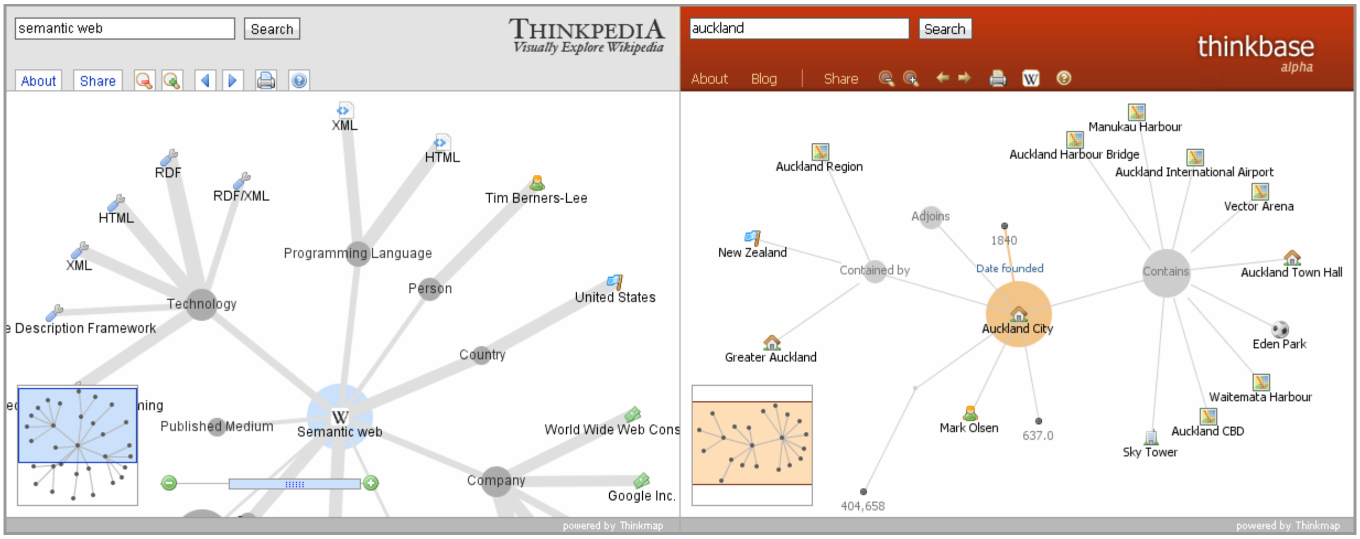
Interactive Visualization Tools for Exploring the Semantic Graph of Large Knowledge Spaces

Christian Hirsch, John Hosking, John Grundy
Department of Computer Science
The University of Auckland
Private Bag 92019, Auckland, New Zealand
{chir008@ec. | john@cs. | john-g@cs.}@auckland.ac.nz



ABSTRACT

While the amount of available information on the Web is increasing rapidly, the problem of managing it becomes more difficult. We present two applications, Thinkbase and Thinkpedia, which aim to make Web content more accessible and usable by utilizing visualizations of the semantic graph as a means to navigate and explore large knowledge repositories. Both of our applications implement a similar concept: They extract semantically enriched contents from a large knowledge spaces (Freebase and Wikipedia respectively), create an interactive graph-based representation out of it, and combine them into one interface together with the original text based content. We describe the design and implementation of our applications, and provide a discussion based on an informal evaluation.

Author Keywords

Semantic Web, Social Web, Wiki, Visualization, User Interface, HCI.

Workshop on Visual Interfaces to the Social
and the Semantic Web (VISSW2009), IUI2009,
Feb 8 2009, Sanibel Island, Florida, USA.

ACM Classification Keywords

H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCTION

This research focuses on the design and implementation of two interactive visualization tools. Both applications are built on top of large knowledge repositories. The first prototype, Thinkbase, is built on top of Freebase. The second prototype, Thinkpedia, is built on top of Wikipedia. The purpose of our applications is to provide a visual navigation and exploration tool for the underlying knowledge space. We aim to provide a proof of concept of how visualizations can improve and support Semantic Web applications.

The remainder of this paper is organized as follows. In the *Background* section we will shortly introduce the concepts of “Web 2.0”, “Semantic Web”, Information Visualization, and their relevance to our work. After clarifying our *Approach and Objective*, we will then describe the two prototypes in the main sections, followed by a *Discussion* and a short section on *Future Work*.

BACKGROUND

In today’s globalized Information Age the problem of *information overload* – having more information available than one can efficiently process – has become a ubiquitous issue. Recent estimates predict that in the next five years more information will be created than has been created in the whole of human history [4]. Most of the information is or will be accessible through the internet and intranet. Tackling the problem of information overload has thus become particularly interesting for the web community.

The “*Web 2.0*” [14] or “Social Web” already addresses the issue of information overload in several ways. The Web 2.0 is a loosely defined set of technologies, tools, and concepts, which has had an enormous impact on how web-based information is processed. Besides new enabling technologies (e.g. XML) and tools (e.g. Wikis), this “New Web” has introduced significant new behavioral and usage patterns like content sharing, personalization, and mass collaboration [16]. The resulting widespread adoption of Wikis and other social software tools transforms the way how information is created and annotated. By using social software, users annotate the content with meta-data in an organic, bottom-up fashion. This enables software agents to better process the information, and as a result more tasks can be delegated to those agents. Search algorithms and recommendation systems are successful examples of how a bottom-up creation of meta-data can help to better cope with an overwhelming amount of information.

The *Semantic Web* [2] represents a further, more recent, approach of addressing the information overload issue. Instead of creating semi-structured meta-data in a bottom-up fashion (as in the Web 2.0), the Semantic Web provides the possibility to formally define meta-data supported by knowledge representation languages (e.g. OWL) and formal specifications (e.g. RDF). The resulting structured content can then not only be understood by humans but also by machines. Therefore more and more tasks can be delegated to software agents. Research for example in the fields like semantic search (e.g. [7]) and semantic recommendation systems (e.g. [22]) is well underway. Further advantages of semantically enriched data, such as interoperability and transformability, allow for better integration of different sources as well as easy transformation between different representations (e.g. different languages).

A further and more general approach of how to cope with information overload can be found in the field of *Information Visualization*. Visualizations provide effective methods for representing and organizing knowledge- and information-rich scenarios [11]. They are tools for knowledge management which make use of the human cognitive processing system in order to create and convey content more efficiently. Information and knowledge visualizations both employ similar techniques. Based on specific mapping rules, they translate resource objects into visual objects, offering easy and comprehensive access to the underlying content [9].

APPROACH AND OBJECTIVE

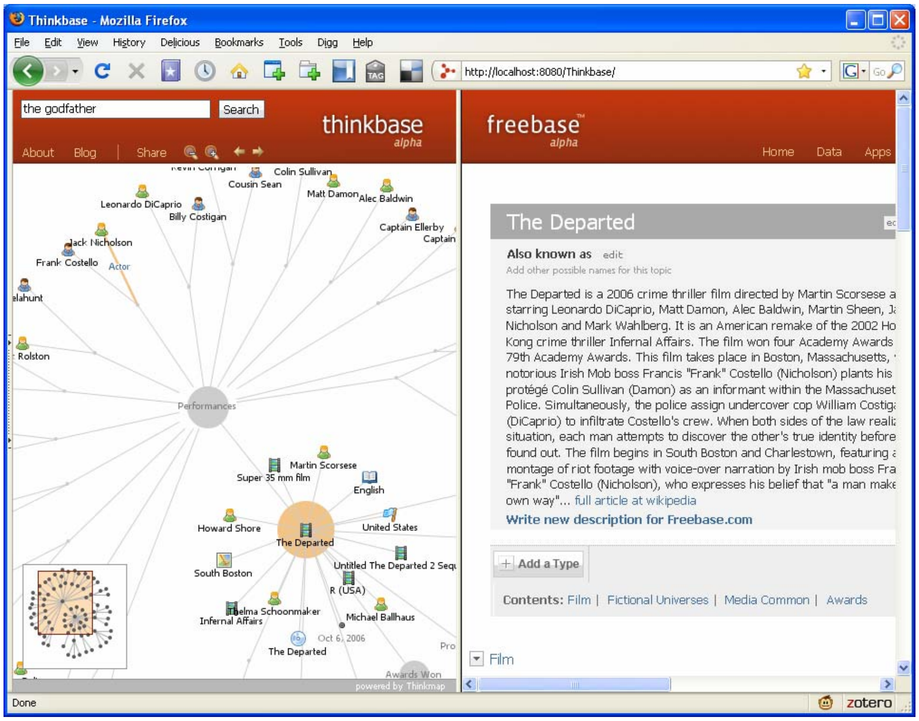
Even though all of the three mentioned approaches above – Web 2.0, Semantic Web, and Information Visualization – attempt to solve the information overload issue differently, there is plenty of space for *synergies*. Approaches of combining the Web 2.0 and the Semantic Web can be seen in two different directions: On the one side, organically grown content within the Web 2.0 (e.g. Wikipedia) is being semantically enriched with the help of natural language processing and knowledge extraction. DBpedia [1] is a well known example of this. On the other side, semantic information repositories start to allow end-users to edit and create semantics in a collaborative wiki-style manner. An example of this is Freebase [5]. One of the better known applications which demonstrate synergies between the Web 2.0 and Information Visualization is Many Eyes [21]. It allows everyone to create, share, and discuss visualizations online. Lastly, there exists a variety of possibilities and approaches to visualize Semantic Web content. This is discussed for example in [6].

Our *approach* when building two prototypes of interactive visualization and exploration tools for large knowledge spaces was to combine elements from all three areas. As knowledge spaces we have chosen Freebase (a “Semantic Wiki”) for the one application, and Wikipedia (semantically enriched) for the other application. At the core of our tools, we utilize one crucial benefit of Semantic Web data: the ability to be easily transformed from one representation into another. More precisely we transform the content from a textual representation into a visual representation. The interactive visualizations are displayed alongside the text-based repositories, providing a focus-plus-context view. The results are applications which present visually enriched user interfaces for Semantic Web content.

The *objective* of our research is to provide a proof of concept of how interactive visualizations can improve Semantic Web applications. This is two-fold. On the one hand, our objective is to demonstrate how it is possible to easily transform Semantic Web content into meaningful visual representations. On the other hand, our objective is to demonstrate how these resulting applications can be used as efficient information discovery tools.

THINKBASE

Our first prototype, *Thinkbase* [18], is a visual navigation and exploration tool for Freebase, an open, shared database of the world’s knowledge [3]. Freebase can also be described as a “Semantic Wiki”. This means its content is semantically enriched, everyone can edit it, and furthermore, the meta-model itself is also editable by everyone. Figure 1 shows the general user interface of Thinkbase (in this case displaying the movie “The Departed”). The application is divided into two frames. The right frame displays the current Freebase topic, which consists of a short textual description as well as all the details in tabular form. The left frame displays an automatically generated, interactive, force directed layout graph of that same topic including all related topics. We have chosen to use the Thinkmap visualization framework to implement this [19]. Thinkmap is a software platform for developing customized visualization interfaces. It consists of loosely coupled components which provide users the ability to retrieve a result set from data sources, and then visualize, navigate, and organize it. The Thinkmap Software Development Kit (SDK) provides ways to easily extend and adjust the suite as well as to integrate it with other web and database technologies.



The Departed

Also known as edit

Add other possible names for this topic

The Departed is a 2006 crime thriller film directed by Martin Scorsese a
starring Leonardo DiCaprio, Matt Damon, Alec Baldwin, Martin Sheen, Ja
Nicholson and Mark Wahlberg. It is an American remake of the 2002 Ho
Kong crime thriller Infernal Affairs. The film won four Academy Awards
79th Academy Awards. This film takes place in Boston, Massachusetts, '
notorious Irish Mob boss Francis "Frank" Costello (Nicholson) plants his
protégé Colin Sullivan (Damon) as an informant within the Massachuset
Police. Simultaneously, the police assign undercover cop William Costig
(DiCaprio) to infiltrate Costello's crew. When both sides of the law reali
situation, each man attempts to discover the other's true identity before
found out. The film begins in South Boston and Charlestown, featuring a
montage of riot footage with voice-over narration by Irish mob boss Fra
"Frank" Costello (Nicholson), who expresses his belief that "a man make
own way"... full article at wikipedia

Write new description for Freebase.com

+ Add a Type

Contents: Film | Fictional Universes | Media Common | Awards

Film

Figure 1. The user interface of Thinkbase.

Thinkbase accesses the Freebase API, retrieves information about the current Freebase topic as well as all related topics, and creates a graph-based visual representation of it with the help of Thinkmap. Each Freebase topic is represented as a node using an icon which corresponds to its type (e.g. person, movie). Edges between those nodes are annotated with the type of the relationship. These labels become visible when hovered by the mouse. For example, Figure 2 shows the Thinkbase graph for “Homer”. There, one can see that the “Place of birth” of “Homer” is “Greece”. Related topics of the same type are combined in an aggregation node (the grey circles) as seen for example for the type “Influenced”. These aggregation nodes can be expanded and collapsed through a context menu, which helps to focus on specific contents while hiding others (e.g. “Quotations”). Further visual cues such as the length of edges, size of aggregation nodes, and text color are used to encode additional information. Users can navigate from node to node by clicking on them. This will refresh the graph as well as the Freebase frame. The graph is animated which will allow for a smooth transition between different visualizations. This helps the users to preserve the “mental map” [12] of the knowledge space. The two alternative representations (textual and visual) of the same underlying body of knowledge enable a focus-plus-context view. This is a further means to support navigation and help users to maintain the mental map. While the textual representation gives a good focus on the current topic, the visual representation allows the user to see the topic embedded in the wider context.

The visual representation in Thinkbase presents a topic-centered view. That is, a specific Freebase topic is at the center of the visualization and connections of all directly related topics are shown around it. As all of the Freebase content is basically one huge graph, this means that at any one time we show a small subset of this whole graph. From

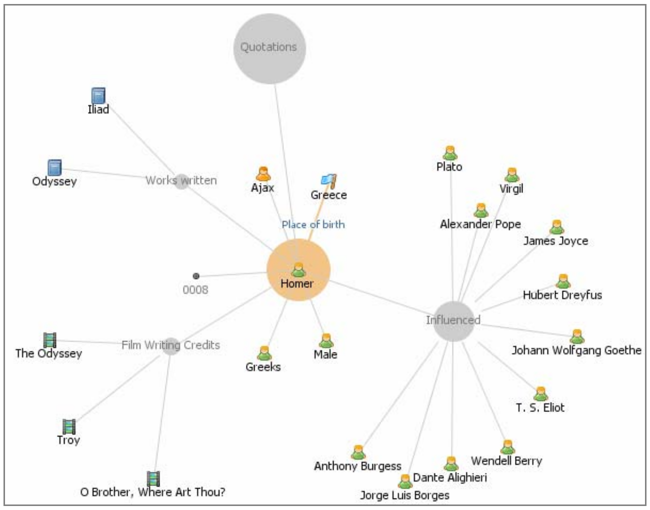


Figure 2. The Thinkbase graph for “Homer”.

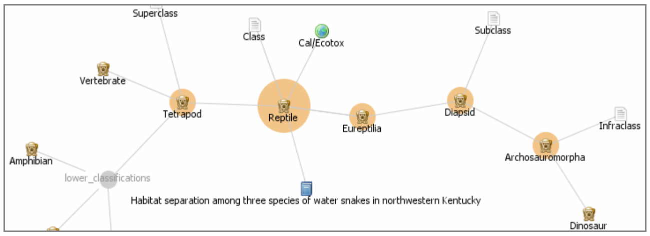


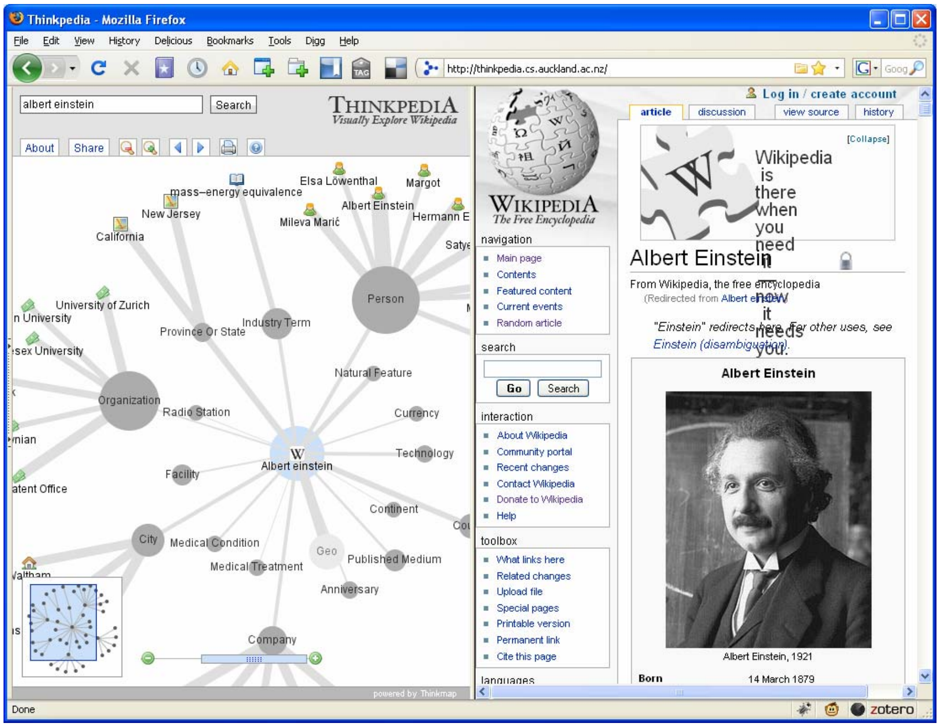
Figure 3. A small extract of the “tree of life”.

a cognitive perspective, it would not make sense to display a very large amount of data [8]. However, we allow users to extend the visualization metaphor by (repetitively) expanding and collapsing not only aggregation nodes but all nodes of the graph to ones liking. This feature gives the user the ability to create unique and informative visualizations. Figure 3 shows an example where the lower and higher classifications of an animal class (here: "Reptile") has been expanded repetitively. The resulting visualization represents a small subset of the tree of life, ranging from "Vertebrate" to "Dinosaur".

Further features of Thinkbase include: zoom functions; the ability to navigate the browsing history; printing the visualization; the possibility to share a direct link to a specific page; and the option to trigger a search of a node in Google or Wikipedia. Our research prototype also provides some functionality to edit the content of Freebase through the visual representation (e.g. add new relationships). This is only possible due to the semantically enriched content.

THINKPEDIA

Our second prototype, *Thinkpedia* [20], is a visual navigation and exploration tool for Wikipedia. The objective for this prototype was to investigate the possibility of creating a similar visual exploration tool as Thinkbase, only for a less structured knowledge space. The "Social Web", of which Wikipedia is a part, has produced a huge amount of interesting content. However, most of it is unstructured or semi-structured. Therefore it is hard for machines to reason with it and, in our case, to automatically translate the content into meaningful visualizations. What is needed is a way to extract semantics from the unstructured contents of Wikipedia. This field of knowledge extraction is a well established research area, and tools like DBpedia [1] demonstrate successful approaches of doing this. We decided to use the SemanticProxy web service which is part of the Calais initiative by Thomson Reuters [15]. The SemanticProxy takes plain text or a URL as input, processes this, and returns the identified concepts and their relationships in RDF format. Using a general "semantifier" like SemanticProxy allows us to easily switch between different MediaWikis (not only Wikipedia) or even other unstructured sources. Figure 4 shows the general user interface of Thinkpedia (in this case displaying the article for "Albert Einstein"). The concept of the application is similar to Thinkbase. It is divided into two frames, the right



Wikipedia is there when you need it — now it needs you.

Albert Einstein

From Wikipedia, the free encyclopedia

(Redirected from Albert einstein)

"Einstein" redirects here. For other uses, see Einstein (disambiguation).

Albert Einstein

Albert Einstein, 1921

Born 14 March 1879

Figure 4. The user interface of Thinkpedia.

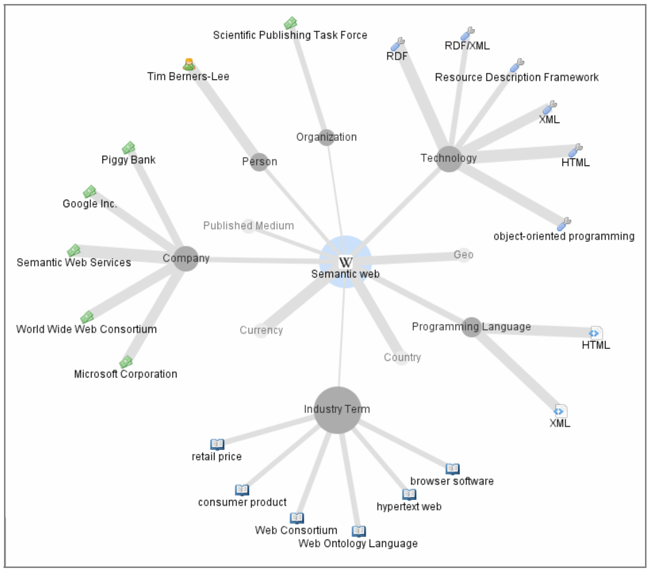


Figure 5. The Thinkpedia graph for "Semantic Web".

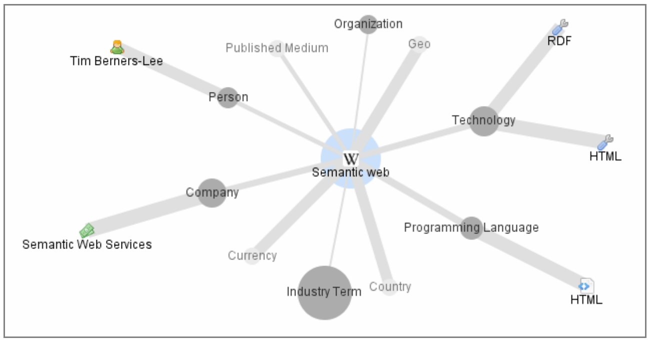


Figure 6. The same graph as Figure 5, reduced to the most relevant concepts.

one displays the current Wikipedia article, the left one displays an interactive, force directed layout graph which was created from the same Wikipedia article with the help of SemanticProxy. We use Thinkmap for this, the same visualization framework as used for Thinkbase.

When a Wikipedia article is requester through Thinkpedia (e.g. through a keyword search), the application first accesses the Wikipedia API in order to retrieve the most relevant article(s). The SemanticProxy API is then used to processes the content (i.e. identifies concepts and their relationships), and returns the result in RDF format. The RDF content is parsed and visualized as an interactive graph. For example, Figure 5 shows the graph for the "Semantic Web" Wikipedia article. The article itself is visualizes as the center node. All identified related concepts are shown around the center. These concepts are things like "Person", "Company", or "Country". Each of these is represented as a node in the graph using an icon which corresponds to its type. Concepts of the same type are combined in an aggregation node which can also be collapsed in order to reduce the amount of information shown. The size of the aggregation nodes corresponds to the number of concepts within this type. One particularly interesting feature of the SemanticProxy is that it annotates each identified concept with a relevance value. This value expresses how relevant the concept is within the processed text. We visually encode this value in form of the edge thickness. The thicker an edge is, the more relevant is its connected concept. For example, in Figure 5 one can see that the "Person" "Tim Berners-Lee" is more relevant to the "Semantic Web" Wikipedia article than any of the "Industry Terms". The edge thickness between the center node and the aggregation node is an average of all the edges going out from the aggregation node. Furthermore, we can use this value for an interactive range slider (see Figure 4 in the lower part of the visualization). This range slider can be increased and decreased which will show more or less relevant nodes in the graph. Figure 6 for example shows the same graph as Figure 5, only that its visible content has been reduced to the most relevant concepts.

Navigating the graph is quite similar to Thinkbase. Clicking on a related node will refresh both the Wikipedia frame as well as the graph. A difference here is that the concepts in the graph do not correspond directly to a Wikipedia page. Clicking on a concept therefore triggers a further Wikipedia search, which returns the most relevant page. Further features of Thinkpedia again include a zoom and printing function, the ability to navigate the browsing history, and the possibility to share a direct link.

DISCUSSION

Clearly, both of our prototypes have their strengths and weaknesses. Some of these are related to the contents and structure of the underlying knowledge repositories (Freebase and Wikipedia). Others are related to how we implement our visual exploration tools on top of those repositories. We conducted a small *informal survey* in order to better discuss potential strengths and weaknesses. Our prototypes depend to a large extend on the efficiency and usability of visualizations. These, however, are fundamentally hard to evaluate. Therefore we chose a quite informal and anecdotal evaluation method proposed in [13]. Instead of giving users a clearly defined task (e.g. finding a specific piece of information) and then measuring the time or accuracy when using different visualization tools, we gave the users one open-ended task and let them report on interesting findings or insights. Seven users participated in the survey, all of them postgraduates or staff members at the Computer Science department of The University of Auckland. Participants were asked to choose any starting topic they are interested in (e.g. their favorite movie, next holiday destination, a famous person) and then explore this topic and its related topics according to their interest. Additionally they were asked to write a short report about how they experienced both of the applications, what kind of insights they gained, as well as notable differences between the tools.

A *general observation* which was made in similar ways by a majority of the participants was that even though Thinkbase provides more structured content, the coverage of its content is rather limited (that is, for many topics the semantic content is still very sparse). Thinkpedia on the other hand has much more coverage but the semantically enriched graph still lacks some structure. Not surprisingly, this also roughly translates into general strengths and weaknesses of the Semantic and Social Web.

More precisely, participants reported that *Thinkbase* is "very well structured", the "[connections] seem very solid", and "navigation felt very natural". Furthermore, the application has been described as "effective and beautiful", and that it is "lot of fun [browsing the content]". On the downside, participants reported that the "richness of content [is] rather less than [the one in] Thinkpedia", e.g. it is "limited for some topics" and not as "full as [one] would have liked". For *Thinkpedia*, participants reported that the "richness of information is much better" and "more comprehensive". "Due to the fuller amount of information available", the application "[gives] an interesting perspective". However, there clearly are weaknesses. Participants found Thinkpedia to be "less solid" and that it sometimes "seems a little bit disorder". Furthermore the visualization presents some "odd mistakes", due to ambiguities within the process of extracting semantics. The implementation of the search function in Thinkpedia is still a little bit flawed and was described as "frustrating".

Insights about which the participants reported were mostly along the lines of discovering related information which they were either not aware of, or which they already knew of, but found noteworthy to see visualized. A typical report would for example look like this: "I found it interesting to see X connected with Y". Exploring content along those kind of connections seems to be a very useful feature. One participant for example described how he navigated from a television show to a city to a state and finally he discovered a "mountain [where he could] go skiing". This relates to a concept called Orienteering [17], which describes a type of search in which the target is not (well) known. Instead of jumping or "teleporting" directly to the target (what is usually the case in keyword search), one rather performs a directed situated navigation. This means a user takes a series of smaller steps while navigating though the information space. Advantages of Orienteering are: it decreases the cognitive load, maintains a sense of location, and gives a better feeling for context. Our applications seem to support such a navigation behavior as it allows starting e.g. with a general topic and then drilling down on it. Lastly, participants reported on the benefits of having information condensed in a visual form. This help to reveal "information that is otherwise difficult to notice when presented in a textual environment". Furthermore one can "easily [see] key words and [does not] need to waste time reading [all the text]".

Graphs are arguable not always the best way to represent large amounts of content, depending on the task a system is meant to support [10]. Instead of simply displaying one "big fat graph", we have focused on several ways to filter the graph (e.g. aggregation nodes and range slider). Furthermore our graph visualizations do not replace but go along with existing user interface approaches (e.g. tabular displays in Freebase). The informal evaluation as discussed above suggests that our approach has several benefits when exploring large knowledge spaces.

FUTURE WORK

Future work of our research will focus on two different areas. Firstly we will further work on improving the existing two prototypes and adding new features. This will include fixing weaknesses identified in the evaluation such as poor search function in Thinkpedia and smaller user interface improvements. Further work might also focus on improving the usability of Thinkbase by adding more advanced filtering mechanisms and giving more control over the display. Improving Thinkpedia might include exploring alternative knowledge extraction tools. Secondly our future work will include extending our concept to further information repositories. We have provided a proof of concept of how visual user interfaces can improve Social and Semantic Web applications. This same concept could be explored for many more applications and domains.

REFERENCES

1. Auer, S., et al., *DBpedia: A Nucleus for a Web of Open Data.* Lecture Notes in Computer Science, 2007. 4825.
2. Berners-Lee, T., J. Hendler, and O. Lassila, *The semantic Web.* Scientific American, 2001. 284(5).
3. Bollacker, K., R. Cook, and P. Tufts, *Freebase: A Shared Database of Structured General Human Knowledge.* Proceedings of the national conference on Artificial Intelligence, 2007. 22(2): p. 1962.
4. Department of Education, Science and Training. *Backing Australia's Ability - An Ongoing Commitment.* 2007. http://backingaus.innovation.gov.au/info_booklet/on_commit.htm.
5. F Freebase. 2008. http://www.freebase.com.
6. Geroimenko, V. and C. Chen, *Visualizing the Semantic Web: Xml-based Internet And Information Visualization*. 2006: Springer.
7. Guha, R., R. McCool, and E. Miller, *Semantic search*, in *Proceedings of the 12th international conference on World Wide Web*. 2003, ACM New York, NY, USA. p. 700-709.
8. Herman, I., G. Melancon, and M.S. Marshall, *Graph Visualization and Navigation in Information Visualization: A Survey.* IEEE Transactions on Visualization and Computer Graphics, 2000: p. 24-43.
9. Jaeschke, G., M. Leissler, and M. Hemmje, *Modeling Interactive, 3-Dimensional Information Visualizations Supporting Information Seeking Behaviors.* in Knowledge and Information Visualization: Searching for Synergies. Springer 2005: p. 119-135.
10. Karger, D. and M.C. Schraefel. *The pathetic fallacy of RDF*. 2006.
11. Keller, T. and S.O. Tergan, *Visualizing Knowledge and Information: An Introduction.* in Knowledge and Information Visualization: Searching for Synergies. Springer 2005: p. 1-23.
12. Misue, K., et al., *Layout Adjustment and the Mental Map.* Journal of Visual Languages and Computing, 1995. 6(2): p. 183-210.
13. North, C., *Toward Measuring Visualization Insight.* IEEE Computer Graphics and Applications, 2006.
14. O'Reilly, T., *What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software.* O'Reilly Media 2005.
15. Reuters, T. *SemanticProxy*. 2008. http://semanticproxy.com.
16. Tapscott, D. and A.D. Williams, *Wikinomics: how mass collaboration changes everything*. 2006 Portfolio.
17. Teevan, J., et al., *The perfect search engine is not enough: a study of orienteering behavior in directed search,* in *Proceedings of the SIGCHI conference on Human factors in computing systems*. 2004, ACM New York, NY, USA. p. 415-422.
18. Thinkbase. 2008. http://thinkbase.cs.auckland.ac.nz.
19. Thinkmap. 2008. www.thinkmap.com.
20. Thinkpedia. 2008. http://thinkpedia.cs.auckland.ac.nz.
21. Viégas, F.B., et al., *Many Eyes: A Site for Visualization at Internet Scale.* IEEE Transactions on Visualization and Computer Graphics, 2007: p. 1121-1128.
22. Ziegler, C.N., L. Schmidt-Thieme, and G. Lausen, *Exploiting semantic product descriptions for recommender systems,* in *Proceedings of the 2nd ACM SIGIR Semantic Web and Information Retrieval Workshop*. 2004.