Universidad
Carlos III de Madrid

Módulo X
Realimentación y expansión de consultas

OpenCourseWare

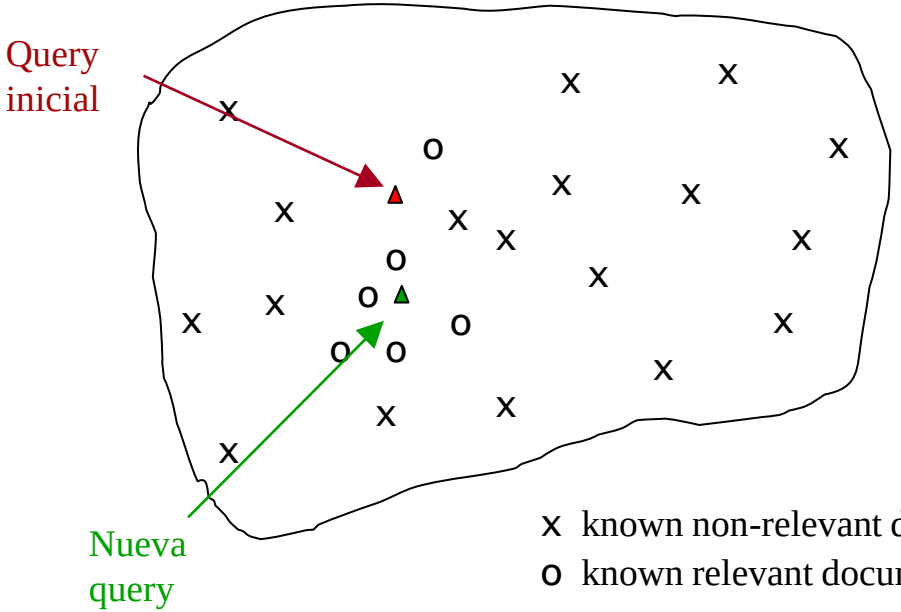
Recuperación y Acceso a la Información

Contenidos

- Mejora de las consultas
  - Relevance Feedback
  - Query Reformulation
  - Query Expansion
    - Wordnet

Relevance Feedback

- Trata de utilizar el “more like this” para mejorar la recuperación
- Etapas
  - El usuario formula una consulta inicial
  - El sistema **devuelve los documentos resultantes**
  - **El usuario le indica al sistema cuáles son relevantes y no relevantes**
  - **El sistema re-calcula los documentos resultantes en base a esta información**



Centroide:

$$\vec{\mu}(D) = \frac{1}{|D|} \sum_{d \in D} \vec{v}(d)$$

Relevance Feedback (II)

- **A) Consulta:** New space satellite applications
- **B) Resultados (en azul=relevantes, en negro=no se sabe)**
  1. 0.539, NASA Hasn't Scrapped Imaging Spectrometer
  2. 0.533, NASA Scratches Environment Gear From Satellite Plan
  3. 0.528, Science Panel Backs NASA Satellite Plan, But Urges Launches of Smaller Probes
  4. 0.526, A NASA Satellite Project Accomplishes Incredible Feat: Staying Within Budget
  5. 0.525, Scientist Who Exposed Global Warming Proposes Satellites for Climate Research
  6. 0.524, Report Provides Support for the Critics Of Using Big Satellites to Study Climate
  7. 0.516, Arianespace Receives Satellite Launch Pact From Telesat Canada
  8. 0.509, Telecommunications Tale of Two Companies

Relevance Feedback (III)

- **C) Nueva consulta**

2.074 new	**15.106 space**
30.816 satellite	**5.660 application**
5.991 nasa	5.196 eos
4.196 launch	3.972 aster
3.516 instrument	3.446 arianespace
3.004 bundespost	2.806 ss
2.790 rocket	2.053 scientist
2.003 broadcast	1.172 earth
0.836 oil	0.646 measure

Relevance Feedback (IV)

- **D) Nuevos resultados**
  2. 0.513, NASA Scratches Environment Gear From Satellite Plan
  1. 0.500, NASA Hasn't Scrapped Imaging Spectrometer
  X. 0.493, When the Pentagon Launches a Secret Satellite, Space Sleuths Do Some Spy Work of Their Own
  X. 0.493, NASA Uses 'Warm' Superconductors For Fast Circuit
  8. 0.492, Telecommunications Tale of Two Companies
  X. 0.491, Soviets May Adapt Parts of SS-20 Missile For Commercial Use
  X. 0.490, Gaping Gap: Pentagon Lags in Race To Match the Soviets In Rocket Launchers
  X. 0.490, Rescue of Satellite By Space Agency To Cost $90 Million

Relevance Feedback. Consulta óptima (I)

- Puede aplicarse relevance feedback tanto en modelos vectoriales como probabilísticos

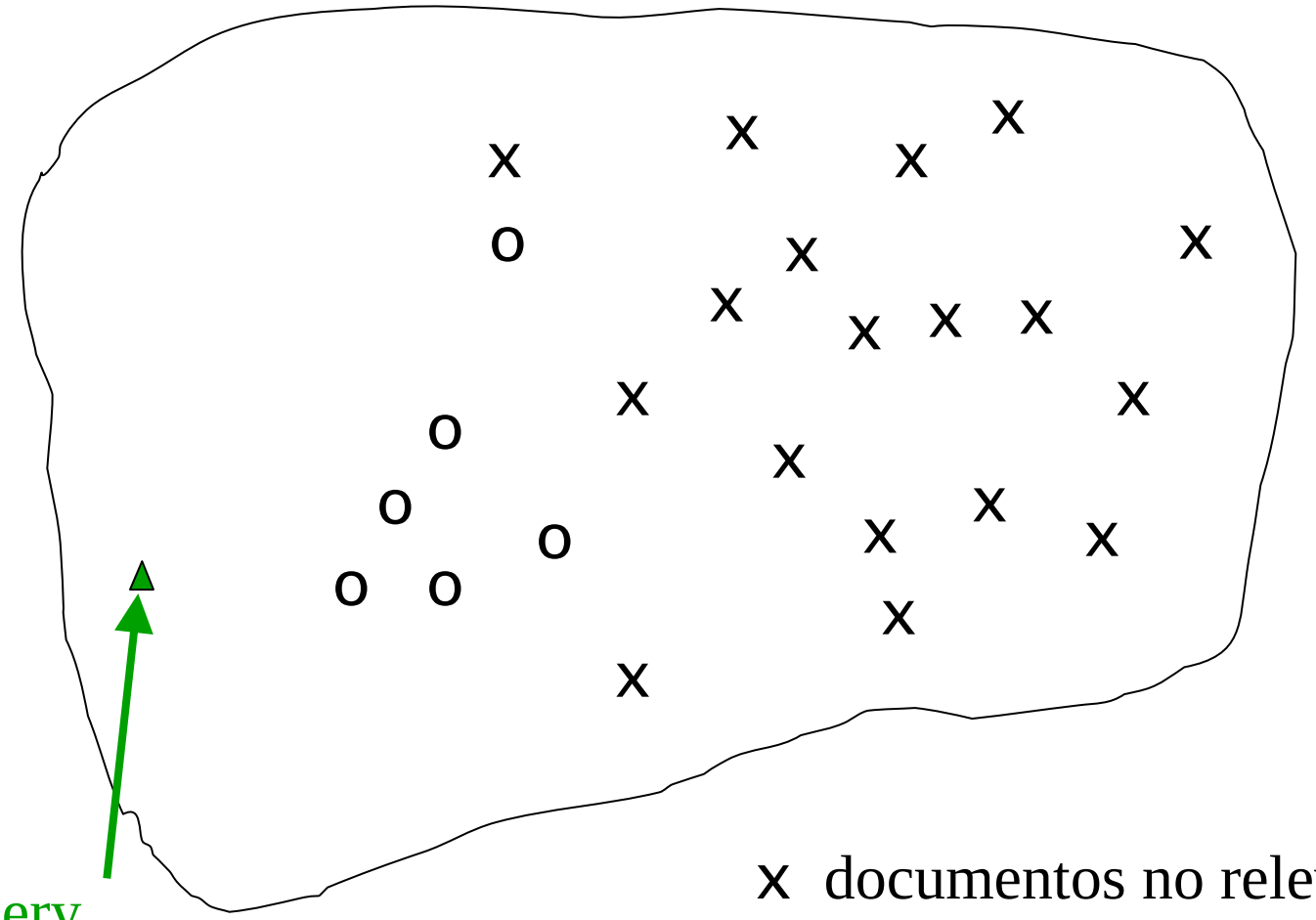
- Implementación en el modelo vectorial (las consultas son vectores)
  - Consulta óptima

$$\vec{q}_{opt} = \max_{\vec{q}}[sim(\vec{q}, C_r) - sim(\vec{q}, C_{nr})]$$

Donde C_r es el conjunto de documentos relevantes en la colección y C_{nr} es el conjunto de no relevantes ($C_{nr} = C - C_r$).

- Pero ¿conocemos todos los documentos relevantes de la colección para cada consulta?

Relevance Feedback.Consulta óptima (II)



Relevance Feedback.Rocchio Algorithm

- Aproximaciones de la consulta óptima en base a la información que conocemos (consulta original, documentos relevantes y no relevantes en el conjunto de resultados)

- The Rocchio Algorithm:

$$\vec{q}_m = \alpha \vec{q}_0 + \beta \frac{1}{|D_r|} \sum_{\vec{d}_j \in D_r} \vec{d}_j - \gamma \frac{1}{|D_{nr}|} \sum_{\vec{d}_j \in D_{nr}} \vec{d}_j$$

  - q_0 es el vector original
  - D_r es el conjunto de documentos relevantes conocidos (distinto de C_r)
  - D_{nr} es el conjunto de documentos no relevantes conocidos (distinto de C_{nr})
  - α, β, γ son constantes para dar más o menos peso a cada conjunto de términos

Relevance Feedback. Ejemplo

- Ejemplo:
  1. Consulta: coche rojo
  2. Documentos resultantes inicialmente:
     - Coche rojo marca citroen
     - Coche rojo madrid ocasión venta
     - Ocasión madrid caniches blancos
  3. El usuario marca como relevantes los dos primeros, y el tercero como no relevante.
  4. Asumiendo que el sistema utiliza sólo TF como pesos y que las constantes $\alpha=\beta=\gamma=1$ **¿cómo sería la nueva consulta generada por el sistema?**
- ¿Qué ocurre si resultan pesos negativos? Ponerlos a 0
  - La ausencia de evidencia no es evidencia de la ausencia

Relevance Feedback. Pesos

- Más peso la consulta original, menos los documentos relevantes, menos los documentos no relevantes.
  - Ejemplo de pesos: α= 1 β= 0.75 γ= 0.15
- Mejores resultados con positive feedback:
  - Puede considerarse γ=0, lo que supondría considerar sólo los documentos relevantes
  - Los no relevantes están muy dispersos

- Ide_Regular: $$\vec{q}_m = \alpha \vec{q} + \beta \sum_{\forall \vec{d}_j \in D_r} \vec{d}_j - \gamma \sum_{\forall \vec{d}_j \in D_n} \vec{d}_j$$

- Ide_Dec_Hi: $$\vec{q}_m = \alpha \vec{q}_0 + \beta \frac{1}{|\mathcal{D}_r|} \sum_{\vec{d}_j \in \mathcal{D}_r} \vec{d}_j - \gamma \cdot \operatorname{argmax}_{\vec{d}_j \in \mathcal{D}_{nr}} sim(\vec{q}_0, \vec{d}_j)$$

Con modelo probabilístico

- Se empieza con el vectorial, luego se ven los documentos y se recalcula.

$$\sum_{k_i[q,d_j]} \log\left(\frac{r_i + 0.5}{R - r_i + 0.5} \times \frac{N - n_i - R + r_i + 0.5}{n_i - r_i + 0.5}\right)$$

	relevante	No relevante	total
Docs evaluados con k	r_i	n_i-r_i	N_i
Docs evaluados sin k	R-r_i	N-n_i-(R-r_i)	N-n_i
Todos	R	N-R	N

Relevance Feedback. Ventajas

- Reduce problemas de polisemia (tiende a incluir términos relacionados)
- Mejora recall y precision
  - Exhaustividad (sobre todo): nuevos términos relacionados se añaden a la consulta. Ej. coche -> motor, accesorios coche, automóvil
  - Precision: damos más peso a términos más específicos y restamos peso o incluso eliminamos términos más ambiguos o que dan más ruido

Relevance Feedback. Inconvenientes

- Requiere intervención del usuario.
  - Estudios muestran que muy pocos usuarios explotan este tipo de opciones en un buscador (aparte del tiempo el usuario puede no querer dar información). Pocos buscadores web lo implementan
  - Funciona mejor en colecciones donde los documentos forman clusters en base al vocabulario
  - Posibles problemas son mayor tiempo computación, peor comprensión de los resultados, más tiempo inicial para el usuario
  - Asume que todos los documentos relevantes contienen los mismos términos y que el vocabulario del que busca y los documentos es el mismo
  - Depende del conjunto inicial de documentos
- Aumenta el tiempo de procesamiento (más términos que procesar)

Relevance Feedback.Automatización

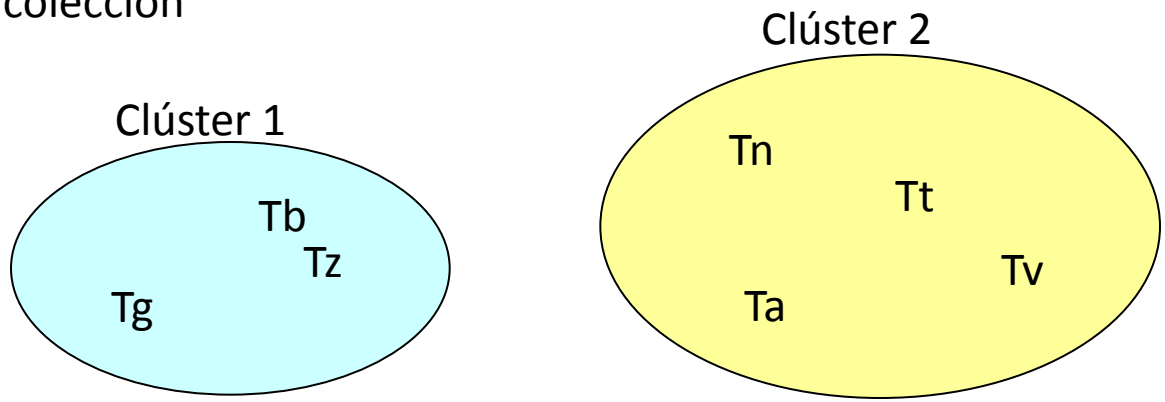
- Se puede usar información histórica del tipo “preguntas similares formuladas”, clicks de los usuarios sobre ciertos documentos ante una consulta (se puede tener en cuenta los términos de los snippets, que es en lo que se basa el usuario para pinchar), etc.
  - Google lo utiliza en páginas similares (se basa en los enlaces)

- Blind Relevance Feedback o Pseudo Relevance Feedback
  - Se **asume** que los primeros k recuperados son relevantes.
  - Se ha probado efectividad en TREC (P@50): no-RF 0.625, RF 0.727
- En ciertos casos puede ser perjudicial
  - Búsqueda por minas de cobre y primeros resultados tratan sobre minas de cobre en Chile. La query reformulada tenderá hacia Chile únicamente
  - Búsqueda por apple (la fruta, no la compañía)

- Clustering (Durante años en http://www.excite.com/ permite seleccionar cluster en “are you looking for?”) **¿cómo?**

Relevance Feedback. Clustering

- Expandir las consultas, pero en lugar de requerir la intervención del usuario se realiza **automáticamente** mediante técnicas de clustering
- Expanden los términos de las **consultas** incluyendo en ellas los términos del **cluster asociado**.
- Estas técnicas se aplican únicamente sobre el conjunto de documentos recuperados (**análisis local**) o sobre la colección completa (**análisis global**)
  - Análisis local: El clustering debe realizarse durante la consulta
  - Análisis global: frecuentemente se basa en el uso de tesauros generados sobre la propia colección



Relevance Feedback. Clustering (II)

- **Índice de equivalencia**

$$E_{ij} = \frac{C_{ij}^2}{C_i C_j}$$

Donde:

C_j es el número de veces que aparece el término j

C_i es el número de veces que aparece el término i

C_{ij} es el número de veces que aparecen conjuntamente i y j

- Chen propuso combinar este mecanismo con la Rochio, considerando no relevante los términos genéricos del dominio

Contenidos

- Mejora de las consultas
  - Relevance Feedback
  - Query Reformulation
  - Query Expansion
    - Wordnet

Query reformulation

Mejorar las consultas del usuario según sugerencias del motor

- Corrección ortográfica (Ej. “Did you mean” de Google. Se basa también en cantidad de resultados)
- Búsquedas más frecuentes (Ej. Google Suggest)
- Información acerca de las transformaciones que el motor realiza (filtrado, normalización, etc)
  - Usuarios especializados
- Consulta de los términos por los que puede buscar
  - Términos indizados
  - Vocabularios controlados (tesauros, ontologías, glosarios, etc.)
    - Mejor si se han utilizado para indizar Ej. MeSH

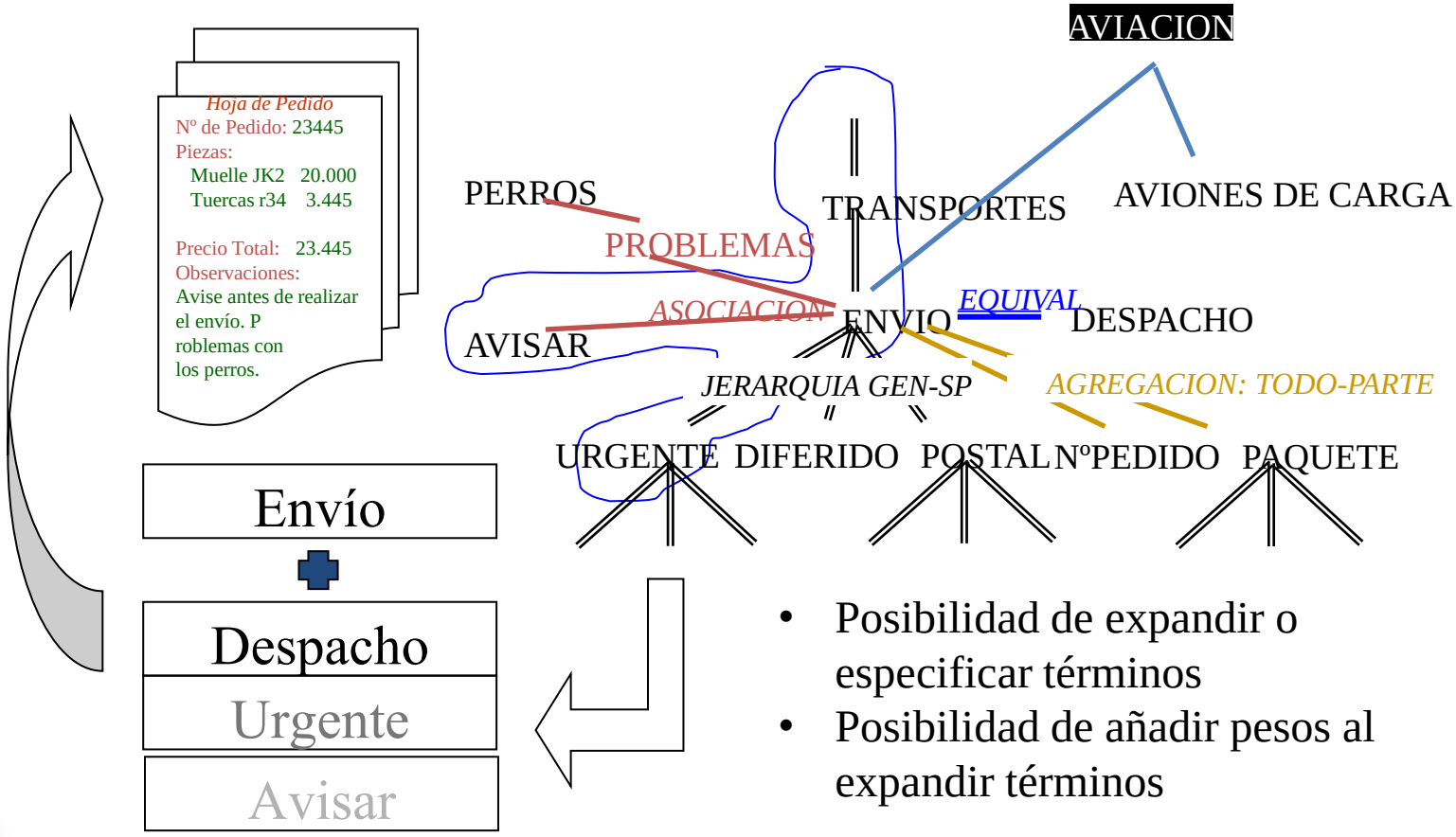
Contenidos

- Mejora de las consultas
  - Relevance Feedback
  - Query Reformulation
  - Query Expansion
    - Wordnet

Query expansion

Expansión (no reformulación) de la consulta por los términos relacionados en base a vocabularios controlados

- A diferencia del RF se realiza sobre la consulta no sobre resultados



- Posibilidad de expandir o especificar términos
- Posibilidad de añadir pesos al expandir términos

Query expansion (II)

- Mejoran exhaustividad, pero suelen empeorar la precisión
  - Normalmente menos peso para los términos expandidos
  - Generan ruido en palabras polisémicas
  - Sólo podemos expandir automáticamente de genérico a específico
    - Documentos que hablen sobre animales, pueden ser documentos que hablen sobre gatos
    - No ayudan a hacer más específica la consulta, sino más genérica (p.e. sinónimos)
- Pueden ser tesauros generados automáticamente sobre la colección, basándose en técnicas de clustering o en técnicas de PLN(p.e Yippy)
  - Clustering por co-ocurrencia de términos no suele mejorar mucho
- Los tesauros pueden ser temáticos (uso de los términos en una o varias áreas de conocimiento) o léxicos (semántica en general)
  - Uno de los más utilizados: **WordNet**

Contenidos

- Mejora de las consultas
  - Relevance Feedback
  - Query Reformulation
  - Query Expansion
    - Wordnet

WordNet

- Base de datos léxica para el inglés, desarrollada en la Universidad de Princeton bajo la dirección de George A. Miller
  - http://wordnet.princeton.edu/
- Sustantivos, verbos, adjetivos y adverbios: ~150.000 palabras
- Elimina las flexiones de las palabras pero no las derivaciones
  - Ej. “went” lo convierte a “go”
  - Pero “interpret”, “interpreter”, “interpretation”, “interpretative dancing” son todas palabras diferentes para WordNet
- Para cada categoría se hacen grupos de sinónimos, que se denominan synsets. Cada synset expresa un concepto semántico. Hay unos 115.000
  - Aproximadamente el 17% de las palabras son polisémicas (una misma palabra en más de un synset) y alrededor de un 40% tiene uno o más sinónimos (palabras diferentes en un mismo synset)

WordNet (II)

- Los synsets se relacionan entre sí mediante relaciones de diferentes tipos
  - Sinonimia
    - Big -> Large
  - Antonimia
    - Big -> Small
  - Hiponimia
    - Car -> Taxi
  - Hiperonimia
    - Car -> Motor Vehicle
  - Meronimia
    - Computer -> CPU
  - Holonimia
    - Leg -> Table

Semantic Relation	Syntactic Category	Examples
Synonymy (similar)	N, V, Aj, Av	pipe, tube rise, ascend sad, unhappy rapidly, speedily
Antonymy (opposite)	Aj, Av, (N, V)	wet, dry powerful, powerless friendly, unfriendly rapidly, slowly
Hyponymy (subordinate)	N	sugar maple, maple maple, tree tree, plant
Meronymy (part)	N	brim, hat gin, martini ship, fleet
Troponomy (manner)	V	march, walk whisper, speak
Entailment	V	drive, ride divorce, marry

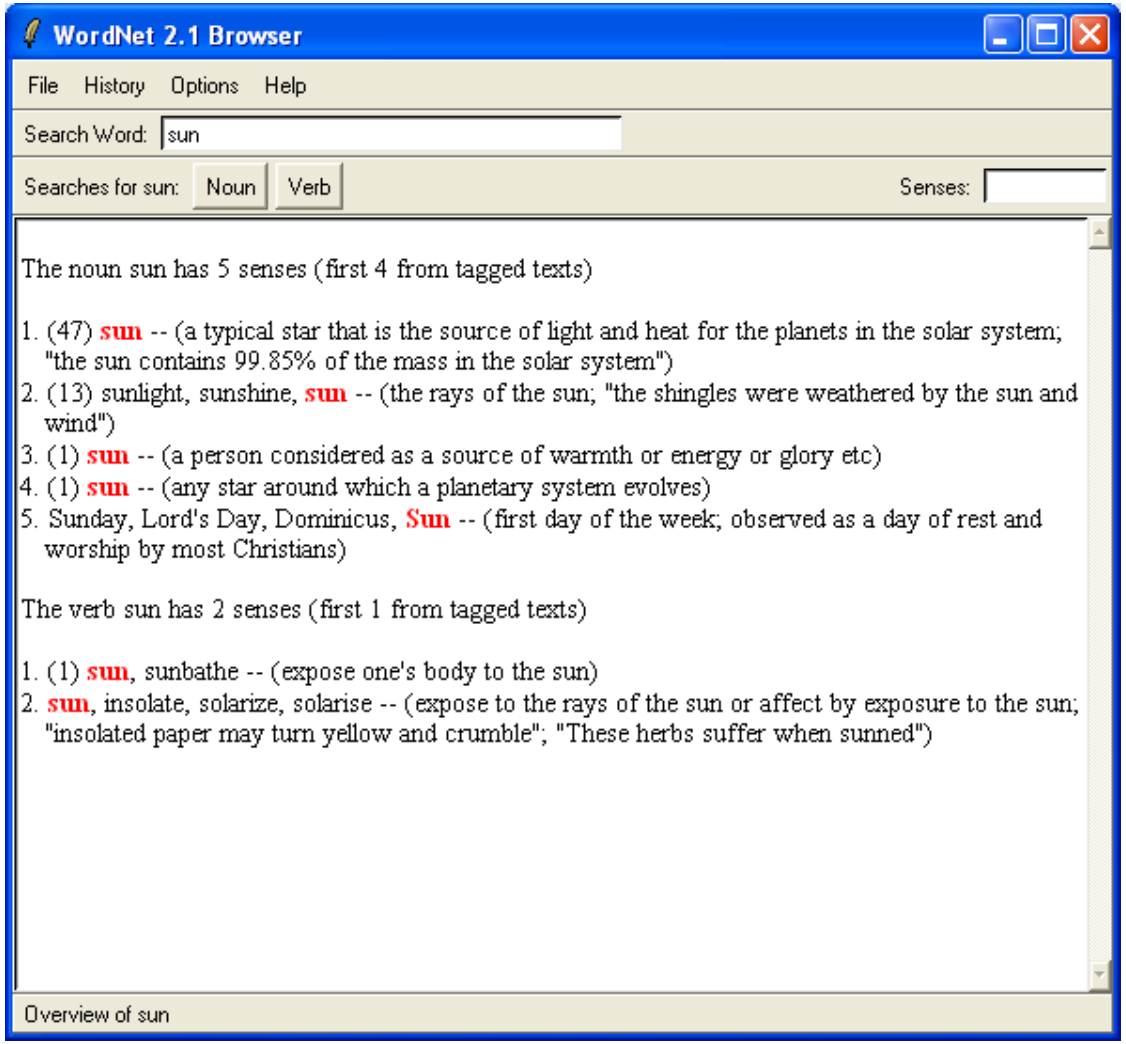
Note: N = Nouns Aj = Adjectives V = Verbs Av = Adverbs

George A. Miller. Wordnet: a lexical database for English. Communications of the ACM , 1995

WordNet (III)

- Tiene información léxica sobre cada palabra, pero es mucho más potente que un diccionario por las relaciones que establece entre ellos
- Indica para cada término su uso más frecuente basándose en la anotación de corpus
  - Esto es muy utilizado para desambiguar términos, así como la relación entre el contexto en el que se encuentra una palabra en un texto y las palabras con las que se relaciona en WordNet
- Numerosos proyectos relacionados. Entre ellos creación de bases de datos similares para otros idiomas y mapeo entre ellas
- Consulta:
  - Puede consultarse on-line: http://wordnetweb.princeton.edu/perl/webwn
  - Interfaz de consulta descargable: http://wordnet.princeton.edu/wordnet/download/
  - Es posible descargarse los ficheros con la información (son ficheros textuales)
  - **Existen numerosas API's para manipular estos ficheros: http://wordnet.princeton.edu/wordnet/related-projects/**

WordNet (IV)


WordNet 2.1 Browser
File History Options Help
Search Word: sun
Searches for sun: Noun Verb
Senses:
The noun sun has 5 senses (first 4 from tagged texts)
1. (47) sun -- (a typical star that is the source of light and heat for the planets in the solar system; "the sun contains 99.85% of the mass in the solar system")
2. (13) sunlight, sunshine, sun -- (the rays of the sun; "the shingles were weathered by the sun and wind")
3. (1) sun -- (a person considered as a source of warmth or energy or glory etc)
4. (1) sun -- (any star around which a planetary system evolves)
5. Sunday, Lord's Day, Dominicus, Sun -- (first day of the week; observed as a day of rest and worship by most Christians)
The verb sun has 2 senses (first 1 from tagged texts)
1. (1) sun, sunbathe -- (expose one's body to the sun)
2. sun, insolate, solarize, solarise -- (expose to the rays of the sun or affect by exposure to the sun; "insolated paper may turn yellow and crumble"; "These herbs suffer when sunned")
Overview of sun

Universidad
Carlos III de Madrid

Módulo X
Realimentación y expansión de consultas

Colaboradores

J.Morato, V.Palacios

M.Marrero, S.Sánchez-Cuadrado, J.Urbano