

Towards a Data-driven Ontology Engineering Framework

Steve LEUNG, Fuhua LIN, Dunwei WEN

School of Computing and Information Systems, Athabasca University, Canada
stevel, oscarl, dunweiw@athabascau.ca

Abstract: The dynamic nature of ontology requires a new methodology to discover the evolving semantics of a particular conceptualization and maintain the novelty of specific ontology with minimal human intervention. This paper posits a conceptual framework that supports a data-driven, iterative, and self-correcting ontology engineering methodology for developing application-oriented and light-weight ontology. The method is being tested and implemented in a work-in-progress intelligent educational system (IES) project that serves an agent-based and ontology-driven academic advising system.

Keywords: Ontology Engineering, Ontology Evolution, Text Categorization, Academic Advising.

Introduction

An ontology consisting of static, consistent and a priori knowledge is an excellent candidate to ensure interoperability among distributed applications. Development of ontology is time-consuming and labor-intensive. In practice, however, ontologies are heavily tied to the applications in which they are involved [11]. Although extensibility is considered to be a significant consideration as noted in development guidelines [3], it is difficult to foresee how the ontologies will be used in future applications. The complexity of ontology engineering is further increased by the many different types of ontologies. As early as 1990s when ontology were innovatively introduced, Guarino recognized the various understandings of ontologies ranging from informal conceptual systems, complex semantics networks, to logical theories [4, 8]. While a complex ontology is able to serve both simple and complex applications, it is difficult to determine the appropriate level of complexity of the ontologies to be developed in the first place. With the pressure of shorter applications development time and the lack of consistent semantics integration standards, sometimes it is easier to develop ontologies from scratch than to reuse existing ontologies, not to mention finding the right ones [9, 10].

Similar to the problem of foreseeing future applications, it is difficult to foresee the future semantics of a catalog of concepts. For some domains, like books or pizzas, the meanings of key concepts probably will not vary significantly. For other domains, especially with elements of human perception, the meanings of the same terms probably will be changed over time. For example, if we want to create an ontology to represent the interests or intentions of people pursuing an academic degree, the exact meanings of the terms that represent a particular interest or intention are likely to be changed over time. The dynamic nature of ontologies has been noted by some authors [5, 6], and a mechanism to accommodate registering, searching and propagating changing ontologies has been

proposed. Yet attention should be paid to the evolving semantics of a particular conceptualization.

While attempts to maintain the novelty of ontology is very valuable, this paper proposes a conceptual framework that supports a data-driven, iterative and self-correcting ontology engineering methodology to developing application-oriented, light-weight ontology. The ultimate goal of the methodology is to minimize the need for human expert intervention by substituting with appropriate and high-quality information. With an automated implementation, multiple iterations can be executed in a short span of time to reflect newly discovered knowledge and achieving self-correcting results.

1. The Conceptual Framework

As shown in Fig. 1, the objective of the framework is to arrive at a versioned ontology (7) that matches as much as possible to the universe (1). The universe is practically a collection of terms and concepts that represents a higher level concept. The higher level concept might, and in most cases, contains several different sub-concepts. For example, the

concept “Career Objective” should include academic, managerial or technical careers. Notice that the sub-concepts are not mutually exclusive. It is common that a single person chooses more than one sub-concepts to represent his own perception of “Career Objective”. The collection is an open set, which means that the entirety of the terms and concepts describing the higher level concepts are not known and subject to changes.

Each sample set (2) is a subset of the universe. Ideally an individual sample represents a natural collection that measures a few sub-concepts of the high level concept. For example, each sample in our study is a human respondent who provides a list of terms that represents his/her own intentions and interests of pursuing a postgraduate degree.

Samples will go through a clustering process to form natural clusters. Different clustering methods may yield different results, but the purpose is to segregate samples into clusters which use different words to describe sub-concepts. In other words, each cluster is a collection of terms that represent the same sub-concepts. Although it is possible that each cluster consists of multiple sub-concepts, a single sub-concept should not exist in different clusters simultaneously. The result is a collection of conceptually independent clusters (4).

The conceptualization process (6) takes two parameters: a list of recognizable terms (5), and the clusters (4). The recognizable terms (5) are a subset of the universe but their meanings are known and unambiguous. Domain experts and peer reviews will be required to create the first iteration of recognizable terms. The conceptualization process aims to define unrecognizable words in terms of the recognizable terms. Assuming homogeneity of clusters, if there is only one unrecognizable term, the term will be recognized as the same meanings of recognizable terms in the same sample. If there is more than one unrecognizable term, a simple majority rule will be used to determine its meanings. The process will count number of samples that links an unknown term to same recognizable term. If the number exceeds an arbitrary threshold, the unknown term will be defined as the recognizable term, and added as a relevant concept to the conceptualized ontology (7). The recognized terms will be added to the list of recognizable terms to form a version of an

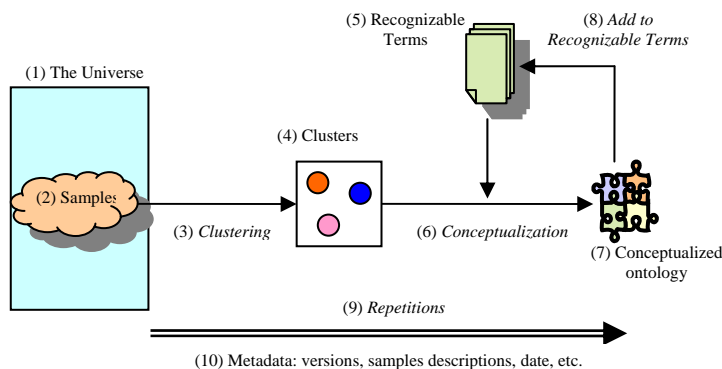


Fig. 1 Conceptual Model of Data-Driven Ontology Engineering Method

ontology (8). All remaining unknown terms will be dropped. The entire process can be repeated for new sample sets (9), and proper metadata should be stamped onto each ontology generated (10).

The following example extracted from an empirical study will illustrate the process. Before the process, a list of recognizable terms is extracted manually from the courses offered by a postgraduate program. The list includes the terms [software applications, uml, java, xml, information management]. Using nearest neighbor method, three of data items that form part of a cluster are:

```
#1: Unknown term = [application development]
    Recognized term = [software engineering, software applications, uml, java, xml]

#2: Unknown term = [application development]
    Recognized term = [information management]

#3: Unknown term = [application development, information system]
    Recognized term = [software development, networking, software engineering, software design]
```

From item #1 and #2, since there is only one unknown term [application development], it is defined as [software applications, uml, java, xml, information management]. In both item #1 and #3, [software engineering] is linked to [application development], it is added to the semantics as we set the voting threshold to one. The final definition of [application development] is [software engineering, software applications, uml, java, xml, information management].

There are several characteristics of the proposed approach to ontology engineering. First, the approach is primarily data-driven and requires minimal human intervention. Human expertise is required at two steps: (1) Maintenance of the list of recognizable terms – human expertise must be used to create the first version of the recognizable terms, and regularly inspect later versions. (2) Sampling – rigorous sampling methodology and field methods should be implemented by survey experts to ensure overall quality.

Second, the framework supports automated processes. All but the first set of recognizable terms are created from data without human intervention.

Third, the approach does not require explicit effort to integrate older and newer versions of the resulting ontology. Results from older versions are already integrated into the list of recognizable terms and in turn fed into the conceptualization process of subsequent iterations. The resulting ontology is easily extensible, although human expertise might be required to judge the quality of the list on a regular basis.

Fourth, the approach is directed towards self-correction and adaptation, provided that quality new samples are used. If the first sample set is not conclusive, further samples should provide more correct results. Moreover, as the higher level concept evolves over time, new samples measuring new sub-concepts can be added to the semantics of the concept, and making the resulting ontology adaptive to new knowledge.

2. An Empirical Study

Like the development of ontologies, the framework originated from a real-life problem and solution called *e-Advisor* [7]. *E-Advisor* is a multiple intelligent agents system developed to tackle individualized study planning problems. Parameters of the planning problem include students' interests and intentions, faculty availability, course prerequisites and requirements, and other administrative obligations.

A major parameter to study plans personalization is career objective and interest of students. Without this knowledge, it is impossible to personalize study plans. During the construction of *e-Advisor*, we adopted the taxonomy of job careers from MSCIS 2006

curriculum [2]. For two reasons the job titles do not necessarily reflect the true intention of learners. First, the perception of the categories between the designers and the students might not be the same. Second, the interests of a student might be a subset of a category, or a combination of subsets of two or more categories. It directs the need for a dynamic representation of the semantics of the job objectives that can be mapped to the courses offered.

An implementation in the form of an intelligent agent based on the proposed approach is being developed to derive an ontology of job objectives for *e-Advisor*. The universe of the job objective problem is a higher level concept “Job Objectives and Personal Interests of pursuing an MSc IS degree” that contains several interrelated sub-concepts. To collect samples, a survey is conducted to ask students for their intentions and interests in the program. A total of 135 valid respondents are collected. The first list of recognizable terms is extracted from the courses offered in the program. Due to the small sample size, the majority threshold is set to one. An ontology including the example described above is constructed.

3. Limitations and Issues

- *Problem Space* - The nature of the problem is to assign and reassign meanings to unknown terms in a collection of concepts and terms that represents dynamic higher level concepts. The iterative and data-driven approach is appropriate to approximate the “true” meanings of each unknown term in the universe. It implies that the nature of the collection is dynamic and contains several sub-concepts. On the other hand, businesses involved in ordering unambiguous products like pizza or books, would not benefit from this method. It is also inappropriate for systems that require logical implications within the ontology. If logical implications are essential for applications to make decisions, human expertise and proofs cannot be substituted by the data-driven and automated approach. Sophisticated ontologies and methods will be needed for this purpose.
- *Integration and Extensibility* - Although the approach provides for automated integration, if the first sample is inadequately or even incorrectly representing the universe, later generations will suffer from the incorrect list of recognizable terms. It is where human expertise is instrumental to judge the quality of the final ontology.
- *Samples and Sampling Methodology* - The conceptual framework takes the assumption that a sample set represents a true subset of the underlying universe. It implies that the collection of samples must follow rigorous survey methodology. Probabilistic sampling, careful definition of population, and proper field methods should be followed as much as possible to support quality results.
- *Clustering Methods and Distance Functions* - There are numerous data clustering methods and distance functions that yield different results. In our study we use the nearest neighbor method that yields larger clusters, and a simple distance function by counting matching word stems. Future explorations should include the testing of different clustering methods and distance functions.
- *Majority threshold* – the choice of the majority threshold will have an impact of the resulting ontology. The higher the threshold, the fewer concepts will be generated.
- *Consistency Maintenance and Versioning* - In theory the resulting ontology has as many versions as the number of sample sets. It is possible that newer versions could be drastically different from current version. Although it is assumed that the users or applications agents will decide which version to be used, abundant information must be embedded in the resulting ontology to support intelligent decision making.

4. Conclusion and Future Work

We have presented a conceptual framework towards a data-driven ontology engineering approach. With minimal human intervention, the approach shows the benefits of automated, fast, self-correcting and adaptive process for building and maintaining a specification of conceptualizations. A work-in-progress implementation of the framework is being developed to serve an academic advising system to illustrate the effectiveness of the approach. Interim result shows that the resulting ontology is useful as an input to an academic advising system and suitable for education applications.

There are several areas to be investigated: First, the results of the implementation should be evaluated by experts and end users. The original purpose of the framework is to achieve fast and iterative ontology generation method from data without too much human intervention. Its usefulness will illustrate the effectiveness of the ontology generated using the data-driven approach on educational applications. A comparison study between theory-based ontology should be carried out to test the applicability of the method.

General guidelines measuring the quality and effectiveness of the framework should be established. It includes, and not limited to, the guidelines to conceptual definition of the universe, conceptual description of sample sets, integration tests and measurement of difference between older and newer versions. Quality of light-weight ontology is tied to applicability, and applicability can be verified by empirical studies of end users' judgment.

Finally, a versioning system that supports the discovery of embedded metadata will be important for applications to choose among different versions of the ontology.

Acknowledgements

The Job Objectives Ontology Maintenance project is funded by Mission Critical Funding from Athabasca University. We specially thank Rory McGreal for his valuable comments.

References

- [1] Devedzic, V. (2002). Understanding Ontological Engineering. *Communications of the ACM*, Vol. 45, Issue 4, pp. 136 – 144.
- [2] Gorgone, J. T., Gray, P., Stohr, E. A., Valacich, J. S., & Wigand, R. T. (2006). MSIS 2006: Model Curriculum and Guidelines for Graduate Degree Programs in Information Systems, *Communication of the Association for Information Systems*.
- [3] Gruber, T. R (1995). Towards Principles for the Design of Ontologies used for Knowledge Sharing. *Int. J. of Human Computer Studies*, 43, pp. 907- 928.
- [4] Guarino, N. (1995). Ontologies and Knowledge Bases: Towards a Terminological Clarification, in N.J.I. Mars (ed.), *Towards Very Large Knowledge Bases*, IOS Press.
- [5] Heflin, J. & Hendler, J. (2000). Dynamic Ontologies on the Web. *Proceedings of 7th National Conference on Artificial Intelligence AAAI-2000*, AAAI/MIT Press.
- [6] Kishore, R., Zhang, H., & Ramesh, R. (2004). A Helix-Spindle Model for Ontological Engineering. *Communications of the ACM*, Vol. 47, No. 2, pp. 69 – 75.
- [7] Lin, F., Leung S., Wen D., Zhang F., Kinshuk & McGreal R. (2007). E-Advisor: A Multi-agent System for Academic Advising. *Agent-Based Systems for Human Learning and Entertainment (ABSHLE)*.
- [8] Mizoguchi, R. (2003). Tutorial on Ontological Engineering – Part 1: Introduction to Ontological Engineering. *New Generation Computing*, Ohm&Springer, Vol. 21, No. 4, pp.365 – 384.
- [9] Noy, N. (2005). Order from Chaos. *QUEUE*, Oct. 2005, pp. 42-49.
- [10] Noy, N. (2004). Semantic Integration: A Survey of Ontology-Based Approaches. *SIGMOD Record*, Vol. 33, No. 4, Dec. 2004, pp. 65 – 70.
- [11] Sebastiani, F. (2002). Machine Learning in Automated Text Categorization, *ACM Computing Surveys*, Vol. 34, No. 1, pp. 1- 47.
- [12] Spyns, P., Meersman, R. & Jarrar, M. (2002). Data Modeling versus Ontology engineering. *SIGMOD Record*, Vol. 31, No. 4, pp. 12 – 17.