Local Word Embeddings for Query Expansion based on Co-Authorship and Citations

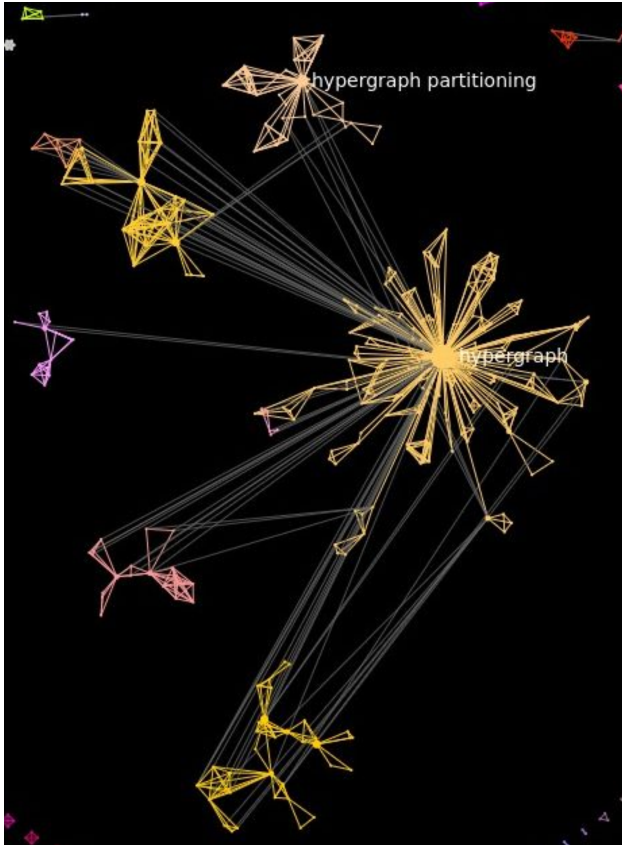
André Rattinger, Jean-Marie Le Goff, Christian Guetl

Introduction

- Learning of representations is a long standing problem
- Mitigate sparsity, learn similarities
- Terms can be used to expand queries
- Our datasets provide limited information for expansion
- Two data sources: publications and patents
- Information from authors and cited documents can help

Collaboration Spotting

- Graph visualization at CERN
- Builds collaborations of multidimensional graphs
- Collaborations are based on search in patents and publications



Query Expansion

- Query can be a poor representation of information need
- Expand query with synonyms and related words
- Effective expansion is done in the fitting context
  - Global context: “latent”: “inherent”, “surpresses”, “innate”
  - Local context: “lsa”, “dirichlet”, “allocation”, “plsi”
- Pseudo relevance feedback can help with this
  - We use it for the selection of training documents

Word Embeddings

- Fixed representation of words (and documents, etc.)
- Neural network based method, recently gained popularity in IR
- Semantically similar terms are close to each other
- Word2Vec, Glove, many more
- Benefits from big datasets while training

Datasets - ACL

- Small dataset of 9,793 research papers, 82 topics
- Scientific publications from the field of computational linguistics
- Supplemented with articles from other authors and from citations
- 33,922 articles in total (for expansion and training, not for retrieval)

Datasets - CLEF-IP

- English subset of the CLEF-IP 2011 collection
- ~420,000 patents, 1350 topics
- Each topic is expressed as a document instead of a query
- Query terms generated from the description (~30 terms)
- No supplementation is size is deemed sufficient
- Patent citation valuable as they added by the author and patent examiner.

Datasets - Overview

- Few relevant documents for both datasets
- Vocabulary size comes from the indexed documents
- ACL, is small CLEF-IP is even smaller (partly caused by english subset)

Name	Topics	Vocab Size	Indexed Docs	Avg. Relevant Docs
ACL	82	329,490	9,793	23.67
CLEF-IP	1,350	2,648,818	420,193	7.2

Experimental Setup

- Pre-processing and indexing
- Initial training of word embeddings
- Retrieval and query expansion

Pre-processing

- Indexing
  - Regex tokenizer, transformation to lower case
  - Stopwords are filtered with SMART stopword list
  - Removed patent specific stopwords
- Word Embeddings
  - Regex tokenization, transformation to lower case
  - Krovetz stemming for ACL, to reduce the overall vocabulary size as the corpus is very small

Learning of word embeddings

- Initial training on whole dataset
- Another model is trained on the English-language edition of Wikipedia
- Initial model is learned, because training many local models is very inefficient
- Settings: minimum frequency of words 8, window size 7, Skip-Gram
- ACL: 20 iterations, CLEF-IP: 5 iterations

Retrieval and Query Expansion (1)

- Set of initial documents is retrieved with inverse document frequency model (InL2) from terrier
- Top k retrieved documents are used as feedback documents
- All available document from authors and citations used for retraining

Retrieval and Query Expansion (2)

- Q be a query issued by the user, $q_1, q_2, \ldots, q_n$
- C be the list of candidate terms for query expansion, represented as $c_1, c_2, \ldots c_k$
- The initial set of C is selected out of all of the terms in the first m relevant documents and the query
- Expanded by all terms from references and authors

Retrieval and Query Expansion (3)

- Stopwords are filtered from candidate terms and they are ranked by Bo1
- Documents are used for retraining word embeddings
- Top k terms for the top ranked candidate terms are generated
- Ranked by Bo1 again

Results

- **Baseline:** retrieval without query expansion applied
- **QE global:** global query expansion with a general purpose query expansion model trained on a dataset from the English-language edition of Wikipedia
- **QE local** locally-trained model
- **QE local ext.** locally-trained model with the extension of reference documents and documents from co-authors.

Results - ACL

- General low retrieval performance, also in reference works, caused by low number of relevant documents
- Improved overall retrieval with local query expansion

Method	MAP	P@5	P@10
Baseline	0.1497	0.2268	0.1683
QE global	0.1502	0.2268	0.1732
QE local	0.1623*	**0.2347**	0.1805
QE local ext.	**0.1713***	0.2314	**0.1822**

Results - CLEF-IP

- Also low retrieval performance
- Improved overall retrieval with local query expansion, but no significant improvements

Method	MAP	P@5	P@10
Baseline	0.0914	0.0630	0.0446
QE local	0.0916	0.0631	0.0448
QE local ext.	**0.0923**	**0.0636**	**0.0455**

Limitations and Future Work

- Implementation of other query expansion methods
- Integration of knowledge bases for retrieval
- Use dataset with more relevant documents on average
- CLEF-IP is based on documents and not queries
- Compare to other query expansion models
- Experiments with the patent classification system
- Different retrieval metrics

Conclusion

- Inclusion of documents that are likely to be relevant provides further information for term selection
- Query expansion increased performance for both datasets, but only significant for the ACL dataset
- CLEF-IP might be problematic because of the low number of relevant documents
- Seems effective for small datasets, but might have a certain number of relevant documents

Thank you