A Machine Learning Approach for Adaptive Information Extraction

Edgar Gonzàlez i Pellicer
Director:Jordi Turmo i Borràs

June 2006

Index

Index

Information Extraction

Information Extraction (IE) is a Natural Language Processing (NLP) area whose goal is to automatically generate structured pieces of information from the relevant information contained in text documents.

Information Extraction

- Structured, semi-structured or free text.
- Documents from a restricted domain.
- A **Scenario of Extraction** is defined a priori:
  - Set of concepts considered relevant.

Sample Scenario of Extraction

An **aircraft crash** *is an event happening in a certain* **site** *and* **date**. *It involves an* **aircraft**, *which is of a* **model**, *belongs to an* **airline** *and is made by a* **manufacturer**. *The* **aircraft** *covers a* **flight**, *taking off from a* **departure** *and landing in a* **destination**.

Sample Document

New York Times - 96/02/07 - 07:22

SEATTLE - It's the phone call no one wants to get, but everyone knows might come one day. It came late Tuesday when Boeing got word that a chartered 757 aircraft crashed shortly after takeoff from the Dominican Republic. All 189 passengers are feared dead. The crash, only the second in the history of the Boeing 757, came less than two months after an American Airlines 757 slammed into a mountain as it approached Cali, Colombia. Four people survived the Dec. 20 crash that killed 160 people. The cause has not yet been determined. After hearing the news of Alas Nacionales Flight 301 Tuesday night, members of Boeing's Air Safety Investigation Group monitored the situation throughout the night and quickly assembled a team of safety experts to be on standby in case they were needed at the crash scene. One Boeing air safety investigator was expected to arrive Thursday in Puerto Plata to assist a team from the National Transportation Safety Board and the Dominican Republic in trying to determine why the two-engine jet crashed. More Boeing engineers will be called in if needed. ...

Sample Document

New York Times - 96/02/07 - 07:22

SEATTLE - It's the phone call no one wants to get, but everyone knows might come one day. It came late Tuesday when Boeing got word that a chartered 757 aircraft crashed shortly after takeoff from the Dominican Republic. All 189 passengers are feared dead. The crash, only the second in the history of the Boeing 757, came less than two months after an American Airlines 757 slammed into a mountain as it approached Cali, Colombia. Four people survived the Dec. 20 crash that killed 160 people. The cause has not yet been determined. After hearing the news of Alas Nacionales Flight 301 Tuesday night, members of Boeing's Air Safety Investigation Group monitored the situation throughout the night and quickly assembled a team of safety experts to be on standby in case they were needed at the crash scene. One Boeing air safety investigator was expected to arrive Thursday in Puerto Plata to assist a team from the National Transportation Safety Board and the Dominican Republic in trying to determine why the two-engine jet crashed. More Boeing engineers will be called in if needed. ...

Linguistic Knowledge for IE

- IE aims at *deep* understanding of texts.
  - Identify relevant information and extract it.
- Systems require significant amounts of linguistic knowledge.
  - Linguistic patterns.
  - Domain, writing style and language specific.
- The acquisition of these patterns and the other required linguistic knowledge can become highly expensive.

Machine Learning for IE

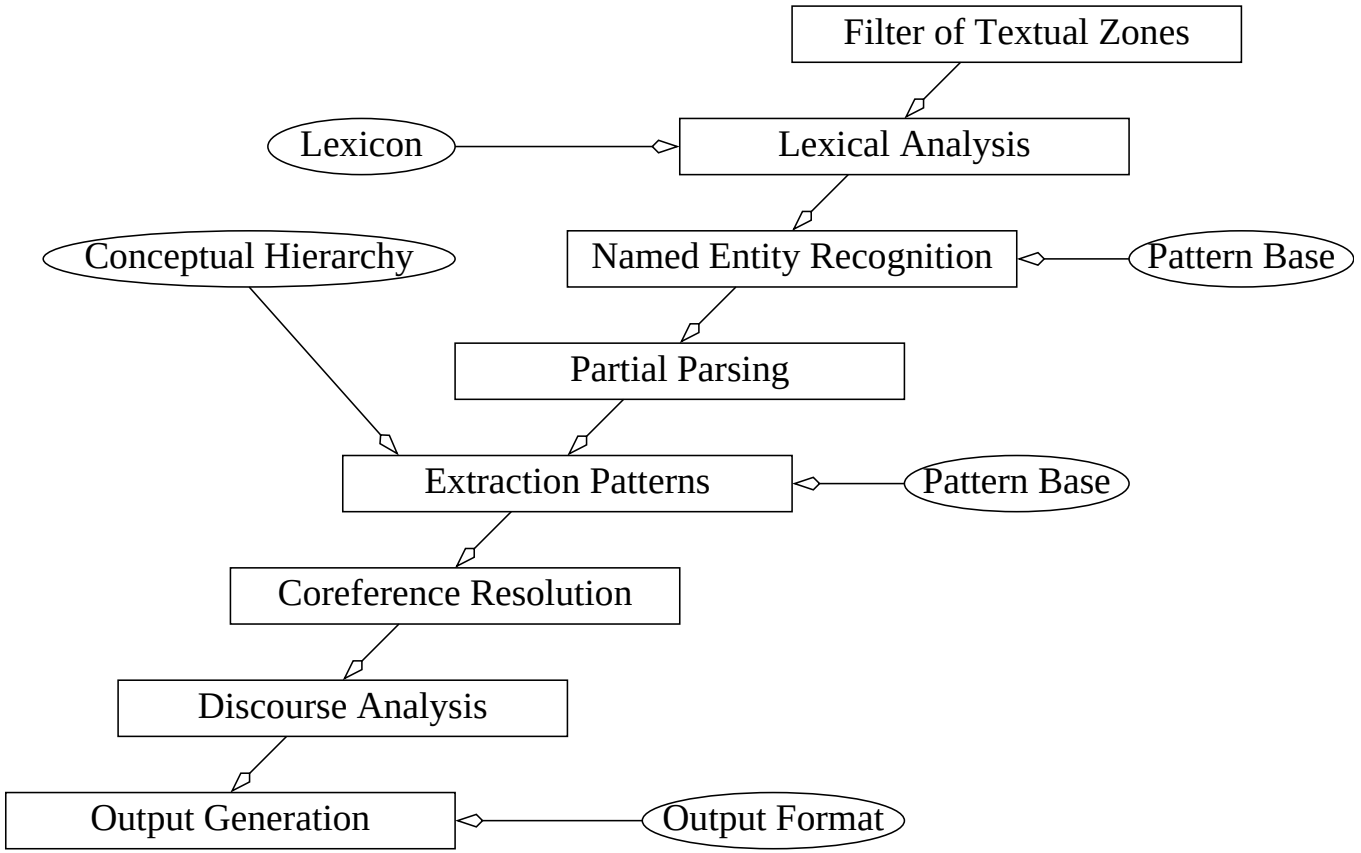
- ML approaches allow the acquisition of knowledge for IE at a lower cost.
- From Supervised Methods to Lightly Supervised Methods.
  - Reduce Elements of Supervision.

Enhancing with Clustering

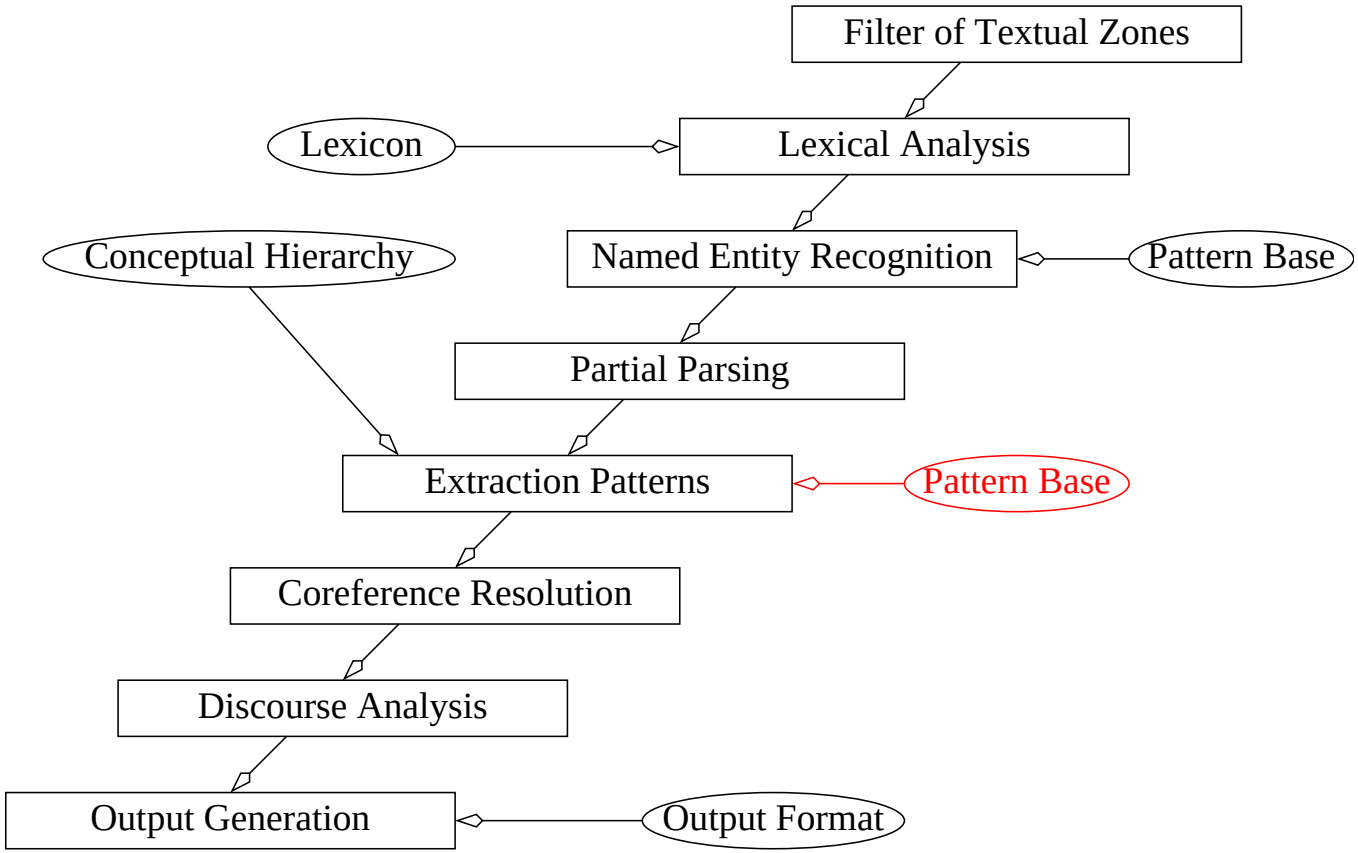
- Use document clustering techniques to further reduce the elements of human supervision.
- Develop a methodology:
  - No annotation.
  - No seeding.
  - Producing good quality IE patterns.
    - For IE and other NLP tasks.

Index

Usual architecture of an IE system



Usual architecture of an IE system



Motivation

- Reduce the cost of the process of knowledge acquisition.
  - Many IE systems use hand-coding by human experts.
- Ease portability to different languages, domains and writing styles.
- Benefit from the increasing number of available document collections.
  - Many of them unannotated.

Knowledge Representations for IE

- Hidden Markov Models.
- Relational Markov Networks.
- Dynamic Bayesian Networks.
- Conditional Random Fields.
- Hyperplane Separators.
- Rules.
  - Linguistic Pattern $\longrightarrow$ Extraction

Knowledge Representations for IE

- Hidden Markov Models.
- Relational Markov Networks.
- Dynamic Bayesian Networks.
- Conditional Random Fields.
- Hyperplane Separators.
- Rules.
  - Linguistic Pattern $\longrightarrow$ Extraction

Why Rules?

- Popular approach.
- Explicit representation of knowledge.
  - Allows inspection (and modification) by users.
  - Allows post-processing.

Sample Rules

- PROTEUS System.

`np(person)` , **x : integer** `years old` , $\rightarrow$ *age* = **x**

... John Doe, 40 years old, ...

`np(person)` , **x : np(position)** , $\rightarrow$ *position* = **x**

... John Doe, General Director, ...

Sample Rules

- PROTEUS System.

`np(person)` , **x : integer** `years old` , $\rightarrow$ *age* = **x**

... John Doe, 40 years old, ...

`np(person)` , **x : np(position)** , $\rightarrow$ *position* = **x**

... John Doe, General Director, ...

Sample Rules

- PROTEUS System.

`np(person)` , **x : integer** `years old` , $\longrightarrow$ *age* = **x**

... John Doe, 40 years old, ...

`np(person)` , **x : np(position)** , $\longrightarrow$ *position* = **x**

... John Doe, General Director, ...

Sample Rules

- PROTEUS System.

`np(person)` , **x : integer** `years old` , $\longrightarrow$ *age* = **x**

... John Doe, **40** years old, ...

`np(person)` , **x : np(position)** , $\longrightarrow$ *position* = **x**

... John Doe, General Director, ...

Sample Rules

- PROTEUS System.

`np(person)` , **x : integer** `years old` , $\longrightarrow$ *age* = **x**

... John Doe, **40** years old, ...

`np(person)` , **x : np(position)** , $\longrightarrow$ *position* = **x**

... John Doe, General Director, ...

Sample Rules

- PROTEUS System.

`np(person)` , **x : integer** `years old` , $\rightarrow$ *age* = **x**

... John Doe, **40** years old, ...

`np(person)` , **x : np(position)** , $\rightarrow$ *position* = **x**

... John Doe, General Director, ...

Sample Rules

- PROTEUS System.

`np(person)` , **x : integer** `years old` , $\rightarrow$ *age* = **x**

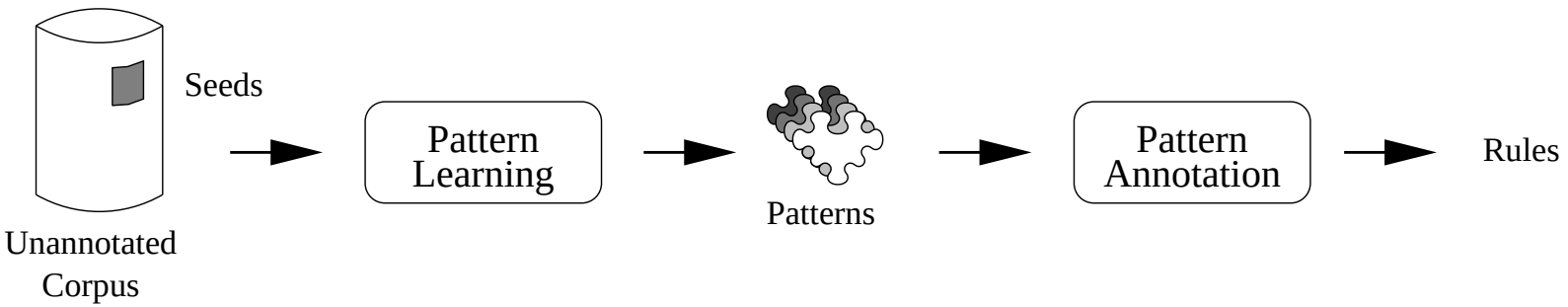
... John Doe, **40** years old, ...

`np(person)` , **x : np(position)** , $\rightarrow$ *position* = **x**

... John Doe, **General Director**, ...

Rule Learning Systems

- Supervised.
- Lightly Supervised.



Elements of *Light* Supervision

- Classification of each Document as Relevant or Not.
- Context Keywords (*name*, *elect*, *CEO*, *company*).
- Word Senses ({*CEO*, *chief executive officer*}).
- Seeds.
  - Words (*Steve Jobs*).
  - Relations (*Steve Jobs - Apple Computer*).
  - Patterns ({*PER*}, *CEO of* {*ORG*}).
  - Documents.

Drawbacks of Lightly Supervised Methods

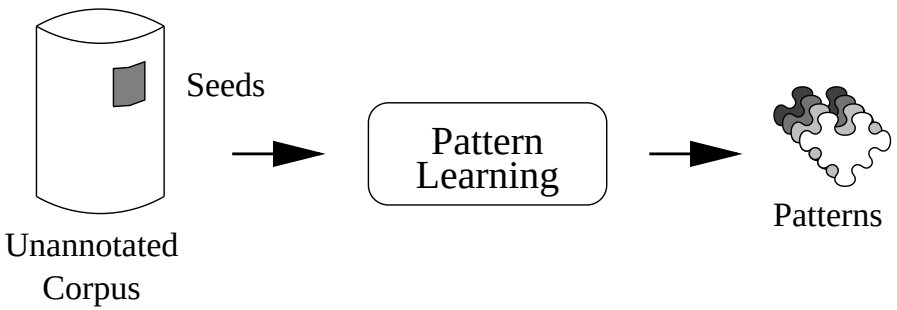
- The supervision requires the user browsing and exploring through an unannotated corpus.
- The user is introducing a strong bias.
- The selection of seed documents relevant to different domains requires an a priori definition of the domains.
- It also requires the availability of collections of documents relevant to each IE task.

Index

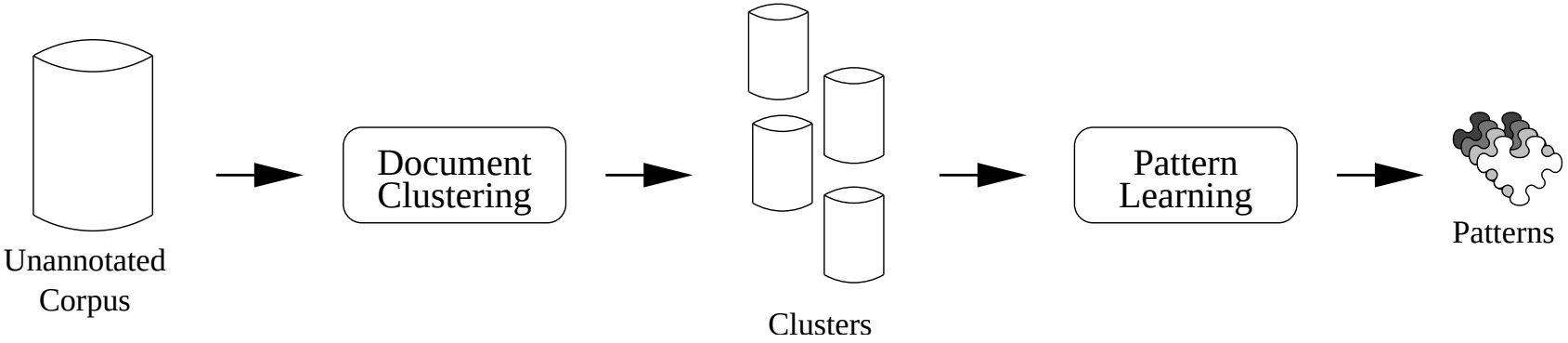
Our Proposal

Combine Document Clustering and Pattern Learning

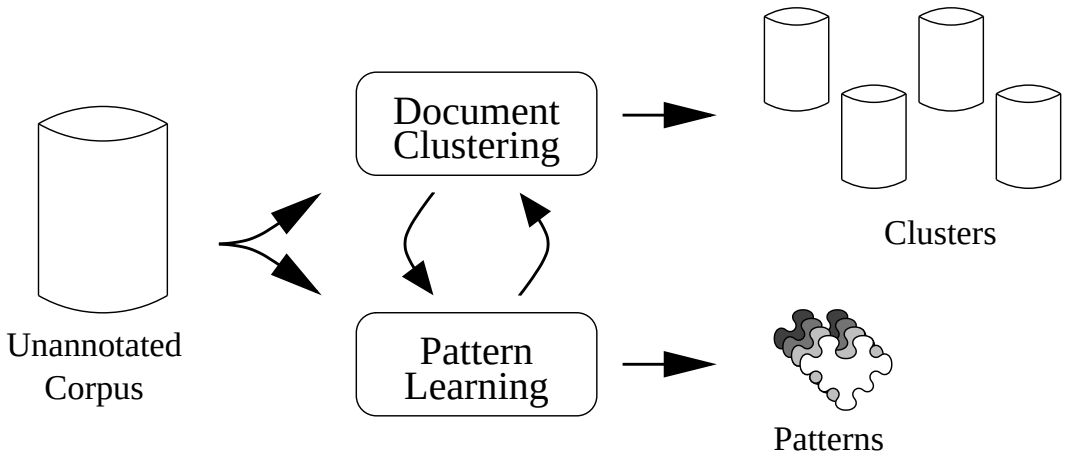
Manual Seeding for Pattern Learning



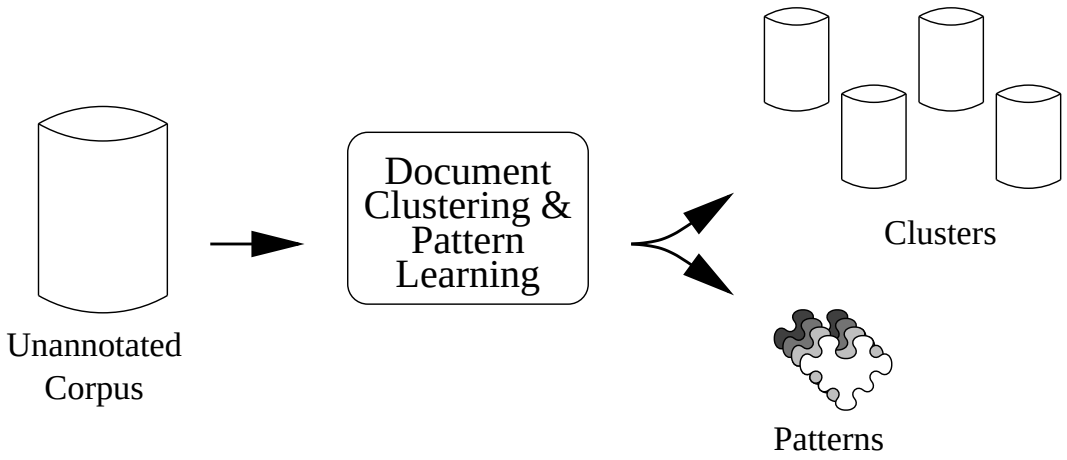
Sequential Clustering and Pattern Learning



Collaborative Clustering and Pattern Learning



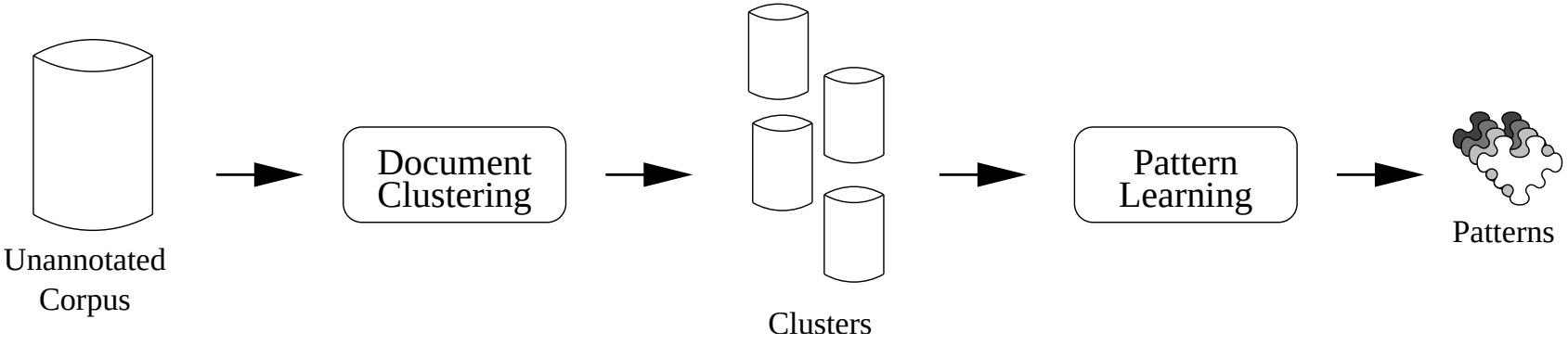
Joint Clustering and Pattern Learning



Semantics of the Learned Patterns

- When human supervision is reduced:
  - There are no starting seeds.
  - There are no starting domains.
  - There is no scenario.
  - Patterns are **open domain**.

Sequential Clustering and Pattern Learning



Pattern Learning Method

- Method of [STA06].
- Set of Meta-Patterns.
  - Grammatical relations.
- Co-training Approach. Train two classifiers:
  - EM and Naive Bayes with the words as features [NMTM00].
  - Decision List with instantiated patterns as features.
- Requires a set of seed documents.
  - Substitute this manual seeding by clustering.

State-of-the-Art Clustering

- Iterative refinement.
  - k-Means [McQ67].
  - EM [DLR77].
  - Divisive Information Theoretical Clustering [DG03].
- Hierarchical.
  - Hierarchical Agglomerative Clustering [MK80].
  - Agglomerative Information Bottleneck [ST99].
- Combination.
  - Graph based [SG02].
  - EM based [TJP05].

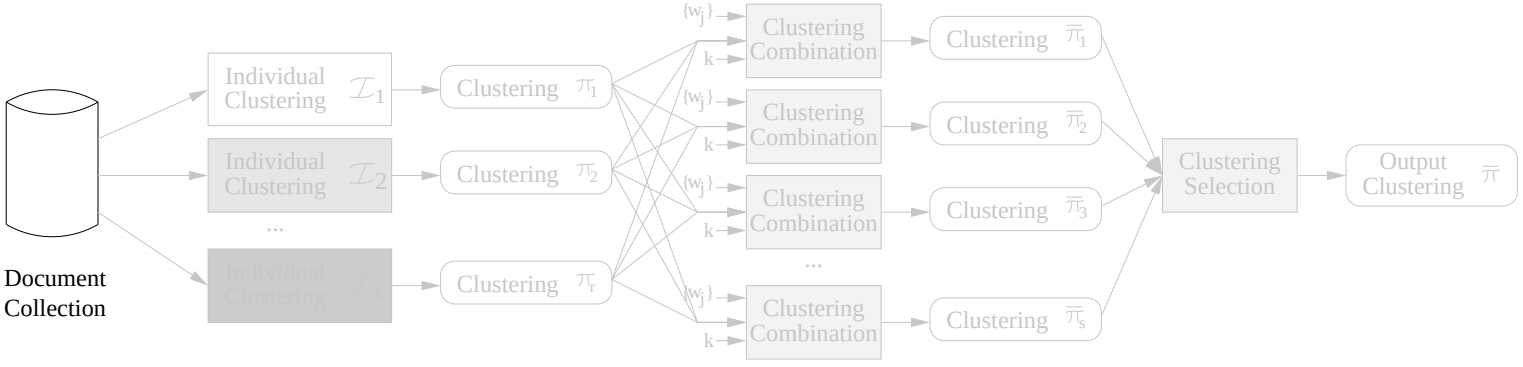
State-of-the-Art Clustering

- Supervised Clustering.
  - The number of clusters k.
  - Set of seed documents.
- Unsupervised Clustering.
  - Classical methods
    - Generate candidates.
    - Select best one.
  - Hybrid methods [STA05].

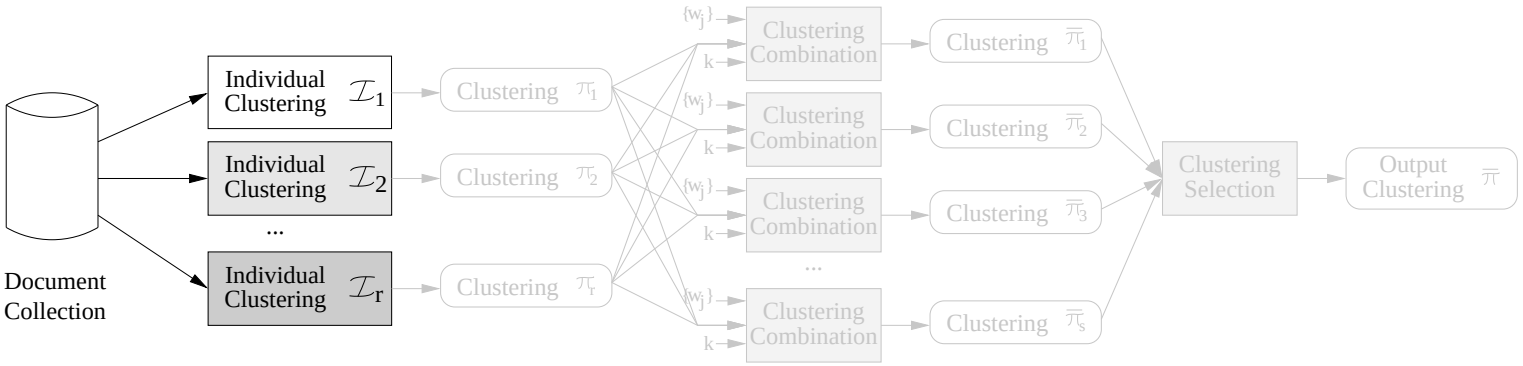
However...

- To our knowledge, no approach integrates:
  - Unsupervised Clustering.
  - Clustering Combination.
  - Weighting.

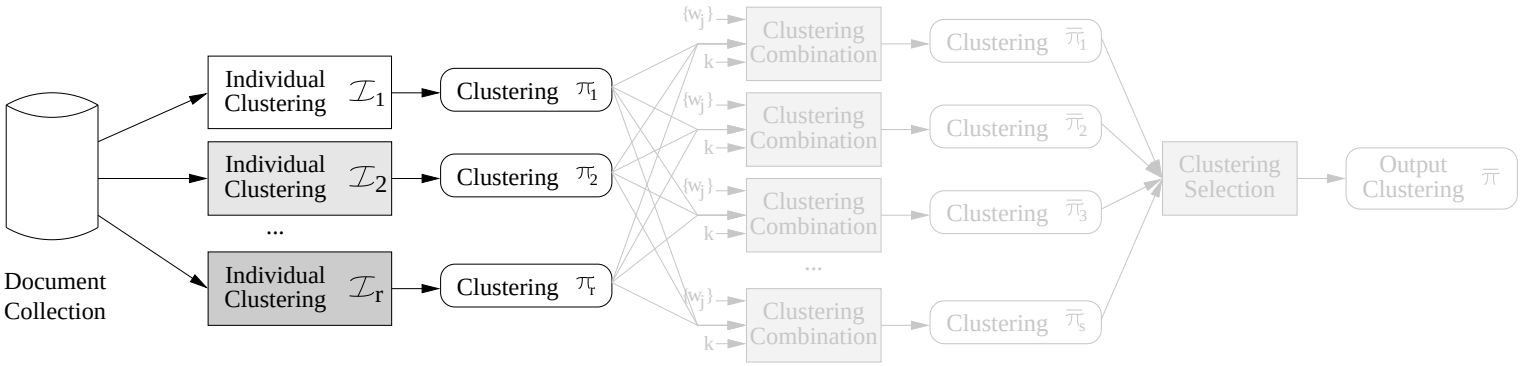
Generic Approach for Weighted Clustering Combination



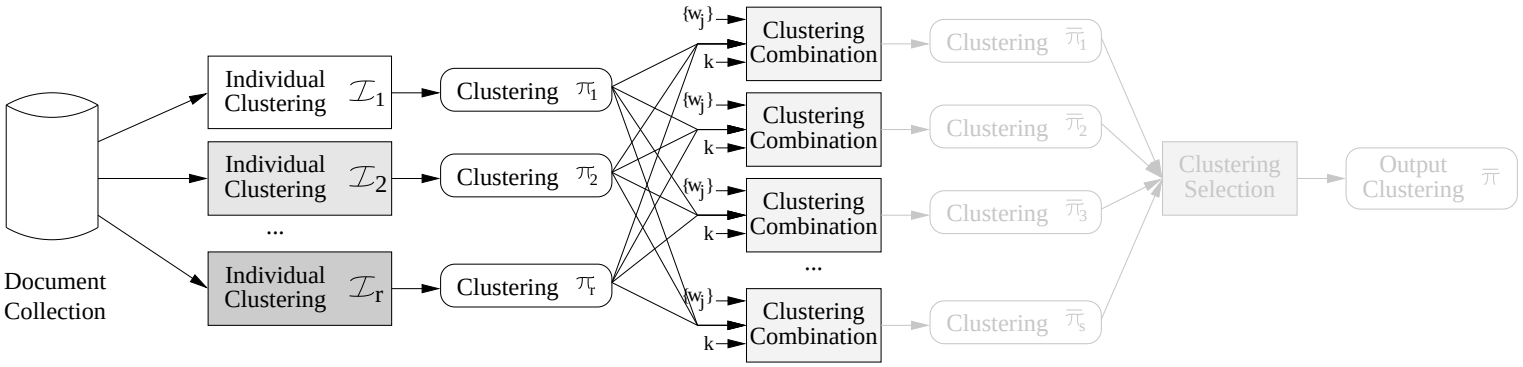
Generic Approach for Weighted Clustering Combination



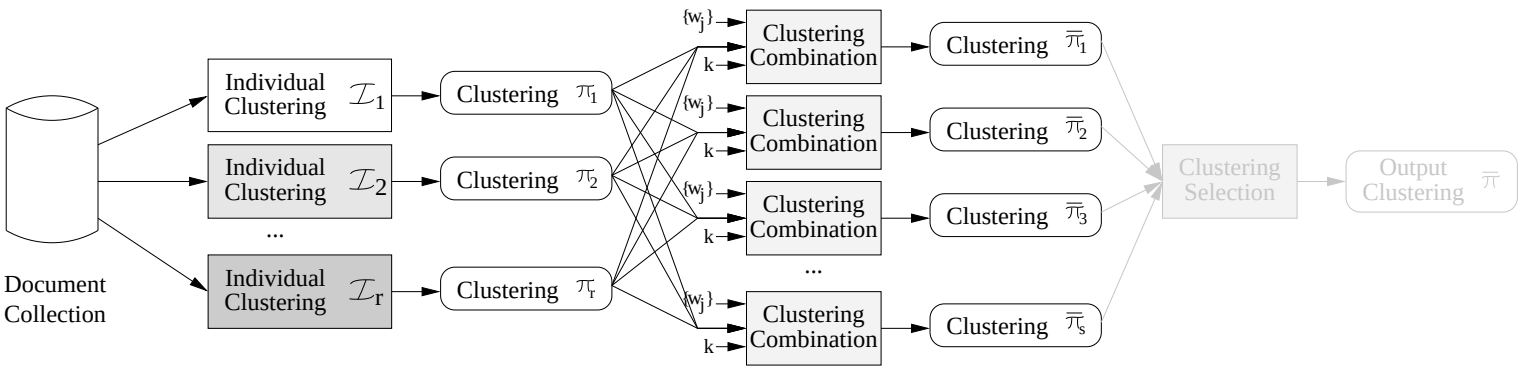
Generic Approach for Weighted Clustering Combination



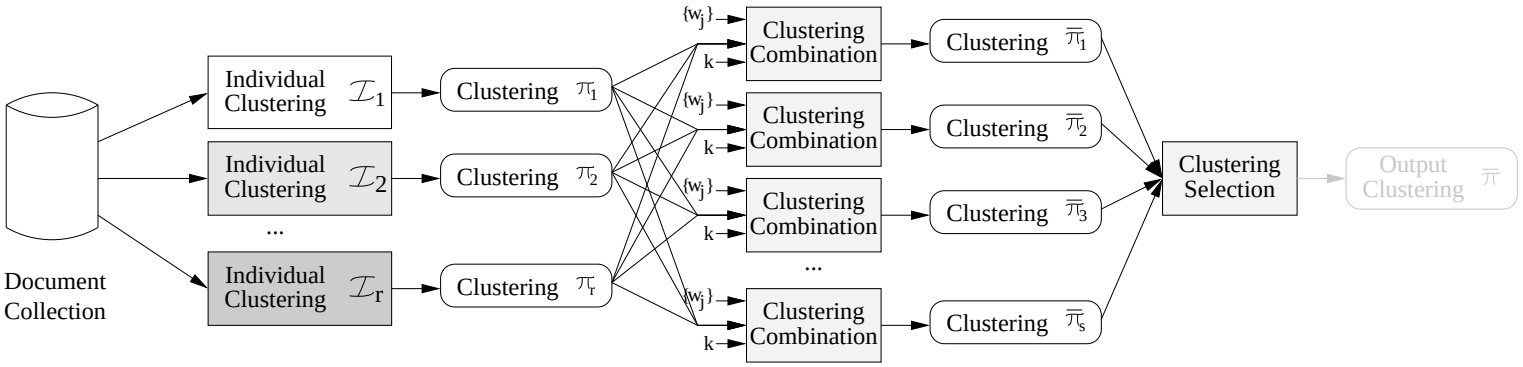
Generic Approach for Weighted Clustering Combination



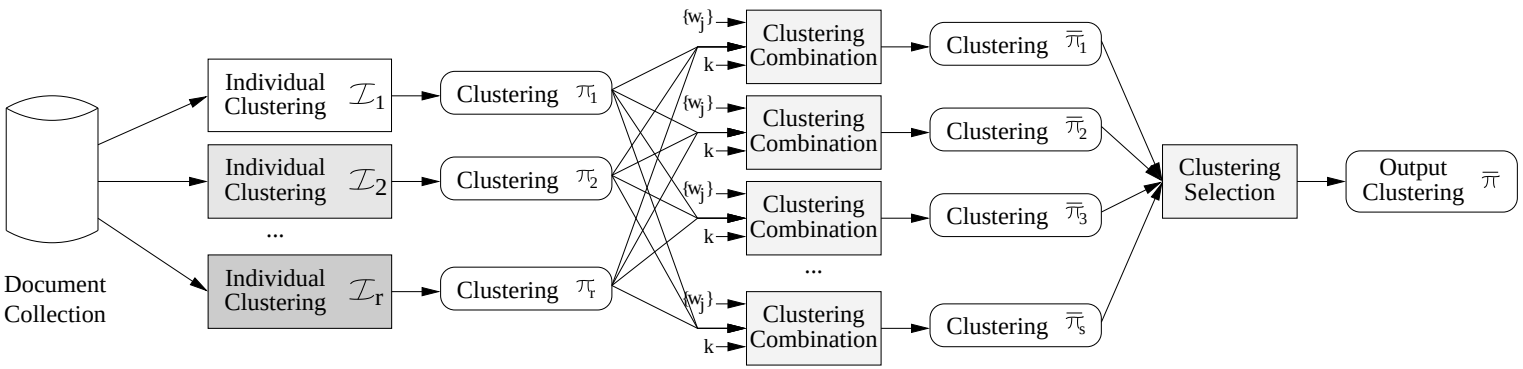
Generic Approach for Weighted Clustering Combination



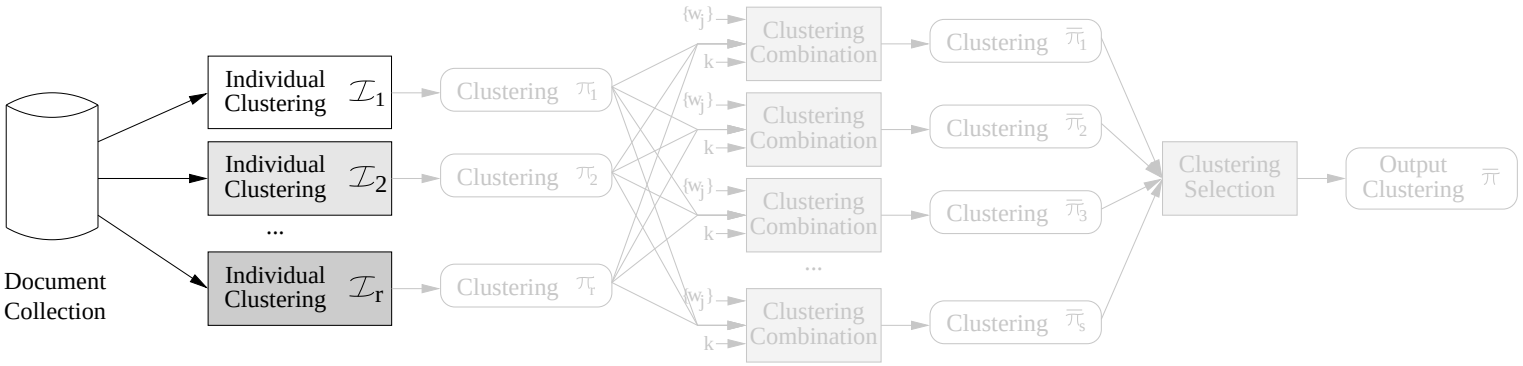
Generic Approach for Weighted Clustering Combination



Generic Approach for Weighted Clustering Combination



Individual Clustering Methods



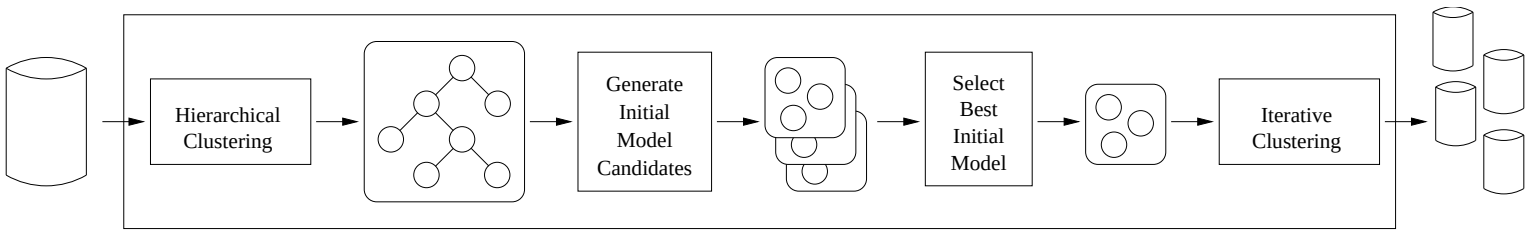
Individual Clustering Methods

Geo Geometric Hybrid Method of [STA05].

IT Information Theoretical Hybrid Method.

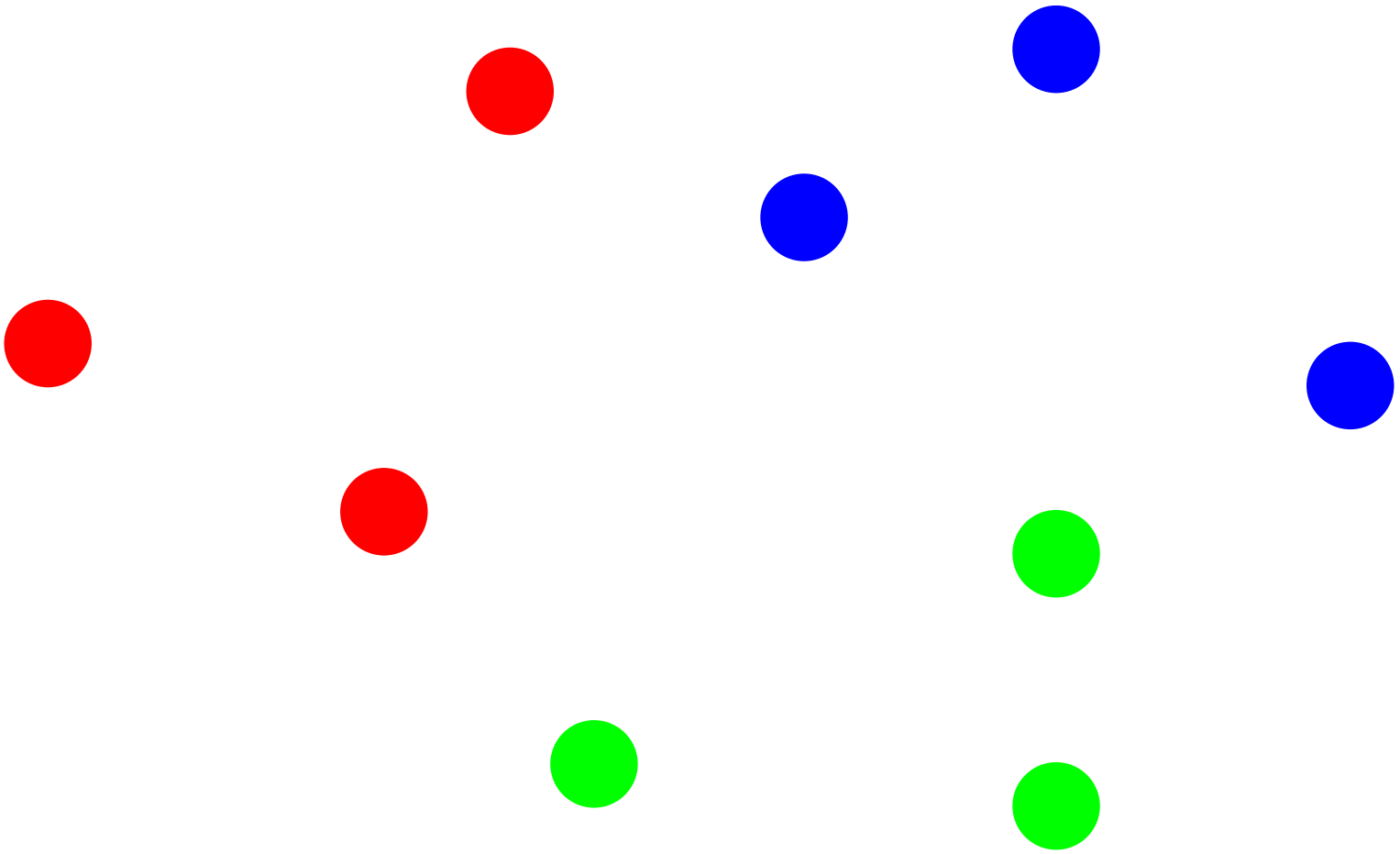
HiIT Hierarchical Method.

Information Theoretical Hybrid Method

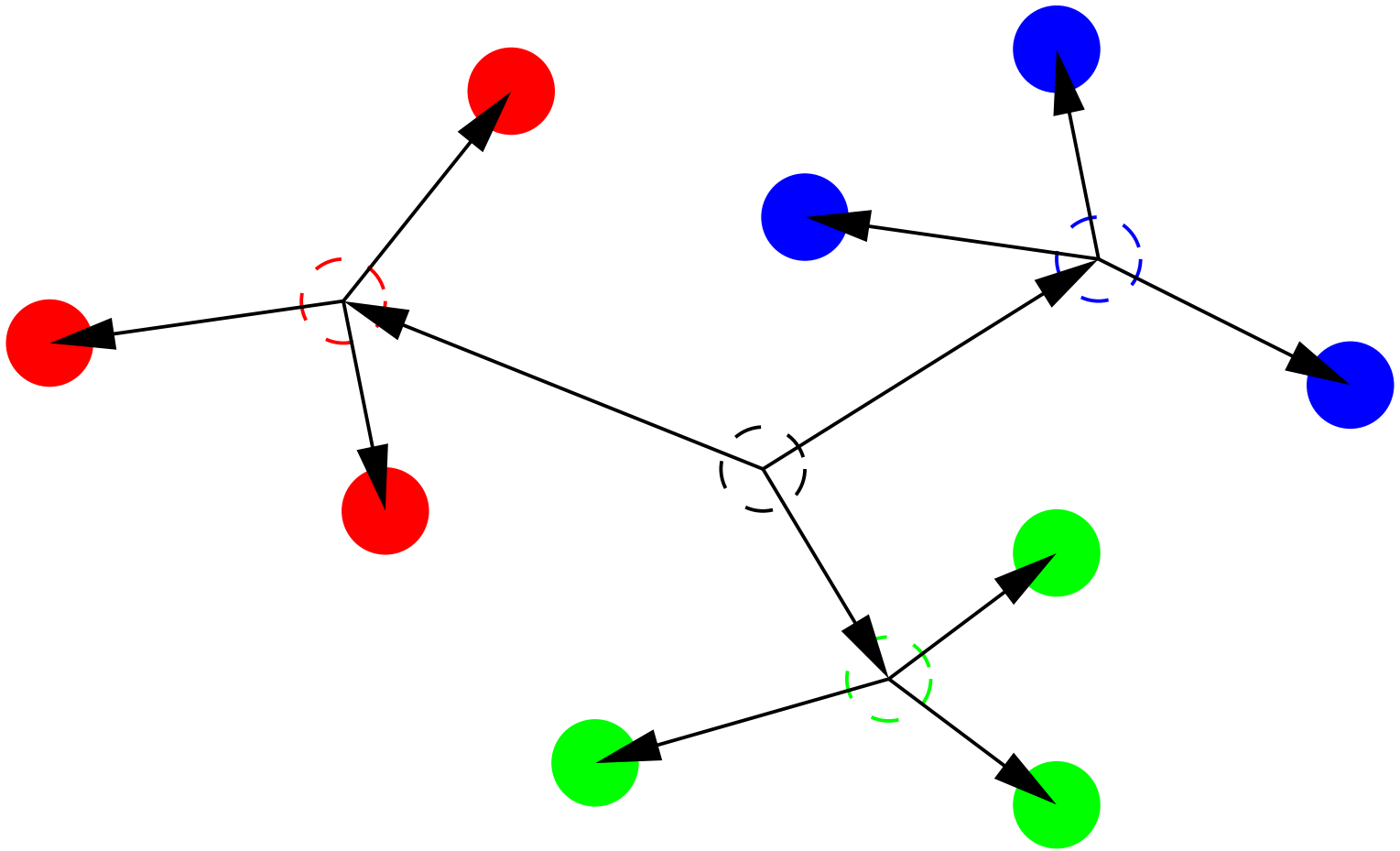


- Probability distribution representation.
- Jensen-Shannon Divergence [Lin91].
- Agglomerative Information Bottleneck [ST99].
- Divisive Information Theoretical Clustering [DG03].
- Code Length Criterion [GT06].

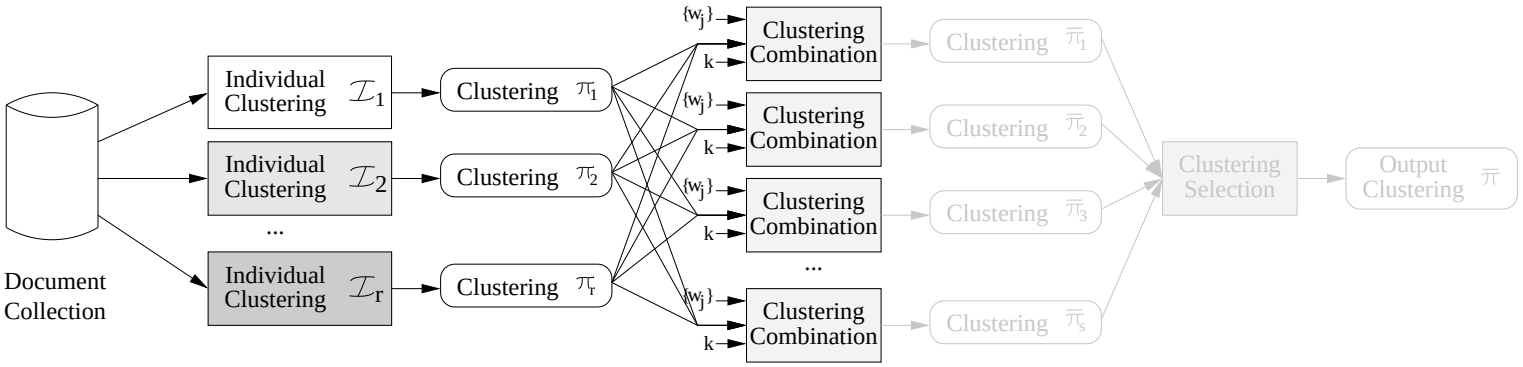
Code Length Criterion



Code Length Criterion



Weighted Combination



Weighted Combination

- Adaptation of two unweighted combination methods:
  - **Gr** Graph Based Method [SG02].
  - **EM** Expectation-Maximization Based Method [TJP05].

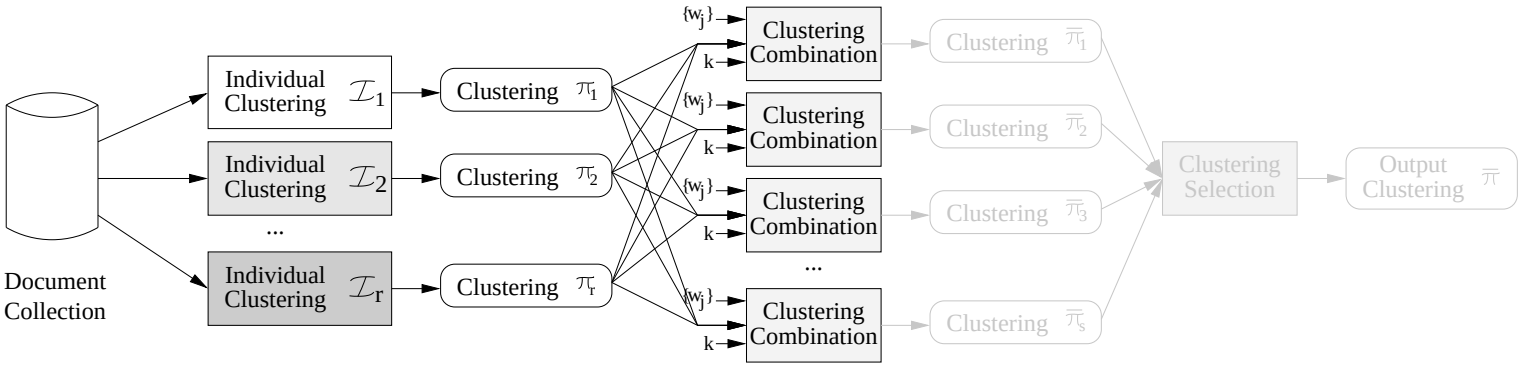
Graph Based Method

- Solve 3 graph partition problems.
  - CSPA
  - HGPA
  - MCLA
- Take the solution with highest Average Normalized Mutual Information.

EM Based Method

- Apply EM with a Naive Bayes Multinomial Model.
  - The cluster labels in each clustering are the features of each element to cluster.
- Use feature salience methods to find the weights [LJF02].

Combination Parameters

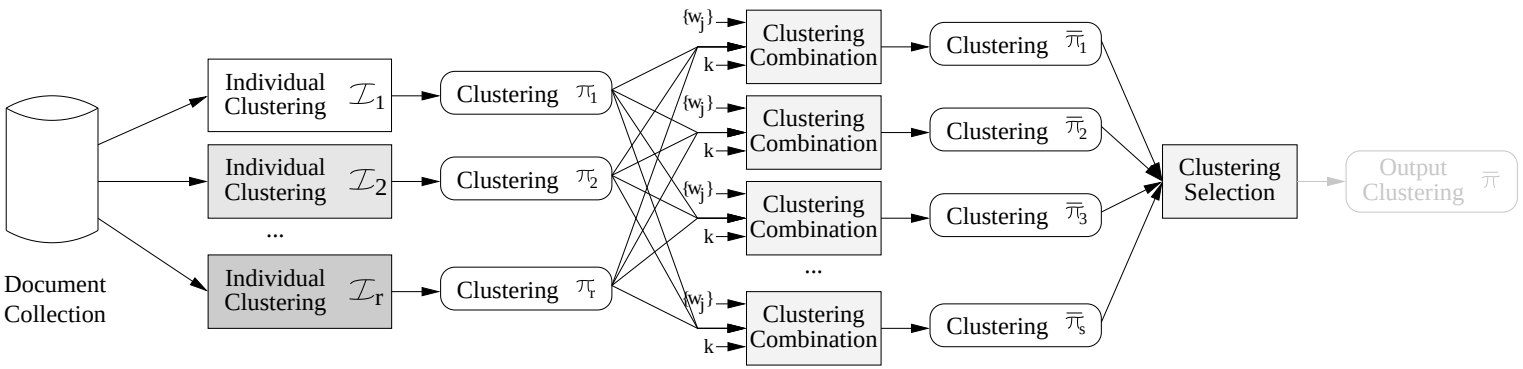


Combination Parameters

- Generate combinations for different $\bar{k}$.
  - And $\{w_j\}$ in Graph Based.

$$\begin{aligned} \bar{k} &\in \{\max(2, \lfloor \min k_j - \sigma_k \rfloor) \ldots \min(\|\mathcal{D}\|, \lceil \max k_j + \sigma_k \rceil)\} \\ w_j &\in \{1 \ldots G\} \end{aligned}$$

Clustering Selection



Best Combination Selection

- Select best using scoring function:
  - ANMI for Graph Based Clustering.
  - Bayesian Information Criterion for EM Based Clustering.

Used Seed Sets

Geo	Complete Clustering Solutions of Geo.
IT	Complete Clustering Solutions of IT.
HiIT	Complete Clustering Solutions of HiIT.
Graph	Graph Based Weighted Combination Method.
EM	EM Based Weighted Combination Method.
Geo.Seeds	Seeds Selected by Geo.
IT.Seeds	Seeds Selected by IT.

Index

Clustering Spontaneous Speech Transcripts

- Checking the suitability of unsupervised document clustering methods on spontaneous speech transcripts.
- [GT05].
- Compare performance of Geo against HAC and EM by themselves.

Evaluation Data

- Manual Transcripts from Switchboard-1 Corpus.

	Documents	Categories
SWB	4876	67
SWB-22	2682	22

Results

Collection	Method	Pur	IPur	F_1	k
SWB	**Geo**	**.54**	**.87**	**.67**	**26.90**
	HAC + C	.07	.99	.14	6.30
	EM5 + C	.10	.91	.19	3.90
SBW-22	**Geo**	**.81**	**.92**	**.86**	**18.00**
	HAC + C	.21	.98	.33	5.50
	EM5 + C	.19	.95	.31	3.60

Results

Collection	Method	Pur	IPur	F_1	k
SWB	**Geo**	**.54**	**.87**	**.67**	**26.90**
	HAC	.35	.93	.50	*id*
	EM5	.16	.88	.27	*id*
SBW-22	**Geo**	**.81**	**.92**	**.86**	**18.00**
	HAC	.52	.95	.67	*id*
	EM5	.32	.89	.46	*id*

Clustering Combination

- Compare performance of individual clustering methods:
  - Geo, IT, HiIT.
- and weighted and unweighted combination methods:
  - Gr, EM.

Evaluation Data

- Document collections from several sources.

	Documents	Categories
APW	5000	11
LATIMES	5000	8
Reuters	3019	93
Reuters10	2545	10
SMART	5467	4

Results

Collection	Method	Pur	IPur	F_1	k
APW	**HilT**	**.63**	**.88**	**.74**	**3**
	Gr.Eq	.71	.73	.72	7
	EM.Eq	.72	.67	.69	8
LATIMES	IT	.75	.61	.67	7
	Gr.Eq	**.75**	**.68**	**.72**	**7**
	EM.Eq	.77	.66	.71	7
Reuters	Geo	.68	.83	.75	8
	Gr.Eq	.70	.86	.77	7
	EM.Eq	**.70**	**.90**	**.79**	**6**
Reuters10	Geo	.77	.85	.81	6
	Gr.Eq	.81	.84	.83	7
	EM.Eq	**.81**	**.88**	**.85**	**6**
SMART	Geo	.91	.77	.83	6
	Gr.Eq	**.91**	**.91**	**.91**	**4**
	EM.Eq	.92	.90	.91	5

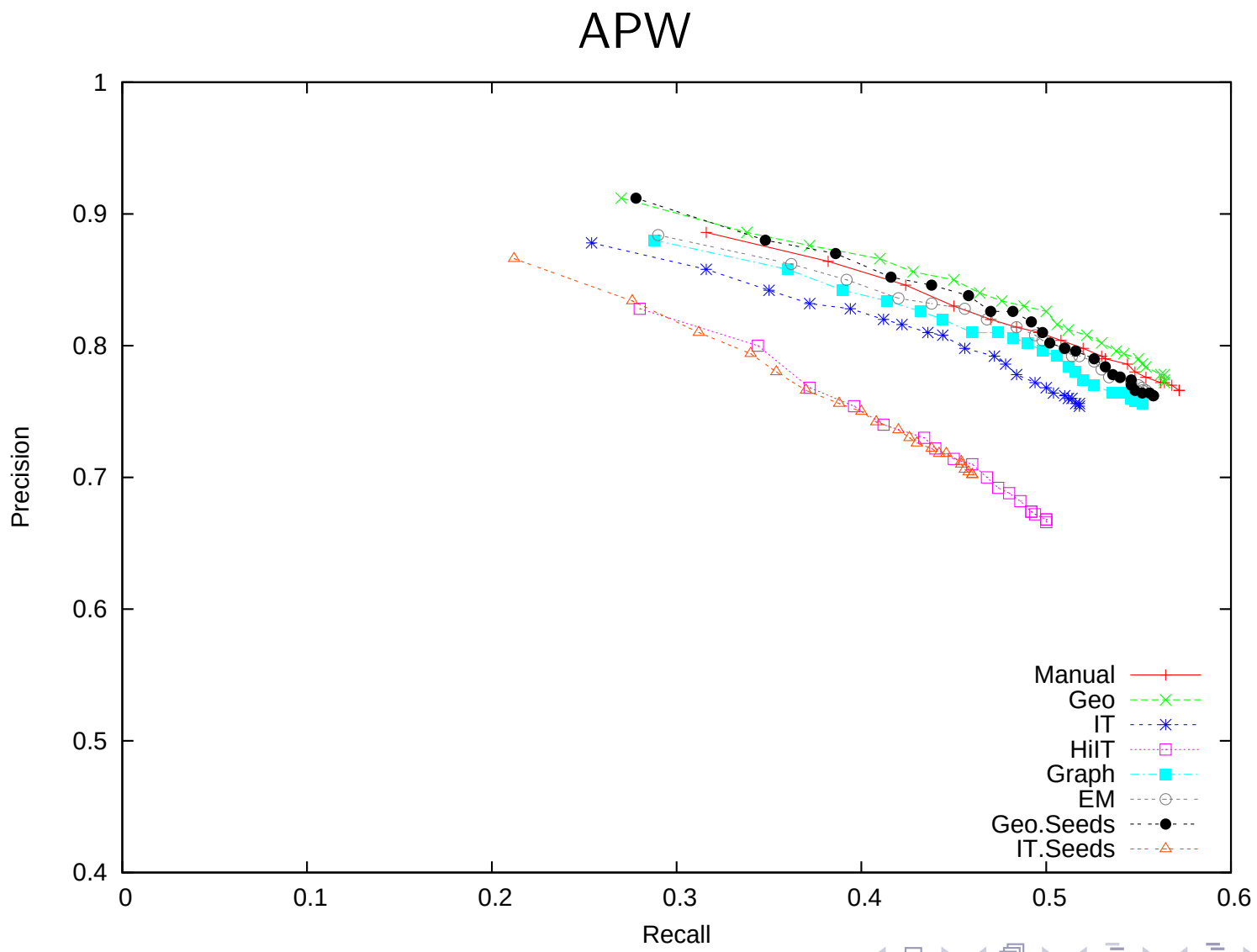
Results

Collection	Method	Pur	IPur	F_1	k
APW	Gr.Eq	.71	.73	.72	7
	Gr.W	.72	.72	.72	7
	EM.Eq	.72	.67	.69	8
	EM.W	.70	.66	.68	7
LATIMES	Gr.Eq	.75	.68	.72	7
	Gr.W	.76	.68	.72	8
	EM.Eq	.77	.66	.71	7
	EM.W	.77	.66	.71	7
Reuters	Gr.Eq	.70	.86	.77	7
	Gr.W	.71	.86	.78	9
	EM.Eq	.70	.90	.79	6
	EM.W	.71	.88	.79	7
Reuters10	Gr.Eq	.81	.84	.83	7
	Gr.W	.82	.85	.83	7
	EM.Eq	.81	.88	.85	6
	EM.W	.81	.88	.85	7
SMART	Gr.Eq	.91	.91	.91	4
	Gr.W	.92	.91	.92	5
	EM.Eq	.92	.90	.91	5
	EM.W	.92	.90	.91	5

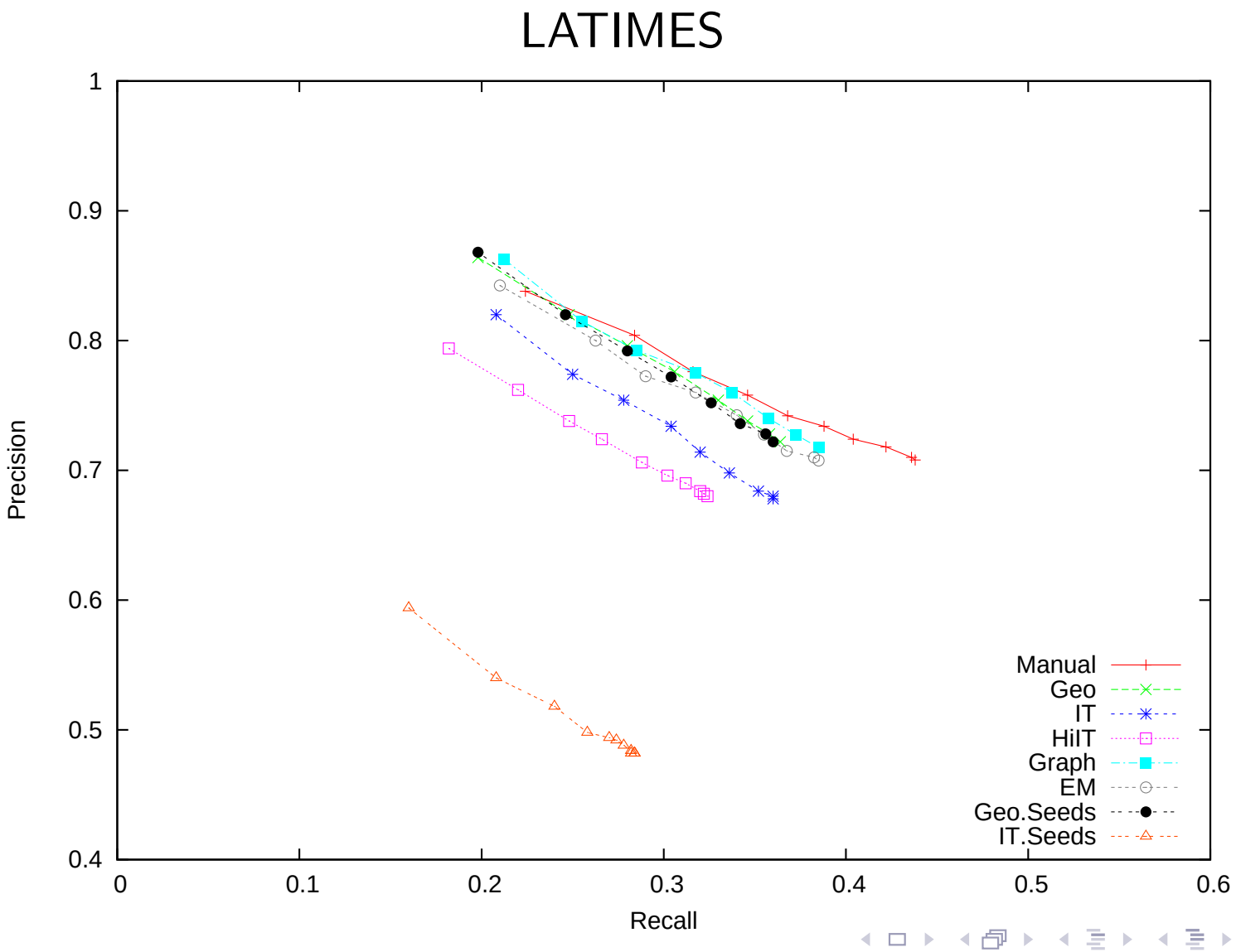
Enhancing Pattern Learning

- Establishing a baseline for the combination of document clustering and pattern learning.
- Evaluation through text classification.
- Same evaluation data as in the previous experiments

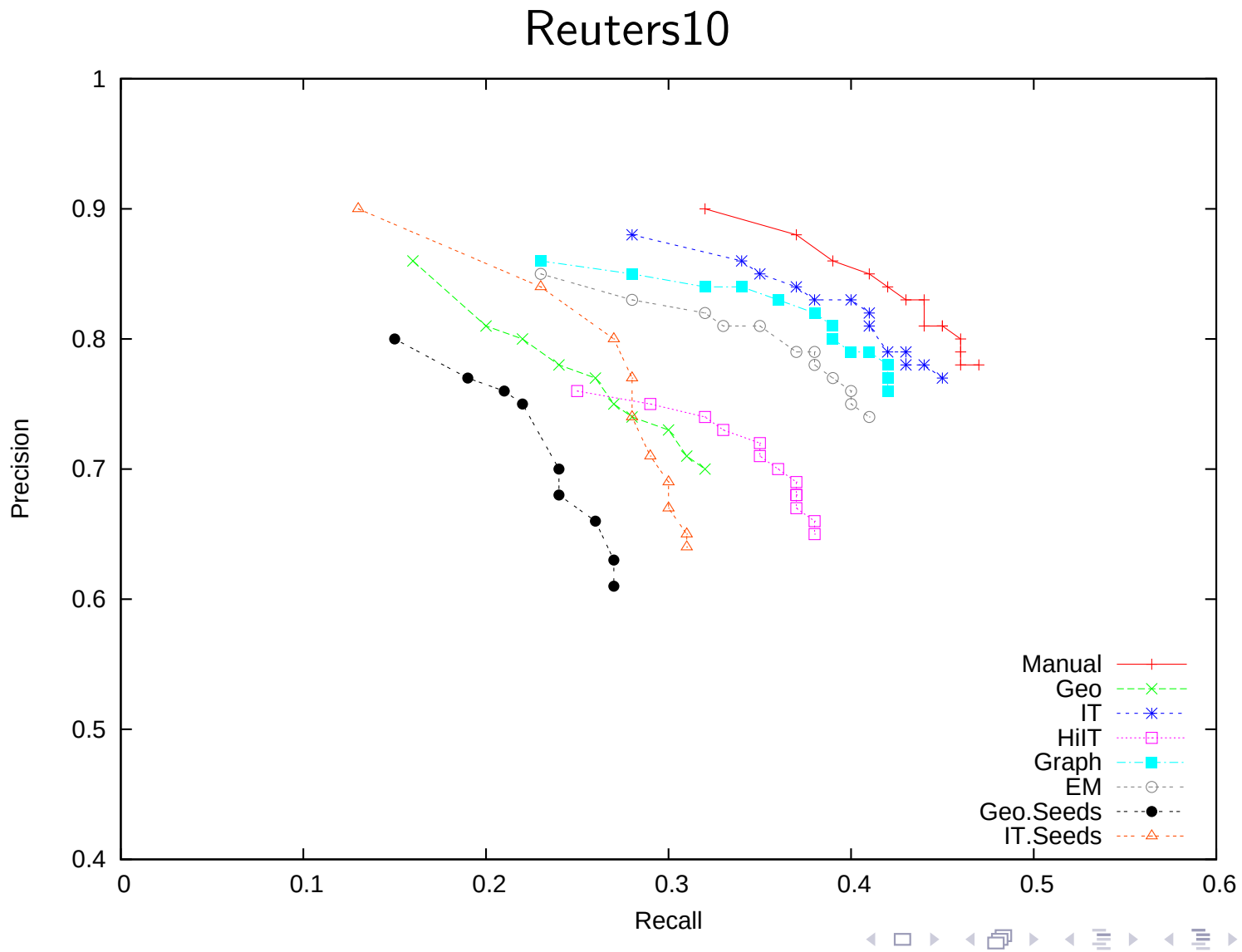
Results


APW
Precision
Recall
1
0.9
0.8
0.7
0.6
0.5
0.4
0
0.1
0.2
0.3
0.4
0.5
0.6
Manual
Geo
IT
HiIT
Graph
EM
Geo.Seeds
IT.Seeds

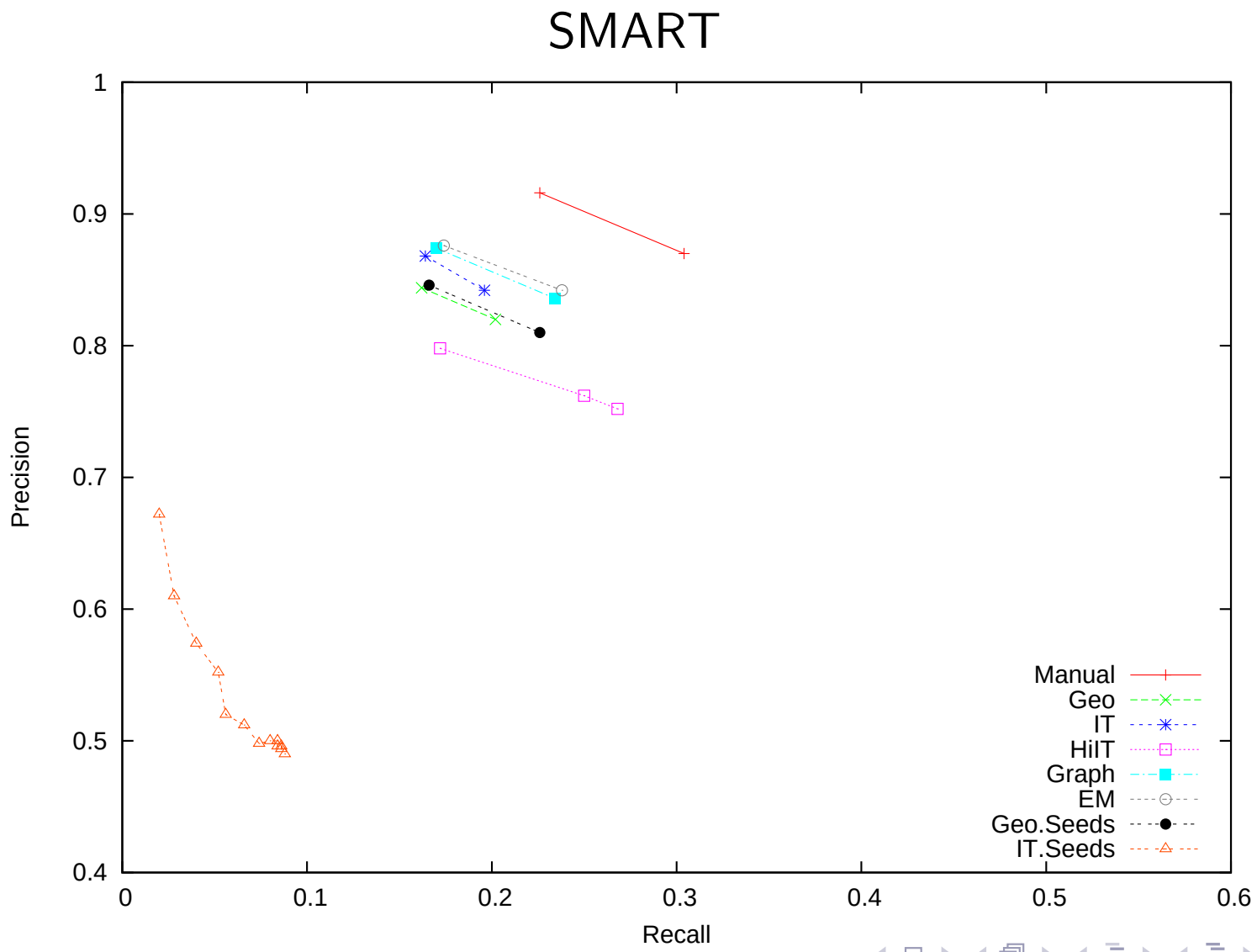
Results


LATIMES
Precision
Recall
1
0.9
0.8
0.7
0.6
0.5
0.4
0
0.1
0.2
0.3
0.4
0.5
0.6
Manual
Geo
IT
HiIT
Graph
EM
Geo.Seeds
IT.Seeds

Results


Reuters10
Precision
Recall
1
0.9
0.8
0.7
0.6
0.5
0.4
0
0.1
0.2
0.3
0.4
0.5
0.6
Manual
Geo
IT
HiIT
Graph
EM
Geo.Seeds
IT.Seeds

Results


SMART
Precision
Recall
1
0.9
0.8
0.7
0.6
0.5
0.4
0
0.1
0.2
0.3
0.4
0.5
0.6
Manual
Geo
IT
HiIT
Graph
EM
Geo.Seeds
IT.Seeds

Sample Learned Patterns

- APW Corpus with EM Seeds.
- `shoot {per}`
  `charge with murder`
- `remove from office`
  `remove president`
- `win title`
  `he {verb} game`

Index

Document Clustering

- E. Gonzàlez and J. Turmo.
  Unsupervised Clustering of Spontaneous Speech Documents.
  In *Proceedings of the 9th European Conference on Speech Communication and Technology (EuroSpeech/InterSpeech)*, 2005.
- E. Gonzàlez and J. Turmo.
  Unsupervised Document Clustering by Weighted Combination.
  Technical report, LSI Department, Universitat Politècnica de Catalunya, 2006.
- E. Gonzàlez and J. Turmo.
  Unsupervised Document Clustering by Combination Methods.
  Submitted to International Joint Conference on Artificial Intelligence (IJCAI) 2007.

Question Answering I

- D. Ferrés, S. Kanaan, A. Ageno, E. Gonzàlez, H. Rodríguez, M. Surdeanu, and J. Turmo.
  Talp-QA System for Spanish at Clef-2004: Structural and Hierarchical Relaxing over Semantic Constraints.
  In *CLEF 2004 Evaluation Campaign*, 2004.
- D. Ferrés, S. Kanaan, E. Gonzàlez, A. Ageno, H. Rodríguez, M. Surdeanu, and J. Turmo.
  Talp-QA System at Trec-2004: Structural and Hierarchical Relaxing over Semantic Constraints.
  In *TREC 2004 Evaluation Campaign*, 2004.
- D. Ferrés, S. Kanaan, E. Gonzàlez, A. Ageno, H. Rodríguez, and J. Turmo.
  The Talp-QA System for Spanish at Clef-2005.
  In *CLEF 2005 Evaluation Campaign*, 2005.

Question Answering II

- D. Ferrés, S. Kanaan, D. Domínguez-Sal, E. Gonzàlez, A. Ageno, M. Fuentes, H. Rodríguez, M. Surdeanu, and J. Turmo.
  Talp-UPC at Trec 2005: Experiments Using a Voting Scheme among three Hetereogeneous QA Systems.
  In *TREC 2005 Evaluation Campaign*, 2005.

Summarization I

- M. Fuentes, E. Gonzàlez, and H. Rodríguez.
  Resumidor de Notícies en Català del Projecte Hermes.
  In *Proceedings of the II Congrés d'Enginyeria en Llengua Catalana*, 2004.
- M. Fuentes, E. Gonzàlez, and J. Turmo.
  Baseline Summarization System for Text including Speech Transcripts.
  European Project CHIL (IP 506909) Deliverable D5.8, 2005.
- M. Fuentes, E. Gonzàlez, H. Rodríguez, J. Turmo, and L. Alonso.
  Summarizing Spontaneous Speech using General Text Properties.
  In *Proceedings of the Crossing Barriers in Text Summarization Research Workshop at RANLP 2005*, 2005.

Summarization II

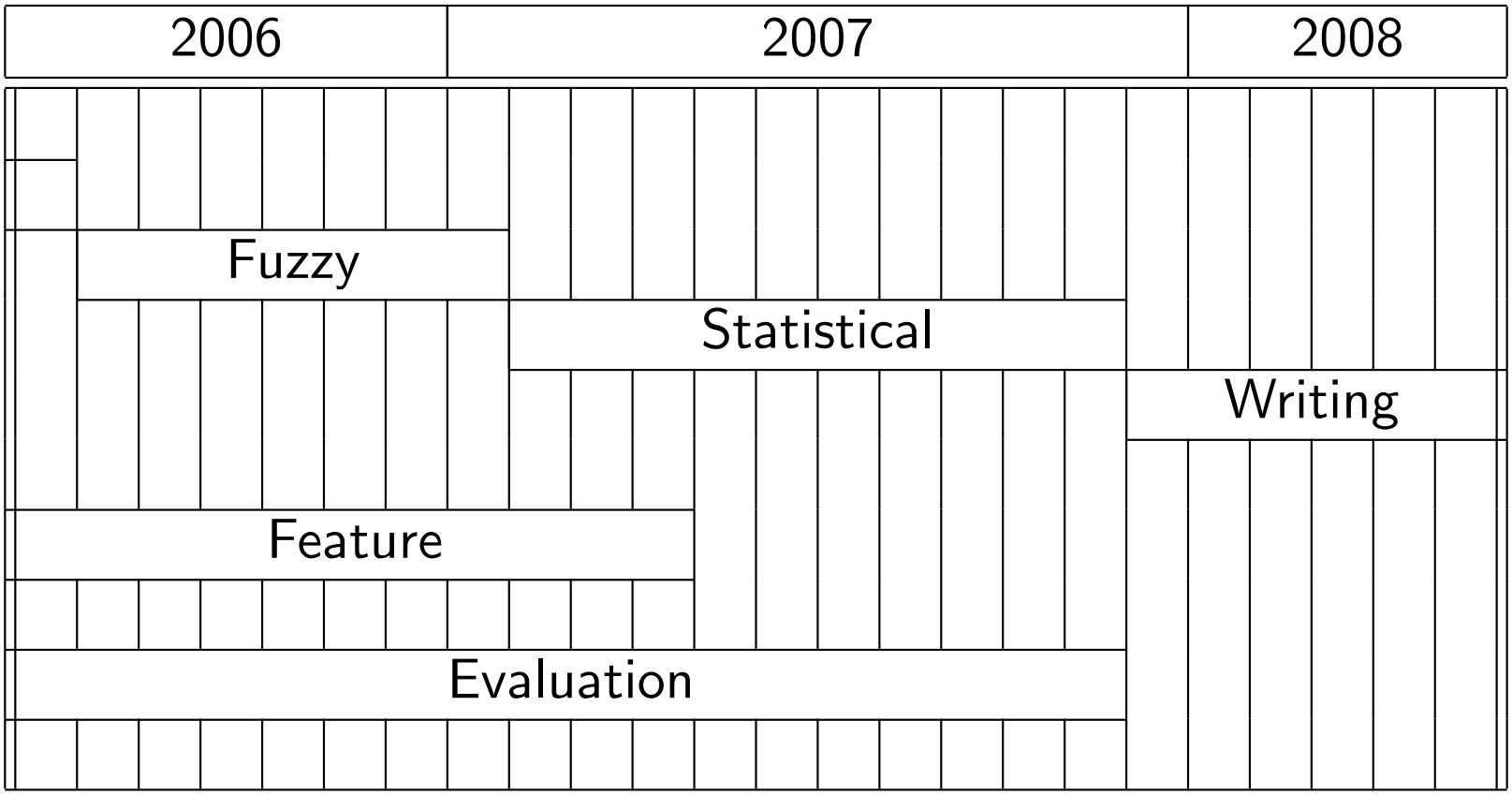
- M. Fuentes, E. Gonzàlez, D. Ferrés, and H. Rodríguez. QaSum-Talp at Duc 2005 Automatically Evaluated with a Pyramid Based Metric.
  In *DUC 2005 Evaluation Campaign*, 2005.

Index

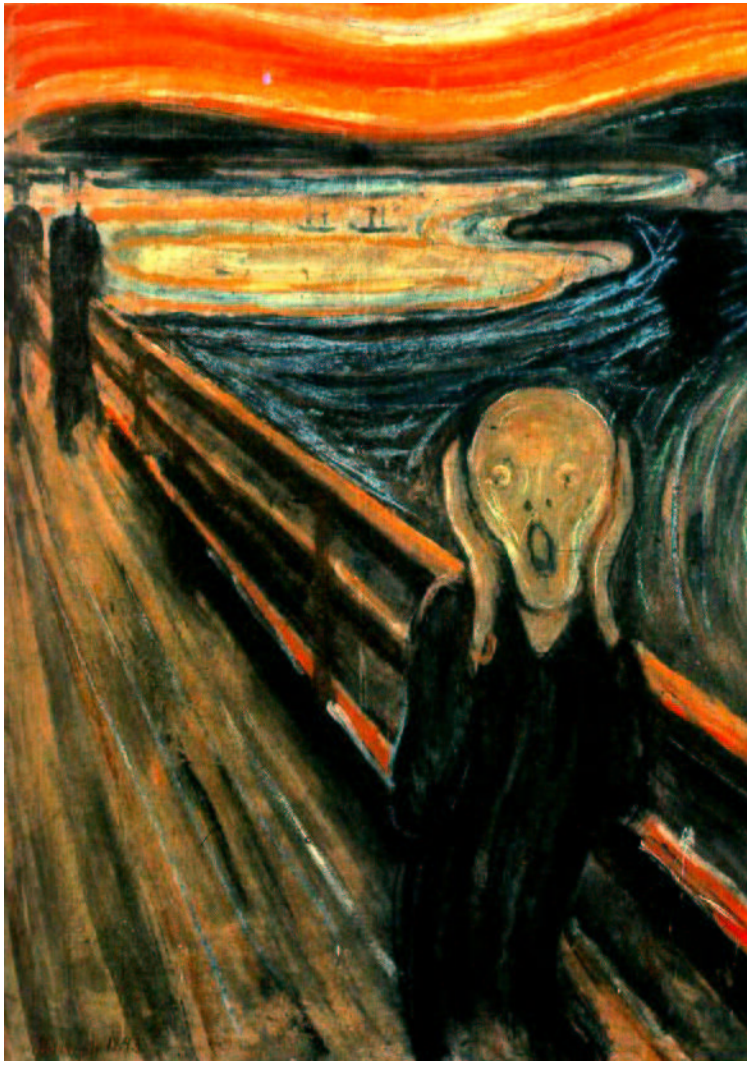
Research Lines

- Improve sequential approach weight detection.
- Add *fuzziness* to clustering.
- Apply statistical rule learning algorithms.
- Analize the features used to represent documents and patterns.
- Evaluate the domain relevance of patterns.
- Evaluate the patterns intrinsically on full IE tasks.
- Evaluate the patterns extrinsically on QA and Summarization.

Schedule


2006
2007
2008
Fuzzy
Statistical
Writing
Feature
Evaluation

Thank You!



Index

Bibliography I


I.S. Dhillon and Y. Guan.
Information theoretic clustering of sparse co-occurrence data.
In *Proceedings of IEEE International Conference on Data Mining*, 2003.


A.P. Dempster, N.M. Laird, and D.B. Rubin.
Maximum likelihood from incomplete data via the EM algorithm.
Royal Statistical Society, Series B, 39(1), 1977.


E. Gonzàlez and J. Turmo.
Unsupervised clustering of spontaneous speech documents.
In *Proceedings of the 9th European Conference on Speech Communication and Technology (EuroSpeech/InterSpeech)*, 2005.

Bibliography II


E. Gonzàlez and J. Turmo.
Unsupervised document clustering by weighted combination.
Technical report, LSI Department, Universitat Politècnica de Catalunya, 2006.


J. Lin.
Divergence measures based on the shannon entropy.
IEEE Transactions on Information Theory, 37(1):145–151, 1991.


M.H. Law, A.K. Jain, and M.A.T. Figueiredo.
Feature selection in mixture-based clustering.
In *Proceedings of the Neural Information Processing Systems Conference (NIPS)*, 2002.

Bibliography III


J. McQueen.
Some methods for classification and anlysis of multivariate observations.
In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967.


M.N. Murty and G. Krishna.
A computationally efficient technique for data clustering.
Pattern Recognition, 12:153–158, 1980.


K. Nigam, A. McCallum, S. Thrun, and T. Mitchell.
Text classification from labeled and unlabeled documents using em.
Machine Learning, 39(2/3), 2000.

Bibliography IV


A. Strehl and J. Ghosh.
Cluster ensembles - a knowledge reuse framework for combining multiple partitions.
Journal of Machine Learning Research, 3:583–617, 2002.


N. Slonim and N. Tishby.
Agglomerative information bottleneck.
In *Proceedings of Advances in Neural Information Processing Systems (NIPS-12)*, 1999.


M. Surdeanu, J. Turmo, and A. Ageno.
A hybrid unsupervised approach for document clustering.
In *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2005.

Bibliography V


M. Surdeanu, J. Turmo, and A. Ageno.
A hybrid approach for the acquisition of information extraction patterns.
In *Proceedings of the EACL 2006 Workshop on Adaptive Text Extraction and Mining (ATEM)*, pages 49–56, 2006.


A. Topchy, A.K. Jain, and W. Punch.
Clustering ensembles: Models of consensus and weak partitions.
IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(12), 2005.