Stochastic models for semi-structured document mining

P. Gallinari

Collaboration with

G. Wisniewski – L. Denoyer – F. Maes

LIP6

University Pierre – Marie Curie - Fr

Outline

- Context
- Generative tree models
- 3 problems
  - Classification
  - Clustering
  - Document mapping
- Experiments
- Conclusion and future work
  - XML Document Mining Challenge

Context - Machine learning in the structured domain

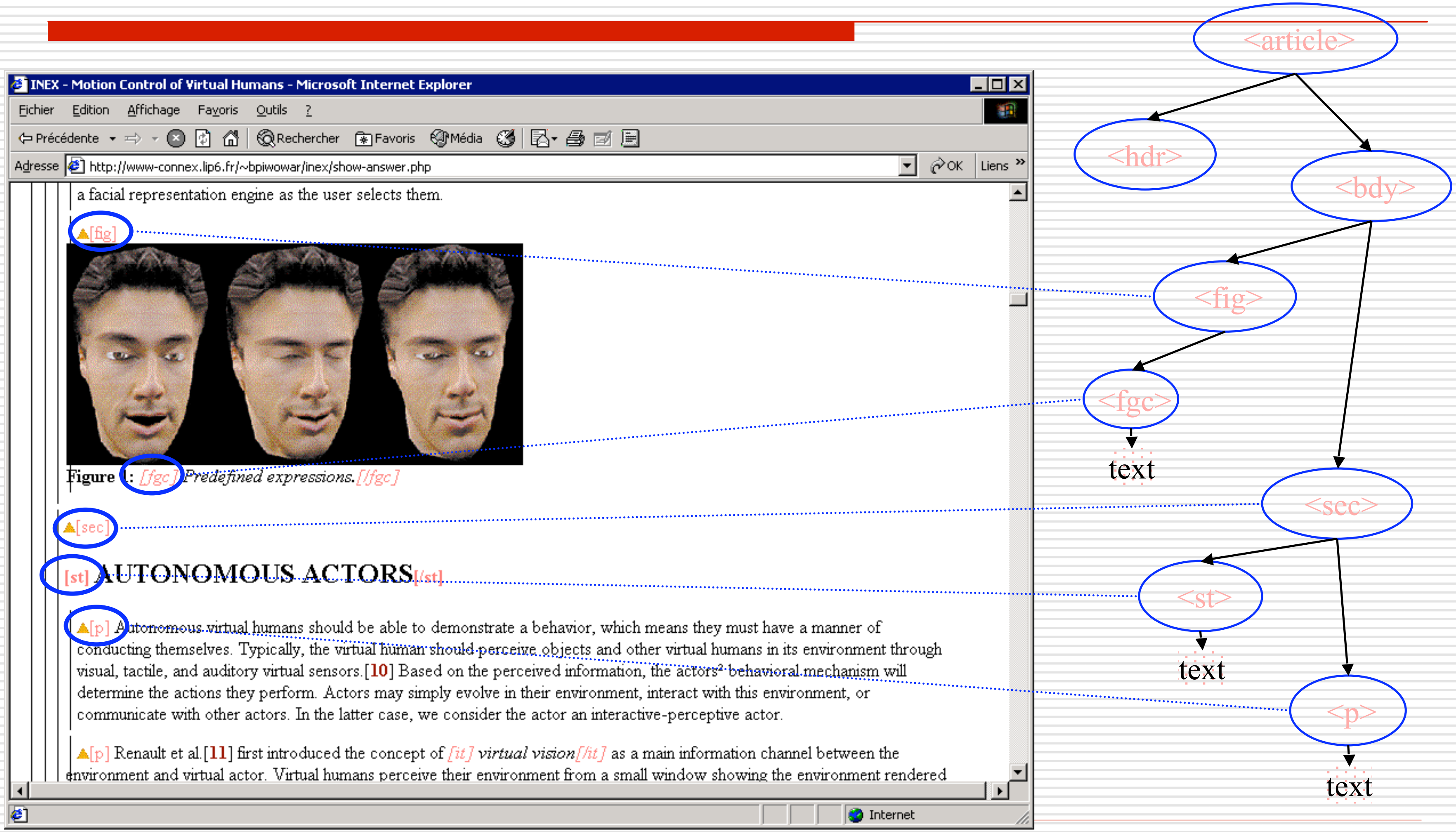
- □ Model, Classify, cluster structured data
  - Domains: Chemistry, biology, XML, etc
  - Models: discriminant e.g. kernels, generative e.g. tree densities
- □ Predict structured outputs
  - Domains: natural language parsing, taxonomies, etc
  - Models: relational learning, large margin extensions
- □ Learn to associate structured representations aka Tree mapping
  - Domains: databases, semi-structured data

Context- Machine learning in the structured domain

- Structure only vs Structure + content
- Central complexity issue
  - □ Representation space (#words, #tags, #relations)
  - □ Search space for structured outputs - idem
  - □ Large corpora

⇨ needs simple and approximate methods

Context-XML semi-structured documents



a facial representation engine as the user selects them.

[fig]

Figure 1: [fgc] *Predefined expressions.*[/fgc]

[sec]

[st] AUTONOMOUS ACTORS[/st]

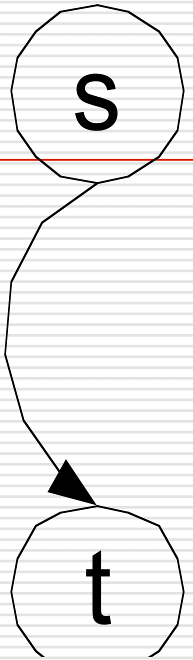
[p] Autonomous virtual humans should be able to demonstrate a behavior, which means they must have a manner of conducting themselves. Typically, the virtual human should perceive objects and other virtual humans in its environment through visual, tactile, and auditory virtual sensors.[10] Based on the perceived information, the actors' behavioral mechanism will determine the actions they perform. Actors may simply evolve in their environment, interact with this environment, or communicate with other actors. In the latter case, we consider the actor an interactive-perceptive actor.

[p] Renault et al.[11] first introduced the concept of [it] *virtual vision*[/it] as a main information channel between the environment and virtual actor. Virtual humans perceive their environment from a small window showing the environment rendered

Outline

- Context
- Generative tree models
- 3 problems
  - Classification
  - Clustering
  - Document restructuration
- Experiments
- Conclusion and future work
  - XML Document Mining Challenge

Document model



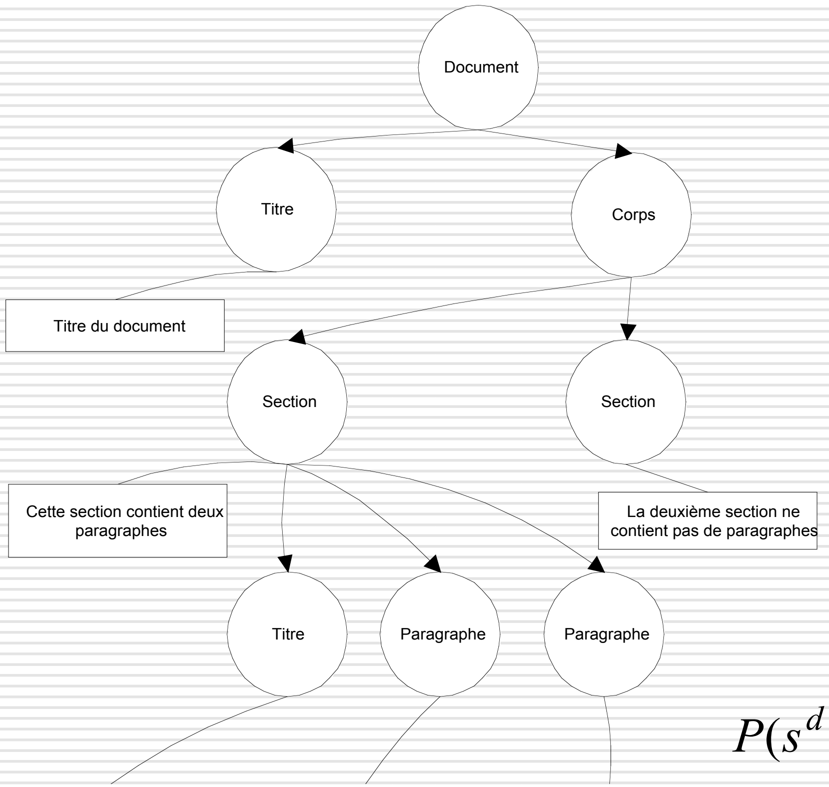
$$d = (s^d, t^d)$$

$$P(D = d / \Theta) = P(S = s^d, T = t^d / \Theta)$$

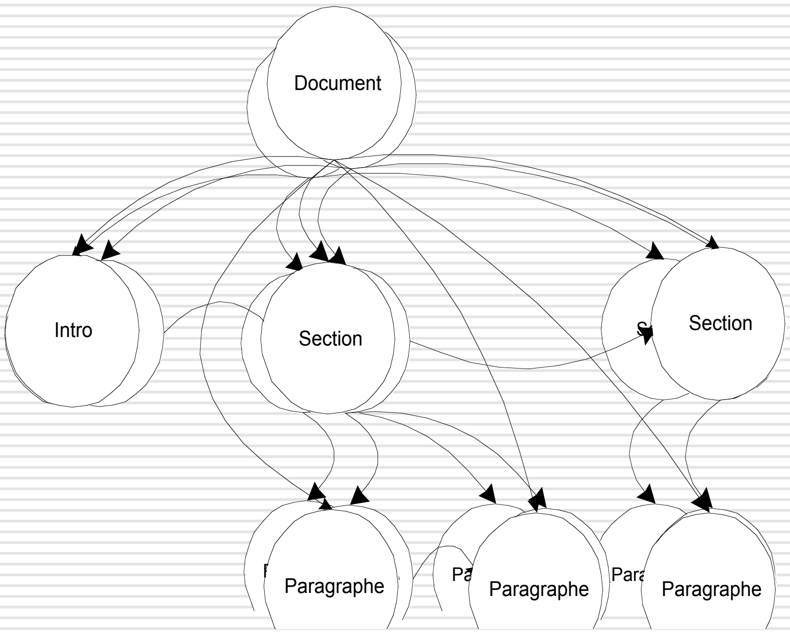
$$= \underbrace{P(S = s^d / \Theta)}_{\text{Structural probability}} \underbrace{P(T = t^d / S = s^d, \Theta)}_{\text{Content probability}}$$

Document Model: Structure

□ Belief Networks


Document
Titre
Corps
Titre du document
Section
Section
Cette section contient deux paragraphes
La deuxième section ne contient pas de paragraphes
Titre
Paragraphe
Paragraphe

$\Longrightarrow$



$$P(s^d) = \prod_{i=1}^{|d|} P\left(s_i^d / label(parent(s_i^d)), label(précédent(s_i^d))\right)$$

Document Model: Content

- □ Model for each node

$$t_d = (t_d^1, \ldots., t_d^{/d/})$$

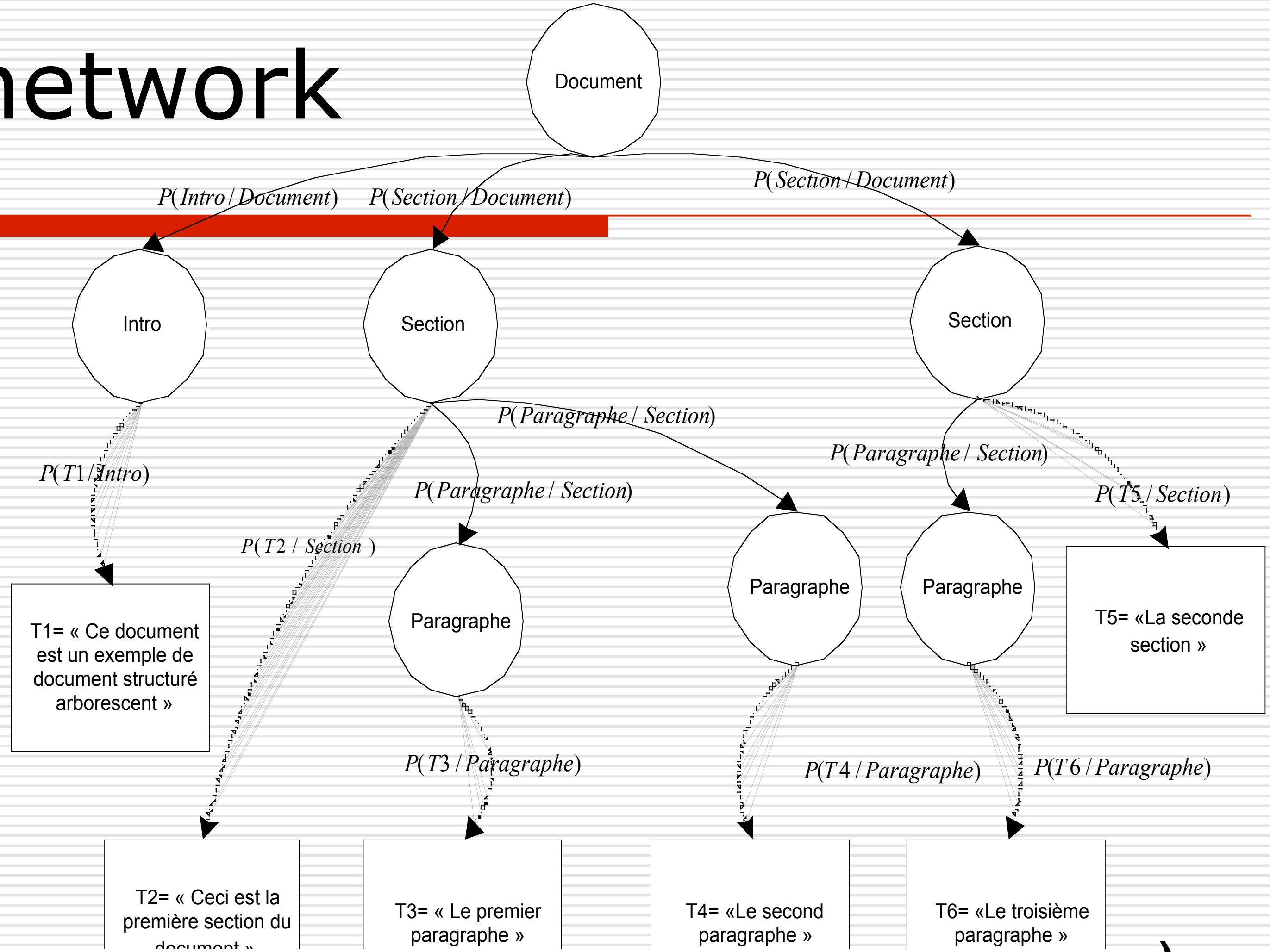
- □ 1st order dependency

$$P(t_d / s_d, \theta) = \prod_{i=1}^{/d/} P(t_d^i / s_d^i, \theta)$$

- □ Use of a local generative model for each label

$$P(t_d^i / s_d^i, \theta) = P(t_d^i / \theta_{s_d^i})$$

Final network



$$P(d) = \left(P(Intro / Document)P(Section / Document)?P(P\arg raphe / Section)^3\right)$$

$$* \, P(T1 / Intro)P(T2 / Section)P(T3 / Paragraphe)$$

$$* \, P(T4 / Paragraphe)P(T5 / Section)P(T6 / Paragraphe)$$

Different learning techniques

- □ Likelihood maximization

$$L = \sum_{d \in D_{TRAIN}} \log P(d/\Theta)$$

$$= \left\{ \sum_{d \in D_{TRAIN}} \log P(s^d / \Theta^s) \right\} + \left\{ \sum_{d \in D_{TRAIN}} \sum_{i=1}^{/d/} \log P(t_i^d / s_i^d, \Theta^t_{s_i^d}) \right\}$$

$$= L_{structure} + L_{contenu}$$

- □ Discriminant learning

$$P(c/x) = \frac{1}{1 + e^{-\log \frac{P(x/c)}{P(x/\bar{c})}}}$$

$$= \frac{1}{1 + e^{-\sum_{i=1}^{n} \log \frac{\theta^{c}_{x_i, pa(x_i)}}{\theta^{\bar{c}}_{x_i, pa(x_i)}}}}$$

- □ Logistic function
  - Error minimization

Fisher Kernel

- Fisher Score :

$$U_X = \nabla_\theta \log P(X/\theta)$$

- **Hypothesis :** The gradient of the log-likelihood is informative about how much a feature « participate » to the generation of an example.

- Fisher Kernel : K(X,Y)=K(Ux,Uy)

Use with the model

$$U_d = \nabla_\Theta \left(\log P(s^d / \Theta^s) + \log P(t^d / s^d, \Theta^t)\right) = \nabla_\Theta \log P(s^d / \Theta^s) + \sum_{l \in \Lambda} \nabla_\Theta \left(\sum_{i / s_i^d = l} \log P(t_i^d / s_i^d, \Theta^t_{tl}) \right)$$

$$U_d \; ? \; ?? \; _? \log P(s^d / ?^s), ? \; _? \; ? \sum_{? i / s_i^d \, ? \, l_1} \log P(t_i^d / s_i^d, ?^t_{tl}) ?, \ldots, ? \; _? \; ? \sum_{? i / s_i^d \, ? \, l_{/?/}} \log P(t_i^d / s_i^d, ?^t_{tl}) ??$$

Sous-vecteur correspondant au gradient sur le modèle de structure

Sous-vecteur correspondant au gradient pour les nœuds de label l1

Sous-vecteur correspondant au gradient pour les nœuds de label $l_{/?/}$

$$? \; _? \log P(t^d / s^d, ?^t)$$

Remark

- Fisher kernels: very large number of parameters
  - On INEX :
    - With flat models : 200 000 parameters
    - With structure models : 20 millions parameters

Conclusion about this faimily of generative models

- □ Natural setting for modeling semi structured multimedia documents
  - ■ Structural probability (Belief network)
  - ■ Content probability (local generative model)
- □ Learning with maximum likelihood, or cross-entropy
- □ Discriminant learning and Fisher Kernel

Outline

- □ Context
- □ Generative tree models
- □ 3 problems
  - Classification
  - Clustering
  - Document restructuration
- □ Experiments
- □ Conclusion and future work
  - XML Document Mining Challenge

Classification

- One model for each category
- 3 XML corpora + 1 multimedia corpus
  - INEX : 12 000 articles from IEEE
    - 18 journals
  - WebKB : Web pages (8K pages)
    - course, department, …7 topics
  - WIPO : XML Documents of patents
    - categories of patents
  - NetProtect (European project) : 100 000 web pages
    - pornographic or not

Categorization : Generative models

		F1 micro	F1 macro
INEX	NB	0.59	0.605
	Structure	**0.619**	**0.622**
WebKB	NB	0.801	0.706
	Structure	**0.827**	**0.743**
WIPO	NB	0.662	0.565
	Structure	**0.677**	**0.604**

Discriminant models

	F1 micro	F1 macro
NB	0.59	0.605
Structure model	0.619	0.622
SVM TF-IDF	0.534	0.564
Fisher kernel	**0.661**	**0.668**
Discriminant learning	0.575	0.600

INEX

	F1 micro	F1 macro
NB	0.801	0.706
Structure model	0.827	0.743
SVM TF-IDF	0.737	0.651
Fisher Kernel	0.823	0.738
Discriminant learning	**0.868**	**0.792**

WebKB

	F1 micro	F1 macro
NB	0.662	0.565
Structure model	0.677	0.604
SVM TF-IDF	0.822	0.71
Fisher Kernel	**0.862**	**0.715**

WIPO

Multimedia model

Director Ang Lee Takes Risks with Mean Green 'Hulk'

LOS ANGELES (Reuters) - Taiwan-born director Ang Lee, perhaps best known for his Oscar-winning "Crouching Tiger, Hidden Dragon," is taking a big risk with the splashy summer popcorn flick ……

FAMILY DRAMA, BIG ACTION

For loyal comic book fans who may think Lee's "Hulk" will be too touchy-feely, think again. " This is a drama, a family drama," said Lee, "but with big action." His slumping shoulders twitch and he laughs…..

	Macroaverage recall	Microaverage recall
NB	89.9 [89.2 ;90.4]	88.4 [87.7 ;89]
Structure model with text	92.5 [91.9 ;93]	92.9 [92.3 ;93.3]
Structure model with pictures	83 [82.2 ;83.7]	82.7 [81.9 ;83.4]
Structure model text and pictures	**93.6** **[93.1 ;94]**	**94.7** **[94.2 ;95.1]**

Classification : conclusion

- □ Structure model is able to handle structure and content information
- □ Both structure and content carry class information
- □ Multimedia categorization
- □ Not in this talk :
  - ■ Categorization of parts of documents
  - ■ Categorization of trees (structure only)

Outline

- Context
- Generative tree models
- 3 problems
  - Classification
  - Clustering
  - Document restructuration
- Experiments
- Conclusion and future work
  - XML Document Mining Challenge

Clustering

- The usual goal is to find groups of similar documents (in a thematic sense)

- The task is different for structured documents :
  - What means “similar documents” :
    - Same structure ?
    - Same content ?
    - Both
  - Open question

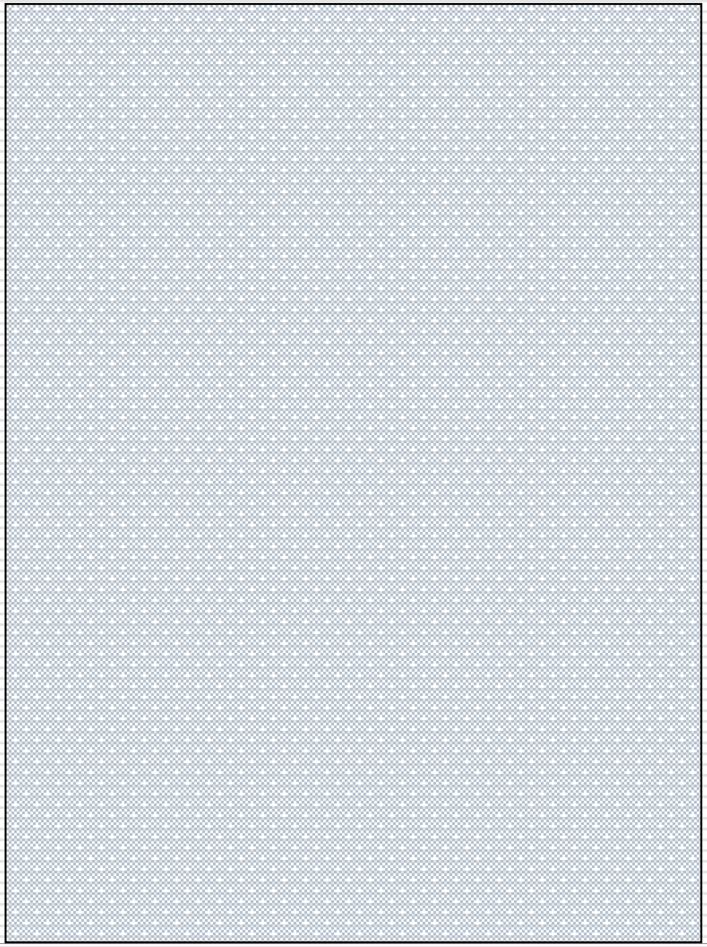
Clustering

- Mixture model :

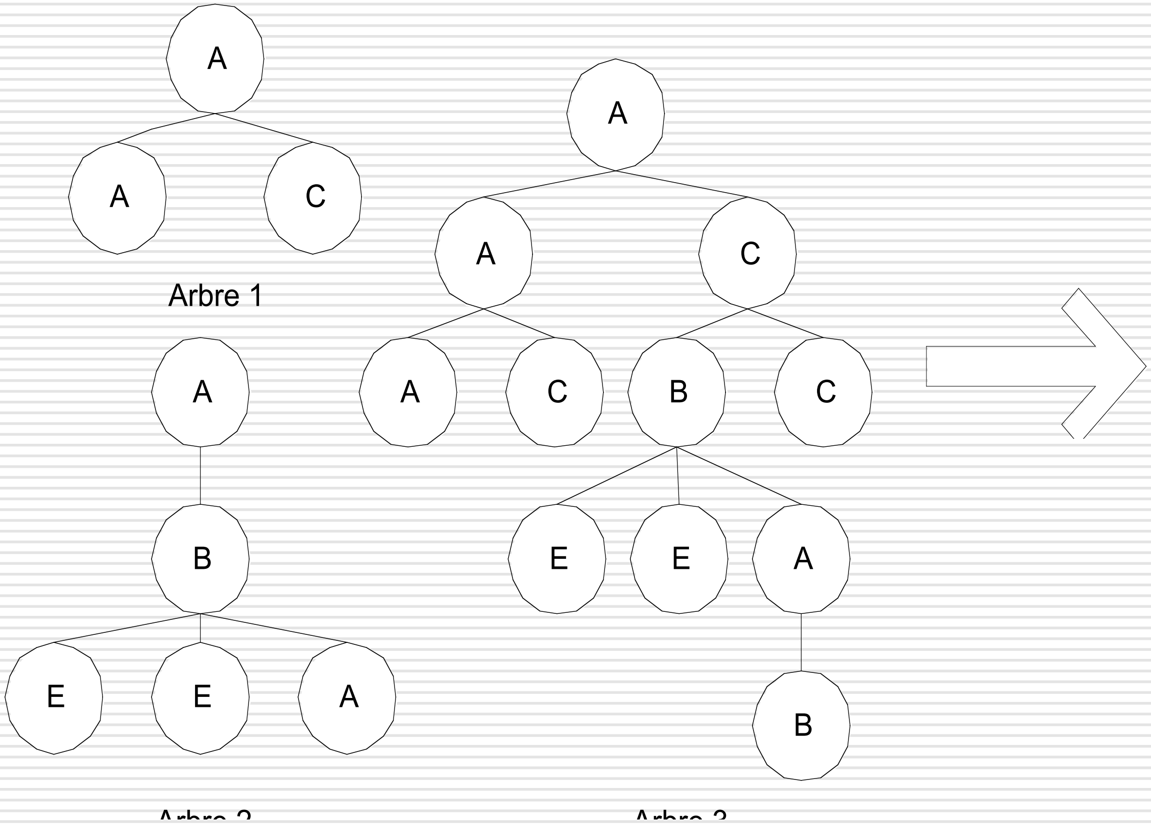
$$P(d/\Theta)=\sum_{i=1}^{/C/}\alpha_{c_i}*P(s^d/\Theta_{c_i})$$

- EM algorithm (CEM)
- Use on the structure (only) using INEX corpus

Different models



The *grammar* model



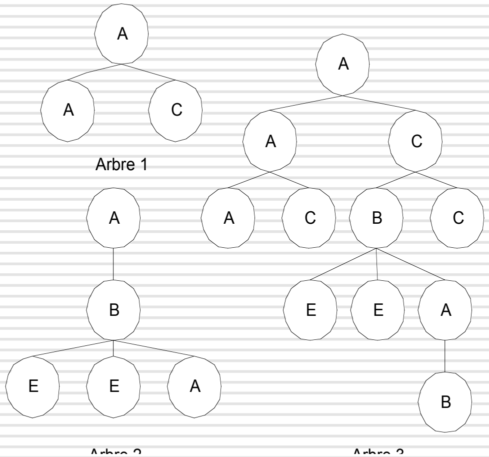
$$P(A, C / A) = \frac{3}{5}$$

$$P(B / A) = \frac{2}{5}$$

$$P(E, E, A / B) = \frac{2}{2}$$

$$P(B, C / C) = \frac{1}{1}$$

Grammar model and DTD



$$P(A,C/A)=\frac{3}{5}$$

$$P(B/A)=\frac{2}{5}$$

$$P(E,E,A/B)=\frac{2}{2}$$

$$P(B,C/C)=\frac{1}{1}$$

$$A \longrightarrow A\,C\,[\frac{3}{5}]$$

$$A \longrightarrow B[\frac{2}{5}]$$

$$B \longrightarrow EEA[1]$$

$$C \longrightarrow BC[1]$$

$$A \longrightarrow A\,C$$

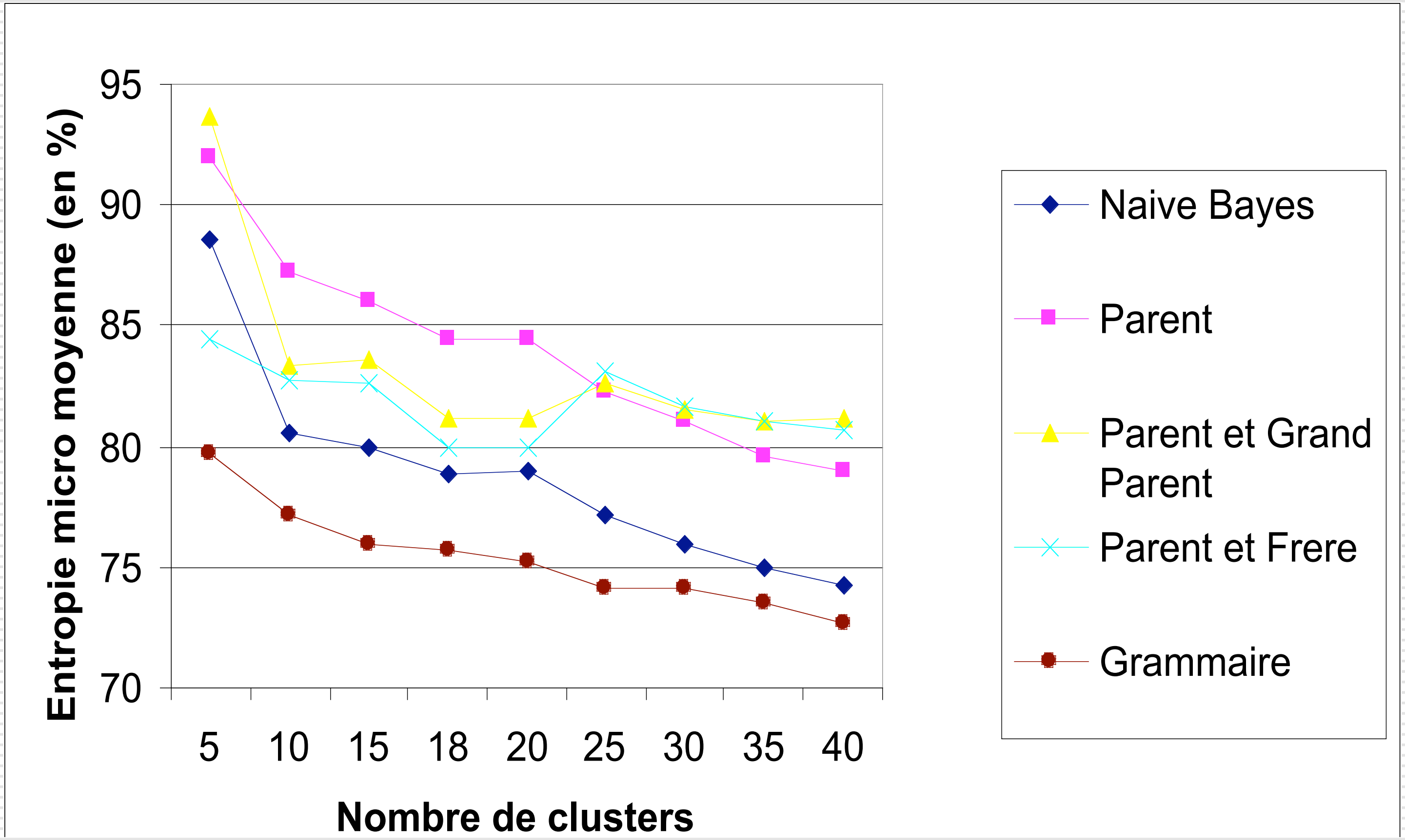
$$A \longrightarrow B$$

$$B \longrightarrow EEA$$

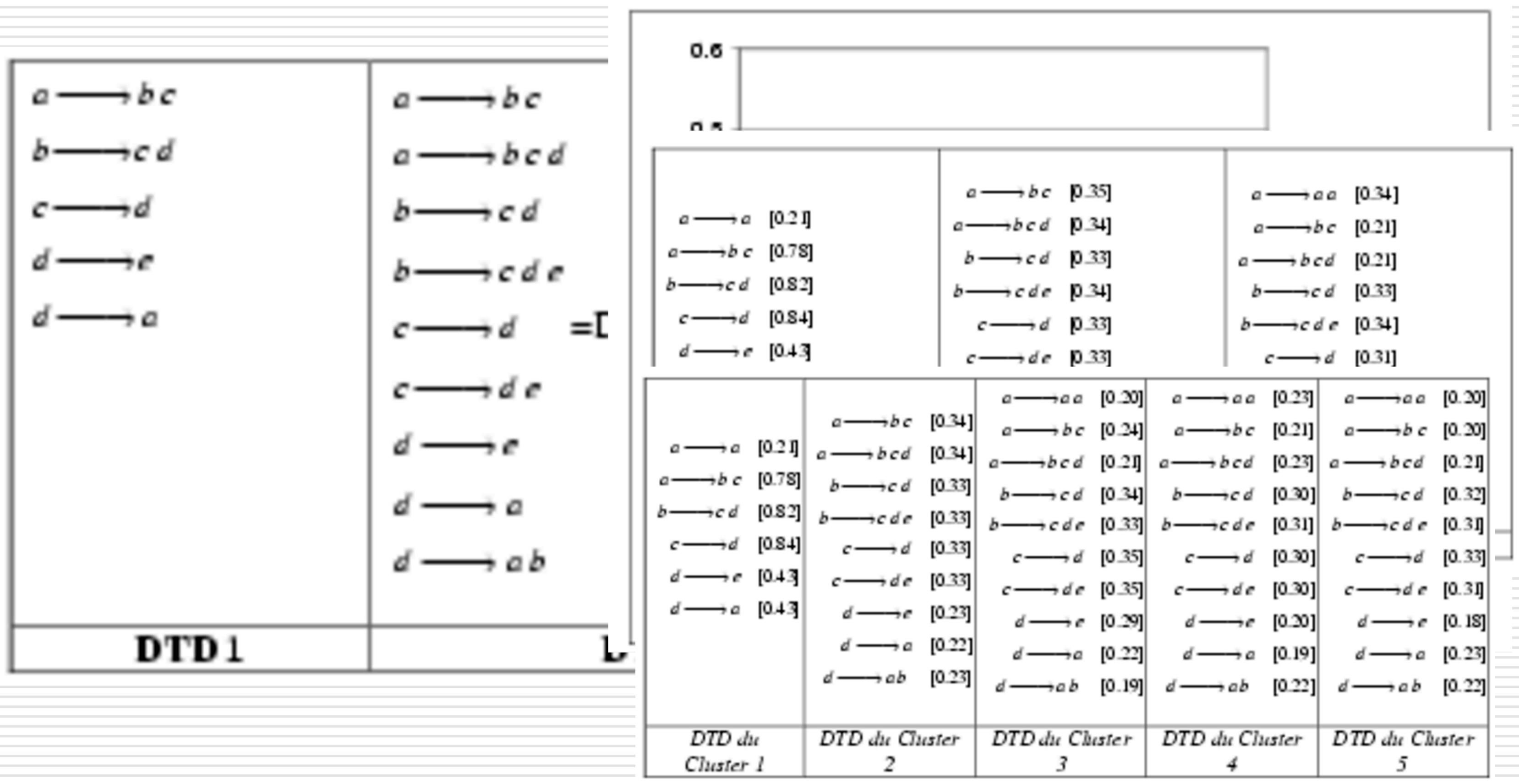
$$C \longrightarrow BC$$

```
<!DOCTYPE A [
<!ELEMENT A  (A,C)>
<!ELEMENT A  (B)    >
<!ELEMENT B  (E,E,A)>
<!ELEMENT  C (B,C)>]>
```

Clustering results



Example of DTDs



$a \rightarrow b\,c$	$a \rightarrow b\,c$
$b \rightarrow c\,d$	$a \rightarrow b\,c\,d$
$c \rightarrow d$	$b \rightarrow c\,d$
$d \rightarrow e$	$b \rightarrow c\,d\,e$
$d \rightarrow a$	$c \rightarrow d$ =D
	$c \rightarrow d\,e$
	$d \rightarrow e$
	$d \rightarrow a$
	$d \rightarrow a\,b$
DTD 1	D

	$a \rightarrow b\,c$ [0.35]	$a \rightarrow a\,a$ [0.34]
	$a \rightarrow b\,c\,d$ [0.34]	$a \rightarrow b\,c$ [0.21]
$a \rightarrow a$ [0.21]	$b \rightarrow c\,d$ [0.33]	$a \rightarrow b\,c\,d$ [0.21]
$a \rightarrow b\,c$ [0.78]	$b \rightarrow c\,d\,e$ [0.34]	$b \rightarrow c\,d$ [0.33]
$b \rightarrow c\,d$ [0.82]	$c \rightarrow d$ [0.33]	$b \rightarrow c\,d\,e$ [0.34]
$c \rightarrow d$ [0.84]	$c \rightarrow d\,e$ [0.33]	$c \rightarrow d$ [0.31]
$d \rightarrow e$ [0.43]		

		$a \rightarrow a\,a$ [0.20]	$a \rightarrow a\,a$ [0.23]	$a \rightarrow a\,a$ [0.20]
	$a \rightarrow b\,c$ [0.34]	$a \rightarrow b\,c$ [0.24]	$a \rightarrow b\,c$ [0.21]	$a \rightarrow b\,c$ [0.20]
$a \rightarrow a$ [0.21]	$a \rightarrow b\,c\,d$ [0.34]	$a \rightarrow b\,c\,d$ [0.21]	$a \rightarrow b\,c\,d$ [0.23]	$a \rightarrow b\,c\,d$ [0.21]
$a \rightarrow b\,c$ [0.78]	$b \rightarrow c\,d$ [0.33]	$b \rightarrow c\,d$ [0.34]	$b \rightarrow c\,d$ [0.30]	$b \rightarrow c\,d$ [0.32]
$b \rightarrow c\,d$ [0.82]	$b \rightarrow c\,d\,e$ [0.33]	$b \rightarrow c\,d\,e$ [0.33]	$b \rightarrow c\,d\,e$ [0.31]	$b \rightarrow c\,d\,e$ [0.31]
$c \rightarrow d$ [0.84]	$c \rightarrow d$ [0.33]	$c \rightarrow d$ [0.35]	$c \rightarrow d$ [0.30]	$c \rightarrow d$ [0.33]
$d \rightarrow e$ [0.43]	$c \rightarrow d\,e$ [0.33]	$c \rightarrow d\,e$ [0.35]	$c \rightarrow d\,e$ [0.30]	$c \rightarrow d\,e$ [0.31]
$d \rightarrow a$ [0.43]	$d \rightarrow e$ [0.23]	$d \rightarrow e$ [0.29]	$d \rightarrow e$ [0.20]	$d \rightarrow e$ [0.18]
	$d \rightarrow a$ [0.22]	$d \rightarrow a$ [0.22]	$d \rightarrow a$ [0.19]	$d \rightarrow a$ [0.23]
	$d \rightarrow a\,b$ [0.23]	$d \rightarrow a\,b$ [0.19]	$d \rightarrow a\,b$ [0.22]	$d \rightarrow a\,b$ [0.22]
DTD du Cluster 1	*DTD du Cluster 2*	*DTD du Cluster 3*	*DTD du Cluster 4*	*DTD du Cluster 5*

Clustering : conclusions

- Mixture model of belief networks
- Different models
- Grammar model is better
  - Able to compute a kind of DTD
- Ill defined problem: clustering of XML documents ?

Outline

- Context
- Generative tree models
- 3 problems
  - Classification
  - Clustering
  - Document restructuration
- Experiments
- Conclusion and future work
  - XML Document Mining Challenge

Structural heterogeneity

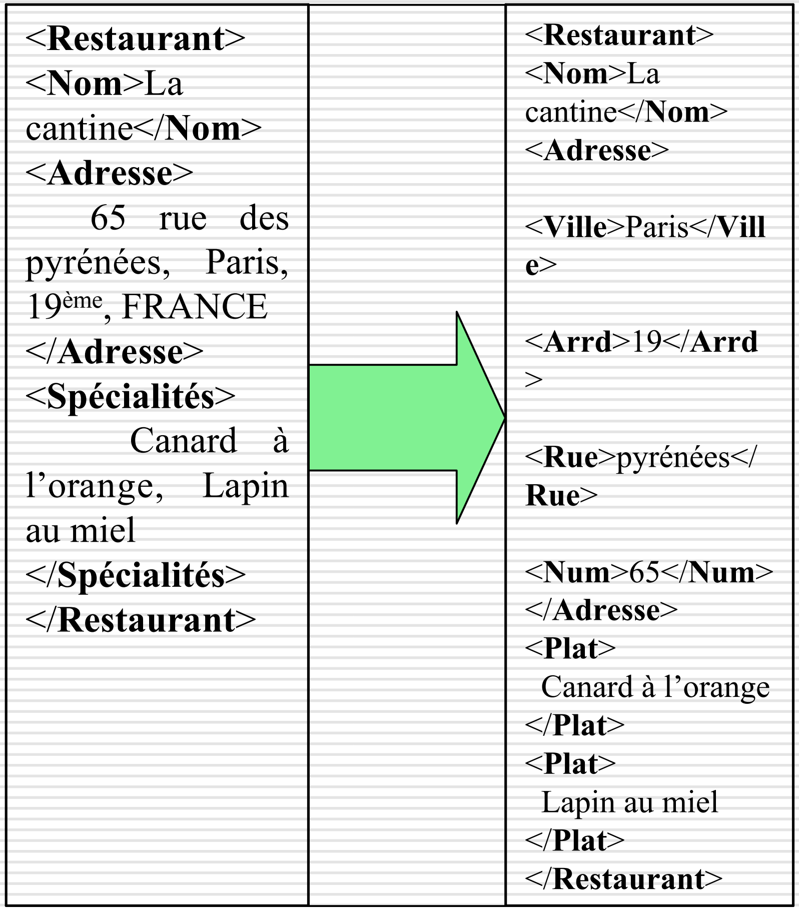
<**Restaurant**> <**Nom**>Tokyo Bar</**Nom**> <**Adresse**> <**Ville**>Paris</**Ville**> <**Arrd**>19</**Arrd**> <**Rue**>Bolivar</**Rue**> <**Num**>127</**Num**> </**Adresse**> <**Plat**>Sushi</**Plat**> <**Plat**>Sashimi</**Plat**> </**Restaurant**>	<**Restaurant**> <**Nom**>La cantine</**Nom**> <**Adresse**> 65 rue des pyrénées, Paris, $19^{\text{ème}}$, FRANCE </**Adresse**> <**Spécialités**> Canard à l'orange, Lapin au miel </**Spécialités**> </**Restaurant**>	<**Restaurant**> <**Nom**>L'olivier</**Nom**> <**Description**> Ce joli restaurant localisé près du métro Jaurès, au 19 du boulevard de la vilette, perdu dans le $19^{\text{ème}}$ arrondissement de Paris propose une cuisine italienne, notamment des pâtes fraîches au 3 fromages. </**Description**> </**Restaurant**>

- □ Problem: Query heterogeneous XML databases or collections, Storage, etc
- □ Needs to know the correspondence between the structured representations

Document mapping problem

- Problem
  - Learn from examples how to map heterogeneous sources onto a predefined target schema
  - Preserve the document semantic
  - Sources: semistructured, HTML, PDF, flat text, etc

Labeled tree mapping problem



```
<Restaurant>
<Nom>La cantine</Nom>
<Adresse>
    65 rue des pyrénées, Paris, 19ème, FRANCE
</Adresse>
<Spécialités>
    Canard à l'orange, Lapin au miel
</Spécialités>
</Restaurant>
```

```
<Restaurant>
<Nom>La cantine</Nom>
<Adresse>
<Ville>Paris</Ville>
<Arrd>19</Arrd>
<Rue>pyrénées</Rue>
<Num>65</Num>
</Adresse>
<Plat>
  Canard à l'orange
</Plat>
<Plat>
  Lapin au miel
</Plat>
</Restaurant>
```

Document mapping problem

- Central issue: Complexity
  - Large collections
  - Large feature space: 10^3 to 10^6
  - Large search space (exponential)

- Approach
  - Learn generative models of XML target documents from a training set
  - Decoding of unknown sources according to the learned model

Learning the correspondence via examples

- Why using ML for structure matching ?
  - Multiple sources: variability, documents do not follow the schema, collection growth, etc
  - Web sources: DTDs, Schema are often unknown or do not exist

Learning correspondence

- Data centered view (Doan et al.)
  - Multiple independent classifier combination
  - Centralized (mediator) or P2P
  - 1:1 or m:n transformations
- Document centered view
  - Document conversion (Xerox)
    - rendering format (HTML, PDF, etc) -> XML predefined DTD format
  - Information retrieval (LIP6)
    - Content and structure queries (e.g. INEX)

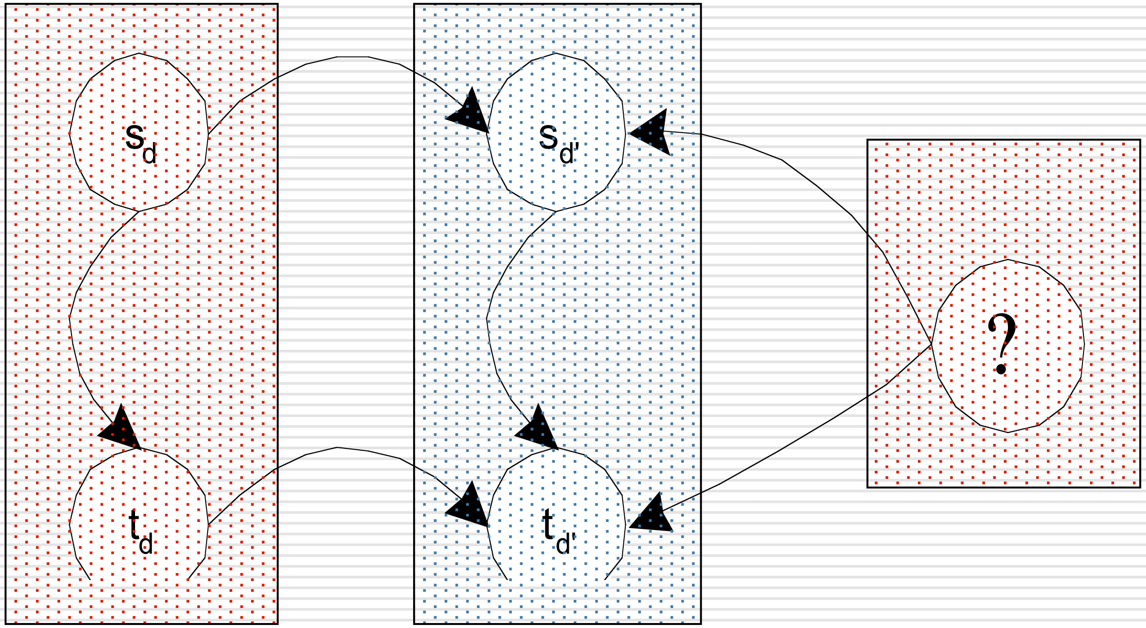
Problem formulation

Given

- S_T a target format
- $d_{sin(d)}$ an input document
- Find the most probable target document

$$d_{S_T} = \underbrace{\arg\max_{d' \in S_T}}_{\text{Decoding}} \underbrace{P(d' \mid d_{S_{in(d)}})}_{\text{Learned transformation model}}$$

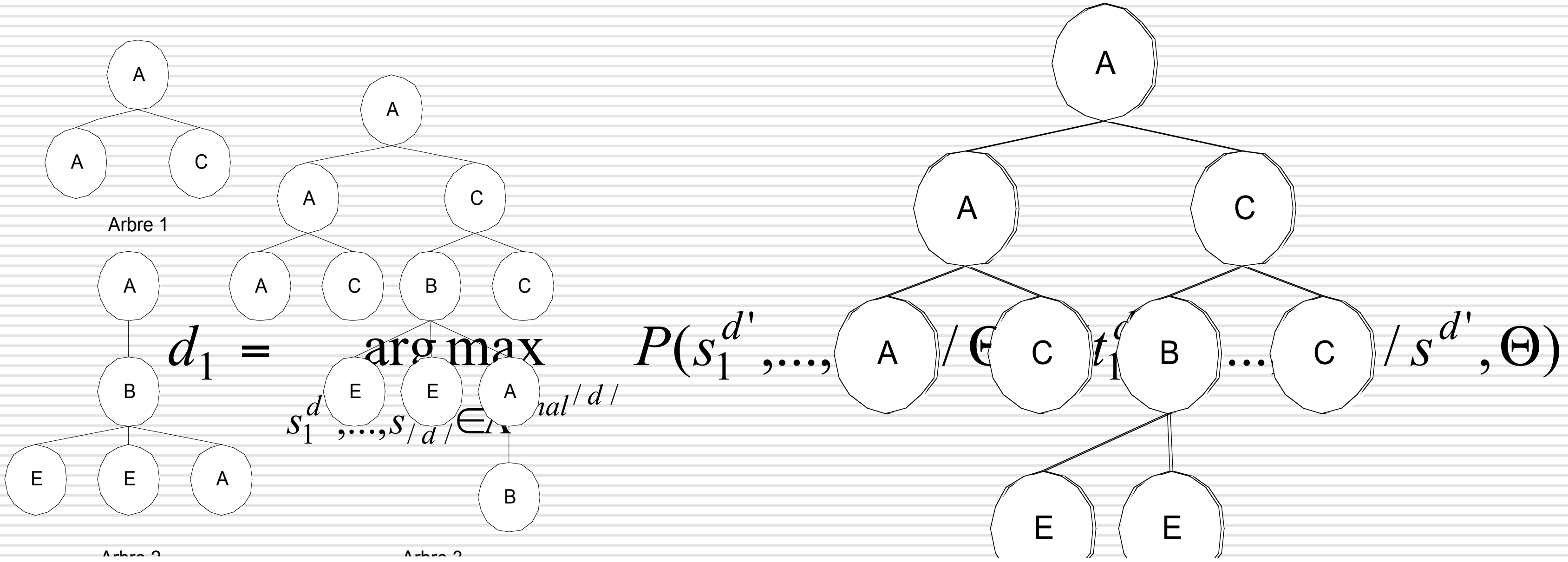
General restructuration model



$$d_1 = \underset{d'}{\operatorname{argmax}} P(s^{d'} / s^d, \Theta) P(t^{d'} / s^{d'}, t^d, \Theta)$$

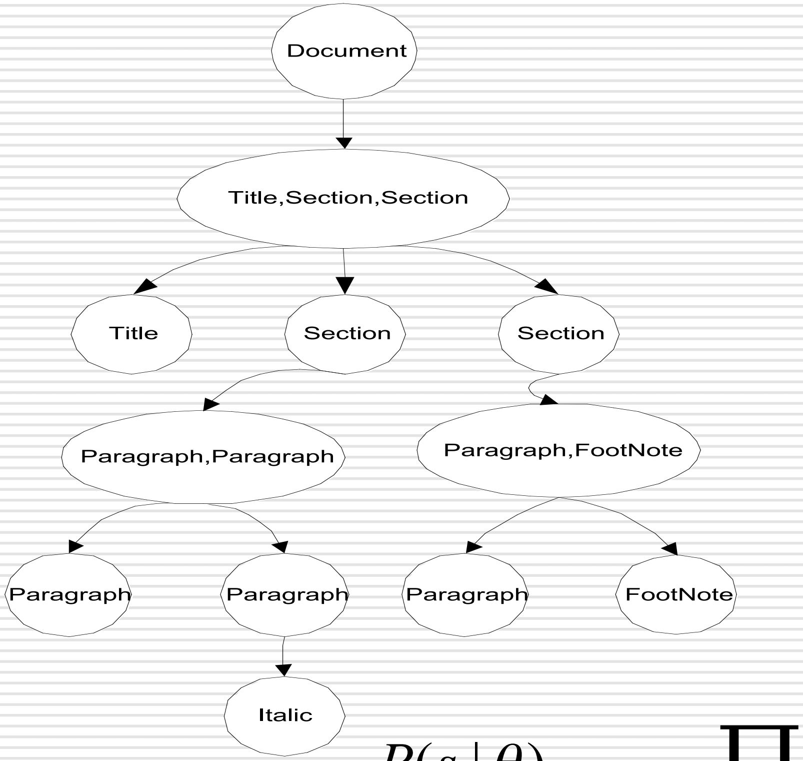
Instance 1 : Label mapping

- Subtask of structure mapping
  - Tree structure remains unchanged
  - Learn to automatically label nodes



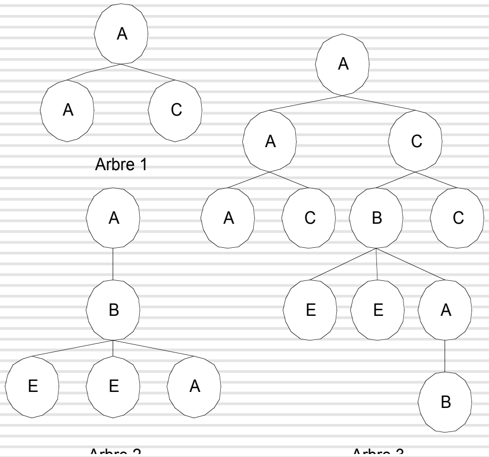
$$d_1 = \arg\max_{s_1^d,\dots,s_{/d/}^d \in \Lambda \ldots al^{/d/}} P(s_1^{d'},\dots, \ldots / \Theta \ldots t_1^{d'} \ldots / s^{d'},\Theta)$$

Document structure model



$$P(s \mid \theta) = \prod_{\text{all nodes } n \text{ in } d} P(childrentags(n) \mid tag(n), \theta)$$

PCFG model



$$P(A, C / A) = \frac{3}{5}$$

$$P(B / A) = \frac{2}{5}$$

$$P(E, E, A / B) = \frac{2}{2}$$

$$P(B, C / C) = \frac{1}{1}$$

$$A \longrightarrow A\,C\,[\frac{3}{5}]$$

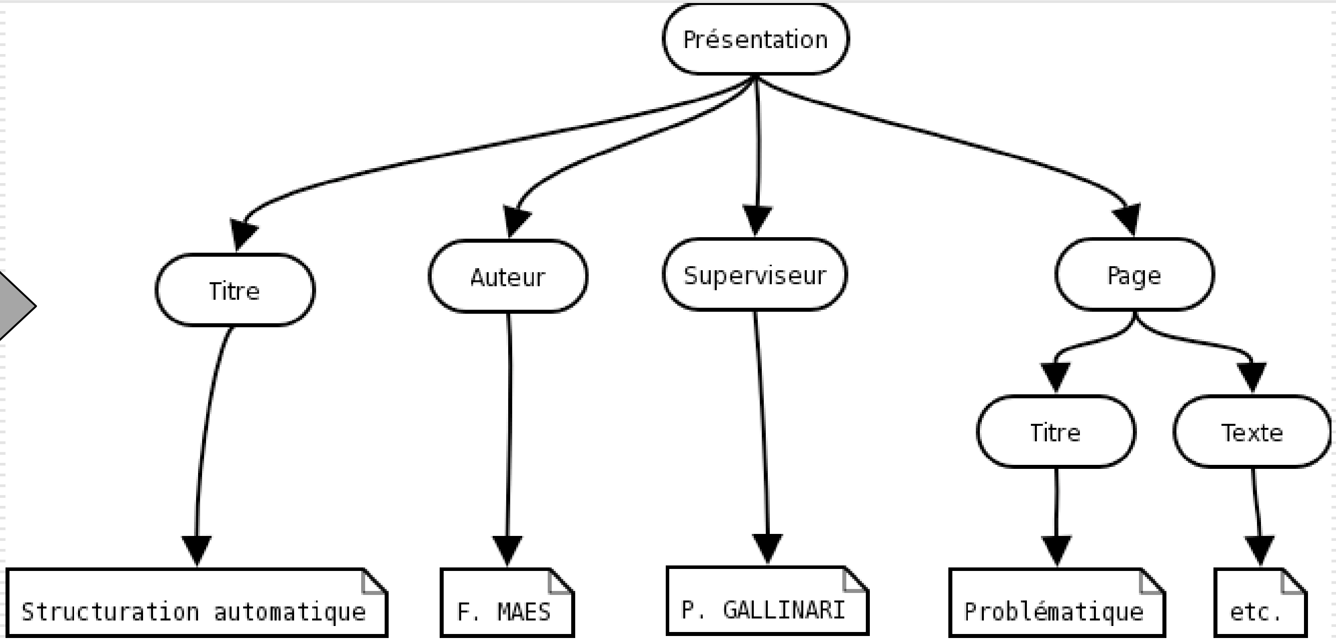
$$A \longrightarrow B[\frac{2}{5}]$$

$$B \longrightarrow EEA[1]$$

$$C \longrightarrow BC[1]$$

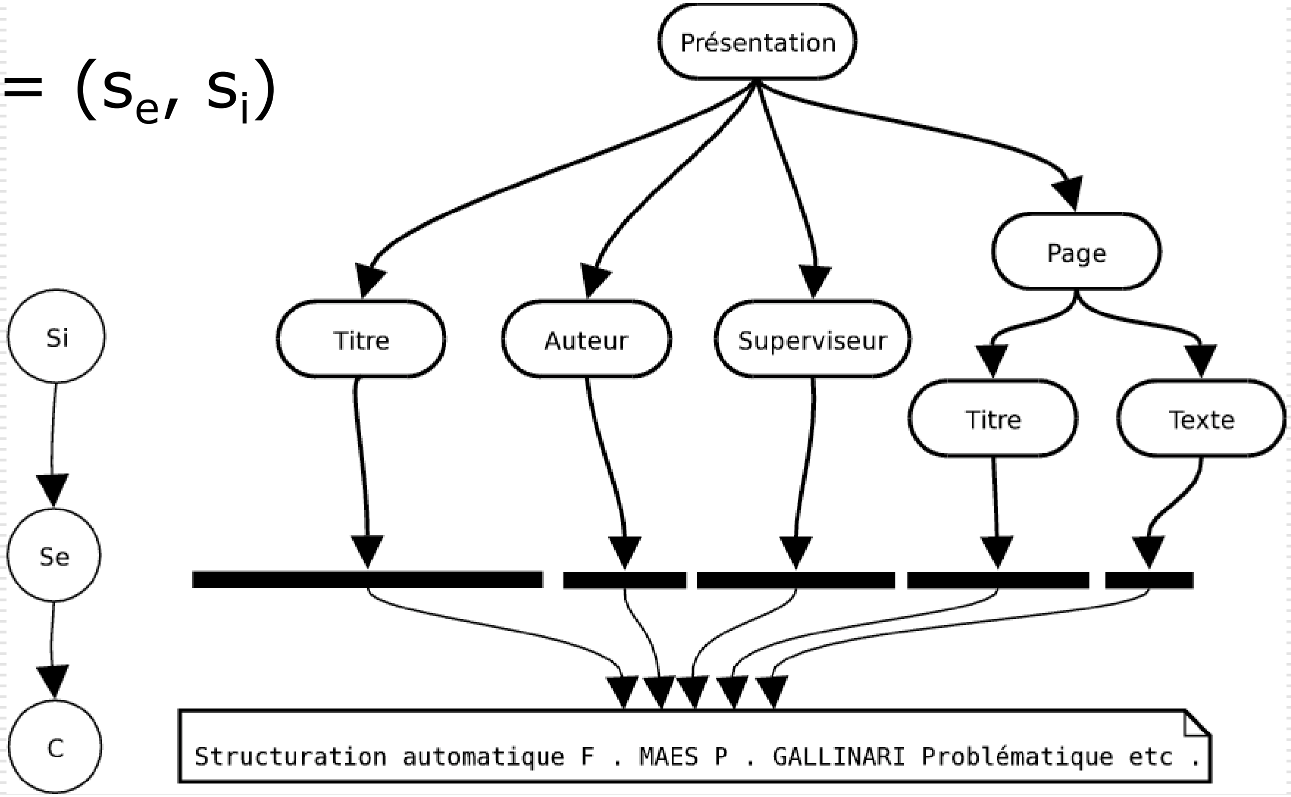
Instance 2: plain text structuring

Structuration
automatique
F
.
MAES
P
.
GALLINARI
Problématiqu
e
etc
.



Stochastic model

d = (c, s) s = (s_e, s_i)


Présentation
Si
Se
C
Titre
Auteur
Superviseur
Page
Titre
Texte
Structuration automatique F . MAES P . GALLINARI Problématique etc .

$$s^* = \underset{(s_i,s_e)\in\mathcal{S}}{\operatorname{argmax}} \frac{P[s_i]P[s_e|s_i]P[c|s_e]}{P[c]}$$

Sub-optimal approach

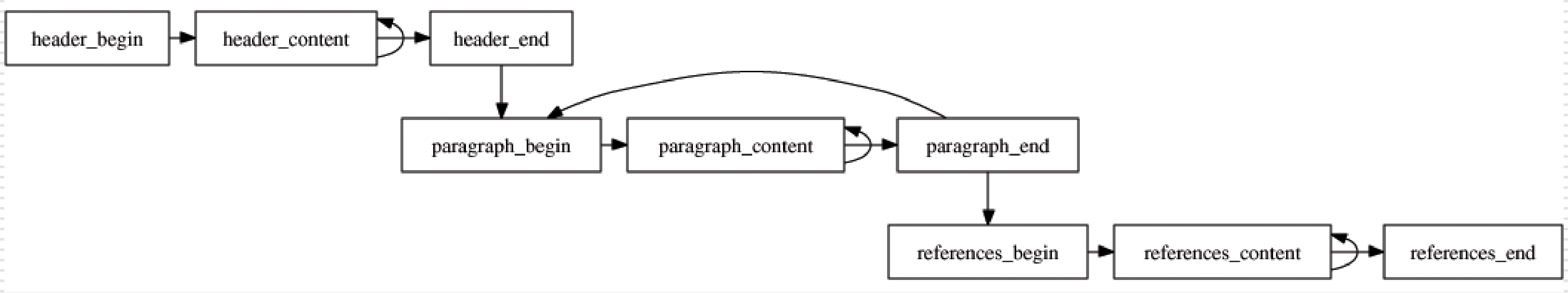
- □ Segmentation and structuration are performed sequentially

$$\max_{(s_i,s_e)\in\mathcal{S}} \log(P[s_i]\cdot P[s_e|s_i]) + \log(P[c|s_e])$$

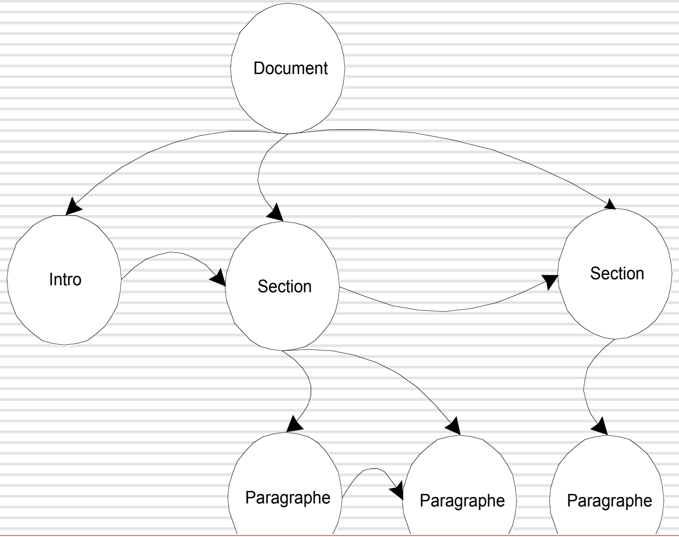
$$\simeq \max_{s_e\in\mathcal{S}_e} \underbrace{\log(P[c|s_e])}_{\text{Segmentation}} + \max_{s_i\in\mathcal{S}_i} \underbrace{\log(P[s_i]\cdot P[s_e|s_i])}_{\text{Structure Extraction}}$$

Models

- Segmentation: HMM



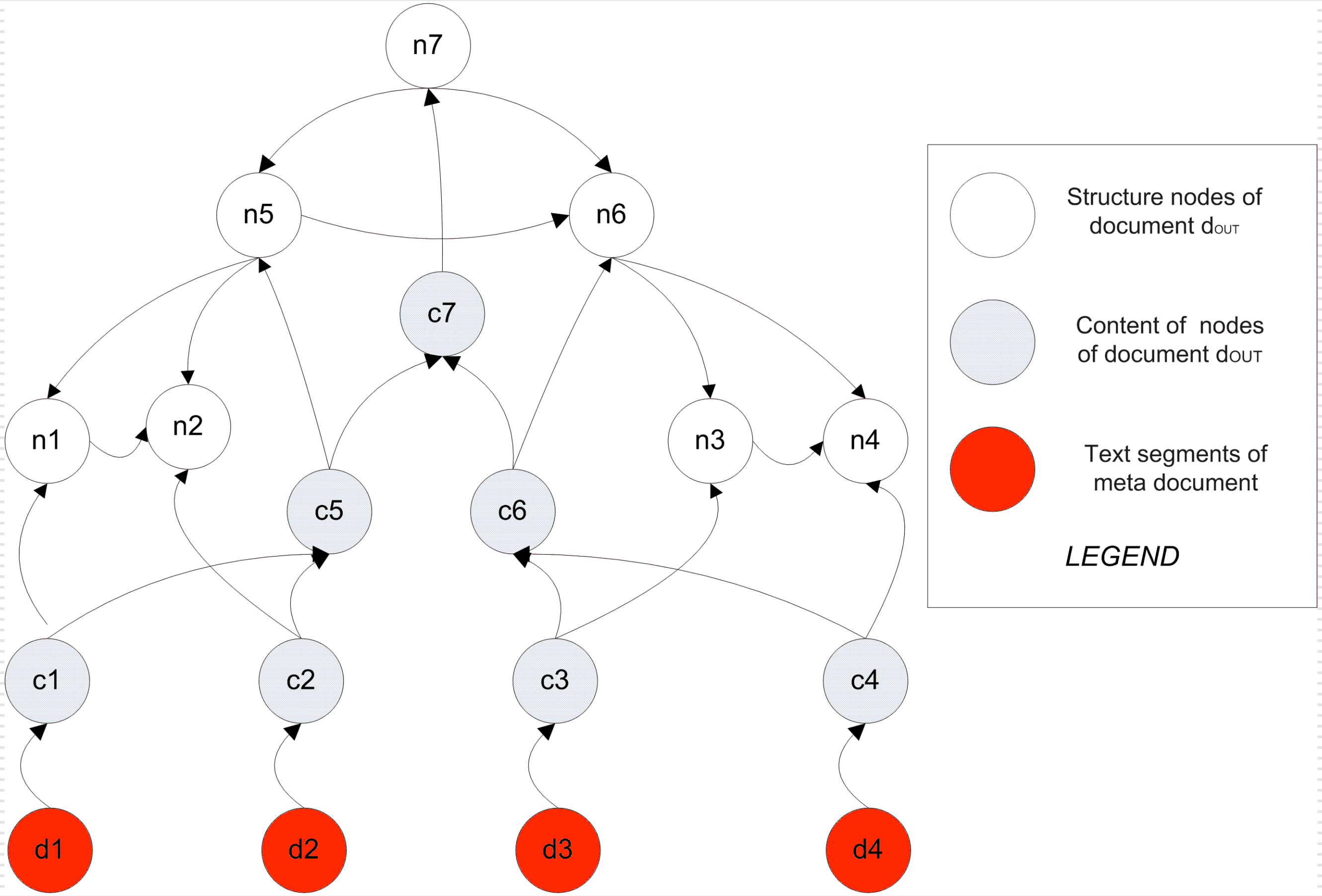
- Structure



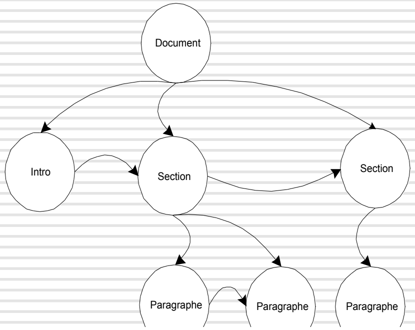
Instance 3 : HTML to XML

- Hypothesis
  - Input document
    - HTML tags mostly for visualization
    - Remove tags
    - Keep only the segmentation (leaves)
  - Transformation
    - Leaves are the same in the HTML and XML document
    - Target document model: node label depends only on its local context
      - Context = content, left sibling, father

Problem representation


n7
n5
n6
c7
n1
n2
n3
n4
c5
c6
c1
c2
c3
c4
d1
d2
d3
d4
Structure nodes of document dOUT
Content of nodes of document dOUT
Text segments of meta document
LEGEND

Model and training



- □ Probability of target tree

$$P(d_T | d_{Sin(d)}) = P(d_T | d_1,\ldots,d_{|d|})$$

$$P(d_T | d_1,\ldots,d_{|d|}) = \prod_{n_i} P(n_i | c_i, sib(n_i), father(n_i))$$

- □ Document model : max-entropy conditional model learned from a training set of target docs

$$P(n_i|c_i, sib(n_i), father(n_i)) = \frac{1}{Z_{c_i,sib(n_i),father(n_i)}} exp\left(< W_{n_i}, F_{c_i,sib(n_i),father(n_i)} >\right)$$

Decoding

- □ Solve

$$d_{S_T} = \arg\max_{d' \in S_T} P(d' |\ d_{S_{in(d)}})$$

$$d_{s_{FINAL}} = \operatorname*{argmax}_{\substack{d_T \text{ such as} \\ (d^1,\ldots,d^{|d|})=(c_1,\ldots,c_{|d|})}} \prod_{n_i \in N_{d_T}} \frac{exp\left(< W_{n_i}, F_{c_i,sib(n_i),father(n_i)} >\right)}{Z_{c_i,sib(n_i),father(n_i)}}$$

- □ Exact Dynamic Programming decoding
  - $O(|\text{Leaf nodes}|^3.|\text{tags}|)$
- □ Approximate solution with LASO (Hal Daume ICML 2005)
  - O(|Leaf nodes|.|tags||tree nodes|)

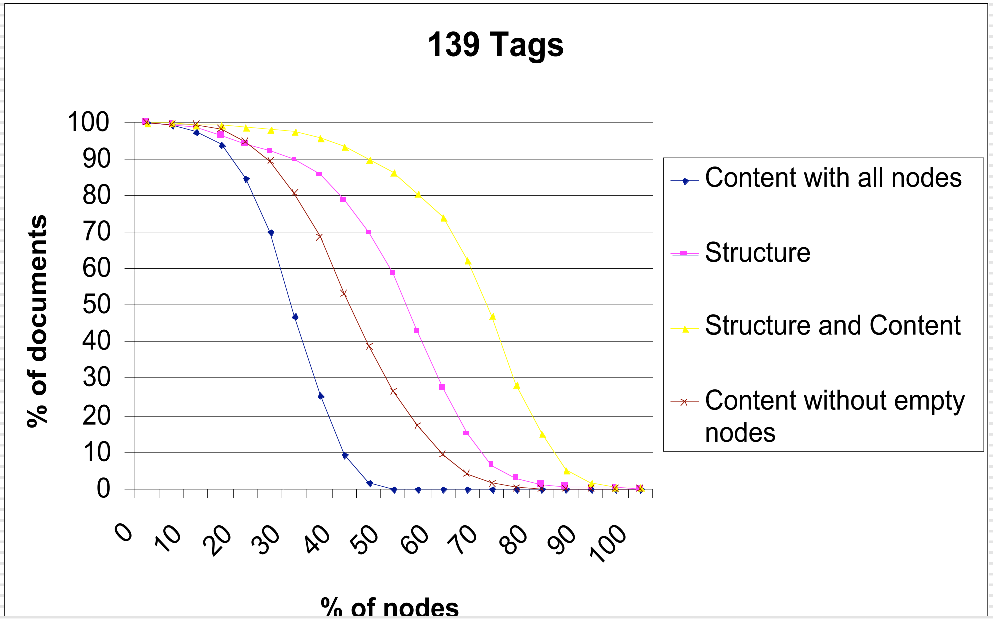
Experiments

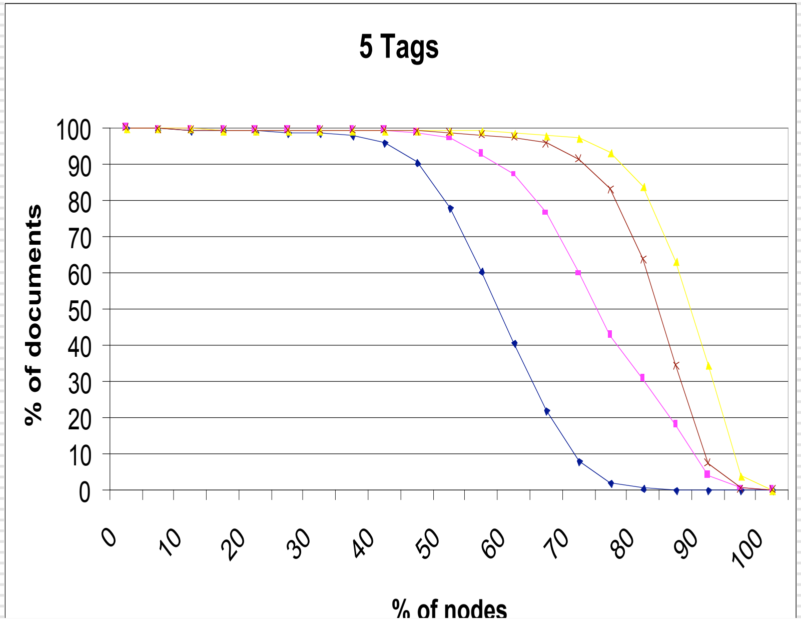
- INEX corpus:
  - IEEE collection (XML) :
    - 12 000 documents (training : 7 800 , Test : 4 200)
    - ≈ 5 000 000 content nodes
    - 139 tags
    - Mean document depth ≈ 7
    - vocabulary : ≈ 22 000 mots
  - test corpus :
    - *Transaction On …series*
    - Unlabeled documents (tags removed)

Instance 1 : Label mapping - results

	Content	Structure	Struct + Content	naïve model
5 tags	58%	72,90%	**86,50%**	79,3%
139 tags	27,80%	49,70%	**65,30%**	9,5%

Instance 1 : IR adapted measure





Instance 2: plain text structuring Results

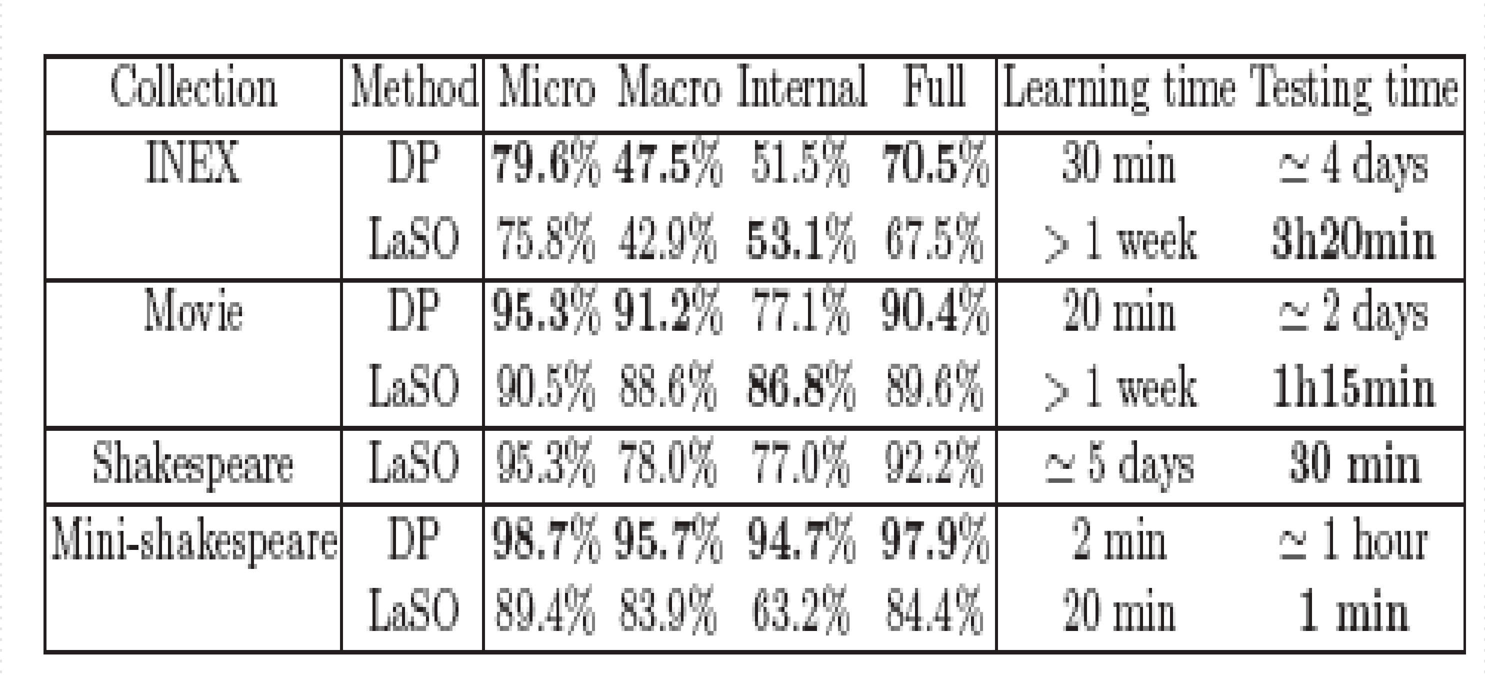
Models	labeling	Segmentation (leafs)	Structuration (internal nodes)
Exact + TMM	92,8 %	75,7%	31,2%
HMM + TMM	91,5%	24,6%	22,8%

- Extreme structuration instance
- Exact + TMM: degraded version of HTML documents structuration

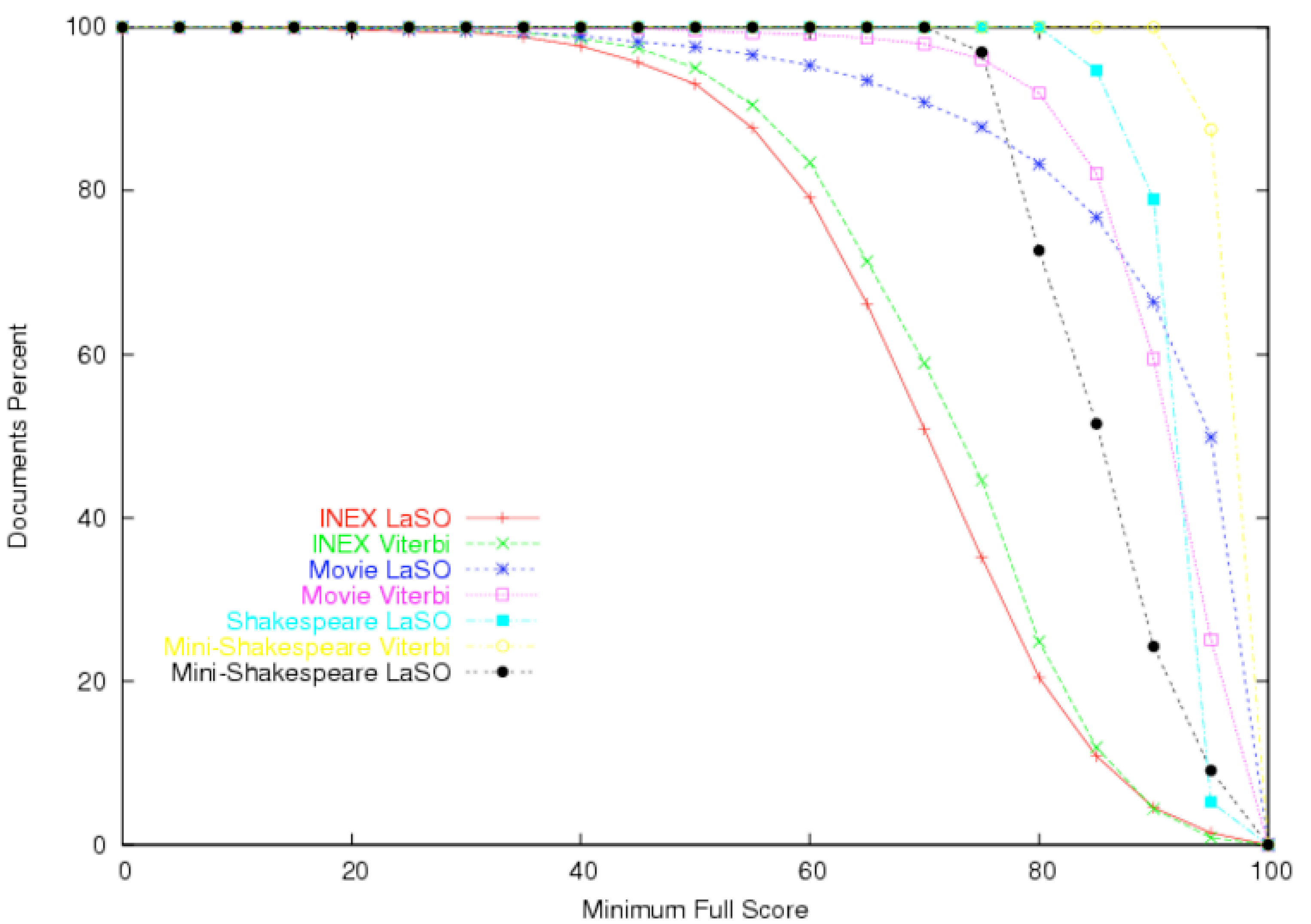
Instance 3 HTML to XML

- IEEE collection / INEX corpus
  - 12 K documents,
    - Average: 500 leaf nodes, 200 int nodes, 139 tags
- Movie DB
  - 10 K movie descriptions (IMDB)
    - Average: 100 leaf nodes, 35 int. nodes, 28 tags
- Shakespeare 39 plays
  - Few doc, but:
    - Average: 4100 leaf nodes, 850 int nodes, 21 tags
- Mini-Shakespeare
  - Randomly chosen 60 scenes from the plays
    - 85 leaf nodes, 20 int. nodes, 7 tags

Performances



Collection	Method	Micro	Macro	Internal	Full	Learning time	Testing time
INEX	DP	**79.6%**	**47.5%**	51.5%	**70.5%**	30 min	$\simeq$ 4 days
	LaSO	75.8%	42.9%	**53.1%**	67.5%	> 1 week	**3h20min**
Movie	DP	**95.3%**	**91.2%**	77.1%	**90.4%**	20 min	$\simeq$ 2 days
	LaSO	90.5%	88.6%	**86.8%**	89.6%	> 1 week	**1h15min**
Shakespeare	LaSO	95.3%	78.0%	77.0%	92.2%	$\simeq$ 5 days	**30 min**
Mini-shakespeare	DP	**98.7%**	**95.7%**	**94.7%**	**97.9%**	2 min	$\simeq$ 1 hour
	LaSO	89.4%	83.9%	63.2%	84.4%	20 min	**1 min**



Documents Percent
Minimum Full Score
0
20
40
60
80
100
0
20
40
60
80
100
INEX LaSO
INEX Viterbi
Movie LaSO
Movie Viterbi
Shakespeare LaSO
Mini-Shakespeare Viterbi
Mini-Shakespeare LaSO

Conclusion

- □ Document restructuration is a new problem
- □ Tree transformation problem of high complexity (content + structure)
- □ Many different instances
- □ Approach based on generative models of target documents

XML Document Mining Challenge 2006

- Challenge
  - INEX-Delos and Pascal networks of excellence
- Three tasks
  - Classification
  - Clustering
  - Document mapping
- 3 XML corpora
  - IEEE collection
  - IMDB (Movie descriptions)
  - Wikipedia in 4 languages
  - Dead line : june 2006
- Web site : http://xmlmining.lip6.fr
- Email : xmlmining@lip6.fr