IMC at the NTCIR-12 IMine-2 Query Understanding Subtask

Jiahui Gu, Chong Feng, Yashen Wang

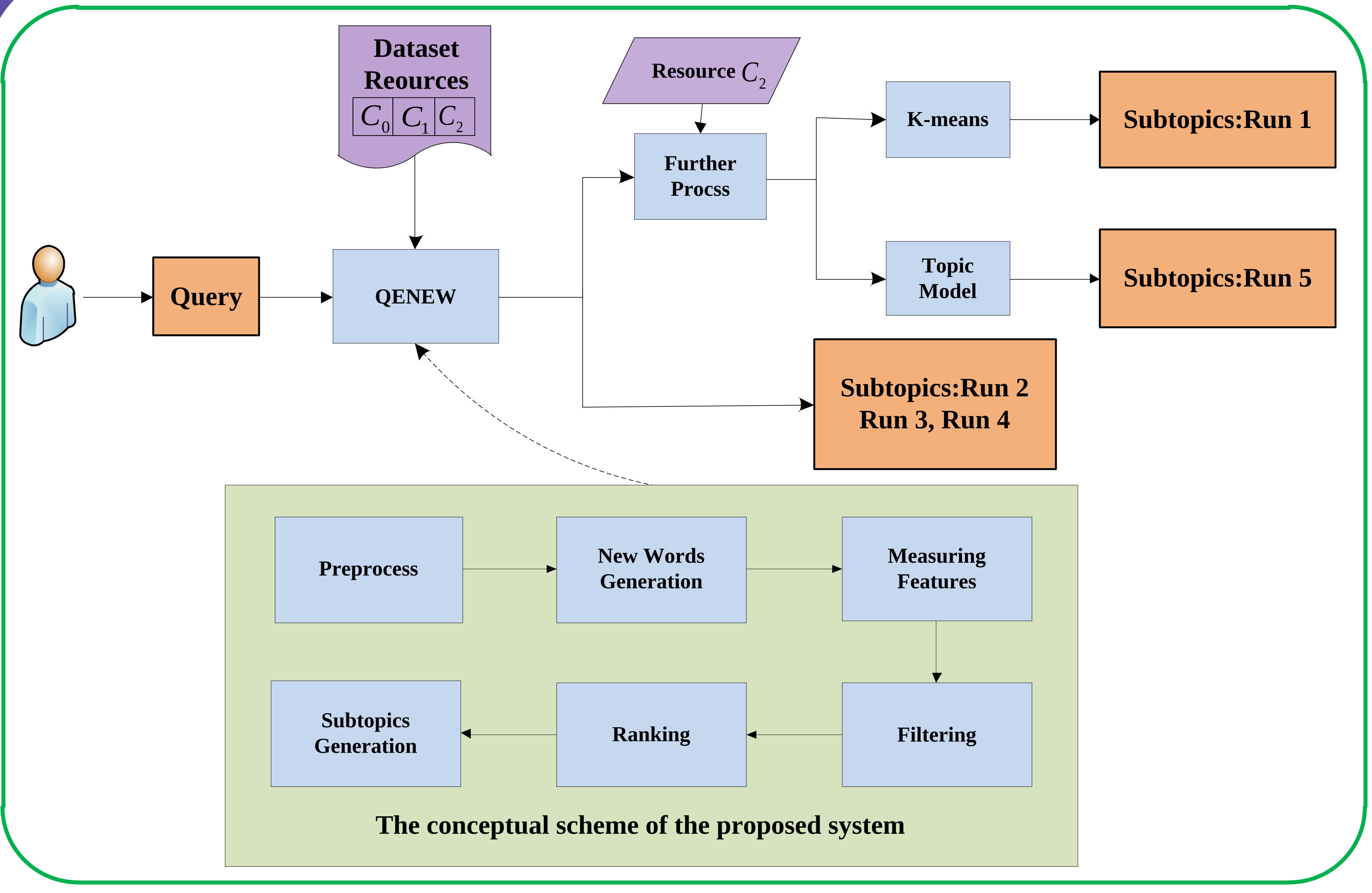
School of Computer Science, Beijing Institute of Technology

{ gujh, fengchong, yswang }@bit.edu.cn

INTRODUCTION

- ➢ Propose a novel framework of QENEW that is **Query Expansion based on New-Word Extraction Algorithm**
- ➢ The extracted new words, whether they exit in the available lexicon or not, are considered as query expansion terms. Then subtopics are generated by linear concatenation of the original query and expansion words
  - ✓ For example, string “*电影院*(*Cinema*)” isn’t viewed as a word. On the contrary, it’s just a sequence of characters at the beginning of our algorithm
- ➢ External Chinese corpus are utilized and crawled from Baidu, Google and Bing
- ➢ K-means algorithm and Topic Model are applied in experiments

OVERVIEW OF THE FRAMEWORK



The conceptual scheme of the proposed system

New Words Generation

- ➢ Extracted by n-gram model

Measuring Features

- ➢ Frequency(F)
- ➢ String Cohesion(SC)
- ➢ String Liberalization(SL)

Ranking

- ➢ $P = \alpha_1 F + \alpha_2 SC + \alpha_3 SL$

Subtopics Generation

- ➢ Subtopics = query+ expansion terms

Dataset

- ➢ C_0: IMine-1 Chinese Web Corpus
- ➢ C_1: Comging from the crawled top five documents in HTML webpages for Baidu, Google and Bing
- ➢ C_2: Comging from the bottom of HTML webpages labeled with “*Related searches*”

Further Process

- ➢ K-means: Adopt the default settings to make diversity clustering of query subtopics
  - Cluster number is between 5 and 10, select the highest frequency of term as the subtopic
- ➢ Topic model: Refers to two topic models to generate subtopic terms
  - Latent Dirichlet Allocation (LDA) and Hierarchical Dirichlet Process(HDP)
  - Set the topic number of LDA and HDP is 8 and use one topic words to describe the corresponding topic. Then, the top 10 topic words with the highest occurrence probabilities among the 16 words are the subtopics of the query

EXPERIMENTS AND RESULTS

Submitted Runs

RunID	**Description**	
	QENEW input	Further Process
IMC-Q-C-1S	C_1,Co	QENEW’s output and C_2 use K-means algorithm to generate the final run
IMC-Q-C-2S	C_1,C_2,Co	QENEW’s output is the final run
IMC-Q-C-3S	C_1,C_2	QENEW’s output is the final run
IMC-Q-C-4S	C_1,C_2,Co	QENEW’s output is the final run. Ranking method utilizes word frequency feature, that is α_1=1,α_2=α_3=0
IMC-Q-C-5S	C_1,Co	QENEW’s output and C_2 use topic model to generate the final run.

Results

Table1: Overall subtopic mining results

RunID	**I-rec@10**	**D#-nDCG@10**	**D-nDCG@10**
IMC-Q-C-1S	0.5685	0.5181	0.4677
IMC-Q-C-2S	0.6172	0.5798	0.5424
IMC-Q-C-3S	0.4403	0.4349	0.4294
IMC-Q-C-4S	**0.6240** 3/16	**0.5869** 2/16	**0.5498** 2/16
IMC-Q-C-5S	0.4325	0.3890	0.3456

CONCLUSIONS

- ➢ Generate query expansion terms based on new words extraction theory
- ➢ The method employ the information entropy theory and statistical language knowledge to measure the words’ features