Thesaurus-Based Feedback to Support Mixed Search and Browsing Environments

Edgar Meij and Maarten de Rijke

ISLA
University of Amsterdam

11th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2007)

Universiteit van Amsterdam

Outline

Outline

1. Introduction
   - Motivation
   - Research Questions
   - Language Modeling
2. Our Algorithm
   - Overview
   - Determining Thesaurus Terms
   - Estimating a Thesaurus-biased Model
   - Interpolating the Original Query Model
3. Results and Discussion
   - Test collection
   - Retrieval Effectiveness
   - Browsing Effectiveness
   - Per-topic Results

UNIVERSITEIT VAN AMSTERDAM

Information Access in the Digital Library

- Vocabulary mismatch
  - Not all authors use the same terms
  - Different authors may use different terms for a single concept or may even denote different concepts with the same term
- Solutions
  - Use the cataloging system/controlled vocabulary/thesaurus of the digital library
  - Apply query enrichment/expansion

Information Access in the Digital Library

- Vocabulary mismatch
  - Not all authors use the same terms
  - Different authors may use different terms for a single concept or may even denote different concepts with the same term
- Solutions
  - Use the cataloging system/controlled vocabulary/thesaurus of the digital library
  - Apply query enrichment/expansion

Information Access in the Digital Library

- Information access in a digital library is usually associated with two tasks:
  - Searching
  - Browsing
- Query expansion can be used to improve the *search* component, whereas a controlled vocabulary may used to enhance the *browsing* component.
- Aim: How can we combine these?

Information Access in the Digital Library

- Information access in a digital library is usually associated with two tasks:
  - Searching
  - Browsing
- Query expansion can be used to improve the *search* component, whereas a controlled vocabulary may used to enhance the *browsing* component.
- Aim: How can we combine these?

Information Access in the Digital Library

- Information access in a digital library is usually associated with two tasks:
  - Searching
  - Browsing
- Query expansion can be used to improve the *search* component, whereas a controlled vocabulary may used to enhance the *browsing* component.
- Aim: How can we combine these?

Research Questions

1. How can we use a language modeling framework to generate thesaurus terms, as well as provide query expansion/pseudo-relevance feedback?
2. What is the impact of the size of the corpus from which feedback terms are being generated?
3. Can our model compete with state-of-the-art IR approaches?
4. How can we assess the quality of the thesaurus terms being proposed for browsing?

Outline

1. Introduction
   - Motivation
   - Research Questions
   - Language Modeling
2. Our Algorithm
   - Overview
   - Determining Thesaurus Terms
   - Estimating a Thesaurus-biased Model
   - Interpolating the Original Query Model
3. Results and Discussion
   - Test collection
   - Retrieval Effectiveness
   - Browsing Effectiveness
   - Per-topic Results

Generative Language Models

- Query-likelihood approach:

$$
\begin{aligned}
P(Q|d) &\propto P(d) \cdot \prod_{q \in Q} P(q|\theta_d) \\
&\propto \prod_{q \in Q} \frac{c(q,d)}{|d|}
\end{aligned}
$$

- With Dirichlet smoothing:

$$
P(Q|d) \propto \prod_{q \in Q} \frac{c(q,d) + \mu P(q|\theta_C)}{|d| + \mu}
$$

Universiteit van Amsterdam

Generative Language Models

- Query-likelihood approach:

$$
\begin{aligned}
P(Q|d) &\propto P(d) \cdot \prod_{q \in Q} P(q|\theta_d) \\
&\propto \prod_{q \in Q} \frac{c(q,d)}{|d|}
\end{aligned}
$$

- With Dirichlet smoothing:

$$P(Q|d) \propto \prod_{q \in Q} \frac{c(q,d) + \mu P(q|\theta_C)}{|d| + \mu}$$

Relevance Models

- Generative language modeling assumes that queries are generated from documents
- Relevance modeling assumes both are generated from an unseen source—a relevance model
- A set of documents *R* is used as a model from which terms are sampled

$$P(w|\hat{\theta}_Q) \propto \sum_{d \in R} P(w|\theta_d) \cdot P(Q|d)$$

Universiteit van Amsterdam

Relevance Models

- Generative language modeling assumes that queries are generated from documents
- Relevance modeling assumes both are generated from an unseen source—a relevance model
- A set of documents R is used as a model from which terms are sampled

$$P(w|\hat{\theta}_Q) \propto \sum_{d \in R} P(w|\theta_d) \cdot P(Q|d)$$

Relevance Models

- Generative language modeling assumes that queries are generated from documents
- Relevance modeling assumes both are generated from an unseen source—a relevance model
- A set of documents R is used as a model from which terms are sampled

$$P(w|\hat{\theta}_Q) \propto \sum_{d \in R} P(w|\theta_d) \cdot P(Q|d)$$

Final Ranking

- Ranking then comes down to calculating the *distance* between $P(w|\theta_Q)$ and $P(w|\theta_d)$ for $w \in V$
- E.g. using the KL-divergence

$$D_{kl}(\theta_Q||\theta_d) = \sum_w P(w|\theta_Q) \cdot \log \frac{P(w|\theta_Q)}{P(w|\theta_d)}$$

Universiteit van Amsterdam

Final Ranking

- Ranking then comes down to calculating the *distance* between $P(w|\theta_Q)$ and $P(w|\theta_d)$ for $w \in V$
- E.g. using the KL-divergence

$$D_{kl}(\theta_Q||\theta_d) = \sum_w P(w|\theta_Q) \cdot \log \frac{P(w|\theta_Q)}{P(w|\theta_d)}$$

Outline

Our Algorithm: Three Steps

1. Determine the thesaurus terms most closely associated with a query
2. Search the documents associated with these thesaurus terms, in conjunction with the query, to look for additional terms to describe the query
3. Interpolate the query model with the found terms

Determining Thesaurus Terms

- For any given query Q, rank the thesaurus terms $m \in M$ according to:

$$
\begin{aligned}
P(m|Q) &= \frac{P(m)P(Q|m)}{P(Q)} \\
&= P(m)\sum_{d} P(Q|d)P(d|m)
\end{aligned}
$$

Estimating a Thesaurus-biased Model

- Then, estimate a thesaurus-biased relevance model by incorporating the top-l thesaurus terms:

$$P(w|\hat{\theta}_Q) \propto \sum_{d \in R} P(w|\theta_d) \cdot P(Q|d) \cdot P(m_1, \ldots, m_l|d)$$

- Assuming the thesaurus terms $m_1, \ldots, m_l$ to be independent and $P(d)$ to be uniform, we obtain

$$P(w|\hat{\theta}_Q) \propto \sum_{d \in R} P(w|\theta_d) \cdot P(Q|d) \cdot \prod_{i=1,\ldots,l} P(d|m_i) \cdot P(m_i)$$

Estimating a Thesaurus-biased Model

- Then, estimate a thesaurus-biased relevance model by incorporating the top-l thesaurus terms:

$$P(w|\hat{\theta}_Q) \propto \sum_{d \in R} P(w|\theta_d) \cdot P(Q|d) \cdot P(m_1, \ldots, m_l|d)$$

- Assuming the thesaurus terms $m_1, \ldots, m_l$ to be independent and $P(d)$ to be uniform, we obtain

$$P(w|\hat{\theta}_Q) \propto \sum_{d \in R} P(w|\theta_d) \cdot P(Q|d) \cdot \prod_{i=1,\ldots,l} P(d|m_i) \cdot P(m_i)$$

Interpolating the Original Query Model

- Finally, the found model is interpolated with the original query using a mixing weight λ to yield the final query model

$$P(w|\theta_Q) = \lambda \cdot \frac{c(w,Q)}{|Q|} + (1-\lambda) \cdot P(w|\hat{\theta}_Q)$$

- When λ is set to 1, a query-likelihood ranking is obtained

Interpolating the Original Query Model

- Finally, the found model is interpolated with the original query using a mixing weight λ to yield the final query model

$$P(w|\theta_Q) \quad = \quad \lambda \cdot \frac{c(w, Q)}{|Q|} + (1 - \lambda) \cdot P(w|\hat{\theta}_Q)$$

- When λ is set to 1, a query-likelihood ranking is obtained

Outline

TREC Genomics 2006

- Passage retrieval from 160k full-text (biomedical) documents
  - We only look at document-level relevance assessments
  - 28 topics
- PubMed
  - Bibliographic database maintained by the National Library of Medicine (NLM)
  - Over 15M entries, containing author information, abstracts, etc. etc.
- Medical Subject Headings (MeSH)
  - 22,997 hierarchically ordered concepts
  - Trained annotators from the NLM assign one or more MeSH terms to every document indexed in PubMed

TREC Genomics 2006

- Passage retrieval from 160k full-text (biomedical) documents
  - We only look at document-level relevance assessments
  - 28 topics
- PubMed
  - Bibliographic database maintained by the National Library of Medicine (NLM)
  - Over 15M entries, containing author information, abstracts, etc. etc.
- Medical Subject Headings (MeSH)
  - 22,997 hierarchically ordered concepts
  - Trained annotators from the NLM assign one or more MeSH terms to every document indexed in PubMed

TREC Genomics 2006

- Passage retrieval from 160k full-text (biomedical) documents
  - We only look at document-level relevance assessments
  - 28 topics
- PubMed
  - Bibliographic database maintained by the National Library of Medicine (NLM)
  - Over 15M entries, containing author information, abstracts, etc. etc.
- Medical Subject Headings (MeSH)
  - 22,997 hierarchically ordered concepts
  - Trained annotators from the NLM assign one or more MeSH terms to every document indexed in PubMed

Retrieval Effectiveness

	λ	MAP		P10	
QL	1	0.359		0.45	
RM (collection)	0.10	0.426	+19%	0.48	+7%
RM (PubMed)	0.35	0.425	+18%	0.48	+7%
MM (collection)	0.05	0.424	+18%	0.48	+7%
MM (PubMed)	0.45	**0.429**	+20%	**0.49**	+9%

Comparison between different query models and a query-likelihood baseline (best scores in boldface.)

Retrieval Effectiveness

	λ	MAP		P10	
QL	1	0.359		0.45	
RM (collection)	0.10	0.426	+19%	0.48	+7%
RM (PubMed)	0.35	0.425	+18%	0.48	+7%
MM (collection)	0.05	0.424	+18%	0.48	+7%
MM (PubMed)	0.45	**0.429**	+20%	**0.49**	+9%

Comparison between different query models and a query-likelihood baseline (best scores in boldface.)

Retrieval Effectiveness

	λ	MAP		P10	
QL	1	0.359		0.45	
RM (collection)	0.10	0.426	+19%	0.48	+7%
RM (PubMed)	0.35	0.425	+18%	0.48	+7%
MM (collection)	0.05	0.424	+18%	0.48	+7%
MM (PubMed)	0.45	**0.429**	+20%	**0.49**	+9%

Comparison between different query models and a query-likelihood baseline (best scores in boldface.)

Browsing Effectiveness

- Concept specificity
- Average distance from each concept to the root of the thesaurus:
  - Collection-based: 4.46
  - PubMed-based: 4.78

Browsing Effectiveness

- TREC 2006 assessors assigned MeSH terms to relevant passages
- Average agreement with the assessors:
  - Collection-based: 2.3/10
  - PubMed-based: 3.0/10
  - *Difference is statistically significant ($p < 0.05$)*

Browsing Effectiveness

- TREC 2006 assessors assigned MeSH terms to relevant passages
- Average agreement with the assessors:
    - Collection-based: 2.3/10
    - PubMed-based: 3.0/10
    - *Difference is statistically significant ($p < 0.05$)*

Summary

- We integrated information inherent in digital libraries into a generative language modeling framework
- Our aim was to facilitate *browsing*, while maintaining and/or improving *retrieval* effectiveness
- While readily providing thesaurus terms, our model outperforms state-of-the-art IR methods when estimated on a sufficiently large corpus