Information visualization for product development in the LIVA project

av Gertrud Berger *, Sándor Darányi **, Johan Eklund , Maivor Hallén *** och Lars Höglund****

**BTJ Sverige AB, Lund*
***Swedish School of Library and Information Science, Högskolan i Borås*
****Bibliotekscentrum Sverige AB*

The LIVA research and development project (2005-2007) was conceived to integrate automatic indexing, automatic categorization, information visualization and information retrieval in library systems managing textual document collections. After a brief overview of some major information visualization methods, the user interface prototype is introduced.

Introduction

LIVA (Library Information Visualization and Analysis) was a research and development project of the Swedish School of Library and Information Science in Borås (SSLIS), Bibliotekscentrum Sverige AB, and BTJ Sverige AB, with funding for 2005-2007 from the Knowledge Foundation (KKS).[1]

Based on the analysis of bibliographic and other metadata, mainly from BTJ's databases, but also from other content providers, the goal of the project was to bring competitive functionality in terms of language technology, classification research, information retrieval (IR) and information visualization (IV) to library systems. In the definition of library systems, we include integrated library systems with Web 2.0 inspired user interfaces, OPACs, union catalogues etc. The novelty the project has brought was the combination and integration of the above functionalities into a working prototype. This includes several interesting possibilities for improved functionality. However in this article focus is on vizualisation, which was one of the goals for the project.

The feedback from customers and users makes it easier to develop user friendly and flexible products and services. Therefore we have linked a reference group of libraries to the project. This group consists of :

- Lund Public Library
- Nordiska Museet
- SCB (Statistics Sweden)
- SÄS (Southern Älvsborg Hospital Library)
- TPB (The Swedish Library of Talking Books and Braille)

In this article, we will focus on the role of information visualization in libraries and information institutions. Our examples will come from text processing only and will be limited to document clustering by different methods as opposed to document classification. We would like to depend on one definition only and distinguish between document classification and document clustering [Sebastiani 2005]. *Text categorization*, also known as *text classification*, is the task of automatically sorting a set of documents into predefined and labelled categories (or classes). *Clustering*, on the other hand, is the categorization of documents into a set of groups, which arise from inter-similarities between the documents. We also distinguish between *partitional clustering*, where no relations between the obtained clusters are stored, and *hierarchical clustering*, where relations between the clusters are stored in a hierarchical (tree) structure. That is, clustering exposes structure inherent in the data.

A brief overview of information visualization

Visual access to classification or IR results is as important and popular as using icons for communication in public spaces: it is a language independent method, therefore it provides the mind with direct access to data, plus using "visual shorthand" for explaining complex relationships implies a high abstracting power. As dealing with textual data, such as a document database, often requires text pre-processing (e.g. stemming, lemmatization, part-of-speech tagging, spelling correction etc.), we can conceive the LIVA product prototype as one that consists of three major parts:

- A large-capacity text processor,
- An analytical component for text categorization and IR, plus
- A front-end visualization component in the form of a graphical user interface.

User interfaces are of primary importance in human-computer interaction, being developed by means of computer graphics and based on insights from cognitive science. By this blend of components, the outcome to using the LIVA prototype is a visual map to semantic and intuitive content.

Information visualization itself is a branch of knowledge visualization as a means of knowledge transfer among humans, to some extent running parallel to scientific visualization. By knowledge, we refer to the set of facts held to be true about the world, Plato's "justified true belief". Both scientific and information visualization are concerned with presenting data to users by means of images, in order to help them to explore, make sense, and communicate about data. Since they have overlapping goals and techniques, there is no clear-cut borderline between the two research domains, however, one may say that scientific visualization deals with data that has a natural geometric structure, whereas information visualization handles more abstract data structures such as trees or graphs.

Information and scientific visualization in a knowledge management context have grown big during the past twenty years. Two good overviews can be recommended to the interested reader as a first step toward taming complexity by visual means: the first offers access, with abstracts, sample images and contact details, to over half a million projects,[2] while the second, called the "periodic table of data visualization methods", manages to create order in the methodological toolbox by falling back on the metaphor of the tabular arrangement of elements in chemistry [Lengler & Eppler 2007].

Library application examples of information visualization

For the companies in the LIVA project it has been important to scan new development within the modern library and information science area. The customer base is the driving force for the continuous development and generates the understanding required to produce effective solutions.

The role of information visualization (IV) in the project, is to find ways to contribute to the companies' development of enhanced products and services for the future. Information visualization has a general importance for libraries and information institutions, focusing on the interaction between evolving technological solutions and developing user needs. Several reports about usability-tests in web-based library catalogues and library web pages, have also confirmed the needs for user-oriented development [Madsen, Gardner & Hofman Hansen 2003], [Abrahamsson & Berg 2007] and [Nygren 2006].

From the 1980s onwards library card catalogues have been replaced by remotely accessible computer databases. The resulting electronic catalogues, called OPACs (Online Public Access Catalogues), have many advantages compared to the card catalogues. They are up to date, show which books are on the shelf or checked out, and let one search easily by keywords, single words or phrases in titles, or other access points.

Nowadays when the library catalogues are accessed on the Internet, one will find even more advantages. Online catalogues include links to full text documents, are integrated with ordering forms, electronic payment systems etc., and different user groups can be given access to different databases and information according to their information needs changing over time.

However, in spite of the dramatic difference they have already made, such online catalogues still need improvements to attract users who want easy ways to find answers to their questions and experiment with a plethora of opportunities [Borgman 1996, Lombardo & Condic 2000, Sridhar 2004, Breeding 2007]. For example when one browses for a book, the traditional result lists in online catalogues still give less information than a visit in a "physical", i.e. non-virtual, library. The reason for this is that the overview of a traditional library in its entire complexity is not easily repeated by an online catalogue.

In the library of the old days, you could "browse and navigate" in the card catalogue, on the shelves and by opening the books. Also the book cover, title, binding and size helped one to pick an item and locate valuable information. Another standard problem is that many users search for something they do not know or have difficulty with spelling out which makes the seeking process unstructured and intuitive. This type of seeking is not supported by online catalogues. But how should we redesign them to give better support to information seekers?

Besides developing better search methods and subject catalogues, their increasing popularity begs for better information visualization tools. These tools can give a browseable and compact overview of the search results, in the form of e.g. topical clusters, graphs, or maps. Further, search results can be displayed in context, showing how the items are related. In what follows we refer to a few good examples of what has been accomplished in libraries and information institutions. Since these examples originally come from information science projects, they have been a valuable source of inspiration for LIVA's own research and development activities.

Easy searching with filtering

Today's library systems are inspired by the ease of use of Web 2.0 trends. It is OK to type one or a few words in natural languages. Close to the result set there are interfaces for improving the search through facets and filters. These interfaces are to a large extent based on the bibliographic information visualized in various ways, sometimes as computer graphics. Users should be able to use all aspects of the available bibliographic information, more or less without knowing that they do so. A number of different visual cues are used when visualizing information. Colour, form and texture are used graphically to express relations between resources. Information that can be used to limit or refine searches are shelf marks, topics, genre, format, library, region, era (e.g. 19th century), language, creator, fiction/non fiction, audience, series and new titles (e.g. in the last week, the last month etc). An example from the State Library of Tasmania[3] is shown in Fig 1. Another example comes from LIBRIS (Fig. 2.).

Clusters

The LIVA project has been doing work on automatic classification and clustering. As said above, to classify a resource means that it is assigned to a category within an existing taxonomy, or classification system. When clustering, one measures the similarity between resources. Those that are similar form natural groups, or clusters. Clusters are formed, based on the variables one is comparing. Apart from subject, they may be clustered with respect to persons, events, temporal or spatial coverage, popularity, format etc., as illustrated in Fig 3.[4] More examples such as Aquabrowser,[5] Grokker,[6] Kartoo,[7] Vivisimo[8] and Tafiti[9] are available at the addresses listed below.

Faceted browsing

Faceted browsing is built upon controlled and consistent data, like classification codes. High speed indexes are created and data are presented in hierarchical or cluster like structures. Browsing is suitable for broad subject queries, because users can be given context specific help. A good example for this type of graphical user interface is the North Carolina State University Libraries OPAC[10] in Fig 4.

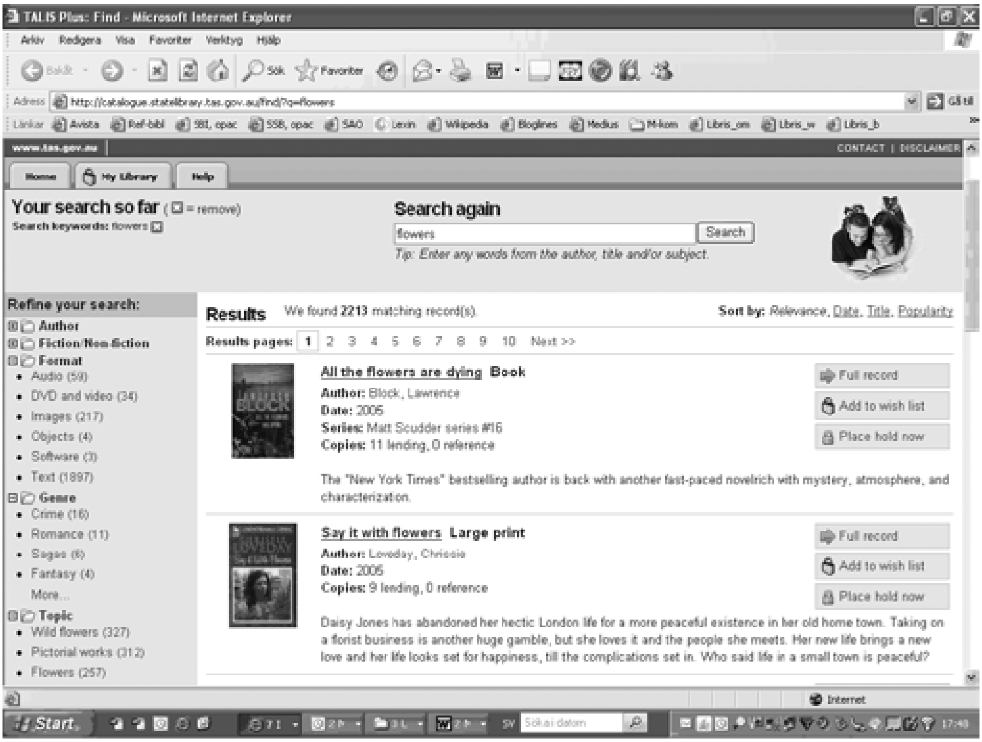

Fig 1: User interface of the TALIS^PLUS online catalogue

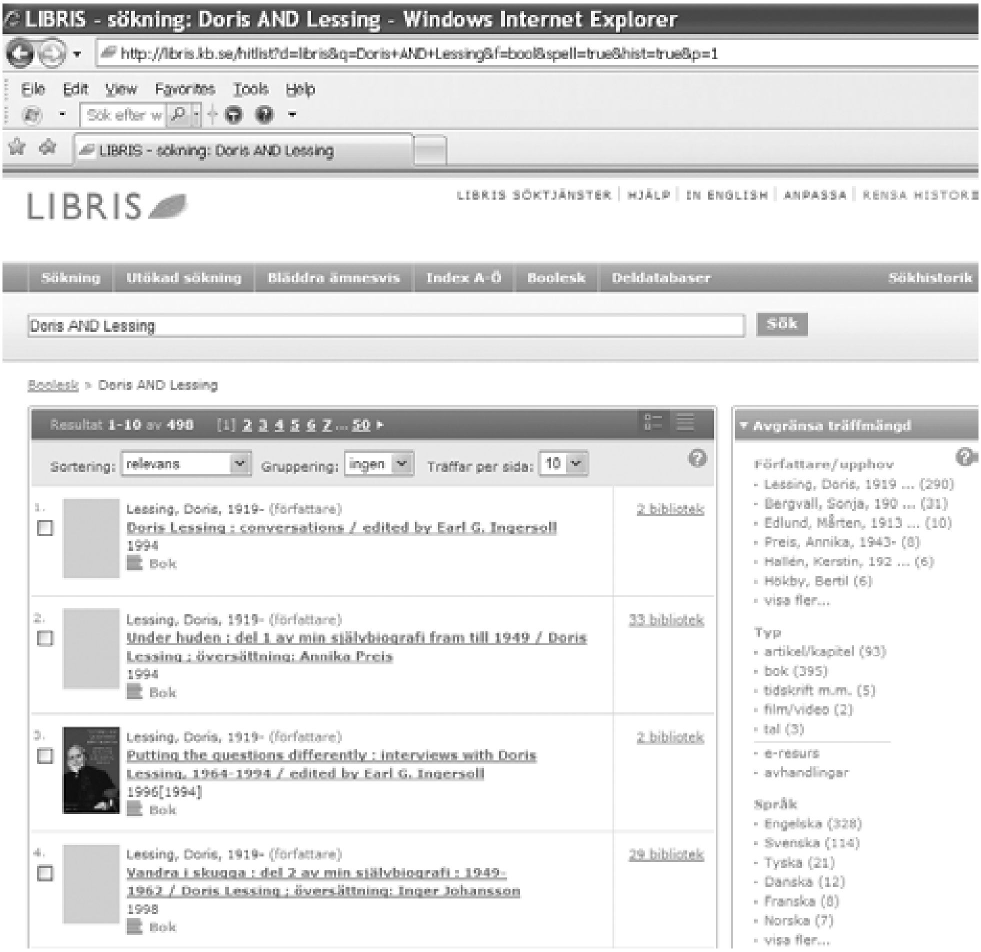

Fig 2: User interface to LIBRIS

Tags and tag clouds

"A tag is a (relevant) keyword or term associated with or assigned to a piece of information (e.g. a picture, a geographic map, a blog entry, or video clip), thus describing the item and enabling keyword-based classification and search of information."[11]. Tags are these days mostly used in folksonomies and can be visualized in tag clouds. Such clouds can be used for a limited number of tags that can aid users getting an overview of resources such as the one shown in Fig 5. This display shows the popularity, frequency, and trends in the usages of words within speeches, official documents, declarations, and letters written by the Presidents of the US between 1776 - 2007 AD.[12] In library systems, tag clouds should usually be an optional part, clickable for a more detailed description.[13]

The information visualization component of the LIVA prototype

While designing the prototype, according to the research and development priorities of the project, we wanted to integrate different tools

Fig 3:
User interface to the Webbrain catalogue

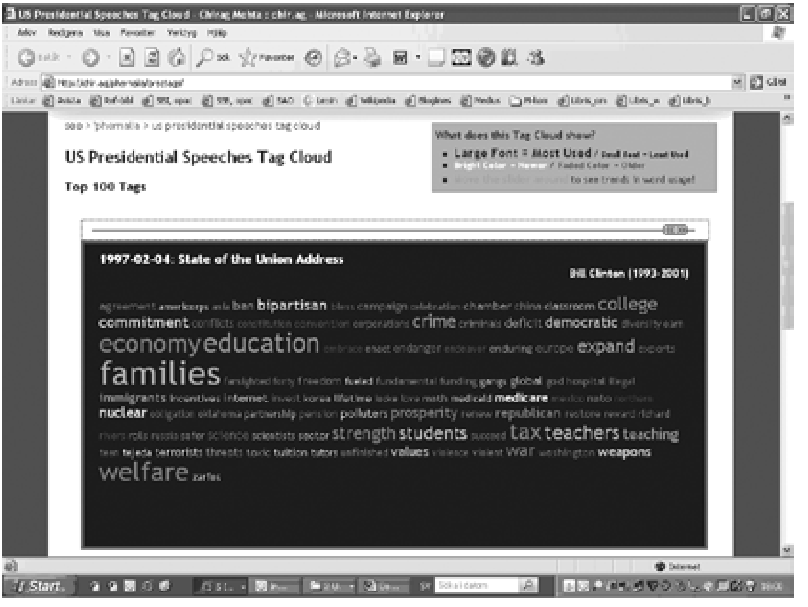

Fig 4:
User interface of the Endeca ProFind™ based integrated library system of NCSU Libraries

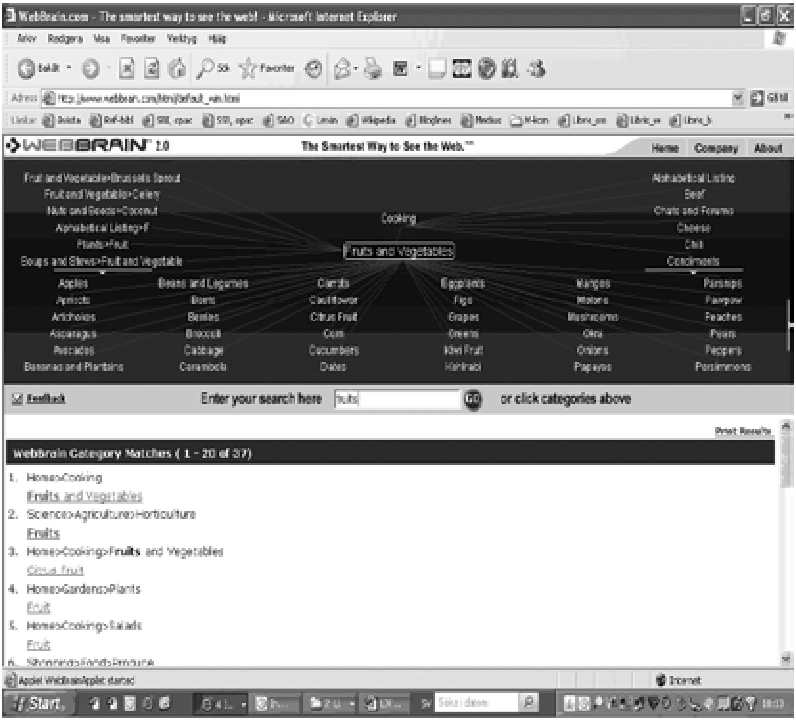

Fig 5: US Presidential speeches aging tag cloud timeline

for automatic document indexing, classification and clustering plus information visualization in a graphical user interface which will help users in both information retrieval and navigation by browsing. Accordingly, several GUI ideas were experimented with and, after some in-house testing, a promising combination of components was developed into several variants and evaluated by students, users and library staff. This user evaluation can form the basis of commercial product development after the project expires. In what follows, first we briefly describe the selection process and then proceed to the introduction of the final result of the project.

As technical background information, after linguistic pre-processing and automatic indexing, we regularly applied clustering and automatic classification methods (latent semantic indexing, principal component analysis, support vector machines), both hierarchical and non-hierarchical, to test data from BTJ, partly relying on the SAB Classification System (Klassifikationssystem för svenska bibliotek). A more detailed first account of our considerations and results was published in *Svensk Biblioteksforskning* [Darányi & Eklund 2007].

In all of our efforts, the crucial step was to apply a visualization metaphor to the semantic content of the test data. We experimented with three such metaphors:

- Document galaxies [Wise 1999],
- Force-directed placement [Walshaw 2001], and
- Contour maps or thematic landscapes [Wise 1999].

As for the parallel by which one expresses represented information, information items as a rule are grouped or ranked based on their similarities, so using e.g. distances for expressing document similarity relies on the metaphor of the document as a *location* in space; expressing similarity by probability considers documents as *events*, and using entropy as the ranking principle of document similarity compares their content to *energy*.

Whereas document galaxies and contour maps support navigation in a database, force-directed placement methods give the user an overview of both the information searching process and single steps of information retrieval. A few snapshots are offered in Figs 6-7.

In Fig 6, we can see 16 so-called subspaces of the complete clustering space, the *x* and *y* axes of the respective coordinate systems standing for different background variables. As background variables are known to represent concepts, these 16 views of the same database help users to access the same data from 16 different combinations of generalized search terms, i.e. concepts. Visual access to clustering space is described in more detail by Preminger [Preminger 2007].

As for Fig 7, here we used a method called quantum clustering for the creation of a thematic landscape based on the probabilities of the index terms in the documents. Red dots represent the documents, and their densities and topical distribution result in contours of the map as if those topics which occur in many documents would create a hilltop, others on a lower level of occurrence a neighbouring meadow, and the least frequent ones a ditch. By looking at such a thematic map to

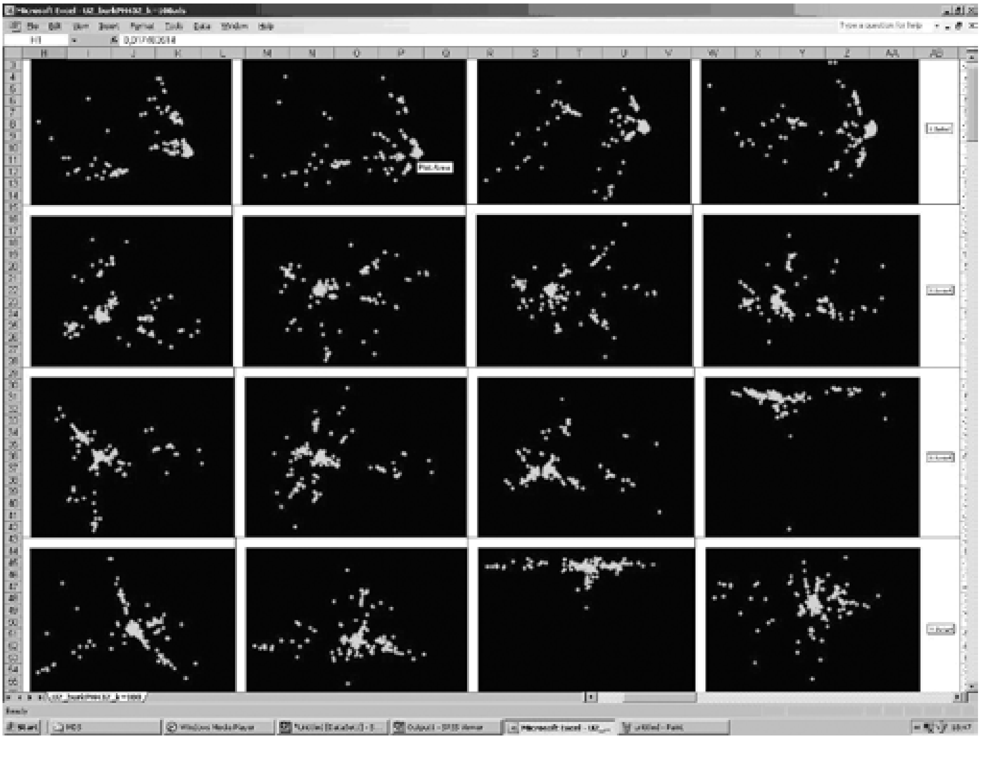

Fig 6: Document galaxies in concept subspaces (BURK-sök® sample Ph class, 432 documents x 1251 index terms, the first 200 documents shown in the space of the first 17 latent variables pairwise arranged)

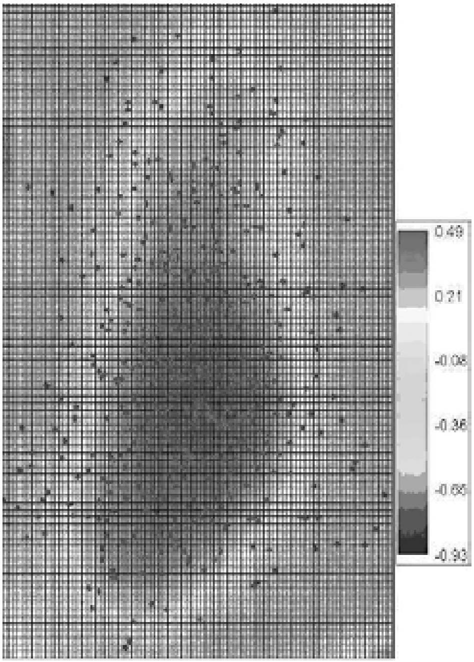

Fig 7: BURK-sök® sample, class Oh [Sociala frågor och socialpolitik], 544 documents x 8928 index terms, potential landscape computed by quantum clustering)

document content, one can easily identify popular or important documents for instance.

As out of these three alternatives, the company representatives in the project, favoured force-directed placement as the method for future GUI development, several versions of the same idea were developed in a prototype to evaluate visual access to different types of information available in the test data (Fig 8).

The user interface of the prototype consists of a hierarchical tree diagram, using force-directed layout for arranging the nodes of the tree, as well as a textual result list. Our main objective for developing the prototype was to obtain empirical information from users regarding the advantages that a graphical interface may provide to facilitate a better understanding of the information structure in the underlying database.

The results of a limited user survey conducted together with the prototype were:

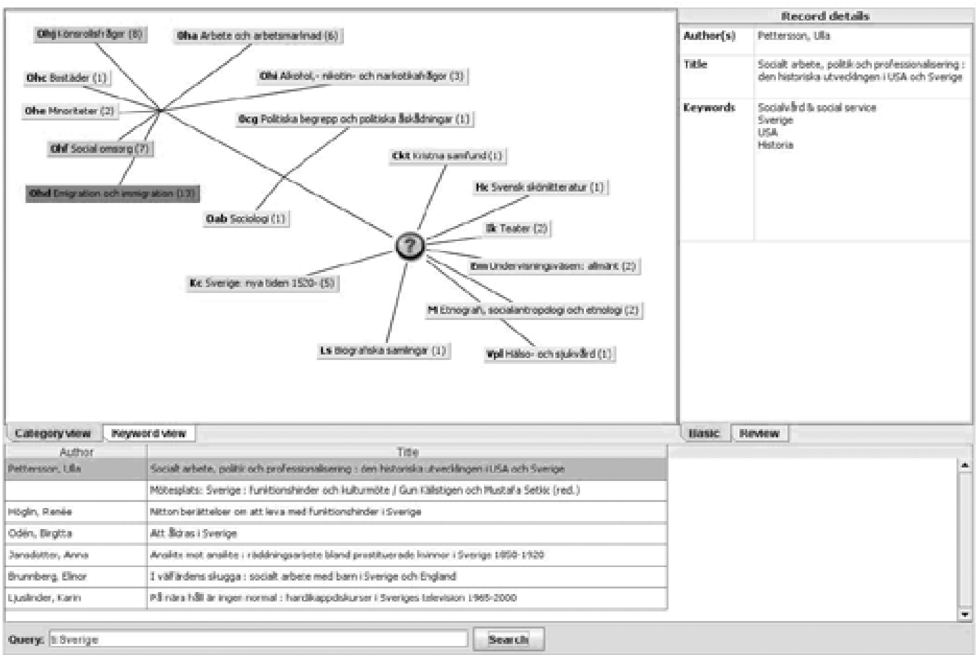

Fig 8: Prototype of the LIVA graphical user interface using force-directed placement

- 77 % of the respondents preferred the graphical interface for result presentation,
- There was a high degree of disagreement between the respondents concerning the question whether the graphical presentation would speed up their work,
- There was a high degree of agreement between the respondents concerning questions about how easy it was to understand the interface and how easy it would be to get used to the interface,
- Among the free text answers a few recurrent remarks were that the graphical and the textual presentations together give a complementary view of the data but that the interface easily gets cluttered when the search result yields many category labels.

A cautious conclusion from this study is therefore that it would add value to a search interface if the results are presented both by textual and graphical means.

Towards Library 2.0.

Web 2.0/Library 2.0 offers new and user oriented ways to build library services. It is driven by technology and users' social networking activities. Functionally, Web 2.0 applications build on the existing Web server architecture, but rely much more heavily on back-end software.

To design an integrated library system (ILS) 2.0 involves questions on how to design a new information service, in a world filled with information services and users seeking information everywhere and everyday. To meet these challenges the libraries need to become as available to virtual users as they are to physical users.

Some libraries have started to experiment with different visualization tools in the Web 2.0/Library 2.0 concept and have moved away from the traditional hitlist orientation. However, there is much more to be done to enhance display and navigation. The display format in the future will probably be a variety and mix of different kinds of results display. As an integrated part of this also user created content will be included. While there is still an open question what kind of visualisation is optimal for different user groups the underlying procedures and improvement show several ways to improve system performance and usability. We hope that the work within the LIVA project and the IV prototype will be one step towards the future library.

References

Abrahamsson, E. & Berg, I. (2007) Hur söker användarna i katalogen? Litteratur- och kunskapsöversikt. At: http://www.kb.se/Dokument/Om/projekt/avslutade/katalogutredning/KU_anvandarstudie.pdf

Borgman, C.L. (1996). Why are online catalogs still hard to use? *Journal of the American Society for Information Science*, 47(7), 493-503.

Breeding, M. (2007). The Birth of a New Generation of Library Interfaces. *Computers in Libraries*, 27(9), 34-37.

Darányi, S. & Eklund, J. (2007). Automated text categorization of bibliographic records. *Svensk Bibiloteksforskning*, 16(2), 1-14, at: http://www.hb.se/bhs/publikationer/svbf/aktuellt.htm

Lengler, R. & Eppler, M. (2007). Towards A Periodic Table of Visualization Methods for Management IASTED. Proceedings of the Conference on Graphics and Visualization in Engineering (GVE 2007), Clearwater, Florida, USA. At .http://www.visual-literacy.org/periodic_table/periodic_table.html [2007-11-06]

Lombardo, S.V. & Condic, K.S. (2000). Empowering users with a new online catalog. *Library Hi Tech* 18(2), 130-141.

Madsen, J. & Gardner, J. & Hofman Hansen, J. (2003): Sådan bliver det elektroniske bibliotek brugervenligt : best practice rapport baseret på usability-test af danske folke-bibliotekers websteder": Köpenhamn: UNI-C (Danmarks IT-Center for Uddannelse og Forskning).

Nygren, E.. (2006). Användarundersökning av bibliotekens hemsidor. LIMIT-projektet. At http://www.lansbiblioteket.se/arkiv/rapport_else_nygren.pdf [2007-11-15]

Preminger, M. (2007). *Uexküll: Multivariate data organization for visual information retrieval.* PhD thesis. Oslo University (in publication).

Sebastiani, F. (2005). Text categorization. In Alessandro Zanasi (ed.), *Text Mining and its Applications*, WIT Press, Southampton, UK, 2005, pp. 109-129, at http://nmis.isti.cnr.it/sebastiani/Publications/Publications.html [2006-09-20]

Sridhar, M.S. (2004). OPAC vs card catalogue: a comparative study of user behaviour. *The Electronic Library* 22(2), 175-183.

Walshaw, C. (2001). A multilevel algorithm for force-directed graph drawing. In Marks,

J. (Ed.): Graph drawing. 8th International Symposium, Colonial Williamsburg 2000. Berlin: Springer, 171-182.

Wise, J.A. (1999). The Ecological Approach to Text Visualization. *Journal of the American Society for Information Science*, 50(13),1224-1233.

Noter

1 For more information about the LIVA project see http://www.hb.se/bhs/liva
2 http://www.visualcomplexity.com/vc//
3 http://catalogue.statelibrary.tas.gov.au/find/?q=flowers
4 http://www.webbrain.com/html/default_win.html
5 http://aqua.queenslibrary.org/
6 http://www.grokker.com/
7 http://www.kartoo.com/uk_index.htm
8 http://vivisimo.com/
9 http://www.tafiti.com/
10 http://www.lib.ncsu.edu/browsesubjects/
11 http://en.wikipedia.org/wiki/Tag_%28metadata%29
12 http://chir.ag/phernalia/preztags/
13 http://chir.ag/phernalia/preztags/