Unidad Ejecutora:

Departamento de Ingeniería e Investigaciones Tecnológicas

Código del proyecto:

C205

Título del proyecto de investigación:

Uso de Minería de Datos para acelerar la recuperación de documentos

Programa de acreditación:

PROINCE (Programa de Incentivos a Docentes Investigadores SPU-ME)

Director del proyecto:

RYCKEBOER, Hugo Emilio

Codirector del proyecto:

BLANCO, Gabriel

Integrantes del equipo:

SPOSITTO, Osvaldo Mario

PROCOPIO, Gastón Emanuel

PRILUSKY, Elisa Mirta

MATTEO, Lorena Romina

MACIAS CORRAL, Patricio Ezequiel

GARGANO, Cecilia Victoria

CASUSCELLI, Mauro Javier

BOSSERO, Julio César

Fecha de inicio:

01/01/2017

Fecha de finalización:

31/12/2018

Informe final

Sumario:

1. Resumen y palabras clave

El objetivo de este proyecto se basó en construir una alternativa innovadora, basada en algoritmos de clasificación, para realizar la búsqueda de documentos relevantes en un tiempo menor de respuesta.

Siguiendo los procesos de un sistema de recuperación de información (SRI), los documentos de un Corpus son transformados en *vectores descriptivos*. Una consulta de usuario es también convertida en otro vector descriptivo. Para obtener un documento que satisfaga la necesidad de información del usuario, el vector de la consulta se debe enfrentar con todo el corpus, en búsqueda de similitudes. Este proceso genera un índice de relevancia, que ordenará la lista de documentos sugeridos que recibe el usuario.

En este trabajo se analiza la posibilidad de fraccionar un corpus de modo tal de reducir la cantidad de documentos a comparar. Para ello, se requiere de dos procesos 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* (MD).

Luego por cada consulta se deben ejecutar dos pasos:

a) Aplicar el algoritmo que direcciona la consulta hacia una de las partes, para
b) Enfrentar la consulta con cada documento de esa parte para determinar su grado de adecuación y posterior posición en la lista de documentos sugeridos.

Los números obtenidos en las simulaciones del primer año fueron promisorios, lo que incentiva seguir investigando para obtener indicadores aún mejores. La cantidad de ideas que fueron generadas es de no acabar. Destacando algunas ideas que deberían contribuir a lograrlo:

- ✓ Decidir por cada consulta la conveniencia de explorar o no los documentos de la franja marginal de los particionados.
- ✓ Recurrir a varios particionados para reducir el problema de frontera

Palabras clave: Particionado. Corpus. Simulación. Índices de eficiencia. Redes neuronales. Centroides.

2. Memoria descriptiva

Introducción

La recuperación de información (RI) [1], es una técnica de la que se disfruta cuando realizan búsquedas en Internet y es pretencioso 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 y no integren un espacio multidimensional estructurado el tiempo de respuesta aumenta ya que es proporcional a la cantidad de documentos.

En este trabajo se analiza la posibilidad de fraccionar un 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 las fracciones 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 tuvo en cuenta que las consultas las 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, la cual justamente ha crecido para elaborar conclusiones sobre universos irregulares.

El desarrollo de un Sistema de Recuperación de Información comprende básicamente tres fases distintas [2]: el Pre-procesamiento, la Modelización y la Utilización. El pre-procesamiento y la modelización, que conlleva las acciones necesarias para transformar los documentos de la colección en una estructura de datos con la información relevante de los documentos ha sido una parte importante de este proyecto y fue explicada en el informe del año anterior.

Se ha centrado en esta fase, la utilización tratando de acortar el tiempo de respuesta. Los resultados fueron expuestas en dos congresos el año 2018: Por un lado en el XXIV Congreso Argentino de Ciencias de la Computación (CACIC 2018), se presentó el escrito baja título: "*Aceleración en la Recuperación de Información utilizando Algoritmos de Minería de Datos de R*" y posteriormente se perfeccionó el método fruto del esfuerzo del segundo año dando lugar al trabajo presentado en el Quinto Congreso Internacional de Educadores en Ciencias Empíricas en Facultades de Ingeniería (ECEFI 2018), titulado "*Recuperación de Información acelerada con Algoritmos de Minería de Datos*". Trabajo que fue seleccionado, junto con otros, para los anales del congreso.

Dada la originalidad de esta línea de trabajo fue presentado como: "*Recuperación de la información"* en el XX Workshop de Investigadores en Ciencias de la Computación (WICC 2018, Universidad Nacional del Nordeste) exponiendo las ideas directrices.

Como este proyecto está inscripto dentro de una línea de trabajo que lleva varios años dedicado a la recuperación de información, restringida al idioma español, se ha encontrado la carencia de grandes Corpora en lengua española de dominio público, con lo cual se encaró además de las tareas específicas, la construcción de un corpus a partir de periódicos digitales, lo que fue presentado como *"Método para la construcción de un Corpus periodístico mediante Expresiones Regulares*" en el 6to. Congreso Nacional de Ingeniería Informática – Sistemas de Información (CONAIISI 2018). En la ciudad de Mar del Plata.

En el informe del primer año se explicó el motivo que llevó a alterar ligeramente el plan de trabajo, por lo cual en este informe se detallan las siguientes tareas:

a) Análisis de métodos de evaluación.
b) Optimización de los algoritmos.
c) Análisis de Resultados.

Es necesario hacer una advertencia inicial pues afecta al desarrollo del presente informe. La idea central con la que se planteó este proyecto, bastante simple por otra parte, "*Particionemos el corpus de documentos y examinemos sólo una parte para tener una respuesta menos precisa pero mucho más rápida*." Lo que no se podía prever sin las debidas simulaciones era la calidad de lo que se obtendría.

El desarrollo de la idea pensaba seguir el esquema típico propuesto por los manuales de Minería de Datos: Particionar con un método No Supervisado y predecir con uno Supervisado, gozando de especial prestigio para esta tarea el uso de Redes Neuronales.

Sobre la marcha surgió la siguiente reflexión: Si el objeto de la predicción es determinar la partición en la cual encontrará la mayor cantidad de elementos afines, o sea, cercanos, es de suponer que la predicción lo llevará a aquella partición que tenga su centroide más cercano, o sea aquella partición a la cual hubiera quedado arrastrada si en lugar de predecir hubiera participado de la acción de particionar.

La conveniencia de utilizar redes o distancias además de su efectividad debe ser juzgada a la luz del costo computacional, pues encaminarse a una partición debe ser ágil para no perder en el proceso la ganancia de tiempo que se espera conseguir.

Si hay ***A*** atributos y se particiona en ***K*** partes con una única red que determine una de las ***K*** pertenencias, detalle a cuestionar más adelante, habrá típicamente:

$$\boldsymbol{H = \sqrt{AK}} \qquad \textbf{(1)}$$

nodos en una capa oculta. Una primera matriz de ***A*** × ***H*** y una segunda de ***H*** × ***K*** lo que lleva a (***A+K***) × ***H*** operaciones lineales más el cálculo de ***H+K*** funciones no lineales propias de la red.

Mirando la alternativa, determinar ***K*** distancias insume del orden de ***K*** × ***A*** operaciones de producto. Si ***A*** fuera 25 y ***K*** 4, ***H*** vale 10 entonces la red hace 290 operaciones lineales, además de las 14 no lineales, y las distancias sólo 80.

En nuestro planteo, el cálculo de distancias era particularmente ágil, por cuanto los documentos y las consultas son normalizados para tener módulo unitario, en estas circunstancias la distancia angular está relacionada con la euclídea.

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

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

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.

Usar distancias a centroides insume menos esfuerzo que predecir con redes. Esta observación no ha sido localizada en libros y en las experiencias numéricas se aplican ambas técnicas y se las compara para convalidar la posibilidad de renunciar a las redes, cosa que efectivamente se ha concluido para el futuro. Además, entrenar redes es un proceso costoso computacionalmente. En cambio, el uso de distancias sólo exige conocer los centroides, esfuerzo lineal en el tamaño del corpus obtenible sin esfuerzo adicional como subproducto del particionado en métodos como K-means.

Sin embargo, se ha efectuado una segunda observación respecto del uso de distancias a centroides: el método K-means toma en cada etapa como centroide, la coordenada que minimiza la suma de los cuadrados de las distancias, lo cual suele coincidir con el centro de gravedad de la partición. Pero si el centroide pretende ser una hipotética muestra más y si todas las muestras cumplen una propiedad el centroide debería cumplirla también.

Usando valores normalizados, estos se encuentran sobre una hiperesfera y el centro de masa estaría en el interior de la misma. Se resolvió por medio de Multiplicadores de Lagrange [3 y 4] la ecuación que determina un elemento de la hiperesfera que minimice la suma de los cuadrados

de las distancias. Lo que introduce un ajuste de los centroides vistos como centros de masa, desigual para cada centroide. Esto lleva a que la determinación de parte basada en menor distancia a centroide de resultado distinto ocasionalmente según de que centroide se hable. La experimentación numérica privilegia ligeramente al uso de los centroides corregidos. La acción de corrección insume $K \times A$ operaciones adicionales lo que no desequilibra su superioridad frente a las redes.

Se podrá observar, en los cuadros que ilustran las experiencias numéricas, entonces una triple manera de resolver el problema de elegir parte.

En un párrafo precedente se adelantó que las decisiones con redes se han tomado sobre la base de una red por partición. De haber delegado en la red la elección de la misma, ella hubiera hecho los redondeos convenientes para proponer finalmente una clase. Con redes individuales que deben decidir con 0 o 1 pero expresado como número real su pertenencia a la clase, se logra tomar la decisión final mirando que red provee el valor más alto.

Hay numerosos casos en los cuales con K valores inferiores a 0,5, el mayor de ellos conduce a la elección acertada de la partición. El conocimiento detallado de la calidad de la decisión ha sido utilizado con provecho en las mejoras introducidas durante este segundo año del proyecto.

a. Análisis de métodos de evaluación

Para poder juzgar la calidad de los resultados obtenidos es necesario disponer de una medida, sobre todo si durante la experimentación se pretende modificar alguna parte del algoritmo. También para ver si la escala de la experimentación altera los resultados. Si variando los tamaños dentro de un rango de valores pequeños y medianos no se aprecia una modificación sensible de la calidad se puede suponer que ellos son extrapolables a tamaños mayores.

Al diseñar una métrica se debe tener presente el contexto dentro del cual se piensa aplicar lo experimentado y tratar de que ello quede absorbido dentro de la métrica. Por otra parte, es de interés que las métricas puedan dar lugar a una apreciación fina y al mismo tiempo den un único número resumen de la apreciación fina, pero apto para decisiones rápidas.

En algún momento se ha señalado que la Recuperación de Documentos no es más que una búsqueda en espacios multidimensionales con escasa estructura interna de apoyo y que por lo tanto las conclusiones de este trabajo son aplicables a otras búsquedas aproximadas. Se han señalado como ejemplos, la venta de vehículos usados, las páginas para relacionar personas con fines matrimoniales, búsqueda de hoteles, etc. Todos estos dominios se encuentran con que ni lo ofrecido, ni lo pedido se pudieron describir con lujo de detalles y que los detalles no estuvieran afectados por ecuaciones personales y elementos de distorsión.

En todos estos dominios se sabe que la elección final será tomada por el consultante después de interiorizarse de detalles que el sistema no supo o no pudo incorporar. La solución práctica adoptada universalmente es contestar por “páginas”, o sea, proveer en lugar de la mejor propuesta un conjunto de propuestas buenas. Dejando que por examen detallado el usuario del servicio de consulta las filtre para finalmente, si no se considera satisfecha recurra a que se le provea un lote o “página” complementaria de la anterior.

El tamaño de la página puede cambiar de un sistema a otro, pero se ha visto que, una vez superado un umbral, no tiene mayor incidencia. Para evitar multiplicidad de tabulaciones, ha sido fijado en 10 sugerencias por página. Examinar 10 sugerencias lleva un tiempo muy grande frente a las velocidades computacionales, de modo tal que, la segunda página podría remediar las deficiencias la primera, deficiencias ocasionadas por haber sido elaborada sobre una partición y no sobre la totalidad del Corpus.

Estos sistemas de recuperación, nunca dan resultados perfectos, ya que en su planteo parten de varias hipótesis estadísticas, que como tales solo acierten cierto grado de verosimilitud. La más importante de las limitaciones, es la hipótesis de que los parámetros recogidos de las entidades ofrecidas: documentos, automóviles, hoteles, etc. describen a éstos a la perfección.

Los usuarios de tales sistemas saben que la respuesta no es ideal y que algunas propuestas recibidas no serán del todo adecuadas al requerimiento que formulan, sin poder discernir si esto se debe a la inevitable imperfección o a la búsqueda restringida a un subconjunto del material disponible.

Hay una característica importante a destacar en la respuesta del sistema. Lo que se obtiene al haber explorado un subconjunto es una sub-lista de la lista que se obtendría al explorar el total. El ordenamiento de los documentos en la respuesta obedece al mismo criterio en ambos casos. Un documento precede a otro en la medida que sea matemáticamente más cercano su descriptor al descriptor de la consulta sin tener en cuenta cuales son los elementos que lo acompañan.

Esta reflexión propone una primera medida: ¿Cuántos documentos de los que hubieran aparecido en la primera página con una exploración total, aparecen en la página armada con la exploración de una parte? La respuesta que es un valor entre 0 y 10.

El siguiente croquis ilustra esta explicación. Por razones de tamaño en él, se ha supuesto páginas de 5 documentos.

El primer vector (tabla 1) ilustra los hipotéticos documentos que hubieran respondido a una consulta convenientemente ordenados por su afinidad con la consulta.

214	307	184	107	223	194	732	156	218	193	318	472	618	286	145	…
A	B	B	C	A	B	B	B	A	C	D	B	D	A	C	
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	…
Página 1					Página 2					Página 3					

Tabla 1. Resultado del supuesto experimento de consulta.

Una búsqueda que sólo examina una partición, por ejemplo: la B, mostraría los documentos ilustrados en el segundo vector.

307	184	194	732	156	472	…
B	B	B	B	B	B	B
1	2	3	4	5	6	..
P á g i n a 1					Página 2…	

Tabla 2. Resultado del experimento con particionado.

Según el primer criterio de evaluación propuesto le tocaría 2 (en una escala de 0 a 5)

El segundo criterio de evaluación que hemos propuesto se pregunta: ¿Cuán lejos queda en el orden total el último que se muestra en la página 1? En este ejemplo se trata del documento 156 y le corresponde la respuesta 8. Generalmente a mayor valor con el criterio 1, le corresponde menor con el criterio 2.

Los siguientes cuadros, tablas 3, 4 y 5, exhiben las frecuencias de las parejas ⟨medida1, medida2⟩ donde las filas corresponden a las medidas 1 y las columnas a la 2, truncado en el valor 30 (el equivalente a 3 páginas de resultados) ya que en los valores bajos de la medida 1, la medida 2 toma valores muy grandes. Fueron computados a partir de muestras del mismo tamaño para particionados en 3, 4 y 5 partes respectivamente.

	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
0											0	0	0	0	0	0	0	1	0	0	3
1										0	0	2	0	0	5	2	3	5	21	10	16
2									0	2	4	2	5	20	14	19	15	19	23	45	27
3								0	2	14	13	29	38	25	33	58	46	47	48	50	32
4							2	20	37	40	40	47	46	47	42	41	36	37	43	24	21
5						32	46	78	65	76	60	55	53	35	27	18	20	7	5	6	13
6					89	100	89	77	61	51	44	30	18	13	15	2	3	2	4	2	2
7				191	163	127	90	38	29	18	8	5	1	3	1	3	0	0	0	0	0
8			349	176	71	45	9	8	5	0	2										
9		618	122	15	2	0	2														
10	1720																				

Tabla 3. Resultado de las frecuencias de las parejas **⟨medida1, medida2⟩**, para particionados en 3 partes.

Las zonas sombreadas señalan parejas que son imposibles de darse. Cuando los valores finales son ceros fueron omitidos. Así se puede observar que con un valor 8 o 9 en la medida 1 sólo aparecieron unos pocos valores bajos de la medida 2.

	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
0											0	0	0	0	0	0	1	2	0	1	1
1										0	1	0	3	0	2	8	1	10	13	13	17
2									0	3	2	5	5	13	13	20	23	37	29	36	43
3								1	8	9	15	29	33	43	46	53	52	54	73	57	39
4							9	12	33	44	51	51	52	72	51	48	46	51	48	44	29
5						22	67	64	90	78	65	72	45	56	43	34	26	24	23	14	19
6					64	83	88	86	95	62	46	46	25	19	8	7	7	11	6	3	2
7				170	140	136	92	50	22	20	18	8	6	9	5	3	0	1	0	1	
8			282	178	87	56	20	4	3	3	0	1	2	0	0	0	0	0	1		
9		538	121	26	4	2	1														
10	1392																				

Tabla 4. Resultado de las frecuencias de las parejas **⟨medida1, medida2⟩**, para particionados en 4 partes.

	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
0											0	0	0	0	0	0	0	1	2	3	2
1										0	0	0	0	0	3	3	4	10	6	10	7
2									0	2	5	6	12	15	15	21	18	31	24	28	51
3								0	5	15	17	29	35	52	46	48	60	60	59	42	56
4							5	25	33	42	55	66	71	74	55	71	54	54	46	46	36
5						30	37	62	77	89	72	99	65	53	50	39	34	23	26	23	25
6					65	83	110	90	83	70	57	41	24	21	14	14	6	6	2	1	6
7				155	140	91	77	61	33	20	21	6	15	8	1	2	1	1	1	1	
8			268	156	93	50	13	12	5	7	0	3	1	0	0	0	1				
9		511	133	28	12	0	1	1													
10	1112																				

Tabla 5. Resultado de las frecuencias de las parejas **⟨medida1, medida2⟩**, para particionados en 5 partes.

Además de servir de ilustración de las medidas propuestas, se observa que conforme el número de partes aumenta los valores iniciales disminuyen y aumentan los finales en cada fila, indicio de una pérdida de calidad en la recuperación. Esto se explica porque una mayor cantidad de partes implica mayor probabilidad de que una consulta caiga cercana a una frontera, situación en la cual es inevitable un resultado deficiente. Esta observación será el punto de partida de las mejoras propuestas este año.

Estas medidas, por supuesto, no reflejan todos los detalles de la calidad del resultado, pero permiten una rápida apreciación. A continuación, tabla 6, se esbozan dos hipotéticas distribuciones con las mismas convenciones del ejemplo donde ambas evalúan en ⟨2,9⟩ siendo claramente mejor la segunda que la primera.

			B	**B**		**B**	**B**	**B**		**B**	
B	**B**				**B**	**B**		**B**		**B**	

Tabla 6. Dos distribuciones de igual valuación.

La tercera medida promedia el valor que se atribuye a cada elemento exhibido. Ese valor es el cociente entre la posición que ocupa en la Respuesta, dividido el lugar que ocuparía en una solución global.

Esto daría:

para la primera (1/4 + 2/5 + 3/7 + 4/8 + 5/9 + 6/11) / 6 = 0,44659692 y
para la segunda (1/1 + 2/2 + 3/6 + 4/7 + 5/9 + 6/11) / 6 = 0,69540645

Para hacer un análisis detallado interesa ver la distribución de las respuestas frente a diversas consultas. Para tener una medida abreviada se ha considerado que si el primer criterio supera o iguala a 7, el usuario no tendrá vivencia de una baja en la calidad. El porcentaje de consultas que cumplen con esta calidad, da una medida global de calidad.

Para el segundo, si no supera a 20, se lo considera satisfactorio. Esto equivale a decir que las 10 respuestas mostradas son un extracto de dos páginas de la respuesta plena.

De los experimentos realizados se descubre que la calidad disminuye con la cantidad de particiones y que es independiente de la cantidad de atributos utilizados para describir consultas y documentos. La tabla 7 reproduce 4 experimentos con 5 atributos y 100 consultas.

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

Tabla 7. Resultado del experimento con 5 atributos y 100 consultas.

La tabla 8 reproduce los experimentos con 7 atributos y 700 consultas utilizando 8 particionados distintos, pero todos en 4 partes.

Exp	0	1	2	3	4	5	6	7	8	9	10	≥7
0	0	6	7	31	52	69	68	89	80	121	177	66.71
1	1	0	16	28	45	52	79	90	100	92	197	68.43
2	1	2	12	28	35	65	77	90	108	100	182	68.57
3	1	3	7	30	51	72	77	96	74	108	181	65.57
4	1	4	9	26	50	76	54	87	75	97	221	68.57
5	0	2	17	35	43	64	81	86	90	86	196	65.43
6	1	2	9	28	34	64	80	93	80	104	205	68.86
7	1	4	8	26	56	67	58	81	87	102	210	68.57

Tabla 8. Resultado del experimento con 7 atributos y 700 consultas.

La tabla 9 reproduce los experimentos con 7 atributos y 700 consultas utilizando 6 particionados distintos, pero todos en 5 partes.

Exp	0	1	2	3	4	5	6	7	8	9	10	≥7
0	2	3	14	33	74	120	113	98	85	83	75	48,71
1	2	3	13	45	73	99	111	83	76	93	102	50,57
2	0	5	26	36	53	78	96	105	111	84	106	58,00
3	4	19	38	45	76	93	104	105	91	68	57	45,86
4	3	21	23	60	73	73	73	64	72	65	173	53,43
5	0	2	29	55	77	89	120	110	80	74	64	46,86

Tabla 9. Resultado del experimento con 7 atributos y 700 consultas.

El promedio general del 50%, confirma lo que era de esperar, que aumentar la cantidad de particiones no es beneficioso y que el camino para mejorar debe ser otro.

b. Optimización de los algoritmos

Tal vez el título propuesto en la planificación hecha hace dos años no fuera el adecuado. En este segundo año se evaluó como mejorar la calidad, expresada por los diversos indicadores, si el 65% no fuera un valor satisfactorio.

En modalidad de estudio es factible recorrer todas las particiones y juzgar cual es la mejor, valor que se puede tomar como referencia para juzgar a los algoritmos, ya sea redes o distancias, que deciden cual es la que debe ser explorada.

Sin embargo, se decidió analizar modificaciones que pudieran mejorar el porcentaje de consultas con respuesta satisfactoria y el precio a ello asociado, ya sea en pre-procesos del corpus como en un menor beneficio en tiempo para la consulta.

Antes de proponer las mejoras se examinó en detalle los resultados obtenidos. Con K=3 y K=5 se elaboraron 4900 consultas y con K=4 5600, Tabla 11, decidiendo la partición a explorar por los tres métodos: mayor valor de red neuronal, menor distancia a centroide baricéntrico1 y menor valor a centroide sobre el casquete. Estos valores fueron tabulados, lo que se ilustra con un trozo de la tabulación en el Anexo I. A partir de ello se separaron aquellos casos en los cuales al menos uno de los tres no toma la decisión óptima.

	K=3	**K=4**	**K=5**
Casos estudiados	4900	5600	4900
Todos coinciden con el óptimo	4199	4697	4244
Se equivoca con la distancia al baricentro	444	648	814
Se equivoca con la distancia a centroide corregido	451	609	724
Se equivoca con las redes	529	664	931

Tabla 10. Resultado del experimento con K= 3,4, y 5.

Volviendo a la reflexión inicial, se ha notado que estos resultados confirman que exigir que el centroide utilizado para decidir, estuviera en el casquete hiper-esférico es conveniente y que decidir por distancias a centroides es superior a decidir por redes neuronales. Esta primera conclusión hizo privilegiar en experimentos posteriores las decisiones basadas sobre distancias a centroides ubicados sobre el casquete hiperesférico.

Tratando de diagnosticar adecuadamente el problema, se separaron los casos que decidieron bien de los que no, tomando el valor que fue utilizado para decidir. En el caso de redes, son

1 Baricentro. Del gr. βαρύς barýs 'pesado, grave' y centro. Según RAE: m. Fís. Centro de gravedad; m. Geom. Punto de intersección de las medianas de un triángulo.

máximos, en el caso de distancias, son mínimos y se promediaron esos valores tal como ilustra la siguiente Tabla.

	K = 3	K = 4	K = 5
Valor promedio de red que decidió bien	0,943959	0,877351	0,879773
Valor promedio de red que decidió mal	0,660186	0,558815	0,581797
Promedio de distancia que decidió bien	0,133913	0,092078	0,084931
Promedio de distancia que decidió mal	0,163708	0,111866	0,102501

Tabla 11. Resultado del experimento con K= 3,4, y 5.

De aquí se dedujo que un valor mejor definido en las Redes o en las Distancias, es señal de una buena definición y decisiones basadas en valores más bajos son propensos a decidir mal. Esto se tuvo en cuenta cuando se dispone de más de un particionado: Elegir a aquel que entregue el valor más alto si son redes o el más bajo si son distancias.

Dos son los motivos que llevan al 65% de resultados "satisfactorios":

- Los errores en las decisiones sugeridas por redes o distancias.
- Que sea imposible una buena decisión porque los documentos requeridos están en más de una partición.

Esta segunda causa influye además en la primera, Redes y Distancias definen con menor decisión la mejor partición.

Cuando los documentos buscados están repartidos en más de una partición se puede considerar que las consultas hayan caído cerca de la frontera de dos o más particiones, de modo tal que fuera inevitable que se viera sólo una parte de los vecinos.

El azar de los documentos puede hacer inclusive que del otro lado de la frontera haya más documentos que del lado consultado.

Dos formas de atacar esto, fueron diseñadas: (a) disponer de dos o más particionados, con la esperanza de que caer en frontera de partes en uno fuera interior para otro; (b) pensar en recubrimientos más que en particionados, de modo tal que cada partición incorpore elementos cercanos de sus vecinos.

a) Múltiples particionados.

Las implantaciones de K-means son sensibles al orden en que reciben los datos. De modo tal que cambiando convenientemente el orden de los vectores representativos de un corpus se puede conseguir diversas propuestas de particionado. Al proponer en que parte de que particionado se debe realizar la búsqueda se puede optar ya sea la que provea la señal de red más elevada o la que tenga más cercano su centroide. Trae como contrapartida un mayor tiempo de pre-proceso del todo el corpus antes de dejarlo operativo y una mayor fragmentación en el almacenamiento del corpus, el cual podría verse fraccionado en forma exponencial siguiendo la distribución de los elementos en los distintos particionados. U, organizar múltiples listas vinculadas sobre los vectores del corpus o listas invertidas ordenados por el lugar físico que ocupe el vector descriptivo, etc.

Experimentos realizados

Sobre un mismo corpus se organizaron 8 particionados y se han procesado 700 consultas, Tabla 13. Luego se extendieron las decisiones sobre el Corpus a utilizar incluyendo 2, 3 o más particiones, eligiendo el particionado y parte cuyo centroide estuviera más cerca. El tiempo de cómputo para elegir parte es proporcional a la cantidad de particionados

consultados. Con sólo elegir entre dos particionados el primer indicador subió de 65% a 80%, continuando el ascenso con más particionados siguiendo una típica ley de rendimientos marginales decrecientes. Desaparecen prácticamente resultados con menos de 3 aciertos en la primera página.

Usando demasiados particionados se estanca el beneficio por la correlación que puede haber entre los mismos. Habiendo hecho 8 particionados se podría experimentar entre las 28 parejas cual es la que da el mejor resultado.

particio	Distribuión detallada de la métrica 1											Métrica 1 abre-		Métrica 2 abre-		Métrica
nado	0	1	2	3	4	5	6	7	8	9	10	viada: 7 o más		viada: pag. 1y2		3 global
0	0	6	7	31	52	69	68	89	80	121	177	467	66,71%	579	82,71%	0,4826
1	1	0	16	28	45	52	79	90	100	92	197	479	68,43%	592	84,57%	0,4829
2	1	2	12	28	35	65	77	90	108	100	182	480	68,57%	604	86,29%	0,4864
3	1	3	7	30	51	72	77	96	74	108	181	459	65,57%	589	84,14%	0,4825
4	1	4	9	26	50	76	54	87	75	97	221	480	68,57%	589	84,14%	0,4918
5	0	2	17	35	43	64	81	86	90	86	196	458	65,43%	588	84,00%	0,4779
6	1	2	9	28	34	64	80	93	80	104	205	482	68,86%	602	86,00%	0,4937
7	1	4	8	26	56	67	58	81	87	102	210	480	68,57%	592	84,57%	0,4916
0 y 1	0	0	2	17	25	37	53	83	101	126	256	566	80,86%	651	93,00%	0,5289
0, 1 y 2	0	0	1	6	12	28	52	63	107	128	303	601	85,86%	676	96,57%	0,5546
0 a 3	0	0	1	4	13	21	36	49	91	118	367	625	89,29%	679	97,00%	0,5798
0 a 4	0	0	1	3	10	11	28	41	80	122	404	647	92,43%	684	97,71%	0,5947
0 a 5	0	0	0	4	10	9	26	40	73	116	422	651	93,00%	685	97,86%	0,6061
0 a 6	0	0	1	4	9	10	26	39	73	113	425	650	92,86%	683	97,57%	0,6072
0 a 7	0	0	1	4	9	10	26	36	73	114	427	650	92,86%	683	97,57%	0,6077

Tabla 12. Resultados obtenidos por múltiples particionados

b) Recubrimientos.

Un recubrimiento es un conjunto de subconjuntos que tiene la propiedad de que la unión de estos sea el todo. Renuncia a que sus partes sean disyuntas. En promedio estos subconjuntos serán más grandes de lo que hubieran sido en un particionado y por lo tanto se pierde parte de la ventaja de reducir el volumen total del corpus a sólo una parte. Una de las formas más sencillas de definir recubrimientos con el debido control de su crecimiento es definir primero un particionado y luego extender esto a elementos vecinos.

No resultó sencillo encontrar una fórmula general para decidir la ampliación teniendo en cuenta que una ampliación demasiado generosa si bien pueda dar una mayor calidad baja el tiempo de la respuesta que es justamente la meta original de este estudio.

Finalmente, dos formas fueron pensadas que dieron un crecimiento moderado y controlado de las partes:
(b1) Una vez terminado el particionado se puede recorrer los documentos y observar las K distancias a los centroides. La menor distancia debiera coincidir con la partición a la cual quedó asignado. Usando esta menor distancia como referencia se puede decidir que aquellas distancias que superan a ésta en un bajo porcentaje puedan producir una pertenencia adicional del documento a otra partición.

El tiempo de proceso es proporcional al tamaño del corpus y a la cantidad de particiones efectuadas. En un mismo recorrido se puede recoger información para distintos porcentajes. Aquí también se produce una fragmentación adicional del corpus en elementos exclusivos de una partición y de las diversas formas de ser compartido por dos o más.

La segunda técnica incorpora a los más cercanos a los documentos de la partición si aún no estuvieran. Conceptualmente, el modo de lograrlo es utilizar los mismos documentos como consulta, la cual ordenará a todos los documentos por su afinidad al analizado y decidir que cierto número de ellos, tomados desde la cabeza de la lista sean forzados, si aún no lo estuvieran, a integrar la partición ampliada. Hacerlo eficientemente es complejo para no enfrentar a cada documento con todos los documentos. Siendo un pre-proceso, no afecta a la velocidad de las consultas reales.

c. Análisis de Resultados

Experimentos realizados

Las dos técnicas para extender las particiones y construir recubrimientos fueron simuladas para visualizar la disminución de la aceleración pretendida. (Ver imagen 1). La primera, basada en distancia de centroides es sencilla de programar y alcanza niveles altos de calidad a costo de sacrificar el beneficio de explorar partes pequeñas adicionales, Si se conformara en reducir el volumen a examinar en medio cuerpo en lugar del cuarto inicial, habría que detenerse en 0,18 de tolerancia.

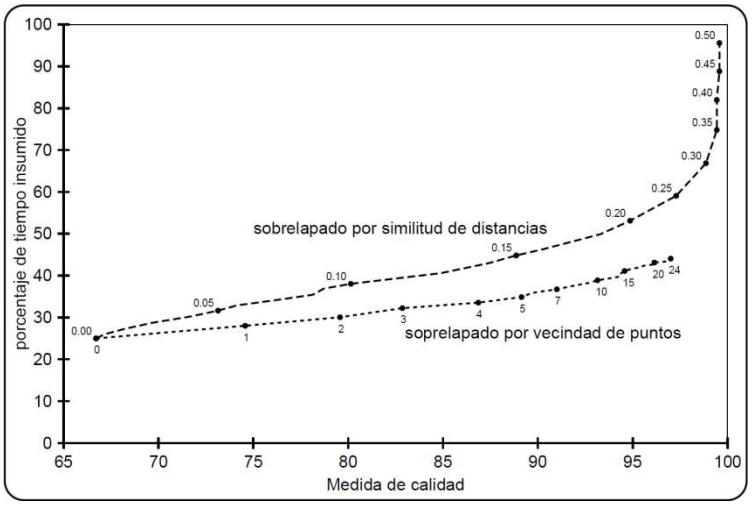

Imagen: 1. Resultados obtenidos en las dos técnicas de extensión

La segunda, basada en incorporar los n documentos más cercanos a uno dado, si es que no estuviera ya en la partición supera a la otra, para igual calidad, en la magnitud de los crecimientos en volumen de los recubrimientos y con la consiguiente ganancia en velocidad de respuesta. Es notable el beneficio con sólo incorporar condicionalmente al más cercano, pasando de 65 a 75%. La pérdida de tiempo fue muy baja ya que para muchos documentos su más cercano ya estaba en la misma partición y por lo tanto no significó un real incremento. Nuevamente se observa una ley de beneficios decrecientes, por lo cual se interrumpió la simulación en 24 vecinos. La superioridad de esta técnica reside en que el criterio de incorporación se corresponde con el criterio de resolver las consultas. La primera amplía sin saber si un documento ya tiene sus vecinos cerca.

d. Otras actividades

El presente proyecto es uno de varios que giran alrededor de la temática Recuperación de Información, habidos en el pasado y que se espera continúen en el futuro.

Uno de los problemas con los cuales se enfrentan los distintos proyectos es la carencia de Corpus Documentales en lengua española. Se comenzó con un Corpus Jurídico y al servicio de una necesidad profesional concreta, la cual proveyó los textos. Se ha procesado una traducción

española de la biblia, de dominio público, pero que no puede ser considerado un corpus voluminoso.

Se decidió recurrir al periodismo que constituye una fuente continua de modo tal que con una labor paciente pero automatizada se puede construir un corpus.

Se estudió la estructura gramatical de los html [8], utilizados por diversos periódicos. Descubriendo las reglas que se repiten en cada noticia por cada uno de los periódicos. Se trató de modelizar la sintaxis de cada uno con expresiones regulares. El lenguaje C# [9], es particularmente poderoso en las acciones de análisis y extracción que se puede lograr con los mismos. La complejidad de las expresiones ilustradas en el cuadro siguiente, que fueron necesarias para uno de los periódicos, da una idea de la magnitud de la tarea.

```
url -> http://www.infobae.com (Ya está establecida)
Direcciones dentro de la pantalla inicial ->
 (?<=href=")(\/\w+)+\/\d{4}\/\d{2}\/\d{2}[^"]+
titulo de la noitica ->  (?<=<meta property="og:title" content=")[^"]+
día de creación de la noticia  ->  (?<=>)\s*\d+ de \w+ de \d+
clasificación  ->  (?<=<meta property="article:section" content=")[^"]+
copete  ->  (?<=subheadline">)[^<]*
cuerpo  ->  <p class="element element-paragraph">.*<\/p>
Pasos adicionales para la obtención del cuerpo
reemplazo <[^>]+>|
reemplazo \s{2,}|
tag  o keywords -> (?<=:tag" content=")[^\"]+
```

Cuadro 1. Ejemplo del uso de las expresiones regulares.

Separado un artículo se lo almacena un archivo con una estructura similar al html pero con rótulos propios.

Carácter	Descripción
<periodico></periodico>	nombre del periódico
<titulo></titulo>	título de la noticia
<url></url>	dirección del artículo.
<dia></dia>	día de emisión del artículo
<diaob> </diaob>	día de obtención por el sistema
<clasificacion> </clasificacion>	otorgada por el portal
<copete> </copete>	copete de la noticia
<cuerpo> </cuerpo>	el contenido del artículo
<tag> </tag>	palabras claves

Cuadro 2. Nómina de Tag utilizados en el trabajo.

El Anexo II ilustra el texto de un artículo, cuyo cuerpo fue amputado en el medio para no ocupar más de una página. La complejidad de las expresiones diseñadas para la extracción da una idea de la magnitud de la tarea realizada. En este momento se dispone ya, de más de 60.000 documentos. Esta tarea fue publicada en CONAIISI 2018 [10].

2.1. Metodología e instrumentos aplicados.

Fueron explicadas en el desarrollo de las tareas planteadas para el segundo año del proyecto.

2.2. Resultados obtenidos:

a. Difusión en congresos, eventos científicos y publicaciones en revistas especializadas.

Dada la originalidad de esta línea de trabajo fue presentado en:

- En el XX Workshop de Investigadores en Ciencias de la Computación (WICC 2018, Universidad Nacional del Nordeste) exponiendo las ideas directrices, con el título "*Recuperación de la información"*

Los resultados de las experimentaciones fueron expuestas en dos congresos en el año 2018:

- En el XXIV Congreso Argentino de Ciencias de la Computación (CACIC 2018), bajo el título: "*Aceleración en la Recuperación de Información utilizando Algoritmos de Minería de Datos de R*" y
- En el Quinto Congreso Internacional de Educadores en Ciencias Empíricas en Facultades de Ingeniería (ECEFI 2018), titulado "Recuperación de Información acelerada con Algoritmos de Minería de Datos".

Como fue necesario construir grandes repositorios de documentos se presentó el tema:

- "*Método para la construcción de un Corpus periodístico mediante Expresiones Regulares*" en el 6to. Congreso Nacional de Ingeniería Informática – Sistemas de Información (CONAIISI 2018). En la ciudad de Mar del Plata.

b. Gestión y formación de recursos humanos (altas y/o bajas).

Direcciones y Tutorías de alumnos de grado y posgrado:

• Spositto, Osvaldo: Director de Tesis de Maestría en informática del Lic. Julio César Bossero.
Título: Estudio Comparativo de Técnicas de Minería de Datos para la predicción de la deserción universitaria, Escuela de Posgrado, Universidad Nacional de La Matanza. Inicio: Junio 2017. Finalizada y defendida el 8/12/2018.

• Spositto, Osvaldo: Director de Tesis de Maestría en informática del Ing. Casuscelli, Mauro Javier.
Título: Estudio comparativo de DBScan, Kmeans con Redes Neuronales en un sistema de recuperación de información. Escuela de Posgrado, Universidad Nacional de La Matanza. Inicio: enero 2018. Etapa de desarrollo.

c. Transferencia efectuada en el marco del proyecto.

- *Los temas investigados sirvieron para reforzar conceptos en la cátedra Inteligencia de Negocios (BI) perteneciente a la carrera Licenciatura en Gestión de Tecnología.*

d. Vinculación con otros grupos de investigación / organismos.

- *Modelos de minería de datos para el diagnóstico de enfermedad de Parkinson mediante el análisis de voz. Argentina. ProInce (2017-2018).*

4.- Conclusión:

Las técnicas ensayadas producen resultados satisfactorios, pero este es un tema que es de no acabar, pues cada vez que se ensaya una nueva alternativa surgen ideas de cómo a su vez mejorarla y dificultades a resolver.

Un tema importante que queda para quienes quieran continuar con este tema, es particionar en más partes y fusionar dos o más entre los más promisorios. Alternativa laboriosa de desarrollar pues de usar redes, estas son difíciles de entrenar si cada una de ellas se entrena con un bajo porcentaje de muestras positivas.

Tener valores individuales de pertenencia es imprescindible para poder elegir las más promisorias. Por otra parte es necesario que las partes sean de tamaños comparables. A lo largo de este proyecto sólo se aprobaron particionados donde el desequilibrio no pasara de un 20%. Conseguir eso en particionados de muchas partes requiere experimentar muchas veces hasta lograrlo. De allí se concluye que de encararlo debería hacerse en más de una etapa.

5.- Referencias bibliográficas.

[1] Martínez Méndez, Francisco. (2004). Recuperación de información: modelos, sistemas y evaluación. Editorial: Murcia Kiosko. ISBN: 84-932537-7-4. Disponible: http://libros.metabiblioteca.org/bitstream/001/227/8/84-932537-7-4.PDF

[2] Oliván, José A. y otros. (2005). Una aproximación al concepto de recuperación de información en el marco de la ciencia de la documentación. Investigación Bibliotecológica, Vol. 20, Núm. 41, julio/diciembre, 2006, México, ISSN: 0187-358X. pp. 13-43. Recuperado en 08/02/2019, de http://www.scielo.org.mx/pdf/ib/v20n41/v20n41a2.pdf.

[3] Marsden, Jerrold Eldon, (2004). Tromba, Anthony J. Calculo vectorial. - 5. Ed. Pearson Educación. España.

[4] García, W. y Fernández Arenas, D. (). El gradiente y el método de los multiplicadores de Lagrange. Licenciatura en Matemáticas y Física. Universidad de Antioquia.Disponible en:https://almagestoudea.files.wordpress.com/2008/07/el-gradiente-y-los-multiplicadores-de-lagrange.pdf

[5] José Hernández Orallo y otros. (2004). Introducción a la minería de datos. Editorial: Pearson. Edición: 2004. ISBN: 84 205 4091 9

[6] Jure Leskovec y otros. (2014) Mining of Massive Datasets. Disponible en:http://infolab.stanford.edu/~ullman/mmds/book.pdf

[7] Libro de actas del XX Workshop de Investigadores en Ciencias de la Computación (WICC 2018, Universidad Nacional del Nordeste). Recuperado en 08/02/2019, de: http://wicc2018.unne.edu.ar/wicc2018librodeactas.pdf

[8] J. D. Gauchat. (2012). El gran libro de HTML5, CSS3 y Javascript ISBN eBook: 978-84-267-1782-5. Disponible en: https://gutl.jovenclub.cu/wp-content/uploads/2013/10/El+gran+libro+de+HTML5+CSS3+y+Javascrip.pdf

[9] Ceballos Sierra, Fco. Javier. (2011). Microsoft C#. Curso de Programación. 2ª Edición. Isbn: 978-84-9964-068-6. Editorial Ra-Ma Editorial.

[10] Osvaldo Spositto, Gastón Procopio y Julio Bossero. (2018). Método para la construcción de un Corpus periodístico mediante Expresiones Regulares. Departamento de Ingeniería e Investigaciones Tecnológicas. UNLaM. Recuperado en 08/02/2019, de: https://www.conaiisi2018mdp.org/memorias/memorias.html#

Apéndice I: Resultado de experimentos Con particiones (K) igual a 3, 4 y 5.

Consulta	particionado	La mejor	La segfunda	Por dis centroides0	Por dis centroides0	Por redes	Evaluación sobre cada partición										VALORES CON LOS CUALES SE DECIDIÓ														
																	Dist. a centroides (promedio)					Dist. a centroides (sobre casquete)					Salidas de red				
							0		1		2		3		4		0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
2	3	3	2	3	3	0	2	41	0	153	4	28	4	26	0	140	0,135	0,268	0,13	0,115	0,226	0,057	0,171	0,053	0,032	0,151	0,432	0,248	0,13	0,345	-0,154
2	5	2	3	3	3	3	0	140	0	153	6	16	4	24	0	326	0,226	0,232	0,128	0,121	0,295	0,151	0,157	0,037	0,03	0,219	0,026	-0,033	0,135	0,856	0,017
3	3	4	2	2	4	4	0	739	1	55	2	96	0	86	7	20	0,463	0,34	0,247	0,291	0,25	0,415	0,253	0,18	0,225	0,177	0,043	-0,028	0,165	0,092	0,728
4	1	1	3	3	3	3	1	111	5	19	0	166	4	22	0	200	0,268	0,189	0,297	0,187	0,316	0,2	0,115	0,232	0,113	0,25	-0,143	0,671	-0,188	0,783	-0,122
4	3	1	0	0	1	1	1	137	9	11	0	168	0	162	0	124	0,229	0,244	0,271	0,266	0,28	0,159	0,143	0,206	0,197	0,21	-0,029	1,154	-0,015	-0,006	-0,104
6	1	1	2	2	2	2	0	372	4	25	4	26	0	236	2	52	0,358	0,197	0,194	0,318	0,23	0,298	0,123	0,12	0,256	0,156	0,01	0,386	0,645	-0,021	-0,021
6	6	2	1	1	1	1	1	58	4	24	5	22	0	486	0	193	0,236	0,192	0,207	0,373	0,312	0,168	0,117	0,131	0,312	0,247	-0,103	0,839	0,413	-0,075	-0,075
7	3	1	3	0	0	1	0	54	7	15	0	154	3	94	0	219	0,235	0,282	0,291	0,275	0,372	0,165	0,186	0,228	0,207	0,312	0,179	0,87	-0,012	0,051	-0,087
7	4	3	1	3	3	0	3	39	3	35	0	78	4	25	0	219	0,227	0,27	0,33	0,209	0,372	0,152	0,193	0,257	0,132	0,312	0,589	0,176	-0,082	0,455	-0,138
8	3	2	4	2	2	4	0	428	0	126	6	17	2	40	2	33	0,285	0,277	0,145	0,178	0,188	0,22	0,181	0,07	0,101	0,109	-0,008	-0,095	0,343	0,371	0,389
8	4	1	4	4	4	1	0	241	8	13	0	155	0	263	2	33	0,269	0,195	0,277	0,29	0,188	0,198	0,11	0,198	0,22	0,109	-0,037	0,712	-0,039	-0,039	0,403
8	5	2	3	0	0	2	2	33	0	215	5	18	3	37	0	155	0,188	0,278	0,196	0,194	0,277	0,109	0,207	0,112	0,11	0,198	0,374	-0,058	0,464	0,322	-0,101
8	6	1	2	1	1	0	2	48	5	20	3	58	0	243	0	38	0,22	0,203	0,232	0,306	0,212	0,151	0,13	0,158	0,239	0,137	0,233	0,232	0,145	0,217	0,173
9	3	1	3	3	3	3	1	112	5	31	2	59	2	33	0	68	0,145	0,196	0,145	0,13	0,187	0,068	0,089	0,07	0,048	0,109	0,078	0,175	0,382	0,431	-0,065
9	4	1	3	3	1	1	2	66	5	16	0	235	3	54	0	68	0,145	0,139	0,238	0,135	0,187	0,062	0,048	0,155	0,05	0,109	0,355	0,907	-0,083	0,272	-0,451
9	5	1	3	3	3	1	0	68	5	31	2	51	3	29	0	235	0,187	0,14	0,142	0,136	0,238	0,109	0,056	0,052	0,046	0,155	-0,233	0,63	0,483	0,345	-0,225
9	6	0	2	0	2	2	4	51	3	25	3	24	0	188	0	198	0,148	0,153	0,149	0,216	0,231	0,072	0,075	0,067	0,141	0,158	0,29	0,172	0,301	0,107	0,13
10	1	4	2	2	2	4	0	135	0	197	2	27	0	248	8	13	0,263	0,288	0,21	0,324	0,214	0,195	0,223	0,137	0,262	0,138	0,095	0,112	0,216	0,123	0,454
11	0	0	1	1	1	1	6	25	4	19	0	134	0	403	0	132	0,147	0,127	0,215	0,278	0,227	0,068	0,046	0,142	0,207	0,157	0,705	0,786	-0,176	-0,165	-0,151
11	1	3	2	3	2	3	0	149	0	134	2	28	8	15	0	384	0,236	0,217	0,133	0,133	0,278	0,165	0,145	0,053	0,053	0,208	-0,2	-0,222	0,769	0,829	-0,177
11	3	1	3	2	2	3	0	471	5	30	2	32	3	24	0	130	0,251	0,211	0,116	0,14	0,229	0,183	0,106	0,038	0,059	0,155	0,02	0,226	0,419	0,437	-0,102
11	5	4	3	3	3	4	0	130	0	203	2	32	3	24	5	30	0,229	0,261	0,164	0,158	0,172	0,155	0,188	0,076	0,07	0,083	-0,207	-0,129	0,481	0,32	0,535
13	2	1	2	2	2	2	0	281	5	28	3	27	2	46	0	147	0,226	0,133	0,116	0,144	0,202	0,149	0,052	0,039	0,058	0,127	-0,051	0,216	0,826	0,065	-0,055
13	3	3	2	2	3	4	0	146	0	229	3	44	6	26	1	27	0,146	0,212	0,108	0,112	0,118	0,069	0,107	0,029	0,029	0,033	-0,006	-0,224	0,311	0,391	0,528
13	4	1	4	4	1	1	0	422	9	12	0	246	0	404	1	27	0,221	0,12	0,206	0,218	0,118	0,146	0,027	0,121	0,141	0,033	-0,09	1,119	-0,166	-0,11	0,248
13	5	3	2	3	3	0	1	27	0	300	3	43	6	26	0	246	0,118	0,219	0,121	0,118	0,206	0,033	0,142	0,028	0,026	0,121	0,666	-0,184	0,445	0,308	-0,235
14	3	1	4	4	4	4	0	100	6	19	0	266	0	139	4	22	0,17	0,247	0,246	0,214	0,166	0,096	0,147	0,179	0,141	0,086	0,069	0,226	-0,088	-0,047	0,84
14	4	4	3	4	3	4	3	42	0	82	0	340	3	39	4	22	0,184	0,213	0,302	0,167	0,166	0,105	0,13	0,226	0,085	0,086	0,181	-0,064	-0,04	0,127	0,796

Apéndice II: Modelo de nota periodística de un portal Web.

<periodico>clarin**</periodico>**

<titulo>A diez años del último título de Messi con la de Argentina**</titulo>**

<url>http://www.clarin.com/deportes/seleccion-nacional/anos-ultimo-titulo-messi-argentina_0_BJLIsxq8Q.html**</url>**

<dia>22 de agosto de 2018**</dia>**

<diaob>22 de agosto de 2018**</diaob>**

<Clasificacion>Deportes**</clasificacion>**

<copete>La Sub 23 brilló en los Juegos Olímpicos de Beijing y ganó la medalla de oro.**</copete>**

<cuerpo>"Me toca decidir a mí y los Juegos Olímpicos es algo que no voy a poder jugar nunca más. Creo que Barcelona entiende lo que pienso y no creo que vaya a haber conflicto o problema alguno por esto". Lionel Messi tenía apenas 20 años el 21 de mayo de 2008, y dejaba en claro su posición. Tres meses después, se coronaba campeón olímpico en Beijing. Hace 10 años, el 23 de agosto en la madrugada argentina, el mediodía chino, el mejor jugador del mundo del siglo XXI no imaginaba que sería su último festejo grande con la celeste y blanca. Nadie lo imaginaba. Barcelona no se resignó fácilmente a cederlo. Si bien el nuevo entrenador Pep Guardiola apoyó al argentino en su pedido, el club acudió al Tribunal de Arbitraje Deportivo (TAS) porque entendía que la

. . . . . .

tras un decepcionante partido en Chile por Eliminatorias el Coco renunció. La posta la tomó Diego Armando Maradona, muy cercano al grupo en Beijing, y Riquelme renunció para siempre al seleccionado. "No tenemos los mismos códigos y no podemos trabajar juntos", disparó contra el Diez. Además de los nombrados, los otros jugadores elegidos por el Checho Batista fueron Oscar Ustari (se lesionó en pleno torneo y en su lugar acudió Nicolás Navarro), Ezequiel Garay, Fabián Monzón, Pablo Zabaleta (el otro mayor), Nicolás Pareja, Fernando Gago, José Sosa, Ever Banega, Ezequiel Lavezzi, Lautaro Acosta y Diego Buonanotte. El plantel olímpico festejando en China. El camino de Argentina hacia la consagración se inició el 7 de agosto con el triunfo por 2-1 ante Costa de Marfil (Messi y Acosta); luego, venció 1-0 a Australia el 10 de agosto (Lavezzi) y se aseguró el primer puesto en el grupo al vencer 2-0 a Serbia (Lavezzi y Buonanotte) el 13 de agosto. En cuartos de final, el 16 de agosto, derrotó 2-1 en tiempo suplementario a Holanda (Messi y Di María). Y en semifinales, el 19 de agosto, despachó 3-0 a Brasil (Agüero 2 y Riquelme)."Nunca le dije que no a la Selección. Y nunca se lo voy a decir. Siempre dije que voy a jugar los partidos de la Selección, en el lugar que sea., porque es lo más lindo que hay", le dijo Messi a Clarín tras la consagración. Para Messi pasaron tres mundiales, dos Copa América, una Copa América Centenario, tres subcampeonatos, más de 100 partidos y 50 goles. A una década de aquel gran festejo. De esta enorme ausencia en el nuevo ciclo que se avecina.
</cuerpo>

<tag>anos, último, titulo, messi, argentina**</tag>**

Anexo III: Copia de artículos presentados en publicaciones periódicas, y ponencias presentadas en eventos científicos.

WICC 2018

XX Workshop de Investigadores en Ciencias de la Computación

RedUNCI

UNNE

Se certifica que:

Ryckeboer Hugo

ha participado en calidad de expositor:

Recuperación de la Información

aceptado en el XX Workshop de Investigadores en Ciencias de la Computación, realizado en la ciudad de Corrientes, los días 26 y 27 de abril de 2018.

LIC. PATRICIA PESADO
COORDINADORA RED UNCI

FACULTAD DE CIENCIAS EXACTAS Y NATURALES Y AGRIMENSURA
UNIVERSIDAD NACIONAL DEL NORDESTE
CORRIENTES

MGTER. GLADYS DAPOZO
COMITÉ ORGANIZADOR UNNE

CACIC 2018
congreso argentino de ciencias de la computación

Se certifica que

Hugo Hugo Ryckeboer

ha participado en carácter de asistente del XXIV Congreso Argentino de Ciencias de la Computación (CACIC 2018), realizado en la ciudad de Tandil del 8 al 12 de octubre de 2018.

Mg. Claudio Aciti
Comité organizador
CACIC 2018

Lic. Patricia Pesado
coordinadora titular
RedUNCI

RedUNCI
FACULTAD DE CIENCIAS EXACTAS
UNICEN

CoNaIISI 2018

CERTIFICAMOS que

Julio Bossero

ha participado en calidad de **EXPOSITOR** del trabajo **"Método para la construcción de un Corpus periodístico mediante Expresiones Regulares"** de los autores **Osvaldo Sositto, Gastón Procopio, Julio Bossero** en el VI Congreso Nacional de Ingeniería en Informática / Sistemas de Información, desarrollado en la Universidad CAECE de Mar del Plata los días 29 y 30 de noviembre de 2018.

Mg . Lucia Rosario Malbermat
Coordinadora Local

Ing.Nelson Roberto Sotomayor
Coordinador del Comité Académico

UNIVERSIDAD TECNOLÓGICA NACIONAL

UTN FACULTAD REGIONAL MENDOZA | SECRETARÍA DE EXTENSIÓN UNIVERSITARIA

CERTIFICADO

QUINTO CONGRESO INTERNACIONAL DE EDUCADORES EN CIENCIAS EMPÍRICAS EN FACULTADES DE INGENIERÍA: ECEFI 2018

Por cuanto

HUGO EMILIO RYCKEBOER - D.N.I. 4.538.872

ha participado como ***expositor*** *de la* ***ponencia***

"Recuperación de Información acelerada con Algoritmos de Minería de datos."

realizado los días 04 y 05 de octubre de 2018. Autorizado por Resolución Nº 654/18 de fecha 01/10/18, en la Universidad Tecnológica Nacional – Facultad Regional Mendoza, por lo que se hace entrega del presente certificado.

Mendoza, Argentina, febrero de 2019.

UNIVERSIDAD TECNOLÓGICA NACIONAL FACULTAD REGIONAL MENDOZA REPÚBLICA ARGENTINA

Ing. CARLOS OSCAR MALLEA
Secretario de Extensión Universitaria
U.T.N. – F.R.M.

Esp. Ing. JOSÉ BALACCO
Decano
U.T.N. – F.R.M.

a) WICC-2018.

RECUPERACIÓN DE LA INFORMACIÓN

Ryckeboer Hugo, Spositto Osvaldo, Bossero, Julio César, Barone Miriam
Departamento de Ingeniería e Investigación Tecnológica
Universidad Nacional de La Matanza
hryckeboer@unlam.edu.ar spositto@unlam.edu.ar
jbossero@unlam.edu.ar mbarone@unlam.edu.ar

RESUMEN

Las Técnicas de Recuperación de Información que responden a inquietudes puntuales, se han hecho populares gracias a los buscadores ofrecidos gratuitamente a quienes recurren al Internet. Y no se hace referencia únicamente a grandes repositorios, sino también a medianos y pequeños, con miles de documentos, donde también es un inconveniente la localización del documento o los documentos que respondan a la inquietud del Usuario.

El grupo posee sus propios motores orientados a corpus estáticos. De las diversas concepciones existentes ha prestado especial atención a la indexación semántica latente conocidas por sus siglas LSI.

Una vez construidos los motores las líneas de investigación se han orientado a enfoques que permitan acelerar los mismos tanto en la búsqueda como en los preprocesos.

Uno de ellos es el uso del procesamiento paralelo, tanto en clúster de máquinas como en el uso de placas de video.

La técnica LSI es particularmente dependiente en su preproceso de un eficiente cálculo de autovalores y autovectores de matrices de gran tamaño, lo que hace incluir en nuestra temática el cálculo numérico.

Se investiga si es factible acelerar la selección de los documentos que responden a un requerimiento por medio de un particionado del corpus basándose en criterios de similitud propia de minería de datos y técnicas de selección de la parte usando redes neuronales.

En este sentido se exponen distintas líneas de trabajo a seguir, teniendo como objetivo diseñar, implementar y probar modificaciones en los procesos de filtrado y ordenamiento de documentos, en un Sistema de Recuperación de Información (SRI), aplicando algoritmos de clustering tradicionales.

Palabras claves: Examinar, Indexar, Buscar, SRI, LSI, Minería de Datos, Agrupamiento, Algoritmos de Clasificación (Browse, Indexing, Search)

CONTEXTO

Esta propuesta de trabajo se lleva cabo dentro de dos proyectos de investigación "*Optimización de la Recuperación de Documentos, usando como técnica base el LSI (Lematización Semántica Latente)",* y el proyecto "*Uso de Minería de Datos para acelerar la recuperación de documento*"

Los cuales son desarrollados por el grupo de investigación del Departamento de Ingeniería e Investigaciones Tecnológicas de la Universidad Nacional de La Matanza, en el marco de investigación del programa PROINCE.

1.INTRODUCCIÓN

Los contenidos de los documentos digitales hoy en día, son una materia prima muy valiosa, tanto para empresas u organizaciones como para simples usuarios. Es por esto que en la Sociedad de la Información se destinan gran cantidad de recursos en almacenar grandes volúmenes de documentos, organizarlos para luego recuperar los adecuados, debido entre otras cosas, a la explosión en el número de fuentes de información disponibles en Internet que sobrepasa a toda información manual. También hay conjuntos de documentos cerrados, por ejemplo, los legislativos, las obras de filósofos famoso, sobre los cuales se desean efectuar consultas puntuales.

Este es uno de los motivos por lo que, desde hace años, se dispone de los denominados Sistemas de Recuperación de Información *(SRI o IRS en inglés Information Retrieval Systems),* que permiten almacenar, buscar y mantener documentos, extendiendo esto a textos, imágenes, vídeos, audios y otros objetos multimedia, los cuales, utilizan técnicas de búsqueda relativas a su contenido, que son específicas para cada tipo de información.

Para evitar una dispersión en un grupo humano pequeño, los proyectos se han centrado sobre información textual en español

Las manifiestas similitudes existentes entre la recuperación de información y otras áreas vinculadas al procesamiento de la información, se repiten en el campo de los sistemas encargados de llevar a cabo esta tarea. Para Salton en [Sal86] "*…la recuperación de información se entiende mejor cuando uno recuerda que la información procesada son documentos…*", con el fin de diferenciar a los sistemas encargados de su gestión de otro tipo de sistemas, como los gestores de bases de datos relacionales. Salton piensa que "…cualquier SRI puede ser descrito como un conjunto de ítems de información (DOCS), un conjunto de peticiones (REQS) y algún mecanismo (SIMILAR) que determine qué ítem satisfacen las necesidades de información expresadas por el usuario en la petición"

Para poder hacer aportes originales en esta temática fue fundamental tener motores completos en estado operativo.

El grupo ha privilegiado la metodología LSI porque además de su conocida habilidad para resolver ambigüedad, equivocidad y sinonimia, provee vectores descriptivos de los documentos y de consultas de menor dimensión lo que beneficia a la minería de datos.

Los motores construidos son lo suficientemente abiertos y flexibles para ser utilizado en docencia y en puestos de investigación.

Cada día se utilizan técnicas más avanzadas de análisis del contenido de los documentos con vistas a mejorar los tiempos de acceso a los documentos y la efectividad del resultado.

Por lo tanto, el problema al que se enfrenta la RI se puede definir como: "Dado un conjunto de documentos, ordenar los documentos de mayor a menor según la relevancia para una determinada necesidad ya expresada como consulta, las limitaciones perceptivas del usuario aconsejan entregar los elementos que encabezan la lista".

Aunque una buena parte del pre proceso de organización y proceso de consultas recurren a técnicas básicas de la computación se pueden señalar algunas áreas que dan lugar a mejoras y optimizaciones las que influirán en la calidad de las prestaciones:

Análisis Textual: es una práctica ya establecida que en lugar de recurrir a la presencia o no de palabras identificadas en la consulta dentro de los documentos, es conveniente reducirlas a lexemas ignorando flexiones propias del desarrollo del texto, pero salvando un concepto común que la palabra encierra en sus distintas formas morfológicas.

Esta actividad en sus detalles es dependiente del idioma y del campo de aplicación.

Proceso del Corpus: Los coeficientes de los vectores que describen la temática de los documentos requieren un ajuste a la luz de la totalidad de los documentos disponibles. Esto en el caso del LSI requiere hallar autovalores y autovectores de elevada dimensión, problema no trivial por los errores de redondeo del cálculo con reales. Continuamente aparecen propuestas que intentan acelerar o mejorar tales cálculos. Algunos cambios algorítmicos son consecuencia de la existencia de jerarquía en las memorias y en las comunicaciones.

Resolución de las Consultas: Una forma de saber el valor de un documento frente a una consulta es enfrentar los vectores representativos de ambos. El tiempo que insume la construcción de la lista ordenada crece con el tamaño del corpus.

Dos líneas de trabajo surgen frente a esta situación:

- Distribuir los documentos en varios procesadores
- Dividir el corpus en conjuntos más pequeños conteniendo documentos "similares" y comenzar la recuperación por la parte más promisoria.

Tanto en el dividir como elegir esta parte lo estamos encarando con técnicas de minería de datos.

Los cálculos de autovalores involucran a todos los coeficientes de la matriz lo que crea un dilema y necesidad de hallar un compromiso entre distribuir el cálculo aumentando las comunicaciones entre procesadores, o concentrarlos para no tener tales recargos de tiempos.

2. LÍNEAS DE INVESTIGACIÓN Y DESARROLLO

Mejorar el trabajo con RI implica, por lo que se ha dicho, considerar diferentes líneas de investigación:

a) Acelerar la velocidad de cómputo, recurriendo a un procesamiento en paralelo
b) Subdividir el corpus en forma inteligente de modo tal que sin gran pérdida de exhaustividad se pueda resolver la consulta examinando una o más partes de la subdivisión, excluyendo a muchas de ellas.
c) Mejorar la lematización disponible del idioma español, dado que la misma no da resultados satisfactorios.

Respectivamente, se señalan las observaciones que serán las inquietudes centrales en estas líneas de investigación:

Los sistemas que operan en gran escala deben recurrir necesariamente al uso en paralelo de varios procesadores. Se estudia la forma de paralelizar algunos algoritmos para acelerar adecuadamente los cómputos.

Dada la posibilidad de extender la selección de documentos a corpus muy voluminosos, existen diversas ideas de subdividir el corpus en grupos aplicando técnicas de agrupamiento. Para que la subdivisión sea efectiva a los fines propuestos se debe agrupar documentos de iguales características y separar los que manifiestamente difieren. En este proyecto se pretende dominar e incorporar estas tecnologías a nuestro prototipo, con la intención de evaluar si la mejora en velocidad compensa una eventual pérdida de exhaustividad a las mismas. Las tecnologías que resuelven el particionado por similaridad se conocen como "clustering" [Her04]. El proyecto pretende subdividir un corpus en varios grupos o clúster, para realizar la búsqueda de documentos pertinentes a una consulta dada. Por lo que la hipótesis principal es que la utilización de técnicas de clustering y de aprendizaje supervisado, en un SRI, acelerará la obtención de documentos pertinentes. Los beneficios alcanzarían a los usuarios directos y a la reducción de recursos computacionales. La subdivisión encierra un riesgo de respuestas no exhaustivas que afecten a la calidad del servicio. Esto se encuentra en etapa de evaluación.

Explorar a fondo las distintas alternativas que se presentan en las diversas etapas de esta solución constituyen el centro de uno de los proyectos en curso.

Las mismas se pueden resumir en:

- ✓ Elegir técnica de clustering o incluso diseñar una nueva.
- ✓ Diseñar el algoritmo que oriente una búsqueda en particular hacia uno o varios de los subconjuntos. Las redes neuronales son una de las técnicas que se vislumbran como promisorias para esta tarea.
- ✓ Proponer técnicas de evaluación en cuanto a cobertura lograda versus tiempo de respuesta.

Pensando en el preproceso está la lematizacion, esta se puede desdoblar o sea repartir hacia distintos procesadores, distintos documentos o inclusive repartiendo párrafos. La suma de la presencia de lexemas en distintos documentos también es paralelizable. La transposición de la tabla documento-termino también es realizable en paralelo diseñando cuidadosamente el algoritmo.

Objetivos Secundarios

De lo enunciado ut-supra se desprenden los siguientes objetivos secundarios:

a) Obtención de un Corpus en español, más voluminoso que el utilizado hasta ahora, para la aplicación de los algoritmos.

b) Evaluar y proponer algoritmos de agrupamiento para el conjunto de documentos disponibles.

c) Ajustar los modelos predictivos de tipo supervisados, para resolver el problema de clasificación.

d) Evaluar el alcance de los resultados a través de las métricas propuestas.

e) Aplicar los modelos desarrollados en nuevos Corpus.

3. RESULTADOS OBTENIDOS/ESPERADOS

Las técnicas de minerías de datos en uso en esta línea de investigación fueron profundizadas en proyectos previos lo que dio lugar a las siguientes publicaciones:

1- "Predicción del riesgo de abandono universitario utilizando métodos supervisados" En colaboración con Edwards, Diego y Pérez, Silvia (UNLaM). Trabajo presentado en el Workshop de la V Jornadas Nacionales y I Latinoamericanas de Ingreso y Permanencia en Carreras Científico – Tecnológicas. Facultad Regional Bahía Blanca. Universidad Tecnológica Nacional. Bahía Blanca. Mayo de 2016. IPECyT 2016

2- "Modelos de minería de datos para el diagnóstico de enfermedad de Parkinson mediante el análisis de voz". En colaboración con el Ing. Osvaldo Spositto, Ing. Gabriel Blanco, Mg. Mónica Giuliano y el Ing. Luis Fernández (UNLaM). Trabajo presentado en el Workshop del V Congreso Nacional de Ingeniería en Informática/Sistemas de Información. Publicación en línea - ISSN. CONAIISI 2017. Santa Fe. Argentina.

3- "Comparación de Algoritmos de Aprendizaje Supervisado para la obtención de perfiles de alumnos desertores". En colaboración con el Ing. Osvaldo Spositto (UNLaM). Trabajo presentado en el Workshop del IV Congreso Nacional de Ingeniería en Informática/Sistemas de Información. Publicación en línea - ISSN 2347-0372. CONAIISI 2016. Salta. Argentina

4- "Una paralización del método de Householder" En colaboración con el Ing. Osvaldo Spositto, Ing. Hugo Ryckeboer. (UNLaM). Trabajo presentado en el XXII Congreso Argentino de Ciencias de la Computación- CACIC 2016- Universidad Nacional de San Luis San Luis.

En el marco de la línea de investigación para acelerar la velocidad de cómputo, recurriendo a un procesamiento en paralelo a lo que respecta a la lematización del idioma español disponible que no daba resultados satisfactorios, se puede concluir que con el uso de los hilos se realiza un procesamiento más rápido que con la forma secuencial.

Finalmente, respecto a los sistemas que operan en gran escala, estos deben recurrir necesariamente al uso en paralelo de varios procesadores. La utilización de los GPU, es el acelerador dominante para realizar procesamiento de cálculos en paralelo, principalmente en matrices, ello tuvo un resultado positivo: los tiempos bajan drásticamente comparando un proceso secuencial en una Pc típica de escritorio contra el procesamiento sobre cualquiera de las GPU.

En cuanto a la comparación entre las diferentes GPU, se observa que los mejores tiempos se obtuvieron para la GPU R9 390X. A medida que haya más documentos, la diferencia de tiempos entre cada una de ellas se va notando considerablemente.

Los documentos de un Corpus son transformados en vectores descriptivos. Una consulta de usuario es también convertida en otro vector descriptivo. Para obtener un documento que satisfaga la necesidad de información del usuario, el vector de la consulta se debe enfrentar con todo el corpus, en búsqueda de similitudes. Este proceso genera un índice de relevancia, que será la salida que recibe el usuario, en forma de lista ponderada. Dividir el Corpus, de modo tal de poder desechar uno o más grupos, debería acelerar la búsqueda.

4. FORMACIÓN DE RECURSOS HUMANOS

Resultados en cuanto a la producción de conocimiento:

Disponer de un buen lematizador del español es una contribución al estado del conocimiento en Recuperación de Información en lengua española.

Resultados en cuanto a la difusión de resultados:

Del mismo modo que en el proyecto precedente se puso un Motor de Búsqueda a disposición de toda la comunidad académica, se hará lo mismo con el lematizador español.

Los resultados en materia de lematización y del uso de "clusters" de computadoras, serán expuestos en Congresos/Revistas.

Los lematizadores no exponen normalmente la metodología con la cual los diseñaron, con lo cual, la exposición de estos detalles, puede ser valiosa para profesionales de otras lenguas.

El armado de "clusters" obliga a resolver problemas prácticos de conexión y administración, lo

que puede acortar el camino a otros investigadores que se inicien en el tema.

Los profesionales de Informática disponen de baja formación lingüística, de modo tal que su participación en este proyecto les abrió nuevos campos de actividades: la lingüística computacional, el manejo de semántica, traducción automática. Todas estas actividades requieren un adecuado manejo morfológico del lenguaje.

Respecto a la minería de Datos se utilizaron parte de los conocimientos en 2 trabajos de tesis desarrolladas en sendas maestrías en informática:

Tesis aprobada: El Soporte Informático y su Aporte para la Inclusión Universitaria

Tesis en curso: Estudio Comparativo de Técnicas de Minería de Datos para la predicción de deserción universitaria

Resultados en cuanto a transferencia hacia las actividades de docencia y extensión:

Los sistemas de IR son eficaces en la medida que diseñen buenas estructuras de datos. Estas son estudiadas en materias intermedias de la Ingeniería de Sistemas, poder ilustrar su uso práctico, beneficiará a los estudiantes.

Del mismo modo, el uso de clusters ilustra los tópicos más avanzados de la Arquitectura de Computadoras, que también integran el Plan de Estudios.

5. BIBLIOGRAFÍA

[Her04] "Introducción a la minería de datos". José Hernández Orallo y otros. Editorial: Pearson. Edición: I. Año 2004

"Análisis Semántico Latente: una panorámica de su desarrollo". René Venegas. Rev. signos [online]. 2003, vol.36, n.53, pp.121-138. ISSN 0718-0934. Pontificia Universidad Católica de Valparaíso. Chile

Disponible en: http://dx.doi.org/10.4067/S0718-09342003005300008.

"Introduction to Modern Information Retrieval". Gerard Salton, Michael J. Michael J. McGill. Ed. McGraw-Hill, Inc. New York, NY, USA. ISBN: 0070544840. 1986

"Automatic Information Organization and Retrieval". Salton, G. McGraw−Hill, N.Y. 1968.

PCI-Express http://pcisig.com/specifications/pciexpress/resources

https://nvlabs.github.io/moderngpu/performance.html

"Clustering de documentos con restricciones de tamaño". Diego Fernando Vallejo Huanga. Trabajo Fin de Máster. Universitario en Gestión de la Información. Escola T. S. d'Enginyeria Informàtica Universitat Politècnica de València. 2015. Disponible: https://riunet.upv.es/bitstream/handle/10251/69089/VALLEJO-ClusteringdeDocumentosconRestricciones deTamaño.pdf?sequence=23

"Clusterdoc, un sistema de recuperación y recomendación de documentos basado en algoritmos de agrupamiento". Marylin Giugni O.;Luis León G.;Joaquín Fernández.Telematique, volumen 9 - número 2 - año 2010. Disponible en:http://publicaciones.urbe.edu/index.php/telematique/article/view/913/ pdf

Lindholm, Erik and Nickolls, John and Oberman, Stuart and Montrym, John, NVIDIA Tesla: A unified graphics *and computing architecture*,IEEE micro, 2008.

b) CoNaIISI-2018.

Método para la construcción de un Corpus periodístico mediante Expresiones Regulares

Osvaldo Spositto, Gastón Procopio y Julio Bossero.
Departamento de Ingeniería e Investigaciones Tecnológicas.
Universidad Nacional de La Matanza
Florencio Varela 1902, San Justo, Prov. Buenos Aires, Argentina
spositto@unlam.edu.ar, gprocopio@unlam.edu.ar, jbossero@unlam.edu.ar

Resumen

Desde hace tiempo la utilización de Corpus lingüísticos ha experimentado una gran evolución, entre otros factores por el uso creciente en proyectos de investigación en torno a los Sistemas de Recuperación de Información (SRI). El Corpus de Noticias en la Web (CONOWEB) se compone de noticias periodísticas de portales nacionales e internacionales, y tiene como objetivo servir como recurso para ser utilizado con algoritmos de minería de datos en un SRI. Este texto presenta los procedimientos lingüísticos y computacionales que se realizaron para el desarrollo del CONOWEB, desde la recolección de textos hasta la construcción de la herramienta. Se buscó basamento teórico-metodológico en la Lingüística Computacional y en las expresiones regulares. El principal propósito de este trabajo es, por tanto, hacer accesible una metodología para la construcción de un Corpus a toda la comunidad investigadora, poco experta en la materia.

1. Introducción

La recuperación de información (RI o IR por sus siglas en inglés Information Retrieval) es un área de investigación que se inició con los trabajos de Salton en los años 60 [1] y recientemente ha experimentado un desarrollo significativo motivado principalmente por el crecimiento de Internet, el incremento de documentación que se puede obtener de ella y la urgencia para localizar respuestas relevantes. Una característica muy importante de la RI es que se ocupa de los problemas de recuperar información por su contenido y no por sus metadatos. Es por este motivo que existen técnicas para recuperar información de diversos tipos: textos, imágenes, archivos de sonido y vídeo, etc. [2].

De la variada cantidad de definiciones sobre RI en este trabajo se adoptará la que dice: *"...la recuperación de información es la búsqueda de material de información (generalmente documentos) de naturaleza no estructurada (generalmente textos) que satisface una necesidad de información a partir de una gran colección de datos (generalmente en los servidores de computadoras locales o en Internet).*"[3]. Por otra parte, Salton en [4] refuerza la idea de que: *"...la recuperación de información se entiende mejor cuando uno recuerda que la información procesada son documentos…*". Otra definición que se utilizará en este trabajo dice que RI es "*...el proceso por el cual las demandas informativas y documentales del usuario son resueltas en un sistema de información, compuesto por un Corpus documental de volumen variable, cuyo tratamiento de indexación y almacenamiento hacen posible su estructuración, interrogación y representación, por medio del empleo de algoritmos matemáticos, estadísticos y semánticos…"* [5].

La construcción de un Corpus Lingüístico (CL) es una metodología de investigación que ha ganado protagonismo en el área de la RI [6 y 7]. Este marco metodológico permite el estudio empírico de grandes bases de datos lingüísticas fundadas en el uso real de la lengua, y la cuantificación del uso de la lengua utilizando métodos estadísticos. Por eso se lo puede concebir como un conjunto amplio y estructurado de ejemplos reales de uso de la lengua. Estos ejemplos pueden ser textos (los más comunes), o muestras orales (generalmente transcriptas).

La Universidad Nacional de La Matanza (UNLaM), a través del DIIT[2] desarrolló en los últimos años varios proyectos de investigación que se enmarcan dentro del Programa de investigación PROINCE (Programa de Incentivos a Docentes Investigadores SPU-ME). Para conseguir llevar adelante estos trabajos fue necesario construir diversos CL. Para los primeros dos trabajos de investigación se elaboró un Corpus con más de 12.000 documentos jurídicos [8].

Para el último proyecto, se construyó uno mediante el acceso a periódicos que se pueden encontrar en Internet de forma gratuita. Sobre la construcción de éste, es el desarrollo de este documento.

2. ¿Qué es un Corpus?

Según Joan Torruella y Joaquim Llisterri en [6], afirman que J. Sinclair, uno de los grandes especialistas en el campo de los Corpus modernos, define un Corpus como: "*…una colección de piezas de lenguaje que se seleccionan y ordenan de acuerdo con criterios lingüísticos explícitos con el fin de ser utilizados como una muestra del idioma…"*. Según esta definición, la informática no tiene que ver con el concepto de "*Corpus*", lo cual es válido, pero hoy en día la informática facilita tanto la organización y la explotación de grandes cantidades de datos que sería impensable crear un Corpus prescindiendo de este medio o herramienta. Por esto, hoy más que hablar de Corpus hay que hablar de Corpus Informatizado, ya que son dos conceptos íntimamente ligados. Así, según el mismo J. Sinclair, un Corpus Lingüístico Informatizado es: *"…un Corpus que está codificado de manera homogénea y estandarizada para tareas de recuperación abiertas. Sus partes constitutivas del lenguaje están documentadas en cuanto a sus orígenes y procedencia…"* [5].

Victoria López Sanjuán, en su trabajo sobre Corpus [9], enumera las ventajas que ofrecen los Corpus, de las que destaca:

- La exactitud del contenido, pues representan datos reales, no datos inventados.

[2] Departamento de Ingeniería e Investigaciones Tecnológicas.

- La facilidad de procesamiento, potencia y eficacia (sobre todo si se los compara con los estudios manuales).
- La obtención de afirmaciones más objetivas que las que se consiguen con la introspección.
- La aportación de evidencia objetiva y de un modo útil de analizarla. Son fuentes de información esenciales para áreas de la lingüística aplicada como la enseñanza de lenguas, traducción automática, reconocimiento del habla, correctores gramaticales o de estilo y un largo etcétera.
- Sirven de asistencia a hablantes no nativos de una lengua.

2.1. Corpus y los SRI

Un Corpus entonces, se puede ver como parte de un conjunto de documentos de texto, los cuales están compuestos por sucesiones de palabras que forman estructuras gramaticales (por ejemplo, oraciones y párrafos). Tales documentos están escritos en lenguaje natural y expresan ideas de su autor sobre un determinado tema. En los SRI para poder realizar operaciones sobre un Corpus, es necesario obtener primero una representación lógica de todos sus documentos, la cual puede consistir en un conjunto de términos, frases u otras unidades (sintácticas o semánticas) que permitan – de alguna manera – caracterizarlos [7].

Para poder ser usado en un SRI, a este Corpus original se le debe aplicar la técnica del Análisis de la Semántica Latente (ASL o LSA por sus siglas en inglés Latent Semantic Analysis), que funciona aproximadamente de la siguiente manera:

El LSA comienza procesando un texto de grandes dimensiones. Este texto contiene miles e, incluso, millones de párrafos (o frases). El Corpus se representa en una matriz cuyas filas contiene todos los términos distintos del Corpus (palabras) y las columnas representan una ventana contextual en la que aparecen esos términos (habitualmente párrafos). De este modo, la matriz contiene sencillamente el número de veces que cada término aparece en un documento. Sobre esta matriz de frecuencias se efectúa una ponderación, con el objetivo de restar importancia a las palabras excesivamente frecuentes y aumentarla a las palabras moderadamente infrecuentes. La razón de esta ponderación es sencilla: las palabras demasiado frecuentes no sirven para discriminar bien la información importante del párrafo y las moderadamente infrecuentes, sí.

El siguiente paso es someter a esta matriz ponderada a un algoritmo llamado Descomposición en Valores Singulares (DVS o SVD por sus siglas en inglés: Singular Value Decomposition). DVS es una técnica de reducción de dimensiones como es el análisis factorial. Este algoritmo se aplica con la idea de reducir el número de dimensiones de la matriz original en un número mucho más manejable (en torno a 300), sin que se pierda la información sustancial de la matriz original.

LSA es una herramienta eficiente para la evaluación de respuestas en los SRI y así lo han demostrado una gran cantidad de investigaciones. Para tener una compilación de estos trabajos se puede leer el escrito de Gabriel H. Tolosa y Fernando Bordignon de la Universidad Nacional de Luján [7].

3. El Corpus de Noticias en la Web.

El Corpus de Noticias en la Web (CONOWEB) nace, como ya se mencionó, dentro de un proyecto de investigación PROInce de la UNLaM, bajo el título: "*Uso de Minería de Datos para acelerar la recuperación de documentos*". Dentro de las tareas de este proyecto se encuentra la construcción de un Corpus voluminoso en español de textos de diversas temáticas, para ser utilizado en un SRI que emplee algoritmos K-Means y Redes Neuronales Artificiales para la recuperación de documentos dada una consulta determinada.

Como se sabe, la World Wide Web o Web, es una colección de documentos hipertextos vinculados entre sí, creando un mundo de información digital que involucra texto, imágenes y sonidos, constituyéndose en uno de los mayores acervos multimedia que integra las tecnologías de comunicación, transmisión de imágenes y sonidos, construyendo una verdadera red de difusión de conocimiento [6]. Estos documentos llamados "*páginas*", son archivos digitales con variados tamaños (número de caracteres) y presentan las siguientes características:

- *Un direccionamiento*. Conocido como Unifom Resourse Locator (URL), localiza el archivo en un equipo conectado a la red.
- *Un protocolo de transferencia*. El Protocolo de transferencia de hipertexto (http) que hace interconexión entre el equipo del usuario y la ubicación donde se encuentra la página (servidor o host)
- *Un lenguaje de marcado estándar*. El mismo estructura y define los componentes de las páginas web, como el Hipertext Markup Language (HTML).
- Utiliza un programa Navegador o Browser. Un tipo de software que permite la visualización de los contenidos que presenta una página web. Estos pueden recorrer una red de documentos vinculados e interpretar el lenguaje HTML para luego mostrarlo en la pantalla.

Las direcciones URL contienen varias partes. Por ejemplo *http://www.unlam.edu.ar*, la dirección tiene: la primera parte - a http: // - detalla qué protocolo de Internet utiliza. La segunda - la parte que generalmente tiene un "www" -, informa qué tipo de recurso de Internet está siendo conectado. La tercera parte - "unb" puede variar en longitud, e identifica el servidor de red a conectar. La parte final identifica un directorio específico en el servidor y una casilla página, documento u otro objeto de Internet.

La mayoría de las páginas están escritas en lenguaje *HTML*. Este lenguaje es una evolución del lenguaje SGML-Standard Generalized Markup Language, un lenguaje estandarizada que utilizó por primera vez las *"marcas", "tags" o "etiquetas*", un conjunto de códigos preestablecidos que definen componentes relacionados con la apariencia y la funcionalidad de la página, además de indicar el inicio y el final de la estructura que compone el documento.

Este lenguaje se construye de un número fijo de etiquetas que definen la apariencia de la página. HTML es un lenguaje muy simple y se puede crear utilizando cualquier editor de texto. Su simplicidad no la limita, pues la misma logra utilizar una gran cantidad de recursos como el uso de marcos (ventanas), y otros recursos multimedia. Una página HTML puede contener etiquetas que especifican direcciones URL de otras páginas, que constituyen los conocidos enlaces o link.

Un ejemplo de código básico HTML es el siguiente:

```
<html>
 <head>
   <title> Esto es un ejemplo </title>
 </head>
 <body>
  <div class="branding-data clearfix">
   <h1> Hola Mundo </h1>
  </div>
 </body>
 </html>
```

Los códigos de composición se acotan con unos caracteres especiales, que permiten diferenciarlos del texto del documento propiamente dicho. Estos símbolos son los corchetes angulares **<** y **>**. Estos códigos no distinguen entre mayúsculas y minúsculas aunque la normalización W3C indique que se debe usar minúsculas.

3.1. Marcadores o etiquetas HTML obligatorios.

Tabla 1. Algunos marcadores HTML.

Marcador	Descripción
<html> </html>	Todos y cada uno de los documentos HTML deben empezar y terminar con este marcador, que sirve para indicar que se trata de un documento HTML.
<!DOCTYPE html>	En html 5 es obligatoria en la primera línea.
<head> </head>	Los documentos HTML se dividen en dos partes, la cabecera y el cuerpo. Los navegadores Web necesitan distinguir entre ambas para poder interpretar correctamente los documentos. En la cabecera se incluye la información general sobre el documento.
<body> </body>	Incluye el contenido real del documento (body o cuerpo). Este marcador tiene también su marcador de terminación con la barra inclinada

Para llevar adelante esta construcción fue necesario contar con un lenguaje computacional para definir exactamente lo que se necesita buscar en la Web [10]. Las expresiones regulares (ER) son un recurso muy utilizado por gran cantidad de programas en los que se hace imprescindible la búsqueda y reemplazo de información.

El resultado más relevante esperado es que a través de las ER, la herramienta sea capaz de hacer una recuperación de la información de forma automática, o sea, que sin necesidad de teclear nuevamente, el algoritmo se encargue de abastecer al Corpus.

4. Expresiones Regulares

Las Expresiones Regulares son patrones utilizados para encontrar una determinada combinación de caracteres dentro de una cadena de texto. En algunos lenguajes como JavaScript, las expresiones regulares también son objetos. Estos patrones se utilizan en los métodos exec y test de RegExp, así como los métodos match, replace, search y split de String. En las siguientes tablas se pueden ver ejemplos de las formas abreviadas y completas y consultas de expresiones regulares [11].

Tabla 2. Formas abreviadas de ER.

Forma abreviada	Forma completa	Resultado
#filename	{prop name=filename} {regex}*.avi {/regex}	Cualquier archivo con extensión .avi (suelen ser archivos de sonido e imágenes en movimiento)
*.avi		

Nota

Si una consulta de propiedad contiene un asterisco (*), un signo de interrogación (?) o una barra vertical (|), se tratará de forma automática como una expresión regular, independientemente del modo que se indique.

Tabla 3. Ejemplos de consultas mediante ER.

Ejemplo	Consulta	Resultado
Grupo	**#filename = *|(ss|,ing|).cxx**	Devuelve todos los documentos con nombres de archivo que acaben en "ss" o "ing" y que tengan la extensión "cxx."
Operador **NOT** (^)	**#filename = |[^f]*.cxx**	Devuelve todos los documentos con nombres de archivo que empiecen por cualquier letra distinta de "f" y que tengan la extensión "cxx."
Operador de rango (–)	**#filename = |[a-c]*.cxx**	Devuelve todos los documentos con nombres de archivo que empiecen por una letra comprendida en el rango a-c y que tengan la extensión "cxx."
Coincidencia de número exacto	**#filename = *s|{2}|.cxx**	Devuelve todos los documentos con nombres de archivo que acaben exactamente con dos caracteres "s" y que tengan la extensión "cxx."
Una coincidencia como mínimo	**#filename = es|{1,|}.cxx**	Devuelve todos los documentos con nombres de archivo que sean combinaciones de la cadena "es" y que tengan la extensión "cxx."
Coincidencias entre dos variables determinadas	**#filename = ci|{2,4}.cxx**	Devuelve todos los documentos con nombres de archivo que sean combinaciones de dos a cuatro ocurrencias de la cadena "ci" y que tengan la extensión "cxx."
Cero o más coincidencias	**#filename = c|*ss.cxx**	Devuelve todos los documentos con nombres de archivo que empiecen con cero o más caracteres "c", acaben en "ss" y tengan la extensión "cxx.".
Cero o una coincidencia	**#filename = c|?ss.cxx**	Devuelve todos los documentos con nombres de archivo que empiecen por ningún o un carácter "c", que acaben en "ss" y que tengan la extensión "cxx.".

4.1. Operadores de Expresiones Regulares

Existen unas normas para el uso de los operadores de expresiones regulares:

- En las consultas, los asteriscos (*), puntos (.) y signos de interrogación (?) se comportan igual que en Windows. El asterisco (*) representa cualquier número de caracteres. El punto (.) representa el final de una cadena y el signo de interrogación (?), cualquier carácter individual.
- Cualquier carácter, excepto el asterisco (*), el punto (.), el signo de interrogación (?) y la barra vertical (|), coincide de manera predeterminada consigo mismo.

- Las expresiones regulares pueden incluirse entre comillas (" ") de apertura y cierre y deben incluirse entre comillas, si contienen un espacio o paréntesis de cierre ( ).

En la tabla 3 se observan los caracteres que adquieren un significado especial, si van precedidos de una barra vertical (|).

Tabla 4. Caracteres con significado especial (|).

Carácter	Descripción
(	Abre un grupo. Debe ir seguido del paréntesis de cierre correspondiente ( )).
)	Cierra un grupo. Debe ir precedido del paréntesis de apertura correspondiente (( ).
[	Abre una clase de carácter. Debe ir seguido del corchete de cierre correspondiente (]).
]	Cierra una clase de carácter. Debe ir precedido del corchete de apertura correspondiente ([).
{	Abre una correspondencia exacta. Debe ir seguido de la llave de cierre correspondiente (}).
}	Cierra una correspondencia exacta. Debe ir precedido de la llave de apertura correspondiente ({).
,	Separa cláusulas OR.
*	Corresponde a cero o más ocurrencias de la expresión precedente.
?	Corresponde a cero o una ocurrencia de la expresión precedente.
+	Corresponde a una o más ocurrencias de la expresión precedente.
Todos los demás	Coinciden consigo mismos.

Corchetes: Si se utilizan entre corchetes ([]), los siguientes caracteres adquieren un significado especial:

Tabla 5. Caracteres con significado especial ([ ]).

Carácter	Descripción
^	Debe ser el primer carácter. Coincide con todo excepto con las clases siguientes.
]	Cierra la clase. Sólo puede ir precedido de un símbolo de intercalación (^).
-	Operador de rango. Precedido y seguido por otros caracteres.
Todos los demás	Coinciden consigo mismos, inician o acaban un rango.

Llaves: Si se escriben entre llaves ({ }), los siguientes caracteres adquieren los significados que se indican a continuación.

Tabla 6. Caracteres con significado especial ({ })

Carácter	Descripción
M	Coincide exactamente con *m* ocurrencias de la expresión precedente (0< m < 256).
m,	Coincide como mínimo con *m* ocurrencias de la expresión precedente (1< m < 256).
m,n	Coincide con entre *m* y *n* ocurrencias de la expresión precedente, ambas inclusive (0 < m < 256, 0 < n < 256).

Para hacer que el asterisco (*), el punto (.) y el signo de interrogación (?) coincidan consigo mismos, se debe escribir entre corchetes. Por ejemplo, para buscar "hola?", se escribiría: hola[?] en la consulta.

5. Metodología.

Aunque hay muchos Corpora[3] disponibles tanto libremente como mediante algún pago, este grupo de investigación se propuso generar su propio Corpus a partir de la búsqueda de noticias en distintos portales gratuitos.

La metodología que presenta este trabajo consta de cuatro etapas principales a seguir:

1) Selección de los portales para la obtención del texto.
2) Compilación (o captura), manipulación, nombramiento de los archivos de textos.
3) La anotación.
4) Almacenado del texto.

5.1. Selección de portales para la obtención del texto.

Como se explicó anteriormente, este artículo comunica la experiencia de compilación de notas periodísticas obtenidas de importantes portales de noticias nacionales y extranjeros. Noticias que datan del año 2017 y 2018 y que contienen, hasta ahora (mes de julio del corriente año), más de 43.578 artículos con un total de 23.281.580 de palabras. En la Tabla 6 se indican los distintos medios visitados y sus respectivas direcciones. Estos medios componen el Corpus desarrollado:

Tabla 7. Medios periodísticos consultados.

Medio	URL
Clarin	https://www.clarin.com.ar
Infobae	https://www.infobae.com
El1	https://www.diariouno.com.ar
La Nación	https://www.lanacion.com.ar
La Tercera	https://www.latercera.com
La Prensa	https://www.laprensa.com.ar
Perfil	https://www.perfil.com
El País	https://www.elpais.com
El Observador	https://www.elobservador.com.uy

Primeramente, se obtienen de la página inicial (portada) todos los links de las noticias más recientes y destacadas de todas las categorías o clasificación que posea dicho portal. Por lo tanto se lee el código html del portal y se busca reconocer las etiquetas (tags) que contengan href seguidos de un "=" y una secuencia de caracteres encerrados entre comillas ("").

De esta forma se obtienen las direcciones tanto de las noticias como también cualquier referencia a las páginas que el Sitio contenga como publicidades o referencias a otros sub-sitios. Por eso se debe tener precaución y realizar un análisis de cómo se conforma la dirección de las noticias. En ocasiones, la dirección contiene algún rubro, número de noticia o una fecha y el nombre del título.

[3] Plural de Corpus.

5.2. La compilación

Dado que cada portal utiliza diferentes nomenclaturas para acceder a la noticia se puede observar que las direcciones son tratadas de diferentes maneras dentro de la portada principal para ingresar a las noticias, ya sea que estas se encuentren de forma relativa (utilizando la dirección actual y re-direcciona a la sección que se indica) o directa (se da un link completo).

Pudiendo entender el comportamiento de las direcciones, se almacenan en una lista para luego ser accedidas y extraer la información necesaria (Se eliminan las imágenes y los epígrafes en caso de que se encuentre en la noticia). Cada portal utiliza diferentes formas de etiquetar cada uno de los elementos de la noticia. Por lo cual, hay que obtener el código html y analizar los patrones de las etiquetas que identifican el título, día de creación, clasificación, copete, cuerpo y etiqueta relacionados de las noticias. En la creación de los documentos se usan una series de tags para clasificar cada parte de la noticia, para ello se emplea el mismo sistema de html:

<nombre identificatorio>

...

</nombre identificatorio>

Las etiquetas que se usarán son para identificar el periódico, título, día de obtención, día el cual se generó el documento realizado por el sistema, clasificación que le otorga el portal a la noticia, copete, cuerpo y los tags (palabras claves) relacionados con la noticia. Sintácticamente poseerán la siguiente estructura:

Tabla 8. Caracteres con significado especial ({ })

Carácter	Descripción
<periodico></periodico>	nombre del periódico
<titulo></titulo>	título de la noticia
<url></url>	dirección del artículo.
<dia></dia>	día de emisión del artículo
<diaob> </diaob>	día de obtención por el sistema
<clasificacion> </clasificacion>	otorgada por el portal
<copete> </copete>	copete de la noticia
<cuerpo> </cuerpo>	el contenido del artículo
<tag> </tag>	palabras claves

5.3. La anotación

En primera instancia se creará un documento de texto que poseerá el nombre del artículo, de esta manera el sistema podrá verificar si el artículo ya está generado, así no se vuelva a crear o pisar.

Con el ingreso ya establecido al diario, a continuación, se inserta la etiqueta <periódico>. Posteriormente la etiqueta de <titulo>. Los títulos se corresponden con la etiqueta "h1", a veces acompañada de una clase que le da un formato determinado.

Para ejemplificar, se toma el título de una de las noticias: **Cuáles son las fortunas familiares más grandes del mundo.** En la siguiente imagen se puede observar el código html respectivo.

```
<h1>
  Cuáles son las fortunas familiares más grandes del mundo
</h1>
```

Imagen 1. Código en html del tag <h1></h1>

En lo que respecta a la etiqueta <url> se lo completa con el link de la dirección al cual pertenece el artículo seleccionado, <dia> y <diaob>. Ambas se componen con el formato de "dd de (mes correspondiente) de aaaa". En la imagen 2 se tiene el código correspondiente. Se debe mencionar que el formato de los días de publicación varía según el portal. En ocasiones se puede encontrar la notación que se ha utilizado y en otros donde se utiliza el formato dd/mm/aaaa hasta aaaa/mm/dd, por lo cual deben ser convertidos a la notación ya mencionada.

```
<span class="byline-date" style="border-left:none;">26 de julio de 2018</span>
```

Imagen 2. Código en html de la fecha

El copete se lo puede encontrar mediante la etiqueta "h2", "span" y "p", al igual que el título, una clase que le da un formato determinado a esto se le adiciona la etiqueta de <copete>. Imagen 3.

```
<span class="subheadline">
  Un grupo formado por 25 clanes familiares que controlan USD 1,1 millón de millones de riqueza, de acuerdo a un ranking elaborado por la agencia Bloomberg
</span>
```

Imagen 3. Código en html del copete

En el caso del cuerpo, el mismo puede ser identificado con la etiqueta de "span" y "p" y un formato en caso de que este tuviera uno. Además, en la extracción de la noticia, algunos portales, contienen palabras o frases que se encuentran resaltadas (utilizando el formato de negrita, cursiva, subrayado o en negrita), como así también palabras que contienen una dirección a otra noticia por medio de una palabra clave o frases, incluso párrafos con un fondo de otro color mediante las etiquetas "strong", "span" o "a" y una clase según la función que se busca realizar.

El cuerpo podía encontrarse seccionado, lo que lleva a unificar cada parte la noticia. En dichas secciones se podían encontrar imágenes con un epígrafe, llamadas a otras noticias que podrían estar relacionadas. Esto se da mayoritariamente cuando se trata un mismo tema varios días consecutivos o es un tema derivado de una situación o problema ocurrente. A lo obtenido se lo guarda bajo la etiqueta <cuerpo>. Por último, se buscan los tags con los nombres claves "Tag" o "Keywords" dentro del html. Estas son extraídas y etiquetadas con <tag>.

Para poder realizar todos estos procedimientos se utilizó Visual Studio 2015 (entorno de desarrollo que proporciona herramientas avanzadas para generar aplicaciones en cualquier tipo de arquitectura). Se programó en C#, lenguaje de programación orientado a objetos desarrollado y estandarizado por Microsoft como parte de su plataforma .NET, que después fue aprobado como un estándar por la ECMA (ECMA-334) e ISO (ISO/IEC 23270). C# fue diseñado para la infraestructura de lenguaje común. Su sintaxis básica deriva de C/C++ y utiliza el modelo de objetos de la plataforma .NET, similar al de Java, aunque incluye mejoras derivadas de otros lenguajes. Las expresiones regulares son compatibles con la mayoría de los lenguajes de programación, por ejemplo, C#, Java, Perl, etc.

Desafortunadamente, cada idioma admite expresiones regulares ligeramente diferentes. Para obtener una breve introducción a este tema se puede consultar el portal de Microsoft[4]. En la siguiente imagen se puede ver parte del código de como se lo extrae de un medio determinado.

```
var url = "http://www.infobae.com";
var textFromFile = (new WebClient()).DownloadString(url);

if (!Directory.Exists(@".\noticias\infobae"))
  Directory.CreateDirectory(@".\noticias\infobae");
  Regex regex = new Regex("href=\"/\\S+\"");
  MatchCollection matches = regex.Matches(textFromFile);
  List<string> subsitios = new List<string>();
  foreach (Match texto in matches)
    subsitios.Add(texto.Value.Replace("href=","").Replace("\"",""));
    subsitios = subsitios.Distinct().ToList();
    foreach (string subdireccion in subsitios) {
      string[] cut = subdireccion.Split('/');
      if (cut.Length > 3 && (cut[1] == "economia" || cut[1] == "tecno" || & _
          cut[1] == "america" || cut[1] == "sociedad" || & _
          cut[1] == "deportes - 2" ||   cut[1] == "politica")) {
          var urlart = url + subdireccion;
          try {
          textFromFile = (new WebClient() {
             Encoding = System.Text.Encoding.UTF8 }).DownloadString(urlart);
                    }
                    catch (Exception)
                    {
                        continue;
                    }

        //obtener el titulo
        //regex de titulo   (?<=< h1 >).* (?=<\/ h1 >)
        regex = new Regex("(?<=<h1>).*(?=<\\/h1>)");
        matches = regex.Matches(textFromFile);
        string nombrearchivo = "";
        regex = new Regex("[\",.:¿?;!;\\/\\|]");
        nombrearchivo = regex.Replace(matches[0].Value, "") + ".txt";
        if (File.Exists(@".\tag\infobae\" + nombrearchivo))
            continue;
          if ((Directory.GetCurrentDirectory() + nombrearchivo).Length > 200)
              nombrearchivo = nombrearchivo.Substring(0, 197 -
              Directory.GetCurrentDirectory().Length) + ".txt";
```

Imagen 4. Fragmente del código en C# utilizado

5.4. Almacenado del texto

Una vez esquematizado y obtenida la información necesaria de cada uno de los artículos que fueron procesados y completados serán almacenados en un repositorio identificando a que diario que pertenece. En la imagen 5 se puede observar la estructura del mismo.

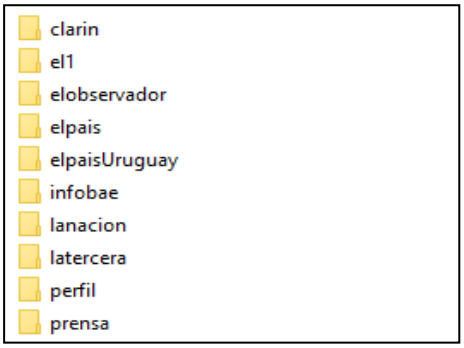

Imagen 5. Estructura de directorio.

Cada noticia será almacenada en un formato de texto plano. Como nombre identificativo será su título. El nombre de este no será superior a los 200 caracteres incluyendo su extensión.

Cada artículo se guardará de esa manera dado que, al momento de realizar un escaneo por un portal determinado, puede ocurrir que se haya tratado con ese artículo. Por lo tanto, no será tomado en cuenta para ser almacenado.

[4] https://docs.microsoft.com/es-es/dotnet/standard/base-types/regular-expression-language-quick-reference

6. Otros recursos relacionados con los Corpus

Existen en la web múltiples recursos relacionados con esta temática. Estos van desde corpus ya construidos hasta herramientas informáticas para incorporar información lingüística a textos. Para obtener Corpus ya compilados en multitud de lenguas cabe destacar estas dos organizaciones [12]:

- ✓ ELRA: http://www.elra.info
- ✓ LDC: http://www.ldc.upenn.edu/

También existen un cierto número de herramientas informáticas de uso libre que permiten añadir información lingüística a textos (p. ej. Información morfológica). Cabe señalar, sin embargo, que muchas de estas herramientas están diseñadas básicamente para el inglés. Su aplicación a otras lenguas es posible, aunque en la mayoría de casos, con un cierto esfuerzo. También hay que destacar que el uso de estas herramientas generalmente requiere de cierta destreza y familiaridad
con el uso de herramientas informáticas. Los recursos más interesantes y con funcionalidades muy diversas son los siguientes:

1. GATE (http://gate.ac.uk/): entorno para el desarrollo de herramientas para el procesamiento del lenguaje.
2. APACHE UIMA: http://incubator.apache.org/uima: herramientas y entorno para desarrollo para el PLN (Procesamiento del Lenguaje Natural).
3. NLTK (http://www.nltk.org/): herramienta concebida para la enseñanza de técnicas de PLN. Incluye también recursos (Corpus, diccionarios, etc.)
4. BYU (Brigham Young University). Creado por Mark Davies, Recursos basados en Corpus. (http://www.Corpusdelespanol.org/)
5. RAE. Real Academia Española. http://www.rae.es/ recursos/ banco-de-datos/corpes-xxi). Catálogo sobre recursos, proyectos, organizaciones etc. relacionados con tecnologías de la lengua.
6. Inicios. (https://inicios.es/Corpus): Listado categorizado de Corpus lingüísticos de español e inglés. Aparece en cada colección en caso de conocerse.

7. Conclusión

Este texto busca mostrar los procedimientos adoptados durante la etapa de desarrollo del CONOWEB. El fin último es explicar a investigadores interesados en la construcción de un corpus lingüístico mediante ER, para ser utilizado en un SRI. Se espera, con ello, colaborar con la propagación de la Lingüística de Corpus como metodología para investigaciones en RI. Como conclusión se desea que esta metodología empleada colabore en la relación interdisciplinaria entre la Lingüística y la Computación, ya que esta última, ha demostrado ser cada vez más necesaria para la búsqueda y recuperación de documentos.

El desarrollo del CONOWEB se hizo a partir de la colaboración entre varios investigadores y docentes de variadas disciplinas en computación. Se les asignó a cada uno de ellos desde la programación de la aplicación y el levantamiento del Corpus, incluyendo el establecimiento de criterios para selección de leyendas y su categorización, formateo y etiquetado.

El propósito del Corpus, aunque no ha sido el foco de este texto, es el uso del mismo en el contexto de un SRI, sirviendo como recurso para la obtención de documentos

utilizando técnicas de Minería de Datos. Este Corpus pretende ser usado de consulta para que profesores y alumnos puedan obtener ejemplos de uso de estructuras lingüísticas, en las distintas carreras del DIIT.

Se pretende a futuro, realizar mejoras tanto para el corpus como para la herramienta, con la inserción de nuevos recursos. Se tiene pensado recopilar nuevos portales de noticias para integrar el Corpus, aumentando su número total de palabras. También se hará una revisión cualitativa de los subtítulos en busca de la corrección de posibles errores de escritura, ortografía o de formato, que pueden perjudicar el funcionamiento del CONOWEB o limitar su uso.

8. Referencias

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

[2] Seco Naveiras, Diego. Técnicas de indexación y recuperación de documentos utilizando referencias geográficas y textuales: Univ. de la Coruña. (2009). Disponible en: <https://citius.usc.es/sites/default/files/ tesis/Tese_DiegoSeco.pdf>. Citado el: 29/07/2018.

[3] Armando Plasencia Salgueiro, A; Ballagas Flores, B .Análisis comparativo de herramientas de recuperación y análisis de información de acceso libre desde una concepción docente. Disponible en: < http://www.scielo.br/ pdf/tinf/v26n3/0103-3786-tinf-26-03-00315.pdf>. Citado el: 29/07/2018.

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

[5] Blázquez Ochando, M. Técnicas avanzadas de recuperación de información. Universidad Complutense de Madrid. (2012). Disponible en: <http://ccdoc-tecnicasrecuperacioninformacion.blogspot.com/2012/02/conceptos-basicos-en-recuperacion-de.html> Citado el: 29/07/2018.

[6] Andrade Figuereido, D. Recuperação da informação: uma análise sobre os sistemas de busca da web. U. de Brasília. (2006) Disponible en: < http://www2.senado.leg.br/bdsf/bitstream/handle/id/70270/Monografia_.pdf> Citado el: 29/07/2018.

[7] Tolosa, G. y Bordignon, R. Introducción a la Recuperación de Información Conceptos, modelos y algoritmos básicos. U. N. de Luján, Argentina. (2007) Disponible en: <http://eprints.rclis.org/12243/1/ Introduccion- RI-v9f.pdf> Citado el: 29/07/2018.

[8] Resúmenes Extendidos 2013. UNLaM. ISBN: 978-987-3806-30-8.

[9] López Sanjuán, V. Integración de los Corpus como herramienta de apoyo en la enseñanza de ESP. Madrid, (2008). Disponible en < http://www.ugr.es/~portalin/ articulos/PL_numero10/9%20Victoria%20Lopez.pdf> ISSN: 1697-7467.Citado el: 29/07/2018.

[10] Avendaño Pérez, C. Diseño e Implementación de un Sistema de Búsqueda en una Colección de Texto Comprimido. Tonantzintla, Puebla. (2004). Disponible en:<https://www.dcc.uchile.cl/~gnavarro/mem/ algoritmos/tesisAvendano.pdf>. Citado el: 29/07/2018.

[11] MDN Web Docs. Disponible en: <https://developer.mozilla.org/es/docs/Web/JavaScript/Guide/Regular_Expressions>. Citado el: 29/07/2018.

[12] Vivaldi, J. Catálogo de herramientas informáticas relacionadas con la creación, gestión y explotación de Corpus textuales. Tradumàtica: traducció i tecnologies de la informació i la comunicació [en línia], 2009,, Núm. 7 . Disponible en: <https://www.raco.cat/index.php/ Tradumatica/article/view/154837/206731>. Citado el: 29/07/2018.

c) CACIC-2018.

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 wij 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.

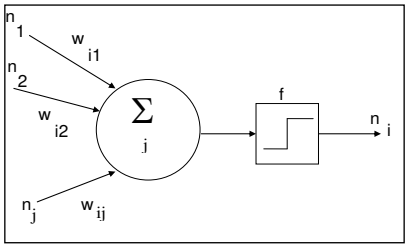

Figura 1. Modelo de neurona de McCulloch y Pitts

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 PROINCE5 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 LSI6). 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 sinonimia7 y a la polisemia8. [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 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. Constructos9 aportes de expresiones regulares. Se realizó una comparación de subcadenas,

[5] Programa de Incentivos a Docentes Investigadores SPU-ME
[6] Por sus siglas en inglés, Latent Semantic Indexing.
[7] La sinonimia es una relación semántica de identidad o semejanza de significados entre determinadas expresiones o palabras.
[8] Una palabra polisémica es aquella que tiene dos o más significados que se relacionan entre sí.
9 Un constructo es una construcción teórica que se desarrolla para resolver un cierto problema científico

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:

uno que particione el corpus utilizando una noción de vecindad o similitud y

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:

aplicar el algoritmo que direcciona la consulta hacia una de las partes, para

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 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:

Exp.	0	1	2	3	4	5	6	7	8	9	10	≥ 7
0	0	1	1	4	10	11	12	15	12	18	1	16	61
1	0	0	0	5	10	12	10	15	18	11	1	19	63
2	0	0	4	4	7	12	18	18	14	10	1	13	55
3	0	3	1	2	6	7	14	12	18	17	1	20	67

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

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:

Exp.	0	1	2	3	4	5	6	7	8	9	10	≥ 7
4	0	3	6	161	39	45	77	69	80	123	177	469

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

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

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	98	65	55	42	32	24	24	22	26	15	21	13	433

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

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

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

[14] 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.

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

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

[17] 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.

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

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

[20] 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

[21] 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

[22] 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

[23] 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

[24] 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.

[25] 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)"

d) **ECEFI-2018.**

Recuperación de Información acelerada con Algoritmos de Minería de Datos.

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

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, mjcasuscelli@gmail.com, jbossero@unlam.edu.ar, lmatteo@unlam.edu.ar, hugor@unlam.edu.ar

Resumen: Para acelerar la respuesta inicial en sistemas de recuperación de información sobre depósitos documentales privados de mediano tamaño fue estudiada la posibilidad de segmentar el mismo y elaborar la respuesta examinando sólo un segmento. En este trabajo se intenta aminorar la pérdida de calidad que ello provocaba.

Palabras claves: Recuperación de información, particionado, recubrimiento, K-means. Minería de datos.

Introducción

La recuperación de información es una técnica de la que se disfruta cuando realizan búsquedas en Internet y sería pretencioso querer introducir mejoras a los buscadores más famosos. No obstante eso, existen depósitos documentales 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 aumenta ya que, si no se ha previsto estructuras complejas, 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 depósito. Una solución es aumentar la potencia de cómputo, pero esto no está al alcance de todos.

En un trabajo previo [4] se analizó la posibilidad de fraccionar el depósito 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.

Tratándose de poblaciones grandes y variadas, tanto en los documentos como en las consultas, ambos imposibles de describir con un patrón regular, evaluar esta propuesta será inevitablemente de un modo estadístico. La tarea aquí comenzada es aplicable a otras situaciones de apareo de objetos.

En la siguiente sección se describe someramente los principios de las tecnologías involucradas, limitado a lo efectivamente utilizado. En la otra sección las tareas de otros investigadores así como algunas pocas ideas afines a lo realizado. Finalmente se expone en detalle las ideas aportadas y su concreción en código. La última sección detalla las experiencias numéricas y el modo de juzgarlas.

Marco Teórico

La recuperación de información (RI o IR por sus siglas en inglés Information Retrieval) es un área de investigación que se inició con los trabajos de Salton en los años 60 [1] y recientemente ha experimentado un desarrollo espectacular motivado por la competencia de los buscadores en Internet y el consiguiente deseo de aplicarlo a repositorios privados. La RI se distingue de otras técnicas de acceso en que recupera archivos por su contenido y no por sus metadatos [2].

Muchos modelos han sido propuestos y usados para representar documentos, entre ellos, el modelo vectorial [3]. Representa las consultas y documentos como vectores y la recuperación de información se hace en función de operaciones entre vectores. El modelo vectorial fue definido por Salton en [1], y es ampliamente usado en operaciones de RI, así como también en operaciones de categorización automática, filtrado de documentos, etc. En este modelo se intenta recoger la relación de cada documento Di, de una colección de N documentos, con el conjunto de las m características de la colección. Formalmente a un documento puede asociarse un vector que expresa la relación del documento con cada una de esas características. Una función del indexado asocia a cada documento Di del corpus D un vector de m características: cik .Ese vector columna Ci identifica en qué grado el documento Di satisface cada una de las m características. En ese vector, cik es un valor numérico que expresa en qué grado el documento Di posee la característica k. El concepto "característica" suele concretarse en la ocurrencia de determinadas palabras o términos en el documento, aunque nada impide tomar en consideración otros aspectos.

La "Indexación Semántica Latente" (LSI por las siglas en inglés de Latent Semantic Indexing) [3] es otro modelo. Parte del modelo vectorial, tratando de superar los problemas de sinonimia10 y a la polisemia11 describe la matriz vectorial en valores propios. Finalmente observa que con los términos de mayor peso logra un buen resultado. LSI ha sido efectivamente usado en RI de la web y en otras aplicaciones

[10] La sinonimia es una relación semántica de identidad o semejanza de significados entre determinadas expresiones o palabras.
[11] Una palabra polisémica es aquella que tiene dos o más significados que se relacionan entre sí.

La Minería de Datos

La minería de datos (MD) es una disciplina ya bien establecida desarrollada en numerosos libros [6]. El objetivo principal de la minería de datos 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.

La similitud puede medirse a través de funciones de distancia, las cuales juegan un papel crucial, ya que individuos cercanos deberían ir para el mismo grupo. Se agrupan los objetos de acuerdo a todas las variables y por ello, una variable irrelevante puede generar ruido en los resultados obtenidos.

Clasificación: Este tipo de tarea sirve para predecir la categoría a la que pertenece un objeto dado. Un conjunto de datos están etiquetados si todos o la mayoría de sus registros contienen un atributo especial llamado atributo de clase o etiqueta [6]. El objetivo de este tipo de atributo es clasificar la muestra para que pertenezca a una categoría. Esta asociación se realiza con base en las características o propiedades de los objetos mediante un patrón frecuente.

La clasificación funciona mediante la descripción del manejo de datos de alguna área determinada. Llamados conjuntos de datos, y los elementos que los componen son conocidos como objetos, registros o muestras. A esta parte se le conoce como fase de entrenamiento, donde el clasificador mediante un análisis "aprenderá" de un conjunto de muestras ya clasificadas y los atributos asociados de la clase.

Una tupla para el aprendizaje de la técnica se le conoce comúnmente como muestras de entrenamiento. Una muestra, X, está representada por un vector de atributo de n-dimensiones, X = (x1, x2, ..., xn), que representa n mediciones hechas sobre los datos de la muestra. Cada atributo representa una característica de X.

Algoritmos Utilizados

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

Algoritmo K-Means. (En español debiera llamarse K-Medias), presentado por MacQueen en 1967, es uno de los algoritmos desarrollados para resolver el problema del agrupamiento. La idea del algoritmo es proporcionar una clasificación de la información de acuerdo con los propios datos, basada en 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.

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. Luego cada iteración realiza dos pasos:

c) Para cada objeto a particionar se determina el centroide más cercano y así se lo etiqueta.
d) Se reposiciona cada centroide utilizando todos los objetos que fueron etiquetados para un mismo centro.

Es usual que esta actividad converja y tanto los desplazamientos que sufren los centroides como los cambios de etiqueta sean cada vez menores. La teoría del método demuestra que la suma de las distancias al cuadrado de los objetos etiquetados a sus respectivos 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. Es un problema de múltiples mínimos locales.

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.

En un sistema neuronal artificial se establece una estructura jerárquica, 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.

El modelo ha evolucionado. Para facilitar el entrenamiento es bueno sustituir la función escalón o de disparo 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.

Antecedentes y Trabajos Relacionados

Este trabajo continúa con la línea de investigación de dos proyectos PROINCE[12] orientado el: primero al estudio del tema y posteriormente a la realización de un prototipo de Sistema de recuperación de la Información, aplicando la metodología del *ISL*. Varios autores de este trabajo, participaron en el otro proyecto dedicado a la predicción de perfiles de alumnos en riesgo de deserción. Aquí se utilizó Minería de Datos y se hizo un análisis comparativo de modelos de clasificación. De la experiencia conjunta surgió: "Uso de Minería de Datos para acelerar la recuperación de documentos".

Se localizaron trabajos afines: Clusterdoc [5], 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. Por último el trabajo de Augusto Cortez Vásquez [6] que aborda el problema de una aplicación usando el algoritmo supervisado: máquinas de soporte vectorial (MVS) en el área de RI.

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 o una estima de su adecuación al requerimiento formulado. Sólo el examen de los documentos por parte del requeridor confirma el mayor o menor acierto del sistema.

Sistemas orientados a volúmenes moderados no quieren programar complejas estructuras de datos que respondan a un ordenamiento por vecindad sino que enfrentan las consultas con la totalidad del corpus, esperando rescatar los más cercanos. Resuelto así, el tiempo es proporcional al tamaño. Para reducir ese tiempo se puede recurrir a un aumento de potencia de cómputo, especialmente usando el paralelismo.

La investigación comenzó con una línea alternativa, fraccionar el corpus, lo que requirió dos algoritmos preparatorios:

c) uno que particione el corpus utilizando una noción de vecindad o similitud y
d) 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 falta de otras sugerencias. 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. Los primeros resultados en esta línea fueron alentadores y publicados [4].

Sin embargo se decidió analizar modificaciones que pudieran mejorar el porcentaje de consultas con respuesta satisfactoria y el precio a ello asociado, ya sea en preprocesos del corpus como en un menor beneficio en tiempo para la consulta.

Una explicación intuitiva del porque no se alcanzó el resultado ideal es que algunas consultas puedan caer cerca de la frontera de dos o más particiones. De modo tal que fuera inevitable que se vieran sólo una parte de los vecinos. El azar de los documentos puede hacer inclusive que del otro lado de la frontera haya más documentos que del lado consultado.

Dos formas de atacar esto fueron diseñadas: (a) disponer de dos o más particionados, con la esperanza que lo que fuera frontera de partes uno fuera interior para otro; (b) pensar en recubrimientos más que en particionados.

a) **Múltiples particionados.** Las implantaciones de K-means son sensibles al orden en que reciben los datos. De modo tal que cambiando convenientemente el orden de los vectores representativos de un corpus se puede conseguir diversas propuestas de particionado. Al proponer en que parte de que particionado se debe realizar la

[12] Programa de Incentivos a Docentes Investigadores SPU-MEdde

búsqueda se puede optar ya sea la que provea la señal de red más elevada o la que tenga más cercano su centroide. Trae como contrapartida un mayor tiempo de preproceso del todo el corpus antes de dejarlo operativo y una mayor fragmentación en el almacenamiento del corpus, el cual podría verse fraccionado en forma exponencial siguiendo la distribución de los elementos en los distintos particionados. U, organizar múltiples listas vinculadas sobre los vectores del corpus o listas invertidas ordenados por el lugar físico que ocupe el vector descriptivo, etc.

b) **Recubrimientos.** Un recubrimiento es un conjunto de subconjuntos que tiene la propiedad de que la unión de estos sea el todo. Renuncia a que sus partes sean disyuntas. En promedio estos subconjuntos serán más grandes de lo que hubieran sido en un particionado y por lo tanto se pierde parte de la ventaja de reducir el volumen total del corpus a sólo una parte. Una de las formas más sencillas de definir recubrimientos con el debido control de su crecimiento es definir primero un particionado y luego extender esto a elementos vecinos. Dos formas fueron pensadas de tener un crecimiento moderado y controlado de las partes: (b1) Una vez terminado el particionado se puede recorrer los documentos y observar las K distancias a los centroides. La menor distancia debiera coincidir con la partición a la cual quedó asignado. Usando esta menor distancia como referencia se puede decidir que aquellas distancias que superan a ésta en un bajo porcentaje puedan producir una pertenencia adicional del documento a otra partición. Esto requiere un preproceso del corpus hasta dejarlo operativo y una fragmentación del corpus en más partes. La segunda técnica incorpora a los más cercanos a los documentos de la partición si aún no estuvieran. Conceptualmente el modo de lograrlo es utilizar los mismos documentos como consulta, la cual ordenará a todos los documentos por su afinidad al analizado y decidir que cierto número de ellos, tomados desde la cabeza de la lista sean forzados, si aún no lo estuvieran a integrar la partición ampliada. Hacerlo eficientemente es complejo para no enfrentar a cada documento con todos los documentos. Siendo un preproceso no afecta a la velocidad de las consultas reales.

Evaluación de los resultados obtenidos

Métricas para evaluar: Es necesario introducir alguna métrica que permita apreciar la calidad del resultado obtenido y ver la incidencia de los parámetros ajustables en ello. Se desarrollaron 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 otros documentos que compiten por lo mismo, estos a lo más quedarán intercalados entre ambos.

Una primera métrica, pensando en la aplicación a búsqueda documental analiza el escenario de uso del sistema, el usuario ingresa la consulta y obtiene una página de sugerencias de lectura, típicamente con 10 documentos. 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 de la primera página hubieran provenido de la partición consultada, número entero que estará entre 0 y 10". 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.

Para tener un número práctico de decisión y no tener que comparar largas estadísticas, ambos fueron condensados en un único valor, para el primero: ""En cuántos casos se muestran en la primera página al menos 7 documentos de la solución global". Para el segundo: "En cuántos casos el décimo elemento mostrado no pasa de la segunda página en la solución global". Finalmente para tener una métrica global sensible a cualquier cambio de posición de los documentos atribuimos a cada documento mostrado el cociente entre el lugar que ocupa al procesar la parte y el lugar que ocuparía de procesar el todo y estas fracciones las promediamos.

particionado	Distribuión detallada de la métrica 1											Métrica 1 abreviada: 7 o más		Métrica 2 abreviada: pag. 1y2		Métrica 3 global
	0	1	2	3	4	5	6	7	8	9	10					
0	0	6	7	31	52	69	68	89	80	121	177	467	66,71%	579	82,71%	0,4826
1	1	0	16	28	45	52	79	90	100	92	197	479	68,43%	592	84,57%	0,4829
2	1	2	12	28	35	65	77	90	108	100	182	480	68,57%	604	86,29%	0,4864
3	1	3	7	30	51	72	77	96	74	108	181	459	65,57%	589	84,14%	0,4825
4	1	4	9	26	50	76	54	87	75	97	221	480	68,57%	589	84,14%	0,4918
5	0	2	17	35	43	64	81	86	90	86	196	458	65,43%	588	84,00%	0,4779
6	1	2	9	28	34	64	80	93	80	104	205	482	68,86%	602	86,00%	0,4937
7	1	4	8	26	56	67	58	81	87	102	210	480	68,57%	592	84,57%	0,4916
0 y 1	0	0	2	17	25	37	53	83	101	126	256	566	80,86%	651	93,00%	0,5289
0, 1 y 2	0	0	1	6	12	28	52	63	107	128	303	601	85,86%	676	96,57%	0,5546
0 a 3	0	0	1	4	13	21	36	49	91	118	367	625	89,29%	679	97,00%	0,5798
0 a 4	0	0	1	3	10	11	28	41	80	122	404	647	92,43%	684	97,71%	0,5947
0 a 5	0	0	0	4	10	9	26	40	73	116	422	651	93,00%	685	97,86%	0,6061
0 a 6	0	0	1	4	9	10	26	39	73	113	425	650	92,86%	683	97,57%	0,6072
0 a 7	0	0	1	4	9	10	26	36	73	114	427	650	92,86%	683	97,57%	0,6077

Tabla 1. Resultados obtenidos por múltiples particionados

Experimentos realizados Se hicieron simulaciones numéricas con córpora aleatorios de tamaños crecientes tanto en cantidad de vectores como en el largo de los mismos, siempre particionando en 4, resultados que fueron publicados en [4].

Aprovechando que K-means no entrega un único resultado, sino que este es dependiente del orden de envío de los datos sobre un pseudo-corpus se realizaron 8 particionados, cada uno evaluado por separado (ver tabla 1), en ella se tabula primero su métrica 1 detallada, los resultados son similares lo que ya fue comprobado en [4] cambiando los parámetros de las simulaciones. Luego se abre la posibilidad de elegir el particionado y parte basándose en la mayor cercanía a los centroides disponibles. Claramente se obtiene mejoras considerables expresadas por cualquiera de las métricas aplicadas.

Las dos técnicas para extender las particiones y construir recubrimientos fueron simuladas para visualizar la disminución de la aceleración pretendida. (Ver figura 2) La primera, basada en distancia de centroides es sencilla de programar y alcanza niveles

altos de calidad a costo de sacrificar el beneficio de explorar partes pequeñas, Si se conformara en reducir el volumen a examinar en medio cuerpo en lugar del cuarto inicial, habría que detenerse en 0,18 de tolerancia.

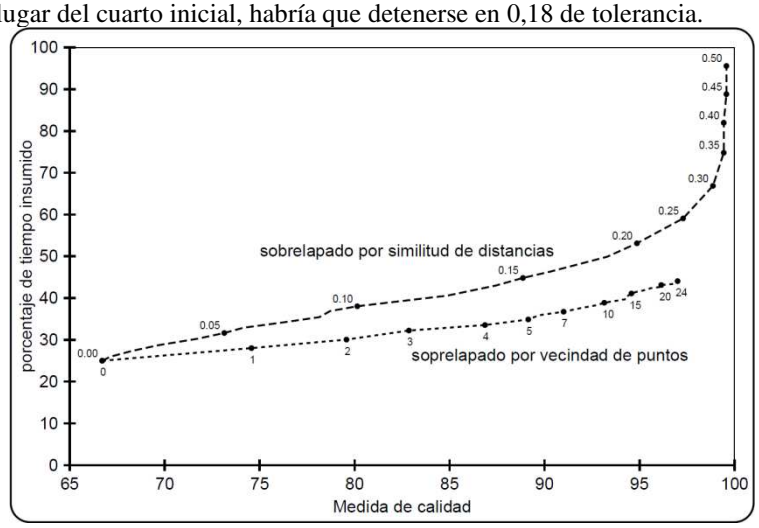

Figura 2. Resultados obtenidos en las dos técnicas de extensión

La segunda, basada en incorporar los n documentos más cercanos a uno dado, si es que no estuviera ya en la partición supera a la otra, para igual calidad, en la magnitud de los crecimientos en volumen de los recubrimientos y con la consiguiente ganancia en velocidad de respuesta.

Conclusiones y futuros trabajos

Los números obtenidos en las simulaciones son promisorios, lo que incentiva seguir investigando para obtener porcentajes aún mejores.

Se destacan algunas ideas que debieran contribuir a lograrlo:

- Decidir por cada consulta la conveniencia de explorar o no los documentos de la franja marginal incorporada.
- 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.

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] Spositto, O. y otros: Aceleración en la Recuperación de Información utilizando Algoritmos de Minería de Datos de R, CACIC 2018 Tandil.

[5] 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

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

Firmantes