REPORT

- Herat Acharya

Overview

This report is about the two papers regarding querying deep web on the web and integrating the results and displaying it to the user in through a unified interface. The papers were: **Toward Large Scale Integration: Building a MetaQuerier over Databases on the Web** and **EntityRank: Searching Entities Directly and Holistically**.

The authors of the first paper, **Toward Large Scale Integration: Building a MetaQuerier over Databases on the Web**, discuss the rapid changes in the way structured data is stored on the Web and the grueling work that accompanies while accessing these structured databases. It's due to these problems that they have developed a new system called the MetaQuerier which unifies all the structured databases and allows users to search the content of the structured databases hidden behind web forms (also called the deep web content) under a unified interface. In order to explain the MetaQuerier in a better fashion I also gave a demonstration of the system implemented as a website called Cazoodle.com. The authors discuss three major subsystems of the MetaQuerier namely Database Crawler, Interface Extraction and Schema Matching. They then move on to discuss the integration of these subsystems, error detection and correction, while doing so also highlight the lessons learnt and challenges faced in focusing their efforts on system integration, i.e. making individual systems work together as a single unit. As I went on with my presentation, I also gave examples illustrating the commonalities of various schemas (web interfaces) across a particular domain, for e.g.: fields like title, author, ISBN are commonly occurring fields across schemas of Books domain. Although the authors do not discuss all the subsystems of the MetaQuerier, one gets a fair idea of the working of MetaQuerier and what they are trying to accomplish.

While the first paper describes the finding the structured content, the second paper i.e. **EntityRank: Searching Entities Directly and Holistically** discusses searching specific content from the unstructured data and unifying the results and displaying to the users according to their score. While the traditional search focuses on document as a data, the Entity Search focuses on entity as a data. They have compared the entity search from the traditional search by illustrating an example. The authors have discussed the core challenges of ranking these entities and a outlining a conceptual framework called *The Impression Model* of ranking these entities. They have also discussed the five characteristics upon which they have constructed The Impression Model namely Contextual, Holistic, Uncertainty, Associative and Discriminative. They went on to deriving an Entity Rank formula which is used to rank the results obtained. One major advantage of this system is that it can be built upon a traditional search engine as it uses the same infrastructure as the traditional search engine. Finally they presented the results of their experiments and also compared and contrasted the statistics about the accuracy of their approach with various other approaches.

Comments

The first paper, **Toward Large Scale Integration: Building a MetaQuerier over Databases on the Web** gives a high level explanation of MetaQuerier and only provides the issues faced and lessons learnt while developing such a system, but in a clear and concise manner. It does not talk about the design of the MetaQuerier in detail nor does it provide any experimental results. Nevertheless they mentioned certain useful insights while developing the system which might prove useful for any future work on this system.

The second paper, **EntityRank: Searching Entities Directly and Holistically** discusses in a clear and concise way of the working of the system and systematically explains the framework implemented for the Entity Rank scheme.

Questions asked during the presentation of the first paper were:

➔ Does all the links provided in Cazoodle.com go directly to the web page containing the information about that particular product?

The answer is Yes, it does go directly to the web page containing the information about the particular product. This is because the MetaQuerier repository contains all the information about the schema of the web site for which the information was queried after that it just becomes a matter of selecting the sources based on the domain and translating the queries of the user to the schema of that website.

Questions asked during the presentation of the second paper were:

➔ Does the system give details about the index on which the entity has been search, for eg: The total number of entities searched?

Although the authors have not discussed the details about how the results would be displayed to the user or what would be shown apart from the rank of the result and the link to the page, but this is quite possible because we have already indexed all the entities and all the information is available to us.

➔ Google also provides entity search, how would you contrast the system used by Google and that of what has been described by the authors? Are they the same?

Yes Google does provide some amount of entity search but the system used by Google and that what has been described by the authors is certainly not the same. Moreover Google provides specialized engines to search these entities, but the authors provide a more unified way of searching entities of different nature like a phone no. or a file of a particular format like pdf.