효율적인 키워드 검색을 지원하는 학습자료의 구조화 방법 연구

김은경*, 최진오*

*부산외국어대학교 컴퓨터공학부

A Study on Structuring Method of Study Data Supporting Efficient Keyword Search

Eun-kyung Kim*, Jin-oh Choi*

*School of Computer Engineering, Pusan University of Foreign Studies

E-mail : jochoi@pufs.ac.kr

요 약

다양한 학습 자료를 저장해두고 검색하는 시스템들은 주로 키워드 검색을 지원하고 있다. 여기서, 키워드 매칭 방식은 같은 분야의 자료라 하더라도 사용자가 입력한 키워드와 정확한 매칭이 되지 않을 경우 검색되지 못하는 문제점을 안고 있다. 또한 학습 테스트를 위한 학습 문제 자료는 키워드로 검색하기에는 포함한 정보의 양이 너무 적어 적용되기 어렵다.

본 논문에서는 이러한 문제점을 해결하기 위하여 학습문서를 입력할 때 문서에 포함되어 있는 각 단어들을 형태소 분석에 의하여 중요 명사들을 추출하고 데이터베이스화하는 기법을 도입하고 미리 마련한 유사한 용어 지식 데이터베이스를 활용하여 지능적이고 효율적인 학습자료 검색 기법을 제안한다.

ABSTRACT

Most reading systems that supply various study data generally support keyword search. But the usual keyword matching techniques have a problem to require the exact keyword matching, and could not find similar field materials. Futhermore, testing materials have too little information to apply the keyword matching search.

To solve these problems, this thesis proposes the method to extract the important keyword from study data and to construct the database automatically when the data are stored at the storage. And using prepared similar terminology database, we suggest the intelligent and efficient technique to find study materials.

키워드

structuring study data, keyword search, computer added education

Ⅰ. 서 론

인터넷을 통한 교육이 보편화됨에 따라 방대한 양의 정보를 컴퓨터에 체계적으로 수록함으로써 정보 사용자가 필요로 하는 정보를 신속하고 정확하게 찾아내는 작업은 매우 중요한 일이 되었다[5]. 그러나 이를 위해 개발된 대부분의 학습시스템들은 계층적으로 단원 또는 분야를 검색하는 방식을 채택하고 있거나 키워드 검색을 지원하고 있다. 키워드 매칭(matching) 방식은 같은 분야의 자료라 하더라도 사용자가 입력한 키워드와 정확한 매칭이 되지 않을 경우 검색되지 못하거나 학습 테스트를 위한 학습문제 자료는 키워드로 검색하기에는 포함한 정보의 양이 너무 적어 적용되기 어려운 문제점이 있다.

이러한 문제점을 해결하기 위해서는 먼저 학습문서를 입력할 때 문서에 포함되어 있는 각 단어들을 형태소 분석에 의하여 중요 명사들을 추출하고 이를 함께 데이터베이스화할 필요가 있다. 둘째, 유사한 용어들을 유사용어지식 데이터베이스를 미리 마련해 두고 학습 자료를 검색할 때 관련이 있는 유사 학습 자료도 검색이 가능하도록 해야 한다. 셋째, 학습문제 자료는 학습문서와 데이터의 구조적 결합에 의하여 학습문서와 함께 검색 될 수 있게 지원할 필요가 있다. 마지막으

로, 검색된 학습 자료는 상기 두 데이터베이스의 매칭 정도에 따라 우선순위를 부여 할 수 있도록 지원되어야 한다.

따라서 본 논문은 학습 자료의 형태소 분석[4] 과정을 통해 문서분석 테이블을 구성하고, 구성된 문서의 클러스터링(clustering) 과정을 통한 키워드 검색 기법을 설계 구현한다. 또한 이러한 검색 시스템을 지원하기 위한 학습 자료의 효율적 구조화 방법을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서 관련 연구를 소개하고, 3장에서 본 논문에서 제안하는 새로운 학습 자료 검색 기법을 소개하고, 그 구현 결과는 4장에서 다룬다. 마지막으로 5장에서 결론과 향후 연구 과제에 대해 논의한다.

II. 관련 연구

지식검색 방법으로는 클러스터링(자동분류)을 이용한 검색방법과 자연어 검색방법이 주로 사용되고 있다. 클러스터링 시스템[1][2]은 단순히 검색결과를 랭킹 기법을 이용하여 순차적으로 보여주는 것이 아니라 의미적으로 유사한 결과들을 적합한 가중치를 적용하여 보여주는 기법이다. 여기서 클러스터링이란 주어진 데이터 집합을 서로 유사성을 가지는 몇 개의 클러스터로 분할 해 나가는 과정으로, 인공지능에서의 분류(classify) 기법을 이용한다.

지식검색의 또 하나의 방법으로 퍼지 집합을 이용한 모델을 많이 적용하고 있고 인덱싱 정보와 별도로 지식베이스를 구축하고 있는 자연어 검색기법이 있다[3]. 그러나, 이러한 접근 방법들은 대용량의 데이터 분류와 검색에 적용가능하나, 구현에 많은 비용과 시간을 필요로 한다. 따라서 본 논문이 대상으로 하는 학습자료 검색시스템에는 적합하지 않다.

본 논문에서는 두 기법을 병합하는 접근방법을 제시한다. 즉, 자연어 검색방법에서 제시하는 지식베이스를 구축하고 유사한 용어들을 묶어 문서를 클러스터링하는 방법을 이용하여 효율적 키워드 검색 기법을 설계하고 구현하였다.

III. 학습 자료 구조화 및 검색 기법

학습 자료를 구조적으로 제시함으로써 데이터베이스의 경제성과 무결성을 높일 수 있다. 본 논문에서 제안하는 학습 자료의 구조는 크게 학습문서의 형태소 분석정보, 유사용어지식정보, 그리고 학습자료정보로 나뉜다. 그림 1은 이 구조의 개념적 모델을 보이고 있다.

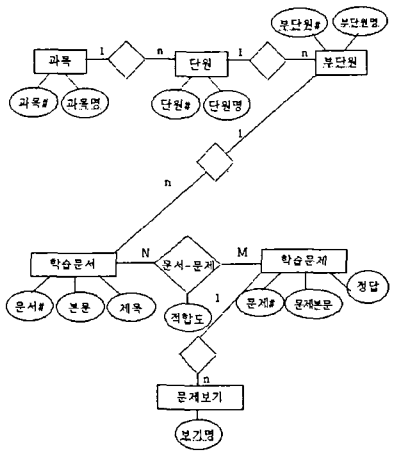


그림 1. 학습자료정보의 개념적 구조

학습 자료를 입력하여 구조화할 때는 문서형태소 분석정보를 함께 구성한다. 형태소 분석정보는 학습문서와 직접적으로, 그리고 학습문제와 간접적으로 연계되어 향후 사용자의 키워드 검색에 대응하게 된다. 유사단어는 유사용어지식정보로부터 사용자가 입력한 키워드를 이용하여 검색해낸다. 학습문서의 형태소 분석정보와 용어 사전 정보의 개념적 구조는 그림 2와 같다.

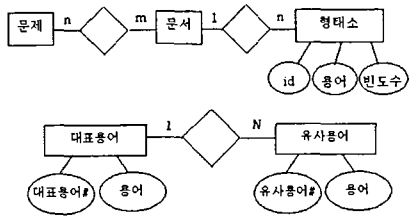


그림 2. 형태소정보와 유사용어 정보의 구조

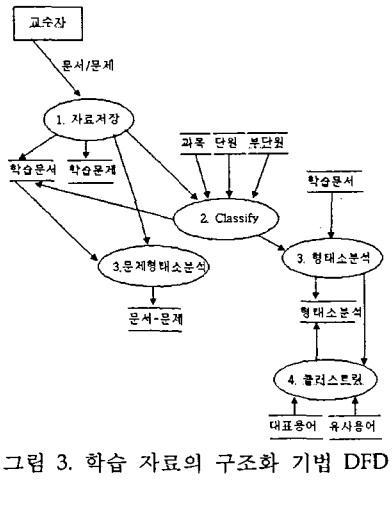


그림 3. 학습 자료의 구조화 기법 DFD

학습 자료의 구조화 알고리즘의 자료 흐름도(DFD)는 그림 3과 같다. 또한 학습 자료의 검색 알고리즘의 자료 흐름도(DFD)는 그림 4와 같다.

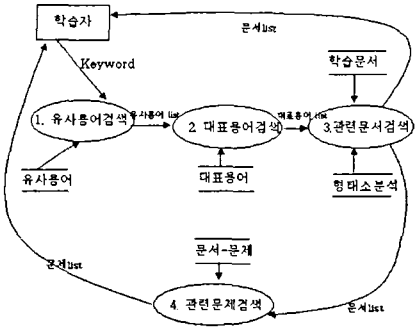
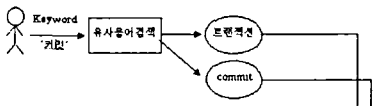


그림 4. 학습 자료의 검색 DFD

본 논문에서 제안하는 학습 자료의 검색 기법의 예를 그림 5에서 보이고 있다. 먼저, 사용자가 키워드 입력하면 입력된 키워드와 유사한 용어들을 유사용어지식정보로부터 모두 추출한다. 다음, 추출된 유사용어들의 대표용어를 형태소 분석정보와 비교하여 관련 학습문서를 구한다. 이때 형태소분석정보 테이블의 빈도수 값과 매칭되는 용어의 개수 값을 이용하여 검색된 문서들의 매칭 정도에 대한 확률 값을 계산할 수 있다. 그림5에서 '커밋'이란 키워드에 의해 문서 100이 문서 300보다 더 정확한 매칭이 됨을 알 수 있다.



id	문서#	용어	빈도수	대표용어#
1	100	transaction	10	트랜잭션
2	100	commit	6	commit
3	100	질의	4	SQL
4	200	록	7	Lock
5	200	락킹	5	Lock
6	300	consistency	12	무결성
7	300	참조키	10	외래키
8	300	transaction	3	트랜잭션

그림 5. 학습 자료의 검색 예

VI. 시스템 설계 및 구현

그림 6은 구현을 위해 설계한 시스템 구조도이다. 설계한 시스템은 크게 사용자와 출제자 인터페이스를 두고 있으며 문서를 구조화하여 저장하기 위한 문서 분석기와 클러스터링을 담당하는 모듈을 포함한다. 시스템의 구현은 IIS Web Sever 5.0에서 HTML, ASP, 그리고 Java Script를 이용하였으며 저장소로서 데이터베이스는 MS-SQL Server 2000을 이용하였다.

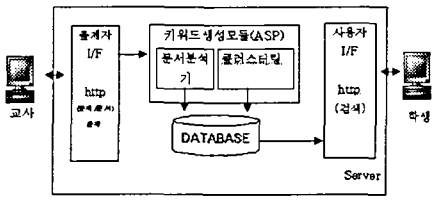


그림 6. 시스템 구조

학습 자료는 문서테이블과 형태소분석정보테이블에 저장되게 된다. 즉, 자료를 형태소 분석 프로그램인 HAM을 이용하여 분석하고 형태소분석정보테이블을 구성한다. 이때 대표용어 정보도 같이 생성된다. 그리고 대표용어정보를 참고하여 형태소분석정보의 대표용어 값이 채워진다. 이 과정을 그림 7에서 보이고 있다.

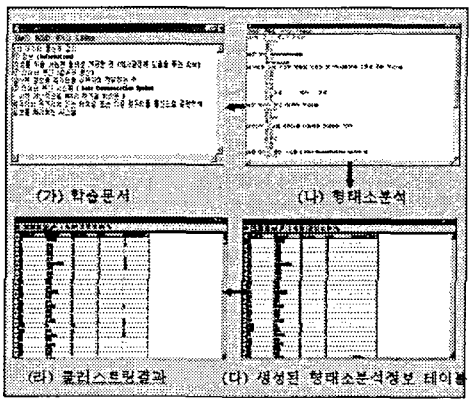


그림 7. 형태소분석정보테이블의 생성과정

그림 8은 학습 자료의 유사 용어 검색을 수행하는 구현 결과 화면이다. 키워드 검색 결과는 그림 9에서 보이고 있다. 그림 9의 검색 결과는 학습 문서와 문제를 모두 포함하고 있다. 그림 10에서와 같이 문제와 문서는 키워드와 유사 용어를 이용하여 상호 연결된 구조를 가지기 때문에 상호 검색을 수행하도록 지원될 수 있다.

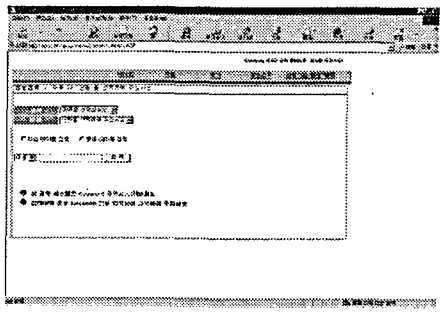


그림 8. 학습자료 검색 화면

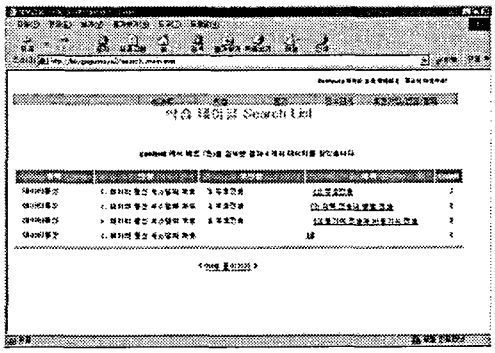


그림 9. 검색 결과 화면

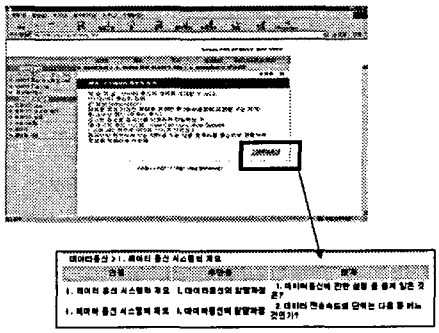


그림 10. 학습 문서와 학습 문제의 연결

V. 결론 및 향후 과제

본 논문에서는 학습 자료를 형태소 분석을 통하여 키워드 용어를 추출하였고 용어의 출현 여부만을 반영하여 출현빈도가 높은 단어에 대해 가중치를 적용하는 방법을 취하였다. 또한 유사용어테이블을 두어 사용자가 검색하고자하는 용어와 유사한 용어의 단어빈도까지 계산하여 검색의 효율성을 높이고자 하였다.

본 논문에서 제시하는 키워드 검색 기법은 용어의 출현의 수에 가중치를 적용하는 방법을 제시한 것이다. 이 방법은 출현하는 용어들이 각 문서의 내용과 밀접하게 관련되어 있음을 전제로 하고 있다. 그러나 용어의 출현이 높다고 해서 그 용어와 문서사이의 관련성이 높다는 것은 엄밀하게 일치하지 않는다.

향후 연구과제로는 용어의 출현 여부만을 이용한 가중치 적용방식에서 확장하여 의미에 기반한 보다 지능적이고 효율적인 가중치 적용 방법에 대한 고찰이 필요할 것으로 판단된다.

참고문헌

[1] 이재윤, 최보영, 정영미, "문헌 자동분류에서 용어가중치 기법에 대한 연구", 20001 한국정보처리학회 춘계 학술발표논문집, 제8권 1호

[2] 정영미, 이재윤, "지식 분류의 자동화를 위한 클러스터링 모형 연구", 정보관리학회지 제18권 2호, p203-230

[3] R. Kosala and H. Blockeel, "Web Mining Research : A Survey" , ACM SIGKDD, Vol 2, p1-5, 2000

[4] 한국어 분석 모듈 HAM version 5.0.0.a, http://nlp.kookmin.ac.kr/HAM/kor/download.html

[5] 맹성현, 주종철, "문서구조화와 정보검색' ,정보과학학회지, 제16권 8호, 1998, p6-15

[6] 김두연, "우리나라 원격교육 현황", 한국정보처리학회지, 제4권 3호, 1997, p4-12.