I.J. Information Technology and Computer Science, **2013, 02, 57-65**
Published Online January 2013 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2013.02.06

Knowledge Discovery in Endangered Species Diversification

Muhammad Naeem
Department of Computer Science, M. A. Jinnah University, Islamabad, Pakistan
naeems.naeem@gmail.com

Sohail Asghar
University Institute of IT, PMAS-Arid Agriculture University, Rawalpindi, Pakistan
sohail.asg@gmail.com

***Abstract*—** Classification of regional territories and countries related to endangered species has been investigated by data mining techniques and graphical modeling using an extensive data set of species. We developed the graphical models (hereafter referred to as 'ESDI') using cosine, jaccard similarity, K Mean clustering and cliques in graph modeling for a large number of countries. Environmental variables associated with species records were identified in context of their diversification to integration with our proposed prototype. We have shown that the problem of finding the most coherent clusters is reducible to finding maximum clique. Key findings include the urge to ameliorate communication about the loss and protection of endangered species and their concerned projects. The proposed framework is presented to serves a portal to knowledge discovery. We have concluded that the proposed framework model and its associated data mining similarity measures can be useful for investigating various scientific and management oriented questions related to protection of endangered species with emphasis on collaboration among regional countries. The rationale behind the proposed approach is that the countries which have been grouped into same clique inherit a lot of argues illustrating common reasons of their struggles towards ecological safety with minimization of perils for endangered species. The development and implementation of a regional approach based on this similar grouping address the actions that could offer significant benefits in achieving their goal for ecological policies. Other critical actions at this clique level include fortifying and elevating harmonization of legal frameworks with emphasis on prevention procedural issues; awareness realizations of endangered species issues and its priority. Such actions will eventually lead towards implementation of essential plans fulfilling co-operative expertise and common endeavors.

***Index Terms*—** Earth and Atmospheric Sciences, Similarity Measures, Document Analysis, Model Classification, Maximum Clique, Statistical Computing

I. Introduction

Let *Sp* = $\{S_1, S_2, S_3, \ldots\ S_n\}$ be a set of documents where each document represent list of endangered species. Similarity measure between these documents is comprised of $\overset{n}{\underset{i=1, j=1}{S_{sp}}} = \{^{\le 1}_{\ge 0}\}$ where similarity is determined between each document of endangered species. The outcome ranges between 0 and 1. Two kinds of similarity measures were considered in this study. These include jaccard similarity measure and Cosine similarity measure. Each of them can be defined such that $X = \{(J_i, C_i)\}$. Every member of the set X denotes the observation in two dimension space which can be clustered into relevant group. Let each cluster be defined as: $C_l[s_i, s_j] = \bigcup_{y \ldots z}^{k \ldots l} x$. There are arbitrary numbers of cluster. Each cluster C_l can be realized into a graph *G* such that $C_l \leftarrow G$. In this study we have considered only those clusters which have high threshold of similarity measure. The crux of this study is to reduce the problem of similarity into graphical model.

Lemma. Given a set of *n* documents *Sp* takes at most *n(n-1)/2* computational steps.

Proof. We need to examine, whether a distinct document takes at most 1 step for calculation of similarity measure with any other member of the document set. The first document needs (*n-1*) steps; the second document needs (*n-2*) steps and so on. The second last document needs only 1 step whereas the last document requires no step to compute the similarity measure. Hence it is proved that for a particular similarity measure, finding the number ρ of each document in the document set is ρ = (n-1) + (n-2) + …+1 = n(n-1)/2 steps.

Let $G = (\{A\}, \{E\})$ be an undirected graph with edge set $\{E\}$ such that a relationship exists between every element from set $\{A\}$. A clique, a subsets $A' \subseteq A$ denoted by (A', E') the subgraph of *G* induced by these sets. A max clique in *G* is a subgraph in which every vertex from set *{A}* has equal degree which is greater than zero. Moreover, a clique in *G* is a subset of edges $X \subseteq E$ such that the graph *G* (A', $E \backslash X$) does not have any node with degree less than one. Throughout the paper, we have the understanding that no such vertex exist which has zero degr*ee*. A clique X is maximal if, total number of edges $e \in X$, computed by $e(e\text{-}1)/2$ gives the highest score among all of the cliques found in graph *G* denoted by *M(G)* the families of maximum cliques for *G*. Note that in any graph *G*, the set M(*G*) is the family of maximal transversals to the family of cliques, i.e. *M(G)* is the family of maximal edge sets containing an edge from every node in the subset of the graph in *G*. Clearly, the above definition implies that the relationship between the number of nodes and their corresponding edges in a clique is not linear but it makes a parabolic plane curve. The main problem we consider in this study is as following:

M (*G*): *Given a graph G, enumerate all maximal cliques within G.*

The analogous problem *M*(*G*) of enumerating maximum clique graphs was considered by Ramsey [1] .

Theorem. The problem of similarity relationship *(S)* among various observations *(O)* extracted from clustering *(C_l)* can be reduced to determination of maximum cliques *M(G)* within a graph *G*. This problem can be solved in incremental polynomial time.

Proof. The proof of theorem is based on a nice characterization of modeling cluster data, which may be of independent interest. The method used for enumeration is the subrgraph method. Let $X = [(A_1, B_1), (A_2,B_2), (A_3,B_3)]$. Each object pair (A_i, B_i) is representing documents with calculated similarity in the same cluster. Let ŋ= $\int_1^3 X dx$ represent the area of this bound. It forms a linear plane conforming the relatedness of object pair. This shows that a path exists among these pairs as well. Such path existence leads to the formulation of a clique within a graph. This shows that the similarity between documents can be reduced to clique identification problem. This also leads to the property that the graph is strongly connected because each node is having connectivity.

In the last decade, very large number of data in earth science has been generated. The growing body of this large scale data is difficult for interpretation and analysis. Data mining and machine learning techniques such as cluster analysis, graph mining, classification and artificial neural networks have been successfully deployed to the problems of feature extraction, segmentation, data modeling and its validation in various domains. Ronald Caose (1991 Nobel-winning economist) has rightly said "*If you torture the data long enough, Nature will confess*". However this torturing, tweaking and mining of large magnitude of data is not straightforward because the size and complexity of data in every field of knowledge has transcended the limits of conventional analysis tools in terms of capacity and efficiency. Data mining has emerged as a technology to squeeze useful knowledge out of avalanche of data. Data mining can be defined as the nontrivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data [2]. Researchers have divided the data mining process into a sequence of steps. First and foremost step is the analysis of the problem domain to identify the objective of the problem and its solution. Secondly, data gathering and exploration is performed for understanding and verification of the data and its quality. In this step, removal of noise, outlier data and adjustment of the missing values is carried out. In the third step, relevant data set is extracted while pruning the unconcerned data. During extraction of relevant data, transformation may also be required to reduce the dimension of the data set. Among various data mining techniques, a relevant algorithm is employed to discover knowledge out of the piles of data set obtained from the previous steps. The representation of this knowledge can be in different shapes like association rules, decision trees or clusters. In the last step, these representations of knowledge are investigated in terms of the domain knowledge for their evaluation.

With the rapid increase in human population and industrialization, biological diversity and ecosystem integrity has faced many threats from natural calamities as well as deliberate human actions. These threats have motivated the research community for deployment of computer-oriented innovative approaches and techniques to assess timely, precise, accurate, and spatially explicit information related to environmental risks for environmental protection [3]. Manley et al., [3] has described that an increasing demand on public sector organizations and conservation practitioners for real-time information dissemination and decision support is particularly required to extenuate the effects of environmental threats and human impacts on the sustainability of natural resources and biodiversity.

The motley of species plays a regulatory role in an environmental and ecological system. A species falls under the category of being endangered when its race is jeopardized with extinction. Extinction is a natural process since beginning of time. Family of dinosaurs, tyrannosaurus is a good known example of extinction. Unluckily, a very large number of plants and animals are on the verge of their extinction due to factors directly related to humans. These catastrophic facts include habitat destruction, introduction of non native organisms, and direct killing including over harvesting and poisoning practices. However, the question arises why we do make efforts to save endangered species as if extinction is naturally inevitable. The answer lies in the fact that there is very large number of plants and animals encountering crucial role in sustainability of human life. Humanly endeavors in saving endangered species directly impact on the quality of human life as well.

Biological diversity can be defined as the variety of organisms found on earth with particular characteristics including range of species, genetic variance within each species, and the associated prominent attributes related to the ecosystems. The advent of modern human civilization has brought loss of species and also their habitats. Such loss is a major cause of disturbing potential equilibrium underlying ecological system. O'Riordan and Kleemann [4] has pointed out that approximately more than one thousand species per year have been observed to be diminishing while the fossil record revealed this number not more than four. This is a clear indication that extinction rate is seriously higher now. A report from UNEP [5] revealed that there are approximately 13 million species in existence however only 15% of them have been identified in detail. Out of these organisms plants, fish, birds and mammals have been mostly discussed. It has also been reported that proliferation in human activities have a dire negative impact in escalating the diminishing process of other species at the rate of 18% mammals, 11% birds, 8% plants and 5% fish [6].

Steinberg [7] described that during the last two decades, a serious and growing concern has been observed centric towards the emergence of widespread environmental consciousness under the global political, social and environmental stewardship particularly in the developing countries. Public opinion polls over the internet or other electronic media are a direct means for measuring the enhancing realization of the environmental protection. This has resulted that in developing countries an effective stratum of citizens are now pleading environmental causes. The records show that in under developed countries, there is a feeble connection found in the law on paper and its implementation. This raises a question: How can impoverish nations be convinced to embrace environmental concerns so that the public preferences can be translated into ameliorated environmental outcomes? The solution lies in institutional initiations leading to collective action while mitigating collective indifference helping in bridging the gap between policy effectiveness and public interests.

In this study, we have highlighted that there is a need to understand the value of endangered species to humans with respect to different political and geographical regions of the world. We have illustrated that there are numerous territories facing same kind of species extinction problem regardless of their geographical diversification. Such revelation confer that there are common human practices responsible for the extinction problem. This study is focused towards demonstration of how data mining techniques can be applied in the field of earth sciences while proposing a framework for discovery of knowledge. The proposed framework *ESDI* is aimed towards the determination of the scarification and stratification of endangered species into their respective regional strata.

Rest of the paper is organized in four sections. In section 2, we have discussed background scientific literature relevant to this study. In section 3, we have introduced *ESDI* our proposed data mining framework. The usefulness of this framework and its analysis has been illustrated in section 4 followed by concluding remarks in section 5.

II. Literature Review

Dennis et. al, [8] argued that survival or extinction of an endangered species is inherently stochastic. They proposed a stochastic model of exponential growth for inferring quantities related to growth rates and extinction probabilities using time series data. Their model was inspired from the biological theory of age or stage-structured populations. The model corroborates the so-called environmental type of stochastic fluctuations and producing a lognormal probability distribution of population abundance. They employed linear regression to evaluate fitness of given dataset calculating maximum likelihood estimates of two unknown parameters. They showed that numerous growth and extinction related attributes are functions of two parameters. However they provided their result to only a few endangered species. Moreover, their model did not consider the possibility of freak catastrophic events like hurricanes, fires, etc. whereas these catastrophic events are usually a robust peril to the species considered by their model.

David and Stockwell [9] introduced an ecological niche modeling algorithm *WhyWhere* for mapping the species distribution. The algorithm employs image processing techniques to effectively sort out large dataset to identify few variables responsible for better prediction of species occurrences. They justified the prime factor in parameterization indicating preliminary success at quickly yielding accurate, scalable and simple models. They showed 14% accuracy increase over another algorithm using two variables on six species. David and Stockwell [9] illustrated that data mining based techniques yields particularly improved results for finding correlations in large datasets in ecological niche modeling domain.

Johan, Loomis and Douglas [10] presented the regression analysis of a factor "willingness to pay" related to the funds for endangered species. They used the parameters such as: annual vs. monthly, one time vs. regular payment, visitor vs. users of donation systems, marine mammals vs. birds species. They established the regression analysis to provide meaningful estimates of anthropocentric benefits of preserving rare and endangered species. They made economic techniques available to perform broad-based benefit-cost analyses of species preservation showing whether the costs are likely to be disproportionate to the benefits.

Lewina and Smolin [11] performed principal component analysis while indicating positive and negative correlation between some specific environmental variables with their impact on a rare and endangered species mollusc. They considered the environmental attributes related to pollution of water by coal mining. Their PCA analysis revealed the invasion of some other species. PCA analysis pointed out a positive correlation between mollusc density and pH value, the concentration of chlorides, the total hardness, alkalinity and total dissolved solids, and a negative correlation between the number of species and phosphates. In this way, Lewina and Smolin [11] exploited the data mining techniques expressing distinctive environmental features of the mining subsidence reservoirs providing a refuge for wildlife.

Wong et. al, [12] proposed a versatile web based decision support system meeting the requirement of Canadian government legislation (2003) to conserve biodiversity. Their system was aimed towards storing; retrieving and interpretational information on species and their critical habitats linking distributed data sources into an integrated system which can manages data providing decision support. They described that their system can provide an effective platform for the delivery of information and services to practitioners of Species at Risk enabling improved decision making employing data mining and modeling functionality.

Domask [13] argued the environmental and ecological policies critically. They described that there is an adequate lack of scientific foundation in currently prevalent policies in many African countries such as Ethiopia, Mali, Zimbabwe, Guinea, and Trinidad? They highlighted that most of the policies are void of influential, mainstream, scientific thinking turning these policies questionable in terms of their credibility and applicability. They pointed out repeated mismatches between policy positions and local realities in the wake of simultaneous trends of globalization, decentralization, and localization interaction with the scientific, managerial practices and the policy processes. Top down approaches were suggested to be reinforced for the renovation of policies.

III. Proposed Framework

From the literature review and investigation of useful application of data mining in various domains, we have concluded that data mining can be successfully applied in this domain. We raised the following research questions in this study:

1. Is it possible to induce most coherent itemset Γ {I_1, I_2, I_3…..I_n} from clustering model € ~ *(S, Cl)* reducible to graphical model $\mathcal{D}\rho$ (v , e). ?
2. What is the significance of groups of territories identified as cliques £ from the graphical model $\mathcal{D}\rho$ (v , e) in perspective of the endangered species?

In order to investigate the above two research questions, we devised a framework comprising of various data mining methodologies. We have divided the framework into many steps in order to deliver an in depth insight to our shrewd readers.

3.1 Collection of Raw Data

We collected the dataset from Earth's Endangered Creatures [14] in form of html pages. A tag parser was developed to extract the relevant data. All of the html pages were combined into a single page while trimming the unwanted stuff from the web pages.

3.2 Pre-processing

In this step, we dealt with the process of data cleansing, feature/data selection followed by data transformation. Basic operations in data cleansing and preprocessing include the removal of noise, handling missing data fields. Fortunately there was no major inconsistency found in data cleaning phase. However duplicate records were adjusted accordingly. The single html page we obtained from the parser contains the features: Species Name, Scientific Name, Group, Range. The feature Range describes the regions in which the specific species were related to. We omit this feature because we separately download all of the html pages for each of the territory. During combining of these web pages into a single web page, the information of the territory was separately conserved resulting in a new feature from existing feature set. Out of these four candidate features, we considered only two useful features: Species Name and Group. Range was already covered by the parser. Data transformation is concerned with mapping data from one form to another. In our case the objective was to prepare the data for text mining. This leads to generating document for each country/ territory.

3.3 Text Mining

In this step we performed text mining on all of the documents obtained from previous step. Term Frequency, Inverse Document Frequency is a famous algorithm usually used in text mining. tf.idf states that if there are many documents then frequency of each word / term is calculated in specific document and also in the entire document. This calculated result set make a sense of vector for each document. Cosine similarity can be applied to the set of these vectors to obtain the correlation effect between documents. Its result always ranges from 0 to 1 where 1 shows maximum correlation and value of 0 indicates minimum correlation between the documents. Cosine Similarity is formally defined as below [15]:

$$Cos(A,B) = \frac{A \bullet B}{\| A \| \times \| B \|} \qquad (1)$$

$$A \bullet B = \sum_{i=1}^{n} A_i \times B_i \qquad (2)$$

$$\| A \| = \sqrt{\sum_{i=1}^{n} (A_i)^2} \qquad (3)$$

$$\| B \| = \sqrt{\sum_{i=1}^{n} (B_i)^2} \qquad (4)$$

Where A and B denotes two document vectors.

Another similarity measure applied over the document vector to find the correlation effect is Jaccard Coefficient Index [16]. We can define it as the ratio of the intersection to union of the two sets. As compared to cosine similarity, it caters the frequency of the each term involved. Figure 1 shows it graphically:

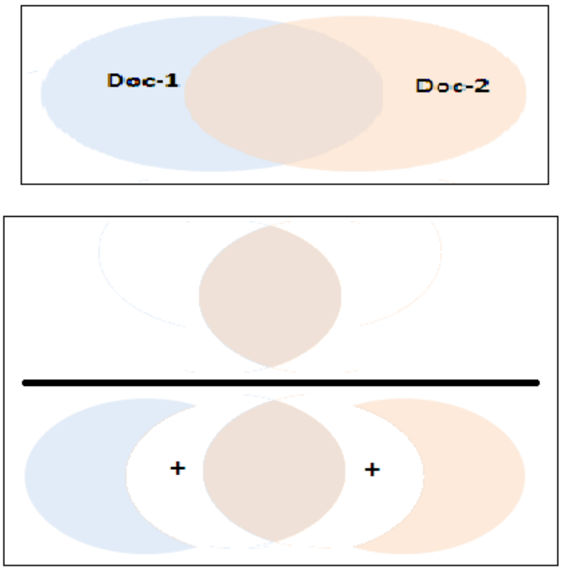

Fig. 1: Jaccard Coefficient in set theory representation

3.4 Clustering

In literature, cluster has been defined as the internal homogeneity while keeping the external separation [17]. Both the similarity and the dissimilarity in clustering are be examinable in a sense of inter-similarity and intra-similarity quantification. Here, we provide some simple mathematical descriptions of clustering used in our approach based on the description in [17].

Given a set of features $F = \{F_1, F_2, \ldots. F_j \ldots\ldots F_n\}$, where $F_j = [f_{j1}, f_{j2}, \ldots. f_{jd}]^T$ є $Я^d$ while every measure f_{ji} is known to be an attribute or dimension variable. K number of partitions of *F* is declared:

- C = {C1, C2, …, Ck}. Where $k \le n$. Every cluster vCi = Ǿ, i ≤ k.
- $\bigcup_{i=1}^{k} C_i = F;$
- $C_i \cap C_j = \phi, i, j \le k, \forall i \neq j$

The underlying distance measuring parameters are numerous. In our case, we used Euclidean, Manhattan, Minkowski and Tchebyschev distance measure. Mathematical detail of which can be described as:

$$Euclidean\ (A, B) = \sqrt{(A_1 - B_1)^2 + (A_2 - B_2)^2} \qquad (5)$$

$$Manhattan(A, B) = \sum_{i=1}^{n} | A_i - B_i | \qquad (6)$$

$$\text{Tchebychev}\,(\text{A, B}) = \max(|\,\text{A}_2 - \text{B}_2\,|, |\,\text{A}_1 - \text{B}_1\,|) \qquad (7)$$

$$Minkowski\,(A, B) = \lim_{p \to \infty} \left(\sum_{i=1}^{n} | A_i - B_i |^p\right)^{\frac{1}{p}} = \max_{i=1}^{n} | A_i - B_i | \qquad (8)$$

Where A and B denotes the member dimension of observation points for which clustering is under consideration. The set of d-dimensional vector in this study was cosine and jaccard similarity coefficient measure. We provided a broad range of seeds ranging from 3 to 99 for each distance measure described above. Each test was performed four times. This results in a very large number of dataset whose regression analysis was made.

3.5 Regression Analysis

In the previous steps, we obtained 4x97x4 = 1552 result set comprising of number of seeds and number of scans at which the centroid was converged. In regression we performed learning a function mapping data item to a real-valued prediction variable discovering the functional relationships between seeds and scans.

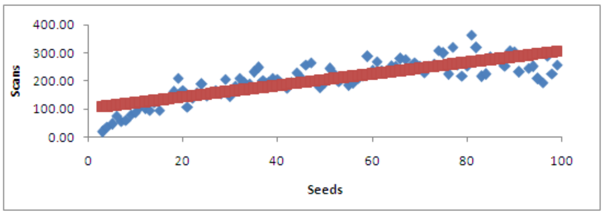

Fig. 2: Euclidean Line Fitting Plot (Actual vs. Predicted)

Figure 2, 3, 4 and 5 represent the graph of linear fitting between actual and predicted scans against number of seeds provided. The figures have been plotted using all of the distance measures we used in our experiment. The equation for linear fitting for figure 2 is computed as:

$$Y = (100.2956751)\ \ X + 2.052065801 \tag{9}$$

The equation 9 is showing the predicted output for scans. If the drawn line is indicating an increase in seeds causing increase in scans then a positive relationship exists between both of the variables. An increase in independent variable causing decrease in dependent variable indicates that both of the variables have a negative relationship where as in any other case no relationship between both of the variables is found.

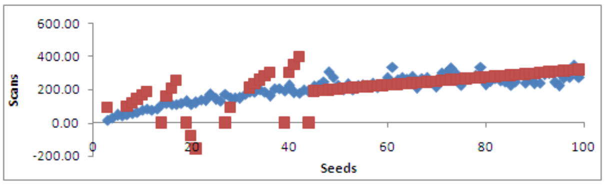

Fig. 3: Manhattan Line Fitting Plot (Actual vs. Predicted)

This is a usual observation that a deviation persist between the actual and predicted observations. The deviation of the actual observations from the predicted plotted line obeying the linear equation is calculated by means of squared error. It is value ranges from 0 to 1. A value of 1 indicates that the observed and predicted values overlaps and value of 0 indicated that both of them have no strong relationship.

Value of change in dependent variable (Scans) can be explained by the change in its corresponding root square. Among all of the four distance measures, these values are 0.683216259, 0.758937, 0.692936 and 0.398407792 respectively.

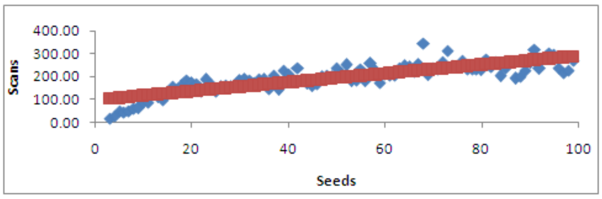

Fig. 4: Tchebyschev Line Fitting Plot (Actual vs. Predicted)

$$Y = (82.9094585)\ \ X + 2.384027193 \tag{10}$$

$$Y = (96.35281664)\ \ X + 1.925717968 \tag{11}$$

$$Y = (221.6724043)\ \ X + 3.325202504 \tag{12}$$

Equation 10, 11 and 12 are predicted linear equations for figure 3, 4 and 5 respectively. A careful observation of these equation and root squared value indicate that Manhattan distance measure was best suited yielding highest confidence and least error while Minkowski distance measure provided least confidence with values of approximately 76% and 40% respectively.

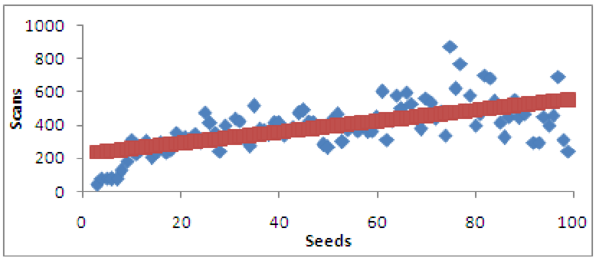

Fig. 5: Minkowski Line Fitting Plot (Actual vs. Predicted)

3.6 Graphical Modelling

In the literature it has been shown that every graph with n vertices and minimum vertex degree δ must have a maximum clique of size at least $\lceil n/(n-\delta)\rceil$ [1]. Moreover this condition is known as the best possible in terms of n and δ [1]. In the previous step, we obtained 1552 result each of which contain 3 to 97 clusters. In graph modeling, we realize these clusters into a graph such that each vertex corresponds to a country. A link exists between two countries if both of them are found in the same cluster. Our objective in this step is to identify the maximum possible cliques. A polynomial time algorithm for finding maximal cliques in graphs was implemented in this part.

IV. Experiment and Analysis

We implemented many data mining techniques, purpose of these applications was tweaking and torturing the data enough to make data confess and yield some useful knowledge either in form of patterns or rules.

Table 1: Sample Dataset of Endangered Species

Territory	Species Name	Group
Albania	Albanian Water Frog	Amphibians
Algeria	Algerian Nuthatch	Birds
Angola	Bellamya monardi	Snails
Antigua	Nectandra krugii	Plants
Oman	Sooty Falcon	Birds
Pakistan	Bigeye Tuna	Fishes
Pakistan	Argali	Mammals
Turkey	Saker Falcon	Birds
USA	Wreathed Cactus Snail	Snails
USA	Earthworm	Worms
Venezuela	Calf Frog	Amphibians
Vietnam	Cycas nongnoochiae	Plants
West Indies	Maria Island Snake	Reptiles
Yemen	Gyraulus cockburni	Snails
Zimbabwe	White Rhinoceros	Mammals
Zimbabwe	African Mahogany	Plants

Table 2: Document Similarity Measures between Territories

Territory		Similarity Measure	
1	2	Cosine	Jaccard
Pakistan	Morocco	0.0581469	0.155495979
	Namibia	0.042059082	0.099744246
	Nepal	0.076083441	0.25382263
	Netherlands	0.042376611	0.080291971
	Nicaragua	0.148638465	0.099547511
	Niger	0.045489509	0.110701107
	Norway	0.054567353	0.090909091
	Iran	0.496098037	0.415662651
	Oman	0.511010224	0.368589744
	Paraguay	0.012520601	0.033419023
	Russia	0.035783591	0.127020785
	Poland	0.029935936	0.097087379
	Puerto Rico	0.016337733	0.047085202
	Qatar	0.48187056	0.28685259

Table I is a sample dataset of endangered species which was pre-processed out of web pages retrieved from Earth’s Endangered Creatures [14]. This dataset contains groups including Amphibian, Birds, Snails, Plants, Fishes, Mammals, Reptiles, Insects and Corals etc. We did not consider these classes of groups individually but for a single territory all of the endangered species were considered. We have considered 175 countries from Earth's Endangered Creatures [14].

Text mining application were applied on all of the documents from Table I using term frequency and inverse document frequency. The document vectors obtained were subjected to cosine and jaccard coefficient similarity. A sample of the results of the text mining similarity measure is depicted in Table II in which we have shown similarity of Pakistan.

Table 3: Cliques identified from K-Mean using Manhattan Distance Measure

Cliques (Manhattan)
Eritrea, Maldives, Singapore, Tuvalu, Wallis and Futuna
Eritrea, Maldives, Qatar, Sudan, Tuvalu
Japan, Nauru, Solomon Islands, Thailand, Taiwan
Kenya, Kuwait, Philippines, Vietnam, Yemen
Kenya, Kuwait, Philippines, Vietnam, Yemen
Eritrea, Maldives, Saudi Arabia, Tuvalu, Wallis and Futuna
Bulgaria, Poland, Russia, Ukraine
Belgium, Iceland, Ireland, Netherlands

The dataset obtained from the calculation of cosine and jaccard similarity was subjected to pruning rate of 19.44% which was taken as the average of jaccard and cosine similarity. This was a mandatory requirement in order to prevent those entries in which weak similarity was observed. The pruning reduced this dataset of 9666 to 1600 records only. This dataset was a good candidate for application of K Mean clustering. We applied clustering algorithm using four different distance measures. It was analyzed that in each cluster, there is a sense of cliques if graph modeling is applied on these clusters. Table 3 is depicting the result of cliques produced out of K Mean clusters using Manhattan distance.

Table 4: Cliques identified from K-Mean using Tchebyschev Distance Measure

Cliques (Tchebyschev)
Cook Islands, Japan, Nauru, Philippines, Thailand,
Japan, Papua New Guinea, Somalia, Wallis and Futuna
Burma, Mauritius, Sudan, Tuvalu
Mayotte, Oman, Solomon Islands, ,Yemen
Cook Islands, Maldives, Philippines, Qatar
Egypt, Marshall Islands, Taiwan, Yemen

Clique 1 from table 3, 5, 6 and clique 2 from table 4 has been explained in figure 6. A Careful analysis of table 3 and figure 6 indicates that Asian, African and European countries mostly were found in their own clique groups. This is in accordance with the natural geographical classification of the area. However this is not same at all the cliques. Table 4 and Table 5 also corroborate this fact in which Japan and Somalia has been placed in same clique. Multiple similarity measures have placed some geographically distant countries in the same clique. There are many underlying reasons. Blickley and Patricelli [18] described one of such reasons that the noise introduced due to human activities in our ecosystem has a direct impact on wildlife. Most well known noise effects include habitat fragmentation and introduction of invasive species. The problem of noise is more concerned at developed countries. These include the definition and implementation of specific action plans such as: an Alpine action plan addressing the grey squirrel, Mediterranean action plan to cope with biological invasions on islands, a Baltic policy related to the treatment of ballast water [19].

1	Eritrea
2	Japan
3	Maldives
4	Papua New Guinea
5	Saudi Arabia
6	Singapore
7	Somalia
8	Tuvalu
9	Wallis and Futuna

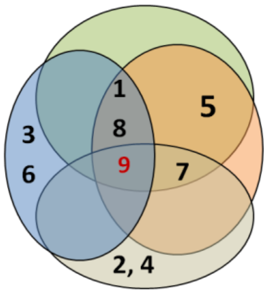

Fig. 6: Clique Cover Explanation

We shall discuss here one particular case relating to South Africa. The country is comparatively much developed in comparison to other African countries. The analysis of cliques reveled that South Africa has shown poor similarity to other African countries. The underlying reason behind this fact is that the ecology and wildlife preservation policies in South Africa are rooted in rural ideas about how to advance the effective use of land without minimum torturing of ecology as reported by Smith et al, [20]. The prevailing practices in southern African countries has long root in an appreciable vision of wildlife.

Table 5: Cliques identified from K-Mean using Euclidean Distance Measure

Cliques (Euclidean)
Eritrea, Saudi Arabia, ,Somalia, Tuvalu, Wallis and Futuna
Eritrea, Maldives, Qatar, Sudan, Tuvalu
Eritrea, Oman, Qatar, Sudan, Tuvalu
Japan, Marshall Islands, Papua New Guinea, Somalia, Wallis and Futuna
Burma, Japan, Solomon Islands, Taiwan, Vietnam
Czech Republic, Bulgaria, Poland, Ukraine

Table 6: Cliques identified from K-Mean using Minkowski Distance Measure

Cliques (Minkowski)
Eritrea, Saudi Arabia, Tuvalu, Somalia, Wallis and Futuna
Eritrea, Maldives, Saudi Arabia, Tuvalu, Wallis and Futuna
Eritrea, Maldives, Qatar, Sudan, Tuvalu
Mariana Island, Mayotte, Samoa, Somalia, Wallis and Futuna
Eritrea, Kuwait, Mariana Island, Oman, Yemen
Japan, Marshall Islands, Papua New Guinea, Somalia, Wallis and Futuna
Burma, Japan, Solomon Islands, Taiwan, Vietnam
Burma, Japan, Solomon Islands, Taiwan, Vietnam
Eritrea, Mariana Island, Mayotte, Oman, Yemen,
Japan, Mayotte, Samoa, Solomon Islands, Somalia

Although Europe is qualified by territorial continuity, however noteworthy biodiversity as well as geographical differences have been observed in terms of species, subspecies, populations and ecosystems. Such biodiversity based approach can be peculiarly useful to illustrate the development of methods to control the policies regarding endangered species. The countries are required to allocate sufficient resources to endangered species issue. Moreover they are required to embrace revised national legal frameworks to eliminate obstacles to the actions compulsory to address the biological invasions; co-ordination mechanism to deal collection and circulation of information, authorization processes and the implementation of mitigation measures, including eradication plans; raise awareness of the endangered species issue; and cooperation to develop and apply a comprehensive policy.

V. Conclusion

We conclude that the proposed framework model and its associated data mining similarity measures can be useful for investigating various scientific and management oriented questions related to protection of endangered species with emphasis on collaboration among regional countries. The rationale behind the proposed approach is that the countries which have been grouped into same clique inherit a lot of argues illustrating common reasons of their struggles towards ecological safety with minimization of perils for endangered species. The development and implementation of a regional approach based on this similar grouping address the actions that could offer significant benefits in achieving their goal for ecological policies. Other critical actions at this clique level include fortifying and elevating harmonization of legal frameworks with emphasis on prevention procedural issues; awareness realizations of endangered species issues and its priority. Such actions will eventually lead towards implementation of essential plans fulfilling co-operative expertise and common endeavors.

References

[1] Ramsey F.P, On a problem of formal logic, [C]. Proc. London Math. Soc., 1930.

[2] Fayyad U.M., Piatetsky-Shapiro G. P., and Smyth, From data mining to knowledge discovery: an overview, Advances in Knowledge Discovery in Data Mining[C]., AAAI Press, Menlo Park, CA, 1996, pp. 1 –34.

[3] Manley, P.N., Zielinski, W.J., Schlesinger, M.D., Mori, S.R., 2004. Evaluation of a multiple-species approach to monitoring species at the ecoregional scale[J].. Ecological Applications 14 (1), 296e310.

[4] Tim O'Riordan & Susanne Stoll-Kleemann, Biodiversity[J]., Sustainability and Human Communities 14 (2002).

[5] United Nations Environmental Program (UNEP) [M]., Global Biodiversity Assesement (1995).

[6] Dennis Pirages and Theresa DeGeest, Ecological Security 141 (2004) [M]., Studies report little on the impact of habitat loss on protozoa, nematodes, and other micro-organisms.

[7] Paul F. Steinberg (2005): From Public Concern to Policy Effectiveness: Civic Conservation in Developing Countries[J]., Journal of International Wildlife Law & Policy, 8:4, 341-365

[8] Brian Dennis, Patricia L. Munholland, Michael Scott, Estimation of growth and extinction parameters for endangered species[J]., Ecological Monographs, 61(2), 1991, pp. 115-143.

[9] David R.B. Stockwell, Improving ecological niche models by data mining large environmental datasets for surrogate models[J]., Ecological Modeling 192 (2006) 188–196.

[10] John B. Loomis, Douglas S. (1996), White, Economic benefits of rare and endangered species: summary and meta-analysis[J]., Ecological Economics 18 (1996) 197-206.

[11] Iga Lewina, Adam Smolin´ skib, Rare and vulnerable species in the mollusc communities in the mining subsidence reservoirs of an industrial area (The Katowicka Upland, Upper Silesia, Southern Poland) [J]. Limnologica 36 (2006) 181–191.

[12] Wong I.W., Bloom R., McNicol D.K., Fong P., Russell R., Chen X., Species at risk: Data and knowledge management within the wildspace Decision Support System[J]. Environmental Modeling & Software 22 (2007) 423-430,

[13] Joseph Domask (2004): Science, Policy Process and Policy Ownership in Africa and The Caribbean[J]., Journal of International Wildlife Law & Policy, 7:3-4, 223-232

[14] Glenn, C. R. 2006. "Earth's Endangered Creatures" (Online). Accessed Aug-2012 at http://earthsendangered.com.

[15] Tan, Pang-Ning; Steinbach, Michael; Kumar, Vipin (2005), *Introduction to Data Mining* [M]., ISBN 0-321-32136-7.

[16] Jaccard, Paul (1901), Étude comparative de la distribution florale dans une portion des Alpes et des Jura [M]. *Bulletin de la Sociéé Vaudoise des Sciences Naturelles* 37: 547–579.

[17] Hansen P. and Jaumard B., Cluster analysis and mathematical programming[J]., *Math. Program.*, vol. 79, pp. 191–215, 1997.

[18] Jessica L. Blickley and Gail L. Patricelli (2010): Impacts of Anthropogenic Noise on Wildlife: Research Priorities for the Development of Standards and Mitigation[J]., Journal of International Wildlife Law & Policy, 13:4, 274-292

[19] Council of Europe/UNEP, Pan - European biological and landscape diversity strategy, Invasive Alien Species[J]., Journal of International Wildlife Law & Policy, 5:3, 291-305.

[20] Geoffrey Wandesforde-Smith , Nicholas S.J. Watts and Arielle Levine (2010): Wildlife Conservation and Protected Areas: Darwin, Marx, and Modern Science in the Search for Patterns That Connect[J]., Journal of International Wildlife Law & Policy, 13:4, 357-374

Authors' Profiles

Muhammad Naeem: Research scholar at department of computer science, M. A. Jinnah University Islamabad Pakistan. His research area is machine learning, structure learning and data mining.

Sohail Asghar: Associate Professor at University Institute of Information Technology, PMAS-Arid Agriculture University Rawalpindi Pakistan. His interest area includes data mining and decision support system

How to cite this paper: Muhammad Naeem, Sohail Asghar,"Knowledge Discovery in Endangered Species Diversification", International Journal of Information Technology and Computer Science(IJITCS), vol.5, no.2, pp.57-65, 2013.DOI: 10.5815/ijitcs.2013.02.06