Semantically Holistic and Personalized Views Across Heterogeneous Information Sources

Cormac Hampson
Knowledge and Data Engineering Group
Trinity College Dublin, Ireland
hampsonc@cs.tcd.ie

Abstract

A consolidated view of the content within different data repositories would facilitate useful operations such as advanced informational discovery. This paper proposes a methodology and an architecture that uses semantic annotations to enable such holistic views across domain specific information sources. The potential for the personalization of this annotated information is described, and the future work necessary to implement the system is elaborated on.

1. Introduction

The exponential growth in digitized content has resulted in a greater need for ways to understand and manipulate large data sets in a more intuitive fashion. Much of this content is semantically ignorant of the information it conveys, and that metadata that does exist is often insufficient to address the needs of highly dynamic environments such as those associated with visual data-mining and context-aware applications [1, 2].

The type of metadata most prevalent is of a kind that is *rigid* in its descriptions not allowing for more flexible and *softer* associations to be made between different content sources. An example of *rigid* metadata would be the use of the tag <timestamp> to give a precise timing of an event, whereas a *softer* tag would be <freshness> which would give an indicator of how recently the event occurred. Furthermore, many systems that do utilize metadata do so in a very application specific manner, not supporting interoperability between multiple heterogeneous information sources. If these issues can be overcome and integrated into a platform that allows user preferences to personalize the information, this system would be an invaluable tool for applications of vastly different domains to enrich the data so central to their operation. For example, if a semantic visualization application utilized such a platform, different information sources could be added and removed as required, with the aggregated content used to find underlying relationships and patterns that could then be personalized for the user.

In accordance to a holistic philosophy the richness of a knowledge base cannot be determined merely from the sum of its component facts alone. Thus much of this richness will remain intangible unless semantic annotations are created to make this layer of meaning more apparent to users. In this research these annotations take the form of what we refer to as *semantic attributes*, which are used for the personalization of diverse information sources. By combining these attributes with a pre-processed version of the data source's original metadata (see section 3.1), processes and trends not previously apparent due to their distribution across different sources become visible, facilitating a holistic view across the data. These views are in essence a consolidated model of the domain's content which the user can query over.

As the prevalence and dependence on electronic data escalates, the need for systems capable of eliciting latent semantics from heterogeneous sources in a holistic fashion will be increased. This paper describes the SPACE platform (Semantically Personalized

Annotations for Content Enhancement) that will be used to investigate the benefits of (semi-) automatically annotating heterogeneous data sources (databases, web services etc) in order to give a holistically and more semantically meaningful view of domain specific information.

2. Related Work

This research touches on a number of fields of research including semantic annotation, semantic inferencing and information classification. Much work in the semantic annotation field has involved annotating heterogeneous sources for the semantic web. As a result many have had degrees of success in the semi-automatic annotation of unstructured data. It is hoped that this research's focus on structured, domain specific data will enable more accurate annotation once the rules for the domain have been created.

UIMA [3] was created by IBM as an open source platform to support knowledge discovery in unstructured information sources. It offers robustness and extensibility, as well as tools for semantic analysis of inputted data. As such it may add considerable value to the data-centric phase of the SPACE platform (see section 3.1). MUSE [4] is an information extraction platform that uses conditional processing for semantic tagging. Reeve and Han [5] describe how MUSE "has shown that rule-based systems can equal the performance of machine learning-based systems". Hence, though machine learning techniques are commonly used in semantic annotation platforms such as Armadillo [6] and Ont-O-Mat [7], its use is not a pre-requisite to success in the annotation field.

In Accenture they created a system to infer semantic attributes for retail data mining [8] Ghani and Fano concluded that the choice of features to classify a domain "should be made with particular applications in mind and that extensive domain knowledge should be used". There will be a necessity for application specific semantic attributes to be included, however it is envisaged that the suite of generic semantic attributes will be applicable to a large number of heterogeneous domains.

The definition of generic semantic attributes will involve classifying data based on its intrinsic properties. Other attempts to do this include Evans and Wurster's [9] classification of the richness of information in terms of *bandwidth, currency, customisation, interactivity, relevance* and *security*. Though their work concentrated on the economics of digital information, it is still one of the few examples of classifying data using generic semantic attributes. The different focus of their research means that despite some of the attributes being generic enough for the purposes of this work (*security* and *currency*), others would only suit very specific applications accessing the SPACE architecture.

Little et al [10] used Dublin Core metadata and pre-defined rules to dynamically generate intelligent multimedia presentations through semantic inferencing. Their media inputs contained metadata a priori, which is similar to the design of the SPACE architecture outlined in the next section. However they did not use this metadata as a basis to enrich the content with another layer of semantics. They also found severe limitations with using Dublin Core for semantic attributes. For instance they found "unqualified Dublin Core too simplistic to infer many rich or interesting semantic relationships". Furthermore they discovered that inferencing rules were hampered by unstructured metadata values and incompatible schemas. These limitations should be overcome in the SPACE architecture by pre-processing the inputted sources before applying semantics, and by using a new suite of semantic attributes tailored to fuzzy inferencing coupled with bespoke attributes for particular domains.

3. Objectives and Initial Design

The main objectives of this Ph.D. research, which is currently in its first year, are as follows:

1. Derive and validate a methodology, and associated technical framework, for producing semantically holistic views across heterogeneous information sources.
2. Build a system that is extensible, allowing structured data from heterogeneous sources to be added and overlaid with semantic attributes.
3. Devise a generic set of semantic attributes and evaluate them as to their flexibility over multiple domains.
4. Investigate domain specific attributes and elicit their relevance compared to generic attributes.
5. Automatically assign appropriate degrees of semantic attributes to inputted data with reference to the domain ontology.
6. Shape the semantic associations made between annotated facts through user preferences.

7. Ensure these associated clusters will be consolidated, semantically clear and provide meaningful hooks for a variety of external applications or services to exploit.

Thus the overall aim of this research is to facilitate a holistic look at domain specific information, towards facilitating advanced exploration and inferencing. A typical scenario involving SPACE's use would be a visualisation application using the semantically enhanced outputs to enable more sophisticated visual data mining to take place. Likewise, semantic web agents could utilize the annotated outputs to refine their searches or give more pertinent recommendations.

In order to achieve this an architecture and emerging methodology have been designed with two distinct phases. The data-centric phase annotates information according to their intrinsic characteristics, and the user-centric phase associates these annotated facts together depending on the preferences of the user. Figure 1 below shows a diagram of the SPACE architecture.

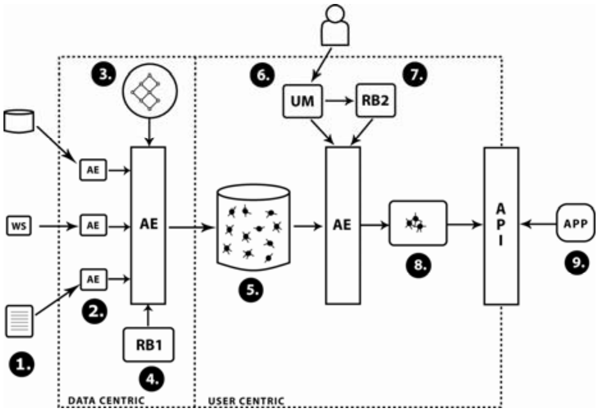

Figure 1. The SPACE Architecture

3.1 Data-Centric Phase

1. All inputs to the system must come from structured data and contain a schema so that the semantics of the information can be understood and utilized by the annotation model. These inputs can be either static or dynamic and will provide the domain specific heterogeneous sources on which the platform performs its annotations. Examples of potential inputs include databases, web services and log files.
2. Using a series of Adaptive Engines [11], these inputs are pre-processed so that they are syntactically in an understandable format for the architecture. Adaptive engines enable the runtime consolidation of multiple models according to a flexible narrative, thus enabling real-time annotation, personalization and semantic association of dynamic information sources.
3. With reference to the domain ontology, input tags representing new information may need to be transformed so that they're compatible with the rule base. Thus if necessary, input metadata will be checked for synonyms and changed appropriately. Key concepts and relationships in the domain will also be identified here.
4. In order to classify all inputted data with semantic attributes, rules are devised by domain experts and knowledge engineers. A generic set of semantic attributes (concepts such as *trust* and *freshness*) that can be used by any domain are made available, as well as bespoke attributes tailored to a specific area. The intrinsic characteristics of incoming data are then cross-referenced with all attributes in order to detect matches, and to fire rules which annotate the information. These annotations may be as simple as an integer weighting, or else a more complex XML description may be required for particular instances.

3.2 User-Centric Phase

5. All annotated outputs from the data-centric phase are stored so that each new annotated fact can be cross-referenced with all others if required by user preferences.
6. Each of the semantic attributes associated with a fact can have its importance increased or decreased depending on the current user preferences. This is achieved by allowing the user model and the second rule base to work in tandem to extract and sequence the stored data relevant to users. The user model contains the overall preferences of the user, as well as their current focus of interest. These preferences and interests play a key role in influencing what annotated facts are associated together.
7. The shaping of semantic attributes and prioritization is enabled by equating user preferences with agendas defined in the rule base.

These agendas are similar to preset functions in a graphic equalizer, influencing which semantic attributes should be stressed and prioritized when associating facts in the working memory. For instance if a person is only interested in recent information and is not much concerned how reliable the source is, the semantic attribute *freshness* will be boosted in importance whereas *trust* will be decreased.

8. The result of this injection of user preferences are clusters of facts associated together in an appropriate fashion.
9. A variety of applications and services will have access to these clusters of data via an API, and use the extra richness to enhance their functionality.

4. Future Work

The next phase in this work is to perform two experiments using the action research methodology. The first will focus on the consumption of heterogeneous data sources via a standard interface, allowing a homogenised version of information to be annotated with semantic attributes. Appropriate semantic attributes (generic and domain specific) will have to be finalised and several technologies integrated in order to complete this experiment. It is anticipated that this experiment will be completed by December 2007.

The second experiment will focus on the personalization and association of semantically annotated data. The integration of user modelling and the AI implementation of rules and agendas for appropriate semantic linking will be the core obstacles to be overcome in this experiment. The results from these experiments will then be used as the foundation for a larger experiment in the final stage of the Ph.D. This will involve the creation of an end-to-end system that allows the outputted data to be consumed by a variety of applications and services. To this end, a semantic visualization application currently in development has already been targeted to consume the outputted clusters from the SPACE platform.

5. Conclusion

This paper has introduced the rationale and design behind the SPACE platform as it stands seven months into the Ph.D. cycle. The motivation and objectives for the research were explained, and an initial design and architecture detailed. Finally, planned experiments for the technical implementation of the architecture and the refinement of the methodology were described.

6. References

[1] E. O'Neill, D. Lewis, K. McGlinn, and S. Dobson, "Rapid user-centred evaluation for context-aware systems," Proceedings of DSV-IS, Dublin, 2006, pp. 220-233.

[2] A. A. Nazari-Shirehjini and F. Klar, "3DSim: Rapid Prototyping Ambient Intelligence," Proceedings of joint sOc-EUSAI, Grenoble, 2005, pp. 303-307.

[3] IBM, "UIMA, An Open, Industrial-Strength Platform for Unstructured Information Analysis and Search." http://www.research.ibm.com/UIMA/

[4] D. Maynard, "Multi-Source and Multi-Lingual Information Extraction," *Expert Update*, vol. 6, 2003, pp. 11-15.

[5] L. Reeve and H. Han, "Survey of semantic annotation platforms," Proceedings of ACM symposium on Applied computing Santa Fe, New Mexico, 2005, pp. 1634-1638.

[6] A. Dingli, F. Ciravegna, and Y. Wilks, "Automatic Semantic Annotation using Unsupervised Information Extraction and Integration," Proceedings of K-CAP Sanibel Island, Florida, 2003.

[7] P. Cimiano, S. Handschuh, and S. Staab, "Towards the Self-Annotating Web," Proceedings of World Wide Web, 2004, pp. 462-471.

[8] R. Ghani and A. E. Fano, "Using Text Mining to Infer Semantic Attributes for Retail Data Mining," Proceedings of International Conference on Data Mining, Maebashi, 2002, pp. 195-202.

[9] P. Evans and T. S. Wurster, *Blown to bits: how the new economics of information transforms strategy*. Boston, Mass: Harvard Business School Press, 2000.

[10] S. Little, J. Geurts, and J. Hunter, "Dynamic Generation of Intelligent Multimedia Presentations through Semantic Inferencing," Proceedings of ECDL, Rome, 2002, pp. 158-175.

[11] O. Conlan, V. Wade, C. Bruen, and M. Gargan, "Multi-Model, Metadata Driven Approach to Adaptive Hypermedia Services for Personalized eLearning," Proceedings of Adaptive Hypermedia, Malaga, 2002, pp.100-111.

7. Acknowledgements

This research has been supported by The Irish Research Council for Science, Engineering and Technology: funded by the National Development Plan