2002년 한국정보처리학회 춘계학술발표논문집 제9권 제1호

영어외래어의 음역어 자동변환을 이용한 검색 시스템

이미란*, 김양택*, 전홍태*, 윤성대**
*부경대학교 전산정보학과
**부경대학교 전자계산학과

e-mail:amuna@tit.ac.kr

A Retrieval System Using the Automatic Transition of the English-Adopted Words into Transliterations

Mi-Ran Lee*, Yang-Taek Kim*, Hong-Tee Jeun*, Sung-Dae Youn**
*Dept. of Computer and Information, Pukyong National University
**Dept. of Computer Science, Pukyong National University

요 약

정보 검색시 질의어가 외래어일 경우에 검색의 재현율은 급격하게 감소된다. 이는 외래어에서 나오는 음역어의 불일치와 영어외래어, 한글음역어는 같은 색인으로 처리가 되지 않기 때문이다. 따라서 본 논문에서는 영어외래어를 한글음역어로 자동 변환시키고, 자동 변환시에는 영어음소에 해당하는 발음값을 한글음소로 모두 변환시킨 다음 조합하였다. 조합된 음역어들은 다시 동치부류 DB에 저장되어, 질의어 검색시 검색어가 동치부류 색인어로 확장되어 검색된다. 제안한 검색시스템의 성능을 평가하기 위해서 재현율을 측정하였다.

1. 서론

정보검색 시스템에서는 무엇보다도 중요한 것이 재현율과 정확율이라 할 수 있다. 검색자가 원하는 정보는 모두 검색되어서 나타내야 하고, 관련 없는 정보는 검색 되어서는 안된다[1].

일반적으로 정보를 검색할 때는 키워드를 사용하게 되는데 컴퓨터 분야 등에서는 프로그램언어 등 외래어 질의어가 많이 나타나게 된다. 이때 외래어는 영어 외에 한국어로 음역된 질의어로도 검색하는 경우가 많다. 예를 들어 자바의 경우 검색자는 검색어를 'Java'로 입력할 수도 있고, 한글 음역어인 '자바'로도 입력 할 가능성이 있다.

기존의 검색 시스템은 입력된 값 그대로 색인되고 일치하는 문서를 찾기 때문에 두 질의어는 같은 결과값를 갖지 못한다. 도서관 등에서 쓰이는 대개의 검색 시스템은 영어 색인어를 임의로 한글음역어로 바꾸고 저장하여 색인처리 하고있다. 그러나 영어는 한글음역어로 바꾸는 대신 한글 음역어는 영어로 바꾸지 않기 때문에 검색 후의 결과값이 여전히 차이가 있다.

본 논문에서는 이러한 문제를 해결하기 위해 영어 외래어를 한글음역어로 자동 변환하는 시스템을 제안하였고, 사용자 모드와 DB관리자 모드로 나누어서 컴퓨터 관련 도서명을 저장하고 검색하는 프로그램을 구현하였다. 사용자 모드에서는 도서명을 검색하여 결과값을 제공하고 DB관리자 모드에서는 서명을 입력하여 DB를 구축하고 동치부류DB를 관리하게 된다.

시스템은 영어 외래어가 입력될 경우 영어음소에 대응하는 한글음소를 조합하여 다수의 음역어를 생성하고, 이는 다시 동치부류 DB에 영어 외래어와 같이 묶여서 저장되고 색인된다.

본 논문의 구성은 2장에서 한국어의 용어 불일치성 해결을 위한 관련연구에 대해 다루고, 3장에서는 본 논문에서 제안하는 시스템의 구성과 처리과정 등에 대해 설명한다. 그리고 4장에서는 검색 시스템을 구현하여 재현율을 측정하여 성능을 평가하였다.

2. 관련연구

영어 외래어를 한글로 음역하기 위한 시스템으로

영어 발음기호를 이용한 외래어 자동표기 시스템이 있다[2]. 이는 영어의 경우 발음기호 전사방식언어로 분류되어 발음기호에 해당하는 한글음소를 조합하여 외래어 및 외국 인명, 지명 등을 외래어 표기 규정에 맞게 표기하였다. 이것은 영어외래어에 대한 현지음을 알아내고 이러한 발음에 근거하여 표기규정에 맞는 음역어를 표기하는 형식이다. 현재 사용되고 있는 음역어는 사람에 따라 다양하게 나타내기 때문에 정보 검색 시스템에서는 자연히 재현율이 떨어지고 영어발음사전이 반드시 필요하다는 단점이 있다.

한글 음역어를 자동 변환하는 방법으로는 STM (Statistical Transliteration Model)을 이용하는 방법이 있다[3]. 이는 음역문제를 주어진 단어에 대해 가장 적절한 음역어를 대응하는 것으로 영어 단어 E에 대한 한국어 단어 K가 대응 될 확률을 $P(K|E)$, 한국어 단어 K에 대한 음역 확률을 $P(K)$로 하면 $P(K|E)$는 식1과 같이 유도된다.

$$argmax_K p(K|E) = argmax_K \frac{p(K)P(K|E)}{P(E)} \quad (식1)$$

동치부류를 구축하는 방법으로 음역어 외에 철자오류도 포함하는 경우가 있다[4]. 음역어나 철자오류의 여부는 형태소 분석 결과를 보고 판단하고, 철자오류 교정 분야에서 올바른 단어후보를 생성하는 과정은 철자오류 패턴의 지식을 표현하는 휴리스틱 규칙에 기초하였다. 철자오류와 표준어 후보간 문맥 유사도를 이용하여 하나의 표준어를 선택하고, 철자오류와 표준어를 동치부류로 구축하는 방법이다.

3. 영어외래어의 음역어 자동 변환을 이용한 검색 시스템

본 논문에서는 검색의 재현율과 정확율을 높이기 위해 동치부류를 사용하였다. 동치부류에 의해 같은 의미를 가진 색인어는 색인어 생성시 같은 그룹으로 생성된다. 동치부류를 구성하는데 있어 동의어, 철자오류, 음역어 등으로 구성될 수 있으나 본 논문에서는 음역어의 동치부류를 구현하였다.

음역어는 영어외래어의 경우에 영어 발음을 한글로 표현하는 것으로, 사용자에 따라 여러 가지 다른 음역어가 사용된다. 영어외래어가 주어지면 이를 다수의 한글 음역어로 자동 변환하여, 다수의 음역어와 영어외래어는 동치부류 DB에 구축된다.

전반적인 시스템의 구성과 동치부류 DB 구축에 대해서는 다음절에서 자세히 기술한다.

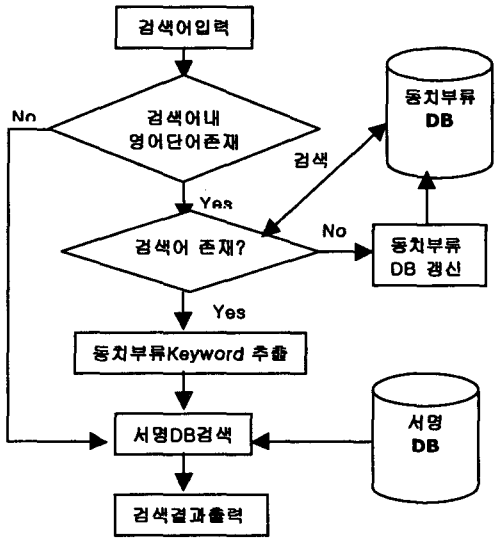
3.1. 시스템의 구성

본 논문에서 구현한 동치부류 시스템의 인터페이스는 시스템관리자 모드, 사용자 모드, DB관리자 모드 세부분으로 구성된다.

먼저 시스템관리자 모드는 ID와 Password를 검증하여 일반이용자는 사용자 모드로, DB를 관리 할 수 있는 권한이 주어진 ID는 DB관리자 모드로 접속할 수 있게 한다.

사용자 모드에서는 검색자가 원하는 질의어를 입력하고 검색결과를 나타내며, "Java"로 입력한 결과와 "자바"로 입력한 결과값은 같다.

사용자 모드에서 검색어 입력시 처리되는 대략적인 과정은 그림 1에 나타내었다.

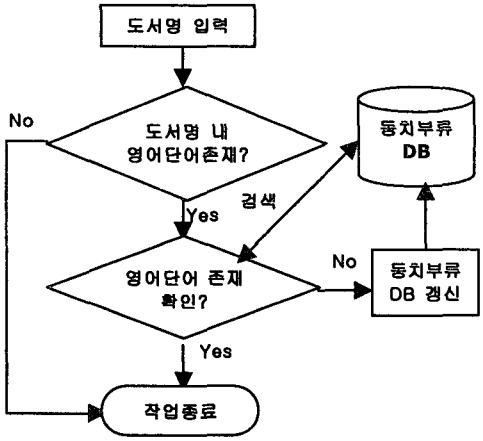


(그림 1) 사용자 모드 개략 순서도

그림 1에서 검색어를 입력하면 검색어 내에 영어단어가 존재하는지 확인하고 존재시에는 동치부류 DB를 확인한다. 동치 부류 DB에 해당하는 영어단어가 존재하면 서명 DB를 바로 검색하게 된다. 즉, 검색어에 대한 색인어들이 동치부류 DB에 이미 존재하므로 색인어들을 추출하고, 추출한 모든 색인어를 이용하여 서명 DB를 검색한 후 도서명을 출력한다. 반대로, 검색어가 동치부류 DB에 존재하지 않을 경우에는 영어단어를 각각의 영어음소에 해당하는 한글음소를 조합하여 동치부류 음역어를 생성하여 동치부류 DB를 갱신한 후, 새로 생성된 색인어를 이용하여 서명 DB를 통해 원하는 도서를 검색한다.

DB관리자 모드에서는 관리자가 서명을 입력하여 서명 DB에 저장을 한다. 그리고 동치부류 DB의 내용을 화면에 표시 및 검색을 할 수 있고 필요 없는 색인어는 삭제 및 커밋(commit)이 가능하도록 설계한다.

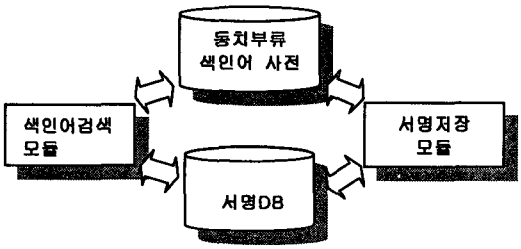
DB관리자 모드에서 도서명 입력시 처리되는 대략적인 과정은 그림 2와 같다.



(그림 2) DB관리자 모드 개략 순서도

그림 2와 같이 우선 도서명을 입력하면 도서명 내에 영어 단어를 분류하여 동치부류 DB에 존재하는지 확인한다. 영어 단어가 존재하지 않으면 바로 서명을 저장하게 되지만 존재할 경우 검색어 입력과 마찬가지로 영어 단어가 동치부류 DB에 존재하는지 확인하고 존재하지 않으면 한글음역어와 함께 동치부류 DB에 저장하고 색인 처리된다.

그림 3은 전체적인 시스템의 구성으로 각각의 색인어 검색 모듈과 서명저장 모듈이 동치부류 색인어 사전과 서명DB를 참조하는 관계를 나타내었다.



(그림 3) 시스템 구성

본 시스템에서 사용되는 주요 데이터 사전인 서명 DB와 동치부류 DB의 테이블 구조를 표 1에 나타내었다.

<표1> 서명DB와 동치부류 DB 테이블 구조

Book_table

Book_id	Num(일련번호)
Book_name	Char

Equ_class_table

Eng_id	Num(일련번호)
Eng_word	Char
Trans_word	Char

표1에서 Book_table은 서명 DB로서 Book_id는 서적에 대한 일련번호 이고 Book_name은 서명을 나타내는 필드이다. 동치부류 DB는 Equ_calss_table로 구현되며 Eng_id는 영어단어에 대한 일련번호이고 Eng_word는 영어 외래어 단어이며 Trans_word는 동치부류에 의해 생성되는 한글 음역어가 된다. 여기에서 Eng_word는 Trans_word에 대해 1:다 관계를 갖는다.

3.2 음역어 자동변환을 이용한 동치부류 DB 구축

영어외래어가 입력되면 한글음역어로 변환시켜주고 다수의 한글 음역어들을 동치부류로 저장하여 서명 검색시 모두 검색어로 사용한다.

동치부류를 이용하여 한글 음역어를 생성하는 과정을 자세히 기술하면, 먼저 영어외래어는 영어음소로 나누어지고 각각의 영어 음소가 발음될 수 있는 다수의 한글음소로 대응된다. 표 2는 영어 음소 a부터 z에 해당하는 한글음소를 나타내고 있다

<표2> 영어음소에 대한 한글음소 변환

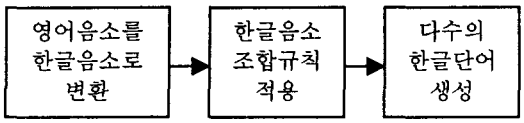
구분	영어 음소	한글음소			
		변환1	변환2	변환3	변환4
모음	A	ㅏ	ㅓ	ㅐㅣ	ㅐ
	E	ㅔ	ㅓ	ㅡ	
	I	ㅣ	ㅏㅣ		
	O	ㅗ			
	U	ㅜ	ㅠ		
자음	B	ㅂ			
	C	ㅋ	ㅅ		
	...				
	Z	ㅈ			
조합	Ea	ㅣ			
	Ng	ㅇ			
	ion	ㅕㄴ			

표 1을 이용하여, 입력된 영어 외래어에 대해 한글 음역어들을 생성하는 예제는 표 3이다. 표 3과 같이 각각 해당하는 한글 음소는 다시 한글음소 조합규칙을 적용하여 조합되며 한글음소 조합규칙을 적용하여 동치부류 색인어들이 생성 된다.

<표3> 동치부류 색인어 DB

입력된 영어외래어	동치부류 색인어
Java	Java, 자바, 제이버, 자베이, 저버...
Programming	Prgramming, 프로그래밍, 프라그라밍, 프레이그레이밍...
Flash	Flash, 플라쉬, 플러시, 플래시, 플레이시...

이상의 동치부류 색인어들을 생성하는 과정을 도식화 하면 그림 4와 같다.



(그림 4) 영어음소의 한글음소 변환과정

영어외래어는 처음 외래어가 들어 올때 여러 가지 음역어로 바뀌어서 사용된다. 우리나라에는 외국어 표기 규정이 있지만, 일반적으로 도서명이라든지 웹 등에서 임의로 음역어를 사용하는 경우가 많다.. 예를 들어 "digital"의 경우 "디지탈", "디지틀", "디지털"등 다양한 음역어로 표기하고 사용한다. 이런 음역어들을 모두 색인어로 처리되어 검색되기 위해서는 동치뷰류를 구축하고 같이 색인처리하는 것이 효과적이다.

결국 사용자가 "디지털"로 검색했을 경우에 "digital", "디지탈", "디지틀", "디지털"등 동치부류에 구축된 단어들이 포함된 서명은 모두 검색되어 진다. 이를 통해 존재하는 서명에서 색인어들을 통해 검색되는 빈도인 재현율과 원하는 서명을 검색하는 정확율을 높일 수 있다.

4. 시스템 구현 및 결과

본 논문에서 구현한 검색 시스템은 Java 2(SDK 1.3.1)을 이용하여 응용 프로그램을 구현하였으며 DBMS는 Oracle 8x를 사용하였다. 그리고 프로그램과 DB 의 연결을 위해 JDBC Connection Type IV(Thin driver : classes111.zip)을 사용하였다.

사용되는 검색어는 컴퓨터 분야의 도서 중 검색율이 높은 영어 외래어 20개를 선정하여 검색하였으며 검색어의 음역어가 포함된 도서명을 저장하였다.

실험을 통한 기존 검색법과 동치부류 DB검색법에 대해 검색된 도서수와 재현율의 결과를 기술하였다.

<표4> 기존검색법과 동치부류DB검색법 결과비교

검색어	기존검색		동치부류DB검색	
	검색된 도서수	재현율	검색된 도서수	재현율
Digital	24	38%	45	82%
Java	22	48%	44	94%
Flash	18	62%	38	98%
Network	25	46%	44	92%
Database	26	35%	52	98%

표 4에서 결과를 보면, 'network'나 'database'와 같은 경우에 서명에서 한국어 단어 '데이터베이스' 혹은 '데이타베이스'로 사용함으로써 재현율에 큰 차이가 있음을 알 수 있다. 그러므로 동치부류DB는 임의의 영어단어에 대해 번역 가능한 한글음역어를 색인어로 대부분 포함하고 있기 때문에, 도서명의 재현율이 증가 되었다.

5. 결론 및 향후연구과제

도서명 검색시 필요한 도서의 재현율을 높이기 위해 검색어의 영어 외래어를 음역어로 자동 변환하고 변환된 색인어들을 동치부류로 묶어서 같이 검색하였다. 음역어 자동 변환 시에는 각각의 영어음소에 해당하는 한글음소로 변환 후 한글음소 조합규칙을 적용하여 다수의 음역어를 생성하였다.

실험을 통한 검색 결과로는 검색어가 포함된 도서명만 검색되는 것보다 제안한 방법이 기존의 방법을 포함하였기 때문에 월등히 재현율이 높았다.

향후 연구과제로는 검색 속도 향상을 위해 동치부류에 포함된 음역어 가운데 검색어로서의 일정시간이상 사용되지 않는 음역어를 자동으로 색인DB에서 삭제 시키는 것과 한글음소조합규칙을 최대한으로 적용시켜 영어 외래어 외에 일반적으로 음역어로 자주 사용하는 영어외국어도 변환할 수 있는 시스템을 구축하는 것이다.

참고문헌

[1] 한국어정보처리연구소, "정보검색시스템", 골드, 1996

[2] 문화관광부 국어정책과, " 외래어 및 한글 자동 표기 시스템(최종연구보고서)" 1999.

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

[4] 윤보현, 박성진, 강현규, "한국어 정보 검색에서 의미적 용어 불일치 완화 방안", 한국정보처리학회 논문지, 제 7 권 제 12 호, pp.3874—3883,2000.