Fecha: 11/07/2006

Desafíos en la clasificación automátizada utilizando sistemas de clasificación bibliotecarios

Kwan Yi
School of Library and Information Science University of Kentucky
USA

Traduccido por: Miren Lorea Elduayen Pereda

Jornadas:	**97. Bibliotecas Nacionales y Tecnología de la información con documentos audiovisuales y multimedia 97 (parte 2).**
Traducción simultánea:	**No**

WORLD LIBRARY AND INFORMATION CONGRESS: 72ND IFLA GENERAL CONFERENCE AND COUNCIL
20-24 August 2006, Seoul, Korea
http://www.ifla.org/IV/ifla72/index.htm

Resumen:
Los principales sistemas de clasificación bibliográfica se han utilizado durante mucho tiempo en el mundo bibliotecario tradicional para clasificar la informacion. Actualmente la clasificación de texto (TC, por sus siglas en inglés) se está convirtiendo en una herramienta popular y atractiva para organizar la información digital. Este artículo revisa estudios y proyectos previos sobre la TC basada en sistemas tradicionales de clasificación bibliotecaria. Resume también el debate sobre los retos que plantea la TC.

1.- INTRODUCCIÓN

El gran incremento en la cantidad de información y de fondos digitales disponibles y la demanda de herramientas de recuperación para gestionar la sobreabundancia de información ha devuelto el interés por las tareas de clasificación automática con la esperanza de reducir el trabajo de manera significativa o incluso de suplirlo parcialmente. Ha habido algunos proyectos y estudios de investigación sobre la viabilidad de la Clasificación de la Library of Congress (LCC) y la Clasificación Decimal de Dewey (DDC) como referencias para la clasificación automática de la información digital.

Una buena aproximación para la organización de la información es clasificar la información recogida de acuerdo a un conjunto predefinido de categorías y recuperar la información relevante buscando a través de la lista de categorías utilizadas. Esta es la forma tradicional de clasificar y localizar ítems bibliotecarios basados en los sistemas de clasificación bibliotecaria. Lo antiguo se revitalizó en el entorno digital con la popularidad de los directorios por materias y los directorios Web. La falta de unos sistemas de clasificación autorizados (completos), sin embargo, se presenta como un problema . La adopción de un sistema de

clasificación bibliotecario tiene un fuerte potencial para llenar este vacío ya que vienen respaldados por su popularidad práctica en las tareas de clasificación tradicional y por su base teórica y sistemática. Ante este problema parece que se está abriendo un campo prometedor de investigación sobre la organización y la recuperación de la información digital.

La situación ha madurado lo suficiente para alcanzar un punto donde es apropiado evaluar los progresos adquiridos en el orden del día de la clasificación automática e identificar los desafíos actuales para establecer la agenda de investigación de los próximos años ya que ha aparecido un amplio rango de estudios y aplicaciones, gran cantidad de proyectos de bibliotecas digitales acabados y no acabados y además se ha aplicado con éxito el aprendizaje automático en el campo de la recuperación de la información.

La sección 2 describe algunos antecedentes en la Clasificación de Textos. La sección 3 describe una panorámica de estudios recientes y proyectos en TC utilizando sistemas de clasificación bibliotecaria. La sección 4 ilustra los retos actuales en la organización de la información automatizada y la sección 5 nos muestra las conclusiones.

2,- ANTECEDENTES.

2.1.- Comprendiendo la clasificación de textos.

La clasificación de textos, surgió como un campo relativamente nuevo que dio un giro a la investigación en el campo de la recuperación de la información. Clasifica los documentos con un conjunto predefinido de categorías sin intervención humana. La tarea es bastante similar a una subtarea de catalogación de materias en bibliotecas tradicionales, pero es más característica en clasificación automática . La Clasificación de Texto (TC) ha llegado a ser mucho más atractiva que nunca ya que la necesidad de herramientas de organización de la información para hacer frente a la vasta cantidad de información digital es cada vez más acuciante.

TC es la actividad de etiquetar textos en lenguaje natural con categorías temáticas tomadas desde un conjunto previamente definido. Hay 3 componentes principales implicados en el proceso de TC. El primero está compuesto por los objetos para ser clasificados, que son los documentos textuales. Sea D= [d1, d2,… , dn] un conjunto de documentos. El segundo componente son las categorías finales. Sea C= [c1, c2, ..., cn] el conjunto de categorías a estudiar. El tercer componente es un algoritmo de representación que actúa como clasificador. Un algoritmo de representación puede ser descrito como una función, tomando un documento como una entrada y produciendo una decisión binaria si el documento encaja en una categoría dada. La función es representada como F: cj [0, 1] donde $1<i<n$ y $1<j<m$. Si el resultado es 1 se interpreta que encaja en una categoría y si es 0 es interpretado como que no encaja en la categoría. Por lo tanto, la realización de una función de representación F y su calidad determinan la ejecución de TC como la función que sirve para medir la relevancia de un documento determinado dentro de una categoría.

Las tareas de TC pueden ser divididas en diferentes tipos, de acuerdo al número de categorías y al numero de etiquetas de categoría. Si hay solo dos clases a considerar, por ejemplo, el valor de m en el conjunto C es igual a 2, esa TC se dice que es una tarea de clasificación binaria. Con más de dos clases, por ejemplo, el valor de m es mayor de 2. Se dice que es una clasificación multicategoría. También si cada documento se asocia con una única etiqueta de

categoría, se llama una clasificación de etiqueta binaria. En clasificación multietiqueta, hay al menos una etiqueta de categoría asociada a cada documento.

2.2 Teoría del Aprendizaje automático en las aplicaciones de TC

Uno de los puntos principales que se está investigando en el campo de la clasificación automatizada es el proceso por el que la máquina (entendida como sistema informático) adquiere el conocimiento necesario para clasificar correctamente. En este campo el enfoque que está primando es el del aprendizaje automático (Machine learning, ML). El planteamiento general en ML consiste en suponer que el ordenador acumula conocimientos gracias a experiencias previas. El escenario de este proceso de aprendizaje automático se podría describir como un proceso sistemático compuesto de 4 elementos (Kubat, Bratleo y Michalski, 1999): experiencias, conocimiento base, algoritmo de aprendizaje y conocimiento final.
Las experiencias como elemento del marco experimental expresarían lo que hay que aprender con vistas al conocimiento que se quiere alcanzar, sería el conocimiento explícito.
El conocimiento base se refiere al conocimiento previo sobre la categoría final que se quiere alcanzar y sería el conocimiento implícito.
La caja negra del escenario del aprendizaje automatizado es el algoritmo de aprendizaje, que representa el método de adquisición del conocimiento.
El conocimiento final es la materialización de lo que un sistema de aprendizaje automatizado adquiere combinando el conocimiento explícito y el implícito.
Se puede utilizar como ejemplo el juego del ajedrez, considerándolo una tarea de aprendizaje. Un ejemplo de rutina de aprendizaje sería la secuencia de cambios de posición en el tablero y las reglas del ajedrez serían el *conocimiento base.* El enfoque del aprendizaje automatizado es similar hasta cierto punto al proceso de aprendizaje humano. Una maquina aprende sobre áreas temáticas o categorias a través de documentos preseleccionados por especialistas en la materia, de la misma manera que nosotros aprendemos leyendo. La colección de documentos preseleccionados se denomina conjunto de ensayo y puede ser implícito o explícito.
Las técnicas de aprendizaje automatizado (ML) se han aplicado a la clasificación de diferentes tipos de documentos: documentos sanitarios (Larkey y Croft, 1996) datos sobre flora (Cui, Heidorn y Zhang, 2002), documentos legales (Thompson, 2001) y documentos web (Chakrabarti, Dom e Indyk, 1998). También se ha investigado en categorías distintas a las materías o las áreas de conocimiento, como en tipos genéricos de documentos: en editoriales, informes, reseñas, artículos de investigación y homepages (Lee y Myaens, 2002), en trabajos de carrera sobre diversas asignaturas (Larkey, 1998), o en filtros para correos-spam (Hidalgo, López y Sanz, 2000).

2.3 Ámbito de la clasificación automatizada

El debate sobre Clasificación de texto se circunscribe a los siguientes puntos:

- Las técnicas de Clasificación de texto y las de Clustering tienen mecanismos parecidos pero clasifican de forma diferente. El concepto de Clustering se parece a TC en que agrupa documentos parecidos. Un clustering se define como: “el grupo de documentos que cumple con un conjunto de propiedades comunes” (Baeza-Yates y Ribeiro-Nieto, 1999). En el clustering no se parte de un conjunto explícito de categorías sino que se buscan características comunes que no se conocían de antemano, mientras que en TC se mide el grado de parecido de un documento con una categoría para evaluar la relevancia del documento dentro de la categoría. La técnica de clustering no evalua los documentos similares según categorías sino respecto a otros documentos.
- Las características no-textuales (externas al texto) no se tienen en cuenta.

- La TC es una clasificación que se basa en el contenido y no en los metadatos o en información estructurada. TC y la clasificación de documentos de la web se caracterizan por utilizar metadatos que no se basan en el contenido como pueden ser los hipervínculos y los datos estructurados.

3. PANORAMA DE LOS ESTUDIOS Y PROYECTOS DE CLASIFICACIÓN AUTOMATIZADA QUE UTILIZAN SISTEMAS DE CLASIFICACIÓN BIBLIOTECARIOS

Vamos a efectuar un repaso a los trabajos más recientes efectuados en el campo de la clasificacion automatizada de documentos digitales que utilizan alguno de los sistemas de clasificación bibliotecarios más importantes.
Uno de los primeros trabajos en este campo puede encontrarse en Larson, 1992. Un conjunto de registros MARC se clasificó en la categoría Z (Bibliografía y Biblioteconomía) de la LCC, basándose en los encabezamientos de título y de materia. Se utilizaron 30471 registros MARC y 286 para comprobaciones. Este trabajo pretendía ayudar a los bibliotecarios a determinar entradas relevantes para elementos no clasificados anteriormente gracias a que proporcionaba una lista de entradas potenciales basadas en encabezamientos de título y de materia. . Existe un trabajo reciente basado en Larson que puede estudiarse en (Frank and Paynter, 2004). Este proyecto asigna encabezamientos de la LCC a los metadatos de recursos de Internet utilizando LCC y los encabezamientos de materia de la LC (LCSH, Library of Congress Subject Headings). 800000 registros se utilizaron como prueba. Un conjunto independiente de 50.000 registros sirvió de tester. La pertinencia de este sistema oscila entre un 55% y un 80%
Los siguientes puntos repasan varios proyectos de clasificación automatizada en los que se adoptó como base algún sistema de clasificación tradicional.

3. 1 Pharos

Pharos es un prototipo de arquitectura de la información que deriva del proyecto Alexandria Digital Library (Dolin, Agrawal and El Abbadi, 1999). Aúna fuentes heterogéneas tanto en contenido como en formato. Se puso en marcha un sistema de clasificación automatizada basado en la LCC con la idea de crear perfiles para información digital de carácter heterogéneo. Se utilizó la técnica Latent Semantic Indexing para los newsgroups y los registros de la LCC. 1,5 millones de registros de la Universidad de California, Santa Barbara, sirvieron de grupo de pruebas. Se extrajeron títulos, encabezamientos de materia y campos LCC. Dentro de de una colección concreta se consideraron los datos de titulo y materia como descripción de categoría y se les asignó una notación de la LCC. Esta relación entre la notación de la LCC y sus descriptores formaron los datos de prueba. Se clasificaron 7214 registros MARC de las 21 clases principales de la LCC Y los resultados experimentales arrojaron un resultado de una mediana entre +- 13.0 y 3.9 y un significado medio de 76 +- 19 para aproximadamente 4200 categorías de la LCC. Se realizó otro estudio preliminar utilizando artículos de más de 2500 newsgroups de USENET no se facilitaron datos sobre la pertinencia de la clasificación ya que se habían incluido artículos que no estaban previamente clasificados.

3.2 Scorpion

La Online Computer Library Center (OCLC) puso en marcha entre 1996 y 1999 el proyecto Scorpion para desarrollar un método automatizado de identificar categorías DDC en obras digitales (Shafer, 2001) para ello utilizaba un método de clustering. Consiste en ponderar las

semejanzas entre un documento nuevo y los clusters previamente definidos siguiendo las clases de la DDC. Se considera el cluster más parecido como el más adecuado para el documento nuevo. Un contador de términos mide las coincidencias. Un grupo de registros bibliográficos sobre recursos de Internet a los que se había asignado manualmente materias de la DDC se utilizó para evaluar los resultados. Los resultados no se publicaron, probablemente porque la comparación no podía ser efectuada correctamenta ya que la indización tradicional se basó unicamente en las frases que describían recursos de Internet. Las conclusiones demostraron que la clasificación automatizada no puede ocupar el lugar de la humana pero si que puede proporcionar una solución rentable para ayudar a los clasificadores

3.3 DESIRE (Kock and Ardö, 2000)

El proyecto DESIRE comenzó en 1996, se trata de un proyecto internacional a gran escala financiado por la Unión Europea para construir un portal temático de recursos de ingeniería. Como prueba los documentos web se clasificaron automaticamente con la clasificación Engineering Information (EI) que se basaba en la coincidencia de términos simples. En la evaluación del sistema con unas 1000 páginas web la pertinencia de la clasificación se comparó con las decisiones del equipo de clasificación. Sobre todo se informó del hecho de que cerca del 60 % de las clasificaciones encajaban mucho mejor con las decisiones humanas. Los mismos datos sobre ingeniería se clasificaron utilizando la DDC con la colaboración de OCLC. En concreto se añadieron algunos encabezamientos LCSH al texto completo. No se informó sobre la clasificación con DDC

3.4 Wolverhampton Web Library (Jenkins et al)

El proyeto Wolverhampton Web Library (WWLib)1 consiste en un motor de búsqueda de documentos web producidos en Gran Bretaña que utiliza la DDC para clasificarlos. Una característica interesante de WWLib es que considera una página web como un ejemplar de una biblioteca y elabora registros catalográficos que describen la información incluyendo el título, el Universal Resource Locator (URL), la clasificación DDC y una descripción de las páginas web recopiladas. Los motores de búsqueda de la web presentan por lo general los resultados en el orden establecido por los usuarios, WWLib por el contrario ordena las páginas web según la DDC. El componente Classifier realiza automaticamente el proceso de clasificar los documentos web basándose simplemente en la coincidencia de términos. Classifier compara una serie de términos extraidos de los documentos y las categorías DDC (Wallis & Burden, 1995). Las palabras que aparecen en los documentos web son “pesadas” según las etiquetas que se les asignan, y se aplica una técnica de stemming. También se tiene en cuenta un método para la relevancia de un documento tanto para una categoría como para la superior para aprovechar las ventajas de la estructura jerárquica de la DDC. La última versión de WWLib está estudiando incluir un conjunto más amplio de categorías de la DDC que incluyan sinónimos. No parece que se haya llevado a cabo un estudio formal para medir la actuación del sistema pero si se publicó un estudio informal con 17 URLs elegidas al azar (WWLib); en este estudio se concluía que 13 de las 17 eran relevantes, pero no se dieron a conocer más detalles del procedimiento seguido como los métodos de evaluación y la selección de los datos.

3.5 Conclusión

La conclusión es que la mayoría de los proyectos de investigación descritos catalogaron y clasificaron automaticamente datos y páginas web pero no texto completo, al menos desde el punto de vista de la aplicación. Desde el punto de vista del sistema de clasificación, tanto la

LCC como la DDC, que son las clasificaciones más extendidas en Norte América, han sido utilizadas en aplicaciones para clasificación. No se explica claramente el razonamiento para elegir determinado tipo de clasificación. La elección de un sistema de clasificación bibliográfico parece estar basada en preferencias personales más que en las tareas a realizar o en la disponibilidad de los datos, al menos eso se desprende de los artículos publicados

4. RETOS PARA LA CLASIFICACIÓN DE TEXTOS EN LOS SISTEMAS DE CLASIFICACIÓN

El debate sobre los retos de TC está compuesto por los elementos del proceso de TC:

4.1 Sistemas de clasificación

Los sistemas de clasificación se desarrollaron para organizar principalmente materiales como libros o publicaciones periódicas, y se han utilizado en bibliotecas tradicionales durante algo más de un siglo.El uso de estos sistemas se ha ampliado al entorno en-línea para organizar la información digital en los casos en los que el papel potencial de los sistemas de clasificacion bibliotecarios se ha explorado como herramienta para organizar, navegar y acceder a la información. Algunos de los sistemas de clasificación universales utilizados en diferentes proyectos son (Koch and Day 1997), LCC, DDC, National Library of Medicine (NLM) y la Clasificación Decimal Universal (CDU)

4. 1.1. Características y Campo

El primer reto se presenta con el tamaño y lo etéreo de las categorías. Aproximadamente existen 100.000 diferentes categorías en la LCC y un número similar en la DDC. Por lo tanto, no parece que sea posible desde el punto de vista logístico el preparar datos provisionales para cada categoría y el construir un sistema de TC que corresponda a cada categoría. Además, los esquemas de clasificación no son estáticos ya que se revisan continuamente junto con las categorías existentes. Y tampoco parece que se usen todas las categorías que se especifican en los sistemas de clasificación.

El segundo reto está en las diferencias que presentan los distintos sistemas de clasificación. Lo que tienen en común es que la característica básica de las categorías es la materia. Sin embargo son muy diferentes en cuanto a la estructura y los sistemas de notación que adoptan.

Para hacer frente a este tipo de problemas hay que:

. Fijar los límites del nivel de categorías según el tema genérico de la aplicación CT

- Cada aplicación de CT se ocupa de un diferente nivel o conjunto de categorías. Una aplicación de historia puede estar interesada en la categoría de historia, mientras que para una aplicación como un directorio web es más interesante un conjunto completo de categorías.

. Implementar las características de una estructura de clasificación

En los sistemas de clasificación las relaciones entre las categorías se reflejan en su estructura jerárquica. Así, la DDC es una clasificación de tipo jerárquico, en la que una categoría dentro de un nivel indica una disciplina o tema más general que una categoría dentro su nivel subordinado. La naturaleza jerárquica de la LCC es parecida a la de la DDC. Un conjunto de categorías principales en la parte más alta del esquema jerárquico representa una lista de disciplinas, y cada una está dividida en subclases para disciplinas más concretas, excepto la E y la F (historia en América), y la Z (bibliografía y biblioteconomía)

4.2. Origen del conocimiento

Para decirlo de forma simple, el conocimiento final adquirido por la máquina deriva directamente del conjunto de datos de prueba introducidos, o sea que, el conocimiento conseguido por el proceso de TC está directamente influido por los datos de prueba. Por lo tanto, sistemas de TC que tienen como objetivo el mismo campo del conocimiento consiguen diferentes conocimientos finales si se utilizan diferentes datos de prueba.
La adquisición de datos de prueba es considerada generalmente como un proceso difícil, laborioso y costoso y en muchas ocasiones inviable. Los datos de prueba están formados por *experiencias* y por conociemto de base. El conocimiento de base se compone de un conjunto de datos generales que se pueden aplicar a materias más amplias, mientras que la experiencia se reserva para un área de conocimiento o una categoría específicos.

4.2.1. Datos de prueba más específicos para ensayo

La investigación en el área de TC basada en ML depende de los datos introducidos en prueba. Hay muy pocos datos de prueba normalizados que se puedan usar como punto de referencia. El desarrollo de mayor cantidad de datos de prueba en disciplinas diferentes incentivaría estudios más competitivos y conduciría a una mejora sustancial de TC.

4.2.2 Desarrollando el conocimiento base

Los estudios para desarrollar herramientas de desarrollo de conocimiento base son numerosos y a menudo se enfrentan a la escasez de los datos de prueba y a la dificultad de acceder a ellos. Los vocabularios controlados y los tesauros son colecciones de términos autorizados en áreas temáticas concretas y además representan cierto grado de relaciones entre los términos. El potencial de utilizar sus definiciones y sus relaciones para alimentar el conocimiento base parece prometedor. También hay que tener en cuenta que puede mejorar significativamente la puesta en marcha de la integración y la interrelación de herramientas de organización del conocimiento diferentes.

4.2.3 Generación automática de datos

Resultaría fundamental el encontrar alternativas a la recogida manual de datos ya que el proceso de generar datos es el más costoso en TC. El entorno de la información está preparado para esto: hay enormes cantidades de información digital disponibles; y también una amplia gama de herramientas y técnicas para procesar la información desarrolladas para el campo de la recuperación de la información. En algunos sistemas de información como directorios web o bases de datos en línea a menudo se encuentran recursos digitales pre-clasificados por profesionales de la información. Para una recogida de datos nuevos puede ser una opción utilizar documentos estrictamente precontrolados .
Cuando un sistema de TC se pone en marcha introduce documentos sin clasificar y produce documentos clasificados. Otra opción consiste en reutilizar documentos clasificados como conjunto de prueba.

4.2.4 Integración de múltiples fuentes

El desarrollo de modelos y de herramientas con el fin de incorporar múltiples fuentes como pueden ser el conocimiento base y el conocimiento explícito mediante diferentes métodos y vías puede ser un camino fundamental para sintetizar ideas que generen materias pertinentes.

4.3 Modelos/Técnicas de Clasificación

Cada vez mayor número de investigadores procedentes de diferentes campos, aunque sobre todo informáticos y documentalistas se está interesando en el desarrollo de herramientas y métodos para la clasificación automatizada de textos. Se ha propuesto y se ha probado una amplia gama de algoritmos inductivos de aprendizaje como Vector Machines, Baysian Belief Network, Decision Trees, y Artificial Neural Networks (Yiming, 1994; Joachims, 1998; Lewis y Ringuette, 1994; Mitchell, 1997). La investigación en el campo de TC se ha centrado basicamente en el desarrollo de técnicas, de métodos y de algoritmos de aprendizaje (Sebastiani, 2002).
Diferentes tipos de modelos de clasificación fundamentan diferentes representaciones del conocimiento y adoptan diferentes modelos de aprendizaje. Los algoritmos de red neuronal representan el conocimiento en forma de gráfico con nodos y vértices, mientras que los algoritmos de inducción de reglas utilizan métodos de condición-acción. Otros métodos de representación del conocimiento han utilizado funciones, programas lógicos y reglas, máquinas de estado finito, gramáticas y sistemas de resolución de problemas.

4.3.1 Técnicas de Indización semántica y modelos de clasificación

El fin último de la clasificación automatizada de texto debería ser la total comprensión del contenido de los documentos textuales. Investigadores del campo de la lingüística computacional y del de la comprensión del lenguaje natural se enfrentaron a este reto ya en el inicio del proceso automático de textos en los años 50.
La TC, que es una subdivisión de la investigación en el campo de la Recuperación de la Información (IR, Information Retrieval), ha adoptado varias de las técnicas y modelos desarrolladas para dicho campo. Ultimamente se ha recurrido a métodos y herramientas del campo de la Inteligencia Artificial, como las redes neurales, el aprendizaje simbólico y los algoritmos genéticos. Las técnicas y métodos de probabilidades para TC han sido más utilizadas en la década pasada.
Los modelos de texto que se manejan actualmente tratan con aspectos del lenguaje relativamente simples (palabras, frases, sustantivos) y los modelos para indizar términos se basan en mecanismos simples de cálculo como los de frecuencia o los de concurrencia. Este tipo de modelos no capturan aspectos de la estructura semántica o de la de relaciones que serían los más relevantes para definir las materias en la Clasificación Automatizada de textos (TC). Los ejemplos típicos de los problemas que generan los métodos no semánticos son los producidos por los sinónimos (diferentes palabras con el mismo significado) o por la polisemia (la misma palabra tiene diferentes significados).
En el campo de la Recuperación de la Información se han producido varios intentos para hacer frente a este problema. El estudio de modelos que incorporen términos que hayan pasado el control de autoridades parece una vía de investigación prometedora.

5. Conclusión

Los sistemas de clasificación bibliotecarios han cumplido diferentes papeles durante el último siglo: herramienta para localizar los fondos en las estanterías, o instrumento para navegar a través de las materias en los catálogos de acceso público en línea (OPAC, Online Public Access Catalog) y actualmente permiten organizar y acceder a recursos de información en entornos digitales . La adopción de sistemas tradicionales de clasificación bibliotecaria en el entorno digital tiene las siguientes ventajas:
(1) Los sistemas especializados de clasificación bibliotecaria se han venido utilizado mayoritariamente en el campo de la organización de la información

(2) Ya hay disponible un rico conjunto de herramientas de organización de la información basadas en sistemas de clasificación bibliotecaria como: vocabularios controlados y encabezamientos de materia
(3) Las descripciones biblográficas de las fuentes de información, los registros, incluyen la asociación entre las fuentes de información y las herramientas bibliográficas utilizadas.

Referencias

Baeza-Yates, Ricardo, and Berthier Ribeiro-Neto. 1999. *Modern Information Retrieval* New York, NY ACM Press.

Chakrabarti, S., B. E. Dom, and P. Indyk. 1998. Enhanced hypertext categorization using hyperlinks. Paper read at The ACM SIGMOD, at Seattle, WA.

Cui, Hong, P. B. Heidorn, and Hong Zhang. 2002. An approach to automatic classification of text for information retrieval. Paper read at The 2nd ACM/IEEE-CS Joint Conference on Digital Libraries, at Portland, Oregon.

Dolin, R., D. Agrawal, and A. El Abbadi. 1999. Scalable collection summarization and selection, at Berkley, CA, USA.

Frank, E., and G. W. Paynter. 2004. Predicting Library of Congress classifications from Library of Congress subject headings. *Journal of the American Society for Information Science and Technology* 55 (3):214-227.

Hidalgo, J. G., M. Maña López, and E. Puertas Sanz. 2000. Combining Text and Heuristics for Cost-Sensitive Spam Filtering. Paper read at The Fourth Computational Natural Language Learning Workshop, at Lisbon, Portugal.

Jenkins, Charlotte, Mike Jackson, Peter Burden, and Jon Wallis. 2006. *Automatic classification of Web resources using Java and Dewey Decimal Classification* [cited May 12 2006]. Available from http://www7.scu.edu.au/1846/com1846.htm.

Joachims, T. 1998. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. Paper read at The 10th European Conference on Machine Learning, April 21-24, 1998, at Chemnitz, Germany.

Koch, Traugott, and Anders Ardö. 2006. *Automatic classification: DESIRE II D3.6a, Overview of results* 2000 [cited May 12 2006]. Available from http://www.lub.lu.se/desire/DESIRE36a-overview.html.

Koch, Traugott, and Michael Day. *The role of classification schemes in Internet resource description and discovery* 1997 [cited. Available from http://www.ub2.lu.se/desire/radar/reports/D3.2.3/

Kubat, Miroslav, Ivan Bratko, and Ryszard S. Michalski. 1999. A Review of Machine Learning Methods. In *Machine Learning and Data Mining: Methods and Applications*, edited by R. S. Michalski, I. Bratko and M. Kubat. Chichester, England: John Wiley & Sons.

Larkey, L. S. 1998. Automatic essay grading using text categorization techniques. Paper read at The 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, at Melbourne, Australia.

Larkey, L. S., and W. B. Croft. 1996. Combining classifiers in text categorization. Paper read at The 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 18-22, 1996, at Zurich, Switzerland.

Larson, R. R. 1992. Experiments in automatic Library of Congress Classification. *Journal of the American Society for Information Science* 43 (2):130-148.

Lee, Yong-Bae, and Sung Hyon Myaeng. 2002. Text genre classification with genre-revealing and subject-revealing features, at Tampere, Finland.

Lewis, D. D., and M. Ringuette. 1994. A Comparison of Two Learning Algorithms for Text Categorization. Paper read at The 3rd Annual Symposium on Document Analysis and Information Retrieval, at Las Vegas, NV.

Mitchell, Tom M. 1997. *Machine Learning*. Boston, MA: McGraw-Hill.

Sebastiani, F. 2002. Machine learning in automated text categorization. *ACM Computing Surveys* 34 (1):1-47.
Shafer, Keith E. 2001. Evaluating Scorpion results. OCLC research project using DDC for automatic subject assignment. *Journal of Library Administration* 34 (3/4):237-44.
Thompson, Paul. 2001. Automatic categorization of case law. Paper read at The 8th International Conference on Artificial Intelligence and Law, May 2001, at St. Louis, Missouri.
Yiming, Yang. 1994. Expert Network: effective and efficient learning from human decisions in text categorization and retrieval, at Dublin, Ireland.