한국어 정보 검색에서 의미적 용어 불일치 완화 방안

윤 보 현[†] · 박 성 진[††] · 강 현 규[†††]

요 약

정보검색시스템은 색인어와 질의어가 정확히 일치하지 않더라도 사용자 질의에 적합한 문서를 검색할 수 있어야 한다. 그러나, 색인어와 질의어간의 용어 불일치는 검색성능의 개선에 심각한 장애요소로 작용해 왔다. 따라서, 본 논문에서는 문서 코퍼스의 단어들간에 자동 용어 정규화를 수행하고, 용어 정규화의 산물을 한국어 정보검색 시스템에 적용하는 방안을 제시한다. 용어 불일치를 완화하기 위해 두가지 용어 정규화, 동치부류와 공기단어 클러스터를 수행한다. 첫째, 음역어, 철자 오류, 그리고 동의어를 위해 문맥 유사도를 이용하여 동치부류로 구축하는 작업이다. 둘째, 상호정보와 단어 문맥의 조합을 이용하여 단어 유사도를 계산하고 문맥 기반 용어를 정규화한다. 그런 다음, K-means 알고리즘을 이용하여 자율 클러스터링을 수행하고 공기단어 클러스터를 구축한다. 본 논문에서는 이러한 용어 정규화의 산물들을 용어 불일치를 완화하기 위해 질의어 확장과정에서 사용한다. 다시 말해서 동치부류와 공기단어 클러스터는 새로운 용어로 질의를 확장하는 자원으로서 사용된다. 이러한 질의확장으로 사용자는 질의어에 음역어를 추가하여 질의어를 포괄적으로 만들거나 특정어를 추가하여 질의어를 세밀하게 만들 수 있다. 질의어 확장을 위해 두 가지 상호보완적인 방법인 용어 제시와 용어 적합성 피드백을 이용한다. 실험결과는 제안된 시스템이 의미적 용어 불일치를 완화할 수 있고, 적절한 유사도 값을 제공할 수 있음을 보여준다. 결과적으로 제안한 시스템이 정보 검색 시스템의 검색 효율을 향상시킬 수 있음을 알 수 있다.

Alleviating Semantic Term Mismatches in Korean Information Retrieval

Bo-Hyun Yun[†] · Sung-Jin Park[††] · Hyun-Kyu Kang[†††]

ABSTRACT

An information retrieval system has to retrieve all and only documents which are relevant to a user query, even if index terms and query terms are not matched exactly. However, term mismatches between index terms and query terms have been a serious obstacle to the enhancement of retrieval performance. In this paper, we discuss automatic term normalization between words in text corpora and their application to a Korean information retrieval system. We perform two types of term normalizations to alleviate semantic term mismatches : equivalence class and co-occurrence cluster. First, transliterations, spelling errors, and synonyms are normalized into equivalence classes by using contextual similarity. Second, context-based terms are normalized by using a combination of mutual information and word context to establish word similarities. Next, unsupervised clustering is done by using K-means algorithm and co-occurrence clusters are identified. In this paper, these normalized term products are used in the query expansion to alleviate semantic term mismatches. In other words, we utilize two kinds of term normalizations, equivalence class and co-occurrence cluster, to expand user's queries with new terms, in an attempt to make user's queries more comprehensive (adding transliterations) or more specific (adding specializations). For query expansion, we employ two complementary methods : term suggestion and term relevance feedback. The experimental results show that our proposed system can alleviate semantic term mismatches and can also provide the appropriate similarity measurements. As a result, we know that our system can improve the retrieval efficiency of the information retrieval system.

† 정 회 원 : 한국전자통신연구원 선임연구원 언어이해연구팀
†† 종신회원 : 한신대학교 정보시스템공학과 교수
††† 종신회원 : 한국전자통신연구원 책임연구원 언어이해연구팀장
논문접수 : 2000년 2월 2일, 심사완료 : 2000년 11월 22일

1. 서 론

한국어 문서에는 색인어와 질의어간에 의미적 용어 불일치가 발생하여 정보검색시스템의 성능을 향상시키는데 장애요소로 작용하여 왔다. 의미적 용어 불일치는 철자오류, 음역어, 동의어 또는 다의어로 인하여 생겨난다.

첫째, 동의어는 하나의 용어 "분석"은 "분해", "분리", 그리고 "분할"과 비슷한 의미를 가진다. 예를 들어 질의어로 "분석"이 입력되면 문서내의 색인어 "분석" 뿐만아니라 "분해", "분리" 그리고 "분할"을 같은 단어로 처리하여야 한다. 또한 질의어와 같은 단어이지만 문서상에 잘못 표기된 철자오류도 용어불일치를 일으킨다. 마지막으로 한국어 문서에는 하나의 외국단어에 대해 여러 개의 음역된 외래어가 존재하므로 의미적 용어 불일치를 일으키는 요인이 되기도 한다.

한국어에서 외래어는 영어와 같은 원시 언어에서 그 원어(origin)를 가지지만, 원시 언어의 한 원어는 하나 이상의 외래어로 음역될 수 있다. 예를 들어, KTSET 2.0에서 다양한 음역의 예는 <표 1>과 같다. 하나의 영어단어 'digital'은 세 개의 외래어 '디지틀', '디지탈', 그리고 '디지털'로 음역되어 문서에서 실제로 사용되고 있다[6]. 따라서 이러한 현상을 고려하지 않는다면, 정보검색시스템의 성능 향상을 기대하기는 어렵다.

〈표 1〉 KTSET 2.0에서 다양한 음역의 예[1]

영어단어	음역된 외래어		
data	데이타(933)	데이터(563)	
digital	디지틀(269)	디지탈(246)	디지털(4)
radio	라디오(26)	레이디오(3)	

둘째, 다의어는 하나의 키워드가 여러 개의 의미를 지녀서 사용자의 질의어와는 의미가 전혀 다른 색인어로 색인된 문서를 검색하게 한다. 예를 들어, 사용자가 컴퓨터 분야의 '바이러스'에 관한 문서를 찾고자 질의를 입력하면, 문서상에는 색인어 '바이러스'를 포함하고 있는 컴퓨터 분야의 문서와 의학 분야의 문서가 함께 검색될 것이다.

유사도 계산 과정에서는 동의어와 다의어로 인하여 다음과 같은 문제가 발생한다. 일반적으로 벡터공간모델은 질의와 문서를 용어집합으로 표현하고, 질의와 문서간에 유사도를 계산한다. 질의와 문서는 식 (1)과 식 (2)와 같이 용어 벡터로 표현된다.

$$D_i=(d_{i1},\cdots,d_{ik},\cdots,d_{it}) \quad (1)$$

$$Q_j=(q_{j1},\cdots,q_{jk},\cdots,q_{jt}) \quad (2)$$

여기에서 계수 d_{ik}와 q_{jk}는 문서 D_i와 질의 Q_j에서 k번째 용어의 값을 나타낸다. 유사도 측정은 식 (3)에 의해 문서의 용어 벡터와 질의의 용어벡터간에 코사인 각을 측정값으로 한다.

$$Sim(D_i,Q_j)=\frac{\sum_{k=1}^{t} d_{ik}\times q_{jk}}{\sqrt{\sum_{k=1}^{t} d_{ik}^2}\times\sqrt{\sum_{k=1}^{t} q_{jk}^2}} \quad (3)$$

이와 같은 과정에서 동의어는 문서검색의 유사도 계산 과정에서 유사도 측정을 과소평가하는 경향이 있다. 즉, 한 용어에 대한 동의어들도 유사도 계산 과정에서 반영되어 문서 D_i나 질의 Q_j가 확장되어 측정되어야 한다. 반대로 한 단어가 여러 개의 의미를 가지는 다의어는 용어 매칭 과정에서 유사도 측정을 과대평가하는 경향이 있다. 즉, 다의어로 인해 질의에 입력된 단어의 의미가 아닌 단어를 포함한 문서까지도 검색될 수 있다. 따라서 검색성능을 향상시키기 위해서는 유사도 측정값을 적당히 계산할 수 있어야 한다.

의미분석을 수행하지 않고서도 다음과 같은 용어 정규화의 산물을 통해 의미적 용어 불일치를 완화할 수 있다.

- 동치부류 구축 : 다양한 음역어, "디지탈", "디지털", 그리고 "디지틀"은 한 부류로 그룹화 된다.
- 공기단어 클러스터 구축 : 관련 단어, "바이러스", "컴퓨터", 그리고 "백신" 그리고 "프로그램"은 하나의 클러스터로 그룹화 된다.

본 논문에서는 의미적 용어불일치를 완화하기 위해 동치부류와 공기 단어 클러스터를 생성하는 용어 정규화를 수행한다. 이러한 용어 정규화의 산물들을 용어 불일치를 완화하기 위해 질의어 확장에 사용하는 방안을 제시한다.

2. 관련연구

동치부류를 구축하기 위한 연구는 크게 두가지로 나

1) 괄호안의 숫자는 발생 빈도이다.

번다. 첫째, 외래어를 포함한 어절을 찾고 통계정보를 이용하여 어절로부터 외래어를 추출한다. 추출된 외래어를 음절 bi-gram과 DLM(Damerau-Levenshtein Metric)을 이용하여 동치 부류를 구축하는 방법[3]이다. 둘째는 영어단어에서 한글 단어로 음역하기 위해 STM (Statistical Transliteration Model)을 이용하는 방법[4]이다. 이 방법은 음역문제를 주어진 단어에 대한 가장 가능한 음역어를 발견하는 것으로 본다. $p(K|E)$는 영어단어 E에 대한 한국어 단어 K의 확률이다. 단어 K에 대한 음역 확률은 $p(K)$이다. 베이지언 규칙에 의해 음역문제는 식 (4)와 같이 변경될 수 있다.

$$argmax_k p(K|E) = argmax_k \frac{p(K)p(E|K)}{p(E)} \quad (4)$$

그러나 이러한 방법들은 외래어 동치부류를 구축하는데는 효율적이지만 철자오류까지 동치부류를 구축하기는 어렵다.

의미적 용어불일치를 완화하기 위해 관련된 용어를 한 부류로 구축하는 용어 정규화를 수행하여 이를 색인이나 질의어 확장에 이용하는 연구가 있다. 이를 위해 다양한 기법, 즉 용어분류(term classification) 기법, 구문문맥(syntactic context) 기법, 어휘영역맵(lexical domain map) 기법, 유사도 시소러스(similarity thesaurus) 기법, 용어연관정보(term correlation information) 기법, 어휘-의미관계(lexical-semantic relationship) 기법 등이 있다.

용어분류 기법[10]은 유사도 값에 따라 용어를 부류(class)로 그룹화한다. 용어간 유사도는 연관가정(association hypothesis)에 기초하여 계산되고 임계치를 정해서 용어를 분류한다. 이러한 부류는 질의어 확장과정에서 검색 용어가 포함된 부류를 질의에 추가하는 방법이다. 이 방법은 가장 단순하지만 효율이 좋지 못하는 단점이 있다. 구문 문맥 방법[2, 12]은 검색 용어를 개선하기 위해서 언어 지식을 이용한다. 용어 관계는 언어지식과 공기단어 통계치에 의해 생성된다. 이 방법은 각 용어를 위한 용어 리스트를 추출하기 위해서 문법과 사전을 이용한다. 질의어확장은 질의어와 가장 비슷한 용어를 찾아 추가함으로써 이루어진다. 이 방법은 초기 질의보다 약간 나은 결과를 보일 수 있다. 어휘영역맵 기법[14]은 수식어-중심어 쌍을 추출하여 가중치가 있는 Jaccard 계산식[13] 형에 의해 용어 유사도를 계산한다. 구축된 맵은 검색성능 향상을 위해 질의확장에 적용된다. 유사도 시소러스 기법[1, 11]은 문서집합의 용어가 색인된 상태에 기초한 용어-용어 유사도 행렬을 이용한다. 확률 방법이 벡터공간 모델에서 주어진 질의에 유사한 용어의 확률을 추정하기 위해서 사용된다. 이 방법은 검색 성능을 상당히 개선한다고 보고되고 있다. 용어연관정보 기법[7]은 단어간의 상호관계를 추정하기 위해 용어 상호관계 정보를 이용하여 주어진 질의에 독립적인 개념들을 찾는 방법이다. 어휘-의미관계 기법[8]은 통계적으로 계산된 문맥 정보를 이용함으로써 획득되는데, 문맥 정보는 상호 정보에 의해 계산된다. 문서집합의 모든 단어 쌍에 대해 의미적 용어 유사도가 계산되고 클러스터링 알고리즘이 비슷한 용어를 그룹화하기 위해서 이용되고 계층적인 구조를 형성한다.

3. 용어 정규화

3.1 동치부류 구축

음역어는 영/한이나 일/한 등의 언어간에 음소 번역된 단어이다. 특히 한국어 음역어에는 한 외국어에 대해 여러 개의 음역어가 존재한다. 이것은 여러 나라마다 음가 체계가 다르기 때문이다. 한국어 문서상에서 하나의 영어 단어에 대해 여러 개의 음역된 단어가 존재하고 있다. 게다가 다양한 형태의 철자오류도 존재하고 있다. 표준어와 음역어를 그룹화할 수 있는 동치부류는 정보검색시스템에 유용하게 사용될 수 있다. 색인을 위해 같은 의미의 음역어들이 용어 불일치를 줄이기 위해서 같이 색인될 수 있고, 질의 처리 과정에서 질의어는 여러 음역어로 확장되어 검색 성능을 향상시킬 수 있다.

본 논문에서는 동의어, 음역어, 그리고 철자오류에 대한 용어불일치를 완화하기 위해 다음과 같은 세 단계를 수행하여 동치부류를 구축한다.

3.1.1 표준어후보 생성 단계

표준어 후보 생성 단계에서는 음역어나 철자오류가 주어지면 음성에러 패턴에 기초하여 규칙(rule)에 의해 표준어 후보를 생성하는 과정을 말한다. 본 논문에서 한 단어에 대한 음역어나 철자오류 여부는 형태소 분석 결과를 보고 판단한다. 즉, 단어가 형태소 분석이 되지 않으면, 음역어나 철자오류로 간주한다. 한편, 철자오류 교정 분야에서 올바른 단어후보를 생성하는 과정은 철자오류 패턴의 지식을 표현하는 휴리스틱 규칙

에 기초한다[5, 9]. 본 논문에서도 음역어나 철자오류에 대한 표준어 후보를 생성하기 위해 철자오류 교정에서의 휴리스틱 규칙과 비슷한 규칙을 이용한다. 그러나 모든 종류의 음역어나 철자오류를 표현하는 규칙은 너무 많은 후보를 생성하기 때문에 고빈도 에러만을 다룰 수 있는 규칙을 구성한다. 실험적으로 음소에러와 몇 개의 철자오류가 자주 발생하는 고빈도 에러였으므로 이들을 다룰 수 있는 규칙만을 이용한다.

규칙은 초성 자음을 변환하는 규칙, 중성 모음을 변환하는 규칙, 종성 자음을 변환하는 규칙으로 나뉘어진다. 초성 자음을 변환하는 규칙의 예는 "현재 음소의 첫 자음이 "ㄲ"이면, 첫 자음 "ㄲ"을 "ㄱ"과 "ㅋ"으로 변환한다."이다. 중성 모음을 변환하는 규칙의 예는 "현재 음소의 중성 모음이 "ㅏ"이면 "ㅐ", "ㅓ", "ㅘ", 그리고 "ㅡ"로 변환한다."다. 마지막으로 종성 자음을 변환하는 규칙의 예는 "중성 모음이 "ㅏ"이고 현재 음소으 종성 자음이 "ㄹ"이면, "ㄹ"을 "ㄹㅎ"으로 변환한다."이다.

3.1.2 문맥 유사도 계산 단계

먼저 음역어나 철자오류 x에 대한 문맥비트벡터 C_x는 식 (5)과 같이 정의된다.

$$C_x=(x_1(b_1),\cdots,x_i(b_i),\cdots,x_n(b_n)) \quad (5)$$

여기에서 b_i는 i번째 문맥단어 x_i가 정해진 윈도우(window)내에 존재하면 1이고, 그렇지 않으면 0이고, n은 사전내의 명사의 총 개수이다.

표준어후보 y에 대한 문맥비트벡터 C_y는 식 (6)와 같이 정의된다.

$$C_y=(y_1(b_1),\cdots,y_j(b_j),\cdots,y_n(b_n)) \quad (6)$$

여기에서 b_j는 j번째 문맥단어 y_j가 정해진 윈도우(window)내에 존재하면 1이고, 그렇지 않으면 0이고, n은 사전내의 명사의 총 개수이다.

음역어나 철자 오류와 표준어 후보의 문맥비트벡터간에 유사도는 식 (7)에 의해 계산된다.

$$Sim(C_x,C_y)=\frac{\sum_{i=1}^{n} b_i\times b_j}{\min(\sum_{i=1}^{n} b_i,\sum_{j=1}^{n} b_j)} \quad (7)$$

식(7)에서 분모 $\min(\sum_{i=1}^{n} b_i,\sum_{j=1}^{n} b_j)$는 표준어후보나 철자오류중의 저빈도를 선택하기 위한 식으로 이 저빈도로 정규화 함으로써 저빈도 음역어나 철자오류를 효율적으로 다룬다. 계산된 문맥유사도가 정해진 임계치 이상이면 표준어와 음역어나 철자오류들을 묶어 동치부류를 구축한다.

3.1.3 동치 부류 구축 단계

동치부류 구축 단계는 하나의 철자오류에 대한 여러 표준어 후보 중에서 철자오류와 표준어 후보간 문맥유사도를 이용하여 하나의 표준어를 선택하고, 철자오류와 표준어를 동치부류로 구축하는 과정이다.

동치부류 구축을 평가하기 위한 인자는 다음과 같다.

$$\text{인식도}=\frac{\text{인식된 표준어의 개수}}{\text{문서에 존재하는 표준어의 총 개수}}$$

$$\text{정확도}=\frac{\text{정확한 표준어의 개수}}{\text{인식된 표준어의 총 개수}}$$

인식도는 얼마나 많은 표준어가 인식되었는지를 평가하고, 정확도는 얼마나 많은 표준어가 정확하게 인식되었는지를 평가하는 인자이다. <표 2>에서 각 유사도에 따른 인식도와 정확도를 보여주고 있으며, 문맥유사도가 0.0보다 클 때, 정확도는 95.1%임을 알 수 있다. 철자 오류와 표준어 후보간 유사도가 0이상일 때 동치부류의 인식도가 76.6%라는 의미는 유사도가 0이상일 때 표준어의 76.6%를 인식했다는 것을 의미한다. 또한 정확도 95.1%는 인식한 표준어 중에서 95.1%가 정확하게 표준어를 인식했다는 것을 의미한다.

〈표 2〉 동치부류 구축 결과

문맥유사도	인식도	정확도
Sim > 0.30	14.3%	98.5%
Sim > 0.0	76.6%	95.1%

동치부류 구축의 예는 <표 3>에서 보여준다. 부류

〈표 3〉 동치 부류 구축의 예

부류	영어 단어	한국어 단어
부류 1	Scheduling	스케쥴링, 스케듈링, 스케주링, 스케줄링
부류 2	Transaction	트랜잭션, 트렌잭션, 트랜잭션, 트란잭션
부류 3	Parameter	파라메터, 파라미터, 파라미타, 파라메타
부류 4	Summation	합계, 합계, 합게
부류 5	Machine	기계, 기기, 기게

1에서 부류 3은 다양한 음역어를 포함한 부류이다. 부류 4는 표준어 "합계"에 대한 철자오류 "함계"와 "합게"를 보여주고 있다. 부류 5는 표준어 "기계"에 대한 동의어 "기기"와 철자오류 "기게"를 보여주고 있다. 이러한 표에서 제안한 방법이 음역어, 철자오류, 그리고 동의어까지도 그룹화할 수 있음을 보인다.

3.2 공기단어 클러스터 구축

문맥기반 클러스터의 기본 아이디어는 다음과 같다. 한 개념의 문맥은 개념간 유사성을 결정하기 위해 사용될 수 있다. 문맥은 여러 가지로 정의될 수 있고, 개념도 마찬가지이다. 가장 단순한 정의는 단어를 개념으로 간주하고 한 단어에 대한 문맥은 그 단어와 동시발생하는 모든 단어이다. 본 논문에서 개념은 명사 그룹이고, 문맥은 개념을 둘러싼 고정 길이 윈도우의 모임으로 정의된다. 명사그룹은 단일명사이거나 복합명사이고, 윈도우 크기는 10이다.

본 논문에서 공기단어 클러스터를 구축하는 방법은 다음과 같이 네단계로 이루어진다.

3.2.1 공기단어 추출

입력 문서가 주어지면, 형태소 분석기와 품사태거[2] [16, 17]는 품사 태깅된 문서를 생성한다. 태깅된 문서에서 윈도우 크기 10내의 명사그룹과 빈도를 추출한다.

3.2.2 상호정보 계산

목적 단어 x와 문맥 단어 y간에 상호 정보는 다음과 같은 식 (8)에 의해 구해진다.

$$MI(x,y)=\log\frac{N\times f(x,y)+1}{f(x)\times f(y)} \quad (8)$$

여기에서 N은 공기단어의 총 개수이다. 이것은 목적 단어 x와 함께 나타나는 문맥 단어 y에 대한 상호 정보 값이다. 동시발생빈도가 0일 경우 상호 정보의 값이 0이 되므로 1을 더한다.

3.2.3 명사 벡터 구축

구해진 상호 정보값이 5이상인 명사들의 명사 벡터를 구축한다. 어떤 명사와 동시발생하는 명사들이 $N_1, N_2, \cdots, N_n$일 경우, 명사 벡터 $x^{(p)}$는 다음과 같이 정의된다.

$$x^{(p)}=(f(N_1,N_p),f(N_2,N_p),\cdots,f(N_n,N_p))$$

여기에서 $x^{(p)}$는 코사인 정규화 인자[13]에 의해 정규화된다.

3.2.4 용어 클러스터링

용어 클러스터링을 위해 본 논문에서는 최단 거리에 의해 클러스터링을 수행하는 방법중의 하나인 K-means 알고리즘을 이용한다. K-means 알고리즘을 이용하는 이유는 K개의 클러스터 이하로 반드시 수렴하기 때문이다. 그러나 적당한 K 값을 정해야 하는 문제가 있다. 본 논문에서는 컴퓨터 과학 분야의 문서를 포함하고 있기 때문에 의미의 종류가 많지 않아 K값을 3으로 정한다. 예를 들어, 질의어 "바이러스"에 대한 공기단어 클러스터는 <표 4>와 같다. 공기단어 클러스터는 질의를 특정화하기 위해서 사용될 수 있다.

〈표 4〉 공기단어 클러스터의 예

Cluster 1	디스켓, 디스크, 컴퓨터, 미켈란젤로, 발견, 백신, 변경, 복사, 신뢰도, 신종, 실행, 악성, 예루살렘, 예방, 이름, 조작, 침입, 침투, 침해, 감염, 퇴치, 금요일, 대책, 등장, 파괴, 파일, 프로그램, 패해
Cluster 2	에이즈, 유전자, 공격, 돌연, 몸, 방어, 변이, 세포, 의구심

4. 질의어 확장

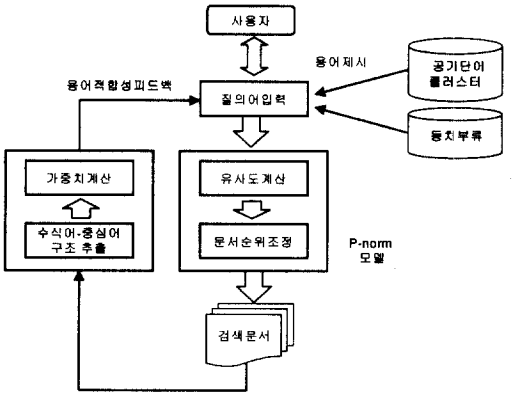
질의에 관련된 용어를 추가하여 재검색하는 것은 색인어와 질의어간에 매칭을 위해 사용되는 용어의 수가 증가함으로써 인식된 적합한 문서의 개수를 증가시킬 수 있다. 따라서 정규화된 용어는 질의의 용어를 세밀하게 또는 포괄적이도록 할 수 있다. 본 논문에서 용어 정규화를 통해 얻어진 동치부류와 공기단어 클러스터는 질의어 확장을 위해 사용된다. 대안으로는 정규화의 산물을 색인과정에서도 이용할 수 있으나 질의어 확장을 위해 사용하는 것이 시스템에 융통성을 부여하기 때문에 질의어 확장과정에서 이용한다. 또한 만약 색인 과정에서 그들을 이용한다면 색인의 양이 매우 커지는 문제가 발생한다. 아울러 용어 정규화의 산물이 부적당하게 구축되었을 경우 부적당한 색인어가 시스템의 성능을 저하시키는 요인이 되기도 한다.

기존의 질의어 확장에 관한 연구는 크게 용어 제시와 용어 적합성 피드백으로 나뉘어 진다. 용어 제시는

2) 한국어 분석기와 품사 태거는 고려대학교 자연어처리 연구실에서 이미 개발된 시스템을 이용한다.

검색 전에 질의를 확장하는 방법이고 용어 적합성 피드백은 검색 후에 재검색을 수행하기 위해 질의를 확장하는 방법이다. 즉, 기존의 연구는 동치부류를 구축하여 검색 전에 질의를 확장하는 용어제시만을 수행하거나 용어 정규화를 수행하여 검색 후에 질의를 확장하는 용어적합성피드백만을 수행한다. 하지만 본 논문에서 제시하는 방법은 동치 부류와 공기단어 클러스터를 구축하여 검색 전/후에 질의를 확장할 수 있도록 용어제시와 용어적합성피드백을 함께 수행할 수 있는 방법이다. 따라서 본 논문에서 제시하는 방법이 기존의 연구방법보다 질의 확장에 있어서 융통성이 있다.

(그림 1)은 본 논문에서 제안하는 검색시스템 구성도이다. 사용자가 질의를 입력하면 동치부류와 공기단어 클러스터를 이용하여 검색 전에 용어제시를 수행하고, 검색 후에 용어 적합성 피드백을 수행한다.



(그림 1) 검색시스템 구성도

4.1 용어 제시

정보검색에서 사용자는 대부분 몇 개의 명사를 나열하는 정도로 짧게 질의한다. 이러한 상황에서 용어에 대한 단어 의미중의성을 해결하기는 매우 어렵다. 또한 초기 질의에 의해 검색된 문서중에 적합한 문서가 없다면 적합성 피드백을 위해 적합한 용어를 추출할 수 없다. 그리하여 검색 전에 용어를 제시하여 사용자로 하여금 선택할 수 있도록 도와주어야 한다.

따라서 본 논문에서는 용어 정규화의 산물인 동치부류와 공기단어 클러스터를 이용하여 사용자의 넓은 의미의 질의 용어를 좁은 의미의 질의 용어로 유도하여 보다 정확한 검색을 수행한다. 첫째, 사용자가 질의어를 입력하면, 이미 구축된 동치 부류를 탐색하여 이 질의어가 속한 동치 부류에 속한 모든 단어를 질의에 첨가하여 질의를 확장한다. 만약 질의어로 "스케줄링"이 입력되었다면, 이러한 동치 부류에 속한 모든 엔트리 "스케쥴링", "스케듈링", 그리고 "스케주링"을 참조하여 질의어에 자동으로 사용자의 적합성 판단없이 추가된다.

둘째, 사용자 질의의 용어에 일치하는 공기단어 클러스터는 사용자에게 제시되고, 사용자가 질의에 추가할 것인지를 판단한다. 예를 들어 만일 사용자가 컴퓨터 바이러스에 관한 문서를 원할 경우 <표 4>의 두 클러스터중 클러스터 1의 용어들을 선택하여 추가하면 된다. 반면에 의학에 관한 문서를 검색하고자 할 경우 사용자는 초기 질의에 클러스터 2의 용어를 첨가하면 된다.

이와 같은 질의어 확장은 입력된 초기 질의에 직접 추가됨으로써 사용자가 원하는 방향으로 검색 결과를 쉽게 유도할 수 있다. 또한 한국어 문서에 존재하는 다양한 음역어나 철자오류를 고려해줄 수 있어 검색의 향상을 가져올 수 있다.

4.2 용어 적합성 피드백

적합성 피드백은 문서 적합성 피드백과 용어 적합성 피드백으로 나뉘어진다. 문서 적합성 피드백은 초기 질의로 검색된 문서를 사용자에게 보여주고 사용자가 적합한 문서를 선택하면, 시스템이 적합한 문서에서 용어를 추출하여 질의에 추가한다. 이 방법은 구현하기 쉽지만 사용자가 용어선택을 전혀 제어하지 못한다. 따라서 본 논문에서는 검색된 상위 5개 문서에서 용어를 추출하고 가중치를 계산하여 사용자에게 보여주고, 사용자가 원하는 방향의 용어를 질의에 추가하는 용어 적합성 피드백으로 질의를 확장한다.

용어 적합성 피드백에서 중요한 것은 시스템이 의미있는 용어를 추출하여 사용자가 선택할 수 있도록 해야 한다. 이를 위해 문장 중에 같은 의미의 내용이 구문적으로 다르게 표현될 때 하나의 일관된 형태, 즉 수식어-중심어(modifier-header) 구조를 추출한다.

따라서 본 논문에서는 복잡한 구문 분석을 수행하지 않더라도 의존관계를 이용하여 수식어-중심어 구조를 추출한다. 의존관계는 수식어-중심어 구조의 수식어와 중심어간의 구문관계를 나타낸다. 본 논문에서는 하나의 절내에 있는 명사들간에 교차(Crossing Branch)를 허용하여 보다 많은 수식어-중심어 구조를 추출한다.

본 논문에서 사용하는 의존관계는 <표 5>와 같다.

〈표 5〉 의존 관계

Types	명사구 수식어	명사구 중심어
Type(1)	격조사가 없는 명사	후위 명사
Type(2)	소유격조사가 붙은 명사	후위 명사
Type(3)	주격조사, 목적격조사, 부사격조사, 관형사격조사, 또는 보조격조사가 붙은 명사	후위 동작성 명사나 상태성 명사
Type(4)	관형사형 어미가 붙은 명사	후위 명사
Type(5)	관형어형 동사가 붙은 명사	후위 명사

<표 1>의 의존 관계를 이용하여 수식어-중심어 구조를 추출하는 과정은 다음과 같다.

(1) 하나의 절내에 있는 단어에 색인번호를 할당한다.
(2) 절을 오른쪽에서 왼쪽으로 스캔해 가면서 만약 현재 단어가 명사상당어구이면, 이 명사를 수식어-중심어 구조의 중심어로 간주하고 단계 (3)으로 간다. 그렇지 않으면 명사상당어구를 찾을 때까지 반복한다.
(3) 단계 (2)에서 발견된 중심어에서 왼쪽으로 절을 스캔한다. 만약 현재 단어가 명사상당어구이면, 이 명사를 수식어-중심어 구조의 수식어로 간주한다.
(4) 단계 (2)에서 발견된 중심어와 단계 (3)에서 수식어로 이루어진 수식어-중심어 구조가 의존관계에 속하는지 판단한다. 속하면 수식어-중심어 구조의 후보로 추출하고, 그렇지 않으면 의미없는 수식어-중심어 구조로 무시한다.

예를 들어, "정보를 검색하는 시스템"에서 "시스템"이라는 중심어를 단계 (2)에서 발견하면 중심어 "시스템"에서 왼쪽으로 절을 스캔한다. 현재 단어가 명사상당어구인 "검색"이므로 이를 수식어로 간주한다. 따라서 "검색시스템"이라는 수식어-중심어구조를 추출한다.

이렇게 추출된 의미있는 용어는 그 문서를 대표하는 정도를 표시하는 가중치가 부여되어야 한다. 가중치 계산을 위해 본 논문에서는 복합명사나 수식어-중심어 구조에 높은 가중치를 부여하기 위해 일반적인 가중치 계산식 $tf \times idf$을 변형한 다음과 같은 수식 (4)을 이용한다.

$$w = tf \times \log(N/n) \times |t| \quad (4)$$

여기에서 tf는 문서 D에서 색인어 t의 발생 회수이고, $\log(N/n)$는 idf로서 색인어 t가 나타나는 문서 D의 개수의 역이고, N은 전체 문서에서의 문서의 수를 의미한다. $|t|$는 용어의 길이이다.

5. 실험 및 평가

실험데이터는 전자와 전산분야에 관련된 논문 초록으로 구성된 KTSET 2.0이다. 문서수는 4,414개와 50개의 질의로 구성되어 있다. 질의문은 평균 4.27개 단어로 이루어지고, 질의문 하나당 평균 적합 문서수는 29개이다.

용어 제시와 용어 적합성피드백의 상대적 성능을 평가하기 위해 다음과 같은 다섯 가지 실험방법을 수행하였다. 실험에 있어서 기존에 제시된 방법을 이용하지 않고 baseline 방법을 사용한 이유는 기존의 연구가 국외 연구로서 색인 및 검색 시스템의 평가가 어렵기 때문이다. 아울러 동치 부류를 구축하는 연구는 이러한 동치 부류를 구축하는 방법론만 제시되었고, 동치 부류를 정보검색의 질의확장에 적용한 기존의 연구는 없으므로 비교가 불가능하다. 따라서 한국어 문서의 색인 및 검색을 위한 Baseline을 정해놓고 이와 비교한다.

- Baseline
  색인어와 검색어를 추출하기 위해 단일 명사와 복합 명사를 추출하고 복합 명사를 단일 명사로 분해하여 가중치를 계산한다. 검색시스템은 확장된 불리언 모델인 P-norm 모델에 기초한다.
- TS-EC(Term Suggestion using Equivalence Class)
  동치부류를 이용한 용어제시 방법
- TS-CC(Term Suggestion using Co-occurrence Cluster)
  공기단어 클러스터를 이용한 용어제시 방법
- TRF(Term Relevance Feedback)
  용어 적합성 피드백
- TS + TRF
  동치부류 및 공기단어 클러스터를 이용한 용어제시와 용어 적합성 피드백을 이용한 방법

〈표 6〉 다양한 문서레벨에 따른 정확률

방법 / 문서레벨	Baseline	TS-EC	TS-CC	TRF	TS + TRF
	정확률	정확률	정확률	정확률	정확률
5 문서	0.9244	0.9633(+3.9%)	0.9544(+0.3%)	0.9721(+4.8%)	0.9743(+5%)
10 문서	0.5072	0.5220(+1.5%)	0.5157(+0.8%)	0.5240(+1.7%)	0.5649(+5.7%)
15 문서	0.3630	0.3811(+1.8%)	0.3501(-1.3%)	0.3879(+2.4%)	0.3947(+3.1)
20 문서	0.3081	0.3224(+1.4%)	0.2602(-4.8%)	0.3457(+3.7%)	0.3581(+5%)
30 문서	0.2559	0.2648(+0.9%)	0.2459(-0.1%)	0.2633(+0.8%)	0.3043(+4.9%)
100 문서	0.2852	0.2322(-5.3%)	0.2025(-0.6%)	0.2242(-6.1%)	0.2924(+0.7%)
200 문서	0.1982	0.2194(+2.1%)	0.1909(-0.8%)	0.2078(+0.9%)	0.2236(+2.5%)
500 문서	0.1968	0.2078(+1.1%)	0.1882(-0.8%)	0.2043(+0.8%)	0.2148(+1.8%)

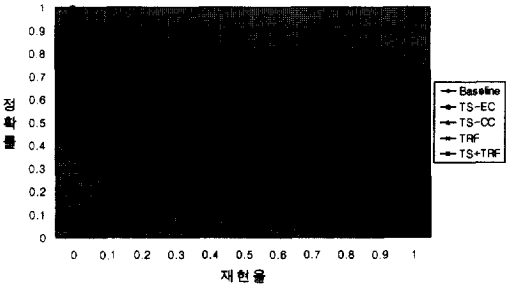
〈표 7〉 다양한 문서레벨에 따른 재현율

방법 / 문서레벨	Baseline	TS-EC	TS-CC	TRF	TS + TRF
	재현율	재현율	재현율	재현율	재현율
5 문서	0.8743	0.8822(+0.8%)	0.9366(+6.2%)	0.9378(+6.3%)	0.9422(+6.8%)
10 문서	0.4761	0.4832(+0.7%)	0.5400(+6.4%)	0.5358(+5.9%)	0.5016(+2.5%)
15 문서	0.3592	0.3526(-0.7%)	0.4111(+5.2%)	0.4069(+4.7%)	0.3737(+1.4%)
20 문서	0.2997	0.2886(+1.1%)	0.3355(+3.6%)	0.3518(+5.2%)	0.3143(+1.5%)
30 문서	0.2442	0.2592(+1.5%)	0.2914(+4.7%)	0.3071(+6.3%)	0.2893(+4.5%)
100 문서	0.2261	0.2288(+0.2%)	0.2618(+3.5%)	0.2774(+5.1%)	0.2487(+2.2%)
200 문서	0.2134	0.2269(+1.3%)	0.2609(+4.7%)	0.2765(+6.3%)	0.2286(+1.5%)
500 문서	0.2134	0.2132(+0%)	0.2609(+4.7%)	0.2765(+6.3%)	0.2286(+1.5%)

"Baseline"은 색인시스템에서 단일명사와 복합명사를 추출하고 복합명사를 단일명사로 분해하며[3][15, 18], $tf \times idf$ 계산식에 의해 가중치를 계산한다. 검색시스템에서는 확장된 불리언 검색 모델인 P-norm 모델을 이용한다. "TS-EC"에 의해 질의문은 평균 5.12개 단어로 확장되었고, "TS-CC"에 의해 질의문은 평균 7.25개 단어로 확장되었고, "TRF"에 의해 질의문은 평균 8.32개 단어로 확장되었다. 또한 "TS+TRF"에 의해 질의문은 평균 11.42개 단어로 확장되었다.

(그림 2)에서는 다섯 가지 방법에 대한 재현율에 따른 정확률을 측정한 실험결과이다. 모든 방법이 평균적으로 "Baseline"방법보다 나은 성능을 보였으나 "TRF" 방법이 상대적으로 적은 향상을 보이고 있다. 하지만 제안한 방법인 "TS+TRF"방법이 가장 나은 성능을 보이고 있다.

<표 6>과 <표 7>에서는 검색된 문서의 다양한 순위에 따른 다섯 가지 방법의 검색결과를 보여준다. "+"의 의미는 "Baseline" 방법에 대한 각 방법의 증가치를 의미하며, "-"는 각 방법의 감소치를 나타낸다. 표에서 "TS-EC"방법만 재현율을 향상시키지 못했고, 나머지 방법은 "Baseline"보다 상당히 성능을 개선시킴을 보여주고 있다. 특히 본 논문에서 제안한 용어제시와 용어적합성 피드백을 함께 사용한 방법이 가장 나은 성능을 보이고 있다. 따라서 본 논문에서 제안하는 질의어 확장 방법이 사용자 검색어휘를 확장하여 사용자가 원하는 방향으로 검색을 유도하는 효과를 얻을 수 있다. 아울러 제안하는 용어제시와 용어적합성피드백 방법은 동의어와 다의어로 인해 발생하는 의미적 용어불일치를 완화할 수 있음을 알 수 있다.



(그림 2) 질의어 확장 실험 결과

"TS-EC"방법이 재현율을 향상시키지 못한 이유는

3) 복합 명사 분해 시스템은 고려대학교 자연어처리 연구실에서 이미 개발된 시스템을 이용한다. 이 시스템은 통계 정보와 선호 규칙을 이용하여 복합 명사를 분해하는데 정확도는 약 96%정도이다.

실험 자료 구축중 적합성 판정 과정에서 음역어나 철자오류를 고려하지 않고서 표준어에만 기인하였기 때문이다. 예를 들어, KTSET의 8번 질의 "멀티미디어 데이터베이스에 대해 알고싶습니다."를 분석해본 결과, "멀티미디어"와 "데이타베이스"를 가지고 있는 문서는 30개였다. 그러나 "멀티미디어"와 "데이터베이스"를 가지고 있는 문서는 14개였다. 14개의 문서중 6개 문서만이 적합성 판정에 포함되었고 8개 문서는 제외되었다. 그러므로 적합성 판정이 정확하다면 "TS-EC"방법도 재현율과 정확률을 향상시킬 수 있다.

따라서 본 논문에서는 10개의 질의를 만들고 이에 대한 적합성 판정을 구축하여 "TS-EC"의 효율성을 제시하기 위해 "Baseline"과 "TS-EC"를 다시 실험을 수행하였다. 부록 1에서 새로운 질의와 확장된 질의를 보여주고 있다. <표 8>은 새로운 질의에 대한 "Baseline"과 "TS-EC"방법에 대한 실험 결과이다. 표에서 정확률은 약간의 향상을 보이나 재현율은 약 17%정도 개선됨을 알 수 있다. 그러므로 적합성 판정만 정확히 구축된다면 본 논문에서 제안한 방법이 성능을 상당히 개선시킬 수 있음을 알 수 있다.

〈표 8〉 새로운 질의에 대한 정확률과 재현율

방법 / 문서레벨	Baseline		TS-EC	
	정확률	재현율	정확률	재현율
5 문서	0.9600	0.9400	0.9900(+0.3%)	0.9800(+0.2%)
10 문서	0.8200	0.5928	0.8577(+3.7%)	0.8000(+20%)
15 문서	0.7575	0.4930	0.7615(+0.4%)	0.6983(+20%)
20 문서	0.7309	0.4597	0.7249(-0.6%)	0.6616(+20%)
30 문서	0.7093	0.4380	0.6903(-1.9%)	0.6335(+19%)
100 문서	0.6953	0.4365	0.6786(-1.7%)	0.6296(+19%)
200 문서	0.6923	0.4365	0.6771(-1.5%)	0.6296(+19%)
500 문서	0.6913	0.4365	0.6766(-1.5%)	0.6296(+19%)

6. 결 론

색인어와 질의어간의 의미적 용어 불일치는 검색성능의 개선에 심각한 장애요소로 작용해 왔다. 따라서 본 논문에서는 색인어와 질의어간에 발생하는 의미적 용어불일치를 완화하기 위해 용어정규화를 수행하여 대화적 질의 확장 방법으로 용어제시와 용어적합성피드백을 이용하는 정보검색방안을 제안하였다.

용어 정규화는 음역어, 철자오류, 그리고 동의어를 동치부류로 구축하기 위해 문맥유사도를 이용하였다. 또한 공기단어 클러스터를 구축하기 위해 상호정보와 K-means 알고리즘을 이용하였다. 또한 이러한 용어 정규화 산물들을 시스템에 융통성을 부여하기 위해 용어 제시와 용어 적합성 피드백에 이용하였다.

실험결과, 본 논문에서 제안하는 용어 제시와 용어 적합성 피드백을 함께 이용하는 방법이 가장 나은 성능을 보였다. 아울러 새로운 질의에 대한 실험으로 "TS-EC"방법이 적합성 판정만 정확히 구축된다면 검색성능을 개선할 수 있음을 보였다. 용어제시와 용어적합성피드백은 사용자 검색어휘를 확장하여 사용자가 원하는 방향으로 검색을 유도하는 효과를 얻을 수 있음을 보였다. 아울러 제안하는 용어제시와 용어적합성 피드백 방법은 동의어와 다의어로 인해 발생하는 의미적 용어불일치를 완화할 수 있음을 보였다. 향후에는 용어제시를 위해 한영사전이나 시소러스와 같은 다양한 자원을 이용하는 방안을 연구할 것이다

참 고 문 헌

[1] Han, C., Fujii, H., Croft, W.B., "Automatic Query Expansion for Japanese Text Retrieval," UMass Technical Report 95-11, 1995.

[2] Grefenstte, G., "Use of syntactic context to produce term association lists for text retrieval," *Proc. of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.89-97, 1992.

[3] Jeong, K. S., Kwon, Y. H., Myaeng, S. H., "Construction of Equivalence Class of Foreign Words through Automatic Identification and Extraction," *Proc. of Natural Language Processing Pacific Rim Symposium*, pp.335-340, 1997.

[4] Kim, B. H. "Automatic Transliteration of the English words into Hangul," Master Dissertation, Public Policy Graduated School, Seogang University, 1997.

[5] Kukich, K., "Techniques for Automatically Correcting Words in Text," *ACM Computing Survey*, Vol.24, No.4, December, 1992.

[6] Lee, J. S., Choi, K. S., "English to Korean Statistical Transliteration for information retrieval," *Computer Processing of Oriental Languages*, Vol.12, No.1,

pp.17-37, 1998.

[7] Mitra, M., Burkely, C., Singhal, A., Cardie, C., "An Analysis of Statistical and Syntatic Phrases," *Proc. of Computer-Assisted Information Searching on Internet*, pp.200-214, 1997.

[8] Myaeng, S. H., Li, M., "Building Term Clusters by Acquiring Lexical Semantics from a Corpus," *Proc. of International Conference of Information and Knowledge Management*, pp.130-137, 1992.

[9] Park, B. R., Yun, B. H., Rim., H. C., "Automatic Identification of Standard words Corresponding to Misspelled Words based on Contextual Simailarity," *Proc. of the Workshop on Information Retrieval with Asian Languages*, pp.116-121, 1998.

[10] Peat, H. J., Willett, P., "The limitation of term co-occurence data for query expansion in document retrieval system," *Journal of the American Society Information Science*, 42(5), pp.378-383, 1991.

[11] Qiu, Y., Frei, H. P., "Concept Based Query Expansion," Proc. of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.160-169, 1993.

[12] Ruge, G., "Experiments on Linguistically-based term associations," Information Processing & Management, 28(3), pp.317-332, 1992.

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

[14] Strzalkowski, T., Perez-Carballo, J., Marinescu, M., "Natural Language Information Retrieval : TREC-4 report," The Fourth Text REtrieval Conference, NIST SP 500-236, 1996.

[15] Yun, B. H., Cho, M. J., Rim, H. C., "A Korean Information Retrieval Model Alleviating Syntactic Term Mismatches," Proc. of Natural Language Processing Pacific Rim Symposium, pp.107-112, 1997.

[16] 김진동, 이상주, 임해창, "어절 띄어쓰기를 고려한 형태소 단위 품사 태깅 모델", 제10회 한글 및 한국어정보처리 학술발표 논문집, pp.3-8, 1998.

[17] 임희석, "어절의 중의성 유형 분류에 근거한 한국어 형태소 분석기", 고려대학교 석사학위 논문, 1993.

[18] 윤보현, 조민정, 임해창, "통계정보와 선호규칙을 이용한 한국어 복합 명사의 분해", 정보과학회논문지, 24권 8호, pp.900-909, 1997.

부 록

새로운 질의 및 확장된 질의

1) 새로운 질의
① 디지탈
② 시소러스
③ 스케쥴링
④ 멀티태스킹
⑤ 파라메터
⑥ 플라즈마
⑦ 휴렛팩커드
⑧ 트랜잭션
⑨ 프로시듀어
⑩ 도큐먼트

2) 확장된 질의
① 디지탈 or 디지틀 or 디지털
② 시소러스 or 디소러스 or 디소오러스
③ 스케쥴링 or 스케줄링 or 스케듈링
④ 멀티태스킹 or 멀티타스킹 or 멀티테스킹
⑤ 파라메터 or 파라메타 or 파라미타
⑥ 플라즈마 or 플래즈마
⑦ 휴렛팩커드 or 휴랫팩커드
⑧ 트랜잭션 or 트란잭션 or 트랜젝션
⑨ 프로시듀어 or 프로시쥬 or 프로시듀어
⑩ 도큐먼트 or 도큐멘트

윤 보 현

e-mail : ybh@etri.re.kr

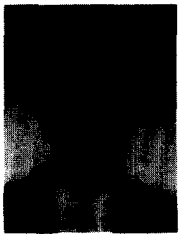
1992년 목포대학교 전산통계학과 (학사)

1995년 고려대학교 컴퓨터학과 (석사)

1999년 고려대학교 컴퓨터학과 (박사)

1999년~현재 한국전자통신연구원 선임연구원 언어이해연구팀

관심분야 : 정보 검색, 자연언어 처리, XML/SGML, 지식정보 처리



박 성 진

e-mail : sjpark@hucc.hanshin.ac.kr

1991년 고려대학교 전산학과 졸업(학사)

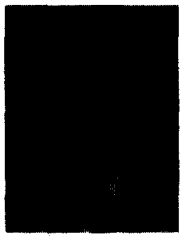
1993년 고려대학교 대학원 전산과학과(이학석사)

1998년 고려대학교 대학원 전산과학과(이학박사)

1998년~2000년 한국전자통신연구원 선임연구원

2000년~현재 한신대학교 정보시스템공학과 조교수

관심분야 : 웹데이타베이스, 데이터웨어하우스 및 데이타마이닝, 정보검색, 컴포넌트공학 등



강 현 규

e-mail : hkkang@etri.re.kr

1985년 홍익대학교 전자계산학과 (학사)

1987년 한국과학기술원 전산학과 (석사)

1992년 정보처리 기술사 자격 취득

1997년 한국과학기술원 전산학과 (박사)

1987년~현재 한국전자통신연구원 책임연구원 언어이해연구팀장

관심분야 : 정보 검색, 자연언어 처리, XML/SGML, 지식정보 처리