Transactions on Machine Learning
and Data Mining
Vol. 9, No.2 (2016) 46-47
© ISSN: 1865-6781 (Journal)
ISBN: 978-3-942952-47-7
IBaI-Publishing ISSN 1864-9734

ibai Publishing
www.ibai-publishing.org

# Editorial

Dr. Geert Gins

Aixial BVBA
Brussels, Belgium

e-mail: ggins@aixial.com

The amount of data currently generated by various activities of the society has never been so big, and is been generated in an ever increasing speed [1]. Managing and gaining insight from these large amount of data is an enormous challenge for researcher, but it is, at the same time, a key award producing competitive business advantages or compelling scientific advances. Living in the era of data deluge, data insight is appearing not only as a challenge but also as necessity in many contexts, from meteorology, genomics, and environmental research, to finance, business, and healthcare, enabling lower cost and better business decisions [2]. When properly captured and analyzed, data may provide unique insight for variety of important decisions including financial market trends, equipment failures, buying patterns, or maintenance cycles.

At the system level, the essential challenge is that new data mining framework needs to consider complex relationships between data, models, and diverse data sources, along with their evolving changes in time and space, and influenced by other often non-measurable factors. The complexity may be expressed additionally through the features of massive, high-dimensional, heterogeneous, unstructured, incomplete, noise data [3].The availability of unprecedentedly large and rich data sets initiate the need for new computational approaches including new or adopted analytical tools to model that increasing complexity. That will allow researchers and many disciplines to test hypotheses, theories, or system behavior that were previously untestable.

Availability of large amount of unstructured data on Internet, influenced advances in web query recommendations as an essential component of a user-oriented search engine. A common fact in Web search is that a user often needs multiple iterations of query refinement to find the desired results from a search engine. The task of query recommendation is to help users formulate queries that better represent user's search intent during Web search interactions. Building a good Web query recommendation system, however, is very difficult due to the fundamental challenge of predicting users' search intent, especially given a limited user context information.

The first article describes new query recommendation strategies based on query relevance graphs [4]. Queries are recommended based on two graphs: a) click graphs to

model usage patterns, and b) similarity graphs to model query similarities. A click graph is constructed as a URL-Query bipartite graph with the assumption that queries accessing similar documents convey similar meaning, even though they might be textually different. The query similarity graph is a directed graph where the vertices represent queries, and edges represent a high Jaccard similarity in text. A normalized query relevance graph is constructed based on the combination of the above two graphs. A depth first search on this derived graph is used to provide query recommendations. The proposed methodology is implemented on AOL search engine data, and later extended to the task of image recognition from tags as well. The paper presents extensive experimentation based on human expert's feedback from 12 research students, and also from an automated Open Directory Project data. A comparisons with the Heat Diffusion algorithm is performed, and the proposed methodology was found to be superior. Performance evaluation of the proposed methodology shows linear growth in the number of queries and URLs, indicating the scalability of the solution.

One of the most time-consuming and labor-intensive tasks in data mining process is preparation of data for analysis and mining, where features selection is an important part of the process. Special attention should be made to the feature selection techniques when the number of features is relatively large comparing to the number of samples. Most of the feature selection research efforts, including traditional methods classified as filters, wrappers or embedded methods, were directed towards improving individual techniques. The authors of the second article in this issue propose an alternative: combining results from multiple feature selection methods, which are relying on disjoint assumptions about the regression function [5]. The proposed approach outputs a union of selected variables which are results from three methods: MIC, lasso regression, and hierarchical network lasso regression. This approach leads to better sensitivity in selection of features than using traditional methods individually.

The preliminary experiments with synthetic and UCI data sets showed that when the assumed model doesn't include certain types of terms (e.g. nonlinear or interactions between the predictors) then wrong predictors will be selected to explain variability coming from these terms. The number of wrong predictors tends to increase as it is increased the number of samples in the $n << p$ set-up. One way to overcome this problem could be to keep the number of selected variables low or significantly lower than the number of samples when regularized regression is used. Low dimensionality will ensure reduced number of false positives.

## References

1. Markus Vattulainen. Preprocessing Optimization for Predictive Classification: Baseline Results from Six Industry Cases. Transactions on Machine Learning and Data Mining, ibai-Publishing, Vol. 9, No. 2 (2016), 47-60.
2. David Kaczynski, Lisa Gandy, Gongzhu Hu. Innovations in News Media: Crisis Classification System. Transactions on Machine Learning and Data Mining, ibai-Publishing, Vol. 9, No. 2 (2016), 61-76.