

Herramientas para construir ontologías

Rocío Abascal, Héctor Jiménez

Departamento de Tecnologías de la Información
Universidad Autónoma Metropolitana (Cuajimalpa)

1. *Motivación*
2. Ontologías
3. Algunas herramientas
 - (a) Terminología
 - (b) Thesauri e identificación de relaciones

Digitalización progresiva

Información { Estructurada Bases de Datos
 { No estructurada Texto, imágenes...

Internet (2003)^a

- 167 terabytes anuales de texto (\approx 167 millones de libros).
- 600 millones de personas tienen acceso.

^acibernauta.elcorreodigital.com, Peter Lyman et al.

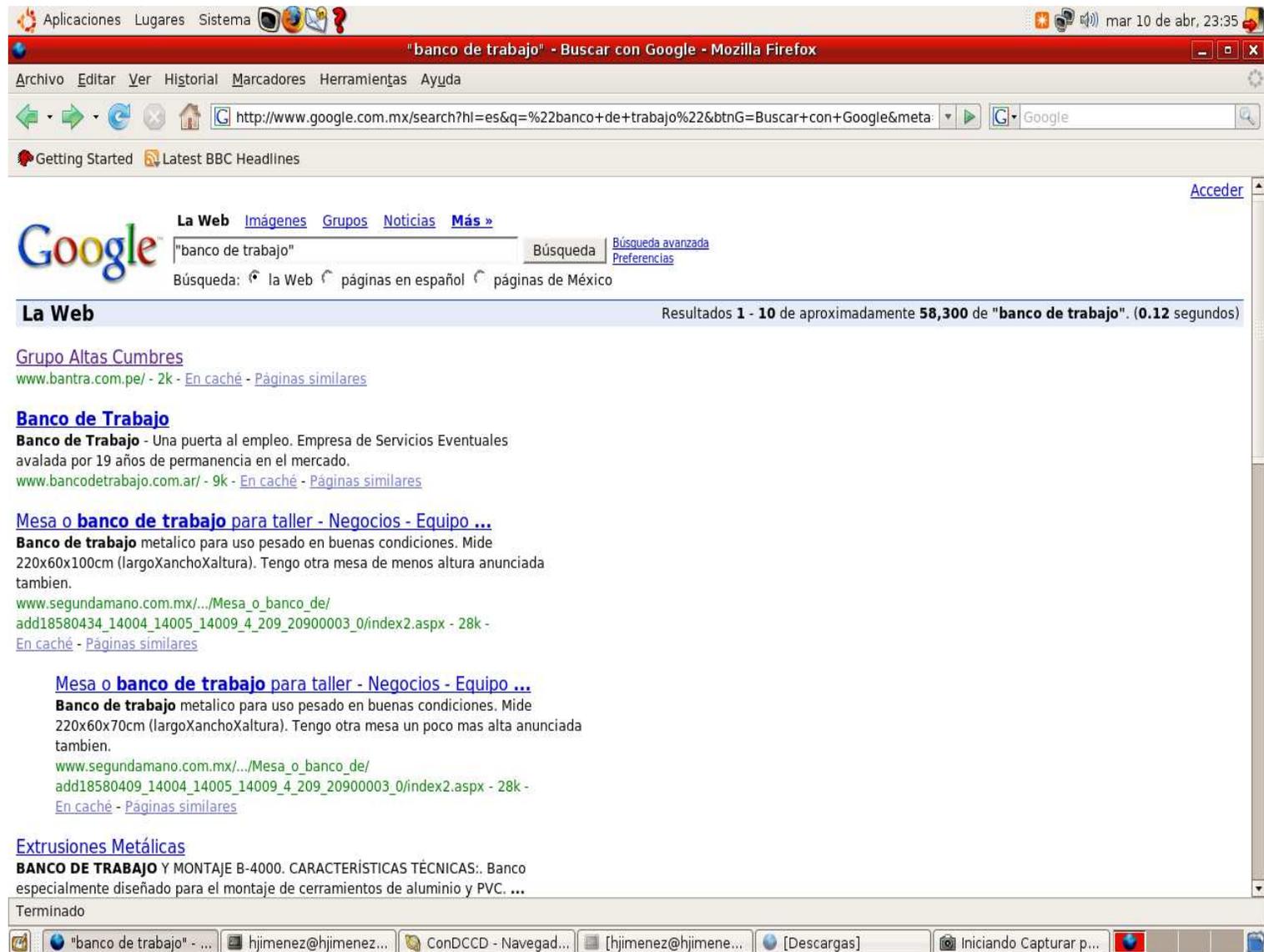
Problemas de Búsqueda

Cadenas. Actualmente hacemos búsquedas dando una cadena de caracteres: `banco` es una palabra ambigua y debe darse información adicional, `banco de trabajo` → 4,430,000 páginas (el buscado aparece primero en la posición 6).

Filtros. Puede hacerse la búsqueda dentro de una página para localizar la información, o la misma página puede contar con una biblioteca digital, glosario, etc. lo cual restringe la búsqueda.

Especificación. Es necesario agregar información semántica a cada elemento de las páginas html. Esto lleva a etiquetar los términos con categorías semánticas (p.e. `gato` → `animal`, etc.)

Se han propuesto diferentes caminos para realizarlo, pero en todos los casos es necesario tratar con el lenguaje natural.



Búsqueda de información: Ambigüedad.

Internet: problema y solución

Informatización

- Información para la toma de decisiones.
- Hábito a la información.
- Crecimiento exponencial de Internet, y captura de datos.

Acceso a información; dominio cultura

1. Hacer expedito el análisis sobre información cultural.
2. Observatorio de las industrias, políticas, y consumos culturales.
3. Navegación con base en una red semántica (\Rightarrow extracción de regularidades a partir de textos sobre cultura).

Contenido

1. Motivación
2. *Ontologías*
3. Algunas herramientas
 - (a) Terminología
 - (b) Thesauri e identificación de relaciones

Ontologías en el campo computacional

*Una **ontología** es un sistema compuesto de la terminología correspondiente a la conceptualización en un dominio, organizada en clases que reúnen instancias, además existen relaciones y funciones sobre clases e instancias. La sistematización de dicha información permite abstraer reglas para incorporar nuevos elementos.*

Una ontología puede verse como un **sistema formal** (constantes, variables, funciones, relaciones, axiomas, y reglas de inferencia), por ejemplo los derivados de la lógica clásica. Aunque también las hay informales donde normalmente existe una definición de cada concepto.

En términos generales, el diseño de una ontología deberá considerar aspectos semánticos del lenguaje.

(<http://babage.dia.fi.upm.es/ontoweb/wp1/OntoRoadMap/index.htm>)

Tipos de ontologías

La variedad de ontologías existentes ha conducido a clasificarlas según su orientación.

Orientación	Características	Ejemplo
Representación del Conocimiento	Mantiene convenciones, traslada conceptos a otros lenguajes	<i>Frame-Ontology</i>
Representación de Alto Nivel	Universal, identificación, no redundancia de conceptos	<i>SUO</i>
Procesamiento del Lenguaje Natural	Synsets (conjunto de sinónimos)	<i>WordNet</i>
General	Independiente del Dominio	<i>CyC</i>
Aplicaciones en un dominio particular	Modelación de problemas	<i>UMLS</i>

Volumen del Conocimiento

El tamaño de una ontología se mide con el número de conceptos, aunque en algunos casos, se pone importancia en otros parámetros; p.e. número de axiomas.

Nombre	# C	# R	# A	# I
<i>Frame-Ontology</i>	10-50	10-50	100-500	-
<i>Standard Upper O</i>	100-500	100-500	-	1K-5K
<i>WordNet</i>	100K	-	0	200K
<i>CyC</i>	>5K	-	-	-
<i>UMLS</i>	5K	50-100	-	-
<i>CHEMICALS</i>	10-50	10-50	10-50	100-500
<i>Onto-Qpr</i>	-	100-200	0	50-100

Número de: conceptos (#C); relaciones (#R); axiomas (#A); e instancias (#I).

Contenido

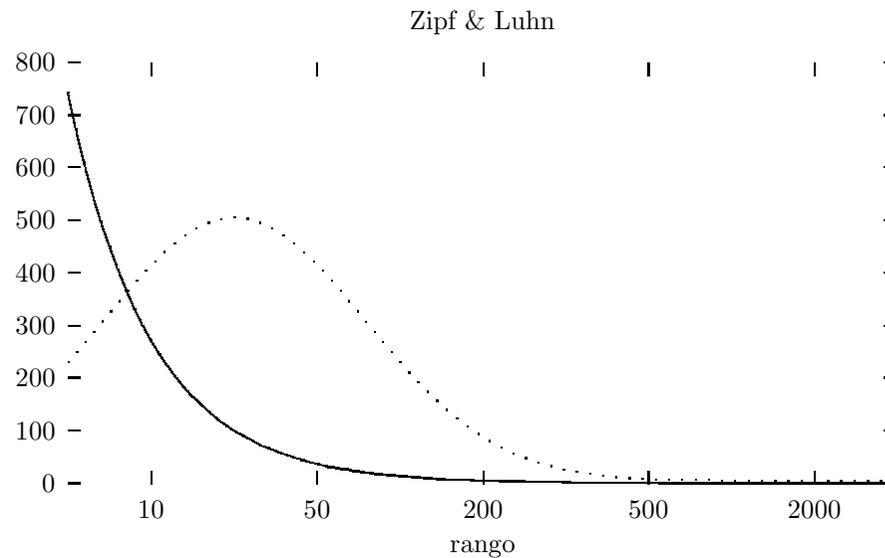
1. Motivación
2. Ontologías
3. Algunas herramientas
 - (a) *Terminología*
 - (b) Thesauri e identificación de relaciones

Terminología

- La terminología es una rama de la lingüística que se ocupa de identificar los términos propios de un determinado sistema de una lengua.
- Además de su aplicación a la construcción de diccionarios tiene impacto en la creación de bibliotecas digitales, traducción automática, y otros sistemas que requieren apoyarse en textos o el habla.
- Con el empleo de los sistemas de cómputo se espera desarrollar sistemas que sean capaces de construir catálogos para todos los dominios conocidos (proyecto *Meaning*).

Términos importantes

Hipótesis de Luhn: Propone acotar los términos de alto contenido semántico. Dichos términos tienen frecuencia media.



Una medición de la importancia

El punto de transición (PT)

Frecuencia que divide al vocabulario de un texto en dos: palabras de alta y baja frecuencia.

Una forma sencilla de calcularlo es:

identificar la frecuencia más baja que no se repita.

Se propone el PT como el centro de la zona de palabras de alto contenido informativo. Se ha reportado que en el 25% de las palabras con frecuencia alrededor del PT se encuentran las *palabras clave* del texto (subconjunto de la terminología).

Aplicación del PT

Intuitivamente, las palabras de mayor frecuencia son las que más representan al texto, y en muchos casos se han empleado así. Sin embargo, esto es aparente. En el siguiente ejemplo se ilustra un caso diferente.

Dado un texto (`cine4.txt`) su vocabulario ordenado por frecuencias resultó:

- 6 romeo
- 5 sangre
- 4 fantasma
- 3 cine
- 2 susana
- 2 (16 palabras adicionales)
- 1 (192 palabras)

En este caso $PT=3$, que corresponde a la palabra `cine`.

Contenido

1. Motivación
2. Ontologías
3. Algunas herramientas
 - (a) Terminología
 - (b) *Thesauri e identificación de relaciones*

Construcción de un *thesaurus*

Un diccionario que contiene por cada entrada una palabra y una serie de palabras o expresiones relacionadas con ella. Cada entrada puede verse como:

palabra REL [...] EXPV [...] EXPC [...] PALF [...].

donde REL significa “palabras relacionadas”; EXPV, “expresiones verbales”; EXPC, “expresiones comunes”; y PALF, “palabras familiares” .

Método (basado en Grefenstette (1996)):

1. Representar cada término por las palabras vecinas.
2. Calcular la similitud entre las representaciones.
3. Elegir los pares con mayor similitud.

Relaciones léxico semánticas

Se emplean tres rasgos:

- Distancia inversa (DIP) en los contextos de coocurrencia de los términos.
- Patrones léxico sintácticos (PLS) que indican una forma de relación.
- Redes de coocurrencia léxica (RCL): estructuración de la importancia de coocurrencia de los términos en supuesta relación.

Tomando las parejas del *thesaurus*:

1. Determinar umbrales con base en una muestra de ejemplos que cumplen una relación (hiponimia, antonimia, etc.).
2. Extraer del resto de los pares aquellos cuya puntuación rebase los umbrales.

Algunos patrones LS

No.	Expresión	Peso
1	Ant1 word*, pero word* Ant2	5
2	desde word* Ant1 hasta word* Ant2	4
3	Ant1 word* [, ;] sino word* Ant2	5
4	Ant1 word{0,4} [y o] word{0,4} Ant2	1

Ejemplos de antonimia

Par	DIP	PLS	RCL	Total	Clase
Moneda-Dinero	0.36	0.02	0.06	0.44	
Absoluto-Relativo	0.76	0.79	1	2.51	Antónimo
Precio-Costo	0.58	0.19	0	0.77	
Compra-Venta	0.73	0.17	0.06	0.96	Antónimo
Producción-Consumo	0.64	0.47	0.26	1.37	Antónimo
Obrero-Trabajador	0	0	0.3	0.3	
Consumidor-Productor	0.75	0.56	0.09	1.4	Antónimo
Máximo-Mínimo	0.68	0.27	0.2	1.15	Antónimo
Corrección-Ajuste	0.84	0.2	0.43	1.47	Antónimo
Positivo-Negativo	0.63	0.81	0.13	1.57	Antónimo
Escasez-Abundancia	1	1	0.59	2.59	Antónimo
Mercancía-Producto	0.29	0.19	0.21	0.69	
Público-Privado	0.65	0.53	0.16	1.34	Antónimo

Desempeño aproximado: 80%.

Resumen

- La atención al problema de acceso a la información requiere herramientas *ad hoc*.
- Los sistemas basados en conocimiento tienen su base principal en el tratamiento automático del lenguaje natural.
- El gran almacén mundial (www) y las herramientas computacionales pueden ayudar a resolver el problema de la informatización.