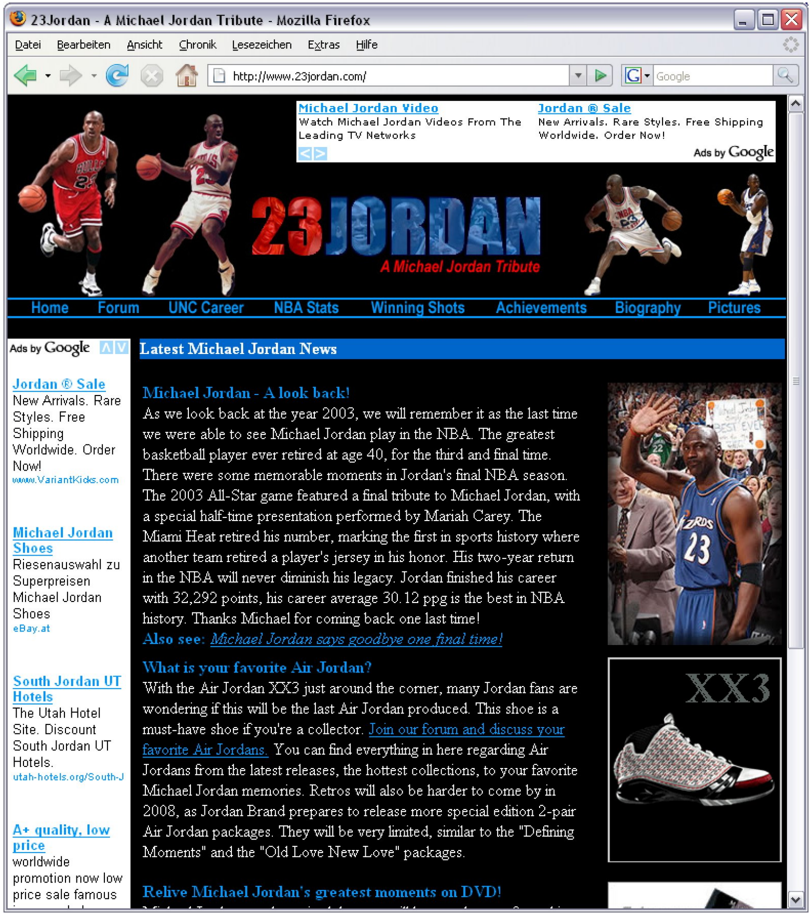
Chapter DM:II (continued)

II. Cluster Analysis

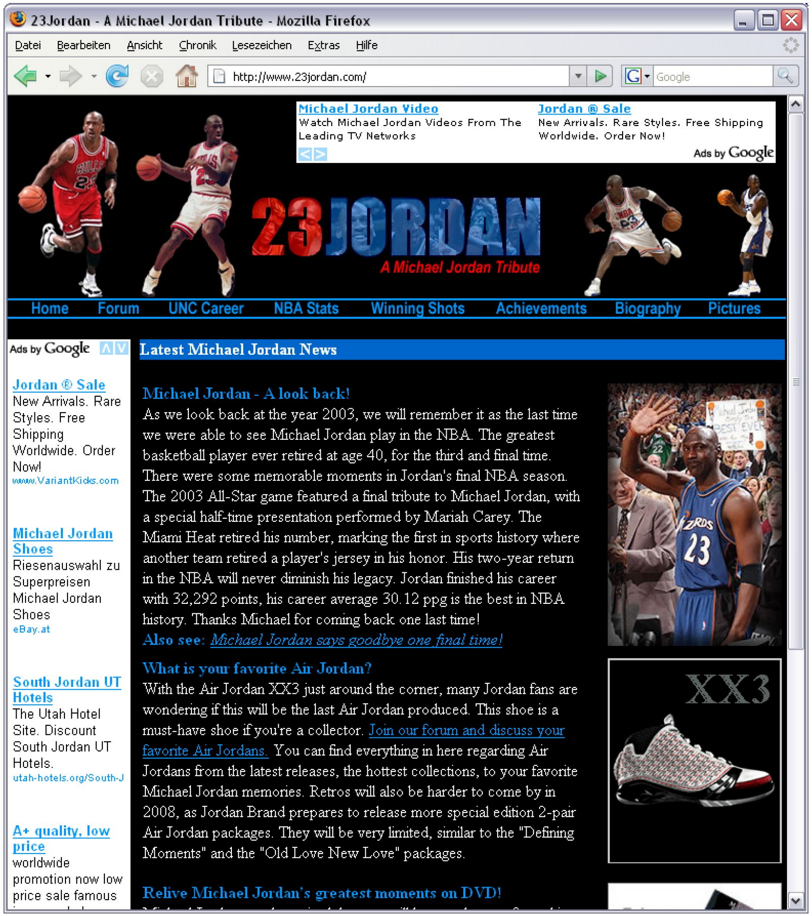
Constrained Cluster Analysis

Person Resolution Task



Constrained Cluster Analysis

Person Resolution Task

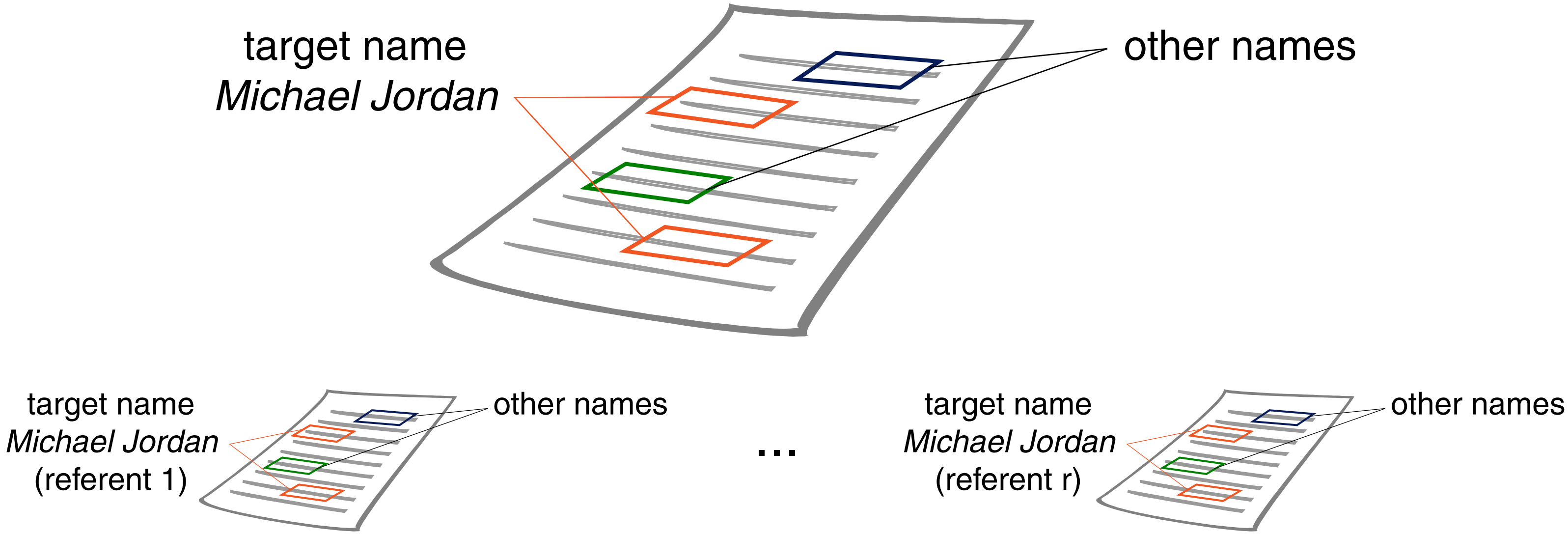


Constrained Cluster Analysis

Person Resolution Task

Constrained Cluster Analysis

Person Resolution Task

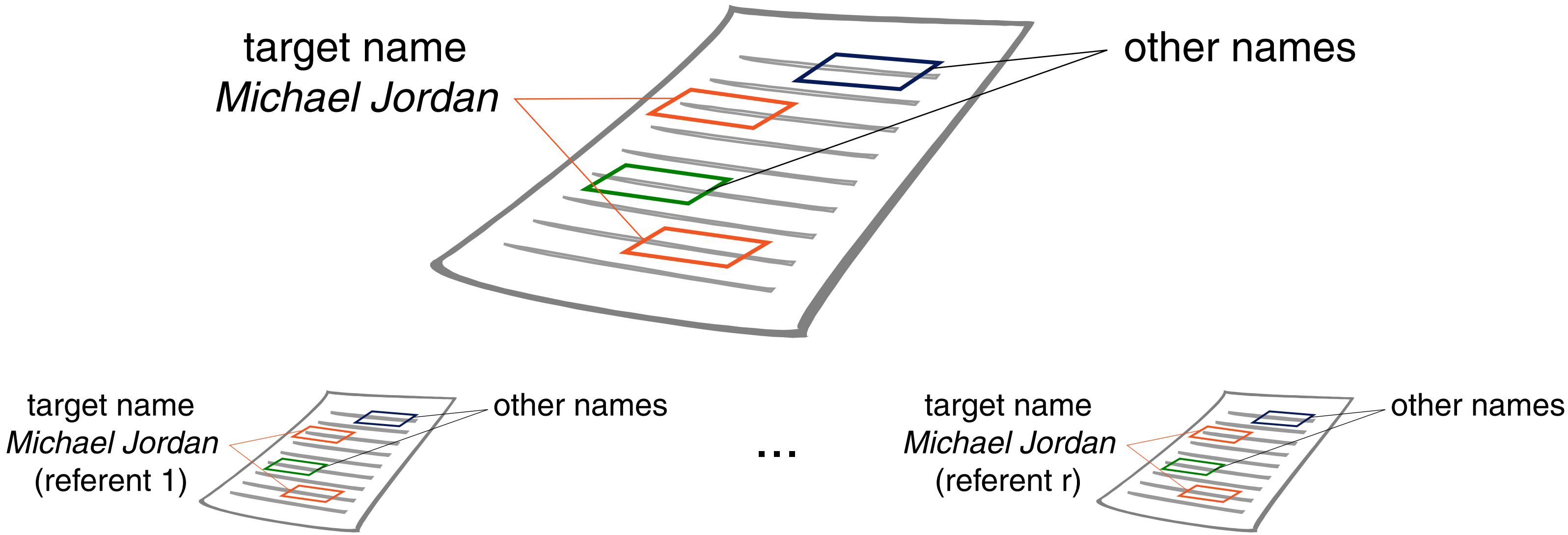


The basket ball player.

The statistician.

Constrained Cluster Analysis

Person Resolution Task



The basket ball player. *The statistician.*

- **Multi-document resolution task:**

Names, Target names:	$N = \{n_1, \ldots, n_l\}$,	$T \subset N$
Referents:	$R = \{r_1, \ldots, r_m\}$,	$\tau : R \to T$, $\lvert R \rvert \gg \lvert T \rvert$
Documents:	$D = \{d_1, \ldots, d_n\}$,	$\nu : D \to \mathcal{P}(N)$, $\lvert \nu(d_i) \cap T \rvert = 1$
A solution:	$\gamma : D \to R$,	s.t. $\tau(\gamma(d_i)) \in \nu(d_i)$

Constrained Cluster Analysis

Person Resolution Task

The basket ball player.

...

The statistician.

- **Facts about the Spock data mining challenge:**

  Target names: $|T| = 44$

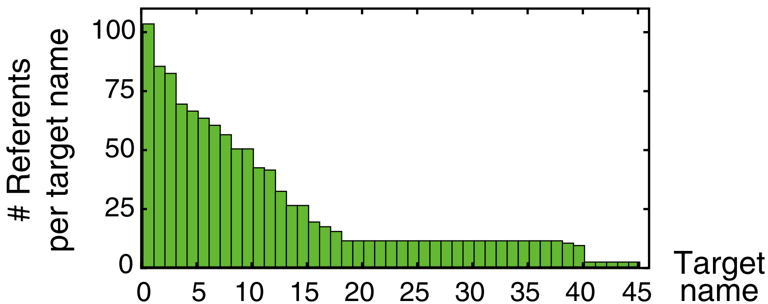
  Referents: $|R| = 1\,101$

  Documents: $|D_{train}| = 27\,000$ (labeled $\approx$ 2.3GB)

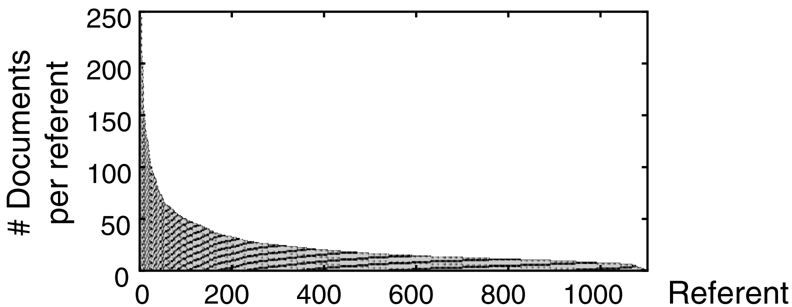
  $|D_{test}| = 75\,000$ (unlabeled $\approx$ 7.8GB)

Constrained Cluster Analysis

Person Resolution Task



- up to 105 referents for a single target name
- about 25 referents on average per target name



- about 23 documents on average per referent

- **Facts about the Spock data mining challenge:**

  Target names: $|T| = 44$

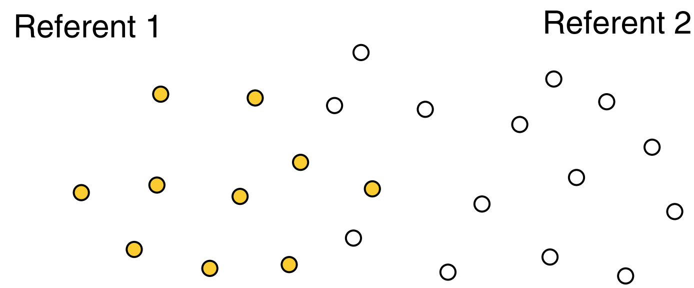
  Referents: $|R| = 1\,101$

  Documents: $|D_{train}| = 27\,000$ (labeled $\approx$ 2.3GB)

  $|D_{test}| = 75\,000$ (unlabeled $\approx$ 7.8GB)

Constrained Cluster Analysis

Applied to Multi-Document Resolution



1. Model similarities ➜ new and established retrieval models:
   - global and context-based vector space models
   - explicit semantic analysis
   - ontology alignment

2. Learn class memberships (supervised) ➜ logistic regression

3. Find equivalence classes (unsupervised) ➜ cluster analysis:

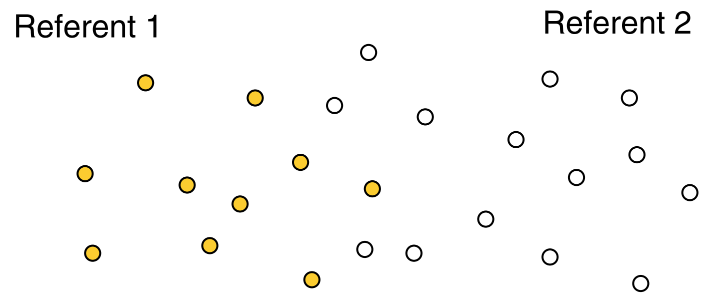
   (a) adaptive graph thinning

   (b) multiple, density-based cluster analysis

   (c) clustering selection by expected density maximization

Constrained Cluster Analysis

Applied to Multi-Document Resolution



1. Model similarities ➜ new and established retrieval models:
   - global and context-based vector space models
   - explicit semantic analysis
   - ontology alignment

2. Learn class memberships (supervised) ➜ logistic regression

3. Find equivalence classes (unsupervised) ➜ cluster analysis:

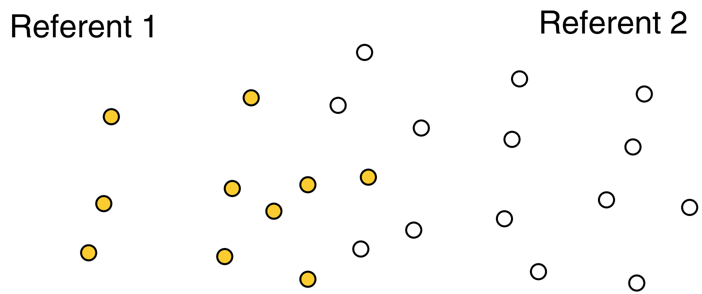
   (a) adaptive graph thinning

   (b) multiple, density-based cluster analysis

   (c) clustering selection by expected density maximization

Constrained Cluster Analysis

Applied to Multi-Document Resolution



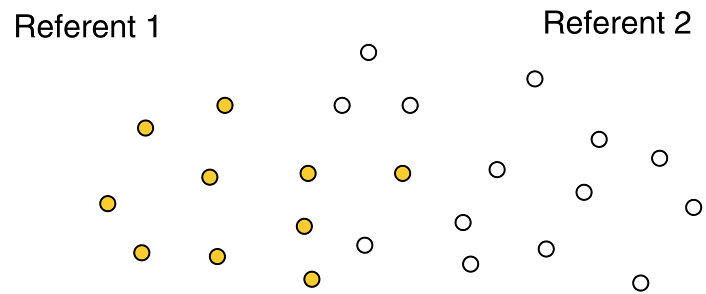
1. Model similarities ➜ new and established retrieval models:
   - global and context-based vector space models
   - explicit semantic analysis
   - ontology alignment

2. Learn class memberships (supervised) ➜ logistic regression

3. Find equivalence classes (unsupervised) ➜ cluster analysis:
   (a) adaptive graph thinning
   (b) multiple, density-based cluster analysis
   (c) clustering selection by expected density maximization

Constrained Cluster Analysis

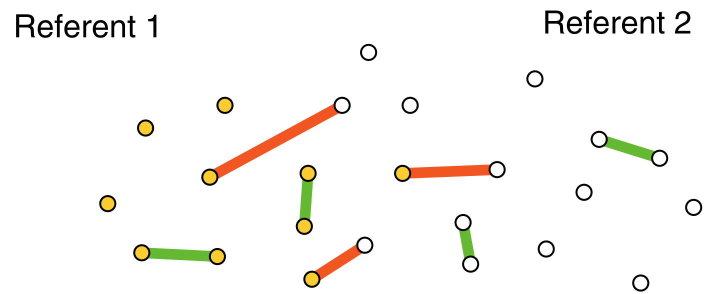
Applied to Multi-Document Resolution



1. Model similarities ➜ new and established retrieval models:
   - ❑ global and context-based vector space models
   - ❑ explicit semantic analysis
   - ❑ ontology alignment
2. Learn class memberships (supervised) ➜ logistic regression
3. Find equivalence classes (unsupervised) ➜ cluster analysis:
   (a) adaptive graph thinning
   (b) multiple, density-based cluster analysis
   (c) clustering selection by expected density maximization

Constrained Cluster Analysis

Applied to Multi-Document Resolution



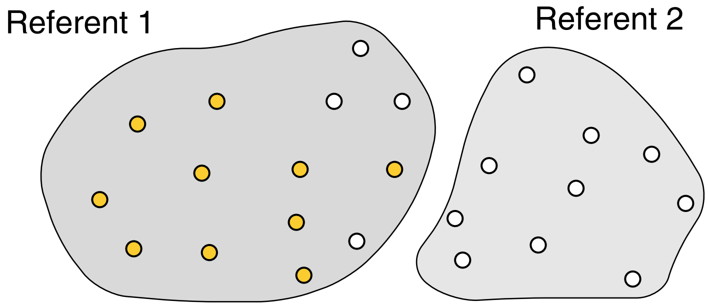
1. Model similarities ➜ new and established retrieval models:
   - global and context-based vector space models
   - explicit semantic analysis
   - ontology alignment

2. Learn class memberships (supervised) ➜ logistic regression

3. Find equivalence classes (unsupervised) ➜ cluster analysis:
   (a) adaptive graph thinning
   (b) multiple, density-based cluster analysis
   (c) clustering selection by expected density maximization

Constrained Cluster Analysis

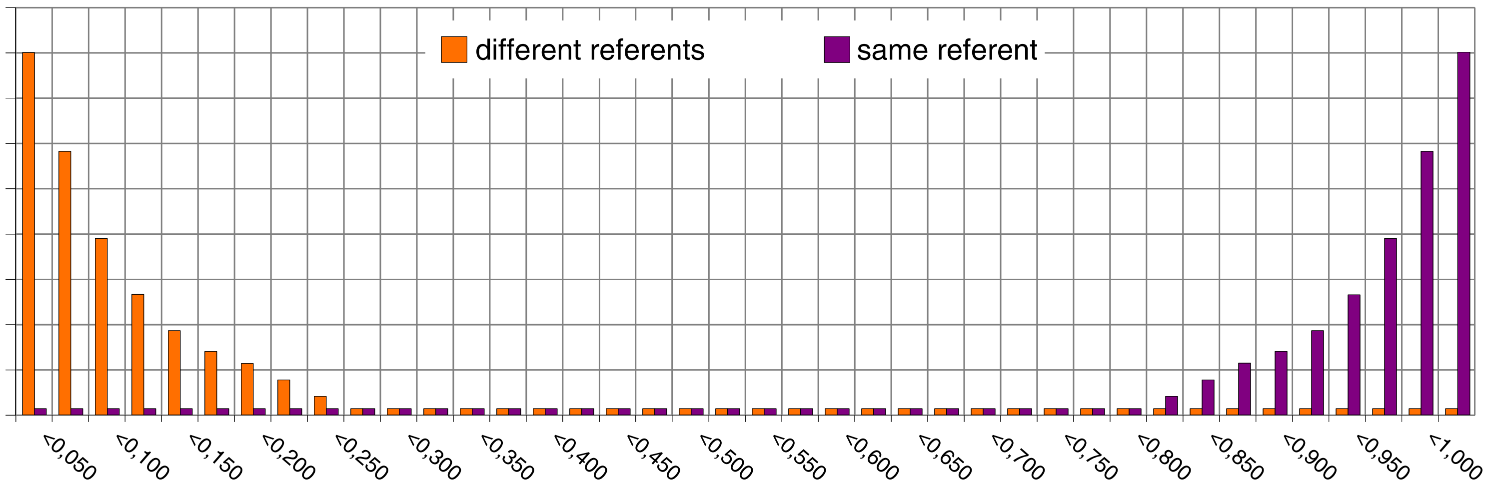
Applied to Multi-Document Resolution



1. Model similarities ➜ new and established retrieval models:
   - global and context-based vector space models
   - explicit semantic analysis
   - ontology alignment
2. Learn class memberships (supervised) ➜ logistic regression
3. Find equivalence classes (unsupervised) ➜ cluster analysis:
   (a) adaptive graph thinning
   (b) multiple, density-based cluster analysis
   (c) clustering selection by expected density maximization

Constrained Cluster Analysis

Idealized Class Membership Distribution over Similarities



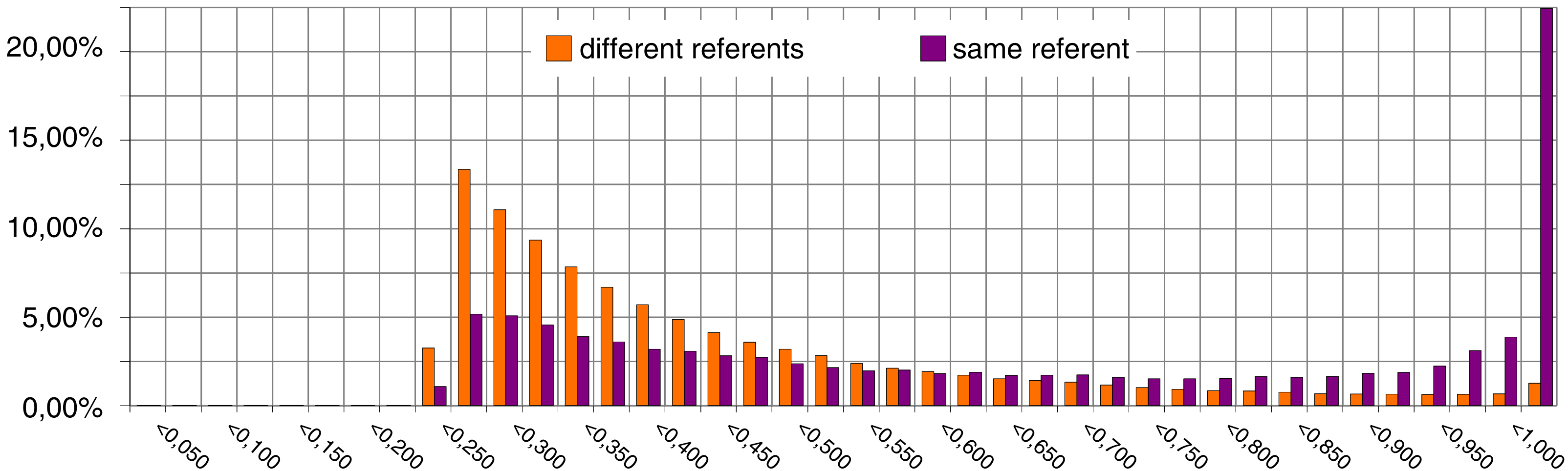
Similarity distributions for document pairs from different referents and same referent.

Logistic regression task:

- sample size: 400 000
- classes imbalance: non-target class : target class ≈ 25:1
- items are drawn uniformly distributed wrt. non-targets and targets
- items are uniformly distributed over the groups of target names

Constrained Cluster Analysis

Membership Distribution under *tf·idf* Vector Space Model

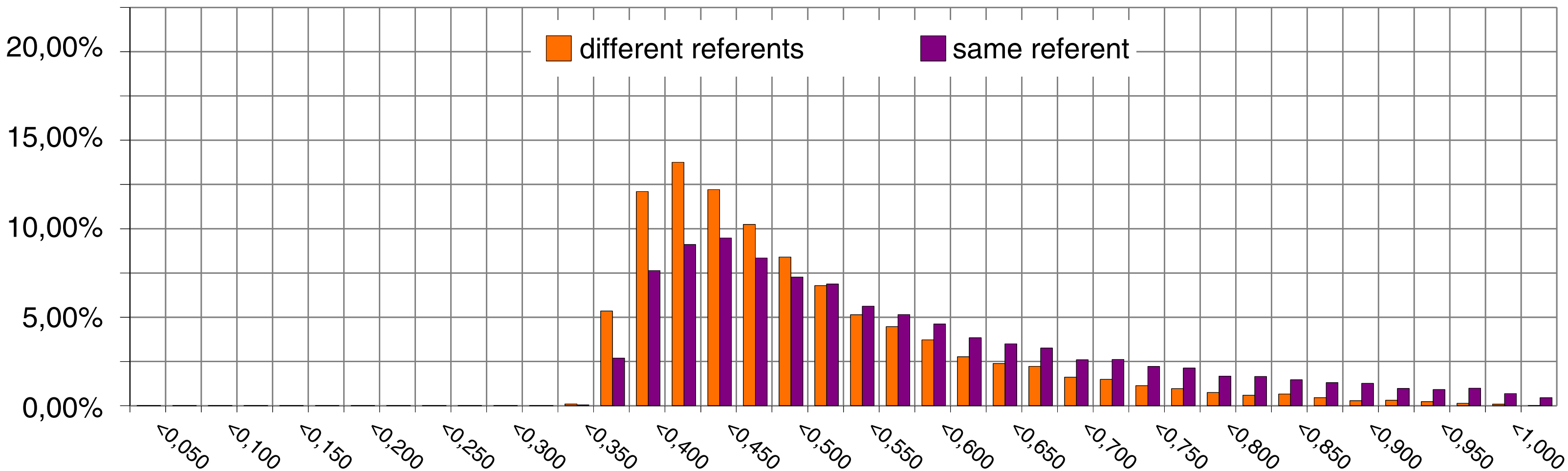


Model details:

- corpus size: 25 000 documents
- dictionary size: 1,2 Mio terms
- stopwords number: 850
- stopword volume: 36%

Constrained Cluster Analysis

Membership Distribution under Context-Based Vector Space Model

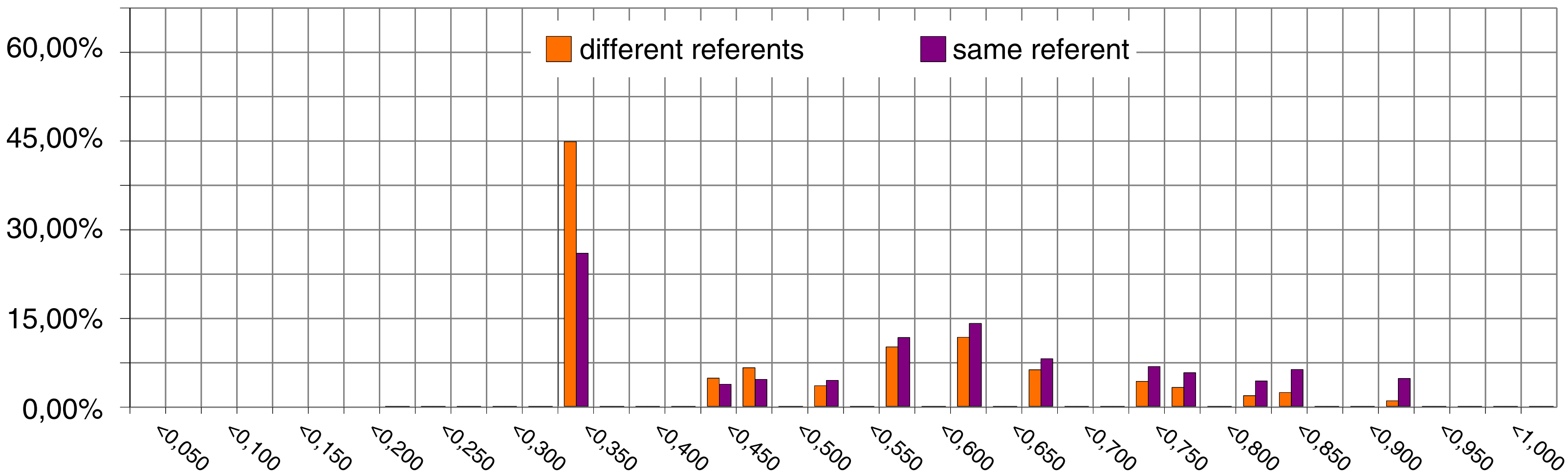


Model details:

- corpus size: 25 000 documents
- dictionary size: 1,2 Mio terms
- stopwords number: 850
- stopword volume: 36%

Constrained Cluster Analysis

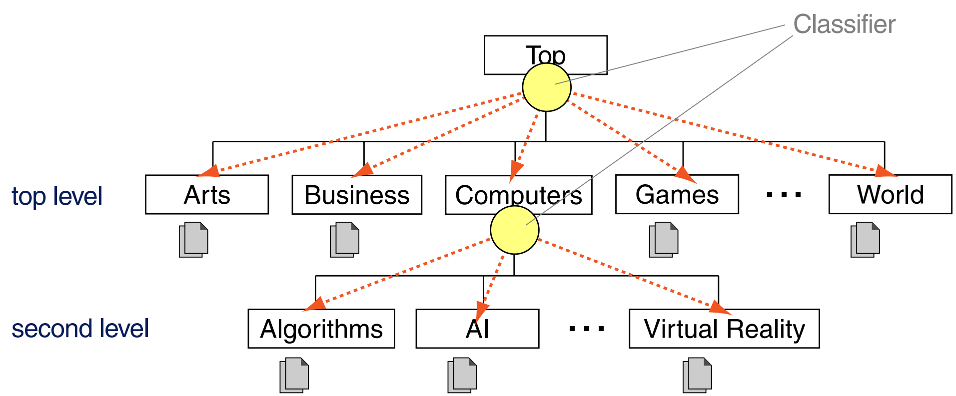
Membership Distribution under Ontology Alignment Model



Model details:

- DMOZ open directory project
- > 5 million documents

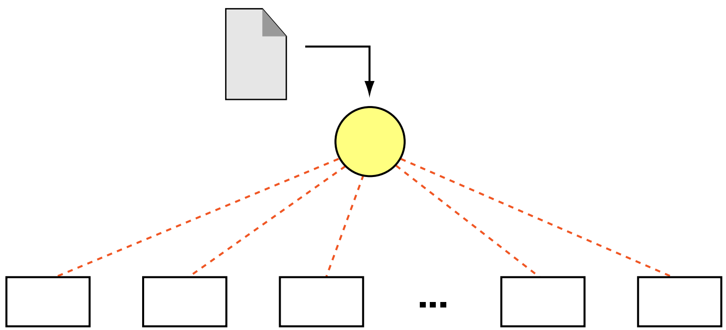
- 12 top-level categories
- 31 second level categories
- ML: hierarchical Bayes
- training set: 100 000 pages



Constrained Cluster Analysis

In-Depth: Multi-Class Hierarchical Classification

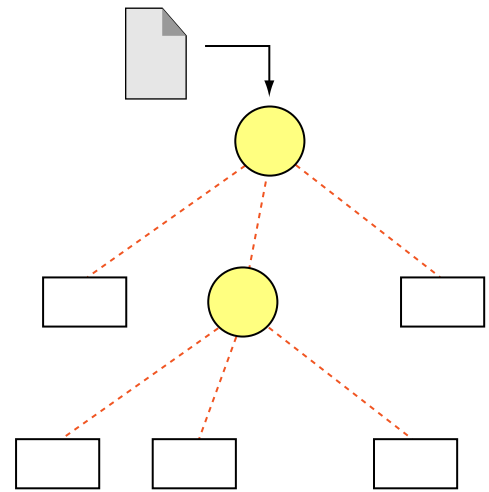
Flat (big-bang) classification



\+ simple realization

– loss of discriminative power with increasing number of categories

Hierarchical (top-down) classification



\+ specialized classifiers (divide and conquer)

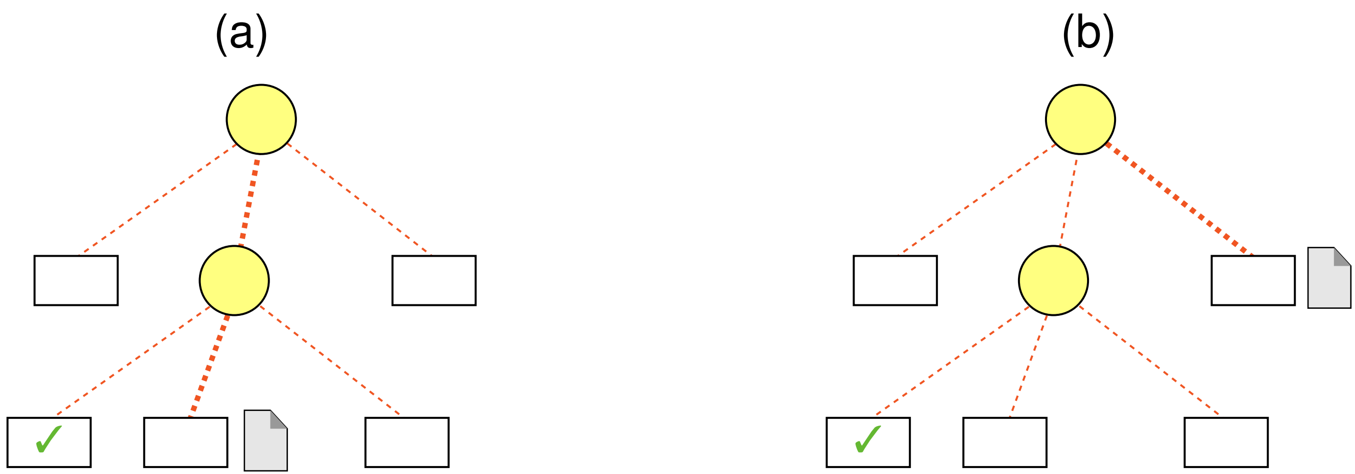
– misclassification at higher levels can never become repaired

Constrained Cluster Analysis

In-Depth: Multi-Class Hierarchical Classification

State of the art of effectiveness analyses:

1. independence assumption between categories
2. neglection of both hierarchical structure and degree of misclassification



Improvements:

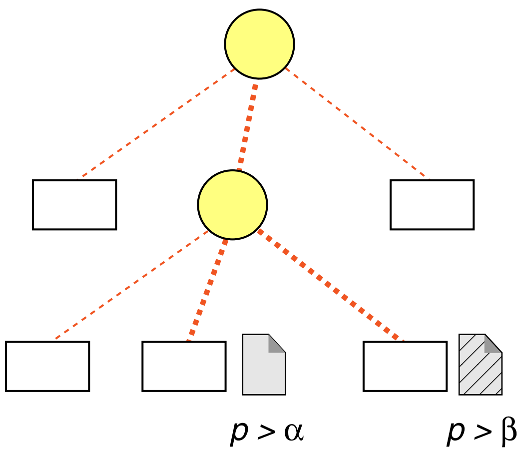
- Consider similarity $\varphi(C_i, C_j)$ between correct and wrong category.
- Consider graph distance $d(C_i, C_j)$ between correct and wrong category.

Constrained Cluster Analysis

In-Depth: Multi-Class Hierarchical Classification

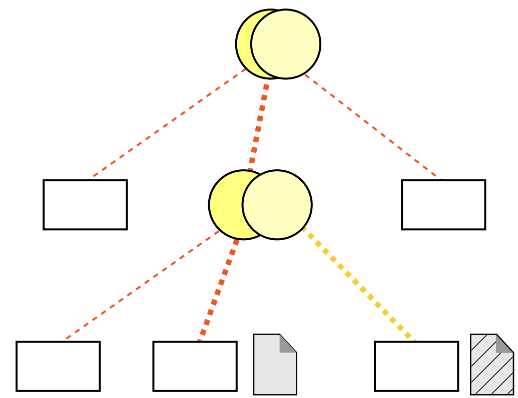
Improvements continued:

Multi-label (multi path) classification



- traverse more than one path and return all labels
- employ probabilistic classifiers with a threshold: split a path or not

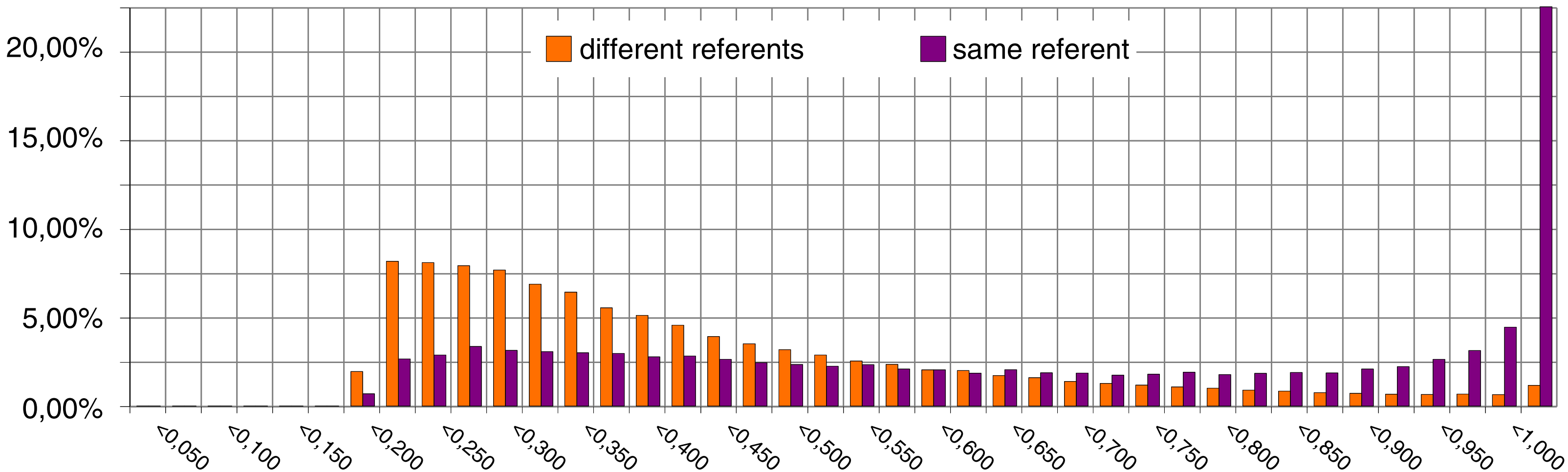
Multi-classifier (ensemble) classification



- classification result is a majority decision
- employ different classifier (different types or differently parameterized)

Constrained Cluster Analysis

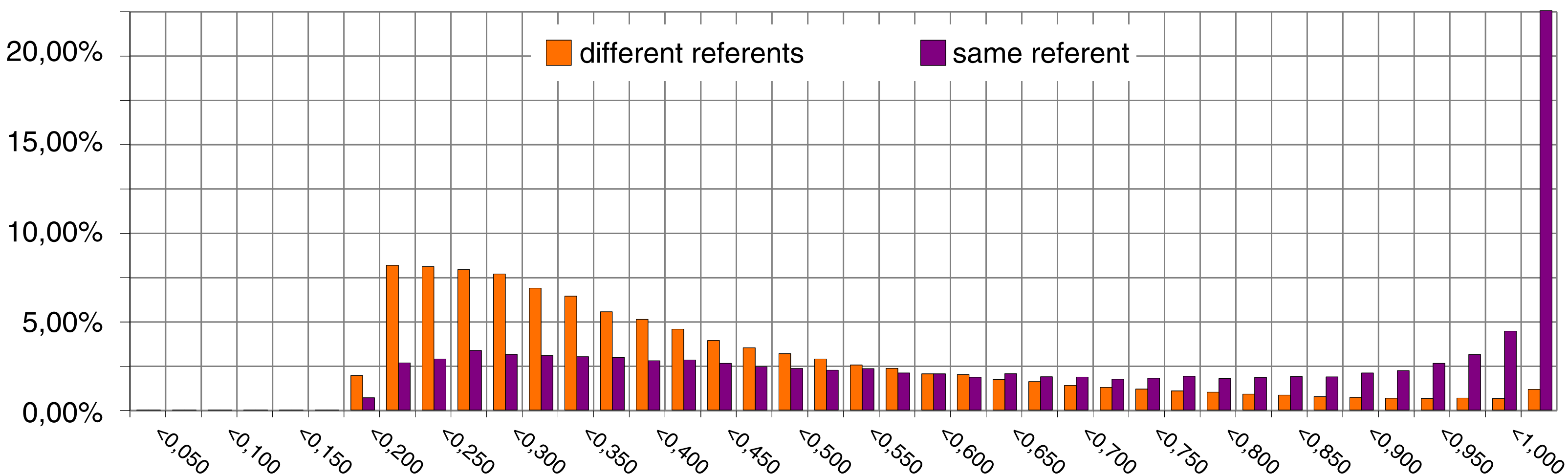
Membership Distribution under Optimized Retrieval Model Combination



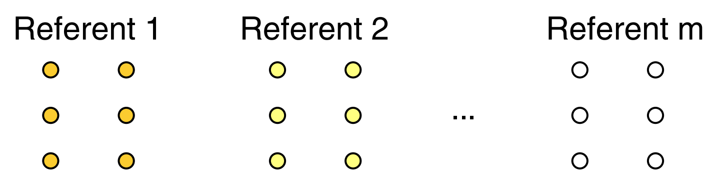
Retrieval Model	$F_{1/3}$ -Measure
tf·*idf* vector space	0.39
context-based vector space	0.32
ESA Wikipedia persons	0.30
phrase structure grammar	0.17
ontology alignment	0.15
optimized combination	**0.42**

Constrained Cluster Analysis

Membership Distribution under Optimized Retrieval Model Combination

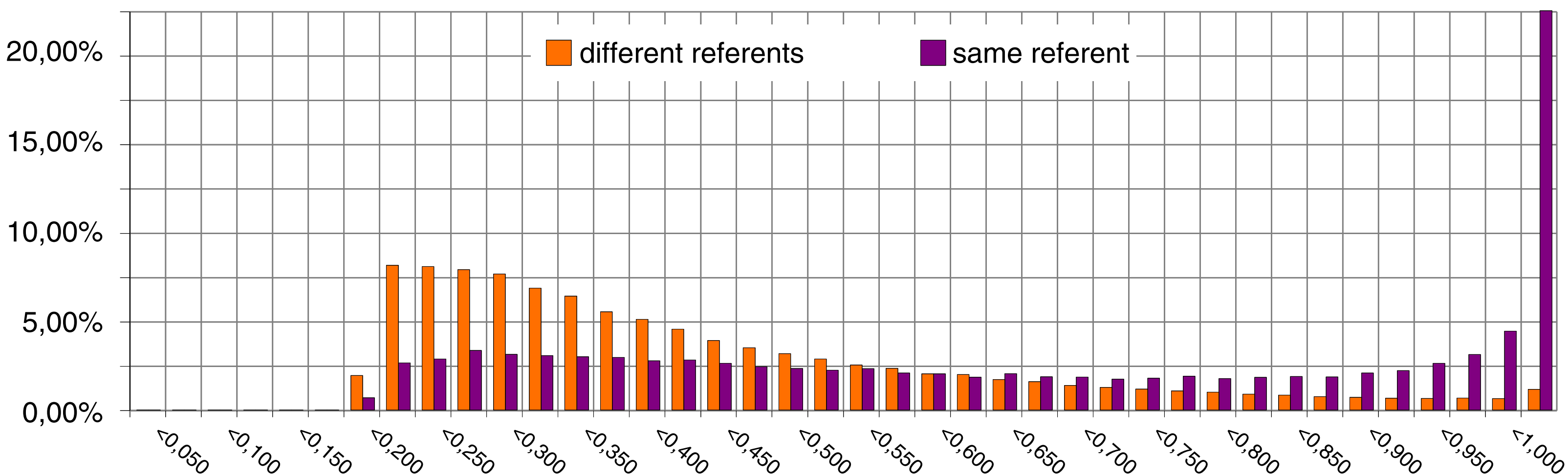


Retrieval Model	$F_{1/3}$ -Measure
tf·idf vector space	0.39
context-based vector space	0.32
ESA Wikipedia persons	0.30
phrase structure grammar	0.17
ontology alignment	0.15
optimized combination	**0.42**

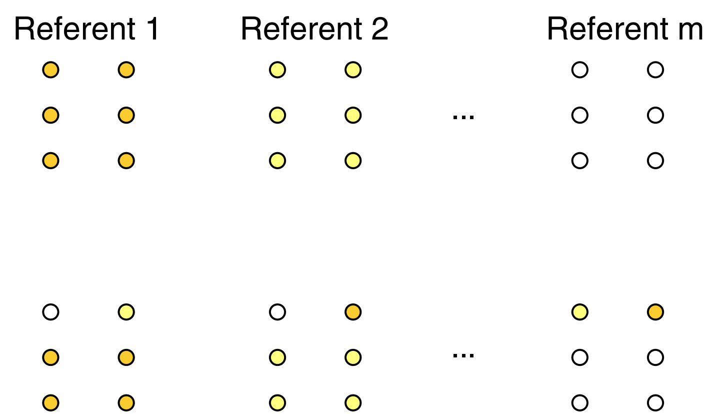


Constrained Cluster Analysis

Membership Distribution under Optimized Retrieval Model Combination

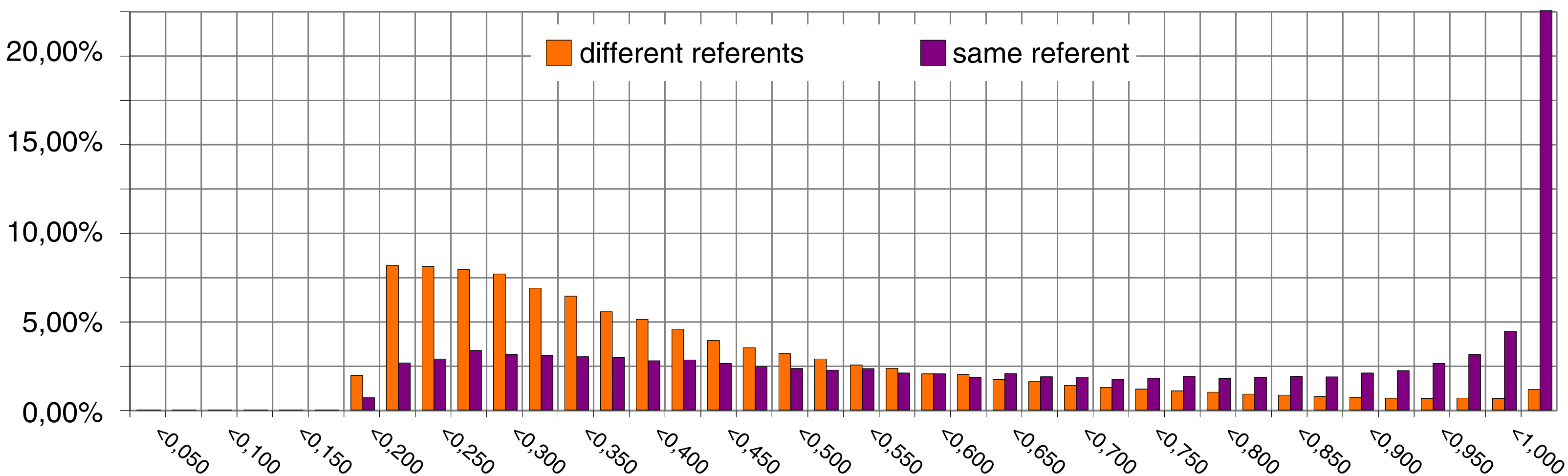


Retrieval Model	$F_{1/3}$ -Measure
tf·idf vector space	0.39
context-based vector space	0.32
ESA Wikipedia persons	0.30
phrase structure grammar	0.17
ontology alignment	0.15
optimized combination	**0.42**

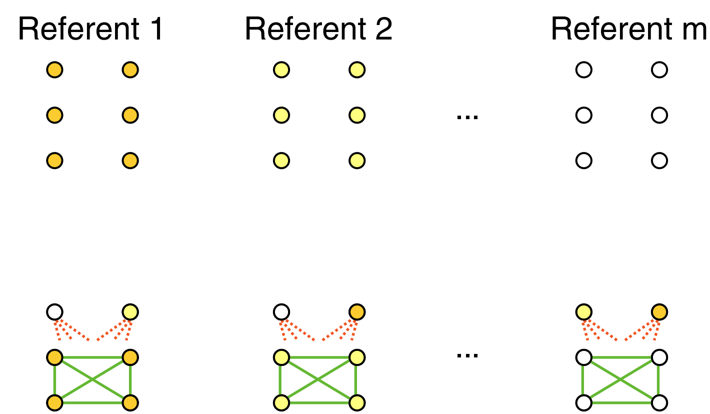


Constrained Cluster Analysis

Membership Distribution under Optimized Retrieval Model Combination

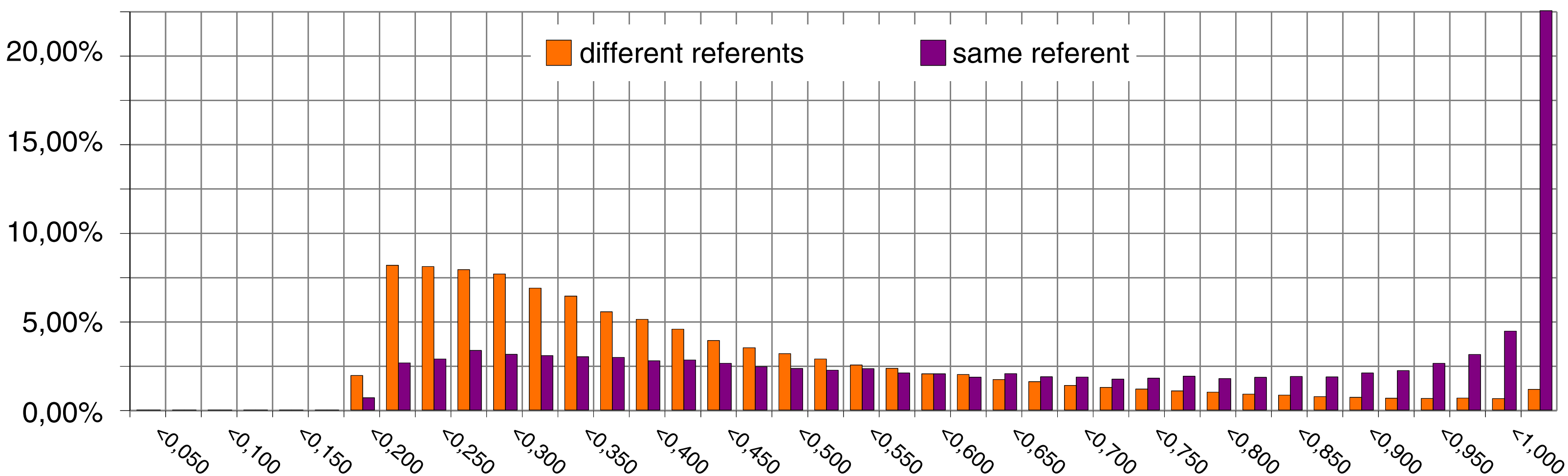


Retrieval Model	$F_{1/3}$ -Measure
tf·idf vector space	0.39
context-based vector space	0.32
ESA Wikipedia persons	0.30
phrase structure grammar	0.17
ontology alignment	0.15
optimized combination	**0.42**



Constrained Cluster Analysis

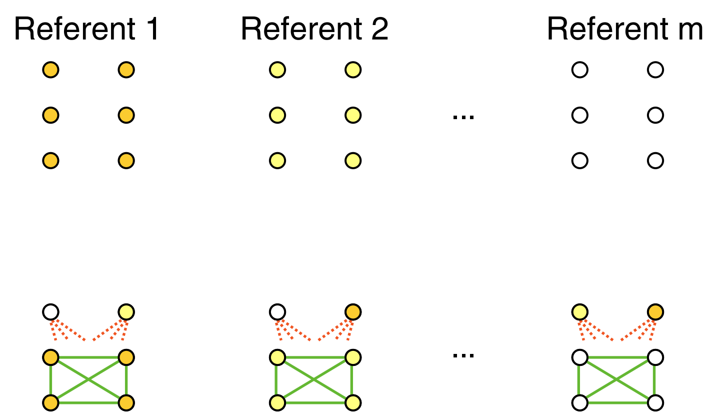
Membership Distribution under Optimized Retrieval Model Combination



In the example:

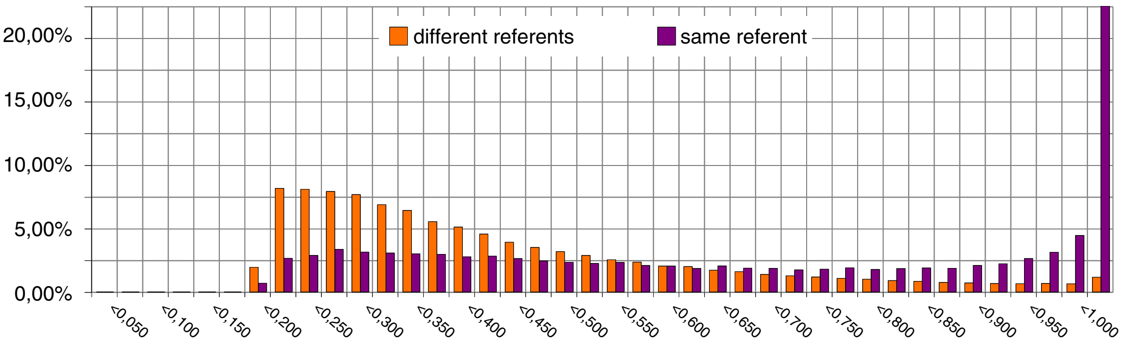
- precision = 0.4
- recall = 0.43
- $F_{1/3}$ = 0.41

(if false negatives are uniformly distributed)

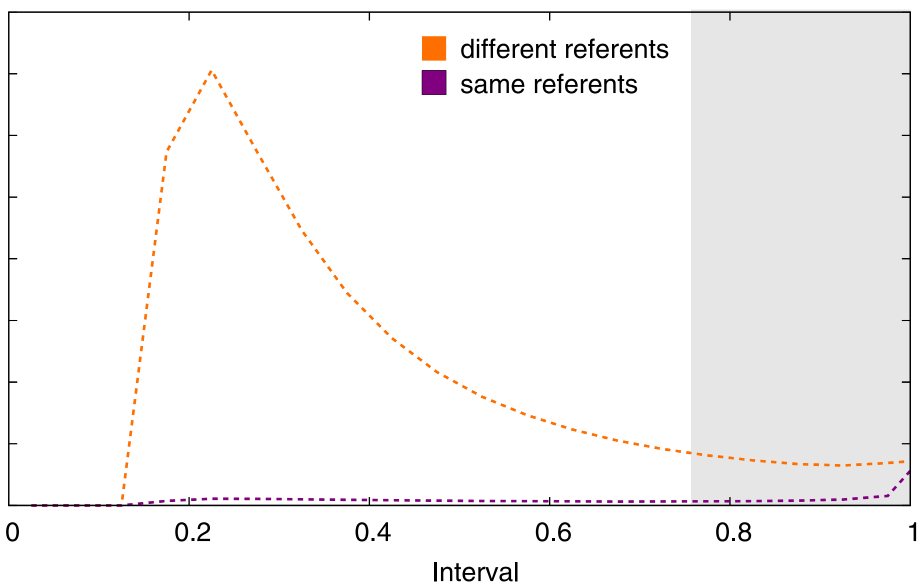


Constrained Cluster Analysis

In-Depth: Analysis of Classifier Effectiveness



Consideration of imbalance:



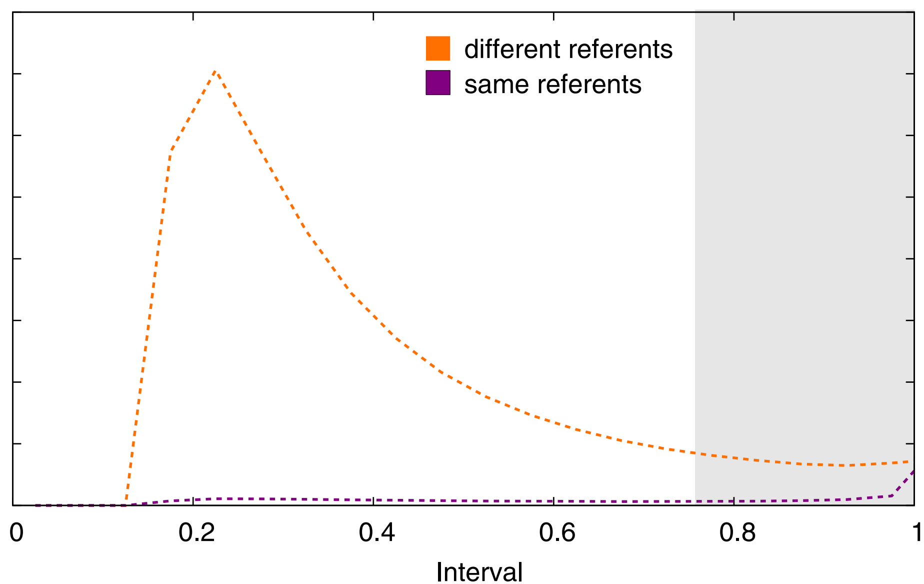
Constrained Cluster Analysis

In-Depth: Analysis of Classifier Effectiveness

- ❑ class imbalance factor (*CIF*) of 25
- ⇒ precision in interval [0.725; 1] for edges between same referents: $\approx$ 0.17

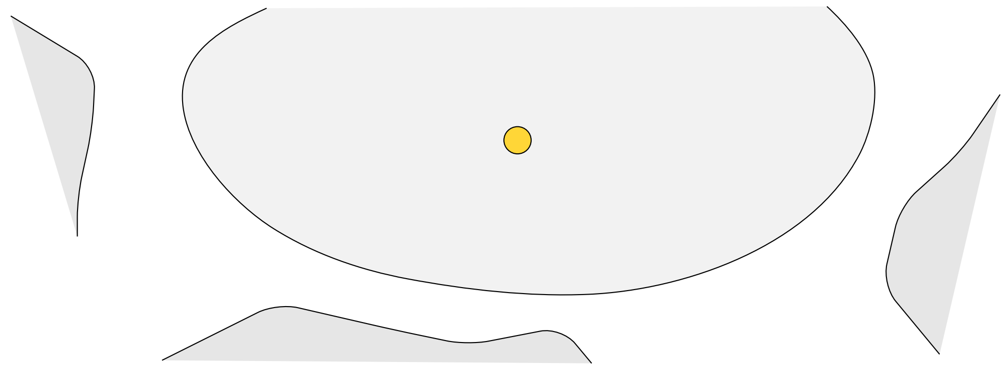
How can $F_{1/3}$ = 0.42 be achieved via cluster analysis?

Consideration of imbalance:



Constrained Cluster Analysis

In-Depth: Analysis of Classifier Effectiveness



Assumption: uniform distribution of referents over documents (here: 25 clusters with $|C| = 23$)

- $\Rightarrow$ $|TP|$ true 1-similarities per cluster (here: 130 @ threshold 0.725)
- $\Rightarrow$ $\frac{|TP|}{|C|}$ degree of true positives per node (here: 11)
- $\Rightarrow$ $|TP|(\frac{1}{\textit{precision}} - 1)$ false 1-similarities per cluster (here: 760)

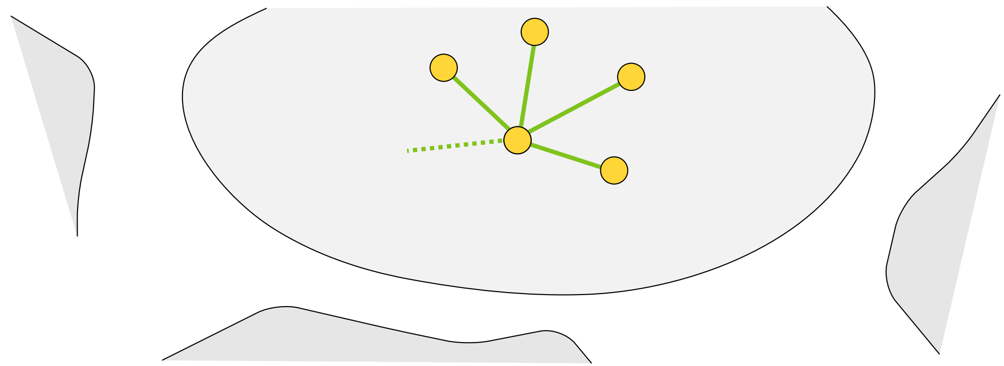
Density-based cluster analysis: effective false positives, FP^*, connect to same cluster

- $\Rightarrow$ analyze $P(|FP^*| > k \mid D, R_{iid})$ (here: $E(|FP^*|) = 2.7$)
- $\Rightarrow$ edge tie factor (*ETF*) specifies the excess of true positives until tie (here: $3 \ldots 5$)

$$ETF = \frac{|TP|}{|C| \cdot E(|FP^*|)}, \qquad \text{effective precision} = \textit{precision} \cdot \frac{CIF}{ETF}$$

Constrained Cluster Analysis

In-Depth: Analysis of Classifier Effectiveness



Assumption: uniform distribution of referents over documents (here: 25 clusters with $|C| = 23$)

- ⇒ $|TP|$ true 1-similarities per cluster (here: 130 @ threshold 0.725)
- ⇒ $\frac{|TP|}{|C|}$ degree of true positives per node (here: 11)
- ⇒ $|TP|(\frac{1}{precision} - 1)$ false 1-similarities per cluster (here: 760)

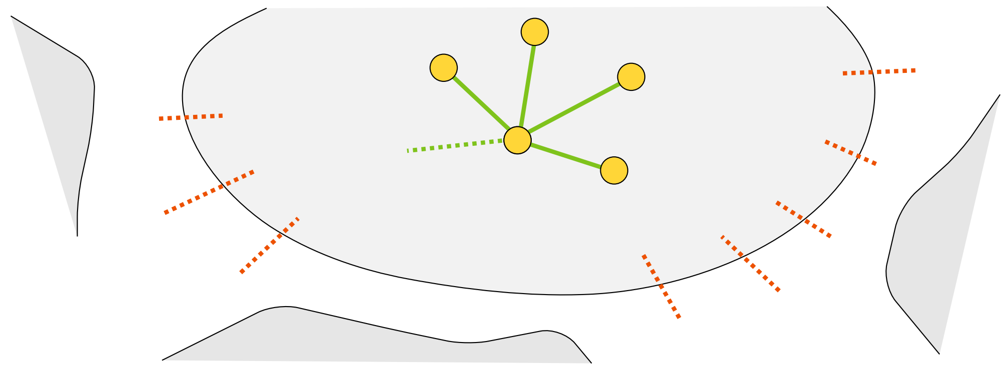
Density-based cluster analysis: effective false positives, FP^*, connect to same cluster

- ⇒ analyze $P(|FP^*| > k \mid D, R_{iid})$ (here: $E(|FP^*|) = 2.7$)
- ⇒ edge tie factor (ETF) specifies the excess of true positives until tie (here: $3 \ldots 5$)

$$ETF = \frac{|TP|}{|C| \cdot E(|FP^*|)}, \qquad \text{effective precision} = precision \cdot \frac{CIF}{ETF}$$

Constrained Cluster Analysis

In-Depth: Analysis of Classifier Effectiveness



Assumption: uniform distribution of referents over documents (here: 25 clusters with $|C| = 23$)

$\Rightarrow$ $|TP|$ true 1-similarities per cluster (here: 130 @ threshold 0.725)

$\Rightarrow$ $\frac{|TP|}{|C|}$ degree of true positives per node (here: 11)

$\Rightarrow$ $|TP|(\frac{1}{precision} - 1)$ false 1-similarities per cluster (here: 760)

Density-based cluster analysis: effective false positives, FP^*, connect to same cluster

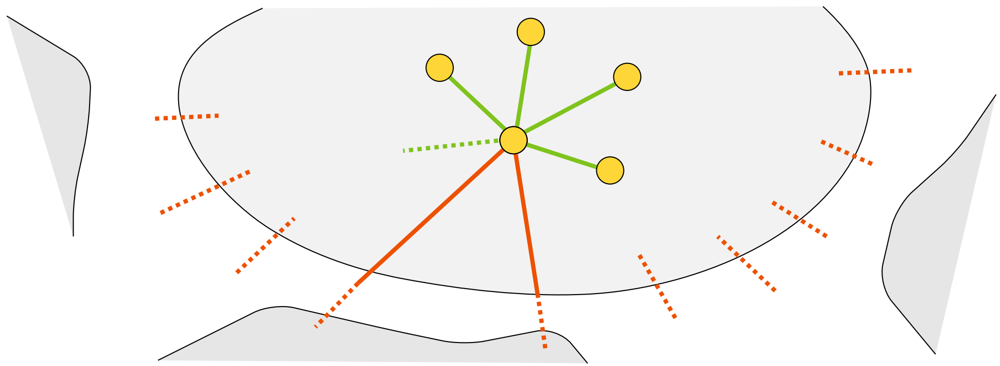
$\Rightarrow$ analyze $P(|FP^*| > k \mid D, R_{iid})$ (here: $E(|FP^*|) = 2.7$)

$\Rightarrow$ edge tie factor (ETF) specifies the excess of true positives until tie (here: $3 \ldots 5$)

$$ETF = \frac{|TP|}{|C| \cdot E(|FP^*|)}, \qquad \text{effective precision} = precision \cdot \frac{CIF}{ETF}$$

Constrained Cluster Analysis

In-Depth: Analysis of Classifier Effectiveness



Assumption: uniform distribution of referents over documents (here: 25 clusters with $|C| = 23$)

$\Rightarrow$ $|TP|$ true 1-similarities per cluster (here: 130 @ threshold 0.725)

$\Rightarrow$ $\frac{|TP|}{|C|}$ degree of true positives per node (here: 11)

$\Rightarrow$ $|TP|(\frac{1}{precision} - 1)$ false 1-similarities per cluster (here: 760)

Density-based cluster analysis: effective false positives, FP^*, connect to same cluster

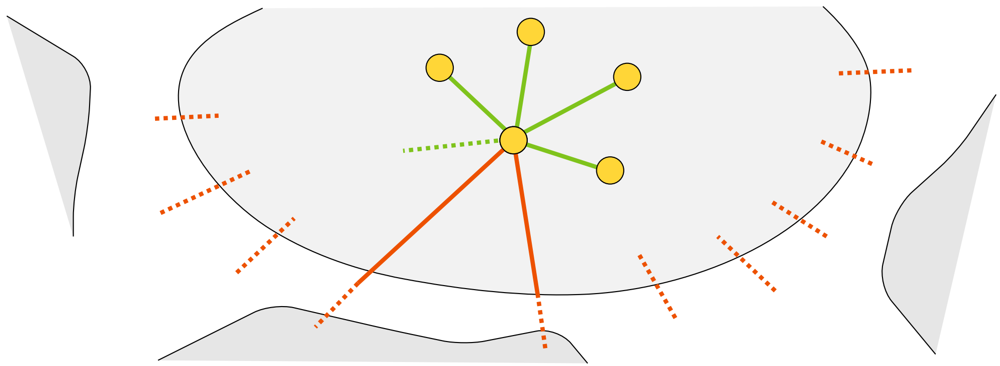
$\Rightarrow$ analyze $P(|FP^*| > k \mid D, R_{iid})$ (here: $E(|FP^*|) = 2.7$)

$\Rightarrow$ edge tie factor (ETF) specifies the excess of true positives until tie (here: $3 \ldots 5$)

$$ETF = \frac{|TP|}{|C| \cdot E(|FP^*|)}, \qquad \text{effective precision} = precision \cdot \frac{CIF}{ETF}$$

Constrained Cluster Analysis

In-Depth: Analysis of Classifier Effectiveness



Assumption: uniform distribution of referents over documents (here: 25 clusters with $|C| = 23$)

$\Rightarrow$ $|TP|$ true 1-similarities per cluster (here: 130 @ threshold 0.725)

$\Rightarrow$ $\frac{|TP|}{|C|}$ degree of true positives per node (here: 11)

$\Rightarrow$ $|TP|(\frac{1}{precision} - 1)$ false 1-similarities per cluster (here: 760)

Density-based cluster analysis: effective false positives, FP^*, connect to same cluster

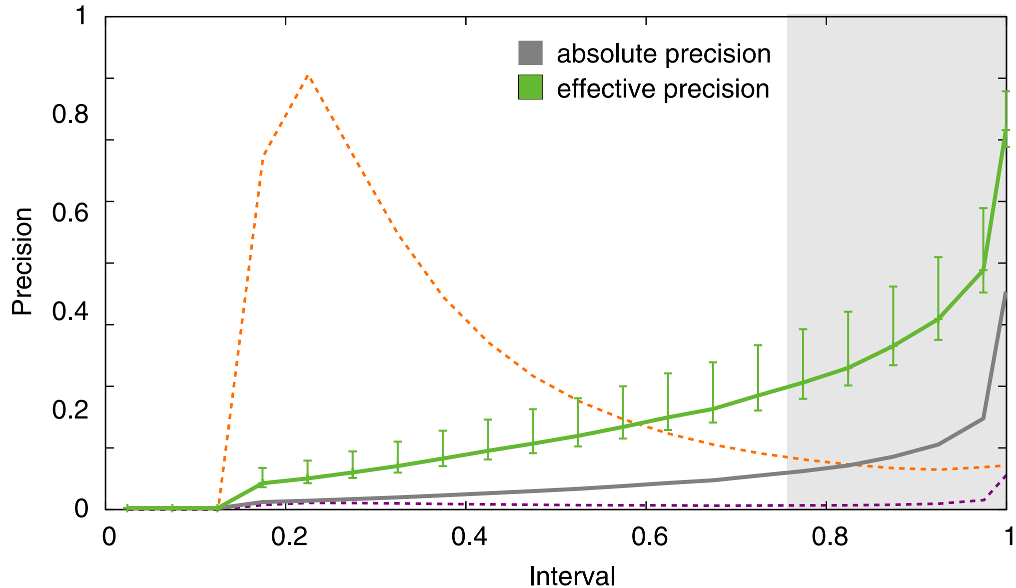
$\Rightarrow$ analyze $P(|FP^*| > k \mid D, R_{iid})$ (here: $E(|FP^*|) = 2.7$)

$\Rightarrow$ edge tie factor (*ETF*) specifies the excess of true positives until tie (here: $3 \ldots 5$)

$$ETF = \frac{|TP|}{|C| \cdot E(|FP^*|)}, \qquad \text{effective precision} = precision \cdot \frac{CIF}{ETF}$$

Constrained Cluster Analysis

In-Depth: Analysis of Classifier Effectiveness



Density-based cluster analysis: effective false positives, FP^*, connect to same cluster

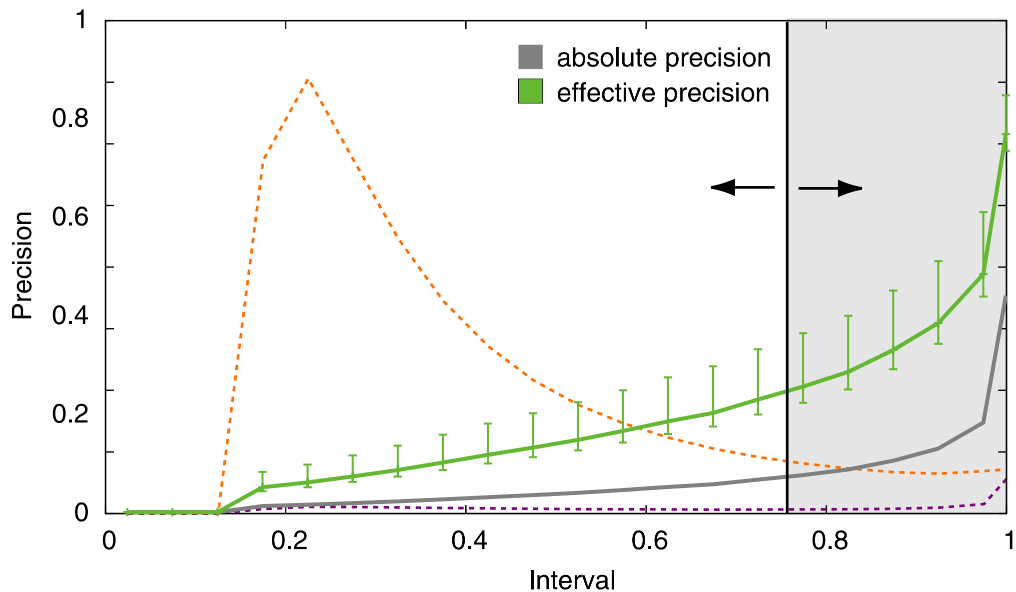
$\Rightarrow$ analyze $P(|\mathit{FP}^*| > k \;\mid\; D, R_{iid})$ (here: $E(|\mathit{FP}^*|) = 2.7$)

$\Rightarrow$ edge tie factor (*ETF*) specifies the excess of true positives until tie (here: $3 \ldots 5$)

$$\mathit{ETF} = \frac{|\mathit{TP}|}{|C| \cdot E(|\mathit{FP}^*|)}, \qquad \text{effective precision} = \mathit{precision} \cdot \frac{\mathit{CIF}}{\mathit{ETF}}$$

Constrained Cluster Analysis

In-Depth: Analysis of Classifier Effectiveness



Determine optimum similarity threshold for class-membership function:

$$\theta^* = \underset{\theta \in [0;1]}{\text{argmax}} \left\{ \frac{1+\alpha}{\frac{ETF}{precision_\theta \cdot CIF} + \frac{\alpha}{recall_\theta}} \right\}$$

θ^* considers co-variate shift, introduces model formation bias and sample selection bias.

Constrained Cluster Analysis

Model Selection: Our Risk Minimization Strategy

Retrieval Model	$F_{1/3}$ -Measure
tf·idf vector space	0.39
context-based vector space	0.32
ESA Wikipedia persons	0.30
phrase structure grammar	0.17
ontology alignment	0.15
optimized combination	**0.42**
Ensemble cluster analysis	**0.40**

Ensemble cluster analysis: higher bias, better generalization.

(1) Do we speculate on a better fit for D_{test}?

(2) Do we expect a significant covariate shift, more noise, etc. in D_{test}?

Constrained Cluster Analysis
Recap

1. Multi-document resolution can be tackled with constrained cluster analysis.
2. Constraints are derived from labeled examples.
3. Class membership function ties constraints to multiple retrieval models.
4. Advanced density-based clustering technology is key.

Constrained Cluster Analysis

References

- Disambiguating Web Appearances of People in a Social Network.
  [R. Bekkerman, A. McCallum. WWW 2005]

- A Bayesian Model for Supervised Clustering with the Dirichlet Process Prior.
  [H. Daumé III, D. Marcu. Journal MLR 2005]

- Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis.
  [E. Gabrilovich, S. Markovitch. IJCAI 2007]

- Unsupervised Discrimination of Person Names in Web Contexts.
  [T. Pedersen, A. Kulkarni. CICLing 2007]

- On Information Need and Categorizing Search.
  [S. Meyer zu Eissen. Dissertation, Paderborn University, 2007]

- Weighted Experts: A Solution for the Spock Data Mining Challenge.
  [B. Stein, S. Meyer zu Eissen. I-KNOW 2008]

- GRAPE: A System for Disambiguating and Tagging People Names in Web Search.
  [L. Jiang, W. Shen, J. Wang, N. An. WWW 2010]

