Master's Project

Online Opinion Correlation for Large-scale Demographics

Sihem Amer-Yahia (LIG/HADAS) and Ahlame Douzal (LIG/AMA)

1 Project Description

Nowadays, users on the social web are more than eager to express their opinion in the form of reviews, ratings, tags, likes, etc. We are also witnessing the increasing availability of users' demographics data. As a result, the need to provide fine-grained analytics of demographics opinion is growing, demanding efficient and scalable methods of opinion aggregation and correlation. While several approaches exist for mining opinions from product reviews or micro-blogs, little attention has been devoted to *the online aggregation and correlation of extracted opinions for different demographics groups (e.g., "Students in Italy") over time.* In our recent work [2, 3], we introduced methods for aggregating user ratings and tags and extracting meaningful demographics patterns such as "Teenage girls consistently like Woody Allen Movies".

In this master's project, we aim to formalize a number of opinion aggregation and correlation functions and explore their efficient implementation. The problem of opinion aggregation relies on two main points. First, given a demographics group described by a set of opinion profiles, the aim is to extract prototypes best representing the main underlying opinion dynamics. For that, many strategies developed for static data as well as clustering approaches for temporal data will be explored [6, 7]. The second focus relies on the description of the extracted prototype dynamics. The challenges for these descriptions are : a) to cover the variability, uncertainty and confidence of the aggregated opinion within demographics groups, b) to select the most informative and confident time periods by segmenting the opinion prototypes. For that, interval-valued, histograms-valued data and temporal segmentation strategies can be investigated [1, 5].

This work aims to study families of metrics for temporal data as those developed in [4] to measure, in a scalable way, correlations between demographics opinion dynamics; and to subsequently reveal the most correlated ones. Finding demographics groups requires the exploration of all possible combinations of values for demographics attributes - a task, that becomes quickly intractable, especially for pairs of groups. Our algorithms in [2, 3] exploit the *lattice structure induced by demographics attributes* in order to prune the search space. We propose to adapt them to the newly developed aggregation and correlation functions.

This work will serve as a basis for the definition of a number of problems and the exploration of their efficient implementation. Examples of such problems are finding demographics groups whose opinion changes over time, or finding pairs of demographics groups with correlated opinions over time, that is, groups that react similarly over time to external events. We plan to co-advise a PhD student on this topic.

2 Related Work

Some recent studies have already made a step along this direction. For instance, "A Demographic Analysis of Online Sentiment during Hurricane Irene" [8] revealed dynamic (temporal) sentiment differences between Southern USA and New England, and at the same time a constant (inherent) difference in the sentiments expressed by males and females. A similar problem of detecting correlations between multiple time series has been approached in the area of data streams using a variety of techniques [11, 9, 10]. Among these, [11] computes correlations using sliding time intervals of specified sizes, composed from a number of sub-intervals of a fixed length. Their work is interesting in particular by the use of *Discrete Fourier Transformation (DFT)* to compute correlations in an approximate and updateable manner.

3 Co-advisors

Ahlame Douzal is an expert in temporal and sequential machine learning and more particularly in temporal segmentation strategies that will serve the formalization of sentiment aggregation and correlation task. Sihem Amer-Yahia is an expert in large-scale indexing and mining of structured datasets, that will serve the task of adapting existing algorithms to enable the newly developed opinion aggregation and correlation functions.

4 Prerequisites

The candidate should have basic knowledge of database management systems and data mining. Programming skills including knowledge of Perl/Python will be a plus. We will be using the MovieLens 10 million ratings dataset [1] that contains user demographics, movie attributes and ratings.

References

[1] A. Chouakria-Douzal. Compression technique preserving correlations of a multivariate temporal sequence. In M. R. Berthold, H.-J. Lenz, E. Bradley, R. Kruse, and C. Borgelt, editors, *Advances in Intelligent Data Analysis*, volume V, pages 566–577. Springer, 2003.

[2] M. Das, S. Amer-Yahia, G. Das, and C. Yu. Mri: Meaningful interpretations of collaborative ratings. In *VLDB*, 2011.

[3] M. Das, S. Thirumuruganathan, S. Amer-Yahia, G. Das, and C. Yu. Who tags what? an analysis framework. *PVLDB*, 5(11):1567–1578, 2012.

[4] A. Douzal-Chouakria and C. Amblard. Classification trees for time series. *Pattern Recognition*, 45(3):1076–1091, 2012.

[5] A. Douzal-Chouakria, L. Billard, and E. Diday. Principal component analysis for interval-valued observations. *Statistical Analysis and Data Mining. Wiley*, 4(2):229–246, 2011.

[6] A. Douzal-Chouakria, A. Diallo, and F. Giroud. Adaptive clustering for time series: application for identifying cell cycle expressed genes. *Computational Statistics and Data Analysis*, 53(4):1414–1426, 2009.

[7] A. Douzal-Chouakria, A. Diallo, and F. Giroud. A random-periods model for the comparison of a metrics efficiency to classify cell-cycle expressed genes. *Pattern Recognition Letters*, 31:1601–1617, 2010.

[8] B. Mandel, A. Culotta, J. Boulahanis, D. Stark, B. Lewis, and J. Rodrigue. A demographic analysis of online sentiment during hurricane irene. In *Proceedings of the Second Workshop on Language in Social Media*, pages 27–36, Montréal, Canada, June 2012. Association for Computational Linguistics.

[9] S. Papadimitriou, J. Sun, and C. Faloutsos. Streaming pattern discovery in multiple time-series. In *VLDB*, pages 697–708, 2005.

[10] S. Papadimitriou, J. Sun, and P. S. Yu. Local correlation tracking in time series. In *ICDM*, pages 456–465, 2006.

[11] Y. Zhu and D. Shasha. Statstream: statistical monitoring of thousands of data streams in real time. In *VLDB*, pages 358–369, 2002.

[1] *http://movielens.umn.edu*