PolicyMiner: From Oysters to Pearls

Hossein Rahmani[1] and Christine Arnold [1]

[1]Faculty of Arts and Social Sciences, Maastricht University
{hossein.rahmani,c.arnold} @ maastrichtuniversity.nl

Abstract. Today, a historically unprecedented volume of data is available in the public domain with the potential of becoming useful for researchers. More than at any other time before, political parties and governments are making data available such as speeches, legislative bills and acts. However, as the size of available data increases, the need for sophisticated tools for web-harvesting and data analysis simultaneously grows. Yet, for the most part researchers who are developing these tools come from a computer science background, while researchers in the social and behavior sciences who have an interest in using such tools often lack the necessary training to apply these tools themselves.

In order to provide a bridge between these two communities we propose a new tool called PolicyMiner. The objective of this tool is twofold: First, to provide a general purpose web-harvesting and data clean-up tool which can be used with relative ease by researchers with limited technical backgrounds. The second objective is to implement knowledge discovery algorithms that can be applied to textual data, such as legislative acts. With our paper we present a technical document which details the steps of data processing that have been implemented in the PolicyMiner. First, the PolicyMiner harvests the raw html data from publically available websites, such as governmental sites, and provides a unique integrated view for the data. Second, it cleans the data by removing irrelevant items, such as html tags and non-informative terms. Third, it classifies the harvested data according to a pre-defined standard conceptual hierarchy relying on the Eurovoc thesaurus. Fourth, it applies different knowledge discovery algorithms such as time series and correlation-based analysis to capture the temporal and substantive policy dependencies of the textual data across countries.

1 Introduction

As a quote from John Naisbitt: “We are drowning in information but starved for knowledge” indicates, there is an inherent tension between large-scale data and our ability to process it and to discover meaningful patterns. The amount of data available from different aspects of life increase every second and the task to mine data and extract useful knowledge becomes more and more challenging. The main goal of the knowledge discovery process is to extract informative knowledge from a large amount of data and to represent it in a human understandable structure. We can understand the knowledge discovery process as a system which takes a certain type of data as input and produces informative knowledge as output. Figure 1 shows three main subprocesses of the entire system. The first subprocess is called "Data Pre-Processing" which takes raw input data and outputs the cleaned version of the data [20, 16]. "Data Cleaning", "Data Integration", "Data Transformation" and "Data Reduction" are the most common used methods in this subprocess [20]. The second subprocess is called "Machine Learning" and its main task is to extract potential informative patterns from the cleaned data. Supervised learning [15] and Unsuprvised learning [9] are two main approaches of Machine Learnining. Supervised learning analyzes the training data and produces an inferred function, which can be used for mapping new examples. The main aim of Unsupervised learning is finding hidden structure in unlabeled data. The last subprocess is called "Data Post-Processing"; it validates and evaluates the extracted patterns [4].

In this paper, we propose a novel knowledge discovery framework called PolicyMiner for harvesting and analyzing the legislative documents. The work discussed in this paper have been developed as part of a larger project, The Policy Votes project [2], which is funded by the Dutch Science Foundation, with Christine Arnold, Mark Franklin and Christopher Wlezien as the Co-PIs. The purpose of this paper s to report the steps followed in the project and to function as a technical documentation for replication purposes. Following the sub-processes shown in Figure 1, our PolicyMiner has two main modules. First, Crawler module which is in charge of harvesting, cleaning and persisting the legislative documents. Second, Analyzer module which its main task is to extract informative knowledge from the cleaned version of data. We discuss these modules in details in the following sections.

2 Crawler Module

In this section, we discuss the "Crawler" module of our PolicyMiner which harvests legislative documents from the governments' websites, cleans the downloaded documents and persists the cleaned documents in the DataBase Management System (DBMS). We used Python programming language to develop this module. Preferring Python over other programming language was due to Pythons strength in rapid prototyping, web scripting, text manipulation and xml processing. Our crawler is capable of harvesting different document types from websites with varied structures. So, flexibility is among the main characteristics of the crawler module. This allows us to easily modify the crawler in order to make it compatible with a range of websites. At the same time this flexibility allows us to easily update the crawler in case the structure of any of the websites we are currently web-harvesting would change. We designed and implemented a general work-flow for harvesting, cleaning and persisting the documents into the database using the Template Pattern [7]. In Template Pattern, we define the skeleton of a process in a method, called template method, which defers some steps to variable points that are defined as external xml configuration files in our

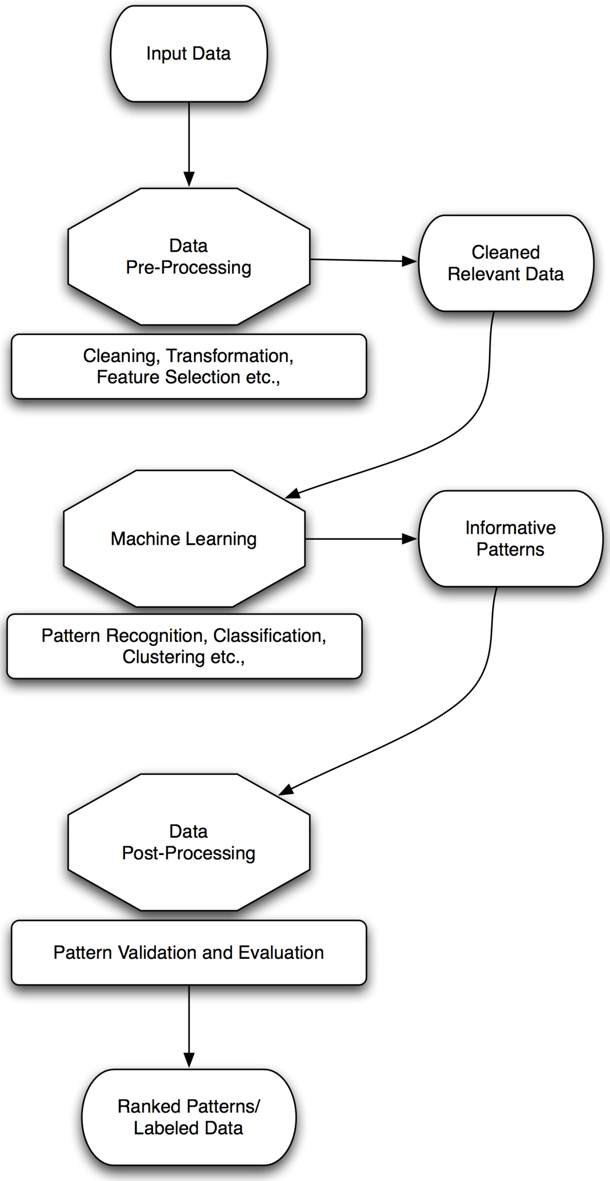

Input Data
Data Pre-Processing
Cleaned Relevant Data
Cleaning, Transformation, Feature Selection etc.,
Machine Learning
Informative Patterns
Pattern Recognition, Classification, Clustering etc.,
Data Post-Processing
Pattern Validation and Evaluation
Ranked Patterns/ Labeled Data

Fig. 1. Three main subprocesses of the knowledge discovery process. The first subprocess is called "Data Pre-Processing" which takes raw input data and output the cleaned version of the data. The second subprocess is called "Machine Learning" and its main task is to extract potential informative patterns from the cleaned data. The last subprocess is called "Data Post-Processing" and validates and evaluates the extracted patterns.

PolicyMiner. So, we could re-define certain steps of an process without changing the structure of the process.

Now, we briefly introduce some basic concepts about the *Input Data*. Figure 2 shows the Entity Relationship Diagram (ER) of the input data. ER Models [5] are useful in describing the dataset in an abstract way. We discuss each element in Figure 2 as follows:

Case 1. Country: Currently, we have implemented our web-harvesting tools for the government websites of 15 European countries. We fetch and analyze the legislative documents of the following countries: Austria, Belgium, Denmark, Germany, Ireland, United Kingdom, France, Greece, Portugal, Italy, Latvia, Spain, Sweden, Luxemburg, Finland and Netherlands.

Case 2. Category: For classifying the documents into policy areas we make use of a multilingual concept hierarchy called Eurovoc [6]. Eurovoc exists in 22 official languages of the European Union (Bulgarian, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Greek, Italian, Hungarian, Latvian, Lithuanian, Maltese, Polish, Portuguese, Romanian, Slovak, Slovene, Spanish and Swedish), as well as Basque, Catalan, Croatian, Russian and Serbian. It has the hierarchical structure as 21 categories in its first layer, 127 categories in the second layer and 7132 categories in the third layer. Table 1 lists high-level categories of Eurovoc concept hierarchy.

Case 3. Document: Legislative documents are the main elements of analysis in this paper. For each document, we know its country.

Case 4. Cat-Doc: We predict at least 6 predefined Eurovoc categories for each document in our DBMS. Cat-Doc elements store the information about the document's concept labels.

2.1 Formal Representation of Input Data

In this section, we describe the formal representation of our input data. If C is the set of all the considerd countries then, for each country $c_i \in C$, we harvest all the related legislative documents $D = \{d_1 \dots d_n\}$. We group the documents according to their published year and the PolicyMiner then assigns to each document $d_i \in D$ a set of weighted policies from the Eurovoc categories. $weight(d_i, p_j)$ means how much document d_i is relevant to policy p_j. Eurovoc policies are in hierarchical order, for each policy p_i, we show its sub categories with $subcats(p_i)$.

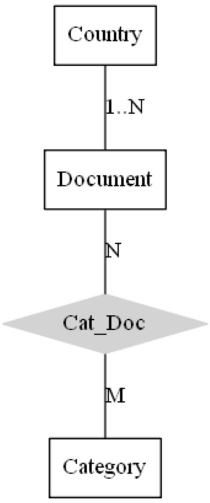


Fig. 2. Entity Relationship Diagram (ERD) of the input data.

Table 1. High-level categories of Eurovoc concept hierarchy.

Index	Category
1	LAW
2	BUSINESS AND COMPETITION
3	PRODUCTION, TECHNOLOGY AND RESEARCH
4	SOCIAL QUESTIONS
5	AGRI-FOODSTUFFS
6	TRADE
7	INDUSTRY
8	EDUCATION AND COMMUNICATIONS
9	INTERNATIONAL RELATIONS
10	FINANCE
11	TRANSPORT
12	POLITICS
13	GEOGRAPHY
14	ENERGY
15	EMPLOYMENT AND WORKING CONDITIONS
16	AGRICULTURE, FORESTRY AND FISHERIES
17	INTERNATIONAL ORGANISATIONS
18	SCIENCE
19	ENVIRONMENT
20	EUROPEAN COMMUNITIES
21	ECONOMICS

3 Analyzer Module

The second module of our PolicyMiner is the Analyzer module and mainly has the task of extracting informative knowledge from the raw persisted data. We choose Java programming language for developing this module because of the availability of a huge amount of open source data analyzer tools and implemented algorithms in this programming language. Similar to the Crawler module, we apply the Template pattern to the Analyzer module too. For each analyzing step, we define a variable point which can be fullfill with different available tools. For example, the main objects in the area of political science are pre-defined policies called Eurovoc categories which are available in different languages. So, our first task in the data analysis step is to classify each document into those pre-defined policies. We define a variable point for a classifier which gets a legislative document as input and generates related eurovoc categories according to the document's content. At the moment, our PolicyMiner uses off-the-shelf tool called JEX [19] to assign 6 weighted policies to each document in our Database. JEX is multi-label classification software that learns from manually labelled data to automatically assign EuroVoc categories to new documents. We beleive that the main advantage of JEX is in its learned model which is built based on the manually classified documents. In future, in the case of better classifier, we could simply switch to the new classifier without any change for the external users of PolicyMiner.

After classifiying the documents to the predefined categories, for each pair (year=y_i, policy=p_j), we could simply sum the weights of all the documents labelled with p_j and assume the result as the policy representor of y_i. Considering the period of time, we could model each policy p_j as a time series data. From now on, we decide to consider just time-series model of policies for our data analysis. Decreasing the size of data in addition to more informative nature of the data are the main advantages of this decision.

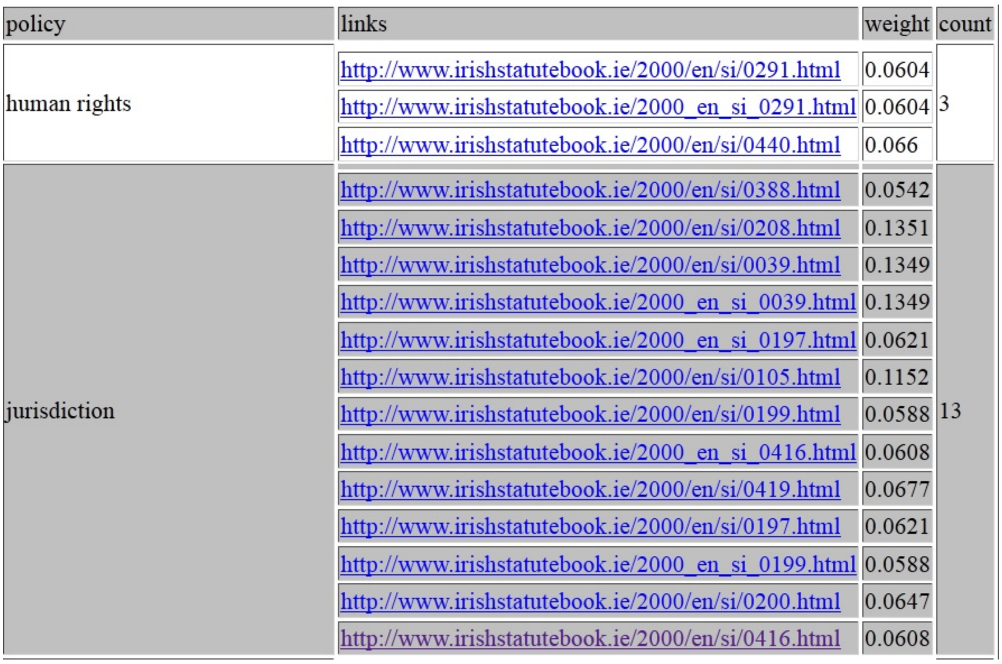
Having transformed the raw legislative document into time-series model of data, now, we could apply several data analysis algorithms including Search Analysis, Trend Analysis, Attention Allocation Analysis, Entropy Analysis, Correlation Analysis, Causality Analysis and Clustering Analysis that are defined and implemented in previous methods [11, 3, 1, 18, 14]. For each of these analysing algorithms, we define a variable point in our Template pattern. In the following subsections, first, we define each analysis method. Second, we apply the analysis method to our dataset. Third, we briefly interpret the results.

3.1 Search Analysis

Retrieving the relevant documents is the first step of any data analysis process. Our PolicyMiner provides easy document retrieval through the three main parameters: $1 - Country$, $2 - Year$ and $3 - Policy$.

Considering the hierarchical order of Eurovoc categories, Figure 3 shows part of the search result for parameter values $Country = Ireland$, $Year = 2000$ and

$Policy = law$. For each policy $p_j \in subcats(law)$, PolicyMiner lists the relevant documents $d_1 ... d_n$ in addition to $weight(d_i, p_j)$ for each document d_i.



policy	links	weight	count
human rights	http://www.irishstatutebook.ie/2000/en/si/0291.html	0.0604	3
	http://www.irishstatutebook.ie/2000_en_si_0291.html	0.0604	
	http://www.irishstatutebook.ie/2000/en/si/0440.html	0.066	
jurisdiction	http://www.irishstatutebook.ie/2000/en/si/0388.html	0.0542	13
	http://www.irishstatutebook.ie/2000/en/si/0208.html	0.1351	
	http://www.irishstatutebook.ie/2000/en/si/0039.html	0.1349	
	http://www.irishstatutebook.ie/2000_en_si_0039.html	0.1349	
	http://www.irishstatutebook.ie/2000_en_si_0197.html	0.0621	
	http://www.irishstatutebook.ie/2000/en/si/0105.html	0.1152	
	http://www.irishstatutebook.ie/2000/en/si/0199.html	0.0588	
	http://www.irishstatutebook.ie/2000_en_si_0416.html	0.0608	
	http://www.irishstatutebook.ie/2000/en/si/0419.html	0.0677	
	http://www.irishstatutebook.ie/2000/en/si/0197.html	0.0621	
	http://www.irishstatutebook.ie/2000_en_si_0199.html	0.0588	
	http://www.irishstatutebook.ie/2000/en/si/0200.html	0.0647	
	http://www.irishstatutebook.ie/2000/en/si/0416.html	0.0608	

Fig. 3. Small part of the search result for parameter values $Country = Ireland$, $Year = 2000$ and $Policy = law$. For each policy $p_j \in subcats(law)$, PolicyMiner lists the relevant documents $d_1 ... d_n$ in addition to $weight(d_i, p_j)$ for each document d_i.

3.2 Trend Analysis and Attention Allocation Analysis

One of the most direct way of individual analysis is to draw the total weight of each policy p through the time period $[t_1 ... t_2]$ in a scatter diagram in which the X axis shows time t_i and the Y axis is Total Weight p ($TW(p)$) calculated using Formula 1. In Formula 1, $D(c_j, y_k)$ lists all the documents in year y_k related to country c_j. The resulting diagram is called "Trend" diagram of p and shows how the government attention to specific policy p changes through the time. The trend analysis figures are useful in extracting the attention trajectory of governments to the different policies. Figure 4 shows Trend analysis for three policies "Science", "Trade" and "Industry" in time period 1938 to 2012. According to Figure 4 we could simply conclude that Trade policy seems more important to Irish government comparing to industry and science. Detailed interpretation of attention intonation for all 21 high-level policy areas is out of scope of this paper and needs comprehensive separate study.

$$TW_{c_j}^{y_k}(p_i) = \sum_{\forall d_m \in D(c_j, y_k)} weight(d_m, p_i) \tag{1}$$

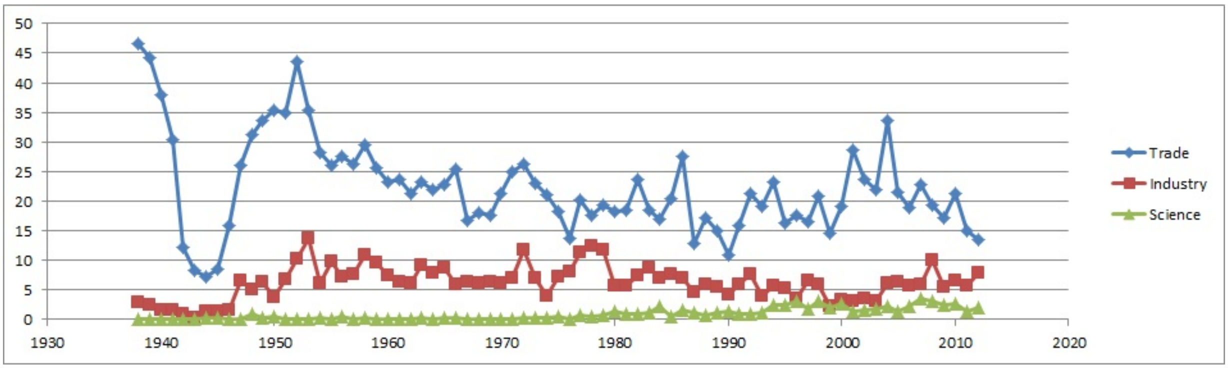


Fig. 4. Trend analysis for parameter values: $Country = Ireland$, $Policy = science, trade, industry$, $TimePeriod = 1938...2012$. The X axis shows different years between 1938 to 2012 and the Y axis is $TW_{c_j}^{y_k}(p_i)$.

To consider the relative attention allocation of all high-level policy areas, new type of analysis method called "Attention Allocation Analysis" is used by previous methods [14, 3, 1]. The result of this analysis is a stacked-area graph in which the total area of the graph represents the government total attention allocation and the region for each policy represents the proportion of the government's attention for that policy. Figure 5 shows Attention Allocation Analysis of Irish government in time period 1938 to 2012 considering all high-level Eurovoc categories shown in Table 1. The impressions one gets from Figure 5 are 1-The rough dominance of "Agri-FoodStuffs" and "Trade" in time period 1938 to 1960. 2-The rough dominance of "Transport" in time period 1960 to 1990. 3-The rough dominance of "Agriculture, Forestry and Fisheries" in time period 1990 to 2012. In a separate study, we could discuss the possible reasons behind the policies attention dominance in those time periods.

3.3 Entropy Analysis

To measure the government's diversity of attention towards 21 high-level policy areas, entropy values are calculated according to classified documents. To measure the entropy values, we use Shannons H information entropy which is used by most of the previous methods [13, 12, 11, 1]. Shannons H is a probabilistic measure of the spread of objects or observations across a defined number of nominal categories. In our case, Shannons H reflects how many different policies

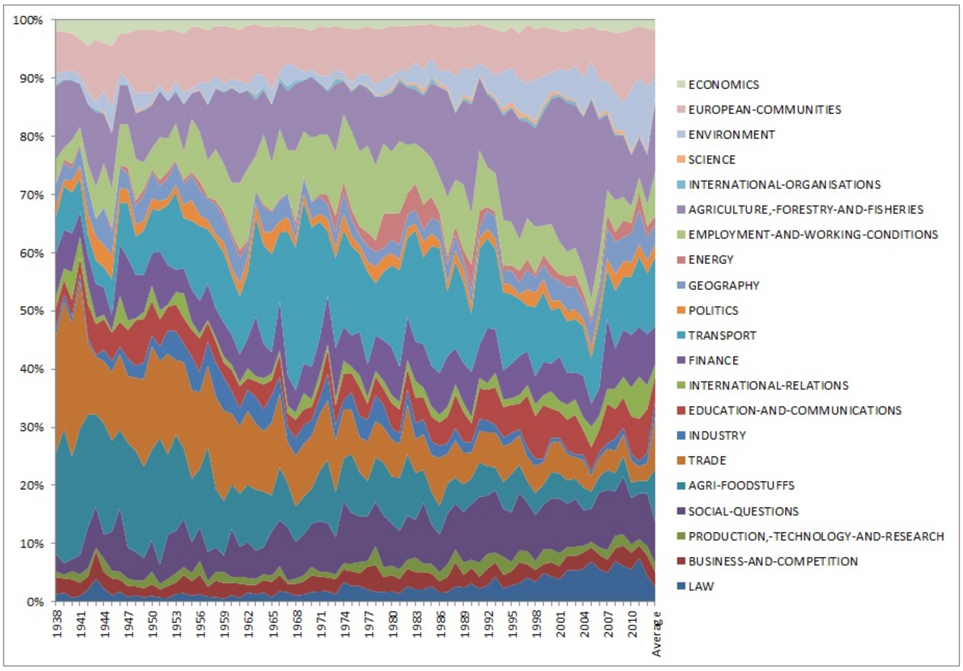


Fig. 5. Attention Allocation Analysis for parameter values $Country = Ireland$, $TimePeriod = 1938...2012$ considering all high-level Eurovoc categories shown in Table 1.

are considered in a dataset, and simultaneously how evenly the government's attentions are distributed among those policies. For a given number of policies, the value of a diversity is maximized when all policies have equal attention values.

Figure 6 shows the Entropy Analysis of Irish government in time period 1938 to 2012. Years 1939 and 2008 have minimum and maximum entropy values, respectively. These results are in line with our Attention Allocation Analysis shown in Figure 5. In $year = 1939$, Irish government's attention is more biased toward "Trade", "Agri-foodstuffs" and "Agriculture, Forestry and Fisheries" policies and this results in minimum entropy value. While in $year = 2008$, Irish government's attention is more evenly distributed among different policies and results in maximum entropy value for this year.

3.4 Correlation Analysis, Causality Analysis and Clustering Analysis

From the political science point of view, it is very important to find out the novel relationships among different policies. Here, we introduce three types of analysis: Correlation analysis, Causality analysis and Clustering analysis to discover hidden relationships among different policies.

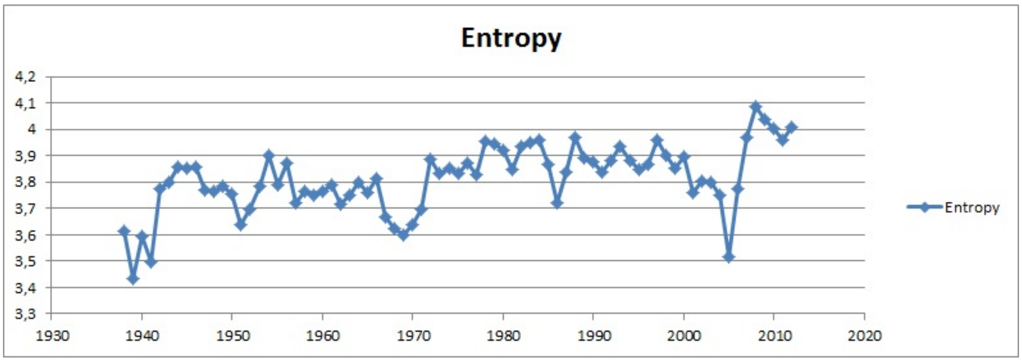


Fig. 6. Entropy Analysis for parameter values: $Country = Ireland$, $TimePeriod =$ 1938...2012. Years 1939 and 2008 have minimum and maximum entropy values, respectively.

Correlation between policies is a measure of how well the policies are following the same trend. The most common measure of correlation in statistics is the Pearson correlation, which shows the linear relationship between two variables. Table 2 shows the 5 most correlated policies for parameter values: $Country = Ireland$, $TimePeriod =$ 2000...2012 using Pearson correlation. The most correlated policy pair is (agri-foodStuffs, trade). Figure 7 shows the Trend analysis for these two policies. As it visible in Figure 7, agri-foodstuffs and trade policies, roughly following the same trend in $TimePeriod =$ 2000...2012.

Table 2. The 5 most correlated policies for parameter values: $Country = Ireland$, $TimePeriod =$ 2000...2012 using Pearson correlation.

policy 1	policy 2	correlation value
AGRI-FOODSTUFFS	TRADE	0.90
POLITICS	EUROPEAN COMMUNITIES	0.84
INTERNATIONAL RELATIONS	ENVIRONMENT	0.82
LAW	TRANSPORT	0.81
EDUCATION AND COMMUNICATIONS	GEOGRAPHY	0.79

The relationship among different policies could be causal which means by increasing/decreasing the attention allocation in one policy, considering a time lag, the attention allocation of another policy increases/decreases. We use Granger Test [8] to calculate the causality relationships among different policies. The Granger causality test is a statistical hypothesis test for determining whether one time series is useful in forecasting another. Figure 8 shows the casuality analysis for parameter values $Country = Ireland$, $TimePeriod =$ 2000...2012, $TimeLag = 1year$. According to Figure 8, Irish government's attention to policy "Agri-foodstuff" provides statistically significant information about future

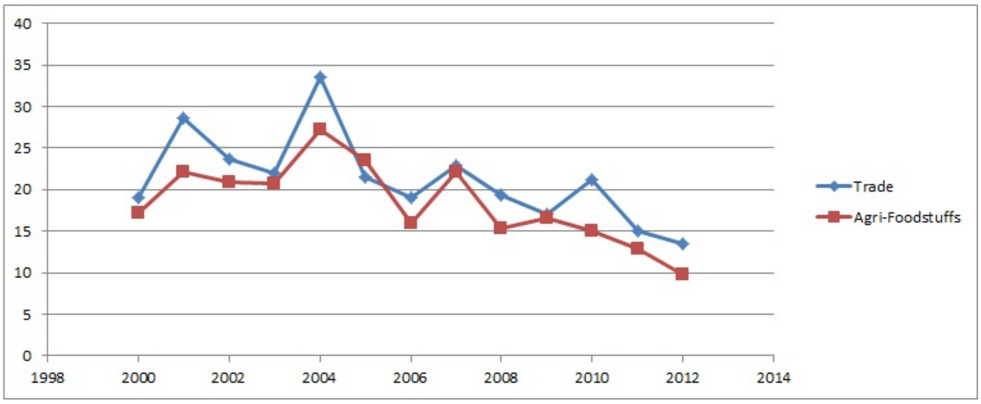


Fig. 7. Trend analysis for parameter values: $Country = Ireland$, $Policy = trade, agri-foodstuffs$, $TimePeriod = 2000...2012$. The X axis shows different years between 2000 to 2012 and the Y axis is $TW_{cj}^{y_k}(p_i)$.

attention values of policies “Finance”, “Geography”, “Employment and Working Condition” and “International Relations”.

Policy Clustering is the task of grouping set of policies in a way that policies in the same group (called cluster) are more similar to each other comparing to those in other clusters. We use WEKA [10] machine learning tool to cluster the policy areas according to K-means algorithm [17]. Figures 9 and 10 show the Irish policies cluster results for $year = 2000$ and $year = 2001$, respectively. One interesting analysis is considering the “Cluster Dynamics” through the time. How different clusters appear, integrate with each other or disappear through the time. For example, considering the cluster containing “Law”, “Environment”, “Geography”, “International Relations” as its members in Figure 9, only relation between “Law” and “Geography” will remain in the cluster analysis in Figure 10.

4 Conclusions

In this paper, we propose a new tool called PolicyMiner to promote the development and deployment of innovative computationally-based research techniques in large-scale data analysis with a focus on applying these to the substantive political science question of political representation in the Europe. The main task of the PolicyMiner project is to analyze the legislative documents of 15 European countries for at least 30 years time period which indicate the scale of the project undoubtedly as a large-scale project in the area of social science. The main task of this project can be divided into the several sub-tasks such as: (1) Locating the relevant documents, (2) harvesting the documents automatically, (3) applying data cleaning, (4) generating novel hypothesis, (5) evaluating the proposed hypothesis by using knowledge discovery tools, (6) interpreting the new findings

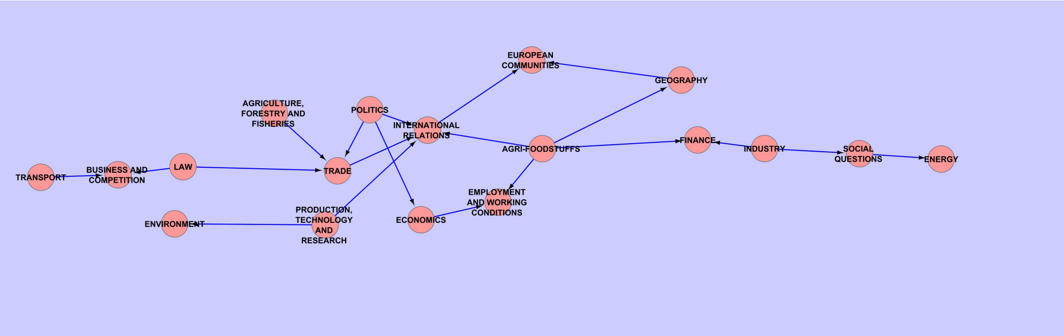


Fig. 8. Casuality analysis for parameter values $Country = Ireland$, $TimePeriod = 2000...2012$, $TimeLag = 1year$.

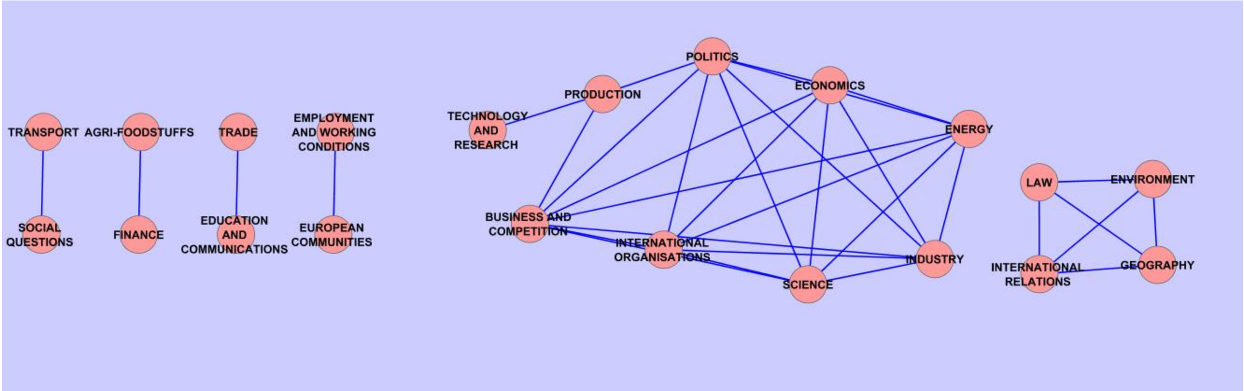


Fig. 9. Cluster Analysis for parameter values $Country = Ireland$ and $Year = 2000$.

and (7) displaying the findings through visualization tools. Handling these sub-tasks in a efficient way makes the nature of this project multidisciplinary. Since there is a need for collaboration of scholars with a political science background to handle sub-tasks 1, 4 and 6 and scholars from a computer science background should mainly provide their expertise for the sub-tasks of 2, 3, 5 and 7.

To handle sub-tasks 2, 3, 5 and 7, our proposed PolicyMiner has two main modules Crawler and Analyzer. The Crawler module is capable of harvesting the legislative documents from the governments' websites, cleaning the harvested data and persisting the cleaned data in DataBase Management System (DBMS). The analyzer module provides the list of predefined algorithms to analyze the legislative documents. So far, we have included the following data analysis capabilities in our PolicyMiner: Search Analysis, Trend Analysis, Attention Allocation Analysis, Entropy Analysis, Correlation Analysis, Causality Analysis and Clustering Analysis.

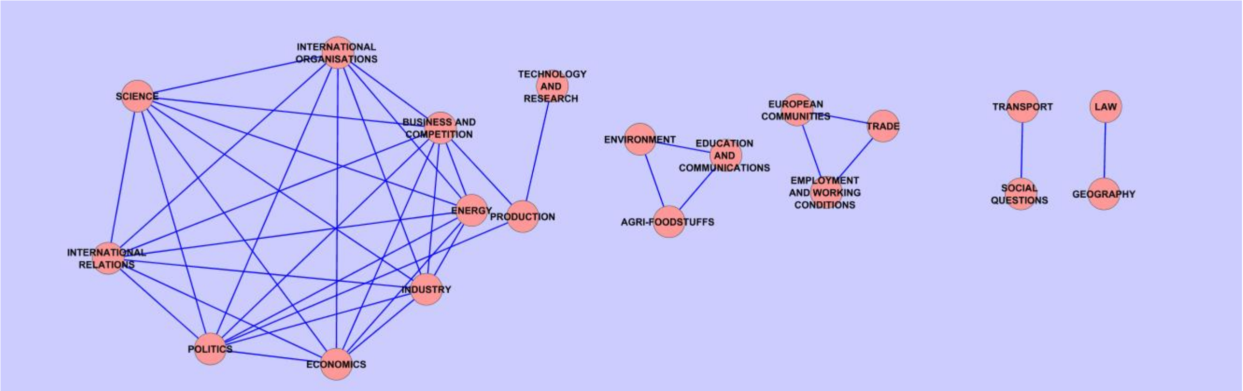


Fig. 10. Cluster Analysis for parameter values $Country = Ireland$ and $Year = 2001$.

Beside the functional capabilities of our proposed PolicyMiner, we aim to consider at least two related non-functional features: Make as much automatic as possible and make as less dependent to human as possible . Having these two features included, we provide the users of our system a semi-automatic approach in which the general work-flow with specific context variable points according to Template Pattern should be defined by the user. The motto here is find what varies and encapsulate it. Both modules Crawler and Data Analyzer will follow this general work-flow and they refer to external configuration files to fill in context-aware variables.

In the end, we believe that our PolicyMiner tool can be a great help for social scientist to gather and analyze the legislative documents. Using the external configuration files, the social scientists could easily tuned PolicyMiner according to their problem definition. As a future work, we could extend our PolicyMiner in two directions Data and Algorithms. We could gather the data related to European Parliment and Public Opinion and then, apply the discussed analysis on them. Regarding the Algorithms extension, we could consider more complex clustering and time-series algorithms in our PolicyMiner.

References

1. Petya Alexandrova, Marcello Carammia, and Arco Timmermans. Policy punctuations and issue diversity on the european council agenda. *Policy Studies Journal*, 40, 2012.
2. Christine Arnold, Mark Franklin, Christopher Wlezien, Eliyahu Sapir, and Hossein Rahmani. Policyvotes:http://www.policyvotes.org/, 2013.
3. Gerard Breeman, David Lowery, Caelesta Poppelaars, Sandra L. Resodihardjo, Arco Timmermans, and Jouke de Vries. Political Attention in a Coalition System: Analysing Queen's Speeches in the Netherlands 1945-2007. *Acta Politica*, 44(1):1–27+, 2009.

4. Ivan Bruha. Pre- and post-processing in machine learning and data mining. In *Machine Learning and Its Applications, Advanced Lectures*, pages 258–266, London, UK, UK, 2001. Springer-Verlag.
5. Peter Pin-Shan Chen. The entity-relationship modeltoward a unified view of data. *ACM Trans. Database Syst.*, 1(1):9–36, March 1976.
6. The EU multilingual thesaurus Eurovoc. Eurovoc, the eu's multilingual thesaurus:http://eurovoc.europa.eu/, 2013.
7. Erich Gamma, Richard Helm, Ralph Johnson, and John Vlissides. *Design patterns: elements of reusable object-oriented software.* Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1995.
8. C. W. J. Granger. Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica*, 37(3):424–438, August 1969.
9. Nizar Grira, Michel Crucianu, and Nozha Boujemaa. Unsupervised and semi-supervised clustering: a brief survey. In *in A Review of Machine Learning Techniques for Processing Multimedia Content, Report of the MUSCLE European Network of Excellence (FP6*, 2004.
10. Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA data mining software: an update. *SIGKDD Explor. Newsl.*, 11(1):10–18, November 2009.
11. Will Jennings, Shaun Bevan, Arco Timmermans, Gerard Breeman, Sylvain Brouard, Laura Chaqués-Bonafont, Christoffer Green-Pedersen, Peter John, Peter B. Mortensen, and Anna M. Palau. Effects of the core functions of government on the diversity of executive agendas. *Comparative Political Studies*, 44(8):1001–1030, August 2011.
12. Will Jennings, Shaun Bevan, Arco Timmermans, Gerard Breeman, Sylvain Brouard, Laura Chaqués-Bonafont, Christoffer Green-Pedersen, Peter John, Peter B. Mortensen, and Anna M. Palau. Effects of the core functions of government on the diversity of executive agendas. *Comparative Political Studies*, 44(8):1001–1030, August 2011.
13. Peter John and Will Jennings. Punctuations and turning points in british politics: the policy agenda of the queen's speech, 1940-2005. *British Journal of Political Science*, 40(3):561–586, April 2010.
14. Bryan D. Jones and Frank R. Baumgartner. Representation and Agenda Setting. *Policy Studies Journal*, 32(1):1–24, 2004.
15. S. B. Kotsiantis. Supervised machine learning: A review of classification techniques. In *Proceedings of the 2007 conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies*, pages 3–24, Amsterdam, The Netherlands, The Netherlands, 2007. IOS Press.
16. S. B. Kotsiantis and et al. Data preprocessing for supervised learning, 2006.
17. J. B. Macqueen. Some methods of classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967.
18. Peter Bjerre Mortensen, Christoffer Green-Pedersen, Gerard Breeman, Laura Chaqués-Bonafont, Will Jennings, Peter John, Anna M. Palau, and Arco Timmermans. Comparing government agendas: executive speeches in the netherlands, united kingdom and denmark. *Comparative Political Studies*, 44(8):973–1000, August 2011.
19. Bruno Pouliquen, Ralf Steinberger, and Camelia Ignat. Automatic annotation of multilingual text collections with a conceptual thesaurus. In *Proceedings of the*

Workshop Ontologies and Information Extraction at (EUROLAN'2003, pages 9–28, 2003.

20. Dorian Pyle. *Data Preparation for Data Mining (The Morgan Kaufmann Series in Data Management Systems)*. Morgan Kaufmann, March 1999.