Information Retrieval from Deep Web Based on Visual Query Interpretation

Radhouane Boughammoura, Department of Sciences of Data Processing, Faculty of Sciences of Monastir, Research Unit MARS, Monastir, Tunisia

Mohamed Nazih Omri, Department of Sciences of Data Processing, Faculty of Sciences of Monastir, Research Unit MARS, Monastir, Tunisia

Lobna Hlaoua, Department of Electronic and Computer Science, High School of Sciences and Technology of H. Sousse, Research Unit MARS, H. Sousse, Tunisia

ABSTRACT

Deep Web is growing rapidly. More than 90% of relevant information in web comes from deep Web. Users are usually interested by products which satisfy their needs at the best prices and quality of service. Hence, user’s needs concerns not only one service but many competitive services at the same time. However, for commercial reasons, there is no way to compare all web services products. Each web service is a black box which accepts queries through its own query interface and returns results. As consequence, users ask separately different web services and spend a lot of time comparing products in order to find the best one. This is a burden for novice users. In this paper, the authors propose a new approach which integrates query interfaces of many web services into one universal web service. The new interface describes visually the universal query and is used to ask many web services at the same time. The authors have evaluated their approach on standard datasets and have proved good performances.

Keywords: Information Retrieval, Query Model, Query Visualization, Query’s Schema Extraction, Schema Integration

INTRODUCTION

Deep Web is a treasury of products and services. This treasury is hidden behind Web forms which give access to data stored in distant databases. Deep Web is very rich in terms of quantity (more than 90% of Web) and quality of service. However they are often disappointed because there is no way to compare products of two competitive deep web services. This drawback is due to two factors: 1) deep web services are offering competitive commercial products, 2) these services are offering query interfaces with different query capabilities. Hence queries submitted to servers have different meaning. In order to find a good
product, the burden of information retrieval is the responsibility of users. This is time consuming and need a considerable cognitive effort. For novice users, this charge is the deadline between them and deep Web.

We try to facilitate information retrieval from deep web to novice users. Our goal is to build a universal web form which allows novice users query many web services at the same time. This web form make possible to formulate only one query and obtaining results from all merged web services at the same time. We aim to create this new web form which must be user friendly and easy to understand for novice users (Figure 1).

1. RELATED WORKS

Approaches of query extraction can be classified in two categories: Approaches based on visual features (Zhang, He, & Chang, 2004; Dörk, Carpendale, Collins, & Williamson, 2008; Cai, Yu, Wen, & Ma, 2003), and approaches based on HTML features (Nguyen, Nguyen, & Freire, 2008; Boughammoura, Omri, & Youssef, 2008; Boughammoura & Omri, 2009).

When we review literature, we distinguish four works that utilize some visual information to extract deep Web data: ViNTS (Zhao, Meng, Wu, Raghavan, & Yu, 2005), ViPERS (Simon & Lausen, 2005), HCRF (Zhu, Nie, Wen, Zhang, & Ma, 2006), and VENTex (Gatterbauer, Bohnsky, Herzog, Krpl, & Pollak, 2007).

ViNTs is based on visual content features on the query result pages in order to capture content regularities denoted as Content Lines, and then, utilize the HTML tag structures to combine them.

ViPER is another approach which extracts data records and uses visual information on a Web page with the help of a global multiple sequence alignment technique. However, in the two approaches, tag structures are still the primary information utilized, while visual information plays a small role.

HCRF is a probabilistic model for both data record extraction and attribute labeling. Compared to our solution, it also uses VIPS algorithm (Zhang, He, & Chang, 2005) to

Figure 1. Architecture of our deep Web Information retrieval system