KDEIM at NTCIR-12 IMINE-2: Intent Mining Through Diversified Ranking of Subtopics

Md Zia Ullah, Md Shajalal, and Masaki Aono

Knowledge Data Engineering and Information Retrieval Laboratory

Toyohashi University of Technology, Toyohashi, Japan

Introduction

Motivation:

Problem with Web Search:

- Ambiguous, vague, or broad queries
- Diverse intents behind the same query
- Heterogeneous information of query
- Popular intent dominates in the search results

Challenges:

- Mining diverse subtopics of the query
- Predicting relevant verticals of the subtopic

e.g. Query: Pluto, Subtopic: Pluto picture, Vertical: image and web

Our Goal:

- Identifying diversified subtopics covering intents of the query
- Classifying subtopic into vertical intents

Overview:

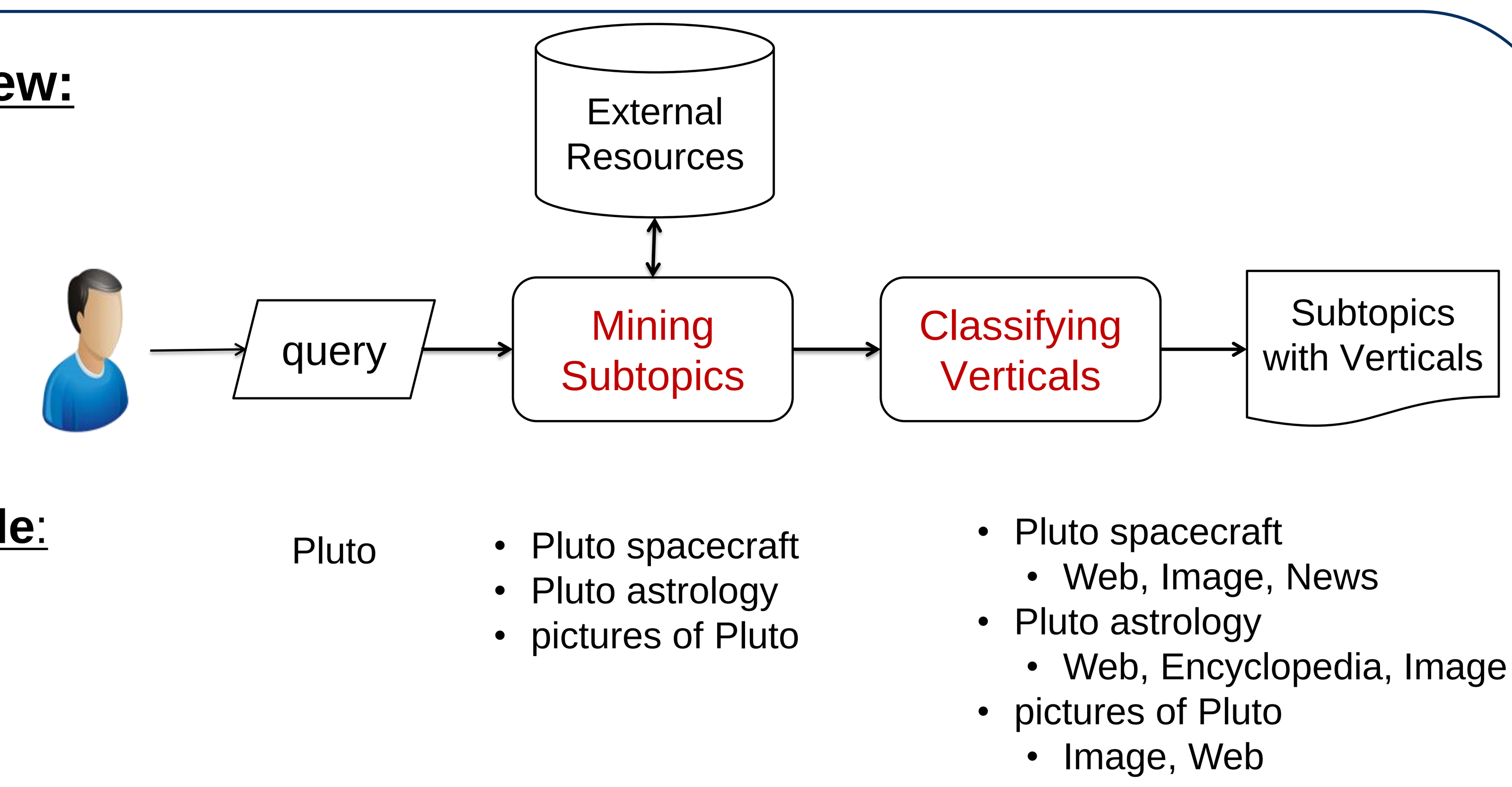


Figure 1: An overview of our subtopic mining framework

Subtopic Mining Framework

Candidate Subtopics:

Resources

- Bing query suggesions and completions
- Google query completions
- Yahoo query completions

Feature Extraction:

- Term frequency based features (DFH, PL2, BM25, etc.)
- Language modeling based features (KL, QLM-JM, SLM-JM, etc.)
- Lexical features (Edit distance, Term overlap, etc.)
- Web hit count based features (NHC, PMI, etc.)

Feature Selection:

Optimization function of Elastic Net:

$$\min_{\beta_0,\beta}\left(\frac{1}{2M}\sum_{i=1}^{M}(y_i-\beta_0-X_i^T\beta)^2+\lambda\sum_{j=1}^{p}\left(\frac{(1-\alpha)}{2}\beta_j{}^2+\alpha||\beta_j||\right)\right)$$

Relevance Estimation:

Linear ranking model is employed to estimate relevance as follows:

$$rel(q,s)=\sum_{k=1}^{N}w_k\ .f_k(q,s)$$

- Random forest is utilized to estimate feature importance w_k
- N is the number of selected features.

Subtopic Diversification:

Diversifying subtopics by balancing relevance and novelty:

$$s_i^{*}=arg\max_{S_i\in R\text{¥}C_i}[\gamma\ rel(q,si)+(1-\gamma)\ novelty(s_i,C_i)]$$

$$novelty(s_i,C_i)=-\max_{s'\in C_i}cosine(s_i,s')$$

$\gamma\in[0,1]$ is a combining parameter.

$novelty(s_i,C_i)$ indicates the novelty of subtopic s_i given the set C_i

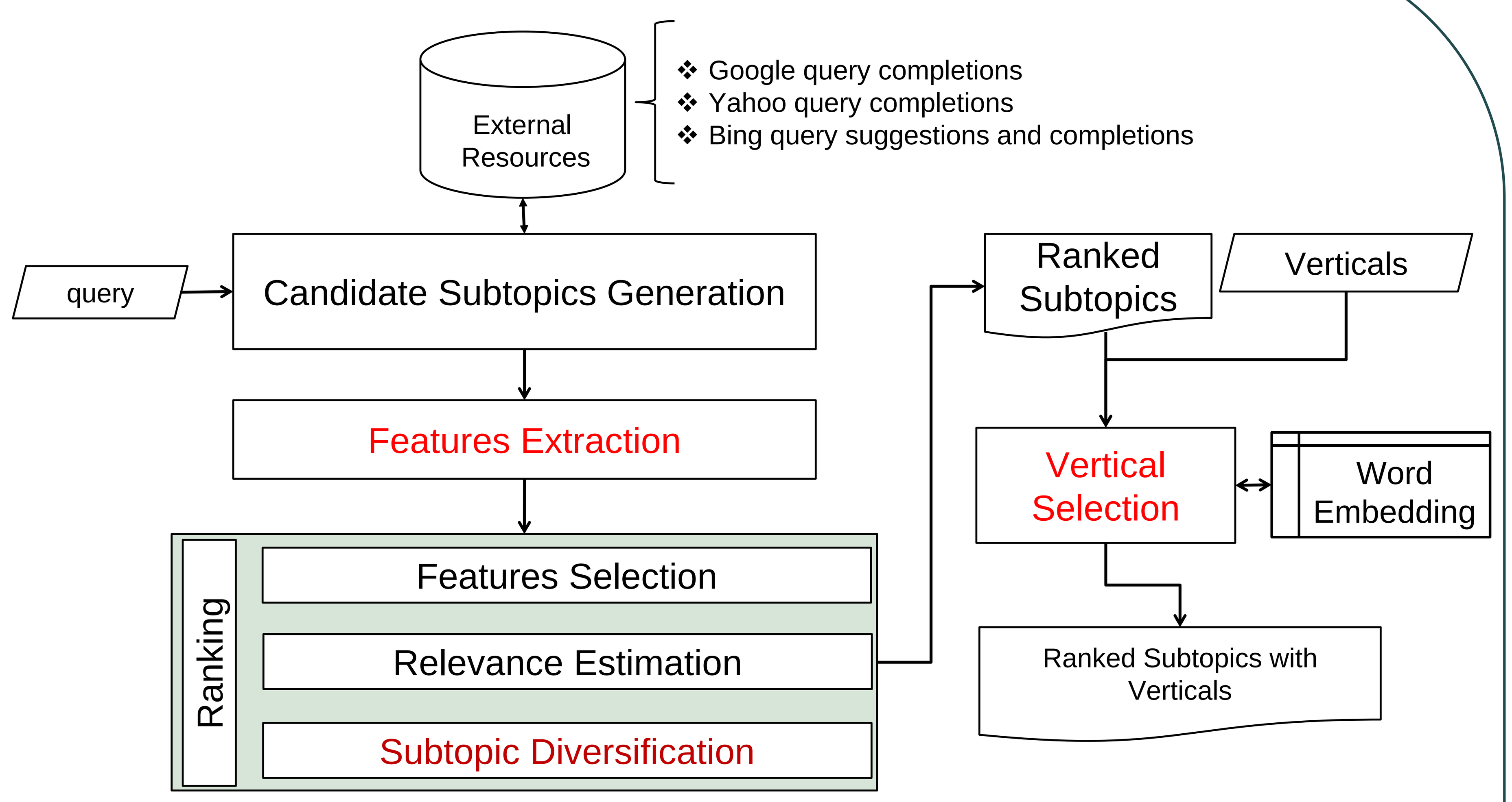


Figure 2: A Vertical Oriented Subtopic Mining Framework

Vertical Selection:

- Verticals:
  - Web, Image, News, QA, Encyclopedia, and Shopping
- Word vectors:
  - 300-dimensional embedding from word2vec (Google News Corpus)
- Vertical representatives:
  - Image vertical: Photo, Album, Gallery, and Artwork

Vertical vector: $\boldsymbol{v_v}=\frac{1}{L}\sum_{l=1}^{L}\boldsymbol{t}_i$ Subtopic vector: $\boldsymbol{v_s}=\frac{1}{P}\sum_{p=1}^{P}\boldsymbol{t}_i$

If $\mathrm{cosine}(v_s\ \ v_v)\geq 0.75$, subtopic s is a type of vertical v

Experiments and Evaluation

Dataset:

Training: NTCIR-10 INTENT-2 English

Testing: NTCIR-12 IMINE-2 English

Subtopic Mining Subtask:

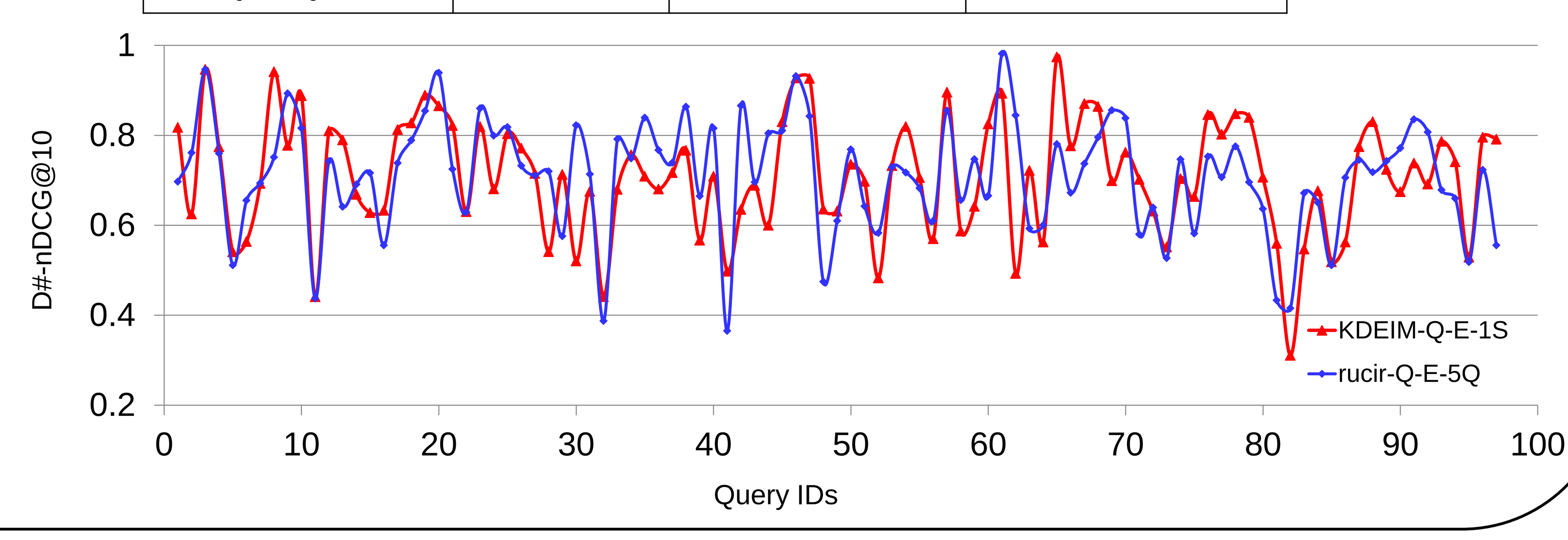
Runs	I-rec@10	D-nDCG@10	D#-nDCG@10
KDEIM-Q-E-1S	**0.7556**	**0.6644**	**0.7100**
KDEIM-Q-E-2Q	0.7556	0.6644	0.7100
KDEIM-Q-E-3Q	0.7458	0.6472	0.6955
KDEIM-Q-E-4S	0.7484	0.5645	0.6565

Query Understanding with Vertical

Runs	V-score	QU-score
KDEIM-Q-E-2Q	**0.3014**	**0.5057**
KDEIM-Q-E-3Q	0.2931	0.4948

Comparison with IMINE-2 participants:

Runs	I-rec@10	D-nDCG@10	D#-nDCG@10
KDEIM-Q-E-1Q	0.7556	0.6644	**0.7100**
ruicir-Q-E-5Q	0.7502	0.6694	0.7098
HULTECH-Q-E-1Q	0.7279	**0.6786**	0.7033
ruicir-Q-E-4Q	**0.7601**	0.5096	0.6348



Conclusion

- Proposed a method for mining diversified subtopics
- Proposed a method for vertical selection exploiting word embedding
- Language modelling and query independent features are effective
- Diversification penalize the noisy and redundant subtopics

Future Work

- Extracting candidate subtopics from other resources
  - Top retrieved documents, Wikipedia, Knowledge graph.
- More semantic features for estimating relevance and novelty
- Effectively using word embedding for vertical classification
- Search result diversification using the mined subtopic and vertical