자연언어 질의 문장의 용어 가중치 부여 기법

강승식[0], †이하규, ‡손소현, ‡문병주, ‡홍기채
국민대학교 컴퓨터학부, 첨단정보기술연구센터,
†성공회대학교 컴퓨터정보공학부, ‡한국전자통신연구원 지식정보센터
sskang@kookmin.ac.kr, hglee@mail.skhu.ac.kr, {shson, bjmoon, gchong}@etri.re.kr

Term Weighting Method for Natural Language Query Sentence

Seung-Shik Kang[0], Hagyu Lee, So-Hyun Son, Gi-Choi Hong, Byung-Joo Moon
School of Computer Science, Kookmin University & AITrc
Division of Computer and Information Science, Sungkonghoe University
Knowledge Information Center, ETRI

요 약

자연언어 질의 문장으로부터 검색어로 사용될 질의어의 추출 및 질의어 가중치를 계산하기 위하여 질의 문장들의 유형을 분석하였으며, 질의어 구문의 특성에 따라 용어들의 가중치를 계산하는 방법을 제안하였다. 용어의 가중치를 부여할 때 띄어쓴 복합명사와 접속 관계 등에 의해 연결된 명사구는 질의어 가중치를 동등하게 적용할 필요가 있다. 질의 문장에서 가중치가 동등하게 적용되는 명사구를 인식하기 위한 목적으로 구현된 명사구 chunking을 수행한 후에 각 용어들에 대한 질의어 가중치를 계산한다. 질의어 가중치를 계산하기 위하여 용어의 유형, 질의 구문의 특성, 문서 유형을 지칭하는 용어, 조사 유형, 용어의 길이 등에 따라 가중치를 조절하는 방법을 사용한다. 용어 유형에 의한 가중치 계산은 추출된 용어의 품사 정보와 전문 용어 사전, 부사성 명사 사전을 이용하였다.

1. 서론

정보 자료의 양이 증가함에 따라 수많은 문서들 중에서 사용자가 필요로 하는 문서만을 검색하는 요구가 증가하고 있다. 사용자에게 적합한 정보 자료만 선별하여 제공하는 것은 검색 결과를 사용자의 요구(검색 의도)를 반영하여 순서화(ranking)하거나 적합 문서(relevant document)를 판별하는 문제이다. 검색 결과의 순서화와 적합 문서를 판별하는 기법은 주로 색인어의 출현 빈도와 문헌 빈도에 의해 색인어의 가중치를 계산하는 통계적인 기법을 이용한다[1,2]. 그런데 통계적인 기법을 이용한 가중치의 계산은 문장 혹은 문서 내에서 색인어의 중요도를 반영하지 못하는 제약이 있다.

질의 문장이나 정보 자료 내에서 질의어(query term) 혹은 색인어(index term)의 중요도를 반영하려면 문서 분석 기법에 의한 색인어의 가중치를 계산해야 한다[3]. 키워드 질의어에서는 사용자가 검색하고자 하는 키워드를 입력하므로 모든 질의어간의 상대적 가중치를 계산하기가 어렵다. 그러나 문장 형태의 자연언어 인터페이스의 질의문에 나타난 질의어들은 검색하고자 하는 키워드와 검색 요구를 위한 상투적인 용어들이 혼합되어 있다.

특히, 키워드 질의어와 다르게 자연언어 질의 문장에서는 사용자가 검색하고자 하는 용어를 신중하게 선택하는 것이 아니라 검색하고 싶은 내용을 자연스러운 문장 형태로 구성하기 때문에 질의 문장에서 추출된 용어들이 정보 자료의 검색에 미치는 비중이 다른 경우가 많다.

키워드 질의어 방식에서는 통계적으로 질의어 가중치를 계산한다. 그러나 자연언어 인터페이스에서 질의 문장에서 추출된 질의어에 대해 통계적인 가중치 계산 방식을 적용하는 것은 적합하지 않다. 왜냐 하면, "~에 대한 문서를 검색해 주세요"에서 '문서', '검색' 등과 같이 통상적으로 검색 요구를 위한 용어들이 질의 문장에 포함되기 때문이다.

문장 단위의 구문 분석 기법은 기계 번역이나 기계 이해 시스템, 정보 추출, 문서 요약 등 다양한 응용 분야에서 시스템의 성능을 향상시키거나 정확도를 개선하는데 중요한 역할을 하고 있다. 이에 비해, 정보 검색에서는 단순히 색인어를 추출하는 목적의 스테밍 혹은 형태소 분석에 의한 색인어 추출, 복합어 인식 등 형태론적 수준의 분석 결과만을 활용하고 있다. 그러나 질의 문장을 사용하는 자연언어 인터페이스에서는 질의 문장으로부터 질의어들을 추출하고 질의어 가중치를 부여하는 문장 단위의 분석 기법이 필요하다.

일반적으로 질의어 문장이나 문서의 내용을 분석하려면 형

태소 분석과 구문 분석, 의미 분석 등이 사용되지만 현재 구문 분석과 의미 분석 기술은 실용적인 시스템에 적용하는데 많은 제약이 있다[2]. 한국어 형태소 분석은 미등록어와 복합명사, 숫자와 수사 문제 등이 완전히 해결될 수 없기 때문에 복합명사 처리 등의 기능이 필요한 응용 소프트웨어에서는 이러한 문제를 해결하는 기능이 추가되어야 한다. 또한, 구문 분석 기술은 일반적으로는 실용적인 응용 시스템에 적용하기가 쉽지 않으나 구문 분석 결과에 오류가 포함되어 있더라도 한국어 정보처리 기능에 영향이 심각하지 않은 분야에서 활용되기도 한다.

문장 단위의 용어 가중치를 부여하는 문제를 해결하는 방법으로 구문 분석 결과를 활용할 수 있다. 그러나 한국어에서는 형태소 분석 결과로부터 특정 구문 유형들을 판별할 수 있으며, 이 때 구문 분석 오류를 고려하지 않아도 되는 장점이 있다.

본 논문에서는 형태소 분석 수준에서 질의 문장을 분석하여 핵심어와 비핵심어를 구별하고 각 질의어에 대한 가중치를 부여하는 질의어 가중치 부여 기법을 제안한다. 정보 자료의 분석에 의한 질의어 가중치는 기존의 출현 빈도와 문헌 빈도에 의한 통계적 기법에 의한 가중치의 계산의 문제점을 개선하여 적합 문서를 판별하거나 검색된 정보 자료의 순서화에 사용될 수 있다.

2. 질의 문장 및 질의어 유형

2.1 질의 문장의 특성

정보 검색 시스템에서 질의어 문장은 키워드 검색 유형과 질의-응답 유형에 따라 질의문의 유형이 매우 다르다. 키워드 검색 유형은 통상적으로 정보 자료를 검색하는데 사용되는 어휘가 빈번하게 사용되는데 비해, 질문에 대한 정답을 찾는 질의-응답과 관련된 질의어 문장은 육하원칙과 관련된 어휘들이 빈번하게 사용된다. 여기서는 질의-응답 검색을 배제하고 키워드 검색과 관련된 질의어 문장의 유형들을 분석한다.

키워드 검색 대신에 질의어 문장을 검색할 때 사용되는 통상적인 구문은 술어로 끝나는 경우와 명사구로 끝나는 경우로 구분된다. 술어로 끝나는 검색 요구 유형은 그림 1과 같다.

~에 대해 알려주세요
~에는 어떤 것이 있나요
~에 대해 알고 싶어요/싶습니다/싶다
~에 관해 알려주세요
~를 찾아주세요
~를 검색해 주세요
~를 보여주세요

그림 1. 검색 요구에 관한 구문 유형

명사구로 끝나는 질의문에서는 주로 검색하고자 하는 키워드에 관한 검색 요구 표현이 포함되어 있다. KTSET에 포함되어 있는 자연언어 질의 문장에서는 그림 2와 같이 검색 요구와 관련된 통상적인 구문들이 빈번하게 사용되고 있다[4].

A에 관한 연구
A를 다룬 논문
A의 B에 대한 연구 결과
A중 B에 대한 연구
A에서 B에 대한 연구
A중 B를 응용한 연구
A중에서 B를 다룬 것
A를 갖고 있는 B
A에 대한 B
A에 대한 B에 대한/관한 연구
A중 B를 위한 C
A를 이용한 B에 관한 연구

그림 2. 자연언어 질의 문장 유형의 예

위 질의 문장 유형에서 A, B, C는 사용자가 검색하고자 하는 질의어에 관한 용어로서 복합어와 접속 관계가 허용된다. "정보통신과 네트워크에 대한 운영체제"의 경우 '정보통신'과 '네트워크'라는 두 개의 용어가 접속되었고, "프로그래밍 언어에 대한 컴파일러의 작성에 대한 연구"는 '~에 대한'이 중첩된 예이다.

2.2 질의어 유형

키워드 질의어는 질의어를 단순히 나열하기 때문에 각 용어들의 가중치를 다르게 부여하기가 어렵다. 즉, 사용자가 원하는 문서를 검색하기 위해 보조로 추가한 질의어와 핵심 용어의 구분이 없다. 따라서 기존 검색 모델에서는 단순히 출현 빈도에 의해 질의어 가중치를 계산한다.

질의 문장에서 추출되는 질의어는 질의 문장을 구성하기 위한 용어, 핵심어, 보조 용어 등으로 구분된다. 질의 문장에서 핵심어와 핵심어를 보조하는 용어를 지칭하는데 사용되는 구문은 다음과 같다.

(1) 핵심어를 지칭하는 구문
~에 대한, ~에 관한, ~에 관련된, ~를 다룬,
~에 대해 다루어진/다룬

(2) 핵심어를 보완(수식)하는 구문
~를 위한, ~를 이용한, ~를 응용한, ~에 사용되는,
~에 이용되는, ~를 대상으로 한/하는,
A 분야에서, A 분야의, A 분야중에서,
A 분야 중에, A에서

핵심어를 지칭하는 구문은 핵심어가 검색된 문서에서 주요어인 문서들을 검색하고자 하는 사용자의 의도가 내포되어 있다. 이에 비해, 핵심어를 보완하는 보조 용어는 핵심어가 검색된 문서들 중에서 보조 용어와 관련된 문서로 제약하게 된다. 따라서 핵심어가 주요어인 문서들 중에서 보조 용어에 관한 문서들의 우선 순위를 높여 주는 형태의 검색이 요구된다.

예를 들어, 질의 문장 "***컴퓨터 이론***중 ***그래프***를 대상으로

하는 ***알고리즘***에 관한 ***연구***"로부터 '컴퓨터 이론', '그래프', '알고리즘',.'연구'라는 4개의 용어가 추출된다. 이 질의 문장에서 검색하고자 하는 정보 자료는 '알고리즘'이고, 알고리즘 중에서도 '그래프 알고리즘'에 관한 문서이다. 그런데 그 중에서도 '컴퓨터 이론' 분야에 속하는 문서로 제한하고 있다. 따라서 '그래프'와 '알고리즘'은 핵심어이고 '컴퓨터'와 '이론'은 보조 용어가 된다. 이 문장에서 '연구'는 단지 문서의 유형을 지정하는 역할을 하고 있으므로 '연구'는 검색어로 사용되지 않는다.

A중에서 B에 대한 <문서유형>
A의 B에 대한 <문서유형>
A에서(의) B에 대한 <문서유형>
<문서유형 1> '연구/논문/문서/기사/보고서/책/것'
<문서유형 2> '방법/기법/이론/실험'

그림 3. 문서 유형을 지정하는 질의 구문

문서 유형을 지정하는 용어들의 예는 그림 3과 같으며, 문서 유형을 지정하는 용어들은 두 가지로 구분된다. 문서 유형을 지정하는 용어들은 키워드 검색에서는 사용되지 않는 용어들이므로 불용어로 처리하거나 가중치를 매우 낮게 부여해야 한다. 이러한 유형에 해당되는 용어들은 그림 4와 같다.

~에 대해/대한 → '대해', '대한'
~에 사용되는, ~에 이용되는 → '사용', '이용'
~를 응용한 → '응용'
~를 대상으로 하는/한 → '대상', '한', '하'
A 분야에서 → '분야'
검색해 주세요 → '검색', '주세'
연구 결과(물) → '연구', '결과', '결', '결과물'
방법/기법(중) → '방법', '기법', '방법중', '기법중'
방식(중), 분야(중) → '방식', '방식중', '분야', '분야중'
중에(서) → '중'
찾아 주라 → '주'
혹은 → '혹'

그림 4. 질의 문장 구성에 관한 불용어

3. 문장 단위 색인어 추출

입력된 질의 문장에 대한 형태소 분석 결과로부터 추출된 용어 정보는 다음과 같다.

- 용어 문자열
- 용어의 출현 위치
- 용어 유형

용어의 출현 위치는 문장내 어절의 위치로서 용어가 추출된 어절의 형태소 분석 결과를 참조하는데 사용된다. 따라서 용어의 출현 위치와 문장 단위 형태소 분석 결과를 일치시키기 위해 사용된다. 용어의 유형은 질의어를 유형에 따라 복합명사, 미등록어, 불용어, 영문자, 숫자 등을 구분하는 표지이다.

```
n = get_terms(sent, terms, kmaresult);
```

함수 get_terms()는 질의 문장 sent에 대한 형태소 분석 결과로 kmaresult를 생성하고 이 결과로부터 용어 terms들을 추출한다. 추출된 용어의 개수는 n에 저장된다. terms에는 추출된 질의어와 질의어의 출현 위치(어절 번호)가 저장되어 질의어의 출현 위치에 의해 형태소 분석 결과인 kmaresult를 참조하게 된다.

질의어 문장에서 용어를 추출하고 용어 가중치를 부여하는 방법은 다음과 같이 두 가지 방법이 가능하다.

① 형태소 분석 결과로부터 직접 용어들을 추출하고 가중치를 부여
② 용어 추출 후에 용어 문자열과 형태소 분석 결과를 이용한 가중치 계산

형태소 분석 결과로부터 직접 용어를 추출하는 방법은 형태소 분석 결과 및 어절 유형, 문맥 정보 등을 직접 이용할 수 있는 장점이 있다. 그러나 붙여쓴 복합명사를 분해하거나 불용어 제거, 특수 색인어 추출, 용언과 독립언 추출 옵션의 적용 등 색인어 추출과 관련하여 추가된 기능들을 별도로 적용해야 하는 단점이 있다. 이러한 단점으로 인하여 본 논문에서는 두 번째 방법으로 색인어 추출에 사용되는 모든 기능을 활용하면서 추출된 용어에 대해 가중치를 계산하는 방식을 취한다.

용어를 추출한 결과로부터 용어 문자열과 용어가 출현한 어절 위치, 그리고 형태소 분석에 의해 구분된 용어의 유형을 질의어 문장 분석에 의해 가중치를 부여하여 핵심어와 비핵심어를 구분할 수 있다. 추출된 용어들과 용어가 출현한 어절 위치를 이용하여 형태소 분석 결과를 참조하려면 형태소 분석 결과의 구조체 항목들을 인지해야 한다.

4. 질의어 가중치 부여 기법

4.1 가중치 부여를 위한 자료 구조

질의어 문장에 출현한 용어를 핵심어와 비핵심어로 구분하거나 용어 가중치를 계산하기 위해서는 입력 문장을 분석하여 그 결과로부터 용어들을 추출한다. 추출된 용어들에 대한 가중치 계산에 필요한 자료 구조는 그림 5와 같다.

```
/* 용어 추출 결과의 저장 구조 */
typedef struct termList {
        char *term;         /* term string */
        int type;           /* term type */
        int tf;             /* term frequency */
        int loc;            /* term 출현 위치 */
        int weight;         /* term weight */
} HAM_TERMLIST, *HAM_PTERMLIST;
```

그림 5. 용어 가중치 부여를 위한 자료 구조

이 자료 구조는 질의어 문장에서 문장 단위의 용어 추출 결과 및 정보 자료의 문서 전체에 대한 용어 추출 결과를 저장하기 위한 목적으로 정의한 것이다. 따라서 문장 단위 용어 추출을 중심으로 정의하였으나 문서 단위 용어 추출에도 사용될 수 있도록 하였다.

질의어 문장으로부터 용어들을 추출하고 용어들에 대한 가중치를 부여하는 순서는 다음과 같다.

1) 질의어 문장에 대한 형태소 분석 및 용어 추출
2) 용어 가중치 부여에 필요한 정보를 저장
3) 추출된 용어와 형태소 분석 결과에 의해 가중치 부여

4.2 명사구 인식

용어 유형에 의한 가중치 부여 기법은 문장 특성을 반영하고 있지 않다. 또한, 용어 유형에 의한 가중치 부여와 문장 패턴 정보에 의한 가중치를 결합하기가 어렵다. 질의어 문장은 일반적인 문장과 구별되는 특성이 있으며, 질의어 문장의 특성 패턴을 이용하여 용어의 가중치를 적용하려면 띄어쓴 복합명사와 명사구 단위로 가중치를 적용해야 한다. 따라서 용어 가중치의 기본 단위가 되는 명사구 인식이 선행되어야 한다.

가중치 부여 단위가 되는 명사구로 간주되는 명사구는 수식어를 제외한 명사들로 제한한다. 또한, 가중치를 동일하게 부여해야 하는 명사의 나열형도 하나의 chunk로 구성한다. 명사구 chunking 대상이 되는 명사구의 유형은 그림 6과 같다.

N N ... N
N의 N
N과(와) N
N 혹은/또는/및 N
N, N
N이나 N

그림 6. 띄어쓴 복합명사 및 나열형 용어

용어 가중치를 부여하는 기본 단위인 띄어쓴 복합명사와 명사구를 인식하기 위하여 명사구 chunking을 한다. 명사구 chunking을 위해 입력 문장의 각 어절들에 대해 chunking tag을 부착한다. 명사구가 시작되는 어절의 태그는 NPCHUNK_BEG를 부여하고, 끝나는 어절은 NPCHUNK_END를 부여한다. 청크의 중간에 위치하는 어절은 NPCHUNK_IN으로 태깅을 하고, '그리고', '또는' 등 나열형이나 '-와/과'에 의한 접속형에 의한 어절들에 대해 동일한 가중치를 적용하기 위하여 나열형과 접속형에 사용되는 어절에는 NPCHUNK_CNJ를 부여한다. 명사구 chunking에 사용되는 tag는 그림 7과 같다.

```
#define NPCHUNK_OUT     0       /* out */
#define NPCHUNK_BEG     1       /* begin */
#define NPCHUNK_IN      2       /* inside */
#define NPCHUNK_CNJ     3       /* 연결어미 */
#define NPCHUNK_END     4       /* end */
#define NPCHUNK_NP1     5       /* 단일명사 */
```

그림 7. 명사구 인식을 위한 명사 유형 정의

4.3 가중치 부여 기준

그림 8은 문장에서 추출된 용어들에 대해 용어 유형에 따라 부여되는 기본적인 가중치이다. 이 가중치는 용어 유형에 따라 상대적으로 차별화하기 위한 것으로서 경험적인 방법으로 가중치를 설정하였다.

TERM_SCORE_TNN은 전문 용어 사전(domain dictionary)에 수록된 용어의 가중치이다. 동일한 용어라고 하더라도 용어의 길이에 따라 가중치에 차등을 두기 위하여 음절 길이에 따라 가중치를 다르게 부여한다. 복합명사 가중치는 붙여쓴 것과 띄어쓴 것의 가중치를 다르게 하고, 붙여쓴 복합명사에 비해 띄어쓴 복합명사의 구성 명사는 가중치를 낮게 부여한다.

```
#define TERM_SCORE_TNN      100.  /* domain dic. */
#define TERM_SCORE_CNN1      90.  /* 붙여쓴 복합명사 */
#define TERM_SCORE_CNN2      80.  /* 띄어쓴 복합명사 */
#define TERM_SCORE_UNREG     70.  /* 미등록 명사 */
#define TERM_SCORE_NOUN      50.  /* 사전 등록 명사 */

#define TERM_SCORE_MNN       10.  /* 수식어 사전 수록 명사 */
#define TERM_SCORE_NSFX      30.  /* 명사 접미사 결합된 명사 */
#define TERM_SCORE_VSFX      20.  /* 용언 접미사 결합된 명사 */
#define TERM_SCORE_NONE      10.  /* 기타 */
```

그림 8. 용어 유형별 가중치 기본 점수

그림 9는 용어가 추출된 어절의 조사 유형에 따라 가중치를 조절하기 위한 것이다. 이 가중치는 기본 가중치에 곱해 주는 값으로써 1.0을 중심으로 가중치를 높여 주거나 낮춰 주는 목적으로 사용된다. 그림 9의 조사 가중치는 질의어 문장에서 추출된 용어의 가중치 계산을 중심으로 정의되었다. 따라서 조사 '에'의 가중치가 2.0으로서 가장 높다. 그 이유는 "A에 대한/관한" 유형은 질의어 문장에서 용어 A가 사용자가 찾고자 하는 핵심어가 되기 때문이다. 즉, 주제어 추출을 위한 아래 조사 가중치와 질의어 분석을 위한 조사 가중치는 차이가 있다.

- '은/는' : 1.0
- '이/가' : 0.8
- '을/를' : 0.8
- '만/도' : 0.5
- '의' : 0.8
- '에/에서' : 0.4
- 기타 조사 : 0.3

```
#define JOSA_EUN        1.4     // '은/는'
#define JOSA_KA         1.3     // '이/가'
#define JOSA_EUL        1.4     // '을/를'
#define JOSA_MAN        1.0     // '만'
#define JOSA_EUI        1.2     // '의'
#define JOSA_AE         2.0     // '에'

#define JOSA_1SYL       1.1     // 1음절 조사
#define JOSA_2SYL       0.8     // 2음절 이상의 조사
#define JOSA_NONE       1.0     // 조사가 없는 경우
#define JOSA_etc        0.4     // 기타 조사

#define EOMI_vsfx       0.2     // '하다/되다/시키다' 등
#define EOMI_etc        0.1     // 기타 어말어미
```

그림 9. 조사 유형에 따른 가중치 비율

4.4 질의어 가중치 부여 알고리즘

질의어 문장에서 추출된 용어와 문장에 대한 형태소 분석 결과로부터 용어 가중치를 부여하는 알고리즘은 그림 10과 같다. 이 알고리즘에서 색인어의 유형에 따른 가중치 부여 기준은 다음과 같다.

- 전문 용어 사전(domain dictionary)에 수록된 용어
  함수 is_domain_noun()
- 시간성 명사 등 수식어 역할을 하는 용어
  함수 is_modi_noun()
- 질의어 문장에 관용적으로 사용되는 용어
  함수 is_query_term()
- 검색하고자 하는 문서의 유형을 지정하는 용어
  함수 is_term_doctype()
- 붙여쓴 복합명사, 띄어쓴 복합명사의 구성 요소

전문 용어 사전에 수록된 용어는 가중치를 높게 부여하고, 시간성 명사 등 수식어 역할을 하는 용어는 가중치를 낮게 부여한다. 전문 용어와 시간성 명사를 제외한 용어들은 용어 유형에 따라 기본 점수를 부여한다.

그림 10은 질의어의 문장 패턴에 의한 용어 가중치 부여 알고리즘이다. 먼저, 입력 문장의 각 어절들에 대해 명사구 chunking을 수행하여 각 어절에 대해 tag를 부여한다. 다음으로 용어 유형에 의한 가중치 적용 기법에 의해 각 용어마다 가중치를 부여한다. 명사구 chunking 정보를 이용하여 각 용어들의 가중치를 조절한다. 문장 패턴 정보를 이용한 용어 가중치 계산 과정은 다음과 같다.

- 명사구 chunking
- 용어 유형에 의한 가중치 계산
- 질의어 구문 유형에 따라 가중치 조절

질의어의 초기 가중치는 전문 용어와 시간성 명사를 제외한 용어들은 붙여쓴 복합명사, '하다/되다' 등 용언화 접미사가 결합된 용어, 미등록어 등 용어 유형에 따라 기본 점수를 부여한다.

```
형태소 분석 및 용어 추출;
명사구 인식(NP chunking);
for (each term in query sentence) {
    용어 유형 가중치로 초기화;
    용어 길이에 의한 가중치 조절;
    조사 유형에 의한 가중치 조절;
    질의 구문에 의한 가중치 조절;
    문서 유형 용어의 가중치 조절;
}
```

그림 10. 질의어 가중치 부여 알고리즘

5. 결론 및 향후 연구

질의 문장에서 추출된 용어들의 가중치를 계산하기 위하여 질의 문장들의 유형을 분석하였으며, 질의 문장의 유형에 따라 질의어 가중치를 계산하는 연구를 수행하였다. 명사구에 대해 질의 문장 유형에 따른 가중치를 동등하게 적용하기 위하여 명사구 chunking을 수행하였고, 각 용어들에 대해 용어 유형에 의한 가중치를 계산하였다.

용어 유형에 의한 가중치 계산은 추출된 명사 유형에 의한 가중치 부여, 용어 길이 특성에 의한 가중치 부여, 전문 용어 사전과 부사성 명사 사전에 의한 용어 가중치 부여, 그리고 조사 유형에 의한 가중치 부여 기법을 사용하였다.

질의어 가중치 기법의 효용성을 검증하려면 기존의 통계적 기법에 의한 질의어 가중치 기법과 비교했을 때 검색 결과의 성능이 얼마나 향상되었는지에 대한 실험 및 검색 엔진에서 실제로 질의되고 있는 다양한 유형의 질의어 문장에 대한 실험에 의해 성능을 개선하는 작업이 필요하다.

질의어 가중치와 더불어 정보 자료에서 추출된 색인어의 가중치를 계산하여 정보 자료의 내용을 대표하는 주제어를 선별하는 색인어 가중치 계산 기법에 의해 정보 검색 결과의 정확도를 향상시킬 수 있을 것으로 추정된다.

6. 참고 문헌

[1] R. Baeza-Yates, B. Ribeiro-Neto, *Modern Information Retrieval*, Addison Wesley, 1999.

[2] Salton, G. and C. Buckley, "The Term-Weighting Approaches in Automatic Text Retrieval", Information Processing and Management, vol. 24, no. 5, pp.513-523, 1988.

[3] 강승식, 이하규, 손소현, 홍기채, 문병주, "조사 유형 및 복합명사 인식에 의한 용어 가중치 부여 기법", 제28회 한국정보과학회 학술발표 논문집(II), 28권 2호, pp.196-198, 2001.

[4] 김재군, 김영환, 김성혁, "한국어 정보검색연구를 위한 시험용 데이터 모음(KTSET) 개발", 제6회 한글 및 한국어 정보처리 학술발표 논문집, pp.378-385, 1994.