Semantic Web Mining and the Representation, Analysis, and Evolution of Web Space

Bettina Berendt,

Andreas Hotho, & Gerd Stumme

Humboldt University Berlin / University of Kassel, Germany

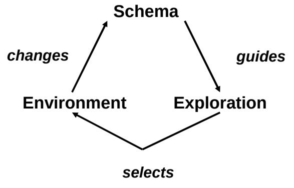
More info: www.berendt.de

The Web is mankind´s largest repository of knowledge ...

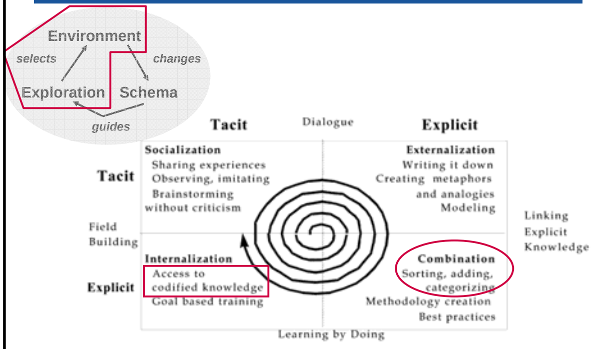
... but knowledge isn´t something that can be “put in a container and then used as the need arises“.

Knowledge is constructed in learning activities.

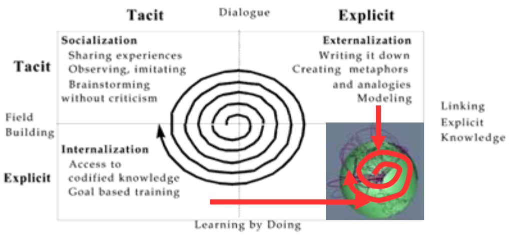
Knowledge$_1$ (“in people‘s minds“) is created by interaction with the environment (e.g., Neisser, 1967)



Knowledge$_2$ (codified) is part of the environment; Learning accesses this knowledge (e.g., Nonaka, 1991)

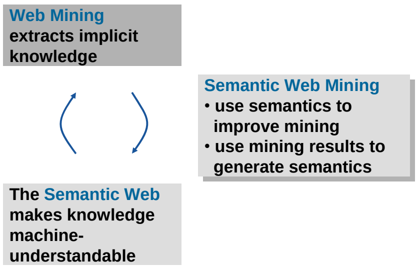


Semantic Web Mining



Approaches to the current Web‘s biggest challenges:

lots of data, human-understandable



Web Mining extracts implicit knowledge

Semantic Web Mining
- **use semantics to improve mining**
- **use mining results to generate semantics**

The Semantic Web makes knowledge machine-understandable

[Berendt, Hotho, & Stumme, *Proc. ISWC* 2002]
[-“- (Eds.), *Proc. WS Semantic Web Mining at ECML/PKDD* 2001 and 2002]
[Berendt, Hotho, Mladenic, van Someren, Spiliopoulou, Stumme (Eds.),
Web Mining: From Web to Semantic Web, 2004]

Agenda

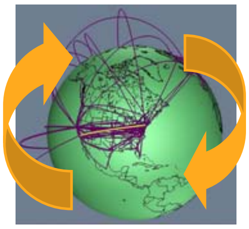
Web Mining

(Semantic) Web

Agenda

Web Mining

Semantic Web



Extracting semantics from Web content & structure – ideas and examples

Using syntactic structure, semi-automatically learn

- **Ontologies (build or extend Yahoo-like taxonomies; Web-scale example *KnowItAll*: „... such as ...“, see Etzioni et al. 2004)**
- **Instances of concepts and relations in a given ontology (ontology population)**
  - **Technique: Information extraction**
  - **From textual information including tables**
    - **Krátky, Andrt, & Svátek**
  - **From visual information including text layout**
    - **Burget; Gatterbauer, Krüpl, Holzinger, & Herzog; Hassan & Baumgartner; Labský, Vacura, & Praks**
  - **From structure (hyperlinks)**
    - **Frivolt & Bieliková**
- **Interactive learning**
  - **Ceresna; Schindler, Arya, Rath, & Slany**
- **Re-using existing conceptualizations**
  - **Švihla & Jelínek**

PS: This is just my understanding of your papers – please send me email if you find I've missed something!

Agenda

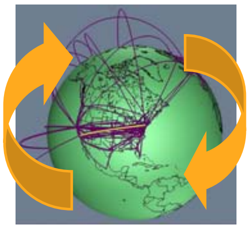
Web Mining

(Semantic) Web

Agenda

Web Mining

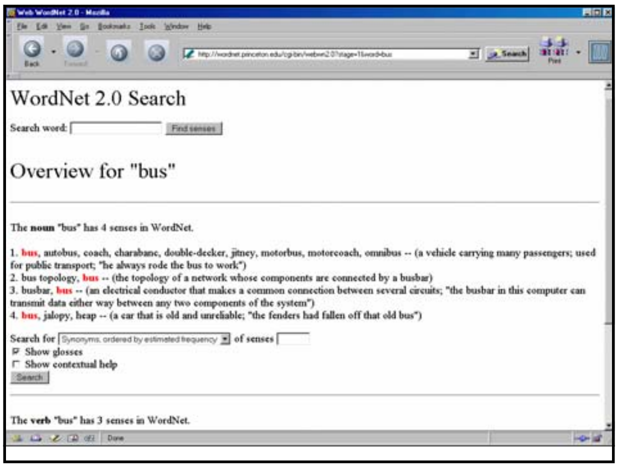
Semantic Web



Exploiting semantics for textual Web resources – ideas and examples

- Allowing expert users to contribute ontologies for semantics-enhanced IE
  - Schindler, Arya, Rath, & Slany
- Using ontologies to build templates for the composition of Web services
  - Svátek & Vacura
- Use ontologies as additional structure on the tokens in a text to
  - disambiguate meaning (e.g., word sense disambiguation: Navigli & Velardi, 2005)
  - reveal additional structure (e.g., clustering: Hotho, 2004)
  - help in the discovery of new ontological structures, or in instance learning (e.g., Navigli & Velardi, 2005; Hotho, 2004; KDD Cup 2005)

R. Navigli & P. Velardi. Structural Semantic Interconnections: a knowledge-based approach to word sense disambiguation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (27-7), 2005.


Web WordNet 2.0 - Mozilla
File Edit View Go Bookmarks Tools Window Help
Back Forward
http://wordnet.princeton.edu/cgi-bin/webwn2.0?stage=1&word=bus
Search
Print
Done

WordNet 2.0 Search

Search word: [ ] Find senses

Overview for "bus"

The **noun** "bus" has 4 senses in WordNet.

1. **bus**, autobus, coach, charabanc, double-decker, jitney, motorbus, motorcoach, omnibus -- (a vehicle carrying many passengers; used for public transport; "he always rode the bus to work")
2. bus topology, **bus** -- (the topology of a network whose components are connected by a busbar)
3. busbar, **bus** -- (an electrical conductor that makes a common connection between several circuits; "the busbar in this computer can transmit data either way between any two components of the system")
4. **bus**, jalopy, heap -- (a car that is old and unreliable; "the fenders had fallen off that old bus")

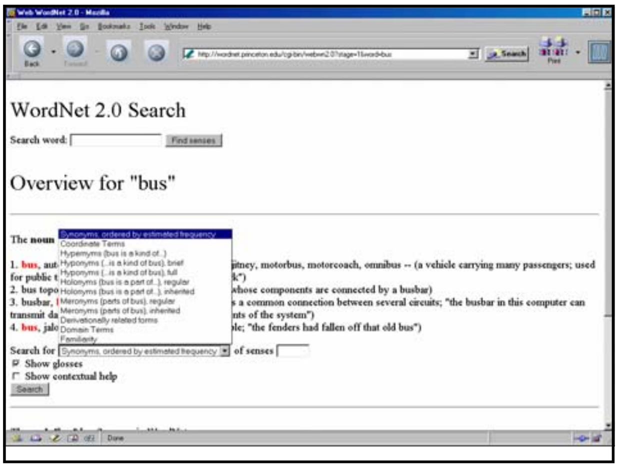
Search for [Synonyms, ordered by estimated frequency ▾] of senses [ ]

☑ Show glosses

□ Show contextual help

Search

The **verb** "bus" has 3 senses in WordNet.


Web WordNet 2.0 - Mozilla
File Edit View Go Bookmarks Tools Window Help
Back Forward
http://wordnet.princeton.edu/cgi-bin/webwn2.0?stage=1&word=bus
Search
Print
Done

WordNet 2.0 Search

Search word: [ ] Find senses

Overview for "bus"

The **noun**

1. **bus**, aut … jitney, motorbus, motorcoach, omnibus -- (a vehicle carrying many passengers; used for public t … k")
2. bus topo … whose components are connected by a busbar)
3. busbar, … s a common connection between several circuits; "the busbar in this computer can transmit da … nts of the system")
4. **bus**, jalo … le; "the fenders had fallen off that old bus")

Synonyms, ordered by estimated frequency
Coordinate Terms
Hypernyms (bus is a kind of...)
Hyponyms (...is a kind of bus), brief
Hyponyms (...is a kind of bus), full
Holonyms (bus is a part of...), regular
Holonyms (bus is a part of...), inherited
Meronyms (parts of bus), regular
Meronyms (parts of bus), inherited
Derivationally related forms
Domain Terms
Familiarity

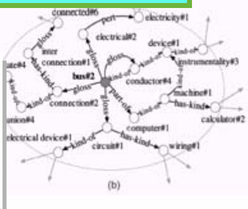
Search for [Synonyms, ordered by estimated frequency ▾] of senses [ ]

☑ Show glosses

□ Show contextual help

Search

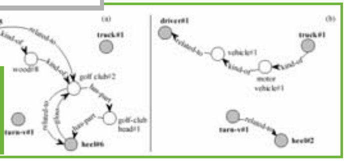
Basic idea: Graphs induced by WordNet + domain labels for synsets + cooccurrence information from annotated corpora + collocations



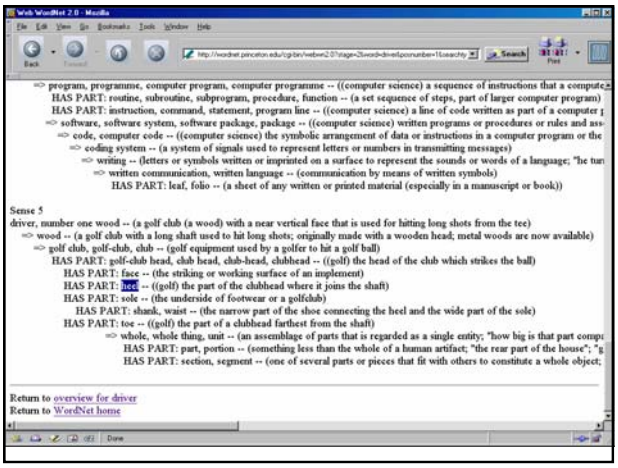
$S_G \rightarrow S_s \mid S_g$ *(all the heuristics)*
$S_s \rightarrow S_1 \mid S_2 \mid S_3$ *(simple heuristics)*
$S_1 \rightarrow E_1 S_1 \mid E_1$ *(hyperonymy/meronymy)*
$E_1 \rightarrow e_{\text{kind-of}} \mid e_{\text{part-of}}$
$S_2 \rightarrow E_2 S_2 \mid E_2$ *(hyponymy/holonymy)*
$E_2 \rightarrow e_{\text{has-kind}} \mid e_{\text{has-part}}$
$S_3 \rightarrow e_{\text{kind-of}} S_3 e_{\text{has-kind}} \mid e_{\text{kind-of}} e_{\text{has-kind}}$ *(parallelism)*
$S_g \rightarrow e_{\text{gloss}} S_s \mid S_4 \mid S_5$ *(gloss)*
$S_4 \rightarrow e_{\text{gloss}} \mid e_{\text{topic}}$ *(gloss, context)*
$S_5 \rightarrow e_{\text{gloss}} e_{\text{is-in-gloss}}$ *(gloss+gloss^{-1})*

FigFig. 2. An excerpt of the context-free grammar for the recognition of semantic interconnections.

(+ weights for each pattern and thus interconnection)



Using SSI for word sense disambiguation (“The driver turned on his heel and went back to the truck.“)


Web WordNet 2.0 - Mozilla
File Edit View Go Bookmarks Tools Window Help
Back Forward
http://wordnet.princeton.edu/cgi-bin/webwn2.0?stage=2&word=driver&posnumber=1&searchty
Search
Print
Done

```
   => program, programme, computer program, computer programme -- ((computer science) a sequence of instructions that a compute
      HAS PART: routine, subroutine, subprogram, procedure, function -- (a set sequence of steps, part of larger computer program)
      HAS PART: instruction, command, statement, program line -- ((computer science) a line of code written as part of a computer p
    => software, software system, software package, package -- ((computer science) written programs or procedures or rules and ass
      => code, computer code -- ((computer science) the symbolic arrangement of data or instructions in a computer program or the
        => coding system -- (a system of signals used to represent letters or numbers in transmitting messages)
          => writing -- (letters or symbols written or imprinted on a surface to represent the sounds or words of a language; "he tur
            => written communication, written language -- (communication by means of written symbols)
               HAS PART: leaf, folio -- (a sheet of any written or printed material (especially in a manuscript or book))

Sense 5
driver, number one wood -- (a golf club (a wood) with a near vertical face that is used for hitting long shots from the tee)
  => wood -- (a golf club with a long shaft used to hit long shots; originally made with a wooden head; metal woods are now available)
    => golf club, golf-club, club -- (golf equipment used by a golfer to hit a golf ball)
       HAS PART: golf-club head, club head, club-head, clubhead -- ((golf) the head of the club which strikes the ball)
         HAS PART: face -- (the striking or working surface of an implement)
         HAS PART: heel -- ((golf) the part of the clubhead where it joins the shaft)
         HAS PART: sole -- (the underside of footwear or a golfclub)
           HAS PART: shank, waist -- (the narrow part of the shoe connecting the heel and the wide part of the sole)
         HAS PART: toe -- ((golf) the part of a clubhead farthest from the shaft)
                => whole, whole thing, unit -- (an assemblage of parts that is regarded as a single entity; "how big is that part comp
                   HAS PART: part, portion -- (something less than the whole of a human artifact; "the rear part of the house"; "g
                   HAS PART: section, segment -- (one of several parts or pieces that fit with others to constitute a whole object;
```

Return to overview for driver
Return to WordNet home

SSI for ontology learning

19

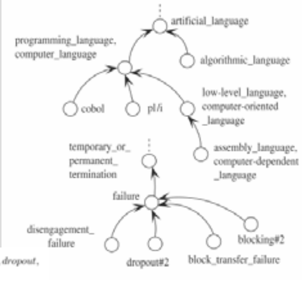
1. Extract pertinent domain terminology:
   - Simple and multiword expressions that consistently occur in domain-related corpora and are not found in other domains (e.g., packet switching network)
2. Web search of available NL definitions from glossaries or documents
3. Use context-free grammar to
   1. filter out non-relevant definitions, based on statistical domain model
   2. parse definitions to extract kind-of information
4. Arrange terms in hierarchical trees
5. Link sub-hierarchies to the concepts of a core ontology (general-purpose: WordNet)
6. Provide the output to domain specialists for evaluation and refinement

```
S → PP ',' NP SEP
NP → N1 KIND1
KIND1 → MOD1 NOUN1
MOD1 → Verb | Adj | Verb ',' MOD1 | Adj ',' MOD1
NOUN1 → Noun
N1 → Art | Adj
SEP → ',' | '.' | Prep | Verb | Wh
PP → Prep NP
```

SSI: Word sense disambiguation and ontology learning

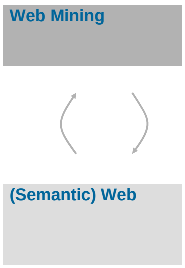
20

- *artifical language*: monosemous in WordNet
- *temporary or permanent termination*: does not exist in WordNet
- *termination*: need to apply WSD (#1: end of a time span, #2: expiration of a contract)
  - ➔ use terms that occur in the subtree and have a lexical correspondent in WordNet
  - ➔ $P = \{disengagement, failure, block, transfer, dropout, temporary, permanent\}.$


artificial_language
programming_language, computer_language
algorithmic_language
cobol
pl/i
low-level_language, computer-oriented _language
assembly_language, computer-dependent _language
temporary_or_ permanent_ termination
failure
disengagement_ failure
dropout#2
block_transfer_failure
blocking#2

Agenda

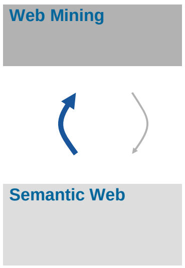
21


Web Mining

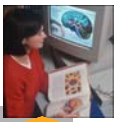
(Semantic) Web

Agenda

22


Web Mining

Semantic Web

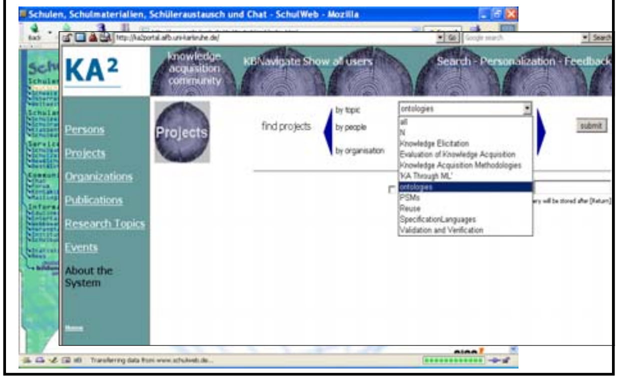


```
...
p3ee24304.dip.t-dialin.net
[19/Mar/2002:12:03:51 +0100
/search.html?t=jane%20auste
3785&ord=asc HTTP/1.0" 200
p3ee24304.dip.t-dialin.net
[19/Mar/2002:12:05:06 +0100
/search.html?t=jane%20austen&m=vide
o&SID=023785&ord=desc HTTP/1.0" 200
8450
p3ee24304.dip.t-dialin.net - -
[19/Mar/2002:12:06:41 +0100] "GET
/view.asp?id=3456&SID=023785
HTTP/1.0" 200 3478
...
```

Under-stand

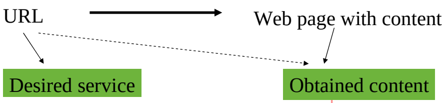
Application: Search in knowledge portals

23


Schulen, Schulmaterialien, Schüleraustausch und Chat - SchulWeb - Mozilla
http://ka2portal.aifb.uni-karlsruhe.de/
Google search
Search
KA²
knowledge acquisition community
KBNavigate Show all users
Search - Personalization - Feedback
Persons
Projects
Organizations
Publications
Research Topics
Events
About the System
Projects
find projects
by topic
by people
by organisation
ontologies
all
N
Knowledge Elicitation
Evaluation of Knowledge Acquisition
Knowledge Acquisition Methodologies
'KA Through ML'
ontologies
PSMs
Reuse
SpecificationLanguages
Validation and Verification
submit
Transferring data from www.schulweb.de...

Ontology-based modelling of behaviour: URLs and application events

24



Definition 24.3.1 *A site model* $M := (\mathcal{S}, \mathcal{C}[, \mathcal{D_S}][, \mathcal{D_C}])$ *is a tuple consisting of a service ontology* $\mathcal{S}$, *a content ontology* $\mathcal{C}$, *and optionally tuples* $\mathcal{D_S}$ *and/or* $\mathcal{D_C}$ *consisting of the service/content information dimensions.*

Definition 24.3.2 *An* atomic application event *for a site model* $M = (\mathcal{S}, \mathcal{C}[, \mathcal{D_S}][, \mathcal{D_C}])$ *is a tuple* $AAE := (S, C)$ *consisting of (1)* S, *the service: a concept or relation from the service ontology* $\mathcal{S}$, *or (if and only if* $\mathcal{D_S}$ *is present in* M*) a tuple whose components are concepts/relations from* $\mathcal{S}$ *describing the service along the dimensions of* $\mathcal{D_S}$, *and (2)* C, *the content: a concept or relation from the content ontology* $\mathcal{C}$, *or (if and only if* $\mathcal{D_C}$ *is present in* M*) a tuple whose components are concepts/relations from* $\mathcal{C}$ *describing the service along the dimensions of* $\mathcal{D_C}$.

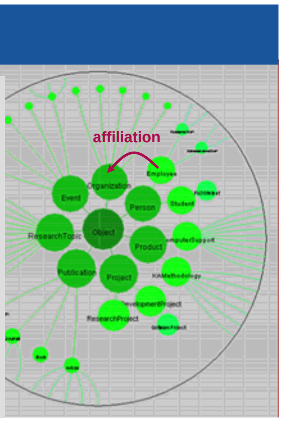
Berendt, B., Stumme, G., & Hotho, A. (2004). Usage mining for and on the Semantic Web. In H. Kargupta, A. Joshi, K. Sivakumar, & Y. Yesha (Eds.), *Data Mining: Next Generation Challenges and Future Directions.* Menlo Park, CA: AAAI/MIT Press.

Definition 24.3.3 *A complex application event CAE is a nonempty (1) set, (2) sequence (i.e., order relation), or (3) another sequential structure (a regular expression, a context-free grammar, ...), whose elements are atomic application events.*

Semantics of requests

Step 1: Domain ontology

- community portal
  ka2portal.aifb.uni-karlsruhe.de
- ontology-based:
  - Knowledge base in F-Logic
  - Static pages: annotations
  - Dynamic pages: generated from queries & KB
  - Queries also in F-Logic
  - Logs contain these queries



[Oberle, Berendt, Hotho, & Gonzalez, *Proc. AWIC* 2003]

Semantics of requests

Step 2: Modelling requests as atomic application events

RESEARCHER
PERSON
PROJECT
PUBLICATION
RESEARCHTOPIC
EVENT
ORGANIZATION
RESEARCHINTEREST
LASTNAME
TITLE
ISABOUT
EVENTS
EVENTTITLE
WORKSATPROJECT
AUTHOR
AFFILIATION
ISWORKEDONBY
PROGRAMCOMMITTEE
EMPLOYS
NAME
RESEARCHGROUPS
EMAIL

An example query with concepts and relations:

FORALL N,PEOPLE <-PEOPLE:
Employee[affiliation->> "http://www.anInstitute.org"]
and PEOPLE:Person[lastName->>N].

Query =
feature vector of concepts + relations

➔

Clustering, Association rules, Classification, ...

Session =
feature vector of concepts + relations, summed over all queries in the session

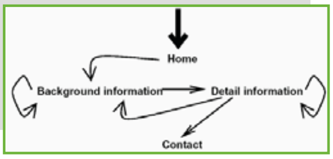
Semantics of sequences

Step 3: Using ontologies of behaviour for info. Extraction – Modelling sequences as composite application events

Composite application events - Example *customer typology*

- **Based on background theory from marketing: the customer buying cycle**
- **Modelled in terms of regular expressions and employed in Web usage mining**
- **Example:**
  knowledge builders

  (as opposed to, e.g., *direct buyers*)

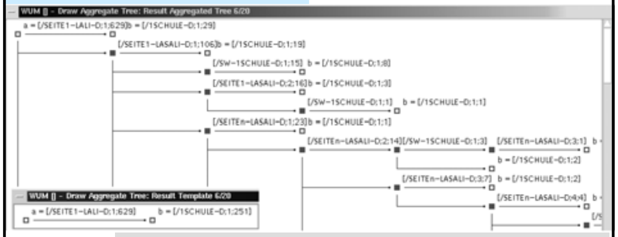


[Moe, *Journal of Consumer Psychology*, 2002]
[Spiliopoulou, Pohle, and Teltzrow, *Proc. Wirtschaftsinformatik 2002*]

2. Semantics of sequences

for Step 3: an interactive tool with a query language

```
select t
from node a b, template a * b as t
where a.url startswith "SEITE1-"
and a.occurrence = 1
and b.url contains "1SCHULE"
and b.occurrence = 1
and (b.support / a.support) >= 0.2
```


```
WUM [] - Draw Aggregate Tree: Result Aggregated Tree 6/20
a = [/SEITE1-LALI-D;1;629]b = [/1SCHULE-D;1;29]
[/SEITE1-LASALI-D;1;106]b = [/1SCHULE-D;1;19]
[/SW-1SCHULE-D;1;15]  b = [/1SCHULE-D;1;8]
[/SEITE1-LASALI-D;2;16]b = [/1SCHULE-D;1;3]
[/SW-1SCHULE-D;1;1]   b = [/1SCHULE-D;1;1]
[/SEITEn-LASALI-D;1;23]b = [/1SCHULE-D;1;1]
[/SEITEn-LASALI-D;2;14][/SW-1SCHULE-D;1;3]  [/SEITEn-LASALI-D;3;1]  b
b = [/1SCHULE-D;1;2]
[/SEITEn-LASALI-D;3;7]  b = [/1SCHULE-D;1;2]
[/SEITEn-LASALI-D;4;4]  b
[/S

WUM [] - Draw Aggregate Tree: Result Template 6/20
a = [/SEITE1-LALI-D;1;629]      b = [/1SCHULE-D;1;251]
```

Tool: www.hypknowsys.de; Data: [Berendt & Spiliopoulou, *VLDB Journal*, 2000]

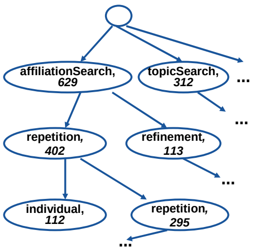
Semantics of sequences

Step 4: Pattern discovery / instance learning

An ontology of composite application events (CAEs)

- **Define templates as regular expressions**
  - **of atomic application events**
  - **of transitions (between atomic application events)**

  ***Ex.* [.search .* individual]**
- **Discover instances by learning a CAE trie**



[Berendt & Spiliopoulou, *VLDB Journal*, 2000]
[Berendt, *Data Mining and Knowledge Discovery*, 2002]

Semantics of sequences

Step 5: Pattern evaluation

Use pattern statistics to

- **derive descriptive measures of CAEs**
  - **support, confidence**
  - **popularity, effectiveness, efficiency**
- **apply inferential statistics to compare CAEs**

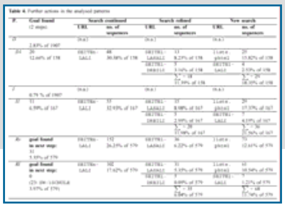


Table 4. Further actions in the analysed patterns

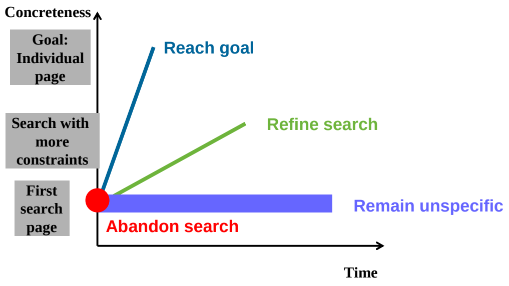
Goal found (2 steps) | Search continued | Search refined | New search | URL | no. of sequences

[Berendt, *Data Mining and Knowledge Discovery*, 2002]

Communication – Visual data mining

Step 6: Mapping an ontological relation over concepts to a linear order; mapping to visual variables

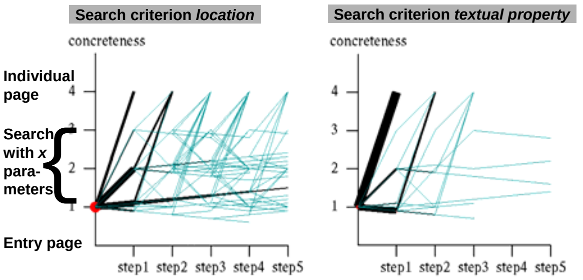
31



Communication – Visual data mining

Step 6 – Example

32

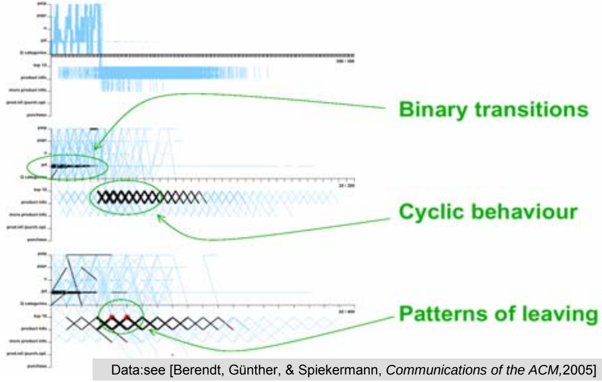


[Berendt, *Data Mining and Knowledge Discovery*, 2002], [Berendt, *Postproc. WebKDD 2001*]

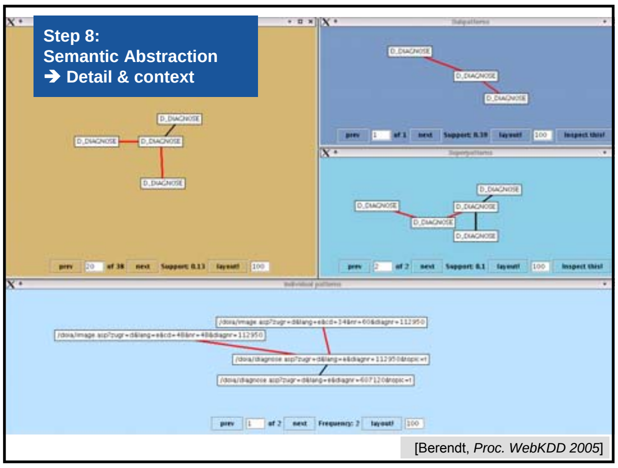
Communication – Visual data mining

Step 7: Visual abstraction ➔ new semantic patterns

33



Data:see [Berendt, Günther, & Spiekermann, *Communications of the ACM*,2005]

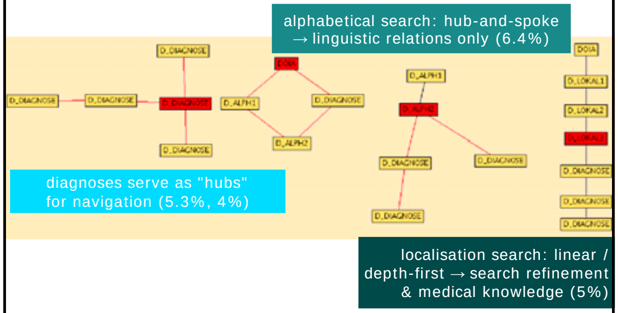


Step 8: Semantic Abstraction ➔ Detail & context

[Berendt, *Proc. WebKDD 2005*]

Case study: Information search in a medical portal*

35



alphabetical search: hub-and-spoke → linguistic relations only (6.4%)

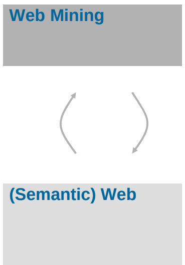
diagnoses serve as "hubs" for navigation (5.3%, 4%)

localisation search: linear / depth-first → search refinement & medical knowledge (5%)

* (20333 requests / 1397 sessions from Web log collected in 2001/2002; preprocessing: mining in concept space, see paper in proceedings)

Agenda

36



Web Mining

(Semantic) Web

Agenda

Web Mining

Semantic Web

contribute

. . .

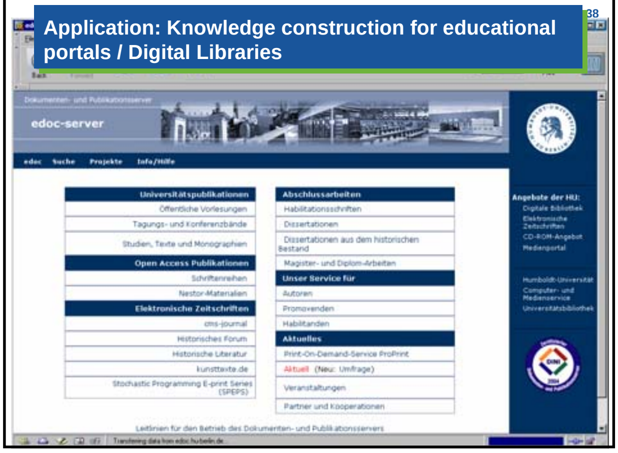
<BIBLIOGRAPHY><FLOAT><PAG
R>136</PAGENUMBER></FLOAT

<HEAD>Literaturverzeichnis</HE

<CITATION WORKTYPE="jou
PUBLISHED="PUBLISHED">
ID="bib-15-">[1]
</CUT><WORKAUTHOR>Aga
Krueger, B. P.; Scholes, G. D.
M.; Yom, J.; Mets, L.; Flem
R.</WORKAUTHOR>U<ARTICLE
>ltrafast energy transfer in LHC-II
revealed by three-pulse photon echo
peak shift
measurements</ARTICLETITLE>,
<WORKTITLE>J. Phys. Chem.
B</WORKTITLE>,
<PUBDATE>2000</PUBDATE>,
<NUMBER>104</NUMBER>,
<PAGES>2908</PAGES>,
</CITATION>

. . .

Application: Knowledge construction for educational portals / Digital Libraries



Dokumenten- und Publikationsserver

edoc-server

Universitätspublikationen
Öffentliche Vorlesungen
Tagungs- und Konferenzbände
Studien, Texte und Monographien

Open Access Publikationen
Schriftenreihen
Nestor-Materialien

Elektronische Zeitschriften
cms-journal
Historisches Forum
Historische Literatur
kunsttexte.de
Stochastic Programming E-print Series (SPEPS)

Abschlussarbeiten
Habilitationsschriften
Dissertationen
Dissertationen aus dem historischen Bestand
Magister- und Diplom-Arbeiten

Unser Service für
Autoren
Promovenden
Habilitanden

Aktuelles
Print-On-Demand-Service ProPrint
Aktuell (Neu: Umfrage)
Veranstaltungen
Partner und Kooperationen

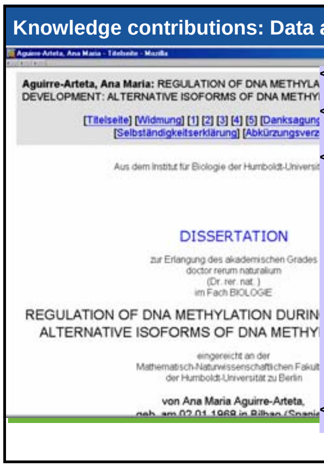
Angebote der HU:
Digitale Bibliothek
Elektronische Zeitschriften
CD-ROM-Angebot
Medienportal

Humboldt-Universität
Computer- und Medienservice
Universitätsbibliothek

Leitlinien für den Betrieb des Dokumenten- und Publikationsservers

Transferring data from edoc.hu-berlin.de...

Knowledge contributions: Data and metadata



Aguirre-Arteta, Ana Maria - Titelseite - Mozilla

Aguirre-Arteta, Ana Maria: REGULATION OF DNA METHYLA
DEVELOPMENT: ALTERNATIVE ISOFORMS OF DNA METHY

[Titelseite] [Widmung] [1] [2] [3] [4] [5] [Danksagung
[Selbständigkeitserklärung] [Abkürzungsverz

Aus dem Institut für Biologie der Humboldt-Universit

DISSERTATION

zur Erlangung des akademischen Grades
doctor rerum naturalium
(Dr. rer. nat.)
im Fach BIOLOGIE

REGULATION OF DNA METHYLATION DURIN
ALTERNATIVE ISOFORMS OF DNA METHY

eingereicht an der
Mathematisch-Naturwissenschaftlichen Fakult
der Humboldt-Universität zu Berlin

von Ana Maria Aguirre-Arteta,
geb. am 02.01.1969 in Bilbao (Spani

<BIBLIOGRAPHY><FLOAT><PAGENUMBER>136</PAGENUMBER></FLOAT>

<HEAD>Literaturverzeichnis</HEAD>

...

<CITATION WORKTYPE="journal" PUBLISHED="PUBLISHED">

<CUT ID="bib-45-">[2] </CUT><WORKAUTHOR>Albrecht, T. F.; Bott, K.; Meier, T.; Schulze, A.; Koch, M.; Cundiff, S. T.; Feldmann, J.; Stolz, W.; Thomas, P.; Koch, S. W.; G&ouml;bel; E. O.</WORKAUTHOR> <ARTICLETITLE>Disorder mediated biexcitonic beats in semiconductor quantum wells</ARTICLETITLE>, <WORKTITLE>Phys. Rev. B</WORKTITLE>, <PUBDATE>1996</PUBDATE>, <NUMBER>54</NUMBER>, <PAGES>4436</PAGES>,

</CITATION> ...

Dissertation Markup Language DiML

http://edoc.hu-berlin.de/diml/dtd/xdiml.dtd

```
...
<!ELEMENT citation (#PCDATA | email | url | note | workauthor | worktitle | articletitle | serialtitle |
address | editor | publisher | edition | volume | number | version | pages | pubdate | bible | court |
law | cut | pagenumber)*>
<!ATTLIST citation
  id ID #IMPLIED
  label CDATA #IMPLIED
  workType (Book | Journal | Misc) #IMPLIED
  published (yes|no) 'yes'>
<!ELEMENT note (#PCDATA | em | u | strong | br | sup | tt | sub | link | name | email | organization
| term | foreign | url | footnote | endnote | glossref | indexref | pagenumber | q | citation | imath |
im)*>
<!ATTLIST note
  id ID #IMPLIED>
<!ELEMENT workauthor (#PCDATA | given | surname | suffix | organization)*>
<!ATTLIST workauthor
  role CDATA #IMPLIED
  ref IDREF #IMPLIED
  id ID #IMPLIED>
...
```

Authoring support for document servers

- **Surveys (ca. 2500 persons; 12-14% response rate) & Web usage mining (ca. 11000 sessions) showed:**
  - **Metadata creation is one of the main barriers for contribution.**
- **Reasons include deficiencies in**
  - **information flow**
  - **understanding and use of structured search**
  - **education in structured writing**
  - **HCI aspects**

➔ Marketing

)
) ➔ Education
)

Intelligent Authoring Tools

[Berendt, Brenstein, Li, & Wendland, *Proc. ETD* 2003]
[Berendt, *Proc. AAAI Spring Symposium KCVC*, 2005]

Consequences of metadata neglect

<BIBLIOGRAPHY><FLOAT><PAGENUMBER>136</PAGENUMBER></FLOAT>

<HEAD>Literaturverzeichnis</HEAD>

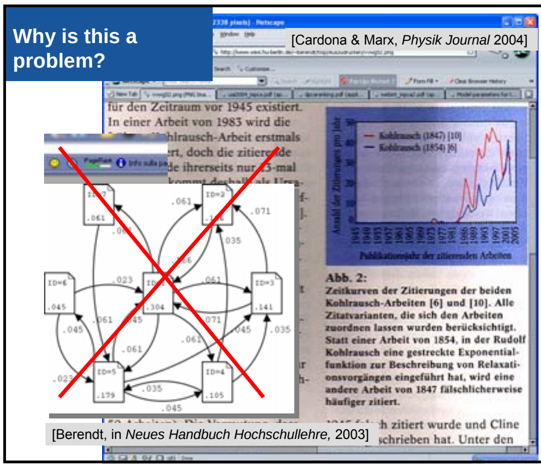
<CITATION WORKTYPE="journal" PUBLISHED="PUBLISHED">

<CUT ID="bib-15-">[1] </CUT><WORKAUTHOR>Agarwal, R.; Krueger, B. P.; Scholes, G. D.; Yang, M.; Yom, J.; Mets, L.; Fleming, G. R.</WORKAUTHOR>U<ARTICLETITLE>ltrafast energy transfer in LHC-II revealed by three-pulse photon echo peak shift measurements</ARTICLETITLE>, <WORKTITLE>J. Phys. Chem. B</WORKTITLE>, <PUBDATE>2000</PUBDATE>, <NUMBER>104</NUMBER>, <PAGES>2908</PAGES>,

</CITATION>

...

Why is this a problem?



für den Zeitraum vor 1945 existiert. In einer Arbeit von 1983 wird die … hlrausch-Arbeit erstmals … rt, doch die zitierende … de ihrerseits nur 35-mal … kommt deshalb als Ursa-

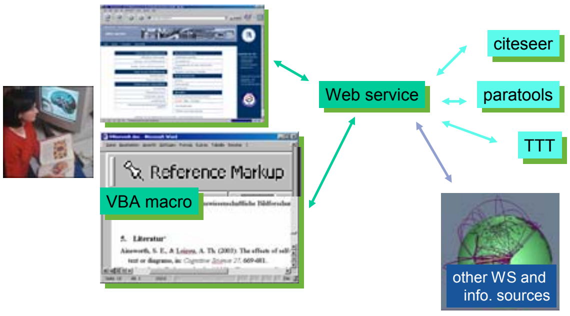
Abb. 2: Zeitkurven der Zitierungen der beiden Kohlrausch-Arbeiten [6] und [10]. Alle Zitatvarianten, die sich den Arbeiten zuordnen lassen wurden berücksichtigt. Statt einer Arbeit von 1854, in der Rudolf Kohlrausch eine gestreckte Exponentialfunktion zur Beschreibung von Relaxationsvorgängen eingeführt hat, wird eine andere Arbeit von 1847 fälschlicherweise häufiger zitiert.

… h zitiert wurde und Cline … schrieben hat. Unter den

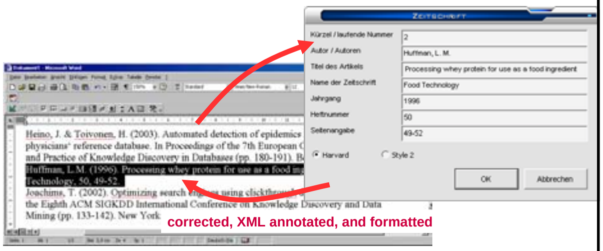
[Cardona & Marx, *Physik Journal* 2004]

[Berendt, in *Neues Handbuch Hochschullehre*, 2003]

System architecture



Usage interface



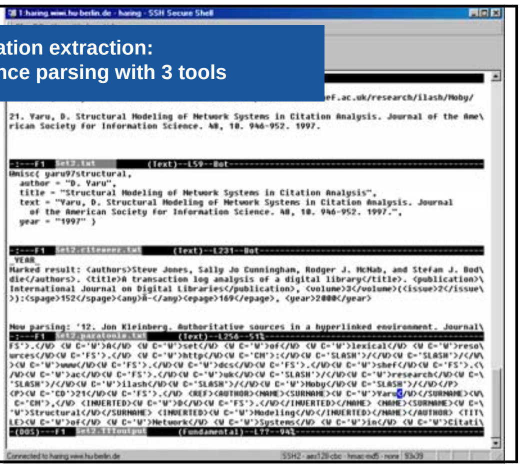
Heino, J. & Toivonen, H. (2003). Automated detection of epidemics … physicians' reference database. In Proceedings of the 7th European C… and Practice of Knowledge Discovery in Databases (pp. 180-191). B… Huffman, L.M. (1996). Processing whey protein for use as a food ing… Technology. 50, 49-52. Joachims, T. (2002). Optimizing search engines using clickthrough … the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 133-142). New York

Kürzel / laufende Nummer	2
Autor / Autoren	Huffman, L. M.
Titel des Artikels	Processing whey protein for use as a food ingredient
Name der Zeitschrift	Food Technology
Jahrgang	1996
Heftnummer	50
Seitenangabe	49-52

● Harvard ○ Style 2

corrected, XML annotated, and formatted

Information extraction: Reference parsing with 3 tools



1:haring.wiwi.hu-berlin.de - haring - SSH Secure Shell
Connected to haring.wiwi.hu-berlin.de
SSH2 - aes128-cbc - hmac-md5 - none 93x39

```
ef.ac.uk/research/ilash/Moby/

21. Yaru, D. Structural Modeling of Network Systems in Citation Analysis. Journal of the Ame\
rican Society for Information Science. 48, 10. 946-952. 1997.


-:---F1  Set2.txt       (Text)--L59--Bot-------------------------------------------------
@misc{ yaru97structural,
  author = "D. Yaru",
  title = "Structural Modeling of Network Systems in Citation Analysis",
  text = "Yaru, D. Structural Modeling of Network Systems in Citation Analysis. Journal
    of the American Society for Information Science. 48, 10. 946-952. 1997.",
  year = "1997" }


-:---F1  Set2.citeseer.txt      (Text)--L231--Bot----------------------------------------
 YEAR_
Marked result: <authors>Steve Jones, Sally Jo Cunningham, Rodger J. McNab, and Stefan J. Bod\
die</authors>. <title>A transaction log analysis of a digital library</title>. <publication>\
International Journal on Digital Libraries</publication>, <volume>3</volume>(<issue>2</issue\
>):<spage>152</spage><any>â-</any><epage>169</epage>, <year>2000</year>


Now parsing: '12. Jon Kleinberg. Authoritative sources in a hyperlinked environment. Journal\
-:---F1  Set2.paratools.txt      (Text)--L256--51%---------------------------------------
FS'>.</W> <W C='W'>A</W> <W C='W'>set</W> <W C='W'>of</W> <W C='W'>lexical</W> <W C='W'>reso\
urces</W><W C='FS'>.</W> <W C='W'>http</W><W C='CM'>:</W><W C='SLASH'>/</W><W C='SLASH'>/</W\
><W C='W'>www</W><W C='FS'>.</W><W C='W'>dcs</W><W C='FS'>.</W><W C='W'>shef</W><W C='FS'>.<\
/W><W C='W'>ac</W><W C='FS'>.</W><W C='W'>uk</W><W C='SLASH'>/</W><W C='W'>research</W><W C=\
'SLASH'>/</W><W C='W'>ilash</W><W C='SLASH'>/</W><W C='W'>Moby</W><W C='SLASH'>/</W></P>
<P><W C='CD'>21</W><W C='FS'>.</W> <REF><AUTHOR><NAME><SURNAME><W C='W'>Yaru</W></SURNAME><W\
 C='CM'>,</W> <INVERTED><W C='W'>D</W><W C='FS'>.</W></INVERTED></NAME> <NAME><SURNAME><W C=\
'W'>Structural</W></SURNAME> <INVERTED><W C='W'>Modeling</W></INVERTED></NAME></AUTHOR> <TIT\
LE><W C='W'>of</W> <W C='W'>Network</W> <W C='W'>Systems</W> <W C='W'>in</W> <W C='W'>Citati\
-(DOS)---F1  Set2.TTToutput      (Fundamental)--L??--94%---------------------------------
```

Paratools-Zitations-Parsing
http://paracite.eprints.org

A database of templates of the form

'_AUTHORS_ (_YEAR_). _TITLE_.
PUBLICATION, _VOLUME_(_ISSUE_):_PAGES_'

each _XXX_ is associated with a regular expression

- Ex.: _YEAR_ ➔ ([[:digit:]]{4})

2 weighting factors

- reliability: how “syntactically fixed“ is a regular expression?
  - Ex.: _URL_ > _TITLE_
- concreteness = number of fixed symbols
  - Ex.: '_AUTHORS_, _PUBLICATION_, in press' > '_AUTHORS_, _PUBLICATION_'

Templates are matched against the reference.

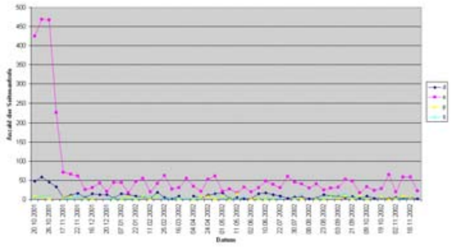
Choose the template with the highest *reliability*, or (if these are equal) with the highest *concreteness*.

Outlook 1: Diversity
(or: Web space and real-life spaces)

Which diagnosis is that?

49

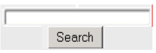
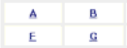
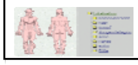
Request frequency for a specific diagnosis in the investigated eHealth portal, depending on time and request language



[Yihune, 2003]

Hypotheses: search preferences

50

Search option	Characteristics	Presumably preferred by
Search engine 	■ little context ■ fast information access ■ no hierarchies	■ Low context ■ Low Uncertainty Avoidance ■ Short-Term oriented ■ Low Power Distance
Alphabetically organized links 	■ large hierarchies	■ High Power Distance
Content-organized links 	■ highest amount of (context) information ■ more time-consuming information access ■ large hierarchies	■ High context ■ High Uncertainty Avoidance ■ Long-Term oriented ■ High Power Distance

(Kralisch & Berendt, *Proc. IWIPS 2004)*

Search behaviour: sample results

51

UA – Uncertainty Avoidance
Cont – Context Specifity
LTO – Long-Term Orientation
PD – Power Distance

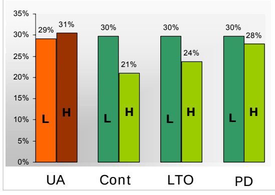
1. **Which search options were used?**

🟩 Expected results

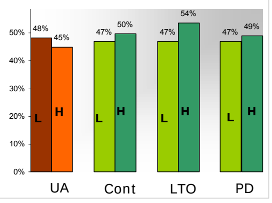
🟥 Unexpected results

- all results significant ($p<0.001$)

search engine:

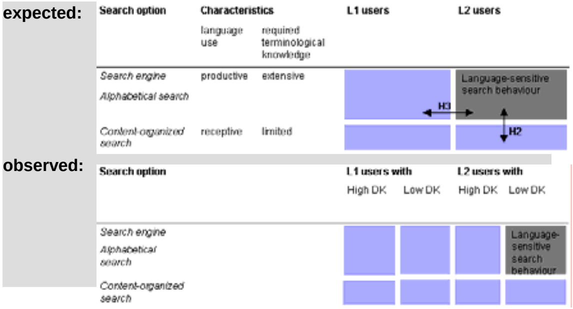


content-organized links:



Interactions between language and domain knowledge

52



expected:

Search option	Characteristics		L1 users	L2 users
	language use	required terminological knowledge		
Search engine *Alphabetical search*	productive	extensive		Language-sensitive search behaviour
Content-organized search	receptive	limited		

observed:

Search option	L1 users with		L2 users with	
	High DK	Low DK	High DK	Low DK
Search engine *Alphabetical search*				Language-sensitive search behaviour
Content-organized search				

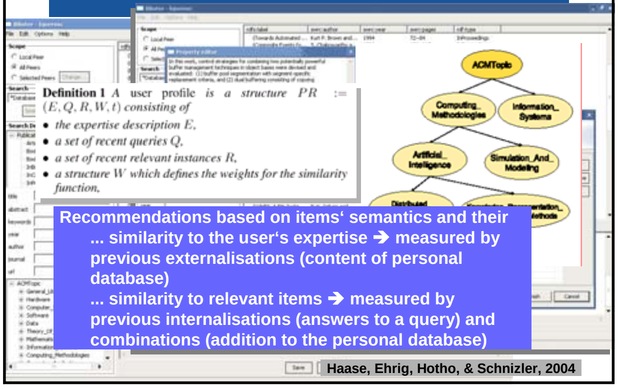
Kralisch & Berendt, *New Review of Hypermedia and Multimedia*, in press)

53

Outlook 2: Community

bibster.semanticweb.org

54

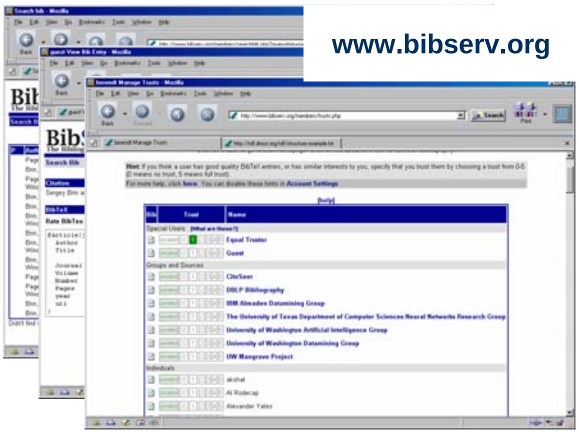


Definition 1 *A user profile is a structure* $PR := (E, Q, R, W, t)$ *consisting of*

- *the expertise description* E,
- *a set of recent queries* Q,
- *a set of recent relevant instances* R,
- *a structure* W *which defines the weights for the similarity function.*

Recommendations based on items' semantics and their
... similarity to the user's expertise ➔ measured by previous externalisations (content of personal database)
... similarity to relevant items ➔ measured by previous internalisations (answers to a query) and combinations (addition to the personal database)

Haase, Ehrig, Hotho, & Schnizler, 2004



www.bibserv.org

bsnendt Manage Trusts - Mozilla

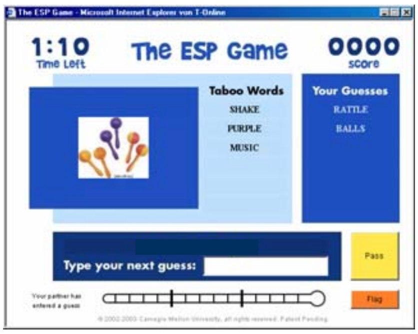
http://www.bibserv.org/members/trusts.php

Hint: If you think a user has good quality BibTeX entries, or has similar interests to you, specify that you trust them by choosing a trust from 0-5 (0 means no trust, 5 means full trust).
For more help, click **here**. You can disable these hints in **Account Settings**.

[help]

Bib	Trust	Name
Special Users [What are these?]		
		Equal Truster
		Guest
Groups and Sources		
		CiteSeer
		DBLP Bibliography
		IBM Almaden Datamining Group
		The University of Texas Department of Computer Sciences Neural Networks Research Group
		University of Washington Artificial Intelligence Group
		University of Washington Datamining Group
		UW Mangrove Project
Individuals		
		akshat
		Al Rodecap
		Alexander Yates

Outlook 3: Fun!



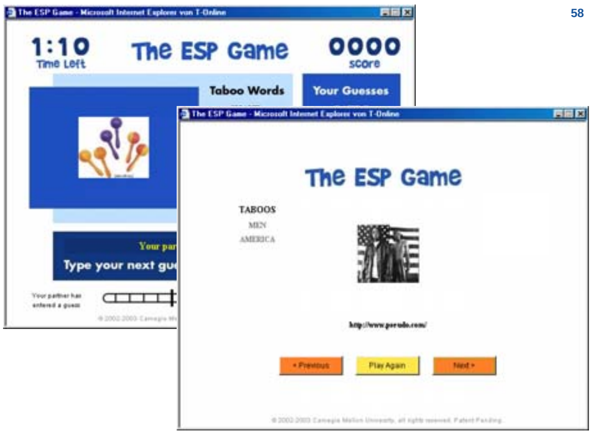
The ESP Game - Microsoft Internet Explorer von T-Online

Taboo Words
SHAKE
PURPLE
MUSIC

Your Guesses
RATTLE
BALLS

Type your next guess:

© 2002-2003 Carnegie Mellon University, all rights reserved. Patent Pending.



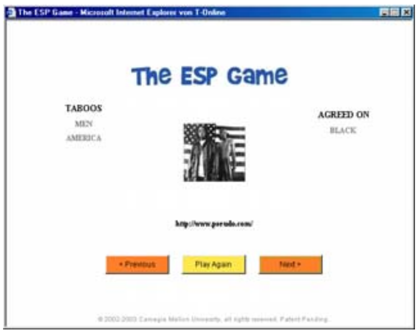
The ESP Game - Microsoft Internet Explorer von T-Online

The ESP Game - Microsoft Internet Explorer von T-Online

TABOOS
MEN
AMERICA

http://www.perudo.com/

© 2002-2003 Carnegie Mellon University, all rights reserved. Patent Pending.



The ESP Game - Microsoft Internet Explorer von T-Online

TABOOS
MEN
AMERICA

AGREED ON
BLACK

http://www.perudo.com/

© 2002-2003 Carnegie Mellon University, all rights reserved. Patent Pending.

?
Outlook 4: Share the initiative – automated Web service search and composition for SWM
?

Thank you for your attention!