An Ontology-Based Data Mining Approach for Predicting the Research Ideas using Past Research in the Wildlife Sector of Sri Lanka

P Premisha[1*], BTGS Kumara[2], EP Kudavidanage[3,] and K Banujan[4,]

[1,2,4]*Department of Computing & Information Systems, Faculty of Applied Sciences, Sabaragamuwa University of Sri Lanka*
[3]*Department of Natural Resources, Faculty of Applied Sciences, Sabaragamuwa University of Sri Lanka*
[*]*ppremisha@std.appsc.sab.ac.lk*

Abstract— *Sri Lanka being a global biodiversity hotspot, places great value for biodiversity owing to ecological, socio-economic, and cultural factors. However, the wildlife of Sri Lanka is critically threatened due to several factors, mainly human activities and needs dire conservation measures. Inadequate knowledge and technical support also hinder wildlife management activities. Findings of wildlife research studies could be integrated into data-driven conservation and management decisions but the current contribution is not satisfactory. .This research work shows a novel data mining approach for finding hidden keywords and automatic labeling for past research work in this domain. We used Latent Dirichlet Allocation (LDA) algorithms to model topics and identify the major keywords also developed an ontology model to represent the relationships between each keyword. Both approaches are also useful for potential research ideas, to identify research gaps and can classify the subjects related to a publication by non-professional related fields. The experiment results demonstrate the validity and efficiency of the proposed method.*

Keywords— **wildlife, LDA, ontology, topic modeling**

I. INTRODUCTION

Wildlife is critical for the sustenance of life on earth. Conserving biodiversity is critical to maintaining a healthy ecological balance in the world.

Sri Lanka is a global biodiversity hotspot consisting of a large variety of fauna and flora. It is one of the main sources of income generation through tourism and other means. The diversity of ecosystems is primarily due to its topographical and climatic heterogeneity, as well as its coastal effect (L.P.Jayatissa, 2012). This rich biodiversity is threatened due to unplanned land use, pollution, overexploitation, etc.

Data from wildlife research can contribute to a large extent is proper conservation and management. However, there is a gap between research and application. Most of the existing research work is not converted into applications while there are many data gaps. Limited numbers of researchers are focussing on the actual research needs from conservation. The selection of research topics is often not compatible with the actual research needs due to multiple reasons. This is a disheartening scenario as there are plenty of opportunities for such work. Inadequate knowledge of the existing research and their applicability, inadequate use of technology, and inability to locate some research are some of the contributing factors. Other than the research published in a known journal, some past research information available online cannot be found properly because they belong to conventional archives, unfortunately.

Increasing public awareness on the values of wildlife and the consequences of losing this heritage can assist conservation to a large extent. To achieve this, we have to simplify the gap between the public and the accessibility to information on wildlife. Technology can play a major role in filling the gap between them.

Mostly wildlife studies aimed to understand species diversity, behavior, and habitat use, and ecology, the role of wildlife in disease transmission, species conservation, population management, and methods to control threats to diversity.

In our study, we concentrate on reviewing past research papers using data mining techniques to provide potential research ideas that can be conducted in the future. To fill the data needs for conservation our solution focuses primarily on semi-automating the finding of research gaps through abstract analysis. Finally, the model includes the most commonly used keywords and question top. This will be a vital milestone for researches as well as wildlife activists to give an eye on recent problems that need a solution urgently.

In technological perspective there was prior work (Zhu, Klabjan and Bless, 2017)(Adhitama, Kusumaningrum, and Gernowo, 2018)(Chowdhury and Zhu, 2019) has shown hierarchical relationship-based latent Dirichlet allocation (hrLDA), a data-driven model of hierarchical topics to derive terminology ontology from a large number of heterogeneous documents. Unlike conventional topic models, hrLDA relies on noun phrases instead of unigrams, considers syntax and text structures,

and enriches topic hierarchies with topic relations. Through a series of experiments, we are demonstrating hrLDA's superiority over established topic models, particularly for hierarchy building.

So we have to deviate past research techniques to come up with our final solution. Some trending techniques used here to improve the outputs. Our research aims to resolve the inadequate application of wildlife research and technologies in the decision-making process.

II. METHODOLOGY

In our research, we used a semi-automated methodology using LDA and Ontology. The text data of the defined domain were collected and pre-processed for the input to LDA algorithms then compared with the ontology graph to the final output. The steps of our methodology defined below.

A. Data Collection

We collected information about past wildlife researches in Sri L Figure 1.Methodology the aid of the Department of Natural Resources, Sabaragamuwa University of Sri Lanka, and an extreme literature survey. After that, we obtained full research papers of selected papers from each domain. We've selectively applied the title and abstract data to the CSV file from those research papers.

B. Data Pre-processing

Data pre-processing is so important because if our data set contained mistakes, redundancies, missing values, and inconsistencies that all compromised the integrity of the set, we need to fix all those issues for a more accurate outcome (Editors, 2019). We performed the following steps:

- Tokenization: Divide the text into sentences, and the sentences into words. Lower case the words and smooth punctuation
- Stop word removal: Delete words that have fewer than 3 letters. All stop words are removed.
- Lemmatizing: Words in the third person are shifted to first-person and verbs shifted to present from past and future tenses.
- Words are stemmed — words are reduced to their root form

C. Topic Modelling-LDA

LDA helped rework the textual data into a format that could act as an input to the LDA model for training. We began by converting the documents to a simple representation of the vectors as a group of words called Bag of Words (BOW) (Rani, Dhar, and Vyas, 2017). First, we translated a list of titles into vector lists, all with vocabulary-capable lengths.

The topic model described in Figure 2 is one of the unsupervised methods; that is, it is a technique of text mining in which the topics or themes of documents can be extracted from a larger collected corpus of documents (Lee *et al.*, 2018). LDA, one of the most popular modeling techniques, is a probabilistic model of a corpus-based on Bayesian models. This is often considered a probabilistic extension of Latent Semantic Analysis (LSA). The LDA's basic idea is that each document has a word distribution that can be defined as.

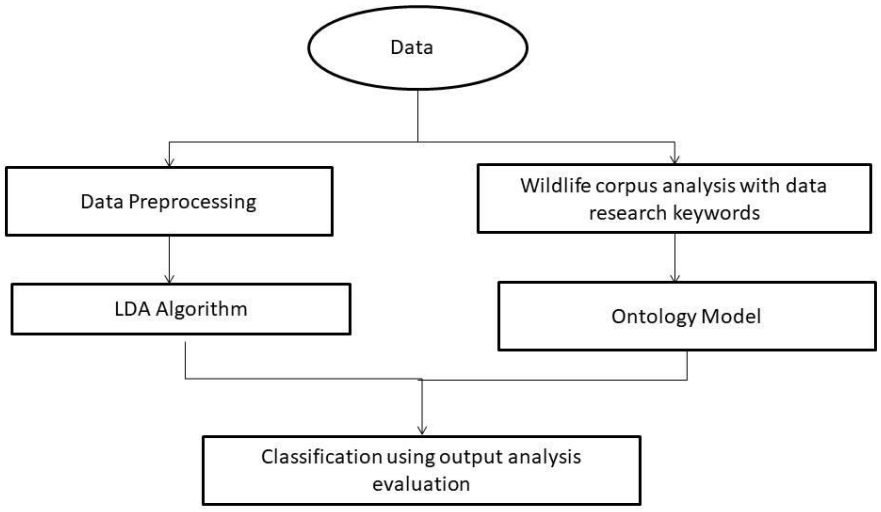

Figure 1. Methodological framework

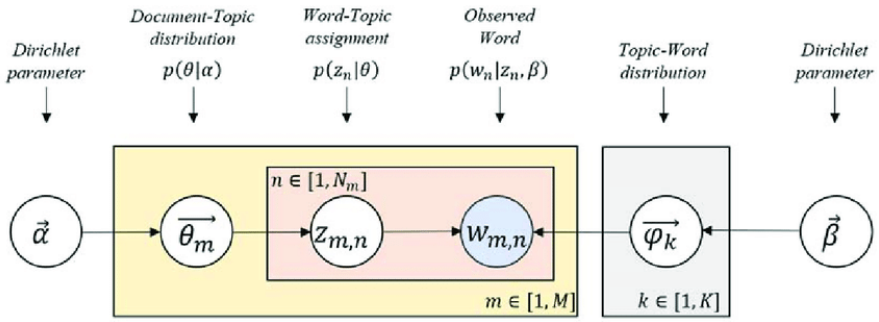

Figure 2. Graphical model for LDA

D. Ontology

Ontologies contain features such as general vocabulary, reusability, machine-readable content, as well as semantic searching, enabling agent interaction, and ordering and structuring information for the Semantic Web application (Movshovitz-Attias and Cohen, 2015). Automated learning is the issue in ontology engineering, such as the lack of a fully automated approach from a text corpus or dataset of different topics to form ontology using machine learning techniques is still present.

The ontology model was finalized using protégé tools, which is the most popular tool of ontology visualization (Hussein *et al.*, 2020). The Protégé 5.5.0 tool is being applied for further development in various disciplines for a better understanding of knowledge with the aid of domain professionals in the wildlife.

E. Comparison

Our interactive, web-based visualization framework, LDAvis, has two key functionalities that allow users to understand the topic-term relationships within a fitted LDA model, as well as several additional features that provide additional perspectives on the model (Adhitama, Kusumaningrum, and Gernowo, 2018). First and foremost, LDAvis allows one to pick a topic to report the words most applicable to the subject. We compared the total term frequency to the approximate term frequency for finding the keywords that appear and are most significant.

F. Evaluation

In our research, the output assessment evaluation method was used to analyze the final output to the overall conclusion. We have defined a new approach to automatic ontology learning, and this method has applied the LDA model to generate topics, and the progress of learned ontology does not need the seed of ontology, but only the document corpus (Lin, 2017).

III. RESULTS

The outcome of this paper was described by using abstract past researches that serve as an input in Sri Lanka. For LDA implementation we used python language. The text used as input is interpreted and tokenized, resulting in a compilation of input nouns, adjectives, and verbs. Furthermore, it eliminates all the stop words like papers.

The tokenized and pruned text is then subjected to the LDA modeling algorithm. That gave production as word sets that could collection contain words that are linked to each other. Such collections of words are classified as various subjects. The LDA model approach is used to arrange, synthesize broad corpus, and to retrieve subjects and words.

Figure 3 and Figure 4 are the final visualizations of the LDA model which shows the overall keyword for each research paper and the essential keyword using the pyLDAvis library in python. This output allowed the detection of hidden keywords from every abstract. To get the output of the pyLDAvis method we used the equation of saliency and relevance to accommodate the keyword distributions.

The intertropical distance map is indicated via multidimensional scaling by our LDA output. In CE literature and inter-topic distance, the top 20 salient keywords.

$$Saliency = frequency \times \left[sum\, p(t|w) \times \log\left(p\left((t|w)/ p(t)\right)\right)\right] \quad (1)$$

Where, t- Topic, Frequency (w) –frequency of word w, p (t|w) - conditional probability: the likelihood that observed word w was generated by latent topic t, p (t) - probability of topic t, sum p (t|w) - summation of the probability of observed word w was generated by latent topic t

This formulation (1) defines (in a theoretical context of information Sense) how informative the specific term w, versus a randomly selected word, is for determining the generating subject. For instance, if a word w appears in all topics, observing the word tells us nothing about the topical mixture of the document; thus the word will obtain a score of low distinctiveness. The saliency (Chuang, Manning and Heer, 2012) of a term is defined by the product:

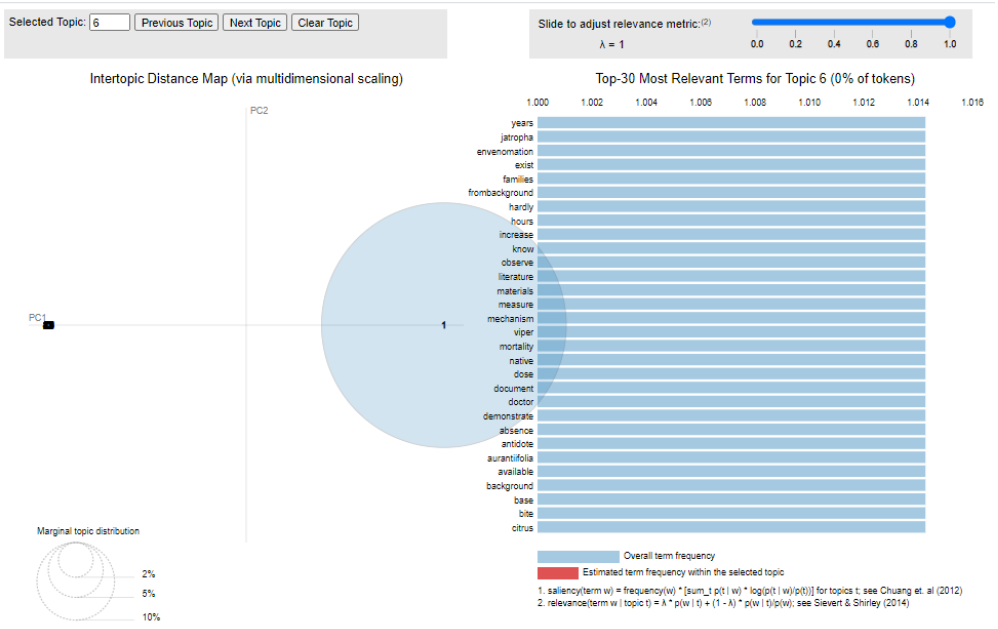

Figure 3. LDA model for overall keyword

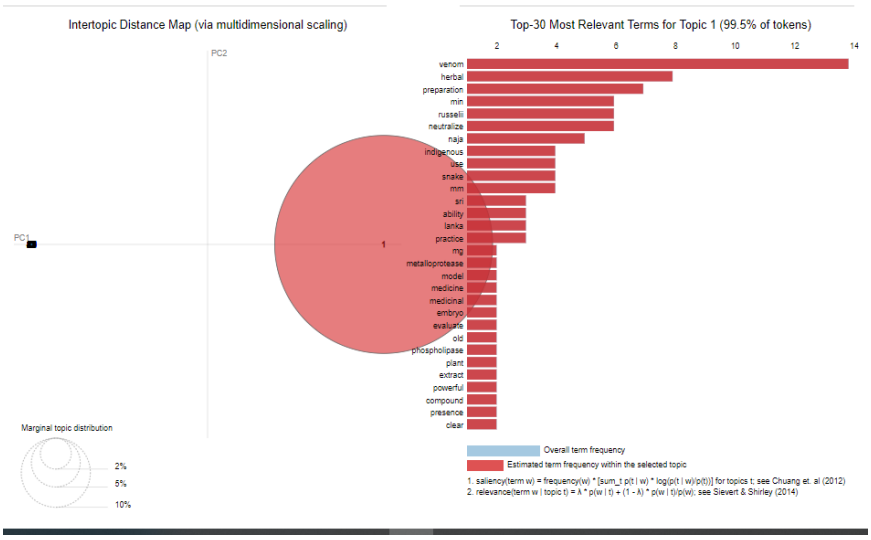

Figure 4. LDA model for estimated keyword

$$Relevance = \lambda * p(w|t) + (1 - \lambda) * p(w|t)/p(w) \quad (2)$$

Where, λ –slide to adjust relevant metric, p (w|t) - conditional probability: the likelihood that observed word w was generated by latent topic t, p (w) –probability of word w (Frasier *et al.*, 2019)

Using this output from LDA we compared the ontology output. Analysed the estimated keywords and their ontology domain formation. The protégé tool used the Sri Lankan wildlife research domain ontology to be developed. The partial view of the final ontology production shown in Figures 5 and 6.

Figure 5.Ontograph partial view

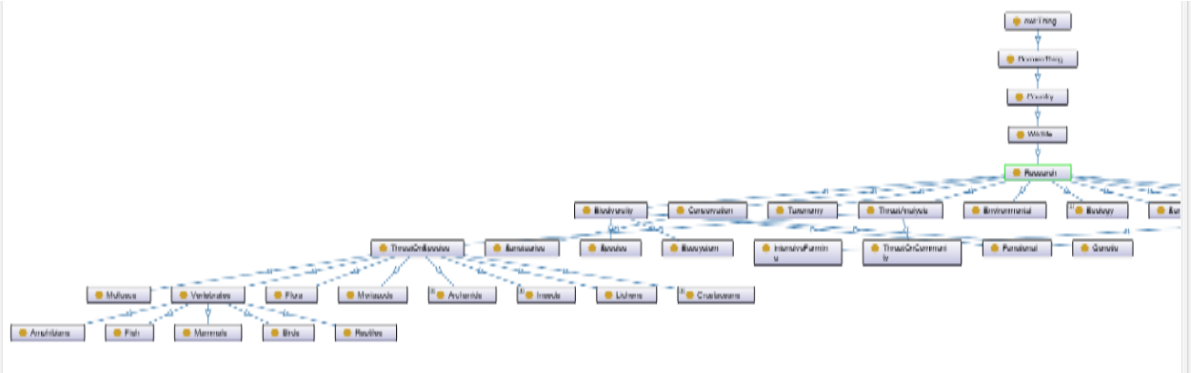

Figure 6.Ontograph partial view

Each research papers' keyword generated by LDA visualization model estimation was analyzed through the ontograph and each paper classification performed.

IV. CONCLUSION AND FUTURE WORKS

In this paper, we suggested a domain-independent and self-learning model, which means that studying ontologies in new domains is very exciting and thus can save considerable time and effort in the acquisition of ontology. For past research papers using learning ontologies, we have established a new approach for automatic classification, and this method applied the LDA model to generate topics, and the progress of learned ontology does not need the seed of ontology, it only requires given document corpus. We generated LDA keywords for selected research abstracts of the past wildlife domain in Sri Lanka. We devise a semi-automated topic labeling for the research papers. The final experiment has proved effective results.

This work reduced the complexity to label the research papers without any domain pre-knowledge. Using this method the hidden keyword and the relations between the keywords also identify to help future research ideas.

In this topic labelling method, there is some inefficient while ontology classification. Because there are several cross path hierarchy moves of keywords identified from LDA. So when we used ontology it collapsed the different path in onto graph. So we will use other classification methods for fully automated our methods.

REFERENCES

Adhitama, R., Kusumaningrum, R. and Gernowo, R. (2018) 'Topic labeling towards news document collection based on Latent Dirichlet Allocation and ontology', *Proceedings - 2017 1st International Conference on Informatics and Computational Sciences, ICICoS 2017*, 2018-Janua, pp. 247–251. doi: 10.1109/ICICOS.2017.8276370.

Chowdhury, S. and Zhu, J. (2019) 'Towards the ontology development for smart transportation infrastructure planning via topic modeling', *Proceedings of the 36th International Symposium on Automation and Robotics in Construction, ISARC 2019*, (Isarc), pp. 507–514. doi: 10.22260/isarc2019/0068.

Chuang, J., Manning, C. D. and Heer, J. (2012) 'Termite: Visualization techniques for assessing textual topic models', *Proceedings of the Workshop on Advanced Visual Interfaces AVI*, (June), pp. 74–77. doi: 10.1145/2254556.2254572.

Editors, N. T. R. (2019) *Technologies in Data Science and Communication*.

Frasier, T. R. *et al.* (2019) 'Bayesian abundance estimation from genetic mark-recapture data when not all sites are sampled: an example with the bowhead whale', *bioRxiv*. Elsevier Ltd, 22, p. 549394. doi: 10.1101/549394.

Hussein, G. *et al.* (2020) 'ONTOLOGY DOMAIN MODEL FOR E-TUTORING SYSTEM', 5(1), pp. 37–44.

L.P.Jayatissa (2012) *Present Status of Mangroves in Sri Lanka, The National Red List 2012 of Sri Lanka; Conservation Status of the Fauna and Flora*.

Lee, J. *et al.* (2018) 'Ensemble modeling for sustainable technology transfer', *Sustainability (Switzerland)*, 10(7). doi: 10.3390/su10072278.

Lin, Z. (2017) 'Terminological ontology learning based on LDA', *2017 4th International Conference on Systems and Informatics, ICSAI 2017*, 2018-Janua(Icsai), pp. 1598–1603. doi: 10.1109/ICSAI.2017.8248539.

Movshovitz-Attias, D. and Cohen, W. W. (2015) 'KB-LDA: Jointly learning a knowledge base of hierarchy, relations, and facts', *ACL-IJCNLP 2015 - 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Proceedings of the Conference*, 1, pp. 1449–1459. doi: 10.3115/v1/p15-1140.

Rani, M., Dhar, A. K. and Vyas, O. P. (2017) 'Semi-automatic terminology ontology learning based on topic modeling', *Engineering Applications of Artificial Intelligence*, 63(August), pp. 108–125. doi: 10.1016/j.engappai.2017.05.006.

Zhu, X., Klabjan, D. and Bless, P. N. (2017) 'Unsupervised terminological ontology learning based on hierarchical topic modeling', *Proceedings - 2017 IEEE International Conference on Information Reuse and Integration, IRI 2017*, 2017-Janua, pp. 32–41. doi: 10.1109/IRI.2017.18.

ACKNOWLEDGEMENT

We acknowledge the Department of wildlife conservation of Sri Lanka for the research permit (WL/3/2/60/15) granted to E. P Kudavidanage.

AUTHOR BIOGRAPHY

Premisha Premananthan is an undergraduate student at the Sabaragamuwa University of Sri Lanka who follows a Bachelor of Science Special degree in Information Systems. She is currently in the final year of her degree program.

Banage T. G. S. Kumara received the Bachelor's degree in 2006 from Sabaragamuwa University of Sri Lanka. He received the master's degree in 2010 from the University of Peradeniya and a Ph.D. degree in 2015 from the University of Aizu, Japan. His research interests include semantic web, web data mining, web service discovery, and composition.

Enoka P. Kudavidanage received a Bachelor's degree in Zoology from the University of Colombo. She received a master's degree in Environmental Sciences from the University of Colombo and a Ph.D. degree for Conservation Biology from the National University of Singapore.

Kuhaneswaran Banujan received his Bachelor of Science degree in 2019 with the Second Class Upper Division from Sabaragamuwa University of Sri Lanka. He is currently attached to the Department of Computing and Information Systems as a Lecturer in Computer Science. His research interests include Data Mining, Knowledge Management, Ontology Modeling, Business Process Simulation.