René Venegas
Clasificación de textos académicos en función de su contenido léxico-semántico
Revista Signos, vol. 40, núm. 63, 2007, pp. 239-271,
Pontificia Universidad Católica de Valparaíso
Chile

Revista Signos,
ISSN (Versión impresa): 0035-0451
revista.signos@ucv.cl
Pontificia Universidad Católica de Valparaíso
Chile

Revista Signos
2007, 40(63)
239-271

Clasificación de textos académicos en función de su contenido léxico-semántico*

René Venegas
Pontificia Universidad Católica de Valparaíso
Chile

Resumen: El objetivo de esta investigación es clasificar, utilizando y comparando dos métodos de categorización automática, los textos académicos incluidos en el Corpus PUCV-2006 perteneciente al trabajo realizado en el proyecto Fondecyt 1060440. Estos métodos están basados en los lexemas de contenido semántico compartidos en el corpus de textos académicos usados en cuatro carreras profesionales de la Pontificia Universidad Católica de Valparaíso, Chile. El corpus PUCV-2006 actualmente está conformado por 652 textos, los que en cantidad total de palabras alcanza a 96.288.874. Para los propósitos de esta investigación, utilizamos una muestra de 216 textos (30.886.081 palabras) divididos en cuatro áreas disciplinares: 26 usados en Ingeniería en Construcción, 31 en Química, 64 en Trabajo Social y 95 en Psicología. Los métodos de clasificación a comparar en esta investigación son Bayes Ingenuo y Máquina de Soporte de Vectores, ambos métodos permiten identificar un pequeño grupo de lexemas compartidos, que una vez pesados estadísticamente, sirven para clasificar un nuevo texto en alguna de las cuatro áreas disciplinares. Los resultados nos permiten establecer que la Máquina de Soporte de Vectores clasifica más eficientemente los textos académicos, con altos valores de precisión y exhaustividad. Con este método podemos identificar automáticamente el dominio disciplinar de un nuevo texto académico en consulta con un alto porcentaje de exactitud (93,9%). Proyectamos usar este método como parte de un análisis multidimensional más acabado del Corpus PUCV-2006.

Palabras Clave: Discurso académico, modelo vectorial, Bayes Ingenuo, Máquina de Soporte de Vectores.

Recibido:
3-VII-2006
Aceptado:
7-XI-2006

Correspondencia: René Venegas (rene.venegas@ucv.cl). Tel.: (56-32) 2273388. Pontificia Universidad Católica de Valparaíso. Av. Brasil 2830, piso 9, Valparaíso, Chile.

* Investigación financiada parcialmente por el Proyecto FONDECYT 1060440.

Academic text classification based on lexical-semantic content

Abstract: The aim of this research is to classify, using and comparing two automatic classification methods, the academic texts included in the PUCV-2006 Corpus belonging to the Fondecyt 1060440 research project. The methods are based on shared lexical-semantic content words present in a corpus of academic texts used in four professional carriers at the Pontificia Universidad Católica de Valparaíso, Chile. The research corpus, nowadays, is constituted by 652 texts with 96.288.874 words. For our purposes, we use a sample of 216 texts (30.886.081 words) divided, as following: 26 used in Construction Engineering, 31 used in Chemistry, 64 used Social Work, and 95 used in Psychology. The classification methods compared in this research are Multinomial Naïve Bayes and Support Vector Machine, both permits to identify a small group of shared words that permit, according statistical weights, to classify a new text into the four disciplinary areas. The results allows us to establish that Support Vector Machine classify in a efficient way academic texts, with high precision and recall values. With this method we are able to identify automaticalLy the disciplinary domain, with a high percentage of accuracy (93,9%), of a new academic text in a query. We project to use this method as part of a more detailed multidimensional analysis of the PUCV-2006 Corpus.

Key Words: Academic discourse, vectorial model, Naïve Bayes, Support Vector Machine.

INTRODUCCIÓN

La investigación que presentamos a continuación se encuentra enmarcada en el proyecto Fondecyt 1060440, denominado "El discurso especializado escrito en el ámbito universitario y profesional: lingüística de corpus y análisis multidimensional". En este proyecto nos proponemos realizar un estudio descriptivo-comparativo de orden lingüístico-textual a partir de los textos que son leídos en el ámbito académico y en el profesional, tanto en las áreas de las Ciencias Básicas y de la Ingeniería como en el de las Ciencias Sociales y Humanas. Para ello, nos encontramos abocados a la recolección y estudio de ocho corpus textuales, que circulan en el nivel académico universitario y en el nivel profesional laboral (ver Parodi, 2007). Para el análisis de estos textos nos hemos propuesto seguir una línea innovadora en los estudios de este tipo a nivel nacional y latinoamericano, esto es, la utilización de la lingüística de corpus y el empleo de herramientas computacionales para el tratamiento de corpus digitales. En consonancia con lo anterior, en este artículo, aportamos a la descripción del Corpus PUCV-2006. En términos más precisos, nuestro objetivo de investigación es clasificar, utilizando y comparando dos métodos de categorización automática, los textos académicos incluidos en este corpus.

Realizar la tarea que nos proponemos nos exige integrar en una buena medida dos ámbitos disciplinares, a veces considerados opuestos: la lingüística, a través del estudio del discurso desde su variante especializada y académica, y la matemática y estadística, a través de su

aplicación en los diversos métodos y técnicas que se utilizan en el procesamiento natural del lenguaje (PNL). En este sentido, concordamos con García (2007), en que es habitual percibir cierto escepticismo entre algunos especialistas cuando se plantean cuestiones lingüísticas obtenidas o acreditadas por medio de procedimientos matemáticos. Sin embargo, esto debiera poder ser superado si se interpreta la ciencia de los números como un conjunto de elementos y reglas que se integran en el seno de un sistema de representación con el que se puede dar cuenta de algunos aspectos de la realidad, a través de la elaboración de modelos explicativos, métodos y técnicas de análisis.

Ante este desafío hemos decidido acercarnos a algunos de estos métodos y técnicas, utilizados en el ámbito de la clasificación automática de documentos. En concreto, en esta investigación, utilizamos para la categorización de los textos del corpus dos métodos, desarrollados a partir de los modelos vectoriales de representación de documentos: el método de clasificación Bayes Ingenuo y el método conocido como Máquina de Soporte de Vectores. Además, los comparamos con el fin de determinar cuál de ellos puede clasificar con mayor exhaustividad y precisión los textos académicos de las cuatro disciplinas en investigación.

Cabe señalar, que la conexión entre estos métodos y los estudios lingüísticos se encuentra en la elección de un criterio particular de clasificación, esto es, los lexemas de contenido semántico compartidos entre los textos de cada una de las áreas. De este modo, utilizamos una categoría lingüística de la semántica léxica para identificar las características propias del discurso de cada una de las disciplinas, plasmadas en la materialidad discursiva. Aquí, se halla subyacente un supuesto que cruza toda esta investigación. En términos más explícitos, suponemos que podemos diferenciar significativamente discursos académicos, a través de la ocurrencia sistemática y particular de un conjunto más bien pequeño de lexemas de contenido semántico compartido. En otras palabras, pensamos que es posible diferenciar los textos, a partir de la idea de que en los textos se encuentran palabras comunes, pero que su uso particular (y la información que ellas aportan) es distinta para cada grupo de textos que representa a las cuatro disciplinas en estudio. En lo particular, esta idea es interesante, pues en términos prácticos, si pensamos en un computador que deba clasificar textos según sus disciplinas, la opción más tradicional sería tratar de identificar las características propias de cada discurso disciplinar que los hacen distintos (por ejemplo, terminología propia); sin embargo, ello requeriría que el sistema computacional tuviera previamente almacenada información léxica y semántica respecto de, por ejemplo, la terminología, a través de glosarios y ontologías, que le permita tener información sobre los aspectos que hacen distintas a cada una de las clases de textos según disciplina. Por el contrario, el supuesto que explicitamos aquí sugiere que solo identificando un grupo limitado y fijo de lexemas se puede clasificar cualquier texto, según su comportamiento en términos de frecuencia y de peso estadísticos.

(Parodi & Venegas, 2004; Cademartori, Parodi & Venegas, 2006; Parodi, 2006a, Venegas, 2005, 2006).

Con el fin de cumplir con todo lo anterior, presentamos en el apartado de antecedentes teóricos una breve discusión crítica sobre los conceptos de texto especializado y académico; así como una explicación de lo que se conoce como clasificación automatizada de textos, del modelo vectorial y de los métodos a utilizar en esta investigación. Luego se presenta un apartado metodológico en el que se explican detalladamente los procedimientos realizados; finalmente se presentan los resultados y las conclusiones obtenidas en esta investigación.

1. Antecedentes teóricos

1.1. El discurso especializado y el discurso académico

Sabemos que en la ciencia no existe siempre consenso en la denominación de los objetos de estudio, debido fundamentalmente a la focalización y delimitación que deben realizar los investigadores al intentar conceptualizar el objeto a estudiar. Normalmente, los abordajes son múltiples en razón de supuestos teóricos divergentes. El concepto de "discurso especializado" no es la excepción. Este ha sido denominado de múltiples maneras, por ejemplo: discurso académico, discurso especial, discurso profesional, discurso técnico, discurso institucional, etc. Alcanzar un relativo orden terminológico y lograr una visión más o menos homogénea tampoco resulta fácil (Ciapuscio, 2000; López, 2002, Parodi, 2007).

Por otra parte, determinar de forma discreta si un texto se clasifica como de especialidad o de tipo general es, sin duda, un problema teórico y descriptivo (Schröder, 1991; Parodi, 2004, 2006b). Hoy en día, la postura predominante está en favor de un *continuum* de textos que se distribuyen de manera progresiva desde un dominio altamente especializado hasta otro extremo mucho más divulgativo y general (Gläser, 1982; Schröder, 1991; Halliday & Martin, 1993; Jeanneret, 1994; Peronard, 1997; Ciapuscio, 1994, 2000; Cabré, 2002; Parodi, 2004, 2006b). En este sentido, Parodi (2004: 10) plantea que:

> "[...] es un hecho que establecer límites precisos entre un tipo de texto y otro es una cuestión de envergadura. Sin importar el foco atencional en uno u otro criterio clasificatorio, siempre existirán casos mixtos o límites; sin embargo, parece ser que el discurso especializado corresponde a una categoría reconocible para cualquier hablante de español".

Gotti (2003), siguiendo la idea del *continuum*, plantea en relación con la naturaleza multi-dimensional del discurso especializado que no existe homogeneidad entre los diferentes lenguajes especializados. Argumenta que las variaciones disciplinares producen no solo connotaciones

léxicas especiales, sino que también a menudo influencian otras opciones (morfosintácticas, textuales y pragmáticas), teniendo además repercusiones en las peculiaridades epistemológicas, semánticas y funcionales de una variedad de discurso especializado.

Ahora bien, desde esta concepción de discurso especializado, avanzaremos hacia el concepto de "discurso académico", el cual –en principio– visualizamos como incrustado en un *continuum* mayor, constituido por el especializado. En este sentido, esta investigación se focaliza en el estudio y descripción del discurso académico de circulación a nivel universitario, en particular, de los textos que se leen en cuatro carreras de la Pontificia Universidad Católica de Valparaíso, Chile (Corpus PUCV-2006).

Desde este foco, el discurso académico se entiende en un sentido restringido, pues –para nosotros– es el ámbito de circulación, los propósitos pedagógicos, los tópicos y los lectores en formación, los que caracterizan esta práctica discursiva. Al mismo tiempo, desde una mirada lingüística propiamente tal, este discurso se identifica a partir de ciertos grupos de rasgos lingüísticos que tienden a co-ocurrir sistemáticamente a lo largo de las tramas textuales. Esto quiere decir que las funciones más comunicativas y sociales, así como el conocimiento disciplinar que se transmite, se manifiestan a través de ciertas selecciones léxicas y retórico estructurales (Parodi, 2004, 2005; Venegas, 2005, 2006).

Dado que nos interesa la indagación del discurso académico escrito desde una metodología vanguardista, entendida desde la lingüística de corpus y el procesamiento del lenguaje natural con herramientas estadístico-computacionales, buscamos una interacción entre la lingüística, la estadística y la informática. Por ello, en lo que sigue, focalizaremos nuestra atención en los aspectos relativos a la clasificación automática de documentos, ya que pensamos que la identificación de una técnica efectiva de clasificación nos permitirá avanzar en una descripción más certera y automatizada de nuestro Corpus PUCV-2006 (ver Parodi, 2007).

1.2. Hacia una clasificación automática de textos

Si bien una clasificación de textos se puede realizar manualmente, tarea que en lingüística se efectúa fundamentalmente en base a criterios multiniveles (lingüísticos, textuales, pragmáticos y funcionales) y que se conoce como tipologización textual (Ciapuscio, 1994, 2000; Parodi, 2005), también es deseable alcanzar un nivel de automatización de estos procedimientos. Esta tarea, desde una aproximación del procesamiento automático del lenguaje, es conocida como clasificación automática de documentos, donde el estándar es construir y usar las llamadas "máquinas de aprendizaje supervisado". El proceso de crear una clasificación automática de textos consiste en descubrir variables que sean útiles en la discriminación de los textos que pertenecen a clases pre-existentes distintas. En particular, los clasificadores (programas que ejecutan algoritmos de clasificación) son entrenados en un grupo de docu-

mentos, previamente clasificados y etiquetados acorde a algún criterio particular (tema, materia, origen, etc.), conformando una clase. De esta manera, el objetivo de estos clasificadores es decidir en qué categoría debe ir cada texto nuevo, partiendo de un esquema de clasificación previo (Figuerola, Zazo & Berrocal, 2000). También se dice que la clasificación o categorización automática de documentos puede ser entendida como una tarea en la cual, en base a la identificación por medios matemático-estadísticos, un documento nuevo es asignado a una clase particular de documentos pre-existentes (Jurafsky & Martin, 2000).

Es importante destacar que la clasificación automática de documentos surge de los estudios realizados en recuperación de información (Salton, 1968; Salton & McGill, 1983; Salton & Bukley, 1988). Para ello, se ha utilizado mayoritariamente corpus textuales en lengua inglesa. En menor medida, y solo en los últimos años, estas técnicas han sido aplicadas a corpus en español (Figuerola, 2000; Figuerola et al., 2000; Martín-Valdivia, García-Vega & Ureña-López, 2001; Zazo, Figuerola, Alonso & Gómez, 2002; Cerviño, García, Calvo & Ceccato, 2004).

En términos prácticos, la utilidad de la clasificación automática de documentos se basa en la posibilidad de poder efectuar una adecuada recuperación de documentos no conocidos, asumiendo que aquellos textos que tratan, por ejemplo, de la misma materia están clasificados juntos, o en sectores cercanos (un ejemplo de herramienta desarrollada usando estos procesos son los filtros de correo basura o *spam*).

Diversas técnicas han sido propuestas, desde hace ya algunos años (Fairthorne, 1961; Hayes, 1963; Salton, 1968). Buena parte de tales técnicas se basan en la utilización de medidas de semejanza (o de disparidad, dependiendo del punto de vista) entre dos documentos. Una síntesis de las más importantes, tanto de unas como de otras, puede encontrarse en Jurafsky y Martin (2000) y Manning y Schütze (2003).

En síntesis, la clasificación automática de documentos puede concebirse como un proceso de "aprendizaje matemático-estadístico", durante el cual un algoritmo implementado computacionalmente capta las características que distinguen cada categoría o clase de documentos de las demás, es decir, aquellas que deben poseer los documentos para pertenecer a esa categoría. Estas características no tienen porqué indicar de forma absoluta e inequívoca la pertenencia a una clase o categoría, sino que más bien lo hacen en función de una escala o graduación. De esta forma, por ejemplo, documentos que posean una cierta característica tendrán un factor de posibilidades de pertenecer a determinada clase, de modo que la acumulación de dichas características arrojará un resultado que consiste en un coeficiente asociado a cada una de las clases ya conocidas. Este coeficiente lo que expresa en realidad es el grado de confianza o certeza de que el documento en cuestión pertenezca a la clase asociada al coeficiente resultante.

Las técnicas empleadas para la clasificación de documentos, como ya mencionamos, son originarias de los métodos clásicos de recuperación de información, es por ello que antes de describir las técnicas a utilizar en esta investigación, presentamos brevemente el modelo vectorial, ya que, sin duda, es la base conceptual para las técnicas de clasificación actuales (Figuerola et al., 2000; Manning & Schütze, 2003). El modelo vectorial fue definido inicialmente por Salton (1968) y es ampliamente usado en operaciones de recuperación de información, así como también en operaciones de categorización automática, filtrado de información, etc. (Zazo et al., 2002; Manning & Schütze, 2003).

Un aspecto nos resulta altamente relevante. Los términos de "aprendizaje", "entrenamiento" y "decisión", en este contexto, no son más que metáforas lexicalizadas que se utilizan acorde con un paradigma cognitivista de la ciencia, en el que, en su versión extrema, la mente funcionaría como un computador. En este artículo, por motivos de focalización y espacio, no discutiremos la validez de tales términos, pero sí queremos plantear claramente que el proceso que desarrollan las máquinas está muy lejos de ser similar al proceso de aprendizaje o de toma de decisión que llevan a cabo los seres humanos, aun cuando algunos resultados de estos procesos automatizados nos puedan ser útiles e incluso, en ocasiones, nos sorprendan.

1.2.1. El modelo vectorial

A continuación presentaremos sucintamente la forma en que es utilizado el modelo vectorial para la recuperación de información. Nos interesa destacar este modelo, porque es la manera más sencilla de explicar cómo se llevan a cabo las operaciones matemático-estadísticas que permiten determinar la similitud entre documentos a partir de las palabras contenidas en ellos, en nuestro caso los lexemas de contenido semántico, y a partir de ellas clasificar documentos (textos académicos) nuevos en clases pre-existentes (disciplinas).

Según Zazo et al. (2002), en el modelo vectorial se intenta recoger la relación de cada documento D_i, de una colección de N documentos, con el conjunto de las m características de la colección. Formalmente, un documento puede considerarse como un vector que expresa la relación del documento con cada una de esas características. La Ecuación 1 da cuenta de esta representación vectorial de un documento.

$$1. \quad D_t \rightarrow \vec{d} = (c_{t1}, c_{t2}, ..., c_{tm})$$

Observamos que el vector identifica en qué grado el documento D_i satisface cada una de las m características. En otras palabras el vector, c_ik es un valor numérico que expresa en qué grado el documento D_i posee la característica k. La noción de "característica" suele concretarse en la ocurrencia de determinadas palabras o términos en el documento, aunque nada impide to-

mar en consideración otros aspectos. Respecto de esto último, cabe señalar que este tipo de procedimientos se han utilizado en el reconocimiento de objetos, donde las características son de carácter viso-perceptual (color, forma, etc.) (Landauer, 2002).

Si se consideran las palabras como características definitorias del documento, el proceso que debe seguir el sistema de clasificación se inicia con la selección de aquellas palabras útiles que permitan discriminar unos documentos de otros. En este punto, debemos señalar que no todas las palabras contribuyen con la misma importancia en la caracterización del documento. Desde el punto de vista lingüístico aplicado a la recuperación o clasificación de documentos, existen lexemas casi vacíos de contenido semántico, como los artículos, las preposiciones o las conjunciones. Estos lexemas son conocidos como palabras funcionales en la tradición lingüística y como *stop words* en el procesamiento de lenguaje natural. Estas palabras, que en español comúnmente son entre 100 y 200, son poco útiles para el proceso de clasificación (Cerviño et al., 2004). También son poco importantes aquellas palabras que por su frecuencia de aparición en toda la colección de documentos pierden su poder de discriminación, es por ello que o son eliminadas o son ponderadas con muy bajo peso estadístico.

Además de la eliminación de las palabras funcionales o poco informativas, en el proceso se pueden incluir aplicaciones léxicas como lematización o extracción de raíces, etiquetado de términos, detección de unidades multipalabra, etc. Todo esto permite reducir la cantidad de palabras a considerar en la matriz de análisis, sin embargo, el problema de estas aplicaciones es la pérdida de información relevante (género y número de las palabras), que puede ser necesario para una clasificación más ajustada a la realidad de los textos.

Una vez seleccionado el conjunto de términos caracterizadores de la colección de documentos, es necesario calcular el valor de cada elemento del vector del documento. El caso más simple es utilizar una aproximación binaria, de forma que si en el documento D_i aparece el término k, el valor c_ik sería 1, y en caso contrario sería 0.

Ahora bien, como se sabe, una palabra puede aparecer más de una vez en el mismo documento y, además, algunas palabras pueden considerarse con más peso estadístico que otras, esto es, más significativas que otras, de forma que el valor numérico de cada uno de los componentes del vector obedece normalmente a cálculos más sofisticados que la simple asignación binaria. Por otro lado, también es importante normalizar los vectores para no privilegiar documentos.

Se han propuesto diversos métodos para calcular el peso de cada término en el vector que representa al documento (Salton & McGill, 1983; Salton & Buckley, 1988; Harman, 1992), pero en general, para estimarlos se parte de dos ideas en cierto sentido contrapuestas: si un término se consigna mucho en un documento, aquel es importante para caracterizar el

documento. Pero si aparece en muchos documentos de la colección, este término no resulta beneficioso para distinguir un documento de los demás, dado su escaso poder discriminatorio, siendo por tanto poco útil para la recuperación o clasificación de nuevos documentos.

Para determinar la capacidad de representación de un término para un documento dado, se calcula el número de veces que este aparece en dicho documento, obteniéndose la frecuencia del término en el documento (*term frequency,* tf). Por otra parte, si la frecuencia de un término en toda la colección de documentos es extremadamente alta, se opta por eliminarlo del conjunto de términos de la colección. Podría decirse que la capacidad de recuperación de un término es inversamente proporcional a su frecuencia en la colección de documentos. Esto es lo que se conoce como idf (*inverse document frequency*). Así, para calcular el peso de cada elemento del vector que representa al documento, se tiene en cuenta la frecuencia inversa del término en la colección, multiplicándola por la frecuencia del término dentro de cada documento (Harman, 1992), esto es:

$$2.\quad w_{ij} = tf_i \cdot idf_j$$

Al respecto, Salton y Buckley (1988) experimentaron con más de 200 sistemas de cálculo de pesos, pero uno de los más utilizados viene dado en la Ecuación 3, que expresa el peso del término j en el documento i.

$$3.\quad w_{ij} = tf_{ij} \cdot \log \frac{N}{df_j}$$

Donde df_j es el número de documentos en que aparece el término j, y N el número de documentos de la colección.

Una aplicación de este proceso realizado para los documentos es el utilizado por los sistemas automatizados de recuperación o clasificación de documentos (bases de datos, por ejemplo) en el que los usuarios realizan consultas en lenguaje natural. Estas consultas pueden considerarse como un documento más, seguramente bastante breve, aunque no siempre. Así pues, el mecanismo de obtención de pesos también se aplica a las consultas, para de esta manera poder disponer de representaciones vectoriales homogéneas de consultas y documentos, que posibiliten obtener el grado de similitud entre ambos documentos, representados como vectores en un espacio multidimensional.

La resolución de la consulta consiste en establecer el grado de semejanza existente entre el vector que representa a la consulta y el vector que representa a cada uno de los documentos de la colección. Para una consulta determinada, cada documento arrojará un grado de similitud determinado; aquellos cuyo grado de similitud sean más elevado se ajustarán mejor a las

necesidades expresadas en la consulta, desde el punto de vista del sistema de recuperación o clasificación de información. No obstante, es el usuario el que debe decidir la relevancia de los documentos recuperados, siendo esta una característica totalmente subjetiva del mismo (Manning & Schütze, 2003).

El modo más simple de calcular la similitud entre una consulta y un documento, utilizando el modelo vectorial, es realizar el producto escalar de los vectores que los representan (Ecuación 4). En la siguiente ecuación se incluye la normalización de los vectores, a fin de obviar distorsiones producidas por los diferentes tamaños de los documentos. El índice de similitud más utilizado es el coseno del ángulo formado por ambos vectores. Para una consulta Q, el índice de similitud con un documento D_i es:

$$4. \quad simil(Q, D_i) = \frac{\sum_{j=1}^{m} p_j d_{ij}}{\sqrt{\sum_{j=1}^{m} p_j^2 \cdot \sum_{j=1}^{m} p_{ij}^2}}$$

Cabe señalar, que existen otros métodos propuestos para calcular la similitud. Algunos ejemplos son: el coeficiente de emparejamiento (*matching coefficient*), el coeficiente de Dice, el coeficiente de Jaccard (o Tanimoto) y el coeficiente de solapamiento (*overlap coefficient*). Una síntesis con la descripción de estas medidas de similitud puede encontrarse en Jurafsky y Martin (2000), Manning y Schütze (2003), Tzoukermann, Klavans y Strzalkowski (2003).

Ahora bien, dado que en nuestro caso nos interesa describir el Corpus PUCV-2006 e identificar las similitudes y diferencias entre las disciplinas consideradas en este corpus, compararemos dos métodos de clasificación supervisada, basados en el modelo vectorial: el Bayes Ingenuo (*Multinomial Naïve Bayes*) y la Máquina de Soporte de Vectores (*Support Vector Machine*), con el fin de conocer cuál de ellas es más eficiente en la clasificación de los textos.

Hemos elegido estos métodos, porque, si bien son comunes en la clasificación de documentos, aún no han sido suficientemente probados en corpus textuales en español (algunas excepciones son Figuerola et al., 2000; Cerviño et al., 2004). Además, porque con estos métodos no se requiere necesariamente un marcaje estructural de los textos de ningún tipo, es decir, los textos no deben ser etiquetados morfosintácticamente; por lo tanto, tienen menor exigencia en cuanto a costo computacional y se pueden implementar más fácilmente en programas de clasificación automática, que sistemas basados en representaciones simbólicas guiadas por la sintaxis (Jurafsky & Martin, 2000).

En el siguiente apartado, explicaremos brevemente las características principales de cada método. Cabe comentar que debido al espacio y focalización de este artículo no profundi-

zaremos en todos los aspectos referentes a los procedimientos y cálculos estadísticos que conciernen a cada uno de ellos. De este modo, para mayor información se recomiendan las siguientes fuentes: para Bayes Ingenuo (Johnson, 2000; Manning & Schütze, 2003; Molina & García, 2004) y para Máquinas de Soporte de Vectores (Cristianini & Shaw-Taylor, 2002; Hsu, Chang & Lin, 2003; Betancourt, 2005).

1.2.2. Clasificador Bayes Ingenuo (BI)

Los clasificadores bayesianos (Duda & Hart, 1973) son clasificadores estadísticos, que pueden predecir tanto las probabilidades del número de miembros de una clase, como la probabilidad de que una muestra dada pertenezca a una clase particular. Este tipo de clasificadores, basados en el teorema probabilístico de Bayes, han demostrado una alta exactitud y velocidad cuando se han aplicado a grandes bases de datos textuales, particularmente en español (Molina & García, 2004; Cerviño et al., 2004; Bordignon, Peri, Tolosa, Villa & Paoletti, 2004).

Diferentes estudios en los que se han comparado diversos algoritmos de clasificación han determinado que el clasificador BI es comparable en rendimiento a un árbol de decisión y a clasificadores de redes neuronales, procedimientos que son mucho más complejos computacionalmente (Molina & García, 2004).

A continuación, se explican los fundamentos de los clasificadores bayesianos y, más concretamente, del clasificador BI.

El objetivo de este método de aprendizaje matemático-estadístico es determinar cuál es la mejor hipótesis (la más probable) dado un conjunto de datos pre-existentes. Si denotamos P(D) como la probabilidad *a priori* de los datos y P(D|h) como la probabilidad de los datos dada una hipótesis, lo que queremos estimar es P(h|D), o sea, la probabilidad posterior de h dado ciertos datos conocidos, de aquí la noción de "probabilidad condicionada". Esto se puede estimar con el teorema de Bayes:

$$5.\ P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)}$$

Para estimar la hipótesis más probable (MAP, [*maximum a posteriori* hipótesis]) se busca el mayor P(h|D) como se muestra en la ecuación siguiente:

$$6.\ h_{MAP} = \arg\max_{h \in H}(P(h \mid D))$$

$$= \arg\max_{h \in H}\left(\frac{P(D \mid h)P(h)}{P(D)}\right)$$

$$= \arg\max_{h \in H}(P(D \mid h)P(h))$$

Ahora bien, como P(D) es una constante independiente de h, se asume que todas las hipótesis son igualmente probables, esto permite entonces concebir la hipótesis de máxima verosimilitud (ML, [*maximum likelihood*]) expresada en la Ecuación 7:

$$7.\ h_{ML} = \arg\max_{h \in H}(P(D \mid h))$$

De modo más particular, el clasificador bayesiano ingenuo se utiliza cuando se quiere clasificar un ejemplo descrito por un conjunto de atributos (ai's) en un conjunto finito de clases (V). Esto es clasificar un nuevo ejemplo de acuerdo con el valor más probable dado los valores de sus atributos. Así, si se aplica la Ecuación 7 al proceso de la clasificación, se obtendrá la Ecuación 8:

$$8.\ v_{MAP} = \arg\max_{v_j \in V}(P(v_j \mid a_1,\ldots,a_n))$$

$$= \arg\max_{v_j \in V}\left(\frac{P(a_1,\ldots,a_n \mid v_j)P(v_j)}{P(a_1,\ldots,a_n)}\right)$$

$$= \arg\max_{v_j \in V}\left(P(a_1,\ldots,a_n \mid v_j)P(v_j)\right)$$

Además, el clasificador BI asume que los valores de los atributos son condicionalmente independientes dado el valor de la clase, por lo que se hace cierta la Ecuación 9 y con ella la 10.

$$9.\ P(a_1,\ldots,a_n \mid v_j) = \prod_i P(a_{i|} \mid v_j)$$

$$10.\ P(v_j \mid a_1,\ldots,a_n) = P(v_j) \times \prod_i P(a_a \mid v_j)$$

En este sentido, con los clasificadores bayesianos ingenuos se asume que el efecto de un valor del atributo en una clase dada es independiente de los valores de los otros atributos. Esta suposición se llama "independencia condicional de clase" (Jurafsky & Martin, 2000; Molina & García, 2004; Bordignon et al., 2004). Ella permite simplificar los cálculos involucrados, siendo por esto que se le considera "ingenuo" (*naïve*) al método y, por lo mismo, sus resultados deben ser entendidos como una simplificación de la realidad.

1.2.3. La Máquina de Vectores de Soporte (MVS)

La Máquina de Vectores de Soporte es un método de clasificación de datos bastante nuevo, con el que diversos investigadores han conseguido buen desempeño de generalización sobre una amplia variedad de problemas de clasificación, destacando recientemente en problemas

de clasificación de textos. En relación a esto último, se le reconoce al método la capacidad de minimizar el error de generalización, es decir, los errores del clasificador sobre nuevos documentos (Cortes & Vapnick, 1995; Hsu, Chan & Lin, 2003; Baldi, Fresconi & Smyth, 2003; Téllez, 2005).

Particularmente la MVS es apropiada para trabajar con datos multidimensionales, tales como representaciones de vectores en un espacio de documentos textuales. En su formulación estándar se trabaja con problemas de clasificación binaria donde el número de clases es restringido a dos, aunque también se puede utilizar para la clasificación multiclases, a través de la reducción del problema de clasificación a sub-problemas de orden binario (Cortes & Vapnick, 1995).

Como ya hemos visto, una tarea normal de clasificación de textos involucra datos para entrenamiento y datos para prueba de un algoritmo a partir de características cuantificables de los textos. Cada unidad textual en el grupo de entrenamiento contiene un "valor de clasificación", designado por una etiqueta de clase, y múltiples atributos o rasgos. El objetivo de la MVS, entonces, es producir un modelo que permita predecir los valores de clasificación (identificar la clase) en la etapa de prueba conociendo solo los atributos (Baldi et al., 2003).

En términos geométricos, el problema que resuelve la MVS es identificar una frontera de decisión linear entre dos grupos, a través de una línea que los separe maximizando el espacio del hiperplano, de modo muy similar a lo que se realiza utilizando el análisis discriminante (Sharma, 1968; Hair, Anderson, Tatham & Black, 1999). Sin embargo, la MVS incluye una operación nueva llamada "truco de kernel", la que le permite realizar separaciones no lineales de los datos y con ello optimizar la clasificación de los mismos. Así, por ejemplo, en la Figura 1, observamos que en un espacio multidimensional las posibilidades de separación de las clases pueden ser múltiples, sin embargo, lo que se necesita es una separación óptima del hiperplano, sustentada por márgenes definidos, que en la práctica serán los vectores de soporte (ver Figura 2).

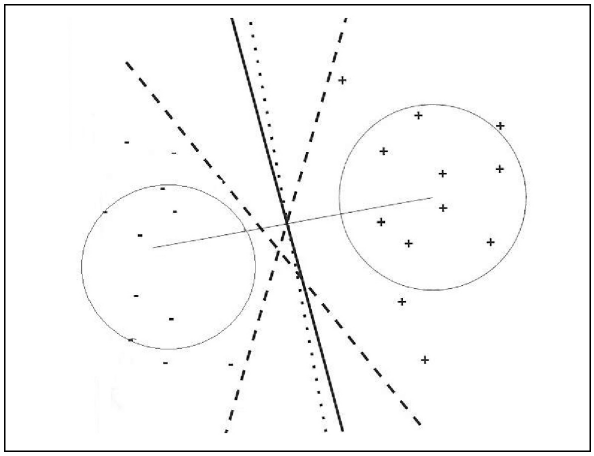

Figura 1. Fronteras de decisión lineal alternativas para un problema de clasificación binario (Baldi et al., 2003: 99).

En estas condiciones, una frontera de edición óptima (BI, por ejemplo) sería aquella que minimice la posterior probabilidad de que un nuevo punto sea mal clasificado y que esta frontera sea el hiperplano ortogonal al segmento que conecta los centros de masa de las dos distribuciones (línea punteada de la Figura 1). Claramente, un hiperplano azaroso que solo por casualidad separe los puntos de entrenamiento (línea segmentada en la Figura 1) puede estar significativamente lejos de una frontera de separación óptima, aportando una generalización muy pobre ante datos nuevos (Baldi et al., 2003), aumentando exponencialmente este problema en la medida en que aumentan la dimensionalidad del espacio de documentos.

Ante esta situación, Vapnick (2000) propone, en su teoría de aprendizaje estadístico, un hiperplano de separación óptima el cual tiene dos propiedades importantes: es único para cada grupo de datos separables linealmente, y el riesgo asociado de sobreestimación es más reducido que para cualquier otro hiperplano de separación. El margen de separación M del clasificador será la distancia entre el hiperplano de separación y el ejemplo de entrenamiento más cercano. De este modo, el hiperplano de separación óptimo es aquel que tenga el máximo margen. Para calcularlo se comienza con la determinación de la distancia de un punto x del hiperplano de separación (ver Figura 2).

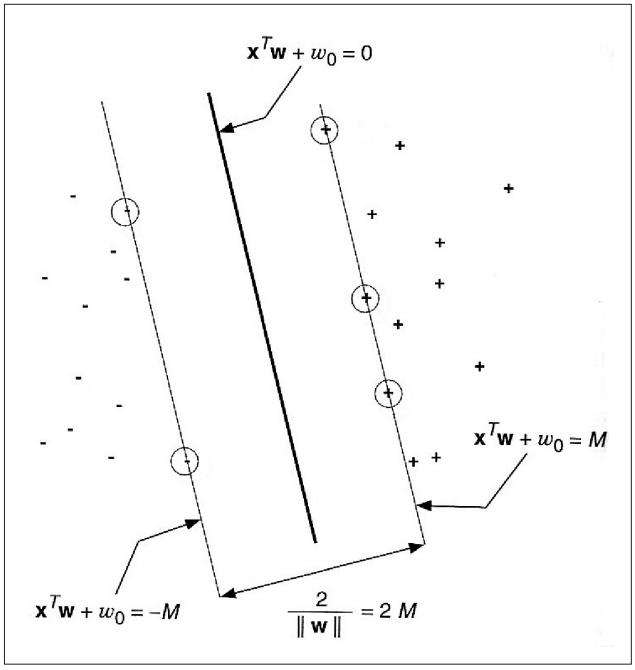

Figura 2. Ilustración del hiperplano de separación óptima y sus márgenes. Los datos en círculo indican los vectores de soporte (Baldi et al., 2003: 100).

Cabe señalar que la determinación final de la separación óptima del hiperespacio involucra una serie de pasos entre los que se considera la aplicación de la función de Langrage y el paquete de optimización estandarizada (para la solución del problema de programación cuadrática), el que debe satisfacer las condiciones de Karush-Kuhn-Tucker (Vapnick, 2000; Christianini & Shaw-Taylor, 2002; Baldi et al., 2003).

Ahora bien, la característica más relevante de la MVS es el uso de las funciones kernel (por ejemplo, lineal, polinominal, función de base radial, sigmoidal) para extender las aplicaciones de la determinación de separación óptima a casos no lineales (Christianini & Shaw-Taylor, 2002). Esto se hace traspasando los datos desde el espacio de entrada *X* a un amplio espacio de características χ mediante una función Φ, y resolviendo el problema de aprendizaje lineal en $\chi(\Phi: X \rightarrow \chi)$. La función real Φ no necesita ser conocida, es suficiente tener una función kernel *k* que calcule el producto interno en el espacio de características: $k(x, y)=\Phi(x)\cdot\Phi(y)$

(Christianini & Shaw-Taylor, 2002; Bautista, Guzmán & Figueroa, 2004). Una ilustración de esto último, es la que se presenta en la Figura 3.

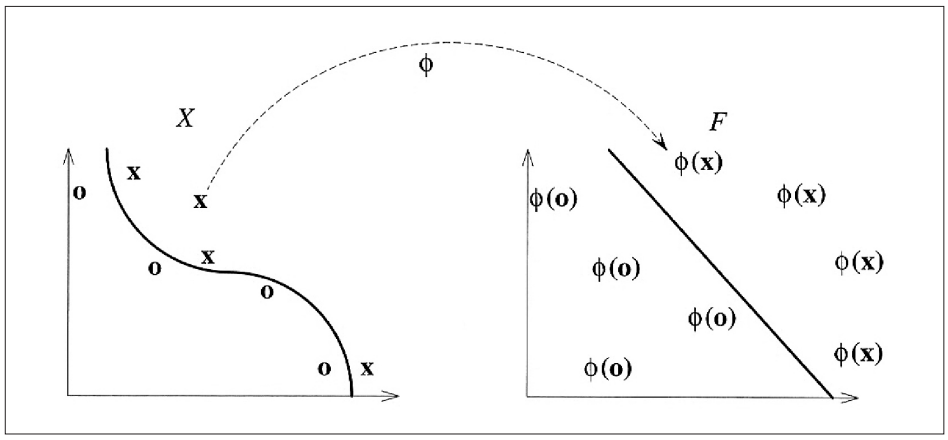

Figura 3. Traspaso de rasgos en el que se simplifica la tarea de clasificación (Christianini & Shaw-Taylor, 2002: 28).

Hasta ahora hemos abordado muy someramente las dos técnicas que serán comparadas en la tarea de clasificación textual que nos hemos impuesto, no obstante, cabe volver a comentar que cada una de ellas tiene una complejidad matemático-estadística que no nos es posible explicar en detalle en este trabajo. En lo que sigue, presentaremos la metodología de trabajo utilizada en la comparación de estos métodos con el fin de conocer cuál de ellos puede brindar resultados más efectivos en cuanto a la clasificación de los textos académicos a utilizar en esta investigación y, de esta forma, poder describir en términos más exhaustivos el Corpus Académico PUCV-2006.

2. Metodología

2.1. Corpus Académico PUCV-2006

Para el desarrollo de esta investigación, aplicamos ambos métodos de clasificación a una muestra del Corpus Académico PUCV-2006, construido en el marco del proyecto Fondecyt 1060440. Este corpus está constituido por casi el 100% de los textos que leen los alumnos de cuatro carreras universitarias de la Pontificia Universidad Católica de Valparaíso, Chile,

durante el desarrollo de sus estudios, a saber: Química Industrial, Ingeniería en Construcción, Trabajo Social y Psicología. Las dos primeras carreras son representantes del área de las Ciencias Básicas y de la Ingeniería y las dos siguientes representantes de las Ciencias Sociales y Humanas. De este modo, el corpus se compone de 652 textos, los que alcanzan a un total de 96.288.874 palabras (ver Parodi, 2007). Para esta investigación, en particular, se ha utilizado solo una parte del corpus, pues a la fecha de la investigación el Corpus Académico PUCV-2006 aún se encontraba en construcción. En la Tabla 1 se comparan porcentualmente el número de textos y de palabras de cada una de las disciplinas del corpus con la muestra utilizada en esta investigación:

Tabla 1. Comparación porcentual del número de textos y palabras utilizados en la muestra de investigación respecto del corpus total.

Corpus académico		Textos			Palabras		
Áreas	**Carreras**	**PUCV_2006**	**Muestra**	**%**	**PUCV_2006**	**Muestra**	**%**
Ciencias Básicas y de la Ingeniería	**Química Industrial**	76	26	34,21	22.103.620	8.209.911	37,14
	Ingeniería en Construcción	92	31	33,70	14.653.760	4.937.665	33,70
Ciencias Sociales y Humanas	**Trabajo Social**	188	64	34,04	29.394.904	7.601.999	25,86
	Psicología	296	95	32,09	30.136.590	10.136.506	33,64
Total		**652**	**216**	33,13	**96.288.874**	**30.886.081**	**32,08**

Como puede observarse en la Tabla 1, la muestra utilizada para esta investigación corresponde al 33,13% de los textos del corpus original y al 32,08% del total de palabras originales, es decir, se trabaja en esta investigación con cerca de un tercio del corpus total. Cabe señalar que la diferencia de porcentajes en este primer momento de descripción de la muestra no afecta la representatividad de los resultados, pues como ya se planteó en el apartado teórico, los datos utilizados en los métodos basados en el modelo vectorial están sujetos a procesos de normalización y ponderación estadística, lo que asegura la comparatividad de los datos.

2.2. Procedimientos metodológicos

Los procedimientos metodológicos realizados en esta investigación pueden ser agrupados en dos grandes etapas: preprocesamiento de los textos y aplicación de cada técnica de clasificación.

2.2.1. Preprocesamiento de los textos

En esta etapa, lo primero que se realizó fue construir un archivo de texto agrupando todos los textos correspondientes a cada carrera, esto se realizó utilizando un programa *ad hoc* construido en Perl (*cambiopala.pl*). De este modo se obtuvieron cuatro archivos de textos con la información textual de las carreras en estudio (QUI.txt, IC.txt, TS.txt, PSI.txt). Cabe señalar que durante este proceso a los textos se les eliminaron los lexemas de tipo funcional, es decir, aquellos que no aportan contenido semántico o *stop words* y cualquier información no lingüística. Luego, cada uno de estos archivos fue subido a la interfaz de etiquetaje e interrogación de corpus textuales El Grial (www.elgrial.cl). Esta herramienta se utilizó con el propósito de identificar con mayor exactitud, todos los sustantivos, verbos y adjetivos presentes en cada uno de los cuatro archivos.

Una vez identificados los lexemas de contenido semántico, se procedió a identificar cuáles de estos eran compartidos entre los cuatro archivos de textos. Este procedimiento se llevó a cabo utilizando otro programa construido en Perl (*batch.pl*). Cabe señalar que no se llevó a cabo ningún proceso de lematización o de extracción de raíces como lo plantean otras investigaciones (Figuerola et al, 2000; Cerviño et al, 2004). Esto último debido a que se proyecta simplificar el proceso de reconocimiento de las palabras para el procesamiento de clasificación en un futuro programa computacional, en este sentido el programa podría "leer y procesar" directamente los nuevos textos de entrada sin tener que llevar a cabo estos procesos intermedios.

Luego de identificados los lexemas de contenido semántico compartidos por los cuatro archivos de texto (2.729 lexemas), se procedió a construir una matriz en la que se consignó la frecuencia de cada uno de los lexemas compartidos por cada uno de los textos que componen los cuatro archivos, es decir, una matriz de 2.729 palabras por 222 textos. Cabe señalar que este procedimiento es muy relevante, pues se realiza bajo el supuesto de que los textos pueden diferenciarse en base al comportamiento (en términos de ocurrencias) de los lexemas compartidos. En la Figura 4 se esquematizan los procedimientos realizados en la etapa de preprocesamiento y los programas asociado a cada subetapa.

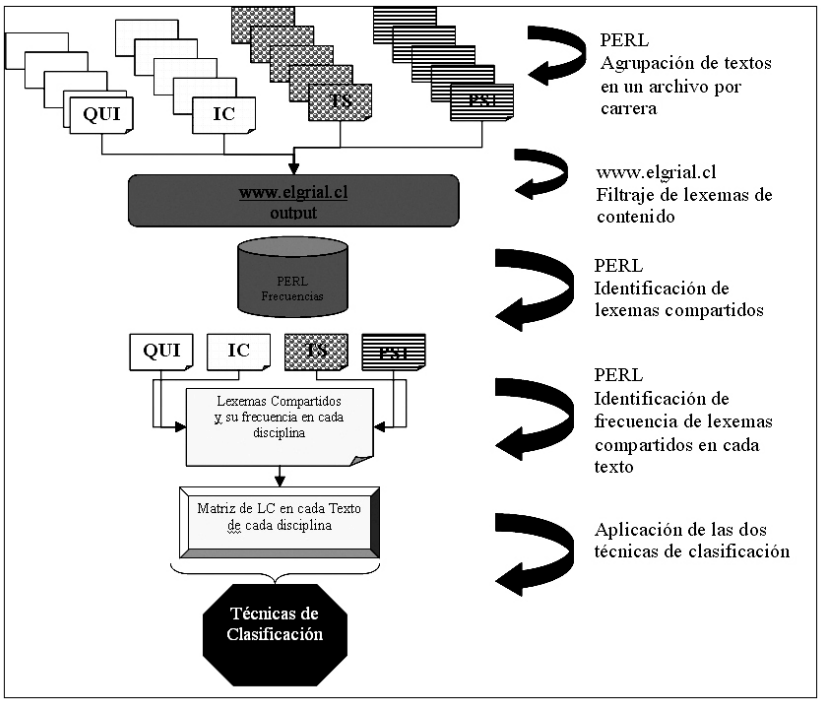

Figura 4. Preprocesamiento de los textos de la muestra para aplicación de las técnicas de clasificación.

2.2.2. Aplicación de las técnicas de clasificación

Como se mencionó más arriba, los métodos de clasificación a comparar en este estudio son: Bayes Ingenuo (BI) y Máquina de Soporte de Vectores (MVS). Ambos métodos requieren, en primera instancia, normalizar los datos de la matriz (esto es, la frecuencia en la que los 2.729 lexemas compartidos aparecen en cada uno de los 222 textos) como requisito para la construcción de los espacios vectoriales correspondientes. En general, y como mencionáramos anteriormente, existen más de 200 sistemas de cálculo de pesos para normalizar los datos. En este caso particular seguimos el modelo de Salton (1968), denominado TFC, donde:

T = frecuencia bruta del término (número de veces en que el término aparece en un documento).

F = LOG(*N*/*n*), esto es, multiplicación del factor original T por un factor de frecuencia inversa de la colección (*N* es el número total de documentos en la colección y *n* es el número de documentos a los cuales un término le es asignado).

C = normalización por coseno, donde cada peso de un término *w* es dividido por un factor que representa el largo de vector euclideano (*Euclidean vector length*).

Una vez concluido este proceso, se obtiene una matriz con las mismas dimensiones que la matriz original, pero con los términos ponderados según su aporte informativo.

Ahora bien, dado que el tamaño de la matriz dificulta el proceso estadístico y computacional, esta debe reducirse a una matriz representativa de la original, para ello se utilizó en el caso del método BI dos técnicas combinadas, las que como resultado, por una parte, entregaran las variables (lexemas de contenido semántico) que permitirán clasificar los documentos y, por otra, desecharán las variables que no aportan nada a la identificación de las clases (en este caso la identificación de cuatro disciplinas a partir de los textos académicos). En este sentido se usó la determinación de valor umbral para la frecuencia de documento (*document frequency thresholding*, DF) y la ganancia informativa (*information gain*, IG) (ver Yang & Pedersen, 1997; Manning & Schütze, 2003). De este modo, aplicando DF e IG se obtuvieron 74 variables útiles para nuestra tarea de clasificación. En el caso de la MVS, solo se requirió utilizar el criterio IG, obteniéndose 515 variables para la clasificación. Además, se utilizó el kernel denominado Función de Base Radial (*Radial Basis Function*, RBF). Cabe señalar, que se ha seleccionado este kernel, pues, en general, se le reconoce una alta eficiencia en la etapa de entrenamiento (Colmenares, 2007).

El siguiente procedimiento, una vez identificados los lexemas de contenido semántico más apropiados para la clasificación de los textos, fue calcular, según cada método, los valores de clasificación. Esto se realizó en tres etapas: a) etapa de entrenamiento, b) etapa de prueba y c) etapa de clasificación general. Para la etapa de entrenamiento se utilizó el 80% de los textos de cada disciplina y, para la de prueba, el restante 20% de los textos.

Posteriormente, se midió la efectividad de las operaciones de clasificación de cada uno de los métodos, utilizando las medidas clásicas de exhaustividad y precisión (Figuerola et al., 2000). Para realizar esto último se utilizan las Ecuaciones 11 y 12:

$$11.\ R = \frac{a}{a+c} \qquad 12.\ P = \frac{a}{a+b}$$

Donde:

- *R* es la exhaustividad
- *P* es la precisión
- *a* es el número de documentos pertenecientes a una clase y adscritos a esa clase (clasificados correctamente)
- *b* es el número de documentos no pertenecientes a una clase pero asignados a esa clase (clasificados incorrectamente)
- *c* es el número de documentos pertenecientes a una clase no asignados a esa clase.

Adicionalmente, se utiliza una medida, también común en trabajos de clasificación, que unifica en los resultados la precisión y la exhaustividad. Esta medida es conocida como medida F (F_β) y se expresa como:

$$13.\ F_\beta = \frac{\left(\beta^2 + 1\right)PR}{\beta^2 P + R}$$

Donde β es un parámetro que permite ajustar la influencia relativa de ambos componentes, precisión y exhaustividad. De este modo, β=1 proporciona igual peso a ambos componentes de la medida. Si β es mayor que uno, la precisión se ve favorecida y si β menor que 1 se ve favorecida la exhaustividad (Manning & Schütze, 2003).

Por último se debe destacar que para los cálculos realizados en ambos métodos se utilizó el programa Statistica. Este programa permite llevar a cabo análisis, administración y visualización de datos. Además, permite llevar a cabo procedimientos de minería de datos e incluye una amplia selección de técnicas de modelamiento, agrupamiento, exploración y clasificación (para mayor información ver http://www.statsoft.com/products/products.htm).

3. Resultados y conclusiones

A continuación, presentaremos los resultados obtenidos acorde con lo planteado en la metodología de trabajo. Para ello, daremos a conocer primero los resultados de la fase de entrenamiento, luego los resultados obtenidos en la fase de prueba de clasificación; más adelante, el resultado de la clasificación considerando todos los textos y, por último, las medidas de evaluación de los métodos.

3.1. Fase de entrenamiento

En primer lugar, presentamos los resultados obtenidos utilizando el método BI. De los 216 textos, se seleccionó el 80% de ellos para la fase de entrenamiento, esto es, 172 textos, y luego se calculó el porcentaje de asignación correcta para cada disciplina.

La Tabla 2 muestra tanto el porcentaje de clasificación (%) como el número de textos clasificados (#) utilizando el método BI, considerando los valores para los 74 lexemas de contenido, considerados por el método como discriminantes para las cuatro áreas.

Tabla 2. Porcentaje de clasificación de textos utilizando BI, en la fase de entrenamiento.

%(#)	IC	QUI	TS	PSI	Total
IC	91,67 (22)	8,33 (2)	0,00 (0)	0,00 (0)	100,00 (24)
QUI	0,00 (0)	100,00 (22)	0,00 (0)	0,00 (0)	100,00 (22)
TS	5,56 (3)	0,00 (0)	83,33 (45)	11,11 (6)	100,00 (54)
PSI	4,17 (3)	25,00 (18)	12,50 (9)	58,33 (42)	100,00 (72)
					(172)

Como se puede observar, en esta fase, los textos académicos correspondientes a Química son clasificados correctamente en un 100% de los casos. Los textos de Ingeniería en Construcción son clasificados correctamente en un 91,67%, errando la clasificación en solo 2 textos. Para el caso de Trabajo Social, el método en esta fase alcanza una clasificación correspondiente al 83,3% de los textos académicos, clasificando erróneamente 9 textos: 6 (11,11%) los ubica en Psicología y 3 (5,56%) en Ingeniería en Construcción. Por último, para Psicología, el método presenta un resultado que alcanza al 58,33% de los textos, adjudicando a todas las otras disciplinas porcentajes variables de textos, destacándose Química, área en la que ubica erróneamente el 25% de los textos. Por último, el porcentaje de clasificación correcta para esta etapa considerando el número de aciertos dividido el total de textos de la fase (22+22+45+42/172) alcanza a un 77,33%.

Para el caso de la MVS, en la etapa de entrenamiento, los resultados porcentuales se presentan en la Tabla 3. En este método también se utilizó el 80% de los textos, sin embargo, en el procedimiento uno de los textos fue rechazado por el sistema de aprendizaje, es por ello que se consideraron 171 textos en esta fase. Para este método, luego de la reducción de variables, se consideró el valor ponderado de 515 lexemas de contenido semántico, variables con las cuales se realizaron las clasificaciones.

Tabla 3. Porcentaje de clasificación de textos utilizando la MVS en la fase de entrenamiento.

%(#)	IC	QUI	TS	PSI	Total
IC	100,00 (24)	0,00 (0)	0,00 (0)	0,00 (0)	100,00 (24)
QUI	4,55 (1)	95,45 (21)	0,00 (0)	0,00 (0)	100,00 (22)
TS	0,00 (0)	0,00 (0)	98,15 (53)	1,85 (1)	100,00 (54)
PSI	0,00 (0)	0,00 (0)	1,41 (1)	98,59 (70)	100,00 (71)
					(171)

Observamos que los textos académicos de Ingeniería en Construcción son clasificados correctamente en un 100% de los casos. En Química 95,45% de los textos son clasificados correctamente, en tanto que solo un texto es clasificado en Ingeniería. Notamos que en estas dos áreas el porcentaje de clasificación es muy bueno con ambos métodos, sin embargo, el orden de acierto entre estas dos disciplinas es inverso (BI clasifica mejor Química y MVS, Ingeniería). En el caso de Trabajo Social, el porcentaje de clasificación acertada es de 98,15%, lo que supera en un 15% a la clasificación realizada por BI. Para Psicología el porcentaje de acierto también es muy alto, alcanzando un 98,59%. En esta disciplina la diferencia entre ambos métodos es sustancial, en tanto que MVS supera en esta área en más de un 40% al método BI. Si calculamos el porcentaje de clasificación acertada general, este alcanza al 98,83%, superando en 21,5 puntos porcentuales al método a BI.

En suma, podemos establecer que, a la luz de los datos presentados en esta fase, el método MVS clasifica mejor los textos académicos de las cuatro disciplinas que el método BI.

3.2. Fase de prueba

En lo que sigue, se presentan comparativamente los resultados obtenidos por ambos métodos en la fase de prueba de los algoritmos entrenados en la etapa anterior. Los porcentajes indican el nivel de predicción en que un texto académico nuevo, no incluido en la etapa de entrenamiento (20% del total de los textos académicos por área), es clasificado correctamente.

En la Tabla 4 se presentan los porcentajes de clasificación obtenidos utilizando el método BI.

Tabla 4. Porcentajes de clasificación utilizando el método BI en la fase de prueba.

%(#)	IC	QUI	TS	PSI	Total
IC	57,14 (4)	14,29 (1)	0,00 (0)	28,57 (2)	100,00 (7)
QUI	0,00 (0)	100,00 (4)	0,00 (0)	0,00 (0)	100,00 (4)
TS	0,00 (0)	0,00 (0)	80,00 (8)	20,00 (2)	100,00 (10)
PSI	8,70 (2)	13,04 (3)	17,39 (4)	60,87 (14)	100,00 (23)
					(44)

Como es posible observar, en esta etapa predictiva el método BI clasifica correctamente todos los textos académicos de Química, al igual que en la etapa de entrenamiento. En Ingeniería el porcentaje de clasificación adecuada baja considerablemente en relación a la etapa previa (34,5% de diferencia). En Trabajo Social, el porcentaje de textos académicos correctamente clasificados alcanza al 80%, porcentaje muy similar al de la etapa anterior. En Psicología, se observa que el porcentaje de textos adecuadamente clasificados es de 60,87%, superando en cerca de un 2% el porcentaje de la etapa previa. Cabe mencionar que, así como en la etapa anterior, en esta área los textos académicos aparecen, según el BI, distribuidos en las otras áreas. El porcentaje general de clasificación adecuada según este método para los textos de las cuatro áreas es de un 68,18%, generándose una diferencia porcentual negativa de 9,1% con la etapa de entrenamiento.

En cuanto a la aplicación del método MVS, en su fase de prueba, cabe señalar que se debió eliminar un texto del grupo de textos de prueba, manteniéndose así el porcentaje correspondiente a un 20% del total de textos a analizar con MVS. De este modo, se trabaja con 43 textos nuevos para la etapa predictiva.

Tabla 5. Porcentajes de clasificación utilizando el método MVS en la fase de prueba.

%(#)	IC	QUI	TS	PSI	Total
IC	57,14 (4)	14,29 (1)	0,00 (0)	28,57 (2)	100,00 (7)
QUI	0,00 (0)	100,00 (4)	0,00 (0)	0,00 (0)	100,00 (4)
TS	0,00 (0)	0,00 (0)	40,00(4)	60,00 (6)	100,00(10)
PSI	0,00 (0)	0,00 (0)	4,55 (1)	95,45 (21)	100,00(22)
					(43)

Como vemos en la Tabla 5, todos los textos académicos del área de Química han sido clasificados correctamente, mejorando levemente (4,55%) respecto de la etapa de entrenamiento e igualando la clasificación realizada por el método BI. En cuanto a Ingeniería, el porcentaje

de clasificación bajó considerablemente respecto de la etapa previa (de un 100% a solo un 57,14%). Cabe señalar que MVS y BI clasifican exactamente en igual proporción porcentual y distribución los textos de Ingeniería en esta etapa. Lo que puede indicar una particularidad común en los textos de prueba de esta área que hacen que ambos algoritmos los clasifiquen de la misma manera. En Trabajo Social, el método MVS, clasificó erróneamente los textos de esta área (60% de los textos clasificados en Psicología). En cuanto a esta área el método BI, en la fase predictiva, superó ampliamente a MVS (80% vs. 40%, respectivamente). Para el área de Psicología, la MVS es capaz de predecir correctamente el 95,45% de los textos académicos, bajando levemente respecto del porcentaje de entrenamiento. Sin embargo, este porcentaje supera ampliamente el 60,87% mostrado por el método BI. El porcentaje general de clasificación de MVS en esta etapa alcanza a un 76,74%, siendo este porcentaje casi 23 puntos porcentuales más bajo que el porcentaje de la etapa entrenamiento; aunque supera en casi un 9% al método BI en la capacidad de predicción de textos académicos cuya "pertenencia disciplinar" es desconocida.

En síntesis, en la fase de prueba ambos método vieron reducida su capacidad de clasificación, sin embargo, MVS presentó un mejor desempeño general. Cabe señalar, que a pesar de lo anterior, este método no pudo clasificar adecuadamente los textos académicos correspondientes al área de Trabajo Social (confundiéndose la clasificación con el área de Psicología), en tanto que el método BI, sí clasificó efectivamente estos textos académicos.

3.3. Fase de clasificación general

En esta tercera fase se consideran la capacidad de clasificación general y la evaluación en términos de exhaustividad y precisión de ambos métodos, en relación a todos los textos pertenecientes al Corpus Académico PUCV-2006.

En la Tabla 6 se presentan los porcentajes de clasificación general de los textos académicos de las cuatro áreas disciplinares en investigación, obtenidos a través del método BI.

Tabla 6. Porcentajes de clasificación utilizando el método BI en la fase de clasificación general.

%(#)	IC	QUI	TS	PSI	Total
IC	83,87 (26)	9,68 (3)	0,00	6,45 (2)	100,00 (31)
QUI	0,00 (0)	100,00 (26)	0,00	0,00 (0)	100,00 (26)
TS	4,69 (3)	0,00	82,81 (53)	12,50 (8)	100,00 (64)
PSI	5,26 (5)	22,11 (21)	13,68 (13)	58,95 (56)	100,00 (95)
					(216)

Como es posible observar, el método BI mantiene, como en todas las etapas anteriores, una correcta clasificación de todos los textos académicos del área de Química. En el área de Ingeniería el porcentaje de clasificación baja respecto de la etapa de entrenamiento, pero mejora notablemente en comparación con la etapa de prueba (26,7% más de textos correctamente clasificados). En cuanto al área de Trabajo Social en esta etapa, se presenta un porcentaje de clasificación muy similar al obtenido en las dos etapas anteriores. En el área de Psicología, también el porcentaje de clasificación es muy similar al obtenido en las otras etapas, aunque levemente mejor que en la etapa de entrenamiento (casi un 1%) y levemente peor que en la etapa de prueba (casi un 2%). Cabe señalar que el área de Psicología es la más débil en términos de la clasificación de textos académicos, utilizando el método BI. Por último, en esta etapa el porcentaje de clasificación general corresponde a 74,54%. Este porcentaje supera levemente los porcentajes obtenidos en las dos etapas anteriores (1,21% respecto de la etapa de entrenamiento y 6,36% respecto de la etapa de prueba).

En relación con el método MVS en esta etapa de clasificación general de los textos, la Tabla 7 muestra los porcentajes obtenidos para cada una de las áreas. Cabe decir que para la ejecución de este método dos textos quedaron fuera (uno perteneciente a la etapa de entrenamiento y otro perteneciente a la etapa de prueba), por lo que se trabaja con 214 textos.

Tabla 7. Porcentajes de clasificación utilizando el método MVS en la fase de clasificación general.

%(#)	IC	QUI	TS	PSI	Total
IC	90,32 (28)	3,23 (1)	0,00 (0)	6,45 (2)	100,00 (31)
QUI	3,85 (1)	96,15 (25)	0,00 (0)	0,00 (0)	100,00 (26)
TS	0,00 (0)	0,00 (0)	89,06 (57)	10,94 (7)	100,00 (64)
PSI	0,00 (0)	0,00 (0)	2,15 (2)	97,85 (91)	100,00 (93)
					(214)

Como podemos notar, en Química, nuevamente obtenemos un alto porcentaje de clasificación correcta (96,15%), siendo este porcentaje muy similar al porcentaje de la etapa de entrenamiento y levemente menor al 100% obtenido en la etapa de prueba. En comparación con el método BI, este método clasifica los textos académicos del área de Química con cerca de 4% más de error, en esta etapa. En cuanto a Ingeniería, es de 90,32%, siendo este porcentaje levemente menor al de la etapa de entrenamiento, pero muchísimo mayor que el porcentaje obtenido en la etapa de prueba (33,18% mejor en la clasificación). Comparado con el método BI, el porcentaje de textos correctamente clasificados es mayor en un 6,45%. Para el área de Trabajo Social se presenta un porcentaje de 89,06% de corrección en la clasificación, siendo

este porcentaje menor en casi 9% al de la etapa de entrenamiento, aunque muchísimo mejor (incremento de un 49%) que el de la etapa de prueba. En relación con el método BI, este porcentaje también es superior en un 6,25%. En Psicología, el porcentaje de clasificación adecuada de los textos alcanza al 97,85%, porcentaje que es levemente menor que el de la etapa de entrenamiento (menos de un 1%) y algo mayor que el de la etapa de prueba (2,40%). En comparación con el método BI, el MVS clasifica un 38,9% mejor los textos académicos de esta área. En términos porcentuales de clasificación general para esta área, el método MVS obtiene un 93,93% de corrección en la clasificación. Este porcentaje es algo menor que el obtenido para la etapa de entrenamiento (cerca de 5%) y mayor que el de la etapa de prueba (17% más). En cuanto a la comparación con el método BI, el MVS clasifica mejor en un 19% los textos académicos en esta etapa.

A modo de síntesis de los resultados obtenidos hasta aquí, presentamos en el Gráfico 1 los porcentajes de textos académicos correctamente clasificados en cada una de las áreas, según cada método utilizado en cada una de las tres fases (_E= entrenamiento; _P= prueba; _G= general).

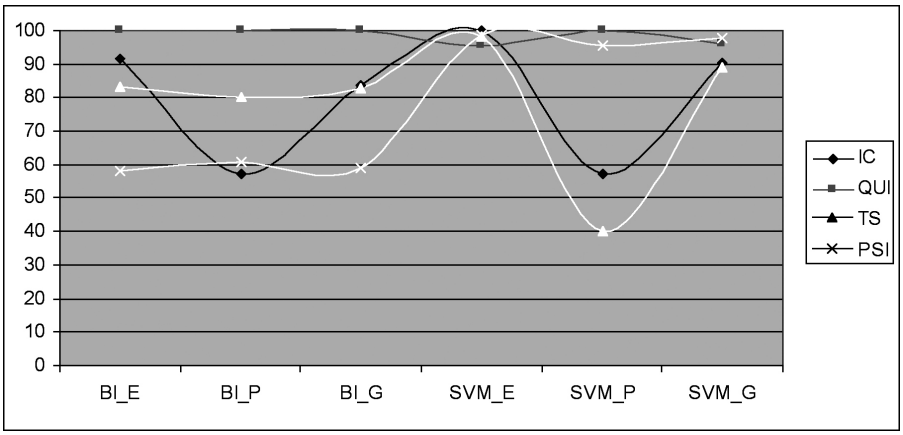

Gráfico 1. Comparación de métodos en cada una de sus fases.

En este gráfico lo primero que se destaca es la consistencia con la que los textos del área de Química son clasificados correctamente, independiente del método utilizado y la fase de clasificación. Esto nos permite aseverar que las características léxico-semánticas de los textos académicos de esta área, los distinguen significativamente de las otras tres áreas disciplinares. Observamos, también, que los textos académicos del área de Psicología son mejor clasi-

ficados por el método MVS que por el BI, en cada una de las tres fases. En cuanto a los textos del área de Trabajo Social, notamos que es el método MVS es el que mejor los clasifica, con excepción de la fase de prueba, en la que BI clasifica mejor; además, la clasificación presenta valores porcentuales muy estables en las tres fases. Para el área de Ingeniería, advertimos que ambos métodos presentan una fuerte variación entre las fases de clasificación, ya que, tanto en BI como en MVS, en la fase de prueba estos métodos clasifican pobremente los textos académicos de esta área (57,14%). Sin embargo, observamos también que el método MVS, en su fase de entrenamiento y general, logra clasificar muy bien los textos de esta área. En tanto el método BI, si bien presenta un alto porcentaje de clasificación en ambas fases, siempre se presenta un menor porcentaje que MVS.

Ahora bien, si nos concentramos en la fase de clasificación general (con todos los textos), se observa una notoria diferencia entre BI y MVS. El primer método, en esta fase, solo logra clasificar mejor los textos académicos de Química (área que, por lo demás, es en general muy estable). En cambio, el método MVS logra clasificar con altos porcentajes de acierto los textos académicos de las cuatro áreas disciplinares en investigación.

3.4. Evaluación de los métodos

En este apartado aplicamos las medidas de evaluación que nos permitan validar los resultados obtenidos por ambos métodos de clasificación. De este modo, como se mencionó anteriormente, se calcula la exhaustividad y precisión; así como también la medida F_β, que nos permite conocer la relación entre estos dos criterios (ver Tabla 8).

Tabla 8. Medidas de validación de los resultados obtenidos por ambos métodos.

Métodos	Bayes Ingenuo			Máquina de Soporte de Vectores		
Medidas de validación	**R**	**P**	F_β	**R**	**P**	F_β
ICC	0,84	0,76	0,8	0,9	0,97	0,93
QUI	1	0,52	0,68	0,96	0,96	0,96
TS	0,83	0,8	0,82	0,89	0,97	0,93
PSI	0,59	0,85	0,7	0,98	0,91	0,94
Promedio	**0,81**	**0,73**	**0,75**	**0,93**	**0,95**	**0,94**

En cuanto a la evaluación del método BI, acorde con las medidas de exhaustividad y precisión, podemos observar en la Tabla 8 que para las cuatro áreas en estudio la exhaustividad del método es mayor que la precisión, siendo esto confirmado por la medida F_β. Esto quiere

decir que, en general, este método tiende a extraer información relevante de los textos más que a entregar un alto grado de precisión en la clasificación.

Por su parte, para MVS se puede observar que el promedio de exhaustividad es levemente menor que el de precisión, siendo ambos mayores que los obtenidos por el método BI, esto implica que el método MVS clasifica de modo preciso y exhaustivo los textos académicos correspondientes a las disciplinas en estudio; además, la medida F nos permite confirmar el equilibrio existente entre la exhaustividad y la precisión para este método.

En consecuencia, con lo observado en los resultados y los métodos de evaluación propuestos, podemos decir, entonces, que el método MVS (con el kernel RBF) clasifica de mejor manera los textos del Corpus Académico PUCV-2006 que el método BI, acorde con un grupo de lexemas de contenido semántico, compartidos entre las áreas, cuya variación sistemática en términos de frecuencia normalizada caracteriza a los textos de cada una de las cuatro disciplinas en estudio.

A modo de conclusión de esta investigación, es posible señalar que a propósito de la noción de discurso académico, incluida en una noción mayor de discurso especializado, hemos integrado, o mejor dicho, puesto en interacción dos campos disciplinares, en apariencia muy disímiles, como son la lingüística y el procesamiento natural del lenguaje. Del primero destacamos la preocupación, cada vez más intensa, por la caracterización de los discursos de especialidad y académicos, a través de criterios descriptivos de orden funcional, pragmático y cognitivo, siendo uno de ellos el criterio léxico-semántico; del segundo, destacamos las posibilidades que nos entregan los métodos de clasificación automática de documentos, en particular, desde el tratamiento de los textos desde un perspectiva vectorial. Nos propusimos describir y clasificar los textos académicos utilizados en cuatro carreras universitarias de la Pontificia Universidad Católica de Valparaíso. Para cumplir con tal objetivo nos planteamos la tarea de implementar y comparar dos métodos de clasificación de documentos, el método BI y el método MVS. De este modo, como se explicó en el apartado de resultados, el método MVS resultó ser el que mejor clasificó los textos académicos de las cuatro carreras, obteniendo, además, altos valores tanto para la exhaustividad como para la precisión de la clasificación.

Además de lo dicho anteriormente, nos resulta interesante que los textos de Química, solo considerando sus lexemas de contenido semántico compartido con las otras áreas, se distinga fuertemente respecto de las otras disciplinas. Esto, sin duda, nos revela la importancia de la selección léxica condicionada por el área de especialidad. Así también, se desprende de este trabajo, considerando solo la fase general de MVS, que las áreas de las Ciencias Básicas y de la Ingeniería, tienden a compartir textos en menor cantidad que lo que lo hacen las áreas de las Ciencias Sociales y Humanas.

Por último, proyectamos, a partir de los métodos utilizados y sus correspondientes resultados, avanzar en la descripción de los textos académicos de estas cuatro disciplinas con mayor profundidad y detalle. De este modo, nos proponemos en un breve plazo utilizar el método MVS en la clasificación de textos académicos, ya no solo considerando el ámbito disciplinar sino que también el tipo textual. Esto nos permitirá conocer con mayor detalle el discurso académico, por una parte, e integrar con mayor énfasis el conocimiento lingüístico (en particular desde el análisis discursivo y textual) y los métodos y técnicas que nos proporciona el procesamiento natural del lenguaje, por otra. De esta forma, además, se busca aportar al desarrollo interdisciplinar de ambas perspectivas, a través de trabajos empíricos con grandes cantidades de textos, tal como lo hemos venido realizando en la Pontificia Universidad Católica de Valparaíso, desde hace ya varios años (Parodi & Venegas, 2004; Venegas, 2005; Cademartori, Parodi & Venegas, 2006; Parodi, 2006a, 2006b; Venegas, 2006).

REFERENCIAS BIBLIOGRÁFICAS

Baldi, P., Frasconi, P. & Smyth, P. (2003). *Modeling the Internet and the web*. Chichester: John Wiley.

Bautista, E., Guzmán, E. & Figueroa, J. (2004). Predicción de múltiples puntos de series utilizando support vector machines. *Computación y sistemas, 7*(3), 148-155.

Betancourt, G. (2005). Las máquinas de soporte vectorial. *Scientia e Técnica, 11*(27), 67-72.

Bordignon, F., Peri, J., Tolosa, G., Villa, D. & Paoletti, L. (2004). *Experimentos en clasificación automática de noticias en español utilizando el modelo bayesiano* [en línea]. Disponible en: http://www.unlu.edu.ar/~tyr/TYR-publica/paper-unlu-bayes-2004.doc

Cabré, M. (2002). Textos especializados y unidades de conocimiento: Metodología y tipologización. En J. García & M. Fuentes (Eds.), *Texto, terminología y traducción* (pp. 122-187). Barcelona: Almar.

Cademártori, Y., Parodi, G. & Venegas, R. (2006). El discurso escrito y especializado: Caracterización y funciones de las nominalizaciones en los manuales técnicos. *Literatura y Lingüística*, 17, 243-265.

Cerviño, U., García, J., Calvo, R. & Ceccatto, A. (2004). *Automatic classification of news articles in Spanish* [en línea]. Disponible en: http://citeseer.ist.psu.edu/beresi04automatic.html

Ciapuscio, G. (1994). *Tipos textuales*. Buenos Aires: EUDEBA.

Ciapuscio, G. (2000). Hacia una tipología del discurso especializado. *Discurso y Sociedad, 2*(2), 39-71.

Colmenares, G. (2007). *Función de base radial. Radial Basis Function (RBF)* [en línea]. Disponible en: http://www.webdelprofesor.ula.ve/economia/gcolmen/programa/redes_neuronales/capitulo4_funciones_bases_radiales.pdf

Cortes, C. & Vapnick, V. (1995). Support vector networks. *Machine Learning*, 20, 273-297.

Cristianini, N. & Shaw-Taylor, J. (2002). *Introduction to support vector machines: And other kernel-based learning methods*. Cambridge: University of Cambridge.

Duda, R. & Hart, P. (1973). *Pattern classification and scene analysis*. New York: Wiley.

Fairthorne, R. (1961). *The mathematics of the classification. Towards information retrieval*. London: Butterwoths.

Figuerola, C., Zazo, A. & Berrocal, J. (2000). *Categorización automática de documentos en español: Algunos resultados experimentales* [en línea]. Disponible en: http://imhotep.unizar.es/jbidi/jbidi2000/14_2000.pdf

Figuerola, C. (2000). La investigación sobre recuperación de la información en español. En E. Gonzalo & V. García (Eds.), *Documentación, terminología y traducción* (pp. 73-82). Madrid: Síntesis.

García, A. (2007). *Los procedimientos matemáticos en estudios de investigaciones lingüísticas. Utilidad y riesgos* [en línea]. Disponible en: http://angarmegia.275mb.com/riesgos_y_beneficios1.htm

Gläser, R. (1982). *The problem of style classification in LSP (ESP)*. Ponencia presentada en el 3rd European Symposium on LSP, Copenhague.

Gotti, M. (2003). *Specialized discourse. Linguistic features and changing conventions*. Bern: Peter Lang.

Hair, J., Anderson, R., Tatham, R. & Black, W. (1999). *Análisis multivariante*. Madrid: Prentice -Hall.

Halliday, M.A.K. & Martin, J.R. (1993). *Writing science: Literacy and discursive power*. London: Falmer.

Harman, D. (1992). Relevance feedback and other query modification techniques. En W. Frakes & R. Baeza-Yates (Eds.), *Information retrieval: Data structures and algorithms* (pp. 241-236). Englewood Cliffs, NJ: Prentice.

Hayes, R. (1963). *Mathematical models in information retrieval. Natural language and the computers*. New York: McGraw-Hill.

Hsu, Ch., Chang, Ch. & Lin, Ch. (2003). *A practical guide to support vector classification* [en línea]. Disponible en: http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf

Jeanneret, Y. (1994). *Écrire la science. Formes et enjeux de la vulgarisation*. Paris: PUF.

Johnson, D. (2000). *Métodos multivariados aplicados al análisis de datos*. México, DF: Thomson.

Jurafsky, D. & Martin, J. (2000). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. New Jersey: Prentice-Hall.

Landauer, T. (2002). On the computational basis of learning and cognition: Arguments from LSA. *The psychology of learning and motivation*, 41, 43-84.

López, C. (2002). Aproximaciones al análisis de los discursos profesionales. *Revista Signos*, *35*(51-52), 195-215.

Manning, C. & Schütze, H. (2003). *Foundations of statistical natural language processing*. Cambridge: MIT Press.

Martín-Valdivia, M., García-Vega, M. & Ureña-López, L. (2003). LVQ for text categorization using a multilingual linguistic resource. *Neurocomputing*, 55, 665-679.

Molina, J. & García, J. (2004). *Técnicas de análisis de datos en aplicaciones prácticas utilizando Microsoft Excel y Weka* [en línea]. Disponible en: http://galahad.plg.inf.uc3m.es/~docweb/ad/transparencias/apuntesAnalisisDatos.pdf

Parodi, G. & Venegas, R. (2004). BUCÓLICO: Aplicación computacional para el análisis de textos. Hacia un análisis de rasgos de la informatividad. *Lingüística y Literatura,* 15, 223-251.

Parodi, G. (2004). Textos de especialidad y comunidades discursivas técnico-profesionales: Una aproximación basada en corpus computarizado. *Estudios Filológicos*, 39, 7-36.

Parodi, G. (2005). Lingüística de corpus y análisis multidimensional: Exploración de la variación en el Corpus PUCV-2003: Una aproximación multiniveles. En G. Parodi (Ed.), *Discurso especializado e instituciones formadoras* (pp. 83-125). Valparaíso: Ediciones Universitarias de Valparaíso.

Parodi, G. (2006a). El Grial: Interfaz computacional para anotación e interrogación de corpus en español. *RLA*. *44*(2), 91-115.

Parodi, G. (2006b). Discurso especializado y lengua escrita: Foco y variación. *Estudios Filológicos*, 41, 165-204.

Parodi, G. (2007). El discurso especializado escrito en el ámbito universitario y profesional: Constitución de un corpus de estudio. *Revista Signos*, *40*(63) (en prensa).

Peronard, M. (1997). ¿Qué significa comprender un texto escrito? En M. Peronard, L. Gómez, G. Parodi & P. Núñez (Comps.), *Comprensión de textos escritos: De la teoría a la sala de clases* (pp. 55-78). Santiago: Andrés Bello.

Salton, G. & Buckley, C. (1988). Term-witghting approaches in automatic text retrieval. *Information Processing and Management*, *24*(5), 513-523.

Salton, G. & McGill, M. (1983). *Introduction to modern information retrieval*. New York: McGraw-Hill.

Salton, G. (1968). *Automatic information organization and retrieval*. New York: McGraw-Hill.

Schröder, H. (1991). Linguistic and text-theorical research on languages for special purposes. A thematic and bibliographical guide. En H. Schröder (Ed.), *Subject-oriented texts. Languages for special purposes and text theory* (pp. 1-48). Berlin: W. de Gruyter.

Sharma, S. (1996). *Applied Multivariate Techniques*. New York: Wiley.

Téllez, A. (2005). *Extracción de información con algoritmos de clasificación* [en línea]. Disponible en: http://ccc.inaoep.mx/~mmontesg/tesis%20estudiantes/TesisMaestria-AlbertoTellez.pdf

Tzoukermann, E., Klavans, J. & Strzalkowski, T. (2003). Information retrieval. En R. Mitkov (Ed.), *The Oxford handbook of computational linguistics* (pp. 530-544). New York: Oxford University Press.

Vapnick, V. (2000). *The nature of statistical learning theory*. New York: Springer.

Venegas, R. (2005). *Las relaciones léxico-semánticas en artículos de investigación científica: Una aproximación desde el análisis semántico latente*. Tesis doctoral, Pontificia Universidad Católica de Valparaíso, Chile.

Venegas, R. (2006). La similitud léxico-semántica en artículos de investigación científica en español: Una aproximación desde el Análisis Semántico Latente. *Revista Signos, 39*(60), 75-106.

Yang, Y. & Pedersen, J. (1997). *A comparative study on feature selection in text categorization* [en línea]. Disponible en: http://citeseer.ist.psu.edu/yang97comparative.html

Zazo, A., Figuerola, C., Alonso, J.L. & Gómez, R. (2002). *Recuperación de información utilizando el modelo vectorial* [en línea]. Disponible en: http://tejo.usal.es/inftec/2002/DPTOIA-IT-2002-006.pdf