A Lexical-based Formal Concept Analysis Method to Identify Missing Concepts in the NCI Thesaurus

Fengbo Zheng*,†, Licong Cui†
*Department of Computer Science, University of Kentucky, Lexington, Kentucky, USA
†School of Biomedical Informatics, University of Texas Health Science Center at Houston, Houston, Texas, USA

***Abstract*—Biomedical terminologies have been increasingly used in modern biomedical research and applications to facilitate data management and ensure semantic interoperability. As part of the evolution process, new concepts are regularly added to biomedical terminologies in response to the evolving domain knowledge and emerging applications. Most existing concept enrichment methods suggest new concepts via directly importing knowledge from external sources. In this paper, we introduced a lexical method based on formal concept analysis (FCA) to identify potentially missing concepts in a given terminology by leveraging its intrinsic knowledge – concept names. We first construct the FCA formal context based on the lexical features of concepts. Then we perform multistage intersection to formalize new concepts and detect potentially missing concepts. We applied our method to the *Disease or Disorder* sub-hierarchy in the National Cancer Institute (NCI) Thesaurus (19.08d version) and identified a total of 8,983 potentially missing concepts. As a preliminary evaluation of our method to validate the potentially missing concepts, we further checked whether they were included in any external source terminology in the Unified Medical Language System (UMLS). The result showed that 592 out of 8,937 potentially missing concepts were found in the UMLS.**

***Index Terms*—Biomedical Terminologies, Concept Enrichment, Formal Concept Analysis**

I. Introduction

A terminology or ontology provides formalized representation of knowledge in a domain, including a set of concepts and the describable relationships among them. In biomedicine, terminologies have played important roles in biomedical research and applications to ensure data consistency and interoperability [1]. For instance, the National Cancer Institute (NCI) Thesaurus, covering knowledge of cancers, genes and therapies [2]–[4], has been widely used as a standard for biomedical coding, knowledge reference, and public reporting for many NCI and other systems [5].

Biomedical terminologies are often incomplete and constantly evolving due to the growing knowledge in biomedicine and new requirements from emerging biomedical applications [6]. During the terminology evolution process, new concepts are regularly added to the newer versions. For instance, the NCI Thesaurus is updated every month with averaging roughly 700 new concepts in each release [7].

This work was supported by the US National Science Foundation (NSF) under grant 1931134 and National Institutes of Health (NIH) under grant R01LM013335. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NSF or NIH. Correspondence: licong.cui@uth.tmc.edu

There are two types of approaches to identify new or missing concepts for the concept enrichment of biomedical terminologies. One type mainly leverages extrinsic knowledge (or external sources). For instance, Chandar et al. developed a similarity-based method that suggests extracted phrases from text corpus as new concepts for the SNOMED CT [8]. Peng et al. analyzed connected matrices from the Gene Ontology (GO) and biological network to identified new terms for the GO [9]. He et al. leveraged alignment between different terminologies to suggest new concepts for the SNOMED CT [10] and NCI Thesaurus [11].

The other type mainly utilizes the intrinsic knowledge within the terminology itself. For example, Jiang and Chute performed Formal Concept Analysis (FCA) based on logical definitions of concepts in the SNOMED CT to search for possible missing concepts [12]. Zhu et al. developed a scalable multistage algorithm called Spark-MCA [13] to deal with the computational challenge of performing large-scale FCA for evaluating concept completeness of the SNOMED CT. A limitation of these two FCA-based approaches is that the potentially missing concepts identified only involve logical definitions and no concept names were provided. Therefore, it is difficult to validate those missing concepts. In previous work [14], we discovered a lexical pattern in non-lattice subgraphs that can reveal missing concepts in the SNOMED CT; and we explored deep learning-based methods to properly name a concept given its lexical components (or a bag of words) [15]. However, our previous work is limited to a specific type of lexical patterns and sub-structures of terminologies, which only revealed a small portion of missing concepts.

In this paper, we introduce a lexical- and FCA-based method to identify potentially missing concepts in the NCI Thesaurus. Lexical features (i.e., words appeared in the concept names) are considered as FCA attributes while generating formal context. Applying multistage intersection of FCA attributes identifies newly formalized bags of words (i.e., FCA formal concepts) that represent missing concepts, which may be further validated through external knowledge. We applied our method to the *Disease or Disorder* (C2991) sub-hierarchy in 19.08d version of the NCI Thesaurus and identified 8,983 potentially missing concepts. We performed a preliminary evaluation and validated that 592 out of 8,983 potentially missing concepts were included in external terminologies in the Unified Medical Language System (UMLS).

II. Background

FCA is a mathematical theory concerned with the formalization of concepts and conceptual thinking [16]. With FCA, we can generate a concept hierarchy from a collection of object and attributes. The input of FCA is *formal context* $K = (O, A, R)$, where O is a set of objects, A is a set of attributes, and R is a binary relation between O and A. The notation $(o, a) \in R$ means that object o has attribute a.

Each formal context K induces two operators: derivation operators $\uparrow$: $2^O \rightarrow 2^A$ and concept-forming operators $\downarrow$: $2^A \rightarrow 2^O$. The operators are defined, for each $X \subseteq O$ and $Y \subseteq A$, as follows:

$$X^{\uparrow} = \{a \in A | \forall o \in X\text{: } (o, a) \in R\},$$

$$Y^{\downarrow} = \{o \in O | \forall a \in Y\text{: } (o, a) \in R\},$$

where $X^{\uparrow}$ is the set of all attributes shared by all objects in X, and $Y^{\downarrow}$ is the set of all objects sharing all attributes in Y.

A formal concept of K is a pair (X, Y) with $X \subseteq O$ and $Y \subseteq A$ such that $X^{\uparrow} = Y$ and $Y^{\downarrow} = X$. The subconcept-superconcept relation between formal concepts is given by $(X_1, Y_1) \leq (X_2, Y_2)$ iff $X_1 \subseteq X_2$ ($Y_2 \subseteq Y_1$). All formal concepts derived from the formal context K together with the subconcept-superconcept relation form a complete lattice [17]. Note that lattice is a desired property for well-structured terminologies.

III. Method

Our method mainly consists of two steps: (1) preprocessing concept names and constructing FCA formal context; and (2) performing FCA via a multistage intersection algorithm to identify potentially missing (or new) concepts in the NCI Thesaurus.

A. Constructing Formal Context

Given a collection of concepts in the terminology, we consider all the concepts as FCA objects O and words appears in the concept names (i.e., lexical features) as FCA attributes A, respectively. With the binary relation $R \subseteq O \times A$ specifying whether concept $o \in O$ contains word $a \in A$, we can construct the FCA formal context $K = (O, A, R)$.

Since words appearing in concept names may have variations (e.g., plural vs. singular forms) or synonyms, we perform attribute/word normalization to create a more robust FCA formal context. For word variations, we normalize words appearing in concept names using LuiNorm [18], a lexical tool provided by the UMLS. For example, "bones" can be normalized to "bone". Regarding word synonyms, we leverage concepts in the NCI Thesaurus with single-word preferred names and single-word synonyms. More specifically, if a word w itself is the preferred name of an NCI Thesaurus concept and has a synonym s that is also a single word, then we maintain a mapping between the synonym s and the preferred name w. This way words with the same meanings can be normalized to their preferred names thus the same attribute.

B. Identifying Potentially Missing Concepts

To derive FCA formal concepts, we leverage the idea of the faster concept analysis introduced in [19], which is to perform multistage intersection on each pair of formal concepts from the initial formal concept set consisting of all objects, until no more new formal concept is generated. The pseudocode of the algorithm is shown in Fig. 1.

Algorithm 1 Identifying Missing Concepts

Input: Formal context (O, A, R)
Output: Missing concept set M
Initialization:
Original set $S_0 \leftarrow \{o^{\uparrow} | o \in O\}$
Initial set $I \leftarrow S_0$
Newly derived formal concept set $N \leftarrow S_0$
while $N \neq \emptyset$
    Last iteration formal concept set $L \leftarrow I$
    for each pair (C_x, C_y) in L
        I.add(Intersetion(C_x, C_y))
    $N \leftarrow (I - L)$
$M \leftarrow I - S_0$

Fig. 1. Pseudocode of identifying potentially missing concepts by multistage intersection.

In practice, for computation convenience, we perform operations on the lexical feature sets (i.e., using FCA attribute sets to represent FCA formal concepts). The initial set of FCA formal concepts is a set of FCA attribute sets, that is, the lexical feature sets of all the original concepts (i.e., $\{o^{\uparrow} \mid o \in O\}$). In the first iteration, we compute the intersection of each pair of FCA attribute sets in the initial set; and if the result is not included in the initial set, we add it into the initial set. We repeat this process until no new FCA attribute set can be derived. Each newly generated FCA attribute set is taken as the lexical feature set of a potentially missing concept among the given concepts. An advantage of using lexical features (or words) as FCA attribute sets is that these words can be further leverage to name the newly discovered concepts.

C. Illustrative Example

Fig. 2 shows a simple example of FCA formal context in a tabular format generated from the concept *Breast Fibroepithelial Neoplasm* (C40405) and its descendants in the NCI Thesaurus. The cells with check marks represent the binary relation between the concepts and their lexical features. Note that word "Tumor" is normalized to "neoplasm", since it is a synonym of *Neoplasm* (C3262) in the NCI Thesaurus.

Given the FCA formal context, the FCA formal concept with attribute set {breast, neoplasm} (see blue cells in Fig. 2) can be derived by intersecting the attribute sets of *Borderline Breast Phyllodes Tumor* (C5316) and *Breast Fibroepithelial Neoplasm* (C40405). Therefore, a concept with lexical feature set {breast, neoplasm} is considered as a potentially missing concept for the given FCA formal context. This example only

	juvenile	fibroepithelial	malignant	breast	fibroadenoma	neoplasm	complex	borderline	pericanalicular	intracanalicular	bengin	phyllode	giant
C3744: Breast Fibroadenoma				✓	✓								
C7575: Breast Phyllodes Tumor				✓		✓						✓	
C4271: Breast Intracanalicular Fibroadenoma				✓	✓					✓			
C4272: Breast Pericanalicular Fibroadenoma				✓	✓				✓				
C4273: Breast Giant Fibroadenoma				✓	✓								✓
C4276: Breast Juvenile Fibroadenoma	✓			✓	✓								
C5194: Breast Complex Fibroadenoma				✓	✓		✓						
C4504: Malignant Breast Phyllodes Tumor			✓	✓		✓						✓	
C5196: Benign Breast Phyllodes Tumor				✓		✓					✓	✓	
C5316: Borderline Breast Phyllodes Tumor				✓		✓		✓				✓	
C40405: Breast Fibroepithelial Neoplasm		✓		✓		✓							

Fig. 2. An example of FCA formal context generated by the concept *Breast Fibroepithelial Neoplasm* (C40405) in the NCI Thesaurus and its descendants in company with their lexical features. Word "Tumor" is normalized to "neoplasm" and word "Phyllodes" is normalized to "phyllode". An FCA formal concept (marked by blue cells) with FCA attribute set {breast, neoplasm} is considered as a potentially missing concept among the given concepts.

intends to illustrate how our method works, and one may have noticed that *Breast Neoplasm* (C2910) is an existing concept in the NCI Thesaurus although it is not among the given concepts. For the actual implementation of our method, we further check if the newly generated concepts are existing in the NCI Thesaurus and ensure the removal of such cases from the list of potentially missing concepts.

IV. Results

A. Summary Result

We applied our method to the sub-hierarchies under *Disease or Disorder* (C2991) in the NCI Thesaurus (19.08d version). Table I shows the numbers of existing concepts, newly generated concepts, and potentially missing concepts respectively for each sub-hierarchy. For example, there are 10,996 existing concepts in the *Neoplasm* (C3262) sub-hierarchy; and FCA generated a total of 8,511 new concepts, among which 7,737 were potentially missing concepts in the NCI Thesaurus.

Note that potentially missing concepts are detected in terms of the given FCA formal context (or the given collection of the input concepts). Therefore, the missing concepts detected in a sub-hierarchy may overlap with those detected in another sub-hierarchy. In total, 8,983 unique potentially missing concepts were identified among these sub-hierarchies.

B. Preliminary Evaluation

We performed a preliminary evaluation to validate the potentially missing concepts identified using the external knowledge in the UMLS. The UMLS integrates millions of biomedical concepts from more than 200 source terminologies, including the GO, SNOMED CT and Medical Subject Headings (MSH), to enable interoperability between biomedical information systems [20].

For each potentially missing concept identified, we checked whether its lexical feature set can be matched to any concept name from the external terminologies in the UMLS. We found 592 out of 8,983 potentially missing concepts are included in the external terminologies in UMLS (see Table I for the number of missing concepts validated via UMLS for each sub-hierarchy). Table II lists 10 examples of validated missing concepts (in the form of lexical feature sets) and matched concept names in the UMLS terminologies.

TABLE I
The numbers of existing concepts, newly generated concepts, potentially missing concepts, and missing concepts validated via UMLS for each sub-hierarchy of *Disease or Disorder* (C2991).

Sub-hierarchy	# of Concepts	# of Newly Generated Concepts		
		Total	# of Potentially Missing	# of Validated via UMLS
C27551: Disorder by Site	13,595	9,114	7,864	451
C3262: Neoplasm	10,996	8,511	7,737	289
C53529: Non-Neoplastic Disorder	4,198	1,279	813	227
C8278: Cancer-Related Condition	578	491	374	28
C4873: Rare Disorder	915	283	196	44
C89328: Pediatric Disorder	528	280	218	20
C28193: Syndrome	907	266	204	31
C3101: Genetic Disorder	159	52	30	6
C2893: Psychiatric Disorder	231	45	29	11
C3113: Hyperplasia	81	24	17	8
C3340: Polyp	110	24	7	2
C35470: Behavioral Disorder	49	19	9	0
C3075: Hamartoma	63	15	6	0
C26684: Radiation-Induced Disorder	25	5	3	0

Since a matching concept may be from multiple UMLS terminologies, we further looked into the terminologies that contributed most to the validation of the 592 identified potentially missing concepts. The top 10 in terms of the number of matched concepts (in parentheses) are listed as follows: Consumer Health Vocabulary - CHV (328), SNOMED CT US Edition - SNOMEDCT_US (245), Read Codes - RCD (135), MedDRA - MDR (124), ICPC2-ICD10 Thesaurus - ICPC2ICD10ENG (101), MSH (97), Metathesaurus Names - MTH (79), MEDCIN (78), Online Mendelian Inheritance in Man - OMIM (55), and Logical Observation Identifiers Names and Codes - LNC (52).

V. Discussion

In this work, we leveraged words in concept names and FCA to detect potentially missing concepts in the NCI Thesaurus. The preliminary evaluation via UMLS-based validation indicates that our method has the potential to identify missing concepts for concept enrichment of the NCI Thesaurus.

However, this work has several limitations that need further improvement. First, the potentially missing concepts detected by our method may not be directly imported into a terminology. This is because different terminologies are developed for disparate purposes and have varying target applications,

TABLE II
TEN EXAMPLES OF VALIDATED MISSING CONCEPTS AND THEIR MATCHED CONCEPTS IN THE UMLS TERMINOLOGIES.

Lexical Feature Set of Missing Concept	Matched Concept (External Terminology)
{carcinoma, papillary, urothelial}	Papillary urothelial carcinoma (SNOMEDCT_US)
{borderline, serous, tumor}	Serous borderline tumor (SNOMEDCT_US)
{intestinal, lymphoma}	Intestinal lymphoma (SNOMEDCT_US)
{adrenal, carcinoma}	Adrenal carcinoma (OMIM)
{in, breast, carcinoma, situ}	breast carcinoma in situ (CHV)
{fossa, piriform}	Piriform Fossa (MSH)
{cellular, pigmentation}	cellular pigmentation (GO)
{b-cell, cutaneous, lymphoma, primary}	Primary cutaneous B-cell lymphoma (MEDCIN)
{gastric, sarcoma}	gastric sarcoma (MEDCIN)
{adenocarcinoma, breast, metaplasia, with}	breast adenocarcinoma with metaplasia (MEDCIN)

and a concept that is essential for a terminology may not be necessary for another. Further reviews and evaluations by the terminology curators are still required to decide whether a concept is meaningful and should be added according to the scope of the terminology and its potential applications.

Although we have found supporting evidence (i.e., matching concept names) in the UMLS for a certain portion of potentially missing concepts, further work is still needed to name the remaining concepts according to their lexical feature sets. We plan to experiment with two ideas. One is to maintain the order or sequence of words in concept names while performing the multistage intersection in FCA. The other is to leverage our previous work [15] on predicting concept names using deep learning approaches given bags of words.

A limitation of using words in concept names as the FCA attributes is that the "subconcept-superconcept" relation derived may be different from the hierarchical IS-A relation in the original terminology. For instance, *Breast Neoplasm* and *Breast* are two new concepts generated based on the FCA formal context in Fig. 2. Although the two concepts have a "subconcept-superconcept" relation in terms of the FCA word attributes, they do not form a valid IS-A relation. In fact, *Breast* locates in a different sub-hierarchy *Organ*. A potential solution to avoid such cases is to use enriched lexical features for a concept, which includes its ancestor's lexical features. This way, the original hierarchical relation will be captured in the initial FCA formal context, and thus the new concepts generated by attribute set intersection will locate within the same sub-hierarchy with the root concept. However, the enriched lexical features may make it more difficult to decide which words to use for naming a concept. To deal with this, we plan to leverage both logical definitions and lexical features to identify and naming missing concepts.

In addition, we only performed a preliminary evaluation to automatically validate potentially missing concepts using UMLS. In future work, we will invite domain experts to perform manual evaluation to validate potentially missing concepts identified by our FCA-based method.

VI. CONCLUSION

In this paper, we introduced a lexical- and FCA-based method that utilizes intrinsic knowledge of a terminology to detect potentially missing concepts. We applied our method to the NCI Thesaurus *Disease or Disorder* sub-hierarchy and identified 8,983 potentially missing concepts. The preliminary evaluation via external validation using UMLS showed encouraging evidence for the effectiveness of our method.

REFERENCES

[1] O. Bodenreider, "Biomedical ontologies in action: role in knowledge management, data integration and decision support," *Yearbook of medical informatics*, p. 67, 2008.

[2] S. De Coronado, M. W. Haber, N. Sioutos, M. S. Tuttle, L. W. Wright *et al.*, "Nci thesaurus: using science-based terminology to integrate cancer research results." in *Medinfo*, 2004, pp. 33–37.

[3] G. Fragoso, S. de Coronado, M. Haber, F. Hartel, and L. Wright, "Overview and utilization of the nci thesaurus," *International Journal of Genomics*, vol. 5, no. 8, pp. 648–654, 2004.

[4] N. Sioutos, S. de Coronado, M. W. Haber, F. W. Hartel, W.-L. Shaiu, and L. W. Wright, "Nci thesaurus: a semantic model integrating cancer-related clinical and molecular information," *Journal of biomedical informatics*, vol. 40, no. 1, pp. 30–43, 2007.

[5] M. A. Haendel, J. A. McMurry, R. Relevo, C. J. Mungall, P. N. Robinson, and C. G. Chute, "A census of disease ontologies," *Annual Review of Biomedical Data Science*, vol. 1, pp. 305–331, 2018.

[6] L. Cui, S. Tao, and G.-Q. Zhang, "Biomedical ontology quality assurance using a big data approach," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 10, no. 4, p. 41, 2016.

[7] "Overview of nci thesaurus," https://wiki.nci.nih.gov/pages/viewpage.action?pageId=7472532, [Online; Accessed October, 2020].

[8] P. Chandar, A. Yaman, J. Hoxha, Z. He, and C. Weng, "Similarity-based recommendation of new concepts to a terminology," in *AMIA Annual Symposium Proceedings*, vol. 2015. American Medical Informatics Association, 2015, p. 386.

[9] J. Peng, T. Wang, J. Wang, Y. Wang, and J. Chen, "Extending gene ontology with gene association networks," *Bioinformatics*, vol. 32, no. 8, pp. 1185–1194, 2016.

[10] Z. He, J. Geller, and Y. Chen, "A comparative analysis of the density of the snomed ct conceptual content for semantic harmonization," *Artificial intelligence in medicine*, vol. 64, no. 1, pp. 29–40, 2015.

[11] Z. He, Y. Chen, S. de Coronado, K. Piskorski, and J. Geller, "Topological-pattern-based recommendation of umls concepts for national cancer institute thesaurus," in *AMIA Annual Symposium Proceedings*, vol. 2016. American Medical Informatics Association, 2016, p. 618.

[12] G. Jiang and C. G. Chute, "Auditing the semantic completeness of snomed ct using formal concept analysis," *Journal of the American Medical Informatics Association*, vol. 16, no. 1, pp. 89–102, 2009.

[13] Z. Wei, C. Licong, and Z. Guo-Qiang, "Spark-mca: Large-scale, exhaustive formal concept analysis for evaluating the semantic completeness of snomed ct," in *AMIA Annual Symposium Proceedings*, vol. 2017. American Medical Informatics Association, 2017, p. 1931.

[14] 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," *Journal of the American Medical Informatics Association*, vol. 24, no. 4, pp. 788–798, 2017.

[15] F. Zheng and L. Cui, "Exploring deep learning-based approaches for predicting concept names in snomed ct," in *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2018, pp. 808–813.

[16] D. I. Ignatov, "Introduction to formal concept analysis and its applications in information retrieval and related fields," in *Russian Summer School in Information Retrieval*. Springer, 2014, pp. 42–141.

[17] B. Ganter and R. Wille, *Formal concept analysis: mathematical foundations*. Springer Science & Business Media, 2012.

[18] "Lexical tools: Luinorm," https://lexsrv3.nlm.nih.gov/LexSysGroup/Projects/lvg/2020/docs/userDoc/tools/luiNorm.html, [Online; Accessed October, 2020].

[19] A. D. Troy, G.-Q. Zhang, and Y. Tian, "Faster concept analysis," in *International Conference on Conceptual Structures*. Springer, 2007, pp. 206–219.

[20] O. Bodenreider, "The unified medical language system (umls): integrating biomedical terminology," *Nucleic acids research*, vol. 32, no. suppl_1, pp. D267–D270, 2004.