색인어 말뭉치 처리를 기반으로 한 웹 정보검색 시스템의 설계

송정동 · 이정현 · 최준혁

평택공과대학 전산정보처리과 교수
인하대학교 전자계산공학과 교수
김포대학 전자계산과 전임강사

요 약

대부분의 정보검색시스템들은 부적절한 색인어들에 의해 가끔 사용자의 의도에 맞지 않는 전혀 다른 검색 결과가 나타난다. 그것은 시스템이 색인어들을 검색하기 위해 그 의미가 아닌, 단지 용어로서만 고려하기 때문이다. 검색 정확도의 증진을 위해 색인어는 연관된 용어 사용 빈도와 역 빈도 사용으로 검색되고 동시 발생어는 원시 문서로부터 추출된다. 결과적으로 색인어는 계산된 상호 정보들을 사용함으로써 그들의 세맨틱에 의해 클러스팅된다.
이 논문은 재현율의 감소없이 클라이언트 사용자 모듈로부터 피드백에 따라 세분된 세맨틱 정보를 사용하여 부적절한 검색 결과를 거절함으로써 검색 효율을 높일 수 있도록 설계하였다.

Ⅰ. 서론

정보검색(IR; Information Retrieval)이란 수집된 정보 또는 정보 자료의 내용을 분석한 뒤 적절히 가공하여 축적해 놓은 정보 파일로부터 이용자의 정보 요구에 적합한 정보를 탐색하여 찾아내는 일련의 과정을 의미한다[14]. 이 과정에서 질의에 사용된 색인어들이 문서에 대하여 어느 정도의 중요도를 가지고 존재하느냐를 기준으로 문서를 사용자의 요구에 적합한 문서들이 먼저 출현되도록 순서화한다. 그러나 실제 순서화된 문서들을 보면 질의한 내용과는 다른 문맥의 문서들이 상위로 순서화되는 경우를 볼 수 있다. 이는 질의 색인어들이 문서에서 다루고 있는 문맥과 반드시 일치하지 않기 때문이다[5].

이러한 문제들은 컴퓨터 관련 분야 종사자들의 경우에는 불리언 질의문을 최적화하여 재작성하거나 각 정보검색 시스템별로 제공하는 기능을 이용하여 관련 정보들에 대한 추가적인 입력을 함으로써 어느 정도 해결을 하고 있다. 그러나 최근 인터넷의 대중화와 더불어 정보검색 시스템의 사용자 층이 컴퓨터 관련 분야 종사자에서 일반인으로 점차 확대되면서 시스템을 이해하지 못하는 다수의 사용자들에게 그 문제점이 크게 부각되고 있다.

현재 국내의 정보검색 연구의 주된 대상으로는 복합명사 처리, 자연언어 입력 처리, 순위 인접 연산 처리 등을 들 수 있으나, 이들에 관한 연구가 몇몇 대학과 연구소를 중심으로 진행되고 있고, 한편으로는 시소러스(thesaurus) 및 개념에 의한 검색 시스템에 관한 연구도 시작 단계에 접어들고 있는 실정이다. 그러나, 이러한 대부분의 연구들이 색인어의 선정에 중점을 두고 진행되고 있으며[7][10][11], 중의성을 가지는 색인어가 발생하였을 때나 형태소 분석시 발생하는 모호한 어휘에 대한 문맥상의 의미적 고려가

논문접수일 : 97년 10월 게재확정일 : 98년 2월

미흡한 실정이다.

이러한 문제를 해결하기 위하여 정보검색 시스템의 정확도를 향상시키게 되면 일반적으로 재현률(recall ratio)과 정확도(precision)가 반비례 관계를 가지기 때문에 재현률 저하가 발생한다[4][14]. 그러나 본 실험에서는 문서를 대표하는 색인어 추출을 기존 검색 시스템이 사용하는 상대 출현 빈도를 이용함으로써 재현률을 저하시키지 않을 수 있었고, 추출된 색인어에 대하여 원 문서를 대상으로 공기 정보를 추출하고 각 공기 단어에 대하여 상호정보량을 계산하여 동음 이의어 색인어를 의미별로 분류함으로써 정확도를 향상시킬 수 있었다.

재현률을 저하시키지 않고 정확도를 향상시키기 위해서는 정보검색 시스템의 서버 측에 사용자의 질의 의도를 피이드백해 주어야 하며, 이를 위하여 클라이언트 측에서는 단순한 브라우징 기능만을 가지는 웹 브라우저 이외에 사용자 파라미터를 서버 측에 전송하고, 사용자 선호도 프로파일을 지속적으로 갱신하는 클라이언트 사용자 모듈을 포함하고 있어야 한다.

본 논문에서는 클라이언트에 사용자 모듈이 추가되고 이와 통신하는 서버측 모듈이 보완된 웹 정보검색 시스템을 제안한다. 본 시스템에서 서버는 클라이언트로 서버 측의 색인어의 의미 정보를 전송하고 클라이언트로부터 사용자의 질의 의도를 피이드백을 받아 사용자의 의도와 상관없는 문서들을 검색 결과에서 제외하고 클라이언트로 제공하는 과정을 수행한다. 실험 및 평가에서는 서버측 색인어의 의미를 자동 추정 분류하여 이를 다시 피이드백 받을 수 있도록 서버와 클라이언트가 통신하는 문서 필터링의 정량적 선택 기법을 실험 및 평가한다.

II. 기존의 정보검색 시스템 그 문제점

정보검색 시스템은 검색되는 정보의 유형에 따라 데이터 검색 시스템(data retrieval system), 참조 정보 검색 시스템(reference retrieval system), 본문 검색 시스템(full-text retrieval system), 질문 응답 시스템(question-answering system), 비디오텍스(videotex) 등이 있다[14][15].

기존의 DBMS와 정보검색의 유사점과 차이점은 다음과 같이 몇 가지로 요약할 수 있다. 첫째, 정보검색에서 다루는 개체(object)는 텍스트형(textual)의 문서를 대상으로 하고, DBMS에서는 레코드(record)와 같은 구조적(structural) 형태의 데이터를 대상으로 한다. 따라서 정보검색에서는 문서를 검색하기 위해서는 문서의 내용을 분석하여 효율적으로 주요어를 추출하는 것이 문제이고, DBMS에서는 수동으로 분석된 문서에 대해 정보를 효율적으로 저장하는 부분이 문제이다. 둘째, 정보검색의 특징은 검색이 확률(probabilistic)에 기반한다는 점이다. 따라서 사용자가 질의시 검색된 문서가 원하는 문서일 수도 있고 아닐 수도 있다. 또한 원하는 문서를 찾지 못할 수도 있다. 하지만 DBMS에서는 결정적(deterministic)이다. 즉, 질의시 주어진 속성-값(attribute-value)과 일치하는 레코드가 있으면 반드시 찾을 수 있다. 셋째, 유사점으로는 정보검색과 DBMS 둘 다 대량의 문서를 대상으로 하는 정보 시스템 이라는 것이다. 또한 문서의 첨가, 변경, 제거 등이 발생할 때 이를 처리하는 데이터 구조와 알고리즘이 필요하게 된다[6].

국내의 정보검색 시스템에서 공통적으로 처리

되는 기능은 구 단위 처리와 불리언 연산, 전문 검색, 복합 명사 처리 등이고, 인접어 연산, 개념, 필드, 본문 요약, 정렬 등의 기능은 전무한 상태이다. 일부 시스템에서 한글 처리 확장 기능으로서, 오류 정정, 유의어 발음 확장 기능 등을 탑재하고 있으나 사전 대조식의 처리로 인하여 한계가 있으며, 입력 문장을 해석하여 검색하는 자연언어 질의 처리 또한 미흡하다.

이러한 정보검색 시스템에서 영어 문서의 경우에는 색인어를 추출하기 위하여 스테밍(stemming) 기법 만을 적용하여도 문제가 없다. 그러나 한국어 정보검색 시스템은 교착어라는 특징 때문에 색인어를 추출하기 위한 연구 및 이와 관련된 색인 기법이 영어권 시스템에 비하여 중요하다.

Ⅲ. 색인과 자동색인

색인이란 어떤 문헌에 대해 그 문헌의 전체적 내용을 나타내거나, 그 문헌을 다른 문서들로부터 구별할 수 있도록 그 문서의 선택 단서가 되는 단어 또는 단어구 등을 추출하는 것을 말한다. 이는 각 문헌을 구분 지을 수 있는 대표 어구를 각 문헌에 부여하는 것을 의미하며, 동시에 검색에 이용될 경우에 유용한 어구를 추출하는 것을 목적으로 한다[15].

종래에는 도서관이나 정보 관리 부서에서 잘 훈련된 사서에 의하여 색인을 수행하였다. 그러나 방대한 양의 문서를 대상으로 정보검색을 수행하기에는 비용이 많이 소요되고, 주어진 시간 내에 색인할 수 있는 문서의 수에 한계가 있으므로 사서에 의한 수동 색인은 적합하지 않았다. 이러한 문제를 극복하기 위하여 대상 문서로부터 검색 시스템에 유용한 문헌의 주제어나 핵심어를 컴퓨터를 이용하여 자동으로 찾아내는 연구를 자동색인이라 한다.

자동색인은 색인어를 추출하는 자연어 처리 기법에 따라 어휘적 단계, 구문적 단계, 어의적 단계의 세 가지로 분류할 수 있는데[14], 어휘적 단계로는 형태소 분석을 통하여 색인어를 추출하고 모호성이 발생하는 어절을 형태소 구조 규칙에 의해 해결하거나[11], 구문적 단계로 구문 해석을 통해 격을 추출해서 적용함으로써 정확한 색인어를 추출하고자 하는 노력이 있었다[16]. 이러한 연구는 색인어에 대한 의미 특성을 고려하지 않은 상태에서 색인어 추출 및 순위 부여를 시도하고 있기 때문에 검색 결과에 대한 정확도 향상을 기대하기는 어렵다.

어의적 단계로는, 한글 키워드의 모호성을 해소하고 정보검색 시스템의 정확도를 향상시키기 위해 키워드 외에 명사구와 간단한 문장을 포함하는 키팩트 개념을 도입하기도 하였다[12]. 키팩트의 경우 색인어 이외에 색인어를 수식하는 간단한 명사구를 추출함으로써 동음 이의어에 의한 모호성은 해결할 수 있었으나, 수식되는 어휘 분석에 한계가 있으므로 결국은 재현률을 저하시키게 된다.

어의 모호성 해소를 위하여 의미적 관련 정보를 백과사전으로부터 습득하여 상호 정보개념을 도입하여 어의 모호성을 해소하는 시도가 있었으나, 백과사전의 표제어와 표제어 풀이 문장에 쓰인 단어들의 의미적 관계를 정확히 하기 위하여 마크업 기호들을 이용하여 수작업에 의해 백과사전을 마킹해야 하는 제약이 따른다[13]. 이러한 제약은 웹 문서의 경우에는 문서의 내용이

수시로 변하기 때문에, 어의 모호성을 해소하기 위해 매번 수작업에 의해 마킹을 한다는 것이 현실적으로 불가능하다는 단점이 있다.

본 논문에서는 문헌을 대표하는 색인어라는 제한된 어휘에 한하여 상호 정보량을 계산하여 공기 단어를 추출하고 추출된 공기 단어를 대상으로 상호 정보량을 계산함으로써 별도의 사전이나 의미 태그가 없이도 색인어의 의미를 추정에 의하여 정확도를 향상시킬 수 있으며, 또한 색인어를 추출하는 과정에서 기존 시스템이 사용하고 있는 상대 빈도에 의한 통계적인 기법을 적용하여 재현률을 저하시키지 않는 시스템을 설계 및 구현한다.

Ⅳ. 시스템 설계

4.1 색인어 추출

정보검색 시스템을 구성하는 각 요소들 중에서 색인어 추출 기법과 색인어 자료구조가 가장 중요한 요소가 된다. 이는 색인어를 추출하는 기법의 경우 시스템의 정확도와 신뢰성을 높여주는 요소이며, 색인어 자료구조는 시스템의 성능을 좌우하게 되기 때문이다. 색인어 추출 알고리즘은 시스템이 운영 중에도 오프라인(off-line)으로 수행이 되기 때문에 속도와 관련되는 항목은 크게 중요시되지 않는 반면 색인어 자료구조는 사용자 질의에 대한 검색 결과를 실시간(real-time)으로 제공해야 하기 때문에 시스템의 자원과 밀접하게 관련이 되며, 또한 시스템 전체의 반응 속도에 결정적인 영향을 미치는 요소이다.

본 4절에서는 시스템의 정확도와 신뢰성을 높이기 위한 색인어 후보 추출기법과 시스템의 성능을 향상시키기 위한 다양한 자료구조를 고찰하고 각 자료구조의 단점을 보완하기 위한 확장된 색인어 자료구조를 설계한다.

색인어 후보를 추출하는 기법은 단순히 조사나 어미를 제거하고 남는 어절들로부터 명사를 추출하는 방법과 형태소 분석후 품사가 명사로 추정되는 어절들을 대상으로 색인어를 선정하는 방법이 있다. 이러한 방법들중 단순히 조사나 어미를 제거하는 방법은 고유명사나 사전에 수록되지 않은 단어를 인식하지 못하게 되므로 잘못된 색인어 후보를 색인어로 선정하는 오류를 범하게 된다.

본 연구에서는 형태소 분석을 통하여 선정되는 색인어 후보를 대상으로 형태소 구조 규칙에 의하여 불필요한 후보를 제거하는 과정을 거쳐 색인어 선정 과정의 오류를 감소시켰다. 형태소 구조 규칙은 1차적으로 불가능한 형태소 규칙이 정의된 유한 상태 오토마타(FSA; Finite State Automata)와 2차적인 우선순위를 가지는 조합 가능한 형태소 구조 규칙이 정의된 유한 상태 오토마타에 의해 수행된다. 오토마타는 텍스트로 표현되는 리스트 구조를 가지며 그 일부는 [그림 1]에 나타나 있다. 프로그램 내부에서는 링크드 리스트에 의한 FSA로 변환되어 수행되지만 유지 보수 측면을 고려하여 텍스트 파일로 표현된다. 불가능한 형태소 분석 결과는 후처리에서 제거되며, 생성 가능한 형태소 규칙중 명사 후보로 추정되는 형태소를 모두 색인어 후보로 중복 선택한다. 이는 정확한 색인어를 추출하기 위한 시스템의 추가적인 오버헤드를 감소시킬 수 있을 뿐 아니라 정보검색 시스템의 특

성상 색인어 후보가 잘못 선정되어도 상대 출현 빈도에 의하여 색인어에서 제외되기 때문이다.

```
{N, ND}, ...
{PN, ND, SF}, {PN, ND}, ...
{IDA, EA}, {IDA, EU, EA}, {IDA, EE}, {IDA, EU, EE}, ...
{SFC, IDA, EPF, EU, EA}, {SFC, IDA, EPF, EU, EE}, ...
...
```

[그림 1] 형태소 구조 규칙의 일부

색인어는 문서에서 키워드가 될 수 있는 단어를 의미한다. 따라서 문서에 존재하는 모든 단어가 색인어가 되지는 않는다. 그러므로 색인어 결정 기법은 문서에서 이용되는 언어에 따라 처리 방법에 차이가 발생한다. 영어권의 색인어 결정 기법이 한글 문서의 색인어 결정 기법으로 적합하지 않은 것은 한국어가 교착어라는 특징뿐 아니라 영어권에서는 심각하게 발생하지 않는 복합명사에 대한 고려가 필요하기 때문이라고 볼 수 있다.

색인어를 선택하는 기법은 통계적 기법과 언어학적 기법, 문헌 구조적 기법 그리고 통제 어휘집 기반 기법 등이 있다. 자동 색인에 적용되는 언어학적 기법은 다시 어휘적 단계, 구문적 단계, 어의적 단계로 나눌 수 있다.

통계적 기법은 단어의 출현 빈도에 근거하여 주제어로서의 중요도를 측정한 다음 색인어를 선정한다. 주제어로서의 중요도를 측정하는 방법은 다양하며, 이 측정 방법의 기본 가설은 단어의 출현 빈도가 높을수록 그 단어가 문헌의 주제를 대표할 확률이 높다는 것이다.

영어권의 시스템은 비교적 어절의 구성이 단순하므로 통계적인 방법을 많이 이용하고 있다. 그러나 교착어의 특징을 가지는 한국어에 대해서 단순하게 통계적인 방법만을 사용하게 되면 시스템의 성능을 저하시키게 된다.

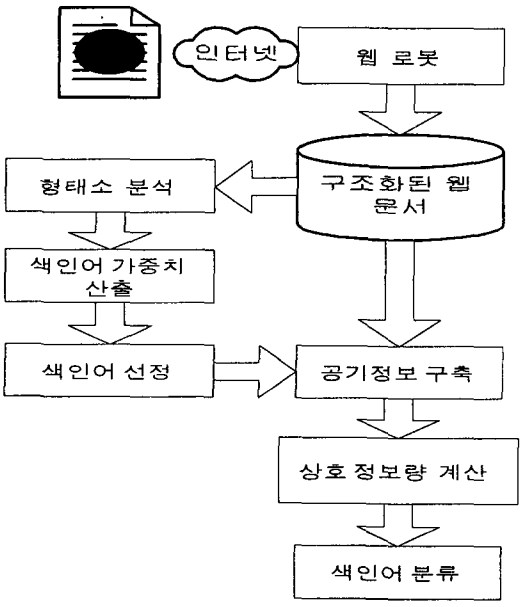
본 논문에서는 한국어의 특징을 반영하기 위하여 형태소 분석을 통해 추출된 품사를 기준으로 형태소 구조 규칙에 의한 후처리를 함으로써 통계적인 방법 뿐 아니라 한국어의 특징을 반영한 구조적 방법을 병행하여 사용하고 있다.

언어학적 기법은 어휘 분석 후 보통 불용어 제거 기법을 적용하며, 사전을 이용하여 단서어를 찾든지 구문 분석을 수행하여 색인어를 찾아내는 기법이다. 단서어 기법은 보통 '결론', '결과', '요약', '입증하다' 등과 같이 문헌의 주제를 축약적으로 표현해 주는 특정 의미의 단어를 찾아 그 단어가 출현한 문장 속에 함께 출현한 단어들을 색인어로 선택하는 방법이다. 또는 구문 분석을 수행한 후 구문의 격을 파악하고 필수 격에 해당하는 단어의 어근을 색인어로 취할 수도 있다.

문헌 구조적 기법은 문서 속에 단어가 나타난 위치에 의해 색인어를 선정하는 방법으로, '서론', '결론', '요약' 등의 제목을 갖는 특정한 부분에 나타난 단어들을 색인어로 선택하는 방법이다.

통제 어휘집 기반은 시소러스와 같은 통제 어휘집을 이용하여 문서에 나타난 단어들을 파악하여 어휘집에 있는 대표 단어로 개념 색인을 할 수 있는 방법이다. 정보검색에서 사용하는 시소러스는 색인시와 검색시에 사용할 수 있다. 색인 작업시에는 적절한 색인어의 선택과 색인어의 통제를 위하여 필요하며, 검색시에는 적절한 탐색 용어의 선택을 위해 필요하다. 그리고 용어 통제 이외에 탐색어의 확장이나 축소를 통하여 검색 효율을 조절하는 데에도 사용한다. 즉, 용어간의 계층 관계 및 연관 관계를 이용하

여 포괄적인 탐색을 하거나 특정한 용어를 사용하여 보다 한정된 탐색을 함으로써 검색 문서의 수를 적절하게 조절할 수 있다. 그러나 시소러스를 만들고 최신성을 유지, 관리하는 데 많은 노력이 필요하다. 최근 자연언어 처리 기술의 발전으로 통제 어휘집을 사용하여 색인하기 보다는 자연언어를 색인하는 것이 효과적일 수 있다. 자연언어 색인은 최신성을 유지하기 용이하고 입력 비용이 저렴하며, 데이터베이스간 교환이 용이하기 때문이다[6].



[그림 2] 색인어 추출 및 분류

본 논문의 색인어 선정을 위한 통계적 기준은 주로 단어의 출현 빈도에 근거하고 있다. 출현 빈도를 직접적으로 이용하는 기준은 단어의 빈도 산출방식에 따라 단순빈도와 상대빈도로 구분한다.

단순빈도는 단어가 어디에 출현했는가에 따라 단어빈도(term frequency), 문헌빈도(document frequency)로 구분한다. 단어빈도(T_f)는 색인대상이 되는 각 문헌 i에 특정 단어 k가 출현한 문헌의 수로,

$D_f = \sum_{i=1}^{n} b_n$이며 $f_{ik} \geq 1$일 때 $b_{ik} = 1$, $f_{ik} = 0$일 때 $b_{ik} = 0$이다.

단순빈도는 문헌집단의 크기나 분석 대상 텍스트의 길이, 또는 단어의 사용빈도를 전혀 고려하지 않은 것으로 실제로 이것만을 색인어 선정 기준으로 사용하기는 어렵다. 따라서 이러한 요인을 고려한 상대빈도가 보다 더 적합한 기준으로 평가되고 있다[14].

본 논문에서는 상대빈도를 구하기 위해서 Sparck Jones가 제시한 다음과 같은 공식을 사용한다[1][2]. 색인어 i가 문서 j에서 가지는 중요도는 다음과 같이 계산된다.

$$W_{ij} = TF_{ij} * (\log_2(N) - \log_2(DF_i) + 1)$$

W_{ij}는 문서 i에서 색인어 j의 중요도이고, TF_{ij}는 문서 i에서 색인어 j의 출현빈도수이고, N은 전체 문헌수이다. DF_i는 단어 i의 문헌빈도(document frequency)이다.

이와 같이 계산된 가중치에 의하여 추출된 색인어에 대하여 원 문서를 대상으로 공기 단어를 구축하게 되며 상호 정보량을 구하는 데 그 전체 과정은 [그림 2]과 같다.

4.2 색인어 자료구조

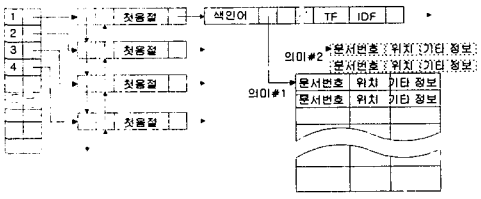
색인어 파일은 색인어 결정 기법으로 결정된 색인어에 대한 정보를 저장하는 파일을 말한다. 이 파일은 색인과 검색할 때 공유되며 빠른 속도와 효율적인 저장 공간을 이용해야 한다.

정보 파일의 구성 기법으로 주로 이용되는 방법은 각 색인어에 대한 참조를 저장한 역파일(inverted file)을 이용하는 방법과 각 문서에 해당하는 색인어 정보를 이진 열(bit stream)로 구성한 비트맵과 요약(signature) 파일을 이용하는 방법으로 크게 나눈다. 역파일을 이용하는 방법에는 동적인 구성, 검색 속도, 저장 공간의 크기 등이 중요한 요소이고 비트맵 및 요약 파일을 이용하는 방법에서는 검색 기법, 저장 공간의 크기 등이 중요한 점이다.

역파일의 구성은 비트맵이나 요약파일 방식과는 달리 색인어에 대한 별다른 가공을 하지 않고 정보를 저장하게 된다. 먼저 문자별 참조파일은 색인어의 첫음절의 순서에 의해 정렬된 색인 용어별 참조 파일의 첫 음절별 위치 정보를 가진다. 색인 용어별 참조 파일은 색인어의 출현 빈도, 가중치 및 문헌 식별자 참조 파일의 관련 위치 정보를 가진다. 최종적으로 문헌 식별자 참조 파일로부터 얻어진 문서 정보에 의하여 해당 색인어를 가지는 문서의 목록을 참조문서 배열을 통하여 얻을 수 있다.

이 방식은 저장 공간 갱신이나 재구성 비율 및 리스트가 많거나 길 때 그들을 합병하는 비용이 많이 든다는 단점이 있지만, 별도의 탐색 없이 바로 검색이 가능한 장점이외에도 주요어들에 대한 가중치나 문서에 대한 정보를 표현할 수 있고, 검색시 질의에 대한 유사도 계산이나 동음이의어 검색이 용이하다. 일반적으로 역파일 기법은 정적인 환경, 즉 질의가 빈번하고 갱신 작업이 거의 없는 환경에 적합하다[6]. 그러나 저장공간 갱신이나 재구성 비율 및 리스트가 길어질 때 시스템의 성능을 저하시키는 단점을 가지고 있다. 특히 색인어에 대한 부가적인 정보를 추가적으로 구성하기 위해서는 역파일 구조의 확장이 필수적이며 시스템의 성능에 영향을 미치지 않아야 한다.

본 논문에서는 이러한 문제를 해결하기 위하여 역파일 자료구조를 색인어의 첫 음절에 의한 동적인 트라이(dynamic trie)로 구성하였다. 뿐만 아니라 색인어는 문서내의 위치와 빈도 정보 외에도 추가적인 정보를 지속적으로 포함할 수 있게 되며 그 구성은 [그림 3]과 같다.



[그림 3] 의미정보를 포함하는 확장된 역파일 자료구조

동적인 트라이는 역파일의 첫 음절과 차일드 노드에 관한 정보만을 저장하고 있기 때문에 초기에 로드해야 하는 파일의 크기가 동적으로 구성하지 않은 역파일의 크기의 0.5%에 지나지 않았다. 또한, 4만여 단어 사전을 기준으로 추출된 2천여 개의 색인어를 644개의 파일로 분리함으로써 잘 사용되지 않는 불필요한 노드를 메모리에 적재하지 않게 되어 시스템의 자원을 효율적으로 사용할 수 있게 되었다. 이러한 동적인 트라이로 구성된 역파일은 시스템이 필요로 하는

노드만을 로드하게 되고 한 번 메모리에 로드된 노드는 같은 프로세스에서 동작되는 모든 인터넷 서버 응용 프로그램이 공유할 수 있으므로 기존 CGI와 비교하였을 때 시스템 자원의 효율적 이용 측면과 성능 향상 측면에서 유리함을 알 수 있었다.

이러한 장점으로 인하여 확장된 역파일 구조는 웹 정보검색 시스템의 성능을 극대화시킬 수 있을 뿐 아니라 서버 자원 공유라는 측면에서 정적인 기존 CGI 방법과 비교하여 시스템의 성능 향상이 용이하고 새로운 정보의 추가나 유지보수시에 시간과 노력이 현저히 감소하는 경제적인 측면도 뛰어나게 된다.

V. 상호 정보량에 의한 색인어 분류

상호 정보량 수식 (5-1)는 단어와 단어의 연관성을 정량적으로 나타내기 위해 사용되어 왔다[3][13]. 크기가 N인 말뭉치에서 단어 x가 사용된 횟수를 각각 $f(x)$, $f(y)$라고 하고 x와 y가 한 문장 내에서 함께 사용된 횟수를 $f(x, y)$라고 했을 때 N이 충분히 크다면 수식 (5-1)는 수식 (5-2)과 같이 근사할 수 있다[8][9][13].

$$MI(x,y) = \log_2 \frac{P(x,y)}{P(x)P(y)} \quad \cdots\cdots(5\text{-}1)$$

$$\simeq \log_2 \frac{f(x,y)}{f(x)f(y)} \quad \cdots\cdots(5\text{-}2)$$

수식 (5-2)는 다음과 같은 의미와 특성을 가진다[6].

첫째, 단어 x와 y가 많은 관계가 있다면, 확률 $p(x, y)$는 $p(x)*p(y)$보다 클 것이며, 결과적으로 $MI(x, y)$값은 0보다 큰 값이 될 것이다.

둘째, 단어 x와 y가 그다지 큰 관계가 없다면, 확률 $p(x, y)$와 $p(x)*p(y)$는 거의 같은 값이 될 것이며, 결과적으로 $MI(x, y)$는 0에 가까운 값을 얻을 것이다.

셋째, 단어 x와 y가 전혀 관계가 없다면, 확률 $p(x, y)$는 $p(x)*p(y)$의 값보다 적은 값이 될 것이며, 결국 $MI(x, y)$의 값은 0이 될 것이다.

넷째, 일반적으로 상호 정보량은 $MI(x, y) = MI(y, x)$가 만족되는 대칭성(symmetric)을 갖는다. 하지만 본 논문에서 적용하는 상호 정보량은 색인어의 공기 단어 추출시에는 대칭성을 허용하지 않는다. 즉, 색인어 x가 가중치 계산에 의하여 선정된 후 일정한 범위의 윈도 내에서 공기단어 y의 출현빈도를 계산하는 값과 색인어 y가 가중치 계산에 의하여 선정된 후 일정한 범위의 윈도 내에서 공기단어 x의 출현빈도를 계산할 때 서로 다른 별개의 값이 나오기 때문이다.

다섯째, 상호 정보량 수식은 단어와 단어 사이의 의존 관계를 확률적으로 나타낸 것이므로 복합어의 정보량을 정량적으로 나타낼 수 있다.

여섯째, 상호 정보량 수식은 발생 빈도가 적은 단어는 단어간 의존 관계의 객관성을 확인할 충분한 근거가 되지 못하므로 비교적 크기가 큰 문서에 적용해야 한다.

상호 정보량을 이용하여 공기 정보를 구축하는 예는 다음과 같다. 아래의 두 예문은 실험 데이터로 사용된 인하대학교 교내 웹 페이지중에서 추출된 문서 중의 일부이다.

예문 1)

> 검색은 페이지 문서 전체를 대상으로 어구나 문장 형태까지도 검색 할수 있다. 대구대 컴퓨터 동아리에서 만든 검색 **엔진**으로 우리나라 한글 검색 **엔진**의 시조격인 검색 사이트이다.

예문 2)

> 디이젤 **엔진**의 실린더 라이너 및 피스톤 링 그로브의 레이저 표면 경화로 고품질 부품 생산을 위한 연구를 수행하고, 자동차용 크랭크 샤프트의 필렛부의 레이저 표면 경화 처리로 피로 수명을 연장하는 연구를 수행한다.

예문 1에서는 '엔진'의 공기 단어로 검색, 문서, 문장, 컴퓨터, 한글 등의 단어가 추출될 것이고, 예문 2에서는 디이젤, 실린더, 피스톤, 레이저, 자동차, 크랭크, 수명 등의 단어가 추출될 것이다.

실험 대상 문서에 대해서 위와 같은 방법으로 추출된 색인어에 대해서 일정한 범위의 윈도(window, 100어절)내에서 공기 단어를 추출하고 출현 빈도를 계산한다. 공기 단어 목록이 작성되면, 색인어와 공기단어 간의 상호정보량을 계산한 후에 색인어 클러스터링(clustering)을 수행한다.

[표 1] 공기 단어 빈도

단 어	빈 도	단 어	빈 도	단 어	빈 도
검색	106	제어	64	진동	37
정보	98	시스템	46	공학	37
연구	96	자동차	41	페이지	36
개발	71	유체	39	…	

'엔진'이라는 색인어에 대하여 구해진 공기 단어중 일부의 빈도는 [표 1]에 제시하였으며 색인어와의 상호정보량은 아래와 같다. 아래 단어의 순서는 빈도순이며, 빈도와 상호정보량이 반드시 일치하지 않는다는 사실을 보인다.

공기 단어간의 상호정보량은 [표 2]과 같다.

[표 2] 공기 단어간의 상호정보량

	제어	정보	유체	진동	자동차
검색	-1.82	0.35	-4.40	-2.87	-2.27
제어		-1.66	0.53	0.27	0.99
정보			-1.51	-3.59	-2.46
유체				0.92	0.43
진동					0.82

상호정보량 계산에 의해 구해진 공기 단어간의 상호정보량은 색인어를 분류하는 기준으로 사용된다.

먼저, '검색'과 '제어'의 상호정보량을 비교하면 -1.82라는 비교적 큰 차이를 보이므로 '검색'과 제어'를 분리한다.

다음 '정보'와 '검색', '정보'와 '제어'의 상호정보량을 비교하였을 때 '정보'라는 단어는 제어와는 많은 차이를 보이지만 '검색'과는 비교적 정보량이 있으므로 '정보'는 '검색'과 같은 부류로 분리된다.

이와 같은 방법에 의해 아래와 같이 '엔진'이라는 색인어는 두가지 의미로 분리됨을 알 수 있다.

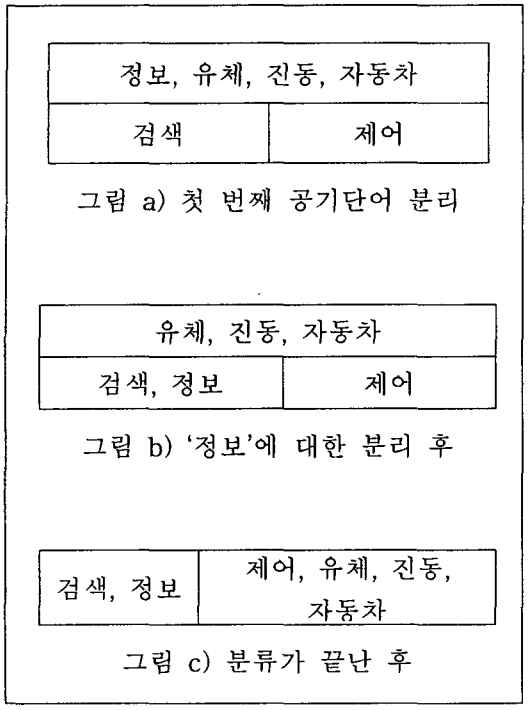


그림 a) 첫 번째 공기단어 분리

그림 b) '정보'에 대한 분리 후

그림 c) 분류가 끝난 후

[그림4] 공기단어 분리

Ⅵ. 실험 및 평가

실험은 인하대학교내의 웹상에 존재하는 4485개의 문서를 대상으로 하였다. 실험 대상 문서로부터 2154개의 색인어 후보가 추출되었으며, 두 글자 어휘는 1390개로 전체 색인어 후보중 64%를 구성하고 있었다. 한글자 색인어 후보와 두 글자를 초과하는 색인어 후보는 그 비율이 각각 6.5%와 29.5%를 차지하고 있었다. 본 실험에는 한 글자 색인어는 한자어와 접두어 및 접미사가 대부분이고 그 의미 중요도가 상대적으로 낮으므로 실험 및 평가에 사용하지 않았다.

두 글자 어휘중 동음 이의어는 166개로 전체 1390개의 어휘중 12%를 차지하고 있었으며, 세 글자 이상 어휘중 동음 이의어는 2%이었다.

정보검색 시스템의 평가는 일반적으로 검색효율, 신속성, 경제성의 세가지 측면에서 수행된다. 검색 효율은 이용자가 요구하는 수준의 정보 서비스를 제공하는 시스템의 능력을 측정하는 것으로 서비스의 질(quality)의 척도이며, 신속성과 경제성은 각각 일련의 정보검색 작업에 소요되는 시간과 경비를 측정하는 것이다[14].

검색 효율은 이용자의 정보 요구 만족도를 측정하는 평가 기준인 반면, 신속성과 경제성은 주로 시스템 운영자의 입장에서 관심을 갖는 평가 기준이라고 볼 수 있으므로 본 논문에서 제안하는 시스템의 평가는 신속성과 경제성 측면을 제외한 검색 효율의 관점에서 수행한다.

본 실험의 평가를 위해서 수식 (6-3) 및 수식 (6-4)와 같은 정확도와 재현률 공식을 사용한다[14]. 재현률과 정확도는 검색 효율 척도 가운데 가장 널리 사용되고 있다. 재현률은 시스템이 소장하고 있는 적합 문헌들 가운데 검색된 적합 문헌의 비율을 말하며, 정확도는 검색된 문헌들 가운데 적합 문헌의 비율을 말한다. 재현률은 시스템이 적합 문헌을 검색해 내는 능력과 관련된다. 즉, 검색 시스템이 대상 문서중에서 적합한 문헌을 검색해 내는 능력을 표현하는 것이다. 정확도는 검색된 문헌들이 얼마나 적합한가를, 시스템이 부적합 문헌을 얼마나 많이 걸러내는지를 나타내는 수치이다.

재현률과 정확도 이외에 누락률과 잡음률 등의 평가 수치가 있으나 재현률과 정확도에 의해 표현 가능한 수치이므로 본 실험에서는 이들 수치는 제시하지 않았다. 본 평가에 사용되는 재현률과 정확도를 산출하는 방법은 아래 수식 (6-3) 및 수식 (6-4)와 같다.

$$재현률 = \frac{검색된\ 적합\ 문헌수}{적합문헌\ 총수} \quad \cdots(6\text{-}3)$$

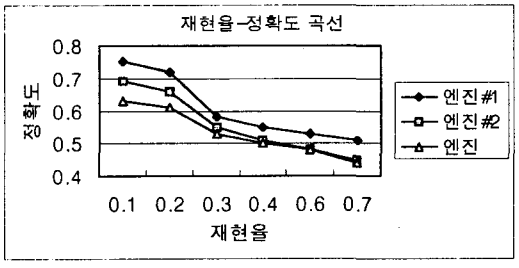
$$정확도 = \frac{검색된\ 적합\ 문헌수}{검색된\ 문서\ 총수} \quad \cdots(6\text{-}4)$$

실험 데이터로부터 추출된 색인어 후보에 대하여 분류를 시도한 결과 96.7%의 성공률을 보였으며, 분류에 실패한 어휘의 예는 [표 3]와 같다.

[표 3] 분류에 실패한 단어의 예

Type 1	기사	자격증, 합격, 기술
		신문, 논문, 방송국, 사회, 경제
		전기, 전자
		위생, 식품
Type 2	저항	반항, 정권, 정치, 정부
		마찰, 출력, 입력, 신호, 측정
		전압, 전류, 회로, 콘덴서

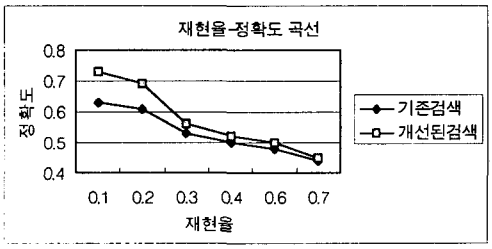
Type 1은 관련 공기 단어들간의 상호정보량이 작아서 과다하게 다른 의미로 분류한 경우이며, type 2는 공기 단어의 출현 빈도는 상당히 높으나 상대적으로 공기하는 빈도의 편차가 작아서 상호정보량에 의한 의미 분류에 실패한 경우이다.



[그림 5] 의미 고려후의 정확도-재현률 곡선

앞에서 예를보인 '엔진'의 경우 의미를 고려하였을 때와 의미를 고려하지 않고 색인어를 사용하였을 때의 재현률과 정확도는 [그림 5]과 같다.

[그림 5]에서 '엔진#1'은 기계에 관련된 자동차 분류의 '엔진' 계열의 의미로 검색을 시도한 경우의 재현률과 정확도 곡선이며, '엔진#2'는 컴퓨터에 관련된 '검색 엔진' 계열의 의미로 검색을 시도한 경우이다. 세 번째의 '엔진'은 의미를 고려하지 않고 검색을 시도한 후에 앞에서 시도한 두 가지 의미로 재현률과 정확도를 각각 측정한 후 평균을 취한 곡선이다.



[그림 6] 색인어 분류후의 재현률-정확도 곡선

[그림 6]은 색인어 의미를 고려하지 않고 재현률과 정확도를 측정한 값과, 본 논문에서 제안하는 방법에 의해 색인어 분류를 시도한 후 재현률과 정확도를 측정하여 도식한 것이다.

실험에서 재현률을 낮출수록 정확도가 더 급격히 증가하는 현상이 발생하였다. 이는 재현률이 낮을 때 선정되는 색인어의 빈도가 많아지고 더 많은 문서에 대해서 공기 단어를 구축하기 때문에 공기 단어의 색인어에 대한 상호정보량과 공기 단어간의 상호정보량의 값이 더 큰 편차를 가지기 때문으로 분석할 수 있다.

지금까지 실험된 결과를 앞에서 제시한 방법

에 따라 평가를 하면 평균적으로 4.5%의 정확도 향상을 보임을 알 수 있다.

Ⅶ. 결론

기존의 연구들이 의미 중의성을 해결하기 위하여 말뭉치에 태그를 주고 학습한후 의미를 추정하거나 일부 의미망과 신경망, 의미 벡터 값을 이용하는 등의 많은 시도가 있었다.

본 논문에서는 정보검색이라는 제한된 범위내에서는 모든 어휘에 대한 의미 정보나 공기 정보에 대한 구축이 필요하지 않다는 점에 착안하여 상대 출현 빈도에 의해 추출된 색인어에 대해서만 공기 단어를 수집하고, 상호 정보량 계산에 의해 별도의 의미 태그가 없이도 색인어를 의미에 따라 분류할 수 있었다.

색인어의 의미 분류는 정보검색 시스템에서 색인어를 추출하는 기법만큼 정확도 향상에 중요한 요소임을 실험을 통해 확인할 수 있었다. 현재 상용 서비스 되고 있는 웹 정보검색 시스템들은 보통 수 천 또는 수 만 건의 검색 결과를 제공하는 데, 이러한 정책은 사용자가 실제로 방문할 수 있는 웹 문서 수와 비교하면 매우 비현실적이며 이제는 재현률보다 정확도를 높이는 방향으로 검색 시스템들의 연구 방향이 변화되어야 한다.

웹 정보검색 시스템의 경우 미등록어가 많이 발생하고 그 내용이 수시로 변하는 특징 때문에 기존의 의미 관련 연구를 적용하기에는 어려운 점이 많지만, 본 논문에서 제안하는 기법은 별도의 의미 사전이나 태그가 없이도 정확도를 4.5% 정도 향상시킬 수 있었으므로 향후 공기 단어 추출에 HTML 태그의 특성을 반영하고 미등록어에 대해서도 클러스터링을 할 수 있는 연구가 뒷받침된다면 상용 웹 정보검색 시스템의 정확도가 향상되어 고품질의 검색 결과를 사용자에게 제공할 수 있을 것이다.

참고 문헌

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

[2] G. Salton, *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*, Addson-Wesley Publising Company, 1989.

[3] R.J. McEliece, *The Theory of Information and Coding*, Addson-Wesley Publising Company, 1977.

[4] William B.Frakes, Ricardo Baeza-Yates, *Information Retrieval*, Prentice-Hall, 1992.

[5] 강현규, 박세영, 최기선, "자동 키워드망과 2단계 문서 순위 결정에 의한 자연어 정보검색 모델," 제7회 한글 및 한국어 정보처리 학술대회, pp.8~12, 1995.

[6] 김판구, *한국어 정보검색을 위한 상호정보량에 기반한 복합어 자동색인*, 서울대학교 컴퓨터공학과 박사학위 논문, 1995.

[7] 김철완, 장재우, "형태소 네트웍을 이용한 한글 문헌의 자동 키워드 추출," 제6회 한글 및 한국어정보처리 학술발표 논문집, pp.363~368, 1994.

[8] 박세영, "멀티미디어 정보검색에서의 한국어 정보처리," 정보과학회지 제 12권 제 8호, pp.60~66, 1994.

[9] 심광섭, "음절간 상호 정보를 이용한 한국어 자동 띄어쓰기," 정보과학회 논문지 제 23권 제 9호, pp.991~1000, 1996.

[10] 안성현, 장재우, "문법형태소 네트워크를 이용한 자동색인 시스템의 설계," 제7회 한글 및 한국어 정보처리 학술대회, pp.13~17, 1995.

[11] 이현아, 홍남희, 이종혁, 이근배, "한국어 형태소 구조규칙에 기반한 색인 시스템의 구현," 정보과학회 봄 학술발표 논문집 Vol.22, No.1, pp.933~936, 1995.

[12] 장호욱, 박세영, "정보검색 시스템의 정확도 향상을 위한 키팩트 개념의 도입," 정보과학회 가을 학술발표 논문집 Vol.22 No.2, pp.651~624, 1995.

[13] 전미선, 박세영, "상호 정보를 이용한 어의 모호성 해소에 관한 연구," 제6회 한글 및 한국어정보처리 학술발표 논문집, pp.369~373, 1994.

[14] 정영미, 정보검색론, 구미무역(주) 출판부, 1992.

[15] 최기선, "한국어 정보검색," 정보과학회지 제12권 제8호, pp.23~32, 1994

[16] 한성현, 박혁로, 최기선, 김길창, "구문해석을 이용한 색인어 자동 추출 시스템," 제2회 한글 및 한국어정보처리 학술발표 논문집, pp.16~23, 1990.

Design of WWW IR System Based on Keyword Clustering Architecture

Jeom-Dong Song*, Jung-Hyun Lee**, Jun-Hyeog Choi***

Abstract

In general information retrieval systems, improper keywords are often extracted and different search results are offered comparing to user's aim bacause the systems use only term frequency informations for selecting keywords and don't consider their meanings. It represents that improving precision is limited without considering semantics of keywords because recall ratio and precision have inverse proportion relation. In this paper, a system which is able to improve precision without decreasing recall ratio is designed and implemented, as client user module is introduced which can send feedbacks to server with user's intention. For this purpose, keywords are selected using relative term frequency and inverse document frequency and co-occurrence words are extracted from original documents. Then, the keywords are clustered by their semantics using calculated mutual informations.

In this paper, the system can reject inappropriate documents using segmented semantic informations according to feedbacks from client user module. Consequently precision of the system is improved without decreasing recall ratio.

* Dept. of Computing & Information Processing, Pyongtaek Institute of Technology
** Dept. of Computer Engineering, Inha University
*** Dept. of Computer Engineering, Kimpo College