Regular paper | Journal of the KIECS. pp. 295-300, vol. 19, no. 1, Feb. 29. 2024, t. 123, pISSN 1975-8170 | eISSN 2288-2189
http://dx.doi.org/10.13067/JKIECS.2024.19.1.295

SNOMED CT 용어체계에서 형제 노드의 유사도 분석 기법

류우석*

Similarity Analysis of Sibling Nodes in SNOMED CT Terminology System

Woo-Seok Ryu*

요 약

본 논문에서는 SNOMED CT 용어체계가 가지는 불완전성을 논의하고 이를 유지하는 방법으로 형제 노드 간 유사성을 평가하는 지표를 제안한다. SNOMED CT는 방대한 양의 의학용어를 포함하고 있으나 계층구조 내 컨셉의 누락 등 온톨로지의 불완전성 문제가 존재한다. 누락된 컨셉 발견을 위해 다수의 노드로 구성된 형제 노드 그룹 내에서의 노드 간 유사도 평가를 위한 지표를 제안하고 유사도가 낮은 그룹을 도출하였다. 2023년 3월 SNOMED CT 국제 배포판에 적용하여 형제 노드 그룹들의 유사도를 분석한 결과 임상적 발견 컨셉의 하위 컨셉들 중 2개 이상의 형제 노드를 가진 29,199개의 형제 노드 그룹의 평균 유사도는 0.81로 나타났다. 반면, 유사도가 가장 낮은 그룹은 중독 컨셉의 자식 컨셉으로 그 유사도는 0.0036으로 확인되었다.

ABSTRACT

This paper discusses the incompleteness of the SNOMED CT and proposes a noble metric which evaluates similarity among sibling nodes as a method to address this incompleteness. SNOMED CT encompasses an extensive range of medical terms, but it faces issues of ontology incompleteness, such as missing concepts in the hierarchy. We propose a noble metric for evaluating similarity among nodes within a node group, composed of multiple sibling nodes, to identify missing concepts, and identify groups with low similarity. Analyzing the similarity of sibling node groups in the March 2023 international release of SNOMED CT, the average similarity of 29,199 sibling node groups, which are sub-concepts of the clinical finding concept and are consist of two or more sibling nodes, was found to be 0.81. The group with the lowest similarity was associated with child concepts of poisoning, with a similarity of 0.0036.

키워드

Incompleteness, Missing Concept, Sibling Node, Similarity, SNOMED CT
불완전성, 컨셉 누락, 형제 노드, 유사도, SNOMED CT

* 교신저자 : 부산가톨릭대학교 병원경영학과
• 접 수 일 : 2023. 11. 18
• 수정완료일 : 2023. 12. 31
• 게재확정일 : 2024. 02. 17
• Received : Nov. 18, 2023, Revised : Dec. 31, 2023, Accepted : Feb. 17, 2024
• Corresponding Author : Woo-Seok Ryu
Dept. of Health Care Management, Catholic University of Pusan,
Email : wsryu@cup.ac.kr

Ⅰ. 서 론

진료, 처치 등 의료 기록 및 통계에 사용되는 표준 용어체계는 정확하고 효과적인 의사소통, 의사결정 및 진료의 연속성을 지원하는 것에 핵심적인 역할을 수행하고 있다. 그 중에서 SNOMED CT는 방대한 의학적 의미를 기반으로 환자 정보의 체계적 기록 및 활용을 위해 전 세계적으로 널리 사용되고 있으며, 국내에서도 그 활용 범위를 넓혀 가고 있다[1].

SNOMED CT 용어체계는 온톨로지 구조로 정의되어 있으며 임상적 의미를 가지는 컨셉(Concept), 그 의미를 문자로 표현하기 위한 설명(Description), 의미 간의 연관성을 표현하기 위한 관계(Relationship)를 포함하고 있으며, 그 중 관계는 컨셉 간 계층 구조를 기술하기 위한 하위형식 관계와 컨셉의 특성을 기술하기 위한 속성 관계로 구분되어 있다[2]. 현재의 SNOMED CT는 수십만 개의 컨셉이 수백만 개의 관계로 매우 복잡하게 연결되어 있음에 따라 온톨로지의 완전성과 일관성을 유지하는 것은 용어체계의 체계적인 활용을 위한 품질 보증(Quality Assurance) 측면에서 매우 중요한 문제이다[3].

본 연구의 선행 연구로서 SNOMED CT 온톨로지 구조의 완전성과 일관성을 유지하기 위한 연구는 다양한 측면에서 이루어지고 있다. 하위형식 관계에서 누락된 컨셉을 찾기 위해 온톨로지 내 비격자(non-lattice) 그래프를 탐지하는 기법이 연구되었으며[4][5], 계층 구조의 컨셉들이 가지는 설명들에 대한 어휘 분석(Lexical Analysis)을 통해 누락된 컨셉을 찾는 기법이 연구되었다[6][7]. 또한, 보건의료 분야에 확대되고 있는 머신러닝 기법을 이용한 연구로[8][9], 딥러닝을 통해 유사 어휘 패턴을 가지는 컨셉을 탐지하기 위한 연구가 제안되었다[10].

본 연구에서는 계층 구조에 따른 그래프의 특성 또는 어휘 특성을 분석하여 컨셉과 컨셉 사이에 누락된 컨셉을 찾는 기존의 방법 대신, 하나의 컨셉에 하위형식 관계로 연결된 다수의 자식 컨셉(Child Concept)들을 분석하여 여러 개의 그룹으로 분리하는 방법을 제시하고자 한다. 이를 위해 자식 컨셉들 간의 유사성을 비교하기 위한 유사도 지표를 제안하고, 이질적인 자식 컨셉으로 구성된 노드 그룹을 제시하여 이를 여러 서브 그룹으로 분리할 수 있도록 하는 것이 본 연구의 목적이다.

본 연구의 흐름은 다음과 같다. 먼저 2장에서는 SNOMED CT 용어체계의 특성을 제시하고, 3장에서는 하나의 컨셉의 자식 컨셉들로 구성된 형제 노드 그룹의 유사도를 계산하기 위한 유사도 지표를 제안한다. 4장에서는 제안한 유사도 지표를 적용하여 유사도가 낮은 컨셉 그룹을 도출하고 5장에서 결론 및 향후 연구를 기술한다.

Ⅱ. SNOMED CT 용어체계

SNOMED CT 용어체계는 고유의 임상적 의미를 컨셉(Concept)으로 정의하고 숫자 형태의 식별자인 컨셉 코드(Concept ID)로 이를 서로 구분한다. 그리고, 이 컨셉은 그 의미를 텍스트로 표현하기 위해 여러 설명(Description)들을 가지고 있다. 컨셉은 관계(Relationship)을 통해 여러 다른 컨셉과의 연관성을 표현하고 있는데, 이 관계는 크게 하위형식(IS-A) 관계와 속성(Attribute)관계로 구분된다. 모든 컨셉들은 하위형식 관계로 계층적으로 연결되어 있는데, 최상위 컨셉인 *"SNOMED CT Concept(138875005)"*에서 계층 구조로 이어진다. 이때의 계층구조는 다중 부모를 허용하는 것이 특징이다.

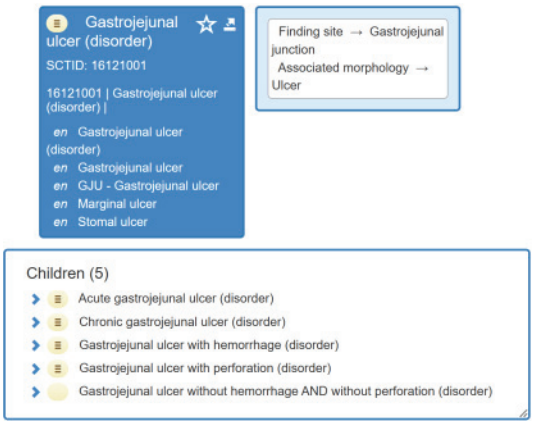

그림 1. SNOMED CT 컨셉의 하위 컨셉 및 속성 예시(SNOMED CT 브라우저)
Fig. 1 An example of SNOMED CT Concept with children and attributes(SNOMED CT browser)

속성 관계는 하나의 컨셉의 특성을 다른 컨셉을 통해 연관시킴으로써 해당 컨셉의 의미를 보다 명확하게 하기 위해 사용된다. 그림 1은 SNOMED 인터네셔널에서 제공하는 SNOMED CT 브라우저[1]에서 위십이지장궤양 컨셉(*Gastrojejunal ulcer, 16121001*)을 조회한 화면이다. 이 컨셉에는 두 가지 속성이 정의되어 있는데 발견 위치(Finding site) 속성의 값은 위십이지장 교차부(Gastrojejunal junction)이고, 연관 형태(Associate Morphology) 속성의 값은 궤양(ulcer)으로로 연관시킴으로써 두 속성을 통해 위십이지장궤양 컨셉의 의미를 보다 명확하게 정의하고 있다. 그리고, 이 컨셉의 자식 컨셉으로 5개의 컨셉이 정의되어 있는데 본 논문에서는 하나의 컨셉에 연결된 자식 컨셉들을 형제 노드 집합으로 정의한다.

III. 형제 노드 그룹의 유사도 정의

이 장에서는 형제 노드 간 유사성과 이질성을 평가하기 위한 지표로서 하나의 형제 노드 그룹 내에 있는 여러 형제 노드 간의 유사도 지표를 제안하고자 한다. 형제 노드들은 동일한 부모 노드들을 가지고 있으며, 각각 여러 속성들을 통해 해당 컨셉의 의미를 구체화하고 있으므로 컨셉들이 동일한 속성으로 정의되어 있을수록 두 컨셉의 의미가 서로 유사하게 정의되어 있다고 가정한다.

본 연구에서 형제 노드간 유사도를 정의하기 위해 형제 노드들이 가지고 있는 공통 속성의 여부를 이용한다. 보다 구체적으로 공통 속성의 개수를 이용한 노드 간 해밍 거리(Hamming distance)를 통해 유사도를 정의한다. 해밍 거리는 두 이진 벡터 간의 차이를 나타내는데 유용한 거리 측정 방법이며, 노드의 속성 타입 여부를 이진화하여 적용이 가능하다. 즉, 컨셉 i가 가지고 있는 속성의 집합을 A_i라고 할 때 컨셉 i와 컨셉 j간의 유사도 s_{ij}는 식(1)과 같이 정의할 수 있다. 이때, $n(A_i)$는 컨셉 i가 가지고 있는 속성 타입의 개수를 의미한다. 이에 유사도 값은 0에서 1 사이의 실수로 표현될 수 있다.

$$s_{ij} = n(A_i \cap A_j) / n(A_i \cup A_j) \quad (1)$$

위에 정의된 두 노드간의 유사도 정의를 확장하면 특정 컨셉과 하위형식 관계로 연결된 형제 노드 그룹의 유사도를 계산할 수 있다. n개의 형제 노드들로 구성된 형제 노드 그룹에 대한 유사도 s_n은 그룹 내 컨셉 쌍 별 유사도 전체의 합을 가능한 조합 수로 나눌 수 있으며 식(2)와 같이 정의한다.

$$s_n = \sum N(A_i \cap A_j) / ({}_nC_2 \times N(A_1 \cup \ldots \cup A_n)) \quad (2)$$

형제 노드 집합의 유사도는 집합에 포함된 각 컨셉 쌍에 대한 유사도 값들의 단순 평균과 차이가 있다. 집합에 포함된 각 컨셉들의 속성이 서로 다를 수 있으므로 집합에 포함된 컨셉들의 모든 속성을 기준으로 각 컨셉 간 해밍 거리를 산출해야 한다. 이를 통해 유사도가 높은 형제 노드 집합과 그렇지 않은 형제 노드 집합을 구분할 수 있다.

IV. 유사도 분석

4.1 분석 방법

본 연구에서 제안한 형제 노드 그룹의 유사도를 이용하여 분석을 수행하기 위해 NIH에서 제공하는 SNOMED CT 배포판을 이용하였다[2]. 본 연구에서 사용한 버전은 2023년 3월 31일에 배포된 국제판(International Release of SNOMED CT)이며, 최상위 컨셉인 *"SNOMED CT Concept"*의 19개 하위 컨셉 중 가장 많은 수의 하위 컨셉을 가진 임상적 발견(*Clinical finding (finding),* 404684003) 컨셉의 하위 컨셉들을 대상으로 유사도 평가를 수행하였다. 이 컨셉의 하위 컨셉들은 질병, 진단 등 의료 및 임상정보 시스템의 다양한 임상적 발견과 관련된 정보를 기술하는 데 사용되는 컨셉이다. MySQL 8.0을 이용하여 배포판을 데이터베이스에 적재시키고 임상적 발견 컨셉의 하위 컨셉들에 대해 모든 형제 노드 그룹들을 생성하고 유사도 계산을 수행하였다.

임상적 발견 컨셉의 하위 컨셉들에 대한 특성은 표

1) https://browser.ihtsdotools.org/

2) https://www.nlm.nih.gov/healthit/snomedct

1과 같다. 전체 컨셉의 개수는 118,729개이며, 형제 노드 그룹의 개수는 총 42,620개로 나타났다. 이때, 형제 노드가 두 개 이상의 컨셉으로 구성된 그룹의 수는 29,199개로 나타났다. 형제 노드 그룹들에 포함된 컨셉의 수는 평균 5.28개, 표준편차는 16.75로 나타났는데, 그 최소값은 1, 최대값은 1,372로 나타났다.

표 1. 임상적 발견 컨셉의 하위 컨셉들의 특성
Table 1. Features of child concepts of clinical finding

Features	Value
the number of concepts	118,729
the number of sibling node groups	42,620
the number of groups with two or more sibling nodes	29,199
average and standard deviation of the number of nodes in groups	5.28±16.75
the node count range in groups	[1, 1,372]

4.2 유사도 분석 결과

두 개 이상의 형제 노드로 구성된 형제 노드 그룹(n=29,199)의 유사도 값을 측정한 결과의 분포는 그림 2와 같다. 전체의 약 57.0%에 해당하는 16,640개의 노드 그룹은 유사도가 0.9이상으로 나타났으며, 0.1로 나눈 유사도 구간에서 가장 높은 비율을 보였다. 하지만, 유사도가 0.5이하인 노드 그룹도 3,764개(12.9%)로 나타났으며, 유사도가 0.1 이하인 노드 그룹은 78개로 나타났다.

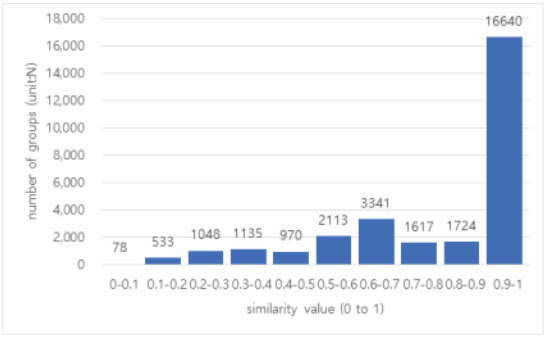

그림 2. 유사도 값에 따른 형제 노드 그룹의 분포
Fig. 2 Histogram of sibling node groups by similarity

표 2는 노드 그룹의 유사도 값을 형제 노드의 개수별로 제시한 결과이다. 유사도 값의 전체 평균은 0.81로 나타났으나, 노드 그룹에 포함된 형제 노드들의 개수가 늘어날수록 유사도는 감소하는 것이 확인되었다. 그리고, 전체 노드 그룹의 5.4%에 해당하는 형제 노드 수가 20개를 초과하는 경우(243개)의 평균 유사도는 0.39에 불과하였다.

표 2 . 형제 노드 수 별 그룹 개수 및 유사도 평균
Table 2. Features of child concepts of clinical finding

the number of nodes in a group	count of groups (n, %)	average similarity
2	8,860 (30.3%)	0.93
3	4,975 (17.0%)	0.87
4	3,485 (11.9%)	0.85
5	2,577 (8.8%)	0.83
6-10	5,209 (17.8%)	0.74
11-20	3,743 (8.7%)	0.61
>20	243 (5.4%)	0.39
Total	29,199 (100%)	0.81

표 2에 제시된 형제 노드 그룹에 포함된 형제 노드 수에 따른 그룹 개수 및 유사도 평균을 형제 노드 수가 20개 이하일 때(전체의 94.6%)와 20개 초과일 때(전체의 5.4%)로 나누어 그림으로 도시한 결과는 각각 그림 3과 그림 4와 같다. 그림 3에서 보는 바와 같이 그룹에 포함된 노드의 개수가 늘어나는 만큼 유사도는 약 0.5까지 선형적으로 감소하는 것을 보이고 있으나, 그림 4에서 그룹 내 노드 개수가 20개를 초과할 때는 노드 개수가 증가하는 것과 별개로 유사도는 매우 불규칙하게 변화하는 것을 확인할 수 있다. 즉, 노드 개수가 20개를 초과하는 형제 노드 그룹에서는 형제 노드 수 증가에 따라 유사도가 선형적으로 감소하지 않으며, 노드 그룹별로 다른 특성을 가지는 것을 확인할 수 있다. 예를 들어, 유사도가 1이면서 노드 개수가 가장 많은 형제 노드 그룹은 약물에 의한 의도적인 중독(*Intentional poisoning caused by drug (disorder)*, 431307001)의 자식 컨셉들로 총 418개의 형제 노드를 가지고 있으나, 해당 컨셉들 모두 "*Causative agent*"와 "*Due to*" 두 개의 속성만을 가지고 있는 것으로 확인되었다.

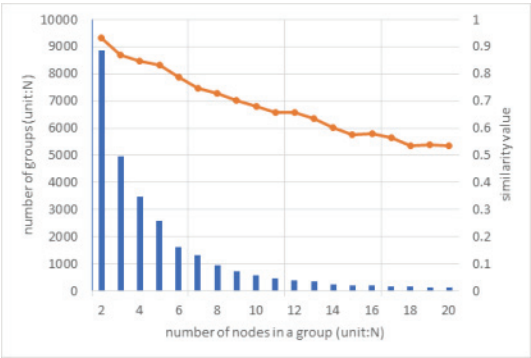

그림 3. 형제 노드 수 별 그룹의 개수 및 평균 유사도 (형제 노드 수 20개 이하)
Fig. 3 the number of groups and average similarity by sibling node count (Lees than 20 sibling nodes)

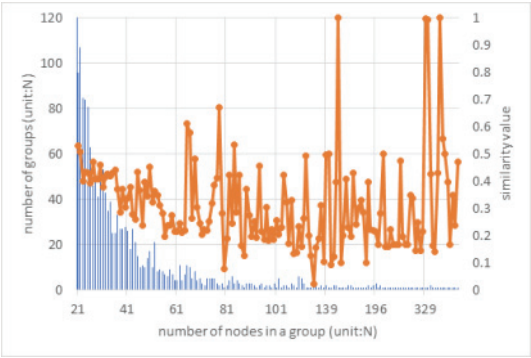

그림 4. 형제 노드 수 별 그룹의 개수 및 평균 유사도 (형제 노드 수 20개 초과)
Fig. 4 the number of groups and average similarity by sibling node count (Exceeding 20 sibling nodes)

전체 형제 노드 그룹을 분석한 결과 유사도가 가장 낮은 하위 5개의 그룹은 표 3에 제시한 내용과 같다. 그 중에서도 가장 유사도가 낮은 노드 그룹은 중독(*Poisoning (disorder)*, 75478009) 컨셉의 자식 컨셉들로 구성된 형제 노드 그룹으로 125개의 컨셉 노드로 구성되어 있으며 유사도 값은 0.0036으로 나타났다. 이때 125개의 컨셉들을 정의하는 속성의 총 개수는 11개이나, 개별 컨셉의 속성 수의 평균은 1.14개에 불과할 만큼 형제 노드들이 서로 다른 속성으로 정의되어 있음을 확인한 만큼 서브 그룹으로의 분리를 고려할 수 있는 것으로 확인되었다.

표 3 . 유사도 하위 5개 노드 그룹
Table 3. Features of child concepts of clinical finding

Parent concept	similarity	# of siblings
Poisoning	0.0036	125
Occupational disorder	0.0096	13
Pain	0.0096	26
Megaloblastic anemia	0.0104	12
Psyhotic disorder	0.0110	13

V. 결 론

본 논문에서는 SNOMED CT 용어체계에서 형제 노드들 간의 유사성을 평가하기 위한 유사도 지표를 제안하였다. 그리고, 제안한 지표를 임상적 발견 컨셉의 하위 컨셉들에 적용하여 형제 노드 수에 따른 유사도 변화를 분석하고 가장 유사도가 낮은 형제 노드 그룹을 도출하였다. 본 연구에서 제안한 유사도는 형제 노드 그룹에 포함된 컨셉들 간에 공유하는 속성의 수를 기반으로 하므로 유사도가 낮은 그룹은 보다 유사도가 높은 서브 그룹으로 분할할 수 있으며 이를 통해 용어체계의 완전성을 더욱 높일 수 있을 것으로 판단된다. 본 연구의 향후 연구로서 유사도가 낮은 형제 노드들을 대상으로 최적의 서브 그룹으로의 분할을 통해 용어 체계의 계층 구조를 보다 명확하게 정의할 수 있는 알고리즘을 개발할 필요가 있으며, 이때 서브 그룹으로의 분할을 위한 누락된 컨셉의 자동화된 생성 또한 필요하다.

감사의 글

이 논문은 2021년도 부산가톨릭대학교 교내연구비에 의하여 연구되었음

References

[1] H. Park, S. Yu, and H. Jung, “Strategies for Adopting and Implementing SNOMED CT in Korea,” *Healthcare Informatics Research*, vol. 27,

no. 1, Jan. 2021, pp. 3-10.
[2] W. Del-Pinto, R. A. Schmidt, and Y. Gao, "Extracting Subontologies from SNOMED CT," In *European Semantic Web Conf.*, Crete, Greece, May 2022, pp. 291-294.
[3] X. Hao, R. Abeysinghe, K. Roberts, and L. Cui, "Logical definition-based identification of potential missing concepts in SNOMED CT," *BMC Medical Informatics and Decision Making*, vol. 23, no. 1, May 2023, pp. 1-18.
[4] L. Cui, W. Zhu, S. Tao, J. T. Case, O. Bodenreider, and G. Q. Zhang, "Mining non-lattice subgraphs for detecting missing hierarchical relations and concepts in SNOMED CT," *J. of the American Medical Informatics Association*, vol. 24, no. 4, July 2017, pp. 788-798.
[5] X. Hao, R. Abeysinghe, F. Zheng, and L. Cui, "Leveraging non-lattice subgraphs for suggestion of new concepts for SNOMED CT," In *2021 IEEE Int. Conf. on Bioinformatics and Biomedicine*, San Diego, USA, Dec. 2021, pp. 1805-1812.
[6] F. Zheng, J. Shi, and L. Cui, "A lexical-based approach for exhaustive detection of missing hierarchical IS-A relations in SNOMED CT," In *American Medical Informatics Association Annual Symp. Proc.*, Virtual Event, Nov. 2020, pp. 1392-1401.
[7] X. Hao, R. Abeysinghe, J. Shi, and L. Cui, "A substring replacement approach for identifying missing IS-A relations in SNOMED CT," In *2022 IEEE Int. Conf. on Bioinformatics and Biomedicine*, Las Vegas, USA, Dec. 2022, pp. 2611-2618.
[8] E. Lim, "IoB Based Scenario Application of Health and Medical AI Platform," *J. of the Korea Institute of Electronic Communication Sciences*, vol. 17, no. 6, Dec. 2022, pp. 1283-1292.
[9] H. Sim and H. Kim, "Development of Type 2 Prediction Prediction Based on Big Data," *J. of the Korea Institute of Electronic Communication Sciences*, vol. 18, no. 5, Oct. 2023, pp. 999-1008.
[10] R. Abeysinghe, F. Zheng, E. V. Bernstam, J. Shi, O. Bodenreider, and L. Cui, "A deep learning approach to identify missing is-a relations in SNOMED CT," *J. of the American Medical Informatics Association*, vol. 30, no. 3, Mar. 2023, pp. 475-484.

저자 소개

류우석(Woo-Seok Ryu)

1997년 부산대학교 컴퓨터공학과 졸업 (공학사)
1999년 부산대학교 대학원 컴퓨터공학과 졸업(공학석사)
2012년 부산대학교 대학원 컴퓨터공학과 졸업(공학박사)
2013년~현재 부산가톨릭대학교 병원경영학과 부교수
※ 관심분야 : 의료정보, 빅데이터, 병렬분산 처리, 머신러닝