ISSN (online) : 2277-5390
ISSN (Print): 2277 - 5382

Automatic extraction of policy networks using snippets and social networks

Vidhya.R*[1] , Ajitha.A[2]

*[1]*M.E Computer Science Engineering, MountZion College of Engineering and Technology, Pudukottai- 622507, Tamil Nadu*
[2]*Assistant Professor, Department of Computer Science Engineering, MountZion College of Engineering and Technology, Pudukottai- 622507, Tamil Nadu*
*[1] vidhyahicet@gmail.com
[2]ajithatensingh@gmail.com

Abstract

Background/Objectives: To automatically extract policy networks by using Snippets and Social networks.
Methods/Statistical analysis: The analysis of policy networks demands a series of difficult and time-consuming manual steps including interviews and questionnaires. The approach involves the process of estimating the strength of relations between actors in policy networks using features like webpage counts, out links, and lexical information extracted from data harvested from the web snippets. The approach extracts the irrelevant documents that affect the performance and accuracy. It is overcome by including the process of investigating machine learning algorithms for selecting the most informative metrics.
Findings: The proposed approach includes metrics such as recovery degree, in-link, broken link, anchor text type and kl-divergence and filtering web data based on relevance and type of source, investigating the applicability of proposed metrics for social networks. This enhances the extraction of policy network more accurately.
Improvements/Applications: Overall extraction of policy network is automatic and accurate while using the proposed approach of using Snippets and Social networks.

Keywords—*Policy networks, relatedness metrics, Spatial proximity, Social Proximity*

1. Introduction

Policy networks are widely used by political scientists and economists to explain various financial and social phenomena, such as the development of partnerships between political entities or institutions from different levels of governance. Policy networks is a set of relatively stable relationships which are of non-hierarchical and interdependent nature linking a variety of actors, who share common interest with regard to a policy and who exchange resources to follow these shared interests acknowledging for co-operation. Policy networks is a mere metaphor to denote the fact that policy-making involves a large number and wide variety of actors, others acknowledge them to valuable analytical tool to analyze the relations between actors interacting with each other in a given policy sector[1].

Automatic extraction of the Policy networks is achieved by information which is collected from the web. The relationship between actors can be estimated by i) the frequency of co-occurrence for each pair of actors in web documents, ii) the lexical contextual similarity between snippets iii) the co-occurrence of hyperlinks. In policy networks, nodes are actors who are involved in a given policy field and their relationship (friendship, co-authorship) is represented in edges [2][3].

After extracting the web documents, the web spam is detected using some technique like AIRWeb [link] competition. It is the definition of features which take different values for spam and non-spam pages. It has focused on quantitative features and also qualitative features to improve web spam detection.

The proposed algorithm uses the web document's metrics and social networks metrics. It increases the efficiency and reduces the error rate and develop high scalable algorithm for social proximity and spatial proximity [4]. Social and Spatial Ranking Query (SSRQ) reports the top-k users in the social networks based on ranking function that incorporates social and spatial distance from the query user.

Analyzing large amount of data for the policy networks has flourished in past years. There are two computational methods are used (i) Text analysis (ii) Social analysis. Most of the text analysis is done by electoral campaigns for identifying voter's profilers, ideology positions, political interaction and conflict contents [5][6]. One of the approaches of text analysis is the hand coding of traditional and highly labor intensive techniques of content analysis. That should be replaced by computerized coding schemes by matching the text to coding dictionaries. Estimating the policy position between two parties can be done by calculating the word scores. It can be done by WORDSCORE [7]. This can calculate the economic policy dimension to particular party. WORDFISH is an algorithm to estimate policy position based on frequencies in text. Extracted positions reflect changes in the party system more accurately than existing time series estimates and also to examine which words are important for placing parities on the left and right [8].

Social network analysis is an important fact of policy network. It can make formal and informal interactions. Extracting the social networks is done by the following steps (i) relation identification (ii) relation labeling [9] and (iii) estimating strength. Relation labeling is based on link based ranking method. Link refers to the sites and listing the site is automated rather than manual work. Generally site is an organized collection of pages on a specific topic maintained by a person or group. Topic of site might be quite broad, it broad range of subject matter and provides a range of service. Site finding is a huge process and also it makes a problem describing them. Inside the site it has a more content and the common features like the frequency of co-occurrence of related pair of terms in web documents, but other features lexical content, key phrase, log files and e-mail information are used to identify the relation.

Social network is an important role in our life which includes knowledge management, information retrieval. That network should be extracted using some technique. One of the techniques is POLYPHONET which employs several advanced techniques to extract relations of persons, to detect groups of persons and also to obtain keywords for a person [10]. It has three main things to do, first to reduce the related method into simple, second to create algorithm for new things like classifying the categories, obtaining and utilizing the person-to-word relationship and finally all the modules are implemented into POLYPHONET. Automatically extracted keypharses are used to describe the relation between entities such as affiliations, roles, locations, part-whole and social relationships. Obtain a local context in which two entities co-occur on the web and accumulate the context of the entity pair, the key idea is clustering all entity pairs according to similarity of their collective contexts.

The influence between two users captures the probability that one user follows the other's actions. Stem from social networks can improve the marketing strategies, for instance, by recommending product to users based on purchases of their contacts [11] [12]. In past graph scores were created for influencing users. But data scientists directly obtain the top-k most influential users from the historical data without intermediate step of constructing influence graph. Some of the techniques are, vertex neighborhoods (Shortest path distance between the users) [13] and the number of common friends [14].

Object associate with multiple domains attributes. Web pages containing geo information and flick photos with geo-tags require query processing on both the spatial and textual domains. It has to apply in IR-tree data structure, which extends the R-tree with inverted files. This index can be used to efficiently support novel types of spatio-textual queries [15],[16]. By observing location-based social networks, they build supervised learning framework based on predictions and common check-ins.

In graph search, the users build the graph with shortest path from source to target vertex. Dijikstra's algorithm starts from source and iteratively expands the network using priority heap until the target reached. Landmark concept were introduced which select a set of vertices as landmark in graph and pre-computes distance from every vertex to each landmark [17]. An approach to compute approximate distance between vertices in a graph is to construct oracles which provide constant query time while having linear space requirements.

2. Relatedness metrics

There are 3 kinds of metrics for computation between actors (i) Page count metric (i) Text based metrics and (iii) Link based metrics. The classical information retrieval techniques and natural language processing mainly consist of two stages (i) Extracting relevant information on a link and constructing complex queries and requesting to a search engine. In the first stage, anchor text is an important source but it doesn't return enough information. So we take other sources like URL- the page that contains the link, the context of anchor text and cached page version of the analyzed sources that can be stored in digital library. Extraction can be done by two main approaches and they are based on frequency (Term Frequency- Inverse Document Frequency) and statistical language model (KL divergence). In the second stage, a link is considered to be recovered if the page pointed by the link is in the set of pages retrieved with some of the queries.

A. *Pagecount based metrics*

Estimate the co-occurrence between the policy actors. In web documents Co-occurrence means capturing the variety of relation among the terms ranging that is both deal with common policy issues and serve with policy functions. For a set of documents indexed by search engine denoted by {D} and its cardinality be |D|. {D_{ai}} is indicates the set of web document indexed by actor ai. { D_{ai} , D_{aj}} represents set of web documents that contains actor ai and aj and its cardinality | D_{ai} , D_{aj}].

Jaccard coefficient: This coefficient computes the similarity between sets. The jaccard coefficient S^P_J between ai and aj is defined as:

$$S_J^P(ai,aj)=\frac{|D_{ai,aj}|}{|D_{ai}|+|D_{aj}|-|D_{ai,aj}|}$$

Dice coefficient: It is slightly same as jaccard coefficient and it defined as:

$$S_D^P(ai,aj)=\frac{2|D_{ai,aj}|}{|D_{ai}|+|D_{aj}|}$$

Mutual information: If the set of documents are random variable then point wise mutual information reflects the dependence between the occurrences of actors [18].

$$S_I^P(ai,aj)=\log\frac{\frac{|D_{ai,aj}|}{|D|}}{\frac{|D_{ai}|}{|D|}\frac{|D_{aj}|}{|D|}}$$

Google-based semantic relatedness: It is another similarity metric and proposed in [19][20]

$$S_R^P(ai,aj)=\frac{\max\{\log|D_{ai}|,\log|D_{aj}|\}-\log|D_{ai,aj}|}{\log|D|-\min\{\log D_{ai}|,\log|D_{aj}|\}}$$

B. *Text-Based metrics*

The proposed text based metric computes the lexical similarity between political actors who appear in snippet. The lexical similarity may occur in syntactic, semantic and topical features for example if two actors share their political activities, then it is expected that their activities mentioned in their lexical surroundings. To extract the lexical features for actor ai, text –based metrics apply a contextual window W (containing W words preceding and W word following the actor)

$|f_{W,L}\ldots\ldots\ldots f_{2,L}, f_{1,L}|ai|f_{1,R} ,f_{2,R} \ldots\ldots\ldots f_{W,R}|$

Where f_j is the jth feature that exist left and right to the context of actor ai and the feature vector is built as $V_{ai,w}$=($v_{ai,1}$,$v_{ai,2,\ldots\ldots\ldots,}$ $v_{ai,N}$). $v_{ai,j}$ is an non negative integer and N is an vocabulary size. Context-based metric S_W^T computes the cosine similarity between the actors ai and aj as follow as

$$S_W^T(ai,aj)=\frac{\sum_{l=1}^{N} v_{ai,l}v_{aj,l}}{\sqrt{\sum_{l=1}^{N}\left(v_{ai,l}\right)^2}\sqrt{\sum_{l=1}^{N}\left(v_{aj,l}\right)^2}}$$

C. *Link based metrics*

Examine the number of hyperlinks which is commonly shared between the two actors that contains the term of interest. Common links denote that political actors share the same interest or point to common links in the networks. It expects that hyperlinks will point to topically relevant web sites and documents. Generally the outlinks are represented in two form either full form where whole path is specified or base form where only the main website address is indicated.

The set of outlinks for actors is represented by {O_{ai}} that appears in the web documents. In this metrics the similarity between the actors can be calculated based on the overlap between the members of their outlink sets.

Google-based semantic relatedness using outlinks(S_G^L)

$$S_G^L(ai,aj)=\frac{\max\{\log|O_{ai}|,\log|O_{aj}|\}-\log|O_{ai,aj}|}{\log|O|-\min\{\log O_{ai}|,\log|O_{aj}|\}}$$

Where O_{ai} , O_{aj} and { $O_{ai,}$ O_{aj}} are set of outlinks for actor ai and aj and jointly for both (intersection of O_{ai} and O_{aj})

D. *Recovery Degree*

For every page the system tries to retrieve all their links and result as, three values are obtained: (i) the number of recovered links (top ten of search) (ii) the number of unrecovered links and (iii) the difference between both previous values. The degree of recovered links can be understood as a coherence measure between the analyzed page, one of its links, and the page pointed by this link [21].

E. *Incoming-Outgoing links*

This is link from spam pages to non spam pages but non spam pages do not link spam pages. Taking advantages of the possibilities of the system to submit queries to search engine, to include new query to request the search engine, how many sites point to the analyzed pages (incoming pages).

F. *Broken Links*

It is common problem for both spam and non spam pages, even when this sort of link has a negative impact in the PageRank. The number of spam pages is higher in almost the whole range of numbers of broken links considered.

G. *Anchor text typology*

The anchor text of many links is usually generated thinking in the context of the search engines instead of the users. We have to select four features in order to measure the number of links that are formed by (i) punctuation marks (ii) digits (ii) a URL and (iv) an empty chain.

H. *Linear fusion of relatedness metrics*

Combine all the features using late integration, relatedness scores from the types of metrics. Linear fusion (S) between two actors is given by

$$S(ai,aj)=\lambda_P S^P(ai,aj)+\lambda_T S^T(ai,aj)+\lambda_l S^T(ai,aj)$$

Where λ_P ,λ_T λ_L are corresponding weights for each metrics.

3. Proposed methodology

The proposed algorithm and metrics have been evaluated on the policy networks. In policy networks, documents, snippets and number of hits are mined from web for pair of actors and relatedness metrics are calculated in the form of correlation. The selection of most descriptive terms or passages from the text is crucial for several tasks. In major cases, ranking of all keywords or sentences are done and then top ranked items are selected as features,

ISSN (online) : 2277-5390
ISSN (Print): 2277 - 5382

Semantic rank is a graph based ranking algorithm where it constructs a semantic graphs using implicit links, based on semantic relatedness of text nodes. It also ranks the consequent nodes using different ranking algorithms [22].

Algorithm 1: Semanticrank

```
Input: a text document collection D, Mode flag
Output: a ranking R of the semantic graph nodes for every document $d_j$ D
Execute(D, mode)
If mode is keyword then
    Identify composite term of length up to 5 words
End if
Compute and index TF-IDF values for all terms
For all $d_j \epsilon$ D do
    G: an initially empty graph
    G: Construct_Semantic_Graph($d_j \epsilon$ D)
    R= Rank_nodes(G)
Construct Semantic Graph( $d_j \epsilon$ D)
G: an initially empty graph
If mode is keyword then
    Initialize G with $K_{dj}$
Else
    Initialize G with $Sen_{dj}$
End if
For all pairs of verticea($v_i$ , $v_j$)do
    If mode is keyword then
        $W_{i,j}= W_{j,i}=\lambda_{vi,vj}.SRT(v_{i,} v_j)$
    Else
        $W_{i,j}= W_{j,i}=SRS(v_{i,} v_j)$
    End if
End for
Return G
Rank nodes G
Execute weighted page rank in G
R= rank vertices of G in descending order of pagerank values
Return R with their page rank values
```

We characterize the relationship between two linked web pages according to different values of divergence. These values are obtained by calculating the Kullback–Leibler (KL) divergence between one or more sources of information from each page. KL divergence can be applied to the anchor text of a link and the title of the page pointed by this link. KL measures the probability of more relevant terms in terms of segment texts and documents.

SVM is supervised learning model with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. It is used to identify the most informative metric more efficiently. In this module, we introduce a machine learning algorithm for selecting the most informative metrics based on the frequent occurrences, as well as, for their fusion, and filtering web data based on relevance. It is used to classify and estimate the weights for the linear fusion of the additional metrics. In policy network, this algorithm is efficient for classification of social users [23].

Social and spatial ranking query (SSRQ) reports the top-k users in the SN based on a ranking function that incorporates social and spatial distance from the query user.

Algorithm 2: top-k users

```
Input: N training samples for classification,
Output: predicted class
For each sample do
    Input[i] =sample
    Foreach i in neural metwork do
        Output [i] = module. ForwardPropagate(input[i])
        Input[i+1]=output[i]
    End
    Predictedclass = criterion (output)
    If training then
        Check error attributes
        For each [k-i] in neural network do
            Output = module. BackwardPropagate
            Input[i+1]=output[i]
        End
    End
End
```

In each input sample, it can be positive, negative or neutral in the data sample. For the number of input training sample we have to perform the analysis of reviews. For each input sample the neuron network perform the output sample based on the prior knowledge. It is used to search the more accurate results along with hidden layer and this layer is used to map the semantic results for corresponding input sample. Hence it avoids the error values also it progress the speed of process more efficiently.

The above mentioned algorithm is used to classify and estimate the weights for the linear fusion of the additional metrics. The NN algorithm achieves better efficiency in finding the most informative metrics. The other metrics are fused using linear fusion technique. In policy network, this algorithm is efficient for classification of social users.

4. Experimental setup

In this section we experimentally evaluate the ranking algorithm and proximity values. All the methods are implemented in Java. We use publicly available dataset like INDIA, CHINA [other countries]. In experimental setup in table I , first to retrieve the document from the web by entering the actors and then to listing the abbreviation from the web documents. To tackle both data sparseness and term ambiguity problems and also a number of lexicalization forms is manually selected for each actor in collaboration with actors. The machine learning algorithm is such as SVM which is used to identify the most informative metrics in more efficiently.

Table 1. Informative metrics

ACTORS	**PAGE COUNT BASED METRICS**			
	Jacaard coefficient	***Dice coefficient***	***Mutual information***	***Google***
India,China	0.690	0.549	0.078	0.431
	TEXTBASED METRICS(COSINE SIMILARITY)			
	0.5211			
India, China	**LINK BASED METRICS**			
	0.2088			
	Linear Fusion			
	0.4627			

After calculate the metrics to select the most informative metrics and to grouping the actors according to the deadline. That dead line should be decided by developer based on required actors. For example the minimal value is 0.6 and who are all having above 0.6 in linear fusion they become group 1 or else they become group 2. Classification is use to easily identify the actor's relationship. Finally it calculates the correlation co-efficient and MSE among the active learning algorithm and manual calculation. That should be describing in previous section. MSE would be reducing and correlation would be increase.

The second part of this work is social network. Recently so many social network are available like facebook, LinkedIn, Flicker, Google+ etc., here I took facebook. First step is to login the Facebook (Fb) Id in search engine. For retrieving the data from the Fb, want to get token from Fb. For that to go facebook developer login and get token from fb to log in account. After entering token, to get the friends details from the facebook. Calculate the social and spatial proximity for each friend with log in person. Like who made a post to me or like my profile, made Comments etc,.

Graph distance should be calculated among the actor and all the users. Euclidean distance is used for calculate the graph distance. Then Support Vector Machine algorithm is to find the top kth person to make interaction with user. Finally the accuracy should be returned.

5. Performance evaluation

To execute the proposed technique and generate various results we use java language in this environment. In this section, the analysis has been done for existing and proposed research work by using algorithms. In this evaluation part, the performance metrics are executed by using existing and proposed method. The performance metrics are such as correlation, MSE values, graph distance and accuracy. The proposed system is shown the higher performance in terms of high correlation, lower MSE values, efficient graph distance and greater accuracy.

The process can be evaluated based on correlation and Mean Square Error (MSE). Correlation is calculated depends on some relatedness scores. That has H= (h_1, h_2,.., h_M)and K =(k_1,k_2,…. k_M) indicates human rated and automatically computed relatedness scores, respectively, where M is the total number of relations. k_i is calculated from the section 3 mentioned metrics.

It scaled as

$$e_i = \frac{2(k_i - k_{\min})}{k_{\max} - k_{\min}} + 1$$

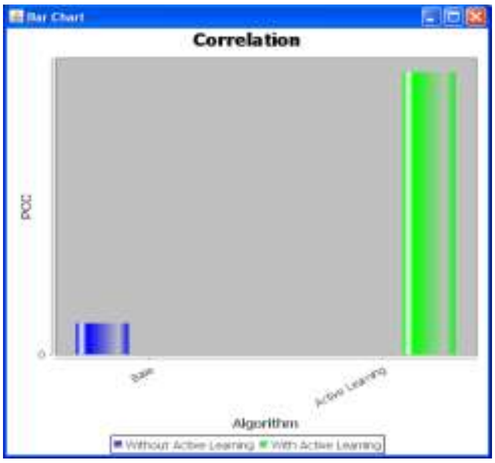
Where k_{min} and k_{min} is minimum and maximum score of metrics and is normalized relatedness score that takes continuous value. Correlation is calculated between human rate and normalized relatedness score by using Pearson coefficient,

$$r_{H,E} = \frac{\sum_{i=1}^{M}(h_i - \overline{H})(e_i - \overline{E})}{\sqrt{\sum_{i=1}^{M}(h_i - \overline{H})^2 \sum_{1=1}^{M}(e_i - \overline{E})}^{2}}$$

Where $\overline{H}$ and $\overline{E}$ denoted the sample mean of H and E. In Mean Square Error (MSE) is used to measure the distance between the human ratings and normalized relatedness averaged over all investigating relations.

The correlation value should be high in the policy network. If the correlation value is high then system is performed in an effective manner.

Figure 1. Correlation among non active learning and active learning algorithm



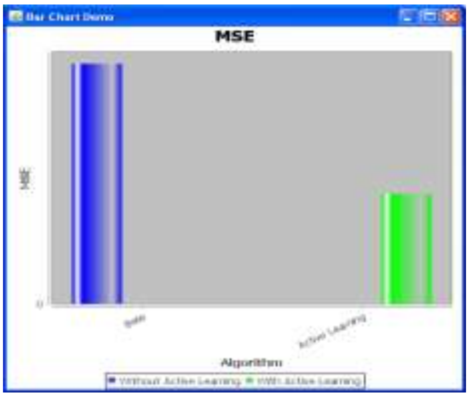
From the above graph in fig.1 we can observe that the comparison of existing and proposed system in terms of correlation metric. In x axis we plot the algorithms and in y axis we plot the correlation values. The correlation values are lower by using existing algorithm. The correlation value is higher by using the proposed algorithm of active learning. From the result, we conclude that proposed system is superior in performance.

In statistics, the mean squared error (MSE) of an estimator measures the average of the squares of the "errors".

$$MSE = \frac{1}{M}\sum_{i=1}^{M}(h_i - e_i)^2$$

Note the MSE values lies between 0 and 4.

Figure 2. MSE among non active learning and active learning algorithm



The accuracy is the proportion of true results (both true positives and true negatives) among the total number of cases examined in Fig.2.

Accuracy can be calculated from formula given as follows

$$Accuracy = \frac{true\ \ positive\ + True\ negative}{TruePositive + True\ negative + False\ \ positive + False\ negative}$$

An accuracy of 100% means that the measured values are exactly the same as the given values.

By using machine learning algorithm the proposed system has shown the highest accuracy rather than previous system.

The social and spatial proximities are calculated based on the graph distance value. The proposed system is shown the aggregate proximity values efficiently and the nearest as well shortest distances are computed effectively.

6. References

1. Tanja A.Borzel. An exploration of the concept and its usefulness in studying European governance. European Integration online papers, 1997; 1, 16.
2. R.A.W. Rhodes. Policy network analysis, John Donne [1611], The First Anniversary. An Anatomy of the World, 1985 edition, 335 line 213.
3. Theodosis Moschoulos, Elias Iosif. Towards automatic extractiojn of policy networks usinf web links and documents. IEEE, 2013; 25(25), 2403-2417.
4. Kyriakos Mouratidis, Jing Li, Yu Tang, Nikos Mamoulis. Joint search by social and spatial proximity. IEEE transactions on knowledge and data engineering. 2015; 27(3), 781-793.
5. Burt L.Monroe. Introduction to the special issue: The statistical analysis of political text. in Advance access publication.2009; 351-355.
6. L.Zhu, Computational Political Science Literature Survey. http://www.personal.psu.edu/luz113. Date accessed: 2013.
7. Michael laver and Kenneth benoit. Extracting policy positions from political texts using words as data. American political science review. 2003; 97, 2.
8. Nick Craswell, David Hawking, Stephen Robertson. Effective site finding using link anchor information. Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval. 2001; 250-257
9. Jonathan B. Slapin and Sven-oliver Proksch. A scling model for estimating time-series party positions from texts. Trinity college, Dublin and University of california, Los Angeles. 2008; 52(3), 705-722.
10. Yutaka Mastsuo, Junichiro Mori and Masahiro Hamasaki. POLYPHONET: An advanced social network extraction system from the web. National institude of Advanced industrial science and technology, United Staes, University of Tokyo, Japan. 2007; 5(4), 262-278.
11. L. Adamic, O. Buyukkokten, E. Adar. Social network caught in the Web. First Monday 2003; 8 (6).
12. Dutta P, Kumaravel A. A Novel Approach to Trust based Identification of Leaders in Social Networks. Indian Journal of Science and Technology. 2016; 9(10), 1-9.
13. Skorobogatov VV, Yurchenko IV, Yurchenko NN, Telyatnik TY, Yermolenko OA. Scientific Methods in Social and Humanitarian Researches. Indian Journal of Science and Technology. 2015; 8 (s10), 1-9.
14. Singla ML, Durga A. How Social Media Gives You Competitive Advantage. Indian Journal of Science and Technology. 2015; 8(S4), 90-95.
15. Khazali MJ, Sargolzaei E, Keikha F. Privacy Preserving Approach of Published Social Networks Data with Vertex and Edge Modification Algorithm. Indian Journal of Science and Technology. 2016; 9(12), 1-8.
16. Ramya GR, Sivakumar PB. Advocacy Monitoring of Women and Children Health through Social Data. Indian Journal of Science and Technology. 2016; 9(6), 1-6.
17. Liu K, Xu L, Zhao J. Co-extracting opinion targets and opinion words from online reviews based on the word alignment model. Knowledge and Data Engineering, IEEE Transactions on. 2015; 27(3), 636-50.
18. D. Bollegala, Y. Matsuo, M. Ishizuka. Measuring Semantic Similarity between Words Using Web Search Engines. Proc. 16th Int’l World Wide Web Conf. 2007. 757-766.
19. R. Cilibrasi, P. Vitanyi. The Google Similarity Distance. IEEE Trans. Knowledge and Data Eng. 2007; 19(3), 370-383.

20. P. Vitanyi. Universal Similarity. Proc. Information Theory Workshop Coding and Complexity. 2005; 238-243, 2005.
21. Lourdes Araujo, Juan Martinez- Romo. Web Spam Detection: Mew Classification Features Based on Qualified Link Analysis and Language Model. IEEE Trans. On Information Forensics and Security. 2010; 5(3), 581-590.
22. George Tsatsaronis and Iraklis Varlamis. Semantic Rank: Ranking Keywords and Sentences using Semantic Graphs, in Norwegian University of Science and technology. 2010; 1074-1082.
23. Priyanka N Guttedar, Pushpalata S. Scene Text Recognition in Mobile Application using K- Mean Clustering and Supporrt Vector Machine. IJARCET. 2015; 4(5), 2492-2496

The Publication fee is defrayed by Indian Society for Education and Environment (iSee). www.iseeadyar.org

Cite this article as:

Vidhya.R, Ajitha.A. Automatic extraction of policy networks using snippets and social networks. *Indian Journal of Innovations and Developments.* 2016; 5(1), January.