How to Summarize Big Knowledge Subjects

Ling Zheng, Yehoshua Perl, James Geller,
Gai Elhanan
NJIT, Newark, NJ USA
{lz265, perl, james.geller, gai.elhanan}@njit.edu

***Abstract*—One manifestation of the "Big Knowledge'' challenge is providing automated tools for summarization of ontology content to facilitate user comprehension. An aggregation approach for the automatic identification and display of major subjects covered by an ontology's content is presented. The results show that our methodology is viable in capturing the "big picture" of ontology content.**

I. Introduction

The purpose of the Big Data to Knowledge (BD2K) initiative is to develop methodologies and techniques for mining knowledge hidden in large data repositories. The next challenge facing the scientific community is to enable effective use of the knowledge discovered in BD2K research [1]. Without orientation into the content of knowledge, no effective or innovative use of the knowledge is possible [2]. However, if the knowledge itself is "Big Knowledge," orientation becomes difficult, due to the structure of knowledge repositories. While data repositories often have a tabular structure, Big Knowledge (BK) is usually organized as a large network, which is harder to comprehend.

In this paper, we refer to Assertional BK (ABK), consisting of triples of two concepts connected by a relationship of a given semantics. ABK repositories are found in ontologies and terminologies. The complex network structure of ABK stems from concepts being unique and participating in several (or many) triples. Using visualization methods where concepts are boxes and relationships are arrows connecting the boxes, SNOMED CT, a widely used clinical terminology with about 320,000 active concepts, appears as an incomprehensibly large and complex network diagram. SNOMED CT is indeed an ABK repository.

II. Methods

It is customary to use *summaries* to obtain an orientation into Big Data. For example, in a large drug repository, there may be *x* Antibiotics, *y* Beta blockers, etc., but how can we summarize ABK to obtain an orientation into its content? In this poster, we concentrate on the orientation aspect of *identifying important subjects* in a large ontology.

We have previously developed the theory of *Abstraction Networks* [3] for summarizing ABK. For SNOMED CT hierarchies, we have developed a kind of Abstraction Network called ***partial area taxonomy*** (taxonomy for short) [4]. Each *node* of a taxonomy represents a unit (group) of concepts that is named by its root concept. This reflects the structure of the terminology well. However, many groups are small (measured by the number of concepts), and one cannot *see the forest for the trees*, i.e., one cannot perceive the summary due to too many small groups. Hence, a more compact summary capturing mainly the large units is needed.

However, if we remove all units below a given size *b*, then a large portion of the knowledge is not accounted for. To remedy this problem, we can aggregate descendant units with fewer than *b* concepts ("small units") into the closest ancestor unit with at least *b* concepts (a "large unit"). By varying the integer parameter *b*, we can control the granularity of the summary, i.e., how large is the smallest unit in the summary and how many large units are in the summary. Previously [5], we defined this as an *aggregate taxonomy*.

However, there is another problem due to the structure of a *partial area taxonomy* and its dependency on the concepts where new relationships are introduced. We discovered that some important subjects disappear (do not appear in the summary at all) due to the small sizes of their units, in spite of having many small related descendant units of the same subject area. For example, the unit of the subject *Specimen from nervous system* has only 12 concepts, but many more descendant concepts belong to this subject area. The reason is that the unit itself, the root of which has many descendant concepts, is small due to the fact that some children or grandchildren have new relationships and thus are not included in the unit of the root, but introduce their own units.

To overcome this difficulty, we define an **aggregated weight** for each unit. Then the aggregated weight equals the sum of the size *x* of the unit itself and the sizes of all its descendant units smaller than *x*. In this way, the decision which "small units" to eliminate from the summary can now be based on the aggregated weight of the subject root. For example, the unit *Specimen from nervous system* has 12 concepts and it does not appear in the aggregate taxonomy when $b>12$. However, its aggregated weight is 42, because it has 22 descendant units with fewer than 12 concepts, summarizing 30 descendant concepts. Considering the aggregated weight, the unit *Specimen from nervous system* will appear in the aggregate taxonomy as long as $b<=42$.

We tested this idea for the *Specimen* hierarchy of SNOMED CT. A domain expert MD with extensive experience in ontologies (G.E.) identified 21 major subjects for *Specimen* as a gold standard list. They were mapped to the closest *Specimen* concepts. The partial area containing each such concept is listed in Table 1, followed by its size and aggregated weight (weight for short). The aggregated weight is used to decide whether the partial area and its subject appear in the aggregate taxonomy if the aggregated weight $>= b$ for various values of *b*. Thus, small units will be eliminated from the aggregate taxonomy based on their aggregated weights rather than their own sizes. In this way, the unit *Specimen from nervous system* will not be eliminated from the summary representation.

III. Results

A subject is identified by our methodology if the corresponding partial area appears in the aggregate taxonomy for parameter *b*. The results and the corresponding recall R, precision P and F values are listed at the bottom of Table 1 for various *b* values. The optimal aggregate taxonomy (Fig. 1) for identifying subjects is obtained for the maximum F value=0.48 for *b*=25. Twelve out of 21 subjects (R=0.57) are identified among the 29 partial areas (P=0.41) of the aggregate taxonomy. The 12 subject partial areas identified are highlighted in yellow in Fig. 1. We note that both recall and precision are important. The recall gives the portion of the original subject list identified, while the precision gives the ratio of the aggregate taxonomy units which qualify as important subjects. We optimize F, which combines both of them symmetrically. Fig. 1 was shown to G.E. who after inspecting it determined that 13 more units (highlighted in pink) qualify as important subjects. The original list should have 21+13=34 subjects, 25 of which were identified. Hence, for this enhancement R is 0.74 (=25/34), P is 0.86 (=25/29) and F is 0.79, which improve the original results. Thus, the methodology was shown successful in automatically identifying most of the subjects in a sample of ABK, an imperative BK challenge, by summarizing the important subjects in a SNOMED CT hierarchy.

Table 1. Identification results for $S = 21$ chosen subjects in aggregate taxonomies with different thresholds *b* (spc is short for specimen and smp for sample).

Subject	Concept	Partial-area	Size (Weight)	*b*=1	*b*=5	*b*=10	*b*=15	*b*=20	***b*=25**	*b*=30
Blood spc	*Blood spc*	*Blood spc*	28 (43)	✓	✓	✓	✓	✓	✓	✓
Body substance smp	*Body substance smp*	*Body substance smp*	63 (498)	✓	✓	✓	✓	✓	✓	✓
Fluid smp	*Fluid smp*	*Fluid smp*	50 (257)	✓	✓	✓	✓	✓	✓	✓
Bone marrow spc	*Bone marrow spc*	*Bone marrow spc*	8 (13)	✓	✓	✓	–	–	–	–
Bone spc	*Spc from bone*	*Musculoskeletal smp*	15 (44)	–	–	–	–	–	–	–
Spc from nervous system	*Spc from nervous system*	*Spc from nervous system*	12 (42)	✓	✓	✓	✓	✓	✓	✓
Dermatological spc	*Dermatological smp*	*Dermatological smp*	8 (30)	✓	✓	✓	✓	✓	✓	✓
Device spc	*Device spc*	*Device spc*	19 (40)	✓	✓	✓	✓	✓	✓	✓
Digestive system spc	*Spc from digestive system*	*Spc from digestive system*	50 (126)	✓	✓	✓	✓	✓	✓	✓
Endocrine system spc	*Endocrine smp*	*Endocrine smp*	10 (26)	✓	✓	✓	✓	✓	✓	–
Genital system spc, male	*Male genital smp*	*Spc from trunk*	132 (489)	–	–	–	–	–	–	–
Genitourinary spc	*Genitourinary smp*	*Spc from trunk*	132 (489)	–	–	–	–	–	–	–
Hair spc, scalp	*Hair spc*	*Dermatological smp*	8 (30)	–	–	–	–	–	–	–
Musculoskeletal spc	*Musculoskeletal smp*	*Musculoskeletal smp*	15 (56)	✓	✓	✓	✓	✓	✓	✓
Skin spc	*Spc from skin*	*Dermatological smp*	8 (30)	–	–	–	–	–	–	–
Soft tissue spc	*Soft tissue smp*	*Soft tissue smp*	21 (92)	✓	✓	✓	✓	✓	✓	✓
Cardiovascular smp	*Cardiovascular smp*	*Cardiovascular smp*	12 (28)	✓	✓	✓	✓	✓	✓	–
Spc from eye	*Spc from eye*	*Spc from head and neck structure*	53 (196)	–	–	–	–	–	–	–
Spc from joint	*Joint smp*	*Musculoskeletal smp*	15 (56)	–	–	–	–	–	–	–
Lesion smp	*Lesion smp*	*Lesion smp*	17 (118)	✓	✓	✓	✓	✓	✓	✓
Stool spc	*Stool spc*	*Body substance smp*	63 (498)	–	–	–	–	–	–	–
# Identified subjects (*C*)				13	13	13	12	12	12	10
# Partial-areas (*A*)				503	89	54	40	35	29	26
Recall (*R* = *C*/*S*)				0.62	0.62	0.62	0.57	0.57	0.57	0.48
Precision (*P* = *C*/*A*)				0.03	0.15	0.24	0.30	0.34	0.41	0.38
F = 2·*P*·*R*/(*P*+*R*)				0.05	0.24	0.35	0.39	0.43	**0.48**	0.43

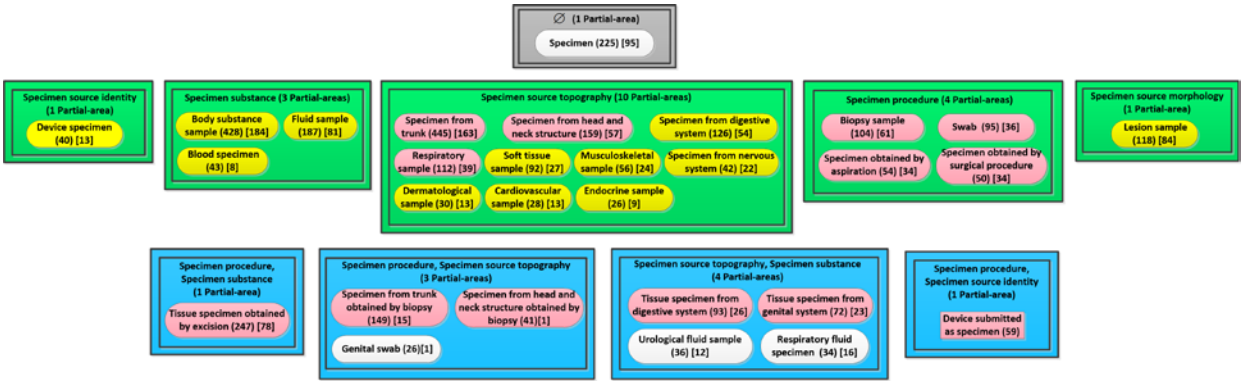

Fig. 1. Aggregate taxonomy for the *Specimen* hierarchy with $b = 25$. The 12 partial-areas corresponding to prescribed subjects are highlighted in yellow. The 13 partial-areas added during the enhancement are highlighted in pink.

References

[1] Geller, J., Perl, Y., Halper, M., et al. The Big Knowledge to Use (BK2U) Challenge. Workshop on Data Science, Learning and Applications to Biomedical and Health Sciences (DSLA-BHS). 2016.

[2] Patel, V. L., Kaufman, D. R., Arocha, J. Conceptual change in the biomedical and health sciences domain. Advances in instructional psychology. 2000; 5:329-392.

[3] Halper, M., Gu, H., Perl, Y., et al. Abstraction networks for terminologies: Supporting management of "big knowledge". Artif Intell Med. 2015; 64(1):1-16.

[4] Wang, Y., Halper, M., Min, H., et al. Structural methodologies for auditing SNOMED. J Biomed Inform. 2007; 40(5):561-581.

[5] Ochs, C., Perl, Y., Geller, J., et al. Using aggregate taxonomies to summarize SNOMED CT evolution. IEEE International Conference on Bioinformatics and Biomedicine (BIBM). 2015:1008 - 1015.