Expert Systems with Applications 36 (2009) 10068–10077

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

# Semantic information integration and question answering based on pervasive agent ontology

Qing-lin Guo [a,b,*], Ming Zhang [b]

[a] School of Computer Science and Technology, North China Electric Power University, Beijing 102206, China
[b] Department of Computer Science and Technology, Peking University, Beijing 100871, China

ARTICLE INFO

*Keywords:*
Question answering
Pervasive agent ontology
Semantic web
Natural language understanding
Concept extraction
Answer extraction

ABSTRACT

The traditional search engines return a large number of relative web pages rather than accurate answers. However, in a question answering system, users could use sentences in daily life to raise questions. The question answering system will analyze and comprehend these questions and return answers to users directly. Aiming at the problems in current network environment, such as low precision of question answering, imperfect expression of domain knowledge, low reuse rate and lack of reasonable theory reference models, we put forward the information integration method of semantic web based on pervasive agent ontology (SWPAO) method, which will integrate, analyze and process enormous web information and extract answers on the basis of semantics. With SWPAO method as the clue, we mainly study the method of concept extraction based on uniform semantic term mining, pervasive agent ontology construction method on account of multi-points and the answer extraction in view of semantic inference. Meanwhile, we present the structural model of the question answering system applying ontology, which adopts OWL language to describe domain knowledge base from where it infers and extracts answers by Jena inference engine, thus the precision of question answering in QA system could be improved. In the system testing, the precision has reached 86%, and recalling rate is 93%. The experiment indicates that this method is feasible and it has the significance of reference and value of further study for the question answering systems.



## 1. Introduction

With the popularization of the Internet and increasingly diverse online information, search engines such as Google, Yahoo, and Baidu have gradually become a part of people's life. However the traditional search engine has many limitations, for instance, it returns a lot of relative web pages rather than accurate answers. In addition, it only regards keywords as index without relating to semantic information, making it difficult to really understand what the user intends to do. In question answering (QA), users could use sentences in daily life to raise questions and the system will return answers to users directly after analyzing and comprehending these questions. Therefore, the question answering system better satisfies the users' requirements. We could make an example to illustrate the difference between the question answering system and online browsing and searching: online browsing is like reading newspaper and searching is like consulting dictionaries; however, question answering is like asking for directions. As a result, compared with online browsing and searching, a question answering system could better express users' needs and adapt to their habits, moreover, it should return information more accurately, quickly and efficiently. It can be said that the question answering system is a new generation of intelligent search engine. Therefore, integrating, analyzing and processing enormous web information in network environment and providing corresponding answers to users' questions are becoming new hot topics in current researches.

The data in web has the features of semi-structure (Zhou & Wang, 2006), heterogeneity and distributivity. Blocking these features and providing uniform model for users are the key questions of web information integration. However, the inherent features of heterogeneity, distributivity, growth and variation of the data in the Internet determine that the structural method is not adaptable to web information integration. With the vigorously generalization of semantic web by W3C, semantics oriented web information integrating method has been the major point of the research on web information integration technology (Brickly & Guha, 2008). One common semantic model must be offered to solve the problem of semantic heterogeneity in semantics oriented web information integration, which is a platform independent model to block semantic heterogeneity among web information. As "the set of

* Corresponding author. Address: Department of Computer Science and Technology, Peking University, Beijing 100871, China.
*E-mail addresses:* qlguo88@sohu.com, qlguo@pku.edu.cn (Q.-l. Guo), mzhang@pku.edu.cn (M. Zhang).

0957-4174/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2009.01.056

concepts and relations among concepts in specific domain" (Holsapple & Joshi, 2007), ontology could efficiently express the general knowledge in specific domain, which is adaptable to be the common semantic model of semantics oriented web information integration.

In recent years, ontology arouses widespread concern in the artificial intelligence domain and it has been applied extensively (Ehrig & Staab, 2004). Adopting ontology knowledge base in QA could better indicate the inherent relations among knowledge, rationalize knowledge organizations, reduce redundant memories and help extract semantic answers. In many applications of web information integration that use enquiry-driven integration mode (KAON, 2008) facing to information sources of multiple domains (Jena, 2008), the users' enquiries usually cover many information sources. Therefore, users usually expect to search information they interested using only one special ontology without caring for answers received from which information sources or through what kind of processing. To meet this demand, a general ontology should be constructed among local domain ontologies, which is called pervasive agent ontology (PAO for short). Pervasive agent ontology and local ontology should collaborate mutually through shared lexical collection, and they should be of mapping relations. On such basis, we put forward the information integrated method of semantic web based on pervasive agent ontology, SWPAO method for short.

Natural language is the most habitual communication tool for human being, and the question answering mode is one of the handiest ways for people to exchange information and study (Bril, Dumas, & Banko, 2006). The popularization and development of network technologies break the limitations of time and space, and especially the integration of semantic web information provides channels for people to ask and answer through human–computer interaction. Therefore, people have been expecting for long a question answering system based on natural language understanding. At the same time, it is not only a hot research subject for many fields such as computer science, linguistics, psychology, cognitive science, etc, but also the most concerned topic in international conferences of Text REtrieral Conference (TREC) and Text Analysis Conference (TAC), and an important R&D project of big companies like Microsoft (Yang & Chua, 2007). It belongs to the category of natural language understanding, which mainly covers philosophy, cognitive science, linguistics, mathematics, and computer science, and it is an interdisciplinary research of many domains. Among them, the research of philosophy could provide research methodology such as ontology; the research of cognitive science could offer relative cognitive model; the study of linguistics could present the law of question answering and other natural language phenomena; the research of mathematics could abstract the rules and laws of question answering and their inherent relations into mathematic description so that they can be formalized and calculated (Horrocks & Patel-Schneider, 2004); computer science provides concrete implementation of target system finally.

It is predicted that the question answering system will be "a killer application of the post search engine", and it will be one of the main ways for human–computer interaction and become a part of our daily life in the future. Therefore, study on the question answering system upon semantic web information is of significance for the academic research and high application value with distinct prospectiveness.

For this purpose, we probe into the information integration and question answering of semantic web based on pervasive agent ontology.

This paper is organized as follows: Section 2 is the construction of pervasive agent ontology. An example of the construction of pervasive agent ontology is given for illustrating that the method and algorithm of PAO constructing is effective. In Section 3, we dissertate concept extraction based on uniform semantic term mining. Section 4 discusses intelligent searching for documents on the semantic web. Section 5 discusses the flexible enquiry on ontology and enquiry rewriting in SWPAO method. In Section 6, we specifically introduce the design of answer extraction module via typical title template and pervasive agent ontology. Section 7 is our experimental evaluation. Sections 8 and 9 are the discussion and conclusion, respectively.

## 2. The construction of pervasive agent ontology

In Web information integration system constructed by applying SWPAO method, a pervasive agent ontology should be provided to users as an access interface and as a common semantic model corresponding to all information sources. Pervasive agent ontology should contain the possible concepts and relations among concepts in users' enquiries and conform to local ontology in semantics as much as possible. Therefore, we put forward three principles in the process of constructing pervasive agent ontology: principle of completeness, principle of cross-correlation and principle consistency.

**Definition 1.** Domain space: domain space can be defined as $\langle D,W\rangle$, $D$ is domain, $W$ is the events set in the domain.

**Definition 2.** Ontology $O$ can be defined as a seven tuple:

$$O := (c, H_C, \Re, H_R, P, A, D)$$

Where $c$ is the set of concept $C$ (instances of rdfs: Class). $H_C \subseteq c \times c$ denotes subsumption relation between concepts (instances of rdfs: subClassOf). $\Re \subseteq c \times c$ denotes the set of binary property relation between concepts (instances of rdfs: objectProperty), whereas $H_R, \subseteq \Re \times \Re$ denotes hierarchy between properties (instances of rdfs: subPropertyOf). $P \subseteq D \times DT$ is a set of data properties for concept (instances of owl: DatatypeProperty), where $DT \subseteq D$ is a datatype set. $A$ is a set of axioms expressed in a logical language and can be used to infer knowledge from existing one. $D$ denotes a domain set containing all instances of concepts $C$. In fact, all elements in ontology $O$ can be expressed as 3-tuple set of the form (subject, predicate, object).

Ontology Mapping between ontology $O_1$ and ontology $O_2$ can be regarded as follows: some entities in $O_1$ is mapped onto at most one entity in $O_2$, vice versa.

**Definition 3.** Conceptual relation: the $n$-tuple relation on $\langle D,W\rangle$ can be defined as:

$$p^n : W \to 2^{D^n}$$

It is the function of the sets about all $n$-tuple relation on $\langle D,W\rangle$.

**Definition 4.** Let $\sum O$ be the vocabulary of the ontology $O$, We use $\varepsilon$ to refer to a null entity. Ontology mapping can be defined by a mapping function $\rho : \sum O_m \to \sum O_t$ such that:

$$\forall s \in \sum O_m \left(\exists t \in \sum O_t : \rho(s) = t \text{ or } \rho(s) = \varepsilon\right)$$

Where $\sum O_m$ and $\sum O_t$ are the source ontology and the target ontology of mapping $\rho$ respectively. Sometimes, we can use $\rho\,(O_1,O_2)$ denotes the mapping function from $O_1$ to $O_2$.

Semantic reasoning over ontology representations can help us to find additional implicit semantic information.

Pervasive agent ontology is constructed according to the local ontology, so its construction process is that of ontology integration in essence. Using the multi-view theory (Sánchez & Moreno, 2008) in requirement engineering as reference and basing on some limits and hypotheses, we put forward the method of pervasive agent