WebLQM: A Web Community Examiner

Jesús Ubaldo Quevedo
Department of Computer Science
University of Houston
Houston, TX 77204-3475
713-743-3338
jquevedo@uh.edu

S. H. Stephen Huang
Department of Computer Science
University of Houston
Houston, TX 77204-3475
713-743-3338
shuang@uh.edu

ABSTRACT
WebLQM is a system with capabilities to locate, query and mine web communities on the Internet. WebLQM has a special way to define the World Wide Web, its contents and relations. This approach unifies several ideas for converting the WWW into a relational model to later extract patterns from cyber communities on the Web using existing mining techniques. Its main purpose is to discover rules associated with the community being studied.

Keywords
Data Mining, Web Query Language, Cyber Communities, World Wide Web

1. INTRODUCTION
It is really amazing the amount of information stored in the World Wide Web estimated in billions [2] of documents. It is definitely a good source for knowledge discovery / Data Mining. Viewing the web as a huge data set, we will apply our data mining techniques to a small manageable sub set called community. Communities are unique groups that share a common interest on a particular topic of interest of the extensive variety of themes exposed on the Internet. Those groups are being referred in several publications as "cyber communities", "web communities" or just "communities". Web communities are generally not explicitly defined on the Internet; therefore, there has been ongoing research to identify them. Search engines do not identify communities properly since they do not consider relevant aspects of them such as connectivity. However, they can be used as a start point to locate communities. Moreover, it has become almost impossible to manage the number of documents provided by a regular search engine when prompted for a specific query. For instance, when asking altavista for information related to "Java", there will be more than 11 million web documents referring to it. It is clear that this amount of data will be impossible to be analyzed by a human being. Intuitively, there may be more than one community allied to Java in this large amount of documents. One community may be associated with either the programming language Java, an island of Indonesia, brewed coffee or even with a sport team named Java.

2. RELATED WORK
Research about collecting information from the World Wide Web has been developed mainly in three areas such as query systems, web mining (resource/knowledge discovery) and community identification. Web query systems like W3QL [4], WebSQL [6], WOQL, [1], WebSSQL [9] and WebLog [5] retrieve information from the World Wide Web simulating some capabilties of a DML. Web Mining is defined in WebML [8], and web communities can be found in [3].

3. WebLQM
WebLQM is a system with a particular definition of the World Wide Web constituted by a web query language that can be oriented to find web communities on the Internet and coming after discover interesting association rules. WebLQM could be used to just query the World Wide Web; however, its distinct functionality will be unnoticed.

3.1 Relational Representation of the Web
Our approach considers two aspects of the WWW. First, the attributes of its content referred as objects or documents, and second, the connectivity among its objects referred as links. While extending previous approaches, we added the necessary attributes and relations to distinguish, to query and mine communities over the Internet. WebLQM first conceives the content of the Web as a set of either readable or unreadable objects. A readable object will be that which can provide us with some useful information about itself, i.e. html files, text files, pictures, Microsoft Documents, etc. On the other hand, an unreadable object gives us no useful information. We may be able to determine its size, time of last modification and other attributes, but that data by itself is considered useless since we are incapable of analyzing/mining it. Therefore, we emphasize our research in readable objects. Then we classify readable objects as either hyper-text documents (html, xhtml, etc.) or multimedia files (still images, video, sounds, ascii, pdf, etc). Next, WebLQM addresses connectivity (links) between two hyper-text documents or WebDocuments as the reference of one document to the other. Connectivity is visualized as a directed graph or digraph G=(V,E) where G is the WebGraph formed by V= { t : t ∈ WebDocuments} and E= { (u,v) : u, v ∈ V }; therefore, (u,v) ∈ E means that there is a direct link from document u to document Only links among documents are considered valid connections since V contains only WebDocuments.WebLQM produces the relations shown in Table 1 that consider both aspects of the WWW established previously. These schema are the "*independent relations*", and they symbolize the relational transformation of the web.

This relation contains all readable **hyper-text documents** in the Internet
Webdocuments (url, title, size, type, last_modified, no_of_tables, no_of_tags, has_applets, no_of_forms, chat_option, download_option, sells_something)
Next, we define relations for readable **multimedia files**
Images (url, image source, title, description, color, texture, fetch_time, last_modified, type)
Sounds (url, sound source, size, type)
Video (url, video source, size, type)

Other objects contained in the **hyper-text document**
No_of_tags (url, title, font, headers)
Tables(url, counter, No_of_rows, No_of_columns)
Forms(url, title, size)
Relations for **connectivity**
Links (from_url, to_url)

Table 1 : Independent Relations

There are other meaningful relations derived from independent relations called "*dependent relations*" Table 2. These new relations are not virtual tables or views like those used in traditional SQL systems.

Community(url,no_of_outgoing_links,no_of_incomming_links)
Includes(document, object, type)

Table 2 : Dependent Relations

3.2 WebLQM - Primitives

WebLQM will need several primitives to perform its purpose of identifying, querying and mining communities in the WWW. They are the basis to generate dependent relations. A complete explanation of WebLQM primitives can be found in [7].

PRIMITIVE	SYNTAX
Mentions	Mentions (Q, q)
Link_to	Link_to(x,y,n)
References_to	References_to(Q, x, n)
Length	Length(Q, n)
Intersect	Intersect(P,Q, R)
Outgoing_links	Outgoing_links(Q, q, n)
Sites	Sites(P,Q)

Table 3 : WebLQM Primitives

Query = { x : Outgoing_links(Q, "www.cs.uh.edu",1),x ∈ Q, References_to(R,x,1), length(R,3) }

This query lists all URLs of pages that are referred by exactly three other pages and are directly accessible from the computer science department of UH.

4. COMMUNITIES

First, we have to define our concept of a "Web Community". A community will be a set of sites that share a common interest on a particular topic. All of them having information related to the topic of interest. They must make references to other members of the group, and they need to be referenced as well by the group. Therefore, every pair of hyper-text documents in the undirected version of G' is connected by a path.

Community = { t : mentions(Q,q) , t ∈ Q, References_to (R, t, n), n=1, Intersect(P,Q,R), Length(P,m), m ≥ 1 }.

This definition of **minimal community** using WebLQM notation states that a page belongs to a community if contains information related to q and it is referred directly by at least one page that also contains q. A stronger definition of community will increase the value of m to be at least either 3 or 4. Increasing the value of m reduces the size of the whole community. Intuitively, we give more relevance to communities that are highly referenced and strongly connected. We could still strength even more the community by increasing the value of n, restricting the number of outgoing links, and considering web sites for set P,Q and R instead of web documents. An example of stronger community would be:

Community = { t : Sites(Q,Community), t ∈ Q, Outgoing_links(Q2,t,n1), n1≥3, Intersect(P1,Q2,Q), Length(P1,n2),n2 ≥ 3, References_to (R, t, n), n=2, Intersect(P,Q,R), Length(P,m), m ≥ 2 }.

5. WebLQM - IMPLEMENTATION

WebLQM first prunes some documents from the web that are inapplicable to the theme of the community. During the pruning stage, it makes use of traditional search engine indexes. Later, it maps some data from the WWW into independent relational schemes and uses those schemes and some built-in primitives to derive additional dependent relations. This last feature allows us to use WebLQM as a pure web query language. A crucial part of the system is the identification of the community that matches the topic q and the WebLQM definition. Finally, WebLQM feeds conventional mining devices with sufficient dependent relations to produce interesting association rules. WebLQM is currently in its first phase of development. There are five basic components of the WebLQM architecture that can be found at [7].

6. REFERENCES

[1] G. Arocena, WebOQL: Exploiting Document Structure in Web Queries Master's Thesis, University of Toronto, 1997.

[2] J. Carriere, and R. Kazman, WebQuery: Searching and visualizing the Web through connectivity, in: Proc. 6th International World Wide Web Conference, 1997.

[3] David Gibson, Jon Kleinberg and Prabhakar Raghavan. Inferring Web communities from link topology. Proc. 9th ACM Conference on Hypertext and Hypermedia, 1998.

[4] D. Konopnicki and O. Shmueli. W3QL: A Query System for the World Wide Web. In Proc. 21st Int. Conf. On very Large Data Bases (VLDB), pages 54-65, Zurich, Switzerland, 1995.

[5] L. Lakshmanan, F. Sadri, and I. Subramanian. A Declarative Language for Querying and Restructuring the Web. In Proc. 6th Int. workshop on research Issues in Data Engineering, New Orleans, 1996.

[6] G. Mihaila. WebSQL - an SQL-like query language for the WWW, MSc. Thesis, University of Toronto, 1996

[7] J.U. Quevedo and S. Huang. Locating, Querying and Mining Web Communities using WebLQM. Technicla Report. Computer Science Department, University of Houston 2001.

[8] O. R. Zaiane: Resource and Knowledge Discovery from the Internet and Multimedia Repositories ,Ph. D. Thesis, Simon Fraser University, March 1999.

[9] C. Zhang, W. Meng, Z. Zhang and Z. Wu. WebSSQL - A Query Language for Multimedia Web Documents. IEEE Conference on Advances in Digital Libraries (ADL'00), Washington, D.C., May 2000.