Web Page Structured Content Detection Using Supervised Machine Learning

Roberto Panerai Velloso, Carina F. Dorneles
{rvelloso, dorneles}@gmail.com

UNIVERSIDADE FEDERAL DE SANTA CATARINA

Table of Contents

Context

- What kind of data are we considering? (loosely structured web data, e.g., e-commerce search result records);
- Why are we considering it? (volume, diversity, usefulness);
- We can take advantage of the structure and leverage it;
- There is about 15+ years of research on the subject;
- And it is still an open problem (it is a hard one, indeed). Why? Mainly due of its diversity and also because it is made available for human consumption, not for machines.

Content vs Noise

When we consider ONLY the structured data from web pages:

- What is content in this context?
- What is noise in this context?
- What is the difference between them? (both have structure)
  - CONTENT: is the data for which a web page as built for, its reason for existing (e.g., search result records);
  - NOISE: all the rest (e.g., template menus, footer data, etc. - they have structure, but they are not content).

Classification

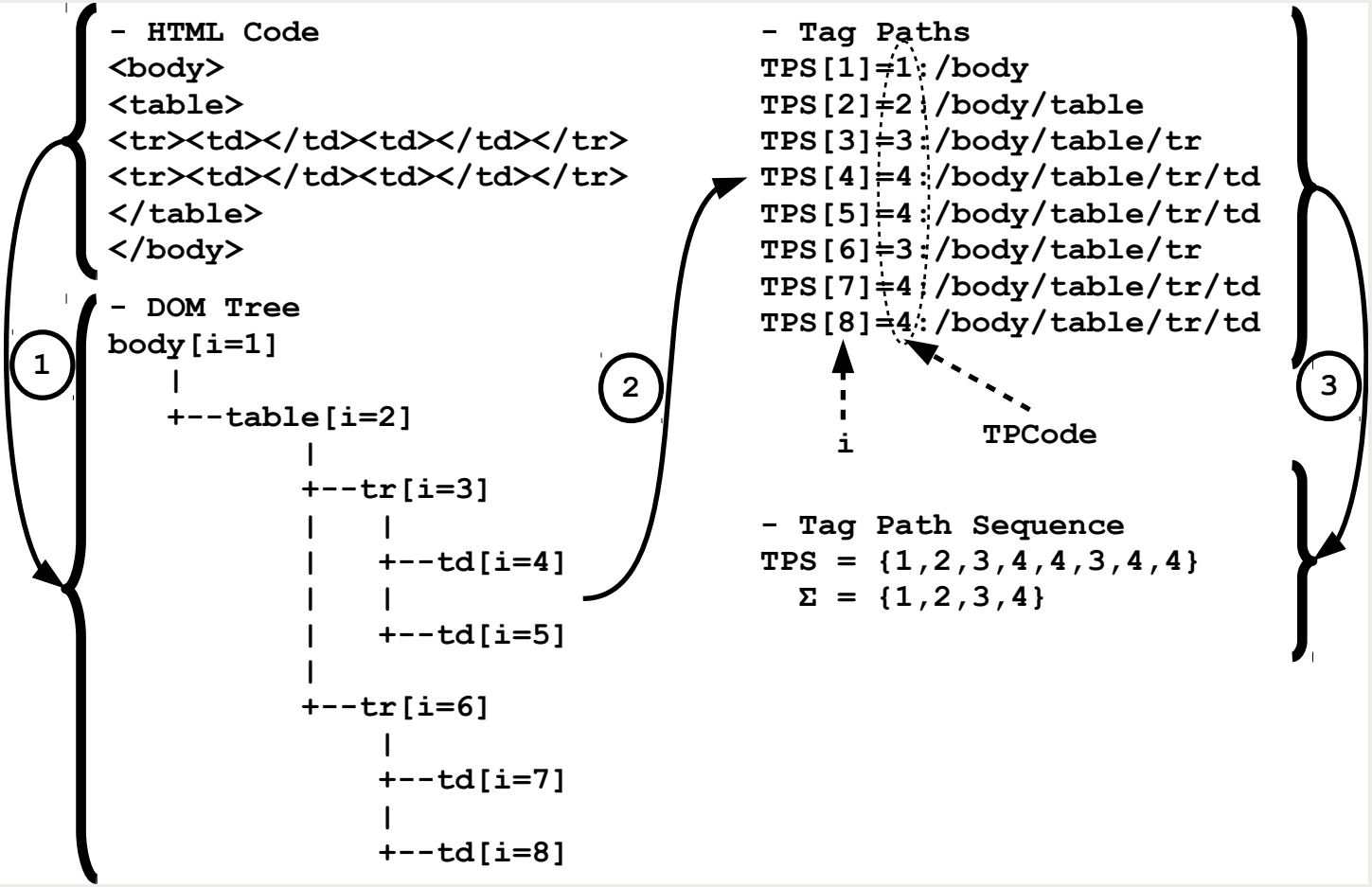
- We have tackled it as a classification problem;
- We have to characterize the problem (what features are important? Where do they come from?);
- We can apply supervised ML to labeled problem instances;

Web Page Sequence Representation

Instead of using DOM tree to represent the web document, we use an alternative sequence representation from which we extract the structured data and its features as well.

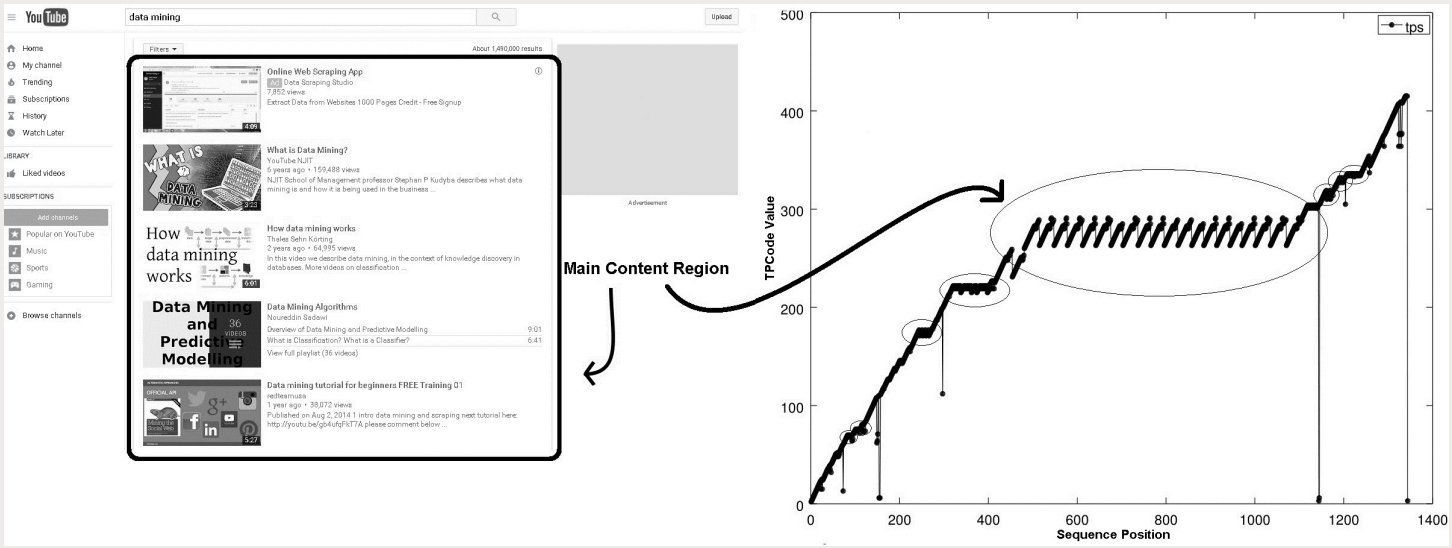
DOM Tree to Sequence

Figure: Tag Path Sequence construction.



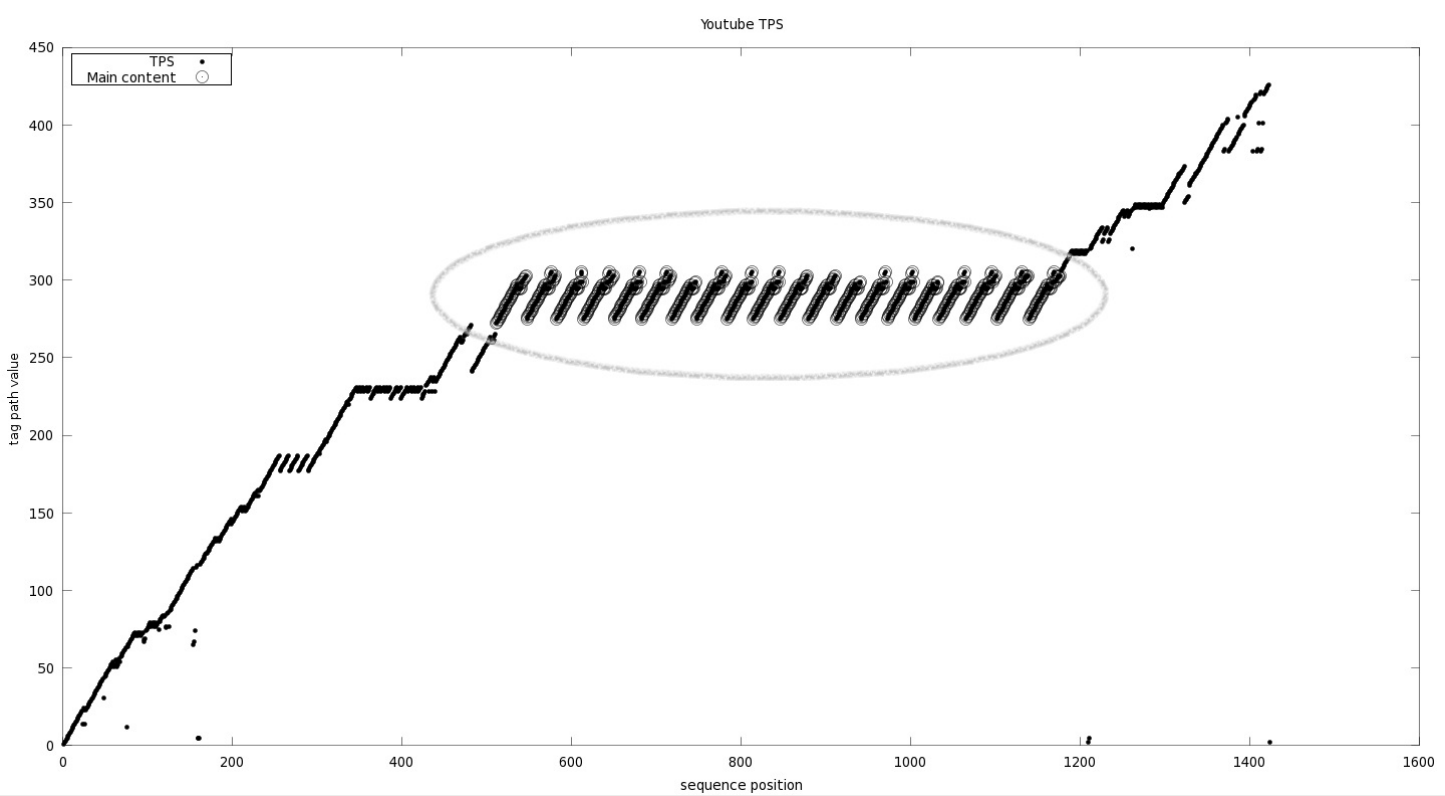
DOM Tree to Sequence

Figure: Tag Path Sequence document representation.

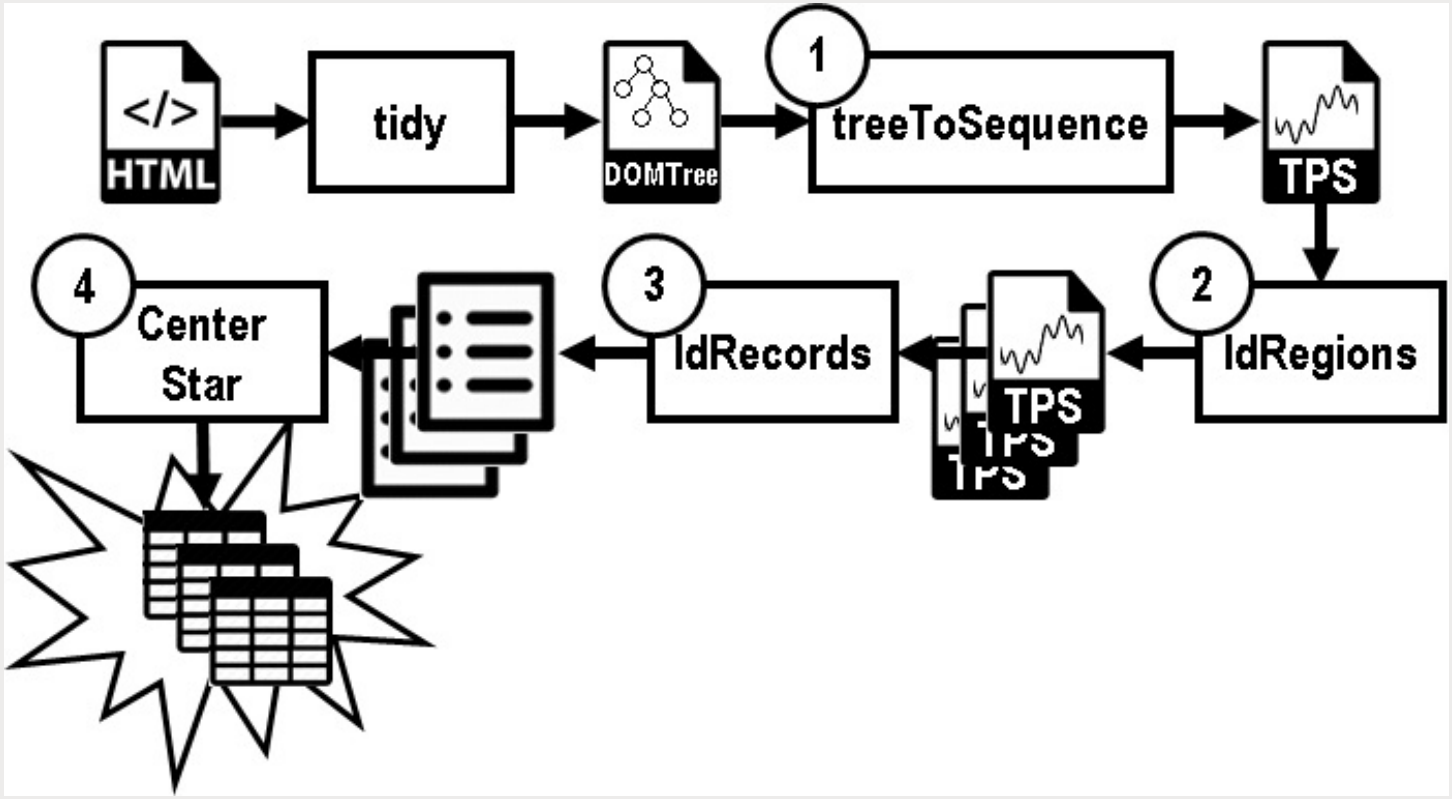


DOM Tree to Sequence

Figure: Tag Path Sequence document representation.



DOM Tree to Sequence

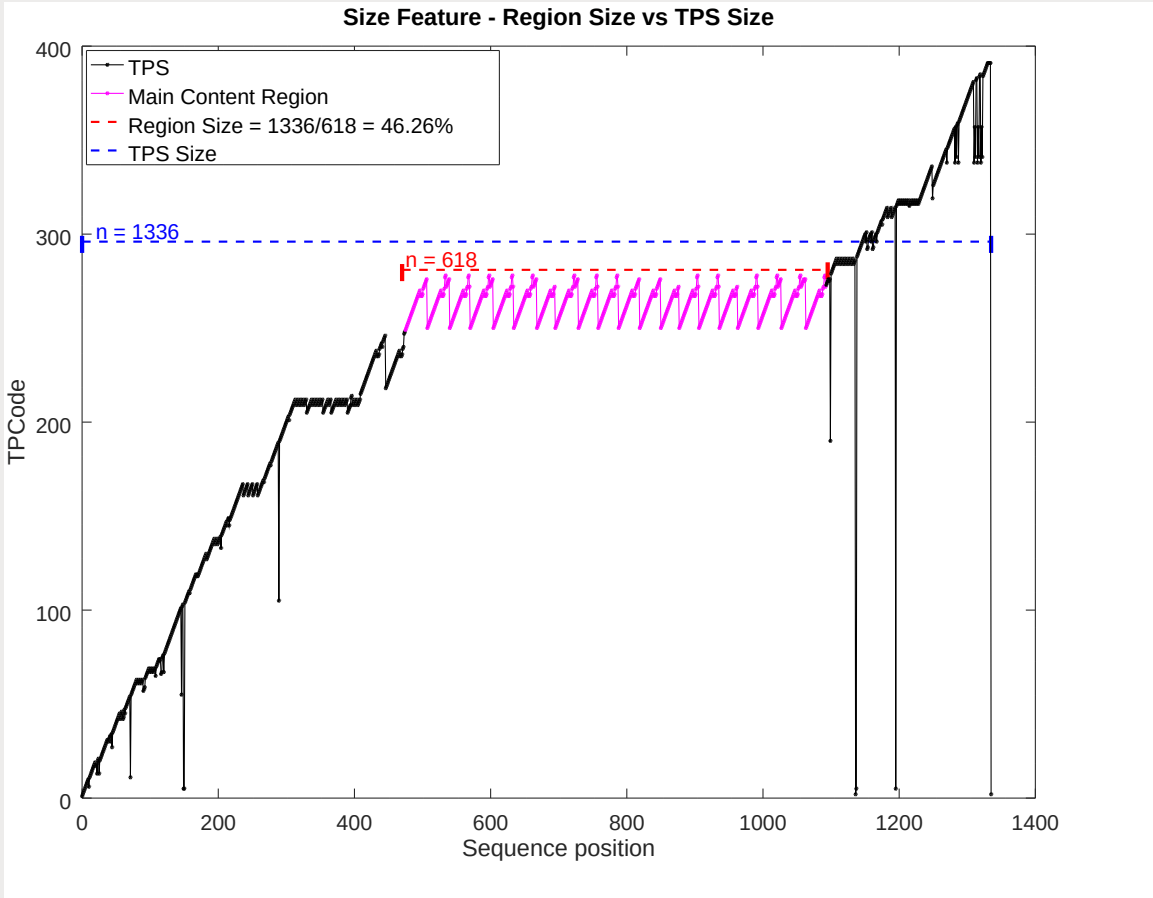


Features

1. **Size Feature**: data region's size (relative to whole sequence);
2. **Center Position Feature**: data region's center position (distance to the center of the sequence);
3. **Horizontal Position Feature**: data region's horizontal position (distance to the end of the sequence);
4. **Vertical Position Feature**: data region's vertical position (distance to the top of the sequence);
5. **Range Feature**: data region's range (relative to the whole sequence);
6. **Record Feature**: ratio between data region's record count and record size;

Size Feature

Figure: Size feature.



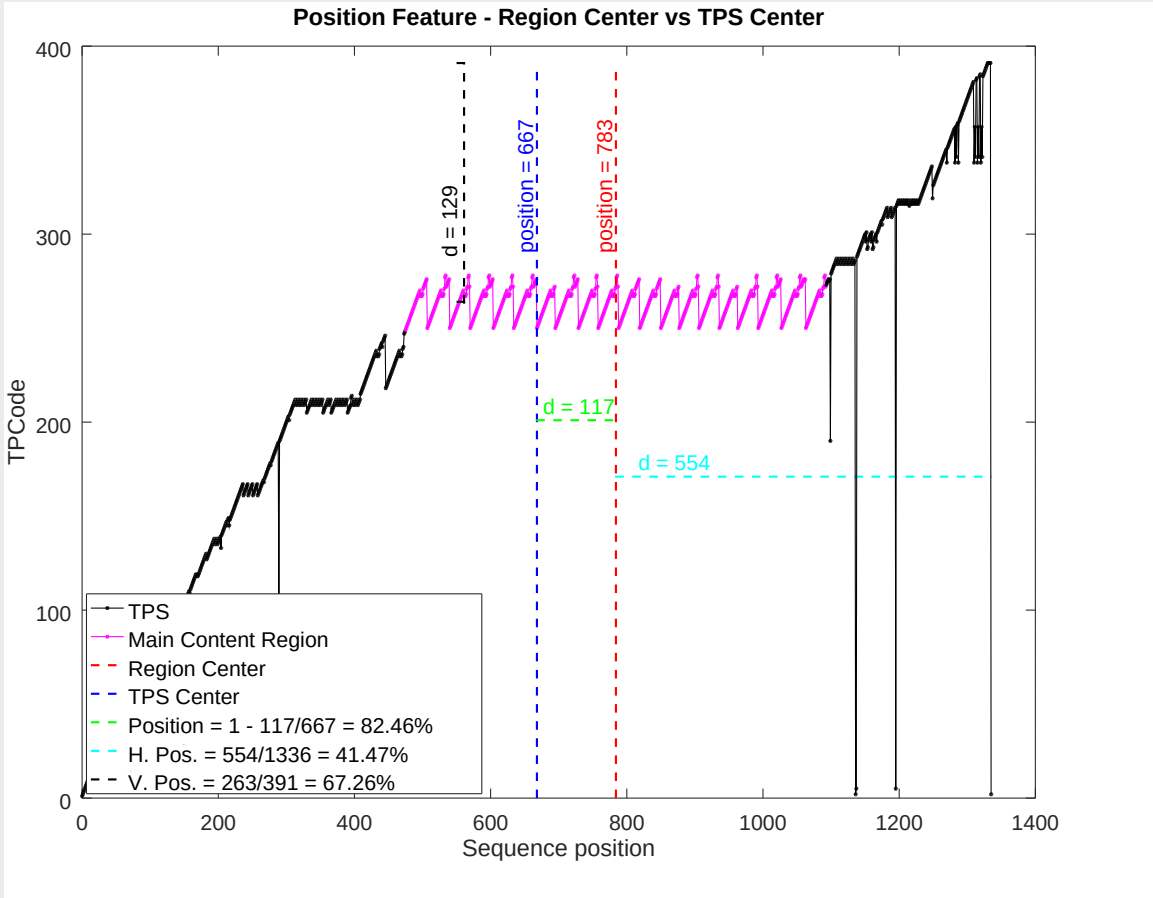
Position Features

- Center Position (how close to the center?);
- Horizontal Position (how close to the end?);
- Vertical Position (how close to the top?).

Vertical and Horizontal features were added to deal with documents without structured content.

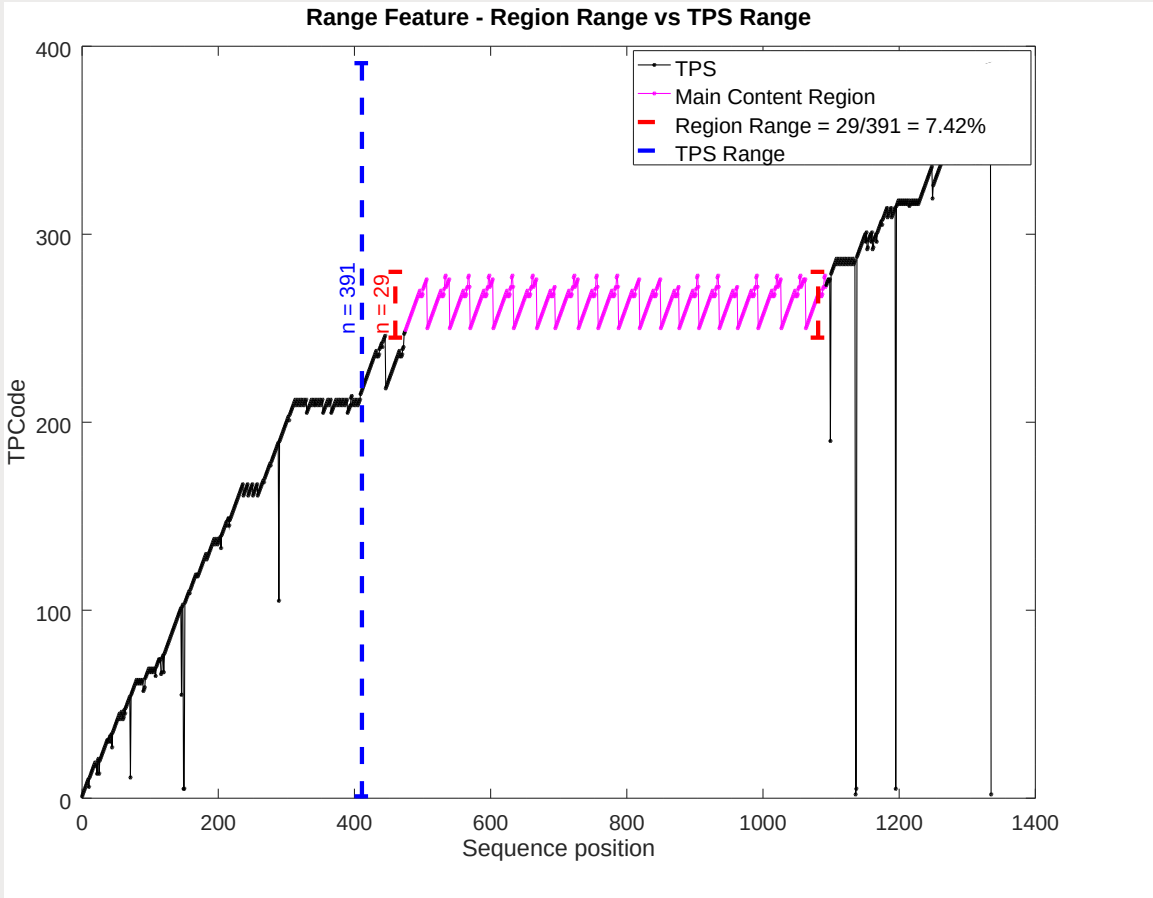
Position Features

Figure: Position features.



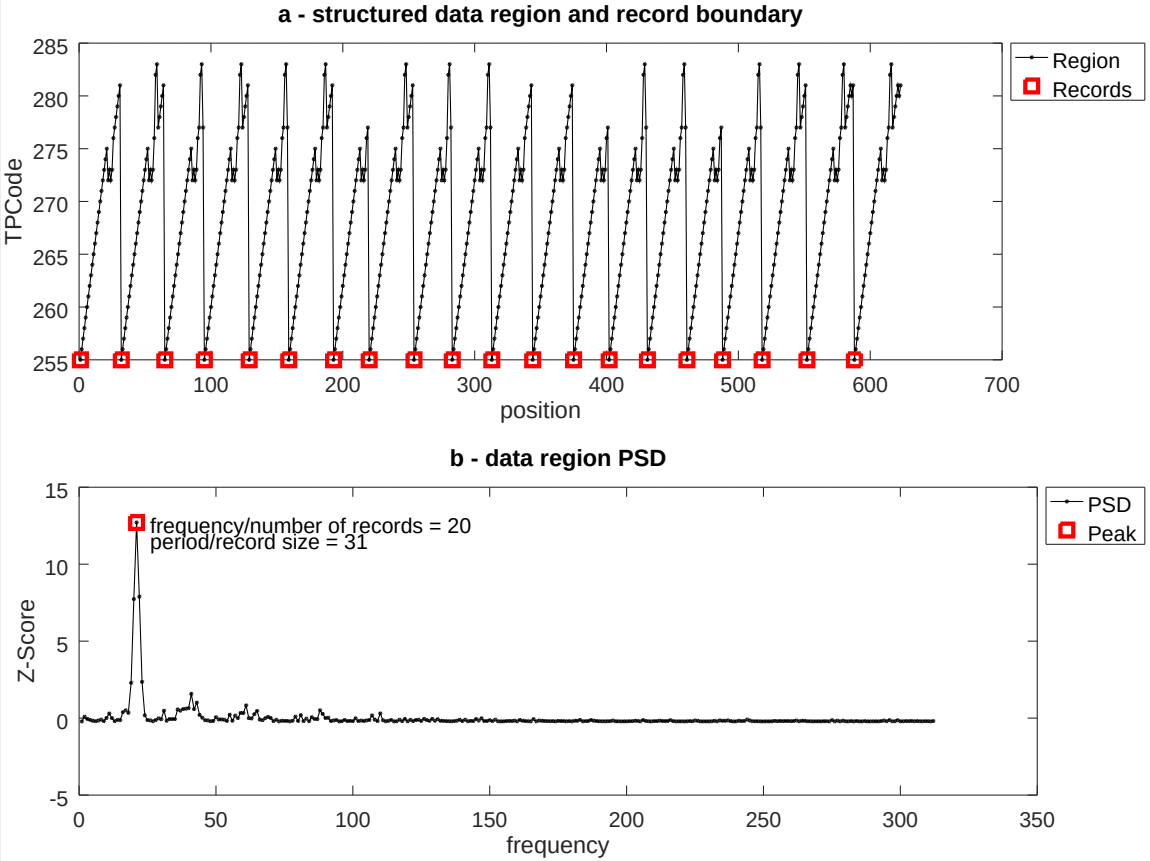
Range Feature

Figure: Range feature.



Record Feature

Figure: Record feature.



Evaluation Scenarios

1. Documents guarantee to contain at least one structured region;
2. Above documents and some more documents without structured content (i.e., textual content).

Table: Input dataset summary

# Content Regions	254	47.65%
# Noise Regions	279	52.35%
Total	533	100%
# Structured documents	266	81.35%
# Unstructured documents	61	18.65%
Total	327	100%

Correlation

Table: Feature correlation

	Size	Record	Range	Horiz.	Vert.
Content & Noise					
Center	**0.63**	0.08	0.45	-0.11	0.01
Size		0.08	**0.68**	-0.02	0.11
Record			0.08	0.04	0.03
Range				-0.01	0.08
Horizontal					**0.85**
Content					
Center	0.58	-0.11	0.21	-0.37	-0.07
Size		-0.10	0.53	-0.12	0.12
Record			-0.04	0.04	-0.04
Range				-0.03	0.08
Horizontal					0.65
Noise					
Center	0.25	-0.01	0.22	-0.15	-0.11
Size		-0.12	0.39	-0.15	-0.11
Record			-0.08	0.01	0.01
Range				-0.13	-0.09
Horizontal					0.89

Statistics

Table: Feature statistics

Feature	Mean	CV	Skewness	Kurtosis
Content & Noise				
Size	0.27	0.87	0.92	2.90
Center	0.57	0.51	-0.18	1.71
Range	0.11	1.12	1.98	7.46
Record	0.40	0.68	0.52	2.15
Vertical	0.59	0.40	-0.49	2.42
Horizontal	0.55	0.47	-0.26	2.14
Content				
Size	0.44	0.49	0.28	2.47
Center	0.74	0.27	-0.81	2.95
Range	0.19	0.74	1.49	5.51
Record	0.46	0.61	0.22	1.89
Vertical	0.63	0.24	-0.80	3.58
Horizontal	0.57	0.26	-0.45	3.63
Noise				
Size	0.12	0.94	2.32	9.37
Center	0.41	0.65	0.58	2.16
Range	0.05	1.40	4.37	26.83
Record	0.34	0.74	0.81	2.74
Vertical	0.56	0.53	-0.15	1.69
Horizontal	0.53	0.61	-0.06	1.44

Feature Importance

Table: Feature importance (vs class)

feature	χ^2	ANOVA
Size	51.6426225	487.5011631
Center	25.8025951	260.9367942
Range	23.1719961	232.4460871
Record	4.7116862	26.5957295
Vertical	1.1694271	12.3825010
Horizontal	0.4337931	3.6350506

Results

- Results obtained on both evaluation scenarios (precision, recall, accuracy and f-score);
- Result comparison with other approaches from the literature.

Only documents with structured content.

Table: Results using dataset containing only structured content documents.

Model	Precision	Recall	Accuracy	F-Score
LR	93.30%	**93.85%**	**93.02%**	**93.57%**
GNB	91.97%	90.55%	90.62%	91.26%
kNN	**93.83%**	92.21%	92.59%	93.01%
SVM	93.60%	92.20%	92.37%	92.90%
EXT	91.88%	91.87%	91.23%	91.88%
GB	90.75%	90.14%	89.74%	90.44%
VOT	92.41%	92.20%	91.71%	92.31%
STCK	92.97%	92.20%	92.06%	92.59%

Full dataset

Table: Results using complete dataset (including unstructured documents).

Model	Precision	Recall	Accuracy	F-score	Drop
LR	87.97%	92.13%	89.45%	90.00%	**-3.57%**
GNB	89.69%	88.99%	89.47%	89.34%	-1.92%
kNN	89.72%	90.56%	90.02%	90.14%	-2.87%
SVM	89.51%	92.12%	90.77%	90.79%	-2.11%
EXT	88.80%	88.43%	88.71%	88.61%	**-3.27%**
GB	90.13%	88.97%	89.65%	89.55%	**-0.89%**
VOT	**90.45%**	**92.52%**	**91.33%**	**91.47%**	-1.38%
STCK	89.93%	91.81%	90.73%	90.86%	-1.73%

Results Comparison

Table: Comparison w/other approaches

Algorit.	Prec.	Recall	F-Score	Acc.
BERyL	n/d	n/d	90.00%	n/d
SIG	92.02%	94.11%	93.05%	n/d
TPC	90.40%	93.10%	91.73%	n/d
MDR	59.80%	61.80%	60.78%	n/d
Our models				
LR	**95.45%**	95.45%	**95.45%**	94.80%
VOT	**93.02%**	90.91%	**91.95%**	90.91%

Conclusion

- We have achieved good results (F-Score);
- We have shown that using supervised ML with only very basic features is feasible for this problem;
- Our proposal's performance is comparable to the state-of-the-art;
- We need a larger dataset to improve confidence in the results. Specially to better evaluate performance degradation in the presence of unstructured content.