논문 2015-10-20

SNOMED CT 브라우저에서 검색 결과의 재구성 기법

(A Restructuring Method for Search Results of SNOMED CT Browser)

류 우 석*
(Wooseok Ryu)

Abstract : SNOMED CT browser is a browsing tool for searching clinical terms in SNOMED CT which is a standard terminology set used worldwide. The search result view of previous browsers merely list up candidate terminologies. The problem is that most of users become confused about how to select an appropriate term from the list. This leads serious waste of medical recoding cost. This paper discusses characteristics of SNOMED CT dataset and proposes a novel design of enhanced result view by restructuring the results using relationships of SNOMED CT concepts. Using the proposed scheme, medical doctors or officers can select appropriate terms more efficiently and can reduce overall recording time.

Keywords : SNOMED CT, Search Browser, Terminology Search, EMR

Ⅰ. 서 론

전자의무기록(Electronic Medical Record) 시스템은 환자의 호소, 검사, 진단, 치료 등 임상 진료 과정에서 발생하는 환자에 관련된 모든 자료를 진료기록지가 아닌 전산화를 통해서 기록, 관리하는 시스템이다. 작성된 진료기록은 단순 보관뿐만 아니라 경영 및 의료 연구의 목적으로 다양한 형태로 분석될 수 있다. 진료기록 정보를 서로 공유하고 교환하기 위해서는 표준화된 의학 용어 체계에 따라 기록되는 것이 필수적이다.

SNOMED CT(Systematized Nomenclature of Medicine-Clinical Terms) [1]는 표준화된 진료기록을 위해 제시되어 있는 대표적인 표준 의학용어 체계이다. 이 용어체계는 약 40만 개의 방대한 의학적 의미(Concept, 컨셉)를 포함함에 따라 이를 이용하기 위해서는 SNOMED CT 브라우저 [2, 3]와 같은 소프트웨어 프로그램을 사용해야 한다. 진료의사, 의무기록사 등이 진료기록을 작성할 때 이 브라우저를 통해서 임상적 상황에 가장 잘 부합하는 컨셉을 찾은 후 그 컨셉을 이용하여 진료기록을 작성하게 된다.

기존에 제시되어 있는 SNOMED CT 브라우저들은 용어의 검색 시 문자열 매칭을 이용한 검색 기능을 제공하고 일치하거나 유사한 컨셉들을 단순 나열하는 수준에 머무르고 있다. 의학 용어의 특성상 이름이 유사한 컨셉이 매우 많으며 그중에서는 이름이 동일한 컨셉들도 존재한다 [4]. 검색 결과가 여러 개가 조회되는 경우 검색 결과 중 의도하는 컨셉을 선택하기가 매우 어려운데, 특히 임상 상황에서는 진료 기록 시간이 촉박함에 따라 빠른 시간에 이를 선택하기가 어려운 문제가 있다.

본 논문에서는 보다 빠른 시간에 컨셉을 검색하고 진료기록 작성자가 의도하는 컨셉을 선택하기 위한 개선된 SNOMED CT 브라우저를 제안한다. 이를 위하여 본 논문에서는 사용자의 혼란을 야기하는 동일 이름의 용어들의 유형을 분석하고 브라우저가 갖추어야 할 요건을 제시한다. 그리고 단순 나열식의 브라우저에서 벗어나 검색 결과 컨셉들의 관계 정보를 이용하여 이를 효율적으로 제시하기 위한 기법과 검색 알고리즘을 제안한다.

본 논문의 구성은 다음과 같다. 2장은 관련 연구로서 기존의 SNOMED CT 브라우저들의 특성과 문제점을 제시하고 3장에서는 SNONED 용어체계의 분석을 통해 브라우저의 요구사항을 제시한다. 4장

*Corresponding Author(wsryu@cup.ac.kr)
Received: 15 Nov. 2014, Revised: 6 Jan. 2015, Accepted: 25 Feb. 2015.
W. Ryu: Catholic University of Pusan

 2015 June 10(3) 165-170
ISSN : 1975-5066
http://dx.doi.org/10.14372/IEMEK.2015.10.3.165

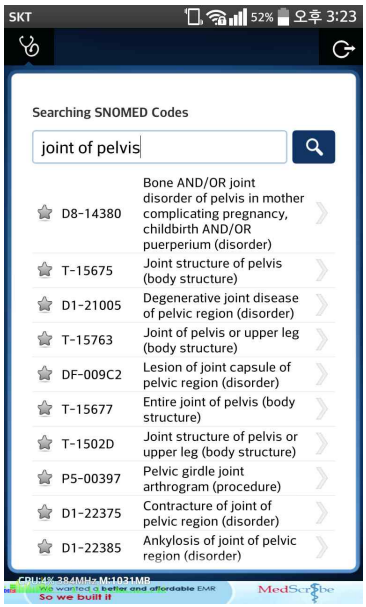

그림 1. MediCode 모바일 브라우저 안드로이드 앱 [6]

Fig. 1 MediCode SNOMED CT mobile browser android app [6]

에서는 효율적인 비교 검색을 위해 검색 결과를 화면에 표시하는 인터페이스를 제안하고 이를 구현하기 위한 알고리즘을 기술한다. 5장에서는 데이터 분석을 통한 제안한 기법의 적정성 평가를 기술하고, 마지막으로 6장에서 결론 및 향후 연구를 기술한다.

II. 관련 연구

SNOMED CT가 가지고 있는 의학용어를 검색하기 위한 SNOMED CT 브라우저는 텍스트 또는 ID를 이용하여 사용자가 원하는 컨셉을 검색하고 브라우징할 수 있는 도구이다. 이 브라우저는 PC 설치용 응용 프로그램 [3], 웹 사이트 [2, 5], 모바일 앱 [6] 등 다양한 형태로 구현 및 배포가 되어 있다. SNOMED CT 용어체계 자체가 주기적으로 업데이트되기 때문에 응용 프로그램보다는 웹 및 모바일 사이트 형태가 많이 구현되어 있다. 그림 1은 Key Management Group에서 제작 배포하고 있는 안드로이드용 앱인 MediCode의 실행 화면이다.

검색 브라우저들의 종류와 형태는 많은 반면에 기능은 서로 상당히 유사하다. 컨셉의 ID를 이미 알고 있다면 ID를 입력해서 해당하는 컨셉을 바로 찾을 수 있으며, 컨셉의 ID를 알지 못한다면 문자열 입력을 통해 해당 문자열에 대한 부분문자열 검색을 통해 용어가 일치하거나 유사한 컨셉들의 목록을 표시하고 검색 결과를 선택하면 해당 컨셉에 대한 상세 정보를 조회하는 형태이다. 모바일 검색 브라우저 역시 웹 검색 브라우저와 전체적으로 유사하나 화면 크기의 제약 때문에 보다 제한된 정보의 조회만 가능하다. 웹 검색 브라우저인 NLM SNOMED CT 브라우저 [2]가 제공하는 기능은 다음과 같이 분류할 수 있다.

- 검색 기능 : 용어 검색, 컨셉 ID 검색, 용어 ID 검색, 검색 옵션(활성 컨셉만 검색, 특정 컨셉의 하위 컨셉만 검색)
- 트리 뷰 : SNOMED CT 계층구조 조회
- 리포트 뷰 : 컨셉 상태(Activity Status), 용어(Description), 관계(Relationship) 정보, 계층구조 정보(Parent, Child, Tree Position List)

모바일 브라우저와 웹 브라우저가 가지는 공통점인 문제점은 문자열을 통해 검색을 수행할 때 입력하는 검색어에 따라 여러 결과가 서로 중복되어서 표시된다는 점이다. 그림 1에서도 하나의 용어에 대해 아주 많은 수의 검색 결과가 나타나는 것을 확인할 수 있다. 그 이유는 동일하거나 유사한 용어를 가진 컨셉이 SNOMED CT에 다수 존재하기 때문이다 [4, 7]. 여기서 문제가 발생하는데 사용자는 단순한 목록만으로는 이를 서로 구분하기 어렵다는 점이다. 검색 결과 중 의도하는 컨셉을 선택하기 위해서는 검색 결과 모두에 대해 상세 정보들을 비교, 조회해야 하는데 시간적으로도 많은 비용이 소모되며 정확하게 이해하기도 어려운 문제가 발생한다.

III. 검색 브라우저의 요구 분석

SNOMED CT 용어체계는 크게 컨셉(Concept), 용어(Description), 관계(Relation)으로 구성되어 있다. 컨셉은 하나의 의학적 의미를 가지며 식별자로서 6~19자리의 숫자형식 ID를 포함한다. 하나의 컨셉은 그 의미를 표현하기 위해 여러 개의 용어(Description)을 포함한다. 그림 2를 예를 들면 심근경색을 의미하는 컨셉 “22298006”은 그 의미의 표현을 위해 7개의 용어를 포함하고 있다. 컨셉을 표현하는 용어가 하나 이상인 것은 동일한 의미 하더라도 지역적, 언어적 특성 때문에 여러 단어들을

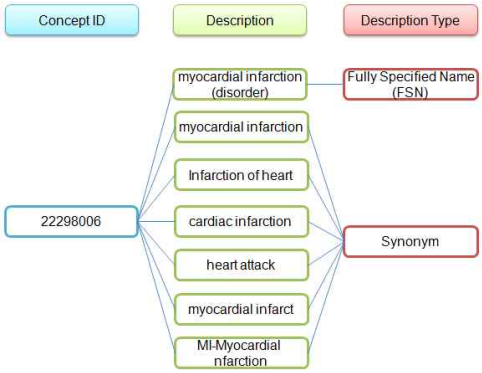

그림 2. SNOMED CT의 컨셉
Fig. 2 Concept in SNOMED CT

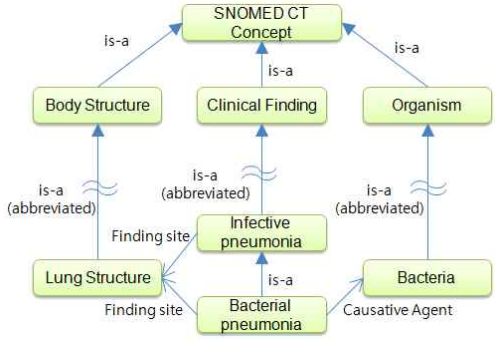

그림 3. SNOMED CT의 관계도 예시
Fig. 3 An example of relationship of SNOMED CT

사용하기 때문이다. 이때, 그 용어들들 FSN(Fully Specified Name)과 동의어(Synonym)로 구분하는데 FSN은 대표 명칭을 의미하고 동의어는 혼용하여 사용해도 무방한 용어들을 의미한다.

SNOMED CT의 컨셉들은 관계(Relationship)를 통해서 서로 유기적으로 연결되어 있다. 관계의 종류는 “finding site”,“causative agent“,”severity“ 등 매우 다양한데, 그중 대표적인 관계는 ”is-a“ 관계이다. 그림 3는 SNOMED CT의 컨셉들에 대한 관계의 예시를 도시하고 있다. 약 40만 여개의 컨셉들은 “SNOMED CT Concept”이라는 루트 컨셉에서 “is-a” 관계를 통해 계층 구조로 연결되어 있다. 그림 3에서 “Body Structure”, “Clinical Finding”, “Organism” 컨셉들은 루트 컨셉과 “is-a” 관계를 통해 직접 연결이 되어 있는데 이를 최상위 컨셉(Top-Level Concept, TLC)라고 부른다. SNOMED CT 용어체계에서는 총 19개의 활성(active) TLC가 정의되어 있다. “Body Structure”에서 “Lung Structure”는 “is-a”로 바로 연결된 부모-자식 관계가 아니라 그림의 축약을 위해 중간에 “is-a”로 연결된 컨셉들을 생략한 것이다.

SNOMED CT의 계층구조는 다중 부모노드를 허용하는 특징이 있다. 즉, 임의의 한 컨셉은 둘 이상의 컨셉으로부터 “is-a” 관계로 연결될 수 있다. 이때 두 컨셉 또한 계층구조상 동일한 위치가 아니라 임의의 위치에 있다. 즉 부모와 자식 컨셉간 관계가 N:N 관계가 있으므로 계층구조 형태로 컨셉들을 도시화하기가 매우 복잡한 특성이 있다.

SNOMED CT의 컨셉들은 그 용어가 다른 컨셉과 동일한 경우가 발생한다. 2014년 1월에 배포된 SNOMED CT 용어체계 배포버전 [8]에 따르면 298,581개의 전체 활성 컨셉(Active Concept) 중 약 8.2%인 24,462개의 컨셉은 다른 컨셉과 그 용어가 서로 동일하며, 동일한 용어를 가진 컨셉 쌍의 수는 총 12,714쌍이 존재하는 특징이 있다 [7].

SNOMED CT 브라우저의 검색 결과 화면에서 동일한 용어를 가진 컨셉들을 표시할 때 이를 단순 목록 형태가 아니라 사용자가 구분할 수 있도록 추가적인 정보와 함께 표시하여야 한다. SNOMED CT 컨셉들은 “is-a” 관계에 따라 계층적으로 서로 연결되어 있으므로, 브라우저에서는 검색 결과를 목록으로 표시하는 대신 사용자의 직관적인 이해를 위해 관계정보를 최대한 활용하여 계층 구조상의 위상(Topology)를 함께 표시하는 것이 필요하다. 이때 두 컨셉 간의 위상 정보를 모두 표현하기는 어려우므로 이를 간략화하여 표시하는 것이 필요하다. 이를 통해 사용자가 검색 결과만을 보고도 빠르게 선택할 수 있도록 하는 것이 필요하다.

Ⅳ. 검색 결과의 재구성 기법

1. 검색 결과 표시 기법

문자열 검색 후 검색 결과를 관계 정보를 이용하여 표현할 때 가장 중요한 부분은 사용자가 직관적으로 검색 결과를 비교할 수 있도록 최소한의 정보를 제공하는 것이다. 이때의 관계 정보는 “is-a” 관계를 의미하며 “is-a”를 이용하여 검색 결과를 계층 구조로 제시할 때 그 크기가 너무 크지 않도록 최소한의 정보만으로 제시되어야 한다는 점이다. 또한, 계층구조에 지나치게 그래픽 요소를 많이 넣으면 구현의 어려움이 발생하므로 가능한 그래픽 요소를 줄이는 것이 필요하다.

그림 4는 동일한 용어의 두 컨셉을 표시하기 위

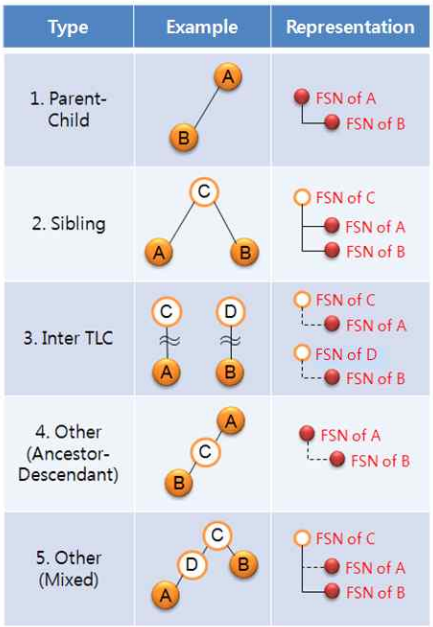

그림 4. 컨셉 쌍의 유형별 결과 표시 방법
Fig. 4 Display methods per case of two concepts with same description

해 계층구조상의 유형을 크게 다섯 가지로 구분하고 이를 간략화하여 표시하기 위한 기법을 제시한 것이다. 각각의 유형별 설명은 다음과 같다.

- 유형 1: 두 컨셉이 부모-자식 관계임을 의미한다. 이때, 목록으로 표현하는 대신 자식 컨셉인 B를 들여쓰기 하여 표현하고 실선으로 연결하면 A와 B의 관계를 직관적으로 파악할 수 있다.
- 유형 2: 두 컨셉이 형제 관계인 경우 이를 표시하기 위한 최소한의 정보로서 A와 B의 부모노드를 함께 제시함으로써 A와 B가 서로 형제라는 것을 표현한다. 단, C는 검색 결과는 아니므로 A 및 B와는 서로 다르게 표현하는 것이 필요한데 이때 글꼴을 다르게 둘 수도 있고, 그림 4과 같이 아이콘 모양을 다르게 표현할 수 있다.
- 유형 3: 두 컨셉의 최상위 컨셉이 서로 다른 경우 컨셉 A의 최상위 컨셉인 C, B의 최상위 컨셉인 D를 같이 표현하면 A와 B를 명확하게 구분할 수 있다. 이때 C와 A의 관계는 조상-손자(Ancestor-Descendant) 관계가 될 수 있으므로 이때는 유형 4의 표시방법을 따른다.
- 유형 4: 두 컨셉이 조상-손자 관계인 경우 기본적으로 손자 컨셉의 기술시 들여쓰기를 하되 유형 1과 구분하기 위하여 점선으로 표시한다. 그래픽 요소를 배제하는 경우라면 다른 특수기호를 추가하여 유형 1과 유형 4를 구분할 수 있다.
- 유형 5: 유형 1, 2, 4가 혼합된 나머지 형태를 의미한다. 이 경우 컨셉 A와 컨셉 B를 서로 구분하기 위하여 공통 조상 컨셉인 C를 함께 제시하고 A와 B는 위상관계에 따라 적절하게 표현한다.

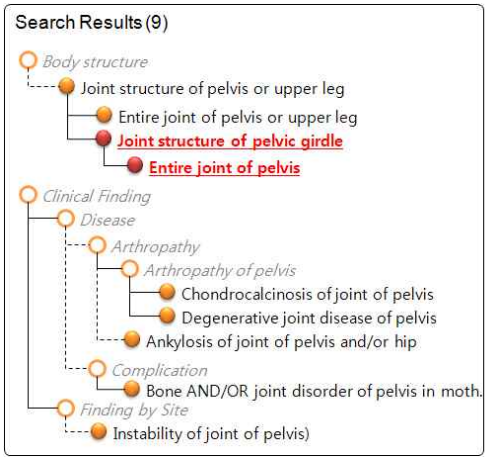

그림 5. 검색 결과의 그래프 표현 예시
Fig. 5 An example of graph representation of search result

그림 4에서 제시한 표현 기법을 적용한 예시는 그림 5과 같다. 그림 1의 NLM SNOMED CT 브라우저에서 “joint of pelvis”를 검색하면 검색 결과로 9건의 컨셉이 목록 형태로 나타나는데, 이를 그림 5과 같이 관계 구조를 함께 도시하면 크게 “Body Structure”와 “Clinical Finding” 두 최상위 컨셉으로 구분하여 도시할 수 있다. 이때 “골반대 구조”와 “전체 골반 관절”에 해당하는 두 컨셉은 검색어와 일치하는 검색 결과이므로 별도의 표시(굵게, 밑줄)을 통해 강조하며 검색어와 일치 또는 유사하지 않지만 계층관계를 위하여 표현하는 컨셉들은 글꼴 및 색상을 달리하여 표시한다. 이 기법을 적용하면 단순 검색 결과 목록과 비교하여 9개의 컨셉을 보다 직관적으로 비교할 수 있으며 결과적으로 의도하는 컨셉의 빠른 선택이 가능하게 된다.

2. 검색 결과 조회 알고리즘

검색결과 조회시 관계 정보를 추출하기 위해서는 두 컨셉이 서로 어떤 관계 유형에 해당하는지를 빠르게 검색할 수 있어야 한다. 이를 위해서는 먼저 각 컨셉에 대해서 루트 컨셉까지 도달하는 경로 정보(Transitive Closure) [9]를 구축하고, 경로에 포함된 컨셉마다 거리 정보(distance)를 추가로 구축한다. 이때 경로 정보는 자기 자신도 포함해야 하는데 이때 자기 자신과의 거리는 0으로 산정한다.

컨셉별 경로 정보가 구축이 되면 임의의 두 컨셉에 대해 경로 정보를 비교함으로써 두 컨셉의 관계 즉, 부모-자식 관계, 형제 관계, 조상-손자 관계, 혼합 관계 등을 산출할 수 있다. 이때 관계 종류의 산출 방법은 먼저 두 컨셉 각각의 경로 정보의 교집합을 수행하여 공통 조상 집합을 구한 후, 공통 조상 중 두 컨셉에 도달하는 거리의 합이 최소가 되는 조상을 구한다. 이때 구한 조상을 “최소 거리 조상”이라고 정의한다면 이를 통해 아래와 같이 유형을 판단할 수 있다.

- 유형 1: 부모-자식 관계는 최소 거리 조상이 두 컨셉 중 하나이고 거리가 1인 경우이다. 이때 최소 거리 조상이 부모 컨셉이 된다.
- 유형 2 : 최소 거리 조상이 두 컨셉의 부모 컨셉인 경우이다.
- 유형 3 : 최소 거리 조상이 루트 컨셉인 경우이다. 이때 각 컨셉의 경로 정보에서 최상위 컨셉의 추출이 가능하다.
- 유형 4 : 최소 거리 조상이 두 컨셉 중 하나이고 거리가 2 이상인 경우이다. 이때 최소 거리 조상이 조상 컨셉이 된다.
- 유형 5 : 위 네 가지 유형에 해당하지 않는 경우이다. 이때 최소 거리 조상과 두 컨셉의 거리를 각각 계산하여 유형 5의 세부 유형을 결정할 수 있다.

V. 재구성 기법의 적정성 평가

본 논문에서 제안한 검색 결과 재구성 기법은 사용자의 편의를 위해 검색 결과에 일부 정보를 추가적으로 제공하는 특성이 있다. 이때 추가되는 정보가 지나치게 많다면 사용자의 혼란이 가중되어 오히려 검색의 편의성이 떨어지게 된다. 이 장에서는 본 기법을 적용할 때 추가적으로 제공되는 정보의 양을 산출하고 이를 비교 분석함으로써 제안한 기법의 적정성을 평가한다.

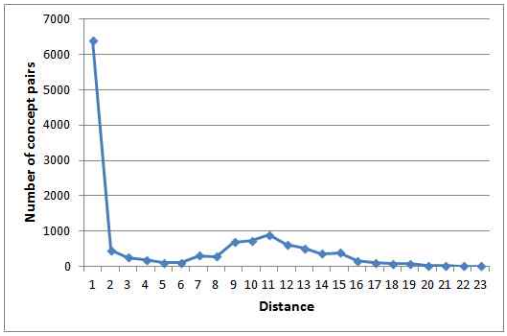

그림 6. 용어가 동일한 두 컨셉간 거리
Fig. 6 Distances of two concepts with same descriptions.

표 1. 용어가 동일한 컨셉의 유형별 건수
Table 1. Case analysis of two concepts with same description

Type	Distance	Count	Ratio
1. Parent-Child	1	6,380	50.2%
2. Sibling	2	383	3.0%
3. Inter TLC	5~23	5,084	40.0%
4. Anc.-Des.	2~4	92	0.7%
5. Mixed	3~16	775	6.1%
Total		12,714	100.0%

데이터 분석을 위한 시스템 환경은 다음과 같다. 2014년 1월에 배포된 SNOMED CT 용어체계 배포버전 [8]을 MySQL Database에 구축하고, 동일한 명칭을 가지는 모든 컨셉들에 대해 거리 정보를 추가로 계산한 후 이를 분석하였다.

그림 6은 동일한 용어를 가진 12,714건의 컨셉 쌍에 대해 컨셉 간 거리를 그래프로 표현한 것이다. 그림에 따르면 거리가 1인 컨셉 쌍이 6,380건이며, 거리가 증가함에 따라 컨셉 쌍의 수가 급격히 증가하나 거리가 7~15인 컨셉 쌍의 수는 소폭 증가하는 특성이 있다. 만일 용어가 동일한 두 컨셉을 표시하기 위해 두 컨셉 간의 계층구조를 모두 표현한다면 거리+1개 만큼의 컨셉들이 표현되어야 하므로 분석 결과 평균 6.55개로 산출된다. 이는 동일한 용어의 두 컨셉을 검색 결과로 표시할 때 평균 4.55개의 컨셉을 추가로 표시해야 함을 의미한다.

표 1은 4장의 재구성 기법에서 제안한 유형별로

컨셉 쌍의 수를 분석한 결과이다. 제안한 재구성 기법에 따르면 유형 2와 유형 5의 경우 공통부모(또는 조상)의 표현을 위해 한 개의 컨셉을 추가로 표현해야 하며 유형 3의 경우 최상위 컨셉의 표현을 위해 두 개의 컨셉을 추가로 결과에 표시해야 한다. 이를 적용하여 검색 결과에 표시해야 하는 컨셉의 수를 계산하면 평균 2.89개로 산출된다. 즉, 1개 미만(0.89개)의 컨셉 만을 추가적으로 표시하는 것을 의미한다. 이는 계층구조상의 모든 경로정보를 표현하는 것과 비교하여 약 20%의 정보만을 이용하므로 비교적 적은 추가정보만으로 사용자에게 검색 편의를 제공할 수 있음을 확인할 수 있다.

Ⅵ. 결 론

SNOMED CT 검색 브라우저는 SNOMED CT 용어체계에 포함된 방대한 양의 컨셉을 검색하는 프로그램이다. 그런데, 이 프로그램은 용어체계의 복잡성으로 인해 동일하거나 유사한 검색결과가 단순한 목록으로 표시되어 진료기록의 작성자가 원하는 컨셉을 빠르게 선택하기가 어려운 문제가 있다. 본 논문에서는 이 문제를 해결하기 위하여 데이터 분석을 통해 요구사항을 도출하고 검색 결과 컨셉들에 대해 컨셉 간 관계 정보를 이용하여 효과적으로 표시하는 기법을 제시하였다. 또한 컨셉 간 관계 유형을 빠르게 계산하기 위한 알고리즘을 함께 제시하였다. 향후 연구로 본 논문에서 제안된 표시 기법과 알고리즘을 임상 환경에 적용 후 임상 실험을 통해 실 사용자 관점에서의 적정성과 유용성을 평가하는 것이 필요하다.

References

[1] V.B. Pinto, C.R.O. Rabelo, I.P.T. Girao, “SNOMED-CT as Standard Language for Organization and Representation of the Information in Patient Records,” Knowledge Organization, Vol. 41, No. 4, pp. 311-318, 2014.

[2] U.S. National Library of Medicine, "UMLS SNOMED CT Browser," Online access http://uts.nlm.nih.gov/ snomedctBrowser.html

[3] The Clinical Information Consultancy Ltd., “CliniClue Xplore,” http://www.cliniclue.com.

[4] S. Lusignan, T. Chan, S. Jones, “Large complex terminologies: more coding choice, but harder to find data - reflections on introduction of SNOMED CT (Systematized Nomenclature of Medicine - Clinical Terms) as an NHS standard,” Informatics in primary care, Vol. 19, No. 3, pp. 3-5, 2011.

[5] International Health Terminology Standards Development Organisation, “The IHTSDO SNOMED CT Browser,” http://browser.ihtsdotools.org.

[6] Key Management Group, “MediCode,” http://play.google.com/store/apps/details?id=kmg.android.medicode.

[7] W. Ryu, “Requirement Analysis of Search Browser for Efficient Searching of Clinical Terminology,” Journal of the Korea Institute of Information and Communication Engineering, Vol. 18, No. 11, pp. 2691-2696, 2014 (in Korean).

[8] U.S. National Library of Medicine, “SNOMED CT Release Files”, Online access http://www.nlm.nih.gov/research/umls.

[9] IHTSDO, "SNOMED CT Technical Implementation Guide," http://www.snomed.org

Wooseok Ryu (류 우 석)

He received the Ph.D. degree in department of Computer Engineering from Pusan National University, Busan, Republic of Korea, in 2012. In March 2013, he joined Catholic University of Pusan, Busan, Republic of Korea, where he is currently an assistant professor. His research interests include health informatics, medical terms, and health big data analysis.

Email: wsryu@cup.ac.kr