

스키마 통합을 위한 시소러스의 활용 방안

이원조^o 이단영 권순덕 고재진
울산대학교 대학원 컴퓨터정보통신공학부
wjlee@mail.ulsan-c.ac.kr^o

Utilization of Thesaurus for Schema Integration

Won-Jo Lee^o Dan-Young Lee Soon-Duk Kwon Jae-Jin Koh
School of Computer Engineering and Information Technology University of Ulsan

요 약

최근 정보기술의 급속한 발달로 기존 정보시스템의 확장을 위한 데이터베이스 재설계에 대한 사용자들의 요구에 직면해 있다. 그러나 기존 데이터베이스 시스템은 다중 서버에 분산되어 있거나, 규모가 방대하여 확장을 위한 재설계에 많은 비용을 지불하고 있다. 따라서 기존 데이터베이스에서 서브스키마들의 정보를 추출하고, 이들간의 유사성 분석을 통하여 의미충돌을 해결하는 통합 방법을 적용하고 있는데, 본 연구에서는 이를 지원하기 위한 서브스키마의 유사성 분석을 위한 시소러스의 활용 방안을 제시한다.

1. 서 론

일반적으로 기존의 데이터베이스 스키마를 분석하는 목적에는 첫째, 기존 정보 시스템을 재구성하거나, 아키텍처의 전환시에 참조하기 위한 정보들의 추출 둘째, 스키마의 재사용을 위한 연관된 참조 스키마들의 라이브러리 구축 셋째, 기존 데이터베이스 내의 정보 흐름과 데이터의 중복을 확인 넷째, 웹서버에 저장된 개념적 정보의 분석이나, 서로 다른 시스템의 상호동작 가능성 점검 등인데, 이러한 분석 정보들을 데이터베이스 설계에 효율적으로 이용하기 위해서는 보다 체계적인 추출, 저장, 검색, 비교 및 기타 분석을 위한 기법이 요구된다.[1][8]

본 논문에서는 기존 데이터베이스에서 서브스키마들의 요소 정보들을 추출하여 서브스키마들 간의 유사성 분석을 위해서 시소러스를 활용하는 방법을 제시하고, 기존의 의미객체 모델을 이용한 서브스키마 통합 방법론[7]에서 의미충돌 해결 알고리즘을 서브스키마 시소러스를 이용하는 알고리즘으로 제안한다. 이러한 과정들은 서브스키마의 유사성 분석을 통한 의미충돌을 해결하기 위한 시소러스 구축의 사전 단계인데, 서브스키마 시소러스의 구축 목적은 정보시스템의 재설계 단계에서 대상 서브스키마들의 구성요소와 특질들에 대한 정보를 이용해서 표준화된 스키마 통합에 활용하기 위한 것이다.[1][2][7]

2. 관련연구

2.1 서브스키마 통합 방법

서브스키마의 통합 목적은 서브스키마간 개별 의미객체의 동일 개념을 표현하고 있는 모든 객체들을 찾아 표준화하여 통합하는데 있다. 그러나 동일 개념을 표현하고 있는 의미객체들을 찾는다는 것은 매우 어려운 과제이다. 첫째, 동일 의미객체를 설계시 설계자 자신의 관점

에서 모델링하기 때문이고, 둘째, 의미객체 모델이 다양한 표현구조를 갖는 매우 복잡하다는 것이다. 셋째, 이름, 구조 등이 서로 호환할 수 없는 표현으로 명시되기 때문이다. 이러한 원인의 결과로 동일 개념의 서로 다른 표현인 충돌을 야기한다. 이러한 충돌을 찾아내어 제거하고, 해결하는 과정이 서브스키마 통합에서 가장 중요한 과제이다. 충돌현상의 발견과 해결을 위한 통합순서는 관계가 가장 높은 순서에 의해 서브스키마간의 의미충돌 분석 및 통합으로 접근하는 것이 더 바람직하다. 따라서 그림1과 같이 친밀도 값이 가장 높은 서브스키마를 우선 충돌분석을 하고 통합한다. 그리고 부분 통합이 완료되면 다시 부분 통합된 서브스키마와 연결된 서브스키마간의 친밀도 값을 구해서 가장 높은 값을 갖는 서브스키마간에 충돌 분석과 통합을 수행한다. 이러한 과정을 모든 서브스키마가 통합될 때까지 반복 수행하여 완료한다. 의미객체 모델을 이용한 서브스키마 통합[7]에서는 서브스키마의 통합 절차를 위한 친밀도를 분석하여 서브스키마를 통합하는 절차는 친밀도 분석 단계, 충돌 발견 및 해소하는 의미 충돌 해결단계, 서브스키마를 합병하는 서브스키마 통합단계의 3단계로 구성하여 전개한다.[3][5][6][7]

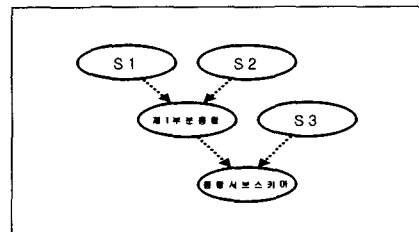


그림 1 서브스키마 통합순서

2.2 서브스키마 유사성 평가방법

이장에서는 서브스키마 유사성들의 발견을 통해서 서브스키마를 분석하기 위한 기법들을 기술하고 있는데, 개념적 서브스키마 S_i 와 S_j 사이의 유사함은 서브스키마들의 디스크립터(descriptor)의 특질들을 근거로 유사성 기준을 적용하여 평가하는데 유사성을 나타내는 특질들의 수가 많을수록 서브스키마간의 유사함이 더욱 높아진다.[1][8] 유사성 값은 단일 용어 특질인 경우 용어간의 값이 되고, 두 개 용어 특질이면 용어간의 평균값을 구하여 적용한다. 유사함 계수는 유사함(서브스키마1, 서브스키마2)로 표시하면 서브스키마의 유사함 계수는 유사성 값의 합계합수와 Dice 함수를 각각 적용해서 구한다.[1][5]

[특질 유사성 합계합수]

$$\text{유사함(서브스키마1, 서브스키마2)} = (2 \times (\text{특질간의 유사성 값의 합계})) / (\text{서브스키마의 총 특질 수})$$

[Dice 함수 적용]

$$\text{유사함(서브스키마1, 서브스키마2)} = (2 \times (\text{유사성 값이 평균 이상인 특질 수})) / (\text{서브스키마의 총 특질 수})$$

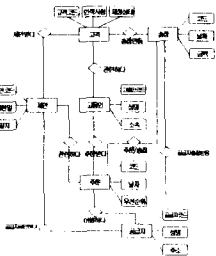


그림 2 주문관리 서브스키마1

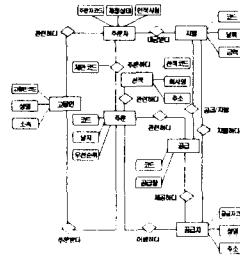


그림 3 주문관리 서브스키마2

예를 들면, 그림2와 그림3의 서브스키마1과 서브스키마2의 유사함을 조사하기 위해 가중치를 가진 특질들 리스트로 디스크립터를 표현하였다.

서브스키마의 유사함을 평가하기 위해서는 먼저 특질들의 용어들을 비교해서 특질들의 용어들을 비교해서 특질 유사성 값을 구해야 된다. 예를 들어 서브스키마1의 특질인 (송장, 금액)과 서브스키마2의 특질인 (지불, 금액)에 대한 유사성 값을 구하는 과정을 보면

-유사성(송장, 지불) = 0.4

-유사성(금액, 금액) = 1 임으로 유사성 평균값은 0.7이 된다.

유사함 계수는 유사함(서브스키마1, 서브스키마2)으로 표시할 수 있으며, 계산함수는 특질 유사성 합계합수와 Dice 함수를 각각 적용해서 구하였다.

-유사성 값 합계합수를 적용하면

$$\text{유사함(서브스키마1, 서브스키마2)} = (2 \times 8.45) / (12 + 16) = 0.6$$

-Dice 함수를 적용하면

$$\text{유사함(서브스키마1, 서브스키마2)} = (2 \times 9) / (12 + 16) =$$

0.64으로 Dice 함수를 적용한 경우 더 큰 결과 값을 얻었다. 이러한 과정을 통해서 그림4와 같은 통합된 서브스키마를 완성하게 된다.[1][8]

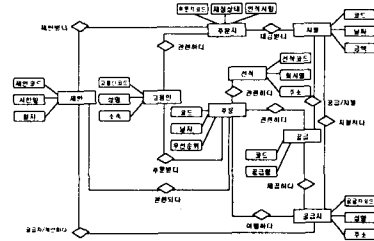


그림 4 통합된 서브스키마

3. 시소러스 구축 방법

3.1 시소러스의 개요

시소러스(thesaurus)란 용어의 사용법과 용어들 사이의 관계에 대한 정보를 제공하는 어휘 도구를 말한다. 용어의 관계성은 일반적으로 상위 개념(BT: broader term), 하위 개념(NT: Narrower Term), 용례 혹은 동의어(UF: Use For Or Synonymous), 관계어(RT: Related Term), 대체어(USE) 등으로 분류되는데, 시소러스는 이러한 관계성을 이용, 탐색시 질의에 포함된 용어의 의미를 확대하기 위해 주로 사용된다. 사용자에게 가장 유사한 혹은 직접적인 동의어나 의미를 부여하고, 이를 알파벳순으로 정렬하므로써, 작가나 시인 및 프로 그래머에게 도움이 되도록 그 용어에 가깝게 되는 알파벳순 항목들의 집합이다. 계산에 있어서 색인, 식별 혹은 분류 등에 사용되며, 데이터 저장과 데이터 은행에서 정보를 꺼내는 데 쓰이는 단어나 용어들의 집합 즉 주 용어들은 레이블, 키워드 혹은 기술부 등으로 서비스를 하며, 이러한 언급이 관련 항목 색인표를 통하여 인용될 때 매우 유용한 색인이 만들어진다. 정보 검색을 위한 키(색인)와 단어간의 관계, 즉 동의어, 하위어, 관련어 등의 관계를 나타낸 색인표인데, 특히 검색엔진에 있어 시소러스는 다음과 같은 기능을 제공한다. 시소러스는 키워드와 그 관계어간의 관련을 정리하여 놓은 일종의 사전이라 할 수 있다. 검색엔진이 시소러스를 가지고 있다면 검색식 입력시 이용자는 키워드의 어미변화나 다른 품사의 단어를 동시에 입력할 필요가 없이 원래의 찾고자 하는 단어만을 입력하면 된다.[15]

3.2 시소러스를 이용한 스키마 통합 방법

정보시스템의 재설계 단계에서 대상 서브스키마들의 통합을 위한 방법론[7]과 ER 모델로부터 스키마의 유사성을 평가[1]하기 위하여 서브스키마로부터 추출된 디스크립터들을 통합된 시소러스에 저장하고, 이를 이용해서

의미충돌 해결을 위한 서브스키마들 간의 유사성 분석에 이용한다면 보다 더 효율적인 전체 스키마의 통합을 효율적으로 지원하게 될 것이다. 따라서 다음 그림5는 여러 가지 서브스키마 통합을 위한 자료로부터 서브스키마 시소러스를 구축하고, 서브스키마 통합 방법론[7]과 서브스키마 유사성 평가방법[1]을 적용하여 전체 스키마를 통합하는 절차를 보여준다.

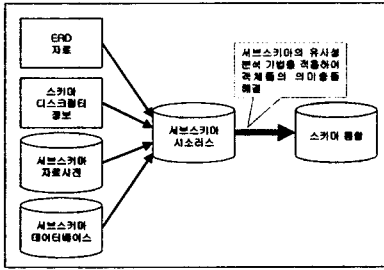


그림 5 전체 스키마 통합 절차

3.3 의미충돌 해결 알고리즘

기존의 서브스키마 통합 방법론[7]에서 의미충돌 해결을 위한 단계에서 두 서브스키마간의 이름 충돌에서 이음동의어와 동음이의어의 발견시 객체와 식별자에 의해 찾는 방법을 제안한 알고리즘을 보면 서브스키마 데이터베이스를 기반으로 충돌해결 절차를 보여주고 있는데, 본 연구에서는 그림7과 그림8은 서브스키마 시소러스를 이용한 이음동의어와 동음이의어의 충돌해결 절차를 제안한다.[7]

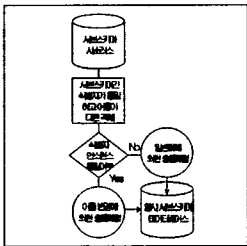


그림 6 시소러스를 이용한 이음동의어 충돌해결 절차

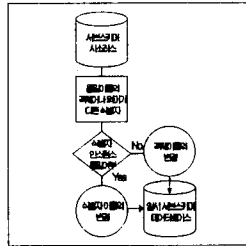


그림 7 시소러스를 이용한 동음이의어의 충돌해결 절차

그림6과 그림7은 이름 충돌해결 절차와 유사하게 구조적 충돌 분석과 카디널리티 충돌해결 절차상에서도 사전에 서브스키마의 통합을 위한 정보들이 시소러스로 구축되어 있기 때문에 서브스키마 데이터베이스를 이용하지 않고, 시소러스를 이용하여 충돌 문제를 해결하고, 임시 서브스키마 데이터베이스에 저장하여 문제 해결이 완료되면 원본 데이터베이스 스키마에 반영하여 전체 스키마의 통합을 완성하게 된다.[2][7]

4. 결론 및 향후과제

본 연구에서는 의미객체 모델을 이용한 서브스키마의 통합 방법론[7]과 개념적 스키마 분석을 위해서 서브스키마들의 디스크립터를 추출하여 기존의 정보시스템을 새로운 아키텍처로 전환하고, 새로운 애플리케이션 개발 시 사용될 참조 모델들의 라이브러리를 구축하는데, 사용하기 위한 목적의 방법론[1]에서 생성된 정보들을 포함하는 서브스키마 분석용 시소러스를 구축하는 사전 단계로서 스키마 통합을 위한 시소러스의 활용 방안과 의미충돌 해결 알고리즘을 제시하였다. 향후 과제는 정보시스템의 통합을 위한 세실계 단계에서 서브스키마의 유사성 분석을 위해서 추출된 서브스키마들의 정보 시소러스를 구축하기 위한 연구가 되어야 하겠다.

5. 참고문헌

- [1] Silvana Castano, Conceptual Schema Analysis : Techniques and Applications, ACM Trans. on Database Systems, Vol. 23, No 3, September 1998
- [2] Silvana Castano, Semantic Dictionary Design for Database Interoperability, University of Milano
- [3] C. Batini, Structuring primitives for a dictionary of entity relationship data schemas, IEEE Trans. Softw. Eng., 1993
- [4] P. Feldman, Entity Model clustering : Structuring a data model by abstraction, Comput. J., 1986
- [5] M. Batini, a Comparative Analysis of Methodologies for Database Schema Integration, ACM Computing Surveys, Vol. 18, No. 4, Dec. 1986
- [6] S. B. Navathe, R. Elmasri, J. Larson, Integrating User Views in Database Design, IEEE Computer, Jan. 1986
- [7] 임병학, 의미객체모델을 이용한 뷰 통합 지원시스템 개발, 한국과학기술원, 1996
- [8] 김홍수, 참조 스키마 생성을 위한 개념적 스키마 분석, 한국OA학회, 2002
- [9] Elmasri, Navathe, Fundamentals of Database Systems Third Edition, Addison-Wesley, 2000
- [10] Chapter 9 Automatic Indexing
- [11] Chapter 10 Audvanced Information-Retrieval Models
- [12] Chapter 3 Data Modeling Using the Entity-Relationship Model
- [13] 이재학, 데이터베이스 디자인과 모델링, 베스트북, 2002
- [14] 이춘식, 데이터베이스 설계와 구축, 한빛미디어, 2002
- [15] <http://www.helloec.net/network/thesaurus.htm>