Clasificación de Información en BigData mediante la utilización de Técnicas de Inteligencia Artificial y Análisis de Redes Sociales

Claudio Delrieux, Damián Barry, Romina Stickar, Renato Mazzanti, Carlos E. Buckle, Rodrigo René Cura, Marcos Zárate

Departamento de Informática, Facultad de Ingeniería,
Universidad Nacional de la Patagonia San Juan Bosco.
Puerto Madryn, Chubut, Argentina. +54 280-4472885 – Int. 116
cad@uns.edu.ar, damian_barry@unpata.edu.ar, romistickar@gmail.com,
renato.mazzanti@gmail.com, carlos.buckle@gmail.com, rodrigo.renecura@gmail.com,
marcosdzarate@gmail.com

Resumen

El presente proyecto se enfoca a evaluar técnicas existentes e implementar desarrollos experimentales que permitan clasificar, ordenar, jerarquizar y analizar información sobre grandes volúmenes de datos heterogéneos (Big Data Analytics), mejorando y/o reformulando dichas técnicas.

Dichas técnicas permitirán establecer las capacidades necesarias con las que debería contar una base de datos de información masiva, tanto desde la perspectiva de almacenamiento y técnicas de indexación, como de distribución de las consultas, escalabilidad y rendimiento en ambientes heterogéneos.

Para ello se utilizarán Corpus científicos existentes de investigadores colaboradores del Centro Nacional Patagónico - CONICET, con la finalidad de contar con expertos que nos permitan realizar y estructurar los mecanismos de clasificación y evalúen los resultados de los desarrollos experimentales en estos Corpus. Adicionalmente a estos se analizará información pública digital en repositorios de portales y plataformas de información, tanto científica como de acceso general.

Palabras clave: big data analytics, extracción, clasificación, ontologías, escalabilidad.

Contexto

En la actualidad existe la necesidad de administrar grandes volúmenes de información no estructurada. Este incremento, debido a la enorme producción de información digital, ya sea desde la perspectiva de información producida en internet como así también la enorme cantidad de información producida por las empresas y organismos que en su gran mayoría, por no ser administradas correctamente, se termina dejando sin uso en algún repositorio de información. Normalmente esta información termina siendo eliminada sin evaluar correctamente su utilidad por parte de una comunidad.

Actualmente los motores de búsqueda, portales de internet y redes sociales han perfilado su enfoque a la gestión de enormes cantidades de información mediante el uso de bases de datos no

estructuradas y distribuidas, denominadas NoSql. Estas bases de datos han tomado relevancia no sólo en el ámbito de la Web 2.0 sino a nivel de las organizaciones y en el plano científico, debido esto a las técnicas tanto de almacenamiento como de recuperación de información (information retrieval).

Transversalidad en la utilización de técnicas aplicadas a este tipo de tecnología: lematización, tokenización, análisis fonético, data mining, sistemas distribuidos, infraestructura, permitiendo integrar distintos enfoques de investigación a una misma problemática. En particular el proyecto se centra en los aspectos de information retrieval, distribución, escalabilidad y disponibilidad.

Introducción

La producción y obtención de información ha pasado a ser uno de los grandes activos de las organizaciones, ya sean públicas, mixtas o privadas. En este sentido el desarrollo y estudio de la generación, administración, explotación, interpretación y clasificación de información se ha convertido en un desafío tecnológico y científico a nivel mundial. Para poder abordarlo, no sólo se requiere del soporte de científicos y tecnólogos en el área de la informática sino además de la integración con investigadores y expertos de distintas áreas vinculadas con las actividades que se desean analizar y comprender, donde a través de la conformación de equipos multidisciplinarios generen verdadero valor a la información circundante.

Por otra parte además en la actualidad existe gran cantidad de Corpus Documentales, artículos, sitios web, librerías digitales, que procesan grandes volúmenes de información, la cual es necesario manejar eficientemente. En suma se ha pasado de hablar de gigabyte de información a hablar con total normalidad del orden de los petabytes [Henderson, 2006; Abouzeid, BajdaPawlikowski, Abadi, Silberschatz, Rasin, 2009; Wei, Pierre, Chiy, 2010].

Esta situación ha generado el desafío de mejorar las herramientas de búsqueda en lo que se denomina “Information Search and Retrieval” utilizando para ello diversas técnicas que necesitan ser evaluadas, re-formuladas y si es posible mejoradas.

Asociado a esta problemática, se suma la necesidad de escalabilidad, disponibilidad y desempeño en el manejo de grandes volúmenes de información, situación que requiere de técnicas de sistemas distribuidos.

El volumen de información hace impensable utilizar mecanismos manuales supervisados para su clasificación, ordenamiento y uso, especialmente como aportes al análisis científico de datos.

Líneas de Investigación, Desarrollo e Innovación

La necesidad de contar con motores de bases de datos que puedan procesar grandes volúmenes de información no estructurada, donde la recuperación inteligente de dicha información pasa a ser crítica para las comunidades. Alternar tanto con técnicas de almacenamiento y distribución de información en ambientes heterogéneos como el análisis con técnicas de datamining para su subsecuente recuperación.

Hoy en día la información es un bien imprescindible de la sociedad, por lo tanto su correcta administración y acceso se convierte en una cuestión indispensable.

Experimentar y optimizar infraestructuras de bases de datos de gran volumen de información heterogénea. Implementar distintos servidores de bases de datos NoSql como: Hadoop / Hbase, Lucene /Solr, Casandra, CouchDB, especialmente por la disponibilidad de uso de Mapreduce como técnica de distribución inteligente de información, permitiendo escalabilidad, disponibilidad y recuperación eficiente e inteligente de la información almacenada.

Este Proyecto de investigación trae aparejadas diversas motivaciones socio-económicas. La sensibilización en la sociedad del uso de este tipo de herramientas y la correcta y eficiente administración de los recursos tanto de hardware como de software que permitan a los interesados contar con la información deseada en el momento requerido, permitiendo la correcta utilización de los mismos.

Los resultados aquí obtenidos podrán ser de directa aplicación tanto en organismos públicos como privados redundando en un beneficio económico no solo al optimizar los recursos tecnológicos sino además, y más importante, contar con información útil para la toma de decisiones.

Objetivos

Objetivos de Investigación y Desarrollo

- Investigar técnicas de Search & Information Retrieval, Data Mining, Knowledge Discovery, Análisis de Redes.
- Determinar la factibilidad y aplicabilidad de los métodos teóricos en los entornos prácticos estudiados, especialmente en lo referido a information retrieval y clasificación de información y clustering.
- Proponer mejoras o nuevas técnicas y/o re-formulaciones a las técnicas existentes para el manejo de recursos, en lo que se refiere a las técnicas de clasificación y clustering.
- Implementar y validar las técnicas y métodos propuestos sobre plataformas de desarrollo concretas.

Objetivos Académicos

- Consolidar mediante el proyecto, un grupo de investigación de la Universidad Nacional de la Patagonia San Juan Bosco sede Puerto Madryn, sobre la disciplina BigData y NoSQL. Este grupo se integra actualmente de 5 miembros, los cuales se encuentran abocados a la investigación a fin de crear nuevos métodos, desarrollos y trabajos de publicación científica para revistas, congresos de orden nacional e internacional.
- Generar espíritu crítico y contacto con las tareas de investigación en los alumnos de la carrera que que participan del proyecto, actualmente el grupo cuenta con 5 alumnos de la Licenciatura en Informática.
- Fomentar, incentivar y difundir las tareas de investigación.
- Mejorar la formación de recursos humanos altamente calificados, con capacidades de investigación y desarrollo. Lograr la categorización de los docentes participantes y la jerarquización del departamento de informática y

de la universidad en todos sus niveles.
- Consolidar las actividades realizadas por el Laboratorio de Investigación en Informática (LINVI) perteneciente al Departamento de Informática de la UNPSJB.
- Interactuar con otros grupos de investigación de las sedes de la universidad y de otras universidades, en tareas conjuntas de investigación y desarrollo, como también en la formación de recursos humanos.
- Incrementar el número de proyectos acreditados y de trabajos publicados por la universidad y la sede.
-

Metas del Proyecto

- Incrementar el número de proyectos acreditados y de trabajos publicados por la universidad y la sede.
- Seleccionar material bibliográfico y generar una base de conocimiento sobre las técnicas y métodos empleados en diferentes modelos de clasificación (Clasificador Bayesiano, Redes Neuronales, etc).
- Investigación de metodologías de razonamiento y explotación del conocimiento. Una vez que se cuente con un mecanismo de integración y alineamiento de ontologías, se procederá al estudio de metodologías de razonamiento y explotación del conocimiento, nuevamente a partir del análisis de requerimientos provisto por los expertos en el dominio, y utilizando estándares como por ejemplo RIF [Krötzsch, 2010] o análisis basado en grafos [Zhao y Ichise, 2012]
- Utilizar el conjunto de datos fácticos de carácter cuantitativo y cualitativo que han sido provistos mediante Data Mining por las investigaciones del CONICET-CENPAT, confrontando estos resultados con los mecanismos de simulación social y Análisis de Redes Sociales.
- Definir métricas que permitan obtener conclusiones relevantes respectos a las técnicas y métodos implementados.
- Armar un banco de pruebas que permita comprobar las distintas implementaciones y métodos utilizados para la clasificación de información. Según la observación de los resultados obtenidos, proponer mejoras y/o reformulaciones.

Referencias

1. Alani, H.: Position paper: “Ontology construction from online ontologies”, Proceedings ofthe 15th International Conference on World Wide Web (WWW '06), 2006.

2. Azza Abouzeid, Kamil BajdaPawlikowski, Daniel Abadi1, Avi Silberschatz, Alexander Rasin: “HadoopDB: An Architectural Hybrid of MapReduce and DBMS Technologies for Analytical Workloads”, VLDB Endowment, 2009.

3. Barceló, J. A. Experimentación y desarrollo de técnicas avanzadas de inteligencia artificial para la simulación computacional de la dinámica social y la evolución histórica. Proyecto I+D HAR2009-12258, 2010-2012, Universidad Autónoma de Barcelona. Ms. 2009a.

4. Barceló, J. A. Computational Intelligence in Archaeology. New York, Hershey, The Igi Group. 2009b.

5. Benslimane, S. M., Benslimane, D., and Malki, M.: "Acquiring OWL ontologies from data-intensive Web sites", Proceedings of the 6th International Conference on WebEngineering (ICWE '06), 2006

6. Davidsson P. Agent Based Social Simulation: A Computer Science View.Journal of Artificial Societies and Social Simulation vol. 5, no. 1. 2002.

7. Escolar, D., Salomón Tarquini, C. y Vezub, J. E. Proyecto I+D "Redes sociales indígenas y formación del estado en Cuyo, Pampa y Patagonia (1850-1900)". ANPCYT, PICT No 2011-1457, 2012-2014. 2011.

8. Euzenat J. and Shvaiko, P.: "Ontology matching", Springer-Verlag, 2007.

9. Gilbert, N. The Simulation of Social Processes. En Modèles et Systèmes Multi- Agents pour la Gestion de l'Environment et des Territoires. N. Ferrand (Ed.). Cemagref Editions, Clermont-Ferrand:121 - 137. 2000.

10. Gilbert, N. Agent-Based Models. London. Sage Publications Ltd. 2008.

11. Hanneman, R. A. y Riddle M. Introduction to social networks methods. University of California, Riverside. 2005.

12. Jennings, N. R. y Wooldridge, M. Applications of intelligent agents. Agent Technology: Foundations, Applications and Markets. 1998.

13. Joshi A., Jain P., Hitzler P., Yeh P., Verma K., Sheth A., and Damova M.: "Alignment-based querying of Linked Open Data". In R. Meersman, H. Panetto, T. Dillon, S. Rinderle-Ma,P. Dadam, X. Zhou, S. Pearson, A. Ferscha, S. Bergamaschi, and I. F. Cruz, editors, On the Move to Meaningful Internet Systems: OTM 2012, Confederated International Conferences. Volume 7566 of Lecture Notes in Computer Science. Springer, Heidelberg, 2012.