

Ontology Research and Development

Part 1 – A Review of Ontology Generation

Ying Ding

Division of Mathematics and Computer Science
Vrije Universiteit, Amsterdam
(ying@cs.vu.nl)

Schubert Foo

Division of Information Studies, School of Computer Engineering
Nanyang Technological University, Singapore
assfoo@ntu.edu.sg

Abstract

Ontology is an important emerging discipline that has the huge potential to improve information organization, management and understanding. It has a crucial role to play in enabling content-based access, interoperability, communications, and providing qualitatively new levels of services on the next generation of Web transformation in the form of the Semantic Web.

The issues pertaining to ontology generation, mapping and maintenance are critical key areas that need to be understood and addressed. This timely survey is presented in two parts. This first part reviews the state-of-the-art techniques and work done on semi-automatic and automatic ontology generation, as well as the problems facing these researches. The second complimentary survey is dedicated to ontology mapping and ontology evolving.

Through this survey, we identified that shallow information extraction and natural language processing techniques are deployed to extract concepts or classes from free-text or semi-structured data. However, relation extraction is a very complex and difficult issue to resolve and it has turned out to be the main impedance to ontology learning and applicability. Further researches are encouraged to find appropriate and efficient ways to detect or identify relations through semi-automatic automatic means.

Keywords

ontology generation, ontology, knowledge representation

1. Introduction

Ontology is the term referring to the shared understanding of some domains of interest, which is often conceived as a set of classes (concepts), relations, functions, axioms and instances (Gruber, 1993). Now in knowledge representation community, the commonly used or highly cited ontology definition is from Gruber (1993): “an ontology is a formal, explicit specification of a shared conceptualization. ‘*Conceptualization*’ refers to an abstract model of phenomena in the world by having identified the relevant concepts of those phenomena. ‘*Explicit*’ means that the type of concepts used, and the constraints on their use are explicitly defined. ‘*Formal*’ refers to the fact that the ontology should be machine readable. ‘*Shared*’ reflects that ontology should capture consensual knowledge accepted by the communities”.

Ontology is a complex multi-disciplinary field that draws upon the knowledge of information organization, natural language processing, information extraction, artificial intelligence, knowledge representation and acquisition. Ontology is gaining popularity and is touted as an emerging technology that has a huge potential to improve information organization, management and understanding. In particular, ontology as the shared and common understanding of the domain that can be communicated between people and application systems, has a significant impact on areas dealing with vast amounts of distributed and heterogeneous computer-based information, such as World Wide Web and Intranet information systems, complex industrial software applications, knowledge management, electronic commerce and ebusiness. For instance, ontology plays the strategic role for agent communication (Hwang, 1999); ontology mapping is the capable way to break the bottleneck of B2B marketplace (Fensel et al., to appear) and ontology is the enabler to improve the intranet knowledge management systems (Kietz, Maedche and Volz, 2000). Ontology itself is an explicitly defined reference model of application domains with the purpose of improving information consistency and reusability, systems interoperability, and knowledge sharing. It describes the semantics of a domain in both a human-understandable and computer-processable way.

The development at the World Wide Web Consortium (W3C) indicates that the first generation of World Wide Web will make its transition in future into the second generation of Semantic Web (Berners-Lee, Hendler and Lassila, 2001; Fensel et al., 2001). The term “Semantic Web” is coined by Tim Berners-Lee, the inventor of the World Wide Web, to describe his vision of the next generation Web that provides much more automated services based on machine-processable semantics of data and heuristics (Berners-Lee and Fischetti, 1999). Ontologies, that provide shared and common domain theories, will be a key asset for this to happen. They can be seen as metadata that explicitly represent semantics of data in a machine-processable way. Ontology-based reasoning services can operationalize such semantics and be used for providing various forms of services (for instance, consistency checking, subsumption reasoning, query answering, and so on). Ontologies help people and computers to access the information they need, and effectively communicate with each other. They therefore have a crucial

role to play in enabling content-based access, interoperability, and communication across the Web, providing it with a qualitatively new level of service: the Semantic Web (Fensel, 2001).

Ontology learning is starting to emerge as a sub-area of ontology engineering due to the rapid increase of web documents and the advanced techniques shared by the information retrieval, machine learning, natural language processing and artificial intelligence communities. The majority of existing ontologies have been generated manually. Generating ontologies in this manner has been the normal approach undertaken by most ontology engineers. However, this process is very time-intensive, error-prone, and poses problems in maintaining and updating ontologies. For this reason, researchers are looking for other alternatives to generating ontologies in a more efficient and effective way. This survey aims to provide an insight into this important emerging field of ontology, and highlights the main contributions of ontology generation, mapping and evolving¹ whose inter-relationships are shown in Figure 1.

The survey is carried out over two parts, namely, the state-of-the-art survey on ontology generation and state-of-the-art survey on ontology mapping and evolving. In this first part of the survey on ontology generation, the areas of semi-automatic or automatic ontology generation will be covered. A subsequent paper will report on the ontology mapping and evolving.

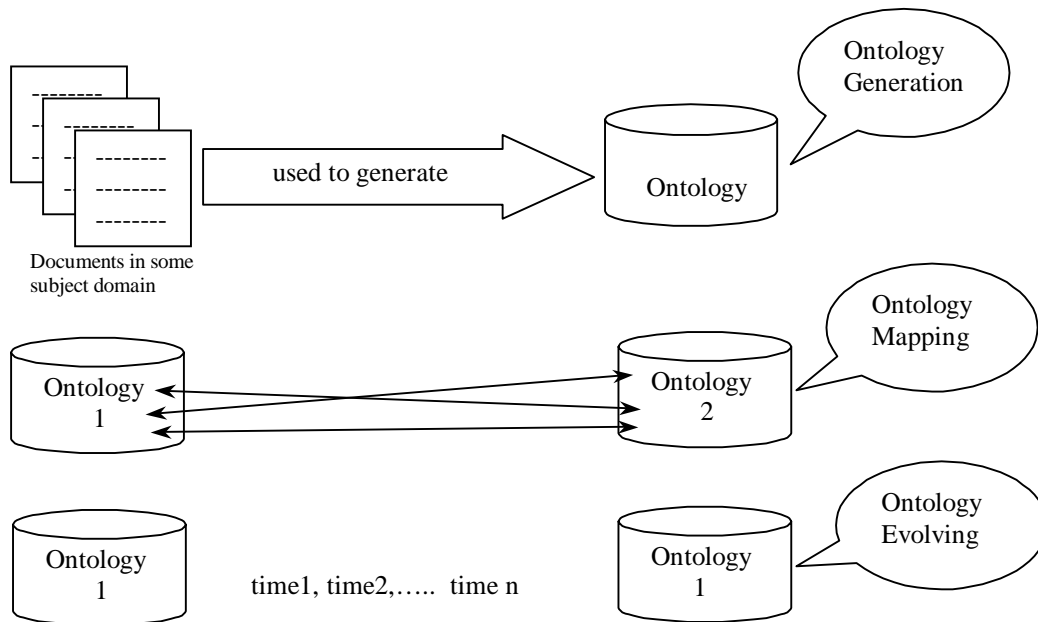


Figure 1. General overview of ontology generation, mapping and evolving

¹ Ontology evolving, although an incorrect use of English language, has become an accepted buzzword in the ontology field, as in ontology mapping and ontology versioning).

2. Ontology generation in general

Although there already exist large-scale ontologies, ontology engineers are still needed to construct the ontology and knowledge base for a particular task or domain, and to maintain and update the ontology to keep it relevant and up-to-date. Manually constructed ontologies are time-consuming, labor-intensive and error-prone. Moreover, a significant delay of updating ontologies causing currency problems actually hinders the development and application of the ontologies.

The starting point for creating an ontology could arise from different situations. An ontology can be created from scratch, from existing ontologies (whether global or local ontologies) only, from a corpus of information sources only; or a combination of the latter two approaches (Uschold, 2000). Various degrees of automation could be used to build ontologies, ranging from fully manual, semi-automated, to fully automated. At present, the fully automated method only functions well for very lightweight ontology under very limited circumstances.

Normally, methods to generate ontology could be summarized as bottom-up: from specification to generalization; top-down: from generalization to specification (e.g. KACTUS ontology); and middle-out: from the most important concepts to generalization and specialization (e.g. Enterprise ontology and Methodology ontology) (Fernandez-Lopez, 1999). Most often, lifting algorithms are used to lift and derive different levels of ontologies from a basic ontology (McCarthy, 1993).

There are also a number of general ontology design principles that are proposed by different ontology engineers over a period of time:

- Guarino (1998) was inspired by philosophical research and proposed a methodology for ontology design known as “Formal Ontology” (Cocchiarella, 1991). This design principle contains theory of parts, theory of wholes, theory of identity, theory of dependence and theory of universals. He summarized the basic design principles that include the need to (1) be clear about the domain; (2) take identity seriously; (3) isolate a basic taxonomic structure; and (4) identify roles explicitly.
- Uschold & Gruninger (1996) proposed a skeletal methodology for building ontologies via a purely manual process: (1) identify purpose and scope; (2) build the ontology via a 3–step process: *ontology capture* (identification of the key concepts and relationships and provision of the definitions of such concepts and relationships); *ontology coding* (committing to the basic terms for ontology (class, entity, relation); choosing a representation language; writing the code); *integrating existing ontologies*; (3) evaluation (see Gomez-Perez etc., (1995)); (4) documentation; (5) guidelines for each of the previous phases. The final resulting ontology should be clear (definitions should be maximally clear and unambiguous), consistent and coherent (an ontology should be internally and

externally consistent), extensible and reusable (an ontology should be designed in such a way as to maximize subsequent reuse and extensibility).

- Ontological Design Patterns (ODPs) (Reich, 1999) were used to abstract and identify ontological design structures, terms, larger expressions and semantic contexts. These techniques can separate the construction and definition of complex expressions from its representation to change them independently. This method was successfully applied in the integration of molecular biological information (Reich, 1999).

Hwang (1999) proposed a number of desirable criteria for the final generated ontology to be (1) open and dynamic (both algorithmically and structurally for easy construction and modification), (2) scalable and interoperable, (3) easily maintained (ontology should have a simple, clean structure as well as being modular), and (4) context independent.

The remaining sections highlights the major contributions and projects that have been reported with respect to ontology generation. In each project, an introduction and background is first provided. This is followed by a description of the methods employed and concluded with a summary of problems that have surfaced during the process of ontology generation.

2.1 InfoSleuth (MCC)

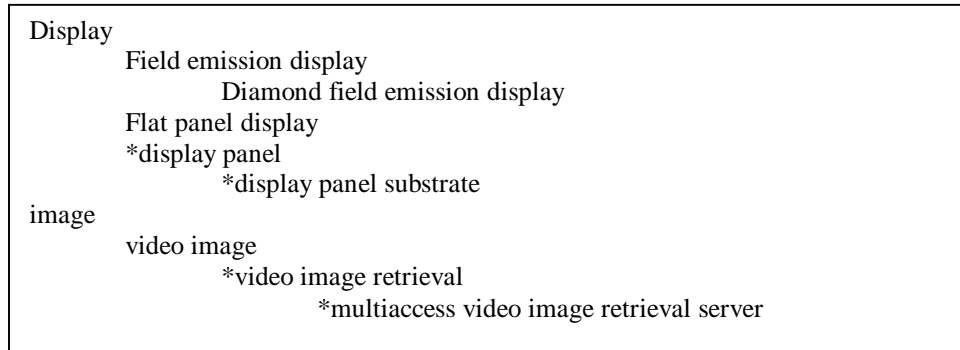
InfoSleuth is a research project at MCC (Microelectronics and Computer Technology Corporation) to develop and deploy new technologies for finding information available both in corporate networks and external networks. It focuses on the problems of locating, evaluating, retrieving, and merging information in an environment in which new information sources are constantly being added. It is the project aiming to build up the ontology-based agent architecture. It has been successfully implemented in different application areas that include Knowledge Management, Business Intelligence, Logistics, Crisis Management, Genome Mapping, environmental data exchange network and so on.

Method

The procedure for automatic generation of the ontology adopted in InfoSleuth is as follows (Hwang, 1999):

- Human experts provide the system with a small number of seedwords that represent high-level concepts. Relevant documents will be collected from the web (with POS-tagged or otherwise unmarked text) automatically.
- The system processes the incoming documents, extracts phrases containing seedwords, generates corresponding concept terms and places them in the 'right' place in the ontology. At the same time, it also collects candidates for seedwords for the next round of processing. The iteration continues until some satisfied results are reached.

- Several kinds of relations are extracted: “is-a”, “part-of”, “manufactured-by”, “owned-by”, etc. The “assoc-with” relation is used to define all relations that are not an “is-a” relation. The distinction between “is-a” and “assoc-with” relations is based on a linguistic property of noun compounds.
- For each iteration, a human expert is consulted to ascertain the correctness of the concepts. If necessary, the expert has the right to make the correction and reconstruct the ontology. Figure 2 shows an example of the structure of the generated ontology:



Notes: * shows the add the extra concept to the end of the upper level concept

Figure 2: Example of the automatic generated ontology from InfoSleuth

In Figure 2, the indentation shows the hierarchy (class and subclass relationships). Thus, ‘field emission display’, ‘flat panel display’ and ‘display panel’ are subclasses of ‘display’. Here one obvious rule to generate this hierarchy is that if the phrase has ‘display’ as the last word in the phrase, this phrase will become the subclass of ‘display’. Likewise, the same rule is applied for ‘image’. Another rule is that if the phrase has ‘display’ as the first word in the phrase, this phrase will also become the subclass of the ‘display’ with the indication of ‘*’, such as ‘display panel’, ‘video image retrieval’ and so on.

This system has a number of special features and characteristics:

- Discover-and-alert: The system expands the ontology with new concepts it learns from the new documents and alerts the human experts of the changes.
- Attribute-relation-discovery: This approach can discover some of the attributes associated with certain concepts. For instance, the method can discover the attributes of physical dimensions or number of pixels and even can learn the range of their possible values. Based on the linguistic characters, “assoc-with” relation can be identified automatically. Since the ontology is organized as hierarchies, attributes are automatically inherited following “is-a” links.
- Indexing-documents: While constructing the ontology, this method also indexes documents for future retrieval, optionally saving the results in a relational database. It collects “context lines” for each concept generated, showing how the concept was used in the text, as well as frequency and co-occurrence statistics for word association discovery and data mining.

- This system allows the users to decide between the precision and completeness through browsing the ontology and inferences based on the is-a relation and associated relations.

The system uses a simple part-of-speech (POS) tagging to conduct superficial syntactic analysis (shallow information extraction techniques). The relationship of the concepts is detected based on the linguistic features. As in any other corpus-based approach, the richer and the more complete data set, the higher will be reliability of the results achieved as a direct result of the applicability of machine learning techniques.

Problem encountered

Generating automatic ontologies in this manner is plagued with several challenges and problems such as:

- Syntactic structural ambiguity: A correct structural analysis of a phrase is important because the decision whether to regard a certain subsequence of a phrase as concept often depends on the syntactic structure. For example, the concept “image processing software” when extracted sequentially becomes image (wrong), image processing (wrong), image processing software (right concept). Likewise, “low-temperature polysilicon TFT panel” becomes “low-temperature” (wrong), “low-temperature polysilicon” (wrong), “low-temperature polysilicon TFT panel” (right concept).
- Recognize different phrases that refer to the same concept. For example, “quartz crystal oscillator” is actually the same as “crystal quartz oscillator”. One possible solution could be to differentiate them via the co-occurrence frequency of these two phrases.
- Proper attachment of adjective modifiers is a possible way to avoid creating non-concepts.
- Word sense problem: A possible way to solve this problem is to make reference to some general linguistic ontologies (such as SENSUS or WordNet (Miller, 1995)) so as to disambiguate different word senses (using human-involvement to select the word sense from SENSUS or WordNet, or through machine learning techniques to learn the patterns).
- Heterogeneous resources as the data source for generating ontology: terminological inconsistencies are common when document sources are diverse which is also a common information retrieval problem. A possible way to solve this is to find the synonyms or similar terms from general linguistic ontologies (such as SENSUS or WordNet) or use the co-occurrence techniques to cluster the concepts based on the high similarities to easily detect the inconsistency.
- The automatically constructed ontology can be too prolific and deficient at the same time. Excessively prolific ontologies could hinder domain expert’s browsing and correction (reasonable choice of seedwords and initial cleaning and training data should limit this risk). On the other hand, automatically generated ontologies could be deficient since they rely on seedwords only. One promising technique could be synonym learning (Riloff & Shepherd, 1997)

2.2 SKC (University of Stanford)

The Scalable Knowledge Composition (SKC) project aims to develop a novel approach to resolve semantic heterogeneity in information systems. The project attempts to derive general methods for ontology integration that can be used in any application area so that it is basically application neutral. An ontology algebra has been developed therefore to represent the terminologies from distinct, typically autonomous domains. This research effort is funded by United States Air Force Office of Scientific Research (AFOSR), with cooperation of the United States Defense Advanced Research Project Agency (DARPA) High-Performance Knowledge Base (HPKB) program.

In this project, Jannink & Wiederhold (1999) and Jannink (1999) converted the Webster's dictionary data to a graph structure to support the generation of a domain or task ontology. The resulting text is tagged to mark the parts of the definitions, similar to the XML (eXtensible Markup Language) structure. According to their research purpose, only head words (<hw> ... </hw>) and definitions (<def> ... </def>) having many-to-many relationships are considered. This resulted in a directed graph that have the two properties that each head word and definition grouping is a node; and each word in a definition node is an arc to the node having that head word.

Method

Jannink & Wiederhold (1999) did not describe the adopted technique in detail in their publication. However, Jannink (1999) mentioned that they used a novel algebraic extraction technique to generate the graph structure and create the thesaurus entries for all words defined in the structure including some stop words (e.g., a, the, and). Ideas from the PageRank algorithm (Page & Brin, 1998) were also adopted. This is a flow algorithm over the graph structure of the WWW that models the links followed during a random browsing session through the web. The ArcRank from PageRank model was chosen to extract relationships between the dictionary words and the strength of the relationship. The attraction of using the dictionary as a structuring tool is that head words are distinguished terms from the definition text which provides the extra information allowing for types of analysis that are not currently performed in traditional data mining, and information retrieval. This method could also be applied to document classification and the relevance ranking of mining queries. The basic hypothesis for this work is that structural relationships between terms are relevant to their meaning. The methodology to extract the relations (the important component of ontology) is achieved through a new iterative algorithm, based on the Pattern/Relation extraction algorithm as follows (Brin, 1998):

- Compute a set of nodes that contain arcs comparable to seed arc set
- Threshold them according to ArcRank value
- Extend seed arc set, when nodes contain further commonality
- If the node set increased in size repeat from the first step.

The algorithm outputs a set of terms that are related by the strength of the associations in the arcs that they contain. These associations were computed according to local hierarchies of subsuming and specializing relationships, and set of terms is related by the kinship relation. The algorithm is naturally self-limiting via the thresholds. This approach also can be used to distinguish senses. For instance, the senses of a word such as *hard*, are distinguished by the choice of association with *tough* and *severe*. Also, ranking the different senses of a term by the strength of its associations with other terms could uncover the principal sense of a term.

Problems encountered

A number of problems have been highlighted during the process of ontology generation in this project:

- Syllable and accent markers in head words
- Misspelled head words
- Mis-tagged fields
- Stemming and irregular verbs (e.g. hopelessness)
- Common abbreviations in definitions (e.g. etc.)
- Undefined words with common prefixes (e.g. un-)
- Multi-word head words (e.g. water buffalo)
- Undefined hyphenated and compound words (e.g. sea-dog)

The interested reader can refer to Jannink (1999) for a more detailed account of the methodology and problems encountered.

2.3 Ontology Learning (AIFB, University of Karlsruhe)

The ontology learning group in AIFB (Institute of Applied Informatics and Formal Description Methods, University of Karlsruhe, Germany) is quite active in the ontology engineering area. They have developed various tools to support ontology generation that include OntoEdit (the ontology editor) and Text-To-Onto (an integrated environment for the task of learning ontologies from text) (Maedche & Staab, 2000 and 2000a).

Extracting ontology from domain data, especially domain-specific natural language free-texts turns out to be very important. Common approaches usually extract relevant domain concepts based on the shallow information retrieval techniques and cluster them into a hierarchy based on statistic and machine learning algorithm. But most of these approaches have only managed to learn the taxonomic relations in ontologies. To detect the non-taxonomic conceptual relationships, for example, the “has Part” relations between concepts, is becoming critical for building good-quality ontology (Maedche & Staab, 2000 & 2000a).

Method

AIFB's approach of ontology generation contains two parts: shallow text processing and learning algorithms. In shallow text processing, techniques have been implemented on top of SMES (Saarbrücken Message Extraction System). SMES is a shallow text processor for German language developed by DFKI (German Research Center for Artificial Intelligence, Germany) (Neumann, et al., 1997). It comprises techniques for tokenizer, lexicon, lexical analysis (morphological analysis, recognition of name entities, retrieval of domain-specific information, part-of-speech tagging) and Chunk parser. SMES uses weighted finite state transducers to efficiently process phrasal and sentential patterns (for the basic knowledge about the above natural language processing techniques, please refer to Grishman, 1997 and Gaizausksa & Wilks, 1998).

The outputs of the SMES are dependency relations found through lexical analysis. These relations were treated as the input of the learning algorithms. Some of the dependency relations did not hold the meaningful relations of the two concepts which could be linked together (co-occurrence) by some mediator (i.e., proposition, and so on). SMES also returns some phrases without any relations. Some heuristic rules have been defined to increase the high recall of the linguistic dependency relations, for instance, NP-PP-heuristic (attaching all prepositional phrases to adjacent noun phrases), sentence-heuristic (relating all concepts containing in one sentence), and title-heuristic (linking the concepts appeared in the title with all the concepts contained in the overall document) (Maedche & Staab, 2000 & 2000a).

The learning mechanism is based on the algorithm for discovering generalized association rules proposed by Srikant and Agrawal (1995). The learning module contains four steps: (1) selecting the set of documents; (2) defining the association rules; (3) determining confidence for these rules; and (4) outputting association rules exceeding the user-defined confidence.

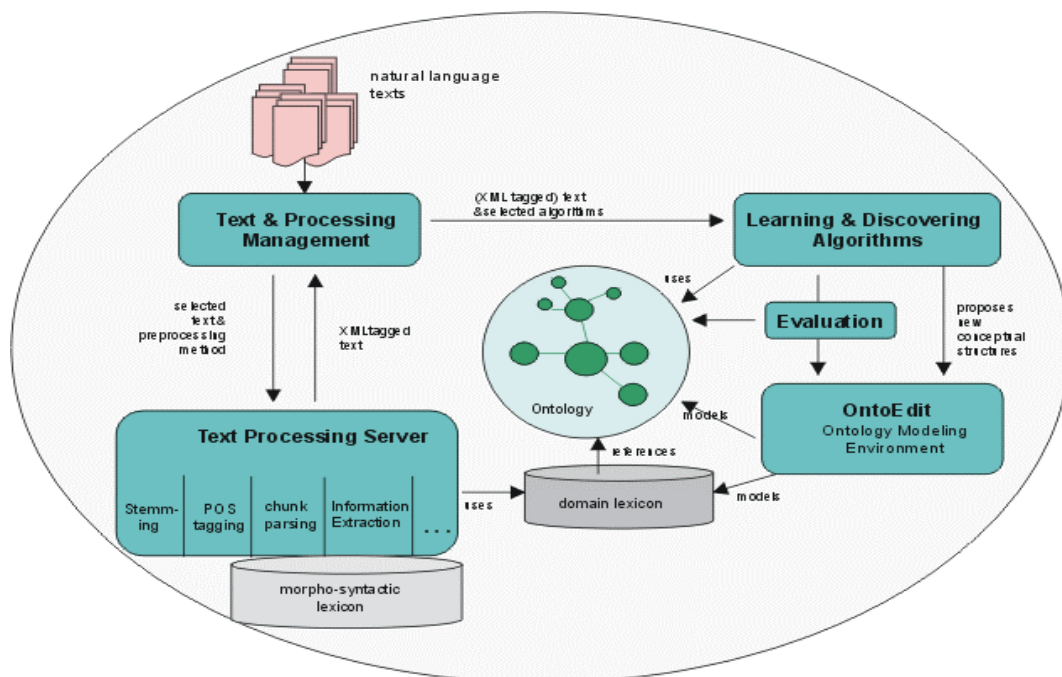


Figure 3. General overview of ontology learning system Text-to-Onto (<http://ontoserver.aifb.uni-karlsruhe.de/texttoonto/>)

AIFB also built up a system to facilitate the semi-automatic generation of the ontologies known as Text-To-Onto (Maedche & Staab, 2000b) as shown in Figure 3. The system includes a number of components: (1) text & processing management component (for selecting domain texts exploited for the further discovery process); (2) text processing server (containing a shallow text processor based on the core system SMES. The result of text processing is stored in annotations using XML-tagged text); (3) lexical DB & Domain Lexicon (facilitating syntactic processing based on the lexical knowledge); (4) learning & discovering component (using various extraction methods (e.g. association rules) for concept acquisition); and (5) ontology engineering environment (supporting in semi-automatically adding newly discovered conceptual structures to the ontology by the ontology engineers (OntoEdit).

Kietz, Maedche and Volz (2000) adopted the above method to build an insurance ontology from a corporate Intranet. First, the GermaNet was chosen as the top-level ontology for the domain-specific goal ontology. Then, domain-specific concepts were acquired using a dictionary that contains important corporate terms that were described in natural language. Then, a domain-specific and a general corpus of texts were used to remove concepts that were not domain-specific through some heuristic rules. Relations between concepts were learned by analyzing the corporate Intranet documents based on a multi-strategy learning algorithm, either from the statistical approach (frequent coupling of concepts in sentences can be regarded as relevant relations between concepts) or from the pattern-based approach.

Problems encountered

A number of problems have been highlighted in the process of ontology generation in this project:

- The lightweight ontology contains too many noisy data. For instance, not every noun-phrase or single term can be considered as the concept or class. The word sense problem generates lots of ambiguity.
- The refinement of these lightweight ontologies is a trickle issue that needs to be resolved in future. For instance, it should be domain experts involved or should also be semi-automatic based on the recursive machine learning algorithms.
- The learning relationship is not trivial. For instance, general relations can be detected from the hierarchical structure of the extracted concepts (or terms). How to identify and regroup the specific or concrete relationships becomes the main hurdle for ontology generation.

2.4 ECAI 2000

Some of the automatic ontology learning researches reported in the Ontology Learning Workshop of ECAI 2000 (European Conference on Artificial Intelligence) are important and appropriate to this survey. A number of shallow natural language processing techniques such as Part-of-Speech tagging, word sense disambiguation, tokenization and so on, are directly relevant. These are used to extract important (high frequency) words or phrases that could be used to define concepts. In the concept formation step, it is usual for some form of general top-level ontologies (WordNet, SENSUS) to be used to assist to extract correctly the final terms and disambiguate the word senses.

Methods

Wagner (2000) addressed the automatic acquisition of selectional preferences of verbs by means of statistical corpus analysis for automatic ontology generation. Such preference is essential for inducing thematic relations, which link verbal concepts to nominal concepts that are selectionally preferred as their complements. Wagner (2000) introduced a modification of Abe & Li (1996)'s method (based on the well-founded principle: Minimum Description Length) and evaluated it by employing a gold standard. It aimed to find the appropriate level of generation of concepts that can be linked to some specific relations. The EuroWordNet database provided information that can be combined to obtain a gold standard for selectional preferences. With this gold standard, lexicographic appropriateness can be evaluated automatically. However, one of the drawbacks of this method was that the learning algorithms were fed with word forms rather than word senses.

Chalendar & Grau (2000) conceived a system SVETLAN, which was able to learn categories of nouns from domain-free texts. In order to avoid general classes, they also considered the contextual use of words. The input data of this system was the semantic domains with the Thematic Units (TU). Domains were sets of weighted words, relevant to represent the same specific topic. The first step of SVETLAN was to retrieve text segments of the original texts associated to the different TUs. Then all the triplets constituted by a verb, the head noun of a phrase and its syntactic role from the parser results in order to produce Syntactic Thematic Units (STUs). The STUs belonging to a same semantic domain were aggregated altogether to learn about a Structured Domain.

Bisson, Nedellec & Canamero (2000) described Mo'K (a configurable workbench) to support the development of conceptual clustering methods for ontology building. Mo'K was intended to assist ontology developers to define the most suitable learning methods for a given task. It provides facilities for evaluation, comparison, characterization and elaboration of conceptual clustering methods.

Faure & Poibeau (2000) discussed how semantic knowledge learned from a specific domain can help the creation of a powerful information extraction system. They combined two systems, SYLEX (Constant, 1996) and ASIUM together, and termed it the "double regularity model", to eliminate the negative impacts of individual systems to yield good results. For instance, this combination could easily avoid a very time-

consuming manual disambiguation step. The special part of their work is that the conceptual clustering process does not only identify the lists of nouns but also augments this list by induction.

Todirascu, et al. (2000) used the shallow natural language processing parsing techniques to semi-automatically built up the domain ontology (conceptual hierarchy) and represent them in Description Logics (DL) which provides powerful inference mechanism and is capable of dealing with incomplete, erroneous data. Different small French corpora have been tested in the prototype. The system was capable of identifying relevant semantic issues (semantic chunks) using minimal syntactic knowledge and the complex concepts were inferred by the DL mechanism. Several tool were employed in the model:

- A POS tagging identifies the content words (nouns, adjectives, verbs) and functional words (prepositions, conjunctions, etc.). The tagger uses a set of contextual and lexical rules (based on prefixes and suffixes identification) learned from annotated texts.
- The sense tagger contains a pattern matcher, a set of patterns (words, lexical categories and syntagms) and their sense assigned by a human expert. The sense is represented by DL concepts. The set of conceptual descriptions was established by a human expert from a list of the most frequent repeated segments and words extracted from a set of representative texts. The pattern matcher annotates each sequence of words matching the pattern with its semantic description.
- The chunk border identification identifies the words and the syntactic constructions delimiting the semantic chunks. This module uses the output of the POS tagger, as well as a set of manually built cue phrases (syntactic phrases containing auxiliaries, composed prepositions etc.). The borders of noun and prepositional phrases (determiners, prepositions) are best candidates for the chunk border.

This research has basically automated the process of creating domain hierarchy based on a small set of primitive concepts defined by the human expert. The expert also has to define the relations of these concepts. As part of future research, focus on the use of document summaries as indexes and integration the system in XML documents were emphasised.

Problems encountered

The main problem arising from these researches pertains to relation extraction. Such relations were defined manually or inducted from the hierarchical structure of the concept classes. A potential solution proposed is to have provision of very general relations, such as “assoc-with”, “is-a” and so on. How to efficiently extract concrete relations for the concept class remain an important and an interesting topic for ontology learning and research.

2.5 Inductive logic programming (University of Texas at Austin (UT))

The Machine Learning group of UT applied the Inductive Logic Programming (ILP) to learn relational knowledge from different examples. Most machine learning algorithms are restricted to feature-based examples or concepts therefore limit themselves for learning complex relational and recursive knowledge. The applications of ILP by this group are extended to various problems in natural language and theory refinement.

Methods

Thompson & Mooney (1997) described a system WOLFIE (Word Learning From Interpreted Examples) that learns a semantic lexicon from a corpus of sentences. The lexicon learned consists of words paired with representations of their meaning, and allows for both synonym and ploysemy. WOLFIE is part of an integrated system that learns to parse novel sentences into their meaning representations. The overview of WOLFIE algorithm is to “derive possible phrase-meaning pairs by sampling the input sentence-representation pairs that have phrases in common, and deriving the common substructure in their representations, until all input representation can be composed from their phrase meanings, then add the best phrase-meaning pair to the lexicon, constrain the remaining possible phrase-meaning pairs to reflect the pair just learned, return the lexicon of learned phrase-meaning pairs”.

Inductive logic programming (ILP) is a growing topic in machine learning to study the induction of Logic programs from examples in the presence of background knowledge (Lavrac & Dzeroski, 1994).

Problems encountered

UT’s systems (Inductive Logic Programming) have the potential to become a workbench for ontological concept extraction and relation detection. It combines the techniques from the information retrieval, machine learning and artificial intelligence for concept and rule learning. However, how to deploy the UT’s methods for ontology concept and rule learning is still an open question that needs to be resolved to make this workbench a feasible proposal.

2.6 Library Science and Ontology

Traditional techniques deployed in library and information science area has been significantly challenged by the huge amount of digital resources. Ontologies, explicit specification of the semantics and relations in a machine-processable way, have the potential to suggest a solution to such issues. The digital library and semantic web communities are at present working hand-in-hand and collaborating to address such special needs in the digital age. The recent European thematic network on digital libraries (DELLOS, www.ercim.org/delos/) and semantic web (ONTOWEB, www.ontoweb.org) jointly proposed a joint sponsorship of a working group on “Content standardization for cultural repositories” within the OntoWeb SIG on Content Management

(www.ontoweb.org/sig.htm) attest to the cooperation and future collaborations of these communities.

Methods

In digital library projects, ontologies are specified or simplified in the form of various vocabularies, including cataloguing codes such as Machine Readable Cataloguing (MARC), thesauri or subject heading lists, and classification schemes. Thesauri and classifications, on the one hand, are used to represent the subject content of a book, journal article, a file, or any form of recorded knowledge. Semantic relationships among different concepts are reflected through broader terms, narrower terms and related terms in thesauri, and a hierarchical structure in classification schemes. On the other hand, a thesauri does not handle descriptive data (title, author, publisher, etc.). In this instance, separate representational vocabularies for the descriptive data such as the Anglo-American Cataloguing Rules 2nd (AACR2) to meet the need for descriptive representation, or the Dublin Core Metadata (<http://www.dublincore.org>) have been used.

The fundamental difference between an ontology and conventional representational vocabulary is the level of abstraction, relationships among concepts, the machine-understandable, and the most important but not the least, the expressiveness nature that can be provided by ontologies. For instance, an ontology can be layered according to different requirements (similar as the design model of the object-oriented programming languages (UML)). So we have upper-level ontology to define the general and generic concept or the schema of the lower-level ontology. Besides, ontology can be functioned as the database schema to define various tasks or applications. In a nutshell, ontology aims to achieve the communication between human and computer, while conventional vocabulary in library and information science fulfils the requirement for the communication among human beings. Ontology promotes standardization and reusability of information representation through identifying common and shared knowledge. Ontology adds values to traditional thesauri through deeper semantics in digital objects, both conceptually, relationally and machine understandably (Kwasnik, 1999).

The University Michigan Digital Library (UMDL) (Weinstein, 1998) maps the UMDL ontology to MARC with either “has” or “of” or “kind-of” or “extended” relationships. In another study, Kwasnik (1999) converted a controlled vocabulary scheme into an ontology citing it as an added-value contribution between ontology and the knowledge representation vocabularies used in libraries and information industries due to the following reasons:

- Higher levels of conception of descriptive vocabulary
- Deeper semantics for class/subclass and cross-class relationships
- Ability to express such concepts and relationships in a description language; and
- Reusability and “share-ability” of the ontological constructs in heterogeneous system
- Strong inference and reasoning functions

Qin & Paling (2001) used the controlled vocabulary at the Gateway to Educational Materials (GEM) as an example for converting it into an ontology. The major difference between the two models is the value-added through deeper semantics both conceptually and relationally. The demand to convert the controlled vocabulary into ontology is due to the limited expressive power of controlled vocabulary, the emerging popularity of agent communication (ontology-driven communication), the semantic searching through the Intranet and Internet, and the content and context exchanges existing in various marketplace of eCommerce. The purpose of such conversions not only reduces the duplication of effort involved in building an ontology from scratch by using the existing vocabularies, but also establishes a mechanism for allowing differing vocabularies to be mapped onto the ontology.

Problems encountered

Three main problems remain in the area of development of ontology in this respect:

- Different ways of modeling the knowledge (Library Science and Ontology) due to the “shallow” and “deeper” semantics that are inherent in these two disciplines;
- Different way of representing the knowledge; For instance, librarian uses the hierarchical tree (more lexical-flavored) to represent thesaurus or catalogues, while Ontology engineer uses mathematical and formal logics to enrich and represent ontologies (more mathematical and logical-flavored)
- Achieving consensus to merge the two to create a standardized means to organize and describe information has a long way to go.

2.7 Others

Borgo, et al. (1997) used the lexical semantic graphs to create ontology (or annotate the knowledge) based on the WordNet. They pointed out the some special nouns, which are always used to represent relations, for instance, part (has-part), function (function-of) and so on. They called these special nouns as “relational nouns” which would facilitate to identify the relations between two concepts.

Yamaguchi (1999) focused on how to construct domain ontologies based on a machine readable dictionary (MRD). He proposed a domain ontology rapid development environment (called DODDLE) to manage concept drift (word sense changes due to different domains). However, no detailed information about the basic theory that they adopted was made available. Nonetheless, it implies some form of concept graph mapping plus user-involved word sense disambiguation based on the WordNet to trim and deal with the concept shift so as to get the very specific small domain ontology from the user input containing several seed words for the domain.

Kashyap (1999) proposed an approach for designing an ontology for information retrieval based on the schemas of the databases and a collection of queries that of interest to the users. Ontology construction from database schema involves many issues, such as

determining primary keys, foreign keys, inclusion dependencies, abstracting details, grouping information from multiple tables, identifying relationships, and incorporating concepts suggested by new database schema. A set of user queries expressing information needs of users could be used to further refine the ontology created which could result in the design of new entities, attributes, class-subclass relationships that are not directly presented in existing database schemas. The generated ontology can be further enhanced by the use of a data dictionary and controlled vocabulary. The approach for ontology construction in this instance, is therefore, to make the full use of the existing sources, such as database schemas, user queries, data dictionaries and standardized vocabularies for proposing an initial set of entities, attributes, and relationships for an ontology.

3. Summary and conclusions

As the first part of the survey of ontology generation, we have examined researches that are related to semi-automatic or automatic ontology generation. Table 1 summarizes the general pattern and characteristics of the various methods adopted by different research groups or researchers along the dimensions of the source data, methods for concept extraction and relation extraction, ontology reuse, ontology representation, associative tools and systems and other special features.

In general, we can observe the following salient points and features in ontology generation to date:

- Source data are more or less semi-structured and some seed-words are provided by the domain experts for not only searching for the source data but also as the backbone for ontology generation. Learning ontology from free-text or heterogeneous data sources are still within the research lab and far beyond the real applications.
- For concept extraction, there already exist some quite-matured techniques (such as POS, word sense disambiguation, tokenizer, pattern matching, etc.) that have been employed in the field of information extraction, machine learning, text mining and natural language processing. The results of these individual techniques are promising as basic entities and should prove most useful in the formation of concepts in ontology building.
- Relation extraction is a very complex and difficult problem to resolve. It has turned out to be the main impedance to ontology learning and applicability. Further researches are encouraged to find appropriate and efficient ways to detect or identify the relations either semi-automatically or automatically.
- Ontologies are highly reused and reusable. Based on a basic ontology, other forms of ontologies may be lifted off to cater to specific application domains. This is important due to the cost of generation, abstraction and reusability.

- Ontologies can be represented as graph (conceptual graph), logic (description logic), web standards (XML), or a simple hierarchy (conceptual hierarchy). Currently there is the standard ontology representation language called DAML+OIL (www.daml.org/2001/03/daml+oil-index) which combines the merits of the description logic, formal logic and web standards.
- A number of tools have been created to facilitate ontology generation in a semi-automatic or manual way. For instance, University of Karlsruhe (Germany) developed and commercialized the semi-automatic ontology editor called OntoEdit (now owned by the Spin-off Company called Ontoprice). Stanford University exploited and provided an ontology-editing environment called Protégé with massive users. University of Manchester owns OilEd - an ontology editor for supporting DAML+OIL².

It is evident that much needs to be done in the area of ontology research before any viable large scale system can emerge to demonstrate ontology's promise of superior information organization, management and understanding. Far beyond ontology generation, evolving and mapping existing ontologies will form another challenging area of work in the ontology field.

² For details about the state-of-the-art of ontology editor, refer to:
<http://www.ontoweb.org/workshop/amsterdamfeb13/index.html>

Table 1. Summary of state-of-the-art ontology generation studies and projects

	InfoSleuth (MCC)	SKC (University of Stanford)	Ontology Learning (AIFB)	ECAI2000	Inductive logic programming (UT)	Library Science and Ontology	Others
Source Data (tagged)	- domain thesaurus - seedwords from expert - free-text doc from Internet (POS tagged automatically)	- Webster's online dictionary (in XML or SGML format)- semi-structured source data	- free-text natural language documents from the Web	- domain-free text - semantic domain with Thematic Units (TU) - annotated texts - primitive concepts from the human experts	- annotated documents - a corpus of sentences	- controlled vocabulary	- machine readable dictionary - schema of database - user queries
Methods for concept extraction	superficial syntactic analysis: pattern matching + local context (noun phrases). Word sense disambiguation is needed	- tag extraction - pattern matching (wrapper or script) - PageRank algorithm	- tokenzier - morphological analysis - name recognition - part-of-speech tagging - chunk parser	- category of nouns - conceptual clustering & induction - shallow natural language processing - POS tagging (contextual and lexical rules)	- slots fillers (rules learning from C4.5 & Rapier) - pattern matching - POS - Token	- subject headings from controlled vocabulary - manually refined concepts	
Methods for relation extraction	Relations were automatically acquired based on the linguistic property of noun components and the inheritance hierarchy. Two kinds: is-a and assoc-with	- ArcRank - an iterative algorithm based on Pattern/Relation extraction algorithm - relations could be learned and refined based on the local graphical hierarchies of subsuming and specialising	- co-occurrence clustering of concepts - mediator (proposition, verb) - heuristic rules based on the linguistic dependency relations - general association rules by machine learning	- selectional preferences of verbs (minimum description length with a gold standard)	-inductive logic programming (machine learning)	- relations from controlled vocabulary (broad term, narrow term, etc.) - manually refined relations	- relational nouns to represent relations
Ontology reuse	Yes (unit of measure, geographic metadata etc.)	Yes (online dictionary)	Yes (Lexicon)	Yes (EuroWordNet)		Yes (controlled vocabulary)	Yes (WordNet, data dictionary, controlled vocabulary)
Ontology representation	Hierarchical structure	Conceptual graph	XML	- conceptual hierarchy - description logic			- conceptual graph
Tool or system associated	None		. SMES . Text-To-Onto . OntoEditor	- SVETLAN - Mo'K - SYLEX - ASIUM	- WOLFIE		- DODDLE
Others	Corpus-based learning		Non-taxonomy relation learning				

References:

Abe, N. & Li, H. (1996). Learning word association norms using tree cut pair models. In *Proc. Of 13th Int. Conf. On Machine Learning*, 1996.

Berners-Lee, T. & Fischetti, M. (1999). *Weaving the Web*. Harper San Francisco, USA. 1999.

Berners-Lee, T., Hendler, J. & Lassila, O. (2001). The Semantic Web. *Scientific American*, May 2001.

Bisson, G., Nedellec, C. & Canamero, D. (2000). Designing clustering methods for ontology building: The Mo'K workbench. In *Proceedings of the First Workshop on Ontology Learning (OL-2000) in conjunction with the 14th European Conference on Artificial Intelligence (ECAI 2000)*. August, Berlin, Germany.

Borgo, S., Guarino, N., Masolo, C., and Vetere, G. (1997). Using a large linguistic ontology for internet-based retrieval of object-oriented components. In *Proceedings of 1997 Conference on Software Engineering and Knowledge Engineering*. Madrid, Knowledge Systems Institute, Snokie, IL, USA.

Brin, S. (1998). Extracting patterns and relations from the world wide web. In *WebDB Workshop at EDBT '98*, 1998.

Chalendar, G. & Grau, B. (2000). SVETLAN: A system to classify nouns in context. In *Proceedings of the First Workshop on Ontology Learning (OL-2000) in conjunction with the 14th European Conference on Artificial Intelligence (ECAI 2000)*. August, Berlin, Germany.

Cocchiarella, N.B. (1991). Formal ontology. In H. Burkhardt & B. Smith (eds.), *Handbook of metaphysics and ontology* (pp. 640-647). Munich: Philosophia Verlag.

Constant, P. (1996). Reducing the complexity of encoding rule-based grammars. December 1996.

Ding, Y. (submitted). Ontology: The enabler for the Semantic Web. *Journal of Information Science*, submitted.

Faure, D. & Poibeau, T. (2000). First experiments of using semantic knowledge learned by ASIUM for information extraction task using INTEX. In *Proceedings of the First Workshop on Ontology Learning (OL-2000) in conjunction with the 14th European Conference on Artificial Intelligence (ECAI 2000)*. August, Berlin, Germany.

Fensel, D. (2001). *Ontologies: Silver Bullet for Knowledge Management and Electronic Commerce*. Springer-Verlag, Berlin.

Fensel, D., Hendler, J., Lieberman, H. and Wahlster W. (Eds., to appear): *Semantic Web Technology*, MIT Press, Boston.

Fensel D., Omelayenko B., Ding Y., Schulten E., Botquin G., Brown M., and Flett A. (2001). *Information Integration in B2B Electronic Commerce*, to appear.

Fernandez-Lopez, M. (1999). Overview of methodologies for building ontologies. In., *Proceedings of IJCAI-99 Workshop on Ontologies and Problem-Solving Methods: Lessons Learned and Future Trends*, in conjunction with the Sixteenth International Joint Conference on Artificial Intelligence, August, Stockholm, Sweden.

Gaizausksa R. and Wilks Y. (1998). Information Extraction: Beyond document retrieval, *Journal of Documentation*, 54(1), 70-105.

Gomez-Perez, A.; Juristo, N. & Pazos, J. (1995). Evaluation and assessment of knowledge sharing technology. In. Mars, N.J. (Ed.). *Towards very large knowledge bases – Knowledge building and knowledge sharing* (pp. 289-296). IOS Press: Amsterdam.

Grishman R. (1997). Information extraction: Techniques and challenges. In M. T. Pazienza, *International Summer School SCIE-97*, Springer-Verlag.

Gruber, T.R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5, 199-220.

Guarino N. (1998). Some Ontological Principles for Designing Upper Level Lexical Resources. *Proc. of the First International Conference on Lexical Resources and Evaluation*, Granada, Spain, 28-30 May 1998.

Hwang, C. H. (1999). Incompletely and imprecisely speaking: Using dynamic ontologies for representing and retrieving information. In. *Proceedings of the 6th International Workshop on Knowledge Representation meets Databases (KRDB'99)*, Linköping, Sweden, July 29-30, 1999.

Jannink, J. (1999). Thesaurus Entry Extraction from an On-line Dictionary. In *Proceedings of Fusion '99*, Sunnyvale CA, July 1999.

Jannink, J. & Wiederhold, G. (1999). Ontology maintenance with an algebraic methodology: A case study. In. *Proceedings of AAAI workshop on Ontology Management*, July, 1999.

Kashyap, V. (1999). Design and creation of ontologies for environmental information retrieval. In., *Proceedings of Twelfth Workshop on Knowledge Acquisition, Modeling and Management*, Alberta, Canada, October, 1999.

Kietz, J.-U., Maedche, A. and Volz, R. (2000): Extracting a Domain-Specific Ontology Learning from a Corporate Intranet. *Second "Learning Language In Logic" LLL Workshop, co-located with the International Conference in Grammere Inference (ICGI'2000) and Conference on Natural Language Learning (CoNLL'2000)*. Lisbon, Portugal, September 13-14, 2000.

Kwasnik, B. (1999). The role of classification in knowledge representation and discovery. *Library Trends*, 48(1), 22-47.

Lavrac, N. & Dzeroski, S. (Eds.) (1994). *Inductive Logic Programming: Techniques and Applications*. Ellis Horwood.

Maedche, A. & Staab, S. (2000). Discovering conceptual relations from text. In, *ECAI 2000. Proceedings of the 14th European Conference on Artificial Intelligence*, IOS Press, Amsterdam, 2000.

Maedche, A. & Staab, S. (2000a). Mining Ontologies from Text. In: Dieng, R. & Corby, O. (Eds). *EKAU-2000 - 12th International Conference on Knowledge Engineering and Knowledge Management*. October 2-6, 2000, Juan-les-Pins, France. LNAI, Springer.

McCarthy, J. (1993). Notes on formalizing context. *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence*, AAAI, 1993.

Miller, G.A. (1995). WORDNET: A lexical database for English. *Communications of ACM* (11), 39-41.

Neumann, G., Backofen, R., Baur, J., Becker, M. & Braun, C. (1997). An information extraction core system for real world german text processing. In, *ANLP'97 – Proceedings of the Conference on Applied Natural Language Processing* (pp. 208-215). Washington, USA.

Page, L., & Brin, S. (1998). The anatomy of a large-scale hypertextual web search engine. In. *Proceedings of the 7th Annual World Wide Web Conference*.

Qin, J. & Paling, S. (2001). Converting a controlled vocabulary into an ontology: the case of GEM. *Information Research*, 6(2).

Reich, J.R. (1999). Ontological Design Patterns for the Integration of Molecular Biological Information. In. *Proceedings of the German Conference on Bioinformatics GCB'99* (pp.156-166), 4-6.October, Hannover, Germany.

Riloff, E. & Shepherd, J. (1997). A corpus-based approach for building semantic lexicons. In *Proc. Internat. Symposium on Cooperative Database Systems for Advanced Applications*, 1999.

Srikant, R. & Agrawal, R. (1995). Mining generalized association rules. In. *Proc. Of VLDB'95* (pp. 407-419), 1995.

Thompson, C.A. & Mooney, R. J. (1997). *Semantic Lexicon Acquisition for Learning Parsers*. Unpublished Technical Note. January 1997.

Todirascu, A., Beuvron, F., Galea, D. & Rousselot, F. (2000). Using description logics for ontology extraction. In. *Proceedings of the First Workshop on Ontology Learning (OL-2000) in conjunction with the 14th European Conference on Artificial Intelligence (ECAI 2000)*. August, Berlin, Germany.

Uschold, M. (2000). Creating, integrating and maintaining local and global ontologies. In. *Proceedings of the First Workshop on Ontology Learning (OL-2000) in conjunction with the 14th European Conference on Artificial Intelligence (ECAI 2000)*. August, Berlin, Germany.

Uschold, M. & Gruninger, M. (1996). Ontologies: principles, methods, and applications. *Knowledge Engineering Review*, 11(2), 93-155.

Wagner, A. (2000). Enriching a lexical semantic net with selectional preferences by means of statistical corpus analysis. In. *Proceedings of the First Workshop on Ontology Learning (OL-2000) in conjunction with the 14th European Conference on Artificial Intelligence (ECAI 2000)*. August, Berlin, Germany.

Weinstein, P. (1998). Ontology-based metadata: transforms the MARC legacy. In. Akscyn, F. & Shipman, F.M. (Edit). *Digital Libraries 98, Third ACM Conference on Digital Libraries*. New York: ACM Press, 254-263.

Yamaguchi, T. (1999). Constructing domain ontologies based on concept drift analysis. In., *Proceedings of IJCAI-99 Workshop on Ontologies and Problem-Solving Methods: Lessons Learned and Future Trends*, in conjunction with the Sixteenth International Joint Conference on Artificial Intelligence, August, Stockholm, Sweden.