Enhancing the Quality of Hierarchical Relations in the NCI Thesaurus to Enable Faceted Query of Cancer Registry Data

Licong Cui[1*], Rashmie Abeysinghe[1,2], Fengbo Zheng[1,2], Shiqiang Tao[6], Ningzhou Zeng[2], Isaac Hands[4], Eric B. Durbin[3,4], Lori Whiteman[5], Lyubov Remennik[5], Nicholas Sioutos[5], Guo-Qiang Zhang[1,6]

[1]School of Biomedical Informatics, University of Texas Health Science Center at Houston, Houston, TX

[2]Department of Computer Science, University of Kentucky, Lexington, KY

[3]Division of Biomedical Informatics, Department of Internal Medicine, University of Kentucky, Lexington, KY

[4]Kentucky Cancer Registry, Lexington, KY

[5]Enterprise Vocabulary Services, Center for Biomedical Informatics & Information Technology, National Cancer Institute

[6]Department of Neurology, McGovern School of Medicine, University of Texas Health Science Center at Houston, Houston, TX

***Corresponding author**: Licong Cui, 7000 Fannin Street, Suite 600, Houston, TX 77030. Telephone: 713-500-3791. Email: licong.cui@uth.tmc.edu

Funding: This work was supported by the National Science Foundation (NSF) through grant IIS-1931134, and the National Institutes of Health (NIH) National Cancer Institute through grant R21CA231904. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NSF or NIH.

ABSTRACT

PURPOSE

To audit and improve the completeness of the hierarchical (or *is-a*) relations of the NCI Thesaurus in order to support its role as a faceted system for querying cancer registry data.

METHODS

We performed quality auditing of the 19.01d version of the NCI Thesaurus. Our hybrid auditing method consists of three main steps: (1) computing non-lattice subgraphs; (2) constructing lexical features for concepts in each subgraph; and (3) performing subsumption reasoning with each subgraph to automatically suggest potentially missing *is-a* relations.

RESULTS

A total of 9,512 non-lattice subgraphs were obtained. 925 potentially missing *is-a* relations in 441 non-lattice subgraphs were identified by our method. 72 out of 176 reviewed samples were confirmed as valid missing *is-a* relations and have been incorporated in the newer versions of the NCI Thesaurus.

CONCLUSION

Auto-suggested changes resulting from our auditing method can improve the structural organization of the NCI Thesaurus in supporting its new role for faceted query.

INTRODUCTION

The Kentucky Cancer Registry (KCR) [1] was established in 1991 at the University of Kentucky Markey Cancer Center (MCC). It is a central cancer registry receiving data about new cancer cases from all healthcare facilities and physicians in Kentucky within 4 months of diagnosis, as required by state law. Despite advances in cancer research over the last several decades, the cancer burden in Kentucky remains severe. According to State Cancer Profiles statistics provided by the National Cancer Institute (NCI) and the Centers for Disease Control and Prevention (CDC), Kentucky is the state that has the nation's highest cancer burden [2, 3]. In 2000, KCR became a part of the NCI's Surveillance Epidemiology and End Results (SEER) program [4, 5]. The SEER registries are considered to be among the most accurate and complete population-based cancer registries in the world that include stage of cancer at the time of diagnosis and patient survival data.

Such cancer registry data have enabled web-based access to the data and analytics tools for cancer research. For example, State Cancer Profiles provide a user-friendly interface for finding cancer statistics for specific states and counties for public health officials and policy makers. KCR has also developed an NCI-funded Apple iOS app, called Cancer Rates [6], to make incidence and mortality information available on mobile devices. However, the interfaces of such query engines do not support sophisticated data exploration such as identifying patient cohorts for the feasibility of clinical trials, and have not achieved usability approaching the levels of those for consumer websites due in critical part to the lack of faceted capabilities [7-9]. Faceted organization and presentation of metadata is the key mechanism that allowed

consumers of websites such as Amazon to quickly narrow down from millions of products to items of interest using dimensions of attributes (e.g., simple facets such as size, color, maker, price range). Faceted systems for querying clinical data is not widely available due to the complexity of data and the mismatch between the ontologies used for organizing and annotating clinical data (such as NCI Thesaurus [10, 11]), and the desired facet structures and properties. Therefore, in the NCI-funded project (R21CA231904), we aim to overcome these challenges and develop OncoSphere, a faceted query engine using the NCI Thesaurus as a nested facet system (NFS) [12] to provide web-based exploration of the Kentucky Cancer Registry data.

A nested facet is a facet that includes a collection of other facets (or sub-facets) as its components [12]. An NFS is a set of nested facets with a hierarchical (or subtype, or is-a) relation among them. The efficacy of an NFS requires the properties of soundness and completeness [12]. Soundness means that all items within each facet are relevant to the facet, that is, for each facet, all the sub-facets listed within the facet is indeed its subtypes. Completeness means that any sub-facets relevant to a specific facet are already contained in and accessible through the facet, that is, there is no missing subtypes for the facet. The soundness and completeness properties of facets directly affect the performance of the query engine in terms of precision and recall. Incomplete facets will reduce recall, while unsound facets will reduce precision. For instance, "*Anaplastic T-Lymphocyte*" is currently not listed as one of the subtypes of "*Neoplastic Large T-Lymphocyte*" (i.e., incomplete facet) in the NCI Thesaurus, and would thus be a missing choice in the corresponding facet for "*Neoplastic Large T-Lymphocyte*." As a consequence, patients with "*Anaplastic T-Lymphocyte*" would not be included for a

cohort of patients with "*Neoplastic Large T-Lymphocyte,*" reducing the query recall. Since OncoSphere relies on the hierarchical structure of the NCI Thesaurus for its faceted query interface, it is essential to ensure the quality of the NCI Thesaurus.

In this paper, we focus on performing quality auditing on the hierarchical structure of the NCI Thesaurus. We develop a hybrid method leveraging a specific substructure called non-lattice subgraph [13] and lexical features of concepts in the non-lattice subgraph to automatically detect missing hierarchical relations in the NCI Thesaurus. The key idea of our method is that non-lattice subgraphs pinpoint to problematic areas which are likely to contain hierarchical quality issues, while lexical features facilitate the identification of potentially missing hierarchical relations in the non-lattice subgraphs through subsumption reasoning.

METHODS

We use the 19.01d version of the NCI Thesaurus. Our hybrid method consists of the following steps: (1) computing non-lattice subgraphs; (2) constructing lexical features; and (3) performing subsumption reasoning.

Computing Non-lattice Subgraphs

Concepts in the NCI Thesaurus are hierarchically organized as a direct acyclic graph (DAG), where a node (or concept) may have multiple parents. A pair of concepts is called a non-lattice pair, if the two concepts share more than one maximal common descendant (or one minimal common ancestor) [14]. Here the maximal common descendant of two concepts *v* and *w* in a DAG is the highest node that has both *v* and *w* as ancestors; and the minimal common ancestor of two concepts in a DAG is the lowest

node that has both concepts as descendants. For instance, in Figure 1, concept 1 and concept 2 have two maximal common descendants: concept 5 and concept 6; therefore, concepts 1 and 2 form a non-lattice pair. Similarly, concept 2 and concept 3 have two maximal common descendants: 5 and 6, and form a non-lattice pair.

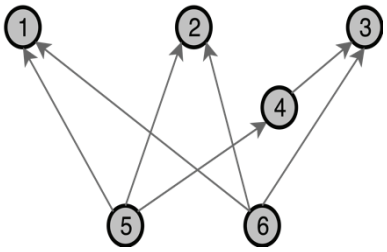

Figure 1. An example of non-lattice subgraph.

A non-lattice pair *P* determines a non-lattice subgraph, which can be obtained by first computing the maximal common descendants of the non-lattice pair, denoted as *mcd(P)*; reversely computing *mcd(P)*'s minimal common ancestors, denoted as *mca(mcd(P))*; and then aggregating the concepts and relations between (and including) any concept in *mca(mcd(P))* and any concept in *mcd(P)* [13]. For instance, given the non-lattice pair *P* = (1, 2) in Figure 1, *P*'s maximal common descendants *mcd(P)* are 5 and 6; computing *mcd(P)*'s minimal common ancestors obtains 1, 2, and 3; by aggregating concepts between {1, 2, 3} and {5, 6}, we have the non-lattice subgraph containing six concepts {1, 2, 3, 4, 5, 6}.

Figure 2 shows an example of non-lattice subgraph in the NCI Thesaurus determined by the non-lattice pair *P* = (C_5, C_6). *P*'s maximal common descendants *mcd(P)* are C_1 and C_2; computing *mcd(P)*'s minimal common ancestors still obtains C_5 and C_6, that is, *P* itself; and

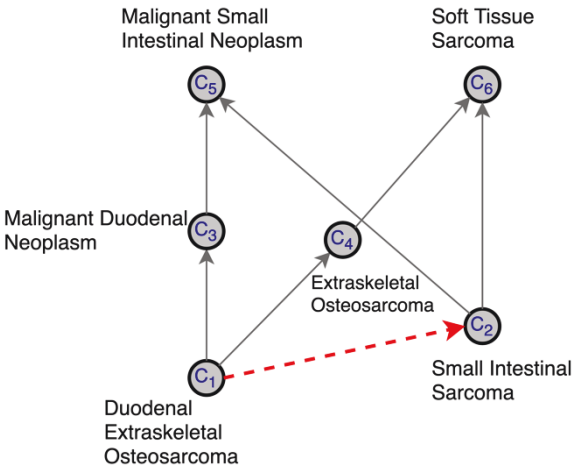

Figure 2. An example of non-lattice subgraph in the NCI Thesaurus (19.01d version).

aggregating concepts in between results in the non-lattice subgraph consisting of six concepts {C_1, C_2, C_3, C_4, C_5, C_6}.

We use an efficient algorithm developed in our previous work [15] to exhaustively compute all the non-lattice subgraphs in the NCI Thesaurus. This algorithm has been tested on large biomedical terminologies in DAG including SNOMED CT, Gene Ontology, and NCI Thesaurus.

Constructing Lexical Features

We create a set of lexical features (or lexical set) for each concept in the non-lattice subgraph. Given a concept *C* in a non-lattice subgraph *G*, we model its lexical set as the words (unordered) appearing in the name of the concept *C* and inherited from the names of *C*'s ancestors in *G*. That is, the concept *C*'s lexical features consist of two parts, where the first part contains the words appearing in the concept *C*'s own name, and the second part contains the words inherited from the names of the concept *C*'s ancestors within the non-lattice subgraph *G*. For example, for concept C_1 = "*Duodenal Extraskeletal Osteosarcoma*" in Figure 2, the first part of lexical features contains the words in its own name, i.e., {*duodenal, extraskeletal, osteosarcoma*}. The second part contains the words inherited from C_1's ancestors (C_3, C_4, C_5, and C_6), that is, {*malignant, duodenal, neoplasm, extraskeletal, osteosarcoma, malignant, small, intestinal, neoplasm, soft, tissue, sarcoma*}. Since we model the lexical features of a concept as a set of words, removing duplicated words in both parts obtains {*duodenal, extraskeletal, osteosarcoma, malignant, neoplasm, small, intestinal, soft, tissue, sarcoma*}, where "*duodenal,*" "*extraskeletal*" and "*osteosarcoma*" appear in the name of the concept C_1 itself; "*malignant*" and "*neoplasm*" are from its ancestors C_3 and C_5;

"*small*" and "*intestinal*" are from its ancestor C_5; and "*soft,*" "*tissue*" and "*sarcoma*" are from its ancestor C_6. Table 1 presents the lexical sets of all concepts in the non-lattice subgraph shown in Figure 2.

Table 1. The lexical sets of concepts in the non-lattice subgraph shown in Figure 2.

Concept	**Lexical set**
C_1	{*duodenal, extraskeletal, osteosarcoma, malignant, neoplasm, small, intestinal, soft, tissue, sarcoma*}
C_2	{*small, intestinal, sarcoma, malignant, neoplasm, soft, tissue*}
C_3	{*malignant, duodenal, neoplasm, small, intestinal*}
C_4	{*extraskeletal, osteosarcoma, soft, tissue, sarcoma*}
C_5	{*malignant, small, intestinal, neoplasm*}
C_6	{*soft, tissue, sarcoma*}

Performing Subsumption Reasoning

We perform subsumption reasoning to detect potentially missing *is-a* relations among the pairs of concepts which are currently not hierarchically related. For each non-lattice subgraph, we first identify pairs of concepts which are currently not hierarchically related; then for each pair of concepts (say C_1 and C_2), we check whether their lexical sets have an inclusion relation as follow: (1) if C_2's lexical set is a proper subset of C_1's lexical set, then we suggest a potentially missing *is-a* relation between C_1 and C_2, i.e., C_1 *is-a* C_2; (2) if C_1's lexical set is a proper subset of C_2's lexical set, then we suggest a potentially missing *is-a* relation between C_2 and C_1, i.e., C_2 *is-a* C_1; otherwise (3) no suggestion will be made.

For instance, for concepts C_1 (*Duodenal Extraskeletal Osteosarcoma*) and C_2 (*Small Intestinal Sarcoma*) in Figure 2, since C_2's lexical set {*small, intestinal, sarcoma, malignant, neoplasm, soft, tissue*} is a proper subset of C_1's lexical set {*duodenal, extraskeletal, osteosarcoma, malignant, neoplasm, small, intestinal, soft, tissue, sarcoma}*, we suggest C_1 *is-a* C_2, that is, "*Duodenal Extraskeletal Osteosarcoma*" is a subtype of "*Small Intestinal Sarcoma*" (see the dashed red arrow in Figure 2). For concepts C_5 (*Malignant Small Intestinal Neoplasm*) and C_6 (*Soft Tissue Sarcoma*) in Figure 2, there is no inclusion between their lexical sets, and thus no suggestion will be made for this pair of concepts.

When performing such subsumption reasoning, we do not make suggestions for certain scenarios which are prone to generate incorrect suggestions, such as concepts containing stop words (e.g., "*and/or*," "*no*," "*not*," "*without*," "*except,*" "by") and lexical sets containing antonyms (e.g., "*small*," "*large*"). In addition, after obtaining all the potential missing *is-a* relations in non-lattice subgraphs, we further remove redundant *is-a* relations that can be inferred by other *is-a* relations.

RESULTS

Using the 19.01d version of the NCI Thesaurus, a total of 9,512 non-lattice subgraphs were obtained. The size of the non-lattice subgraph ranges from 4 to 644. Our hybrid method detected 925 potentially missing *is-a* relations in 441 non-lattice subgraphs. Figure 3 shows two examples of the identified missing *is-a* relations in two non-lattice subgraphs in size of 4 and 5, respectively.

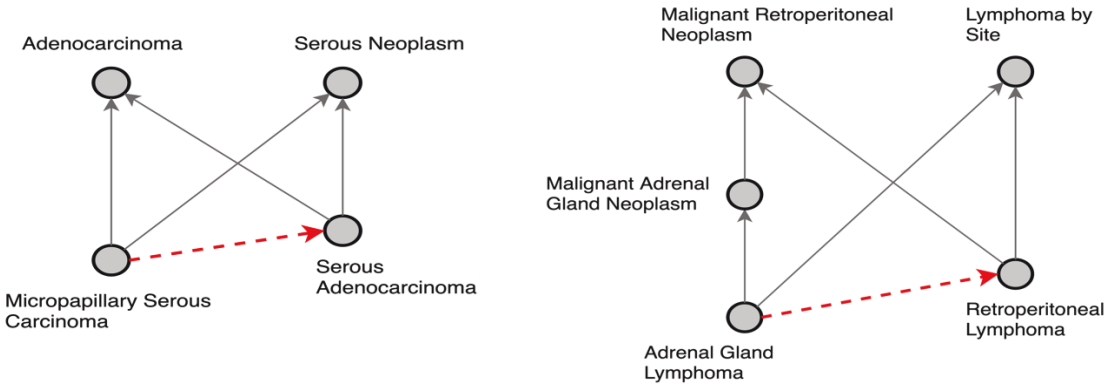

Figure 3. Left: A non-lattice subgraph of size 4 suggesting a missing *is-a* relation: "*Micropapillary Serous Carcinoma*" *is-a* "*Serous Adenocarcinoma*." Right: A non-lattice subgraph of size 5 suggesting a missing *is-a* relation: "*Adrenal Gland Lymphoma*" *is-a* "*Retroperitoneal Lymphoma*."

Preliminary Evaluation

For evaluation, we provided the NCI Enterprise Vocabulary Services (EVS), by which the NCI Thesaurus is managed, with 253 potentially missing *is-a* relations in non-lattice subgraphs of size less than or equal to 15. These non-lattice subgraphs were visualized in PDF and organized in terms of the sub-hierarchies to facilitate the EVS experts' review and evaluation. Table 2 shows the number of the potentially missing *is-a* relations identified by our method according to the sub-hierarchies. The EVS experts reviewed four sub-hierarchies: "*Disease, Disorder or Finding*," "*Activity*," "*Abnormal Cell*," and "*Anatomic Structure, System, or Substance*." The sub-hierarchy "*Disease, Disorder or Finding*" contains 136 potentially missing is-a relations, among which 50 were verified as valid by EVS experts. In total, 72 out of 176 reviewed samples were confirmed as valid missing *is-a* relations and have been incorporated in the newer versions of the NCI Thesaurus. Table 3 lists ten examples of valid missing *is-a* relations verified by EVS experts (see Supplemental Material for the comprehensive list).

Table 2. The number of potentially missing is-a relations detected in non-lattice subgraphs (size <=15) in terms of the sub-hierarchies in the NCI Thesaurus (19.01d version), as well as the number of valid missing is-a relations verified by EVS experts. "-" indicates that the samples in the corresponding sub-hierarchy were not reviewed by EVS experts.

Sub-hierarchy	**# potentially missing *is-a* relations**	**# valid missing *is-a* relations**
Disease, Disorder or Finding	136	50
Drug, Food, Chemical or Biomedical Material	31	-
Activity	20	16
Experimental Organism Diagnosis	18	-
Abnormal Cell	16	6
Manufactured Object	14	-
Anatomic Structure, System, or Substance	4	1
Conceptual Entity	4	-
Property or Attribute	3	-
Gene Product	3	-
Molecular Abnormality	2	-
Biological Process	1	-
Biochemical Pathway	1	-

Table 3. Ten examples of valid missing *is-a* relations verified by EVS experts.

Sub-hierarchy	**Missing *is-a* relation**
Disease, Disorder or Finding	Micropapillary Serous Carcinoma *is-a* Serous Adenocarcinoma
Disease, Disorder or Finding	Adrenal Gland Sarcoma *is-a* Retroperitoneal Sarcoma
Disease, Disorder or Finding	Acoustic Schwannoma *is-a* Benign Ear Neoplasm
Disease, Disorder or Finding	Penile Bowenoid Papulosis *is-a* Male Reproductive System Precancerous Condition
Disease, Disorder or Finding	Congenital Pulmonary Lymphangiectasia *is-a* Pulmonary Vascular Disorder

Disease, Disorder or Finding	Gingival Ecchymosis *is-a* Oral Hemorrhage
Activity	Sigmoidoscopy with Biopsy *is-a* Diagnostic Colonoscopy
Activity	Nucleic Acid Hybridization *is-a* Genetic Technique
Abnormal Cell	Malignant Clear Cell Oncocyte *is-a* Adenocarcinoma Clear Cell
Abnormal Cell	Neoplastic Parathyroid Gland Clear Cell *is-a* Neoplastic Endocrine Clear Cell

DISCUSSION

Erroneous Suggestions and Potential Solutions

Although our hybrid method was able to suggest valid missing *is-a* relations, it sometimes makes erroneous suggestions for ambiguous cases. For instance, our method suggested "*Metastatic Malignant Neoplasm in the Pancreas*" is a subtype of "*Metastatic Malignant Pancreatic Neoplasm.*" This suggestion is invalid since the latter concept refers to metastatic malignant neoplasms that originate from the pancreas and spread to other anatomical sites. This is an erroneous pattern of "*malignant neoplasm in X site*" versus "*malignant X site neoplasm*" that our lexical-set-based method was not able to differentiate. A potential solution to avoid such erroneous suggestions is to add "*in*" as a stop word. However, adding it as a stop word would miss valid suggestions such as "*Metastatic Malignant Neoplasm in the Sellar Region*" *is-a* "*Metastatic Malignant Neoplasm in the Central Nervous System*." This relates to the challenge that the degrees of generating erroneous suggestions vary for different stop words. For future improvement, we plan to explore machine learning-based approaches to train the model on positive and negative samples for different stop words and test whether the model can differentiate the degrees or even automatically learn new stop words in

addition to what we have used. Another potential solution is to leverage the roles defining the concepts, to automatically facilitate the subsumption reasoning and differentiate the ambiguous concepts.

Erroneous Suggestions Indicating Problematic Existing Relations

For certain scenarios, erroneous suggestions made by our method further reveal problematic relations existing in the NCI Thesaurus. For example, our method suggested an invalid relation: "*Nerve Plexus*" *is-a* "*Peripheral Nerve*." However, the existing relations which were leveraged by our method to generate the suggestion are: "*Nerve Plexus*" *is-a* "*Peripheral Nervous System Part*" and "*Nerve Plexus*" *is-a* "*Nerve*" (see the non-lattice subgraph in Figure 4). Since nerve plexus has peripheral nerves as parts but is not itself a nerve, this invalid suggestion revealed an incorrect existing relation in the NCI Thesaurus: "*Nerve Plexus*" *is-a* "*Nerve*" (see the link with a bolded cross in Figure 4). This indicates that our method may help with further identification of problematic *is-a* relations in the NCI Thesaurus in addition to missing *is-a* relations.

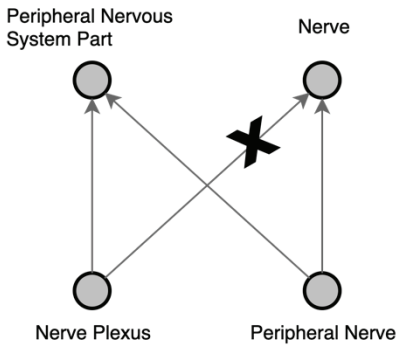

Figure 4. An example of scenarios where erroneous suggestions further reveal problematic existing relations.

Comparison with Related Work

In our previous work [16], we have used six predefined lexical patterns in non-lattice subgraphs to identify potentially missing *is-a* relations in the NCI Thesaurus. In this work, we directly utilize the lexical sets of concepts to perform subsumption reasoning with no need to predefine lexical patterns. More importantly, our method in this work identifies previously undiscovered missing *is-a* relations in the NCI Thesaurus. In

another related work [17], we leveraged non-lattice subgraphs and concepts names to perform subsumption reasoning for suggesting potentially missing *is-a* relations in SNOMED CT. In this work, we model the lexical features of a concept in the NCI Thesaurus using not only the name of the concept itself, but also the names of the concept's ancestors.

Limitations

Although our hybrid method is capable of revealing valid missing *is-a* relations, it only touches upon a small portion of non-lattice subgraphs in the NCI Thesaurus (441 out of 9,512), leaving the remaining non-lattice subgraphs untapped. New methods are needed to uncover the potential quality issues in the untapped non-lattice subgraphs. We will further leverage the roles defining the concepts to facilitate the quality auditing task. Regarding the use of stop words in our method, although it can avoid making erroneous suggestions, it may also miss valid suggestions as mentioned above. Additional research is needed to specifically handle concepts containing stop words.

Another limitation is that our evaluation was preliminary in two aspects. (1) The evaluation was based on non-lattice subgraphs with size of less than or equal to 15., since the EVS experts would be visually overwhelmed with large graphs to manually review. (2) The EVS experts did not review all the provided potentially missing *is-a* relations. We plan to provide EVS experts with a random sample of potentially missing *is-a* relations detected from a newer version of the NCI Thesaurus after improving our method's ability to distinguish ambiguous concepts.

CONCLUSION

In this paper, we developed a hybrid method to automatically suggest potentially missing *is-a* relations in the NCI Thesaurus. Auto-suggested changes resulting from our auditing method can improve the structural organization of the NCI Thesaurus in supporting its new role for faceted query.

REFERENCES

[1] Kentucky Cancer Registry. https://www.kcr.uky.edu/

[2] The 20 States with the Highest Cancer Rates. https://www.thehealthy.com/cancer/states-with-the-highest-cancer-rates/

[3] United States Cancer Statistics: Data Visualizations. https://gis.cdc.gov/Cancer/USCS/DataViz.html

[4] NCI's Surveillance Epidemiology and End Results (SEER). https://seer.cancer.gov/

[5] Hayat MJ, Howlader N, Reichman ME, Edwards BK. Cancer statistics, trends, and multiple primary cancer analyses from the Surveillance, Epidemiology, and End Results (SEER) Program. The oncologist. 2007;12(1):20-37.

[6] Cancer Rates. https://itunes.apple.com/us/app/cancer-rates/id1049312556?mt=8

[7] D. Tunkelang. Faceted search (synthesis lectures on information concepts, retrieval, and services). Morgan and Claypool Publishers, 2009.

[8] M. A. Hearst. Clustering versus faceted categories for information exploration. Communications of the ACM, 49(4):59-61, 2006.

[9] M. Hearst. Design recommendations for hierarchical faceted search interfaces. In ACM SIGIR workshop on faceted search, pp. 1-5, 2006.

[10] S, Haber MW, Sioutos N, Tuttle MS, Wright LW. NCI Thesaurus: using science-based terminology to integrate cancer research results. In Medinfo 2004, pp. 33-37.

[11] Sioutos N, de Coronado S, Haber MW, Hartel FW, Shaiu WL, Wright LW. NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information. Journal of biomedical informatics. 2007;40(1):30-43.

[12] Zhang GQ, Tao S, Zeng N, Cui L. Ontologies as nested facet systems for human–data interaction. Semantic Web.(Preprint):1-8.

[13] Cui L, Zhu W, Tao S, Case JT, Bodenreider O, Zhang GQ. Mining Non-Lattice Subgraphs for Detecting Missing Hierarchical Relations and Concepts in SNOMED CT. Journal of the American Medical Informatics Association 2017;24(4): 788-798.

[14] Zhang GQ, Bodenreider O. Large-scale, exhaustive lattice-based structural auditing of SNOMED CT. AMIA Annu Symp Proc;2010:922-26.

[15] Zhang GQ, Xing G, Cui L. An efficient, large-scale, non-lattice-detection algorithm for exhaustive structural auditing of biomedical ontologies. Journal of biomedical informatics. 2018;80:106-19.

[16] Abeysinghe R, Brooks MA, Talbert J, Cui L. Quality Assurance of NCI Thesaurus by Mining Structural-Lexical Patterns. AMIA Annual Symp Proc; 2017:364-373.

[17] Cui L, Bodenreider O, Shi J, Zhang GQ. Auditing SNOMED CT hierarchical relations based on lexical features of concepts in non-lattice subgraphs. Journal of biomedical informatics. 2018;78:177-84.

Figure Legends

Figure 1. An example of non-lattice subgraph.

Figure 2. An example of non-lattice subgraph in the NCI Thesaurus (19.01d version).

Figure 3. Left: A non-lattice subgraph of size 4 suggesting a missing *is-a* relation: "*Micropapillary Serous Carcinoma*" *is-a* "*Serous Adenocarcinoma*." Right: A non-lattice subgraph of size 5 suggesting a missing *is-a* relation: "Adrenal Gland Lymphoma" *is-a* "Retroperitoneal Lymphoma."

Figure 4. An example of scenarios where erroneous suggestions further reveal problematic existing relations.