

Representing Genre-Specific Websites

Alexander Mehler and Rüdiger Gleim
Bielefeld University

Agenda

1. Motivation
2. Extracting Websites
3. Logical Hypertext Document Structure
4. Services
5. Conclusion

Motivation

Web Structure Mining

graph topology (Adamic 1999)

- small worlds

Categorizing macro structures
(Amitay et al. 2003):

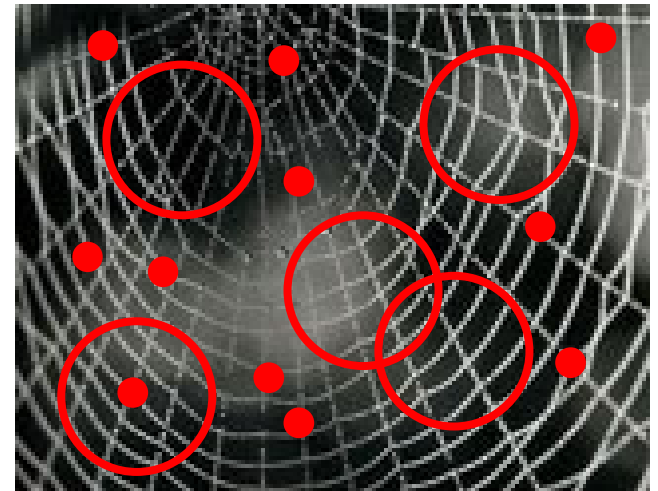
- ... of directories vs. corporate sites vs. online shops ... by means of the link structure of their pages

Categorizing web pages (Fürnkranz 1998):

- HTML-Markup, meta-tags and link structure as additional input

Segmenting web pages (Mizuuchi und Tajima 1999; Rehm 2004):

- identifying logical sections in single pages



Motivation

Observation

- web documents with similar function / content tend to have similar structures
- ➔ these structures are not directly addressed by segmentation and categorization of single web pages
- ➔ this problem relates to any effort of utilizing web mining in order to extract and explore genre specific corpora of web documents

Call for Participation

- [Submission Categories](#)
- [Conference Themes](#)
- ...
- [Conference Chairs](#)

Conference Themes

- [Theme 1](#)
- [Theme 2](#)
- [Theme 3](#)
- ...

Submission Categories

- [Full Papers](#)
- [Short Papers](#)
- [Posters](#)
- ...

Call for Participation

Submission Categories

- [Full Papers](#)
- [Short Papers](#)
- [Posters](#)
- ...

Conference Themes

- [Theme 1](#)
- [Theme 2](#)
- [Theme 3](#)
- ...

Conference Chairs


- ...

The Fifth International Conference on Logical Aspects of Computational Linguistics

Logical Aspects of Computational Linguistics, LACL '05, english - Mozilla

Datei Bearbeiten Ansicht Gehe Lesezeichen Tools Fenster Hilfe

Zurück Vor Neu laden Stopp <http://www.labri.fr/Recherche/LLA/signes/LACL/cfp.htm> Suchen Drucken

 **LaBRI - C.N.R.S.**
Université Bordeaux 1 Université

Program Committee **Schedule**

Program **...**

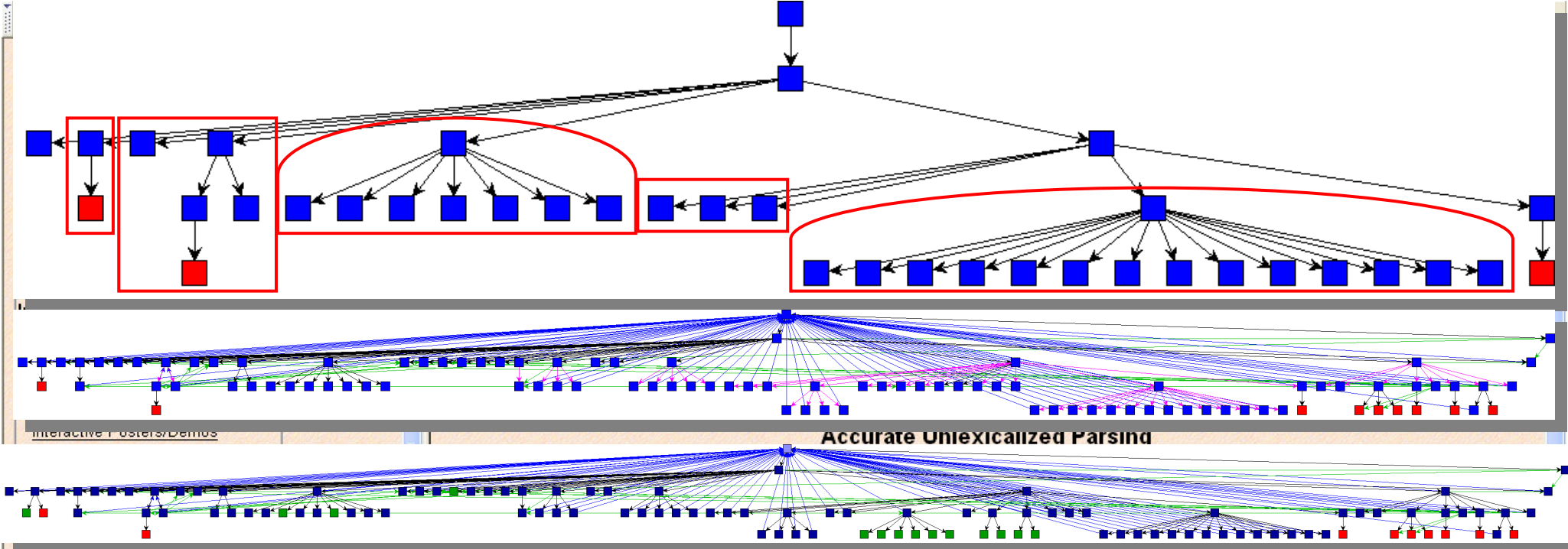
General Information

LACL 2005
Fifth International Conference on
Logical Aspects of Computational Linguistics
28-29-30 April 2005
Bordeaux
<http://lacl.labri.fr/>
Submission deadline: monday 3 january 2005

- [LACL](#)
- [Topics](#)
- [Submissions](#)
- [Dates](#)
- [C...](#)
- [P...](#)
- [Organizing committee](#)

Link nicht gefunden: "+"

41st Annual Meeting of the Association for Computational Linguistics



Life Time Achievement Award
Interactive Posters/Demos
Student Workshop
Sponsors (PDF)
Exhibitors
Conf
Reg
Accommodations
Venues
Equip
Supp
People behind ACL03
Links
ACL
AFNLP
Archives
Previous Announcements
ACL 2003 News Letter No. 7

* [Information on ACL2004](#) *

Spain, 2004
<http://www.acl2004.org>

General Information

Program

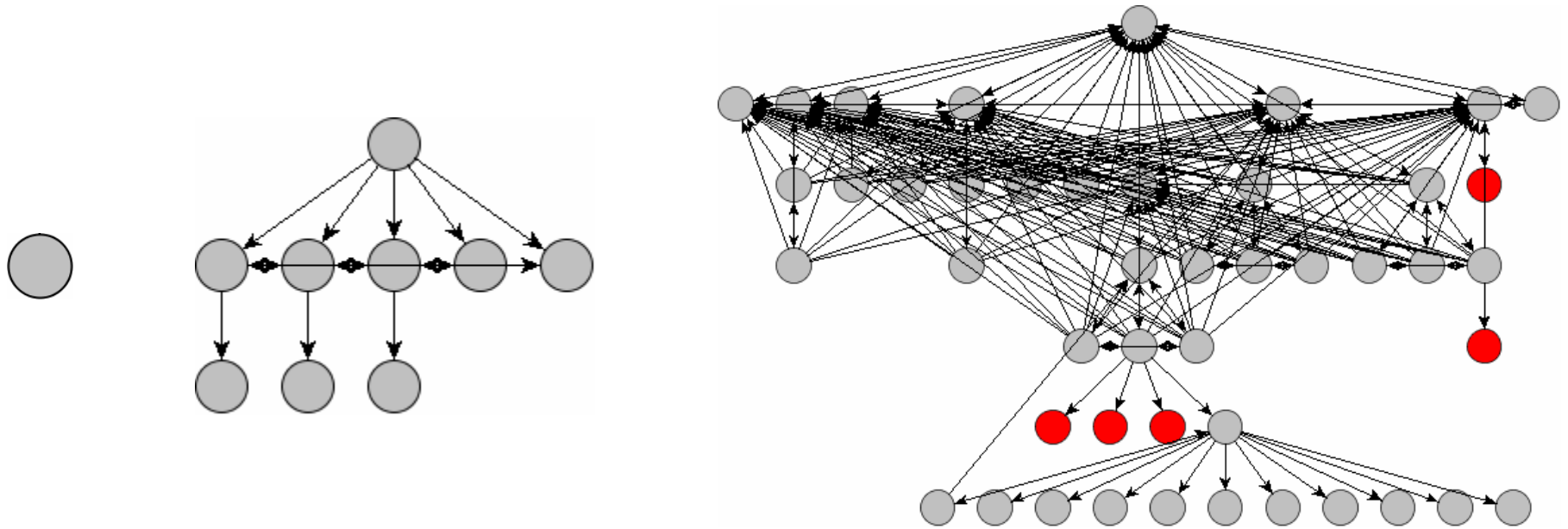
Program Committee

Schedule

...

Motivation

Variety of Tree-like and Graph Structures



Motivation

Polymorphism

- The same expression unit manifests several categories.
- The same category is distributed over several units.
- n:m relation between manifest form and manifested structure
- exploring genre specific web-corpora cannot be performed as function learning, but has to be performed as structure learning

Motivation

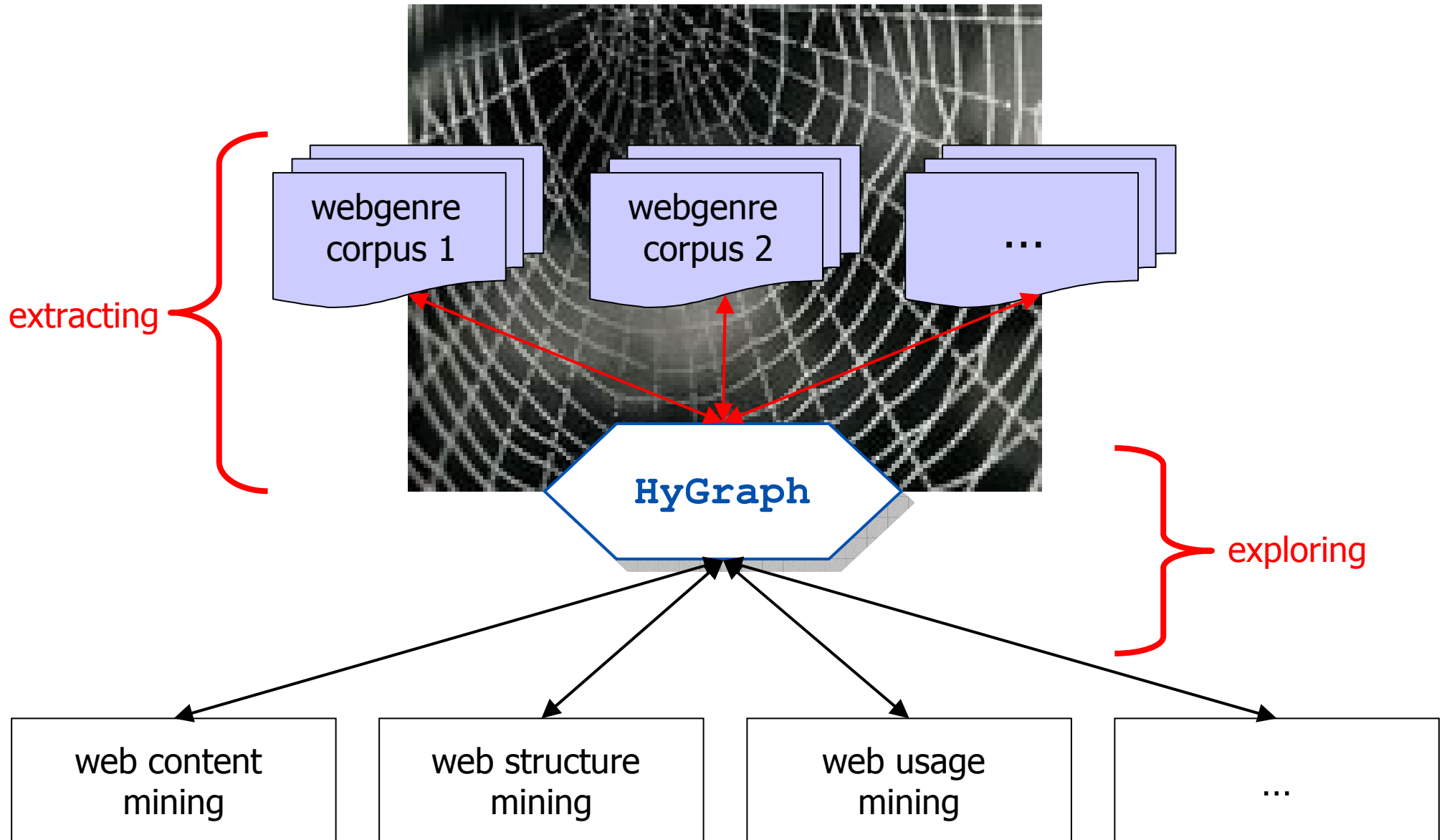
Webgenres

- electronic encyclopaedia
- personal academic homepage
- conference website
- corporate site
- online shop
- ...

Agenda

1. Motivation
2. Extracting Websites
3. Logical Hypertext Document Structure
4. Services
5. Conclusion

Extracting Websites



Extracting Websites

Tasks

- automatic extraction of corpora of websites of certain webgenres
 - generic representation of web documents
 - corpus management and maintenance
 - visualization of document structures
 - classification of hypertext graphs
 - corpus conversion for machine learning tools (**Weka**, **LibSVM**, ...)
- HyGraph as an interface of ML-related approaches to webgenres

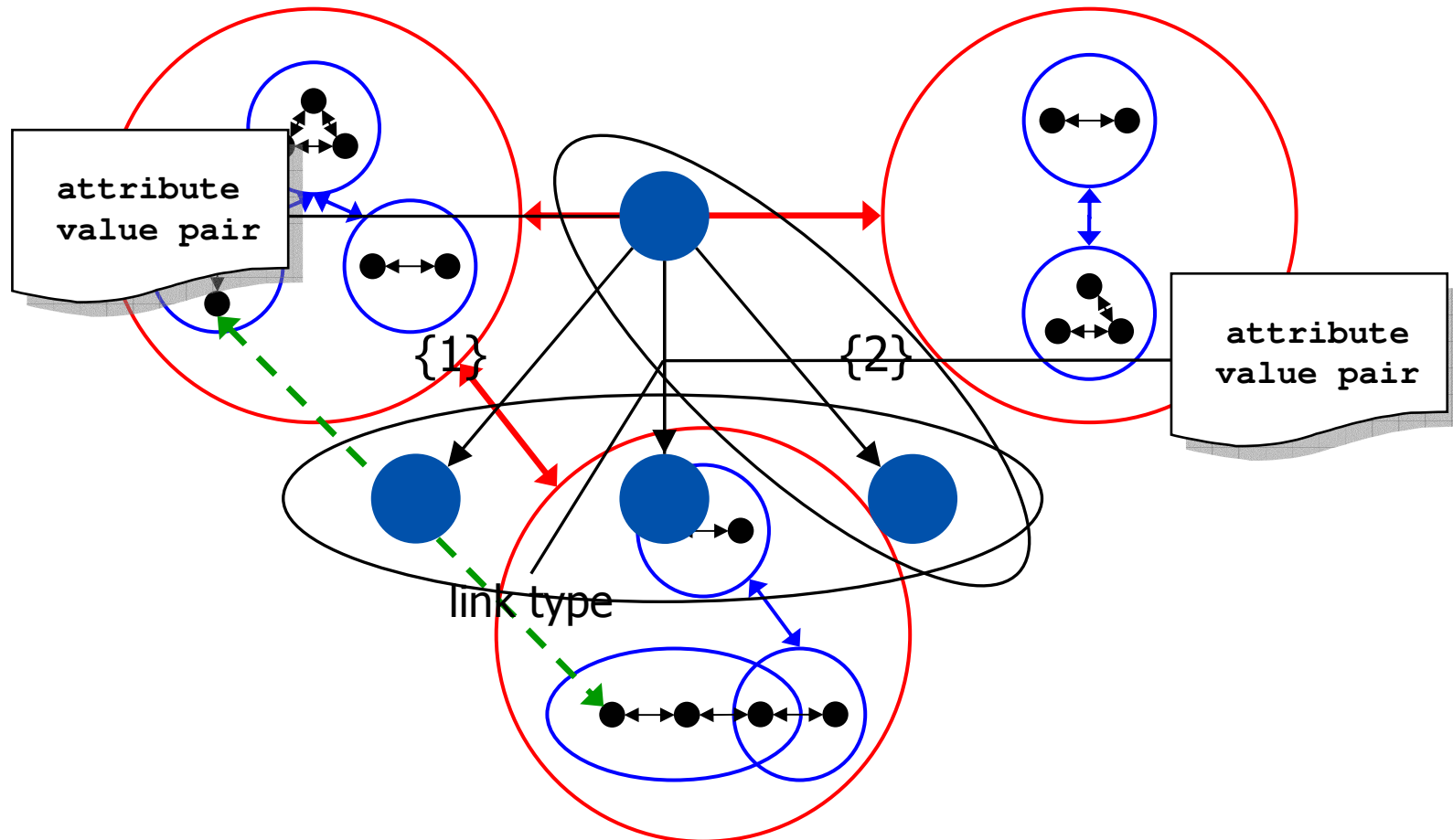
Extracting Websites

Representing Graphs – Requirements

- expressive power for mapping graphs of varying types
 - typing and attributing nodes and edges
 - using flexible, extensible standards of information modelling
- Graph eXchange Language (Winter et al. 2002)

Extracting Websites

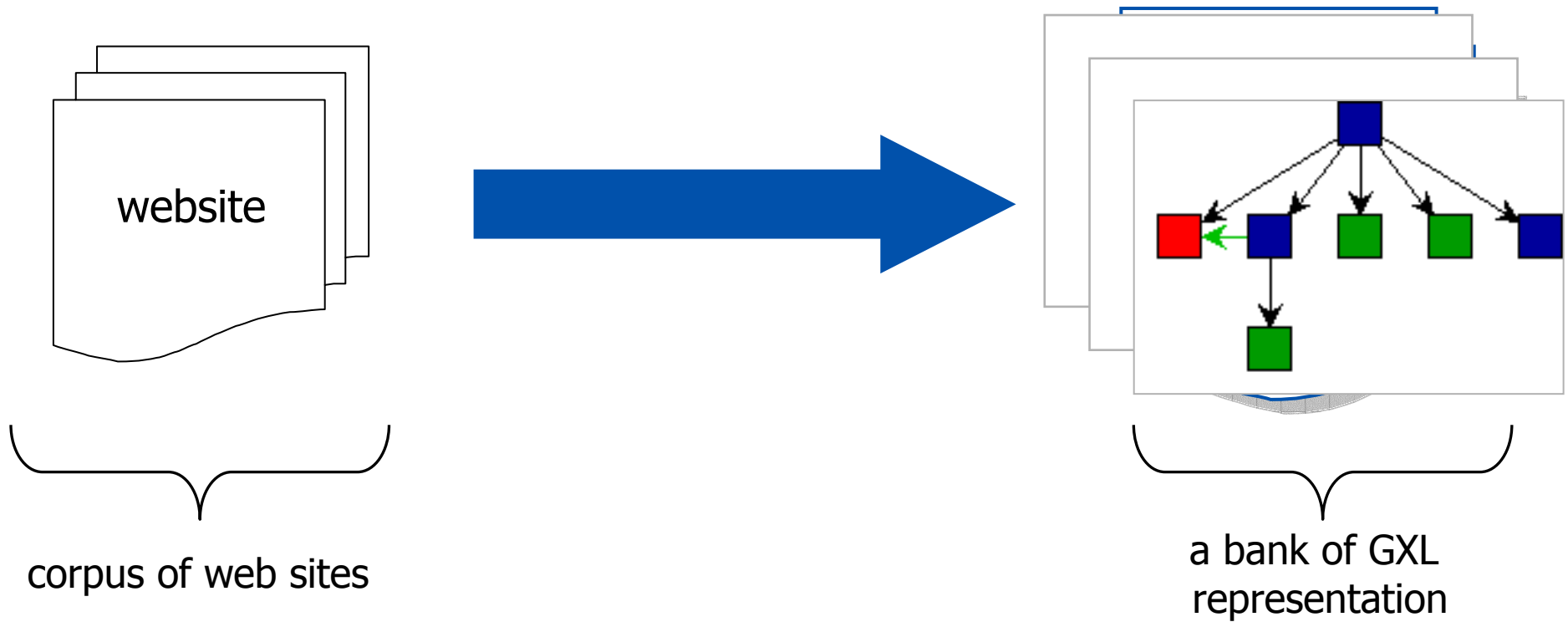
A Representation Format for Hypertext Graphs



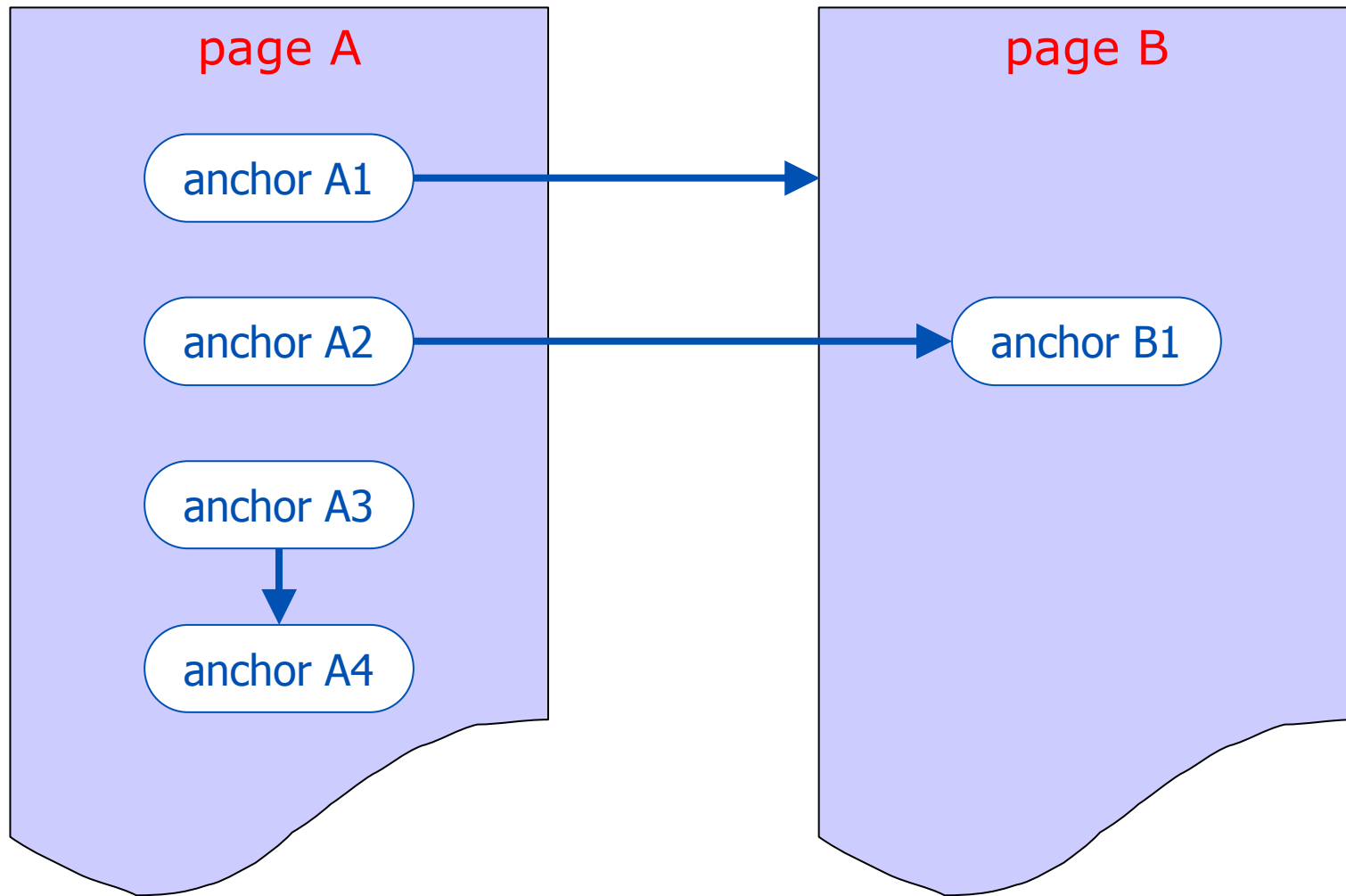
Extracting Websites

Using the GXL ...

- for representing the document structure of instances of webgenres
- as a uniform representation format of ML algorithms on web-based data



Extracting Websites

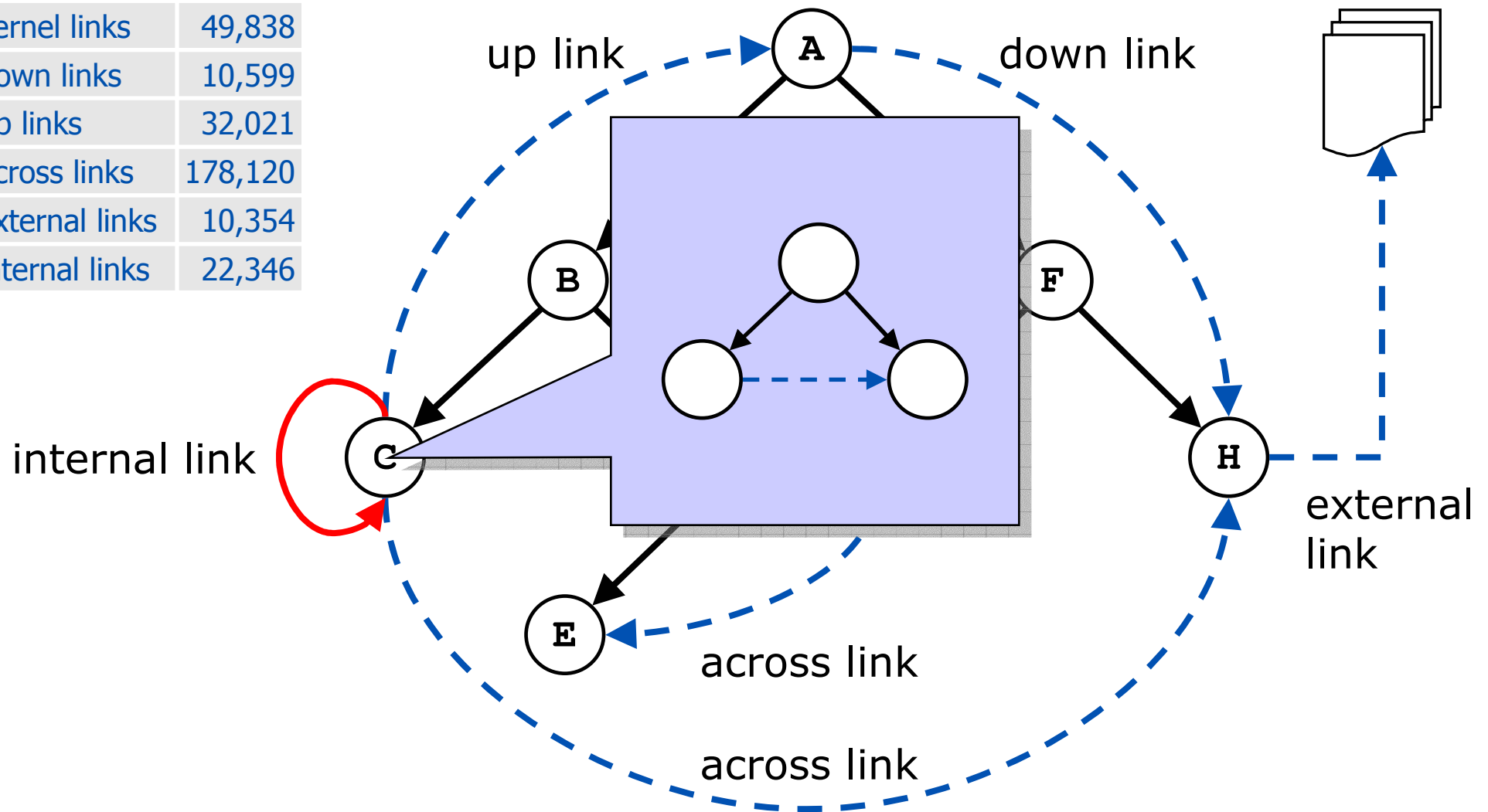


page internal link

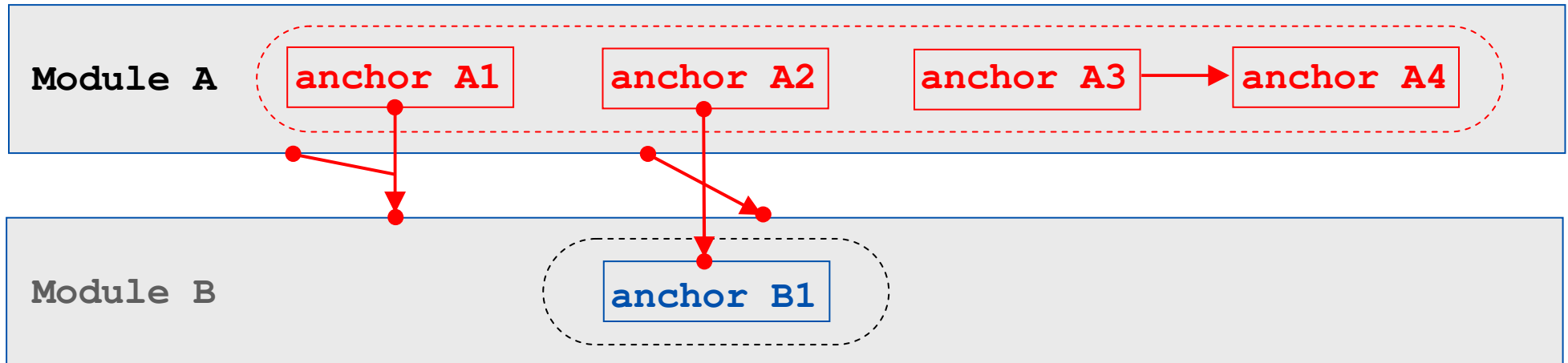
Extracting Websites

Hierarchical Kernel Structure

kernel links	49,838
down links	10,599
up links	32,021
across links	178,120
external links	10,354
internal links	22,346



Extracting Websites

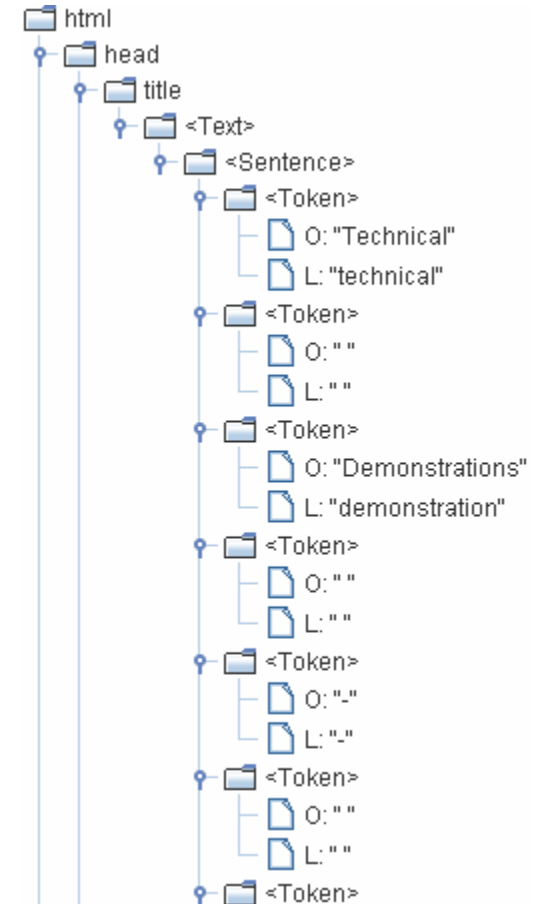
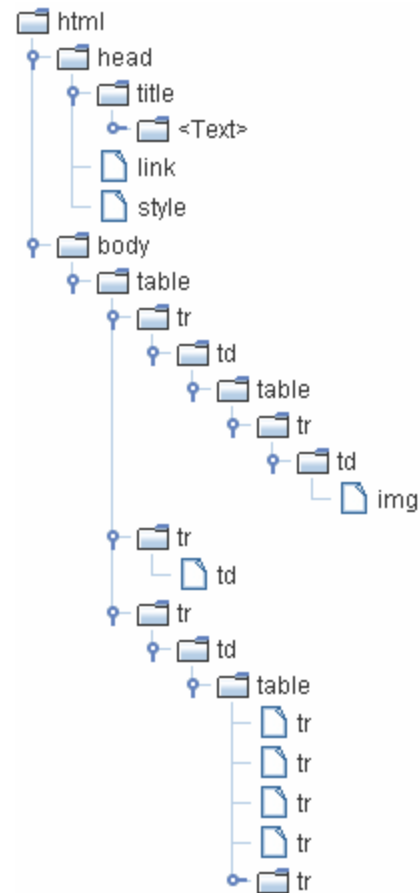
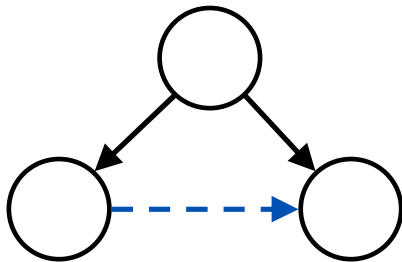


```
<gxl xmlns:xlink="http://www.w3.org/1999/xlink">
  <graph hypergraph="true" edgemode="directed" id="HyperGraph0">
    <node id="ModuleA">
      <graph hypergraph="false" edgemode="directed" id="EmbGraph0">
        <node id="AnchorA1">...</node>
        <node id="AnchorA2">...</node>
        <node id="AnchorA3">...</node>
        <node id="AnchorA4">...</node>
        <edge id="edge0" from="AnchorA3" to="AnchorA4"/>
      </graph>
    </node>
  </graph>
</gxl>
```

Extracting Websites

Hierarchical Hypergraphs

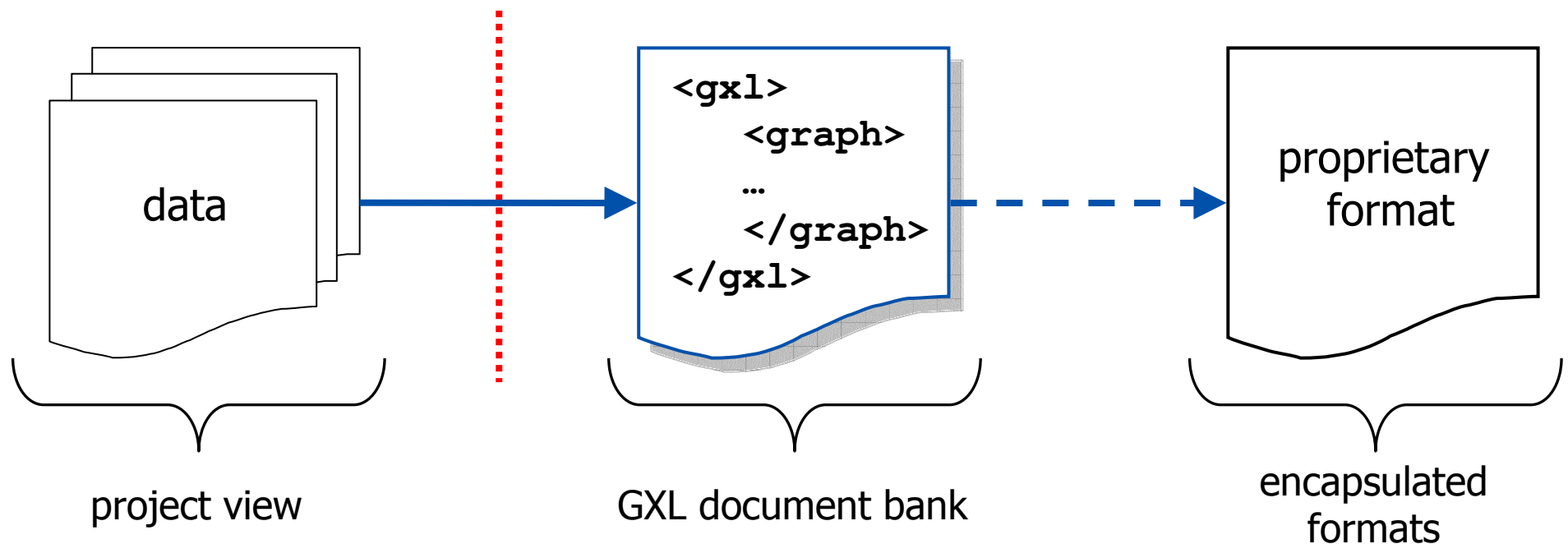
- Website structure
- DOM structure
- linguistic structure



Extracting Websites

Corpus Conversion

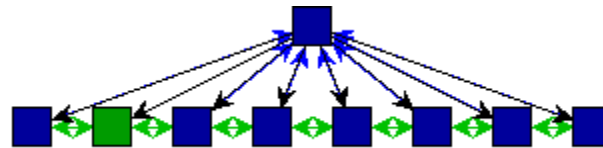
Encapsulating the variety of data formats used in ML



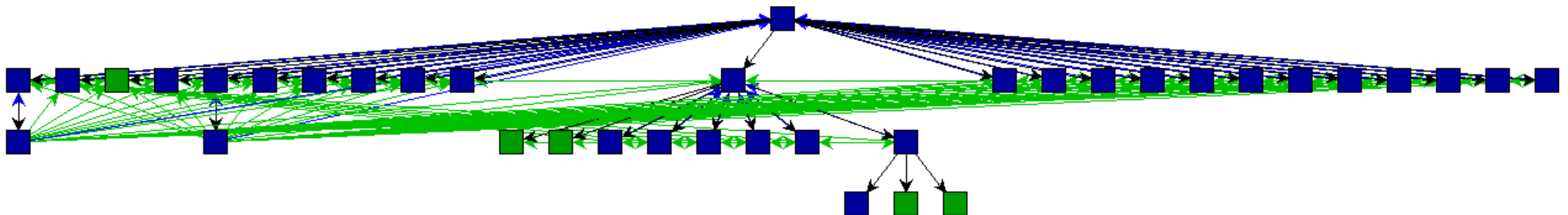
Extracting Websites

Temporal Variability

ACM Multimedia 2005, February 2, 2005



ACM Multimedia 2005, July 11, 2005



→ representing websites as families of hypertext graphs

Extracting Websites

Conference Websites

Source:

- conference calendar of ACM, IEEE, and IFIP

Languages:

- mainly English

Event Types:

- conferences, workshops, seminar, meeting, course, ...

Size:

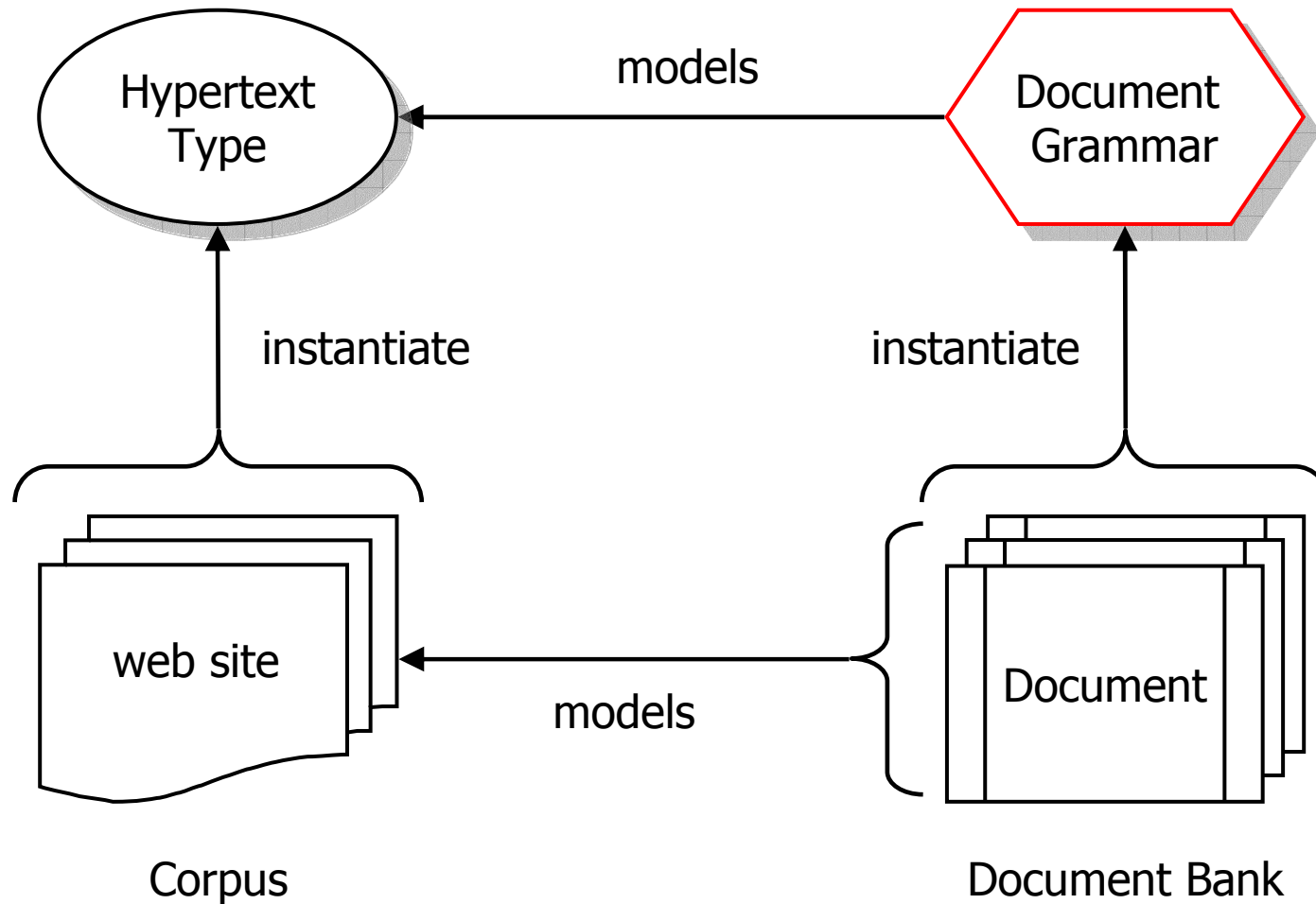
- 1,096 (English conference websites minus spam URLs)

Variable	Value
Number of Sites	1,096
Number of Pages	50,943
Number of Links	303,278
Maximum Depth	23
Maximum Width	1,035
Average Size	46
Average Width	38
Average Height	2

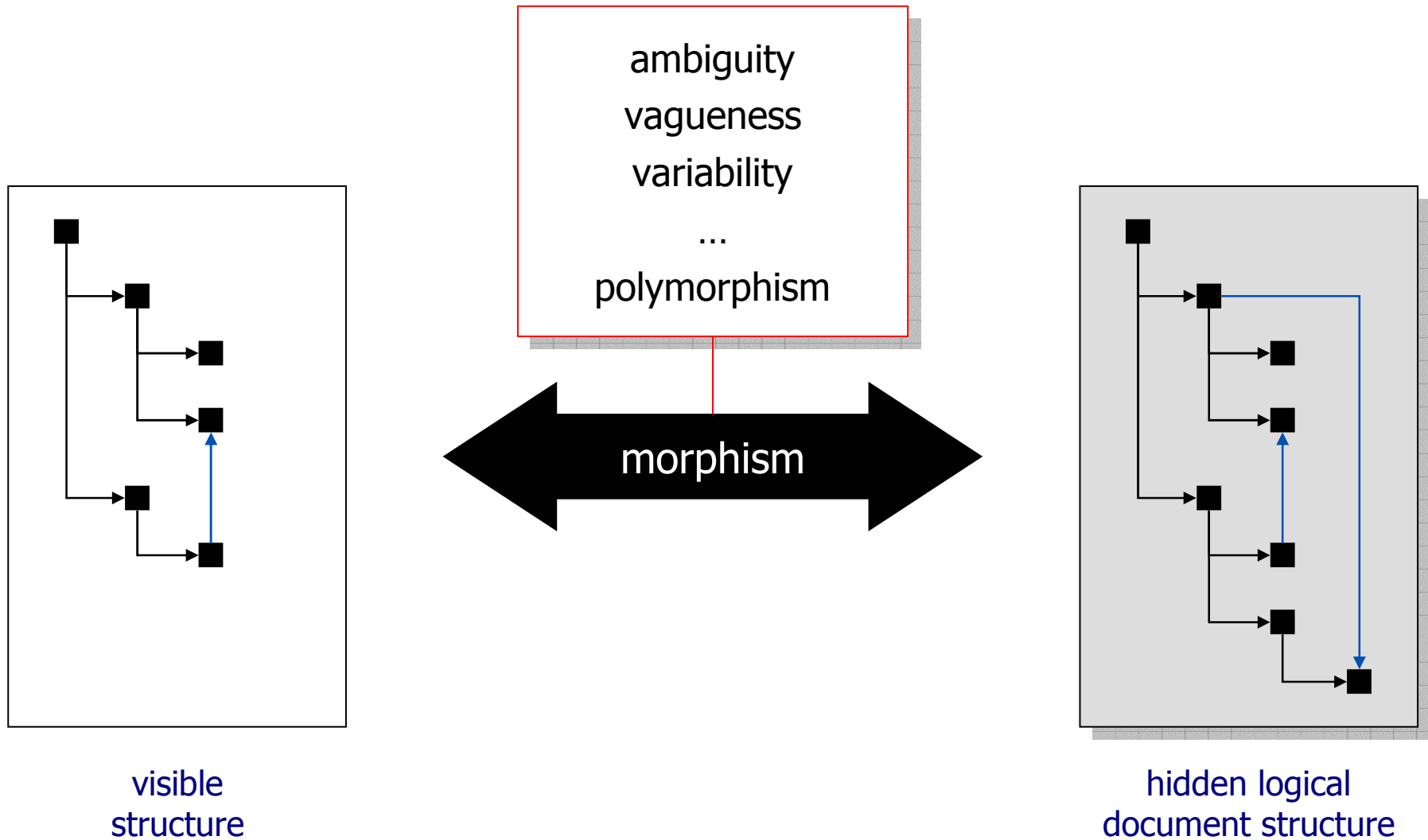
Agenda

1. Motivation
2. Extracting Websites
3. Logical Hypertext Document Structure
4. Services
5. Conclusion

Logical Hypertext Document Structure



Logical Hypertext Document Structure



Logical Hypertext Document Structure

Implications

- The visual structure of a web site is an informational uncertain manifestation of the underlying logical hypertext document structure.
- indispensable step of web structure mining: reconstruction of this logical document structure.

→ Aim:

Inducing a probabilistic document grammar which predicts the similarity of web sites of similar function, but differing form.

Agenda

1. Motivation
2. Extracting Websites
3. Logical Hypertext Document Structure
4. Services
5. Conclusion

Services

Building Reference Corpora of Certain Webgenres

- extracting websites using the GXL
 - preprocessing linguistic structures
 - providing an interface for performing corpus linguistic operations
- conference websites as a starting point

Agenda

1. Motivation
2. Extracting Websites
3. Logical Hypertext Document Structure
4. Services
5. Conclusion

Conclusion

- extracting genre specific corpora of web documents
 - informational uncertainty
 - structure learning instead of function learning
 - building up of large reference corpora of webgenres
- DFG funded research project

„induction of probabilistic web document grammars“

as part of the research group

„texttechnological information modelling“

Publication

```
@inproceedings{Mehler:Gleim:2005,  
  author={Alexander Mehler and R{"u}diger Gleim},  
  title={Polymorphism in Generic Web Units. {A} corpus  
    linguistic study},  
  booktitle={Corpus Linguistics Conference Series},  
  volume={1},  
  number={1},  
  issn={1747-9398}  
  year={2005}  
}
```