Overview of Topic Map Construction Approaches

Nebrasse Ellouze\textsuperscript{1}, Mohamed Ben Ahmed\textsuperscript{1}, Elisabeth Métais\textsuperscript{2}

\textsuperscript{1}Ecole Nationale des Sciences de l’Informatique, Laboratoire RIADI
Université de la Manouba, 1010 La Manouba
\textsuperscript{2}Laboratoire Cedric, CNAM
292 rue Saint Martin, 75141 Paris cedex 3, France
nebrasse.ellouze@riadi.rnu.tn
mohamed.benahmed@riadi.rnu.tn
metais@cnam.fr

Abstract

Topic Maps standard (ISO-13250) has been gradually recognized as an emerging standard for information exploration and knowledge organization. One advantage of topic maps is that they enable a user to navigate and access the documents he wants in an organized manner, rather than browsing through hyperlinks that are generally unstructured and often misleading. Nowadays, the topic maps are generally manually constructed by domain experts or users since the functionality and feasibility of automatically generated topic maps still in progress. In this paper, we give an overview of Topic Map building approaches. These approaches take as input different data types: structured documents, structured knowledge, unstructured documents and semi-structured data, and propose different techniques to build a Topic Map such as merging, mapping from RDF to TM and learning techniques. Some other research works are dedicated to cooperative Topic Map building and another research area deals with automatic generation of TM from XML documents.

1. Introduction

Studies have shown that Topic Map \cite{1} authors face major difficulties in constructing topic maps similar to the difficulties associated with ontology construction. For users with no prior experience, generating topic maps manually can be a hard and time consuming task. The major difficulties that they faced were related to ontology building, i.e. to content conceptualization and classification, and to identifying and naming topics and relationships between these topics. Indeed, in most cases, TM authors had to deal with large and complex information systems involving a great diversity of resources, concepts and actors. Thus, TM construction can be very costly and can quickly become a bottleneck in any large-scale application if recourse is not made to automatic or semi-automatic building approaches.

This paper gives an overview of Topic Maps construction proposed for building and generating Topic Maps. In section 2, we give a brief introduction to the Topic Map model. In section 3, we present some related works about TM building. The last section aims to be a conclusion of the state of the art in the TM building area.

2. Topic Maps

Topic Maps are an ISO standard which allows to describe knowledge and to link it to existing information resources. They are intended to enhance navigation in complex and heterogeneous data sets. The key features of Topic Maps are: topics representing entities of the modelized domain, and associations connecting topics and identifying the role played by each topic in the association. Topics can be identified by names and characterized by occurrences (intrinsic properties). Two specific binary associations are standardized for any TM: Class-instance association and superclass-subclass association. The Topic Map standard is like an intelligent extension of the index of books with key features as topics, associations between topics, and occurrences of topics. It is very flexible in merging and extending different sets of Topic Maps. Topic Maps were designed to solve the problem of large quantities of unorganized and heterogeneous information in a way that can be optimized for navigation.
We can observe a growing interest in the use of Topic Maps for modeling and sharing knowledge, and consequently tools and methods have been proposed to ease their management. There are some commercial tools [2] compliant with the ISO 13250 [1] TM standard and with the XML Version 1.0 Specification [3] and available for creating, manipulating and publishing Topic Maps from Ontopia, Infoloom, Empolis and other vendors. There is even an effort to standardize the API used by vendors, called Topic Map API (TMAPI). There are also some open source tools [2], we can classify these tools according to their main functionalities and the support brought to the user, we distinguish three categories: TM engines like TM4J, tinyTIM and XTM4XMLDB used to load TM from XML documents, store TM in databases and modify and access TM, TM navigators such as TMNav and TM editors like TM4L Editor and Viewer.

3. Related Works

Some approaches have been proposed to build, manage and maintain Topic Maps. We notice that, very few Topic Maps have to be developed from scratch, since, in most cases; construction approaches are based on existing data sources and use a priori knowledge (such as domain ontologies, thesaurus) to build and populate the Topic Maps. These approaches take as input different data types: structured documents, databases, unstructured documents and semi-structured data. We notice that there are many sources of topic map data, much of this such as XML documents may be leveraged very efficiently through automated processes, some of it can be mapped directly to topic maps such as RDF metadata. We note also that manual enrichment can add considerable value to the TM building process; in fact, most topic map creation approaches are a combination of auto-generation and manual enrichment [4].

Some other approaches propose to use learning techniques and Natural Language Processing (NLP) techniques to extract topics and associations from textual documents. Learning methods can be applied with different automation levels: manual, semi-automatic, automatic. Some research works are dedicated to cooperative TM building. Another research area deals with merging Topic Maps.

3.1. Mapping from RDF to Topic Maps

Some research work has been proposed to automatically extract Topic Maps constructs by leveraging existing metadata in RDF [5] format [6], [7]. The approach proposed by [7] is a three-step procedure for extracting knowledge from different data sources (relational databases, web sites, enterprise information systems, etc) first in the form of an RDF model and finally as a topic map. The first step consists on recognizing subjects which means find occurrences related to a given subject in the source data. This can be used to extract statements about the subject that can be useful in the knowledge base. The next step is to extract those statements as RDF statements. In fact, an RDF model consists of statements about resources, which are often called triples. A statement has three parts: subject, the resource the statement is about, property, the property being assigned to the subject, and value, the value assigned to the subject. The last step is to map the RDF statements into topics and topic characteristics. In this step, they have to decide whether RDF statements are to be mapped into subject identities, names, occurrences or associations. Finally, since knowledge is extracted from multiple data sources, they propose to create topic maps for the individual data sources and then merge the resulting topic maps.

3.2. Automatic Generation of Topic Maps from XML documents

We can point out some propositions dedicated to automatic generation of TM from XML documents [8], [9], [10], [11]. The approach described in [8] proposes to use XSLT for Topic Map generation over sets of XML resources. The first stage is hand authoring a relatively invariant ontology Topic Map. This consists of defining the ontology of types and associations that capture the data model for a particular subject domain. The second stage is generating additional Topic Maps through an algorithmic process (XSLT) applied to XML document instances. The third is hand authoring those things not captured in the first two stages. This consists of the capture of information not directly discernible from the markup, or stored in non-XML resources. The resultant Topic Maps are merged giving a Topic Map that can be as rich as if completely hand authored. Topic Map merging enables these generated XTM to be combined with topical information that can't be extracted using a style-sheet.

Another approach presented in [9] describes a research prototype of topic maps repository. The main purpose of this repository is to promote knowledge structure sharing and reuse in representing and organizing information resources. Knowledge structure may include existing ontologies, thesauri and existing Topic Maps. As shown in figure 1, the overall architecture of the repository includes wrappers created to convert disperse knowledge structures into an integrated XML schema used in the repository. For that, they incorporate hierarchical enhancement in the
TM structure in order to support hierarchical relationships in thesauri such as BT (Broader Term), NT (Narrower Term), TT (Top Term). Using wrappers, they can convert ISO 2788-based thesauri, RDF-based taxonomies, or XML-based ontologies to the repository. For each new standard, only a new wrapper needs to be defined by introducing more elements in the internal XML schema. In this case, topic maps created through this knowledge repository can be easily export to XTM-based topic maps, and XTM topic maps can also be imported to the repository.

The repository is implemented as a relational database (using MYSQL). Then, they convert data in the database to XML documents using DOM-based servlets. They propose also a web-based authoring tool supported by Java servlets. Through the interface, a user can create topic maps without having to know the syntaxes of topic maps and XML. The user can add or modify topics, occurrences, associations and other related information directly on the web.

![Figure 1. Major components of the knowledge repository](image)

The approach proposed in [10] presents the Topic Map Builder, a processor that extracts topics and relations from instances of a family of XML documents. As shown in figure 2, the TM-Builder is an XSL stylesheet that receives an XML document as input and generates another XML file that contains a Topic Map. The algorithm of the TM-Builder is a three-step algorithm: (1) Initially, for the given ontology creates all the topics types, occurrences roles, occurrences types and associations types; (2) During a document tree traversal, for each association, define the: association type and association members; (3) For each element in the source that is seen as a topic, create the topic ID, topic type, topic names and topic occurrences. This TM-Builder is strongly dependent on the resources structure. So, to extract a topic map for different collections of information resources, they have to implement several TM-Builders, one for each collection. To overcome this inconvenience, they have created an XML abstraction layer for TM-Builders. This layer is composed by one specification in XSTM (XML Specification for Topic Maps), a new XML language that enables to specify topic maps for a class of XML documents.

![Figure 2. The TM-Builder architecture](image)

### 3.3. Co-construction of Topic Maps with Different Users: Collaborative Approaches

The proposed approaches are based on co-construction of Topic Maps with many users [12], [13], [14], [15]. The approach presented in [13] describe an e-learning environment, called BrainBank Learning, a web application for construction of individual topic maps as a mean for learning, where the learners can construct their own learning ontology during a course or a complete study. It works with standard Internet browsers and users enter the application through individual accounts. Topics that the learner meets during education activities are entered and described. The topics can then be connected by linking phrases to form propositions or associations: The learner creates his own associated network of topics that represents his knowledge. This way of documenting in the learning process is good for the learner’s understanding of the area of study (placing knowledge in a context), as well as navigating and overview of the acquired knowledge later on. To further describe topics and associations, digital resources such as documents, pictures, movie clips and sound clips can be attached to the topics. These resources can be either linked to or uploaded to and stored in BrainBank application.

Zaheer and Cahier in [14] extend the TM model and propose the Hypertopic model created by Tech-CICO lab and the Agoré software tool based on this model. It is an all-purpose Knowledge Management approach, that they have called “Socio Semantic Web”, to help communities to formulate, to publish, or to broadcast knowledge, especially scientific knowledge in the field of Human and Social Sciences. The HyperTopic model is a knowledge representation language allowing to build Hypertopic maps using a few basic concepts such as Entity, Point of view, Topic, Association, Resource, Standard attribute. The cooperation model to co-build the map is based the “Knowledge-Based Market Place” (KBM) model with 3 predefined roles (“reader”, “contributor”, “semantic editor”). Through this platform, community members can describe and find domain entities and collections, by designing and browsing “multi-point of view” knowledge maps.
Every contributor may declare the characteristics of an entity following an index structure made of several tree diagrams. Thus, the community would build a dynamic and collective meaning.

3.4. Topic Maps Learning

The approach defined in [16] intends to extract knowledge from web sites to help users find relevant information on the Web. The construction process starts by defining the profile of a TM (and later applying it to Web sites). These profiles characterize Topic Maps and help evaluate their relevance to users' information needs. Second, the analysis identifies topics that have no interest, semantically speaking, which allows to "clean" the TM. That Topic Maps - or Web sites - characterization, filtering and clustering are deduced from the results of a conceptual classification algorithm based on Formal Concept Analysis and Galois connections. TM characterization is based on calculating statistics for every object of the topic map. They propose to compute a weighted mean of these statistics. Each object has a weight which is assigned according to its importance in the topic map. The clustering algorithm consist on grouping topics which share common properties into clusters in order to provide different level of detail (or scales) of the topic map. Figure 3 shows the TM analysis algorithm:

![Figure 3. TM Analysis Algorithm](image)

3.5. Merging Topic Maps

Merging has been defined in the XML Topic Maps specification. A topic map application may use multiple topic maps. Each topic map may emanate from a different source, generated by a different technique or written in a different syntax. Merging takes place on the basis of subject identity or topic names. When two topic maps are merged, topics that represent the same subject should be merged to a single topic and the resulting topic has the union of the characteristics of the two original topics. As we mentioned above, there are other approaches such as [7], [10] and [12] that use the merging technique.

Another approach described in [20] is essentially based on merging technique. It consists on semantic integration of Web-based learning resources using the Topic Maps standard. It is a three steps approach including learning objects representation, enriching learning object with semantics and semantic integration of distributed learning resources. In fact, these learning resources belong to different repositories. Learning resources of a learning repository are represented in Topic Maps knowledge base. Semantics of a learning resource is incorporated in the knowledge base as links to concepts of DAML + OIL ontology. This is made for each learning repository leading to one topic map for each repository. Semantic integration consists to merge all these topic maps into a single and coherent one.

In the next table, we show a summary of the most relevant TM construction approaches.
Table 1. Summary of Topic Maps building approaches

<table>
<thead>
<tr>
<th>Approaches</th>
<th>Main goal</th>
<th>Main techniques used</th>
<th>Reuse other structured knowledge</th>
<th>Input data</th>
<th>Tool associated</th>
<th>Evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gronno [7]</td>
<td>Automatic generation of TM via RDF statements</td>
<td>- Mapping RDF to TM</td>
<td>No</td>
<td>Relational Databases</td>
<td>Information not available in papers</td>
<td>Information not available in papers</td>
</tr>
<tr>
<td>Lin and Qin [9]</td>
<td>Design knowledge repository to integrate and share existing knowledge structures</td>
<td>TM enhancement with hierarchical structure</td>
<td>No</td>
<td>Thesauri Ontologies Topic Maps</td>
<td>XML stylesheets for displaying TM</td>
<td>User (future work)</td>
</tr>
<tr>
<td>Librelotto and al [10]</td>
<td>A new automatic TM extractor from XML documents</td>
<td>-Collaborative build -Learning methods</td>
<td>No</td>
<td>XML documents</td>
<td>TM Extractor developed by the authors</td>
<td>User</td>
</tr>
<tr>
<td>Zaher and al [14]</td>
<td>HyperTopic Model: Collaborate a TM with different actors</td>
<td>-Cooperative build</td>
<td>No</td>
<td>Web pages</td>
<td>Agorae tool</td>
<td>-User -Tests</td>
</tr>
<tr>
<td>Dicheva and Dichev [15]</td>
<td>TM4L: Environment for creating and browsing educational Topic Maps</td>
<td>-Classification (using the ontological structure to index the repository content) -Collaborative build</td>
<td>Domain ontology</td>
<td>Learning repositories</td>
<td>TM4L</td>
<td>-User -Expert -Tests</td>
</tr>
<tr>
<td>Le grand and Soto [16]</td>
<td>Navigation and information retrieval on the Web</td>
<td>-Clustering techniques -Statistical methods</td>
<td>Ontology</td>
<td>Web sites</td>
<td>3D visualisation tool developed by the authors</td>
<td>User</td>
</tr>
<tr>
<td>Folch and Habert [17]</td>
<td>Semantic acquisition and structuring of distributed and dynamic resources</td>
<td>NLP techniques</td>
<td>No</td>
<td>Domain texts</td>
<td>Zellig ALCESTE text mining tools</td>
<td>User</td>
</tr>
<tr>
<td>Ouziri [20]</td>
<td>Semantic integration of web-based learning resources</td>
<td>-TM Enrichment -TM Merging</td>
<td>Ontology</td>
<td>Learning repositories</td>
<td>Information not available in papers</td>
<td>User</td>
</tr>
</tbody>
</table>

4. Discussion and Conclusions

This paper presents an overview of the most relevant Topic Maps construction approaches taken as input different sources like text, structured data, document metadata etc. Based on the state of the art, we notice that there are many techniques used to generate Topic Maps from different sources. Some of these approaches take as input XML documents and propose to apply automated processes to leverage these documents; other approaches propose to map directly RDF metadata to topic maps. We note also that collaborative construction involving different users are very appropriate to the TM building process. In addition, we notice that most topic map construction approaches are a combination of auto-generation enrichment techniques and merging process.

In all proposed approaches for TM construction, we notice that automated Topic Map construction is not
enough developed. Indeed, the construction of a TM can be very costly and can quickly become a bottleneck in any large-scale application if recourse is not made to automatic methods. Problems of maintenance and coherence may arise when the TM is applied to heterogeneous and multilingual information sources, as manual construction and maintenance can not keep pace with any incoming documents.

Also, the main lacks for all the approaches presented in this paper are that they are dedicated to a specific domain and there are not approaches or techniques for building Topic Maps from multilingual resources. In fact, the multilingual aspect is not landed in all existing approaches which support a monolingual environment. The problem is that the large majority of resources available today are written in various languages, and these resources are relevant even to non-native speakers. In this case, it is difficult for users to find relevant information in documents using their native language. At least, we note that the majority of the TM construction approaches do not propose techniques for the evaluation of the resulting Topic Map.

5. References


