Information Retrieval from the Web: an Interactive Paradigm

Massimiliano ALBANESE
Pasquale CAPASSO
Antonio PICARIELLO
Antonio Maria RINALDI

Dipartimento di Informatica e Sistemistica
Università di Napoli "Federico II"
Napoli, Italy

Introduction

- The goal of an information retrieval system is that of finding the most relevant information to a user query, possibly providing a compact answer
  - Users don't want to go through large result sets in order to find what they are actually looking for
    - Targeted answers to their queries should be computed, even if their interests are either poorly defined or inherently broad
- A classical approach in which the search engine returns a ranked list of documents containing the keywords in the query is not suitable anymore for today's information retrieval challenges

Contribution

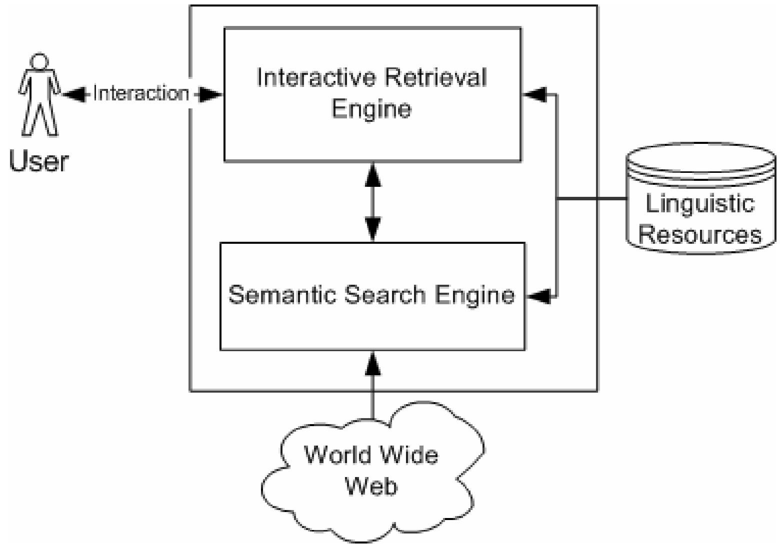
- An approach for designing a web retrieval system capable to find the desired information through several interactions with the users
- The proposed approach
  - allows to overcome the problems deriving from ambiguous or too vague queries
  - uses semantic search and topic detection techniques
- The results of the very experiments on a prototypal system are reported

The approach

- The answer of a search engine to a user query may be thought as the *engine's model* of the user's idea of what is considered relevant
- If the user finds out that most of the retrieved documents are not relevant (*the model is wrong*), she gives up and usually tries to rewrite the query
  - In order to prevent this to happen, the system should have the capability of *understanding* if the user query is too much general, thus automatically *trying to refine* it
    - User's feedback is used to adjust the *engine model*

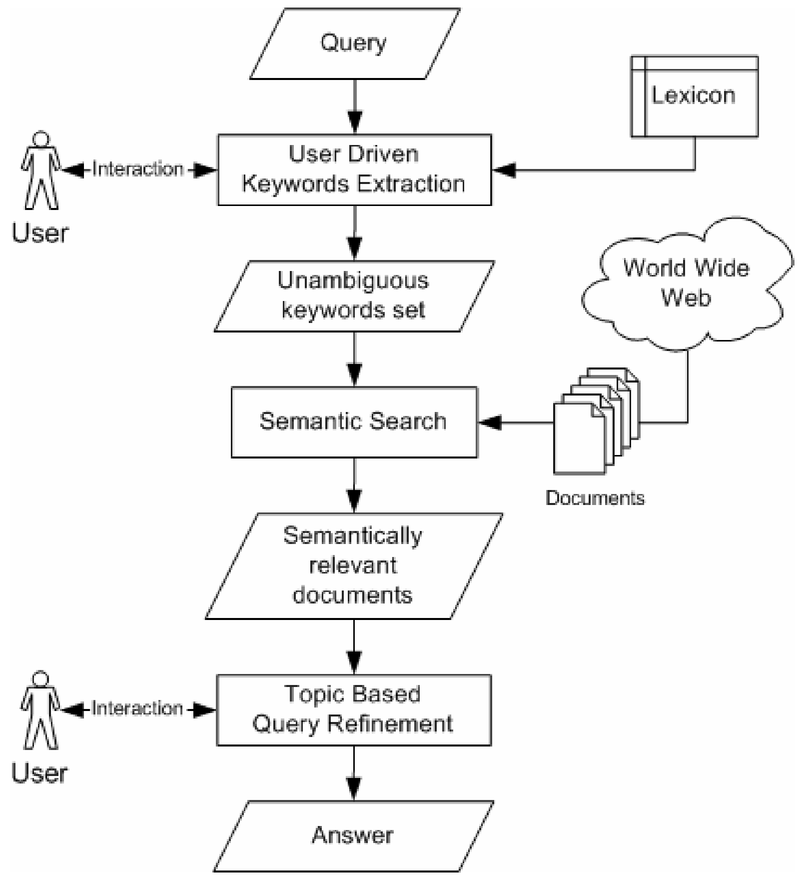
System Architecture

- Users submit queries to the *Interactive Retrieval Engine*
  - Both keyword and natural language queries are allowed
- The *Interactive Retrieval Engine*
  - accesses the web through the *Semantic Search Engine*
  - interacts with the users in order to clarify and refine queries
- The system relies on a set of linguistic resources
  - *WordNet* is used as both a dictionary and a semantic network



The retrieval process

- Keywords are derived from the query and disambiguated through user's feedback
- The *Semantic Search Engine* uses the unambiguous keyword set for retrieving document semantically relevant to the query
- Topics are identified and the query is further refined, based on further user feedback



User Driven Keyword Extraction

- *Part of Speech* (PoS) *tagging* is applied to user queries
- The PoS tagging is improved through
  - *Named Entities Recognition*
  - Heuristics to disambiguate ambiguous PoS assignment
- The results of PoS tagging allow to distinguish between keyword and natural language queries
  - In both cases a set of keywords is derived
    - Keyword whose meaning is not clear from the context need to be disambiguated through user feedback

Example of disambiguation

- What do you mean by “car”?

Glosses from WordNet

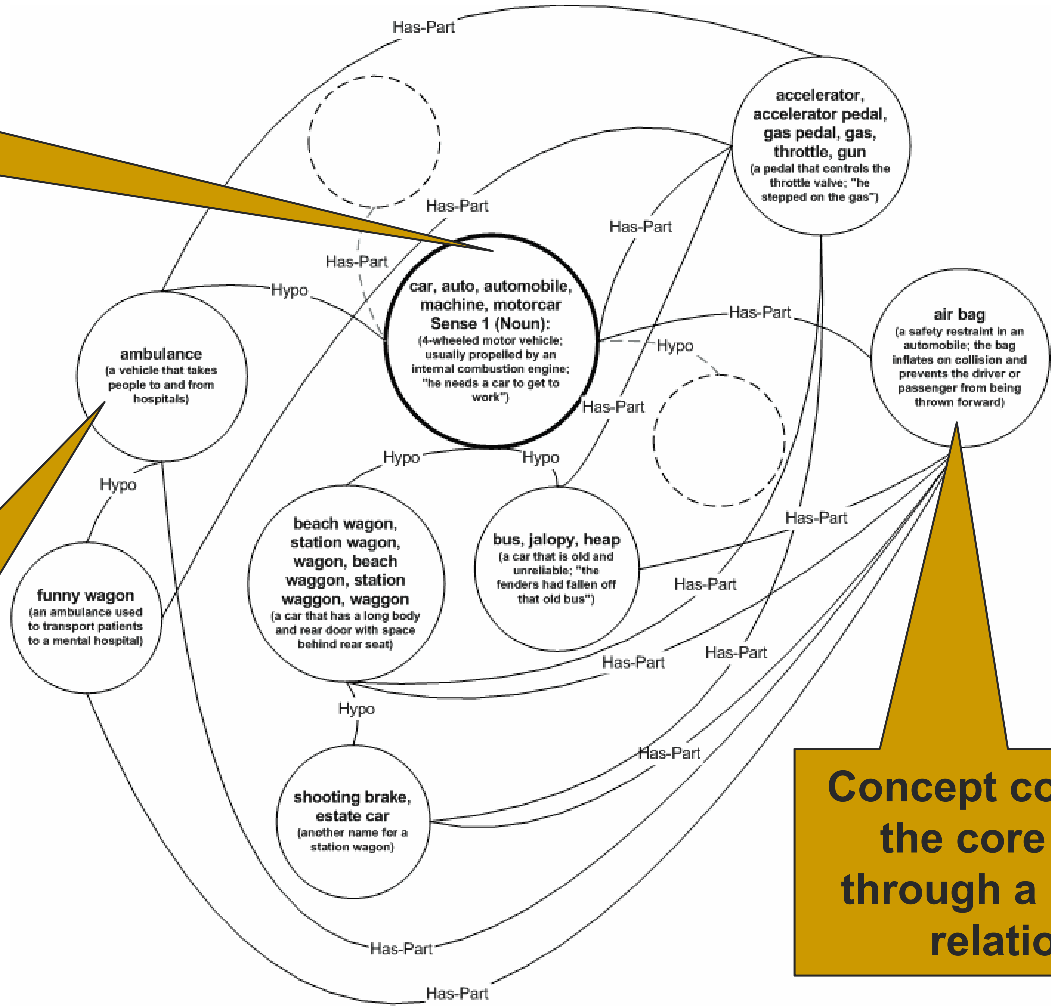
1. “4-wheeled motor vehicle; usually propelled by an internal combustion engine”
2. “a wheeled vehicle adapted to the rails of railroad”
3. “a conveyance for passengers or freight on a cable railway”
4. “car suspended from an airship and carrying personnel and cargo and power plant”
5. “where passengers ride up and down”

Semantic Search

- Semantic search capabilities are needed to overcome the limitations of traditional search engines, that are mainly keyword based
  - Ontologies are fundamental to achieve this goal
- The Semantic Search Engine presented in this work is based on *Dynamic Semantic Networks* (DSN)
  - A DSN is a semantic network dynamically built around one or more concepts that are central to a specific context
  - We build DSNs by extracting a subgraph from the complete graph of WordNet

Dynamic Semantic Network

Concept corresponding to sense #1 of the word “car”



Concept connected to the core concept through a “hyponymy” relationship

Concept connected to the core concept through a “has-part” relationship

Semantic Relatedness
Preliminary definition

- Given a DSN, we define the length l of the path between two terms/concepts as

  $$l = \min_j \sum_{i=1}^{h_j} \frac{1}{\sigma_j}$$

  where j spans over all the paths between the two considered terms, h_j is the number of hops in the j-th path and σ_j is the weight assigned to relations in the j-th path

Semantic Relatedness

Definition

- The *Semantic Relatedness* between two terms/concepts is defined as

$$W = e^{-\alpha l} \frac{e^{\beta d} - e^{-\beta d}}{e^{\beta d} + e^{\beta d}}$$

where

  - *l* is the length of the path between the terms
  - *d* is the depth of their subsumer
  - $\alpha \geq 0$ and $\beta > 0$ are two scaling parameters whose values have been defined by experiments

Semantic Search Engine

- The Semantic Search Engine
  - retrieves documents from the web using traditional search engines
    - keywords derived from the user query are used to this aim
  - evaluates the semantic relatedness of such documents w.r.t. the DSN built around the disambiguated keywords
    - documents showing a semantic relatedness greater than a given threshold are returned

Topic Based Query Refinement

- The idea of *Topic Based Query Refinement* is that of
  - Recognizing a set of topics from the set of documents returned by the semantic search engine
  - Asking the user for the topic she is interested in
  - Returning the subset of the semantic search results that match the topic
- To this aim we define a function (discriminating power) that allows to evaluate which topics are most suitable for selecting small document subsets

Discriminating Power

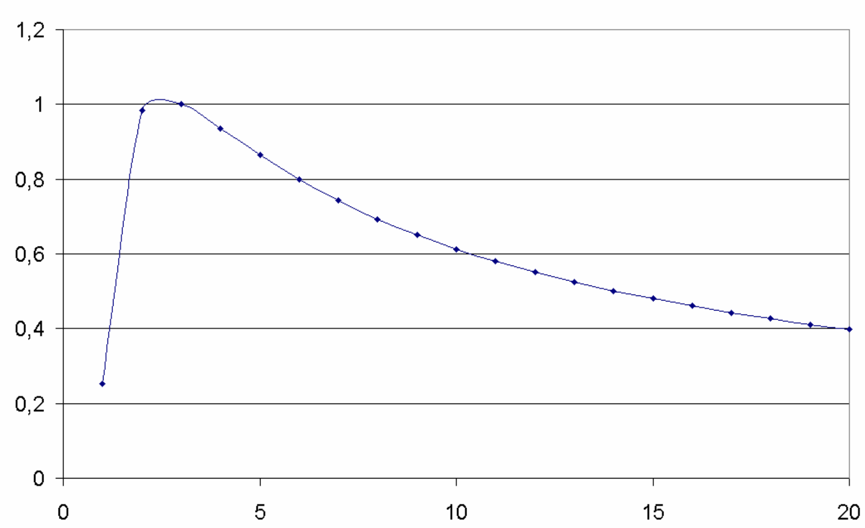
Preliminary considerations

- We empirically found out that the ability of an identified topic to select a small document subset is affected by
  - The fraction of documents matching the topic
  - The average frequency of occurrence of the topic into the matching documents
  - The length of the topic expressed as the number of words

Discriminating Power
Frequency and length

- The way that frequency and length affect the discriminating power can be taken into account through a function like the one sketched below

$$f(x) = \frac{\log x}{x^{\alpha}}$$



Discriminating Power

Definition

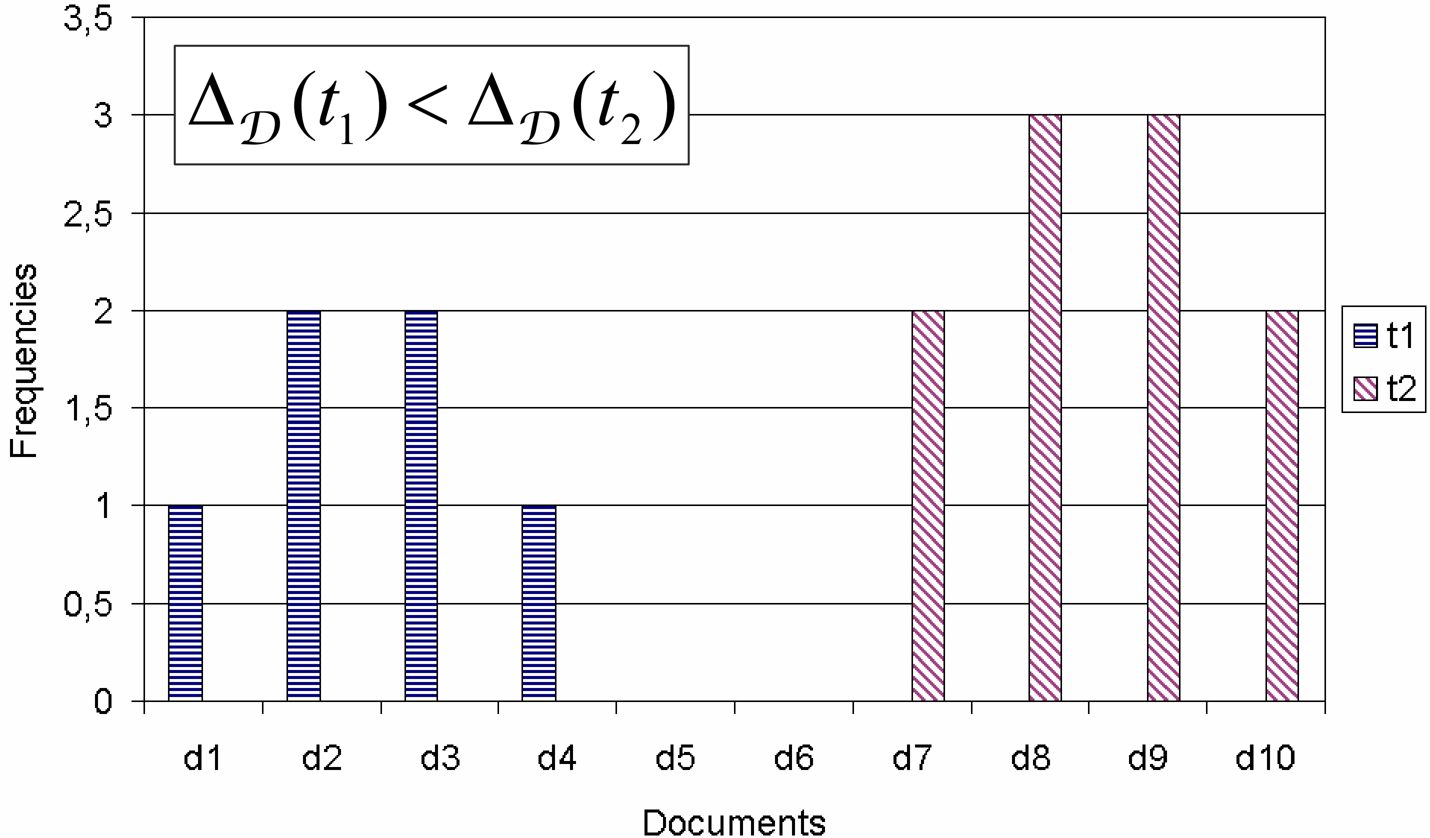
- Given a set $\mathcal{D}$ of documents and a topic t, we define the discriminating power Δ of t in $\mathcal{D}$ as

$$\Delta_{\mathcal{D}}(t) = -\log(p) \cdot \frac{\log(f + \Delta f)}{f^{\alpha}} \cdot \frac{\log(w + \Delta w)}{w^{\beta}}$$

  where
  - p is the fraction of documents matching t
  - f is the average frequency of t over the matching documents
  - w *is* the number of words in t
  - Δf and Δw are used to prevent Δ to be zero when $f = 1$ or $w = 1$
  - α and β are used to regulate the slop of the curve

Discriminating Power

Example



Experimental Results (1/2)

- We considered the worst case of very vague single-keyword queries
  - This case corresponds to the one in which the user doesn't clearly specify from the beginning what she's actually looking for
  - In such cases a traditional search engines would return millions of results

Query	# documents
car	154,000,000
museum	46,200,000
music	283,000,000
photography	45,500,000
soccer	24,900,000
train	39,600,000

Experimental Results (2/2)

- Topic identified for q_1="car" and q_2="museum"
  - Each identified topic allows to select a very small subset of documents
  - Precision is high: each returned document contain the desired information

Topic	Δ	P	f	w
used car values	0.623	2	4	3
car reviews	0.612	2	4	2
find new cars	0.599	1	2	3
car loan calculator	0.599	1	2	3
premium cars	0.589	1	2	2
midsize cars	0.599	1	2	3
msn autos	0.573	1	3	2
dollar rent a car	0.560	1	2	4

q_1 = "car"

Topic	Δ	P	f	w
national museum	0.736	2	7	2
bishop museum	0.695	1	4	2
nobel prize	0.625	1	5	2
asian art museum	0.617	1	3	3
design museum	0.607	1	3	2
american museum	0.571	2	4	2
san francisco museum	0.509	1	2	3
science museum	0.500	1	2	2

q_2 = "museum"

Conclusions

- We have presented an information retrieval system based on an interactive paradigm
  - we have extended a classic search engine with some semantic capabilities and query refinements techniques, trying to dynamically understand user's interests
- We have also described some preliminary experiments on a prototypal system
- Further investigation should be devoted first to conduct a more extensive experimentation and then to integrate management of other kinds of media into the system