IV Jornadas Nacionales de Sociología "La Argentina de la crisis. Desigualdad social, movimientos sociales, política e instituciones". La Plata, 23 al 25 de noviembre de 2005. Mesa: "Bibliotecas, archivos y redes de información"

Publicado en las Actas de la Jornadas (2006) Versión CD.

Matemática y sociedad, una relación determinante en la elección de temas de investigación. El caso de la Ciencia de la Información.

Claudia M. González *

César O. Archuby *

Resumen

Se presenta un tema de llamativa vigencia en un contexto cambiante como el actual ya que, a pesar de que han pasado más de tres décadas desde los primeros trabajos de L. Santaló, S. Papert y otros, sobre el tema, todo parece seguir como entonces. Una sociedad que abrió sus puertas a las tecnologías de la información y las comunicaciones en muy poco tiempo, permanece inmune a los embates de los profesores de matemática. Se expondrá sobre experiencias de aplicaciones de la matemática en facultades de humanidades de universidades españolas y argentinas, en el marco de los Estudios Métricos de la Información (EMI), y sobre experiencias de incorporación de los EMI a los estudios de Bibliotecología y Ciencia de la Información. Se presentará la problemática de la utilización de modelos matemáticos en temas de recuperación de información en ambiente de bases de datos y en Internet, así como en la evaluación de la efectividad de los sistemas de recuperación de información.

Palabras clave: Ciencia de la Información - Formación profesional - Analistas simbólicos - Recuperación de Información - Modelos matemáticos

* Universidad Nacional de La Plata. Departamento de Bibliotecología.

Introducción

En la sociedad, los distintos tipos de trabajo dependen de políticas nacionales e internacionales en cuya determinación el rol de nuestras Universidades es mínimo. Sin embargo, las Universidades pueden influir de manera indirecta en las políticas de estado a partir de la formación de recursos humanos con destino a la investigación, a la docencia y al ejercicio profesional. Por ese motivo orientamos el presente trabajo al análisis de la conveniencia y factibilidad de la formación de profesionales de la Ciencia de la Información conforme a la realidad de la Sociedad de la Información y el mundo globalizado.

Por último, daremos algunos ejemplos puntuales de aplicación de los EMI a la recuperación de información, actividad integradora de los dos frentes de investigación más activos en las ciencias de la información, en los que se utilizan las capacidades y conocimientos citados en el título.

1. Tipos de trabajo, procedimientos y computadoras.

Tal vez una de las diferencias más características de los denominados "primer mundo" y "tercer mundo" es que mientras en el primero se diseña productos y servicios y se especifica procedimientos de fabricación y prestación, por cuya venta se cobra los correspondientes derechos, en el último se fabrica, se presta servicios y se paga por los citados derechos.

Además puede verse que en el primer caso la casi totalidad de la materia prima, conocimiento y capacidades, es inagotable, incluso crece con el uso, y el proceso de producción no degrada el ambiente, mientras que en el último la mayor parte de la materia prima está constituida por recursos naturales no renovables y los procesos de producción degradan el ambiente.

Una educación centrada en la adquisición de capacidades relacionadas con la ejecución de procedimientos y tareas, dará como resultado una sociedad capacitada para construir productos y para prestar servicios.

Una educación superior centrada en la adquisición de capacidades relacionadas con el descubrimiento y especificación de procedimientos, dará como resultado una sociedad capacitada para diseñar productos y servicios y para especificar los procedimientos de producción, mantenimiento y uso, así como la capacitación correspondiente.

Si bien esto último es valioso aún para el caso en que los ejecutores de los procedimientos son seres humanos, su valor es mucho mayor cuando pensamos en las computadoras como ejecutoras.

El uso de procedimientos es tan antiguo como la aritmética y los algoritmos ya que el algoritmo de la suma no es más que un procedimiento que aplicamos mecánicamente, para cuyo uso fuimos entrenados mecánicamente en la escuela básica. Tanto los alumnos como los maestros, desconocíamos la suma y el sistema de numeración hasta el punto de no acertar con la respuesta ante preguntas como “por qué me llevo uno y por qué pongo el uno donde lo pongo”, sin embargo hacíamos la suma con rapidez y sin errores. Ese es el valor de los procedimientos en relación con la ejecución de tareas.

La educación formaba buenos *ejecutores* de procedimientos y *entrenadores* para la *ejecución* de procedimientos, formación pertinente para la época en que la mayoría del trabajo lo hacían seres humanos.

La creación de procedimientos era una tarea excepcional llevada a cabo por pocas personas y dirigida a ejecutores humanos. Esas personas, generalmente destacados ejecutores en su especialidad, en algún momento decidían especificar el

procedimiento mediante el cual ejecutaban la tarea en cuestión; esta decisión los llevaba a enfrentar un problema para el que no estaban preparados ya que es muy distinto ser capaz de sumar muy bien y rápido usando el procedimiento de la suma, que especificar el procedimiento usado de forma tal que pueda ser entendido y ejecutado por otras personas. La educación formal no se ocupó del tema en nuestro país.

Sin embargo, estamos en una época en que la mayoría del trabajo se *ejecuta* con fuerte participación de las computadoras llegándose al protagonismo total en algunos casos, como en los servicios de Internet o en la industria robotizada.

Si en el trabajo ha cambiado el *ejecutor,* puede la educación continuar preparando a la sociedad para *ejecutar* tareas? Si así lo hace estará capacitando a seres humanos para competir con las computadoras, con resultado fácilmente predecible ya que la capacidad de ejecución de tareas por parte de las computadoras crece sin cesar año tras año, mientras que la del hombre permanece constante, salvo la de aquellos que se apoyan en la utilización de procedimientos usados por las computadoras, tema del que comienza a ocuparse la educación, aunque tarde, sin mayor intensidad y centrada en el instrumento.

En lo relacionado con el desarrollo de técnicas y métodos para la resolución de problemas y ejecución de tareas, propios de cada especialidad, con asistencia de computadoras o con el objetivo de automatizarlos, esto es en la creación de procedimientos, la educación formal aún no interviene.

La especificación de procedimientos incluye a la programación de computadoras, la única diferencia es el *ejecutor*. El análisis de tarea, de la pedagogía, el análisis de sistemas y el análisis de computación, tienen base común. Como puede suponerse el diseño del procedimiento debe contemplar las capacidades del ejecutor y su especificación debe utilizar un lenguaje comprensible para el mismo, de allí que la

mayor diferencia se encuentre en las diferentes capacidades y lenguajes del ejecutor en cuestión.

Los programadores de computadoras se especializan en trasladar al lenguaje de las computadoras las especificaciones de procedimientos realizadas por especialistas en los procesos relacionados con la tarea en cuestión.

Los citados especialistas, por ejemplo algunos profesionales de la bibliotecología, especifican tales procedimientos en algún metalenguaje textual, gráfico o híbrido dirigido a los programadores, aunque cada día son más los que especifican directamente en algún lenguaje de computadora, tendencia que debería mantenerse ya que su origen está en la creciente potencia de las computadoras lo que facilita el desarrollo de sistemas cuyo uso directo por parte de los usuarios cada vez es más fácil.

Los mismos procedimientos, debidamente adaptados, pueden estar dirigidos a personas que ejecutarán la tarea, en cuyo caso el lenguaje de especificación sería el natural. También se incluye el caso en que la especificación dirigida a una persona corresponde a una tarea que será ejecutada por una computadora operada por la citada persona.

2. Los analistas simbólicos. Nuestra experiencia en su formación.

Estas actividades profesionales, análisis de tarea y especificación de procedimientos para que computadoras o personas ejecuten tareas, conforman un nuevo tipo de trabajo y, consecuentemente, un nuevo tipo de profesionales que podemos incluir entre los trabajos "simbólico-analíticos", una de las tres grandes clases de trabajadores según la clasificación de Robert Reich[1] para Estados Unidos, a fines del siglo xx.

Esta clasificación comprende tres clases de trabajos: los servicios rutinarios de producción, los servicios en persona y los servicios simbólico-analíticos. En la mayoría de las organizaciones y actividades se encuentran presentes los tres tipos de trabajo, en franco descenso el primero y en ascenso los dos siguientes de los cuales el último, los servicios simbólico-analíticos, tienen la mayor necesidad de atención en la educación superior y pertinencia con nuestra cátedra "Tratamiento Automático de la Información", por lo que hemos tomado esa dirección para la elección de estrategias didácticas dirigidas la formación de los estudiantes.

Estas estrategias incluyen, en la primera etapa, la adquisición de destrezas básicas de tipo general y nivel introductorio en la operación de productos de la tecnología y revisión de trabajos especialmente seleccionados con el objetivo de observar problemas y soluciones a su alcance.

En la segunda etapa se incluye la adquisición de destrezas aplicadas, orientadas a proyectos, de tipo específico y nivel intermedio o avanzado, la revisión de trabajos con objetivos de reingeniería y/o perfeccionamiento, el autoaprendizaje asistido orientado a la autonomía, la organización cooperativa para encarar proyectos de mayor envergadura y la organización del trabajo en modalidad no presencial, todo ello en el marco del desarrollo de un proyecto de aplicación de los productos de la tecnología a la resolución de problemas propios cuya solución incluirá, necesariamente, la especificación de procedimientos de algún tipo, administración integral del proyecto incluida.

La incorporación a la cátedra en calidad de auxiliares, para los alumnos, y de adscriptos para los graduados, así como la realización de seminarios de licenciatura y tesinas, permiten alcanzar objetivos más ambiciosos.

Es importante destacar que los conocimientos y capacidades relacionados con la especificación de procedimientos son comunes a la mayoría de las profesiones hecho que facilita la integración de equipos multidisciplinarios. Esto último es particularmente útil en la relación con los profesionales de la informática, temática transversal cuyo aprovechamiento exige una clara determinación de roles en la que es imprescindible separar las etapas de análisis de tarea y especificación de procedimientos, de la programación de computadoras.

Por último, las capacidades relacionadas con el análisis de tareas permiten optimizar la adquisición de productos, adquiriendo sólo lo conveniente, ya que la separación de las partes facilita la identificación de aquellos módulos que no somos capaces de realizar, o que no estimamos conveniente hacerlo en la oportunidad, evitando la costosa adquisición de "sistemas monolíticos"[2]. Estos costos no son sólo de dinero, sino de dependencia de los centros de desarrollo de productos cuyos objetivos e intereses rara vez coinciden con los de los compradores[3].

Cuando la adquisición de paquetes completos, capacitación y mantenimiento incluidas, se generaliza a todo el país y se convierte en una política de estado, se reduce la calidad y cantidad del trabajo interno, se malgasta el dinero de la sociedad y se reduce o elimina el rol de la Universidad como consultor. Si bien esto nos hace pensar en la década pasada en nuestro país, el problema también existe en el hemisferio norte como podemos ver en las decisiones de D. Swanson[2] en la Universidad de Chicago, como resultado de la aparición las microcomputadoras, hace ya casi 30 años. En forma similar actuaron importantes organizaciones como UNESCO y la OIT, entre otras.

3. Especificación de procedimientos e informatización de procesos.

Los procedimientos han sido, son y serán, a modo de brazo de palanca lógica, una de las más poderosas herramientas que potencian las capacidades previas de las personas, como ya se comentó en párrafos anteriores con respecto a la suma, pero que podemos extender al conjunto de las actividades. Una organización, un grupo, una profesión o una persona ven crecer su capacidad de prestaciones a la sociedad a la que pertenecen ya que, realizando el mismo esfuerzo, obtienen mayores productos lo que es reconocido por la sociedad con beneficio para todos.

Hace ya muchos años, en 1920, mucho tiempo antes de la aparición de las computadoras, John Dewey[4] recordaba la utopía de Francis Bacon que, en el siglo XVII, " ... *aspiraba incluso a la idea bastante absurda de un método tan perfeccionado que pudieran suprimirse las diferencias de habilidad natural de hombre a hombre*.....". Es probable que Bacon se viera a sí mismo o a un grupo selecto de colaboradores, especificando procedimientos dirigidos a ejecutores humanos, sin embargo el absurdo que vio Dewey antes de la aparición de las computadoras, dejó de serlo unos treinta años después, cuando se escribió el primer programa de computadora.

La diferencia entre quien sabe ejecutar una tarea y quien sabe especificar el procedimiento para su realización, es determinante a la hora de decidir la informatización de la citada tarea. Como sabemos, las computadoras sólo saben ejecutar muy rápidamente una cantidad de tareas básicas, pequeñas, de pobre aplicación al mundo real. Si queremos que una tarea compleja y/o grande sea realizada por una computadora, es necesario que dividamos la tarea en pequeñas y simples partes (método top-down) y/o que hagamos crecer la capacidad de la computadora en la dirección de la tarea en cuestión (método botom-up).

Las profesiones que sean capaces de especificar los procedimientos de ejecución de los tramos "..rutinarios de producción.." verán crecer en cantidad y calidad sus prestaciones a la sociedad, liberándose simultáneamente de tareas constantemente repetitivas.

4. Los modelos como instrumentos del trabajo intelectual.

En un texto que forma parte del material didáctico de la cátedra Tratamiento Automático de la Información, encontramos expresiones referidas al uso de los modelos en la ciencia como las siguientes:

> *"El acelerado proceso de desarrollo de la ciencia y la tecnología en el siglo pasado, acompañado por el tiempo cada vez menor entre el hecho científico y la apropiación de los resultados por parte de la sociedad, ha convertido en habitual lo que antes fue excepcional: el uso de técnicas, métodos, literatura, en general diversos instrumentos del trabajo intelectual, propios del ambiente científico, por parte de las distintas profesiones. En este proceso, los modelos de todo tipo juegan un papel relevante. Entenderemos el término modelos en el sentido de "representaciones simplificadas de la realidad" con objetivos específicos, incluidos los didácticos.*
>
> *Si bien la proliferación de poderosas computadoras y la decisión de la sociedad de meter dentro de ellas todo lo que se pueda, con la consecuente necesidad de adaptación previa dirigida al nuevo omni-ejecutor, han puesto de moda a los modelos, su utilización y la de las metáforas y analogías es tan antigua como la civilización: filósofos, sabios, líderes sociales y religiosos, escritores, científicos y todos aquellos dedicados al trabajo intelectual han dejado huella de su uso en textos religiosos, mitológicos, filosóficos y científicos."* [4]

Más adelante, ya refiriéndose a las aplicaciones a la documentación:

Así ocurre con el modelo de Zipf [6], que sólo toma en cuenta palabras y fonemas y, más aún, con el de Shannon[7] ya que éste no toma en cuenta a las palabras sino a los

caracteres que las forman y presupone igual frecuencia para todos ellos, lo que no se cumple en ninguna lengua, tras su objetivo de lograr máxima eficiencia en la comunicación exclusivamente.

Sin embargo, estos modelos, además de utilizarse para lo que fueron construidos, sirvieron de apoyo para otros, como el de Luhn[8], quien asoció el significado de las palabras a sus frecuencias y ubicaciones en un texto, sentando las bases del análisis automático del contenido allá por 1957, en lo que podemos tomar como el paso inicial de las computadoras hacia el procesamiento "inteligente" de textos. Unos años antes(1953) Mandelbrot[9], desde lo abstracto de la matemática, generalizó y perfeccionó el modelo de Zipf , y más tarde Booth[10] (1967) desde lo concreto de la ciencia de la información justificó la relación de Luhn entre significado y frecuencia, base del estudio del fenómeno de transición de Pao-Goffman [11](1977) que permite delimitar las zonas.

Tal vez lo más importante que hay que entender en relación con la utilización de modelos como paso previo a la informatización y más aún en el caso de la automatización, es que no se trata de simplificar la técnica humana para que la computadora pueda hacer el trabajo ya que este intento ha sido causa de los mayores fracasos en el tema ya que se logra que las computadoras hagan lo mismo que antes hacían las personas, y con el mismo procedimiento, por lo que el avance es menor y suele ser acompañado por consecuencias indeseables, como el desempleo y la competencia hombre-máquina.

Si emprendemos realizaciones anteriormente imposibles por razones cualitativas o cuantitativas, estaremos avanzando sin sustituir trabajo humano. Además, en este contexto, el método usado para ejecutar las tareas suele tener muy poca similitud con el utilizado por los ejecutores humanos, como en el caso de la selección de palabras que representen el contenido de un texto.

5. Modelos matemáticos y algorítmicos en la educación básica y superior: experiencias EMI

Cuando logramos especificar el modelo hallado en lenguaje matemático, lógico y/o numérico, habremos salvado la brecha entre las mínimas capacidades de las computadoras y las complejidades de la realidad: todo se reduce ahora a la ejecución de una gran cantidad de operaciones aritméticas y lógicas, que las computadoras llevarán a cabo con rapidez y exactitud a muy bajo costo: es el rubro en el que ofrecen sus mejores prestaciones. Sólo queda por delante un proceso de ajuste de los parámetros del modelo que lo acerque a la realidad tanto como sea necesario, tarea que se lleva a cabo mediante sucesivas pruebas sobre casos con resultados conocidos y aceptados como válidos, como es habitual en la investigación científica y técnica.

Cuando no se logre lo anterior recurriremos a los modelos algorítmicos: deberemos programar, ya que si no logramos modelar matemáticamente alguna parte de la tarea no tendremos más remedio que especificar, en algún lenguaje de computadora, cada acción que se deba ejecutar.

Para las tareas en las que no se logre ninguna de las dos soluciones recurriremos al trabajo manual, perdiendo la posibilidad de apoyar nuestras actividades profesionales en la potencia de trabajo de las computadoras. En este caso se reducirá el radio de acción de las profesiones y el rumbo de la investigación será errático ya que cambiará su dirección cada vez que aparezcan en el camino los citados modelos.

Resulta claro entonces la importancia que asignamos a la adquisición de conocimientos y capacidades en temas correspondientes a la creación y especificación de modelos matemáticos y algorítmicos de tareas propias de nuestra actividad profesional.

En el punto 2 de esta presentación, describimos la experiencia de nuestra cátedra pero quedaría por analizar las dificultades propias de la introducción de tales técnicas en ambientes humanísticos y sociales sin embargo, por razones de espacio nos remitiremos a las presentaciones realizadas en 2003 a las Primeras Jornadas Platenses de Bibliotecología de La Plata[12] y en 2004 al VI Encuentro de Docentes de Escuelas de Bibliotecología y Ciencias de La Información del Mercosur, Ebcim 2004[13]

Una de las causas de las dificultades mencionadas en el párrafo anterior puede estar en el arrastre de los conocidos y persistentes problemas relacionados con el aprendizaje de la matemática en la escuela. Al respecto debemos mencionar en el plano internacional las experiencias de Seymour Papert [14] en el MIT, porque la metodología LOGO integra la modelización matemática con la algorítmica, y en el plano nacional las propuestas de Luis Santaló[15,16] que hace más de treinta años incluía la enseñanza de la estadística y de la teoría de la información en la escuela media, porque integra la matemática y las ciencias de la información.

En cuanto a experiencias consolidadas en nuestro país debemos citar a la correspondiente a la Universidad de Mar del Plata donde, tanto en la modalidad presencial como en distancia,[17] los Estudios Métricos de la Información (EMI) tienen carácter de asignatura obligatoria para la Licenciatura, y los estudios de doctorado correspondientes al convenio con la Universidad de Granada se apoyan mayoritariamente en tal temática.

En el ámbito iberoamericano se destaca la Red Temática sobre Estudios métricos de la Información (RTEMI), encabezada por la Universidad Carlos III de Madrid, que entre 2002 y 2004 integró a profesores de tres universidades americanas y tres españolas en el dictado de seis cursos de perfeccionamiento en la materia.

En la Universidad Nacional de La Plata se acaba de firmar un convenio de cooperación con la Universidad de Granada, iniciadora de este tipo de estudios en las escuelas de nuestro país hace casi una década.

6. Modelos matemáticos aplicados a los métodos de Recuperación de Información

Cuando se habla de Recuperación de Información (RI), se hace referencia al área de la Ciencia de la Información concerniente a la representación, almacenamiento, organización y acceso a los ítems de información. Se puede decir que esta adquiere su carácter de científica, a partir de los estudios de evaluación sobre lenguajes de indización, llevados acabo en la década del 50 y del 60. La cantidad de temas que ha abordado desde entonces son variados, pero tal como sostiene Ellis[18], esta aparente diversidad puede reducirse a dos aproximaciones conceptuales diferentes. La primera, marcada por los estudios de evaluación (Cranfield I y Cranfield II), y que se suele denominar "enfoque algorítmico", se concentra en el estudio de los sistemas, como representaciones del conocimiento almacenado. La segunda, iniciada a mediados de los 70, tiene un interés cognitivista y se focaliza en las personas y en el modelado de sus estructuras de conocimiento.

La recuperación de información entendida como la tarea que realizan los sistemas automáticos, se ubica dentro del primer paradigma e implica dos actividades relacionadas pero diferentes: la indización y la búsqueda. La primera se refiere a la manera en que los documentos y las interrogaciones de los usuarios son representadas a los fines de la recuperación; y la segunda se refiere a la forma en que el archivo es examinado y los ítems son obtenidos para responder a una búsqueda dada. Los métodos y técnicas automáticos utilizados para llevar a cabo ambos procesos son muy variados.

En este trabajo, se pretende mostrar a través de ejemplos diferentes posibles aplicaciones de modelos matemáticos a la búsqueda e indización.

En el inicio, con el uso de las tarjetas perforadas, se introduce el concepto de ***postcordinación*** en la RI. Esto es, la posibilidad de combinar arbitrariamente los términos de indización de un documento en el momento de la búsqueda, de manera de satisfacer ciertas condiciones establecidas por la ***lógica de conjuntos*** (Boole). La idea principal de este modelo es que los documentos y las interrogaciones de los usuarios pueden ser representadas por uno o más términos. Cada término debe pensarse como un conjunto cuyos elementos son los documentos que lo tienen asignado como término de indización. Cuando se expresa un término en la interrogación, para el sistema se está expresando el conjunto de los documentos que lo contienen. Si se piensan los términos como conjuntos con elementos, se puede entonces establecer operaciones entre ellos. Los términos de la interrogación son susceptibles de ser enlazados por los operadores pertenecientes al álgebra de Boole (Y (AND) Intersección - O (OR) Suma o unión - NO (NOT) Resta o negación).

En la década del 60 los esfuerzos se concentran, no tanto en los procesos de búsqueda, sino más bien en agilizar la tarea de asignación de términos a los documentos. Con el desarrollo de las técnicas de ***indización automática*** se introduce la idea de la utilización del lenguaje natural en la construcción de la representación documental. Esto implicó, por un lado, el desarrollo de técnicas automáticas de carácter lingüístico que permitieran reducir la variabilidad del lenguaje de los textos. Por otro lado, el concepto general de que el ***cómputo de la frecuencia de las palabras*** utilizadas en un documento puede ser usado para determinar su significado. La justificación de

esta idea la brindó H. Luhn[8], quien sostuvo que normalmente la persona que escribe repite ciertas palabras para argumentar sus ideas y existe una muy baja probabilidad de que una palabra dada refleje más de una noción así como también, existe una muy baja probabilidad de que un autor utilice diferentes palabras para reflejar una misma noción (se debe pensar aquí que se está hablando de la terminología científico-tecnológica). Aún cuando el autor por razones estilísticas emplea la sinonimia, corre en busca de alternativas legítimas y cae en repeticiones al reforzar su idea principal. Basándose además en un trabajo previo desarrollado por el lingüista G. Zipf [6], quien demostró que cualquier texto presenta una distribución de frecuencia de palabras similar: una poca cantidad de palabras tienen una frecuencia muy alta, mientras que una gran cantidad tienen una frecuencia muy baja (cercana o igual a 1); Luhn sostiene que la significación de cualquier texto se encuentra en la parte media de la distribución.

En la década siguiente se afianzan las investigaciones que, siempre basándose en el cómputo de la frecuencia de las palabras, desarrollan las técnicas de ***ponderación de los términos***. Dichas técnicas son ampliamente usadas y se aplican tanto a la indización automática como al desarrollo de nuevos métodos de búsqueda. Si bien en un principio sólo se enfocó el problema desde la perspectiva de los documentos individuales, luego se introduce el factor de la colección a la que dichos documentos pertenecen, como forma de poder modelar matemáticamente dos variables que afectan a la recuperación de información: la especificidad de los términos utilizados y la exhaustividad en la indización. Si se analiza la indización de un documento, la exhaustividad está dada por la cobertura que hacen los términos asignados de los diferentes temas que trata el documento. Mientras que la especificidad de uno de esos términos es el nivel de detalle con el cual este representa al concepto. La relación que presenta el concepto de

exhaustividad con la ponderación de términos, parte de interpretar que la exhaustividad aumenta si se le asignan más términos de indización al documento. Cuando el número de términos del vocabulario de indización es constante, la probabilidad de que el documento sea recuperado crece. Por ello se sostiene que el aumento de la exhaustividad mejora la performance de tipo "recall" del sistema de recuperación. Ahora bien, cuando la descripción de contenido incluye mayor cantidad de términos, cconsecuentemente crecerá la frecuencia de utilización de algunos de ellos. Esto es inevitable con vocabularios controlados (tamaño constante), pero también es aplicable a la extracción de palabras del texto, especialmente si se aplica stemming (debe recordarse que esta técnica reduce la variabilidad del lenguaje natural). Entonces, la cantidad de palabras del vocabulario no crece pareja con el crecimiento del número de documentos indexados: la extracción de más palabras del documento hará que aumente la frecuencia de la palabra más que generar palabras nuevas. Cuanto más exhaustiva sea la indización, más términos se usarán y su frecuencia de uso aumentará. Esto provoca que el término se transforme en menos efectivo para la recuperación dado que no discrimina. Hace que los documentos no se puedan distinguir entre sí. Por otra parte, la especificidad es una propiedad de un término de indización en particular. Según Spark Jones[19], es una característica semántica de los términos de indización: un término es más o menos específico si su significado es más o menos detallado o preciso. Cuando se construye un vocabulario de indización se toman diversas decisiones acerca del poder de discriminación de cada término de acuerdo con su propiedad descriptiva, por ejemplo: la decisión de incluir "infusiones" o de incluir "té", "café" o "cacao". El término más general: "infusiones", será asignado a muchos más documentos que si se tratara de alguno de los términos específicos siempre que se construya el vocabulario eligiendo exclusivamente una de las dos alternativas. Un aumento en la especificidad de

los términos utilizados aumenta la performance de "precisión" del sistema de recuperación y desde el punto de vista estadístico, cuanto más precisos sean los términos, menos frecuentes serán en la colección. Pensando en una posible manera de medir o representar algorítmicamente ambos conceptos, Spark Jones redefine a la exhaustividad de la descripción de un documento como el número de términos que contiene, y la especificidad de un término como el número de documentos que lo contienen. Esto dio origen a la ***función de ponderación IDF*** (Inverse Document Frequency) donde se asume que la importancia del término es inversamente proporcional al número total de documentos en los que dicho término fue asignado. Esta función combinada con la frecuencia total del término en el documento, dio origen a la función conocida como ***TF*IDF*** la más usada en todos los sistemas desarrollados posteriormente.

En la misma década del 70, G. Salton[20], desarrolla el sistema SMART. Con él, se produce un cambio importante en la concepción de la Recuperación de Información al introducir el ***modelo vectorial.*** Su principal diferencia está dada por la manera de representar el espacio documental. Cada documento tiene uno o más términos asignados, pero la representación del documento en el sistema está dada por un ***vector***, cuyos componentes serán los ***pesos*** (resultado de la aplicación de alguna función de ponderación) que reflejen la importancia de cada término en ese documento en particular. Lo importante es que cada documento está representado no solamente por los términos que contiene, sino también por los que no contiene (peso=0). Se dice que es una representación n-dimensional, donde *n* es la totalidad de los términos de indización del sistema. Así, se puede representar la colección de documentos mediante una ***matriz término/documento.*** Si se piensa en cada documento como un punto en el espacio *n*-

dimensional, donde su posición está determinada por las coordenadas que indican los componentes del vector; realizando una abstracción, se puede convenir que existen conglomerados de puntos en aquellas zonas del espacio donde existen documentos de temáticas más similares. Una gran parte del desarrollo de este modelo está dedicada a los ***cálculos de similaridad*** entre vectores. Esto se debe, en primer término, a que el modelo propone que la interrogación sea representada también como un vector de pesos susceptible de ubicarse en ese espacio n-dimensional. De esta manera, los puntos de los documentos que están más cerca del punto de la interrogación serán los más probablemente relevantes y deberán aparecer en los primeros lugares del orden de la salida. Se suele determinar un umbral que actuará en el sistema como valor del radio de una circunferencia que marca el límite de los documentos que se recuperarán. Esta misma medida, que aquí se presenta como una función de equiparación aplicable en el mecanismo de búsqueda, también es utilizada en el modelo vectorial para producir agrupamiento de documentos o ***"clusters"***. Este tipo de técnicas se aplica para mejorar la performance del sistema ya que los documentos muy similares entre sí son relevantes para las mismas búsquedas, por lo que, al conformar un grupo ante la lógica del sistema se mejora notablemente los tiempos de respuesta. Esto es un avance en el sentido de que ningún modelo de recuperación hasta ahora había superado la relación interrogación/documentos para incorporar en el modelo la relación inter-documentos. Esta técnica de agrupamiento es conocida como ***clasificación automática***.

A mediados de los años 70 surgió un nuevo enfoque para los SRI. Investigadores como W. Cooper, S. Robertson [21], K. Sparck Jones, retoman una idea que con anterioridad habían expuesto M. Maron y J. Kuhns[22] en 1960. Suponiendo que la principal función de los sistemas de RI es ordenar los documentos de la colección en

orden decreciente de probable relevancia ante la necesidad de información de un usuario, estos autores desarrollan lo que se conoce como ***modelo probabilístico*** en RI. Como se sabe, la probabilidad se aplica a cuestiones que implican un cierto grado de incertidumbre y consiste en obtener una estimación numérica de la posibilidad de que suceda o no suceda un determinado hecho. Entonces, dada la consulta *q* se tratará de ***estimar la probabilidad de que el usuario considere relevantes los documentos d1***, d_2, al d_3, etc. Si el valor del cálculo de la probabilidad de d_1 es mayor que el de d_2, entonces d_1 será más relevante que d_2. Se asume que existe dentro de la colección un conjunto *R* de documentos relevantes ante la consulta *q* y un complemento de *R* de documentos no-relevantes. De la misma manera que en el modelo anterior, el documento está representado por un vector de términos con valores binarios: 1 si el término está presente, 0 si está ausente. Entonces la probabilidad de que el vector d_j sea relevante para la consulta q_j será $P(R \mid d_j)$ y $P(R \mid d_j)$ de que no lo sea. El problema es que no se conoce el valor del conjunto R inicial, que en principio el modelo lo considera como un conjunto ideal. Está claro que no se puede saber previamente si el documento en cuestión es relevante, entonces, habrá que encontrar una manera de estimarlo. Para ello se plantea la similitud entre d_j y la interrogación *q* como la relación de la probabilidad de que sea relevante y que no lo sea. Dado que no se puede calcular la probabilidad de R condicionada a un d_j en particular, se utiliza el teorema de Bayes para hacer la inversión, entonces se expresa la probabilidad ya no de R condicionada a d_j, sino la probabilidad de d_j condicionada a R. Esto es, la probabilidad de un documento d_j elegido aleatoriamente del conjunto R de los documentos relevantes. A partir de este principio general se construye una función de equiparación más refinada en la que se calcula la probabilidad de cada término a partir de: el peso que tiene cada término en el vector de la interrogación, el peso del término en el vector del documento, la suma de

las siguientes relaciones: a) probabilidad de que el término esté presente dentro de los documentos relevantes / probabilidad de que no lo esté y b) probabilidad de que esté ausente dentro de los documentos no-relevantes / probabilidad de que no lo esté. Este es un proceso iterativo en el cual, en su comienzo, se debe asumir que la probabilidad es constante para todos los términos de indización presentes en los documentos relevantes, por ejemplo 0.5, y que la distribución de los términos en los documentos no-relevantes es similar a la distribución de los términos en la totalidad de la colección (este valor es conocido dado que se conoce el número total de documentos). Luego, estos valores se mejoran al progresar en las iteraciones.

En la actualidad, con el desmesurado crecimiento de la Web, la mejora de los métodos de recuperación ha planteado nuevos desafíos. La naturaleza hipertextual propia del nuevo medio, ha sido probablemente la principal fuente de soluciones más eficientes. En este sentido fue paradigmático el desarrollo del buscador Google presentado por dos estudiantes de la Universidad de Stanford a mediados de los 90. S. Brin y L. Page [23,] quienes idearon un algoritmo que empleaba un nuevo método para realizar el ***ordenamiento de las páginas***. Como se sabe, el tema del ordenamiento es fundamental en los Sistemas de Recuperación de Información de carácter global ya que está ampliamente comprobado que el usuario sólo revisa, como máximo, los primeros 20 vínculos. La propuesta se fundamenta en que la importancia de una página web, en cierta medida, está dada por la ***cantidad de páginas que registran enlaces a ella***. Así, dada una página A, se cuantifica la cantidad de vínculos entrantes (cantidad de páginas que tienen vínculos hacia ella) y la cantidad de vínculos salientes. Pero además, se considera que existen vínculos con diferente nivel de importancia, y que ello debe ser tenido en cuenta, ya que sería deseable que se mejorara la ponderación de una página

con buenos vínculos, aunque pocos, que la de otra con muchos pero desechables. Entonces, una página será bien valorada si la suma de las calificaciones de cada uno de sus vínculos es alta. Para calcular la calificación de cada página en particular, se suman las relaciones entre cada una de las calificaciones que tienen las páginas que corresponden a los vínculos entrantes y la cantidad de vínculos salientes de cada una de ellas. Esto establece que un nuevo vínculo entrante eleva la calificación, pero a su vez, cuantos más vínculos salientes posee una página, menos beneficiará a la calificación de las páginas a las que se enlaza. Es una calificación recursiva que depende de la calificación de la totalidad de las páginas restantes. El algoritmo pretende modelar también el comportamiento de los usuarios al navegar en la web. Brin y Page sostienen que alguien puede elegir al azar entre los vínculos contenidos en la página actual, o saltar al azar a cualquier página de la Red ingresando la dirección correspondiente. Se supone que sigue un enlace de la página en que está con probabilidad *d*, o salta a cualquier página de la red con probabilidad *1 - d*. Parece razonable suponer que *d* sea mayor a *0.5* ya que se tiende a usar más los vínculos que allí están, antes que hacer una nueva elección al azar.

Conclusión

A partir de lo expuesto anteriormente concluimos que:

* Es necesario encarar la formación de profesionales de la Ciencia de la Información alrededor del manejo de problemáticas simbólico analíticas.
* Tal formación necesariamente presupone preparar a las personas en el análisis de tareas y en la especificación de procedimientos, evaluación y capacitación incluidas.
* La calidad del análisis y la correspondiente especificación de tareas y procesos, estarán fuertemente condicionadas por la base de conocimientos y capacidades

relacionados con modelos matemáticos y algorítmicos. Dichos conocimientos y capacidades son comunes a la mayoría de las profesiones y su manejo facilita la integración profesional.

* La Recuperación de Información es un área temática amplia que emplea una diversidad de métodos y técnicas como los citados, que facilitan la comprensión de la base tecnológica de la sociedad de la información y la hacen particularmente apta para el tipo de formación propuesta.

Bibliografía

1. REICH, R.; El Trabajo de la Naciones, Vergara, 1993, cap. 14.
2. SWANSON, D.; Miracles, microcomputers, and librarians, Library-Journal., 107(11) 1982.
3. ARCHUBY,C.; Reflexiones sobre el uso de la informática y sus productos, Boletín de la Facultad de Humanidades y Ciencias de la Educación, Año II,Nro.5, La Plata, julio de 1994.
4. DEWEY, J.; La reconstrucción de la filosofía, Aguilar, Bs.As, 1970, p.102. Primera ed. 1920.
5. ARCHUBY, C. Modelos, analogías, metáforas y equivalencias, como instrumentos del trabajo intelectual. Material didáctico de la cátedra Tratamiento Automático de la Información, Inédito, La Plata, 2001.
6. ZIPF, G.K.; The Psicho-Biology of Language, Houghton-Mifflin, Boston, 1935.
7. SHANNON, C.E.; A Mathematical Theory of Communication. Bell System Technical Journal, 623-656, 1948.
8. LUHN, H.P.; The automatic creation of literature abstracts, IBM Journal, April, 159-165, 1958.
9. MANDELBROT, B.; http://biblioteca.uam.es/paginas/medicion-calidad.htm Theory mathématique de la loi d´Estoup-Zipf, Institute de Statistique de l´Universite, Paris, 1957.
10. BOOTH, A.; A “Law” of Occurrences for Words of Low Frequency, Information and Control, 10, 386-393, 1967.

11. PAO, M.L.; Automatic Text Analysis Based on Transition Phenomena of Word Occurrences, JASIS, Vol.29, Nro.3, 1978.

12. ARCHUBY, C.; Bibliotecarios, Tecnologías de la Información, Matemática, y la Ley del 90-10. Primeras Jornadas Platenses de Bibliotecología. La Plata, 8-10/09/2003.

13. ARCHUBY, César; La enseñanza de temáticas básicas para los estudios métricos de la información en las escuelas de bibliotecología. La experiencia de la UNMdP, en VII ENCUENTRO DE DIRECTORES, VI ENCUENTRO DE DOCENTES DE ESCUELAS DE BIBLIOTECOLOGÍA Y CIENCIAS DE LA INFORMACIÓN DEL MERCOSUR, Ebcim, 2004.

14. PAPERT, S. ; El desafío a la mente, Ed. Galápago, Bs.As.,1981.

15. SANTALÓ, L. ; La enseñanza de la matemática en la escuela media, Docencia, Bs.As ,1981.

16. ----------; Matemática y Sociedad, Docencia, Bs. As.,1980.

17. ARCHUBY, César ; BAZÁN, Claudia. Elementos de Bibliometría. -- 3a. ed. -- [Documento electrónico hipertextual]. -- Mar del Plata : Universidad Nacional de Mar del Plata. Facultad de Humanidades. Departamento de Documentación, 2003. -- 1 disco compacto : col. ; 5 1/4 plg. -- Contenidos de enseñanza del Seminario C de la Licenciatura en Bibliotecología y Documentación a Distancia.

18. ELLIS, D.; Progress and problems in information retrieval. Library Association, London, 1996.

19. SPARCK JONES, K; A statistical interpretation of term specificity and its applications in retrieval, Journal of documentation, 28, 11-21, 1972.

20. SALTON, G.; Introduction to modern information retrieval, McGraw-Hill, New Cork, 1983.

21. ROBERTSON, S.E. and SPARCK JONES, K.; Relevance weighting of search terms, JASIS, 129-145, 1976.

22. MARON, M.E. and KUHNS, J.L.; On relevance, probabilistic indexing and information retrieval, Journal of the ACM, 7, 216-244, 1960.

23. PAGE, L., BRIN, S., MOTWANI, R. WINOGRAD, T. The PageRank Citation Ranking: Bringing Order to the Web. Stanford Digital Library Technologies Project, 1998.