StreamDiv Project Report

Steve, Nick, Nathan Ong, Zaeem, Panos Chrysanthis and Alexandros Labrinidis
Department of Computer Science
University of Pittsburgh, Pittsburgh, Pennsylvania 15260
Email:

***Abstract*—For this project we explore different designs for a diversity operator in stream processing environments. In addition, we factor in the impact of recency and user ranking of incoming data in the end result. The system model of our work is a stand-alone stream processing system, with a single stream of data as input. Processing is modeled as a pipeline of operators, which process incoming tuples and forward results to the next operator. The output is not a single result (as in the case of traditional databases), but periodical results that reflect the most sensible outcome (according to the algorithm) based on time-windows.**

As far as different dimensions of the problem are concerned, ranking is considered as a set of user-defined properties that are either attached to incoming tuples (annotate), or are used for filtering out tuples. Turning to the diversity operator, we abstract it as a stateful operator, with time- based sliding windows. The operator's frequency of result production can be one of the following: (i) a temporary snapshot of the most diverse tuples in the current time-window produced every time a tuple is received, (ii) a set of the most diverse tuples among all the tuples that have been received in a user-defined epoch. In the latter case, we consider how many tuples are replaced and which ones every time a new tuple arrives. For the former case, we consider a static approach (a pre-defined fraction p of the tuples are replaced after every epoch), and a dynamic one (the fraction p of the tuples replaced is proportional to the δ in diversity gained compared to the replacement fraction of the previous epoch). As regards to replacement policy, we have the alternatives of random replacement, and least recently tuple received replacement.

I. Introduction

Data creation and collection only continues to increase, leaving a large burden on analysts to come up with a way to quickly and efficiently comb through the data to locate trends and correlations. There tends to be a focus on data-specific approaches with frameworks that work well for certain kinds of data (but not well on others CITATION NEEDED). There also is a need in data analysis to provide relevant but diverse data points as a digest of the full data set. Several offline versions have been created [1] CITATIONS NEEDED, but lack the speed to work in a streaming environment. We present StreamDiv, a streaming framework built on top of a previously developed diversity system for offline database systems, PrefDiv[1], to provide queriers of a data streaming systems a top-k set that is most relevant and diverse, adding recency as a possible influence on the digested set.

II. Preliminaries

Elaborate more on diversity, relevance and recency, and possibly related work...

III. Problem Definition

Formal definition of the problem...

IV. StreamDiv Schemes

Discuss the incremental and batch replacement algorithms here. Also explain how the recency and relevancy are combined.

V. Results

VI. Conclusion

Acknowledgment

The authors would like to thank...

References

[1] X. Ge, P. K. Chrysanthis, and A. Labrinidis, "Preferential diversity," in *Proceedings of the Second International Workshop on Exploratory Search in Databases and the Web*, ser. ExploreDB '15. New York, NY, USA: ACM, 2015, pp. 9–14. [Online]. Available: http://doi.acm.org/10.1145/2795218.2795224