Changing Vision for Access to Web Archives

Zeynep Pehlivan
LIP6
University P. and M. Curie
Paris, France
zeynep.pehlivan@lip6.fr

Anne Doucet
LIP6
University P. and M. Curie
Paris, France
anne.doucet@lip6.fr

Stéphane Gançarski
LIP6
University P. and M. Curie
Paris, France
stephane.gancarski@lip6.fr

ABSTRACT

Since late 90s, there has been a large investment in web archiving. Accessing these huge information sources is getting more and more attention. Web archive users profiles differ from casual web users profiles. Archive users need to analyze, evaluate and compare the information which requires complex queries with temporal dimension. These queries can not be performed by currently proposed access methods: wayback machine, full-text search and navigation. In this paper, we address this requirement by proposing a data model and a temporal query language for web archives which take into account different topics in web pages and the issues related to web archiving.

In our approach, a captured web page is visually segmented into semantic blocks. A concrete block notion is introduced to represent these different semantic blocks. A concrete block is a triplet: frame block which keeps properties of a block, the content (textual and:or non-textual) and the importance accorded to a block. Each of them is timestamped with a period called *validity*. A web page, identified with an url, is a set of concrete blocks and a web site is a set of pages. Pages and sites are generated dynamically by manipulating concrete blocks when needed. Operators for data manipulation, navigation and ranking are also proposed.

Keywords

Web Archiving, Access to web archives, Temporal Query Languages, Temporal Data Model

1. INTRODUCTION

Web archives, in short *WAC*s, aim to preserve the history of large portions of the web. They represent thus a huge information source, potentially greater than the web itself. This makes web archiving an active research area with numerous opened issues like crawler optimization, storage models, etc. An overview of main issues is presented by Masanès [21]. It is interesting to notice that the researches on accessing web archives focus on extension of existing access methods to web.

For instance, the "Wayback Machine" [27] which allows users to see captured versions of web pages over time, is the best known access method to web archives. The others methods currently provided are well-known web access methods: full-text search and navigation within the requested time range. These methods are powerful for casual users, who search the web for general information and represent the largest proportion of web users. According to the "Web Archiving User Survey" [24], the web archive users profiles consist of historians, journalists, lawyers, students etc. more than casual users. It is also underlined that the main reason for using web archives is research activity. Web archive users need to analyze, compare and evaluate the information. In order to achieve this, web archive systems must provide tools to execute complex queries.

Consider a researcher who studies how French media covered the event "earthquake at Haiti in 2010" over last year. At the very beginning of her research she would like to know the number per month of web pages, in domain *.fr*, referring to earthquake. This kind of queries can not be performed by wayback machine, full-text search or navigation methods but with an efficient query language. A number of possible user scenarios over web archives are listed and illustrated with technical requirements in "Use cases for access to Internet Archives" [6]. A significant number of them requires complex query capability to be handled.

Multi-topics and noisy information on a web page affect the search performance. In recent years, there has been an increasing interest in web based searching by taking these different topics as units of retrieval and by eliminating noisy information like copyrights, navigation bars etc. (e.g [12, 15, 28]). To achieve this, different page segmentation algorithms are proposed (e.g [20, 14, 16]). They partition web pages into non-overlapping hierarchical blocks where each block deals with a different topic or only contains noisy information. As web archives are temporal collections of web pages, block-based approach needs to be extended with temporal dimension in order to be used in web archive search.

The role of search engines for web users is non-negligible. On the other hand, as seen in use cases [6], most of the time, they are not sufficient to meet the need of web archive users. New approaches to explore web archives with complex queries are needed. An appropriate query language, that enables temporal search besides content queries and structural queries (not only for the hypertext structure, also for web page's internal structure) is especially required for formulat-

TWAW 2011, March 28, 2011, Hyderabad, India.

ing complex queries. There are a number of query languages proposed for querying web data (e.g WebSQL, W3QL, WebLog, WebOQL, etc.) but most of them are not suitable for web archives due to a lack of temporal dimension and a lack of handling challenges related to web archives. Actually, our aim is to integrate the database approach and the IR approach.

In this paper, we present a conceptual model for web archives and basics of a query language based on this model. In our approach, visual blocks, which are extracted from a page, are used as unit of retrieval rather than a whole page. Each element of the model is timestamped with a period called *validity*. Operators for data manipulation, navigation and ranking are proposed. The query language that we propose for web archives:

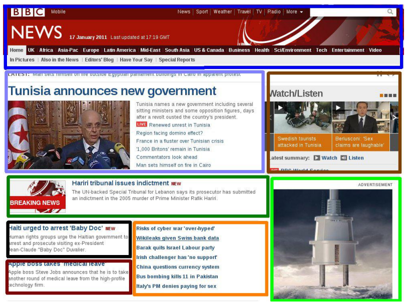
- Enables block-based search
- Takes into account incompleteness
- Eliminates duplicates
- Enables temporal ranking and grouping
- Is user-friendly

We organize our paper as follows. In Section 2, we present the features of the query languages for web archives. In Section 3, we present the related works. In the following section, conceptual model is presented with related operators. Before the conclusion in Section 5, two use cases are explored in Section 5.

2. FEATURES OF WAC QUERY LANGUAGE

The overarching design goal of our approach is to offer a query language for web archive users. In this section, we describe the features of this query language and the rationale behind them.

Block-Based Search: Today, web pages contain various topics. A typical example is the web pages of newspapers/TV channels like www.bbc.co.uk/news (Figure 1) where multiple blocks with unrelated topics are marked with different colors. A web page with matched query terms in the same region is more relevant than a web page with matched terms distributed over the entire page. Besides increasing keyword search performance, the segmentation gives a structure to the web page for structural queries.



BBC Mobile News | Sport | Weather | Travel | TV | Radio | More

NEWS 17 January 2011 Last updated at 17:19 GMT

Home | UK | Africa | Asia-Pac | Europe | Latin America | Mid-East | South Asia | US & Canada | Business | Health | Sci/Environment | Tech | Entertainment | Video

In Pictures | Also in the News | Editors' Blog | Have Your Say | Special Reports

LATEST: Man sets himself on fire outside Egyptian parliament buildings in Cairo in apparent protest

Tunisia announces new government

Tunisia names a new government including several sitting ministers and some opposition figures, days after a revolt ousted the country's president.

LIVE Renewed unrest in Tunisia

Region facing domino effect?

France in a fluster over Tunisian crisis

'1,000 Britons' remain in Tunisia

Commentators look ahead

Man sets himself on fire in Cairo

Watch/Listen

Swedish tourists attacked in Tunisia

Berlusconi: 'Sex claims are laughable'

Latest summary: Watch Listen

BBC World Service

BREAKING NEWS

Hariri tribunal issues indictment NEW

The UN-backed Special Tribunal for Lebanon says its prosecutor has submitted an indictment in the 2005 murder of Prime Minister Rafik Hariri.

ADVERTISEMENT

Haiti urged to arrest 'Baby Doc' NEW

Human rights groups urge the Haitian government to arrest and prosecute visiting ex-President Jean-Claude "Baby Doc" Duvalier.

Apple boss takes 'medical leave'

Apple boss Steve Jobs announces that he is to take another round of medical leave from the high-profile technology firm.

Risks of cyber war 'over-hyped'

Wikileaks given Swiss bank data

Barak quits Israel Labour party

Irish challenger has 'no support'

China questions currency system

Bus bombing kills 11 in Pakistan

Italy's PM denies paying for sex

Figure 1: Multi-topics in a web page

Previous works [20, 14] show that a page can be partitioned into multiple blocks and, often, the blocks in a page have a different importance. The importance weights are accorded to different blocks inside a web page according to their location, area size, content, etc. Using a web page as unit of retrieval does not take into account these different regions and their importance. In our model, we add this notion to enrich our query language. With the block based search and the importance, we are able to answer queries like "Find pages mentioning Obama and Sarkozy in same block" or "Find pages mentioning Obama in their most important blocks" which returns more relevant information and also helps to reduce the number of results.

Incompleteness and Temporal Coherence: Due to the limited sources, all web archives are incomplete (i.e they do not contain all possible versions of all the pages on the web) and we should query them as they are. If a user asks for a version at *t*, and if the archive does not have versions at *t* but it has the versions at *t-2* and at *t+2*, the access model should decide or support different choices to get the closest version to *t*.

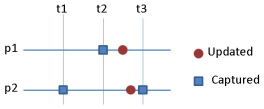


Figure 2: Temporal Coherence

Temporal coherence is another issue in web archiving. The main reason is the dynamic structure of the web. It changes continuously in an unpredicted and unorganized manner. The web sites politeness constraints and the limited allocated resources do not allow archiving a whole web site at once, at the same moment. For example, in Figure 2, for a temporal navigation starting from the version of p1 at *t2*, p2 at *t1* is coherent, while p2 at *t3* is incoherent. The access model should be able to find the most coherent version.

Temporal coherence and incompleteness lend to broken or defected links which disable a complete navigation and bring access to a standstill. An access model for web archives should take into account these issues.

Duplicates: Web archives has much more duplicated contents than the web itself. In web archives, that kind of duplicates can occur in two different cases: two versions of the same URL crawled at different times or the same content is pointed by several different URLs. According to [18], 25% of the documents in web archives are exact duplicates. Duplicates complicate the search and the result visualization.

Temporal Ranking and Grouping: Ranking and grouping are most common ways to deliver query results for the large-scale data. For web archives, temporal dimension must be included in both ranking and grouping process. Ranking has also a dynamic nature in WACs. For example, a page archived mentioning "Obama" at 1999 should not have the same ranking result when querying at 2000 and at 2010.

Temporal Logic: Support for temporal logic operators enriches the language. For example, a researcher who wants to analyze the effects of the arrest of Julian Assange can execute a query: "Find pages linked to wikileaks.org after Julian Assange's arrest in London".

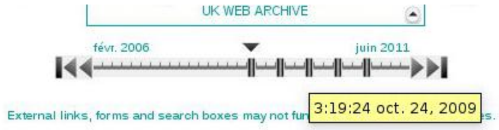
User-Friendliness: This query language will be used by researchers and casual web users. So it should enable users to find information without long-term training. Simple queries (like keyword search) should be expressed straightforwardly. Complex syntax should be used only to express more complex queries. We believe that with an advanced GUI, users can avoid from writing "codes".

3. RELATED WORKS

In this section, we briefly summarize related works in three different areas concerning the access to web archives: Web archiving, query languages for web and block-based search

Web Archiving

To explore web archives, traditional access methods, i.e navigation and full-text search are proposed by web archiving initiatives [3, 4, 5]. In fall of 2001, Internet Archive(IA) [3] launched its collaborated project with Alexa Internet called "Wayback Machine" [27]. It allows users to go back in time and view earlier versions of a web page for a given URL. The inconvenience of this method is the necessity of knowing exact URLs. There is another way of navigating in web archives : navigation between different versions. Some web archive initiatives propose a navigation tool like UK Web Archives(UKWAC) [5] as seen in Figure 3 to facilitate the navigation between versions. By using the cursor, users can browse different versions easily.



External links, forms and search boxes may not fun s.

Figure 3: UKWAC navigation tool

The increasing number of national web archives, diversity of existing works led to the establishment of the International Internet Preservation Consortium (IIPC) [2] in Paris at 2003. The aim is to develop common standards, tools and techniques for web archiving. One of the current projects of IIPC, called WERA, is an archive access solution for searching and navigating web archives. It allows a full-text search besides wayback machine style search and it is based on the NWA(Nordic Web Archive) toolset [19] and the NutchWAX [26] full-text indexer. NutchWAX is extension of Nutch (an open source search engine based on Lucene java for searching and indexing) for searching web archive collections.

Today, most of web archive initiatives use the wayback machine to support URL indexing and search. NutchWAX is used to enable full-text indexing and search. Our motivation is focused on enabling complex queries which can not be performed by existing methods.

Query Languages for Web

A number of Web query languages have been developed in the past (e.g WebSQL, W3QL, WebLog, WebOQL, etc.) [17]. All those languages are intended for online queries on the web. From the perspective of web archiving, the most important inconvenience of those query languages is their lack of temporal dimension and the lack of handling challenges related to web archives like temporal coherence, incompleteness etc.

WebBase [23] project at Stanford University is a web repository project that aims to manage large collections of web pages and to enable web based search. A web warehouse is interpreted simultaneously as a document collection, as a directed graph and as a set of relations. For web based search, a query language with the notions of ranking and ordering is proposed. However, one copy of each page at a time is archived, thus, no temporal dimension is provided in that project.

A web warehousing system called WHOWEDA(Warehouse of Web Data) [10] proposed by Web Warehousing and Data Mining group at the Nanyang Technological University in Singapore aims to store and manipulate web information. It stores extracted web information as web tables and provides web operators to manipulate those tables. Its data model is based on nodes (pages) and links (hyperlinks) objects. Links do not have any time-related attribute. Any change in the last-modified time of the web document results in a new node. Content of a web document is represented as "node data tree". For HTML documents, it is a HTML DOM tree. The flexibility of HTML syntax might cause mistakes in DOM tree structure. In addition, however, DOM tree is powerful for presentation in the browser, it is not introduced for description of the semantic structure of the web page. In WHOWEDA, the internal semantic structure of a page is not modeled, thus it only allows queries to be specified on the whole content of the document. WHOWEDA only focuses on user interested web sites which constitute a much smaller scale than web archives.

Block-Based Search:

Using semantic blocks in web pages as an unit of information retrieval is an active research area. The vector space model customized with importance and permeability (the indexing of neighbors blocks) is proposed in [11] without temporal dimension. In [13], after segmenting each page into non overlapping blocks, an importance value is assigned to each block which is used to weight the links in the ranking computation. A block-based language model is proposed in [28]. As far as we know, there is no approach with temporal dimension in block-based search, ranking and language modeling. Our approach is based on visual page segmentation of web pages [22] and uses an importance model proposed in [25].

In conclusion, wayback machine, full-text search and navigation are the only applied solutions to access to web archives. Most of the web query languages do not contain temporal dimension for querying historical data. Different topics in a web page are not handled in search except recent works like [12, 15, 13] which suffer from lack of temporal dimension to be used for web archives. In our approach, we propose a data model which takes into account different regions of a web page with associated importance and temporal dimension.

4. WAC QUERY LANGUAGE

The details of WAC query language are described in this section. After briefly introducing the interpretation of temporal dimension in Section 4.1, we describe our data model in Section 4.2 and list the operators in Section 4.3.

4.1 Temporal Dimension

4.1.1 Time on the Web

Integrating time into data models, query languages and database management system implementations still attracts the attention of researchers. We focus here on two questions: which time to use and how to represent it.

In the context of web archiving, there are different types of temporal information: time in content, HTTP headers, crawled time.

Temporal expressions can be found embedded in the content of a web page. Those expressions are explained in three categories in [8]: Explicit, Implicit and Relative. Temporal expressions that can be converted directly to chronons are explicit expressions (e.g "December 24,2010..."). Names of holidays or events which can be anchored in timeline are considered as implicit expressions like "Christmas Day 2010". Relative expressions can not be anchored in timeline directly like "today", "on Wednesday" etc. Extraction of temporal expressions from documents are active research field. Temporal information in content is excluded from our model.

The temporal fields in HTTP headers contain `Date` , `Last-Modified` and `Expires`. Servers send `Last-Modified` header with date time value when the content was changed last time. `Last-Modified` header is mostly used as a weak validator. According to the definition of HTTP specification, a validator that does not always change when the resource changes is a weak validator. The meaning of `Last-Modified` depends on the implementation of the origin server and the nature of the original source (files, database gateways, virtual objects, etc.) [1]. The header `Expires` indicates when a document stops being fresh. If it is used correctly, this header indicates a validity period for the document and an exact date for recrawling. But it is an unreliable tool: many servers do not provide the header or provide with zero or with low expiration delay.

HTTP header `Date` represents the date and time at which a web page is returned from a HTTP server upon request by the HTTP client. Crawled time is the time that crawlers capture the snapshot of a page. Web crawlers set the value of the `Date` field in crawled documents to the date that a document is crawled, unless they are configured otherwise. It is the most trustworthy time because it does not depend on host servers configurations. In our model, crawled time is used as the basic time dimension but we should underline the fact that our model supplies also other choices.

Our model uses Allen's interval-based representation of temporal data [7] where intervals are addressed as primitive time elements. Each element is temporally stamped by a period, called *validity*, defined as a time interval $[t_s, t_e)$ where t_s represents a starting time point and t_e is an ending time point. *now* represents the current time and is assumed to be larger than any timestamps.

4.1.2 Time in queries

There are two kinds of temporal queries in temporal information retrieval: time-point and time-interval. In time-point queries, a time point t is given in the query and the results should contain data whose validity contains t. In time-interval queries, an interval [t1,t2) is given in the query and data whose validity overlaps with [t1,t2) are returned as result. A time-interval query is also called a time slice query.

In our model, all queries are treated as time interval queries. For example, a time point query like "2010/08/19" is converted to time-interval [2010/08/19,2010/08/20).

4.2 Conceptual Model

In our model, each concept has its temporal and non-temporal definition. Non-temporal definitions do not contain temporal attributes and are denoted by $concept^-$. We follow the example in Figure 4 to illustrate the main features of the model. In this example, we have a web page crawled at t1,t2 and t3 without structural changes. The page is segmented in three blocks marked with green, blue and pink. In our approach, visually segmented blocks are called *concrete blocks*. Each concrete block has a frame block, content and importance. A page is a set of concrete blocks and a website is a set of pages.

Frame Block

Web page segmentation returns a set of non-overlapping hierarchical blocks. Frame Block (fb) keeps properties of a block: the url to which the block belongs, a Dewey identifier that indicates its place in the page structure and its validity interval. If a frame block disappears and reappears later, it is considered as a new frame block. A frame block is defined as follows:

$$fb = (URL, DeweyID, [fbt_s, fbt_e))$$

$$fb^- = (URL, DeweyID)$$

For the in Figure 4, frame blocks are:
$blue : (www.bbc.co.uk/news, 1, [t1, now))$
$pink : (www.bbc.co.uk/news, 2.1, [t1, now))$
$green : (www.bbc.co.uk/news, 2.2, [t1, now))$

Content

There are two kinds of contents: non-textual and textual. Non-textual content are images, videos etc. in a web page. In our model, treating non-textual content is out of our scope, thus, if the content has non-textual content, *binary* is introduced to provide support for further works. Textual content is treated as a bag of words (after elimination of stop words, stemming etc.) in a block. A *validity* attribute allows to trace the content changes over its frame block. Its validity should be included in the validity of its frame block. A block b can contain only one textual content but several non-textual contents. A content of a block b is defined as:

- if b has only textual content

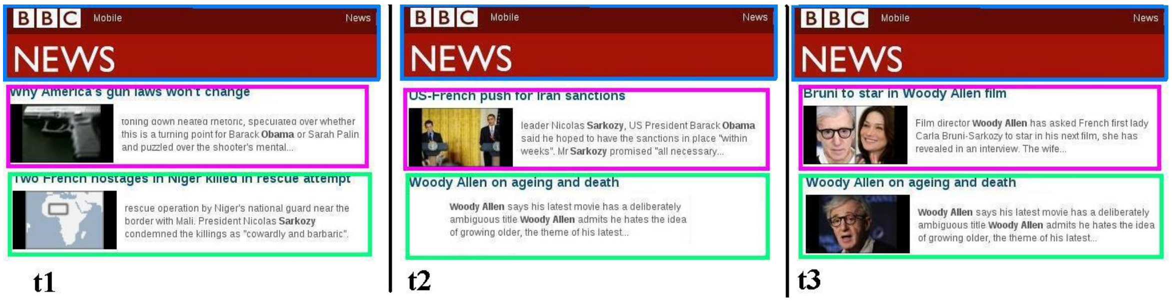
$$c = (text, [ct_s, ct_e))$$
$$c^- = (text)$$

- if b has only non-textual content

$$c = \{(binary_1, [ct_s, ct_e)), ..., (binary_n, [ct_s, ct_e))\}$$
$$c^- = \{(binary_1), ..., (binary_n)\}$$

- if b has textual and non-textual content

$$c = \{(text, [ct_s, ct_e)), ..., (binary_n, [ct_s, ct_e))\}$$
$$c^- = \{(text), (binary_1), ..., (binary_n)\}$$



BBC Mobile News

NEWS

Why America's gun laws won't change

toning down heated rhetoric, speculated over whether this is a turning point for Barack **Obama** or Sarah Palin and puzzled over the shooter's mental...

Two French hostages in Niger killed in rescue attempt

rescue operation by Niger's national guard near the border with Mali. President Nicolas **Sarkozy** condemned the killings as "cowardly and barbaric".

BBC Mobile News

NEWS

US-French push for Iran sanctions

leader Nicolas **Sarkozy**, US President Barack **Obama** said he hoped to have the sanctions in place "within weeks". Mr **Sarkozy** promised "all necessary...

Woody Allen on ageing and death

Woody Allen says his latest movie has a deliberately ambiguous title **Woody Allen** admits he hates the idea of growing older, the theme of his latest...

BBC Mobile News

NEWS

Bruni to star in Woody Allen film

Film director **Woody Allen** has asked French first lady Carla Bruni-Sarkozy to star in his next film, she has revealed in an interview. The wife...

Woody Allen on ageing and death

Woody Allen says his latest movie has a deliberately ambiguous title **Woody Allen** admits he hates the idea of growing older, the theme of his latest...

Figure 4: Example

In the example in Figure 4, contents are listed as followed:

$$\left.\begin{array}{l}((News),[t1,now))\\(binary,[t1,now))\end{array}\right\}blue$$

$$\left.\begin{array}{l}((Obama,America,gun,laws...),[t1,t2))\\((binary,[t1,t2)))\end{array}\right\}pink$$

$$\left.\begin{array}{l}((French,hostage,Niger...),[t1,t2))\\(binary,[t1,t2))\end{array}\right\}green$$

$$\left.\begin{array}{l}((Sarkozy,Obama,US,French),[t2,t3))\\(binary,[t2,t3))\end{array}\right\}pink$$

$$\left.\begin{array}{l}((Woody,Allen,ageing...),[t2,now))\\(binary,[t3,now))\end{array}\right\}green$$

$$\left.\begin{array}{l}((Sarkozy,Carla,Woody,Allen...),[t3,now))\\(binary,[t3,now))\end{array}\right\}pink$$

Importance

The blocks in a page have different importances as explained in Section 2. We define the importance as:

$$i = (alpha, [it_s, it_e))$$

$$i^- = (alpha)$$

The importance of a block, denoted *alpha*, depends on its location, area size, content, etc. It is calculated according to the model proposed in [25]. Validity of importance is not equal to the validity of the content in the same block. Although its content stays unchanged, the importance of a block can change by adding / deleting links or images, or by updating blocks size in the page.

For the example in Figure 4, Importances are listed as followed:

$(0.4, [t1, now))$ blue	$(0.6, [t2, now))$ pink
$(0.2, [t1, t2))$ pink	$(0.3, [t2, t3))$ green
$(0.1, [t1, t2))$ green	$(0.4, [t3, now))$ green

Concrete Block

Concrete Block is a region in a web page. It is defined as follows:

$$cb = (fb, \{c\}, \{i\})$$

In this triplet, *fb* is a frame block, $\{c\}$ is a set of its content ordered by validity attribute and $\{i\}$ is a set of its importance ordered by validity attribute.

In Figure 4, Concrete Block for the green block is described as followed:

$$\left(green, \left\{\begin{array}{c}(French, hostage..., [t1,t2))\\(binary,[t1,t2))\\(Woody, ageing..., [t2,now))\\(binary,[t3,now))\end{array}\right\}, \left\{\begin{array}{c}(0.1,[t1,t2))\\(0.3,[t2,t3))\\(0.4,[t3,now))\end{array}\right\}\right)$$

If we want to eliminate temporal nesting in data, for example to find different versions, T-FLAT operator can be used. Concrete block after applying T-FLAT operator is called *temporally flattened concrete block*, and denoted as *čb*. It has non-temporal triplet of frame block, its content and its importance and a validity which is equal to the intersection of validities of elements in triplet.

$$\check{c}b = \left((fb^-, c^-, i^-), [\check{c}bt_s, \check{c}bt_e)\right)$$

For an example above, T-FLAT operator returns temporally flattened concrete blocks as follows:

$$\left(green, \left\{\begin{array}{c}(French, hostage...)\\(binary)\end{array}\right\}, 0.1, [t1,t2)\right)$$

$$(green, (Woody, ageing...), 0.3, [t2,t3))$$

$$\left(green, \left\{\begin{array}{c}(Woody, ageing...)\\(binary)\end{array}\right\}, 0.4, [t3,now)\right)$$

A snapshot of a concrete block at a given time t is a non temporal triplet valid at t.

$$cb^t = (fb^-, \{c^-\}, i^-)$$

Page

A page is a set of concrete blocks. It is built dynamically from concrete blocks for a given URL when needed. Validity of a page is the union of all concrete blocks' validity. A page is defined as:

$$p_{url} = \{cb1, cb2, cb3...\}$$

In our example (Figure 4), we have one page with three versions. This page is built as follows:

$$p_{url} = \left\{\begin{array}{c}(blue,(News,[t1,now)),(0.4,[t1,now))),\\ \left(pink, \left\{\begin{array}{c}(((Obama,gun,laws...),[t1,t2))\\((binary,[t1,t2)))\\((Sarkozy,Obama...),[t2,t3))\\(binary,[t2,t3))\\((Sarkozy,Woody,...),[t3,now))\\(binary,[t3,now))\end{array}\right\}, \left\{\begin{array}{c}(0.1,[t1,t2))\\(0.3,[t2,t3))\\(0.4,[t3,now))\end{array}\right\}\right)\\ \left(green, \left\{\begin{array}{c}(French,hostage...,[t1,t2))\\(binary,[t1,t2))\\(Woody,ageing...,[t2,now))\\(binary,[t3,now))\end{array}\right\}, \left\{\begin{array}{c}(0.1,[t1,t2))\\(0.3,[t2,t3))\\(0.4,[t3,now))\end{array}\right\}\right)\end{array}\right\}$$

A snapshot of a page is a set of concrete block snapshots for a given URL valid for a requested time and defined as follows:

$${p_{url}}^t = \{cb1^t, cb2^t, cb3^t..., cbn^t\}$$

Site

A website is a set of web pages that are addressed relative to a common URL. In our approach, websites, as well as pages, are built dynamically by using regular expressions to find out which pages belong to which websites.

$$s_{regex} = \{p_{url1}, p_{url2}, p_{url3} ... p_{urln}\}$$

A snapshot of a website at t is a set of pages valid at t.

$$s_{regex}{}^{t} = \{p^{t}_{url1}, p^{t}_{url2}, p^{t}_{url3} ..., p^{t}_{urln}\}$$

Links

Link represents the hyperlink in the page. It points to a page from a frame block and defined as follows:

$$l = (label, type, from, to, [lt_s, lt_e))$$

where:

- *label* is a link label as shown here: $< ahref = \text{“}URL'' >$ *linklabel* $< /a >$.
- *type* is used to distinguish global, internal and local hyperlinks. A hypertext link in an HTML document is said to be:
  - interior if the destination document coincides with the source document (ex:href=“#anchorname”)
  - local if the destination and the source documents are different but in the same domain (ex:href=“/news/art.html”)
  - global if the destination and the source documents are located on different servers.
- *from* attribute corresponds to a frame block
- *to* attribute corresponds to an URL.

In our example (Figure 4), links are listed as followed:

(“Mobile”, local, blue, “/news/mobile”, [t1, now))
(“News”, interior, blue , “#”, [t1, now))
(“Why America’s gun laws won’t change”, local, pink, “/news/politics-25698422”, [t1, t2))
(“US-French push for Iran sanctions”, local, pink, “/news/politics-2457913”, [t2, t3))
(“Bruni to star in Woody Allen film”, local, pink, “/news/cinema-18698422”, [t3, now))
(“Two French hostages in Niger”, local, green, “/news/world-4598422”, [t1, t2))
(“Woody Allen on ageing and death”, local, green, “/news/cinema-2659874”, [t2, now))

4.3 Operators

The proposed language consists of classical set operators with their temporal extensions, relational operators, navigation operators, interval-based temporal logic operators, aggregate operators, ranking and grouping operators. Due to the lack of space, only query operators related to web archive issues are detailed in this section. Table 1 lists the suite of operators.

InBlock

InBlock is one of our logical full-text operators like or-select(OR), and-select(AND), not-select(NOT) etc. It finds matches that satisfy full-text selection in the same block. It is used combined with other logical full-text operators.

For example, assume that we are looking for information about Woody Allen’s new movie where Carla Bruni Sarkozy participates (Figure 4). A query like “Sarkozy and Allen” will return version at t2 and version at t3. In fact, information in version at t2 is not relevant. But a query like “Sarkozy InBlock-AND Allen” will return only version at t3 which has more relevant information.

Wayback

It returns all the versions of a page identified by its URL for a given period.

$$WAYBACK(URL, [ts, te))$$

Fixdate

This operator is used at the beginning of the queries to fix the date interval for all queries in a session. For example, if the user wants to work over the data of January 2000, after calling FIXDATE($[2000-01-01, 2000-02-01)$), she can call other operators without specifying the temporal attributes (*e.g* WAYBACK(url)).

Nearest/Recent/Both

These operators are used to deal with incompleteness explained in Section 2. For example, in Figure 4, if it is asked for the version at t where $t1 < t < t2$, we need to make an assumption over the version at t. Three different operators are proposed:

- NEAREST: it returns the nearest time by minimizing $|t - tx|$
- RECENT: it returns the closest time before t. It is the default operator, if the user does not specify another one.
- BOTH: it returns a time interval constructed with the most closest time before t and after t.

For Figure 4, a query WAYBACK(URL,t) will be executed:

- with NEAREST as WAYBACK(URL,t2), assume that $|t - t2| < |t - t1|$
- with RECENT as WAYBACK(URL,t1)
- with BOTH as WAYBACK(URL, t1) ∪ WAYBACK(URL, t2)

Navigation

Operator out_b finds the set of pages pointed in one step from the set of concrete blocks (CB) by following any of the links valid at a given period. Operator *out* uses out_b to find the set of pages reachable in one step from the set of pages (P) at a given period. For that, it finds the set of concrete blocks foreach page in P and calls out_b foreach *CB*.

Operator *in* finds the set of concrete blocks that points to a set of page by links valid for a requested period.

Operators $jump^{+}$ and $jump^{-}$ return a set of pages reachable, respectively, incoming and outgoing direction in one to n steps by following links valid at a given period. It is a combination of *in* and *out* operators with iteration.

Collapse/Expand COLLAPSE, also referred in literature as coalesce, combines the tuples, which have the same non-temporal values and consecutive or overlapping validities, into one tuple with validity that is the union of the constituent validities.

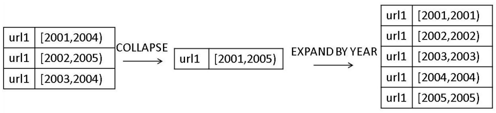


Figure 5: Collapse/Expand

EXPAND expands a tuple into several tuples by splitting its validity into consecutive validities in a given scale. In our approach this scale can be following keywords: YEAR, MONTH, DAY. These operators can be combined with *IN period* to limit the range. In Figure 5, EXPAND BY YEAR IN [2001,2004) returns the first three tuples.

5. USE CASES

In this section, to illustrate how the different operators work in our approach, we use two examples and construct the corresponding queries.

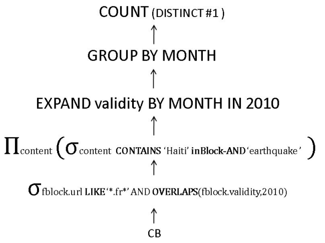

COUNT (DISTINCT #1)
GROUP BY MONTH
EXPAND validity BY MONTH IN 2010
Π content (σ content CONTAINS 'Haiti' inBlock-AND 'earthquake')
σ fblock.url LIKE '*.fr*' AND OVERLAPS(fblock.validity,2010)
CB

Figure 6: Query Example 1

Example 1: We extend the example that we gave in the introduction. A social researcher who studies how French media covered the event "earthquake at Haiti in 2010" over last year wants to know the number per month of *different regions* in web pages in domain .fr referring to earthquake by eliminating duplicates. In that case, if at t, there were two different articles in "lemonde.fr", it is counted as 2 instead of 1 (whole page). Figure 6 shows the query graph for this example.

First, we need to find all concrete blocks in domain .fr which are valid at 2010. LIKE is used to make a string comparison. Then, with CONTAINS operator, we find contents which mention given keywords (Haiti, earthquake). EXPAND operator is used to group by month over validity of content. COUNT and DISTINCT operators are used to find out the number of different regions. By using GROUP BY operator with url, we can limit the count to web pages.

Example 2: Our second example is based on finding broken links from a given url X in a given period (2000 in our example). Figure 7 illustrates the query graph for this example. We need to underline the fact that these broken links are not the result of incompleteness but HTTP 404 error while crawling.

6. CONCLUSION AND FUTURE WORKS

In this paper, we addressed the problem of accessing information in web archives. We presented a conceptual model as the basis of a query language for web archives. The operators to support queries are also described. In our model, we take into account different topics in web pages by using visual blocks as an unit of retrieval with accorded importance. Block-based approach is used for information retrieval on the web, however, as far as we know, it is never used with temporal dimension. Navigation operators with temporal dimension let users to execute queries over web archives temporal hyperlink structure. The model and operators enriched with the temporal dimension allow querying web archives powerfully.

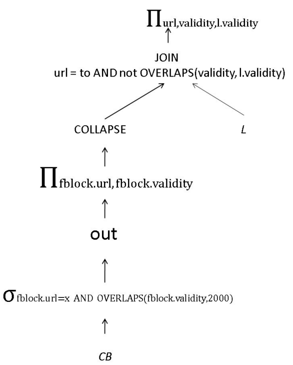

Π url,validity,l.validity
JOIN
url = to AND not OVERLAPS(validity, l.validity)
COLLAPSE
L
Π fblock.url, fblock.validity
out
σ fblock.url=x AND OVERLAPS(fblock.validity,2000)
CB

Figure 7: Query Example 2

Our approach is in the early stage of development. Our first priority is to express the language in algebraic form. Next steps will be the implementation with an appropriate user-friendly syntax. We want to underline the fact that in this paper we clarify the requirements of WAC query language formally. It can be implemented as a new query language or as an extension of an existing query language. We will also work on ranking functions which take into account the block-based structure and temporal dimension. By using the existing temporal indexing [9] and block-based indexing approaches [11], we intend to propose a hybrid indexing model. Once the proposed query language will be fully implemented, our attention will focus on query optimization strategies.

7. REFERENCES

[1] Html protocol, http://www.w3.org/protocols/rfc2616/rfc2616-sec14.html.

[2] Iipc, international internet preservation consortium, http://netpreserve.org/about/index.php.

[3] Internet archive, http://www.archive.org/index.php.

[4] Pandora, australia's web archive, http://pandora.nla.gov.au/.

[5] Uk web archive, http://pandora.nla.gov.au/.

[6] Use cases for access to internet archives. Use case report, IIPC, 2006.

[7] J. F. Allen. Maintaining knowledge about temporal intervals. *Commun. ACM*, 26:832–843, Nov. 1983.

[8] O. Alonso, M. Gertz, and R. Baeza-Yates. On the value of temporal information in information retrieval. *SIGIR Forum*, 41:35–41, Dec. 2007.

[9] K. Berberich, S. Bedathur, T. Neumann, and G. Weikum. A time machine for text search. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 519–526. ACM Press, 2007.

[10] S. S. Bhowmick, S. K. Madria, and W. K. Ng. *Web Data Management.* Springer, 1 edition, Nov. 2003.

[11] E. Bruno, N. Faessel, H. Glotin, J. L. Maitre, and M. Scholl. Indexing by permeability in block structured

OPERATOR	DESCRIPTION
Select	corresponds to well-known select operator in the relational algebra. It returns a subset of data which satisfies given predicates. Temporal select operation is the same as the non-temporal selection but it is extended with additional predicates for temporal comparison
Project	corresponds to its well-known non-temporal version in the relational algebra
Coherent	applies an estimation function over the archive and finds more coherent version for a given page
Rank	applies a ranking function over a bag
Cartesian Product	computes the Cartesian product of two bags
Temporal Cartesian Product	returns a temporal Cartesian product of two bags
Union	returns the union of two bags
Temporal Union	returns the union of temporal elements where their validities overlap
Intersection	returns the intersection of two bags
Temporal Intersection	returns the intersection of temporal elements where their validity overlap
Difference	returns all the elements of the first bag which are not in the second bag
Temporal Difference	in addition to non-temporal Difference, it checks and removes overlapping temporal parts from bags
Join	performs an inner-join on two non-temporal bags based on predicates
Temporal Join	performs an inner-join on two bags based on predicates
Distinct	removes duplicates for non temporal bags
Temporal Distinct	in addition to non-temporal Distinct, it removes duplicates by taking into account overlapping temporal parts
GroupBy	creates groups of tuples sharing some attribute value
Aggregate	COUNT, MIN, MAX are supported
Forward / Backward	for a given element returns the version after/before
T-FLAT	eliminates the temporal nesting in a bag
Logical Full-Text Operators	OR, AND, MILD-NOT, NOT, ORDERED, IN-BLOCK are supported
Contains	realizes keyword search over index servers
OrderBy	sorts a set into a specific order
Temporal Operators	Allen's thirteen temporal operators [7] are supported
Diff	finds differences between two versions of a web page

Table 1: List of Operators

web pages. In *Proceedings of the 9th ACM symposium on Document engineering*, DocEng '09, pages 70–73, New York, NY, USA, 2009. ACM.

[12] E. Bruno, N. Faessel, J. L. Maitre, and M. Scholl. Blockweb: An ir model for block structured web pages. In *Content-Based Multimedia Indexing, 2009. CBMI '09. Seventh International Workshop on*, pages 219–224, 2009.

[13] D. Cai, X. He, J. Wen, and W. Ma. Block-level link analysis. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '04, pages 440–447, New York, NY, USA, 2004. ACM.

[14] D. Cai, S. Yu, J. Wen, and W. Ma. VIPS: a vision-based page segmentation algorithm. Technical Report MSR-TR-2003-79, Microsoft Research, 2003.

[15] D. Cai, S. Yu, J. Wen, and W. Ma. Block-based web search. In *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 456–463. ACM Press, 2004.

[16] Y. Cao, Z. Niu, L. Dai, and Y. Zhao. Extraction of informative blocks from web pages. In *Proceedings of the 2008 International Conference on Advanced Language Processing and Web Information Technology*, pages 544–549, Washington, DC, USA, 2008. IEEE Computer Society.

[17] D. Florescu, A. Levy, and A. Mendelzon. Database techniques for the World-Wide web: A survey. *SIGMOD RECORD*, 27:59–74, 1998.

[18] D. Gomes, A. L. Santos, and M. J. Silva. Managing duplicates in a web archive. In *Proceedings of the 2006 ACM symposium on Applied computing*, SAC '06, pages 818–825, New York, NY, USA, 2006. ACM.

[19] T. Hallgrimsson and S. Bang. Nordic web archive, 2004.

[20] M. Kovacevic, M. Diligenti, M. Gori, M. Maggini, and V. Milutinovic. Recognition of common areas in a web page using visual information: a possible application in a page classification. In *in the proceedings of 2002 IEEE International Conference on Data Mining (ICDM'02*, page 250. IEEE Computer Society, 2002.

[21] J. Masanés. *Web Archiving.* Springer Berlin Heidelberg, 2006.

[22] Z. Pehlivan, M. Ben-Saad, and S. Gançarski. Vi-DIFF: understanding web pages changes. In *Database and Expert Systems Applications*, volume 6261 of *Lecture Notes in Computer Science*, pages 1–15. Springer Berlin / Heidelberg, 2010.

[23] S. Raghavan and H. Garcia-Molina. Complex queries over web repositories. In *Proceedings of the 29th international conference on Very large data bases - Volume 29*, VLDB '2003, pages 33–44. VLDB Endowment, 2003.

[24] M. Ras and S. van Bussel. WEB ARCHIVING user survey. Technical report, National Library of the Netherlands (Koninklijke Bibliotheek), Netherlands, 2007.

[25] R. Song. Learning block importance models for web pages. In *In Intl. World Wide Web Conf. WWW*, pages 203–211. ACM Press, 2004.

[26] M. Stack. Full text search of web archive collections. In *In IWAW*, 2006.

[27] B. Tofel. Wayback for accessing web archives. In *IWAW'07*, Vancouver, British Colombia, Canada, 2007.

[28] Y. Zhang, K. Tanaka, J. X. Yu, S. Wang, M. Li, S. Li, S. Huang, G. Xue, and Y. Yu. Block-Based language modeling approach towards web search. In *Web Technologies Research and Development - APWeb 2005*, volume 3399 of *Lecture Notes in Computer Science*, pages 170–182. Springer Berlin / Heidelberg, 2005.