Automated Incremental Building of Weighted Semantic Web Repository[*]

Martin Řimnáč and Roman Špánek

Summary. The chapter introduces an incremental algorithm creating a self-organizing repository and it describes the processes needed for updates and inserts into the repository, especially the processes updating estimated structure driving data storage in the repository. The process of building repository is foremost aimed at allowing the well-known Semantic web tools to query data presented by the current web sources. In order to respect features of current web documents, the relationships should be at least weighted by an additional indirect criteria, which allow the query result to be sorted accordingly to an estimated quality of data provided by web sources. The relationship weights can be based on relationship soundness or on the reputation of the source providing them. The extension of the relationships by the weights leads to the repository able to return a query result as complete as possible, where (possibly) inconsistent parts are sorted by the relationships weights.

1 Introduction

With the amount of data available on the Web rapidly growing during the past years, the ability to find relevant data on the Web by current methods mostly based on approaches known from information retrieval becomes even more hard but also crucial task.

Martin Řimnáč and Roman Špánek
Institute of Computer Science, AS CR, Prague, Czech Republic
e-mail: {rimnacm,spanek}@cs.cas.cz

Roman Špánek
Technical University of Liberec, Hálkova 6, Czech Republic

[*] The work was supported by the project 1M0554 ”Advanced Remedial Processes and Technologies” and partly by the Institutional Research Plan AV0Z10300504 “Computer Science for the Information Society: Models, Algorithms, Applications”.

A. Abraham et al. (Eds.): Foundations of Comput. Intel. Vol. 6, SCI 206, pp. 265–296.
springerlink.com

The *information retrieval engines* [1] build large indices storing a word occurrence in (Web) documents, and data searching algorithms use these indices to find the documents corresponding to an end-user query given as a list of keywords.

The first information retrieval engines used only a simple algorithm for evaluating queries returning just a set of links to (relevant) documents. Link relevance was evaluated by a cosine measure, which compared a vector representing words in a document with a vector representing keywords in a query. The amount of returned links and their various quality were the most severe drawbacks.

A solution has been brought by *Google* that has proposed a novel approach sorting the links according to source quality estimated indirectly by so called the *Page-Rank* [2]. The Page-rank is based on the assumption that sound (interesting and high quality) documents are referenced more by (sound) documents. In this way, the Page-Rank tries to extend the classical cosine measure by a quality consideration in order to put the most relevant and high quality documents at top positions in the query result.

Today information retrieval engines used on the Web mostly return thousands of links, but only few of them is analyzed by a human user. Therefore seeking a complete information on the web is currently quite a difficult task.

The *Semantic Web vision* [3] is the inspiration for many currently studied approaches for searching relevant data on the Web. The Semantic Web documents provide data as (well defined) relationships between well-defined entities, which are called resources. While resource meaning is often defined in the external ontologies, data relationships are provided by independent local sources. This leads to a need for preprocessing queries by software agents analyzing all relevant documents and trying to provide a complete answer to the user query. Searching data on the Semantic Web makes engines to manage inverse indices of resources and their related relationships (instead of words) provided by a given document. The limited amount of Semantic Web documents is one of the drawbacks of the current Semantic Web.

Most of current web documents are presented in human user friendly form. This fact leads to a strong motivation for developing new machine learning methods handling current web documents and analyzing their content in order to estimate meaning of the documents. These methods [4, 5] try to propagate estimated relationships from Web documents into *extensional definitions* of Semantic Web resources. All the estimated relationships are stored in the proposed *repository*. As the web documents presenting the relationships are often updated, all proposed methods handling processes in the repository are designed incrementally.

Since the repository includes relationships from various web documents, basic *data integration* should be supported in the repository. Using semiautomatic approaches to estimate *integration rules*, a special kind of relationships connecting sources, may cause a proposal including conflicting rules. Therefore all the rules should be weighted by measures based on analysis