Web Searching with Entity Mining at Query Time

P. Fafalios[1], I. Kitsos[1], Y. Marketakis[1], C. Baldassarre[2], M. Salampasis[3], Y. Tzitzikas[1]

1: Institute of Computer Science, FORTH-ICS, and
Computer Science Department, University of Crete, GREECE
2: Food and Agriculture Organization of the United Nations (FAO UN, Rome)
3: Vienna Univ. of Technology, Institute of Software Technology and Interactive Systems

FOUNDATION FOR RESEARCH AND TECHNOLOGY HELLAS

IRF Conference, Vienna, July 2012

The topic in one slide

Exploratory search *services that bridge the gap between the responses of "non semantic" search systems (e.g. keyword search) and semantic information (e.g. SPARQL endpoints)*

through a innovative dynamic entity-based integration approach that is performed at query time (key point: no human effort !)

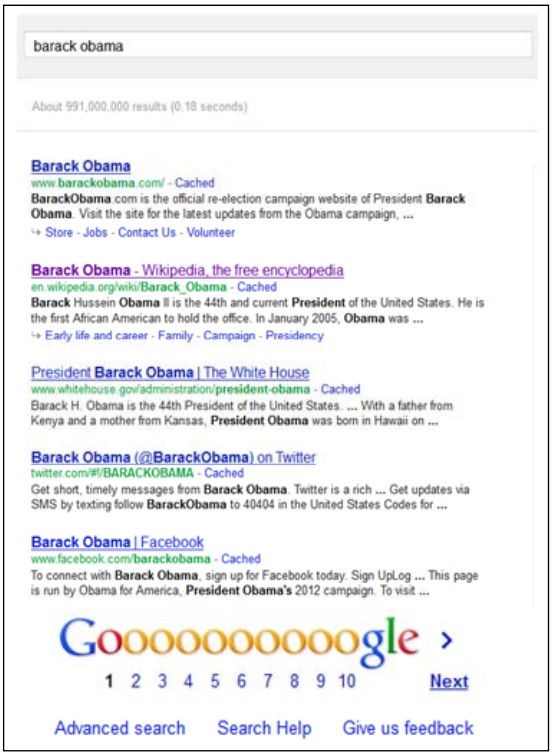
The provided services analyze (in various ways) and semantically enrich the returned results for satisfying recall-oriented information needs (snippet-based results clustering, entity name mining, semantic enrichment, gradual restriction of results etc.).

Outline

- Motivation
- Our approach
  - The Process
  - Challenges and how we tackled them
- Experimental Results
  - Entity Ranking
  - Contents Mining vs. Snippet Mining
  - Exploiting Linked Open Data
- Case Studies
  - Fisheries and Aquaculture Publications
  - Patent Search
- Conclusion and Future Work

Motivation (1/2): Common *Web Search Engines*

- Common Web Search Engines return a ranked list of pages
- Users
  - have to explore the answer linearly
  - tend to look only at the first page of results
  - rarely exploit the available metadata (or advanced search forms)
  - adopt a trial-and-error approach



Motivation (2/2): Emerging *Entity Search Engines*

- Existing **Entity Search Services**:
  - Provide the user with **entities** and **relationships** between these entities
  - Do not provide links to web pages
- Nevertheless:
  - Are still in their infancy (not used in many common and real world tasks)

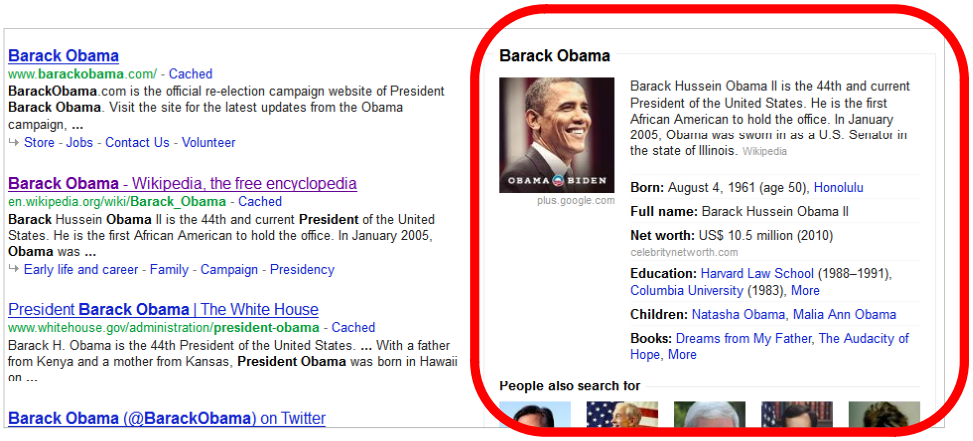
Our Approach

Do not change the way users search for information!

Enrich the classical (keyword based) interaction scheme of WSEs with Named Entity Mining (NEM) at query time.

Combine the pros of both families of systems

Similar (but much more simple) functionality recently announced (May'12) by Google**:**
"Google Knowledge Graph"



From an information integration point of view

- *Entity names* are used as the "**glue**" for automatically connecting documents with data (and knowledge).
- This approach does not require deciding or designing an integrated schema/view, nor mappings between concepts as in knowledge bases, or mappings in the form of queries as in the case of databases.
- Entities can be identified in documents, data, database cells, metadata attributes and knowledge bases.

Our Process

We focus on the following process

1. *Discover entities in the top hits of keyword search results*
2. *Group the entities according to their categories*
3. *Exploit the entities in faceted search-like interaction scheme (session-based)*
4. *Exploit Semantic Data (the LOD cloud) for further semantic information (also for 1)*

Prototype

- Automatically connects knowledge with documents at query time
- No preprocessing
- No indexing

Prototype (cont)

- Exploitation for restricting the focus

Possible Approaches for enriching Web Search and Entity Search

1. Off-line NEM over the entire corpus
   - The corpus must be available
   - We have to build an appropriate index
   - The size of the index could be in the scale of the corpus!

   Most existing systems — *- T. Cheng et al. [2007]* *- EntityCube* *- MediaFaces*

2. Off-line NEM over the top hits of the frequent queries*
   - We do not have to apply NEM at the entire collection
   - Reduced computational effort and storage space
   - Applicable also at meta-search level

   WWW'12 paper

3. Real-time NEM over the snippets of the top hits *default choice*
4. Real-time NEM over the contents of the top hits *on demand*

   The focus of this paper

* Fafalios et al., *"Scalable, Flexible and Generic Instant Overview Search"*, WWW' 12

Challenges & Questions

Challenges

- Real-time response
  - Perform Entity mining over the snippets
    - Compare the results of mining over *snippets* versus over *full contents*
- Selection/Ranking of entities

  The system may discover numerous entities, but the UI has limited space
  - Develop and evaluate an entity ranking method
    - Show only the top-10 entities. Show more entities (all) on demand.
  - Exploit Linked Open Data

Entity Ranking – Our approach

- We focus on ranking methods that:
  - Do not rely on any log analysis (this is aligned with the dynamic nature of our approach)
- Candidate ranking methods:
  1. Count the documents in which the entity appears (i.e. its frequency) and take into account the rank of the documents that contain the entity.

$$Score_{rank}(e) = \sum_{a \in docs(e)} ((|A| + 1) - rank(a))$$

Set of returned hits

Its position in the answer

  2. Take into account the words of the entity name and the query string, and tolerate small differences with Edit Distance.
  3. Consider both perspective: adopt the harmonic mean of the above scores.

On Exploiting Linked Open Data (1/2)

"Exploiting LOD is more dynamic, affordable and feasible than an approach that require the system to store and maintain its own KB"

Our objectives

- Allow the user to get some basic information about an entity, without needing to submit new queries. Ability to continue browsing the related entities
- Offer a flexible configuration method. For specifying the categories and entities that are interesting for the application at hand

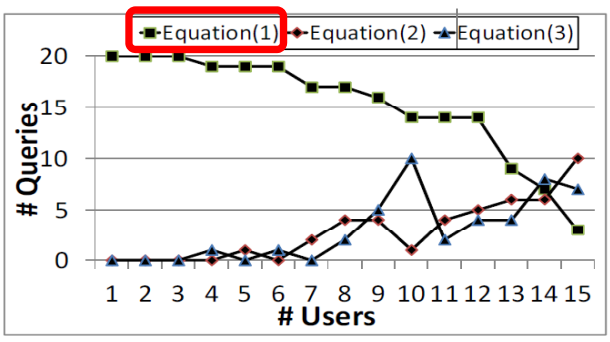
We offer this on demand: By clicking the icon the system checks if that entity lies in a LOD dataset (by performing a SPARQL query); If yes, it collects more information about that entity.

Configurability: The system allows specifying one or more appropriate LOD datasets for each category of entities. E.g. *GeoNames* for entities in "Location", *DBpedia* for entities in "Organization", "Person" and "Location", etc.)

Ability to define new categories and the queries that return entities of interest.

EXPERIMENTAL RESULTS

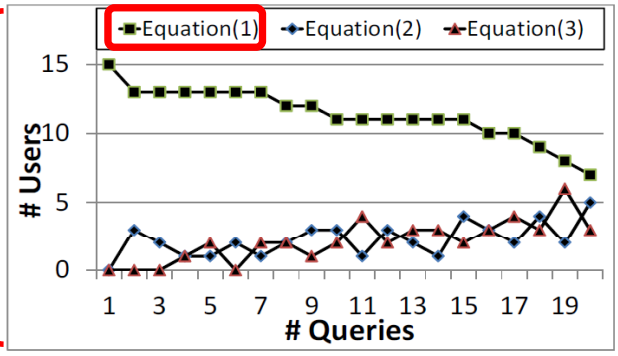
Entity Ranking – Evaluation by Users



Aggregated preferences for each user*

- *15 users*
- *20 queries*
- *3 ranking methods*
- *For each query, user can select one, two, or all rankings.*

Aggregated preferences for each query*

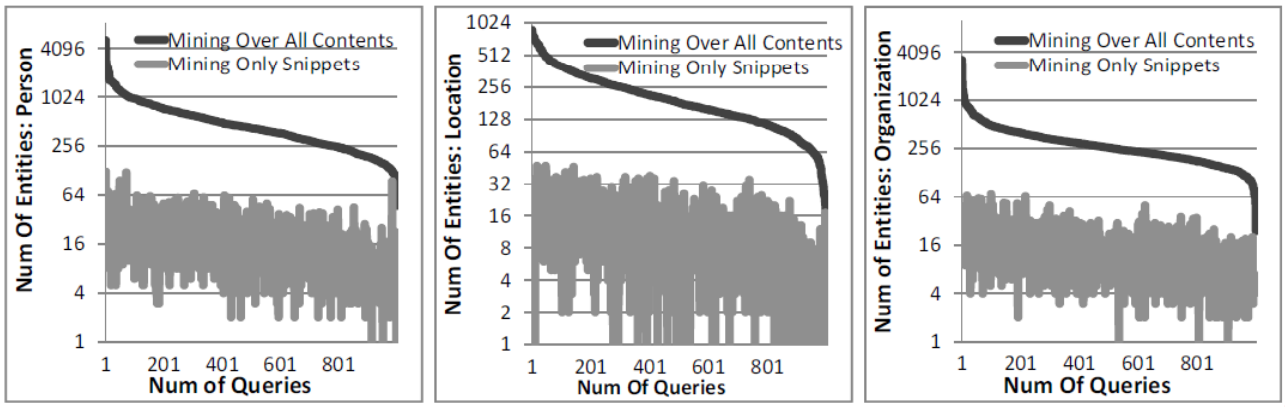


**The aggregation was based on the plurality ranking (by considering only the most preferred options)*

Contents Mining vs. Snippet Mining (1/3)

- Number of identified entities over all contents vs. over only snippets.

- 1000 queries
- NEM over the top-50 hits

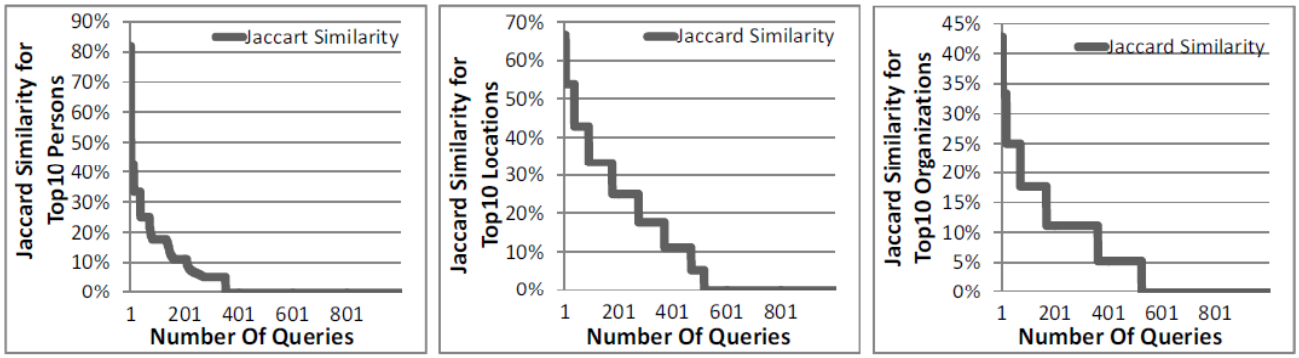


Contents mining yields around 20 times more entities than snippet mining

Contents Mining vs. Snippet Mining (2/3)

- Similarity of the top-10 entities

** Adopting the 1st Ranking approach*



Jaccard Similarity:

- *< 30% for the majority of the queries*
- *0% for about 60% of the queries*

Contents Mining vs. Snippet Mining (3/3)

- Computational and Memory Costs

	Snippet Mining	Contents Mining
Entities per hit:	1.2	10.1
Overall time*:	1.5 seconds	78 seconds
Main memory footprint for one query (for the top-50 hits):	37 MB	300 MB

about **36%** of total time

about **60%** of total time

** Overall time = retrieve results + **download contents** + **apply NEM at results** + apply NEM at query + create representations + rank entities*

CASE STUDIES

Case Study: Fisheries and Aquaculture Publications (1/2)

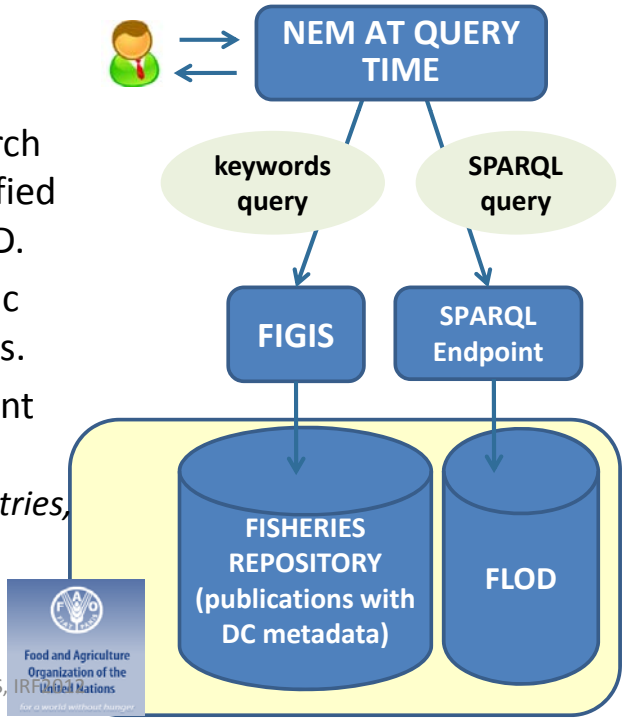
Domain:

- **FAO publications** about fisheries and aquaculture

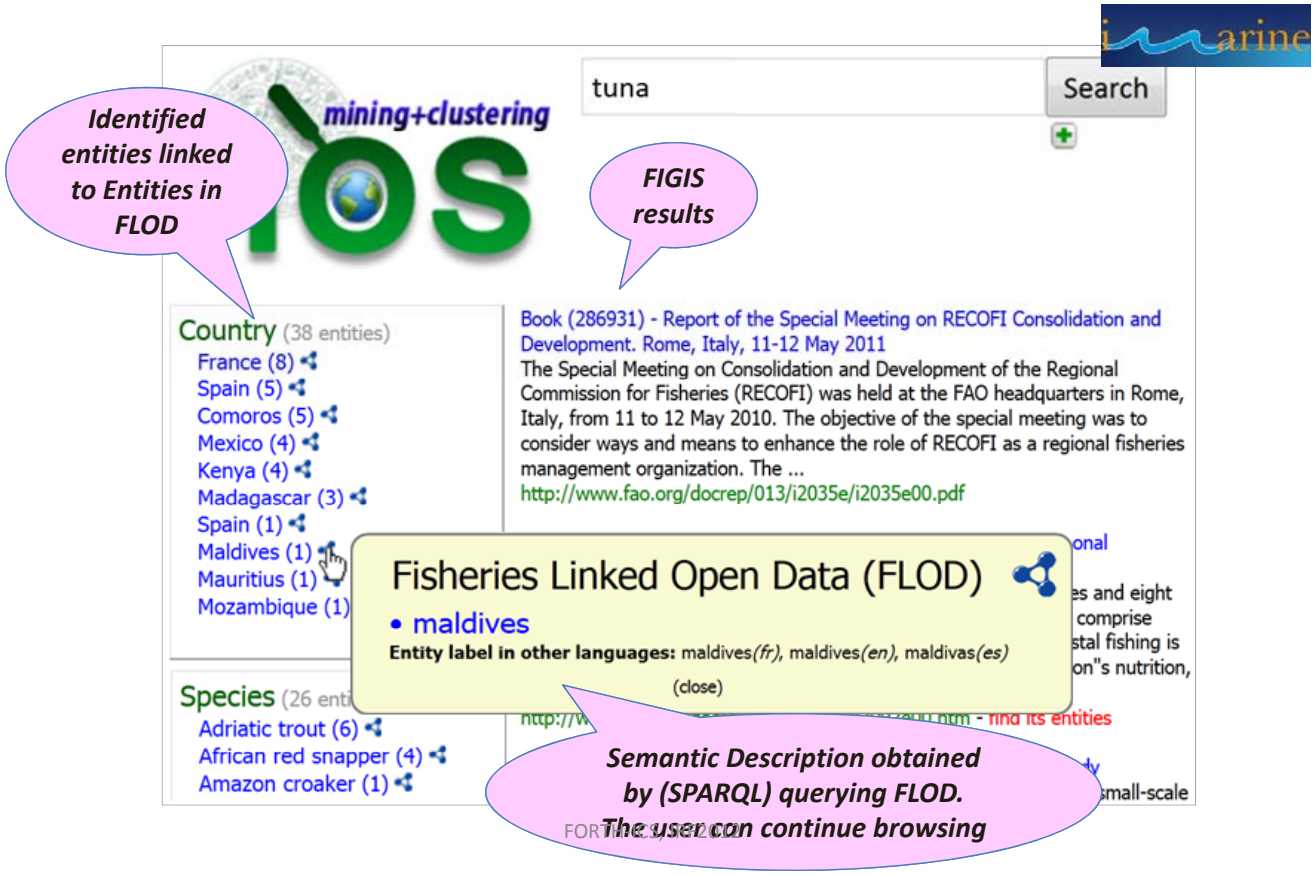
Objective:

- Identify how to enrich keyword search with entity mining where the identified entities are linked to entities in FLOD.
- Create and serve (dynamic) semantic description for the identified entities.

We have identifies the following relevant categories of entities which are semantically described in LOD: *Countries, Water Areas, Regional Fisheries Bodies, Marine Species*



Case Study: Fisheries and Aquaculture Publications (2/2)



Case Study: Patent Search (professional search in general)

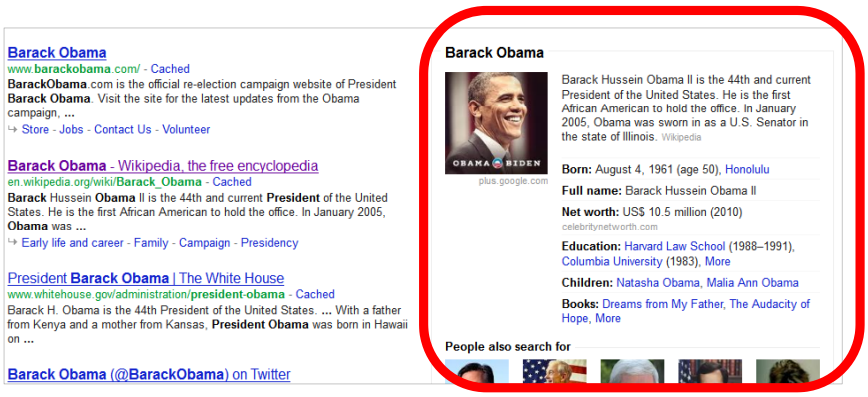
Entity identification and analysis could become a significant aid to professional searches and can be seen (together with other text analysis technologies) as becoming the cutting edge of information retrieval science.

Patent Search

- Missing relevant documents is unacceptable in patent search *(recall oriented search procedure).* Retrieval of all relevant documents is usually necessary
- The related documents contain plenty of named entities
  - *Companies*
  - *Countries*
  - *Persons*
  - *Product types*
  - *Laws*

Related Work: Google Knowledge Graph

although announced after the submission of this paper



Google Knowledge Graph tries to "understand" user's query and presents a semantic description of what user is maybe searching.

- It does NOT locate entities in the search results.
- It does NOT group the results according to the discovered entities.
- It is NOT a recall oriented search
- If the query is not a known entity, the user does NOT get any semantic description

Concluding Remarks

- Enhancing web searching with Named Entity Mining *(at query time, without any preprocessing):*
  - Gives the user an overview of the answer space
  - Allows the user to restrict his focus on a part of the answer
  - Is convenient for user needs that require collecting entities
- Real-time NEM over snippets is feasible and yields about 1.2 entities per snippet
- NEM over contents is more time consuming, but mines much more entities
- String similarity between the query and the entity name does not improve entity ranking (in our setting)
- The top-10 entities derived from snippet mining are quite different from those derived from contents mining (< 30% Jaccard similarity)
- By exploiting LOD for enriching the semantic descriptions of identified entities
  - the user gets useful information about one entity without submitting new queries
  - the user can start browsing the entities that are linked to that entity.

Issues for Further Research

- For the case of *contents mining*: take into account the *local score* of an entity in the document at entity ranking
- Comparatively evaluate with users the top-10 entities from snippets mining with the top-10 from contents mining.
- Long term vision:
  - Mine not only correct entities but probably entire conceptual models (and entity relationships)
  - Support the interaction paradigm of faceted search over such (crispy or fuzzy) semantic models

Thank you for your attention

Demos:

http://www.ics.forth.gr/isl/ios