ANÁLISIS COMPARATIVO DE MOTORES DE BÚSQUEDA: UN ENFOQUE BASADO EN TÉCNICAS DE CALIDAD DE DATOS

Tesis de Licenciatura en Ciencias de la Computación
Departamento de Computación
Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires

Agosto de 2005

Autoras:
Mariela Rodríguez L.U.: 69/93
Miriam Rosenberg L.U.:339/95
{mr9z, mrosenbe}@dc.uba.ar

Directora:
Lic. Mónica Bobrowski

Web

Expresamos nuestro agradecimiento a todas las personas que de alguna manera colaboraron y ayudaron a realizar esta tesis y culminar nuestra carrera:

A nuestra directora Lic. Mónica Bobrowski, por su dedicación y valiosos aportes. A la Dra. Martina Marré, quien participó activamente como directora en los comienzos y en el desarrollo de este trabajo.

A Nicolás Bruno, por abundantes comentarios de dudosa relevancia y fundamentalmente por explicarnos pacientemente cómo redactar esta oración.

A todas las personas que participaron en el experimento, por su gran aporte a esta investigación.

A nuestros compañeros y profesores de la U.B.A., por los inolvidables momentos compartidos durante este período. A nuestros amigos, por alentarnos constantemente.

Para finalizar, les agradecemos especialmente a nuestros padres Sara, Jaime, Mirtha y Héctor, a Fabián y a nuestros familiares por su ayuda, apoyo y contención durante la carrera y el resto de nuestras vidas.

RESUMEN

La World Wide Web (WWW) es una nueva fuente de información en ambientes académicos e industriales. A diferencia de servicios tradicionales, la WWW presenta características únicas en cuanto al proceso de recuperación de la información. No sólo es órdenes de magnitud más grande que bibliotecas digitales tradicionales, sino que su estructura es altamente descentralizada y su contenido cambia rápidamente. Adicionalmente, el conjunto de usuarios de la WWW no está restringido a expertos en el área de Recuperación de la Información. Por estos motivos, una nueva generación de herramientas, denominadas motores de búsqueda, ha aparecido para ayudar a usuarios a encontrar información relevante en la WWW.

Los motores de búsqueda son sistemas complejos y dependen de muchas variables. Evaluar la calidad de un motor de búsqueda o comparar dos motores alternativos es una tarea complicada. Esta tesis define un método que permite medir la calidad de motores de búsqueda de la WWW, facilitando el análisis comparativo de los mismos. Específicamente, se aplica una técnica de Ingeniería de Software, la metodología GQM (*Goal Question Metric*), para comprender qué se desea evaluar y cómo será implementada esta evaluación. El método desarrollado en este trabajo de investigación está basado en funcionalidades y métricas de calidad de datos y se centra en consultas de texto.

Luego de analizar diferentes buscadores desde el punto de vista técnico y principalmente desde la perspectiva de calidad del usuario, se muestra que los motores de búsqueda actuales no alcanzan a cumplir requisitos mínimos de calidad. Esta tesis identifica puntos débiles en el proceso de recuperación de la información usando motores de búsqueda y propone soluciones tendientes a mejorar la calidad de los mismos.

ABSTRACT

The World Wide Web (WWW) is a new information source in both academic and business environments. The WWW differs from traditional services as it has unique characteristics with regards to the information retrieval process. Besides being orders of magnitude bigger than traditional digital libraries, its structure is highly decentralized and its content changes rapidly. Additionally, the user set of the WWW is not restricted to experts in Information Retrieval. As a result of this, a new generation of tools known as search engines appeared to help users find relevant information on the Web.

Search engines are complex systems that depend on many variables. Evaluating their quality in isolation or comparing them is a hard task. This thesis defines a method to measure the quality of Web search engines, facilitating the comparative analysis between them. Specifically, a Software Engineering technique known as Goal Question Metric (GQM) is applied to understand what will be evaluated and how this evaluation will be implemented. The method developed in this investigation is based on Data Quality features and metrics, and focuses on text queries.

After analyzing different search engines from both the technical side and from the end user's quality perspective, this study proves that state of the art search engines do not meet minimum quality requirements. This thesis identifies weak points in the information retrieval process using search engines, and proposes solutions to improve their quality.

ÍNDICE

RESUMEN i
ABSTRACT ii
ÍNDICE iii
CAPITULO 1. INTRODUCCION 1
1.1 Introducción 1
1.2 Internet vs. WWW 2
1.3 Cronología 5
1.4 Motivo y Objetivo de la Investigación 6
1.4.1 Objetivo 7
1.5 Estructura del Trabajo 8
1.5.1 Etapas 8
CAPITULO 2. EL PROBLEMA DE LA CALIDAD DE DATOS 10
2.1 Introducción 10
2.2 ¿Qué Es la Calidad? 10
2.3 Discusión: la Calidad del Software 12
2.4 La Importancia de la Calidad de Datos 13
2.5 Problemas de Calidad de Información 14
2.5.1 Factores Causantes 14
2.5.2 Problemas de Calidad de Datos 15
2.6 Métricas de Calidad 17
2.6.1 Métodos de Evaluación de Métricas 18
2.7 Calidad de Datos y Motores de Búsqueda 19
2.7.1 Cuando la Mala Calidad Pasa Inadvertida 20
CAPITULO 3. RECUPERACION DE LA INFORMACION DE LA WEB 21
3.1 Introducción 21
3.2 La Perspectiva Histórica 21
3.3 Recuperación de la Información y la Web 22
3.4 Arquitectura de Motores de Búsqueda 22
3.4.1 Arquitectura Centralizada 22
3.4.2 Arquitectura Distribuida 24
3.5 Componentes de un Motor de Búsqueda 25
3.5.1 GUI 25
3.5.2 Interface de Búsqueda 25
3.5.3 Interface de Resultados 27
3.5.4 El Ranking 27
3.6 Técnicas de *Crawling* 31
3.7 Indexación 31
3.8 Métricas para R.I.W. 32
3.8.1 Métricas sobre el Corpus de la Web 33
3.8.2 Métricas sobre los Motores de Búsqueda 40
3.9 Problemas de R.I.W. 45
CAPITULO 4. METRICAS: *GOAL QUESTION METRIC* 52
4.1 Introducción 52
4.2 ¿Qué Son las Métricas? 52

4.3 Objetivos de las Métricas 53
4.4 Representación Teórica 53
4.4.1 Las Reglas de la Relación 53
4.5 *Goal Question Metric* 54
4.5.1 Nivel Conceptual (Objetivo) 55
4.5.2 Nivel Operacional (Preguntas) 55
4.5.3 Nivel Cuantitativo (Métricas) 55
4.5.4 *Template* para la Definición de Objetivos [FP97], [S98], [BR88]: 55
4.5.5 Pasos del *Framework* GQM 55
CAPITULO 5. DERIVACION DE METRICAS 57
5.1 Introducción 57
5.2 Entidad a Describir 57
5.3 Categorías de un Motor de Búsqueda - Atributos Generales 57
5.4 Explotación de las Categorías - Atributos Específicos 59
5.5 Funcionalidades Deseables de un Motor de Búsqueda 61
5.5.1 Funcionalidades desde la Perspectiva del Usuario 62
5.5.2 Funcionalidades desde la Perspectiva Técnica 65
5.6 Aplicación de GQM para la Derivación de Métricas 67
CAPITULO 6. DISEÑO DEL EXPERIMENTO 75
6.1 Introducción 75
6.2 Definición del Conjunto Mínimo de Funcionalidades de Usuario 75
6.2.1 Justificación de la Elección del Conjunto Mínimo de Funcionalidades 76
6.3 Definición del Conjunto Mínimo de Funcionalidades Técnicas 76
6.3.1 Justificación de la Elección del Conjunto Mínimo de Funcionalidades 77
6.4 Definición de Calidad de un Motor de Búsqueda 77
6.4.1 La Perspectiva del Usuario 77
6.4.2 Perspectiva Técnica 78
6.5 Factores que Influyen en la Calidad de los Motores de Búsqueda 79
6.6 Objetivo 80
6.6.1 Justificación de la Elección del Conjunto de Motores de Búsqueda 80
6.7 Objeto de Estudio 81
6.8 Marco Muestral 81
6.8.1 Análisis 81
6.9 Diseño del Experimento 81
6.10 Ejecución del Experimento 82
6.11 Formulario del Experimento 82
CAPITULO 7. RECOLECCION Y ANALISIS DE DATOS 83
7.1 Introducción 83
7.2 Recolección de Datos 83
Google 84
AllTheWeb 87
MSN 90
Altavista 93
About 96
Recolección de Datos Adicionales 98
7.3 Análisis de Datos 104
7.4 Evaluación de la Hipótesis 109

7.4.1 Comparación de los Motores de Búsqueda por Funcionalidad 111
7.5 Mejoras Propuestas 112
7.6 Conclusiones del Experimento 116
CAPITULO 8. CONCLUSIONES 117
8.1 Introducción 117
8.2 Conclusiones de la Investigación 117
8.3 Trabajos Futuros 119
BIBLIOGRAFIA 121
II.1 Introducción 147
II.2 Usuarios 147
II.3 Experiencia Previa 147
II.4 Confiabilidad 148
II.5 Interfaz Gráfica de un Motor de Búsqueda 150
II.6 Interfaz de Resultados de un Motor de Búsqueda 153
Sinopsis Comparativa 161
Apéndice I. FORMULARIO DEL EXPERIMENTO 129
Apéndice II. RECOLECCION DE DATOS 147
Apéndice III. INTERFACES DE MOTORES DE BUSQUEDA 161
Apéndice IV. RECUPERACION DE LA INFORMACION 175
Apéndice V. ANALISIS ESTADISTICO 192

CAPITULO 1. INTRODUCCION

"What is needed for a modern world educational structure is a new centralising and unifying organ, which I have called the Permanent World Encyclopaedia"

Orson Wells

1.1 Introducción

Los medios de información y los recursos disponibles para acceder a la misma se han multiplicado y diversificado en los últimos años, incluyendo Internet, la *World Wide Web* (WWW o Web) y sus motores de búsqueda. Uno de los principales factores limitantes para la búsqueda de información de calidad es el tiempo. Actualmente, la industria sabe que tener la información correcta en el momento adecuado aporta grandes beneficios. Estos datos suelen ser almacenados en grandes bases de datos y administrados mediante aplicaciones de software. Sin embargo, no es suficiente contar con buenas aplicaciones. Una empresa precisa buena información para lograr sus objetivos. Esta meta se torna difícil de alcanzar al considerar la gran cantidad de información existente, sus fuentes y sus niveles de calidad.

La Web es una de las principales fuentes de búsqueda de información utilizada en la academia y la industria. Esto ha contribuido al crecimiento exponencial en el uso de la misma en un corto plazo. Se estima que la cantidad de datos disponible en la WWW es de 167 terabytes. Asimismo, otros medios como audio, imágenes y video también están disponibles. Es por ello que la Web puede ser vista como una gran base de datos sin estructura, para la cual se precisan herramientas eficientes de manejo, recuperación y filtro de información. Al utilizar un motor de búsqueda como herramienta, la calidad de los documentos recuperados se encuentra altamente influenciada por los problemas mencionados anteriormente, tales como el tamaño y la falta de estructura de la Web, la experiencia del usuario al interactuar con los buscadores, la calidad del motor y de los documentos de Internet y el tiempo como factor limitante, entre otros. Por lo tanto, para determinar qué buscador es el más adecuado para utilizar en un caso particular, es preciso comparar diferentes motores. Esto se puede lograr evaluando su calidad, de acuerdo a las características particulares que se buscan en este caso, de manera que al usar un buscador se obtengan los "mejores" datos de acuerdo a lo que se precisa. Este análisis comparativo se debe realizar con una muestra significativa de usuarios reales y en base a las funcionalidades de los buscadores agrupadas por categorías. Para ello se necesitan métricas que ayuden a evaluar la calidad de los motores de búsqueda, tanto en lo relacionado con la recuperación de la información como en lo que respecta a la interacción con los usuarios. Tal como describen Fenton y Pfleeger en [FP97], medir es realizar un mapeo desde el mundo empírico al formal o relacional. Una métrica es el número o símbolo asignado a una entidad por medio de este

mapeo para caracterizar a un atributo. Por lo tanto, disponer de métricas para los motores de búsqueda permite compararlos y así determinar cuál es el adecuado para obtener la información que cada tipo de usuario precisa.

Hoy en día no existen métodos estándar y optimizados para medir la calidad de los buscadores. Actualmente, se utilizan técnicas muy complicadas de evaluar y casi imposibles de llevar a la práctica porque se basan principalmente en la precisión y recuperación, otorgándole menor importancia a otras funcionalidades de los buscadores, como la interfase de usuario (GUI) y la completitud. Esto se debe a que las métricas más conocidas para analizar la calidad de un buscador involucran la evaluación de todo el universo de búsqueda (i.e.: todos los documentos de la Web), ente muy amplio y casi imposible de evaluar en la práctica, dadas las dimensiones de la WWW y su continuo crecimiento.

Además de los aspectos técnicos de los motores de búsqueda, la experiencia del usuario en cuanto a la utilización de los mismos es un campo muy importante en el que hay todavía mucho por investigar. En general, las publicaciones científicas y las mejoras que se realizan a los motores se basan en el punto de vista técnico, dejando a un lado las necesidades del usuario y su experiencia. Es por ello que el eje y aporte principal de este trabajo consiste en analizar la calidad de los buscadores desde la perspectiva de los usuarios.

1.2 Internet vs. WWW

Internet y la *World Wide Web* son términos que suelen ser confundidos con frecuencia. Dado que éstos son clave para esta tesis, se definen a continuación:

Internet Gran red de redes que están interconectadas; una infraestructura de redes de comunicación. Hoy en día, Internet facilita que millones de computadoras conectadas a ella alrededor del mundo, se comuniquen entre sí intercambiando información a través de sus protocolos. La figura 1.1 muestra un diagrama del diseño de Internet. Este diseño jerárquico consiste en las subredes principales demarcadas por circunferencias que se denominan sistemas autónomos (S.A.). Los S.A. también se conocen como los *backbones* (i.e., columnas vertebrales), los cuales representan continentes como Europa, EE.UU. y Asia. Estos se conectan entre sí mediante *routers* y líneas de alto ancho de banda, tal como se observa en la figura. Las redes regionales se conectan a los *backbones* y a su vez, las LANs (i.e., redes de área local) a las regionales. El protocolo que facilita la comunicación entre los componentes se denomina *Internet Protocol* (IP).

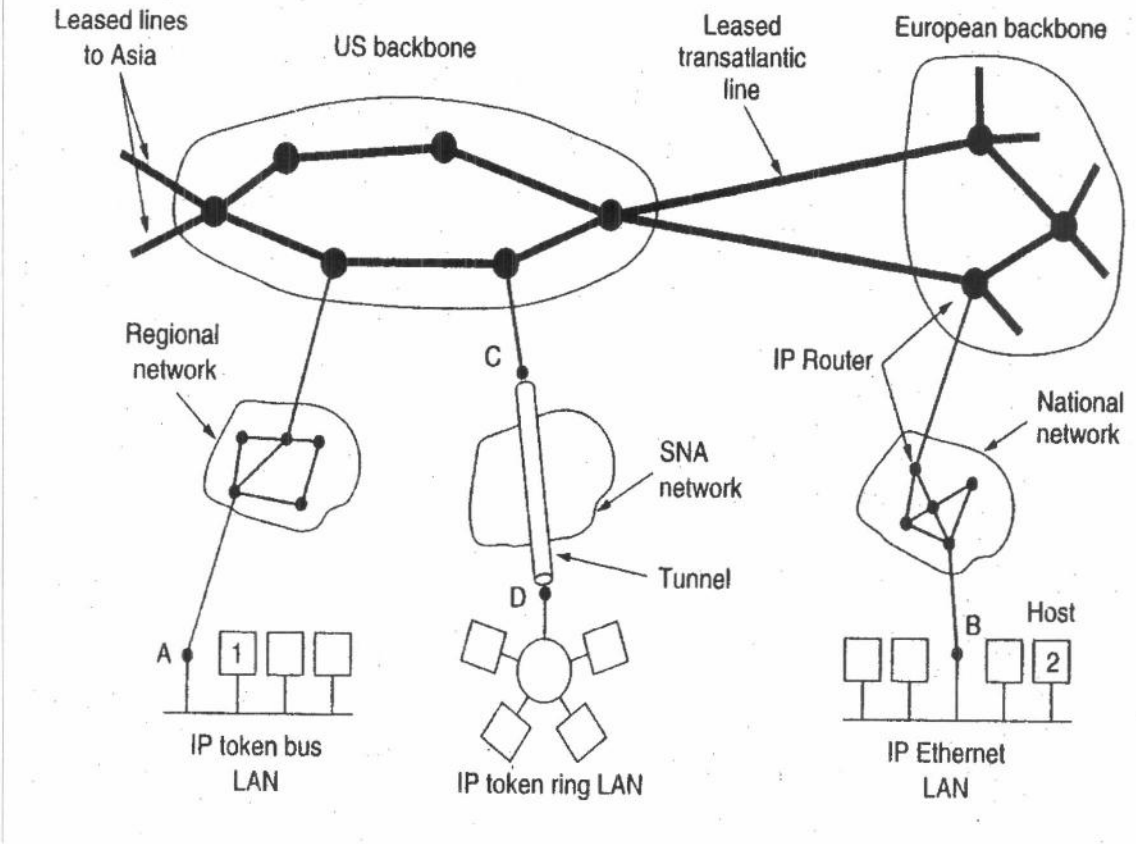


Figura 1.1 – Arquitectura de Internet [TA96].

***World Wide Web* (WWW)** Una de las tantas aplicaciones construidas sobre Internet, así como e-mail y FTP, entre otras. El objetivo de la WWW es facilitar el acceso público de usuarios y computadoras a la información almacenada en las computadoras pertenecientes a esta gran red. Esta información se transfiere mediante el protocolo HTTP. La WWW fue concebida en 1989 por Tim Berners-Lee del Instituto CERN [BL89].

Los usuarios acceden a las páginas o documentos de la WWW a través de un explorador, como Firefox, Netscape o Internet Explorer. Las páginas se conectan entre sí mediante hiperenlaces [WE05].

La siguiente figura muestra la arquitectura de la WWW tal como fue propuesta por Berners-Lee entre 1989 y 1990. La flecha central representa los estándares comunes utilizados en la Web (URL, HTTP y negociación del formato de datos).

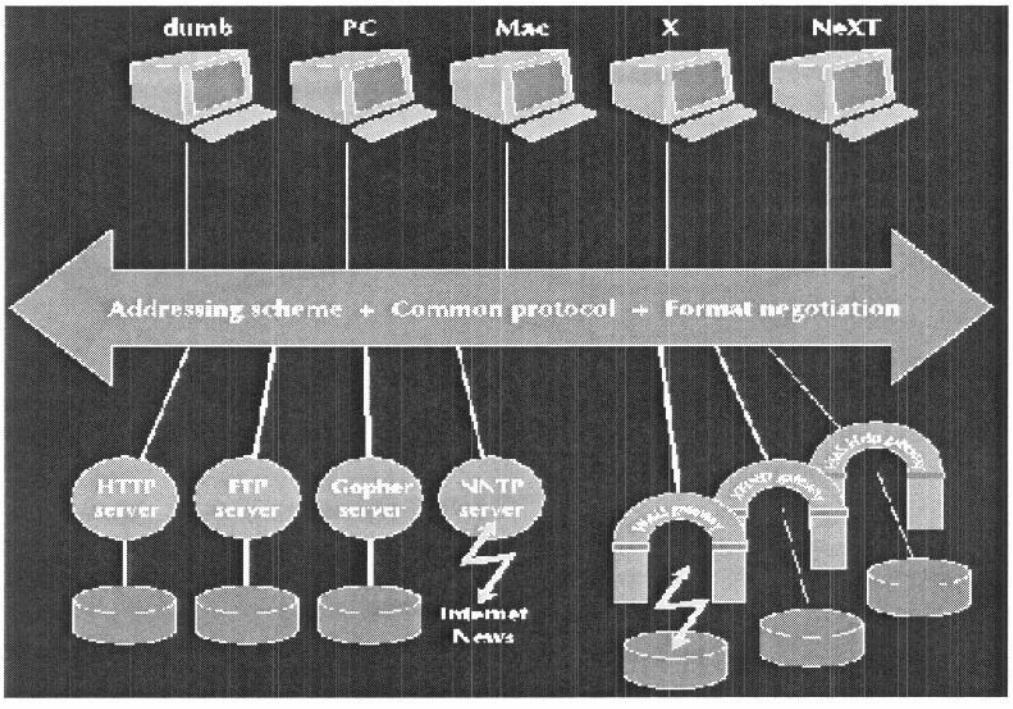


Figura 1.2 – Diagrama original de la arquitectura de la WWW [BL90].

1.3 Cronología

La siguiente figura describe los hitos clave en la historia de Internet y la WWW. Los datos fueron obtenidos de [ZA05].

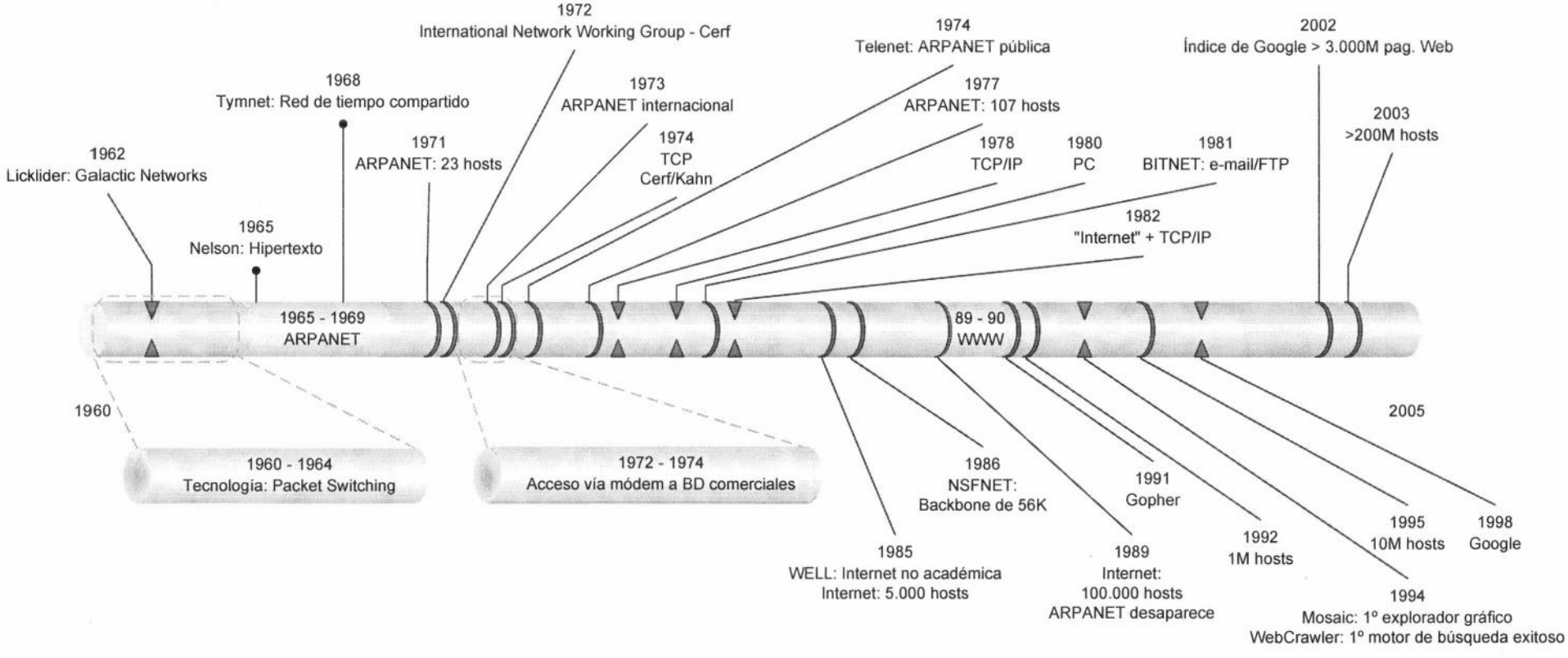


Figura 1.3 – Cronología de Internet y la WWW.

1.4 Motivo y Objetivo de la Investigación

Esta tesis define un método para medir la calidad de los motores de búsqueda de la Web, facilitando de este modo el análisis comparativo de los mismos. Este método está basado en funcionalidades y métricas de calidad de datos y se centra en consultas de texto. Tal como se mencionó anteriormente, este trabajo se enfoca en los motores de búsqueda, analizándolos principalmente desde la perspectiva de calidad del usuario. Se analiza también el punto de vista netamente técnico pero con un menor nivel de profundidad.

A continuación se mencionan los principales problemas relacionados con Recuperación de la Información, la interacción de los usuarios con los buscadores y la calidad de los datos. Estos problemas influyen directamente en la calidad de la información de la Web y, de este modo, constituyen el motivo de este trabajo.

Existen varios problemas relacionados con la recuperación de la información en la Web:

- El volumen de los datos en sí mismo
- La actualización de los datos
- La interacción de los usuarios con los sistemas de recuperación
- Otros

Algunos problemas relacionados con los usuarios son:

- Interacción con las funcionalidades de los motores de búsqueda
- Especificación de la consulta
- Interpretación de la consulta mediante el análisis de los resultados devueltos

Algunos problemas relacionados con los datos son:

- Datos distribuidos
- Alto porcentaje de datos volátiles
- Gran volumen
- Falta de estructura
- Redundancia
- Falta de editorialización que podría implicar falsedad de documentos
- Datos incorrectos
- Datos imprecisos
- Datos obsoletos

Estos problemas afectan directamente la calidad de los datos en la Web. Asimismo, la calidad de datos en general es un problema por resolver en Ingeniería de Software. Como fue mencionado anteriormente, la academia y la industria dependen fuertemente de la calidad de los datos que utilizan, ya que estos aportan a la toma de decisiones y son fundamentales para su desarrollo.

Dado que este trabajo se centra en la calidad de los motores de búsqueda, es preciso estudiar el área de Recuperación de la Información (R.I.), cuyo objetivo es satisfacer la necesidad del usuario mediante la obtención de datos relevantes a su problema. A partir de la investigación realizada en este campo surgieron diversas métricas como *precisión* y *recuperación*, las cuales

estiman la calidad de los resultados obtenidos por un motor de búsqueda. Estas métricas se definen a continuación:

- Precisión = $\frac{\#(\text{documentos relevantes})}{\#(\text{documentos recuperados})}$

- Recuperación = $\frac{\#(\text{documentos relevantes recuperados})}{\#(\text{documentos relevantes conocidos})}$

Si bien estas métricas son las más importantes en el campo de R.I., resultan imposibles de evaluar actualmente en el universo de la WWW. Esto se debe a que es necesario conocer todo el universo de búsqueda para calcularlas, es decir, todas las páginas Web. Por lo tanto, es necesario obtener nuevas métricas que se adecuen al estado del arte de los motores de búsqueda. Con este objetivo se estudia y aplica la metodología *Goal Question Metric* (GQM) [BR88, FP97, SB99], la cual permite derivar métricas en función de lo que se desea evaluar. De esta manera, se obtienen nuevas fórmulas para medir la calidad de los datos de un motor de búsqueda.

1.4.1 Objetivo

Este trabajo propone una metodología o forma de trabajo para comparar y evaluar motores de búsqueda en base a la calidad de los datos que proporcionan y a la experiencia del usuario al interactuar con los mismos.

El objetivo es por un lado analizar la parte técnica y por otro centrarse de forma exhaustiva en la experiencia del usuario y en cómo mejorarla. Para ello, se definen métricas de calidad que abarcan ambos puntos de vista.

A fin de proponer un modelo de calidad de motores de búsqueda, es necesario comprender qué se desea evaluar y luego definir métricas que permitan hacerlo. Para lograr este objetivo, se aplica una metodología conocida de Ingeniería de Software: GQM (*Goal Question Metric*), tal como fue indicado anteriormente.

En el marco de este trabajo, se determinan dos conjuntos de funcionalidades de calidad deseadas para motor de búsqueda:

- Funcionalidades de usuario
- Funcionalidades técnicas

La calidad de un motor se define en base a lo siguiente:

- Elección de un subconjunto de funcionalidades esenciales para el mismo.
- Definición de niveles para las funcionalidades.
  - De este modo, un motor pertenecerá a un nivel específico para cada funcionalidad o atributo. Un nivel especifica las características que posee un motor de búsqueda para una funcionalidad determinada. Es decir, un buscador está en el nivel Y para la funcionalidad X si cumple todas las características de Y para X. Se define asimismo el conjunto minimal de funcionalidades que un motor debe poseer y los

diferentes niveles para las mismas. Los niveles se derivan de las métricas, las cuales fueron obtenidas mediante GQM.

1.5 Estructura del Trabajo

Esta tesis se divide en las siguientes etapas a fin de alcanzar el objetivo detallado anteriormente.

1.5.1 Etapas

1. Estudio y análisis de calidad de datos y de recuperación de la información en la Web.
2. Definición de categorías y funcionalidades deseables de un motor de búsqueda desde el punto de vista del usuario y desde el punto de vista técnico.
3. Aplicación de GQM para la derivación de métricas sobre las funcionalidades definidas.
4. Definición del conjunto minimal de funcionalidades de usuario y técnicas deseables en un motor de búsqueda. Definición de niveles para cada funcionalidad perteneciente a los conjuntos minimales.
5. Diseño de un experimento sobre el conjunto minimal de funcionalidades de usuario.
6. Evaluación del experimento.
7. Análisis y conclusiones sobre el experimento.
8. Sugerencias para trabajos futuros.

A fin de cubrir las etapas mencionadas, este documento se estructura como se indica en la tabla 1.1.

Capítulo	Título	Descripción
1	Introducción	Descripción del motivo y objetivo de la tesis y breve introducción teórica.
2	El Problema de la Calidad de Datos	Introducción a Calidad de Datos.
3	Recuperación de Información de la Web	R.I. aplicado a la World Wide Web (R.I.W.), métricas de RIW y problemas de los motores de búsqueda.
4	Métricas – *Goal Question Metric*	Definición de métricas, su utilización y descripción del método GQM.
5	Derivación de Métricas	Especificación de funcionalidades y aplicación de GQM a los atributos de un motor de búsqueda.
6	Diseño del Experimento	Descripción del experimento.
7	Análisis y Recolección de Datos	Recolección de la información obtenida mediante el experimento. Comparación de los motores de búsqueda y demostración de la hipótesis definida en el capítulo 6.
8	Conclusiones	Conclusiones de la tesis e ideas para trabajos futuros.
Bibliografía	Bibliografía	Bibliografía utilizada en esta tesis.
Apéndice I	Formulario del Experimento	Experimento realizado por los usuarios.
Apéndice II	Recolección de Datos	Datos planos recolectados como resultado del experimento. Estos datos se analizan en profundidad en el capítulo 7.
Apéndice III	Interfaces de Motores de Búsqueda	GUI de los buscadores utilizados en el experimento.
Apéndice IV	Recuperación de la Información	Introduction a R.I.
Apéndice V	Análisis Estadístico	Breve definición de las medidas de estadística descriptiva utilizadas en el capítulo 7.

Tabla 1.1 – Descripción de los capítulos de esta tesis.

CAPITULO 2. EL PROBLEMA DE LA CALIDAD DE DATOS

"The bitterness of poor quality
remains long after the sweetness
of meeting the schedule has been forgotten."

Anonymous

2.1 Introducción

Tal como se mencionó en el capítulo anterior, este trabajo consiste en analizar la calidad de los motores de búsqueda en base a la calidad de los datos que éstos proporcionan a sus usuarios. Por este motivo, es necesario estudiar qué se entiende por calidad de datos.

En consecuencia, este capítulo describe el problema de la calidad de datos, sus dimensiones y su relación con los sistemas de software. Se estudian asimismo métricas de calidad y su aplicación para la evaluación de datos, particularmente en lo vinculado a la información en la Web. Finalmente, se discute la importancia de la calidad de los datos en Internet, los problemas de recuperación de información (R.I.) de calidad y la posibilidad de aplicar métricas tradicionales de calidad a R.I. y a motores de búsqueda.

2.2 ¿Qué Es la Calidad?

Las definiciones del término "calidad" varían según los individuos y profesiones involucradas. Se recurre primero al Diccionario de la Real Academia Española para buscar una definición "objetiva":

> I. **Calidad** Propiedad o conjunto de propiedades inherentes a una cosa, que permiten apreciarla como igual, mejor o peor que las restantes de su especie. (...) 2. En sentido absoluto, buena calidad, superioridad o excelencia. (...) 4. Condición o requisito que se pone en un contrato [RAE].

Las siguientes definiciones están relacionadas con la calidad de software:

> II. **Calidad de software** La totalidad de aspectos y características de un producto o servicio que se sustentan en su capacidad de cumplir las necesidades especificadas o implícitas [I86].
>
> III. **Calidad de software** El producto debe cumplir con su especificación [C79].
>
> IV. **Calidad de software** Cumplir con los requerimientos de alguna persona.

Calidad es valor para alguna persona. Valor es aquello que está dispuesto a pagar para obtener sus requerimientos [W92].

V. **Calidad de software** Aquellas funcionalidades de los productos que cumplen con las necesidades del usuario y por lo tanto proveen la satisfacción del cliente [JG99].

VI. **Calidad de software** Libre de deficiencias [JG99].

Finalmente, la calidad de datos se define de la siguiente forma:

VII. **Calidad de datos** Un dato o colección de datos X es de mayor (o mejor) calidad que un dato o colección de datos Y si X cumple con las necesidades de usuario mejor que Y [Red96].

La definición II. indica que la calidad de software se basa en gran parte en los atributos o características de un producto.

De acuerdo a III., se debe prestar especial atención a la etapa de relevamiento de requerimientos a fin de obtener un software de calidad. Los requerimientos deben definirse formal y explícitamente a fin de poder medir la calidad de software resultante al final del proyecto y, de este modo, asegurar que los requerimientos se cumplieron adecuadamente. De igual manera, se debe controlar la evolución de los mismos en cada etapa del ciclo de desarrollo para asegurar que éstos continúan en la senda adecuada. En resumen, la calidad de un software se puede establecer si y sólo si se definen y controlan requerimientos formales. Tal como lo indica Sommerville [Som01], ya no es aceptable crear productos de baja calidad y reparar las deficiencias luego de la entrega al cliente. El software debe ser tratado como cualquier otro producto en lo que respecta a calidad (electrodomésticos, autos, etc.)

La cuarta definición agrega un aspecto muchas veces olvidado: los requerimientos deben ser acordados y estar claros tanto para el usuario como para el equipo de desarrollo. Si se cumplen los requerimientos de calidad importantes para el programador pero no los del usuario, el sistema final no habrá cumplido su objetivo.

La definición V. está orientada al nivel de satisfacción del cliente y, en último término, a incrementar las ganancias de una empresa. Paradójicamente, a fin de obtener un producto de mayor calidad se debe invertir más en el mismo, lo que aumenta su costo y, consecuentemente, disminuye la ganancia final. Sin embargo, dado que mayor calidad implica menor costo de mantenimiento, el retorno sobre la inversión (ROI) es positivo.

La penúltima definición se refiere a la ausencia de errores como un factor determinante de la calidad. Esto se puede vincular a la definición anterior al considerar que las fallas afectan directamente la satisfacción del usuario final y también el mantenimiento del producto a largo plazo (i.e., el costo para la empresa).

La última es una definición operacional que nos muestra que la calidad de datos es subjetiva, ya que los clientes son los jueces de la misma. Asimismo, los usuarios finales de la

información no siempre pueden indicar sus requerimientos claramente y éstos evolucionan con el transcurso del tiempo, modificando así su definición de calidad. Los atributos que determinan la calidad de los datos se analizan en la sección 2.6. Luego de definirlos, se deben establecer métricas para los mismos, de modo de poder determinar si los datos poseen el nivel de calidad deseado (e idealmente establecido en los requerimientos) para cada una de las características. Ciertas definiciones y métricas para los atributos varían dependiendo del medio en el que sea utilizado el software y los datos asociados al mismo. Por ejemplo, los *problemas de formato*, la *volatilidad* de los datos y las métricas utilizadas para analizarlos, no son las mismas para un motor de búsqueda que para una central nuclear. Disponer de medidas de calidad para los atributos permite determinar la calidad del software al que pertenecen.

2.3 Discusión: la Calidad del Software

Como se observó en la sección anterior, la calidad posee más de un aspecto. Las diferentes perspectivas de la misma varían dependiendo de las personas, profesiones e industrias involucradas. Los niveles de abstracción y los puntos de vista son múltiples, contribuyendo a que cada interlocutor tenga una interpretación propia del significado de la calidad. Es decir, ésta tiende a ser subjetiva. Luego de haber definido calidad en general, calidad de software y de datos, se profundiza en este punto la relación entre la calidad y el software.

Asimismo, la definición de calidad y sus aspectos son diferentes para las diversas profesiones y disciplinas. Por ejemplo, la calidad de un sillón no se define del mismo modo que la calidad de una computadora. Las definiciones y niveles de calidad varían incluso dentro de una misma profesión: en Ciencias de la Computación, la calidad de un procesador de palabras tiene otros atributos que la de un motor de búsqueda.

La ambigüedad existente al referirse a "alta" o "baja" calidad suele facilitar conversaciones coloquiales pero no es aceptable para el ámbito profesional. Se deben establecer factores de calidad y métricas formales para los mismos. Como lo indica Tom DeMarco en [DM82], "no se puede controlar lo que no se puede medir". El control de calidad resultante permite establecer un plan para alcanzar los niveles de calidad deseados.

La "baja" calidad domina la mayor parte de la industria. Los usuarios reciben frecuentemente facturas que incluyen cargos por servicios cancelados, como "facturación detallada", o boletas de servicios de teléfonos celulares que fueron cancelados hace más de sesenta días. Los ancianos suelen escuchar la ya famosa frase "se cayó el sistema" cuando acuden al banco a cobrar sus jubilaciones y el correo continúa enviando correspondencia para el ex dueño del departamento que un familiar compró hace un año. Estas situaciones son la consecuencia de la ausencia de niveles de calidad adecuados en los sistemas de software. Al analizar la variación del costo del arreglo de errores en el software con respecto al tiempo, se obtiene el siguiente gráfico:

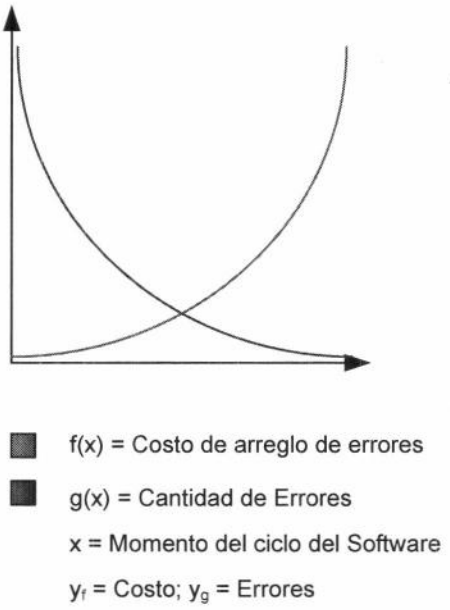


Figura 2.1 - Errores en el *software* y su costo con respecto al momento en el proceso

Además del costo de arreglo de errores o problemas de calidad, se debe considerar: el costo de envío de correspondencia a destinatarios incorrectos, el costo de los sistemas de soporte a usuario, el precio de los materiales utilizados innecesariamente y los que se precisan para reparar los errores cometidos, el tiempo utilizado incorrectamente, los sueldos de empleados invertidos en corregir los errores ocasionados por la baja calidad y las ventas perdidas mientras el sistema no está disponible, entre otros. Asimismo, los costos de la ausencia de calidad afectan la credibilidad e imagen de la empresa que utiliza el software en cuestión.

2.4 La Importancia de la Calidad de Datos

Habiendo discutido qué es calidad en general y qué es calidad de software, se diserta a continuación sobre la importancia de la calidad de la información, la cual constituye una parte fundamental de la calidad del software y, por lo tanto, el eje central de este capítulo.

Los seres humanos interactúan diariamente con información, tanto a nivel personal como profesional o científico. Se depende de la misma tanto para aspectos relevantes como para ínfimas partes de nuestras vidas. La información facilita la toma de decisiones, el aprendizaje y el avance científico, entre muchas otras cosas. Estas decisiones no son mejores que los datos en los que están basadas [Red98]. Por lo tanto, es crucial disponer de datos de calidad para tomar decisiones informadas y correctas que pueden determinar el futuro de una empresa o la certeza de una investigación científica.

Originalmente, los datos fueron subestimados y considerados como problemas externos a la industria del software. No se tomaban en cuenta como parte de los bienes o recursos de una

empresa ni se analizaba su posible utilización como una ventaja competitiva [BMY99]. Consecuentemente, un análisis de calidad de un producto no incluía la información o bases de datos que éste utilizaba. Por lo tanto, dicho estudio de calidad resultaba parcial y muy probablemente inexacto.

Redman [Red96] resume en *Data Quality for the Information Age* las consecuencias o problemas que provoca la baja calidad de datos:

- **Persistencia:** la mala calidad de datos no desaparece sin intervención
- **Impacto negativo en la toma de decisiones**
- **Impacto negativo en el éxito de una organización**
- **Induce costos altos e innecesarios**
- **Disminuye la satisfacción con el trabajo**
- **Crea desconfianza** dentro de la organización
- **Dificulta la reingeniería**
- **Afecta las estrategias organizacionales** de largo plazo
- **Disminuye la ventaja competitiva**

Habiendo comprendido la importancia de la calidad de la información, se estudian a continuación los problemas vinculados a la misma, particularmente enfocados en la industria del software.

2.5 Problemas de Calidad de Información

Tal como fue indicado anteriormente, la información es uno de los bienes más valiosos, tanto en el ámbito académico como en el empresarial. Sin embargo, la poca atención que se le brinda resulta en su deterioro paulatino y, últimamente, en su baja calidad.

2.5.1 Factores Causantes

Los principales causantes de problemas de calidad de datos en el software son:

1) Carencia de requerimientos de calidad de datos en la etapa de análisis de requerimientos.
2) Falta de consideración de la información que utilizan los sistemas de software en la fase de validación y verificación de los mismos.
3) Carencia de métricas formales para calidad de datos.
4) Ausencia de evaluaciones periódicas de calidad.

Estos factores se explican en detalle a continuación:

1) Muy raramente se consulta al cliente sobre el nivel de calidad que éste requiere para la información utilizada por su nuevo sistema de software. Esta brecha en el relevamiento provoca grandes inconvenientes a largo plazo, ya que es mucho más complicado modificar un sistema ya codificado que crearlo desde el principio con los requerimientos de calidad de información en mente.

2) Al no considerar los datos en la etapa de análisis de requerimientos, el flujo natural del ciclo de software determina que éstos tampoco se incluyan en la fase de validación y verificación de un sistema. Este hecho empeora aún más la situación existente, ya que permite que información de baja calidad pase inadvertida y no se corrija.

3) Una de las principales consecuencias de no considerar a los datos como un recurso esencial es que no se aplican a ellos métricas formales de calidad similares a las que se utilizan para medir la calidad del software. Esto se relaciona con el punto anterior, ya que las métricas se utilizan en la etapa de validación y verificación. Sin embargo, se analizan en un punto separado ya que aunque se incluya la información en esta etapa del ciclo de software, no se cuenta con medios formales para hacerlo.

4) Sobre el final del desarrollo de un sistema de software comienza la etapa de mantenimiento, la cual permite mejorar el sistema a través del tiempo y arreglar fallas, defectos y errores. El mantenimiento raramente involucra los datos utilizados. Esto provoca que la baja calidad existente empeore con el transcurso del tiempo hasta que los datos se vuelvan totalmente obsoletos.

Se concluye así que los factores causantes de problemas de calidad de datos son variados y están relacionados entre sí. Si bien se analiza normalmente la calidad de los sistemas industriales, pocas veces se consideran los datos utilizados. Es la calidad de estos últimos la que puede afectar negativamente la performance de sistemas que de otro modo serían de alta calidad.

2.5.2 Problemas de Calidad de Datos

Se describen a continuación los problemas de calidad de datos en los que se incurre cuando la información utilizada no es la adecuada.

a. **Duplicación** Datos repetidos, desincronizados o distribuidos a través del sistema. Esto deriva en múltiples acciones para la misma información. En el caso de clientes, una misma persona puede recibir múltiples llamadas, entregas, cartas, correos electrónicos, facturas, entre otros, lo que deteriora la credibilidad de la empresa e incrementa el costo de comunicación con sus clientes. En el ámbito académico puede provocar inexactitudes en cálculos como el promedio o la mediana. La duplicación se puede prevenir verificando que los datos no existan en el sistema en el momento de su ingreso. En el caso de una base de datos ya existente, se deben identificar n-uplas de información repetida realizando búsquedas por campos fácilmente identificables, como número de teléfono, autor o dirección de correo electrónico.

b. **Incorrectitud** Datos falsos o ingresados con errores. Ejemplos típicos son las direcciones (de correo estándar o electrónico): personas que se mudan o se van del país y sin embargo siguen recibiendo correspondencia a su nombre en sus antiguas direcciones. Otros casos incluyen apellidos tipeados con faltas de ortografía, lo que puede ocasionar que cuando un cliente llame al soporte de una empresa, su información no se pueda recuperar para atenderlo adecuadamente, actualizar sus datos y acceder a su archivo

histórico. En síntesis, disponer de información incorrecta puede resultar en el envío de material a una institución o persona equivocada y en la toma de decisiones inadecuadas, entre otras consecuencias.

c. **Incompletitud** El nivel de detalle no es el adecuado para representar todos los hechos apropiadamente. Hay datos faltantes o abreviados.

d. **Ambigüedad** Ciertos datos tienen más de un significado posible.

e. **Problemas de Formato** Los problemas de formato de datos son variados. A continuación, se presentan algunos de los ejemplos más comunes.

- Sr. Juana García — Género incorrecto
- Sra. Nuria Dezavalía — El apellido correcto es de Zavalía (distinta inicial)
- Sr. JUAN GARCÍA — Capitalización incorrecta

f. **Desactualización** Si los datos no se actualizan con el transcurso del tiempo y con una frecuencia adecuada [BMY99], se tornan irrelevantes rápidamente. Esto implica que no son confiables y, por lo tanto, que no se puede depender de los mismos para la toma de decisiones. Este problema en particular puede causar diversos inconvenientes, como el intentar contactar a una persona fallecida para una campaña de *telemarketing* y provocar así una mala experiencia. Por ejemplo, cuando el cónyuge debe explicar por teléfono esta situación y solicitar que los datos se actualicen.

g. **Inconsistencia** El conjunto de información disponible para X contiene ítems contradictorios entre sí. Si un comercio se muda a un nuevo local, olvidarse de actualizar su dirección y teléfono a tiempo en todas las fuentes de información en las que se lo menciona puede ocasionar la pérdida y/o caída de ventas, que la mercadería y correspondencia no llegue a destino o malgastar el presupuesto de mercadeo.

h. **Subjetividad** La información no es objetiva, lo cual resulta en distintas interpretaciones o incluso la toma de decisiones incorrectas debido a que la opinión de una persona se incluye junto con los datos. Un ejemplo típico de este problema son los datos proporcionados mediante campañas publicitarias o la prensa "amarilla".

i. **Irrelevancia** En ocasiones, los usuarios son bombardeados con información que no es de su interés (e.g., correos electrónicos no deseados o *spam*).

j. **Localización** Los datos están disponibles en idiomas inadecuados para su consumo. Otros casos posibles: la información se presenta en varios lenguajes en un mismo repositorio o la fonética utilizada no es la más apropiada (e.g., *New York* en vez de Nueva York en una base de datos en castellano).

k. **Volatilidad** La existencia de un conjunto de datos en un tiempo determinado no garantiza su persistencia a través del tiempo. El siguiente es un ejemplo reciente de problemas de persistencia de datos en los sistemas de software. La empresa proveedora de servicio de Internet de una de las autoras de esta tesis fue comprada por otra

compañía. Unos días después de anunciada esta adquisición, esta usuaria observó que algunos de sus correos electrónicos habían sido borrados. Con el paso de los días, más mensajes fueron borrados sin advertencia. La autora decidió llamar al soporte técnico de la empresa, el cual le informó que la nueva compañía decidió cambiar la política de persistencia de datos: los correos electrónicos no movidos a carpetas diferentes de la "Bandeja de Entrada" se borrarían automáticamente al mes de su recepción. Los usuarios no fueron advertidos del cambio, perdiendo así mensajes valiosos que no pudieron ser recuperados.

La próxima sección introduce las métricas de calidad de datos y cómo éstas se pueden evaluar para determinar el nivel de calidad de datos de una organización. Esto es relevante ya que en el marco del presente trabajo se derivan métricas que permiten evaluar los motores de búsqueda.

2.6 Métricas de Calidad

El proceso de establecer métricas para el software está directamente relacionado con la calidad de los datos a utilizar, ya que para tomar las decisiones correctas se precisan buenos datos. La creciente utilización del software en todos los ámbitos de nuestras vidas incrementa la importancia de la información que éstos administran. De este modo, poder validar los datos utilizados y medir su calidad resulta cada vez más crucial. Bobrowski, Marré y Yanquelevich sugieren utilizar los siguientes atributos para determinar si los datos en cuestión son de calidad [BMY99]:

Correctitud	Los datos fueron obtenidos siguiendo reglas de definición de métricas.
Exactitud	Similitud entre los datos y el valor real.
Precisión	Los datos son expresados con el nivel de fidelidad y certeza necesario.
Objetividad	Los datos no están afectados por opiniones subjetivas.
Completitud	Toda la información necesaria está disponible.
Medida	La cantidad de información es adecuada.
Relevancia	La información es significativa e interesante para el usuario.
Comprensibilidad	El usuario comprende los datos.
Consistencia	Los datos almacenados no se contradicen entre sí.
Edad	Los datos almacenados tienen una identificación temporal (i.e., *time-stamp*) asociada a los mismos.
Actualidad	La información es actualizada periódicamente para que no pierda su validez con el transcurso del tiempo.
Replicabilidad	El conjunto de datos obtenido empíricamente es reproducible.
Confiabilidad	Los usuarios consideran los datos como válidos o verdaderos.
Valorabilidad	La información es valiosa desde la perspectiva del usuario.
Administrabilidad	La información puede ser manejada fácilmente.
Disponibilidad	Los datos son accesibles por los usuarios del sistema en todo momento.

Seguridad Los datos sólo están disponibles para los usuarios apropiados.

Se deben tener en cuenta los atributos de calidad de datos no sólo para el análisis de la información sino fundamentalmente antes de iniciar el proceso de recolección, de modo que todas las etapas del proyecto puedan validar y verificar la calidad de los mismos. Esto facilita su aplicación para determinar objetivos y tendencias y, asimismo, a fin de obtener conclusiones válidas y útiles para la organización que los administra.

2.6.1 Métodos de Evaluación de Métricas

El programa para la Administración Total de la Calidad de Datos (TDQM) del *Massachussets Institute of Technology* (MIT) [TDQ05] desarrolló un método para evaluar la calidad de los datos de una organización en base a un conjunto de métricas como el que se presenta en este capítulo.

Se debe comenzar evaluando el status quo, i.e., la calidad actual de la información. Esto permite generar una base de partida para poder medir el progreso a través del tiempo y la magnitud del éxito de este proceso. Huang, Lee y Wang de TDQM describen tres formas funcionales que las empresas interesadas en mejorar la calidad de sus datos pueden utilizar a fin de desarrollar un conjunto de métricas de calidad. Las formas funcionales se encuentran fuera del alcance de este trabajo y se detallan en [HLW98].

A continuación, se analiza la importancia de cada métrica y su calidad en base a las opiniones de los usuarios de la información en cuestión. El objetivo es cubrir todos los tópicos principales de información y priorizar las áreas que requieren atención inmediata. De este modo, se obtiene una base para medir las mejoras a través del tiempo.

El próximo paso consiste en utilizar los resultados de la evaluación y enfocarse en proyectos que ayuden a mejorar las áreas más importantes para los usuarios que poseen una calidad más baja. Asimismo, se definen las métricas de éxito.

Luego, se establece una hipótesis sobre la causa de cada problema significativo de calidad y se analiza la causa de raíz (*Root Cause Analysis*) para probar la hipótesis, seguido de una evaluación de costo-beneficio para asegurar que el problema se pueda resolver dentro de los parámetros de costo aceptados por la organización.

Una vez que se creó un plan de implementación detallado para resolver el o los problemas de calidad de datos, se lo ejecuta.

Finalmente, se reevalúan los datos un año más tarde para medir el progreso en cuanto al mejoramiento de la calidad de los datos.

Este proceso es bastante abstracto a simple vista pero su seguimiento formal garantiza que los datos no vuelvan a perder calidad en un futuro cercano.

2.7 Calidad de Datos y Motores de Búsqueda

Este capítulo culmina con una evaluación de la relación entre calidad de información y motores de búsqueda, ya que éstos constituyen el tema central de este trabajo. A los efectos del análisis de calidad de motores de búsqueda y de los datos que éstos proporcionan en particular, se considera el conjunto de documentos de Internet como el conjunto universal de datos a evaluar.

Un simple análisis permite determinar que este conjunto no cumple con varios de los atributos de calidad de datos propuestos anteriormente. En principio, los documentos de la Web no son necesariamente consistentes entre sí.

> Ejemplo: al buscar *aljazeera* en Google, el primer resultado obtenido el día en que el sitio oficial de AlJazeera fue inaugurado en Internet era www.aljazeera.info, el cual no es correcto, ya que www.aljazeera.net es el sitio oficial. Este problema de calidad de datos provocó inconvenientes en la radio durante una entrevista con un empleado de este medio. El locutor del programa en cuestión había accedido al sitio incorrecto a través de Google, el cual contenía información tergiversada. Como consecuencia, se generó una larga discusión sobre la objetividad del sitio cuando el problema real se debió a que el buscador había retornado un documento irrelevante para la búsqueda en la primera posición del *ranking*, documento que a su vez era inconsistente con respecto al sitio oficial de AlJazeera.

Uno de los problemas relacionados con calidad de datos en los motores de búsqueda es que los usuarios de estos sistemas son muy variados y sus niveles de conocimiento son diferentes entre sí. Esto, unido al crecimiento exponencial en la cantidad de usuarios, obstruye el poder acordar cuáles son los requerimientos del software. En base a la definición IV. de calidad, la calidad final es más baja como consecuencia de no poder satisfacer los requerimientos de los usuarios. Asimismo, la gran diferencia entre sus niveles de conocimiento provoca más problemas de calidad de datos. Por ejemplo, si un usuario no domina el idioma inglés, la calidad de los documentos retornados es menor desde su perspectiva, ya que la mayoría de ellos están escritos en este idioma, y éstos no pueden ser aprovechados por el usuario. Si se realiza una búsqueda avanzada configurando el idioma de los documentos a retornar, muy probablemente se pueden perder documentos cruciales para la búsqueda en cuestión ya que gran cantidad de los documentos más importantes sólo están disponibles en inglés. Ciertos motores de búsqueda atacaron este problema de calidad. Google [G97], por ejemplo, agrega un enlace junto a algunos documentos retornados para los cuales está disponible su traducción al idioma del usuario.

Un problema clave de calidad de datos en la Web es la desactualización de la información. La gran mayoría de las páginas no se actualiza con la frecuencia necesaria. Esto reduce la confiabilidad y precisión de los documentos que la utilizan, como informes que preparan alumnos basándose en datos recolectados de sitios de Internet. A menudo se utiliza esta información asumiendo que es correcta sin verificar la fuente ni cuándo ésta se revisó o actualizó por última vez.

Los problemas de calidad de datos se analizan aún en mayor detalle en el capítulo 3 – Recuperación de Información de la Web – ya que se precisa introducir conceptos de R.I. y

R.I. Web antes de abordarlos en profundidad.

2.7.1 Cuando la Mala Calidad Pasa Inadvertida

¿Por qué los usuarios de motores de búsqueda no se dan cuenta de los problemas de calidad de los mismos del mismo modo en que advierten problemas de calidad en otros sistemas de software (e.g., los sistemas que generan facturas para los servicios públicos)?

- Al ser gratuitos, la expectativa de calidad de los usuarios es baja y la importancia que se les concede es menor. Consecuentemente, la frustración que se siente cuando no se obtienen los resultados deseados o cuando los mismos no aparecen en el orden correcto es menor.
- Los buscadores no se utilizan tan frecuentemente como otros servicios. Este uso esporádico obstaculiza la acumulación de malas experiencias y el tenerlas presentes cuando se vive una nueva.
- Los buscadores son mucho más cómodos que tener que viajar hasta una biblioteca, buscar los libros que se precisan, fotocopiarlos y devolverlos a tiempo para evitar multas o suspensiones. A medida que el uso intensivo de los motores de búsqueda se incremente con el tiempo, estos factores de calidad asomarán a la superficie con mayor intensidad.
- La mayoría de las veces, los motores de búsqueda se utilizan para recuperar información coloquial como acceder a diarios en línea, encontrar la página Web de una empresa, comprar artículos, buscar direcciones y teléfonos o fotos de personajes famosos. Estas son "búsquedas simples" y muchas veces realizadas con alta frecuencia, lo que facilita una mejor performance de los motores de búsqueda que en búsquedas más complejas y menos habituales. Al ser un medio de información relativamente nuevo para la mayoría de sus usuarios, la expectativa de que la mayor cantidad de los documentos relevantes para un tema estén indexados y sean retornados para una búsqueda es baja. La mayoría de los usuarios todavía está "contenta" con este nuevo y conveniente medio de información.

Estos factores no indican que la calidad de los motores de búsqueda no tenga que mejorar. Por el contrario, a medida que sus usuarios los utilicen más frecuentemente y comiencen a depender más de los mismos, la ausencia de niveles adecuados de calidad será más notable a simple vista, lo que resultará en un menor nivel de satisfacción del cliente y, consecuentemente, en mayores demandas de calidad.

El apéndice IV introduce Recuperación de la Información, detallando sus conceptos básicos y su arquitectura. El capítulo 3 estudia los motores de búsqueda y el proceso de Recuperación de Información de la WWW.

CAPITULO 3. RECUPERACION DE LA INFORMACION DE LA WEB

"We really are about answers
and not about links"

Yusuf Mehdi, Vicepresidente Senior de MSN

3.1 Introducción

El capítulo anterior introdujo los problemas de calidad de datos y el apéndice IV hizo lo propio con Recuperación de la Información (R.I.), abarcando sus principales modelos y métricas. En este capítulo se aborda R.I. desde la perspectiva de la *World Wide Web* (R.I.W.) y se analizan los principales problemas de calidad de la información que se recupera de la Web.

La WWW creció exponencialmente desde que Tim Berners-Lee y su equipo la crearon a fines de la década del ochenta [BL89]. El corpus de texto de la Web se calcula en el orden de los 167 TB [LiV03]. Es por ello que se puede analizar la Web desde el punto de vista de R.I. al considerarla como una base de datos gigante. Para lograrlo, se precisa adaptar o incluso crear nuevas herramientas para evaluar, buscar y recuperar información de bases de datos de esta magnitud. Consecuentemente, las distinciones entre recuperación de datos y R.I. que fueran mencionadas en el apéndice IV - Recuperación de la Información, serán también la base de este capítulo.

Así como esta tesis se centra en el análisis de documentos de texto en particular, el presente capítulo estudia la R.I. de este tipo de documentos de la Web. También se analiza el texto sintácticamente, buscando términos y patrones seleccionados por el usuario en el corpus de la Web. Se deja como idea para trabajos futuros la búsqueda mediante lenguaje natural y en directorios y meta buscadores.

3.2 La Perspectiva Histórica

Los catálogos de tarjetas, como los utilizados por bibliotecas, fueron los primeros motores de búsqueda. La recuperación estaba limitada a los datos básicos de un libro, como título o autor o, en el mejor de los casos, a citas o resúmenes del mismo. Estos funcionaban como *proxies* o "poderes" del libro en cuestión.

Con el progreso en la industria del hardware y las redes de computadoras, los sistemas de recuperación de información evolucionaron hasta incluir en las respuestas el texto completo de los ítems recuperados, como los sistemas DIALOG y Lexis/Nexis [FOA00]. Sin

embargo, una diferencia clave es que estos no fueron accesibles al público general hasta hace pocos años atrás.

Los *proxies* son versiones resumidas de documentos que permiten al usuario proveer *relevance feedback*[1] rápidamente y acelerar así el proceso de recuperación de documentos relevantes.

3.3 Recuperación de la Información y la Web

Existen tres métodos para recuperación de información de la Web [BR99]:

Motores de búsqueda
Directorios
Explotación de Hipervínculos

El primer método consiste en indexar parte de los documentos de la Web como una base de datos de texto completo. El segundo clasifica los documentos de la Web en categorías y el tercero utiliza la estructura de hipervínculos de la Web para las búsquedas. Esta tesis se centra en el análisis del primer método para R.I.W.: los motores de búsqueda.

3.4 Arquitectura de Motores de Búsqueda

Dado que uno de los objetivos de este trabajo es derivar métricas para las funcionalidades técnicas de los buscadores, es necesario comprender su arquitectura a fin de poder evaluar cómo funcionan desde el punto de vista técnico. Esta sección describe las arquitecturas existentes para un motor de búsqueda de la Web y el funcionamiento de sus componentes.

A diferencia de los motores de búsqueda originales de R.I. que no utilizaban índices[2], las búsquedas en la Web no acceden a los documentos sino sólo a sus índices. Acceder al texto implica un costo mucho mayor que afecta al usuario en el caso de realizar búsquedas remotas (lentitud) o a los dueños de los motores en el caso de poseer copias locales de todos los documentos (costo del hardware y su administración). Esto implica que los lenguajes de búsqueda y los algoritmos de evaluación e indexación deben ser adaptados para funcionar en este ambiente.

3.4.1 Arquitectura Centralizada

La arquitectura centralizada de *crawler*-indexador es la más utilizada por los motores de búsqueda. Esta plantea una división del balance de la carga entre las distintas tareas de un buscador, uno de los aspectos más importantes de una buena arquitectura.

Crawler Agentes de software que atraviesan la WWW con el fin de enviar páginas nuevas o actualizadas al servidor del motor de búsqueda, quien las indexará para facilitar su acceso. El *crawler* corre localmente enviando pedidos a servidores Web remotos.

[1] Ver apéndice IV - Recuperación de la Información para la definición de *relevance feedback*.
[2] Ver la sección 3.6.1 para la definición de índice.

Sinónimos: robot, araña, viajante, caminante, *knowbot.*

Índice Listado centralizado que se utiliza para responder búsquedas de usuarios de la WWW.

El siguiente gráfico resume la arquitectura centralizada de un buscador típico. Generalmente, el motor que busca en el índice consume más del 75% de los recursos del sistema [BR99].

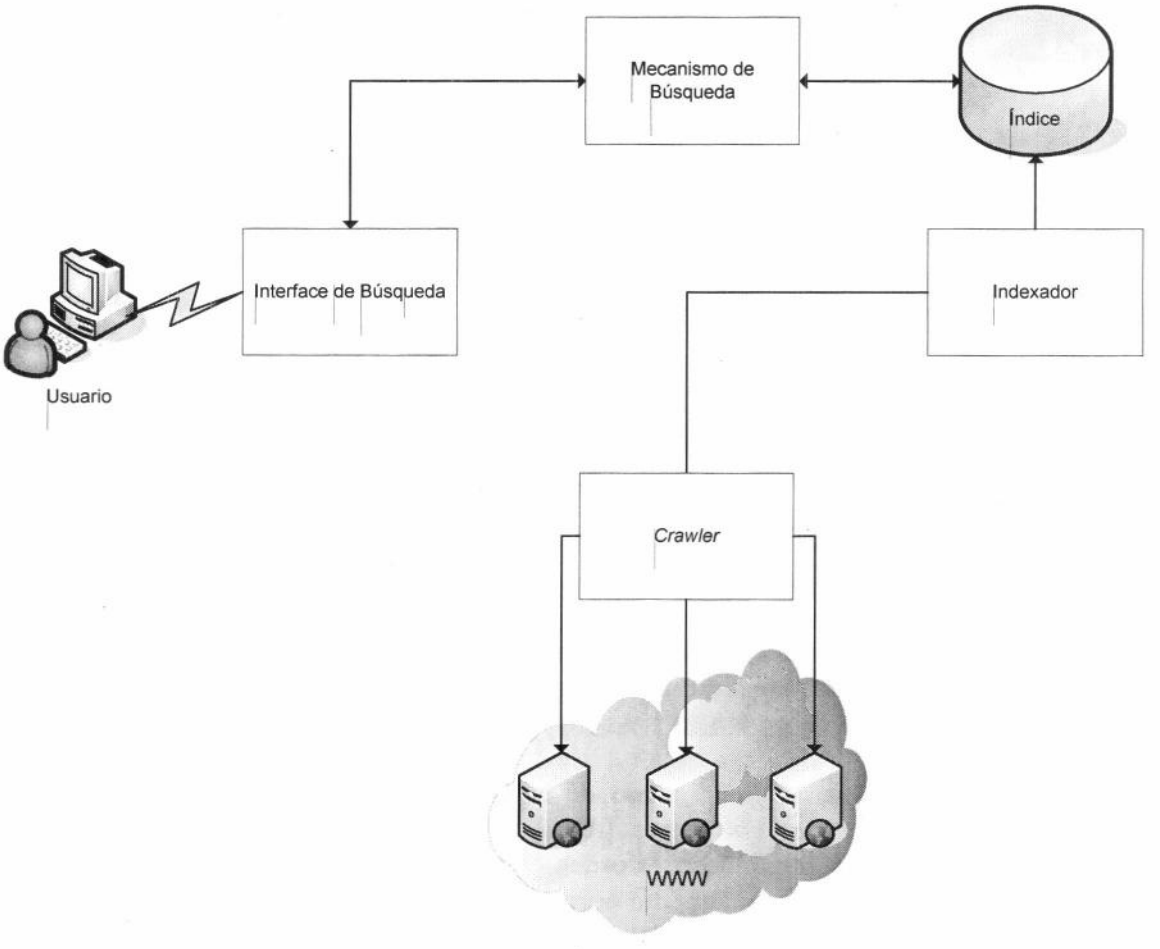


Figura 3.1 – Arquitectura de un Motor de Búsqueda.

Debido a la naturaleza dinámica de la WWW, uno de los problemas clave de esta arquitectura es la recolección de información, ya que ésta se desactualiza con frecuencia y el tráfico de los servidores Web y la WWW en general es muy alto. El volumen de los datos también complica a esta arquitectura, al nivel de plantear dudas sobre su adaptabilidad al rápido crecimiento de la Web.

No todos los motores de búsqueda usan un motor interno creado por ellos mismos. Muchos utilizan los servicios de otros. Por ejemplo, Yahoo le provee los resultados a Altavista. La siguiente figura ilustra las empresas proveedoras de otros buscadores:

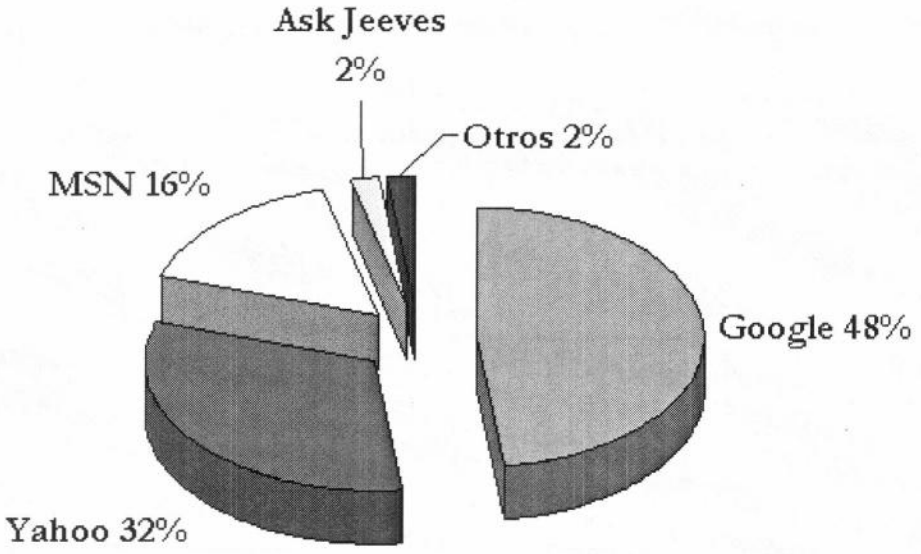


Figura 3.2 – Principales proveedores de buscadores en 2/2005 [SEW05].

Durante el desarrollo de esta investigación, MSN Search cambió de proveedor, abandonando Inktomi luego de muchos años para dedicarse a su propio algoritmo de búsqueda [MSN05].

Si bien la mayoría de los buscadores sigue concentrada en documentos en inglés y se desarrolla en los EE.UU., cada vez se implementan más motores locales para distintos idiomas. Su beneficio es que los resultados de las búsquedas en idiomas diferentes al inglés tienen una mayor relevancia en estos motores que en los otros. Por ejemplo, se suelen obtener mejores resultados al realizar una búsqueda relacionada con el castellano o Argentina en www.ubbi.com.ar que en www.altavista.com.

Algunos motores cambian el mecanismo de la búsqueda, utilizando métodos como las entrevistas a la manera de Ask Jeeves [AJ05]. Otros utilizan algoritmos de relevancia originales, como listar los resultados por orden de popularidad entre los usuarios en vez de considerar la importancia del documento con respecto a la búsqueda realizada [TE05]. Por último, algunos motores se centran en temas específicos, e.g., CiteSeer [CS05] (publicaciones de Ciencias de la Computación) y Feedster (buscador de weblogs) [FS05].

3.4.2 Arquitectura Distribuida

La arquitectura distribuida es una variante del diseño clásico de *crawler*-indexador. Una de las ventajas de esta arquitectura es que es más eficiente para recolectar y distribuir información que la centralizada, lo cual constituye a su vez una de sus desventajas, i.e., requerir la coordinación de varios servidores Web.

Esta arquitectura resuelve tres de los problemas principales de la alternativa centralizada, a saber:

a. La carga de los servidores Web se incrementa debido a los pedidos de varios *crawlers*.
b. El trabajo independiente de los *crawlers* resulta en que la misma información es recolectada por cada uno de ellos sin coordinación conjunta de los motores de búsqueda.
c. La mayor parte de los documentos recuperados por los *crawlers* se descarta luego de la indexación, lo cual resulta en un incremento innecesario del tráfico de la WWW.

Harvest [HA94] es la implementación más importante de la arquitectura distribuida hasta la fecha, ya que resuelve los tres puntos anteriores mediante el uso de componentes nuevos:

- **Recolectores**
- **Agentes**
- **Replicadores**
- ***Cache* de objetos**

La arquitectura distribuida no es utilizada en este trabajo, por lo cual no se analiza en mayor profundidad. [BR99] y [CHO04] la describen en detalle.

3.5 Componentes de un Motor de Búsqueda

Esta sección estudia los componentes principales de un buscador que posee una arquitectura centralizada: GUI, mecanismo de búsqueda y ranking, crawler e indexador.

3.5.1 GUI

Los componentes principales de la GUI de los motores de búsqueda son:

- Interface de búsqueda
  - Estándar
  - Avanzada
- Interface de resultados

3.5.2 Interface de Búsqueda

Uno de los factores que contribuye particularmente a la creación del hábito al utilizar un buscador es que una misma búsqueda no significa lo mismo en dos motores diferentes.
Las búsquedas pueden construirse con una simple secuencia de palabras o haciendo uso del lenguaje de comando avanzado provisto por los motores.

Los usuarios pueden aprovechar el GUI y los lenguajes de comando para mejorar la performance de sus búsquedas mediante los siguientes métodos:

- Especificidad: utilizar tantos términos como sea posible
- Utilizar operadores "booleanos" (+/AND, -/NOT, etc.). E.g., excluir documentos que contienen ciertas palabras.
- Restricciones: ingresar palabras clave que filtren los resultados de las búsquedas o disminuir el alcance mediante la utilización de las funcionalidades de búsqueda

avanzadas. E.g., páginas que pertenecen a dominios que no son de interés o en URLs que contienen una cadena de caracteres predeterminada. El apéndice III incluye las interfaces de búsqueda utilizadas en el experimento de este trabajo.

El diseño de la interface de búsqueda es una parte clave de un buscador ya que los usuarios invierten mucho tiempo buscando datos en la Web. La tabla 3.10 lista la cantidad de tiempo promedio que un usuario tipo dedica a los motores de búsqueda por mes.

Se listan a continuación las funcionalidades principales de una GUI de búsqueda [CHO04]:

Funcionalidad de búsqueda	Notas
Búsqueda "booleana"	Utilizar operadores "booleanos" junto a los términos (*AND, OR, NOT*, +término, -término, "*all the words*", "*none of these words*", "*any of these words*").
Búsqueda por proximidad	Exigir que un término esté cerca de otro en los resultados (*NEAR*).
Búsqueda de frases	Ingresar una frase entre comillas para que sólo los documentos que la contienen sean retornados.
Búsqueda jerárquica	Búsqueda de términos de acuerdo a su ubicación en el documento (título, encabezado, etc.)
Búsqueda restringida	Pedir que sólo documentos que cumplen con ciertas condiciones sean retornados (e.g., páginas que fueron actualizadas en cierto momento, se almacenan en determinado dominio o tipo de dominio, etc.)

Tabla 3.1 – Funcionalidades de interfaces de búsqueda

Lenguajes de Búsqueda por Enlaces

Los paradigmas más modernos de R.I.W. se basan en el uso de enlaces para la realización de búsquedas. Estos modelos no son muy conocidos ya que no se utilizan en motores de búsqueda comerciales y aún tienen problemas de performance.

El concepto principal se basa en visualizar la Web como un grafo etiquetado [Abi97] y analizar no sólo el contenido de los nodos (i.e., páginas Web) que está representado mediante un esquema de datos sino también la estructura de sus ejes o enlaces durante la búsqueda. Mediante este modelo, un usuario puede buscar todas las publicaciones sobre R.I.W. que son accesibles desde http://citeseer.ist.psu.edu atravesando a lo sumo dos enlaces. Los lenguajes de búsqueda tradicionales no poseen funcionalidades que automaticen este tipo de búsquedas.

Lenguajes de búsqueda como WebSQL o W3QL permiten localizar documentos que contengan ciertas características estructurales junto con enlaces entrantes que responden a

una estructura de grafos determinada. Esto se logra utilizando expresiones regulares de caminos de grafos dentro de las búsquedas [KS95] [MMM97]. Si bien estos lenguajes plantean mejoras muy interesantes para R.I.W. y algunos están acompañados de GUIs, éstos no fueron creados con el usuario final en mente sino para la automatización de búsquedas mediante algoritmos.

3.5.3 Interface de Resultados

El apéndice III incluye la GUI de resultados de los cinco buscadores utilizados en el experimento para una misma búsqueda.

Como se observa en ese apéndice, una página de resultados típica dispone de las siguientes funcionalidades para los documentos retornados:

- Ranking
- Enlace
- Resumen "inteligente" o *snippet* (i.e., idealmente contiene las palabras de la búsqueda resaltadas para que el usuario pueda evaluar su relevancia)
- Tamaño de la página
- *Cache*
- Traducción
- Propagandas de documentos relevantes (*"sponsored links"*)
- Fecha de indexación
- Última fecha de actualización
- Cantidad total de documentos retornados
- Tiempo requerido para buscar los documentos
- Ofrecimiento de búsqueda de documentos similares

Algunos motores son más flexibles que otros, otorgando al usuario la posibilidad de personalizar la información mostrada en esa pantalla, así como el orden de los documentos (por fecha, relevancia, dominio, etc.) o la cantidad de documentos retornados por página.

La interface de resultados también contiene la de búsqueda, permitiendo así refinar la misma si los resultados no son satisfactorios.

La sección siguiente describe una de las funcionalidades expuestas a través de la interface de resultados: el ranking de documentos.

3.5.4 El Ranking

Esta sección describe los algoritmos de ranking utilizados por los motores de búsqueda y no se centra en los algoritmos clásicos compartidos con la recuperación de información de bases de datos tradicionales, ya que éstos fueron descriptos en el apéndice IV.

Los modelos de lógica de Boole y vectoriales descriptos en el apéndice IV así como variaciones de los mismos, son los más utilizados para la funcionalidad de ranking. A fin de minimizar el tiempo de retorno de resultados, esta funcionalidad debe ser utilizada accediendo sólo al índice de los documentos.

Los algoritmos de ranking y las mejoras implementadas a los mismos por los buscadores populares no son muy conocidos ya que constituyen el secreto de su éxito. Estas razones dificultan su comparación y el análisis de métricas como la recuperación a través de distintos motores.

Además de los algoritmos de ranking clásicos de tf-idf descriptos en el apéndice IV, Yuwono y Lee [YL96] proponen los siguientes:

- Extensión "Booleana"
- Extensión Vectorial
- Más Citado

Los algoritmos de extensión "booleana" y vectorial son, como indica su nombre, ampliaciones de los originales "booleano" y vectorial de modo que el conjunto de respuestas también incluye las páginas apuntadas por los documentos retornados en los algoritmos clásicos y las que contienen enlaces a los mismos. El último algoritmo se basa en los términos de las páginas que apuntan a los sitios Web de los resultados. Los experimentos de Yuwono y Lee determinaron que Extensión Vectorial es mejor en cuanto a precisión (75% en promedio) y recuperación.

Una diferencia importante entre los algoritmos de ranking utilizados por bases de datos tradicionales y los de R.I. Web (R.I.W.) es que los más importantes entre los segundos utilizan la información de los enlaces para determinar el orden de los elementos del conjunto de respuestas, e.g., cantidad de enlaces que apuntan a una página. También se mide la relación entre páginas observando si estas son apuntadas por las mismas páginas o si tienen muchos enlaces en común.

A continuación, se describen las técnicas más utilizadas para los algoritmos de ranking en R.I.W.

- WebQuery
- HITS
- PageRank
- Tf/idf[3]

WebQuery

Este algoritmo toma los resultados de una búsqueda como entrada y retorna un nuevo ranking en base a la conectividad de cada elemento, i.e., cuán conectado está cada documento en la Web. Luego agrega las páginas que están muy conectadas a los elementos del nuevo ranking [BR99].

[3] Tf/idf se describe en el apéndice IV – Recuperación de la Información.

HITS

HITS, *Hypertext Induced Topic Search*, utiliza el algoritmo de Kleinberg que se basa en los conceptos de *hubs* y autoridades.

Sea S un conjunto de páginas que apunta o es apuntado por páginas que pertenecen a la respuesta de una búsqueda.

Definición 3.1: *Hubs* Páginas que tienen muchos enlaces salientes hacia S.

$$H(p) = \sum_{u \in S | p \to u} A(u)$$

Definición 3.2: Autoridades Páginas que tienen muchos enlaces entrantes desde S.

$$A(p) = \sum_{v \in S | v \to p} H(v)$$

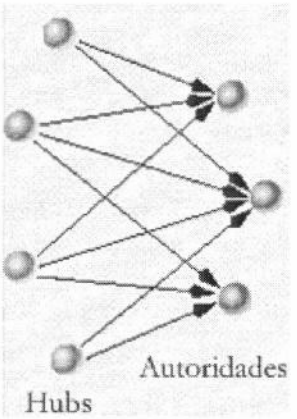


Figura 3.3 – Algoritmo de Kleinberg: hubs y autoridades [EP05].

Este sistema crea un ciclo de retroalimentación positiva, ya que se generan mejores autoridades a través de enlaces entrantes que provienen de buenos *hubs* y también mejores *hubs* a través de enlaces salientes a buenas autoridades.

Uno de los problemas principales de HITS es que no funciona con la "Web oculta" (i.e., páginas generadas dinámicamente) ni con enlaces repetidos al intentar controlar el tamaño de S. Entre las modificaciones a HITS se encuentra el análisis del contenido de cada página y la consecuente asignación de un puntaje para la misma. Este puntaje puede utilizarse como multiplicador en las sumatorias a fin de evitar que los resultados se expandan a temas muy generales que a veces incluyen la respuesta original.

Experimentos con HITS demostraron que tanto recuperación como precisión mejoran significativamente para los primeros diez resultados [BH98].

PageRank

PageRank modela la situación de un usuario navegando por la Web que salta a otra página aleatoria con probabilidad q o que hace clic en un enlace del mismo sitio con probabilidad 1-q. Se toma como axioma que el usuario nunca retorna a una página navegando por los sitios ya visitados hacia atrás. Larry Page y Sergey Brin [BP98] descubrieron que este proceso se puede simular mediante una cadena de Markov, a partir de la cual se puede calcular la probabilidad estacionaria de estar en una página determinada.

Cadena de Markov

Sea una colección de variables aleatorias $\{X_t\}/t \geq 0$ que tiene la propiedad de que, dado el presente, el futuro es condicionalmente independiente del pasado. Esta propiedad se representa mediante la siguiente fórmula:

$$P(Xt = j \mid X_0 = i_0, X_1 = i_1, \ldots, X_{t-1} = i_{t-1}) = P(Xt = j \mid X_{t-1} = i_{t-1})$$

Si una secuencia de Markov de variables aleatorias X_n toma los valores discretos $a_1, \ldots, a_n$, entonces se puede afirmar que

$$P(x_n = a_{in} \mid x_{n-1} = a_{in-1}, \ldots, x_1 = a_{i1}) = P(x_n = a_{in} \mid x_{n-1} = a_{in-1})$$

y la secuencia x_n es una cadena de Markov.

El valor obtenido para la probabilidad de cada página se utiliza para el algoritmo de ranking de Google [GG05].

Sea C(a) el número de enlaces salientes de la página a. Sean $p_1, \ldots, p_n$ las páginas que tienen enlaces hacia a.

$$p_1, \ldots, p_n \rightarrow a \rightarrow O / |O| = C(a)$$

Luego el PageRank de a o PR(a) se define como:

$$PR(a) = (1-q) + q\sum_{i=1}^{n} \frac{PR(p_i)}{C(p_i)}$$

Donde q es típicamente 0,15 y es estipulado por el sistema y P es la cantidad total de páginas. Como se puede observar, el peso o ranking de las otras páginas es normalizado por la cantidad de enlaces de la misma.

PageRank se corresponde con el autovector principal de la matriz normalizada de enlaces de la Web y se calcula utilizando un algoritmo iterativo. La matriz en cuestión es también la matriz de transición de la cadena de Markov.

De acuerdo a Cho, García Molina y Page, es mejor realizar el *crawling* de la Web con este algoritmo de ordenamiento que utilizando otros mecanismos [CGP98].

A medida que Google se popularizaba, los administradores de sitios Web aprendieron cómo lograr que sus páginas ocuparan posiciones más elevadas en el ranking de este buscador Además de utilizar títulos, palabras clave y *meta tags* significativos, éstos se concentraron en intentar que páginas populares agregaran enlaces a sus sitios, lo cual resulta en un mejor posicionamiento en el ranking a largo plazo. Estas prácticas no son deseables ya que alteran la percepción del usuario final en cuanto a la relevancia real de los resultados.

3.6 Técnicas de *Crawling*

La técnica de *crawling* más simple consiste en comenzar con un enlace o URL e ir descubriendo otros recursivamente mediante BFS (*breadth-first search*) o DFS (*depth-first search*). Algunos motores recorren extensivamente los sitios siguiendo estos algoritmos, mientras que otros sólo toman una muestra de los mismos. El conjunto inicial puede ser aleatorio, estar compuesto de URLs populares o de páginas enviadas recientemente por sus usuarios o administradores.

El beneficio de utilizar BFS para el *crawling* es permite acceder fácilmente a más información sobre un tema en particular para los sitios que están estructurados en base a temas relacionados. Por otra parte, el recorrido de los sitios es superficial e incrementa significativamente el tráfico para cada sitio. DFS soluciona la superficialidad de BFS pero es demasiado "angosto" ya que sólo accede al enésimo enlace de cada página durante cada iteración. Es decir, tarda demasiado tiempo en indexar un sitio completo.

La razón por la que se hace tanto hincapié en disminuir el tráfico de la Web es que éste se incrementa en gran escala a medida que transcurre el tiempo. La tabla 3.8 muestra la cantidad de búsquedas realizadas por día en los motores de búsqueda principales. Los datos fueron obtenidos entre enero y febrero de 2003 [SEW05].

Los sitios que no pertenecen a la muestra inicial o que no son registrados en los buscadores podrán igualmente ser alcanzados por el *crawler*. Sin embargo, el tiempo de espera aumenta considerablemente, pudiendo llegar hasta un par de meses. Algunos motores amortiguan este problema identificando el tiempo de actualización de cada sitio y de este modo "aprenden" cada cuánto deben visitarlos. Sin embargo, este proceso insume mucho tiempo hasta que se torna efectivo [CLW97]. Otra forma de mitigar este problema es identificar los sitios más visitados y realizar el *crawling* de los mismos más frecuentemente.

La sección 3.7 describe los problemas principales del *crawling* en R.I.W.

3.7 Indexación

El apéndice IV describe los índices más comunes en el ámbito de R.I. En esta sección, se detalla cómo éstos son aplicados a R.I.W. y qué otras alternativas existen.

La mayoría de los motores de búsqueda utiliza variantes del índice invertido. Los documentos también pasan por el proceso de normalización que elimina el ruido, como el que proveen espacios extra, puntuación, signos gramaticales y *stopwords*. Estos pasos ayudan a controlar el tamaño del índice. Luego se agregan los *meta tags*, los que acompañan al índice proveyendo la información básica sobre las páginas que será de utilidad para el usuario (i.e., título, fecha, secciones, autor, tamaño, resumen, etc.) Cuando un usuario realiza una consulta, el buscador retorna estos datos básicos junto con el URL de cada documento pero mantiene en memoria la información completa para la búsqueda a fin de no tener que recalcularla si el usuario decide acceder a algunas de las páginas retornadas.

Mediante la utilización de las técnicas de indexación de R.I.W., se puede obtener un índice equivalente al 30% o menos del tamaño de las páginas originales [WMB94] (e.g., alrededor de 11 TB para la cantidad de páginas indexadas por Google) y al 10% si además se lo comprime.

El algoritmo de búsqueda en el índice invertido ordenado es casi exclusivamente la búsqueda binaria. Dado que la mayoría de las búsquedas involucra más de un término, se deben realizar múltiples búsquedas en el índice y combinar sus resultados a fin de obtener la respuesta final. Lo mismo ocurre para búsquedas de prefijos. Una alternativa es calcular sólo los primeros *n* resultados, donde *n* es el número de documentos que un usuario revisa en promedio y calcular los próximos documentos sólo a medida que se solicitan. Las búsquedas de expresiones regulares o con comodines son secuenciales. Sin embargo, como el tamaño del vocabulario es mucho menor que el del texto plano (e.g., 10 MB para 2 GB de texto), este tipo de búsquedas sólo requiere unos pocos segundos.

Otra opción disponible es el índice invertido completo, en el cual los términos apuntan no sólo al documento sino a su(s) ocurrencia(s) dentro de la página. La ventaja de este tipo de índice es que disminuye la complejidad de búsquedas de contexto, como las de frases y proximidad. La desventaja: el espacio extra requerido. La alternativa intermedia entre los índices invertidos y los invertidos completos es hacer que las palabras apunten a páginas o a bloques de texto del mismo tamaño. Esto facilita que el índice sea más chico que el completo mientras que las búsquedas por proximidad y de frases no se tornan mucho más complejas. Los índices por bloques son más utilizados en las arquitecturas distribuidas, ya que estas proveen el marco más adecuado para dividir el índice en partes.

Hasta aquí se han estudiado los componentes principales de los buscadores. La siguiente sección describe las métricas principales que se utilizan para R.I.W.

3.8 Métricas para R.I.W.

Una de las diferencias clave entre las bibliotecas digitales y la Web es que esta última es mucho más dinámica, lo que dificulta en gran medida hallar métricas adecuadas para evaluarla.

Dado que este trabajo de investigación deriva nuevas métricas para comprender mejor los atributos principales de los motores de búsqueda, se detallan en esta sección las métricas

principales utilizadas en la actualidad para medir la Web: métricas sobre el corpus, i.e., independientes de un buscador y métricas que se aplican a un motor de búsqueda en particular. Asimismo, se consideran métricas adicionales para R.I.W., más allá de las mencionadas en el apéndice IV. Vale aclarar que las métricas de R.I. se aplican también a R.I.W.

3.8.1 Métricas sobre el Corpus de la Web

a. Número de sitios Web

Como se observa en el gráfico siguiente [NET05], el tamaño de la Web medido en cantidad de sitios (o *hostnames*) se triplicó entre enero de 2000 y enero de 2001.

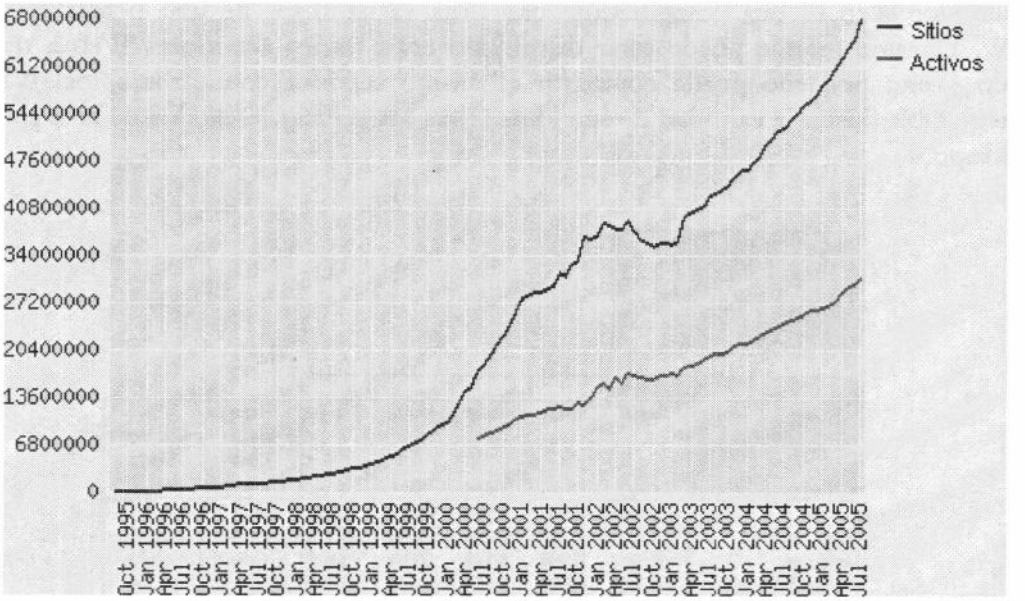


Figura 3.4 – Número de sitios Web a través de todos los dominios (1995 - 2005).

La encuesta de Netcraft de julio de 2005 refleja un total de 67.571.581 sitios Web, lo que significa un incremento de 2.760.000 sitios con respecto al mes anterior. Se cree que la causa de la explosión de esta métrica se debe principalmente al creciente uso de Internet por parte de PYMEs y a la creación de nuevos weblogs.

b. Número de dominios en la WWW

Un dominio Web ubica una organización o entidad en Internet utilizando un URL, dirección IP y tipo de servidor *host* (www, etc.). La URL incluye el dominio de alto nivel, el cual es una referencia al tipo de organización (.com, .gov, etc.). El dominio de segundo nivel define a la organización o entidad (e.g.: uba) junto con el de alto nivel [WI05].

De acuerdo a WHOIS [WHO05], durante el desarrollo de esta tesis existían 47.501.249 dominios registrados en la Web. Este número creció más de un 47% desde el 2003, año en el cual había "sólo" 31.987.198 dominios registrados.

c. Número de servidores Web

Un servidor Web es una computadora que aloja los archivos de uno o más sitios de la WWW. Un sitio grande puede estar distribuido entre varios servidores Web a través del mundo y uno pequeño puede compartir el mismo servidor con otros sitios [WI05]. La siguiente tabla ilustra la cantidad de servidores para los principales nodos de la Web a través del tiempo.

Country	(All)
Hosting Model	(All)
Active Per Comp.	(All)
Major OS	(All)
OS	(All)

Computers	Date						
DNS Owner	Jun-01	Sep-01	Dec-01	Mar-02	Sep-02	% Change Since Last Quarter	% Change Since Jun 01
uu.net	5,792	5,512	5,551	6,729	7,029	+4%	+21%
krnic.net	4,751	4,275	5,064	6,109	6,212	+2%	+31%
apnic.net	4,378	3,998	4,372	5,854	5,926	+1%	+35%
verio.net	6,238	5,693	5,512	5,740	5,418	-6%	-13%
rackshack.net		668	1,208	2,121	4,926	+132%	0%
schlund.de	1,056	693	812	1,448	4,458	+208%	+322%
rackspace.com	2,390	2,471	2,807	3,260	4,095	+26%	+71%
kornet.net	1,421	1,644	2,178	2,711	2,864	+6%	+102%
dialtoneinternet.net	1,816	1,967	2,243	2,493	2,862	+15%	+58%
sprintlink.net	1,881	1,729	2,114	2,531	2,837	+12%	+51%
interbusiness.it	1,735	1,591	1,931	2,438	2,624	+8%	+51%
dtag.de	1,390	1,888	1,659	2,296	2,567	+12%	+85%
qwest.net	962	1,025	1,428	1,836	2,464	+34%	+156%
att.net	2,416	2,353	2,505	2,605	2,179	-16%	-10%
interland.net	1,822	1,355	1,541	1,647	2,029	+23%	+11%

Tabla 3.2 – Número de servidores Web entre 6/2001 y 9/2002 [NET05].

d. Cantidad de *hosts* de la Web

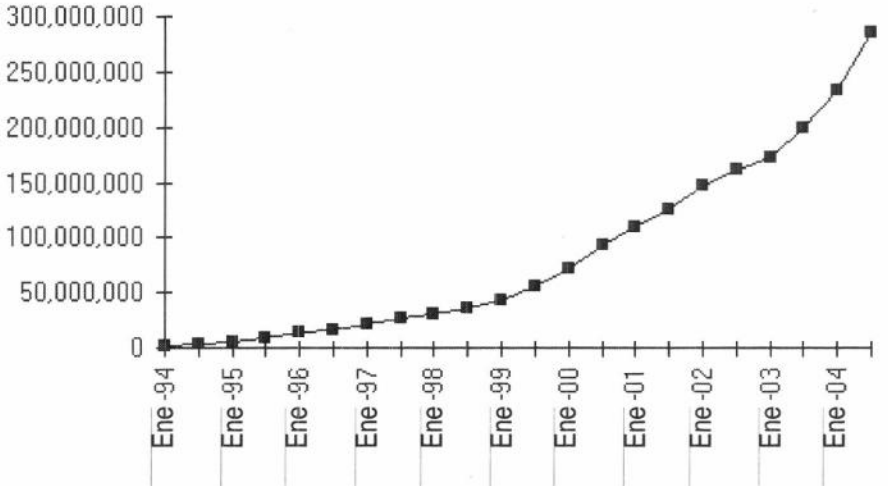


Figura 3.5 – Número de *hosts* en la WWW entre 1/1994 y 1/2004 [IDS05].

e. Formatos de documento más populares

Al utilizar como métrica de R.I.W. los formatos más utilizados para páginas Web, se puede comprender uno de los problemas principales de R.I.W.: la falta de un proceso de editorialización en la creación de documentos Web.

Los formatos más utilizados son los siguientes [HMI03]:

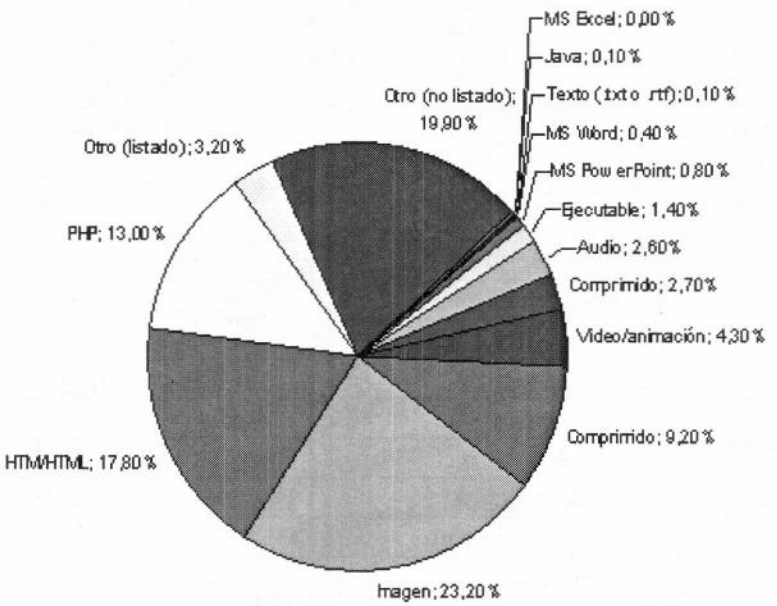


Figura 3.6 – Composición de la WWW por tipo de archivo.

Como se observa más arriba, HTML es el formato más popular entre los documentos de texto. Sin embargo, los estándares de la especificación de HTML no son siempre observados en la creación de documentos para la Web (e.g., en la mayoría de los casos, los documentos no incluyen una definición de tipo al comienzo de los mismos). Este tipo de documentos es generalmente pequeño (18.7 KB es la media, la mayoría contiene sólo una o dos imágenes chicas) e incluye ocho enlaces, mayormente locales, en promedio. Consecuentemente, los enlaces que apuntan a una página x provienen mayormente de servidores locales, incluso al considerar la página principal de un sitio.

f. Cantidad promedio de enlaces salientes por sitio Web

Esta métrica es muy interesante ya que podría influenciar el algoritmo de ranking de los motores de búsqueda Lamentablemente, actualmente no se dispone de estadísticas al respecto.

El crawler de un buscador podría medir la cantidad promedio de enlaces salientes en la Web si llevara la cuenta de la cantidad de enlaces de cada sitio a sitios externos. Una vez que se dispone de esta información, obtener el promedio de las cantidades recolectadas durante la fase de *crawling* permitiría obtener el valor de esta métrica.

g. Sitios Web más referenciados

Los **sitios más referenciados externamente**[4] son los siguientes:

	2000		2002
1	www.microsoft.com	1	www.adobe.com
2	www.netscape.com	2	www.microsoft.com
3	www.geocities.com	3	www.geocities.com
4	members.aol.com	4	www.netscape.com
5	www.yahoo.com	5	members.aol.com
6	www.adobe.com	6	www.yahoo.com
7	www.amazon.com	7	www.amazon.com
8	www.altavista.com	8	www.google.com
9	members.tripod.com	9	www.macromedia.com
10	www.macromedia.com	10	www.cnn.com
11	www.angelfire.com	11	www.altavista.com
12	www.cnn.com	12	www.angelfire.com
13	www.excite.com	13	members.tripod.com
14	www.real.com	14	www.mapquest.com
15	www.lycos.com	15	www.weather.com
16	members.xoom.com	16	www.apple.com
17	www.linkexchange.com	17	www.excite.com
18	Home.earthlink.net	18	www.lycos.com

[4] Es decir, desde un sitio Web hacia otro.

19	ourworld.compuserve.com	19	www.real.com
20	www.apple.com	20	www.hotbot.com

Tabla 3.3 – Sitios con más enlaces entrantes de la Web [OCL05].

Esta métrica ayuda a comprender cómo está interconectada la Web. Como se observó, ciertos algoritmos utilizan los enlaces entre sitios para calcular los resultados de las búsquedas. De este modo, los sitios más "enlazados" y los que contienen más enlaces a otras páginas se tornan en el centro del grafo de la Web, en el eje de interconexión del universo de discurso.

La siguiente figura muestra el grafo de los enlaces que conectan al Departamento de Computación de la FCEN con el resto de la *World Wide Web*.

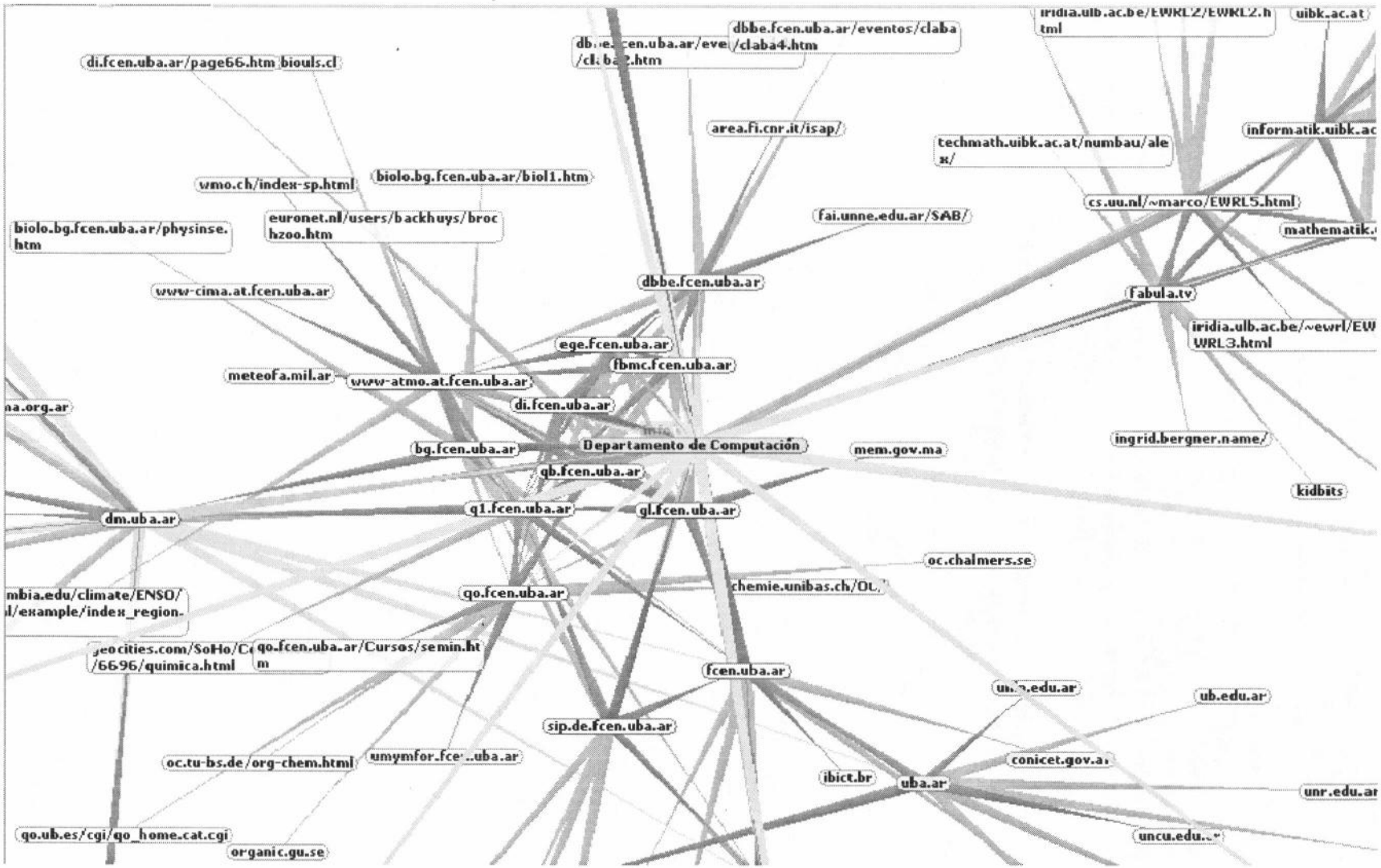


Figura 3.7 – Red de enlaces para el sitio Web del Departamento de Computación de la U.B.A. [TOU05].

h. Tamaño de la WWW

Al considerar, como se mencionó anteriormente, que el tamaño promedio de una página HTML es de 18.7 KB y que existen más de 8.000.000.000 de páginas Web, se observa que la Web almacena al menos 144 TB de datos.

Una investigación realizada en la Universidad de Berkeley en 2003, indica que el tamaño de la Web "superficial" (i.e., accesible por buscadores) era de 167 TB en ese momento mientras que la Web "profunda" (i.e., los sitios Web con un *back-end* de bases de datos que crean páginas por demanda) alcanzaba los 91.850 TB [LiV03].

i. Frecuencia relativa de idiomas

Si se analizan los idiomas utilizados en la Web, se encuentra que hay más de cien en uso. El inglés sigue siendo el más utilizado por un gran margen pero los puestos siguientes han variado bastante en los últimos diez años. La siguiente figura muestra los resultados de O'Neil, Lavoie y Bennett para esta métrica tal como fueron publicados por D-Lib Magazine en Abril de 2003.

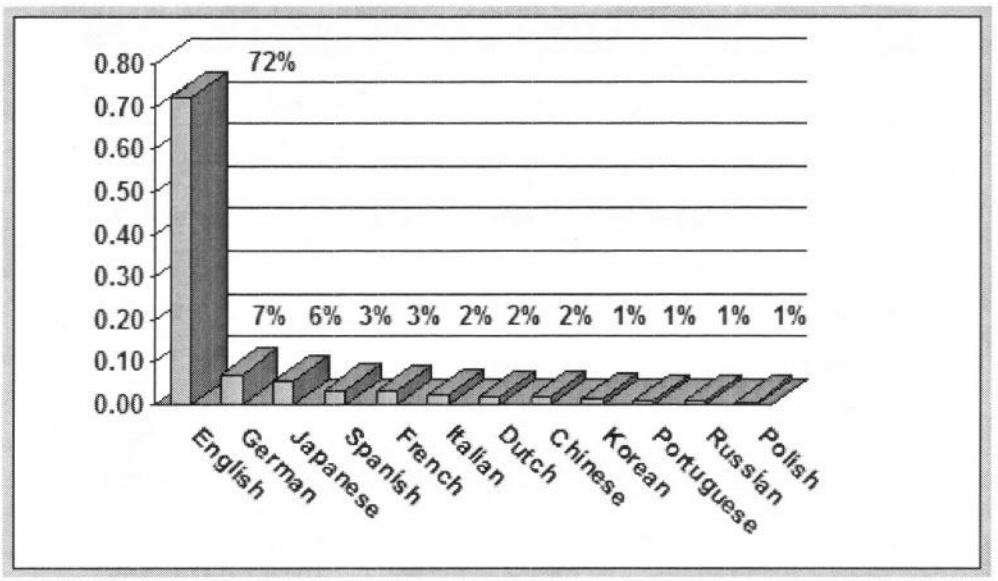


Figura 3.8 - Distribución del uso de idiomas en las páginas Web durante 2002 [OLB03].

j. Tamaño de los documentos

En el siguiente gráfico se puede observar que si bien la mayoría de los documentos de la Web son pequeños, hay una cantidad significativa de textos grandes. Es por ello que la distribución tiene una larga cola hacia la derecha [CV96].

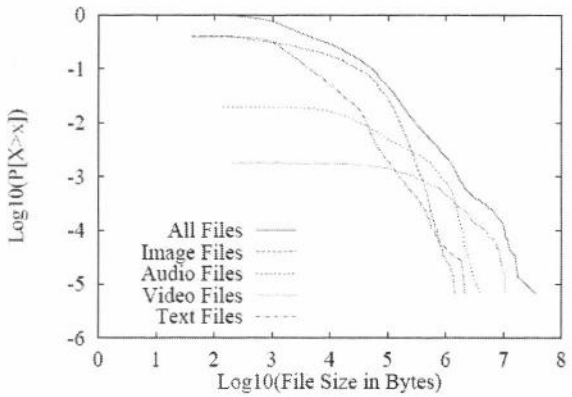


Figura 3.9 – Distribución de cola derecha para los tipos de archivo principales de la Web.

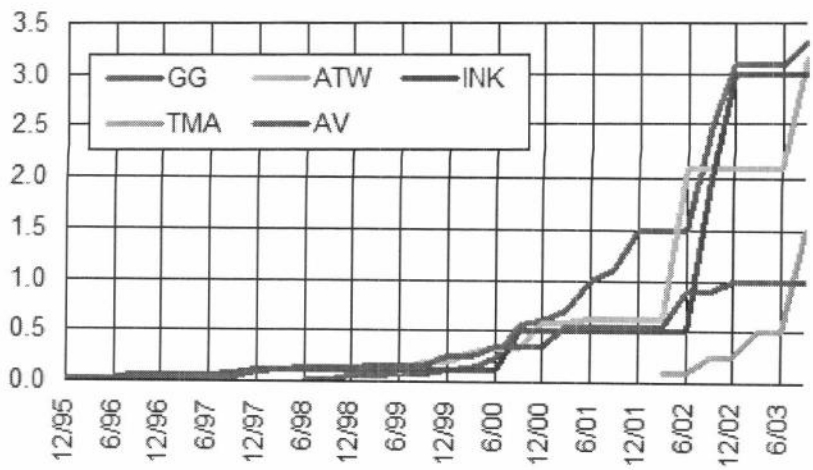
La siguiente tabla muestra la distribución de los distintos tipos de archivo de la WWW de acuerdo a su tamaño.

Tamaño del documento	Tipo de archivo predominante
x<50 KB	Texto, imagen
50 KB ≤x≤300 KB	Audio
x≥300 KB	Video

Tabla 3.4 – Tipos de archivos de la Web clasificados por tamaño

3.8.2 Métricas sobre los Motores de Búsqueda

a. Cantidad de páginas Web indexadas



GG: Google; ATW: AllTheWeb; INK: Inktomi; TMA: Teoma; AV: Altavista

Figura 3.10 – Miles de millones de páginas indexadas por los principales buscadores entre 12/1995 y 9/2003 [SEW05].

La WWW crece a razón de más de veinte millones de páginas por mes [BR99]. El número de páginas es un claro ejemplo de una métrica afectada por el dinamismo de la Web. El valor de esta métrica siempre será estimado. Esta estimación se calcula midiendo la cantidad de páginas indexadas por cada motor de búsqueda y ponderando ese número por una estimación del porcentaje de solapamiento entre los buscadores [BB98].

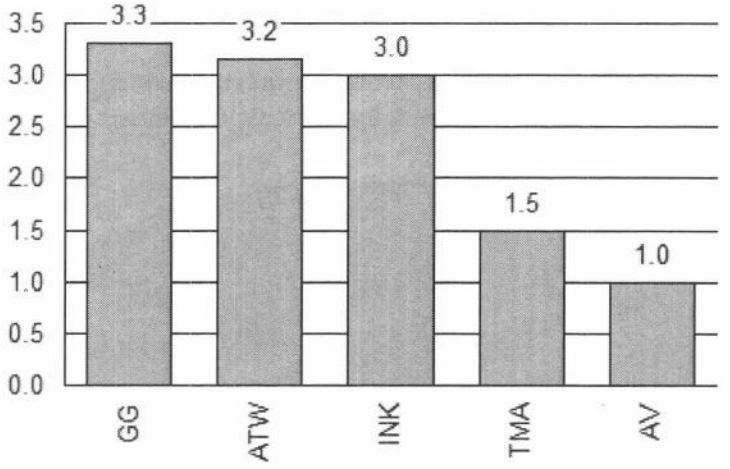


Figura 3.11 – Miles de millones de páginas indexadas por los principales motores en 9/2003 [SEW05].

La siguiente tabla muestra el tamaño del índice para los principales buscadores y directorios tal como fue reportado en noviembre de 2004.

M.B./Directorio	Tamaño declarado
Google	8.100.000.000
MSN	5.000.000.000
Yahoo	4.200.000.000 (estimado)
Ask Jeeves	2.500.000.000

Tabla 3.5 – Cantidad de páginas indexadas por los principales buscadores en 11/2004 [SEW05].

b. Profundidad de *crawling*

Esta métrica indica el nivel promedio de *crawling* de una página que realiza el motor, es decir, cuántos KB "lee" un motor por página. Como se observa en la tabla 3.2, la profundidad de *crawling* es similar para los principales motores de búsqueda y cinco veces mayor para el principal directorio de la Web (Yahoo).

M.B./Directorio	Profundidad de página
Google	101 KB
MSN	150 KB
Yahoo	500 KB
Ask Jeeves	>101 KB

Tabla 3.6 – Profundidad de *crawling* de los buscadores [SEW05].

c. Distribución de búsquedas

El gráfico de torta de la figura 3.12 representa el porcentaje de búsquedas realizadas en los Estados Unidos durante mayo de 2004 para los principales motores de búsqueda.

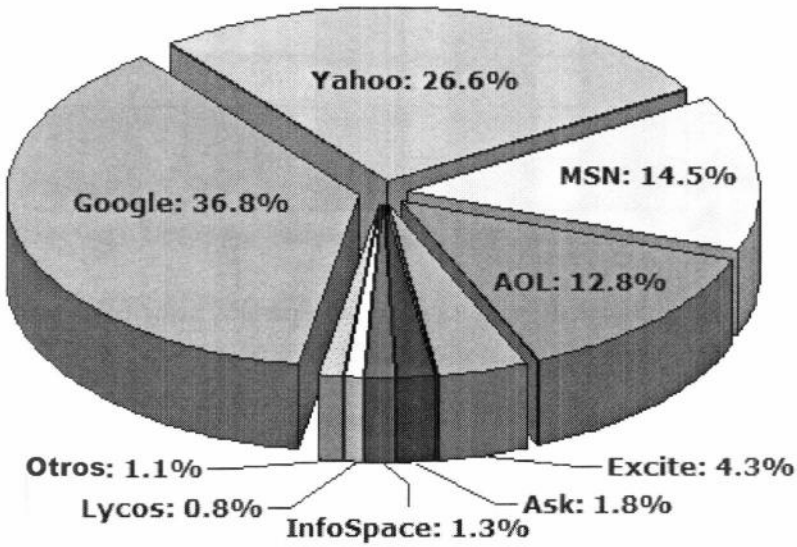
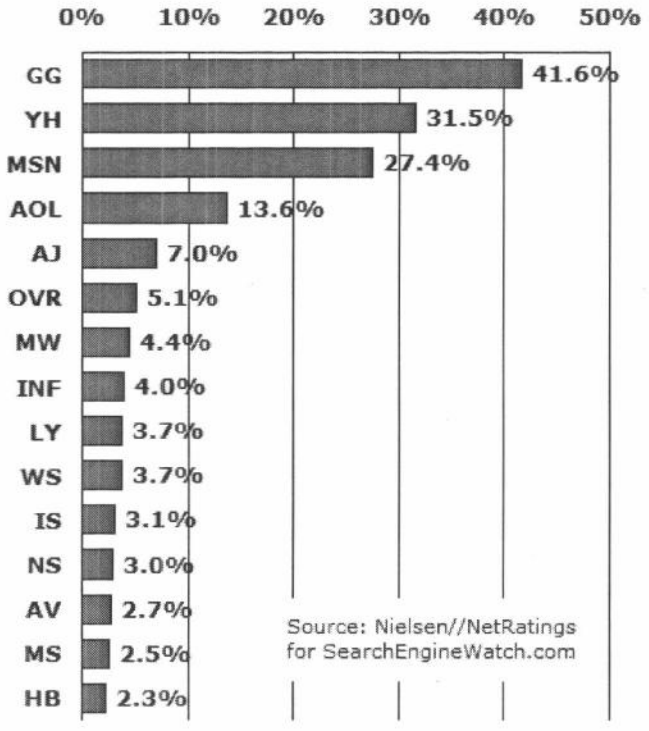


Figura 3.12 – Distribución de búsquedas realizadas en EE.UU. durante 2004 [SEW05].

El siguiente gráfico ilustra la popularidad de los principales motores de búsqueda durante junio de 2004:



Abreviaturas: GG=Google, YH=Yahoo, MSN=MSN Search, AOL=AOL, AJ=Ask Jeeves. OVR=Overture, MY=MyWay.com, INF=Information.com, LY=Lycos Networks, WS=WebSearch.com, IS=InfoSpace Networks, NS=Netscape Search, AV=AltaVista, MS=Microsoft.com, HB=HighBeam.com.

Figura 3.13 – Buscadores más utilizados (6/2004) [SEW05].

Finalmente, se presentan las estadísticas globales de utilización de buscadores entre 1/2003 y 5/2004:

Buscador	5/04	11/03	5/03	1/03
Google	56.4%	56.1%	55.2%	54.7%
Yahoo!	21.1%	21.5%	21.7%	22.1%
MSN Search	9.2%	9.4%	9.6%	9.5%
AOL Search	3.8%	3.7%	3.8%	3.7%
Terra Lycos	2.0%	2.3%	2.6%	2.8%
Altavista	1.7%	1.9%	2.2%	2.5%
Ask Jeeves	1.7%	1.6%	1.5%	1.5%

Tabla 3.7 – Distribución del uso de los buscadores entre 1/2003 y 5/2004 [CLI05].

d. Cantidad de búsquedas por día

La tabla 3.8 muestra la cantidad de búsquedas realizadas por día en los motores principales. Los datos fueron obtenidos entre enero y febrero de 2003 [SEW05].

M.B.	Búsquedas/día	Notas
Google	250.000.000	Búsquedas realizadas en Google y en sus sitios asociados
Overture	167.000.000	Búsquedas realizadas en Overture y en sus sitios asociados
Inktomi	80.000.000	
LookSmart	45.000.000	Incluye búsquedas realizadas mediante todos los sitios asociados a LookSmart, incluyendo MSN Search
FindWhat	33.000.000	Búsquedas realizadas en FindWhat y en sus sitios asociados
Ask Jeeves	20.000.000	Búsquedas realizadas en Ask Jeeves, Ask UK, Teoma y en sus sitios asociados
AltaVista	18.000.000	
FAST	12.000.000	

Tabla 3.8 – Cantidad de búsquedas realizadas por día en los principales buscadores.

e. Búsquedas más frecuentes

Las siguientes tablas listan las búsquedas más populares realizadas en Google durante 2004. Como se puede observar, artistas y personajes famosos ocupan los primeros puestos entre las búsquedas generales. Con respecto a las tecnológicas, los fondos de pantalla y un software libre para bajar música de la Web son los más buscados.

Búsquedas Generales	#	Búsquedas de Tecnología
Britney Spears	1	Wallpaper
Paris Hilton	2	Kazaa
Christina Aguilera	3	MP3
Pamela Anderson	4	Spybot
Chat	5	Linux
Games	6	
Carmen Electra	7	
Orlando Bloom	8	
Harry Potter	9	
MP3	10	

Tablas 3.9 – Búsquedas más populares realizadas en Google durante 2004.

f. Tiempo de búsqueda por mes

La tabla 3.10 ilustra el tiempo que invierte un usuario promedio por mes en buscar información en los motores de búsqueda.

Motor de búsqueda	Minutos/mes
Google	29:57
AOL Search	28:28
Netscape	13:09
InfoSpace	11:41
Yahoo	11:04
Web Search	8:06
MSN Search	7:39
Ask Jeeves	6:29
AltaVista	6:27
My Way	5:11
Overture	3:25
Lycos Network	2:53
MS Search	2:22
HighBeam	1:36
Information	0:50

Tabla 3.10 – Cantidad de tiempo promedio que un usuario invierte realizando búsquedas por mes.

Si bien la mayoría de los buscadores lista el tiempo insumido por una búsqueda ,determinada en su GUI de resultados, no se encontraron métricas formales que calculen el promedio de los tiempos de búsqueda por motor.

La mayoría de las métricas descriptas en esta sección están principalmente orientadas a la parte técnica o arquitectural, dejando generalmente de lado el punto de vista del usuario. Por este motivo, en los capítulos siguientes se derivan nuevas métricas de R.I.W. teniendo en cuenta las dos perspectivas pero analizando con más detalle el punto de vista del usuario, proponiendo así una visión diferente a la actual.

3.9 Problemas de R.I.W.

La siguiente lista describe los problemas principales que enfrenta R.I.W. en la actualidad.

1. El lenguaje de búsqueda a utilizar (i.e., lógica de Boole) no se corresponde con el lenguaje natural de los usuarios, sino que es un lenguaje artificial. Cuatro de cada

cinco búsquedas no utilizan operadores, lo cual disminuye la especificidad de las mismas, retornando una mayor cantidad de documentos irrelevantes.

2. El universo de discurso de los motores de búsqueda de la Web es generalmente muy grande y no se encuentra reducido a un área de conocimiento en particular. Esto dificulta las búsquedas en gran medida.
3. "*Do not worry about a large numbers of results: the best ones come first!*" [AA98]. El corpus de documentos de la Web es varios órdenes de magnitud mayor que el de una biblioteca digital o base de datos similar. Los algoritmos de búsqueda que son efectivos para estos últimos deben ser adaptados para manejar este nuevo ambiente que, además de contar con más documentos de tipos o formatos más variados, tiene un tamaño acumulado mucho mayor. E.g., la página principal de Google indicaba en febrero de 2005 que el índice de este motor almacenaba 8.058.044.651 páginas Web.

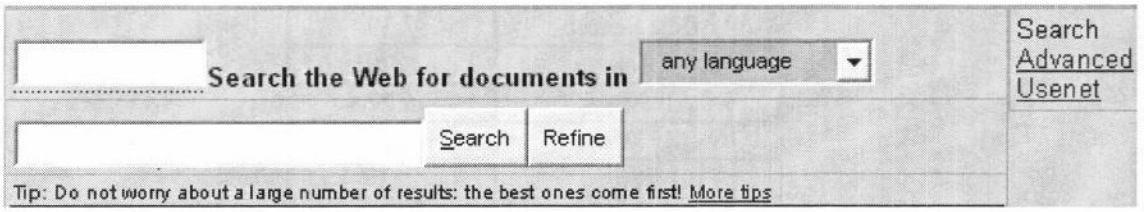


Figura 3.14 – GUI de búsqueda de Altavista en 1998.

4. Uno de los problemas clave de R.I. que se ve acentuado en el caso de la Web es que un ser humano raramente podrá anticipar los términos exactos que otro usaría al buscar o incluso indexar un documento en particular [Fur87]. Esta falta de consistencia dificulta aún más el proceso de búsqueda, indexación y recuperación de documentos relevantes.

   El sitio http://www.cantfindongoogle.com lista necesidades de usuarios junto con búsquedas infructuosas realizadas en Google para intentar satisfacerlas. Otros usuarios agregan comentarios para ayudar a sus pares a encontrar lo que precisan. La siguiente tabla muestra un ejemplo del contenido de esa página:

Can't Find On Google

Posted by	Date posted	What You're Really Looking For	Comments
richie	January 29, 2005 21:01:09	**Really looking for:** windows CE c++ programming board or online tutorial, or source code, etc. **Queries tried:** just about every combo, string, etc. too numerous to mention... **Comments:** i have found links to boards and tutorials, but all the links end up being mostly useless.	Comments (0)

Figura 3.15 – Sitio "*Can't Find on Google*".

5. Indexación: la indexación de un documento involucra la elección de términos que lo describan. Dado que la audiencia de la Web es tan amplia, la elección de términos resultará generalizada de modo de cubrir el amplio espectro de usuarios (i.e., la audiencia del sistema) pero a su vez resultará poco específica o significativa para un usuario en particular. El objetivo no es sólo describir un documento sino también distinguirlo de los demás [Bel00].
6. Indexación: La vista lógica de los documentos indexados no es conocida. Distintos motores de búsqueda utilizan distintas estrategias para la misma (e.g., eliminación de stopwords, stemming, etc.) Si el usuario conociera las estrategias utilizadas, podría ingresar mejores búsquedas para aumentar la relevancia de los resultados.
7. *Crawling*: Uno de los problemas principales de toda técnica de *crawling* es que estas generan tráfico adicional en la red. Esto se debe a que no contemplan, en su mayoría, el trabajo conjunto de distintos *crawlers*. Por lo contrario, cada uno de estos componentes trabaja independientemente, resultando en una redundancia mayor en lo que respecta a la recuperación de información. Una solución a este problema consiste en dividir la Web en secciones de acuerdo a nombres de dominio o ubicación geográfica y asignarle a cada *crawler* o grupo de *crawlers* una sección en particular.
8. *Crawling*: Las páginas recuperadas por el *crawler* y consecuentemente indexadas se vuelven obsoletas rápidamente. Cuando un usuario realiza una búsqueda, los documentos retornados no fueron determinados como relevantes en base a su contenido en tiempo real, i.e., en el momento de la búsqueda, sino a su contenido indexado, el cual probablemente ya no existe o no es el mismo que en el momento de la indexación. Las páginas indexadas pueden haber sido recolectadas por el *crawler* hasta un par de meses atrás. Los motores de búsqueda intentan amortiguar este problema indicando la fecha de indexación junto a cada resultado. Baeza y Yates [BR99] estiman que el porcentaje de enlaces inválidos almacenados en los buscadores se encuentra entre el 2 y el 9%. Este porcentaje es bastante elevado si se considera que los *crawlers* más avanzados visitan hasta diez millones de páginas Web por día. [SEW05]. Cho, García Molina y Page descubrieron que se puede mejorar el *crawling* mediante el uso conjunto de esquemas de ordenamiento y PageRank. Estos facilitarían el acceso a los mejores documentos primero, lo que resultaría en un índice con documentos más relevantes y útiles para el usuario final [CGP98].
9. *Crawling*: hoy en día no es posible acceder a la "Web oculta" (i.e., páginas de acceso protegido, páginas dinámicas, etc.). Esto resulta en que muchos documentos no puedan ser indexados y por lo tanto retornados como resultados para las búsquedas. Este es uno de los problemas más complicados que afecta a R.I.W.
10. Relevancia: este es uno de los problemas clave de los motores y se considera que lo seguirá siendo por mucho tiempo. Aun en los mejores motores como Google o AllTheWeb, la recuperación y precisión de los resultados es pobre en muchos casos. Como ejemplo se cita el siguiente extracto escrito por el dueño de un *weblog* [BO05]. Esta persona observó que los buscadores atraen usuarios de la WWW a su sitio personal como respuesta a búsquedas determinadas para las cuales no hay información relevante en su página. En vez de quejarse de la relevancia de los motores de búsqueda, decidió ayudarlos y cambiar el texto de su sitio para que las páginas relevantes pudieran ser encontradas más fácilmente.

I see a couple hundred google searches that lead to pages in this blog. Queries that lead to my site (on the first page of results) include restaurant related ones like *steak houses in boston*, *restaurants near copley square*, *bugaboo creek price filet* and *weblog for restaurants*, as well as game related ones like *ma ni ki*, *adam and eva game review*, *family boardgames "market size"*, *typo cwali* and a variety of others. To that end, I've tried to write my entries with search engines in mind to some extent. The goal is not to raise the ranking of a particular page, but to make the relevant pages findable.

Despite all of that it's clear some visitors arrive here with queries my pages don't directly answer. In many cases, I'm sure they find the answers they are seeking on other sites. However, in cases where I present something close to the answer but insufficient

Figura 3.16 – *Weblog* de tecnología

11. Relevancia: se observa que la relevancia de los resultados es menor para búsquedas relacionadas con culturas o idiomas diferentes al inglés. Esto es particularmente cierto para motores de búsqueda desarrollados en países anglosajones.
12. Inconsistencia: Una misma búsqueda representada por una secuencia de términos no significa lo mismo en distintos motores de búsqueda. Algunos retornan la unión de las páginas que tienen todos los términos de la búsqueda primero (e.g., Google) y los que tienen uno o más a continuación mientras que otros sólo retornan el segundo conjunto (e.g., HotBot y Altavista).

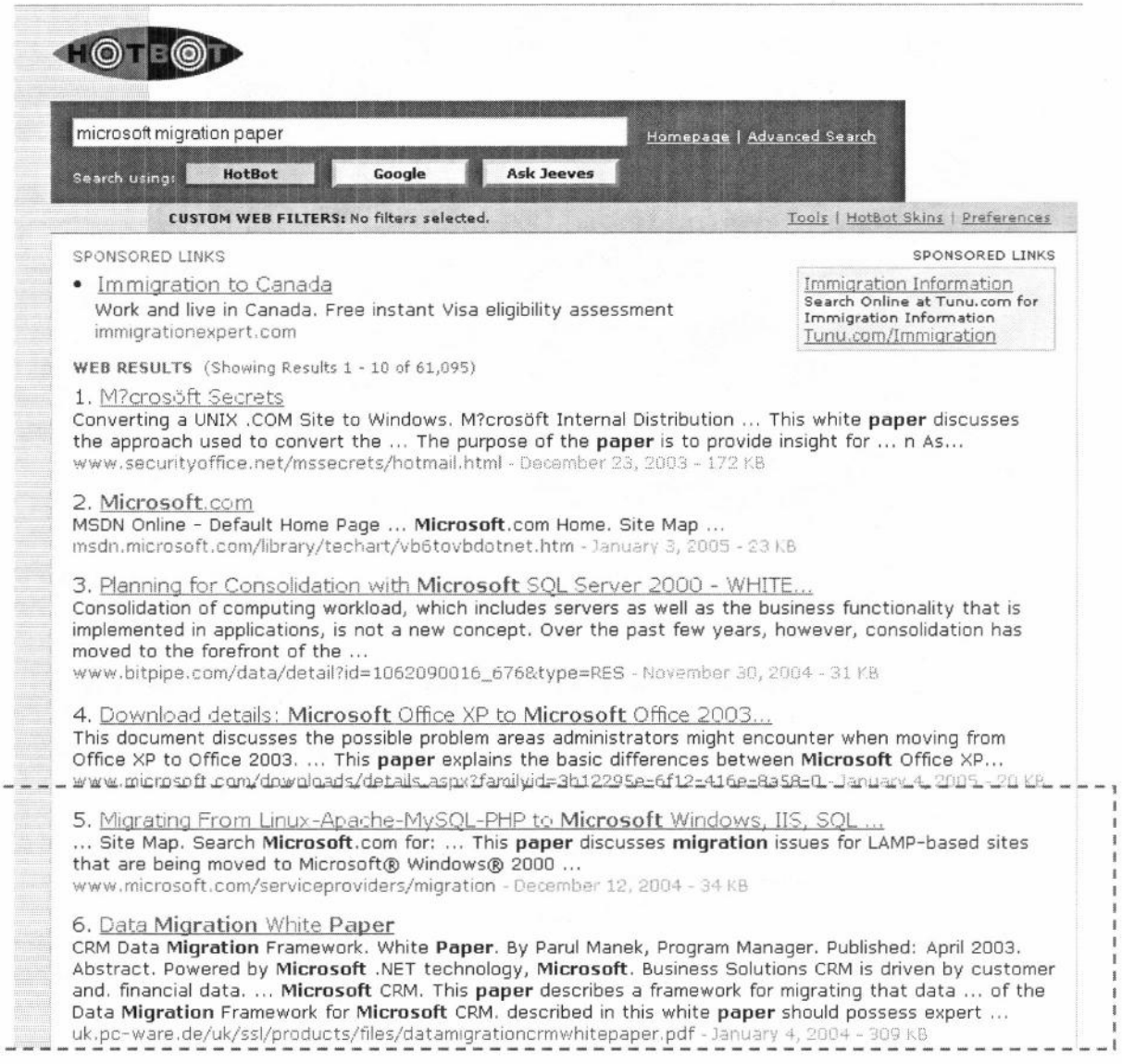


Figura 3.17 – El algoritmo de ranking de HotBot no retorna primero los documentos que contienen todas las palabras de una búsqueda

Además, problemas en el algoritmo de ranking dan lugar a resultados mal ordenados aun en los motores cuyo algoritmo de ranking clasifica primero los documentos que contienen más términos de una búsqueda:

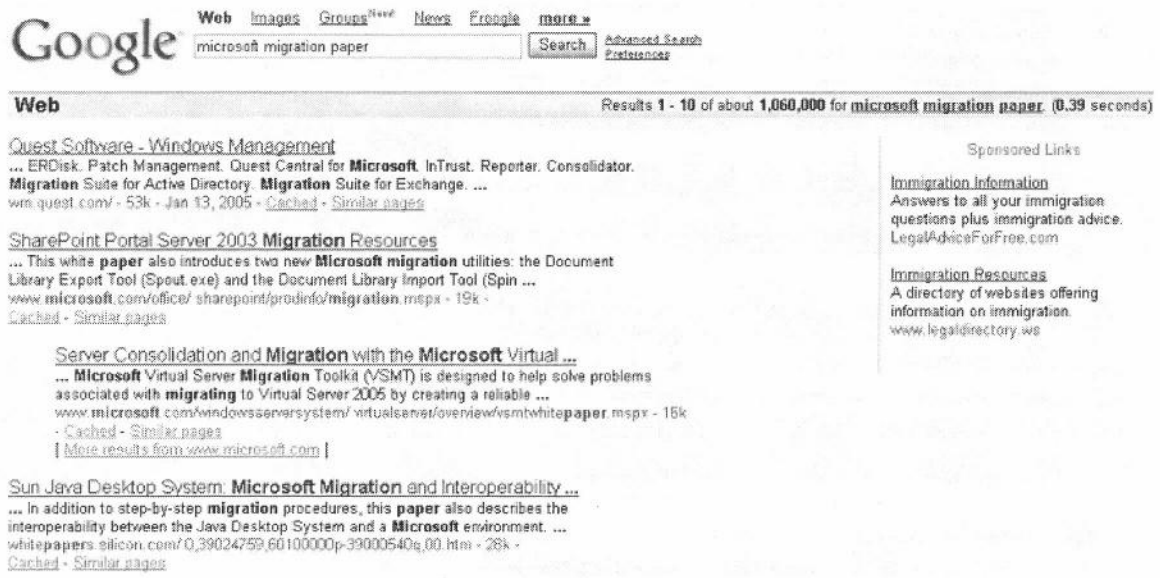


Figura 3.18 – Problemas del algoritmo de ranking de Google: los documentos que contienen todas las palabras de una búsqueda no son retornados primero.

13. Ingreso de búsquedas: Los usuarios no comprenden lo que significa utilizar un conjunto de palabras para realizar búsquedas [BR99]. Esto dificulta la elección de los términos y operaciones a realizar sobre los mismos, resultando en una cadena de búsqueda no ideal.
14. Semántica
    a. Pérdida de contexto parcial debido a la indexación y posterior normalización del índice. Si se ingresa la siguiente búsqueda:

       Escuela de la Armada

       ésta se traduce como *escuela de la armada*, perdiendo así parte de su semántica, lo que resultará en páginas adicionales incluidas en la respuesta que no serán de interés.
    b. Búsqueda de sinónimos. Si bien el índice típico de un motor de búsqueda almacena sinónimos de los términos de su vocabulario, la recuperación de documentos que contienen sinónimos de las palabras de la búsqueda no tiene hoy el nivel de performance esperado (i.e., los documentos que contienen sinónimos deberían ser retornados como relevantes pero esto no ocurre en la medida adecuada).
    c. Una serie corta de términos no captura adecuadamente la semántica de un documento de texto. Esto afecta la precisión y recuperación de un motor de búsqueda.
    d. Polisemia. Cuando un término tiene más de un significado, este puede generar resultados que lo contengan para los cuales su significado no es el deseado. Este es un problema muy difícil de resolver, ya que determinar el contexto de un término dentro de una búsqueda corta y concisa es una tarea muy compleja si no imposible.

15. Errores: errores ortográficos, tanto en la búsqueda como en las páginas Web, agregan complejidad a R.I.W. Se estima que entre el diez y el veinte por ciento de los resultados se pierde debido a este problema y a 14.b [BR99].
16. Errores: los errores de tipeo en búsquedas y documentos generan una pérdida de hasta un 50% de los resultados deseados.

La siguiente tabla resume los problemas principales de R.I.W. mencionados anteriormente:

Problemas de R.I.W.	
Datos	**Interacción usuario – M.B.**
Calidad	Gramática del lenguaje de búsqueda
Dispersión	Análisis de resultados
Volatilidad	Errores en cadenas de búsqueda (ortográficos, de tipeo, etc.)
Volumen	Desconocimiento del proceso de indexación resulta en búsquedas empobrecidas
Falta de estructura	Funcionalidades similares se comportan inconsistentemente en diversos buscadores.
Redundancia	
Heterogeneidad	
Pérdida de semántica	
Polisemia	
Errores ortográficos	
Errores de tipeo	

Tabla 3.11 – Problemas de R.I.W.

Este capítulo definió la arquitectura de un motor de búsqueda, sus principales métricas y los problemas de R.I.W. El capítulo 4 introduce el método utilizado en este trabajo para derivar métricas de R.I.W. desde la perspectiva del usuario: *Goal Question Metric*. El capítulo 7 - Recolección y Análisis de Datos, propone mejoras a los problemas de R.I.W. citados en esta sección en base a los datos recolectados mediante el experimento.

CAPITULO 4. METRICAS: *GOAL QUESTION METRIC*

"You cannot control what you cannot measure."

DeMarco, 1982

4.1 Introducción

Este trabajo desarrolla un método que permite comparar la calidad de los motores de búsqueda desde la perspectiva del usuario. Por lo tanto, es necesario comprender primero qué es lo que se desea evaluar a fin de poder medir su calidad. Para lograrlo, se utiliza la metodología *Goal Question Metric* (GQM) [BR88, FP97, SB99], la cual permite derivar métricas en base a un objetivo buscado.

Este capítulo sirve como introducción teórica al capítulo 5 - Derivación de Métricas. Aquí se describe qué son las métricas, cómo se derivan mediante *Goal Question Metric* (GQM) y cómo se utilizan.

4.2 ¿Qué Son las Métricas?

Las métricas forman parte de nuestra vida cotidiana. Entre otras aplicaciones, sirven para comparar precios, el tamaño de la ropa, la temperatura ambiental, la altura de una persona, etc. Las métricas ayudan a comprender el mundo que nos rodea.

Una entidad es cualquier objeto o evento del mundo real. Las entidades se describen identificando las características o atributos más importantes que la diferencian de otras. Las métricas, por otra parte, proveen información sobre los atributos de una entidad. Por lo tanto, es incorrecto decir que se miden entidades, ya que lo que realmente se mide son los atributos de las mismas.

Definición 4.1: Una medida es el proceso de asignar un número o un símbolo a atributos de entidades del mundo real, de forma tal que describan a la entidad de acuerdo a reglas definidas claramente [FP97].

Al describir las entidades utilizando atributos, generalmente se definen los atributos con símbolos o números. Por ejemplo, la edad de una persona se puede definir con un número expresado en años. El tamaño de un vaso de gaseosa se puede medir en chico, mediano o grande. Estos símbolos y números son abstracciones utilizadas para reflejar nuestras percepciones del mundo real. Con la definición de números y símbolos se intenta preservar las relaciones que se observan entre las entidades [FP97]. Por lo tanto, para poder evaluar una entidad es necesario conocer y analizar sus atributos.

La exactitud de una métrica depende del instrumento utilizado para medirla. La altura puede

ser medida de forma exacta utilizando adecuadamente un metro. Pero existen algunas métricas que no son exactas ya que sus mediciones dependen del juicio de las personas que las utilizan. Como ejemplo se puede citar la relevancia de un documento retornado por un motor de búsqueda.

4.3 Objetivos de las Métricas

Cada métrica debe tener una meta claramente definida. Los objetivos de las métricas deben ser claros, específicos y estar relacionados con la necesidad del usuario, ya que ésta puede diferir según el perfil de cada persona o grupo de personas involucrado. Los objetivos indican cómo se debe recolectar la información a utilizar para evaluar las métricas.

4.4 Representación Teórica

La representación teórica de las métricas busca formalizar la percepción humana del mundo real. Las métricas obtenidas representan atributos de entidades del mundo y por ende la manipulación de estos datos debe preservar la relación entre las entidades.

Debido a la forma en que percibimos el mundo, los seres humanos tratamos de comprenderlo comparando los objetos que lo componen. Por ejemplo, la relación "es mayor que" es una relación empírica para la edad. Se puede observar que cierta persona es mayor que otra sin necesidad de preguntarle su edad. Sin embargo, esta observación refleja un conjunto de reglas implícitas impuestas sobre el conjunto de personas observadas.

Definición 4.2: Medición es el mapeo desde el mundo empírico al mundo relacional (i.e., formal) [FP97]. Consecuentemente, una métrica es el número o símbolo asignado a una entidad mediante un mapeo relacional para caracterizar un atributo.

En las relaciones empíricas no siempre se obtienen los mismos valores al medir los mismos atributos, especialmente cuando estas relaciones reflejan una preferencia personal. En estos casos, se puede realizar una evaluación subjetiva pero el resultado no es necesariamente una métrica desde la perspectiva de la teoría de medición.

4.4.1 Las Reglas de la Relación

Se explicó cómo se utilizan las métricas para caracterizar a los atributos de una entidad. Para ello se toma la entidad del mundo real y se intenta comprenderla. De esta manera, se puede deducir que el mundo real es el dominio de la relación y el mundo matemático es el codominio. Cuando se relaciona un atributo con el mundo matemático, se puede optar por varios tipos de mapeos y codominios (números reales, números enteros, símbolos, etc.).

Hasta aquí se explicó teóricamente para qué sirven las métricas pero falta saber cómo se las puede definir teniendo en cuenta las entidades a evaluar y sus objetivos.

4.5 *Goal Question Metric*

Como se mencionó anteriormente, es necesario determinar cómo definir métricas para una entidad a evaluar. Para ello se introduce la metodología *Goal Question Metric* (GQM), la cual es utilizada en el siguiente capítulo como herramienta para la derivación de nuevas métricas para un motor de búsqueda (entidad).

Basili y Rombach [BR88] son los creadores del método GQM. La característica más importante de este método es que se centra en los objetivos [SB99]. Según los autores, siempre debe existir una razón para realizar las mediciones y se debe rechazar la noción de un conjunto fijo de métricas. Un conjunto fijo consiste en las métricas que ya estaban definidas sin tener en cuenta las metas del proyecto. Por ejemplo, decir que se va a medir la cantidad de líneas del código fuente de un proyecto. Es por ello que estos investigadores proponen un método que ayuda a construir un conjunto de métricas específicas para los objetivos y el ambiente de una organización. Todo proceso de medición debe comenzar tratando de responder cuál es el propósito de la medición y no qué métricas se deben utilizar.

En otras palabras, esto significa que para mejorar la organización, el proceso o el proyecto en cuestión, se deben definir objetivos que se puedan medir. A través de GQM, se refinan los objetivos en preguntas y luego en métricas que suministren toda la información necesaria para responder las preguntas. El método GQM provee un plan de medición que abarca un conjunto particular de problemas y un conjunto de reglas para interpretar los datos. La interpretación determina si se alcanzaron o no los objetivos.

El método GQM provee un *framework* que incluye tres niveles o pasos:

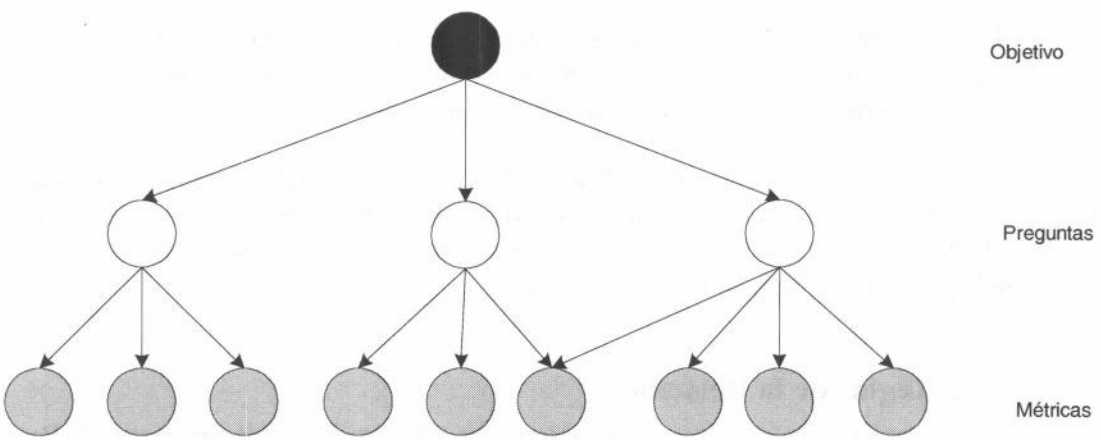


Figura 4.1 – Método *Goal-Question-Metric.*

Como se observa en la figura 4.1, el framework GQM tiene tres niveles [S97]:

4.5.1 Nivel Conceptual (Objetivo)

Una meta u objetivo se define teniendo en cuenta el objeto a estudiar. Un objeto puede ser un producto, un proceso, un recurso, un proyecto o una empresa.

4.5.2 Nivel Operacional (Preguntas)

Se utiliza un conjunto de preguntas para caracterizar la forma de evaluar y alcanzar los objetivos específicos. Las preguntas tratan de caracterizar el objeto a medir teniendo en cuenta el punto de vista de la persona que utilizará la métrica.

4.5.3 Nivel Cuantitativo (Métricas)

Un conjunto de datos es asociado con cada pregunta de forma que la misma se responda de manera cuantitativa. Los datos pueden ser:

- Objetivos: Si los datos sólo dependen del objeto a ser medido y no del punto de vista de quien realiza la medición.
- Subjetivos: Si los datos dependen del objeto a ser medido y del punto de vista de quien realiza la medición.

De esta manera, utilizando GQM se generan métricas relacionadas con los objetivos. El objetivo aporta el propósito para recolectar datos y las preguntas dicen cómo deben ser utilizados los datos.

Todos los objetivos y preguntas deben estar asociados al menos con un par de entidades y atributos. Es decir, un objetivo deriva en una pregunta y la respuesta a la pregunta requiere que se midan algunos atributos de una entidad (o posiblemente varios atributos de una entidad, o varios atributos de varias entidades) [FP97].

4.5.4 *Template* para la Definición de Objetivos [FP97], [S98], [BR88]:

Dada la importancia de la definición de los objetivos, Basili y Rombach recomiendan un *template* que se divide en tres componentes:

- Propósito: (Caracterizar, evaluar, predecir, motivar, etc.) el (proceso, producto, modelo, etc.) para lograr (entenderlo, evaluarlo, manejarlo, mejorarlo, etc.)
  Ejemplo: Evaluar el proceso de mantenimiento de un producto para mejorarlo.
- Perspectiva: Examinar el (costo, eficiencia, correctitud, defectos, cambios, etc.) desde el punto de vista del (desarrollador, gerente, cliente, usuario, etc.)
  Ejemplo: Examinar el costo del proyecto desde el punto de vista del gerente.
- Ambiente: El ambiente consiste en factores del proceso, factores de las personas, métodos, herramientas, restricciones, etc.
  Ejemplo: El staff de mantenimiento se encuentra poco motivado porque sus programadores tienen acceso limitado a las herramientas.

4.5.5 Pasos del *Framework* GQM

Existen cuatro pasos básicos en GQM [S98]:

1. Identificación de objetivos
   1.1. Definición de objetivos
   1.2. Redacción de preguntas
   1.3. Derivación de métricas
2. Proceso de planificación de las mediciones
   2.1. Identificación de la hipótesis
   2.2. Diseño del plan de medición
3. Realización de las mediciones
   3.1. Recolección de datos
   3.2. Validación de datos
4. Análisis e interpretación

Cada uno de estos pasos es ejecutado en este trabajo en los capítulos siguientes. La distribución de los mismos se describe a continuación:

Paso 1 – Identificación de Objetivos: En el capítulo 5 - Derivación de Métricas, se analizan los atributos deseables de un motor de búsqueda desde los puntos de vista técnico y de usuario y se intenta hallar métricas que ayuden a comparar diferentes motores de búsqueda mediante la aplicación del método GQM.

Paso 2 – Proceso de Planificación de las Mediciones: En el capítulo 6 – Diseño del Experimento, se presentan los objetivos del experimento para el cual se planificaron los atributos a medir con respecto a los motores de búsqueda y la forma en que éstos serían medidos. El experimento se basa en la perspectiva del usuario, ya que tal como se observó en las métricas recolectadas en el capítulo 3 - R.I. Web, esta perspectiva generalmente no es tenida en cuenta al medir los atributos de los motores de búsqueda.

Paso 3 – Realización de las Mediciones: En el capítulo 7 - Recolección y Análisis de Datos, se recolecta la información obtenida mediante el experimento y se realizan las mediciones.

Paso 4 – Análisis e Interpretación: En la sección 7.4 se comparan los motores de búsqueda y se presentan las conclusiones del experimento con respecto a su hipótesis.

CAPITULO 5. DERIVACION DE METRICAS

"Measure what is measurable,
and make measurable what is not so"

Galileo Galilei (1564 – 1642)

5.1 Introducción

Como se indicó en los capítulos anteriores, la meta de este trabajo es evaluar los motores de búsqueda desde las perspectivas técnica y de usuario. Para lograr este objetivo, es necesario determinar cuáles son los atributos que caracterizan a los motores de búsqueda y desarrollar métricas para un subconjunto de atributos que se desea estudiar. Estas métricas permiten comprender y comparar distintos buscadores mediante el estudio de sus atributos.

Consecuentemente, este capítulo se centra en los atributos que caracterizan a un motor de búsqueda y en la aplicación del método *Goal Question Metric* para la derivación de métricas.

5.2 Entidad a Describir

Como se observó en el capítulo anterior, las métricas se aplican a los atributos de las entidades. Es decir, se describe una entidad identificando las características más importantes que la diferencian de otra entidad y analizando estas características cuantitativamente.

En este trabajo, la entidad está representada por un motor de búsqueda. Sus atributos generales se describen a continuación.

5.3 Categorías de un Motor de Búsqueda - Atributos Generales

Los siguientes atributos generales caracterizan a los motores de búsqueda. Estos se listan sin seguir un orden en particular.

Amigabilidad
Un motor de búsqueda es amigable para un usuario si éste lo encuentra fácil de usar. Es decir, si entiende cómo realizar las búsquedas y qué devuelve el buscador. Asimismo, el usuario comprende y utiliza las características especiales del motor de búsqueda.

Performance
Un motor de búsqueda es eficiente si utiliza sus recursos adecuadamente.

Ejemplo: El índice de un buscador está preparado para responder una consulta en el tiempo esperado por el usuario.

Correctitud
Un motor de búsqueda es correcto si cumple con las expectativas del usuario en cuanto a la relevancia y accesibilidad de los documentos recuperados.

Extensibilidad de Búsquedas
Facilidad de ampliar consultas realizadas o por realizar mediante la utilización de diversas estrategias que permiten al buscador obtener mejores resultados.

Evolutibilidad
Un motor de búsqueda es adaptable a modificaciones y novedades que surjan en el ambiente en el que se utiliza y en el estado del arte de R.I.W.

Visibilidad
La información recuperable puede ser obtenida de todo el universo de análisis (i.e., la Web). Es decir, se pueden encontrar todos los documentos relevantes para una consulta que existen en la Web.

Personalización
Adecuación del buscador a las necesidades y preferencias del usuario. Esta funcionalidad está relacionada con la amigabilidad, ya que el personalizar ciertas preferencias del motor de búsqueda permite al usuario sentirse más cómodo y a gusto con la herramienta que está utilizando.
Ejemplo: Elección del idioma del motor de búsqueda.

Confiabilidad
El motor de búsqueda es confiable si el usuario puede obtener la información que necesita sin dudar de su veracidad.

Educativo
La definición de esta categoría se provee desde dos puntos de vista:

- Motor de búsqueda: Mejoramiento constante de los procesos de búsqueda y recuperación de la información a través de la interacción de sus componentes con los usuarios y los sitios Web.
- Usuario: Mejoramiento del proceso de búsqueda de información a través de su interacción con los motores de búsqueda.

Granularidad
Grado de especificidad de la información alcanzable por un motor de búsqueda.
Ejemplo: El motor de búsqueda de Medline (www.medline.com) tiene alta granularidad ya que sólo devuelve documentos relacionados con la medicina.

Integridad

Imparcialidad en la determinación de la relevancia de los documentos. El algoritmo de ranking no modifica su estrategia teniendo en cuenta los sitios que alteran sus páginas para obtener un mejor posicionamiento en el resumen de resultados.

Asociatividad
Agrupamiento de la información recuperada en base a distintos factores.

Usabilidad
El motor de búsqueda y la información recuperada son utilizables por el usuario.
Ejemplo: Poder acceder a los documentos que resultan relevantes mediante la realización de una búsqueda.

Interoperabilidad
El motor de búsqueda es capaz de interactuar con otros sistemas.
Ejemplo: Se puede incluir el motor de búsqueda dentro de un sitio Web y realizar búsquedas desde el mismo.

Nota: Un subconjunto de las definiciones anteriores está inspirado en los trabajos de [GMJ91] y [BMY98].

Como se observa, estos atributos son muy genéricos y sirven para realizar una primera categorización de las características de la entidad a evaluar. Pero para poder diferenciar mejor los motores de búsqueda mediante sus atributos, es necesario un mayor nivel de detalle.

5.4 Explotación de las Categorías - Atributos Específicos

A continuación, cada categoría es explotada de forma tal que los atributos de los motores de búsqueda sean más específicos. Los atributos descriptos para cada categoría forman parte de las funcionalidades deseables de un motor de búsqueda.

Amigabilidad

- Facilidad de realización de la búsqueda
- Identificación eficiente de documentos relevantes
- Sugerencias ante novedades sobre búsquedas frecuentes
- Sugerencias relacionadas con tareas en ejecución
- Clasificación de resultados
- Contacto
- Almacenamiento de información de búsquedas precedentes
- Utilización de perfiles y contexto

Performance

- Eficiencia
- Caché

- Utilización de *idf* y *tf* [5]
- Almacenamiento de *tf* en el vocabulario
- Almacenamiento de información de búsquedas precedentes
- Eficiencia de respuesta

Correctitud

- Exactitud
- Disponibilidad

Extensibilidad de Búsquedas

- Búsqueda por sinónimos
- Búsqueda restringida
- Búsqueda aproximada
- Búsqueda de frases
- Búsqueda por proximidad
- Búsqueda de expresiones regulares
- Búsqueda de patrones extendidos
- Búsqueda de páginas relacionadas
- Búsquedas especiales
- Análisis lingüístico

Evolutibilidad

- Escalable
- Optimización
- Adaptabilidad a tecnología de componentes
- Adaptabilidad a tecnología de sitios Web
- Adaptabilidad al cambio

Visibilidad

- Completitud
- Alcance

Personalización

- Intervención del usuario en la confección del ranking
- Personalización de motores de búsqueda
- Formato del resultado

Confiabilidad

- Confiabilidad
- Disponibilidad

[5] Ver apéndice IV – Recuperación de la información para la definiciones de *idf* y *tf*.

Educativo

- Ofrecimiento de servicios adicionales al usuario
- Sugerencias al usuario ante respuestas nulas
- *Relevance feedback*
- Sugerencias ante novedades sobre búsquedas frecuentes
- Sugerencias relacionadas con tareas en ejecución

Granularidad

- Buscadores genéricos – Buscadores especializados

Integridad

- Objetividad
- Tratamiento de *spamming* [6]

Asociatividad

- *Clustering* de documentos
- *Clustering* de sitios
- Clasificación de resultados

Usabilidad

- Exactitud
- Disponibilidad
- Envío de páginas para actualización del índice
- Multilenguaje
- Personalización de motores de búsqueda

Interoperabilidad

- Importación de motor de búsqueda

Se observa que algunos atributos pertenecen a más de una categoría.

5.5 Funcionalidades Deseables de un Motor de Búsqueda

Como se indicó en la sección anterior, cada categoría posee un conjunto de funcionalidades que las describen. Estas funcionalidades o atributos caracterizan a los motores de búsqueda (i.e., las entidades).

En el capítulo anterior, se observó que los objetivos de GQM pueden diferir según el perfil de cada persona. Por lo tanto, se divide el conjunto de funcionalidades teniendo en cuenta dos tipos de perfiles: usuario y técnico. El punto de vista del usuario es aquel que se centra en qué obtiene del motor de búsqueda en lugar de cómo lo obtiene y, por lo tanto, está

[6] Tratamiento de *spamming* se explica en la sección 5.6.

relacionado con la expectativa y experiencia del usuario. El punto de vista del técnico es el que se centra en cómo se comportan las funcionalidades del motor de búsqueda.

A continuación, se describe cada funcionalidad según los tipos de perfiles definidos.

5.5.1 Funcionalidades desde la Perspectiva del Usuario

1. **Facilidad de realización de la búsqueda:** Simplificar la tarea del usuario en cuanto a indicarle al buscador qué es lo que realmente desea buscar. Esto está relacionado con el diseño del GUI [H00] entre otros, ya que éste determina cómo se realizan las búsquedas.
2. **Identificación eficiente de documentos relevantes**: Poder encontrar rápidamente los documentos relevantes para el usuario entre los devueltos por el buscador. El objetivo principal es lograr que las respuestas estén ordenadas por el orden de relevancia que les daría el usuario. Esto está relacionado con el diseño del GUI [H00] entre otros, ya que éste determina cómo se presentan las respuestas al usuario.
3. **Sugerencias ante novedades sobre búsquedas frecuentes**: El motor de búsqueda propone al usuario nuevos documentos para búsquedas que realiza con asiduidad.
4. **Sugerencias relacionadas con tareas en ejecución**: El motor de búsqueda plantea al usuario posibles búsquedas vinculadas a la tarea que esté realizando en determinada oportunidad. Ejemplo: Al realizar una presentación utilizando *PowerPoint* sobre un tema en particular, el motor sugiere una búsqueda sobre el tópico o bien recupera documentos y los muestra por defecto.
5. **Clasificación de resultados**: El motor de búsqueda permite agrupar los resultados por palabras clave, facilitando de este modo la búsqueda entre los documentos retornados.
6. **Contacto**: El usuario de un motor de búsqueda puede comunicarse con los administradores del mismo a fin de hacerle llegar comentarios, quejas, dudas o sugerencias.
7. **Eficiencia de respuesta**: El tiempo que se tarda en buscar los documentos a recuperar para una consulta determinada es aceptable para el usuario. Se toma en cuenta el período comprendido entre el ingreso de la búsqueda y la obtención de los resultados. No involucra la elección de los documentos relevantes.
8. **Exactitud**: Los documentos recuperados ante una consulta son relevantes para el usuario.
9. **Disponibilidad**: Los documentos recuperados para una consulta pueden ser accedidos sin inconvenientes. Entre los problemas posibles se encuentran los siguientes: el servidor que contiene el documento no está en funcionamiento, la página tarda mucho tiempo en cargar (*timeout*), enlaces caídos, sitios redireccionados o que dejaron de existir, etc. De este modo, en algunos casos la información puede no estar disponible a pesar de pertenecer a las respuestas.
10. **Búsqueda por sinónimos**: Extensión de la consulta realizada mediante la utilización de sinónimos. Al ingresar la búsqueda, el motor puede ampliarla incorporando sinónimos de las palabras clave y así obtener una gama mayor de documentos recuperados sobre el tema de la consulta.
11. **Búsqueda restringida**: Ver Tabla 5.1 – Restricciones para búsquedas.

Tipo de Restricción	Detalle

Perfiles	Perfil del usuario que realiza la búsqueda. Ejemplos: técnico/no técnico, por edades, etc.
Antigüedad de documentos	Restringir los resultados a documentos posteriores o previos a cierta fecha o pertenecientes a un rango previamente determinado (Por ejemplo, los documentos se comparan con la fecha de creación o fecha de ultima actualización)
Dominios – Universo de análisis	Tipos de dominio de los sitios que contienen los documentos a recuperar (seleccionar en qué parte de la Web buscar: sitios académicos, comerciales, toda la Web, etc.)
Ubicación geográfica	País de procedencia de los documentos
Idioma	Lenguaje en que están escritos los documentos
Cantidad de páginas	Restringir la cantidad de páginas Web que contendrán los resultados a devolver por una consulta
Cantidad de resultados	Cantidad de *hits* a recuperar
Tamaño del texto	El tamaño del texto correspondiente a los enlaces devueltos como respuesta no debe superar cierta medida
Exactitud	Restricción por porcentaje de similitud entre la búsqueda realizada y cada respuesta. Ejemplo: Buscador de Sun.

Tabla 5.1 - Restricciones para búsquedas.

12. **Búsqueda aproximada**: Realización de la búsqueda y recuperación de documentos en los cuales la palabra clave aparece con errores. Estos pueden ser ortográficos o de separadores, entre otros. Si el buscador no omitiera los documentos en los cuales hay un error de este tipo, se obtendría un mayor porcentaje de documentos relevantes recuperados (i.e., aumenta la *recuperación*).
13. **Búsqueda de frases**: La consulta debe recuperar documentos que contengan una frase determinada, tal como fue ingresada por el usuario.
14. **Búsqueda por proximidad**: Se precisa conocer la ubicación relativa de las palabras clave en el texto a fin de determinar qué documentos se consideran relevantes y cuáles no. De acuerdo a la separación entre las palabras de la consulta se puede decidir si un documento debe o no ser incorporado a la respuesta. En algunos casos, evitar exigir que ciertos términos de una búsqueda estén cerca puede provocar el retorno de muchos documentos irrelevantes (i.e., disminuye la *recuperación*). El usuario puede indicar la distancia máxima que debe existir en el texto entre las palabras de la consulta.
15. **Búsqueda de expresiones regulares y patrones extendidos**: Este tipo de búsqueda es adecuado en los casos en que se precisa hallar un conjunto de cadenas que debe ser

especificado mediante herramientas más potentes. Los patrones extendidos son clases de caracteres, caracteres opcionales o repetibles y comodines, entre otros.

16. **Búsqueda de páginas relacionadas**: El motor de búsqueda devuelve páginas relacionadas con las recuperadas para aumentar el espectro de resultados. Es decir, si se está buscando "enfermedades de perros", se puede además mostrar documentos relacionados con "enfermedades de mamíferos domésticos". Una forma de implementar este atributo es agregar una opción al margen de los documentos recuperados que permita visualizar los documentos similares a éstos. Esta funcionalidad esta relacionada con la similitud de documentos[7]. Si las respuestas retornadas son escasas o nulas, se incluye con los documentos recuperados un resumen de los que están relacionados con los mismos. Ejemplo: "*Google Similar Pages*".

Google
... Advertise with Us - Business Solutions - Services & Tools - Jobs, Press, & Help ©2004 **Google** - Searching 4,285,199,774 web pages.
www.**google**.com/ - 3k - Mar 26, 2004 - Cached - » **Similar pages** «

Figura 5.2 – Google *Similar Pages.*

17. **Búsquedas especiales**: Consultas extendidas mediante el agregado de palabras clave del buscador. Con la utilización de estas palabras clave y el agregado de parámetros, entre otras posibilidades, se pueden buscar páginas que contengan cierta *clase* de un lenguaje de programación, un vínculo a otra con un URL determinado, similares a una dada, que contengan la palabra o expresiones especificadas en el texto de un hipervínculo, que cuenten con imágenes cuyo nombre contenga las palabras pasadas como parámetro, etc. Ejemplo: "Funciones especiales para búsquedas" de *Altavista.*
18. **Análisis lingüístico**: Capacidad de lematizar (analizar las variantes de una palabra), etiquetar (distinguir entre distintas partes del discurso) y detectar frases comunes [N00]. Ciertos motores de búsqueda también incluyen el análisis distintivo por mayúsculas y minúsculas. Ejemplo: *Altavista.*
19. **Completitud**: Una fracción significativa de los documentos relevantes existentes en la Web para una consulta determinada es recuperada. Esta funcionalidad está condicionada por el *crawling.*
20. **Personalización de motores de búsqueda**: Facilidad de modificar ciertas características del diseño del GUI de los buscadores para que se presente de un modo personalizado de acuerdo a los gustos del usuario. Ejemplo: "*My* MSN".
21. **Formato del resultado**: Personalización del formato de las respuestas. Una de las características deseables es "ver información de página completa" que permite ampliar el resumen, caché o *snippet* presentado al usuario para cada documento recuperado.
22. **Confiabilidad**: Los documentos recuperados por el buscador son aceptados como válidos por parte del usuario. En el caso de la Web, la confiabilidad está muy relacionada con la calidad de los datos. El principal obstáculo para lograr la confiabilidad necesaria es la falta de un proceso de editorialización de documentos. Es por ello que los datos presentados pueden ser falsos, inválidos (ser demasiado viejos, etc.), estar pobremente

[7] Ver Apéndice IV – Recuperación de la Información.

escritos o contener muchos errores de diferentes tipos (ortográficos, de tipeo, gramaticales, de OCR, etc.) [BR99].

23. **Ofrecimiento de servicios adicionales al usuario**: El objetivo es ayudarlo a que aprenda a buscar cada vez mejor lo que desea, de modo de evitar que los motores de búsqueda continúen eternamente intentando "adivinar" lo que necesita. Algunos ejemplos: *newsletters*, tutoriales, consejo del día, archivos para bajar de la Web, páginas accesibles desde el motor con información para mejorar las búsquedas, cambios en el diseño del GUI, sugerencias para modificar la búsqueda (al estilo "¿Estabas buscando...?" de MSN), entre otros.
24. **Sugerencias al usuario ante respuestas nulas**: Proponer cambios en algunos de los términos involucrados cuando el conjunto de respuestas es vacío, asumiendo posibles errores de tipeo entre otras causas.
25. **Buscadores genéricos – Buscadores especializados:** Facilidad de elegir entre diversos motores de búsqueda o entre distintas partes de uno determinado, dependiendo de si se desea realizar una consulta sobre un tema específico o sobre uno genérico. Si se trata de un solo buscador, éste puede determinar la granularidad de la búsqueda. Ejemplo: el motor de búsqueda de *NEC Research Institute* (http://citeseer.ist.psu.edu/), el cual sirve para realizar búsquedas específicas de artículos científicos en el área de Ciencias de la Computación.
26. **Objetividad**: El ranking de documentos recuperados se realiza de manera objetiva, sin influencia de acuerdos comerciales o de otro tipo. Esta funcionalidad está ligada a *tratamiento de spamming*[8].
27. **Multilenguaje**: Traducción de páginas a idiomas conocidos por el usuario.
28. **Importación de motor de búsqueda**: Posibilidad de acceder al buscador desde el sitio de una empresa, no mediante un enlace sino a través de un motor embebido en las páginas de la *intranet* de la misma, entre otras posibilidades. Ejemplos: *Altavista*, *Google*.

5.5.2 Funcionalidades desde la Perspectiva Técnica

1. **Almacenamiento de información de búsquedas precedentes**: El motor de búsqueda permite guardar la información concerniente a x búsquedas ya realizadas, donde x es un parámetro numérico que puede ser fijado por el usuario.
2. **Utilización de perfiles y contextos:** El motor de búsqueda considera el perfil del usuario y el contexto en el cual realiza sus consultas [EU01].
3. **Eficiencia**: Los algoritmos utilizados por el motor de búsqueda tienen alta performance, funcionan adecuadamente y el tiempo de respuesta es aceptable.
4. **Caché**: Almacenamiento de las páginas en la memoria caché del buscador con el fin de recuperarlas para los usuarios como una copia de seguridad, en caso que el servidor de una página falle temporalmente o que se desee acceder rápidamente al contenido de los sitios. Si el servidor no funciona correctamente, la memoria caché del motor de búsqueda puede ser una alternativa viable para acceder al documento deseado. Además, generalmente el caché puede ser accedido mucho más rápido que entrando al sitio Web mediante el enlace devuelto como parte de la respuesta. De todos modos, se debe tener en cuenta que al utilizar este medio la información disponible puede no estar actualizada. Esta funcionalidad se puede deshabilitar a pedido del sitio Web. La razón de estos

[8] *Tratamiento de spamming* se explica en la sección siguiente.

pedidos suele ser que el caché no se actualiza convenientemente o que no presenta la información que los dueños del sitio desean mostrar, ya que en ciertos motores el caché está compuesto por un resumen de la página en cuestión.

5. **Utilización de *idf* y *tf***: Determinar la finalización de la búsqueda al alcanzar cierto nivel predeterminado para el *idf* y/o el *tf*.
6. **Almacenamiento de *tf* en el vocabulario**: La frecuencia de los términos es incluida en el vocabulario a fin de reducir el orden de complejidad del algoritmo de búsqueda.
7. **Optimización**: Los componentes del buscador son óptimos con el objetivo de aumentar su eficiencia y eficacia.
8. **Adaptabilidad a tecnología de componentes**: Adecuación a los cambios tecnológicos que puedan llegar a sufrir los componentes del motor de búsqueda (i.e., *crawler*, indexador, interface de usuario, motor de consultas, entre los más importantes).
9. **Adaptabilidad a tecnología de sitios Web**: Adecuación a los cambios que surgen en cuanto a las tecnologías de sitios Web. Por ejemplo, cuando apareció la tecnología multimedia se incorporó a ciertos buscadores una funcionalidad que permite hallar archivos de sonido, imágenes, etc.
10. **Adaptabilidad al cambio**: Adecuación del *crawler* a la frecuencia de actualización de los sitios para mantener el índice al día.
11. **Alcance**: Indexación de toda la Web en todo momento.
12. **Intervención del usuario en la confección del ranking**: El motor de búsqueda toma en cuenta las preferencias del usuario al momento de confeccionar el ranking de R.I.
13. ***Relevance Feedback*/Capacidad de Aprendizaje**: Aprendizaje de las búsquedas realizadas exitosamente. Una de las posibilidades de aprendizaje es que si el usuario decide entrar al sitio correspondiente al iésimo enlace de las respuestas, el motor de búsqueda puede almacenar los clics y así "aprender" de las elecciones realizadas por el usuario. Asimismo, permite confirmar estadísticamente la efectividad en la confección del ranking. Otro caso se presenta cuando el usuario desea obtener documentos similares a los recuperados, lo que le permite al buscador aprender cuál de las respuestas fue la más relevante para el usuario. Ejemplos: "*more like this*" o "*more pages from this site*" de *Altavista*.
14. **Tratamiento de *spamming***: Se distingue entre las páginas, diferenciándolas de entre las que realizan *spamming*. En este caso, el término *spamming* se refiere a los administradores de sitios Web que agregan palabras clave intencionalmente a fin de que el sitio sea considerado más relevante que otros al armar el *ranking*. También consiste en el envío múltiple de la misma página con diferentes URLs para que la posibilidad de ser contabilizada para el proceso de R.I. sea mayor.
15. ***Clustering* de documentos**: Agrupamiento de documentos similares en un mismo *cluster*. Esta funcionalidad se encuentra relacionada con similitud de documentos[9]. Ejemplo: *Vivísimo* [VV05].
16. ***Clustering* de sitios**: Agrupamiento de páginas recuperadas de acuerdo a si pertenecen o no a un mismo sitio. Algunos de los formatos posibles: unificación en un solo enlace o indentación. Ejemplos: *Excite* y *Google*, respectivamente.
17. **Escalable**: Factibilidad de mantener un buen comportamiento y performance con respecto al crecimiento de la Web. Aspectos a considerar: actualización de los índices, porcentaje de la Web que es capaz de indexar y frecuencia de *crawling*, entre otros.

[9] Ver Apéndice IV – Recuperación de la información.

18. **Envío de páginas para actualización del índice**: Los usuarios que administran un sitio Web pueden enviar nuevas páginas o actualizaciones de las mismas para ser consideradas por el *crawler*. Esta funcionalidad puede ser ejecutada periódicamente y de manera automática, buscando nuevos sitios o actualizaciones.

En esta sección se detallaron los atributos que caracterizan a los motores de búsqueda. Estos fueron analizados desde las perspectivas de usuario y técnica. Por lo tanto, el paso siguiente es determinar cuáles son las principales funcionalidades que se desean evaluar y cómo se las puede medir. Para ello se aplica el método *Goal Question Metric* (GQM), el cual fue explicado en el capítulo anterior.

5.6 Aplicación de GQM para la Derivación de Métricas

A continuación, se aplica el método GQM a fin de obtener métricas para las funcionalidades de las secciones 5.5.1 y 5.5.2. La aplicación de este método se realiza de acuerdo al formato de Bobrowski, Marré y Yankelevich [BMY99].

Las primeras funcionalidades a las cuales se aplica GQM son aquellas que se consideran como las más destacadas desde el punto de vista del usuario. En el capítulo siguiente se explica el motivo de esta elección, dado que dichos atributos forman parte de lo que se denomina el conjunto mínimo de funcionalidades de usuario.

OBJETIVO	PREGUNTA	MÉTRICA	TÉCNICA
O1 **Objeto**: Conjunto de buscadores **Propósito**: Evaluar **Funcionalidad**: Exactitud **Perspectiva**: Usuario **Entorno**: - Conjunto fijo de buscadores - Una consulta	**P1** ¿Cuántos documentos recuperados son relevantes?	**M1** Porcentaje de documentos relevantes recuperados.	Número de documentos relevantes recuperados*100/ Número de documentos recuperados
	P2 ¿El conjunto de documentos recuperados tiene el nivel de exactitud necesaria?	**M2** Niveles de Exactitud: 1. La consulta SÓLO retorna los documentos deseados. 2. La consulta retorna los documentos deseados en las primeras posiciones de los resultados. 3. La consulta retorna el o los documentos que el usuario desea. 4. La consulta retorna documentos que son aceptables para el usuario, si bien no son exactamente lo que estaba buscando. 5. La consulta retorna en su mayoría documentos que contienen los términos ingresados pero que no satisfacen la necesidad del usuario. 6. La consulta no retorna documentos • Es preferible que el motor no retorne documentos si no cuenta con ninguno relacionado con la búsqueda a que retorne. 7. La consulta no retorna documentos relacionados con la búsqueda.	Experimento
O2 **Objeto**: Conjunto de buscadores **Propósito**: Evaluar **Funcionalidad**: Completitud **Perspectiva**: Usuario **Entorno**: - Conjunto fijo de buscadores - Una consulta	**P3** ¿El conjunto de documentos relevantes para la consulta es recuperado en su totalidad?	**M3** Porcentaje de documentos relevantes existentes que fueron recuperados.	Número de documentos relevantes recuperados*100/ Número de documentos relevantes existentes
	P4 ¿La completitud de los documentos recuperados son todos los existente en el universo de búsqueda?	**M4** Niveles de Completitud Dada una consulta: 1. Los documentos devueltos son todos los existentes en el universo de búsqueda para esa consulta. 2. Los documentos devueltos son todos los existentes en la base de datos del motor de búsqueda para la misma. 3. Los documentos devueltos no son todos los existentes en la base de datos del motor de búsqueda para esa consulta determinada. 4. Los documentos devueltos contienen toda la información que el usuario requiere para la consulta. 5. Los documentos devueltos poseen menos información que la deseada por el usuario. 6. Los documentos devueltos no poseen información relacionada con la búsqueda. 7. La búsqueda no retorna documentos.	Investigación académica

O3 **Objeto**: Conjunto de buscadores **Propósito**: Evaluar **Funcionalidad**: Facilidad de realización de la búsqueda **Perspectiva**: Usuario **Entorno**: - Conjunto fijo de buscadores - Una consulta	**P5** ¿El tiempo que le insume a un usuario pensar y confeccionar la consulta adecuada para lo que desea buscar es aceptable?	**M5** Cantidad de tiempo necesario para confeccionar la consulta adecuada para la búsqueda a realizar / Cantidad de tiempo que involucra realizar todas las consultas hasta obtener la definitiva.	$\sum_{1}^{ncr} tcci$ tcci: tiempo de confección de la iésima consulta
	P6 ¿Comprende el usuario como realizar la consulta deseada teniendo en cuenta la sintaxis de motor de búsqueda?	**M6** Niveles de Facilidad de Realización de la Búsqueda Dada una consulta determinada: 1. El usuario ingresa la consulta correctamente la primera vez que lo intenta. a. "Correctamente" significa respetar la gramática del lenguaje de consulta al expresar mediante el mismo lo que se desea buscar. 2. El usuario debe ingresar la consulta entre dos y cinco veces hasta hacerlo correctamente. 3. El usuario no logra ingresar la consulta correctamente luego de cinco intentos.	Experimento
O4 **Objeto**: Conjunto de buscadores **Propósito**: Evaluar **Funcionalidad**: Identificación eficiente de documentos relevantes **Perspectiva**: Usuario **Entorno**: - Conjunto fijo de buscadores	**P7** ¿El tiempo que le insume a un usuario encontrar entre las respuestas el primer documento relevante para la consulta es aceptable?	**M7** Cantidad de tiempo que insumió la búsqueda entre las respuestas hasta encontrar el primer documento relevante para el usuario / Cantidad de tiempo que involucra el recorrer todas las respuestas	Tiempo en segundos para encontrar el primer documento relevante/ $\sum_{1}^{n} ti$ ti: tiempo necesario para analizar cada respuesta
	P8 ¿El tiempo que le insume a un usuario encontrar entre las respuestas todos los documentos relevantes para la consulta es aceptable?	**M8** Cantidad de tiempo que insumió la búsqueda entre las respuestas hasta encontrar todos los documentos relevantes para el usuario	Tiempo en segundos para encontrar todos los documentos relevantes

- Una consulta - Conjunto de respuestas a la consulta	**P9** ¿Encuentra el usuario los documentos relevantes a partir del resumen devuelto por el motor de búsqueda?	**M9** Niveles de Identificación eficiente de documentos relevantes ☑ El usuario encuentra los documentos relevantes sin tener que recorrer más de 1 página de resultados. ☑ Los documentos son devueltos en orden de relevancia para la consulta. o Esto implica que el usuario estará satisfecho con el orden en el que aparecen los documentos recuperados. ☑ El usuario comprende el resumen devuelto por el motor de búsqueda. **Nota:** Para esta funcionalidad nos parece adecuado enumerar diversos factores y considerar que a mayor cantidad de características tildadas, se corresponde un mayor nivel de calidad para esta funcionalidad. Esto es debido a que existen diversos factores, no necesariamente relacionados, que afectan la identificación eficiente de documentos relevantes. Por ende los niveles de calidad son los siguientes: 1. El motor de búsqueda cumple con las tres propiedades enumeradas. 2. El motor de búsqueda cumple con dos de las tres propiedades enumeradas. 3. El motor de búsqueda cumple con una de las tres propiedades enumeradas. 4. El motor de búsqueda no cumple con ninguna de las propiedades enumeradas.	Experimento
O5 **Objeto**: Conjunto de buscadores **Propósito**: Evaluar **Funcionalidad**: Disponibilidad **Perspectiva**: Usuario **Entorno**: - Conjunto fijo de buscadores - Una consulta - Conjunto de respuestas a la consulta	**P10** ¿Los documentos recuperados son accedidos adecuadamente?	**M10** Promedio de documentos accedidos exitosamente.	# documentos accesibles / # enlaces devueltos para la consulta
		M11 Promedio de documentos que no se han podido acceder o que son demasiado lentos.	# documentos no exitosos / # enlaces devueltos para la consulta
		M12 Niveles de Disponibilidad 1. Todos los documentos devueltos se encuentran disponibles. 2. Todos los documentos relevantes retornados se encuentran disponibles. 3. El buscador dispone de un resumen (o caché) para los documentos relevantes no disponibles. 4. Algunos documentos relevantes devueltos no están disponibles ni en caché. 5. Ningún documento relevante devuelto está disponible. 6. Ningún documento relevante devuelto está disponible ni en caché.	Experimento

O6 **Objeto**: Conjunto de buscadores **Propósito**: Evaluar **Funcionalidad**: Confiabilidad **Perspectiva**: Usuario **Entorno**: - Conjunto fijo de buscadores	**P11** ¿Cree el usuario que los documentos que recupera el motor de búsqueda son aceptables desde su punto de vista?	**M13** Niveles de Confiabilidad 1. El usuario utiliza siempre el mismo motor de búsqueda. 2. El usuario utiliza siempre más de un motor de búsqueda para una misma consulta. 3. El usuario no usa siempre el mismo motor de búsqueda. 4. El usuario no utiliza un motor de búsqueda en particular. 5. El usuario no utiliza motores de búsqueda para realizar consultas sobre sitios de Internet. a. Libros sobre sitios de Internet por temas (al estilo de una guía telefónica). b. Sólo acude a sitios conocidos. c. Accede a nuevos sitios por referencia. d. No utiliza Internet para buscar información o no lo utiliza del todo.	Experimento

Tabla 5.2 – GQM – Funcionalidades de usuario.

El segundo conjunto de funcionalidades al cual se aplica GQM consiste en aquellas que se consideran como las más destacadas desde el punto de vista técnico. En el capítulo siguiente se explica el motivo de esta elección, ya que forman parte del conjunto mínimo de funcionalidades técnicas.

OBJETIVO	PREGUNTA	MÉTRICA	TÉCNICA
O7 **Objeto**: Conjunto de buscadores **Propósito**: Evaluar **Funcionalidad**: Alcance **Perspectiva**: Usuario **Entorno**: - Conjunto fijo de buscadores - Usuario	**P12** ¿Cuál es el porcentaje de los documentos de la Web que indexa el buscador en determinado momento?	**M14** Niveles de Alcance 1. El motor de búsqueda indexa todos los documentos existentes en la Web en todo momento. 2. El motor de búsqueda indexa todos los documentos existentes en la Web en un momento determinado. 3. El motor de búsqueda indexa todos los documentos existentes en la Web para una consulta determinada. 4. El motor de búsqueda no indexa todos los documentos relevantes existentes para una consulta determinada. 5. El motor de búsqueda no indexa los 10 documentos más relevantes para una consulta determinada 6. El motor de búsqueda no indexa los documentos relevantes para una consulta determinada.	Investigación académica utilizando motores de búsqueda
O8 **Objeto**: Conjunto de buscadores **Propósito**: Evaluar **Funcionalidad**: Utilización de perfiles y contextos **Perspectiva**: Usuario **Entorno**: - Conjunto fijo de buscadores - Usuario	**P13** ¿El motor de búsqueda utiliza información del usuario al realizar las consultas?	**M15** Niveles de Utilización de perfiles y contextos ☑ El motor utiliza perfiles y contextos para realizar la búsqueda [EU01]. ☑ El usuario puede configurar su propio perfil mediante las funcionalidades del motor. ☑ El motor utiliza perfiles de usuario automáticos [EU01]. **Nota**: Para esta funcionalidad nos parece adecuado enumerar diversos factores y considerar que a mayor cantidad de características tildadas, se corresponde un mayor nivel de calidad para esta funcionalidad. Esto es debido a que existen diversos factores, no necesariamente relacionados, que afectan la utilización de perfiles y contextos. Por ende los niveles de calidad son los siguientes: 1. El motor de búsqueda cumple con las tres propiedades enumeradas. 2. El motor de búsqueda cumple con dos de las tres propiedades enumeradas. 3. El motor de búsqueda cumple con una de las tres propiedades enumeradas 4. El motor de búsqueda no cumple con ninguna de las propiedades enumeradas.	Investigación académica

O9 **Objeto**: Conjunto de buscadores **Propósito**: Evaluar **Funcionalidad**: *Relevance Feedback* / Capacidad de Aprendizaje **Perspectiva**: Usuario **Entorno**: - Conjunto fijo de buscadores - Conjunto de consultas	**P14** ¿Existe alguna mejora en la búsqueda?	**M16** Niveles de *Relevance Feedback* / Capacidad de Aprendizaje 1. El usuario advierte una mejora al utilizar un motor que aprovecha esta técnica. 2. El motor mejora la búsqueda utilizando *relevance feedback*. 3. El motor de búsqueda pide información al usuario para poder aprender sobre lo que el usuario quiere y por lo tanto mejorar la búsqueda. 4. El motor de búsqueda no utiliza *relevance feedback* ni solicita información al usuario para mejorar los resultados de las consultas.	Investigación académica
O10 **Objeto**: Conjunto de buscadores **Propósito**: Evaluar **Funcionalidad**: *Clustering* de documentos **Perspectiva**: Usuario **Entorno**: - Conjunto fijo de buscadores - Usuario	**P15** ¿Cuál es el tamaño promedio de un cluster de documentos?	**M17** Promedio de tamaño en memoria de todos los clusters de una consulta	$\sum_{i=0}^{n} \frac{tamaño(i)}{n}$ n: cantidad de clusters
	P16 ¿Cuál es la cantidad de clusters de documentos devueltos para una consulta?	**M18** Promedio de la cantidad de clusters devueltos	$\sum_{i=0}^{n} \frac{cantclusters(i)}{n}$ n: consultas realizadas
	P17 ¿El motor de búsqueda utiliza *clustering* de documentos?	**M19** Niveles de *Clustering* de documentos 1. El motor de búsqueda utiliza *clustering* de documentos para retornar los resultados. 2. El motor de búsqueda utiliza *clustering* de documentos para indexarlos y almacenarlos. 3. El motor de búsqueda no utiliza *clustering* de documentos.	Investigación académica
O11 **Objeto**: Conjunto de buscadores **Propósito**: Evaluar **Funcionalidad**: Caché **Perspectiva**: Usuario **Entorno**: - Conjunto fijo de buscadores - Usuario	**P18** ¿Se utiliza un caché para los documentos no activos?	**M20** Niveles de *Caché* 1. El buscador dispone de un resumen (o caché) para los documentos 2. El buscador dispone de un resumen (o caché) para los documentos no disponibles. 3. El buscador no dispone de un resumen (o caché) para los documentos.	Investigación académica

Tabla 5.3 – GQM – Funcionalidades técnicas.

Como se mencionó anteriormente, en el capítulo siguiente se definen las funcionalidades más destacadas y se especifica el experimento ejecutado para medirlas utilizando un subconjunto de las métricas derivadas en las tablas precedentes.

CAPITULO 6. DISEÑO DEL EXPERIMENTO

"The principle of science, the definition, almost, is the following:
The test of all knowledge is experiment.
Experiment is the sole judge of scientific "truth"."

Richard Feynman (The Feynman Lectures on Physics, Vol. 1)

6.1 Introducción

Tal como se indicó en los capítulos anteriores, el propósito de este trabajo es evaluar la calidad de los motores de búsqueda principalmente desde la perspectiva del usuario, sin olvidarse del punto de vista técnico. Por este motivo, en el capítulo 5 – Derivación de Métricas, se seleccionaron las funcionalidades de los motores de búsqueda y un conjunto de métricas para las funcionalidades más destacadas. La evaluación de estas métricas permite determinar la calidad de los atributos de los motores de búsqueda y, de esta manera, compararlos entre sí.

El objetivo de este capítulo, por lo tanto, es definir las funcionalidades más destacadas para esta investigación, teniendo en cuenta las dos perspectivas estudiadas. Se define asimismo qué significa que un buscador sea de calidad y se especifica el experimento que se ejecutó para evaluar el conjunto de funcionalidades mínimas desde la perspectiva del usuario.

En el capítulo anterior se describieron los motores de búsqueda a través de un conjunto de atributos o funcionalidades. Dado que el conjunto definido resulta muy amplio, se trabaja en este capítulo con el conjunto mínimo de funcionalidades destacadas que, según el criterio de este trabajo, hacen a la calidad de un motor de búsqueda. Se definen a continuación dos conjuntos de funcionalidades, uno desde el enfoque del usuario y el otro desde el punto de vista técnico.

6.2 Definición del Conjunto Mínimo de Funcionalidades de Usuario

El conjunto mínimo de funcionalidades de calidad de un motor de búsqueda desde la perspectiva del usuario es el siguiente:

1. **Exactitud**: Los documentos recuperados ante una consulta son relevantes para el usuario.
2. **Completitud**: El motor de búsqueda recupera una fracción significativa de los documentos relevantes existentes en la Web para una consulta determinada.
3. **Facilidad de realización de la búsqueda:** La interfaz del motor de búsqueda simplifica la tarea del usuario en cuanto a indicarle al buscador lo que desea buscar.
4. **Identificación eficiente de documentos relevantes**: Los documentos relevantes pueden encontrarse rápidamente entre los devueltos por un buscador.

5. **Disponibilidad**: Los documentos recuperados para una consulta son accedidos sin inconvenientes.
6. **Confiabilidad**: Los documentos recuperados por el buscador son tomados como verdaderos (i.e., válidos) por el usuario.

6.2.1 Justificación de la Elección del Conjunto Mínimo de Funcionalidades

A continuación, se explica el motivo de la selección de las funcionalidades anteriores como el conjunto mínimo de atributos que un motor de búsqueda de calidad debe poseer desde la perspectiva del usuario.

Las funcionalidades **exactitud y completitud** abarcan las métricas conocidas de recuperación de la información, i.e., precisión y recuperación. Estas funcionalidades no pueden ser excluidas porque indican si la recuperación de los documentos es exacta y completa.

Se sabe que hoy en día no se puede evaluar la completitud en la práctica, ya que no es posible establecer cuáles son todos los elementos del conjunto de documentos relevantes para una consulta dada. Sin embargo, el hecho de no poder evaluarla empíricamente no significa que **completitud** no se deba tener en cuenta como una funcionalidad sumamente importante.

La funcionalidad **identificación eficiente de documentos relevantes** forma parte de los atributos intrínsecos de un motor de búsqueda, ya que no es eficiente que el buscador devuelva los documentos que el usuario precisa pero que éste no los pueda encontrar eficazmente en la colección de respuestas.

La **facilidad de realización de la búsqueda** es una característica interesante para todo motor de búsqueda, ya que mientras más clara sea para el usuario la forma correcta de ingresar una consulta, el motor podrá resolverla mejor y, de esta forma, obtener resultados más exactos.

Disponibilidad forma parte del conjunto mínimo debido a que desde la perspectiva de este trabajo es esencial poder acceder a los documentos devueltos por el motor de búsqueda.

Si bien la **confiabilidad** no afecta a los documentos devueltos por el motor de búsqueda, ésta está relacionada con el motivo por el cual los usuarios eligen y utilizan un buscador determinado. Esto implica que el usuario la considera parte de su definición de calidad. Es por ello que se la incluye en el conjunto de funcionalidades mínimas.

6.3 Definición del Conjunto Mínimo de Funcionalidades Técnicas

El conjunto mínimo de funcionalidades de calidad de un motor de búsqueda desde el punto de vista técnico es el siguiente:

1. **Alcance**: Indexación de toda la Web en todo momento.
2. **Utilización de Perfiles y Contextos**: El motor de búsqueda considera el perfil del usuario y el contexto en el cual realiza sus consultas como parte integral del criterio para determinar los resultados.
3. ***Relevance Feedback*/Capacidad de Aprendizaje**: El buscador aprende de las búsquedas realizadas exitosamente y de su interacción con el usuario.
4. ***Clustering* de documentos**: Agrupamiento de documentos similares en un mismo *cluster*.
5. **Caché**: Almacenamiento de páginas Web en la "memoria caché del buscador" como una copia de seguridad, con el fin de recuperarlas para los usuarios en caso de que el servidor de la página falle temporalmente o que se desee acceder rápidamente al contenido de los sitios.

6.3.1 Justificación de la Elección del Conjunto Mínimo de Funcionalidades

A continuación, se explica el motivo de la selección de las funcionalidades anteriores como el conjunto mínimo de atributos que un motor de búsqueda de calidad debe poseer desde la perspectiva técnica.

La funcionalidad **alcance** es elemental y no puede quedar fuera del conjunto mínimo técnico ya que ayuda a determinar el porcentaje de páginas de la Web indexado por el motor de búsqueda.

Si bien **utilización de perfiles y contextos** no es una de las funcionalidades más exploradas, su utilización contribuye al mejoramiento de la relevancia de los documentos recuperados para una consulta.

Por otra parte, ***relevance feedback*/capacidad de aprendizaje** es un atributo necesario en un motor de búsqueda debido a que su utilización permite mejorar la performance de búsquedas futuras y evitar cometer los mismos errores.

El ***clustering* de documentos** permite mostrar los resultados en base a una clasificación determinada y, por lo tanto, facilita la búsqueda de documentos relevantes en el resumen de resultados.

Finalmente, la funcionalidad **caché** facilita el acceso a los documentos inaccesibles en el momento de una consulta.

Dado que la finalidad de este trabajo es medir la calidad de los buscadores desde las dos visiones estudiadas, es indispensable definir qué significa que un motor de búsqueda sea de calidad. Esta definición constituye la base del experimento realizado.

6.4 Definición de Calidad de un Motor de Búsqueda

6.4.1 La Perspectiva del Usuario

Dada la siguiente **propiedad de calidad** desde el punto de vista del usuario:

Propiedad 6.1: Un buscador es de calidad desde el punto de vista del usuario si y sólo si posee el conjunto mínimo de las funcionalidades de usuario y estas funcionalidades cumplen con los niveles especificados a continuación:

Funcionalidad	Nivel	Definición del Nivel[10]
Exactitud	2	La consulta retorna los documentos deseados en las primeras posiciones de los resultados.
Completitud	4	Los documentos devueltos contienen toda la información que el usuario requiere para la consulta.
Facilidad de realización de la búsqueda	1	El usuario ingresa la consulta correctamente la primera vez que lo intenta.
Identificación eficiente de documentos relevantes	1	El usuario encuentra los documentos relevantes sin tener que recorrer más de una página de resultados. Los documentos son devueltos en orden de relevancia para la consulta. El usuario comprende el resumen devuelto por el motor de búsqueda.
Disponibilidad	2	Todos los documentos relevantes retornados se encuentran disponibles.
Confiabilidad	1	El usuario utiliza siempre el mismo motor de búsqueda.

Tabla 6.1 - Definición de niveles de calidad - Funcionalidades de usuario.

El nivel seleccionado para indicar que cada funcionalidad es de calidad es el mínimo requerido desde el punto de vista de este trabajo para satisfacer dicha funcionalidad desde la perspectiva del usuario.

6.4.2 Perspectiva Técnica

Dada la siguiente **propiedad de calidad** desde el punto de vista técnico:

Propiedad 6.2: Un buscador es de calidad desde el punto de vista técnico si y sólo si posee el conjunto mínimo de las funcionalidades técnicas y estas funcionalidades cumplen con los niveles especificados a continuación:

[10] Las definiciones fueron obtenidas del capítulo 5 – Derivación de métricas.

Funcionalidad	Nivel	Definición del Nivel[11]
Alcance	2	El motor de búsqueda indexa todos los documentos existentes en la Web en un momento determinado.
Utilización de perfiles y contextos	1	El motor utiliza perfiles y contextos para realizar la búsqueda [EU01]. El usuario puede configurar su propio perfil mediante las funcionalidades del motor. El motor utiliza perfiles de usuario automáticos.
Relevance feedback/ Capacidad de aprendizaje	3	El motor de búsqueda pide información al usuario para poder aprender sobre lo que el usuario quiere y, por lo tanto, mejorar la búsqueda.
Clustering de documentos	1	El motor de búsqueda utiliza *clustering* de documentos para retornar los resultados.
Caché	2	El buscador dispone de un resumen (o caché) para los documentos no disponibles.

Tabla 6.2 - Definición de niveles de calidad - Funcionalidades técnicas.

El nivel seleccionado para indicar que cada funcionalidad es de calidad es el mínimo requerido desde el punto de vista de este trabajo para satisfacer dicha funcionalidad desde la perspectiva técnica.

6.5 Factores que Influyen en la Calidad de los Motores de Búsqueda

Antes de definir el objetivo del experimento es necesario enumerar los factores fundamentales que influyen en la calidad de los motores de búsqueda, ya que éstos están relacionados con la hipótesis a probar y con las funcionalidades definidas en el capítulo 5 – Derivación de Métricas.

Generalmente, las siguientes características de los M.B. influyen en los resultados obtenidos:

1. Las búsquedas deben ser refinadas más de una vez para obtener el resultado deseado.
2. La cantidad de documentos relevantes devueltos es menor que la de documentos irrelevantes. Es decir, se debe invertir mucho tiempo buscando los primeros entre la colección devuelta.
3. La interfaz no es suficientemente amigable y no se aprovecha el perfil de cada usuario, o lo que se utiliza no alcanza para obtener una búsqueda mejorada [EU01].
4. Las búsquedas avanzadas resultan muy complejas para usuarios que no dominan el lenguaje booleano.
5. Debido al estado del arte de los buscadores, no es posible recuperar todos los documentos existentes en la Web que son relevantes para la búsqueda.

[11] Las definiciones fueron obtenidas del capítulo 5 – Derivación de métricas.

6. Los usuarios suelen utilizar motores de búsqueda no adecuados para ciertas consultas por hábito. En ciertos casos, es más conveniente utilizar buscadores especializados en el tema de la consulta en particular para de este modo alcanzar resultados óptimos o por lo menos más convenientes.
7. Los resultados devueltos no siempre son accesibles o no están actualizados.

En base a lo expresado, se define en la siguiente sección el objetivo del experimento.

6.6 Objetivo

El experimento tiene como objetivo probar la siguiente hipótesis:

Hipótesis: Los siguientes motores de búsqueda no cumplen la propiedad 6.1:

1. Google (www.google.com)
2. AllTheWeb (www.alltheweb.com)
3. MSN (search.msn.com)
4. Altavista (www.altavista.com)
5. About.com (www.about.com)

6.6.1 Justificación de la Elección del Conjunto de Motores de Búsqueda

1. **Google**: Se eligió este motor ya que según *Search Engine Watch* [SEW05], este buscador tiene una reputación muy buena y es el más utilizado entre los usuarios que buscan información en la Web. Además, es el buscador que indexa más páginas Web y la relevancia de los documentos que recupera es elevada.
2. **AllTheWeb**: De acuerdo a *Search Engine Watch*, este buscador es actualmente el segundo en cantidad de páginas indexadas. A pesar de ello, AllTheWeb no pertenece al conjunto de motores de búsqueda más usados. Por lo tanto, se seleccionó para el experimento porque tiene un gran alcance en la Web pero no es muy utilizado.
3. **MSN:** Este buscador ha crecido mucho recientemente tanto en popularidad como en indexación de la Web.
4. **Altavista**: Altavista nació en 1995 y por varios años fue el "Google de hoy en día", ya que proveía resultados relativamente relevantes en base al estado del arte de los buscadores de la época y tenía un grupo de usuarios muy fieles. Lamentablemente, la compañía le dio más importancia a otras iniciativas y, como consecuencia, su relevancia fue disminuyendo al igual que el alcance de su *crawler*. Últimamente, Altavista ha vuelto a enfocar sus estrategias corporativas en R.I.W. [SEW05]. El motivo de la elección es su rol histórico y el haber sido considerado número uno durante los comienzos de R.I.W.
5. **About.com**: La elección de este motor de búsqueda se debe a que no es conocido y no aparece en ninguno de los rankings publicados en *Search Engine Watch*.

Las estimaciones del tamaño del índice y la popularidad de cada uno de los buscadores elegidos para el experimento son un buen complemento para la justificación anterior. Estos se encuentran en el capítulo 3 – Recuperación de Información de la Web.

6.7 Objeto de Estudio

Como se mencionó anteriormente, la entidad a evaluar es un motor de búsqueda.
Los motores de búsqueda evaluados fueron definidos en la sección 6.6 - Objetivo.

6.8 Marco Muestral

La población estudiada está compuesta por personas de nacionalidad argentina con diferentes intereses y conocimientos que interactuaron con algún motor de búsqueda en el pasado.

6.8.1 Análisis

El experimento se realizó entregando a los usuarios un formulario con una serie de consignas que indicaban las tareas a realizar en el conjunto de motores de búsqueda establecido por la hipótesis.

6.9 Diseño del Experimento

El objetivo del experimento consistió en verificar la hipótesis definida en la sección 6.6. Para lograrlo, se evaluó si los motores de búsqueda seleccionados cumplían con los niveles requeridos para las funcionalidades del conjunto mínimo desde el punto de vista de usuario, pidiéndole a los encuestados que analizaran los 50 primeros documentos devueltos para cada búsqueda.

Para alcanzar la meta planteada, se definió un conjunto de consignas que permitió recolectar los datos necesarios para obtener los niveles para cada funcionalidad definida en el conjunto mínimo de funcionalidades de usuario. El conjunto de preguntas se dividió en tres grupos.

La meta del primer grupo fue la de analizar cómo interactúa el usuario con las interfaces de los motores de búsqueda. De esta forma, se logró evaluar la funcionalidad **facilidad de realización de la búsqueda**.

El segundo conjunto de preguntas evaluó la interacción del usuario con el resumen devuelto por el buscador. Esto permitió calcular los niveles de las funcionalidades **exactitud, identificación eficiente de documentos relevantes** y **disponibilidad**.

El último grupo de preguntas permitió conocer cuál es el nivel de **confiabilidad** de los motores de búsqueda evaluados.

Se debe notar que la funcionalidad **completitud** no se evaluó, ya que para medirla se necesita conocer todo el universo de documentos y en la actualidad no existe ningún buscador que indexe toda la Web.

6.10 Ejecución del Experimento

El experimento se llevó a cabo en los primeros meses de 2004. Participaron veinte personas de distintos perfiles cuyas edades se encuentran entre los quince y los sesenta años.

Los pasos seguidos fueron los siguientes:

1. Se entregó a cada usuario el formulario del experimento.
2. Los usuarios tuvieron permitido realizar cualquier tipo de consulta sobre el formulario.
3. Las mediciones que se realizaron se corresponden con las métricas derivadas en el capítulo 5 – Derivación de Métricas para las funcionalidades pertenecientes al conjunto mínimo de funcionalidades de usuario.
4. Se analizó la información obtenida en los formularios para determinar los niveles de las funcionalidades evaluadas. Esto se hizo mapeando de forma directa una respuesta o combinando un conjunto de respuestas de los encuestados para obtener el nivel en que se halla cada funcionalidad. Este estudio se encuentra en el capítulo 7 – Recolección y Análisis de Datos.
5. Obtención de conclusiones (capítulo 8 - Conclusiones).

6.11 Formulario del Experimento

El formulario entregado a los usuarios se encuentra en el Apéndice I - Formulario del Experimento.

CAPITULO 7. RECOLECCION Y ANALISIS DE DATOS

"Data should be collected with a clear purpose in mind.
Not only a clear purpose but a clear idea
as to precise way in which they will be analysed do as
to yield the desired information."

Moroney, 1950

7.1 Introducción

En este capítulo se analizan los resultados del experimento utilizando las medidas de tendencia central y de dispersión detalladas en el apéndice V. El mismo se estructura de la siguiente forma:

En la primera parte, se muestran los datos obtenidos para las funcionalidades estudiadas e información adicional que ayuda a comprender el comportamiento y necesidades de los usuarios por medio de histogramas. En segundo lugar, se analizan los datos teniendo en cuenta los niveles definidos para cada funcionalidad en el capítulo 5 – Derivación de Métricas. Luego se proponen mejoras para que los buscadores alcancen los niveles óptimos propuestos en este trabajo para las funcionalidades estudiadas. Para finalizar, se exponen las conclusiones generales observadas dentro del marco del experimento realizado.

En el apéndice II – Recolección de Datos, se encuentran los datos recolectados durante el experimento.

7.2 Recolección de Datos

Para comenzar con el estudio, se analiza primero el porcentaje de usuarios que conocía los motores de búsqueda del experimento con anterioridad.

El 100% de los usuarios conocía Google, el 75% conocía Altavista, el 70% conocía MSN y ninguno de los usuarios conocía los motores de búsqueda AllTheWeb o About.

La figura 7.1 ilustra esta relación:

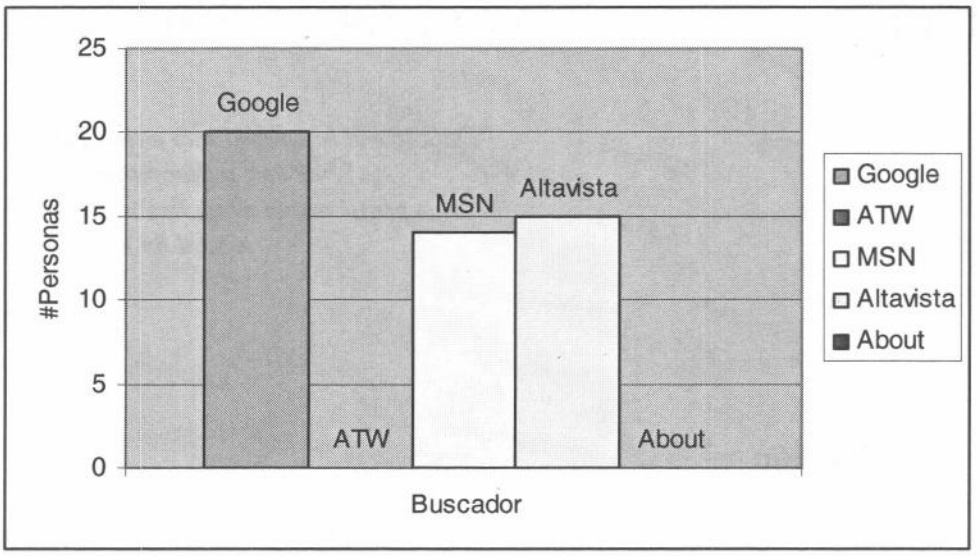


Figura 7.1 – Cantidad de usuarios que ya conocía los motores de búsqueda del experimento.

Los datos obtenidos en este punto se corresponden con lo esperado, ya que si bien AllTheWeb es el segundo motor de búsqueda en cobertura de la Web, este motor no tiene gran repercusión en Argentina. El caso de About también resultó acorde a lo esperado, ya que no es muy conocido en la Argentina.

Tal como lo indican los resultados obtenidos, el resto de los motores de búsqueda evaluados son populares entre los usuarios argentinos.

Los siguientes gráficos ilustran los niveles de las funcionalidades para cada uno de los buscadores. Las definiciones de las funcionalidades y sus niveles fueron introducidos en el capítulo 5 – Derivación de Métricas.

Google

Funcionalidad: Exactitud

La figura 7.2 muestra que la mayoría de los datos recogidos posiciona a Google en el nivel tres de *exactitud*. Por lo tanto, se puede afirmar que, en general, los usuarios que participaron del experimento obtuvieron los documentos deseados.

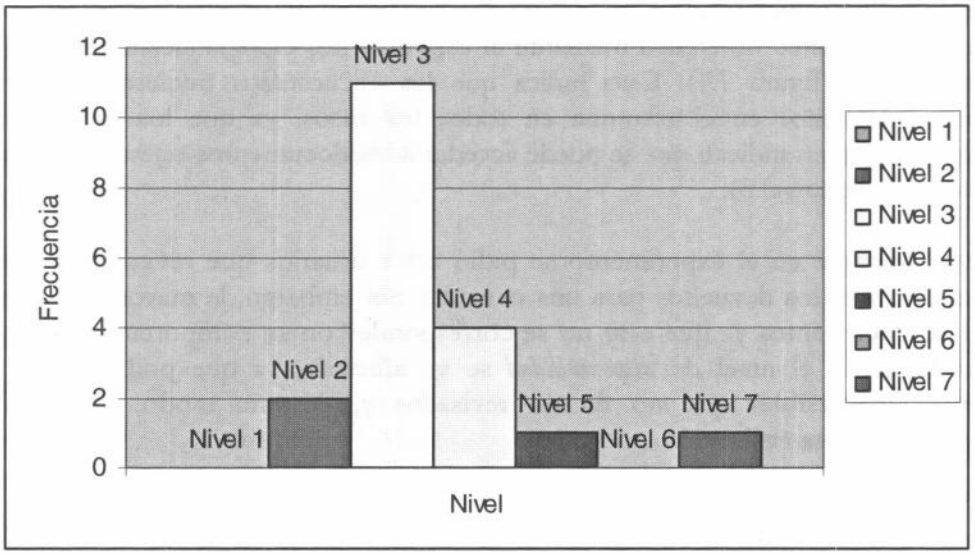


Figura 7.2 – Exactitud de Google.

Funcionalidad: Identificación Eficiente de Documentos Relevantes

La figura 7.3 muestra que la mayoría de los datos recogidos en el experimento ubica a Google en el nivel dos para *identificación eficiente de documentos relevantes*, aunque hay también una gran cantidad de observaciones en los niveles uno y tres. Esto indica que, en general, los usuarios pudieron encontrar los documentos relevantes de forma relativamente eficiente dentro del resumen de resultados.

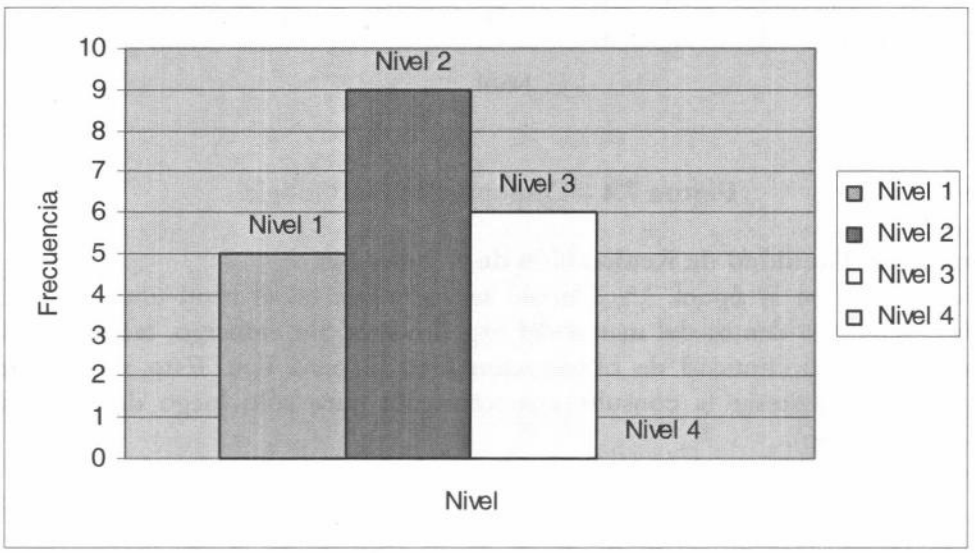


Figura 7.3 – Identificación eficiente de documentos relevantes en Google.

Funcionalidad: Disponibilidad

De acuerdo a los datos obtenidos mediante el experimento, Google alcanza el primer nivel para *disponibilidad* (figura 7.4). Esto indica que los encuestados pudieron acceder a los documentos devueltos en el resumen en todos los casos, ya que los niveles restantes obtenidos (dos y tres) indican que se puede acceder a los documentos relevantes, aunque sea utilizando el caché (nivel 3).

Se debe aclarar que en el experimento se pidió a los usuarios que revisaran los primeros cincuenta documentos devueltos para una consulta. Sin embargo, la mayoría no revisó esta cantidad de documentos ya que esto no se corresponde con su comportamiento en la vida real. Por lo tanto, el nivel de *disponibilidad* se ve afectado, ya que podría haber habido documentos inaccesibles que no fueron revisados y, de este modo, se consideraron "accesibles" para esta evaluación.

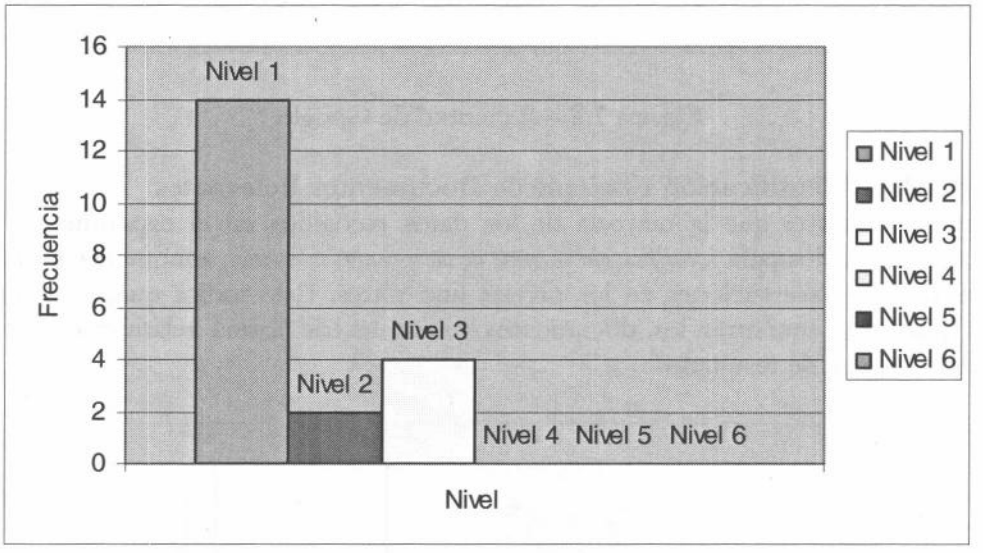


Figura 7.4 – Disponibilidad en Google.

Funcionalidad: Facilidad de Realización de Búsquedas

Como se observa en la figura 7.5, Google se encuentra en el nivel uno para *facilidad de realización de búsquedas* dentro del marco del experimento. Sin embargo, también se advierte claramente una gran cantidad de observaciones en el nivel dos. Esto indica que varios usuarios lograron ingresar la consulta correctamente pero sólo luego de varios intentos (cinco como máximo).

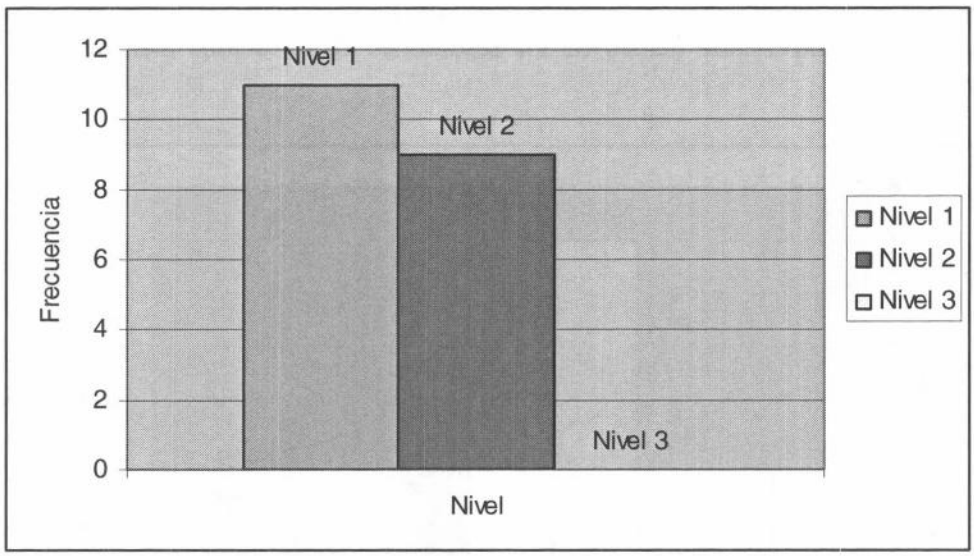


Figura 7.5 – Facilidad de realización de búsquedas en Google.

Funcionalidad: Confiabilidad

La siguiente figura muestra que la mayoría de los datos recogidos mediante el experimento ubica a Google en el nivel uno con respecto a *confiabilidad*. Esto indica que la mayoría de los usuarios confía en Google para realizar sus búsquedas.

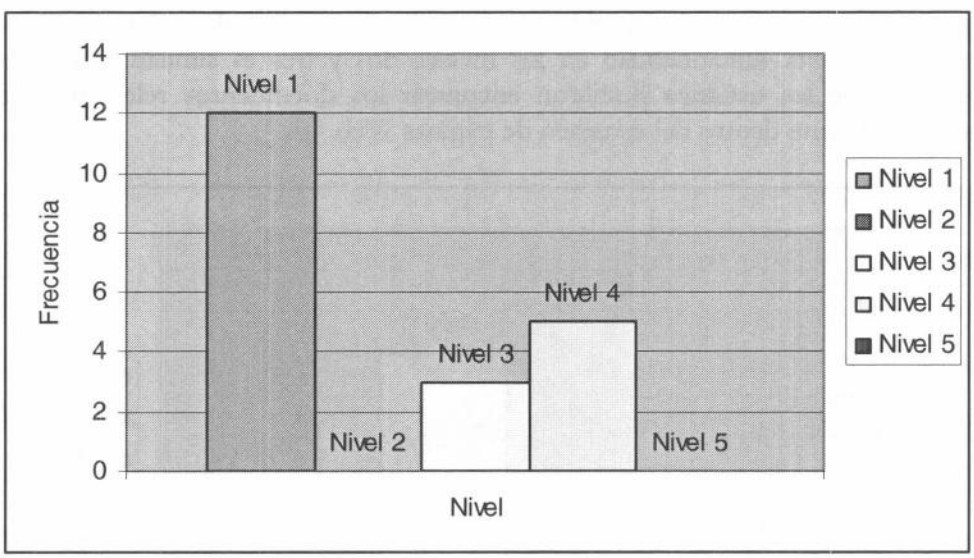


Figura 7.6 – Confiabilidad de Google.

AllTheWeb

Funcionalidad: Exactitud

La figura 7.7 muestra que la mayoría de los datos obtenidos mediante la evaluación posiciona a AllTheWeb en el nivel tres de *exactitud*. Sin embargo, las observaciones están distribuidas a

través de todos los niveles excepto por el primero. Se observa que varios encuestados no encontraron nada interesante dentro de los resúmenes de resultados (i.e., nivel 4, nivel 5, nivel 6 y nivel 7).

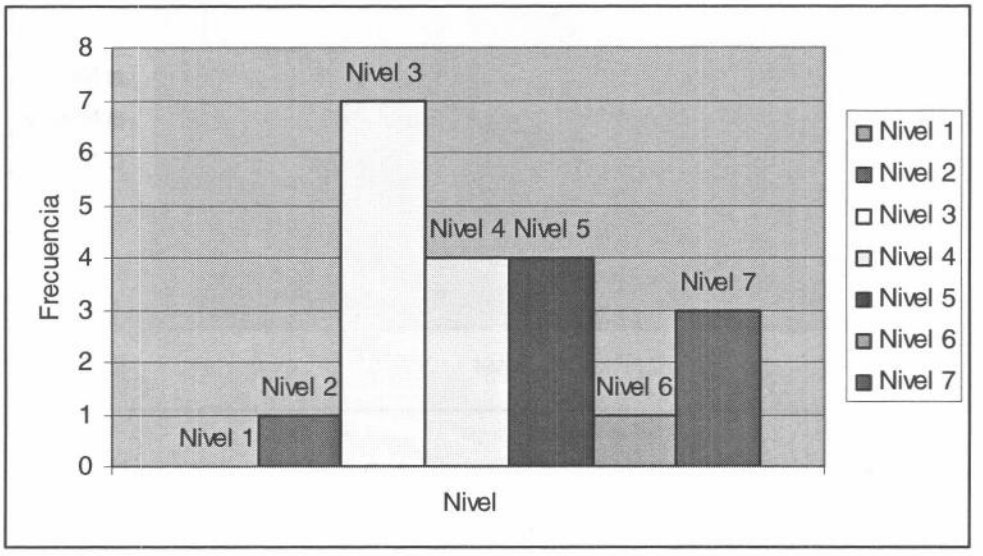


Figura 7.7 – Exactitud de AllTheWeb.

Funcionalidad: Identificación Eficiente de Documentos Relevantes

La figura 7.8 muestra que la cantidad de observaciones recolectadas por medio del experimento para esta funcionalidad en los niveles dos y tres es similar. Por lo tanto, se puede afirmar que los usuarios pudieron encontrar los documentos relevantes de forma relativamente eficiente dentro del resumen de páginas Web.

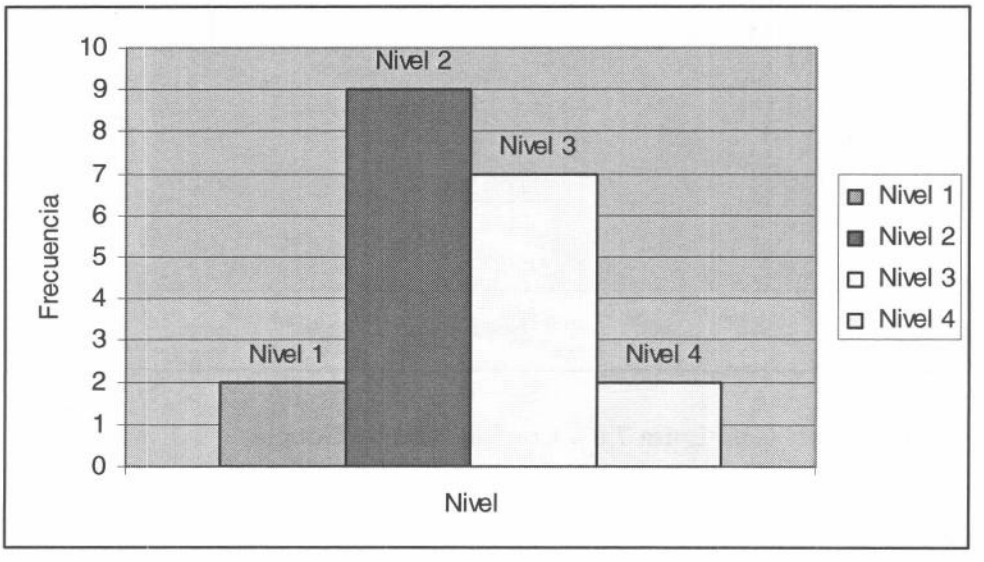


Figura 7.8 – Identificación eficiente de documentos relevantes en AllTheWeb.

Funcionalidad: Disponibilidad

La figura 7.9 muestra que la mayoría de los datos recogidos a través del experimento posiciona a AllTheWeb en el primer nivel de *disponibilidad*. Es decir, la mayoría de los encuestados pudo acceder a los documentos retornados.

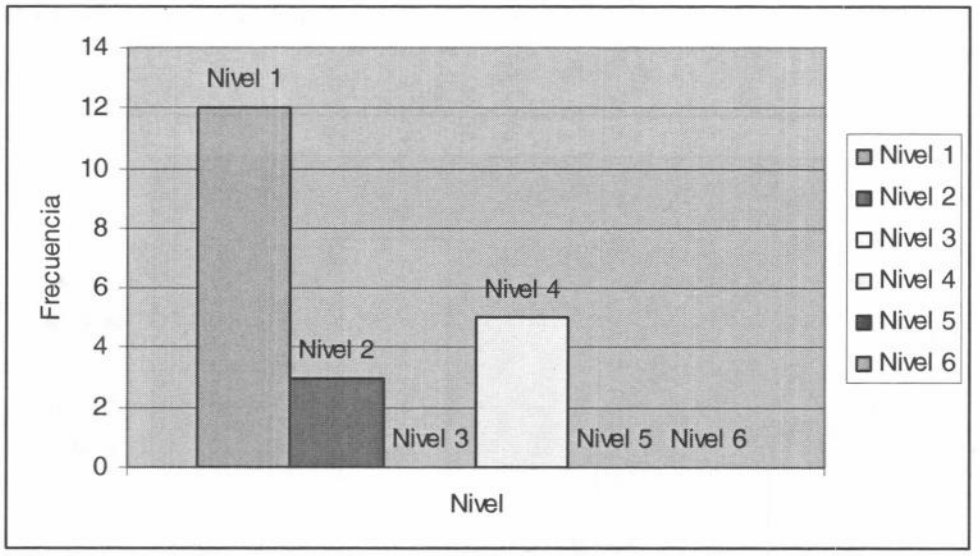


Figura 7.9 – Disponibilidad en AllTheWeb.

Funcionalidad: Facilidad de Realización de Búsquedas

La figura 7.10 muestra que la mayoría de los datos obtenidos a través del experimento determina que AllTheWeb está entre los niveles uno y dos en cuanto a la *facilidad de realización de búsquedas* (i.e., el usuario logró ingresar la consulta correctamente luego de pocos intentos (cinco como máximo)).

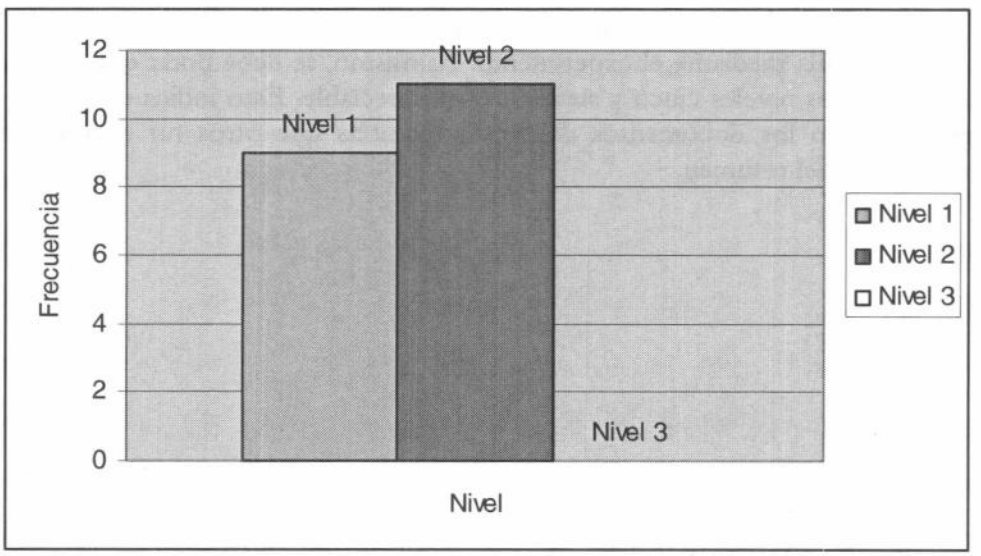


Figura 7.10 – Facilidad de realización de búsquedas en AllTheWeb.

Funcionalidad: Confiabilidad

La figura 7.11 muestra que AllTheWeb se encuentra en el nivel cuatro de *confiabilidad* dentro del marco del experimento. Es decir, ninguno de los entrevistados confía en este buscador. Es interesante destacar este resultado, ya que el nivel de *confiabilidad* es bajo a pesar de que este buscador es el segundo en indexación de la Web. Una de las explicaciones para este resultado es la baja popularidad de AllTheWeb en Argentina, lo cual afecta la confiabilidad del mismo.

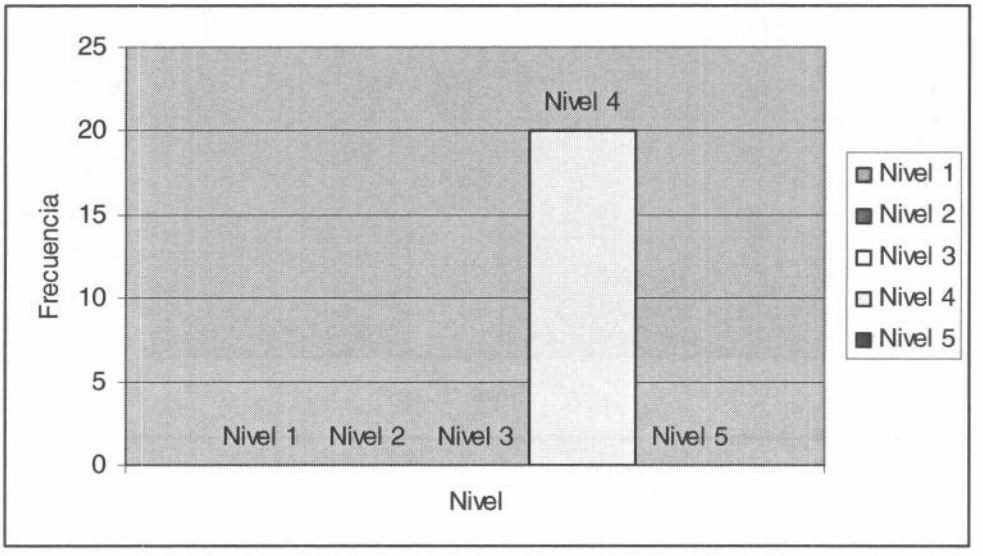


Figura 7.11 – Confiabilidad de AllTheWeb.

MSN

Funcionalidad: Exactitud

La figura 7.12 muestra que MSN se encuentra en el nivel tres de *exactitud* en base a la información obtenida mediante el experimento. Asimismo, se debe notar que la cantidad de observaciones en los niveles cinco y siete no es despreciable. Esto indica que la mayoría de los usuarios obtuvo los documentos deseados, mientras que otros no encontraron nada interesante dentro del resumen.

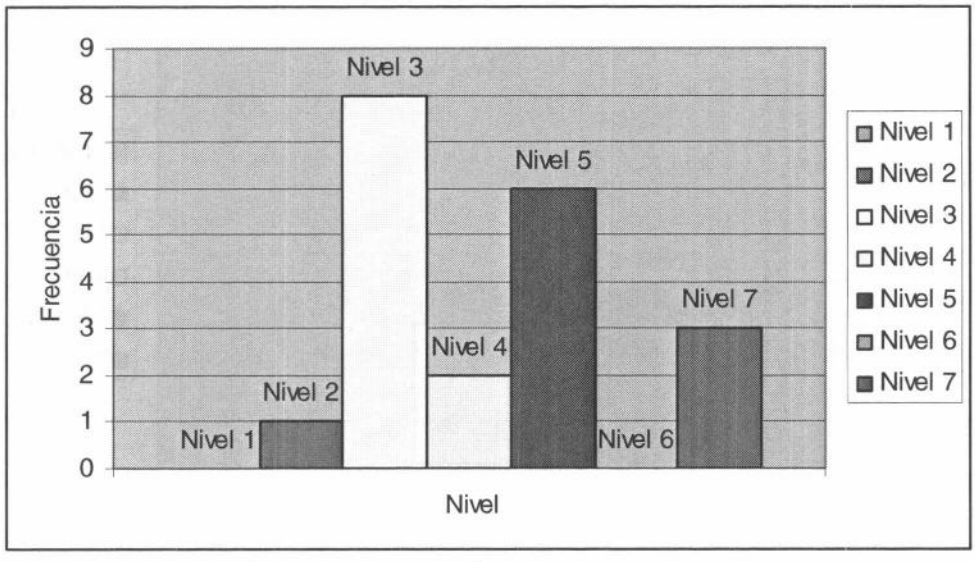


Figura 7.12 – Exactitud de MSN.

Funcionalidad: Identificación Eficiente de Documentos Relevantes

De acuerdo a la figura 7.13, MSN está en el tercer nivel para *identificación eficiente de documentos relevantes* dentro de los datos obtenidos en el experimento. Esto significa que a los usuarios no les resultó fácil encontrar los documentos relevantes dentro de los resultados.

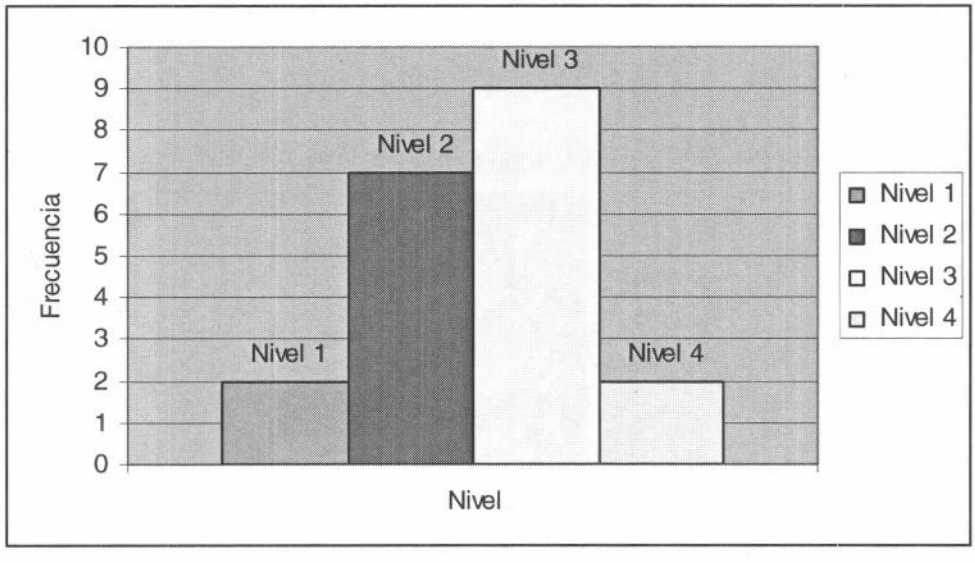


Figura 7.13 – Identificación eficiente de documentos relevantes en MSN.

Funcionalidad: Disponibilidad

MSN se encuentra en el nivel uno en cuanto a *disponibilidad* ya que la mayoría de los usuarios que participó del experimento pudo acceder a los documentos devueltos en el resumen.

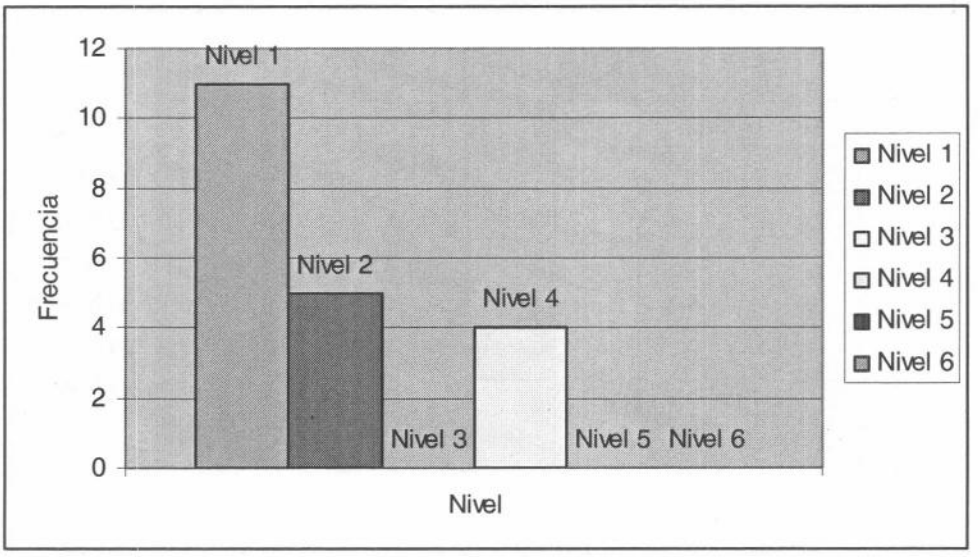


Figura 7.14 – Disponibilidad en MSN.

Funcionalidad: Facilidad de Realización de Búsquedas

La siguiente figura indica que la *facilidad de realización de búsquedas* para MSN está entre los niveles uno y dos. Por lo tanto, se puede afirmar que los encuestados lograron ingresar las consultas correctamente luego de pocos intentos (cinco como máximo).

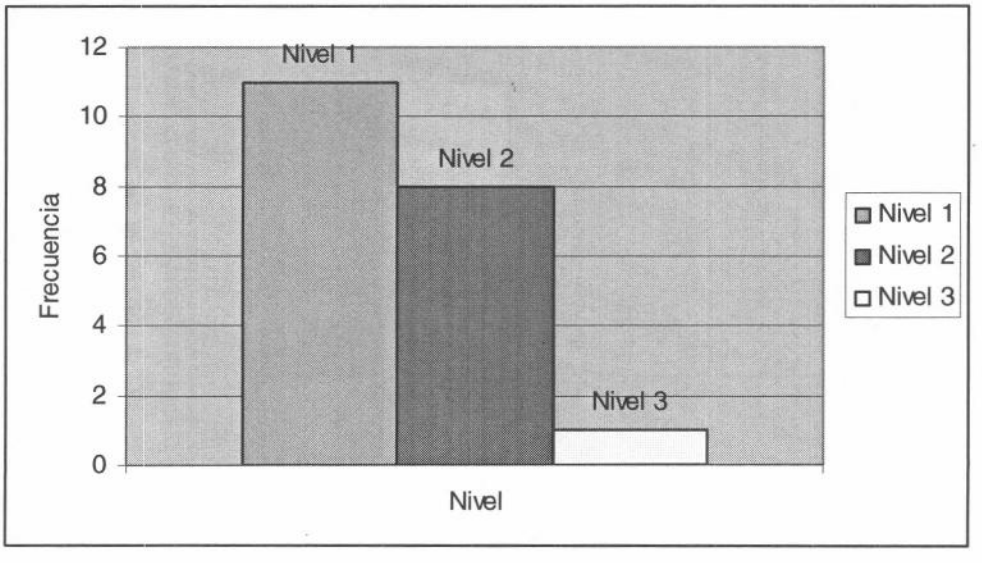


Figura 7.15 – Facilidad de realización de búsquedas en MSN.

Funcionalidad: Confiabilidad

En la siguiente figura se observa que el 95% de los encuestados no confía en MSN para realizar sus búsquedas.

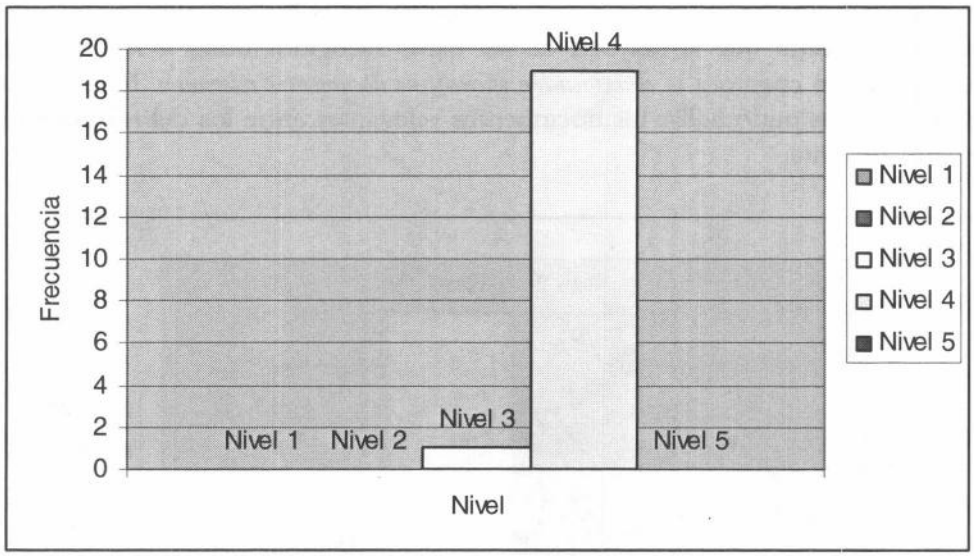


Figura 7.16 – Confiabilidad de MSN.

Altavista

Funcionalidad: Exactitud

La mayoría de los encuestados obtuvo los documentos que deseaba al realizar búsquedas en Altavista. La cantidad de observaciones en el nivel tres es claramente superior al número de observaciones en otros niveles.

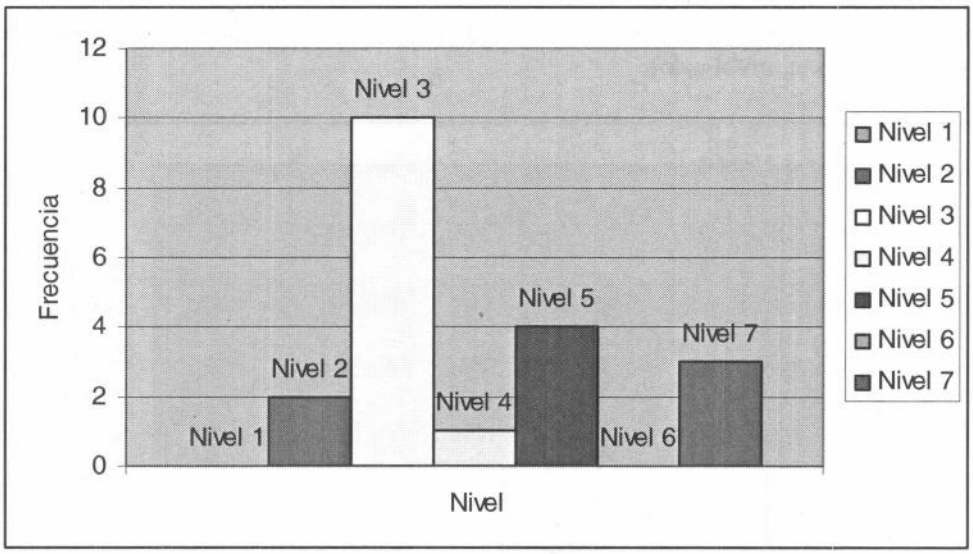


Figura 7.17 – Exactitud de Altavista.

Funcionalidad: Identificación Eficiente de Documentos Relevantes

La figura 7.18 muestra que la mayoría de los datos recogidos ubica a Altavista entre los niveles dos y tres en cuanto a la *identificación eficiente de documentos relevantes*. Es decir, una gran parte de los usuarios pudo hallar los documentos relevantes entre los del resumen de forma relativamente eficiente.

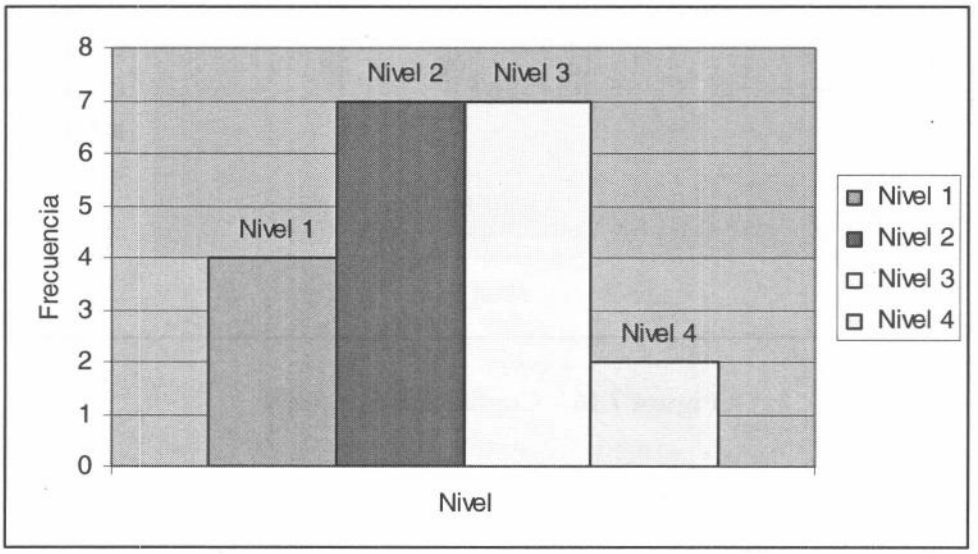


Figura 7.18 – Identificación eficiente de documentos relevantes en Altavista.

Funcionalidad: Disponibilidad

Como fue observado para los motores de búsqueda anteriores, también Altavista provee una alta disponibilidad (i.e., nivel uno):

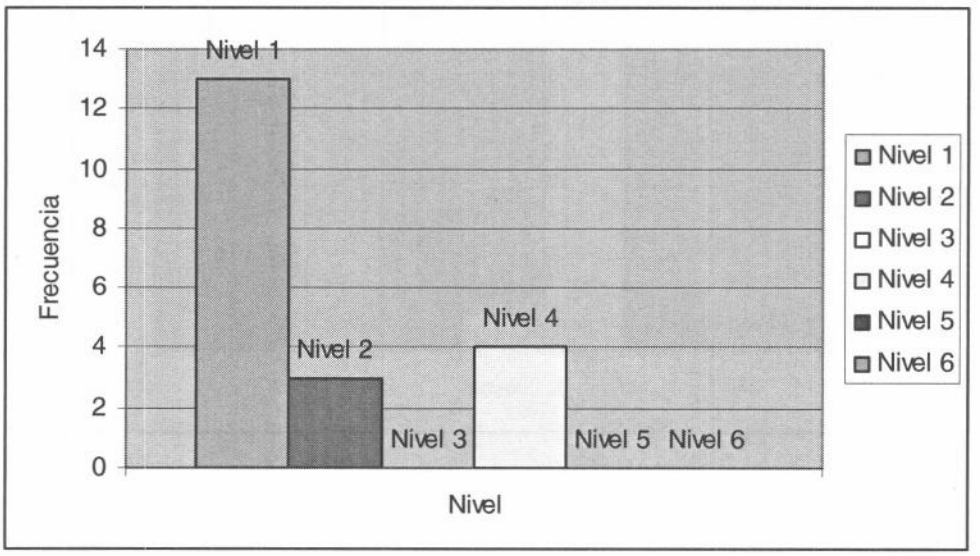


Figura 7.19 – Disponibilidad en Altavista.

Funcionalidad: Facilidad de Realización de Búsquedas

La figura 7.20 muestra que la mayoría de los usuarios que intervino en el experimento pudo realizar búsquedas en Altavista con relativa facilidad, aunque en ciertos casos éstas tuvieron que ser repetidas más de una vez.

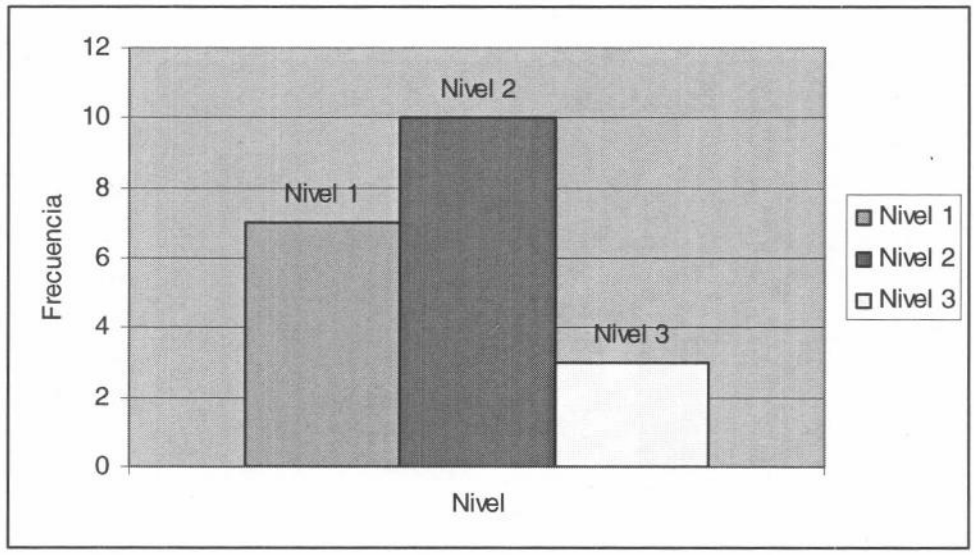


Figura 7.20 – Facilidad de realización de búsquedas en Altavista.

Funcionalidad: Confiabilidad

Como se observa a continuación, el 90% de los encuestados no confía en Altavista para realizar sus búsquedas, a pesar de que este buscador fue uno de los mejores durante muchos años.

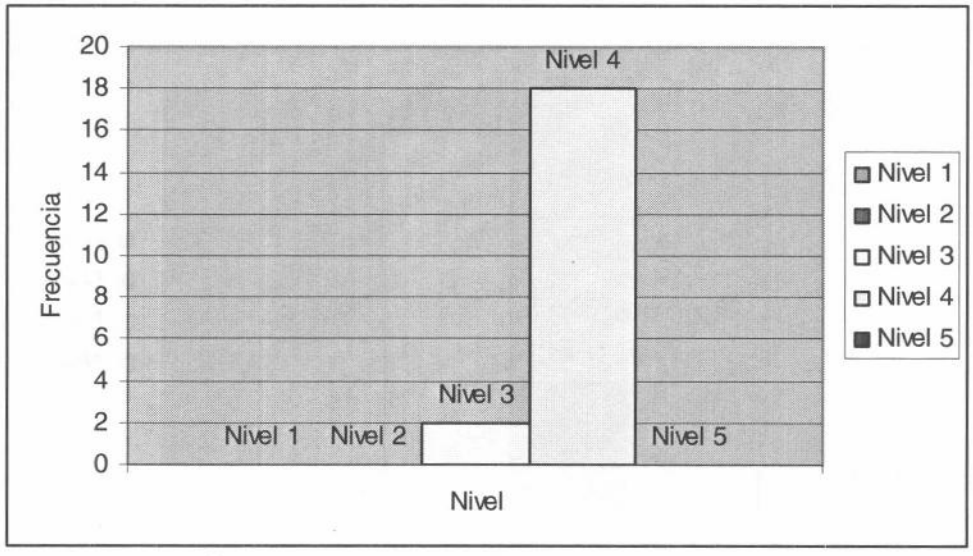


Figura 7.21 – Confiabilidad de Altavista.

About

Funcionalidad: Exactitud

En base al siguiente gráfico se puede determinar que la recuperación de documentos relevantes por parte de About es realmente pobre para la mayoría de los encuestados.

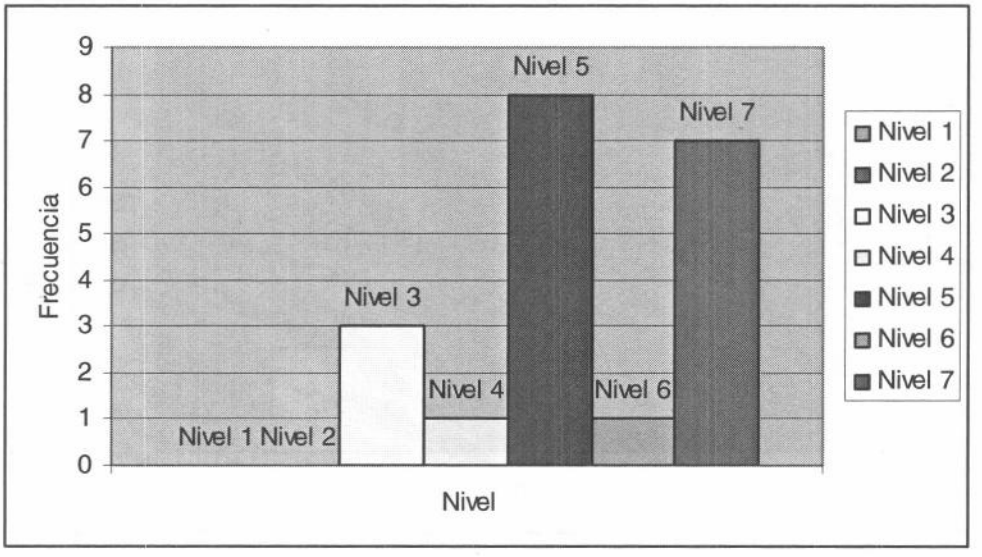


Figura 7.22 – Exactitud de About.

Funcionalidad: Identificación Eficiente de Documentos Relevantes

Condiciéndose con la funcionalidad anterior, la siguiente figura demuestra que los usuarios del experimento tuvieron dificultades para hallar los documentos relevantes dentro del resumen devuelto por este buscador.

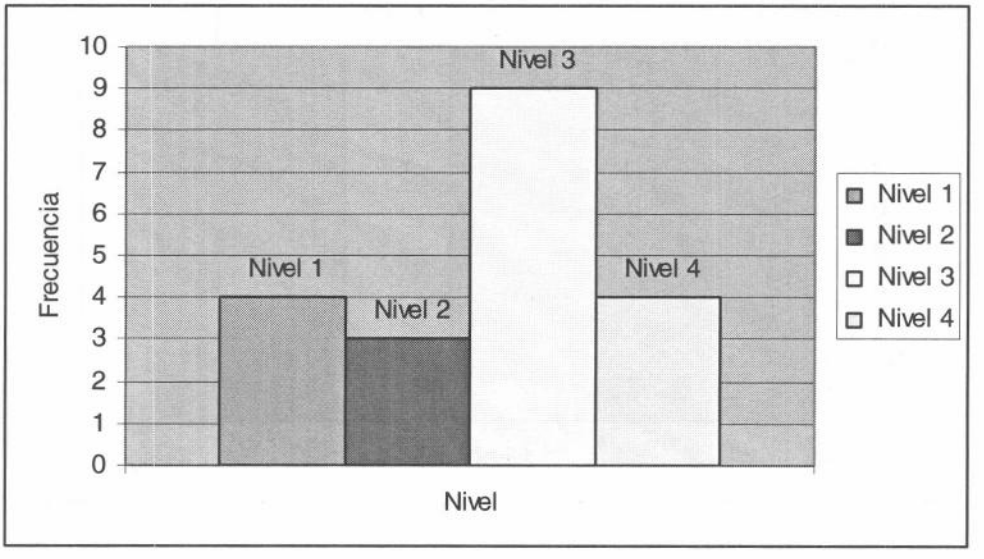


Figura 7.22 – Identificación eficiente de documentos relevantes en About.

Funcionalidad: Disponibilidad

Al igual que todos los buscadores anteriores, la accesibilidad de los documentos retornados por About es de alto nivel.

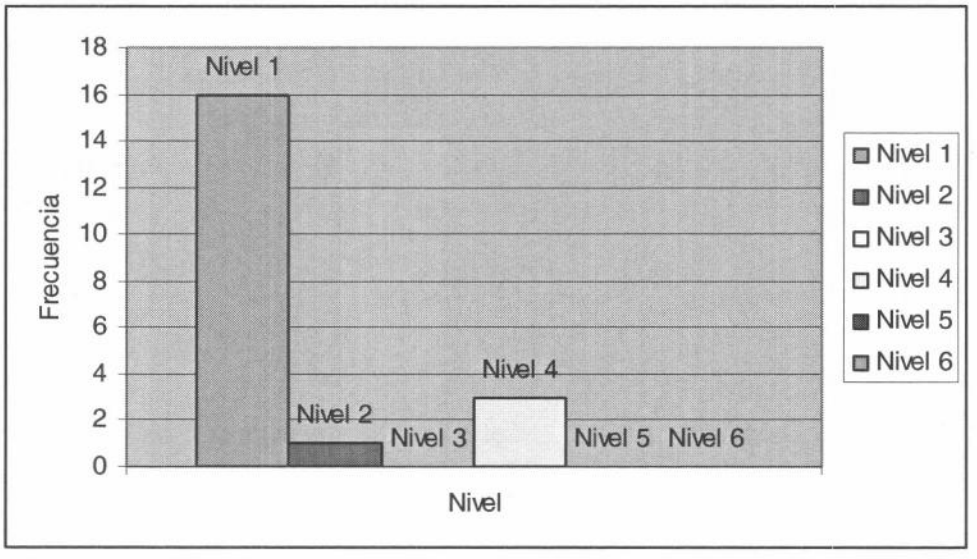


Figura 7.23 – Disponibilidad de About.

Funcionalidad: Facilidad de Realización de Búsquedas

Los usuarios ingresaron sus consultas unas pocas veces para alcanzar resultados satisfactorios, tal como se puede observar en la figura 7.24.

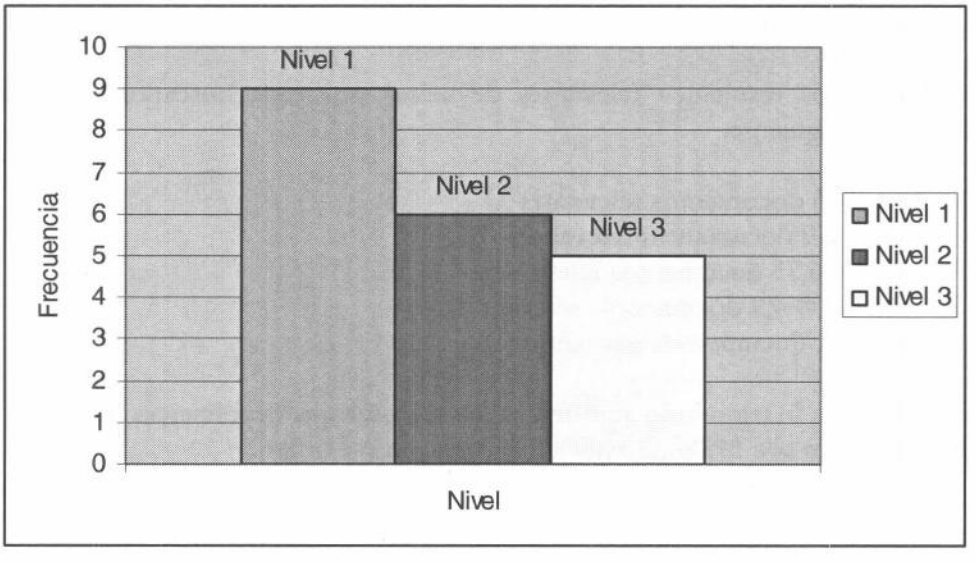


Figura 7.24 – Facilidad de realización de búsquedas en About.

Funcionalidad: Confiabilidad

Como se puede anticipar al evaluar los resultados anteriores, la figura 7.25 muestra que ninguno de los usuarios confía en este buscador.

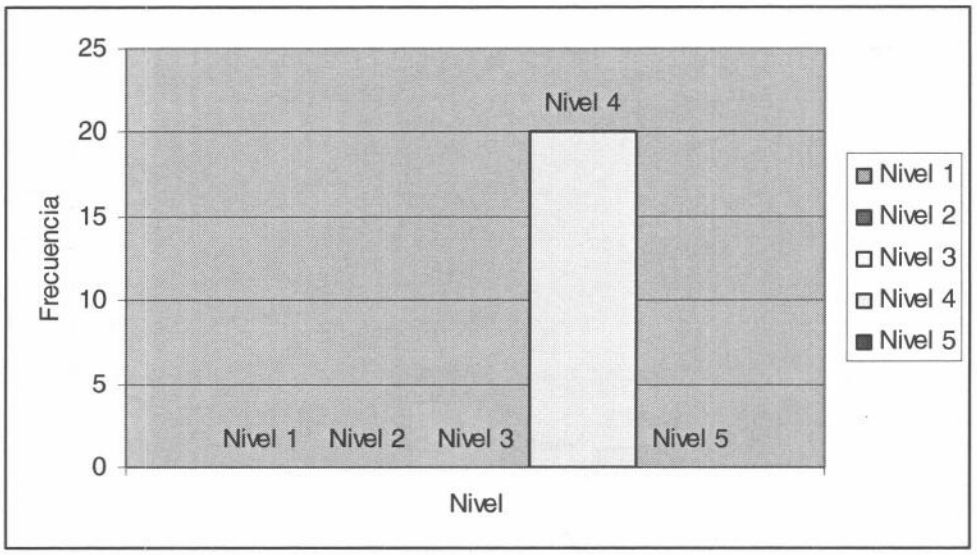


Figura 7.25 – Confiabilidad de About.

Recolección de Datos Adicionales

En esta sección se analizan otros datos que se recolectaron mediante el experimento. Aunque no forman parte de las funcionalidades evaluadas, son interesantes para observar el comportamiento de los usuarios y los buscadores.

El promedio de los resultados relevantes devueltos por cada buscador evaluado en el experimento es el siguiente:

1. MSN = 7,45 documentos relevantes
2. Google = 6,9 documentos relevantes
3. Altavista = 6,75 documentos relevantes
4. AllTheWeb = 4,9 documentos relevantes
5. About = 2,1 documentos relevantes

La figura 7.26 ilustra lo expresado aquí arriba: los motores que devolvieron la mayor cantidad de páginas relevantes son MSN, Google y Altavista, en ese orden.

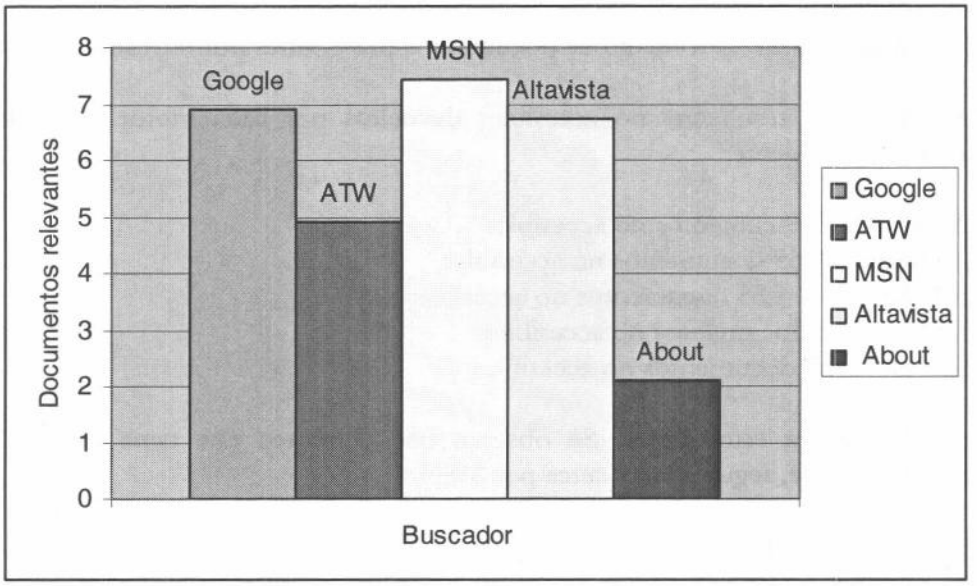


Figura 7.26 – Promedio de documentos relevantes por buscador.

El promedio de los resultados no relevantes devueltos por cada buscador es el siguiente:

1. Google = 42,2 documentos irrelevantes
2. AllTheWeb = 38,75 documentos irrelevantes
3. Altavista = 38 documentos irrelevantes
4. MSN = 35,55 documentos irrelevantes
5. About = 31,35 documentos irrelevantes

La figura 7.27 ilustra estos resultados. Cabe destacar que el hecho de que los resultados anteriores sean muy parecidos para todos los buscadores se debe a que sólo se evaluaron los primeros 50 documentos.

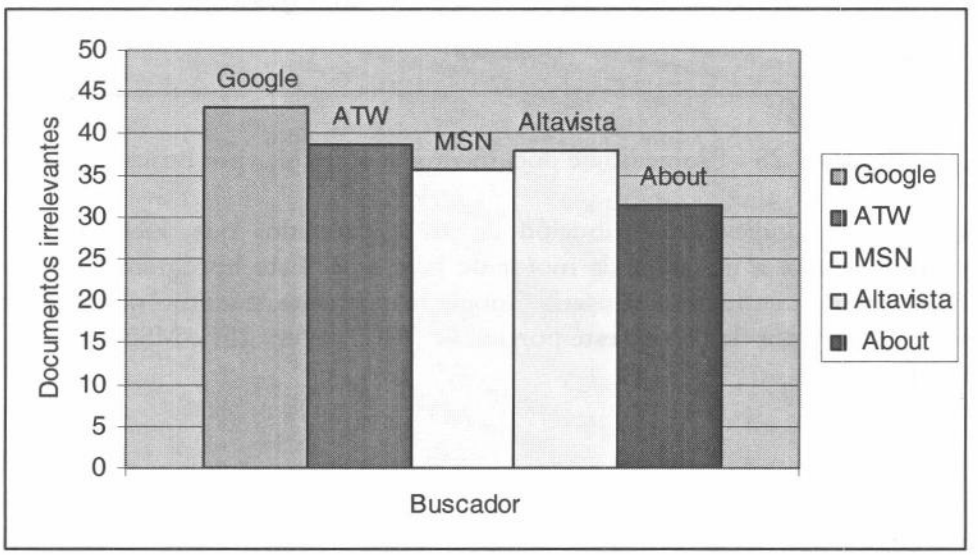


Figura 7.27 - Promedio de documentos irrelevantes por buscador.

El promedio de los resultados no accesibles devueltos por cada motor evaluado en el experimento es el siguiente:

1. About = 2,7 documentos no accesibles
2. Altavista = 1,65 documentos no accesibles
3. AllTheWeb = 1,55 documentos no accesibles
4. MSN = 1,05 documentos no accesibles
5. Google = 0,9 documentos no accesibles

La figura 7.28 ilustra estos datos. Se observa que el motor que tiene menos páginas inaccesibles es Google, seguido muy cerca por MSN.

Se debe notar que la suma de los promedios de los documentos relevantes más los no relevantes más los no accesibles no es 50 en todos los casos debido a que los motores retornaron menos de 50 documentos por búsqueda en algunos casos.

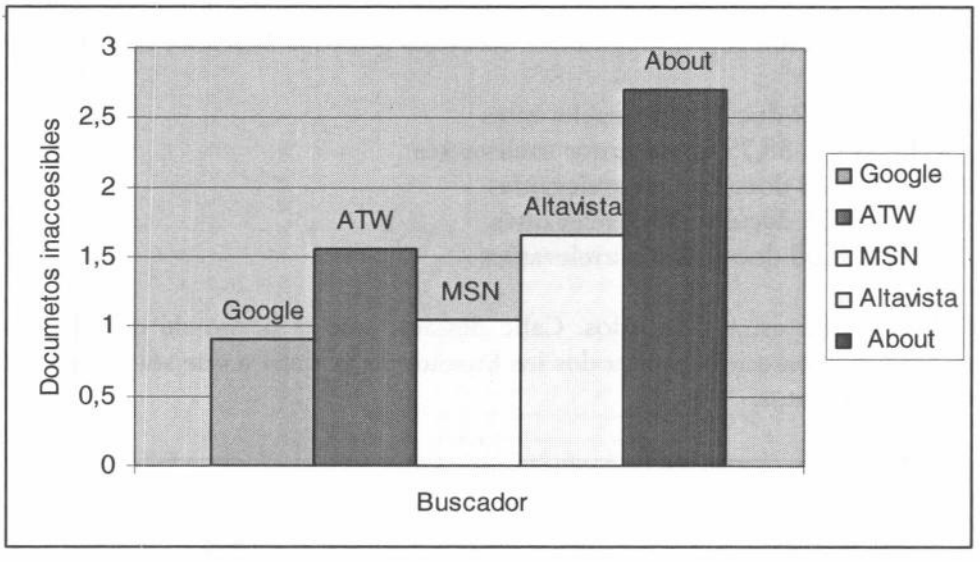


Figura 7.28 - Promedio de documentos inaccesibles por buscador.

El siguiente gráfico indica la distribución de los encuestados que, luego de realizar el experimento, volverían a utilizar cada motor de búsqueda. Este histograma expresa que el 100% de las personas entrevistadas usaría Google nuevamente, mientras que el resto de los buscadores está muy por debajo de este porcentaje (AllTheWeb: 25%, MSN: 45%, Altavista: 45% y About: 10%).

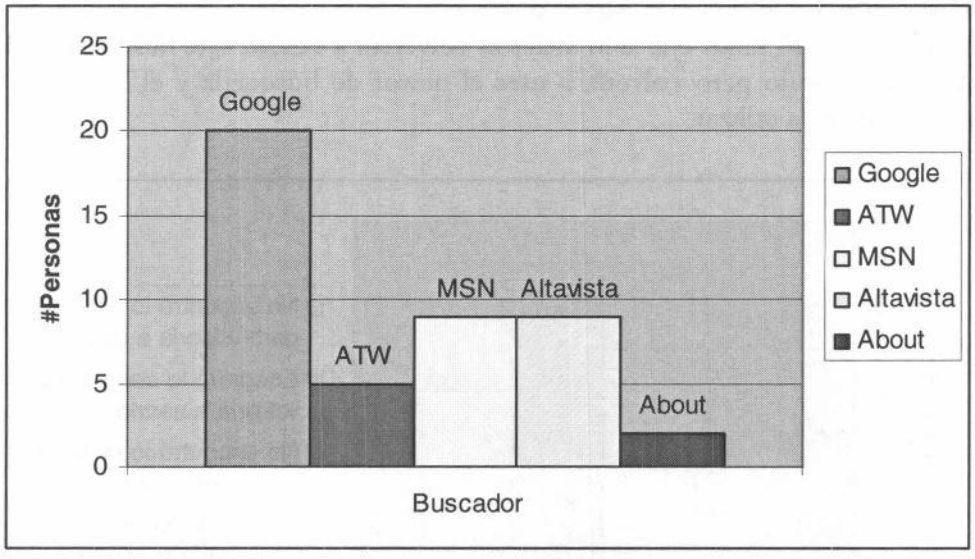


Figura 7.29 – Cantidad de personas que volverían a usar el buscador.

Relación Relevancia – Elección de Buscador/Uso

En esta sección se correlacionan los resultados obtenidos, analizando si los usuarios que obtuvieron los documentos deseados mediante un motor de búsqueda volverían a utilizarlo.

En el siguiente gráfico se ilustra la relación anterior para Google. En este caso, tanto las personas que no encontraron lo deseado como las que sí, volverían a utilizar este motor de búsqueda.

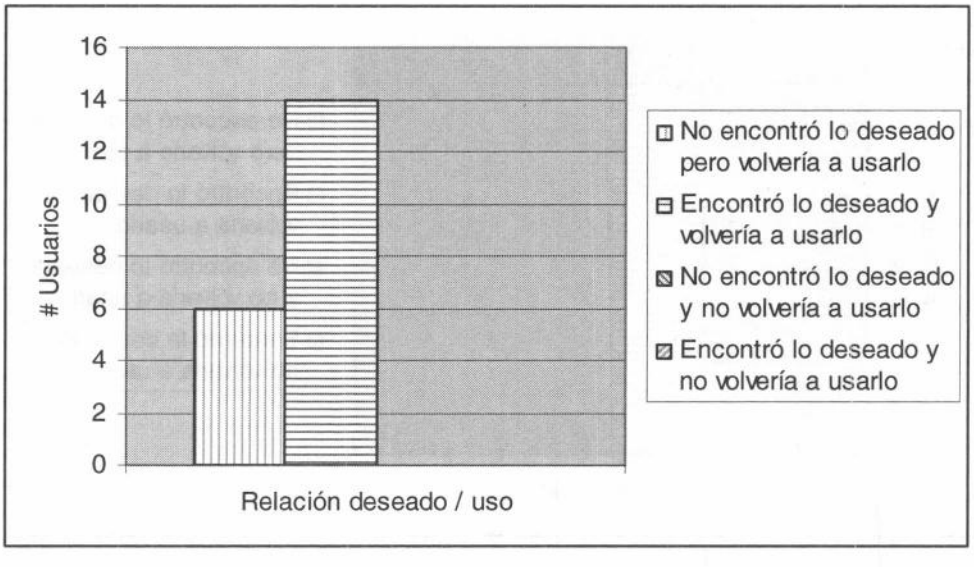


Figura 7.30 – Relación Relevancia – Elección de Buscador/Uso en Google.

Al analizar la misma relación para AllTheWeb se observa lo opuesto: ni los usuarios que no encontraron lo deseado ni los que sí lo hicieron volverían a utilizar este buscador. Sólo el 5% no encontró lo deseado pero volvería a usar el motor de búsqueda y el 20% encontró lo buscado y lo volvería a utilizar.

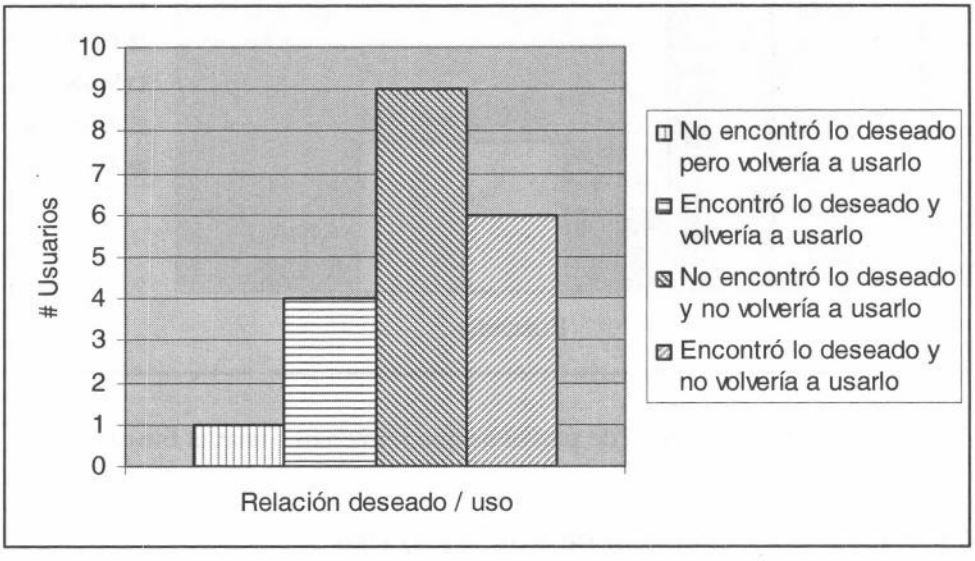


Figura 7.31 - Relación Relevancia – Elección de Buscador/Uso en AllTheWeb.

Con respecto a MSN, la mayoría de las personas que no encontraron lo deseado no volverían a utilizarlo mientras que las que sí encontraron documentos relevantes lo volverían a usar.

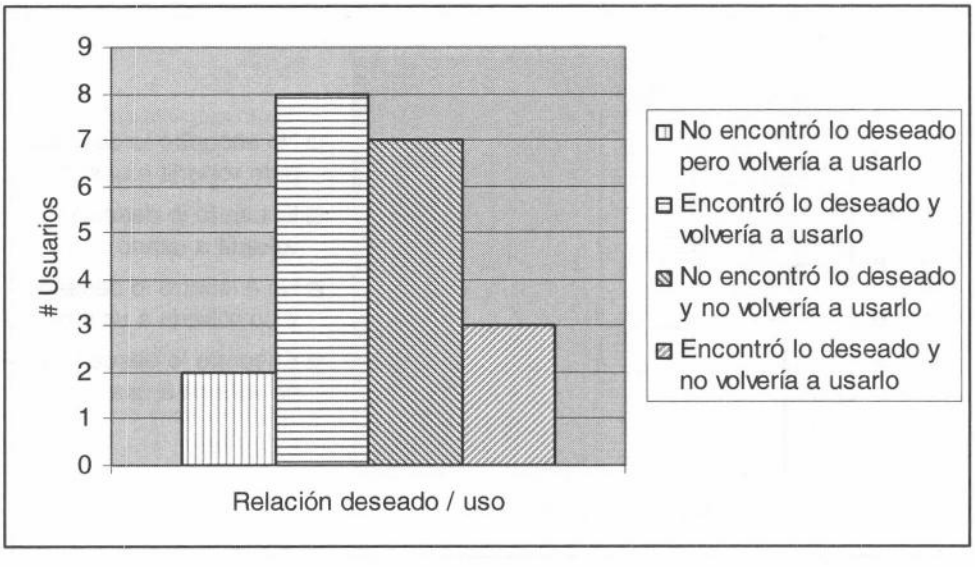


Figura 7.32 - Relación Relevancia – Elección de Buscador/Uso en MSN.

La figura 7.33 muestra que cuatro de cada diez usuarios que obtuvieron lo buscado mediante Altavista lo volverían a utilizar. Por otra parte, el 30% de los que encontraron lo deseado no lo volverían a utilizar.

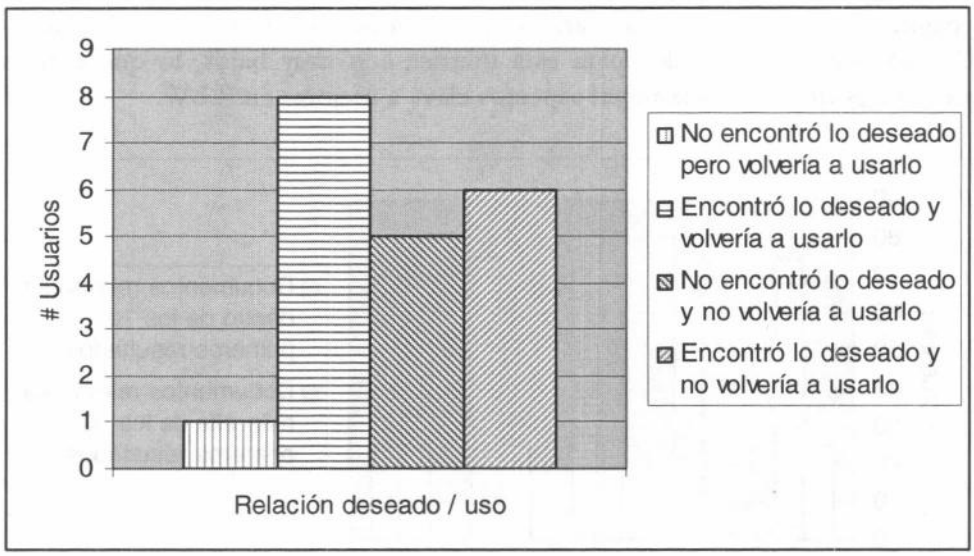


Figura 7.33 - Relación Relevancia – Elección de Buscador/Uso en Altavista.

Finalmente, la mayoría de los individuos que participó del experimento no obtuvo páginas relevantes al utilizar About. Entre ellos, el 70% no lo volvería a utilizar, un porcentaje bajo considerando que una gran fracción de los resultados no fue interesante para los usuarios. Por otra parte, sólo el 10% de los que encontraron lo que buscaban lo volverían a utilizar en el futuro.

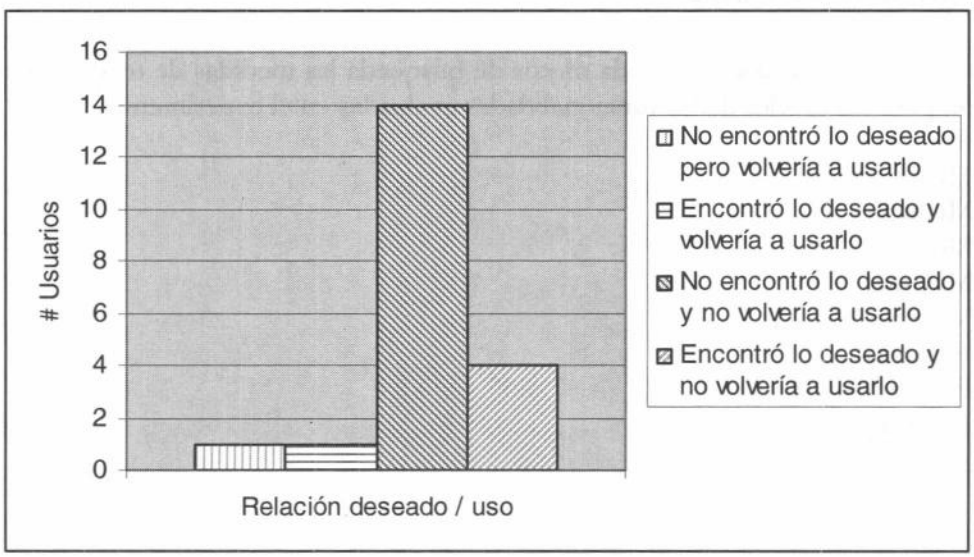


Figura7.34 - Relación Relevancia – Elección de Buscador/Uso en About.

Relevancia

En esta sección se analiza el porcentaje de documentos relevantes dentro de las diez primeras posiciones del ranking, siempre dentro del marco del experimento. Este es un aspecto muy importante, ya que la mayoría de los usuarios espera encontrar lo deseado en la primera página de resultados y muy raramente revisa más de una página. De acuerdo a la figura 7.35, los valores obtenidos para esta métrica son muy bajos, lo que convierte a la relevancia y el algoritmo de ranking en aspectos clave a mejorar en R.I.W.

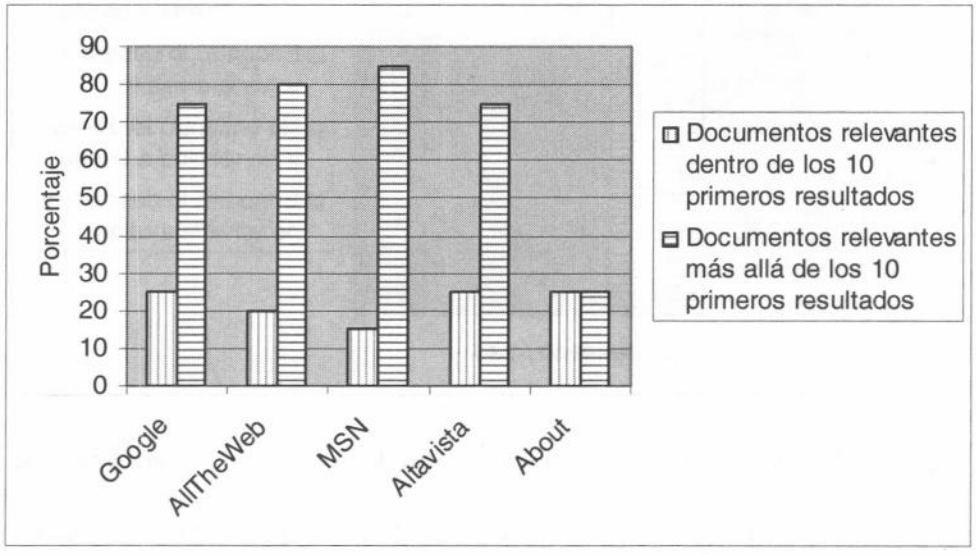


Figura 7.35 - Documentos relevantes por buscador.

7.3 Análisis de Datos

A continuación, se calculan para cada motor de búsqueda las medidas de tendencia central y dispersión para los niveles de las funcionalidades evaluadas en el experimento:

- Media
- Mediana
- Moda
- Máximo
- Mínimo
- Varianza
- Desviación estándar
- MidRange
- Rango

También se calcularon los coeficientes de Pearson y de Fisher, aunque el primero se descartó por las restricciones de la distribución para su utilización[12].

En este análisis se utilizó la mediana para definir el nivel de cada funcionalidad. Pero para poder decir en qué medida de tendencia central había que basarse se utilizó la moda, la media, la varianza, la desviación estándar y el coeficiente de Fisher. Se presentan también los datos adicionales obtenidos durante la evaluación para que el lector saque sus propias conclusiones.

[12] Las medidas de tendencia central y los coeficientes de Pearson y de Fisher se explican en el Apéndice V – Análisis Estadístico.

Google	Media	Mediana	Min	Max	Moda	Varianza	Desviación estándar	Desviación Media	MidRange	Rango	Coeficiente Pearson	Coeficiente de Fisher
Nivel de Exactitud	3,45	3	2	7	3	1,207895	1,099043	0,785	4,5	5	0,37254902	1,029387
Identificación eficiente de documentos relevantes	2,05	2	1	3	2	0,576316	0,759155	0,57	2	2	0,08675799	-0,01716
Nivel de Disponibilidad	1,5	1	1	3	1	0,684211	0,82717	0,7	2	2	0,73076923	0,313548
Nivel de facilidad de realización de búsqueda	1,45	1	1	2	1	0,260526	0,510418	0,495	1,5	1	1,72727273	0,495
Nivel de Confiabilidad	2,05	1	1	4	1	1,839474	1,356272	1,26	2,5	3	0,57081545	0,532186

Tabla 7.1 – Google – Medidas de tendencia central y de dispersión.

AllTheWeb	Media	Mediana	Min	Max	Moda	Varianza	Desviación estándar	Desviación Media	MidRange	Rango	Coeficiente Pearson	Coeficiente de Fisher
Nivel de Exactitud	4,3	4	2	7	3	2,326316	1,525226	1,26	4,5	5	0,55882353	0,832666
Identificación eficiente de documentos relevantes	2,45	2	1	4	2	0,681579	0,825578	0,695	2,5	3	0,66023166	0,044307
Nivel de Disponibilidad	1,9	1	1	4	1	1,673684	1,293709	1,08	2,5	3	0,53773585	0,874768
Nivel de facilidad de realización de búsqueda	1,55	2	1	2	2	0,260526	0,510418	0,495	1,5	1	-1,7272727	-0,495
Nivel de Confiabilidad	4	4	4	4	4	0	0	0	4	0	0	0

Tabla 7.2 – AllTheWeb – Medidas de tendencia central y de dispersión.

MSN	Media	Mediana	Min	Max	Moda	Varianza	Desviación estándar	Desviación Media	MidRange	Rango	Coeficiente Pearson	Coeficiente de Fisher
Nivel de Exactitud	4,25	4	2	7	3	2,302632	1,517442	1,275	4,5	5	0,54285714	0,838417
Identificación eficiente de documentos relevantes	2,55	3	1	4	3	0,681579	0,825578	0,695	2,5	3	-0,6602317	-0,04431
Nivel de Disponibilidad	1,85	1	1	4	1	1,397368	1,182103	0,935	2,5	3	0,60828625	0,789435
Nivel de facilidad de realización de búsqueda	1,5	1	1	3	1	0,368421	0,606977	0,55	2	2	1,35714286	3
Nivel de Confiabilidad	3,95	4	3	4	4	0,05	0,223607	0,095	3,5	1	-1	-0,02804

Tabla 7.3 – MSN – Medidas de tendencia central y de dispersión.

Altavista	Media	Mediana	Min	Max	Moda	Varianza	Desviación estándar	Desviación Media	MidRange	Rango	Coeficiente Pearson	Coeficiente de Fisher
Nivel de Exactitud	3,95	3	2	7	3	2,576316	1,605091	1,34	4,5	5	0,36874362	1,440688
Identificación eficiente de documentos relevantes	2,35	2	1	4	3	0,871053	0,933302	0,785	2,5	3	-0,7462236	0,019449
Nivel de Disponibilidad	1,75	1	1	4	1	1,460526	1,208522	0,975	2,5	3	0,51351351	0,953423
Nivel de facilidad de realización de búsqueda	1,8	2	1	3	2	0,484211	0,695852	0,56	2	2	-0,4130435	1,68
Nivel de Confiabilidad	3,9	4	3	4	4	0,094737	0,307794	0,18	3,5	1	-1,0555556	-0,04653

Tabla 7.4 – Altavista – Medidas de tendencia central y de dispersión.

About	Media	Mediana	Min	Max	Moda	Varianza	Desviación estándar	Desviación Media	MidRange	Rango	Coeficiente Pearson	Coeficiente de Fisher
Nivel de Exactitud	5,4	5	3	7	5	2,042105	1,429022	1,18	5	4	0,19587629	-0,35764
Identificación eficiente de documentos relevantes	2,65	3	1	4	3	1,081579	1,03999	0,855	2,5	3	-0,323601	-0,21201
Nivel de Disponibilidad	1,5	1	1	4	1	1,210526	1,100239	0,8	2,5	3	0,41304348	1,097497
Nivel de facilidad de realización de búsqueda	1,8	2	1	3	1	0,694737	0,833509	0,72	2	2	1,15151515	4,08
Nivel de Confiabilidad	4	4	4	4	4	0	0	0	4	0	0	0

Tabla 7.5 – About – Medidas de tendencia central y de dispersión.

Como se observa en las tablas anteriores, el coeficiente de Fisher es distinto de cero salvo por unas pocas excepciones[13]. Como consecuencia, las distribuciones son asimétricas y, debido a esto, la medida de tendencia central que se utiliza para el análisis es la **mediana**. Esta se evalúa en la siguiente sección.

7.4 Evaluación de la Hipótesis

En esta sección se compara la hipótesis del capítulo 6 con los resultados obtenidos. Las conclusiones que siguen se basan en el experimento realizado y, por lo tanto, son válidas únicamente dentro de ese marco.

A continuación se recuerda la hipótesis expuesta en la sección 6.4:

Hipótesis: Los siguientes motores de búsqueda no cumplen la propiedad 6.1.

Motores de búsqueda

1. Google (www.google.com)
2. AllTheWeb (www.alltheweb.com)
3. MSN (search.msn.com)
4. Altavista (www.altavista.com)
5. About.com (www.about.com)

Propiedad 6.1: Un buscador es de calidad desde el punto de vista del usuario si y sólo si posee el conjunto de las funcionalidades de usuario mínimas y estas funcionalidades cumplen con los niveles especificados a continuación:

Funcionalidad	Nivel	Definición[14]
Exactitud	2	La consulta retorna los documentos deseados en las primeras posiciones de los resultados.
Completitud	4	Los documentos devueltos contienen toda la información que el usuario busca para la consulta.
Facilidad de realización de la búsqueda	1	El usuario ingresa la consulta correctamente la primera vez que lo intenta.
Identificación eficiente de documentos relevantes	1	El usuario encuentra los documentos relevantes sin tener que recorrer más de 1 página de resultados. Los documentos son devueltos en orden de relevancia para la consulta. El usuario comprende el resumen devuelto por el motor de búsqueda.
Disponibilidad	2	Todos los documentos relevantes devueltos se encuentran disponibles.
Confiabilidad	1	El usuario utiliza siempre el mismo motor de

[13] El apéndice V – Análisis Estadístico contiene información adicional al respecto.
[14] Las definiciones fueron obtenidas del capítulo 6 – Derivación de métricas.

		búsqueda (como única opción).

Tabla 6.1 - Definición de niveles de calidad - Funcionalidades de usuario.

En la tabla 7.6 se presentan los valores obtenidos para cada funcionalidad, de los cuales se desprende lo siguiente:

Con respecto a *exactitud*, ninguno de los buscadores alcanzó el nivel deseado (nivel 2). Se observa que tanto Google como Altavista llegaron al tercer nivel. Si bien la mayoría de los usuarios encontró todos los documentos que deseaba, éstos no se encontraban en las primeras posiciones, lo cual es un requisito distintivo para el segundo nivel. El nivel de AllTheWeb y MSN es 4, lo que significa que algunos de los documentos recuperados son aceptables para el usuario aunque no constituyan exactamente lo deseado. Por último, About cumple tan sólo con el nivel cinco. Es decir, la mayoría de los documentos recuperados contiene algunos de los términos buscados pero estos documentos no son relevantes.

En cuanto a *identificación eficiente de documentos relevantes*, nuevamente ninguno de los buscadores alcanzó el nivel deseado (nivel 1). Google, AllTheWeb y Altavista lograron el nivel 2, el cual requiere que la búsqueda en el resumen de los documentos relevantes no sea muy dificultosa para los usuarios, aunque ésta podría ser mejor de cumplir con las tres propiedades pedidas por este nivel. Por otro lado, MSN y About alcanzaron el nivel 3, i.e., la búsqueda en el resumen de los documentos relevantes es un tanto dificultosa para los usuarios y podría ser mejor si cumpliera con las tres propiedades pedidas para este nivel.

Para *disponibilidad* se obtuvo un mayor nivel al esperado para todos los buscadores, lo cual significa que todos los documentos del resumen pudieron ser accedidos por el usuario. Esto es mejor que lo que pide la hipótesis, que es que todos los documentos relevantes puedan ser accedidos por el usuario.

Facilidad de realización de búsquedas alcanza el nivel deseado para Google y MSN (nivel 1). Esto significa que el usuario comprende cómo debe ingresar la consulta y que lo logra en menos de tres intentos. Se puede deducir que el usuario comprende la sintaxis de consulta y que le resulta fácil utilizarla. En cambio, AllTheWeb, Altavista y About alcanzaron el nivel 2, lo que significa que el usuario comprende cómo debe ingresar la consulta y que lo logra en menos de cinco intentos. Se concluye que el usuario comprende básicamente la sintaxis para realizar las consultas al utilizar estos buscadores pero que le resulta un poco dificultoso hacerlo.

Finalmente, el único buscador que obtuvo el nivel deseado para *confiabilidad* (nivel 1) fue Google. Es decir, los usuarios realmente confían en este buscador. El resto de los buscadores obtuvieron un nivel mucho menor (nivel 4), lo que indica que los usuarios no confían en ellos.

Funcionalidad	Valor hipótesis	Google	AllTheWeb	MSN	Altavista	About
Nivel de Exactitud	**2**	3	4	4	3	5
Nivel de Identificación eficiente de documentos relevantes	**1**	2	2	3	2	3
Nivel de Disponibilidad	**2**	1	1	1	1	1
Nivel de Facilidad de realización de búsquedas	**1**	1	2	1	2	2
Nivel de Confiabilidad	**1**	1	4	4	4	4

Tabla 7.6 – Niveles de las funcionalidades.

7.4.1 Comparación de los Motores de Búsqueda por Funcionalidad

Para culminar esta evaluación, se agrupan los motores de búsqueda según el nivel obtenido en el experimento para cada funcionalidad, de forma de poder compararlos más fácilmente entre sí.

Como se observa en la siguiente tabla y dentro del marco del experimento, los motores de búsqueda a utilizar son Google y Altavista si se desea obtener una mayor *exactitud*.

Exactitud	**Nivel 1**	**Nivel 2**	**Nivel 3**	**Nivel 4**	**Nivel 5**	**Nivel 6**	**Nivel 7**
Motores de búsqueda			Google Altavista	AllTheWeb MSN	About		

Tabla 7.7 – Comparación de niveles de *exactitud*.

La tabla 7.8 indica que cuando se desea obtener un mayor nivel de *identificación eficiente de documentos relevantes*, los mejores motores de búsqueda son Google, AllTheWeb y Altavista.

Identificación eficiente de documentos relevantes	**Nivel 1**	**Nivel 2**	**Nivel 3**	**Nivel 4**
Motores de búsqueda		Google AllTheWeb Altavista	MSN About	

Tabla 7.8 – Comparación de niveles de *identificación eficiente de documentos relevantes*.

Por otra parte, todos los motores proveen el mismo nivel de *disponibilidad*:

Disponibilidad	**Nivel 1**	**Nivel 2**	**Nivel 3**	**Nivel 4**	**Nivel 5**	**Nivel 6**

Motores de búsqueda	Google AllTheWeb MSN Altavista About					

Tabla 7.9 – Comparación de niveles de *disponibilidad.*

También se observa que Google y MSN proveen las interfaces de búsqueda más simples en cuanto a interacción con el usuario.

Facilidad de realización de búsquedas	**Nivel 1**	**Nivel 2**	**Nivel 3**
Motores de búsqueda	Google MSN	AllTheWeb Altavista About	

Tabla 7.10 – Comparación de niveles de *facilidad de realización de búsquedas.*

Para *confiabilidad*, el mayor nivel de *confiabilidad* está en manos de Google, como se observa a continuación.

Confiabilidad	**Nivel 1**	**Nivel 2**	**Nivel 3**	**Nivel 4**	**Nivel 5**
Motores de búsqueda	Google			AllTheWeb MSN Altavista About	

Tabla 7.11 – Comparación de niveles de *confiabilidad.*

7.5 Mejoras Propuestas

Esta sección propone mejoras a los motores de búsqueda de forma de alcanzar los niveles deseables formulados en el capítulo 5 – Derivación de Métricas para las funcionalidades evaluadas en el experimento.

Exactitud

- Mejorar el algoritmo de ranking para que éste utilice el contexto en el cual el usuario realiza la búsqueda y tenga en cuenta sus preferencias [EU01].
- El usuario debe poder indicar el contexto de una búsqueda para que el motor pueda comprenderlo.
- Incorporar *relevance feedback:* los usuarios no deberían tener que refinar sus búsquedas manualmente cambiando los términos de la misma sino que deberían poder indicar cuáles son los documentos más relevantes entre los recuperados. Es decir, los buscadores comerciales deberían hacer un mayor uso de las técnicas de *relevance feedback.* Esto ejecutaría una nueva búsqueda automáticamente, utilizando la retroalimentación del

usuario como punto de refinamiento, por ejemplo, buscando documentos similares a todos los marcados con "+" y lo más disimilares posible a los marcados con "-". El motor debe aprender de las elecciones de los documentos relevantes del usuario para mejorar búsquedas futuras. El buscador debe permitir que el usuario califique las respuestas obtenidas para mejorar las búsquedas futuras.

- Incluir los sinónimos de las palabras que pertenecen a la consulta en las búsquedas y permitir al usuario elegir los sinónimos que cree conveniente utilizar.
- Mejorar el algoritmo de *crawling* buscando entre las páginas no por DFS o BFS sino de acuerdo a una heurística que determine cuáles son los enlaces más importantes de cada documento y realizando el *crawling* en ese orden. La heurística podría estar basada en el ranking de términos más utilizados para búsquedas en la Web. Esto permitiría que las páginas más relevantes para los usuarios sean indexadas primero.
- Indexación: mejorar este proceso para que la falta de editorialización de los documentos de la Web no afecte tan significativamente la recuperación de resultados, es decir, al usuario final.
- Relevancia: uno de los problemas que se deben atacar para mejorarla es el del no aprovechamiento de la semántica de los documentos. Este problema se mejoraría utilizando la Web Semántica, la cual intenta responder a los problemas de semántica y polisemia de los motores tradicionales. Tim Berners-Lee, el creador de la WWW, escribió la publicación original sobre este tema junto con Hendler y Lassila [BHL01]. El concepto principal de la Web semántica consiste en que, así como los seres humanos, las computadoras pueden descubrir el significado de los datos en la Web con el fin de alcanzar objetivos útiles para los usuarios. La relevancia de las búsquedas podría ser mejorada si los buscadores comprendieran la semántica de los documentos que indexan además de construir un índice de términos y documentos. Se intenta crear un motor que discrimine los resultados de sus búsquedas mediante el uso de la semántica. Esto sería aun más beneficioso en los casos en los que no se está seguro del objetivo de la búsqueda. La Web semántica va más allá de los índices y las palabras clave, permitiendo a los usuarios buscar por contenido y categorías.

Identificación eficiente de documentos relevantes

- Clasificar los resultados de una búsqueda, de forma que el usuario pueda evaluar el conjunto de resultados y descartar los conjuntos que no le interesen. E.g., Vivísimo [VV05] y Clusty [CLU05].
- Mejorar el ranking como se propuso en *exactitud* implica que los documentos relevantes van a encontrarse en las primeras posiciones. De esta forma, van a ser más fáciles de encontrar.
- Modificar el resumen de resultados, de manera que éste explique semánticamente los documentos o provea una contextualización de los términos que pertenecen a una búsqueda.

Disponibilidad

- Utilizar caché de documentos, el cual permite acceder a documentos que no se podrían evaluar si el servidor no se encontrara disponible.
- Antes de devolver el resumen con los documentos relevantes, analizar si los mismos (o por lo menos los primeros x documentos) están disponibles. En caso negativo,

eliminarlos del resumen o presentarlos ante el usuario de forma tal que se pueda distinguir qué documentos son accedidos mediante su caché.

- Si el usuario hace clic en un resultado no disponible, el navegador le puede indicar al buscador que la página no está respondiendo. De esta manera, el motor podrá tener en cuenta esta información para no listar el documento como respuesta en futuras búsquedas o, alternativamente, presentarlo indicándole al usuario que lo está accediendo del *cache*.

Facilidad de realización de búsquedas

Si bien no se encontraron problemas para esta funcionalidad mediante el experimento, se proponen a continuación algunas mejoras para la misma:

- Al realizar la consulta, pedirle a los usuarios que ingresen un contexto para la misma.
- Educar al usuario de forma que realice búsquedas más eficientes. A medida que el usuario ingresa los términos de una búsqueda, los motores de búsqueda deberían ayudarlo a construir una consulta adecuada para sus necesidades. Hoy en día, la mayoría de los usuarios no sabe cómo mejorar sus búsquedas (agregando tantos términos como sea posible, restringiendo los dominios, países, fechas e idiomas de los documentos, etc.) Google está probando actualmente una herramienta que tiene como objetivo ayudar a resolver este problema: *Google suggest*. Esta herramienta funciona de la siguiente forma: mientras el usuario ingresa las palabras clave, *Google suggest* intenta adivinar qué es lo que se desea buscar y ofrece sugerencias de forma interactiva para ayudar a armar la consulta. La siguiente figura muestra cómo funciona esta herramienta.

Figura 7.36 – *Google Suggest Beta.*

El siguiente paso que esta herramienta todavía no resuelve, consiste en enseñarle al usuario a utilizar la mayor cantidad de términos posible así como operadores booleanos, ya que, como fue observado anteriormente, la falta de los mismos en cuatro de cada cinco búsquedas colabora con la disminución de la relevancia. El avance ideal es la automatización de estos pasos por parte del buscador. Éste podría detectar quién es el usuario y filtrar automáticamente los documentos en los que éste está interesado en base a su perfil. Al mismo tiempo, debería ser fácil para el usuario sobrescribir los atributos de su perfil para una búsqueda determinada.

Por otra parte, MSN Search provee una funcionalidad relacionada con su nueva versión, llamada *Search Builder* [SB05]. Su objetivo es ayudar al usuario a construir una búsqueda correcta para mejorar la relevancia de los resultados. Para alcanzar esta meta, MSN Search provee una interfaz de ingreso de búsquedas simplificada que facilita el acceso a las funcionalidades de restricción y refinamiento de consultas. La siguiente figura ilustra *Search Builder.*

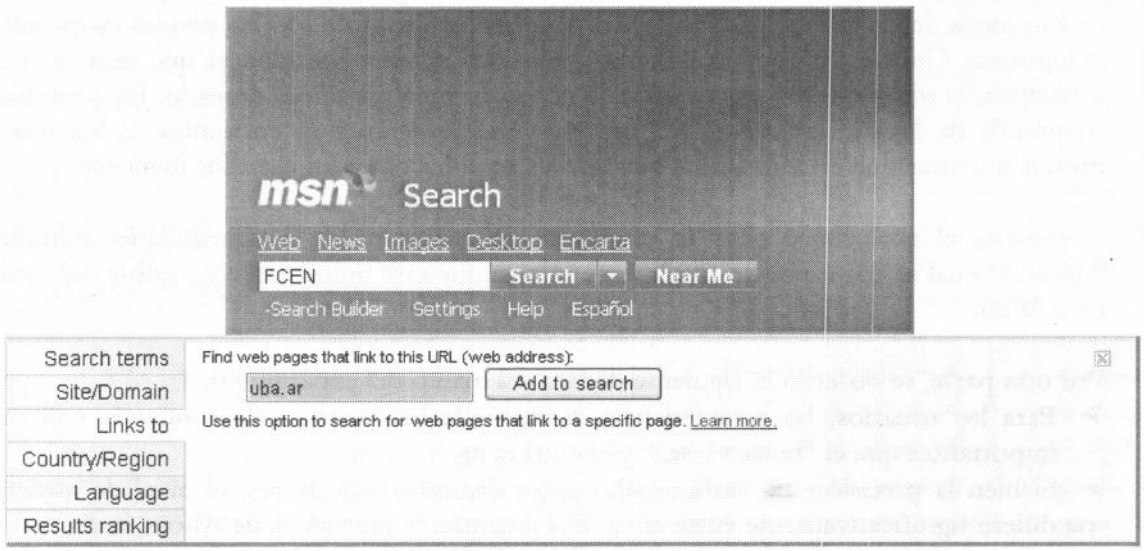


Figura 7.37 – MSN Search: *Search Builder.*

Confiabilidad

- Los documentos devueltos deben estar disponibles en los idiomas que el usuario comprende.
- Las interfaces deben estar disponibles en idiomas que el usuario entiende.
- Mejorar la *exactitud* para que el usuario encuentre documentos relevantes y así aumente su nivel de *confiabilidad* con respecto al buscador.

Como se observa en los párrafos anteriores, un motor de búsqueda debe contar con el conjunto de funcionalidades deseables descriptas en el capitulo 5 a fin de ser un buscador de calidad. La mayoría de las funcionalidades está relacionadas entre si y, por lo tanto, una mejora en una de ellas afecta el rendimiento de las otras. E.g., las mejoras propuestas para *exactitud* incluyen a las funcionalidades de *relevance feedback* y perfiles y contextos.

Para culminar este capítulo, se presentan las conclusiones generales del experimento.

7.6 Conclusiones del Experimento

Como se mencionó anteriormente, el nivel de *disponibilidad* para todos los motores de búsqueda fue mayor al esperado. Esto se debe en gran medida a que los usuarios no evaluaron 50 documentos por búsqueda.

Al observar detenidamente las secciones anteriores, se puede determinar que ningún motor de búsqueda de los cinco elegidos para el experimento cumple con los niveles estipulados en la hipótesis. Google es el que más se aproxima a los niveles deseados, ya que sólo en *exactitud* e *identificación eficiente de documentos relevantes* posee un nivel menor al deseado. Esto también se desprende de las tablas 7.11 a 7.15, ya que Google siempre se encuentra en los primeros niveles mientras que el resto de los buscadores se ubica entre los estadíos menores.

Asimismo, el motor que peor se comporta en cuanto a las funcionalidades mínimas es About, lo cual se corresponde con lo esperado ya que este motor tiene un pobre cubrimiento de la Web.

Por otra parte, se observó lo siguiente dentro del marco del experimento:

- Para los usuarios, las características técnicas de los motores de búsqueda son menos importantes que el "boca a boca" y el marketing.
- Si bien la precisión varía mucho entre distintos buscadores, el nivel de *confiabilidad* no difiere significativamente entre ellos. Por ejemplo, la relevancia de About es muy inferior a la de MSN mientras que sus niveles de *confiabilidad* son similares.

Para finalizar, se observa lo siguiente:

En el marco del experimento realizado, se mostró que se cumple la hipótesis planteada. Es decir, ninguno de los motores de búsqueda evaluados cumple con la propiedad 6.1 del capítulo 6.

CAPITULO 8. CONCLUSIONES

"Cuando se proclamó que la Biblioteca abarcaba todos los libros,
la primera impresión fue de extravagante felicidad.
Todos los hombres se sintieron señores de un tesoro intacto y secreto.
No había problema personal o mundial cuya
elocuente solución no existiera: en algún hexágono."

Jorge Luis Borges – La Biblioteca de Babel – Ficciones

8.1 Introducción

En este capítulo se presentan las conclusiones de la investigación realizada en esta tesis, tanto desde el punto de vista empírico como desde la perspectiva teórica. Asimismo, se proponen ideas para trabajos futuros relacionados con el tema de este trabajo.

8.2 Conclusiones de la Investigación

Dado que la WWW crece con mucha rapidez y que el volumen de información almacenada es realmente grande, satisfacer las necesidades de los usuarios es una tarea muy difícil para los motores de búsqueda. Este trabajo se desarrolló a fin de analizar en detalle el nivel de los buscadores y compararlos, analizándolos mediante un enfoque basado en calidad de datos. Con este propósito se definieron objetivos, preguntas y métricas mediante el método GQM y se identificaron los atributos más importantes desde el punto de vista de esta tesis, así como la forma en la que se deben medir.

Dentro del marco del experimento realizado, se comprobó que la hipótesis es válida ya que los motores de búsqueda evaluados no satisfacen los niveles de calidad especificados para el conjunto de funcionalidades mínimas de la propiedad 6.1. En consecuencia, se observó que la calidad de los buscadores estudiados es inferior a la deseada. Es decir que el estado del arte de los motores de búsqueda no es aún muy avanzado. Esto se observa al obtener resultados muy poco relevantes para las consultas realizadas por los usuarios en áreas de su expertez. Uno de los motivos más importantes es que los motores de búsqueda no tienen en cuenta el contexto ni el perfil de los usuarios que realizan las consultas ni utilizan *relevance feedback* adecuadamente. Considerar estos aspectos mejoraría la relevancia ya que acotaría el universo de búsqueda a un área más reducida. Asimismo, las búsquedas serán más efectivas sólo cuando el usuario pueda acceder de cierta manera a la clasificación, vocabulario de meta datos y esquemas de indexación de los buscadores. No se hace referencia a un acceso literal a los mismos sino a acceder a términos específicos pertenecientes al vocabulario del buscador como respuesta a palabras clave utilizadas para una búsqueda particular.

Por otra parte, y tal como lo indican Baeza Yates y Ribeiro Neto en *Modern Information Retrieval* [BR99], "los motores de búsqueda aún retornan mucha paja junto con la aguja". Las mejoras sugeridas en el capítulo 7 para los problemas de R.I.W. analizados en el capítulo 3,constituyen la base de las propuestas para aumentar la calidad de los motores de búsqueda junto con las funcionalidades mínimas (técnicas y de usuario) definidas en el capítulo 6. Del mismo modo, incluir el resto de las funcionalidades definidas en el capítulo 5 ayudaría a mejorar la calidad de los buscadores.

De acuerdo a los resultados del experimento, una característica adicional que afecta a la relevancia y que está vinculada al uso del contexto y el perfil del usuario es el orden en que se retornan los documentos relevantes. Muy frecuentemente, los motores de búsqueda evaluados encuentran documentos relevantes pero los listan más allá de los diez primeros resultados. De este modo, los usuarios no pueden aprovecharlos ya que la mayoría suele consultar sólo la primera página de respuestas (i.e., las primeras diez) y con menor frecuencia la segunda. Si los buscadores utilizaran perfiles y contextos para el algoritmo de ranking, podrían posicionar mejor en el resumen los documentos a los que responden a la necesidad del usuario.

Entre las conclusiones obtenidas del experimento, se observa que los usuarios confían ampliamente en Google y consideran seguir utilizándolo, incluso aquéllos que no encontraron lo que deseaban. Esto sólo ocurrió con Google, ya que los usuarios indicaron que no confían en los otros motores de búsqueda evaluados. Las respuestas obtenidas para *confiabilidad* con respecto a AllTheWeb (ATW) y About se condicen con lo que se esperaba, ya que ningún usuario conocía estos buscadores con anterioridad y el hábito es uno de los atributos clave que impactan la *confiabilidad.* Es decir, utilizar dichos motores durante el experimento no alcanzó para que los usuarios confiaran en sus respuestas. Esto se vio reforzado por el hecho de que la calidad de sus respuestas estuvo por debajo del nivel esperado.

Con respecto a la *identificación eficiente de documentos relevantes*, se observó que el usuario generalmente comprende el resumen devuelto por un motor de búsqueda, aunque éste podría mejorar agregando más información sobre el contexto del documento y su estructura. Es fundamental que los documentos más relevantes se encuentren en la primera página de resultados, lo que ocurrió en la menor parte de los casos. Lo mismo sucedió con el orden en que fueron devueltos los documentos, ya que éste coincidía muy raramente con la percepción del usuario en cuanto a relevancia.

En cuanto a la funcionalidad *facilidad de realización de búsquedas*, se advirtió que los usuarios comprenden en general cómo realizar una consulta sintácticamente, obteniendo la definitiva en un máximo de cinco intentos. Sin embargo, utilizan muy ocasionalmente las opciones avanzadas o los operadores booleanos. El uso de estas facilidades permitiría acotar la búsqueda pero los usuarios no se animan o no comprenden cómo hacerlo. Para mejorar esta situación, los motores de búsqueda deberían ayudarlos de forma interactiva o con páginas de ayuda más claras para que puedan realizar consultas más refinadas y precisas.

Todos los motores evaluados alcanzaron el nivel de calidad requerido por la hipótesis para *disponibilidad.* Como se mencionó en el capítulo 7, en el experimento se les pidió a los

usuarios que revisaran los primeros cincuenta documentos devueltos para una consulta. No obstante, la mayoría no revisó esta cantidad de documentos ya que esto no se corresponde con su comportamiento en la vida real. Este hecho aumentó artificialmente el nivel de *disponibilidad* porque podría haber habido documentos inaccesibles que no sólo no fueron revisados sino que se consideraron accesibles para el experimento.

Uno de los problemas que se encontraron al trabajar con las funcionalidades es la imposibilidad de evaluar la *completitud* empíricamente, ya que actualmente no es factible determinar el universo de discurso completo para una búsqueda. Si bien se analizó la posibilidad de realizar la unión de los primeros cincuenta resultados devueltos por los cinco buscadores, ésta fue descartada por su costo y porque en la mayoría de las ocasiones los documentos devueltos diferían entre los buscadores.

Se ha observado la importancia del idioma en la selección del motor de búsqueda a utilizar, ya que muchos usuarios indicaron que no volverían a usar los motores de búsqueda que no poseen una interfaz gráfica (GUI) en castellano. Por otro lado, es importante que se pueda seleccionar el país y el idioma sobre el cual se va a realizar la consulta, de forma tal de poder reducir el universo de documentos, favoreciendo la relevancia y la localidad de resultados.

Por otra parte, así como AllTheWeb es el segundo motor en cuanto a cantidad de páginas indexadas después de Google [SEW05], el experimento confirmó que éste está por debajo de los niveles de calidad obtenidos para Google. E.g., el nivel de *exactitud* de ATW fue tan sólo cuatro, es decir, algunos de los documentos recuperados son aceptables para el usuario aunque no exactamente los deseados o relevantes. Esto muestra que a pesar de tener un gran índice, el algoritmo de ranking no es lo suficientemente exacto. La baja precisión y recuperación de este buscador pueden estar altamente afectadas por el idioma de las consultas, ya que la gran mayoría de ellas se realizó en castellano. Este es un problema a resolver en R.I.W.; el experimento confirmó que la performance de los atributos principales de calidad de usuario es menor para búsquedas realizadas en idiomas distintos al inglés.

Entre los motores de búsqueda evaluados, Google - el mejor posicionado en Search Engine Watch - es el que más se acerca a los niveles de calidad definidos en este trabajo. Por otra parte, el motor que no se encuentra en ningún ranking (About), es el que posee los niveles más bajos para las funcionalidades seleccionadas. Es decir que dentro del marco evaluado, el experimento comprobó los resultados de Search Engine Watch para las funcionalidades desde la perspectiva de usuario, las cuales no son evaluadas extensivamente por esta organización.

En la próxima sección se proponen ideas de trabajos que se sugieren como continuación de esta tesis.

8.3 Trabajos Futuros

En esta sección se analizan áreas interesantes para seguir investigando a partir de este trabajo.

- El principal trabajo a continuar es evaluar el conjunto minimal de funcionalidades técnicas para un motor de búsqueda. Esto permitirá comparar los motores de búsqueda no sólo por sus atributos funcionales sino también por sus atributos técnicos.

- Otro punto importante es investigar en profundidad alternativas para evaluar los niveles de *completitud* empíricamente, ya sea utilizando como universo de búsqueda la unión de los cincuenta primeros resultados o un nuevo método para medir este atributo. Esto permitirá tener una mejor noción de la relevancia alcanzada por cada motor de búsqueda.

- Se sugiere continuar con la derivación de métricas para las funcionalidades que no se estudiaron, de forma tal de tener una gama más grande de atributos para comparar los buscadores.

- Esta tesis se centra en el análisis de motores de búsqueda para búsquedas de texto. Se podría realizar un análisis similar para otros tipos de documento como imágenes y video. Asimismo, sería interesante proponer mejoras para R.I.W. para este tipo de documentos.

- Este trabajo se concentró en los motores de búsqueda, dejando como idea para trabajos futuros la posibilidad de realizar una evaluación análoga en directorios y meta buscadores. También se sugiere realizar una investigación similar a la indicada en el punto anterior para directorios y meta buscadores y otra para búsquedas en lenguaje natural.

- La Web semántica es un nuevo campo que debería afectar positivamente a la precisión y la recuperación en R.I.W. Estudiar en profundidad la estructura y los meta datos de los documentos Web, más allá de los términos que los componen, es clave para el éxito de los buscadores del futuro. Sería interesante analizar este tema y su relación con R.I.W. en profundidad.

- Otro campo a estudiar es la próxima generación de motores de búsqueda, la que tendrá a su alcance la "Web profunda" además de la superficial. Esto facilitará el acceso a páginas dinámicas y restringidas a usuarios específicos y proveerá resultados mejor organizados, ya sea mediante *clustering* o un mejor ordenamiento de documentos de acuerdo a los intereses de un usuario en particular. Además de mejorar el ranking de relevancia, la inteligencia de los buscadores jugará un papel mayor descubriendo conocimiento, meta datos e información de contexto adicional para las búsquedas.

- Finalmente, se propone estudiar en profundidad los problemas de R.I.W. mencionados en el capítulo tres y proponer soluciones adicionales para los mismos.

Lo bueno de un procedimiento científico es que nunca hace perder tiempo a los demás: también trabajar siguiendo el surco de una hipótesis científica para descubrir después que hay que refutarla es hacer algo útil bajo el impulso de una propuesta precedente. Si nuestra tesis además sirve para animar a alguien a efectuar otras experiencias, habremos logrado algo útil. [E77]

BIBLIOGRAFIA

[AA98] Altavista: http://www.altavista.com, 1998.

[AB05] About: http://www.about.com, 2005.

[Abi97] Abiteboul, S.: Querying Semi-Structured Data, *International Conference on Database Theory (ICDT)*, No. 1186 en LNCS, pp. 1-18, Grecia, 1997.

[AJ05] Ask Jeeves: http://www.askjeeves.com, 2005.

[AL05] Alexa: http://www.alexa.com, 2005.

[ALL05] AllTheWeb: http://www.alltheweb.com, 2005.

[B96] Beizer, B.: Software System Testing and Quality Assurance, pp. 27-32, 1996.

[BB98] Bharat, K., Broder, A.: A Technique for Measuring the Relative Size and Overlap of the public Web Search Engines, *Proceedings of the 7º WWW Conference*, pag. 379-388, 1998.

[BDMS94] Bowman, M., Danzig, P., Manber, U., Schwartz, M.: Scalable Internet Resource Discovery: Research Problems and Approaches, http://citeseer.nj.nec.com/cache/papers2/cs/6678/http:zSzzSzthor.csie.ntu.edu.twzSznotebookzSzreviewed_paperzSzreferenceszSzharvest-projzSzRD_ResearchProblems_Jour.pdf/bowman94scalable.pdf, 1994.

[Bel00] Belek, Richard: Finding Out About – A Cognitive Perspective on Search Engine Technology and the WWW, *Cambridge University Press*, 2000.

[BH98] Birmingham, W., Durfee, E. H., Mullen, T., Wellman, M.: The Distributed Agent Architecture of the University of Michigan Digital Library, *AAAI Spring Simposium on Information Gathering from Heterogeneous, Distributed Environments, AAAI Press*, 1995.

[BHL01] Berners-Lee, T., Hendler, J., Lassila, O.: I know what you mean... the semantic Web: a new form of Web content that is meaningful to computers will unleash a revolution of new possibilities, *Scientific American,* vol.284, i5, pag.34 (10), mayo de 2001.

[BL89] Berners-Lee, Tim: Information Management: A Proposal, *CERN*, 1989.

[BL90] Berners-Lee, Tim: The World Wide Web: Past, Present and Future, http://www.w3.org/People/Berners-Lee/1996/ppf.html, 1990.

[BL97] Blum, A., Langley, P.: Selection of Relevant Features and Examples in Machine Learning, *Special Issue of Artificial Intelligence on 'Relevance'*,

http://citeseer.nj.nec.com/cache/papers2/cs/677/http:zSzzSzwww.isle.orgzSz~langleyzSzpaperszSzrelevance.aij.pdf/blum97selection.pdf, 1997.

[BMY98] Bobrowski, M., Marré, M., Yankelevich, D.: A Software Engineering View of Data Quality, *Proceedings of 2nd European Quality Week*, Bruselas, Bélgica, noviembre de 1998.

[BMY99] Bobrowski, M., Marré, M., Yankelevich, D.: A Homogeneous Framework to Measure Data Quality, *Conference on Information Quality*, Boston, EE.UU., 1999.

[BO05] Weblog de tecnología: http://server.boardgamestuff.com:8000/blog/Technology/Blog/BlogAudience.html, 2005.

[BP98] Brin, S., Page, L.: The Anatomy of a Large-Scale Hypertextual Web Search Engine, *Proceedings of the 7th International WWW Conference*, Australia, abril de 1998.

[BR88]: Basili, V., Rombach, D.: The TAME Project: Towards Improvement-Oriented Software Environments, *IEEE Transactions on software engineering*, vol. 14, No. 6, junio de 1998.

[BR99] Baeza-Yates, R., Ribeiro Neto, B.: Modern Information Retrieval, *Addison-Wesley*, 1999.

[BW97] Baldonado, M., Winograd, T.: SenseMaker: An Information-Exploration Interface Supporting the Contextual Evolution of a User's Interests, *Proceedings of CHI*, http://citeseer.nj.nec.com/cache/papers2/cs/5893/http:zSzzSzwww-diglib.stanford.eduzSzdiglibzSzWPzSzPUBLICzSzDOC115.pdf/baldonado97sensemaker.pdf, 1997.

[C1] Web Searching Tips: http://searchenginewatch.com/facts/, 2005.

[C10] What People Search for: http://searchenginewatch.com/facts/searches.html, 2005.

[C11] Search Spotlight: http://searchenginewatch.com/facts/spotlight.html, 2005.

[C12] Web Admin's Guide to Site Search Tools: http://www.searchtools.com/info/guide.html, 2005.

[C13] Explore the Internet!: http://www.amdahl.com/internet/meta-index.html, 2005.

[C14] Data Quality Objectives: http://dqo.pnl.gov/, 2005.

[C15] The Invisible Web: Searchable Databases and Hidden Sites: http://www.websearch.about.com/internet/websearch/library/weekly/aa061199.htm, 2005.

[C16] Search Engine Reviews: http://searchenginewatch.com/resources/reviews.html, 2005.
[C17] Impact of Data Quality on the Web User Experience: http://www.useit.com/alertbox/980712.html, 2005.

[C18] Comparación de las Medidas de Tendencia Central: http://www.cca.org.mx/dds/cursos/estadistica/html/m10/comoparacion_medidas.htm, 2005.

[C19] Medidas de Forma y Concentración: http://www3.uji.es/~mateu/tema4-d37.doc, 2005.

[C2] How Search Engines Work: http://searchenginewatch.com/webmasters/work.html, 2005.

[C20] Medidas de Asimetría: http://thales.cica.es/rd/Recursos/rd99/ed99-0278-01/est_des7.html, 2005.

[C21] Medidas de Tendencia Central y de Dispersión: http://server2.southlink.com.ar/vap/MEDIDAS.htm, 2005.

[C22] Unidad 3 – Estadística Descriptiva: http://www.uaq.mx/matematicas/estadisticas/xu3.html, 2005.

[C3] Search Engine Glossary: http://searchenginewatch.com/facts/glossary.html, 2005.

[C4] Search Engine Math: http://searchenginewatch.com/facts/math.html, 2005.

[C5] Boolean Searching: http://searchenginewatch.com/facts/boolean.html, 2005.

[C6] Power Searching for Anyone: http://searchenginewatch.com/facts/powersearch.html, 2005.

[C7] Search Utility Reviews & Articles: http://searchenginewatch.com/resources/utilities.html, 2005.

[C79] Crosby, P.: Quality is Free, *McGraw Hill*, 1979.

[C8] The Search Engine Report – Popular Articles: http://searchenginewatch.com/sereport/bysubject.html, 2005.

[C9] Search Engine Sizes: http://searchenginewatch.com/reports/sizes.html, 2005.

[CDGKRR98] Chakrabarti, S., Dom, B., Gibson, D., Kleinberg, J., Raghavan, P., Rajagopalan, S.: Automatic Resource Compilation by Analyzing Hyperlink Structure and Associated Text, http://citeseer.nj.nec.com/cache/papers2/cs/883/http:zSzzSzwww.cs.princeton.eduzSzcourseszSzarchivezSzspring98zSzcs598bzSzauthoritative_system.pdf/chakrabarti98automatic.pdf, 1998.

[CGP98] Cho, J., García Molina, H., Page, L.: Efficient Crawling through URL Ordering, *7th WWW Conference*, Brisbane, Australia, abril de 1998.

[CHO04] Chowdhury, G.: Introduction to Modern Information Retrieval, 2° edición, *Facet Publishing*, 2004.

[CLI05] Clicz.com: http://www.clickz.com/stats/stats_toolbox/article.php/156101, 2005.

[CLU05] Clusty: http://www.clusty.com, 2005.

[CLW97] Coffman, E., Liu, Z., Weber, R.: Optimal Robot Scheduling for Web Search Engines, *Technical Report 3317, INRIA*, Francia, diciembre de 1997.

[CN96]: Crowder, G., Nicholas, C.: Resource Selection in CAFE: an Architecture for Network Information Retrieval, http://citeseer.nj.nec.com/cache/papers2/cs/865/http:zSzzSzwww.cs.umbc.eduzSz~crowderzSzpubszSzsigir96.pdf/crowder96resource.pdf, 1996.

[CS05] CiteSeer.IST – Scientific Literature Digital Library: http://citeseer.ist.psu.edu, 2005.

[CV96] Crovella, M., Bestavros, A.: Self-Similarity in World Wide Web Traffic: Evidence and Posible Causes, *ACM Sigmetrics Conference on Measurement and Modeling of Computer Systems*, pp. 160 – 169, mayo de 1996.

[CV98] Crovella, M., Barford, P.: Generating Representative Web Workloads for Network and Server Performance Evaluation, *ACM Sigmetrics Conference on Measurement and Modeling of Computer Systems*, pp. 151 – 160, julio de 1998.

[DM82] DeMarco, Tom: Controlling Software Projects: Management, Measurement, and Estimates, *Prentice Hall*, 1982.

[DP94] De Bra, P., Post, R.: Information Retrieval in the World-Wide Web: Making Client-Based Searching Feasible, http://citeseer.nj.nec.com/cache/papers2/cs/299/http:zSzzSzwww.cern.chzSzPapersWWW94zSzreinpost.pdf/information-retrieval-in-the.pdf, 1994.

[E77] Eco, U.: Cómo se hace una tesis, *Gedisa Editorial*, 1997.

[EP05] e-Publishing: Website Marketing, Algoritmo de Kleinberg, http://www.e-publishing.de/online/suchmaschinen/die_echten/suchmaschinen_bewerten.html, 2005.

[ES96] Etzioni, O., Selberg, E.: The MetaCrawler Architecture for Resource Aggregation on the Web, http://www.cs.washington.edu/homes/speed/papers/ieee/ieee-metacrawler/ieee-metacrawler.html, 1996.

[EU01] Efron, A., Urtubey, F.: Personalización de Búsquedas Web Usando Perfiles y Contextos, Tesis de Licenciatura, *Departamento de Ciencias de la Computación, F.C.E.N., U.B.A.*, 2001.

[EZ98] Etzioni, O., Zamir, O.: Web Document Clustering: A Feasibility Demonstration, http://citeseer.nj.nec.com/cache/papers2/cs/18/http:zSzzSzzhadum.cs.washington.eduzSzzamirzSzsigir98.pdf/zamir98web.pdf, 1998.

[F98] Frank, A.: Metamodels for Data Quality Description, http://citeseer.nj.nec.com/cache/papers2/cs/3225/ftp:zSzzSzgi27.geoinfo.tuwien.ac.atzSzfrankzSz2690_dataquality_short.pdf/metamodels-for-data-quality.pdf, 1998.

[FOA00] Belek, R.: Finding Out About: A Cognitive Perspective on Search Engine Technology and the WWW, *Cambridge University Press*, 2001.

[FP97] Fenton, N., Pfleeger, S.: Software Metrics – A Rigorous and Practical Approach, 2da. Edición, Edición Revisada, *PWS Publishing Company*, 1997.

[FS05] Feedster: http://www.feedster.com, 2005.

[FS98] Faulstich, L., Spiliopoulou, M.: WUM: A Web Utilization Miner, http://citeseer.nj.nec.com/cache/papers2/cs/1261/http:zSzzSzwww.inf.fu-berlin.dezSz~faulsticzSzpubzSz1998zSzWebDB98.pdf/spiliopoulou98wum.pdf, 1998.

[Fur87] Furnas, G. Landauer, T., Gomez, L., Dumais, S.T.: The Vocabulary Problem in Human-System Communication, *Communications of the ACM, 30(11):964-971*, 1987.

[G97] Guy, L.: Strategies for Searching for Information on the Internet, *International Data Resources Workshop*, http://dpls.dacc.wisc.edu/www_searchers.html, 1997.

[GG05] Google: http://www.google.com, 2005.

[GMJ91] Ghezzi, C., Mandrioli, M., Jazayeri, D.: Software Engineering, pp. 17-36, *Addison Wesley*, 1991.

[GL98] Giles, L., Lawrence, S.: Searching the World Wide Web, *Science*, pp. 98-100, vol. 280, http://citeseer.nj.nec.com/cache/papers2/cs/7956/http:zSzzSzwww.neci.nj.nec.comzSz~lawrencezSzpaperszSzsearch-science98zSzsearch-science98.pdf/lawrence98searching.pdf, abril de 1998.

[GS98] Gertz, M., Schmitt, I.: Data Integration Techniques Based on Data Quality Aspects, http://citeseer.nj.nec.com/cache/papers2/cs/4941/http:zSzzSzwwwiti.cs.uni-magdeburg.dezSzfdb98zSzonline-proczSz1-19.pdf/data-integration-techniques-based.pdf, 1998.

[H00] Hearst, M.: Next Generation Web Search: Setting Our Sites, *IEEE Data Engineering Bulletin,* http://www.sims.berkeley.edu/~hearst/papers/data-engineering/, septiembre de 2000.

[HA94] Bowman, C., Danzig, P., Hardy, D., Manber, U., Schwartz, M.: The Harvest Information Discovery and Access System, http://archive.ncsa.uiuc.edu/SDG/IT94/Proceedings/Searching/schwartz.harvest/schwartz.harvest.html, 1994.

[HLW98] Huang Huan-tsae, Lee, Yang W., Wang, Richard: Quality Information and Knowledge, p. 82, *Prentice Hall*, 1998.

[HMI03] Proyecto How Much Information 2003, http://www.sims.berkeley.edu/research/projects/how-much-info-2003/internet.htm, 2003.

[HW97] Hauptmann, A., Wactlar, H.: Indexing and Search of Multimodal Information, http://citeseer.nj.nec.com/cache/papers2/cs/146/http:zSzzSzwww.cs.cmu.eduzSzafszSzcszSzuserzSzalexzSzdocszSzidvlzSzicassp97.pdf/hauptmann97indexing.pdf, 1997.

[I86] ISO 8402-1986.

[IDS05] Internet Systems Consortium – Internet Domain Survey, http://www.isc.org/index.pl?/ops/ds/, 2005.

[J91] Jones, C., Applied Software Measurement, pp. 253-307, *McGraw Hill*, 1991.

[JG99] Juran, J., Godfrey, A. B.: Juran's Quality Handbook, 5ta. Edición, *McGraw Hill*, 1999.

[K95] Krulwich, B.: Learning User Interests across Heterogeneous Document Databases, http://citeseer.nj.nec.com/cache/papers2/cs/24/ftp:zSzzSzvenera.isi.eduzSzsimszSzsss95zSzkrulwich.pdf/krulwich95learning.pdf, 1995.

[KS95] Konopnicki, D., Shmueli, O.: W3QS: A Query System for the World Wide Web, *Proceedings of VLDB'95,* pp. 54 – 65, Zurich, Suiza, septiembre de 1995.

[L96] Lager, M.: Spinning a Web Search, http://www.library.ucsb.edu/untangle/lager.html, 1996.

[LiV03] Lyman, P., Varian, H.: How Much Information 2003?, *Universidad de Berkeley*, 2003.

[LYN97] Lynch, C.: Searching the Internet, *Scientific American*, vol. 276, i.3, p.52, marzo de 1997.

[MMM97] Mendelzon, A., Mihaila, G., Milo, T.: Querying the World Wide Web, *International Journal on Digital Libraries, pp. 54 – 67*, http://citeseer.nj.nec.com/cache/papers2/cs/2959/http:zSzzSzwww.math.tau.ac.ilzSz~mataszSzcourseszSzpaperszSzwebsql.pdf/mendelzon97querying.pdf, abril de 1997.

[MSN05] MSN Search: http://search.msn.com, 2005.

[N00] Navarro, G.: Recuperación de la Información: Algoritmos, Estructuras y Aplicaciones en la Web, *Escuela de Ciencias Informáticas, Departamento de Computación, Universidad de Buenos Aires*, julio de 2000.

[NET05] Netcraft Surveys: http://news.netcraft.com/archives/2003/04/14/netcraft_hosting_provider_server_count_available.html, 2005.

[OCL05] Web Characterization Project: http://www.oclc.org/research/projects/archive/wcp/stats/linkage.htm, 2003.

[OLB03] Trends in the Evolution of the Public Web: http://www.dlib.org/dlib/april03/lavoie/04lavoie.html, 2003.

[RAE] Diccionario de la Real Academia Española: http://www.rae.es.

[Red96] Redman, T.: Data Quality for the Information Age, *Artech House*, 1996.

[Red98] Redman, T.: The Impact of Poor Data Quality on the Typical Enterprise, *Communications of the ACM*, vol. 41, no. 2, pp. 79-82, febrero de 1998.

[RS99] Rosenthal, A., Sciore, E.: Helping Data Suppliers and Consumers Negotiate the Details: A Database View Approach, http://citeseer.nj.nec.com/cache/papers2/cs/6473/http:zSzzSzwww.cs.bc.eduzSz~sciorezSzpaperszSznegot.pdf/rosenthal99helping.pdf, 1999.

[S97]: Wang, Q: Data Gathering and Analysis, http://sern.ucalgary.ca/courses/seng/621/W97/qianw/present.html, 1997.

[S98]: Shepperd, M: Goal Question Metric, http://dec.bournemouth.ac.uk/ESERG/mshepperd/SEMGQM.html, 1998.

[SB05] MSN Search - Search Builder: http://search.msn.com/?qb=1, 2005.

[SB99] van Solingen, R., Berghout, E.: The goal/question/metric method, a practical method for quality improvement of software development, http://is.twi.tudelft.nl/gqm/GQM%20Guide%20non%20printable.pdf, *McGraw-Hill*, 1999.

[SEI00] Software Engineering Process Management Program – Frequently Asked Questions, http://www.sei.cmu.edu/programs/sepm/process-faq.pdf, *Carnegie Mellon University*, otoño de 2000.

[SEW05] Sullivan, D.: Search Engine Watch, http://www.searchenginewatch.com, 2005.

[SMLab] SMLab: GQM Method Application, http://ivs.cs.uni-magdeburg.de/sw-eng/us/java/GQM/main.shtml, 2005.

[SOM01] Sommerville, Ian: Software Engineering, 6° edición, *Addison Wesley*, 2001.

[TA96] Tanenbaum, A.: Computer Networks, 3° edición, *Prentice Hall*, 1996.

[TDQ05] Total Data Quality Management Institute (TDQM): http://web.mit.edu/tdqm/www/index.shtml , *Massachusetts Institute of Technology*, 2005.

[TE05] Teoma: http://www.teoma.com, 2005.

[TOU05] Google Browser: http://www.touchgraph.com/TGGoogleBrowser.html, 2005.

[VV05] Vivisimo: http://vivisimo.com, 2005.

[W92] Weinberg, G.: Quality Software Management, *Dorser House*, 1992.

[WE05] Webopedia: http://www.webopedia.com/DidYouKnow/Internet/2002/Web_vs_Internet.asp, 2005.

[WHO05] WHOIS.net: www.whois.net, 2005.

[WI05] Whatis?com: http://whatis.techtarget.com, 2005.

[WMB94] Witten, I., Moffat, A., Bell, T.: Managing Gygabytes: Compressing and Indexing Documents and Images, *Van Nostrand Reinhold*, Nueva York, EE.UU., 1994.

[YL96] Yuwono, B., Lee, D.: WISE: A World Wide Web Resource Database System, 1996.

[ZA05] Zakon, R.: Internet Timeline, http://zakon.org/robert/internet/timeline/, 2005.

APÉNDICE I. FORMULARIO DEL EXPERIMENTO

Análisis de Calidad de Motores de Búsqueda en Internet

El propósito del experimento es poder evaluar el uso de los motores de búsqueda y poder recolectar información útil para mejorarlos. Para esto se le pide responder algunas preguntas sobre los motores de búsqueda que deseamos evaluar y realizar una consulta en cada uno de los buscadores. De parecerle relevantes ya sea por el resumen o por el texto del enlace devuelto, tómese el tiempo para abrir cada uno de los documentos y leerlos.

Importante: A fin de optimizar su tiempo, lea este documento completamente antes de contestar las preguntas"

Los motores de búsqueda a analizar en este experimento son[15]:

1) Google www.google.com
2) AllTheWeb www.alltheweb.com
3) MSN www.msn.com
4) Altavista www.altavista.com
5) About www.about.com

Experimento – Parte I

1)

Motor de búsqueda	Google	AllTheWeb	MSN	Altavista	About
¿Utilizó anteriormente este buscador?					

2) Para cada motor de búsqueda mencionado anteriormente, realice una consulta distinta sobre un tema de su expertez para el que considere que puede encontrar información útil en Internet:

1er. Motor de Búsqueda

I. Indique el motor de búsqueda que está utilizando:

Google	AllTheWeb	MSN	Altavista	About

II. Acerca de la página de ingreso de búsquedas:
¿Le resulta fácil [16] interactuar con la misma? Si la respuesta es negativa, por favor indique cuáles son las dificultades que se le presentan.

III. Ingrese la búsqueda deseada en el motor de búsqueda y transcríbala a continuación.

IV. Indique cuánto tiempo le insumió aproximadamente escribir la consulta con la cual obtuvo los resultados deseados (la consulta definitiva). Si fue más de una consulta, considere el tiempo total que le insumió escribir todas las consultas.

Tiempo = _____

V. Indique cuántas veces tuvo que reformular la consulta hasta llegar a la definitiva.

Cantidad de veces = _____

2do. Motor de Búsqueda

I. Indique el motor de búsqueda que está utilizando:

Google	AllTheWeb	MSN	Altavista	About

II. Acerca de la página de ingreso de búsquedas:
¿Le resulta fácil [17] interactuar con la misma? Si la respuesta es negativa, por favor indique cuáles son las dificultades que se le presentan.

III. Ingrese la búsqueda deseada en el motor de búsqueda y transcríbala a continuación.

[16] Ver glosario para definición de fácil.
[17] Ver glosario para definición de fácil.

IV. Indique cuánto tiempo le insumió aproximadamente escribir la consulta con la cual obtuvo los resultados deseados (la consulta definitiva). Si fue más de una consulta, considere el tiempo total que le insumió escribir todas las consultas.

Tiempo = _____

V. Indique cuántas veces tuvo que reformular la consulta hasta llegar a la definitiva.

Cantidad de veces = _____

3er. Motor de Búsqueda

I. Indique el motor de búsqueda que está utilizando:

Google | AllTheWeb | MSN | Altavista | About

II. Acerca de la página de ingreso de búsquedas:
¿Le resulta fácil[18] interactuar con la misma? Si la respuesta es negativa, por favor indique cuáles son las dificultades que se le presentan.

III. Ingrese la búsqueda deseada en el motor de búsqueda y transcríbala a continuación.

IV. Indique cuánto tiempo le insumió aproximadamente escribir la consulta con la cual obtuvo los resultados deseados (la consulta definitiva). Si fue más de una consulta, considere el tiempo total que le insumió escribir todas las consultas.

Tiempo = _____

V. Indique cuántas veces tuvo que reformular la consulta hasta llegar a la definitiva.

Cantidad de veces = _____

4to. Motor de Búsqueda

I. Indique el motor de búsqueda que está utilizando:

Google | AllTheWeb | MSN | Altavista | About

[18] Ver glosario para definición de fácil.

II. Acerca de la página de ingreso de búsquedas:
¿Le resulta fácil[19] interactuar con la misma? Si la respuesta es negativa, por favor indique cuáles son las dificultades que se le presentan.

III. Ingrese la búsqueda deseada en el motor de búsqueda y transcríbala a continuación.

IV. Indique cuánto tiempo le insumió aproximadamente escribir la consulta con la cual obtuvo los resultados deseados (la consulta definitiva). Si fue más de una consulta, considere el tiempo total que le insumió escribir todas las consultas.

Tiempo = _____

V. Indique cuántas veces tuvo que reformular la consulta hasta llegar a la definitiva.

Cantidad de veces = _____

5to. Motor de Búsqueda

I. Indique el motor de búsqueda que está utilizando:

Google	AllTheWeb	MSN	Altavista	About

II. Acerca de la página de ingreso de búsquedas:
¿Le resulta fácil[20] interactuar con la misma? Si la respuesta es negativa, por favor indique cuáles son las dificultades que se le presentan.

III. Ingrese la búsqueda deseada en el motor de búsqueda y transcríbala a continuación.

IV. Indique cuánto tiempo le insumió aproximadamente escribir la consulta con la cual obtuvo los resultados deseados (la consulta definitiva). Si fue más de una consulta, considere el tiempo total que le insumió escribir todas las consultas.

Tiempo = _____

V. Indique cuántas veces tuvo que reformular la consulta hasta llegar a la definitiva.

Cantidad de veces = _____

[19] Ver glosario para definición de fácil.
[20] Ver glosario para definición de fácil.

Experimento – Parte 2

Elija una de las consultas realizadas en el punto anterior y realícela en cada uno de los motores de búsqueda indicados. Responda las siguientes preguntas para cada motor de búsqueda.

Consulta elegida	

Google

Indique dentro de los primeros cincuenta resultados devueltos:

a. ¿Le resulta difícil[21] comprender la página de resultados? Marque sólo una respuesta.

Nivel	√
1. Muy difícil	
2. Difícil	
3. Normal	
4. Fácil	
5. Muy fácil	

Si seleccionó 1, 2 ó 3, indique cuáles son las dificultades que se le presentan:

b.

i. ¿Está satisfecho[22] con las primeras 50 respuestas obtenidas? Marque sólo una respuesta.

Nivel	√
1. Completamente satisfecho	
2. Muy satisfecho	
3. Satisfecho	
4. Poco satisfecho	
5. Insatisfecho	

[21] Ver glosario para definición de difícil.
[22] Ver glosario para definición de satisfecho.

ii. Si seleccionó 3, 4 ó 5, ¿desearía que el contenido de las mismas tuviera un mayor nivel de detalle?

Sí
No

iii. ¿Obtuvo toda la información deseada para la consulta en cuestión?

Sí
No

c. Indique la cantidad de documentos retornados.

Cantidad de respuestas = _____

d. Clasifique las primeras 50 respuestas obtenidas:

i.

	Categoría	Cantidad
1.	Me interesan	
2.	No me interesan	
3.	Inaccesibles	

ii. Indique para los documentos inaccesibles

	Posición	¿Le interesa?

NOTA: Recuerde, por favor, acceder a todos los documentos que le parezcan interesantes.

e. Dentro de los primeros 50 documentos obtenidos, indique la posición del/de los documento(s) más valioso(s) según su opinión.

Posiciones = _____

f. Indique la cantidad de documentos del resumen que haya evaluado (i.e.: que haya ingresado al documento en cuestión)

Cantidad de respuestas evaluadas = ___ de 50

g. Indique la cantidad de páginas en las cuales encontró información de su interés pero sólo luego de acceder a **enlaces adicionales** a partir de la misma.

h. Indique su nivel de satisfacción [23] para con el motor de búsqueda según su interacción con el mismo. Marque sólo una respuesta.

Nivel	√
1. Completamente satisfecho	
2. Muy satisfecho	
3. Satisfecho	
4. Poco satisfecho	
5. Insatisfecho	

¿Volvería a utilizar este motor de búsqueda? ¿Por qué?

i. Analice los documentos que considera importantes para su búsqueda, ¿cree que el orden en que fueron devueltos es el correcto, según el grado de importancia de los mismos?

Sí
No

AllTheWeb

Indique dentro de los primeros cincuenta resultados devueltos:

a. ¿Le resulta difícil [24] comprender la página de resultados? Marque sólo una respuesta.

Nivel	√
1. Muy difícil	
2. Difícil	
3. Normal	
4. Fácil	
5. Muy fácil	

Si seleccionó 1, 2 ó 3, indique cuáles son las dificultades que se le presentan:

b.

i. ¿Está satisfecho [25] con las primeras 50 respuestas obtenidas? Marque sólo una respuesta.

[23] Ver glosario para definición de satisfacción.
[24] Ver glosario para definición de difícil.
[25] Ver glosario para definición de satisfecho.

Nivel	√
1. Completamente satisfecho	
2. Muy satisfecho	
3. Satisfecho	
4. Poco satisfecho	
5. Insatisfecho	

ii. Si seleccionó 3, 4 ó 5, ¿desearía que el contenido de las mismas tuviera un mayor nivel de detalle?

Sí
No

iii. ¿Obtuvo toda la información deseada para la consulta en cuestión?

Sí
No

c. Indique la cantidad de documentos retornados.

Cantidad de respuestas = _____

d. Clasifique las primeras 50 respuestas obtenidas:

i.

Categoría	Cantidad
1. Me interesan	
2. No me interesan	
3. Inaccesibles	

ii. Indique para los documentos inaccesibles

Posición	¿Le interesa?

NOTA: Recuerde, por favor, acceder a todos los documentos que le parezcan interesantes.

e. Dentro de los primeros 50 documentos obtenidos, indique la posición del/de los documento(s) más valioso(s) según su opinión.

Posiciones = _____

f. Indique la cantidad de documentos del resumen que haya evaluado (i.e.: que haya ingresado al documento en cuestión)

Cantidad de respuestas evaluadas = ___ de 50

g. Indique la cantidad de páginas en las cuales encontró información de su interés pero sólo luego de acceder a **enlaces adicionales** a partir de la misma.

h. Indique su nivel de satisfacción [26] para con el motor de búsqueda según su interacción con el mismo. Marque sólo una respuesta.

Nivel	√
1. Completamente satisfecho	
2. Muy satisfecho	
3. Satisfecho	
4. Poco satisfecho	
5. Insatisfecho	

¿Volvería a utilizar este motor de búsqueda? ¿Por qué?

i. Analice los documentos que considera importantes para su búsqueda, ¿cree que el orden en que fueron devueltos es el correcto, según el grado de importancia de los mismos?

Sí
No

MSN

Indique dentro de los primeros cincuenta resultados devueltos:

a. ¿Le resulta difícil[27] comprender la página de resultados? Marque sólo una respuesta.

[26] Ver glosario para definición de satisfacción.
[27] Ver glosario para definición de difícil.

Nivel	√
1. Muy difícil	
2. Difícil	
3. Normal	
4. Fácil	
5. Muy fácil	

Si seleccionó 1, 2 ó 3, indique cuáles son las dificultades que se le presentan:

b.

i. ¿Está satisfecho [28] con las primeras 50 respuestas obtenidas? Marque sólo una respuesta.

Nivel	√
1. Completamente satisfecho	
2. Muy satisfecho	
3. Satisfecho	
4. Poco satisfecho	
5. Insatisfecho	

ii. Si seleccionó 3, 4 ó 5, ¿desearía que el contenido de las mismas tuviera un mayor nivel de detalle?

Sí
No

iii. ¿Obtuvo toda la información deseada para la consulta en cuestión?

Sí
No

c. Indique la cantidad de documentos retornados.

Cantidad de respuestas = _____

[28] Ver glosario para definición de satisfecho.

d. Clasifique las primeras 50 respuestas obtenidas:

i.

Categoría	Cantidad
1. Me interesan	
2. No me interesan	
3. Inaccesibles	

ii. Indique para los documentos inaccesibles

Posición	¿Le interesa?

NOTA: Recuerde, por favor, acceder a todos los documentos que le parezcan interesantes.

e. Dentro de los primeros 50 documentos obtenidos, indique la posición del/de los documento(s) más valioso(s) según su opinión.

Posiciones = _____

f. Indique la cantidad de documentos del resumen que haya evaluado (i.e.: que haya ingresado al documento en cuestión)

Cantidad de respuestas evaluadas = ___ de 50

g. Indique la cantidad de páginas en las cuales encontró información de su interés pero sólo luego de acceder a **enlaces adicionales** a partir de la misma.

h. Indique su nivel de satisfacción [29] para con el motor de búsqueda según su interacción con el mismo. Marque sólo una respuesta.

Nivel	√
1. Completamente satisfecho	
2. Muy satisfecho	
3. Satisfecho	
4. Poco satisfecho	
5. Insatisfecho	

¿Volvería a utilizar este motor de búsqueda? ¿Por qué?

[29] Ver glosario para definición de satisfacción.

i. Analice los documentos que considera importantes para su búsqueda, ¿cree que el orden en que fueron devueltos es el correcto, según el grado de importancia de los mismos?

Sí
No

Altavista
Indique dentro de los primeros cincuenta resultados devueltos:

a. ¿Le resulta difícil [30] comprender la página de resultados? Marque sólo una respuesta.

Nivel	√
1. Muy difícil	
2. Difícil	
3. Normal	
4. Fácil	
5. Muy fácil	

Si seleccionó 1, 2 ó 3, indique cuáles son las dificultades que se le presentan:

b.
i. ¿Está satisfecho[31] con las primeras 50 respuestas obtenidas? Marque sólo una respuesta.

Nivel	√
1. Completamente satisfecho	
2. Muy satisfecho	
3. Satisfecho	
4. Poco satisfecho	
5. Insatisfecho	

ii. Si seleccionó 3, 4 ó 5, ¿desearía que el contenido de las mismas tuviera un mayor nivel de detalle?

Sí
No

[30] Ver glosario para definición de difícil.
[31] Ver glosario para definición de satisfecho.

iii. ¿Obtuvo toda la información deseada para la consulta en cuestión?

Sí
No

c. Indique la cantidad de documentos retornados.

Cantidad de respuestas = _____

d. Clasifique las primeras 50 respuestas obtenidas:

i.

Categoría	Cantidad
1. Me interesan	
2. No me interesan	
3. Inaccesibles	

ii. Indique para los documentos inaccesibles

Posición	¿Le interesa?

NOTA: Recuerde, por favor, acceder a todos los documentos que le parezcan interesantes.

e. Dentro de los primeros 50 documentos obtenidos, indique la posición del/de los documento(s) más valioso(s) según su opinión.

Posiciones = _____

f. Indique la cantidad de documentos del resumen que haya evaluado (i.e.: que haya ingresado al documento en cuestión)

Cantidad de respuestas evaluadas = ___ de 50

g. Indique la cantidad de páginas en las cuales encontró información de su interés pero sólo luego de acceder a **enlaces adicionales** a partir de la misma.

h. Indique su nivel de satisfacción[32] para con el motor de búsqueda según su interacción con el mismo. Marque sólo una respuesta.

[32] Ver glosario para definición de satisfacción.

Nivel	√
1. Completamente satisfecho	
2. Muy satisfecho	
3. Satisfecho	
4. Poco satisfecho	
5. Insatisfecho	

¿Volvería a utilizar este motor de búsqueda? ¿Por qué?

i. Analice los documentos que considera importantes para su búsqueda, ¿cree que el orden en que fueron devueltos es el correcto, según el grado de importancia de los mismos?

Sí
No

About

Indique dentro de los primeros cincuenta resultados devueltos:

a. ¿Le resulta difícil [33] comprender la página de resultados? Marque sólo una respuesta.

Nivel	√
1. Muy difícil	
2. Difícil	
3. Normal	
4. Fácil	
5. Muy fácil	

Si seleccionó 1, 2 ó 3, indique cuáles son las dificultades que se le presentan:

[33] Ver glosario para definición de difícil.

b.

i. ¿Está satisfecho [34] con las primeras 50 respuestas obtenidas? Marque sólo una respuesta.

Nivel	√
1. Completamente satisfecho	
2. Muy satisfecho	
3. Satisfecho	
4. Poco satisfecho	
5. Insatisfecho	

ii. Si seleccionó 3, 4 ó 5, ¿desearía que el contenido de las mismas tuviera un mayor nivel de detalle?

Sí
No

iii. ¿Obtuvo toda la información deseada para la consulta en cuestión?

Sí
No

c. Indique la cantidad de documentos retornados.

Cantidad de respuestas = _____

d. Clasifique las primeras 50 respuestas obtenidas:

i.

Categoría	Cantidad
1. Me interesan	
2. No me interesan	
3. Inaccesibles	

ii. Indique para los documentos inaccesibles

Posición	¿Le interesa?

[34] Ver glosario para definición de satisfecho.

e. Dentro de los primeros 50 documentos obtenidos, indique la posición del/de los documento(s) más valioso(s) según su opinión.

Posiciones = _____

f. Indique la cantidad de páginas en las cuales encontró información de su interés pero sólo luego de acceder a **enlaces adicionales** a partir de la misma.

g. Indique la cantidad de documentos del resumen que haya evaluado (i.e.: que haya ingresado al documento en cuestión)

Cantidad de respuestas evaluadas = ___ de 50

h. Indique su nivel de satisfacción [35] para con el motor de búsqueda según su interacción con el mismo. Marque sólo una respuesta.

Nivel	√
1. Completamente satisfecho	
2. Muy satisfecho	
3. Satisfecho	
4. Poco satisfecho	
5. Insatisfecho	

¿Volvería a utilizar este motor de búsqueda? ¿Por qué?

i. Analice los documentos que considera importantes para su búsqueda, ¿cree que el orden en que fueron devueltos es el correcto, según el grado de importancia de los mismos?

Sí
No

[35] Ver glosario para definición de satisfacción.

Experimento – Parte 3

4. Marque sólo una respuesta. Usted…

Categoría	√	Indique que buscador / buscadores
1. Utiliza un único buscador.		
2. Utiliza siempre el mismo buscador para un área determinada.		
3. Utiliza motores de búsqueda indistintamente.		
4. No utiliza motores de búsqueda de Internet para buscar información.		
5. No utiliza Internet.		

5. Describa las características que usted desearía estén presentes en un motor de búsqueda, teniendo en cuenta los aspectos sugeridos a continuación. Consulte con el encuestador si tiene dudas sobre este punto.

- Página de ingreso de la búsqueda
  - Botones:
  - Ayuda:
  - Lenguaje de búsqueda:
  - Elementos presentes en la página:
  - Elección del tipo de búsqueda (Web, imágenes, grupos, noticias, etc.):
  - Otros (por favor indicar):
- Página de resultados
  - Enlaces:
  - Resumen de sitios:
  - Caché:
  - Árbol de categorías para cada respuesta:
  - Traducción de sitios:
  - Otros (por favor indicar):

¡Muchas gracias por su participación en este experimento!

Glosario

Fácil:

- No tuvo dificultades para interactuar con la página de ingreso de búsquedas
- No necesitó buscar ayuda o si la buscó, fue simple encontrarla
- Los botones los encuentra rápidamente.
- Los botones están en los lugares que esperaba.
- Otro (por favor indicar): _____

Difícil:

- Tuvo problemas para interactuar con la página de resultados
- Necesitó buscar ayuda o la ayuda no fue simple de encontrar.
- No comprende rápidamente la página de resultados
- Acceder a los documentos no es intuitivo
- El resumen de documentos, de estar disponible, no es útil para decidir si el documento es de su interés o no.
- Los enlaces para acceder a páginas siguientes no siempre están en los lugares que esperaba.
- Otro (por favor indicar): _____

Satisfecho:

- Los documentos que usted deseaba obtener pertenecen al subconjunto de los primeros 50 resultados retornados.
- Usted considera que el orden de las respuestas es el apropiado.
- Los primeros 50 documentos son accesibles (i.e.: no hay enlaces rotos).
- Usted considera que el conjunto de los primeros 50 resultados es completo.
- Otro (por favor indicar): _____

Satisfacción:

- No tuvo dificultades para interactuar con la página de ingreso de búsquedas
- No necesitó buscar ayuda o si la buscó, fue simple encontrarla
- No tuvo problemas para interactuar con la página de resultados
- Acceder a los documentos es intuitivo
- El resumen de documentos esta disponible y es útil para decidir si el documento es de su interés o no.
- Los enlaces para acceder a páginas siguientes están en los lugares que esperaba.
- Los documentos que usted deseaba obtener pertenecen al subconjunto de los primeros 50 resultados retornados.
- Usted considera que el orden de las respuestas es el apropiado.
- Los primeros 50 documentos son accesibles (i.e.: no hay enlaces rotos).
- Usted considera que el conjunto de los primeros 50 resultados es completo.

APÉNDICE II. RECOLECCION DE DATOS

II.1 Introducción

El objetivo de este apéndice es presentar los datos planos recolectados durante el experimento.

II.2 Usuarios

En el experimento participaron 20 usuarios de nacionalidad argentina que utilizaron con anterioridad algún motor de búsqueda.

II.3 Experiencia Previa

La tabla II.1 ilustra la distribución de los usuarios que conocían los buscadores utilizados en el experimento con anterioridad.

M.B.	# Usuarios
Google	20
AllTheWeb	0
MSN	14
Altavista	15
About	0

Tabla II.1 - Usuarios que conocían los M.B. con anterioridad.

II.4 Confiabilidad

La tabla II.2 muestra los datos planos obtenidos con respecto a *confiabilidad*. La tabla siguiente ilustra el mapeo correspondiente a los niveles de la funcionalidad evaluada.

	Google	**Yahoo**				
1. Utiliza un único buscador.	usr1, user2, usr4, usr6, usr8, usr10, usr11, usr12, usr13, usr15, usr19, usr20	usr7, usr16				
2. Utiliza siempre el mismo buscador para un área determinada.	**Medline**	**Ubbi**	**Google**	**Altavista**	**MSN**	**Otros**
	usr3	usr5	usr5, usr14, usr17	usr14, usr17	usr17	usr18
3. Utiliza motores de búsqueda indistintamente.	**Yahoo**					
	usr9					
4. No utiliza motores de búsqueda de Internet para buscar información.	Nadie marcó esta opción					
5. No utiliza Internet.	Nadie marcó esta opción					

Tabla II.2 - Datos planos recolectados para *confiabilidad.*

	Google	AllTheWeb	MSN	Altavista	About
Usuario1	1	4	4	4	4
Usuario2	1	4	4	4	4
Usuario3	4	4	4	4	4
Usuario4	1	4	4	4	4
Usuario5	3	4	4	4	4
Usuario6	1	4	4	4	4
Usuario7	4	4	4	4	4
Usuario8	1	4	4	4	4
Usuario9	4	4	4	4	4
Usuario10	1	4	4	4	4
Usuario11	1	4	4	4	4
Usuario12	1	4	4	4	4
Usuario13	1	4	4	4	4
Usuario14	3	4	4	3	4
Usuario15	1	4	4	4	4
Usuario16	4	4	4	4	4
Usuario17	3	4	3	3	4
Usuario18	4	4	4	4	4
Usuario19	1	4	4	4	4
Usuario20	1	4	4	4	4

Tabla II.3 - *Confiabilidad* de cada buscador.

II.5 Interfaz Gráfica de un Motor de Búsqueda

Las tablas II.4 a II.8 presentan los datos recolectados con respecto a la interacción de los usuarios con la interfaz de los motores de búsqueda.

Pregunta	Usr 1	Usr 2	Usr 3	Usr 4	Usr 5	Usr 6	Usr 7	Usr 8	Usr 9	Usr 10	Usr 11	Usr 12	Usr 13	Usr 14	Usr 15	Usr 16	Usr 17	Usr 18	Usr 19	Usr 20
II - Le resulta fácil interactuar	si	si	si	si	si	si	si	si	si	si	no	si	si	si	si	si	si	si	si	si
III - Tiempo para confeccionar consulta	20m	1m	5s	1s	5s	5m	1m	1m	5m	1m	40s	15s	1s	1:10 (m:s)	1s	1m	4m	1s	1m	30s
IV - Cantidad de veces	3	2	1	2	1	3	1	2	2	1	1	2	1	2	1	2	1	1	1	2
Nivel de facilidad de realización de búsqueda	2	2	1	2	1	2	1	2	2	1	1	1	1	2	1	2	1	1	1	2

Tabla II.4 - Google – Interacción con la interfaz gráfica.

Pregunta	Usr 1	Usr 2	Usr 3	Usr 4	Usr 5	Usr 6	Usr 7	Usr 8	Usr 9	Usr 10	Usr 11	Usr 12	Usr 13	Usr 14	Usr 15	Usr 16	Usr 17	Usr 18	Usr 19	Usr 20
II - Le resulta fácil interactuar	si	si	si	si	Si	si	Si	si	no	si	no	si	si	no	si	si	si	si	si	si
III - Tiempo en confeccionar consulta	10m	3m	5s	2s	10s	11m	5m	1m	4m	1m	20s	5s	1s	20s	45s	10s	13m	1s	4m	5s
IV - Cantidad de veces	1	4	1	4	1	4	3	2	3	1	2	1	1	1	2	1	2	1	2	3
Nivel de facilidad de realización de búsqueda	1	2	1	2	1	2	2	2	2	1	2	1	1	1	2	1	2	1	2	2

Tabla II.5 AllTheWeb – Interacción con la interfaz gráfica.

Pregunta	Usr 1	Usr 2	Usr 3	Usr 4	Usr 5	Usr 6	Usr 7	Usr 8	Usr 9	Usr 10	Usr 11	Usr 12	Usr 13	Usr 14	Usr 15	Usr 16	Usr 17	Usr 18	Usr 19	Usr 20
II - Le resulta fácil interactuar	si	si	si	no	si	si	Si	si	no	Si	si	si	si	no	no	si	si	si	si	si
III - Tiempo en confeccionar consulta	5m	6m	2m	10s	10s	2m	1m	5m	10m	2m	10s	15s	1s	10s	10m	1m	5m	1	2m	1m
IV - Cantidad de veces	1	3	3	3	1	1	1	5	3	3	1	1	1	1	9	2	1	1	1	3
Nivel de facilidad de realización de búsqueda	1	2	2	2	1	1	1	2	2	2	1	1	1	1	3	2	1	1	1	2

Tabla II.6 - MSN – Interacción con la interfaz gráfica.

Pregunta	Usr 1	Usr 2	Usr 3	Usr 4	Usr 5	Usr 6	Usr 7	Usr 8	Usr 9	Usr 10	Usr 11	Usr 12	Usr 13	Usr 14	Usr 15	Usr 16	Usr 17	Usr 18	Usr 19	Usr 20
II - Le resulta fácil interactuar	no	si	si	no	si	si	si	si	si	si	si	si	no	si	no	si	si	si	si	Si
III - Tiempo en confeccionar consulta	30m	1m	5s	10s	10s	4m	2m	2m	2m	2m	1m	5s	1m	03:30	1m	30s	14m	1	1m	30s
IV - Cantidad de veces	2	1	1	3	2	2	2	2	1	4	3	1	6	3	3	1	4	1	1	(+ de 5)
Nivel de facilidad de realización de búsqueda	2	1	1	2	2	2	2	2	1	2	2	1	3	3	2	1	2	1	1	3

Tabla II.7 - Altavista – Interacción con la interfaz gráfica.

Pregunta	Usr 1	Usr 2	Usr 3	Usr 4	Usr 5	Usr 6	Usr 7	Usr 8	Usr 9	Usr 10	Usr 11	Usr 12	Usr 13	Usr 14	Usr 15	Usr 16	Usr 17	Usr 18	Usr 19	Usr 20
II - Le resulta fácil interactuar	si	si	si	no	Si	si	si	no	no	si	si	si	si	no	no	si	no	si	si	no
III - Tiempo en confeccionar consulta	5m	10m	2m	5m	10s	3m	10m	7m	5m	2m	20s	5s	1s	30s	3m	30s	7m	1s	1 s	5s
IV - Cantidad de veces	1	9	2	7	2	1	10	6	2	4	1	1	1	1	7	1	2	1	1	3
Nivel de facilidad de realización de búsqueda	1	3	2	3	2	1	3	3	2	2	1	1	1	1	3	1	2	1	1	2

Tabla II.8 - About – Interacción con la interfaz gráfica.

II.6 Interfaz de Resultados de un Motor de Búsqueda

Las tablas II.9 a II.13 presentan los datos recolectados sobre los resultados obtenidos en los buscadores.

Referencia	Descripción
F	Fácil
MF	Muy Fácil
N	Normal
S	Satisfecho
CS	Completamente satisfecho
MS	Muy satisfecho
PS	Poco satisfecho
I	Insatisfecho
D	Difícil
MD	Muy Difícil

Pregunta	Usr 1	Usr 2	Usr 3	Usr 4	Usr 5	Usr 6	Usr 7	Usr 8	Usr 9	Usr 10	Usr 11	Usr 12	Usr 13	Usr 14	Usr 15	Usr 16	Usr 17	Usr 18	Usr 19	Usr 20
a - comprender la página de resultados	4 - F	5 - MF	3 - N	4 - F	5 - MF	5 - MF	3 - N	4 - F	5 - MF	5 - MF	5 - F	4 - F	5 - MF	4 - F	3 - N	4 - F	4 - F	4 - F	3 - N	4 - F
bi -Satisfecho con las respuestas	3 - S	1 - CS	3 - S	3 - S	1 - CS	3 - S	3 - S	3 - S	1 - CS	2 - MS	2 - MS	2 - MS	1 - CS	3 - S	3 - S	2 - MS	5 - I	1 - CS	3 - S	5 - I
bii – Mayor detalle	SI	NO	SI	SI	NO	NO	SI	SI	NO	NO	NO	NO	NO	NO	SI	NO	SI	NO	SI	SI
biii - Obtuvo la información deseada	NO	SI	NO	NO	SI	SI	NO	SI	SI	SI	SI	SI	SI	SI	SI	SI	NO	SI	SI	NO
d - # doc interesantes	4	4	1	4	1	4	2	2	7	1	12	20	9	16	2	8	1	23	17	0
e - Posiciones doc interesantes	1,3,4,8	1,2,3,13	1	10,11,13,14	1	4,2,5,9	13,28	2,3	3,13,15,17,18,20,21	9	1,2,3,4,10,15,18	1,2,4,14,30,32,33	1,2,3,6,7,8,9,10,24,	1 AL 15	3,4	3,4,6,8,10,11,12,20	2	13,14,16,17,18,21,22,23,28,29,32,35,36,38,40,41,42,43,4,6,7,8,10	3,4,5,6,9,10,12,14,16,18,23,25,27,35,39,40,45	0
d - # doc inaccesibles	0	0	0	0	0	1	0	0	0	0	1	0	0	11	0	0	1	2	2	0
d - # doc no interesantes	46	46	49	46	49	45	48	48	43	49	37	30	41	23	48	42	48	25	31	50
h - Satisfacción MB	3 - S	1 - CS	3 - S	3 - S	1 - CS	3 - S	2 – MS	3 -S	1 - CS	2 - MS	1 - CS	2 - MS	1 - CS	2 - MS	2 - MS	1 - CS	5 - I	1 - CS	3 - S	5 - i
hi - Volvería a usarlo	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI
i – Orden	SI	SI	NO	SI	SI	SI	NO	SI	NO	NO	NO	SI	SI	SI	NO	NO	SI	SI	NO	NO
c - # respuestas	12100	18700	94	164	43400	53	806	457	178	100	8380	4250	3450	4540	51300	22500	25	44200	2850	25400
g – Enlaces adicionales	0	1	0	3	0	0	0	3	0	3	2	0	0	2	5	-	-	-	-	-
dii – Poción doc inaccesibles	0	0	0	0	0	1	0	0	0	0	11	-	-	42,33,39	-	-	5	3,48	-	-
dii - Le interesa?	0	0	0	0	0	SI	0	0	0	0	SI	-	-	NO, NO, SI	-	-	SI	NO,NO	-	-

Nivel de Exactitud	4	3	4	4	2	3	4	3	3	3	3	4	3	2	3	3	5	3	3	7
Identificación eficiente de documentos relevante	1	2	2	2	1	1	3	1	3	2	3	2	2	3	2	2	1	2	3	3
Nivel de Disponibilidad	1	1	1	1	1	3	1	1	1	1	3	1	1	3	1	1	3	2	2	1

Tabla II.9 - Google – Resultados de búsquedas.

Pregunta	Usr 1	Usr 2	Usr 3	Usr 4	Usr 5	Usr 6	Usr 7	Usr 8	Usr 9	Usr 10	Usr 11	Usr 12	Usr 13	Usr 14	Usr 15	Usr 16	Usr 17	Usr 18	Usr 19	Usr 20
a - comprender la página de resultados	3 - N	5 - MF	3 -N	4 - F	4 - F	n/a	4 - F	4 - F	5 - MF	3 - N	4 - F	4 - F	5 - MF	3 - N	3 - N	4 - F	4 - F	4 - F	3 - N	2 - d
bi -Satisfecho con las respuestas	4 - PS	4 - PS	3 - S	3 - S	3 - S	5 - I	5 - I	4 - PS	1 - CS	3 - S	4 - PS	4 - PS	5 - I	2 - MS	5 - I	4 - PS	5 - i	1 - CS	3 - S	5 - I
bii - Mayor detalle	SI	SI	SI	NO	SI	SI	SI	NO	NO	SI	SI	NO	NO	NO	SI	NO	SI	NO	NO	SI
biii - Obtuvo la información deseada	NO	SI	NO	SI	SI	NO	NO	SI	SI	SI	NO	SI	NO	SI	NO	NO	NO	SI	SI	NO
d - # doc interesantes	4	2	1	2	1	0	1	2	13	1	2	10	0	11	1	3	0	29	15	0
e - Posiciones doc interesantes	1,5,6,8	4,5	10	8, 11	3	0	34	18,48	4,6,7,8,10,13,20,23,27,29,30,35,41	8	14,1	25,20,1,6,9,13	0	3,4,7,9,12,13	24	3,6	0	1,2,3,4,5,6,7,8,10,11,15,16,18,20,22,23,24,25,26,29,31,34,35,38,39,40,42,43,47	1,5,11,14,16,17,19,20,22,27,28,38	0
d - # doc inaccesibles	6	1	0	1	0	0	0	0	0	0	0	0	0	15	0	1	0	2	2	3
d - # doc no interesantes	40	47	49	47	49	0	49	48	37	49	48	40	50	24	49	46	4	19	33	47
h - Satisfacción MB	4 - PS	4 - PS	4 - PS	4 - PS	3 - S	5 - I	4 - PS	4 - PS	2 - MS	3 - S	3 - S	4 - PS	4 - PS	3 - S	5 - I	4 - PS	5 - I	1 - CS	3 - S	5 - i

hi - Volvería a usarlo	NO	NO	NO	SI	NO	NO	NO	NO	SI	NO	NO	NO	NO	NO	NO	NO	NO	SI	SI	SI
i - Orden	NO	NO	NO	SI	NO	NO	SI	NO	SI	NO	NO	NO	NO	NO	NO	NO	SI	SI	NO	NO
c - # respuestas	30123	393078	17	109	104343	0	829	208	60	100	58	2287	97580	9109	143744	33932	4	98866	7423	24700
g - Enlaces adicionales	0	0	0	0	0	0	0	0	0	2	1	2	0	1	1	2	-	-	-	-
dii - Poción doc inaccesibles	1,8,33	9	0	1	0	0	0	0	0	0	-	-	-	22,39,49	-	8	-	27, 28	-	10,18,20
dii - Le interesa?	NO	SI	0	SI	0	0	0	0	0	0	-	-	-	NO, SI,NO	-	SI	-	NO, SI	-	NO,NO,NO
Nivel de Exactitud	4	5	4	3	3	6	4	3	3	3	4	5	7	3	5	5	7	2	3	7
Ident. eficiente de docs. relevante	2	2	2	2	2	4	2	3	1	2	3	3	3	3	3	2	2	1	3	4
Nivel de Disponibilidad	2	4	1	4	1	1	1	1	1	1	1	1	1	4	1	4	1	4	2	2

Tabla II.10 - AllTheWeb – Resultados de búsquedas.

Pregunta	Usr 1	Usr 2	Usr 3	Usr 4	Usr 5	Usr 6	Usr 7	Usr 8	Usr 9	Usr 10	Usr 11	Usr 12	Usr 13	Usr 14	Usr 15	Usr 16	Usr 17	Usr 18	Usr 19	Usr 20
a - comprender la página de resultados	3 - N	5 - MF	3 - N	4 - F	5 - MF	4 - F	3 - N	4 - F	2 - D	3 - N	4 - F	4 - F	5 - MF	5 - MF	2 - MD	4 - F	4 - F	4 - F	3 - N	4 - F
bi -Satisfecho con las respuestas	4 - PS	3 - S	3 - S	4 - PS	1 - CS	3 - S	4 - PS	3 - S	5 - I	4 - PS	4 - PS	3 - S	2 - MS	2 - MS	3 - S	5 - I	5 - I	4 - PS	3 - S	5 - I
bii - Mayor detalle	SI	SI	SI	NO	NO	SI	SI	NO	SI	SI	SI	NO	NO	NO	SI	NO	NO	NO	NO	SI
biii - Obtuvo la información deseada	NO	SI	NO	SI	SI	NO	NO	SI	NO	NO	SI	SI	SI	SI	SI	NO	SI	NO	SI	NO
d - # doc interesantes	10	7	1	2	1	2	2	7	0	0	5	21	6	34	3	1	11	20	16	0
e - Posiciones doc interesantes	2,3	2,4,5	6	11 , 37	3	5	1 , 22	2 , 3 , 38	0	13	1,3,4 ,13,1 4	1,2,3 ,4,5, 6,9,1 0,11	5,6,7 ,9,12 ,15	1 AL 19	1,27, 34	5	6,8,1 0,14, 17	2,10,12,14,16,18, 19,22,23,24,25,2 8,31,38,43,44,45, 48,49,50	1,3,4,7,10,1 1,15,17,21,2 2,23,32,40,4 6,47	0
d - # doc inaccesibles	5	2	0	0	0	0	0	0	0	0	0	1	0	6	1	1	2	2	1	0
d - # doc no interesantes	35	41	49	48	49	24	48	43	1	39	10	28	44	10	45	49	37	28	33	50
h - Satisfacción MB	4 - PS	3 - S	5 - I	4 - PS	2 - MS	2 - MS	4 - PS	3 - S	5 - I	4 - PS	3 - S	3 - S	2 - MS	1 - CS	3 - S	5 - I	2 - MS	4 - PS	3 - S	4 - PS
hi - Volvería a usarlo	NO	SI	NO /SI	NO / SI	SI	NO	NO	SI	NO	NO	SI	NO	SI	SI	NO	NO	SI	NO	SI	SI
i - Orden	NO	SI	NO	NO	NO	NO	SI	NO	NO	NO	NO	NO	NO	SI	NO	NO	SI	NO	NO	NO
c - # respuestas	3770	59	17	56	1872 5	26	196	181	1	39	15	662	2077	1304	3068 3	8866	74	24582	1781	29955
g - Enlaces adicionales	-	-	-	1	-	-	-	-	-	-	1	-	2	-	2	-	1	-	-	-
dii - Poción doc inaccesibles	-	16-19	-	-	-	-	-	-	-	-	-	11	-	20,2 8,38	20	5	7,16	6,42	-	-
dii - Le interesa?	-	NO-NO	-	-	-	-	-	-	-	-	-	NO	-	NO, NO, NO	SI	SI	SI,SI	SI,SI	-	-
Nivel de	5	3	4	5	3	5	5	3	7	7	5	3	3	2	3	5	3	4	3	7

Exactitud																				
Identificación eficiente de documentos relevante	2	1	2	3	2	2	2	3	4	3	3	3	3	1	4	2	2	3	3	3
Nivel de Disponibilidad	2	2	1	1	1	1	1	1	1	1	1	2	1	2	4	4	4	4	2	1

Tabla II.11 - MSN – Resultados de búsquedas.

Pregunta	Usr 1	Usr 2	Usr 3	Usr 4	Usr 5	Usr 6	Usr 7	Usr 8	Usr 9	Usr 10	Usr 11	Usr 12	Usr 13	Usr 14	Usr 15	Usr 16	Usr 17	Usr 18	Usr 19	Usr 20
a - comprender la página de resultados	3 - N	4 - F	5 - MD	4 - F	4 - F	4 - F	3 – N	4 - F	5 - MF	4 - F	4 - F	4 - F	5 – MF	2 - D	3 - N	4 - F	4 - F	4 - F	3 - N	2 -D
bi -Satisfecho con las respuestas	3 - S	3 - S	3 - S	4 - PS	1 - CS	3 - S	4 - PS	3 - S	1 - CS	3 - S	5 - I	3 - S	5 - I	4 - PS	3 - S	2 - MS	2 - MS	4 - PS	3 - S	5 - I
bii - Mayor detalle	SI	SI	SI	NO	SI	NO	SI	NO	NO	NO	SI	NO	NO	SI	SI	NO	NO	NO	NO	SI
biii - Obtuvo la información deseada	SI	SI	SI	SI	SI	SI	NO	SI	SI	SI	NO	SI	NO	NO	SI	NO	SI	SI	SI	NO
d - # doc interesantes	12	6	3	2	1	6	2	9	10	1	0	35	0	2	1	12	10	12	11	0
e - Posiciones doc interesantes	1,2	11, 14, 3	4,8,14	2,8	1	1, 7 , 11	6 ,16	1, 2, 8	1,3,5 ,17,19,21, 31,42	10	0	44,3,5,6,7,9, 10,12,14,17, 18,19,20,21, 22,23,24,27, 28,29,30,36, 39,40,47,48, 50	0	2	7	1,2,3,4, 5,6,7,9, 11,12,16,17,	1,7,14, 16	2,3,4,5, 9,13,18 ,24,25, 38,41,44	1,2,7,9, 11,14,15,16,18, 31,46	0
d - # doc inaccesibles	8	0	2	0	0	0	0	0	0	0	0	2	0	11	0	0	3	3	4	0
d - # doc no interesantes	30	44	45	48	49	18	48	41	40	49	4	13	50	37	49	38	37	35	35	50
h - Satisfacción MB	3 - S	3 - S	3 - S	4 - PS	1 - CS	3 - S	4 - PS	3 - S	4 - CS	3 - S	3 - S	3 - S	5 - I	5 - I	3 - S	2 - MS	2 - MS	5 - I	3 - S	5 - I

hi - Volvería a usarlo	SI	NO	SI	NO	SI	NO	NO	SI	SI	NO	NO	SI	NO	NO	NO	SI	SI	NO	SI	NO
i - Orden	SI	NO	NO	SI	SI	SI	SI	SI	SI	NO	NO	SI	NO	NO	NO	NO	SI	NO	NO	NO
c - # respuestas	5033	51367	31	48	3624	24	337	204	151	100	4	4010	82647	1722	9059	1656	73	16158	759	11200
g - Enlaces adicionales	-	1	-	1	-	-	-	-	4	3	0	5	0	1	4	-	1	-	-	-
dii - Poción doc inaccesibles	-	-	9 - 12 -	-	-	-	-	-	-	-	-	11,35	-	38,26,32	-	-	3,11,9	33,45,39	-	-
dii - Le interesa?	-	-	SI-NO	-	-	-	-	-	-	-	-	SI,NO	-	NO,NO,NO	-	-	SI,SI,NO	NO,NO,SI	-	-
Nivel de Exactitud	2	3	3	5	2	3	5	3	3	3	7	3	7	5	3	4	3	5	3	7
Ident. eficiente de docs. relevante	1	3	4	1	1	2	2	1	2	2	3	2	3	3	2	3	2	3	3	4
Nivel de Disponibilidad	2	1	4	1	1	1	1	1	1	1	1	4	1	2	1	1	4	4	2	1

Tabla II.12 - Altavista – Resultados de búsquedas.

Pregunta	Usr 1	Usr 2	Usr 3	Usr 4	Usr 5	Usr 6	Usr 7	Usr 8	Usr 9	Usr 10	Usr 11	Usr 12	Usr 13	Usr 14	Usr 15	Usr 16	Usr 17	Usr 18	Usr 19	Usr 20
a - comprender la página de resultados	1 - MD	5 - MF	3 - N	4 - F	5 - MF	N/A	5 – MF	2 - D	3 - F	4 - F	4 - F	4 - F	5 - MF	5 - MF	2 - D	4 - F	3 - N	4 - F	3 - N	2 -D
bi -Satisfecho con las respuestas	5 - I	4 - PS	5 - I	5 - I	1 - CS	5 - I	3 – S	3 - S	4 - PS	3 - S	5 - I	5 - I	1 - CS	5 - I	5 - I	5 - I	5 - I	5 - I	5 - I	5 - I
bii - Mayor detalle	SI	SI	SI	NO	NO	SI	SI	SI	SI	SI	SI	SI	NO	NO	NO	NO	SI	NO	SI	SI
biii - Obtuvo la información deseada	NO	NO	NO	NO	SI	NO	NO	SI	NO	NO	NO	NO	SI	SI	NO	NO	NO	SI	NO	NO
d - # doc interesantes	0	2	0	0	1	0	3	8	1	0	1	3	3	2	3	1	0	14	0	0
e - Posiciones doc	0	1, 11	0	0	2	0	5,9,3	2,3,3	3	0	13	2,3,4	5,6,9	1,2	1,24	2	0	9,10,11,17,2	0	0

interesantes							6	3										5,27,28,30,31,35,39,40,41,54		
d - # doc inaccesibles	0	3	0	0	0	0	0	0	0	0	0	0	0	48	1	0	0	2	0	0
d - # doc no interesantes	50	45	4	4	49	50	47	42	2	50	49	3	47	0	46	49	2	34	4	50
h - Satisfacción MB	5 - I	5 -I	5 - I	5 -I	3 - S	I - 5	2 – MS	3 - S	4 - PS	4 - PS	4 - PS	3 - S	3 - S	5 - I	5 - I	5 - I	5 - I	5 - I	5 - I	5 - I
hi - Volvería a usarlo	NO	NO	NO	NO	NO	NO	SI	NO	NO	NO	NO	NO	SI	NO	NO	NO	NO	NO	NO	NO
i - Orden	NO	NO	NO	NO	SI	NO	SI	SI	NO	NO	NO	SI	SI	SI	NO	NO	NO	NO	NO	NO
c - # respuestas	150	72355	4	4	139	0	111	153	3	100	60	6	1477	110	10000	8666	2	18508	4	62
g - Enlaces adicionales	-	2	-	-	-	-	-	-	2	10	1	6	2	-	2	1	-	-	-	-
dii - Poción doc inaccesibles	-	6,12,17	-	-	-	-	-	-	-	-	-	-	-	3 AL 50	17	-	-	21,32	-	-
dii - Le interesa?	-	SI,NO,NO	-	-	-	-	-	-	-	-	-	-	-	NO	SI	-	-	SI,SI	-	-
Nivel de Exactitud	7	5	7	7	3	6	4	3	5	7	5	5	3	5	5	5	7	5	7	7
Ident. eficiente de docs. relevante	4	3	3	3	1	4	2	3	2	3	3	1	1	1	4	2	3	3	3	4
Nivel de Disponibilidad	1	4	1	1	1	1	1	1	1	1	1	1	1	2	4	1	1	4	1	1

Tabla II.13 - About – Resultados de búsquedas.

APÉNDICE III. INTERFACES DE MOTORES DE BUSQUEDA

Sinopsis Comparativa

Google - http://www.google.com

GUI de Búsqueda

GUI de Búsqueda Avanzada

GUI de Resultados

Google Web Images Groups News Froogle Local more »

Logia Lautaro Integrantes 1810 1880 masoneria Search Advanced Search Preferences

Web Results **1** - **5** of **5** for **Logia Lautaro Integrantes 1810 1880 masoneria Fundacion Iglesia Politica 1760**. (**0.46** seconds)

[PDF] La **Iglesia** como lugar de contienda en los mundos colonial y **...**
File Format: PDF/Adobe Acrobat - View as HTML
... en la **Iglesia** del Río de la Plata (**1760-1810**). **...** generó un duro conflicto entre la **Iglesia** y el **...** de los maestros artesanos se casaran con **integrantes** de su **...**
www.er-saguier.org/pdfs/TomosAll/Tomo_08all.pdf - Supplemental Result - Similar pages

[PDF] Estenssoro, Juan Carlos (1993): "La pIstica colonial y sus **...**
File Format: PDF/Adobe Acrobat - View as HTML
... Néstor (1993): Historia de la **Iglesia** en Santiago **...** emergencia y desarrollo del Estadoargentino (**1810**-1930)", en **...** de la democracia argentina, **1880**-1930", Anuario **...**
www.er-saguier.org/pdfs/Tomo_14/0-XIV-BIB--A-Z.pdf - Supplemental Result - Similar pages

[PDF] Autor: Historiador Jorge Núñez INTRODUCCION Si hubo un período de **...**
File Format: PDF/Adobe Acrobat - View as HTML
entre ellas Cádiz, donde funcionaba la **Logia Lautaro**, de tan importante **...** organización **politica** del Estado hasta las relaciones con la **Iglesia** y desde los **...**
edufuturo.com/imageBDE/EF/6905.EcuadorSigloXIX.pdf - Similar pages

[PDF] AZ
File Format: PDF/Adobe Acrobat - View as HTML
1861: Inauguración del nuevo reloj de la **Iglesia** Matriz. **...** del Ateneo de Montevideo contribuyó a la **fundación** de la Federación Feminista pro Paz **...**
www.montevideo.gub.uy/nomenclatura/nomenclatura1.pdf - Similar pages

0##### 1##### 2##### 3##### 4##### 5##### 6##### 7##### 8##### 9 **...** - [Translate this page]
Jorge Litvak, los nuevos **integrantes** de esta entidad iniciaron **...** de la **Masonería** en Chile#13-05-2003#/undin2/actuales/noti3016.shtml##La **Logia** Fraternal **...**
www.uchile.cl/undin2/infotema - 979k - Cached - Similar pages

Logia Lautaro Integrantes 1810 1880 I Search

Search within results | Language Tools | Search Tips | Dissatisfied? Help us improve

AllTheWeb - http://www.alltheweb.com

GUI de Búsqueda

GUI de Búsqueda Avanzada

Domain Filters - Filter results from specific domains (com, gov, dell.com, etc.)

Include results from

Exclude results from

Find results from a specific geographic region

Only find results from Any

File format - Only find results that are Any format

Date - Only find results updated ☐ after 1 January 1980 ☐ before 26 February 2005

Presentation - Display 10 results per page

SEARCH

Save Settings | Clear Settings | Load Saved

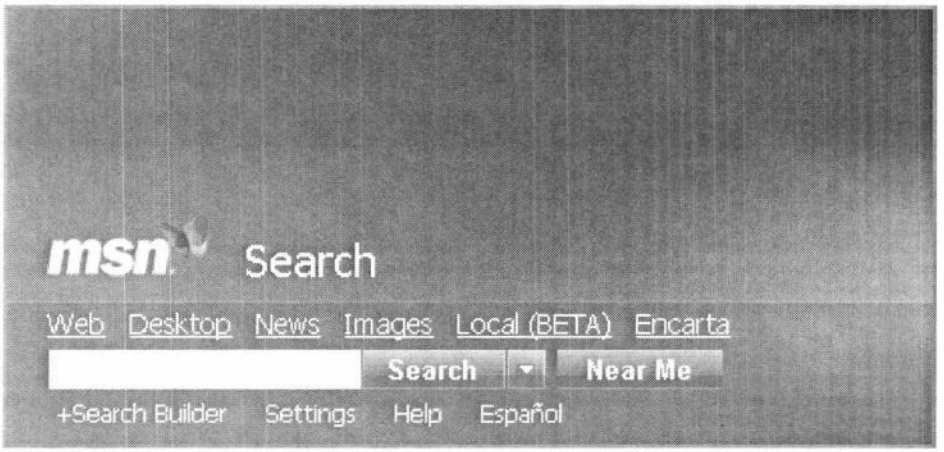
OPERA compliant

Home :: Submit Site :: About Us :: Help

GUI de Resultados

MSN Search – http://search.msn.com

GUI de Búsqueda

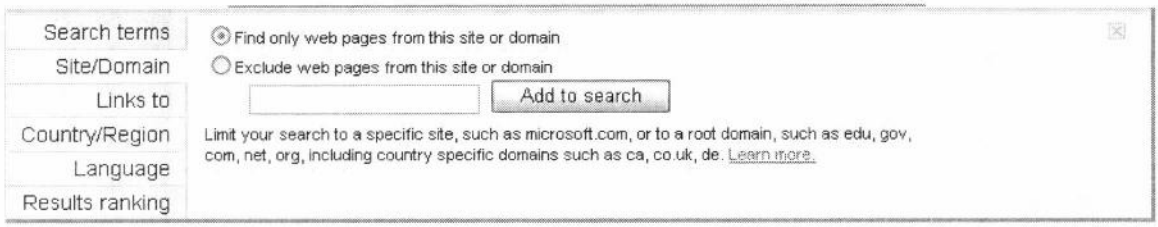


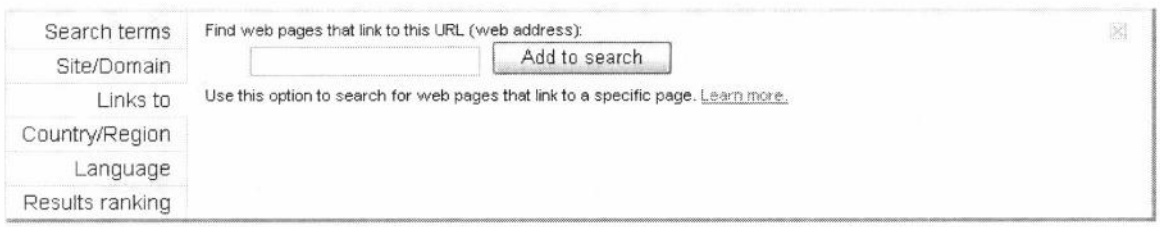
MSN Home · My MSN · Hotmail · Messenger · About MSN Search

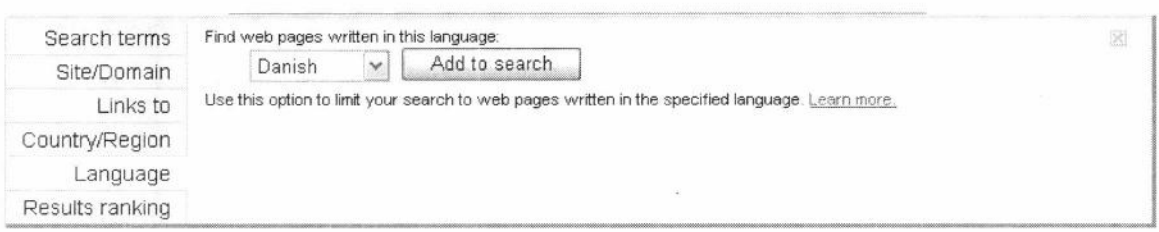
Make MSN Search your homepage

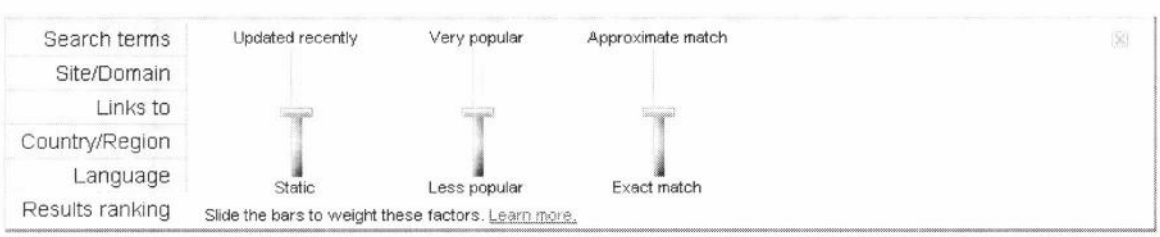
© 2005 Microsoft. MSN Privacy

GUI de Búsqueda Avanzada









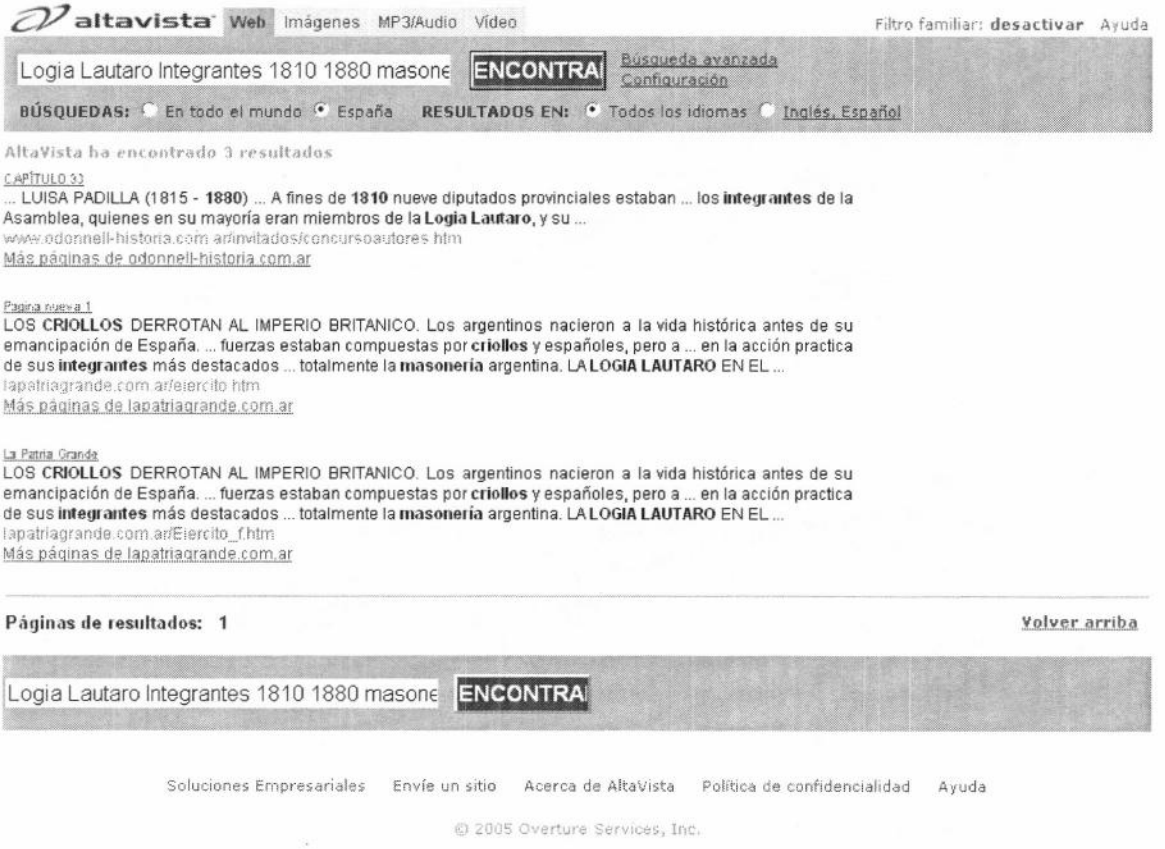
GUI de Resultados

Altavista – http://www.altavista.com

GUI de Búsqueda

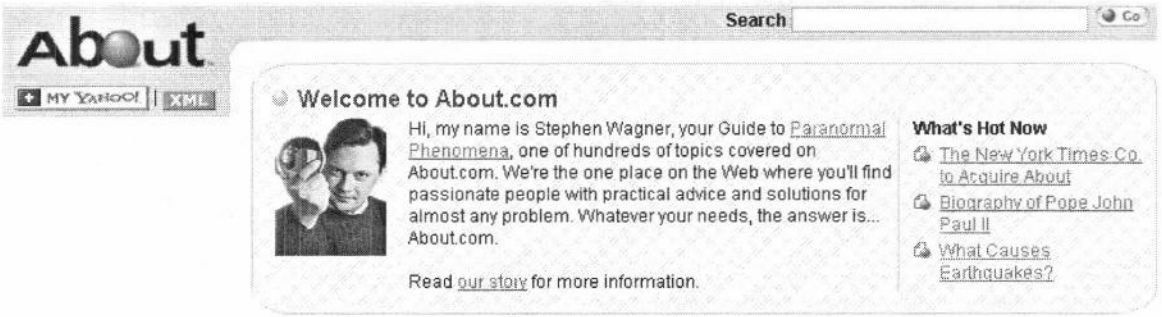
GUI de Búsqueda Avanzada

GUI de Resultados

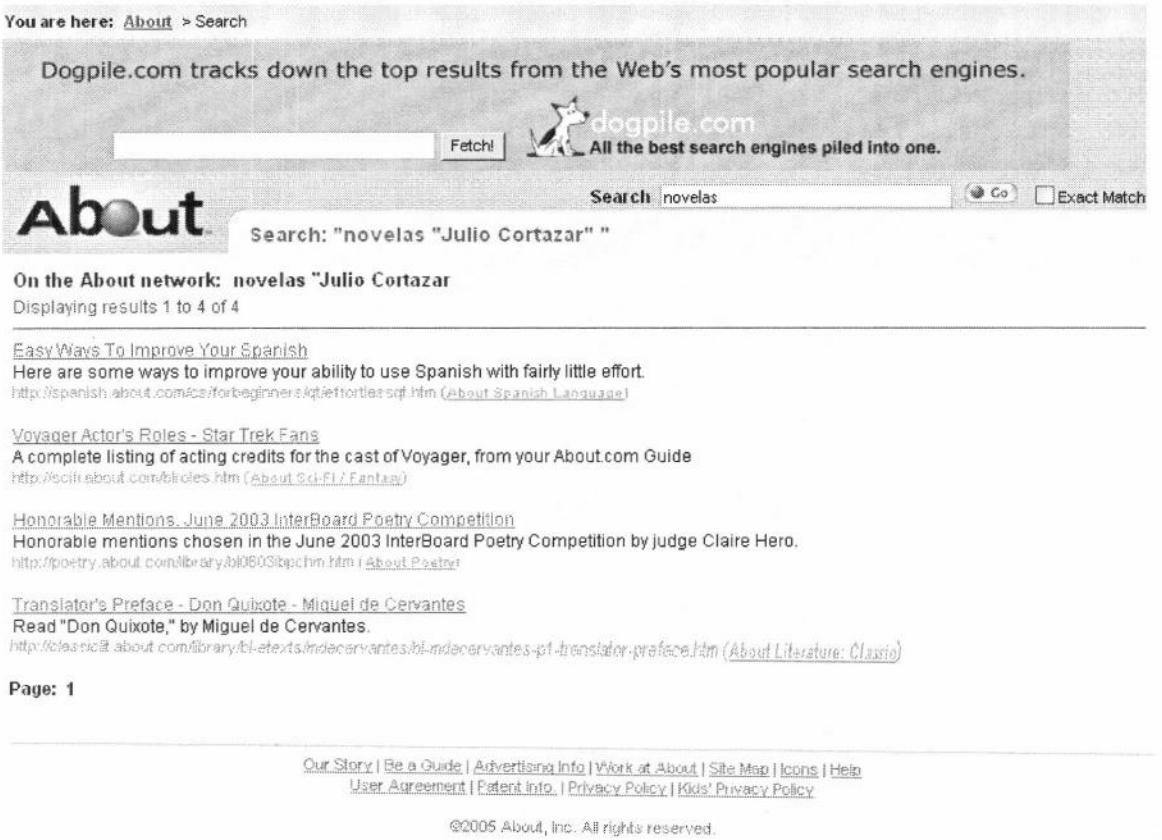


About – http://www.about.com

GUI de Búsqueda



GUI de Resultados



APÉNDICE IV. RECUPERACION DE LA INFORMACION

Introducción

Este apéndice abarca los conceptos principales de recuperación de la información, una ciencia que se dedica a la representación, almacenamiento, organización y acceso a la información. Además, ayuda a comprender varios términos utilizados en esta tesis.

¿Qué es la Recuperación de la Información?

El objetivo principal de un sistema de Recuperación de la Información (R.I.) es recuperar información útil y relevante para el usuario. Es importante tener en cuenta que recuperar información es distinto de recuperar datos, ya que un sistema de recuperación de datos tiene como objetivo obtener de un conjunto de documentos, aquellos que contienen las palabras clave buscadas por el usuario. Por lo tanto, un error en uno de los documentos recuperados implica que el sistema ha fallado. Otra característica muy importante es que los sistemas de recuperación de datos cuentan con datos estructurados y con una semántica bien definida. Por otra parte, un sistema de R.I. tiene como objetivo obtener información relevante sobre un tema. Obtener documentos inexactos no se considera necesariamente una falla. Estas diferencias se deben a que en los sistemas de R.I. se utiliza generalmente el lenguaje natural, el cual no posee siempre una estructura bien definida y la semántica es ambigua [BR99].

Al mismo tiempo, un sistema de R.I. tiene que satisfacer la necesidad de información del usuario, interpretando el contenido de los documentos y devolviéndolos en orden de acuerdo al grado de relevancia de los resultados con respecto a la consulta del usuario. Interpretar un documento implica analizarlo sintáctica y semánticamente de manera de poder realizar una correspondencia con la necesidad de información del usuario. El problema es cómo decidir si un documento es relevante. Para resolver este problema, R.I. utiliza distintos modelos de recuperación de la información, los cuales son introducidos en la sección IV.4.

La relevancia es el punto más importante de R.I. ya que su objetivo principal es recuperar todos los documentos que sean relevantes a la consulta del usuario e idealmente sólo devolver documentos relevantes (i.e.: no retornar documentos no deseados entre las respuestas).

En el pasado, el área de recuperación de la información estaba ligada mayormente a los bibliotecarios y expertos en información, pero a comienzos de 1990 hubo un cambio con la introducción de la World Wide Web. Encontrar información útil en la Web no es una tarea simple o fácil, ya que la ausencia de un modelo de datos bien definido colabora para que la información y la estructura de la misma sean generalmente de baja calidad. Estos problemas atrajeron interés por R.I.

Desde el punto de vista de Ciencias de la Computación, el problema de R.I. consiste en construir índices eficientes, lograr un alto desempeño en el procesamiento de las consultas del usuario y desarrollar algoritmos que ordenen los documentos y que mejoren la calidad del conjunto de respuesta. Desde otros puntos de vista, el problema de R.I. consiste en estudiar el comportamiento de los usuarios, entender sus necesidades y

determinar cómo éstas afectan la organización y estructura del sistema de R.I. [BR99].

Este apéndice se concentra en R.I. desde el punto de vista de las Ciencias de la Computación.

El Proceso de Recuperación de la Información

Antes de iniciar el proceso de recuperación de información, se deben definir:

a) los documentos a utilizar
b) las operaciones a realizar sobre el texto (las cuales transforman el documento original en una vista lógica del mismo)
c) el modelo del texto (la estructura del texto y que elementos pueden ser recuperados)

Luego se construye un índice para el texto. Un índice es una estructura de datos que permite realizar búsquedas sobre un gran volumen de información (ver figura IV.1 [BR99]).

Una vez indexados los documentos, se puede comenzar el proceso de recuperación. Primero el usuario especifica lo que necesita mediante una consulta, la cual es analizada gramaticalmente y transformada por el módulo de "operaciones de texto". Luego, la consulta es procesada para obtener los documentos relevantes. Este proceso de búsqueda se realiza rápidamente debido a que el índice ha sido construido con anterioridad. A continuación, el conjunto de documentos es ordenado según el grado de relevancia de sus elementos y presentado al usuario. Una vez que el usuario selecciona el conjunto de documentos que le resulta interesante, comienza el ciclo opcional de retroalimentación del usuario (*feedback*). Este ciclo consiste en reformular la consulta basándose en la selección del usuario entre los documentos recuperados. De esta manera, es posible obtener una representación más adecuada de las necesidades del usuario.

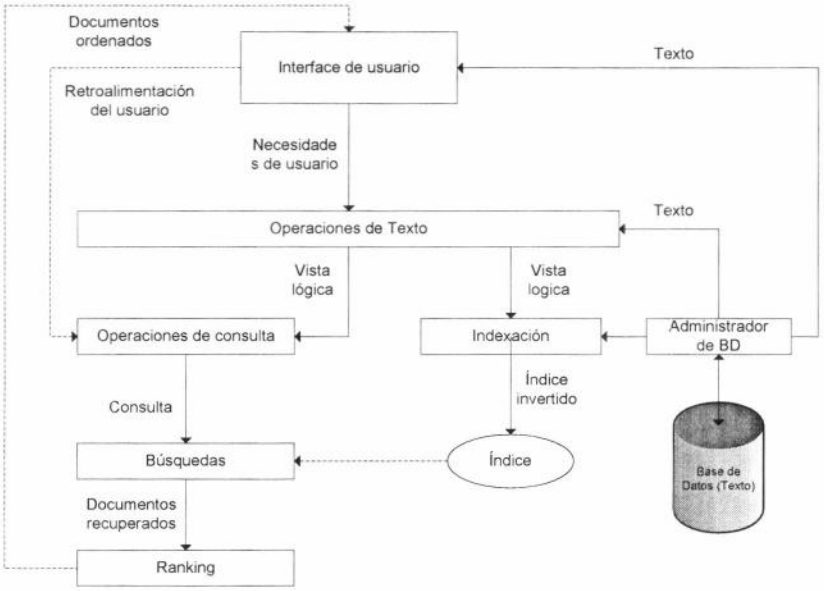


Figura IV.1 – El proceso de recuperación de la información.

Uno de los problemas comunes a los sistemas de R.I. son sus interfaces, ya que generalmente el usuario no puede especificar la información que necesita, sino que se lo obliga a realizar una consulta que debe tener la representación específica del sistema de R.I. Dado que la mayoría de los usuarios desconoce el lenguaje de consulta, éstas resultan inadecuadas y no representan con exactitud y precisión la necesidad del usuario. Por lo tanto, no sorprende que al realizar una búsqueda con una consulta inadecuada, se obtenga un conjunto de documentos poco relevantes. Esto es lo que sucede muy a menudo en la Web [BR99].

Otro problema de R.I. consiste en predecir qué documentos serán relevantes y cuáles no. Esta decisión depende del algoritmo que realiza el ranking, el cual establece un orden dentro del conjunto de documentos recuperados. Los documentos que se encuentran en una posición alta (de acuerdo al orden) tienen mayor probabilidad de resultar relevantes. El algoritmo de ranking trabaja con premisas relativas a la noción de documentos relevantes. Distintos conjuntos de premisas utilizan distintos modelos de recuperación de la información. La elección del modelo determina cómo se realiza la predicción de si un documento es relevante o no, es decir que la noción de relevancia esta dada por la implementación del sistema.

Modelos

Como se mencionó en el párrafo anterior, los modelos de recuperación de la información establecen cómo se determina la relevancia de los documentos. A continuación, se introducen los modelos clásicos de R.I.

Taxonomía de los Modelos de Recuperación de la Información

Los modelos clásicos de R.I. son el booleano, el vectorial y el probabilístico.

En el modelo booleano, los documentos son representados como conjuntos. En el modelo vectorial los documentos y las consultas son representados por vectores de t dimensiones. Y en el modelo probabilístico, los documentos y las consultas son representados basándose en la teoría de probabilidades.

Otros modelos, como los estructurados, hacen referencia a la estructura del texto. Estos modelos son: listas no solapadas y nodos proximales (ver figura IV.2).

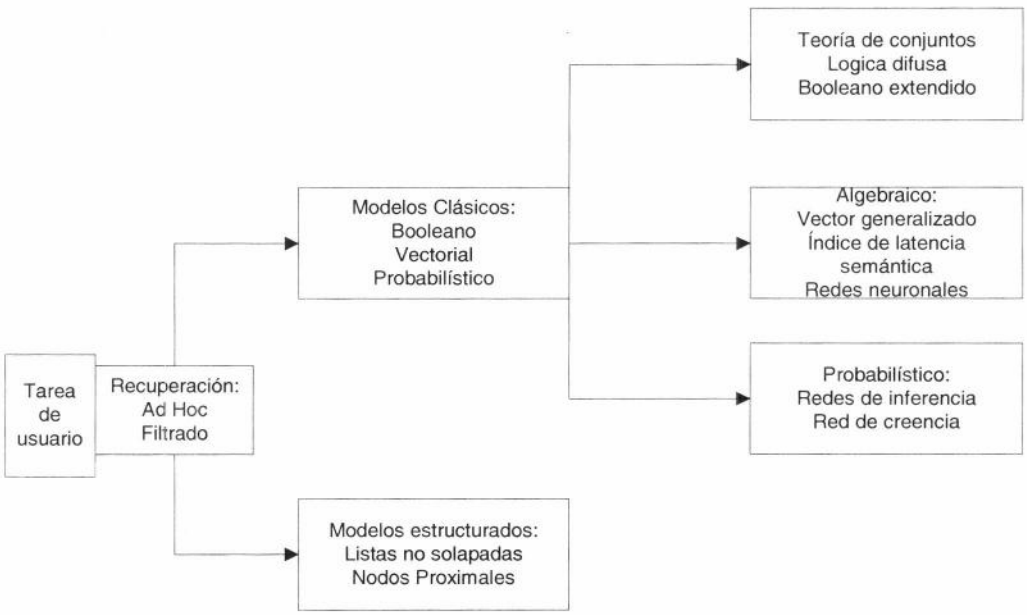


Figura IV.2 – Taxonomía de Modelos de Recuperación de la Información [BR99].

¿Qué es un Modelo de R.I.?

Un modelo de recuperación de la información es una 4-upla $[\mathbf{D},\mathbf{Q},\mathbf{F},\mathbf{R}(q_i,d_j)]$ donde

1) **D** es un conjunto que contiene las representaciones de los documentos de la colección.
2) **Q** es un conjunto que contiene la representación de la información que necesita el usuario (consultas).
3) **F** es un *framework* utilizado para modelar la representación de los documentos, las consultas y la relación entre ambos.
4) $\mathbf{R}(q_i,d_j)$ es la función que asocia un número con una consulta $q_i \in \mathbf{Q}$ y una representación de un documento $d_j \in \mathbf{D}$. Esta función de *ranking* define un orden entre los documentos con respecto a la consulta q_i.

Ejemplos de F

Para el modelo booleano, F está compuesto del conjunto de documentos y las operaciones sobre conjuntos.
Para el modelo vectorial, F está compuesto del espacio vectorial de t dimensiones y las operaciones algebraicas sobre vectores.

Conceptos Generales

Antes de continuar con el estudio específico de cada modelo, es necesario introducir algunos conceptos generales.

En recuperación de la información, los documentos son representados mediante un conjunto de palabras clave llamadas términos, cuya semántica es utilizada para inferir sobre qué trata el documento.

Se debe notar que no todos los términos utilizados en la representación del documento sirven de igual modo para definirlo, ya que algunos pueden ser más ambiguos que otros. Se asocia a cada término dentro del documento un peso que cuantifica la importancia del mismo a la hora de representar la semántica del documento.

A continuación, se definen formalmente algunos conceptos utilizados en el resto del capítulo.

Definición:
Sea t la cantidad de términos del índice y ki un término. Sea $K = \{k_1,...,k_t\}$ *el conjunto de todos los términos del índice. A cada término ki se le asocia un peso dentro del documento* d_j, *llamado* $w_{i,j} > 0$. *Si un término no aparece en el documento, su peso es 0* $(w_{i,j} = 0)$. *Además, a cada documento* d_j *se le asigna un vector de términos, representado por* $\overrightarrow{d_j} = (w_{1,j}, w_{2,j},..., w_{i,j})$. *Sea* g_i *la función que devuelve el peso asociado al término* k_i *en cualquier vector t-dimensional* $(g_i(\overrightarrow{d_j})) = w_{i,j})$

Modelo Booleano

El modelo booleano se basa en la teoría y el álgebra de Boole. Es simple y fácil de entender. Las consultas se especifican utilizando expresiones booleanas, es decir los términos están conectados mediante los operadores AND, OR, NOT.
El modelo booleano considera que el término está presente en el documento o no, con lo cual el peso de los términos es binario (0,1).

Definición:
Para el modelo booleano, los pesos son binarios: $w_{ij} \in \{0,1\}$. *Una consulta q es una expresión booleana. Sea* $\vec{q}_{dnf}$ *la forma normal disyuntiva para la consulta q. Sea* $\vec{q}_{cc}$ *cualquiera de los componentes conjuntivos de* $\vec{q}_{dnf}$. *La similitud de un documento* d_j *para una consulta q se define como:*

$$sim(d_j, q) = \begin{cases} 1 & Si\ \exists \vec{q}_{cc} \mid (\vec{q}_{cc} \in \vec{q}_{dnf}) \wedge (\forall_{ki}\, g_i(\overrightarrow{d_j}) = g_i(\vec{q}_{cc})) \\ 0 & En\ otro\ caso \end{cases}$$

Si bien este modelo es simple, un documento es relevante o no, lo que indica que no hay noción de orden entre los documentos recuperados. Por lo tanto, no se puede determinar si un documento es más relevante que otro.

Modelo Vectorial

El modelo vectorial se basa en el booleano pero trabaja con la noción de que un documento puede ser más relevante que otro para el usuario. Este cambio se logra asociando a cada término de la consulta y del documento un peso no booleano. Los pesos de los términos son utilizados para calcular el grado de similitud entre la consulta y los documentos. Por lo tanto, se puede obtener una lista de documentos ordenados por su grado de similitud con respecto a la consulta. Finalmente, se obtiene un conjunto de documentos que cumplen más adecuadamente con las necesidades del usuario que el conjunto devuelto por el modelo booleano [BR99]. Un punto en contra de este modelo, es que es menos expresivo que el modelo booleano.

Definición:

Para el modelo vectorial, el peso $w_{i,j}$ asociado al par (k_i,d_j) es mayor a cero y no es binario. Además, los términos de la consulta también son pesados. Sea $w_{i,q}$ el peso asociado al par (k_i,q) donde $w_{i,q} > 0$. Luego, se define el vector $\vec{q}$ como $\vec{q} = (w_{1,q}, w_{2,q}, \ldots, w_{t,q})$ donde t es la cantidad de términos en el índice. Como antes, el vector del documento $\vec{d}_j$ es representado como $\vec{d}_j = (w_{1,j}, w_{2,j}, \ldots, w_{t,j})$

En este modelo, un documento d_j y una consulta q son representados con vectores de t-dimensiones (ver figura IV.3)). Por lo tanto, para obtener el grado de similitud de d_j y q se puede utilizar la correlación entre los vectores d_j y q, para la cual se utiliza el coseno del ángulo entre los dos vectores, como sigue:

$$sim(dj,q) = \frac{\vec{dj} \bullet \vec{q}}{|\vec{dj}| \times |\vec{q}|}$$

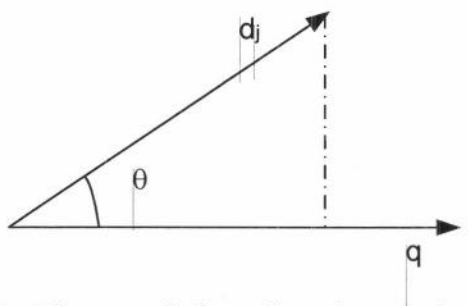


Figura IV.3 – El coseno de θ es adoptado como la sim(d_j,q).

El grado de similitud se encuentra entre 0 y 1, de manera que este modelo, en lugar de decirnos si un documento es relevante o no, ordena los documentos de acuerdo al grado de similitud con la consulta y retorna al usuario todos los documentos que poseen un grado de similitud mayor a un umbral prefijado. Este umbral define el grado aceptable de similitud.

Además, como mencionan Baeza Yates y Ribeiro Neto en [BR99], es posible analizar el problema de R.I. como un problema de clustering viendo a los documentos como un conjunto C de objetos y la consulta del usuario como la especificación del conjunto A.

Entonces el problema de R.I. se reduce a encontrar los documentos que están en el conjunto A, es decir, los que cumplen la especificación del usuario. En este caso es necesario resolver dos temas:

1) Se deben determinar las propiedades que describen mejor a los objetos que pertenecen al conjunto A. Estas propiedades permiten cuantificar la similitud intra-clustering, la cual es cuantificada como la frecuencia de un término k_i dentro del documento d_j. Esta medida se denomina tf y sirve para indicar cuán bien un término describe al documento.
2) Se deben especificar las propiedades que describen mejor a los objetos que pertenecen al conjunto C. Estas propiedades permiten cuantificar la disimilitud inter-clustering, la cual es cuantificada como la inversa de la frecuencia de un término k_i dentro del conjunto de documentos. Esta medida se denomina idf. El idf sirve para determinar si un término es realmente bueno como palabra clave, ya que si el término aparece demasiado en los documentos de la colección, éste no será útil para distinguir un documento relevante de otro no relevante.

Como conclusión, se observa que si un término aparece demasiado en un documento, se supone que es importante para ese documento y por lo tanto el tf crece, pero si éste aparece en muchos documentos, entonces resulta inútil para distinguir documentos entre si, con lo cual el idf decrece.

Definición:
Sea N la cantidad de documentos y n_i el número de documentos donde aparece el término k_i. Sea $freq_{i,j}$ la frecuencia del término k_i en el documento d_j. Entonces la frecuencia normalizada $f_{i,j}$ del término k_i en el documento d_j es

$$f_{i,j} = \frac{freq_{i,j}}{\max_l \times freq_{l,j}}$$

donde el máximo se calcula entre todos los términos del documento d_j. Si el término k_i no aparece en el documento d_j, entonces $f_{i,j}=0$. Además, se define la inversa de la frecuencia de k_i como

$$idf_i = \log \frac{N}{n_i}$$

Dadas las definiciones de $f_{i,j}$ e idf_i, se concluye que el mejor esquema para el peso de los términos es el siguiente

$$w_{i,j} = f_{i,j} \times idf_i$$

Este modelo permite mejorar la eficiencia en la recuperación al asociar un peso a los términos, permite recuperar documentos que concuerdan parcialmente con la consulta y además, los documentos son devueltos de manera ordenada según el grado de similitud con la consulta. Una desventaja del modelo vectorial, es que asume la independencia de los términos pero en la práctica, considerar la dependencia de los términos podría ser una desventaja, ya que su utilización desmedida podría disminuir mucho la eficiencia [BR99].

Modelo Probabilístico

Este modelo intenta resolver el problema de R.I. utilizando un *framework* probabilístico. Se basa en que dada una consulta, existe un conjunto que contiene exactamente los

documentos relevantes (conjunto ideal). Por lo tanto, y tal como fue explicado anteriormente con el problema de *clustering*, se debe especificar cuáles son las propiedades del conjunto. El problema es que no se sabe con exactitud cuáles son esas propiedades en el momento en que se realiza la consulta. De este modo, se debe especular en primera instancia sobre las mismas. Esto permite tener una primera aproximación a la especificación del conjunto de respuesta. Luego, el usuario interactúa con el sistema para ir mejorando los documentos recuperados, eligiendo los documentos relevantes del conjunto obtenido, de forma tal que el sistema pueda refinar la búsqueda. La idea es que al repetir varias veces el proceso, el sistema se va acercando al conjunto "ideal" de respuestas.

Definición:

Para el modelo probabilístico, los pesos de los términos son valores binarios $w_{i,j} \in \{0,1\}$, $w_{i,q} \in \{0,1\}$. *Una consulta q es un subconjunto de términos. Sea* R *el conjunto de los documentos conocidos (o inicialmente supuestos) como relevantes. Sea* $\overline{R}$ *el complemento de* R *(el conjunto de los documentos no relevantes). Sea* $P(R \mid d_j)$ *la probabilidad que el documento* d_j *sea relevante para la consulta q y* $P(\overline{R} \mid d_j)$ *la probabilidad que dj sea no relevante para q. La similitud* $sim(d_j,q)$ *del documento* d_j *y la consulta q se define como*

$$sim(d_{j,q}) = \frac{P(R \mid \vec{d_j})}{P(\overline{R} \mid \vec{d_j})}$$

La ventaja de este modelo es que ordena los documentos de forma decreciente, según su probabilidad de ser relevantes. Su desventaja principal es que se necesita suponer en principio cómo realizar la separación entre documentos relevantes y no relevantes. Además, este modelo no utiliza la frecuencia de los términos en un documento y supone la independencia entre los términos.

Los modelos listas no solapadas y nodos proximales se encuentran fuera del alcance de esta tesis.

¿Cómo Resolver Algunas Falencias de los Modelos Clásicos?

Para solucionar el problema de no capturar la correlación entre los términos se pueden utilizar ciertos tipos de análisis lingüístico. Por ejemplo:

- Utilizar distintas técnicas para descubrir la correlación entre los términos. Los modelos alternativos que se detallan a continuación y las técnicas de expansión de consultas (que se explicaran dentro del marco de operaciones sobre las consultas) apuntan a esto.

Modelos Alternativos

- Extensiones al modelo booleano
  - Booleano extendido
  - Conjuntos difusos

- Extensiones al modelo vectorial
  - Vectorial generalizado
  - LSI: *Latent Semantic Indexing*
  - Redes neuronales

- Extensiones al modelo probabilístico
  - Redes bayesianas
  - Redes de inferencia bayesiana

Estos modelos son costosos de implementar y no siempre dan buenos resultados. Por esta razón no son muy populares, exceptuando LSI, redes neuronales, redes bayesianas. [N00].

Características de los Documentos

Dentro de los documentos, las palabras poseen una distribución que se intenta aproximar mediante la ley de Zipf, la cual captura la distribución de las frecuencias de las palabras en un documento.

Ley de Zipf

La frecuencia de la iésima palabra de un documento es $1/i^{\theta}$ veces mayor que la palabra más frecuente. Esto implica que en un texto de n palabras con un vocabulario V, la iésima palabra más frecuente aparece $n/(i^{\theta} H_V(\theta))$ veces. En este caso, $H_V(\theta)$ es el número armónico de orden θ del conjunto V y se define de la siguiente manera:

$$H_V(\theta) = \sum_{j=1}^{V} \frac{1}{j^{\theta}}$$

Por lo tanto, la suma de todas las frecuencias es n.

Otra característica de los documentos es la cantidad de palabras distintas. Este conjunto se llama vocabulario. Para predecir el crecimiento del tamaño del vocabulario de un texto en lenguaje natural, se usa la ley de Heaps.

Ley de Heaps

El vocabulario de un texto de n palabras es de tamaño $V = K\eta^{\beta} = O(\eta^{\beta})$, donde K y β dependen del texto.

$$H_V(\theta) = \sum_{j=1}^{V} \frac{1}{j^{\theta}}$$

Operaciones sobre los Documentos

Como se dijo anteriormente, existen palabras que representan mejor la semántica del documento que otras. Por eso, es importante analizar el preprocesamiento del texto, cuyo principal objetivo es encontrar los términos que serán utilizados en el índice.

El preprocesamiento de documentos puede verse como el proceso de controlar el tamaño del vocabulario (índice), ya que se espera que un vocabulario controlado mejore la eficiencia de la recuperación. Sin embargo, a los usuarios les puede resultar más difícil interpretar el resultado de la consulta. Por ejemplo: Si un usuario sabe que existe un documento que contiene la frase "La casa del Rey", al realizar la consulta podría no encontrarlo en las primeras posiciones debido a que el vocabulario no contiene las

palabras "la" ni "del". Por este motivo, muchos motores de búsqueda no realizan ninguna transformación en el texto e indexan todo el documento.

A continuación, se describen los pasos principales que se realizan durante el preprocesamiento del texto [BR99]:

1) Análisis lexicográfico del texto
   Este proceso convierte una cadena de caracteres en una cadena de palabras, que serán las candidatas a pertenecer al índice. Es decir, el objetivo principal de esta fase es el de identificar palabras. En este punto se consideran los espacios, los números, los guiones, la puntuación y las mayúsculas y minúsculas.
2) Eliminación de *stopwords*
   Una palabra que aparece en el 80% de los documentos, no debería pertenecer al índice, ya que no permite diferenciar los documentos entre si. A este tipo de palabras se las conoce como *stopwords*. Como ejemplos se pueden citar los artículos, preposiciones y conjunciones.
3) Lematización
   El objetivo es obtener el tronco de la palabra (para no perderse variantes de la misma), de forma tal que también se puedan recuperar los textos que contienen variaciones de la palabra, e.g.: plurales, gerundios y tiempos verbales.
4) Selección de términos para el índice.
   Una estrategia conocida es utilizar los sustantivos del documento como términos del índice, ya que la mayor parte de la semántica está contenida en los mismos.
5) Construcción de tesauros
   Un simple tesauro consiste en una lista de palabras con sus respectivos sinónimos. El motivo fundamental para la construcción de tesauros se debe a que permite realizar una mejor indexación y tener un vocabulario controlado. Un vocabulario controlado tiene como ventaja la normalización del concepto de indexación, la reducción de ruido, identificar los términos con una semántica clara y realizar la recuperación basándose en el concepto en lugar de palabras.

Otro tipo de operaciones de texto es la compresión, la cual reduce el espacio requerido para el almacenamiento del texto. Sin embargo, su desventaja es el costo de codificar y decodificar el texto comprimido.

Operaciones sobre las Consultas

La recuperación basada en palabras clave puede ser de baja calidad por dos razones: la primera es que la consulta del usuario puede contener muy pocas palabras, lo cual implica que el contexto de búsqueda se encuentra pobremente caracterizado. La segunda es que el conjunto de palabras clave elegido para un documento puede no representarlo semánticamente. El primer problema se trata realizando transformaciones en la consulta, como expandirla o utilizar retroalimentación (*relevance feedback*). Estas técnicas se explican a continuación.

Relevance Feedback

La estrategia mas común para reformular las consultas es el ciclo denominado *relevance feedback*, en donde el usuario, luego de examinar los documentos marca los relevantes. La idea principal consiste en la selección de términos importantes de los documentos relevantes marcados por el usuario y en resaltar la importancia de estos términos en la

reformulación de la consulta. Se espera que la nueva consulta se acerque más a los documentos relevantes y se aleje de los no relevantes.

Es decir, puede preguntársele al usuario cuáles documentos de los recuperados son relevantes y cuáles no, y volver a pesar todo en función de la respuesta.
El objetivo es acercar el vector de consulta al conjunto de documentos relevantes y alejarlo del conjunto de documentos no relevantes.

Las ventajas que ofrece *relevance feedback* en la estrategia de reformulación de la consulta son las siguientes:

- Protege al usuario de los detalles del proceso de reformulación de la consulta, porque todos los usuarios tienen que seleccionar (según su opinión) los documentos relevantes.
- Divide la tarea de búsqueda en una secuencia de pasos menores que son más fáciles de comprender.
- Provee un proceso que permite enfatizar algunos términos (los relevantes) y des-enfatizar otros (los no relevantes) [BR99].

Expansión de la Consulta

Durante el ciclo de *relevance feedback*, el usuario examina los documentos mejor clasificados (i.e.: los que se encuentran en posiciones más altas del ranking) y los separa en dos clases: los relevantes y los no relevantes. Esta información es luego usada para seleccionar nuevos términos para la expansión de la consulta. Por lo tanto, existe una noción de *clustering* en la estrategia de *relevance feedback*. De acuerdo a esta noción, los documentos que el usuario marco como relevantes contienen términos que pueden ser útiles para describir un *cluster* grande de documentos relevantes. En este caso la descripción del *cluster* se construye de forma interactiva con el usuario. También existe un enfoque automático para obtener la descripción del *cluster* de documentos relevantes el cual generalmente se basa en la identificación de los términos relacionados con la consulta. Esos términos se obtienen utilizando sinónimos, lematización, o términos que están cercanos a la consulta en el documento.

Es posible expandir una consulta para mejorar la relevancia del resultado. Un modelo posible es considerar un espacio vectorial donde los términos son los puntos y los documentos las coordenadas. Si dos términos tienden a aparecer en los mismos documentos, sus vectores serán cercanos.

Dada una consulta formulada por términos, éstos se consideran vectores y la consulta se expande con todos los vectores (términos) cercanos a los de la consulta original.

Índices

Como se indicó anteriormente, los índices juegan un papel muy importante en la realización de búsquedas.

Una forma de realizar la búsqueda para una consulta simple es recorrer el texto secuencialmente (u *online*), lo cual significa que el texto no ha sido preprocesado. Esta

opción es apropiada en documentos pequeños y muy volátiles o cuando no se cuenta con demasiado espacio para el índice.

Una segunda opción es construir estructuras de datos que permitan acelerar las búsquedas (i.e.: índices). La construcción de índices es apropiada cuando la colección es grande y los documentos son semiestáticos

A continuación, se describen las principales técnicas de indexación utilizadas en R.I.

Índices Invertidos

Un índice invertido es un mecanismo para indexar texto y está orientado a palabras.
Se compone de un vocabulario (conjunto de términos distintos del texto) y una lista de posteo u ocurrencias, la cual contiene las posiciones en las que aparece cada término [N00].

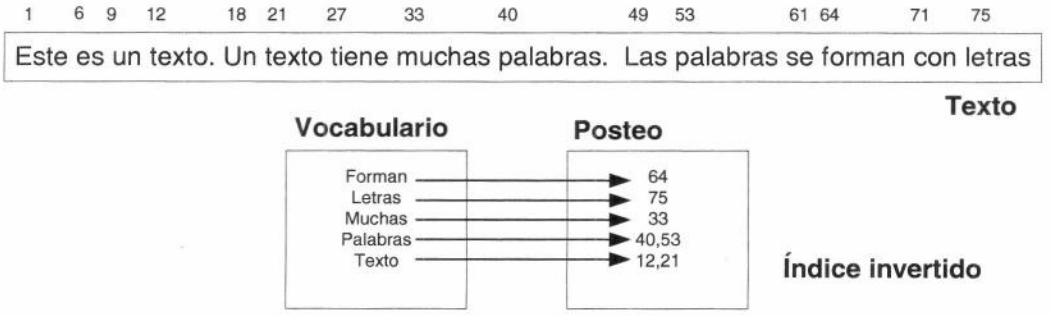


Figura IV.4 – Un texto y su índice invertido.

En los modelos booleano y probabilístico, la lista de posteo de cada término se encuentra preferentemente ordenada en forma creciente de documentos (cantidad de documentos).

En el modelo vectorial, se debe guardar el tf correspondiente en cada entrada de posteo para poder determinar la relevancia del término en el documento. Asimismo, se debe guardar el idf correspondiente en cada entrada del vocabulario, para poder determinar la importancia del término en la colección.

A fin de reducir espacio se puede utilizar un índice de bloques, mediante el cual se divide el texto en bloques y en la lista de posteo se apunta al bloque en donde se encuentra la palabra en lugar de apuntar a la posición exacta. Es importante notar que dentro de cada bloque que contiene el término hay que realizar una búsqueda secuencial para hallarlo y que además el texto debe estar siempre disponible, i.e.: esta estructura no se puede utilizar para la Web.

Arreglo de Sufijos

Un arreglo de sufijos es un índice que no necesita que el texto esté formado por palabras. Los arreglos de sufijos son una implementación eficiente de árboles de sufijos. Permiten realizar búsquedas complejas y obtener eficientemente una respuesta. Este tipo de índice ve al texto como una cadena grande y cada posición en el texto es considerada como un

sufijo. Por lo tanto puede recuperar rápidamente cualquier subcadena del texto, buscar expresiones regulares y realizar búsqueda aproximada [N00].

En R.I. tradicional, un arreglo de sufijos puede resultar útil para lo siguiente:
- Análisis lingüístico especializado
- Búsqueda de frases de varias palabras
- Búsqueda de patrones que trascienda palabras

Los arreglos de sufijos requieren un espacio cercano al 40% del texto, son más costosos de construir y mantener que un índice invertido y sólo son superiores a esté ultimo en la búsqueda de patrones.

Para finalizar este apéndice, se estudian las métricas mas conocidas de R.I., así como algunas métricas alternativas.

Métricas: ¿Cómo Determinar qué Modelo de R.I. es Mejor?

En los sistemas de R.I. es necesario determinar cuán preciso es el conjunto de respuestas. Para realizar esta evaluación, es necesario tener: un conjunto de documentos, un conjunto de consultas y un conjunto de respuestas relevantes (seleccionado por expertos) para cada una de las consultas.

Dada una estrategia de recuperación S, la evaluación cuantifica la similitud entre el conjunto de documentos recuperados por S y el conjunto de documentos relevantes provisto por expertos. Esto provee una estimación de bondad de S.

A continuación se explicarán las principales métricas y cómo se utilizan.

Precisión y Recuperación

En la mayoría de los casos, la efectividad de un sistema de R.I. es medida utilizando las métricas de precisión y recuperación (ver figura IV.5).

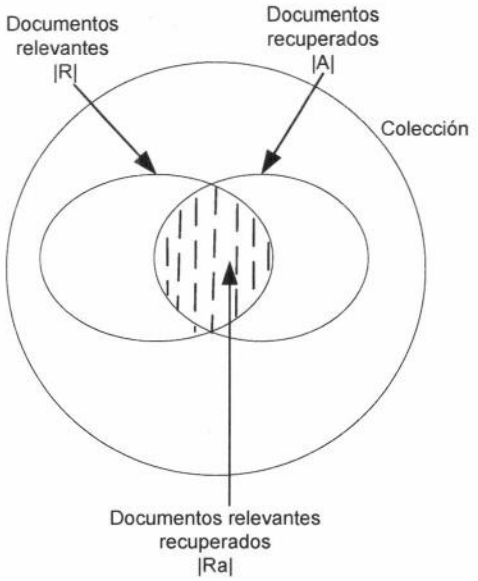


Figura IV.5 – Recuperación y precisión.

Sea

|C|	cantidad de elementos de la colección
|R|	cantidad de elementos del conjunto de documentos relevantes
|A|	cantidad de documentos recuperados para una consulta
|Ra|	cantidad de documentos relevantes recuperados

Recuperación indica cuántos documentos relevantes se recuperaron

$$\text{Recuperación} = \frac{|Ra|}{|C|}$$

Precisión indica cuantos documentos recuperados son relevantes

$$\text{Precisión} = \frac{|Ra|}{|A|}$$

Una baja precisión indica que el sistema devolvió al usuario muchos documentos irrelevantes. Mientras que una alta recuperación indica que el sistema no puede distinguir claramente los documentos relevantes.

Para lograr una alta precisión alcanza con devolver el documento más relevante, pero en ese caso la recuperación sería cercana a 0. Por otro lado, para obtener una alta recuperación alcanza con retornar todos los documentos del conjunto pero la precisión sería cercana a 0. Según la aplicación, puede ser preferible una u otra métrica.

Se debe notar que este tipo de mediciones no se pueden realizar adecuadamente en la Web, ya que el conjunto de documentos esta continuamente variando, no existe un vocabulario controlado, los documentos son muy diferentes entre sí, tanto en estructura

como en contenido y, además, actualmente no existe ningún motor de búsqueda que indexe completamente la Web. Por estos motivos, es imposible medir la recuperación en la Web. Esto se puede estimar mediante muestreo [N00].

Métricas Alternativas

A pesar de la popularidad de las métricas de recuperación y precisión, éstas no son siempre las más adecuadas. A continuación se describen métricas alternativas obtenidas de [BR99].

Harmonic Mean

Esta métrica combina precisión y recuperación.

$$F(j) = \frac{2}{\frac{1}{r(j)} + \frac{1}{P(j)}}$$

Donde r(j) es la recuperación del documento j y P(j) es la precisión del documento j.
La función F [0,1], es 0 cuando ningún documento relevante fue recuperado y 1 cuando todos los documentos relevantes fueron recuperados. Determinar el valor aceptable de F puede interpretarse como un compromiso entre recuperación y precisión.

E

Esta métrica también combina la precisión y la recuperación. Su objetivo es permitirle al usuario especificar que le interesa más, si la recuperación o la precisión.

$$E(j) = \frac{1+b^2}{\frac{b^2}{r(j)} + \frac{1}{P(j)}}$$

Donde r(j) es la recuperación del documento j y P(j) es la precisión del documento j y b es el parámetro utilizado por el usuario para indicar la importancia de la recuperación o la precisión.

Para

- b=1, E se comporta como F(j).
- b > 1, indica que el usuario esta más interesado en la precisión.
- b < 1, indica que el usuario esta más interesado en la recuperación.

Métricas Orientadas al Usuario

La recuperación y la precisión asumen que el conjunto de documentos relevantes es el mismo para todos los usuarios. Sin embargo, cada usuario podría tener una interpretación diferente con respecto a la relevancia de un documento. Para corregir este problema se propusieron varias métricas como *coverage*, *novelty*, *relative recall* y *recall effort* [BR99].

Sea R el conjunto de documentos relevantes para una consulta, A el conjunto de respuestas obtenido y U el conjunto de documentos relevantes conocido previamente por el usuario. La intersección de los conjuntos A y U conforma el conjunto de documentos relevantes conocidos por el usuario que fueron devueltos. Sea $|R_k|$ el número de documentos en el conjunto y sea $|R_u|$ el número de documentos relevantes desconocidos previamente por el usuario para la consulta. La figura IV.6 ilustra este caso.

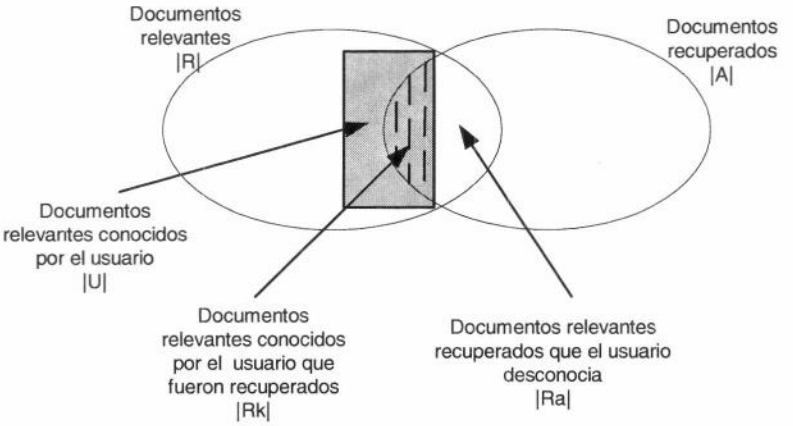


Figura IV.6 – *Coverage* y *Novelty.*

- ***Coverage*** muestra la relación entre documentos relevantes recuperados que el usuario conocía y los documentos relevantes realmente recuperados.

$$"\text{cov}\,erage" = \frac{|Rk|}{|U|}$$

- ***Novelty*** muestra la relación entre documentos relevantes recuperados que el usuario desconocía y los documentos relevantes recuperados.

$$novelty = \frac{|Ru|}{|Ru| + |Rk|}$$

Un valor alto de *coverage* indica que el sistema encuentra la mayoría de los documentos que el usuario esperaba ver. Un valor alto de *novelty* indica que el sistema revela al usuario una gran cantidad de documentos relevantes que desconocía.

- ***Relative recall*** indica la relación entre los documentos relevantes recuperados y los documentos relevantes que el usuario esperaba. Cuando el usuario encuentra los documentos que él esperaba, *relative recall* es igual a 1.

- ***Recall effort*** muestra la relación entre los documentos relevantes que el usuario esperaba y el número de documentos que tuvo que examinar para tratar de encontrarlos.

Otras Métricas

- ***Expected search length***: Es utilizada para conjuntos débilmente ordenados.
- ***Satisfaction***: Tiene en cuenta sólo los documentos relevantes recuperados.
- ***Frustration***: Tiene en cuenta sólo los documentos no relevantes.

Los conceptos expuestos en este apéndice son utilizados en los capítulos 3 – R.I. Web, 5 – Derivación de Métricas y 7 – Recolección y Análisis de datos.

APÉNDICE V. ANALISIS ESTADISTICO

V.1 Introducción

Este apéndice describe brevemente las medidas de estadística descriptiva utilizadas en el capítulo 7 - Recolección y análisis de datos.

V.2 Medidas de estadística descriptiva

A continuación, se definen las medidas de tendencia central y de dispersión.

Definición V.1: Medida de tendencia central: Medida que se refiere al punto medio de una distribución.

Definición V.2: Medida de dispersión: Medida que se refiere a la extensión de los datos en una distribución. Es decir, al grado en que las observaciones se distribuyen.

V.2.1 Medidas de tendencia central

En esta sección se explican las medidas de tendencia central.

Definición V.3: Media ($\overline{X}$): El promedio aritmético que se obtiene al sumar cada una de las observaciones y dividirlas por el número de observaciones realizadas.

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{k} x_i$$

Definición V.4: Mediana (Me): La observación central de una muestra. Para obtenerla, se ordenan las observaciones de menor a mayor. Si el tamaño de la muestra es:

- Impar: este valor es el que está en la posición

$$\frac{n+1}{2}$$

- Par: la mediana es la media de los dos valores centrales

$$\frac{x_i - x_{i+1}}{2}$$

Definición V.5: Moda (Mo): El valor que se presenta con mayor frecuencia en una serie de datos. Cuando ningún valor se repite, la serie no tiene moda; cuando varios valores se repiten

con igual frecuencia, la serie tiene más de una moda. Esto se conoce como distribución bimodal, trimodal, etc.

Definición V.6: Máximo (max): El valor máximo de las observaciones.

Definición V.7: Mínimo (min): El valor mínimo de las observaciones.

V.3 Medidas de dispersión

A continuación se describen las medidas de dispersión.

Definición V.8: Varianza (S^2): Diseminación de los valores en torno a la media, expresado en unidades cuadradas. Para calcular la varianza, se divide la suma de los cuadrados de las distancias entre la media y cada elemento de la muestra por el número total de observaciones.

$$S^2 = \frac{1}{n}\sum_{i=n}^{n}\left(xi - \bar{x}\right)^2$$

Definición V.9: Desviación estándar (S): Es la raíz cuadrada de la varianza y nos indica la dispersión de los valores alrededor de la media.

$$S = \sqrt{S^2}$$

Definición V.10: Rango: Diferencia entre el valor máximo y mínimo de las observaciones.

$$\text{Rango} = \max - \min$$

Definición V.11: MidRange: Suma del valor mínimo más el valor máximo divido dos.

$$\text{MidRange} = \frac{\min - \max}{2}$$

V.3.1 Comparación de las medidas de tendencia central

Las ventajas y limitaciones de la media, la moda y la mediana para describir un conjunto de datos dependen estrictamente de la distribución de la muestra. Siempre que sea posible, se prefiere la media para describir la tendencia central, aunque algunas distribuciones se describen mejor por medio de la moda y la mediana. A continuación, se determina qué tipo de tendencia conviene utilizar según la distribución de las observaciones.

En una **distribución normal**, la media, moda y la mediana tienen un valor idéntico. Esto se debe a que una distribución normal es perfectamente simétrica y la curva que la representa tiene un solo punto máximo (moda) que también se encuentra en el centro. Así, la media es la métrica preferida de tendencia central para distribuciones normales debido a su facilidad de cálculo.

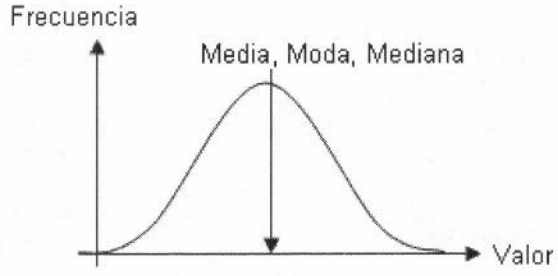


Figura V.1 – Distribución Normal

En las **distribuciones asimétricas** o sesgadas, la media no es la mejor medida de tendencia central. A mayor asimetría de datos, mayor utilidad tendrá la mediana, porque ésta estará más cerca del "valor promedio real" para las observaciones. La asimetría puede ser positiva o negativa tal como se observa en las siguientes figuras.

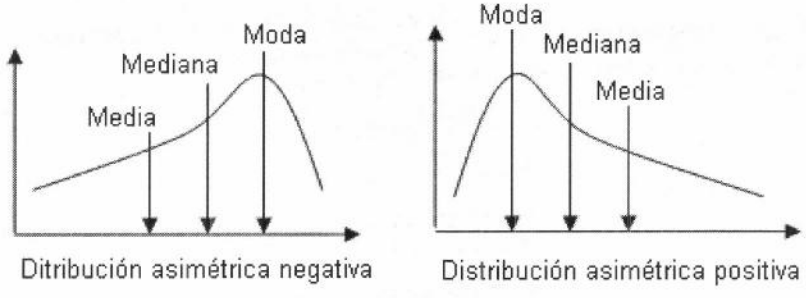


Figura V.2 - Distribución asimétrica

Existen varias medidas para la asimetría de una distribución. Una de ellas es el **coeficiente de asimetría de Pearson** para distribuciones campaniformes, unimodales y moderadamente asimétricas:

$$As = \frac{(\bar{x}) - Mo}{S}$$

Este coeficiente es igual a cero cuando la distribución es simétrica, positivo cuando la asimetría es a derecha y negativo cuando hay asimetría es a izquierda.

Otra medida de asimetría es el **coeficiente de asimetría de Fisher**, que se define de la siguiente manera:

$$A_f = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^3}{n * S^3}$$

La interpretación del coeficiente de Fisher es la misma que la del coeficiente de Pearson: si la distribución es simétrica vale cero y éste es positivo o negativo cuando exista asimetría a derecha o izquierda, respectivamente.

V.4 Gráficos

Para finalizar con este apéndice, se define el término histograma.

Definición V.12: Histograma: Gráfico de una distribución de frecuencias para una variable cuantitativa. En el eje horizontal se representan los intervalos de datos, marcándose de manera continua las fronteras entre cada uno de éstos. De esta manera, el histograma se compone de rectángulos cuya cantidad coincide con el número de intervalos considerados. El ancho de la base de cada uno de los rectángulos es siempre el mismo y coincide con las fronteras de los intervalos. Su altura corresponde a la frecuencia de cada intervalo.

ANÁLISIS COMPARATIVO DE MOTORES DE BÚSQUEDA: UN ENFOQUE BASADO EN TÉCNICAS DE CALIDAD DE DATOS

Tesis de Licenciatura en Ciencias de la Computación
Departamento de Computación
Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires

Agosto de 2005

Autoras:
Mariela Rodríguez L.U.: 69/93
Miriam Rosenberg L.U.:339/95
{mr9z, mrosenbe}@dc.uba.ar

Directora:
Lic. Mónica Bobrowski