Automatic Text Annotating: Methodology

Djamel A. Zighed, Hakim Hacid, Vincent Pisetta

Laboratoire ERIC, Université de Lyon2

September 2005

RUIG project on Social Dialogue Regimes: Progress report 2

METHODOLOGY OF WORKING

I. Corpus

The corpus is divided into groups:

- labelled texts (known violations)
- unlabelled texts (with unknown violations)

Texts with known violations are the 71 texts from which ILO has extracted examples of violations.
We have worked with this corpus during all the study.

II. Terminology extraction

Our aim here, is to extract words and associations of words while keaping their meaning. For that, we use two tools which are described in the following.

1. BRILL

The first step is "tagging". This step aims to assign for each word its grammatical tag. We use a software called Brill[1].

Example of tagging:

the committee reiterates its previous comments concerning the need to guarantee the right of association

the/NN committee/NN reiterates/VBZ its/PRP$ previous/JJ comments/NNS concerning/VBG the/DT need/NN to/TO guarantee/VB the/DT right/NN of/IN association/NN

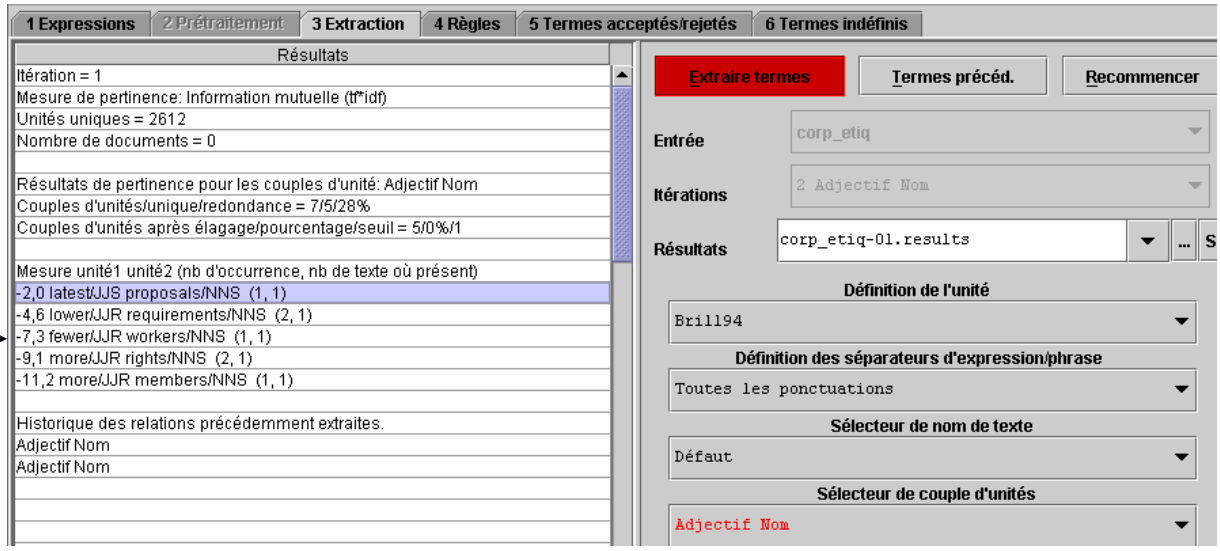
Each word is followed by his tag (NN = noun, DT = determinant, JJ = adjective.....)

2. EXIT

The second phase is to extract "candidate terms". We use the software EXIT [2]. It extracts all binary relations in the form

- noun→verb
- adjective→noun

….



C. An example of EXIT

Some candidate terms

We have to do a selection of these candidate terms by statistical criteria (mutual information and frequency[2]).

3. Concepts

Thanks to the list of candidate terms, we can determine concepts. Concepts are groups of candidate terms with a similar signification. Here is our list of concepts:

" workers' organisations"
workers organisation
association of workers
workers organization
seafarers organisation
workshop occupation
workers section
persons working

"registration rights"
formation of a trade union
registration of a trade union
right to form
right to join
freedom of association
enterprise trade union
public servants
freedom to associate
servants exercising

"right to strike"
right to strike
consent to a strike
illegality of strike
declare a strike
call a strike
calling a strike
strike action
organization of strike
prohibits a strike
prohibit a strike

"financial aspect"
financial independence
financial activitie
financial management
financial assistance
financial contribution
financial autonomy

"trade union"
trade union

"foreign workers"
foreign workers

"service"
minimum service
public service

"protection"
property rights
protection of association
property of the association
right to property

"monopoly-pluralism"
monopoly
pluralism
unity
new organisations
trade union status

"right to organize"
hold meeting
right to organize
right of association
prohibition of political activities

"election"
vote
elect member
nomination
union officer
union leader
representative members
trade union office
absolute majority

"security"
health
safety
security staff
absenteeism
work stoppage
vital need

"arbitration"
compulsory arbitration
arbitration procedure

"actions"
collective action
protest action

"punition"
penalty
penalties
rebuke
reprimand
imprisonment
legislative measure

III. Modelisation and generalisation's tools

1. Data representation

We need a crosstable to analyse texts and find violations. For each text, we have calculated concepts frequency. So, at each time we encounter a candidate term of a concept, we add "1" to the concept frequency. Our crosstable is like the one illustrated one here after:

Example:

	"punition"	*"actions"*	*"arbitration"*	*"security"*	*"service"*
South Africa2000	0	0	0	0	0
South Africa1998	0	0	0,80820063	0	0
Russia2002	0	0	0	1,41026063	0,74827419
Russia2000	1,9414151	0	0	0	0,74827419
Russia1998	5,17710693	0	0	0	0,74827419
Russia1996	2,58855346	0	0	0	0,74827419
Russia1995	3,23569183	0	0	0	0,74827419
Russia1994	3,23569183	0	0,80820063	0	0
Russia1991	1,9414151	1,55022835	0,80820063	0	0
Poland2002	0	0	0	0	0
Poland2000	0	0	0	0	0,74827419

N.B : frequency are corrected by a statistical method (TF / IDF [3])

2. Generalisation's tools

We have used two different tools to predict violation presence : decision trees[4] and relative neighborhood graphs[5].

2.1 Decison Trees

The aim of decision trees is to find one/more concept(s) which can do a discrimination of presence or absence of a violation. Let us take an example: we try to predict if a "trade union pluralism" violation is present. The corresponding decision tree is illustrated in the figure here after

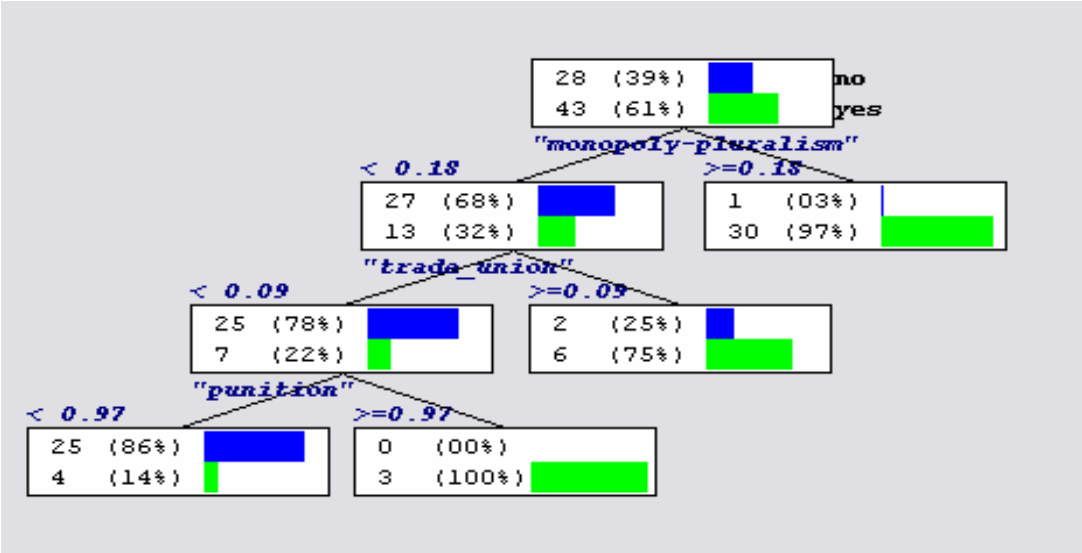


FIG.1 Trade union pluralism's decision tree

As we can see, 43 of the 71 labelled texts have realized a violation of "trade union pluralism". The concept "monopoly-pluralism" can do a good discrimination. In fact, when the coordinate of a text is higher than 0.18 to the concept "monopoly-pluralism", there's a probability of 97% (30/31) that this text contains a violation of "trade union-pluralism". Interpretation is the same for all tree nodes of the tree.
We can extract rules from this tree as the following ones:

- IF "Monopoly-pluralism" >=0,18 THEN Trade union pluralism=*true (probability=97%)*
  OR
- IF "Monopoly-pluralism" <0,18 AND "trade-union" >=0,09 THEN Trade union pluralism=*true (probability=75%)*
  OR
- IF "Monopoly-pluralism" <0,18 AND "trade-union" <0,09 AND "punition">=0,97 THEN Trade union pluralism=*true*
  ELSE
  Trade union pluralism=*false*

Here after other decision trees using other concepts are presented:

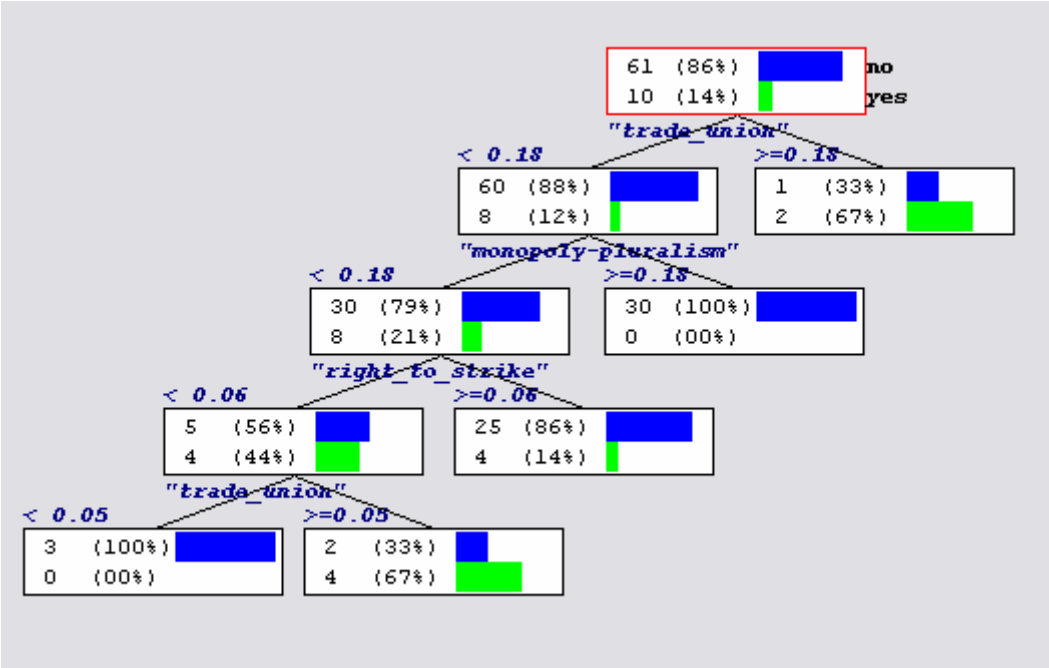


FIG.2 Violation « Protection of property »

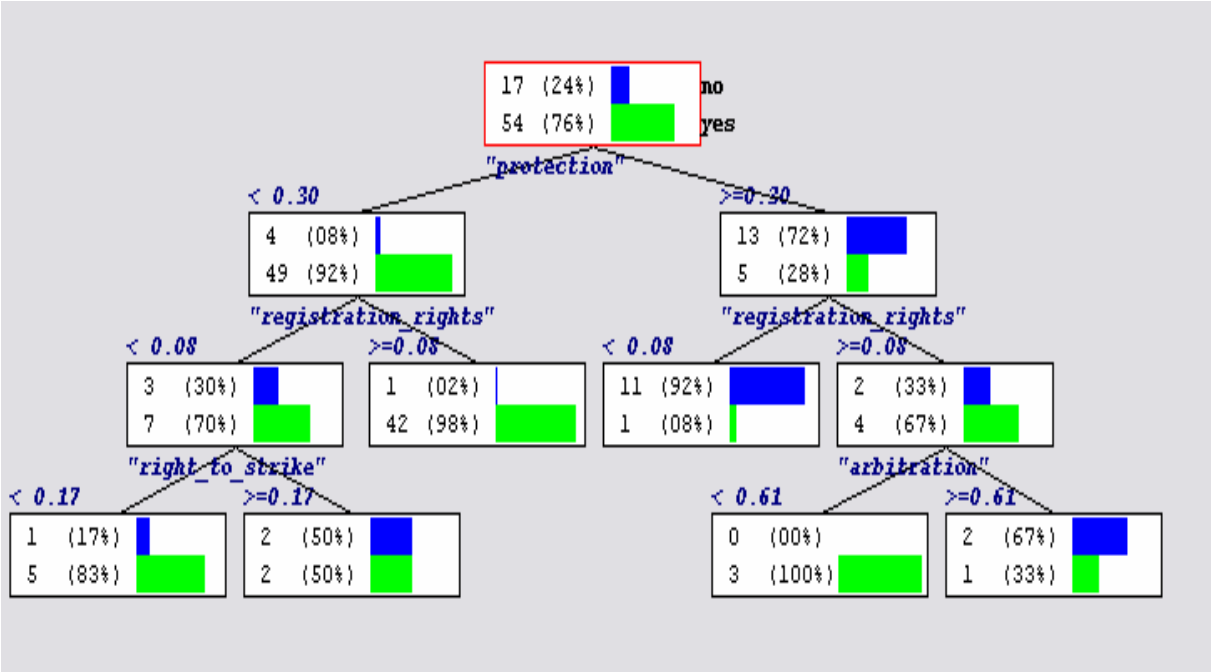


FIG.3 Violation « Right to establish and join workers organisation »

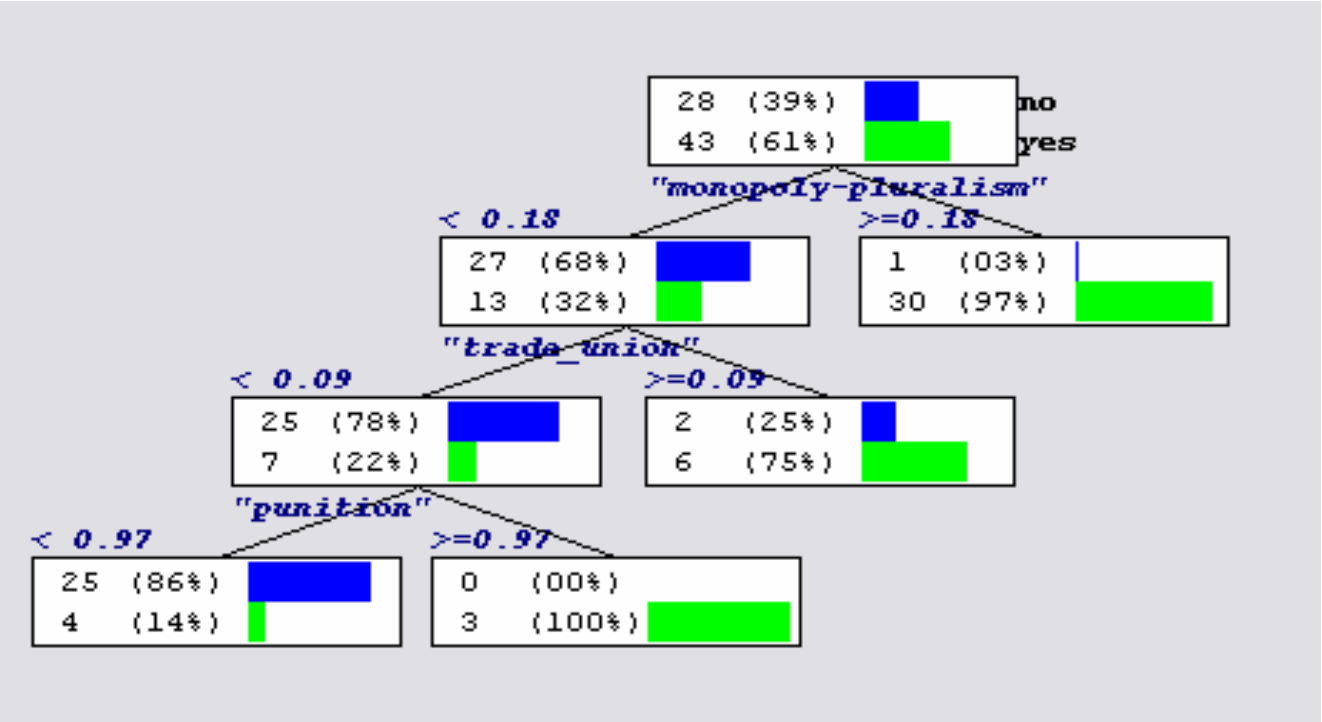


FIG.4 Violation "Trade union pluralism"

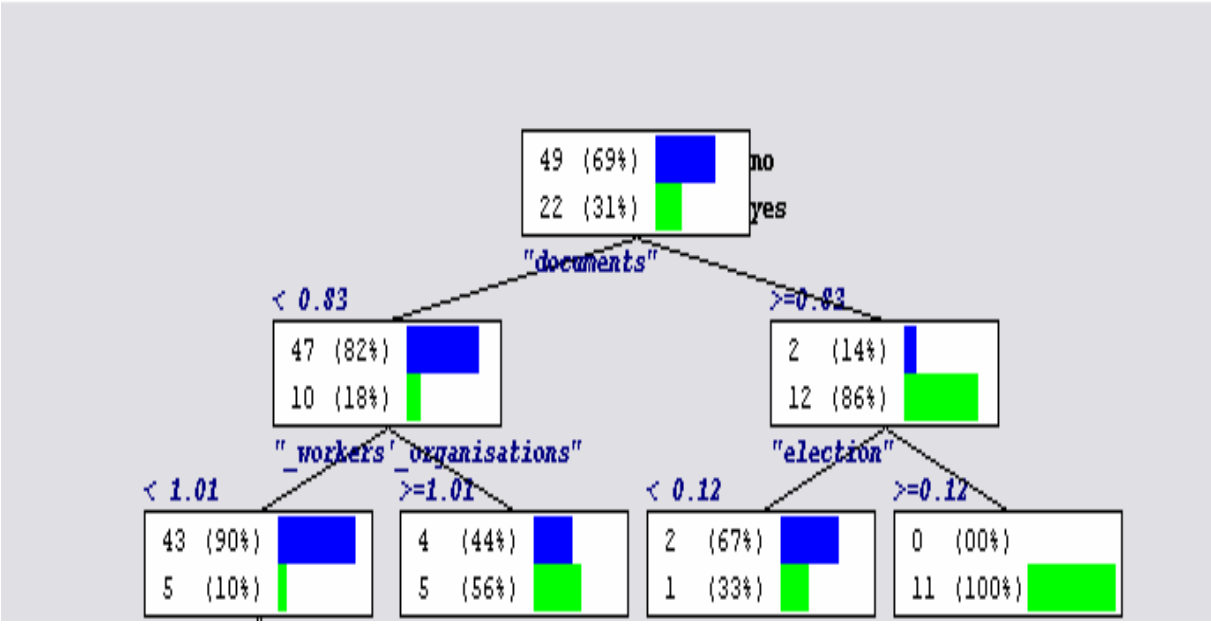


FIG.5 Violation "Establishment and registration of workers' organisations "

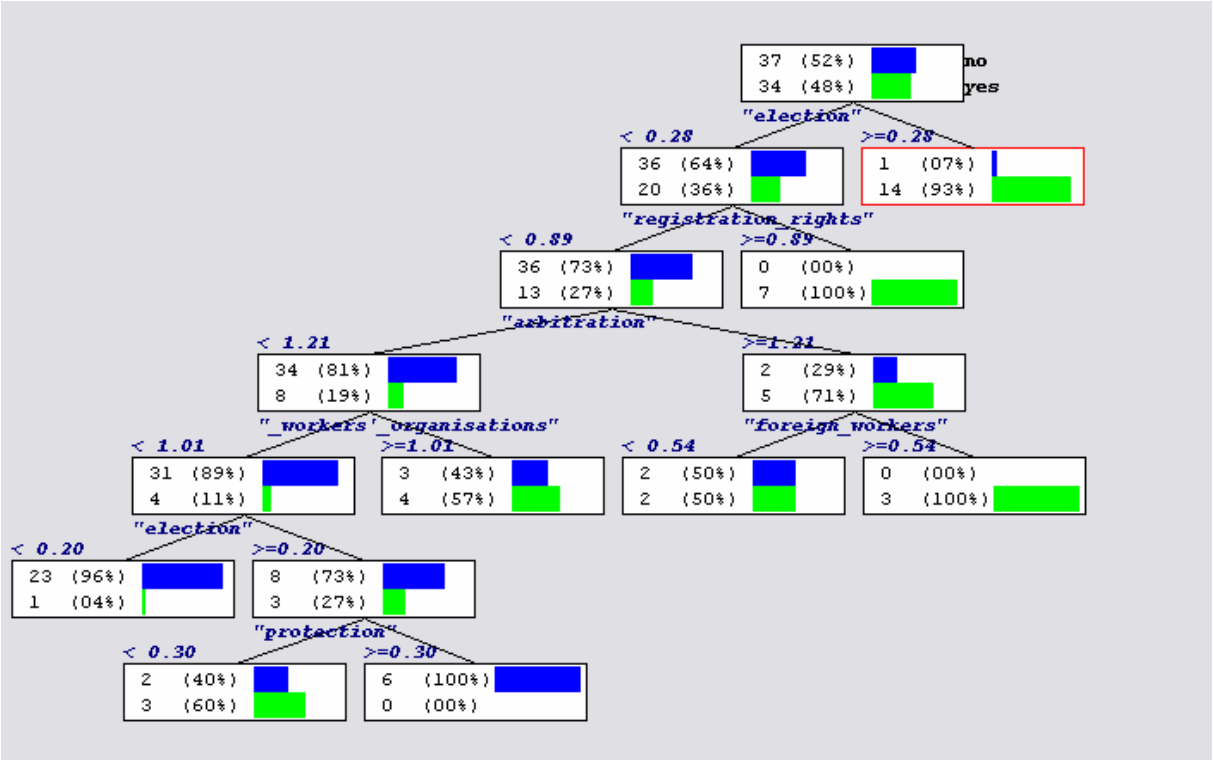


FIG.6 Violation " Election of representatives / Eligibility criteria"

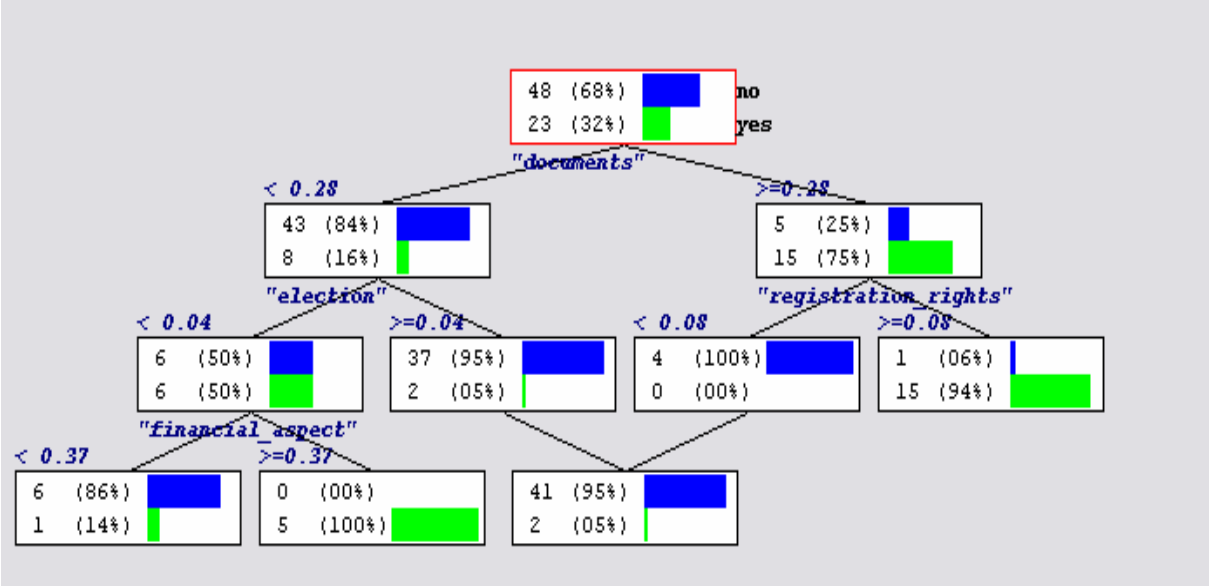


FIG.7 Violation « Administrative/Financial Independence »

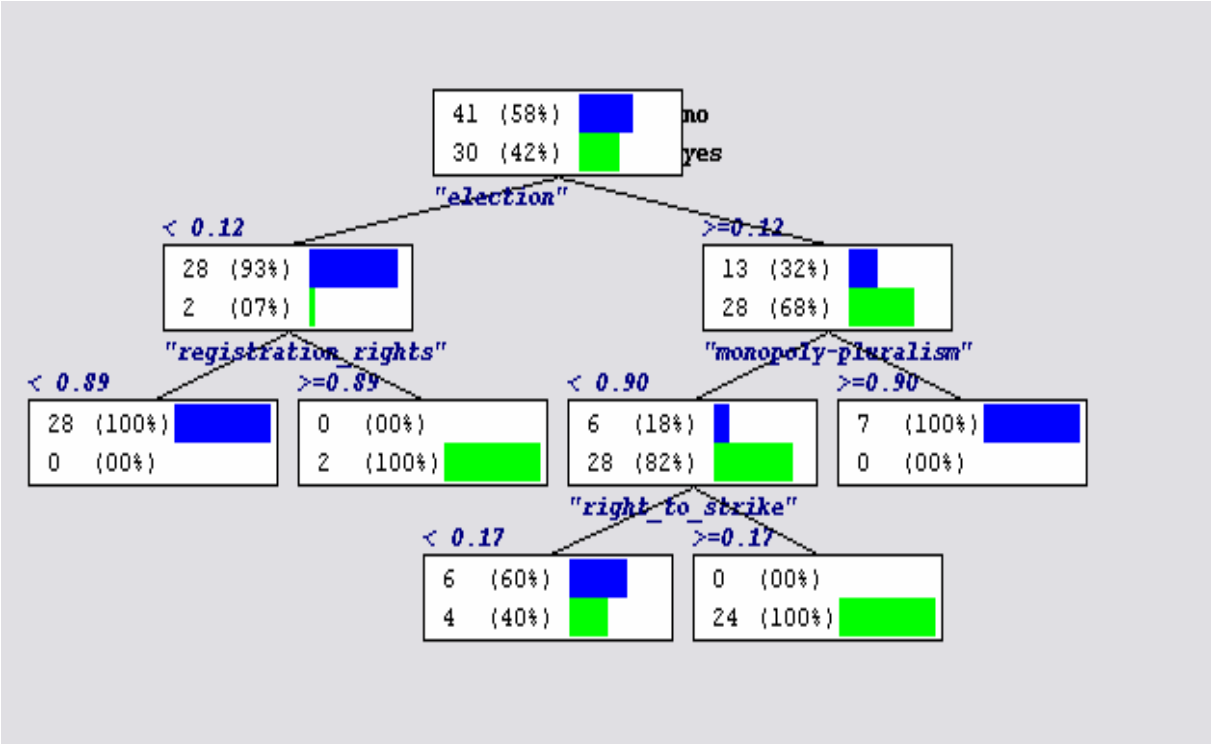


FIG.8 Violation « Organisation of activities »

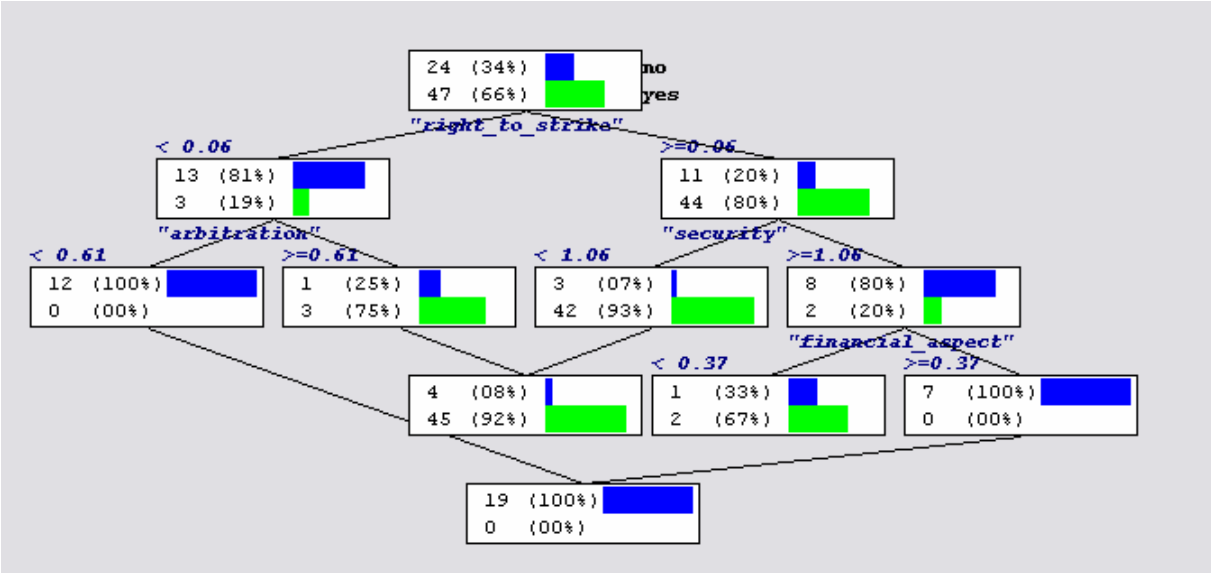


FIG.9 Violation "Restriction on the right to industrial action"

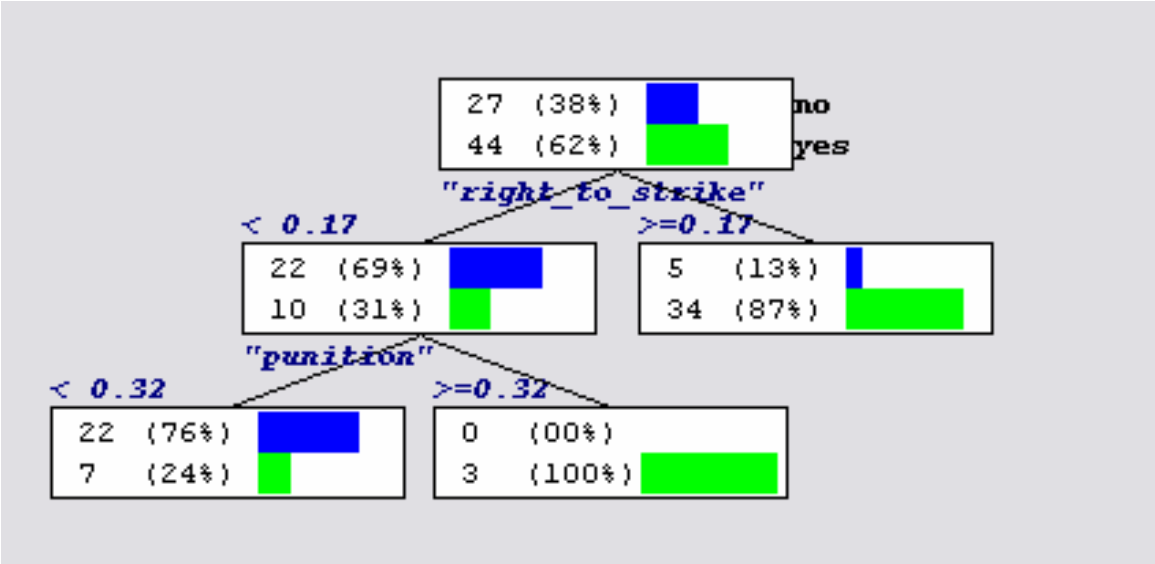


FIG.10 Violation "Conditions for lawful industrial action"

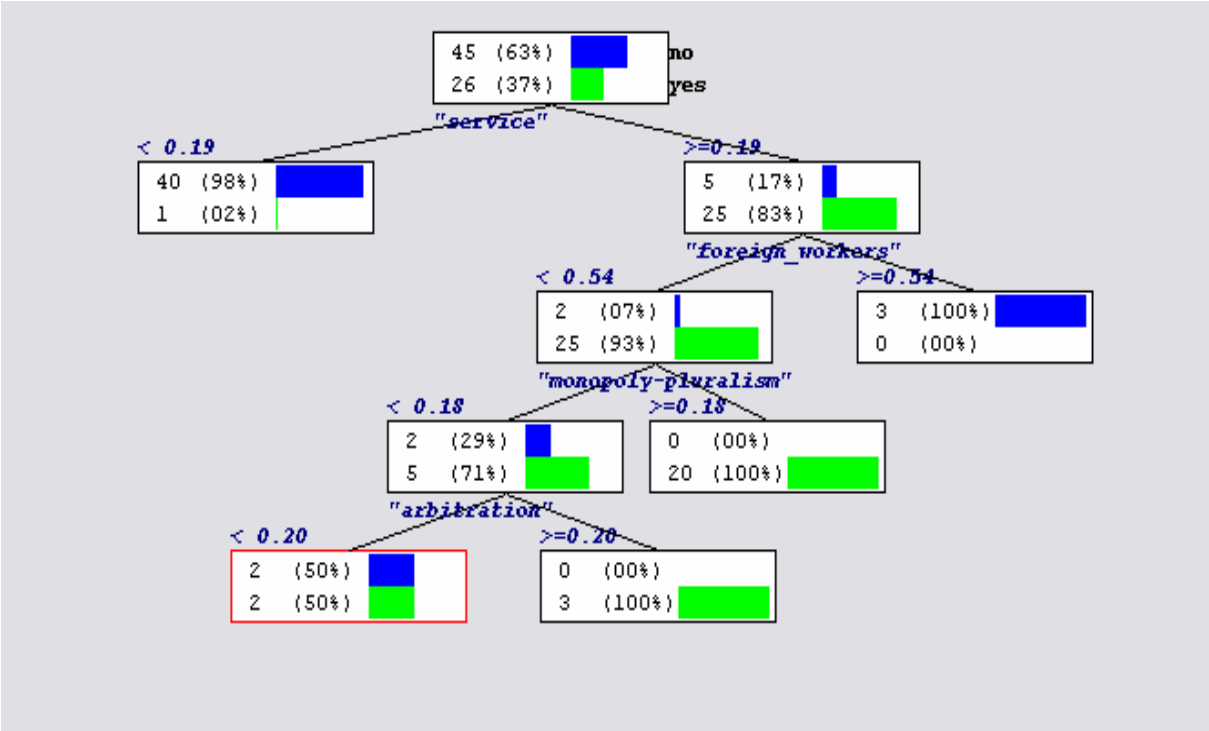


FIG.11 Violation "Minimum service"

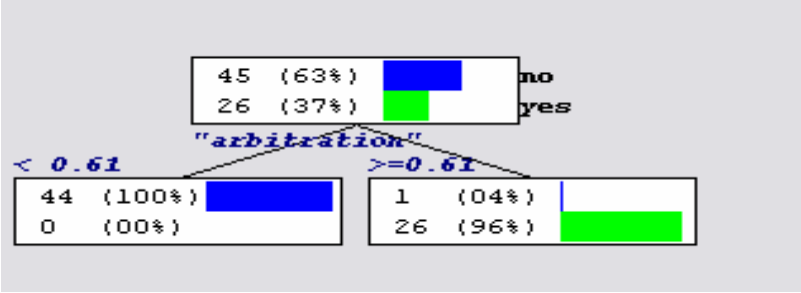


FIG.12 Violation "Compulsory arbitration in the context of industrial action"

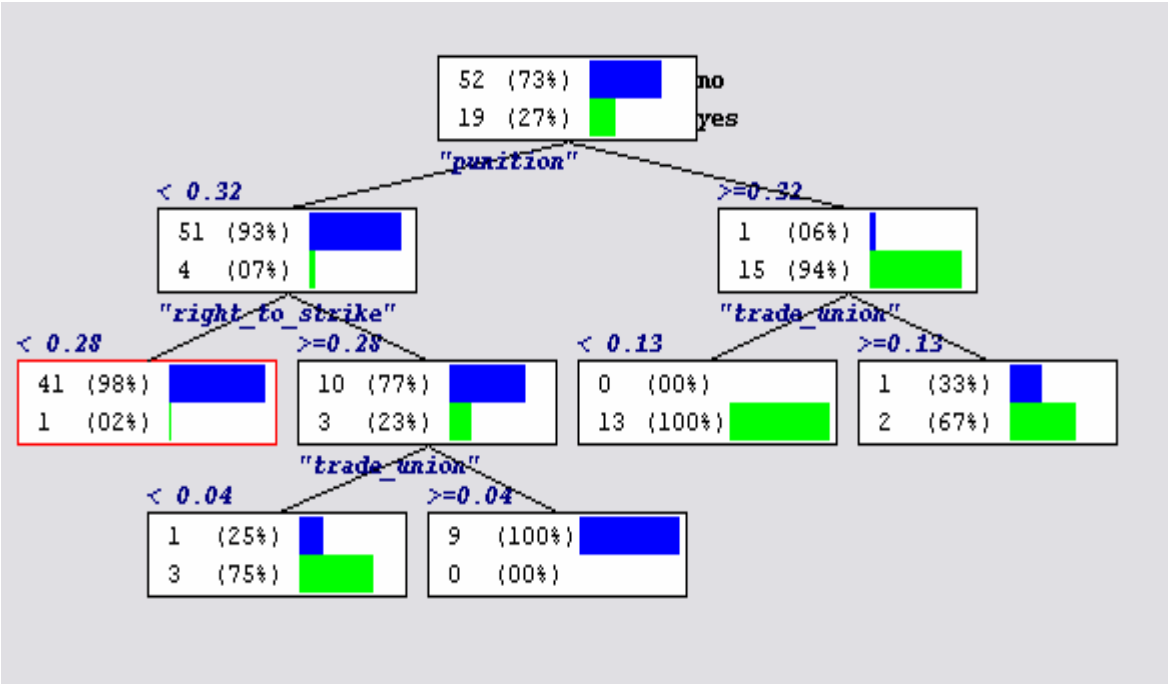


FIG.13 Violation "Penalties for instigation of, or participation in industrial action"

Thanks to decision trees, we're able to predict for a new text (unlabelled) what are its violations. Here is a list of predicted violations for countries:

Automatic Texts Annotating

countries	*candidate classes*	*probability*
Jamaica 1995	Right to establish and join workers' organisations	98%
	Election of representatives / Eligibility criteria	50%
	Restrictions on the right to industrial action	67%
	Conditions for lawful industrial action	87%
	Compulsory arbitration in the context of industrial action	96%
Japan 1995	Right to establish and join workers' organisations	98%
	Trade union pluralism	100%
	Election of representatives / Eligibility criteria	100%
	Organisation of activities	100%
	Restrictions on the right to industrial action	67%
	Conditions for lawful industrial action	87%
	Penalties for instigation of, or participation in, industrial action	100%
Hong-Kong 2004	Right to establish and join workers' organisations	98%
Ghana 1996	Right to establish and join workers' organisations	98%
Kuweit 1997	Right to establish and join workers' organisations	98%
Antigua and Barbuda 2001	Right to establish and join workers' organisations	98%
	Trade union pluralism	100%
	Organisation of activities	67%
	Restrictions on the right to industrial action	67%
	Conditions for lawful industrial action	87%
	Minimum service	100%
Djibouti 2003	Protection of property	67%
	Right to establish and join workers' organisations	98%
	Trade union pluralism	75%
	Election of representatives / Eligibility criteria	93%
	Organisation of activities	100%
	Restrictions on the right to industrial action	67%

Mauritania 1997	Right to establish and join workers' organisations	98%
	Election of representatives / Eligibility criteria	93%
	Organisation of activities	100%
	Restrictions on the right to industrial action	67%
	Minimum service	50%
	Compulsory arbitration in the context of industrial action	96%
Saint Lucia 1997	Right to establish and join workers' organisations	83%
Myanmar 2001	Right to establish and join workers' organisations	98%

As we said in the beggining of this section, we used two strategies for predicting the labels of unlabelled texts.The first one, described above, is decison trees. The second one is relative neighboorhood graph which represent the next paragraph.

2.2 Neighborhood graphs

Neighbourhood graphs are very much used in various systems. Their popularity is due to the fact that the neighbourhood is determined by coherent functions which reflect, in some point of view, the mechanism of the human intuition. Their use is varied from information retrieval systems to geographical information systems. Neighbourhood graphs, or proximity graphs, are geometrical structures which use the concept of neighbourhood to determine the closest points to a given point. For that, they are based on dissimilarity measures[6].

In a relative neighborhood graph $G_{rng}(\Omega, \varphi)$, two points (α, β) in Ω^2 are neighbors if they check the relative neighborhood property defined hereafter. Let $H(\alpha, \beta)$ be the hyper-sphere of radius $\delta(\alpha, \beta)$ and centred on α, and let $H(\beta,\alpha)$ be the hyper-sphere of radius $\delta(\beta,\alpha)$ and centred on β. $\delta(\alpha, \beta)$ and $\delta(\beta,\alpha)$ are the dissimilarity measures between the two points α and β. $\delta(\alpha, \beta) = \delta(\beta,\alpha)$. Then, α and β are neighbors if and only if the lune $A(\alpha, \beta)$ formed by the intersection of the two hyper-spheres $H(\alpha, \beta)$ and $H(\beta,\alpha)$ is empty [5]. Formally:

$$A(\alpha, \beta) = H(\alpha, \beta) \cap H(\beta,\alpha) \text{ Then } (\alpha, \beta) \text{ in } \in \varphi \text{ iff } A(\alpha, \beta) \cap \Omega = \phi$$

Figure 14 illustrates the relative neighborhood graph.

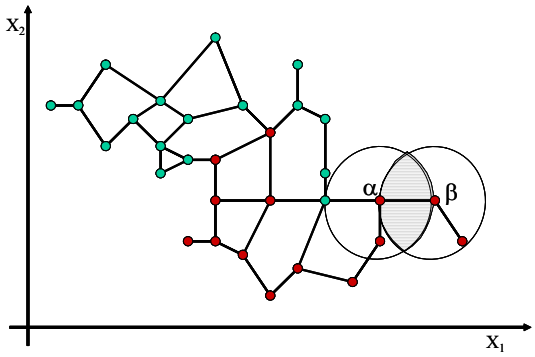


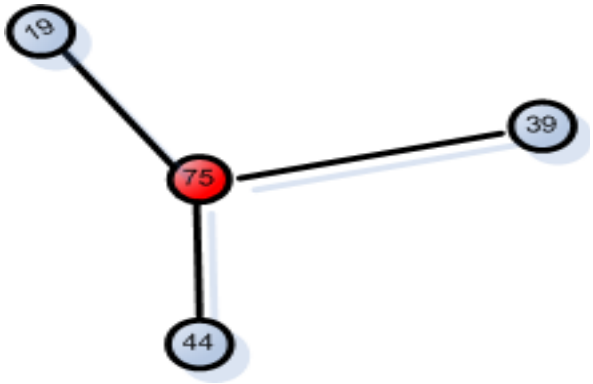
Fig. Relative neighborhood graph.

Application of neighborhood graph on the corpus.

In order to be able to set a label for unknown item, we build a basic graph using the labelled texts. Then, we took the unlabelled items, we insert each one sequentially and we apply a decision making function to label these items.
The decision making function is simple. Indeed, we just calculate the probability of the presence of a concept in the neighborhood of the inserted item.
Here after, illustrations and the obtained results using this approach.
In the figures presented hereafter, the red node in each sub-graph, represents the unlabelled item and the blue ones represent the labelled neighboors.

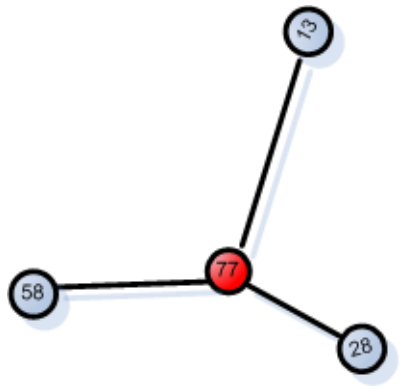
Automatic Texts Annotating

Queries	Candidate Classes	Probability
Jamaica 1995	compulsory arbitration in the context of industrial action	100.00%
	election of representatives / eligibility criteria	100.00%
	right to establish and join workers' organisations	100.00%
	conditions for lawful industrial action	66.67%
	minimum service	66.67%
	organisation of activities	66.67%
	trade union pluralism	66.67%
	restrictions on the right to industrial action	33.33%

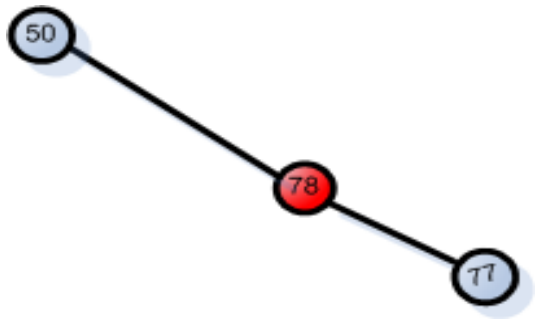


Japan 1995	conditions for lawful industrial action	100.00%
	election of representatives / eligibility criteria	100.00%
	establishment and registration of workers' organisations	100.00%
	organisation of activities	100.00%
	penalties for instigation of or participation in industrial action	100.00%
	restrictions on the right to industrial action	100.00%
	right to establish and join workers' organisations	100.00%

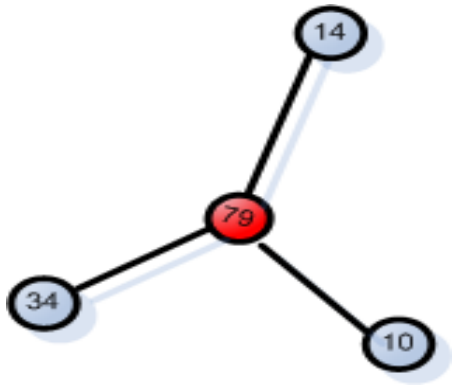
Hong-Kong 2004	right to establish and join workers' organisations	100.00%
	administrative/financial independence	66.67%
	restrictions on the right to industrial action	66.67%
	trade union pluralism	66.67%
	conditions for lawful industrial action	33.33%
	dissolution or suspension of workers' organisations	33.33%
	election of representatives / eligibility criteria	33.33%
	establishment and registration of workers' organisations	33.33%
	organisation of activities	33.33%
	penalties for instigation of or participation in industrial action	33.33%
	protection of property	33.33%



Ghana 1996	restrictions on the right to industrial action	100.00%
	right to establish and join workers' organisations	100.00%
	administrative/financial independence	50.00%
	conditions for lawful industrial action	50.00%
	election of representatives / eligibility criteria	50.00%
	establishment and registration of workers' organisations	50.00%
	trade union pluralism	50.00%

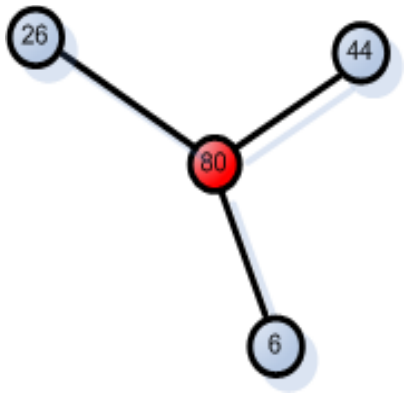


Kuweit 1997	right to establish and join workers' organisations	100.00%
	protection of property	66.67%
	trade union pluralism	66.67%
	election of representatives / eligibility criteria	33.33%
	organisation of activities	33.33%
	restrictions on the right to industrial action	33.33%

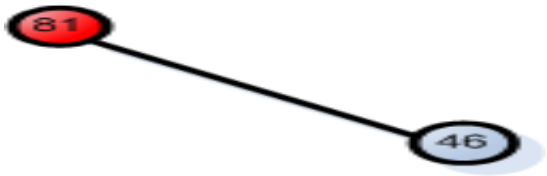


Antigua and Barbuda 2001	right to establish and join workers' organisations	100.00%
	conditions for lawful industrial action	66.67%
	minimum service	66.67%
	organisation of activities	66.67%
	restrictions on the right to industrial action	66.67%

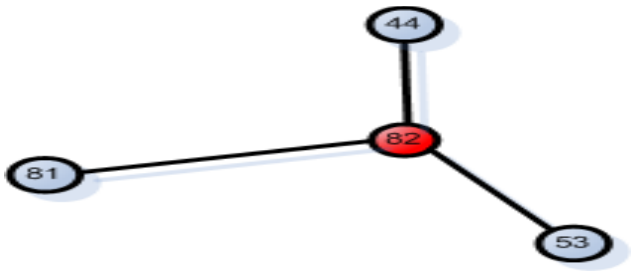
	trade union pluralism	66.67%
	administrative/financial independence	33.33%
	compulsory arbitration in the context of industrial action	33.33%
	dissolution or suspension of workers' organisations	33.33%
	election of representatives / eligibility criteria	33.33%
	penalties for instigation of or participation in industrial action	33.33%
	right to liberty and security of person / right to a fair trial	33.33%
	right to life and physical integrity	33.33%



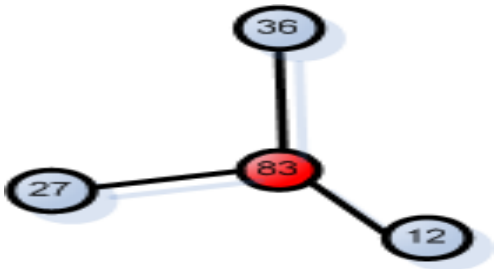
Djibouti 2003	compulsory arbitration in the context of industrial action	100.00%
	election of representatives / eligibility criteria	100.00%
	minimum service	100.00%
	organisation of activities	100.00%
	trade union pluralism	100.00%



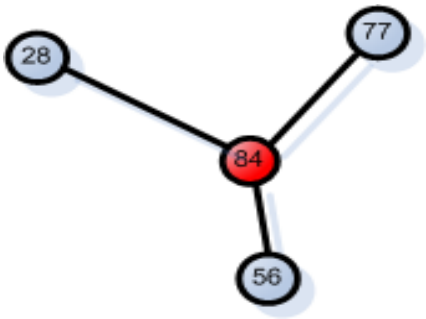
Mauritania 1997	election of representatives / eligibility criteria	100.00%
	organisation of activities	100.00%
	compulsory arbitration in the context of industrial action	66.67%
	conditions for lawful industrial action	66.67%
	minimum service	66.67%
	right to establish and join workers' organisations	66.67%
	trade union pluralism	66.67%
	establishment and registration of workers' organisations	33.33%
	penalties for instigation of or participation in industrial action	33.33%
	restrictions on the right to industrial action	33.33%



Saint Lucia 1997	right to establish and join workers' organisations	100.00%
	administrative/financial independence	66.67%
	dissolution or suspension of workers' organisations	33.33%
	minimum service	33.33%
	protection of property	33.33%
	restrictions on the right to industrial action	33.33%
	trade union pluralism	33.33%



Myanmar 2001	administrative/financial independence	100.00%
	restrictions on the right to industrial action	100.00%
	right to establish and join workers' organisations	100.00%
	trade union pluralism	66.67%
	conditions for lawful industrial action	33.33%
	dissolution or suspension of workers' organisations	33.33%
	election of representatives / eligibility criteria	33.33%
	establishment and registration of workers' organisations	33.33%
	organisation of activities	33.33%
	penalties for instigation of or participation in industrial action	33.33%



Legends:
Labelled texts

ID Text	Country and date
6	Russia 1996
10	Poland 2002
12	Poland 1998
13	Poland 1996
14	Poland 1995
19	Nicaragua 2000
26	Indonesia 2003
27	Indonesia 2002
28	Indonesia 2001
34	Greece 1998
36	Greece 1993
39	Egypt 2003
44	Egypt 1996
46	Egypt 1991
50	Costa Rica 1996
52	Costa Rica 1993
53	Costa Rica 1991
56	Bangladesh 2000
58	Bangladesh 1996

Unlabelled texts

ID Text	Country and Date
75	Jamaica 1995
76	Japan 1995
77	Hong-Kong 2004
78	Ghana 1996
79	Kuweit 1997
80	Antigua and Barbuda 2001
81	Djibouti 2003
82	Mauritania 1997
83	Saint Lucia 1997
84	Myanmar 2001

Conclusion and future work

The finality of this work is to propose a predictive model able to set for unlabelled items their corresponding labels according to a set of labelled texts. An automatic learning approach was adopted. In the first step (data preparation) we processed a set of 71 texts in rder to extract potentially important concepts. This is done using BRILL and EXIT. The second step (learning step) we used two predictive models, decision trees and neighborhood graphs, to label the texts.

The results seems to be interesting, the two models give, approximatly, the same conclusions. From the automatic learning point of view, these results are meaningfull. However, our approach as well as we obtained results have to be validated by experts.

As future work, we plane, after the validation of the proposed approach, to use much items in the learning step (processing more significant items). Also, the generalization of the proposed models (issued from decision trees and/or neighbohood graphs), is a logical continuation.

References

1. E. Brill (1995). “Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging”. Computational Linguistics, 2(4):543-565.

2. T. Heitz, Y. Kodratoff, M.Roche « EXIT: Un système itératif pour l'extraction de la terminologie du domaine à partir de corpus spécialisés ». Dans Proceedings of JADT'04 (International Conference on Statistical Analysis of Textual Data), volume 2, pages 946-956.

3. Brunet-Manquat F. « Description et conception d’une plate-forme robuste combinant des analyseurs d’énoncés », journal en ligne ISDM (informations, savoirs, décisions et médiations), vol13, 12 pages, février 2004.

4. D. A. Zighed et R.Rakotomalala « Graph d’induction et data mining », Hermès, 2000

5. G. T. Toussaint. « The relative neighborhood graphs in a finite planar set ». Pattern recognition, 12:261–268, 1980.

6. G. T. Toussaint. « Some insolved problems on proximity graphs”. D. W Dearholt and F. Harrary, editors, proceeding of the first workshop on proximity graphs. Memoranda in computer and cognitive science MCCS-91-224. Computing research laboratory. New Mexico state university Las Cruces, 1991.