Preface

The LarKC PhD symposium is an annual event that the Large Knowledge Collider (LarKC) Consortium organized. The main objectives of this symposium series is to provide a communication platform for young researchers (especially PhD students) on their recent progresses in the EU FP-7 LarKC project, Web-scale reasoning and the Semantic Web in general. The seminar is open and free to everyone who is interested.

The 2010 LarKC PhD Symposium is the 2nd symposium in this series. The 1st symposium is jointly held with STI PhD Seminar 2009 in Berlin. The participants of that events all agree that they learned a lot from each other and that is one of the most important reason why we have this event this year. During year 2 of the LarKC project, the consortium has many progresses on Web-scale reasoning and search, ranging from new selection and reasoning strategies to real-world use cases. Many of them are from PhD students and young researchers in this consortium. We are very proud to have these researchers report their recent results in the 2nd LarKC PhD symposium. In addition, we are very pleased to see that there are external plug-in contributions outside the LarKC consortium in the form of close collaboration with the LarKC members.

The speakers for the 2nd LarKC PhD symposium are from China, Germany, Italy, the Netherlands, UK, etc. We are please to have several talks which cover a wide range on Semantic Web, Machine Learning, and AI in general. The topics focus but not limited to: Natural Language Interfaces to Ontologies, segmentation strategies for Web-scale data, Machine learning meets the Semantic Web, selection strategies, parallel and Contrastive reasoning for the Semantic Web, and Semantic Web-enabled Recommender System.

Some of the speakers are still in their PhD program, hence we are very pleased to have several senior members from the LarKC consortium to make comments and suggestions to their future research in the area of Web search and reasoning. More importantly, the speakers will learn from each other during their communications in the symposium.

November 10, 2010

Zhisheng Huang (Vrije University Amsterdam, the Netherlands)
Yi Zeng (Beijing University of Technology, China)
Reto Krummenacher (STI Innsbruck, Austria)
Speakers

**Danica Damljanovic** is a Research Associate in the GATE team, Natural Language Processing (NLP) Group, Department of Computer Science, University of Sheffield. Her interests include applying NLP and Information Retrieval to semantic web applications, usability of Natural Language Interfaces, Question answering systems, and application of semantic web technologies to E-Tourism.

*Title: Natural Language Interfaces to Ontologies*

**Matthias Assel** is from Department Intelligent Service Infrastructures, High Performance Computing Center Stuttgart (HLRS). He is the leader of Work Package 5 in the LarKC Project. His actual research focus rests on scalable applications/algorithms for context-sensitive data access, distributed data and text mining, and contextual content search as well as intelligent infrastructures for next-generation computing systems.

*Title: Intelligent Segmentation of Large Datasets*

**Yan Wang** is a PhD student at International WIC Institute (WICI), Beijing University of Technology. She is interested in medical literature search and reasoning. Her research interests include Semantic Web and Web Intelligence.

*Title: Interleaving reasoning and selection by knowledge summarization*

**Jacopo Urbani** is a PhD student in Distributed Systems and Semantic Web at the Vrije Universiteit Amsterdam, the Netherlands. His current research focuses on large scale reasoning in the Semantic Web. He is the winner of ESWC 2010 best student paper award, IEEE SCALE 2010 challenge at CCGrid 2010.

*Title: Scalable and parallel reasoning in the Semantic Web*

**Jun Fang** is currently a lecture in the school of automation at Northwestern Polytechnical University, China. His current work focuses on Semantic Web and ontology reasoning, especially on scalable reasoning and reasoning with inconsistent ontologies.

*Title: Contrastive Reasoning with Inconsistent Ontologies*
Speakers

Yi Huang
is a staff scientist at Siemens Corporate Research and Technology and is doing his PhD at Ludwig Maximilian University of Munich, Germany. His research interests focus on statistical machine learning, text mining, information retrieval, and the Semantic Web.

Title: Relation prediction using machine learning in semantic domains

Daniele Dell'Aglio
has been working since 2008 as junior researcher at CEFRIEL. His research activities focus on Semantic Web technologies, recommender systems, location-aware and mobile applications. He is currently participating in the Urban Computing research activities of LarKC FP7 project and in research activities related to Web service and recommender systems in SOA4All and Indenica FP7 projects. In the past he participated in the ended Service-Finder FP7 project.

Title: ANSWERS: A Novel Semantic Web-Enabled Recommender System
Agenda

- 9:00-9:05: Opening
- 9:05-9:35 Danica Damljanovich (USFD), Natural Language Interfaces to Ontologies
- 9:35-10:05 Matthias Assel (HLRS), Intelligent Segmentation of Large Datasets
- 10:05-10:35 Yan Wang (WICI), Interleaving reasoning and selection by knowledge summarization
- 10:35-11:00 Tea Break
- 11:00-11:30 Jacopo Urbani (VUA), Scalable and parallel reasoning in the Semantic Web
- 11:30-12:00 Jun Fang (NPU), Contrastive Reasoning with Inconsistent Ontologies
- 12:00-12:30 Yi Huang (SIMENS), Relation prediction using machine learning in semantic domains
- 12:30-13:00 Daniele Dell'Aglio (CEFRIEL), ANSWERS: A Novel Semantic Web-Enabled Recommender System
- 13:00: Closing Remark
Table of Contents

Natural Language Interfaces to Ontologies (Danica Damljanovich) ........................................... 1
Intelligent Segmentation of Large Datasets (Matthias Assel) ......................................................... 2
Interleaving reasoning and selection by knowledge summarization (Yan Wang, Zhisheng Huang, Yi Zeng) ........................................................................................................................................... 3
Scalable and parallel reasoning in the Semantic Web (Jacopo Urbani) ........................................... 4
Contrastive Reasoning with Inconsistent Ontologies (Jun Fang, Zhisheng Huang) ....................... 7
Relation prediction using machine learning in semantic domains (Yi Huang, Volker Tresp) ........ 8
ANSWERS: A Novel Semantic Web-Enabled Recommender System (Daniele Dell'Aglio) ............ 9
Natural Language Interfaces to Ontologies

Danica Damljanovic

Department of Computer Science, University of Sheffield
Regent Court, 211 Portobello Street, S1 4DP, Sheffield, UK
{D.Damljanovic}@dcs.shef.ac.uk

1 Abstract

Accessing structured data in the form of ontologies currently requires the use of formal query languages (e.g., SeRQL or SPARQL) which pose significant difficulties for non-expert users. One way to lower the learning overhead and make ontology queries more straightforward is through a Natural Language Interface (NLI). While there are existing NLIs to structured data with reasonable performance, they tend to require expensive customisation to each new domain. Additionally, they often require specific adherence to a pre-defined syntax which, in turn, means that users still have to undergo training.

We study the usability of NLIs from two perspectives: that of the developer who is customising the NLI system, and that of the end-user who uses it for querying. We investigate whether usability methods such as feedback and clarification dialogs can increase the usability for end users and reduce the customisation effort for the developers. To that end, we have developed two systems, QuestIO and FREyA.

QuestIO uses a very shallow NLP, but uses ontology extensively when translating the natural language into SeRQL/SPARQL. It performs well for very narrow domains and manually crafted ontologies on the set of simple questions including conjunction and disjunction. FREyA supports querying more ontologies at once, and does not rely on the ontology lexicalisations only, but develops mechanisms for extending the available lexicon through the interaction with users. It employs a semi-supervised learning approach in order to improve its performance with time.

Unlike QuestIO, which is fully automatic and does not give the opportunity to the user to validate the systems interpretations, in FREyA, the user is in control and can significantly influence the final result. When evaluated using the same dataset, FREyA outperformed all other similar systems (although it required dialog with the user).

Authors Danica Damljanovic is a Research Associate in the GATE team, Natural Language Processing (NLP) Group, Department of Computer Science, University of Sheffield. Her interests include applying NLP and Information Retrieval to semantic web applications, usability of Natural Language Interfaces, Question answering systems, and application of semantic web technologies to E-Tourism.
Intelligent Segmentation of Large Datasets

Matthias Assel
High Performance Computing Center Stuttgart (HLRS), University of Stuttgart,
Nobelstrasse 19 70569 Stuttgart, Germany
assel@hlrs.de

The large amount of digital information and the time needed to locate, access and process this data greatly impacts the work of scientists. Especially in the High Performance Computing (HPC) domain, data grows extremely rapidly and even faster than the bandwidth and the processing power of modern supercomputers. Implicitly and at the same time the size of single datasets increases irresistibly and obviously faster than networks and platforms can deal with. Today, datasets are already distributed across multiple storage points so as to distribute work load, but this makes access to it sometimes even more complicated and tedious. What is more, retrieval and analysis of data is getting more and more time and power consuming, for example applications that require only part of the data typically have to handle the full dataset first. For these reasons, significant improvements in the data management field are necessary to increase research productivity in solving complex scientific problems [1].

To approach an intelligent segmentation, a couple of issues have to be taken into account: 1) data must be equipped with additional metadata, in particular information about the data structure are essential to process the data at a semantic level 2) data and application context must be accessible from everywhere and at any time and 3) partitioning (clustering) algorithms must be able to efficiently process terabyte-scale datasets as well as consider the multidimensional dependencies between such metadata, the contextual information and the data itself [2].

In this research, the goal was to study existing clustering concepts and to come up with a new methodology to partition large datasets into smaller context- and content-aware blocks, which can then be processed independently and concurrently by different parts of the application. The main contribution of this work thereby consists in an incremental clustering algorithms that considers application-specific requirements (i.e., generic and specific dependencies between data and application logic) in order to extract particular segments dynamically at runtime and/or on a timely basis.

References

Interleaving reasoning and selection by knowledge summarization

Yan Wang\(^1\), Zhisheng Huang\(^2\), Yi Zeng\(^1\)

\(^1\) International WIC Institute, Beijing University of Technology, China

\(^2\) Department of Artificial Intelligence, Vrije University Amsterdam, The Netherlands

\texttt{wang.yan@emails.bjut.edu.cn, yizeng@bjut.edu.cn}

\texttt{huang@cs.vu.nl}

Abstract

Interleaving reasoning and selection by knowledge summarization is a method to deal with the scalability in large-scale data searches and reasoning. Many existing methods have been developed to improve searches on large-scale data, which require high performance hardware support (with emphasizes on Parallelization and Cloud architectures). Because the scale of the semantic data grows faster than the hardware performance, only rely on the improvements of hardware and architecture innovation will not always work, especially in resource limited environments. Interleaving reasoning and selection by knowledge summarization does not rely on high-performance hardware. Instead of a global check on all the data, this method only explores partial data by using heuristic strategies. To achieve this goal, some preprocess need to be done offline, including dividing and summarization. The dividing process cut the original dataset into small chunks, and the summarization process gives the heuristic information to measure how far the target chunks are. For each query, the heuristic information is calculated with a function to estimate the possibility of the answers appearing in a chunk. When a query comes, the heuristic information will be used to find the chunks with high possibility of partially answering the query. Then the query will be executed against the found chunks, in descending order of their possibility. Interleaving reasoning and selection does not require completeness, so searches based on knowledge summarization can stop at any time, and the most possible chunks will be selected first to answer the query. This method enables personal computer or even mobile devices have the ability to deal with large-scale data, and some experiments on the RDF version of the MEDLINE dataset has been done to prove the proposed method. The summarization is to describe the original dataset with a quite small size. It can be expressed as ontologies in different perspectives. With these ontologies, and other external ontologies related to a query, this method can process the query with reasoning features. Based on this method, a prototype system named Knowledge Intensive Summarization System (KISS) is developed and the system indicates that the proposed method is potentially effective.
Scalable and parallel reasoning in the Semantic Web

Jacopo Urbani

Department of Computer Science, Vrije Universiteit Amsterdam,
j.urbani@few.vu.nl

Abstract. The current state of the art regarding scalable reasoning consists of programs that run on a single machine. When the amount of data is too large, or the logic is too complex, the computational resources of a single machine are not enough. We propose a distributed approach that overcomes these limitations and we sketch a research methodology. A distributed approach is challenging because of the skew in data distribution and the difficulty in partitioning Semantic Web data. During the presentation, we will present initial results which are promising and suggest that the approach may be successful.

1 Problem statement

Most of the current reasoners are programs that are executed on a single machine. The scalability of these approaches is limited by the physical resources of the single machine. The size of the Semantic Web has grown to the point where this limitation notably affects the performance of the reasoners. Therefore, in order to realize the vision of Semantic Web, a scalable and efficient way to reason over an ever-growing amount of data is crucial.

The scalability is typically evaluated on two aspects: computational complexity (i.e. the ability to perform more complex tasks) and input size (i.e. the ability to process a larger input). It is essential that the reasoning process is scalable regarding both aspects. The research question we pose is:

We aim to research algorithms which allow complex and scalable reasoning over a large amount of data (billions of statements).

The reasoning results should be accessed with interactive queries. In this context we identify two subtasks: reasoning and querying. Depending on the type of reasoning, these two tasks can be either independent from each others or strongly interlinked. When possible, we intend to put more emphasis on the task of reasoning than on the task of querying, but both tasks are important because if the results of the reasoning are unavailable to the end user then the entire process becomes meaningless.
2 Proposed approach

Our purpose is to reason over a very large amount of data which cannot be handled by a single machine. To overcome this problem, we propose a distributed approach where the reasoning task is executed simultaneously on several independent machines. We assume that we have complete control of the machines in which the data is processed.

We will consider only monotonic rule-based reasoning. The motivation behind this choice lies on several considerations:

- in the Web, the data is distributed and we cannot retract existing facts;
- there are already some standardized rule sets (RDFS, OWL Horst, OWL 2 RL) that are widely used and which allow us to compare our work with already existing one;
- currently there is no distributed rule-based reasoner which implements a complex logic on a very large scale.

A distributed approach is potentially more scalable than a single machine approach because it can scale on two dimensions: the hardware and the number of the machines. However, it is more challenging to design because:

- In the Semantic Web, the data is not uniformly distributed, but instead there is an high data skew which generates load balancing problems;
- In rule-based reasoning the data must be joined together many times in order to find a possible derivation. Doing joins over distributed data is one of the most challenging tasks in a distributed environment.

Our approach aims to limit the exchange of the data (which is expensive) and try instead to move the computation (which is cheap) because rule based reasoning is mainly a data intensive task. There are several distributed programming models, like MapReduce [1] or Satin [5] which reduce the data communication and efficiently distribute the computation.

In some cases, there are additional technical problems introduced by a distributed approach. For example, the nodes must communicate to each other using an efficient protocol, otherwise the performance will suffer. For our work, we intend to use the Ibis framework [4] to ease the development of our approach. The Ibis framework offers many tools like IPL [2] or JavaGAT [3] which handles many technical aspects of the communication between the nodes and the heterogeneity of the system.

In our context, rule based reasoning can be applied either in a forward way or in a backward way. In forward reasoning the algorithm first materializes all the closure and then the data is queried in a database fashion. This approach is ideal when we have a dataset which does not change frequently and we need to query it extensively. It can be problematic if the closure is very large or if the data changes too frequently.

In backward reasoning the derivation is calculated on the fly when the data is queried. This approach is generally more complex and it makes the queries
much slower, however it has the advantage that it works if the data changes often or if the complete closure is too large to materialize.

Our purpose is to apply the current programming models to the different types of reasoning (forward, backward, or a combination of the two) and to research what model is the most efficient and under which conditions. In case none of the existing programming model suits well for our purpose, our research will try to design a new programming model which overcomes the limitations of the existing ones.

References

Contrastive Reasoning with Inconsistent Ontologies

Jun Fang$^1$ and Zhisheng Huang$^2$

$^1$ School of Automation, Northwestern Polytechnical University, China.
$^2$ Department of Computer Science, Vrije Universiteit Amsterdam, The Netherlands

Abstract. We present a framework for answering queries over inconsistent ontologies by using contrastive reasoning, the reasoning of contrasts which are expressed as contrary conjunctions like the word “but” in natural language.

We argue that contrastive answers are more informative for reasoning with inconsistent ontologies, as compared with the usual simple boolean answer, i.e., either “yes” or “no”. We propose a general approach to obtaining contrastive answers from inconsistent ontologies, including an algorithm for computing contrastive answers. The proposed framework has been implemented in the system CRION (Contrastive Reasoning with Inconsistent ONtologies) as a reasoning plug-in in the LarKC (Large Knowledge Collider) platform.

We report several experiments in which we apply the CRION system to some realistic ontologies. This evaluation shows that contrastive reasoning is a useful extension to the existing approaches of reasoning with inconsistent ontologies.

Brief Bio

Jun Fang is currently a lecture in the school of automation at Northwestern Polytechnical University, China. His current work focuses on Semantic Web and ontology reasoning, especially on scalable reasoning and reasoning with inconsistent ontologies.
Relation Prediction in Semantic Domains using Multivariate Prediction

Yi Huang and Volker Tresp

Siemens AG, Corporate Research and Technology, Munich, Germany

Abstract In many semantic domains data is often incomplete. A huge amount of relations and literal values of interest are missing. For instance, in a semantic social web, the income and interests of a particular person might be influence factors for recommending items to this person (e.g., movies), but this information is not always available. Another example are gene-disease databases, where many gene-disease relations haven’t been recorded as they have not been discovered yet, but they might be highly interesting for biological research. In this talk we explore the capability of machine learning to uncover potentially true relations and to predict missing values.

We have developed a machine learning approach using multivariate prediction, termed SUNS (statistical unit node set). First, in a given semantic domain, we define a clean statistical setting: a statistical unit and a population. Then, we transform data to a data matrix. The resulting data matrices are typically high-dimensional and sparse. Next, we apply multivariate prediction approaches on data matrices. In the talk we investigate four different multivariate prediction approaches based on matrix completion, i.e., reduced rank penalized regression (RRPP), singular value decomposition (SVD), non-negative matrix factorization (NNMF) and latent Dirichlet allocation (LDA). All approaches estimate unknown matrix entries via a low-rank matrix approximation. After matrix completion the entries are interpreted as certainty values that the corresponding triples are true.

We extend the approach to be able to define multi-population. Motivated by a Bayesian perspective we combine the prediction of the trained models based on each population. This extension makes the approach more suitable and more effective for learning in semantic domains due to the semantic nature.

In the talk we focus on an important life science problem, i.e., gene-disease-relationship. We apply the basic SUNS approach and the extended multi-population approach on various life science databases of the Linked Data. To our knowledge this is the first serious application of machine learning to the Linked Data. We show that in one side, the extended model gives better results than the basic model and on the other side, the SUNS approaches achieve compromising prediction accuracy.

Authors Yi Huang is a staff scientist at Siemens Corporate Research and Technology and is doing his PhD at Ludwig Maximilian University of Munich, Germany. His research interests focus on statistical machine learning, text mining, information retrieval, and the Semantic Web. Contact him at yihuang@siemens.com.

Volker Tresp is the head of a research team in machine learning at Siemens Corporate Research and Technology. His research interests include machine learning, data mining, information extraction and the semantic web. Tresp has a PhD from Yale University. Contact him at volker.tresp@siemens.com.
ANSWERS: A Novel Semantic Web-Enabled Recommender System

Daniele Dell’Aglio
CEFRIEL – ICT Institute Politecnico di Milano, Italy
daniele.dellaglio@cefriel.it

Abstract

Recommender Systems and Semantic Web are two research fields that acquired more and more relevance in the last few years.

On the one hand, Recommender System research is aimed to develop algorithms and instruments to identify and to select items of interest for users; in particular, Knowledge-based Recommender Systems work building knowledge base describing items and users and processing it in order to compute recommendations. The usual barrier of this approach is, therefore, the availability of the necessary knowledge and its maintenance over time.

On the other hand, Semantic Web community studies how to model, link and reason on big amounts of structured data. Semantic Web technologies can be of great help for Recommender Systems: not only knowledge technologies and languages can be employed to build the knowledge base, but the large availability of open and linked data about a growing variety of fields and topics, published on the Web of Data, further simplifies the modeling step for recommender systems.

In my talk I present my concept of Semantic Web-enabled Recommender System, based on the retrieval from the linked data Web of the necessary pieces of knowledge about items and users. I illustrate the general structure of this new family of Knowledge-based Recommender Systems and I explain how I concretely followed this approach to develop a tool to recommend Web services on top of the LarKC platform. I also offer some considerations about the strengths, the current limitations and the possible extensions of this proposal.

Author

Daniele Dell’Aglio has been working since 2008 as junior researcher at CEFRIEL. His research activities focus on Semantic Web technologies, recommender systems, location-aware and mobile applications. He is currently participating in the Urban Computing research activities of the LarKC FP7 project and in research activities related to Web service and recommender systems in the SOA4All FP7 project. In the past he participated in the completed Service-Finder FP7 project.