

Expansión de consultas utilizando recursos lingüísticos para mejorar la recuperación de información

Mg. Claudia Deco

Departamento de Sistemas e Informática
Facultad de Ciencias Exactas, Ingeniería y Agrimensura
Universidad Nacional de Rosario

Rosario, Argentina



Recuperación de Información - 1

Buscar material en bibliotecas ó centros de información

- Fichero manual
- Catálogo informatizado de cada biblioteca
- **Especialista:** encargado de expresar la necesidad de información del usuario mediante una estrategia de búsqueda
- Problemas
 - Tiempo** elevado de **búsqueda** (ir de biblioteca en biblioteca)
 - Información desactualizada e insuficiente**

Recuperación de Información - 2

Búsqueda actual

- Catálogo online de cada biblioteca
- Bancos de datos remotos online

Continúa presente el especialista en ciencias de la información

- Internet
 - Acceso a bases de datos
 - Páginas web
 - Buscadores (Altavista, Yahoo, Google, ...)
- Problemas
 - Ausencia** del especialista
 - Información en exceso**
 - Tiempo elevado de búsqueda**

Recuperación de Información - 3

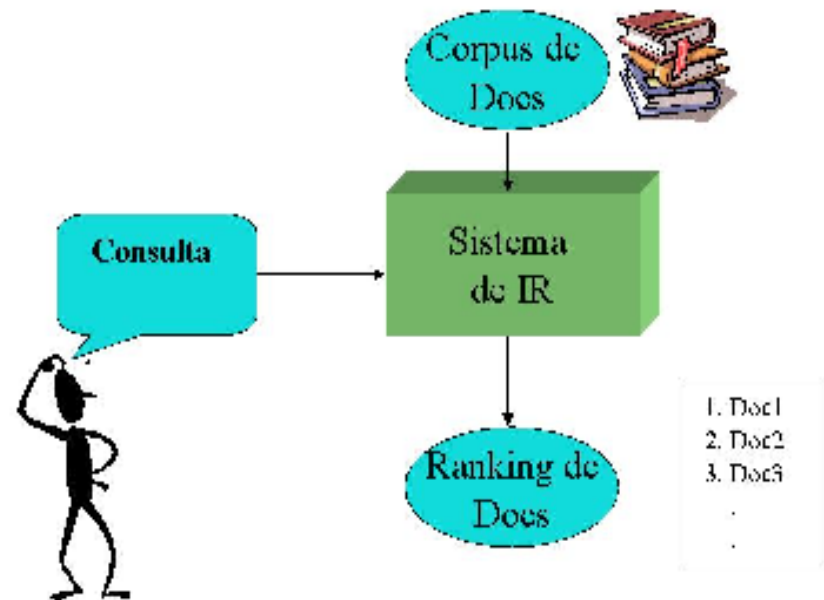
Tarea típica de la RI

Dados:

- ◆ Un **corpus** de documentos textuales en lenguaje natural.
- ◆ Una **consulta** de usuario en la forma de un string de texto.

Encontrar:

- ◆ datos con la **mejor coincidencia** con el patrón dado
- ◆ **ordenados** según su **relevancia** para la consulta





Problemas

- ◆ Encontrar **información útil** en la **Web** es frecuentemente una **tarea difícil**.
- ◆ **Problemas** para la búsqueda en la web.
 - con los **datos**.
 - con los **usuarios**.



Problemas

Problemas con los datos

- **Datos distribuidos.**
- **Datos volátiles.** (dinámica de Internet)
- **Gran volumen.** (crecimiento exponencial de la Web).
- **Datos no estructurados y redundantes** (30% duplicado).
- **Calidad de los datos.** (no hay proceso ni control editorial)
- **Datos heterogéneos.** (estructural, semántica)

Problemas con los usuarios

- **Cómo especificar la consulta**
- **Cómo interpretar las respuestas obtenidas**



Algunas estadísticas

- ◆ promedio de **palabras por consulta**: 2 palabras
- ◆ **operadores lógicos por consulta**: 0,4.
- ◆ el 80% de los usuarios **no modifica su consulta inicial**
- ◆ el 85% **ve sólo la primera página** de la respuesta.
- ◆ el 85% de los usuarios **utiliza motores de búsqueda**
- ◆ los usuarios **no están conformes** con los resultados obtenidos



Algunas estadísticas

Esto indicaría

- la mayoría de los **usuarios desconoce técnicas de RI**
 - tiene **dificultad de expresar claramente** su necesidad de información
- ⇒ **no obtienen los resultados deseados.**



Ejemplo Motivador

“Utilización de la aspirina en el cáncer”

Nuestro enfoque es
la expansión de los conceptos involucrados:

{ aspirina, cáncer }

- Dominio: Cáncer: ¿medicina? ¿horóscopo? ¿astrología? ...
- Especificidad: ¿Interesa todo tipo de cáncer ó algún tipo particular?
- Equivalencias semánticas: Existen sinónimos de cáncer que pueden no recuperarse, por ejemplo neoplasma.

Expansión de consultas

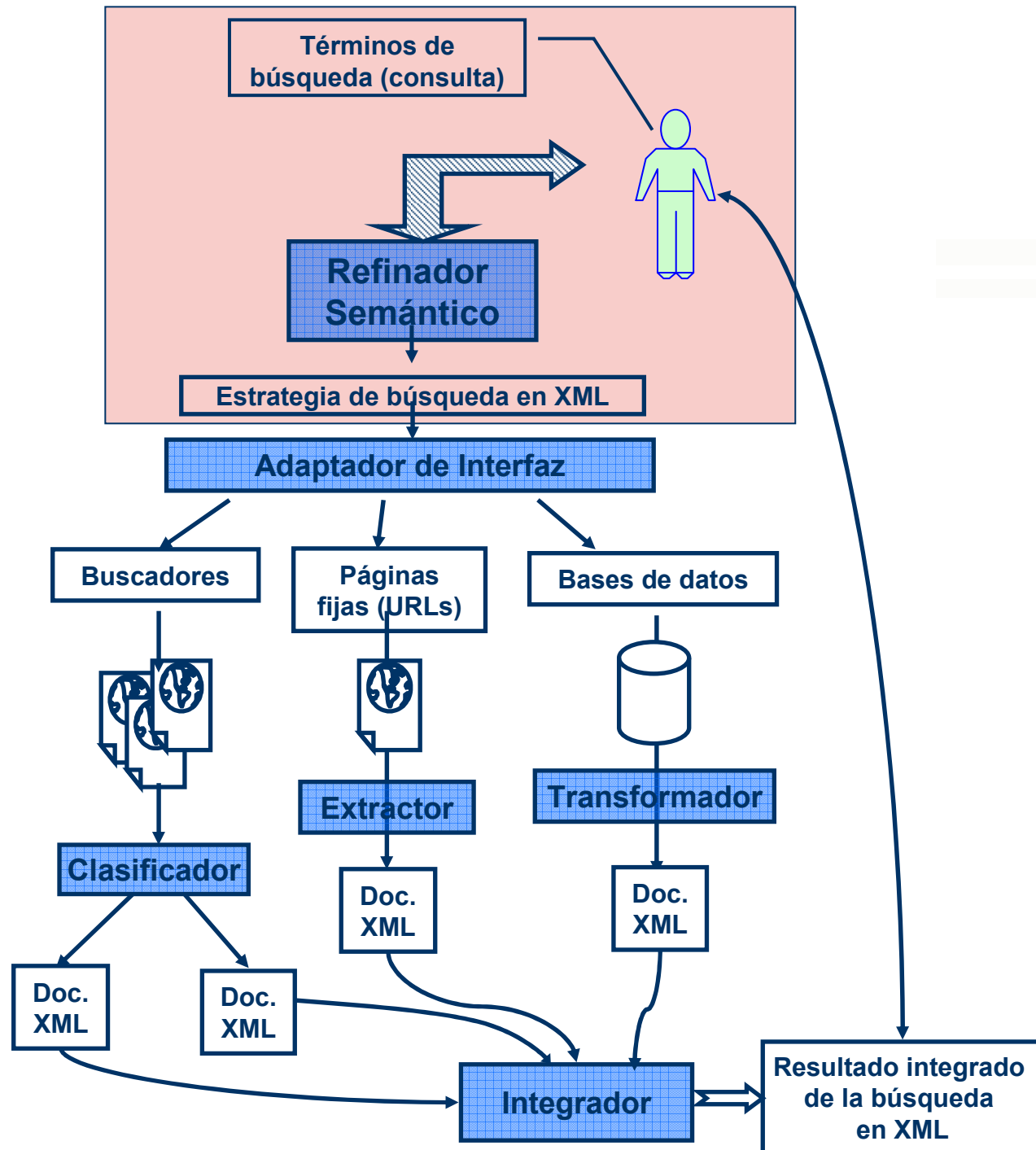
- ◆ es el proceso de suplementar la consulta original con términos adicionales.
- ◆ es un método para mejorar la recuperación.
- ◆ puede ser desarrollada manual, automática o interactivamente.

Se propone una **expansión de la consulta semiautomática**:

El **sistema sugiere** términos

y es el **usuario** quien **decide** cuál ó cuáles representan su interés de búsqueda.

Contexto





Refinamiento Semántico para Recuperación de Información desde la Web

Que le permita al usuario

- **Desambiguar** sentidos de los conceptos
- **Seleccionar** conceptos jerárquicamente relacionados
- **Expandir** semánticamente cada concepto

para preparar una Estrategia de Búsqueda adecuada a su necesidad de información

Estrategia de Búsqueda

Es una **expresión lógica** compuesta por distintos conceptos combinados con conectores lógicos AND, OR y NOT.

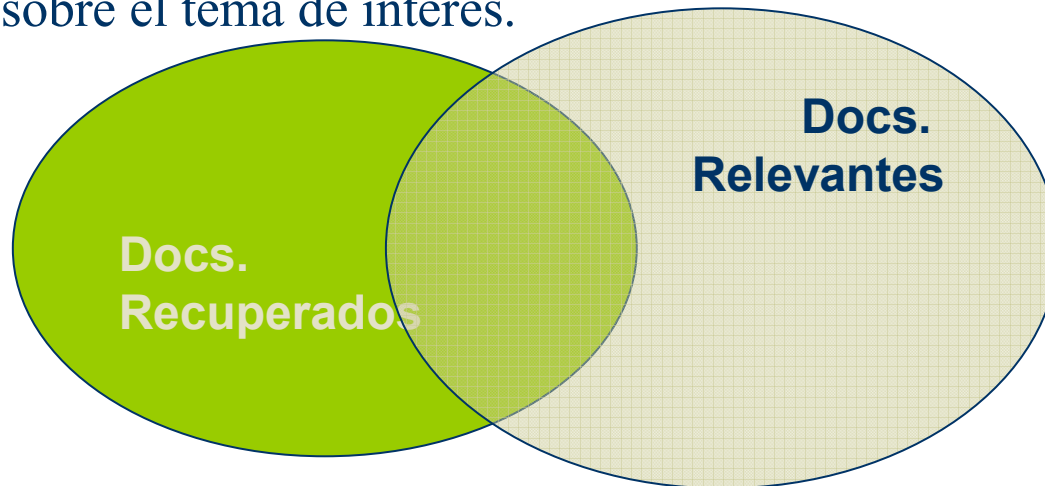
- ◆ Indicadores para evaluar:

$$\text{Precisión} = \frac{\text{Número de documentos relevantes recuperados}}{\text{Número de documentos recuperados}}$$

$$\text{Recall} = \frac{\text{Número de documentos relevantes recuperados}}{\text{Número total de documentos relevantes}}$$

Realizada una búsqueda:

el conj. de docs recuperados **no coincide** totalmente con el conj. de docs relevantes sobre el tema de interés.



Una **búsqueda** será **óptima** cuando estos **dos conjs. coincidan**

→ todos los docs recuperados sean relevantes
y todos los docs relevantes sean recuperados.



Estrategia de Búsqueda

En su preparación pueden ocurrir contingencias debido a:

- el uso de términos **ambiguos** o no específicos
- la utilización de términos **demasiado específicos**,
- la **falta** de conceptos
- el uso incorrecto de la disyunción (**OR**) y la conjunción (**AND**)
- el uso incorrecto de la negación (**NOT**)
- el uso incorrecto de **paréntesis**
- **no** incluir **sinónimos** suficientes,
- errores de tecleo,
- errores de deletreo (“color” y “colour”),



Recursos lingüísticos que se utilizan en la RI

- Diccionarios
- Diccionarios multilinguales
- Ontologías
- Tesoros



Recursos lingüísticos que se utilizan en la RI

Diccionarios

- Indican las **distintas acepciones** de un término
- Permiten su **expansión** con **sinónimos**
- y sugieren términos **relacionados jerárquica y/o semánticamente**

Diccionarios multilinguales

- Permiten **traducir** un concepto a otros idiomas.



Recursos lingüísticos que se utilizan en la RI

Tesauros

- Permiten el control del vocabulario para **representar en forma unívoca** cada concepto.
- **Términos relacionados** semántica y genéricamente, los cuales cubren un **dominio específico** del conocimiento.
- **Interrelaciones:** jerárquicas, de afinidad y preferenciales (sinónimos – homónimos)



Recursos lingüísticos que se utilizan en la RI

Ontologías

- Permiten representar el conocimiento en la web.
- Definen conceptos y relaciones de algún dominio.
- Consisten de términos, sus definiciones y axiomas.
 - Los axiomas permiten **inferir conocimiento** que no esté indicado explícitamente en la taxonomía de conceptos.



Ejemplos

WordNet (www.cogsci.princeton.edu/~wn/)

- ◆ Es un **sistema de referencia léxica** online
 - incluye sinónimos, variantes de deletreo, ampliación de siglas, y para ciertos términos su escritura en otros idiomas.
 - tiene relaciones jerárquicas: muestra para cada término
 - ◆ sus términos específicos ó hipónimos, y
 - ◆ su término más amplio ó hiperónimo.

WordNet

Results for Hyponyms of noun "cancer"

Sense 1

cancer, malignant neoplastic disease --

(any malignant growth or tumor caused by abnormal and uncontrolled cell division; it may spread to other parts of the body through the lymphatic system or the blood stream)

=> lymphoma --

(a neoplasm of lymph tissue that is usually malignant; one of the four major types of cancer)

=> carcinoma --

(any malignant tumor derived from epithelial tissue; one of the four major types of cancer)

=> liver cancer, cancer of the liver --

(malignant neoplastic disease of the liver usually occurring as a metastasis from another cancer; symptoms include loss of appetite and weakness and bloating and jaundice and upper abdominal discomfort)

=> adenocarcinoma, glandular cancer, glandular carcinoma --

(malignant tumor originating in glandular epithelium)

=> prostate cancer, prostatic adenocarcinoma --

(cancer of the prostate gland)

=> breast cancer --

(cancer of the breast; one of the most common malignancies in women)

.....



MeSH

"lung cancer" is not a MeSH term, but it is associated with the MeSH term **Lung Neoplasms**

Lung Neoplasms : Tumors or cancer of the LUNG.

Term **Lung Neoplasms** appears in more than one place in the MeSH tree.

All MeSH Categories

Diseases Category

Neoplasms

Neoplasms by Site

Thoracic Neoplasms

Respiratory Tract Neoplasms

Lung Neoplasms

Carcinoma, Bronchogenic

Coin Lesion, Pulmonary

Pancoast's Syndrome

Pulmonary Blastoma

¿Qué diferencia hay entre estos recursos?

- ◆ **Diccionario:** todos los sinónimos son representativos y tratados por igual.
- ◆ **Tesauro:** se tiene una **palabra clave** preferida y **representativa** del conjunto de sinónimos para cada concepto.
- ◆ **Ontologías:** se agregan los **axiomas**, que permiten realizar **inferencias sobre conceptos**.
- ◆ En **BD:** terminología está **controlada** y el uso de **tesauros** permite obtener un resultado más preciso.
- ◆ En **Web:** terminología **no** está **controlada**. Por esto el uso de **diccionarios** y de **ontologías** es más adecuado.



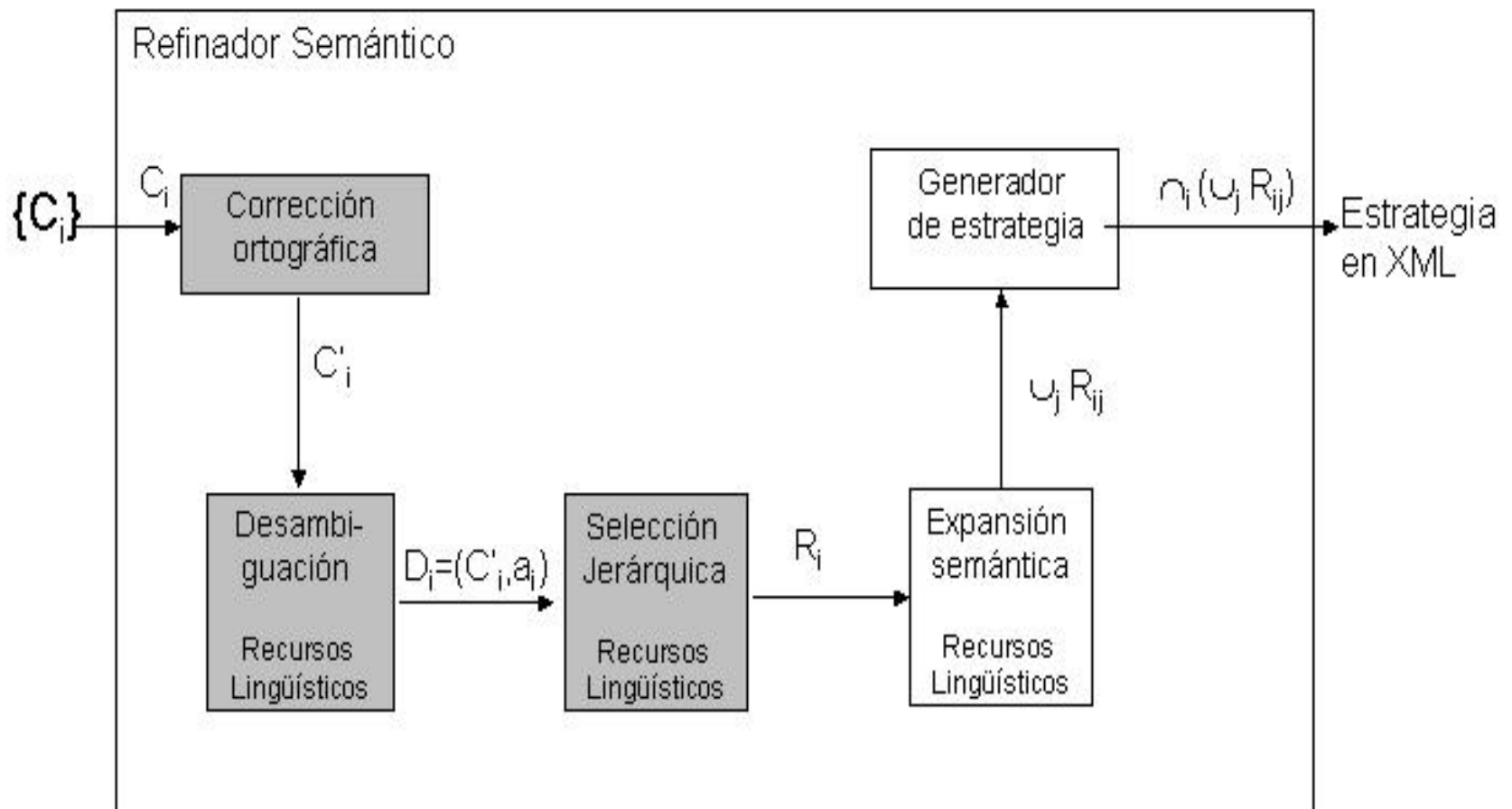
Refinamiento semántico

Una búsqueda que involucra los conceptos

$\{ C_1 , C_2 , \dots , \text{not } C_h , \dots , C_n \}$

resulta en una estrategia de búsqueda.

Arquitectura para la preparación de la Estrategia de Búsqueda





Refinamiento semántico



- Se representa en XML.
 - Porque es el estándar de intercambio de datos
 - La estrategia es independiente de la sintaxis de los lenguajes de consulta a las fuentes.
 - Bases de datos, Páginas web, Buscadores

Ventaja: el usuario no necesita conocer las distintas sintaxis de buscadores
- Un adaptador de interfaz transforma este XML a las sintaxis de las distintas fuentes.

Ejemplo resuelto con el RS

Para la búsqueda que involucra los conceptos:

cáncer de pulmón , aspirina , tratamiento

Una estrategia sería

(lung neoplasms OR lung cancer OR cáncer de pulmón OR carcinoma of the lungs)

AND

(aspirina OR aspirin OR ácido acetil salicílico)

AND

(tratamiento OR treatment)

dependerá de los RL utilizados



y el
XML
será

```
<estrategia>
```

```
<concepto 1>
```

```
<ampliación 1>cáncer de pulmón</ampliación 1>
```

```
<ampliación 2>lung cancer</ampliación 2>
```

```
<ampliación 3>lung neoplasms</ampliación 3>
```

```
<ampliación 4>carcinoma of the lungs </ampliación 4>
```

```
</concepto 1>
```

```
<concepto 2>
```

```
<ampliación 1>aspirina</ampliación 1>
```

```
<ampliación 2>aspirin</ampliación 2>
```

```
<ampliación 3>ácido acetil salicílico</ampliación 3>
```

```
</concepto 2>
```

```
<concepto 3>
```

```
<ampliación 1>tratamiento</ampliación 1>
```

```
<ampliación 2>treatment</ampliación 2>
```

```
</concepto 3>
```

```
</estrategia>
```



Prototipos



◆ Utilización de

- estándares y recomendaciones del grupo W3C y
- lenguajes y recursos libres disponibles en la web
 - Selección de jerarquía y la expansión semántica: MeSH – **WordNet**
 - Lenguaje para implementar: PHP
 - Buscador: Google - **Yahoo!**



Experimentación

- ◆ Se realizaron consultas.
- ◆ Niveles de usuarios: Inexperto, Medio, Experto.
- ◆ Cada usuario describió su interés de búsqueda
- ◆ Cada usuario planteó la estrategia directamente a un buscador
- ◆ Se planteó la estrategia generada por el refinador semántico
- ◆ Se registró:
 - la cantidad de documentos resultantes, y
 - la cantidad de docs que respondían al interés del usuario en los primeros 50 docs. \Rightarrow *precisión*

Karting Race	431.000	21	- Karting Race	"Karting Race"	3.660	26	Inexperto
Oedipus Complex	87.900	22	- Oedipus Complex	"Oedipus Complex" OR "Oedipal Complex"	53.100	33	Medio
bulimia treatment	403.000	47	- treatment - bulimia	treatment AND (bulimia OR "binge-eating syndrome")	408.000	47	Medio
software distributor usa	941.000	23	- software - distributor - usa	(software OR "software system" OR "software package" OR package) AND (distributor OR distributor) AND (usa OR "u.s.a." OR us OR "u.s." OR "United States of America" OR "United states")	2.250.000	21	Medio
software distributor usa	941.000	23	- software - distributor - usa	(software OR "software system" OR "software package" OR package) AND (distributor OR distributor) AND ("United States of America" OR "United states")	1.520.000	25	Medio
food without sugar	2.680.000	13	- food	"diabetic diet"	89.800	19	Inexperto
Yamaha virago model	34.400	14	- yamaha virago - model	"yamaha virago" AND model	7.800	48	Inexperto
penicillin discovery	76.200	7	- penicillin - discovery	penicillin AND (discovery OR find OR uncovering)	78.900	7	Medio

Algunos Resultados

- ◆ El usuario no utiliza la búsqueda por frases.
Ejemplo: Gabriel García Márquez

El uso de frases

- **aumenta** la **precisión**, y
 - **disminuye** la **cantidad** de docs recuperados.
- ◆ La corrección ortográfica realizada por el refinamiento
 - **aumenta** la cantidad de documentos recuperados y
 - **aumenta** la precisión.

Ejemplo: escherichia colli (12.000)

escherichia coli (8.500.000)

Resultados de la experimentación

Consultas con términos más específicos

treatment y Hodgkin

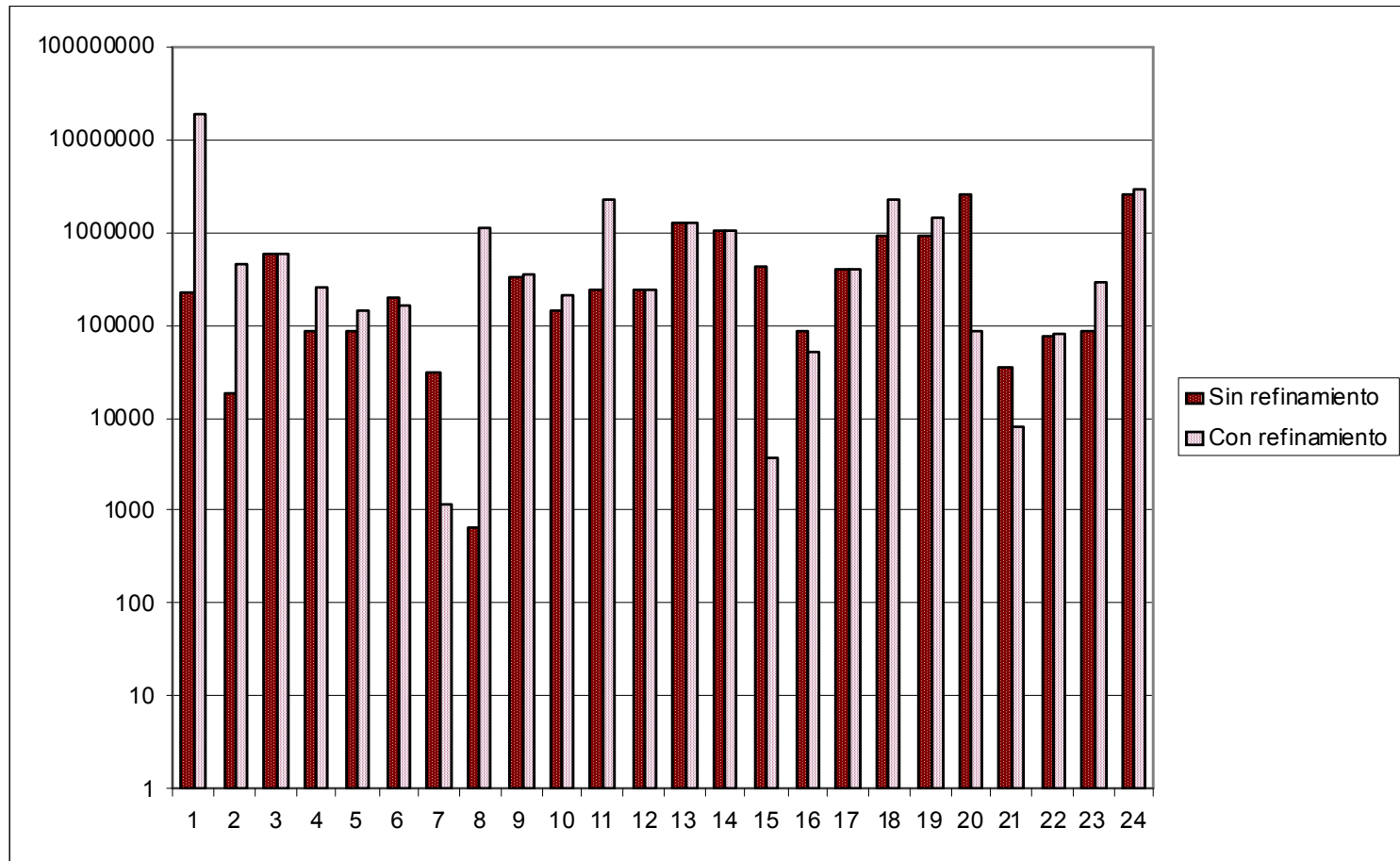
movimiento por jerarquía conceptual asociada

lymphoma

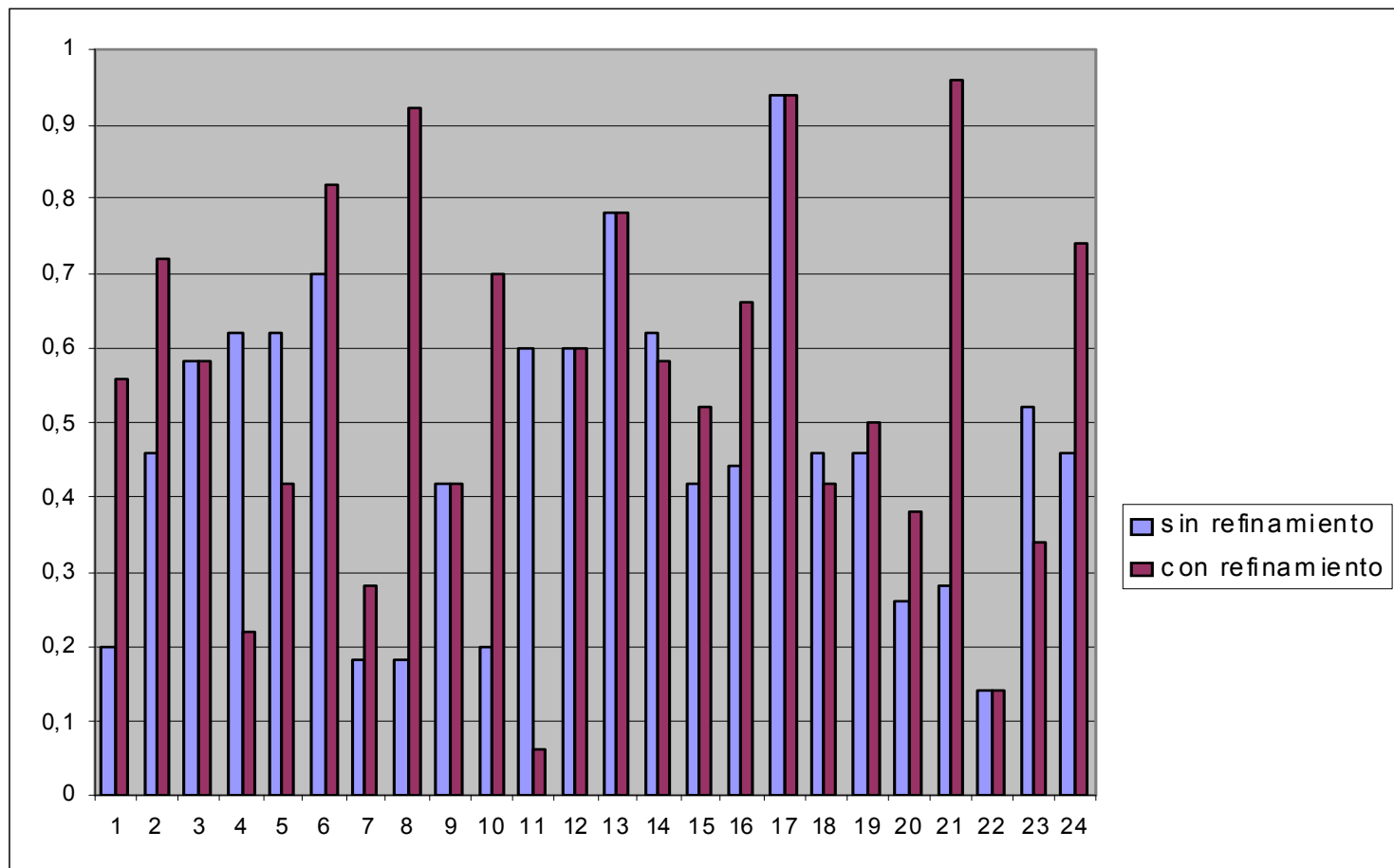
“hodgkin’s disease”

- ◆ **disminuye** la **cantidad** de docs recuperados.
- ◆ **aumenta** la **precisión**.

Cantidad de documentos resultantes con y sin refinamiento semántico en escala logarítmica



Precisión en los primeros 50 documentos, con y sin refinamiento semántico



Resultados de la experimentación

- ◆ Promedio cantidad de docs recuperados y la precisión sobre los primeros 50 resultados sin y con RS:

	Recuperados	Precisión
Sin refinamiento	533811,67	0,46
Con refinamiento	651726,67	0,55
	22,09 %	19,03 %

- ◆ El RS mejora:
 - la cantidad de documentos recuperados en un 22,09 %
 - y la precisión en un 19,03 %.



Conclusiones



El refinamiento

- ◆ Mejora la **cantidad de docs recuperados**
 - Al expandir cada concepto con sinónimos y términos relacionados
- ◆ Mejora la **precisión en los primeros n docs**
 - Al permitir al usuario moverse por jerarquías conceptuales



Conclusiones

- ◆ El RS **resuelve problemas** relacionados con las contingencias:
 - correcto uso de la disyunción y la conjunción,
 - uso correcto de paréntesis,
 - inclusión de sinónimos y variantes de escritura,
 - utilización de términos específicos,
 - uso correcto de la negación y
 - errores de tecleo.
- ◆ El refinamiento es **semiautomático** porque un esfuerzo inicial por parte del usuario le evita a posteriori la lectura y el descarte de los docs.



Se está trabajando en

- Búsqueda personalizada de objetos de aprendizaje
 - Utilización de perfiles de usuario.
 - Selección automática de recursos lingüísticos adecuados
 - Expansión multilingual.
 - Utilización de ontologías con axiomas. Incorporar conceptos obtenidos a través de la inferencia.
 - Feedback de relevancia. Mejorar estrategia incorporando conceptos extraídos de docs marcados como relevantes por el usuario.



Muchas gracias....