Computers in Biology and Medicine 66 (2015) 337–342

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Clinic expert information extraction based on domain model and block importance model

Yuanpeng Zhang [a], Li Wang [a], Danmin Qian [a], Xingyun Geng [a], Dengfu Yao [b], Jiancheng Dong [a,*]

[a] Department of Medical Informatics, Medical School, Nantong University, 19 Qixiu Road, Nantong 226001, Jiangsu Province, China
[b] Research Center of Clinical Medicine, Affiliated Hospital Nantong University, Nantong 226001, China

ARTICLE INFO

Article history:
Received 27 November 2014
Accepted 12 July 2015

Keywords:
Information extraction
Clinic expert information
Domain model
Block importance model
SVM

ABSTRACT

To extract expert clinic information from the Deep Web, there are two challenges to face. The first one is to make a judgment on forms. A novel method based on a domain model, which is a tree structure constructed by the attributes of query interfaces is proposed. With this model, query interfaces can be classified to a domain and filled in with domain keywords. Another challenge is to extract information from response Web pages indexed by query interfaces. To filter the noisy information on a Web page, a block importance model is proposed, both content and spatial features are taken into account in this model. The experimental results indicate that the domain model yields a precision 4.89% higher than that of the rule-based method, whereas the block importance model yields an F1 measure 10.5% higher than that of the XPath method.

1. Introduction

Many patients need to choose a suitable doctor when they need medical care, but there is not enough information or data to help them make a right choice. Nowadays, most hospital portals display information of their clinic experts, such as professional title, department, specialty and clinic time. If the expert information from all hospital portals of one or more cities are collected and integrated into databases of electronic health records (EHRs), it will provide great convenience for community residents who need hospital care. However, most of this information, hidden behind the Web, is not directly shown on Web pages. This information is known as the Deep Web [1,2]. Therefore, this paper focuses on the crawling of the Deep Web.

The only way to retrieve information from the Deep Web is through Web Query Interfaces (WQIs). Fig. 1 depicts this process. On Web pages, WQIs always appear as HTML forms that allow accessing Hidden-Web databases [3]. Two problems generally need to be dealt with when extracting information from the Deep Web. The first one is the identification of Web sites containing WQIs, and the second is to extract useful information from response pages searched by WQIs. In Cope et al., the authors put forward some rules to identify WQIs [4]. For example, they state that a HTML form is considered as a WQI if it contains some text input controls ($<input\ type=text>$) and attribute-words. Unfortunately, this method needs further improvements because it cannot distinguish search engines from WQIs. For the information extraction problem, it is critical to filter noisy information such as navigation information, advertisement information, version information, etc. Yan Fu et al. adopted a series of rules called XPath to distinguish useful content blocks from noisy clutters [5]. This method was tested in five different Web sites and resulted in average integrality of 92% and average accuracy of 83.2%. However, this method has a basic precondition that the Web pages should have a common layout. Therefore, this method does not have general applicability.

There are three main contributions in this paper:

- A domain model is defined to identify WQIs. The model can identify the domain to which an unknown interface belongs, and provide domain keywords to fill in WQIs.
- In order to improve the performance of the domain model, a virtual MapReduce cluster based on VMware ESX2.0 and Hapood version 0.18.0 is established. This cluster provides a parallel computing environment.
- A block importance model is proposed to filter noisy blocks on a Web page. Both content and spatial features are taken into account in this model.

Although our paper focuses on a clinic domain, it also can be generalizable to other domains. For example, vocabularies such as

* Corresponding author.
E-mail address: dongjc@ntu.edu.cn (J. Dong).

http://dx.doi.org/10.1016/j.compbiomed.2015.07.009
0010-4825/

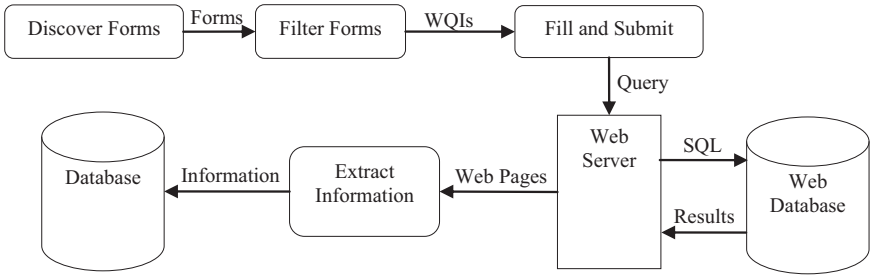

Fig. 1. Process of Deep Web extraction.

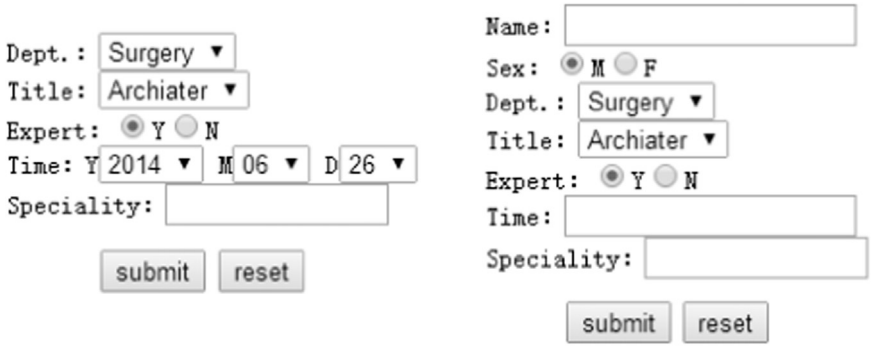

Fig. 2. Two WQIs in clinic domain.

"From", "To", "Departing time", "Returning time", etc. are usually used to depict the attributes of air ticket query interfaces. Obviously, these attributes are convergent through clustering, therefore, a domain model can be established for extracting air ticket information. In addition, this model also can be used in the book domain, job domain, automobile domain, etc.

2. Method

2.1. Domain model

Researchers from UIUC manually collected 477 WQIs in 8 areas through the Google engine and Web directory service, performed a statistical analysis of these WQIs and came to the following conclusions [6].

- Attributes of each WQI are finite.
- Although there are numerous WQIs in each domain, vocabularies that depict the attributes of WQIs are convergent through clustering.

Based on these two unique features, a domain model that can describe the attributes of WQIs is proposed. The definition of the domain model is as follows.

Definition. The domain model is an ordered attribute tree that is defined as a 11-tuple, i.e. $DM=(V, v_0, E, \Delta, TP, N, Lb, Val, tf, R, \leq)$, where

V–a node set
v_0–a root node,$v_0 \in V$
E –an edge set (linking a parent node with a child node)
Δ–a character set
TP –a function returning the type of a node. ($V \rightarrow$ {(*radio button*, *check box*, *text box*, *select list*)*})
N–a function returning the name of a node. ($V \rightarrow \{\Delta^*\}$)
Lb–a function returning the label of a node. ($V \rightarrow \{\Delta^*\}$)
Val–a function returning the default value of a node. ($V \rightarrow \{\Delta^*\}$)
tf–a function returning the frequency of occurrences of a node in all WQIs. ($V \rightarrow \{N^*\}$) (N represents natural number)
R–a function returning the relationship between a node and its father node. ($V \rightarrow$ {*range,part,group,constraint*})
$\leq$ –represents the order of nodes in the node set, $(u, v) \in \leq$ means that u appears before v

2.2. Construction of domain model

Based on the above definition, the construction of a domain model can be described as follows. A WQI is chosen from one domain as an original domain model, then it is combined with other WQIs in this domain in order to enlarge and enrich the original one. This process is repeated until the domain model becomes stable. This combination process should comply with the following four rules.

Assuming u is a node belonging to the original domain mode and v is a new node, then

- *Add rule*. If the semantics of v is different from other nodes in the original domain model, then add a tree (v is the root node) to the original domain model.
- *Update rule*. If the semantics of v is similar to u, then update *TP list, N list, Lb list* and *Val list* of u with *TP, N, Lb* and *Val* of v.
- *Refine rule*. If the semantics of v is similar to u, and v contains some attributes that do not appear in u, then add v to the domain model as a child node of u.
- *Generalize rule*. If the semantics of v can generalize several nodes {*u1,u2…un*} in the original domain model, then add v to the domain model as a child node of these nodes' parent. At the same time, take {*u1,u2…un*} as children of v.

For example, Fig. 2 shows a typical example of two WQIs in the clinic domain, we can merge these two WQIs using the above rules. Fig. 3 shows the merging result.

2.3. Storage of domain model

A perfect and stable domain model can be established through clustering of WQIs. However, it should be stored before we use. Extensible Markup Language (XML) [7] is chosen as the storage