Personalized Concept-based Search & Exploration on the Web of Data using Results Categorization

Melike Sah and Vincent Wade

CNGL- Global Intelligent Content
Trinity College Dublin, Ireland
Outline

- Challenges & Motivation
- Personalized Concept-based Search
  - Walk through
  - Categorisation & Personalisation
- Evaluations
- Conclusions and Future Work
Web of Data (WoD)

Structured data

Relationships to other data

How to access relevant information?

Billions of triples
Challenges for Search over WoD

- Increasing challenge of finding and exploring relevant information on ever expanding WOD
- Support information queries & data gathering (~80% of Web queries)
- Today’s LoD search engines present results as ranked lists (e.g. Sindice, Watson)
- Need for more ‘Guided, Personalised support’ for users
  - Provide more scaffolded, informed navigation
  - Increase Result Precision
  - Increase User Satisfaction
Guided or Faceted search/browse interfaces is another approach for exploratory search

- Facet generation is **bound** to specific datatype/object properties

- Typically applied in closed domains since it requires high data completeness and consistent markup across the whole corpus

- Can suffer **Scalability & performance** issues (due to dynamic conjunctive clauses)

- Difficult to generate useful facets for **large and heterogeneous** data
Results clustering and personalized search are popular methods in Information Retrieval (IR).

**Results clustering:**
- Organize results into categories
- Useful for results exploration & disambiguation
- Some categories are widely used e.g. Google categories, Yahoo Directories, Open Directory Project

**Personalized search:**
- improve retrieval effectiveness by adapting results to context/interests of individual users
- Also personalized search and interactions can become more important as the size of LOD increases especially for dialogic interfaces
Proposal: Personalised Concept-based Search Mechanism for the WoD

Objective:
... to evaluate novel personalized search and exploration strategies for WoD:

Approach
- Develop Innovative combination of results categorization and personalized IR
- Use a novel fuzzy retrieval model for automatic categorization of LOD resources into UMBEL (Sah & Wade 2012)
- Group the results related to the same concepts to form concept lenses, to assist end user exploratory search and browsing
- Perform runtime personalisation based on User interactions (i.e. clicks on a concept lens or a search result) to adapt the results and results’ display to individual needs
Personalized Search with Concept Lenses

kilareny sightseeing
Personalized Search with Concept Lenses

Total of 305 results are found.

Result Categories:
- Town
- Hotel Organization
- Bed And Breakfast
- National Park
- Walking Generic
- Horseback Riding
- Lake
- Tourism
- Tourist Attraction
- Picture Image
- Golf
- Restaurant Organization
- British Cuisine
- Irish Food
- Hamburger Sandwich
- Mountain

Suppose we choose this lens
Personalised re-ranking of Concept lenses

Result re-ranking and Query expansion based on selection of lens

Tourist Attraction

<table>
<thead>
<tr>
<th>Horseback Riding</th>
<th>Tourism</th>
<th>Golf</th>
<th>Mountain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Walking Generic</td>
<td>National Park</td>
<td>Bar Place</td>
<td>Hotel Organization</td>
</tr>
<tr>
<td>Bed And Breakfast</td>
<td>Hamburger Sandwich</td>
<td>British Cuisine</td>
<td>Irish Food</td>
</tr>
<tr>
<td>Picture Image</td>
<td>Restaurant Organization</td>
<td>Italian Cuisine</td>
<td>Mexican Cuisine</td>
</tr>
<tr>
<td>Thai Cuisine</td>
<td>Continental Cuisine</td>
<td>Cafe Organization</td>
<td>Town</td>
</tr>
<tr>
<td>Lake</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Tourist Attraction

Things to do in Killarney – 60 Killarney Attractions - TripAdvisor
http://www.tripadvisor.ie/Attractions-g186612-Activities-Killarney_County_Kerry.html
★★★★★ 2,685 reviews and photos of 60 things to do in Killarney, Ireland...

More Results

Gap of Dunloe - Killarney - Reviews of Gap of Dunloe - TripAdvisor
http://www.tripadvisor.ie/Attraction_Review-g186612-d659775-Reviews-Gap_of_Dunloe-Killarney_County_Kerry.html
Gap of Dunloe, Killarney: See 315 reviews, articles, and 153 photos of Gap of Dunloe, ranked No.2 on TripAdvisor among 60 attractions in Killarney sim 0.5

Meeting of the Waters - Killarney - Reviews of Meeting of the Waters - TripAdvisor
http://www.tripadvisor.ie/Attraction_Review-g186612-d215897-Reviews-Meeting_of_the_Waters-Killarney_County_Kerry.html
Meeting of the Waters, Killarney: See 7 reviews, articles, and 3 photos of Meeting of the Waters, ranked No.28 on TripAdvisor among 60 attractions in Killarney sim 0.5

Killarney Guided Walks - Killarney - Reviews of Killarney Guided Walks - TripAdvisor
http://www.tripadvisor.ie/Attraction_Review-g186612-d1857777-Reviews-Killarney_Guided_Walks-Killarney_County_Kerry.html
Killarney Guided Walks, Killarney: See 38 reviews, articles, and 37 photos of Killarney Guided Walks, ranked No.2 on TripAdvisor among 27 attractions in Killarney sim 0.5

Killarney Riding Stables - Killarney - Reviews of Killarney Riding Stables - TripAdvisor
http://www.tripadvisor.ie/Attraction_Review-g186612-d3249993-Reviews-Killarney_Riding_Stables-Killarney_County_Kerry.html
Killarney Riding Stables, Killarney: See 8 reviews, articles, and 5 photos of Killarney Riding Stables, ranked No.12 on TripAdvisor among 60 attractions in Killarney sim 0.5

See Also

<table>
<thead>
<tr>
<th>Horseback Riding</th>
<th>Tourism</th>
<th>Golf</th>
<th>Mountain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Walking Generic</td>
<td>National Park</td>
<td>Bar Place</td>
<td>Hotel Organization</td>
</tr>
<tr>
<td>Bed And Breakfast</td>
<td>Hamburger Sandwich</td>
<td>British Cuisine</td>
<td>Irish Food</td>
</tr>
<tr>
<td>Picture Image</td>
<td>Restaurant Organization</td>
<td>Italian Cuisine</td>
<td>Mexican Cuisine</td>
</tr>
<tr>
<td>Thai Cuisine</td>
<td>Continental Cuisine</td>
<td>Cafe Organization</td>
<td>Town</td>
</tr>
<tr>
<td>Lake</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Query: Killarney sightseeing

Non adaptive concept lenses presentation

The user selects a concept lens and starts results exploration

Personalization of lenses and results based on user interaction

Personalized Re-Ranking of Concept Lenses

Result Categories:
- Town
- Hotel Organization
- Bed And Breakfast
- National Park
- Walking Generic
- Horseback Riding
- Lake
- Tourism
- Tourist Attraction

Tourist Attraction

**Things to do in Killarney – 60 Killarney Attractions - TripAdvisor**
http://www.tripadvisor.ie/Attractions-g186612-Activities-Killarney_County_Kerry.html
★★★★★ 2,685 reviews and photos of 60 things to do in Killarney, Ireland.... RANK:11

More Results

**Gap of Dunloe - Killarney - Reviews of Gap of Dunloe - TripAdvisor**
http://www.tripadvisor.ie/Attraction_Review-g186612-d659775-Reviews-Gap_of_Dunloe-Killarney_County_Kerry.html
Gap of Dunloe, Killarney: See 315 reviews, articles, and 153 photos of Gap of Dunloe, ranked No.2 on TripAdvisor among 60 attractions in Killarney.

**Meeting of the Waters - Killarney - Reviews of Meeting of the Waters - TripAdvisor**
http://www.tripadvisor.ie/Attraction_Review-g186612-d215897-Reviews-Meeting_of_the_Waters-Killarney_County_Kerry.html
Meeting of the Waters, Killarney: See 7 reviews, articles, and 3 photos of Meeting of the Waters, ranked No.28 on TripAdvisor among 60 attractions in Killarney.

**Killarney Guided Walks - Killarney - Reviews of Killarney Guided Walks - TripAdvisor**
http://www.tripadvisor.ie/Attraction_Review-g186612-d1857777-Reviews-Killarney_Guided_Walks-Killarney_County_Kerry.html
Killarney Guided Walks, Killarney: See 38 reviews, articles, and 37 photos of Killarney Guided Walks, ranked No.2 on TripAdvisor among 27 attractions in Killarney.

See Also
- Horseback Riding
- Tourism
- Golf
- Mountain
- Walking Generic
- National Park

Personalized Concept Lenses Suggestion
The Architecture and Workflow

WEB INTERFACE

Context-based user modelling; interactions + categorization

User clicks onto a result

Interactive personalization; completely implemented at client-side – requires no login & scalable performance

SERVER

Searching the Linked Open Data (using Sindice)

Set of Linked Open Data resources

Categorization exists?

YES

Categorization of resources into UMBEL concepts and supertype categories using fuzzy retrieval model

NO

Index of categorizations

Results with UMBEL concepts and supertypes

Concept lenses formation

Results re-ranking + expansion

Query + lens labels

Lenses suggestion

Store

Query re-ranking + expansion + lens labels

Query + lens labels

Query + lens expansion
Personalised Concept Based Search

- **Unique** personalized concept-based search
- Results categorization as a tool for personalized lenses/results re-ranking & query expansion.

Personalization occurs:
- When a user selects a concept lens; all lenses are re-organized, relevant results are included and relevant lenses are suggested
- When a user clicks onto a result; using interactive personalization. Last $N$ clicks of the users within a search session are used to add relevant results

- **Non-intrusive, privacy preserving and scalable,**
  - No user login required and personalization on client-side
- **Adaptable** and can be **plugable** into any LOD search engine
- Only requires UMBEL categorizations that is achieved by our fuzzy retrieval model
Categorization of LOD Resources (Results) using UMBEL

• What is UMBEL?
  • A broad conceptual vocabulary!
  • Simpler version of OpenCyc, which is manually developed over 20 years
  • Provides coherent hierarchy of 25,000 broad concepts

• Why UMBEL?
  • Coherent hierarchy of concepts
  • 32 top-level Supertype – good for organization of results
  • Alternative labels (synonyms, alternate names, etc.) – good for matching
  • Connected to linked datasets OpenCyc, DBpedia, schema.org, GeoNames – can be used for further results display, e.g. Google maps, DBpedia definition/links, etc.
How Categorization Works

• Extract **semantic information** from LOD resources and match to UMBEL descriptions using a fuzzy retrieval model.
  • label+type & label+type+subject provide the best results (see ESWC paper Sah & Wade 2012)
  • Categorization can be applied offline or dynamically. For scalability, new LOD resources are incrementally categorized and indexed

• A resource may match to any concept in 5-depth UMBEL hierarchy (~25,000 concepts), which is different from the general personalized search approaches
  • Generally top 1-2 levels are used but it can model only **general user interests**
  • In our approach, specific **UMBEL concept** matches can be used to understand user’s **specific interests** (e.g. Winter Sports, Skiing) and **Supertype concepts** can be used to understand user’s **general interests** (e.g. Sports)
Represent Results with UMBEL Concepts and Syntactic Information

- Each result is associated with a set of UMBEL concepts, their Supertypes and terms (textual content extracted from abstract and labels – stop words removed, terms are stemmed)
- We use vector space representation to represent this information
  - Allows scalable performance
  - Allows robust and efficient similarity measure for personalized lenses/results re-ranking;
  - Thus each result contains a vector of concepts, Supertypes and terms; for scalability we process the top $K$ results ($K = 100$)
- We use only hierarchical relationships (vector of Supertype) since UMBEL does not contain semantic relatedness relations between concepts
- We alleviate this issue to an extend by
  - Extracting data from subject of LOD resources, e.g. Both Pope and Vatican resources may share Christianity and Catholic subjects
  - To include semantically related subjects, we associate each LOD resource with 1-3 UMBEL concepts and their Supertype categories. Semantically related subjects share more concepts and Supertypes. Only the most confident categorization is used for concept lenses formation
Typical approaches to User Data Gathering

• **Sources:** Relevance feedback, implicit relevance feedback, desktop data, social Web or user’s context

• Relevance feedback requires time. Desktop data/social web data often contains enough information about general user interests
  - Relying on all past interests is trick. It requires identification of subset of interests and a timeline, e.g. a user is not interested in Florence hotels after booking a room
  - In long-term user profiling, usually server-side login & storage requires – **privacy issues.** In client-side storage, the user profile may be dislocated to multiple access devices

• In the context-based user modeling, only the **current available information** within the current search context is utilized (i.e. query, query context, clicked results, etc.).
  - Benefit: System only deals with few number of interests hence **performance is scalable**
  - Drawback: Past interests are lost but not all past interests are useful or identification of related interests can be challenging
Our Context-based User Modelling

- Only click data within the current search session is used
- The system copes with changes of search domain from categorization;
  - In a refined query, probable that similar concepts/supertypes will occur in new results
  - If the search topic changes completely, categorization in ~25,000 UMBEL concepts will not be the same. Fortunately, supertypes can be used to understand general interests
  - By analyzing the last N clicks on concepts/supertype concepts and the system can find similar LOD resources that share related concepts
- We represent user’s information need using; vector of concepts (specific interests), vector of Supertypes (broad interests) & vector of terms (term interests)
- For user profiling, we track clicks onto: (i) concept lenses – personalization based on user’s local choices & (ii) last N results – interactive personalization
  - User’s interest to concept lenses; all results under a lens combined to create vectors of concepts, Supertypes & terms. We use frequency as weight to compute the shared information between lenses
    $$\sum_{z=1}^{m} l_z = \sum_{i=1}^{n} \sum_{j=1}^{k} r_i c_j \rightarrow Vc(l_z) = (w(c_1, l_z), w(c_2, l_z), ..., w(c_t, l_z))$$
  - For interactive personalization, we capture the user’s information need as vectors of concepts, Supertypes and terms from the last N clicks (similar as concept lenses)
Re-Organization of Concept Lenses

• Personalisation (reordering) of concept lenses based on the user’s local choices. Thus, conceptually relevant concept lenses move to the top of the list

• Compare similarity of the selected lens to other concept lenses using the cosine similarity of concept, Supertype and term vectors:

\[ \text{sim}(l_1, l_2) = \frac{\vec{V}_1(l_1) \cdot \vec{V}_2(l_2)}{|\vec{V}_1(l_1)| \cdot |\vec{V}_2(l_2)|} \]

• Concept similarity: If lenses share specific concepts, it is more likely they are relevant

• Supertype similarity: Computes shared broad concepts, e.g. “mountain” and “lake”

• Term similarity: Syntactic similarity. Can be noisy but guarantee some level of similarity
Evaluations showed that the concept vector similarity alone provided the best precision. The best results were obtained when all similarity measures were combined (combined semantic and syntactic similarity).

Lenses are displayed in decreasing relevance order.
Results Re-Ranking and Concept Lenses Suggestion

- We apply results re-ranking in two cases:
  (a) when the user selects a concept lens from the results list for exploration
  (b) when the user clicks onto a result (LOD resource) within a concept lens
- In both cases, re-ranked results are included within the interacted concept lens. This allows in-context exploration of more results.
- In case (a), we compare concept vector of the selected concept lens with the top $K$ results ($K=100$):
  - We compare concept vectors since results matching at specific concepts are more likely to be relevant compared to supertype or term similarities (we only have user’s interest for a concept lens). Results are added in decreasing order, a threshold can be used.
- In case (b), we use the click history of the user. User’s specific concept, supertype and term interests (from the last $M$ results clicks) are compared with top $K$ result vectors ($K=100$):

$$\frac{\sum_{i=1}^{K} \alpha \cdot \text{sim}(Vc(r_i), Vc(u)) + \beta \cdot \text{sim}(Vsc(r_i), Vsc(u)) + \delta \cdot \text{sim}(Vt(r_i), Vt(u))}{\alpha + \beta + \delta}$$

- If a relevant result belongs to another concept, then the concept lens is suggested.
Query Refinement

• Query adaptation is applied in two cases:
  • When the user selects a concept lens from the results list for exploration
  • When the last two consecutive result clicks share the same concept

• We expand the original query with the concept label the user interested

• It is a simple approach, but works well since UMBEL categorizations provide specific concept names to clarify the meaning of the query with the user feedback

• In both cases, more results are included in the context of the interacted concept lens, so that the user can explore more relevant results in context
Evaluations

• In traditional IR, there are public benchmarks (e.g. TREC) but no standard evaluation benchmarks for semantic search evaluations

• Created a benchmark dataset using LOD resources, which is available for validation and comparison, https://www.scss.tcd.ie/melike.sah/tourismdataset.zip

• We measured personalized search efficacy using precision @top M concept lenses and @top N results

  • We focused on precision for top M concept lenses & N results to measure improved precision for top results/
For evaluation we chose tourism domain, which suits well for data gathering and results exploration

Indexed a particular dataset for stable and comparative evaluations

- Popular search queries were investigated from Google Trends and searched over the WoD using Sindice to gather URIs
- ~500 URIs from DBpedia, GeoNames, Trip Advisor and ookaboo
- RDF descriptions, their UMBEL/supertype categorizations were indexed offline by the proposed fuzzy retrieval model to carry the experiments
- We selected 20 queries, (no navigational queries). Such small size queries have been used before to determine indicative results in semantic search
- Top concept lenses and results returned by the queries, were manually assigned relevant or irrelevant based on search intent. The same dataset was used for non-adaptive baseline system
Example Queries from the Dataset

- killarney sightseeing
- killarney waterfall
- killarney park things to do
- killarney lake
- killarney mountain
- killarney outdoor activities
- killarney island
- killarney irish food
- killarney bar pub
- ring of kerry

...  e.g. Search Intent is finding tourist attractions in the area
• Most LoD result categorization happens apriori. Thus, we computed dynamic average time required to generate personalized results, i.e. lenses re-organization, results re-ranking and query expansion following a lens selection.

• For each query, average of 5 runs used. Experiments run on Windows 7 computer, 2.2GHz CPU and 7.90GB RAM.
Lens Re-Ranking: Precision @top M Concepts

Precision @top M concepts over all queries for various lenses re-ranking similarity measures

- Concept similarity
- Supertype similarity
- Term similarity
- Combined concept and supertype similarities
- Combined concept and term similarities
- Combined concept, supertype and term similarities
Result Personalisation: Precision @top N

Precision @top N results over all queries for various personalization strategies:

- **lense re-ranking**
- **combined lense re-ranking and results re-ranking**
- **combined lense re-ranking and query expansion**
- **combined lense re-ranking, results re-ranking and query expansion**

Graph showing Precision at various top N values (P@3, P@5, P@7, P@10, P@15, P@20) for different personalization strategies.
Average vs Min order
Precision @top M Concepts

Precision @top M concepts over all queries for non adaptive lenses order - average vs min order

- min rank lense order
- average rank lense order

Precision

- P@1
- P@3
- P@5
- P@7
- P@10
- P@15
Comparison of Non Personalised vs Personalised Concept Re-ranking Precision @top M Concepts

Precision @top M concepts over all queries; personalized vs non adaptive lenses order (min)

- non-adaptive (min order)
- personalized

Precision

P@1  P@3  P@5  P@7  P@10  P@15
Comparison of Personalised vs Non-Personalised Result Reranking Precision over all queries Precision @ top N Results
Conclusions and Future Work

Indicative results

• Novel concept-based personalized search and exploration mechanism for the Web of Data
  • Results categorization as the basis of personalized concept lenses re-ordering, results re-ranking and query expansion
  • Our personalization approach is non-intrusive, privacy preserving and scalable
  • Proposed approach is adaptable and can be plugged on top of any LOD search engine

• Evaluation of Personalized concept-based search outperformed ranked list and non-adaptive concept-based search but further, larger performance & user evaluation trials are required

Future Work

• Perform scaled user studies to evaluate usability of our approach
• Data quality, trust and graph popularity can be considered in rankings
Thank You, Questions?

Contact:
Vincent.Wade@TCD.IE
SupplementarySlides
Categorization of LOD Resources (Results) using UMBEL

• What is UMBEL?
  • A broad conceptual vocabulary!
  • Simpler version of OpenCyc, which is manually developed over 20 years
  • Provides coherent hierarchy of 25,000 broad concepts

• Why UMBEL?
  • Coherent hierarchy of concepts
  • 32 top-level Supertype – good for organization of results
  • Alternative labels (synonyms, alternate names, etc.) – good for matching
  • Connected to linked datasets OpenCyc, DBpedia, schema.org, GeoNames – can be used for further results display, e.g. Google maps, DBpedia definition/links, etc.
Typical approaches to User Data Gathering

- **Sources**: Relevance feedback, implicit relevance feedback, desktop data, social Web or user’s context

- Relevance feedback requires time. Desktop data/social web data often contains enough information about general user interests
  - Relying on all past interests is trick. It requires identification of subset of interests and a timeline, e.g. a user is not interested in Florence hotels after booking a room
  - In long-term user profiling, usually server-side login & storage requires – privacy issues. In client-side storage, the user profile may be dislocated to multiple access devices

- In the context-based user modeling, only the **current available information** within the current search context is utilized (i.e. query, query context, clicked results, etc.).
  - Benefit: System only deals with few number of interests hence performance is scalable
  - Drawback: Past interests are lost but not all past interests are useful or identification of related interests can be challenging
Re-Organization of Concept Lenses

- Personalisation (reordering) of concept lenses based on the user’s local choices. Thus, conceptually relevant concept lenses move to the top of the list.
- Compare similarity of the selected lens to other concept lenses using the cosine similarity of concept, Supertype and term vectors:

\[
sim(l_1, l_2) = \frac{\vec{V}_1(l_1) \cdot \vec{V}_2(l_2)}{|\vec{V}_1(l_1)| |\vec{V}_2(l_2)|}
\]

- Concept similarity: If lenses share specific concepts, it is more likely they are relevant.
- Supertype similarity: Computes shared broad concepts, e.g. “mountain” and “lake”.
- Term similarity: Syntactic similarity. Can be noisy but guarantee some level of similarity.
• Evaluations showed that the concept vector similarity alone provided the best precision. The best results were obtained when all similarity measures were combined (combined semantic and syntactic similarity).

• Lenses are displayed in decreasing relevance order.
Query Refinement

- Query adaptation is applied in two cases:
  - When the user selects a concept lens from the results list for exploration
  - When the last two consecutive result clicks share the same concept

- We expand the original query with the concept label the user interested

- It is a simple approach, but works well since UMBEL categorizations provide specific concept names to clarify the meaning of the query with the user feedback

- In both cases, more results are included in the context of the interacted concept lens, so that the user can explore more relevant results in context
### Example Queries from the Dataset

<table>
<thead>
<tr>
<th>Killarney Sightseeing</th>
<th>e.g. Search Intent is finding tourist attractions in the area</th>
</tr>
</thead>
<tbody>
<tr>
<td>Killarney Waterfall</td>
<td></td>
</tr>
<tr>
<td>Killarney Park Things to Do</td>
<td></td>
</tr>
<tr>
<td>Killarney Lake</td>
<td></td>
</tr>
<tr>
<td>Killarney Mountain</td>
<td></td>
</tr>
<tr>
<td>Killarney Outdoor Activities</td>
<td></td>
</tr>
<tr>
<td>Killarney Island</td>
<td></td>
</tr>
<tr>
<td>Killarney Irish Food</td>
<td></td>
</tr>
<tr>
<td>Killarney Bar Pub</td>
<td></td>
</tr>
<tr>
<td>Ring of Kerry</td>
<td></td>
</tr>
</tbody>
</table>

......
Precision @top M concepts over all queries for non adaptive lenses order - average vs min order

- min rank lense order
- average rank lense order

Precision

P@1  P@3  P@5  P@7  P@10  P@15