韓國 컴퓨터情報學會 論文誌
第11卷 第1號, 2006. 3.

2006-11-1-4-5

거리 제한을 이용한 색인 시스템

박 찬 이*, 김 상 복**

An Index System using Restrictive Distance

Chan-Ee Park *, Sang-Bok Kim **

요 약

한본 논문에서는 단어 가중기법에 거리 개념을 도입한 색인 기법을 제안한다. 본 색인 기법은 질의어와 문서를 대표하는 색인의 대부분은 복합명사 혹은 인접한 두개 이상의 명사 또는 명사구가 많으며 이들 명사간의 거리가 멀면 멀수록 색인으로 선택되는 비율이 줄어드는 점을 착안하여, 이를 기존의 가중치 부여 기법으로 색인어 후보를 선정하고, 후보들 간의 거리가 3어절 이내의 후보를 최종 색인으로 선정하였다. 이 방법을 이용하여 신문기사, 학술논문, 웹문서 등 100여종의 문서를 대상으로 실험한 결과 신문기사 92.03%, 학술논문 95%, 웹문서는 73..33%의 정확율을 보였다.

Abstract

In this paper, we propose index method introducing distance concept in word by a method weighting word. This index method is frequent representing an inquiry word and document index and compound noun or more than two adjoin nouns or noun phrase, the farther the distance between these nouns, the fewer selected ratio decreases in index point is the aiming, this choose guide word candidate by existent weight grant method and distance between candidates chose candidate finally in index within 3 sentences. Using in these way I document of 100 kinds of newspaper, scientific treatise, web document and so on, showed the correctness rate resulted of newspaper 92.03% scientific treatise 95% web document 73.33%

▸ Keyword : automatic indexing(자동 색인), distance(거리), weight of word(단어 가중치)

• 제1저자 : 박찬이 • 교신저자 : 김상복
• 접수일 : 2006.02.01, 심사완료일 : 2006.03.13
* 경 상대학교 컴퓨터과학과 박사수료, ** 경상대학교 컴퓨터과학과 교수, 경상대학교 컴퓨터정보통신연구소 연구원

Ⅰ. 서 론

인터넷 기반의 웹서비스가 일반화됨에 따라 정보검색을 위해 다양한 종류의 검색엔진들이 개발되어 사용되고 있다. 현재 인터넷에서 사용하고 있는 정보검색서비스 종류는 크게 주제별 검색엔진, 단어별검색엔진, 메타검색엔진 등이 있다. 더 나아가 유전자 알고리즘 및 퍼지 이론을 이용한 추론기법을 이용한 검색엔진[1]도 연구되고 있다. 검색사이트는 많은 발전을 거듭하고 있으며 대부분의 검색엔진이 이들 3가지 유형의 검색엔진의 특성을 포함하고 있다. 이들 3가지 검색엔진 중 단어별검색엔진은 로봇 에이전트에 의해서 각 사이트에서 수집한 정보를 데이터베이스에 저장하고, 이를 사용자가 입력한 키워드에 따라 원하는 정보를 검색하는 방식이다. 검색엔진은 사용자의 질의어에 대해 웹 데이터에서 검색하여 그 결과를 표시한다. 그러나 웹 데이터의 본문을 검색한다는 그 자체만으로도 많은 시간을 소요하게 되므로 비효율적이다. 그 대신 웹문서의 내용을 대표하는 주제어를 색인하여 이를 데이터베이스로 일차적으로 구축한 후 검색하여 검색결과를 되돌려주는 것이 일반적이다.

색인은 크게 수동색인과 자동색인으로 구분할 수 있다. 자동색인은 다시 전문가에 의해 선정된 어휘집을 이용한 통제어색인과 형태소분석[2,3,4,5,6]과 같은 기법을 이용하여 문서에 나타난 어휘만을 색인으로 추출하는 자연어색인으로 나눈다. 마지막으로 자연어색인기법은 단일어색인과 복합어색인으로 나누어지진다.

색인 선정은 다양한 통계적 기법을 이용하여 가중치를 부여하여 일정 한계치를 초과하는 것들만을 색인으로 선정하는 방법[2, 3, 4,5]을 이용한다. 이 방법은 단어의 출현빈도수를 이용하는 방법과 특정 단어와 관련하여 관련문헌, 비관련문헌을 분리하는 문헌분리[7]를 이용하는 단어의 문헌 분리 능력에 의한 방법, 확률분포에 의한 방법 마지막으로 정보이론에 근거한 방법 등이 있다.

색인의 대상은 명사나 명사구 혹은 전치사구가 될 수 있는데, 이들을 분리하기 위해서 초보적 검색 시스템에서는 문장 부호나 특정 조사, 어미 등을 단서로 분리하였으나, 요즘 대부분의 검색시스템은 형태소 분석기법[2,3,4,5,6]을 이용하고 있다. 단위명사와 복합명사는 형태소 분석을 통해 얻을 수 있으며, 명사구나 전치사구는 간단한 구문분석기법[4]을 통해 얻을 수 있다. 그런데 형태소 분석을 통해 얻어진 연속된 단위명사를 결합하여 복합명사로 인식하는 것은 문제가 되지 않으나, 복합명사를 단일어로 분리하는 과정[8,9]에는 많은 문제점을 내포하고 있다. 예를 들어 “한국통신”, “현대건설” 등과 같은 고유명사를 복합명사로 인식하여 “한국”, “통신”, “현대”, “건설” 등으로 분리하는 문제가 발생할 수 있다. 그리고 붙여 쓴 복합명사를 단위명사로 분리할 때, “대학생선교회”를 “대학생”, “선교회”로 분리하는 것이 맞지만, 프로그램에 의해 자동 분리될 경우 “대학”, “생선”, “교회”로 분리되는 중의적 분리 문제가 발생할 수 있다. 그러므로 복합명사의 정확한 분해는 명사를 색인으로 추출 관리하는 검색엔진에서는 검색 성능을 좌우하는 중요한 요인이 된다.[2,8,9]

본 논문에서는 가장 널리 이용되고 있는 검색엔진인 야후코리아, 네이버, 엠파스을 통해 복합명사 분해와 검색 성능을 분석하고, 검색 성능을 높일 수 있는 새로운 색인 시스템을 제안하고자 한다.

본 논문의 구성은 2장에서는 현재 연구 및 운영되고 있는 색인 시스템의 현황을 설명하고 3장에서는 거리 개념을 적용한 색인 시스템의 구현 방법과 실험 결과를 설명하고 4장에는 실험 결과에 대한 결론과 향후 과제를 제시하였다.

Ⅱ. 관련연구

검색엔진의 역할은 웹상의 많은 문서들을 수집하고 이를 가공하여 데이터베이스에 저장하고 사용자의 질의에 따라 데이터베이스를 검색하여 그 결과를 사용자에게 제공하는 것이다. 그런데 모든 문서의 내용을 데이터베이스에 저장하고 이를 검색하는 것은 많은 시간이 소요되고 처리해야할 데이터가 기하급수적으로 증가하므로 비현실적이다. 문서의 내용을 잘 나타내면서 처리할 데이터는 양은 줄임으로써 검색엔진의 효율을 높이는 방법이 강구되어야 한다.

웹문서를 분석한 후 해당 문서의 특성을 표현하는 주요 개념을 추출하여 각 문서를 대표하도록 한 것을 색인이라고 한다. 색인을 이용하면 방대한 양의 문서로부터 사용자가 원하는 자료만을 선택할 수 있어서 사용자가 직접 접근해야 하거나 탐색해야 하는 문서의 수를 줄일 수 있다. 특히 검

색 시스템에서 문서에 대한 직접적인 탐색을 하기보다는 이들 문서를 대표하는 색인을 대상으로 탐색 및 처리하게 되므로 데이터의 양을 줄일 수 있으며, 적은 양의 데이터 처리로 인해 처리에 소요되는 시간이 짧아지게 되므로 검색엔진의 효율이 높아지게 된다.

자동색인은 단어별 검색 시스템뿐만 아니라 주제별 검색 시스템에서 문서를 자동 분류하는데 있어 문서의 특성을 분류기 및 기계학습에 사용되는데 문서의 특성을 추출하는 방법으로 자동색인을 사용한다. 그리고 문서 자동요약시스템에서 대표 색인어를 추출하여 그와 관련된 문장을 추출하여 요약하는데도 이용된다.[10] 그러므로 자동색인은 형태소 분석과 함께 다른 응용시스템의 기초로 사용됨으로 매우 중요한 위치를 차지하고 있다.

과거에는 문서에 대한 색인을 선정하는 색인 작업은 색인에 대한 전문지식을 갖춘 훈련된 색인자 또는 주제 전문가가 자신의 지식을 기초로 하여 문헌을 분석한 후 임의의 색인어를 부여하거나 통제어휘집을 사용하여 통제된 용어 중에서 적합한 색인어를 선택하는 방법으로 이루어졌다. 하지만 방대한 정보가 쏟아지고 있는 시점에서 전문지식과 경험을 갖춘 색인 전문가가 절대적으로 부족하고 신속한 문서처리를 위한 비용이 엄청나게 증가하는 이유로 전통적인 수작업에 의한 색인 작업은 현실적으로 역부족이다. 또한 수작업을 통해 색인 작업을 함으로써 문서에 대한 색인자나 색인 작업 시점 등에 따라 서로 다른 색인어를 선택하는 색인어 선정의 일관성 문제[10]가 발생하게 된다.

이러한 문제를 해결하기 위해 컴퓨터를 이용한 자동색인 기법이 출현하게 되었다. 자동색인 기법은 컴퓨터가 수집된 문서를 분석하여 각 문서의 주제를 대표할 수 있는 단어나 단어구를 자동으로 추출하고 이를 해당 문헌의 색인어로 부여하는 기법이다. 색인어를 선정하는 기준에 따라 통계적 기법, 언어학적 기법, 문헌 구조적 기법으로 구분할 수 있다. 언어학적 기법은 형태소 분석 수준에서 이루어지는 자동색인 기법과 구문 분석 수준에서 이루어지는 자동색인으로 나누어진다. 형태소 분석에 의한 자동 색인은 형태소 분석 결과로부터 주제를 찾아내는 명사나 구를 식별하는 방법[2,4]이고, 구문 분석 의한 자동 색인은 문장의 구문 구조를 분석한 다음 전치사구나 명사구 등을 이루는 단어군을 찾아내고 이 가운데 빈번하게 나타나는 구를 선택하는 방법[4]이다.

그러나 통계적 기법에 의한 색인은 일단 색인어 후보가 추출되었다고 가정하고 후보 색인으로부터 문헌을 대표할 수 있는 색인어를 선택하기 위한 기준으로 문헌과 문헌과의 유사성, 단어의 문헌 분별력 등을 이용한 색인기법[7]이다. 형태소 분석에 의한 자동색인은 문헌으로부터 색인어 후보가 될 수 있는 단어만을 추출하는 문서 내의 색인과정이다. 자동색인은 형태소 분석에 의해 후보 색인을 우선 추출하고 통계적 기법에 의해 색인어를 선정한다.

위와 같은 색인 기법들에 의해 생성된 색인의 요소 중 명사에 대한 처리가 필요하다. 명사는 단위명사 또는 복합명사로 구분된다. 그러나 복합명사는 다양한 단위명사와 결합될 수 있으므로 생성될 수 있는 복합명사의 수는 무한하다고 할 수 있다. 그러므로 이들 복합명사를 색인의 기본 단위로 하였을 때 축적된 문서보다 색인공간이 크게 되므로 비현실적이다. 따라서 검색엔진의 색인시스템의 색인 기본 요소는 단위명사가 적합하며, 복합명사는 형태소 분석과정에서 단위명사로 분해하여 색인 선정 과정에 적용된다. 특히 복합명사의 분해는 형태소 분석과정에서의 중요한 한 분야로서 다루어지고 있으며, 현재 복합명사 분해에 대한 다양한 방법론[2,8,9]이 연구되고 있다.

III. 색인 시스템 제안

검색엔진의 초기 버전에서는 검색의 효율을 높이기 위해 각 검색엔진 각각 나름대로의 검색 연산자들이 많이 있었다. 그러나 이들 연산자들은 검색엔진마다 그 의미나 사용법 등이 다르고 복잡하다. 결국 복잡한 연산자를 사용하는 검색 인터페이스는 점차 사라지고 자연어 형태의 문장이나 명사구, 단어들의 나열 형태의 질의어를 사용하여 원하는 자료를 검색하게 되었다. 사용자들이 입력하는 질의어의 유형을 보면 명사구, 인접 단어의 나열이 많다. 이러한 사용자의 질의어 유형 변화에 가장 적합한 것은 명사구 또는 전치사구를 색인의 단위로 하는 색인시스템이나 자연어를 기반으로 하는 색인시스템일 것이다. 그러나 이들은 구문 분석 및 의미 분석 등 복잡한 절차를 거쳐야만 한다.

형태소 분석을 통해 명사를 추출하고, 복합명사는 단위명사로 분해하며, 미등록어는 복합명사 분해 알고리즘을 통해 복합명사를 판별하고 단위명사로 분해하여 단어가중기법을 이용하여 주제어를 선정한다. 마지막으로 단위명사들 사이의 거리를 계산하여 일정 거리 미만의 단위명사들을 최종 선정하는 방법인 거리개념을 이용한 색인 시스템이다.

3.1 거리 계산 조건

주제어들 간의 거리 계산은 다음과 같은 규칙을 따른다. 첫째, 거리 계산에 사용되는 위치를 나타내는 단위는 문장과 어절로 한다. 질의어의 대부분은 단어를 나열하는 형태이거나 구의 형태가 대부분이므로 좌표는 (문장, 어절)형태로 하는 것이 적합하다. 두 번째, 복합명사에서 분리된 단위명사간의 거리는 (0, 0)이다. 예를 들면 "정보문화"이라는 복합명사를 분해화면 "정보", "문화"라는 두 단위명사로 분리된다. 이들은 실제 문서에는 하나의 어절을 이루고 있으므로 앞의 두 단위명사간의 거리는 (0,0)이다. 세 번째, 명사간의 유효 거리는 같은 문장 내로 한다. 비록 거리가 아주 가깝더라도 서로 다른 문장에 위치하고 있다면 두 명사간의 의미상의 연결 고리가 만들어지기 힘들다. 그러므로 두 명사 사이에는 의미상 관련이 없는 것이 많다. 다섯 번째, 사용자의 질의에서 비롯된 주제어의 역배열로 생기는 문제도 있을 것이다. 이 경우 초창기 네이버에서는 "키워드1 - 거리 키워드2" 형태의 연산을 지원했으나 본 논문에서는 일치하지 않는 것으로 한다.

3.2 전처리

웹문서에서 링크 및 강조된 영역은 주제어로 선택될 가능성이 가장 높은 부분으로 이후 이어지는 가중치 계산단계 이전에 이에 대한 전처리가 필요하다. 본 논문에서는 주제어 영역 테이블(Subject Area Table)을 두고, 링크 및 강조된 영역을 아래 방법에 따라 다음과 같은 전처리 단계를 거쳐 주제어 영역을 테이블에 (영역 시작 문장, 절, 영역 종료 문장, 절)의 형태로 등록하여 위치에 따른 가중치 계산 단계에서 이를 적용하도록 하였다. 단 영역이 중첩되거나 특정 영역에 포함되는 경우 넓은 영역만 등록하였다. 그리고 만일 영역이 교차할 경우 보다 앞에 위치한 영역을 우선 등록하며, 보다 뒤쪽에 있으면서 교차하는 영역은 앞 영역 다음부터 영역의 끝까지를 등록하였다.

먼저 불필요한 태그를 삭제한다. 특히 <SCRIPT> .... </SCRIPT>, <STYLE> ... </STYLE>, <!-- ... ---> 등 태그 영역의 모든 내용 및 아래에 명기되지 않은 태그는 삭제한다.

</BR> 태그와 <P> ... </P>에 대한 처리가 필요하다. 먼저 하나의 </BR>태그가 삽입되어 있을 경우 문자의 전후 사정을 고려하지 않고 삭제한다. 단 연속된 </BR>태그의 경우 단락을 바꾼 것으로 간주 문장이 끝난 것으로 하며 </BR> 태그 이전 문자가 문장부호 "."가 없으면 이를 삽입하고 연속된 </BR> 태그를 삭제한다. 그리고 <P> 태그 이전에 문장부호 "."가 없으면 문장이 끝나고 새로운 문장이 시작되므로 문장부호 "."를 삽입하고 <P> 태그를 삭제한다. 그리고 이후의 </P> 태그를 "."로 치환한다. 만일 단독으로 </P> 태그가 삽입되어 있을 경우 태그만 삭제할 뿐 "."는 삽입하지 않는다.

<TITLE>...</TITLE>, <STRONG>...</STRONG>, <B>...</B> 등은 명시적 강조 구문으로 주제를 나타내는 문장이나 주제어일 가능성이 높다. 그러므로 이 부분은 태그를 삭제한 후 주제어 영역 테이블에 등록한다. 단 앞의 태그 중 <STRONG> 및 <B> 태그의 경우 단순 강조로도 사용될 수 있으므로 내포된 문자가 명사 및 구 형태를 띠지 못하면 무시한다. 그리고 링크는 현재 문서와 다른 사이의 연관성을 나타내므로 이 또한 주제어가 내포될 가능성이 높으므로 전자와 같이 태그를 삭제한 후 주제어 영역 테이블에 등록한다.

<FONT>...</FONT>의 경우 문자의 생상 및 크기 두 가지의 강조효과를 가지고 있다 색상은 문서의 기본 색상과 달라야 하며 크기는 시스템 폰트보다 큰 경우 태그를 삭제한 후 주제어 영역 테이블에 등록한다.

스타일시트에 의한 문서의 강조는 매우 다양하므로 어떠한 태그에 어떠한 스타일이 적용되어 있는지 그리고 문서 전체에 적용되는 스타일이 있는지를 살펴야 한다. 특히 문서 전체에 적용되는 스타일은 어떤 특정 태그에 적용된 스타일이 강조를 위한 것인지 아닌지를 판별하는 기본이 된다. 만일 이 부분이 존재하지 않으면 웹 탐색 프로그램의 기본값으로 해야 할 것이다. 여기서 색상은 기본스타일 또는 시스템의 기본 설정 값과 다르면 강조된 것으로 하며 문자의 크기는 기본 스타일 또는 기본 설정 값보다 클 경우, 진함의 정도는 기본 스타일 또는 기본 설정 값보다 진할 경우 강조된 것으로 보아 주제어 영역 테이블에 등록한다.

이렇게 하여 주제어가 포함되어 있을 가능성이 높은 부분을 표시한 기호 외에 다른 기호가 일체 배제된 오직 텍스트만을 추출하여 1차 주제어 선정을 위한 입력으로 한다.

3.3 가중치 계산

가중치 계산은 1차 주제어를 선정하기 위해 문서 내의 각 문장의 위치와 문장 내의 용어의 유형 및 조사정보 그리고 빈도수를 이용하여 가중치를 계산한다. 그리고 본 논문에서는 각 단계별 가중치 범위는 0~1의 배정도 실수를 가지도록 한다. 각 단계별로 계산된 가중치는 최종 주제어를 얻기 위해 각 단계별 가중치를 적용하여 최종 가중치를 계산한다.

첫째, 문서 내의 단위 명사 출현 빈도만을 이용하여 가중치를 계산한다. 먼저 문서 내에서의 각 단위 명사의 빈도(Ti)를 구하고 이 중 빈도가 가장 큰 것(Tmax)을 구해 식 1을 이용하여 0과 1사이의 가중치(W)를 계산한다.

$$W = \frac{T_i}{T_{max}} \quad \text{(식1)}$$

둘째, 어절 단위에서 유형에 따라 가중치를 계산하기 위하여 복합명사, 미등록어, 보통명사, 1음절 명사, 부사성 명사, 숫자포함 용어 순으로 가중치를 부여한다. 이는 문서에서 유형별 명사들의 수를 구하고 이들 중 주제어로 선정된 명사들의 수를 구해 주제어 선정 비율을 구하였다. 그리고 이 중 최대값을 기준으로 각 유형별 주제어 선정 비율을 차감하여 가중치를 계산하였다. <표 1>은 신문기사와 기타 웹문서 그리고 국립국어연구원과 한글학회의 국문소설 및 자료 등 400여건의 문서 중 100건의 문서를 임의 선정하여 각 문서에서 주제어로 수용된 명사들을 조사한 결과 아래와 같은 결과를 얻었다.

표 1. 유형별 가중치
Table 1. Weight according to the kind of noun

구분	복합 명사	미 등록어	보통 명사	1음절 명사	부사성 명사	숫자포함 명사
명사수	376	296	117	40	37	75
비율	0.4	0.31	0.12	0.04	0.04	0.08
가중치	1	0.79	0.31	0.11	0.10	0.2

강승식은 유형별 가중치[2]를 부여함에 있어 복합명사에 1.0, 미등록어에 0.8, 보통명사는 음절수에 따라 0.3에서 0.8의 가중치를 부여하고 1음절 명사와 부사성 명사는 0.8, 그리고 숫자 포함 용어에는 0.2의 가중치를 부여하였다. 위와 같이 조사한 결과에서도 거의 동일한 결과를 얻었다. 단 복합명사는 분해되어 단위 명사 단위로 색인이 구성되므로 분해된 단위명사에 가중치를 부여하되, 동일 단위명사가 이미 존재할 경우 가중치가 높은 복합명사에서 분해된 단위명사의 가중치를 가진다. 그리고 조사의 역할에 따라 계산된 가조사의 가중치 또한 위의 100건의 문서에서 선정된 주제어들이 포함된 어절에서 조사의 출현 빈도를 조사하여 각 조사가 차지하는 비율을 조사하고, 각 조사의 빈도를 최고 빈도수로 나누어 <표 2>과 같은 가중치를 얻었다.

표 2. 조사 출현 빈도에 의한 가중치
Table 2. Weight according to the frequency of postpositional word

조사	은/는	이/가	의	을/를	만/도	에 /에서	기타
빈도	203	182	159	148	101	84	64
비율	0.22	0.19	0.17	0.16	0.11	0.09	0.07
가중치	1	0.90	0.78	0.73	0.50	0.41	0.32

셋째, 나열식 문서를 제외한 대부분의 문서는 주제어가 포함된 문장이 문서의 서두와 끝 부분에 위치하게 된다. 또한 웹문서의 경우 강조효과가 주어진 부분 및 링크 등은 문서를 대표하는 주제어일 가능성이 아주 높다. 하지만 이것은 독자의 주관에 따른 것일 뿐, 기계적으로 그 범위를 정하기 위한 기준을 제시하기란 매우 어렵다.

(그림 1)은 위의 100여건의 문서로부터 주제어의 출현 위치를 조사하여 문서의 크기에 따른 상대적 위치를 비율로 환산하여 주제어의 분포를 조사한 것이다. 여기서 문서를 신문기사, 웹문서, 학술논문 등 세 가지 부문으로 나누어 분포를 조사하였는데, 신문 기사는 문서의 처음 15% 범위에 주제어가 많이 분포하며, 웹문서는 처음 10% 범위에, 학술논문은 처음 25% 범위와 후반 10% 범위에 주제어가 많이 포함한 것으로 조사하였다.

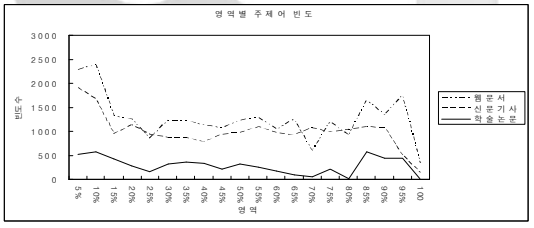

그림 1. 영역별 주제어 빈도
Fig 1. Frequency of subject word according to each area

그러나 문서의 유형별 특성에 따라 빈도수와 고빈도의 위치가 조금씩 다르다. 그러므로 통합하여 평균 빈도를 이용하는 것보다 각 문서의 유형의 영역별 누적빈도를 이용하여 가중치를 구하는 것이 바람직하다. 아래의 <수식 2>와 같이 문서의 크기(St)와 구간의 수(CN)을 이용하여 어절의 위치(p)의 구간(N)을 계산하고, 구간(N)의 누적빈도(TN)를 각 구간별 누적빈도 중 최대 빈도(Tmax)로 나누어 가

중치(식3)를 계산하였다. 단 어절의 위치(p)가 전처리 단계에의 마아크업 구간에 있다면 그 가중치는 1로 지정하였다.

$$N = \frac{p * CN}{S_t} \quad \text{(식2)}$$

$$P_w = \frac{T_N}{T_{max}} \quad \text{(식3)}$$

```
Function Weight_As_Position( CN, St, p )
    Begin
    If Is_Subject_Area_Using_Binary_Search( p )
        Then Weight_As_Position ← 1
        Else
            Begin
            N ← Int( p * CN / St )
            Tmax ← Max_Frequency_of_Subject_in_All_Area
            TN ← Frequency_of_Area( N )
            Weight_As_Position ← TN / Tmax
            End
    End
```

이상 앞의 세 가지의 가중치 계산 방법을 이용하여 각 가중치를 구하여 아래와 같이 각 부문별 중요도를 곱하고 합산하여 최종 가중치(Wt)의 구성 식은 아래 식4와 같다.

$$W_t = T_w \times 2 + S_w \times 3 + P_w \times 5 \quad \text{(식4)}$$

Tw : 단어 출현 빈도에 따른 가중치
Sw : 유형별 가중치
Pw : 위치에 따른 가중치

4.3.1 1차 주제어 선정

계산된 가중치를 바탕으로 주제어를 선정하는 기존의 방법은 가중치 순으로 정렬된 목록에서 일정 순위까지 주제어로 선정하는 방식을 취하였다. 이렇게 하면 문서의 분량 및 높은 가중치를 가지는 후보들이 탈락될 가능성이 높다. 그래서 본 논문에서는 위와 같이 정적인 선정방식이 아니라 (그림 2)에 나타난 것처럼 최대 가중치와 최소 가중치와 각 가중치별 최대 가중치와 최소 가중치로 향하는 일정한 속도(V)의 방향 벡터를 구하고 이 두 벡터의 합이 최대인 가중치를 선정 기준점으로 정해 기준 가중치 이상을 1차 주제어로 선정하는 방식을 이용하였다.

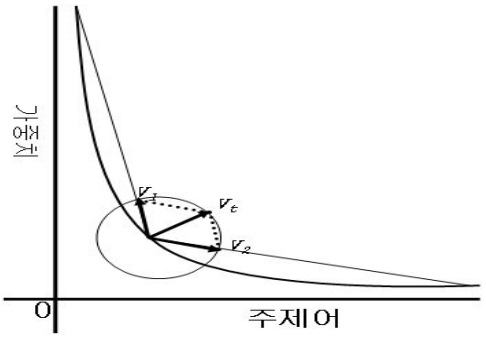

그림 2. 벡터를 이용한 1차 주제어 선정
Fig 2. Select 1st subject word using vector

앞의 가중치 계산방법에 따라 가중치를 계산하고 이를 내림차순으로 정렬하여 아래 좌표평면 상에 표시하면 위와 같은 곡선 형태로 나타나게 된다. 그리고 최대 가중치와 각 가중치를 잇는 직선의 기울기를 a, 최소 가중치와 각 가중치를 잇는 직선의 기울기를 b라고 하고, 두 직선의 교점을 원점으로 옮겼을 때, 속도가 V인 각 방향 벡터의 식은 아래 식5와 식6과 같다. 단 아래 벡터의 좌표는 직선 $y = ax$ 또는 $y = bx$의 직선상에 있어야 하며 최대 가중치 또는 최소 가중치와 가까워야 한다. 이와 같이 구해진 벡터의 합(식7)이 최대인 가중치까지를 1차 주제어로 선정한다.

$$\overrightarrow{V_1} = \left(\pm\sqrt{\frac{V^2}{a^2+1}}, \pm\sqrt{\frac{a^2V^2}{a^2+1}}\right) \quad \text{(식5)}$$

$$\overrightarrow{V_2} = \left(\pm\sqrt{\frac{V^2}{b^2+1}}, \pm\sqrt{\frac{b^2V^2}{b^2+1}}\right) \quad \text{(식6)}$$

$$\overrightarrow{V_t} = \overrightarrow{V_1} + \overrightarrow{V_2} \quad \text{(식7)}$$

3.3.2 최종 주제어 선정

최종 주제어를 선정하기 위해 1차 주제어로 선정된 최종 주제어 후보를 대상으로 각 후보가 출현한 위치를 어절 단위로 추출하여 후보색인어와 문정번호, 문장에서의 어절 위치를 가지는 위치정보를 순서쌍으로 구성 문장번호와 어절 위치를 기준으로 오름차순 정렬하여 위치정보 리스트를 구성하고, 각 후보가 다른 후보로부터 선택되었는지를 판별하기 위해 색인어와 후보 선택 수를 순서쌍으로 하는 선택정보 리스트를 구성하였다. 최종 주제어는 이 리스트에서 다른 후보로부터 선택된 것만 선정하도록 하였다.

표 3. 최종 주제어 선정을 위한 리스트
Table 3. List for selecting final subject word

후보 색인	선책 수	후보 색인	문장 번호	어절 위치
가정	0	가정	1	1
교육	0	교육	1	1
개국	0	학부모	2	1
시조	0	교육	2	2
개발	0	형태	2	3
차관	0	학교	3	1
객원	0	학원	3	2
교수	0	가정	3	6
거인	0	교육	4	2
설화	0	인식	4	3
검찰	0	정부	4	6
•		•		
•		•		
•		•		

위와 같이 <표 3>과 같이 구성된 리스트를 바탕으로 각 후보 색인과 거리를 계산하여 일정한 범위 속하는 후보를 선택하고 선택된 후보 선택 정보 리스트의 선택 수를 증가시켜 해당 후보가 선택되도록 하였다. 그리고 각 단위 명사와의 거리를 보면 두 개 이상의 단위명사를 붙여 쓴 복합명사의 경우 거리는 0이며 띄어 쓴 경우 1로 계산하였다.

검색엔진을 이용하는 사용자의 검색어로 미루어 볼 때, 고유명사, 복합명사 그리고 두 단위명사를 띄어 쓴 경우가 대부분이다. 그러나 두 개 이상의 단위명사가 분리되어 쓰일 경우, 단위명사 사이에 앞 혹은 뒤에 오는 단위명사를 수식하여 그 의미를 더욱 강조하는 동사, 관형사, 부사, 형용사 등이 올 수 있다. 이들 수식 어절은 그 표현에 따라 그 수가 무한히 늘어날 수 있다. 그러므로 수식 어절의 경계를 찾기란 매우 어렵다. 본 논문에서는 이를 반영하여 명사 선택 범위를 같은 문장 내에서 명사를 수식하는 어절을 포함하여 거리가 3어절 이내로 제한하였다.

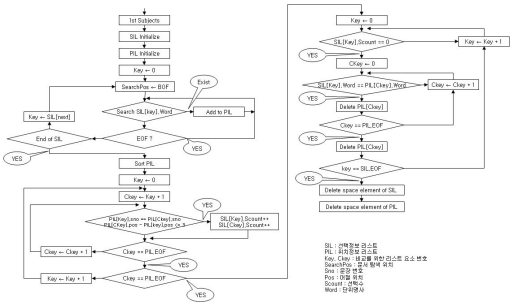

그림 3 최종 주제어 선정 과정
Fig 3. Flowchart for selecting final subject word

주제어로 선정된 후보들을 이용하여 각 후보의 선택 정보를 가지는 리스트를 구성한다. 그리고 이 리스트를 바탕으로 문서 내에서 각 후보가 위치한 문장과 어절 위치를 모두 추출하여 위치 정보 리스트에 추가하고 문장과 어절을 기준으로 오름차순 정렬하여 준비한다. 처음에는 위치 정보 리스트의 첫 번째 요소를 중심으로 그 다음 요소들을 탐색하여 거리가 같은 문장의 3어절 이내에 있으면 선택 정보 리스트에 중심요소와 탐색된 요소의 선택 수를 증가시킨다. 그렇지 않으면 더 이상 탐색을 하지 않고 중심요소를 위치 정보 리스트의 다음 요소로 지정하고 앞에서 수행했던 탐색 및 선택 수를 증가시키는 작업을 한다. 이러한 작업을 중심요소가 맨 위치 정보 리스트의 맨 마지막 요소가 될 때까지 반복 수행한다.

위의 탐색 및 선택 수 등록 작업이 수행 된 후 선택정보 리스트의 각 요소들을 조사하여 요소의 선택 수가 0인 요소는 삭제함과 동시에 위치 정보 리스트에서도 삭제한다. 그리고 삭제되지 않은 선택 정보 리스트의 각 요소들을 최종 주제어로 선정하여 전체 색인 리스트에 등록하고 위치 정보 리스트의 각 요소는 문서 정보 리스트에 등록한다.

이상 최종 주제어를 선택하는 절차를 살펴보았다. 여기서 전체 색인 리스트 외에 문서 정보 리스트를 따로 보관하는 것은 질의어를 이용한 검색에서 두 개 이상의 키워드가 동일 문서에 출현 했을 때, 전후 관계 및 정확도를 판단하는 근거로 사용하기 위함이다.

3.4 실험 및 분석

본 논문에서 제안한 색인 선정 기법은 형태소 분석을 그 토대로 하고 있다. 현재 사용이 자유로운 형태소 분석 모듈이 많이 있다. 이 중 강승식의 형태소 분석 모듈HAM은 동적 라이브러리 형태로 되어 있어 본 논문의 자동 색인 모듈에 쉽게 적용할 수 있는 장점을 가지고 있다. 그래서 형태소 분석 단계에서는 HAM을 이용하였다. 그리고 형태소 분석을 거쳐 품사 태깅된 문서에서 복합명사를 추출 및 분해 알고리즘을 이용하여 분해하였다. 그리고 위 주제어 선정 절차에 따라 최종 주제어를 선정하는 자동 색인 프로그램을 구현하였다.

본 논문에서 제안한 자동 색인 시스템의 결과 분석 및 평가는 수동색인과 자동 색인 시스템에 의해 얻어진 색인을 비교하여 색인 시스템의 성능을 평가하였다. 그리고 1차 주제어 선정 단계에서의 성능과 최종 주제어 선정 단계의 성능을 비교하여 각 유형의 문서에 따른 성능 향상을 조사하였다. 수작업으로 구성된 색인(Hi)을 바탕으로 자동 색인 시스템에 의해 얻어진 색인(Ai)을 비교 일치하는 색인의 수를 구해 일치율과 비일치율을 구하고 이를 바탕으로 시스템의 성능을 평가할 수 있는 정확율을 구하였다.

$$\text{일치율} = \frac{count(Ai \in Hi)}{count(Ai)} \quad \text{(식8)}$$

$$\text{비일치율} = 1 - \text{일치율} \quad \text{(식9)}$$

$$\text{정확률} = \frac{count(Ai \in Hi)}{count(Hi)} \quad \text{(식10)}$$

위의 식을 이용하여 아래와 같이 신문기사, 웹문서, 학술논문, 그리고 국문 자료 등의 문서 분류하여 문서들의 평균 일치율, 비일치율, 정확율을 아래의 <표 3>와 같이 보였다.

표 4. 색인 시스템의 성능
Table 4. efficiency of Indexing System

구분	1차 주제어 선정 단계			최종 주제어 선정 단계		
	일치율	비일치율	정확률	일치율	비일치율	정확률
신문 기사	0.7283	0.2717	0.9018	0.7419	0.258	0.9203
웹 문서	0.5672	0.4328	0.6053	0.647	0.353	0.7419
학술 논문	0.6251	0.3749	0.7	0.7916	0.1666	0.9500

위 결과를 보면 1차 주제어 선정 단계와 최종 주제어 선정 단계의 성능의 차이를 보면 신문기사는 1.85%, 웹문서는 13.66%, 학술논문은 25%의 성능 향상을 나타냈다. 이는 기존 문서의 내의 통계적 가중치를 적용한 자동 색인에 비해 높은 정확률을 보였다. 그리고 최종 주제어 선정 단계에서 학술논문 및 신문기사는 아주 높은 정확률을 가지는데 이는 내용 전달을 위해 주제어가 초반부 및 후반부에 위치하는 일정한 패턴을 가지고 있기 때문이다. 그리고 웹문서의 경우 앞의 두 문서에 비해 훨씬 낮은 정확률을 가지는데 이는 문서의 표현 방법이 너무나 다향하며 이미지 및 ActiveX 등 다양한 시각적 효과를 위한 매체의 이용이 급격히 늘어나고, 색인 시스템의 근원이 되는 텍스트의 비중이 낮아짐에 그 원인이 있다.

아래의 <표 5>에 본 논문에서 제안한 자동 색인 기법과 정천영의 자동 색인 기법[11]의 성능을 비교하였다. 그리고 비교 대상은 논문으로 한정하였다. 이는 기존 색인 시스템의 대부분이 논문 초록을 대상으로 실험 및 평가를 진행해 왔기 때문이다.

표 5. 자동 색인 시스템의 비교
Table 5. Compare cappacity of auto indexing System

구분	일치율	비일치율	정확률
제안된 색인시스템	0.7916	0.1666	0.9500
기존 색인시스템	0.5462	0.4538	0.8993

위 표에서 알 수 있듯이 본 논문에서 제안한 색인 시스템에서의 정확률이 기존의 것보다 더 나은 성능 향상을 보였다.

Ⅳ. 결론

현재 많이 이용하고 있는 검색엔진의 경우 복합명사 분해 분문에서는 많은 발전을 거듭하여 매우 정확하였다. 그러나 질의어에 대한 검색 결과는 그 정확성이 많이 떨어지는 것으로 조사되었다. 이는 검색 로봇이 변경된 웹문서의 정보를 적절하게 반영하지 못하여 페이지 오류가 나거나 전혀 다른 문서가 나오는 것에서 1차적 원인이며, 검색엔진에 탑재되어 있는 기존 색인 시스템의 낮은 성능으로 인해 두 개 이상의 어절로 이루어진 질의어에 대해 각 어절의 관련성 보다 단순히 존재 여부만을 판별하여 보여주므로 전혀 다른 내용의 문서가 검색되는 것에 그 원인을 찾을 수 있다.

한국어 정보검색의 그 근간을 이루고 있는 것은 형태소 분석이다. 이를 기반으로 구문 분석, 의미 분석 및 자동 색인에 대한 많은 연구가 이루어지고 있다. 본 논문에서도 주제어를 추출하는 과정에서 형태소 분석 기법을 이용하고 있다. 그러나 아직 형태소 분석에서 복합명사의 분해 분야는 분해 후 의미의 모호성과 여러 가지의 단위 명사로 분해 될 수 있어 지금도 활발한 연구가 진행되고 있다.

자동 색인 시스템 분야에서도 많은 연구가 진행되고 있으며 색인 구축에 대한 여러 가지 방법들이 나오고 있다. 색인의 주목적은 사용자의 질의에 대해 얼마나 정확한 내용을 빠르게 검색하느냐에 있다. 본 논문에서는 기존 색인 방법에서 불필요한 색인어를 배제하고 각 색인어 사이의 관계를 고려하여 서로 연관성이 높은 색인어들만을 추출하였다. 본 논문에서 제안한 색인 시스템을 통해 네 종류의 문서들을 조사한 결과 신문 기사는 92.03%, 학술 논문은 95%의 아주 높은 정확율을 나타내었다. 그러나 웹문서는 73.33%의 비교적 낮은 정확율을 나타내었는데, 이는 문서의 표현 유형이 너무나 다양하고, 텍스트가 아닌 너무나 다양한 멀티미디어 요소를 동원하여 어떠한 규칙에 얽매이지 않고 시각적 효과를 중시하는 경향으로 인해 색인을 추출하는데 많은 어려움이 있다.

위 결과를 보면 웹문서의 정확율이 현저히 떨어져 전체 색인 시스템의 성능을 떨어뜨리는 요인이 되고 있는데, 이를 극복하기 위해서는 다음 두 가지의 연구 과제가 있다. 첫째, Table(표)의 이용방법에 있다. 일반적인 표는 본 논문과 같이 자료를 정형화하여 표현하는 용도로 사용하나 웹문서에서는 내용의 배치 및 시각적 표현을 위해 복잡한 Table 구조를 가진다. 이러한 구조에서 정형화된 자료를 가진 표를 분리하는 연구와 이미지, Flash, 기타 ActiveX 등의 미디어로부터 색인 요소를 추출하는 연구가 필요하다.

참고문헌

[1] 하창승, 류길수, “사례기반 추론을 이용한 지능형 웹검색 에이전트의 설계 및 구현”, 한국컴퓨터정보학회 논문지 8권, 1호, 2003

[2] 강승식 “한국어 형태소 분석과 정보 검색”, 홍릉과학출판사, pp.1-581, 2003

[3] 채영숙 최성필 서정현, “자동 색인을 위한 한국어 형태소 분석기의 실제적인 구현 및 적용”, 정보처리학회논문지B, 9권, 5호, pp.689-700, 2002

[4] 서은경, “구문 . 통계적 기법을 이용한 한국어 자동색인에 관한 연구”, 정보관리학회지, 10권, 1호, pp.97-124, 1993

[5] 김용성 우선미 유춘식 유철종 이종득 권오봉, “자연어처리 , 통계적 기법, 적합성 검증을 이용한 자동색인 시스템에 관한 연구”, 정보처리학회논문지, 5권, 6호, pp.1552-1562, 1998

[6] 송재관, 박찬곤, “기계번역용 한국어 품사에 관한 연구”, 한국컴퓨터정보학회 논문지 5권 4호, 2000

[7] 장한명 김재현 박정기, “문헌분리가에 의한 한글 문헌 자동 색인 시스템 구현에 관한 연구”, 한양대학교 기초과학연구소 기초과학논문집 13권, pp.31-38, 1994

[8] 백현철 류방 김상복, “멀티미디어 정보처리 : 접사정보 및 선호패턴을 이용한 복합명사의 역방향 분해 알고리즘”, 멀티미디어학회논문지, 7권, 3호, pp.418-26, 2004

[9] 이현민 박혁로, “복합명사의 역방향 분해 알고리즘”, 정보처리학회논문지B 8권, 4호, pp.357-364, 2001

[10] 이영자 “주제분석기법으로서의 자동색인”, 한국도서관정보학회지, 12권, 0호, pp.61-96, 1985

[11] 정천영 장수진 진성일, “한국어 자동색인 시스템 설계에 관한 연구”, 충남대학교 자연과학연구소, 19권, 1호, pp10-pp18, 1992

저 자 소 개

박 찬 이

1993 진주교육대학교 초등교육학과 학사

2001 진주교육대학교 컴퓨터교육학과 석사

현재 : 경상대학교 컴퓨터과학과 박사과정 수료, 진주교육대학교부설 초등학교 교사

<관심분야> 유무선통신, 한국어 정보처리, 멀티미디어통

김 상 복

1989 중앙대학교 전자공학과 박사

현재 : 경상대학교 컴퓨터과학과 교수, 경상대학교 컴퓨터 정보통신연구소 연구원

<관심분야> 멀티미디어 통신, 한국어 정보처리, 컴퓨터 구조