Page Segmentation by Web Content Clustering

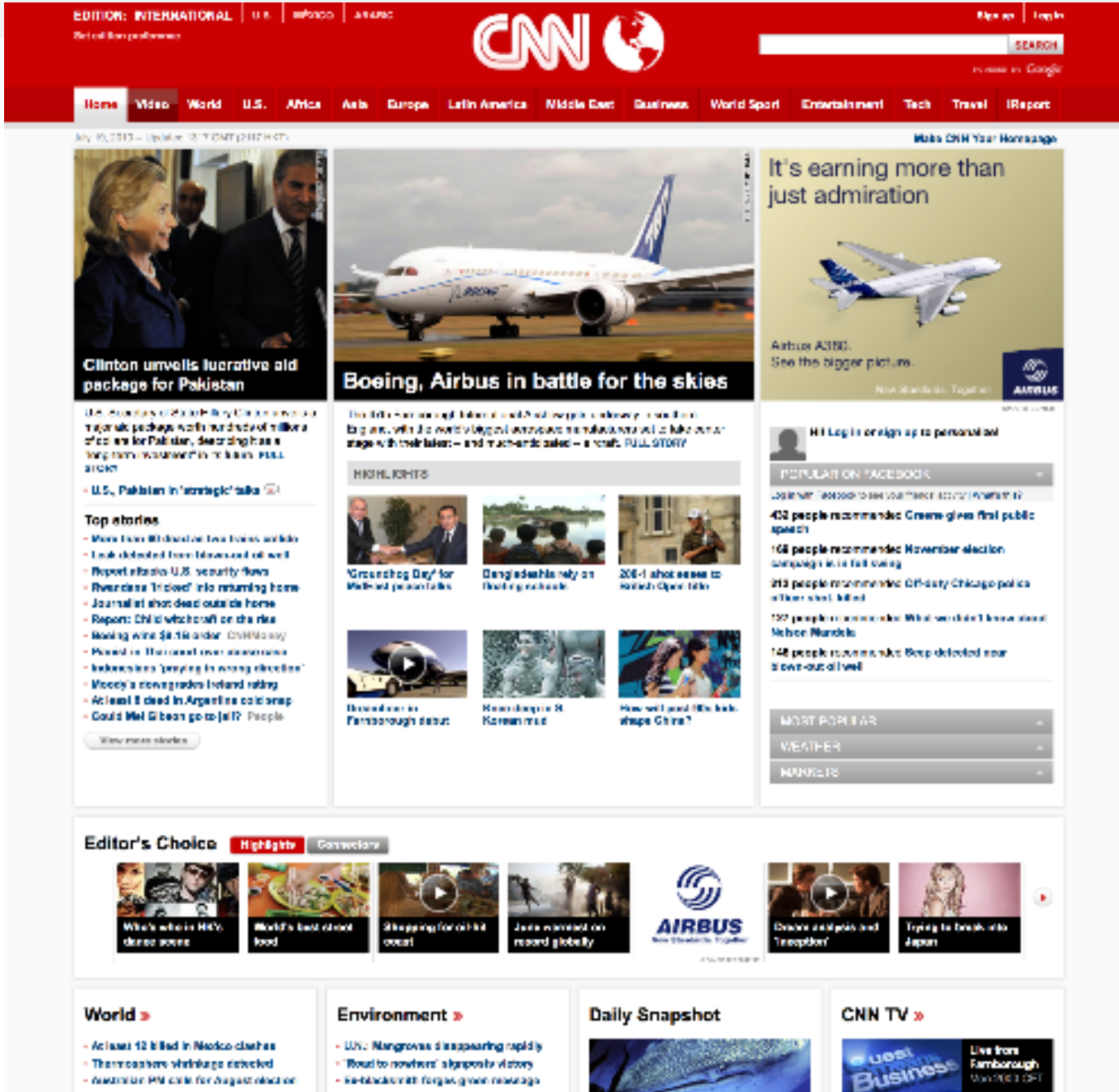
Sadet Alcic

Heinrich-Heine-University of Duesseldorf
Department of Computer Science
Institute for Databases and Information Systems

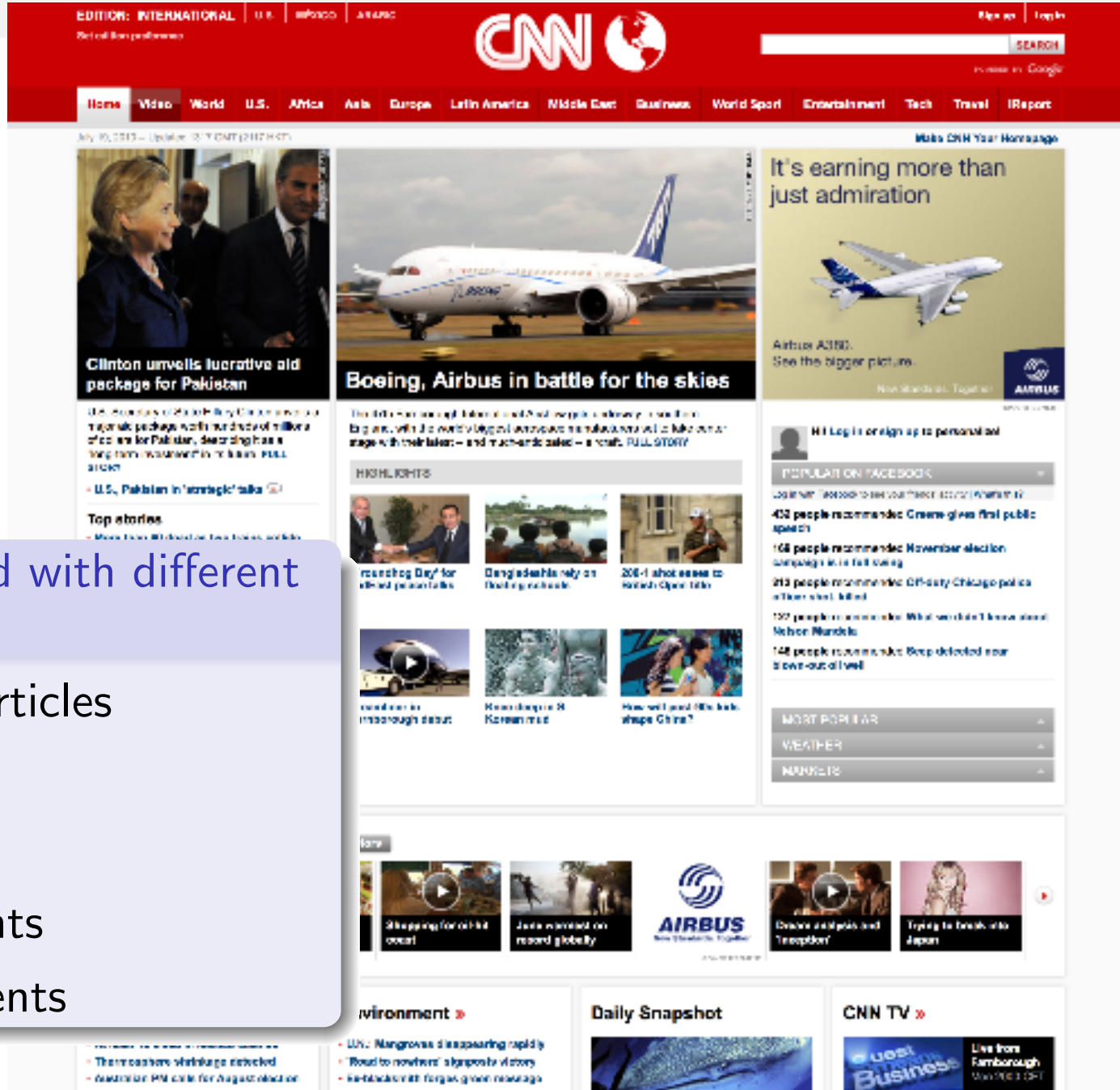
May 26, 2011

Outline

Motivation


EDITION: INTERNATIONAL
CNN
SEARCH
Home
Video
World
U.S.
Africa
Asia
Europe
Latin America
Middle East
Business
World Sport
Entertainment
Tech
Travel
iReport
Clinton unveils lucrative aid package for Pakistan
Boeing, Airbus in battle for the skies
It's earning more than just admiration
Airbus A380.
See the bigger picture.
AIRBUS
Top stories
Editor's Choice
Highlights
Connections
AIRBUS
World »
Environment »
Daily Snapshot
CNN TV »

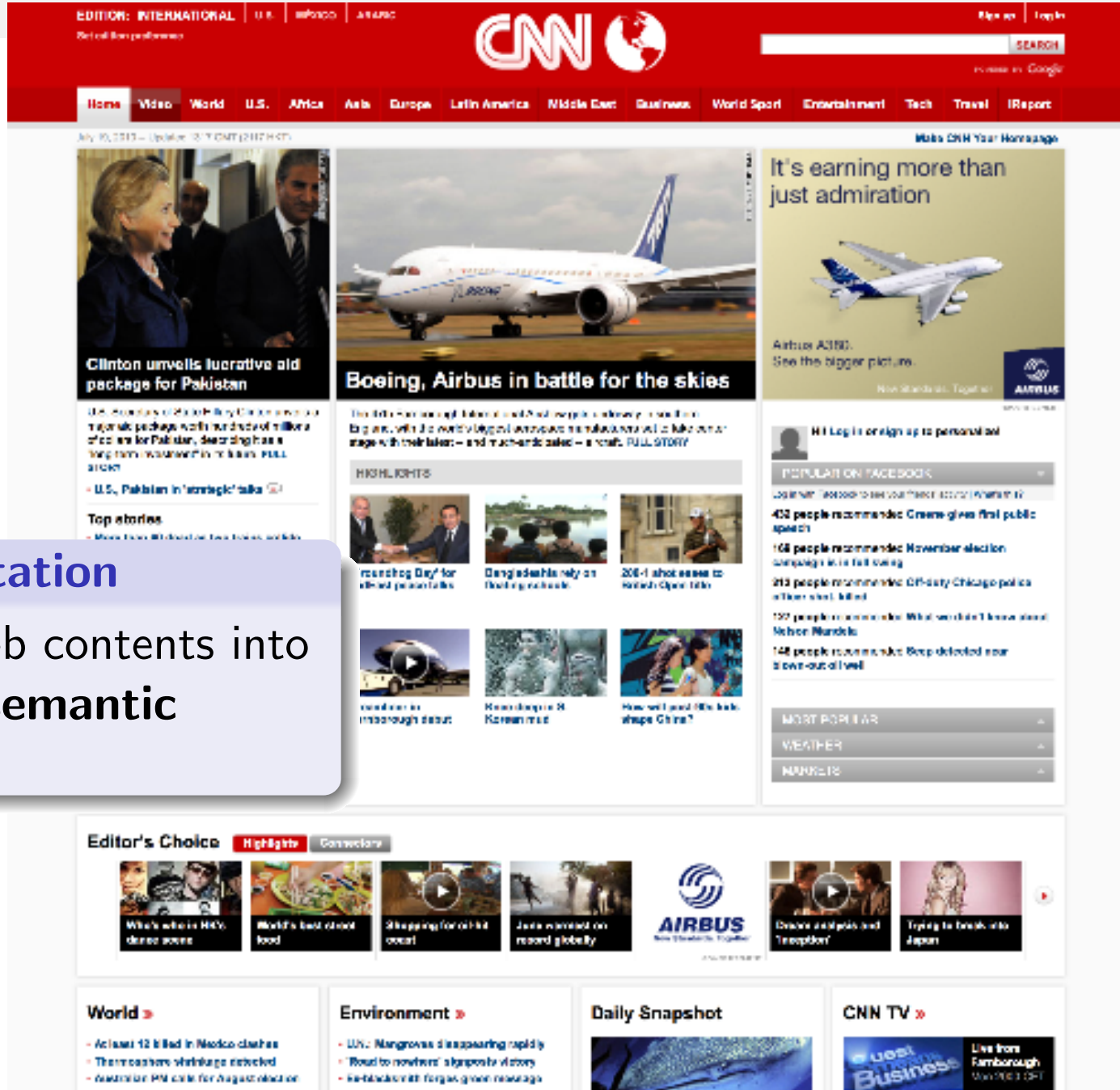
Motivation



Web Page is cluttered with different contents

- Different news articles
- Link lists
- Commercials
- Template elements
- Functional elements

Motivation

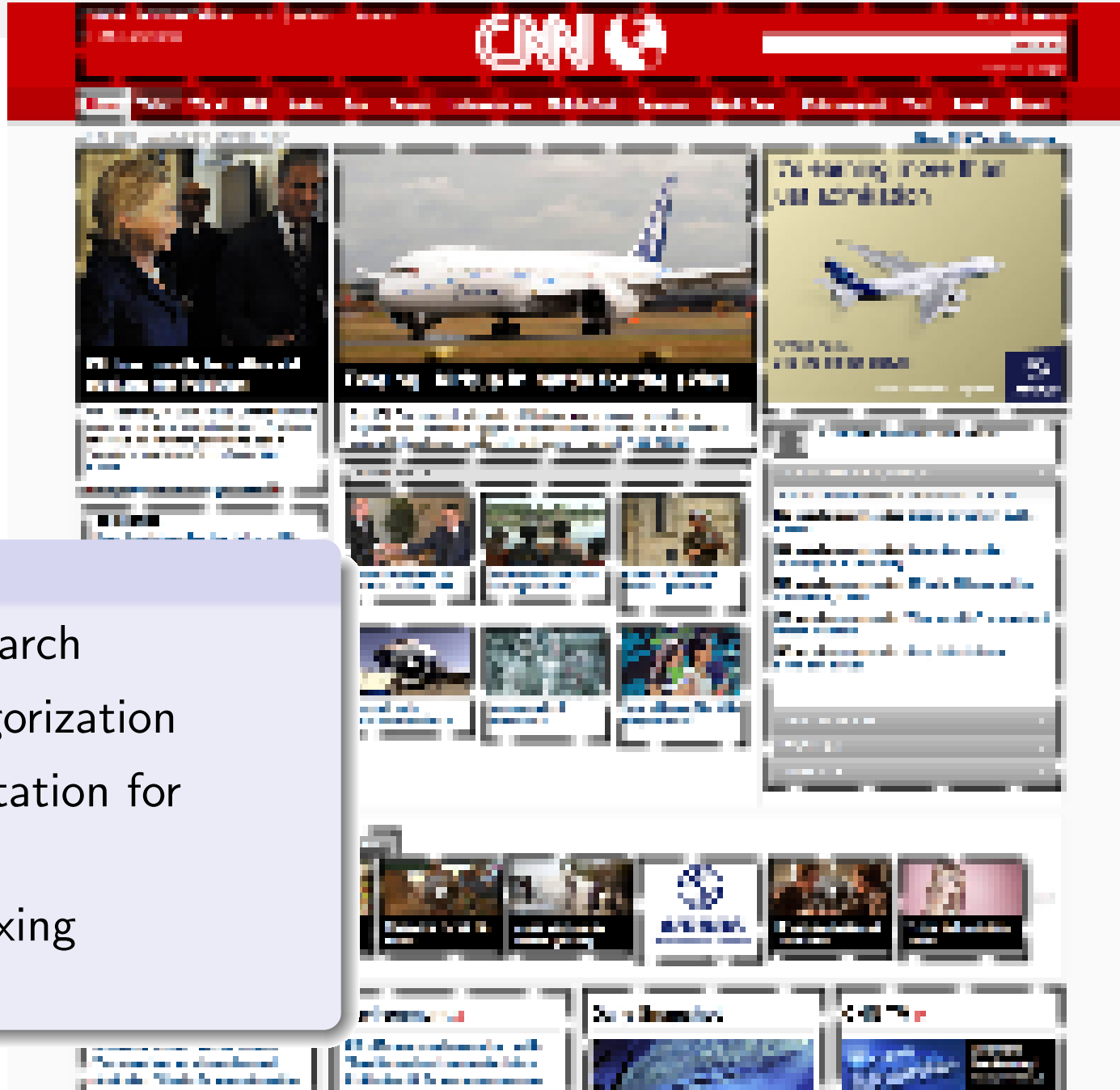


Web Page Segmentation

- Separation of web contents into **structural** and **semantic** cohesive blocks

Motivation

Motivation



Applications

- Web Content Search
- Web Page Categorization
- Web Page Adaptation for Mobile Devices
- Web Image Indexing
- ...

Overview of Related Work to Web Page Segmentation

- **TOP-DOWN page segmentation:**
  - KDD'02: Lin and Ho. *Discovering Informative Content Blocks from Web Documents* (Table properties)
  - APWeb'03: Cai et al. *Extracting content structure for web pages based on visual representation* (Heuristic rules on visual and DOM properties)
  - TKDE'05: Kao et al. *Web Intrapage Informative Structure Mining Based on DOM* (Term entropy based on heuristics)
- **BOTTOM-UP page segmentation:**
  - CIKM'02: Li et al. *Using Micro Information Units for Internet Search* (Heuristic rules)
  - WWW'08: Chakrabarti et al. *A graph-theoretic approach to webpage segmentation* (Graph partitioning)
  - CIKM'08: Kohlschuetter and Nejdl. *A densitometric approach to web page segmentation* (Partitioning of a histogram of text density)

Overview of Related Work to Web Page Segmentation

- **TOP-DOWN page segmentation:**
  - KDD'02: Lin and Ho. *Discovering Informative Content Blocks from Web Documents* (Table properties)
  - APWeb'03: Cai et al. *Extracting content structure for web pages based on visual representation* (Heuristic rules on visual and DOM properties)
  - TKDE'05: Kao et al. *Web Intrapage Informative Structure Mining Based on DOM* (Term entropy based on heuristics)

Basic Idea

- Start with complete Page as initial block
- Decide for each block:
  - should the block be separated?
  - if yes, where to separate?

! Based on heuristics

Overview of Related Work to Web Page Segmentation

Basic Idea

- Start with smallest content units (e.g., DOM leafs)
- group them to blocks of coherent content
- How?

- **BOTTOM-UP page segmentation:**
  - CIKM'02: Li et al. *Using Micro Information Units for Internet Search* (Heuristic rules)
  - WWW'08: Chakrabarti et al. *A graph-theoretic approach to webpage segmentation* (Graph partitioning)
  - CIKM'08: Kohlschuetter and Nejdl. *A densitometric approach to web page segmentation* (Partitioning of a histogram of text density)

Overview of Related Work to Web Page Segmentation

Our Approach

- belongs to BOTTOM-UP methods
- DOM leafs are used as basic web objects
- Idea!: group web objects to blocks by **clustering**

- **BOTTOM-UP page segmentation:**
  - CIKM'02: Li et al. *Using Micro Information Units for Internet Search* (Heuristic rules)
  - WWW'08: Chakrabarti et al. *A graph-theoretic approach to webpage segmentation* (Graph partitioning)
  - CIKM'08: Kohlschuetter and Nejdl. *A densitometric approach to web page segmentation* (Partitioning of a histogram of text density)

Web Page Segmentation by Clustering

Page Segmentation by Clustering

General Definition: Clustering

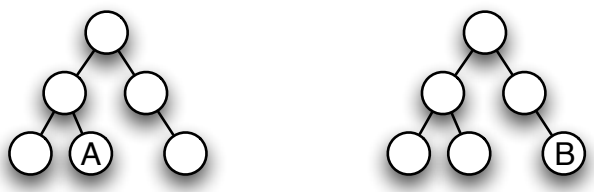
- Clustering is *the process of organizing objects into groups whose members are similar in some way*
- A cluster is therefore a collection of objects which are **similar** between them and are **dissimilar** to the objects belonging to other clusters

Open questions addressed in this work

- How can the similarity (or dissimilarity) of web objects be estimated?
- Which representation is best suitable to represent web objects?
- Which clustering method should be applied for clustering?

Different Representations of Web objects

- Geometric Representation
  - web browser puts every object of a web page in a 2-dim plane
  - extract the **bounding rectangle** for each object
- Semantic Representation
  - elements in DOM contain some textual contents
  - extract **keywords** from the corresponding text
- DOM-based Representation
  - each object is a node in the DOM tree of the page
  - use the **position of the object in DOM tree** to characterize it



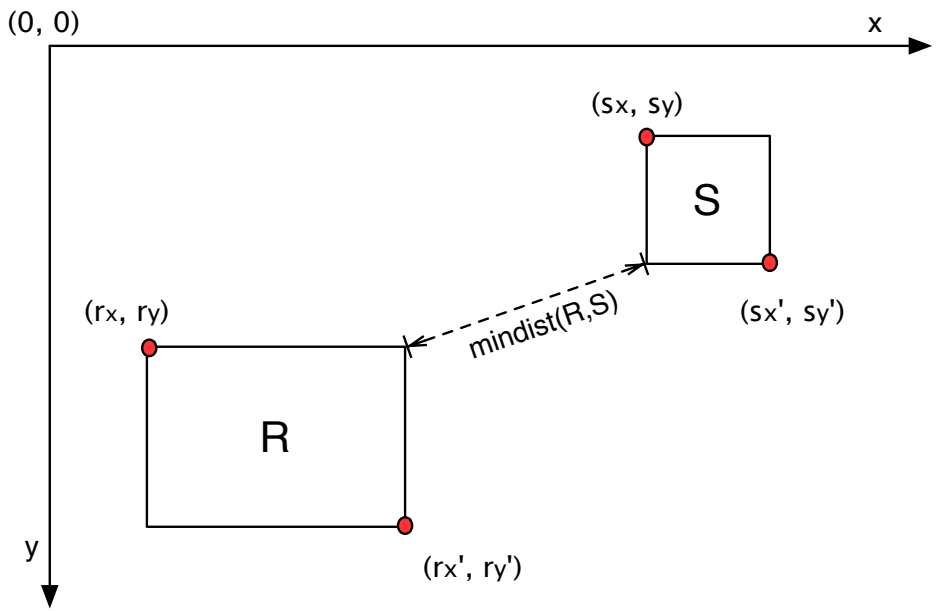
$\Rightarrow$ **Different distance measures are possible**

Geometric Distance

- Let $R = [(r_x, r_y), (r_x', r_y')]$ and $S = [(s_x, s_y), (s_x', r_y')]$ be two bounding rectangles
- The geometric distance of R to S is given by

$$dist(R,S) = \left(\sum_{i \in x,y} {t_i}^2\right)^{\frac{1}{2}}, \text{ with } t_i = \begin{cases} r_i - s_i' & \text{if} \quad r_i > s_i' \\ s_i - r_i' & \text{if} \quad r_i' < s_i \\ 0 & \text{if} \quad \text{otherwise.} \end{cases}$$

- Visually:

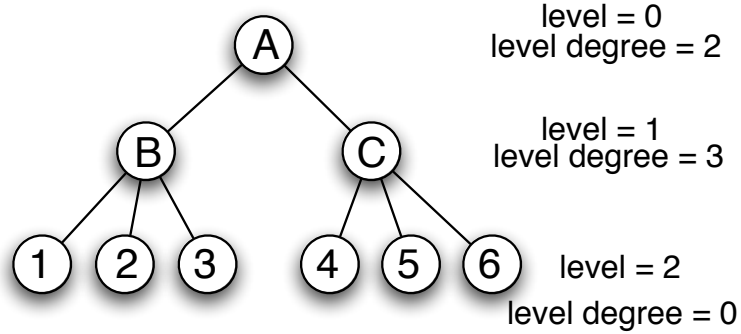


Semantic Distance

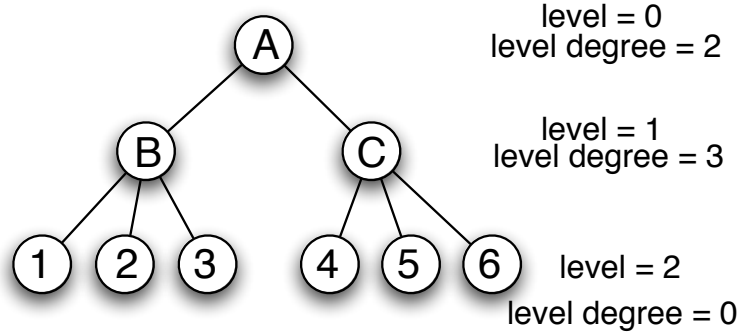
- Given $T_1 = (\text{dog, run, street})$, $T_2 = (\text{puppy, walk, road})$
- Cosine Similarity Measure (Information Retrieval)
  - Lexical word-to-word matching $\rightarrow sim(T_1, T_2) = 0$
- to strict: e.g. synonym and hyponym relationships are not considered
- Instead: text similarity measure based on WordNet [Corley 05]
  - Words are mapped to concepts in WordNet (concept-to-concept matching)

$$sim(T_1, T_2) = \frac{\sum_{w_i \in T_1} maxSim(w_i, T_2) \cdot idf(w_i)}{\sum_{w_i \in T_1} idf(w_i)}$$

DOM-based Distance



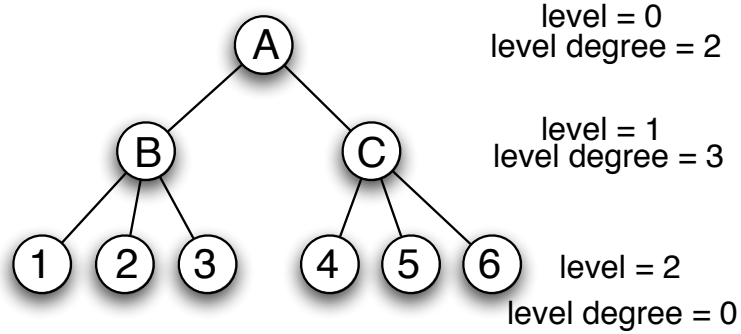
DOM-based Distance



Requirements

- Nodes under same parent are closer than nodes under different parent
- Nodes on higher tree level are closer that nodes on lower level

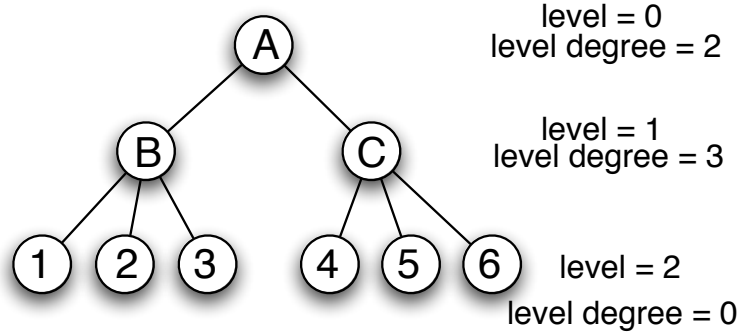
DOM-based Distance



- Traverse DOM-tree in preorder traversing:

$$P = (\text{Ⓐ}, \text{Ⓑ}, ①, ②, ③, \text{Ⓒ}, ④, ⑤, ⑥)$$

DOM-based Distance

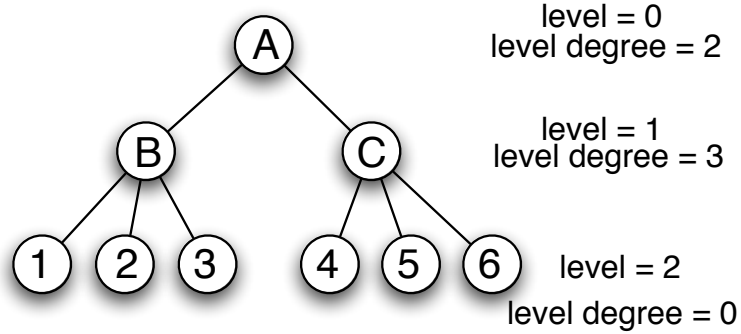


- Traverse DOM-tree in preorder traversing:

$$P = (\text{Ⓐ}, \text{Ⓑ}, ①, ②, ③, \text{Ⓒ}, ④, ⑤, ⑥)$$

- For each element in P define a weight w_{p_i} that expresses the costs needed to reach p_i from its predecessor in P

DOM-based Distance

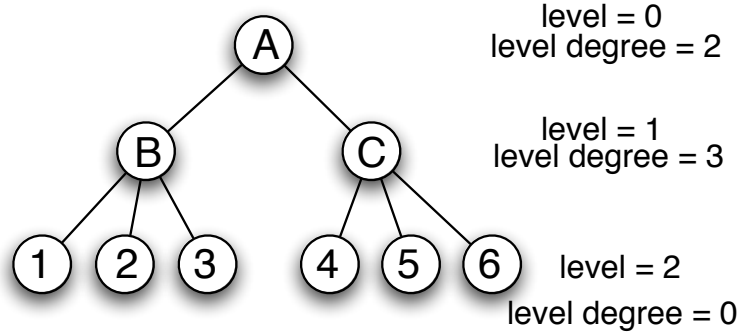


- The distance between $p_a, p_b \in P$, (wlog. $a < b$) is defined as:

$$d(p_a, p_b) = \sum_{i=a+1}^{b} w_{p_i}$$

Example: $d(②, ④) = w_3 + w_C + w_4$

DOM-based Distance



- The weight w_i of a node $p_i \in P$ depends on the level l and the level degree d_l of p_i:

$$w(l) = \begin{cases} c & : \quad d_l = 0 \\ d_l \cdot w(l+1) & : \quad d_l > 0, \end{cases} \tag{1}$$

e.g., $w(2) = c$, $w(1) = 3 * w(2) = 3c$, $w(0) = 2 * w(1) = 6c$,

Clustering Methods

- Partitioning Clustering
  - **k-medoid** (similar as k-means, but cluster representatives are real objects)
- Agglomerative Hierarchical Clustering
  - **single link** method applied to compute distance between sets of objects
- Density-based Clustering
  - **DBSCAN** variant (able to find clusters of different density levels)

Evaluation Studies

Distance-Matrix Visualization

- A distance matrix contains all pairwise distances of the objects to be clustered, e.g.

	a	b	c
a	0	1.9	1.1
b	1.9	0	2.3
c	1.1	2.3	0

WIMS'11

International Conference on Web Intelligence, Mining and Semantics
MAY 25 - 27, 2011

57-65

- Home
- Organisation
- Call for Papers
- Submission
- Registration
- Program
- Important Dates
- Accommodation
- Venue



66-68

ORGANISED BY 44-54

WESTERN NORWAY RESEARCH INSTITUTE
VESTLANDSFORSKING

Proceedings will be published by

Association for Computing Machinery

Selected papers from WIMS'11, after further revisions, will be published in the special issues of the following journals.

International Journal of Metadata, Semantics and Ontologies

International Journal of Web Services Practices

International Journal of Information Retrieval Research

International Journal of Computer Science & Applications

Some selected extended papers from WIMS'11 will be considered for **Elsevier (Morgan Kaufmann)** book [pending approval].

55-56

Welcome

29-43

The International Conference on Web Intelligence, Mining and Semantics (WIMS'11) will be organised under the auspices of Western Norway Research Institute.

This is the first in a new series of conferences concerned with intelligent approaches to transform the World Wide Web into a global reasoning and semantics-driven computing machine. Next conferences in this series, WIMS'12 and WIMS'13, will take place in Craiova (Romania) and Madrid (Spain) respectively.

The conference will provide an excellent international forum for sharing knowledge and results in theory, methodology and applications of Web intelligence, Web mining and Web semantics. The program will feature several keynote and invited talks, from academia and the industry.

The purpose of the WIMS'11 is:

- To provide a forum for established researchers and practitioners to present past and current research contributing to the state of the art of Web technology research and applications.
- To give doctoral students an opportunity to present their research to a friendly and knowledgeable audience and receive valuable feedback.
- To provide an informal social event where Web technology researchers and practitioners can meet.

Scientific American article by Tim Berners-Lee, Ora Lassila and James Hendler was published in May 2001. The WIMS'11 conference is an appropriate event to reflect on the 10 years - successes and misses >>

NEW Call for Breaking News Abstracts

WIMS'11 Flyer

Important Dates

1-19

Electronic submission of papers/posters (Extended):
~~**November 20, 2010**~~

Tutorial and Workshop proposals due:
December 15, 2010 NEW

Notification of paper acceptance:
January 15, 2011

Registration opens:
February 1, 2011

Camera-ready of accepted papers:
February 15, 2011

Breaking news - Abstract submission:
March 1, 2011 NEW

Registration closes:
May 10, 2011

Conference:
May 25 - 27, 2011

CONTACT: 20-28

Vestlandsforsking,
PO Box 163,
NO-6851 SOGNDAL, NORWAY

Phone: +47 916 85 607
Fax: +47 947 63 727
E-mail: wims11 @ vestforsk.no

SPONSOR WIMS'11!

WIMS'11 International Conference on Web Intelligence, Mining and Semantics
MAY 25 - 27, 2011

57-65

Home
Organisation
Call for Papers
Submission
Registration
Program
Important Dates
Accommodation
Venue



66-68

ORGANISED BY 44-54

WESTERN NORWAY RESEARCH INSTITUTE
VESTLANDSFORSKING

Proceedings will be published by

Association for Computing Machinery

Selected papers from WIMS'11, after further revisions, will be published in the special issues of the following journals.

International Journal of Metadata, Semantics and Ontologies

International Journal of Web Services Practices

International Journal of Information Retrieval Research

International Journal of Computer Science & Applications

Some selected extended papers from WIMS'11 will be considered

Welcome 29-43

The International Conference on Web Intelligence, Mining and Semantics (WIMS'11) will be organised under the auspices of Western Norway Research Institute.

This is the first in a new series of conferences concerned with intelligent approaches to transform the World Wide Web into a global reasoning and semantics-driven computing machine. Next conferences in this series, WIMS'12 and WIMS'13, will take place in Craiova (Romania) and Madrid (Spain) respectively.

The conference will provide an excellent international forum for sharing knowledge and results in theory, methodology and applications of Web intelligence, Web mining and Web semantics. The program will feature several keynote and invited talks, from academia and the industry.

The purpose of the WIMS'11 is:

- To provide a forum for established researchers and practitioners to present past and current research contributing to the state of the art of Web technology research and applications.
- To give doctoral students an opportunity to present their research to a friendly and knowledgeable audience and receive valuable feedback.
- To provide an informal social event where Web technology

Important Dates 1-19

Electronic submission of papers/posters (Extended):
~~**November 20, 2010**~~

Tutorial and Workshop proposals due:
December 15, 2010 NEW

Notification of paper acceptance:
January 15, 2011

Registration opens:
February 1, 2011

Camera-ready of accepted papers:
February 15, 2011

Breaking news - Abstract submission:
March 1, 2011 NEW

Registration closes:
May 10, 2011

Conference:
May 25 - 27, 2011

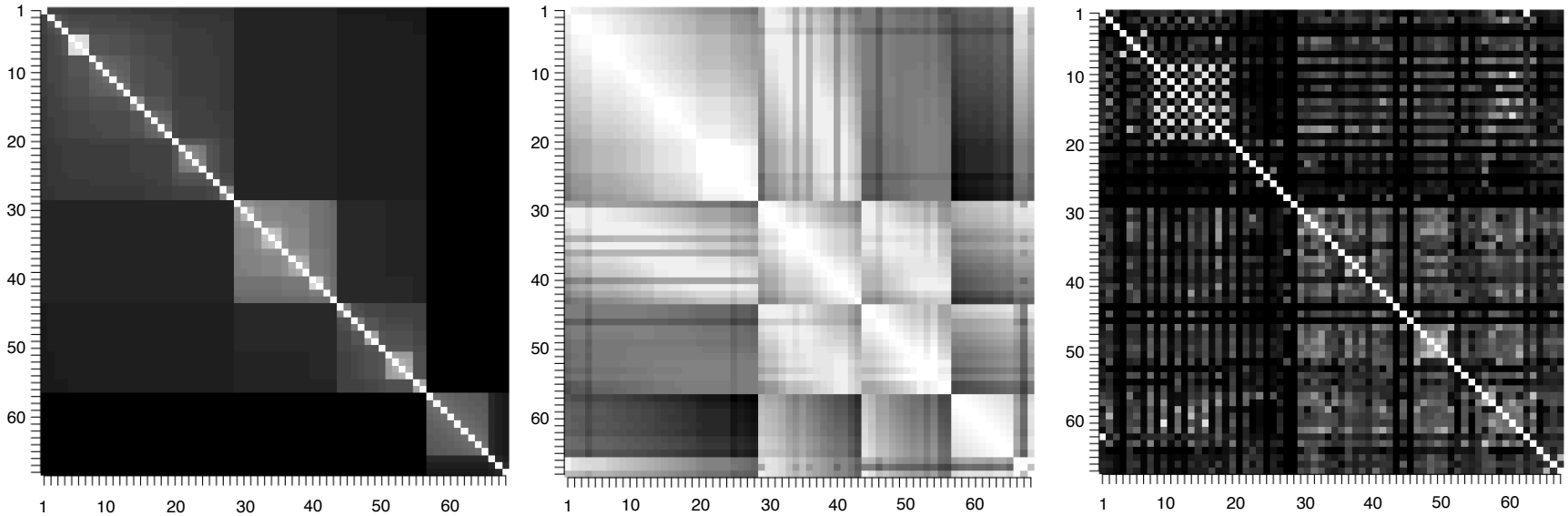
20-28

ng,
NDAL, NORWAY
3 85 607
3 727
@ vestforsk.no
WIMS'11!

Numbers

- indicate the position of particular web objects in the source code

Distance-Matrix Visualization



- left and bottom axe represent the web objects ordered by their appearance on the web page
- each pixel represents a distance value
- white means lowest distance, black means highest distance
- bright squares in the diagonal indicate possible page blocks

WIMS'11

International Conference on Web Intelligence, Mining and Semantics
MAY 25 - 27, 2011

57-65

- Home
- Organisation
- Call for Papers
- Submission
- Registration
- Program
- Important Dates
- Accommodation
- Venue



66-68

ORGANISED BY 44-54

WESTERN NORWAY RESEARCH INSTITUTE
VESTLANDSFORSKING

Proceedings will be published by

Association for Computing Machinery

Selected papers from WIMS'11, after further revisions, will be published in the special issues of the following journals.

International Journal of Metadata, Semantics and Ontologies

International Journal of Web Services Practices

International Journal of Information Retrieval Research

International Journal of Computer Science & Applications

Some selected extended papers from WIMS'11 will be considered for **Elsevier (Morgan Kaufmann)** book [pending approval].

55-56

Welcome

29-43

The International Conference on Web Intelligence, Mining and Semantics (WIMS'11) will be organised under the auspices of Western Norway Research Institute.

This is the first in a new series of conferences concerned with intelligent approaches to transform the World Wide Web into a global reasoning and semantics-driven computing machine. Next conferences in this series, WIMS'12 and WIMS'13, will take place in Craiova (Romania) and Madrid (Spain) respectively.

The conference will provide an excellent international forum for sharing knowledge and results in theory, methodology and applications of Web intelligence, Web mining and Web semantics. The program will feature several keynote and invited talks, from academia and the industry.

The purpose of the WIMS'11 is:

- To provide a forum for established researchers and practitioners to present past and current research contributing to the state of the art of Web technology research and applications.
- To give doctoral students an opportunity to present their research to a friendly and knowledgeable audience and receive valuable feedback.
- To provide an informal social event where Web technology researchers and practitioners can meet.

Scientific American article by Tim Berners-Lee, Ora Lassila and James Hendler was published in May 2001. The WIMS'11 conference is an appropriate event to reflect on the 10 years - successes and misses >>

NEW Call for Breaking News Abstracts

WIMS'11 Flyer

Important Dates

1-19

Electronic submission of papers/posters (Extended):
~~**November 20, 2010**~~

Tutorial and Workshop proposals due:
December 15, 2010 NEW

Notification of paper acceptance:
January 15, 2011

Registration opens:
February 1, 2011

Camera-ready of accepted papers:
February 15, 2011

Breaking news - Abstract submission:
March 1, 2011 NEW

Registration closes:
May 10, 2011

Conference:
May 25 - 27, 2011

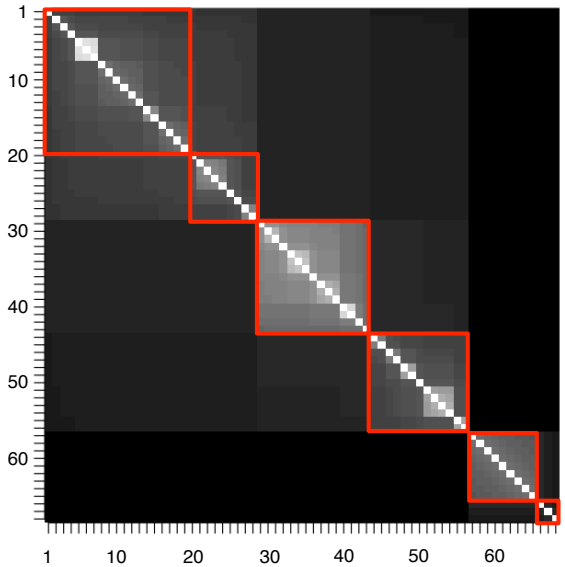
CONTACT: 20-28

Vestlandsforsking,
PO Box 163,
NO-6851 SOGNDAL, NORWAY

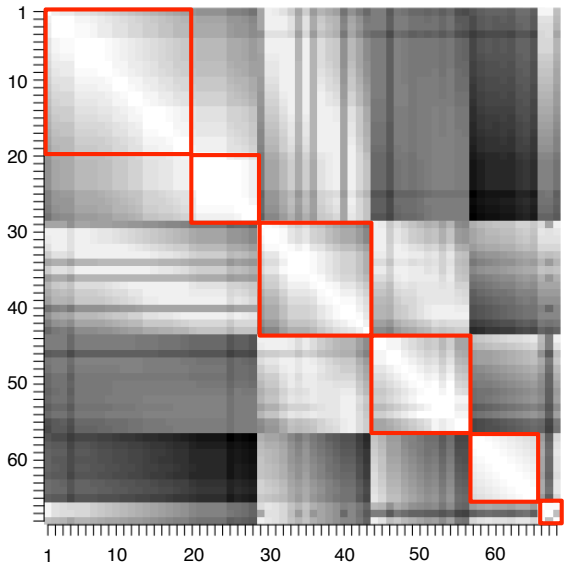
Phone: +47 916 85 607
Fax: +47 947 63 727
E-mail: wims11 @ vestforsk.no

SPONSOR WIMS'11!

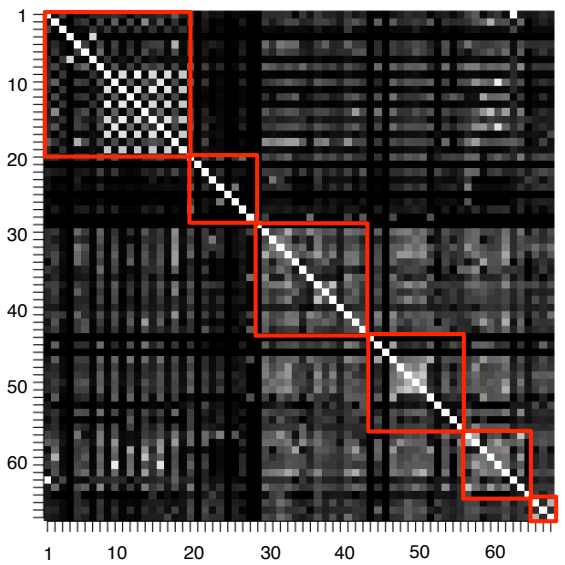
Distance-Matrix Visualization



a) DOM-Distance

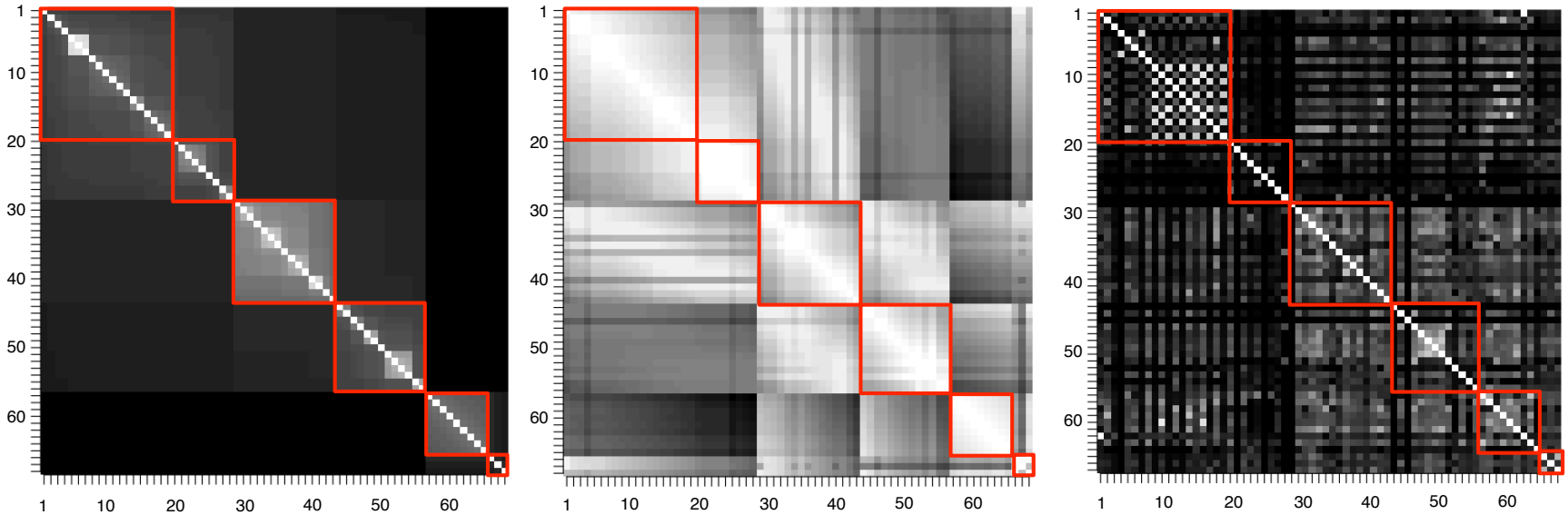


b) Geometric-Distance



c) Semantic-Distance

Distance-Matrix Visualization

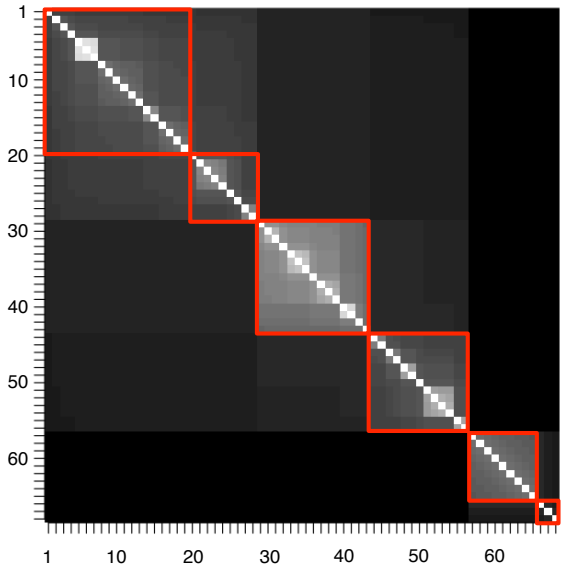


a) DOM-Distance b) Geometric-Distance c) Semantic-Distance

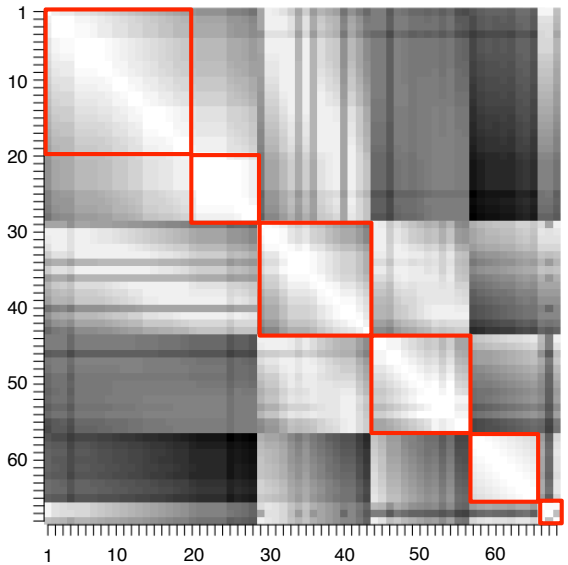
Results:

- DOM-distance has good correspondence

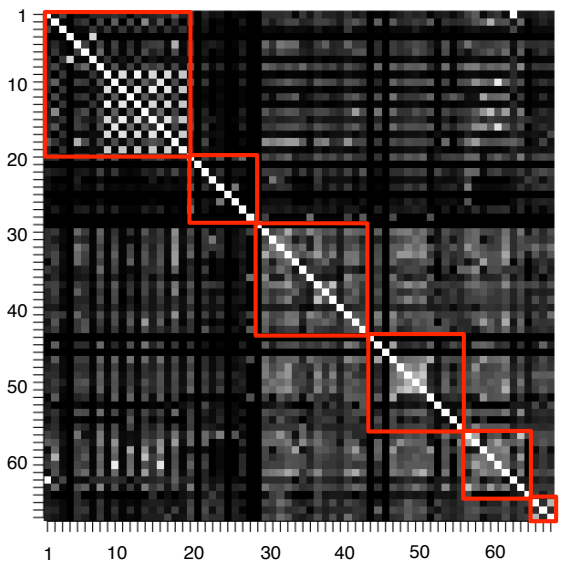
Distance-Matrix Visualization



a) DOM-Distance



b) Geometric-Distance



c) Semantic-Distance

Results:

- DOM-distance has good correspondence
- Geometric distance has some correspondence, but there are other bright rectangles

WIMS'11

International Conference on Web Intelligence, Mining and Semantics
MAY 25 - 27, 2011

57-65

- Home
- Organisation
- Call for Papers
- Submission
- Registration
- Program
- Important Dates
- Accommodation
- Venue



66-68

ORGANISED BY 44-54

WESTERN NORWAY RESEARCH INSTITUTE
VESTLANDSFORSKING

Proceedings will be published by

Association for Computing Machinery

Selected papers from WIMS'11, after further revisions, will be published in the special issues of the following journals.

International Journal of Metadata, Semantics and Ontologies

International Journal of Web Services Practices

International Journal of Information Retrieval Research

International Journal of Computer Science & Applications

Some selected extended papers from WIMS'11 will be considered for **Elsevier (Morgan Kaufmann)** book [pending approval].

55-56

Welcome

29-43

The International Conference on Web Intelligence, Mining and Semantics (WIMS'11) will be organised under the auspices of Western Norway Research Institute.

This is the first in a new series of conferences concerned with intelligent approaches to transform the World Wide Web into a global reasoning and semantics-driven computing machine. Next conferences in this series, WIMS'12 and WIMS'13, will take place in Craiova (Romania) and Madrid (Spain) respectively.

The conference will provide an excellent international forum for sharing knowledge and results in theory, methodology and applications of Web intelligence, Web mining and Web semantics. The program will feature several keynote and invited talks, from academia and the industry.

The purpose of the WIMS'11 is:

- To provide a forum for established researchers and practitioners to present past and current research contributing to the state of the art of Web technology research and applications.
- To give doctoral students an opportunity to present their research to a friendly and knowledgeable audience and receive valuable feedback.
- To provide an informal social event where Web technology researchers and practitioners can meet.

Scientific American article by Tim Berners-Lee, Ora Lassila and James Hendler was published in May 2001. The WIMS'11 conference is an appropriate event to reflect on the 10 years - successes and misses >>

NEW Call for Breaking News Abstracts

WIMS'11 Flyer

Important Dates

1-19

Electronic submission of papers/posters (Extended):
~~**November 20, 2010**~~

Tutorial and Workshop proposals due:
December 15, 2010 NEW

Notification of paper acceptance:
January 15, 2011

Registration opens:
February 1, 2011

Camera-ready of accepted papers:
February 15, 2011

Breaking news - Abstract submission:
March 1, 2011 NEW

Registration closes:
May 10, 2011

Conference:
May 25 - 27, 2011

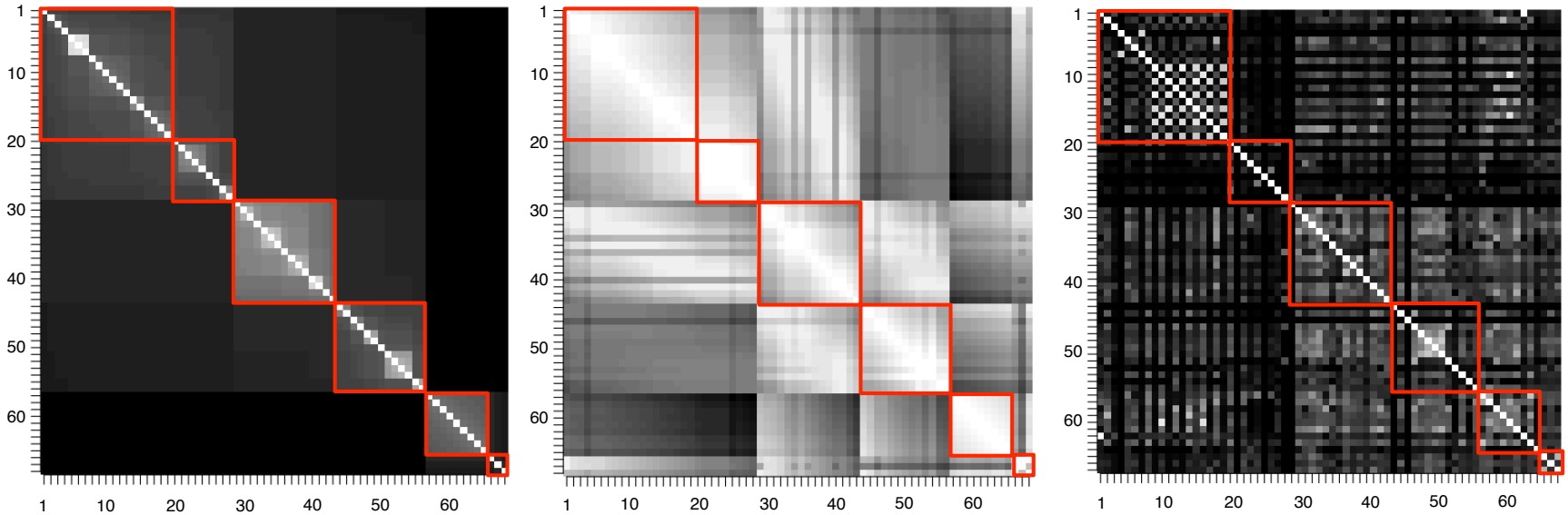
CONTACT: 20-28

Vestlandsforsking,
PO Box 163,
NO-6851 SOGNDAL, NORWAY

Phone: +47 916 85 607
Fax: +47 947 63 727
E-mail: wims11 @ vestforsk.no

SPONSOR WIMS'11!

Distance-Matrix Visualization



a) DOM-Distance b) Geometric-Distance c) Semantic-Distance

Results:

- DOM-distance has good correspondence
- Geometric distance has some correspondence, but there are other bright rectangles
- Semantic distance has almost no correspondence

Evaluation - Clustering Performance

Dataset

- 78 web documents from 8 different categories from Yahoo! directory
- 23,819 web contents (in average 305 per web page)
- Web contents clustered manually by 3 volunteers
- Ground truth is combination of all three proposals

Method

- For each combination of clustering method & distance function
  - compute clustering of the web contents into page blocks
- Ground Truth Clustering (GT), Computed Clustering (C)

Evaluation - Performance Measure

Based on Contingency Table for pairs of objects:

	Same cluster in C	Different cluster in C
Same cluster in GT	f_{11}	f_{10}
Different cluster in GT	f_{01}	f_{00}

Performance Measure

$$Rand\ statistic = \frac{f_{00} + f_{11}}{f_{00} + f_{01} + f_{10} + f_{11}}$$

Average Rand Statistic Results

	DOM-based	geometric	semantic
partitioning	0.45	**0.47**	0.25
hierarchical	**0.52**	0.41	0.24
density-based	**0.61**	0.43	0.27

Results:

- Rand statistic values are similar to the results of Distance Matrix Visualization
- DOM distance reaches highest values with DB clustering
  - the distances between together belonging objects are varying in the metrical space derived by DOM distance
  - DB clustering is able to find clusters of different densities

Conclusion and Future Work

Conclusion

- Web Page Segmentation by Clustering was presented
- three different distance function for web objects based on geometric, semantic and DOM properties
- three clustering methods from different categories: partitioning, hierarchical and density-based clustering
- best clustering accordance to ground truth with DOM-Distance and DB clustering

Future Work

- combination of distance measure (linear, multiplicative, ?)
- comparison to other Web Page Segmentation methods from literature
- application of Web Page Segmentation to Web Image Context Extraction (paper accepted, to be published)

Thank You!