



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 \neq WAC Users

Data & Queries

Queries	Ranked Keyword Search		IR Systems
	Complex & Structured	Database Systems	
		Structured	Unstructured
		DATA	

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 Algorithmes
- ▶ Relevance
- ▶ Noisy Information
 - ▶ Advertisement, navigation bars, decoration stuffs, interaction forms, copyrights, and contact information
- ▶ Popular IR method for web
- ▶ Giving structure to pages

Blocks Example

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

LIVE BBC World Service



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.

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

ADVERTISEMENT

Importance

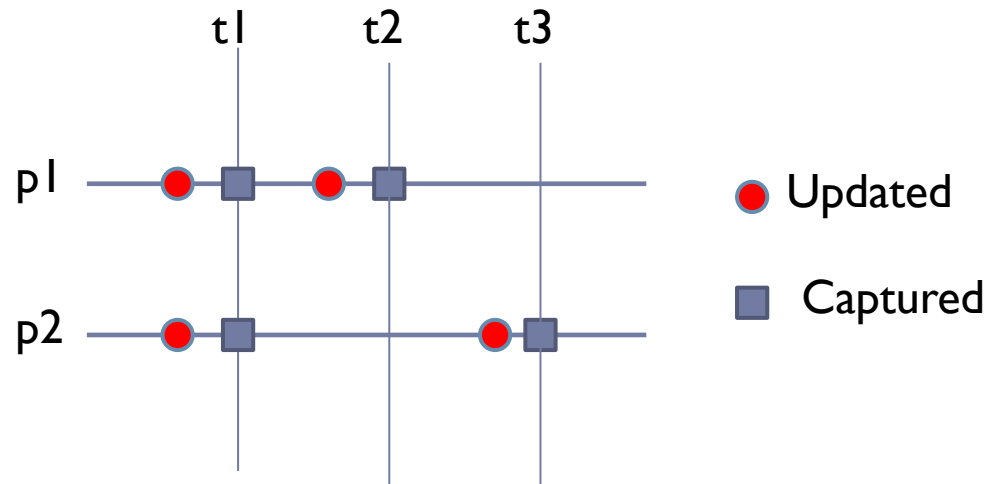
- ▶ **Block Importance**
 - ▶ The block on the center of a page is more important than one on the header
- ▶ **Other Parameters**
 - ▶ Page rank, page depth in the site etc.



[Song 2004]

Incompleteness

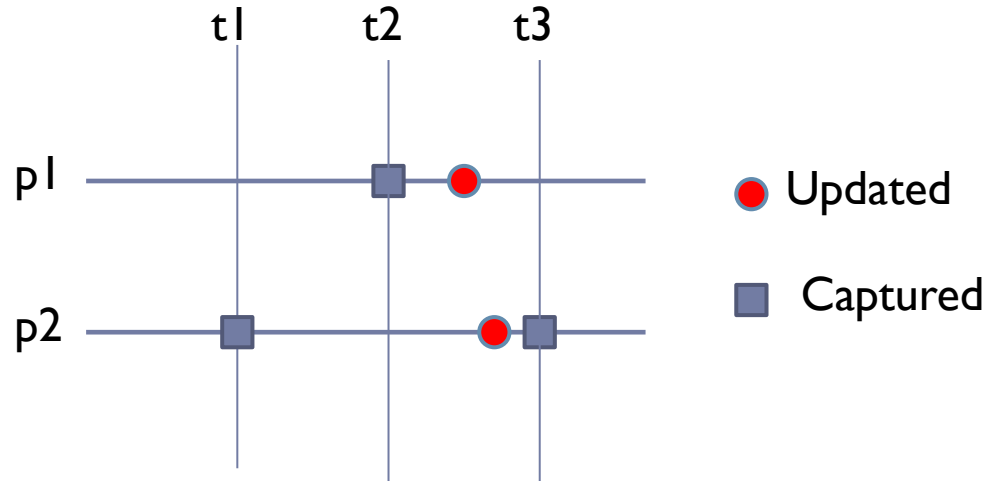
- ▶ Containing all the the versions of all pages



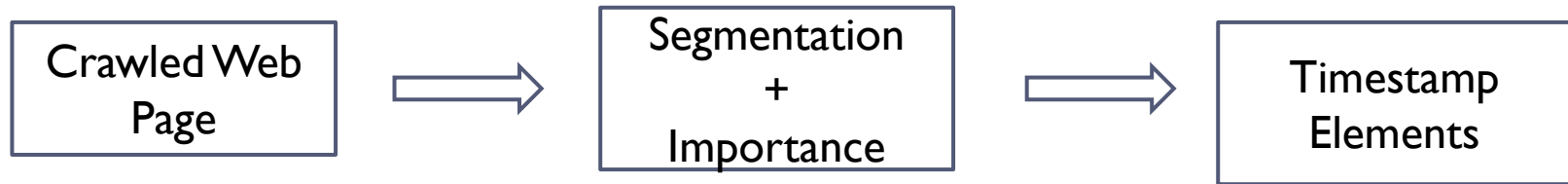
- ▶ Impossible to have a complete archive due to the large number of pages to crawl and the limitations of resources

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

$$[t_{\text{start}}, t_{\text{end}})$$

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

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



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*))

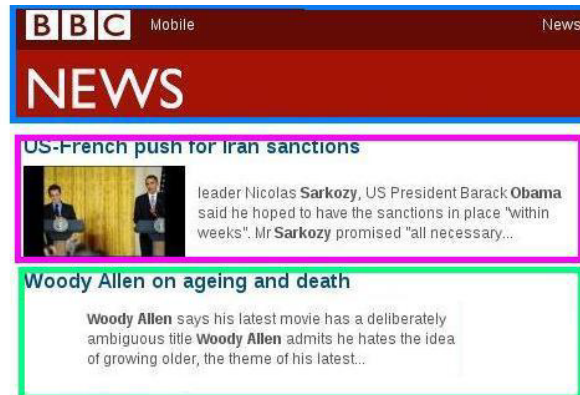
Content

- ▶ **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

Content (2)



t1



t2



t3

$((Woody, Allen, ageing...), [t2, now))$
 $(binary, [t3, now))$ } *green*

$((Sarkozy, Carla, Woody, Allen...), [t3, now))$
 $(binary, [t3, now))$ } *pink*



Importance

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

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



$(0.4, [t1, now))$ blue

$(0.2, [t1, t2))$ pink

$(0.1, [t1, t2))$ green

$(0.6, [t2, now))$ pink

$(0.3, [t2, t3))$ green

$(0.4, [t3, now))$ green

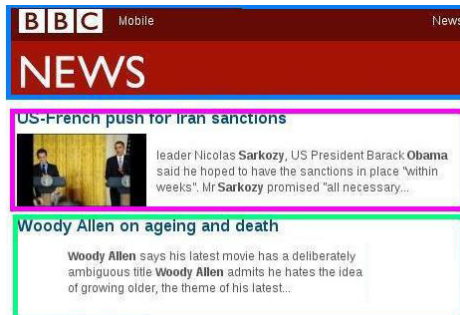
Concrete Block

- A region in a web page

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



t1



t2



t3

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

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

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

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

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

Site

- ▶ A site is a set of web pages

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

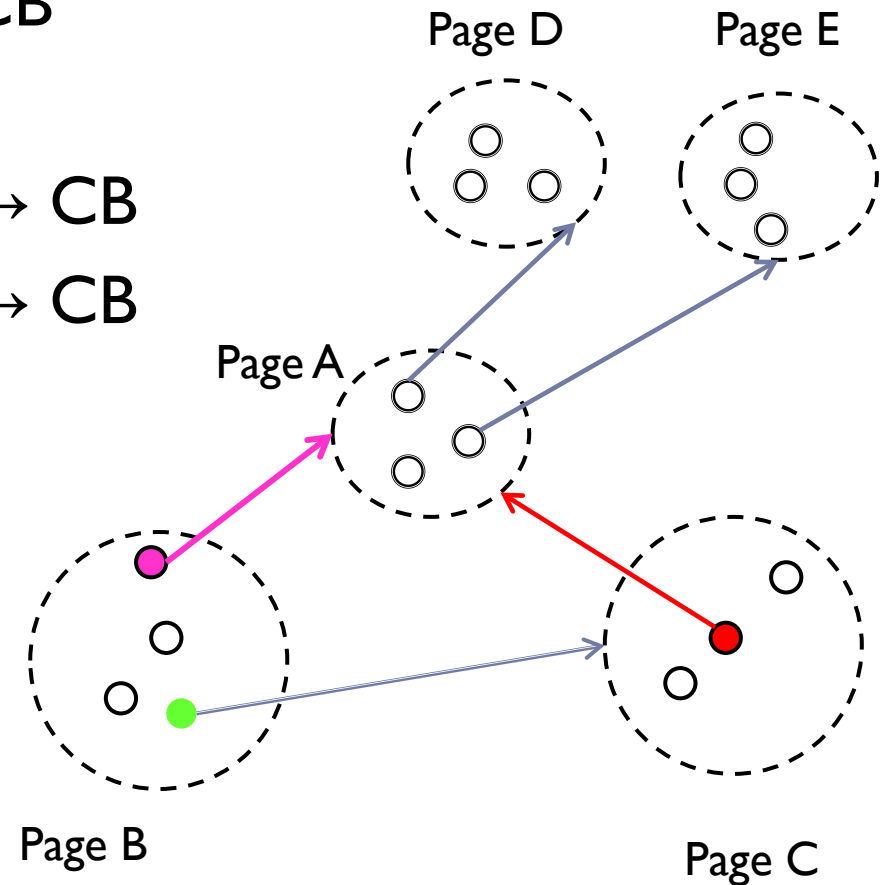
- ▶ A snapshot of a site is a set of snapshot of pages
- ▶ Built dynamically

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

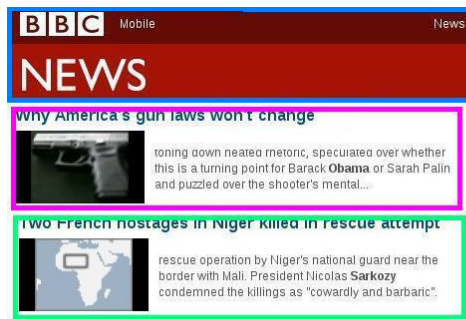
Navigation Operators

- ▶ $\text{Out}, \text{out}_b : \text{CB} \times \text{period} \rightarrow \text{CB}$
- ▶ $\text{In} : \text{CB} \times \text{period} \rightarrow \text{CB}$
- ▶ $\text{jump}^+ : \text{CB} \times \text{int} \times \text{period} \rightarrow \text{CB}$
- ▶ $\text{jump}^- : \text{CB} \times \text{int} \times \text{period} \rightarrow \text{CB}$



In-Block

- ▶ Logical Full-Text Operator
- ▶ Example « Woody IN-BLOCK AND Sarkozy »



t1



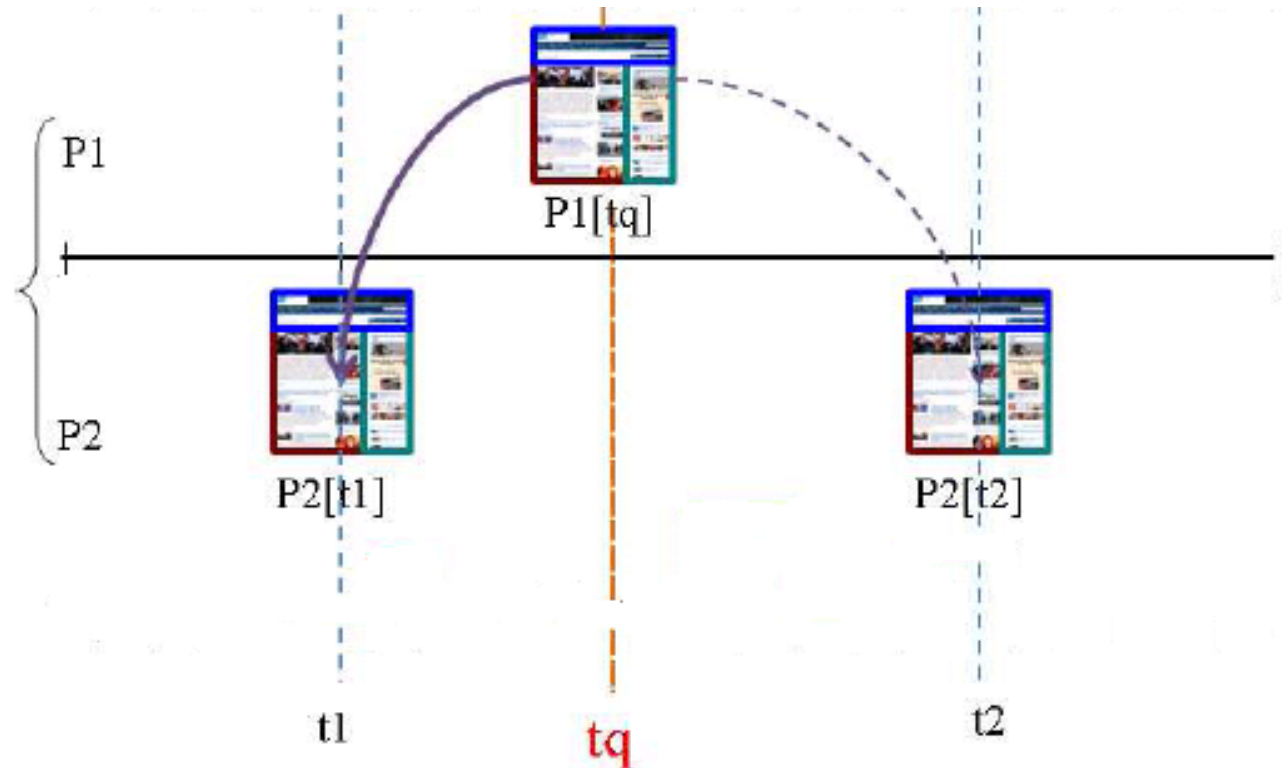
t2



t3

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

- ▶ WebSQL [*Mihaila 1996*], WebOQL [*Arocena 1997*]
- ▶ 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

THANK YOU

