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*—Data mountains continue to grow in size, increasing the difficulty in retrieving useful analysis from them. In addition, generated data is frequently plagued by sameness; the data that is retrieved tends to follow the same pattern, helpful in simple prediction tasks or outlier detection, but not necessarily helpful in characterizing the data from a stream. Users tend to know in a general sense what they are searching for in their datasets and streams, but may not know the full space of relevant results. Diversity helps alleviate this problem, but faces its own challenges when combined with relevancy and recency in streaming contexts.**

In our work, we explore different designs for a diversity operator in a stream processing environment. In addition, we factor in the impact of recency and user-defined relevancy by ranking incoming data. Users will specify a size for a representative top-k set for the data seen from the stream. This top-k set is maintained by the system and updated based on recency, relevancy, and diversity from continuous streaming data. We consider two types of tuple-replacement schemes, Incremental and Batch, for maintaining the top-k set, and find CONCLUSION.

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.

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.

The pipeline is as follows: from a single input stream, a timestamp is appended to the tuple based on the time the system received it. The tuple is then sent to a Relevancy Operator where it is given a score based on its relevance, which is appended to the tuple. Then the Diversity Operator takes the tuple, and uses it to maintain the top-k representative list, which is then sent to the user based on his or her defined window interval. Maintenance of the top-k set is governed by the replacement policy, *Incremental* or *Batch*. The Incremental Replacement Scheme attempts to replace tuples in the top-k set as tuples come into the system, while the Batch Replacement Scheme replaces buffered tuples in a user-specified time or tuple-based window.

II. Related Work

Our system in some senses are close to the top-k publish/subscribe system. There are some existing work on such system (e.g. [2], [3], [4], [5]), in all these systems an newly published item trigger a subscription only if the score of this newly published item are higher then one or more items that are currently in the top-k published item list.

In [3], [4], [5], the relevance score of an item will remain the same until the lifetime of this item ends. Once an item being discarded an most relevant item from outside of the top-k list that are still with in it's own life period will be chosen to replace this newly discarded item. [3], [4], [5] are different from ours, since we consider time as part of the relevancy score, hence the score of an item will constantly changing.

[2] is another top-k publish/subscribe system that consider geo-tagged tweets as published items, and the corresponding Points of Interest as the subscriptions. In [2] , the time is part of the relevance score, which is the same as in our system. However, our work is different from [2] , because on top of the relevancy and recency, we also considered the diversity, which introduced a new trade-off between diversity and relevance, therefore with this new challenge this existing approach is inapplicable.

In addition, [6] is consider to be the most similar to our approach, as it consider relevancy, recency and diversity for top-k subscription. It is different from our work since in [6] the author consider diversity as as part of the ranking score, such that diversity is a score that calculated by the MAXSUM algorithm that defined as following:

Definition 1: MaxSum *generates a subset of R with the maximum $f = \Sigma_{p_i,p_j \in S} dist(p_i, p_j)$ where* dist *is some distance function, $p_i \neq p_j$ for all subsets with the same size.*

Our work uses a modified version of PrefDiv as the diversification algorithm, which is a complete different diversification algorithm that doesn't require the heavily computation of MAXSUM algorithm. Moreover, [6] is build as a pure experiment system without the use of any existing real world streaming system. In contrast, our system is builded based on the existing streaming system STORM.

III. Preliminaries

Both diversity and relevance have been explored to quite some depth within the broader context of *representative data*

exploration techniques. The diversity of a set of tuples is usually defined using a distance measure between pairs of tuples. In the case of numerical attributes, this could be the Euclidean distance between the data points, whereas in other cases, this could be a function specific to the type of data we are dealing with. Formally, let $d(t_i, t_j)$ denote the distance between two tuples t_i and t_j. One possible measure of diversity of a set S is the sum of pairwise distance of all the points in S and is defined as

$$div(S) = \sum_{i=1}^{|S|} \sum_{j>i}^{|S|} d(t_i, t_j) \quad (1)$$

The objective of diversification algorithms is the pick a subset $S*$ with a fixed size k from a larger set D such that the diversity of set $S*$ is maximum. In other words

$$S^* = \underset{S \subseteq D, |S|=k}{argmax}\ div(S) \quad (2)$$

Such an optimization problem has been shown to be NP hard. Several heuristics, however, have been proposed in the research literature that aim to achieve diversity close to the optimal. While diversity in general is not query specific, the relevance of a set S is usually based on the similarity of the tuples in S to the user query posed to the database. Again, for numerical attributes, this similarity could be the proximity of a tuple to a desired, possibly non-existent, tuple or point in the sample space as per the user query. The overall relevance of the set S is usually defined as the sum of the individual relevance scores of the points within the set. Techniques to extract representative data that is most diverse and also contains relevant results usually employ the MMR, or the Maximum Marginal Relevance, scheme, in which the objective function is the weighted sum of the diversity and relevance score of the subset. As both diversity and relevance are orthogonal, and sometimes conflicting, objectives, the weights in these schemes are usually pre specified based on problem requirements and properties of the dataset.

Most of the literature focuses on the problem of diversification and relevancy in the framework of traditional databases, where the data is stored on a disk and all of the data is available to the representative data extraction algorithm as it looks through the entire haystack to pick out a small *representative* subset. The luxury of looking at the entire dataset is not afforded in a streaming system where the system only sees incoming windows of tuples and has to process them on the fly. The incoming data is not stored in its entirety because, at high input rates, the amount of data would quickly exceed the storage capacity. Even if the system has enough storage space to store all the data that it receives, the timing requirements in the streaming system are quite strict. The algorithm would be required to produce the top k attributes at fixed intervals and looking at the entire dataset may take enough time that the system fails to meet its deadline. In this work we look at the problem of diversification and relevance maximization in the context of streaming databases and develop schemes that would work in an online setting, producing high quality results at fixed intervals or as soon as new tuples arrive in the system.

Another aspect that becomes more prominent in streaming systems is the recency of the tuples produced. There may be applications that would prefer to see more recent results at the expense of diversity and/or relevance. One reason for this could be to observe trends in user tweets, for example in the case of twitter data. Our schemes also include the recency of tuples as an objective. We also explore the tradeoff between all these three objectives and their impact on the quality of results produced.

IV. Problem Definition

A. Input Output Relationship

Let k be the number of tuples of S, where $S = \{t_0, t_1, \ldots, t_{k-1}\}$ representing the top-k set. Let n be the number of tuples of N, where $N = \{t'_0, t'_1, \ldots, t'_{n-1}\}$ representing the set of incoming tuples from the stream. (Note for the Incremental Replacement Scheme, $n = 1$.) Our goal is to take some tuples from N to replace some tuples in S based on recency, relevancy, and diversity.

B. Recency

Recency is defined as a monotonically decreasing score, computed as some formula relating the tuple's timestamp of arrival to the system T_A and the current system time T_C. We choose two different formulas to capture a stronger and weaker emphasis on the recency score of a tuple $Rel(t)$.

1) $Rec(t) = -|T_C - T_A|$
2) $Rec(t) = e^{-|T_C - T_A|}$

The first provides a linear decay, allowing older tuples the possibility to persist longer in the top-k set. The second provides an exponential decay, which heavily prefers newer tuples. In our preliminary experiments, we use the exponential decay version.

C. Relevancy

Relevancy $Rel(t)$ is a user-defined function regarding what tuples are considered relevant in a user's query. In our preliminary experiments, we do a simple keyword comparison. Given a user-defined set of keywords, we check to see the number of keywords that are found in the tweet.

D. Diversity

Diversity is a complex operation divided into three parts: Replacement Policy, Distance Function, and Distance Threshold. The descriptions of the two different Replacement Policies are found in Section V. The Distance Function is a user-defined function of two tuples $d(t_1, t_2)$ providing a single score that describes the distance between the two tuples in the attribute space. The Distance Threshold, d_{thresh} describes the maximum score between two tuples to consider them to be *neighbors*, i.e. similar to each other. The space that surrounds a single tuple t_0 with a distance of up to d_{thresh}, $(\forall t' \in \mathbb{T}, d(t_0, t') \leq d_{thresh})$ indicates the *neighborhood* that the tuple could represent, if chosen. The Distance Threshold is currently a user-defined value, and as a result is dependent on the chosen Distance Function and the attribute space. In our preliminary experiments, we use Cosine Similarity as our distance function, which generates the representation of the word co-occurrences of the pair of tweets in vector-space, then using the dot-product of the two tuple vectors as the distance between them.

V. StreamDiv Schemes

We discuss two replacement algorithms for maintaining the top k set as new tuples arrive. The first is an incremental replacement algorithm the processes each individual tuple as it arrives and decides whether to discard the new tuple or replace it with a tuple already in the top k list. The second, called the batch replacement algorithm is a more general version of the incremental algorithm and processes an incoming batch of more than one tuples to maintain the top k list.

A. Incremental Algorithm

The algorithm for the incremental replacement scheme is as follows:

Algorithm 1 Incremental

Require:
A set of k tuples S, a newly arrived tuple t, a relevance and recency combined score t_s of t and a radius r.

Ensure:
A top-k representative set S.

1: **if** $|T| < k$ **then**
2: $S = S \cup o_i$
3: **else**
4: **if** $\forall$ tuple $t_i \in T, d(t_i, t) > r$ **then**
5: $S = S \cup t$
6: Discard tuples with lowest combined score from S.
7: **end if**
8: **if** $\forall$ tuple $t_i \in S, |d(t_i, t) \leq r| = 1$ **then**
9: $S = S \cup t$
10: Discard tuples with lowest combined score between t and the tuple that eliminates t.
11: **end if**
12: **if** $\forall$ tuple $t_i \in T, |d(t_i, t) \leq r| > 1$ **then**
13: $S = S \cup t$
14: Discard tuples with lowest combined score between t and those tuples that eliminate t.
15: **end if**
16: **end if**
17: **Return** S

B. Batch Replacement Algorithm

TODO!

VI. Results

In this section we present some initial results we gathered from our preliminary experiments. We used Apache Storm v0.9.4 and Java OpenJDK v1.7. The topology we utilized in our experiments consisted of a Relevancy operator, followed by a Diversity Operator. As far as our dataset is concerned, we gathered tweets using Twitter's Streaming API, and accumulated 75000 tweets. Even though the Streaming API provides a plethora of attributes for each Tweet, we kept only the creation date and the text. The reason for that is because those are the only attributes important for our experiments (creation date for recency and text for relevance). During the gathering process, we ignored tweets that had non-ASCII characters, since those do not contain usable information and posed problems in storage.

TABLE I. Parameter Configuration

PARAMETER	VALUE
α	0.5
p	0.5
Radius	0.8
Dataset size	75000

In order to get a first impression on StreamDiv's effectiveness, we decided to compare it with the offline version. We compare StreamDiv with the offline implementation of combined relevance-and-diversity, which is named PrefDiv. The latter performs the same algorithm, but it reads data from a database and has a global view of the tweets. Our experiments were designed in a way to measure StreamDiv's performance on (i) diversity, (ii) relevance, and (iii) intensity.

For diversity, we calculated for each pair of tweets the Cosine Similarity distance. There is no particular reason that we opted for the previous metric, and the usage of alternative distances is orthogonal to our experiments. In terms of relevance, we used a simple approach, which annotated each tweet with a score, based on a given set of keywords. The score is just the number of keywords found in the text of the tweet. The result show in figures 1, 2, 3 are based on the statistic of the final top-k list of each model. Hence, for the incremental algorithm the result are constant.

Table I illustrated different parameters used in our experiments for both incremental and batch replacement version of StreamDiv, such that α is the scaling factor between relevancy and diversity, Radius is a parameter that use to distinguish between the similar and dissimilar items, for two tweets that has the cosine similarity larger then Radius we consider them as dissimilar.

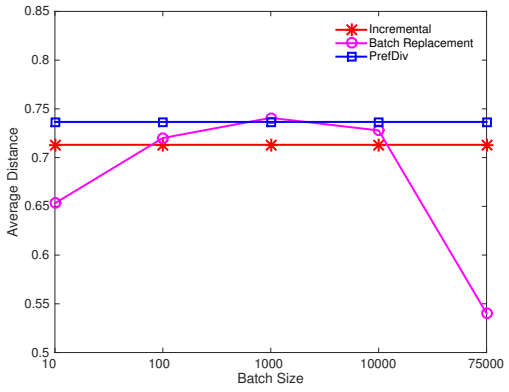

Fig. 1. Average Distance on different batch windows.

Figure 1 depicts the average distance achieved by using different batch sizes on StreamDiv. The incremental version of StreamDiv is not affected by the batch size, because it will always replace the same percentage of elements on every tuple. Therefore, the distance remains constant. The previous happens also in relevance and intensity. Turning to the batch version of StreamDiv, it is interesting how greater batch sizes, produce worse results in terms of diversity. The degradation of diversity is expected since the replacement policy of our prototype is naive at this point. In detail, the same percentage

of tuples is discarded at every epoch. Hence, even if we achieve higher diversity at some point, a large number of tuples will be discarded, and the average distance will drop.

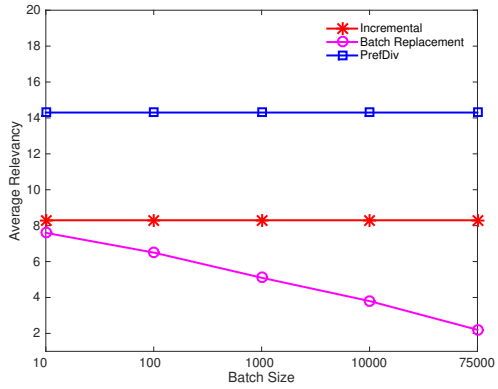

Fig. 2. Average Relevancy on different batch windows.

Turning to relevance, we monitored similar results with average distance (depicted in Fig. 2). The number of tuples that will be replaced increases with the batch size. Therefore, the batch approach will more likely throw away important tuples. The same story appears in normalized intensity (see Figure 3). We expect that with the dynamic reconfiguration of the percentage, we will achieve better results.

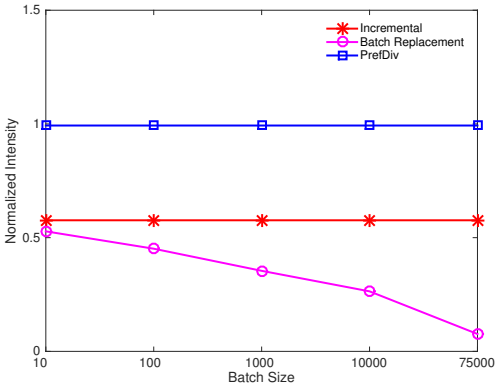

Fig. 3. Normalized Intensity on different batch windows.

VII. Future Work

In order for the paper to be published, below are list of things that we believe need to be improved:

A. Implement dynamic reconfiguration of p for batch approach

In order to make the windows replacement scheme work, we defined p as the coefficient of replacement, such that $0 \leq p \leq 1$. For every newly incoming window w_1, $k * p$ tuples from the top-k set will be replaced with the tuple in w_1. One naive way to implement the p is to choose a fixed number, so that for every new window we only replace a fixed portion of tuples. However the drawback for this type of approach is obvious, as it is almost impossible to find the optimal p that suits every input steam and window. In order to overcome this drawback, we purpose to implement a dynamic coefficient of replacement, such that:

$$f(p) = (\frac{Score_{new}}{Score_{old}}) * p \tag{3}$$

$$p = \begin{cases} f(p) & \text{if } 0 \leq p \leq 1 \\ p & \text{if otherwise} \end{cases} \tag{4}$$

Where $Score_{new}$ is the total score for all tuples in the new windows and $Score_{old}$ is the total score for all tuples in the old windows.

By adopt the dynamic coefficient of replacement, StreamDiv will be able to reflect the underlying trends of changes of intensity value for all successor windows.

B. Implement smarter batch replacement policy

As the performs of StreamDiv will be heavily depends on the policy of how we conduct batch replacement, we need to carefully study the best way to perform the batch replacement. Currently, we only adopted a very naive approach for the batch replacement policy, in which for every newly incoming window we will run the incremental replacement algorithm $k * p$ times. Such replacement policy does not guarantee the optimal performance.

In order to improvement the performance of the batch replacement policy. We purpose to implement following scheme. For each batch b_i, once a newly incoming tuple from b_i are being insert into the top-k set, we will mark it as non-removable, so that it will not be replaced by any tuple that are also from b_i even if it is a neighbors of other successor tuples in b_i. This way we can ensure a desired amount of tuples will be replaced for every batch, and instead of stop the replacement for current batch when $k * p$ times incremental algorithm are performed, the replacement will keeping running until $k * p$ tuples are being replaced. This way, the number of tuples being replaced in this windows will be more closely bind with the coefficient of replacement p. We believe by adopting both smarter batch replacement policy and dynamic reconfiguration we will be able to significantly improve the perform of StreamDiv.

C. Re-run experiments for a larger dataset.

One thing we have to admit is that our current experiment are very preliminary, since we have only used 75000 tweets as our current data set for the experiment, and our current data set are gathered using only the stop words (e.g. the, a, is) as keywords, hence the data set is lacking focuses. In order to improve our experiment we purpose to use more meaningful keywords along with more tweets data. So that the experiment can have greater convincingness.

Moreover, currently we only measure the statistics (e.g. average distance, normalized intensity and average relevancy) of the final result of each model, which means we only calculate all of the statistics once for the final top-k list of each model after the entire excitation is finished. However, for the future work, we would like to measure the average performance over each batch for both incremental and batch

replacement models for different batch sizes. With this more detailed set of experiments, the performance different between different models can be more clear.

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

[2] L. Chen, G. Cong, X. Cao, and K.-L. Tan, "Temporal spatial-keyword top-k publish/subscribe," in *ICDE*, 2015.

[3] P. Haghani, S. Michel, and K. Aberer, "Evaluating top-k queries over incomplete data streams," in *CIKM*, 2009, pp. 877 – 886.

[4] ——, "The gist of everything new: Personalized top-k processing over web 2.0 streamsh," in *CIKM*, 2010, pp. 489 – 498.

[5] K. Pripuic, I. P. arko, and K. Aberer, "Top-k/w publish/subscribe: Finding k most relevant publications in sliding time window w," in *DEBS*, 2008, pp. 127 – 138.

[6] L. Chen and G. Cong, "Diversity-aware top-k publish/subscribe for text stream," in *SIGMOD*, 2015.