TECHNICAL UNIVERSITY OF CRETE

Department of Electronics & Computer Engineering

Automatic creation of policy networks using Web documents

a thesis by

Bakirtzoglou Vagios

Supervisor Committee

Potamianos Alexandros (Supervisor)
Digalakis Vassilios
Petrakis Euripides

to my family

Table of Contents

List of Tables

List of Figures

ACKNOWLEDGMENTS

I would like to express my sincere regards to my advisor Dr. Alexandros Potamianos for his invaluable guidance.

I am also grateful to the members of the supervisor committee.

I would also like to thank Iosif Elias for his unreserved support.

Finally, I would like to thank my friends and especially my family for their love and sacrifices which made for me.

Introduction

Nowadays, there is a huge amount of information in the Web for any kind of research activities. Especially after the development of the Semantic Web, even more information has become available online and be structured into semantic classes, called ontologies. This is a crucial fact as information of Web is someway arranged according to similarity measures to different ontologies, which makes them easier to be mined. The semantic relationship between words plays a significant role in many occasions. For instance, query expansion in search engines is a popular research sector. In this case, words which have strong semantic relationship with the search's words are used also in the search so as to a better recall can be achieved. Thus, semantic similarities comprise a useful tool for the community of computer science.

On the other hand, the ways of using efficiently this available information have not developed in a satisfactory level for all research sectors. For example, some of the text mining techniques, which are developed by the information retrieval community, are not used to give a solution to a variety of problems that till now are being solved manually. One of these sectors of research that is being conducted manually is the creation of policy networks.

Unfortunately, manually constructing a policy network is a tedious, costly and time-consuming task, which also requires expert knowledge. Another disadvantage is the complete change of implementation in order to create a policy network for another country or language. Furthermore, the interest of examining the progress of the policy network through time requires the decrease of the time that is consumed by such a task. All these reasons raise the need for an automatic and unsupervised procedure.

The goal of this thesis is to automatically exact the policy network using web documents. This method relies to the hypothesis that the more similar the contexts of two words are, the stronger is the semantic relationship between these words. In other words, the policy terms that appear in similar lexical contexts are semantically similar. This method bases on this assumption and uses some metrics in order to exact the semantic similarity among political entities. The lexical environment was the only text's feature that was explored.

Moreover, the main contribution of this thesis is to evaluate the use of such an automatic and unsupervised method with reliable results that have already been produced. Thus, the method is evaluating for the first three experiments in comparison with the policy networks that have already been exacted by the political sector's researches, whereas in the final experiment, which concerns the creation of mapping among the most famous Greek political entities (politicians, political parties) with the similarities that excluded by a questionnaire which was issued in Greek citizens.

In this method two kinds of metrics are used. The first one is the page-count-based metrics which compute the semantic similarity with the processing of the co-occurrences of target words in web documents. The second one is the context-based metrics estimate the degree of semantic similarity among the political entities, by the proximity of the contexts' similarity.

The automatic method using Web resources presents a number of advantages which may lead to its use as a standard, after its optimization. One of this method's advantages is the exponential decrease of time needed to extract a policy network manually. This is a very important factor, as the research in political scene needs to be completed in short time of period, so as to be examined through time. Moreover, the language independence of the method is one of the crucial characteristics. For example, there are not differences of implementing this method to a variety of countries and languages. Finally, it does not require any expert knowledge in order to be conducted and comprises itself a cost-effective approach.

In the next chapters, the definitions of metrics and theoretical "tools" which used for the implementation of this method (Chapter 1) follow, the elaboration on the related work (Chapter 2) concerning either the researches which conducted from the political sciences' sector or from the community of computer sciences. In the same chapter is presenting the description of the unsupervised automatic method and in the following chapters are stated the experimental procedure (Chapter 3), the evaluation of the method (Chapter 4) and finally some conclusions and proposed future work (Chapter 5). At the end some figures, the questionnaire which used and the meaning of the abbreviation of the terms that used in experiments is presented in the Appendix.

1 Unsupervised Semantic Similarity Metrics

1.1 Introduction

This chapter introduces some basic material on unsupervised semantic similarity metrics. Firstly, it is important to understand the meaning of the term semantic. This word derives from Greek "σημαντικός" which means significant, from "σημαίνω" which means "to signify" and from "σήμα" which means "sign, mark, token". Thus, semantic is the meaning of a word which can be understood by its contexts. Additionally, semantic similarity defines as the degree in which two words have similar meaning. In this thesis the basic "tool" so as to extract the distances among policy entities is the use of metrics that estimate how strong the semantic relationship between terms is. These metrics can be separated into two categories; the page-count-based metrics and the fully-text based metrics (1). The first kind of metrics considers only the page counts returned by a search engine (2). The second kind of similarity metrics downloads a number of the top ranked documents and computes "wide-context" similarity among words using a "bag-of-words" model. These metrics rely on the concept that the similarity of context implies similarity of the target words. It is assumed that words, which appear in similar lexical environment (left and right contexts), have a close semantic relation (3), (4), (5). This assumption was advocated as "a word is characterized by the company it keeps" by J.R Firth (6) which is known in linguistics as the distributional hypothesis.

1.2 Page-count-based similarity metrics

The basic idea under this approach is that the words' co-occurrence is likely to indicate some kind of semantic relationship between words. A quick approximation of word co-occurrence can be estimated exploring the web. However, the number of pages in which a certain word pair co-occurs, does not express a direct semantic similarity. Moreover, it is reasonable to take into account the number of documents that include the each pair component individually for normalization purposes. In other words, for a word pair, we need to know the information that the two words share, normalized by the degree of their independence. Thus, in order to define each one of the four co-occurrence measures used in this work, we have to define the following elements (2) of Table 1:

$\boldsymbol{\{D\}}$**:** a set of containing the whole document collection that is indexed and accessible by a web search engine
$\boldsymbol{\|D\|}$**:** the number of documents in collection $\{\boldsymbol{D}\}$
$\boldsymbol{w_1}$**:** a term
$\boldsymbol{\{D\|w_1\}}$**:** a subset of $\{\boldsymbol{D}\}$, documents indexed by w_1
$\boldsymbol{\{D\|w_1, w_2\}}$**:** a subset of $\{\boldsymbol{D}\}$, documents indexed by w_1 and w_2
$\boldsymbol{f(D\|w_1)}$**:** the fraction of documents in $\{\boldsymbol{D}\}$ indexed with w_1
$\boldsymbol{(D\|w_1, w_2)}$**:** the fraction of documents in $\{\boldsymbol{D}\}$ indexed with w_1 and w_2

Table 1.Elements needed in co-occurrence measures' definition

1.2.1 Jaccard Coefficient

The Jaccard coefficient is a measurement which estimates the similarity (or distance) between sets. It is used a variation of the Jaccard coefficient defined as:

$$\mathbf{Jaccard}(\mathbf{w_1}, \mathbf{w_2}) = \frac{f(D|w_1, w_2)}{f(D|w_1) + f(D|w_2) - f(D|w_1, w_2)} \quad \textbf{(1.2.1)}$$

In probabilistic terms, Equation (1.2.1) finds the maximum likelihood estimate of the ratio of the probability of finding a document where words w_1 *and* w_2 co-occur over probability of finding a document where either w_1 *or* w_2 occurs. If w_1 and w_2 are the same word then the Jaccard coefficient is equal to 1 (absolute semantic similarity) whereas if these two words never co-occur in a document then Jaccard coefficient is equal to 0.

1.2.2 Dice Coefficient

The Dice coefficient is related to Jaccard coefficient and is defined as:

$$\mathbf{Dice}(\mathbf{w_1}, \mathbf{w_2}) = \frac{2f(D|w_1, w_2)}{f(D|w_1) + f(D|w_2)} \quad \textbf{(1.2.2)}$$

In the same way, Dice coefficient is equal to 1 when w_1 and w_2 is the same word and 0 when the two words never co-occur.

1.2.3 Normalized Google Distance

The Normalized Google Distance (NGD) (7) is a distance measurement between two sets. It is used a variation of NGD in this thesis, defined as follows:

$$\mathbf{NGD}(\mathbf{w_1}, \mathbf{w_2}) = \frac{\max\{\log(f(D|w_1)), \log(f(D|w_2))\} - \log(f(D|w_1, w_2))}{\log|D| - min\{\log(f(D|w_1)), \log(f(D|w_2))\}} \quad \textbf{(1.2.3)}$$

The range of NGD is between 0 and ∞. If $w_1 = w_2$ or if $w_1 \neq w_2$ but

$f(D|w_1) = f(D|w_2) = f(D|w_1, w_2) > 0$, then $\mathrm{NGD}(\mathrm{w_1}, \mathrm{w_2}) = 0$, which means that w_1 and w_2 has the absolute semantic similarity. On the other hand, if $f(D|w_1) = 0$, then we have $f(D|w_1, w_2) = 0$, so $\mathrm{NGD}(\mathrm{w_1}, \mathrm{w_2}) = \frac{\infty}{\infty}$, which we take to be 1 by definition. In this occasion w_1 and w_2 has similarity equal to 0. Moreover, NGD is symmetric but *is not a metric* as it can violate the triangle inequality in some circumstances.

For example, if $w_3 = w_1 \cup w_2$, $w_1 \cap w_2 = 0$, $w_1 = w_2 \cap w_3$, $w_2 = w_2 \cap w_3$, and $|w_1| = |w_2| = \sqrt{|D|}$ are chosen, then $f(D|w_1) = f(D|w_2) = f(D|w_1, w_3) = f(D|w_2, w_3) = \sqrt{|D|}$, $f(D|w_3) = 2\sqrt{|D|}$ and $f(D|w_1, w_2) = 0$. This means that $\mathrm{NGD}(\mathrm{w_1}, \mathrm{w_2}) = \infty$ and also means that $\mathrm{NGD}(\mathrm{w_1}, \mathrm{w_3}) = \mathrm{NGD}(\mathrm{w_2}, \mathrm{w_3}) = \frac{2}{\log(|\mathrm{D}|)}$. Unfortunately, this violates the triangle inequality.

1.2.4 Mutual Information

If we consider the occurrence of words w_1 and w_2 as random variables X and Y, respectively, then the point wise mutual information (MI) among X and Y measures the mutual dependence between the appearances of words w_1 and w_2 (9). The maximum likelihood estimate of MI is

$$\mathbf{MI(X,Y)} = \mathbf{log}\frac{\frac{f(D|w_1,w_2)}{|D|}}{\frac{f(D|w_1)}{|D|}\frac{f(D|w_2)}{|D|}} \quad \mathbf{(1.2.4)}$$

Mutual information measures the information that variables X and Y share. It quantifies how the knowledge of one variable reduces the uncertainty about the other. For instance, if X and Y are independent, then knowing X does not give any information about Y and the mutual information is 0. For X = Y, the knowledge of X gives the value of Y without uncertainty and the mutual information is 1. Note that the fractions of documents are normalized by the number of documents indexed by the search engine, |D|, giving a maximum likelihood estimate of the probability of finding a document in the web that contains this word.

1.3 Cosine Similarity

Cosine similarity is a "bag-of-words" model, which is based on the distributional hypothesis. Thus, we examine the context similarity in order to compute the similarity between words. The right and left context of length WS are considered for a word and the feature (word) vector $[v_{WS,L} \dots v_{2,L}\ v_{1,L}]w[v_{1,R}v_{2,R} \dots v_{WS,R}]$ is created, where $v_{i,L}$ and $v_{i,R}$ represent the i^{th} word to the left and to the right of w respectively. The feature vector for every word w is defined as $T_{w,WS=}(t_{w,1}, t_{w,2}, \dots, t_{w,N})$ where $t_{w,i}$ is a non-negative integer and WS is the context window size. Note that the feature vector size is equal to the vocabulary size N, i.e., we have a feature for each word in the vocabulary V. The i^{th} feature value $t_{w,i}$ reflects the occurrences of vocabulary word v_i within the left or right context window WS. This feature value is set according to several schemes, known from bibliography and are presenting in Table 2.

$w = tf/max\,\{tf\}$
$w = IDF = log(^N/_n)$
$w = tf * IDF = tf * log(^N/_n)$
$w = tf * IDF = tf * log((N-n)/n)$

Table 2.Possible values of feature $t_{w,i}$

where tf as the term frequency, $max\,\{tf\}$ as the maximum term frequency in a document, N as the number of documents in a collection and n as the number of documents containing a query term. Finally, the cosine similarity (cosine angle) between two words w_1 and w_2 is computed from the cosine similarity of their feature vectors $T_{w1,WS}$ and $T_{w2,WS}$ respectively:

$$\boldsymbol{CS_{WS}(w_1, w_2) = \frac{\sum_{i=1}^{N} t_{w_{1,i}} t_{w_{2,i}}}{\sqrt{\sum_{i=1}^{N}\left(t_{w_{1,i}}\right)^2}\sqrt{\sum_{i=1}^{N}\left(t_{w_{2,i}}\right)^2}}} \quad \textbf{(1.3)}$$

1.4 Contextual Similarity metrics

The computational model of semantics is referred as word-space model by Hinrich Schütze (8). A model that measures the semantic relationship between words is defined with respect to the vocabulary which forms an n-dimensional space, where n is the cardinality of the vocabulary. Thus, each vocabulary's word can be considered as one dimension. This model's concept is that the semantic similarity can be represented as proximity in vocabulary's dimensional space.

Spatial proximity between words as a representation of their semantic similarity seems to be very intuitive and naturally derived with respect to the way that human conceptualizes similarities. This geometric metaphor of meaning has been pointed out by the work of Lackoff and Johnson (9), (10). They state that metaphors form the raw base of abstract conceptualization. Moreover, they argue that these metaphors are used by human mind for reasoning about abstract and complex phenomena, such as natural language and semantics.

This physical tendency of human mind has the result of placing the conceptual locations of words with similar meaning to be "near" each other, while the dissimilar words are placed "far apart". Of course, a sole word in a high-dimensional space gives no additional information for deeper understanding of the word. The space must be populated with other words in order to apply the proximity as an indicator of similarity. The geometric metaphor of meaning conceptualizes the words as locations in a word-space and the similarity is considered as the proximity between the locations (11).

1.4.1 The distributional hypothesis of semantic similarity

The word-space model provides not only a spatial representation of meaning, but also naturally suggests a way to build the model. The model's only requirement is that the words have to be used without the need of a priori knowledge or constraints about the underlying semantics. Statistical approaches are valuable tools in order to learn the distributional properties of words with an unsupervised way. This framework is suitable for measuring the proximity as is reflected by the distributional similarity. Hence, the concept of the model is the estimation of the semantic similarity by the distributional hypothesis of semantically similar words. This hypothesis assumes that words with similar contexts have similar meaning. One of the first studies of the distributional hypothesis is the work of Rubenstein and Goodenough (12), who stated that “words which are similar in meaning occur in similar contexts”. Schütze and Pedersen (13) re-phrased this hypothesis, considering the data sparseness problem, as “words with similar meanings will occur with similar neighbors if enough text material is available”. The linguist Zelling Harris (14),(15) initially believed that it is possible to typologize the whole of linguistic phenomena using their distributional behavior without intrusion of other features. Later, he extended his distribution-based analysis, considering that in many cases the meaning goes beyond the formal linguistic theory, affected by many extralinguistic factors, such as social situations. Even in these cases, Harris suggested that a distributional correlation will always exist between the extralinguistic factors and the influenced linguistic phenomena. The basis of his work is that the differences of meaning are usually characterized by differences of distribution. For instance, if two words word1 and word2 are more different in meaning that word1 and word3 then will often be found that the distributions of word1 and word2 may be more different than the distributions of word1 and word3. To sum up, there is a correlation between the difference of meaning and the difference of contexts.

In 1965, the earliest proof of the distributional hypothesis was conducted by Rubenstein and Goodenough (12). They compared the contextual similarities of 65 noun pairs with synonymy scores assigned by students. In their work they pointed out that there is a correlation between the degree of semantic similarity between a pair of words and the degree to which their contexts are similar. Moreover, Rubenstein and Goodenough, state that the generalization of the above conclusions is dependent on factors like vocabulary size and homogeneity of content. Three decades later (1991) Miller and Charles (16) repeated the experiment of Rubenstein and Goodenough using 30 of the 65 pairs and they reached similar results, supporting the distributional hypothesis. Thus, the distributional hypothesis seems to formalize a useful tool, operating on the broad notion of semantic similarity (11).

1.4.2 Several Schemes

In this work a variation of cosine similarity measure is used in order to compute the similarity between the political entities. This metric is based on the distributional hypothesis, which was mentioned before. Thus, the degree of similarity of the terms’ context leads to the estimation of the similarity between the terms. This metric CS_{WS} computes the “wide-context” similarity using a “bag-of-words” model (2).

In “bag-of-words” models a context window size (WS) is selected for each word w in the vocabulary. The right and left contexts of length WS in the corpus are considered for word w, e.g.$[v_{WS,L} \ldots\ v_{2,L}\ v_{1,L}]w[v_{1,R}v_{2,R} \ldots\ v_{WS,R}]$, where $v_{i,L}$ and $v_{i,R}$ represent the i^{th} word to the left and to the right of w respectively. The feature vector for every word w is defined as $T_{w,WS=}(t_{w,1}, t_{w,2}, \dots, t_{w,N})$ where $t_{w,i}$ is a non-negative integer and WS is the context window size. Note that the feature vector size is equal to the vocabulary size N, i.e., we have a feature for each word in the vocabulary V. The i^{th} feature value $t_{w,i}$ reflects the occurrences of vocabulary word v_i within the left or right context window WS. This feature value is set according to one of the several schemes [Binary (Bin.), a Logarithm of Term frequency (Log (freq)) or the normalized Logarithm of Term Frequency Scheme], that are described below.

1.4.2.1 Binary Scheme

As far as the binary scheme is concerned, the $t_{w,i}$ equals to 1 if the word v_i appears within the left or right WS size context for the word w and 0 if the word v_i does not appear to the left or right context of word w. Finally, the similarity between two words w_1 and w_2 is computed from the cosine similarity (Eq. 1.3) of their feature vectors $T_{w1,WS}$ and $T_{w2,WS}$ respectively (5), (17).

1.4.2.2 Term Freq Scheme

In this scheme the $t_{w,i}$ takes positive values equal to the frequency of the appearances of word v_i within the left or right WS size context for the word w and 0 if the word v_i does not appear to the left or right context of word w. Finally, the similarity between two words w_1 and w_2 is computed from the cosine similarity of their feature vectors $T_{w1,WS}$ and $T_{w2,WS}$ respectively and is described by the above stated Eq. (1.3).

1.4.2.3 Term Log (Freq) Scheme

In this scheme the $t_{w,i}$ equals to the logarithm of the frequency that the word v_i appears within the left or right WS size context for the word w and 0 if the word v_i does not appear to the left or right context of word w. The similarity between two words w_1 and w_2 is computed from the cosine similarity of their feature vectors $T_{w1,WS}$ and $T_{w2,WS}$ as it is seemed to the Eq. (1.3). This scheme performs better if we have a large number of documents because for frequencies close to 1 the use of logarithm normalizes the $t_{w,i}$ to zero.

1.4.2.4 Term Normalized Log (Freq) Scheme

In this occasion the value of $t_{w_j,i}$ differs from the Log (Freq) Scheme as it is normalized over the logarithm of the frequency that word w_j appears in the document. This scheme performs more efficiently as the occurrences are balanced after the normalization. This Scheme is described by Eq. (1.4.2.4):

$$\boldsymbol{CS_{WS}(w_1,w_2)} = \frac{\sum_{i=1}^{N} t_{w_{1,i}} t_{w_{2,i}}}{\sqrt{\sum_{i=1}^{N}\left(\frac{t_{w_{1,i}}}{\log f(w_1)}\right)^2}\sqrt{\sum_{i=1}^{N}\left(\frac{t_{w_{2,i}}}{\log f(w_2)}\right)^2}} \quad \mathbf{(1.4.2.4)}$$

1.5 Summary

The basic advantage of these methods is the fact that they are fully unsupervised. This allows the automatic conduction of the research without the need of experts' knowledge. Moreover, the decrease of the time that is consumed for the completion of manually policy creation task is another crucial advantage of these semantic similarity metrics, which gives the option of examining the influence to a procedure during time.

On the other hand, page-count-based similarity metrics, despite their simplicity, are not a reliable measure of co-occurrence of two words, as they present lots of drawbacks (1). Firstly, page-count-based metrics ignore the position of a word in a document. Thus, even though two words may appear in a web page, it does not imply that they are really related. Moreover, page counts of a word with more than one senses, might contain a combination of all its senses. For instance, page counts for the word *apple* contain page counts for *apple* as a fruit and *apple* as a company. Moreover, given the scale and noise in the Web, some words might occur arbitrarily on some pages. For all these reasons, page-count-based metrics are unreliable when measuring semantic similarity. On the other hand, fully-text-based metrics cognize the distinctiveness of these occasions and thus they are more efficient measures of semantic similarity, relatively with page-count-metrics.

2 Related Work in policy network construction

This chapter contains the description of the main contribution of this thesis, which is the automatic and unsupervised method in order to create policy networks. The resources that used of this method for policy networks' extraction derived from the Web. This huge amount of "hidden" information can be found with the use of a search engine and be processed with natural language processing techniques in order to estimate a semantic relationship among the political entities. The concept of the method is based on the contextual hypothesis which states the fact that similar contexts imply similar meaning.

2.1 Introduction

Social networks recently attracted considerable interest. With the intention of utilizing social networks for Semantic Web, several studies from computer science community have examined automatic extraction of social networks. Particularly, one interesting sector of social networks is the political social networks. Various systems have been developed up to this target from Mika (18) and Matsuo et al.(19). However, the results of these methods have no criterion of evaluation.

The contribution of this thesis is the implementation and the evaluation of an unsupervised automatic method for the creation of policy networks. The concept of this task is the re-creation of policy networks that have already been extracted manually. The manual creation of these political social networks has been implemented in the framework of the European study about the decentralization degree of the Second Community Support Framework (CSF). Unfortunately, the manual extraction of these studies was tedious, costly and time-consuming and also required expert knowledge. For this reason, an automatic method is more preferable if its results are reliable. Thus, the results of the automatic and unsupervised method are evaluated in comparison with the results of the manual method. More specifically, two policy networks are extracted. The first one concerns the regional policy network of the South Aegean Region, in Greece (20) and the second one the Mid-West Region of Ireland (21).

2.2 Related Work

In this epoch social networks provoke great interest in research. This may be caused by the apparition of commercialization of social networks by companies such as Facebook, MySpace and others. Another domain of social networks is the policy networks. This sector also concentrates great interesting of research from the community of political sciences. Unfortunately, the manually implemented methods used by this community constitute a costly and time-consuming task. This fact attracted the interest of the community of computer science to create an unsupervised method. The innovation of this method is the use of lexical features from the web in order to extract policy networks with an automatic and unsupervised way.

2.2.1 Methods in Political Sciences

Networks are usually applied either for evaluation purposes or in order to compare maps. They might also be useful in representing the state of the art in some field (22). An example of the use of networks in evaluation purpose is the task that was conducted by the team of Getimis and Demetropoulou. In their article they tried to evaluate the level of decentralization of the second Community Support Framework (CSF) to the policy-making structures of the Southern Aegean, one of Greek regions (20). For this reason, they constructed a map in which the actors are located according to their decentralization degree, as shown in Figure 1. For example political entities that cooperated with the majority of the actors, like Regional Secretariat (RS), are in the centre of the graph.

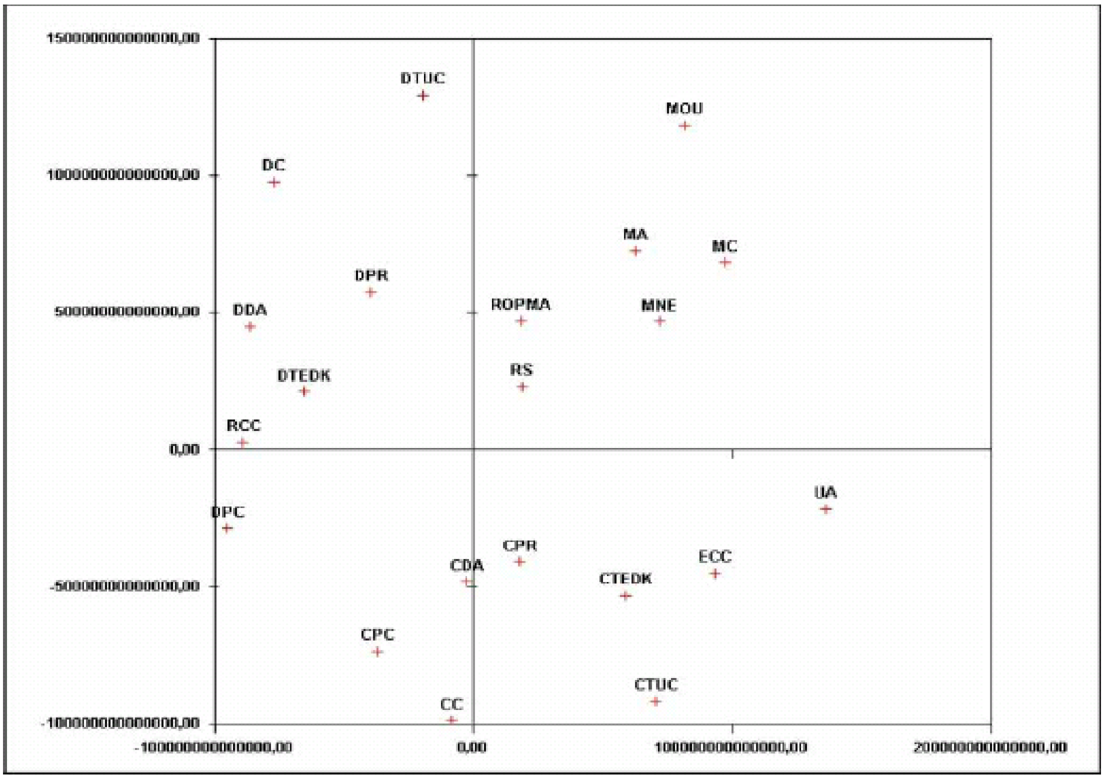


Fig 1.Policy network in Southern Aegean Region of Greece

This task was completed after two years of research. In this time a number of interviews were conducted, questionnaires were sent to be completed by the actors of the graph and several time-consuming procedures took place in order to extract the adjacency matrix of Figure 5, which used for the creation of the policy network. The same project was assigned by the European Commission to several groups of political sciences in order to create the policy networks of other countries such as Hungary (South Transdanubia Region) and Ireland (Mid-West Region) (23), (21) whose results are presented in Figures 2 and 3 respectively.

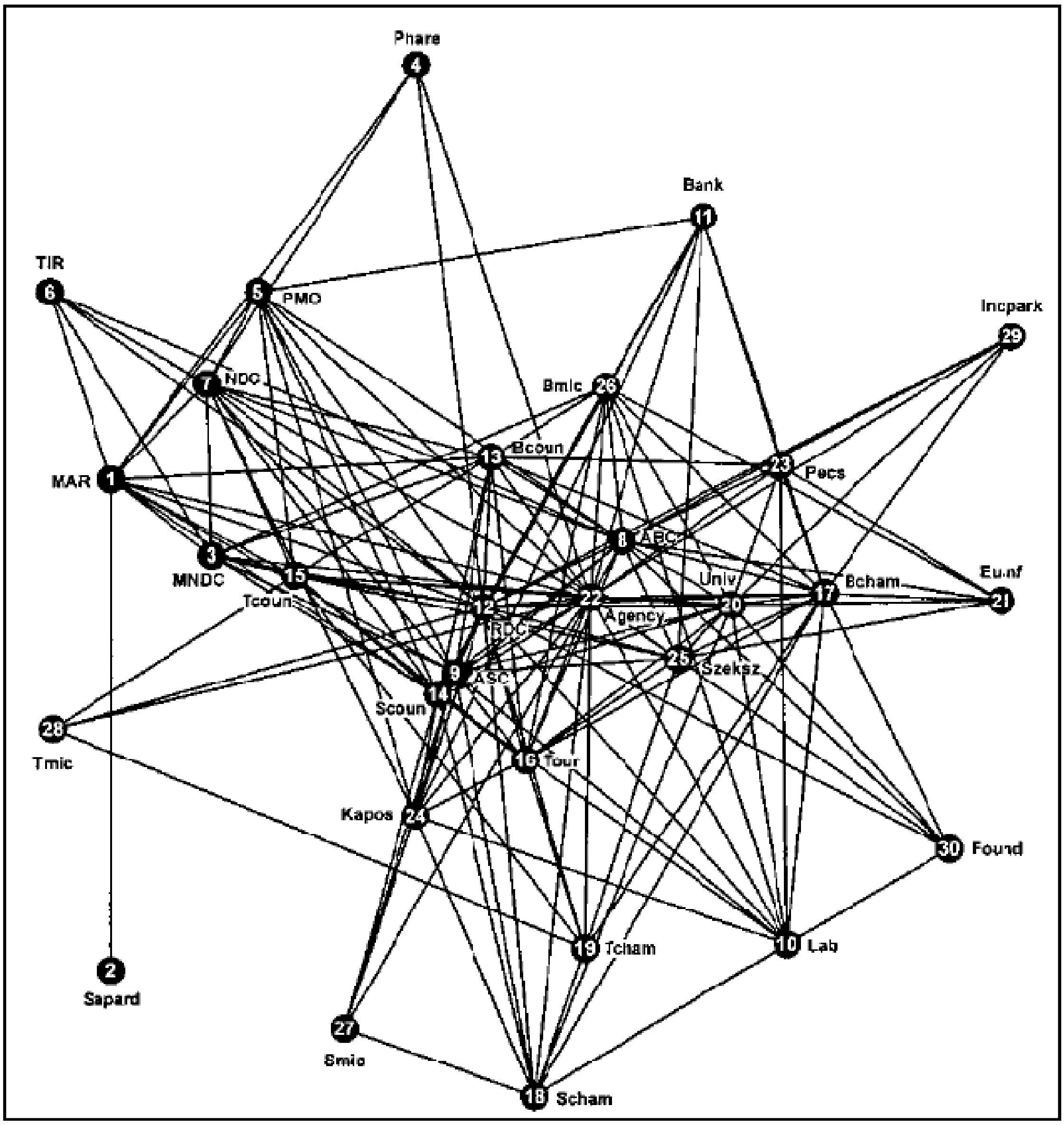


Fig 2.Policy network of Transdanubia

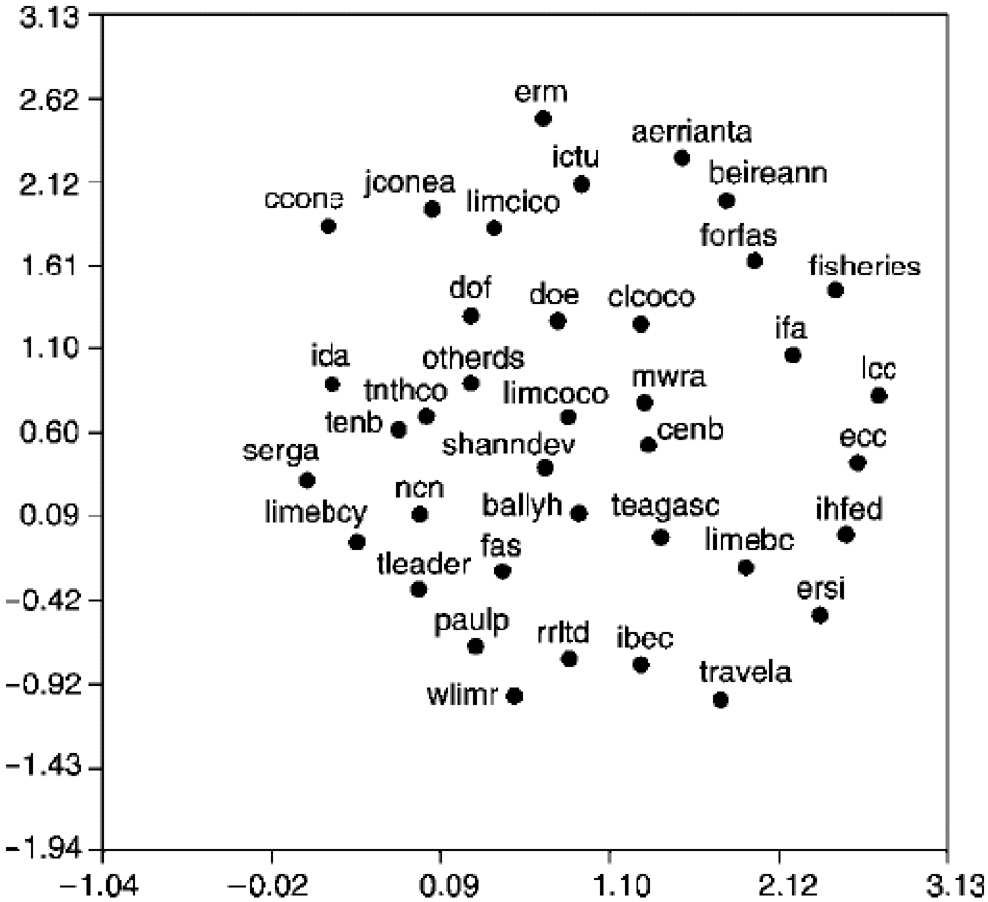


Fig 3.Policy Network of Mid-West region

The edges in Figure 2 mean that there is relationship between the two actors-nodes that are linked. These policy networks constructed with the use of adjacency matrixes, which resulted from regional and federal surveys. Computer Science Community tries to put in practice methods that implement this part of policy network construction fully automatic and unsupervised.

2.2.2 Methods in Computer Science

Lots of research has been conducted in the area of social networks' extraction. Most of the models that try to simulate the human model in order to compute semantic similarities among terms are based on the contextual hypothesis. This assumption defines that similar contexts imply similar meaning of terms. Several Systems have created that use web resources in order to automatically construct such a network. Firstly, in 1997 Kautz and Selman developed a social network extraction system from the Web, called Referral Web (24). This system addressed co-occurrences of names in web documents with the use of a search engine. The query that was used was in the form of "name1 AND name2" and if there were lots of page counts, this meant that the two people presented great semantic similarity. Moreover, the fact that a path from a person to another person could be obtained automatically having used the system is really interesting. Recently, Mika developed a system for extraction, aggregation and visualization of online social networks for a Semantic Web community, called Flink (18). Similarly to Referral Web, this system also uses co-occurrence analysis in order to extract similarities. Thus, a given set was inserted in the system and with the use of a search engine page counts for individual names as well as co-occurrences of the pair were obtained. Furthermore, Harada et al. (25) developed a system which could extract names and person-to-person semantic relationships from the Web, using co-occurrences. Faloutsos et al. (26) extracted a social network of 15 million persons from 500 million web documents using their co-occurrences within a window of 10 words. Knees et al. (27) classified artists in genres using co-occurrences of names and keywords of music in the top 50 ranked pages retrieved by a search engine. In 2002, Matsuo et al. created a social extraction system from the Web, called POLYPHONET. This system targeted to extract relations of persons, to group persons and obtain keyword for a person. They used matching coefficient and Jaccard coefficient in order to obtain co-occurrences. Another idea, which they experimented, was the use of a threshold. If the threshold was greater than page counts (the threshold equaled to 30) then the similarity was set to zero. This heuristic increased the efficiency of POLYPHONET. In 2006, Mori, Matsuo and their groups succeeded to extract labels which described the relations among entities (19). The extraction of the labels was succeeded by the clustering of the similar entity pairs according to their collective contexts in web pages. The labels, which described the relations among entities, were the result of the clustering procedure. This was also the first research in extracting a policy network with an unsupervised method. Their system is presented in Figure 4.

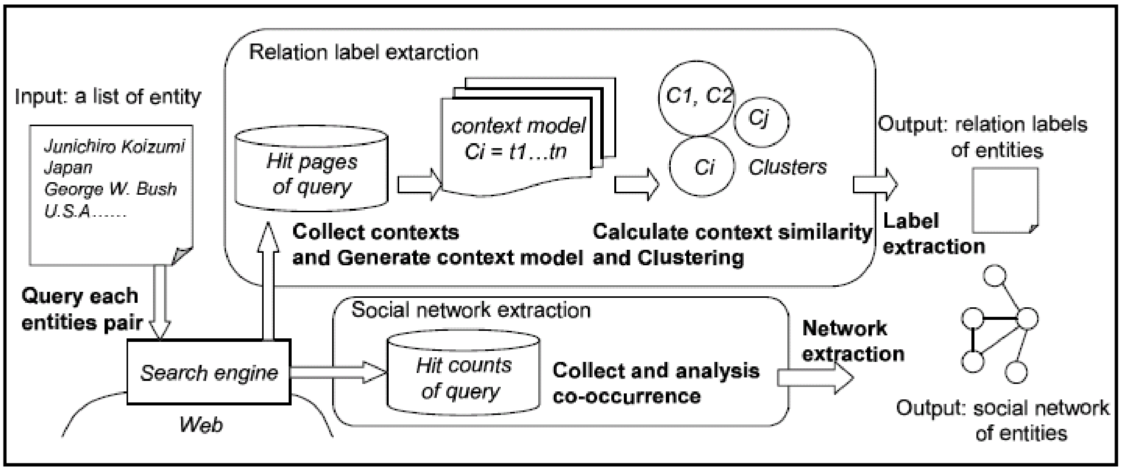

Relation label extarction
Input: a list of entity
Junichiro Koizumi
Japan
George W. Bush
U.S.A......
Hit pages of query
context model Ci = t1...tn
C1, C2
Cj
Ci
Clusters
Output: relation labels of entities
Collect contexts and Generate context model
Calculate context similarity and Clustering
Label extraction
Query each entities pair
Social network extraction
Search engine
Web
Hit counts of query
Collect and analysis co-occurrence
Network extraction
Output: social network of entities

Fig 4.Automatic system of extracting descriptive labels for social networks (19)

2.3 Our approach

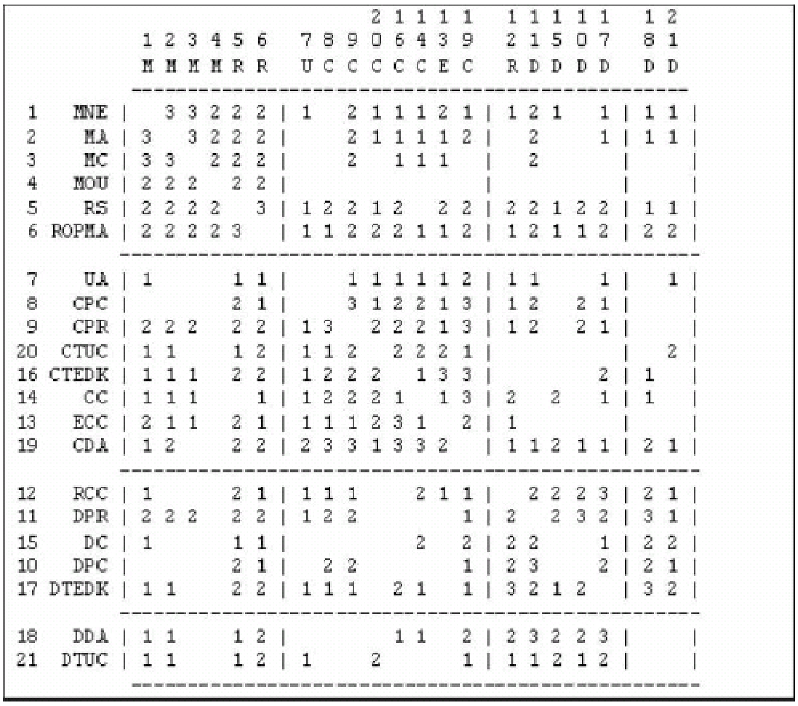
The main purpose of this thesis is to automatically create the policy networks which were manually constructed by the community of political sciences for the cases of Greece and Ireland. Thus, there is the chance of evaluating our method with reliable results.

Moreover, page-count-based metrics and fully-text-based metrics are used in order to exact the similarities between political entities. It is expected that the variance of cosine similarity that succeeds the best correlation of the unsupervised methods in bibliography (2) will work more efficiently than the other metrics. Finally, there is the possibility of comparison between the mapping of the similarities that result from the automatic and the manual method.

3 Unsupervised automatic method for policy networks creation

3.1 Automatic creation of policy network of Southern Aegean, Greece

The goal of this experiment is to exact the similarities among all pairs that exists in the adjacency matrix (Figure 5) of the study that was conducted by the community of political sciences (20) for the region of Southern Aegean.



```
                                  2 1 1 1 1    1 1 1 1 1    1 2
            1 2 3 4 5 6    7 8 9 0 6 4 3 9    2 1 5 0 7    8 1
            M M M M R R    U C C C C C E C    R D D D D    D D
           ------------------------------------------------------
 1    MNE |   3 3 2 2 2 | 1   2 1 1 1 2 1 | 1 2 1   1 | 1 1 |
 2     MA | 3   3 2 2 2 |     2 1 1 1 1 2 |   2     1 | 1 1 |
 3     MC | 3 3   2 2 2 |     2   1 1 1   |   2       |     |
 4    MOU | 2 2 2   2 2 |                 |           |     |
 5     RS | 2 2 2 2   3 | 1 2 2 1 2   2 2 | 2 2 1 2 2 | 1 1 |
 6  ROPMA | 2 2 2 2 3   | 1 1 2 2 2 1 1 2 | 1 2 1 1 2 | 2 2 |
           ------------------------------------------------------
 7     UA | 1       1 1 |     1 1 1 1 1 2 | 1 1     1 |   1 |
 8    CPC |         2 1 |     3 1 2 2 1 3 | 1 2   2 1 |     |
 9    CPR | 2 2 2   2 2 | 1 3   2 2 2 1 3 | 1 2   2 1 |     |
20   CTUC | 1 1     1 2 | 1 1 2   2 2 2 1 |           |   2 |
16  CTEDK | 1 1 1   2 2 | 1 2 2 2   1 3 3 |         2 | 1   |
14     CC | 1 1 1     1 | 1 2 2 2 1   1 3 | 2   2   1 | 1   |
13    ECC | 2 1 1   2 1 | 1 1 1 2 3 1   2 | 1         |     |
19    CDA | 1 2     2 2 | 2 3 3 1 3 3 2   | 1 1 2 1 1 | 2 1 |
           ------------------------------------------------------
12    RCC | 1       2 1 | 1 1 1     2 1 1 |   2 2 2 3 | 2 1 |
11    DPR | 2 2 2   2 2 | 1 2 2         1 | 2   2 3 2 | 3 1 |
15     DC | 1       1 1 |           2   2 | 2 2     1 | 2 2 |
10    DPC |         2 1 |   2 2         1 | 2 3     2 | 2 1 |
17  DTEDK | 1 1     2 2 | 1 1 1   2 1   1 | 3 2 1 2   | 3 2 |
           ------------------------------------------------------
18    DDA | 1 1     1 2 |         1 1   2 | 2 3 2 2 3 |     |
21   DTUC | 1 1     1 2 | 1     2       1 | 1 1 2 1 2 |     |
           ------------------------------------------------------
```

Fig 5.Adjacency matrix with similarities between political entities (20)

The possible values of this matrix could be 0 (for no semantic similarity between two political entities) till 4 (perfect similarity between two political entities). As the metrics that were used require either the web page counts or web documents for exacting the similarity degree, the need of a search engine was crucial. Therefore, Yahoo Search Engine was used via Perl module of Yahoo Search (yahoo search API). Moreover, for each one of the 21 political terms there is more than one expression. Due to this fact, the query which considered more efficient and was used to the search engine was the below stated:

$(\text{“term}1_{\text{exprA}}\text{” OR } \ldots\ldots \text{ OR “term}1_{\text{exprN}}\text{”) AND (“term}2_{\text{exprA}}\text{” OR } \ldots\ldots \text{ OR “term}2_{\text{exprN}}\text{”)}$

The good precision ensured with the use of quotes around the words of an expression and a good recall achieved by the use of all the possible expression for a political entity among “OR” statement. As a result, either the page counts or the URLs of the top ranked pages are returned for every pair. In the case of page-count metrics, whose procedure is simpler than the contextual metrics, the page counts returned from the search of each pair were used, with the page counts for the individual political term of the pair and in some occasions the number of documents that are indexed of the search engine and the similarity was exacted. Whereas, in the case of fully-text-based metrics, there are more "steps" till the completion of the experimental process. The top 100 ranked web documents had to be downloaded, filtered by HTML tags and be concatenated into one document per pair. In the case of Greek experiment was observed the following phenomenon. Some writers of the documents which downloaded forgot to change the mode from English language to Greek language. Hence, when they realized this fact, they erased all the letters till the capital letter of a word, for example the first letter of a sentence or the first letter of a name. This happens because there are 14 common capital letters between Greek and English alphabet. So, the writer could not distinguish that the error was happened also for these letters. This fact caused problems to the processing of the text. This solved by the creation of a pre-filtering script, which transformed all these English capital letters to the corresponding Greek capital letter. Afterwards, all the letters of the document were transformed into capital letters. This leaded to a more comfortable way of string matching. Furthermore, the words that consisted the terms of political entities were transformed into the unique word "term_i". This procedure is presented in Figure 6. Finally, another improvement could be the extraction of the stop words that the web documents contained. A Greek stop word list was found, and these words were excluded from the web documents (28).

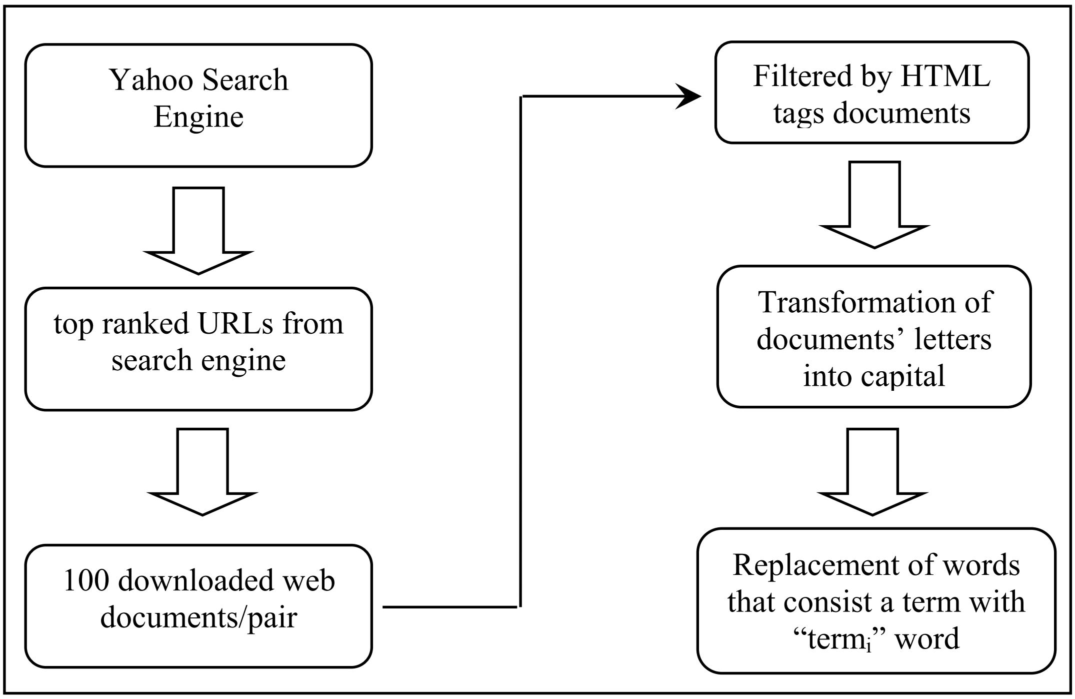
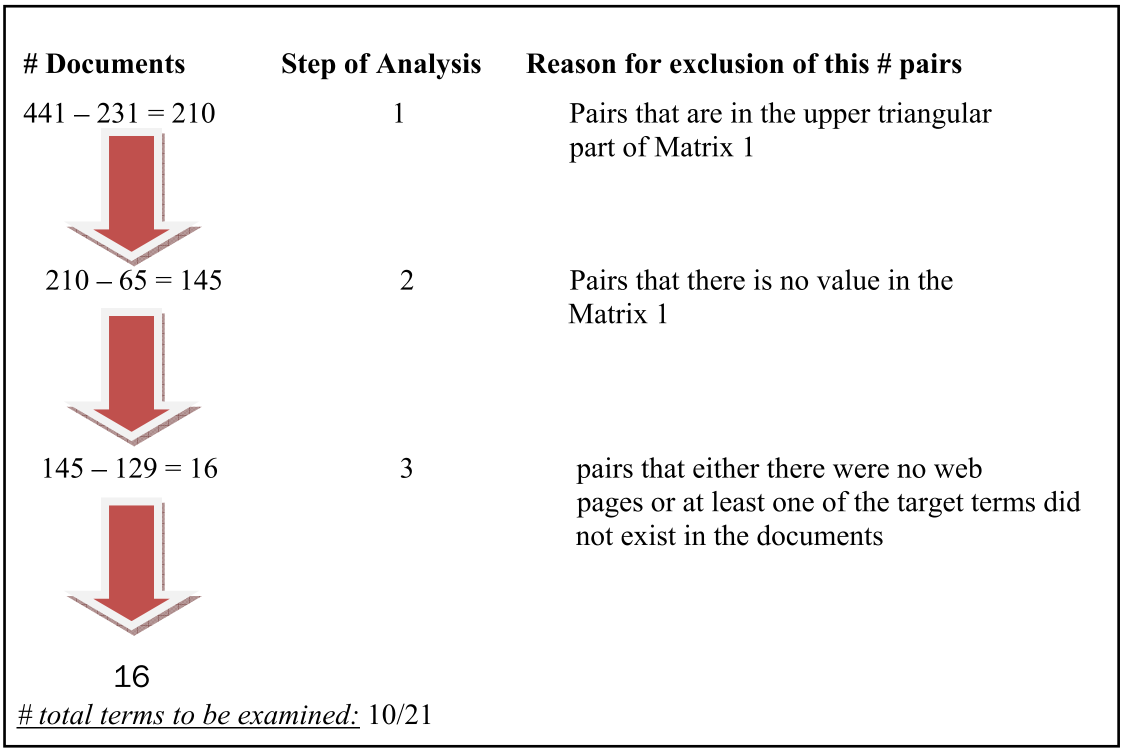

Yahoo Search Engine
top ranked URLs from search engine
100 downloaded web documents/pair
Filtered by HTML tags documents
Transformation of documents' letters into capital
Replacement of words that consist a term with "term_i" word

Fig 6.Experimental procedure for creation of final form of web documents for each pair

In the case of Southern Aegean's region, the dimension of the adjacency matrix is 21x21 elements. Thus, there are 441 documents (one document per pair) to be examined. An analysis to these documents was conducted, in order to examine their quality. At the end of the analysis, the number of the pairs which decided to be examined was 16. Firstly, from the 441 initial number of pairs were excluded 231 documents (as only the upper triangular matrix has to be examined). The format of matrix is item – item, thus the diagonal of matrix and the lower triangular matrix is useless information (i.e. the upper triangular part is symmetric of lower triangular part). In other words, there are duplicate values. The remaining number of documents is 210. From these documents 65 were excluded, as there were not values in the Matrix 1 for these pairs, which means that the degree of the relation did not examine by the manual method. The remaining number to the current step of the analysis is 145. In the final step of this analysis, 129 pairs were excluded, as either there were no web pages for these pairs or at least one of the target terms did not exist in the documents. To sum up, only 16 pairs of the initial number (210 pairs) were examined by the unsupervised method. The number of separate policy entities concluded in these 16 pairs is 10. This analysis is presented in Figure 7 in order to be easily understood by the readers.



# Documents	Step of Analysis	Reason for exclusion of this # pairs
441 – 231 = 210	1	Pairs that are in the upper triangular part of Matrix 1
210 – 65 = 145	2	Pairs that there is no value in the Matrix 1
145 – 129 = 16	3	pairs that either there were no web pages or at least one of the target terms did not exist in the documents
16		

# total terms to be examined: 10/21

Fig 7.Number of documents during quality analysis to the documents

Finally, the similarities that produced by the metrics have continuous value range. In order to transform these similarities from continuous range (0,1) to discrete values (0,1,2,3,4) we apply a heuristic method. At start, we assign the sorted array of similarities to array'. In order to divide the elements of array into 3 parts, it was assumed that the distribution of values 1, 2 and 3 is uniform. Therefore, two thresholds were needed. The first threshold has the value of the element in $rank\left[\frac{length(array\prime)}{3}\right]$ of the sorted array' and the second threshold has the value of the element in $rank\left[\frac{2\cdot length(array\prime)}{3}\right]$. This heuristic method is presented to the Figure 8.

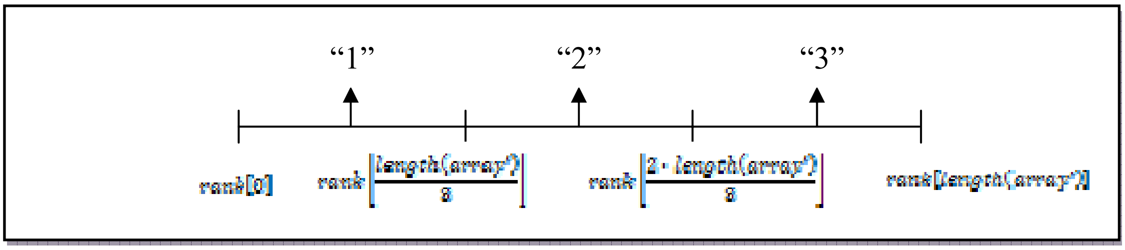


Fig 8.Heuristic of transformation into discrete value range

3.2 Automatic creation of policy network of Mid-West Ireland, Ireland

This experiment is much similar to the previous one. The difference is that the terms that were inserted in the search engine, were from the Mid-West region of Ireland. In this case, the values of the similarities among all pairs presented in the adjacency matrix below (Figure 9) of the study that was conducted by the community of political sciences (21) for the region of Mid-West Ireland.

		1	2	21	17	5	25	26	19	34	14	37	23	13	3	32	33	22	12	11	35	24	36	10	4	31	7	8	20	6	30	9	28	29	15	16	27	18	38
		s	s	p	l	i	n	t	t	d	t	e	r	l	n	j	d	b	i	i	o	w	i	t	f	c	f	a	c	f	t	b	e	i	l	c	l	l	e
1	shanndev		3	3	3	3	3	3	3	2	3		2	3	3	1	2	3			2	1	2	1	2		1	3	3	2			2	2	3	3	2	3	2
2	serega	3			2		3	1	1	2	3			2			2																						
21	paulp	3	2	2		1	2		2	2				1	2		1	3	2		2	3	2		2										2		2		
17	limebcy	3	2	2		1	2		2	2				2	1			2			2		1		2				1								2		
5	ida	3			1		3	1	2	2	3			2	2		1	2											1										
25	ncn	3	3	1	2	3		3	3		3		1	1	3			1			3	1	3	1	3					3									
26	tleader	3	1	1		1	3		3	2	3		2	1	1			3		1		1		1	1						1							1	
19	tenb	3	1		2	2	3	3		3	3			2	1				2	2	3	1		1	2				1	1						2			
34	dof	2	2		2	2		2	3		3	1		2	3	3	2				2		2						2	2		2			2	2		2	
14	tnthco	3	3			3	3	3	3	3				3	3		3	2			3				3			1							1	1			
37	erm									1							1																						
23	rrltd	2					1	2						1			1	2				2		3	3				2							3			
13	limcoco	3	2	1	2	2	1	1	2	2	3		1		3		3	3	1	1	3	2	2	3	2		1	1	2	1	1	1	1	1	3	3	1	2	
3	nwra	3		2	1	2	3	1	1	3	3			3			2	2	3	2	3		1	3	3		2	2	3	2	3	3	2	3	3	3	2	1	3
32	jconea	1								3							3		1	1	3		1			1	1					1							
33	doe	2	2	1		1				2	3	1	1	3	2	3		2	3	3	3		2	1	2				3	1					3	3			
22	ballyh	3		3	2	2	1	3			2		2	3	2		2			2	3	3		3	2				2	2	1	2		1	2	2		2	
12	ictu			2					2					1	3	1	3												1										
11	ifa							1	2					1	2	1	3	2						2															
35	otherds	2		2	2		3		3	2	3			3	3	3	3	3						2					3										
24	wlimr	1		3			1	1	1				2	2				3						3															
36	ibec	2		2	1		3			2				2	1	1	2							2					3										
10	teagasc	1					1	1	1				3	3	3		1	3		2	2	3	2		3		3		3									3	
4	fas	2		2	2		3	1	2		3		3	2	3		2	2						3					3										
31	ccone															1																							
7	fisheries	1												1	2	1								3															
8	aerrianta	3									1			1	2														1										
20	cenb	3			1	1			1	2			2	2	3		3	2	1		3		3	3	3			1		1		1	1			3		1	
6	forfas	2					3		1	2				1	2		1	2											1										
30	travela							1						1	3			1																					
9	beireann									2				1	3	1		2											1										
28	ecc	2												1	2														1										
29	ihfed	2												1	3			1																					
15	limcico	3		2						2	1			3	3		3	2																					
16	clcoco	3							2	2	1		3	3	3		3	2											3										
27	lcc	2		2	2									1	2																								
18	limebc	3						1		2				2	1			2						3					1										
38	ersi	2													3																								

Fig 9.Adjacency matrix with similarities between political entities of Mid-West Ireland (21)

The possible values of this matrix could be 0 (for no semantic similarity between two political entities) till 4 (perfect similarity between two political entities). As the metrics that were used require either the web page counts or web documents for exacting the similarity degree, the need of a search engine was crucial. Therefore, Yahoo Search Engine was used via Perl module Yahoo Search API. Moreover, for each one of the 37 political terms (the term *other departments* was excluded for implementation's convenience) there are more than one expressions. Due to this fact, the query that was used to the search engine was the one that succeeds good precision. The form of this query is the below stated:

$$\left(\text{“term1}_{\text{exprA}}\text{” OR } \ldots\ldots \text{OR “term1}_{\text{exprN}}\text{”}\right)\text{AND } (\text{“term2}_{\text{exprA}}\text{” OR } \ldots\ldots \text{OR “term2}_{\text{exprN}}\text{”})$$

In the same manner, the good precision ensured with the use of quotes around the words of an expression and a good recall succeeds by the use of all the possible expression for a political entity among "OR" statement. As a result, either the page counts or the URLs of the top ranked pages are returned for every pair. In the case of page-count metrics, whose procedure is simpler than the contextual metrics, we used the page counts returned from the search of each pair, the page counts for the individual political term of the pair and in some occasions the number of documents that are indexed of the search engine and the similarity was exacted. Whereas, in the case of fully-text-based metrics, there were more "steps" till the completion of the experimental process. The top 100 ranked web documents have to be downloaded, filtered by HTML tags and be concatenated into one document per pair. Moreover, all the letters of the documents were transformed into capital letters. This leaded to a more comfortable way of words comparison. Furthermore, the words that consist the terms of political entities were replaced by the unique word "term$_i$". This procedure is presented in Figure 10. The same experiment was executed for the 1000 top ranked web pages. Finally, another improvement could be the extraction of the stop words that the web documents contained. A stop word list was found, and these words were excluded from the web documents.

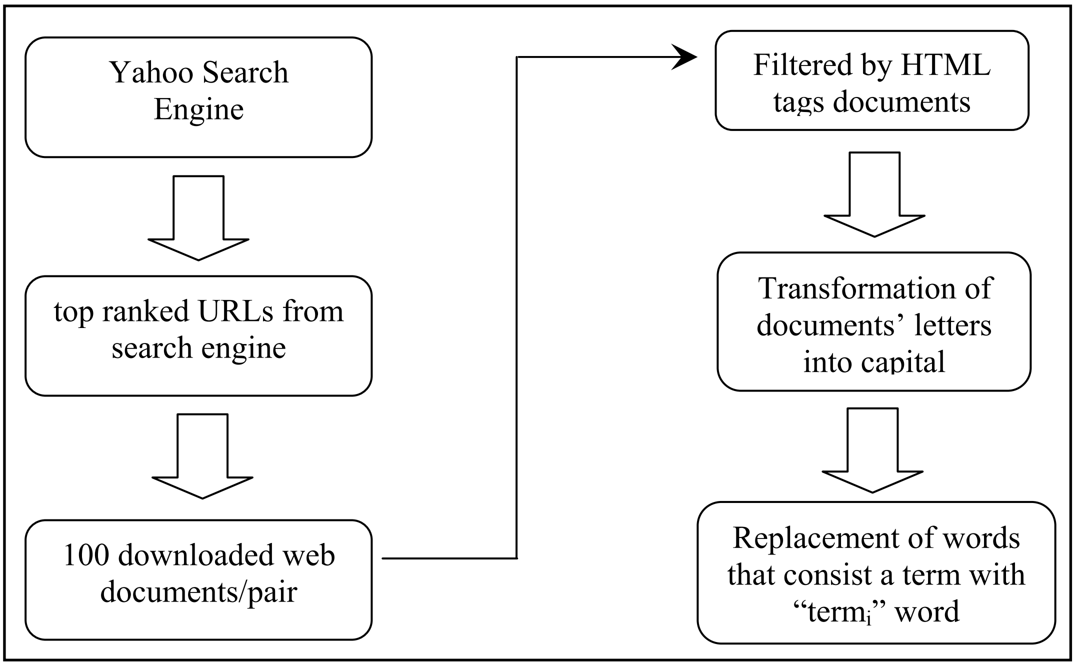

Yahoo Search Engine
top ranked URLs from search engine
100 downloaded web documents/pair
Filtered by HTML tags documents
Transformation of documents' letters into capital
Replacement of words that consist a term with "term$_i$" word

Fig 10.Experimental procedure for creation of final form of web documents for each pair

In the case of Mid-West region of Ireland, the dimension of the adjacency matrix is 37x37 elements. Thus, there are 1369 documents (one document per pair) to be examined. An analysis to these documents was conducted, in order to examine the quality of the final documents. At the end of the analysis, the number of the pairs which decided to be examined was 116. This happened because from the 1369 initial number of pairs, were excluded 703 pairs (as only the upper triangular matrix has to be examined). The upper triangular and the lower triangular part of matrix contain the same information and the elements of the diagonal are equal to 1. Hence, there is no reason to be examined twice. The remaining number of documents is 666. From these documents 349 were excluded, as either there were no web pages for these pairs or at least one of the target terms did not exist in the documents. Furthermore, from the remaining 317 pairs, were excluded 201, as there were not values for them in the matrix of Figure 9, which means that the degree of the relation did not examine by the manual method. The remaining number to the current step of the analysis is 116. To sum up, only 116 pairs of the initial number (666 pairs) is examined by the unsupervised method. The number of separate policy entities concluded in these 116 pairs is 31. This analysis is presented in Figure 11 in order to be easily understood by the readers.

# Documents	Step of Analysis	Reason for exclusion of this # pairs
1369 – 703 = 666	1	Pairs that are in the upper triangular part of Matrix 1
666 – 349 = 317	2	pairs that either there were no web pages or at least one of the target terms did not exist in the documents
317 – 201 = 116	3	Pairs that there is no value in the Matrix 2
116		

# total terms to be examined: 10/21

Fig 11.Number of documents during quality analysis to the documents

Finally, the similarities that produced by the metrics have continuous value range. In order to transform these similarities from continuous range (0,1) to discrete values (0,1,2,3,4) we apply a heuristic method. At start, we assign the sorted array of similarities to array'. In order to divide the elements of array into 3 parts, we assume that the distribution of values 1, 2 and 3 is uniform. Therefore, we need two thresholds. The first threshold has the value of the element in $rank\left[\frac{length(array\prime)}{3}\right]$ of the sorted array' and the second threshold has the value of the element in $rank\left[\frac{2\cdot length(array\prime)}{3}\right]$. This heuristic method is presented to the Figure 12.

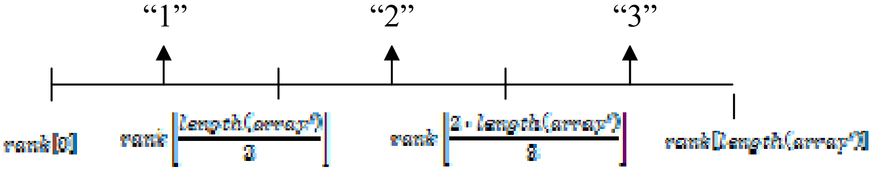


Fig 12.Heuristic of transformation into discrete value range

3.3 Influence in Mid-West Ireland's policy network by time

The significance of the automatic and unsupervised method becomes more obvious by the completion of this experiment. The decrease of the task's time gives the potential of investigating the influence of time to a policy network. The changes during a certain period of time can be outstanding. Moreover, political phenomena are interesting to be examined through time.

Unfortunately, it is an undeniable fact that the majority of the web pages which concern political entities are dynamic. In other words, the web developer has the option of updating the context of the existent site. Thus, it is not an easy procedure to separate the contexts of a web page with time criterion. That was the main reason of conducting this experiment only by using the page-count-metrics.

On the framework of this experiment, two different kinds of queries tried. The first one was in the form "term1" AND "term2" +"/year", for example "Shannon Development" AND "Paul Partnership" +"2004". The punctuation mark consists a simple way of disambiguate the number 2004 from the year 2004. The second query that used to the Yahoo Search Engine was formed as

"term1" AND "term2" +"year1/year2", as it is a common fact that formal documents that concern political entities are referred for academic years.

3.4 Similarities between Greek political entities

It is an undeniable fact, that the interest about political entities is focused essentially to the significant political parties. A simple evidence of this fact is that the majority of statistical researches deal with political parties' scenarios. Moreover, the semantic similarity between political rivals consist an interesting theme. Thus, it seemed really exciting of the possible success of this method in this subject to be investigated.

The experiment estimated the semantic similarity between three kinds of pairs. These pairs have the form of elements in Table 3:

Political party 1 – Political party 2
Politician 1 – Politician 2
Politician – Political party

Table 3.Form of political entities' pairs

The political entities that experimented are presented below in Table 4:

Political Entity	Short Description of this political entity
Konstantinos A. Karamanlis (the junior)	Konstantinos A. Karamanlis is the prime Minister of Greece and the political leader of the Greek political party that is called New Democracy (ND)
George A. Papandreou (the junior)	George A. Papandreou is the Leader of the Opposition in Greece and political leader of the Greek political party called Pan-Hellenic Socialist Movement (PASOK)
Aleka Papariga	Aleka Papariga is the General Secretary of the Communist Party of Greece (KKE)
Alexis Tsipras	Alexis Tsipras is Greek left wing politician and currently the chairman of the Coalition of the Radical Left (SYRIZA) political party
Georgios Karatzaferis	Georgios Karatzaferis is the president of the Popular Orthodox Rally (LAOS), the Greek nationalist/radical right-wing populist party
New Democracy (ND)	New Democracy (ND) is the main center-right political party in Greece whose members form the Greek government
Pan-Hellenic Socialist Movement (PASOK)	The Pan-Hellenic Socialist Movement, better known as PASOK is a Greek centre-left political party
Communist Party of Greece (KKE)	The Communist Party of Greece better known by its acronym, KKE (usually pronounced "koo-koo-eh" or "kappa-kappa-epsilon"), is the communist party of Greece and the oldest party in the Greek political scene
Coalition of the Radical Left (SYRIZA)	The Coalition of the Radical Left, commonly known by its Greek abbreviation ΣΥΡΙΖΑ (SYRIZA), is a coalition of left political parties in Greece
Popular Orthodox Rally (LAOS)	The Popular Orthodox Rally or The People's Orthodox Rally often abbreviated to ΛΑ.Ο.Σ. (LA.O.S.) as a pun on the Greek word for people, is a Greek right-wing populist political party

Table 4.The political entities that experimented automatically with the unsupervised method

To exclude the similarities among the pairs of the above stated entities, both of the metrics were used. Similarly to the previous experiments, Yahoo Search API was used as a search engine and the query that was used had the below stated form:

$$\left(\text{“term1}_{\text{exprA}}\text{” OR } \dots\dots \text{OR “term1}_{\text{exprN}}\text{”}\right)\text{AND (“term2}_{\text{exprA}}\text{” OR } \dots\dots \text{OR “term2}_{\text{exprN}}\text{”).}$$

In this occasion the expressions that inserted in the query were either multiple of ways that the political entities are known in Greek language or multiple of ways that the political entities are known in English language. This fact consists a case in which is seemed the advantage of the language independence that this method provides.

Unfortunately, the results of this method had to be compared with reliable results for the similarities of this pairs. The problem was that these similarities did not exist in the bibliography. Thus, in the framework of this consequential experiment, were created a questionnaire, whose form is stated in the Appendix, and was possible to compare the mean value of the similarities that excluded from the 27 questionnaires compared with the results of the method.

4 Results of Experimental Procedure

The major contribution of this thesis is the opportunity of testing this method with results that produced manually by the sector of political sciences. The criteria that measure the success of the method and the results of the experimental procedure are described below.

4.1 Evaluation Criteria

In order to have an understandable measure from the comparison between the results of the method and the reliable results of each occasion we use three criteria. These are correlation coefficient, mean squared error and mapping using multidimensional scaling. The first two measures give a percentage which shows the level of the match between the results of the reliable method and the results of the unsupervised method (quantitative measure), whereas the quality of the results of the third measure can be evaluated qualitatively.

4.1.1 The Correlation Coefficient

If we have a series of n measurements of random variables X and Y written as x_i and y_i where $i = 1,2, \ldots, n$, then the correlation coefficient can be used to estimate the correlation of X and Y. The analysis of correlation coefficient in the form of summary is also known as the "sample correlation coefficient" (29). The correlation coefficient is then the best estimate of the correlation of X and Y. The sample correlation coefficient is defined in Eq. (4.1.1):

$$r_{xy} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\ \sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}} \qquad \text{Eq. (4.1.1),}$$

where x_i and y_i are the samples of X and Y and n is the number of the samples.

The correlation is defined only if both of the standard deviations are finite and both of them are nonzero. It is a corollary of the Cauchy-Schwarz inequality that the correlation cannot exceed 1 in absolute value.

The correlation is 1 in the case of an increasing linear relationship, −1 in the case of a decreasing linear relationship, and some value between -1 and 1, indicating the degree of linear dependence between the variables. The closer the coefficient is to either −1 or 1, the stronger the correlation between the variables.

If the variables are independent then the correlation is 0, but the converse is not true because the correlation coefficient detects only linear dependencies between two variables. Here is an example: Suppose the random variable X is uniformly distributed on the interval from −1 to 1, and $Y = X^2$. Then Y is completely determined by X, so that X and Y are dependent, but their correlation is zero; they are uncorrelated. However, in the special case when X and Y are jointly normal, uncorrelatedness is equivalent to independence (29).

4.1.2 Mean Squared Error Measure

The mean squared error (MSE) between the samples of random variables X and Y is defined as:

$MSE(x, y) = \mathcal{E}((x_i - y_i)^2)$, where $\mathcal{E}$ is the average value and x_i, y_i are the samples of random variables X and Y, with $i = 1,2, \dots n$.

The mean squared error coefficient is used to compare two models; the unbiased method with the smaller MSE is generally interpreted as the best one. The smallest value that a method can succeed is zero.

4.1.3 Mapping using Multidimensional Scaling

Multidimensional scaling (MDS) is a set of related statistical techniques often used in information visualization for exploring similarities or dissimilarities in data. An MDS algorithm starts with a matrix of item – item similarities and then assigns a location of each item in a low-dimensional space, suitable for graphing. In this thesis the adjacency matrixes of "policy entity" – "policy entity" transformed into a two dimensional map. The concept of this kind of algorithms is to minimize a loss function created by the weights of similarity matrix. The cardinality of the matrix should be between 4 and 20 in order to have satisfactory results. Moreover, there are multidimensional scaling algorithms that refer to metric spaces and others that refer to non-metric spaces. In our case, metric multidimensional algorithms are used to all metrics, except the NGD measure, which is not a metric. The result of the execution of algorithm to the matrix similarity is the transformation of the n-dimensional space of the model (each word of the vocabulary corresponds to one dimension) into a two-dimensional graph, in order to can be easily observed by the human "model". Thus, this useful tool provides the ability of creating a policy network, just with the use of the similarity matrix estimated by the unsupervised method.

Moreover, due to the excluded pairs from our analysis in the first three experiments and the answer to the questionnaires in a subset of the pairs, there are missing values in the adjacency matrixes. Thus, these missing values have to be completed in order to have the graph created. The method that the missing values are completed is the assignment of the mean squared error of the values that exist in the matrix, so as not to affect the result of visualization.

4.2 Results

In this section, the results of each one of the experiments are presented according to the three evaluation criteria. The most important evaluation method considered the correlation coefficient. The multidimensional scaling is a qualitative measure, which comprises the result of this thesis.

4.2.1 Automatic creation of policy network of Southern Aegean, Greece

The creation of the Southern Aegean's policy network implemented with the use of all metrics. Thereafter, the similarity matrixes of each one of the metrics evaluated in comparison with the adjacency matrix of Figure 5. Moreover, the results according to the two quantitative evaluation criteria have been exacted during the 4 steps of analysis, which described in chapter 3.1.

4.2.1.1 Correlation Coefficient Criterion

In this case, not only the correlation coefficient matters, but also the window size (WS) for which is succeeded. It is also obvious the fact that the more the analysis level increases, the greater correlation coefficient is succeeded.

The first metric which is presented is the cosine similarity using the **binary scheme**. Firstly, correlation coefficient is computed among the similarities – elements of the upper triangular sub-matrixes (**step 1** of analysis). Secondly, the pairs that do not have values in the matrix of Figure 1 are excluded (**step 2** of analysis). Unfortunately, the method cannot compute similarity of the pairs that does not exist in web documents. So, these pairs are excluded (**step 3** of analysis) and for another time the correlation coefficient is computed. Finally, the continuous value range of similarities is transformed into the discrete values 1,2 and 3 (**step 4** of analysis). This non-linear transformation succeeds the greater correlation coefficient of all steps of analysis equals to $\mathbf{0,5898}$. The progress of the values of correlation coefficient for each one of the analysis levels is presented in Table 5, whereas the progress of the correlation coefficient for the four steps and all the values of window size is presenting in Figure 13.

metric	*step 1*	*step 2*	*step 3*	*step 4*
binary	0,1680	0,0456	0,5738	**0,5898**
WS	110	3	3	**4**

Table 5.Correlation coefficient for binary scheme for all levels of analysis

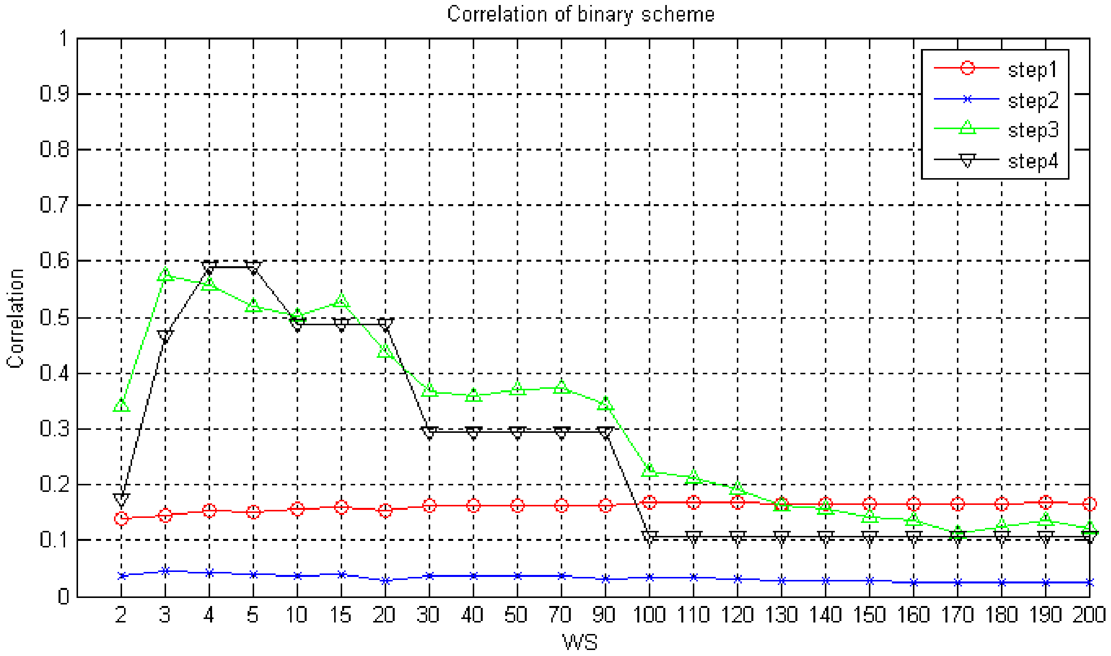


Fig 13.Correlation of all steps for binary scheme

The results for the same steps of analysis, which described previously, are stated this time for the **log frequency scheme**. Therefore, the progress of the values of correlation coefficient for each one of the analysis levels is presented in Table 6, whereas the progress of the correlation coefficient for the four steps and all the values of window size is presenting in Figure 14, with greatest correlation equals to **0, 5898** for the fourth step of analysis.

metric	*step 1*	*step 2*	*step 3*	*step 4*
log(freq)	0,1692	0,0481	0,5744	**0,5898**
WS	100	3	3	**4**

Table 6.Correlation coefficient for log(freq) scheme for all levels of analysis

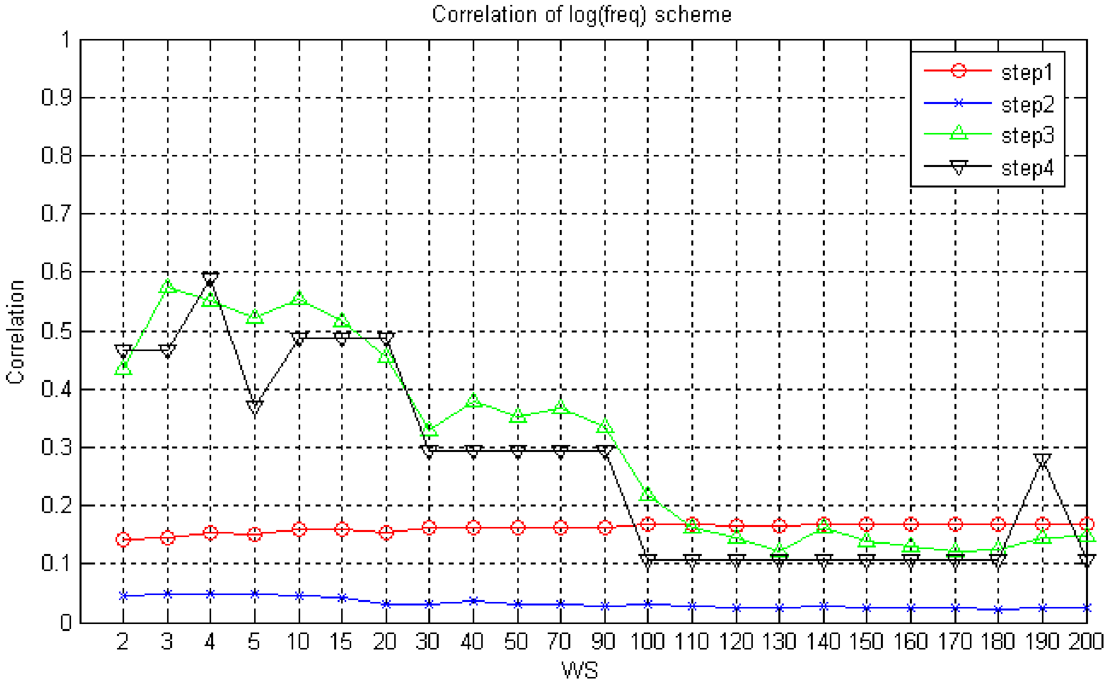


Fig 14.Correlation of all steps for log(freq) scheme

Moreover, one of this thesis' contributions is the testing of a new variance, the normalized logarithm of terms' frequency (Elias Iosif introduced this idea) in the concept of automatic creation of policy networks. The results of this implementation are stated in Table 7, whereas the results of the analysis' steps are shown and in Figure 15, where it is presented the changes of correlation due to the increase of the window size.

metric	*step 1*	*step 2*	*step 3*	*step 4*
norm_log(freq)	0,1694	0,0472	0,5827	**0,5898**
WS	100	3	3	**3**

Table 7.Correlation coefficient for normalized log(freq) scheme for all levels of analysis

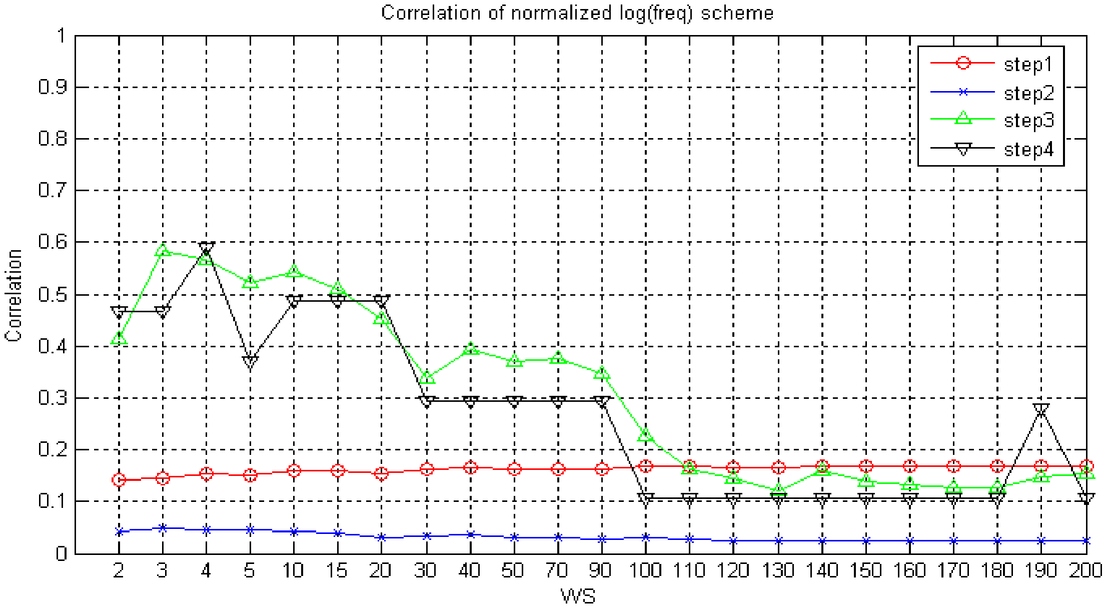


Fig 15.Correlation of all steps for normalized log(freq) scheme

To sum up, the results of the three schemes concerning the contextual metrics are stated in the Table 8 and Figure 16. It is obvious that in step 4 of analysis all the schemes succeed the same correlation coefficient ($\mathbf{0,5898}$) for window size equals to 4. It is assumed that the values of each one of the schemes in step 3 are not similar, but the relative position remains the same. That is the reason why they have equal value of correlation coefficient.

Scheme	*step 1*	*step 2*	*step 3*	*step 4*
binary	0,1680	0,0456	0,5738	**0,5898**
log(freq)	0,1692	0,0481	0,5744	**0,5898**
norm_log(freq)	0,1694	0,0472	**0,5827**	**0,5898**

Table 8.Correlation coefficient for all schemes

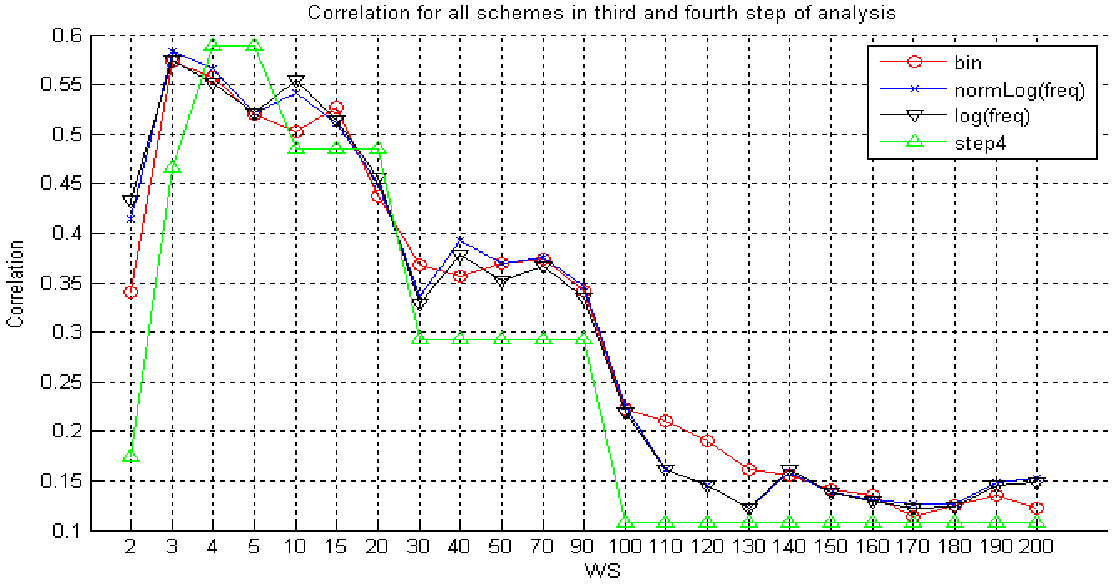


Fig 16.Correlation for third and fourth steps of all schemes

Finally, the Table 9 summarizes the results of correlation from the similarity matrix estimated by the page-count-based metrics.

Scheme	step 1	step 2	step 3	step 4
jaccard	0,1763	0,1739	0,2799	0,4069
dice	0,1823	0,1790	0,2883	**0,4069**
ngd	0,2089	0,2261	0,2261	0,2089

Table 9.Correlation coefficient for all page-count-metrics

There are lots of details to the results of Table 9 that are interesting to be explained. Firstly, as far as ngd metric is concerned the steps 2 and 3 of analysis resulted the same correlation coefficient. This happens because ngd does not result any zeros even if the page count for the pair is zero. Thus, the similarity array of step 2 equals with the similarity array of step3. The ngd metric seems to have worse results than the other metrics, but the cardinality of the similarity array of ngd metric in step 3 is 145, whereas the cardinality of similarity array in step 3 of analysis for jaccard and dice metrics is 40. If we would like to compare these metrics, we could do this in step 1, where there is estimated the correlation between all the similarity matrix resulted by the metrics. Hence, it is obvious that ngd succeeds better correlation coefficient of the other two metrics.

4.2.1.2 Mean Squared Error Criterion

This situation is similar in the case of MSE criterion. It is observed that for the forth step all the schemes have the same efficiency. This may also be for the fact that the relative order of the values do not change from one scheme to the other. It is also clear that in step 4, binary scheme has the minimum MSE ($\mathbf{0,2661}$) for window size equals to 4. The fourth step of analysis results the best (smallest) MSE ($\mathbf{0,1830}$) for window size equals to 10.

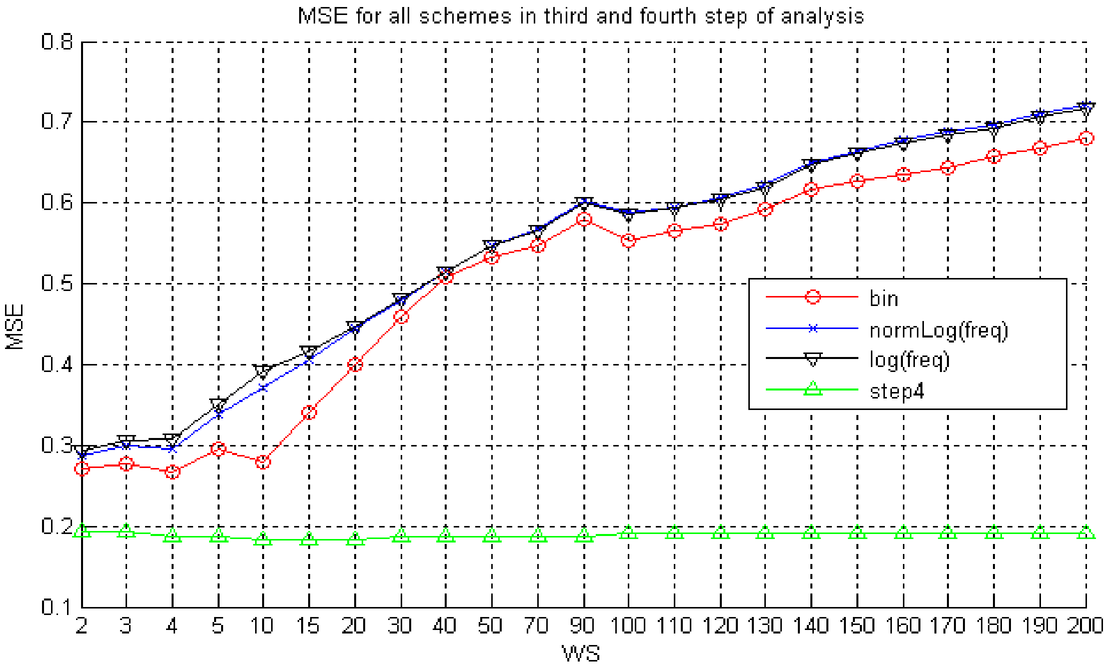


Fig 17.MSE for third and fourth steps of all schemes

Finally, the results of the page-count-based metrics are presented in Table 10.

Scheme	step 1	step 2	step 3	step 4
jaccard	0,00039	0,00057	0,0021	0,2234
dice	0,0011	0,0016	0,0058	0,2234
ngd	0,3173	0,3305	0,3305	0,1974

Table 10.MSE for all page-count-based metrics

The best results according to MSE criterion is the jaccard coefficient, because in step 1 of analysis it has the minimum MSE. The comparison of the three metrics is done in step 1, as the cardinality of the similarity array is the same for all metrics. For another time, ngd metric has the same values in steps 3 and 4 because all the values of pairs are not zero. This means that the similarity array of step 3 and of step 4 is the same.

4.2.1.3 Multidimensional scaling criterion

In this section are presenting the maps that resulted by the adjacency matrixes with the use of the best correlated with the real similarities metrics. Thus, in Figure 18 the results of logarithm term frequency scheme are presented.

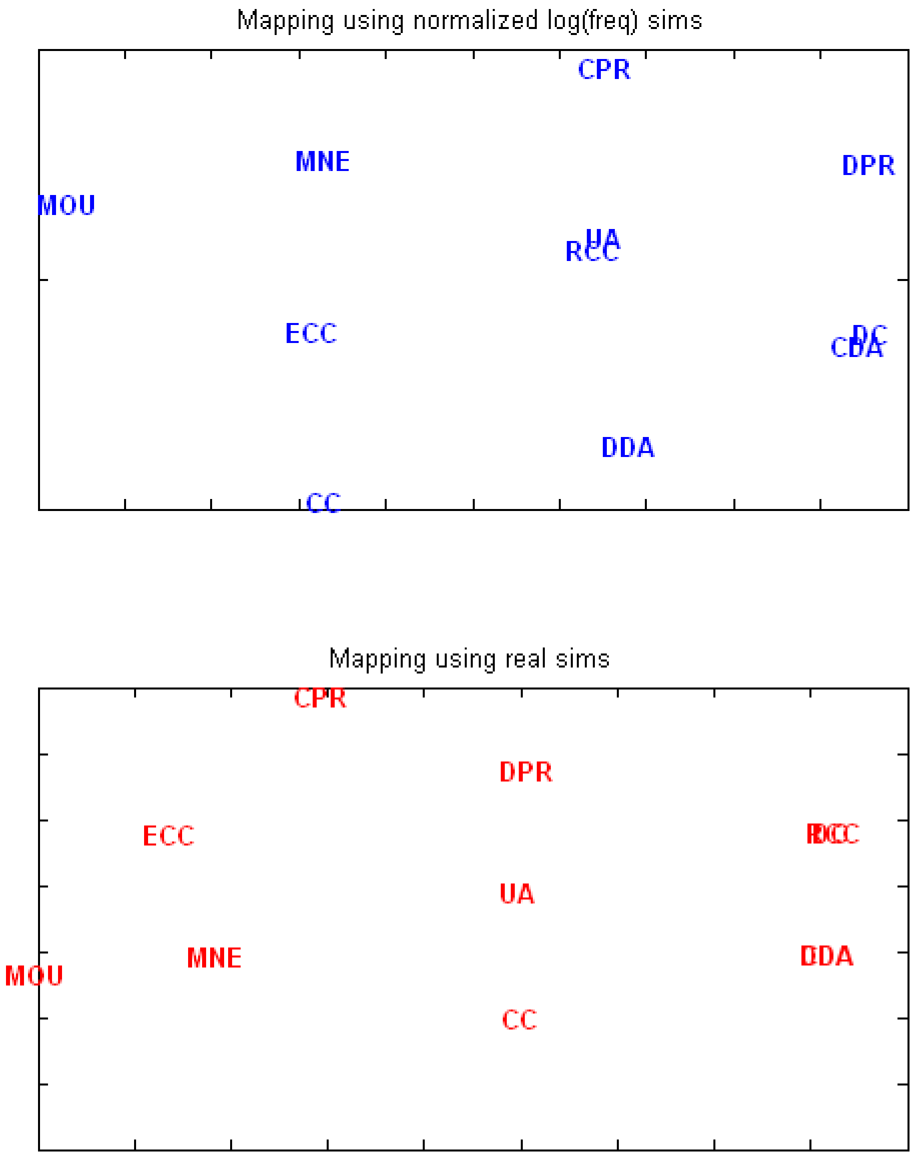


Fig 18.Maps by the normalized log(freq) scheme in contrary with mapping of political sciences survey

The results are really exciting. The mapping of the estimated similarities is very similar to this of the similarities exacted by the manual method. It is noticed in Figure 18 that the term CSF Managing Organization Unit (MOU) is distinguished by the other policy terms with closer to Ermoupolis City Council (ECC) and Ministry of National Economy either in the map of the unsupervised method or in the map of manual method. The meanings of policy terms' abbreviations are presented in the Appendix.

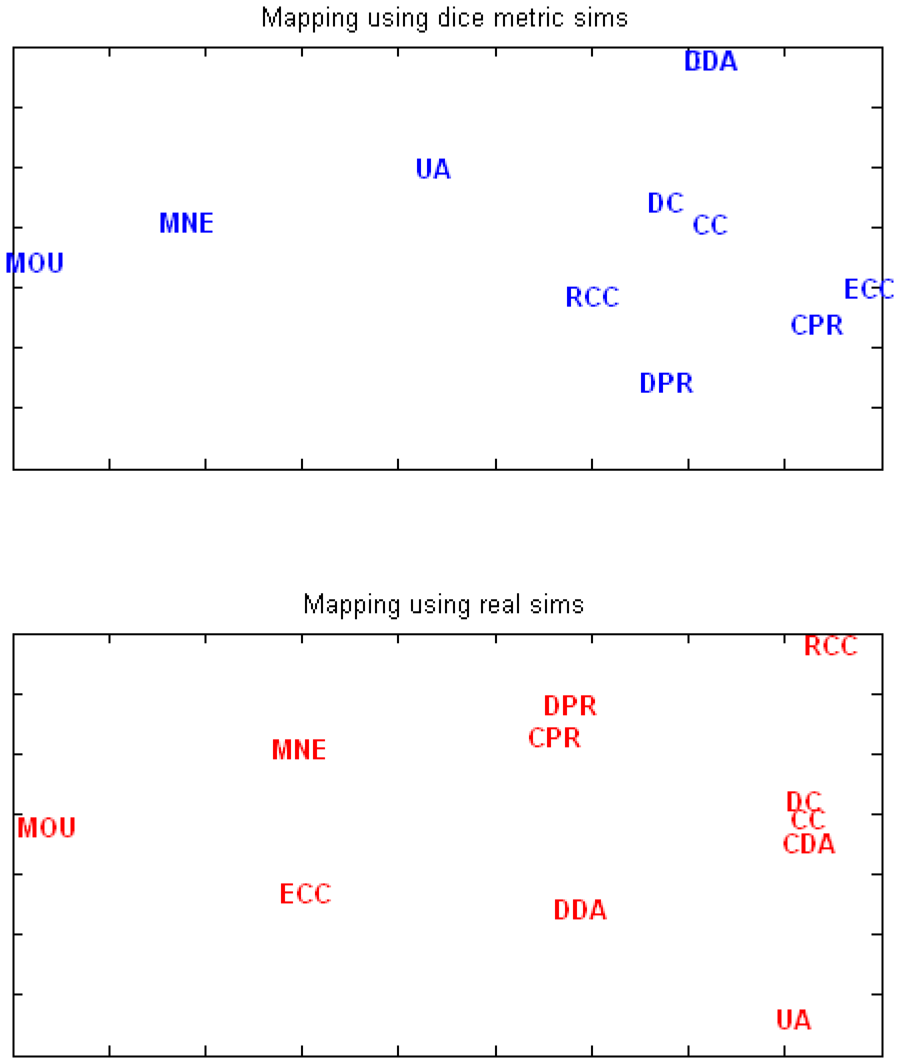


Fig 19.Mapping using similarities of Dice metric

As it is presumed, the result of dice coefficient is not of the same quality that the contextual metric result is. This is a logical result, as the correlation that achieved by the contextual metrics is higher than the page-count-based metrics.

4.2.2 Automatic creation of policy network of Mid-West Ireland, Ireland

The same procedure that described above, re-implemented for the automatic creation of another policy network, the Mid-West Ireland's policy network. The similarity matrix was exacted by the use of all metrics. The correlation of these metrics, the MSE and the multidimensional scaling follows. In this experiment the correlation coefficient and the MSE is computed between the similarity matrix which computed by the political sciences research (Figure 9) and the similarity matrix that computed via the automatic method for each one of the metrics.

Moreover, the results according the two quantitative evaluation criteria have been exacted during the 4 steps of analysis, which described in chapter 3.2. Firstly, the context-based metrics are presented for all steps of analysis and the page-count metrics are following.

4.2.2.1 Correlation Coefficient criterion

The first scheme's results are the binary's scheme correlation coefficient. As it can be seen in Table 11, the best level of correlation (**0, 2222**) is succeeded in fourth step of analysis for window size equals to 100, as it is seemed in Figure 20.

metric	*step 1*	*step 2*	*step 3*	*step 4*
binary	0,2204	0,2071	0,2119	**0,2222**
WS	8	2	40	**100**

Table 11.Correlation coefficient for binary scheme for all levels of analysis

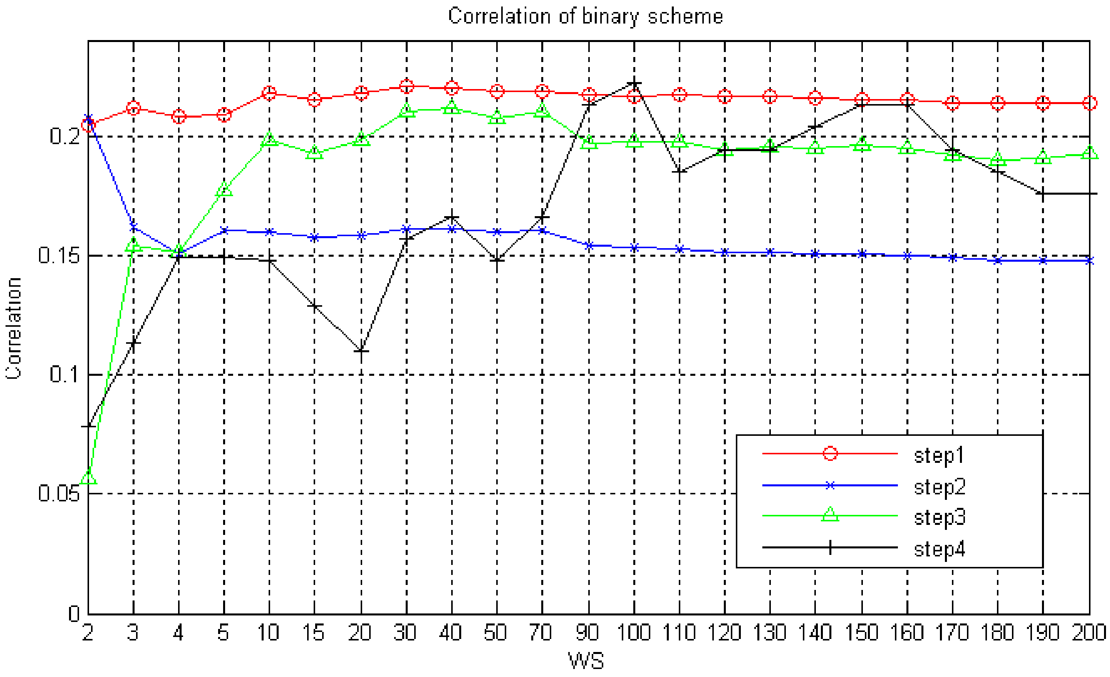


Fig 20.Correlation of all steps for binary scheme

The next scheme that experimented is the **logarithm of term frequency.** Its results for all steps of analysis are shown in the Table 12 and the correlation coefficient for all window sizes is presented in Figure 21. As it is noticed, the maximum correlation coefficient (0,2613) is succeeded in step 3 of analysis for window size equals to 20. It seemed that in the case of implementation of logarithm of term frequency in Mid-West Irish region's policy network, the transformation of similarities from continuous range of values into discrete range of values worsened the correlation result.

metric	step 1	step 2	step 3	step 4
log(freq)	0,2267	0,2136	**0,2613**	0,2502
WS	20	2	**20**	100

Table 12.Correlation coefficient for log(freq) scheme for all levels of analysis

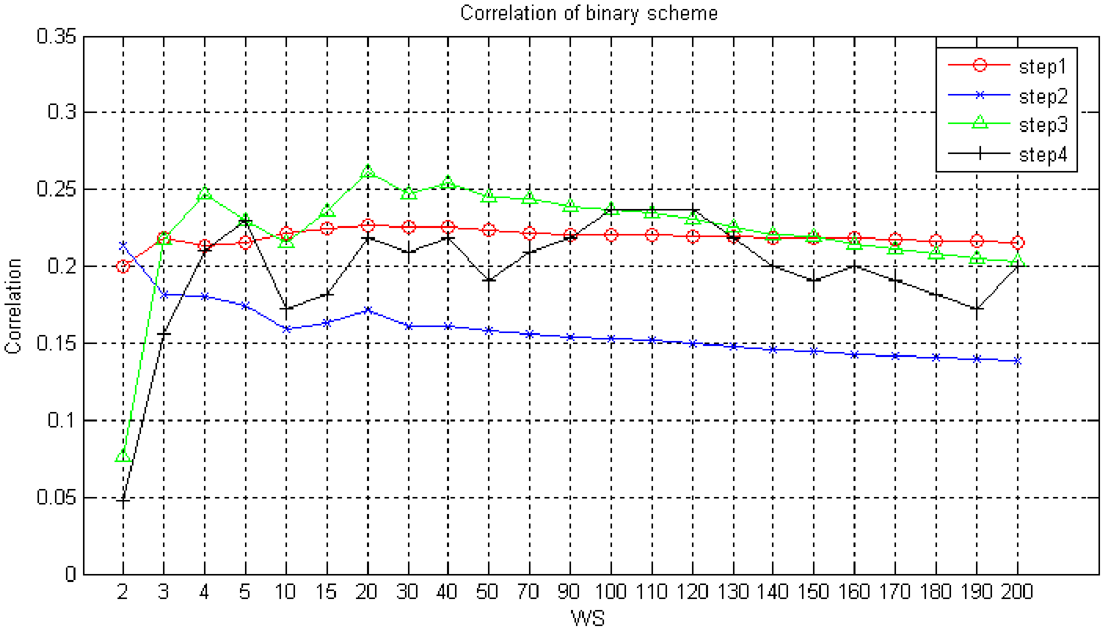


Fig 21.Correlation of all steps for log(freq) scheme

Finally, the correlation coefficient for all windows sizes and all step of analysis is excluded with the use of **normalized logarithm term frequency**. This scheme balances the frequency according to the number that a term appears in a document. The results of this scheme appear in the Table 13, where it can be noticed that the maximum correlation coefficient is succeeded in the third step of analysis for window size equals to 40, as it is obvious if Figure 22 is observed.

metric	step 1	step 2	step 3	step 4
norm_log(freq)	0,2252	0,2067	**0,2498**	0,2316
WS	20	2	**40**	40

Table 13.Correlation coefficient for normalized log(freq) scheme for all levels of analysis

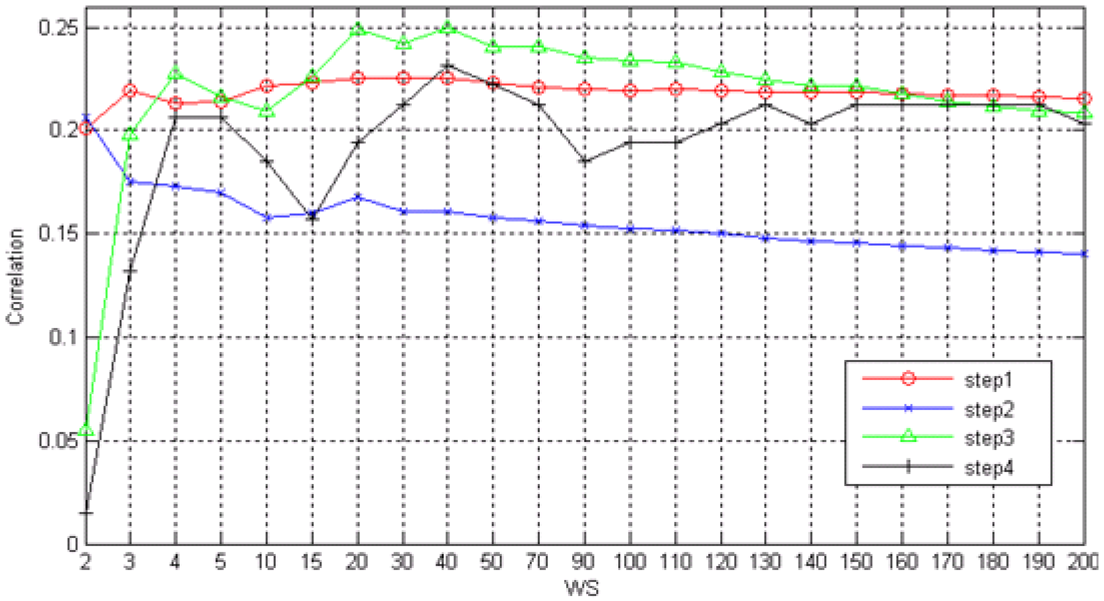


Fig 22.Correlation of all steps for normalized log(freq) scheme

The results concerning correlation coefficient for all the schemes are summarized in Table 14 for comparison reasons.

metric	step 1	step 2	step 3	step 4
binary	0,2204	0,2071	0,2119	0,2222
log(freq)	0,2267	0,2136	**0,2613**	0,2502
norm_log(freq)	0,2252	0,2067	0,2498	0,2316

Table 14.Comparison of correlation results among schemes used in contextual metrics

Thus, it is clear enough that the best results are succeeded with the use of logarithm of term frequency scheme for all steps of analysis, with the best of all (**0, 2613**) for the third step of analysis.

Finally, the MSE results of page-count-based metrics are presented in Table 15.

Scheme	step 1	step 2	step 3	step 4
jaccard	0,1175	0,0426	0,0388	0,0726
dice	0,1233	0,0477	0,0436	0,0726
ngd	0,2696	0,0592	0,0592	**0,0732**

Table 15.Correlation coefficient of all page-count-metrics

There are also lots of details to be mentioned concerning the results of Table 15. Firstly, as far as ngd metric is concerned the step 2 and 3 of analysis resulted the same correlation coefficient. This happens because NGD does not result any zeros even if the page count for the pair is zero. Thus, the similarity array of step 2 equals with the similarity array of step3. The NGD metric seems to have better results than the other metrics, even if the cardinality of the similarity array of ngd metric in step 3 is 226, whereas the cardinality of similarity array in step 3 of analysis for jaccard and dice metrics is 204. In other words, ngd metric is more efficient even if it tries to estimate the similarity for more pairs than the other two metrics do. Hence, it is obvious that NGD succeeds better correlation coefficient of the other two metrics.

4.2.2.2 Means Square Error Criterion

The MSE criterion of the binary scheme, as it can be seen by the Figure 23, inccreases when the window size increases. Moreover, an interesting fact is that in step 4, where we quantize in some way the similarity matrix's values, the MSE remains constant to 0,27 . This happens because even if the values of similarities change, their relative ranks in the sorted similarity array remain the same. The situation is similar for the other schemes, as it can be seen in Figures 24 and 25, which present the logarithm of the term frequency scheme and the normalized logarithm term frequency scheme respectively. Finally, the exact results for every step of analysis for each one of the schemes are stated in Table 16.

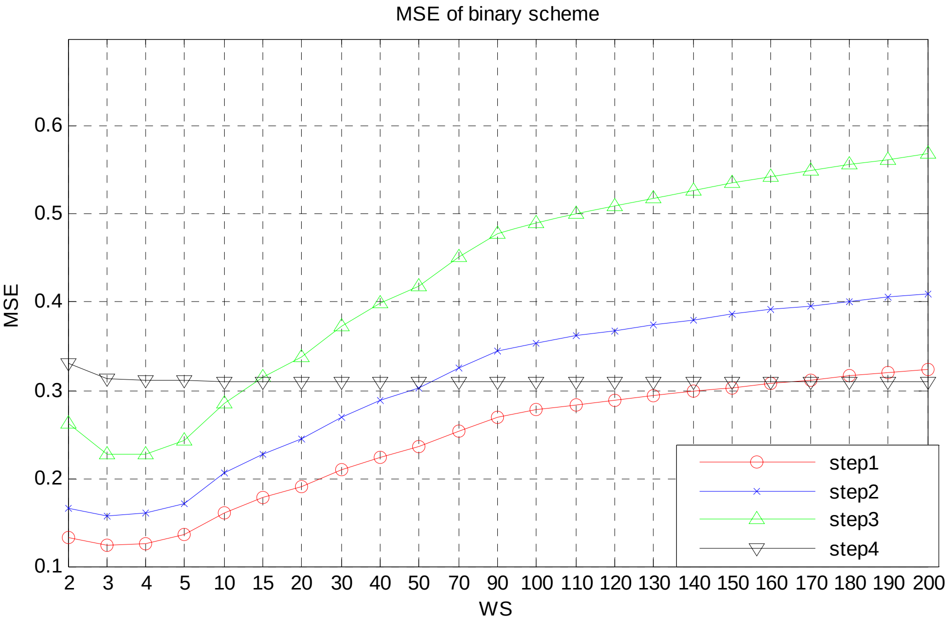


Fig 23.MSE vs. WS for binary scheme

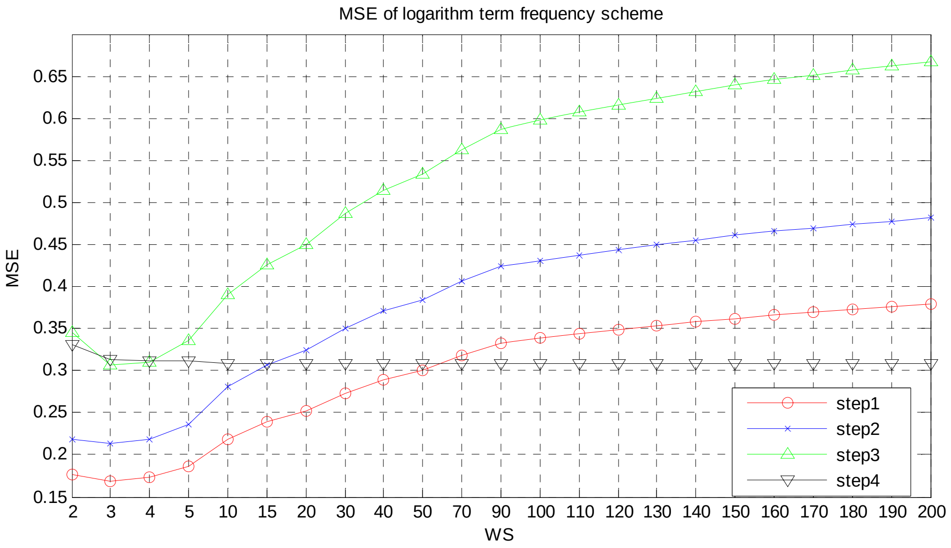


Fig 24.MSE vs. WS for log(freq) scheme

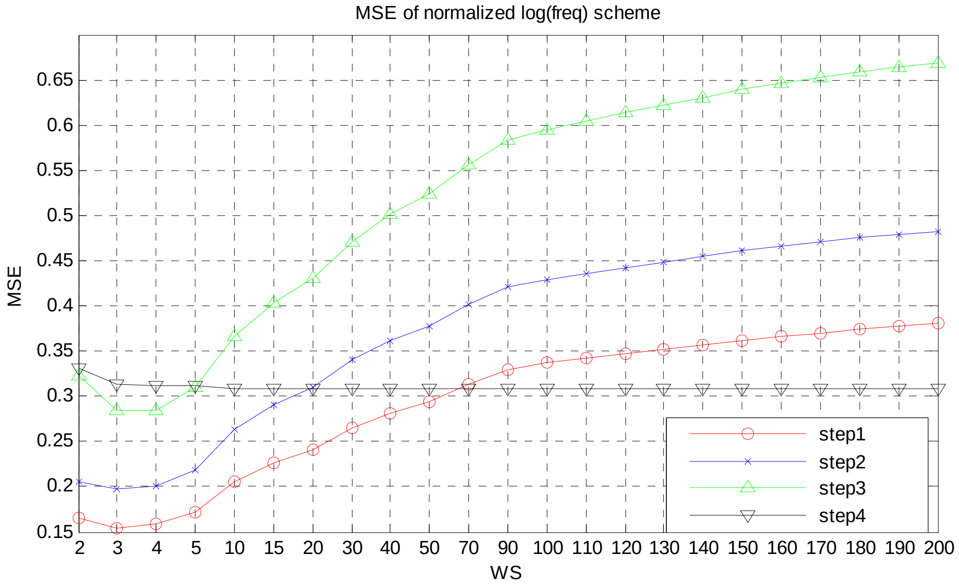


Fig 25.MSE vs. WS for normalized log(freq) scheme

metric	*step 1*	*step 2*	*step 3*	*step 4*
binary	0,1238	0,1575	**0,2267**	0,3090
log(freq)	*0, 1680*	*0, 2134*	*0,3072*	*0,3090*
norm_log(freq)	0,1542	0,1971	0,2837	0,3090

Table 16.MSE for all schemes and steps of analysis

According to the second evaluation criterion, for another time, binary scheme comprise the best scheme, as it succeeds the minimum MSE ($\mathbf{0,2267}$) at step 3 of analysis.

Finally, the results of the *page-count-based* metrics according to MSE criterion are presented in Table 17.

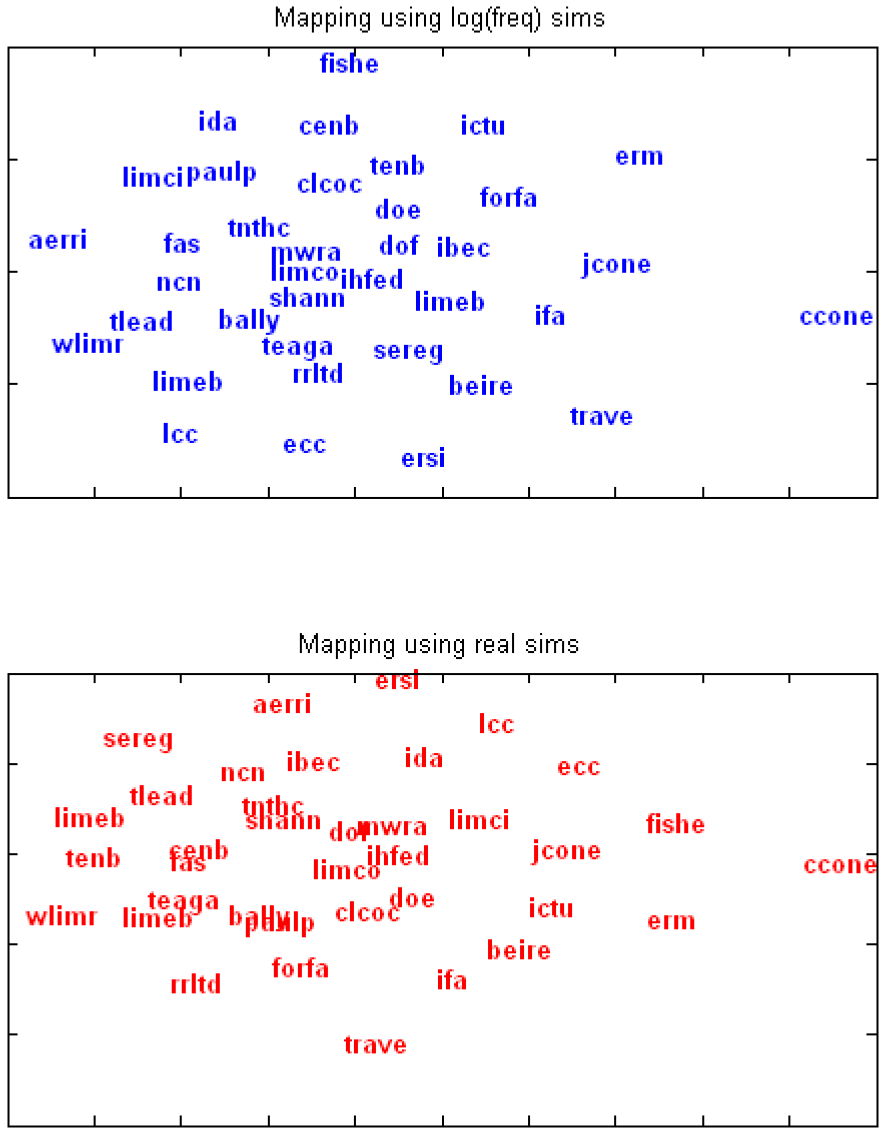
metric	*step 1*	*step 2*	*step 3*	*step 4*
jaccard	0,00016	0,00034	**0,00037**	0,3024
dice	0,0004	0,0010	0,0010	0,3024
ngd	0,1181	0,1502	0,1502	0,2970

Table 17.MSE for all page-count-based metrics

There are lots of details to the results of Table 18 that are interesting to be explained. Firstly, NGD and dice metric resulted for steps 2 and 3 of analysis the same MSE coefficient. This happens because NGD does not result any zeros even if the page count for the pair is zero. Thus, the similarity array of step 2 equals with the similarity array of step 3. The NGD metric seems to have worse results than the other metrics. The minimum MSE is resulted by the jaccard coefficient ($\mathbf{0,00037}$) for step 3 of analysis. The MSE of the fourth step seems really worse than the MSE of the third step. This happens because in step 4 we have discrete values 0, 1, 2, 3 and 4, whereas the values of the three first steps of analysis are between 0 and 1.

4.2.2.3 Mapping using Multidimensional Scaling

In this section the policy network's mapping are presenting for each one of metrics' results. It is observed in the Figures 26 and 27 the results of the best scheme of contextual metrics and the results of the best page-count-based metric. It is also noticed that the fully-text-based metrics are more efficient than page-count-based metrics.



Mapping using log(freq) sims

Mapping using real sims

Fig 26.Maps by the log(freq) scheme in contrary with mapping of political sciences survey

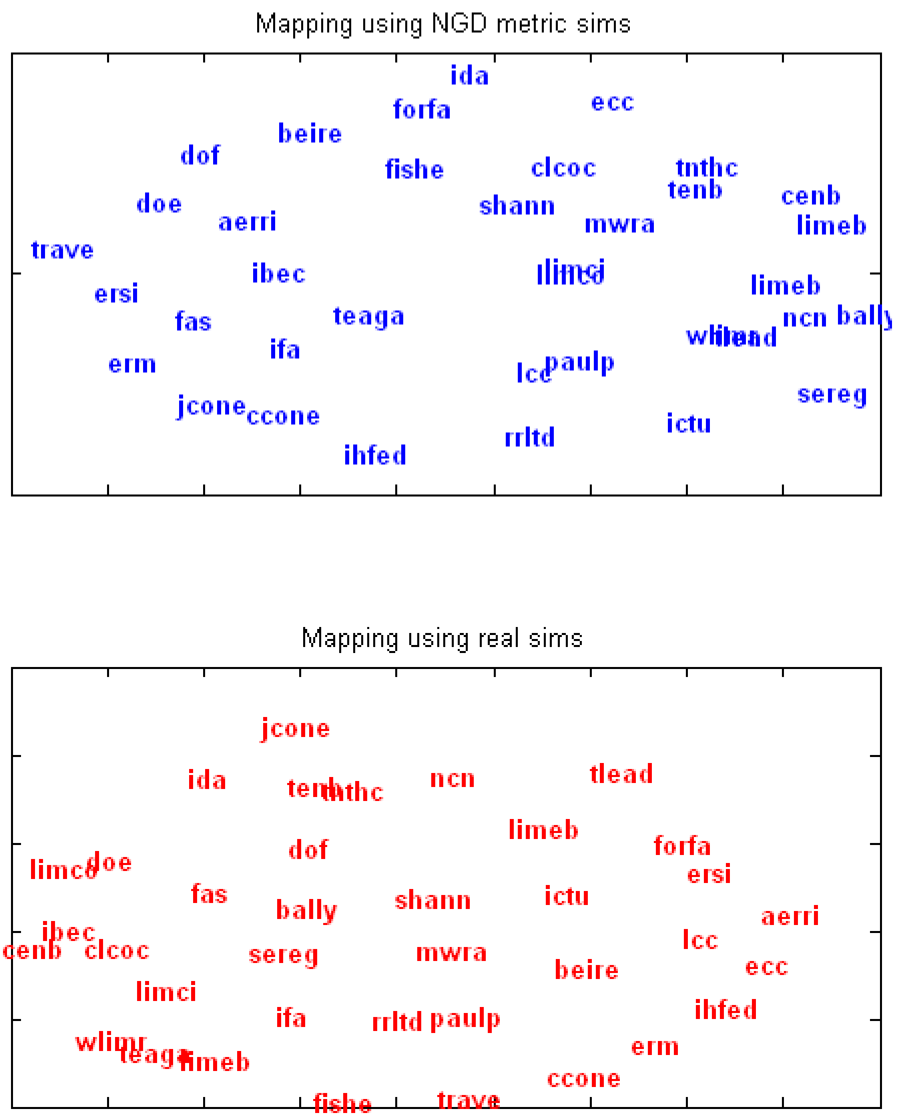


Fig 27.Maps by the NGD metric in contrary with mapping of political sciences survey

4.2.3 Influence in Mid-West Ireland's policy network by time

In this section, the only criterion is the correlation between the similarity matrix which computed by the political sciences sector and the similarity matrix which exacted by the automatic method with the use of page-count-based metrics.

As it was mentioned in experimental description, the experiment conducted with the use of two different queries to the search engine. The results for all page-count-based metrics and for each one of the queries are following figures.

Query type: "term1" AND "term2" +"/year"

Dice coefficient:

The level of correlation coefficient that is succeeded from dice coefficient results, as it is observed in Figure 28, is about $\mathbf{0,11}$. On the other hand, the fact that in years 1990 – 1995 dice coefficient succeeds maximum correlation coefficient is very exciting. It is important to be mentioned that the Second Community Support Framework took place in years 1994 – 1999. So, the research, whose result is the similarity matrix (Figure 5), refers to the years of the Second CSF. The 0,11 degree of succeeded correlation coefficient represents a satisfactory result. On the other hand, a greater correlation coefficient succeeded in comparison with the 0,0726 that is achieved with the use of the kind of query that did not conclude the information of year. It seems that the method adapts up to a point to the related year that the research of the political sciences refers.

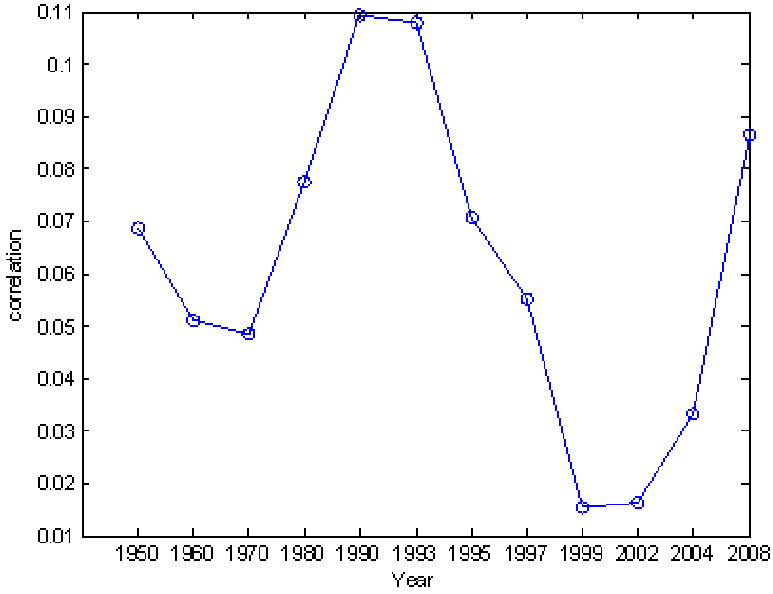


Fig 28.The influence of time to Mid-West Irish policy network by dice coefficient results

Normalized Google Distance:

The NGD “metric”, as it can be seen in Figure 29, gives maximum correlation coefficient at about $\mathbf{0,22}$, but this occurs for years 2008 and not for the years that the research of political sciences sector refers to. This happens because the NGD “metric” depends of the number of indexed by the search engine web pages. In the experiment, this number considered ten billion pages, which is the number of pages that Yahoo indexes nowadays. Thus, it is absolutely normal that the maximum correlation coefficient succeeds for the year 2008.

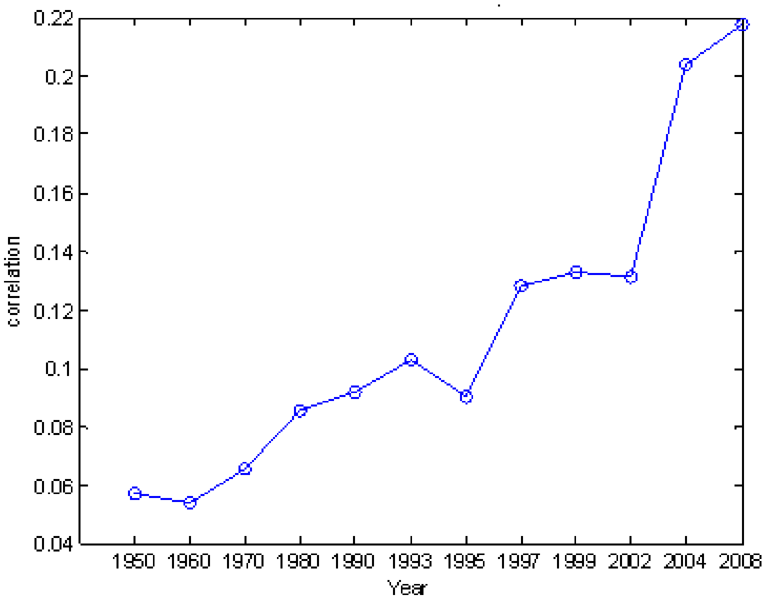


Fig 29.The influence of time to Mid-West Irish policy network by NGD "metric" results

Jaccard Coefficient:

The Jaccard coefficient, as it can be observed in Figure 30, has similar results to dice coefficient. This is quite normal, as the two metrics are by definition very similar. Thus, this similarity metric also succeeds greater correlation coefficient for the years of the Second CSF.

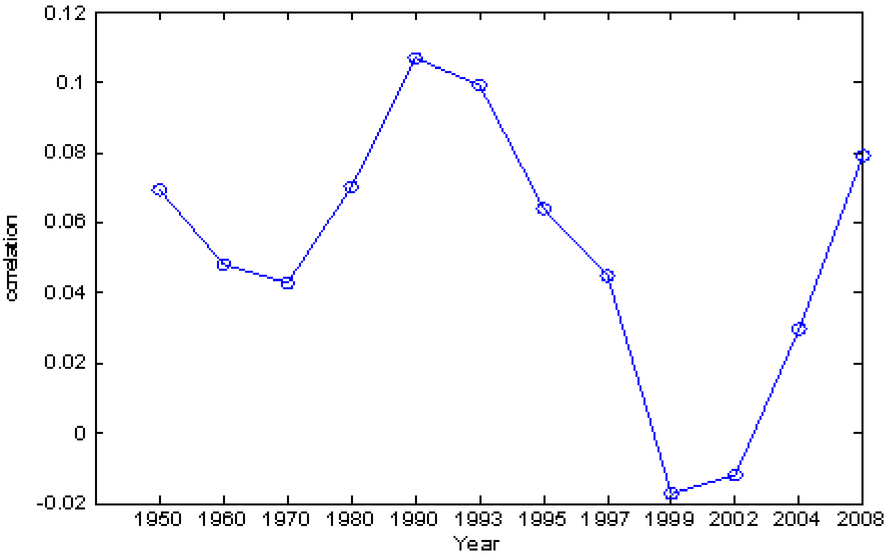


Fig 30.The influence of time to Mid-West Irish policy network by Jaccard coefficient results

Query type: "term1" AND "term2" +"year1/year2"

Dice coefficient:

With the use of the second type of query, as it is observed in Figure 31, the maximum correlation coefficient is about $\mathbf{0,18}$. This occurs for the academic year 2007/2008. For the years that the Second CSF took place, there is also comparable correlation, at about $\mathbf{0,15}$. The fact that in year 2008 the correlation is the greatest may depend to the fact that there are many more web pages nowadays.

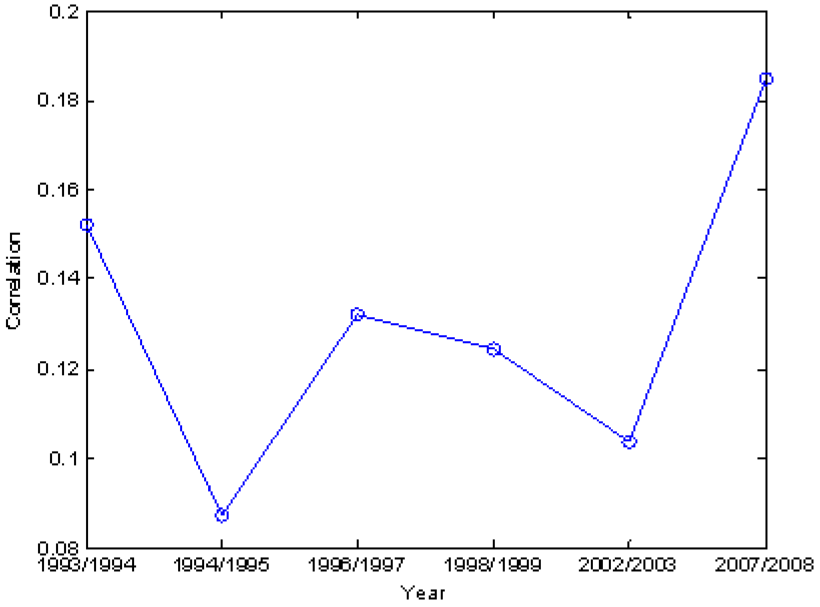


Fig 31.The influence of time to Mid-West Irish policy network by dice coefficient results

Normalized Google Distance:

The NGD "metric", as it can be noticed in Figure 32, gives the maximum correlation coefficient at years 1996/1997. This is quite interesting, as the automatic method seems to react in a different way when the referred years change. Moreover, it is more correlated in the years that the Figure 5 is really referred to, which are the years that the results of Second CSF came up.

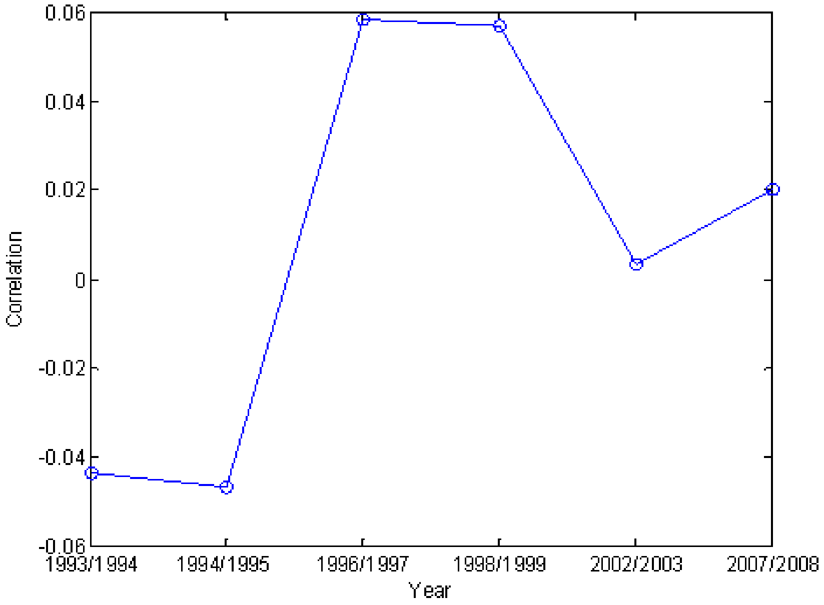


Fig 32.The influence of time to Mid-West Irish policy network by NGD "metric" results

Jaccard Coefficient:

The Jaccard coefficient, as we can observe in Figure 33, has similar results to dice coefficient. This is quite normal, as the two metrics are by definition very similar. The results of this metric are stable for all the years of the experiment.

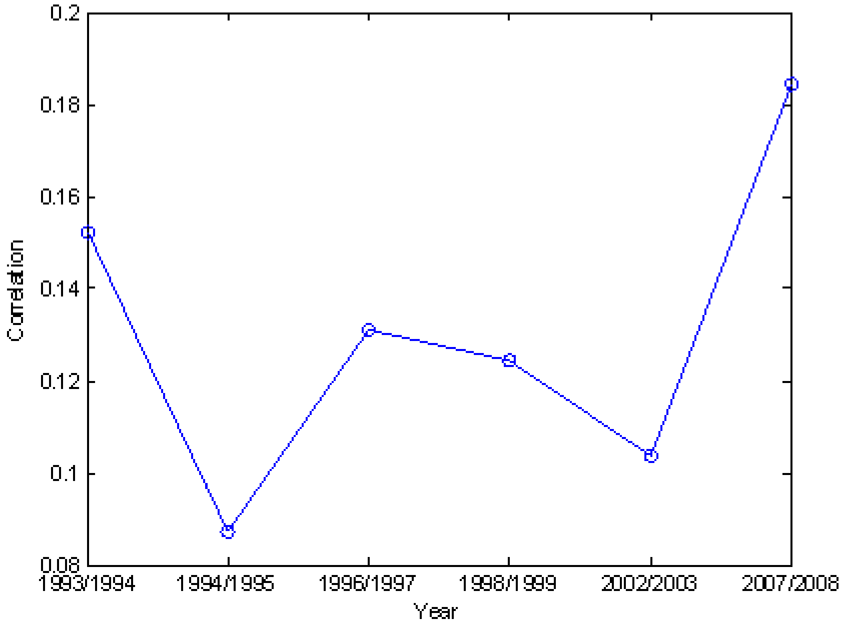


Fig 33.The influence of time to Mid-West Irish policy network by Jaccard coefficient results

To sum up, the change of the referred year to the automatic method seems to provoke changes to the results. This may mean that the automatic method response in a different way, when the parameter year changes. Moreover, the first type of query seems to have better results either to the level of the correlation or to the matching of year reference. Finally, it has to be mentioned that for the years 1950, 1960 and 1970 the levels of the correlation is low. This may depend to the fact that the web resources that refer to these years are limited.

4.2.4 Similarities between Greek political entities

This experiment tries to prove that human model about semantics and the automatic NLP model are correlated in a considerable degree. The correlation coefficient was computed by the comparison between the similarities that exacted by the given answers to the questionnaires of 27 people and the similarities that exacted by the automatic method. The exact questionnaire is presented in the Appendix. The correlations that each one of the methods succeeds are presented in Table 18. The best page-count-based metric is the MI, as it succeeds correlation coefficient equals to $\mathbf{0,2757}$. Moreover, the best results are succeeded by the contextual metrics, especially when binary scheme is used, which result the maximum correlation coefficient $(\mathbf{0,5373})$ and the minimum MSE $(\mathbf{0,0737})$. In literature, the logarithm term frequency scheme succeeds better correlation. In this case, binary scheme is more efficient. This may depend to the fact that the number of web documents was limited and it is known that in such cases binary scheme has better results.

	dice	jaccard	ngd	MI	bin	log(freq)
Correlation	0,2588	0,2898	0,2471	0,2757	0,5373	0,5138
MSE	0,1435	0,1256	0,0871	42,5276[1]	0,0737	0,0786

Table 18.Correlation and MSE between similarities from questionnaire and similarities from all metrics

Moreover, the experiment confirmed the fact, which is known by the literature, that contextual metrics are more efficient model that the page-count-based metrics. Bollegala et al. (1) mentions that page-count-based metrics, despite their simplicity, present several drawbacks, when they are used alone to compute co-occurrence of two words. Firstly, this type of metrics ignores the position of words in the document. Thus, even though two words can exist in a web document but not be related. Secondly, page count of a word with several meanings can contain a combination of all its senses. For instance, George Papandreou is the grandfather of George Papandreou the junior and they are related to the past and present political scene of Greece respectively. Thus, this ambiguity may reduce the efficiency of the metric.

Finally, given the scale and noise in the Web, some words occur randomly in the same page. On the other hand, fully-text-based metrics do not ignore these cases and this is the reason why they succeed better results. Furthermore, the window size that fully-text-based metrics have their best results is valuable information. In Figures 34 and 35 can be seen the changes of MSE and correlation coefficient vs. window size respectively.

[1] The MSE has no meaning as the MI has values to the range (0,∞) while the values of questionnaires are in range (0,1)

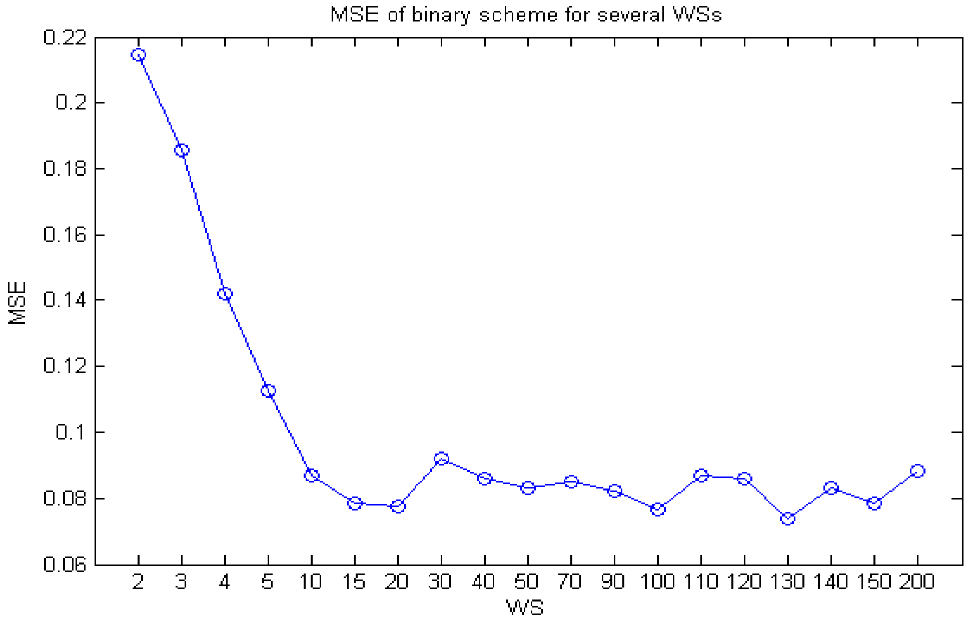


Fig 34.MSE vs. WS for binary scheme

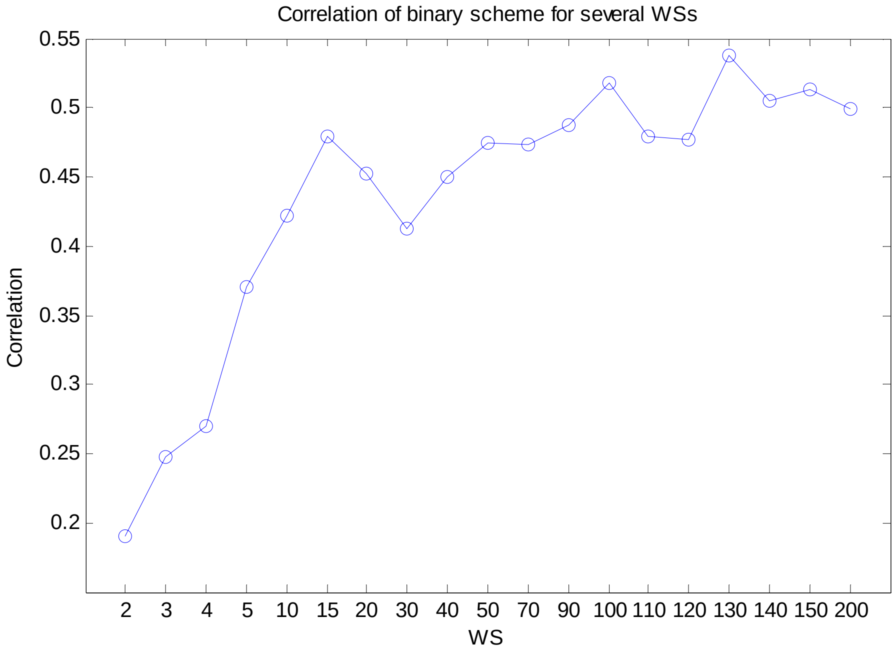


Fig 35.Correlation vs. WS for binary scheme

As it is observed by the Figures 34 and 35, the minimum MSE and the maximum correlation coefficient are succeeded for window size equals to 130. The same results are presented in Figures 36 and 37 for the logarithm term frequency scheme.

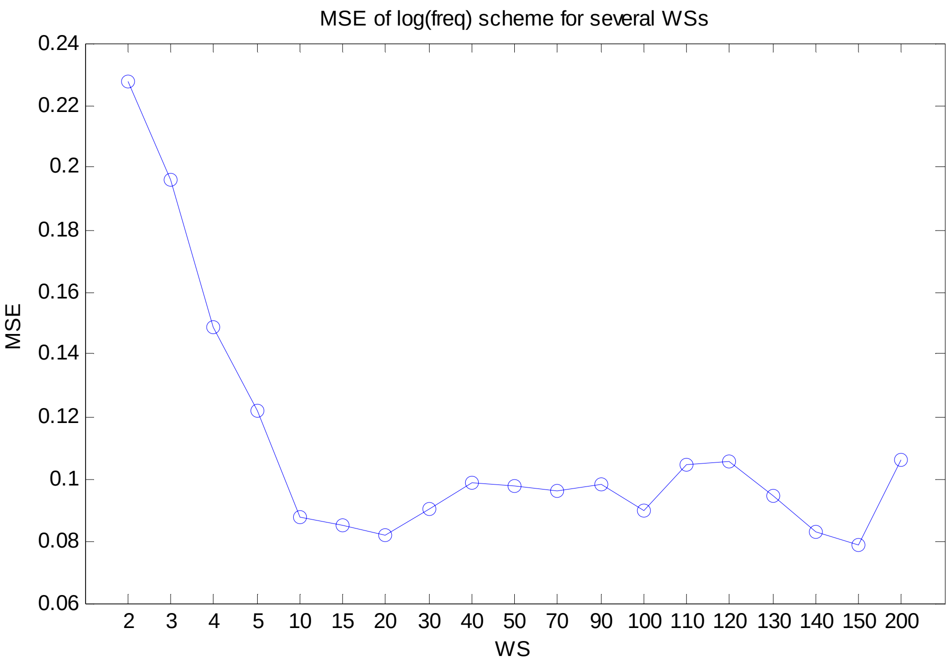


Fig 36.MSE vs. WS for log(freq)scheme

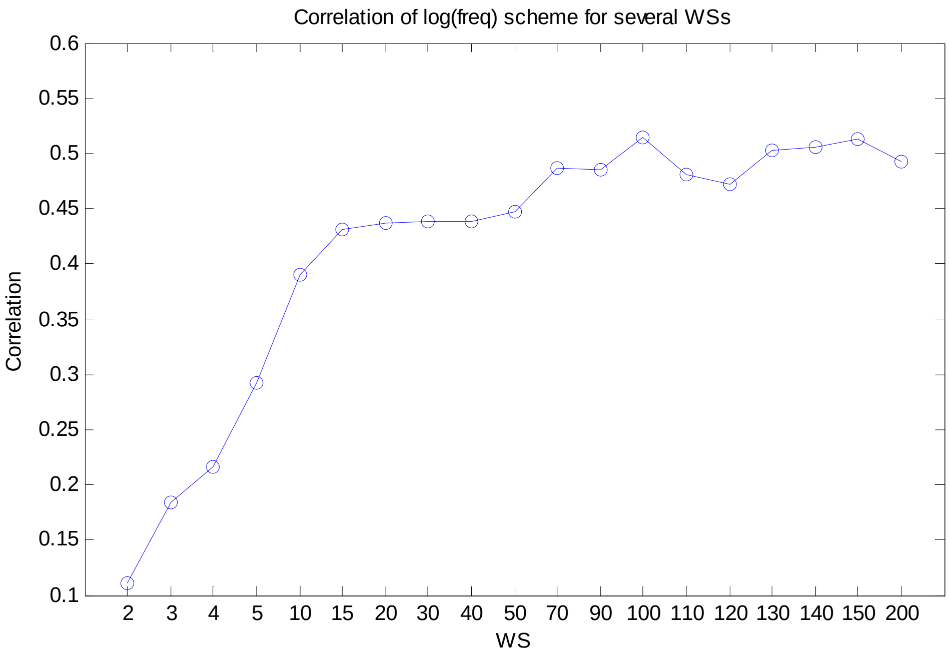


Fig 37.Correlation vs. WS for log(freq)scheme

In this occasion, the logarithm term frequency metric is more efficient for window size 100 according the best correlation coefficient, whereas is more efficient for window size 150 according to MSE criterion. The greater efficient for large window sizes is not a common phenomenon according to the bibliography. Iosif et al. (2) showed that the best correlation results for the measurement of co-occurrence between two nouns are succeeded for small window sizes (2 or 3). This difference may depend to the fact that the syntactic description in political documents defers dramatically from the one in simple nouns. Moreover, there is the assumption that the space that can describe better political entities has more dimensions.

The maximum correlation coefficient that is succeeded by the binary scheme up to 0,5373 is influenced in a great level by the people's opinion that completed the questionnaire. The automatic method has a great ability to simulate the humans' decisions. It has to be mentioned that it is not an obvious procedure to give a similarity in such complex political entity pairs. Thus, people did not consider that the three pairs which are marked with italics in Table 19, do present considerable semantic similarity. In my point of view, the political rivals (opposite political parties) have great semantic similarity. For example, two nouns that have opposite meaning have great similarity. This is the reason why the correlation coefficient is only to 0,5373 . If we exclude these three pairs the correlation of the remaining pairs reaches the **0, 9325**.

Political entity pair	metric sims	human sims
Karamanlis K. – Papandreou G.	0,5906	0,675926
Karamanlis K. – Papariga A.	*0,7716*	*0,185185*
Karamanlis K. – Karatzaferis G.	0,7174	0,648148
Papandreou G. – Karatzaferis G.	0	0,259259
Papariga G. – Karatzaferis G.	*0,8803*	*0,268519*
Tsipras A. – Karatzaferis G.	0	0,212963
ND – PASOK	0,7513	0,740741
ND – KKE	*0,6002*	*0,259259*
ND – LAOS	0,7592	0,694444
PASOK – SYRIZA	0,6245	0,601852
KKE – SYRIZA	0,6002	0,62963
Papandreou G. – PASOK	0,7359	0,888889
Papandreou G. – KKE	0	0,111111
Tsipras A. – SYRIZA	0,7583	0,916667
Karatzaferis G. – ND	0,4417	0,694444

Table 19.Similarities for pairs of questionnaire from automatic method and humans

The mapping of this policy network has created with multidimensional scaling. In this case, the similarity adjacency matrix is sparse, due to the fact that the questions in the questionnaire were only 15, whereas there are 100 similarities (adjacency matrix 10x10) which had to be completed. Thus, the missing values replaced with the MSE of the existed values, so as do not affect the creation of the graph. The results of the page-count-based metrics are stated in Figure 38, whereas the fully-text-based results stated in Figures 39 and 40.

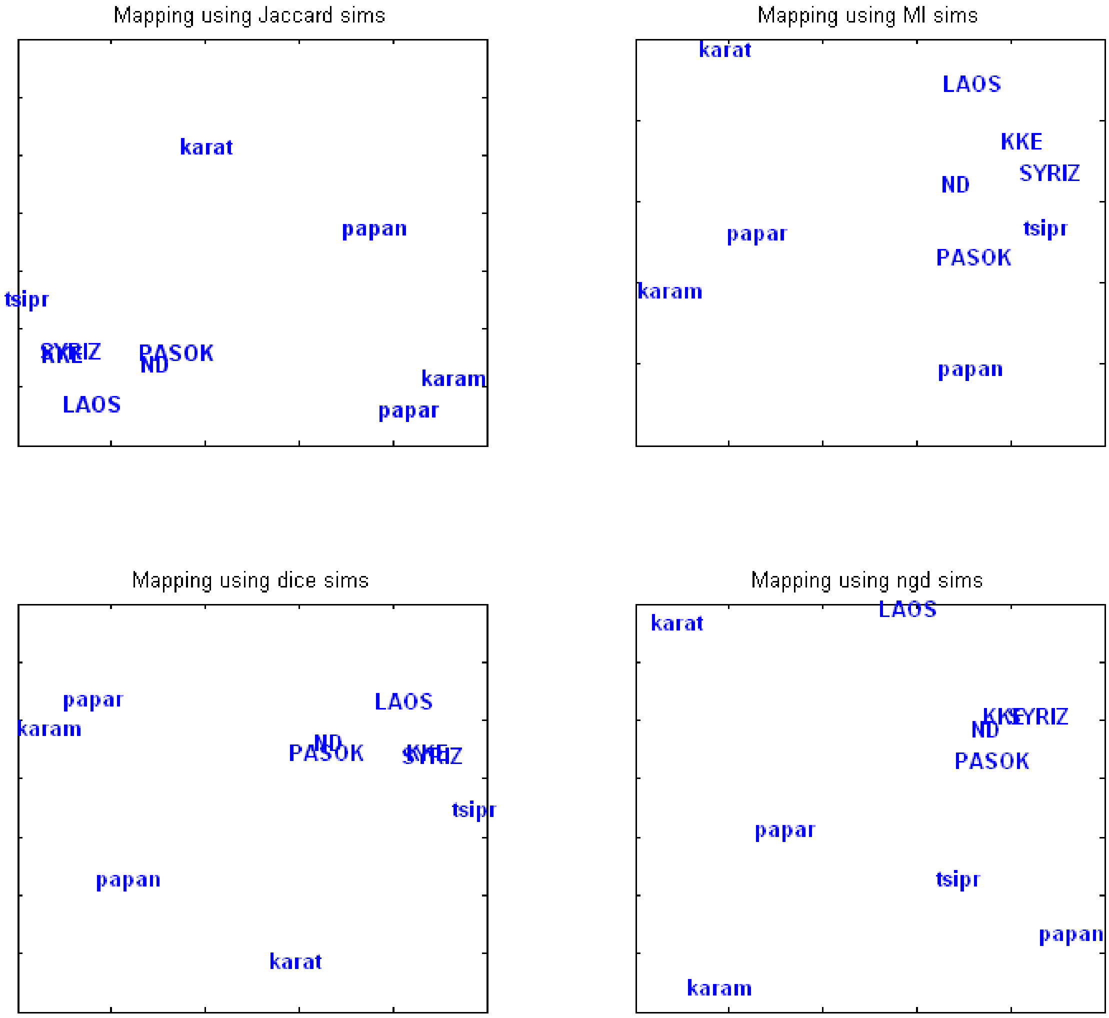


Fig 38.Mapping for all page-count-based metrics

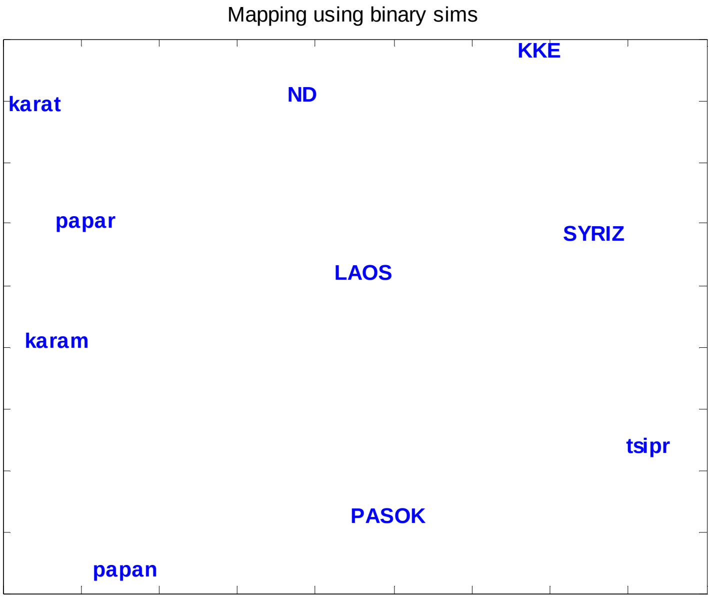


Fig 39.Mapping using binary scheme matrix similarities

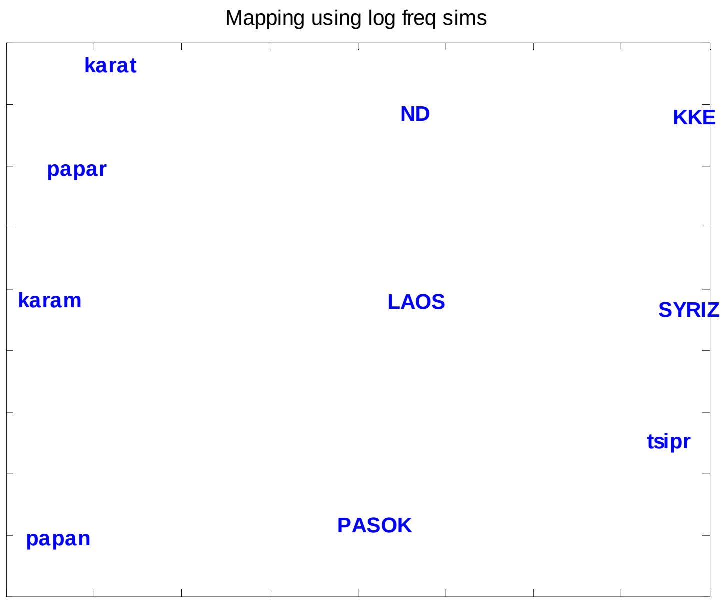


Fig 40 Mapping using log freq scheme matrix similarities

5 Conclusions and future work

5.1 Conclusions

The completion of this thesis raised several conclusions. Firstly, the experiments pointed out that there is potential in developing a fully unsupervised and automatic method for the creation of policy networks. This method has several advantages such as the language independence, independence of experts' knowledge, automatically processing in order to estimate the semantic similarities between political entities. The assumption that similar contexts imply strong semantic relationship was confirmed in the case of political entities up to a point. This may depend to the fact that there were limited web documents in the case of Greek policy network. However, the immense information which is provided by the Web seemed to be exploited in extracting semantic similarity by this unsupervised method.

This work was developed according to the experimental procedure which is described in Chapter 3. The programming language which was used is Perl. The flexibility on processing the natural language of the web documents with the use of regular expressions makes Perl a powerful and useful "tool" for these purposes. The need of a search engine is crucial in order to implement this method. Thus, Yahoo Search Engine used by its Perl module.

Moreover, it is interesting enough that in the implementation of this method in the Greek language, the following phenomenon was often observed; the first capital letter of a Greek name or a word that started a sentence was Latin and the other letters were written in Greek. This happens for example when the writer of a document writes an English word and follows a Greek word whose first letter has to be written in capital. In this case, if the writer forgot to change the language into Greek when he realized the event he would erase all the letters except the first capital letter[2]. The result is a word whose first capital letter is a Latin one, whereas the other letters are written in Greek. This provoked some problems to the development of this method, but it was easily resolved with a Perl script which transformed capital Latin letters into the corresponding capital Greek letters. Additionally, the number of documents for the Greek political entity pairs was restricted. That was the reason why only 16 of the 210 pairs were examined. There is a possibility that the results could be better if was valuable a larger number of web documents for these pairs.

As far as the more efficient metric in the case of Southern Aegean's policy network creation is concerned, the normalized logarithm term frequency scheme, which comprises a variance of cosine similarity, succeeded the best correlation coefficient ($\mathbf{0,5827}$). Furthermore, the transformation of the range of values from continuous $(0,1)$ into discrete values $(0, 1, 2, 3, 4)$

[2] There are 14 capital letters which are common either in Greek or English language. Thus, it is much difficult to be distinguished.

seems to be more efficient, as the correlation coefficient is estimated between two similar approaches. There is ground for improvement as the distribution of the degrees is not really uniform as it was considered. Greater correlation coefficients were achieved with the transposition of the thresholds. Unfortunately, these results cannot be reliable as the distribution is not a known factor and it is not constant for all cases. Moreover, the mapping was developed for the best method and its results reached a satisfactory level.

In the case of the Irish policy network creation, the logarithm of term frequency scheme succeeded the maximum value for the correlation coefficient criterion $(0,2613)$. The quantization of the continuous range of values did not improve the efficiency of method.

Moreover, as it is referred in the literature, although page-count-metrics are a simpler method, they present lots of drawbacks comparing to fully-text-based metrics. This fact is confirmed by this thesis, as for all the sets of experiments page-count-based metrics are less efficient according to the correlation coefficient evaluation criterion.

Another important conclusion is the study about the influence of time to the creation of the policy network of Mid-West Ireland. It is found that the best correlation was achieved for the years that the second CSF was in progress. This means that the metrics respond to the change of time with efficient results comparing to other years.

Finally, the most impressing results concern the creation of the network among the most famous Greek political parties and their leaders. The correlation that is succeeded in this case is $0,5373$. In my opinion, this result is plasmatic and the method responds great to the answers of humans. Unfortunately, the answers that were given for pairs among the communist party and two right political parties were misunderstood by the people who completed the questionnaire. Hence, their answer was according to the political believes of these parties and not according to the semantic distance that was asked. This influenced the evaluation of the method, as political rivals have strong semantic relationship.

5.2 Further Improvements

There are several additions that could be done in order to improve the method’s efficiency. First of all, another search engine, for example Google, that could be used or the combination of them, in order to increase the restricted number of documents that there were in our experiments.

Moreover, the extraction of the stop words could be done not by an already developed stop word list, whose content is general, but with the extraction of the most frequent words in our downloaded documents. This could increase the quality of the Web.

Another possible improvement, which may boost the dynamic of this method, could be the use of a stemmer. Especially in Greek language, where the forms of a word are so much, the use of a stemmer could increase the performance, as for example the model could realize that word politician have the same semantic with word politicians. At the moment, the model of the developed method does not have this ability.

The development of a metric that could be a better estimator in the case of policy network extraction could comprise a theme for extended research. This could be the subject of following research to the domain of policy networks' creation. For instance, a metric that considers how close to the target word is a context word and separates the weight coefficient according to this position could be experimented in a future work.

Another improvement could be the implementation of heuristics that could smartly exploit the specific parts of a web page and give a bigger weight coefficient to them. For example, the title of the web document or words that appear in bold or italics should have greater semantic meaning, as they are highlighted by the writer.

Finally, the way that missing values fact was resolved from our approach could be improved. Several methods could be experimented where the qualitative measure of mapping have more sensible results.

Appendix

A) The questionnaire:

Ερωτηματολόγιο

Το ερωτηματολόγιο αυτό δημιουργήθηκε στα πλαίσια της διπλωματικής εργασίας με θέμα:

«Αυτόματη δημιουργία γράφων πολιτικών φορέων με την χρήση εγγράφων από το Διαδίκτυο»

Οδηγίες για την συμπλήρωση του ερωτηματολόγιου:

Στο συγκεκριμένο ερωτηματολόγιο καλείστε να βαθμολογήσετε την ομοιότητα ως προς πραγματική πολιτική στάση, πιθανές πολιτικές συμμαχίες κτλ ανάμεσα σε ζεύγη πολιτικών οντοτήτων (κόμματα, πολιτικοί κτλ).
Παρακαλώ διαβάστε μία φορά όλα τα ζεύγη κάθε κατηγορίας και στην συνέχεια βαθμολογήστε την σημασιολογική ομοιότητα με βαθμούς από μηδέν (0 - για μικρή ομοιότητα) εώς και τέσσερα (4 – μεγάλη ομοιότητα).

Κατηγορία 1. Πολιτικός1 – Πολιτικός2

- Κωνσταντίνος Καραμανλής – Γεώργιος Παπανδρέου

○	○	○	○	○
0	**1**	**2**	**3**	**4**

- Κωνσταντίνος Καραμανλής – Αλέκα Παπαρήγα

○	○	○	○	○
0	**1**	**2**	**3**	**4**

- Κωνσταντίνος Καραμανλής – Γεώργιος Καρατζαφέρης

○	○	○	○	○
0	**1**	**2**	**3**	**4**

- Γεώργιος Παπανδρέου – Γεώργιος Καρατζαφέρης

○	○	○	○	○
0	**1**	**2**	**3**	**4**

- Αλέκα Παπαρήγα – Γεώργιος Καρατζαφέρης

○	○	○	○	○
0	**1**	**2**	**3**	**4**

- Αλέξης Τσίπρας – Γεώργιος Καρατζαφέρης

○	○	○	○	○
0	**1**	**2**	**3**	**4**

Κατηγορία 2. Πολιτικό κόμμα1 – Πολιτικό κόμμα2

- Ν.Δ – ΠΑ.ΣΟ.Κ

○	○	○	○	○
0	**1**	**2**	**3**	**4**

- Ν.Δ – Κ.Κ.Ε

○	○	○	○	○
0	**1**	**2**	**3**	**4**

- Ν.Δ – ΛΑ.Ο.Σ

○	○	○	○	○
0	**1**	**2**	**3**	**4**

- ΠΑΣΟΚ – ΣΥ.ΡΙΖ.Α

○	○	○	○	○
0	**1**	**2**	**3**	**4**

- Κ.Κ.Ε – ΣΥ.ΡΙΖ.Α

○	○	○	○	○
0	**1**	**2**	**3**	**4**

Κατηγορία 3. Πολιτικός – Πολιτικό κόμμα

- Γεώργιος Παπανδρέου – ΠΑ.ΣΟ.Κ

○	○	○	○	○
0	**1**	**2**	**3**	**4**

- Γεώργιος Παπανδρέου – Κ.Κ.Ε

○	○	○	○	○
0	**1**	**2**	**3**	**4**

- Αλέξης Τσίπρας – ΣΥ.ΡΙΖ.Α

○	○	○	○	○
0	**1**	**2**	**3**	**4**

- Γεώργιος Καρατζαφέρης – Ν.Δ

○	○	○	○	○
0	**1**	**2**	**3**	**4**

B) Figures

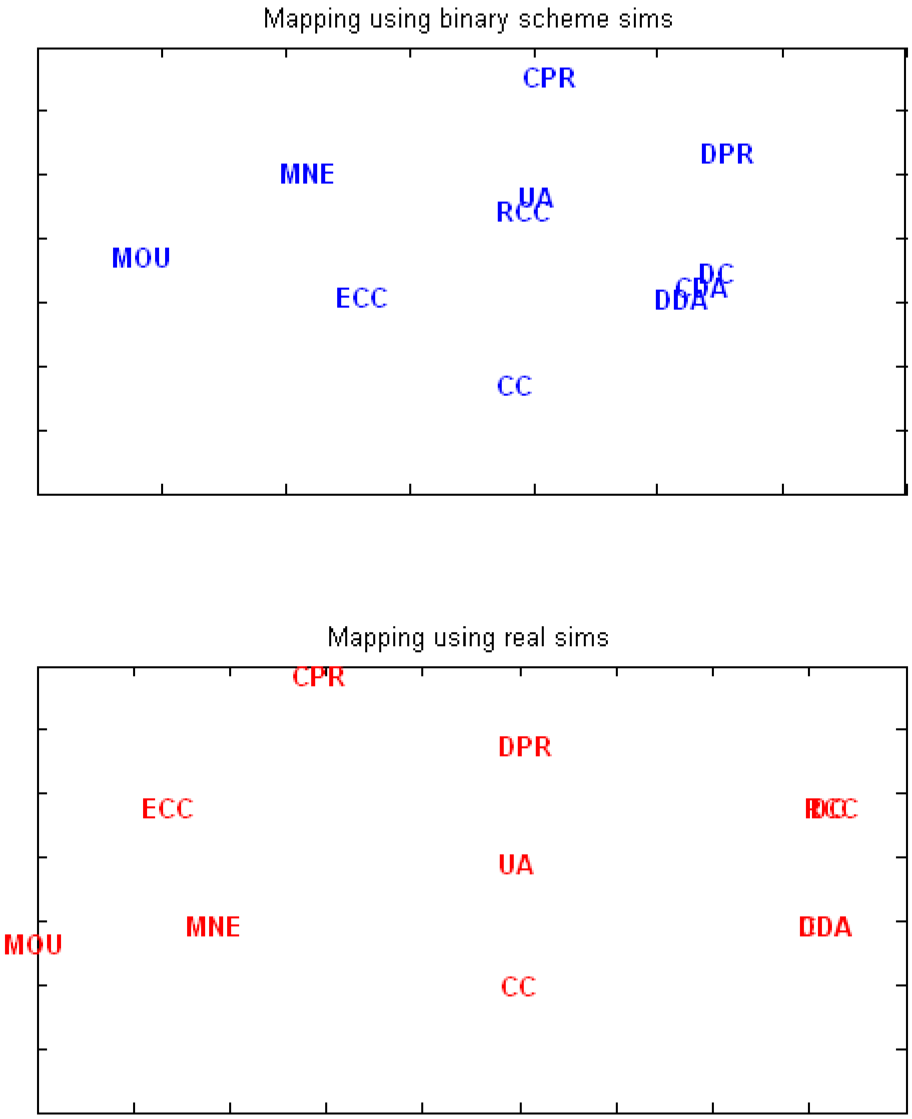


Fig 41.Maps by the binary scheme in contrary with mapping of Greek political sciences survey

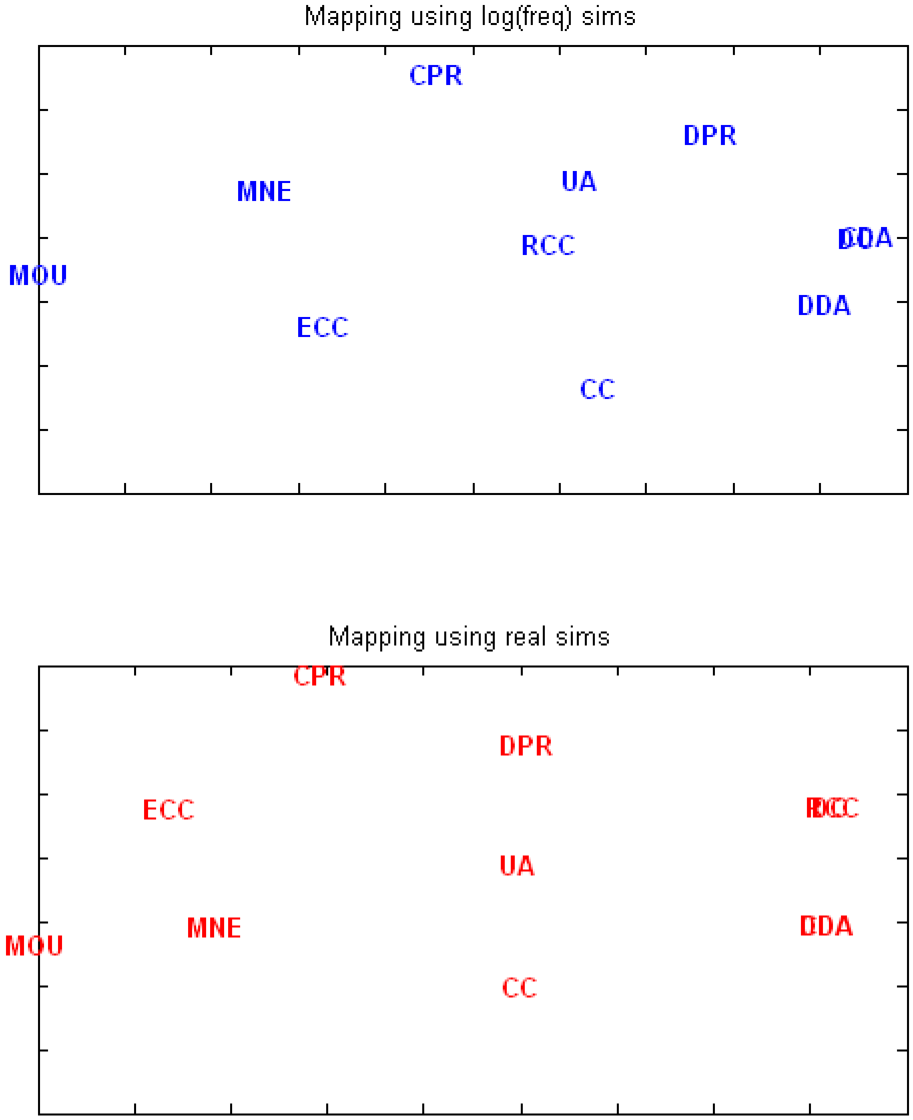
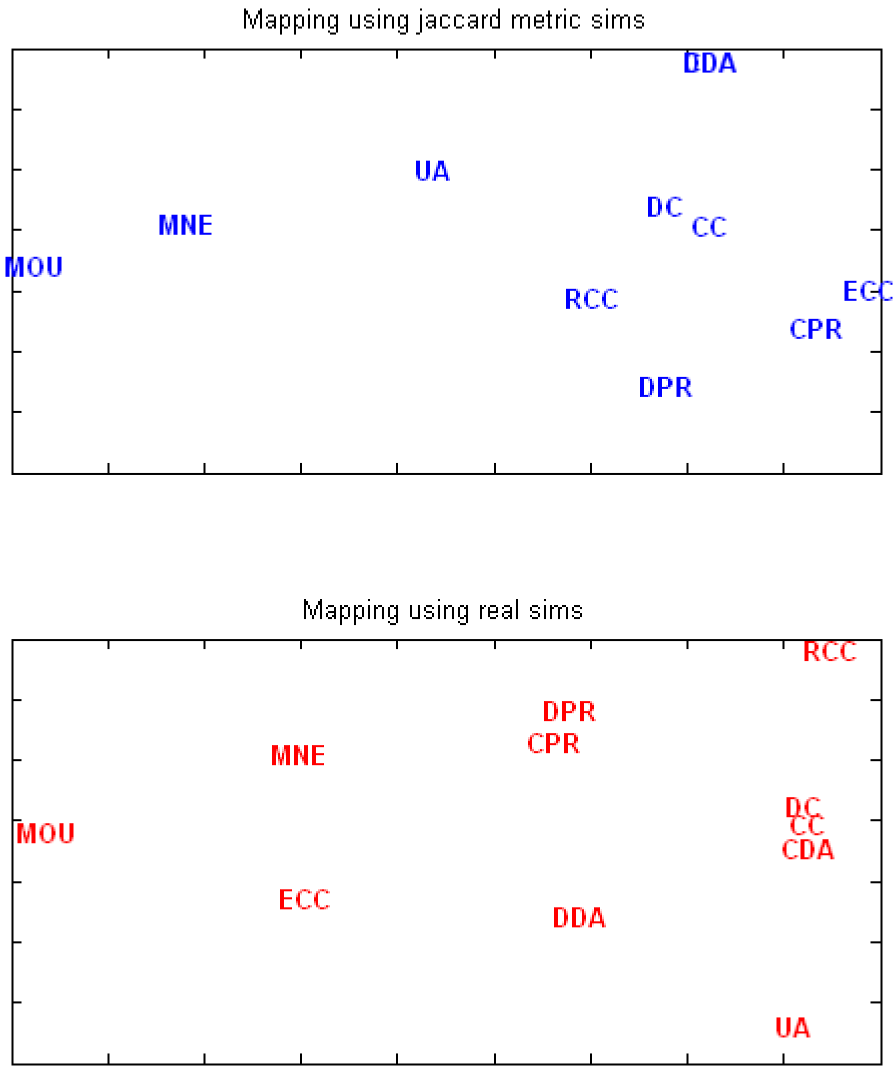


Fig 42.Maps by the log(freq) scheme in contrary with mapping of Greek political sciences survey


Mapping using jaccard metric sims

Mapping using real sims

Fig 43.Maps by the jaccard metric in contrary with mapping of Greek political sciences survey

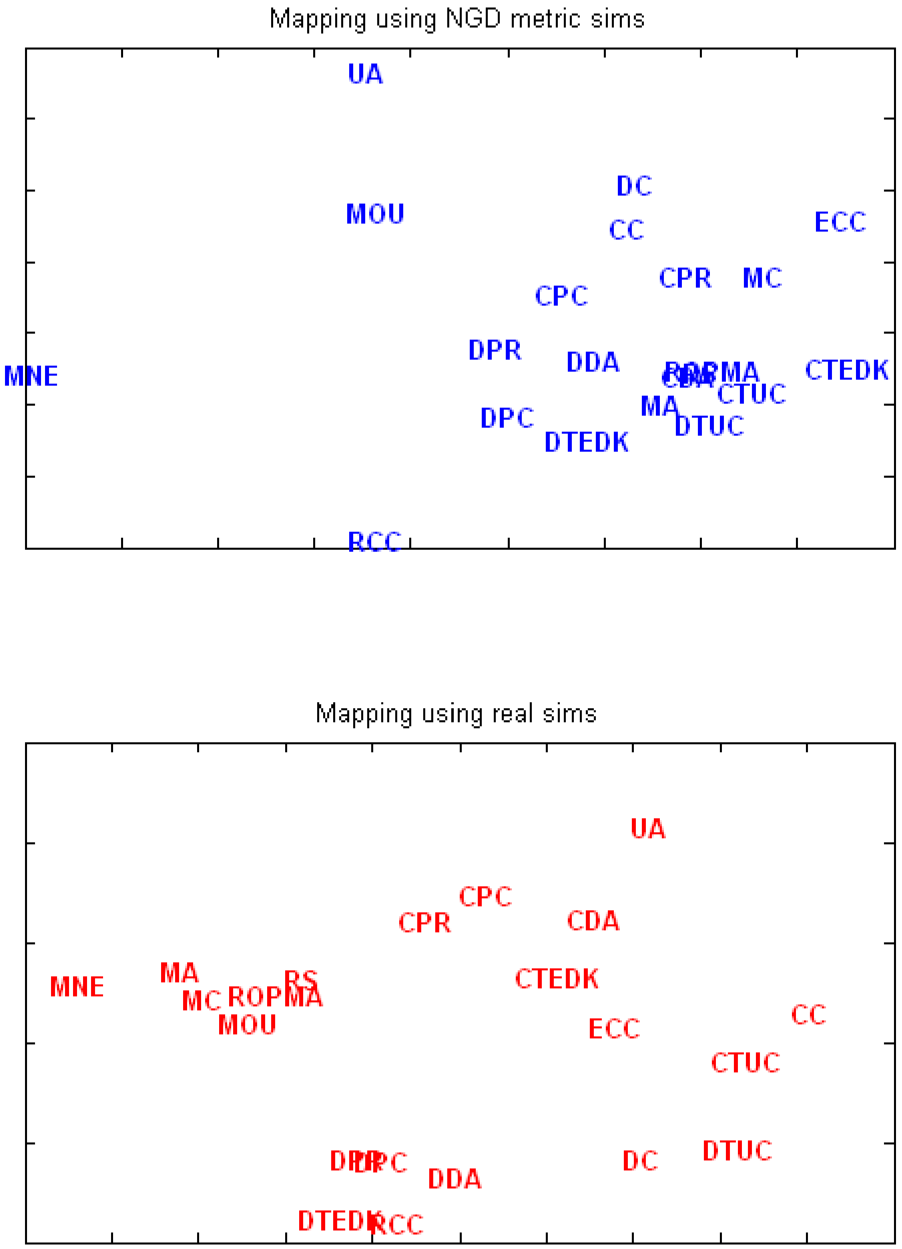


Fig 44.Maps by the ngd metric in contrary with mapping of for all terms of Greek political sciences survey

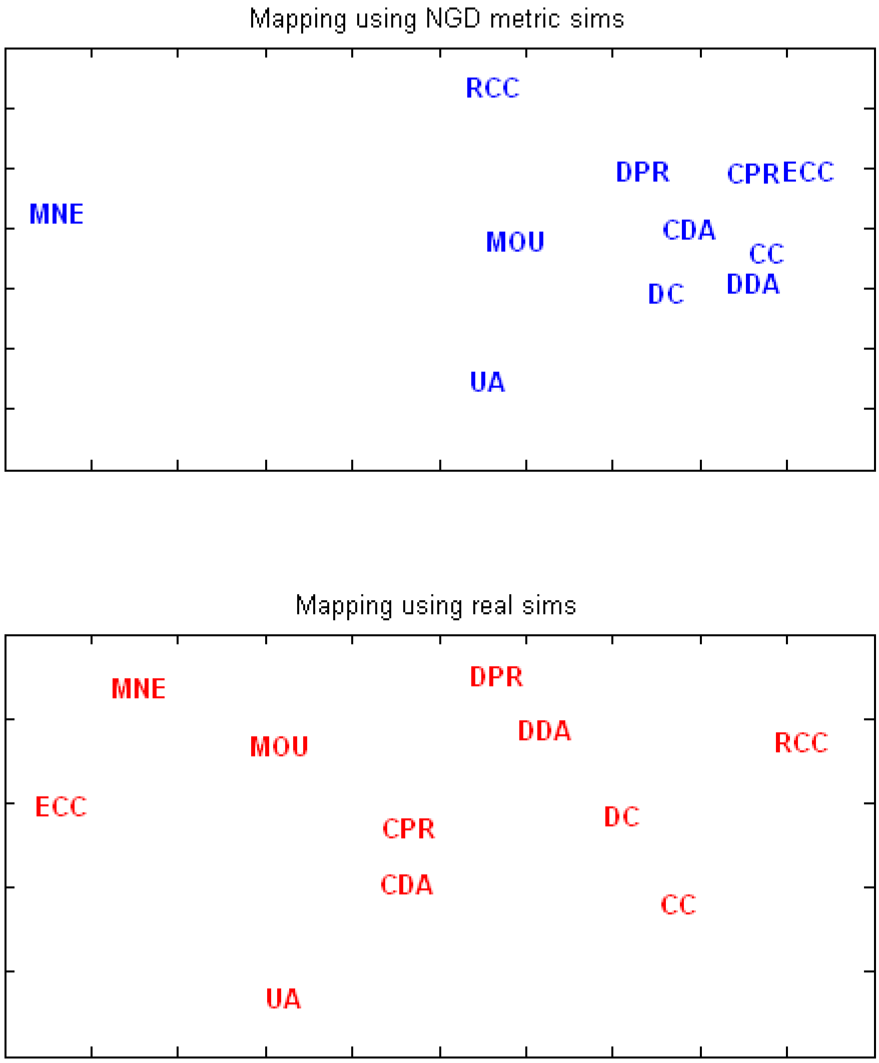


Fig 45.Maps by the ngd metric in contrary with mapping of for 10 terms of Greek political sciences survey

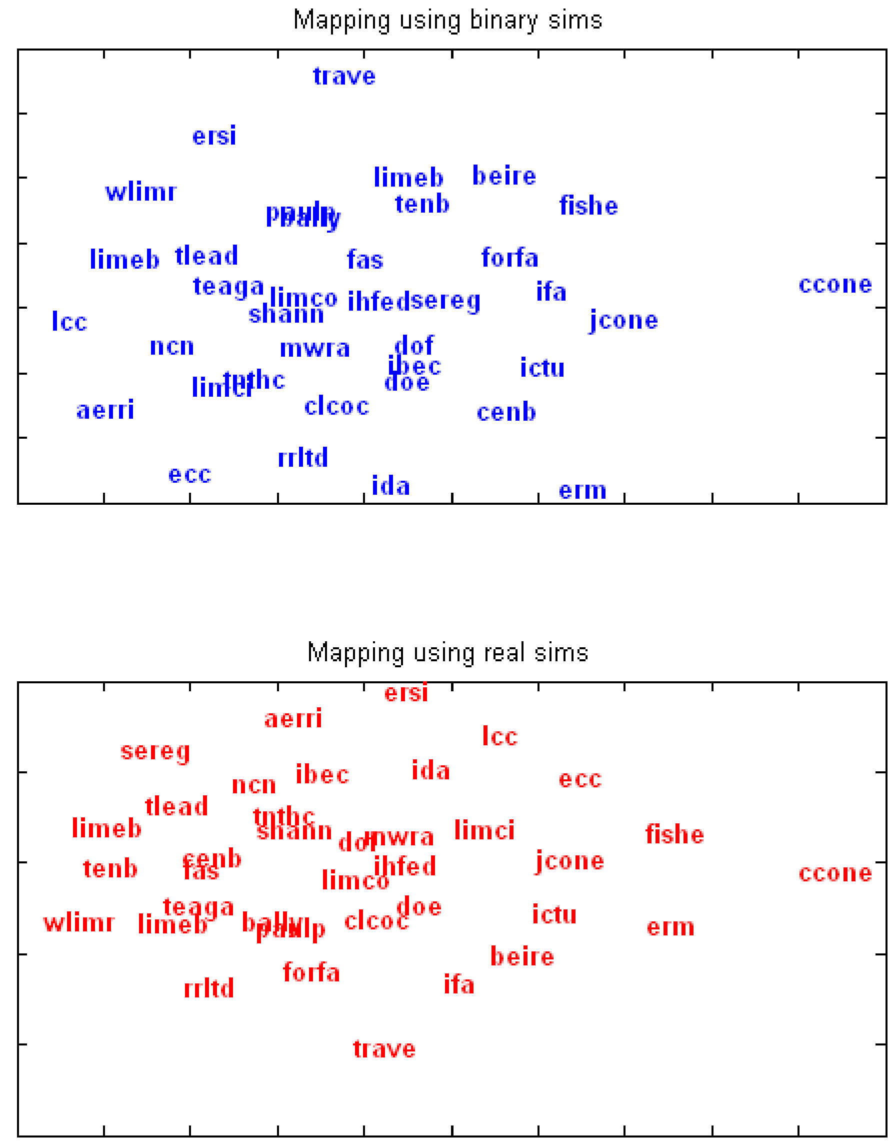


Fig 46.Maps by the binary scheme in contrary with mapping of political sciences survey

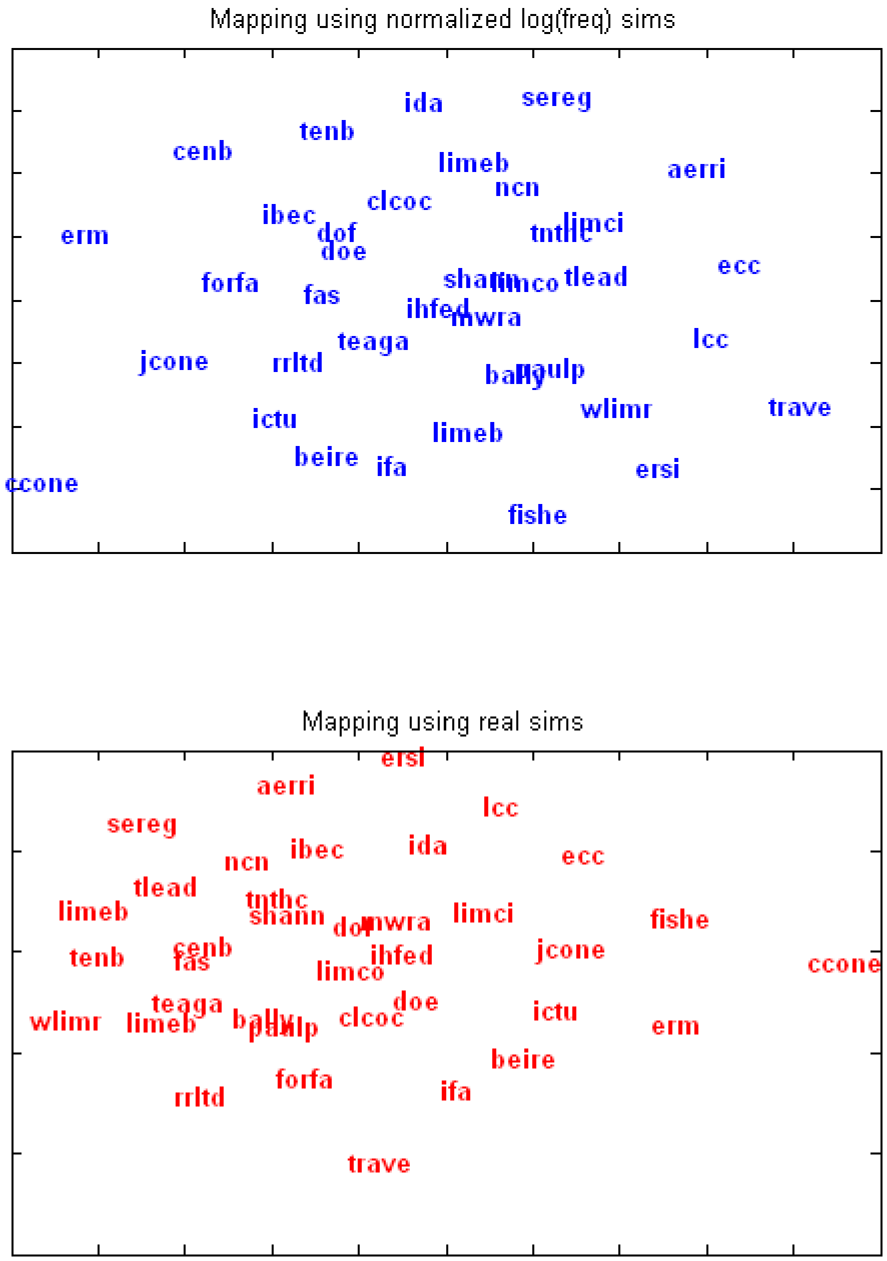


Fig 47.Maps by the normalized log(freq) scheme in contrary with mapping of Irish political sciences survey

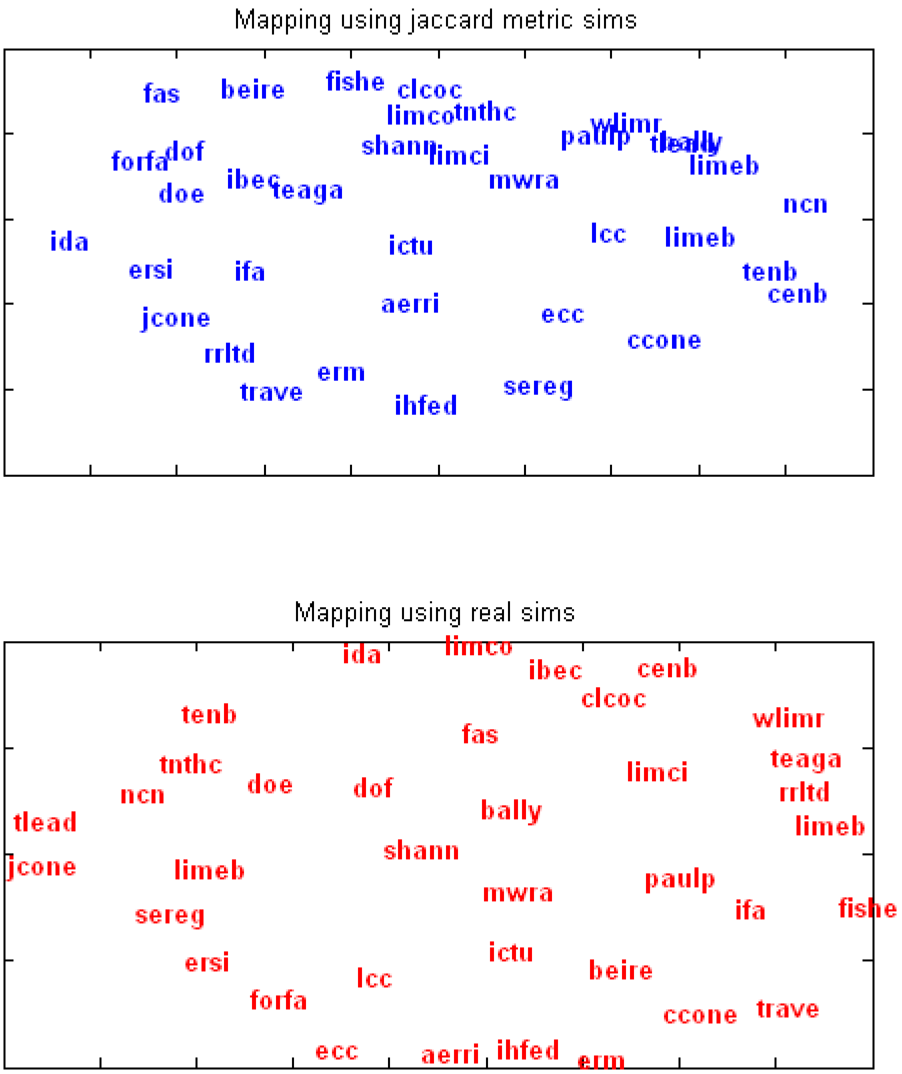
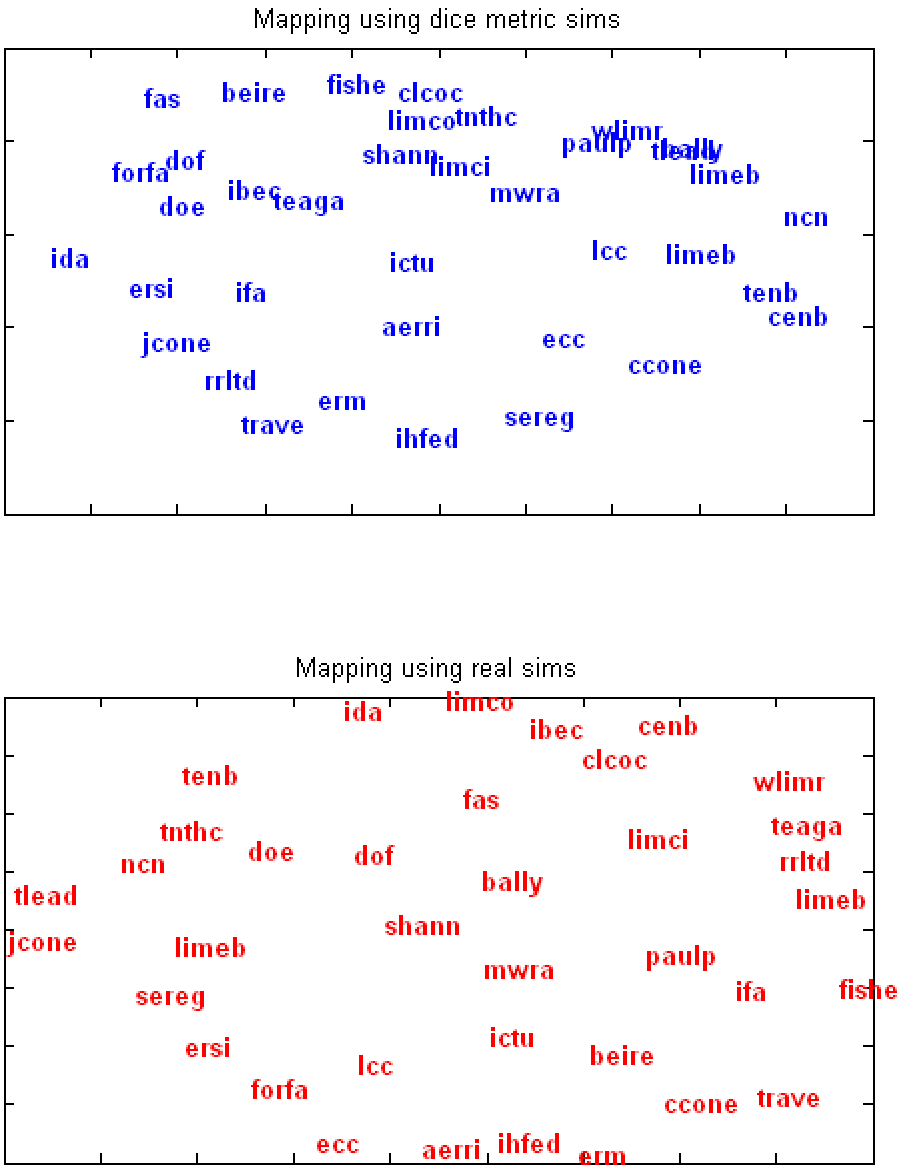


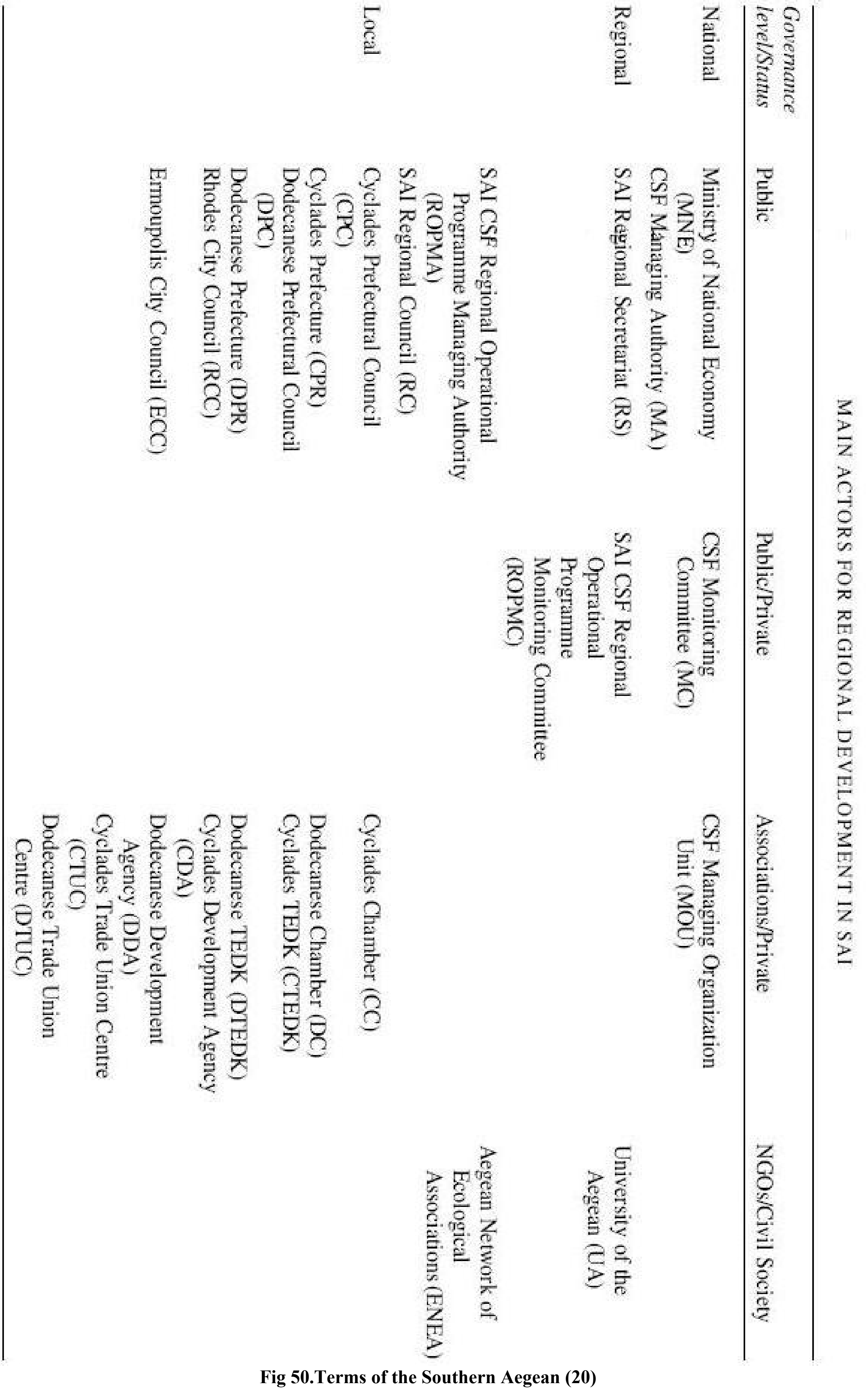
Fig 48.Maps by the jaccard metric in contrary with mapping of Irish political sciences survey


Mapping using dice metric sims

Mapping using real sims

Fig 49.Maps by the jaccard metric in contrary with mapping of Irish political sciences survey

C) Table with the meaning of all Greek policy terms' abbreviation

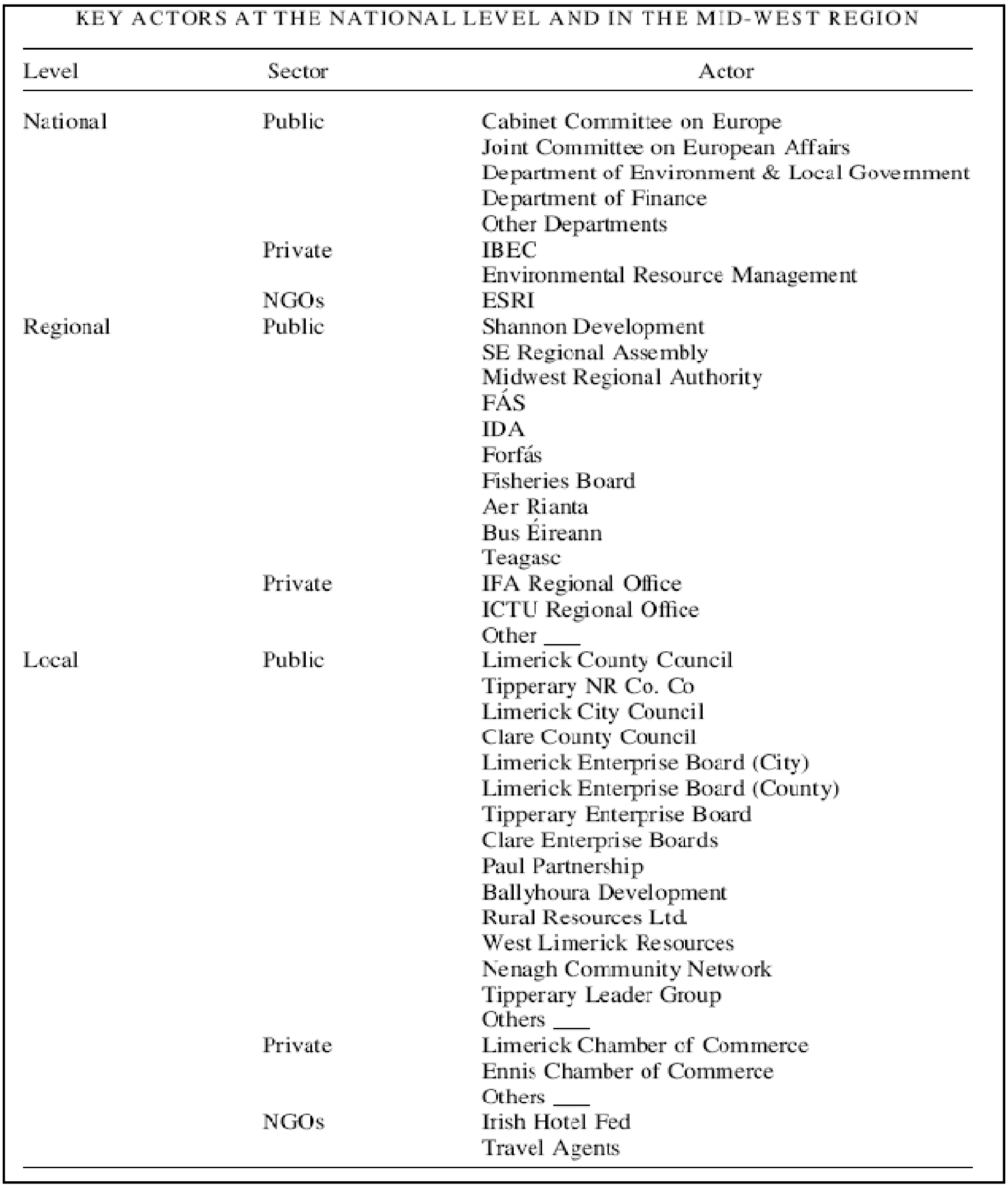


MAIN ACTORS FOR REGIONAL DEVELOPMENT IN SAI

Governance level/Status	Public	Public/Private	Associations/Private	NGOs/Civil Society
National	Ministry of National Economy (MNE) CSF Managing Authority (MA)	CSF Monitoring Commitee (MC)	CSF Managing Organization Unit (MOU)	
Regional	SAI Regional Secretariat (RS) SAI CSF Regional Operational Programme Managing Authority (ROPMA) SAI Regional Council (RC)	SAI CSF Regional Operational Programme Monitoring Committee (ROPMC)		University of the Aegean (UA) Aegean Network of Ecological Associations (ENEA)
Local	Cyclades Prefectural Council (CPC) Cyclades Prefecture (CPR) Dodecanese Prefectural Council (DPC) Dodecanese Prefecture (DPR) Rhodes City Council (RCC) Ermoupolis City Council (ECC)		Cyclades Chamber (CC) Dodecanese Chamber (DC) Cyclades TEDK (CTEDK) Dodecanese TEDK (DTEDK) Cyclades Development Agency (CDA) Dodecanese Development Agency (DDA) Cyclades Trade Union Centre (CTUC) Dodecanese Trade Union Centre (DTUC)	

Fig 50.Terms of the Southern Aegean (20)

D) Table with the meaning of all Irish policy terms' abbreviation



KEY ACTORS AT THE NATIONAL LEVEL AND IN THE MID-WEST REGION

Level	Sector	Actor
National	Public	Cabinet Committee on Europe
		Joint Committee on European Affairs
		Department of Environment & Local Government
		Department of Finance
		Other Departments
	Private	IBEC
		Environmental Resource Management
	NGOs	ESRI
Regional	Public	Shannon Development
		SE Regional Assembly
		Midwest Regional Authority
		FÁS
		IDA
		Forfás
		Fisheries Board
		Aer Rianta
		Bus Éireann
		Teagasc
	Private	IFA Regional Office
		ICTU Regional Office
		Other ___
Local	Public	Limerick County Council
		Tipperary NR Co. Co
		Limerick City Council
		Clare County Council
		Limerick Enterprise Board (City)
		Limerick Enterprise Board (County)
		Tipperary Enterprise Board
		Clare Enterprise Boards
		Paul Partnership
		Ballyhoura Development
		Rural Resources Ltd
		West Limerick Resources
		Nenagh Community Network
		Tipperary Leader Group
		Others ___
	Private	Limerick Chamber of Commerce
		Ennis Chamber of Commerce
		Others ___
	NGOs	Irish Hotel Fed
		Travel Agents

Fig 51.Terms of the Mid-West Ireland (21)

Bibliography

1. *Measuring Semantic Similarity between Words using Web Search Engines.* Bollendala, D., Matsuo, Y., Ishizuka, M. 2007, WWW, pp. 757-766.
2. *Unsupervised Semantic Similarity Computation usingWeb Search Engines.* Potamianos A., Iosif E. 2007, IEEE Computer Society, pp. 381-387.
3. *Auto-induced Semantic Classes.* Pargelis, A., Fosler-Lussier, E., Lee, C., Potamianos, A., Tsai, A. s.l. : Speech Communication, 2004, Vol. 43.
4. *Semi-Automatic Acquisition of Domain-Specific Semantic Structures.* Siu, K.C., Meng, H.M,. s.l. : Proc. EUROSPEECH , 1999.
5. *Combining Statistical Similarity Measures for Automatic Induction of Semantic Classes.* Pangos, A., Iosif, E., Potamianos, A., Fosler-Lussier, E. s.l. : IEEE Automatic Speech Recognition and Understanding Workshop, 2005.
6. *Papers in linguistics.* Firth, J. R. 1957.
7. *The Google Similarity Distance.* Cilibrasi, L. , Vitanyi, P. s.l. : IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, 2007, Vol. 19. 3.
8. *Word Space.* Schutze, H. Stanford : Proc. Conference on Advances in Neural Information Processing Systems (NIPS), 1993. 94305-4115.
9. Lackoff, G., Johnson, M.,. *Philosophy in the Flesh: The Embodied Mind and its Challenge to Western Thought.* 1997.
10. *Metaphors we live.* Lackoff, G., Johnson, M.,. s.l. : University of Chigago Press, 1980.
11. Sahlgren, M.,. *The Word-Space Model: using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces.* 2006.
12. *Contextual Correlates of Synonymy.* Rubenstein H., Goodenough, B.J.,. s.l. : Communications of the ACM, 1965, Vol. 8.
13. *Information Retrieval Based on Word Senses.* Schutze, H., Pedersen, J.,. s.l. : Proc. 4th Annual Symposium on Document Analysis and Information Retrieval, 1995.
14. *Distributional Structure.* Harris, Z. s.l. : Papers in Structural and Transformational Linguistics, 1970.
15. *Mathematical Structures of Language.* Harris, Z. s.l. : Interscience Publishers, 1968.
16. *Contextual Correlates of Semantic Similarity.* Miller, G., Charles, W.,. s.l. : Language and Cognitive Processes, 1991, Vol. 6.
17. *Unsupervised Combination of Metrics for Semantic Class Induction.* Iosif, E., Tegos, A., Pangos, A., Fosler-lussier, E., Potamianos, A. s.l. : IEEE/ACL Spoken Language Technology Workshop, 2006.
18. *Flink: Semantic Web technology for the extraction and.* Mika, P. s.l. : Elsevier B.V., 2005, Vol. 3.
19. *Extracting Relations in Social Networks from the Web using Similarity between Collective Contexts.* Mori, J., Tsujishita, T., Matsuo, Y., Ishizuka, M.
20. Demetropoulou L., Getimis P. Towards new forms of regional governance in Greece: the Southern Aegean Islands. *Regional & Federal Studies.* September 01, 2004, pp. 355-378.
21. *Ireland's pragmatic adaptation to regionalization: the Mid-West region.* Rees, N., Quinn, B.,Connaughton, B. 3, s.l. : Regional & Federal Studies, 2004, Vol. 14.

22. *Knowledge Graphs and Network Text Analysis* . Popping, R. 91, s.l. : Social Science information, 2003, Vol. 42.
23. *Institutional 'legacies' and the shaping of regional governance in Hungary.* ILONA, I . PALNE KOVACS, PARASKEVOPOULOS, C., J. and HORVATH, G.,Y.,. s.l. : Regional & Federal Studies, 2004.
24. *The hidden Web.* Kautz, H., Selman, B, Shah, M.,. 2, s.l. : AI magazine, 1997, Vol. 18.
25. *Finding authoritative people from the web.* Harada, M., Sato, S., Kazama, K. s.l. : Joint Conference on Digital Libraries, 2004.
26. *Fast discovery of connection subgraphs.* Faloutsos, C., McCurley, K., S., Tomkins, A. s.l. : ACM SIGKDD, 2004.
27. *Artist classification with web-based data.* Knees, P., Pampalk, E., Widmer, G.,. s.l. : 5th International Conf. on Music Information Retrieval (ISMIR), 2004.
28. *Engineering and Utilizing a Stopword List in Greek Web Retrieval.* Lazarinis, Fotis. Mesolonghi : Wiley InterScience, 2007.
29. Papoulis, A.,. *ΠΙΘΑΝΟΤΗΤΕΣ ΚΑΙ ΤΥΧΑΙΕΣ ΜΕΤΑΒΛΗΤΕΣ ΚΑΙ ΣΤΟΧΑΣΤΙΚΕΣ ΔΙΑΔΙΚΑΣΙΕΣ, 3η ΕΚΔΟΣΗ.* ΘΕΣ/ΝΙΚΗ : ΕΚΔΟΣΕΙΣ ΤΖΙΟΛΑ, 2002. 960-7219-34-1.
30. *Contextual Correlates of Synonymy.* Herbert, R., Goodenough, B.J. s.l. : Communications of the ACM, 1965, Vol. 8.
31. *Mining Text using Keywords Destributions.* Feldman, R., Dagan, I. s.l. : Intelligent Information Systems, 1998.
32. *Word Assosiation Norms, Mutual Information, and Lexicography.* Church, W.K., Hanks, P. s.l. : Computanional Linguistics, 1990, Vol. 16.