문자 인식에 의해 구축된 한글 문서 데이터베이스에 대한 정보 검색

이 준 호[†] · 이 충 식[††] · 한 선 화[†††] · 김 진 형[††††]

요 약

문자 인식에 의해 구축된 문서들은 키보드 입력에 의해 구축된 문서들에 비하여 다수의 오류를 포함한다. 따라서 이러한 문서들로부터 원하는 정보를 검색하기 위해서는 다수의 오류를 포함하고 있는 문서들에 대한 효과적인 자동 색인 방법이 요구된다. 본 연구에서는 개별 문자 인식률 90% 수준의 문자 인식기에 의해 구축된 한글 문서 데이터베이스로부터 원하는 정보를 효과적으로 검색하기 위한 자동 색인 방법에 대하여 살펴본다. 실험 결과는 문자 인식에 의해 구축된 한글 문서 데이터베이스에 대해서는 형태소 단위 색인법과 2-gram 기반 색인법이 유사한 수준의 검색 효과를 제공함을 보여준다.

Retrieving Information from Korean OCR Text Database

Joon-Ho Lee[†] · Chung-Sik Lee[††] · Sun-Hwa Hahn[†††] · Jin-Hyung Kim[††††]

ABSTRACT

The texts constructed with Optical Character Recognition(OCR) contain more errors than those constructed with keyboard typing. Therefore, in order to retrieve useful information from OCR texts, we need to develop an effective automatic indexing method. In this paper, we investigate automatic indexing methods that can retrieve information effectively from Korean OCR text database with the character-level recognition ratio of 90%. Experimental result shows that 2-gram indexing provides similar retrieval effectiveness to morpheme-based indexing for the Korean OCR text database.

1. 서 론

지난 30년 동안 과학과 기술 분야의 급속한 발전은 수많은 주제들에 대해 방대한 양의 정보가 생성되는 정보화 사회를 탄생시켰으며, 원하는 정보에 대한 정확하고 빠른 접근은 정보화 사회를 살아가는 현대인들에게 성공의 여부를 결정짓는 중요한 요소가 되었다. 이러한 정보화 사회에서 대용량의 데이터로부터 원하는 정보의 발견을 도와주는 정보 검색 시스템의 중요성이 널리 인식되고 있다[10].

정보 검색 시스템을 통하여 유용한 정보들을 제공하기 위해서는 수 백년의 역사를 통해 생성되어 왔고, 현재도 활발히 생성되고 있는 활자화된 문서들을 전산화하여 데이터베이스를 구축해야 한다. 현재까지는 많은 경우에 키보드 입력과 같은 수작업을 통하여 활자화된 문서들의 전산화 작업을 수행하고 있으나, 이러한 방식은 너무도 비효율적이고 처리할 수 있는 문서의 양에도 한계가 있다. 따라서 활자화된 문서들을 전

† 정 회 원 : 숭실대학교 컴퓨터학부 교수
†† 비 회 원 : 한국과학기술원 전산학과
††† 정 회 원 : 연구개발정보센터 선임연구원
†††† 정 회 원 : 한국과학기술원 전산학과 교수

논문접수 : 1998년 9월 4일, 심사완료 : 1999년 2월 6일

산화하여 데이터베이스를 구축할 수 있는 효율적인 방법이 요구된다.

문자 인식은 활자화되어 있거나 손으로 작성된 문서를 인식하여 컴퓨터에서 사용할 수 있는 내부 표현 방식으로 변환시켜 주는 분야이다. 문자 인식 기술은 스캐너를 이용하여 입력된 문서 영상에 포함된 문자들을 자동으로 인식하는 수단을 제공함으로써, 방대한 양의 문서들을 효율적으로 전산화할 수 있는 수단을 제공한다. 그러나 문자 인식에 의해 구축된 문서 데이터베이스는 키보드 입력에 의해 구축된 데이터베이스에 비하여 많은 오류를 포함한다. 따라서 문자 인식에 의해 구축된 문서 데이터베이스로부터 원하는 정보를 검색하기 위해서는 다수의 오류를 포함하고 있는 문서들에 대한 효과적인 검색 방법의 개발이 요구된다.

정보 검색 시스템의 중요한 역할 중의 하나는 검색된 각각의 문서에 대하여 순위 결정 방법(Ranking)을 적용하는 것이다. 문서 순위 결정 방법은 문서와 질의 사이의 관련 정도를 나타내는 유사도(Similarity)를 계산하고, 계산된 유사도에 따라 문서에 순위를 부여한다. 높은 순위를 갖는 문서일수록 질의에 대한 만족도가 크며, 사용자는 높은 순위를 갖는 문서를 우선적으로 검토함으로써 필요한 정보를 얻는데 소모되는 시간을 최소화할 수 있다[4].

정보 검색 모델들 중의 하나인 벡터 공간 모델은 문서와 질의를 가중치가 부여된 색인어들의 벡터로 표현하고, 문서 벡터와 질의 벡터의 내적(Inner Product)으로 문서와 질의 사이의 유사도를 계산한다[7,11]. 일반적으로 문서나 질의의 내용을 표현하는 색인어들은 이를 추출하는 색인 방법에 따라 달라지므로, 벡터 공간 모델에서 계산되는 유사도의 질은 사용되는 색인 방법에 의해 크게 영향을 받는다.

한글 정보 검색에서 개발된 효과적인 색인 방법으로는 형태소 단위 색인법[12,18,19,20,22]과 n-gram 기반 색인법[6,17]이 있다. 본 연구에서는 개별 문자 인식률 90% 수준의 문자 인식기에 의해 구축된 한글 문서 데이터베이스에 대해 형태소 단위 색인법과 n-gram 기반 색인법을 적용시켜 벡터 공간 모델을 기반으로 문서들을 검색함으로써, 문자 인식에 의해 구축된 한글 문서 데이터베이스를 통한 정보 검색의 타당성을 검토하고, 기존의 색인 방법 중 어느 방법이 문자 인식에 의해 구축된 한글 문서 데이터베이스에 적합한지를 알아보고자 한다.

2. 한글 문서에 대한 자동 색인 방법

전통적인 색인 작업은 훈련된 사서나 주제 전문가에 의해 수작업으로 수행되어 왔다. 그러나 수작업에 의한 색인은 많은 시간과 비용을 필요로 할 뿐 아니라, 색인자의 주관에 따라 색인어의 양이나 질이 달라지는 일관성 결여의 문제점을 지니고 있다. 이러한 문제점을 극복하기 위해 컴퓨터를 이용하여 문서를 분석하여 색인어를 추출하는 자동 색인 방법들이 정보 검색 분야에서 개발되어 왔다[1,9].

한글 문서의 자동 색인을 위한 가장 오래된 방법은 어절 단위 색인법이다. 어절 단위 색인법은 문서나 질의에서 불용어를 제외한 모든 어절들을 색인어 후보로 간주하고, 각 어절로부터 색인어의 부분으로써 무의미한 조사, 어미, 접미사와 같은 비색인 분절(Non-indexable Segment)을 제거한 나머지 색인 분절(Indexable Segment)을 색인어로 선택한다[14-16]. 그러나 어절 단위 색인법은 복합 명사를 단순 명사들로 분리하지 못하기 때문에, 문서가 많은 수의 복합 명사들을 포함하고 있을 경우에 검색 효과가 저하되는 경향이 있다. 이러한 문제를 해결하기 위해 다음에서 설명되는 형태소 단위 색인법과 n-gram 기반 색인법이 개발되었다.

2.1 형태소 단위 색인법

형태소 단위 색인법은 문장 분석의 정도에 따라 형태소 해석만을 이용하는 방법과 구문 해석을 통한 방법이 있다. 형태소 해석만을 이용하는 방법은 문장의 모든 어절들에 대해 형태소 해석을 수행하여 단순 명사들을 색인어로 선정한다[12,18]. 구문 해석을 통한 방법은 문장 단위의 구문 해석을 수행하여 문장에서 중요한 의미를 갖는 특정한 명사나 명사구를 색인어로 선정한다[19,20,22].

형태소 분석에 의한 색인은 복합 명사의 띄어쓰기를 잘 처리할 수 있으며, 검색 효과에 있어서도 좋은 결과를 보여준다. 그러나 형태소 해석 및 구문 분석을 위해 형태소 사전이나 격틀 사전과 같은 많은 언어 정보를 필요로 하는 문제점을 지니고 있다. 특히, 형태소 사전의 경우 개발에 많은 시간과 비용이 요구되며, 대상 문서에 따라 사전도 달라져야 한다. 또한, 사전에 등록되지 않은 미등록어가 문서 내에 존재할 경우 검색 효과의 저하를 초래할 수 있다. 특히 과학 기술 분야에서는 많은 전문 용어들이 미등록되어 있는 경우가 많아 형태소 분석 방식의 한계로 지목되고 있다.

2.2 N-gram 기반 색인법

N-gram 기반 색인법[6,17]은 어절 단위 색인법[14-16]에 의해 추출된 각 색인 분절에 대해 *n*-gram 방법[2,3]을 접목시킨 색인 방법이다. *N*-gram이란 인접한 *n*개의 음절을 말한다. 예를 들면, '프로그래밍'이란 단어에 대해 2-gram은 '프로', '로그', '그래', '래밍'이며, 3-gram은 '프로그', '로그래', '그래밍'이다. *N*-gram 기반 색인법은 어절 단위 색인법에 의해 추출된 색인어의 음절 수가 *n*보다 작은 경우에는 추출된 색인 분절 전체를 색인어로 선정하고, 큰 경우에는 추출된 색인 분절의 *n*-gram들을 색인어로 선정한다.

N-gram 기반 색인법은 형태소 해석을 수행하지 않기 때문에 형태소 단위 색인법에서와 같은 복잡한 문장 해석 규칙과 언어 정보의 개발을 요구하지 않으면서도, 어절 단위 색인법에서의 복합 명사 띄어쓰기 문제를 완화함으로써 단순 명사를 추출할 수 있는 형태소 단위 색인법과 유사한 수준의 검색 효과를 제공한다. 또한 철자 오류나 일관성이 없는 외래어 표기 문제를 적절히 극복할 수 있다. 예를 들면, 문서가 '정보검섹'이라고 철자가 틀린 어절을 포함하고 있고, 사용자는 '정보검색'을 질의로 입력하였다고 가정하자. 2-gram 기반의 색인법은 이들 문서와 질의에 대해 각각 {'정보', '보검', '검섹'}, {'정보', '보검', '검색'}의 벡터 표현을 생성한다. 따라서 문서에 '검섹'이라는 철자 오류가 있더라도 2개의 색인어들이 일치하므로, 문서는 질의에 관련된 문서로서 검색될 수 있다.

3. 문자 인식에 의한 데이터베이스 구축

3.1 데이터베이스 구축 환경

본 논문에서는 한글 정보 검색의 연구를 위해 널리 사용되고 있는 KT 테스트 컬렉션[13]에 포함된 논문들을 문자 인식기를 이용하여 데이터베이스로 구축하였다. KT 테스트 컬렉션은 정보과학회논문지, 한국정보과학회 학술발표대회논문집, 정보관리학회지에 수록된 1000편의 논문들에 대해 저자, 발행년도, 국문 및 영문 초록 등을 수작업으로 입력하여 구성한 작은 규모의 데이터베이스로, 30개의 질의와 각각의 질의에 적합한 문서들의 리스트를 포함하고 있다.

실험을 위해 KT 테스트 컬렉션에 포함된1000편의 문서들 중에서 정보관리학회지 2권 2호에 수록된 논문 6편과 11권 1호에 수록된 한편의 논문을 제외한 993편의 논문을 스캔하였다. 스캔된 논문들 중에서 '한글 요약이 없는 논문'과 '한글 요약에 한자가 섞여 있는 논문' 23편은 문자 인식 과정에서 제외되었으며, 따라서 총 970편의 논문으로 데이터베이스를 구축하였다. 문자 인식기로는 상용 문자 인식기 중에서 우수한 문자 인식 결과를 보여준 Speed Reader 1.2를 사용하였으며, 문자 인식에 의해 구축된 데이터베이스의 평균 개별 문자 인식률은 90.54%였다[21].

3.2 구축된 데이터베이스의 특성

문자 인식에 의해 구축된 데이터베이스는 키보드 입력에 의해 구축된 데이터베이스에서 나타나는 유형과는 종류가 다른 오류들을 포함한다. 키보드 입력에 의해 구축된 데이터베이스에서 나타나는 오류는 키보드를 잘 못 쳐서 발생하는 오타 오류와 본문의 단어를 잘 못 읽어서 발생하는 오류가 대종을 이룬다. 반면에 문자 인식에 의해 구축된 데이터베이스에서 나타나는 오류는 빈칸이 없는 곳에 빈칸을 집어넣거나, 하나의 문자를 둘로 나누어 인식하는 등 사람이 생각치 못하는 유형의 오류들이 자주 발생한다. 문자 인식 기술을 사용하여 데이터베이스를 구축하는 과정에서 발생한 오류의 유형은 다음과 같다.

- 하나의 문자가 다른 하나의 문자로 인식되는 오류
- 하나의 문자가 인식되지 않고 없어지는 경우
- 없는 문자가 나타나는 경우
- 연속된 여러 문자가 다른 하나의 문자로 인식되는 경우
- 하나의 문자가 연속된 여러 문자로 인식되는 경우
- 공백 문자가 추가되는 경우

이들 오류 유형들 중에서 특히 검색 성능에 심각한 영향을 미치는 것은 공백 문자가 추가되는 오류이다. 공백 문자가 추가됨으로써 문장의 성분이 달라지고, 전혀 엉뚱한 색인 어절이 생성될 수 있기 때문이다.

4. 문자 인식에 의해 구축된 데이터베이스에 대한 검색 실험

본 연구에서는 다음의 두 가지 문제에 대한 답을 구하기 위해 정보 검색 실험을 수행하였다. 첫째, 문자 인식에 의해 구축된 데이터베이스로부터 원하는

문서에 대한 검색을 효과적으로 수행할 수 있는가? 둘째, 문자 인식에 의해 구축된 데이터베이스에 대한 자동 색인 방법으로는 어떠한 방법이 효과적인가? 이에 대한 해답을 얻기 위해 문자 인식에 의해 구축된 데이터베이스와 키보드 입력에 의해 구축된 데이터베이스에 대해 형태소 단위 색인법과 2-gram 기반 색인법을 적용하여 자동 색인을 수행하고, KT 테스트 컬렉션에 포함된 30개의 질의에 대한 검색 효과를 측정하였다.

4.1 검색 효과 측정 방법

정보 검색 시스템의 검색 효과는 일반적으로 재현율과 정확률로서 평가된다[8]. 재현율(Recall)은 문서 집합에서 사용자가 원하는 문서를 어느 정도 검색하였는가를 나타내며, 정확률(Precision)은 검색된 문서들 중에서 사용자가 원하는 문서가 얼마나 포함되어 있는가를 나타낸다. 예를 들어, 전체 문서 집합에 200개의 문서가 저장되어 있고, 이 문서 집합 속에 사용자가 입력한 질의에 적합한 문서가 5개 있다고 가정하자. 이때 사용자가 검색 시스템을 사용하여 6개의 문서를 검색하였고 검색된 문서 중에서 4개의 문서가 질의에 적합한 문서라고 하면, 재현율과 정확률은 각각 0.8과 0.67이 된다. 문서 순위 결정 방법을 제공하는 검색 시스템은 보간 기법을 사용하여 고정된 재현율에 대한 정확률을 계산할 수 있는데, 본 연구에서는 고정된 11개의 재현율 (0.0, 0.1, ..., 1.0)에 대한 모든 질의의 정확률을 평균한 값을 나타내는 11-포인트 평균 정확률을 이용하여 검색 효과를 측정하였다.

4.2 SMART 시스템

본 논문에서는 코넬 대학에서 38년 간에 걸쳐 개발된 SMART 시스템을 이용하여 정보 검색 실험을 수행하였다. SMART 시스템은 벡터 공간 모델을 기반으로 하며, 문서와 질의 모두 다음과 같은 벡터로 표현된다[7,11].

$$di = (w_{i1}, w_{i2}, \cdots, w_{in})$$

여기에서 d_i는 문서 또는 질의를 표현하고, w_{ik}는 문서 d_i에서 색인어 t_k의 가중치이다. SMART 시스템에서 이러한 벡터들은 색인 방법을 적용하여 생성된 색인어들에 가중치를 부여함으로써 생성되며, 문서 또는 질의에 나타나지 않는 색인어들에 대해서는 가중치 0이 할당된다.

문서 또는 질의에 대한 벡터들이 형성된 이후의 검색 과정은 벡터들의 연산에 의해 이루어진다. 문서 d가 $(w_{d1}, w_{d2}, \cdots, w_{dn})$로 표현되고, 질의 q가 $(w_{q1}, w_{q2}, \cdots, w_{qn})$로 표현되었을 때, 문서 d와 질의 q 사이의 유사도는 다음과 같이 두 벡터들의 내적으로 계산된다.

$$Sim(d, q) = \sum_{i=1}^{n} (w_{di} \times w_{qi})$$

위의 식에서 알 수 있듯이 문서와 질의 사이의 유사도는 색인어들의 가중치에 의해 결정되기 때문에 가중치 부여 기법은 검색 효과에 영향을 미치는 중요한 요소이다[5]. 본 연구에서는 우수한 검색 효과를 제공하는 것으로 알려진 다음과 같은 가중치 기법을 문서와 질의에 대하여 사용하였다.

$$w_{ik} = \frac{\left(0.5 + 0.5\frac{tf_{ik}}{\max if}\right) \times \ln\frac{N}{n_k}}{\sqrt{\sum_{j=1}^{n}\left[\left(0.5 + 0.5\frac{tf_{ij}}{\max if}\right) \times \ln\frac{N}{n_k}\right]^2}}$$

여기에서 w_{ik}는 i번째 문서의 k번째 색인어의 가중치이고, tf_{ik}는 i번째 문서에서 k번째 색인어가 출현하는 빈도수이며, n_k는 전체 문서 집합에서 k번째 색인어를 포함하고 있는 문서수이다.

4.3 검색 실험 결과

<표 1>과 <표 2>는 각각 키보드 입력과 문자 인식에 의해 구축된 데이터베이스에 대하여 '인공지능 문자인식'이라는 질의를 수행한 결과를 보여준다. 키보드 입력에 의해 구축된 데이터베이스에 대하여 '인공지능 문자인식'이라는 질의를 수행한 결과, 541번 문서는 질의에 대한 유사도가 두 번째로 높은 것으로 검색되었다. 541번 문서에 대한 문자 인식률은 98.04%이었으며, 문자 인식에 의해 구축된 데이터베이스로부터는 질의에 대한 유사도가 가장 높은 문서로 검색되었다. 또한, 623번과 854번 문서의 인식률은 각각 76.95%와 57.89 %로서 문자 인식 결과가 좋지 않았으나, 키보드 입력과 문자 인식에 의해 구축된 데이터베이스로부터 유사한 순위로 검색되었다. 그러나, 인식률이 69.25%인 544번 문서의 경우, 키보드 입력에 의해 구축된 데이터베이스로부터의 검색시에는 상위 순위 15개 문서에 포함되었으나, 문자 인식에 의해 구축된 데이터베이스로부터의 검색시에는 상위 순위15개

문서에 포함되지 않았다. 이는 문자 인식 오류가 발생한 문자에 따라서 검색에 중요한 단서를 잃어버릴 수 있음을 보여준다.

〈표 1〉 키보드 입력에 의해 구축된 데이터베이스에 대한 검색 결과
〈Table 1〉 Retrieval result from keyboard input database

문서번호	유사도	문서 제목
623	0.21	분산인공지능 시스템을 위한 Speech-act 에이전트
541	0.20	문자인식 시스템의 올바른 reject점 설정 방법
404	0.16	자율형 이동로보트의 IN-DOOR NAVIGATION
854	0.15	불변 특징을 이용한 숫자인식
154	0.14	인쇄체 한글문자의 인식을 위한 계층적 신경망
544	0.14	효과적인 한글 문서 판독시스템 구현
641	0.13	개선된 HMM을 이용한 화자인식 시스템의 성능 향상
383	0.13	적응학습을 이용한 온라인 한글 문자인식 시스템의 구현
934	0.13	문헌정보학 영역 지식기반시스템에서의 지식표현
75	0.13	야외에 위치한 인공물을 인식하기 위한 지식기반형 방법
192	0.13	적응학습법에 의한 문자집합별 온라인 인식
344	0.13	인쇄체 한글 문자인식을 위한 특징성능의 비교
150	0.12	문서 영상에서 문자와 비문자의 분리추출방법
100	0.12	한글에 적합한 획 해석에 의한 연속 필기 한글의 On-line인식
547	0.11	생물학적 뉴런 구조를 가진 개선된 퍼지 퍼셉트론 알고리즘

〈표 2〉 문자인식에 의해 구축된 데이터베이스에 대한 검색 결과
〈Table 2〉 Retrieval result from OCR database

문서번호	유사도	문서 제목
541	0.20	문자인식 시스템의 올바른 reject점 설정 방법
404	0.17	자율형 이동로보트의 IN-DOOR NAVIGATION
623	0.17	분산인공지능 시스템을 위한 Speech-act 에이전트
854	0.16	불변 특징을 이용한 숫자인식
383	0.14	적응학습을 이용한 온라인 한글 문자인식 시스템의 구현
154	0.13	인쇄체 한글문자의 인식을 위한 계층적 신경망
641	0.13	개선된 HMM을 이용한 화자인식 시스템의 성능 향상
934	0.13	문헌정보학 영역 지식기반시스템에서의 지식표현
75	0.13	야외에 위치한 인공물을 인식하기 위한 지식기반형 방법
344	0.13	인쇄체 한글 문자인식을 위한 특징성능의 비교
192	0.12	적응학습법에 의한 문자집합별 온라인 인식
100	0.12	한글에 적합한 획 해석에 의한 연속 필기 한글의 On-line 인식
150	0.12	문서 영상에서 문자와 비문자의 분리추출방법
948	0.11	구문 및 어의분석을 통한 한국어 자동색인
853	0.11	훈련된 규칙인식과 설명을 위한 FCES의 설계

키보드 입력과 문자 인식에 의해 구축된 데이터베이스에 대한 검색 효과를 정량적으로 측정하기 위해, 형태소 단위 색인법과 2-gram 기반 색인법을 적용하여 자동 색인을 수행하고, KT 테스트 컬렉션에 포함된 30개의 질의에 대한 검색 효과를 측정하였다. <표 3>은 실험 결과를 보여주며, 다음과 같이 분석될 수 있다.

〈표 3〉 문자인식 데이터베이스에 대한 한글 색인 방법의 검색 효과
〈Table 3〉 Retrieval effectiveness of Korean indexing methods for OCR database

	형태소 단위 색인 (키보드입력)	2-gram 기반 색인 (키보드입력)	형태소 단위 색인 (문자인식)	2-gram 기반 색인 (문자인식)
0.0	0.8837	0.8381	0.8627	0.8093
0.1	0.8226	0.7814	0.7990	0.7581
0.2	0.6957	0.6677	0.6602	0.6158
0.3	0.5945	0.5501	0.5135	0.4998
0.4	0.5445	0.5052	0.4416	0.4412
0.5	0.5062	0.4448	0.3936	0.3965
0.6	0.4152	0.3799	0.2882	0.3178
0.7	0.3578	0.3443	0.2190	0.2776
0.8	0.2805	0.2637	0.1591	0.2058
0.9	0.2014	0.1680	0.1234	0.1346
1.0	0.1213	0.1083	0.0966	0.1075
11-포인트 평균	0.4930	0.4592	0.4142	0.4144

- 키보드 입력에 의해 구축된 데이터베이스에 대해서는 형태소 단위 색인법이 2-gram 기반 색인법보다 높은 검색 효과를 제공하고 있다. 그러나, 문자 인식에 의해 구축된 데이터베이스에 대해서는 형태소 단위 색인법에 의한 검색 효과가 많이 저하되어 2-gram 기반 색인법과 유사한 수준의 검색 효과를 제공하고 있음을 보여준다. 이는 문서에 존재하는 오류로 인하여 형태소 분석이 적절히 수행될 수 없음을 암시한다.
- 2-gram 기반 색인법은 문서에 존재하는 오류로 인한 검색 효과의 저하가 형태소 단위 색인법에 비하여 크지 않음을 알 수 있으며, 그 이유는 다음과 같이 설명될 수 있다. '인공지능 문자인식'이라는 문자열 영상을 문자 인식한 결과, 한 문자에 오류가 발생하여 '인공지농 문자인식'이라고 인식되었다고 가정하자. 이러한 문자열에 대해 2-gram 기반 색인법은 '인공', '공지', '지농', '문자', '자인', '인식'이라는 색인어들을 생성한다. 반면, 문자 오류가 발생하지 않았다면, '인공', '공지', '지능', '문자', '자인', '인식'이라는 색인어들이 생성된다. 따라서, 7개의 색인어들 중에서 1개의 색인어만이 다르기 때문에, 유사도 계산 결과는 문자열에 포함된 오류에 크게 영향받지 않을 것이다.

5. 결 론

문자 인식 기술은 문헌 정보들을 효율적으로 전산화할 수 있는 수단을 제공한다. 그러나, 인쇄 품질이 조악한 문서들의 문자 인식 결과들은 많은 오류를 포함하기 때문에, 현재 많은 정보 서비스들은 문서들을 스캔한 영상들을 사용자에게 제공하고 있다. 이와 같은 방식은 문서들의 본문에 대한 검색과 같은 기능을 제공할 수 없다. 이러한 문제점을 극복하는 방법들 중의 하나는 문자 인식을 통해 문서 영상을 텍스트로 변환하고, 변환된 텍스트에 대한 검색을 수행한 후, 검색된 텍스트에 해당하는 문서 영상을 사용자에게 제공하는 방법이다. 이때 중요한 문제는 문자 인식 오류를 많이 포함하고 있는 문서들에 대한 검색 방법에 대한 연구이다.

본 연구에서는 개별 문자 인식률 90% 수준의 문자 인식기에 의해 구축된 한글 문서 데이터베이스에 대해 형태소 단위 색인법과 n-gram 기반 색인법을 적용시켜 벡터 공간 모델을 기반으로 문서들을 검색함으로써, 문자 인식에 의해 구축된 한글 문서 데이터베이스에 대한 정보 검색의 타당성을 실험을 통하여 검토하였다. 실험 결과, 키보드 입력에 의해 구축된 데이터베이스에 대해서는 형태소 단위 색인법이 2-gram 기반 색인법보다 높은 검색 효과를 제공하였으나, 문자 인식에 의해 구축된 데이터베이스에 대해서는 형태소 단위 색인법에 의한 검색 효과가 많이 저하되어 2-gram 기반 색인법과 유사한 수준의 검색 효과를 제공하였다. 또한, 문자 인식에 의해 구축된 데이터베이스에 대한 검색 효과가 키보드 입력에 의해 구축된 데이터베이스에 대한 검색 효과에 크게 뒤지지 않는다는 사실은 문자 인식을 이용한 자동 문서 입력 및 검색에 관한 연구의 가능성을 밝게 하고 있다.

참 고 문 헌

[1] C.W. Cleverdon, "Optimizing Convevient On-line Access to Bibliographic Databases," Information Service and Use, Vol.4, No.1, pp.37-47, 1984.

[2] W.B. Cavnar, "N-Gram-Based Text Filtering for TREC-2," The 2nd Text Retrieval Conference (TREC-2), NIST Special Publication 500-215, pp.171-179, 1994.

[3] M. Damashek, "Gaushing Similarity with N-Grams : Language-Independent Categorization of Text," Science, Vol.267, pp.843-848, 1995.

[4] J.H. Lee, M.H. Kim and Y.J. Lee, "Ranking Documents in Thesaurus-Based Boolean Retrieval Systems," Information Processing & Management, Vol.30, No.1, pp.79-91, 1994.

[5] J.H. Lee, "Combining Multiple Evidence from Different Properties of Weighting Schemes," ACM SIGIR Conference on Research and Development in Information Retrieval, pp.180-188, 1995.

[6] J.H. Lee and J.S. Ahn, "Using N-Grams for Korean Text Retrieval," ACM SIGIR Conference on Research and Development in Information Retrieval, pp.216-224, 1996.

[7] G. Salton, A. Wong and C.S. Yang, "A Vector Space Model for Automatic Indexing," Communications of the ACM, Vol.18, No.1, pp.613-620,

1975.

[8] G. Salton and M.J. McGill, 'Introduction to Modern Information Retrieval,' McGraw-Hill, Inc., 1983.

[9] G. Salton, "Another Look at Automatic Text Retrieval," Communications of the ACM, Vol.27, No.7, pp.648-656, 1986.

[10] G. Salton, "Historical Note : The Past Thirty Years in Information Retrieval," Journal of the American Society for Information Science, Vol.38, No.5, pp.375-380, 1987.

[11] G.Salton, 'Automatic Text Processing : The Transformation, Analysis, and Retrieval of Information by Computer,' Addison Wesley, 1989.

[12] 강승식, 권혁일, 김동렬, "한국어 자동 색인을 위한 형태소 분석 기능", 한국정보과학회 봄학술발표논문집, 제22권 제1호, pp.930-932, 1995.

[13] 김성혁, 서은경, 이원규, 김명철, 김영환, 김재군, "자동 색인기 성능 시험을 위한 Test Set 개발", 정보관리학회지, 제11권 제1호, pp.81-102, 1994.

[14] 김영환, '한글 한자 혼용문의 자동 색인 시스템', 한국과학기술원 석사학위논문, 1982.

[15] 안현수, "한글 문헌 자동 색인에 관한 실험적 연구", 정보관리학회지. 제3권 제2호, pp.108-306, 1986.

[16] 예용회, "국내 문헌 정보 검색을 위한 키워드 자동 추출 시스템 개발", 정보관리연구, 제23권 제1호, pp.39-62, 1992.

[17] 이준호, 안정수, 박현주, 김명호, "한글 문서의 효과적인 검색을 위한 n-gram기반의 색인 방법", 정보관리학회지, 제13권 제1호, pp.47-63, 1996.

[18] 이현아, 홍남희, 이근배, "한국어 형태소 구조 규칙에 기반한 색인 시스템의 구현", 한국정보과학회 봄학술대회발표논문집, 제22권 제1호, pp.933-936, 1995.

[19] 정진성, '단일 문서내에서의 언어 및 통계 정보를 이용한 자동 색인', 한국과학기술원 석사학위논문, 1992.

[20] 최기선, "구문 및 의미 분석을 통한 한국어 자동 색인," 정보관리학회지 제8권 제2호, pp.96-107, 1991.

[21] 한선화, 이충식, 이준호, 김진형, "문자 인식 기술을 이용한 데이터베이스 구축에 대한 연구", 정보처리학회지 제출중.

[22] 한성현, '구문해석을 이용한 색인어 자동 추출 시스템의 설계와 구현', 한국과학기술원 석사학위논문, 1991.

한 선 화

e-mail : shhahn@kordic.re.kr

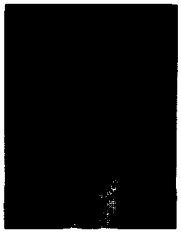
1987년 성균관대학교 정보공학과 (학사)

1989년 한국과학기술원 전산학과 (석사)

1997년 한국과학기술원 전산학과 (박사)

1997년~현재 연구개발정보센터 선임연구원

관심분야 : 데이터베이스/마이닝, Intelligent Tutoring, 에이전트, HCI



이 준 호

e-mail : joonho@computing.soongsil.ac.kr

1987년 서울대학교 전산학과(학사)

1989년 한국과학기술원 전산학과 (석사)

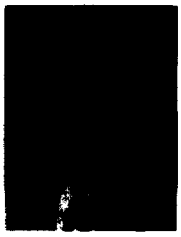
1993년 한국과학기술원 전산학과 (박사)

1993년~1994년 한국과학기술원 인공지능연구센터 연구원

1994년~1997년 연구개발정보센터 선임연구원

1997년~현재 숭실대학교 컴퓨터학부 부교수

관심분야 : 정보검색, 정보시스템, 데이터베이스



이 충 식

e-mail : cslee@ai.kaist.ac.kr

1994년 한국과학기술원 전산학과 (학사)

1996년 한국과학기술원 전산학과 (석사)

1996년~현재 한국과학기술원 박사과정

1998년~현재 동경공과대학 방문학생

관심분야 : 패턴인식, Neural Network, Genetic Algorithm

김 진 형

e-mail : jkim@cs.kaist.ac.kr

1971년 서울대학교 공과대학(학사)

1973년~1976년 과학기술연구소 (KIST) 전산실 연구원

1976년~1977년 미 California State, 도로국, 프로그래머

1979년 UCLA 전산학과(석사)

1981년~1985년 Hughes Research Center, Malibu, Senior Computer Sceintist

1983년 UCLA 전산학과(박사)

1990년~1991년 미 IBM Watson Research Center 초빙 연구원

1985년~현재 과학기술원 전산학과 교수

1991년~현재 과학재단 지정 과학기술원 인공지능 연구센터 부소장

1995년~현재 출연(연) 연구개발정보센터 소장

1997년~현재 공학한림원 회원

관심분야 : 문자인식, 지능형 인터페이스, 인공지능