Learning Concept Hierarchies through Probabilistic Topic Modeling

V S Anoop⁵⁶, S Asharaf⁵⁶, Deepak P⁵⁶

⁵⁶Data Engineering Lab, Indian Institute of Information Technology and Management - Kerala (IIITM-K), Thiruvananthapuram 695 581, India, Contact: anoop.res15@iiitmk.ac.in

⁵⁶Indian Institute of Information Technology and Management - Kerala (IIITM-K), Thiruvananthapuram 695 581, India, Contact: asharaf.s@iiitmk.ac.in

⁵⁶Queen’s University, Belfast, UK, Contact: deepaksp@acm.org

With the advent of semantic web, various tools and techniques have been introduced for presenting and organizing knowledge. Concept hierarchies are one such technique which gained significant attention due to its usefulness in creating domain ontologies that are considered as an integral part of semantic web. Automated concept hierarchy learning algorithms focus on extracting relevant concepts from unstructured text corpus and connect them together by identifying some potential relations exist between them. In this paper, we propose a novel approach for identifying relevant concepts from plain text and then learns hierarchy of concepts by exploiting subsumption relation between them. To start with, we model topics using a probabilistic topic model and then make use of some lightweight linguistic process to extract semantically rich concepts. Then we connect concepts by identifying an "is-a" relationship between pair of concepts. The proposed method is completely unsupervised and there is no need for a domain specific training corpus for concept extraction and learning. Experiments on large and real-world text corpora such as BBC News dataset and Reuters News corpus shows that the proposed method outperforms some of the existing methods for concept extraction and efficient concept hierarchy learning is possible if the overall task is guided by a probabilistic topic modeling algorithm.

**Keywords**: Concept Extraction, Natural Language Processing, Probabilistic Topic Models, Semantic Web, Subsumption Hierarchy Learning, Text Mining.

1. INTRODUCTION

Due to rapid growth of text producing and consuming applications, numerous tools and techniques were introduced in the recent past for extracting useful patterns from unstructured text. These patterns are crucial for organizations to discover knowledge out of it and aid in making intelligent decisions. As the amount of such data grows exponentially, already available algorithms performs poor on the scalability and performance aspects. But there are still a lot of avenues where text data is yet to be exploited fully and thus we need new and efficient algorithms to tackle this situation. Platforms such as social networks, e-commerce websites, blogs and research journals generate such data in the form of unstructured text and it is essential to analyze, synthesis and process such data for efficient retrieval of useful information.

In text mining, concepts are defined as a sequence of words that constitute real or imaginary entities. Extraction of such entities are non-trivial for applications such as automated ontology generation [1], document summarization [2] and aspect oriented sentiment analysis [3] to name a few. This is the era of data explosion thus it is very difficult to store, process, manage and most importantly to extract knowledge out of it. To overcome this short-fall, a significant amount of research has been carried out in the recent past for leveraging underlying thematic and semantic structure from
Learning Concept Hierarchies through Probabilistic Topic Modeling

posed algorithm, false positive is the number of extracted concepts that are not truly human authored concepts and false negative is the human authored concepts that are missed by the concept extraction method. Using these measures, we have compared our proposed method against some of the existing concept extraction algorithms and the result is shown in Table 4.

From the performance graph shown in Figure 4, it is clear that our proposed algorithm extracts more concepts as the number of topics are increasing. The other baseline algorithms such as ACE and ICE performs poor when the number of topics are increased randomly. This shows that the proposed algorithm outperforms the baseline algorithms when extracting real-world concepts from large number of statistically generated topics.

Table 4
Comparison of ACE, ICE and our proposed method

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACE</td>
<td>0.2372</td>
<td>0.2689</td>
<td>0.2517</td>
</tr>
<tr>
<td>ICE</td>
<td>0.7113</td>
<td>0.8147</td>
<td>0.7595</td>
</tr>
<tr>
<td>Proposed</td>
<td>0.8165</td>
<td>0.8901</td>
<td>0.8516</td>
</tr>
</tbody>
</table>

7. CONCLUSIONS AND FUTURE WORK

This paper proposed a novel framework for extracting close to real world concepts from large collection of unstructured text documents which is guided by a probabilistic topic modeling algorithm. Proposed method also deals with learning a subsumption hierarchy which exploits “is-a” relationships among identified concepts which is extensively used in ontology generation. Experiments conducted on large datasets such as Reuters and BBC news corpus shows that the proposed method outperforms...
some of the already available algorithms and better concept identification is possible with this framework.

Because of the promising end results, we are interested to work mainly on the directions of measuring the scalability of proposed framework by using more large datasets. Apart from the basic subsumption hierarchy which depicts "is-a" relation, our future work will be on leveraging other relations that exist between concepts we would like to so that a this framework can automate the complete ontology generation process.

REFERENCES


**Anoop V S** is a full time Ph.D Research Scholar at Data Engineering Lab, Indian Institute of Information Technology and Management - Kerala (IIITM-K), Thiruvananthapuram, India. He received his Masters in Computer Applications (MCA) - from IGNOU and Master of Philosophy in Computer Science from Cochin University of Science and Technology (CUSAT), Kerala in 2014. He has several publications in international journals, conference proceedings and book chapters. His research interests include Information Retrieval, Text Mining and NLP.

**Asharaf S** is an Associate Professor at Indian Institute of Information Technology and Management - Kerala (IIITM-K), Thiruvananthapuram, India. He received his Ph.D and Master of Engineering degrees in Computer Science and Engineering - from Indian Institute of Science, Bangalore. His areas of interest include Algorithms, Business Models and Software Systems related to Data Mining, Data Analytics, Information Retrieval, Computational Advertising, Soft Computing and Machine Learning.

**Deepak Padmanabhan** is a Lecturer (Asst. Professor) in Computer Science at Queens University Belfast, UK. He completed his M.Tech and Ph.D from Indian Institute of Technology Madras, all in Computer Science. His current research interests include Data Analytics, Similarity Search, Information Retrieval and Natural Language Processing. He has published over 40 research papers across major venues in Information and Knowledge Management. He is a Senior Member of the IEEE and ACM.