Primera práctica de laboratorio de Sistemas de Información para la Web

Motivación

Un buscador web es un sistema informático que permite a los usuarios buscar información en la Web. Para ello los usuarios introducen consultas textuales en las que tratan de plasmar su necesidad de información. El buscador devuelve una lista de resultados que puede ser una mezcla de páginas web, imágenes o ficheros. Además, algunos buscadores proporcionan información estructurada así como ayudas a la búsqueda como consultas relacionadas u otras preguntas realizadas por otros usuarios. Para realizar estas tareas los buscadores exploran de manera automática la web, así como bases de datos o repositorios de datos estructurados.

Jo Kristian Bergum
@jobergum

Search is such a fascinating problem, one day we'll be able to produce great results for queries like these.

From the Amazon shopping query dataset:

"I need something timeless and that can go with any outfit, not a thin fabric but not thick either because I get warm easily"

22 Aug 2022 • 13:36

Descripción del ejercicio

En este ejercicio se proponen **12 necesidades de información** propuestas por Gayo-Avello y Brenes (2009). Dichas tareas tienen **niveles de dificultad diferentes**: unas son, a priori, fáciles de resolver y otras son, a priori, difíciles. En el segundo caso es posible que el usuario debe realizar varias consultas, explorar distintas páginas web para aprender nueva información y, en consecuencia, reformular la consulta inicial y/o resolver la necesidad en varios pasos.

Las tareas se describen en inglés tal y como fueron publicadas originalmente ya que uno de los objetivos de este ejercicio es reflexionar sobre el modo en que los buscadores ofrecen distintas perspectivas (y funcionalidades) en base al idioma utilizado.

El objetivo del ejercicio es resolver las 12 tareas con las siguientes variaciones:

- Con consultas en **inglés** y usando **Google**.
- Con consultas en **castellano** y usando **Google**.
- Con consultas en **inglés** y usando **otro buscador** (p.ej. Bing o DuckDuckGo)
- Con consultas en **castellano** y usando **otro buscador** (el mismo que en la tarea anterior)

Listado de tareas

Reproducción original de las tareas tal y como fueron descritas por Gayo-Avello y Brenes (2009).

`greyhound` *(task 1)* Find the page displaying the route map for Greyhound buses. *Joachims et al. [48].* A priori trivial.	`cornell` *(task 2)* The founder of Cornell University used to live close to campus –near University and Stewart Avenue. Does anybody live in this house now? If so, who? *Joachims et al.[48]. Slightly rephrased.* A priori difficult.
`baeza` *(task 3)* Find the homepage of Ricardo Baeza, not the director of Yahoo! Research Barcelona but another fellow countryman of him. *By the authors.* A priori difficult.	`time-machine` *(task 4)* Which actor starred as the main character in the original Time Machine Movie? *Joachims et al. [48].* A priori difficult.
`ny-mountains` *(task 5)* Which are the tallest mountains in New York? *Joachims et al. [48]. Slightly rephrased.* A priori difficult.	`purple-cow` *(task 6)* How can you follow the international bestselling author of Purple Cow on Twitter? *By the authors.* A priori difficult.
`1000-acres` *(task 7)* Find the homepage of the 1000 Acres Dude Ranch. *Joachims et al. [48].* A priori trivial.	`antibiotic` *(task 8)* What is the name of the researcher who discovered the first modern antibiotic? *Joachims et al. [48].* A priori difficult.
`michael-jordan` *(task 9)*	`emeril` *(task 10)*

Find the homepage of Michael Jordan, the statistician.

Joachims et al. [48].

A priori difficult.

Find the homepage of Emeril, the chef who has a television cooking program.

Joachims et al. [48].

A priori trivial.

`strangelove` *(task 11)*

This picture of Homer Simpson is a reference to a famous scene from a classic movie.

- What is the title of that episode?
- And the title of the movie?
- What's the name of the character in the original movie?
  - And the name of the actor who played it?

By the authors.

A priori difficult.

`cmu` *(task 12)*

Find the homepage for graduate housing at Carnegie Mellon University.

Joachims et al. [48].

A priori trivial.

Entregable

Documento de Word estructurado en las partes siguientes:

a) Reflexión razonada sobre las siguientes cuestiones relativas al funcionamiento de buscadores web:

1. ¿Cuál crees que es la diferencia entre las tareas fáciles y difíciles a priori?
   ¿Por qué crees que algunas tareas son más fáciles para el buscador y otras difíciles?
   ¿Qué características tienen las tareas a priori difíciles (en 2009) que ahora resultan más fáciles?

2. ¿Qué diferencias supone realizar las consultas en inglés y castellano para unas y otras tareas? ¿Qué diferencias supone realizar las consultas en inglés y castellano a la hora de obtener información enriquecida del buscador (consultas relacionadas, infoboxes, etc.)?
   ¿Existe alguna tarea que es virtualmente irresoluble por culpa del idioma? ¿Cuál? ¿Por qué crees que sucede esto?

3. ¿Cuáles son las principales diferencias entre los buscadores comparados respecto a ambas clases de tareas?
   ¿Cuáles son las principales fuentes de datos estructurados que explota cada buscador?
   ¿Qué consecuencias crees que tiene esa dependencia de fuentes externas?

4. ¿Existe alguna tarea en la que no hayas usado consultas textuales sino otro tipo de información, p.ej. Imágenes?
   ¿Qué diferencias has percibido al resolver esa tarea sin usar texto?

b) Resolución de tareas

Para cada tarea y variación (inglés/Google, castellano/Google, …) se deberá especificar:

- El código de la tarea (p.ej. *Greyhound*, *cornell*, *baeza*, …)
- Las consultas que introduce el usuario (de manera literal)
- Las páginas que explora el usuario (indicando URL y título) así como si resultaron pertinentes para resolver la tarea o no. Para ello deberá indicarse brevemente por qué (o por qué no) la página fue útil.
- Cualquier "ayuda" extra facilitada por el navegador: ya sean infoboxes, respuestas concretas a las consultas, consultas similares, preguntas que han hecho otros usuarios, etc.
- La respuesta final que se ofrece para la tarea (esta puede ser una página web o bien un dato concreto)

c) Apéndice

Todo el material anotado para cada configuración y tarea deberá aportarse como apéndice en un documento que se entregará en el campus virtual.

Referencias bibliográficas

- Gayo-Avello, D., & Brenes, D. J. (2009). Making the road by searching-a search engine based on swarm information foraging. *arXiv preprint arXiv:0911.3979*.
- Joachims, T., Granka, L., Pan, B., Hembrooke, H., & Gay, G. (2017, August). Accurately interpreting clickthrough data as implicit feedback. In *ACM SIGIR Forum* (Vol. 51, No. 1, pp. 4-11). ACM.