제19회 한국정보처리학회 춘계학술발표대회 논문집 제10권 제1호 (2003. 5)

의미적 의료정보 통합을 위한 UMLS와 LOINC DB 기반의 연관 값 지식베이스 개발

김태우, 홍동완, 윤지희
한림대학교 컴퓨터공학과
{dataminer, dwhong, jhyoon}@hallym.ac.kr

Development of an Associative Value Knowledge Base based on UMLS & LOINC Database for Semantic Medical Information Integration.

Tae-Woo Kim, Dong-Wan Hong, Jee-Hee Yoon
Dept of Computer Engineering, Hallym University

요 약

최근 다양한 의료정보 시스템이 개발되어, 그 사용이 급증하고 있다. 이 들 각각의 의료정보 시스템에서 발생, 축적된 의료정보는 분산 이질의 형태를 가지며, 또한 같은 의미를 갖는 의료정보가 각기 다른 구조와 용어로 기술되어 축적되는 것이 일반적이다. 이와 같이 개별적으로 개발, 활용되어 온 의료정보를 웹 상에서 통합하여, 단일화 된 의료정보 검색 기능을 제공하기 위해서는 이들 의료정보의 의미적 연관성을 고려한 정보의 통합, 검색 기술의 개발이 필수적이다. 본 논문에서는 의미적 의료정보의 통합을 위한 UMLS와 LOINC 데이터베이스 기반의 연관 값 지식베이스의 설계 및 개발 방식을 제안한다. 웹 상에 존재하는 각종 분산 이질 형태의 의료정보는 XML을 공통 데이터 구조로 하여 통합되며, 정보 통합의 과정에서 연관 값 지식베이스를 참조하여 의미적 관련도가 높은 의료정보(구조 정보와 내용 정보)는 상호 연결되어, 진정한 의미의 정보 통합을 구현하게 된다. 지식베이스는 용어별로 식별자, 요소명, 연관값, 복수형, 동의어, 한글 이름 등의 필드를 가지며, 현재 상담, 처방, 보험, 의료용어, 증상, 임상결과 등 적용분야별로 작성된 연관값 지식베이스가 구현되어 있다.

1. 서 론

IT(Information Technology)산업의 발전과 함께 의료정보 시스템의 개발과 사용이 급증함에 따라, 의료정보가 다량으로 생성, 축적되고 있으며, 그에 따른 의료정보의 효율적 공유 및 활용에 대한 요구가 높아지고 있다. 그러나 기존의 의료정보 시스템에 구축되어 있는 환자의 기초 데이터, 의료장비에서 산출된 DICOM(Digital Imaging and Communication in Medicine)[1]자료, 의료정보 표준인 HL(Health Level)7[2] 데이터 등 각종 이질 분산 형태를 갖는 의료 정보를 웹 상에서 통합하여 단일화 된 의료정보 검색 기능을 제공하기 위하여 해결되어야 할 많은 문제점이 산재되어 있다.

웹 상에 존재하는 분산 이질 정보를 효율적으로 통합하기 위한 방법 중 하나로 웹 상의 정보 공유/교환 표준으로 자리 잡고 있는 XML(Extensible Markup Language)[3]을 사용한 미디에이터 시스템 개발에 관한 연구를 들 수 있다[4]. 본 연구실에서는 의료정보 통합을 위한 미디에이터 시스템 HMS(Hallym Mediator System)[5]를 개발 중에 있으며, HMS에서는 웹 상에 산재되어 있는 DTD/XSD 기반의 XML 문서, HL7 메시지, 객체관계형 데이터베이스 등 이질 저장 구조를 갖는 데이터를 통합하기 위한 방법으로 XML Schema를 공통 데이터 모델로 사용하며, 사용자에게 XML 기반의 공통 뷰(view) 기능과 단일화된 검색 기능을 제공하고 있다.

분산이질의 의료정보 통합을 위하여 고려되어야 할 중요한 문제점 중의 하나로 의미적 의료정보의 연관 문제를 들 수 있다. 각각 다른 목적으로 또는 다른 형태로 발생, 축적되어 온 의료정보는 일반적으로 동일한 의미를 갖는 의료정보라 하더라도 구조적, 내용적으로 각기 다른 형태로 기술되어 축적될 수 있다. 따라서 이 들 의료정보의 의미적 연관성을 고려한 정보의 통합, 검색 기술의 개발이 필요하다. 예를 들어 서로 다른 의료서비스 제공자가 입력한 의료용어 및 증상, 임상결과 등이 같은 의미를 나타내지만 다른 용어 및 표기법으로 기술될 수 있으므로, 정보를 통합하는 과정에서 이 들의 구조적 내용적 상호 연관성 관계를 추출, 해결하여야 한다. 본 논문에서는 이와 같은 문제를 해결하기 위한 방법으로서 의미적 관련이 깊은 데이터에 대한 연관성 정보를 포함하는 연관 값 지식베이스의 구축 방안을 제시하고 이를 이용한 통

본 연구는 정보통신부의 정보통신 기초기술연구지원사업(정보통신 연구진흥원)(과제번호 : C1-2002-146-0-3)으로 수행한 연구 결과임.

합 시스템 개발 사례를 보인다.

연관 값 지식베이스는 다음과 같은 두 가지 목적으로 사용된다. 첫 번째, 의료 정보의 구조적 통합을 위한 정보로 사용된다. 즉, 서로 다른 의료정보를 XML 스키마 공통 구조로 변형한 후, 이를 통합하는 과정에서 구조 정보를 나타내는 XML 태그의 엘리먼트 명 사이의 연관성을 참조하여 같은 의미를 갖는 정보가 서로 다른 구조로 표현되는 것을 방지한다. 두 번째로 의료문서의 내용 데이터에 대한 의미적 통합을 위한 기본정보로 사용된다. 즉 동일한 개념을 가진 의료정보가 의료문서마다 다른 용어로 기록되어 있다면, 연관 값 정보를 이용하여 이를 단일화하는데 사용할 수 있다.

지식베이스는 UMLS(Unified Metathesaurus Language System)[6]와 LOINC(Logical Observation Identifiers Names and Codes)[7] 데이터베이스를 활용하여 연관값 정보를 추출하였으며, 지식베이스의 각 항목은 '식별자', '요소명', '연관값', '복수형', '동의어', '한글이름' 등으로 구성된다. 지식베이스는 B+트리 인덱스 파일 구조를 가지며, 활용 분야에 따른 분야별 지식베이스로 구별되어, 현재 환자의 상담, 처방, 보험, 의료용어 및 증상, 임상결과에 대한 연관값 지식베이스가 구현되어 있다.

2. 관련 연구

다양한 의료용어 체계를 사용하고 있는 의료 환경에서는 같은 질병 또는 증상을 나타내는 용어 및 코드의 표현 범위가 다르게 정의되어 있으며, 이와 같은 표준용어의 부재는 의료기관 사이의 자료교환 및 정보공유의 근본적인 제약 원인으로 지적되고 있다. 이를 위한 다각적인 연구 및 활동이 수행되고 있으며, 국내에서도 통합용어모델(UMLS)에 한글이름과 코드를 추가하여 데이터베이스를 생성하는 작업과 이를 사용한 의료용어 통합 검색 시스템 개발에 관한 연구[9] 등이 진행되고 있다.

본 연구실에서 개발 중인 HMS(Hallym Mediator System)[5]는 가상 접근기법의 웹 정보 통합/검색 시스템으로 공통 데이터 모델로 XML을 사용한다. 이질 저장구조의 데이터 통합을 위하여 XML Schema를 공통 스키마로 사용하며, 구조 분석 및 연관성 파악을 위하여 내부적으로 공통 데이터 구조(common data structure)를 사용한다. 공통 데이터 구조는 복합 필드 구조를 갖는 트리형으로 경로정보, 연관 값, 노드 명, 노드 속성, 소스 매핑정보 등을 포함하여, 여기서 연관 값은 다른 문서와 통합 과정에서 노드 간 의미적 연관성을 파악하는데 사용된다. 즉, 연관 값 정보는 상호 연관성이 높은 노드(엘리먼트)들에게 같은 연관 값이 할당되어, 정보 통합 과정에서 이 들을 하나의 구조로 재구성하는데 사용되며, 또한 조상 노드의 연관 값 정보를 계승 받아 이름 충돌 문제를 해결하는데 사용된다.

3. 의료정보 통합을 위한 지식 구성요소

연관 값 지식베이스는 의료정보의 구조적 통합과 내용적 의미 통합을 위하여 사용된다. 지식베이스는 이와 같이 사용목적에 따라 크기 두 가지 지식베이스로 분류되며, 각각의 지식베이스는 다시 응용 분야에 따라 개별적으로 작성된다. 현재 구조 통합을 위한 지식베이스로는 상담, 처방, 보험을 위한 지식베이스가 구현되어 있으며, 내용적 의미 통합을 위해서는 의료용어 및 증상, 임상결과를 위한 지식베이스가 구현되어 있다. 지식베이스의 각 항목은 식별자, 요소 명, 연관값, 복수형, 동의어, 한글이름 등으로 이루어진다. 그림 1.에 상담정보의 구조 통합에 사용되는 지식베이스의 구성 예의 일부를 보인다. 식별자는 각 요소마다 유일한 값이 할당되며, 상호 연관이 깊거나 같은 개념의 요소에는 모두 같은 연관 값이 할당된다. 예를 들어 그림 1의 예에서 'consultant'의 식별자 E0018685는 'consultant' 요소와 같은 의미를 가진 모든 요소의 식별자를 대표하여 연관 값으로 설정된다. 복수형은 요소 명에 's'를 붙인 것이므로 요소와 같은 식별자가 할당되는 것으로 간주한다. 이와 같이 생성된 지식베이스의 연관 값은 요소 간의 의미적 연관성을 자동으로 파악하는데 사용되지만, 그러나 요소가 깊은 계층 구조로 표현되어 연관성 파악이 어려운 경우나, 요소명의 단수, 복수형을 명확히 구분하고자 하는 경우 등, 의미 통합에 문제가 발생할 수 있는 경우에는 관리자의 중간 제어가 필요하다.

ID	Element_Name	Associative_V	Plural	Synonym	Hangul
E0018685	consultant	E0018685	consultants	E0035889,E0027284,E0029663,E0101154	자문의사
E0035889	consultant physician	E0018685	consultant physicians	E0018685,E0027284,E0029663,E0101154	상담의사
E0027284	physician	E0018685	physician	E0018685,E0035889,E0029663,E0101154	내과의사
E0029663	doctor	E0018685	doctor	E0018685,E0035889,E0027284,E0101154	의사
E0101154	doctor name	E0018685	doctor names	E0018685,E0035889,E0027284,E0029663	의사이름들

그림 1. 지식베이스 구성 예

4. 연관값 추출

4.1 지식 생성 자원의 특징

지식베이스 생성을 위한 자원으로서 UMLS(Unified Metath esaurus Language System)[6]와 LOINC(Logical Observati on Identifiers Names and Codes)[7] 데이터베이스를 활용 하였다.

UMLS는 미국 NLM(National Library of Medicine)에서 장기과제로 개발하고 있는 통합용어시스템이며 지식자원으로는 SPECIALIST Lexicon, Metathesaurus, Semantic Network가 있다. SPECIALIST Lexicon 데이터베이스는 SPECIALI

ST Natural Language Processing System(NMP)[8]에 필요한 어휘정보를 제공하기 위해 개발되었고 일반 영어 단어와 생체의료 용어를 포함하고 있다. Metathesaurus는 같은 의미를 갖는 의료 용어 및 증상에 관한 정보를 포함하고 있어 의료데이터가 의미하는 범위와 의료데이터 간 포함관계를 쉽게 파악할 수 있다. 저장 구조는 문자열의 집합이 어휘를 형성하고, 어휘의 집합은 하나의 의료용어와 증상에 대한 개념을 뜻하는 메타시소러스 형식을 가진다.

LOINC 데이터베이스는 Regenstrief 연구소에서 개발하였으며 화학, 혈액학, 미생물학, 세균학 분야에서 사용하는 임상결과에 사용되는 용어에 관한 정보를 포함하고 있다. 임상결과를 reporting system과 care system 사이에 전자적으로 교환하기 위해 LOINC 코드를 사용하며, 구성요소로는

LOINC_NUM, Component, Short Common Name, Class등을 갖는다. LOINC 데이터 교환을 위한 통신 프로토콜은 HL7 표준 메시지이다. 표준 HL7 V2.x와 V3 메시지에서는 임상결과를 OBX 컴포넌트와 observation 태그정보로 사용하며, HL7 CDA(Clinical Document Architecture)에서는 임상결과를 content 태그정보에 사용한다[2].

4.2 구조 통합을 위한 지식베이스 생성

구조 통합을 위한 지식베이스는 응용 분야에 따라 개별적으로 구성된다. 각 응용 분야별로 사용되는 요소 명을 파악하기 위하여 HL7 CDA 상담, 처방문서와 HIPAA[2] 등을 사용하였다. 이 정보를 기초로 하여 각 요소와 동의어가 될 수 있는 요소 등을 정의하며, 동의어 정의에 대한 예를 그림 2.에 보인다. 그림 2.에서 'consultant'는 'consulting physician' 또는 'physician'과 'name'의 합성어 등과 같은 의미이며 대/소문자의 구별이 없이 같은 의미로 정의되어, 구체적으로는 그림 1.의 예와 같이 연관 값 정보가 설정된다.

의료문서의 요소 명으로 사용되는 의료용어와 어휘를 자동 생성하기 위하여 UMLS 소스 서버에서 제공하는 SPECIALIST Lexicon 데이터베이스를 사용한다. SPECIALIST Lexicon 데이터베이스 중 LRAGR(Lexical Agreement Inflection) 파일은 의료용어와 일반어휘에 관한 정보를 일반 사전식 표기법으로 제공하며, 파일의 구성 예는 그림 3.과 같다. 이 중 'EUI'(Entry Unique Identifier) 필드와 'STR'(String Term) 필드를 자동 추출하여 지식베이스의 식별자와 요소 명으로 사용할 수 있다. 또한 요소의 복수형은 STR 필드에서 자동 검색되어 지식베이스에 추가되며, 동의어는 관리자에 의하여 STR 필드에서 수동으로 추출되어 지식베이스에 추가된다. 한글이름 필드는 의학용어 사전 또는 표준의료 데이터를 기반으로 요소를 번역한 값으로 입력되며, 특히 한글이름에 대한 식별자는 'E' 다음에 정수 1을 추가하여 영문 이름과 구별한다.

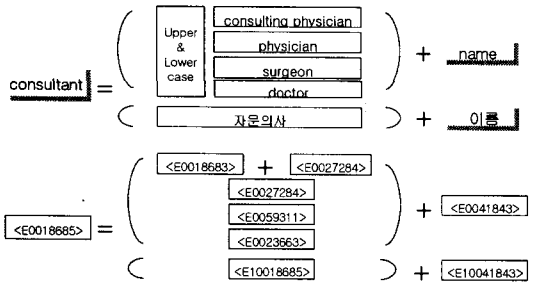

그림 2. Consultant 요소의 동의어 정의

EUI	STR	SCA	AGR	CIT	BAS
E0018685	consultants	noun	count(thr_plur)	consultant	consultant
E0018685	consultant	noun	count(thr_sing)	consultant	consultant

그림 3. LRAGR파일에서 지식으로 사용 된 필드, 실제 데이터

4.3 내용 정보 통합을 위한 지식베이스 생성

내용 정보 통합을 위한 지식베이스로서 의료용어 및 증상과 임상결과를 위한 지식베이스를 각각 구현하였다. 지식베이스 자동 생성을 위하여 UMLS의 Metathesaurus 데이터베이스와 HL7 표준메시지의 일부 전송정보로 사용되고 있는 LOINC 데이터베이스를 사용하였다.

4.3.1 Metathesaurus 데이터베이스에서의 연관 값 추출

Metathesaurus 데이터베이스는 생물의학 분야에서 각기 다른 뜻으로 사용된 용어에 대한 의미, 속성, 분류에 관한 정보를 포함하고 있다. Metathesaurus 데이터베이스 중 MRCON (Concept Names) 파일을 사용하며, 구체적인 파일 예를 그림 4.에 보인다. STR 필드의 내용을 요소 명으로 추출하며, SUI 필드를 식별자로 추출한다. STR 필드에서 추출한 요소 명의 동의어는 MRCON 파일로부터 다음과 같은 과정을 걸쳐서 추출한다. 우선, 영어로 된 용어를 선택하기 위해서는 'LAT'(Language of Term) 필드의 값이 ENG인 것을 선택한다. 다음, 우선순위가 높은 용어와 동의어에 대해 정리되어 있는 'TS'(Term Status) 필드의 값이 P(Preferred Name), p(suppressible preferred name), S(Synonym)와 s(suppressible synonym)를 선택한다. 다음, 같은 의미의 의료용어 및 증상을 추출하기 위하여 메타시소러스 방식의 CUI 필드와 LUI 필드를 사용한다. 즉 서로 다른 LUI 값을 갖는 항목이 같은 CUI 값을 가지면, 이 들 LUI 값을 갖는 항목은 같은 의미를 나타내는 정보로서 같은 연관 값을 갖도록 설정한다. 그림 4.의 파일을 메타시소러스 방식으로 표현하면 그림 5.와 같이 나타낼 수 있다.

CUI	LAT	TS	LUI	STT	SUI	STR	LRL
C0242379	ENG	P	L0024592	VO	S1917080	Neoplasm malignant:lung	3
C0242379	ENG	P	L0024592	VO	S1936615	Neoplasm malignant, lung	3
C0242379	ENG	S	L0288901	PF	S1676974	Cancer of Lung	0
C0242379	ENG	S	L0288901	VO	S0362068	Cancer, Lung	0
C0242379	ENG	S	L0288901	VO	S0678353	Lung cancer	3
C0242379	ENG	S	L0592664	PF	S0684514	Malignant tumor of lung	3
C0242379	ENG	S	L0592664	VO	S0684913	Malignant tumour of lung	3
C0242379	ENG	S	L1404832	PF	S1677109	Pulmonary Cancer	0
C0242379	ENG	S	L1404832	VWP	S1677108	Cancers, Pulmonary	0

그림 4. MRCON 파일에서 지식에 사용 된 필드, 실제 데이터

Concept (CUI)	Terms (LUIs)	Strings (SUIs)
C0242379	L0027649	S1917080 Neoplasm malignant lung
		S1936615 Neoplasm malignant, lung
	L0288901	S1676974 Cancer of Lung
		S0362068 Cancer, Lung
		S0678353 Lung cancer
	L0024585	S0684514 Malignant tumor of lung
		S684913 Malignant tumour of lung
	L1404832	S1677109 Pulmonary Cancer
		S1677108 Cancers, Pulmonary

그림 5. MRCON 파일의 Metathesaurus 표현 예

4.3.2 LOINC 데이터베이스에서의 연관 값 추출

LOINC 데이터베이스의 구성 예를 그림 6.에 보인다. LOINC 데이터베이스에는 COMPONENT 필드의 값과 관련 있는 항목을 'RELAT_NMS' 필드와 'RelatedNames2' 필드의 값으로 제공하고 있다. 'RELAT_NMS' 필드와 'RelatedNames 2' 필드의 값은 하나의 임상결과에 대해 축약어와 동의어 정보를 포함하며, 값의 구분은 세미콜론으로 되어 있다. 이들 필드 값은 Null 값을 가질 수 있으며, 두 필드의 값이 모두 지정되어 있다면 같은 의미로 해석한다. 여기에서 'RELAT_NM
S' 필드와 'RelatedNames2' 필드의 값이 요소명으로 각각 추출되며, 같은 연관값으로서 LOINC_NUM 필드 값을 할당하여 연관값 지식베이스를 자동 생성한다.

LOINC_NUM	COMPONENT	RELAT_NMS	RelatedNames2
24358-4	HEMOGRAM PANEL	Null	Cell count: PNL: BLOOD: WB: WHOLE BLOOD: ISTAT: ER IS: CBC: CBC2: CBC3: Complete Blood Count

그림 6. LOINC에서 추출한 HEMOGRAM PANEL

5. 연관값 관리

연관값 관리 엔진(Associative Value Management Engine)은 다음과 같은 3가지 구성요소로 이루어져 있으며, 연관값 지식베이스를 참조하여 관련 요소 간의 의미적 자동 연결 기능을 수행한다. 각 요소의 기능은 다음과 같다. 그림 5.는 각 구성 요소의 연계성을 보인다.

(1) 요소 추출기(Element Extractor) : 소스 스키마 추출기(Source Schema Extractor)[11]가 각 의료데이터베이스에서 추출한 XML Schema문서를 전송받아 문서상의 모든 요소명을 추출한다.

(2) 연관값 생성기(Associative Value Generator) : 추출된 요소가 기존의 연관값 지식 데이터베이스의 요소 명, 복수형, 동의어, 한글이름 중에서 하나 이상 일치할 경우 해당 연관값을 공통 뷰 관리기[12]와 소스 스키마 관리기[12]로 전송한다. 기존의 지식베이스에 요소명이 없다면, SPECIALIST Lexicon 데이터베이스에서 STR 필드의 값이 요소와 같은 문자열을 검색하여 새로운 항목을 지식베이스에 추가시켜 재사용 한다.

(3) 연관값 지식 데이터베이스(Associative Value Knowledge Database) : 지식베이스는 응용 분야별로 구성, 저장되어 있으며, 문서의 성격에 따라 해당 지식베이스가 선택, 활용된다.

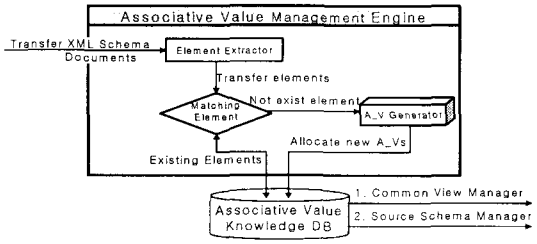

그림 7. 연관값 관리 엔진의 구조도

6. 결론 및 향후 연구과제

본 연구에서는 분산이질 의료정보의 구조적, 내용적 정보통합을 위한 연관값 지식 데이터베이스의 개발 방식을 보였다. 연관값은 통합 노드 구조에서 관계가 깊은 노드를 자동 연결하기 위한 정보로 사용되어 구조 통합에 활용되며, 같은 의미의 의료용어 및 증상, 임상결과 등의 지식은 분야별 의료정보의 내용 통합에 활용된다. 연관값 관리 엔진은 지식베이스의 연관값을 사용하여 효율적인 지식 관리기능을 제공하며, 새로운 용어에 대한 연관값 자동 생성 기능을 담당한다.

현재, 지식베이스 기능 강화를 위한 표준 인터페이스 설계 작업과 지식베이스의 성능 평가 작업을 수행 중이며, 아울러 HL7 CDA의 ADT(Admission Discharge Transmission), Orders, VHR(Virtual Hospital Round), Bill 등 표준 문서 해석을 통한 지식베이스 추가 생성 작업을 수행하고 있다.

참고문헌

[1] "Digital Imaging and Communication in Medicine(DICOM)," http://medical.nema.org/

[2] "Health Level 7," http://www.hl7.org/library/standards.cfm

[3] "Extensible Markup Language(XML),"http://www.w3.org/XML/

[4] Chaitan, B., Amarnath. G., etc. "XML-Based Information Mediation with MIX(MIX mediator system)," ACM SIGMOD International Conference no Management of Data, 1999, http://feast.ucsd.edu/Projects/MIX/

[5] 양정욱, 홍동완, 이덕형, 윤지희, "웹 정보 통합 및 검색을 위한 XML기반 미디에이터 시스템의 개발," 한국데이터베이스 학회 2001년 춘계 Conference, pp. 281-294, 2001.

[6] "Unified Metathesaurus Language System Knowledge Source Server," http://umlsks1.nlm.nih.gov

[7] "Logical Observation Identifiers Names and Codes," http://www.loinc.org

[8] "Lexical Systems Group," http://umlslex.nlm.nih.gov/

[9] 한승빈, 최진욱 "UMLS(Unified Medical Language System)기반의 정보시스템의 설계," 대한의료정보학회지, Vol. 8, 보완본 1호, pp. 120-122, 2002.

[10] 홍동완, 박진만, 김태우, 윤지희 "연관값을 이용한 XML의 공통 뷰 설계 및 구현," 정보과학회 추계 학술대회 논문지, Vol. 29, No. 2, pp. 202-204, 2002.

[11] 홍동완, 박진만, 김태우, 윤지희 "이질 저장 구조상의 의료 데이터에 대한 효율적인 검색 및 관리기법," 대한의료정보과학회지, Vol. 8, 보완본 1호, pp. 209-212, 2002.

[12] 노관준, 김태우, 박진만, 홍동완, 윤지희 "의미적 정보통합을 위한 지식기반의 의료정보 미디에이터 시스템," 한국데이터베이스 학회 2003년 춘계 Conference.