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

Masaharu YOSHIOKA

Hokkaido University

Background and Objectives

- Background
  - Subtopic mining for supporting users to find out more focused retrieved results
  - Utilization of user query log and external resource for subtopic mining
- Objectives
  - Evaluation of external resource for subtopic mining candidates
    - query analysis results
    - Wikipedia

Approach

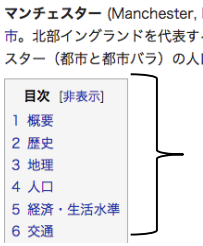
- Selection of subtopic candidates from external resource
  - Wikipedia and query analysis result
- Evaluation of appropriateness of the subtopic candidates
  - Use topic model for checking diversity and representativeness of the candidates
- Vertical intent analysis
  - Subtopic keyword base
  - Analysis 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


マンチェスター (Manchester, IPA: ['mæn,tʃɪstə][ヘルプ/ファイル]) は、イングランドの北西部、グレーター・マンチェスターに位置する都市。北部イングランドを代表する都市であり、イギリスで9番目の都市である。1853年に市制 (英語版) が布かれ、2011年の時点で、マンチェスター（都市と都市パラ）の人口は49万人である。2011年の近郊を含む都市圏人口は224万人であり、同国第3位である[2]。

目次 [非表示]
1 概要
2 歴史
3 地理
4 人口
5 経済・生活水準
6 交通

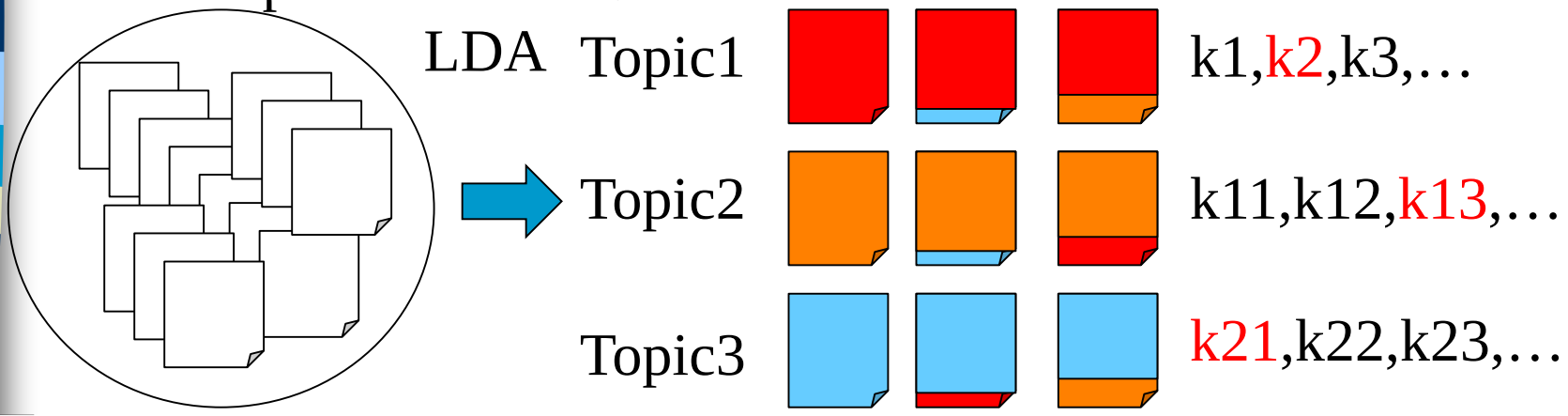
Subtopic candidates (概要,歴史,地理,人口,…)

Subtopic Candidates (Query Analysis)

- Query suggestion data provided by organizers (Bing, Google, Yahoo!)
- Query analysis data: coclick and cotopic (30 candidates from each data set)
- 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)

Selection of Subtopic Candidates

- Usage of topic model (LDA)
  - 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)
  - Select candidates based on its representativeness in the topic documents.



Assign Subtopic Candidates for Representative Topic of Topic Model

- Construct LDA topic model with topic size = 30
- Representativeness of the keyword for the subtopic
  - $dtr_{tid,did}$:Variational Dirichlet parameters for each document (did)
  - $Dt_{tid} = \{did | dtr_{tid,did} > 0\}$
  - Dr_{sid}:Document set retrieved by sid

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

Assign Subtopic Candidates for Representative Topic of Topic Model

- Assign most representative topic for each subtopic candidate
  - $qtr_{tid,sid}$:Variational Dirichlet parameters for each subtopic candidate (sid)

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

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

Selection of Subtopic candidates

- 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.

Vertical Intent Estimation

- Subtopic 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

Vertical Intent Estimation(cont.)

- 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)
- Rests are categorized as Web

Experimental Results

- Implementation of the system
  - Tokenizer: JUMAN normalization by using utilization of 代表表記 (Normalized form)
  - Document retrieval system: groonga that supports snippet generation, phrase based retrieval http://groonga.org
  - LDA: LDA implemented by Prof. Blei. http://www.cs.princeton.edu/~blei/lda-c/

Variation of Submitted Run

- Target document
  - All retrieved documents provided by organizers: Q1,Q2,Q3
  - For the queries with two or more keywords, documents that don’t have all query keywords in given window (checked by snippet) are excluded: Q4, Q5
- Subtopic candidates
  - All candidates (Query suggestion, Query analysis, Wikipedia) : Q1, Q4
  - Query suggestion + Wikipedia: Q2, Q5
  - Wikipedia only: Q3

Evaluation Results

- J-3Q (Wikipedia based candidates only) is significantly worse than others ($p < 0.01$)
- Utilization of Query analysis results is slightly improve the performance but it is not significant
- Target document selection is almost no effect

	J-1Q	J-2Q	J-3Q	J-4Q	J-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

Resource Used for Subtopic Candidates(J-1Q)

- Subtopic candidates from query analysis is frequently selected for representative subtopic
- Most of subtopic candidates selected from Wikipedia is not selected as oracle subtopic candidates

	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)

Discussion

- Candidates from Wikipedia is not a good one, even though those candidates seems to be reasonable subtopic.
- It may be necessary to take into account the representativeness of the subtopic.

Summary

- Query analysis results is a good resource to estimate good subtopics.
- However, candidates from Wikipedia seem to be reasonable subtopic, but it is not selected as oracle subtopics.
- Further analysis including failure analysis, effect of parameter is necessary.