Efficient Auto-Generation of Taxonomies for Structured Knowledge Discovery and Organization

Deepak Ajwani
Nokia Bell Labs
Dublin, Ireland
deepak.ajwani@nokia-bell-labs.com

Sourav Dutta
Nokia Bell Labs
Dublin, Ireland
sourav.dutta@nokia-bell-labs.com

Pat Nicholson
Nokia Bell Labs
Dublin, Ireland
pat.nicholson@nokia-bell-labs.com

Luca Maria Aiello
Nokia Bell Labs
Cambridge, United Kingdom
luca.aiello@nokia-bell-labs.com

Alessandra Sala
Nokia Bell Labs
Dublin, Ireland
alessandra.sala@nokia-bell-labs.com

ABSTRACT

This tutorial introduces the audience to the latest breakthroughs in the area of interpreting unstructured content through an analysis of the key enabling scientific results along with their real-world applications. With technical presentations of problems like named-entity disambiguation and dynamically updating the knowledge hierarchy with domain-specific vocabulary, it would provide the fundamentals to the building-blocks of various applications in Artificial Intelligence, Natural Language Processing, Machine Learning, and Data Mining.

CCS CONCEPTS

• **Information systems** → *Data extraction and integration*; *Ontologies*; *Information extraction*; Information systems applications;

ACM Reference Format:
Deepak Ajwani, Sourav Dutta, Pat Nicholson, Luca Maria Aiello, and Alessandra Sala. 2018. Efficient Auto-Generation of Taxonomies for Structured Knowledge Discovery and Organization. In *HT '18: 29th ACM Conference on Hypertext and Social Media, July 9–12, 2018, Baltimore, MD, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3209542.3212476

1 INTRODUCTION

Future Big Data systems are expected to showcase enriched cognitive abilities for data and pattern discovery for large-scale analytics on vast amounts of linked structured and unstructured multi-modal data. This would usher in the next-generation functionalities for e-commerce, transportation, IoT and smarter health-care.

However, progress in the area of data science lies at the confluence of semantic search, reasoning, knowledge representation, algorithm engineering, natural language processing and machine learning. To this end, the proposed tutorial will provide the audience with the latest breakthroughs and state-of-the-art techniques for knowledge discovery and their organization for applications like semantic linking and contextual interpretation. We further present how linked knowledge hierarchies can be compared on both structural and semantic subsumption similarities. Further, such cognitive blocks should be highly accurate and scalable, depicting just-in-time prediction and computationally cheap updates.

As a real-world manifestation, we discuss the application of techniques to novel analytical avenues like: (1) analyzing "sound maps" of urban areas to extract relationship between soundscapes, emotions and perceptions; (2) creation of dictionary for urban smell to analyze how different categories (e.g., industry, transport) correlate with air quality; and (3) retrieving topically related multimedia content segments for faster ingestion of information.

Finally, as food for thought, the tutorial will also highlight future directions of work and various open challenges.

Keywords: *Linked Knowledge Hierarchies, Entity Linking, Word Embeddings, Graph Measures, Katz Centrality, Topic Labeling*

Tutorial Outline

The tutorial would impress the importance of structure knowledge hierarchy and enable the attendees to gain an insight as to how a taxonomy can be mined from unstructured or semi-structured corpus of text using co-occurrence graphs, statistical methods and hierarchical clustering methods. The detailed outline is as follows:

(1) Introduction – Semantic Linking, Knowledge Repositories, and Linked Data discovery
   – Ontologies and Knowledge Hierarchies
   – RDF structure and Linked Data
   – Taxonomy structure: Directed Acyclic Graphs
   – TF-IDF, LDA [3], classifiers [6], word embeddings [13]
(2) Application areas that leverage taxonomies
   – Semantic relationship and Topical relatedness
   – Community detection in social media
(3) Efficiency trade-offs
   – Semantic interpretation
   – Accuracy and Scalability
   – Just-in-time prediction
   – Fast updates
(4) Dynamically updating taxonomies
   – Induction of taxonomies [1, 7]
   – Breaking cycles in noisy hierarchies [15]
   – Evolution of new concepts and word senses

HT '18, July 9–12, 2018, Baltimore, MD, USA

ACM ISBN 978-1-4503-5427-1/18/07.
https://doi.org/10.1145/3209542.3212476

- Named entity linking for existing concepts [4, 11]
- Measures to capture new concepts

(5) Unsupervised placement of new concepts in taxonomies
- Rule based techniques
- Probabilistic approaches

(6) Supervised placement of new concepts in taxonomies
- Syntactic Features
- Semantic Features
- Graph Features
- Integrating features using learning-to-rank [12]
- Discussion of efficiency trade-offs
- Identifying Wikipedia categories for emerging concepts

(7) Efficient comparison of taxonomies
- Structural overlap measures
- Tree-edit distance [2] and graph similarity measures [10]
- Fowlkes-Mallows measure [5]
- Katz similarity scores [9] and their aggregation
- Discussion of efficiency trade-offs

(8) Domain-specific taxonomies for smarter applications
- Assigning human-readable topical tags to documents [8]
- Linking related multi-media contents
- Taxonomies for different senses: sound, visual, smell [14]

(9) Conclusion and Future Directions

Tutorial Length: 1.5 hours.

2 CONCLUSION

Linked data such as structured knowledge hierarchies provide invaluable source of information pertaining to concepts, their relationships, and dependency structure. This tutorial discusses efficient techniques for induction of taxonomies, and their subsequent dynamic updation to reflect emerging concepts. We show how current techniques in text mining, graph processing and machine learning can be leveraged by breaking complex learning models into smaller models. Such techniques would directly impact the representation, enrichment, and management for analysis of evolution and influence in semantic graphs and networks.

REFERENCES

[1] Luca Maria Aiello, Rossano Schifanella, Daniele Quercia, and Francesco Aletta. 2016. Chatty maps: constructing sound maps of urban areas from social media data. *Royal Society open science* 3, 3 (2016), 150690.

[2] P. Bille. 2005. A survey on tree edit distance and related problems. *Theoretical Computer Science* 337, 1 (2005), 217–239.

[3] David M. Blei. 2012. Probabilistic Topic Models. *Commun. ACM* 55, 4 (April 2012), 77–84.

[4] Paolo Ferragina and Ugo Scaiella. 2010. TAGME: on-the-fly annotation of short text fragments (by wikipedia entities). In *Proceedings of the 19th ACM Conference on Information and Knowledge Management, CIKM 2010, Toronto, Ontario, Canada, October 26-30, 2010*. ACM, 1625–1628.

[5] E. B. Fowlkes and C. L. Mallows. 1983. A Method for Comparing Two Hierarchical Clusterings. *J. Amer. Statist. Assoc.* 78, 383 (1983), 553–569.

[6] S. R. Gunn. 1998. *Support Vector Machines for Classification and Regression*. Technical Report. Univ. of Southampton, USA.

[7] Victoria Henshaw. 2013. *Urban smellscapes: Understanding and designing city smell environments*. Routledge.

[8] Ioana Hulpus, Conor Hayes, Marcel Karnstedt, and Derek Greene. 2013. Unsupervised Graph-based Topic Labelling Using Dbpedia. In *WSDM*.

[9] Leo Katz. 1953. A new status index derived from sociometric analysis. *Psychometrika* 18, 1 (1953), 39–43.

[10] Danai Koutra, Neil Shah, Joshua T. Vogelstein, Brian Gallagher, and Christos Faloutsos. 2016. DeltaCon: A principled massive-graph similarity function with attribution. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 10, 3 (2016).

[11] Tiep Mai, Bichen Shi, Patrick K. Nicholson, Deepak Ajwani, and Alessandra Sala. 2017. Scalable Disambiguation System Capturing Individualities of Mentions. In *Language, Data, and Knowledge - First International Conference, LDK 2017, Galway, Ireland, June 19-20, 2017, Proceedings (Lecture Notes in Computer Science)*, Vol. 10318. Springer, 365–379.

[12] Brian McFee and Gert Lanckriet. 2010. Metric Learning to Rank. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*. 775–782.

[13] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and Their Compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*. 3111–3119.

[14] Daniele Quercia, Rossano Schifanella, Luca Maria Aiello, and Kate McLean. 2015. Smelly maps: the digital life of urban smellscapes. *arXiv preprint arXiv:1505.06851* (2015).

[15] Jiankai Sun, Deepak Ajwani, Patrick K. Nicholson, Alessandra Sala, and Srinivasan Parthasarathy. 2017. Breaking cycles in noisy hierarchies. In *ACM Conference on Web Science*. 151–160.