

Chapter 7.5

Probabilistic Models for the Semantic Web: A Survey

Livia Predoiu

University of Mannheim, Germany

Heiner Stuckenschmidt

University of Mannheim, Germany

ABSTRACT

Recently, there has been an increasing interest in formalisms for representing uncertain information on the Semantic Web. This interest is triggered by the observation that knowledge on the web is not always crisp and we have to be able to deal with incomplete, inconsistent and vague information. The treatment of this kind of information requires new approaches for knowledge representation and reasoning on the web as existing Semantic Web languages are based on classical logic which is known to be inadequate for representing uncertainty in many cases. While different general approaches for extending Semantic Web languages with the ability to represent uncertainty are explored, we focus our attention on probabilistic approaches. We survey existing proposals for extending semantic web languages or formalisms underlying Semantic

Web languages in terms of their expressive power, reasoning capabilities as well as their suitability for supporting typical tasks associated with the Semantic Web.

INTRODUCTION

The Semantic Web is an extension of the World Wide Web that allows for expressing the semantics and not only the markup of data. By means of the representation of the semantics of data, new and not explicitly stated information can be derived by means of reasoners. In this way, software agents can use and integrate information automatically. As common web languages like (X)HTML and XML are not enough for this purpose (Decker et al., 2000), Semantic Web languages have been standardised (RDF, RDF Schema and OWL), proposed (e.g. WRL, SWRL) and new ones are still being devised. However, most languages that are intended for us-

DOI: 10.4018/978-1-60566-028-8.ch005

age on the Semantic Web are deterministic and cannot represent uncertainty. Currently, there is a growing interest in probabilistic extensions of Semantic Web languages. People start to realize that there is inherently probabilistic knowledge that needs to be represented on the Semantic Web. In the following, we briefly describe five areas where probabilistic information plays a role in the context of the Semantic Web.

Representing inherently uncertain Information: Not all of the information that needs to be represented on the Semantic Web is given in terms of definite statements. E.g. statistical information can provide insights to data to be shared on the Semantic Web. Ontological information attached with statistical values like the percentage of people in a population that are of a certain age can help answer queries about the correlation between this age and a certain chronic disease. There are many situations in which the use of this statistical information could be used to improve the behaviour of intelligent systems. An example would be a recommender System that points the user to certain information based on information about the age group.

Ontology Learning: The manual creation of ontologies has been identified as one of the main bottlenecks on the Semantic Web. In order to overcome this problem several researchers are investigating methods for automatically learning ontologies from texts. Existing approach normally use a combination of NLP and text mining techniques (Maedche & Staab, 2004). Typical tasks are the detection of synonyms and of subclass relations using clustering techniques and association rule mining. In both fields, the result of the mining process can be interpreted in terms of a probabilistic judgement of the correctness of the learned relation.

Document Classification: Document Classification can be seen as a special case of ontology learning called Ontology population. Today a major part of the information on the web is present in terms of documents (Web Pages, PDF Docu-

ments etc.). A common way of linking documents to knowledge encoded in ontologies is to assign individual documents to one or more concepts representing its content. Different machine learning techniques have been applied to this problem (Sebastiani, 2002). The most commonly used is the use of naïve Bayes classifiers that estimate the probability of a document belonging to a topic based on the occurrence of terms in sample documents.

Ontology Matching: Different sources often use different ontologies to organize their information. In the case of documents, these are often classified according to different topic hierarchies. In order to be able to access information across these different sources, semantic correspondences between the classes in the corresponding ontologies have to be determined and encoded in mappings that can be used to access information across the sources. Recently, a number of approaches for automatically determining such mappings have been proposed (Euzenat & Shvaiko, 2007). Some of the most successful ones use machine learning techniques to compute the probability that two classes represent the same information.

Ontology Mapping Usage for Information Integration: The usage of the mappings that have been found by matchers as explained in the paragraph above is currently mainly deterministic. Although the mappings are attached with a confidence that expresses how sure the matcher is that the mapping holds, the usage of those mappings consists of a preprocessing step: All mappings that have a confidence value above that threshold are considered deterministically true and all mappings that have a confidence value below that threshold are considered deterministically false. However, there is evidence that this kind of usage is error prone, especially when mappings are composed over several ontologies.

The five examples above clearly demonstrate the importance of probabilistic information in the context of the Semantic Web. E.g., in order to use the learned structures effectively, we need ways

31 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:

www.igi-global.com/chapter/probabilistic-models-semantic-web/37722

Related Content

Leveraging Blockchain Technology in the Value Chain and the Supply Chain: Cases on Blockchain Applications Within Chinese Firms

Youssef Elhaoussine, Yuhan Huand Yihao Ma (2023). *Blockchain Applications in Cryptocurrency for Technological Evolution* (pp. 173-195).

www.irma-international.org/chapter/leveraging-blockchain-technology-in-the-value-chain-and-the-supply-chain/315974

Evaluating Social Interaction and Support Methods over Time

Birgitta Maria Kopp (2016). *Web Design and Development: Concepts, Methodologies, Tools, and Applications* (pp. 1283-1301).

www.irma-international.org/chapter/evaluating-social-interaction-and-support-methods-over-time/137398

Theoretical Foundations of a CSCL Script in Persistent Virtual Worlds According to the Contemporary Learning Theories and Models

Nikolaos Pellas (2016). *Web Design and Development: Concepts, Methodologies, Tools, and Applications* (pp. 1598-1633).

www.irma-international.org/chapter/theoretical-foundations-of-a-cscl-script-in-persistent-virtual-worlds-according-to-the-contemporary-learning-theories-and-models/137413

Information Acquisition and Recall in Location-Aware and Search Engine Retrieval Systems

Sorin Adam Matei, Lance Madsen and Robert Bruno (2010). *International Journal of Information Technology and Web Engineering* (pp. 32-52).

www.irma-international.org/article/information-acquisition-recall-location-aware/44921

Spectral-Spatial Classification of Hyperspectral Image Based on Support Vector Machine

Weiwei Yang and Haifeng Song (2021). *International Journal of Information Technology and Web Engineering* (pp. 56-74).

www.irma-international.org/article/spectral-spatial-classification-of-hyperspectral-image-based-on-support-vector-machine/272027