

Online and Distance Learning: Concepts, Methodologies, Tools, and Applications

Lawrence Tomei
Robert Morris University, USA



INFORMATION SCIENCE REFERENCE

Hershey • New York

Assistant Executive Editor: Meg Stocking
Acquisitions Editor: Kristin Klinger
Development Editor: Kristin Roth
Senior Managing Editor: Jennifer Neidig
Managing Editor: Sara Reed
Typesetter: Sharon Berger, Jennifer Neidig, Sara Reed, Laurie Ridge, Jamie Snavelly, Michael Brehm,
Elizabeth Duke, and Diane Huskinson
Cover Design: Lisa Tosheff
Printed at: Yurchak Printing Inc.

Published in the United States of America by
Information Science Reference (an imprint of IGI Global)
701 E. Chocolate Avenue, Suite 200
Hershey PA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@igi-pub.com
Web site: <http://www.igi-pub.com/reference>

and in the United Kingdom by
Information Science Reference (an imprint of IGI Global)
3 Henrietta Street
Covent Garden
London WC2E 8LU
Tel: 44 20 7240 0856
Fax: 44 20 7379 0609
Web site: <http://www.eurospanonline.com>

Copyright © 2008 by IGI Global. All rights reserved. No part of this publication may be reproduced, stored or distributed in any form or by any means, electronic or mechanical, including photocopying, without written permission from the publisher.

Product or company names used in this set are for identification purposes only. Inclusion of the names of the products or companies does not indicate a claim of ownership by IGI Global of the trademark or registered trademark.

Library of Congress Cataloging-in-Publication Data

British Cataloguing in Publication Data

A Cataloguing in Publication record for this book is available from the British Library.

All work contributed to this encyclopedia set is new, previously-unpublished material. The views expressed in this encyclopedia set are those of the authors, but not necessarily of the publisher.

Chapter 2.8

Semantics for the Semantic Web: The Implicit, the Formal, and the Powerful

Amit Sheth

University of Georgia, USA

Cartic Ramakrishnan

University of Georgia, USA

Christopher Thomas

University of Georgia, USA

ABSTRACT

Enabling applications that exploit heterogeneous data in the Semantic Web will require us to harness a broad variety of semantics. Considering the role of semantics in a number of research areas in computer science, we organize semantics in three forms—implicit, formal, and powerful—and explore their roles in enabling some of the key capabilities related to the Semantic Web. The central message of this chapter is that building the Semantic Web purely on description logics will artificially limit its potential, and that we will need to both exploit well-known techniques

that support implicit semantics, and develop more powerful semantic techniques.

INTRODUCTION

Semantics has been a part of several scientific disciplines, both in the realm of Computer Science and outside of it. Research areas such as information retrieval (IR), information extraction (IE), computational linguistics (CL), knowledge representation (KR) artificial intelligence (AI), and data(base) management (DB) have all addressed issues pertaining to semantics in their own ways.

Most of these areas have very different views of what “meaning” is, and these views are all built on some meta-theoretical and epistemological assumptions. These different views imply very different views of cognition, of concepts, and of meaning (Hjørland, 1998). In this chapter, we organize these different views to three forms of semantics: implicit, formal, and powerful (a.k.a. soft). We use these forms to explore the role of semantics that go beyond the narrower interpretation of the Semantic Web (that involve adherence to contemporary Semantic Web standards) and encompass those required for a broad variety of semantic applications. We advocate that for the Semantic Web (SW) to be realized, we must harness the power of a broad variety of semantics encompassing all three forms.

IR, IE, and CL techniques primarily draw upon analysis of unstructured texts in addition to document repositories that have a loosely defined and less formal structure. In these sorts of data sources, the semantics are *implicit*.

In the fields of KR, AI, and DB, however, the data representation takes a more formal and/or rigid form. Well-defined syntactic structures are used to represent information or knowledge where these structures have definite semantic interpretations associated with them. There are also definite rules of syntax that govern the ways in which syntactic structures can be combined to represent the meaning of complex syntactic structures. In other words, techniques used in these fields rely on *formal semantics*.

Usually, efforts related to formal semantics have involved limiting expressiveness to allow for acceptable computational characteristics. Since most KR mechanisms and the relational data model are based on set theory, the ability to represent and utilize knowledge that is imprecise, uncertain, partially true, and approximate is lacking, at least in the base/standard models. However, there have been several efforts to extend the base models (e.g., Barbara, Garcia-Molina, & Porter, 1992). Representing and utilizing these types

of more powerful knowledge is, in our opinion, critical to the success of the Semantic Web. Soft computing has explored these types of powerful semantics. We deem these *powerful (soft)* semantics as distinguished, albeit not distinct from or orthogonal to *formal* and *implicit semantics*.

More recently, semantics has been driving the next generation of the Web as the Semantic Web, where the focus is on the role of semantics for automated approaches to exploiting Web resources. This involves two well-recognized, critical enabling capabilities—ontology generation (Maedche & Staab, 2001; Omelayenko, 2001) and automated resource annotation (Hammond, Sheth, & Kochut, 2002; Dill et al., 2003; Handschuh, Staab, & Ciravegna, 2002; Patil, Oundhakar, Sheth, & Verma, 2004), which should be complemented by an appropriate computational approach such as reasoning or query processing. We use a couple of such enabling capabilities to explore the role and importance of all three forms of semantics.

A majority of the attention in the Semantic Web has been centered on a logic-based approach, more specifically that of description logic. However, looking at past applications of semantics, it is very likely that more will be expected from the Semantic Web than what the careful compromise of expressiveness and computability represented by description logic and the W3C adopted ontology representation language OWL (even its three flavors) can support. Supporting expressiveness that meet requirements of practical applications and the techniques that support their development is crucial. It is not desirable to limit the Semantic Web to one type of representation where expressiveness has been compromised at the expense of computational property such as decidability.

This chapter is not the first to make this above observation. We specifically identify a few. Uschold (2003) has discussed a semantic continuum involving informal to formal and implicit to explicit, and Gruber (2003) has talked about informal, semi-formal, and formal ontolo-

gies. The way we use the term *implicit semantics*, however, is different compared to Uschold (2003) insofar as we see implicit semantics in all kinds of data sets, not only in language. We assume that machines can analyze implicit semantics with several, mostly statistical, techniques. Woods has written extensively regarding the limitations of first-order logics (FOLs)—and hence description logics, or DLs—in the context of natural language understanding, although limitations emanating from rigidity and limitation of expressive power, as well as limited value reasoning supported in DLs, can also be identified:

Over time, many people have responded to the need for increased rigor in knowledge representation by turning to first-order logic as a semantic criterion. This is distressing, since it is already clear that first-order logic is insufficient to deal with many semantic problems inherent in understanding natural language as well as the semantic requirements of a reasoning system for an intelligent agent using knowledge to interact with the world. (Woods, 2004)

We also recall Zadeh's long-standing work (such as Zadeh, 2002), in which he extensively discussed the need for what constitutes a key part of the "powerful semantics" here. In essence, we hope to provide an integrated and complementary view on the range of options. One may ask what the uses of each of these types of semantics are in the context of the Semantic Web. Here is a quick take.

- *Implicit semantics* is either largely present in most resources on the Web or can easily (quickly) be extracted. Hence mining and learning algorithms applied to these resources can be utilized to extract structured knowledge or enrich existing structured formal representations. Since formal semantics intrinsically does not exist, implicit semantics is useful in processing data sets or

corpus to obtain or bootstrap semantics that can be then represented in formal languages, potentially with human involvement.

- *Formal semantics* in the form of ontologies is relatively scarce, but representation mechanisms with such semantics have definite semantic interpretations that make them more machine-processable. Representation mechanisms with formal semantics therefore afford applications the luxury of automated reasoning, making the applications more intelligent.
- *Powerful (soft) semantics* in the form of fuzzy or probabilistic KR mechanisms attempt to overcome the shortcomings of the rigid set-based interpretations associated with currently prevalent representation mechanisms by allowing for representation of degree of membership and degree of certainty. Some of the domain knowledge human experts possess is intrinsically complex, and may require these more expressive representations and associated computational techniques.

These uses are further exemplified later on using Semantic Web applications as driving examples. In the next section we define and describe *implicit, formal and powerful (soft) semantics*.

TYPES OF SEMANTICS

In this section we give an overview of the three types of semantics mentioned. It is rather informal in nature, as we only give a broad overview without getting in depth about the various formalisms or methods used. We assume that the reader is somewhat familiar with statistical methods on the one hand and description logics/OWL on the other. We present a view of these methods in order to lead towards the necessity of powerful (soft) semantics.

Implicit Semantics

This type of semantics refers to the kind that is implicit from the patterns in data and that is not represented explicitly in any strict machine processable syntax. Examples of this sort of semantics are the kind implied in the following scenarios:

- Co-occurrence of documents or terms in the same cluster after a clustering process based on some similarity measure is completed.
- A document linked to another document via a hyperlink, potentially associating semantic metadata describing the concepts that relate the two documents.
- The sort of semantics implied by two documents belonging to categories that are siblings of each other in a concept hierarchy.
- Automatic classification of a document to broadly indicate what a document is about with respect to a chosen taxonomy. Further, use the implied semantics to disambiguate (does the word “palm” in a document refer to a palm tree, the palm of your hand, or a palm-top computer?).
- Bioinformatics applications that exploit patterns like sequence alignment, secondary and tertiary protein structure analysis, and so forth

One may argue that although there is no strict syntactic and explicit representation, the knowledge about patterns in data may yet be machine processable. For instance, it is possible to get a numeric similarity judgment between documents in a corpus. Although this is possible, this is the only sort of processing possible. It is not possible to look at documents and automatically infer the presence of a named relationship between concepts in the documents.

Even though the exploitation of implicit semantics draws upon well-known statistical techniques,

the wording is not a mere euphemism, but meant to give a different perception of the problem.

Many tools and applications for implicit semantics have been developed for decades and are readily available. Basically all machine learning exploits implicit semantics, namely clustering, concept and rule learning, hidden Markov models, artificial neural networks, and others. These techniques supporting implicit semantics are found in early steps towards the Semantic Web, such as clustering in the Vivisimo search engine, as well as in early Semantic Web products, such as metadata extraction on Web Fountain technology (Dill et al., 2003), automatic classification, and automatic metadata extraction in Semagix Freedom (Sheth et al., 2002).

Formal Semantics

Humans communicate mostly through language. Natural language, however, is inherently ambiguous—semantically, but also syntactically. Computers lack the ability to disambiguate and understand complex natural language. For these reasons, it is infeasible to use natural language as a means for machines to communicate with other machines. As a first step, statements or facts need to be expressed in a way that computers can process them. Semantics that are represented in some well-formed syntactic form (governed by syntax rules) is referred to as *formal semantics*. There are some necessary and sufficient features that make a language formal and by association their semantics formal. These features include:

- **The notions of model and model theoretic semantics:** Expressions in a formal language are *interpreted* in *models*. The structure common to all models in which a given language is interpreted (the *model structure* for the model-theoretic interpretation of the given language) reflects certain basic presuppositions about the “structure of the world” that are implicit in the language.

- **The principle of compositionality:** The meaning of an expression is a function of the meanings of its parts and of the way they are syntactically combined. In other words, the semantics of an expression is computed using the semantics of its parts, obtained using an interpretation function.

From a less technical perspective, formal semantics means machine-processable semantics where the formal language representing the semantics has the above-mentioned features. Basically, the semantics of a statement are unambiguously expressed in the syntax of the statement in the formal language. A very limited subset of natural language is thus made available for computer processing. Examples of such semantics are:

- The semantics of subsumption in description logics, reflecting the human tendency of categorizing by means of broader or narrower descriptions.
- The semantics of partonomy, accounting for what is part of an object, not which category the object belongs to.

Description Logics

Recently, description logics have been the dominant formalisms for knowledge representation. Although DLs have gained substantial popularity, there are some fundamental properties of DLs that can be seen as drawbacks when viewed in the context of the Semantic Web and its future. The *formal semantics* of DLs is based on set theory. A concept in description logics is interpreted as a set of things that share one required common feature. Relationships between concepts or roles are interpreted as a subset of the cross-product of the domain of interpretation. This leaves no scope for the representation of degrees of concept membership or uncertainty associated with concept membership.

DL-based representation and reasoning for both schema and instance data is being applied in Network Inference's Cerebra product for such problems as data integration. This product uses a highly optimized tableaux algorithm to speed up ABox reasoning, which was the bane of description logics. Although a favorable trade-off between computational complexity and expressive power has been achieved, there is still the fundamental issue of the inability of DLs to allow for representation of fuzzy and probabilistic knowledge.

Powerful (Soft) Semantics

The statistical analysis of data allows the exploration of relationships that are not explicitly stated. Statistical techniques give us great insight into a corpus of documents or a large collection of data in general, when a program exists that can actually "pose the right questions to the data," that is, analyze the data according to our needs. All derived relationships are statistical in nature, and we only have an idea or a likelihood of their validity.

The above-mentioned formal knowledge representation techniques give us certainty that the derived knowledge is correct, provided the explicitly stated knowledge was correct in the first place. Deduction is truth preserving. Another positive aspect of a formal representation is its universal usability. Every system that adheres to a certain representation of knowledge will understand, and a well-founded formal semantics guarantees that the expressed statements are interpreted the same way on every system. The restriction of expressiveness to a subset of FOL also allows the system to verify the consistency of its knowledge.

But here also lies the crux of this approach. Even though it is desirable to have a consistent knowledge base, it becomes impractical as the size of the knowledge base increases or as knowledge from many sources is added. It is rare that human experts in most scientific domains have

a full and complete agreement. In these cases it becomes more desirable that the system can deal with inconsistencies.

Sometimes it is useful to look at a knowledge base as a map. This map can be partitioned according to different criteria, for example, the source of the facts or their domain. While on such a map the knowledge is usually locally consistent, it is almost impossible and practically infeasible to maintain a global consistency. Experience in developing the Cyc ontology demonstrated this challenge. Hence, a system must be able to identify sources of inconsistency and deal with contradicting statements in such a way that it can still produce derivations that are reliable.

In the traditional bivalent-logic-based formalisms, we—that is, the users or the systems—have to make a decision. Once two contradictory statements are identified, one has to be chosen as the right one. While this is possible in domains that are axiomatized, fully explored, or in which statements are true by definition, it is not possible for most scientific domains. In the life sciences, for instance, hypotheses have to be evaluated, contradicting statements have promoting data, and so forth. Decisions have to be deferred until enough data is available that either verifies or falsifies the hypothesis. Nevertheless, it is desirable to express these hypotheses formally to have means to computationally evaluate them on the one hand and to exchange them between different systems on the other.

In order to allow the sort of reasoning that would allow this, the expressiveness of the formalism needs to be increased. It is known that increasing the expressive power of a KR language causes problems relating to computability. This has been the main reason for limiting the expressive power of KR languages. The real power behind human reasoning, however, is the ability to do so in the face of imprecision, uncertainty, inconsistencies, partial truth, and approximation. There have been attempts made in the past at building KR languages that allow such expressive power.

Major approaches to reasoning with imprecision are: (1) probabilistic reasoning, (2) possibilistic reasoning (Dubois, Lang, & Prade, 1994), and (3) fuzzy reasoning. Zadeh (2002) proposed a formalism that combines fuzzy logic with probabilistic reasoning to exploit the merits of both approaches. Other formalisms have focused on resolving local inconsistencies in knowledge bases, for instance the works of Blair, Kifer, Lukasiewicz, Subrahmanian, and others in annotated logic and paraconsistent logic (see Kifer & Subrahmanian, 1992; Blair & Subrahmanian, 1989). Lukasiewicz (2004) proposes a weak probabilistic logic and addresses the problem of inheritance. Cao (2000) proposed an annotated fuzzy logic approach that is able to handle inconsistencies and imprecision; Straccia (e.g., 1998, 2004) has done extensive work on fuzzy description logics. With P-CLASSIC, Koller, Levi, and Peffer (1997) presented an early approach to probabilistic description logics implemented in Bayesian Networks. Other probabilistic description logics have been proposed by Heinsohn (1994) and Jaeger (1994). Early research on Bayesian-style inference on OWL was done by Ding and Peng (2004). In her formalism, OWL is augmented to represent prior probabilities. However, the problem of inconsistencies arising through inheritance of probability values (see Lukasiewicz, 2004) is not taken into account.

The combination of probabilistic and fuzzy knowledge under one representation mechanism proposed in Zadeh (2002) appears to be a very promising approach. Zadeh argues that fuzzy logics and probability theory are “complementary rather than competitive.” Under the assumption that humans tend to linguistically categorize a continuous world into discrete classes, but in fact still perceive it as continuous, fuzzy set theory classifies objects into sets with fuzzy boundaries and gives objects degrees of set membership in different sets. Hence it is a way of dealing with a multitude of sets in a computationally tractable way that also follows the human perception of the

world. Fuzzy logic allows us to blur artificially imposed boundaries between different sets. The other powerful tool in soft computing is probabilistic reasoning. Definitely in the absence of complete knowledge of a domain and probably even in its presence, there is a degree of uncertainty or randomness in the ways we see real-world entities interact. OWL as a description language is meant to explicitly represent knowledge and to deductively derive implicit knowledge. In order to use a similar formalism as a basis for tools that help in the derivation of new knowledge, we need to give this formalism the ability to be used in abductive or inductive reasoning. Bayesian-type reasoning is a way to do abduction in a logically feasible way by virtue of applying probabilities. In order to use these mechanisms, the chosen formalism needs to express probabilities in a meaningful way, that is, a reasoner must be able to meaningfully interpret the probabilistic relationships between classes and between instances. The same holds for the representation of fuzziness. The formalism must give a way of defining classes by their membership functions.

A major drawback of logics dealing with uncertainties is the required assignment of prior probabilities and/or fuzzy membership functions. Obviously, there are two ways of doing that—manual assignment by domain experts and automatic assignment using techniques such as machine learning. Manual assignments require the domain expert to assign these values to every class and every relationship. This assignment will be arbitrary, even if the expert has profound knowledge of the domain. Automatic assignments of prior values require a large and representative dataset of annotated instances, and finding or agreeing on what is a representative set is difficult or at times impossible. Annotating instances instead of categorizing them in a top-down approach is tedious and time consuming. Often, however, the probability values for relationships can be obtained from the dataset using statistical

methods, thus we categorize these relationships as implicit semantics.

Another major problem here is that machine learning usually deals with flat categories rather than with hierarchical categorizations. Algorithms that take these hierarchies into account need to be developed. Such an algorithm needs to change the prior values of the superclasses according to the changes in the subclasses, when necessary. Most likely, the best way will be a combination of both, when the domain expert assigns prior values that have to be validated and refined using a testing set from the available data.

In the end, powerful semantics will combine the benefits of both worlds: hierarchical composition of knowledge and statistical analysis; reasoning on available information, but with the advantage over statistical methods that it can be formalized in a common language and that general purpose reasoners can utilize it, and with the advantage over traditional formal DL representation that it allows abduction as well as induction in addition to deduction.

It might be argued that more powerful formalisms are already under development, such as SWRL (Straccia, 1998), which works on top of OWL. These languages extend OWL by a function-free subset of first-order logics, allowing the definition of new rules in the form of Horn clauses. The paradigm is still that of bivalent FOLs, and the lack of function symbols makes it impossible to define functions that can compute probability values. Furthermore, SWRL is undecidable. We believe that abilities to express probabilities and fuzzy membership functions, as well as to cope with inconsistencies, are important. It is desirable (and some would say necessary) that the inference mechanism is sound and complete with respect to the semantics of the formalism and the language is decidable. Straccia (1998) proves this for a restricted fuzzy DL; Giugno and Lukasiewicz (2002) prove soundness and completeness for the probabilistic description logic formalism P-SHOQ(D).

So far, this powerful semantic and soft computing research has not been utilized in the context of developing the Semantic Web. In our opinion, for this vision to become a reality, it will be necessary to go beyond RDFS and OWL, and work towards standardized formalisms that support powerful semantics.

CORRELATING SEMANTIC CAPABILITIES WITH TYPES OF SEMANTICS

Building practical Semantic Web applications (e.g., see TopQuadrant, 2004; Sheth & Ramakrishnan, 2003; Kashyap & Shklar, 2002) require certain core capabilities. A quick look at these core capabilities reveals a sequence of steps towards building such an application. We group this sequence into two categories as shown in Table 1 and identify the type of semantics utilized by each.

APPLICATIONS AND TYPES OF SEMANTICS THEY EXPLOIT

In this section we describe some research fields and some specific applications in each field. This list is by no means a comprehensive list, but rather samples of some research areas that attempt solve problems that are crucial to realizing the Semantic Web vision. We cover *information integration*, *information extraction/retrieval*, *data mining*, and *analytical applications*. We also discuss *entity identification/disambiguation* in some detail. We associate with each of the techniques in these research areas one or more of the types of semantics we identified earlier.

Information Integration

There is, now more than ever, a growing need for several information systems to interoperate

in a seamless manner. This sort of interoperation requires that the syntactic, structural, and semantic heterogeneities (Hammer & McLeod, 1993; Kashyap & Sheth, 1996) between such information systems be resolved. Resolving such heterogeneities has been the focus of a lot of work in schema integration in the past. With the recent interest in the Semantic Web, there has been a renewed interest in resolving such heterogeneities. A survey of schema matching techniques (Rahm & Bernstein, 2001) identifies a wide variety of techniques that are deployed to solve this problem.

Schema Integration

A look at the leaf nodes and the level immediately above it, in the classification tree of schema matching techniques in Rahm and Bernstein (2001), reveals the combination of the technique used and the type of information about the schema used for matching schemas. Depending on whether the schema or the instances are used to determine the match, the type of information harnessed varies. Our aim is to associate one or more types of semantics (from our classification) with each of the bulleted entries at the leaf nodes of the tree shown. Table 1 does just that.

Entity Identification/Disambiguation (EI/D)

A much harder, yet fundamental (and related) problem is that of *entity identification/disambiguation*. This is the problem of identifying that two entities are in fact either the same but treated as being different or that they are in fact two different entities that are being treated as one entity. Techniques used for *identification/disambiguation* vary widely depending on the nature of the data being used in the process. If the application uses unstructured text as a data source, then the techniques used for EI/D will rely on *implicit semantics*. On the other hand, if EI/D is being

attempted on semi-structured data, the application can, for instance, disambiguate entities based on the properties they have. This implies harnessing the power of *formal* or *semi-formal semantics*. As listed in Table 1, the constraint-based techniques are ideally suited for use in EI/D when semi-structured data is being used. Dealing with unstructured data will require the use of the techniques listed under linguistic techniques.

Information Retrieval and Information Extraction

Let us consider information retrieval applications and the types of data they exploit. Given a

request for information by the user, information retrieval applications have the task of processing unstructured (text corpus) or loosely connected documents (hyperlinked Web pages) to answer the “query.” There are various flavors of such applications.

Search Engines

Search engines exploit both the content of Web documents and the structure *implicit* from the hyperlinks connecting one document to the other. Kleinberg (1998) defines the notions of hubs and authorities in a hyperlinked environment. These notions are crucial to the structural analysis and

Table 1. Some key semantic capabilities and the type of semantics exploited

	Capabilities	Implicit Semantics	Formal Semantics	Possible Use of Powerful (Soft) Semantics
Bootstrapping Phase (building phase)	Building ontologies either automatically or semi-automatically	Analyzing word co-occurrence patterns in text to learn taxonomies/ontologies (Kashyap et al., 2003)		Using fuzzy or probabilistic clustering to learn taxonomic structures or ontologies
	Annotation of unstructured content wrt. these ontologies (resulting in semantic metadata)	Analyzing word occurrence patterns or hyperlink structures to associate concept names from and ontology with both resources and links between them (Naing, Lim, & Goh, 2002)		Using fuzzy or probabilistic clustering to learn taxonomic structures or ontologies OR Using fuzzy ontologies
	Entity disambiguation	Using clustering techniques or support vector machines (SVMs) for entity disambiguation (Han, Giles, Zha, Li, & Tsioutsoulis, 2004)	Using an ontology for entity disambiguation	Using fuzzy KR mechanisms to represent ontologies that may be used for disambiguation
	Semantic integration of different schemas and ontologies	Analyzing the extension of the ontologies to integrate them (Wang, Wen, Lochovsky, & Ma, 2004)	Schema-based integration techniques (Castano, Antonellis, & Vimercati, 2001)	
	Semantic metadata enrichment (further enriching the existing metadata)	Analyzing annotated resources in conjunction with an ontology to enhance semantic metadata (Hammond et al., 2002)		This enrichment could possibly mean annotating with fuzzy ontologies

continued on following page

Table 1. continued

	Capabilities	Implicit Semantics	Formal Semantics	Possible Use of Powerful (Soft) Semantics
Utilization Phase	Complex query processing		Hypothesis validation queries (Sheth, Thacker, & Patel, 2003) or path queries (Anyanwu & Sheth, 2002)	
	Question answering (QA) systems ¹	Word frequency and other CL techniques to analyze both the question and answer sources (Ramakrishnan, Chakrabarti, Paranjpe, & Bhattacharya, 2004)	Using <i>formal</i> ontologies for QA (Atzeni et al., 2004)	Providing confidence levels in answers based on fuzzy concepts or probabilistic
	Concept-based search ¹	Analyzing occurrence of words that are associated with a concept, in resources	Using hypernymy, partonymy, and hyponymy to improve search (Townley, 2000)	
	Connection and pattern explorer ¹	Analyzing semi-structured data stores to extract patterns (technique in Kuramochi & Karypis, 2004, applied to RDF graphs)	Using ontologies to extract patterns that are meaningful (Aleman-Meza, Halaschek, & Sahoo, 2003)	
	Context-aware retriever ¹	Word frequency and other CL techniques to analyze resources that match the search phrase	Using formal ontologies to enhance retrieval	Using fuzzy KR mechanisms to represent context
Utilization Phase	Dynamic user interfaces ¹		Using ontologies to dynamically reconfigure user interfaces (Quan & Karger, 2004)	
	Interest-based content delivery ¹	Analyzing content to identify concept of content so as to match with interest profile	User profile will have ontology associated with it which contains concepts of interest	
	Navigational and research (Guha, McCool, & Miller, 2003) search	Navigational searches will need to analyze unstructured content	Discovery style queries (Anyanwu & Sheth, 2002) on semi-structured data which is a combination of implicit and formal semantics	Fuzzy matches for research search results

Table 2. Techniques used for schema integration and the type of semantics they exploit

	Type of Information Used	What Does it Mean?	Types of Semantics Exploited
Linguistic Techniques	Name Similarity	Using canonical name representations, synonymy, hypernymy, string edit distance, pronunciation, and N-gram-like techniques to match schemas' attribute and relation names	<i>Implicit Semantics</i> are exploited by string edit distance, pronunciation, and N-gram-like techniques. <i>Formal Semantics</i> are exploited by synonymy, etc.
	Description Similarity	Processing natural language descriptions associated with attributes and relations	<i>Implicit Semantics</i> are exploited by the NLP techniques deployed.
	Word Frequencies of Key Terms	Using relative frequencies of keywords and word combinations at the instance level	<i>Implicit Semantics</i>
Constraint Based Techniques	Type Similarity	Using information about data types of attributes as an indicator of a match between schemas	<i>Formal Semantics</i>
	Key Properties	Using foreign keys, part-of relationships, and other constraints	<i>Formal Semantics</i>
	Graph Matching	Treating the structure of schemas as graph algorithms to determine match degree; between graphs are used to match schemas.	Combination of <i>Implicit</i> and <i>Formal Semantics</i>
	Value Patterns and Ranges	Using ranges of attributes and patterns in the value of attributes as an indicator of similarity between the corresponding schemas	<i>Implicit Semantics</i>

the eventual indexing of the Web. A modification of this approach aimed at achieving scalability is used by Google (Brin & Page, 1998). Google has fairly good precision and recall statistics. However, the demands that the Semantic Web places on search engine technology will mean that future search engines will have to deal with information requests that are far more demanding. Guha et al. (2003) identify two kinds of searches:

- **Navigational searches:** In this class of searches, the user provides the search engine with a phrase or combination of words which s/he expects to find in the documents. There is no straightforward, reasonable interpretation of these words as denoting a concept.

In such cases, the user is using the search engine as a navigation tool to navigate to a particular intended document. Using the domain knowledge as specified in relevant domain ontology can enable an improved semantic search (Townley, 2000).

- **Research searches:** In many other cases, the user provides the search engine with a phrase that is intended to denote an object about which the user is trying to gather/research information. There is no particular document that the user knows about that s/he is trying to get to. Rather, the user is trying to locate a number of documents, which together will give her/him the information s/he is trying to find.

We believe that research searches will require a combination of *implicit semantics*, *formal semantics*, and what we refer to as *powerful semantics*.

Question Answering Systems

Question answering systems can be viewed as more advanced and more “intelligent” search engines. Current question-answering systems (Brin & Page, 1998; Etzioni et al., 2004; Ramakrishnan et al., 2004) use Natural Language Processing (NLP) and pattern matching techniques to analyze both the question asked of the system and the potential sources of the answers. By comparing the results of these analyses, such systems attempt to match portions of the sources of the answer (for instance, Web pages) with the question, thereby answering them. Such systems therefore still use data like unstructured text and attempt to extract information from it. In other words the semantics are *implicit* in the text and are extracted from this text. To facilitate question answering, Zadeh (2003) proposes the use of an epistemic lexicon of world knowledge, which would be represented by a weighted graph of objects with uncertain attributes; in our terminology this is the equivalent of an ontology using *powerful semantics*.

Data Mining

The goal of data mining applications is to find non-trivial patterns in unstructured and structured data.

Clustering

Clustering is defined as the process of grouping *similar* entities or objects together in groups based on some notion of similarity. Clustering is considered a form of *unsupervised learning*. The applications of clustering use a given similarity metric and, as a result of the grouping of data

points into clusters, attempt to use this information (*implicit semantics*) to learn something about the interactions between the clustered entities. The sort of information sought from the clustered data points may range from simple similarity judgments as in query-by-example (QBE) document retrieval systems or systems aimed at extracting *more formal* semantics from the underlying data, as is the aim of semi-automatic taxonomy generation.

Semi-Automatic Taxonomy Generation (ATG)

As described in Kashyap et al. (2003), the aim of Automated Taxonomy Generation is to hierarchically cluster a document corpus and extract from the resulting hierarchy of clusters a *sequence* of clusters that best captures all the levels of specificity/generality in the corpus, where this sequence is ordered by the value of the specificity/generality measure. This is then followed by a node label extraction phase, where each cluster in the sequence is analyzed to extract from it a set of labels that best captures the topic its documents represent. These sets of labels are then pruned to reduce the number of potential labels for nodes in the final output hierarchy.

Association Rule Mining

An example of an association rule is given in Agrawal, Imielinski, and Swami (1993) and Agrawal and Srikant (1994) as follows: 90% of the transactions in a transaction database that involve the purchase of bread and butter together also have the purchase of milk involved. This is an example of an application where occurrence patterns of attribute values in a relational database (*implicit semantics*) are converted in association rules (*formal semantics*).

Analytical Applications

These come under the purview of applications that support complex query processing. It would be reasonable to hypothesize that search engines of the future will be required to answer analytical or discovery style queries (Guha et al., 2003; Anyanwu & Sheth, 2002). This is in sharp contrast to the kinds of information requests today's search engines have to deal with, where the focus is on retrieving resources from the Web that may contain information about the desired keyword. In this current scenario the user is left to sift through vast collections of documents and further analyze the returned results. In addition to querying data from the Web, future search engines will also have to query vast metadata repositories. We discuss one such technique in the following section.

Complex Relationship Discovery

As described in Anyanwu and Sheth (2002):

Semantic Associations capture complex relationships between entities involving sequences of predicates, and sets of predicate sequences that interact in complex ways. Since the predicates are semantic metadata extracted from heterogeneous multi-source documents, this is an attempt to discover complex relationships between objects described or mentioned in those documents. Detecting such associations is at the heart of many research and analytical activities that are crucial to applications in national security and business intelligence.

The datasets that Semantic Associations operate over are RDF/RDFS graphs. The semantics of an edge connecting two nodes in an RDF/RDFS graph are *implicit*, in the sense that there is no explicit interpretation of the semantics of the edge other than the fact that it is a predicate in a statement (except for *rdfs:subPropertyOf* edges that

represent data type properties—for which there is model-theoretic (formal) semantics). Hence the RDF/RDFS graph is composed of a combination of *implicit* and *formal* semantics. The objective of Semantic Associations is therefore to find all contextually relevant edge sequences that relate two entities. This is in effect an attempt to combine the *implicit* and *formal* semantics of the edges in the RDF/RDFS graph in conjunction with the context of the query to determine the *multifaceted (multivalent) semantics* of a set of “connections” between entities. We view this *multivalent semantics* as a form of *powerful semantics*. In the context of search, Semantic Associations can be thought of as a class of research searches or discovery-style searches.

CONCLUSION

We have identified three types of semantics and in the process assorted key capabilities required to build a practical semantic application involving Web resources. We have also qualified each of the listed capabilities with one or more types of semantics, as in Table 1. This table reveals some very basic problems that need to be solved for an application to be termed “semantic.” It is clear from this table that *entity disambiguation*, *question answering capability*, *context-based retrieval*, and *navigational and research (discovery) style query capability* require the use of all three types of semantics. Therefore by focusing research efforts in representation mechanisms for *powerful (soft) semantics* in conjunction with fuzzy/probabilistic computational methods supporting techniques that use implicit and formal semantics, it might be possible to solve some of the difficult but practically important problems. In our opinion the current view taken by the Semantic Web community is heavily biased in favor of *formal semantics*. It is clear, however, that the focus of effort in pursuit of the Semantic

Web vision needs to move towards an approach that encompasses all three types of semantics in representation, creation methods, and analysis of knowledge. If the capabilities that we identified do in fact turn out to be fundamental capabilities that make an application semantic, these capabilities could serve as a litmus test or a standard against which other applications may be measured to determine if they are “semantic applications.”

REFERENCES

- Agrawal, R., Imielinski, T., & Swami, A.N. (1993). Mining association rules between sets of items in large databases. In P. Buneman & S. Jajodia (Eds.), *In Proceedings of the 1993*.
- Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules. In *Proceedings of the 20th International Conference on Very Large Databases*, Santiago, Chile.
- Aleman-Meza, B., Halaschek, C., & Sahoo, S. (2003). *Terrorist-related assessment using knowledge similarity*. LSDIS Lab Technical Report, December.
- Anyanwu, K., & Sheth, A.P. (2002). The p operator: Discovering and ranking associations on the Semantic Web. *SIGMOD Record*, 31(4), 42-47.
- Atzeni, P., Basili, R., Hansen, D.H., Missier, P., Paggio, P., Paziienza, M.T., & Zanzotto, F.M. (2004, June). Ontology-based question answering in a federation of university sites: The MOSES case study. In *Proceedings of the 9th International Conference on Applications of Natural Language to Information Systems (NLDB'04)*, Manchester, UK.
- Barbará, D., Garcia-Molina, H., & Porter, D. (1992). The management of probabilistic data. *IEEE Transactions on Knowledge and Data Engineering*, 4(5), 487-502.
- Blair, H.A., & Subrahmanian, V.S. (1989). Paraconsistent logic programming. *Theoretical Computer Science*, 68, 135-154.
- Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. In *Proceedings of the 12th International World Wide Web Conference*.
- Cao, T.H. (2000). Annotated fuzzy logic programs. *International Journal on Fuzzy Sets and Systems*, 113, 277-298.
- Castano, S., Antonellis, V.D., & Vimercati, S.D.C. (2001). Global viewing of heterogeneous data sources. *IEEE Transactions on Knowledge and Data Engineering*, 13(2), 277-297.
- Dill, S., Eiron, N., Gibson, D., Gruhl, D., Guha, R., Jhingran, A., Kanungo, T., Rajagopalan, S., Tomkins, A., Tomlin, J.A., & Zien, J.Y. (2003). Semtag and seeker: Bootstrapping the Semantic Web via automated semantic annotation. In *Proceedings of the 12th International Conference on World Wide Web* (pp. 178-186). ACM Press.
- Ding, Z., & Peng, Y. (2004, January 5-8). A probabilistic extension to ontology language OWL. In *Proceedings of the Hawaii International Conference on System Sciences*, Big Island, Hawaii.
- Dubois, D., Lang, J., & Prade, H. (1994). Possibilistic logic. In D.M. Gabbay et al. (Eds.), *Handbook of logic in artificial intelligence and logic programming* (Vol. 3, pp. 439-514). Oxford: Oxford University Press.
- Etzioni, O., Cafarella, M., Downey, D., Kok, S., Popescu, A.-M., Shaked, T., Soderland, S., Weld, D.S., & Yates, A. (2004). Web-scale information extraction in knowitall (preliminary results). *WWW*, 100-110.
- Giugno, R., & Lukasiewicz, T. (2002). P-SHOQ(D): A probabilistic extension of SHOQ(D) for probabilistic ontologies in the Semantic Web. In *Proceedings of the European Conference on Logics in Artificial Intelligence*.

- Gruber, T. (2003, March 26-27). It is what it does: The pragmatics of ontology. Invited talk at *Sharing the Knowledge—International CIDOC CRM Symposium*, Washington, DC. Retrieved from tomgruber.org/writing/cidoc-ontology.htm
- Guha, R., McCool, R., & Miller, E. (2003, May). Semantic search. In *Proceedings of the 12th International World Wide Web Conference*, Budapest, Hungary.
- Hammer, J., & McLeod, D. (1993). An approach to resolving semantic heterogeneity in a federation of autonomous, heterogeneous database systems. *Journal for Intelligent and Cooperative Information Systems*.
- Hammond, B., Sheth, A., & Kochut, K. (2002). Semantic enhancement engine: A modular document enhancement platform for semantic applications over heterogeneous content. In V. Kashyap & L. Shklar (Eds.), *Real-world Semantic Web applications* (pp. 29-49). IOS Press.
- Han, H., Giles, C.L., Zha, H., Li, C., & Tsioutsoulis, K. (2004). Two supervised learning approaches for name disambiguation in author citations. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL 2004)* (pp. 296-305).
- Handschuh, S., Staab, S., & Ciravegna, F. (2002, October 1-4). S-CREAM—semi-automatic CREATION of metadata. In *Proceedings of the European Conference on Knowledge Acquisition and Management (EKAW-2002)*, Madrid, Spain. Berlin: Springer-Verlag.
- Heinsohn, J. (1994). *Probabilistic description logics* (pp. 311-318). UAI.
- Hjorland, B. (1998). Information retrieval, text composition, and semantics. *Knowledge Organization*, 25(1/2), 16-31.
- Jaeger, M. (1994). *Probabilistic reasoning in terminological logics* (pp. 305-316). KR.
- Kashyap, V., Ramakrishnan, C., Thomas, C., Bassu, D., Rindfleisch, T.C., & Sheth, A. (2003). *TaxaMiner: An experimentation framework for automated taxonomy bootstrapping*. Technical Report No. UGA-CS-TR-04-005, Computer Science Department, University of Georgia, USA.
- Kashyap, V., & Sheth, A. (1996). Semantic heterogeneity in global information systems: The role of metadata, context and ontologies. *Cooperative Information Systems*.
- Kashyap, V., & Shklar, L. (Eds.). (2002). *Real-world Semantic Web applications—Volume 92: Frontiers in artificial intelligence and applications*.
- Kifer, M., & Subrahmanian, V.S. (1992). Theory of generalized annotated logic programming and its applications. *Journal of Logic Programming*, 12, 335-367.
- Kleinberg, J. (1998). Authoritative sources in a hyperlinked environment. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*.
- Koller, D., Levy, A., & Pfeffer, A. (1997). P-CLASSIC: A tractable probabilistic description logic. In *Proceedings of the 14th National Conference on Artificial Intelligence (AAAI-97)* (pp. 390-397).
- Kuramochi, M., & Karypis, G. (2004). Finding frequent patterns in a large sparse graph. In *Proceedings of the SIAM International Conference on Data Mining (SDM-04)*.
- Lukasiewicz, T. (2004). *Weak nonmonotonic probabilistic logics, principles of knowledge representation and reasoning*. KR.
- Maedche, A., & Staab, A. (2001). Ontology learning for the Semantic Web. *IEEE Intelligent Systems*, 16(2), 72-79.
- Naing, M.-M., Lim, E.-P., & Goh, D.H.-L. (2002). Ontology-based Web annotation framework for

- hyperlink structures. In *Proceedings of WISE Workshops 2002* (pp. 184-193).
- Omelayenko, B. (2001). Learning of ontologies for the Web: The analysis of existent approaches. In *Proceedings of the International Workshop on Web Dynamics*.
- Patil, A., Oundhakar, S., Sheth, A., & Verma, K. (2004, May). METEOR-S Web service annotation framework. In *Proceedings of the World Wide Web Conference* (pp. 553-562). New York.
- Quan, D., & Karger, D.R. (2004). *How to make a Semantic Web browser in WWW*.
- Rahm, E., & Bernstein, P.A. (2001). A survey of approaches to automatic schema matching. *VLDB Journal*, 10, 4.
- Ramakrishnan, G., Chakrabarti, S., Paranjpe, D., & Bhattacharya, P. (2004). Is question answering an acquired skill? In *Proceedings of the 13th International Conference on the World Wide Web 2004*.
- Sheth, A., Bertram, C., Avant, D., Hammond, B., Kochut, K., & Warke, Y. (2002). Managing semantic content for the Web. *IEEE Internet Computing*, (July/August), 80-87.
- Sheth, A.P., Thacker, S., & Patel, S. (2003). Complex relationships and knowledge discovery support in the InfoQuilt system. *VLDB Journal*, 12(1), 2-27.
- Sheth, A.P., & Ramakrishnan, C. (2003). Semantic (Web) technology in action: Ontology-driven information systems for search, integration and analysis. *IEEE Data Engineering Bulletin*, 26(4), 40-48.
- Straccia, U. (2004). *Uncertainty and description logic programs: A proposal for expressing rules and uncertainty on top of ontologies*. Technical Report 2004-TR-14.
- Straccia, U. (1998). A fuzzy description logic. In *Proceedings of AAAI-98, 15th Conference of the American Association for Artificial Intelligence*.
- Townley, J. (2000). The streaming search engine that reads your mind. Retrieved August 10, 2000: smw.internet.com/gen/reviews/searchassociation/
- Uschold, M. (2003). Where are the semantics in the Semantic Web? *Artificial Intelligence*, (Fall).
- Wang, J., Wen, J.-R., Lochovsky, F.H., & Ma, W.-Y. (2004). Instance-based schema matching for Web databases by domain-specific query probing. In *Proceedings of the 2004 Conference on VLDBs*.
- Woods, W.A., (2004, June 2-5). Meaning and links: A semantic odyssey. *Principles of Knowledge Representation and Reasoning: In Proceedings of the 9th International Conference (KR2004)* (pp. 740-742).
- Yen, J. (1991). Generalizing term subsumption languages to fuzzy logic. *IJCAI*, 472-477.
- Zadeh, L.A. (2002). Toward a perception-based theory of probabilistic reasoning with imprecise probabilities. *Journal of Statistical Planning and Inference*, 105, 233-264.
- Zadeh, L.A. (2003). From search engines to question-answering systems—the need for new tools. In *Proceedings of the 1st Atlantic Web Intelligence Conference*.

ADDITIONAL ONLINE RESOURCES

www.networkinference.com/Assets/Products/Cerebra_Server_Datasheet.pdf

www.topquadrant.com/documents/TQ04_Semantic_Technology_Briefing.PDF

This work was previously published in the International Journal on Semantic Web & Information Systems, Volume 1, Number 1, pp. 1-18, copyright 2005 by Idea Group Publishing (an imprint of IGI Global).