HUKB at NTCIR-12 IMine-2 task: Utilization of Query Analysis Results and Wikipedia Data for Subtopic Mining

Masaharu Yoshioka e-mail: yoshioka@ist.hokudai.ac.jp
Graduate School of Information Science and Technology, Hokkaido University

Motivation

- Subtopic mining for supporting users to find out more focused retrieved results
- Evaluation of external resource for subtopic mining candidates : query analysis results and Wikipedia

Approach

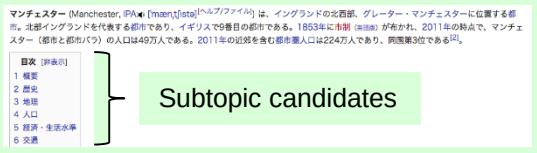
- Selection of subtopic candidates from external resource
  - Wikipedia : Yahoo! abstract and disambiguation
  - Query analysis results: Query suggest, co-topic and co-click (provided by organizers)
  - Keywords that characterize each topic of the topic model (LDA)
- Evaluation of appropriateness of the subtopic candidates
  - Use topic model for check diversity
  - Check representativeness of the candidates by using number of articles in the initial retrieval results
- Vertical intent analysis
  - Subtopic keyword base
  - Analysis of the type of retrieved pages

Subtopic candidates

- Wikipedia
  - Check existence of the disambiguation articles (that belongs to Wikipedia category “曖昧さ回避”(disambiguation))

Use article title in the list for subtopic candidates
e.g., Concurrent Versions System, コンビニエンスストア, CVS/ファーマシー… for CVS

  - Yahoo! abstract: short description + list of chapters



- Query Analysis
  - Query suggestion data provided by organizers (Bing, Google, Yahoo!)
- Keywords from topic model
  - When the system fails to generate 10 subtopics from above candidates, top 10 characteristic keywords are examined for each topic in the topic model (LDA)

Vertical intent

- Keyword base estimation
  - Comparing subtopic candidate and original query to keywords
  - Image is verified existence of original keyword in Alt of Img tag

Intent	Keywords
Image	イラスト(illustration), 画像(image), 絵(picture), 似顔絵(portrait), 写真(Photo), 壁紙(Wallpaper)
News	news, ニュース(news), 記事(article)
Shopping	注文(Order), 価格(price), 値段(price), 料金(price), 相場 (market price), 費用(Cost), 安い(Cheap), 格安(Very cheap)
QA	質問(Question), 方法(Method)
Encyclopedia	内容(Contents), 意味(Meaning), 知識(Knowledge), 辞書 (Dictionary), Wikipedia
Web	動画(Video), Youtube

- Check by retrieved results
  - Shopping: existence of keywords 注文(order), 支払(payment), 送料 (shipping charge), 買い物カゴ(cart) in the page. (5 or mote than half)
  - News: check host of url list with news site list constructed by using Open Directory and Google news (10 or more than half)
- Rest are categorized as Web

Selection of subtopic candidates

- Target document
  - Initial retrieval results provided by organizers
  - Snippet based selection for two or more words query (all query keywords should exist in a given text window; calculated by snippet selection algorithm)
- Usage of topic model (LDA)
  - Construct LDA topic model with topic size = 30
  - Variational Dirichlet parameters for each document (did) is used for evaluating representativeness of the subtopic in the document. $dtr_{tid,did}$
  - Variational Dirichlet parameters for each subtopic candidate (sid) is used for evaluating representativeness of the subtopic in the document. $qtr_{tid,sid}$
  - Document set related to tid $Dt_{tid} = \{did | dtr_{tid,did} > 0\}$
  - Document set retrieved by sid:Dr_{sid}
  - Three groups for subtopic candidates
    - Candidates from Wikipedia and query analysis with higher $qtr_{tid,sid} > 0.3$
    - Other candidates from Wikipedia and query analysis
    - Keywords from topic model
  - Sort subtopic candidates by Rtrep for each group and pick candidates from the first group
    - Exclude candidates whose representative topics are also selected or whose retrieved results are similar to the selected one.

$$Frep_{tid,sid} = \frac{|Dt_{tid} \cap Dr_{sid}|}{|Dt_{tid} \cup Dr_{sid}|} \times |Dt_{tid}|$$

$$Rt_{sid} = \underset{tid \in \{tid | qtr_{tid,sid} > 0\}}{\operatorname{argmax}} Frep_{tid,sid}$$

$$Rtrep_{sid} = Frep_{Rt_{sid},sid}$$

Disucssion

- Candidates from query analysis is a good resource for subtopic candidate generation, but ones from Wikipedia is not good, even though those candidates seems to be reasonable subtopic
- It may be necessary to take into account the representativeness of the subtopic.

Experimental Results

- Submitted runs
  - Different target document all (1Q,2Q,3Q) and snippet based selection(4Q,5Q)
  - Different candidates
    - All candidates: 1Q+4Q
    - Query suggestion + Wikipedia: 2Q+5Q
    - Wikipedia only: 3Q

	1Q	2Q	3Q	4Q	5Q
I-rec@10	0.646	0.632	0.497	**0.653**	0.645
D-nDCG@10	**0.507**	0.475	0.368	0.505	0.470
D¥#-nDCG@10	0.576	0.553	0.433	**0.579**	0.557
V-Score	**0.535**	0.480	0.384	**0.535**	0.481
QU-Score	**0.556**	0.517	0.408	0.557	0.519

- Number of subtopic candidates used in 1Q

	Selection	Unique selection
Yahoo coclick	260/368 (0.71)	145/218 (0.67)
Yahoo cotopic	318/447 (0.71)	145/217 (0.67)
Query suggestion	220/314 (0.70)	80/124 (0.65)
Wikipedia abstract	23/41 (0.56)	21/39 (0.54)
Wikipedia alternatives	5/15 (0.33)	3/10 (0.3)
Topic model	36/115 (0.31)	36/115 (0.31)