Survey Paper on Discovering Groups by Analyzing Web Documents and Links

Vijay Ghute[1] Prof. S. R. Durugkar[2]
P.G. student[1] ,Assistant Professor[2],
Department of Computer Engineering
Late G.N.Sapkal College of Engineering, Anjaneri, Nasik [1, 2]
University of Pune

Abstract:

Policy groups are broadly used by many areas to discuss variety of issues. The analysis of policy groups requires a series of difficult and time-consuming manual steps including interviews and questionnaires.

As we will analyze everyone's area of interest and will try to analyze all the documents then deriving the conclusion about those users i.e how many users are interested in similar documents.

In this paper, we have survey after studying various papers, identifying the strength of relations which may exists between users in policy groups using feature extraction process from data retrieved from the web links and documents.

As we are saying features it may include webpage, web documents, and it's extraction is nothing but counting the incoming and outgoing links and also extracting lexical information about those documents.

Then at the end we have proposed what we are thinking about our proposed approach.

The main focus is on extracting the various documents, web pages and concluding by forming 'n' groups which may be interested in common subject or area.

Keywords: clustering, Information Retrieval, Policy groups, Features, Web documents.

I.INTRODUCTION TO POLICY GROUPS

We will consider this term "network" as a group which is often used to describe clusters i.e. set of peoples of different types of opinions and strategic views, who are related in the following few fields such as political, social, and economic spheres.

Fig 1.1 Groups of Peoples

Actually Policy Network organizes debates and conducts research on policy and political challenges that present all governments and political parties with urgent dilemmas, either because sustainable solutions remain elusive, or because there are political barriers to their implementation. This concept is from U.S. but here in our proposed system what we wish to do is to form the 'n' groups so that we will understand how many peoples are interested in common area.

Typically, policy groups can be identified through a manual procedure performed by experts. Identification of actors, web links, and documents i.e., analyzing a policy group's structure, requires various techniques and broad and time-consuming manual collection of data through interviews and questionnaires in which training data will be collected through asking some questions to end users and after receiving the answers of those questions we will try to evaluate those answers.

During the manual identification of groups, many subjective factors may be present, because this procedure relies strongly on the human subjects that participate in the interviews. Such factors include personal opinions, the person's willingness to participate, and even cultural issues. Overall, policy network identification currently requires a "large scale

investment" that does not always "lead to breathtaking empirical and theoretical results".

Here in this approach we will try to find out the behavior of end user i.e. in which area – subject area, domain he / she is interested. And then we will find the pairs at initial stage then formation of groups will starts.

The term 'group' is usually used to explain clusters of different kinds of users who are linked together in political, social or economic life. Networks may be loosely structured but still capable of spreading information or engaging in collective action.

II. SURVEY ON SOCIAL NETWORKS

A **social network** is a social structure made up of a set of social actors and a set of the relations between these actors. The social network perspective provides a set of methods for analyzing the structure of whole social entities as well as a variety of theories explaining the patterns observed in these structures. The study of these structures uses social network analysis to identify local and global patterns, locate influential entities, and examine network dynamics.

Fig 2.1 Social Network

Social networks and the analysis of them is an inherently interdisciplinary academic field which emerged from variety of fields where idea is to form 'n' clusters of data or given objects.

2.1 Use of Text Based Metrics

Assuming given the two web documents or links are similar, and characterizing their differences is useful in many tasks, including retrieval of information, updating some records, and knowledge base editing. We can describe a number of text based similarity metrics that characterize the relation between semantic web documents and evaluate these metrics for specific cases of similarity that we have to identify.

In addition to determining the similarity between given two Web documents, we generate a value which lies in between documents that have been identified as having a relationship with other data or object.

This proposed idea need to follow following steps:

1. Automatically identifying actors participating in policy groups means before categorization we need to realize those end users.
2. Next step is to apply some machine learning algorithmic steps such as calculation of term frequency and inverse document frequency
3. Filtering web data based on relevance and type of source i.e. filtration of those 'n' documents and identifying users as per their interest

2.2 Similar Aspect "Opinion Mining"

Opinion mining is the computational study of people's opinions, appraisals, attitudes, and emotions toward entities, individuals, issues, events, topics and their attributes. The task is technically difficult and practically very useful as we are identifying persons according to their opinions, attitudes and sentiments.

For example, in businesses or marketing we always wish to find out consumer opinions about their products and services. Potential customers also want to know the opinions of existing users before they use a service or purchase a product.

Like association rule in which we will predict the association in between the 'n' products. Means with the help of Bayesian Networks / diagrams we will show the relation existing in between various products or objects. Similar case is applicable in or proposed system where we will predict the user "A" is relates with user "B".

Hence for 'n' user if we will do the same thing can form various groups based on some similarity.

III.IMPLEMENTATION ISSUES

3.1 Issues need to be discussed

- **The frequency of co-occurrence**

The frequency of co-occurrence for each pair of actors / person in given web documents, web links.

- The lexical contextual similarity between snippets of web documents in which the actors appear, and the co-occurrence of hyperlinks

Present in web documents that contain the descriptions of some users. For each type of feature and for their combinations, a variety of similarity metrics are used to estimate the link strength for each pair of users in this scenario what we wish to show is the prediction in between users and document present in training set. By using such type of approach we will draw a graph in which directly it will be possible for

use to show one kind of dependencies which exists in between user and documents.

The proposed method aims to be efficient and reduce human biases. A policy group can be considered as a type of graph in which the nodes will represent the users involved in a given policy field, while their relations are represented by the 'n' edges. In social networks, nodes usually correspond to persons and edges represent relations among them built on a ground of mutual understanding, which can take several forms, such as friendship.

3.2 Term frequency:

It is nothing but the checking of word's occurrence in one document. Means we will check how many times a user relates with the particular documents.

3.3 Inverse Document Frequency:

It is opposite to the term frequency where we are finding the word's occurrence in multiple documents.

Means we will check in our proposed system to how many and which documents a user is associated.

Both the above mentioned terms are very useful in our project because we need to find out the relationship in between users and documents.

3.4 Lexicalization of Users

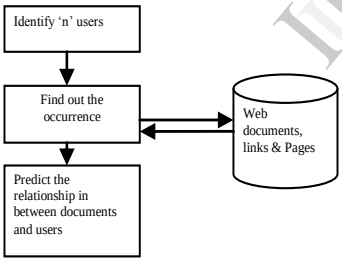

Fig 3.1 Proposed System

A critical step for the successful extraction of policy group is the derivation of the lexicalized forms that describe each of the users as we are analyzing every user and 'n' documents.

Lexicalizations are usually multiword terms or abbreviations, e.g., user "border security force" can be lexicalized as "border security troop" or abbreviated as "BSF." Using only the official names of users often returns very few relevant documents, while certain lexicalizations can be overly general, returning many irrelevant documents. Therefore idea is to maximize the relevant criteria by retrieving more relevant documents so that we will achieve the higher accuracy and efficiency for our system.

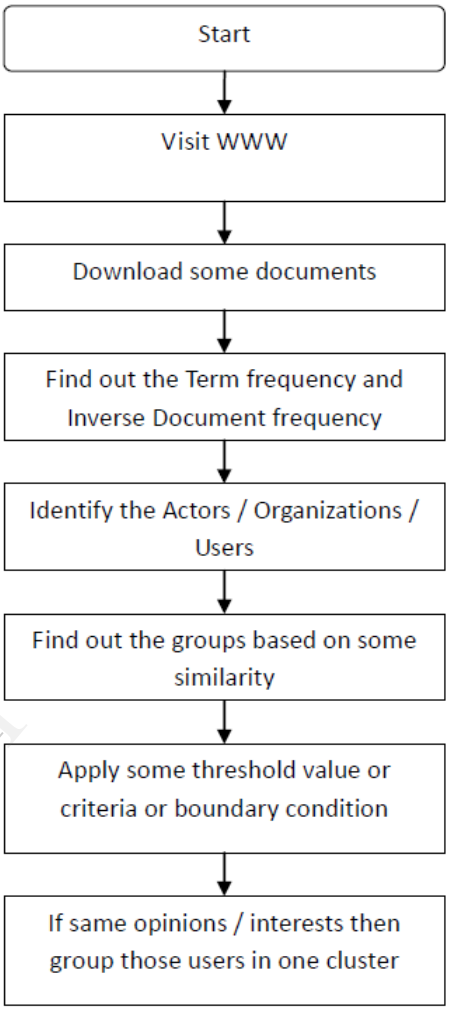

Fig 3.2 Flow of Proposed System

3.5 Focus on executing conjunctive query

In the theory of relational databases, a Boolean conjunctive query is a conjunctive query without distinguished predicates. Such a query evaluates to either true or false depending on whether the relations in the database contains the appropriate tuples of values.

We will have to consider following types of information after query execution at any search engine or information retrieval application:

1) Page counts,

2) URLs of those web documents which are retrieved, and

3) need to consider the outlinks and to grab those outlinks of the web documents we will have to introduce one extractor module which will extract the web pages and will find out the URLs and outlinks from that documents or web page.

Conclusion

In this way what we have to do is that first of all the documents need to be extracted and then after identifying the various users will try to analyze the opinions of each user.

So it will be possible for use to group those users into one cluster. Idea is to form clusters of users who will have some common opinion about something.

REFERENCES

1. Theodosis Moschopoulos, Elias Iosif, Student Member, IEEE, Leeda Demetropoulou, Alexandros Potamianos, Senior Member, IEEE, and Shrikanth (Shri) Narayanan, Fellow, IEEE •\Toward the Automatic Extraction of Policy Networks Using Web Links and Documents.
2. J. Peterson and E. Bomberg, Desicion-Making in the European Union. Palgrave Macmillan, 1999.
3. L. Zhu, Computational Political Science Literature Survey, http://www.personal.psu.edu/luz113, 13.
4. B. Monroe and P. Schrodt, ―Introduction to the Special Issue: The Statistical Analysis of Political Text,‖ Political Analysis, vol. 16, no. 4, pp. 351-355
5. P. Kenis and V. Schneider, Policy Networks and Policy Analysis:Scrutinizing a New Analytical Toolbox, pp. 25-59, Westview Press, 1991.