Homophilic Self Organizing Feature Maps: Finding Topics on Socially Connected Data

Bernardo Simões

***Abstract*—With the evolution of social media platforms, the amount of unlabeled information has gone skyrocketing. The process of labeling this kind of information is evermore complex. Typical approaches used on the WEB for Topic Detection and Tracking cannot be directly applied due to the small amount of text produced per tweet, orthographic errors, abbreviations and so on.**

In this thesis, we propose and analyze a new form of topic detection and tracking on social networks. By leveraging the social relations between authors of the gathered content, and apply them to the clustering process.

In order to achieve this, we proposed some modifications to the artificial neural network and clustering algorithm — Self Organizing Maps.

***Index Terms*—topic detection, twitter, self-organizing maps, classification, clustering**

I. Introduction

WITH the evolution of social network websites like Facebook and Twitter, the amount of pertinent content about a specif issue is increasing dramatically, which calls for new ways to make sense and catalog this data. The usage of social networks for branding quality and on-line marketing is specially compelling since 19% of all tweets [2] and 32% of blog posts [5] are about brands or products. Nevertheless, finding topic sensitive information on social networks is extremely complicated due to the fact that documents have very little content, slang vocabulary, orthographic mistakes and abbreviations. Asur and Huberman [1] successfully predicted box-office revenues by monitoring the rate of creation of new topics based on debuting movies. Their work outperformed some traditional market-based predictors.

Thus, academic and enterprise worlds started looking at Machine Learning (ML) for new ways to achieve revenue or simply explore and discover patterns in data.

Using unsupervised ML, Le et al. [4] was able to achieved 81.7% accuracy in detecting human faces, 76.7% accuracy when identifying human body parts and 74.8% accuracy when identifying cats. He used a 9-layered locally connected sparse auto-encoder with pooling and local contrast normalization on a large dataset of images (the model has 1 billion connections, the dataset has 10 million 200x200 pixel images downloaded from the Internet). This dataset was trained using model parallelism and asynchronous SGD on a cluster with 1,000 machines (16,000 cores) during three days. Even though the amount of computing power used in this project was of several order of magnitude, it is remarkable how an unsupervised algorithm could achieve such results.

Social Media Analytics is another raising topic that draws from Social Network Analysis [3], ML, Data Mining [7], Information Retrieval (IR) [6], and Natural Language Processing (NLP). As stated Melville et al. [5], 32% of the 200 million active bloggers write about opinions on products and brands, while 71% of 625 million Internet users read blogs and 78% of respondents put their trust in the opinion of other consumers. In comparison, traditional advertising is only trusted by 57% of consumers. This kind of data drives companies to Social Media Analytics as a way to know what people are saying on the web about their companies and products. This new worry has brought to life a lot of new startups like Sumalor ThoughtBuzz, but also solutions from the old players like IBM and SAS.

II. Conclusion

The conclusion goes here.

References

[1] Asur, S. and Huberman, B. (2010). Predicting the future with social media. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, volume 1, pages 492–499.

[2] Jansen, B. J., Zhang, M., Sobel, K., and Chowdury, A. (2009). Twitter power: Tweets as electronic word of mouth. *J. Am. Soc. Inf. Sci. Technol.*, 60(11):2169–2188.

[3] Knoke, D. and Yang, S. (2008). *Social network analysis*, volume 154. Sage.

[4] Le, Q. V., Ranzato, M., Monga, R., Devin, M., Corrado, G., Chen, K., Dean, J., and Ng, A. Y. (2012). Building high-level features using large scale unsupervised learning. In *ICML*. icml.cc / Omnipress.

[5] Melville, P., Sindhwani, V., and Lawrence, R. (2009). Social media analytics: Channeling the power of the blogosphere for marketing insight. *Proc. of the WIN*, pages 2–6.

[6] Salton, G. and McGill, M. J. (1983). Introduction to modern information retrieval.

[7] Witten, I. H. and Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.