Documents Search Using Semantics Criteria

Santiago Cotelo
Luis Chiruzzo
Alejandro Makowski
Dina Wonsever

Instituto de Computación - Facultad de Ingeniería, Universidad de la República - Montevideo, Uruguay

Lucene

NLTK
+
WordNet

Stanford University NLP Natural Language Processing

Apache Solr

General Characteristics

Current information retrieval (IR) systems frequently rely on a bag of words model. This means: the user specifies some words, and the engine looks for documents that contain one or more occurrences of those words. The set of documents that match the query are ordered according to different criteria such as number of occurrences of each word or popularity of each document.

This simple model implies that the user will not be able to accurately specify which concepts she is looking for. For example, a user might enter the query *"american civilian and afghan solider"*, but she might end up getting results about *"american soldier and afghan civilian"*.

We created a more expressive language for queries, based in a simplified version of first-order logic, that lets the user specify different syntactic and semantic constraints so as to describe more precisely the documents she is looking for. The relevance of a document is scored using the dependency analysis of the document and other linguistic information.

Google bing DuckDuckGo Kosmix Freebase hakia Powerset

Problem

We studied current web search engines to see how much they understand the user's query. The criteria we analyzed is the following: Is it possible to associate terms in a query? Is it possible to disambiguate the polysemous terms? Is it possible to perform a synonym expansion of the terms? Is it possible to specify operators such as negation of the terms? Is it possible to apply time constraints to the terms?

We concluded that no current web search engine includes these characteristics in a satisfactory way.

	Google	bing	DuckDuckGo	hakia
Association between objects and attributes	☆☆☆	★☆☆	☆☆☆	☆☆☆
Polysemous words	☆☆☆	☆☆☆	★☆☆	☆☆☆
Temporal expressions	☆☆☆	☆☆☆	★☆☆	★☆☆
Synonym expansion	★★☆	☆☆☆	☆☆☆	★☆☆
Negation	☆☆☆	☆☆☆	☆☆☆	☆☆☆

SemQL Query Language

USER QUERY	SemQL QUERY
People born between 1930 and 1944	**born(x) ^ between(x, value="1930", value ="1944")**
American civilian and Afghan soldier	**american(x) ^ civilian(x) ^ afghan(y) ^ soldier(y)**
Cats that are not white	**cat(x) ^ !white(x)**

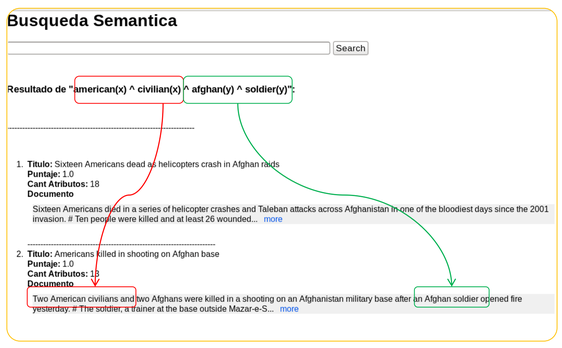
The language, which we call SemQL, provides a way to specify the following constraints:

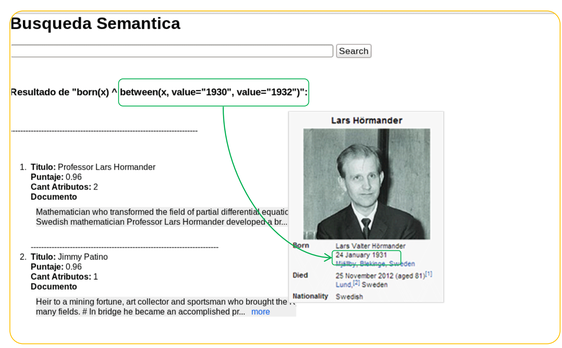
- Association of objects and attributes
- Temporal semantics
- Negation of attributes
- Synonyms expansion

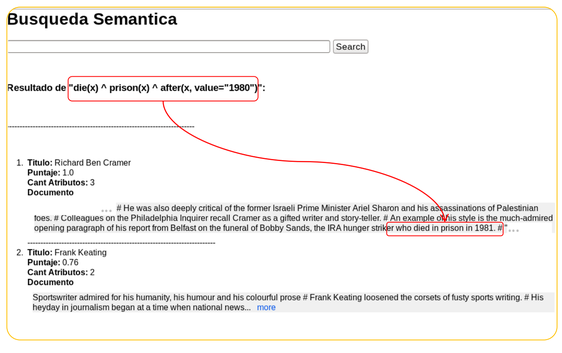
Document Retrieval and Sorting

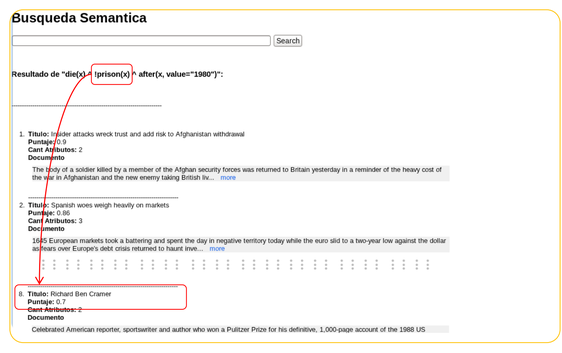
We created:

- A module that retrieves documents based on the SemQL terms used as keywords and their synonym expansion.
- An algorithm that sorts the retrieved documents depending on a scoring function that compares the sentences in the document to the original SemQL query.
- The relevance score is calculated taking into account the following factors: object hit, attribute hit, temporal expression hit, explicit negation hit, explicit negation miss, attribute miss, related term (synonyms)









UNIVERSIDAD DE LA REPUBLICA URUGUAY

FACULTAD DE INGENIERIA