**PROPOSAL ABSTRACT:**

| | |
|---|---|
| **Name of Principal Investigator:** | Juan Francisco Zamora Osorio |
| **Proposal Title:** | Clustering Distributed and High Dimensional Document Collections |

Describe the main issues to be addressed: goals, methodology and expected results. **The maximum length for this section is 1 page** (Verdana font size 10, letter size is suggested).

As a consequence of the explosive growth of the WEB and the used of social networks the task of automatic text clustering is becoming of greater importance due to the necessity to categorize increasing volumes of data (e.g. Tweets and News). In several scenarios, these are generated in a descentralized manner, e.g. collected documents by multiple machines in the context of search engines [1], even in different geographical zones. The generation of large amounts of documents surpases nowadays the capacity of personal computers and even of hiperformance computers. One way to deal with the problem of large volumes of data consists in dividing the collection in different machines and then to seize the local computation capabilities to obtain those groups or categories that represent the document collection. By means of the latter, first the the cost of storage is decreased, and later also the computation cost is reduced, because each machine processes less amount of data, and even parallel approaches can be used. Finally, the large size of the document collection makes the transmission over the network of these data to a centralized node for its storage and analysis, impractical. The Distributed Clustering attacks this point by identifying the groups of the complete collection by means of the integration of the local clusters created in each machine into a single model.

Existen importantes contribuciones en la tarea de agrupamiento de datos distribuídos. Sin embargo, gran parte de estos esfuerzos ha considerado conjuntos de datos de baja dimensionalidad en su representación vectorial, alcanzando como máximo una cantidad de dimensiones del orden de las centenas. Para el caso del texto, su representación vectorial es naturalmente de alta dimensionalidad, debido al tamaño de los vocabularios (cantidad de palabras distintas en la colección). Esta característica hace difícil su procesamiento debido a que muchos métodos ampliamente usados (e.g. basados en K-Means o en densidad) escalan respecto de la cantidad de datos, pero su costo es cuadrático respecto de la cantidad de atributos del espacio de características.

En este proyecto se desarrollaran nuevas técnicas de agrupamiento de documentos de texto capaces de operar sobre colecciones documentales repartidas en varios nodos de una red de computadores, aprovechando además las capacidades de computo individuales de cada nodo. Para esto, los métodos a desarrollar deberan ser capaces de integrar en un solo modelo aquellas categorías de documentos identificadas en distintas máquinas. Por otra parte, deberan aprovechar el paralelismo existente en los procesadores de cada máquina y que además sean eficientes en términos de la cantidad de pasadas realizadas sobre cada sub-colección de documentos. Esto último, debido a que cada fragmento de la colección también puede tener un tamaño considerable, e.g. varias veces la cantidad de memoria RAM disponible en cada máquina. Para validar los métodos propuestos se utilizaran varias colecciones documentales usadas en problemas de clasificación de documentos, mejora de recuperación en motores de búsqueda y agrupamiento de texto. Para medir la calidad de los modelos generados se contrastarán aquellos obtenidos de manera distribuída con los obtenidos centralizadamente (en los casos que sea posible debido al tamaño de las colecciones), usando medidas estándar en agrupamiento tales como *Rand score*, *Mutual Information*, *Homogeneity*, *Completeness score* y la media armónica de estos últimos dos que corresponde a la *V-Measure*. Las colecciones documentales disponibles no han sido diseñadas para probar algoritmos distribuídos, pero pueden ser fácilmente distribuídas en la validación como ya lo han hecho trabajos recientes.

[1] See for example how does Google operate to index textual content in`https://www.google.com/insidesearch/howsearchworks/crawling-indexing.html` (01.08.2016)