

Automatic Learning and Populating Ontologies from Web Data

Sergej Sizov
University of Koblenz
Information Systems and Semantic Web

Semantic Web and Machine Learning

What can Machine Learning do for the Semantic Web?

1. Learning Ontologies
(even if not fully automatic)
2. Annotation by Information
Extraction
3. Learning to map between
ontologies
4. Deep Annotation:
Reconciling databases and
ontologies
5. Duplicate recognition

What can the Semantic Web do for Machine Learning?

1. Lots and lots of tools to
describe and exchange data
for later use by Machine
Learning methods in a
canonical way!
2. Using ontological structures
to improve the Machine
Learning task
3. Provide background
knowledge to guide Machine
Learning

Outline

1. **Pattern-based annotation through knowledge on the Web**
2. **Restrictive distributed Machine Learning methods for automatic organization of data collections**
3. **Clustering Concept Hierarchies from thematically focused document collections**

PART 1:

Pattern-based annotation through knowledge on the Web

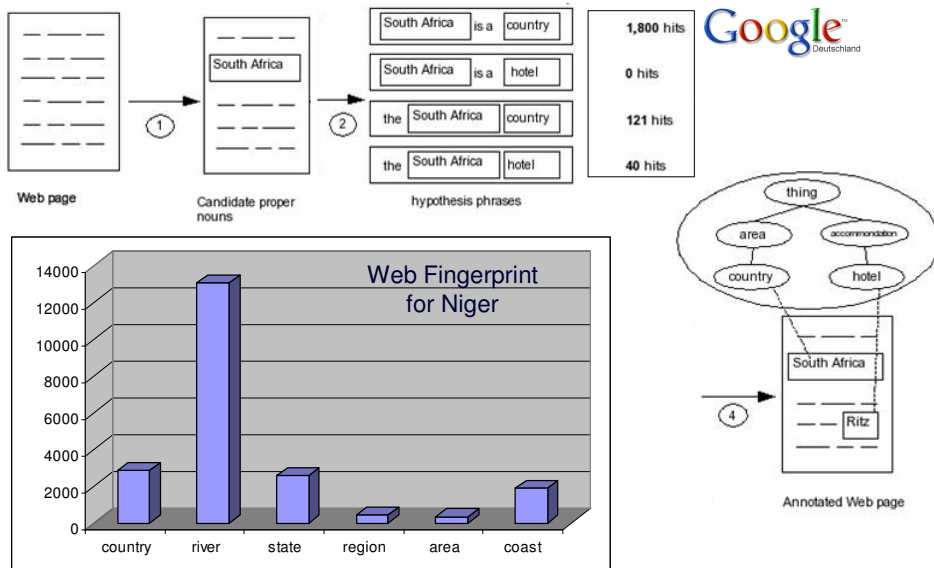
PANKOW: Pattern-based ANnotation through Knowledge On the Web

- Hearst: <INSTANCE> such as <CONCEPT>, ...
- DEFINITE1: the <INSTANCE> <CONCEPT>
- DEFINITE2: the <CONCEPT> <INSTANCE>
- APPPOSITION: <INSTANCE>, a <CONCEPT>
- COPULA: <INSTANCE> is a <CONCEPT>

Examples:

- the Niger country
 - the country Niger
 - Niger, a country in Africa
 - Niger is a country in Africa
- instanceOf(Niger, country)
 Or
 subconcept(Niger, country)

PANKOW Process



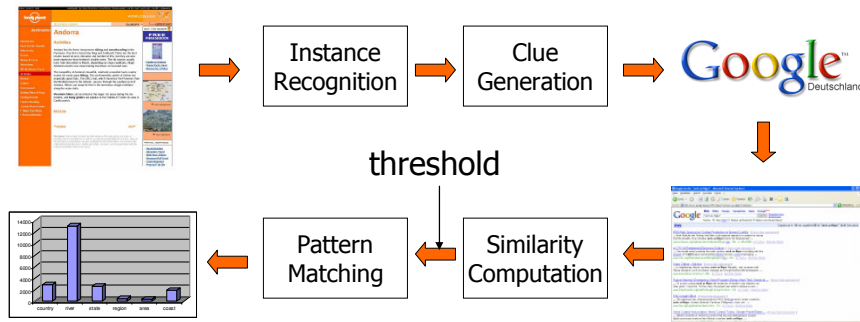
Drawbacks:

- did not take into account the *context* of the Web page to be annotated;
- search is brute-force, a number of patterns are generated is linear in the size of the ontology
>> not scalable

The context-oriented approach: C-PANKOW

- **Contextualize the pattern-matching** by taking into account the similarity of the Google-abstract in which the pattern was matched and the one to be annotated
- **Download a fix number n of Google-abstracts** matching so-called *clues* and analyze them linguistically, matching the patterns offline:
 - match more complex structures
 - more efficient as the number of Google-queries only depends on n
 - more offline processing, reducing network traffic

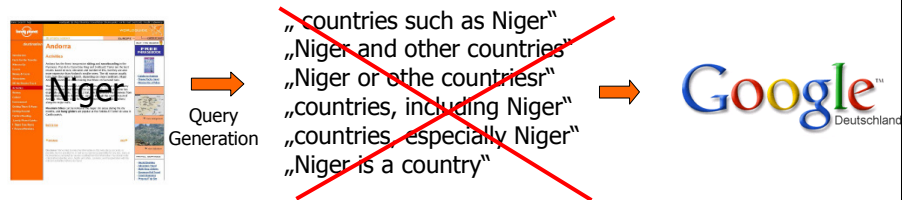
C-PANKOW Process



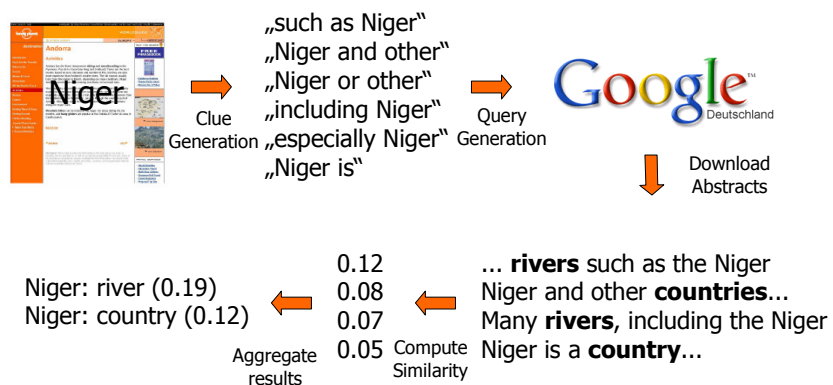
Instance Recognition

- Part-of-speech tag the web page to be annotated
- Consider sequences of upper-case proper nouns as potential instances
- Some heuristics to spot complex instances:
 - Pyramids of Gizeh
 - Torre del Oro

PANKOW



C-PANKOW



PART 2:

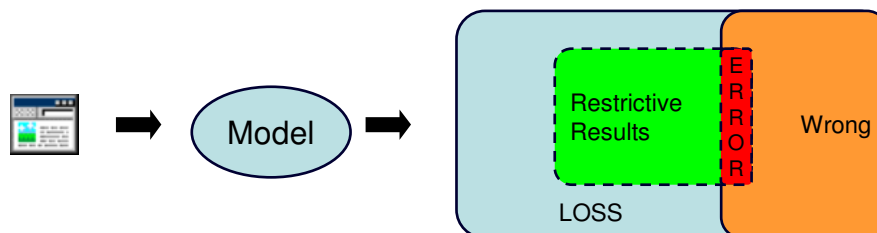
Restrictive Methods and Meta Methods for Organizing Document Collections

Restrictive Methods: Motivation

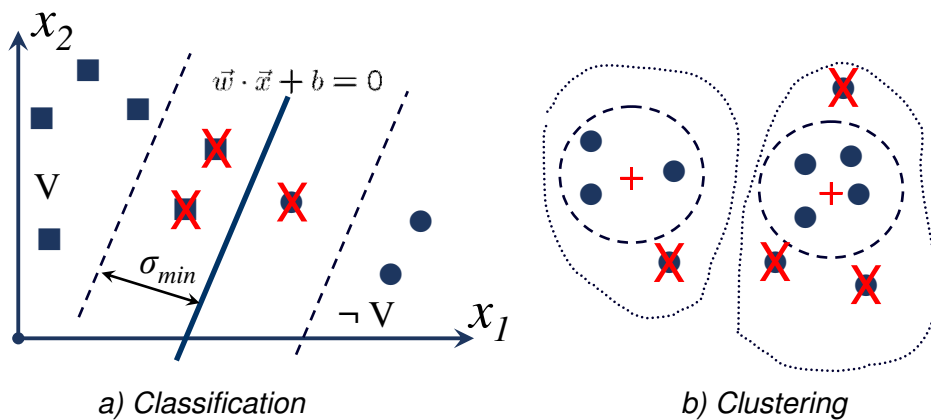
Typical tasks:

- Repository filtering (accurate classification)
- Organization of topics or concepts (clustering)

Idea: organize a **subset** of the repository; dismiss uncertain documents



Restrictive Base Methods



Problems with base methods:

- sparse training data
- initialization/tuning parameters, model anomalies, etc.

P2P Scenario

- Multiple users share topic(s) of interest, can cooperate
- Peers form an overlay network (e.g. Epidemic or Chord)
- Each peer maintains the local data repository (e.g. results of the focused crawl)
- **The data sharing across peers is unwanted**
 - privacy reasons
 - increased network load
 - computational overhead
- **Idea:** peers cooperate by exchanging **models** (centroids, hyperplanes) rather than local data. Received models are used to construct the **meta model**.

Meta Methodology

- given: set of methods $V = \{v_1, \dots, v_L\}$, assignments $res(v_i, d) \in \{-1; 1\}$ for document d and topic/cluster T
- meta result (restrictivity by thresholds t_1 and t_2 , tuning by weights $w(v_i)$):

$$meta(d) = \begin{cases} +1 & \text{if } \sum_i res_i(d) \cdot w(v_i) > t_1 \\ -1 & \text{if } \sum_i res_i(d) \cdot w(v_i) < t_2 \\ 0 & \text{otherwise} \end{cases}$$

Special cases:

- "Unanimous Decision"
- "Voting"
- "Weighted Average" (e.g., weighted by some quality estimator)

P2P Scenario

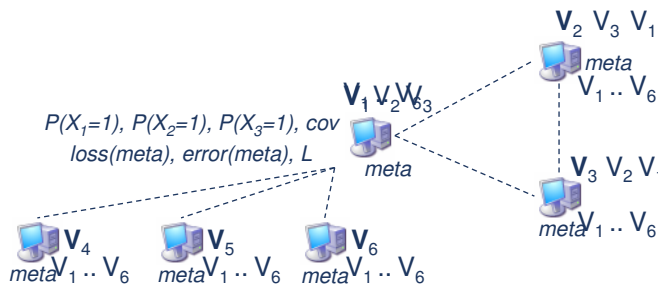
Given: set of methods $V = \{v_1, \dots, v_L\}$, „unanimous decision“

$$X_i = \begin{cases} 1 & \text{if } v_i \text{ assigns document correctly} \\ 0 & \text{otherwise} \end{cases}$$

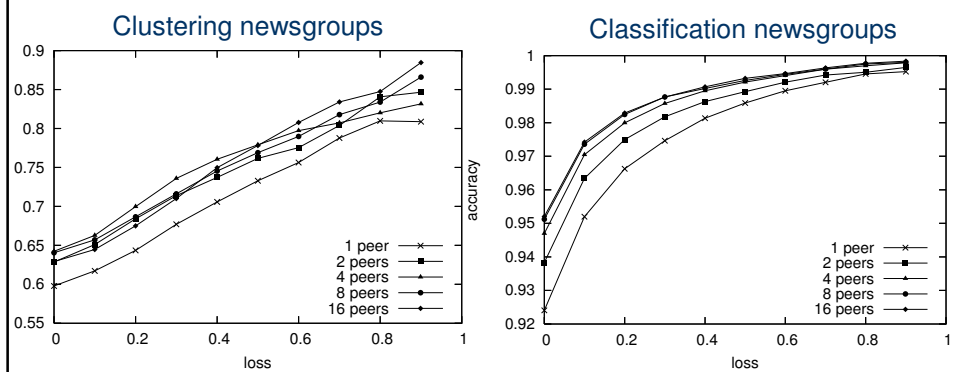
$$P(X_1 = 1, \dots, X_L = 1) = P(X_1 = 1) \cdot \prod_{i=1}^{L-1} \frac{P(X_i = 1)P(X_{i+1} = 1) + cov(X_i, X_{i+1})}{P(X_i = 1)}$$

$$error(meta) = P(X_1 = 0, \dots, X_L = 0 | X_1 = \dots = X_L)$$

$$loss(meta) = 1 - P(X_1 = \dots = X_L)$$



P2P Evaluation



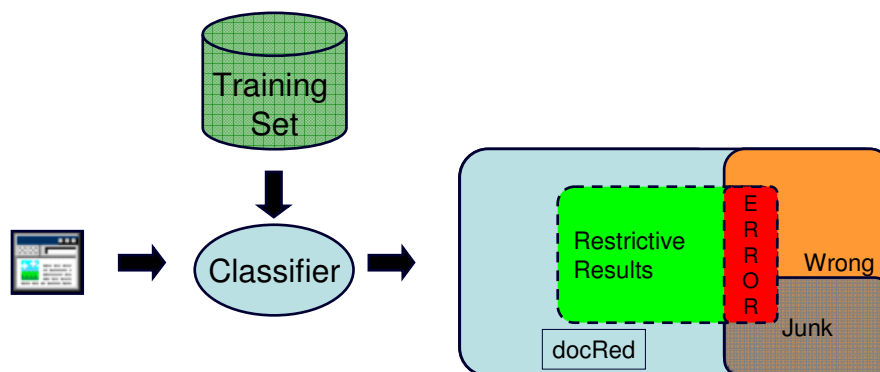
Reference collections:

- WebKB: 8282 HTML pages: "student", "faculty", "staff", etc.
- Newsgroups: 17847 postings from 20 newsgroups
- Reuters-21578. Contains Reuters newswire stories ("earn", "grain", "trade", etc.).
- IMDB. Contains 6853 movie descriptions ("drama", "horror", etc.)

Quality measures:

- Classification: validation on separated samples
- Clustering: best possible overlap between topics and clusters

Restrictive Methods: Junk Elimination



Junk Elimination (2)

		classification		
		A	B	0
real class	A	AA	AB	A0
	B	BA	BB	B0
junk		JA	JB	J0

$$docred = \frac{|A0| + |B0| + |J0|}{|U|}$$

$$\text{maximize } junkred = \frac{|J0|}{|JA| + |JB| + |J0|}$$

$$\text{minimize } error = \frac{|AB| + |BA| + |JA| + |JB|}{|AB| + |BA| + |JA| + |JB| + |AA| + |BB|}$$

$$\text{minimize } loss = \frac{|A0| + |B0|}{|AB| + |BA| + |AA| + |BB| + |A0| + |B0|}$$

Junk Elimination (3)

Model:

- Probability $p < 0.5$ to misassign doc from pos/neg for all k classifiers
- 50% of docs are Junk
- junkDoc is assigned to pos/neg with prob. 0.5
- $c < p(p-1)$ (no perfect correlation)

$$junkred = 1 - \left(\frac{c+1/4}{1/2} \right)^{k-1}$$

$$loss = 1 - (1-p) \left(\frac{c+(1-p)^2}{1-p} \right)^{k-1} - p \left(\frac{c+p^2}{p} \right)^{k-1}$$

$$error = \frac{1}{2} \frac{p \left(\frac{c+p^2}{p} \right)^{k-1} + \left(\frac{c+1/4}{1/2} \right)^{k-1}}{\left(\frac{c+1/4}{1/2} \right)^{k-1} + p \left(\frac{c+p^2}{p} \right)^{k-1} + (1-p) \left(\frac{c+(1-p)^2}{1-p} \right)^{k-1}}$$

$$junkred \xrightarrow[k \rightarrow \infty]{\text{monotonically}} 1$$

$$error \xrightarrow[k \rightarrow \infty]{\text{monotonically}} 0$$

$$loss \xrightarrow[k \rightarrow \infty]{\text{monotonically}} 1$$

$junkred > loss$

$$\frac{1 - junkred}{1 - loss} \xrightarrow[k \rightarrow \infty]{\text{monotonically}} 0$$

PART 3:

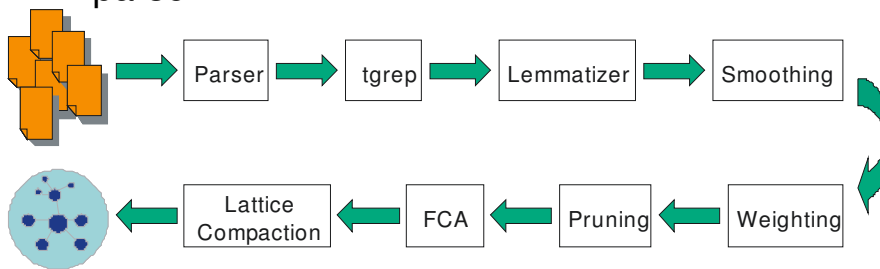
Clustering Concept Hierarchies from Text

Clustering Concept Hierarchies from Text

- Observation: ontology engineers need information about the **effectiveness**, **efficiency** and **trade-offs** of different approaches
- Similarity-based
 - agglomerative/bottom-up
 - divisive/top-down: Bi-Section-KMeans
- Set-theoretical
 - set operations (inclusion)
 - FCA, based on Galois lattices

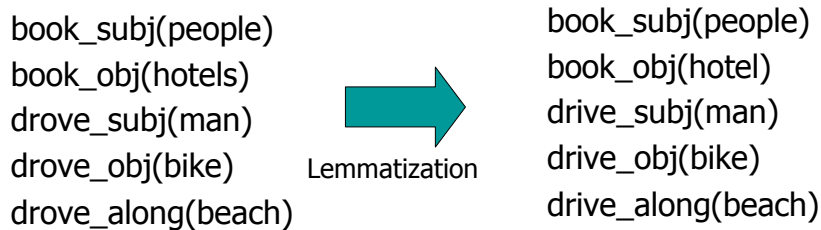
Context Extraction

- extract syntactic dependencies from text
 ⇒ verb/object, verb/subject, verb/PP relations
 ⇒ car: drive_obj, crash_subj, sit_in, ...
- LoPar, a trainable statistical left-corner parser:



Content Extracrion: Example

- People book hotels. The man drove the bike along the beach.



Weighting (threshold t)

$$P(n | v_{\text{arg}})$$

- Conditional:

$$P(n | v_{\text{arg}}) \cdot \log\left(\frac{P(n | v_{\text{arg}})}{P(n)}\right)$$

- Hindle:

$$S_R(v_{\text{arg}}) \cdot P(n | v_{\text{arg}}) \cdot \log\left(\frac{P(n | v_{\text{arg}})}{P(n)}\right)$$

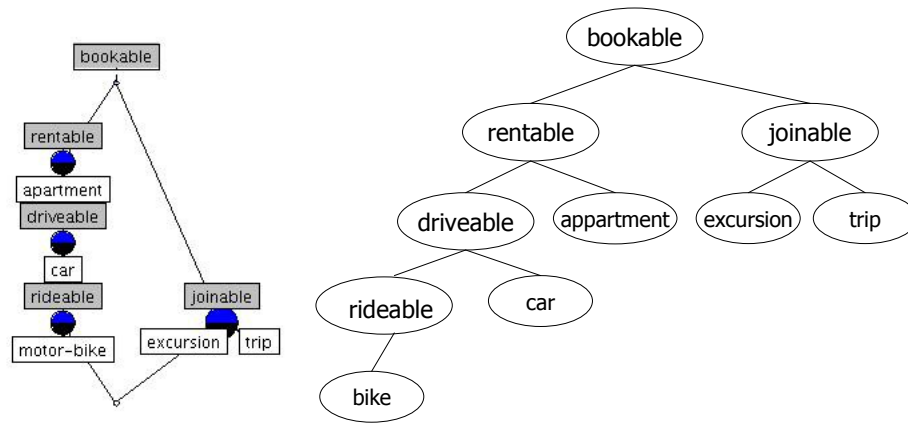
- Resnik:

$$S_R(v_{\text{arg}}) = \sum_{n'} P(n' | v_{\text{arg}}) \cdot \log\left(\frac{P(n' | v_{\text{arg}})}{P(n')}\right)$$

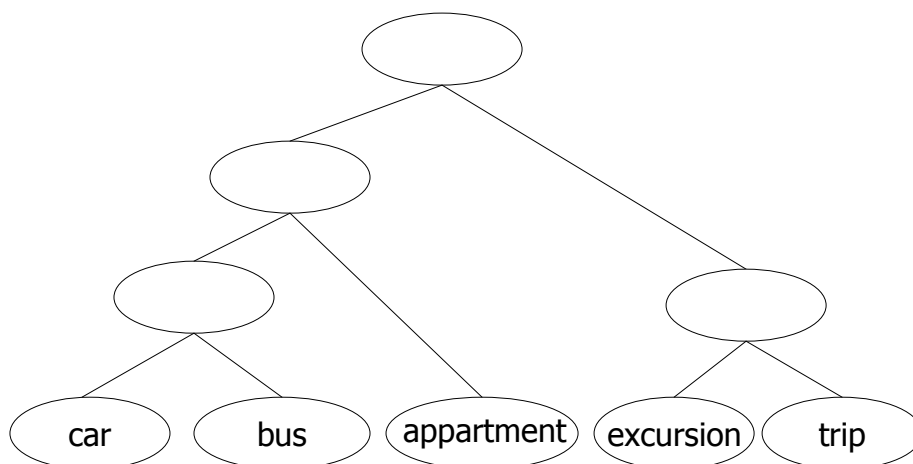
Tourism Formal Context

	bookable	rentable	driveable	rideable	joinable
apartment	X	X			
car	X	X	X		
motor-bike	X	X	X	X	
excursion	X				X
trip	X				X

Tourism Lattice & Concept Hierarchy



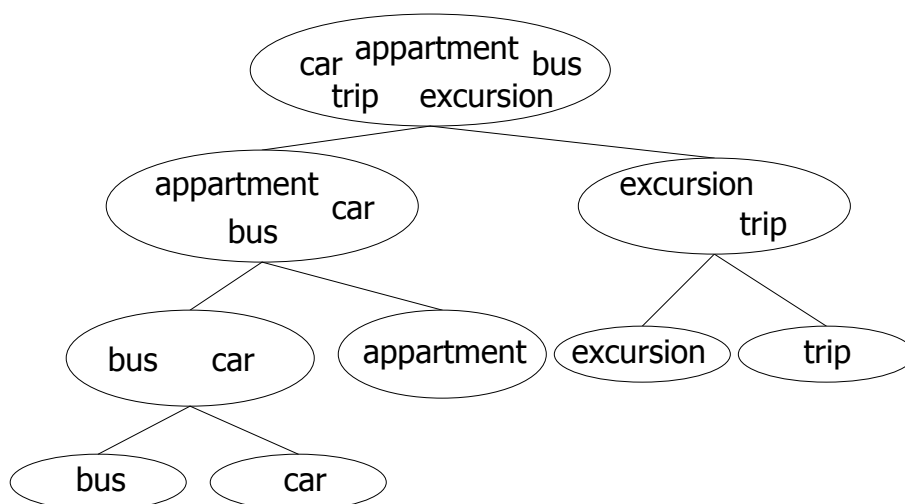
Agglomerative/Bottom-Up Clustering



Linkage Strategies

- Complete-Linkage:
 - consider the two most dissimilar elements of each of the clusters => $O(n^2 \log(n))$
- Average-Linkage:
 - consider the average similarity of the elements in the clusters => $O(n^2 \log(n))$
- Single-Linkage:
 - consider the two most similar elements of each of the clusters => $O(n^2)$

Bi-Section-KMeans



Conclusion

Conclusion

- Pattern-based annotation through knowledge on the Web
 - Pankow: using statistics from large-scale search engines for annotation
 - C-Pankow: more flexible queries, in-depth analysis of returned results, context-specific annotation
- Restrictive methods for data organization
 - Restrictive base methods and meta methods
 - Peer-to-peer application scenario
 - Junk elimination
- Automatic generation of ontologies
 - effectiveness, efficiency and trade-offs
 - FCA, based on Galois lattices
- Connections to other projects and activities
 - Application scenarios for annotation and ontology generation
 - Collaborative data organization (e.g. annotation) using meta methodology
 -

Thank you

References

- Reinberger, M.-L., & Spyns, P. (2005). Unsupervised text mining for the learning of dogma-inspired ontologies. In Buitelaar, P., Cimiano, P., & Magnini, B. (Eds.), *Ontology Learning from Text: Methods, Evaluation and Applications*.
- Sergej Sizov, Stefan Siersdorfer: Automatic Document Organization in a P2P Environment. 28th European Conference on Information Retrieval Research (ECIR), London, 2006.
- Philipp Cimiano, Andreas Hotho, Steffen Staab: Comparing Conceptual, Divise and Agglomerative Clustering for Learning Taxonomies from Text. ECAI 2004: 435-439
- P. Cimiano, A. Pivk, L. Schmidt-Thieme and S. Staab, Learning Taxonomic Relations from Heterogenous Evidence. In Buitelaar, P., Cimiano, P., & Magnini, B. (Eds.), *Ontology Learning from Text: Methods, Evaluation and Applications*.
- Sergej Sizov, Stefan Siersdorfer: Restrictive Clustering and Metaclustering for self-organizing Document Collections. 27th Annual International ACM SIGIR Conference, Sheffield, 2004.
- Sergej Sizov, Stefan Siersdorfer, Gerhard Weikum: Goal-oriented Methods and Meta Methods for Document Classification and their Parameter Tuning. 13th Conference on Information and Knowledge Management (CIKM), Washington D.C., USA, 2004
- Sabou M., Wroe C., Goble C. and Mishne G., Learning Domain Ontologies for Web Service Descriptions: an Experiment in Bioinformatics, In *Proceedings of the 14th International World Wide Web Conference (WWW2005)*, Chiba, Japan, 10-14 May, 2005.
- M. Ciaramita, A. Gangemi, E. Ratsch, J. Saric, I. Rojas. Unsupervised Learning of semantic relations between concepts of a molecular biology ontology. *IJCAI*, 659ff.