Streaming Hierarchical Clustering for Concept Mining
DSCI560

Instructor： Young Cho, Ph.D.

Yifan Yang

8386626867

INTRODUCTION and HIERARCHICAL PARTITIONING

Highlights from what I read:

The approach is applicable to multilingual documents and multiple encodings, which can be automatically identified and converted into a common structure

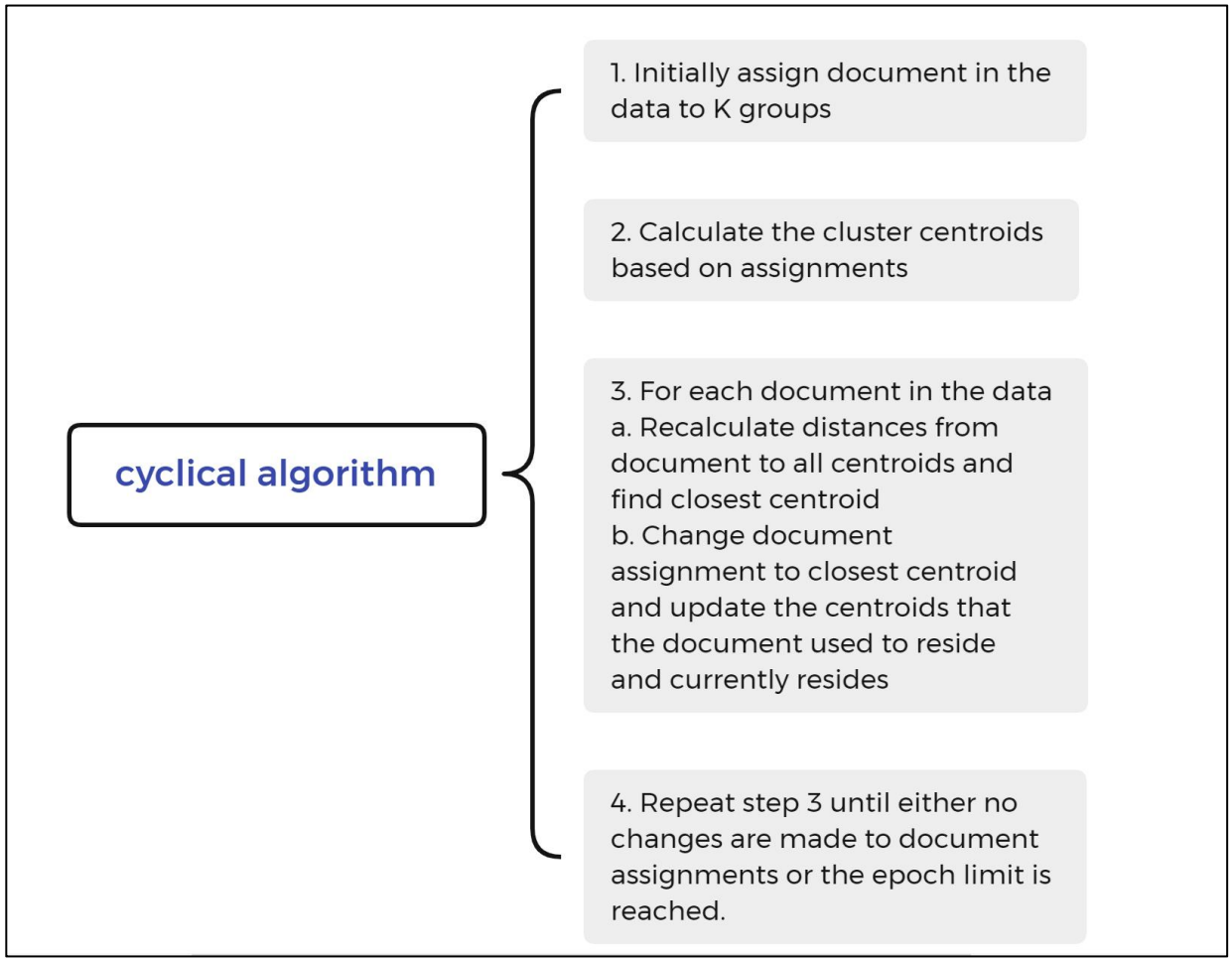
have developed novel, hardware-accelerated approaches to detecting known and unknown content, at line speeds

The AFE system is a High Speed Content classification system that works in three stages to classify flows of TCP traffic.

Two standard approaches can be taken to the problem organizing a collection of documents represented as fixed length vectors of high dimensionality hierarchically – agglomerative (bottom-up), and divisive (top-down).

$$\vec{c}(V) = \frac{1}{|V|}\sum_{\vec{v}\in V}\vec{v}\,, \qquad affinity(V,\vec{x}) = \vec{c}(V)\bullet\frac{\vec{x}}{|\vec{x}|}\,, \qquad score(V) = \sum_{\vec{v}\in V} affinity(V,\vec{v})\,.$$

EXPERIMENTAL RESULTS



HARDWARE DESIGN

Highlights from what I read:

Hierarchical partitioning was designed to be easily implemented on FPGAs without floating-point numbers. FPGAs can be used to implement floating point calculations, however the amount of resources needed to implement floating point arithmetic can reduce the amount of parallelism available.

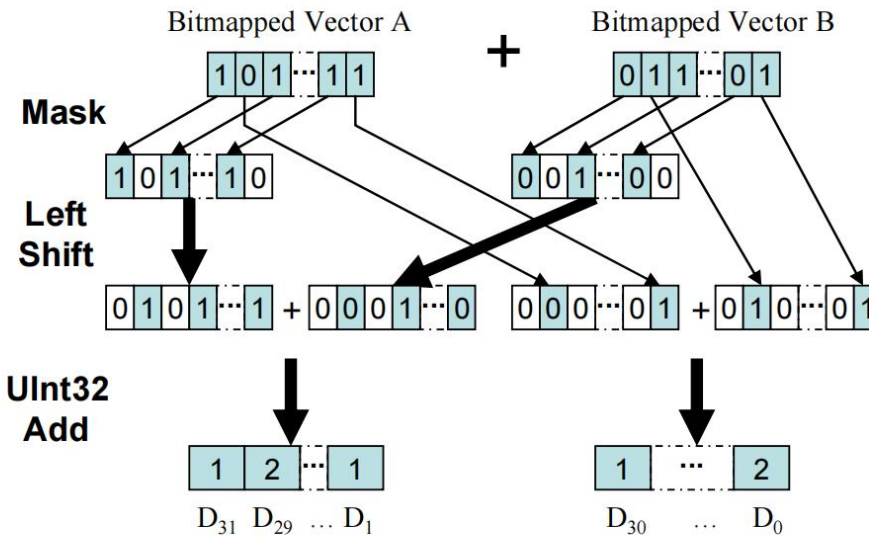


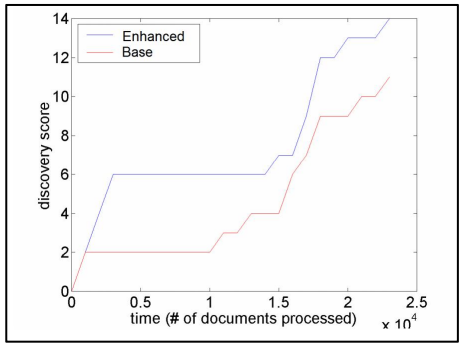
Figure 4: 32-dimension vector sum using Integer operations. The vector sum is accomplished using 6 instructions instead of 32x2=64 instructions.

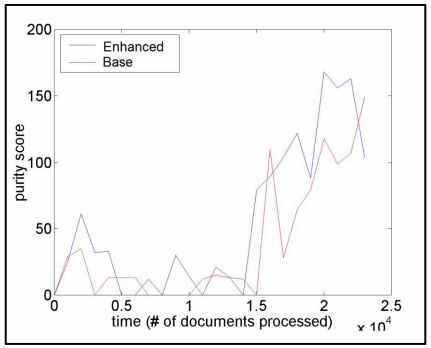
STREAMING HIERARCHICAL PARTITIONING

How to apply hierarchical partitioning clustering methods to handle evolving document streams. This method assumes that the collection of documents to be clustered is infinite and that the documents are presented one after another in sequence and need to be dynamically integrated into the current concept hierarchy.

STREAMING EXPERIMENTAL RESULTS

The experimental results are based on news group data and simulate concept drift according to the designed scheme.
Documents are randomly shuffled, but only half of the newsgroups' documents are initially presented. After even spacing, a newsgroup is gradually introduced into the distribution (thus gradually reducing the density of old newsgroups).
Uniformly distributed noise data from the experiment (chaff).
The ground-truth data from the experiment are displayed through charts to visualize the concept drift.
These experimental results demonstrate the performance of the streaming hierarchical partitioning clustering algorithm in handling simulated concept drift on newsgroup data.





CONCLUSION

Highlights from what I read:

The algorithms are designed to be implemented in hardware and capable of handling extremely high data ingestion rates. Experimental results show that the non-naive streaming hierarchical partitioning clustering method outperforms the naive variant in concept discovery. Future work will focus on integrating clustering into classification systems

Q1: What are the advantages of non-naive streaming hierarchical partitioning clustering methods over naive variants in concept discovery?

A1: The advantage of the non-naive streaming hierarchical partitioning clustering method over the naive variant in concept discovery lies in its ability to identify new concepts more efficiently. Through the analysis of experimental results, non-naive methods are better able to handle concept drift, i.e., concepts in the document stream may change over time. This approach is able to dynamically adjust clustering results, identifying new concepts and integrating them into the current concept hierarchy.