LNCS 3053

Christoph Bussler
John Davies
Dieter Fensel
Rudi Studer (Eds.)

The Semantic Web: Research and Applications

First European Semantic Web Symposium, ESWS 2004
Heraklion, Crete, Greece, May 2004
Proceedings

Springer

Christoph Bussler John Davies
Dieter Fensel Rudi Studer (Eds.)

The Semantic Web: Research and Applications

First European Semantic Web Symposium, ESWS 2004
Heraklion, Crete, Greece, May 10-12, 2004
Proceedings

Springer

Volume Editors

Christoph Bussler
Digital Enterprise Research Institute (DERI)
National University of Ireland
Galway, Ireland
E-mail: Chris.Bussler@DERI.ie

John Davies
British Telecommunications plc
Orion 5/12, Adastral Park
Ipswich IP5 3RE, UK
E-mail: john.nj.davies@bt.com

Dieter Fensel
Digital Enterprise Research Institute (DERI)
National University of Ireland
Galway, Ireland
University of Innsbruck, Austria
E-mail: dieter.fensel@deri.org

Rudi Studer
University of Karlsruhe
76128 Karlsruhe, Germany
E-mail: studer@aifb.uni-karlsruhe.de

Library of Congress Control Number: Applied for

CR Subject Classification (1998): H.4, H.3, C.2, H.5, I.2, K.4, D.2

ISSN 0302-9743
ISBN 3-540-21999-4 Springer-Verlag Berlin Heidelberg New York

Springer-Verlag is a part of Springer Science+Business Media

springeronline.com

Printed in Germany

Typesetting: Camera-ready by author, data conversion by Boller Mediendesign
Printed on acid-free paper SPIN: 11007722 06/3142 5 4 3 2 1 0

Preface

These proceedings contain the papers accepted for presentation at the First European Semantic Web Symposium (ESWS 2004) held on Crete, Greece, May 10–12, 2004.

Given its status as an inaugural event, the organizers were delighted to receive 79 high-quality submissions. Most papers were reviewed by at least three referees, with the review results coordinated by the academic and industrial track chairs. In total, 27 papers were accepted for the academic track and 6 papers were accepted for the industrial track. The papers span a wide range of topics from the Semantic Web area, from infrastructure and ontology engineering to applications.

The high quality of this symposium is due to the efforts of many people. Jos de Bruijn in particular worked hard in a number of areas, including submissions management, publicity and the poster program. We would also like to thank Martin Doerr for local arrangements, Johannes Breitfuss for the WWW site, the Program Committee and additional reviewers for their invaluable support and the sponsors for their financial support.

March 2004

Christoph Bussler
John Davies
Dieter Fensel
Rudi Studer

Organization

Organizing Committee

Academic Track:	Dieter Fensel (DERI, Austria)
	Rudi Studer (University of Karlsruhe, Germany)
Tutorial Program and Proceedings:	John Davies (BT, UK)
Industrial Track and Demo Chair:	Christoph Bussler (DERI, Ireland)
Local Arrangements and Posters:	Jos de Bruijn (DERI, Austria)
	Martin Doerr (ICS-FORTH, Greece)

Scientific Program Committee

Andreas Eberhart	University of Karlsruhe, Germany
Asun Gomez-Perez	Universidad Politécnica de Madrid, Spain
Carole Goble	University of Manchester, UK
Christine Golbreich	University of Rennes, France
Daniel Schwabe	PUC-Rio, Brazil
Deborah McGuinness	Stanford University, USA
Dunja Mladenic	J. Stefan Institute, Slovenia
Enrico Franconi	Free Univ. Bozen, Italy
Enrico Motta	The Open University, UK
Frank van Harmelen	Vrije Universiteit Amsterdam, The Netherlands
Guus Schreiber	Vrije Universiteit Amsterdam, The Netherlands
Hamish Cunningham	University of Sheffield, UK
Heiner Stuckenschmidt	Vrije Universiteit Amsterdam, The Netherlands
Ian Horrocks	University of Manchester, UK
Jerome Euzenat	INRIA, Rhône-Alpes, France
John Davies	BT, UK
Mark Musen	Stanford University, USA
Martin Doerr	ICS-FORTH, Greece
Michael Kifer	University at Stony Brook, USA
Nigel Shadbolt	University of Southampton, UK
Norman Sadeh	Carnegie Mellon University, USA
Raphael Volz	University of Karlsruhe, Germany
Richard Benjamins	iSOCO, Spain
Riichiro Mizoguchi	Osaka University, Japan
Robert Meersman	Vrije Universiteit Brussels, Belgium
Stefan Decker	DERI, Ireland
Stefanos Kollias	National Technical University of Athens, Greece
Steffen Staab	University of Karlsruhe, Germany
Wolfgang Nejdl	University of Hannover, Germany
Ying Ding	DERI, Austria
Yolanda Gil	ISI USC, USA
York Sure	University of Karlsruhe, Germany

Industrial Program Committee

Alain Leger	France Telecom, France
Alistair Duke	BT, UK
Arian Zwegers	European Commission, Brussels
David Trastour	HP, UK
Dean Allemang	Boston University, USA
Frank Leymann	IBM, Germany
Jürgen Angele	Ontoprise, Germany
Kim Elms	SAP, Australia
Leo Obrst	MITRE, USA
Mari Georges	ILOG, France
Mike Dean	BBN, USA
Naso Kiryakov	Sirma AI, Bulgaria
Peter Smolle	NetDynamics, Austria
Ralph Traphoener	Empolis, Germany
Satish Thatte	Microsoft, USA
Stuart Williams	HP, UK
Vipul Kashyap	National Library of Medicine, USA

Additional Reviewers

Arthur Stutt	The Open University, UK
Atanas Kiryakov	Sirma AI, Bulgaria
Bo Hu	University of Southampton, UK
Daniel Oberle	University of Karlsruhe, Germany
Daniele Turi	University of Manchester, UK
Dmitry Tsarkov	University of Manchester, UK
Donovan Artz	ISI – USC, USA
Dumitru Roman	DERI, Ireland
Erik van Mulligen	Vrije Universiteit Amsterdam, The Netherlands
Farshad Hakimpor	The Open University, UK
Francis Kwong	University of Manchester, UK
Gary Wills	University of Southampton, UK
Holger Wache	Vrije Universiteit Amsterdam , The Netherlands
Hongsuda Tangmunarunkit	University of Manchester, UK
Janez Brank	J. Stefan Institute, Slovenia
Jeen Broekstra	Vrije Universiteit Amsterdam, The Netherlands
Jeff Z Pan	University of Manchester, UK
Jesús Barrasa Rodríguez	Universidad Politécnica de Madrid, Spain
Jos de Bruijn	DERI, Austria
Lei Li	University of Manchester, UK
Marc Spraragen	ISI – USC, USA
Maria Vargas-Vera	The Open University, UK
Mariano Fernandez Lopez	Universidad Politécnica de Madrid, Spain
Martin Dzbor	The Open University, UK
Matthew Horridge	University of Manchester, UK

Michael Stollberg	DERI, Austria
Mithun Sheshagiri	Carnegie Mellon University, USA
Mitsuru Ikeda	Osaka University, Japan
Nicholas Gibbins	University of Southampton, UK
Ning Zhang	University of Manchester, UK
Olivier Brunet	INRIA Rhône-Alpes, France
Oscar Corcho	Universidad Politécnica de Madrid, Spain
Peter Spyns	Vrije Universiteit Brussel, Belgium
Philip Cimiano	University of Karlsruhe, Germany
Raphaël Troncy	INRIA Rhône-Alpes, France
Ronald Ashri	University of Southampton, UK
Ronny Siebes	Vrije Universiteit Amsterdam, The Netherlands
Rubén Lara	DERI, Austria
Sudhir Agarwal	University of Karlsruhe, Germany
Terry Payne	University of Southampton, Germany
Uwe Keller	DERI, Austria
Y. Kitamura	Osaka University, Japan
Yannis Kalfoglou	University of Southampton, UK

Table of Contents

Ontology Engineering

Ontology Matching and Mapping

Ontology-Based Querying

Ontology Merging and Population

Infrastructure

Semantic Web Services

Service Discovery and Composition

Data for the Semantic Web

Knowledge Representation

Applications

Content Management

Information Management and Integration

Towards On-the-Fly Ontology Construction - Focusing on Ontology Quality Improvement

Naoki Sugiura[1], Yoshihiro Shigeta[1], Naoki Fukuta[1], Noriaki Izumi[2], and Takahira Yamaguchi[1]

[1] Shizuoka University, 3-5-1 Johoku, Hamamatsu, Shizuoka 432-8011, Japan,
sugiura@ks.cs.inf.shizuoka.ac.jp,
http://mmm.semanticweb.org
[2] National Institute of AIST, 2-41-6, Aomi, Koto-ku, Tokyo, Japan

Abstract. In order to realize the on-the-fly ontology construction for the Semantic Web, this paper proposes DODDLE-R, a support environment for user-centered ontology development. It consists of two main parts: pre-processing part and quality improvement part. Pre-processing part generates a prototype ontology semi-automatically, and quality improvement part supports the refinement of it interactively. As we believe that careful construction of ontologies from preliminary phase is more efficient than attempting generate ontologies full-automatically (it may cause too many modification by hand), quality improvement part plays significant role in DODDLE-R. Through interactive support for improving the quality of prototype ontology, OWL-Lite level ontology, which consists of taxonomic relationships (class - sub class relationship) and non-taxonomic relationships (defined as property), is constructed efficiently.

1 Introduction

As the scale of the Web becomes huge, it is becoming more difficult to find appropriate information on it. When a user uses a search engine, there are many Web pages or Web services which are syntactically matched with user's input words but semantically incorrect and not suitable for user's intention. In order to defeat this situation, Semantic Web[1] is now gathering attentions from researchers in wide area. Adding semantics (meta-data) to the Web contents, software agents are able to understand and even infer Web resources. To realize such paradigm, the role of ontologies[2][3] is important in terms of sharing common understanding among both people and software agents[4]. On the one hand, in knowledge engineering field ontologies have been developed for particular knowledge system mainly to reuse domain knowledge. On the other hand, for the Semantic Web, ontologies are constructed in distributed places or domain, and then mapped each other. For this purpose, it is an urgent task to realize a software environment for rapid construction of ontologies for each domain. Towards the on-the-fly ontology construction, many researches are focusing on

J. Davies et al. (Eds.): ESWS 2004, LNCS 3053, pp. 1–15, 2004.

automatic ontology construction from existing Web resources, such as dictionaries, by machine processing with concept extraction algorithms. However, even if the machine produces ontologies automatically, users still need to check the output ontology. It may be a great burden for users to check all the correctness of the ontology and modify it, especially if the scale of automatically produced ontology is large. Considering such situation, we believe that the most important aspect of the on-the-fly ontology construction is that how efficiently the user, such as domain experts, are able to check the output ontology in order to make Semantic Web contents available to the public. For this reason, ontologies should be constructed not fully automatically, but through interactive support by software environment from the early stage of ontology construction. Although it may seem to be contradiction in terms of efficiency, the total cost of ontology construction would become less than automatic construction because if the ontology is constructed with careful interaction between the system and the user, less miss-construction will be happened. It also means that high-quality ontology would be constructed. In this paper, we propose a software environment for user-centered on-the-fly ontology construction named DODDLE-R (Domain Ontology rapiD DeveLopment Environment - RDF[5] extension). The architecture of DODDLE-R is re-designed based on DODDLE-II [6], the former version of DODDLE-R. Although DODDLE-II has already provided interactive support for ontology construction, the system architecture is not well-considered and sophisticated. The DODDLE-R system is modularized into machine-processing module and user-interaction module in order to separate pre-processing part and user-centered quality management part specifically. Especially, to realize the user-centered environment, DODDLE-R dedicates to the quality improvement part. It enables us to develop ontologies with interactive indication of which part of ontology should be modified. The system supports the construction of both taxonomic relationships and non-taxonomic relationships in ontologies. Additionally, because DODDLE-II has been built for ontology construction not for the Semantic Web but for typical knowledge systems, it needs some extensions for the Semantic Web such as OWL (Web Ontology Language) [7] import and export facility. DODDLE-R supports OWL-Lite level ontology construction because if we think of user-centered ontology construction, OWL-DL or OWL-Full sounds too complicated for human to understand thoroughly. DODDLE-R contributes the evolution of ontology construction and the Semantic Web.

2 System Design of DODDLE-R

Fig. 1 shows the overview of DODDLE-R. The main feature of DODDLE-R is the modularized two parts - pre-processing part and quality improvement part. In pre-processing part, the system generates the basis of the ontology, a taxonomy and extracted concept pairs, by referring to WordNet[8] as an MRD (Machine Readable Dictionary) and domain specific text corpus. A taxonomy is a hierarchy of IS-A relationship. Concept pairs are extracted based on cooccurrency by using statistic methods. These pairs are the candidates which has

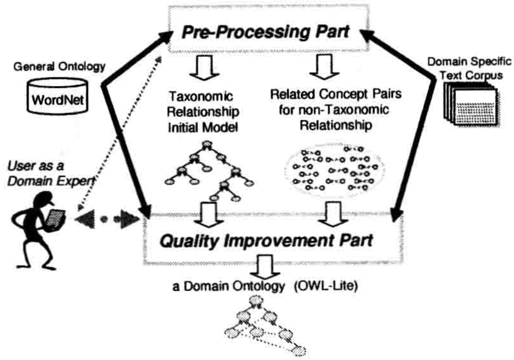


Fig. 1. DODDLE-R overview

significant relationships. A user identifies some relationship between concepts in the pairs. In quality improvement part, the prototype ontology produced by pre-processing part is modified by a user through interactive support by the system.

2.1 Pre-processing Part

In pre-processing part, the system generates the basis of output ontology for further modification by a user. Fig. 2 describes the procedure of pre-processing part. This part consists of three sub-parts: input concept selection, taxonomy building, and related concept pair acquisition. First, as input of the system, several domain specific terms are selected by a user. The system shows a list of noun concepts in the domain specific text corpus as candidates of input concept. At this phase, a user also identifies the sense of terms to map those terms to concepts in WordNet.

For building taxonomic relationship (class - sub class relationship) of an ontology, the system attempts to extract "best-matched concepts". That is, "concept matching" between input concepts and WordNet concepts is done, and matched nodes are extracted, and then merged at each root nodes. To extract related concept pairs from domain specific text corpus as a basis of identifying non-taxonomic relationships (such as "part-of" relationship), statistic methods are applied. In particular, WordSpace[9] and an association rule algorithm[10] are used in this part and these methods attempt to identify significantly related concept pairs.

Construction of WordSpace WordSpace is constructed as shown in Fig.3.
1. Extraction of high-frequency 4-grams Since letter-by-letter co-occurrence information becomes too much and so often irrelevant, we take term-by-term co-occurrence information in four words (4-gram) as the primitive to make up co-

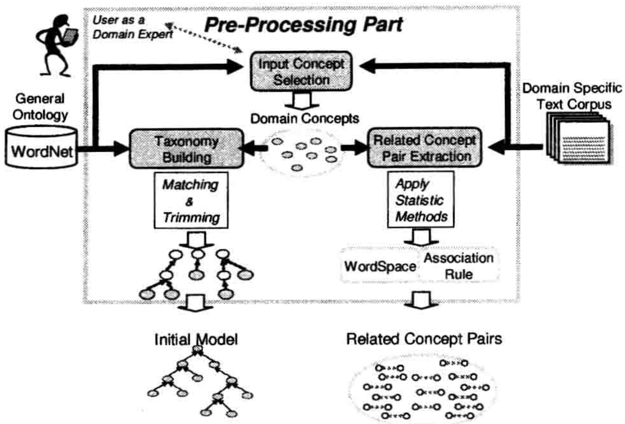


Fig. 2. Pre-processing Part

occurrence matrix useful to represent context of a text based on experimented results. We take high frequency 4-grams in order to make up WordSpace.

2. Construction of collocation matrix A *collocation matrix* is constructed in order to compare the context of two 4-grams. Element $a_{i,j}$ in this matrix is the number of 4-gram f_i which comes up just before 4-gram f_j (called *collocation area*). The collocation matrix counts how many other 4-grams come up before the target 4-gram. Each column of this matrix is the *4-gram vector* of the 4-gram f.

3. Construction of context vectors A *context vector* represents context of a word or phrase in a text. A sum of 4-gram vectors around appearance place of a word or phrase (called *context area*) is a context vector of a word or phrase in the place.

4. Construction of word vectors A word vector is a sum of context vectors at all appearance places of a word or phrase within texts, and can be expressed with Eq.1. Here, $\tau(w)$ is a vector representation of a word or phrase w, $C(w)$ is appearance places of a word or phrase w in a text, and $\varphi(f)$ is a 4-gram vector of a 4-gram f. A set of vector $\tau(w)$ is WordSpace.

$$\tau(w) = \sum_{i \in C(w)} \Big(\sum_{f \text{ close to } i} \varphi(f) \Big) \tag{1}$$

5. Construction of vector representations of all concepts The best matched "synset" of each input terms in WordNet is already specified, and a sum of the word vector contained in these synsets is set to the vector representation of a concept corresponding to a input term. The concept label is the input term.

6. Construction of a set of similar concept pairs Vector representations of all concepts are obtained by constructing WordSpace. Similarity between concepts is obtained from inner products in all the combination of these vectors. Then we

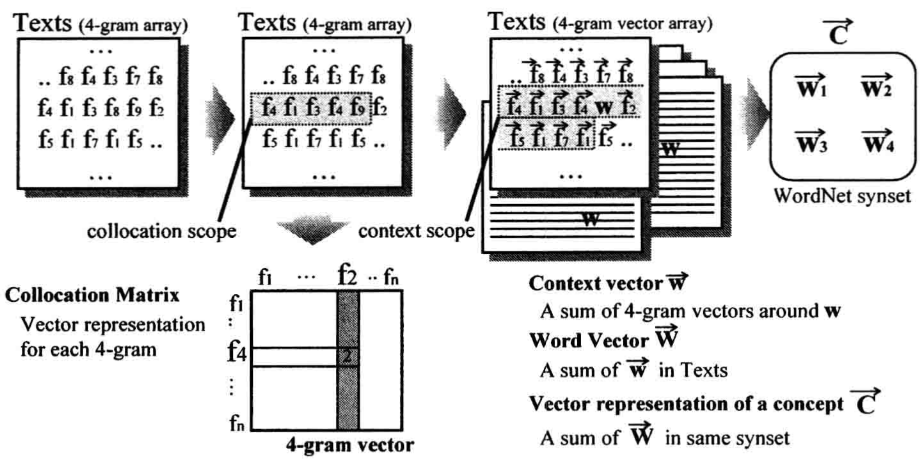


Fig. 3. Construction flow of WordSpace

define certain threshold for this similarity. A concept pair with similarity beyond the threshold is extracted as a similar concept pair.

Finding Association Rules between Input Terms The basic association rule algorithm is provided with a set of transactions, $T := \{t_i \mid i = 1..n\}$,where each transaction t_i consists of a set of items, $t_i = \{a_{i,j} \mid j = 1..m_i, a_{i,j} \in C\}$ and each item $a_{i,j}$ is form a set of concepts C. The algorithm finds association rules $X_k \Rightarrow Y_k : (X_k, Y_k \subset C, X_k \cap Y_k = \{\})$ such that measures for support and confidence exceed user-defined thresholds. Thereby, support of a rule $X_k \Rightarrow Y_k$ is the percentage of transactions that contain $X_k \cup Y_k$ as a subset (Eq.2)and confidence for the rule is defined as the percentage of transactions that Y_k is seen when X_k appears in a transaction (Eq.3).

$$support(X_k \Rightarrow Y_k) = \frac{\mid \{t_i \mid X_k \cup Y_k \subseteq t_i\} \mid}{n} \tag{2}$$

$$confidence(X_k \Rightarrow Y_k) = \frac{\mid \{t_i \mid X_k \cup Y_k \subseteq t_i\} \mid}{\mid \{t_i \mid X_k \subseteq t_i\} \mid} \tag{3}$$

As we regard input terms as items and sentences in text corpus as transactions, DODDLE-R finds associations between terms in text corpus. Based on experimented results, we define the threshold of support as 0.4% and the threshold of confidence as 80%. When an association rule between terms exceeds both thresholds, the pair of terms are extracted as candidates for non-taxonomic relationships.

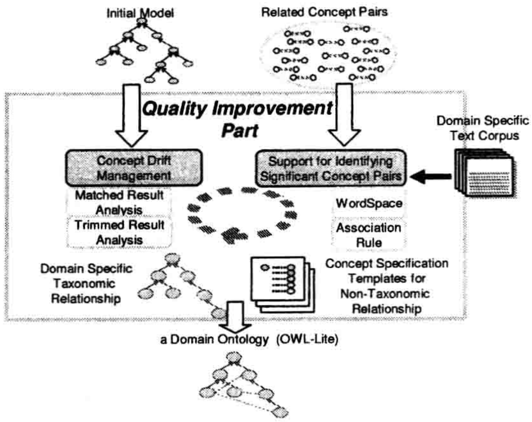
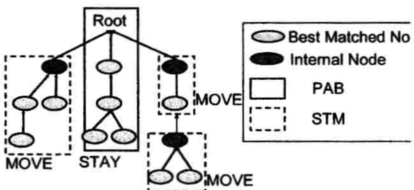


Fig. 4. Quality improvement part



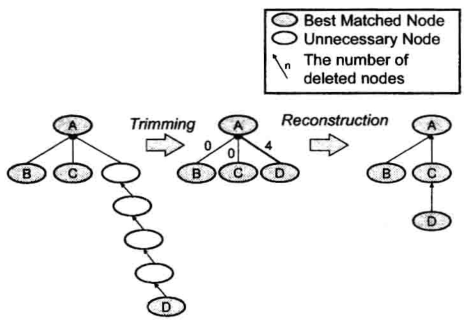


Fig. 5. Matched Result Analysis **Fig. 6.** Trimmed Result Analysis

2.2 Quality Improvement Part

In order to improve the quality of the pre-processed ontology, the quality improvement part works interactively with a user. Fig. 4 shows the procedure of this part. Because the pre-processed taxonomy is constructed from a general ontology, we need to adjust the taxonomy to the specific domain considering an issue called Concept Drift. It means that the position of particular concepts changes depending on the domain. For concept drift management, DODDLE-R applies two strategies: Matched Result Analysis (Fig. 5) and Trimmed Result Analysis (Fig. 6).

In Matched Result Analysis, the system divides the taxonomy into PABs (PAths including only Best matched concepts) and STMs (SubTrees that includes best-matched concepts and other concepts and so can be Moved) and indicates on the screen. PABs are paths that include only best-matched concepts that have senses suitable for the given domain. STMs are subtrees of which root is an internal concept of WordNet and its subordinates are all best-matched concepts. Because the sense of an internal concept has not been identified by a user yet, STMs may be moved to other places for the concept adjustment to the domain. In addition, for Trimmed Result Analysis, the system counts the number of internal concepts when the part was trimmed. By considering this number as the original distance between those two concepts, the system indicates to move the lower concept to other places.

As a facility for related concept pair discovery, there are functions that allow users to attempt some ways to improve the quality of extracted concept pairs through trial and error by changing parameters of statistic methods. Users can re-adjust the parameters of WordSpace and association rule algorithm and check the result. After that, the system generates "Concept Specification Templates" from by using the results. It consists of some concept pairs which have considerable relationship considering the result value of statistic methods.

By referring to the constructed domain specific taxonomic relationship and the "Concept Specification Templates", a user develops a domain ontology.

3 Implementation

In this section, we describe the system architecture from the aspect of system implementation. DODDLE-R support environment for ontology construction is realized in conjunction with MR^3 (Meta-Model Management based on RDF(S)[11] Revision Reflection) [12]. MR^3 is an RDF(S) graphical editor with meta-model management facility such as consistency checking of classes and a model in which these classes are used as the type of instances. Fig. 7 shows the relationship between DODDLE-R and MR^3 in terms of system implementation. Both MR^3 and DODDLE-R are implemented in Java language (works on Java 2 or higher). MR^3 is implemented using JGraph[13] for RDF(S) graph visualization, and Jena 2 Semantic Web Framework[14] for enabling the use of

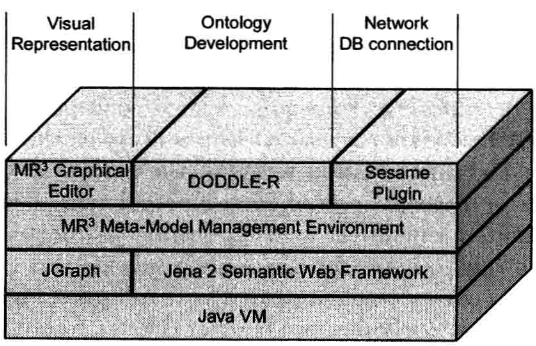


Fig. 7. DODDLE-R architecture