Report
on
ACM SIGIR Workshop on "Semantic Web"
SWIR 2003
Toronto, Canada
August 1, 2003

Ying Ding[#], Cornelis J. van Rijsbergen[*], Iadh Ounis[*], Joemon Jose[*]

[#] University of Innsbruck, Austria
[*] University of Glasgow, UK

1. Introduction

This was the first workshop combining semantic web and information retrieval (IR). So far, the two communities were working independently, whereas they address more or less common issues. The main aim of the workshop was to bring together these two disparate communities.

The workshop has been subsidized by the OntoWeb Consortium (www.ontoweb.org, EU Thematic Network on Semantic Web). We have received 15 papers in total. We had experts representing both the semantic web and the IR fields as reviewers. All the submitted papers were thoroughly reviewed by at least 4 people representing both fields. Based on these reviews, we selected 7 very good papers dealing with various aspects of both information retrieval and semantic web. In addition, we scheduled two invited talks, spanning two important aspects of the semantic web. Dr Arturo Trujillo, of the Canon Research Centre Europe Ltd, has kindly chaired the workshop. The workshop attracted an audience of more than 30 people, who were encouraged to interact with the authors and the invited speakers.

2. Purpose & Motivation

The emergence of the Internet has brewed the revolution of information storage and retrieval. Tons of web pages are being appended to the Internet on a daily basis, and this causes to explode the amount of available information on the Web dramatically. These dynamic data strongly hinder the efficiency and the effectiveness of current searching facilities. Although various indexing, cataloguing and search systems are easily accessible from the web, their functions to retrieve relevant information and manage knowledge are still very limited. Therefore, effective or intelligent search for information on this massive information resource becomes highly critical.

Tim Berners-Lee created the vision of a semantic web that enables automated information access and use, based on machine-processable semantics of data. In his informal 'Semantic Web Road

Map' note[1], he outlined possible future directions for the evolution of the World Wide Web. These ideas have been met with growing enthusiasm of researchers and developers world-wide, both in academia and in industry. They encourage the integration of ongoing efforts from different disciplines, involving specialists in Information Retrieval, Natural Language Processing, Information Extraction, Knowledge Representation, Artificial Intelligence and Databases. These efforts aim at capturing the semantics of digital content of all sorts and origins, and making current Web to be easily accessed and precisely retrieved.

Information retrieval can benefit from building ontologies and other semantic structures, making it possible to have a better understanding of the application domains and the user queries. Indeed, semantic web technologies could be a basis for intelligent retrieval. On the other hand, the information retrieval field has a lot to bring to the semantic web community, based on the 30-years of research and development in the context of very large collections of documents. This workshop aimed to bring researchers from the two communities together.

3. Background Information on Semantic Web

The web was initially designed for direct human processing. With its current structure, machine-based approaches to web applications are not possible unless its content is transformed into a machine-readable format encoding semantic information. Ontologies are the backbone technology for the semantic web and - more generally - for the management of formalized knowledge within the technical context of distributed systems. They provide machine-processable semantics of data and information sources that can be communicated between different agents (software and people).

The semantic web has to go beyond the simple bag-of-words approach currently used by search engines and get closer to the meaning of the texts. One of the most important things missing in current web technologies is a link between the texts and external lexical resources encoding semantic information. The following common problems can be easily foreseen in conventional keyword-based information retrieval systems:

- **Searching information**: Existing keyword-based search retrieves irrelevant information due to term ambiguity, and misses information when material related to similar concepts is stored under quite different terms.
- **Extracting information**: Currently, people have to browse and read extensively in order to extract relevant information from textual or other representations. Software agents do not possess the common-sense knowledge required to assist effectively in tasks of this type, let alone automate them. Moreover, they fail to integrate information from different sources.
- **Maintaining** large repositories of weakly structured text is a difficult and time-consuming activity.
- **Adaptation** and dynamic reconfiguration of information repositories (e.g. websites) according to user profiles or other aspects of relevance, hinges on automatic document generation and is not yet fully mastered.

Ontologies that provide shared, common domain theories will be a key enabler to tackle these problems. Ontologies can be regarded as metadata that explicitly represent the semantics of data

[1] http://www.w3.org/DesignIssues/Semantic.html

for machine processing. Once the data on the web is understandable and represented in a machine-processable format, further information, not necessarily explicit in the text, can be gathered via ontologies. This would provide the basis for extracting information, reasoning around the stored data, answering questions, and other applications. By making explicit the link between the form and the content of information, ontologies help people and computers access the information they need, and effectively communicate with each other. They have a crucial role in enabling content-based *access*, *interoperability*, and *communication* across the web.

Semantic web technologies and especially the use of ontologies, are expected to enable a much higher degree of automation and scalability in performing operations pertaining to the above mentioned tasks. For instance, in order to keep weakly structured collections consistent, or to generate information presentations from semi-structured data, the semantics of these collections and data must not only be machine-accessible but also machine-processable. In other words, the semantics must be represented based on formal ontologies.

Semantic web will provide intelligent access to heterogeneous, distributed information, enabling software products to mediate between user needs and the information sources available. Web services can be accessed and executed via the web. However, all these service descriptions are based on semi-formal natural language descriptions. Therefore, the human programmer needs to be kept in the loop and the scalability as well as economy of web services are limited. Bringing them to their full potential requires their combination with semantic web technology. This technology will provide mechanization in service identification, configuration, comparison and combination**.** Semantic web enabled web services have the potential to change our life to a much higher degree than the current web already has. Semantic retrieval to support web services will be the promising vision in the foreseeable future.

4. Topics

Topics of this workshop are centred on semantic web and information retrieval and how to combine both in order to achieve mutual benefit and high impact. The workshop aimed to address the following areas: Ontology-based Information Retrieval or Semantic Information Retrieval; Metadata in Information Retrieval; Ontology Learning based on Information Extraction and Natural Language Processing Technologies; Automatic Indexing and Cataloguing; Information Retrieval and Web Services; Languages, Tools and Methodologies for Semantic Annotations; Knowledge Portals; Semantic Web Mining; Semantic Web Searching and Querying; Semantic Web for Multimedia Retrieval; Semantic Web for Multilingual Information Retrieval; Semantic Web for Digital Library; Semantic Interoperability and Integration; Semantic Web and Peer-to-Peer; Semantic Web for Information Visualization; Peer-to-Peer for Information Retrieval; User Studies for Semantic Web; Business Applications and Best of Practice.

5. Presentation Summaries

As mentioned in the introduction, the event included the schedule of two invited talks:

- A talk by Atanas Kiryakov from OntoText, Bulgaria. OntoText is a company that has successfully combined natural language processing with semantic web technologies. Atnas Kiryakov gave a talk on Semantic indexing and retrieval. He presented a vision about a holistic system allowing annotation, indexing, and retrieval of documents with respect to real-world entities. An operational system based on this, called KIM, has been developed. KIM shows the possibility to enhance current information extraction and information retrieval system by using semantic web technologies.

- Dr Christoph Bussler from Digital Enterprise Research Institute (DERI), National University of Ireland, Galway, is the renowned expert in the semantic web services area. He gave a contribution on message mediation in composite web services. Web services powered by semantic web technology are the pioneering and challenging applications in the semantic web area. This contribution was meant to introduce us to this trendy application.

Furthermore, the event included the presentation of the following 7 papers:

J. Mayfield and T. Finin (The Johns Hopkins University, University of Maryland, USA): *Information Retrieval on the Semantic Web: Integrating inference and retrieval.*

The authors discussed some very interesting issues regarding the combination of information retrieval techniques with current semantic web technologies. In a nutshell, the authors aim to adopt the inference mechanism currently used in semantic web field to information retrieval, in order to improve retrieval performance.

A. Ankolekar, Y.W. Seo and K. Sycara (Carnegie Mellon University, USA): *Investigating Semantic Knowledge for Text Learning*

The authors explore the hypothesis that the use of an ontology can be useful to extract features for document clustering. They compared two methods of statistical feature extraction against manually created ontologies for a text classification task. Their experiments on a standard text collection (Reuters-21578 database), considered as a reference in the domain of text classification, show a limited advantage of ontology-based classification as compared to standard approaches.

A. Hotho, S. Staab and G. Stumme (University of Karlsruhe, Germany): *Wordnet improves Text Document Clustering*

Hotho and al. investigate and evaluate different strategies for using ontological information, derived from WordNet, to supplement document-clustering algorithms. It is shown how the representation of documents is enriched to incorporate related concepts obtained from WordNet. These representations are then used to form cluster of documents by means of a variant of the KMeans clustering algorithm. Various ways of using the information from WordNet for document expansion are evaluated. The outcome of the experiments shows some promising improvements from the use of such expressive representations.

C. Ciorascu, I. Ciorascu and K. Stoffel (Université de Neuchâtel, Switzerland): *Knowler-Ontological Support for Information Retrieval*

The authors presented an information retrieval system, which allows complex queries that require searches through WordNet. It proposes the use of the ontology language (OWL) for interchanging semantic information in large-scale IR systems. The authors have built an OWL ontology, which includes knowledge from Wordnet, stemming information and a whole document base. Experiments on time-performance access to the ontology are reported. The evaluation shows that a very good scalability is achieved by using persistency at both storage and reasoning levels.

P.H. Meland, J. Austvik and J. Heggland (Norwegian University of Science and Technology, Norway): *Using Ontologies and Semantic Networks with Temporal Media*

The authors discussed the annotation of video and audio data, describing a system for annotating and retrieving such data based on RDF ontologies. The most recent standards and approaches related to both multimedia and semantic web knowledge representation are properly introduced and discussed. Authors show how their system integrates and uses this knowledge.

J. Klavans, S.D. Popper and R. Passonneau (Columbia University, USA): *Tackling the Internet Glossary Glut: Automatic Extraction and Evaluation of Genus Phrases*

This work discussed the automatic evaluation of a method for extraction of term definitions, based on a gold standard created through human annotation of test data. The authors briefly described a natural language processing system that extracts genus phrases from Internet documents. The method has immediate applications in the automatic construction of thesauri, dictionaries and ontologies. Furthermore, the authors discuss a few insights in the problems entailed by the development of a gold standard for genus phrases.

C. Brewster, F. Ciravegna and Y. Wilks (University of Sheffield, UK): *Background and Foreground Knowledge in Dynamic Ontology Construction: Viewing Text as Knowledge Maintenance*

The authors identified important issues around the topic of learning ontologies/taxonomies from document sets. They provided a nice introduction regarding the relations between text, backgound and foreground knowledge. Within this framework they present and discuss a possible approach for ontology extraction from text, supported by external - richer and better-structured - sources of knowledge. In particular, authors discussed several issues regarding the (im)possibilities of identifying background knowledge. This contribution was backed up by an interesting, informal testing (correctly identified by the authors).

6. Conclusions

This workshop was a pleasant experience with the participation of people from two different research fields. The fusion of these communities sparked off strong intellectual debate and interactions. This encouraged participants to continue their research addressing both the topics. We believe there was a strong interest in this SWIR 2003 workshop particularly given the adverse circumstances prevalent at the time of paper submission. This suggests there is a case for continuous effort along these lines.