The Cluster Hypothesis: Ranking Document Clusters

Oren Kurland
Faculty of Industrial Engineering and Management
Technion

* Based on joint work with Fiana Raiber
* This work has been supported by, and carried out, at the Technion-Microsoft Electronic Commerce Research Center
Is search a solved problem?

Query = “oren kurland dblp”

Search #1

**DBLP: Michael Bendersky**

Eyal Kikon, Oren Kurland, Michael Bendersky: Utilizing inter-passage and inter-graph similarity update Fri Feb 17 20:20:27 2012 CET by the DBLP Team — Data released under the terms of the Creative Commons Attribution License - CC BY 3.0 DE

**DBLP: David Carmel**

Oren Kurland, Anna Shtok, David Carmel, Shay Hummel: A Unified Framework for Document Retrieval and Recommendation update Mon Feb 20 23:32:27 2012 CET by the DBLP Team — Data released under the terms of the Creative Commons Attribution License - CC BY 3.0 DE

**DBLP: Geoff Hulten**

from DBLP and Google Scholar. Related People Matthew Richardson; AnHai Do; Alon Y. Halevy
dblife.cs.wisc.edu/person/Geoff_Hulten

**Language Model Information Retrieval with Document Expansion**

ametminer.org/viewpub.do?pid=511245

**29. SIGIR 2006: Seattle, WA, USA - WHAT's NEW! — ACM SIGIR**


www.sigmod.org/sigmod/dblp/db/conf/sigir/sigir2006.html

**Relevance-Based Language Models - ArnetMiner - Academic ...**

Authors: Inna Gelfer Kalmanovich Oren Kurland Organization: SIGIR Developing filtering systems.

www.arnetminer.org/viewpub.do?pid=595277

Search #2

**DBLP: Oren Kurland**

www.informatik.uni-trier.de/~ley/db/indices/a-tree/k/Kurland-Oren.html


**DBLP: Liron Zighelnic**

www.informatik.uni-trier.de/~ley/db/indices/a-tree/z/Zighelnic-Liron.html

Liron Zighelnic, Oren Kurland: Query-drift prevention for robust query expansion update Mon Oct 1 18:15:59 2012 CET by the DBLP Team — Data released

**DBLP BibTeX Record 'conf/sigir/HummelSRKC12'**

dblp.uni-trier.de/rec/bibtex/conf/sigir/HummelSRKC12

@inproceedings{DBLP:conf/sigir/HummelSRKC12, author = {Shay Hummel and Anna Shtok and Fiana Raiber and Oren Kurland and ...}

**CiteSeerX — Corpus Structure, Language Models, and Ad Hoc**

citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.9.5663

Download: http://www.cs.cornell.edu/home/lee/papers/clustir DBLP CACHE [cs/0405044] Corpus structure, language models, and ad hoc ...

**DBLP - CS Bibliography ... Oren Kurland Lillian Lee**

**Search results for "query similarity" — FacetedDBLP**

dblp.13s.de/Topics/Topics/query%20similarity%26rrc=30

Michael Bendersky, Oren Kurland: Re-ranking search results using document graphs. SIGIR 2008: DBLP DOI BibTeX RDF: passage language models, paths, documents ...
The ad hoc retrieval task

Rank documents in a corpus by their relevance to the information need expressed by a query
Example: Vector space model
(Salton ’68)

- query = “Technion”
- $\vec{q} = \langle 0, ... , 0, 1, 0, ... , 0 \rangle$

- document = “Faculty Technion Student Technion”
- $\vec{d} = \langle 0, ... , 0, 1, 0, ... , 1, 0, ... , 2, 0, ... , 0 \rangle$

- $score(d; q) \triangleq cos(\vec{q}, \vec{d})$

- Term weighting scheme: TF.IDF
  - TF: the number of occurrences of a term in the document
  - IDF: The inverse of the document frequency of the term
Web search engines

- Use a variety of relevance signals
  - The textual similarity between the page and the query (query-dependent)
  - The textual similarity between the query and the anchor text of pages that point to the page (query-dependent)
  - The PageRank score of the page (query-independent)
  - The clickthrough rate for the page (query-independent)
  - ...
- Learning-to-rank (Liu ’09)
The document-query similarity

- Relevance is determined based on whether the document content satisfies the information need expressed by the query
- The document-query similarity is among the most important features for ranking pages in Web search engines (Liu ’09)

Back to classical information retrieval?
The cluster hypothesis

“Closely associated documents tend to be relevant to the same requests”

(Jardine & van Rijsbergen ’71, van Rijsbergen ’79)

Operational consequence:

Relevant documents should be more similar to each other than to non-relevant documents
Cluster-based results interface

Clustering the pages
Automatically labeling the clusters
Finding the highest quality clusters
Cluster-based document retrieval

1. Query
2. Initial list of documents
   - Clustering method
3. Set of clusters
   - Cluster ranking method
4. Ranking of clusters
   - Each cluster is replaced with its documents
5. Ranking of documents
### The optimal cluster problem
(Kurland & Domshlak ’08)

<table>
<thead>
<tr>
<th>Doc-query similarity</th>
<th>p@5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>WSJ</td>
</tr>
<tr>
<td><strong>Standard LM</strong></td>
<td>53.6</td>
</tr>
<tr>
<td><strong>Relevance Model</strong></td>
<td>58.8</td>
</tr>
<tr>
<td><strong>Optimal Cluster</strong></td>
<td>81.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Query expansion</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>WSJ</td>
</tr>
<tr>
<td><strong>Relevance Model</strong></td>
<td>58.8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Oracle experiment</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>WSJ</td>
</tr>
<tr>
<td><strong>Optimal Cluster</strong></td>
<td>81.5</td>
</tr>
</tbody>
</table>
The cluster ranking task

- Estimate the probability that cluster $C$ is relevant to query $Q$:
  \[
  p(C|Q) = \frac{p(C,Q)_{\text{rank}}}{p(Q)} = p(C,Q)
  \]

- Estimate $p(C,Q)$ using Markov Random Fields
Markov Random Fields

- Define a graph $G$:
  - Nodes – random variables representing $Q$ and $C$’s documents
  - Edges – dependencies between the variables
  - $L(G)$ – set of cliques in $G$
  - $l$ – clique
  - $\psi_l(l)$ – potential defined over $l$
  - $Z$ – normalization factor

$$p(C,Q) = \frac{\prod_{l \in L(G)} \psi_l(l)}{Z}$$

- A common instantiation of potential functions:
  - $f_l(l)$ – feature function defined over $l$
  - $\lambda_l$ – weight associated with $f_l(l)$

$$\psi_l(l) \overset{\text{def}}{=} \exp(\lambda_l f_l(l))$$
ClustMRF

- A linear (in feature functions) cluster ranking function that depends on the graph $G$

\[
p(C, Q)_{\text{rank}} = \sum_{l \in L(G)} \lambda_l f_l(l)\]

- Next:
  - Determine the clique set $L(G)$
  - Associate feature functions with cliques
The $\text{I}_{\text{QD}}$ clique

- Contains the query $Q$ and a single document $d$ in cluster $C$
- Consider query-similarity values of $C$’s documents independently

$$f_{\text{geo-qsim}}(l_{\text{QD}}) \overset{\text{def}}{=} \log \text{sim}(Q, d) \frac{1}{|C|}$$

$\text{sim}(\cdot, \cdot)$ is an inter-text similarity measure

$|C|$ – number of documents in $C$

(cf., Liu&Croft ’08)
The $l_{QC}$ clique

- Contains the query $Q$ and all $C$'s documents
- Induce information from relations between query-similarity values of $C$'s documents

\[ f_{A-qsim}(l_{QC}) \overset{\text{def}}{=} \log A(\{\text{sim}(Q,d)\}_{d \in C}) \]

\[ A \in \{\text{min, max, stdv}\} \]
The $l_C$ clique

- Contains only $C$'s documents
- Induce information based on query-independent properties of $C$'s documents (e.g., PageRank score, ratio of stopwords to non-stopwords)

$$f_{A-P}(l_C) = \log A(\{P(d)\}_{d \in C})$$

$A \in \{\text{min, max, geometric mean}\}$

$P$ is a query-independent document (quality) measure
Empirical evaluation

Query

Markov Random Field (MRF) (Metzler & Croft '05)

Initial list of 50 documents

Nearest Neighbor clustering (Griffiths et al. '86)

Set of clusters

ClustMRF

Each cluster is replaced with its documents

Ranking of clusters

Ranking of documents
Comparison with the initial ranking

- **Init**: Markov Random Field (Metzler & Croft ’05)
- **ClustMRF** (Our algorithm)
- Statistically significant differences with **ClustMRF**
Comparison with other cluster ranking methods

- **AMean**: Arithmetic mean of query similarity values (Liu & Croft '08)
- **GMean**: Geometric mean of query-similarity values (Liu & Croft '08, Seo & Croft '10)
- **ClustRanker**: Uses measures of document and cluster biases (Kurland '08)
- **ClustMRF** (Our algorithm)

♦ Statistically significant differences with ClustMRF

![Graph showing MAP comparison between ClustMRF and other methods](chart)

1. **GOV2**
2. **ClueWebB**
Comparison with automatic query expansion

- **RM3** (Abdul-Jaleel et al. ’04)
- **ClustMRF** (Our algorithm)
- ♦ Statistically significant differences with ClustMRF

![MAP comparison chart for GOV2 and ClueWebB](chart.png)

- **GOV2**: ClustMRF vs. RM3
- **ClueWebB**: ClustMRF vs. RM3
Diversifying search results
Diversifying search results

- **MMR** (Carbonell & Goldstein '98) and **xQuAD** (Santos et al. '10)
  iteratively re-rank the initial list
- In each iteration a document is scored by:

\[
\beta \text{ Relevance to the query} + (1 - \beta) \text{ Diversity with respect to documents already selected}
\]

Relevance estimates:
- **Init (MRF)**
- **ClustMRF**
- ♦ Statistically significant differences with **ClustMRF**

![Graph showing α-NDCG@20 for MMR and xQuAD on ClueWebB](attachment:image.png)
TREC (Text REtrieval Conference)

- Annual “competition” of search approaches
- In 2013, the Technion won the first place in the Web retrieval track according to most evaluation criteria
  - ClustMRF was used to re-rank the results of a highly effective learning-to-rank approach
Summary

- We presented a novel cluster ranking approach that
  - outperforms state-of-the-art document-based retrieval methods
  - outperforms state-of-the-art cluster-based retrieval methods
  - can be used to improve the performance of results-diversification methods