# Learning to Predict the Future using Web Knowledge and Dynamics

Kira Radinsky
Technion - Israel Institute of Technology
*kirar@cs.technion.ac.il*

**Abstract**

Mark Twain famously said that "the past does not repeat itself, but it rhymes." In the spirit of this reflection, we present novel algorithms and methods for leveraging large-scale digital histories and human knowledge mined from the Web to make real-time predictions about the likelihoods of future human and natural events of interest.

The Web is a dynamic being, with constantly updating content, which is entangled with sophisticated user behaviors and interactions. Some of these behaviors have the ability to convey current trends in the present, e.g., economical growth (predicting automobile sales based on query volume [6]), popular movies [4], and political unrest [1, 3, 5]. We mine the ever-changing Web content and user Web behavior. We show that, not only the dynamics itself can be predicted, but also that it can be used for future real-world event prediction.

We mine decades of news reports (1851 – 2010) from the New York Times (NYT), and describe how we can learn to predict the future by generalizing sets of concrete transitions in sequences of reported news events. In addition to the news corpora, we leverage data from freely available Web resources, including Wikipedia, FreeBase, OpenCyc, and GeoNames, via the LinkedData platform [2]. The goal is to build predictive models that generalize from specific sets of sequences of events to provide likelihoods of future outcomes, based on patterns of evidence observed in near-term Web activities. We propose the methods as a means of generating actionable forecasts in advance of the occurrence of target events in the world.

This thesis is one of the first works to demonstrate general, unrestricted artificial-intelligence prediction capacity. We present methods derived from heterogeneous Web sources to make knowledge-intensive reasoning about causality and future event prediction, using both automatic feature extraction and novel algorithms for generalizing over historical examples.

## References

[1] S. Asur and B. A. Huberman. Predicting the future with social media. In *ArxiV*, 2010.

[2] C. Bizer, T. Heath, and T. Berners-Lee. Linked data – the story so far. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 2009.

[3] M. Joshi, D. Das, K. Gimpel, and N. A. Smith. Movie reviews and revenues: An experiment in text regression. In *Proceedings of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL HLT)*, 2010.

[4] Kalev. Culturomics 2.0: Forecasting large-scale human behavior using global news media tone in time and space. *First Monday*, 15(9), 2011.

[5] G. Mishne. Predicting movie sales from blogger sentiment. In *Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI) Spring Symposium*, 2006.

[6] H. Varian and H. Choi. Predicting the present with google trends. *Technical Report*, 2009.