Post-Analysis of Keyword-based Search Results using Entity Mining, Linked Data and Link Analysis at Query Time

Pavlos Fafalios and Yannis Tzitzikas

fafalios@ics.forth.gr tzitzik@ics.forth.gr

Foundation for Research and Technology – Hellas (FORTH)
Institute of Computer Science (ICS)
Information Systems Laboratory (ISL)

University of Crete
Computer Science Department
Greece

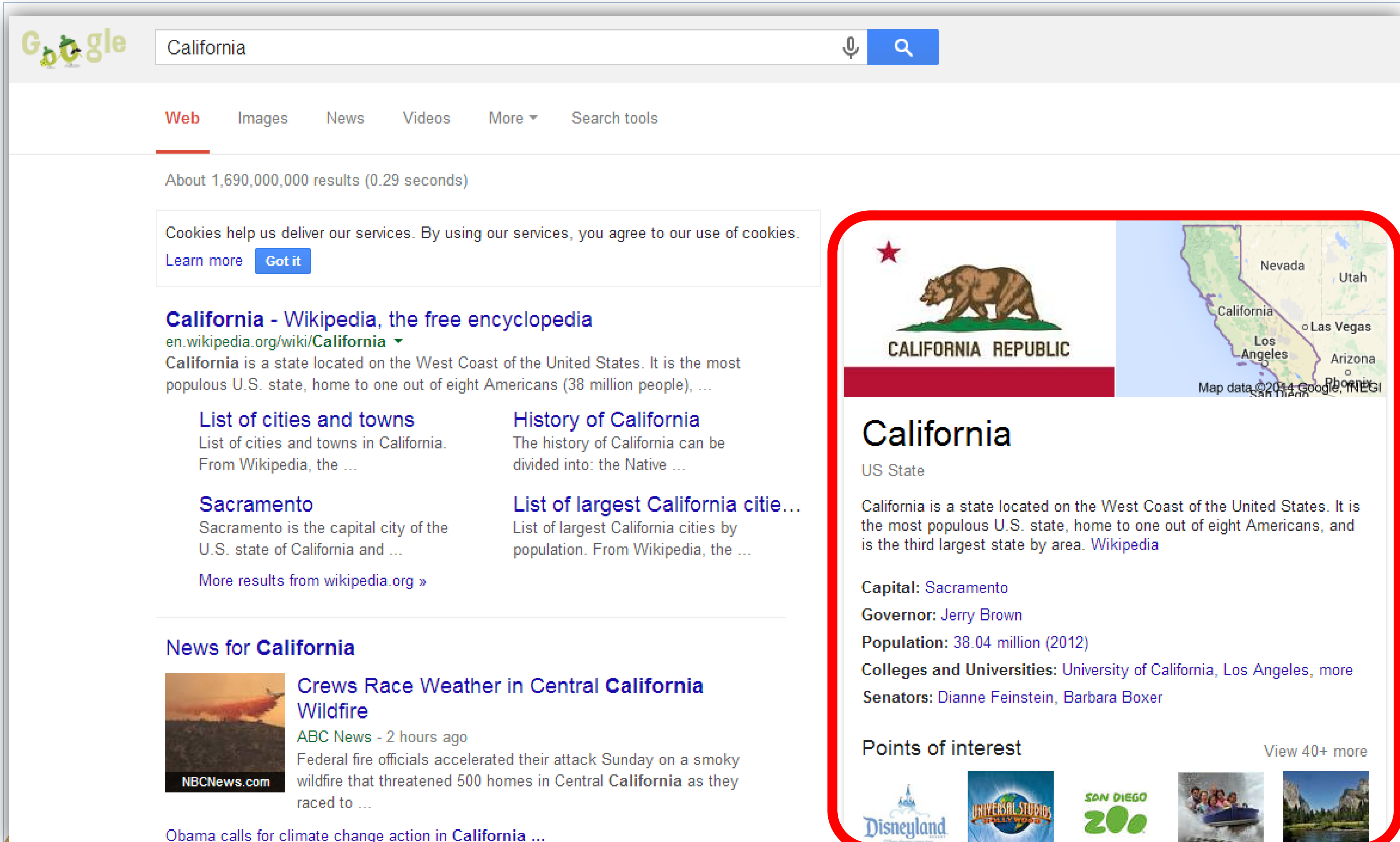
Outline

- Introduction
  - *motivation / challenges / contribution / context*
- A Link Analysis-based approach
  - *ranking entities & properties / producing top-k semantic graphs*
- Evaluation
  - *usefulness / effectiveness / feasibility / scalability*
- Conclusion and Future Research

Motivation

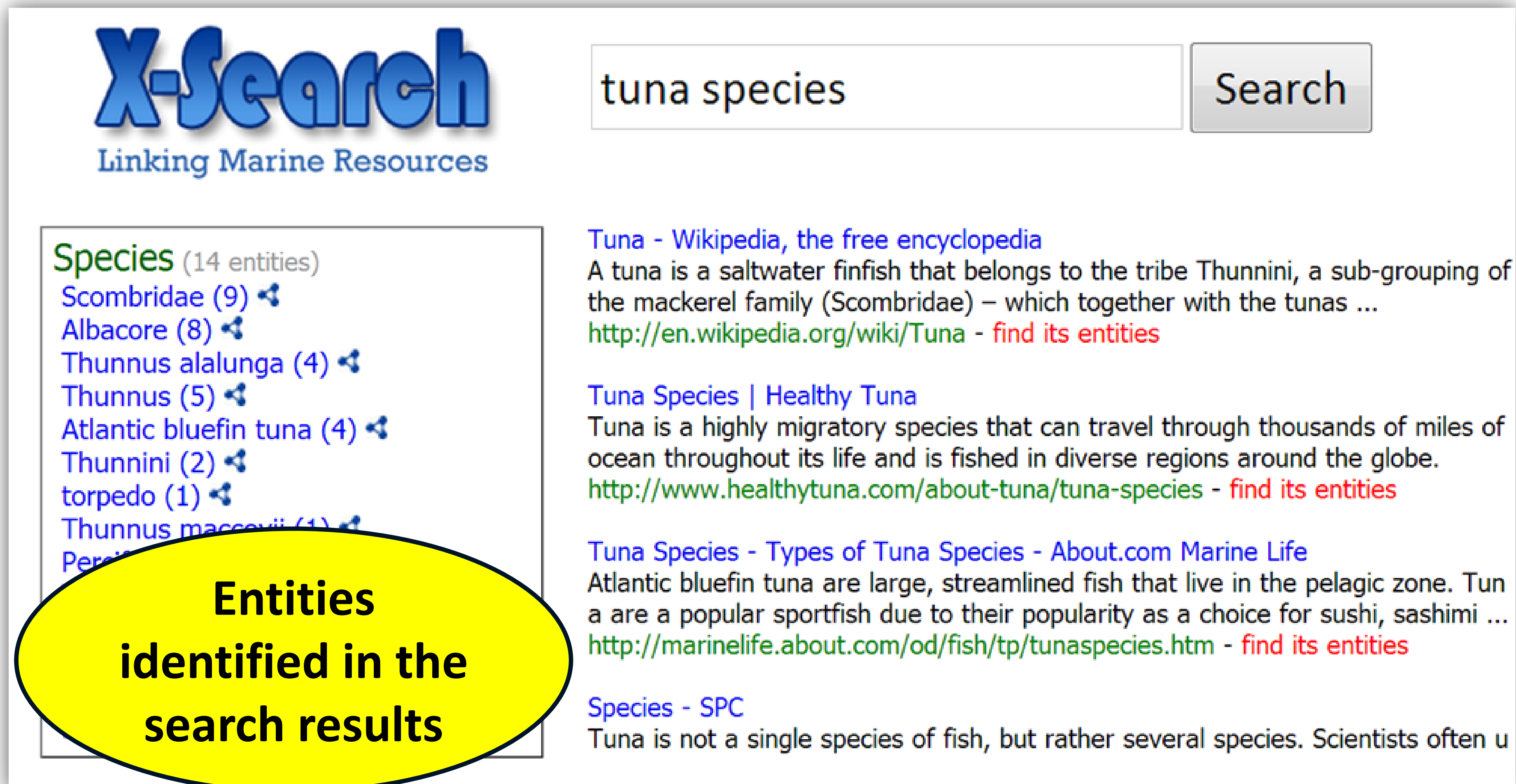
- Most search methods are appropriate for *focalized search*
  - They make the assumption that users can accurately describe their information need and that they are interested only in the top hits
- A high percentage of search tasks are *exploratory*
  - Focalized search leads to inadequate interactions and poor results
  - Specially in **Professional Search**
- Faceted and Dynamic Taxonomies is an approach towards that direction
  - but its applicability over *distributed* and *heterogeneous* sources of various *structuring complexity* is an open challenge

Motivation


California
Web
Images
News
Videos
More
Search tools
About 1,690,000,000 results (0.29 seconds)
Cookies help us deliver our services. By using our services, you agree to our use of cookies.
Learn more
Got it
California - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/California
California is a state located on the West Coast of the United States. It is the most populous U.S. state, home to one out of eight Americans (38 million people), ...
List of cities and towns
List of cities and towns in California. From Wikipedia, the ...
History of California
The history of California can be divided into: the Native ...
Sacramento
Sacramento is the capital city of the U.S. state of California and ...
List of largest California citie...
List of largest California cities by population. From Wikipedia, the ...
More results from wikipedia.org »
News for California
Crews Race Weather in Central California Wildfire
ABC News - 2 hours ago
Federal fire officials accelerated their attack Sunday on a smoky wildfire that threatened 500 homes in Central California as they raced to ...
NBCNews.com
Obama calls for climate change action in California ...
CALIFORNIA REPUBLIC
Nevada
Utah
California
Las Vegas
Los Angeles
Arizona
California
US State
California is a state located on the West Coast of the United States. It is the most populous U.S. state, home to one out of eight Americans, and is the third largest state by area. Wikipedia
Capital: Sacramento
Governor: Jerry Brown
Population: 38.04 million (2012)
Colleges and Universities: University of California, Los Angeles, more
Senators: Dianne Feinstein, Barbara Boxer
Points of interest
View 40+ more
Disneyland
SAN DIEGO ZOO

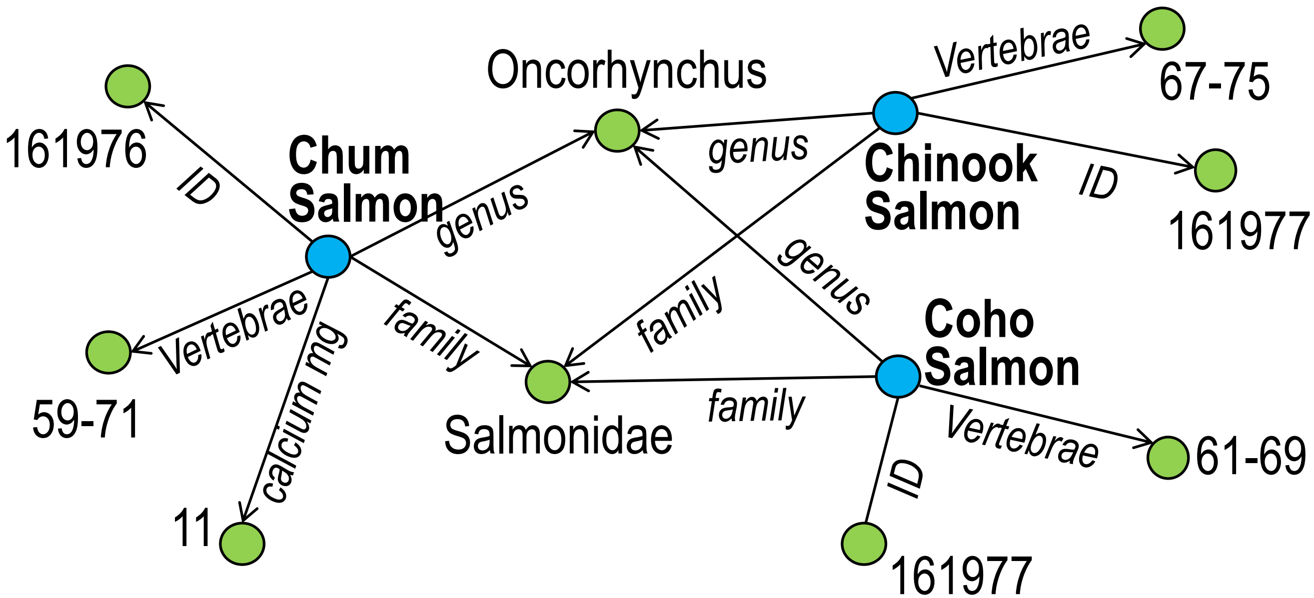
Motivating *(marine-related)* example

iMarine FP7



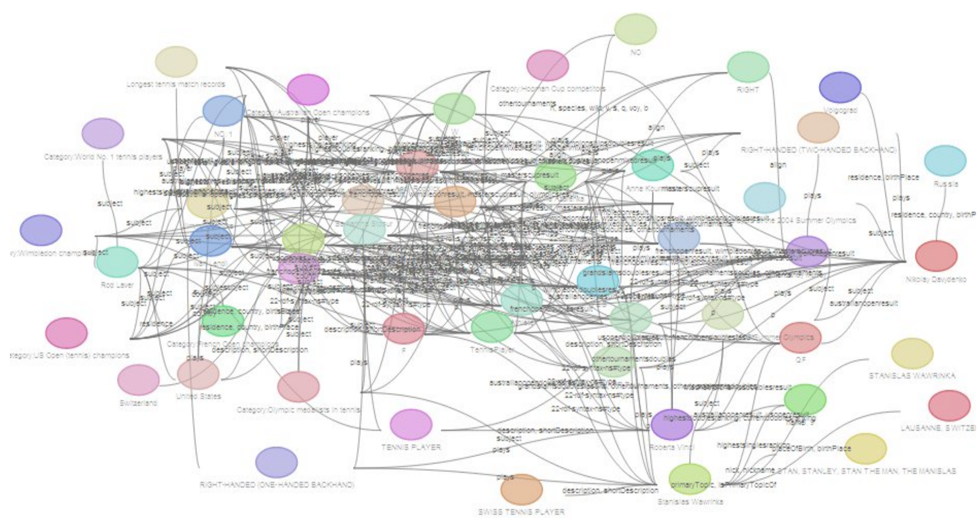
Motivating *(marine-related)* example

- The structured knowledge that is available (e.g. LOD) is not exploited
  - Properties (e.g. genus, family, kingdom, …)
  - Related entities (e.g. predators, preys, water areas, binomial authority, …)
  - Categories/Classes (e.g. Fish, Eukaryote, Fish of Hawaii, …)
- Some entities may share one or more common properties or related entities



Challenges

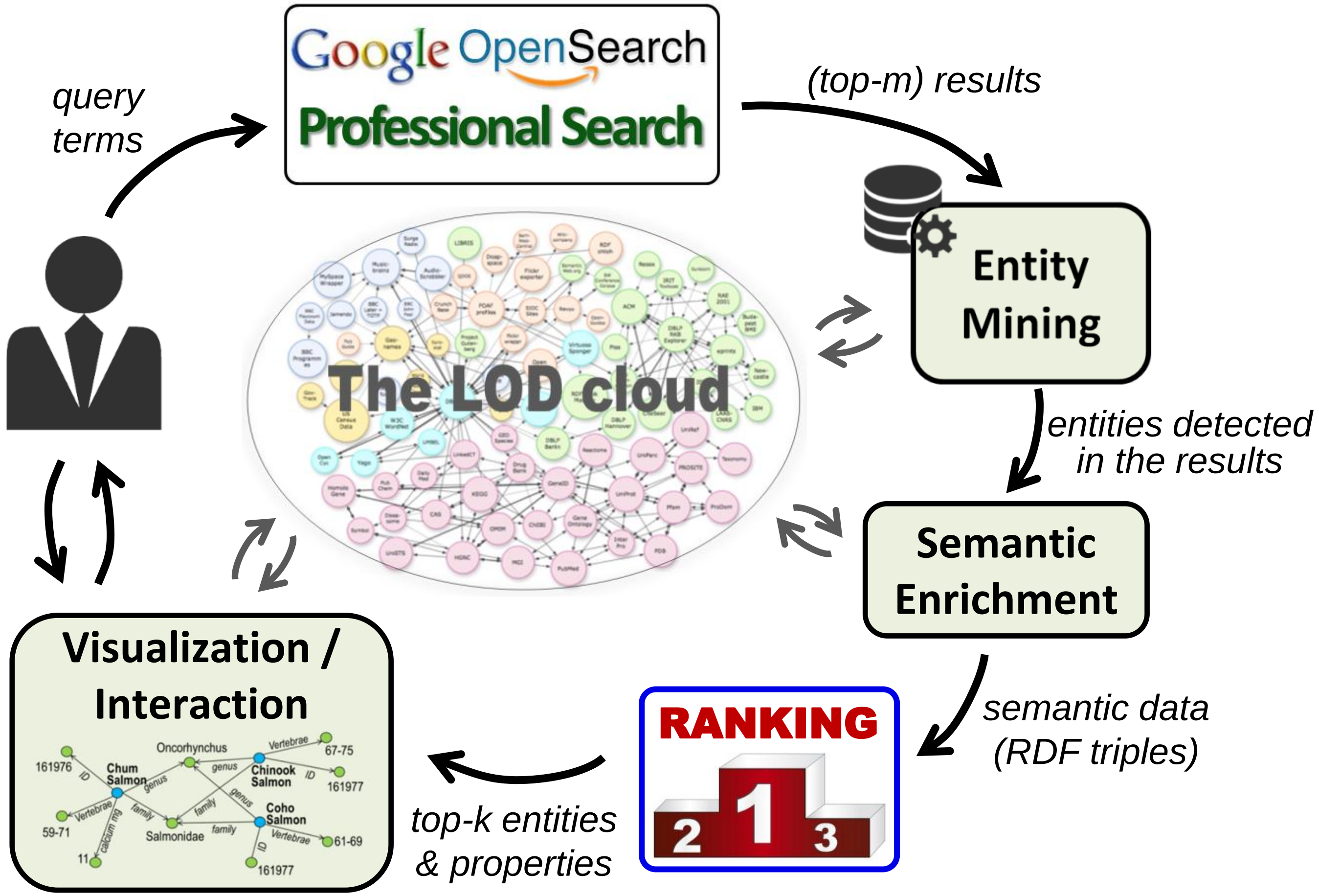
- The number of identified entities can be high
- The amount of structured information that is available for these entities can be high
  - *their associations, properties and categories*
- Need for **ranking** all this semantic information
  - *promote and present to the end-users the <u>most important</u> entities, associations and properties*



Contribution

- We propose a general method for semantic post-processing of search results in which
  - The search results are connected with data and knowledge at query time with no human effort
  - **Named entities** are used as the "glue" for automatically connecting documents (i.e. search results) with data and knowledge.
- We propose a **Link Analysis**-based method for ranking entities and properties
  - This method identifies and promotes the important semantic information
  - The result is exploited for producing and showing **Top-K Semantic Graphs**
- This approach:
  - exploits associations
  - is general and configurable
  - provides a way of making the LOD accessible to the end-users

Context – The Process



Outline

- Introduction
  - *motivation / challenges / contribution / context*
- A Link Analysis-based approach
  - *ranking entities & properties / producing top-k semantic graphs*
- Evaluation
  - *usefulness / effectiveness / feasibility / scalability*
- Conclusion and Future Research

Formalization of Structured (Semantic) Knowledge

- Structured knowledge available in the LOD or queryable through a SPARQL endpoint:
  - RDF URI references: *U*
  - Blank Nodes: *B*
  - Literals: *L*
- A triple $(s_{ubject}, p_{redicate}, o_{bject}) \in (U \cup B) \times U \times (U \cup B \cup L)$ is called an RDF triple
- An RDF Knowledge Base (KB) K, or equivalently an RDF graph G, is a set of RDF triples
- For an RDF graph G_i we shall use U_i, B_i, L_i to denote the **U**RIs, **B**lack nodes and **L**iterals that appear in the triples of G_i

The SEGIE

- The method is based on an RDF graph (called **SEGIE**) that we construct dynamically:

Semantically **E**nriched **G**raph of **I**dentified **E**ntities

q ⇨ *A* ⇨ *E* ⇨ out(e), in(e)

query terms — top-L results — entities identified in A

entity mining using Gate Annie

$\forall e \in E$

$out(e) = \{ o \mid (e, p, o) \in KB \}$

$in(e) = \{ s \mid (s, p, e) \in KB \}$

⇩

S

$E \cup out(E) \cup in(E)$

⇦

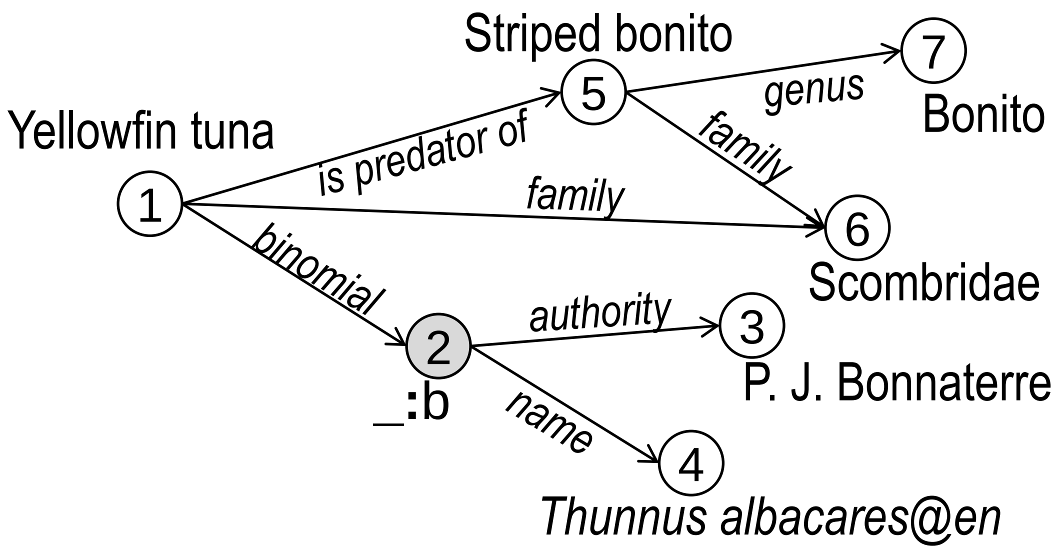
$$X = (E_x, P_x)$$

$E_x = S$

$P_x = directed\ links\ (properties)$

The SEGIE

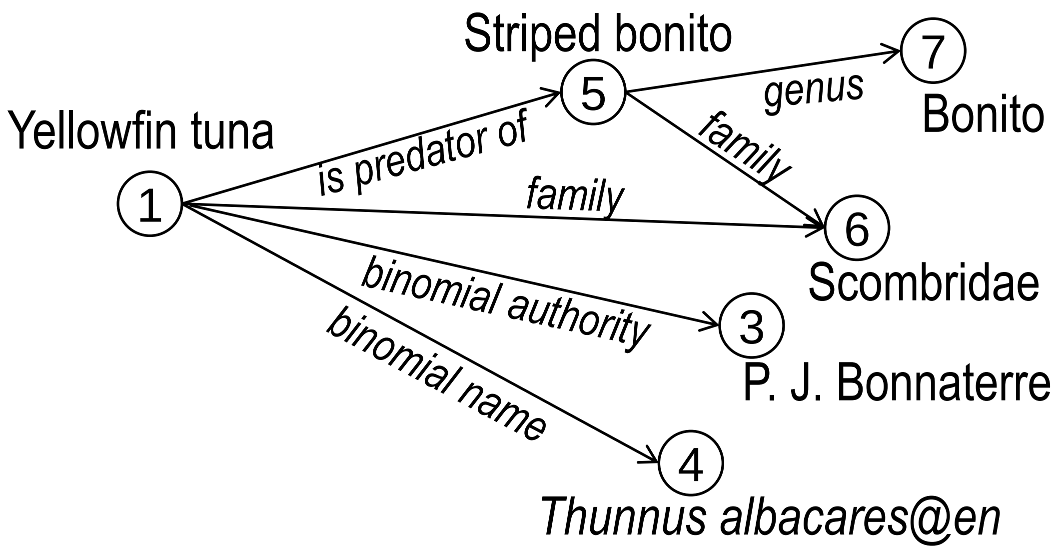
- A simple graph of entities:



- In case the object (or subject) is a blank node (e.g. the node **_b**) we include in the graph the set out(**_b**) (or in**(_b)** respectively) and not the blank node **_b**.
- In that case, we concatenate the names of the properties that are merged.

The SEGIE

- A simple graph of entities:



- In case the object (or subject) is a blank node (e.g. the node **_b**) we include in the graph the set out(**_b**) (or in**(_b)** respectively) and not the blank node **_b**.
- In that case, we concatenate the names of the properties that are merged.

The State Transition Graph

- A graph over which a <u>random walk model</u> can be applied
  - *Its nodes correspond to states and its edges to transitions*
- An edge in the graph represents a property that connect two entities
  - If a property connects two entities, then the two entities are *semantically* bi-connected *(i.e. the difference lies in how we name the property)*
  - For each directed edge that connects two entities ($e_1 \rightarrow e_2$), we consider also the edge of the opposite direction ($e_2 \rightarrow e_1$)
- If an entity e is connected with an entity e' with multiple properties, then e' is probably important for e
  - We specify edge weights: we collapse multiple directed edges that connect two entities into a single one but with higher weight

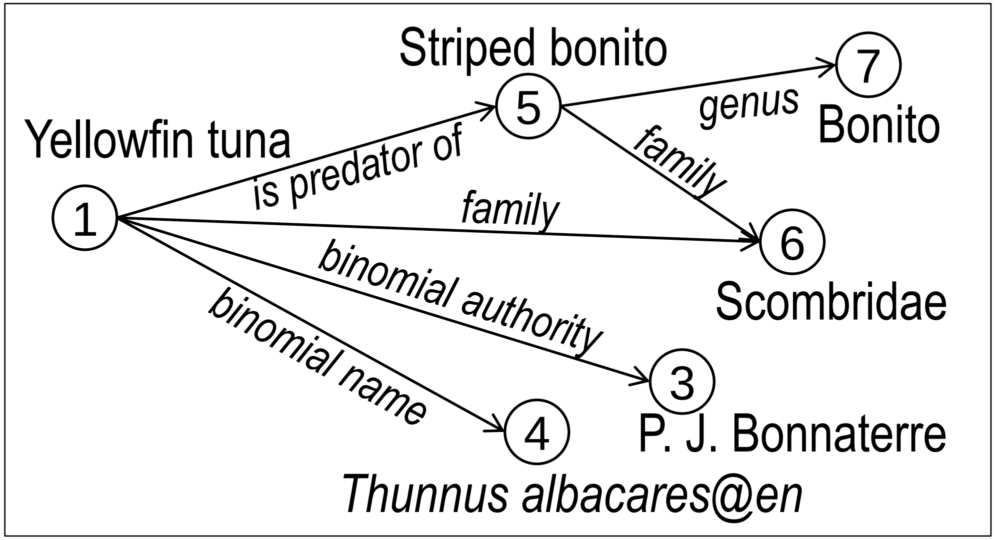
$$\mathrm{w}(\mathrm{e}, \mathrm{e}') = \frac{|props(\mathrm{e}, e')|}{|o(\mathrm{e})|}$$

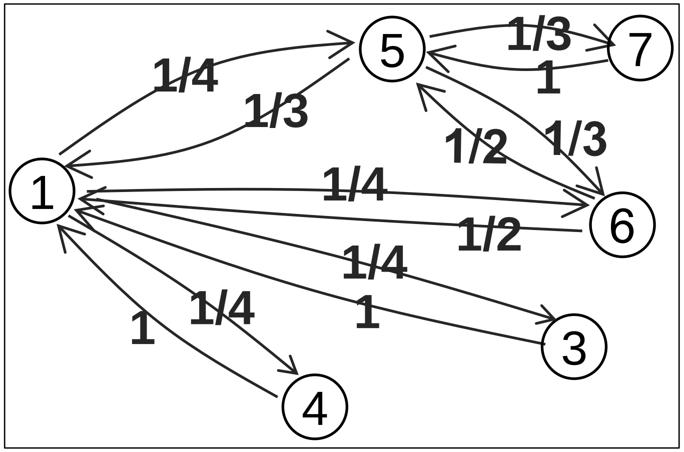
Set of directed edges that connect e with e'

Set of outgoing directed edges of e

The State Transition Graph

The SEGIE

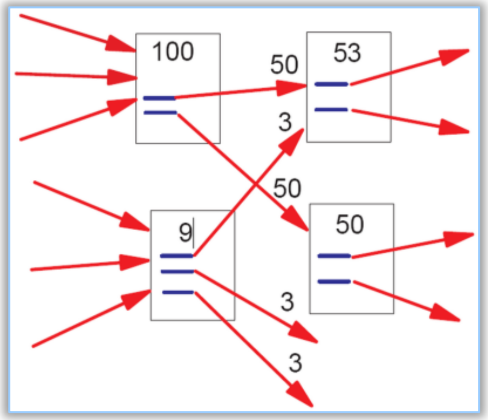




The State Transition Graph (STG)

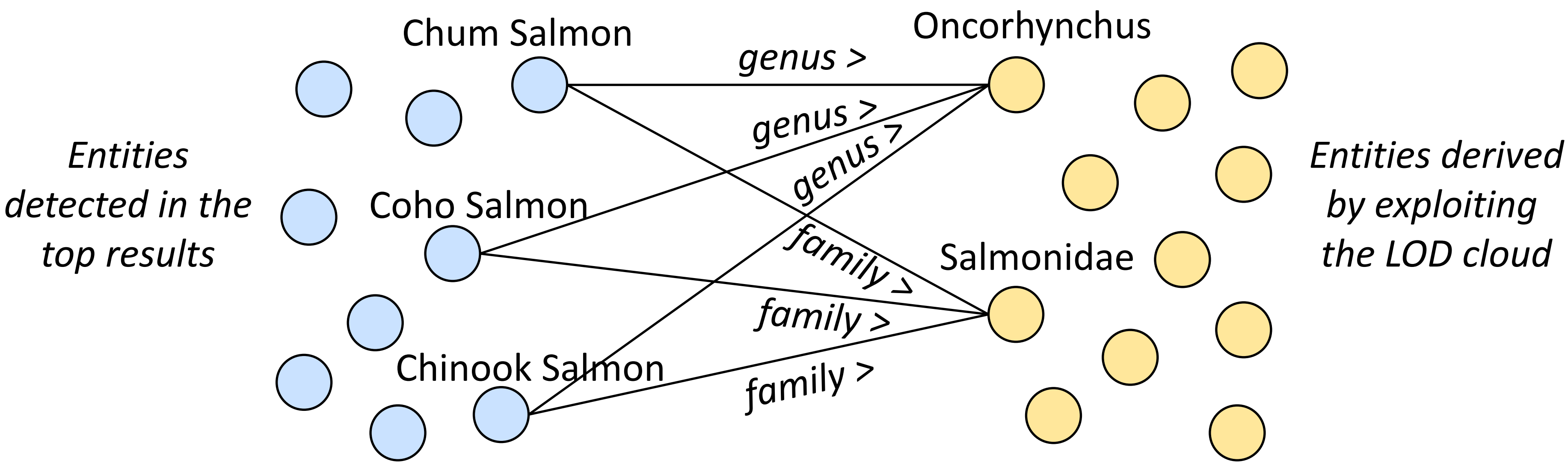
Analyzing the State Transition Graph

- Apply Link Analysis, e.g. a PageRank-like algorithm for identifying the most important entities, properties and associations
- This approach has been successful in Web Search
  - The important web pages are pointed by several other important web pages
  - The importance of a certain web page influences and is being influenced by the importance of some other web pages
- In our problem, an entity or property is considered important (and thus we must present it to the user) if several other important entities point to it.



Analyzing the State Transition Graph

- Apply Link Analysis for deriving the top-k graphs, i.e. the graphs containing the *k* most important entities, where:
  - vertices correspond to entities,
  - edges correspond to associations among entities.



Top-5 graph

Analyzing the State Transition Graph

An entity, i.e. a vertex in the graph

decay factor

The set of (directed) edges that connect e with e'

$$r(e) = \boxed{q \cdot Jump(e)} + \boxed{(1-q) \cdot \sum_{e' \in i(e)} \frac{|props(e',e)|}{|o(e')|} r(e')}$$

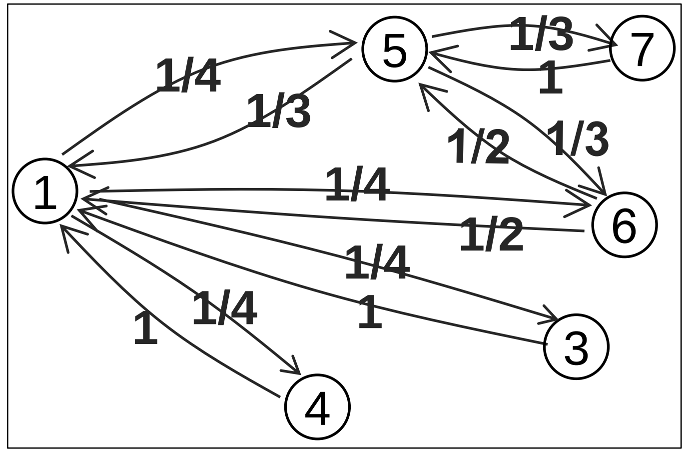
*Expresses the probability of a random jump to e, and can be defined as **Jump(e) = 1 / |S|** if we assume **uniform** distribution*

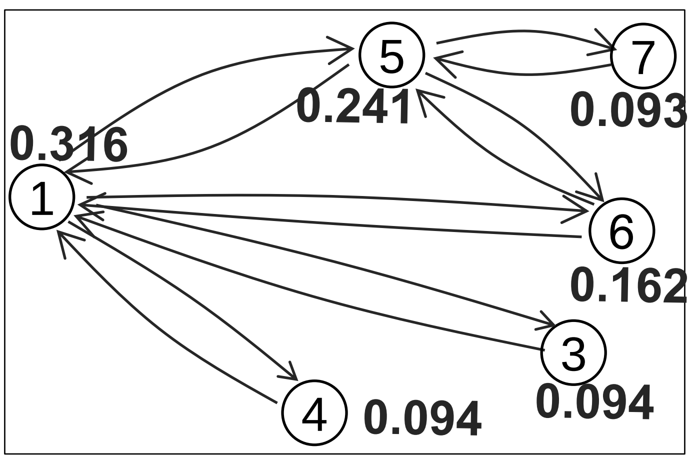
The set of ingoing (directed) edges of e

The set of outgoing (directed) edges of e'

- The scores can be computed iteratively and iterations should be run to convergence
  - *The number of iterations required for convergence is empirically O(logn), where n is the number of edges*

Analyzing the State Transition Graph





1 Yellowfin tuna ≈ 0.316

5 Striped Bonito ≈ 0.241

6 Scombridae ≈ 0.162

3 4 P.J.Bonnaterre, Thunnus *albacares* ≈ 0.094

7 Bonito ≈ 0.093

Performing 10 iterations, decay factor = 0.15

Promoting the Top-ranked Hits

- Ranking is very important in web searching
- We consider that the top results in the ranked list probably contain more useful entities than the last results
  - since they are considered better results for the current query terms
- We bias PageRank:
  - $r(e) = q\ \boxed{\boldsymbol{Jump(e)}} + (1-q) \cdot \sum_{e' \in i(e)} \frac{props(e',e)}{|o(e')|} r(e')$
  - We score higher the entities that have been discovered in the first results than those discovered in the last results

$$\boxed{Jump(e) = \frac{HitScore(e)}{\sum_{e' \in E} HitScore(e')}}$$

An entity discovered in the top-L search results

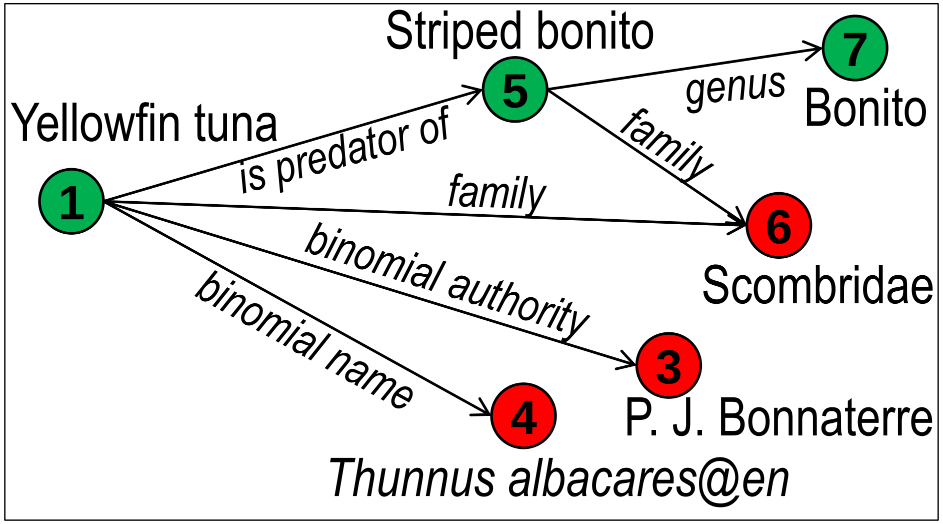
Number of top hits

Position of 'a' in the answer

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

The results in which entity 'e' has been identified

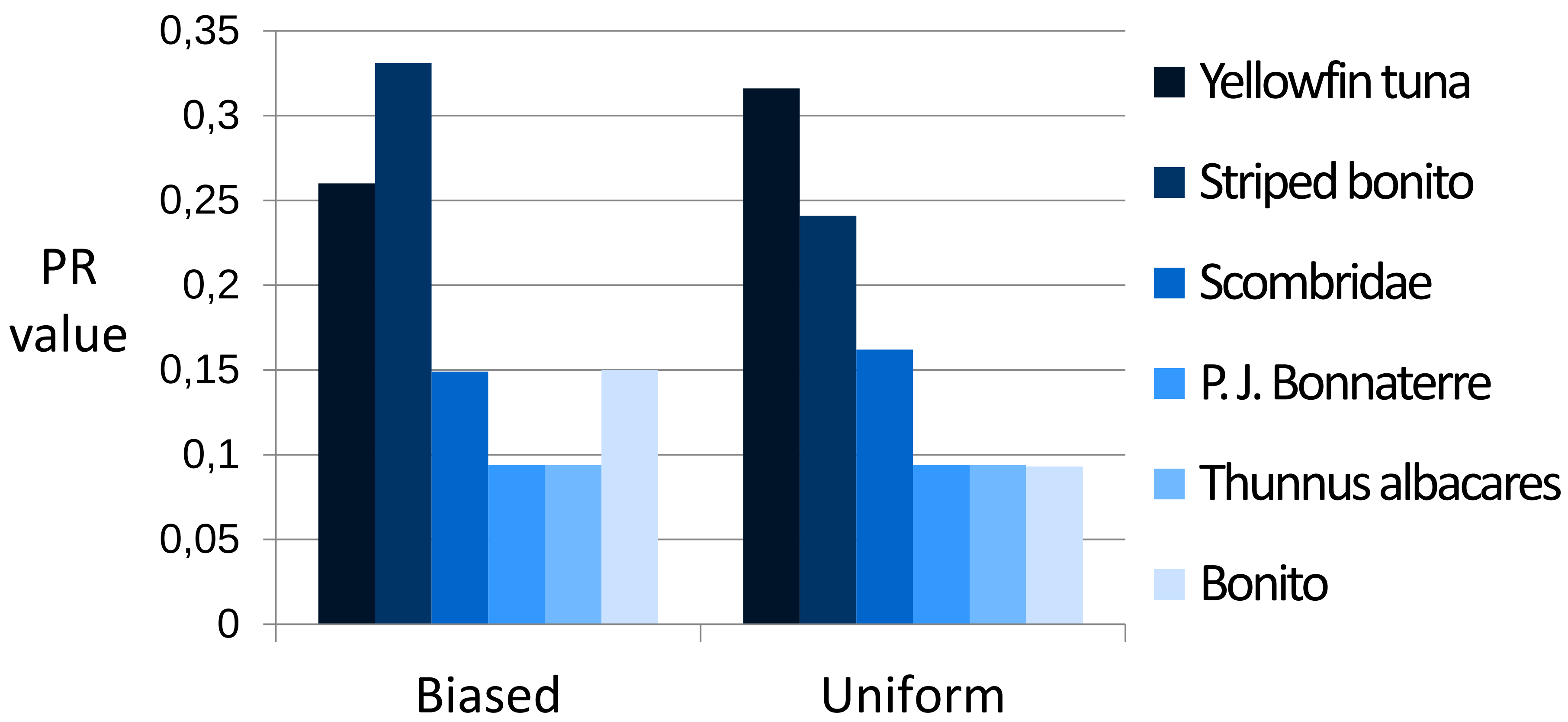
Promoting the Top-ranked Hits – Example



We perform entity mining in the top-10 results and get the following results:

- **Striped bonito** (node 5) was detected in the **1st**, **2nd** and **3rd** result
- **Bonito** (node 7) was detected in the **1st** and **3rd** result
- **Yellowfin tuna** (node 1) was detected in the **8th** result only
- **P.J. Bonnaterre** (node 3), **Thunnus albacares** (node 4) and **Scombridae** (node 6) were **not** detected in the top-10 results (they were derived by exploiting the LOD cloud)

Promoting the Top-ranked Hits – Example



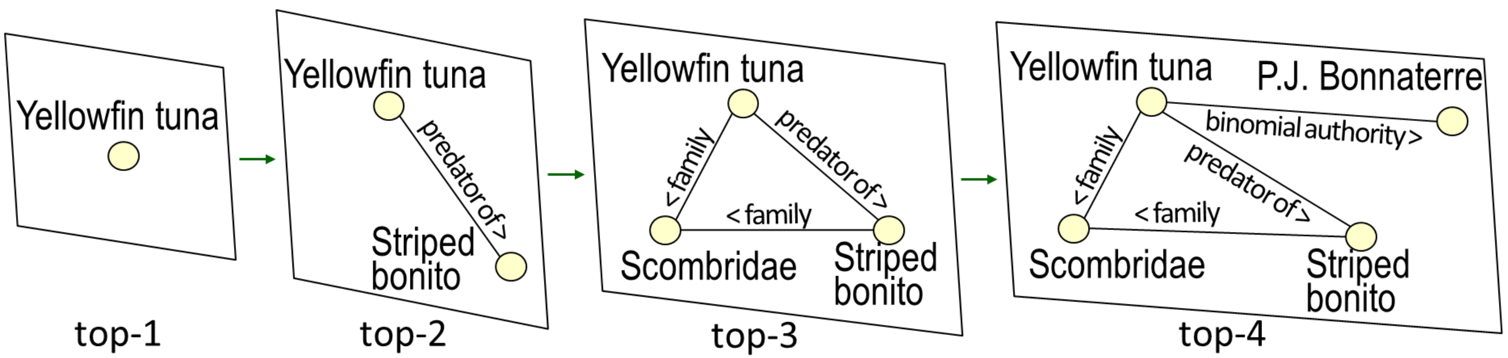
Performing 10 iterations, decay factor = 0.15

Biased Jumps

- Exploit the biased approach for supporting also other kinds of “promotion”/personalization
  - Promotion of entities coming from a particular KB
  - Promotion of entities of one or more RDF classes
  - Personalized / Collaborative promotion of entities
    - *E.g., according to user context*

Visualization: the Top-K graph

- The system can return the top-K graph for any K from 1 to number of nodes produced
  - <u>Vertices:</u> the K most highly ranked nodes
  - <u>Edges:</u> the edges that connect the K most highly ranked nodes
- The user is free to increase or reduce the value of K



- This graph **complements** the query answer with useful information regarding the <u>connectivity</u> of the identified entities
- Several user actions could be supported over these graphs
  - *Integration in the search process*

Outline

- Introduction
  - *motivation / challenges / contribution / context*
- A Link Analysis-based approach
  - *ranking entities & properties / producing top-k semantic graphs*
- Evaluation
  - *usefulness / effectiveness / feasibility / scalability*
- Conclusion and Future Research

Usefulness: Survey on the marine domain

- **Objective:**
  - Get a first feedback for the usefulness of the proposed approach
  - Study whether the depiction of associations among the derived semantic information can help the users in an exploratory search process
- Survey based on a questionnaire (Google Form)
  - We asked persons related to the marine domain *(mainly marine biologists)* to answer a few questions related to 5 particular queries
  - Each query corresponds to a different <u>query type</u> [Pound et al, WWW 2010]:
    - **Entity query:** *yellowfin tuna*
    - **Type query:** *jack fishes*
    - **Attribute query:** *chum salmon genus*
    - **Relation query:** *zander and walleye*
    - **Other keyword query:** *fishing in Hawaii*

Survey on the marine domain – Setting

- For each query:
  - We retrieved the top-100 snippets returned by Bing
  - We performed entity mining in these snippets
  - By exploiting DBpedia, we retrieved the incoming and outgoing properties of each entity URI
  - We applied the proposed approach for deriving the top-5 semantic information
  - We depicted the derived semantic information as:
    - A **top-5 list**
    - A **top-5 graph**

Survey on the marine domain – Questions

- **[Q1]** In an exploratory search process regarding the query <here_the_query>, how would you prefer to see the top-5 entities and properties related to that query?
  - Only the LIST is enough | Only the GRAPH is enough | I would like to see BOTH | I do not want to see neither the list nor the graph
- **[Q2]** In an exploratory search process regarding the query <here_the_query>, do you believe that the appearance of a graph of semantic information related to the search results can help the user during his/her search process?
  - Yes | Maybe Yes, it depends on the interaction model & the quality of the graph visualization | Maybe No | No
- We distributed the questionnaire to marine biologists and persons working on marine-related projects
  - *…who have knowledge on marine species*

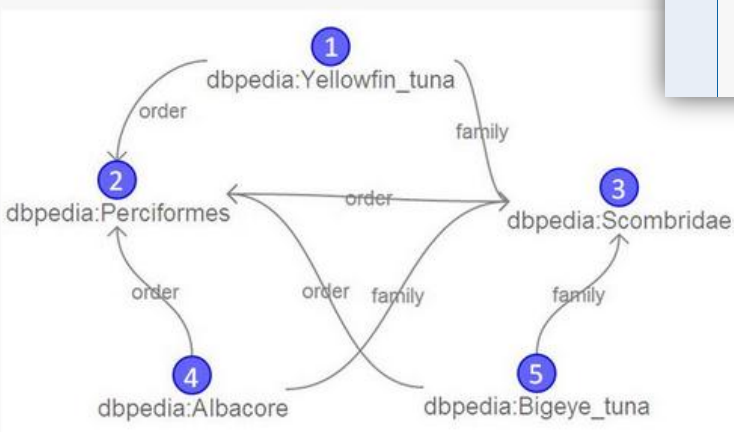
Survey on the marine domain – Questionnaire

Query "yellowfin tuna" (1/5)

Top-5 List

dbpedia:Yellowfin_tuna
dbpedia:Perciformes
dbpedia:Scombridae
dbpedia:Albacore
dbpedia:Bigeye_tuna

Top-5 Graph



In an exploratory search process regarding the query "yellowfin tuna", how would you prefer to see the Top-5 entities and properties related to that query? *

Please do not consider the quality of the visualization of the list and the graph.

- ○ Only the LIST is enough
- ○ Only the GRAPH is enough
- ○ I would like to see BOTH
- ○ I do not want to see neither the list nor the graph

Query "yellowfin tuna" (1/5)

In an exploratory search process regarding the query "yellowfin tuna", do you believe that the appearance of a GRAPH of semantic information related to the search results can help the user during his/her search process? *

- ○ Yes
- ○ Maybe Yes. It depends on the interaction model and the quality of the visualization of the graph
- ○ Maybe No
- ○ No

Survey on the marine domain – Results

- 30 subjects participated in the user study
  - 22 to 60 years old
  - 6 countries
  - 12 organizations

- **[Q1]** In an exploratory search process regarding the query <here_the_query>, how would you prefer to see the top-5 entities and properties related to that query?

QUERY	ONLY LIST	ONLY GRAPH	BOTH	NO LIST, NO GRAPH
yellowfin tuna	23%	30%	43%	3%
jack fishes	13%	37%	47%	3%
chum salmon genus	17%	37%	43%	3%
zander and walleye	13%	43%	40%	3%
fishing in Hawaii	23%	37%	30%	10%

Survey on the marine domain – Results

- **[Q2]** In an exploratory search process regarding the query <here_the_query>, do you believe that the appearance of a graph of semantic information related to the search results can help the user during his/her search process?

QUERY	YES	MAYBE YES	MAYBE NO	NO
yellowfin tuna	23%	63%	10%	3%
jack fishes	27%	67%	7%	0%
chum salmon genus	33%	57%	7%	3%
zander and walleye	33%	53%	13%	0%
fishing in Hawaii	30%	43%	23%	3%

Effectiveness: Comparative evaluation of ranking schemes

- User study regarding the marine domain
- **Objective:**
  - evaluate the effectiveness of the proposed (PageRank-based) ranking scheme
- Comparative evaluation of:
  - Proposed biased PageRank algorithm (**BiPR**)
  - Plain PageRank algorithm (**PR**)
  - Spreading Activation (**SA**)
    [a=0.85, firing threshold=0.00001, initial activation of an entity e = jump(e)]
- We deployed a Web application which implements the proposed functionality
  - Keyword-based queries
  - Entity mining in the top-100 snippets returned by Bing
  - Fish Species (from Dbpedia) as the entities of interest
  - Exploiting DBpedia for retrieving the properties of the identified entities
  - The users could submit their own queries

Comparative evaluation of ranking schemes – Setting

- For each submitted query, the system presents three <u>top-10 lists</u> of ranked semantic information related to the results
  – the one next to the other with random display order
  – each one is produced by one of the aforementioned ranking schemes (BiPR, PR and SA)
- The user can evaluate each ranking by selecting one of the following options:
  – 1 (poor)
  – 2 (not bad)
  – 3 (good)
  – 4 (very good)
  – 5 (excellent)
- The user can inspect many top-K lists for several values of K

Comparative evaluation of ranking schemes – Application

Guidelines

- You will see three top-10 lists. Each list has been produced by applying a different ranking algorithm on the retrieved semantic information. Specifically, 4813 entities and properties were ranked.
- By taking into account the submitted query `(tuna species)`, investigate the three lists and provide a score for each ranking algorithm. ***Please consider and judge which of the three rankings display the more relevant entities in the top positions (and/or which rankings display few irrelevant entities).***
- Recall that the displayed semantic information has been derived by exploiting only DBpedia, so please, during the evaluation, do not judge the *completeness* of the displayed information, but just its ranking.
- The display-order of the three candidate algorithms derives randomly.
- You can inspect several top-k lists of the candidate algorithms (e.g. top-5, top-15, top-20, etc.) by clicking the corresponding number above the lists.

Show Top-5 |Top-10 |Top-15 |Top-20 |Top-25 |Top-30 |Top-35 |Top-40 |Top-45 |Top-50

TOP-10 LIST	TOP-10 LIST	TOP-10 LIST
1. dbpedia:Perciformes	1. dbpedia:Perciformes	1. dbpedia:Perciformes
2. dbpedia:Scombridae	2. dbpedia:Scombridae	2. dbpedia:Shark
3. dbpedia:Haddock	3. dbpedia:Albacore	3. dbpedia:Scombridae
4. dbpedia:Thunnus	4. dbpedia:Thunnus	4. dbpedia:Haddock
5. dbpedia:Pacific_bluefin_tuna	5. dbpedia:Yellowfin_tuna	5. dbpedia:Turbot
6. dbpedia:Skipjack_tuna	6. dbpedia:Atlantic_bluefin_tuna	6. dbpedia:Skipjack_tuna
7. dbpedia:Turbot	7. dbpedia:Shark	7. dbpedia:Swordfish
8. dbpedia:Wahoo	8. dbpedia:Haddock	8. dbpedia:Thunnus
9. dbpedia:Atlantic_bluefin_tuna	9. dbpedia:Skipjack_tuna	9. dbpedia:Yellowfin_tuna
10. dbpedia:Yellowfin_tuna	10. dbpedia:Bigeye_tuna	10. dbpedia:Bigeye_tuna
evaluate the ranking:	evaluate the ranking:	evaluate the ranking:

Comparative evaluation of ranking schemes – Results

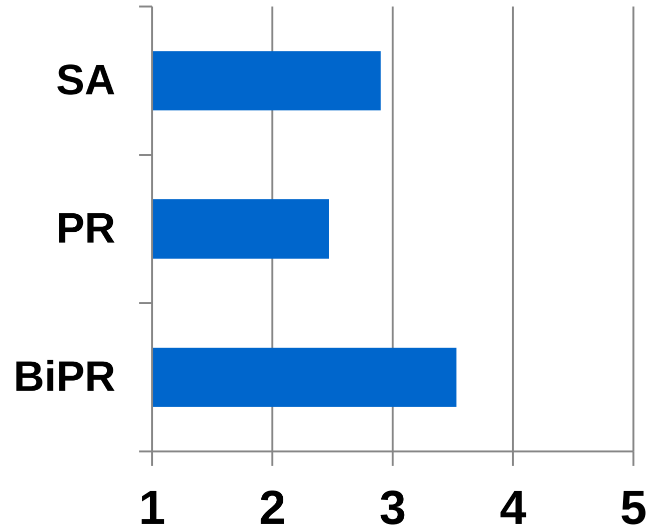
- 17 subjects performed the evaluation
  - Part of those who completed the survey
- 51 queries were submitted, for each query in average:
  - 11.5 entities detected in the search results
  - 4685 triples were derived from DBpedia
  - 2031 entities and properties had to be ranked

The full results, the derived semantic data for each submitted query, and the top-200 rankings as produced by each one of the three ranking algorithms, are available to download through: http://139.91.183.72/x-ens-2/fullEvalResults.zip

Comparative evaluation of ranking schemes – Results

Average scores:

- Biased PageRank (BiPR): **3.53/5** (good to very good)
- PageRank (PR): **2.47/5** (not bad to good)
- Spreading Activation (SA): **2.9/5** (almost good)



Efficiency

- Real-time entity mining using Gate Annie in the top-100 snippets costs about 1 second [Fafalios et al. 2012, Fafalios et al. 2013]
  - 10 ms / snippet
- We measure the time for:
  1. Creating the SEGIE (accessing DBpedia's online N3/Turtle files at real-time)
  2. Creating the STG
  3. Running PageRank
  4. Creating a top-500 graph
- We run the experiments for various numbers of randomly selected entities belonging to 10 randomly selected RDF classes
  - *In real setting, the randomly selected entities correspond to entities discovered in the search results*

- For achieving accuracy we repeated the experiments 20 times.

- The experiments were carried out using an ordinary laptop with processor Intel Core i5 @ 2.4Ghz CPU, 4GB RAM and running Windows 7 (64 bit). The implementation is in Java 1.7 and for the creation and the management of the graphs we use the Java Universal Network/Graph Framework (JUNG)

Efficiency – Results

#entities	SEGIE #vertices	SEGIE #edges	STG #edges	Top-500 Graph #edges
50	2,573	3,790	7,580	889
100	4,133	6,193	12,386	1,493
500	20,743	34,816	69,632	3,471
1,000	49,954	84,893	169,786	3,411
10,000	528,815	995,981	1,991,962	3,421

#entities	SEGIE creation time	STG creation time	Time for Running PageRank	Top-500 creation time
50	1.4 sec	28 ms	194 ms	42 ms
100	2.9 sec	95 ms	329 ms	68 ms
500	13 sec	298 ms	1.7 sec	343 ms
1,000	27 sec	480 ms	3.9 sec	552 ms
10,000	258 sec	8 sec	58 sec	22 sec

Efficiency – Time dependencies

Task	Time depends on:
1) Retrieving (top) results	Underlying search system
2) Performing entity mining	i) Number of (top) results, ii) part of the answer upon which we perform entity mining (e.g. snippets or full contents), iii) number of categories for which we detect entities
3) Creating X-Graph (i.e. retrieving LOD related to the detected entities)	i) Number of detected entities, ii) underlying knowledge bases, iii) categories of the detected entities
4) Creating the STG	Number of triples in X-Graph (i.e. number of edges)
5) Running PageRank	i) Number of iterations, ii) number of edges in STG
6) Creating the Top-K graph	Number of vertices and edges in X-Graph

Scalability and Reliability

- Existing publicly available Knowledge Bases are not reliable
  - They mainly serve demonstration purposes
  - Their efficiency and availability changes over time
- For big number of detected entities the process can be time consuming.
  - Solution: retrieve LOD only for the top-m (e.g. m=100) detected entities as returned by the plain entity mining approach [Fafalios et al. 2012]
  - The top entities are those that lie in the most of the top-ranked results, therefore they are probably the more important
  - In this way, we can bound the maximum response time
- In a real application:
  - The underlying KBs may not be publicly available
  - A dedicated **Warehouse** can be constructed that will serve the application
  - Distributed infrastructure

Outline

- Introduction
  - *motivation / challenges / contribution / context*
- A Link Analysis-based approach
  - *ranking entities & properties / producing top-k semantic graphs*
- Evaluation
  - *usefulness / effectiveness / feasibility / scalability*
- Conclusion and Future Research

Conclusion

- General, flexible and adaptive method for semantic post-processing of search results which is based on Entity Mining and Linked Data
  - This approach shows how search results can be integrated with external sources of structured (semantic) information
- Link Analysis-based method for selecting the semantic information that better characterizes the search results
  - Biased PageRank-style ranking algorithm *(based on the rank of the results that contained the entity)* on a directed multigraph containing as nodes the identified entities and their neighbors and as edges those produced considering the properties as bidirectional transitions
- The produced top-k semantic graphs:
  - Allow users to instantly inspect information that may lie in different places and that may be laborious and time consuming to locate
  - Provide useful information about the context of the identified entities
  - Allow the users to get a more sophisticated overview and to make better sense of the results
  - Make the LOD accessible to the end-users

Conclusion

- Survey for the marine domain:
  - The majority of participants:
    - *would like to see a graph representation of the top entities regardless the type of the submitted query*
    - *believe that the appearance of a graph of semantic information related to the search results can help them during an exploratory search process*
- Comparative evaluation of ranking schemes:
  - The proposed PageRank-based ranking scheme produces more preferred rankings compared to other link analysis-based algorithms
- Efficiency:
  - 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 KBs
  - We expect this limitation to get overcome in the near future
  - In the meanwhile, we can use caching / indexing / dedicated warehouses / distributed infrastructure

Future Research

- Interaction Model
  - *Integration in the search process*
- Visualization of the top-k semantic graphs
- Evaluation of other ranking schemes

Prototype:

http://139.91.183.72/x-ens-2

Thank you

Q | A