Discussion on Web Search and Querying at SUM 2010

Amedeo Napoli
LORIA – CNRS (UMR 7503) – INRIA Nancy Grand Est – Nancy Université
BP 70239, 54506 Vandœuvre-lès-Nancy
Email: Amedeo.Napoli@loria.fr
http://www.loria.fr/~napoli/

SUM 2010 — Toulouse, September 28th
Before going to Toulouse, I would like to get some information about...

- A book on Claude Nougaro
- A biography of Claude Nougaro
- An autobiography of Claude Nougaro
- A book written by Claude Nougaro
- A book on the poetry of Claude Nougaro
- A songbook of Claude Nougaro
- ...

and I search the Web for finding some answers...
The Web would not have been so successful without the existence of search engines, but many problems remain:

- High recall and low precision.
- Low or no recall.
- Results are highly sensitive to vocabulary.
- Results are (usually) single Web pages rather than classified Web pages, i.e. not really “information retrieval” but rather “location finding”.
- The user must browse the result of a search for selecting the “correct” documents where to extract the requested information.
- Search engines are isolated applications.
How to improve things?

- With **explicit metadata** and **annotations**: for being **accessible** and **processable** by software agents, the **content** of a document has to be explicitly represented as a **structured description** with an associated **semantics**.

- An **intelligent manipulation** of documents is based on the understanding of the content of the documents w.r.t. domain knowledge.

- In the context of **Semantic Web, ontologies** are set to play a key role in establishing a **common terminology** between software and human agents, thus ensuring that different agents have a shared understanding of terms.
The semantic Web “cake”
In computer science, the term **ontology** refers to an engineering artifact which:

- constitutes a **model** of some part of the world,
- introduces a **specific vocabulary** and **specifies relative meaning**,  
- is formalized within a **knowledge representation language** (e.g. description logics),
- is usually intended to provide a **coherent view** of a domain and to assist **query answering**, **reasoning**, and **problem-solving**.
Guiding Document Manipulation with Domain Knowledge

- **Reasoning operations**: subsumption and satisfiability within classification-based reasoning and case-based reasoning.
- **Information extraction** and **information retrieval**.
- **Data mining** and **text mining** for analyzing and classifying documents with respect to their content, and for **ontology engineering**.
- **Methods for ontology engineering**:
  (i) constructing an ontology manually,
  (ii) reusing existing ontologies (alignment),
  (iii) using semi-automatic methods based on KDD processes.
Elements for Discussion: Web search needs KDD, Knowledge Representation, and Reasoning

The KDD process is iterative, interactive, and guided by an analyst.

Data
↓ selection and preparation of data
↓ cleaning and formatting the data
	Prepared data
↓ data mining operations
↓ numerical and symbolic methods
	Discovered patterns
↓ interpretation / evaluation
↓ representation of discovered patterns
	Knowledge units
↓ Knowledge systems (problem-solving, ontologies)
How can we define Web Search and Querying?
What do we expect from Web Search and Querying?

Web Search applies on a very large collection of documents, actually an open universe, needing to pay attention to:
- needs for a guided search
- taking into account scalability: algorithms, data organization,
- taking into account the open world
More Elements for Discussion

▶ **Guided Search**: ontologies and annotations for guiding and improving the search, representation of the document content,

▶ **Scalability**: efficient search algorithms (based on annotation or indexing), data classification through e.g. Formal Concept Analysis (FCA) for organizing data and the result of a search and then navigating this result (**Web Clustering Engines**).

▶ **Open World**: which kinds of problems are appearing and will have to be solved?

▶ **Correctness, completeness, and precision** of the answers.
Some elements for discussion w.r.t. four papers

- **Semantic Web Search** (d’Amato et al.): using standard Web Search engines with ontological background knowledge plus inductive reasoning for offline ontology compilation (guided search concerns).

- **Dealing with Plethoric Answers** (Bosc et al.): query-oriented cooperative systems for providing minimal sets of correct and useful answers by introducing predicates and fuzzy cardinalities (query answering and scalability concerns).

- **Aggregate queries and aggregate constraints** (Flesca et al.): computing range-consistent answers of aggregate queries with aggregate constraints (query answering and guided search concerns).

- **Ontology matching** (Wang et al.): structure-based similarity for improving ontology matching (ontology engineering concerns for guiding search).
Some other directions...

- **Web clustering engines** (Carpineto et al., ACM Computing Surveys, 2009)
- **Semantic search using graph-structured semantic models for supporting search process** (Tran - Haase - Studer, ICCS 2009)
- **Semantic Wiki search** (Haase et al. ESWC 2009)
- **Ontology searching and querying** with Swoogle, a “Metadata engine for the Semantic Web” (SHOE, SIRIO etc.).
- ...