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 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 permit the development of specific information gathering systems. In this article, we present this novel approach based on these guiding ideas, and a generic software architecture, named AGATHE, which is a full-fledged scalable multi-agent system. [Article copies are available for purchase from InfoSci-on-Demand.com]

Keywords: Agent Software; Information Retrieval; Ontologies

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 et al., 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 relate 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 (Steel, 2001; McCallum et al., 1999).

Another argument pleading for a restricted domain in IR relates to Information Extraction (IE). Generally, IE works over textual documents collections (Musleia, 1999; Embley et al., 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 (Kushmeric, 1999). 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.

In the last years, we have conducted research with these research hypotheses, and produced ontology-based restricted-domain cooperative information gathering multi-agents accordingly, that permit the development of a specific information gathering systems e.g. the MASTER-Web system (Freitas & Bittencourt, 2003). In this article, we present a novel approach based on these guiding ideas, a generic software architecture, named AGATHE, which is a full-fledged scalable multi-agent system, which includes a better-designed and organized agent topology. The article is organized as follows: in section 2, we introduce the concept of cooperative information gathering, and the interest of using agents and different kinds of ontologies to develop intelligent gathering systems on one or more restricted domains of the Web. Section 3 introduces the AGATHE system, a multi-agent architecture for development of such intelligent gathering systems on the Web, its objectives, its architecture with its three main subsystems, and its general functioning. AGATHE is inspired from an early prototype developed in Brazil and named MASTER-Web (Freitas & Bittencourt, 2003). Sections 4 to 6 present in detail these subsystems composing AGATHE system: the Search Subsystem, the Extraction Subsystem and the User Subsystem. Section 7 presents some implementation details of the prototype in progress and some results. Finally, we conclude with some research perspectives.

COOPERATIVE INFORMATION GATHERING AND ONTOLOGIES

Suggested by Oates et al. (1994), the concept of “Cooperative Information Gathering” (CIG), is based on the distributed problem solving paradigm for the fields of multi-agent systems (MAS) and distributed artificial intelligence (DAI) (Huhns & Singh, 1994) (Nwana, 1996). CIG involves concurrent, asynchronous discovery and composition of information spread
Related Content

Mobilization Techniques Utilized by Leading Global E-Commerce Sites
[www.igi-global.com/article/mobilization-techniques-utilized-by-leading-global-e-commerce-sites/152315?camid=4v1a](www.igi-global.com/article/mobilization-techniques-utilized-by-leading-global-e-commerce-sites/152315?camid=4v1a)

Exploring Decision Rules for Sellers in Business-to-Consumer (B2C) Internet Auctions
[www.igi-global.com/article/exploring-decision-rules-sellers-business/1897?camid=4v1a](www.igi-global.com/article/exploring-decision-rules-sellers-business/1897?camid=4v1a)

RFID Systems: Applications vs. Security and Privacy Implications
[www.igi-global.com/chapter/rfid-systems-applications-security-privacy/9325?camid=4v1a](www.igi-global.com/chapter/rfid-systems-applications-security-privacy/9325?camid=4v1a)
Semiautomatic Derivation and Use of Personal Privacy Policies in E-Business
www.igi-global.com/article/semiautomatic-derivation-use-personal-privacy/1836?camid=4v1a