Análisis de Documentos para Medicina Basada en Evidencia. Una propuesta con Big Data Analytics

Document Analysis for Evidence Based Medicine. A proposal with Big Data Analytics

Gustavo Verduzco Reyes [*]
Ernesto Bautista Thompson [†]
Jorge A. Ruiz Vanoye [‡]
Alejandro Fuentes Penna [§]

Recibido el 05 de Mayo de 2017. Aceptado el 12 de Julio de 2017.

Resumen

Big data analytics es una tecnología que contempla el almacenamiento, administración y análisis de grandes volúmenes de datos, su aplicación en el campo de la salud ha sido reciente. El presente trabajo se centra en una propuesta de análisis de documentos de tres fases: Diagnóstico, Análisis y Evaluación. La propuesta forma parte de un trabajo de tesis en desarrollo, la cual ha identificado como resultado las fuentes de datos para su análisis como Cochrane, ACP Journal y PUBMED. También las técnicas de análisis a utilizar como Máquinas de vector de soporte, *Naive Bayes* y *Cluster k-means*, así como las técnicas de evaluación matriz de confusión y curva ROC. Esta propuesta es un apoyo en la toma de decisiones para los profesionales de la salud, al permitirles realizar mejores diagnósticos médicos.

Palabras Clave: Técnicas de análisis, Big Data Analytics, Medicina basada en evidencia.

Abstract

[*] Estudiante del Doctorado en Planeación Estratégica y Dirección de Tecnología: gustavo.verduzco@upaep.edu.mx
[†] Profesor Investigador Externo UPAE: eb_thompson@yahoo.com
[‡] Profesor Investigador Universidad Autónoma del Estado de Hidalgo: jorge@ruizvanoye.com
[§] Profesor Investigador Universidad Autónoma del Estado de Hidalgo: alexfp10@hotmail.com

Big data analytics is a technology that includes the storage, administration and analysis of large volumes of data, its application in the field of health has been recent. This paper focuses on a proposal for analysis of three-phase documents: diagnosis, analysis and evaluation. The proposal forms part of a thesis work under development, which has identified the data sources for its analysis as Cochrane, ACP Journal, PUBMED. Also the analysis techniques to use as Support vector machines, Naive Bayes and Cluster k-means. As well as evaluation techniques such as confusion matrix and ROC curve. This proposal is a support in the decision making for the health professionals, allowing them to make better medical diagnoses.

Keywords: Algorithms, Big Data Analytics, Evidence-based Medicine

Introducción

Big data se refiere a la explosión en cantidad (en ocasiones calidad) de los datos disponibles y potencialmente relevantes, en gran parte por el resultado del registro sin precedentes de datos y las tecnologías de almacenamiento (Diebold, 2012). *Big data analytics* se ha utilizado para describir la relación entre datos y técnicas de análisis de datos que requieren el manejo de terabytes o exabytes de datos y donde es indispensable la tecnología para almacenar, administrar y analizar datos. *Big data analytics* se ha aplicado en varios campos como análisis financiero, análisis político, *marketing*, tendencias sociales, etc.

Recientemente ha habido un especial interés en la aplicación de *big data analytics* en el campo de la salud. Por ejemplo, Belle et al. (2015) examinaron tres campos de la salud, el procesamiento de imágenes, el procesamiento de señales y la información genómica. Por otro lado, Simpao, Ahumada, & Rehman (2015) señalan que el Instituto de Calidad de Anestesía y el grupo de Multicentrico Perioperatorio han generado una gran base de datos que al aplicarle análisis de datos les ha permitido realizar una evaluación predictiva de riesgos, apoyo a decisiones clínicas y gestión de recursos. Chawla & Davis (2013) proponen un modelo de predicción y administración de enfermedad centrado en el paciente. El modelo está basado en una técnica de minería de datos llamada filtrado colaborativo.

Por otro lado, la Medicina Basada en la Evidencia (MBE) es la utilización consciente, explícita y juiciosa de la mejor evidencia científica clínica disponible para tomar decisiones sobre el cuidado de los pacientes (Sackett et al.,1996). Para Mellis (2015) la MBE busca aprovecharse de la experiencia del médico y combinar sus conocimientos adquiridos en la formación profesional y sustentada en la práctica clínica, junto con la mejor evidencia externa disponible producto de la investigación científica. Por todo el mundo se ha generado mucha evidencia médica útil que puede ser tratada con técnicas de *big data analytics* y tal como se menciona en Brennan & Bakken (2015) la exploración de esta clase de información podría llevarse a cabo a través del agrupamiento de datos, visualización, técnicas de reducción de dimensionalidad, modelado y predicción de datos. Aunque hay varias fuentes de datos de medicina basada en la evidencia como Medline, PubMed y Cochrane, no se han encontrado o reportado en la literatura científica, estudios que apliquen el *big data analytics* a este campo específico de la salud (Montori & Guyatt, 2008).

La MBE aboga para que un médico se valga del análisis de la literatura médica para realizar un mejor diagnóstico y presentar un dictamen al paciente que acude a sus servicios. Revisar toda la literatura sería exhaustivo para un médico, sin embargo, si se realiza de forma automatizada a través de *big data analytics*, no solo se beneficiaría al médico, quien incrementaría su bagaje de conocimientos, sino principalmente al paciente. Por tanto, el objetivo de esta investigación es presentar una propuesta metodológica que integre las técnicas de minería de datos y estadística, que permitan extraer información relevante de documentos de literatura en MBE para la toma de decisiones médicas. Esto conlleva algunas implicaciones como analizar las fuentes de información médica en MBE, seleccionar aquellas técnicas de analítica de datos y estadística

para clasificar y seleccionar la información y valorar los resultados de los análisis y compararlos con los existentes en el mercado.

Las preguntas de investigación son las siguientes: ¿puede el *big data analytics* contribuir a mejorar la práctica de la medicina basada en evidencia? ¿existe alguna combinación de técnicas de analítica de datos y estadística que permita extraer la información más relevante de MBE?, ¿los resultados obtenidos con las técnicas analíticas aplicados a la MBE son mejores que las herramientas existentes?

El llevar a cabo esta investigación se justifica porque actualmente no se ha reportado en la literatura la aplicación de técnicas de analítica de datos en MBE, así mismo hay una carencia de combinación de estas técnicas que permitan filtrar la información de forma automática y entendible para el uso de los profesionales de la salud.

1. Marco Teórico

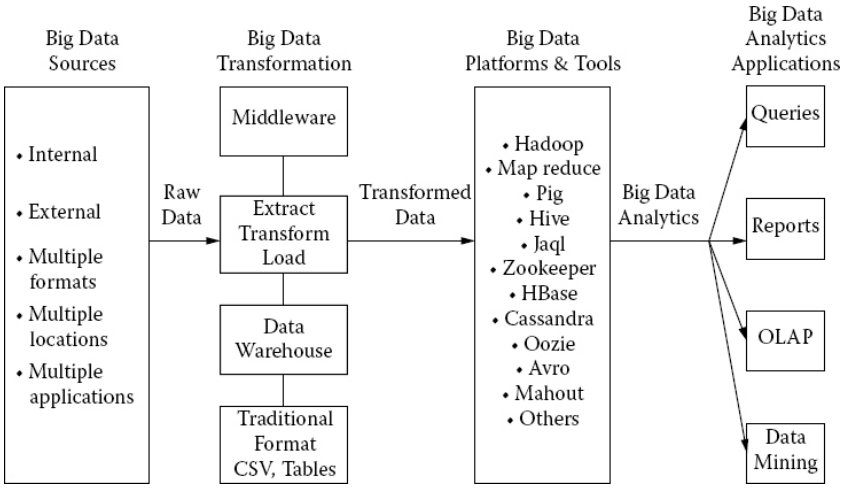
Big data se refiere a grandes conjuntos de datos cuyo formato puede estar de manera estructurada y no estructurada. A manera de ejemplo, tenemos los historiales del Internet, correos electrónicos, documentos de textos, transacciones comerciales, anuncios comerciales, análisis médicos, páginas web, etc. En Jiang & Leung (2015) se presentan cinco características que lo distinguen: veracidad, velocidad, variedad y volumen.

- *Veracidad*, se refiere a la confiabilidad de los datos.
- *Velocidad*, es la rapidez con la que los datos son generados y almacenados.
- *Valor*, se enfoca a la utilidad de los datos.
- *Variedad*, se refiere a los tipos de formatos de datos y los contenidos.
- *Volumen*, se refiere a la cantidad de datos.

Esta información requiere ser analizada a través de técnicas estadísticas y científicas, por tanto el big data analytics es el proceso de extraer información de un conjunto grande de datos, de forma ordenada, para hacer predicciones y estimados acerca de futuros resultados (Larose et al., 2015).

En su libro, Kudyba (2013) presenta la arquitectura de *big data analytics* (ver Figura 1), en la que se destaca el proceso que se sigue para tratar un gran conjunto de datos. Primero, los datos son adquiridos de varias fuentes, que pueden ser internas, es decir datos de una misma organización, o externas, de otras organizaciones. Luego son preparados para ser tratados, a través de software especializado para unificar diversos formatos de archivos, sean estos estructurados o no estructurados.

Figura 1. Arquitectura de Big Data Analytics



Fuente: Kudyba, 2013.

Posteriormente se aplican una serie de herramientas tecnológicas para almacenar y manipular esos datos. Entre las herramientas más conocidas de software libre y propietario destacan las siguientes:

• *Hadoop:* Tecnología desarrollada originalmente por la empresa yahoo® para su buscador. Procesa datos a gran escala. Está formado por dos componentes, MapReduce para el procesamiento de datos y un Administrador de Archivos Distribuido (DFS).

•*Google Bigtables*®: Es un sistema distribuido para grandes volúmenes de datos. Puede manejar petabytes de datos en datos en cientos de servidores.

•*PIG*: Es un lenguaje de alto nivel para flujo de datos. Puede usarse por los programadores para la analítica de datos. Puede usar paralelización para grandes volúmenes de datos.

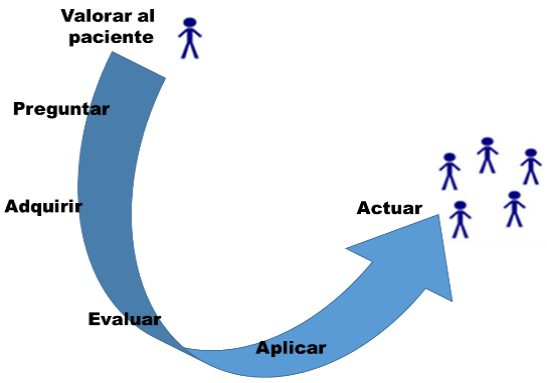
•*MAHOUT*: Es un biblioteca escalable para aprendizaje de máquina. Desde 2014 MAHOUT ha reemplazado a MapReduce en algunas aplicaciones debido a que es más eficiente en ejecución que MapReduce.

•*NoSQL* (Not Only SQL): Es un ambiente de base de datos No relacional. Organiza de forma rápida la información de un alto volumen de datos.

Finalmente se aplica la analítica de datos, en forma de consultas, reportes, técnicas de minería de datos, técnicas estadísticas. Los resultados de la analítica son presentados de forma gráfica mediante distintos tipos de diagramas, árboles, diagramas de pastel, líneas, 3D, etc.

En lo que respecta a la medicina basada en la evidencia (MBE), ésta consiste en aprovecharse de la experiencia del médico y complementarla con estudios científicos validados y probados que contribuyan a mejorar el diagnóstico de un médico y que por ende permitan generar un mejor tratamiento para un paciente. En Mellis (2015) hallamos una descripción más detallada del proceso de la MBE (ver Figura 2) la cual consiste en valorar, preguntar, adquirir, evaluar, aplicar y actuar.

Figura 2. Proceso de la Medicina Basada en Evidencia



Fuente: Mellis, 2015

•*Valorar al paciente:* Valorar clínicamente al paciente y reconocer las lagunas de información.

•*Preguntar:* Hacer una pregunta clínica en base a las lagunas de información, se puede usar el formato población, comparación, resultado.

•*Adquirir:* Adquirir la evidencia usando una o más bases de datos adecuadas y una estrategia de búsqueda eficiente.

•*Evaluar:* Evaluar la evidencia, determinar la importancia de los resultados, en la precisión de los mismos.

•*Aplicar:* En esta fase se mide la aplicación de la evidencia de la investigación en relación al paciente, cómo se ajusta el problema de investigación, la población de estudio y pregunta clínica del paciente.

•*Actuar:* Supone tener confianza en la evidencia, se tiene un riesgo mínimo de sesgo. El resultado es importante para el paciente (hospitalización, terapia, calidad de vida, etc.). El médico ha evaluado los riesgos y beneficios para el paciente, le presenta las opciones que puede elegir o no elegir ninguna.

Georgiou (2002) señala que debe haber bibliotecas digitales de salud confiables y extensas que permitan ser explotadas mediante tecnologías de la información para realizar una buena práctica de la medicina basada en la evidencia. Aunque actualmente hay bibliotecas que concentran mucha evidencia médica como Medline, Pubmed, Cochrane, así como también buscadores como google®, cada vez se requieren mecanismos tecnológicos más eficientes que permitan explotar la información para el uso médico y la mejor atención de los pacientes (Montori & Guyatt, 2008).

De hecho, Spink et al. (2004) realizaron un estudio acerca de cómo la gente hace consultas médicas o relacionadas con la salud, usando buscadores de salud tanto especializados como no especializados y hallaron que generalmente las personas no encuentran resultados esperados, por lo que exploran solo las primeras 10 o 20 páginas web que contienen resultados médicos.

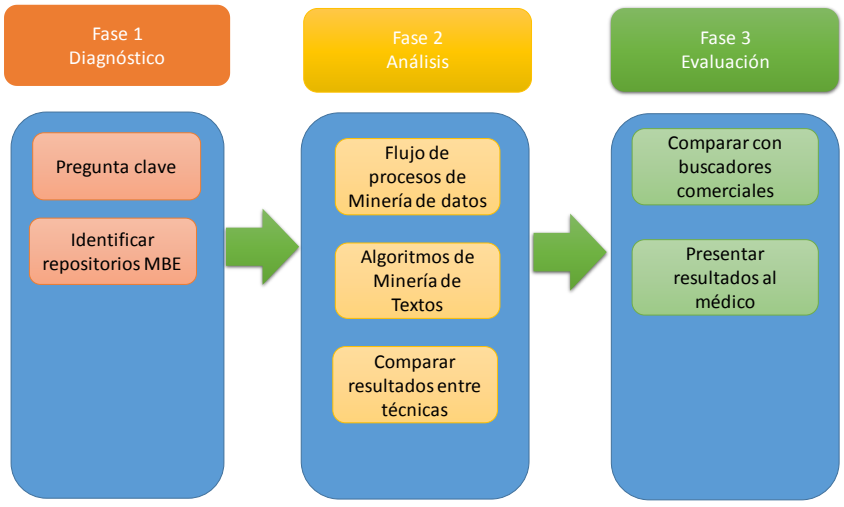
En la literatura científica se reportan algunos esfuerzos por analizar documentos médicos, por ejemplo, Zhang et al. (2014) presenta una herramienta de búsqueda basada en el API del buscador Bing y habilitando una función de dispersión agrupación, usa también la librería Weka Machine-Learning para la minería de textos y los *clusters* de datos. Esto ayuda a los usuarios a agrupar los resultados y hacer análisis en espacios de búsqueda muy grandes. Britt et al. (2008) proponen un software usando dos técnicas para ordenar y agrupar datos de documentos, una es la indexación semántica latente (LSI) que es un modelo de espacio vectorial; la otra es un analizador de textos general (GTP), el cual es un ambiente de software para generar modelos LSI. Dicho ambiente de software soporta también la frecuencia de documentos y la normalización de documentos. Lam et al. (2016) analizan el problema del trastorno de sueño (SD) tomando como base 3,720 artículos relacionados con el sueño de la biblioteca PubMed, aplicando minería de texto en esos documentos usando MetaMap y los cluster en combinación

con regresión logística, observando una tendencia a la alza en documentos que apuntaban a términos como Insomnio y parasomnia, en tanto que a la baja el término relacionado con la respiración.

2. Metodología

Tal como menciona Cazau (2006), este trabajo es una investigación exploratoria con enfoque de análisis cuantitativo, ya que en la revisión del estado del arte no se encontraron trabajos que aborden las técnicas de *big data analytics* aplicada al análisis de documentos de medicina basada en la evidencia, por tanto, nos hallamos ante un nicho de oportunidad para ser abordado. Se ha desarrollado una propuesta que integra técnicas de analítica de datos aplicadas a la medicina basada en evidencia (ver Figura 3) la cual consta de tres fases: Diagnóstico, Análisis y Evaluación, las cuales se describen a continuación.

Figura 3. Propuesta de Análisis de documentos en MBE



Fuente: elaboración propia

Fase 1. Diagnóstico

En esta fase, el médico valora la condición del paciente y en base a su experiencia determina la posible enfermedad que lo afecta. No todas las consultas al médico requerirán que éste realice una consulta a las bases de datos especializadas en medicina basada en la evidencia (MBE), sino aquellas en las que existan lagunas de información.

Pregunta clave.- En base a un diagnóstico preliminar del paciente, el médico valora e identifica cuál es la laguna de información. Entonces genera una pregunta clave en forma de oración que será la consulta a los repositorios de medicina basada en evidencia.

Identificar repositorios de Medicina Basada en Evidencia.- En vista de que hay varias fuentes de información de medicina basadas en evidencia como Cochrane, ACP Journal Club, así como no especializados tales como PUBMED, es necesario elegir uno de los repositorios, debido a que el tiempo de búsqueda dependerá de forma directa del repositorio elegido y del número de estudios analizados.

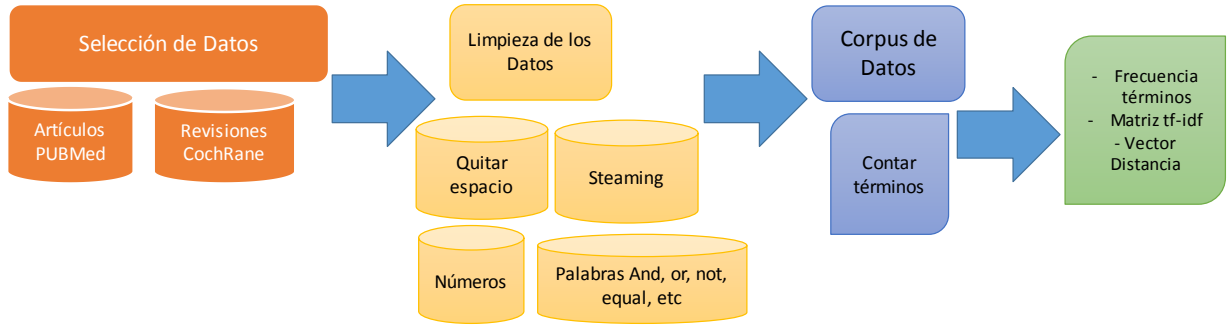
Fase 2. Análisis

En esta fase se realiza un análisis de los contenidos de los repositorios de medicina basada en la evidencia; asimismo, se llevan a cabo las pruebas a través de algoritmos de selección y clasificación.

Flujo de Procesos de Minería de textos.- Hay varios modelos de minería de datos en la industria como CRISP-DM (Cross Industry Standard Process for Data Mining), compuesto por seis fases: entendimiento del negocio, entendimiento de los datos, preparación de los datos, evaluación y desarrollo (Hiltbrand, 2013); SEMMA (Sample, Explore, Modify, Model, Assess), el cual consta de cinco fases Muestra, Explorar, Modificar, Modelar, Evaluación (Balkan & Goul, 2010);

Modelo de dos cuervos o el proceso KDD (Knowledge Discovery in Databases), descubrimiento de conocimiento en bases de datos. Nuestro flujo de procesos de análisis de textos (ver figura 4) incluye un proceso lineal similar al que se realiza en KDD.

Figura 4. Proceso de minería de datos



Fuente: Elaboración propia

Selección de datos.- Implica elegir las bases de datos que servirán para llevar a cabo los análisis. Pueden ser especializados como Cochrane, ACP Journal o no especializados como PUBMED.

Limpieza de datos.- La base de datos se prepara para ser analizada, se eliminan los espacios en blanco, así como las palabras que no agregan significado a las oraciones como conectores and, equal, or, entre otros. También se busca el origen de las palabras para reducir términos, por ejemplo kids, se reduce a kid.

Corpus de datos.- La base de datos está lista para ser analizada a través de los algoritmos de minería de textos. Para agilizar el proceso los datos son almacenados en vectores y matrices de memoria.

Preparación de datos para análisis.- Se generan las matrices de datos como frecuencia de términos (tf) por documento, frecuencia de cada término en relación con todos los demás documentos (tf-idf).

Algoritmos de Minería de textos.- Incluye las pruebas a los algoritmos de minería de textos existentes, entre los que destacan los árboles de decisión, naives bayes, clustering (ver tabla 1). Las pruebas incluyen número de documentos analizados, velocidad de respuesta e índice de certeza. Los resultados derivados de cada uno se podrán combinar para tener una mejor perspectiva del problema que se está analizando. En la tabla 1 se presentan algoritmos de aprendizaje de máquina.

Tabla 1. Algoritmos de Aprendizaje de máquina

Algoritmo de Aprendizaje de Máquina	Descripción
Filtrado Colaborativo	Permiten crear perfiles de los gustos de los usuarios y luego empatarlos con otros usuarios que tuvieron gustos afines. Estas técnicas se emplean mucho por las tiendas de ventas on-line o de renta de películas. Se los considera también como técnicas de recomendación.
Árboles de decisión	Se emplean para tomar decisiones en base a dos posibles soluciones si/no, mediante interacciones sucesivas y ponderados por un valor, se va llegando a una solución.
Naive Bayes	Algoritmo basado en el conocido teorema de Bayes. Se emplea generalmente como un clasificador de datos.
Clustering	Estos algoritmos permiten dividir los datos en subconjuntos de datos llamados *clusters*. Hay varias técnicas K-means clustering, Spectral Clustering

Fuente: Elaboración propia

Comparar resultados entre técnicas.- Esta sección del análisis es clave, los resultados son medidos a través de medidas estadísticas de tal forma que se pueda elegir una sola técnica o combinación de técnicas de minería de datos. En un principio, la técnica que muestre el nivel más alto de precisión en la búsqueda será la candidata a ser elegida. A fin de medir la eficiencia de los algoritmos de análisis de documentos se aplicaran dos técnicas de medida, la matriz de confusión (Corso, 2009) y la curva ROC (Prati, Batista, & Monard, 2008).

Fase 3. Evaluación

En esta fase los resultados derivados de las técnicas de minería de textos son comparados con productos disponibles en el Internet, sean comerciales o libres. Luego se presentan al médico los resultados.

Optimización de técnicas.- Hay algunos productos comerciales en Internet, como PUBMED del gobierno de Estados Unidos, donde las consultas a diversos temas arrojan resultados de distintos estudios que abordan el tema. Se harán experimentos que permitan analizar el comportamiento de varias técnicas de análisis, los resultados se compararán con los generados por la herramienta comercial PUBMED.

Presentar resultados al médico.- Los resultados se presentan en forma de una lista lineal o mediante un gráfico de bloques con todos los documentos que guardan relación directa con la pregunta clave que se planteó en la fase 1. El médico podrá ver el contenido del estudio que mejor le convenga y entonces combinar su conocimiento con los resultados de la minería de textos en salud. Esto ayudará al médico a presentar un mejor diagnóstico al paciente.

Cabe señalar que la propuesta de análisis de documentos es relevante en vista de que integra la medicina basada en evidencia y la parte de la analítica de datos empleada en big data. Por otro

lado, la propuesta está diseñada para elegir las mejores técnicas de análisis de documentos que permitan obtener los resultados más óptimos para el filtrado de documentos en MBE.

3. Resultados

El análisis de documentos en salud está cobrando mucha importancia, máxime porque al aplicar técnicas de analítica de datos se ha encontrado información muy útil que de otra forma no se hubiera descubierto. Por ejemplo, Rojas et al., (2016) hicieron una revisión de 74 artículos médicos, para conocer cómo se aplica el proceso de minería de datos en salud, hallaron que los temas más estudiados fueron las enfermedades oncológicas y cirugía. Por su parte Uramoto et al. (2004) se enfocaron en analizar la enfermedad de la leucemia en 1,051 resúmenes de documentos de MEDLINE, sus hallazgos lo condujeron a encontrar unas proteínas para combatir dicha enfermedad por medio del desarrollo de un fármaco. Los trabajos antes citados dan cuenta de la importancia de contar con una fuente confiable de datos a analizar, en la presente propuesta en desarrollo, ya se han identificado varios repositorios de documentos médicos. Hay tres fuentes disponibles y libres al público en general. El primero es PubMed (www.ncbi.nlm.nih.gov/pubmed/); este sitio web de salud no especializado en medicina basada en evidencia, contiene alrededor de 27 millones de registros de documentos médicos de distintos temas de salud como cáncer, cardiología, instrumentación médica, pediatría, etc. El segundo es ACP Journal Club (www.acpjc.org/), sitio especializado en medicina basada en la evidencia, contiene el resumen de la mejor medicina interna de más de 130 revistas clínicas e incluye el período desde 1991 hasta 2008. La última fuente de datos es Cochrane (www.cochrane.org/), que también es un portal especializado en medicina basado en la evidencia y contiene cientos de

documentos con temas tan diversos como el cáncer, el corazón y la circulación, la neurología, las enfermedades mentales, la gastroenterología, entre otros.

Por otro lado, Cerrito & Cerrito (2006) emplearon el software estadístico SAS Text Miner® para analizar los registros médicos de un periodo de alrededor de 6 meses; cuando aplicaron el algoritmo cluster hallaron que a los pacientes con enfermedades similares se les daba tratamiento totalmente diferente, lo cual afectaba el costo de sus medicamentos. Así mismo, Raja et al. (2008) examinaron 1,500 registros médicos de pacientes, lo que les permitió identificar 18 clusters, resultado que apuntaba a hallar pacientes con enfermedades similares. Santiso et al. (2016) hallaron que es posible pronosticar si un paciente tiene alguna reacción adversa a un medicamento empleando dos técnicas de análisis: Bayes y Bosque aleatorio. Lemke & Mueller (2003) examinaron 200 casos de pacientes, emplearon el software propietario KnowledgeMiner® y aplicaron clasificación binaria para detectar y prevenir problemas del corazón. Como hemos visto, es muy importante seleccionar técnicas de análisis de documentos; para nuestro caso en la propuesta que está en desarrollo, emplearemos al menos tres técnicas que al combinarlas se espera mejoren la búsqueda en relación a buscadores comerciales como PUBMED; las técnicas son Máquinas de vector de soporte, Naive Bayes y cluster k-means.

Los estudios citados dan cuenta de que es posible realizar análisis de documentos médicos con resultados plausibles, por tanto la propuesta presentada en este trabajo, misma que se encuentra en desarrollo como una tesis, tiene el potencial de convertirse en un analizador de documentos de MBE para encontrar información relevante para los profesionales de la salud.

4. Discusión

Actualmente la cantidad de información médica disponible en Internet es asombrosamente grande, sin embargo, tanto médicos como pacientes requieren que esta información esté disponible de forma sencilla y en un formato que sea fácil de comprender. Aunque hay una docena de buscadores médicos, es importante que los profesionales de la salud conozcan las debilidades, fortalezas y confiabilidad de cada buscador. Aunque no siempre los médicos tienen el tiempo para hacer esta tarea (Cañedo, 2011). Tal como observamos en el marco teórico, para lograr que la medicina basada en evidencia esté al alcance médico, se requiere contar con toda una infraestructura tecnológica que permita almacenar, administrar y analizar estos datos o documentos, bondades que ofrece el *big data analytics*.

Se presentó una propuesta de análisis de documentos para medicina basada en evidencia que cuenta con tres fases: Diagnóstico, Análisis y Evaluación. Los hallazgos apuntan a que es factible hacer la carga de datos de distintas fuentes, dentro de estas bases de datos de documentos hay tanto especializadas como no especializadas. En cuanto al Pre procesamiento, al aplicar los filtrados de datos, estos reducen la cantidad de información a analizar. En el caso de la analítica aunque aún no se aplicaron las técnicas de análisis por ser un trabajo en desarrollo, muchos de los algoritmos ya han sido probados con éxito en otros trabajos como los que se citaron a lo largo de esta investigación. En el caso de los resultados, aún hace falta aplicar la matriz de confusión y curva ROC para comparar el rendimiento de las distintas técnicas.

Aunque la propuesta de análisis de documentos en MBE cubre la analítica de datos, no aborda el almacenamiento masivo de datos expresado en terabytes o exabytes. Esto se debe a que la investigación está limitada por los recursos disponibles de cómputo, para este efecto, una computadora con cuatro núcleos, 4 Gb de memoria y 750 Gb de disco duro, características limitadas para instalar una herramienta que gestione grandes volúmenes de datos como PIG o

Hadoop. Sin embargo, el trabajo podrá ser escalado a una herramienta de almacenamiento masivo de datos.

5. Conclusiones

En este trabajo se ha presentado una propuesta de análisis de documentos en MBE. Aplicar esta propuesta tendría un impacto importante ya que reduciría el tiempo que invierte un médico a analizar un gran conjunto de estudios médicos en MBE, ya que tendría a su alcance solo los documentos importantes que guarden relación con el tema a tratar con el paciente. También, contribuiría a que los diagnósticos del especialista en salud sean más asertivos y por tanto el tratamiento más efectivo para los pacientes. Aunque la propuesta tiene la limitación de que no contempla el uso de las herramientas de almacenamiento masivo de datos en la forma de un clúster de cómputo, un trabajo futuro sería realizar pruebas a las herramientas libres de almacenamiento y administración de datos de *big data*.

Referencias

BAKAN, S. y GOUL, M. (2010). "Advances in Predictive Modeling: How In-Database Analytics Will Evolve to Change the Game". *Business Inteligence Journal*, 15(2). Retrieved from https://www.researchgate.net/profile/Sule_Balkan/publication/264905640_Advances_in_Predictive_Modeling_How_In-Database_Analytics_Will_Evolve_to_Change_the_Game/links/53f5c3870cf22be01c3faa29.pdf

BELLE, A., THIAGARAJAN, R., SOROUSHMEHR, S. M. R., NAVIDI, F., BEARD, D. A., & NAJARIAN, K. (2015). "Big Data Analytics in Healthcare". *BioMed Research International*, 2015, 1–16. https://doi.org/10.1155/2015/370194

BRENNAN, P. F., & BAKKEN, S. (2015). "Nursing needs big data and big data needs nursing". *Journal of Nursing Scholarship*, 47(5), 477–484.

BRITT, B. L., BERRY, M. W., BROWNE, M., MERRELL, M. A., & KOLPACK, J. (2008). "Document classification techniques for automated technology readiness level analysis". *Journal of the Association for Information Science and Technology*, 59(4), 675–680.

CAÑEDO, ANDALIA, R. (2011). "Los buscadores en la recuperación de información en salud". ACIMED, 22(3), 219–236.

CAZAU, P. (2006). *Introducción a la investigación en ciencias sociales.* Lima. Editorial Universidad Ricardo Palma. Retrieved from http://www.academia.edu/download/37844523/cazau_-_metodologia.pdf

CERRITO, P., & CERRITO, J. C. (2006). "Data and text mining the electronic medical record to improve care and to lower costs". *In Proceedings of SUGI* (Vol. 31, pp. 26–29). Retrieved from https://pdfs.semanticscholar.org/a4e0/0a006becd0df35163c1d8a4b612dcc7cea07.pdf

CHAWLA, N. V., & DAVIS, D. A. (2013). "Bringing big data to personalized healthcare: a patient-centered framework". *Journal of General Internal Medicine*, 28(3), 660–665.

CORSO, C. L. (2009). *Aplicación de algoritmos de clasificación supervisada usando Weka. Córdoba:* Universidad Tecnológica Nacional, Facultad Regional Córdoba. Retrieved from http://www.investigacion.frc.utn.edu.ar/labsis/Publicaciones/congresos_labsis/cynthia/CNIT_2009_Aplicacion_Algoritmos_Weka.pdf

DIEBOLD, F. X. (2012). *On the Origin (s) and Development of the Term'Big Data'.* Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2152421

GEORGIOU, A. (2002*).* "Data, information and knowledge: the health informatics model and its role in evidence-based medicine". *Journal of Evaluation in Clinical Practice*, 8(2), 127–130.

HILTBRAND, T. (2013). "Behavior-Based Budget Management Using Predictive Analytics". *The Business Intelligence Journal*, 18(INL/JOU-12-26713). Retrieved from http://www.osti.gov/scitech/biblio/1072389

JIANG, F., & LEUNG, C. (2015). *A Data Analytic Algorithm for Managing, Querying, and Processing Uncertain Big Data in Cloud Environments.* Algorithms, 8(4), 1175–1194. https://doi.org/10.3390/a8041175

KUDYBA, S. P. (2013). *Big Data, Mining, and Analytics: Components of Strategic Decision Making* - Books24x7. Retrieved May 10, 2017, from http://ezproxy.upaep.mx:2070/toc.aspx?bookid=61774

LAM, C., LAI, F.-C., WANG, C.-H., LAI, M.-H., HSU, N., & CHUNG, M.-H. (2016). *Text Mining of Journal Articles for Sleep Disorder Terminologies.* PloS One, 11(5), e0156031.

LAROSE, DANIEL T. & LAROSE, CHANTAL D.. (2015). *Data Mining and Predictive Analytics*, Second Edition - Books24x7. Retrieved May 10, 2017, from http://ezproxy.upaep.mx:2070/toc.aspx?bookid=63513

LEMKE, F., & MUELLER, J.-A. (2003). "Medical data analysis using self-organizing data mining technologies". *Systems Analysis Modelling Simulation*, 43(10), 1399–1408.

MELLIS, C. (2015). "Evidence-based medicine: What has happened in the past 50 years?*".* *Journal of Paediatrics and Child Health*, 51(1), 65–68.

MONTORI, V. M., & GUYATT, G. H. (2008). "Progress in evidence-based medicine". Jama, 300(15), 1814–1816.

PRATI, R. C., BATISTA, G., & MONARD, M. C. (2008). "Curvas ROC para avaliação de classificadores". *Revista IEEE América Latina*, 6(2), 215–222.

RAJA, U., MITCHELL, T., DAY, T., & HARDIN, J. M. (2008). "Text mining in healthcare. Applications and opportunities". *J Healthc Inf Manag*, 22(3), 52–6.

ROJAS, E., MUNOZ-GAMA, J., SEPÚLVEDA, M., & CAPURRO, D. (2016). "Process mining in healthcare: A literature review". *Journal of Biomedical Informatics*, 61, 224–236.

SACKETT, D. L., ROSENBERG, W. M., GRAY, J. M., HAYNES, R. B., & RICHARDSON, W. S. (1996). "Evidence based medicine: what it is and what it isn't". *British Medical Journal Publishing Group*. Retrieved from http://www.bmj.com/content/312/7023/71.short

SANTISO, S., CASILLAS, A., PÉREZ, A., ORONOZ, M., & GOJENOLA, K. (2016). "Document-level adverse drug reaction event extraction on electronic health records in Spanish". *Procesamiento del Lenguaje Natural*, 56, 49–56.

SIMPAO, A. F., AHUMADA, L. M., & REHMAN, M. A. (2015). "Big data and visual analytics in anaesthesia and health care". *British Journal of Anaesthesia*, aeu552.

SPINK, A., YANG, Y., JANSEN, J., NYKANEN, P., LORENCE, D. P., OZMUTLU, S., & OZMUTLU, H. C. (2004). "A study of medical and health queries to web search engines*". Health Information & Libraries Journal*, 21(1), 44–51.

URAMOTO, N., MATSUZAWA, H., NAGANO, T., MURAKAMI, A., TAKEUCHI, H., & TAKEDA, K. (2004). "A text-mining system for knowledge discovery from biomedical documents". *IBM Systems Journal*, 43(3), 516–533.

ZHANG, Y., BROUSSARD, R., KE, W., & GONG, X. (2014). "Evaluation of a scatter/gather interface for supporting distinct health information search tasks". *Journal of the Association for Information Science and Technology*, 65(5), 1028–1041.30.