NTCIR-8 GeoTime at Osaka Kyoiku University - Hierarchical Index for Geographic Retrieval -

SATO, Takashi
Information Processing Center
Osaka Kyoiku University
4-698-1 Asahiga-oka
Kashiwara, Osaka, JAPAN
+81-72-978-3823
sato@cc.osaka-kyoiku.ac.jp

FUKUZAWA, Yuu
Department of Arts and Sciences
Osaka Kyoiku University
4-698-1 Asahiga-oka
Kashiwara, Osaka, JAPAN
+81-72-978-3823
fukuyuu@ss.osaka-kyoiku.ac.jp

ABSTRACT

We retrieved topics that contained the geographic and temporal information at NTCIR-8 GeoTime task. Employing morphological analysis, 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.

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information Search and Retrieval – *Information filtering, Query formulation, Search process.*

General Terms

Experimentation, Performance, Measurement.

Keywords

Information Retrieval, Index, Geographic Hierarchy.

1. INTRODUCTION

We retrieved topics that contained the geographic and temporal information at NTCIR-8 GeoTime task. Employing morphological analysis, temporal and geographic information are extracted from GeoTime collection. The index that represents a geographic hierarchy is made from the geographic information.

As for retrieval, weight of terms is increased if they match the name of person. We expand the terms by using Weblio(the thesaurus dictionary site).

In the experiment, we confirmed that the effect of the geographic hierarchyical index when topics included term of wide area district.

2. INDICES

Added to *n*-gram (long and varying length gram coded in fix bytes) based indices[1], which our group usually use for information retrieval, we made time and geographic indices using MeCab[2] morphological analysis system..

2.1 Temporal Index

We extract temporal information of the following form from morphological analysis.

(1)**年　(2)**年**月　(3) **年**月**日　(4)**月
(5) **月**日　(6) **日

However, the search noise occur when (1), (4), and (6) are used to represent the time width like "Ten years". Then, terms which are proceeded or followed by characters gave on Table 1 were excluded from the index.

Table 1. Filter of Temporal Information

Position	Character
Before (1),(4),(6)	約, 今後, 過去, 懲役, 震災
After (1),(4),(6)	間, 前, 後, 中, ほど, 程, 先, 以上, 以内, 未満, 連続, ぶり, 代

2.2 Geographic Index

If MeCab analyses a sentence including geographic information "アメリカのニューヨークで・・・", its output becomes as shown in Figure 1. 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.

```
%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
```

Figure 1. Example of Morphological Analysis of Geographic Information

2.3 Geographic Hierarchical Index

We also made an index which represents hierarchical structured of the geographic information. We used the Japanese geographic hierarchy shown in Figure 2 because we used Japanese Mainichi news as Collection.

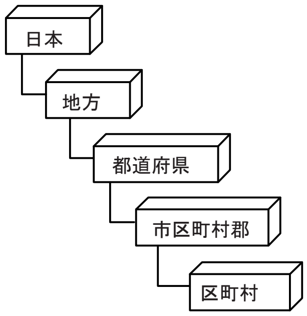

Figure 2. Geographic Hierarchy

The hierarchical structure was made by the ZIP code of Japan Post Group[3]. We quote a part of it in Table 2. The example of geographic hierarchical structure represented by tree is shown in Figure 3.

Table 2. Part of ZIP Code of Japan Post Group

ZIP	Prefecture	City, Ward, 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)

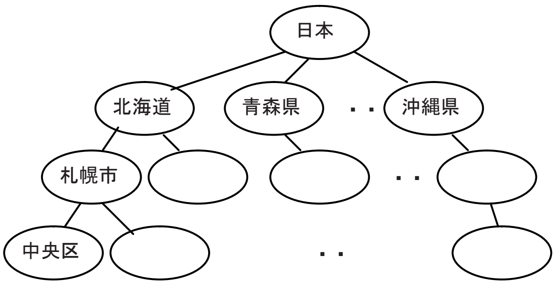

Figure 3. Example of Geographic Hierarchical Tree

Moreover, 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 3.

Table 3. Regulation Filter for Geographic Information

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
加	カナダ
露	ロシア
朝 朝鮮	朝鮮人民共和国
豪州	オーストラリア

The result of query, which includes wide area region term, is the sub tree of which root matches the term(See Figure 4).

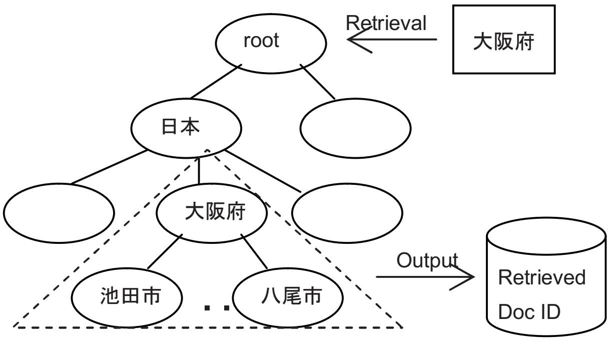

Figure 4. Retrieval of Geographic Hierarchical Tree

3. TERM EXTRACTION FROM TOPICS

3.1 Extraction of Retrieval Term

We extracted retrieval terms from the NARRATIVE tag of TOPICS. Because NARRATIVE sentences are short (around two rows), we not put different weight between retrieval terms by frequency.

3.2 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[4]. 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 in Table 4.

3.3 Thesaurus Expansion

The terms have been enhanced by using the thesaurus of dictionary site Weblio[5]. Table 5 shows example of <TOPIC ID="GeoTIme-0001">.

Table 4. Example of Term Weight Including Person's Name

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

Table 5. Example of Term Expansion

Term	Expanded Term
都市	とし 大都市 市街地 都会
死亡	しぼう 亡 卒去 憤死 死 死去 死因 物故 病死 絶命 逝去

4. EXPERIMENTAL RESULTS

4.1 Indexing

We made each index of *n*-gram, temporal information, country name, regional name, and geographic hierarchy from the collection. Figure 5 shows our retrieval system. The indexing time is indicated in Table 6.

Table 6. Indexing Time

Index	Time(sec)
n-gram	114
Temporal and Geographic	809
Geographic Hierarchical	3,046

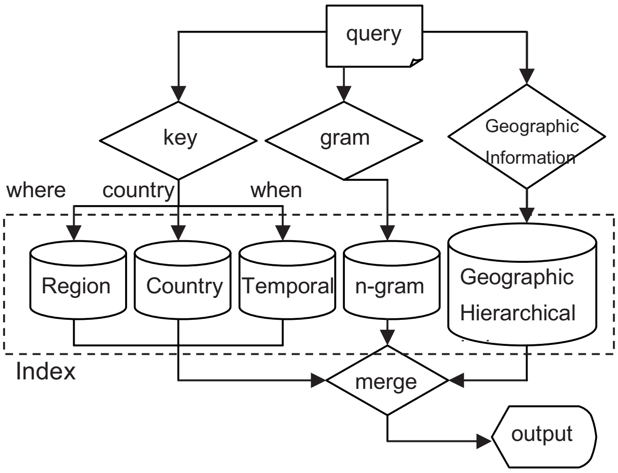

Figure 5. Retrieval System of OKSAT

4.2 Query Using Geographic Hierarchy

Because there was no query using a geographic hierarchy in GeoTime TOPICS, we 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.

5. ANALYSYS 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. Therefore, in addition to the retrieval of individual index such as n-gram, word, geographic, and temporal index, we should have merged their similarity more carefully. Although we increased the weight of proper nouns such as name of person or location in topics, we should have recognized them more precisely by using much more online dictionaries etc.

6. CONCLUSIONS

We retrieved topics that contained the geographic and temporal information at NTCIR-8 GeoTime task. Employing morphological analysis, 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.

7. REFERENCES

[1] Sato, T., Fast full text retrieval using gram based tree structure, *Proc. ICCPOL '97*, Vol.2, pp.572-577 (1997).

[2] MeCab, http://mecab.sourseforge.net/

[3] JP, http://www.post.japanpost.jp/

[4] SPYSEE, http://spysee.jp/

[5] Weblio, http://thesaurus.weblio.jp/