CSE 718 – SEMINAR REPORT

BUILDING STRUCTURED WEB COMMUNITY PORTALS:
A TOP-DOWN, COMPOSITIONAL, AND INCREMENTAL APPROACH

Presenter – Yogesh A. Vaidya
yavaidya@buffalo.edu

Overview:

Structured Web Community Portals are portals that extract and integrate information from raw Web Pages to present a unified view of entities and relationships in the community. These portals can thus provide users with powerful capabilities for searching, querying, aggregating, browsing, and monitoring community information.

The paper first comments on the disadvantages of approaches used by commercial domain community portals: lack of a standardized process and domain specific development. The paper then comments on the complex machine learning approaches used by the researcher community to construct structured community portals. After discussing these limitations the authors propose new Top-Down, Compositional and Incremental Approach for building Structured Web Community Portals. The paper concludes by comparing the accuracy (precision, recall and F1 measure) results obtained from the case study portal (DBLife) versus other research community portals like Rexa.

This paper recommends a Top-Down, Compositional, and Incremental approach to build structured web community portals. This approach emphasizes on exploiting the following common Web community characteristics:

- Web communities often exhibit 80-20 phenomenon. Hence it is recommended to start with 20% of most prominent sources, because these names are often already covers 80% of interesting community activities.
- Within a community, significant collections of entity/relation names are often available with just a manual scrapping of certain Web sources.
- Entity names are often designed to be as mutually exclusive as possible to avoid confusion by community members. Even the clash of names is limited to a small percentage of names in a small set of sources. Thus harder or stricter logic can be applied to these ambiguous sources while keeping logic simple for bulk of the sources.
- Most new and interesting sources/entities/relations will eventually be mentioned within the community as a form of "advertisement" at one of the sources. Thus web portal can be expanded by simply monitoring these specific sources.
- Entities and relations need not be extracted from all the possible sources but only a subset of sources can be considered for specific entities and relations. This facilitates simple operator implementations, better accuracy and efficiency.

To aid development of Web Structured Community Portals a CIMPLE workbench is developed. The workbench facilitates easy and quick development since a portal builder can make use of built-in facilities and the workbench comes provides implementations of various operators.

As a proof of concept, DBLife a structured web community portal focusing on database community was developed. The portal was developed in relatively small timeframe and is easy to maintain and gives results with good accuracy even though it uses simple operators.

Detailed Comments:

The paper provides proper explanation/reasons while commenting on the disadvantages of the current approaches. The paper also gives proper justification backed by results obtained while proposing the new approach. The paper focuses on keeping things simple by exploiting the common web community characteristics and smart source selection.

In the details of case study that has been provided a clear mention of simplicity of operators and approach used is made. The results obtained are highly accurate and are better than that given by contemporary research community portals (like Rexa) even when research community portals use complex machine learning techniques for entity extraction and matching.

For the questions raised as to how relevant source selection is done and how monitoring for new sources is done: A tool called RankSource is provided which gives the community builder relevant sources in the decreasing order of their relevance. A detailed approach as to how to identify potential new sources of information is also given with an example in case study.

The paper however does not discuss in detail (just makes a brief reference) on how help can be solicited from community members in a mass collaboration fashion and thus making portal maintenance easier and distributed.