Entity and Knowledge Base-oriented Information Retrieval

Presenter: Liuqing Li
liuqing@vt.edu

Digital Library Research Laboratory
Virginia Polytechnic Institute and State University
Blacksburg, VA 24061
Nov 08, 2018
Outline

• Paper List
• Background
• Approach
• Evaluation
Outline

- Paper List
- Background
- Approach
- Evaluation
**Paper List**

- State-of-the-art
- Major groups and members
  - James Allan’s group (UMA), Jeffrey Dalton
  - Jamie Callan’s group (CMU), Chenyan Xiong

<table>
<thead>
<tr>
<th>Paper</th>
<th>Author</th>
<th>Conference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entity query feature expansion using knowledge base links</td>
<td>Jeffrey Dalton, Laura Dietz, James Allan</td>
<td>SIGIR’14</td>
</tr>
<tr>
<td>Esdrank: Connecting query and documents through external semi-structured data</td>
<td>Chenyan Xiong, Jamie Callan</td>
<td>CIKM’15</td>
</tr>
<tr>
<td>Query expansion with Freebase</td>
<td>Chenyan Xiong, Jamie Callan</td>
<td>ICTIR’15</td>
</tr>
<tr>
<td>Bag-of-Entities representation for ranking</td>
<td>Chenyan Xiong, Jamie Callan, Tie-Yan Liu</td>
<td>ICTIR’16</td>
</tr>
<tr>
<td>Paper</td>
<td>Author</td>
<td>Conference</td>
</tr>
<tr>
<td>-------</td>
<td>--------</td>
<td>------------</td>
</tr>
<tr>
<td>[5] An empirical study of learning to rank for entity search</td>
<td>Jing Chen, Chenyan Xiong, Jamie Callan</td>
<td>SIGIR’16</td>
</tr>
<tr>
<td>[7] Explicit Semantic Ranking for Academic Search via Knowledge Graph Embedding</td>
<td>Chenyan Xiong, Russell Power, Jamie Callan</td>
<td>WWW’17</td>
</tr>
<tr>
<td>[8] Word-entity duet representations for document ranking</td>
<td>Chenyan Xiong, Jamie Callan, Tie-Yan Liu</td>
<td>SIGIR’17</td>
</tr>
<tr>
<td>[10] End-to-end neural ad-hoc ranking with kernel pooling</td>
<td>Chenyan Xiong, et al.</td>
<td>SIGIR’17</td>
</tr>
</tbody>
</table>
Outline

- Paper List
- Background
- Approach
- Evaluation
## Background

<table>
<thead>
<tr>
<th>Information Retrieval</th>
<th>Document Retrieval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pseudo Relevance Feedback &amp; Query Expansion (PRF &amp; QE)</td>
<td></td>
</tr>
<tr>
<td>Learning to Rank (LeToR)</td>
<td></td>
</tr>
</tbody>
</table>

| Entity-related                      |                                 |
|-------------------------------------|                                 |
| Entity Retrieval                   |                                 |
| Knowledge Base                     |                                 |
| Entity Linking Systems             |                                 |

**Entity and Knowledge Base-oriented Information Retrieval**
Background – Document Retrieval

1. User Interface
2. User need
3. Query
4. Featured query
5. Operations on textual data
6. Ranking Documents
7. Document Files
8. Document Preprocessing and Indexing
• Inverted Indexing

<table>
<thead>
<tr>
<th>Words</th>
<th>Document</th>
<th>Inverted Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>67</td>
<td>0</td>
<td>{(0,3)}</td>
</tr>
<tr>
<td>diabet</td>
<td>1,2</td>
<td>{(1,4),(2,3),(2,7)}</td>
</tr>
<tr>
<td>female</td>
<td>0</td>
<td>{(0,6)}</td>
</tr>
<tr>
<td>hyperten</td>
<td>0,1</td>
<td>{(0,8),(1,6)}</td>
</tr>
<tr>
<td>mother</td>
<td>2</td>
<td>{(2,1)}</td>
</tr>
<tr>
<td>niddm</td>
<td>0</td>
<td>{(0,8)}</td>
</tr>
<tr>
<td>no</td>
<td>1</td>
<td>{(1,3)}</td>
</tr>
<tr>
<td>old</td>
<td>0</td>
<td>{(0,5)}</td>
</tr>
<tr>
<td>patient</td>
<td>0,1,2</td>
<td>{(0,0),(1,1), (2,0),(2,4)}</td>
</tr>
<tr>
<td>sister</td>
<td>2</td>
<td>{(2,5)}</td>
</tr>
<tr>
<td>year</td>
<td>0</td>
<td>{(0,4)}</td>
</tr>
</tbody>
</table>
Background – PRF & QE

Initial Query → Information Retrieval System → Returned Document

Reformulated Query → Sampling (1) → Pertinent Document

Rewriting of the Query (3) → Pertinent Terms → Extraction of Evidences (2)

Pertinent Terms
Background — Learning to Rank

- Feature Extractor
- Labeler
- Learning Algorithm
- Ranking System

Training Corpus:
- \((q_1, d_1)\)
- \((q_1, d_2)\)
- \(\vdots\)
- \((q_{|Q|}, d_{|D|})\)

Test Corpus:
- \((q, d_1)\)
- \((q, d_2)\)
- \(\vdots\)
- \((q, d_{|D|})\)

\(f(q, d_1)\)
\(f(q, d_2)\)
\(\vdots\)
\(f(q, d_{|D|})\)

- Vectorized instances with judgments
- Ranking function: \(f\)
• Features are IMPORTANT!
  • Query-independent features
    • Depend only on documents (e.g., PageRank or doc length)
  • Query-dependent features
    • Depend both on document content and query
  • Query level features
    • Depend only on query (e.g., number of words in a query)
Background — Entity Retrieval

• Goal
  • Retrieve entities (e.g., Wikipedia pages) with a couple of queries

• Previous tasks
  • INEX Entity Ranking Track (2009)
  • INEX Linked Data Track (2012)

• Document retrieval with entity annotations
Background — Knowledge Base
Background — Knowledge Base

• Freebase Sub-graph

![Freebase Sub-graph Diagram]

Table 1: Freebase facts used.

<table>
<thead>
<tr>
<th>Fact</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>mld</td>
<td>Unique Id</td>
</tr>
<tr>
<td>key:en</td>
<td>English Key (if any)</td>
</tr>
<tr>
<td>rdfs:label</td>
<td>Object’s Name</td>
</tr>
<tr>
<td>/common/topic/description</td>
<td>Text Description</td>
</tr>
<tr>
<td>rdf:type</td>
<td>Category (if any)</td>
</tr>
</tbody>
</table>

Figure 1: A sub-graph of Freebase.

Background – Entity Linking Systems

• FACC1 (Google)
  • Freebase Annotations of the ClueWeb Corpora
  • 11 billion entity annotations in 800 million documents

• TagMe (A3 Lab)
  • Identify short-phrases and link them to Wikipedia
  • Demo: https://tagme.d4science.org/tagme/
Background – Entity Linking Systems

• CMNS (Hasibi et al., Norwegian University)
  • Entity Linking in Queries: Tasks and Evaluation
  • ICTIR’15

{France, FIFA world cup}

or

{France national football team, FIFA world cup}
Background – Entity Linking Systems

- CMNS (Hasibi et al., Norwegian University)
  - Entity Linking in Queries: Tasks and Evaluation
  - ICTIR’15

Semantic mapping

query → Mention detection → set of mentions → Candidate entity ranking → ranked list of entities → Interpretation finding → interpretations

Interpretation Finding
The goal of entity linking in queries
Background – Entity and KB-oriented IR

- Workflow
Outline

• Paper List
• Background
• Approach
• Evaluation
Approach

• Paper
  • Entity query feature expansion using knowledge base links. SIGIR’14.

• Strengths
  • Most previous work focuses on text, mainly using unigram concepts
  • This paper enriches queries with features from entities and their links to knowledge bases

• Knowledge Base (Wikipedia, Freebase)
• Entity Annotation (FACC1)
Approach

• Entity Query Feature Expansion (EQFE)
  • Focuses on words, entities, types, categories
  • Entity linking the query provides very precise indicators but may also miss many of the relevant entities
  • Entity expansion in PRF may make a query noisy

```c
#combine(
    #sdm(   obama family tree  )
    #sdm(   [Barack_Obama] [Family_Tree]  )
    #sdm(   {US President} {Politician})
    #sdm(   [Michelle_Obama] [Ireland] [Kenya])
)
```

Approach

- **EQFE**

---

\[ f_{\text{ExplWiki}}(Q, W) = \sum_{M} \left( \sum_{E} p(W|E)p(E|M) \right) p(M|Q) \]

\[ f_{\text{KB}}(Q, E) = \frac{1}{Z} \exp s_Q(E) \]

\[ f_{\text{RM}}(Q, E) = \sum_{D} \left( \sum_{M} p(E|M)p(M|D) \right) p(D|Q) \]

Approach

• Paper
  • Esdrank: Connecting query and documents through external semi-structured data. *CIKM’15.*

• Strengths
  • Treats the vocabularies, terms and entities from external data as objects
  • Treats the external objects as latent layer between query and documents

• Knowledge Base (Freebase)
• Entity Annotation (TagMe, FACC1)
• EsdRank
  • Latent-ListMLE, a LeToR algorithm, models the objects as latent space between query and documents

\[
U = \{u_{11}, ..., u_{ij}, ..., u_{nm}\} \quad V = \{v_1, ..., v_j, ..., v_m\}
\]

• How to find related objects?
  • Query Annotation, Object Search, Document Annotation

Approach

- EsdRank
  - Selected Features

<table>
<thead>
<tr>
<th>Query - Objects</th>
<th>Objects - Documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Object Selection Score</td>
<td>Textual Similarity</td>
</tr>
<tr>
<td>Textual Similarity</td>
<td>Ontology Overlap</td>
</tr>
<tr>
<td>Ontology Overlap</td>
<td>Graph Connection</td>
</tr>
<tr>
<td>Object Frequency</td>
<td>Document Quality</td>
</tr>
<tr>
<td>Similarity with Other Objects</td>
<td></td>
</tr>
</tbody>
</table>

Approach

• Paper
  • Query expansion with Freebase. *ICTIR’15.*

• Strengths
  • Two methods of identifying the entities associated with a query
  • Two methods of using those entities to perform query expansion

• Knowledge Base (Freebase)
• Entity Annotation (FACC1)
Approach

- Unsupervised Expansion
  - Linking Freebase Objects to the Query
    - Link by Search
      - Query q is issued to the Google Search API to get its ranking of objects O with ranking scores
    - Link by FACC1
      - Select related objects from the FACC1 annotations in top retrieved documents
  - How to rank objects?

\[ r_f(o_k) = \sum_{d_j \in D} tf(d_j, o_k) \log \frac{|F|}{df(o_k)} \]

Approach

• Unsupervised Expansion
  • Selecting Expansion Terms from Linked Objects
    • Select by PRF
      • Tf-idf based PRF on linked objects’ descriptions
    • Select by Category
      • Naïve Bayesian Classifier
      • Probability of a term belonging to a category

<table>
<thead>
<tr>
<th></th>
<th>Link by Search</th>
<th>Link by FACC1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Select by PRF</td>
<td>FbSearchPRF</td>
<td>FbFaccPRF</td>
</tr>
<tr>
<td>Select by Category</td>
<td>FbSearchCat</td>
<td>FbFaccCat</td>
</tr>
</tbody>
</table>

Approach

• Paper
  • Bag-of-Entities representation for ranking. *ICTIR’16*.
• Strengths
  • Evaluate three different entity linking systems
• Knowledge Base (Freebase, Wikipedia)
• Entity Annotation (FACC1, TagMe, CMNS)
Approach

• Bag-of-Entities Representation
  • Entity annotation (FACC1, TagMe, CMNS)
  • Use a bag-of-entities to represent a query or document
  • Apply two basic ranking models
    • Coordinate Match (COOR)
      \[ f_{COOR}(q, d) = \sum_{e: \tilde{E}_q(e) > 0} 1(\tilde{E}_d(e) > 0) \]  
    • Entity Frequency (EF)
      \[ f_{EF}(q, d) = \sum_{e: \tilde{E}_q(e) > 0} \tilde{E}_q(e) \log(\tilde{E}_d(e)) \]

Approach

• Bag-of-Entities Representation
  • Pre-retrieved documents re-ranking
  • Study different entity annotation methods
  • Compare bag-of-words and bag-of-entities methods

• Paper
  • An empirical study of learning to rank for entity search. *SIGIR’16.*

• Strengths
  • Learning to rank methods are as powerful for ranking entities as for ranking documents

• Knowledge Base (DBpedia with RDF triples)

• Entity Annotation (None)
Approach

• LeToR for Entity Search
  • How to represent entities?
    • Name, Cat, Attr, RelEn, SimEn
  • Features for query-entity pairs
    • Language model
    • BM25
    • Coordinate match
    • Cosine similarity
    • Sequential dependency model (SDM)
    • Fielded sequential dependency model (FSDM)

Approach

• Paper

• Strengths
  • Combines query entity linking and entity-based document ranking
  • A joint learning-to-rank model

• Knowledge Base (Wikipedia)
• Entity Annotation (FACC1)
Approach

• Joint Semantic Ranking
  • Spotting, Linking and Ranking
    • Surface forms and candidate entities
    • Feature Generation

<table>
<thead>
<tr>
<th>Surface Form Features ($\phi_s$)</th>
<th>Entity Features ($\phi_e$)</th>
<th>Entity-Document Ranking Features ($\phi_r$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Linked Probability</td>
<td>(1) Commonness</td>
<td>(16) BM25, Coordinate Match, TFIDF and</td>
</tr>
<tr>
<td>(1) Surface Form Entropy</td>
<td>(2) Max and Mean Similarity</td>
<td>language model with Dirichlet smoothing</td>
</tr>
<tr>
<td>(1) Top Candidate Entities Margin</td>
<td>with Query Words</td>
<td>from entity’s textual fields (name and</td>
</tr>
<tr>
<td>(1) Surface Form Length and Coverage</td>
<td>(2) Max and Mean Similarity</td>
<td>description) to document’s fields (title</td>
</tr>
<tr>
<td>(1) Surface Form Coverage</td>
<td>with Other Query Entities</td>
<td>and body)</td>
</tr>
</tbody>
</table>

Table 1: Spotting, linking, and ranking features. Surface Form Features are extracted for each spotted surface form. Entity Features are extracted for each candidate entity from each spot. Entity-Document Ranking Features are extracted for each entity-document pair. The number in brackets is the dimension of the corresponding feature group.

Approach

- Joint Semantic Ranking
  
  - Joint Learning to Rank

  \[
  f_s(s_i) = w_s^T \phi_s(s_i). \quad \text{learns the importance of the surface form}
  \]

  \[
  f_e(e_j^i) = w_e^T \phi_e(e_j^i). \quad \text{learns the importance of the aligned entity}
  \]

  \[
  f_r(e_j^i, d) = w_r^T \phi_r(e_j^i, d). \quad \text{learns the ranking of document for the entity}
  \]

  \[
  f(q, d|\theta, S, E) = \sum_{i=1}^{M} \sum_{j=1}^{N} f_s(s_i) \cdot f_e(e_j^i) \cdot f_r(e_j^i, d)
  \]

  \[
  \theta^* = \arg\min_{\theta} \sum_{q} \sum_{d^+, d^- \in D_q^{+, -}} [1 - f(q, d^+|S, E) + f(q, d^-|S, E)]_+
  \]

Approach

• Paper
  • Explicit Semantic Ranking for Academic Search via Knowledge Graph Embedding. *WWW’17.*

• Strengths
  • Represents queries and documents in the entity space and ranks them based on their semantic connections from their knowledge graph embedding

• Knowledge Base (Freebase)
• Entity Annotation (FACC1, CMNS)
• Explicit Semantic Ranking

Each element in the matrix
\[ s(e_i, e_j) = \cos(V(e_i), V(e_j)) \]
max-pooling along the query dimension
\[ \tilde{S}(d) = \max_{e_i \in E_q} s(e_i, \vec{E}_d) \]
bin-pooling (histogram)
\[ B_k(q, d) = \log \sum_j I(st_k \leq \tilde{S}_j(d) < ed_k) \]

• **Paper**
  - Word-entity duet representations for document ranking. *SIGIR’17.*

• **Strengths**
  - Creates a word-entity duet framework
  - Creates an attention-based ranking model

• **Knowledge Base (Freebase)**

• **Entity Annotation (TagMe)**
Approach

• Word-Entity Duet Representation
  • Words and entities in both query and documents

Table 1: Ranking features from query words to document words (title and body) ($\Phi_{Qw-Dw}$).

<table>
<thead>
<tr>
<th>Feature Description</th>
<th>Dimension</th>
</tr>
</thead>
<tbody>
<tr>
<td>BM25</td>
<td>2</td>
</tr>
<tr>
<td>TF-IDF</td>
<td>2</td>
</tr>
<tr>
<td>Boolean OR</td>
<td>2</td>
</tr>
<tr>
<td>Boolean And</td>
<td>2</td>
</tr>
<tr>
<td>Coordinate Match</td>
<td>2</td>
</tr>
<tr>
<td>Language Model (Lm)</td>
<td>2</td>
</tr>
<tr>
<td>Lm with JM smoothing</td>
<td>2</td>
</tr>
<tr>
<td>Lm with Dirichlet smoothing</td>
<td>2</td>
</tr>
<tr>
<td>Lm with two-way smoothing</td>
<td>2</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>18</strong></td>
</tr>
</tbody>
</table>

• Word-Entity Duet Representation
  • Words and entities in both query and documents

Table 2: Ranking features from query entities (name and description) to document words (title and body) ($\Phi_{Qe-Dw}$).

<table>
<thead>
<tr>
<th>Feature Description</th>
<th>Dimension</th>
</tr>
</thead>
<tbody>
<tr>
<td>BM25</td>
<td>4</td>
</tr>
<tr>
<td>TF-IDF</td>
<td>4</td>
</tr>
<tr>
<td>Boolean Or</td>
<td>4</td>
</tr>
<tr>
<td>Boolean And</td>
<td>4</td>
</tr>
<tr>
<td>Coordinate Match</td>
<td>4</td>
</tr>
<tr>
<td>Lm with Dirichlet Smoothing</td>
<td>4</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>24</strong></td>
</tr>
</tbody>
</table>

Approach

• Word-_entity Duet Representation
  • Words and entities in both query and documents

Table 3: Ranking features from query words to document entities (name and description) ($\Phi_{Qw-De}$).

<table>
<thead>
<tr>
<th>Feature Description</th>
<th>Dimension</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top 3 Coordinate Match on Title Entities</td>
<td>6</td>
</tr>
<tr>
<td>Top 5 Coordinate Match on Body Entities</td>
<td>10</td>
</tr>
<tr>
<td>Top 3 TF-IDF on Title Entities</td>
<td>6</td>
</tr>
<tr>
<td>Top 5 TF-IDF on Body Entities</td>
<td>10</td>
</tr>
<tr>
<td>Top 3 Lm-Dirichlet on Title Entities</td>
<td>6</td>
</tr>
<tr>
<td>Top 5 Lm-Dirichlet on Body Entities</td>
<td>10</td>
</tr>
<tr>
<td>Total</td>
<td>48</td>
</tr>
</tbody>
</table>

Approach

• Word-Entity Duet Representation
  • Words and entities in both query and documents

Table 4: Ranking features from query entities to document’s title and body entities ($\Phi_{Qe-De}$).

<table>
<thead>
<tr>
<th>Feature Description</th>
<th>Dimension</th>
</tr>
</thead>
<tbody>
<tr>
<td>Binned translation scores, 1 exact match bin, 5 soft match Bins in the range [0, 1).</td>
<td>12</td>
</tr>
</tbody>
</table>

• Word-Entity Duet Representation
  • Attention-based ranking model

\[ F_w(w_i) = W_w^m \cdot R_w(w_i, \cdot) + b_w^m \]
\[ F_e(e_j) = W_e^m \cdot R_e(e_j, \cdot) + b_e^m. \]

\[ f(q, d) = F_w \cdot \alpha_w + F_e \cdot \alpha_e. \]

\[ l(q, D) = \sum_{d \in D^+} \sum_{d' \in D^-} [1 - f(q, d) + f(q, d')]_+. \]

Approach

- Paper
  - End-to-end neural ad-hoc ranking with kernel pooling. *SIGIR’17.*

- Strengths
  - Kernel pooling
  - CNN
  - Could be entity or knowledge base-based
Approach

• Kernel Pooling

Approach

• Kernel Pooling

\[ f(q, d) = \tanh(w^T \phi(M) + b) \]

Learning to Rank

\[ \phi(M) = \sum_{i=1}^{n} \log \tilde{K}(M_i) \]

Soft-TF Features

\[ \tilde{K}(M_i) = \{K_1(M_i), ..., K_K(M_i)\} \]

Kernel Pooling

\[ K_k(M_i) = \sum_j \exp(-\frac{(M_{ij} - \mu_k)^2}{2\sigma_k^2}) \]

RBF Kernel

\[ M_{ij} = \cos(\vec{v}_{t_i}, \vec{v}_{t_j}) \]

Translation Matrix

\[ t \Rightarrow \vec{v}_t \]

Word Embedding

Approach

- Kernel Pooling

\[
g(M_{ij}) = \sum_{k=1}^{K} \frac{g(K_k(M_i)) \times \sigma_k^2}{(\mu_k - M_{ij}) \exp\left(\frac{(M_{ij} - \mu_k)^2}{-2\sigma_k^2}\right)}
\]

Approach

- N-grams

Outline

- Paper List
- Background
- Approach
- Evaluation
Evaluation

• Datasets
  • ClueWeb09
    • [https://lemurproject.org/clueweb09/](https://lemurproject.org/clueweb09/)
    • Entire Dataset: 4,780,950,903 Unique URLs
    • ClueWeb09-B: first 50 million English pages (Popular)
  • ClueWeb12
    • [https://lemurproject.org/clueweb12/](https://lemurproject.org/clueweb12/)
    • Entire Dataset: 733 million pages
    • ClueWeb12-B13: about 50 million pages (Popular)
Evaluation

• Queries
  • TREC 2009-2012: 200 queries (ClueWeb09-B)
  • TREC 2013-2014: 100 queries (ClueWeb12-B13)
  • Ad hoc task and diversity task

```xml
<topic number="2" type="faceted">
  <query>french lick resort and casino</query>
  <description>
    Find information on French Lick Resort and Casino in Indiana.
  </description>
  <subtopic number="1" type="nav">
    Find the homepage for French Lick Resort and Casino.
  </subtopic>
  <subtopic number="2" type="inf">
    What casinos are located within a day's drive of French Lick Resort and Casino?
  </subtopic>
  <subtopic number="3" type="inf">
    What jobs are available at French Lick Casino and Resort?
  </subtopic>
  <subtopic number="4" type="inf">
    Are there discounted packages for staying at French Lick Resort and Casino?
  </subtopic>
</topic>
```
Evaluation

• Metrics
  • TREC Official: NDCG@20 and ERR@20
  • More: MAP@K, P@K
  • Re-rank: Win / Tie / Loss

\[
DCG_n = \sum_{i=1}^{n} \frac{rel_i}{\log_{2+1}(i)}, \quad NDCG_n = \frac{DCG_n}{IDCG_n},
\]

\[
AveP = \frac{\sum_{k=1}^{n} (P(k) \times rel(k))}{\text{number of relevant documents}}, \quad MAP = \frac{\sum_{q=1}^{Q} AveP(q)}{Q}
\]
Questions?

Liuqing Li

3rd year PhD student in Digital Library Research Laboratory
GTA in CS4984 / CS5984 (Big Data Text Summarization)
Email: liuqing@vt.edu
Website: http://liuqing.dlib.vt.edu/main