International Journal of Information Processing, 10(3), 1-11, 2016
ISSN : 0973-8215
IK International Publishing House Pvt. Ltd., New Delhi, India

Learning Concept Hierarchies through Probabilistic Topic Modeling

V S Anoop[a], S Asharaf[b], Deepak P[c]

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

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

[c]Queen'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 shortfall, a significant amount of research has been carried out in the recent past for leveraging underlying thematic and semantic structure from

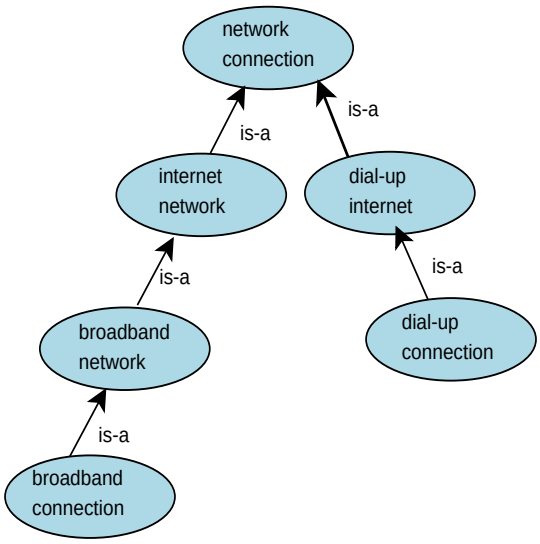

Figure 2. Part of a subsumption hierarchy learned using Algorithm 2

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

Algorithm	Precision	Recall	F1
ACE	0.2372	0.2689	0.2517
ICE	0.7113	0.8147	0.7595
Proposed	**0.8165**	**0.8901**	**0.8516**

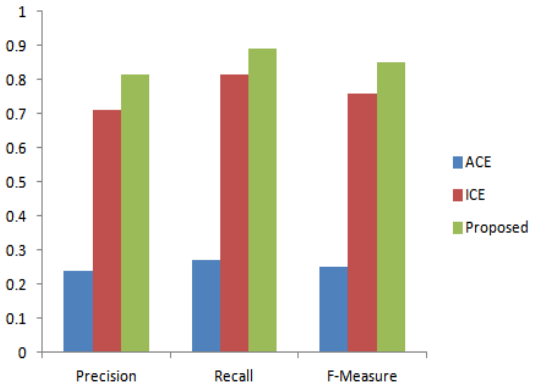

Figure 3. Precision, Recall and F-measure comparison of ACE, ICE and proposed method

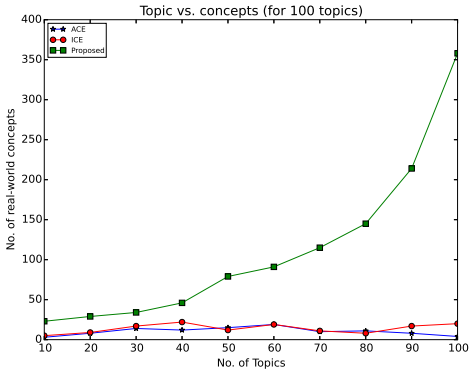

Figure 4. Performance of algorithms on 100 topics (No. of human interpretable concepts generated)

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

1. Pospiech, Sebastian, Martin Pelke and Robert Mertens. Semi-automated Ontology Creation for Semantic Search in Business Process Exploration *IEEE Tenth International Conference on Semantic Computing (ICSC)*, 2016.
2. Marujo, Lus, *et al.*,. Exploring Events and Distributed Representations of Text in Multi-Document Summarization. *Knowledge-Based Systems*, 94:33–42, 2016.
3. Manek AS, Shenoy P D, Mohan MC and Venugopal K R. Aspect Term Extraction for Sentiment Analysis in Large Movie Reviews using Gini Index Feature Selection Method and SVM Classifier. *World Wide Web*, pages 1–20, 2016.
4. Steyvers M and Griffiths T. Probabilistic Topic Models. *Handbook of Latent Semantic Analysis*, 427(7):424–440, 2007.
5. Sanderson M and Croft B. Deriving Concept Hierarchies from Text. *In Proceedings of the* 22^{nd} *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 206–213, 1999.
6. Lindsey RV, Headden III WP and Stipicevic MJ. A Phrase-Discovering Topic Model using Hierarchical Pitman-yor Processes. *In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 214–222, 2012.
7. El-Kishky A, Song Y, Wang C, Voss CR and Han J. Scalable Topical Phrase Mining from Text Corpora. *Proceedings of the VLDB Endowment*, 8(3):305–316, 2014.
8. Wang X, McCallum A and Wei X. Topical n-grams: Phrase and Topic Discovery, with an Application to Information Retrieval. *In Proceedings of Seventh IEEE International Conference on Data Mining (ICDM 2007)*, pages 697–702, 2007.
9. Ramirez P M, Mattmann C A. ACE: Improving Search Engines *via* Automatic Concept Extraction. *In Proceedings of the 2004 IEEE International Conference on Information Reuse and Integration, IRI 2004,* pages 229–234, 2004.
10. Turney P D. Learning Algorithms for Key Phrase Extraction. *Information Retrieval.*, 2(4):303–316, 2000.
11. Parameswaran A, Garcia-Molina H and Rajaraman A. Towards the web of Concepts: Extracting Concepts from Large Datasets. *In Proceedings of the VLDB Endowment.*, 3(1-2):566-577, 2010.
12. Gelfand B, Wulfekuler M and Punch WF. Automated concept extraction from plain text. *In AAAI 1998 Workshop on Text Categorization.*, pages 13-17, 1998.
13. Rajagopal D, Cambria E, Olsher D and Kwok K. A graph-based approach to commonsense concept extraction and semantic similarity detection. *In Proceedings of the 22nd international conference on World Wide Web companion*, pages 565-570, 2013.
14. Krulwich B and Burkey C. Learning User Information Interests through Extraction of Semantically Significant Phrases. *In Proceedings of the AAAI Spring Symposium on Machine Learning in Information Access*, pages 100–112, 1996.
15. Witten I H, Paynter G W, Frank E, Gutwin C and Nevill-Manning CG. KEA: Practical Automatic Key Phrase Extraction. *In Proceedings of the Fourth ACM Conference on Digital Libraries.*, pages 254-255, 1999.
16. Song M, Song IY and Hu X. KPSpotter: A Flexible Information Gain-based Key Phrase Extraction System. *In Proceedings of the* 5^{th} *ACM International Workshop on Web Information and Data Management.*, pages 50–53, 2003.
17. Frantzi K, Ananiadou S and Mima H. Automatic Recognition of Multi-word Terms: the c-value/nc-value Method. *International Journal on Digital Libraries.*, 3(2):115–130, 2000.
18. Hofmann T. Probabilistic Latent Semantic Indexing. *In Proceedings of the* 22^{nd} *Annual International ACM SIGIR Conference on Research and Development in Information Re-*

trieval, pages 50–57, 1999.
19. Blei DM, Ng AY and Jordan MI. Latent Dirichlet Allocation. *In Journal of Machine Learning Research.*, 3:993–1022, 2003.
20. Dumais ST. Latent Semantic Analysis. *Annual Review of Information Science and Technology.*, 38(1):188–230, 2004.
21. Hofmann T. Probabilistic Latent Semantic Indexing. *In Proceedings of the* 22^{nd} *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 50–57, 1999.
22. Blei DM, Ng AY and Jordan MI. Latent Dirichlet Allocation. *Journal of Machine Learning Research.*, 993-1022, 2003.
23. Bird S. NLTK: The Natural Language Toolkit. *In Proceedings of the COLING/ACL on Interactive Presentation Sessions*, pages 69–72, 2006.
24. Lewis D, Yang Y, Rose T and Li F. RCV1: A New Benchmark Collection for Text Categorization Research. *Journal of Machine Learning Research*, 5:361-397, 2004.
25. Greene D and Cunningham P. Practical Solutions to the Problem of Diagonal Dominance in Kernel Document Clustering. *In Proceedings of the* 23^{rd} *International Conference on Machine Learning*, pages 377–384, 2006.

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.