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 Resumen de la tesis doctoral del Investigador Responsable

El uso creciente de la Internet y la amplia propagación de tecnologías de la información para la generación de contenido textual en medios de noticias y redes sociales han posibilitado la producción de grandes volúmenes de texto. Debido a que generalmente estas cantidades de datos no caben en la memoria principal de los computadores usados actualmente y a que el uso de memoria secundaria resulta prohibitivo por sus altas latencias, se necesitan técnicas especiales para procesar y extraer información valiosa a partir de estos volúmenes masivos de texto.

Los métodos tradicionales de agrupamiento tiene problemas para operar sobre texto cuya representación computacional es altamente dimensional y además cuando la colección completa no cabe en memoria principal. Un enfoque para lidiar con estas dificultades consiste en usar técnicas capaces de generar un resultado realizando una sola pasada sobre cada dato o documento. Esto significa que solamente un vector representando a un documento se encuentra cargado en memoria principal a la vez, se extrae lo que se necesite de él y luego se descarta de la memoria. En este trabajo, se propone un método de clustering de una sola pasada capaz de operar eficientemente sobre datos con representaciones altamente dimensionales y de volúmenes masivos bajo restricciones de cantidad de memoria principal y acceso a memoria secundaria. Para este fin, se propone el uso de una estructura de datos que permite mantener versiones reducidas en tamaño de los datos originales y que además permite la estimación de similitud entre todos ellos mediante dos procedimientos eficientes también propuestos en este trabajo. Esta estructura resumen que se propone aprovecha las propiedades teóricas y de eficiencia de dos familias de funciones que permiten el procesamiento de una pasada sobre cada documento y la reducción de su uso de memoria, además de disminuir la cantidad de pares de documentos sobre los cuales se calcula su similitud. Adicionalmente, esta estructura de datos permite una construcción rapida de una red de documentos poco conexa, a partir de la cual se extraeran los grupos finales de documentos usando un método de disección. En ausencia del método propuesto, la construcción de esta red de documentos también sería posible, pero implicaría un alto costo computacional y de memoria, dado que la cantidad de cálculos de similitud entre pares de documentos y conexiones sería muy alta (cuadrática en la cantidad de documentos de la colección), además de tener que mantener todos los vectores asociados a los documentos cargados en memoria principal. En contraste con esto último, un procesamiento de la colección en una sola pasada hace factible la tarea de extracción de información sobre grandes bases de datos documentales, usando únicamente operaciones en memoria principal. En consecuencia, dado que la estimación de similitud no se realiza sobre todos pares de documentos (solamente entre los que probablemente son cercanos) y además que esta estimación se realiza sobre versiones reducidas de estos, este método permite procesar grandes volúmenes de datos altamente dimensionales.

Para evaluar esta propuesta, se usaran colecciones de texto reales generadas en distintos dominios (noticias, reportes técnicos, reglamentos gubernamentales y resúmenes de artículos científicos). Finalmente, el desempeño obtenido será comparado con el obtenido por algoritmo de disección de una red de documentos construída con similitudes exactas entre todos los documentos. Además de los desempeños alcanzados, se muestra que es posible alcanzar una disminución significativa en el espacio usado por los datos originales al emplear sus versiones reducidas, sin afectar la notoriamente la calidad de los agrupamientos obtenidos en varias colecciones documentales.

2 Justificación de la elección del Investigador Patrocinante

El área de investigación del Dr. Allende-Cid es aprendizaje automático en ambientes con datos cambiantes y distribuidos. Específicamente, ha realizado contribuciones en el problema de regresión sobre bases de datos distribuidas, usando vecindarios de máquinas establecidos mediante medidas de similitud entre las leyes de probabilidad subyacente de estas fuentes de datos. Durante los dos últimos años, el Dr. Allende-Cid ha generado 4 trabajos en revistas de corriente principal (ISI) y además se adjudicó un proyecto de investigación financiado por CONICYT el año 2015. Otro aspecto relevante en el desempeño de este investigador es su colaboración en un proyecto internacional con un grupo de investigación de Sistemas Distribuídos de la Universidade Federal de Alagoas (Brasil).

El Dr. Allende-Cid es uno de los pocos especialistas nacionales en el procesamiento de datos distribuidos, área hacia la que deseo extender mi línea de investigación. Por lo tanto, espero que como resultado de esta colaboración se genere sinergia, la cual tenga como consecuencia final el desarrollo de una nueva línea de investigación en el país.