가변 속성 포스팅 구조의 설계

안후영*, 이승희*, 박영호*, 이종훈**
*숙명여자대학교 멀티미디어과학과
** 한국전자통신연구원 소셜미디어서비스연구팀
e-mail : hyahn85@sookmyung.ac.kr, ariseshine@sookmyung.ac.kr, yhpark@sookmyung.ac.kr, mine@etri.re.kr

A Design of Variable Attributes Posting Structures

Hoo Young Ahn*, Seung Hee Lee*, Young Ho Park*, Jong Hoon Lee**
*Dept. of Multimedia Science, Sookmyung Women's University
** Social Media Service Team, ETRI

요 약

최근, 이질적인 형태의 멀티미디어 데이터들의 증가와 함께, 멀티미디어 콘텐츠들의 저장 및 검색의 중요성이 대두되고 있다. 기존의 검색 엔진들은 대부분 텍스트 데이터만을 대상으로 하고있으며, 특별한 도메인에서는 객체 지향 데이터베이스, 객체 관계 데이터베이스 등 별도의 데이터베이스를 활용하여 검색에 사용하고 있다. 본 논문에서는 다양한 멀티미디어 콘텐츠들의 저장 및 색인에 유연한 가변 속성 포스팅 구조를 제안한다. 제안하는 가변 포스팅 구조는 벡터의 개념을 사용하여 포스팅의 속성을 추가할 수 있게 하였다. 본 논문에서 제안하는 포스팅 구조는 이질적인 형태의 멀티미디어 콘텐츠들을 각 콘텐츠들의 속성에 따라 가변적으로 저장 구조를 정의하고, 이에 따라 인덱스를 구축할 수 있는 확장성 있는 방안이다.

1. 서론

최근 멀티미디어 데이터들의 방대한 증가와 디지털 콘텐츠 종류들의 다양화가 가속되고 있다. 이를 통해 다양한 형태의 디지털 콘텐츠들의 효율적인 저장 및 검색의 문제가 대두되고 있다. 본 논문에서 정의하는 디지털 콘텐츠란 텍스트 문서, 동영상, 음악, 이미지 등을 의미한다. 기존의 검색 엔진들은 대부분 관계형 데이터베이스(Relational Databases)를 이용하여 디지털 콘텐츠들을 관리하였다[1, 2]. 또한 관리할 데이터의 특성에 따라 객체지향 데이터베이스(Object Oriented Database), 객체 관계형 데이터베이스(Object Relational Database) 등을 별도로 사용하여 디지털 콘텐츠들을 관리하였다[3]. 그러나 위와 같은 데이터 저장 및 인덱스 방법은 특정 상용 데이터베이스들에서 제공하는 한정된 인덱스 구조이기 때문에 다양한 형태의 디지털 콘텐츠들을 자유롭게 관리하고 인덱싱 하는데 어려움이 많다.

본 논문에서는 이와 같은 문제를 해결하고자 가변 속성 역 인덱스 구조를 제안한다. 본 논문에서 제안하는 가변 속성 역 인덱스 구조는 기본적인 역 인덱스의 개념을 확장하여 설계하였다. 역 인덱스란 인덱스를 기준으로 데이터를 만드는 것이 아니라, 데이터를 기준으로 인덱스를 뽑아 생성하는 텍스트 기반 검색 엔진에서 주로 사용되는 색인 방법이다[4]. 역 인덱스의 핵심은 포스팅에 있다. 본 논문에서는 포스팅의 구조를 사용자가 정의하여 인덱싱을 수행할 수 있도록 하였다.

본 논문의 공헌은 다음과 같다.

- 제안하는 가변 속성 포스팅 구조는 멀티미디어 새로운 형태의 멀티미디어 콘텐츠들의 관리에 유연한 장점을 가진다.
- 가변 속성 포스팅 구조는 콘텐츠의 형태에 제한되지 않는 인덱스 구조이므로 향후 멀티미디어 데이터 관리에 확장성을 가진다.

본 논문의 구성은 2 장에서는 본 연구와 관련된 연구들을 소개하고, 3 장에서는 본 논문에서 제안하는 가변 속성 포스팅 구조에 대하여 소개 하며, 특히, 가변 포스팅 구조에 대하여 설명하며, 마지막으로 4 장에서는 본 논문의 결론을 내린다.

2. 관련 연구

역 인덱스란 텍스트 문서를 대상으로, 문서 내에 있는 단어들을 인덱스의 키(Key)로하여, 단어의 위치와 빈도수를 포스팅으로 한 정보 검색의 대표적인 인덱스이다. 역 인덱스는 대용량 문서에 대한 빠른 검색을 지원하는 정보 검색 분야에서 검증된 검색 인덱

* 본 연구는 지식경제부 IT 원천기술개발 사업의 일환으로 수행하였음. [2008-F-043-01, 장소/사회적 관계 인지형 Social 미디어 서비스 기술]

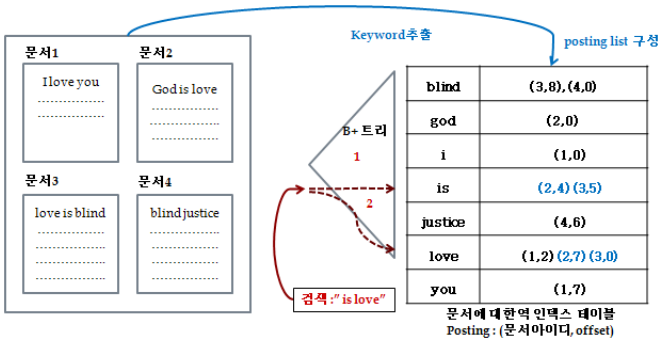
스로, 그 활용 범위가 다양하다[4-5]. 본 논문에서는 역 인덱스의 주요 구조인 포스팅을 가변적으로 설계하여, 다양한 멀티미디어 콘텐츠들의 색인이 가능한 확장성 있는 검색 인덱스를 제안한다.

3. 본론

본 장에서는 본 논문에서 제안하는 가변 속성 포스팅 구조에 대하여 자세히 설명한다.

3.1 가변 속성 포스팅 구조

대용량 데이터 저장 및 검색에 효율적인 인덱스로서 역 인덱스가 많이 사용되고 있다. 역 인덱스는 텍스트 문서를 대상으로 문서 내의 모든 단어를 인덱스로 하는 색인 방법이다. 그림 1 은 역 인덱스 구조를 보인다.



(그림 1) 역 인덱스 구조

그림 1 의 왼쪽에는 인터넷 상의 다양한 문서들이 존재한다. 대표적으로 문서 1~4 에는 I, love, you, God, is, blind, justice 의 단어들이 속해있다. 역 인덱스란 이러한 단어들을 모두 인덱스의 키(Key)로 하여 오른쪽과 같은 역 인덱스 테이블을 생성한다. 역 인덱스 테이블은 문서에서의 단어들을 키로하여, 테이블 내의 포스팅(Posting)에 문서의 아이디와 해당 단어의 출현 위치(Offset)을 저장한다. 역 인덱스의 포스팅은 문서의 아이디와 오프셋을 구성요소로 하며, 해당 단어가 문서들 내에서 발견 됨에 따라 포스팅들이 여러 개 생기며, 이러한 포스팅 들을 포스팅 리스트라 한다. 즉, 포스팅 리스트란 포스팅들의 집합을 의미한다.

역 인덱스는 위의 B+트리와 역 인덱스 테이블로 구성된 인덱스이다.

위의 인덱스를 통하여 만약, 사용자가 “is” 와 “love” 의 키워드를 인덱스를 통하여 검색하는 경우 , B+트리를 통하여 “is”와 “love” 포스팅 리스트를 찾게 된다. 포스팅 리스트를 찾으면, 각 단어들의 포스팅 리스트 요소들을 비교하며, 문서 아이디를 비교한 후, 같은 문서 내에 있는 검색어 들이 검색 결과의 우선순위가 되며, 같은 문서에 있다면, 오프셋이 가까울수록 우선 순위가 높게 된다.

본 예제에서는 문서 2 와 3 이 검색되고, 두 단어의 오프셋이 더 가까우므로 문서 2 의 검색 순위가 문서 3 보다 높은 결과를 가진다.

가변 속성 포스팅 구조란 그림 1 의 역 인덱스 구조의 포스팅 구조의 속성들이 가변적인 타입을 가지게 하여 저장 대상의 특성에 따라 스키마를 가변적으로 설계하여 저장 및 검색 가능하도록 하는 인덱스이다. 본 논문에서 제안하는 가변 포스팅에 대해서는 3.2 절에서 자세히 설명하도록 한다.

3.2 가변 포스팅

가변 포스팅이란, 역 인덱스 구조의 포스팅 구조를 멀티미디어 콘텐츠의 특성에 따라 저장 스키마를 다르게 하여 저장하는 유연한 저장 구조이다. 만약, 이미지 콘텐츠를 대상으로 검색 포스팅 구조를 구축한다면, 그림 1 의 포스팅 구조를 이미지 파일의 속성들(Attributes)인 Posting : (이미지 이름, 색상, 명도, 채도, 파일 형태)로 스키마를 설계한다. 이는 동영상, 음악 콘텐츠 등도 각 콘텐츠의 속성에 따라 저장 스키마를 다르게 하여, 검색 대상의 종류와 상관 없이 본 논문에서 제안하는 포스팅 구조를 사용하면, 모든 형태의 콘텐츠들을 색인 할 수 있도록 한다.

4. 결론

본 논문은 다양한 멀티미디어 콘텐츠들을 저장하고, 색인할 수 있는 유연한 구조인 가변 속성 포스팅 구조를 제안하였다. 제안된 구조는 이질적인 멀티미디어 콘텐츠들의 저장 및 검색이 가능하도록 하는 하부 저장 구조로, 확장성 있는 검색 기술로 활용될 수 있을 것으로 사료된다.

참고문헌

[1] S. Heinz and J. Zobel. Efficient single-pass index construction for text databases. *JASIST*, Vol. 54, No. 8, pp. 713-729, 2003

[2] V. Hristidis and Y. Papakonstantinou. Discover: Keyword search in relational databases. In Proc. 28th International Conference on Very Large Data Bases , Hong Kong, pp. 670-681, August, 2002

[3] A. Guttman. R-trees: a dynamic index structure for spatial searching. In Proc. *1984 ACM SIGMOD international conference on Management of data*, pp. 47-57. ACM Press, 1984

[4] Cutting and J. Pedersen, “Optimizations for dynamic inverted index maintenance,” In Proc. *13th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 405-411, 1990

[5] 정병수, 이혜자,” 대용량 XML 문서의 효율적인 질의 처리를 위한 세그먼트 기반 역 인덱스,” . *한국서비스학회지*, Vol. 7, No. 3, pp. 145-157, 2008

[5] 서치영, 이상원, 김형주,” XML 문서에 대한 RDBMS에 기반을 둔 효율적인 역 색인 기법,” . *한국정보과학회논문지*, Vol. 30, No. 1, pp. 27-40, 2003