

An ontology-driven approach to accessing and understanding statistical information in the GovStat Project

Maria Cristina Pattuelli

School of Information and Library Science, University of North Carolina at Chapel Hill, CB# 3360, 100 Manning Hall, Chapel Hill, NC 27599-3360. Email: pattm@email.unc.edu

Junliang Zhang

School of Information and Library Science, University of North Carolina at Chapel Hill, CB# 3360, 100 Manning Hall, Chapel Hill, NC 27599-3360. Email: junliang@email.unc.edu

Introduction

The GovStat project (<http://ils.unc.edu/govstat/>) is a project which aims to facilitate end user access to statistical information through a series of interactive interface tools within the Statistical Knowledge Network (SKN) (Marchionini, Haas, Plaisant, Shneiderman, & Hert, 2003). More than simple retrieval tools, these tools are also being developed to help users understand the statistical information available from U.S. government agency web sites.

The complexity of the domain of statistics represents a challenge to information access, comprehension, and use, especially for users with limited or no statistical expertise. Technical and domain-specific terminology (e.g., standard error, Consumer Price Index), term variants (e.g., revenue, income), and complex concepts (e.g., seasonal adjustment) make it difficult for the non-expert user to formulate effective queries and understand the data retrieved.

The use of knowledge organization systems such as thesauri, taxonomies, and ontologies has proven to be effective in supporting traditional information retrieval systems. The application of these systems to the Web is receiving increasing interest (Tudhope & Koch, 2004). In particular, ontologies are seen as the key technology to support the semantic development of the web (Berners-Lee, Hendler, & Lassila, 2001).

Development

Ontologies are conceptual models which represent the knowledge of a domain by defining concepts and relationships held between them. Ontologies can be used as intermediate semantic technologies to support a number of system functions and applications. Increasing the semantic capability of the SKN interactive tools with the use of an ontology enhances information access by improving the quality of the search results, providing context, and augmenting the understanding of statistical information.

The various functions supported by an ontology in the context of the SKN are summarized in figure 1.

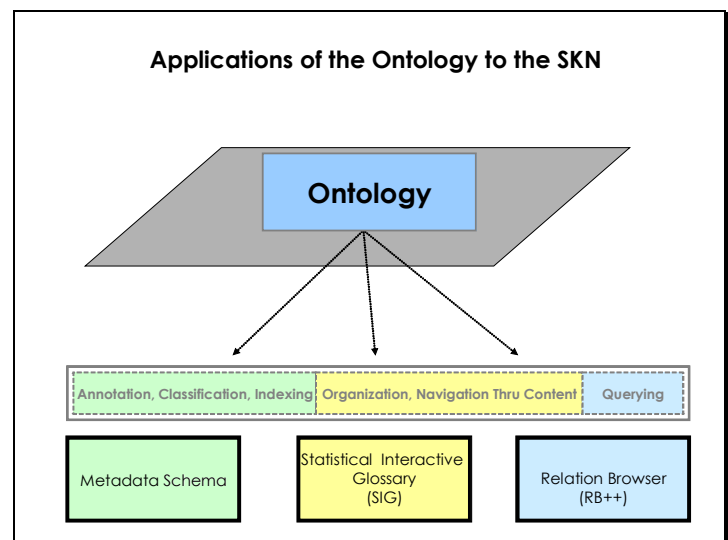


Figure 1.

Our current research focuses on the application of the ontology to the Relation Browser (RB++). The RB++ is a graphical interface tool that enables users to browse and search large information collections (Zhang, 2004). The RB++ provides an overview of the collection by displaying multi-faceted categories. Search capabilities are also built into the interface. By typing keywords into the search boxes, users are able to search the entire collection, as well as filter returned results.

The ontology was developed using Protégé (<http://protege.stanford.edu>), a well-established open-source ontology editor (Knublauch, 2003). Protégé

provides a Java API that enables developers to integrate the ontology into other applications. The RB++, which was developed using Java, utilizes the API to interact with the ontology.

Working on the back-end of the RB++, populated with approximately 17000 webpages from the Bureau Labor Statistics (BLS), the ontology functions as a controlled vocabulary to support query expansion and to facilitate the task of retrieving BLS webpages relevant to the user. In this instance, the ontology-driven approach helps to reconcile non-expert users' terms with the agencies' vocabulary by expanding user queries in order to include terms more frequently used by agencies. For our initial test, two small sets of concepts and the relationship between them have been identified and modeled (Figure 2).

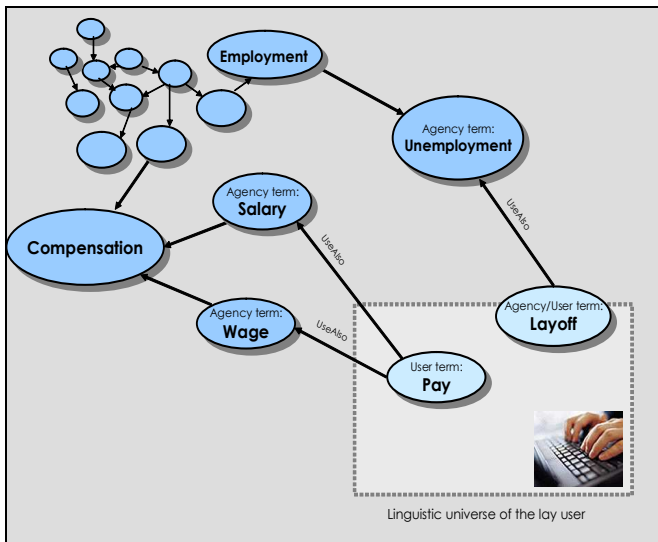


Figure 2.

For example, when a user enters the search term *pay*, instead of retrieving documents that only contain the word *pay* or nothing if *pay* is not a term used by an agency, the user will retrieve documents that also include the word(s) *wage* and/or *salary*, more typical agency terms. The ontology will also help find documents where agencies themselves have multiple terms per concept.

The line between the linguistic universe of the lay users and the agencies often blurs (e.g., *layoff* is both a user and an agency term). Typically, an authority file, that establishes a clear hierarchy between preferred and non-preferred terms, is used. However, such an approach could result in information loss. We opted instead for relating variant terms in semantic associations that would bridge the

gap between user and agency's vocabularies in a more comprehensive way.

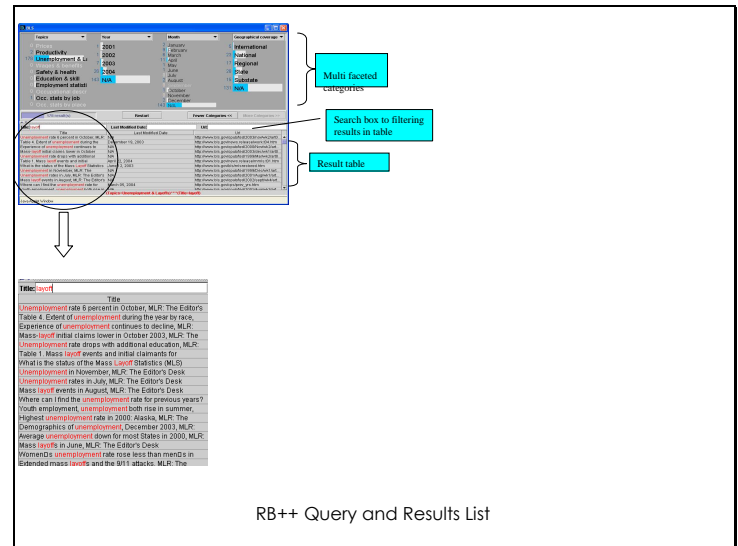


Figure 3.

As illustrated in figure 3, the results from a user's search for the keyword *layoff* will automatically return results which include not only documents containing *layoff*, but also those pages which contain the related agency term *unemployment*. Both terms are highlighted in red giving the user a visual cue that indicates the expansion of query terms.

Conclusions and future work

An initial evaluation of performance shows that the RB++, augmented with the ontology, appears to improve the quality of search results.

Future work includes the further development of the ontology vocabulary and the definition of additional types of relationships (e.g., part-whole and disjoint) to enable and support search capability beyond synonym expansion. The performance of the enhanced RB++ will be tested through a series of usability tests.

Acknowledgments

The GovStat Project is supported by NSF grant EIA 0131824. We would also like to thank the entire GovStat team for helpful comments and discussions.

References

- Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. *Scientific American*, 284(5), 34-43.
- Knublauch, H. (2003). *An AI tool for the real world. Knowledge modeling with Protégé*. Retrieved 02/09/2004, from <http://www.javaworld.com/javaworld/jw-06-2003/jw-0620-protege.html?>
- Marchionini, G., Haas, S. W., Plaisant, C., Shneiderman, B., & Hert, C. (2003). Toward a Statistical Knowledge Network. In *Proceedings of the National Conference on Digital Government Research, dg.o2003. Digital Government Research Center*.
- Tudhope, D., & Koch, T. (2004). New Applications of Knowledge Organization Systems. *Journal of Digital Information*, 4(4).
- Zhang, J. (2004). Relational Browser ++: An interface for exploring and searching large information collections. *The National Conference on Digital Government Research. Seattle, WA*.