Aceleración en la Recuperación de Información utilizando Algoritmos de Minería de Datos de R.

Osvaldo Spositto, Hugo Ryckeboer, Mauro J. Casuscelli,
Lorena Matteo, Julio Bossero.

Departamento de Ingeniería e Investigaciones Tecnológicas.
Universidad Nacional de La Matanza, Prov. Buenos Aires, Argentina
Florencio Varela 1902, San Justo, Prov. Buenos Aires, Argentina
spositto@unlam.edu.ar, hugor@unlam.edu,ar, mcasuscelli@alumno.unlam.edu.ar,
lmatteo@unlam.edu.ar, jbossero@unlam.edu.ar.

Resumen. Para acelerar la respuesta inicial en sistemas de recuperación de información sobre depósitos documentales privados de mediano tamaño se estudia la posibilidad de segmentar el mismo y elaborar la respuesta examinando sólo un segmento. Se analiza la pérdida de calidad que ello provoca. Las herramientas para fraccionar y elegir segmento provienen de la algoritmia de la minería de datos y se eligió el lenguaje R por tener ya incorporado los algoritmos básicos y ser un lenguaje que completo permite escribir el código que los vincula.

Palabras clave: Recuperación de Información, K-Means, Redes Neuronales, LSI, Distancia Euclidea.

1 Introducción

La recuperación de información es una técnica de la que se disfruta cuando se realizan búsquedas en Internet y es pretensioso intentar introducir mejoras a los buscadores más famosos. No obstante eso, existen depósitos documentales (corpus) privados que no se desea exponer al público y sobre los cuales se tiene interés en tener un sistema de recuperación. En la medida que tales depósitos aumentan de tamaño el tiempo de respuesta sube ya que es proporcional a la cantidad de documentos.

En todo sistema que interactúa con el hombre, hay que tener en cuenta la psicología de éste, el cual quiere respuestas con sensación de instantáneas y contra esta característica conspira el crecimiento del corpus. Una solución es aumentar la potencia de cómputo, pero esto no está al alcance de todos.

En este trabajo se analiza la posibilidad de fraccionar el corpus de modo tal de reducir el tiempo sin gran desmedro en la calidad de la primera respuesta que entrega el sistema frente a un requerimiento. Se propone que los segmentos contengan documentos afines de modo tal que muchas consultas queden resueltas por examen de un solo segmento aunque esa respuesta adolezca de algunos documentos.

También la propuesta tiene en cuenta que las consultas la realiza una persona y que habiendo varios documentos válidos en la respuesta inicial el usuario estará ocupado dando tiempo al sistema de perfeccionarla para cuando solicite las siguientes páginas.

Tratándose de poblaciones grandes, tanto los documentos como las consultas, ambos imposibles de describir con un patrón regular, evaluar esta propuesta será inevitablemente de un modo estadístico. Las herramientas para realizar esta tarea fueron sacadas de la minería de datos (MD), la cual justamente ha crecido para elaborar conclusiones sobre universos irregulares.

La tarea aquí comenzada se presta para futuras investigaciones, y además la técnica expuesta en un modo abstracto es aplicable a otras situaciones de apareo de objetos.

En la sección dos se describe someramente los principios de las tecnologías involucradas, limitado a lo efectivamente utilizado. En la sección tres se describe las tareas afines de los investigadores y algunas pocas ideas sobre lo realizado. En la cuartea, se detalla la idea y su concreción en código. Finalmente en la última sección se detallan las experiencias numéricas y el modo de juzgarlas.

2 Marco Teórico

El presente trabajo intenta poner la minería de datos al servicio de las necesidades de la recuperación de la información. La explicación de los algoritmos se reduce a lo necesario para facilitar la comprensión de la sección 4.

2.1 Recuperación de Información

Por recuperación de información (RI) se conoce una disciplina que ayuda a ubicar dentro de un repositorio de documentos los que mejor puedan resolver las necesidades intelectual del usuario del sistema de recuperación de información (SRI) [1].

Aunque el nombre de la disciplina quedó establecido así sería más adecuado verlo como un sistema que provee una lista ordenada de sugerencias de lectura [9], esperando que el usuario por examen de los mismos encuentre el material, no necesariamente sólo información, buscado. A diferencia de otros técnicas no utiliza, o al menos no primariamente, los metadatos del documento [2].

Como ya se mencionó, el conjunto de documentos sobre la cual se hará la selección se denomina el corpus y los sistemas que brindan el servicio, los buscadores. La idea que guía esta actividad es que las consultas y los documentos esperados comparten un mismo vocabulario. Afinando esta idea y teniendo en cuenta las inflexiones que sufren las palabras por necesidades gramaticales se ha pasado rápidamente a que más que palabras concretas conviene atenerse a los lexemas.

Si cada lexema se toma como una dimensión del espacio del habla, habrá lexemas con distinta frecuencia lo que hace que cada documento queda representado por un vector en este gigantesco espacio y allí también se representa la consulta.

Con una conveniente medida de distancia se puede lograr un ordenamiento de los documentos del corpus en función de la pregunta formulada y esto constituye la respuesta teórica al requerimiento formulado.

En la práctica no se entrega de una vez la lista total, imposible de manejar

intelectualmente, sino trozos, por ejemplo 10 documentos por vez, comenzando por los más promisorios. Después de examinar sus títulos y abrir los más promisorios bajo la óptica del investigador pasa a otra hoja si no encontró algo apropiado o por el contrario reformula su consulta.

Distintos sistemas difieren por el modo detallado con el cual construyen los vectores representativos, se los llama modelos. Los coeficientes son todos positivos y lo más sencillo es hacerlo booleano, unos si el lexema aparece, ceros si no.

Mejor que ello contar las ocurrencias, a las cuales se les aplican diversas correcciones. Esto se conoce como el modelo vectorial [1]. Los vectores son ralos ya que según la temática tratada aparecen unos lexemas y otros no.

Finalmente recurriendo a una descomposición en valores singulares (DVS) se logra evitar los errores que introducen la polisemia y la sinonimia. Los vectores resultantes son densos pero la experiencia indica que pueden reducirse sus dimensiones a unos pocos centenares. Se lo conoce como el modelo LSI (Latent semantic indexing).

Las consultas se deben volcar en el mismo modelo que se haya aplicado al corpus, como si fueran minúsculos documentos y enfrentar su vector representativo con los vectores de cada documento para luego ordenarlos, o al menos obtener el trozo inicial de lo que sería un vector ordenado. El tiempo de proceso es proporcional a la cantidad de documentos del corpus. El objetivo de esta investigación supone la existencia de algunos de estos modelos y es indiferente respecto de la calidad de los mismos.

2.2 La Minería de Datos

La minería de datos es una disciplina ya bien establecida y desarrollada en numerosos libros [6]. El objetivo principal es extraer nuevos conocimientos a partir de datos. Existen diversos tipos de métodos para extraer el conocimiento, estos métodos se agrupan de acuerdo al tipo de tarea que realizan. Las principales tareas son: clasificación, regresión, agrupación y asociación [6]. Tal vez sea más adecuado decir que organizan a través de estas tareas la información disponible para facilitar el conocimiento de la situación por parte de los usuarios de estos sistemas, las decisiones que deben tomar y su eventual delegación en sistemas automatizados. A continuación se explican de manera resumida dos de ellas que son las que fueron usadas en esta investigación.

Agrupación. También conocida como segmentación (en inglés se la conoce como clustering), es una técnica que permite analizar y examinar datos que no se encuentran etiquetados, formando conjuntos de grupos a partir de su similitud [6]. Los que comparten un mismo grupo recibirán una misma etiqueta, distinta de la de otros grupos. Los objetos a clasificar poseen propiedades sobre las cuales se puede definir un criterio de similitud o distancia. Especificar tales criterios se simplifica si las características son numéricas, con valores en un conjunto conceptualmente continuo. Esta situación la tenemos en los modelos de representación de documentos, salvo en el booleano, prácticamente en desuso. Las etiquetas son arbitrarias, optando muchos sistemas en aplicar números naturales consecutivos, carentes de todo significado adicional. Para nuestra aplicación la función distancia deberá armonizar con aquella que usa el ordenador de documentos en la recuperación.

Clasificación. Este tipo de tarea predice la categoría a la que pertenece un objeto dado. Se debe basar sobre el conocimiento de las categorías de otros objetos ya clasificados. De algún modo intenta ubicarlos en el grupo que contiene otros cercanos a él. Los métodos de clasificación extraen características de los objetos que ya están ubicadas en categorías durante un pre-proceso para agilizar las posteriores clasificaciones.

2.3 Algoritmos Utilizados

A continuación se da explica los dos algoritmos elegidos para realizar este proyecto.

Algoritmo K-Means. (En español debiera llamarse K-Medias), presentado por MacQueen [13] en 1967, es uno de los algoritmos desarrollados para resolver el problema del agrupamiento. La idea del algoritmo es proporcionar una clasificación de información de acuerdo con los propios datos, basada en análisis y comparaciones entre sus valores numéricos. Así, el algoritmo proporcionará una clasificación automática sin la necesidad de supervisión humana, es decir, sin pre-clasificación existente. Debido a esta característica, se considera como un algoritmo del tipo No Supervisado [6].

Es un algoritmo iterativo, parte de *K*, valor propuesto por el usuario, puntos en el espacio multidimensional de las características, que llamaremos centroides. La forma de elegir los centroides iniciales varía según distintas implantaciones que tiene el método. De allí en más cada iteración realiza dos pasos:

a) Por cada objeto a particionar se calcula cual es el centroide más cercano y se lo etiqueta como perteneciente a él.
b) Actualización centroides: se actualiza la posición del centroide de cada etiqueta tomando como nuevo centroide la posición del promedio de los objetos pertenecientes a dicho grupo.

Es usual que esta actividad converja y los cambios que sufren los centroides y acorde a ello los cambios de etiqueta sean cada vez menores. La teoría del método demuestra que la suma de los cuadrados de las distancias de los objetos etiquetadas a sus respectos centroides disminuye en cada paso.

Visto como problema matemático la solución no es única y puede ser aconsejable reiniciarla con nueva elección de centroides iniciales.

El comportamiento del algoritmo está influenciado por:

- El número de centroides (*K*) elegidos.
- La elección de los centroides iniciales.
- El orden en que las muestras son presentadas, en el caso de inicialización autónoma.
- Las propiedades geométricas de los datos.

Redes neuronales artificiales. Las Redes Neuronales Artificiales (RNAs o ANNs, en inglés, Artificial Neuronal Networks), son modelos computacionales que surgieron como intento de conseguir formalizaciones matemáticas acerca de la estructura y el comportamiento del cerebro humano. Simulan un aprendizaje a través de la experiencia. Los algoritmos desarrollados alrededor de esa idea resultaron útiles para resolver muchas situaciones de las cuales se posee un conocimiento insuficiente para

plantear una solución rigurosa. Evaluados estadísticamente logran un gran porcentaje de aciertos.

Los elementos básicos de un sistema neuronal biológico son las neuronas, agrupadas en redes compuestas por millones de ellas y organizadas a través de una estructura de capas [6]. En un sistema neuronal artificial puede establecerse una estructura jerárquica similar, posiblemente más regular que las biológicas. Las neuronas de una capa reciben estímulos solamente de las neuronas de la capa previa, si la hubiera y si no del exterior. A su vez su salida es enviada con distinto grado de intensidad a las neuronas de la capa siguiente, si las hubiera, de forma tal que una RNA puede concebirse como una colección de procesadores elementales (neuronas artificiales), conectados entre sí o bien a entradas externas y con una salida que permite propagar la señal por múltiples caminos.

Modelo de McCulloch-Pitts. Propuesto en 1943 y de salida binaria, la cual calcula la suma ponderada de sus entradas producidas por otras unidades, y da como salida un uno (1) si aquella se encuentra por encima de un umbral, o un cero (0) si está por debajo.

La figura 1 ilustra esquemáticamente una neurona según el modelo de McCulloch-Pitts, En ella se supone *j* entradas que llegan atenuadas por un coeficiente w_{ij} a una *i*-ésima neurona. Su estímulo neto es la suma de tales entradas, la función escalón tiene un umbral, si la suma lo supera da 1 si no, da 0.

Los valores de los coeficientes de esta y demás neuronas se ajustan para que tenga la red el comportamiento deseado.

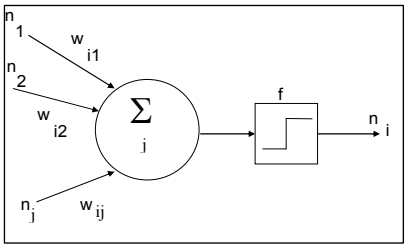

Figura 1. Modelo de neurona de McCulloch y Pitts

Este modelo ha evolucionado. Para facilitar el entrenamiento es bueno sustituir el escalón por una función analítica pero con alta pendiente en lo que sería el umbral. De esa forma enfrentando la red con casos de resultado conocido se modifican en sucesivas rondas los coeficientes, esperando que luego, frente a casos nuevos entregue el resultado correcto. Todas las salidas internas pasan a ser números reales, pasando a entero la salida final si la aplicación lo exige. En la aplicación aquí desarrollada requerimos salida real sin redondeo.

3 Antecedentes y Trabajos Relacionados

Este trabajo continúa con la línea de investigación de los proyectos PROINCE[1] C151 y C177, cuya temática se orientó: primero al estudio del tema y posteriormente a la realización de un prototipo de Sistema de recuperación de la Información, aplicando la metodología conocida como *Indexación Semántica Latente* (ISL o LSI[2]). Éste utiliza un proceso numérico llamado *Descomposición en Valores Singulares* para identificar patrones en las relaciones entre los términos contenidos en una colección de textos no estructurados y solucionar el problema de la sinonimia[3] y a la polisemia[4]. [12]

Varios autores de este trabajo, participaron en otro proyecto relacionado con la Minería de Datos: Análisis Comparativo de Modelos de Clasificación de Minería de Datos (Data Mining). Su aplicación en la predicción de perfiles de alumnos en riesgo de deserción. Proyecto PROINCE C176, en los años. 2015-2016. En la actualidad, los mismos, están llevando a cabo, desde el año pasado, una investigación cuyo título es: Uso de Minería de Datos para acelerar la recuperación de documentos.

En trabajos anteriores [9], Clusterdoc, es un sistema de recuperación y recomendación de documentos que está dirigido a usuarios con necesidades de búsqueda de información, que a través de algoritmos de agrupamiento divide el conjunto de datos en pequeños grupos con características comunes, lo cual permite minimizar el espacio de búsqueda y proporcionar información adaptada a los intereses del usuario.

Respecto a las RNA en [10], se presenta un trabajo muy completo, que nos introduce en este algoritmo en cuanto a sus definiciones, principios y tipología, en una aplicación concreta en el campo de la recuperación de la información. Se concluye aquí, que existen varias aplicaciones que explotan las características de las RNA y las aplican en nuestro campo de estudio, estableciendo que se encuentran todavía limitaciones muy grandes. El principal problema, citan, "…consiste en el volumen de procesamiento de información necesario…"y que esto se debe en parte "…a que la mayoría de las aplicaciones aquí citadas simulan el funcionamiento de una red masivamente paralela mediante un ordenador secuencial con arquitectura Von Neumann. Estas simulaciones no explotan la principal característica de las redes, el procesamiento paralelo…". Este trabajo concluye diciendo que a pesar de estas limitaciones las técnicas basadas en RNA aplicadas a la recuperación de la información, constituyen un campo de investigación muy prometedor.

Por último tenemos el trabajo de Augusto Cortez Vásquez y otros [11], que aborda el problema de una aplicación usando el algoritmo supervisado: máquinas de soporte vectorial (MVS) en el área de recuperación de información. El objetivo de esta propuesta es crear un modelo que permita etiquetar un texto con una categoría predefinida dado un conjunto de documentos D y un conjunto de categorías C, se trata de encontrar una función que haga corresponder a un documento *d* tomado de D, una

[1] Programa de Incentivos a Docentes Investigadores SPU-ME
[2] Por sus siglas en inglés, Latent Semantic Indexing.
[3] La sinonimia es una relación semántica de identidad o semejanza de significados entre determinadas expresiones o palabras.
[4] Una palabra polisémica es aquella que tiene dos o más significados que se relacionan entre sí.

categoría determinada *c* en C. Algo similar a una parte de nuestro trabajo que etiqueta cada documento con un número de grupo o clúster. En este proyecto se utilizó el análisis lexicográfico para identificar los lexemas. Constructos[5] aportes de expresiones regulares. Se realizó una comparación de subcadenas, para determinar el grado de semejanza entre dos textos a mayor subcadena en común mayor es el grado de semejanza. Por último se diseñó de la función kernel que fue utilizada, la cual determinó la eficacia de la MVS construida.

4 Uso de minería de datos en la recuperación de documentos.

El objetivo de todo sistema de recuperación de documentos es sugerir una lista de documentos ordenados de acuerdo a la probabilidad estimada de ser adecuados al requerimiento formulado. Sólo el examen de los documentos por parte del requeridor confirma el mayor o menor acierto del sistema. A esto se agrega la impaciencia propia del modo acelerado en que se vive que quiere la respuesta como si fuera instantánea.

Tal como se señaló al describir la tecnología subyacente, conforme los corpus aumentan de tamaño ese tiempo aumenta por ser este proporcional al tamaño. Para reducir ese tiempo se puede recurrir a un aumento de potencia de cómputo los que encaran ese camino recurren especialmente al paralelismo y en especial a las placas de video.

En esta investigación se analiza una línea alternativa, fraccionar el corpus. Esto requiere dos algoritmos preparatorios:

a) uno que particione el corpus utilizando una noción de vecindad o similitud y
b) el entrenamiento de un algoritmo de clasificación que direccione la consulta hacia la parte más promisoria.

Ambos servicios los estudia y provee la minería de datos.

Luego por cada consulta se debe ejecutar dos pasos:

c) aplicar el algoritmo que direcciona la consulta hacia una de las partes, para
d) enfrentar la consulta con cada documento de esa parte para determinar su grado de adecuación y posterior posición en la lista de documentos sugeridos.

Si la parte elegida efectivamente contiene un alto porcentaje de los documentos que hubieran encabezado la lista de haber hecho el proceso sobre la totalidad el usuario no sentiría demasiado la baja de la exhaustividad. Evidentemente habrá consultas cuya respuesta completa esté repartida entre varias partes, pero mientras haya en la página inicial suficientes documentos representativos de la respuesta ideal para que el usuario los examine hay tiempo de procesar otras partes y mostrarle cuando solicite la segunda página lo que hubiera faltado en la primera.

Se puede destacar algunas armonías que debe haber entre los cuatro procesos aquí señalados: Los procesos (b) y (c) se deben realizar sobre elementos ubicados en un mismo espacio conceptual. El entrenamiento del algoritmo de clasificación trabaja sobre representaciones vectoriales de los documentos. El clasificador debe recibir la consulta expresada en el mismo espacio de representación lo que aconseja someterlo a las mismas transformaciones que sufren los documentos. Por otra parte es necesario que el proceso (a) use la misma fórmula de distancia que (d). Así, si la consulta está

5 Un constructo es una construcción teórica que se desarrolla para resolver un cierto problema científico

cerca de un elemento de una partición, estará en términos comparativos cerca de todos. Se puede destacar que los procesos relacionados con clasificación podrían no utilizar la totalidad de los vectores que describen a los documentos buscando un compromiso entre velocidad y precisión.

Una manera de introducir uniformidad en el espacio de los documentos es escalar sus vectores descriptivos para tener módulo unitario. En esas condiciones las dos medidas intuitivas de distancia, una basada en el coseno del ángulo entre las direcciones y la euclídea son equivalentes desde el punto de vista práctico, como surge de la siguiente deducción aplicada a dos versores a y *b* de dimensión *d*

Partiendo del cuadrado de la distancia euclídea se llega al doble de la distancia medida a partir del coseno del ángulo entre dos versores:

$$\sum_{k=1}^{d}(a_k - b_k)^2 \;=\; \sum_{k=1}^{d}(a_k^2 - 2a_k b_k + b_k^2) \;=\; 2 - 2\sum_{k=1}^{d} a_k b_k \;=\; 2\,(1 - \sum_{k=1}^{d} a_k b_k)$$

la suma de los cuadrados en el primer paso intermedio da 1 por ser versores y cuando se quiere usar el coseno del ángulo como distancia debe ser complementado a 1 para que direcciones paralelas tengan distancia nula.

5.1 Evaluación de los resultados obtenidos

Métricas para evaluar . Es necesario introducir alguna métrica que permita apreciar la calidad del resultado obtenido y para ello hay varias propuestas. Conviene tener presente que la lista de documentos exhibidos al procesar una partición es una sublista de la lista que provee el corpus completo manteniendo el orden de ésta. Pues un documento precede a otro por su mayor afinidad sin tener en cuenta cuales son los elementos que lo acompañan.

Una primera, pensando en la aplicación a búsqueda documental analiza el escenario de uso del sistema, el usuario ingresa la consulta y obtiene una página con 10 documentos sugeridos. Mientras los examina el sistema puede seguir procesando una o más particiones adicionales y enriquecer una segunda página. Una primera estadística elegida es: “De haber procesado el corpus completo cuantos de los documentos hubieran provenido de la partición consultada, número entero que estará entre 0 y10”. Planteado en sentido inverso podría ser el décimo elemento que muestra cuán lejos estaría en la lista resultado si se hubiera procesado el corpus completo.

Experimentos realizados Se hicieron 4 experimentos con vectores de características de 5 elementos generados al azar en el rango [0, 1, 0). Los pseudo-corpora tienen 120 vectores, que fueron particionados en 4 partes y sobre ellos se procesaron 100 consultas con vectores de las mismas características. Tanto en los corpora como en las consultas los vectores se normalizaron a módulo 1.0.

Después se hizo un experimento adicional con un pseudo-corpus con 1200 vectores de largo 7 y 700 consultas.

Primera evaluación. Se la describe en forma tabular:

Tabla 1: Primer criterio de evaluación aplicado a 4 experimentos con vectores de largo 5

Exp.	0	1	2	3	4	5	6	7	8	9	10	≥ 7

0	0	1	1	4	10	11	12	15	12	18	16	61
1	0	0	0	5	10	12	10	15	18	11	19	63
2	0	0	4	4	7	12	18	18	14	10	13	55
3	0	3	1	2	6	7	14	12	18	17	20	67

Se observa que en la primera tabla que alrededor del 60% de los casos se tienen 7 o más de los documentos que hubieran entrado en la primera página de haber procesado el corpus completo.

La experiencia se repitió sobre el pseudo-corpus de 1.200 vectores de 7 elementos haciendo 700 consultas sobre el mismo, los resultados fueron porcentualmente similares:

Tabla 2: Primer criterio de evaluación aplicado a un experimento con vectores de largo 7

Exp.	0	1	2	3	4	5	6	7	8	9	10	≥ 7
4	0	3	6	31	49	75	67	89	80	123	177	469

Segunda evaluación. Se la describe en forma tabular pero incompleta ya que el décimo elemento exhibido al computar sólo la partición elegida puede estar muy alejado en una evaluación total, los valores posibles comienzan en 10

Tabla 3: Segundo criterio de evaluación aplicado a un experimento con vectores de largo 7

Exp.	10	11	12	13	14	15	16	17	18	19	20	21	22	≤ 14
0	20	13	8	8	6	6	5	4	3	7	3	3	2	55
1	19	4	9	8	6	8	4	4	4	2	1	3	7	46
2	13	6	6	10	9	5	2	8	4	2	6	4	1	44
3	20	13	8	8	6	6	5	4	3	7	3	3	2	55
4	177	89	67	55	45	32	32	24	24	22	16	25	11	433

Se observa que aproximadamente la mitad tiene su décimo elemento no más atrás de la posición 14 y en el experimento grande supera el 60%

6 Conclusiones y futuros trabajos

Los números obtenidos en las simulaciones son promisorios lo que incentiva seguir investigando para obtener porcentajes aún mejores. Al contemplar las tablas de resultados hay que tener presente que las recuperaciones se habrían obtenido en el 25% del tiempo de proceso que hubiera insumido un proceso del corpus completo.

Se destacan algunas ideas que debieran contribuir a obtener una mejora evaluación:

- En lugar de particionar recubrir el corpus con K partes, lo que resolvería el problema de documentos cercanos a la frontera de dos o más partes
- Particionar en más partes y fusionar dos o más entre los más promisorios.
- Probar con otros algoritmos de particionado prefiriendo aquellos que logren partes más equilibradas.
- Superponer dos particionados de distinta semilla y unir las partes que uno u otro hubieran recomendado.

Agradecimientos. A Cecilia Gargano por su contribución en algunos cálculos

iniciales.

Bibliografía

[1] Salton, G.: Automatic Information Organization and Retrieval. McGraw−Hill, N.Y. (1968).

[2] Seco Naveiras, D.: Técnicas de indexación y recuperación de documentos utilizando referencias geográficas y textuales. Universidade da Coruña. Dep. de Computación. (2009), <http://ruc.udc.es/dspace/handle/2183/ 7172>. Citado el: 20/06/2018.

[3] Salton, G.; McGill, M.J.: Introduction to Modern Information Retrieval, New York: McGraw-Hill, (1983).

[4] Manning, C., & Schutze, H. Chapter 11.: Probabilistic Context Free Grammars. En: Foundations of Statistical Natural Language Processing. (1999).

[5] Zazo Rodríguez Á. F. y otros.: Diseño de un motor de recuperación de la información para uso experimental y educativo. Facultad de Documentación Universidad de Salamanca (2000). <http://bid.ub.edu/04figue2.htm>. Citado el: 20/06/2018.

[6] Hernández Orallo, J., Ramírez Quintana, M.J. Ferri Ramírez, C.: Introducción a la Minería de Datos. Pearson, ISBN: 84 205 4091 9. (2005).

[7] Bedregal Lizárraga, C.: Agrupamiento de Datos utilizando técnicas MAM-SOM. Universidad Católica San Pablo. (2008). :
http://personales.dcc.uchile.cl/~cbedrega/publications/Tesis.pdf

[8] Vallejo Huanga, D.: Clustering de documentos con restricciones de tamaño. Universitario en Gestión de la Información. (2015). https://riunet.upv.es/bitstream/handle/10251/69089/ Vallejo-Clustering de Documentos con RestriccionesdeTamaño.pdf?sequence=23

[9] Giugn, M.: Clusterdoc, un sistema de recuperación y recomendación de documentos basado en algoritmos de agrupamiento. Telematique, vol 9 - nro 2. (2010). http://www.redalyc.org/pdf/ 784/78415900002.pdf

[10] de Moya Anegón, F.: La aplicación de Redes Neuronales Artificiales (RNA): a la recuperación de la información. Revistes Catalanes Obert. (1998). http://www.raco.cat/index.php/Bibliodoc/article/view/56630

[11] Cortez Vasquez, A.: Categorización de Textos mediante Máquinas de Soporte Vectorial. Revista de Investigación de sistemas e Informática. Universidad Nacional Mayor de San Marcos Facultad de Ingenieria de Sistemas e Informática. (2003). ISSN 1816-3823. http://revistasinvestigacion.unmsm.edu.pe/index.php/sistem/article/viewFile/5711/4942

[12] Venegas, R.: Análisis Semántico Latente: una panorámica de su desarrollo. Pontificia Universidad Católica de Valparaíso. Chile. Revista signos [online]. (2003), vol.36, n.53, ISSN 0718-0934. http://dx.doi.org/10.4067/S0718-09342003005300008.

[13] MacQueen, J. B.: Some Methods for classification and Analysis of Multivariate Observations. En: Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability", Berkeley, University of California Press, 1:281-297 (1967)"