

STARzoom– An Interactive Visual Interface to a Semantic Database

Per Bruno, Viktor Ehrenberg and Lars Erik Holmquist

Viktoria Institute

Viktoriagatan 3

413 11 Göteborg, Sweden

cl4pbrun@cling.gu.se, cl4vehre@cling.gu.se,

leh@informatics.gu.se

ABSTRACT

STARzoom is a visualisation of a semantic hierarchical database utilising the hypernym structure from *WordNet*. It is also the search tool for that same database with which the user interacts in trying to visually chisel out a search query. The database contains the semantics from a large number of documents. The feedback from the system being the texts semantically closest to the query.

Keywords

Information retrieval, WordNet, hypernyms, semantic clustering, direct manipulation.

INTRODUCTION

Search engines usually come in two flavors: the direct query-reply system (e.g. Alta Vista) and the topically hierarchical system (e.g. Yahoo), *STARzoom* is a little bit of both. The user navigates through a hierarchy of concepts based on a semantic tree, rating from very general to very specific. The documents however are not (as in the topically hierarchical system) the leafs of the hierarchy, but are pointed to from all over the tree. By picking interesting concepts it is possible for the user to create a query that at once returns documents matching the chosen concepts, combined using simple conjunctive logic (AND).

We have created a semantically indexed database of about 700 articles from the Wired News service [4]. The texts have been chosen randomly and range from all the different categories (culture, politics, technical, general and business) that are used at Wired, and if you choose for example typically political concepts you will end up with quite a lot of documents from the political category. Still there is a certain blurring of the data in the information retrieval since articles about politics can also be about technology. Instead of trying to find specific words, the user has to deal with semantic concepts.

VISUAL INFORMATION RETRIEVAL

The usual procedure for querying a database is to first formulate a complex (often SQL) query and then wait for the system to process it. We are suggesting a different approach: By browsing the database (a key VIS principle[1]), progressively building and refining the queries, getting instant feedback to every addition, the user can not only not ask about things that are not there, but also see what the database actually contains.

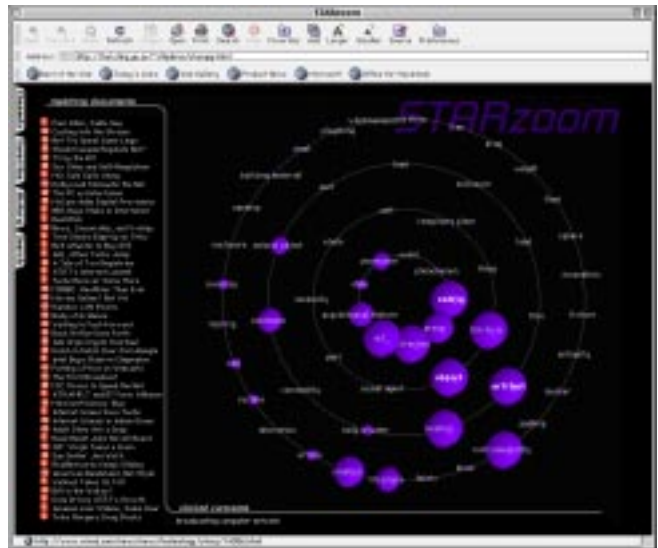


Fig 1: Screenshot with "Broadcasting" and "Computer Network" expanded.

A FIRST VIEW: CONCEPT BROWSING

The first thing that meets the eye is a circle, placed on that circle are a few blue dots or stars. These stars represent the first and highest abstraction level of concepts in the text collection. The size of each star indicates how much information the entire document collection contains in that semantic field. If one of the stars are clicked, a new circle orbiting the first will appear holding the stars representing the concepts of the next, lower level of abstraction.

Several different systems have been tried in visualizing complex hierarchical structures, e.g. the Hyperbolic Tree Browser[2]. There is however a natural tradeoff between over-

view and detail. Since detail is of great importance in this specific system (how many documents does that concept contain? what is that label saying? etc.) overview had to step back, we could not possibly show the whole structure at once. What is shown on the screen is a subjective view, everything not connected to the last clicked concept is pruned away.

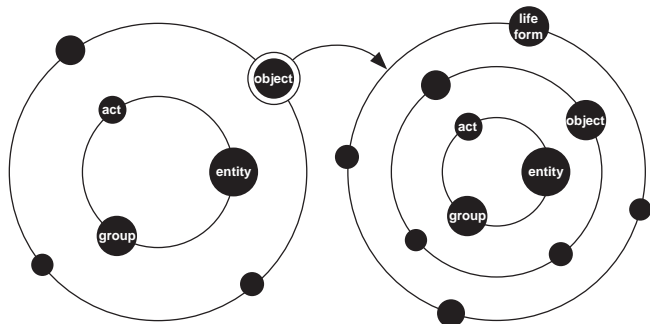


Fig 2. Simple view of how the node “object” is expanded to show its child nodes.

The System in Use

Imagine we are interested in the latest advances in television broadcasting, a subject most likely to occur in the WiredNews text collection. The first thing to do is to locate some “tv-concepts” by expanding the stars most likely to hold such. Broadcasting is found under “artifact”, “object”, and so on down to “medium” and by making a right click on the star labeled “broadcasting” it is turned red. On the left hand side of the screen all the documents talking about broadcasting in some way appear, ordered by how many words in the chosen concepts they contain. The articles seem to be about different aspects of broadcasting, we decide that we are interested in watching tv over the internet, so we locate “network” as well. The articles to the left now have to talk about both broadcasting and networks. By deselecting network, expanding it and choosing “computer network” instead we narrow down the scope a bit causing the list of articles to reorganise, disposing documents about other kinds of networks. We sharpen the focus on “broadcasting” as well by expanding it and picking “television” instead. This leaves us with just a few documents. The best matches in this case are obviously all about television and the Internet. The first article, for example, is about Paul Allen of Microsoft buying a cable television network to experiment with interactive television. Some of the other articles are named “TV By The Bit” and “Net TV’s Speak Same Lingo”. An interesting detail is that the matches are from several different WiredNews categories.

THE DATABASE

To be able to search a document collection by concept rather than by word one has to find a way to extract semantic information from each document, thereby creating a “profile” of it. Our system uses profiles constructed by extracting WordNet

hypernym chains[3] from the words in the documents. The idea of hypernyms is based on the way the human mind makes associations between different concepts, and the way it structures information based on those associations; we know that a dog is a mammal and that a mammal is an animal. The hypernym chains are used to form semantic tree structures leading from wide scope (root) to precision (leaves). In those tree structures nodes at every level of abstraction points to all the documents containing that node or concept, “dog” for example points to a subset of what “animal” points at.

CONCLUSIONS

The STARzoom visualization method presents a different solution to presenting hierarchical data structures. From any state of the system the user can see all the parent nodes of the current level of nodes and their siblings, all of which are instantly accessible at any time. The downside of this is that other branches of interest can not be presented at the same time, as in for example in the Hyperbolic Trees[2].

FUTURE WORK

STARzoom has, as we have pointed out, been created as an interface to a semantic database. We believe that STARzoom has a more general quality though. It could easily be used to visualize all kinds of hierarchical structures, such as file systems or the structure of web sites. The mechanisms used to interrelate semantic concepts can be used to handle hyperlinks or aliases. Depending on what is visualized, different changes could be made to the interface, such as hiding irrelevant choices or actually showing everything, all depending on the complexity of the data visualized.

ACKNOWLEDGMENTS

We would like to thank Jussi Karlgren of SICS and Ivan Bretan of Telia Research for their help in delivering the ideas behind this system.

REFERENCES

1. Ahlberg, C. and Shneiderman, B. Visual Information Seeking: Tight Coupling of Dynamic Query Filters with Starfield Displays. *Proceedings of the ACM CHI'94 Conference. 1994.*
2. Hyperbol Shneiderman, B. Visual Information Seeking: Tight Coupling of Dynamic Query Filters with Starfield Displays. *Proceedings of the ACM CHI'94 Conference. 1994.*
3. Wordnet Shneiderman, B. Visual Information Seeking: Tight Coupling of Dynamic Query Filters with Starfield Displays. *Proceedings of the ACM CHI'94 Conference. 1994.*
4. WiredNews Shneiderman, B. Visual Information Seeking: Tight Coupling of Dynamic Query Filters with Starfield Displays. *Proceedings of the ACM CHI'94 Conference. 1994.*