



LarKC

The Large Knowledge Collider

a platform for large scale integrated reasoning and Web-search

FP7 – 215535

D4.5.2 Implemented Plug-ins for Reasoning by Committee

Coordinator: Volker Tresp (Siemens)
With contributions from: Yi Huang (Siemens)
Quality Assessor: Florian Steinke (Siemens)
Quality Controller: Zhisheng Huang (VUA)

Document Identifier:	LarKC/2008/D4.5.2/Vx.x
Class Deliverable:	LarKC EU-IST-2008-215535
Version:	version 1.0
Date:	October 1, 2011
State:	final
Distribution:	public



EXECUTIVE SUMMARY

The previous deliverable D4.5.1 concerns the exploration of the applicability of committee machines to reasoning. It provides background material on committee machines and proposes ideas how to apply committee machines to LarKC reasoning. Based on those ideas we implement a plug-in for reasoning by committee, *CommitteeReasoner*. This deliverable documents the design and implementation of the plug-in. The plug-in is implemented on the LarKC platform v2.5.



DOCUMENT INFORMATION

IST Project Number	FP7 – 215535	Acronym	LarKC
Full Title	The Large Knowledge Collider: a platform for large scale integrated reasoning and Web-search		
Project URL	http://www.larkc.eu/		
Document URL			
EU Project Officer	Stefano Bertolo		

Deliverable	Number	4.5.2	Title	Implemented Plug-ins for Reasoning by Committee
Work Package	Number	4	Title	Reasoning and Deciding

Date of Delivery	Contractual	M42	Actual	30-September-11
Status	version 1.0		final	<input checked="" type="checkbox"/>
Nature	prototype <input checked="" type="checkbox"/> report <input type="checkbox"/> dissemination <input type="checkbox"/>			
Dissemination Level	public <input checked="" type="checkbox"/> consortium <input type="checkbox"/>			

Authors (Partner)	Volker Tresp (Siemens), Yi Huang (Siemens)			
Resp. Author	Volker Tresp (Siemens)		E-mail	volker.tresp@siemens.com
	Partner	Siemens	Phone	+49 (89) 63649408

Abstract (for dissemination)	The previous deliverable D4.5.1 concerns the exploration of the applicability of committee machines to reasoning. It provides background material on committee machines and proposes ideas how to apply committee machines to LarKC reasoning. Based on those ideas we implement a plug-in for reasoning by committee, <i>CommitteeReasoner</i> . This deliverable documents the design and implementation of the plug-in. The plug-in is implemented on the LarKC platform v2.5.
Keywords	Committee Machine, Reasoning

Version Log			
Issue Date	Rev No.	Author	Change



PROJECT CONSORTIUM INFORMATION
















Participant's name	Partner	Contact
Semantic Technology Institute Innsbruck, Universitaet Innsbruck	 	Dieter Fensel Semantic Technology Institute (STI), Universitaet Innsbruck, Innsbruck, Austria Email: dieter.fensel@sti-innsbruck.at
AstraZeneca AB		Bosse Andersson AstraZeneca Lund, Sweden Email: bo.h.andersson@astrazeneca.com
CEFRIEL - SOCIETA CONSORTILE A RESPONSABILITA LIMITATA		Emanuele Della Valle CEFRIEL - SOCIETA CONSORTILE A RESPONSABILITA LIMITATA Milano, Italy Email: emanuele.dellavalle@cefriel.it
CYCORP, RAZISKOVANJE IN EKSPERIMENTALNI RAZVOJ D.O.O.		Michael Witbrock CYCORP, RAZISKOVANJE IN EKSPERIMENTALNI RAZVOJ D.O.O., Ljubljana, Slovenia Email: witbrock@cyc.com
Höchstleistungsrechenzentrum, Universitaet Stuttgart		Georgina Gallizo Höchstleistungsrechenzentrum, Universitaet Stuttgart Stuttgart, Germany Email : gallizo@hlrs.de
MAX-PLANCK GESELLSCHAFT ZUR FOERDERUNG DER WISSENSCHAFTEN E.V.		Lael Schooler, Max-Planck-Institut für Bildungsforschung Berlin, Germany Email: schooler@mpib-berlin.mpg.de
Ontotext AD		Atanas Kiryakov, Ontotext Lab, Sofia, Bulgaria Email: naso@ontotext.com
SALTLUX INC.		Kono Kim SALTLUX INC Seoul, Korea Email: kono@saltlux.com
SIEMENS AKTIENGESELLSCHAFT		Volker Tresp SIEMENS AKTIENGESELLSCHAFT Muenchen, Germany Email: volker.tresp@siemens.com
THE UNIVERSITY OF SHEFFIELD		Hamish Cunningham, THE UNIVERSITY OF SHEFFIELD Sheffield, UK Email: h.cunningham@dcs.shef.ac.uk
VRIJE UNIVERSITEIT AMSTERDAM		Frank van Harmelen, VRIJE UNIVERSITEIT AMSTERDAM Amsterdam, Netherlands Email: Frank.van.Harmelen@cs.vu.nl
THE INTERNATIONAL WIC INSTITUTE, BEIJING UNIVERSITY OF TECHNOLOGY		Ning Zhong, THE INTERNATIONAL WIC INSTITUTE Mabeshi, Japan Email: zhong@maebashi-it.ac.jp
INTERNATIONAL AGENCY FOR RESEARCH ON CANCER		Paul Brennan, INTERNATIONAL AGENCY FOR RESEARCH ON CANCER Lyon, France Email: brennan@iarc.fr
INFORMATION RETRIEVAL FACILITY		John Tait, INFORMATION RETRIEVAL FACILITY Vienna, Austria Email: john.tait@ir-facility.org



TABLE OF CONTENTS

LIST OF TABLES	6
1 INTRODUCTION	7
2 THE COMMITTEEREASONER PLUG-IN	8
2.1 Theory of Approaches	8
2.2 Implementation	9
2.3 Parameters	9
3 CONCLUSIONS	11



LIST OF TABLES

2.1	The sense of the parameters α and β by especial values 0, 1 and 0.5. . .	9
-----	---	---



1. Introduction

In last years the idea of committee decision has been successfully demonstrated over and over again. An example is the winner of the NETFLIX challenge that combines different recommendation engines to form the best ranking. Another example is IBM's Watson applying a committee decision for achieving the answer.

Very often the data required to answer a query is distributed in several data sources. Querying information from the Linked Open Data (LOD) is a typical such situation. The standard approach is to attempt to interlink the different data bases such that queries can be applied to the joint domains. We are all aware of the problems and the complexity of this approach. In contrast, committee machines suggest a different approach:

Instead of combining the data by interlinking data sources, combine the knowledge returned in query results.

We assume that for each data source a SPARQL endpoint is defined. We apply a given SPARQL query to all SPARQL endpoints in parallel and then combine the results in form of a committee machine. For instance, if the query is “*Is Siemens the largest German electronic company?*”, different SPARQL endpoints might return different answers. The reason could be that they employ different reasoners, different ontologies or different instances. The term “largest” might mean the largest number of employees in an ontology, while in another ontology it might refer to the largest revenue or the largest market value. Similarly, we might consider companies which have businesses in Germany or companies with its headquarter in Germany. Moreover, data sources might be incomplete and ontological background rules might be faulty.

Deliverable D4.5.1 [1] provides an overview of committee machines. This helps us understand the background and the state of the art of committee machines. In D4.5.1 we also proposed two ideas how to apply reasoning by committee (see Chapter 5): causal independence committee and variance-based committee. Based on those ideas, we have developed a plug-in for reasoning by committee, called *CommitteeReasoner*, on the LarKC platform v2.5. This document describes the technical aspects of the plug-in, such as its input and output and how it works internally.



2. The CommitteeReasoner Plug-in

In this chapter we explore the technical aspects of the CommitteeReasoner plug-in¹. First, we discuss the two approaches implemented in the plug-in. Then we describe how the plug-in is implemented. Finally, we comment on the parameters set up for the plug-in.

2.1 Theory of Approaches

In the previous deliverable D4.5.1 [1] we suggested two approaches to reasoning by committee. In this section we briefly review these approaches.

For instance, we want to know whether the statement s such as “Gene G is associated with disease D ” is true or not. We assume that for doing that we need to query N data sources, where $N > 1$. So we apply a reasoner to each of these data sources. Note that the reasoners are not necessarily different. First, we define two parameters $0 < \alpha_i \leq 1$ and $0 \leq \beta_i < 1$ for reasoner $i \in \{1, ..N\}$. The first parameter α_i is a measure of the soundness of reasoner i , while $1 - \alpha_i$ stands for the unsoundness of the reasoner.² The second parameter β_i is a measure of the incompleteness of SPARQL endpoint i . Therefore, $1 - \beta_i$ represents the completeness of SPARQL endpoint i . From a probabilistic point of view, β_i means that even if in a SPARQL endpoint the reasoner cannot prove a statement to be true, this statement in fact would be true with the probability β_i . The parameters α_i and β_i reflex the subjective belief of the expert (who uses the plug-in) in the soundness and the incompleteness of the results generated by reasons, respectively. Table 2.1 explains intuitively what they mean by the especial values 0, 0.5 and 1.

We combine the results (true or unknown) returned by N reasoners in two ways:

- **Causal independence committee:**

We define $q_i = 1 - \alpha_i$, when the reasoner i returns true, while $q_i = 1 - \beta_i$, when the reasoner i cannot prove the validness of statement s and returns unknown.

The combination expression is

$$P(s = true) = 1 - \prod_i q_i$$

where P stands for the probability of the statement s being true. Note that if any reasoner is sound, then $P(s = true) = 1$.

- **Variance-based committee:**

We define $var_i^\alpha = \alpha_i(1 - \alpha_i)$ and $var_i^\beta = \beta_i(1 - \beta_i)$. We then define $c_i = \frac{1}{var_i^\alpha + \epsilon}$ and $f_i = \alpha_i$, if the reasoner i returns true; $c_i = \frac{1}{var_i^\beta + \epsilon}$ and $f_i = \beta_i$, if the reasoner

¹<http://wiki.larkc.eu/LarkcProject/MachineLearningPlugins?action=AttachFile&do=view&target=CommitteeReasoner.larkc>

²Here, we do not explicitly distinguish the soundness of reasoners from the soundness of data sources, more strictly speaking, from the soundness of the results inferred by reasoners based on “dirty” data sources. In LarKC we assume that all logic-based reasoners are sound which means that the soundness of reasoners is trivial. The unsoundness of data sources is exactly the reason for defining the parameter α . This unsoundness can be caused by the (partially) incorrect interpretation of ontologies, faulty rules and contradictory facts as described in the introduction.



α_i	$\rightarrow 0$	The results of reasoner i are mostly wrong.
	0.5	The soundness of the reasoner i is unknown, i.e., a random answer.
	1	The reasoner i is sound.
β_i	0	The data source i is complete.
	0.5	The completeness of the data source i is unknown, i.e., a random answer.
	$\rightarrow 1$	The data source i is mostly empty.

Table 2.1: The sense of the parameters α and β by especial values 0, 1 and 0.5.

i returns unknown. ϵ is a small positive number. Roughly speaking, var_i is the variance of i -th reasoning result. c_i stands for the confidence of the result. var_i and c_i are inverse proportional. c_i increases in two cases: 1. reasoner i tends to be sound or data source i is incomplete to a great extent; 2. reasoner i is rather unsound or data source i tends to be complete. f_i is equal to α resp. β and can be considered as a weighting factor emphasizing the first case.

The combination expression is

$$P(s = true) = \frac{1}{Z} \sum_i c_i f_i$$

where $Z = \sum_i c_i$ is a normalization factor.

Which of the two approaches should be chosen is dependent on the query. In general, the causal independent committee is more suitable for the query whose results are inferred by causality. For example, it should be applied in medical domains where there are many causal relationships among influence factors. The variance-based approach returns roughly the average certainty of the reasoners in committee. We advice that to figure out the preferred approach for a given domain one should try both approaches in empirical experiments.

In the plug-in we implemented both approaches. A parameter *approach* indicates which approach is applied (see 2.3).

2.2 Implementation

The input of the plug-in is an array of the reasoning results $R = [r_1, \dots, r_N]$ with the length of N . The value of the array elements r_i is either **true** or **unknown**, which means that the reasoner proves the given statement to be true or fails to prove respectively. Again, we assume that a statement can only be proven but cannot be disproven, since this is the situation considered in LarKC. It is worth noting that the order of this array must be the same as that of the parameter arrays α and β . Then, the plug-in applies a committee machine as described in 2.1. The outcome is the probability of the truth value of the statement.

2.3 Parameters

The CommitteeReasoner plug-in can be configured with the following parameters:



- **alpha** is an array and its elements α_i are the probabilities that if reasoner i proves a statement to be true, this statement in fact is true. For a sound reasoner, we obtain $\alpha_i = 1$.
- **beta** is an array and its elements β_i are the probabilities that if reasoner i cannot prove the truth of a statement, this statement in fact is true. For a complete reasoner, we obtain $\beta_i = 0$.
- **approach** indicates which committee decision is taken. Its values **causality** and **variance** stand for causal independence committee and variance-based committee, respectively.
- **N** is the number of reasoners. This should be equal to the length of **alpha** and **beta**.

The order of the parameters **alpha** and **beta** must be identical and must be the same as that of the array of the reasoning results that form the input to the `CommitteeReasoner`.

There are different options for determining the parameters α_i and β_i .

- α_i and β_i can be set equal for all i
- α_i and β_i might be constant for a given i (different competencies of SPARQL endpoints), regardless of the query
- α_i and β_i can be specific both to i and to the query (query specific competencies of SPARQL endpoints)

It is interesting if it would be possible to find good algorithms for specifying the competence of a query endpoint for a given query.

Listing 2.1: An example for parameter setting in a workflow

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix larkc: <http://larkc.eu/schema#> .
@prefix committee: <http://larkc.eu/plugin/committee#> .

_:reasoner a <urn:eu.larkc.plugin.Reasoner.CommitteeReasoner> .

_:reasoner larkc:hasParameter _:param .
_:param committee:N "2" .
_:param committee:approach "causality" .
_:param committee:alpha "0.8 0.9" .
_:param committee:beta "0.1 0.3" .

<urn:committee> a <urn:eu.larkc.endpoint.push> .
<urn:committee> larkc:links _:path .
_:path a larkc:Path .
_:path larkc:hasInput _:reasoner .
_:path larkc:hasOutput _:reasoner .
```



3. Conclusions

This deliverable addresses the implementation of a reasoner by committee *CommitteeReasoner*. In the plug-in we implemented two approaches suggested in the previous deliverable D4.5.1 [1]. In this document we claimed that committee machines are a different approach for handling queries over heterogenous data sources, compared to the data integration approaches which interlink the data sources. The document briefly reviews our proposed committee approaches and describes the implementation of the plug-in as well as its parameters. The plug-in usage is illustrated by a simple workflow example.

The application of committee machines for reasoning in LarKC is based on the assumption of unsoundness of the reasoners and of the incompleteness of the data sources. For a given query, the result returned by each reason, i.e., each element of the input array of the plug-in, is a single answer `true` or `unknown`. Examples are “*Is Siemens the largest German electronic company?*” or “*Is gene G associated with disease D?*” Interesting future work will be to extend the plug-in so that it can also deal with queries that return a list of results with corresponding confidences, for example for queries like “*Which are the largest 10 German electronic companies?*” or “*Which genes are mostly associated with disease D?*”



REFERENCES

- [1] D4.5.1 strategies and design for reasoning by committee. Technical report. http://www.larkc.eu/wp-content/uploads/2008/01/LarkC_D4.5.1-Strategies-%Design-for-reasoning-by-committee.pdf.