Changing Vision for Access to Web Archives (WAC)

Zeynep PEHLIVAN       Anne DOUCET      Stéphane GANçARSKI
Introduction

- WACs

- Access to WACs
  - Full-Text Search (with temporal dimension)
  - Navigation
  - Wayback Machine

- Web Users ≠ WAC Users
Data & Queries

Keyword search
Complex query operations (information synthesis)
Ranked results!

TEMPORAL DIMENSION
Outline

- Features of Query Language for WAC (WACQL)
- Data Model
  - Modelling Time
  - Modelling WAC
- Operators
- Related Works
- Conclusion
Features of WACQL

- Eliminates duplicates (Distinct)
- Enables temporal ranking and grouping
- Is user-friendly
- Enables block-based search
- Takes into account incompleteness and temporal coherence
Block-Based Search

- Diversity of web page content
  - Segmentation Algorithms

- Relevance

- Noisy Information
  - Advertisement, navigation bars, decoration stuffs, interaction forms, copyrights, and contact information

- Popular IR method for web

- Giving structure to pages
Importance

- **Block Importance**
  - The block on the center of a page is more important than one on the header

- **Other Parameters**
  - Page rank, page depth in the site etc.

[Song 2004]
Incompleteness

- Containing all the versions of all pages
- Impossible to have a complete archive due to the large number of pages to crawl and the limitations of resources
Temporal Coherence

- A collection is considered coherent, if it reflects the real state of the collection, at least, at one point in time.
Data Model

- Modelling Time
- Modelling WAC
Modelling Time

- Different time in WACs
  - Last-modified time, Date, Expires
  - Content time
  - Crawl time

- Allen's interval-based representation of temporal data
  \([t_{start}, t_{end})\)
Elements

- Frame Block
- Content
- Importance
- Links
- Concrete Block

* Each element has temporal and non-temporal definition
Frame Block

Keeps properties of a block:
- Url to which it belongs
- Dewey identifier that indicates its place in the block hierarchy
- Validity interval

\[ fb = (URL, DeweyID, [fb_{t_1}, fb_{t_2})] \]
Content

- Non-textual (images, videos etc.)
  - Represented by « binary »
  - More than one allowed for each block

- Textual
  - Represented by bag of words
  - Only one for each block
(((Woody, Allen, ageing...), [t2, now]))
(binary, [t3, now])
}
green

(((Sarkozy, Carla, Woddy, Allen...), [t3, now]))
(binary, [t3, now])
}
pink
Importance

- Calculated according to the importance model of [Song 2004]
- Depends on block’s location, area size, content, etc.

\[ i = (\alpha, [t_s, t_e]) \]

- \( (0.4, [t_1, \text{now}]) \) blue
- \( (0.2, [t_1, t_2]) \) pink
- \( (0.1, [t_1, t_2]) \) green
- \( (0.6, [t_2, \text{now}]) \) pink
- \( (0.3, [t_2, t_3]) \) green
- \( (0.4, [t_3, \text{now}]) \) green
A region in a web page

\[ cb = (f_b, \{ c \}, \{ i \}) \]

\[
\left( \text{green}, \begin{cases} (\text{French, hostage,..., } [t_1, t_2]) \\ (\text{binary, } [t_1, t_2]) \\ (\text{Woody, ageing,..., } [t_2, \text{now}]) \\ (\text{binary, } [t_3, \text{now}]) \end{cases} \right), \begin{cases} (0.1, [t_1, t_2]) \\ (0.3, [t_2, t_3]) \\ (0.4, [t_3, \text{now}]) \end{cases} \right)
\]

\[ cb^t = (f_b^-, \{ c^- \}, i^-) \]
Links

- Hyperlinks in web pages
- Type: Global, local, interior

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

- (“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])
Page

- A page is a set of concrete blocks

\[
p_{url} = \begin{cases} 
    \text{blue}, (News, [t1, now]), (0.4, [t1, now])), \\
    \text{pink}, \begin{cases} 
        \text{Obama, gun, laws...}, [t1, t2) \\
        \text{Sarkozy, Obama...}, [t2, t3) \\
        \text{Sarkozy, Woody...}, [t3, now) \\
    \end{cases}, \\
    \text{green}, \begin{cases} 
        \text{French, hostage...}, [t1, t2) \\
        \text{Woody, ageing...}, [t2, now) \\
    \end{cases}
\end{cases}
\]

- A snapshot of a page is a set of snapshot of concrete blocks
- Built dynamically
Site

- A site is a set of web pages

\[ s_{regex} = \{ purl_1, purl_2, purl_3 \ldots purl_n \} \]

- A snapshot of a site is a set of snapshot of pages

- Built dynamically
Operators

- **Time Operators**
  - Allen’s 13 interval operators \( op: period \times period \rightarrow bool \)
  - T-Union, T-Intersect, Minus, Collapse / Expand

- **Unary Operators**
  - Select, Project, Group By

- **Set Operators**
  - Union, Intersect, Difference

- **Aggregate Operators**
  - Sum, Average, Count, Max, Min

- **Ordering Operators**
  - Rank, Order By

- **WAC specific operators**
  - FixDate, Wayback, Coherent

- **Navigation Operators**
Navigation Operators

- **Out, out\_b**: \( CB \times \text{period} \rightarrow CB \)
- **In**: \( CB \times \text{period} \rightarrow CB \)
- **jump\+:** \( CB \times \text{int} \times \text{period} \rightarrow CB \)
- **jump\-**: \( CB \times \text{int} \times \text{period} \rightarrow CB \)
In-Block

- Logical Full-Text Operator
- Example « Woody IN-BLOCK AND Sarkozy »
Coherent (Temporal Coherence)

- Find the most coherent version for navigation
NEAREST/RECENT/BOTH (Incompleteness)

- **NEAREST**: it returns the nearest time by minimizing $|t - t_x|$.
- **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 $t$. 
Related Work (1)

- **Web Archiving**
  - IIPC
  - Wayback Machine (IA) [Tofel 2007]
  - NutchWAX [Stack 2006]

- **Web Based Query Languages**
  - WebBase [Raghavan 2003], WHOWEDA [Bhowmick 2003]

- **Block-Based Search**
  - Block- Based Indexing [Bruno et al. 2009]
  - Block- based IR Model [Li et al. 2004]
  - Block-based link Analysis [Cai et al 2004]
Related Work (2)

- **In conclusion**
  - To access to WACs: Wayback + Full-text search + Navigation
    - No complex queries
    - Does not take into account different topics
  
- **Web Based Query Languages**
  - No temporal dimension
  - Does not take into account different topics

- **Block-based Search**
  - No temporal dimension
  - No complex queries
Conclusion & Future Works

- **WAC Query Language**
  - Visual blocks as an unit of retrieval
  - Temporal dimension
  - Complex queries
  - Ranked keyword queries

- **Future Works**
  - Implementation
  - Temporal Block Based Indexing
  - Temporal Block IR Model