Bruno, Emmanuel; Le Maitre, Jaques; Murisasco, Elisabeth

Hypertextual Navigation in the SgmlQL Language: Browsing, Querying and Restructuring Web-like Networks

Computación y Sistemas, vol. 2, núm. 4, abril-junio, 1999, pp. 165-168

Instituto Politécnico Nacional
Distrito Federal, México

Available in: http://www.redalyc.org/articulo.oa?id=61520402
Hypertextual Navigation in the SgmlQL Language: Browsing, Querying and Restructuring Web-like Networks

Emmanuel Bruno, Jacques Le Maitre, and Elisabeth Murisasco
GECT, Equipe Systèmes d'Information MultiMédia
Université de Toulon et du Var, BP 132, 83957 La Garde cedex France
{bruno, lemaître, murisasco}@univ-tln.fr

Article received on November 15, 1998; accepted on April 19, 1999

Abstract

Due to the growth of networks, the notion of document has evolved to the notion of hyperdocument. It is more and more usual to fragment documents in a set of files referencing each others. These files can be distributed over a network. The manipulation of the whole document requires the ability to browse this graph-like structure. That is why we need tools that are able to browse these structures, to query them and eventually to restructure them. SgmlQL is an OQL-like functional language which manipulates SGML documents. We present in this paper an extension of this language to express a controlled navigation on Web-like networks.

Keywords:
Query language, SgmlQL, Web, SGML, XML.

1 Introduction

Due to the growth of networks, the notion of document has evolved to the notion of hyperdocument. It is more and more usual to fragment documents in a set of files referencing each others. These files can be distributed over a network. The manipulation of the whole document requires the ability to browse this graph-like structure. Manual navigation remains the more usual way to browse it either from a known URL or from the output of an index server. Nevertheless, it is difficult to avoid digressions and even getting "lost in the cyberspace"; it is mandatory to read each accessed document to check its reliability. Moreover, it is difficult to take advantage of hypertext specific organizations such as lists or trees. That is why we need tools that are able to browse these structures, to query them and eventually to restructure them.

Some languages have been proposed to deal with documents distributed over a network (Konopnicki and Shmueli, 1995; Laksmanan et al., 1996; Mendelzon and Milo, 1997; Vercoutre and Paradis, 1998). Most of them focus more on the expression of the navigation between documents than on the operators to query reached documents. Our proposition is to add navigation capabilities to an existing query language for documents: SgmlQL (Le Maitre et al., 1996), therefore every component of each document can be used to control the navigation. Moreover all these components can also be extracted and reused in new documents. SgmlQL is one of the languages (Blake et al., 1994; Christophides et al., 1994; Yan and Anneveinkel, 1994) developed to manipulate SGML (Goldfarb, 1990) documents.

SgmlQL is an OQL-like functional language which can be used with any kind of document from which a tree-like structure can be extracted, in particular SGML, HTML, XML (XML, 1998) or even RTF. SgmlQL enables to extract parts of documents and also to construct new documents by means of tree traversing and tree transformation operators. Its distinctive feature is that it
does not use a transcription into a classical database schema, and that only a partial knowledge of the structure of the document is needed to query it.

We present in this paper an extension of the SgmIQL language. The purpose of this extension is to express path filters and to use them by means of a specific operator. This operator controls the navigation through a set of constraints on the followed path, on the links and on the content of traversed documents. Section 2 shows how a path filter is expressed and used in a SgmIQL query. Section 3 focuses on the problem of cycles. Section 4 concludes.

2 An Operator for Controlled Navigation

In a hypertext, hand-browsing is the usual way to access documents. However it is difficult to move on a graph-like structure with a sequential access, while keeping the hypertext structure in mind and not to get lost because of the huge number of documents and the existence of cycles. Moreover, it is not possible to ensure that the navigation is exhaustive and that the constraints which the user wants to apply, are continually checked.

2.1 Path Filters

In (Bruno, 1998), we proposed to separate the start of the navigation from the path to follow. Therefore, we view a path like a list of pairs \((\text{link}, \text{node})\).

The \text{link} component catches all the data needed to define the relation between the source and the target documents. This component is represented as an SGML element provided with the attributes \text{base} and \text{href}, which refer to the source and the target documents, \text{label}, which labels the link, \text{rel} and \text{rev}, which specify the semantics of the relation between the source document and the target document and its inverse.

The \text{node} component is built from the document which is viewed like a node of the graph. The root tag is enriched with the following attributes: \text{url} which identifies the document, \text{mimetype}, \text{length} and \text{modif}, which give its date of creation or its date of last update.

We can now define path filters as regular expressions:

- A link filter is either \(\rightarrow\) which filters any link, or \(=\) which filters a null link.
- A node filter is either \(_\) which filters any document, or an url which filters the referenced document.
- If \(F_1\) is a link filter and \(F_2\) is a node filter then \((F_1, F_2)\) is a path filter. It filters any path \((l, d)\) such that \(l\) is filtered by \(F_1\) and \(d\) is filtered by \(F_2\).
- If \(F_1\) and \(F_2\) are path filters then \(F_1 \wedge F_2\) filters any path with is filtered either by \(F_1\) or by \(F_2\).
- If \(F\) is a path filter then \(F^*\) filters either the empty path or a path composed of \(n\) \((n>0)\) consecutive subpaths filtered by \(F\). To impose a bound on the time \(t\) and the length \(l\) of the navigation, an \(F^*\) filter can be suffixed by \([t, l]\). These two parameters ensure that the number of filtered paths is finite.

For instance, the expression \((\rightarrow, _\), \(\rightarrow\), \(_\)) filters a path of length 2. Variables can be bound to any component of a path filter using the keyword \text{as}. These variables are used to define constraints on the followed links or on the content of the documents accessed during the navigation. They can also be used to extract some components of these links and documents. Note that variables in an \(F^*\) filter cannot be seen outside \(F\) (see section 2.3).

2.2 The Navigate Operator

In order to produce the set of all paths matching a given pattern from a start document, we propose a specific operator called \text{navigate}. The expression \("\text{navigate } u F\"\) starts from the url \(u\) and for each path filtered by \(F\), instantiates the variables of \(F\). Therefore, the \text{navigate} operator produces a list of environments. For instance:

\[
\text{navigate } \"http://www.univ-tln.fr\" \\
(\rightarrow \text{as } l_1, _\text{as } d_1)
\]

- filters all the paths of length 1 starting at the url \(http://www.univ-tln.fr\),
- produces the list \([\{\langle l=v_{11}, \langle d=v_{12}\rangle, \ldots, \langle l=v_{n1}, \langle d=v_{n2}\rangle\}\rfloor\] if \(n\) paths are filtered and if for each filtered path \(i, v_{1i}\) is a link element and \(v_{2i}\) a node document.

This operator can be used in the \text{from} clause of a SgmIQL query. For example:

\[
\text{Q1:}
\begin{align*}
\text{select} & \quad \text{first TITLE within } d \\
\text{from} & \quad \text{navigate } \"http://www.univ-tln.fr\" \\
& \quad (\rightarrow \text{as } l_1, _\text{as } d) \\
\text{where} & \quad l_1 \rightarrow \text{label match } \text{"java"}
\end{align*}
\]

returns the list of the titles of the documents reached from the Web server of the University of Toulon through a link whose label contains the word “java”.

The user can use a easy-to-read presentation: the expression \(\rightarrow \text{as } \text{var}\) is the same as the expression \text{var} and the expression \(_\text{as } \text{var}\) is the same as the expression \text{var}.

The query Q1(bis) built on top of query Q1 shows how new SGML elements or documents can be constructed.

\[
\text{Q1(bis)}
\begin{align*}
document body: \\
\text{let } \$\text{result } =
\end{align*}
\]
2.3 Guards

In the above mentioned example, the constraint on the link is set in the where clause. Setting the constraints in this way delays their checking after the generation of each path. This causes two kinds of problems: optimization (it is better to check constraints during the navigation) and expression (when arbitrary length filters have to be constrained). One solution can be to add guards to filters as it is done with the Caml functional language (CAML). A guard is a condition added to a filter in order to check constraints during the navigation. In SgmIQL, a guard is introduced by the clause when. For example, using a guard, query Q1 can be optimized as follows:

\[ (Q2) \]
\[
\text{select first TITLE within } \$d \\
\text{from navigate } \text{"http://www.univ-ltn.fr"}
\]
\[
(\$1 \text{ when } \$1\rightarrow\text{LABEL} \text{ match } \text{"java"}, \$d)
\]

The label of each link is now checked during the navigation, and only needed documents are loaded in main memory.

Let \( F^* \) be a filter. Guards are mandatory to set constraints on \( F \). Let us consider, for example a linked list of documents (each document contains a link to the next one). The query selecting the title of each document from this list can be expressed as follows:

\[ (Q3) \]
\[
\text{select first TITLE within } \$d \\
\text{from navigate } \$url\_start
\]
\[
(\$1 \text{ when } (\$1\rightarrow\text{REL} \text{ match } \text{"next"}), \ldots)^* \quad (1) \\
(=, \$d) \quad (2)
\]

Lines (1) and (2) provide access to each document within the list (the variable \$d will successively contain each document). Line (1) filters each path starting at \$url\_start of length 0 up to \( n \) (number of documents in the list). For each filtered path, line (2) gives access to the last document in the path (null link =).

2.4 Using Index Servers

Finally, like WebSQL (Mendelson and Milo, 1997), SgmIQL enables to call an index server (Altavista for instance) to select documents about a given topic by means of a specific operator: about. For example, the following query selects the title of the documents dealing with Java beans referenced by documents about Java provided by an index server:

\[ (Q4) \]
\[
\text{select first TITLE within } \$d \\
\text{from $url about } \text{"java", navigate $url}
\]
\[
(\$1 \text{ when } (\$1\rightarrow\text{LABEL} \text{ match } \text{"beans"}), \$d)
\]

3 Dealing With Cycles

In a hypertext a lot of documents are referring to each other. Consequently it is necessary to be able to deal with cycles to avoid infinite navigation. The first solution could be to forbid traversing twice the same document. This is easy to implement but is semantically insufficient because the same document can be differently processed at different steps along the navigation.

The best solution would be to take into account the way each document is accessed: not only the link reaching the document but also the whole navigation before. This needs an in-depth study of the semantics of the links. This remains to be done. Between these two solutions we have chosen to keep track of the links reaching a document and to avoid returning to the same document by the same link.

Note that if a dangling link occurs along a navigation (the target document does not exist) this path is left and next one is tried.

4 Conclusion

Because of the increasing use of hyperdocuments over networks, it is essential to provide tools that make their localization, extraction and reuse easier.

We have presented in this paper an operator (Bruno, 1998) in order to manage fragmented data easily. This operator uses a path filter with associated constraints, to carry out a controlled navigation over a hypertext from either a start point (a valid URL) or the URL returned by an index server. Combined with standard SgmIQL operators, the user can express a query which:

- specifies a navigation where conditions are set on the followed path, and on the traversed links and documents,
- extracts hypertext fragments,
• reuses these fragments with classical SgmlQL operators to build new documents (or hyperdocuments).

This paper also focuses on the problem of cycles and presents a detection mechanism. Finally, a language to express constraints with guards is proposed: \( i \) to check these constraints in arbitrary length paths; \( ii \) to check constraints as soon as possible during the navigation. All the functionalities described in this paper are implemented and validated (SgmlQL). Other languages like WebSQL (Mendelzon and Milo, 1997), Weblog (Laksmanan et al., 1996), W3QL (Konopnicki and Schmueli, 1995) or RIO (Vercoustre and Paradis, 1998) deal with the same set of problems. Our extension is close to WebSQL or RIO. However our \texttt{navigate} operator is part of the SgmlQL language. Therefore a query can constrain any hypertext component and build a new document, without any translation in a classical database schema. The perspectives of our work are to produce new hypertext in order to offer dynamic views (which could be specific to user profiles), to add connecting capabilities with servers processing documents to add meta-data and to study the semantics of links, in particular to solve the cycles problem.

References


XML http://www.w3.org/TR/REC-xmll

Yan T. W. and Annevelink J., “Integrating a Structured-Text Retrieval System with an Object-Oriented Database System”, Proceedings of the 20th Conf. on Very Large Data Bases (VLDB ’94), Santiago (Chile), 1994.

Emmanuel Bruno is a Ph.D. student in Computer Science at the University of Toulon. His research deals with information systems and databases. In particular, he studies a way to query semi-structured distributed and fragmented documents. He is a member of the SIMM (Multimedia Information Systems) team.

Jacques Le Maitre is a Professor in Computer Science at the University of Toulon. He leads the SIMM (Multimedia Information Systems) research team of the GECT laboratory. His research interest include databases, structured documents and functional languages.

Elisabeth Murisasco is an Assistant Professor in Computer Science at the University of Toulon. Her current research topics are textual databases and distributed hypertext databases. She is a member of the SIMM (Multimedia Information Systems).