WEB HARVESTING

Wolfgang Gatterbauer
Computer Science and Engineering
University of Washington, USA

SYNONYMS

web data extraction, web information extraction, web mining

DEFINITION

Web harvesting describes the process of gathering and integrating data from various heterogeneous web sources. Necessary input is an appropriate knowledge representation of the domain of interest (e.g. an *ontology*), together with example instances of concepts or relationships (*seed knowledge*). Output is structured data (e.g. in the form of a relational database) that is gathered from the Web. The term *harvesting* implies that, while passing over a large body of available information, the process gathers only such information that lies in the domain of interest and is, as such, relevant.

MAIN TEXT

The process of web harvesting can be divided into three subsequent tasks: (1) *retrieving data*, which involves finding relevant information on the Web and storing it locally. This requires tools for searching and navigating the Web, i.e. crawlers and means for interacting with dynamic or deep web pages, and tools for reading, indexing and comparing the textual content of pages; (2) *extracting data*, which involves identifying relevant data on retrieved content pages and extracting it into a structured format. The important tools that allow access to the data for further analysis are parsers, content spotters and adaptive wrappers; (3) *integrating data*, which involves cleaning, filtering, transforming, refining and combining the data extracted from one or more web sources, and structuring the results according to a desired output format. The important aspect of this task is organizing the extracted data in such a way as to allow unified access for further analysis and data mining tasks.

The ultimate goal of web harvesting is to compile as much information as possible from the Web on one or more domains and to create a large, structured knowledge base. This knowledge base should then allow querying for information similar to a conventional database system. In this respect, the goal is shared with that of the Semantic Web. The latter, however, tries to solve extraction *à priori* to retrieval by having web sources present their data in a semantically explicit form.

Today's search engines focus on the task of finding content pages with relevant data. The important challenges for web harvesting, in contrast, lie in

extracting and integrating the data. Those difficulties are due to the variety of ways in which information is expressed on the Web (*representational heterogeneity*) and the variety of alternative, but valid interpretations of domains (*conceptual heterogeneity*). These difficulties are aggravated by the Web's sheer size, its level of heterogeneity and the fact that information on the Web is not only complementary and redundant, but often contradictory too.

Another important research problem is the optimal combination of automation (high recall) and human involvement (high precision). At which stages and in which manner a human user must interact with an otherwise fully automatic web harvesting system for optimal performance (in terms of speed, quality, minimum human involvement, etc.) remains an open question.

CROSS REFERENCE*

information retrieval, data extraction, web data extraction, web data extraction systems, fully automatic web data extraction, web scrapers, wrappers, data integration, Semantic Web

REFERENCES*

Optional. A list of 1-3 citations that give the reader a place to find more information.

[1] Fabio Ciravegna, Sam Chapman, Alexiei Dingli, and Yorick Wilks. Learning to harvest information for the Semantic Web. In *Proc. 1st European Semantic Web Symposium (ESWS 2004)*, volume 3053 of *Lecture Notes in Computer Science*, pages 312–326. Springer, 2004.

[2] Valter Crescenzi and Giansalvatore Mecca. Automatic information extraction from large websites. *Journal of the ACM*, 51(5):731–779, 2004.

[3] Oren Etzioni, Michael J. Cafarella, Doug Downey, Stanley Kok, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, and Alexander Yates. Web-scale information extraction in KnowItAll: (preliminary results). In *Proc. 13th international conference on World Wide Web (WWW 2004)*, pages 100–110. ACM, 2004.