ADDITIONAL INFORMATION:
-Summary of your Doctoral Thesis
-Sponsoring Researcher Choice Justification
The maximum length for this section is 1 page (Verdana font size 10, letter size is suggested)

1 Summary of Doctoral Thesis

The increasing use of Internet and the wide spread of information technologies for textual content generation in news media and social networks have enabled the production of large volumes of text data. Since these large amounts of data do not generally fit in the main memory and the usage of secondary storage results prohibitive due to the high latencies involved, special techniques to process and extract valuable information from these massive volumes of text are needed.

Traditional clustering algorithms have problems to operate over text data whose computational representation is high dimensional and also when the document collections do not fit in main memory. An approach to deal with these issues consists in using clustering techniques capable of generating an output by making a single pass or scan on each input document. In this work, we propose a single-pass text clustering method capable of dealing with high dimensional and massive document collections under limited main memory space and secondary memory access constraints. To that end, we propose a summary data structure or sketch that allows to maintain reduced versions of the input documents and two efficient similarity estimation procedures that consider only the pairwise computation between near documents. The proposed sketch exploits the properties of two special families of functions that perform a single pass processing over each document and then obtain a reduced version of its computational representation. Additionally, this data structure enables a fast construction of a sparse similarity graph of documents, from which the final clusters are extracted by a repeated bisection method. Without the proposed technique, the construction of this (dense) document graph would involve the computation of similarity between every pair of documents within the collection, which would require a quadratic number of operations together with a storage of all document vectors in main memory. In contrast, a single-scan processing of text collections makes the information extraction task feasible on very large databases by solely performing operations in main memory. In turn, the similarity estimation only between near documents, allows to avoid several expensive operations between high dimensional vectors.

In order to evaluate this proposal, real text collections extracted from different domains (news media, technical reports, government reports and medical abstracts) are employed. Finally, the obtained performance is compared against a similar algorithm that builds a document graph by using exact similarities and then partitions it. Besides the attained performance scores, it is shown that a significant reduction of the original space dimensionality (up to 95%) is achieved without affecting the quality of the defined clusters.

2 Sponsoring Researcher choice justification

The main research activities of Dr. Allende-Cid are focused on automatic learning of algorithms in dynamic environments with distributed data. Specifically, he has contributed in the regression problem over distributed data collections, using machine neighborhoods built by using similarity measures between the estimated probability densities underlying he different data sources. In the last two years Dr. Allende-Cid has produced four articles published in indexed journals (ISI). Dr. Allende-Cid also started a project for initiation in research funded by FONDECYT (The National Fund for Scientific and Technological Development) at 2015. Allende-Cid also engaged in collaboration with the Distributed Systems research team at the *Universidade Federal de Alagoas* in Brazil.

It is very important to highlight that Dr. Allende-Cid is one of the few national research experts in distributed data processing, which is the field onto which I wish to guide my research. Therefore I expect as a result of this collaboration an powerful sinergy that ease the development of a new research line in Chile.