Using UMLS Metathesaurus Relations for Managing Biomedical Knowledge from MEDLINE Citations

Liqin Wang
Mentor: Olivier Bodenreider

Abstract
Assertional knowledge captured by SemRep from MEDLINE citations could be overwhelming, unorganized, and of different level of granularity. The hierarchical relations from the UMLS Metathesaurus was used to cluster similar concepts, to estimate the information content of concepts to determine the most representative concepts for each cluster, and to aggregate information within a cluster. In this work, we analyze the effectiveness of using UMLS relations for the management of biomedical knowledge by applying this to the task of summarizing biomedical knowledge.

1. Introduction

MEDLINE citations contain an overwhelming amount of biomedical knowledge, part of which has been captured by SemRep (Rindflesch, 2003 and 2005) and represented as subject-predicate-object predications. This structured representation of knowledge is simple and compatible by design with similar relations from the UMLS Metathesaurus. By linking with the biomedical literature, this structure representation enables a variety of advanced applications, including information retrieval, abstractive summarization of biomedical literature (Fiszman, 2009), literature-based knowledge discovery (Miller 2012), and clinical question answering (Chambliss, 1996; Demner-Fushman and Lin, 2007). However, several issues have hindered the using or managing those semantic predications from MEDLINE citations. As we understand, the information needs to be effectively organized and analyzed in order to be useful. However, from a simple query of biomedical literature for a given topic or domain, it could turn to a sheer number of predications. Those predications could cover many domains, like genetics, disease, pharmacology, procedure, etc., but in an unorganized way. Another issue is that information is of different level of granularity. In fact, many predications are somewhat connected, although they are actually not sharing any arguments, simply because their arguments exhibit differences in granularity. For the same reason, some predications can be redundant. For example, the predication {ACE Inhibitor TREATS Heart failure} is somewhat redundant with {Enalapril TREATS Heart failure} since {Enalapril ISA ACE Inhibitor}. Moreover, the aggregation of knowledge at a coarser level is sometimes required in order to increase the confidence of a given assertion. Previous studies have focused on condensing large graphs of predications based on the principles of relevance, connectivity, novelty, and saliency. However, many patterns represented in these graphs could not be revealed without enriching these graphs with hierarchical information.

In this study, we start with a graph built from semantic predications, and enrich the graph with UMLS hierarchical relations which could actually categorize those concepts into clusters, and estimate the information content of the concepts for a better understanding of the level of granularity. The enriched graph performs as an underground for information aggregation which

the graph will be condense and summarized. We believe that this study would benefit for the management of large and unorganized graphs and further used to support advanced applications relying on such graphs.

2. Background

2.1. Unified medical language system (UMLS)

UMLS is a repository of biomedical vocabularies developed at National Library of Medicine (Lindberg 1993, Bodenreider 2004). It has been regarded as a key medical knowledge resource in the biomedical domain. UMLS consists of three main knowledge sources: the Metathesaurus, the Semantic Network, and the SPECIALIST Lexicon. UMLS Metathesaurus is the largest knowledge source of these three, which integrating up to 169 source terminologies into a unified terminology system, including SNOMED CT, ICD, RxNorm, Medical Subject Headings (MeSH). The Metathesaurus essentially contains definitional knowledge, for example, congestive heart failure finding_site_of cardiac ventricle. One kind of definitional knowledge is hierarchical relations indicated by Parent (PAR), Children (CHD), Narrower Than (RN), and Broader Than (RB), likr, *congestive heart failure PAR heart failure*. Another knowledge source in the UMLS is Semantic Network, which provides high-level categories used to categorize every Metathesaurus concept and all possible relations between any two concepts in the biomedical domain. Currently, semantic network contains 133 semantic types and 54 relations, which constitute 6952 semantic predications, e.g., {Pathologic Function co-occurs_with Mental or Behavioral Dysfunction}. Semantic groups (McCray 2001) further categorize these semantic types into 15 groups, such as Disorders, Devices, Procedures, Anatomy, etc. One concept can belongs to multiple semantic types; however, it can only belong to one semantic group.

2.2. MEDLINE

MEDLINE is a bibliographic database that contains journal citations for biomedical literature from the world around. Currently, there are over 20 million citations in the MEDLINE database. MEDLINE is semi-structured database; the titles and abstracts of biomedical articles in the database are free-text.

MEDLINE has gained many researchers' interests during the past two decades; a main reason is that it contains a vast amount of biomedical knowledge. The research questions under activities are including document retrieval, clinical question answering (Sneiderman, 2007), knowledge extraction (Mendonca, 2000, Craven, 1999) and knowledge discovery. Along with the growth of MEDLINE database, good query strategies and summarization of retrieved biomedical literature become a necessary. As well, users are not satisfied with a list of documents, instead, they prefer an application could answer their question by using the data from MEDLINE. All these needs are motivating many researches and application developments on top of MEDLINE.

2.3. SemRep and SemMedDB

SemRep (Rindflesch 2003 and 2005) is a knowledge-based natural language processing tool that is developed at US National Library of Medicine. The main function of the SemRep is to

capture the relations between concepts that have been identified by MetaMap (Aronson 2010) from narrative sentence, and generate semantic predications.

The SemRep has been used to extract semantic predications from MEDLINE citations (including titles and abstracts). All these predications are stored in a repository called SemMedDB. There are, currently, over 57 million semantic predications in the database. 93% of predications are associative predications (or, none “ISA” predication). SemMedDB essentially contains assertional knowledge meaning most predications are only true in certain context. For example, the predication {Enalapril treats congestive heart failure} might be true only for a certain population.

Facilitated by the availability of SemRep that the free-text sentence could be represented as semantic predications, Fiszman and his colleagues are working on abstractive summarization. They proposed methods to identify the important predications based on the principles of relevance, connectivity, novelty, and saliency, which thus is able to condense large group by remove relative unimportant piece of knowledge.

2.4. Information content

The information content (IC) is a measure of the amount of information a concept contained in certain context. It provides a numerical score to estimate the degree of generality/specialty of a concept. This quantitative measure improves the understanding of the concepts when they come with similar meanings. In fact, the most common use of IC is the computation of semantic similarity of pairs of terms. Resnik firstly applied IC to the computation of semantic similarity [Resnik, 1995], where he computed the IC by using the propagated frequency of the concept according to the taxonomical hierarchy. Sanchez [Sanchez, 2011] proposed a new approach for the IC computation, which is using ontology since it has defined and organized the concept in a meaningful way. This method has been evaluated on the task of calculating semantic similarity and it outperforms others methods so far.

The IC of *a* concept was defined by Sanchez et al as:

$$IC(a) = -\log\left(\frac{\frac{|leaves(a)|}{|subsumers(a)|} + 1}{max_leaves + 1}\right)$$

Where max_leaves represents the number of leaves corresponding to the root node of the hierarchy; $leaves(a)$ is the number of concepts that are lower than concept *a* in the hierarchy; and $subsumers(a)$ is a set of subsumers of concept *a* (including *a*).

$$subsumers(a) = \{c \in C | a \leq c\} \cup \{a\}$$

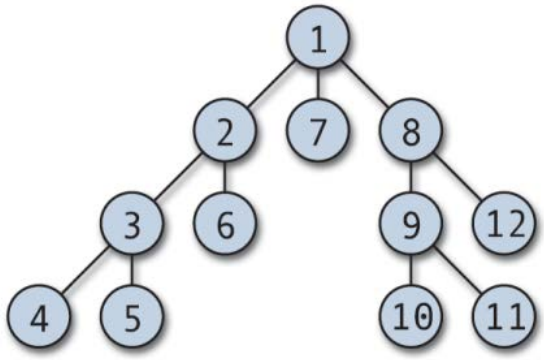

Fig 1. An make up graph for the computation of information content

For example, we calculated the IC of concept 2 (see Fig.1.) by using the formula above, then the IC(2) is: $IC(2) = -log_{(\max_leaves+1)}\left(\frac{\frac{leaves(2)}{subsumers(2)}+1}{\max_leaves+1}\right) = -log_{12}\left(\frac{\frac{4}{1+1}+1}{11+1}\right) = 0.558$. In order to normalize the value from [0,1], we set the base of log as the (max_leaves+1).

3. Methods

Our method has three processes. First, we build a graph based on the semantic predications which are retrieved from SemMedDB. The query expansion can be used to expand the graph with more relevant information. The second part of the process is to enrich the graph with proper number of edges and nodes, where the edges are actually UMLS upper-level hierarchical relations and nodes are those concepts as part of the hierarchical relations. Afterwards, concepts of hierarchical relations will be grouped in to a cluster. Third, for the aggregation, we will propagate information from the descendant concepts to a same ancestor concept in each cluster, as well as establishing new relations between the cluster and the topic concepts. Those processes will be given a more detailed explanation in the following paragraphs.

3.1. Build an original graph

There are two primary approaches to query SemMedDB, document-oriented and concept-oriented query. Document-oriented query is to first retrieve relevant documents for a given topic, and then to retrieve the semantic predications from SemMedDB for those documents. The concept-oriented query is to retrieve all semantic predications from SemMedDB if the predication contains this concept. The difference between graphs built from these two queries is that the former does not have central concept while the later one have pre-defined central concept(s). Our method will be generable to both query approaches. But, for the sake of simplicity, we build the graph based on concept-oriented query.

First of all, we take congestive heart failure (CHF) as our topic; CHF is a condition in which the heart can no longer pump enough blood to the rest of the body. Query expansion is a process that expands the query where we are not only retrieving semantic predications based on one concept but also its hyponyms. Hyponyms of a concept are indicated in the UMLS

Metathesaurus through hierarchical relations. For the sake of simplicity, we skip the query expansion and only retrieve associative predications for the concept of CHF from SemMedDB. For this initial study, we restricted the predicate to “TREATS” and “PREVENTS”. We also remove uninformative predications based on the novelty information provided by SemMedDB. In this process, we also obtain the occurrence of each predication in the entire database. After this, we build an original graph from those associative predications.

3.2. Graph enrichment and clustering

Graph enrichment is a process to enrich the graph with proper connections by adding adequate nodes and edges in order to be able to cluster those concepts of the original graph. We will explain this process into several sub-processes.

3.2.1. Adding UMLS hierarchical relations

Excluding the central concept, there would be no connection between any two concepts in the original graph. In order to identify any possible connections between any two concepts, we add UMLS upper-level hierarchical relations.

3.2.2. Pruning

Prune is a process to remove any new added concepts if they are not sharing any ancestors with the concepts from the original graph, since they are less meaningful to the graph if they do not help with the clustering of those original concepts.

3.2.3. Breaking up large clusters

When the top-level concept of cluster is too general, we will consider breaking them up in order to maintain the cluster as meaningful as possible. However, before any real breaking up process, we need to measure the IC of concepts in order to determine whether they are general enough for breaking up the clusters that are containing them. Currently, there is no threshold of IC to separate the concepts into general or special.

3.2.4. Clustering

Cluster is the process to cluster concepts that are hierarchical related or of same direct ancestor will be put into a same cluster. In the case that one concept can have more than one ancestor, thus, we allow one concept have membership to multiple clusters.

3.3. Information aggregation

Aggregation is the process that aggregate individual concepts into collections which could condense the graph proportional. The information of these concepts in a cluster will propagate to the most representative concepts in the graph. Usually we define the most representative concepts as the highest level concept in a cluster.

After propagate the information to a single concept assuming concept A, we then break up all the relations between the concepts in the cluster and the concept that all these concepts are

related to, assuming the concept B. Then, a new relation will be established between the cluster and concept B. We can name the cluster with the name of most representative concept B in that cluster, so instead of having many relations between the concepts in the cluster with concept A, we will only have A and B as well as the information from all the concepts in the cluster.

After the aggregation, we could actually condense the graph tremendously and the graph would be something looks like Fig 2.

4. Results

With the query of SemMedDB by the concept of congestive heart failure, we retrieve 924 concepts and 971 predications among of which 95.78% are “TREATS” predications. We use those concepts and predications to build an original graph (see Fig. 3) that every concept only connect to congestive heart failure and the strength of the edges represents the occurrence of predication in the SemMedDB.

In order to see more details, we separate the steps of enrichment. First, we retrieved all the directional relations between any two concepts in the graph, which is 1060 hierarchical relations and affecting on totally 617 concepts in the graph (Fig. 4). Thus, there are still 308 concepts not connecting to any other concepts which could be seen as a central circle in Fig. 3. After adding hierarchical relations, the concepts that are closely related will automatically form clusters.

Then, we add one-hop hierarchical relations for all the concepts in the graph (excluding the CHF) and also new concepts involved in these hierarchical relations. In this step, we found that 1652 concepts could be added to the graph. However, after the pruning process, we found that only 527 of 1652 concepts are remaining in the graph and over 150 concepts became connected through those 527 concepts, while 150 concepts were still not connected to any other concepts from the original graph (see Fig. 5).

5. Discussion

We have demonstrated that the use of UMLS hierarchical relations to enrich graphs from SemMedDB. According preliminary results, we have seen that most concepts in the graph could be connected to others after the enrichment.

There are many remaining issues in this work which we try to solve in the near future. For example, we could not determine the threshold for the cut-off of IC when determine the breaking up of large clusters. We also would like to develop a use cases for demonstrating the usefulness of information aggregation for navigation, knowledge extraction, summarization, etc.

6. Conclusion

We conclude out work that graph enrichment and aggregation would potentially facilitate the management of large and unorganized graphs. They will further benefit some advanced applications relying on such graphs, including knowledge extraction, knowledge summarization, etc.

Acknowledgments

This research was supported in part by an appointment to the NLM Research Participation Program sponsored by the National Library of Medicine and administered by the Oak Ridge Institute for Science and Education.

References

1. Sahoo SS, Zeng K, Bodenreider O, Sheth AP. From "glycosyltransferase" to "congenital muscular dystrophy": Integrating knowledge from NCBI Entrez Gene and the Gene Ontology. Stud Health Technol Inform (Proc Medinfo) 2007;129(Pt 1):1260-1264.
2. Kelly Zeng, Olivier Bodenreider. Integrating the umls into an rdf-based biomedical knowledge repository. AMIA 2007 Poster.
3. Brachman RJ. What IS-A is and isn't: an analysis of taxonomic links in semantic networks. Computer. 1983;16(10):30–6.
4. Bodenreider O, Burgun A, Rindflesch TC. Lexically-suggested hyponymic relations among medical terms and their representation in the UMLS. Proc Conf Terminology and Artificial Intelligence. 2001:11–21.
5. Han Zhang, Marcelo Fiszman, Dongwook Shin, Christopher M Miller, Graciela Rosemblat, Thomas C Rindflesch. Degree centrality for semantic abstraction summarization of therapeutic studies. J Biomed Inform; 2011: 830-838.
6. Marcelo Fiszman, Thomas C. Rindflesch, Halil Kilicoglu. Abstraction Summarization for Managing the Biomedical Research Literature. Proceedings of the Workshop on Computational Lexical Semantics. 2004:76–83.
7. Fiszman M, Demner-Fushman D, Kilicoglu H, Rindflesch TC. Automatic summarization of MEDLINE citations for evidence-based medical treatment: a topic-oriented evaluation. J Biomed Inform. 2009 Oct;42(5):801-13. Epub 2008 Nov 5.
8. Miller, Christopher M.; Thomas C. Rindflesch; Marcelo Fiszman; Dimitar Hristovski; Dongwook Shin; Graciela Rosemblat; Han Zhang; Kingman P. Strohl. 2012. A closed literature-based discovery technique finds a mechanistic link between hypogonadism and diminished sleep quality in aging men. SLEEP 35(2):278-85.
9. Rindflesch, TC.; Fiszman, M.; Libbus, B. Semantic interpretation for the biomedical research literature. In: Chen, H.; Fuller, S.; Hersh, W.; Friedman, C., editors. Medical informatics: knowledge management and data mining in biomedicine. Springer; New York: 2005. p. 399-422.
10. Rindflesch TC, Fiszman M. The interaction of domain knowledge and linguistic structure in natural language processing: Interpreting hypernymic propositions in biomedical text. J Biomed Inform 2003;36(6):462–77. [PubMed: 14759819]
11. Chambliss ML, Conley J. Answering clinical questions. J Fam Pract. 1996 Aug; 43(2):140-2.
12. Demner-Fushman, D., Lin, J. Answering clinical questions with knowledge-based and statistical techniques. Computational Linguistics, 2007, 33(1): 63-103.
13. J.J. Cimino, G. O. Barnett. Automatic Knowledge Acquisition from MEDLINE. Methods of Information in Medicine, 1993
14. E. A. Mendonça, J. J. Cimino. Automated knowledge extraction from MEDLINE citations. Proc AMIA Symp. 2000: 575–579.
15. Charles A Sneiderman, Dina Demner-Fushman, Marcelo Fiszman, Nicholas C Ide, Thomas C Rindflesch. Knowledge-based Methods to Help Clinicians Find Answers in MEDLINE. J Am Med Inform Assoc 2007;14:772-780
16. Marcelo Fiszman, Thomas C. Rindflesch, Halil Kilicoglu. Abstraction Summarization for Managing the Biomedical Research Literature. Proceedings of the Workshop on Computational Lexical Semantics. 2004:76–83.

17. Han Zhang, Marcelo Fiszman, Dongwook Shin, Christopher M Miller, Graciela Rosemblat, Thomas C Rindflesch. Degree centrality for semantic abstraction summarization of therapeutic studies. J Biomed Inform; 2011: 830-838.
18. Fiszman M, Demner-Fushman D, Kilicoglu H, Rindflesch TC. Automatic summarization of MEDLINE citations for evidence-based medical treatment: a topic-oriented evaluation. J Biomed Inform. 2009 Oct;42(5):801-13. Epub 2008 Nov 5.
19. Nikolai Daraselia*, Anton Yuryev, Sergei Egorov, Svetalana Novichkova, Alexander Nikitin, Ilya Mazo. Extracting human protein interactions from MEDLINE using a full-sentence parser. Bioinformatics (2004) 20(5):604-611
20. Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic acids research. 2004;32(suppl 1):D267.
21. Zeng Q, Cimino JJ. Automated knowledge extraction from the UMLS. Proc AMIA Symp. 1998:568-72.
22. Mendonca EA, Cimino JJ. Automated knowledge extraction from MEDLINE citations. Proc AMIA Symp. 2000:575-9.
23. Craven M, Kumlien J. 1999. Constructing biological knowledge bases by extracting information from text sources. Proc. Int. Conf. Intel. Syst. Mol. Biol., 7th, Heidelberg, 1999, pp.77-86. Menlo Park, CA:AAAI Press.
24. Chen ES, Hripcsak G, Xu H, Markatou M, Friedman C. Automated acquisition of disease drug knowledge from biomedical and clinical documents: an initial study. J Am Med Inform Assoc. 2008 Jan-Feb;15(1):87-98. Epub 2007 Oct 18.
25. A.R. Aronson, F.M. Lang. An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc, 17 (3) (2010), pp. 229–236.
26. David Sanchez, Montserrat Batet, David Isern. Ontology-based information content computation. Knowledge-based systems 24(2011) 297-303.
27. P. Resnik, Using information content to evaluate semantic similarity in a taxonomy, in: Proc. of 14th International Joint Conference on Artificial Intelligence, IJCAI 1995, Morgan Kaufmann Publishers Inc., Montreal, Quebec, Canada, 1995, pp. 448–453.
28. McCray, A. T.; Burgun, A.; and Bodenreider, O. 2001. Aggregating UMLS semantic types for reducing conceptual complexity. In Medinfo, volume 10, 216–20.

Appendix

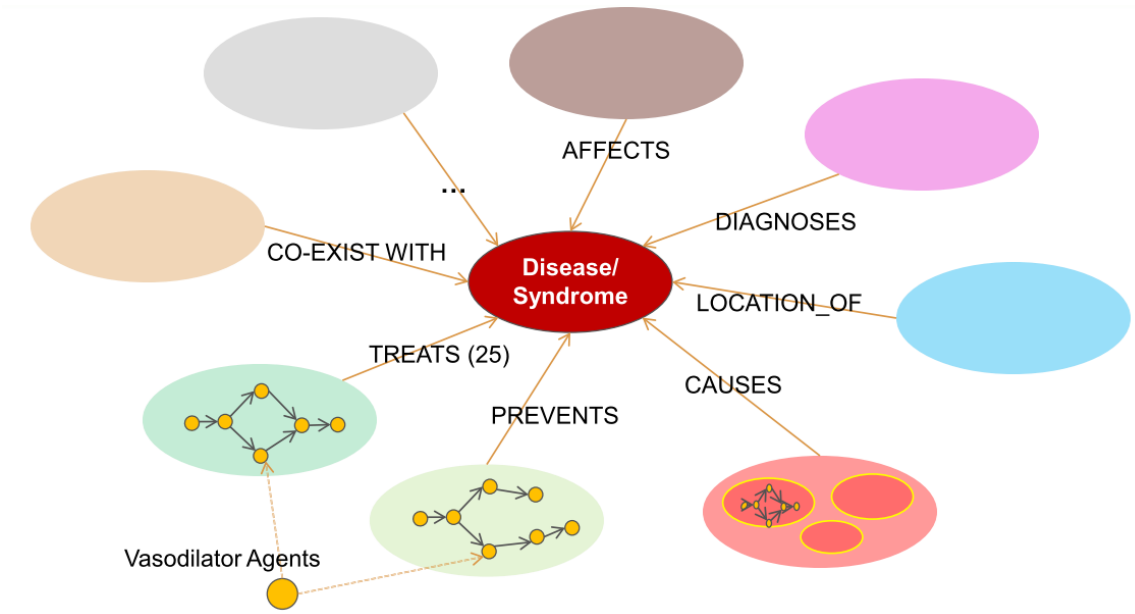

Fig.2. Graphical aggregation architecture according to the predicate types

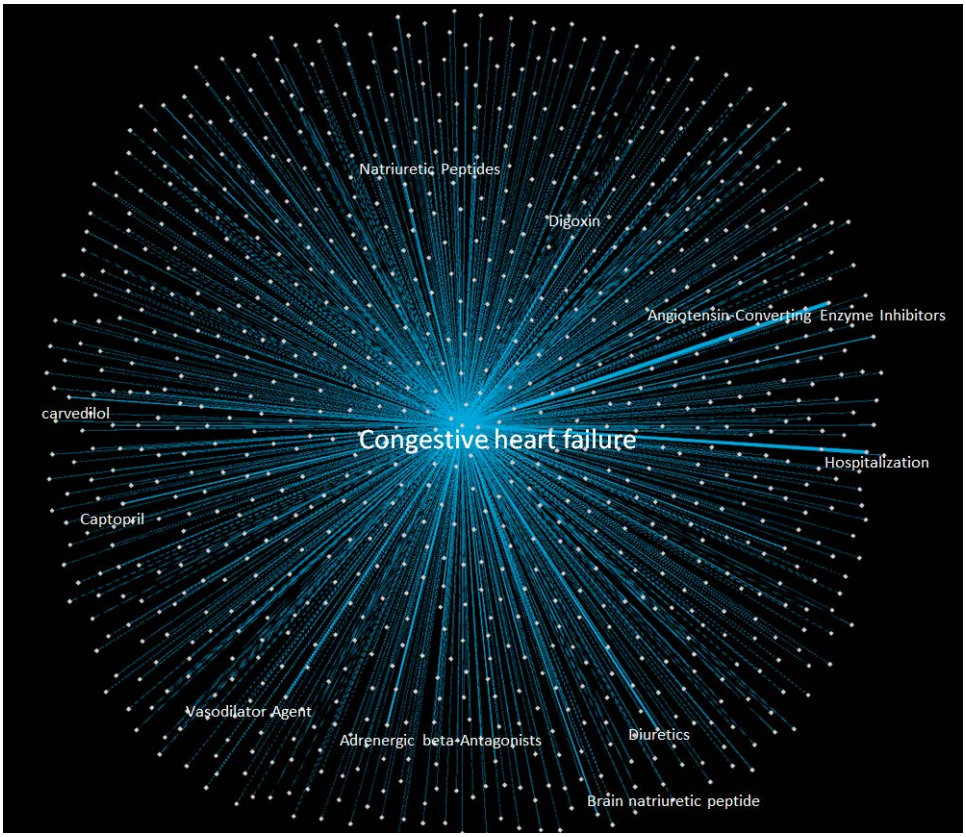

Fig.3. An original graph

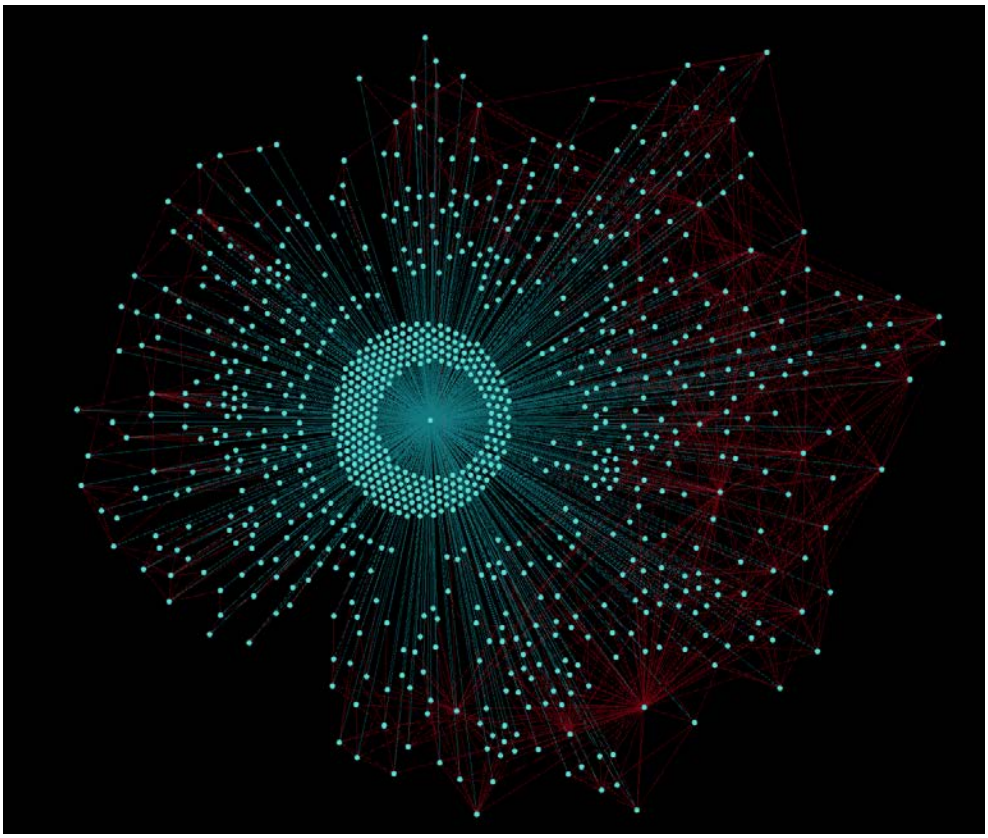

Fig.4. Enriched graph with hierarchical relations among concepts of the original graph

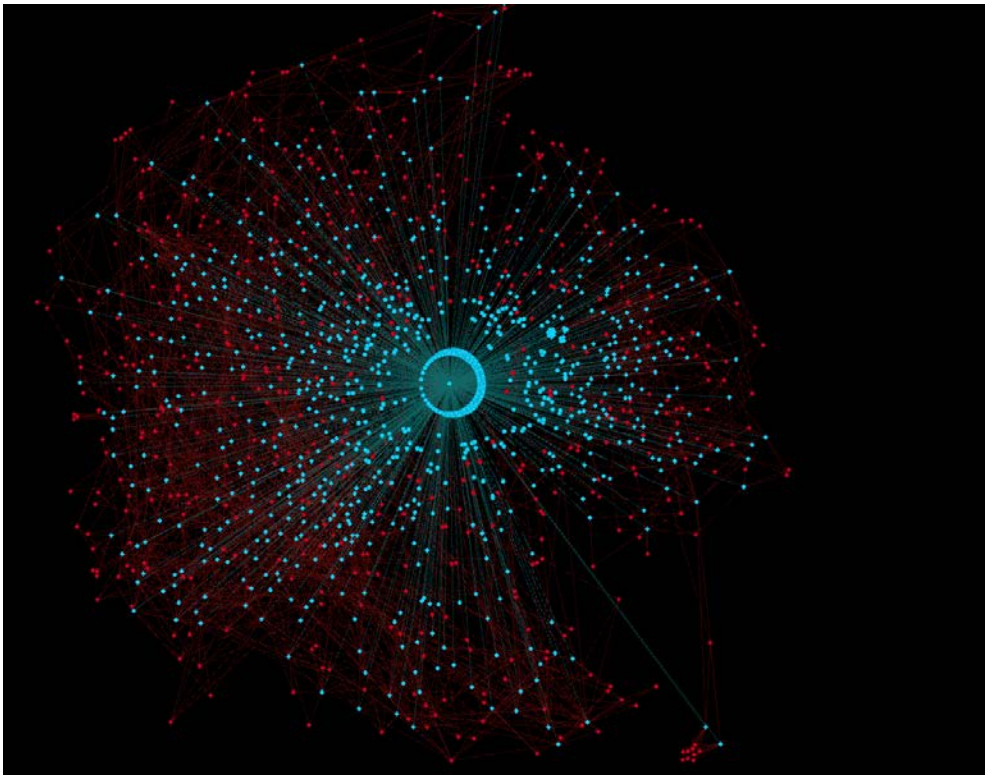

Fig.5. Enriched graph with one-hop hierarchical relations after pruning