An Ontology-Based System for the Marketing Information Management

Eduard BARBU
Center for Mind/Brain Sciences, University of Trento, Italy
E-mail: eduard_barbu@yahoo.com

Abstract. In the new globalisation context the competition for expanding to new markets becomes sharper for Small and Medium Sized Enterprises. We answer their information needs by building the AMI-SME system. The AMI-SME system is an ontology-based system for the management of marketing information. In the context of this system we show how specially built domain ontologies aid the organization of Small and Medium Sized Enterprises information in a meaningful way. Based on our experience in building the main ontology of the system we argue that the untrained user cannot commit to a formal ontology.

Introduction

In the new globalisation context the competition for expanding to new markets becomes sharper for Small and Medium Sized Enterprises (SME). Lacking financial resources they cannot afford to hire consultant companies for making detailed market studies. Nevertheless, competitiveness requires to access reliable and detailed information about the target market before competitors.

There are many factors that play an important role in SME decision making process: the strength of the competitors already present in the market, the buying power of the potential customers, the legislation of the target market, etc.

The aim of the craft European Project Analysis of Marketing Information for Small and Medium Sized Enterprises (AMI-SME) is to develop a software solution that supports SME in their internalization process. Currently, for gathering information about the potential markets SME relay on Web Searching Engines like Google and Yahoo. Even if the storage and indexing power of these web searching engines is impressive they lack any support for managing the information relevant to the user. The AMI-SME system, a flexible meta-searching engine coupled with a series of useful components for information extraction, overcomes these limitations using a library of ontologies.

We show how specially built marketing ontologies help marketing experts and naïve users alike to organize their information in a meaningful way.

This paper has the following organization: in the first part, an overview of the AMI-SME system will be given. In the second one the process of building the General Marketing Ontology, the main system ontology, is presented. The third part outlines the functions that ontologies have in the system. Based on the experience of the users of the AMI-SME

1 We use the term ontology interchangeably to refer to the terminological box or to the terminological and assertional box. The first usage comes from Knowledge Representation tradition and is the proper usage. The second comes from W3C OWL recommendations.
system we argue in the last part of the paper that the untrained user will never really commit to formal ontologies. The inability of the untrained user to commit to a formal ontology is a problem that the future Semantic Web should face.

1 An Overview of the System

The AMI-SME system is a flexible meta-searching engine coupled with a series of useful components for information extraction. A higher-level schema of the system is presented in the figure 1.

![Figure 1: A high level schema of the AMI-SME system.](image)

This schema is meant to give the reader an overview of our system. It is not intended to show all system components and their interaction.

A central concept of the system is that of information source. An information source is any data repository or any software that exposes an interface for querying a data repository. Examples of information source are: web search engines, database systems, the user file system, etc. At the moment of writing this paper the system supported information sources were: the web searching engines Google, Yahoo and A9, the local file system and the Internal Document Database.

Unlike the queries formulated in the web searching engines, AMI-SME queries are defined inside projects. A project can be thought of as a high level context that binds together the user query, his/her personal ontology and the documents he/she is interested in. In the system any project is identified by its unique name. To take as example a company that wants to sell computers to France, a good name for the project related to this activity would be “Computer Selling in France”. The most successful queries related to this project could be saved for further run.
After the query is executed over all information sources in the system a ranked list of documents is returned as answer. One can select the documents he/she is interested in for future processing. After the selection is performed the system triggers a series of transformations over the document content. Therefore the document is automatically summarized, interesting meta-data as title and author are extracted, NER (Name Entity Recognition) performed, etc. All the ontologies in the system are kept in the Ontology Library. The user loads his/her personal ontology and annotates his/her documents with concepts and instances from the ontology. The interesting documents can be saved in the Internal Document Database. The document metadata like the user comments, the automatically generated summary, will be saved in the Document Annotation database. The Ontology Library, the Document Annotation Database and the Internal Document Database are linked such that all the information related to a document could be retrieved easily.

2 The General Marketing Ontology

This section will present the main ontology built in the AMI-SME project. The ontology is called General Marketing Ontology (GMO) and formalizes the marketing domain for a company that wants to internationalize. To build GMO a three-stage methodology was followed. In the first stage the domain and the purpose of the ontology were made precise. In the second one the main concepts in the domain were discovered using relevant domain resources such as technical dictionaries and tools for semi-automatic building of ontologies. In the third stage the language for representing the ontology has been chosen and the ontology effectively modeled. We detail below each of the above mentioned steps.

2.1 Domain and purpose of the ontology.

The intuition behind GMO is the following: any company that wants to internationalize needs a detailed perspective of the target market. The potential competitors should be discovered, the strength and the weaknesses of the products they sell should be known, and the buying power of the potential customers should be taken into consideration. Once all the details about the target market are known the company could start developing its business plan. GMO contains concepts like “Market”, “Competitor”, “Product”, etc. GMO will be instantiated in different ways by different companies. A company that sells operating systems will add as instances of the concept “Competitor” its rival companies that are different from the rival companies of a company that sells cheese. Likewise the instances of the concept “Product” will be different for two companies that do not sell the same products.

The purpose of the ontology becomes clear if we think that GMO is used in the context of classical searching engines. There are three main usages that ontologies have in our system:

1. They assist the user in better formulating his/her query.
2. They are used for labeling the user documents with relevant concepts and instances.
3. They are used in NER and summarization tasks.

The usage of ontologies is detailed in the corresponding section.

2.2 Concept identification
The concept identification is the phase where the main concepts and their relationships for the domain to be modeled are identified. For achieving this for the marketing domain we used the following sources:

1. Philip Kotler’s well known book *Marketing Management* [5]. This is an authority book and its reading is more than beneficial for any attempt to formalize the marketing domain. From this book we “borrowed” the top level vision on the market. For example, the intended meaning of the “market” as the community of buyers as opposed to the outdated definition of the market as place where buyers and sellers were gathered is taken from here. A central part of our ontology is constituted by Porter’s five forces. Michael Porter [8] developed a framework that models the industry as being influenced by five forces: Supplier Power, Barriers to Entry, Threat of Substitutes, Buyer Power, and Rivalry. The intelligent manager seeking to defeat its competitors can use Porter’s model for understanding the marketing context.

2. The second source for the relevant concepts in our ontology is a set of marketing dictionaries found on the web. These dictionaries were used as a corpus for TextToOnto [6] a tool for semi-automatic concept acquisition. TextToOnto combines different methods for learning ontology from text. These methods include: conceptual clustering and formal concept analysis, sub-categorization frames for learning arbitrary relations, and rule mining techniques for learning association between terms. TextToOnto found some useful concepts that helped filling the gaps in our ontology.

3. The third source was the competency questions formulated by the potential users of the system. The competency questions were analyzed and the concepts that could be used for answering these questions were identified.

4. For enriching the ontology with synonymous terms we used Princeton WordNet version 2.0. Ontoling Plug-in, a plug-in for Protégé, allows for the extension of the ontology with definitions and terms from Princeton WordNet and similar lexical resources. Being a resource for general language, Princeton WordNet is not so useful for building domain specific ontologies. Its definitions are many times loose and its distinctions, even if relevant from a lexical semantics perspective, are not justified from an ontological point of view. Despite these problems, Princeton WordNet provided us some synonymous terms difficult to find otherwise.

At the end of this stage the ontology had around 300 concepts taxonomically arranged and 150 relations and attributes. Because the users of our system use the ontology directly we demanded their feedback. They found the ontology difficult to understand and required a lighter version of it. In the last section of our paper the reasons of their decision and its possible consequences for the future Semantic Web will be presented. After the interaction with the users we reduced the number of concepts in the ontology to 100 and the number of relations and attributes to 50. For modeling the ontology we used both SemTalk and Protégé tools.

2.3 The language for representing the ontology

In choosing the adequate language for expressing the ontology we confronted two different and contradictory requirements. On the one hand we wanted the built ontology to be reusable on the Semantic Web. This means that we should express the ontology in RDFS or OWL. OWL is a much better language for representing ontologies than RDFS, so we were biased toward it. To be more precise, we wanted to stick to OWL-DL the complete and decidable fragment of First Order Logic. On the other hand, we wanted to satisfy the
requirements of the clients of our system who wanted an easily understandable ontology. They do not want to see constructs like disjoint or enumerated classes, equivalence classes or properties, hierarchies of properties or cardinality restrictions. To reconcile these two points of view we opted for supporting a limited set of constructs from OWL. This set of constructs should be sufficiently rich to express all the statements that an AMI-SME user wants to express but not more. The subsystem that deals with ontology handling incorporates the functions of a mini ontology editor. It supports the following functions:

1. State that a concept is in the IS-A relation with another concept (for example state that “Administrative Area” IS-A “Area”). In this way new concepts can be added to the ontology. Because the addition of concepts to the ontology is not a task for the untrained user, the concept “UserLabel” has been added to GMO. All user concepts are added as subclasses of the concept “UserLabel”. In this way we prevent the logic of the ontology to be broken.

2. State that two concepts are related via a certain relation (e.g.: the concept “Product” is related via the relation “hasSeller” with the concept “Seller”).

3. State that a certain concept has a certain attribute (e.g.: the concept “Product” has the attribute “price”).

4. Provide alternative notations for a certain concept. (In English the notation for the concept “Product” is “Product”, in Italian is “Prodotto”.) In this way GMO becomes a multilingual ontology. The non-English user has only to add translations of the concepts in his/her native language.

5. Add synonyms for the concepts in the ontology.

6. State that a concept has a certain instance (“IBM” is an instance of the concept “Company”).

7. Add statements about instances (“Dell” competes with “Apple” or Dell hasBudget “$500,000,000”).

8. Add comments in the chosen language to concepts, instances, attributes and relations.

3 The role of ontologies in the system

Currently AMI-SME Ontology Library comprises GMO and other ontologies which describe the products sold by two industry partners. We had to choose between modeling the information specific to products in the GMO or outside it. If we had decided to add the description of particular products in GMO, the generality of the ontology would have been lost. The AMI-SME system is not restricted to marketing. For changing the domain, specific ontologies could be loaded in the Ontology Library. In this way the AMI-SME system can be easily adapted to physics, culinary art, travel, biomedicine, etc.

In the AMI-SME system the ontologies have the following roles:

1. The annotation role. The concepts and instances in the ontologies are used for document annotation. The annotation process takes profit of the taxonomic part of the ontology. Thus, the annotations explicitly provided by the user are easily extended. For example if a user annotates a certain document with the label Microsoft, the system will look in his/her personal ontology and see that “Microsoft” is an instance of the concept “Competitor” and the concept “Competitor” IS-A “Company”. It can thus infer that the implicit annotation of the found document is “Microsoft->Competitor->Company”. After the documents are annotated in this way and saved in the Internal Document Database they can be viewed by browsing the ontology. The user labeling of documents is used as input for training a Naïve Bayesian classifier one of the most effective algorithm for text
classification [7]. For properly training the Naïve Bayesian classifier we made the assumption, correct in our system, that the instances in the ontology which are the categories of the Naïve Bayesian Classifier are distinct. After the Naïve Bayesian classifier is learnt, the documents newly brought in the system are automatically classified and their annotation automatically extended as specified above.

2. The NER and summarization role. The NER sub-system in AMI-SME has two parts: the first part is a Gazetteer Based Name Entity Recognizer. The gazetteers are files that define a set of related items. We have gazetteers for: countries, companies, cities, etc. These files together with the gazetteer definition file reside in a directory inside the AMI-SME system. The definition file assigns a name to each new gazetteer added to the system. The entry: “company.lst: company” is an example of an entry in the definition file. It assigns to the gazetteer “company.lst” the name ‘company’. The gazetteers are loaded and run over document contents. If the items from the gazetteers are found the documents that contain them are annotated with the combination “gazetteer name, gazetteer item”. For example, if the item “Microsoft” from the gazetteer “Company” is found in a certain document, then the document is automatically annotated with the label “Company, Microsoft”.

The second part of the NER subsystem is based on the user’s personal ontology. The ontology instances are searched in the documents. In GMO there is an entry Named “GazetteerConcept”. Whenever an item from a gazetteer is found in a document, the name of the gazetteer is added as subconcept of “GazetteerConcept” and the found items are added as its instances. Next time the user can retrieve these documents by ontology browsing. There is no guarantee that the intended meaning of the concepts is the same as that of the gazetteer name. Thus, it would be wrong to match the concept names with the gazetteer names and add the items of the gazetteer as instances of the concepts. The ontology concepts and instances are used in the summarization task. The summarization component of the AMI-SME system is a language independent sentence extractor. From the point of view of the genre the summary produced by the AMI-SME system is a generic summary which favors the virtual important information for the user. We will not describe our summarization system here because this is beyond the purpose of our paper. We stress only that it has four components. Based on different criteria each component assigns scores to the sentences of the document to be summarized. One of these components assigns higher scores to sentences that contain concepts and instances from ontology. All the scores computed by the four components are combined and the sentences that have highest scores are selected in the summary.

3. Assisting the user to formulate his/her query role. Because the ontology is bound to a certain project it reflects a particular user interest. In the fictive project “Computer Selling in France” the ontology loaded in the project will reflect the user’s interest related to selling computers in France. The ontology will presumably contain the competitors of the user company in France; the types of computers the user company is selling, etc. After the ontology will grow enough it can act as a repository of potential useful keywords for query building. The user query can be formulated by ontology browsing and concept/instance selection.

4 Semantic Web and formal ontologies
The frustrating experience the users have using web search engines like Google or Yahoo was one of the reasons for proposing the new Semantic Web. The documents returned by these search engines, it is said, have in many cases nothing to do with the user’s information needs. The cause of this behaviour is identified in the lack of common semantics between men and machines. The current markup of the web pages is meaningful for human agents but meaningless for the machines. To bridge the gap between men and machines it is proposed that the semantics of the future web to be specified by formal ontologies (expressed in languages like OWL). The web documents will be annotated with the concepts from ontologies and all assertion on the Semantic Web will follow the RDF data model. Because all agents, humans and machines alike, have a common vocabulary and the languages they communicate with each other will be unambiguous, the problems like bad results when searching the web will disappear.

The contention that the search on the Semantic Web will deliver much better results than the search on the current web remains to be proved. Semantic Web technologies were used successfully for small collection of documents. In all cases the documents in the system were semi-automatically or even fully manually annotated [3]. Despite the recent progress in Information Extraction illustrated by systems based on GATE [2] like KIM [4], the annotation task involves a lot of effort. It is certain that the present annotation technology is not scalable to the size of the web.

We want to point another problem that Semantic Web would face: the untrained user does not understand and therefore cannot commit to formal ontologies.

According to the Semantic Web proponents [1] high quality formal ontologies will be essential for the success of the Semantic Web. A high quality formal ontology is an ontology that passes the essential reasoning tests. In the design phase reasoning ensures the consistency of the ontology and computes the concept hierarchy. In this phase the inconsistent assertion should be resolved and the ontology builder must check that the derived taxonomy matches his/her intuition. In the usage phase the reasoning process checks the consistency of the assertions in those web documents that are annotated with the ontology vocabulary. Therefore a high quality ontology should be a consistent ontology that matches the user intuition and formal enough for being “understood” by the computer.

Formal ontologies reflect the intuition of the ontologist, a trained individual that masters the language of logic. Unfortunately it does not mirror the intuition of the untrained users of ontologies. The inability of untrained individuals to understand the meaning of some logical constructs present in the OWL language has been pointed before [9]. We want to underline a more serious problem: the lack of understanding form the part of average computer users of some basic concepts related to ontology. The users of our system, all of them experienced in web searching, had difficulties in understanding a light ontology like GMO. The most difficult features to understand were:

- The semantics of the IS-A relation. For them the IS-A relation is a relation that links highly similar concepts. When our users add concepts to the ontology they tend to add them in total disagreement with the semantic of the IS-A relation. For example, most of the users added the concept “MarketExpert” under the concept “Market” and not under the concept “Person”.
- The semantics of the relations themselves. They wanted to use the same relation names over and over even if the ontological meaning of the relation was different. They tend to use the same name when in the common language the relation was lexicalized with the same name. For example the concept “Company” was linked by the relation “has” with the concepts “Manager” and “Competitor”.
The distinction between concepts and instances. They cannot properly distinguish concepts from instances. For most of them the sub-concepts of a certain concept are the instances of the respective concept.

The distinction between attribute, relation and concept is hard to grasp. For example “color” should have been a property of a certain product in one of our product ontologies. When confronted to the problem of adding the label “color” to the ontology the users added it as a concept.

It should be clear enough from these remarks that an untrained user does not understand a formal ontology. For him/her an ontology is a collection of labels loosely coupled. Their notion of ontology resembles more the psycholinguist’s semantic memory than to formal ontologies. Without training, the average user will never really commit to a formal ontology. Or if he/she does, his/her commitment will be only a superficial one and the communication with the machine agents will fail. The extensive focus on the machine agents losing from sight the human agents will be a mistake that the Semantic Web community should avoid. The consequence of this mistake will be the inability of formal ontologies to bridge the communication gap between humans and machines.

As a positive note I would add that the users of our system where very interested in ontologies. As pointed before they wanted to add new concepts and instances to ontologies even if in most cases they failed to model them properly.

5 Conclusion

Up to now the information related to market obtained by Small and Medium Sized Enterprises was hard to manage. To overcome this obstacle we developed the AMI-SME system, a meta-searching engine coupled with a series of useful components for information extraction.

The system uses a library of ontologies to manage the information related to market. The core of the library of ontologies is the General Marketing Ontology, an ontology that formalizes the marketing domain for a company that wants to internationalize. The users of AMI-SME system have the possibility to personalize the light ontologies provided by the experts, so enhancing the capability of the system. Our users were enthusiastic about using the ontologies even if they did not fully understand them. A lesson learnt is that the average user has difficulties understanding even light ontologies like those used by the AMI-SME system. Unless the scientific community copes with this problem, the Semantic Web idea will fail.

The AMI-SME system features were agreed with the project partners from industry. Being a system that should fulfill the purposes of the industry partners, the developers paid distinguished attention to industry partners’ requirements. Every time the developers finalized the implementation of a feature, industry partners tested the system. Because the system is still in the testing phase we did not perform a comparative testing of the AMI-SME system with other similar systems. However the users of AMI-SME used delivered prototypes for solving real problems and gave us their feedback, which is encouraging. Soon before the end of the project a more thorough evaluation of the system will be performed. There will be tested features like usability, speed, relevance, etc. The objective of this final testing phase is to perform a benchmarking of the system relative to similar systems.

The AMI-SME architecture allows for an easy integration of other information management components and allows for the easy extension of the library of ontologies. In the future, taking advantage of the component-oriented architecture of the system we will enhance the system by integrating other components and ontologies.
6 Acknowledgments

The work presented here has been realized during the European funded project Analysis of Marketing Information for Small and Medium Sized Enterprises. I want to thank all project partners for their efforts in the project and for the helpful discussions during our meetings.

References:


