Administración

Desarrollo de una aplicación destinada a la clasificación de información textual y su evaluación por simulación

Cristal Karina Galindo Durán[1]
Mihaela Juganaru-Mathieu[2]
Carlos Áviles Cruz[3]
Héctor Javier Vázquez[4]

Resumen

En el presente trabajo se propone una aplicación de cómputo destinada a la clasificación de documentos textuales, basada en el algoritmo de los K vecinos más próximos. Después de presentar brevemente el diseño y las funcionalidades de la aplicación, se presentan los resultados de la simulación de pruebas de clasificación de documentos.

[1] Maestría en Ciencias de la Computación, División de Ciencias Básicas e Ingeniería (DCBI), Universidad Autónoma Metropolitana, Unidad Azcapotzalco (UAM-A). Ave. San Pablo 180, Col. Reynosa Tamaulipas, México D.F., C.P. 02200, Tel. 55-26-72-66-76, cdgalindod@gmail.com
[2] Laboratoire en Sciences et Technologies de l'Information; Institut H. Fayol, École Nationale Supérieure des Mines de St Étienne, 42023 ST ÉTIENNE Cedex 2, France, mathieu@emse.fr
[3] Depto. de Electrónica,DCBI, UAM-A. Ave. San Pablo 180, Col. Reynosa Tamaulipas, México D.F., C.P., 02200, Tel. 5318-9550 (ext 1026), Fax 5394 6843, caviles@correo.azc.uam.mx ,
[4] Depto. de Sistemas, DCBI, UAM-A. Ave. San Pablo 180, Col. Reynosa Tamaulipas, México D.F., C.P. 02200, Tel. 5318-9532 (ext 109), Fax 5394 4534, hjv@correo.azc.uam.mx

Abstract

In the present work a software application is proposed to perform classification of textual documents, based on the use of the K Nearest Neighbors. The design and the functionalities of the software application are presented as well as the results of simulation tests to classify documents.

Palabras clave: clasificación, información textual, K-vecinos más próximos, simulación
Key words: classification, textual information, K-nearest neighbors, simulation

I. Introducción

Es difícil imaginar una organización sin un sistema de información formal o informal. Hoy en día, muchos sistemas de información se han computarizado con la finalidad de que diferentes niveles de una organización compartan gran cantidad de información relacionada con sus diferentes actores (internos o externos) y con sus múltiples actividades, en particular aquellas relacionadas con la toma de decisiones.

Para el desarrollo de estos sistemas de información, diversos autores (Ackoff, 2003, Martin, 2009; Moreno, 2000) resaltan la importancia de obtener claridad en cuanto a los objetivos del sistema de información y a los distintos conceptos relacionados con la información y su manejo; ya que esto resulta fundamental para obtener una mayor definición de las funciones, funcionalidades e, incluso, de la arquitectura del sistema de información. Vázquez (2007), apoyado en la propuesta de Ackoff (1989), recuerda la importancia de distinguir entre datos, información y conocimiento. Los datos son símbolos, o unidades "atómicas", producto de la observación; los mismos representan objetos, eventos y propiedades; al establecer asociaciones entre los datos, por ejemplo preguntando, ¿quién?, ¿qué?, ¿dónde?, ¿cuándo? y ¿cuánto?, se genera información útil para decidir qué hacer, pero no, para decidir cómo hacerlo. Las respuestas a ¿cómo? constituyen el conocimiento. Sin embargo, es importante aclarar que no necesariamente a partir de la información o del conocimiento se pueden obtener los datos o la información que permitieron su construcción (Carlisle, 2007), pues al relacionar, esto es, al interactuar, los datos o la información entre sí, puede surgir un nuevo subsistema conceptual con nuevas propiedades, no siempre identificables en los elementos aislados que lo formaron.

En lo que respecta al manejo de información, los procedimientos de clasificación y de agrupación resultan fundamentales en los procesos de toma de decisiones, ya que ayudan a identificar, distinguir, e incluso a establecer, criterios para evaluar distintas alternativas para tratar determinada situación o problemática. Un ejemplo de clasificación es cuando se cuenta con una base de datos o información de perfiles de clientes en la que se consideran distintos atributos relacionados con su capacidad de pago. Cada perfil es una clase. Si otros clientes solicitan un préstamo hipotecario, el proceso de clasificación consiste en determinar la clase a la que pertenecen y decidir si se les otorga o no el préstamo. Si no se cuenta con información previa el proceso consiste en agrupar las nuevas solicitudes de préstamo tomando en cuenta los atributos relacionados con su capacidad de pago, identificar los grupos y establecer una jerarquía entre los grupos.

Del entendimiento de los conceptos relacionados con la información y su manejo puede surgir, por ejemplo, un sistema destinado sólo a la memorización de datos, un sistema de administración de base de datos que permita generar información, un sistema de clasificación, un sistema de comunicación o bien un sistema "experto" que pueda responder algunas preguntas de tipo ¿cómo? En este artículo se propone desarrollar una aplicación para el manejo de datos e información textual.

En la actualidad existen grandes avances en cuanto a los sistemas de información para el manejo de datos e información numérica. Sin embargo, aún se requieren

esfuerzos para desarrollar sistemas para el manejo de información textual. En una organización, es frecuente, observar gran cantidad de datos, información y conocimientos en forma textual dispersos en documentos impresos, muchos de éstos olvidados en los archivos de la empresa. Cuando se encuentran documentos en forma digital, resulta que algunos de éstos fueron hechos con programas que ya no existen, con formatos obsoletos o archivados en soportes magnéticos distintos, esto es, no es fácil recuperarlos. Por esta dispersión de la información y por la poca integración de los sistemas de información se generan procesos de decisión poco eficaces y eficientes (Ackoff, 2003).

En particular, en el presente trabajo, se propone una aplicación para el manejo de información destinada a la clasificación de documentos.

En la sección II se presenta información general sobre el proceso de clasificación y en particular sobre el algoritmo de los K vecinos más próximos.

En la sección III se presenta el diseño de la aplicación propuesta y las diferentes etapas del proceso de clasificación:

- Creación de la base de entrenamiento
- Ingreso de los documentos a clasificar
- Aplicación del algoritmo de los K-vecinos más próximos para clasificar cada documento en las clases predefinidas en la base de entrenamiento

En la sección IV se estudian opciones para evaluar el desempeño del clasificador. Es decir, para un conjunto de documentos dado, evaluar si éstos quedan en las clases preestablecidas. Sin embargo es necesario considerar, para su validación, distintos factores, como el número de documentos en la base de entrenamiento, la cantidad de clases de éstos y el total de los vecinos más próximos. De aquí la necesidad de realizar pruebas por simulación. En este artículo se entiende por simulación el proceso de diseño del modelo de un sistema y la realización de experiencias para evaluar y comprender el funcionamiento de éste (Shannon, 1988).

II. Proceso de clasificación

La clasificación es un proceso de categorización de información. Para establecer las categorías, se requiere elaborar una base de entrenamiento previa, la cual contiene la información de las diferentes clases. Si se cuenta con un objeto y se desea saber si pertenece a una de las clases, se identifican los atributos más significativos y se evalúa qué tan semejantes son a los atributos de la base de entrenamiento. En el caso que existan varias opciones de clases, se establece una jerarquía para poder decidir la clase a la que será asignado el objeto (Hernández, 2004). Para la identificación del elemento por clasificar, se aplican diferentes métricas (medidas de distancia) que indican qué tan similar es un objeto con respecto de alguna clase de la base de entrenamiento. Si las características son numéricas, existen diversas métricas; la más utilizada es la distancia euclidiana; sin embargo, en el caso de información textual en donde sólo es posible contar las propiedades, éstas son discretas y se necesita establecer otras funciones (Feldman y Sanger, 2007).

A lo largo de este proceso, el usuario interviene y da seguimiento a las diferentes etapas. Es decir, el proceso se produce de manera supervisada.

Los diferentes algoritmos de clasificación se pueden agrupar en tres grandes tipos:

- Los clasificadores paramétricos se basan en la estimación de parámetros de las distribuciones de probabilidad que representan las clases de estudio mediante distribuciones multinomiales, mezcla de multinomiales o por ejemplo una combinación de modelos (Bernouilli, gaussiano y multinomial) (Mesa, 2008) .
- Los clasificadores no paramétricos son aquellos que se basan en la estimación directa sobre la probabilidad a posteriori de pertenecer a una clase. Uno de los clasificadores más simples es el clasificador de los vecinos más próximos, llamado comúnmente KNN (*K-Nearest Neighbors*). Éste consiste en establecer el número de vecinos más próximos del objeto por clasificar (Moreno, 2004)
- Los clasificadores artificiales son clasificadores que se basan en la aplicación de diferentes técnicas de inteligencia artificial para el reconocimiento de patrones; el más conocido es el modelo del perceptrón generalizado o multicapa (Barandela, 2001)

En este trabajo se usará el clasificador no paramétrico de los K-vecinos más próximos, debido a que es uno de los algoritmos más sencillos.

El método consiste en establecer la cantidad de vecinos más próximos del objeto por clasificar dentro de la base de entrenamiento. Cuando el nuevo objeto se presenta al sistema de aprendizaje, éste se clasifica según la distancia más cercana (Mora, 2008). Los vecinos más próximos a un objeto se obtienen, en caso de atributos numéricos, mediante diferentes distancias sobre los *n* posibles atributos. La mejor elección de los *k* depende fundamentalmente de los datos; generalmente, valores grandes de los *k* reducen el efecto de ruido en la clasificación, pero crean límites entre clases parecidas (Clark y Boswell, 2000).

III. Desarrollo de la aplicación de clasificación

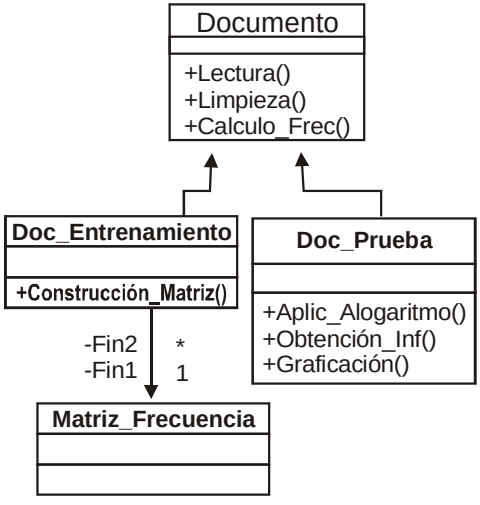
Para realizar esta aplicación se utilizó Matlab,[5] el cual es un *software* matemático que ofrece un entorno de desarrollo integrado (IDE) con un lenguaje de programación interpretado propio (llamado lenguaje M). El programa Matlab permite manejar diferentes estructuras de datos, generar resultados en forma gráfica de calidad, disponer de subrutinas probadas y bibliotecas de funciones preestablecidas. Esto resulta útil para el desarrollo rápido de los algoritmos y demostración de aplicaciones de cómputo. Es importante señalar que esta aplicación forma parte del estudio de factibilidad para el desarrollo, en lenguaje Java,[6] de una aplicación de manejo de información a gran escala para 60 000 documentos (Galindo, 2011).

En esta aplicación se propone un diseño considerando tres clases: **Documento**, **Doc_Entrenamiento**, y **Doc_Prueba** las cuales se relacionan de acuerdo con la Figura 1.

- La clase **Documento** posee un método de *Lectura* que permite leer los archivos de texto plano (*Lectura*) y un método de *Limpieza*, el cual permite eliminar símbolos y caracteres poco relevantes para el usuario. El método de cálculo de frecuencias (*Calculo_Frec*) invoca al

[5] http://www.mathworks.com/
[6] http://www.java.com y un IDE http://netbeans.com

Figura 1.
Diagrama de clases con sus principales métodos.



método construcción de la matriz de frecuencias de la clase **Doc_Entrenamiento.**

- La clase **Doc_Entrenamiento** contiene el método construcción de la matriz de frecuencias (*Construccion_Matriz*) a partir del cual se aplica el método para generar la matriz de frecuencias.
- La clase **Doc_Prueba** posee el método *Aplic_Algoritmo* el cual permite construir la matriz de distancias y aplicar el método de K vecinos más próximos. En el método *Obtencion_Inf* se define a que clase pertenece el objeto a clasificar (en este caso un documento). El método de *Graficacion* permite visualizar los resultados en forma de grupos.

En la figura 2 se presenta el funcionamiento de la aplicación tomando como base estas clases y la definición del proceso de clasificación. Con el fin de presentar las principales funcionalidades de esta aplicación, y por la facilidad de acceso a documentos, se decidió usar documentos de texto de una organización editorial, la cual desea clasificar artículos de revistas en distintas categorías. Para fines demostrativos se supone que desean clasificar esos artículos (definición de atributos), tomando como atributos los signos de puntación y signos aritméticos, en una de las siguientes categorías: Sociales (clase 1), Ingeniería (clase 2) y Medicina (clase 3).

Figura 2.
Proceso de clasificación

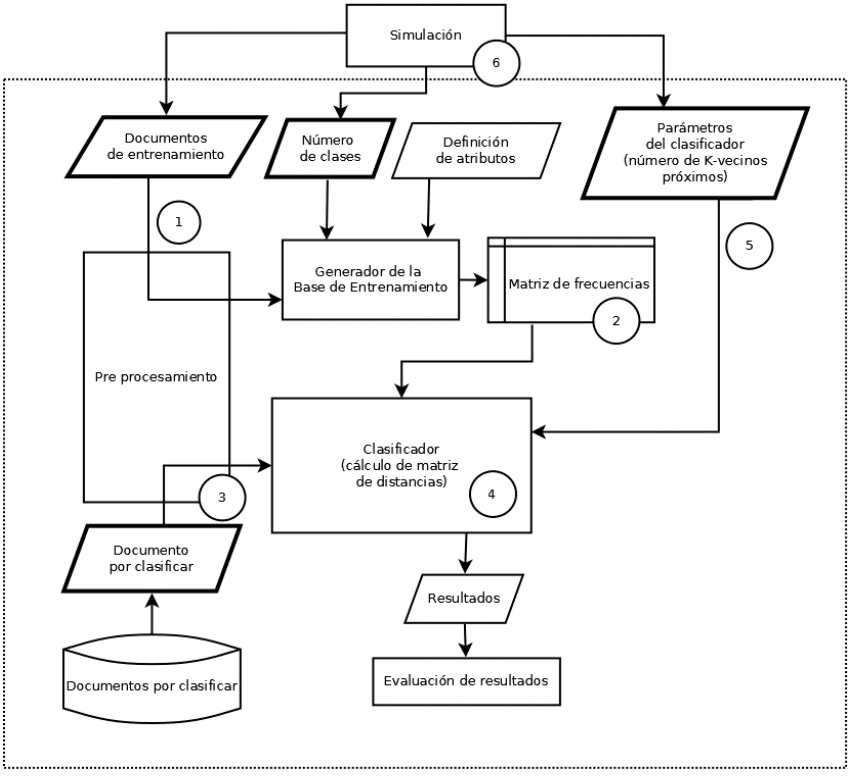


Tabla 1.
Matriz de frecuencias

Documentos	**.**	**,**	**“**	**(**	**)**	**?**	**!**	**-**	**:**	**;**	**<**	**>**	**+**	*****	**/**	**=**	***# Pala***	***Clase***
1	413	694	0	229	228	0	0	65	71	81	3	5	31	2	5	87	8170	1
2	315	736	0	46	46	4	0	54	23	37	0	0	0	0	13	0	8179	1
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
50	483	866	0	113	116	1	0	51	46	28	0	0	0	2	11	0	10871	1
51	217	183	0	57	59	0	0	72	25	11	0	0	0	0	7	6	6450	2
52	738	345	0	243	244	0	0	33	68	26	1	1	0	0	48	24	7399	2
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
100	507	275	0	110	118	0	0	09	29	11	0	1	2	18	10	10	9053	2
101	588	379	0	110	115	1	0	35	86	44	27	2	0	60	27	53	5503	3
102	390	305	0	55	55	3	0	43	59	25	0	0	0	2	16	6	5695	3
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
150	377	356	0	46	57	3	0	93	83	77	0	0	0	0	4	0	5196	3

En la primera etapa (etapa 1 en la figura 2), que se realiza sólo una vez, la aplicación inicia con la lectura de la información previa (documentos de entrenamiento) y el procesamiento supervisado de los documentos. Una vez que el usuario considera que los documentos están listos se obtiene la matriz de frecuencias (etapa 2 en la figura 2), desarrollada en la tabla 1, con las características establecidas. Esta matriz de frecuencias es la base de conocimiento previo, llamada comúnmente “base de entrenamiento”.

En la segunda etapa (etapa 3 en la figura 2), que se repite para cada documento por clasificar, el usuario proporciona el documento por clasificar y se generan las frecuencias, descritas en la tabla 2 tomando en cuenta las mismas características usadas para construir la base de entrenamiento.

Tabla 2.
Frecuencias del documento por clasificar

.	**,**	**“**	**(**	**)**	**?**	**!**	**-**	**:**	**;**	**<**	**>**	**+**	*****	**/**	**=**	**# Pala**
422	374	0	120	120	0	0	297	54	46	0	0	2	0	235	49	6645

Una vez obtenida la matriz de frecuencias y las frecuencias del documento por clasificar se procede al cálculo de la matriz de distancias (etapa 4 en la figura 2) descrita en la tabla 3.

Tabla 3.
Matriz de distancias

Distancia	***Clase***
0.2445	1
0.1311	1
.	.
.	.
.	.
0.5610	2
0.5450	2
.	.
.	.
.	.
0.5363	3
0.5362	3

Posteriormente se ingresa el número de vecinos más próximos que se desea obtener (etapa 5 en la figura 2) y en base al número indicado, se procede a obtener las distancias menores. Una vez seleccionados, se contabiliza, para decidir en qué clase aparece el mayor número de atributos y se obtiene así la clase en la que se clasifica el documento

Tabla 4.
Matriz de distancia con los 10 vecinos más cercanos

Distancia (x 10^{-3})	***Clase***
0	**3**
306.8241	**3**
347.7959	1
369.1734	2
389.6858	2
404.6863	1
410.8771	3
421.5780	1
426.9344	3
439.1070	2

Para este ejemplo se propusieron 10 vecinos (aunque se recomienda usar un número de vecinos impar para evitar ambigüedades), los cuales se muestran en la tabla 4. Como resultado, el programa nos indica a qué clase pertenece el documento, es decir, si pertenece a Sociales, Ingeniería o Medicina. Para este ejemplo en particular el documento pertenece a la clase 3, es decir, a Medicina.

IV. Evaluación de la aplicación

Para realizar la evaluación del desempeño de un clasificador existen diversas estrategias, las más utilizadas son (Mesa, 2008):

- Re sustitución
- Validación simple
- Validación cruzada (*Cross- Validation*)

1.- Re sustitución

Es una técnica de evaluación de desempeño de la clasificación en donde la misma base de datos de aprendizaje se usa para la prueba. Este tipo de evaluación proporciona una medida optimista de clasificación con un mínimo error.

2.- Validación simple

Esta técnica (también se conoce como *hold out*.) de evaluación de desempeño es una parte integral del proceso de entrenamiento, el cual consiste en dividir los datos en tres grupos:

- Conjunto de datos de entrenamiento
- Conjunto de datos de validación
- Conjunto de datos de prueba

3.- Validación cruzada

En esta técnica, también llamada "deja los k fuera" (*leave k out*) o "manten los k fueras" (*hold k out*), se toma un porcentaje en forma aleatoria (del total de la base), para el aprendizaje y resto para la prueba. Generalmente el porcentaje tanto para el aprendizaje como para la prueba es del 50%. El tiraje y prueba aleatorio se repite un cierto número de veces. El resultado final será el promedio sobre las n-realizaciones.

Como resultado de aplicar un método de validación, se obtiene una matriz de confusión. La matriz de confusión ideal presenta una matriz con "unos" en la diagonal, como se presenta en la tabla 5.

Tabla 5.
Matriz de confusión ideal

Clase 1	*Clase 2*	...	*Clase n*
1	**0**	...	**0**
0	**1**	...	**0**
0	**0**	...	**0**
0	**0**	...	**1**

V. Pruebas de simulación y resultados

Las pruebas se realizan mediante simulación (punto 6 en la figura 2), utilizando los mismos documentos de prueba. Para el experimento se consideró el número de documentos en la base de entrenamiento, el número de clases de éstos y el total de vecinos más próximos. En estas pruebas, la selección de los documentos de prueba se realiza en forma secuencial, sin embargo también es posible realizarla en forma aleatoria. El método utilizado para la evaluación del desempeño de esta aplicación es el método de validación cruzada, el cual toma para cada una de las clases 30 documentos de cada clase para la realización de la prueba.

Un ejemplo de la matriz de confusión obtenida se presenta en la tabla 6.

Tabla 6.
Matriz de confusión obtenida con 80 documentos de entrenamiento, 15 vecinos y 3 clases

	Clase 1	**Clase2**	**Clase 3**
Clase 1	0.86	0	0.14
Clase 2	0	1	0
Clase 3	0.07	0	0.93

Tabla 7.
Resultados de la clasificación

	Documentos en cada clase de la base de entrenamiento																	
	50						80						125					
Clases de entrenamiento	**2** (sociales e ingeniería)			**3** (sociales, ingeniería y medicina)			**2** (sociales e ingeniería)			**3** (sociales, ingeniería y medicina)			**2** (sociales e ingeniería)			**3** (sociales, ingeniería y medicina)		
Vecinos más próximos del objeto por clasificar	**5**	**15**	**25**	**5**	**15**	**25**	**5**	**15**	**25**	**5**	**15**	**25**	**5**	**15**	**25**	**5**	**15**	**25**
Resultado	0.75	0.8	**0.90**	0.74	0.82	**0.90**	0.85	**0.91**	**0.92**	0.81	**0.90**	**0.91**	0.83	**0.92**	**0.94**	0.86	**0.90**	**0.93**

Los resultados, modificando el número de documentos de la base de entrenamiento, la cantidad de clases de la base de entrenamiento y el total de vecinos más próximos del documento por clasificar, se presentan en la tabla 7.

El resultado se mide con la razón del número de documentos bien clasificados (es decir aquellos que se encuentran en la clase a la que pertenecen) respecto del número total de documentos por clasificar. Por ejemplo, si se consideran dos clases y se desea clasificar 60 documentos (en teoría 30 para cada clase), considerando 5 vecinos más próximos, se obtiene, después de aplicar el algoritmo, que sólo 50 fueron bien clasificados. El resultado aparece entonces dividiendo 50 entre 60, es decir, 0.75.

De los resultados de esta simulación, ver tabla 7, se observa que, entre mayor sea el tamaño de la base de entrenamiento y mayor el número de vecinos más próximos, se obtiene una mejor clasificación de los documentos.

VI. Conclusiones

El presente trabajo presenta una aplicación enfocada a la clasificación de documentos de texto no estructurado o texto plano. Este es un caso que no considera la jerarquía o la importancia del contenido de los documentos; en realidad la información puede presentarse en diferentes niveles de importancia y no necesariamente tiene el mismo valor en la organización. Para incluir el nivel de importancia de la información contenida en los documentos, sería necesario considerar un formato de texto, como el formato XML (Harold, 2004). Este formato es ampliamente usado en la actualidad por los beneficios que presenta; por ejemplo, la posibilidad de estructurar contenido con distintos niveles de importancia y distinguir así entre las diferentes jerarquías de información textual.

Es importante resaltar que la clasificación requiere de información previa, la cual no siempre está disponible

en una organización. En este caso el proceso de agrupación resulta más adecuado ya que no necesita conocimiento previo.

En lo que respecta a las pruebas de simulación, se observa que, para poder llevar a cabo una buena clasificación, es importante partir de una buena base de entrenamiento y para ello se aconseja contar con el mismo número de elementos en cada una de las clases, ya que si alguna clase cuenta con mayor número de elementos, el algoritmo de clasificación parece ser menos efectivo, favoreciendo a la clase que posea mayor número de elementos.

Por supuesto, hay que considerar los resultados de esta simulación con reserva, ya que los documentos de prueba se seleccionaron de forma secuencial, y esto genera resultados deterministas. Debido a la dificultad de disponer de datos de entrenamiento representativos para simular todas las situaciones posibles, es importante incluir aleatoriedad en el proceso: por ejemplo tomar documentos de manera aleatoria, definir atributos en forma aleatoria, etc. Otra opción sería comparar los resultados usando otros métodos de validación.

En la bibliografía, se reportan diversas aplicaciones para la clasificación de documentos (Téllez, 2005). Sin embargo, esta experiencia ha permitido realizar una aplicación funcional, obtener una mayor comprensión del proceso de construcción de matrices de frecuencias y de distancias a partir de información textual; y contar con una aplicación que permita realizar experimentos de simulación para evaluar los efectos de diversos parámetros sobre el resultado de la clasificación. Estos resultados son una etapa importante en el estudio de factibilidad y es parte fundamental para el desarrollo de una aplicación de gran escala para tratar documentos XML que posean una estructura interna (Galindo, 2011).

Agradecimientos

Mihaela Juganaru-Mathieu agradece el apoyo de la Doctora Silvia González Brambila, Coordinadora de la Maestría en Ciencias de la Computación, de la División de Ciencias Básicas e Ingeniería (DCBI), de la Universidad Autónoma Metropolitana, Unidad Azcapotzalco (UAM—A), así como el del Doctor Nicolás Domínguez Vergara, profesor y ex jefe del Departamento de Sistemas, de la misma División.

Héctor Javier Vázquez agradece al doctor Russell L. Ackoff (1919, 2009) por todas sus enseñanzas sobre los diferentes conceptos de sistemas, por sus publicaciones, conferencias, cursos, investigaciones, pláticas y conversaciones personales. Así mismo agradece al profesor Germán Sergio Monroy Alvarado por su amistad, su paciencia y motivación continua para estudiar las metodologías de sistemas. Finalmente, Héctor Javier Vázquez no deja de agradecer, por el apoyo otorgado para realizar sus estudios de Maestría y Doctorado en Francia. En particular agradece, a la Universidad Autónoma Metropolitana, Unidad Azcapotzalco, por las distintas licencias sin goce de sueldo, concedidas; y por otro lado, agradece, por el apoyo económico otorgado, al Ministerio de Educación Superior y de la Investigación (MESR) y a la Agencia Nacional del Empleo de Francia (ANPE), ambas instituciones del Gobierno de Francia.

Bibliografía

• Ackoff, R. L. (1989). *From Data to Wisdom*, Journal of Applied Systems Analysis, Vol. 16.

• Ackoff, R. L. (2003). *Redisigning Society*, Stanford University Press, Stanford.

• Barandela, G. E. (2001). *Corrección de la muestra para el aprendizaje del perceptrón multicapa*. Revista Iberoamericana de Inteligencia Artificial, 2-9.

• Galindo, D. C. K., Juganaru-Mathieu M. y Vázquez H. J. (2011). *Specification Design for an XML Mining Configurable Application*, International MultiConference of Engineers and Computer Scientists (aceptada para su presentación y publicación, http://www.iaeng.org/IMECS2011/publication.html).

• Carlisle, J. P. (2007). *A Look into the Relationship Between Knowledge Management and the Knowledge Hierarchies*, Proceedings of the 40th Hawaii International Conference on System Sciences (http://www.computer.org/portal/web/csdl/doi/10.1109/HICSS.2007.19).

• Clark, P., y Boswell, R. (2000). *Data Mining. Practical Machine Learning Tools and Tecniques with Java Implementations*. Morgan Kaufmann Publishers.

• Feldman, R., y Sanger, J. (2007). *The Text Mining Handbook. Advanced Aproaches in Analyzind Unstructured Data*. Cambridge University Press.

• Harold, E. R., y Means, W. S. (2004). *XML in a Nutshell*. O'Reilly Media.

• Hernández, J., Ramírez Quintana, M. J., y Ramírez F. C. (2004). *Introducción a la Minería de Datos*. Pearson Prentice Hall.

• Martin, J. N. (2009). *Knowledge Generation in the Enterprise Using Information and Data Systems*. Second International Symposium on Engineering Systems, MIT, Cambridge, Massachusetts, Junio 15-17, (http://esd.mit.edu/symp09/submitted-papers/martin-paper.pdf).

• Mesa, F. D. (2008). *Algoritmos de Aprendizaje Continuo Mediante Selección de Prototipos para Clasificadores Basados en Distancias*. Tesis de la Universitat Jaume I, Castellón, España.

• Mora, F., J., Morales España, G., y Barrera Cárdenas, R. (2008). *Evaluación del clasificador basado en los K Vecinos más Cercanos para la Localización*. Ingeniería e Investigación , 81-86.

• Moreno, F. S. (2004). *Clasificadores Eficaces Basados en Algoritmos Rápidos de Búsqueda del Vecino más Cercano*. Tesis de Doctorado de la Universidad de Alicante, España.

• Moreno, O. A. (2000). *Diseño e Implementación de un Lexicón Computacional para Lexicografía y Traducción Automática*. Estudios de Lingüística del Español, Vol 9. http://elies.rediris.es/

• Shannon, R. E. (1988) *Simulación de Sistemas*, Trillas, Ciudad de México, D.F.

• Téllez, V. A. (2005) *Extracción de Información con Algoritmos de Clasificación,* Tesis de Maestría del Inaoe (http://ccc.inaoep.mx).

• Vázquez H. J., Martínez A. F. J., Monroy A. G. S. (2007). *Más allá del Conocimiento: un Enfoque Sistémico*. Administración y Organizaciones, 23-38.