# Ontology-driven data management with Topic Maps,

**Frederic Andres, Rajkumar Kannan**

*Abstract*-- **Our study involves associating a context with semantic indexes and descriptors based on topic maps for efficient ontology-driven data managelent (e.g. blogs, cooperative platforms, digital libraries, document versioning tools, etc.). We believe that supporting contextual semantics can add a set of properties that vary according to the use of ontological topic maps (TM). Their use is beneficial for users who intend to exchange semantic knowledge in an application domain. Existing semantic layers in current systems are not yet capable of supporting relevant contextual semantic description. In this paper, we aim at extending the many-sorted algebra formalizing the topic maps layer in order to support 5W1H contexts (What, Why, Where, Who, When and How) using TMBLOG system as case study.**

## I. INTRODUCTION

In order to retrieve relevant information, internal search engines are commonly used to navigate through index databases of posting management systems (PMS: blogs, cooperative platforms, digital libraries, document versioning tools, etc.). Retrieval methods vary from one engine to another but are mainly related to the attributes defined in each application even if the PMS being searched is based on a simple index database (related to the posted data), augmented (index with hierarchical metadata) database, or multiple/integrated index databases. For instance, in library indexing (e.g. Bliss, Dewey, Goettingen, LC, Ranganathan, Riders, etc.), classifications are done by librarians without necessarily understanding the resource (books, documents, reports, etc.) contents. The used indexes are based on standard library attributes and methodologies; consequently, information retrieval is restricted because it is limited to attributes without considering the content. Inspired by the model proposed in [8], our goal is to integrate topic maps (TM) into the retrieval process in order to provide relevant and richer results by allowing the user to search outside the known and/or predefined attributes and norms of classifications by including document content. In this paper, we study two issues. The first issue is related to TM creation from documents and management of this creation in a collaborative environment by integrating six contextual parameters. The second issue is connected to the improvement of information retrieval in TMs. We typify the related problems with concrete examples in the following sections.

The rest of this paper is organized as follows. Section 2 describes the related work on TMs and information retrieval. Section 3 presents our approach and management of topic maps for semantic augmented information. It also shows our architecture to

improve semantic management using TMs. Section 4 describes a case study regarding contextual semantic management. Section 5 concludes this study and discusses ongoing work.

## II. RELATED WORK

Several studies exist on multimedia[12] data management and retrieval. For instance, the search engine NIX (NASA Image eXchange: http://nix.nasa.gov) allows users to search NASA's online image and photo collections over the web. Retrieval is done on texts associated to images by using Boolean operators. Other famous search engines (like Google, Altavista, Yahoo!, Amore, and MSN) use similar methods to search the multimedia content of web pages. The main differences between these engines are in how they automatically locate textual descriptors in the web pages according to their position and importance to the image to be described. The drawbacks of using traditional textual-based methods for multimedia retrieval have been identified by the scientific community and other users. In fact, there are several directions of research attempting to extend current methods with ontologies and taxonomies to rewrite user queries and improve application results. Nevertheless, to the best of our knowledge, none of the existing methods can utilize topic maps for semantic augmented information search and retrieval. The major benefit for end-users of being able to do so would be access to seamless knowledge under ISO standard 13250.

Several studies have used topic maps to visualize, adapt, and otherwise represent information to the user. For instance, the ENWiC (EduNuggets Wiki Crawler) project [1] for students provides an interface to wikis "*in order to take advantage of the large information repositories*". It represents the structure of a wiki as a topic map (automatic creation of instances, associations, and topic) but it only makes use of TMs for adaptation of existing databases. Krötzsch et al.'s work [2] on Wikipedia claimed that information searching in Wikipedia was primitive. Although their solution used semantic technology, it goes without saying that imposing semantic technology on an already existing information base will not achieve much without initially considering the underlying information. An approach to annotation of the Wikipedia website is provided in reference [13]. It consists of creating a layer of information on the existing information that does not directly affect the structure or content of the existing information. In a related work, the KendraBase project [14] was conceived to enable people with diverse ideas to collaborate for knowledge

[1] In this section, we cite examples from information management in multimedia databases, blogs, and wikis because this sort of information management is comparable to information management using topic maps (i.e., we are not attempting to draw a border line between these terminologies).

[2] We use the word multimedia to describe documents including several data types such as text, audio, video, and image.

creation. The project was intended to be "a semantic wiki / database with auto form generation for data input and queries". The Edunuggets project [5] considered how to provide a personalized knowledge repository in a learning environment.. In references [1, 8], the overall objective was to support on-line teaching and learning by: (a) providing a means to evaluate and annotate available information, (b) providing a context for the organization of the available information, and (c) supporting learners' access. Their work was a way of reorganizing and "repacking" existing information. The approach of the System for Universal Media Searching (SUMS) is not just limited to the organization of the information base; it extends to "reorganization" of the information base according to the user's knowledge. SUMS is a tool for finding, retrieving, and organizing material on the Internet aiming at linking personal local collections of facts with the external electronic world [10]. In "Towards a Semantic Web for Culture", Prof. Kim Veltman asserted that "Culture is about both objects and the commentaries on them; about a cumulative body of knowledge; about collective memory and heritage. … In this context, the science of meaning (semantics) is necessarily much more complex than semantic primitives".

For all these reasons and for efficient enhanced information retrieval, we believe that there is need to combine information creation and information search for a specific objective in a collaborative environment.

## III. SEMANTIC AUGMENTED INFORMATION MANAGEMENT

The usage of TMBLOG [8] has pointed out an innovative layer based on topic maps to enrich multimedia postings with metadata and to extract semantic spatial-temporal semantics from those postings. Topic maps provide a new kind of semantic structure for spatial-temporal postings storage, navigation, and visualization. However, the topic map management including contextual semantic support influences the efficiency of the retrieved information. Let us review the related issues.

### *A. Topic Maps Management Issues*

Two issues need to be considered in order to manage semantic augmented information efficiently in topic maps:

- The automatic creation phase of topic maps according to the documents' structure and content";
- The enhancement of topic maps toward their usages as topic maps are sets of subject proxies;

Table 1: Scenario of information search in a Topic Map

| Context of information search | Fixed parameters | | | Representation | Value range Where u=user,d=document,t=time] |
|---|---|---|---|---|---|
| | User | Document | Time | | |
| (1) Topic map of all users of all documents irrespective of time | | | | $\iiint dUdDdT$ | $[0 \leqslant u < \infty,\ 0 \leqslant d < \infty,\ 0 \leqslant t < \infty]$ |
| (2) Topic map of all users of all documents at a specific time | | | X | $T\iint dUdD$ | $[0 \leqslant u < \infty, 0 \leqslant d < \infty]$ |
| (3) Topic map of all users of a document at all times | | X | | $D\iint dUdT$ | $[0 \leqslant u < \infty, 0 \leqslant t < \infty]$ |
| (4) Topic map of all users of a document at a specific time | | X | X | $DT\int dU$ | $[0 \leqslant u < \infty]$ |
| (5) Topic map on all documents used by a user all the time | X | | | $U\iint dDdT$ | $[0 \leqslant u < \infty, 0 \leqslant d < \infty]$ |
| (6) Topic map on all documents used at a specific time by a user | X | | X | $UT\int dD$ | $[0 \leqslant d < \infty]$ |
| (5) Topic map on a document used by a user any time | X | X | | $UD\int dT$ | $[0 \leqslant t < \infty]$ |
| (6) Topic map by a user of a document at a specific time | X | X | X | UDT | [u=1,d=1,t=1] |

The basis of *enhanced* information retrieval in topic maps is the creator point of view regarding the target documents. Table 1 introduces a possible search scenario associated with the type of information that can be derived in TMs. The scenario assumes that there is no subclass in the three parameters. In essence, we want to show that the analysis of information built on three parameters will necessarily entail inter-parameter, intra-parameter and subclass analysis to improve the search using topic maps. Let us consider the following example wherein a topic map on *D* documents is created in an institution by *N* employees over a period of *T* years. Suppose first that a user produces only one topic map per year from only one document. Over the period of five years (T=5), the user will have created five topic maps and thus the N employees will have 5x*N* topic maps. In contrast, if all the employees can make one topic map on each of 1000 documents per year, there will be (5x1000) x N topic maps after five years. Now, if the topic maps are not restrained by the number of times, by the number of creators, or by the number of documents involved, the situation becomes very complex (NxTxD potential topics maps) particularly when all these parameters values are high. The problem that readily comes to mind is how to manage the large number of topic maps.

Having presented the key issues associated with the creation and management of topic maps, we now present our approach regarding contextual topic maps and our related architecture to handle them.

### *B. Architecture and Contextual Topic Maps Issues*

Our approach consists of integrating several parameters that will reflect not just the organization of information but the content to permit comprehensive access to information in contextual topic maps. The analysis proposed in reference [9] is still possible in the context of TM usage. Figure 1 presents an overview of the interactions of the constituents

involved in topic map creation and subsequent topic–related information searches. The information system consists of documents, topic maps, and semantic databases. Below, we will not discuss the content of these databases but only their relative usage.

The functional architecture is composed of two tasks:

- The creation task
- The topic maps search task.

In the first task, the user creates a topic map from a document. The retrieved document can include topic maps previously created from other documents. The retrieved documents' content becomes the subjects of the user's topic maps. The topics are analyzed by the topic analyzer shown in Figure 1. When this is done, the local topic maps are sent to the topic maps storage with references to the semantic database ["with reference to the semantic database" is grammatical but I'm not sure what it refers to in the sentence. For instance, "When this is done with reference to the semantic database", "the local topic maps with reference to the semantic database", "are sent with reference to the semantic database", or "the topic maps storage with reference to the semantic database."]. The semantic database itself may be a creation from previous topic map sessions. Note that the user's topic maps must fulfill all the contextual parameters (detailed in the following sections). Users must also supply one or more descriptive subjects to each of these parameters.

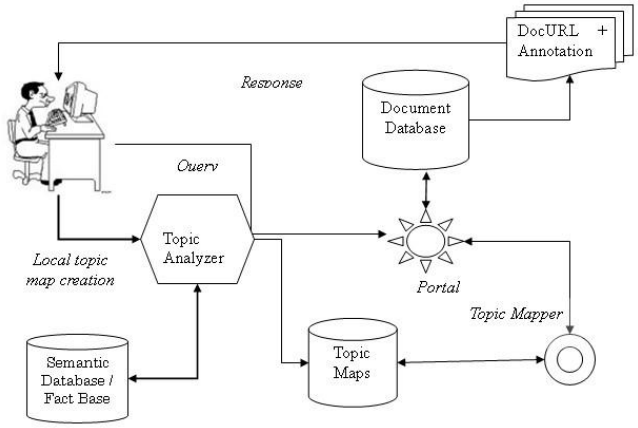


Fig.1. Functional architecture

The topic mapper directly relates the search terms of the users within the topic maps. Users can search using any of the concepts they are familiar with from the list coming from the topic maps.

## IV. CASE STUDY OF TMBLOG

### *A. Resource algebra*

The resource algebra enables various users to share their documents as part of a collaborative platform such as a blog. Resource semantic type and functions in the TMs are directly represented using the appropriate data type and functions supported by the resource algebra. This algebra has two targets. First, it is a semantic interface between scientists who are able to reduce the semantic gap and to strength the metadata bridging them. Second, it facilitates a "collaborative intersection" of scientists using TMs integrating high-level semantics. Let us recall the notion of many sorted algebra [5]. Such algebra consists of several sets of values and a set of operations (functions) between these sets. It consists of two sets of symbols called sorts (e.g. topic, pdf, rtf, lsi_sm) and operators (e.g. tm_transcribe, semantic_similarity); the function sections constitute the signature of the algebra. The second order signature is based on two coupled many-sorted signatures where the top-level signature provides kinds (set of types) as sorts (e.g. DATA, RESOURCE, SEMANTIC_DATA) and type constructors as operators (e.g. set). To illustrate the approach, we assume the following simplified many-sorted algebra:

**Kinds** DATA, RESOURCE, SEMANTIC_DATA, TOPIC_MAPS, SET
**Type** constructor

| | |
|---|---|
| -> DATA | topic |
| -> RESOURCE | pdf, rdt, htm, xml, cvs, jpeg, tiff // resource document type |
| -> SEMANTIC_DATA | lsi_sm, mpeg7_sm, dc_sm, vra_sm, cdwa_sm, ecai_sm, objectid_sm // Semantic and metadata vectors |
| -> TM | tm(topic maps) |
| TM ->SET | set |

**Unary operations**

$\forall$ resource in RESOURCE,
resource → sm: SEMANTIC_DATA,tm **tm_transcribe**
$\forall$ sm in SEMANTIC_DATA sm → set(tm) **semantic_similarity**

The notion sm:SEMANTIC_DATA is to be read as "some type sm in SEMANTIC_DATA," and it means there is a typing mapping associated with the tm_transcribe operator. Each operator determines the result type within the kind of SEMANTIC_DATA depending on the given operand resource types.

Binary operations

$\forall$ tm in TOPIC_MAPS, $(tm)^+ \rightarrow tm$ **topicmaps_merging**

$\forall$ sm in SEMANTIC_DATA , $\forall$ tm in TOPIC_MAPS,
sm,tm tm → tm **semantic_merging**

$\forall$ topic in DATA, $\forall$tm in TOPIC_MAPS,
set(tm) x (topic → bool) → set(tm ) **select**

The semantic merging operation takes two or more operands that are all TM values. The select statement takes an operand type set (tm) and a predicate of the type topic as input and returns a subset of the operand set fulfilling the predicate. From the implementation point of view, the resource algebra is an extensible library package providing a collection of resource data types and operations for domain-oriented resource computation (e.g. cultural field).

The most important concepts in blog management according to [6] are postings access and postings management. Although postings management has been regarded as organizing postings "constituents" (such as information types), we believe that an organization of postings content can facilitate retrieval. Postings searches can be based on one algorithm or the other. The applicability of algorithms used for postings searches depends on the content of the information base or on the organization of the underlying information base.

## V. CONCLUSION

The above proposal can greatly improve the specificity of information creation and information research and thereby improve the access rate. We showed how a topic map can be created bearing in mind its usage in "enhanced information retrieval". We proposed a contextual query methodology based on topic maps. The next step of this work is on combining TMBLOG with a collaborative research project on a semantic tracking platform [15].

## ACKNOWLEDGMENTS

The research presented in this paper would not have been possible without the support and advise of respected professors and colleagues at the National Institute of Informatics (Japan). The authors would like to thank NII for providing the necessary resources to carry out this research.

## REFERENCES

[1] Espiritu C., Stroulia E., and Tirapat T., ENWiC: Visualizing WIKI semantics as Topic Maps: An automated topic discovery and visualization tool, In Proceedings of the 8th International Conference on Enterprise Information Systems 23 - 27, May 2006, Paphos pp. 35-42 URL:http://www.cs.ualberta.ca/~stroulia/ EduNuggets/enwic-iceis2006.pdf

[2] Krötzsch M. Vrandečic D. and Völkel M., Wikipedia and the Semantic Web: The missing Links, Proceedings of Wikimania, Frankfurt, Germany, 2005. url:http://www.aifb.uni-karlsruhe.de/WBS/mak/pub/wikimania.pdf

[3] Güting, R. H., Gral: an extensible relational database system for geometric applications. In Proceedings of the 15th international Conference on Very Large Data Bases (Amsterdam, The Netherlands). Very Large Data Bases. Morgan Kaufmann Publishers, San Francisco, CA, pp. 33-44, 1989.
[4] Haase, K., Context for semantic metadata. In Proceedings of the 12th Annual ACM international Conference on Multimedia (New York, NY, USA, October 10 - 16, 2004). MULTIMEDIA '04. ACM Press, New York, NY, pp. 204-211, 2004.
[5] Jari K., and Stroulia E., EduNuggets: an intelligent environment for managing and delivering multimedia education content, In Proceedings of the 8th international conference on Intelligent user interfaces, Miami, Florida, USA, pp. 303-306, 2003, ISBN:1-58113-586-6, ACM Press.
[6] Le Grand B., and Soto M., Visualisation of the Semantic Web: Topic Maps Visualisation, Proceedings of the Sixth International Conference on Information Visualisation (IV'02), 2002.
[7] Naito, M., and Andres, F., Application Framework Based on Topic Maps, Lecture Notes in Computer Science, Volume 3873, Feb 2006, pp. 42-52, DOI 10.1007/11676904_4, URL http://dx.doi.org/10.1007/11676904_4 Charting the Topic Maps Research and Applications Landscape: First International Workshop on Topic Map Research and Applications, TMRA 2005, Leipzig, Germany, October 6-7, 2005, Revised Selected Papers Editors: Lutz Maicher, Jack Park ISBN: 3-540-32527-1.
[8] Rajbhandari, S., Andres, F., Naito, M., and Wuwongse, V., Semantic-augmented support in Spatial-Temporal Multimedia Blog Management, the International Conference on Topic Maps Research and Applications (TMRA 2006), Leipzig, Germany, October 11-12, 2006, Lecture Notes in Artificial Intelligence, Revised selected papers (LNAI4438), Leveraging the Semantics of Topic Maps, pp. 215-226, ISSN 0302-9743, ISBN 978-3-540-71944-1.
[9] Rath Holger H., The Topic Maps Handbook, Empolis Arvato Knowledge Management, Gütersloh, Germany, 2003
[10] Veltman Kim H., Towards a Semantic Web for Culture, Journal of Digital Information, Volume 4, Issue 4, 2004.
[11] Veltman Kim H., 1995, Electronic Media in the Study of Alberti, Congrès International Leon Battista Alberti, Paris, 1995. http://www.mmi.unimaas.nl/people/Veltman/veltmanarticles/1995%20Electronic%20Media%20in%20the%20Study%20of%20Alberti.pdf
[12] Topic Maps: http://www.topicmaps.org/xtm/
[13] Semantic Wiki Wiki Web: http://www.c2.com/cgi/wiki?SemanticWikiWikiWeb
[14] KendraBase : http://www.kendra.org.uk/wiki/wiki.pl?KendraBase
[15] Kawtrakul, A., Permpool, T., Yingsaeree, C., and Andres, F., A Framework of NLP based Information Tracking and related Knowledge Organizing with Topic Maps, in Z. Kedad et al. (Eds): NLDB2007, LNCS No. 4592 Natural Language Processing and Information Systems, Springer-Verlag, pp. 272-283, ISBN 978-3-540-73350-8, 12th

International Conference on Applications of Natural Language to Information Systems (NLDB 2007), June 27-29, 2007, CNAM, Paris, France.

[16] Kannan R., Andres F., and Guetl C., DanVideo: an MPEG-7 authoring and retrieval system for dance videos, Multimedia Tools and Applications, Springer Netherlands, ISSN 1380-7501, 10.1007/s11042-009-0388-3, pp.545-572, october, 2009.