Efficient Search Result Diversification via Query Expansion Using Knowledge Bases

Raoul Rubien, Hermann Ziak, **Roman Kern**

Know-Center Graz

TIR, 2015-09-01

What?
Research Questions

Background - What

Main research question

- Does **query expansion** help to increase the **diversity** of search results and to which extend?

Secondary research question

- What role does the **query formulation** play in this process?

Why?

What are the implications?

Background - Why

What is the current state?

- Most of the current methods for diversification **rearrange** a long search result list
- By picking results, which match certain criteria via a **cost function**
- Thereby **discarding** a number of search results
- The computation of the cost function is often also **computationally complex**

What is then changed?

- Query expansion does not require to alter the search result list
- It is therefore far more efficient

Context & System

What is our setting and how does our system work?

System - General

General background

- We are developing a **vertical aggregated search system**
- Where search engines are treated as **black boxes**
- Queries are **automatically generated** out of the current user's context
- **Latency** does play an important role

What did motivate us to work on this topic?

- Query expansion techniques are known to **increase recall**
- In literature we found some hints, that it also helps for diversity
- But no systematic comparison

System - Query Expansion

Query expansion strategy

- Our query expansion methods rely on **pseudo relevance feedback**
  1. Take the original query
  2. Conduct a search and collect the results
  3. Create a set of candidate terms out of the results
  4. Rank the candidate terms and define cut-off point
  5. Add the top candidate terms to the query
- The expanded query is then submitted to the search engine

System - Query Expansion

Search the query expansion index

- Our system is capable to use **different systems** for pseudo relevance feedback and for searching
- Currently, we use an external knowledge base just for query expansion
- Specially build **Wikipedia** index
  - Split each Wikipedia article into paragraphs
  - Facets: title, paragraph title, paragraph content
- Allow partial matches, restrict to a number of search results

System - Query Expansion

Candidate selection

- Collect terms from all facets
- Rank the terms according to score $s(t)$
- Select the top k terms

$$s(t) = \sum_{i \in S} \sum_{f \in F} DFR(boost(f) * score(d_i))$$

System - Query Formulation

How is the final query being constructed?

- The way how the expansion terms are added to the query depends on the **capabilities** of the search engine
- We implemented two strategies
  1. A simple baseline, disjunction of all terms

$$OrigQueryTerms\ OR\ ExpTerm_1\ OR\ ...\ OR\ ExpTerm_n$$

  2. The grouping method, expanded terms are grouped

$$OrigQueryTerms\ OR\ (ExpTerm_1\ OR\ ...\ OR\ ExpTerm_n)$$

Evaluation

How did we obtain our results?

Evaluation - Overall Approach

Evaluation goals

- Measure the amount of diversification
- Secondary, compare the different query formulation strategies

Evaluation strategy

1. Compute the search results without query expansion
2. Compute search results using a state-of-the-art diversification technique
   - And compute the diversification against the unexpanded query
3. Compute the search results with query expansion
   - And compute the diversification against the unexpanded query
4. Compare the amount of diversification b/w the two diversification strategies

Evaluation - Reference System

Comparison system

- Implemented a state-of-the-art diversification algorithm - **IA-Select** (Intent Aware - Select)
  - Explicit diversification of search result
  - Requires a weighted mapping for the query to a classification scheme
  - Plus a weighted mapping of the results to the same classification scheme

Note: IA-Select is restricted to items from the original result list, while the search result list with the expanded query may contain many additional results.

Evaluation - Query Set

Query set for evaluation

- Collected queries from **query logs**
  - Including manually entered queries
  - Including automatically generated queries out of users' context
- Manually cleaned and removed duplicates
- Final set consists of 70 queries
- Assignment to categories conducted manually

Evaluation - Measure

Measure of diversity

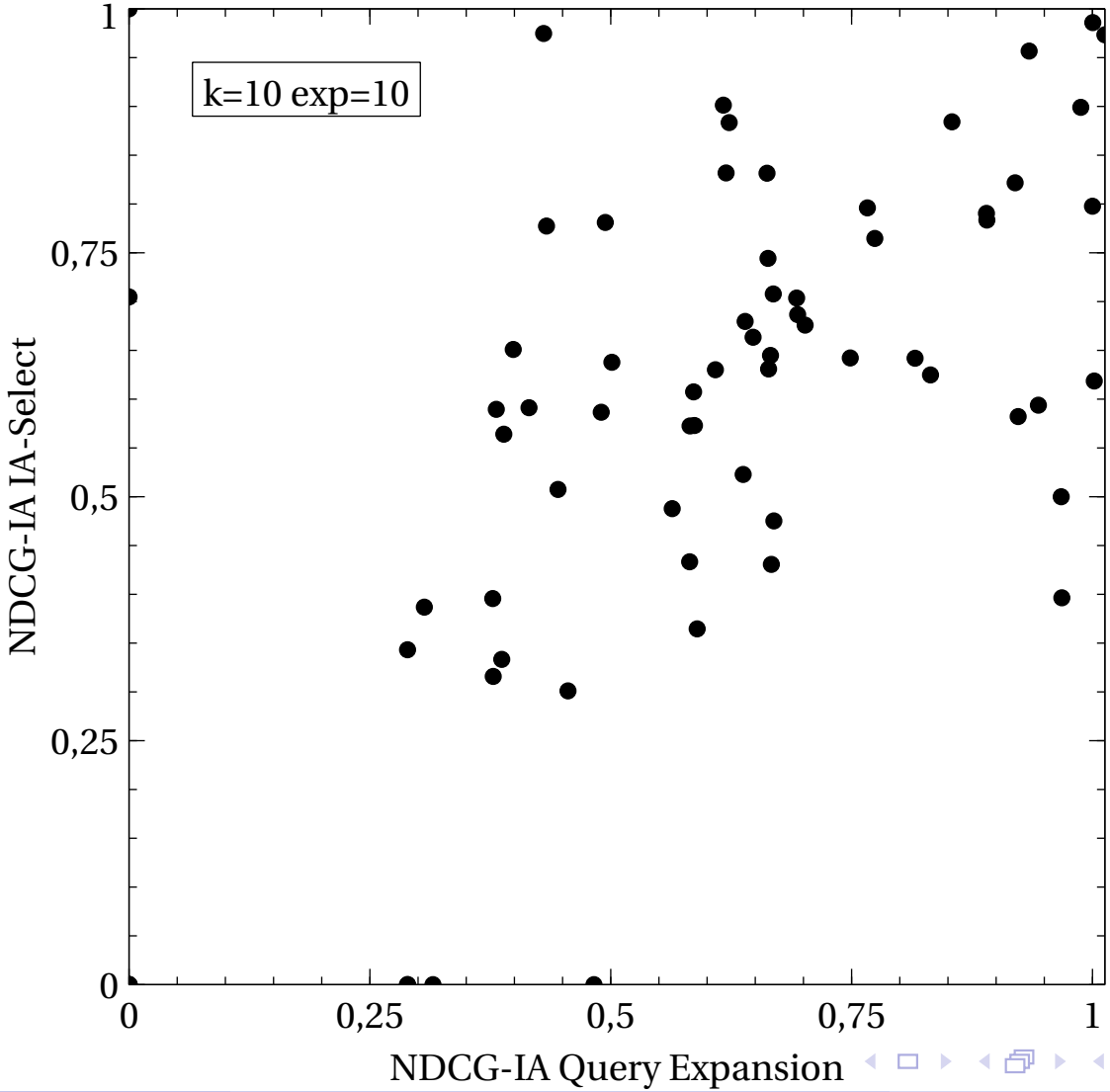
- Means to measure the amount of of diversification
- **NDCG-IA** (Normalized Discounted Cumulative Gain - Intent Aware)
- Modification of the NDCG measure
- Compares two search result lists (the unexpanded query is always taken as reference)
- Compute $NDCG\_IA@k(R_{IA}(q_i))$ and $NDCG\_IA@k(R_{QE}(q_i'))$ for all $q_i \in Q$

Results

The results of the evaluation and discussion

Results

Comparison of the amount of diversification



Results

Comparison of query formulation strategy

Strategy	Pearson's r	Spearman's rho	Kendall's tau
Simple	0.46	0.42	0.30
Grouped	**0.59**	**0.55**	**0.41**

Conclusions

Summary & Outlook

Conclusions

Summary

- Query expansions tweak the search results to contain more diversity
  - $\rightarrow$ Both **efficient** and **effective**
- Number of query terms does play a role - 10 a good starting point
- The actual **query formulation strategy** plays an even bigger role

Future work

- Investigate on more advanced query formulation strategies, e.g. weighting of terms

The End

Thank you for your attention!