

Chapter DM:II (continued)

II. Cluster Analysis

- ❑ Cluster Analysis Basics
- ❑ Hierarchical Cluster Analysis
- ❑ Iterative Cluster Analysis
- ❑ Density-Based Cluster Analysis
- ❑ Cluster Evaluation
- ❑ Constrained Cluster Analysis

Constrained Cluster Analysis

Person Resolution Task

23Jordan - A Michael Jordan Tribute - Mozilla Firefox

http://www.23jordan.com/

Michael Jordan Video
Watch Michael Jordan Videos From The Leading TV Networks

Jordan @ Sale
New Arrivals. Rare Styles. Free Shipping Worldwide. Order Now!

Ads by Google

23JORDAN

A Michael Jordan Tribute

Home Forum UNC Career NBA Stats Winning Shots Achievements Biography Pictures

Ads by Google

Latest Michael Jordan News

Jordan @ Sale
New Arrivals. Rare Styles. Free Shipping Worldwide. Order Now!
www.VariantKids.com

Michael Jordan Shoes
Riesenauswahl zu Superpreisen
Michael Jordan Shoes
eBay.at

South Jordan UT Hotels
The Utah Hotel Site. Discount South Jordan UT Hotels.
utah-hotels.org/South-J

A+ quality, low price
worldwide promotion now low price sale famous

Michael Jordan - A look back!
As we look back at the year 2003, we will remember it as the last time we were able to see Michael Jordan play in the NBA. The greatest basketball player ever retired at age 40, for the third and final time. There were some memorable moments in Jordan's final NBA season. The 2003 All-Star game featured a final tribute to Michael Jordan, with a special half-time presentation performed by Mariah Carey. The Miami Heat retired his number, marking the first in sports history where another team retired a player's jersey in his honor. His two-year return in the NBA will never diminish his legacy. Jordan finished his career with 32,292 points, his career average 30.12 ppg is the best in NBA history. Thanks Michael for coming back one last time!
Also see: [Michael Jordan says goodbye one final time!](#)

What is your favorite Air Jordan?
With the Air Jordan XX3 just around the corner, many Jordan fans are wondering if this will be the last Air Jordan produced. This shoe is a must-have shoe if you're a collector. [Join our forum and discuss your favorite Air Jordans.](#) You can find everything in here regarding Air Jordans from the latest releases, the hottest collections, to your favorite Michael Jordan memories. Retros will also be harder to come by in 2008, as Jordan Brand prepares to release more special edition 2-pair Air Jordan packages. They will be very limited, similar to the "Defining Moments" and the "Old Love New Love" packages.

Relive Michael Jordan's greatest moments on DVD!



Constrained Cluster Analysis

Person Resolution Task

23Jordan - A Michael Jordan Tribute - Mozilla Firefox

http://www.23jordan.com/

Michael Jordan Video
Watch Michael Jordan Videos From The Leading TV Networks

Jordan @ Sale
New Arrivals. Rare Styles. Free Shipping Worldwide. Order Now!

23JORDAN
A Michael Jordan Tribute

Home Forum UNC Career NBA Stats Winning Shots Achievements Biography Pictures

Latest Michael Jordan News

Jordan @ Sale
New Arrivals. Rare Styles. Free Shipping Worldwide. Order Now!
www.VariantKids.com

Michael Jordan Shoes
Riesenauswahl zu Superpreisen
Michael Jordan Shoes
eBay.at

South Jordan UT Hotels
The Utah Hotel Site. Discount South Jordan UT Hotels.
utah-hotels.org/South-J

A+ quality, low price
worldwide promotion now low price sale famous

Michael Jordan - A look back!
As we look back at the year 2003, we will remember it as the last time we were able to see Michael Jordan play in the NBA. The greatest basketball player ever retired at age 40, for the third and final time. There were some memorable moments in Jordan's final NBA season. The 2003 All-Star game featured a final tribute to Michael Jordan, with a special half-time presentation performed by Mariah Carey. The Miami Heat retired his number, marking the first in sports history where another team retired a player's jersey in his honor. His two-year return in the NBA will never diminish his legacy. Jordan finished his career with 32,292 points, his career average 30.12 ppg is the best in NBA history. Thanks Michael for coming back one last time!
Also see: [Michael Jordan says goodbye one final time!](#)

What is your favorite Air Jordan?
With the Air Jordan XX3 just around the corner, many Jordan fans are wondering if this will be the last Air Jordan produced. This shoe is a must-have shoe if you're a collector. [Join our forum and discuss your favorite Air Jordans](#). You can find everything in here regarding Air Jordans from the latest releases, the hottest collections, to your favorite Michael Jordan memories. Retros will also be harder to come by in 2008, as Jordan Brand prepares to release more special edition 2-pair Air Jordan packages. They will be very limited, similar to the "Defining Moments" and the "Old Love New Love" packages.

Relive Michael Jordan's greatest moments on DVD!



Michael Jordan | EECS at UC Berkeley - Mozilla Firefox

http://www.eecs.berkeley.edu/Faculty/Homepages/jordan.htm

EECS
ELECTRICAL ENGINEERING AND COMPUTER SCIENCES
COLLEGE OF ENGINEERING UC Berkeley

Michael Jordan
Professor

Research Areas
Artificial Intelligence (AI)
Biosystems & Computational Biology (BIO)
Control, Intelligent Systems, and Robotics (CIR)
Signal Processing (SP)
Statistical Machine Learning

Research Centers
Center for Intelligent Systems (CIS)
Reliable, Adaptive and Distributed systems Laboratory (RAD Lab)

Biography
Michael Jordan is Professor in the Department of Electrical Engineering and Computer Science and the Department of Statistics at the University of California, Berkeley. He received his Masters from Arizona State University, and earned his PhD in 1985 from the University of California, San Diego. He was a professor at the Massachusetts Institute of Technology from 1988 to 1998. He has published over 250 research articles on topics in computer science, statistics, electrical engineering, molecular biology and cognitive science. His research in recent years has focused on probabilistic graphical models, kernel machines, nonparametric Bayesian methods and applications to problems in information retrieval, signal processing and bioinformatics. Prof. Jordan was named a Fellow of the American Association for the Advancement of Science (AAAS) in 2006. He is a Fellow of the IMS, a Fellow of the IEEE and a Fellow of the AAAI.

Selected Publications
• A. D'Aspremont, L. El Ghaoui, M. Jordan, and G. Leventhal. *Robust semidefinite programming for sparse PCA*

Contact Information
731 Soda Hall
jordan@cs

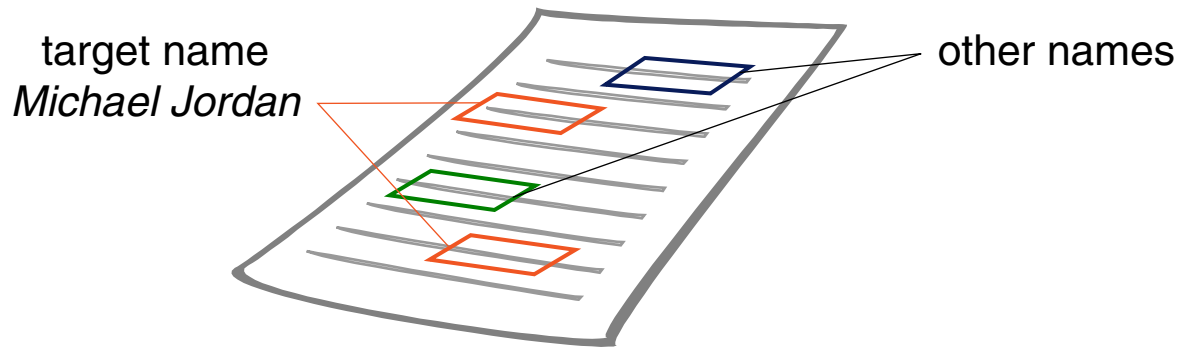
Office Hours
3:00-4:00, Thursday, 731 Soda
3:00-4:00, Tuesday, 401 Evans

Research Support Officer
Judy Tam
684 Soda
642-9575
jtam@erso

Personal Homepage
Projects
Publications
Dissertations

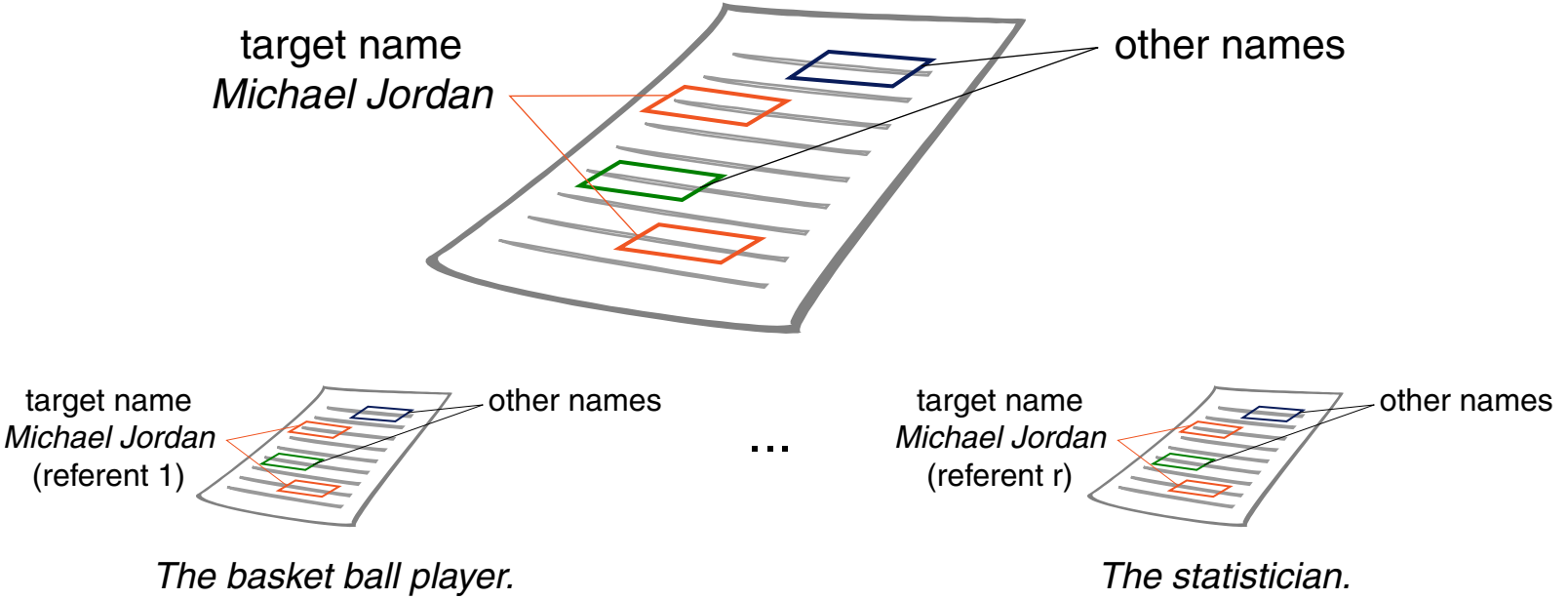
Constrained Cluster Analysis

Person Resolution Task



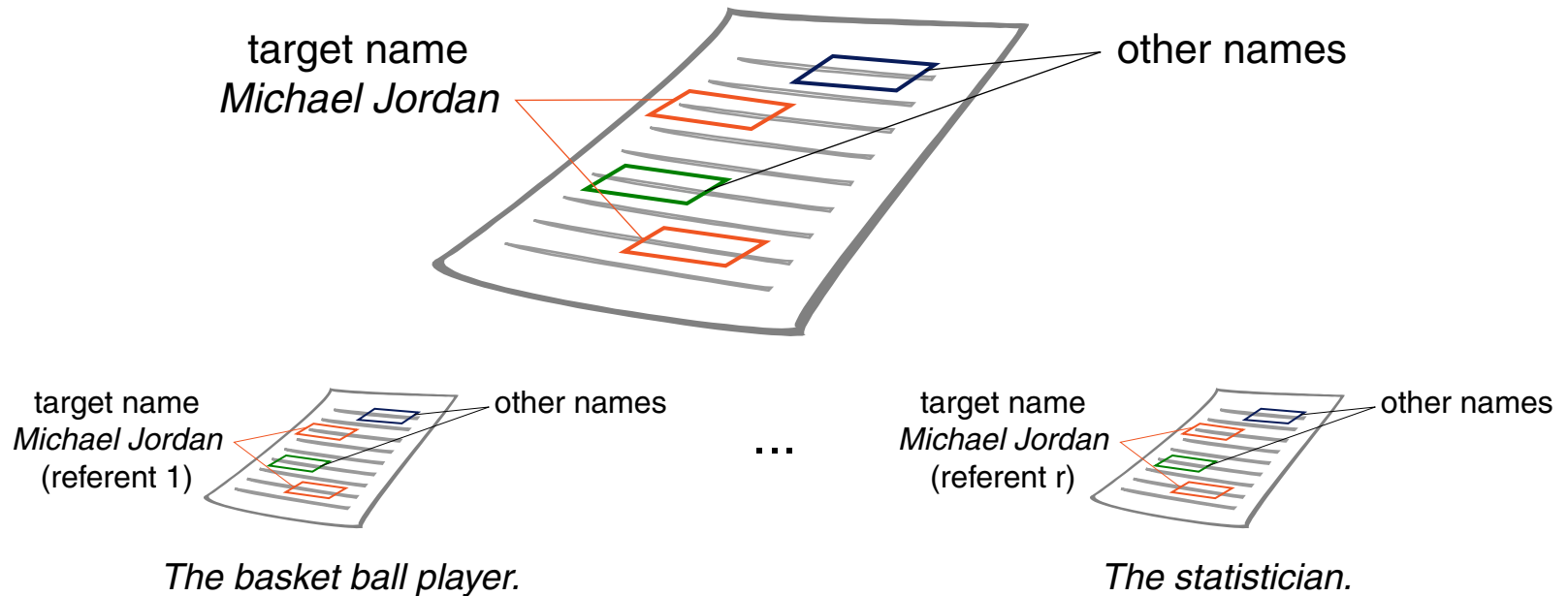
Constrained Cluster Analysis

Person Resolution Task



Constrained Cluster Analysis

Person Resolution Task



❑ Multi-document resolution task:

Names, Target names: $N = \{n_1, \dots, n_l\}, \quad T \subset N$

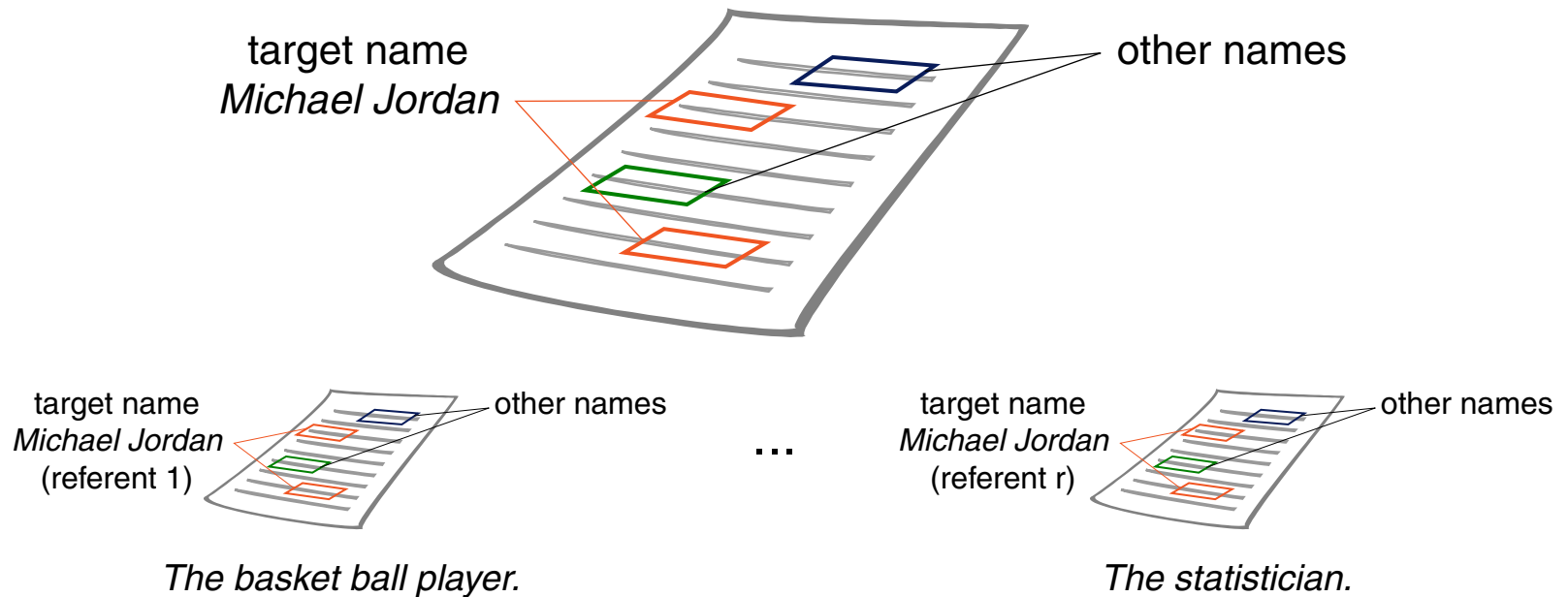
Referents: $R = \{r_1, \dots, r_m\}, \quad \tau : R \rightarrow T, \quad |R| \gg |T|$

Documents: $D = \{d_1, \dots, d_n\}, \quad \nu : D \rightarrow \mathcal{P}(N), \quad |\nu(d_i) \cap T| = 1$

A solution: $\gamma : D \rightarrow R, \quad \text{s.t. } \tau(\gamma(d_i)) \in \nu(d_i)$

Constrained Cluster Analysis

Person Resolution Task



❑ Facts about the Spock data mining challenge:

Target names: $|T| = 44$

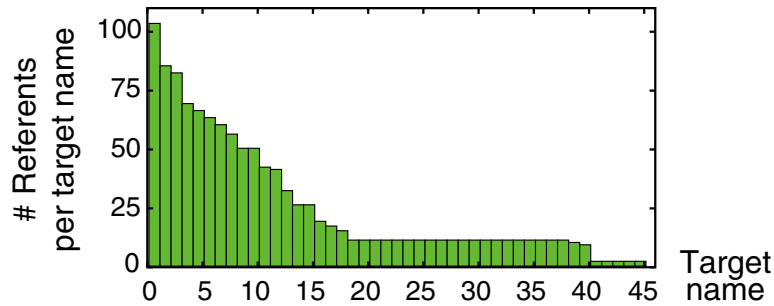
Referents: $|R| = 1\,101$

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

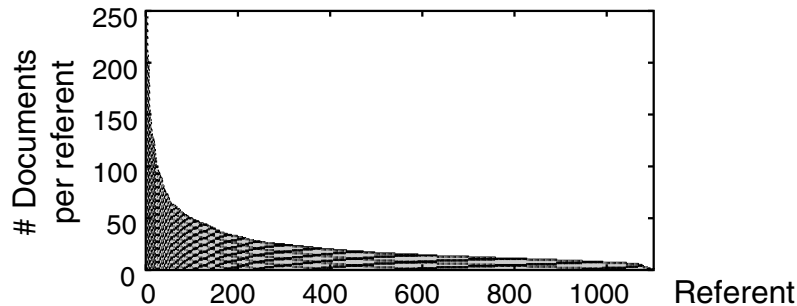
$|D_{test}| = 75\,000$ (unlabeled $\approx 7.8\text{GB}$)

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: $|T| = 44$

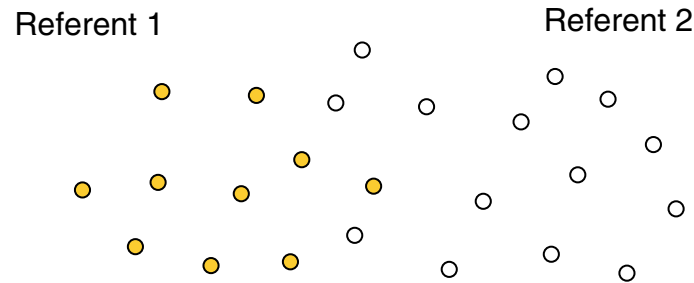
Referents: $|R| = 1\,101$

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

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

Constrained Cluster Analysis

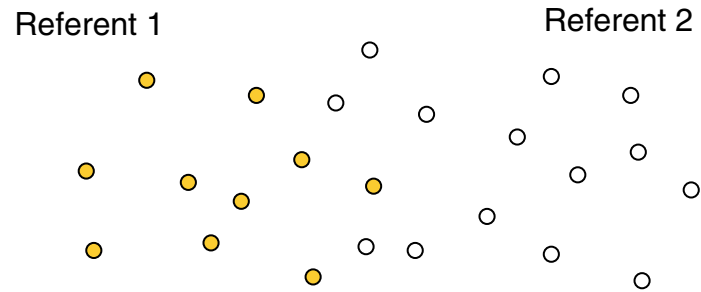
Applied to Multi-Document Resolution



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

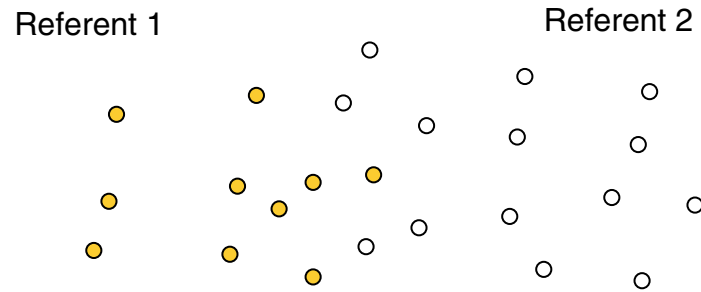
Applied to Multi-Document Resolution



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

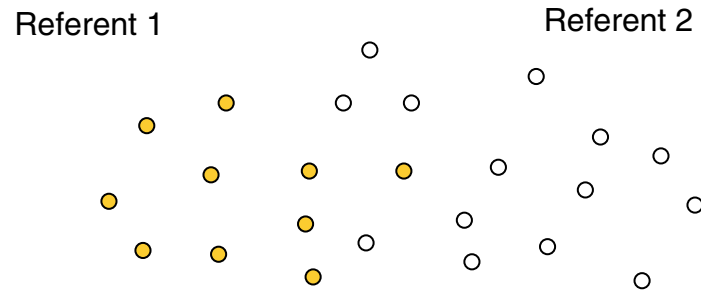
Applied to Multi-Document Resolution



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

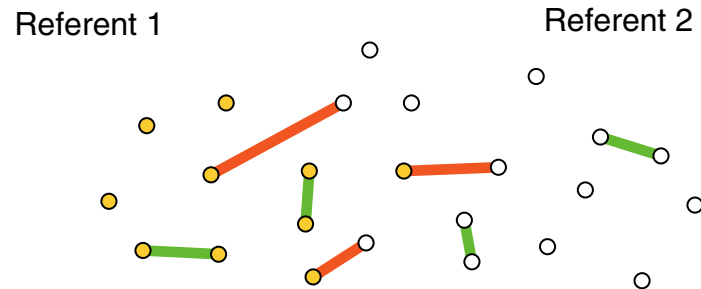
Applied to Multi-Document Resolution



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

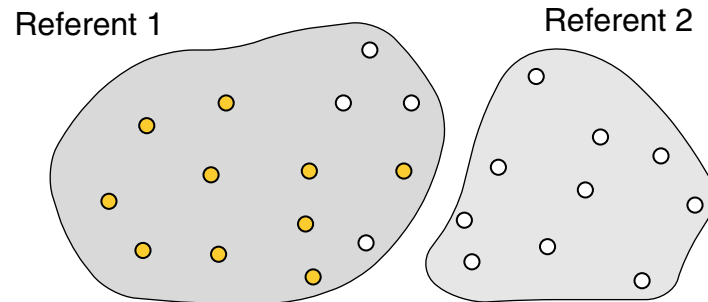
Applied to Multi-Document Resolution



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

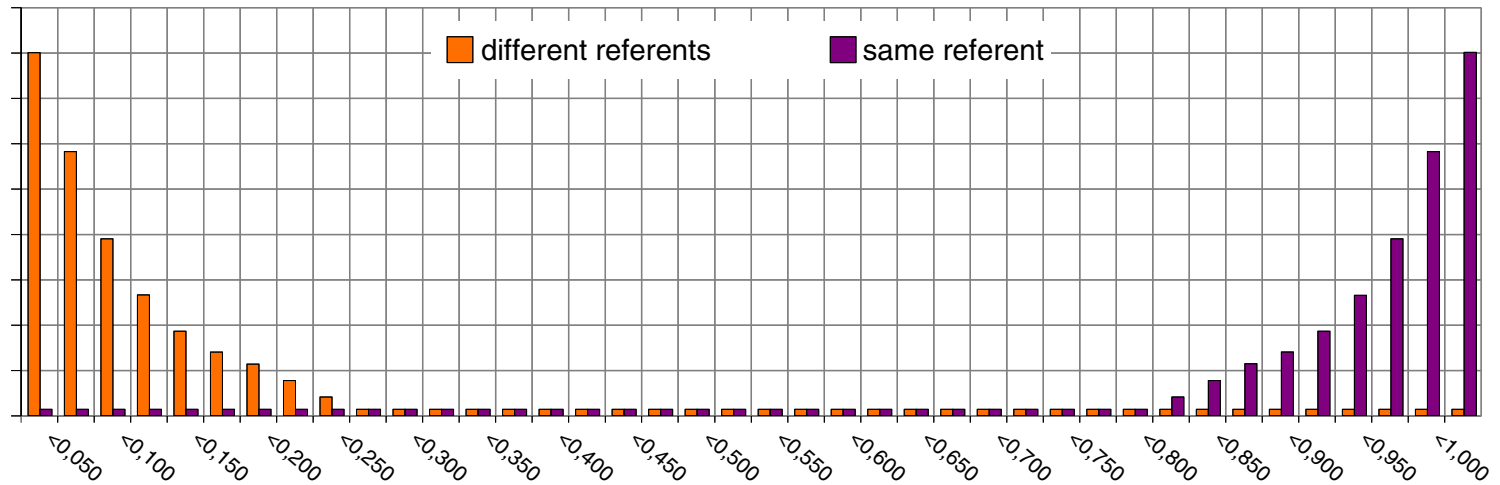
Applied to Multi-Document Resolution



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



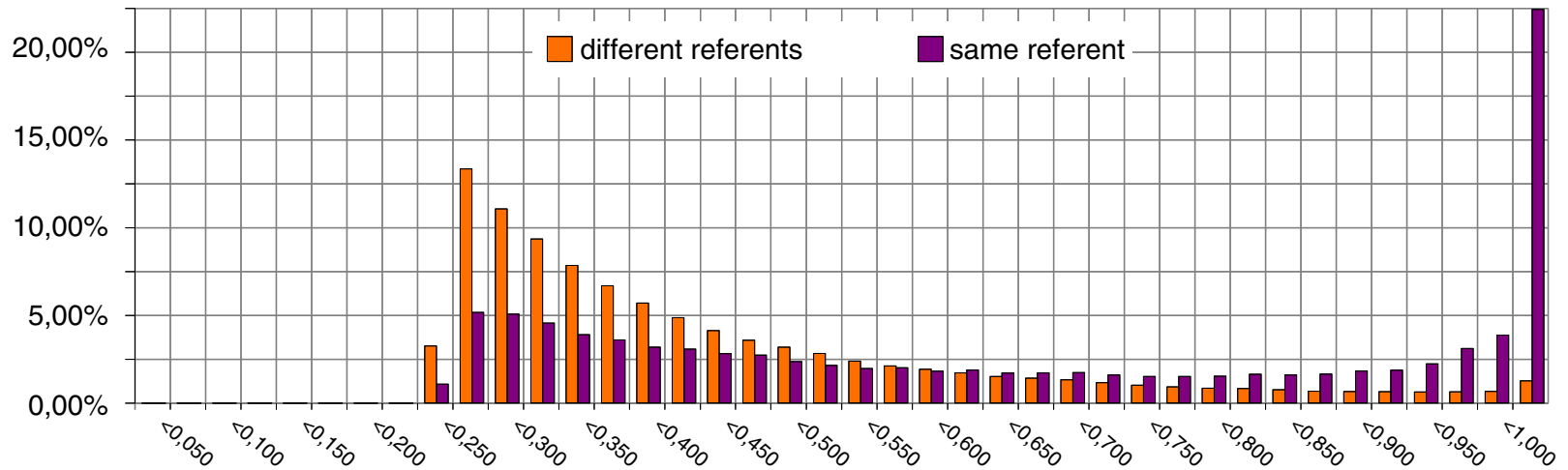
Similarity distributions for document pairs from **different referents** and **same referent**.

Logistic regression task:

- ❑ sample size: 400 000
- ❑ classes imbalance: **non-target class** : **target class** \approx 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

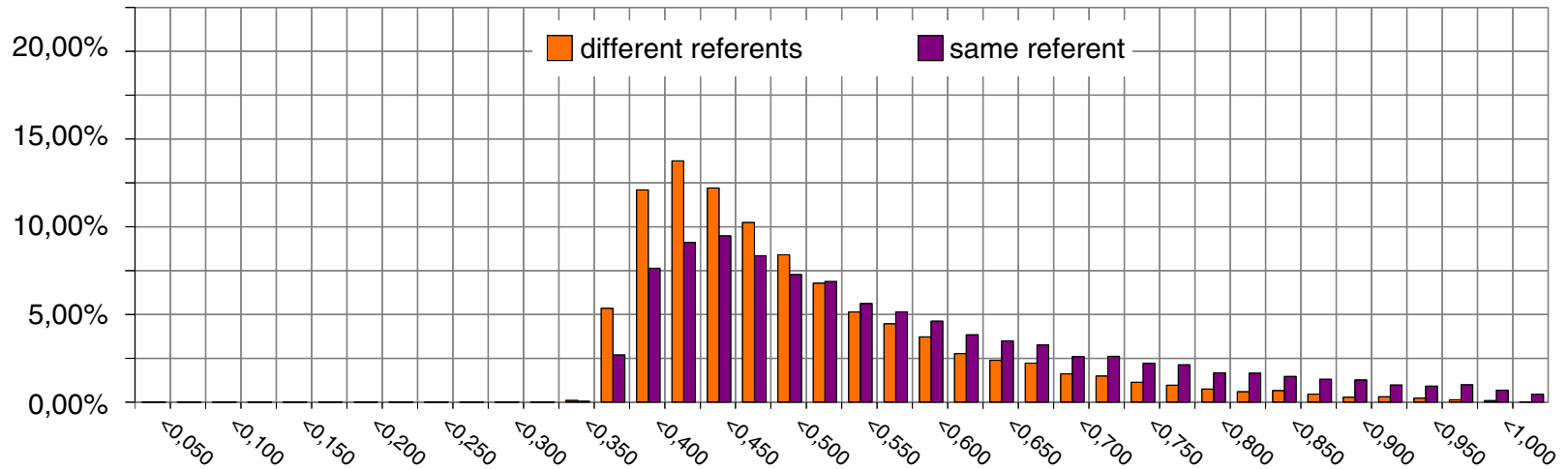


Model details:

- ❑ 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

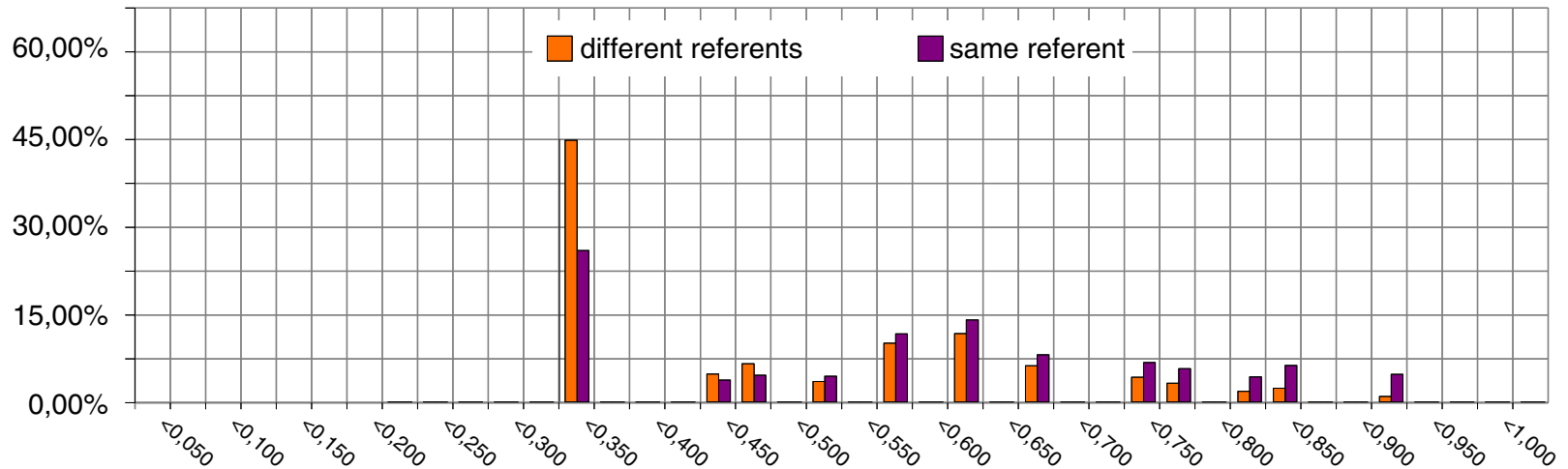


Model details:

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

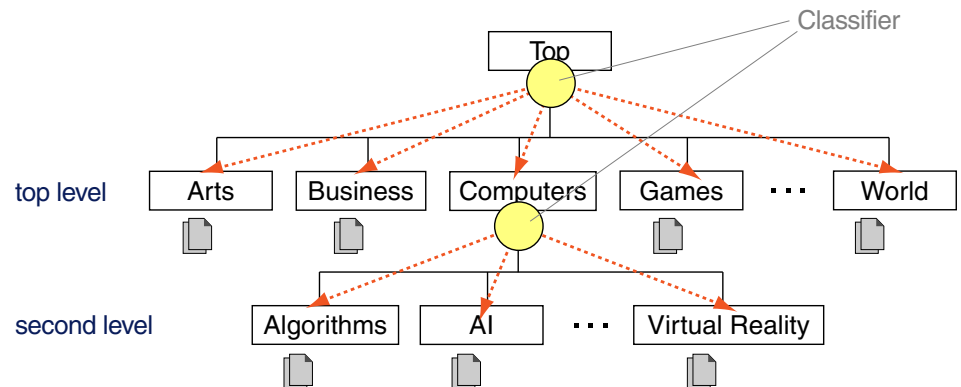
Constrained Cluster Analysis

Membership Distribution under Ontology Alignment Model



Model details:

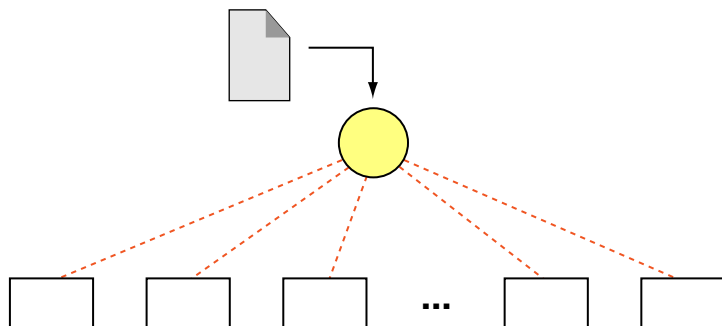
- ❑ DMOZ open directory project
- ❑ > 5 million documents
- ❑ 12 top-level categories
- ❑ 31 second level categories
- ❑ ML: hierarchical Bayes
- ❑ training set: 100 000 pages



Constrained Cluster Analysis

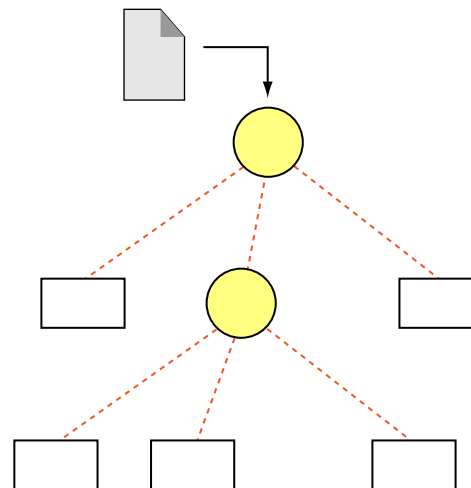
In-Depth: Multi-Class Hierarchical Classification

Flat (big-bang) classification



- + simple realization
- loss of discriminative power with increasing number of categories

Hierarchical (top-down) classification



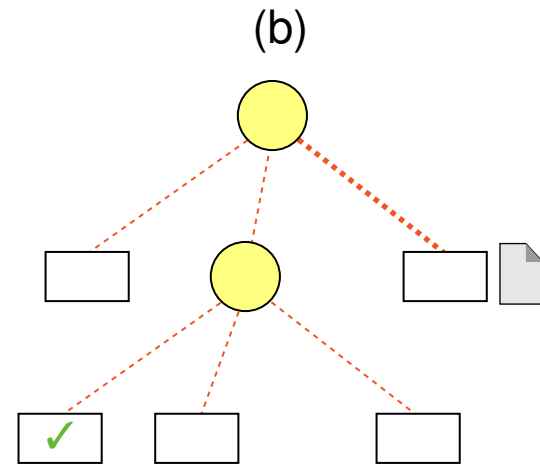
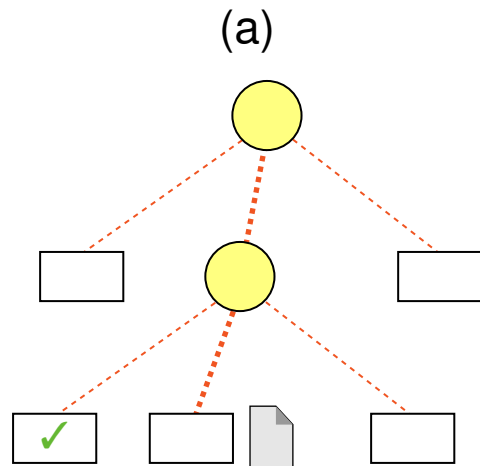
- + specialized classifiers (divide and conquer)
- misclassification at higher levels can never become repaired

Constrained Cluster Analysis

In-Depth: Multi-Class Hierarchical Classification

State of the art of effectiveness analyses:

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



Improvements:

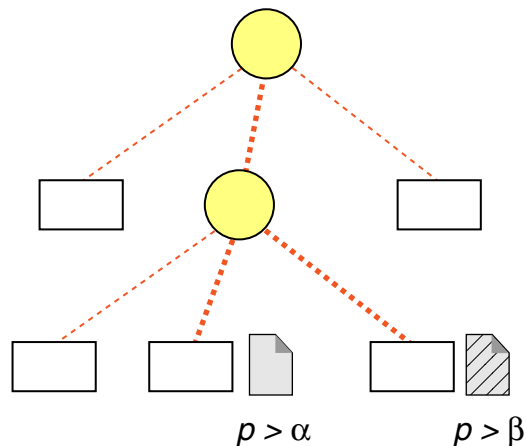
- ❑ Consider similarity $\varphi(C_i, C_j)$ between correct and wrong category.
- ❑ Consider graph distance $d(C_i, C_j)$ between correct and wrong category.

Constrained Cluster Analysis

In-Depth: Multi-Class Hierarchical Classification

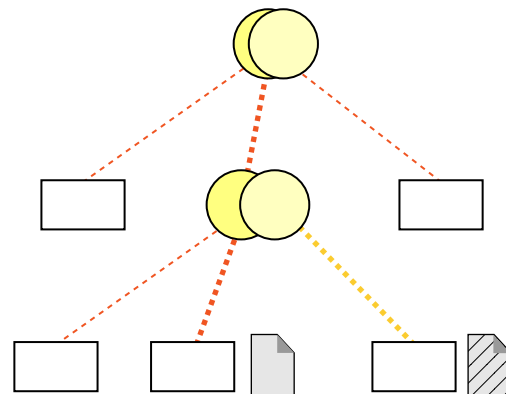
Improvements continued:

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

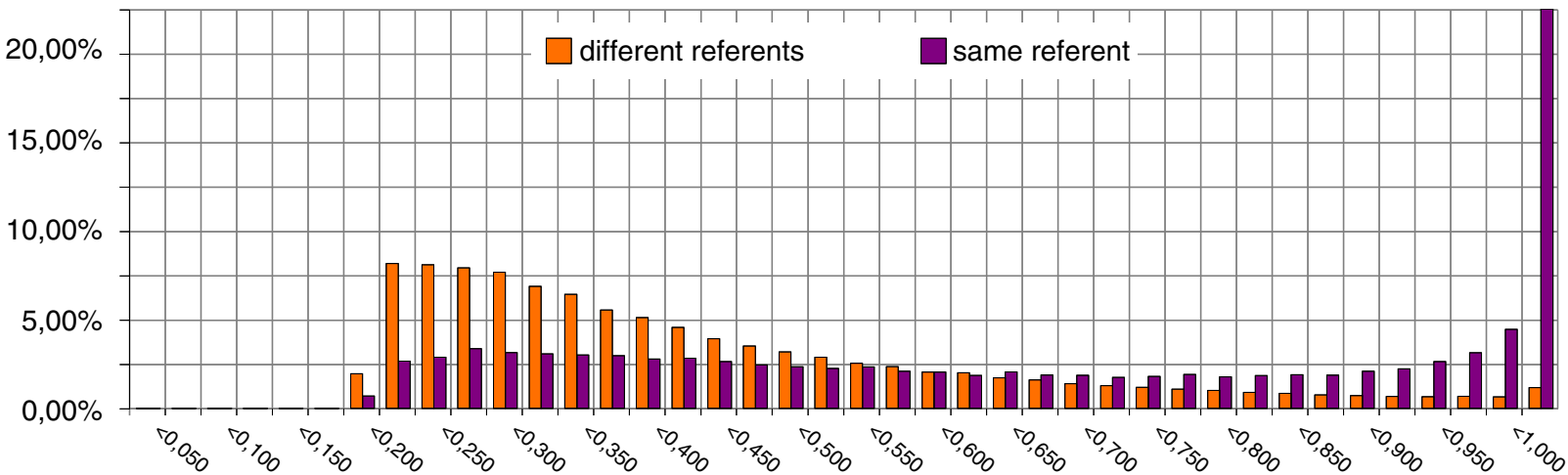
Multi-classifier (ensemble) classification



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

Constrained Cluster Analysis

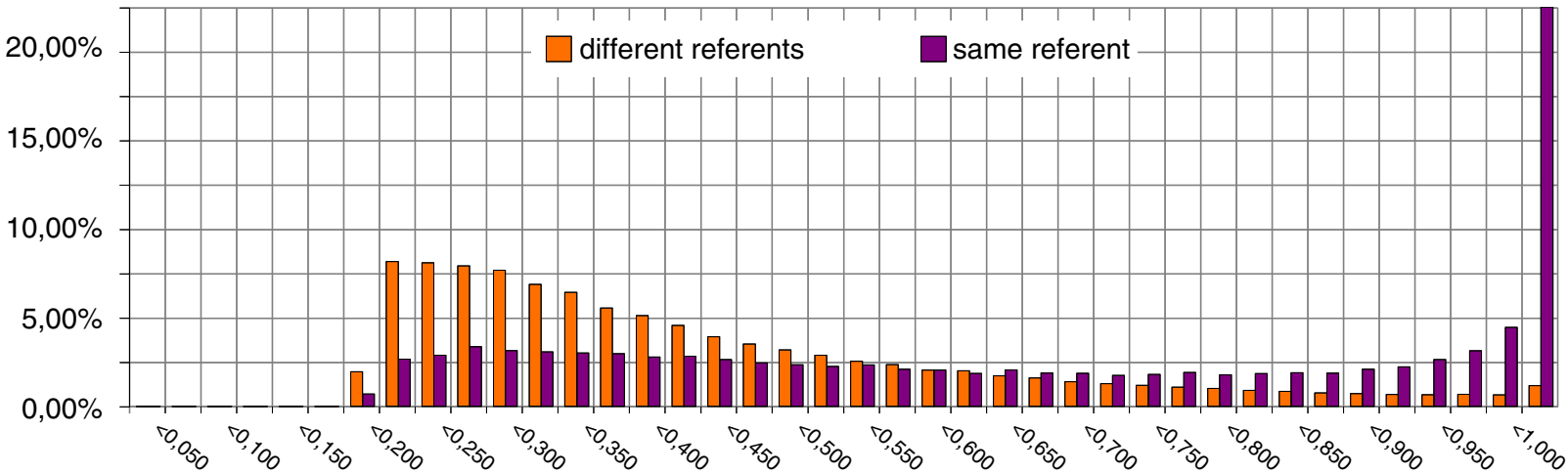
Membership Distribution under Optimized Retrieval Model Combination



Retrieval Model	$F_{1/3}$ -Measure
<i>tfidf</i> 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

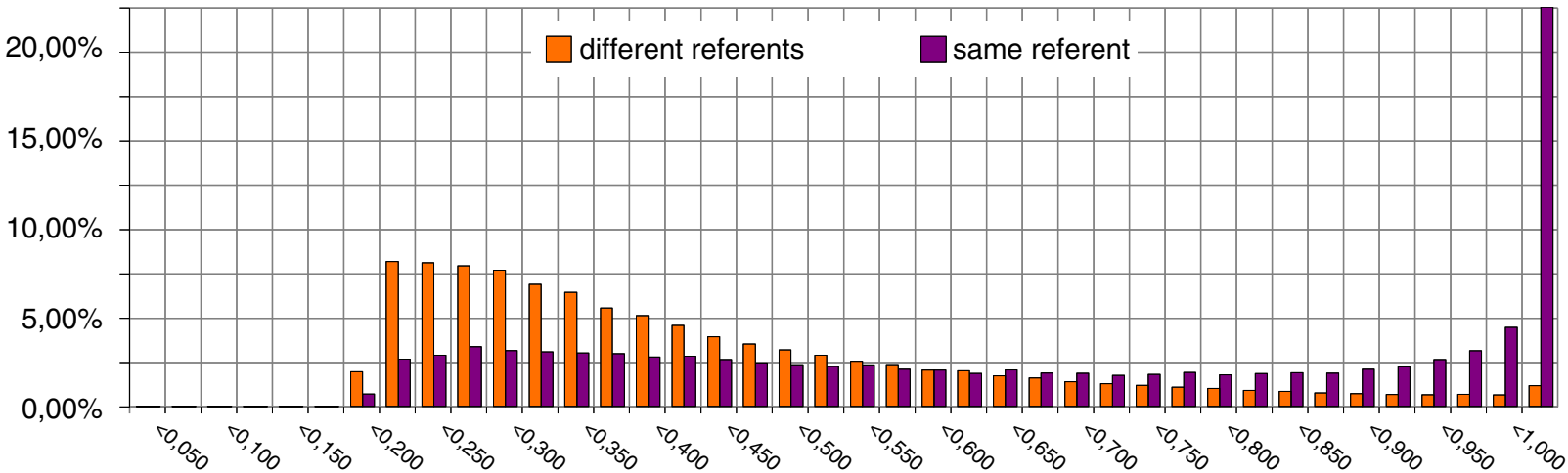


Retrieval Model	$F_{1/3}$ -Measure
<i>tfidf</i> vector space	0.39
context-based vector space	0.32
ESA Wikipedia persons	0.30
phrase structure grammar	0.17
ontology alignment	0.15
optimized combination	0.42



Constrained Cluster Analysis

Membership Distribution under Optimized Retrieval Model Combination

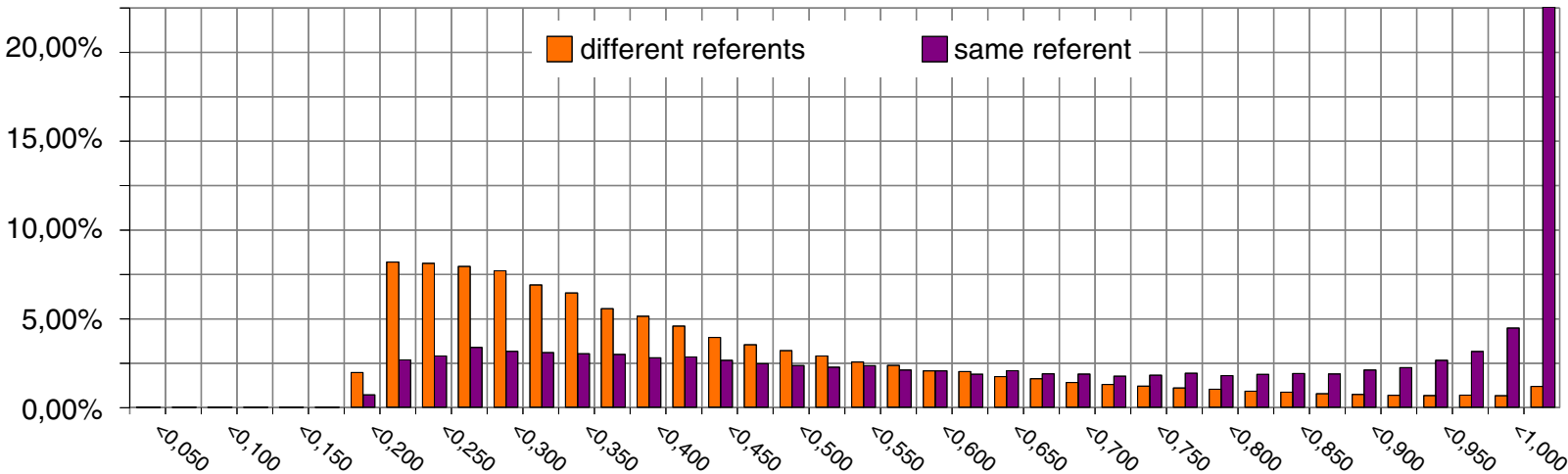


Retrieval Model	$F_{1/3}$ -Measure
<i>tfidf</i> vector space	0.39
context-based vector space	0.32
ESA Wikipedia persons	0.30
phrase structure grammar	0.17
ontology alignment	0.15
optimized combination	0.42

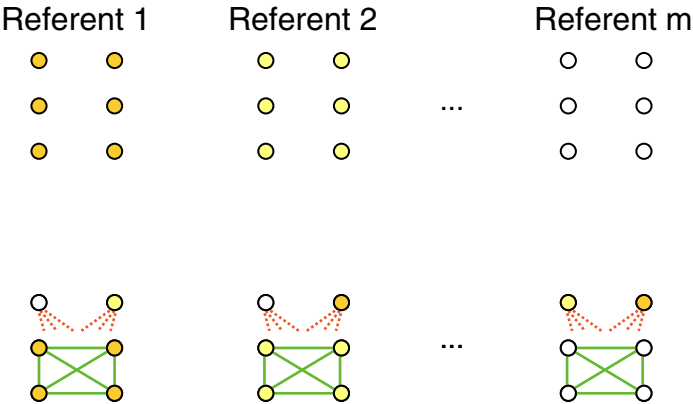
Referent 1		Referent 2		...	Referent m	
●	●	●	●		○	○
●	●	●	●	...	○	○
●	●	●	●		○	○
○	●	○	●		●	●
●	●	●	●	...	○	○
●	●	●	●		○	○

Constrained Cluster Analysis

Membership Distribution under Optimized Retrieval Model Combination

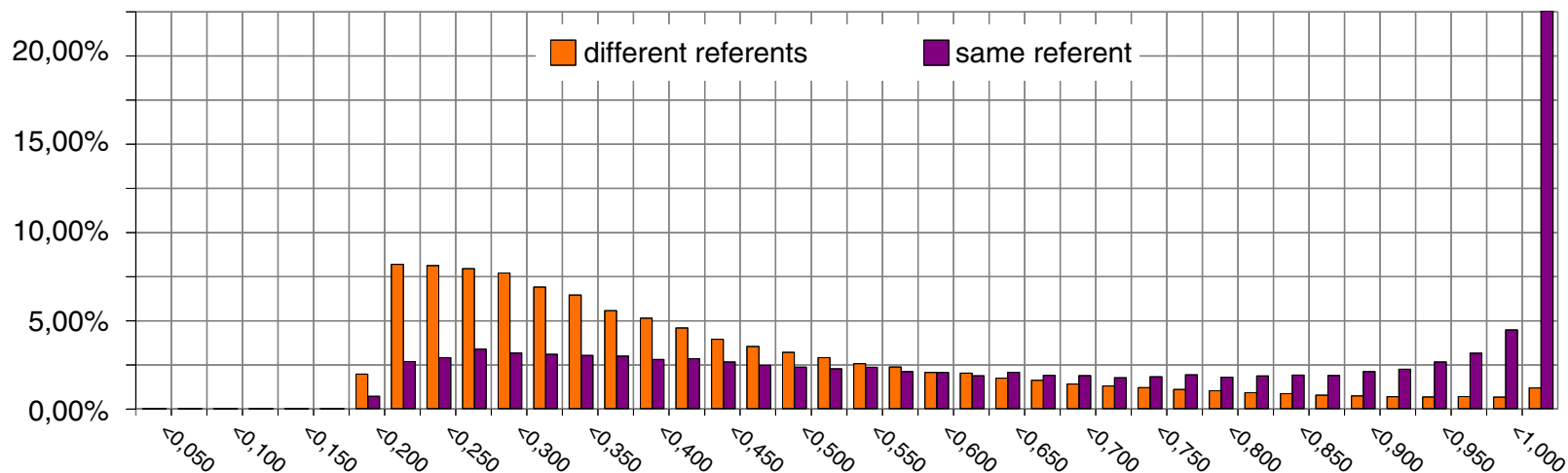


Retrieval Model	$F_{1/3}$ -Measure
<i>tfidf</i> vector space	0.39
context-based vector space	0.32
ESA Wikipedia persons	0.30
phrase structure grammar	0.17
ontology alignment	0.15
optimized combination	0.42



Constrained Cluster Analysis

Membership Distribution under Optimized Retrieval Model Combination



In the example:

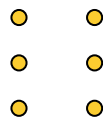
□ precision = 0.4

□ recall = 0.43

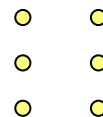
□ $F_{1/3} = 0.41$

(if false negatives are uniformly distributed)

Referent 1

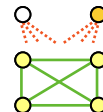
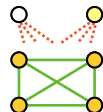
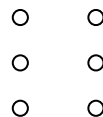


Referent 2

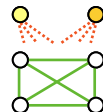


...

Referent m

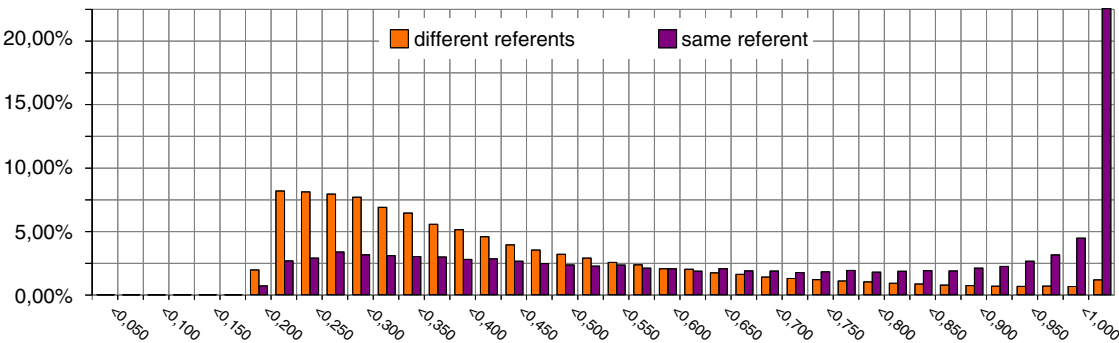


...

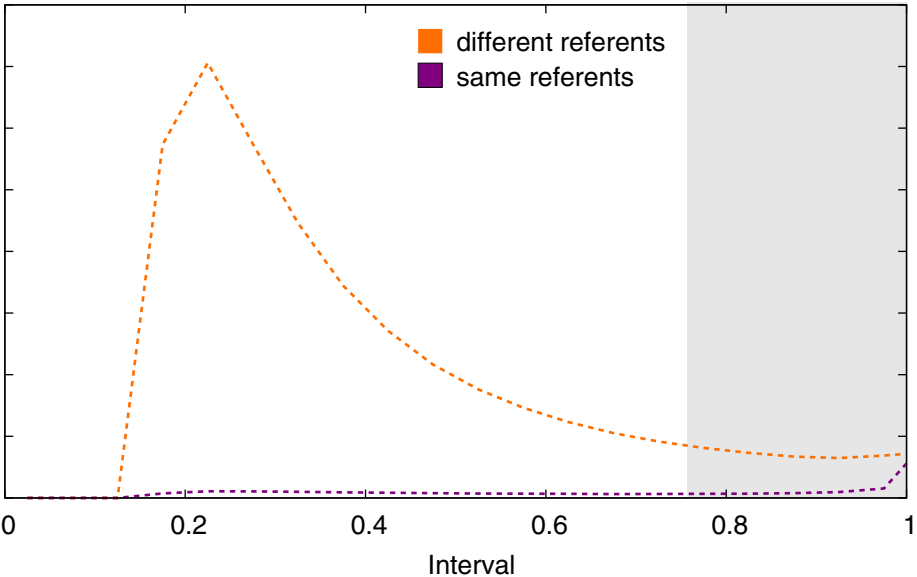


Constrained Cluster Analysis

In-Depth: Analysis of Classifier Effectiveness



Consideration of imbalance:



Constrained Cluster Analysis

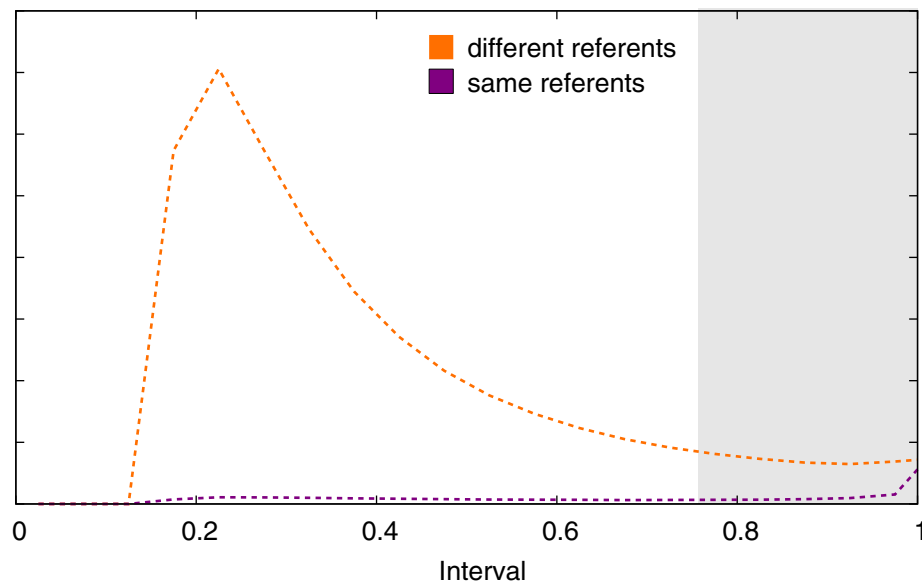
In-Depth: Analysis of Classifier Effectiveness

□ class imbalance factor (CIF) of 25

⇒ precision in interval $[0.725; 1]$ for edges between same referents: ≈ 0.17

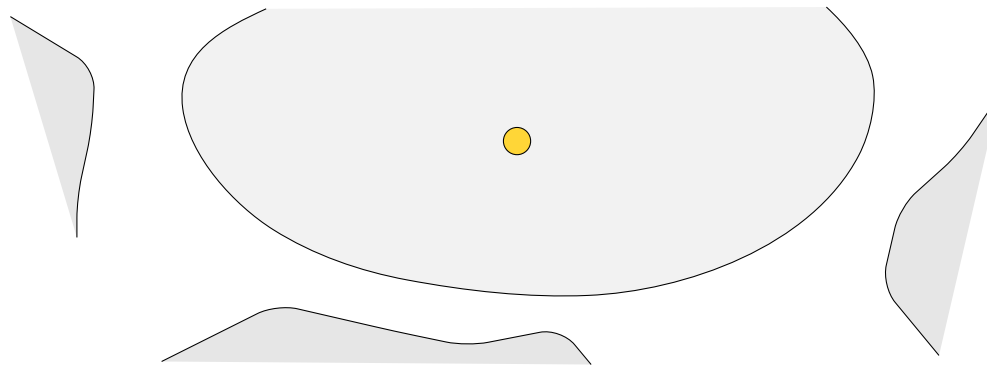
How can $F_{1/3} = 0.42$ be achieved via cluster analysis?

Consideration of imbalance:



Constrained Cluster Analysis

In-Depth: Analysis of Classifier Effectiveness



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

⇒ $|TP|$ 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: $E(|FP^*|) = 2.7$)

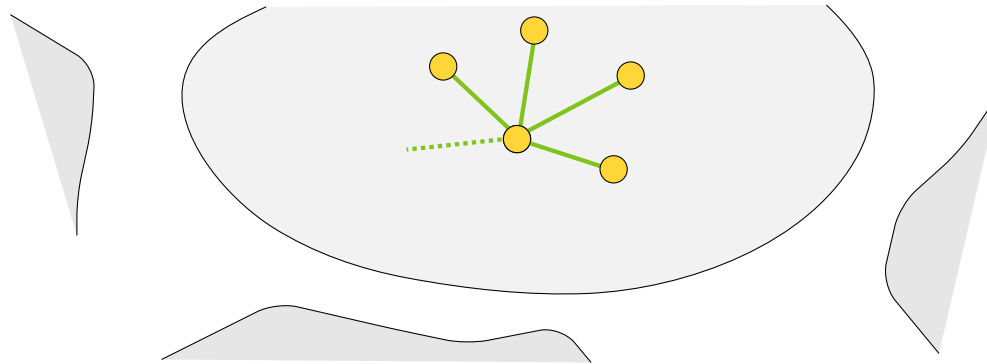
⇒ edge tie factor (ETF) specifies the excess of true positives until tie (here: 3...5)

$$ETF = \frac{|TP|}{|C| \cdot E(|FP^*|)},$$

$$\text{effective precision} = \text{precision} \cdot \frac{CIF}{ETF}$$

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: $E(|FP^*|) = 2.7$)

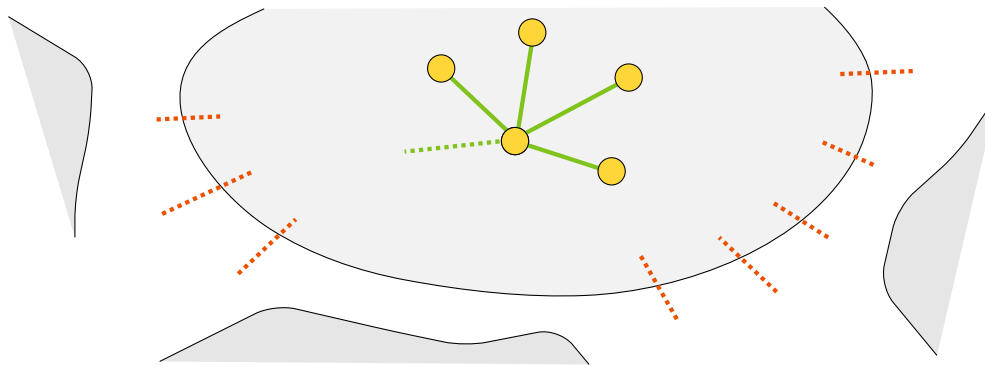
⇒ edge tie factor (ETF) specifies the excess of true positives until tie (here: 3...5)

$$ETF = \frac{|TP|}{|C| \cdot E(|FP^*|)},$$

$$\text{effective precision} = \text{precision} \cdot \frac{CIF}{ETF}$$

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: $E(|FP^*|) = 2.7$)

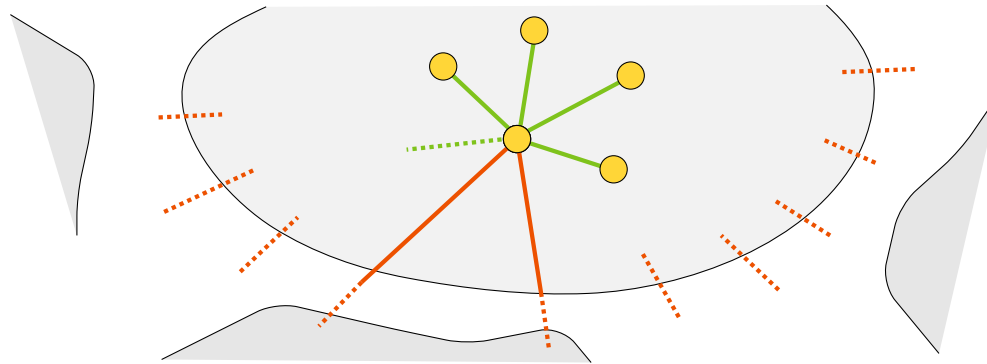
⇒ edge tie factor (ETF) specifies the excess of true positives until tie (here: 3...5)

$$ETF = \frac{|TP|}{|C| \cdot E(|FP^*|)},$$

$$\text{effective precision} = \text{precision} \cdot \frac{CIF}{ETF}$$

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: $E(|FP^*|) = 2.7$)

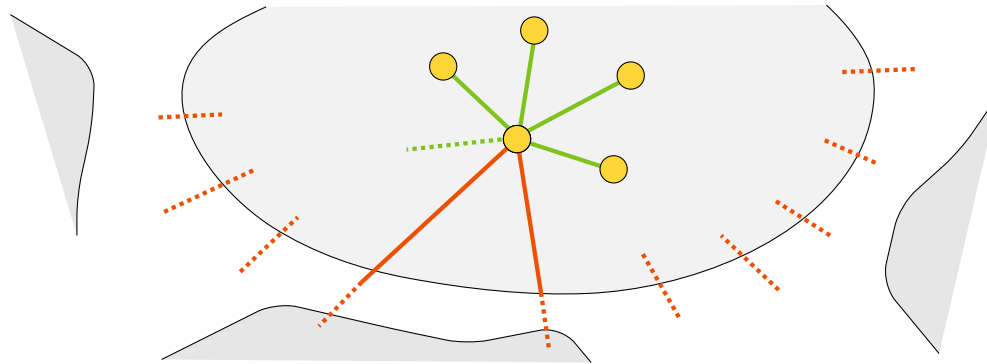
⇒ edge tie factor (ETF) specifies the excess of true positives until tie (here: 3...5)

$$ETF = \frac{|TP|}{|C| \cdot E(|FP^*|)},$$

$$\text{effective precision} = \text{precision} \cdot \frac{CIF}{ETF}$$

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: $E(|FP^*|) = 2.7$)

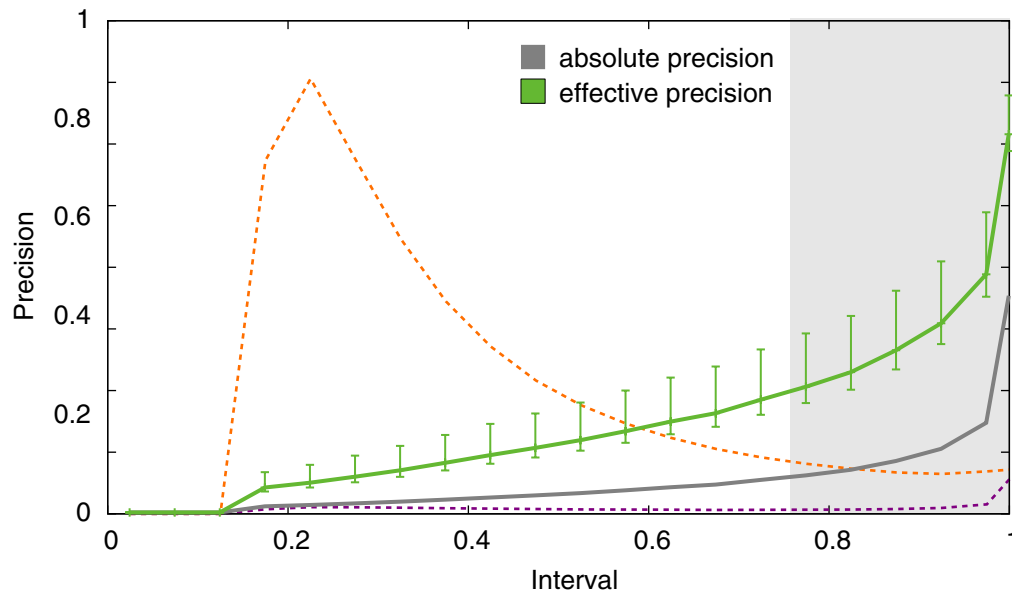
⇒ edge tie factor (ETF) specifies the excess of true positives until tie (here: 3...5)

$$ETF = \frac{|TP|}{|C| \cdot E(|FP^*|)},$$

$$\text{effective precision} = precision \cdot \frac{CIF}{ETF}$$

Constrained Cluster Analysis

In-Depth: Analysis of Classifier Effectiveness



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

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

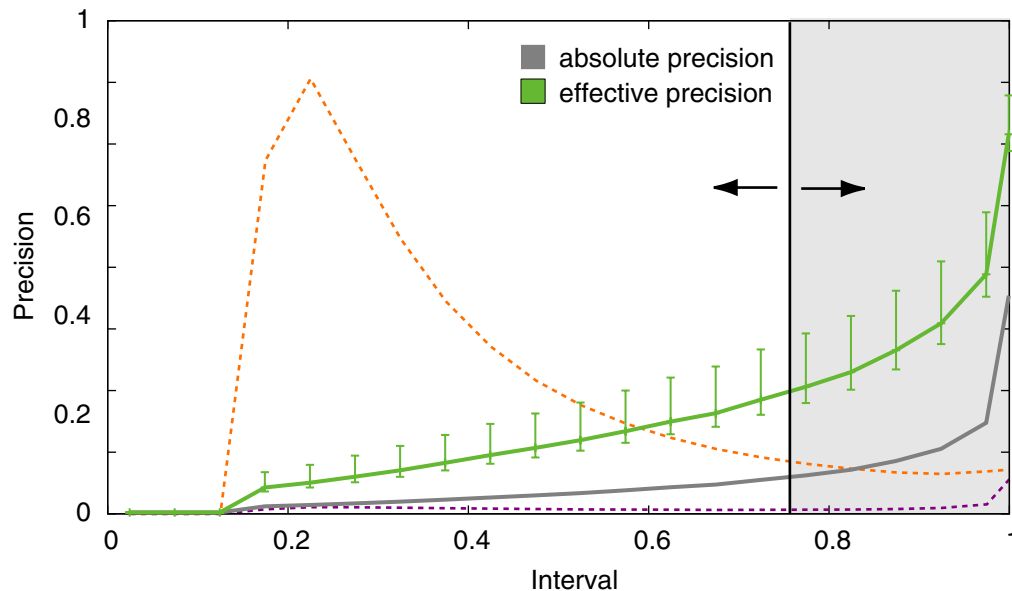
⇒ edge tie factor (ETF) specifies the excess of true positives until tie (here: $3 \dots 5$)

$$ETF = \frac{|TP|}{|C| \cdot E(|FP^*|)},$$

$$\text{effective precision} = \text{precision} \cdot \frac{CIF}{ETF}$$

Constrained Cluster Analysis

In-Depth: Analysis of Classifier Effectiveness



Determine optimum similarity threshold for class-membership function:

$$\theta^* = \operatorname{argmax}_{\theta \in [0;1]} \left\{ \frac{1 + \alpha}{\frac{ETF}{\text{precision}_\theta \cdot CIF} + \frac{\alpha}{\text{recall}_\theta}} \right\}$$

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
<i>tfidf</i> 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]