Handling Information Overload: Automatic Generation of Wikipedia Articles

Vikrant Yadav, Faisal Khan
Indian Institute of Technology, Roorkee
{vikrantiitr1, faisal.khan.ggn}@gmail.com

ABSTRACT

The exponential growth of information on the web over the years has lead to the problem of information overload, i.e. amount of information present on web is beyond the processing capacity of any system. Thus, a need arises to have a single resource to properly cover as well as to have an up to date information about a topic. The popular website Wikipedia does the same in a structured manner thus giving a comprehensive coverage about any topic. However, Wikipedia is crowdsourced and thus can't cover everything. So, a mechanism is required which takes large number of documents from the web and gives a Wikipedia like structured and detailed information about any topic in real-time.

For automatic generation of such an article, a structure-aware approach has been proposed by Sauper et. al. (2009) [2]. They generated templates for different categories and for a given topic, retrieved information from the web using pre-learned queries. However, their approach lacks to utilize semantic relationships between the information under similar sections in different articles. Instead of just clustering section titles to generate templates and queries, our study suggests that a topic model like Replicated Softmax [1] or Deep Boltzmann Machines [3] can be used to create better templates and queries. Also, we generate semantically similar queries for each pre-learned query using DBPedia and the results are then combinedly re-ranked using our algorithm. This results in high-quality information from web and a coherent and comprehensive article.

BODY

Semantic relationships among existing Wikipedia articles can be used to generate high-quality structured information on a given topic.

REFERENCES

[1] R. Salakhutdinov and G. E. Hinton. Replicated softmax: an undirected topic model. In *NIPS*, volume 22, pages 1607–1614, 2009.

[2] C. Sauper and R. Barzilay. Automatically generating wikipedia articles: A structure-aware approach. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1*, ACL '09, pages 208–216, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[3] N. Srivastava, R. Salakhutdinov, and G. E. Hinton. Modeling documents with deep boltzmann machines. *CoRR*, abs/1309.6865, 2013.

Volume 3 of Tiny Transactions on Computer Science