

Interactive Visualization and Navigation of Complex Terminology Systems, Exemplified by SNOMED CT

Erik Sundvall^{a,1}, Mikael Nyström^a, Håkan Petersson^a, Hans Åhlfeldt^a
^a Department of Biomedical Engineering, Linköpings universitet, Sweden

Abstract. Free-text queries are natural entries into the exploration of complex terminology systems. The way search results are presented has impact on the user's ability to grasp the overall structure of the system. Complex hierarchies like the one used in SNOMED CT, where nodes have multiple parents (IS-A) and several other relationship types, makes visualization challenging. This paper presents a prototype, *TermViz*, applying well known methods like "focus+context" and self-organizing layouts from the fields of Information Visualization and Graph Drawing to terminologies like SNOMED CT and ICD-10. The user can simultaneously focus on several nodes in the terminologies and then use interactive animated graph navigation and semantic zooming to further explore the terminology systems without losing context. The prototype, based on Open Source Java components, demonstrates how a number of Information Visualisation methods can aid the exploration of medical terminologies with millions of elements and can serve as a base for further development.

Keywords: Terminology, Information Visualization, SNOMED CT, Medical Informatics

1. Introduction

Large terminology systems with complex intertwined structure can be hard to navigate and get acquainted with. Expandable trees like in Figure 1, from the *Clue* browser bundled with the SNOMED CT² distribution, work fine for some tasks, but have limitations when showing structures allowing multiple parents. It is also easy to run out of screen space and lose context if trying to expand and focus on several nodes at once. The goal of the *TermViz* prototype is to provide an alternative terminology exploration tool. The prototype development described here is a project that intersects the research fields of *Information Visualisation*, *Graph Drawing* and *Medical Terminology Systems*.

Information Visualisation (IV) aims to reduce the cognitive effort required to understand abstract information by engaging the human visual perception system, which is often under-utilized in more

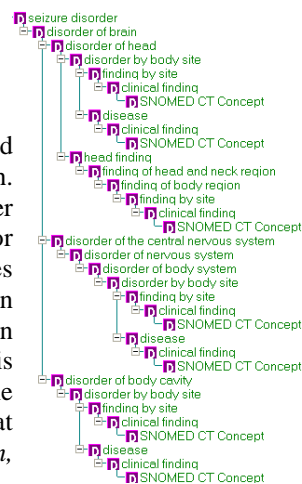


Figure 1: Expandable tree.

¹ Corresponding author: Erik Sundvall, Linköping University, Dept. of Biomedical Engineering, SE-581 85 Linköping, Sweden. erisu@imt.liu.se, <http://www.imt.liu.se/~erisu/>

² <http://www.snomed.org/>

traditional text-oriented systems. Humans can rapidly scan, recognise, and recall images, and can easily detect changes in size, colour, shape, movement, or texture [1]. If structural changes are smoothly animated at a rate that the perceptual system can track, rather than instantly changed, then the cognitive effort of getting oriented in a new modified scene is reduced [2]. An enjoyable user interaction is considered a key element in IV [1]. IV's importance in medical applications has been highlighted by Chittaro [3].

Graph Drawing (GD) is a subfield of mathematics and computer science concerning for example graph theory and layout algorithms. GD also presents algorithms designed to fulfil "aesthetics criteria" like minimizing the number of edge crossings, edge bends and total graph area [4, 5]. The *Graphviz*³ suite is a common GD tool used for example when visualizing parts of the *Gene Ontology*⁴ and *Semantic Web*⁵. Most of the Graphviz-based systems do however mainly present static non-interactive graphs. A static graph generator without edge crossing minimization is also available in *Clue*.

The data sets (terminology systems) in TermViz are directed graphs or networks where the data elements have inherit relations and thus fit into the subfield **Graph Visualisation** [6] which is the intersection of *Information Visualisation* and *Graph Drawing*.

The objective of this paper is to present and discuss the features of the TermViz prototype.

2. Material and methods

The terminology systems to be visualized were converted to and stored in *RDF*⁶ format. Components from the visualization toolkit, *prefuse*⁷, were extended, modified and stringed together to create the visualizations. A surrounding graphical user interface (GUI) with menus, buttons etc. and a caching graph loader bridging the terminology storage and *prefuse* was created for the application. The storage is accessed remotely over a network connection. This enables us to have several visualizing front-ends (including Java Applets) accessing a single terminology server. The architecture also allows accessing and aggregating information from multiple networked information storages of different kinds (for example RDF storages, Relational Databases and UMLS⁸) by creating suitable graph loaders and defining appropriate queries.

The *SNOMED CT* version used was delivered as tables in three "flat files" containing concepts, descriptions and relations. *TermColl* is a collection of English and Swedish versions of five⁹ terminology systems, prepared and imported into a tree-structured database by one of the authors (MN). *TermColl* has so far been used for

³ <http://www.graphviz.org/>

⁴ <http://www.geneontology.org/GO.tools.shtml>

⁵ <http://www.w3.org/2001/11/IsaViz/>

⁶ Resource Description Framework, <http://www.w3.org/RDF/> We used Kowari as storage, <http://kowari.org>

⁷ <http://prefuse.org/> - *prefuse* is intentionally spelled in lower case by its creators.

⁸ <http://www.nlm.nih.gov/research/umls/>

⁹ **ICD-10** and **ICF** (by WHO) **MeSH** (by NLM). **NCSP** (NOMESCO Classification of Surgical Procedures, by the Nordic Medico-Statistical Committee). **KSH97-P** (Swedish Primary Health Care Version of ICD-10, by the Swedish National Board of Health and Welfare).

generating a medical English-Swedish dictionary [7]. Mappings¹⁰ from SNOMED CT to ICD-10 were also converted to RDF and stored.

*prefuse*⁷ is a toolkit for interactive visualisation aimed at Java programmers. It divides the visualization task into a sequence of logical steps that can all be modified by the programmer. Usability studies have been conducted to ensure the toolkits' effectiveness and usability [8].

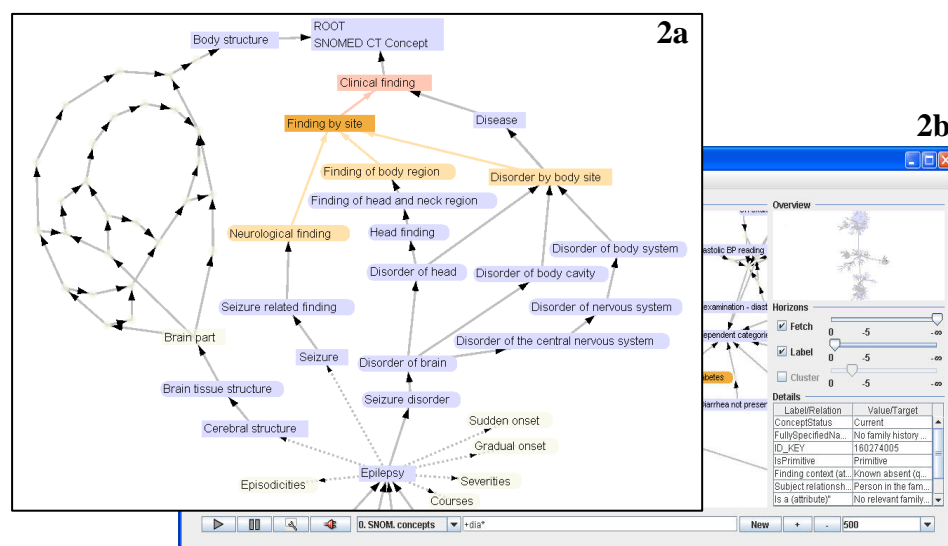


Figure 2a: A subset of concepts related to “Epilepsy” in SNOMED CT. Solid lines represent IS-A relations. Purple¹¹ nodes have been focused by user selections. When the cursor enters a node it becomes highlighted in orange, incoming links/edges turn yellow and outgoing turn red. **Figure 2b:** Parts of the surrounding GUI.

3. Prototype Description¹¹

In TermViz the user can start with an empty screen or by showing a set of predefined root nodes. A node can be expanded by manually selecting to show in- and/or outbound links. Some queries can return thousands of nodes and some SNOMED CT concepts have thousands of incoming links, so the number of hits returned by a query or expansion is bounded by a user-selectable upper limit.

Queries: By entering query text in the field at the bottom of the screen (Figure 2b) the user can find nodes matching the query. The query syntax¹² of the Lucene search engine used resembles that of Google. When issuing a query the results are added to the already visible nodes. Upon query submission the contents of the text and limit fields are inserted into one of the predefined query templates and then submitted for execution. The available query templates have descriptive short labels that are shown in the drop-down list in front of the query field. Templates can for example perform the

¹⁰ Crossmaps from SNOMED CT to ICD-10 delivered with the SNOMED CT distribution

¹¹ If you are reading a copy without colours, refer to corresponding author's homepage for color images.

¹² <http://lucene.apache.org/java/docs/queryparsersyntax.html>

task “Return all SNOMED CT concept nodes that have any description containing the word supplied in the search field”.

Advanced users can modify or add new query templates and preselect a suitable graph loader for the query. This was designed with the purpose of making TermViz an easily evolvable system that can be extended by end users [9].

Query templates can take a node ID as input instead of free-text terms. Such queries can be used for tasks like “show all ICD-10 nodes that are crossmapped from this selected SNOMED CT node”.

Focus sets and automatic graph traversal: A query returns a set of nodes and puts them in a *focus set*. Nodes can also be manually added or removed from a focus set. In Figure 2a the outgoing IS-A links, from all focused (purple¹¹) nodes, have been automatically climbed as far as possible so that a natural root node has been reached. During the climb newly found nodes are continuously being loaded and visualized, and every node is assigned a *degree-of-interest* (DOI) value that decreases when the number of steps from focused nodes increases. The number of steps to climb can be limited by the slider marked *fetch*. The slider adjusts at what DOI-level to stop fetching.

In Figure 2a the node “epilepsy” has been selected and expanded to also show other relationships than IS-A. To avoid cluttering the view, the nodes added this way are not expanded further unless they are manually focused (added to the focus set) by being clicked. *Cerebral structure* and *Seizure* have been manually focused in Figure 2a, resulting in an IS-A climb.

Rendering and Layout: Different renderers (templates) can be used to convey information about different types or states of relations and nodes. Different line patterns, line widths and fill and colours can be configured.

By hovering over an edge or node, details about it are shown in the *Details* view at the bottom right of Figure 2b. A highlighting of the item and its neighbours occurs simultaneously (see Figure 2a).

Semantic zooming [6] adjusts the amount of information displayed, while geometric zooming only adjusts size. A simple form of semantic zooming is also illustrated in Figure 2a where only the focused nodes (purple) and their closest neighbours have text labels. The *Label* slider can be used to decide to what DOI-level labels should be rendered. This way we compress the structure between *Brain Part* and *Body Structure*.

The only layout algorithm currently used in the TermViz prototype is a force based physics simulation. Nodes exert “anti-gravity” repelling each other and links act as springs pulling connected nodes towards each other. A “drag force” (friction) is also active to stabilize the system. The simulation usually results in a fairly balanced self organizing graph structure where the forces balance each other. The force simulation can be stopped at any time using the pause button and nodes can be rearranged manually. In Figure 2a such manual rearrangements has been done to reduce size and improve readability. Individual nodes can also be pinned down at specific positions during force simulation. This is currently illustrated by using sharp instead of rounded node corners. Force simulation parameters can be adjusted during runtime. Increasing the anti-gravity may for example increase readability if the graph is too dense.

4. Discussion

TermViz is useful as a terminology search and browsing tool in its current state, but there is ample room for improvements.

Schneiderman's task list: Schneiderman [1] lists seven desired tasks that an IV system should perform (written in *italics* below). Many of these tasks are accomplished by TermViz.

Smooth *zooming* and panning is available by mouse operations. *Filtering* is available in both queries and the semantic zooming functions. *Details-on-demand* are shown by hovering over the node or edge of interest. Expansion of individual node relations can also be preformed.

The *overview* task is partly accomplished by the available parallel zoomed out view, but no general overview of the entire data collection is available. A limited structural overview can be created by executing a query fetching the root nodes of the terminologies and their descendants a number of steps down. The possibility to *relate* items is partly inherit in the node-link structure of the application area and it is also possible to issue queries to find relations. Schneidermans *history* and *extraction* tasks are not yet available in the prototype.

Problems with force based layouts: Layouts based on force simulations have inherit problems by being non-deterministic. Different runs of a layout algorithm should ideally not produce radically different results since that violates the desired objective preserving the mental map of the user. [6] Algorithms describing how to create more predictable force based layouts are available [10] but have not yet been implemented and tested in TermViz.

The *Graph Drawing* aesthetic rule of "minimizing the number of edge crossings" has been shown to be a prioritized rule for attaining readable graphs [12]. The force based layout algorithms are not optimized reduce crossings but perform fairly well in many situations due to the laws of physics.

Most of the TermColl terminologies are actually simple trees where nodes have a single parent. There are several more efficient layouts than force layout for such graphs. DOI-trees [13] would probably work well in this application. The possibility for, and consequences of, combining graphs with different layouts on the same screen (including mapping links between the graphs) would be interesting to investigate. Ideally there should be many layouts available for the user to choose from.

Semantic zooming and rendering improvements: By grouping sets of nodes and edges with low DOI and replacing the whole group with a special cluster node, the number of visible items can be reduced. This is a semantic zooming method, referred to as *clustering*, that can improve readability and performance. The structure between *Brain Part* and *Body Structure* in Figure 2a could for example be turned into one or a couple of cluster nodes. Clustering would improve TermViz and its implementation is planned.

Adding symbols, patterns, more colours and shapes to the rendering would make it possible to convey more detailed information in the main graph.

Future and Applications: Obvious future development possibilities of TermViz would be to implement the above discussed improvements. Other possibilities would be tailoring TermViz to enable visualization of UMLS⁸, FMA¹³ and Gene Ontology/OBO¹⁴ projects. It would also be possible to enable editing functions to support maintenance and authoring of terminologies and mappings between them. Methods from the prototype are currently being adapted and applied to other tools; Users of an archetype editor [14] will be supported in finding and selecting appropriate terminology bindings when creating archetypes¹⁵ intended for structured clinical data entry. The ontology alignment tool SAMBO¹⁶ has also started incorporating methods from the TermViz project.

Acknowledgements

This work was performed in the framework of the EU-funded Network of Excellence entitled Semantic Interoperability and Data Mining in Medicine. We would also like to thank J. Heer et al, the authors of *prefuse*.

References

- [1] Shneiderman B. The eyes have it: A task by data type taxonomy for information visualizations. Proceedings of the 1996 IEEE Symposium on Visual Languages; 1996 Sep 3-6; Boulder, Colorado, USA. Washington DC: IEEE Computer Society; 1996. p. 336-343
- [2] Robertson GG, Mackinlay JD, Card SK. Cone trees: Animated 3D visualizations of hierarchical information. CHI '91: Proceedings of the SIGCHI conference on Human factors in computing systems; 1991 Apr 27 - May 2; New Orleans, Louisiana, USA. New York: ACM Press; 1991. p. 189-194
- [3] Chittaro L. Information visualization and its Application to Medicine. *Artificial Intelligence in Medicine*. 2001;22(2):81-88.
- [4] Battista GD, Eades P, Tamassia R, Tollis IG. *Graph drawing: Algorithms for the visualization of graphs*. Prentice Hall; 1998
- [5] Kaufmann M, Wagner D, editors. *Drawing Graphs: Methods and Models (Lecture Notes in Computer Science, 2025)* Springer-Verlag; 2001
- [6] Herman I, Melançon G, Marshall MS. Graph Visualization and Navigation in Information Visualization: a Survey. *IEEE Transactions on Visualization and Computer Graphics*. 2000 Jan-Mar;6(1):24-43
- [7] Nyström M, Merkel M, Ahrenberg L, Petersson H, Åhlfeldt H. Creating a Medical English-Swedish Dictionary Using Interactive Word Alignment. Submitted to *BMC Medical Informatics and Decision Making*
- [8] Heer J, Card CK, Landay JA. *prefuse: a toolkit for interactive information visualization*. In Proc. CHI 2005, Human Factors in Computing Systems; 2005 Apr 2-7, Portland, Oregon, USA. Available from: <http://jheer.org/publications/2005-prefuse-CHI.pdf>
- [9] Fischer G. Beyond "Couch Potatoes": From Consumers to Designers and Active Contributors. *First Monday*. 2002 Dec; 7(12). Available from: http://www.firstmonday.org/issues/issue7_12/fischer/
- [10] Bertault F. A force-directed algorithm that preserves edge-crossing properties. *Information Processing Letters archive*, Volume 74 , Issue 1-2 , April 2000
- [12] Purchase HC. Which Aesthetic has the Greatest Effect on Human Understanding? *Lecture Notes In Computer Science*; Vol. 1353. In Proc. 5th International Symposium on Graph Drawing Pages: 248 – 261, 1997
- [13] Heer J, Card SK. DOITrees revisited: Scalable, Space-Constrained Visualization of Hierarchical Data. In Proc. *Advanced visual interfaces 2004*, Gallipoli, Italy. P. 421-424 Available from: <http://jheer.org/publications/2004DOITree-AVI.pdf>
- [14] Forss M, Hjalmarsson J. Development of an archetype editor: A tool for modelling structure in electronic health records. (Master's thesis, 2006) <http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-6205>

¹³ <http://sig.biostr.washington.edu/projects/fm/>

¹⁴ <http://www.geneontology.org/> and <http://obo.sourceforge.net/>

¹⁵ Archetypes as defined by openEHR, <http://www.openehr.org/>

¹⁶ <http://www.ida.liu.se/~iislab/projects/SAMBO/>