OUTLINE

- Context
- A few words about Exploratory Search
- An overview of research prototypes from ISL
  - m1: faceted search using static and dynamic metadata
  - m2: instant overview search
  - m3: entity mining over documents
  - m4: configurability of entity mining
  - m5: application in professional search
  - m6: application in an research infrastructure
  - m7: preference-enriched faceted search
  - m8: top-k semantic graphs
Users should enjoy their data
EXPLORATORY SEARCH

Wikipedia:

“Exploratory search is a specialization of information exploration which represents the activities carried out by searchers who are either:

a) unfamiliar with the domain of their goal (i.e. need to learn about the topic in order to understand how to achieve their goal)
b) unsure about the ways to achieve their goals (either the technology or the process)
c) or even unsure about their goals in the first place.

Consequently, exploratory search covers a broader class of activities than typical information retrieval, such as investigating, evaluating, comparing, and synthesizing, where new information is sought in a defined conceptual area; exploratory data analysis is another example of an information exploration activity. Typically, therefore, such users generally combine querying and browsing strategies to foster learning and investigation.”
Therefore...

- Ranking is not enough for exploratory search

Not enough!
EXPLORATORY SEARCH:
FROM FINDING TO UNDERSTANDING

Research tools critical for exploratory search success involve the creation of new interfaces that move the process beyond predictable fact retrieval.

From the earliest days of computers, search has been a fundamental application that has driven research and development. For example, a paper published in the inaugural year of the *IBM Journal* 36 years ago outlined challenges of text retrieval that continue to the present [4]. Today’s data storage and retrieval applications range from database systems that manage the bulk of the world’s structured data to Web search engines that provide access to petabytes of text and multimedia data. As computers have become consumer products and the Internet has become a mass medium, searching the Web has become a daily activity for everyone from children to research scientists.
Some Common Requirements for Effective Exploratory Search

- Allow easy and fast access even to low ranked hits
- Allow browsing and inspecting the found hits in groups (according to various criteria)
- Offer overviews of the search results
  - Compute and show descriptions and count information for the various groups, or other aggregated values
- Allow gradual restriction/ranking of the search results
Research prototypes from ISL related to exploratory search

- They are presented like a “story” organized in milestones that correspond to activities of ISL (Information Systems Laboratory) of FORTH-ICS (2009-now)
MITOS is WSE built from scratch. Apart from the classical WSE functionality, Mitos offers faceted search over the results of the submitted queries.

- It supports facets corresponding to metadata attributes of the web pages (static metadata), as well as facets corresponding to the outcome of snippet-based clustering algorithms (a kind of dynamic metadata).
- The user can then restrict his/her focus gradually, by interacting with the resulting multidimensional structure through simple clicks.
THE MITOS WSE (2009)

- **By clustering**
  - architecture (6)
  - contact (10)
  - content (10)
  - copyright notice csd (13)
  - course content english (7)
  - csd (14)
  - department (8)
  - forth (38)
  - health telematics network (6)
  - ics (39)
  - information (11)
  - network (8)
  - physical (5)
  - science (22)
  - ΕΠΕ ΤΕΧΝΙΚΩΝ ΑΝΩΔΡΟΤΩΝ (22)
  - REST (3981)

- **By domain**
  - gr (4067)

- **By date**
  - 2008 (479)
  - 2007 (694)
  - 2006 (1340)
  - 2005 (184)
  - 2004 (105)
  - 2003 (82)
  - 2002 (88)
  - 2001 (28)
  - 2000 (13)
  - 1999 (4)
  - 1998 (6)
  - 1997 (1)
  - Unknown (1042)

- **By filetype**
  - application/msword (16)
  - application/pdf (1476)
  - application/vnd.ms-powerpoint (29)
  - text/html (2546)

- **By language**
  - Any (UTF-8) (18)
  - Greek (1209)
  - Latin-1 (Europe, Latin America, Caribbean, Canada, Africa) (944)
  - Latin-2 (Central and Eastern Europe) (4)
  - Unknown (1892)

N. Manolis and Y. Tzitzikas (ESWC'11)
We can focus on "By date" facet, clicking the "2009" label.

A user wants to get information about Information Systems Laboratory.

8558 initial results
We can further limit the results, by selecting one of the clusters (they were recomputed for the new focus). The results of the selected group are loaded in the results’ panel and all facets are updated.
With only 2 clicks, we have limited the results to 5 hits.
Evaluation with Users (main results):
- Faceted search, combining dynamically and statically mined metadata
  - lead to much improved task completeness with much less user interactions
- was more preferred by the users (advanced and plain ones) and lead to greater satisfaction, than plain clustering or faceted interfaces

Most Important Related Publications
**PreScan**: Automated extraction of file-embedded metadata from file systems.
MORE ABOUT PRESCAN

Features

- Automatic Scanning of file systems
- Automatic Format Identification and Extraction of Embedded Metadata
- Support for Human-entered/edited Metadata
- Periodic Re-Scannings without loosing the human-provided metadata
- Referential Integrity services

Then we questioned ourselves:
- *why not offering this functionality during query typing, i.e. a kind of richer autocompletion service?*

This resulted to what we called **Instant Overview Search (IOS)**.

**The idea:**
- For the frequent queries, pre-compute and store not only the first page of results, but also the analysis of these hits

**Technical challenge**
- Since the amount of information that has to be stored for each query is higher (and obviously does not fit in main memory) we devised a **partitioned trie-like index** for efficiency (plus a dedicated cache)
We want to find information about the life of Marilyn Monroe
(and probably its connection to Pavlos?)
However, we are not sure for the spelling of her name.
So, we start typing "mari".
List of query’s suggestions.

First page of results of the top suggestion “marilyn”

We can continue typing the query. Instantly new suggestions are shown

Cluster Label Tree of the top suggestion “marilyn”
We selected the suggestion “marilyn monroe”. The results’ first page and cluster label tree for this suggestion were loaded immediately.

By clicking a label, the results of the specific cluster are loaded in the results panel.
We can exploit this technique for any kind of pre-processing of search results (e.g. metadata-based faceted search, snippet-based clustering, entity mining, etc)
**IOS Indexes**

Hosted in main memory (based on requests and space)

Precomputed information (always in hard disk)

![Diagram of IOS Indexes]

**Average Retrieval Time ≈ 135ms**

*Experiments* over a server running on a **modest personal computer**, with a synthetic query log of **1 million** distinct queries and synthetic precomputed information of **1 Terabyte**
**Key results**
- A partitioned trie-based index structure that can efficiently support recommendations for millions of distinct queries even with modest hardware
  - One can provide instant access to large amount of data, utilizing the existing resources, without requiring more hardware
- A hybrid caching policy (70% static and 30% dynamic) seems to be the more appropriate choice yielding a throughput increment of around 80% and a 25% speedup

**Demo**
  - Select the system “Instant Entity Mining + Clustering (over Bing)”

**Related Publications**
Then we questioned ourselves:

- *why not exploiting LOD in the context of entity mining of the search results?*

**Motivation**

- LOD contains plenty of information about Named Entities (their names, attributes, relationships with other entities, etc)

**Output**

- IOS Entity Mining
  - LOD is used as source for Named Entity Recognition
  - LOD is used for providing more information about the identified entities
IOS ENTITY MINING (2012)

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

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

Nicolas Sarkozy
Current President of France
Birth date: 1955-01-28
Birth place: Paris, France
Profession: Lawyer

Barack Obama - Wikipedia, the free encyclopedia
Barack Hussein Obama II (born August 4, 1961) is the 44th and current President of the United States. He is the first African American to hold the office. Obama ...

http://en.wikipedia.org/wiki/Barack_Obama

Barack Obama
Barackobama.com is the official re-election campaign website of President Barack Obama. Visit the site for the latest updates from

Barack Obama | Facebook
Quote of the day: “This is a good first step, but it is only a step. Congress needs to pass the rest of my American Jobs Act so that can create jobs and put ... https://www.facebook.com/barackobama

http://www.facebook.com/barackobama - find its entities


Person
(1427 entities)
Barack Obama
Michelle Obama
George W. Bush
Ann Dunham
Craig Robinson
Joe Biden
John Mccain
Kennedy
Sarkozy
Clinton

Organization
(842 entities)
Harvard
White House
Congress
University of Hawaii
Columbia University

Find its entities

mine only snippets
• Exploitation for restricting the focus
Then we questioned ourselves:

- why not allowing the user to configure himself the entities of interest by exploiting LOD (again in the context of entity mining of the search results)?

Outcome

- X-ENS (eXplore ENtities in Search)

Related Publications

There is no standard covering such configurations. We have defined our own Configuration format. This is the first .. “gap”
X-ENS (SIGIR 2013)

http://139.91.183.72/x-ens

Semantic Entity Enrichment (close)

Properties of: Andy Roddick

Description
Andrew Stephen "Andy" Roddick (born Aug 30, 1982) is an American professional tennis player and a former World No. 1. He is..

Depiction

BirthPlace
Omaha, Nebraska

BirthDate
1982-08-30

Entiti
Then we questioned ourselves:

- **why not applying and testing this in the context of a professional search system?**

**Outcome**

- Application in **patent search**. Missing relevant documents is unacceptable in patent search (*recall oriented search procedure*). Retrieval of all relevant documents is usually necessary.
- Patents contain plenty of named entities of various kinds:
  - *Companies, Countries, Persons, Product types, Laws, etc*
- Inclusion of **PerFedPat** System
  - In collaboration with Mike Salampasis
The proposed functionality:
- offers a tight integration of different search tools with the main retrieval engine,
- connects the search results (i.e. patents) with data and knowledge,
- can be exploited by any patent search system (i.e. it acts as a service over a ranked list of results)
- The time that we have to pay is proportional to the number of the top results that we want to “explore” (≈ 1.5 sec / 100 results)

Related Publications
- P. Fafalios, M. Salampasis and Y. Tzitzikas, Exploratory Patent Search with Faceted Search and Configurable Entity Mining, 1st International Workshop on Integrating IR technologies for Professional Search, in conjunction with ECIR'13, Moscow, Russia, March 2013
Then we questioned ourselves:

- why not applying this in another domain of professional search in the context of a real and operating EU research infrastructure?

Outcome

- X-Search in the context of the ongoing iMarine Research Infrastructure project
XSearch: Semantic post-processing of search results

- Web browsing
- Query terms
- (top-L) results (+ metadata)
- Entities / contents

Visualization/Interaction
- Faceted search, entity exploration, annotation, top-k graphs, etc.

Semantic Analysis
- Grouping
- Ranking
- Retrieving more properties

Semantc post-processing of search results

The LOD cloud

MarineTLO Warehouse

Entity Mining

http://www.i-marine.eu/
EXAMPLE: X-SEARCH deployed in an Operational Research Infrastructure (2012-now)

Semantically Enriched Results

Query: tuna
In Collections: FIGIS

Mined Entities
- FAOCountry(24)
  - Republic of...(1)
  - Viet nam(1)
  - Venezuela(2)
  - Yugoslavia(2)
  - Senegal(1)
- Species(8)
  - eastern Pacific(1)
  - yellowtail a...(1)
  - Ara(1)
  - pantropical ...(1)
  - Indo-Pacific(1)
- WaterAreas(3)
  - Mediterranean(1)
  - Atlantic(1)

Object Metadata
- Thunnus albacares (Bonnaterre, 1788) - Fact sheet
  - Yellowfin tuna (Venezuela), Ca bo Vang (Viet nam), Tuna zutorperka (Yugoslavia)... There are important yellowfin tuna fisheries throughout tropical and subtropical seas. The most major surface fishing techniques for yellowfin tuna in the Pacific, even though this method
- URL: http://www.fao.org/figis/flod/entities/codedentity/3e6d22db-1f06-437d-ac4a-9d3c8b8950f5
- Value: yellowtail amberjack

Textual Clustering
- Root(15)
  - fact sheet(27)
  - thunnus(8)
  - stenella(4)
  - linnaeus_fac..(3)

Semantic Entity Exploration
- URL: http://dbpedia.org/resource/Thunnus_albacares
- Value: Yellowtail amberjack
Properties of: Yellowtail amberjack
- Type:
  - Animal (open)
  - Thing (open)
  - Species (open)
- SameAs:
  - Seriola Islandi (open)
- Subject:
  - C ateon:Fish of the Red Sea (open)
  - C ateon:Fish of the Indian Ocean (open)
  - C ateon:Seriola (open)
- BinomialAuthority:
  - Georges_Cuvier (open)
  - Achille_Valenciennes (open)
- Class:
  - Actinopterygi (open)
  - Chondrichthyes (open)
- Genus:
  - Seriola (open)
- Kingdom:
  - Animal (open)
- Order:
  - Perciformes (open)
- Phylum:
  - Chordata (open)
- Depiction:
  - Seriola_ Islandi.jpg (open)
  - Thumbnail (open)
  - 2000px-Seria_ Islandi.jpg (open)
Then we questioned ourselves:

- What about the ordering of facets, terms and objects? Should the user only restrict the focus? Why not allowing the user to change the order based on his/her preferences?

Outcome:

- A framework for preferences over multi-dimensional and hierarchical information spaces
- An extension of the interaction model of faceted search with preferences
- The Hippalus system that realizes it
SYSTEM: HIPPALUS (2013)

- Allows faceted browsing and also supports Preferences
  - User actions specify the ranking of the information space
  - Gradual preference specification
  - Automatic resolution of conflicts
  - Different preference composition modes
    - E.g. if the user defines the desired ordering wrt each dimension, then the first block of the ranked objects is the skyline

Hippalus: Preference-enriched Faceted Exploration

P. Papadakos\textsuperscript{1,2} and Y. Tzitzikas\textsuperscript{1,2}
HIPPALUS: INTERACTION OVER A KB OF 50 CARS
HIPPALUS: FDT INTERACTIONS

- Object restriction
- Value expansion
- Mouse over
**Hippalus: Preference Actions**

Cars ordered with priority on manufacturer

**Facets**
- **Acceleration** (8)
- **Body_Type** (8)
- **Doors** (8)
- **Drive_System** (8)
- **Engine_Power** (8)
- **Engine_Torque** (8)

Object restriction

**In focus: 8 objects Number of buckets: 8**
- Peugeot-207-ID33
- BMW-1-ID7
- BMW-3-ID8
- Alfa-Romeo-Brera-ID1
- Audi-TT-ID6
- Saab-9-3-ID36
- Alfa-Romeo-8C-ID3
- Mercedes-Benz-SL-ID24

**In focus: 50 objects Number of buckets: 50**
- Hyundai-i30-ID17
- Hyundai-i10-ID16
- Kia-Ceed-ID18
- Ford-Ka-ID
- Land-Rover-ID
- Citroën-ID

**In focus: 50 objects Number of buckets: 33**
- Fiat-Punto-ID
- Mitsubishi-ID2
- Hyundai-i20-ID20
- Toyota-Yaris-ID14
- Volkswagen-ID18

**Level 2:**
- **Price_Euros**

Yannis Tzitzikas, Information Systems Laboratory
MILESTONE 8 FROM DIMENSIONS TO GRAPHS

Then we questioned ourselves:

- *So far we have seen services for getting and exploiting multidimensional spaces over the search results. But what if the notion of dimension cannot be defined, or in case there are too many? What can be done without having to configure entity types?*

Outcome

- A semantic post-processing of results that does not yield a multidimensional space but a graph.

Challenges

- Graph construction and exploitation for identifying the important (useful for the user) nodes and relationships.
**Top-K Semantic Graphs**

- **Query Terms**
- **Google Open Search Professional Search**
- **The LOD Cloud**
- **Entity Mining**
  - Entities of interest identified in the results
  - Retrieving properties and related entities
  - Ranking
- **Semantic Analysis**
  - Semantic data
- **Visualization/Interaction**
  - Faceted search, top-k semantic graphs, entity exploration, ...

Yannis Tzitzikas, Information Systems Laboratory
The system can return the top-K graph for any K from 1 to number of nodes produced

- **Vertices**: the K most highly ranked nodes
- **Edges**: the edges that connect the K most highly ranked nodes

The user is free to increase or reduce the value of K

Example (from a real domain):
This graph:
- can complement the query answer with useful information regarding the connectivity of the identified entities.
- allows users to instantly inspect information that may lie in different places and that may be laborious and time-consuming to locate.
- provides useful information about the context of the identified entities.
- allows the users to get a more sophisticated overview and to make better sense of the results.

---

**Top-K Semantic Graph**

![Graph](image-url)
**Prototype (2014)**

http://139.91.183.72/x-ens-2/

---

**TOP-5 LIST**

1. dbpedia:Yellowfin_tuna
2. dbpedia:Perciformes
3. dbpedia:Scombridae
4. dbpedia:Albacore
5. dbpedia:Thunnus
### Evaluation (main results)

- **Usefulness – Survey for the marine domain**
  - The majority of participants believe that the appearance of a graph of semantic information related to the search results can help them during an exploratory search process.

- **Effectiveness – Comparative evaluation of ranking schemes:**
  - The proposed PageRank-based ranking scheme produces more preferred ranking compared to other link analysis-based algorithms.

- **Efficiency – Case study over online DBpedia**
  - The exploitation of LOD can be supported at query-time.
  - For up to 100 detected entities we can offer the proposed functionality at real-time, even if we query an online KB (like DBpedia).

- **The major bottleneck is the reliability and performance of online SPARQL endpoints**
  - We expect this limitation to get overcome in the near future.
  - In the meanwhile, we can use caching / indexing / dedicated warehouses / distributed infrastructure.

### Related Publications:

- P. Fafalios and Y. Tzitzikas, Post-Analysis of Keyword-based Search Results using Entity Mining, Linked Data and Link Analysis at Query Time, IEEE 8th International Conference on Semantic Computing (ICSC’14), Newport Beach, California, USA, June 2014.
HIPPALUS DEMO

- With Firefox version 8+ try
  http://www.ics.forth.gr/isl/Hippalus
CONTACT PERSON: YANNIS TZITZIKAS (HTTP://WWW.ICS.FORTH.GR/~TZITZIK)