Automated Ontology Learning and Validation using Hypothesis Testing

Michael Granitzer[1], Arno Scharl[1,3], Albert Weichselbraun[2], Thomas Neidhart[3], Andreas Juffinger[3], and Gerhard Wohlgenannt[2]

[1] Know-Center Graz, Austria mgrani@know-center.at
[2] Vienna University of Economics and Business Administration, Austria albert.weichselbraun@wu-wien.ac.at,wohlg@ai.wu-wien.ac.at
[3] Graz University of Technology, Austria {scharl,tneidhart,ajuffinger}@tugraz.at

Semantic Web technologies in general and ontology-based approaches in particular are considered the foundation for the next generation of information services. While ontologies enable software agents to exchange knowledge and information in a standardized, intelligent manner, describing todays vast amount of information in terms of ontological knowledge remains a challenge.

In this paper we describe the research project AVALON - Acquisition and VALidation of ONtologies, which aims at reducing the knowledge acquisition bottleneck by using methods from ontology learning in the context of a cybernetic control system. We will present techniques allowing us to automatically extract knowledge from textual data and formulating hypothesis based upon the extracted knowledge. Based on real world indicators, like for example business numbers, hypotheses are validated and the result is fed back into the system, thereby closing the cybernetic control system's feedback loop. While AVALON is currently under development, we will present intermediate results and the basic idea behind the system.

1 Introduction

Real-world applications that provide shared meaning understandable for machines and humans alike require semantic technologies, whose increasing importance is reflected by the Gartner Group's prediction that lightweight ontologies will be part of 75% of application integration projects. Tim Berners-Lee's vision of the Semantic Web [1] goes beyond lightweight ontologies and requires a network of trust, in which technology is capable of distinguishing reliable knowledge from collections of trivial data. Implementing this vision, however, involves the transformation of massive amounts of existing information, as well as the validation of the extracted semantics reliability and

validity with regards to the real world they describe. Semantic services not only encompass the World Wide Web, but also smaller corporate intranets and their integration into a global scheme. To ensure trust in the extracted knowledge, semantic technologies also need to cope with the highly dynamic contextual change of information.

Addressing these challenges, the research project AVALON develops a new generation of adaptive knowledge acquisition and management services that test semantic hypotheses by (i) automatically extracting knowledge from heterogeneous, unstructured information sources, (ii) discovering semantic associations within the automatically extracted knowledge base, and (iii) validating the extracted knowledge on observable real-world indicators. The ontology-supported testing of semantic hypotheses will increase the credibility and usefulness of the continuously evolving knowledge base. AVALON's evaluation and quality assurance processes utilise (semi-)automatic feedback loops to align extracted knowledge with external indicators and the expertise of individuals. Such feedback loops are particularly useful in highly dynamic environments like the World Wide Web. Therefore AVALON's methodology is based on a cybernetic control system [6]: A system with an internal knowledge base monitors real-world indicators and uses the knowledge base to recommend a particular action. If the decision-maker accepts the systems recommendation, his or her action affects the real world. By measuring changes in the real-world indicators resulting from taken actions, the knowledge base can be updated and refined automatically.

AVALON implements such an adaptive process for Web-based resources, automatically extracting semantics from unstructured and structured information sources, and validating the extracted knowledge on real-world indicators.

2 Conceptual Approach

AVALON starts by extracting and populating ontologies[4] from textual resources using state of the art techniques (see Section 3). Afterwards, hypothesis can be formulated upon the extracted ontology and tested statistically against the observable real-world indicators that describe out-comes of the decision making process. Based on this testing the knowledge base and subsequently the ontology can be validated and refined (see also Figure 1 for an overview.).

The formulation of ontology-based hypotheses deserves particular attention. The granularity of the hypothesis directly relates to the granularity of the ontology. In general, we distinguish two types of hypotheses:

[4] We distinguish here between a knowledge base and an ontology as stated in [9]. An ontology defines the domain schema, while the knowledge base contains instances of schema concepts

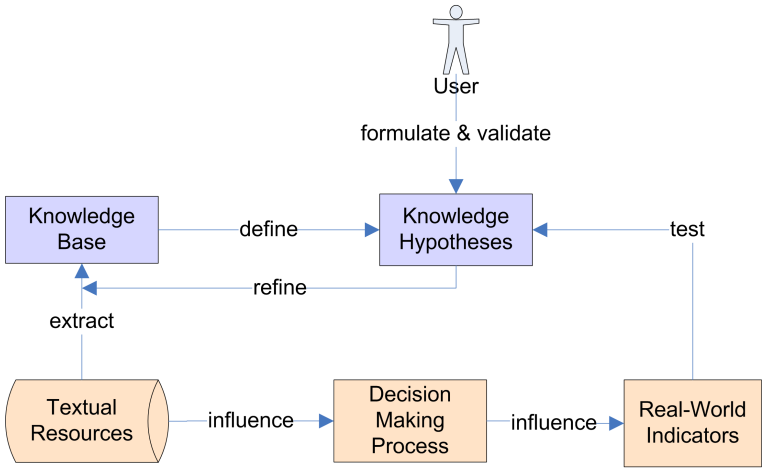

Fig. 1. Conceptual Model of AVALON

- Testing relationships stored in the ontology verifies whether the internal representation is accurate; the results strengthen or weaken dependencies among ontological concepts. This type of test can be performed automatically by choosing various hypotheses and comparing them to measured indicators. Rather than following a random or brute force approach, AVALON is based on heuristics for choosing the most relevant hypotheses for advancing the knowledge base.
- Testing of new relationships reveals new knowledge and refines the internal representation. If users analyzing the knowledge base find irregularities or patterns of interest, hypotheses can be formulated interactively to test those patterns. Through this process, AVALON does not only support knowledge discovery, but also the immediate validation of this knowledge based on a dynamic data set.

As an example, consider AVALON is crawling RSS feeds and blog entries of recent developments and trends in IT industries. After pre-processing the textual resources, relevant concepts (e.g. Persons, Topics etc.) and relations are extracted, together with instances of companies populated in the knowledge base. Taking into account stock quotes as real world indicators, AVALON is capable of combining concepts associated with companies and their loss/gain in stock quotes. This allows validating hypotheses of the form "*Innovative products represent success factors of Web 2.0 companies*". Extending the extraction and population methods to recognize persons and job titles (which can be easily done by means of information extraction) allows us, for example, to include Chief Executive Officers (CEO's) of companies, and to validate hypotheses that postulate that CEOs significantly influence corporate success.

More formally, we have an ontology O in the form of a graph, a set of instances I assigned to parts of the ontology and real world indicators R as

properties of parts of the instances. A hypothesis H is formulated in terms of concepts and relations of the ontology O and therefore can be seen as sub graph of the ontology $H \subset O$. To support user-defined hypothesis (see above) additional relations may be defined by the user of the system, to extend the hypothesis. Evaluating a hypothesis is done by selecting a hypothesis H and by splitting the instances into a set of instances $I_H \subset I$ satisfying the hypothesis and a set of instances hypothesis $I_{\overline{H}} \subset I$ not satisfying the hypothesis. Applying statistical significance tests on the indicators R estimate the correctness of a hypothesis.

In the above scenario for example, companies, properties describing a company and employees are defining the ontology. Thus, the knowledge base consists of instances of people who work for a company with specific properties. The occurrence function now counts, how often particular relationships or further on patterns exist. Statistically testing the assigned indicators of companies w.r.t to properties of the company and/or people working in the company, leads to a validation of the pattern and therefore of the hypothesis.

Iterating these steps allows creating an accurate and trustworthy knowledge base with a minimal amount of user interaction. Users will have access to knowledge facts and analysts will have the opportunity to formulate hypotheses and test them automatically. Thus, AVALON is capable of dynamically mapping external processes into an internal representation.

3 Technical Approach and Algorithms

From an algorithmic point of view, AVALON rests upon three pillars:

- *Determining the domain structure via ontology learning from text*: Unstructured information sources such as Web sites, Wikis, RSS Feeds and Blogs have to be analysed and converted into an ontology. Similar systems like for example the weblyzard [8] KIM platform [10] or Text2Onto [2] demonstrate the feasability to enrich unstructured information. In AVALON, the user defines relevant domain concepts via a seed ontology. AVALON extends this ontology in multiple iterations (c.f. [8]).
- *Populating the Knowledge Base via Information Extraction*: AVALON uses methods from information extraction for populating the knowledge base and attaching indicators to knowledge base instances. Our approach here is similar to [3, 7], where gazetteers are defined via instances in the knowledge base and rules consider ontological relationships (e.g. a CEO IS-A Person). Thus, we can bootstrap information extraction easily and consider domain relations during the extraction process. Finally, results are fed back implicitly to improve the ontology's accuracy and enlarge the knowledge base.
- *Selecting hypotheses via graph mining*: Hypothesis selection is a critical task in the context of the systems accuracy and performance. Selecting all

possible hypothesis is computational intractable, therefore requiring intelligent selection mechanisms. Finding groups of instances sharing similar structures in the ontology can be seen as clustering of graph structures [5] and will allow us to reduce the number of necessary hypothesis tests. For example, assume that we have populated our ontology with companies and attached indicators to them. By clustering those companies based on their relationship to other concepts (e.g. topics, persons etc.), groups of similar companies can be defined. Those groups may serve as starting point for testing hypothesises based on for example comparing stock quote indicators of one group to all other groups of similar companies.

From a technical point of view, computational power and storage capabilities are critical for the success of AVALON. Large scale computing is necessary in order to enable ontology learning on very large document sets, and to allow successive hypothesis testing. Fortunately, current developments in the Open Source community allow us to use distributed approaches originally pioneered by Google. The Hadoop project[5] provides a Open Source computing platform based on Google's Map&Reduce algorithm[4]; a functional programming model allowing large scale distributed computing on several hundreds of machines. Especially pre-processing and natural language processing of documents can be greatly enhanced using this platform.

In the current stage of our project, we are focusing on ontology learning and population using distributed computing based on Map&Reduce. Machine learning algorithms for learning the type of relations are in development and provide promising first results. In the next phase, AVALON will mine relevant sub-graphs to assist in formulating meaningful hypotheses. Finally, integrating real world indicators and user testing will reveal the capabilities of our conceptual model.

4 Conclusion and Outlook

Acquiring and validating knowledge is essential in any knowledge management system. In this paper we have presented the AVALON approach to acquiring and validating knowledge in a dynamic, iterative fashion. Combining textual resources with real-world indicators will help to continually improve the acquired knowledge. As defined in the theory for cybernetic control systems, combining very heterogeneous sources of information which have a relationship in the real world can lead to a accurate internal representation of the world. Based on this feedback process, changes in the real world can be reflected effectively in the internal representation. We consider the conceptual model of AVALON as a step towards flexible and dynamic knowledge-based systems that avoid the cumbersome process of manually scanning large doc-

[5] *http://lucene.apache.org/hadoop/*

ument sets. This will enhance the user's access to existing archives, and facilitate the deduction of new knowledge

Acknowledgement

The project results have been developed in the AVALON (Acquisition and Validation of Ontologies) project. AVALON is financed by the Austrian Research Promotion Agency (http://www.ffg.at) within the strategic objective FIT-IT under the project contract number 810803 (http://kmi.tugraz.at/avalon). The Know-Center is funded by the Austrian Competence Center program Kplus under the auspices of the Austrian Ministry of Transport, Innovation and Technology (www.ffg.at), and by the State of Styria.

References

1. T. Berners-Lee, J. Hendler, and O. Lassila. The semantic web. *Scientific American*, 284(5):34–43, 2001.
2. P. Cimiano and J. Völker. Text2onto - a framework for ontology learning and data-driven change discovery. In A. Montoyo, R. Munoz, and E. Metais, editors, *Proceedings of the 10th International Conference on Applications of Natural Language to Information Systems (NLDB)*, volume 3513 of *Lecture Notes in Computer Science*, pages 227–238, Alicante, Spain, JUN 2005. Springer.
3. F. Ciravegna, A. Dingli, Y. Wilks, and D. Petrelli. Adaptive information extraction for document annotation in amilcare. In *SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 451–451, New York, NY, USA, 2002. ACM Press.
4. J. Dean and S. Ghemawat. Mapreduce: Simplified data processing on large clusters. November 2004.
5. L. Getoor and C. P. Diehl. Link mining: a survey. *SIGKDD Explor. Newsl.*, 7(2):3–12, 2005.
6. F. Heylighen and C. Joslyn. Cybernetics and second-order cybernetics. In R. A. Meyers, editor, *Encyclopedia of Physical Science & Technology*. Academic Press, 3 edition, 2001.
7. J. Iria, F. Ciravegna, P. Cimiano, A. Lavelli, E. Motta, L. Gilardoni, and E. Mönch. Integrating information extraction, ontology learning and semantic browsing into organizational knowledge processes. In *Proceedings of the EKAW Workshop on the Application of Language and Semantic Technologies to support Knowledge Management Processes, at the 14th International Conference on Knowledge Engineering and Knowledge Management*, OCT 2004.
8. W. Liu, A. Weichselbraun, A. Scharl, and E. Chang. Semi-automatic ontology extension using spreading activation. *Journal of Universal Knowledge Management*, 0(1):50–58, 2005.
9. A. Maedche and S. Staab. Ontology learning for the semantic web. *IEEE Intelligent Systems*, 16(2):72–79, 2001.
10. B. Popov, A. Kiryakov, D. Ognyanoff, D. Manov, and A. Kirilov. Kim a semantic platform for information extraction and retrieval. *Nat. Lang. Eng.*, 10(3-4):375–392, 2004.