Domain Ontology Learning from the Web

Candidate: David Sánchez
Institution: Computer Science & Mathematics Department, University Rovira i Virgili, Catalonia, Spain
Supervisor: Antonio Moreno
Year awarded: 2007
URL: http://deim.urv.cat/~itaka/CMS/media/dsanchezThesis.pdf
doi: 10.1017/S0269888909990300

Abstract

Ontology Learning is defined as the set of methods used for building from scratch, enriching or adapting an existing ontology in a semi-automatic fashion using heterogeneous information sources. This data-driven process uses text, electronic dictionaries, linguistic ontologies and structured and semi-structured information to acquire knowledge. Recently, thanks to the enormous growth of the Information Society, the Web has caught the interest of researchers who have started to use it as the corpus to develop Information Extraction and Knowledge Acquisition methodologies. This thesis presents novel approaches to learn Domain Ontologies from the Web. The adaptation of several well-known learning techniques to the Web corpus and the exploitation of particular characteristics of the Web environment composing an automatic, unsupervised and domain-independent approach distinguishes the present proposal from previous works. The solution is designed as an integrated and incremental learning system that is able to construct domain ontologies from scratch. It relies on linguistic regularities (linguistic patterns) of the English language and the information distribution computed from the Web to extract and select ontological entities.

With respect to the ontology building process, the following methods have been developed: (i) extraction and selection of domain-related terms, organizing them in a taxonomical way; (ii) discovery and label of non-taxonomical relationships between concepts; (iii) additional methods for improving the final structure, including the detection of named entities, class features, multiple inheritance and also a certain degree of semantic disambiguation. The full learning methodology has been implemented in a distributed agent-based fashion, providing a scalable solution. It has been evaluated for several well-distinguished domains, obtaining reliable ontologies suitable for knowledge-intensive environments such as the Semantic Web. Finally, several direct applications have been developed, including automatic structuring of digital libraries and web resources, and ontology-based Web Information Retrieval.

A Framework for Summarization of Multi-topic Web Sites

Candidate: Yongzheng Zhang
Institution: Faculty of Computer Science, Dalhousie University, Halifax, NS, Canada
Supervisors: Evangelos Milios and Nur Zincir-Heywood
Year awarded: 2007
URL: http://www.cs.dal.ca/~yongzhen/publication/paper/phd_thesis.pdf
doi:10.1017/S0269888909990312

Abstract

Web site summarization, which identifies the essential content covered in a given Web site, plays an important role in Web information management. However, straightforward summarization of an entire Web site, which is large and with diverse content, may lead to a summary heavily biased to a subset of main topics covered in the target Web site. In this thesis, we propose a two-stage framework for effective summarization of multi-topic Web sites. The first stage identifies the main topics covered in a Web site and the second stage summarizes each topic separately.

In order to identify the different topics covered in a Web site, we perform both text- and link-based clustering. In text-based clustering, we investigate the impact of document representation and feature selection on the clustering quality. In link-based clustering, we study co-citation and bibliographic coupling. We demonstrate that text-based clustering based on the selection of features with high variance over Web pages is reliable and that outgoing links can be used to improve the clustering quality if a rich set of cross links is available.

Each individual cluster computed above is summarized using an extraction-based summarization system, which extracts key phrases and key sentences from source documents to generate a summary. The performance of such an extraction-based Web site summarization system depends on its underlying key phrase extraction method. Hence, we conduct a user study to investigate five alternative key phrase extraction methods. Results show that the best method combines linguistic constraints with frequency over the corpus adjusted to take into account nesting of terms. Another important component in an extraction-based summarization system is the key sentence extraction. To this end, we design and develop a classification approach in the cluster summarization stage. The classifier uses statistical and linguistic features to determine the topical significance of each sentence.

Finally, we evaluate the proposed system via a user study. We demonstrate that the proposed clustering summarization approach significantly outperforms the single-topic summarization approach for any given Web site summarization task.