NTCIR-8 GeoTime at Osaka Kyoiku University

— Hierarchical Index for Geographic Retrieval —

Takashi SATO Yuu FUKUZAWA (Osaka Kyoiku University)

[1] Overview

- Made n-gram Index, Temporal Index, and Geographic Index for J-J subtask.
- Geographic Hierarchical Index is made from ZIP code of Japan Post Group.
- Using SPYSEE (person retrieval site), person's names are extracted from topics.
- Confirmed that the effect of the geographic hierarchical index when topics included term of wide area region.

[2] Temporal Index

- Extracted temporal information of the following form.
  - (1) **年　(2) **年**月　(3) **年**月**日
  - (4) **月　(5) **月**日　(6) **日

- Search noise occur when (1), (4), and (6). Then, terms which are proceeded or followed by specific characters were excluded from the index.

Position	Character
proceeded	約, 今後, 過去, 懲役, 震災
followed	間, 前, 後, 中, ほど, 程, 先, 以上, 以内, 未満, 連続, ぶり, 代

[3] Geographic Index

- MeCab analyses a sentence including geographic information “アメリカのニューヨークで・・・”.

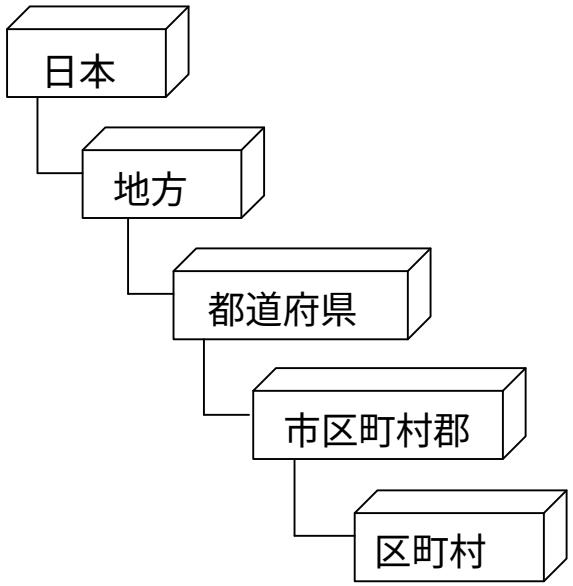
```
%mecab

アメリカのニューヨークで・・・
アメリカ　　名詞, 固有名詞, 地域, 国, *, *, アメリカ, アメリカ, アメリカ
の　　助詞, 連体化, *, *, の, ノ, ノ
ニューヨーク　　名詞, 固有名詞, 地域, 一般, *, *,ニューヨーク,ニューヨーク,ニューヨーク
で　　助詞, 格助詞, 一般, *, *, で, デ, デ
```

- The region is analyzed as "国(country)" and "一般(general regions)".
- Using these analyses, a country index and a general region index were made.

[4] Geographic Hierarchical Index

- Also made an index which represents hierarchical structured of the geographic information. We used the Japanese geographic hierarchy shown because we used Japanese Mainichi news as Collection.

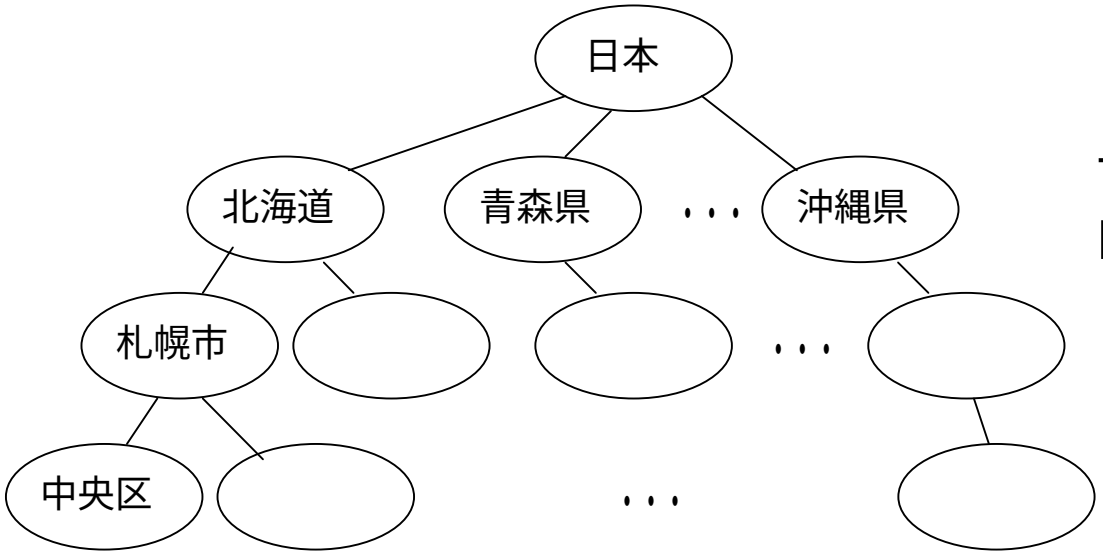


Geographic Hierarchy

- The hierarchical structure was made by the ZIP code of Japan Post Group.

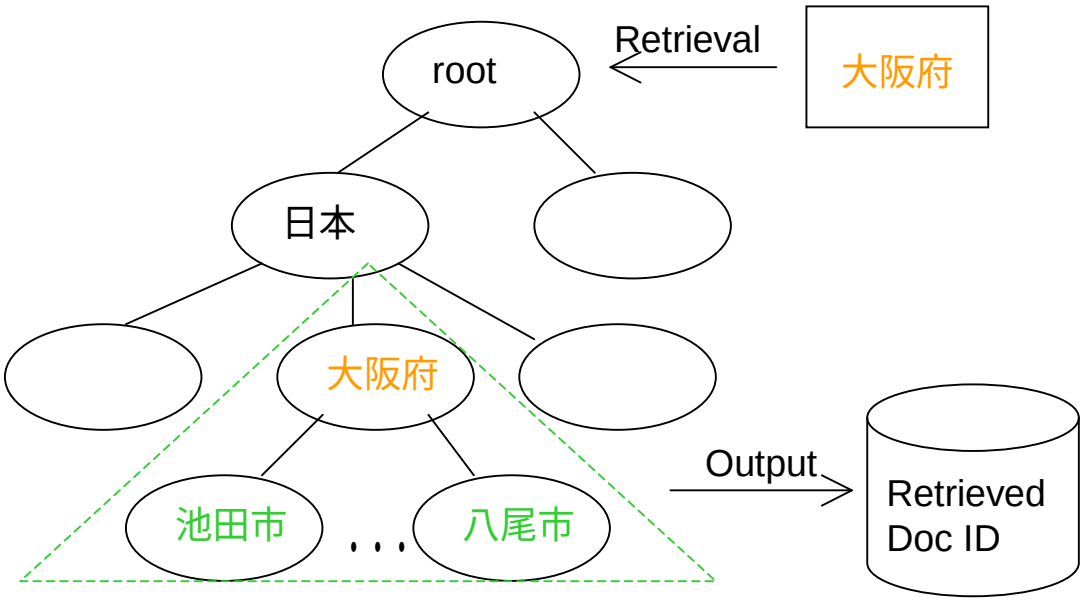
ZIP	prefecture	city, word, town, village	town, region
064-0941	北海道	札幌市中央区	旭ヶ丘
060-0041	北海道	札幌市中央区	大通東
060-0042	北海道	札幌市中央区	大通西 (1-19丁目)

Part of ZIP code of Japan Post Group



The example of geographic hierarchical structure

- The result of query, which includes wide area region term, is the sub tree of which root matches the term.



- For the case when the same region is expressed in different such as "アメリカ" and "米国", we regulated them using Table below.

Region Name	Regulated Region Name
米 米国 アメリカ合衆国 合衆国 U.S.A. U.S.	アメリカ
欧州	ヨーロッパ
英 英国	イギリス
仏	フランス
中 中華人民共和国	中国
日	日本
独	ドイツ
伊	イタリア
韓	韓国

[5]Term Extraction from Topics

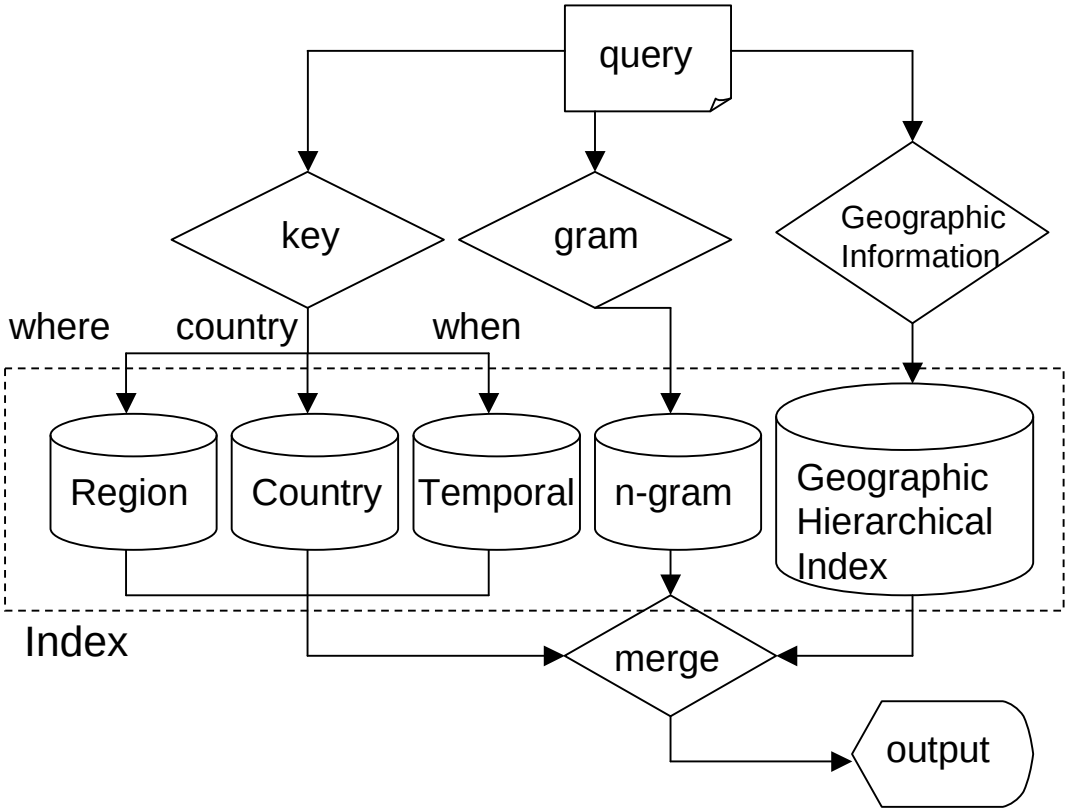
- Extraction of Retrieval Term
  - Extracted retrieval terms from the NARRATIVE tag of TOPICS. Because NARRATIVE sentences are short (around two rows), we do not put different weight between retrieval terms by frequency.
- Extraction of Person's Name
  - In the morphological analysis, the name of a person was not properly analyzed. Therefore, we judged that the term is the name of a person when it matches to the name of a person retrieval site SPYSEE. The word judged to be a name of the person increases weight by a factor of ten. The example of <TOPIC ID="GeoTIme-0001"> is shown.

Term	Weight
アストリッド・リンドバーグ	0.769230
都市	0.076923
児童書作家	0.076923
死亡	0.076923

Example of Term Weight Including Person's Name

[6] Experimental Results

- Retrieval System
  - Made each index of *n*-gram, temporal information, country name, regional name, and geographic hierarchy from the collection.
  - Figure shows our retrieval system.



- Query Using Geographic Hierarchy
  - No query using a geographic hierarchy in GeoTime TOPICS.
  - Prepared additional query "近畿地方の積雪について知りたい (I wanted to know the snowfall in the Kinki region)".
  - Against a wide area of region Kinki, we confirmed that regions of lower hierarchy of Kinki were retrieved.
  - For instance, <DOCNO>JA-020212127</DOCNO> includes name of prefectures in the Kinki province "Shiga Prefecture", "Hyogo Prefecture", and "Kyoto Prefecture" though this document doesn't contain the word of "Kinki" province.
  - Effectiveness was confirmed by being retrieved it in 2nd place when using a geographic hierarchical index though it was 30th place when it was not used.
- Analysis of Results
  - We obtained good precision in the first half of topics whose relevant documents are few.
  - However, precision is low even in comparatively easy topics when there are many relevant documents.
  - In addition to the retrieval of individual index such as *n*-gram, geographic, and temporal index, we should have merged their similarity more carefully.

[7] Conclusions

- Retrieved topics that contained the geographic and temporal information at NTCIR-8 GeoTime task.
- Temporal and geographic information are extracted from GeoTime collection.
- The index that represents a geographic hierarchy is made from the geographic information.
- In the experiment, we confirmed that the effect of the geographic hierarchical index when topics included term of wide area region.