

Capítulo 1: Recuperación de Información, Internet y sus servicios.

Introducción: La recuperación de información (en inglés, Information Retrieval, IR), visto del punto de vista de las Ciencias de la Computación es una disciplina que se ocupa de la representación, almacenamiento, organización y acceso a los elementos de información. El objetivo de la recuperación de la información es obtener información que pueda ser útil o relevante para el usuario. Se imagina la gran cantidad de datos que hoy en día se está transmitiendo o recibiendo alguno de los servidores de Google con el fin de darnos las “coordenadas” de la información asociada a esos datos.?. En este capítulo presentamos la Recuperación de la información como disciplina científica, cuya importancia radica en la relevancia. En particular, la IR trata de encontrar objetos (normalmente documentos, pudiendo ser imágenes, u otro objeto) cuyo carácter son de tipo no estructurado (generalmente de texto) y que satisfacen una necesidad de información (por parte de los usuarios) dentro de grandes colecciones (normalmente digitalizadas y/o almacenadas en computadoras), llamadas Corpus.

Sin embargo, la problemática es mucho más extensa, tal como lo define Gerard Salton, 1968:

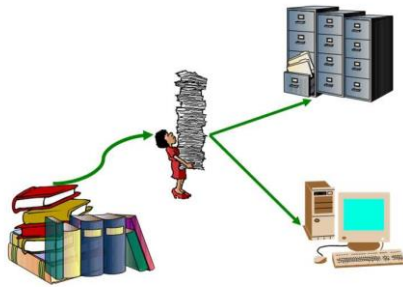
“Information retrieval is a field concerned with the structure, analysis, organization, storage, and retrieval of information.”.

Una vez realizada la recuperación de información, es necesario realizar una serie de procesos, entre ellos, almacenarla, gestionarla, administrarla, transformarla o filtrarla (según necesidad), priorizarla, para finalmente visualizar la información rescatada.

Todos estos procesos por sí sólo siguen siendo un problema constante a lo largo de la historia de la ciencia de la computación, y ahora con mayor razón dada la importancia y magnitud de la información que pudiese haberse recuperado. El cómo se recupera la información y las herramientas tecnológicas que lo apoyan es otro de los temas que se abordan en este capítulo, en donde claramente Internet y la Web juegan un rol superlativo en esta etapa. Ya que el diseño y el despliegue de motores de búsqueda como Google deben responder a consultas basadas en palabras clave mediante la extracción de resultados que incluyen punteros a páginas Web, pero con una capacidad y rapidez para escalar y administrar miles de millones de páginas indexadas dispersas en la Web.

El interés de mostrar el cómo se recupera la información y los procesos asociados a ello tiene que ver con evidenciar los esfuerzos realizados por las ciencias de la computación en términos de mostrar la existencia de interesantes algoritmos, estructuras de datos,

metodologías de recuperación, herramientas y aplicaciones computacionales que lo hacen un importante tema a tratar en el contexto de las TIC. En este segmento se estudiarán los fundamentos y técnicas para la IR como problema general en grandes bases documental, y dejar presentada la impronta en la web. Sin embargo, dejamos planteada la pregunta con la Fig. 1, que será lo más conveniente de hacer?



Según la Real Academia Española define

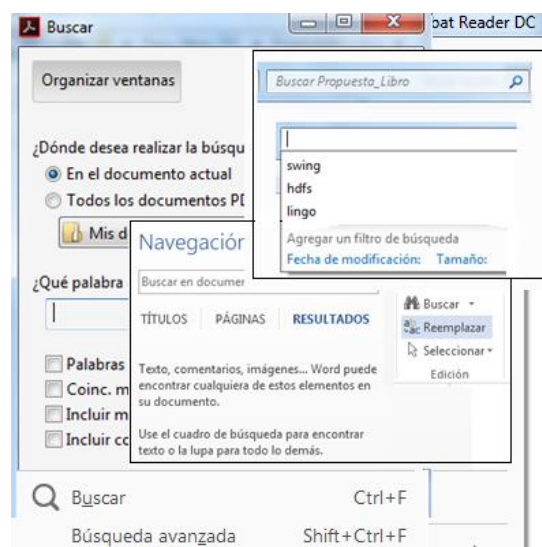
Buscar.

1. Hacer algo para hallar a alguien o algo. Por ejemplo, Estoy buscando un libro.
2. Hacer lo necesario para conseguir algo. Por ejemplo, Busca trabajo.

Recuperar

1. Volver a tomar o adquirir lo que antes se tenía.
2. Volver a poner en servicio lo que ya estaba inservible.

¿Buscar o Recuperar?. En realidad ambas definiciones no satisfacen los requerimientos de buscar o recuperar como la informática o la computación lo define. Y en este sentido lo que algunos sistemas ofrecen son servicios de búsqueda en Word, Adobe Acrobat y del propio sistema operativo Windows en la búsqueda de documentos, tal como se aprecia más abajo.



Y así se tienen diversos editores de texto o entornos de desarrollo capaces de proporcionar ayuda, como también los sistemas operativos, por ejemplo Windows.

Como ya vimos lo fundamental de la IR es encontrar objetos (normalmente documentos) de un carácter no estructurado (generalmente de texto) que satisfacen una necesidad de información (por parte del usuario) dentro de grandes colecciones (normalmente digitalizadas y/o almacenadas en computadoras, llamadas Corpus). Sin embargo, la búsqueda que nosotros realizamos en nuestro trabajo tiene que ver con los medios y recursos tecnológicos que poseemos, es así como Internet y los motores de búsqueda nos ofrecen una excelente oportunidad como usuario. En este segmento por lo general usaremos el motor de búsqueda de Google para ejemplificar los procesos luego de haber planteado una consulta según se vio en Capítulo de FORMAS DE BUSCAR.

Antes de plantearnos la consulta asociada a la búsqueda de información requerida hay que considerar:

El viaje de una consulta de búsqueda comienza mucho antes de que escribas tu búsqueda en Google, ya que usa robots de Software, conocidos como arañas web, que buscan páginas web para luego incluirlas en resultados de búsqueda de Google. Google almacena datos acerca de estas páginas en centros de datos. <https://www.google.com/intl/es-419/about/datacenters/gallery/#/>, Chile ya cuenta con un centro de datos de Google, extendiendo así a 13 centros de datos que tiene alrededor del mundo. Este está ubicado en la comuna de Quilicura y es el primero en instalarse en el hemisferio sur. Funciona a través de la instalación de planta fotovoltaica, una de las más importantes de Latinoamérica, siendo su tarea principal el resguardar toda la información del mundo.

Los Google Data Center son instalaciones especialmente creadas por la empresa para almacenar y gestionar sus servidores, donde guarda la información de sus usuarios, en plataformas como Gmail, YouTube o Android, tras ser dividido y encriptado para que se resguarde al máximo de la confidencialidad.

En este contexto, la Web bien se puede considerar como un gran libro con trillones de páginas, cuyo desafío principal es construir su índice para facilitar la búsqueda ante la solicitud de los usuarios o lectores. Los índices superan los 100 millones de Gigabytes, dedicando más de 1 millón de horas de procesamiento para elaborar este índice.

Cabe señalar que a medida que estás escribiendo la consulta, el proceso de búsqueda ya está activando los algoritmos de Google con el propósito de buscar la información que deseas para cuando se inicie tu búsqueda. Es así como antes de ofrecer resultados, la consulta de búsqueda pudo haber pasado por diferentes centros de datos o Data Center.

Ya cuando empiezas a escribir la búsqueda, Google Instant predice qué estás buscando y comienza a mostrar resultados, sin necesidad de presionar Enter. De esta manera, optimiza el tiempo de la respuesta. Los resultados provienen o están basadas en lo que buscan otras personas y en el contenido de páginas web que Google indexa.

<https://support.google.com/websearch/answer/186645?hl=es-419>

Luego de aquello, el algoritmo analiza tu consulta y usa filtros para decidir qué páginas y que contenido constituyen la respuesta más relevante. Algunos de ellos son:

- i. La cantidad de sitios web que establecen vínculos con un sitio en particular y la autoridad de estos vínculos.
- ii. Sinónimos de tus palabras clave de búsqueda.
- iii. El corrector ortográfico, etc.

Una vez realizada la consulta y el proceso interno de indexación, filtrado y más es la hora de mostrar los resultados al instante. Esta imagen describe lo que antes se mencionaba

Recuperación de la Información

Con la evolución exponencial de la web en la década de los 90, el campo de la recuperación de la información ha pasado de ser un área de interés minoritario a convertirse en foco de atención para empresas, instituciones, y cientos de millones de usuarios ávidos por información. <http://www.evolutionoftheweb.com/?hl=es>

No cabe duda que el volumen de datos que maneja la web y la falta de un modelo de datos subyacente representan un obstáculo importante para resolver las búsquedas en forma más acertadas a las necesidades del usuario, de allí que muchas técnicas que se habían desarrollado durante décadas anteriores a la aparición de la web, junto con otras nuevas y más específicas, subyacen a la efectividad y popularidad de los buscadores actuales, pero aún con limitaciones importantes.

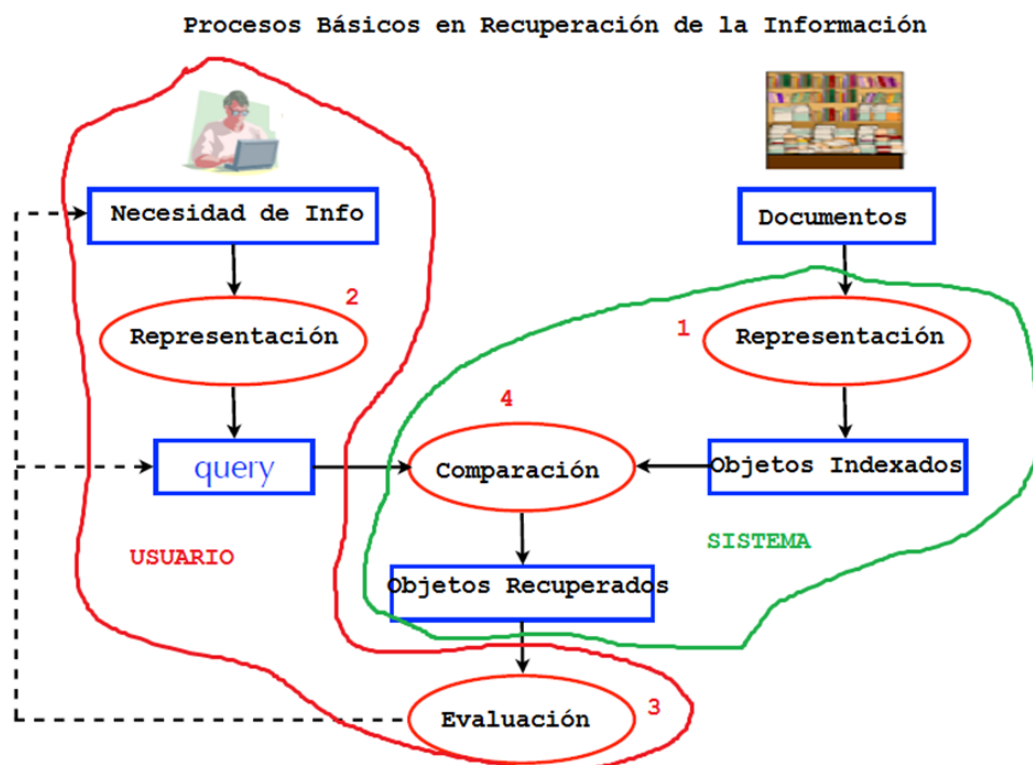
Un sistema de recuperación de información en específico textual lleva a cabo las siguientes tareas para responder a las consultas de un usuario:

1. *Indexación de la colección de documentos*: en esta fase se genera un índice que contiene las descripciones de los documentos. Normalmente, cada documento es descrito mediante el conjunto de términos que, hipotéticamente, mejor representa su contenido.
2. *Formular una consulta* (usando algún lenguaje en particular, por parte del usuario) el sistema la analiza, y si es necesario la transforma, con el fin de representar la necesidad de

información requerida por el usuario del mismo modo que el contenido de los documentos recuperados.

3. *El Sistema de recuperación compara la descripción de cada documento con la descripción de la consulta, y presenta al usuario aquellos documentos cuyas descripciones más se “asemejan” a la descripción de su consulta.*
4. *Los resultados se muestran en función de su relevancia, es decir, ordenados en función del grado de similitud entre las descripciones de los documentos y de la consulta.*

Ha continuación en **Fig. 1** se muestra un modelo genérico de los procesos que intervienen en la IR.



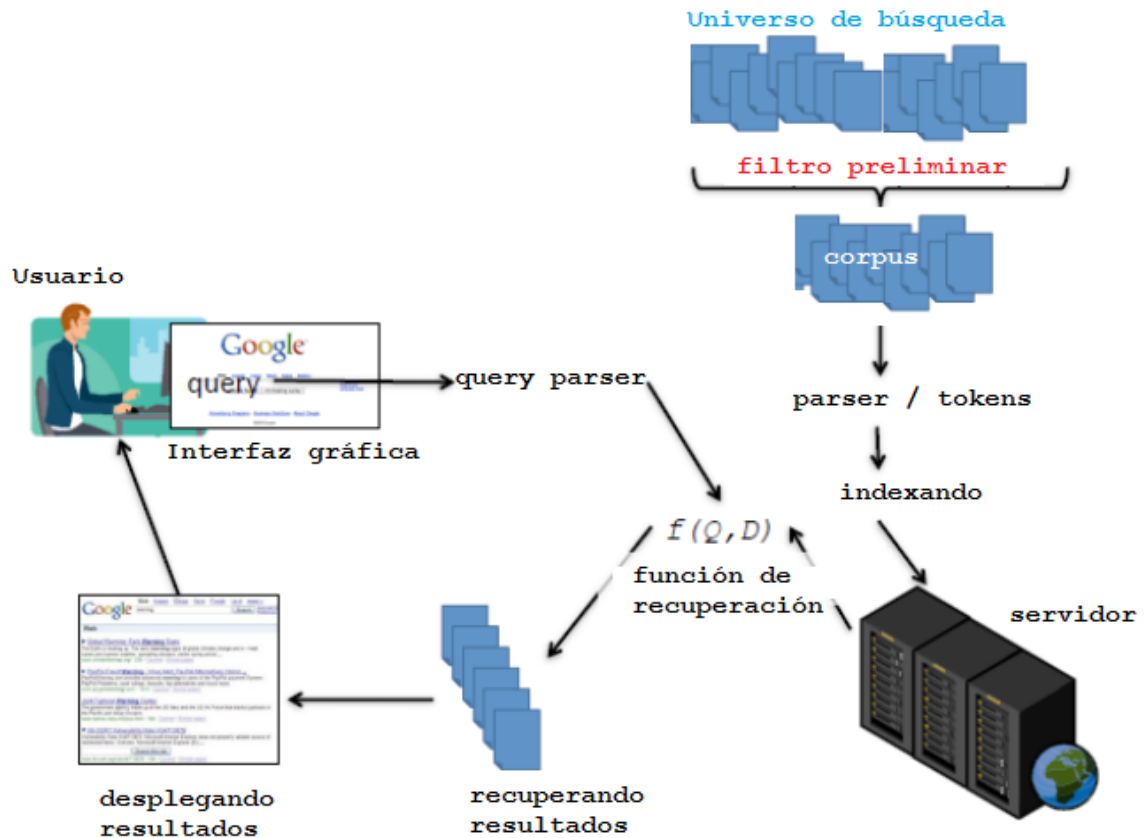
Documentos son las unidades hacia las que se construye el sistema de IR. Estructuralmente, los documentos pueden ser de naturaleza homogénea o heterogénea. Por ejemplo, un sistema orientado a documentos homogéneos podría ser un sistema de IR sobre archivos de biblioteca (en este caso los documentos son archivos estructuralmente idénticos u *homogéneos*).

Para el caso heterogéneo podemos tomar como referencia al buscador web Google, el cual indexa páginas web, documentos PDF, presentaciones PowerPoint y muchos otros formatos *heterogéneos*.

Corpus es el conjunto de documentos recuperables en el sistema. El corpus se puede caracterizar según varios criterios, los cuales dependen de la naturaleza del sistema. La constitución del corpus se refiere al proceso mediante el cual se fijan los criterios que han de guiar el diseño del corpus y, de acuerdo con ellos, se recopilan los textos. La selección de los textos que formarán parte de un corpus se puede efectuar según criterios internos o criterios externos, o bien estáticos o dinámicos. En general, un corpus puede ser:

a) Interno vs. Externo: dependiendo si el almacenamiento de los documentos está controlado por el sistema o fuera de su control. Esto es que se centran en la aparición de patrones o elementos diferenciadores de la variedad lingüística de un texto, como p. ej. la longitud de las oraciones. Destacan entre estos criterios, el tema: dominio o ámbito (tópico) al que pertenece el texto, que corresponde a un parámetro que clasifica un texto a partir de su contenido. Véase también la lista de temas que utiliza el BNC (British National Corpus) <http://ota.ox.ac.uk/desc/2554>. Por ejemplo, el BNC se guio por el tema (*criterio interno*), el medio de publicación y la fecha (*criterios externos*) para elegir los textos escritos que iban a conformar el corpus. <http://www.natcorp.ox.ac.uk/corpus/>. CREA (Corpus de Referencia Español Actual) <http://corpus.rae.es/creanet.html>, utiliza criterios muy similares: tema como criterio interno y medio de publicación, fecha y, además, procedencia geográfica como criterios externos.

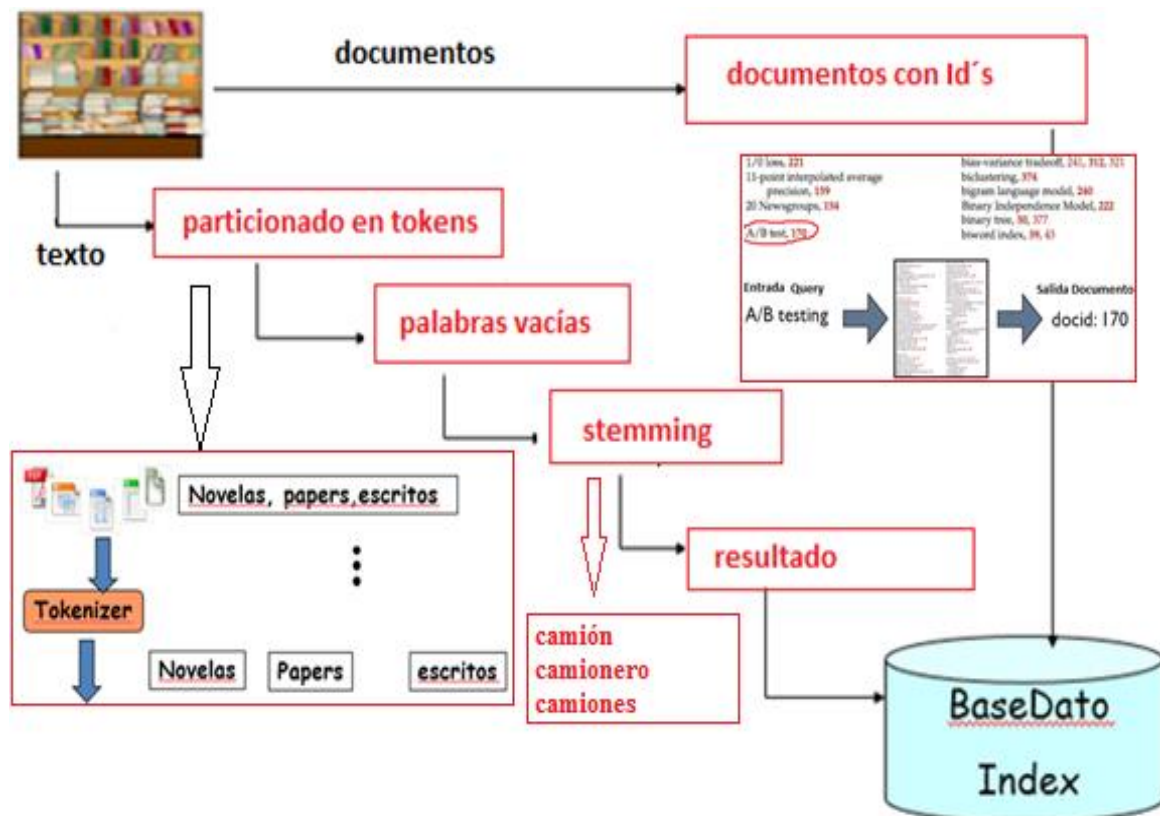
b) Estático vs. Dinámico: en base a si el contenido de sus documentos cambia o permanece inmutable en el tiempo. Para asegurar la representatividad de un corpus también hay que tener en cuenta los cambios debidos al tiempo, de ahí la necesidad de efectuar actualizaciones. Se distingue en este sentido entre corpus estáticos y corpus dinámicos. Por otra parte, atendiendo a las categorías textuales, no es lo mismo un corpus general que uno especializado. El primero, al tener como objetivo proporcionar una imagen lo más completa de la lengua, suele contener una amplia gama de géneros, mientras que el segundo se limita a un dominio (p. ej. el Derecho) o a un género (p. ej. el periodístico). Sin embargo, ambos tipos de corpus deberán mantener cierta proporcionalidad entre los tipos textuales que pretenden representar (p. ej. un corpus especializado en Derecho deberá contener en la debida proporción textos legales, resoluciones judiciales, etc.). En la práctica se ve como sigue Fig. 2, describir la figura



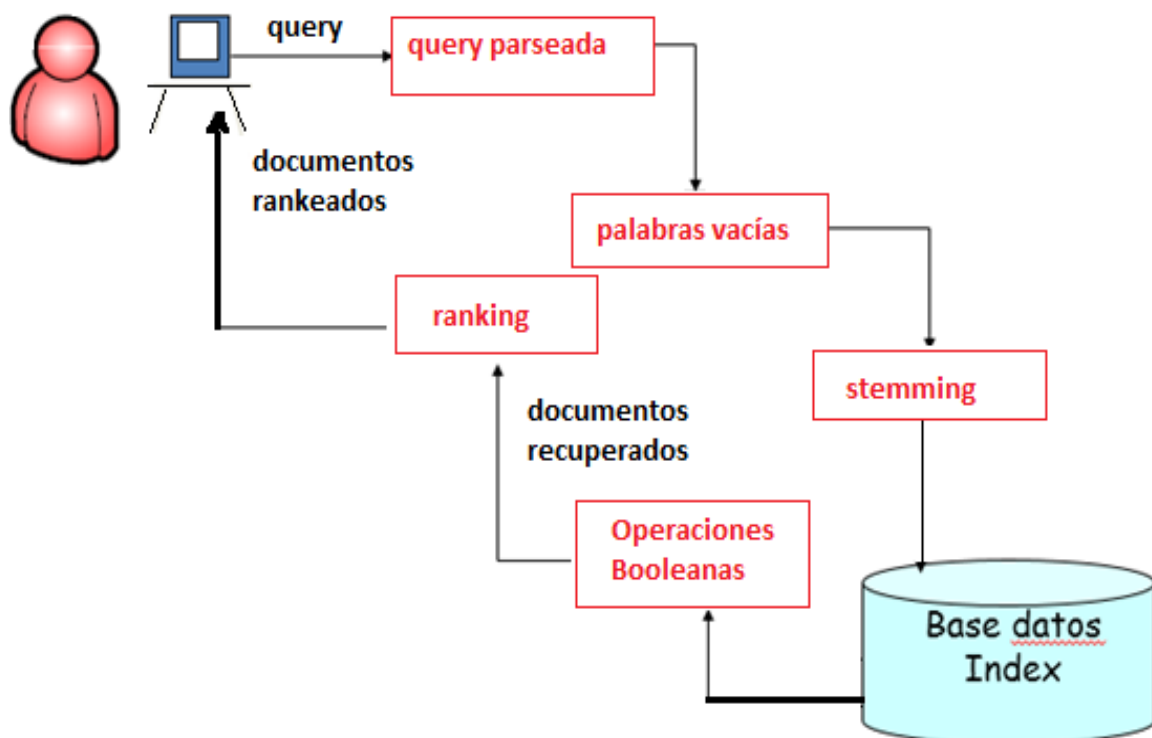
Modelo genérico en la Recuperación de la Información en datos textuales (NoSQL)

Un modelo de recuperación de la información especifica el detalle de:

- Representación de la Documentación
- Representación de las Query o consultas
- Relevancia de los documentos recuperados.



Una vista desde la colección de datos (corpus)



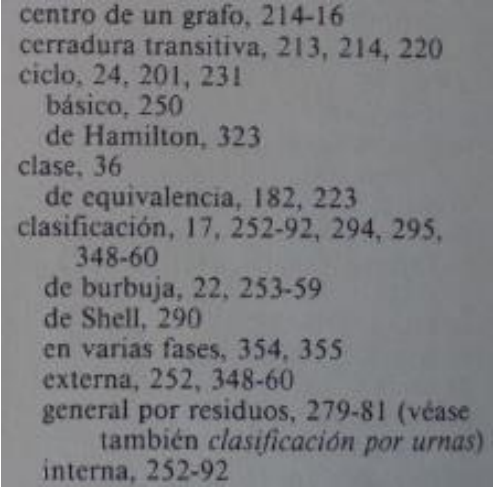
Una vista desde el usuario

Los sistemas de recuperación basados en términos de índice se apoyan en la idea fundamental de que tanto el contenido de los documentos como la necesidad informativa del usuario pueden representarse con términos de índice.

Ejemplo: Supongamos que tenemos algún libro de texto tomado al azar. Si usted necesita alguna información acerca de algún tema, por ejemplo activación de energías, lo natural es que abrirá el índice y averiguará esa palabra. El índice invertido le indicará los números de página donde esa palabra se explica, así como otras páginas en donde aparece.

Ahora, si va a realizar una búsqueda lineal ordinaria, podrá tardar horas para llegar a esa página, pero según sea la estructura de datos a usar puede ser muy rápido, apenas una cuestión de segundos. Entonces, ¿qué hace un índice normal?. Por supuesto, justo enfrente de ella, asigna el número de página relacionada con el tema, notar que la astucia no va con la utilidad en el ámbito de la búsqueda y extracción de información, pero buenos en este otro ámbito.

En la práctica, la consulta obedece a saber de “*clase de equivalencia*”, que es cercano a “*clase*” entregando el valor 182, 223 y que corresponde al n° de la página en donde el tema se encuentra. La ventaja de este método es bastante determinístico, en el sentido que no deja espacio para plantearse otro tipo de temas que no sean los que están o son similares en el index.



centro de un grafo, 214-16
cerradura transitiva, 213, 214, 220
ciclo, 24, 201, 231
 básico, 250
 de Hamilton, 323
clase, 36
 de equivalencia, 182, 223
clasificación, 17, 252-92, 294, 295,
 348-60
 de burbuja, 22, 253-59
 de Shell, 290
 en varias fases, 354, 355
 externa, 252, 348-60
 general por residuos, 279-81 (véase
 también *clasificación por urnas*)
 interna, 252-92

En este sentido, surgen una serie de interrogantes, por ejemplo, problemas de inconsistencia entre indizadores (por ejemplo, sinonimia, polisemia, etc.), en donde se requieren índices de concordancia, ya que dos personas ante una búsqueda pueden asignar diferentes palabras al

mismo concepto, y la misma palabra puede aparecer en documentos que traten temas diferentes, por ejemplo, “*vendo coche usado*” vs. “*automóvil de segunda mano*”.

Este es otro ejemplo más de las dificultades que se tienen hasta hoy con la IR. Pero quedémonos con saber los procesos internos que intervienen. Por ejemplo,

1) *Análisis del texto* para determinar el tratamiento que se realizará sobre números, guiones, signos de puntuación, tratamiento de mayúsculas y/o minúsculas, etc. Por ejemplo, Separación de palabras y 'localización', Carácter espacio, punto, comas,..., Caracteres de puntuación. Tratamiento de acentos. (Importante en otras fases del proceso léxico), Tratamiento de números, Almacenamiento en mayúsculas/minúsculas, Eliminación de palabras vacías, muy frecuentes y muy poco frecuentes.

Con todo esto se pretende reducir el número de términos con valores muy pocos significativos para la recuperación, como podría ser para cuando usamos palabras vacías, ya que poseen poca capacidad semántica, o bien disminuir el tamaño de la base de datos.

2) *Aplicación de lematización* (stemming) sobre los términos resultantes para eliminar variaciones morfo-sintácticas y obtener lemas, esto es, elegir convencionalmente una forma para remitir a ellas todas las que derivan de su misma familia por razones de economía. Así, camión- sería la raíz de camionero, camiones, camiós, camiois, etc., y garraf- la de garrafón, garrafa, garrafiña, etc. Palabras que son variaciones morfológicas con un significado prácticamente idéntico. Desde luego existen otras formas de lematización.

3) *Selección de términos* que serán considerados términos índice (sustantivos, nombres propios).

4) *Utilización de Tesauros*. Tesauro es la lista de palabras o términos controlados empleados para representar conceptos. Puede ayudar tanto en el proceso de indización como en el de búsqueda de información (expansión de consultas).

No se pretende ser exhaustivo respecto al tema de IR, sino dejar claramente establecido la diversidad de procesos que intervienen, entre ellos, lenguajes formales, estructuras de datos, bases de datos, algoritmos, neurociencia en la recuperación, etc. Y con ello dar una explicación a los usuarios de cómo logra Google, entre otros motores de búsqueda realizar la búsqueda de palabras o conceptos.

<http://infolab.stanford.edu/~backrub/google.html>

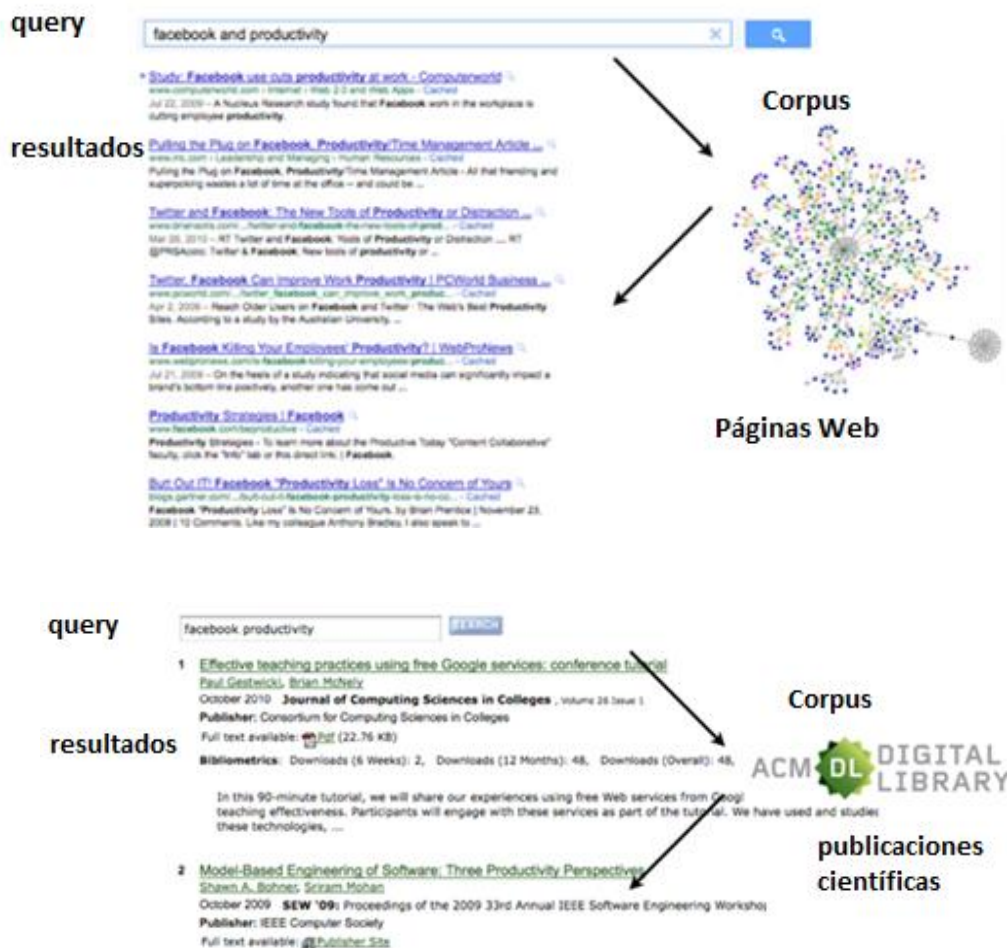
Para mayor información de cómo trabaja Google en particular, vea el enlace

<https://www.google.com/intl/es-419/insidesearch/howsearchworks/algorithms.html>

<https://www.google.com/intl/es-419/insidesearch/features/>

En ellos podrá configurar la (o las) búsquedas de la forma que Ud. lo desee, pero luego volveremos sobre este tema.

Ejemplos: Dada una descripción textual de necesidad de información (usualmente llamada **Query**), una colección de documentos textuales (usualmente llamada **Corpus**), y la satisfacción del usuario frente a las necesidades de información (usualmente llamada relevancia, traducida en **resultados**).



query

mexican food

resultados

Places for mexican food near Chapel Hill, NC

- Bandito's Mexican Cafe & Cantina** • 14 reviews • Place page
www.banditoscafe.com • 159 1/2 East Franklin Street, Chapel Hill • (919) 967-5048
- Las Poltronas Mexican Restaurant** • 9 reviews • Place page
www.lospoltronas.net • 220 West Rosemary Street, Chapel Hill • (919) 932-4301
- Montecito Mexican Restaurant** • 17 reviews • Place page
montecitomexicanrestaurant.com • 237 South Elliot Road, Chapel Hill • (919) 969-8750
- Margarita Cantina** • 19 reviews • Place page
www.margaritacantina.com • 1129 Weaver Dairy Road, Chapel Hill • (919) 942-4745
- Qdoba Mexican Grill** • 19 reviews • Place page
www.qdoba.com • 100 West Franklin Street, Chapel Hill • (919) 929-8996
- Cinco de Mayo** • 11 reviews • Place page
www.cincomayorrestaurant.net • 1502 East Franklin Street, Chapel Hill • (919) 929-8566
- Chipotle Mexican Grill** • 15 reviews • Place page
www.chipotle.com • 301 W. Franklin St., Chapel Hill • (919) 942-2091

corpus



query

Mi_Solr_2016

resultados

- mi_solr
- Mi_Solr_cloud
- Mi_Solr_erico
- Mi_Solr_general
- solr_home_Einfuehrung
- twitter
- MI-Sem_SS2012_Suchmaschinen
- Praxis_solr
- Solr in Action
- Solr_Final_Report

corpus



query

twitter and productivity

resultados

neenjames Neen James
Productivity tip: Follow ppl on Twitter that inspire, challenge and inform you - delete the clutter!
4 minutes ago

mr_Ostentatious Jason Pitts
Took a day off from twitter to increase my productivity and ended up having a productive day!
1 hour ago

adamwiebe Adam Wiebe
Social media at work is here. Be wary of what is and is not productive. <http://lnkd.in/DW3z8J>
3 hours ago

ViggosDaddy Gert van der Linde
A brief look: To tweet, or not to tweet? - How does Twitter affect our productivity, influence and how informe... <http://tinyurl.com/l3wbz3m>
6 hours ago

corpus



tweets

Por lo tanto, la recuperación básica de la información corresponde a recuperar documentos desde una colección, los que están dispuestos de una forma bien particular en respuesta a una consulta del usuario.

Dentro de todas las aplicaciones que tiene la IR, quisiera mencionar algunos ámbitos en donde la recuperación de la información se encuentra inserta. Uno de ellos es la **Auditoría Informática** que es un proceso llevado a cabo por profesionales especialmente capacitados para el efecto, y que consiste en recoger, agrupar y evaluar evidencias para determinar si un sistema de información salvaguarda el activo empresarial, manteniendo la integridad de los datos, y que cumple eficazmente con los fines de la organización, utilizando eficientemente los recursos, cumpliendo con las leyes y regulaciones establecidas.

Finalmente, el auditor tras su labor, emite un informe, determinando que los procesos son los adecuados y si cumplen o no con determinados objetivos o estrategias, estableciendo así los cambios o recomendaciones que se deberían realizar para la consecución de los mismos. Generalmente son profesionales de las ciencias de la ingeniería por ejemplo, Ingeniero Civil, Ejecución informático o Ingeniero en Computación, con sólidos conocimientos en Seguridad de la Información, Redes TCP/IP, Fraude Financiero y Delitos Informáticos.

En un contexto judicial, puede emitir un informe pericial informático, y con ello, aportar pruebas que permitan a los actores judiciales (fiscales, jueces o abogados defensores), llegar a una certeza jurídica respecto de lo allí encontrado, y así poder determinar su liberación por falta de pruebas o bien avalar las responsabilidades individuales del encausado. Como se puede apreciar la importancia de este profesional en lo que respecta a la IR en el ámbito judicial hoy en día es de vital importancia, ya que permite esclarecer la autoría o no de algunos delitos informáticos, entre otros.

Dado que la información es un bien sensible, transable y estratégico de las personas, empresas u organizaciones, entonces bien pueden verse afectada en diversos ámbitos legales, entre otros, la intervención de la misma puede vulnerar principios que están en la Ley de propiedad intelectual (Ley no. 17.336 ; Ley no. 19.166), Ley de Habeas Data (Ley 19.628 sobre Protección a la Vida Privada), Firma Digital (Ley no. 19.799, Ley sobre Documento Electrónico; Servicios de Certificación de Firma Electrónica), delitos informáticos (Ley no. 19.223, Delitos Informáticos; Sistemas de Información) y otras, que están vinculadas con el quehacer de este profesional.

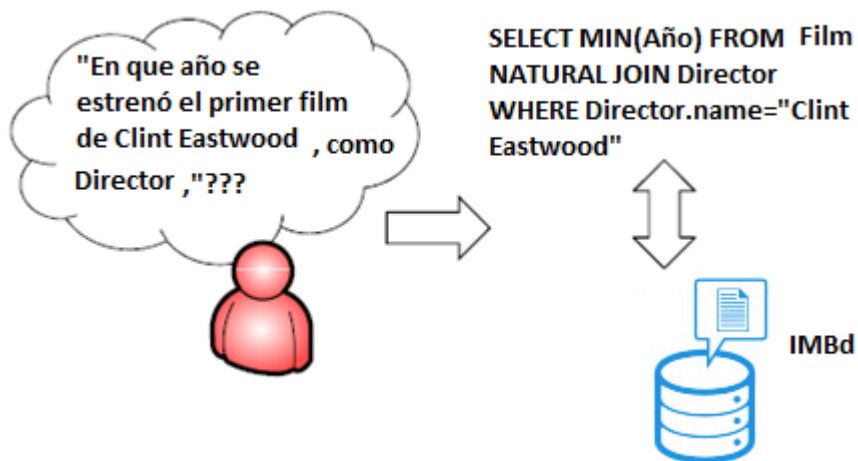
Es por ello que con el crecimiento espectacular de la web, los buscadores con sus motores de búsqueda (o meta-buscadores) se han convertido en puertas de entrada indispensables para cualquier usuario, y por ende este usuario debe estar preparado para tener una conducta que lo haga ser responsable, íntegro y eficiente dentro de su quehacer actual como estudiante y a futuro como profesional.

En lo particular, hoy en día almacenar información no es un problema, más aún los dispositivos en los últimos 10 años han bajado de costo en forma extraordinaria, y los servicios en las nubes (Cloud Computing) permiten en forma relativamente confiable almacenar información confidencial, aunque la normativa legal chilena dista mucho de regular esta situación, no así la IR que sigue siendo un problema en vías de solución, con luces y sombras.

La computación en la nube (en inglés, Cloud Computing), conocida también como servicios en la nube, informática en la nube es un paradigma que permite ofrecer servicios de computación a través de una red, que usualmente es Internet. Para ello, se cuenta con empresas que ofrecen este servicio a través de servidores que atienden las peticiones en cualquier momento, a las cuales se puede acceder mediante una conexión a internet desde cualquier dispositivo móvil o fijo ubicado en cualquier lugar. Algunos ejemplos, Dropbox desarrollado por Dropbox, Google Drive desarrollado por Google, iCloud desarrollado por Apple, entre otros.

Porque no buscar en una base de datos?

En principio se habló de un modelo genérico en la IR en datos textuales (NoSQL), se hace este énfasis porque bien podría consultarse con un lenguaje SQL a una base de datos que contenga tablas relacionales con Filmes, Directores, Actores, etc, y donde una consulta no sería tan sofisticada a como ya se ha venido planteando, sino que obedecería a plantearse una consulta de la siguiente manera:



Pero si la información viene así.....oh!? .. y ahora... QUE?

| Año | Título original | Títulos en español | Nota |
|-------------------------|---|--|---|
| 1966 | what's up, Tiger Lily? | Lily, la Tigresa | versión reeditada y doblada al inglés del filme japonés |
| 1969 | Take the money and run | Toma el dinero y corre / Robó, huyó y lo pescaron | |
| 1971 | Men of crisis: The Harvey wellinger story | | Cortometraje, telefilme |
| Bananas | Bananas | La locura está de moda | |
| 1972 | Everything you always wanted to know about sex (But were afraid to ask) | Todo lo que siempre quiso saber | |
| 1973 | Sleeper | El dormilón | |
| 1975 | Love and death | La última noche de Boris Grushenko / Amor y muerte | |
| 1976 | The Front | La Tapadera | |
| 1977 | Annie Hall | Annie Hall / Dos extraños amantes | |
| 1978 | Interiors | Interiores | |
| 1979 | Manhattan | Manhattan | |
| 1980 | Stardust memories | Recuerdos / Recuerdos de una estrella / Polvo de estrellas | |
| 1982 | A midsummer night's sex comedy | La comedia sexual de una noche de verano / Comedia sexual de una noche | |
| 1983 | Zelig | Zelig | |
| 1984 | Broadway Danny Rose | Broadway Danny Rose | |
| 1985 | The purple rose of Cairo | La rosa púrpura de El Cairo | |
| 1986 | Hannah and her sisters | Hannah y sus hermanas | |
| 1987 | Radio days | Días de radio | |
| September | September | Septiembre | |
| 1988 | Another woman | Otra mujer / La otra mujer | |
| 1989 | Oedipus wrecks, | episodio de New York stories | Edipo reprimido, episodio de Historias de Nueva York |
| Crimes and misdemeanors | Delitos y faltas / Crímenes y pecados | | |
| 1990 | Alice | Alice | |
| 1992 | Shadows and fog | Sombras y niebla | |
| Husbands and wives | Maridos y mujeres / Maridos y esposas | | |
| 1993 | Manhattan murder mystery | Misterioso asesinato en Manhattan / Un misterioso asesinato en Manhattar | |
| 1994 | Bullets over Broadway | Balas sobre Broadway / Disparos sobre Broadway / Balas sobre Nueva York | |
| Don't drink the water | Los USA en zona rusa | Telefilme | |
| 1995 | Mighty Aphrodite | Poderosa Afrodita | |
| 1996 | Everyone says I love you | Todos dicen I Love You / Todos dicen te quiero / Todos dicen que te amo | |
| 1997 | Deconstructing Harry | Desmontando a Harry / Los secretos de Harry / Los enredos de Harry | |
| 1998 | celebrity | El precio del éxito | |
| 1999 | Sweet and lowdown | Acordes y desacuerdos / Dulce y melancólico / El gran amante | |
| 2000 | Small time crooks | Granujas de medio pelo / Ladrones de medio pelo / Pícaros ladrones | |

https://es.wikipedia.org/wiki/Anexo:Filmograf%C3%ADa_de_Woody_Allen, se paso a archivo de texto y quedo así.

O bien la versión textual de Gerard Salton en Wikipedia, que se ve así, así como otros archivos, por ejemplo, un típico ServeLog, o de un típico servicio Twitter, que pueden ser parte de un corpus.

Archivos de un típico servicio Twitter

```
34952194402811904Save BBC World Service from Savage Cuts http://www.petitionbuzz.com/petitions/savews
34952186328784896a lot of people always make fun about the end of the world but the question is.."ARE U READY FOR IT?..
34952041415581696ReThink Group positive in outlook: Technology staffing specialist the ReThink Group expects revenues to be
34952018120409088'Zombie' fund manager Phoenix appoints new CEO: Phoenix buys up funds that have been closed to new business
34952008683229185Latest:: Top World Releases http://globalclassified.net/2011/02/top-world-releases-2/
34951899295920129CDT presents ALICE IN WONDERLAND - Catonsville Dinner has posted 'CDT presents ALICE IN WONDERLAND' to the.
34951860221648896Territory Manager: Location: Calgary, Alberta, CANADA Job Category: bu... http://bit.ly/e3o7mt #jobs
34951846736953344BBC News - Today - Free school funding plans 'lack transparency' - http://news.bbc.co.uk/today/hi/today/new
34951766319706112Manchester City Council details saving cuts plan: http://bbc.in/fYPYPC ...Depressing. Apparently we're 4th
34951749731090432http://bit.ly/e0ujdP, if you are interested in professional global translation services
34951546160553984Fitness First to float but isn't the full service model dead ? http://bit.ly/evflEB
34951513591783424David Cook ! http://bit.ly/fkj2gk has the mostest beautiful smile in the world!
34951452208136192Piss off. Cnt stand lick asses
34951399884197888BEWARE THE BLUE MEANIES: http://bit.ly/hu8iJz #cuts #thebluemeanies
34951141590568960Como perde os dentes no World Of Warcraft - Via Alisson http://ow.ly/1beBPo
34951099060461568How exciting! RT @BunchesUK: Hello! What's happening in your world? We're all gearing up for #Valentines wi
34951007502995456I'd very much appreciate it if people would stop broadcasting asking me to add people on BBM.
34950989601574912@samanthaprabu sam i knw u r a cricket fan r u watching any of the world cup matches
```

Con estos dos ejemplos, podemos afirmar que no es fácil con las herramientas existentes poder siquiera plantearse una consulta en SQL y el uso de bases de datos relacionales. Y evidentemente, existen muchos datos y de las más diversa índole, por ejemplo:

Datos de las cajas negras, que son dispositivos en helicópteros, aviones, y permiten captar audios de la tripulación, de las torres de vuelo, y la información sobre el rendimiento de la aeronave, entre otros.

Datos de Medios de Comunicación Social, son datos que se recuperan de Facebook, o Twitter con opiniones de millones de personas en todo el mundo, sobre algún caso en particular, originando problemas muy interesantes de resolver. Por ejemplo, aprender acerca de la respuesta de campañas publicitarias o promociones.

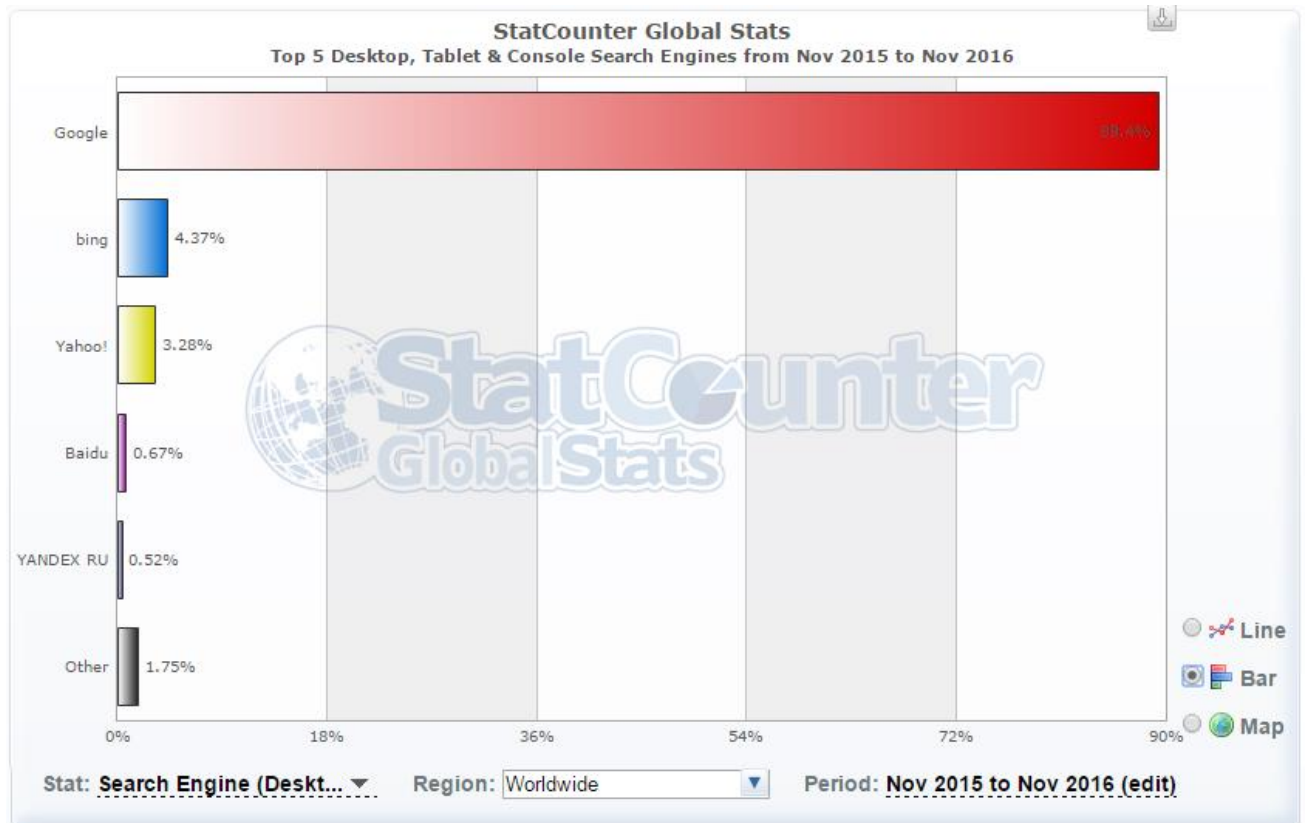
Datos de Bolsa de Valores, son datos que contienen información acerca de la toma de decisiones ante algún negocio.

Datos de una Red Eléctrica, son datos que entregan el suministro de electricidad consumida por un nodo en particular con respecto a una estación base.

Datos de Transporte, que son datos que incluye un modelo de datos, la capacidad, tiempos de recorrido, la distancia y la disponibilidad de un bus. (Por ejemplo, el transantiago)

Entre los motores de búsqueda de datos más importantes esta Google. Aquí se muestra la visión general del mercado de los motores de búsqueda.

<https://www.youtube.com/watch?v=BNHR6IQJGZs>



http://gs.statcounter.com/#search_engine-ww-monthly-201511-201611-bar

Con toda la diversidad de datos, podemos afirmar que los datos son en general de tres tipos. *Datos estructurados*, esto es datos relacionales usando bases de datos relacionales. *Semi estructurado* de datos, con datos XML. *Datos no estructurados*: Word, PDF, Texto, registros medios, etc. Y en cada uno de estos ámbitos debemos tener la forma de como indexarlos, para su posterior análisis basado en la consulta realizada.

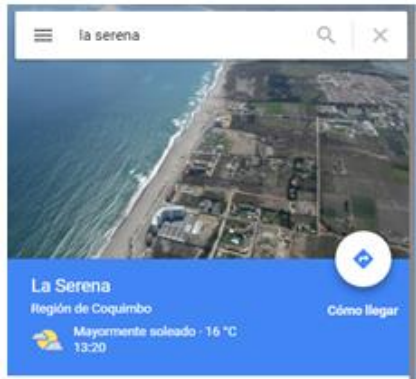
De ahí la importancia de herramientas que puedan indexar la información de textos o en lenguaje natural (sin formato) para luego en base a consultas de un tipo bien particular recuperar la información. Aunque la información puede ser muy diversa, tal como se puede apreciar más abajo. Ya que, se solicitó información sobre La Serena y nos muestra su Descripción, Imágenes, Ubicación, etc.



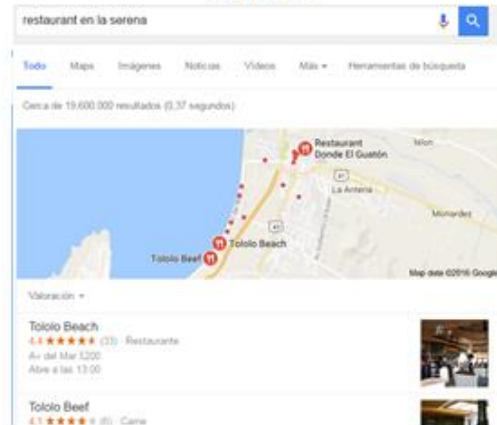
Descripción



Imágenes

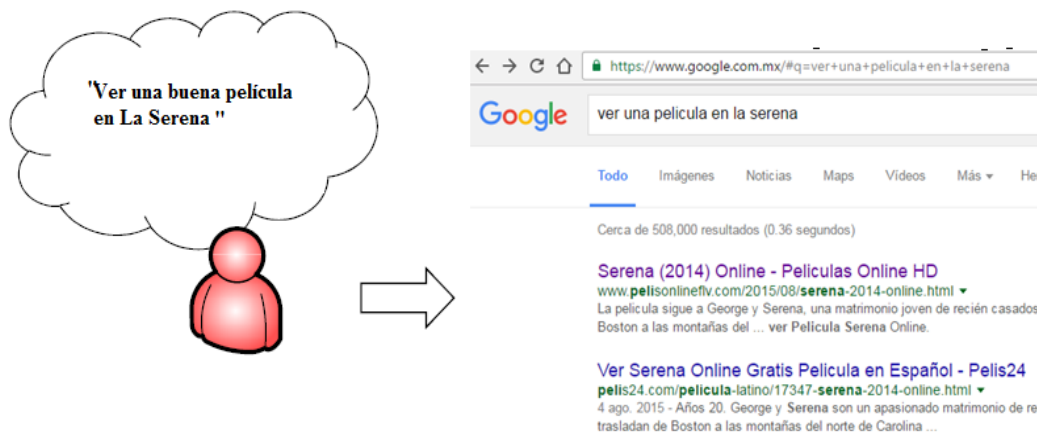


Ubicación

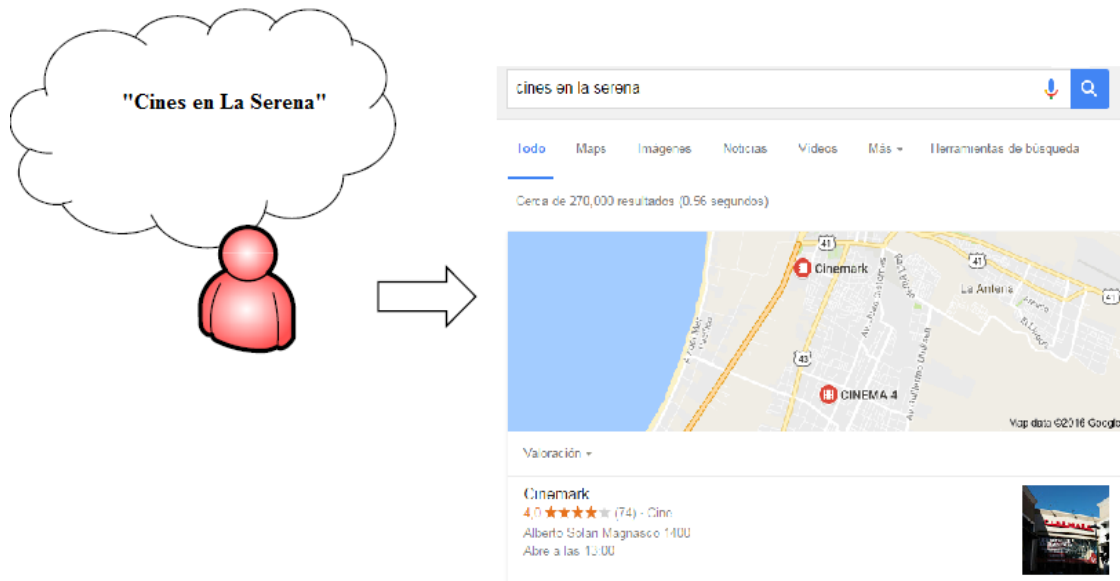


Listado lugares para comer

Pero la realidad puede ser aún más dura!!, ante una consulta plausible de necesidad de información y planteada de buena forma a los motores de búsqueda, bien puede suceder que los buscadores no puedan determinar el contexto, ya que no distingue entre La Serena, la ciudad y Serena, la película, tal como se muestra a continuación.



De manera que debemos nosotros mismos, el usuario, acometer una nueva consulta, y esperar que la respuesta coincida con la necesidad de información que anda buscando. Por lo tanto se detecta un problema de los sistemas de recuperación, y por otra parte del usuario, sino sabe exactamente que anda buscando, ya que de lo contrario perderá su tiempo.



Otro ejemplo, ante la consulta “*Madonna and Child*”, podemos ver este resultado



Las métricas en IR son valores que permiten medir la percepción de "éxito" del sistema de IR frente a una consulta. En (Baeza-Yates et al., 1999; Manning et al., 2008), se presentan dos medidas básicas de IR.

Para explicar la relevancia, se puede afirmar que muchos documentos (redundantes) satisfacen al usuario; el usuario no los quiere todos, sino algunos de ellos, esto es los documentos más relevantes, el mejor. Se origina entonces un ranking con documentos conocidos entre los *relevantes /no-relevantes*, junto con una evaluación métrica como salida con un marcador de calidad. La métrica es una medida de calidad que operan sobre un ranking de conocimientos o tópicos de documentos relevantes y no relevantes, de esta manera se define (**P**) precisión y (**R**) recall como una métrica de evaluación, sin embargo, esto ocurre siempre y cuando se esté

aplicando el modelo booleano. En rigor, si se tiene un conjunto de documentos relevantes (**REL**) y el conjunto de documentos recuperados (**RET**), entonces Precisión, es la proporción de documentos recuperados que son relevantes, mientras que Recall, es la proporción de documentos relevantes que son recuperados.

$$\mathbf{P} = \frac{|RET \cap REL|}{|RET|} \quad \mathbf{R} = \frac{|RET \cap REL|}{|REL|}$$

Por ejemplo, en P si se recupera un porcentaje de documentos (ítems) que son relevantes, digamos 100 recuperados, y 25 relevantes, entonces P posee 25% de P. En cambio si por otra parte la proporción de documentos relevantes que son recuperados es de 50 buenas respuestas en el sistema, con 25 recuperados, entonces tenemos un 50% de R. Ideal es 100% de P y 100% de R.

Frecuencia de términos TF-IDF.

TF-IDF significa "Término Frecuencia, Inverse Document Frequency", es una manera de contar la importancia de las palabras (o "términos") en un documento basado en la frecuencia con la que aparecen en varios documentos. Intuitivamente, si una palabra aparece con demasiada frecuencia en los documentos, es importante. Darle a la palabra una alta puntuación parece ser bueno, pero si una palabra aparece en muchos documentos, no es un identificador único, luego se le asigna a la palabra una puntuación más baja.

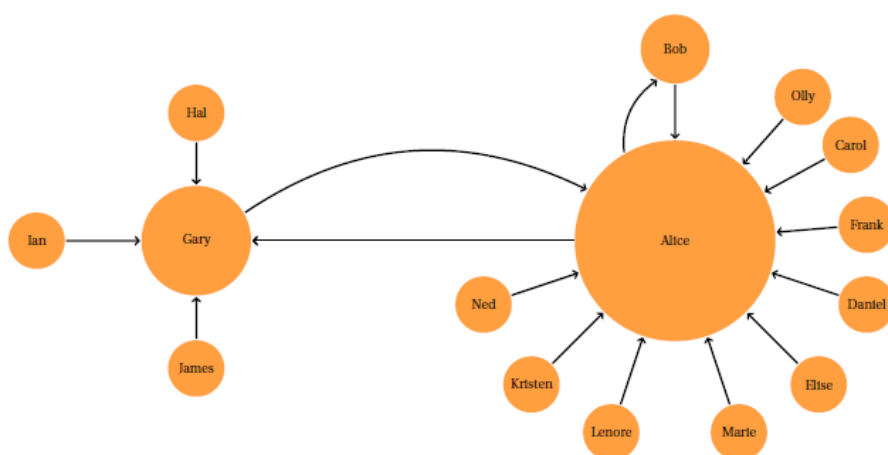
Por lo tanto, palabras comunes como "el" y "para", que aparece en numerosos documentos, será reducida. Las palabras que aparecen con frecuencia en un único documento será ampliado. El cálculo de TF-IDF permite establecer un peso de los términos de la colección.

Algunas reflexiones: ¿Cómo puede procesar una empresa los reclamos que le dejan sus usuarios en sus portales Web? ¿Cómo podemos agrupar los comentarios emitidos en un foro y analizar las opiniones de la gente?, entre otras.

Otra aplicación interesante se refiere al estudio o análisis de los sentimientos con datos de Twitter. Y es aquí donde Text Mining puede extraer conocimiento a partir del texto genérico. La minería de textos es un área multidisciplinaria basada en la recuperación de información, minería de datos, aprendizaje automático, estadísticas y la lingüística computacional. Como la mayor parte de la información (más de un 80%) se encuentra actualmente almacenada como

texto, se cree que la minería de textos tiene un gran valor comercial para responder las preguntas que antes nos planteábamos.

A menudo, queremos saber quién es la persona más importante de la red. El tema “importancia” puede no ser tan sencilla como parece, ya que existen varias dimensiones a lo largo de las cuales se puede considerar importante. Las medidas de importancia en las redes sociales se denominan “medidas de centralidad”. En este modelo podemos interpretar el concepto de centralidad. ¿Quién es más popular Alice o Gary?.



Extraído del libro Twitter Data Analytics (SpringerBriefs in Computer Science) 2014th Edition Shamanth Kumar, Fred Morstatter, Huan Liu, 2013.

La **Ley de Zipf** es uno de los más importantes, ya que está relacionado con saber la frecuencia con la que una palabra aparece en un texto, en general afirma que un pequeño nº de palabras son utilizadas con mucha frecuencia, mientras que frecuentemente ocurre que un gran nº de palabras son poco empleadas.

En los años cuarenta, el lingüista George Zipf se dio cuenta de que las palabras y su número de apariciones en textos, seguían alguna ley especial. La palabra más utilizada ocuparía el número uno en el ranking, el número dos se corresponde con la segunda palabra más veces repetida, etc. Así, se guardaba una estrecha relación entre el número de apariciones de las palabras más frecuentes. La primera palabra más utilizada aparecía el doble de veces que la segunda y tres veces más que la tercera, y sigue el patrón según esta norma.

Ejemplo. Moby Dick; o bien, The Whale by Herman Melville, la palabra más frecuente fue “**the**” con 14.620 apariciones, la segunda es “**of**” con 6.732 apariciones, y la tercera “**and**” aparece 6.502 veces, la cuarta es “**a**” con 4.776 apariciones.

Los Top 30 de Moby Dick

| Rango | Término | Frecuencia | Rango | Término | Frecuencia |
|-------|---------|------------|-------|----------|------------|
| #1 | the -> | 14620 | #16 | for -> | 1646 |
| #2 | of -> | 6732 | #17 | was -> | 1646 |
| #3 | and -> | 6502 | #18 | all -> | 1543 |
| #4 | a -> | 4776 | #19 | this -> | 1443 |
| #5 | to -> | 4706 | #20 | at -> | 1335 |
| #6 | in -> | 4230 | #21 | whale -> | 1232 |
| #7 | that -> | 3099 | #22 | by -> | 1226 |
| #8 | it -> | 2535 | #23 | not -> | 1171 |
| #9 | his -> | 2530 | #24 | from -> | 1105 |
| #10 | i -> | 1989 | #25 | on -> | 1073 |
| #11 | he -> | 1878 | #26 | him -> | 1067 |
| #12 | but -> | 1823 | #27 | so -> | 1066 |
| #13 | with -> | 1770 | #28 | be -> | 1064 |
| #14 | as -> | 1753 | #29 | you -> | 946 |
| #15 | is -> | 1750 | #30 | one -> | 925 |

El término más frecuente representa el 10% del texto, el segundo término más frecuente representa el 5% del texto, el tercero más frecuente representa el 3%, etc. Una de las aplicaciones típicas es en el campo de la economía urbana, la dinámica demográfica y en particular la distribución del tamaño de las ciudades, en donde ha sido un tema de investigación que ha atraído mucho la atención durante las últimas décadas.

En la literatura y en particular en el tema de recuperación de la información se ha enfocado en mostrar si empíricamente se cumplen las leyes de **Zipf** y otras.

En particular, la ley de Zipf establece que el tamaño de las ciudades sigue una cierta distribución (llamada distribución de Pareto con coeficiente igual a 1). En la práctica esto significa que la ciudad de mayor tamaño (por ejemplo, Santiago de Chile) debería ser dos veces más grande que la segunda ciudad en cuestión (Antofagasta), y tres veces más que la tercera (Valparaíso), y así sucesivamente....

Big data, macro datos o datos masivos es un concepto que hace referencia al almacenamiento de grandes cantidades de datos y a los procedimientos usados para encontrar patrones repetitivos dentro de esos datos.

El fenómeno del **Big data** también se denomina a veces datos a gran escala. Empecemos por tratar de aclarar "qué es *Big Data*". Se llama *Big Data* a la gestión y análisis de enormes volúmenes de datos que no pueden ser tratados de manera convencional, ya que superan los límites y capacidades de las herramientas de software habitualmente utilizadas para la captura, gestión y procesamiento de datos.

Recordemos que el Data Warehouse es una tecnología para el manejo de la información construido sobre la base de optimizar el uso y análisis de la misma utilizado por las organizaciones para adaptarse a los vertiginosos cambios en los mercados. Su función esencial es ser la base de un sistema de información gerencial, es decir, debe cumplir el rol de integrador de información proveniente de fuentes funcionalmente distintas (Bases Corporativas, Bases propias, de Sistemas Externos, etc.) y brindar una visión integrada de dicha información, especialmente enfocada hacia la toma de decisiones por parte del personal jerárquico de la organización.

Para ello, se usan tecnología y sistemas OLTP (On-Line Transaction Processing) y OLAP (On-Line Analytical Process). En resumen, son aplicaciones que se encargan de analizar datos del negocio para generar información táctica y estratégica que sirve de soporte para la toma de decisiones, en donde las transacciones OLTP utilizan **Bases de Datos Relacionales u otro tipo de archivos**, OLAP logra su máxima eficiencia y flexibilidad operando **sobre Bases de datos Multidimensionales**.

De manera que Big Data es la evolución de la tecnología y dado los menores costos del almacenamiento han hecho que los volúmenes manejados por estas aplicaciones hayan aumentado de manera muy importante. Sin embargo, a los conceptos de **Volumen, Variedad y Velocidad** (3Vs) se les ha incluido hoy en día nuevas características como son la **Veracidad y Valor del dato** (5Vs)

Se habla de Big Data cuando los volúmenes superan la capacidad del software habitual para ser manejados y gestionados. Por otro lado, se habla de grandes volúmenes para cuando nos referimos a tratamientos de Terabytes o Petabytes de datos que pueden surgir de por ejemplo logs, imágenes satelitales, o en general información obtenida por diferentes Redes Sociales, en donde el número cada vez mayor de dispositivos electrónicos conectados, la explotación de sensores permiten conocer los movimientos y hábitos de vida, de información externa de diversas fuentes, etc.

La información que procesan los Datawarehouse es información estructurada que ha pasado por numerosos filtros de calidad para poder garantizar que la información de salida tiene una precisión y una exactitud determinada, sin embargo, cuando hablamos de *Big Data* nos referimos a información que puede estar semiestructurada o no tener ninguna estructuración.

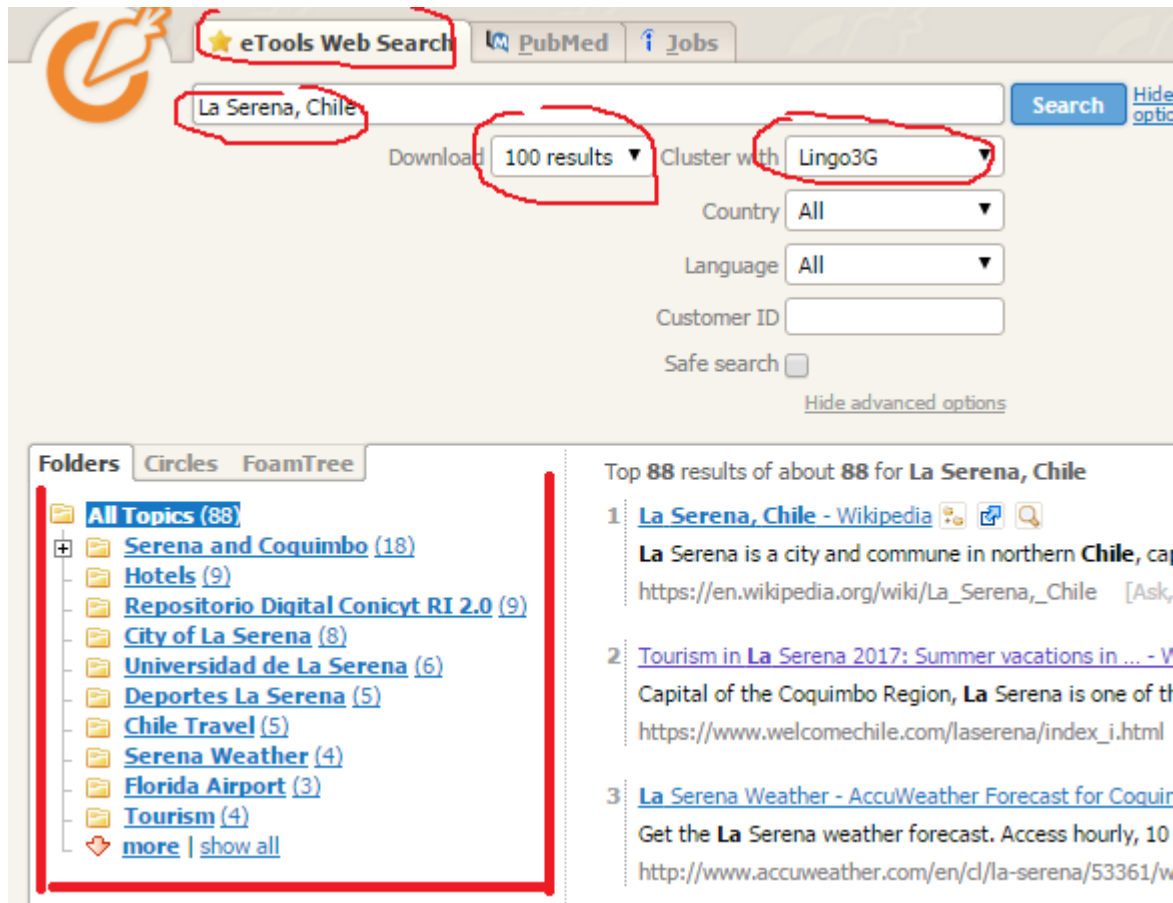
La gestión y administración de esta información desestructurada precisa de una tecnología diferente y permite tomar decisiones basadas en información que tiene importantes grados de inexactitud.

Finalmente, se añade el valor. La importancia del dato para el negocio, saber que datos son los que se deben analizar, es fundamental. Tanto que ya se empieza a hablar del científico de datos (KPI experto, recordar que un **KPI (key performance indicator)**, conocido también como indicador clave o medidor de desempeño o indicador clave de rendimiento), un profesional con perfil científico, tecnológico...pero con visión de negocio. y es aquí, en donde tiene sentido nuestra asignatura.

La trama técnica en las consultas: Consultas aproximadas: Un K-gram es una subsecuencia de n elementos de una secuencia dada. El estudio de los n-grama es interesante en diversas áreas del conocimiento. La forma en que extraemos los gramas se tiene que adaptarse al ámbito que estamos estudiando y al objetivo que tenemos en mente. Se puede usar gramas para casi todos los ámbitos. K-gram: Un inverted index para una secuencia de k caracteres contenidos en una palabra. Por ejemplo, 3-grams para “index”: \$in, ind, nde, dex, ex\$ (donde \$ es un carácter especial que denota la partida o final de una palabra). Para todo k-gram encontrado en el diccionario, el k-gram index tiene un puntero a todas las palabras que contiene el **k-gram**. Por ejemplo, dex → {index, **d**exterity, ambid**ex**trous}.

K-means es un método de agrupamiento, que tiene como objetivo la partición de un conjunto de n observaciones en k grupos en el que cada observación pertenece al grupo cuyo valor medio es más cercano. Es un método utilizado en minería de datos. La agrupación del conjunto de datos puede ilustrarse en una partición del espacio de datos en celdas de Voronoi.

Lingo3G es un componente de software que organiza las colecciones de documentos de texto en carpetas temáticas llamados clústeres que están claramente jerárquica de. En tiempo real, de forma totalmente automática y sin bases de conocimiento externo.



<https://carrotsearch.com/lingo3g-overview.html>

lo muestra metabuscadores y que se refieren a motores de búsqueda más sofisticados que permiten incluir la búsqueda en otros motores de búsqueda. Por ejemplo,

Respecto a la visualización de la información recuperada

Nube de palabras. Una nube de palabras o nube de etiquetas (Word cloud) es una representación visual de las palabras que conforman un texto, en donde el tamaño es mayor para las palabras que aparecen con más frecuencia. Uno de los principales usos es la visualización de las etiquetas de un sitio web, de modo que los temas más frecuentes en el sitio se muestren con mayor prominencia, ante una demanda de información. Las etiquetas son palabras clave que suelen estar ordenadas alfabéticamente o, en ocasiones, agrupadas semánticamente.

Problemática: Tengo un texto y deseo saber el ámbito aproximado de su contenido, expresado en una nube de palabras con el fin de acotar el sentido del texto. **VIRIs (Visual Information Retrieval Interfaces).**

<https://www.jasondavies.com/wordcloud/#%2F%2Fdns.uls.cl%2F~ej%0A>



<http://tagcrowd.com/>

Conclusión del capítulo:

De manera que la IR nos lleva a preguntarnos. ¿Cómo es un motor de búsqueda?. Cómo el motor de búsqueda predice información relevante o desecha la no- relevante?. ¿Qué tan bueno es el motor de búsqueda?. Entonces ...cómo predecir la relevancia?. Es decir, ¿qué tipo de evidencias pueden ser usadas para predecir que un documento es relevante a una query?