2001년 한국정보처리학회 춘계 학술발표논문집 제8권 제1호

데이타마이닝을 이용한 전문 검색엔진의 설계 및 구현

황보윤*, 김병찬, 김영지, 문현정, 우용태
창원대학교 전자계산학과
E-mail : laghby@hyunse.co.kr

A Design and Implementation of Expert Search Engine Using DataMining

Youn Hwang-Bo*, Chan-Byung Kim, Young-Ji Kim,
Hyeong-Jeong Mun, Yong-Tae Woo
Dept. of Computer Science, Changwon University

요 약

본 논문에서는 데이타마이닝 기법을 이용하여 지능형 전문 검색엔진을 설계하고 사용자 인터페이스를 구현하였다. 먼저, 컴퓨터 분야의 전문 용어에 대하여 연관 규칙 탐사 알고리즘을 이용하여 의미적으로 연관된 용어들끼리 클러스터로 구성하였다. 전문 용어별로 구성된 클러스터는 본 논문에서 제안한 지식베이스 테이블에 저장하여 의미적으로 연관된 용어를 포함하는 웹 문서를 검색하는 과정에서 이용하였다. 검색과정에서는 사용자가 제시한 키워드와 관련된 전문 용어들간의 연관정도를 가중치로 부여하여 연관 정도가 높은 웹 문서순으로 출력하였다. 제안된 방법을 통하여 사용자가 제시한 키워드와 의미적으로 연관된 웹 문서를 효과적으로 검색할 수 있었다.

1. 서론

웹의 급속한 팽창에 따라 검색엔진에서 색인하고 있는 웹 문서의 수가 기하급수적으로 증가하고 있다. 이러한 검색엔진의 성능을 평가하기 위한 척도로 색인된 웹 문서의 규모나 검색 속도를 주로 사용하고 있다. 하지만 대부분의 검색엔진에서는 키워드 매칭 방식에 의해 색인된 문서를 검색하는 방식을 사용하는 관계로 사용자가 원하는 개념적인 지식 정보를 포함하는 문서를 효과적으로 검색하기 어렵다.

최근에는 키워드 매칭에 의한 단순 정보 검색보다는 지식 정보를 효율적으로 검색하기 위한 지능형 검색엔진이나 지식 정보를 체계적으로 관리하기 위한 지식탐사시스템이 활발하게 연구되고 있다. 대량의 문서로부터 지식 정보를 검색하기 위한 연구 방향으로 데이타마이닝(Datamining) 기법을 이용하여 대량의 문서로부터 지식 정보를 탐색하기 위한 텍스트 마이닝(Text Mining) 기법과 웹 문서로부터 지식 정보를 검색하기 위한 웹 마이닝(Web Mining) 기법 등에 대한 연구가 진행되고 있다[1].

본 논문에서는 데이타마이닝 기법을 이용하여 지능형 전문 검색엔진을 설계하고 사용자 인터페이스를 구현하였다. 이를 위해 데이타마이닝 기법에서 사용하는 연관 규칙 탐사 알고리즘을 이용하여 컴퓨터분야의 전문 용어들끼리 클러스터링한 지식베이스 테이블을 구성하였다. 지식베이스 테이블은 전문 용어와 의미적으로 연관된 용어들끼리 클러스터한 테이블이다. 그리고 지식베이스 테이블을 이용하여 관련된 용어들간의 연관 정도에 따라 지식 정보를 검색할 수 있는 지능형 전문 검색엔진을 구현하였다. 제안된 방법을 통하여 사용자가 제시한 키워드와 의미적으로 연관된 전문 용어를 동시에 포함하는 웹 문서를 우선적으로 출력할 수 있다. 즉, 관련된 전문 용어들간의 연관 정도를 가중치로 부여하여 키워드와 의미적으로 연관 정도가 높은 웹 문서순으로 검색할 수 있다. 이러한 방법을 통하여 기존의 검색엔진에서 사용하는 키워드 위주의 검색 결과보다 관련된 지식 정보를 포함한 정보를 효과적으로 검색 할 수 있었다. 제안된 방법의 효율성을 검정하기 위하여 컴퓨터 분야의 전문 사이트를 대상으로 관련 웹 문서를 수집하여 인덱싱하고, 사용자 인터페이스를 구현하였다.

2. 웹 마이닝(Web Mining)

웹 마이닝은 웹 상의 데이터로부터 잠재적으로 유용하고 알려지지 않은 정보 또는 지식을 발견하기위한 전반적인 과정을 말한다. 또한 웹 문서와 서비스로부터 자동적으로 정보를 추출하기 위하여 데이타마이닝 기법을 적용하기 위한 기법을 의미한다[2]. Madria, et al.[3]과 Borges and Levene[4]는 웹 마이닝 기법을 웹 컨텐트 마이닝(Web Content Mining), 웹 구조 마이닝(Web Structure Mining) 그리고 웹 사용 마이닝(Web Usage Mining)으로 구분하였다.

3. 지식베이스 테이블을 이용한 전문 검색엔진

다음 그림 1은 본 논문에서 제안한 지식베이스 테이블을 이용한 전문 검색엔진의 전체적인 구성도이다.

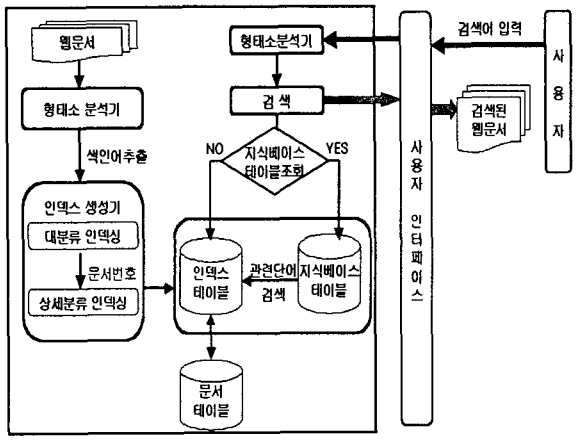

그림 1. 전문 검색엔진의 전체적인 구성도

3.1 지식베이스 테이블 구성을 위한 전처리 과정

초기 지식베이스 테이블은 컴퓨터 관련 논문 240편을 대상으로 구성하였다. 실험 대상 문서에 대한 형태소 분석을 통하여 문서에서 출현하는 모든 용어를 추출하였다. 형태소 분석은 공개용 형태소 분석기인 HAM4.0a[5]를 사용하였다. 그리고 컴퓨터 용어 사전에 수록된 전문 용어를 기준으로 컴퓨터 분야의 용어를 추출하였다. 일반적으로 학술 논문에서 사용하는 전문 용어는 저자에 따라 영어나 한국어를 혼용하거나 영문 용어를 한글화하는 과정에서 차이가 있는 관계로 동의어 사전을 구성하여 전문 용어를 통일하였다. 즉, 'netwotk', '네트워크'와 같은 용어나 '데이타베이스', '데이터베이스'와 같이 동일한 의미의 단어는 동의어로 취급하여 하나의 용어로 통일하였다.

전체 문서에서 출현하는 절대 빈도수가 매우 적은 관계로 연관 규칙 탐사 대상이 되지 않는 용어와 용어들의 분포도가 매우 큰 관계로 무의미한 연관 규칙을 발생시키는 용어를 특이 용어로 취급하여 연관 규칙 탐사 과정에서 배제시켰다[6]. 그리고 공통적으로 출현하는 단어에 대한 가중치를 조정하기 위하여 TF*IDF 알고리즘을 적용하였다. TF*IDF 알고리즘은 역 문서 빈도수를 단어의 빈도수와 같이 적용함으로써 그 문서를 대표하는 단어들을 효율적으로 찾을 수 있는 알고리즘이다[7].

3.2 인덱싱 기법

본 논문에서 제안한 전문 검색엔진의 인덱싱 기법은 오라클 관계형 데이타베이스를 기반으로 인덱스 정보와 문서 정보를 데이타베이스 테이블에 저장하는 방식을 사용하였다. 이 방식은 파일 시스템을 위주로 개발된 기존 검색엔진의 문제점을 개선하여 대용량의 정보를 효율적으로 관리할 수 있다.

인덱싱 과정은 검색 속도를 개선하기 위하여 대분류 인덱싱 과정과 각 문서에 대한 색인어와 내용 정보를 인덱싱하여 실제 정보 검색과정에서 사용하기 위한 상세분류 인덱싱으로 구성된다. 대분류 인덱싱은 지식베이스 테이블을 구성하는 전문 용어별로 가장 관련된 우선 순위순으로 문서 번호를 구성하는 과정이다. 다음 그림 2는 지식베이스 테이블을 이용하여 전문 용어별로 가장 연관된 문서 번호순으로 대분류 인덱스를 구성하는 개념도이다.

상세분류 인덱싱은 사용자가 제시한 검색어를 포함하는 실제 문서를 검색하기 위해 문서 테이블을 구성하는 과정이다. 문서 테이블에는 색인어, 문서 URL, 제

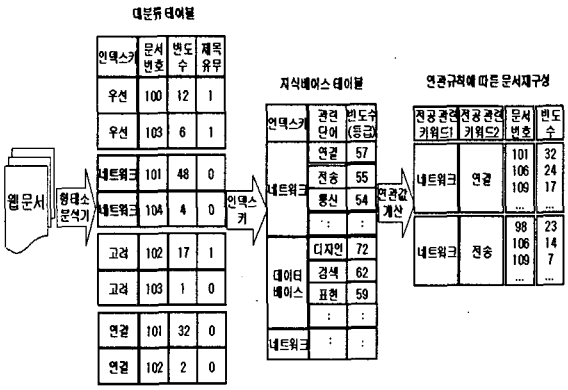

그림 2. 지식베이스 테이블을 이용한 대분류 인덱싱

목 정보, 문서 길이, 파일명, 문서 내용과 문서 내용에 포함된 전공 관련 키워드 등과 같은 정보가 저장된다. 여기서 전공관련 키워드는 형태소 분석시 컴퓨터 용어 사전을 참조하여 컴퓨터 관련 전문 용어들을 추출하여 저장하였다.

다음 그림 3은 상세분류 인덱싱을 구성하는 그림이다.

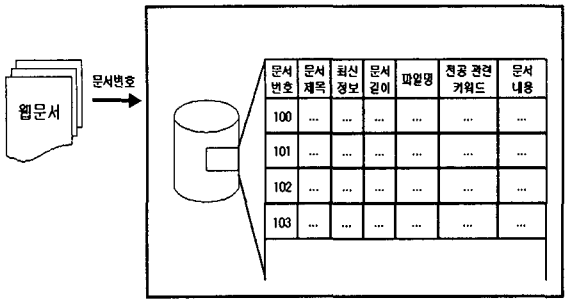

그림 3. 상세 분류 인덱싱

3.3 지식베이스 테이블을 이용한 사용자 질의 처리

본 논문에서 제안한 지식베이스 테이블은 관련된 용어들간의 연관 정도에 따라 지식 정보를 검색하기 위해 구성하였다. 지식베이스 테이블은 전문 용어별로 의미적으로 관련된 클러스터를 저장하기 위한 테이블이다. 지식베이스 테이블의 구성은 데이타마이닝 기법에서 사용하는 연관 규칙 탐사 알고리즘을 적용하여 추출한 관련 용어간의 클러스터 정보와 동일한 클러스터에 속한 용어들간의 연관 정도를 가중치로 나타낸 항목으로 구성된다. 지식베이스 테이블을 이용하여 검색어로 제시된 키워이 드와 의미적으로 연관된 전문 용어를 동시에 포함하는 웹 문서를 우선적으로 검색할 수 있다.

3.4 웹 문서와 용어 클러스터간의 유사도 계산

텍스트마이닝에서는 문서간의 유사도를 측정하기 위하여 다양한 형태의 유사계수를 사용하고 있다. 유사계수는 대상에 따라 거리계수, 연관계수, 상관계수, 확률적 유사계수 등으로 구분된다[8]. 또한 문서를 자동적으로 분류하기 위해 코사인, 다이스, 자카드 계수 등과 같은 유사계수가 사용되고 있다. 이러한 유사계수는 문서, 용어 또는 사용자의 질의어 등의 요소에 따라 문서간의 유사도를 측정하기 위해 사용되고 있다.

본 논문에서는 키워드와 의미적으로 연관된 웹 문서 그룹을 검색하기 위하여 지식베이스 테이블에 저장된 해당 용어에 대한 클러스터와 유사도를 계산하기 위하여 다음과 같은 유사도 계산식을 제안하였다.

$$decending(\forall\ D_k,\ \sum_{i=1}^{n} \min(KC_0, RC_i) * f_i) \quad (1)$$

D_k : 입력 키워드 포함 문서
n : 문서에 출현하는 관련 키워드의 총 개수
KC_0 : 각 문서별 입력 키워드 빈도수
RC_i : 각 문서별 i번째 관련 키워드 빈도수
f_i : 지식베이스테이블에서 입력 키워드와 i번째 관련 키워드의 빈도수

위 식(1)에 의해 구한 결과 값은 해당 웹 문서와 용어 클러스터간의 유사도를 의미한다. 따라서 사용자가 제시한 키워드를 포함하는 웹 문서의 집합을 대상으로 유사도를 각각 계산하여 내림 차순으로 출력하면 키워드와 가장 의미적으로 연관된 웹 문서를 우선적으로 검색할 수 있다.

4. 지식베이스 테이블을 이용한 검색 실험

본 논문에서 제안한 전문 검색엔진을 이용하여 컴퓨터 관련 용어를 대상으로 다양한 검색 실험을 하였다. 다음 표 1은 네이버 검색엔진과 엠파스 검색엔진 그리고 제안한 시스템에서 "네트워크" 키워드에 대해 검색한 상위 5개 웹 문서에 포함된 연관된 용어의 출현 빈도 수를 비교한 결과이다. 본 시스템에서 검색된 웹 문서에서 포함된 관련 용어의 빈도 수가 훨씬 높은 것을 알 수 있다.

표 1. "네트워크" 키워드와 관련된 출현 빈도수

	문서	키워드	관련어				
		네트워크	연결	전송	통신	프로토콜	접근
네이버	문서 1	30	1	1	1	2	0
	문서 2	15	0	0	0	0	0
	문서 3	47	1	0	4	0	0
	문서 4	19	0	2	2	1	0
	문서 5	15	0	0	0	6	0
엠파스	문서 1	15	0	0	0	0	0
	문서 2	16	0	0	1	0	0
	문서 3	11	0	0	0	1	0
	문서 4	21	1	1	1	2	0
	문서 5	9	0	0	0	0	0
제안 시스템	**문서 1**	**229**	**40**	**18**	**22**	**19**	**58**
	문서 2	**70**	**28**	**39**	**30**	**33**	**0**
	문서 3	**79**	**27**	**41**	**30**	**34**	**0**
	문서 4	**9**	**24**	**54**	**12**	**5**	**5**
	문서 5	**22**	**19**	**39**	**63**	**12**	**2**

기존 검색엔진과 비교한 실험 결과에서처럼 본 논문에서 제안한 지식베이스 테이블을 이용하여 키워드와 연관된 정보를 효과적으로 검색할 수 있었다. 즉, 키워드를 포함하는 단순한 웹 문서에 대한 검색이 아니라 키워드와 개념적으로 연관된 용어를 포함하는 지식 정보를 검색할 수 있다. 또한 이러한 지식 정보 검색 기능을 이용하여 대량의 지식 문서를 관리할 수 있는 지식관리시스템 구축이 가능할 것이다.

다음 그림 4는 검색 화면을 이용하여 "네트워크" 키워드를 검색한 결과이다. 검색 결과에서처럼 본 시스템에서는 키워드를 포함하는 웹 문서만 출력하는 기존 검색엔진과 다르게 키워드에 대한 연관성에 의한 검색

결과를 출력할 수 있었다. 또한 키워드와 연관된 용어 그룹을 동시에 출력하여 사용자들이 연관된 용어 그룹에 대한 검색 결과를 반복적으로 이용하여 추가적인 관련 정보를 효율적으로 검색할 수 있다.

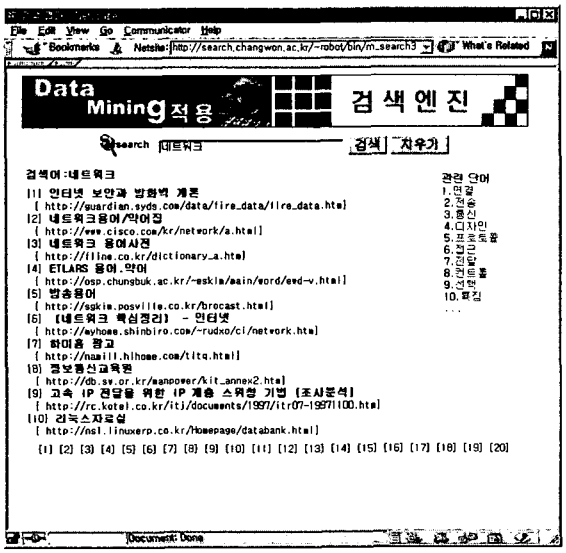

그림 4. "네트워크" 키워드에 대한 검색 결과

5. 결 론

본 논문에서는 데이타마이닝 기법중의 하나인 연관 규칙 탐사 기법을 이용하여 전문 검색엔진을 설계하고 사용자 인터페이스를 구현하였다.

연관 규칙 탐사 알고리즘을 이용하여 전문 용어별로 의미적으로 연관된 용어끼리 클러스터로 구성하였다. 전문 용어별로 구성된 클러스터를 이용하여 본 논문에서 제안한 지식베이스 테이블을 구성하였다. 그리고 사용자가 제시한 키워드와 관련된 웹 문서를 검색하는 과정에서 지식베이스 테이블을 이용하였다. 즉, 사용자가 제시한 키워드와 관련된 전문 용어들간의 연관정도를 가중치로 부여하여 의미적으로 연관 정도가 높은 웹 문서순으로 출력하였다. 컴퓨터 분야의 웹 문서에 대한 실험 결과를 통하여 사용자가 제시한 키워드와 의미적으로 연관된 웹 문서를 효과적으로 검색할 수 있었다.

앞으로 다양한 분야에 대한 연관성 분석을 통하여 일반적인 문서로부터 지식 정보를 효율적으로 검색할 수 있는 일반적인 지능형 검색엔진의 개발에 대한 연구를 진행할 계획이다.

참고문헌

[1] R. Kosala and H. Blockeel, "Web Mining Research : A Survey", ACM SIGKDD, Vol 2, pp. 1-5, 2000.

[2] O. Etzioni, "The world wide web: Quagmire or gold mine", Communications of the ACM, Vol.39, No.11, pp.65-68, 1996.

[3] S. K. Madria, S. S. Bhowmick, W. K. Ng and E.-P. Lim, "Research issues in web data mining", In Proceedings of Data Warehousing and Knowledge Discovery, First International Conference, DaWaK '99, pp. 303-313, 1999.

[4] J. Borges and M. Levene, "Data mining of user navigation patterns", In Proceedings of the WEBKDD'99 Workshop on Web Usage Analysis and User Profiling, August 15, 1999, San Diego, CA, USA, pp.31-36, 1999.

[5] 강승식, "HAM: 한국어 형태소 분석 라이브러리", http://ham.hansung.ac.kr

[6] 이정화, 남상엽, 문현정, 우용태, "데이타마이닝 기법을 이용한 효율적인 전문 용어 클러스터링," '2000 한국 데이타베이스 학술대회 논문집, pp.210-215, 2000.

[7] Thorsten Joachims, "A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization", CMU-CS-96-18, March 1996.

[8] 한승희, "문헌 클러스터링을 위한 유사계수간의 연관성 측정", 한국정보관리학회 학술대회 논문집, pp.25-28, 1999.