Unsupervised Relation Extraction from Web Documents

Towards
Interactive
Dynamic
Information
Extraction

G. Neumann, N. Reithinger, H. Hemsen,
K. Eichler, M. Löckelt, A. Horbach
LT lab, DFKI, Saarbrücken, Germany

EFRE

DFKI

An IE system can be seen as an interface between a template and text fragments

- Offline/static IE:
  - Relevant information in form of templates (entities & relations) and relevant corpus is given to the IE system
- Approaches:
  - Manually implemented rule-based IE systems
  - Automatically induced data-driven IE systems

Current IE systems are too inflexible

- An IE system needs an exact definition of a template
  - it must be known in advance how information is structured for a certain application AND paraphrased in documents
  - usually one IE system handles one template type
- IE systems are realized by means of a set of sub-components making use of simple and static information flow
- IE systems have no way of adapting themselves to the dynamics in information changes, e.g., to adapt the template structure and mapping rules

We need IE systems which emerge on specific user request

- User and IE system must interact
  - Different users have different interest/knowledge
  - User (goal-directed), IE system (data-oriented)
  - Dynamics of user request and document space
- IE system must be adaptive
  - Open (no fixed template structures, multiple templates)
  - Preemptive (predict all possible interesting template structures)
  - On-line (do on-demand and user-driven/personalized)

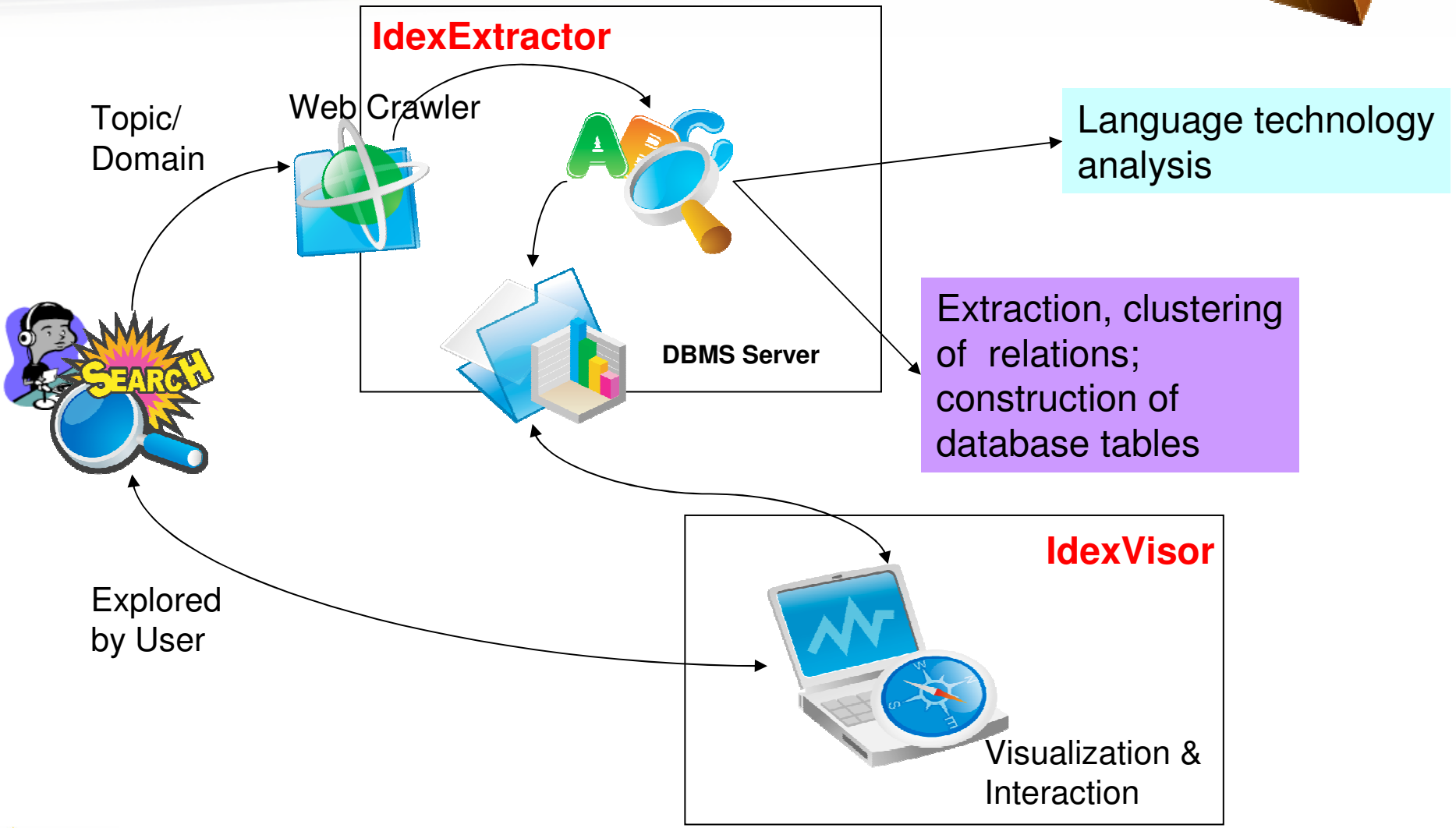
Interactive Dynamic Information Extraction

- Scientific motivation
  - Dynamic recognition, extraction, visualization of knowledge from the Web
  - Research & Development in the field: on-demand IE
- Economic motivation
  - Unveiling of relevant hidden relation, e.g., as for risk analysis
  - Dynamic configuration of IE systems
  - Developers/users can exploit knowledge together with the system

Technology Roadmap

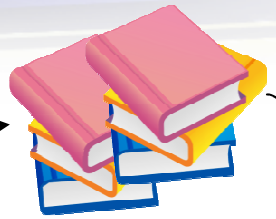
- Innovative combination of
  - On-demand IE
  - Unsupervised machine learning
  - Visualization
  - Interactive search

IDEX – Interactive Dynamic IE System



IDEX: Language technology components

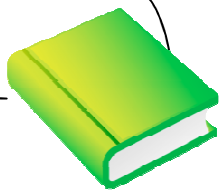
Topic document-crawling



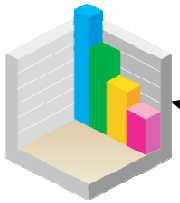
Conversion

- Sentence boundaries
- NE recognition
- Coreference

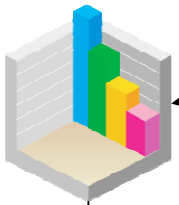
Language filtering



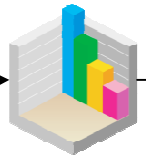
Dependency analysis



Relation Extraction



Filtering of relations



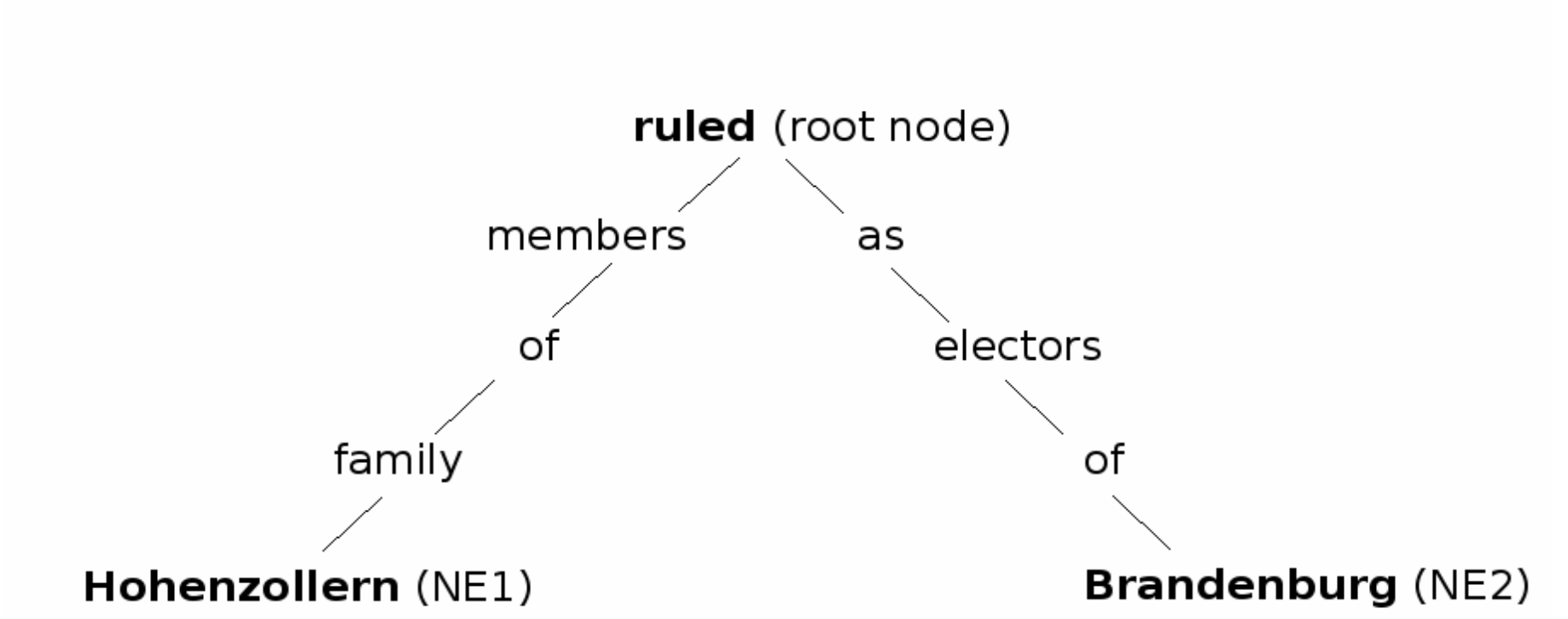
Clustering

Relation extraction

- We define a sentence to be of potential interest if it has at least two NEs
  - skeletons (simplified dependency trees) are extracted,
  - i.e., for each NE pair the common root element depending on the dependency parse tree is identified

- Information based on dependency types is collected
  - verb + its subject(s), object(s), preposition(s) with arguments and auxiliary verb(s)
  - At least subject or object has to be an NE
  - Relations with only one argument are filtered out

Skeleton for the sentence: „Subsequent members of the Hohenzollern family ruled until 1918 in Berlin, first as electors of Brandenburg“



Relation clustering

- Match of verb infinitives? Or in same synonym set?
- Token overlap between subjects/objects?
- Comparison of auxiliary verbs, prepositions and preposition arguments?
- Number of NEs that match?

$\Rightarrow$ results weighted and if defined threshold exceeded put into same cluster

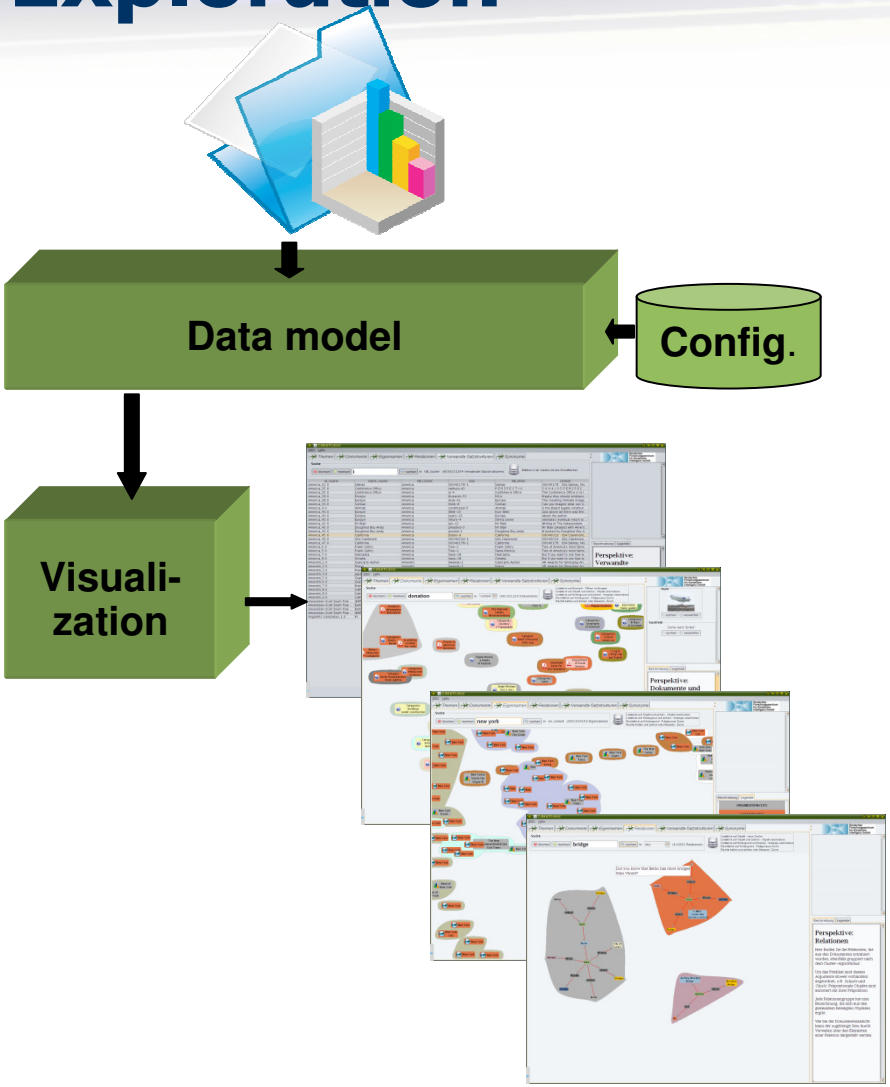
IdexExtractor: Experiments and results

- Test corpus: „Berlin central station“
  - 1068 web pages
  - 55255 sentences
  - 10773 relation instances
  - 306 clusters (two or more instances) – 81 clusters with identical instances
    - 121 consistent (i.e., all instances in the cluster express a similar relation)
    - 35 partly consistent (i.e., more than half of the instances in the cluster express a similar relation)
    - 69 not consistent

Types of clusters

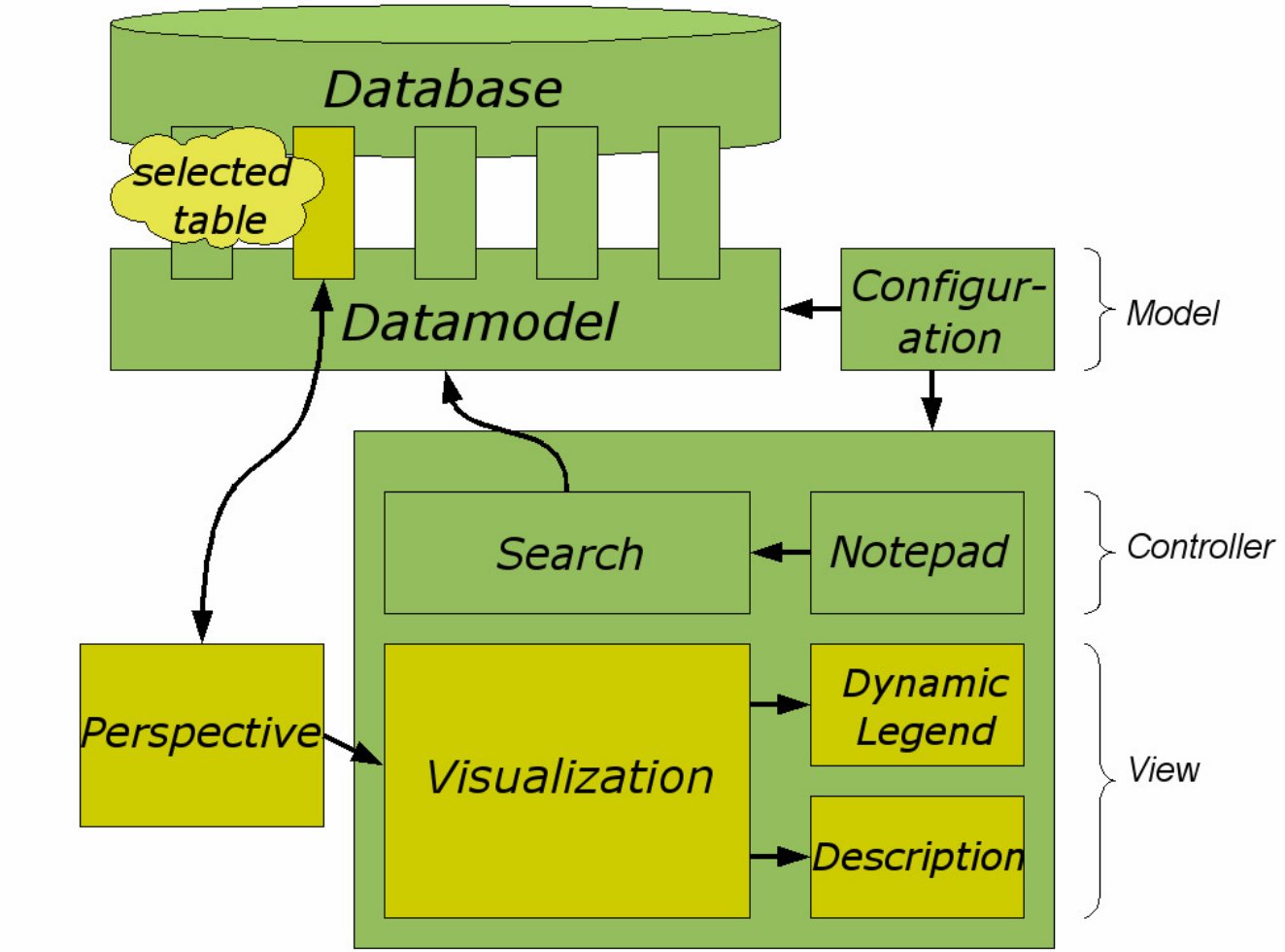
- Relation paraphrases (18 clusters)
  - *accused(Mr Moore, Disney, In letter)*
  - *accused(Micheal Moore, Walt Disney Company)*
- Different instances of same pattern (76 clusters)
  - *operates(Delta, flights, from New York)*
  - *offers(Lufthansa, flights, from DC)*
- Relations about same topic (27 clusters)
  - *rejected(Mr Blair, pressure, from Labour MPs)*
  - *reiterated(Mr Blair, ideas, in speech, on March)*
  - *created(Mr Blair, doctrine)*

IdexVisor: Interactive Information Exploration



- Source
  - the extracted tables
- Goal/function
  - Search
  - interaction
  - exploration
- Features
  - separation of the data model from the database
  - interactions and visualizations fitted to the data

IdexVisor Architecture



Evaluation of IdexVisor

- Qualitative evaluation: 7 users, average age 33 years, 4 male, 3 female
- 4 corpus-related questions had to be solved via interaction with the system

Question	Possible Answers	⌀
How did you like the introduction ?	1=useless/5=helpful	4,42
How useful is the system?	1=useless/5=helpful	4,14
Do you think you might use such a system in your daily work?	1=no/5=yes	4,14
How do you judge the computed information?	1=useless/5=very informative	3,71
How do you judge the speed of the system?	1=very slow/5=very fast	4,42
How do you judge the usability of the system?	1=very laborious/5=very comfortable	3,42
Is the graphical representation of the results useful?	1=totally not/5=very useful	3,57
Is the graphical representation appealing?	1=totally not/5=very appealing	3,71
Is the navigation useful in the system ?	1=totally not/5=very useful	3,57
Is the navigation intuitive in the system?	1=totally not/5=very intuitive	3,57
Did you have any problems using the system?	1=heavy/5=no difficulties	4,28

Results of the Evaluation of IdexVisor

- All users were able to answer the questions
- The search speed was judged generally as „fast“
- Difficulties with the interaction: more complex interface than current search engines („Google“ syndrome)
  - Parts of the user interface were overlooked or actually not recognized
  - Difficulties to use different perspectives and to coordinate the results of different perspectives.

Future work

- IdexVisor
  - More simple/consistent presentation
    - trade-off between intuitiveness and features
  - Integration of dialog functionality
    - QA-cycles, but strongly driven from system perspective
- IdexExtractor
  - Focused web crawling
    - More complex queries, credibility
  - Speed
    - Online clustering, parallelism

IDEX Results

- Innovative combination of:
  - Unsupervised IE
  - Visualization
  - Interactive search
- Evaluation shows feasibility
  - Dynamic IE on web sites
  - Positive assessment of interactive information exploration
- Only few other similar projects
  - Etzioni (U. Washington), Sekine (U. New York)
  - IDEX combines on-demand IE with complex visual interaction
  - However, there is a trend towards unsupervised IE, cf. upcoming conferences, e.g., ECAI, Coling, WWW

Thank you for your attention