Chapter 12
AGATHE-2:
An Adaptive, Ontology-Based Information Gathering Multi-Agent System for Restricted Web Domains

Bernard Espinasse
Aix-Marseilles University, France

Sébastien Fournier
Aix-Marseilles University, France

Fred Freitas
Universidade Federal de Pernambuco, Brazil

Shereen Albitar
Aix-Marseilles University, France

Rinaldo Lima
Universidade Federal de Pernambuco, Brazil

ABSTRACT

Due to Web size and diversity of information, relevant information gathering on the Web turns out to be a highly complex task. The main problem with most information retrieval approaches is neglecting pages' context, given their inner deficiency: search engines are based on keyword indexing, which cannot capture context. Considering restricted domains, taking into account contexts, with the use of domain ontology, may lead to more relevant and accurate information gathering. In the last years, we have conducted research with this hypothesis, and proposed an agent- and ontology-based restricted-domain cooperative information gathering approach accordingly, that can be instantiated in information gathering systems for specific domains, such as academia, tourism, etc. In this chapter, the authors present this approach, a generic software architecture, named AGATHE-2, which is a full-fledged scalable multi-agent system. Besides offering an in-depth treatment for these domains due to the use of domain ontology, this new version uses machine learning techniques over linguistic information in order to accelerate the knowledge acquisition necessary for the task of information extraction over the Web pages. AGATHE-2 is an agent and ontology-based system that collects and classifies relevant Web pages about

DOI: 10.4018/978-1-60960-132-4.ch012

a restricted domain, using the BWI (Boosted Wrapper Induction), a machine-learning algorithm, to perform adaptive information extraction.

INTRODUCTION

Because of the size of the Web and the diversity of accessible information, to gather relevant information from the Web turns out to be a highly complex task. Without taking explicitly into account the search context, the majority of the current approaches of information retrieval (IR) let escape many forms of organized information of the Web, for example, specific domains or "clusters" of information.

However, the field known as Symbolic Artificial Intelligence (AI) has faced a similar challenge in the past. During the seventies, researchers from this field tried to produce systems that could cope with inference capabilities about everything. The lesson learned (Newell, Shaw, & Simon, 1959) was that the use of knowledge-based systems is feasible only over restricted domains, which led to the relative success of the expert systems. This policy is also valid for the IR field. Indeed, the evaluation of the IR systems is mainly carried out over homogeneous corpora, whose texts relates to only one subject and often come from the same source, and not from text sets with diverse contents and writing styles, as it is the case of those available on the Web. This fact is also besides at the origin of the development in IR of specialized search engines (Mc Callum et al, 1999).

Another argument pleading for a restricted domain in IR relates to Information Extraction (IE). Generally, IE works over textual documents collections (Muslea, Minton, & C. Knoblock, 1998). The task consists in extracting data starting from specific classes of Web pages (Gaizauskas & Robertson, 1997). It concerns the identification of specific fragments from a document, which should constitute the core of its semantic contents (Kushmerick, 1999a). The main goal of IE is to populate databases about specific domains - such as Tourism, Academia, etc - regrouping information coming from many Web pages spread over geographically distributed sites. These databases save users' work on finding, checking and comparing the data which then can be easily queried by users.

Taking such a specific domain context into account enables better data processing (Etzioni et al., 2004). It is the case of the extraction of majority of information from a given class of pages (for example the value of the dollar from a currency exchange rates page, subjects of interest of a researcher from his homepage and so on). Another advantage is to make possible for the users to carry out queries combining, in particular, search keys relative to various classes of pages, allowing complex requests (the search of the papers published in a certain whole of conferences, for example). Thus, it is possible to build sophisticated applications in order to gather Web information from specific domains. With the "Tourism" cluster, for example, applications could retrieve, extract, and classify data about hotels, passage tickets, and cultural events.

On the other hand, it is widely known that Machine Learning (ML) algorithms simplify the development of IE programs; these algorithms have been utilized to automate extraction rules' production. In recent times, many IE systems had been developed following a three-step procedure: (1) Recognizing relevant information in the text (2) Extracting this information (3) Storing it in an organized structure or in a database (Kushmerick, 1999b; Siefkes & Siniakov, 2005).

In the last years, we have conducted research with these research hypotheses, and produced ontology-based restricted-domain cooperative information gathering software agents accordingly, that permit the development of a specific information gathering systems e.g. the MASTER-