Universidad Nacional de Luján

Trabajo Final de Licenciatura en Sistemas de Información

Selección de Recursos Distribuidos en Ambientes Dinámicos

Autor: Santiago Banchero

Director: Mg. Gabriel H. Tolosa

Co-Director: Mg. Fernando R. A. Bordignon

2010

A mi esposa,

mis padres y hermanos.

Agradecimientos

Agradezco a mis padres por enseñarme que con esfuerzo y dedicación todo es posible y por alentarme siempre y en todo momento en mi carrera de estudiante. A mi esposa, por su infinita paciencia y contención durante el tiempo que desarrolle este trabajo. También a mis hermanos, por estar siempre junto a mi.

Un agradecimiento especial a Mariano Felice, inseparable amigo en todos esos años de estudiante.

A mi director Gabriel Tolosa, por estar siempre y en todo momento dispuesto a brindar una mano, por ser un ejemplo de profesionalismo y por su incondicional apoyo en todos estos años de estudiante. De igual manera, quiero agradecer a mi codirector Fernando R. A. Bordignon por sus infinitos consejos y por su incansable aliento en mi desarrollo profesional.

Resumen

Históricamente, el hombre ha trabajado incansablemente para organizar su información de manera que esta esté siempre y en todo momento disponible. Siempre impulsado por la necesidad de representar lo relacionado con su entorno a fin de plasmar su evolución. El inmenso volumen de información existente en la actualizad acarrea consigo algunos problemas, como es el de acceder a información que se necesita. La Recuperación de Información (*Information Retrieval*) o RI es el área dentro de las ciencias de la computación que se encarga de estudiar y proveer soluciones al problema mencionado.

En este trabajo se trabajo se aborda el problema de la ir pero en otro contexto, donde los recursos de información no se encuentran en un único lugar sino que están distribuidos. La Recuperación de Información Distribuida (RID) es un área de investigación que se ha desarrollado fuertemente en los últimos años a partir de la expansión de los repositorios de información textual en diferentes organizaciones (empresas, universidades, etc.) y la aparición y rápido desarrollo de nuevos servicios de información en Internet, como – por ejemplo – *blogs*, periódicos digitales y *wikis*.

Existen tres problemas que caracterizan la RID, estos son: cómo representar los recursos, cómo seleccionar los recursos distribuidos y cómo fusionar los resultados. En este trabajo se profundizará en una nueva metodología para la construcción de descripciones de recursos y en el desarrollo de una nueva alternativa de selección de recursos que contempla la frescura de los datos. En este sentido se definió un nuevo modelo de selección de recursos CORI-T cuyos resultados fueron muy satisfactorios al ser comparados con el modelo de referencia.

Organización del Trabajo

En el Capítulo 1 se introducen los conceptos básicos de Recuperación de Información, describiendo problemáticas y modelos tradicionales. También se explican los ambientes de información dinámicos y se introduce el concepto de Web Oculta.

El Capítulo 2, profundiza en los conceptos básicos de la Recuperación de Información Distribuida, explicando en detalle las problemáticas principales y modelos para abordarlas. Se incluye también una descripción profunda de los conceptos de ambientes cooperativos y no cooperativos.

En el Capítulo 3 se aborda el concepto de Bases de Datos Textuales. Se explica en detalle cómo se distribuyen los términos, cómo se incrementa el vocabulario y se caracteriza el *dataset* utilizado para los experimentos de este trabajo.

El Capítulo 4 describen los modelos de selección de recursos ajustados por tiempo. Aquí se muestra en modelo de selección propuesto y se analizan los resultados de su rendimiento comparando los contra el modelo de referencia centralizado.

Por último, en el Capítulo 5 se concluye el trabajo con algunas consideraciones finales y trabajos futuros.

Tabla de Contenidos

Capítulo 1

Introducción

Históricamente, el hombre ha organizado su información de diferentes formas para posteriormente recuperarla y usarla, principalmente, impulsado por la necesidad de representar lo relacionado con su entorno a fin de plasmar su evolución. Este proceso tiene su origen con la escritura, que ha sido históricamente la herramienta para la transmisión de conocimiento. Un ejemplo típico que ilustra cómo se ha organizado la información a lo largo de los años son los índices de libros. Esta es una de las estructuras de datos más antigua que se conoce y ha posibilitado encontrar información en forma rápida y precisa.

En el mismo sentido, la evolución de los diferentes medios para dar soporte a la escritura alcanzan hoy en día un sin número de soluciones donde la representación en forma digital de la información permite almacenarla y distribuirla en forma masiva. El crecimiento del volumen de información es continuo, la representación comprende desde la utilización de archivos de texto plano almacenados en repositorios soportados en DVD hasta bitácoras, librerías digitales y repositorios accesibles a través de la Web.

El inmenso volumen de información existente en la actualizad [Lyman, 2000] acarrea consigo algunos problemas, como es el de acceder a información que se necesita. La Recuperación de Información (*Information Retrieval*) o RI es el área dentro de las ciencias de la computación que se encarga de estudiar y proveer soluciones al problema mencionado.

La RI no es una disciplina nueva dentro de las ciencias de la computación sino que, por el contrario, tiene su origen en la década del 50. En el campo de la ciencia de la

información, se puede definir la RI como un área de estudio que intenta resolver el problema de encontrar y *rankear* documentos relevantes, generalmente de texto y no estructurados, que satisfagan la necesidad de información de un usuario – expresada en un determinado lenguaje de consulta [Croft, 1987].

Concretamente, trata de la representación, almacenamiento, organización, acceso a grandes volúmenes de información y permite a los usuarios alcanzar fácilmente la información que es de su interés [Baeza y Ribeiro, 1999]. En la actualidad, su rol adquiere un papel más relevante debido al grado de utilidad de la información y ha evolucionado notoriamente con la consolidación de la World Wide Web (Web) como herramienta de difusión.

Tradicionalmente, la información se presenta en forma de texto, esto implica que RI también puede ser aceptada como sinónimo de recuperación de texto o de documentos [Ingwersen, 2002], independientemente si se habla de texto plano, información bibliográfica u otro formato de texto. Además en los últimos años el contexto de la RI se ha extendido a ambientes multimedia que involucran el almacenamiento y recuperación de imágenes, sonidos, componentes de software, documentos de oficina, etc. [Lew, 2006].

Antes de plantear las problemáticas de la RI se debe introducir el concepto de Sistema de Recuperación de Información (SRI) , según [Peña, 2003] es “el conjunto formado por el hardware, los programas y los datos que facilitan la localización de los documentos adecuados para la necesidad de información”.

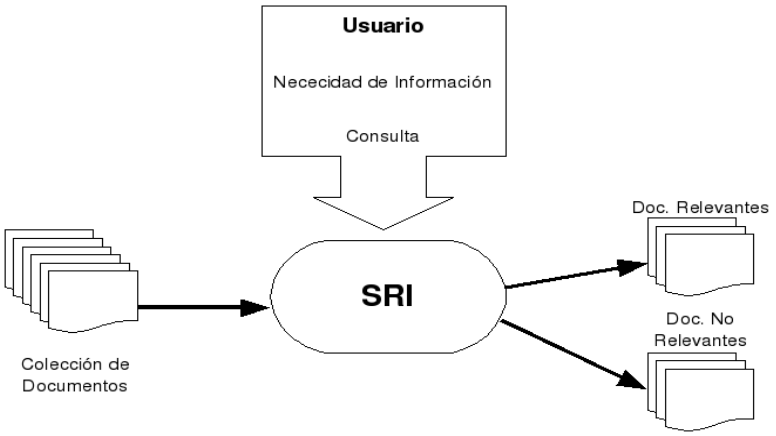


Figura 1.1: Problemática de la RI

En la figura 1.1 se muestra en forma gráfica la problemática de la RI y cómo se relaciona cada una de las entidades participantes del proceso de recuperación. Desde el punto de vista del usuario comienza a partir de una necesidad de información, que debe ser mapeada a un formato de consulta válido para el SRI. Por otro lado, el sistema tomará como entradas el conjunto (o colección) de documentos y la consulta para poder establecer qué documentos son relevantes y cuales no.

La problemática de la RI puede ser abordada desde dos puntos de vista, el computacional, donde el mismo recae en la construcción de estructuras de datos capaces de soportar y agilizar la tarea de búsqueda y recuperación. Por otro lado, también puede ser tratado desde el punto de vista de un ser humano donde se debe analizar la necesidad concreta de los usuarios. A continuación se listan las problemáticas a abordar por los sistemas de RI [Baeza y Ribeiro, 1999].

- Procesar las diferentes colecciones de documentos que existen y que tratan sobre diferentes temas para poder lograr una representación confiable definiendo una estructura de datos.

- Representar la necesidad de información del usuario que es planteada al sistema

de recuperación de información (SRI) en forma de consultas textuales. La tarea de compatibilizar una necesidad de información del usuario, expresada como consulta, con la representación de una colección de documentos es el principal problema con que se enfrenta la RI.

- Realizar una comparación entre ambas representaciones (necesidad de información y documento). El SRI debe ser capas de retornar las la mayor cantidad de referencias a los documentos más relevantes, que son aquellos que satisfacen – total o parcialmente – la necesidad expresada anteriormente.

1.1. Sistemas de Recuperación de Información

Cuando se combinan el soporte físico (hardware), los programas y los datos que permiten localizar los documentos que son relevantes para la necesidad de información de un usuario se tiene un SRI completo [Peña, 2003]. En la figura 2 se muestra la arquitectura completa de un SRI y se puede ver cómo interaccionan cada uno de sus componentes.

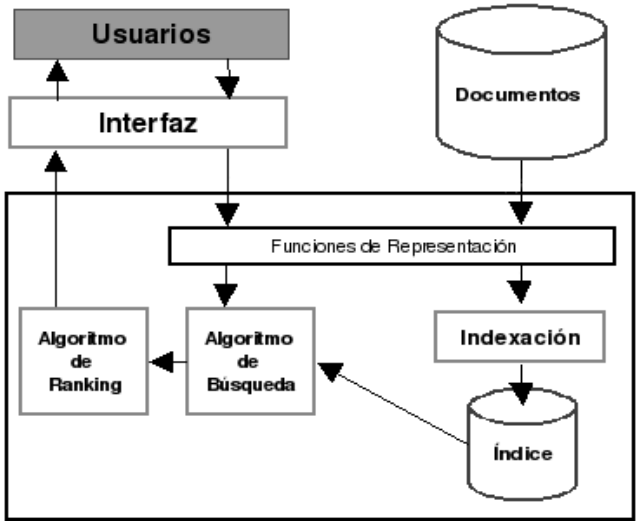


Figura 1.2: Arquitectura de un SRI

Los documentos utilizados por el sistema de recuperación son tratados como unidades individuales, sin formato y deben ser procesados para poder extraer de ellos algún tipo de información útil para el proceso de recuperación. Este proceso es de carácter meramente estadístico y está basado fundamentalmente en el análisis de frecuencias, el mismo no se corresponde con el área de Procesamiento de Lenguaje Natural.

Cada uno de los documentos está formado por un conjunto de elementos, así se lo puede dividir – de lo general a lo particular en – párrafos, oraciones y términos. Todo este volumen de información se encuentra almacenados en repositorios en forma digital.

El conjunto de documentos con los que interactúa el sistema se lo denomina *corpus*, colección o base de datos textual o documental (en adelante se usará indistintamente cualquiera de ellas). El SRI realiza una representación lógica de estos documentos para poder caracterizarlos tomando como unidades los términos, oraciones u otras unidades.

La construcción de los índices es el proceso que se deriva de la representación lógica de los documentos, son fundamentalmente estructuras de datos que sirven para almacenar y para dar soporte a las búsquedas. Una vez construidos los índices es posible deshacerse de los documentos o no dependiendo del tipo de SRI que se tenga. Si solo se conservan las referencias pueden eliminarse los documentos y lo que se tiene es un sistema referencial que solo provee la ubicación del documento (la referencia). Un ejemplo parcial de este tipo de sistemas es el de Google[1] que guarda – entre otros datos – las referencias a los documentos, aunque también guarda documentos en su *cache*. Otro ejemplo de sistema referencial es el de las bibliotecas donde no se tiene el contenido del libro sino que existe una referencia al mismo en una ficha o algún otro medio. Por otro lado, si se almacenan los documentos completos lo que se tiene un SRI documental [Abadal, 2005].

Los algoritmos de búsqueda son los encargados de evaluar si los documentos "se

[1] http://www.google.com

corresponden" de alguna manera con la consulta ingresada por el usuario y deben ser retornados. Esto se puede realizar mediante medidas de similitud a partir de las cuales se establece la relevancia de cada documento [Korfhage, 1997]. El objetivo es retornar aquellos documentos que puedan responder la consulta. Por otro lado, el algoritmo de *ranking* será el encargado de calcular la relevancia de cada uno de los documentos y construir una lista ordenada con el *score* de cada uno, donde el primer ítem es el mejor ponderado y se asume más relevante.

En este proceso de *rankeo* se debe considerar una de las cuestiones más controversiales de los SRI que es – como se mencionó anteriormente – el concepto de relevancia, dado que el objetivo del mismo es retornar la mayor cantidad de documentos que satisfagan la necesidad de información del usuario. El problema radica en que la noción de relevancia es un juicio personal que depende – en parte – de la subjetividad del usuario. Dada esta situación se asocia el concepto de relevancia con el de similitud donde se establece una medida de distancia para comparar un documento con una consulta. En [Greisdorf, 2000] se plantea que si bien es complejo el número de problemas que presenta la relevancia no se ha encontrado otro sustituto práctico que sirva como criterio para medir la efectividad de los SRI.

En la actualidad la RI se ha convertido en una disciplina de gran importancia, esto se debe al valor que tiene la información y el constante crecimiento de grandes repositorios digitales – públicos y privados – que existen en diferentes organizaciones. La RI provee las herramientas necesarias para poder acceder a la información que se necesita en el menor tiempo posible.

1.2. Evaluación de SRI

Las primeras evaluaciones que se realizaron a un SRI datan de los años cincuenta y

fueron realizadas en el marco de los proyectos *Cranfield* (desarrollados en el *Cranfield Institute of Technolgy*) donde se sentaron las bases metodológicas para la evaluación de los sistemas de recuperación de información. El más destacado fue el dirigido por el profesor Cleverdon en el año 1957, que fue el mentor de la evaluación introduciendo las medidas de exhaustividad y precisión que aun hoy se utilizan [Chowdhury, 1999].

Otro proyecto que marco el rumbo de la disciplina fue el desarrollado por Gerard Salton, cuyo resultado distintivo fue el sistema SMART [Salton, 1983]. Este se distinguió por implementar métodos de indexación automática, clasificación de documentos, identificación de los documentos a recuperar en base a la similitud con la necesidad de información del usuario y por la automatización de procesos para generar mejores ecuaciones de búsquedas. A partir de éste, la evaluación adquirió una importancia mayor.

Hoy en día, el foro de intercambio científico más importante en el área de evaluación de la recuperación de información son las conferencias TREC[2] (*Text REtrieval Conference*). Estas conferencias son auspiciadas por NIST[3] (*National Institute of Standards and Tecnology*) y por el *Information Tecnology Office of Defense Advanced Research Projects Agency* (DARPA[4]) y tienen como objetivos la definición de estándares para la evaluación de SRI.

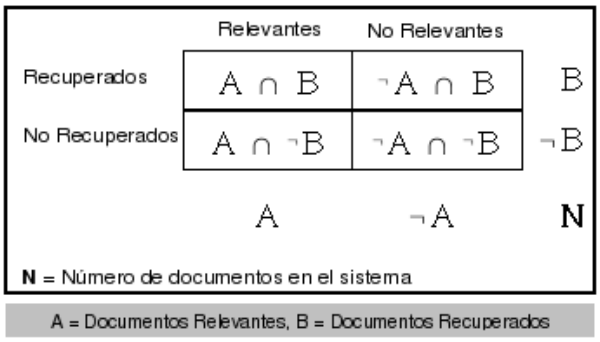
Para analizar la calidad del SRI es posible estudiar si los resultados retornados para una consulta están o no relacionados con esta. Es decir, la evaluación consiste en determinar – en un tiempo y espacio razonable – qué tan preciso y exhaustivo es un SRI dado la cantidad de documentos relevantes que este retorna.

Las medidas de evaluación más utilizadas son las basadas en relevancia, donde al realizar una recuperación de información, los resultados posibles son:

[2] http://trec.nist.gov/
[3] http://www.nist.gov/
[4] http://www.darpa.mil/

- Documentos relevantes con el tema que trataba la necesidad de información.
- Documentos que no son relevantes con la necesidad de información.

También puede ocurrir que se dejen de recuperar documentos que sí eran relevantes a la necesidad de información, y otro conjunto de documentos que no lo eran. Esta distribución de resultados es planteada por [van Rijsbergen, 1999] y resumida en siguiente "tabla de contingencia":



Conociendo ya los posibles conjuntos solución se definen las medidas Precisión (*precision*) y Exhaustividad (*recall*) que son los criterios de evaluación estándar para evaluar un SRI [Cleverdon, 1968]. Precisión mide cuantos documentos relevantes han sido recuperados, simplemente se divide el total de documentos relevantes recuperados (RR) sobre el total de documentos recuperados (R).

$$P = \frac{RR}{R}$$

Exhaustividad es la proporción de documentos relevantes que fueron recuperados sobre el total de documentos relevantes que existen en la colección (REL).

$$E = \frac{RR}{REL}$$

Esta medida es la más difícil de calcular, ya que no es simple conocer el denominador (cantidad de documentos relevantes en la colección) [Martinez, 2004]. La principal dificultad se presenta cuando se trabaja en recuperación de información en la Web, ya que es prácticamente imposible establecer el tamaño y por ende qué cantidad de documentos relevantes puede haber para satisfacer una necesidad de información.

Entre ambas medidas existe una fuerte relación, en sentido inverso, donde cuanto mayor es el valor de la precisión menor va a ser el valor de la exhaustividad [Buckland, 1994]. Por lo tanto, si el sistema es más exhaustivo entonces comenzará a perder precisión. Esto se debe a que la salida de un SRI es un conjunto aproximado de documentos donde pueden existir documentos relevantes como no relevantes. De otra forma, si el sistema fuera más preciso se recuperaran todos documentos relevantes pero corremos el riesgo de que deseche documentos con información relevante.

A continuación se plantea un ejemplo para ilustrar cómo se puede evaluar distintos sistemas de recuperación de información utilizando las medidas de exhaustividad (E) y precisión (P). Dada una colección de 50 documentos y contamos con un *query* de prueba que para el cual existen 10 documentos que son relevantes.

En la tabla 1.1 se muestran cuales son los valores resultantes de E y P, en qué orden aparecen cada uno de los documentos relevantes y como los ordenó en la respuesta cada uno de los tres sistemas a evaluar (SRI_i $\forall$ i = 1,2,3). El cálculo se realizó sobre los primeros 20 documentos retornados.

	SRI_1			SRI_2			SRI_3		
j	w	E	P	w	E	P	w	E	P
1	1	0,10	1,00	1	0,10	1,00	1	0,10	1,00
2			0,50			0,50	2	0,20	1,00
3	2	0,20	0,67			0,33			0,67
4			0,50	2	0,20	0,50	3	0,30	0,75
5			0,40	3	0,30	0,60			0,60
6	3	0,30	0,50	4	0,40	0,67	4	0,40	0,67
7			0,43			0,57	5	0,50	0,71
8			0,38	5	0,50	0,63			0,63
9	4	0,40	0,44	6	0,60	0,67			0,56
10			0,40			0,60			0,50
11	5	0,50	0,45	7	0,70	0,64	6	0,60	0,55
12			0,42			0,58	7	0,70	0,58
13			0,38			0,54			0,54
14	6	0,60	0,43	8	0,80	0,57	8	0,80	0,57
15			0,40	9	0,90	0,60	9	0,90	0,60
16			0,38	10	1,00	0,63			0,56
17	7	0,70	0,41			0,59			0,53
18	8	0,80	0,44			0,56			0,50
19	9	0,90	0,47			0,53	10	1,00	0,53
20	10	1,00	0,50			0,50			0,50

Tabla 1.1: Donde: E = Exhaustividad, P = Precisión, w = Nro. de documento relevante

Los resultados obtenidos de la evaluación se presentan en el gráfico 1 a través de las curvas de Exhaustividad y Precisión, donde se ve que el sistema 3 es el más preciso de los tres mientras que el número 2 alcanza la mayor exhaustividad.

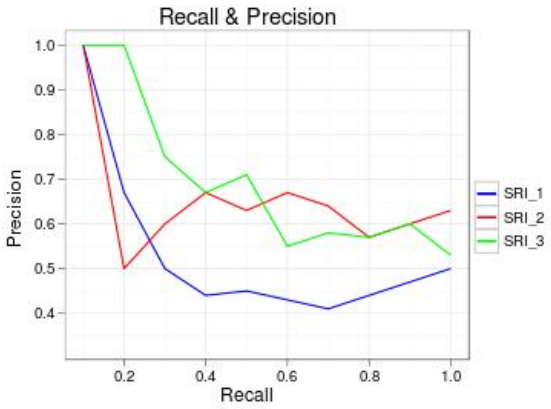


Gráfico 1.1: Curvas de Exhaustividad y Precisión, comparativa de tres SRI

1.3. Modelos de RI

Los modelos empleados en RI permiten realizar una comparación ponderable entre las consultas que se le realizan al sistema y los documentos de la colección para poder identificar cuáles son relevantes. Existen varios modelos de RI, la utilización de uno u otro determina el tipo de indexación que se realizará del corpus [Peña, 2003]. A continuación se describen, brevemente, los modelos clásicos de RI. Estos son:

1.3.1 Modelo Booleano

Es el modelo más simple y está basado en álgebra de *Boole*. Los documentos son indexados por una serie de términos que forman una bolsa de palabras (*bag of word*), que no son *palabras vacías*[5]. Aquí cada uno de los términos describe de igual manera al documento, es decir, no se distingue entre términos más o menos representativos.

El usuario formula la consulta utilizando términos y operadores booleanos (AND, OR,

[5] Son las palabras por las que el usuario no realizará una búsqueda, por lo tanto no se indexan. Por ejemplo los artículos, adverbios, pronombres, preposiciones, conjunciones, etc.

NOT). En base a los términos por los que fue indexado un documento el sistema cotejará la consulta con los índices y retornará las referencias a los documentos que cumplan exactamente con la consulta.

El problema que tiene este modelo es que no puede ordenar los documentos retornados por relevancia. Ese problema se soluciona con el modelo booleano extendido [Salton, 1983b].

1.3.2 Modelo de Espacio Vectorial

Es uno de los primeros modelos que surgió, fue ideado por Gerard Salton [Salton, 1968]. Los documentos y las consultas son representados como vectores de un espacio de n-dimensiones donde cada dimensión corresponde a un término indexado.

Luego a través de una medida de distancia se puede conocer cuáles son los documentos similares entre si. De igual manera se trabaja con la consulta, que es mapeada como un documento más. Este modelo permite que los documentos seleccionados sean ordenados por relevancia, utilizando las diferentes medidas de semejanza entre el documento y la consultas.

1.3.3 Modelo Probabilístico

Este modelo está fundamentado en el cálculo de la probabilidad de que un documento sea relevante para una consulta dada. Para esto se utilizan cuatro conjuntos donde se representan:

- documentos relevantes
- documentos recuperados
- documentos relevantes recuperados
- documentos no relevantes y no recuperados

La idea fundamental de este modelo es que dada una consulta del usuario hay un conjunto de documentos que contiene solo documentos relevantes. Este es denominado el conjunto de respuesta "ideal". Como resulta difícil conseguir este conjunto el modelo propone una interacción con en usuario donde este genera una descripción probabilística del conjunto de documentos relevantes. A través de sucesivas interacciones con el SRI se va mejorando el rendimiento de la recuperación.

Si bien permite realizar *ranking* por relevancia es muy complejo lograrlo y además los resultados son inferiores a los del modelo vectorial.

1.4. Introducción al Recuperación de Información Distribuida y su diferencia con la RI tradicional

Los motores de búsqueda tradicionales para la Web y las redes corporativas utilizan modelos de almacenamiento donde los documentos son copiados a una base de datos centralizada, luego son indexados para poder realizar búsquedas. Sin embargo, existen varias limitaciones para este modelo centralizado ya que los documentos a copiar no siempre se encuentran disponibles o porque se trata de información propietaria, tiene un alto costo monetario, o simplemente no puede ser alcanzado por el modelo.

La Recuperación de Información Distribuida (DIR) [Callan, 2000] es un área de investigación que se ha desarrollado sostenidamente en los últimos años a partir de la expansión de los repositorios de información textual en diferentes organizaciones (empresas, universidades, etc.) y la aparición y rápido desarrollo de nuevos servicios de información en Internet, como – por ejemplo – *blogs*, periódicos digitales y *wikis*. Este tema será tratado en profundidad en el capítulo 2.

Por otro lado, el área de RI tradicional ha sido pionera en la tarea de buscar y *rankear*

documentos relevantes a partir de una necesidad de información del usuario [Callan, 2000; Si, 2005], operando – como ya se menciono – sobre grandes volúmenes de información, generalmente en documentos de texto y tradicionalmente bajo *esquemas centralizados*.

Este tipo de soluciones, basadas en un índice centralizado, adolecen de ciertos problemas entre los que se pueden mencionar la ausencia de escalabilidad para trabajar con grandes colecciones y la complejidad que involucra el proceso de actualización, que dada la continua evolución de éstas en el tiempo, dificulta sobremanera poder contar con una representación actualizada.

Los índices centralizados, utilizados por los SRI tradicionales, donde se representa y consolida toda la colección de documentos se basan – en general – en la representación de los documentos a partir de la frecuencia de sus términos. A esta representación se la conoce como *bag-of-words* [Hawking y Thistlewaite, 1999]. Cuando se trabaja en ambientes de búsquedas distribuidas no es posible disponer de toda la información publicada en los distintos repositorios en un único índice centralizado. Esto se debe a que bajo un esquema de búsqueda distribuido entran en juego factores como el volumen de datos, la frecuencia de actualización, la disponibilidad, etc. que dificultan la construcción de índices centralizados.

1.5. Evolución de las redes de datos

Las redes de datos forman parte de uno de los principales recursos con que se compone un SRI. A través de ellas se consigue intercomunicar e intercambiar información entre un sin fin de computadoras conectadas, contribuyendo así al aprovechamiento de los recursos disponibles en cada uno de los extremos de la red.

Internet nace como un proyecto experimental del Departamento de Defensa americano que pretendía nuclear a un conjunto de redes del ámbito científico y universitario. ARPANET[6] (*Advanced Research Projects Agency* NET) fue el nombre con el que se conoció a la red que se convertiría en Internet. El principal objetivo que tenía era el de desarrollar una red que no se debilitara ante la pérdida de alguno de sus nodos. Tampoco se debería ver debilitado su rendimiento ante el incremento de nuevos equipos al sistema.

En el marco de estos proyectos que surgieron de ARPANET, con la creación de diferentes redes según las temáticas tratadas, se desarrolla la familia de protocolos TCP/IP que se convirtió en el estándar para la interconexión entre redes.

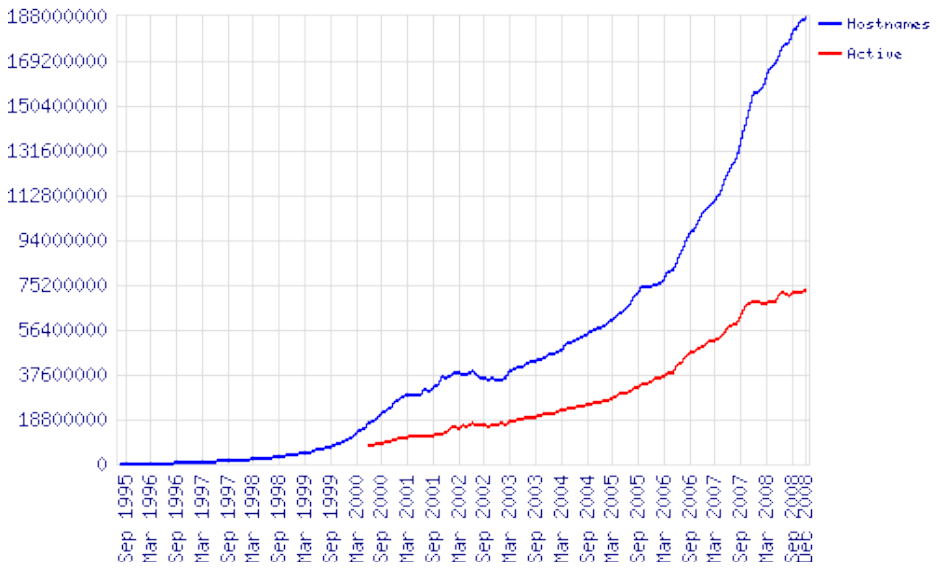


Gráfico 1.2: Total de sitios (en todos los dominios) desde agosto de 1995 a diciembre de 2008

Una de las principales característica de la red de redes (o Internet), es que no se ajusta a ningún tipo de computadora, o tipo de red, ni tecnología de conexión y tampoco a los medios físicos utilizados. Además es descentralizada, cada una de las redes que la

[6] http://en.wikipedia.org/wiki/ARPANET

compone mantiene su independencia y se une de forma cooperativa a las demás redes. La heterogeneidad está controlada por la familia de protocolos TCP/IP.

Desde su aparición en 1989 la web se ha convertido en uno de los servicios más utilizados y de mayor crecimiento. Según NetCraft[7], en diciembre de 2008 el total de sitios Web aproximadamente es de 186.727.854 millones de sitios. En el gráfico 1.2 se muestra como ha evolucionado la Web en cantidad de sitios desde agosto de 1995 a diciembre de 2008[8].

1.6. Una aproximación a la Web Oculta

El marcado crecimiento de la Web en el último tiempo demanda constantemente soluciones de búsqueda más efectivas, para que sus usuarios puedan acceder a información que sea relevante. Los motores de búsqueda convencionales, como pueden ser Google[9] o MSN[10], basados en técnicas de recolección de páginas web (*crawling*) [Cho, Garcia-Molina, Page 1998] proveen una alternativa efectiva para el acceso a algunos tipos de información. La estrategia de estos buscadores es la de construir bases de datos centralizadas, indexando el contenido y poniéndola a disponibilidad para sus usuarios. Este tipo de información, que puede ser obtenida por los motores de búsqueda convencionales forma parte de la *Web visible.*

Por otro lado, existe una gran parte de la Web que no puede ser accedida por los motores de búsqueda convencionales, y por lo tanto, hay muchas fuentes de información que no pueden ser copiadas y agregadas a una base de datos centralizada. Este tipo de información forma parte de lo que se conoce como Web oculta [Kautz, 1997].

---

7 http://news.netcraft.com
8 http://news.netcraft.com/archives/2008/12/24/december_2008_web_server_survey.html
9 http://www.google.com
10 http://www.msn.com

La *Web oculta* contiene gran diversidad de contenidos de diferentes temáticas, idiomas y formatos. Gran parte de los documentos existentes no están estructurados, es decir, están en texto plano mientras que posee también contenido estructurado y semi-estructurado, como pueden ser formatos derivados XML u otros lenguajes de marcado.

1.7. Ambientes de Información Dinámicos

En los últimos años se ha desarrollado un nuevo espacio de publicación de información que fue acompañado por la evolución de la Web 2.0 [O' Reilly, 2005]. Este nuevo ambiente ha transformado radicalmente la forma en que se publica y se accede a la información, brindando herramientas para la sindicación a través de diferentes formatos (RSS, ATOM, RDF, etc.) como así también proveyendo de servicios de búsqueda para acceder a estos nuevos repositorios caracterizados principalmente por su alta tasa de actualización.

Los sistemas de sindicación de contenido tienen hoy en día un papel preponderante en la evolución de la Web. Las nuevas formas de publicación utilizando servicios de feeds a través de formatos de intercambio basados en XML, como son: RSS [Beged-Dov, 2000] y ATOM [The Internet Society, 2005] son las claves de este fenómeno que tiene como principales usuarios a los blogs, periódicos y páginas personales, entre otros.

Todos los formatos de sindicación utilizados en la actualidad, sin excepción, están construidos a partir de XML (*Extensible Markup Language*). Se puede definir XML como un conjunto de reglas que especifica una sintaxis usando etiquetas que se utilizan para fraccionar un documento en partes y que permite identificar cada una de las partes de ese documento con ellas. Así puede ser usado para detallar otros dominios específicos - por ejemplo los formatos de sindicación.

XML fue diseñado para facilitar el intercambio y almacenamiento de datos. Permite incorporar metadatos aportando un valor agregado a los datos almacenados en su estructura.

La lógica de estas metodologías de gestión de contenido fue aportada por las técnicas *push* and *pull* [Datta, 2008]. Por un lado, *push* describe a los sistemas de distribución de contenidos a través de *Internet*, donde la información viaja del servidor al cliente. La tecnología *pull* es la que permite a los usuarios tomar los contenidos utilizando por ejemplo, suscripciones. La sindicación de contenidos está relacionada con ambas tecnologías ya que el usuario se suscribe a un canal de interés y, por otro lado, el proveedor de canales distribuye sus contenidos a los usuarios que se encuentran suscriptos.

Los antecedentes que existen en relación a los diferentes formatos de sindicación datan de 1997 con *Channel Definition Format* [Ellerman, 1997], que fue creado por Microsoft para Internet Explorer 4.0. También del mismo año el esta el formato creado por *Dave Winer, ScriptingNews* [Winer, 2006].

El formato RDF [Lassila, 1999], un estándar de metadatos del W3C[11], fue la base del formato RSS (*RDF Site Summary*), desarrollado por Dan Lobby de Netscape en el año 1999. De este formato derivaron versiones subsecuentes, desarrolladas por UserLand, que convergieron en las utilizadas actualmente que son RSS 1.0 y RSS 2.0 (*Really Simple Syndication*) [Cann, 2006]. El objetivo del formato RSS es el intercambio de contenido sin tener que visitar el sitio Web que ha producido la información, solo es necesario contar con una herramienta de agregación de noticias.

ATOM surgió en el 2003 y contó con el impulso de Google con lo que se volvió muy popular [Sayre, 2005]. Este formato también esta basado en XML y consta de un conjunto de entradas, denominadas *feeds*, que contienen distintos metadatos. La idea

[11] http://www.w3.org/

inicial fue crear un único estándar de sindicación para terminar con la confusión de versiones existentes, pero fracasó. Aun se utilizan ambos formatos – RSS y ATOM – indistintamente.

Existe todo un nuevo espacio de publicación que puede ser accesible a través de servicio de sindicación de contenido, esto permite trabajar de manera opuesta a la idea original de publicar en un sitio Web [Hammond, 2004] en el que los usuarios deban obligatoriamente visitar. Por ejemplo, diarios como Clarín[12] o La Nación[13], entre otros, utilizan estas herramientas para presentar noticias a sus lectores. El primer diario en utilizar sindicación mediante canales RSS fue el New York Times, este fue un hecho clave para la consolidación de esta tecnología.

También existen buscadores verticales que operan sobre espacios acotados como: Technorati[14], Google Blog Search[15], Feedster[16] y operan bajo un enfoque centralizado.

En este trabajo se plantea la utilización de canales de sindicación como alternativa al *crawling*, donde el factor tiempo es muy importante. En otras palabras, es una alternativa que permite incorporar a los SRI el factor de actualidad de la información publicada y también una opción para la construcción de colecciones en ambientes pseudo-cooperativos como son los de sindicación.

[12] http://www.clarin.com
[13] http://www.lanacion.com
[14] http://www.technorati.com/
[15] http:// blogsearch.google.com
[16] http://www.feedster.com

1.8. Objetivos

El área de recuperación de información distribuida ha sido protagonista de diversos proyectos de investigación en el último tiempo. La evolución de las redes de datos y los nuevos ambientes de información dinámicos le han dado un nuevo contexto donde poder adaptar y aplicar modelos tradicionales de la disciplina.

Las primeras investigaciones en RID se realizaron operando principalmente en ambientes cooperativos, siendo muy pocas las que se ocuparon de ambientes no cooperativos. En este trabajo de investigación se pone énfasis en ambientes "híbridos" o semi-cooperativos donde se proponen nuevas alternativas para las dos primeras problemáticas – construcción de descripciones de recursos y selección de recursos.

En relación a la primer problemática de RID, la representación de recursos, la obtención de descripciones en ambientes controlados y cooperativos es una tarea simple. En ellos se encuentran disponibles todas sus estadísticas y su tamaño (cantidad de términos, frecuencia, tamaño en cantidad de documentos, etc.). No ocurre lo mismo en ambientes no cooperativos, no es posible contar con esta información y por lo tanto debe ser estimada. En este trabajo se propone y desarrolla un modelo de base de datos textual orientado a ambientes dinámicos, con fuentes heterogéneas y con una alta tasa de actualización.

Con la segunda problemática de RID, selección de recursos, la frecuencia de actualización de las fuentes es una problemática que no se contemplaba todavía a la hora de realizar la selección de recursos. Entonces, se propone y evalúa una versión extendida del algoritmo de selección de recursos CORI [Callan, 1995], y se lo adapta para que incorpore al cálculo del *score* de *rankeo* la actualidad de las publicaciones de una fuente. Esto en un ambiente dinámico y con fuentes de información heterogéneas.

Por lo tanto, el objetivo principal de este trabajo es proponer y evaluar un modelo de recuperación de información distribuida para operar en ambientes dinámicos en los cuales la actualidad de la información es una característica para la relevancia.

Para el alcance de este objetivo se cumplimentaron las siguientes tareas específicas:

- Diseñar e implementar una herramienta que permita generar descripciones de recursos de fuentes heterogéneas en español.
- Diseñar e implementar las modificaciones del algoritmo de selección de recursos CORI para trabajar con fuentes con una alta tasa de actualización.
- Adaptar una medida de evaluación que permita comparar los resultados obtenidos con la versión original del algoritmo vs la versión modificada.
- Diseñar un experimento de prueba basado en un conjunto de consultas para someter a evaluación al algoritmo modificado.