자동색인에서 단어의 품사와 빈도를 이용한 색인후보어 발췌

Extraction of the Latent Index Terms Using the Word Frequency and Part of Speech in Automatic Indexing

이태영, 남궁황 (전북대학교 문헌정보학과)

Tae Young Lee, Hwang Namgoong
Dept. of Library and Information Science, Chonbuk National University

본 논문에서는 적합한 색인어를 자동으로 추출해 내기 위해 잘 알려진 통계적 기법과 구문분석적 기법을 혼용하였다. 적용결과를 검색효율로 나타내지 않고 각 방법에 따라 추출된 단어들을 실증적으로 보여주어 성능에 대한 판단을 유도하였다. 빈도나 품사가 단독으로 사용된 것보다 동시에 적용된 것이 보다 좋은 결과를 가져왔다.

1. 서론

사람의 손을 거치지 않고 컴퓨터가 자동으로 색인을 만들어 주는 방법에 대한 연구 및 기법개발이 그 동안 많이 수행되었다. 이들 기법을 크게 분류하면 통계적 기법과 구문분석적 기법으로 나눌 수 있다. 통계적 기법의 통계 대상은 문헌에 출현한 특정 단어의 빈도 수와 특정 단어가 출현한 문헌의 빈도 수이다. 가장 기본적인 척도는 단어빈도(TF), 문헌빈도(DF), 장서빈도(CF)이며, 이것들을 근간으로 상대적인 조합이 이루어져 만들어지는 상대빈도 방법들이 다수 실험되어져 왔다.

구문분석적 기법에서 색인어 발췌 수단으로 많이 사용된 것은 단어의 품사와 단어에 내포된 주제의미의 강약을 활용하는 방식이었다. 품사방식에서는 명사(N)를 발췌의 기준으로 삼고 명사가 두 개 이상 이어서 쓰여지는 "N N, N의 N"과 같은 명사구를 색인어로 상정하였다. 그리고 주제의미의 강약에서는 의미가 약한 단어와 기능어들을 불용어휘로 처리하는 방법을 주로 택하였다.

본 연구는 그동안 실험적으로 잘 알려져 왔던 통계적 기법에 속하는 여러 모형들과 이들 모형에 품사 및 불용어리스트를 사용하는 구문분석적 기법이 가미될 때 어느 정도의 효과를 가져오는지를 살펴보고자 한다.

2. 색인어 추출 방법

2.1 구문분석적 방법

문장 내에서 명사(N), 또는 명사구(N N, N의 N 등)로 쓰인 단어(구)들을 발췌하여 색인어로 올리는 방법이 많이 사용된다. 또한 특정한 정보를 발췌하기 위해서 여러 가지 품사가 사용되는 예를 볼 수 있다. W. Le 등은 시간에 관련되는 정보를 추출하기 위해서 시간에 관련된 동사의 정보 즉, 동사의미, 시제 등을 이용하였다. 그리고 의미적인 측면에서 메타시소러스를 적용한 사례도 있었다.

2.2 통계적 방법

일반적으로 자동색인을 설명하는 공식들을 측정 모형의 대상으로 삼았다. 기본적인단순빈도 척도로는 단어빈도(TF), 문헌빈도(DF), 장서빈도(CF)가 있으며, 이들 단순빈도 척도를 서로 가미하여 만든 상대빈도 공식인 TF/DF, TF/CF, TF/Pi(Pi는 문헌 i 내의 단어의 총빈도), TF/CF · Pi 등이 있다. 그리

고 Jones가 제안한 역문헌빈도 공식과 Edmundson과 Wyllys가 제안한 f-r, f/r, f/f+r 등을 고려해 볼 수 있다. 여기서 f는 한 문헌 내에서 단어 k의 상대빈도(=TF/Pi)를 의미하고, r은 전체 문헌집단에서의 상대빈도(=CF/ $\sum$Pi)를 말한다.

2.3 제안 방법

통계적 방법과 구문적 방법을 혼합하여 측정한다. 단어의 출현빈도로 여러 가지 공식을 만들어 색인어후보들을 골라내는 방법은 이미 추출되는 단어들이 명사란 전제를 달고 있다. 궁극적으로 통계와 구문이 이미 혼성되어 있는 것이나 다름이 없다. 본고에서는 단순히 명사의 성분는 물론 다른 품사들의 표징적인 정보를 이용하고자 한다. 빈도에 의한 방법으로 발췌된 색인어후보에 같은 문장, 또는 절에서 출현한 (1) 용언의 어미와 조사의 표징, (2) 용언의 중의성, (3) 용언이나 용언형식으로 쓰인 단어의 의미적 단서, (4) 부사 단서어, (5) 음절수(복합명사의 복합정도 감안) 등을 포함시켜 보다 정제된 색인어후보들이 추출되는가를 살펴본다. 또한 시소러스를 사용하였을 때를 가정한 모의실험도 한다.

위의 (3), (4)는 자동초록에서 적용하는 방법이며, (1)과 (2)는 이와 유사한 방법이라고 할 수 있다. 용언어미인 '하였으니까'는 원인을 이야기하며 결과를 기대하는 표징을 나타낸다. 조사 '만은'은 유일성을 나타내어 "가, 는, 을"들과는 다른 특별성을 준다. 그리고 "내세우다, 걸어가다"는 "세우다, 걷다"를 보다 확실히 표현한 단어이다. 따라서 이들 단어와 같이 쓰인 색인어후보는 보다 강한 상황적 의미를 지녔다고 할 수 있다. (표1, 표2 참조)

<표1> 용언의 어미와 조사

용언 : -(었, 였)기 때문에, -(였으)니까, -지만, -야(만), 부정 등 조사 : 만(은), 뿐만 아니라, 도 등

<표2> 중의 용언

걸어가다, 내세우다, 받아들여지다 등

3. 실험 및 결과

3.1 실험 방법

본고에서 실험은 다음과 같은 방법으로 전개하였다.

(1) 실험에 쓰일 연구대상 문서는 문헌정보학에 관련된 논문 40편에서 각 편마다 15문장씩(연속하여 출현하는 문장들)을 발췌하여 40개의 문서를 만들었다. 한 문장은 평균적으로 23-25단어들을 갖고 있었다. 이들 문서를 10개 문서단위로 나누어서 각 문서단위들을 측정 문헌 집단으로 삼고, 일차적으로 단어들의 산포도를 예측한 후 분산이 중간적인 문서집단을 여러 가지 측정에 사용하여 각각의 모형에 대한 정보를 제공한다.

(2) 명사를 중심으로 한 ①N, ②N +N, ③N의 N을 발췌하여 그 출현 상황을 살펴보고 통계적 빈도와의 합성 시에 어떤 결과가 오는지 관찰한다.

(3) 통계적 기법의 단순빈도에서 사용하는 ①TF, ②DF, ③CF, ④역문헌빈도, ⑤f-r, ⑥f/f+r 등을 측정하여 빈도값이 중간적인단어들이 10개 내지 20개 정도 출력되도록 상하 한계치를 정하고 그 출력 결과를 살펴본다(가급적 명사가 출력되도록 한다).

(4) 2.3절에서 제안되어진 ①용언어미 표징, ②용언의 중의성, ③용언(형)의 의미성, ④부사 단서어, ⑤음절수(3복합어 이상과 6음절 이상)를 고려하여 출현한 색인어후보들을 살펴보고, 통계적 빈도와 합성하였을 때 어떤 결과가 오는지 관찰한다.

3. 2 실험결과

실험대상인 문서들의 총 출현 단어 수는 1901개(조사는 명사에 포함)였다. 최고 빈도의 TF는 16번(업무), DF는 10번, CF는 68번(있다)이었다. 이들 문서 중에서 가장 표준적인 정보를 제공한다고 생각되는 B문헌에 대한 측정현상을 살펴보면 다음과 같다. B문서의 조사를 생략한 띄어쓰기를 기준으로 한 단어의 수는 205개이고, 색인어의 대상인 명사는 <표3>과 같다.

<표3> B문헌에 출현한 명사

도서관, 책, 것, 장식물, 경영, 자료, 보존,측면, 이용, 대전제, 고대, 실존사회, 일부, 특권계급, 왕, 귀족, 부호, 학자, 성직자, 일반대상, 유기적, 조직체, 기관, 외적, 내적, 면, 도서관인, 지식자원, 일체, 편견, 간섭, 유혹, 개척자, 정신, 일상, 난관, 열정, 인내, 용기, 희망, 속, 민족, 인류, 기억, 사회발전, 운영주체, 책임, 우리, 직업적, 행위, 바탕, 비판적, 자기성찰, 심리적, 각성, 폐가제, 자료실, 개가제, 이용자, 직원, 변화, 자료, 만족도, 절대적, 도서, 위치, 비대출시, 경우, 장점, 사본시대, 수, 쇠사슬, 결점, 위주, 근대, 사상, 저항, 구성원, 봉사, 궁극적, 목적, 열람방식, 완전개가제, 이상적

(1) 표3의 결과에서 'N N'과 'N의 N'으로 출현한 명사들 만을 선정하면 표 4와 같이 전체 명사수가 84개에서 34개로 줄어든다.

<표4> 'NN"과 N의N 명사

도서관, 경영, 자료, 보존, 고대, 실존사회,일부, 특권계급, 유기적, 조직체, 일체, 편견, 개척자, 정신, 인류, 기억, 운영주체, 직업적, 행위, 비판적, 자기성찰, 심리적, 각성, 폐가제, 자료실, 개가제, 직원, 변화, 도서, 위치, 비대출, 이용자, 위주, 근대

(2) TF로 빈도값 3에서 6사이에서 나타나는 단어들은 표5와 같다.

<표5> TF 측정 단어

개가제(3), 도서관인(4), 때문에(3), 수(6), 아니다(3), 없다(5),위하다(3), 이용하다(4),이용자(3), 자료(4), 책(3),하다(4)

(괄호안의 숫자는 출현빈도임)

(3) DF로 빈도값이 2에서 3사이에서 나타나는 단어들은 표6과 같다.

<표6> DF 측정 단어

가지다(2),각성(3), 개가제(2), 결코(2), 경영(2), 기관(2), 변화(2), 봉사(2), 선택하다(3), 소장하다(3), 우리(2), 원하다(3), 위주(2), 일부(3), 일(2), 즉(2), 행위(2), 희망(2)

(4) CF로 빈도값이 3에서 6사이에서 나타나는 단어들은 표7과 같다.

<표7> CF 측정 단어

가장(6), 각성(3), 개가제(4), 결코(3), 도서관인(4), 도서(6), 때문(6), 선택하다(3), 소장하다(4), 우리(3), 운영하다(5), 원하다(4),위주(3), 이상적(4), 일부(5), 즉(3), 책(3)

(5) 역문헌빈도로 빈도값이 7인 단어들은 아래의 표8과 같다.

<표8> 역분헌빈도 측정 단어

고대, 공개하다, 공평하다, 귀족, 부호, 살아있다, 성직자, 수, 실존사회, 왕, 이용하다, 일반대상, 자료실, 책임, 특권계급, 폐가제, 학자

(6) f-r과 f/f+r의 상하 한계치가 각각 1과 0.009, 0.7과 0.6 사이에 오는 단어들을 재현하면 표9와 같다.

<표9> f-r과 f/f+rF 측정 단어

f-r : 개가제, 것, 도서관, 도서관인, 때문, 수, 아니다, 없다, 이용하다, 자료, 책 f/f+r: 것, 그, 도서관, 되다, 때, 소장하다, 원하다, 위하다, 이상, 이용자, 자료

(7) TF/CF, TF/CF · Pi, f/r은 각각 1-1, 0.0045-0.005, 8.7-8.7의 한계치에서 79개의 동일한 단어들을 추출해 내었고 TF/DF도 1-1의 한계치에서 거의 비슷한 단어들을 출력하였다.

(8) 용언어미 표징, 용언의 중의성, 용언형의 의미성, 부사 단서어, 음절수를 근거로 한 색인어 후보들을 살펴보면 다음과 같다.

(가) 용언어미 표징

①"-만을 위한 것이 아니고" : 고대, 실존사회, 일부, 특권계급, 왕, 귀족, 부호, 학자, 성직자, 일반대상, 도서관

②"-기 때문에" : 도서관, 유기적, 조직

체, 기관, 외적, 내적, 면, 사본시대, 책, 수, 것, 쇠사슬,

③"-할 수 없는 -지만" : 도서관, 경영자료, 보존, 측면, 일, 이용, 대전제

(나) 용언의 중의성

①"받아들여지다" : 개가제, 결점, 이용자, 위주, 근대, 도서관, 사상, 것,('-지만'도 포함되어 있음)

(다) 용언형의 의미성 :

①"이상적이다" : 도서관, 구성원, 자료, 봉사, 것, 궁극적, 목적, 열람방식, 완전개가제, 이상

(라) 음절수 : 해당하는 단어가 없음

위와 같이 기본적인 구문분석 방법과 통계 방법들을 동원하여 실험한 결과, 명사를 이용한 발췌방법은 많은 수의 색인어후보가 등장을 하며, 통계 공식들은 공식에 따라 추출 되는 단어 수를 어느 정도 조절할 수 있으나 명사 발췌율이 평균 약 60%이었다. 발췌된 명사들 중에서 불용어 리스트에 포함되어야 할 명사의 수는 11개로 추출된 명사 수의 20% 수준이다. 용언어미 표징 등으로 출현한 명사들은 46개가 되었는데, <표5, 6, 7, 8, 9>에서 제시한 단어들과 중복되지 않는 것은 "유기적, 조직체, 기관, 외적, 내적, 면, 사본시대, 쇠사슬, 보존, 측면, 일, 대전제, 개가제, 결점, 근대, 사상, 구성원, 궁극적, 목적, 열람방식, 완전개가제"등이 있다.

또한 표4("NN, N의 N"의 방법)와 비교를 하면 "왕, 귀족, 부호, 학자, 성직자, 일반대상, 측면, 일, 이용, 대전제, 쇠사슬, 면, 사본시대, 보존, 측면, 내적, 외적, 기관, 책, 수, 것, 개가제, 결점, 위주, 근대, 사상, 구성원, 봉사, 궁극적, 목적, 열람방식, 완전개가제, 이상" 등과 같은 단어들이 중복되지 않았다. 중복되지 않은 단어들 중에 "사본시대, 열람방식, 완전개가제"는 강력한 주제 표의어라고 생각되는데 명사와 통계를 이용한 방법에서는 누락되어 있다. 반면에 "측면, 일, 면, 수, 위주, 결점 등"은 색인어가 되기에는 주제표의가 상당히 낮은 단어들이다. 이들 주제표의가 상당히 낮은 단어들을 제거하면(불용어리스트 또는 시소러스 이용) 전체적으로 표10과 같은 단어들이 출현할 수 있다.

<표10> 색인어후보 단어

개가제, 고대, 근대, 귀족, 도서관, 도서관인, 봉사, 부호, 사본시대, 성직자, 실존사회, 완전개가제, 열람방식, 이용자, 자료, 자료실, 책, 특권계급, 폐가제

4. 결 론

띄어쓰기 기준으로 205개 단어 정도의 짤막한 문서들을 실험대상으로 삼았고, 명사 위주로 색인어 발췌를 하는 구문적 기법과 출현빈도에 근거하여 색인어를 추출하는 통계적 기법을 적용하여 나타나는 현상을 살펴보았다. 명사만 추출하였을 때 84개 단어가 출현하였고, 이중적으로 쓰인 명사구를 파악한 결과 34개의 단어를 얻을 수 있었다. 여기서 표10을 고려하면 주제표의가 낮은 단어들이 각각 65, 15개씩 과다 출현하였다.

그리고 일반적인 통계적 기법들을 적용하여 출현빈도가 중간적인 단어들(되도록이면 명사가 많이 재현될 수 있는 한계치 적용)을 10-20개 출력시켜 보았다. 그 결과 평균적으로 40% 정도는 명사가 아닌 단어들이 출력되었다. 또한 명사 중 불용어휘의 양은 20% 가량 되었다. 본고에서 제안한 용언어미 표징, 용언(형) 의미성 등으로 발췌된 명사들은 46개가 되었고 그 중에서 많은 수의 단어가 주제표의가 낮은 단어들이었다. 그렇지만 명사와 통계를 이용한 방법에서 출현되지 못했던 강한 주제표의어를 나타내는 단어를 발견할 수 있었다. 이러한 관점에서 구문분석 기법이 보완 되고 통계적 방법과 결합하였을 때 더 좋은 결과를 얻는 자동색인작성 방법이 출현되리라 생각된다.

참고문헌

정영미, 1997. [정보검색론], 구미무역.

Salton, G., 1989. [Automatic Text Processing], Addison Wesley.

Wright, L.W. etc., "Hierarchical Concept Indexing of Full Text Documents in the Unified Medical Language System @Information Sources Map", JASIS, 50-6, (1999), pp.512-23.

Li, W. etc., "Toward Automatic Chinese Temporal Information Extraction", JASIST, 52-9, (2001), pp.748-62.