AG5 Oberseminar

MAX-PLANCK-GESELLSCHAFT

Automatic ontology extraction for document classification

Student:

Natalia Kozlova

Supervisors:

Prof. Gerhard Weikum

Martin Theobald

Overview

- Introduction
- Framework description
- Ontology creation
- Results
- Conclusions and future work

Problem description

MAX-PLANCK-GESELLSCHAFT

Classification using direct matching

- Lexical matching is loose in terms in capturing meaning
- Synonymy, polysemy and word usage pattern problems
- Nothing to do with unknown words

Ontology can help

- Matching by sense, fighting synonymy, polysemy & …
- Stronger concepts, multi-word concepts allowed
- Possible to infer meaning of unknown concept
- No precision loss with fewer training docs

Why not WordNet?

- WordNet usually offers much more then necessary
- WordNet is very broad, no topic specificity
- No weights

We want to get:

- More topic-specific ontology using complex concepts
  - can we generate reusable corpora-independent heuristics?
- Taxonomies from chosen strongly correlated parts of ontology
  - from small sets provided by user
- More precise document classification in the end

Framework description

- Take & study corpora
- Create Ontology
  - Choose concepts
  - Extract relations
  - Distinguish relations
  - Weight relations
  - Prune ontology
  - do .. while (satisfied)
- Plug in classifier
- Classify new documents
  - Use structural features

Hierarchy example:

- Fine arts
- Mathematical and natural sciences
  - Astronomy
  - Biology
  - Computer science
    - Databases
    - Programming
    - Software engineering
  - Chemistry
  - …
- …

Overview

- Introduction
- Problem description
- Ontology creation
  - Corpora description
  - Concepts extraction
  - Relations extraction
  - Ontology pruning
- Results
- Conclusions and future work

Wikipedia summary

- Contains about 350000 articles, content is very broad; created by many authors
- Internal markup is documented
- Wiki links contain titles of target document and possible "anchor"
  - [[America | United States]]; [[United States]]
- Constructions considered
  - [[Paris]], [[Paris, Tennessee]], [[Paris (god)]]
- Considered structural elements as
  - sections' headings; tables;
  - enumerations; lists;
  - elements in-doc positions and in-section positions;

Framework in general

- Extract concepts
- Parse Wiki documents again with the sliding window
  - Store terms, compute frequencies;
  - Marked known concepts;
- Apply heuristics to reveal relations between concepts
  - Edge types - Hypernyms (i.e. broader sense), hyponyms (i.e. kind of), meronyms (i.e. part of), see also, similar to …
- Quantify relations
  - Edge weights – probability of co-occurrence
- Apply heuristics to “clean” concept’s set

Concepts extraction

- Article titles are concepts. We distinguish:
  - S-Terms. Come from document titles. The most confident.
  - A-Terms. Related to S- ones and share the sense with S-terms. For a given S-term, A-terms are extracted from anchors of the links in documents that refer to S-term.
  - NT-Terms. Appear in the document text as links, but these links have no target documents.
  - E-Terms. Emphasized terms. The additional source for meaningful phrase terms.
- Processing rules form a “policy”

Relations extraction: heuristics

- Synonyms:
  - redirection, same target doc ID
  - anchors
- Hypernyms (and hyponyms)
  - concepts, appeared in parenthesis to the concept near
  - concepts, appeared after comma to the one before
  - hierarchically related concepts with both sides existed
- Unspecified
  - section names
  - links inside doc (to some extent, usually unspecified)
  - artificial concepts for “empty” links added
  - hierarchically related concepts, others
- See also, similar to
  - Found in appropriate sections by names (flexible)

Relations extraction: examples

- Structure analyses was applied on docs with
  - words like “classification” in the anchors
  - words like “topic” in the titles
  - words like “type” in the anchors and titles
  - words like “list of”
  - words with parenthesis
- Example
- Title: Canidae (level 1)
  - Genus *Canis* (level 2)
    - **Wolf**, *Canis lupus* (level 3)
      - **Domestic** Dog, *Canis lupus familiaris* (level 4…)
      - **Dingo**, *Canis lupus dingo*
      - …many other subspecies
    - **Red Wolf**, *Canis rufus* (level 3)
    - **Coyote**, *Canis latrans*
    - **Golden Jackal**, *Canis aureus* ..

Doc: **Automobile**

… Car classification…

Doc: car class-tion

Microcar

Sub-compact

Sedan …

Doc: **microcar**

A **microcar** is a particularly and unusually small automobile.

Pruning relations

- The similarity measure is given by
  - $P(B|A) = P(A \cap B) / P(A)$
- Imagine the number of possible interconnections between 400 00 documents
- The resulting ontologies contain some noise
- Different strategies of pruning:
  - Cut off results, produced by certain heuristics
  - Cut off results, where relationship is not “approved” by the certain level of IDF for target concept. The cut-off level can be chosen.
  - Cut off relations that are not “important” for current concept:
  - $Imp_{c\text{->}Cd} = \alpha\, IO(c,Cd) + \beta\, OO(c,Cd) + \gamma OI(c,Cd) + \sigma\, sim(c,Cd)$

Disambiguation & Mapping strategy

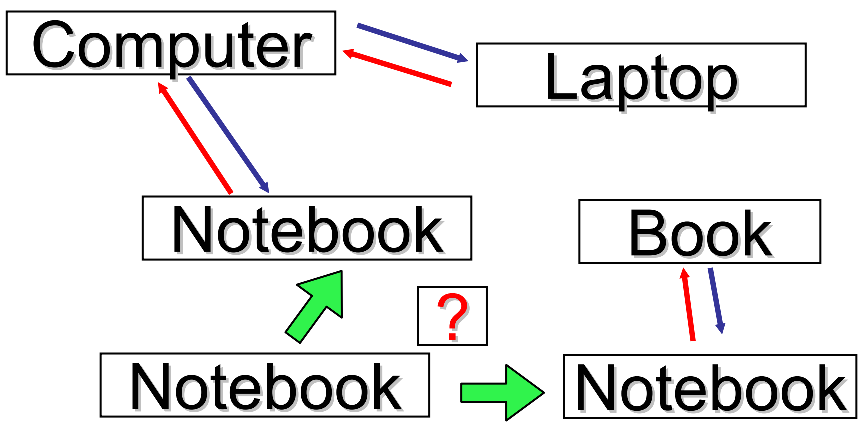
- Map tags to senses
  - Take tag word(-s) and get sets of **senses** for them from ontology
  - Compare tag context *t* and term context *s* using cosine measure (i.e.)
  - Map tag to sense with highest similarity in context

<computer>
<notebook>
<brand>Dell
<ram>512</ram>…

context(<tag>**)** =(text content (name, subordinate elements, their names))

context(term**)** =(hypernyms, hyponyms, meronyms, description)

$$s' = \arg\max_{s'}(sim(con(t), con(s')) \mid s' \in senses_{onto})$$



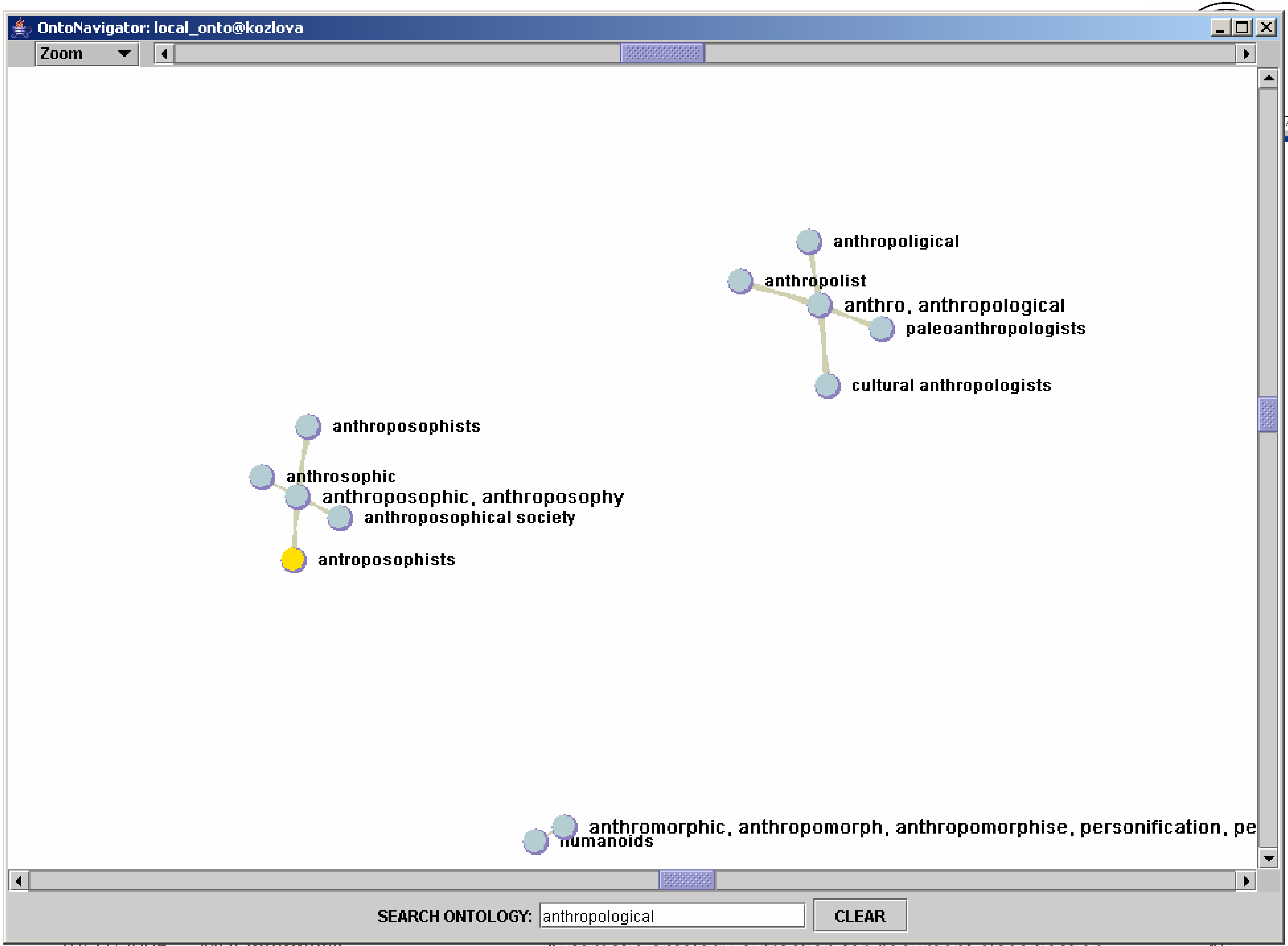
- Result: infer semantics from current context

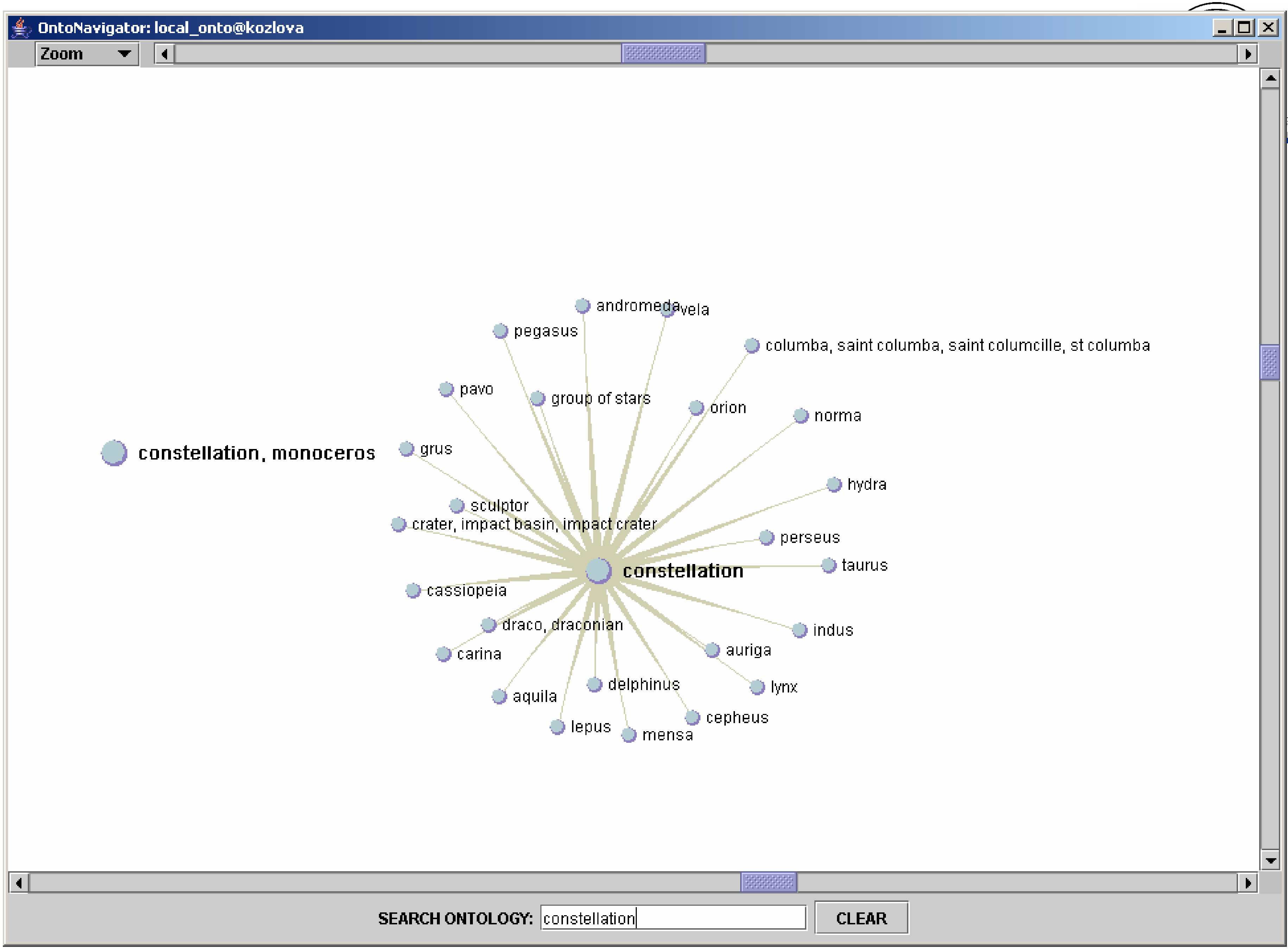
Overview

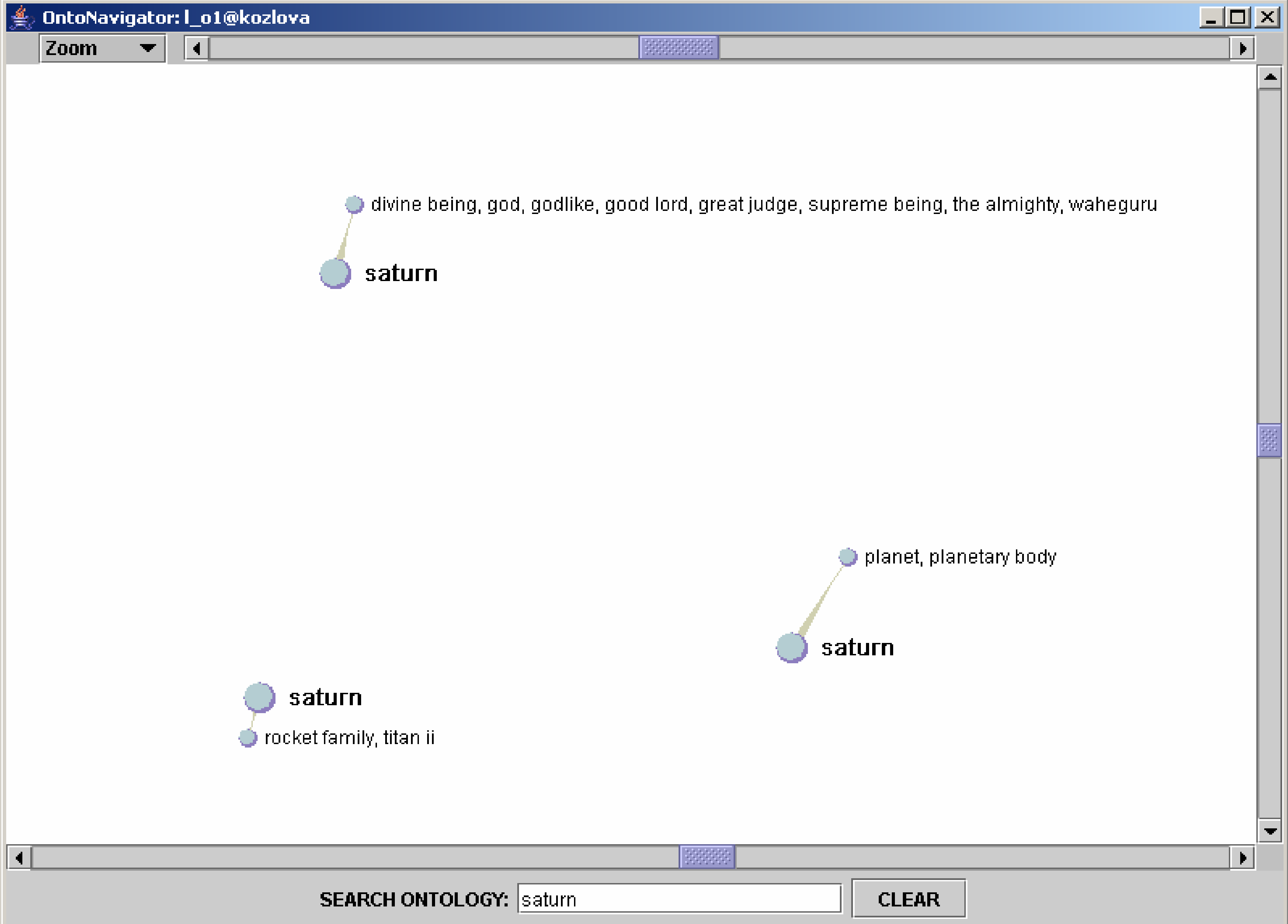
- Introduction
- Problem description
- Ontology creation
- Results
  - How it looks like
  - Experiments
- Conclusions and future work

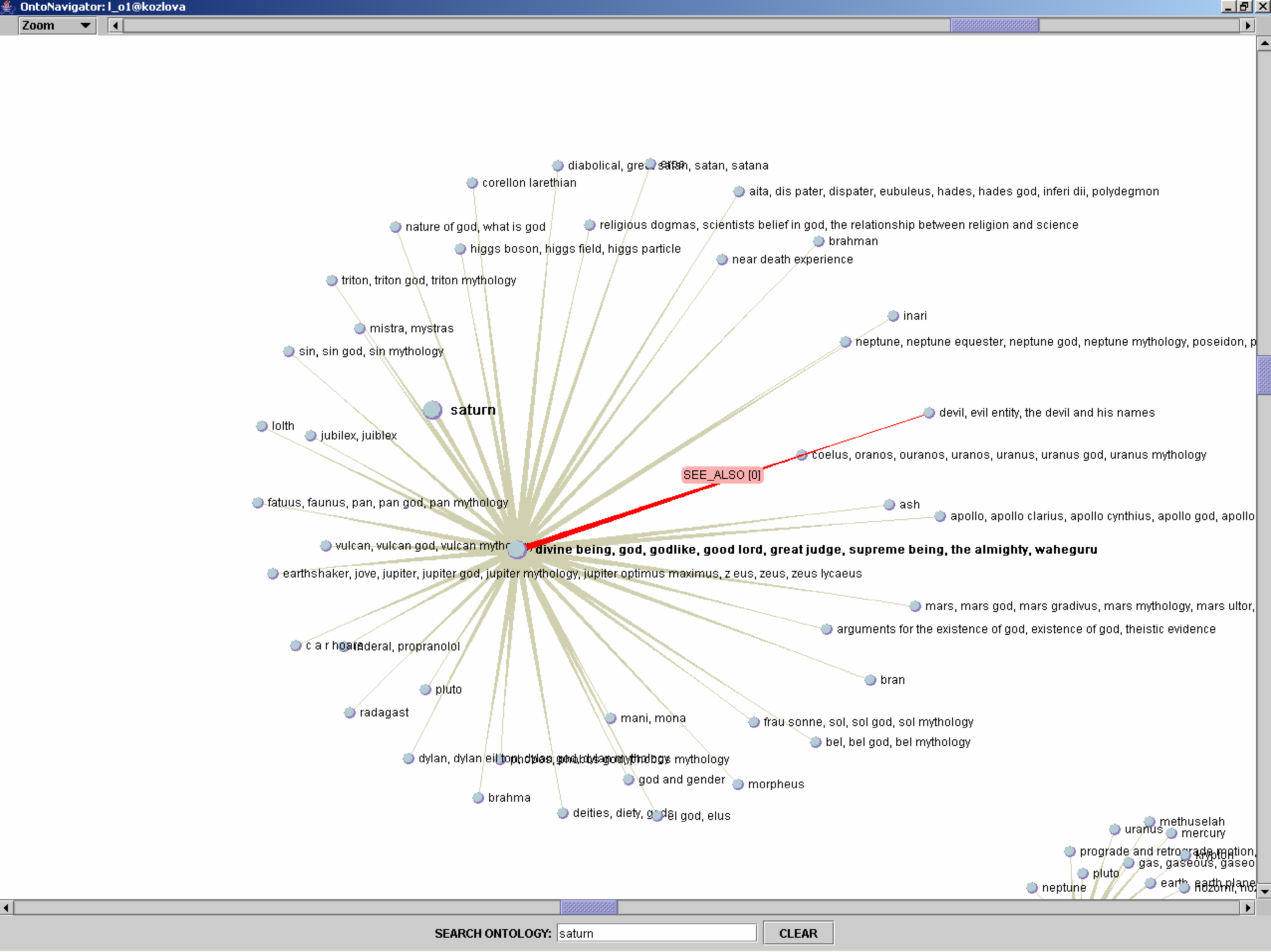
Some statistics

- *Complete set* of concepts has size 365 000, the *working set* has about 313 000
- Sliding window parsing the size of 4 was used
- For each sequence
  - match in unstemmed set, if no
  - match in stemmed set.
    - some terms have more than 1 match
- For each term all its positions stored
  - ~$29*10^6$ of terms found in ~440 000 docs
  - ~1 610 000 of distinct terms
  - Terms stored in stemmed form
- Number of relations
  - Strong ~ 70 000
  - Weak – can use up to ~1 500 000 directed


OntoNavigator: local_onto@kozlova
Zoom
anthropoligical
anthropolist
anthro, anthropological
paleoanthropologists
cultural anthropologists
anthroposophists
anthrosophic
anthroposophic, anthroposophy
anthroposophical society
antroposophists
anthromorphic, anthropomorph, anthropomorphise, personification, pe
humanoids
SEARCH ONTOLOGY:
anthropological
CLEAR


OntoNavigator: local_onto@kozlova
Zoom
constellation, monoceros
andromeda
vela
pegasus
columba, saint columba, saint columcille, st columba
pavo
group of stars
orion
norma
grus
hydra
sculptor
crater, impact basin, impact crater
perseus
taurus
constellation
cassiopeia
draco, draconian
indus
auriga
carina
delphinus
lynx
aquila
cepheus
lepus
mensa
SEARCH ONTOLOGY:
constellation
CLEAR


OntoNavigator: l_o1@kozlova
Zoom
divine being, god, godlike, good lord, great judge, supreme being, the almighty, waheguru
saturn
planet, planetary body
saturn
saturn
rocket family, titan ii
SEARCH ONTOLOGY:
saturn
CLEAR



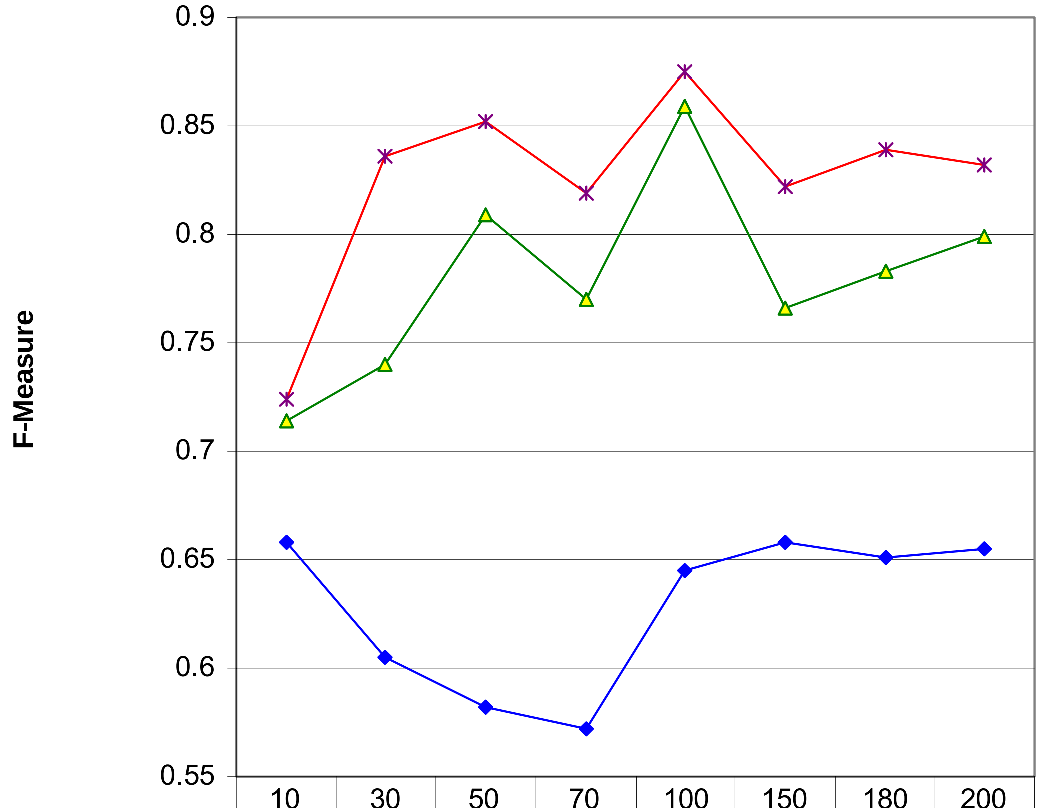
OntoNavigator: l_o1@kozlova
Zoom
corellon larethian
aita, dis pater, dispater, eubuleus, hades, hades god, inferi dii, polydegmon
nature of god, what is god
religious dogmas, scientists belief in god, the relationship between religion and science
brahman
higgs boson, higgs field, higgs particle
near death experience
triton, triton god, triton mythology
inari
mistra, mystras
neptune, neptune equester, neptune god, neptune mythology, poseidon, p
sin, sin god, sin mythology
saturn
devil, evil entity, the devil and his names
lolth
jubilex, juiblex
coelus, oranos, ouranos, uranos, uranus, uranus god, uranus mythology
SEE_ALSO [0]
fatuus, faunus, pan, pan god, pan mythology
ash
apollo, apollo clarius, apollo cynthius, apollo god, apollo
divine being, god, godlike, good lord, great judge, supreme being, the almighty, waheguru
earthshaker, jove, jupiter, jupiter god, jupiter mythology, jupiter optimus maximus, z eus, zeus, zeus lycaeus
mars, mars god, mars gradivus, mars mythology, mars ultor,
arguments for the existence of god, existence of god, theistic evidence
bran
pluto
radagast
mani, mona
frau sonne, sol, sol god, sol mythology
bel, bel god, bel mythology
god and gender
morpheus
brahma
methuselah
mercury
pluto
neptune
SEARCH ONTOLOGY:
saturn
CLEAR

- We created several ontologies of different size and constitution.
- We analyzed the performance of ontology-driven classification with regard to these ontologies.

Rule	**LO1**	**LO3**	**LO4**	**LO5**
G-HYP	14255	14255	14255	0
Ex-HYP	60507	14324	60507	0
S-HYP	8874	0	8874	0
SS-HYP	4613	0	4613	0
T-UNSPEC	0	0	0	254492
L-UNSPEC	0	0	0	326442
SIMTO	0	0	0	0
UNSPEC	124372	0	0	0
TOPLIST	55302	0	0	0

Experiments: Base line

- Reuters collection, classification with two classes: *Acq* and *Earn*
- 150 test documents, trainig set size varies from 10 to 200
- Naïve Bayes (NB) and SVM classification performed
- Different settings for ontology-driven classification

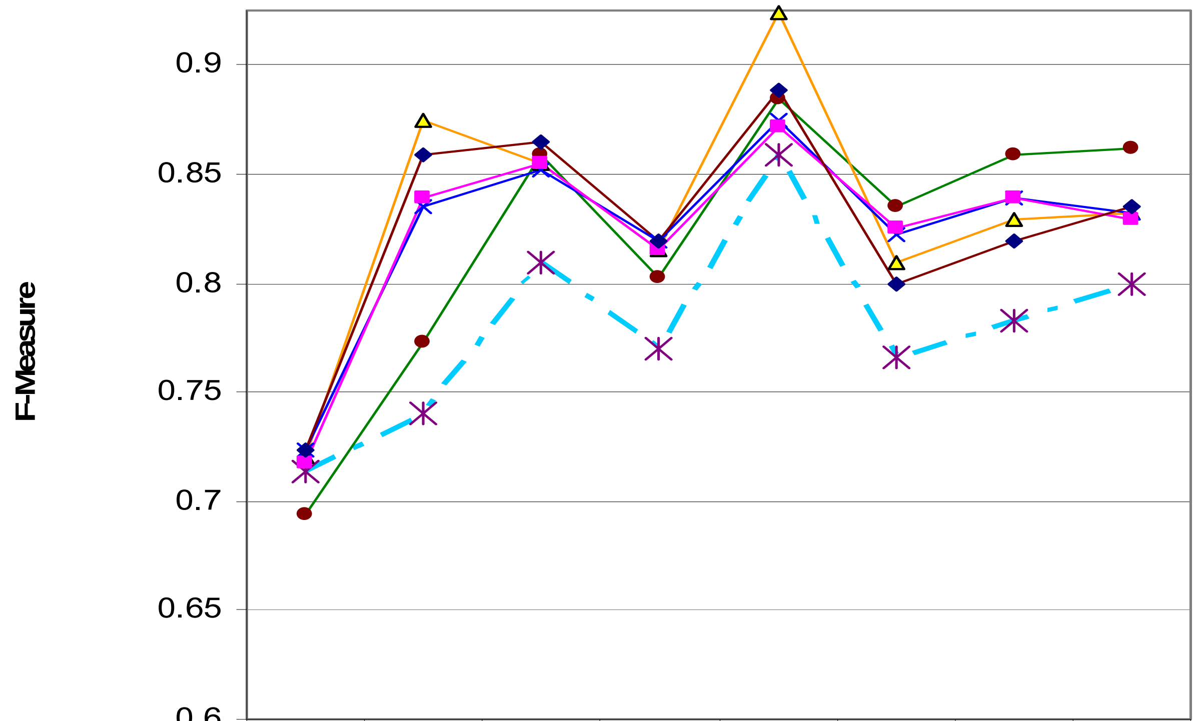


	10	30	50	70	100	150	180	200
Baseline NB	0.6579	0.6053	0.5822	0.5724	0.6447	0.6579	0.6513	0.6546
Baseline SVM	0.7138	0.7401	0.8092	0.7697	0.8586	0.7664	0.7829	0.7993
Baseline SVM+D+I	0.7237	0.8355	0.852	0.8191	0.875	0.8224	0.8388	0.8322

Training set size

Experiments: SVM+D

- SVM with ontology-driven terms disambiguation

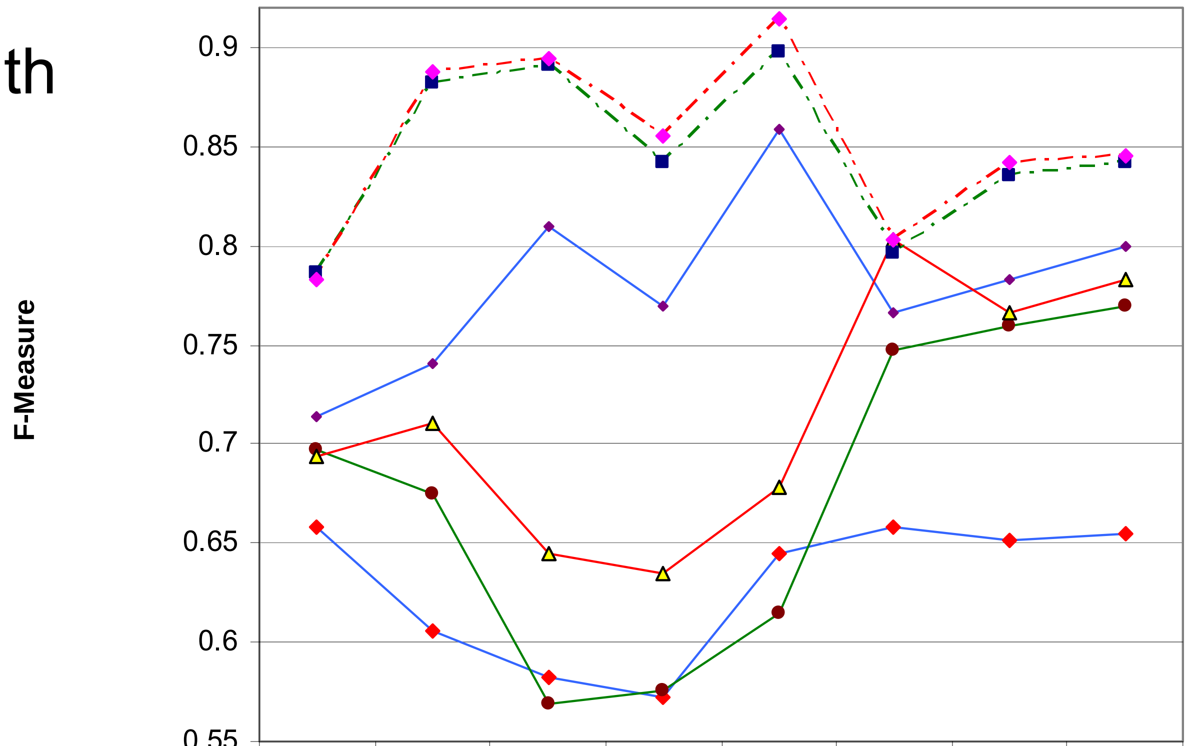


	10	30	50	70	100	150	180	200
Baseline SVM	0.7138	0.7401	0.8092	0.7697	0.8586	0.7664	0.7829	0.7993
WN SVM+D	0.6941	0.773	0.8586	0.8026	0.8849	0.8355	0.8586	0.8618
LO1 SVM+D	0.7204	0.875	0.8553	0.8158	0.9243	0.8092	0.8289	0.8322
LO3 SVM+D	0.7237	0.8355	0.852	0.8191	0.875	0.8224	0.8388	0.8322
LO4 SVM+D	0.7171	0.8388	0.8553	0.8158	0.8717	0.8257	0.8388	0.8289
LO5 SVM+D	0.7237	0.8586	0.8651	0.8191	0.8882	0.7993	0.8191	0.8355

Training set size

Experiments: SVM+P+D

- NB and SVM with ontology-driven phrases extraction

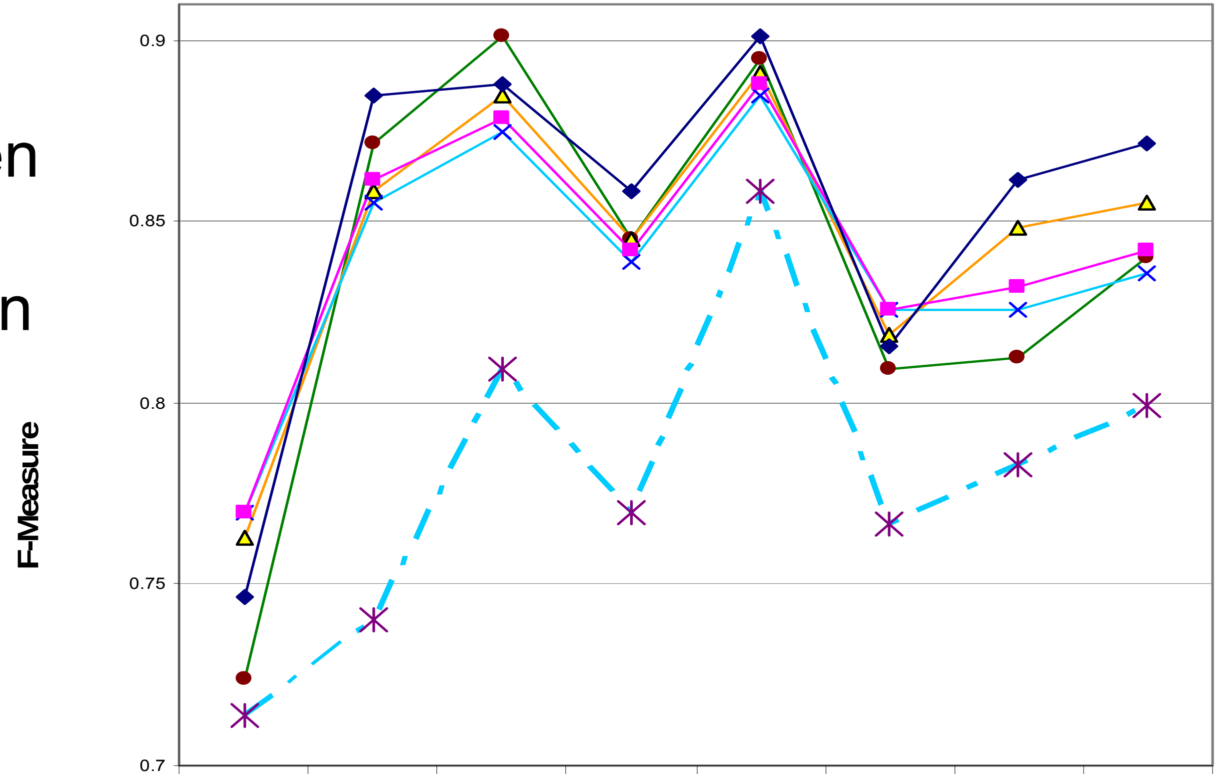


	10	30	50	70	100	150	180	200
Baseline NB	0.65789	0.60526	0.58224	0.57237	0.64474	0.65789	0.65132	0.65461
Baseline SVM	0.71382	0.74013	0.80921	0.76974	0.85855	0.76645	0.78289	0.79934
WN NB+P+D	0.69737	0.67434	0.56908	0.57566	0.61513	0.74671	0.75987	0.76974
LO1 NB+P+D	0.69408	0.71053	0.64474	0.63487	0.67763	0.80263	0.76645	0.78289
WN SVM+P+D	0.78618	0.88158	0.89145	0.84211	0.89803	0.79605	0.83553	0.84211
LO1 SVM+P+D	0.78289	0.88816	0.89474	0.85526	0.91447	0.80263	0.84211	0.84539

Training set size

Experiments: SVM+P+D

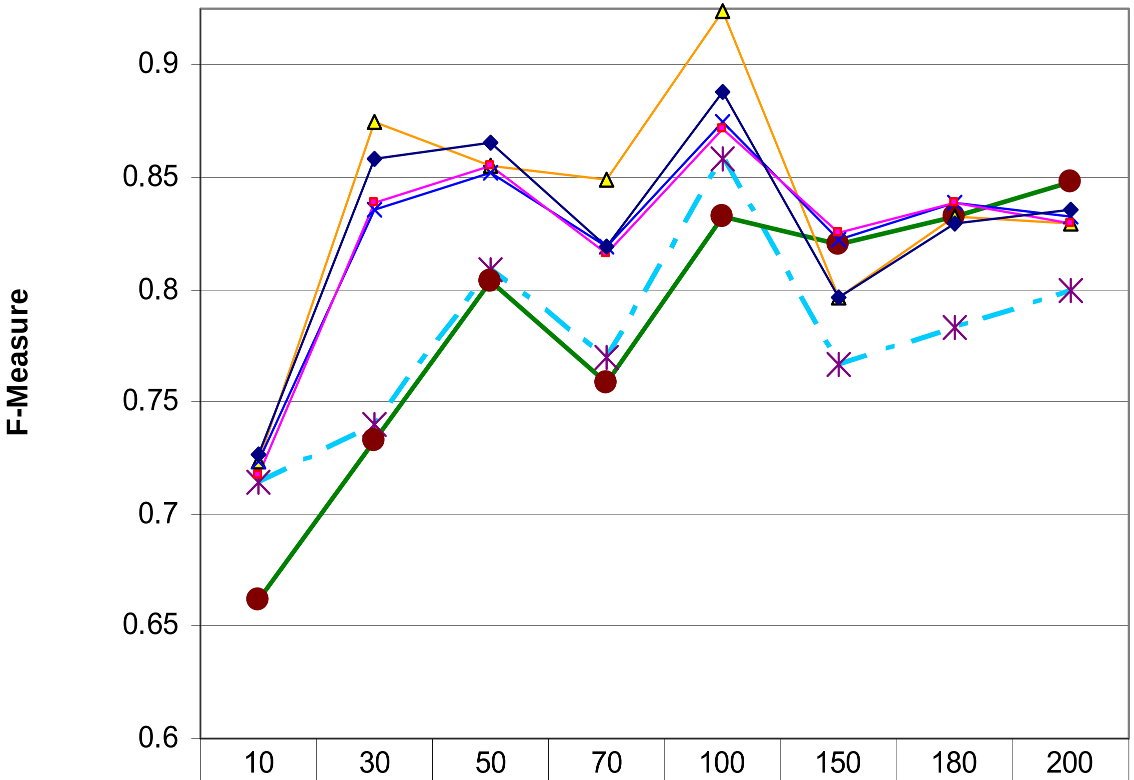
- SVM with ontology-driven terms disambiguation and phrases detection



Training set size	10	30	50	70	100	150	180	200
Baseline SVM	0.71382	0.74013	0.80921	0.76974	0.85855	0.76645	0.78289	0.79934
WN SVM+P+D	0.72368	0.87171	0.90132	0.84539	0.89474	0.80921	0.8125	0.8402
LO1 SVM+P+D	0.76316	0.85855	0.88487	0.84539	0.89145	0.81908	0.84868	0.85526
LO3 SVM+P+D	0.76974	0.85526	0.875	0.83882	0.88487	0.82566	0.82566	0.83553
LO4 SVM+P+D	0.76974	0.86184	0.87829	0.84211	0.88816	0.82566	0.83224	0.84211
LO5 SVM+P+D	0.74671	0.88487	0.88816	0.85855	0.90132	0.81579	0.86184	0.87171

Experiments: SVM+D+I

- SVM with ontology-driven terms disambiguation and incremental mapping

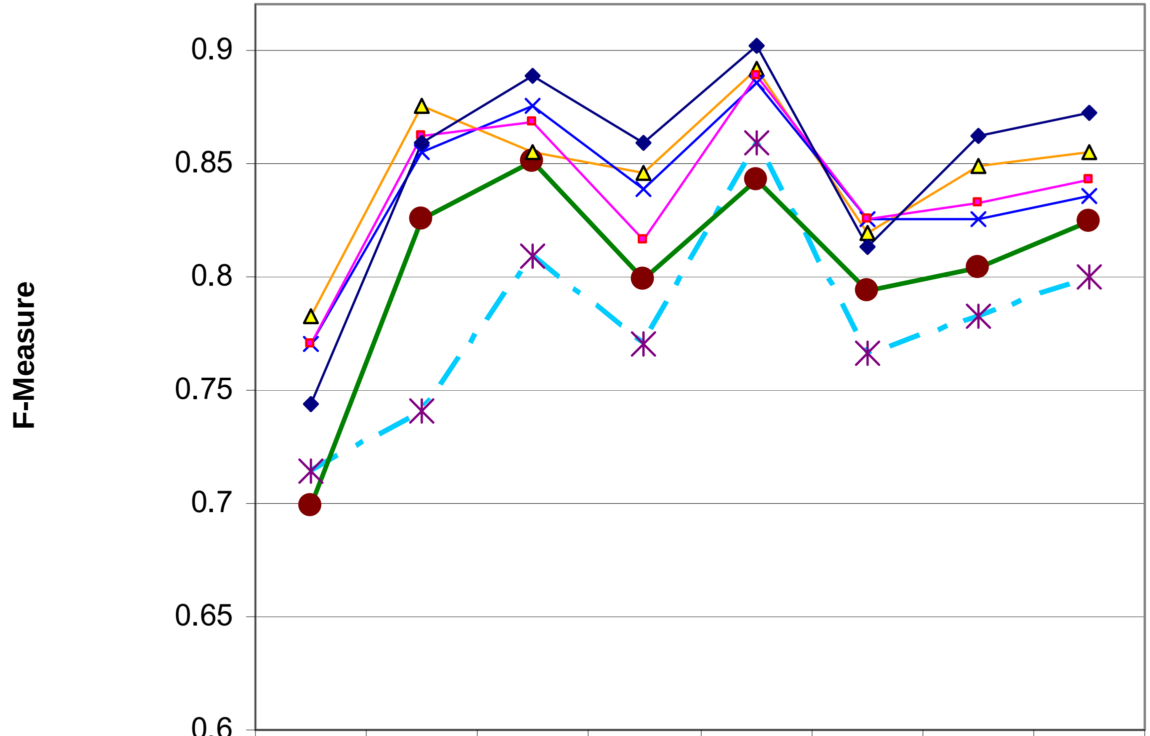


	10	30	50	70	100	150	180	200
Baseline SVM	0.71382	0.74013	0.80921	0.76974	0.85855	0.76645	0.78289	0.79934
WN SVM+D+I	0.66197	0.73298	0.80348	0.75826	0.83217	0.82017	0.83217	0.84808
LO1 SVM+D+I	0.72368	0.875	0.85526	0.84868	0.92434	0.79605	0.83224	0.82895
LO3 SVM+D+I	0.72368	0.83553	0.85197	0.81908	0.875	0.82237	0.83882	0.83224
LO4 SVM+D+I	0.71711	0.83882	0.85526	0.81579	0.87171	0.82566	0.83882	0.82895
LO5 SVM+D+I	0.72697	0.85855	0.86513	0.81908	0.88816	0.79605	0.82895	0.83553

Training set size

Experiments: SVM+P+D+I

- SVM with ontology-driven terms disambiguation, phrases detection and incremental mapping



	10	30	50	70	100	150	180	200
Baseline SVM	0.7138	0.7401	0.8092	0.7697	0.8586	0.7664	0.7829	0.7993
WN SVM+P+D+I	0.6991	0.8252	0.8507	0.7986	0.8427	0.7939	0.8034	0.8243
LO1 SVM+P+D+I	0.7829	0.875	0.8553	0.8454	0.8914	0.8191	0.8487	0.8553
LO3 SVM+P+D+I	0.7697	0.8553	0.875	0.8388	0.8849	0.8257	0.8257	0.8355
LO4 SVM+P+D+I	0.7697	0.8618	0.8684	0.8158	0.8882	0.8257	0.8322	0.8421
LO5 SVM+P+D+I	0.7434	0.8586	0.8882	0.8586	0.9013	0.8125	0.8618	0.8717

Training set size

Conclusion

Ontology is better for:

- Matching by sense, fighting synonyms, polysemy problems
- Complex concepts;
- Inferring meaning of unknown concept

Concept-based classification boosts classification results

- Synonyms detection
- Incremental mapping for unknown concepts

Advantages of the framework, suggested

- Provides a methodology for automatic ontology creation
- Can be easily enhanced with new rules

Future work

- More elaborated ontology-pruning techniques
- Statistical relation detection
- Possible further applications
  - Query disambiguation
  - Training on small, user-specific topic directories
  - Classification of heterogeneous data sources

The end

- Thank you for attention!
- Questions?