을 위한 형태소 분석 기능," 한국정보과학회 봄 학술발표논문집, pp.929-932, 1995.

[4] 강승식, "음절 정보와 복수어 단위 정보를 이용한 한국어 형태소 분석," 서울대학교 박사학위 논문, 1993.

[5] 김민정, 권혁철, "한국어 특성을 이용한 자동 색인 기법," 한국정보과학회 가을 학술발표논문집, pp.1005-1008, 1992.

[6] 강승식, "한국어 형태소 분석을 위한 복합 명사의 인식 방법," 한국인지과학회 춘계 학술발표논문집, pp.175-189, 1993.

[7] 이승선 외, "Compact TRIE Index(CompTI) : 한국어 전자사전을 위한 데이타베이스 구조," 한국정보과학회 논문지, 22 권 1 호, pp.3-12, 1995.

[8] 박혁로 외, "언어 및 통계 정보를 이용한 색인어의 중요도 계산," 한국정보과학회 봄 학술발표논문집, pp.635-638, 1992.

[9] G. Salton and M.J. McGill, Introduction to Modern Information Retrieval, McGraw-Hill, New York, 1983.

[10] William B. Frakes and Ricardo Baeza-Yates, Information Retrieval Data Structure & Algorithms, Prentice Hall, 1992.