Towards an Intelligent system for managing Historical Archive Documents

Akrivi Katifori*

National and Capodistrian University of Athens, Department of Informatics and Telecommunication, Panepistimioupolis, Ilissia, 157 84
vivi@di.uoa.gr

Abstract. This thesis attempts to investigate issues related to the support of information retrieval in digital archives through the use of ontologies and focuses on archives that are related to concepts and entities that change with the passage of time. The problems studied belong to the areas of visualization and information retrieval. More specifically, the focus of the work has been:

- The study and grouping of existing ontology
- The experimental evaluation of ontology visualization methods in order to record their advantages and disadvantages for non-ontology expert users exploring ontologies.
- The study of the possibilities of supporting user document and task management by using a personal ontology.
- The implementation and evaluation of an algorithm using the personal ontology with weights to suggest concepts relevant to those that the user encounters in his/her current task.

This paper briefly describes the thesis objectives and proposed approach.

Keywords: ontology visualization, visualization evaluation, ontology with temporal characteristics, personal ontology, personal interaction management

1 Introduction

The recent progress in the area of digital libraries and the semantic web has lead to new ways of digitizing, organizing and presenting library material, enhanced with the incorporation of semantics. More and more organizations, libraries and document repositories opt for digitizing their material, either for internal use or for publishing it through the web. The great variety of digitized material has brought new needs and several research issues have arisen.

Digitized historical archives (HAs) could be considered as a special case of digital libraries; they have however, characteristics that differentiate them. In particular, the digitization process in the context of HAs is inherently more demanding than the

* Dissertation Advisor: Constantin Halatsis, Professor

equivalent in common digital libraries, mainly due to the large volume of the original material and its poor preservation state, as well as to the convoluted and archaic handwriting often found in documents of HAs. At the best case, keywords or other metadata (creation date, author etc) will be available [3]. Commonly, documents in a HA are fitted into a categorization scheme, which has proven to provide little or no help at all for information retrieval purposes, as it is typically compiled by archivists to suit archiving purposes. As a result, even browsing becomes very difficult without the help of the experienced archive personnel, which mainly relies on their conceptual model of the archive, rather than on some explicit representation of knowledge about the archive content and tools offering guidance and automation for search tasks.

As HAs constitute a very important source for historical research, in this work we attempt to investigate the historical researchers' needs and propose a set of tools to assist them in their research in digitized archives. These tools are based on an ontology, a construct that presents an overview of the domain related to a specific area of interest and may be used for browsing and query refinement. Ontologies model concepts and relationships in a high level of abstraction, providing rich semantics for humans to work with and the required formalism for computers to perform mechanical processing and reasoning.

This thesis has investigated the current state of the art of digital historical archive applications and proposed an ontology based approach to providing advanced functionality to the users. The main results include a detailed analysis and evaluation of existing ontology visualization methods, a tool to support the browsing and visualization of a historical archive ontology, the creation of a personal ontology to support personal information management and the development of a spreading activation algorithm for context inference on top of this ontology.

The following sections briefly outline related work and the proposed approach and the paper concludes with main conclusions and future work

2 Related Work

2.1 Digital Archive Systems

The recent great digitization effort has resulted in the creation of numerous Digital Libraries that may be accessible by historians.

The **European Culture Heritage Online**[2] (ECHO) is a collection of digital libraries of 50 scientific and cultural institutions worldwide, which contribute cultural heritage content as well as scholarly metadata. Access to digitized material, the majority of which concerns philosophy and science, is possible either by inserting key words or by browsing in thematic categories.

The **Perseus Digital Library**[3] of Tufts University contains primary material and secondary sources for research in the humanities, which are accessible by browsing or

[2] http://echo.mpiwg-berlin.mpg.de/home
[3] http://www.perseus.tufts.edu/

by inserting keywords in simple or advanced search. The digital library of Perseus includes various collections, such as the Classics Collection, the Renaissance Collection, the Bolles Collection, the California Collection, the Upper Midwest Collection, the Tufts History and the Boyle's Papers. The site also offers historical information on the related areas. A lot of the primary material is also available in text fromat. For the material in Greek, a transliterated version with comments was chosen, based on various bibliographic sources.

There are other many academic and cultural institutions that have started to digitize primary material, making it available in the form of PDF images. Greek digital libraries such as **Pergamos**[4] and **Hellinomnimon**[5]**,** both of the National and Kapodistrian University of Athens, **Anemi**[6], of the University of Crete, **KENEF**[7], of the University of Ioannina, and others, offer simple and quick access to rich collections of digitized material of relevance to Modern Greek Studies scholars. Most of this material is old and rare. The researcher can browse the digital representations of old prints, manuscripts and visual material for historical, biographic and bibliographic information.

The **National Library of Greece** has developed a digital library[8] that is accessible through the Web. This digital library contains five Greek newspapers in digitized form, covering the period from the end of 19th to the middle of the 20th century. The material appears in PDF form. Each page of the newspapers corresponds to one PDF document. Newspaper issues are accessible either by browsing a calendar for each newspaper or by inserting the desirable date as a keyword. In addition to offering access to the digitized material, this digital library offers the researcher a useful real-time OCR tool for searching into the digitized material. Its major disadvantages are: a) that it does not return all the relevant results, and, b) that the interface of the display of the results is rather inconvenient for the user, since he/she has to open all the relevant pages in order to find what he/she is looking for.

All these efforts do not offer functionality beyond simple keyword search and browsing of the catalogs of the archive collections.

2.2 Ontology Visualization and Evaluation

There are evaluations of hierarchical visualizations, like the ones in [3] and [2]; however, up to this point, comparative evaluations concerning the effectiveness of ontology visualization methods in different contexts and with different user groups have not been conducted. The preliminary results of a survey using questionnaires related to ontology editing tools and ontology visualization are presented in [4].

---

[4] http://pergamos.lib.uoa.gr/dl/index
[5] http://www.lib.uoa.gr/hellinomnimon/
[6] http://anemi.lib.uoc.gr/
[7] http://www.kenef.phil.uoi.gr/en/index.php
[8] http://www/nlg.gr

2.3 Personal Information Management and Spreading Activation

Spreading activation is not a new concept in semantic networks related research. There is a number of proposed applications of spreading activation, especially in the area of information retrieval [11].

Crestani in [12] proposes the use of spreading activation on automatically constructed hypertext networks in order to support web browsing. In this case, constrained spreading activation is used in order to avoid spreading through the whole network, as is the case with our implementation. Liu et al [15] use spreading activation on a semantic network of automatically extracted concepts in order to identify suitable candidates for expanding a specific domain ontology..

Hasan [13] proposes an indexing structure and navigational interface which integrates an ontology-driven knowledge-base with statistically derived indexing parameters, and the experts' feedback into a single spreading activation framework to harness knowledge from heterogeneous knowledge assets.

Neural networks and in particular Hopfield Networks [14] attempt to approach and simulate the associative memory again by using weighted nodes but at a different level. In this case, the individual network nodes are not separate concepts by themselves, but rather, in their whole, are used to represented memory states. This approach corresponds to the neuron functions of the human brain, whereas ours attempts to simulate the human memory conceptual network functions.

3 Proposed Approach

The thesis proposed approach on supporting the historian when using a historical archive to perform historical research has been based on the use of an ontology to model the metatada related to the archive as well as secondary information related to the domain.

Available visualization methods have been investigated to support the use of ontologies as browsing tools and a new visualization tool for ontologies with temporal information has been proposed.

Lastly, the issue of personal ontology management and ontologies has been explored with the modeling of the personal ontology and the design and implementation of a spreading activation algorithm on top of the personal ontology.

The following sections present the proposed approach in more detail.

3.1 Ontology Visualization Evaluation

An important step in the thesis research was the sturdy of the existing ontology visualization techniques and the comparative evaluation of four ontology visualization methods that followed.

The first phase was the study and grouping of existing ontology visualizations that have been proposed in the existing literature or implemented in ontology editing tools [5].

The focus of the ontology visualization evaluation that followed [6] [7] was on users of varying expertise on ontology browsing in an ontology whose structure is unknown but whose concepts are to some extent familiar. The ontology models the domain of a University Historical Archive and it is rich with information concerning the evolution of entities, persons and institutions throughout the Organization history. Particular attention has been given to tasks related to the representation of entity evolution.

The results of this evaluation have been analyzed with respect to two main aspects of the experiment, i.e. the evaluation of the examined visualization methods and the evaluation of the ontology itself. Furthermore, several more issues have been recorded, which are related to the methods users employ for dealing with various browsing tasks and the effects of ontology visualization on ontology learning.

3.2 TimeViz -Visualizing Temporal Ontologies

Taking into account the digital HA characteristics and historian needs, we propose a set of tools to aid historical research in the context of an HA. These tools are based on the ontology of the organization the material of which the archive contains.

An ontology was chosen as the means for organizing knowledge since it offers a number of advantages. The creation of an ontology for an HA is complex process, since the digitized archive documents are not in text format, making automatic concept extraction impossible, and the concepts that must be captured may vary among different time periods. For compiling the ontology, the user-centric methodological approach presented in [1] was used. Structured interviews with university academic and administrative personnel, queries made to the archive, existing ontologies, archive categorizations, selected archive material and other available sources, like books, yearbooks etc, were employed in an iterative process.

In order for the user to benefit from all the available meta-data information concerning time in the context of the archive ontology, there is a need for an appropriate visualization tool. TimeViz allows the user to navigate in the ontology or select specific entities in order to view their course in time. Protégé [9] was selected as the ontology management tool the functionality of which would be augmented. This is due to a number of advantages it provides related to expendability.

The timeViz Protégé plug-in consists of four parts:

An explorer – like view of the ontology, namely the one provided by the Protégé Class Browser. Classes with temporal attributes are designated by a small clock next to their description.

An **Instance View** window were all the instances of a selected entity are presented. Instances that have evolved along the time axis (any attribute has changed its value) are again designated by a small clock next to their instance icon.

Figure 1. The Instance View with clocks next to the instances that have changed during the selected time period.

A **Timeline** that is visualized as an horizontally placed bar at the top of the main visualization window.

The **Main Visualization** window which uses a node-link style visualization for the representation of the ontology.

The **Entity Evolution** window. The user may select an entity or instance and view its course over the selected period.

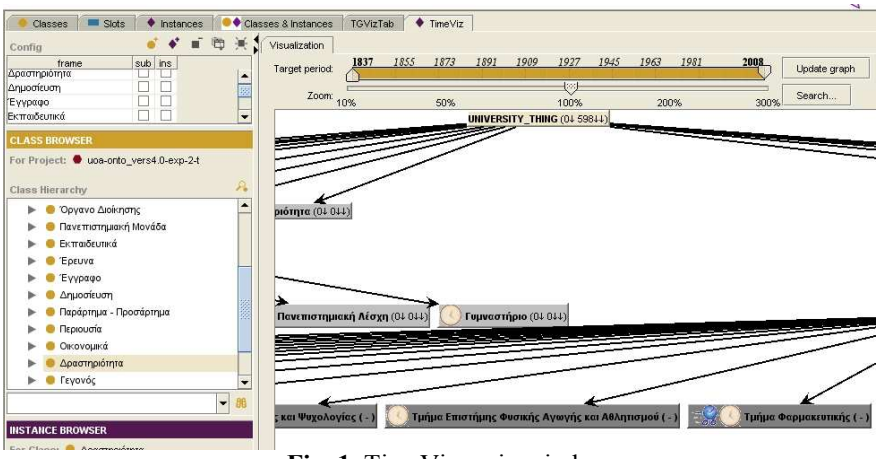

Fig. 1. TimeViz main window

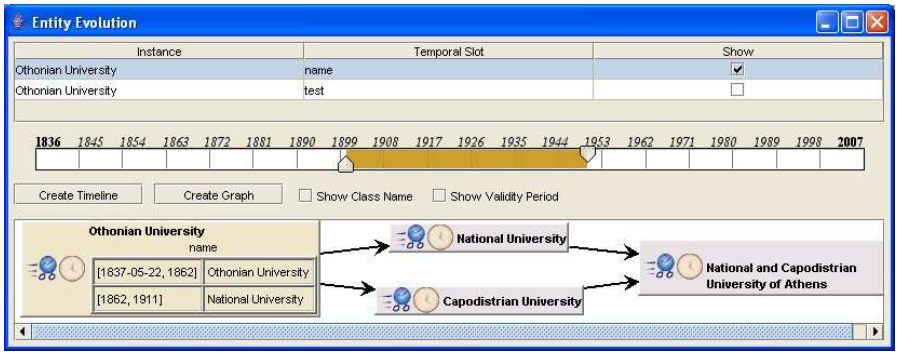

Fig. 2. TimeViz entity evolution window

More information on the design and implementation of the TimeViz tool for visualizing ontologies with temporal characteristics may be found in [8][9][10].

3.3 Personal Ontologies and Spreading Activation

In the context of any PIM/TIM system the personal ontology has a very important part to play. On one hand, it may constitute a useful repository of information related to many aspects of the user's personal and professional life. With the appropriate interface the ontology may become an easily customizable repository of information

that may serve as a memory complement for the user. On the other hand, coupled with intelligent mechanisms, the ontology may become invaluable for context inference in the process of supporting user tasks through task inference.

To this end, we have created an ontology for the user's personal collection domain. This ontology has been created taking into account existing profile models in applications, as well as related research in the area of profiling. The study of the possibilities of supporting user document and task management by using a personal ontology [16][17].

In the course of user interaction with a system, multiple timescales can be noted, which roughly correspond to the ones described in Table 1.

Table 1. Multiple timescales of the human memory

human memory	timescale	mechanism	brain effect
long term	indefinite	physical	synapse growth
short term (Miller's 7±2)	10-30 seconds	electrical	neuron firing
mezzanine (no "proper" name)	minutes to hours	chemical	long-term potentiation LTP

First, there are the contents of the personal ontology and the available information sources that roughly correspond to human long-term memory. Corresponding to the short/working memory are the things the system has to store regarding the current user task – for example, the contents of the email the user has just opened, the text the user has just selected, the web page just visited, or the form field being completed. Finally, there are the things the user has been recently doing (other pages visited, documents seen, etc.) that roughly correspond to the mezzanine memory. This recent history is important as, for example, if the user has recently viewed a web site about an upcoming event and then goes to a travel website it is likely that the place to be visited is that of the event.

These different levels could be dealt with in a spreading activation framework by simply fading memories over time so that entities recently encountered multiple times become increasingly highly "activated". However, with a single mechanism it is hard to create a balance between having recent things be more active (the place just mentioned in an email) than important general things (the user's address), whilst on the other hand not having them crowd-out the longer-term things.

The thesis outlines a spreading activation over a personal ontology framework to be used in the context of a Personal Interaction Management System. The human brain and the theories related to the different levels of human memory and spreading activation have been the incentive of this work.

The implementation and evaluation of an algorithm using the personal ontology with weights to suggest concepts relevant to those that the user encounters in his/her current task [18][19][20][21][22].

4 Conclusions

This thesis has proposed a set of ontology-based methodologies and tools to support the historian in information retrieval as well as to organize his/her personal archive. Preliminary evaluations of the created tools have been concluded with positive results.

The next steps include the testing of the proposed methods in working applications, the design of further evaluations of cognitive and other issues related to ontologies and the development of the spreading activation algorithm with new features like taking advantage of short term relation weights.

Acknowledgments. This thesis has been concluded under the PENED 2001 framework.

References

1. Torou, E., Katifori, A., Vassilakis, C., Lepouras, G., Halatsis, C., Creating an Historical Archive Ontology: Guidelines and Evaluation, Proceedings of ICDIM 2006, December 06-08, 2006, Bangalore, India
2. Wiss, U., Carr, D., and Johnson, H., 1998, Evaluating Three – Dimensional Visualization Designs: a Case Study of Three Designs. In Proceedings of the Second International Conference on Information Visualisation (IV'98), p. 137.
3. Kobsa, A. 2004. User Experiments with Tree Visualization Systems. In IEEE Symposium on Information Visualization (INFOVIS'04), 9-16
4. Ernst, N. A., and Storey, M.-A., 2003, A Preliminary Analysis of Visualization Requirements in Knowledge Engineering Tools, University of Victoria, Victoria, CHISEL Technical Report August 19, 2003
5. A. Katifori, C. Halatsis, G. Lepouras, C. Vassilakis and E. Giannopoulou, "Ontology Visualization Methods – A survey", *ACM Computing Surveys*, Vol. 39, No. 4, October 2007
6. A. Katifori, E. Torou, C. Halatsis, C. Vassilakis, and G. Lepouras, "A Comparative Study of Four Ontology Visualization Techniques in Protégé: Experiment Setup and Preliminary Results", IV 2006
7. A. Katifori, E. Torou, C. Halatsis, C. Vassilakis and G. Lepouras, "Selected Results of a Comparative Study of Four Ontology Visualization Methods for Information Retrieval tasks", Proceedings of the RCIS 2008, June 3-6 2008, Marrakech, Morocco
8. A. Katifori, E. Torou, C. Vassilakis and C. Halatsis, "Supporting Research in Historical Archives: Historical Information Visualization and Modeling Requirements", Proceedings of IV 08
9. A. Katifori, E. Torou, C. Vassilakis, G. Lepouras, C. Halatsis and E. Daradimos, "Historical Archive Ontologies – Requirements, Modelling and Visualization", Proceedings of the RCIS 2007 Conference
10. A. Katifori, C. Vassilakis, G. Lepouras, C. Halatsis and E. Daradimos, "Visualizing a Temporally – Enhanced Ontology", Proceedings of the AVI '06, May 23-26, 2006, Venice, Italy
11. Crestani F.. Application of spreading activation techniques in information retrieval. Artificial Intelligence Review, 11(6):453–482, December 1997

12. Crestani, F. Retrieving documents by constrained spreading activation on automatically constructed hypertexts. In Proceedings of EU- FIT 97- Fifth International Congress on Intelligent Techniques and Soft Computing, pp. 1210-1214, Aachen, Germany.
13. Hasan, M. (2003). A Spreading Activation Framework for Ontology-enhanced Adaptive Information Access within Organisations. In Proceedings of the Spring Symposium on Agent Mediated Knowledge Management AMKM 2003, Stanford University, California, USA
14. Hopfield, J. (1982). Neural networks and physical systems with emergent collective computational properties. Proceedings of the National Academy of Sciences of the USA, 79:2554 - 2588.
15. Liu, W., Weichselbraun, A., Scharl, A. & Chang, E., (2005). Semi-Automatic Ontology Extension Using Spreading Activation, Journal of Universal Knowledge Management, no. 1 (2005), pp. 50 - 58
16. M. Golemati, A. Katifori, C. Vassilakis, G. Lepouras and C. Halatsis, "Creating an Ontology for the User Profile: Method and Applications", Proceedings of the RCIS 2007
17. A. Katifori, C. Vassilakis, I. Daradimos, G. Lepouras, Y. Ioannidis, A. Dix, A. Poggi and T. Catarci, "Personal Ontology Creation and Visualization for a Personal Interaction Management System", *Proceedings of PIM, CHI* 2008
18. A. Dix, T. Catarci, B. Habegger, Y. Ioannidis, A. Kamaruddin, A. Katifori, G. Lepouras, A. Poggi and D. Ramduny-Ellis, "Intelligent context-sensitive interactions on desktop and the web", Proceedings of workshop on Context in Advanced Interfaces at AVI2006, Venice, Italy, May 23, 2006
19. A. Dix, A. Katifori, A. Poggi, T. Catarci, Y. Ioannidis, G. Lepouras and M. Mora, "From Information to Interaction: in Pursuit of Task-centred Infornation Management", DELOS Conference 2007, December 2007, Tirrenia, Pisa
20. V. Katifori, A. Poggi, M. Scannapieco, T. Catarci, and Y. Ioannidis, "OntoPIM: how to rely on a personal ontology for Personal Information Management", *Proceedings of the* First Workshop on the Semantic Desktop, November 2005
21. A. Katifori, C. Vassilakis and A. Dix, "Using Spreading Activation through Ontologies to Support Personal Information Management", Proceedings of CSKGOI, within IUI, 2008
22. A. Katifori, C. Vassilakis, and A. Dix, "Ontologies and the Brain: Using Spreading Activation through Ontologies to Support Personal Interaction", Cognitive Systems Research, Special Issue on Brain Informatics, 2009