Augmenting search with a Semantic Visual Graph

Edwin de Jonge (ejne@cbs.nl) , Olav ten Bosch (obos@cbs.nl)
Statistics Netherlands

Abstract: *Statistics Netherlands produces statistics on a wide spectrum of topics about Dutch society. All statistical results are published for free on the internet via the StatLine database, currently containing over 100 million statistical facts. Searching and finding information in such a large database is not trivial. This searching problem is aggravated because statistical institutes traditionally treat the statistical areas of interest differently. For example economic and social statistics are rarely made by the same statisticians. Many statistical areas however overlap and relate to one other. Each statistical overlapping area though has its own jargon, homonyms and synonyms, which makes searching difficult. However the end-user can't be expected to know all these domain specific knowledge and should be provided with a powerful search engine that helps him find relevant statistics easily.*

The search problem can be alleviated using semantic web technology (XTM) and SVG visualization. This paper briefly describes a new approach to this concept of semantic search focused on SVG visualization and implementation. This search mechanism combines an advanced search engine for statistics with a Semantic Visual Graph (SVG). Scalable Vector Graphics is the natural choice for its implementation because of the possibility to use animation, textPath and callback functionality. The paper gives an example where visualization is an integral part of the semantic searching process. The visualization enables a user not only to specify his search via keywords but also to search associatively, which makes users user find information more intuitively.

Keywords: SVG, Semantic search, XTM

1. Introduction

Statistics Netherlands currently offers a full text search service on the online statistical database StatLine [5,7,8,9]. This database contains over 100 million statistical facts. Although StatLine contains mainly data and no text documents this search approach works reasonably well for most users. However we think that this search mechanism can be improved. Due to the number and complexity of semantic relationships present in the statistical database, we believe that the best way to improve the capabilities of the search engine is to apply semantic techniques.

As the semantic web evolves more and more techniques come around that can be used to model statistical metadata and that can be used by an intelligent search engines to present

intuitive results. We implemented a prototype semantic search engine that uses SVG for visualization. Before describing the SVG visualisation we implemented, we first give a brief description of the concept behind the semantic search technique.

2. The concept of a Semantic Search Technique

First of all, it is useful to note that national statistical institutes have certain particular areas of interest: such as social statistics, business statistics and environmental statistics. These statistical themes or domains contain topics that in itself may be closely related to topics from other themes. For example, statistics on households or dwellings are much related to statistics about motherhood, fatherhood, number of children, residence, inhabitants. However, using an ordinary text based search engine, users searching these domains may not find statistics that are expressed in a different terminology but are semantically close. This brings up the idea to extend a text based search engine with semantic knowledge that is able to catch these relationships and use it to present semantically related statistics to the end-user.

A well-known text search problem is the homonym and synonym problem. Homonyms can give many unexpected results. For example the search result for the Dutch word "banen" contains data on labor statistics ("banen" = jobs) but also from recreational statistics ("banen" = swimming lanes). Synonyms can give too little results. Many statistical domains use synonyms and homonyms. We try to alleviate these problems by augmenting a (full text) search with semantic information. Every search result contains a semantic visual graph that can be used to narrow or divert the search.

Our approach resembles the approach of KartOO [1] and the aquabrowser[2], but also differs in an important aspect. KartOO and Aquabrowser use statistical text analyzing techniques to extract semantic distances between concepts. This distance metric only expresses that concepts are semantically close to each other or not. In our approach, we use an explicit semantic model, where the types of links between concepts can be different. Inhabitants (topic) live in (association) municipalities (topic). So besides a semantic distance we also have types of associations available to express relationships in statistical metadata. For a more detailed explanation of the concepts being used, see the work carried out in cooperation with Delft University of technology [3].

2.2 Using Xml Topic Maps

The semantic information described in the previous section is stored in a xml topic map (XTM) [6]. Xml topic maps is a semantic web technology comparable to RDF/OWL [5]. Basically a topic map is made out of Topics (the subjects), Associations (the typed links between topics) and Occurrences (references to instances outside a topic map). XTM differs from RDF is some aspects. One important pro for XTM is that it supports scope or context. In XTM scopes can be used to define a context in which a property of a topic or association is valid. They can be used to differentiate between homonyms. During topic map navigation scope can be used to restrict the visibility of the number of topics and associations to current scope.

Input for our semantic search is a very large Topic map made of the topics available in the statistical database. This topic map is mainly generated from the available metadata in our

Statistical database and can be manually augmented with other topics and associations. StatLine contains associations like part-of, type-of and subtype-of. It also contains topic types as statistical population, statistical classification. The number of topics generated from the database is in the order of 10,000. Because of performance reasons, the prototype implementation uses only small parts of the topic map while searching. Nevertheless, this topic map being used is large enough to experiment with the results of the search engine for certain specific statistical domains.

2.3 Semantic Visual Graph

The semantic visual graph is a SVG based visualization of the semantically related concepts of the search keywords fed into the search engine. It consists of topic nodes and association edges. The semantic visual graph is a restricted graph visualization of the topic map: it shows a small part of the complete topic map. The center of the graph shows the current topics of interest, surrounded by their related concepts. Topics that are beyond a predefined horizon are not shown. Typically for XTM is that the associations also have names and can be typed. Therefore the visual graph must show the names of the association. We use the <textPath> functionality of SVG.

Beyond the obvious reasons we choose SVG as visualization because Statistics Netherlands has some experience using SVG. StatLine already contains a cartographic SVG module and we have some SVG visualizations of demographics and economical monitors.

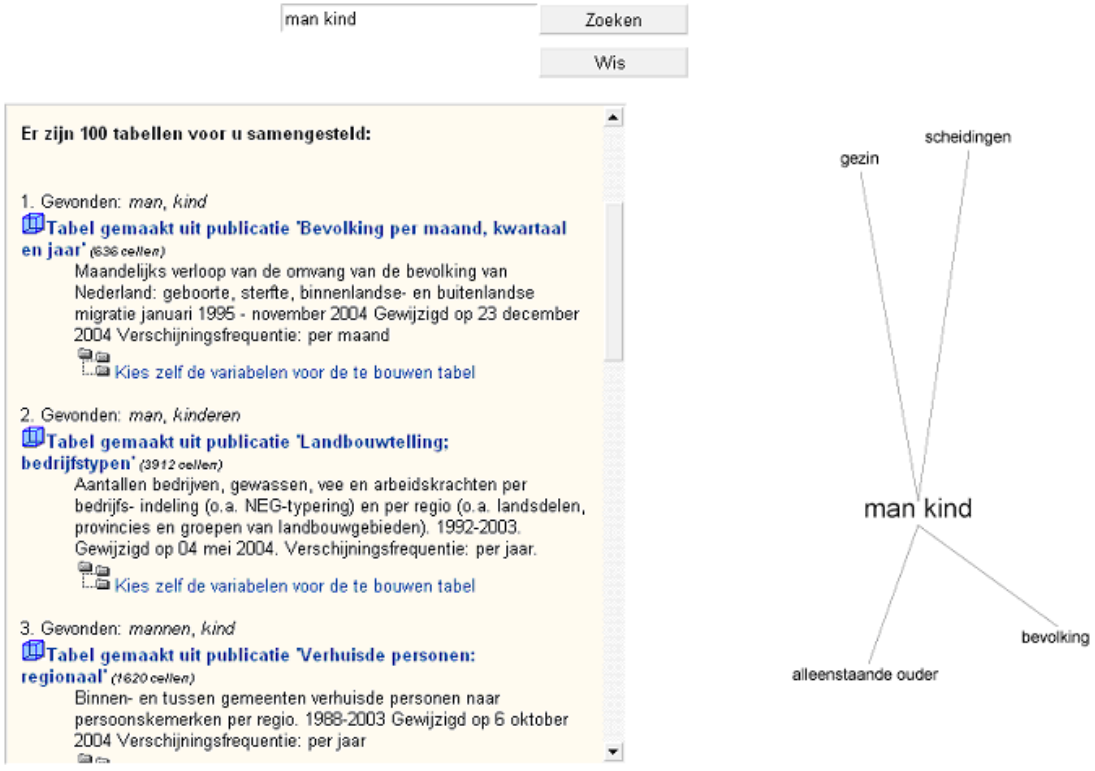

Figure 1: Prototype of semantic search (Dutch)

The overall topic map for StatLine is large having over 10,000 topics. When a user clicks on a topic, this topic becomes the central topic. This means that the SVG graph updates itself with new topic information for the new central topic. This is accomplished with JavaScript

callback. The JavaScript calls for a new Semantic Visual Graph for the new topic in focus and updates the current SVG with this information. The graph then animates from the previous central topic to the new central topic. The semantic visual graph is interactive. It responds to mouse over events. A link will show its name (type) and a topic will show some explanation. Figure 1 shows a screenshot of the prototype implementation.

3. Technical Preliminary Results

A student from TU Delft implemented a prototype of the Semantic Visual Graph using ASP.NET C# and javascript. The current full text search of StatLine is used as a WebService to do the text searching. From the extensive metadata of the StatLine database a topic map was generated using a .NET topic map engine. The prototype is an ASPX application that generates an HTML page with an embedded SVG semantic visual graph. When the user submits one or more search keywords the HTML page is updated with the search results from the WebService. The embedded SVG is updated with the new search words. The SVG animates from the current state to the new state. A user can click on topics and associations. Clicking a topic will change and execute the search query: meaning that a user browses the semantic web of StatLine and is shown during browsing what figures are available for the current topic of interest.

Topics are visualized as a node with text. Associations are shown as an edge. To reduce visual "clutter" names of associations are not shown until the user hovers over the edge. The layout algorithm of the graph is a spring box model where every association acts as a spring between the connected topics. The central topics are connected to the center of the graph with a spring of length 0. The current implementation calculates the spring configuration on the server, but we nearly finished an implementation that is purely client side (javascript based).

We have plan to make the prototype even more interactive by using AJAX: asynchronous javascript calls to the webserver: the semantic visual graph can update itself while the user is typing search keywords.

4. Expected improvements

We expect the following improvements of a full text search augmented with a semantic visual graph.

4.1 Semantic focus

The visual search allows for semantic focus and finding concepts that wouldn't be found using normal search. Semantic focus tries to focus the search of the user on topics known to StatLine. It helps the user to formulate his query. For example if a user tries to search for "man child" (see figure) the topic "family" and "single parent" will be shown in the semantic visual graph. Using normal full text search these suggestions would not be available.

4.2 Associative search

The visual search engine allows for associative searching. Associative searching allows the user to find topics that are related to the search words, that may be of interest. During search semantically nearby topics are shown. Adding topics to the search query narrows the search results. Replacing (parts of) the search query diverts or generalizes it. For example if a user searches for "bicycle industry" the topics "automobile industry" and "vehicle industry" will be shown. Although not intended they may be of interest to the user.

4.3 User tests

These examples show that the visual semantic search engine is an improvement compared to traditional search methods. However we would like to quantify this more exactly. To measure the improvement the following tests can be conducted [see 3] with two groups of testers. One group uses current search, the other the augmented search: This kind of test is still to be carried out.

- Ask both groups to find a specific StatLine publication. Can both groups find the publication in the current StatLine system and in the new StatLine system with SVG?
- Speed, when searching for a publication we measure how long it takes to find the results.
- Usefulness, does the Semantic Visual Graph really help beginners to find the requested information? And can the experienced users also find other interesting publications / information that would be missed in the current system and now is visible with the use of a Topic Map.
- Which one is easier to use?
- Scope, does scope create an area where the publication can be found. And does the user find it helpful that the Topic Map can have different views, beginners and experienced users.

5. Conclusions

We feel that augmenting a full text search with a semantic visual graph is certainly visually attractive and seems promising. SVG as language for visualization seems a useful and natural way to express the semantic nearby relationships as well as to implement the navigation techniques to be used. Further research, based on practical usability tests with different statistical user groups, has to show the exact direction to go, but preliminary experiments have shown that a visual semantic extension to a statistical search engine makes it easier to find certain semantically related statistical information.

Acknowledgements

Thanks to Anand Kapoerchan for his work on the prototype search engine.

Bibliography

[1] *KartOO.com, meta search engine* http://www.kartOO.com
[2] *Dutch Library site* http://zoeken.bibliotheek.nl/?q=vector%20graphics

[3] *Topic Maps for the StatLine database* MSc article TU Delft, Anand Karpoerchan, internal report
[4] *Resource Description Framework (RDF)* http://www.w3.org/RDF
[5] *StatLine* http://statline.cbs.nl
[6] *XTM specification homepage* www.topicmaps.org
[7] *New ways of disclosing StatLine, a large statistical database* E.de Jonge. International Marketing and Statistical Output Database Conference, Annapolis Sep 2002
[8] *Handling Time Dependence of MetaData and Data in StatLine4* J.A. Reedijk, E. de Jonge, O. ten Bosch, International Marketing and Statistical Output Database Conference, Oslo Sep. 2004
[9] *New Developments in Information Processing at Statistics Netherlands* O. ten Bosch, Meeting on the Management of Statistical Information Technology (MSIS), Geneva February 2003;
http://www.unece.org/stats/documents/ces/ac.71/2003/7.e.pdf