DaimlerChrysler

Profile: NLP in Information Retrieval

Daniel Sonntag RIC/AM

MPI -SB 28.5.2004

Agenda

- Multimedia Data Retrieval (Image/Text)
- NLP components in Question Answering/Schema Mapping
- Multiword term indexes
- Connection Wordnet<->Framenet

MM Databases: Introduction

- Multimedia databases have to store numeric, image, video, audio, text, graphical, temporal, relational and categorical data.
- Attention in many application areas:
  - Medical information systems
  - Geographic information systems
  - E-commerce
  - Digital libraries
- We will draw attention on special purpose database files within the DC corporate group with regard to data mining databases.

Current architecture vs. ORDM requirements: Complete Data Model

Object-Relational Data Model	DB2	Oracle	SQL-Server	Informix
New additional basis data types for new application domains	●	●	●	●
Copies of basic data types with new type names	●	●	●	●
Data types for external data.	●	●	-	●
Basic types variants (i.e. structured types)	●	●	-	-[1]
Collection types (List, Set, Multiset)	-	●[2]	-	●[3]
Reference types that objects can be referenced	●	●	-	-
Type hierarchies of objects	●	●	-	-
Type hierarchies of tables	●	-	-	●
Typed tables for typing complete data entries.	●	●	-	●
User defined routines (functions) (UDR(F)) that can be registered in the DBMS and be used as operators for data types.	●	●	●	●

Current architecture vs. ORDM requirements

- Unstructured Image Data
  - different kinds like paintings, drawings, photographic pics, satellite images, architectural, facial ...
  - digital file formats like WAV, AU, GIF, JPG, MPEG with different compression and quality rates.
- Unstructured Text Data
  - string of arbitrary size, in linguistic terms containing words, sentences, paragraphs as logical units
  - in DB own internal representation format, converted from RTF, PDF, PS ...

Theoretical evaluation

- Comparison of object-relational and **multimedia text features**

Query expansion operator	DB2	Oracle	SQL Server	Informix
Fuzzy term matches to include words that are spelled similarly to the query term.	●	●	-	●
Taxonomy search to include more specific or more general terms.	●	●[1]	-	-
Proximity search to test whether two words are close to each other, i.e. near positions.	●	●	●	●
Related term matches to expand the query by related terms defined in a thesaurus.	●	●	●	●
Term replacement to replace a term in a query with a preferred term defined in a thesaurus. Could also be used for synonym searches.	●	●	●	●

DaimlerChrysler

Theoretical evaluation

- Comparison of object-relational and **multimedia text features**

Linguistic query expansion operator	DB2	Oracle	SQL Server	Informix
Stem match to search for terms that have the same linguistic stem as the query term, e.g. runs->run, running ->run	●	●	●	-
Translation match to search for translated terms in a different language, defined by a thesaurus.	-	●	-	-
Soundex match to find phonetically similar words computed by the soundex algorithm.	●	●	●	-
Text summarization Automatic summarization of documents based on key words and related sentences/paragraph (pseudo-semantic processing).	-	●	-	-
Theme search/extraction Automatic extraction of the text theme that can then be searched for.	-	●	-	-
Decomposition match to decompose complex words into their stems.	●	●[1]	-	-

Extraction methods

DAIMLERCHRYSLER

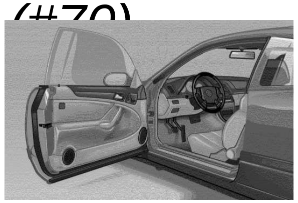
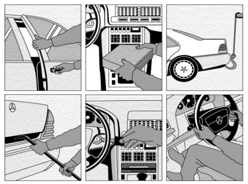
	Concept level	Feature extraction method	DB2	Oracle	Discovir
Color global	1/2	Global color histogram	●	●	●
	1/2	Global average color	●	-	●
	2	Color moment	-	-	●
	2	Color coherence vector	-	-	●
Color local	3	Local color histogram	-	●	●
	3	Local average color	●	-	-
Texture global	2	Homogeneity	-	-	●
	2	Entropy	-	-	●
	2	Probability	-	-	●
	2	inverse differential moment	-	-	●
	2	differential moment	-	-	●
	2	Contrast	●	-	-
	2	Edge direction	●	-	-
	2	Granularity/fineness	●	●	●
	2	Edge frequency	-	-	●
	2	Length of primitives/texture	-	-	●
Texture local	3	Locality of texture	-	●	-
Shape global	2	Geometric moment	-	-	●
	2	Eccentricity	-	-	●
	2	Invariant moment	-	-	●
	2	Legendre moment	-	-	●
	2	Zernike moment	-	-	●
	2	Edge direction histogram	-	-	●
	2	Color-based segmentation	-	●	-
Shape local	3/4	Locality of Shape	-	●	●

Practical evaluation: Case study

- *DC Media service (#50)*


- *Cardetect (#30)*

- *DC internal car image data (#70)*

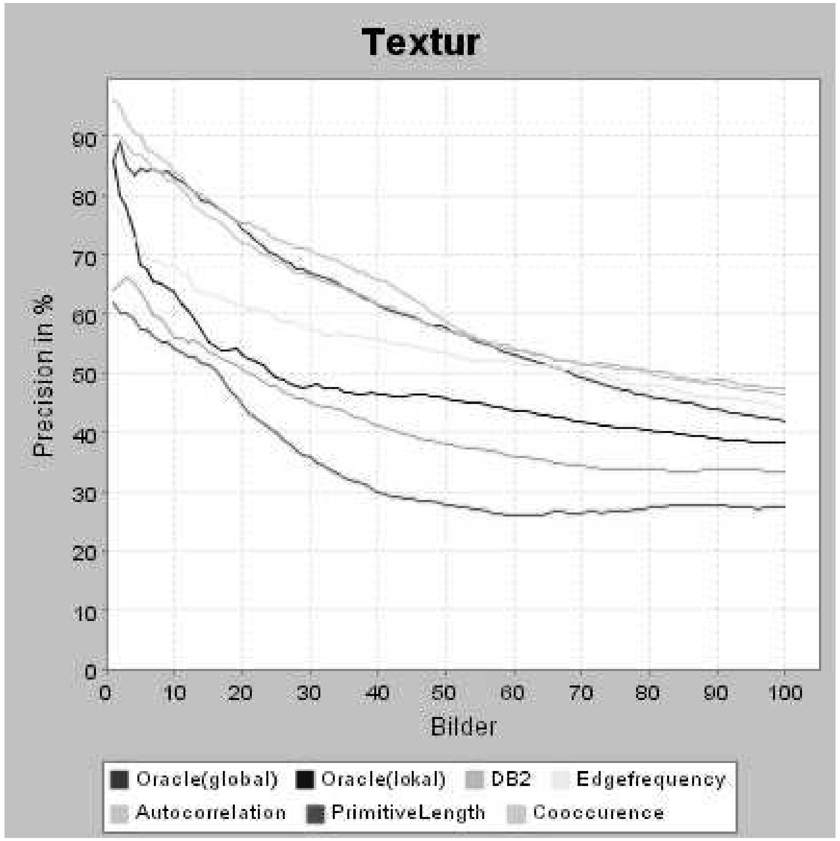



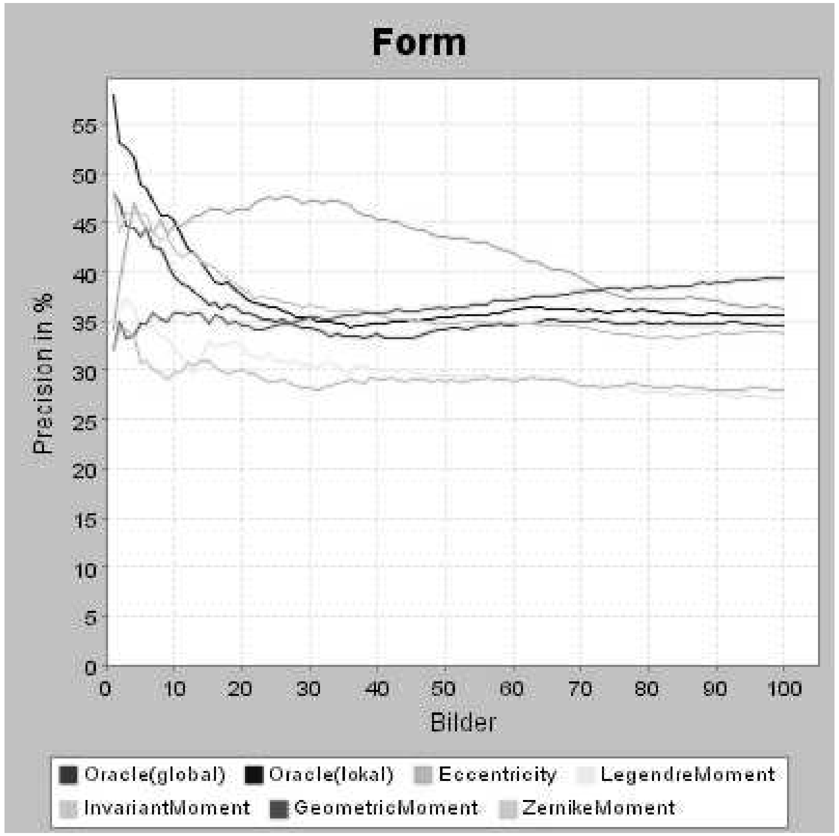
- Rear cars (#400)


Practical evaluation: Case study

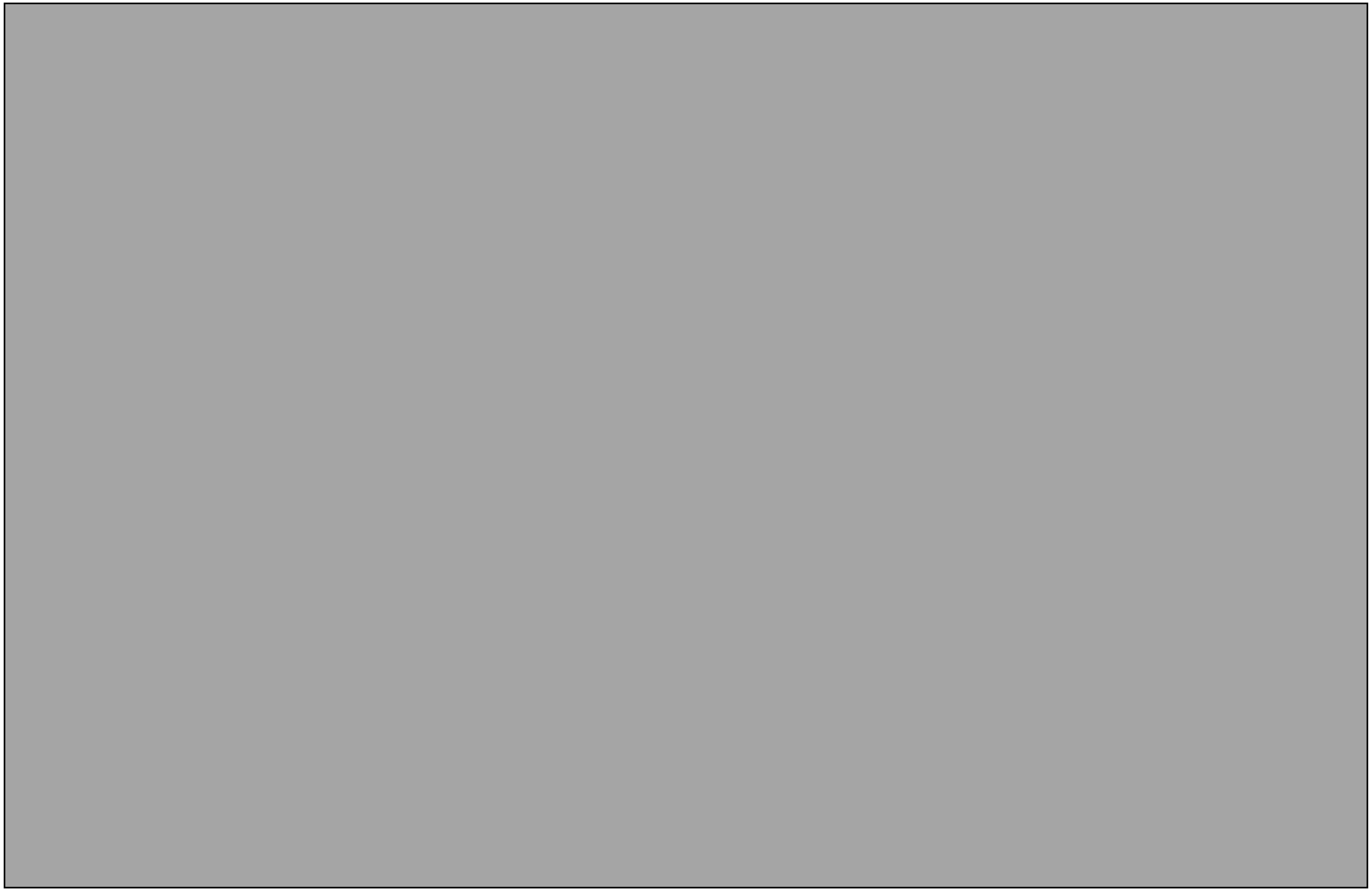
- Evaluation measures (#8):
  - *Precision:* Precision measures the proportion of documents in the result set that are actually relevant.
  - *Recall:* Recall measures the proportion of all the relevant documents in the collection that are in the result set.
  - *Effectiveness:* This measure takes the relative order of retrieved documents into account.
  - Accuracy, Reciprocal Rank, Interpolated Average Precision, F-Measure, Fallout.

Practical evaluation: Case study





Feature concepts



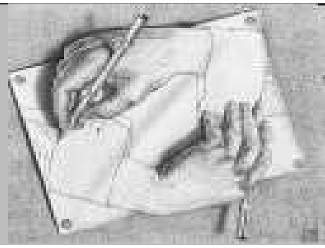
Challenges for MM databases

- Special data types for media types
- **Feature extraction and selection**
  - extractable vs. perceptible vs. interpretable (semantic gap)
- Query system and language
- Similarity search
- Realtime retrieval

Proposed DCX conceptual architecture

DaimlerChrysler

Further Reading Material for MM

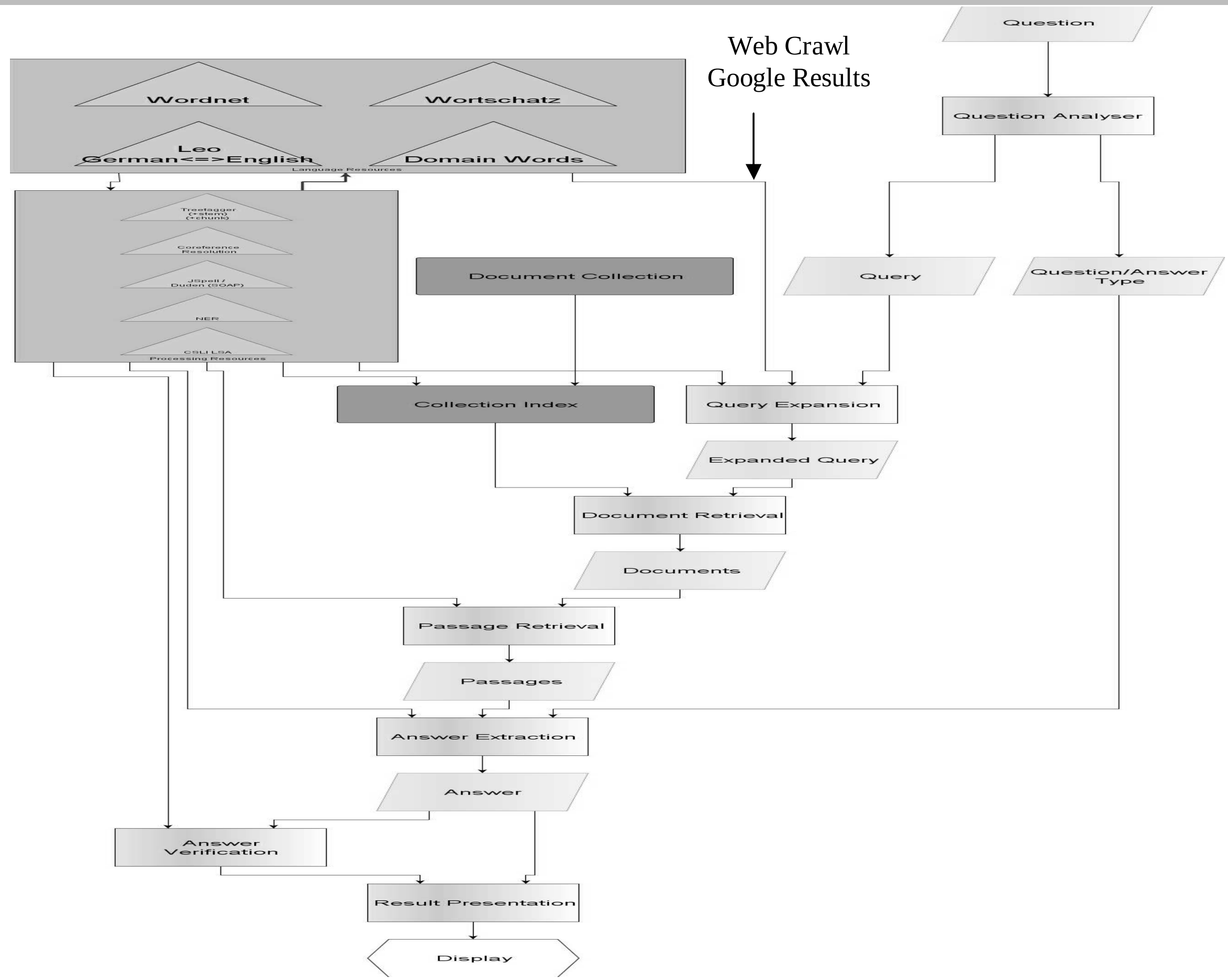


- MultimediaDatabases, State-of-the-art report, Daniel Sonntag, RIC/AM, (2004).
- Analyse kommerzieller ORDB-Bild-Retrieval-Systeme, Diplomarbeit, Doreen Pittner, (2004).
- Image Databases, Search and Retrieval of Digital Imagery, edited by Vittorio Castelli and Lawrence D. Bergman (2003)
- Ingo Schmitt, Retrieval in Multimedia-Datenbanksystemen, Institut für Technische und Betriebliche Informationssysteme, Otto-von-Guericke-Universität Magdeburg, to appear (2004).

Question Answering

LRs: Wordnet, Wortschatz,
Leo,Domain Dics

PRs: Tagger, Chunker,
Duden (Soap)
NER, LSA



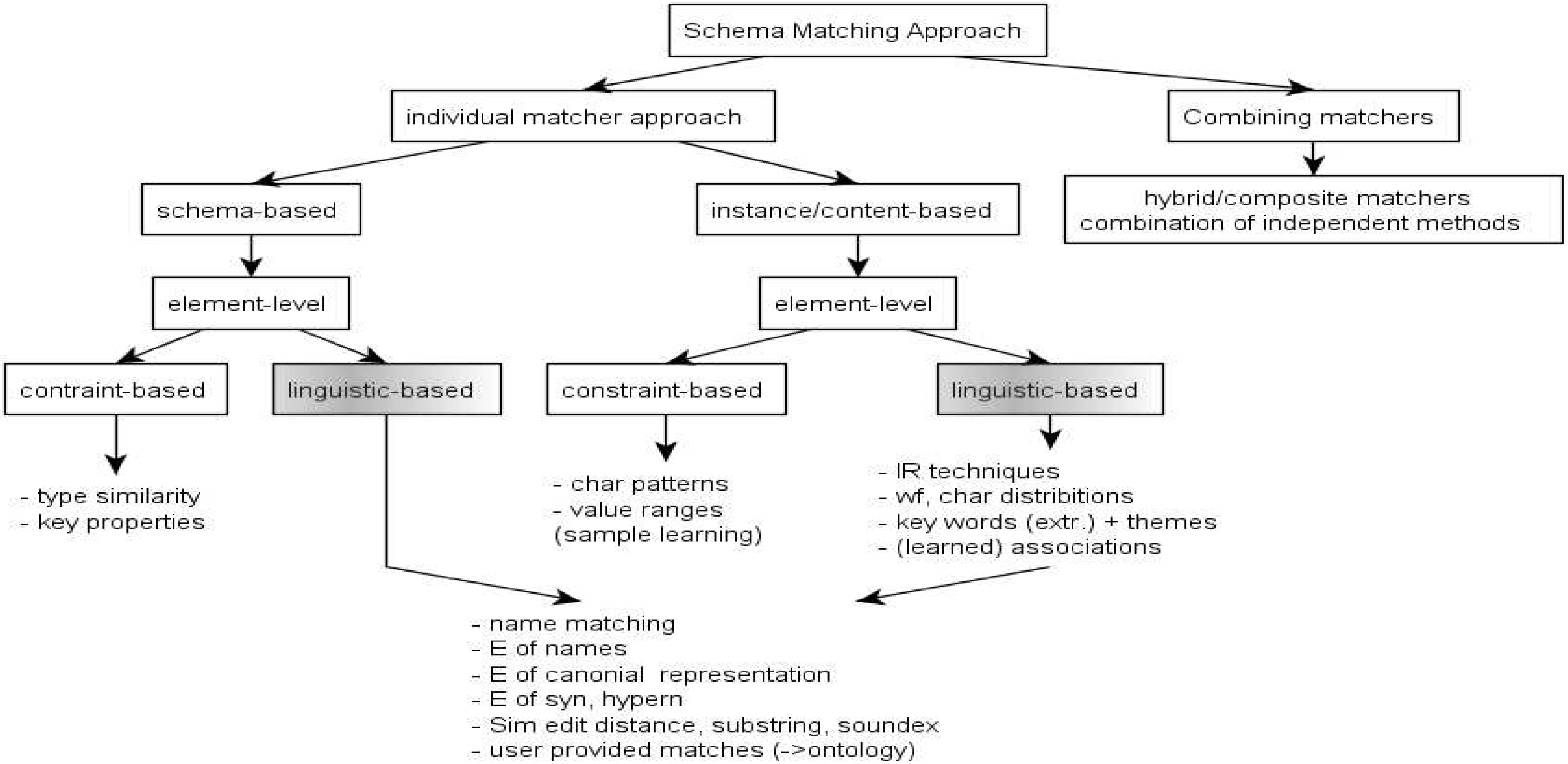
Schema Matching Problems

- External schemas (beside complexity)
  - unknown synonyms
  - unknown hyponyms
  - foreign-language data material
  - cryptic schemata (# attr < n)

-> false positives/false negatives

- label-based, instance-based, and structure-based mapping
- Match cardinality: 1:n, n:1
  - Parsing rules, (De)composition

Schema Matching Approaches [RB01] [FN04]



DSTAT: pattern matching

- a -> abcdefghijklmnopqrstuvwxyzABCDEFGHIJKLMNOPQRSTUVWXYZäöüÄÖÜß
- d -> 0123456789
- s -> !§$%&()={[]}?*+-_.:,'#~@"/\
- o -> any other character

- **ULM_CMP2_S_ADDR_ORG.CONFLICT_ID**
- Type: VARCHAR2, Size: 15
- Patterns: d1 -> 237624 (99.974%)
- d1s1a3s1d4 -> 11 (0.005%)
- d1s1a1d1a1s1d3 -> 8 (0.003%)
- d1s1a1d1a1s1d4 -> 5 (0.002%)
- d1s1a2d1s1d4 -> 5 (0.002%)
- d1s1a3s1d3 -> 5 (0.002%)
- ...

Multiword term indexes

- Similarity function: s(x,y) := s(f(x), f(y));
  - s(x,y): class-based approaches -> thesaurus-based similarity
  - s(x,y): distributional approaches -> clustering, KNN

  word co-occurrence patterns -> class co-occurrence patterns ?

  - f(x), f(y): add dimensions, new/replacing document (content) descriptors

    -> **add MWU/MWE, but which ones? coverage, coding**

    -> **formalism, DB, textual XML, SGML, FS, typed FL?**

- MWU/MWE Induction (*Computational Terminology, TE, RL*):
  - use knowledge-free methods -> subtype collocation finders
  - Central question: Which collocations are suitable MWU/MWE ?

Multiword term indexes

- Collocation finding:
  - *knock (at) door, make up*, *Buenos Aires*, *prime minister*, (*to turn off the power)*,
  - Problems for German: decomposition (syntactic)
  - Problems for English: verb-particle constructions
    - *knock off, tell off, cook off*
  - **segmentation-driven**: collocation = byproduct of segmenting stream of symbols.
  - **Word-based knowledge-driven**: linguistic patterns: N *de* N (regex), linguistic phenomena: NPs
  - **Word-based probabilistic**: word combination probabilities

Multiword term indexes: Prob. MWU Finder/collocation finder [SJ01]

Frequeny-based
vs.
Information-based

METHOD	FORMULA
Frequency (Guiliano, 1964)	f_{XY}
Selectional Association (Resnik, 1996)	$\frac{P_{X\vert Y} * MI_{XY}}{\sum_Z Pr_{Z\vert Y} * MI_{ZY}}$
Log-likelihood (Dunning, 1993; Daille, 1996)	$-2\log \frac{[P_X P_Y P_{\overline{X}} P_{\overline{Y}}]^{f_Y}}{[P_{XY} P_{\overline{XY}}]^{f_{XY}} [P_{X\overline{Y}} P_{\overline{X}Y}]^{f_{\overline{XY}}}}$
Student's t-Score (Church and Hanks, 1990)	$\frac{f_{XY} - \xi_{XY}}{\sqrt{f_{XY}(1-(f_{XY}/N))}}$

METHOD	FORMULA
Pointwise Mutual Information (MI) (Fano, 1961; Church and Hanks, 1990)	$\log_2 (P_{XY} / P_X P_Y)$
Symmetric Conditional Probability (Ferreira and Pereira, 1999)	$P_{XY}^2 / P_X P_Y$
Chi-squared (χ^2) (Church and Gale, 1991)	$\sum_{\substack{i \in \{X, \overline{X}\} \\ j \in \{Y, \overline{Y}\}}} \frac{(f_{ij} - \xi_{ij})^2}{\xi_{ij}}$
Z-Score (Smadja, 1993; Fontenelle, et al., 1994)	$\frac{f_{XY} - \xi_{XY}}{\sqrt{\xi_{XY}(1-(\xi_{XY}/N))}}$
Dice Formula (Dice, 1945)	$2 f_{XY} / (f_X + f_Y)$

Multiword term indexes

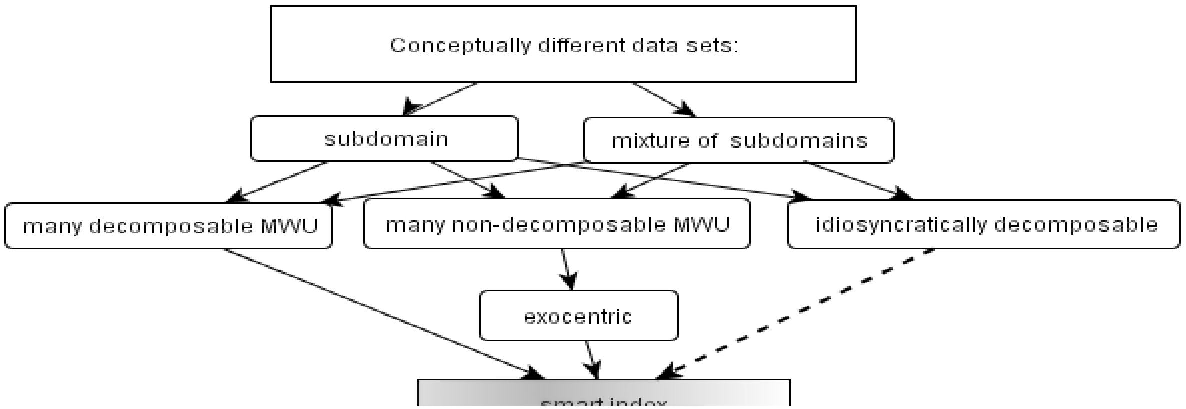
- Which collocations are suitable MWU/MWE = Which collocations need a definition ?
  - Linguist's answer (Sproat):
    - *Simply expanding the dictionary to encompass every word one is ever likely to encounter is wrong: it fails to take advantage of regularities.*
- MWUs are ...
  - non-substitutable: *compact disc* vs. # *densely-packed disk*
  - AND/OR non-compositional: **m(cd) != ded(m(c) , m(d))**
  - AND/OR non-modifiable: # *disk that is compact*

Multiword term indexes

- Idea: Extraction + Recognition in once [SA04];

  (instead of: coll. Finder + hyponymy testing (LSA): **s(f(m,h), f([h|m]))**

- Two goals:
  - Technological expr. are fairly compositional: *filter, oil filter* (-> **ontology***) vs.* Good MWU are non-compositional *(->* **terminology***)*



Multiword term indexes by Mining Sequential Patterns

- Determinative compound (endocentric)

6: INTERVAL@NN => **SERVICE**@NN Supp = 0.82, Conf = 72.64, Cov = 1.13, Lift = 5.52

8: 7500@CD MILE@NN => **SERVICE**@NN Supp = 0.43, Conf = 90.7, Cov = 0.48, Lift = 6.9

14: 22500@CD MILE@NN => **SERVICE**@NN Supp = 0.13, Conf = 86.35, Cov = 0.15, Lift = 6.57

16: 6000@CD MILE@NN => **SERVICE**@NN Supp = 0.46, Conf = 71.55, Cov = 0.64, Lift = 5.44

24: 7x500@CD MILE@NN => **SERVICE**@NN Supp = 0.26, Conf = 89.73, Cov = 0.29, Lift = 6.82

30: 3750@CD MILE@NN => **SERVICE**@NN Supp = 0.54, Conf = 99.02, Cov = 0.54, Lift = 7.53

40: 30000@CD MILE@NN => **SERVICE**@NN Supp = 0.41, Conf = 91.72, Cov = 0.45, Lift = 6.98

- (Possessive) compound (exocentric)

80: **BLOWER**@NN => WIRING@NN Supp = 0.18, Conf = 52.01, Cov = 0.35, Lift = 195.89

81: **BLOWER**@NN => MOTOR@NN Supp = 0.25, Conf = 70.07, Cov = 0.35, Lift = 137.45

82: **BLOWER**@NN MOTOR@NN => WIRING@NN Supp = 0.17, Conf = 67.64, Cov = 0.25, Lift = 254.75

Multiword term indexes

- Extensions: Expansion of collocation (sets)
  - -> strongly associated words
  - Exploiting linguistic theory for finding associated words
    - Systematic Polysemy
    - Metonymie

-> **Frame Elements**

WordNet and FrameNet

- WordNet problem:
  - no syntagmatic relations, e.g. “tennis problem”.
- FrameNet help:
  - *Documents the range of semantic and syntactic combinatory possibilities (valences) of each word in each sense.*
  - Valence descriptions:
    - **Frame Elements (e.g. Patient)**
    - Grammatical Functions (e.g. Object)
    - Phrase Type
  - Connection: Wordform Type ^= Wordform (Framenet)
  - Connection ?: Synsets OR Lexical Unit ^= Frame Elements (Framenet)

Exploiting FrameNet

- Information Retrieval <-> Information Extraction
  - Automatic Frame Element Labeling is questionable!
    - too difficult in conception, only example sentences

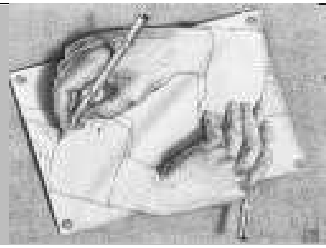
-> Frame Labeling

-> Exploit frame relations

-> Exploit documented element associations -> thesaurus-based

similarity

Further Reading Material



- [FN04] Felix Naumann, Schema Mapping Tutorial, HU Berlin/DC Ulm 2004.
- [RB01] Erhard Rahm and Philip Bernstein, A survey of approaches to automatic schema matching, VLDB Journal 10(4), 2001.
- [SJ01] Patrick Schone and Daniel Jurafsky, Is Knowledge-free induction of Multiword Unit Dictionary Headwords a Solved Problem?
- [SA04] Daniel Sonntag and Markus Ackermann, Multiword Expression Learning for Automatic Classification, to appear 2004.
- [TB02] Timothy Baldwin et al., An Empirical Model of Multiword Expression Decomposability