Visualization of tourism information using WordNet

Nakatoh, Tetsuya
Research Institute for Information Technology, Kyushu University

Yin, Chengjiu
Research Institute for Information Technology, Kyushu University

Hirokawa, Sachio
Research Institute for Information Technology, Kyushu University

Matsuura, Hiroki
Graduate School of Information Science and Electrical Engineering, Kyushu University

出版情報：Proceedings of 2011 3rd International Conference on Awareness Science and Technology, iCAST 2011, pp.412-417, 2011-12-01
バージョン：
権利関係：

Visualization of Tourism Information using WordNet

Tetsuya Nakatoh*, Chengjiu Yin*, Hiroki Matsuura† and Sachio Hirokawa*
*Research Institute for Information Technology
Kyushu University
Hakozaki6-10-1 Fukuoka JAPAN 811-8581
Email: {nakatoh,yin,hirokawa}@cc.kyushu-u.ac.jp
†Graduate School of Information Science and Electrical Engineering
Kyushu University
Hakozaki6-10-1 Fukuoka JAPAN 811-8581
Email: matsuura.hiroki.775@s.kyushu-u.ac.jp

***Abstract*—In recent years, with the development of the Internet, and the rapidly increasing number of tourism portal sites and blogs, we can obtain a variety of tourist information on the Internet. If we have a specific need, we can obtain the required information through checking the retrieved results one by one. However, in order to see the whole trend of the results, it is necessary to analyze and visualize the document group. In this paper, by storing a set of tourism-related future words in WordNet, which is a concept thesaurus, we proposed a method using information from WordNet to visualize a document group as a conceptual graph. With the structure of a thesaurus, this method enables us to understand the contents of document group at a glance. Furthermore, by comparing the thesaurus structures, which are obtained from different background document groups, we can grasp the differences.**

I. Introduction

Generally, a tourist resort has each special feature. In order to draw tourists, hotels and tour companies are making an effort to make many people know such special features through media, campaign, etc. Moreover, in recent years, much tourism information is offered in tourism portal sites to those who gather tourism information on WWW. Probably, tourists are actually enjoying the special feature of each tourist resort investigated by WWW.

On the other hand, the tourist who actually went to the tourist resort may discover the special feature which generally is not known, and may enjoy it. Or they may have evaluation which is different from information about the known special feature. Such the special feature and information that were experienced by the tourist were shared only among very close persons. However, they are exhibited by blog articles, such as a travel record, now in many cases. Such information is profitable also for people who want to choose the tourist resort which should be visited, the tourist agent who wants to improve the quality of service, and the local government which wants to find out the special feature of new tourism.

Since there is little information by a tourist's experience exhibited by a blog article as compared with the information on other, it is hard to be found out. However, we have already reported the difference in the type of the information between the formal tourism information which a local government offers, and the tourism information written on blog articles in AROB [1]. Thus we think that comparison of the document group from which a background differs can also extract the information by a small number of tourist's experience.

We have proposed the method of visualizing and analyzing the information on a document group using the graph of the relation between the concepts [2], [3]. In this paper, we propose the method of visualizing and analyzing the information on a document group by mapping the feature word on WordNet which is a concept thesaurus. Thereby, contents of the document group can be understood at a glance. Furthermore, the difference in the information on documents can be directly identified by mapping the information in the same thesaurus. Tourism information was visualized for the verification. By using the proposal system, the difference between hotels and Japanese inn, and the difference between Kyushu 7 prefectures were illustrated using the thesaurus.

II. Related Work

The tourism industry is one of the industries that suffered big influence of internet. Sightseeing information was available only in a special travel agent before. Now, everyone can easily obtain them thanks to the Internet. We can find sightseeing information on Web in (a) tourism portal sites, in (b) general web pages, and in (c) blog sites. There are several systems and researches intended for each targets. [4] proposes a recommendation and a clustering system and shows their effectiveness for tourism portals. [5] proposes a natural language interface for tourism search engine. [6] shows "keymaps" that visualizes co-occurrences of keywords in tourism documents. [7] analyzes the patterns in HTML documents that characterize the occurrences of NEs(Named Entity), such as the name of the location and the name of the touristic events. [8], [9] study the clue words that can be used to extract tourism related NES. [10] reports the characteristic keywords that distinguish tourism blogs from other general blogs. Most purposes of these existing studies are extraction of tourism information. It is important in order for a tourist

to gather the information on a travel. On the other hand, there is demand to the overall analysis of tourism information. [11] proposes the method to extract and classify strong points in sightseeing area as support techniques to develop sightseeing area. We are aiming at the construction of the overall analysis engine for blogs of tourism information.

We use **Japanese WordNet** [12] which translated **WordNet** [13] into Japanese as a model of a concept thesaurus in this paper. As research using WordNet, there are research on an automatic classification of documents and research on query expansion. [14] used synonym in WordNet as feature data for the document classification by machine learning. [15] made classification performance improve further using hypernym of WordNet. [16] performed query expansion using synonym and hypernym of WordNet. Furthermore, [17] proposed the method of extracting the word for query expansion from the semantic definition of Wordnet.

[18] visualized document data with the potential topic structure. It is visualization of each document and the topic, and the target differs from this proposal which is visualization of a document group. [19] proposed the method of constructing concept graph from a document group. Concept graph is constructed by determining the hierarchical order of the feature word contained in a document group according to the frequency of appearance. It differs in a method from this proposal which visualizes a document group using the hierarchical order of a thesaurus.

III. Basic Data

In this paper, we used 1303 blog articles about tourism of Kyushu area linked from the Kyushu Tourism Promotion Organization[1]. Section III-A describes our method that extract the content from each blog article. Section III-B shows basic analysis of the data obtained from the contents.

A. Wrapper with URL Prefix and Path Prefix

The tourism information used by our preceding studies [1], [20], [21] was 312 event information offered in the Kyushu Tourism Promotion Organization. Since each of that event information was described by the common template, the contents have been extracted by making an exclusive wrapper. On the other hand, 1303 blog articles that are the targets of this paper are on 145 individual sites, and they were not written by a common template. Therefore, extraction of contents from each blog article is not easy. There are a lot of existing researches(e.g., [22]) and tools[2] in the field of the generation of the rapper that extracts demanded contents from the Web document. However, the technology to an individual example is needed because there is no versatile method. Actually, the blog article on a single site did not necessarily use a common template for the target blog of this paper also. We generated the wrapper automatically using a URL prefix and a path prefix in this paper. For example, URL of each article in a site with most articles had the following common prefixes. "`/mt/archives/`*year*`/`*month*`/post`*No*`.html`".

[1] `http://www.welcomekyushu.com/`

[2] `http://search.cpan.org/~miyagawa/Web-Scraper/lib/Web/Scraper.pm`

TABLE I
Blog Sites

Ranking	Number of Articles	URL of Site
1	84	http://asobo-saga.jp
2	80	http://colors1.blog32.fc2.com
3	77	http://blogs.yahoo.co.jp
4	69	http://smplus.jp
5	69	http://betsubala.seesaa.net
6	69	http://ameblo.jp
7	62	http://cottonblog.seesaa.net
8	54	http://blog.livedoor.jp
9	51	http://bakudankozo.blog118.fc2.com
10	50	http://yamatoimpulse.blog50.fc2.com
11	47	http://blog.goo.ne.jp
12	40	http://tokaipia.com
13	33	http://siosai.cocolog-nifty.com
14	29	http://www.okota.net
15	28	http://plaza.rakuten.co.jp
16	24	http://matsuno.otemo-yan.net
17	23	http://mojako.miyachan.cc
18	22	http://hotomeki.blog68.fc2.com
19	21	http://blog.baliyoka.net
20	19	http://yakinikufukuoka.blog40.fc2.com

20 high ranks of the sites with a lot of articles are shown in Table I.

The HTML files with such a common prefix were made into one group. The contents of the blog articles of the group can be extracted by X-path (`/html/body/div[2]/`). Such a path is calculated by the following methods. The number of HTML files in a group is assumed to be n. All HTML files in the group are converted into the HTML trees. Let t_i be i-th HTML tree of them (where $1 \geq i \geq n$). The path from root to a node is expressed as p. Let $tail(t_i, p)$ be the number of child nodes of p in t_i. Now, $tail(p)$ is calculated as a total of $tail(t_i, p)$ of HTML tree t_i of the group, i.e., $\mathrm{tail}(p) = \sum \mathrm{tail}(p, t_i)$. Common-Template-Likeness of path p was defined as $\mathrm{score}(p) = \mathrm{tail}(p)/\mathrm{depth}(p)$. Then, path p wherein $score(p)$ is the maximum has been extracted as a common template. Therefore, contents of 1303 tourism blog articles were collected.

The examples of extraction path are shown in Fig. 1.

B. Basic Analysis

The target of the analysis is set of tourism blog articles in Kyushu area. First, the feature word of each prefecture has been extracted as a basic analysis. They are shown in Table II. The word with the mark * is a proper noun, and a lot of famous places in each prefecture appear in them. However, a mutual comparison is difficult in this table.

Table III shows the most frequent 20 feature words. It contains a lot of general words related to the blog such as comment, track-back, page, diary and post. Moreover, it

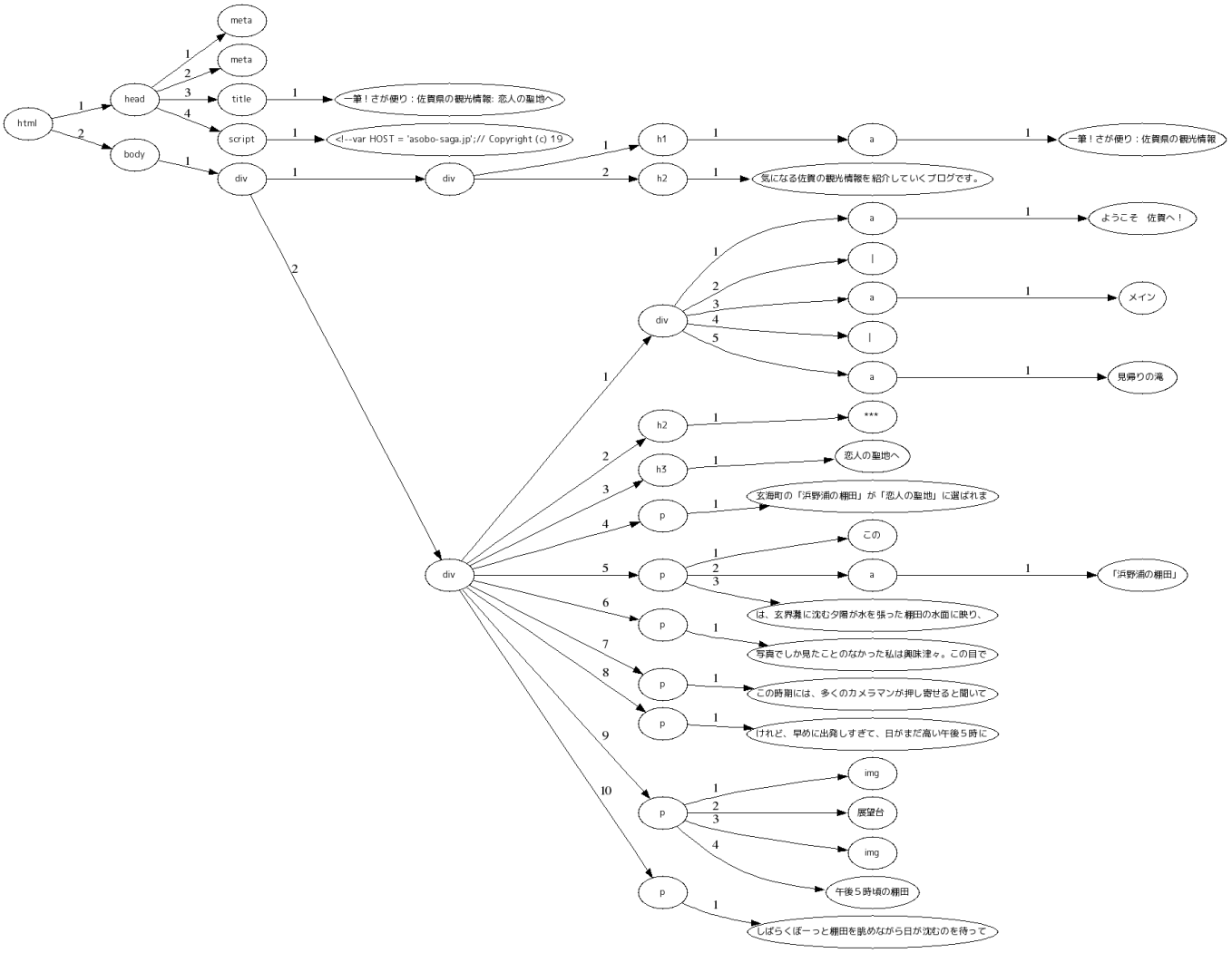


Fig. 1. Path which Specifies the Contents

contains also quite many food-related feature words such as shop, eat, noodle, taste and bite.

TABLE III
MOST FREQUENT 20 FEATURE WORDS

photo(743), shop(667), comment(587), page(520), back(514), feel(512), track(512), diary(509), eat(509), noodle(488), look(466), good(464), people(448),post(446), before(433), taste(430), blog(396), time(390), information(387), bite(383)

IV. FEATURE WORDS ON WORDNET

The feature words extracted in Sec. III-B contain the feature of each prefecture in Kyushu area, such as a tourism resort. However, they are insufficient to understand the relation of the feature words, and the overall feature including it. In this paper, we propose the visualization system of the feature words by putting them on WordNet. Then, we can understand not only the feature but those interrelations at a glance.

The hierarchical structure of concepts on WordNet is very large, and complicated. Therefore, even if the all are visualized, we can not understand it at a glance. Then, it is necessary to choose the part to visualize appropriately. Moreover, a concept on WordNet and a word do not have the relation of 1 to 1. Therefore, we have to assign a word appropriately to a concept. This section shows the details.

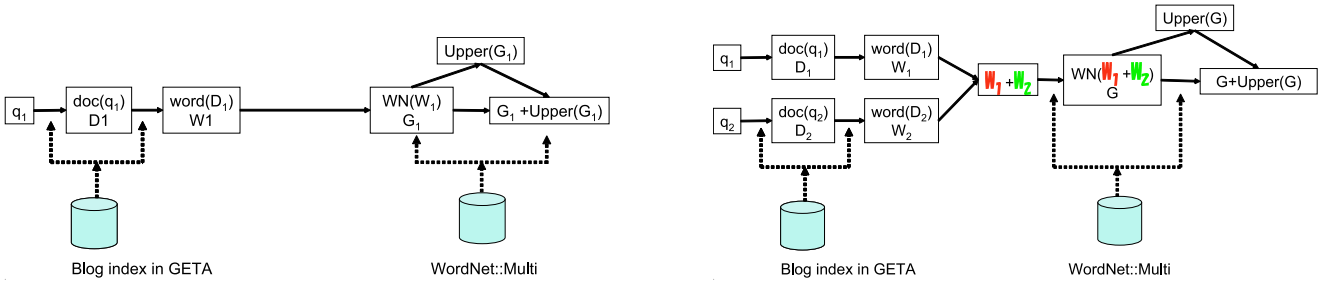
A. single graph

The feature words extracted from the document set are made visible according to the thesaurus of WordNet. The document set is the search results obtained by the query according to the purpose.

The concept in WordNet has **synset** (synonymous-words set), and is described by peculiar ID. When feature word w is contained in synset of the concept p, the p is used as the node of w. Although each node has $\mathrm{label}(p)$ displayed, $\mathrm{label}(p) = w$ in this case. Generally p has superordinate concepts. A set of the adjacent superordinate concept of p is described as $\mathrm{wnUp}(p)$. In concept p' obtained not as a

TABLE II
FEATURE WORDS IN EACH PREFECTURE

Prefecture	# article	Feature Words
KUMAMOTO	186	*KUMAMOTO “press navi” sponsor *YAMATO *MOTO castle *KUMA SAN
FUKUOKA	178	*FUKUOKA *HAKATA ward *TENJIN central TEL Influenza town report program-king
NAGASAKI	112	*NAGASAKI haku chinese *DEJIMA CHAMPON-noodle *CHINA magistrate *SUWA
OITA	93	*OITA *YUFUIN *BEPPU *SAITAMA berry sex *amaba-blog-top “Heavens whole country” diary
MIYAZAKI	219	*MIYAZAKI *ZAKI camp epidemic hoof *KOKUBARU smell haste male
KAGOSHIMA	202	*KAGOSHIMA communication sending *DUSKIN *SATSUMA *SHUICHI usefull *KUSHIKINO absorption category
SAGA	116	*SAGA “land near the castle” doll “kokoiko” festival main *KANZAKI



(a) single graph (b) comparative graph

Fig. 2. Structure of System

feature word but as the adjacent superordinate concept of p ($p' \in \mathrm{wnUp}(p)$), $\mathrm{label}(p')$ is the first word of **synset** which p' has.

Construction of a graph mainly consists of two processes. First, let W be the highest score K feature words. The superordinate concepts of each feature word in W are collected within the range of W. As a result, a hierarchical relation of feature words is obtained. Next, the superordinate concepts of each feature word in W are also collected exceeding the range of W. Thereby, the relation of the feature words which did not have the hierarchical order within W is obtained.

As mentioned above, correspondence of a concept and a word is not necessarily 1 to 1. Therefore, the word displayed on different concept nodes may become the same. It is not desirable as a graph which shows the relation of words. Two following procedures were used to prevent this. (1) About the subroutine `Up` which adds a branch between present node p and the superordinate node r, the branch is added only in $\mathrm{label}(p) \neq \mathrm{label}(r)$. (2) When the node p and q have same label ($\mathrm{label}(p) = \mathrm{label}(q)$) and they have common adjacent superordinate concept r, two branch $r \leftarrow p$ and $r \leftarrow q$ cannot be distinguished. Therefore, q and $r \leftarrow q$ are deleted by $\mathrm{CutFork}$.

The outline of this system is shown in figure 2(a), and the algorithm is shown in figure 3.

B. comparative graph

In order to compare two search results, we construct graph individually from them by Sec. IV-A, and may just look at them simultaneously. However, more intelligible visualization can be performed by merging these two graph by the same node. The merge is possible at the following processings. First, the set union of each K feature word set is made the new feature word set W. The whole graph is generated about this W. Furthermore, in order to visualize the difference between two graphs clearly, they are classified by color. The color of each node is calculated as follows. Let U and V be two search-results document groups to compare. Each node is concept p_i in WordNet, and has set of words $w_1, w_2, ..., w_n$ as synset. About a certain node, appearing probability $\Pr(u, w_i)$ of the word w_i in U is compared with appearing probability $\Pr(v, w_i)$ of the wotd w_i in V. The node with many words that satisfy $0.8\Pr(u, wi) > \Pr(v, wi)$ displays in red, and the node with many words that satisfy $0.8\Pr(u, wi) < \Pr(v, wi)$ displays in green. The node with many words which have few differences of appearance ratio is displayed in gray.

The outline of this system is shown in figure 2(b).

V. CASE ANALYSIS AND CONSIDERATION

This section explains the comparison and analysis by this system using examples. The used data is 1309 blog articles about the tourism on Kyushu area registered into WebPage of the Kyushu Tourism Promotion Organization, as explained in Sec. III.

A. Comparison of Japanese Inn and Hotel

Fig. 4(a) is visualization of the search results which used the query “ryokan”(Japanese inn) by our proposed system that is explained at Sec. IV-B. The feature words were extracted by using GETA [3], and this analysis used the highest score 30 feature words. That is, $K = 30$ in the algorithm of Fig. 3. Similarly, Fig. 4(c) is visualization of the search results

[3] `http://geta.ex.nii.ac.jp/e/index.html`

```
sub simple(query){
  W = word(doc(query),K)
  foreach wi in W {
    synsets = wnSynset(wi)
    map {lambda p {map {lambda q. Up(p,q)}} wnUp(p) } synsets
  }
  CutFork
  tops = Top(Edge)
  map {lambda p {map {lambda q. UpAll(p,q)}} wnUp(p) } tops
}

sub Up(p,q){
  if (c2w(q)*W != empty){
    addNode(q)
    addEdge(q<-- p) if (label(p) != label(q))&&( p<--p not in Edge)
    map {lambda r. Up(q,r)} wnUp(p)
  } else {
    map {lambda r. Up(p,r)} wnUp(q)
  }
}

sub UpAll(p,q){
  if (c2w(q) != empty){
    addNode(q)
    addEdge(q<-- p) if (label(p) ne label(q))&&( p<--p not in Edge)
    map {lambda r. Up(q,r)} wnUp(p)
  } else {
    map {lambda r. Up(p,r)} wnUp(q)
  }
}
```

Fig. 3. Algorithm

which used the query "hotel", and Fig. 4(b) is visualization by the comparative graph (Sec. IV-B. The node including many feature words from the query "Japanese inn" is displayed in red, and the node including many feature words from the query "hotel" is displayed in green.

As feature words from the query "Japanese inn" displayed in red, there are "area", "place", "sea", "close", "profile" and "inn". As feature words from the query "hotel" displayed in green, there are "room", "night" and "sightseeing". General terms, such as "blog", "news", "spa" and "person" are displayed in gray. From this graph, Japanese inns are assumed to be community-base, and hotels are assumed to be the base of tourism.

B. Comparison of Seven Prefectures in Kyushu

In this section, the tourism information on Kyushu 7 prefectures was compared mutually. Using all the combination of 2 prefectures, the system of Sec. IV-B generated comparative graphs. If all combination are visualized, the matrix of 7×7 as shown in Fig. 5 is generable. A general situation analysis is possible by seeing the whole, and a detailed individual analysis is possible by pay attention to a specific grid.

The following views are made as global analysis. (i) It has many features (there are many red nodes). (ii) It is mutually alike (there are many gray nodes). (iii) It has almost no feature (there are many green nodes). For example, since the line of Fukuoka Prefecture has many red nodes in Fig. 5, Fukuoka Prefecture supports (i), i.e., it can identify that Fukuoka Prefecture has many features compared with other prefectures.

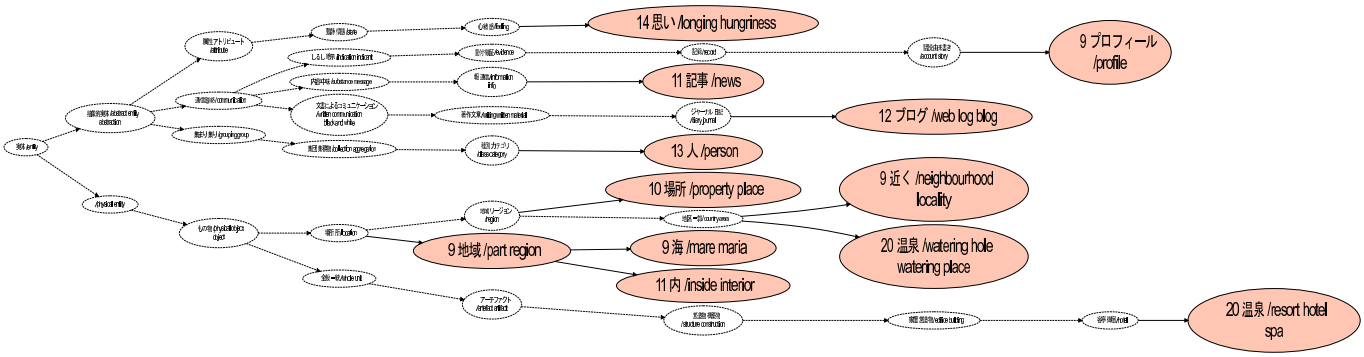
Moreover, each analysis is possible by seeing a specific grid. For example, in the graph of Kumamoto vs Fukuoka Prefecture, the castle is red. That is, Kumamoto Prefecture has the feature about a "castle" compared with Fukuoka Prefecture. Maybe, it seems Kumamoto Castle. On the other hand, since there is almost no red node in the line of Miyazaki Prefecture, Miyazaki Prefecture has almost no feature on tourism. However, in the whole graph of the line of Miyazaki Prefecture, the "taste" node is red in common. That is, the special feature of Miyazaki Prefecture of excelling another prefectures is "taste."

VI. Conclusion and Future Words

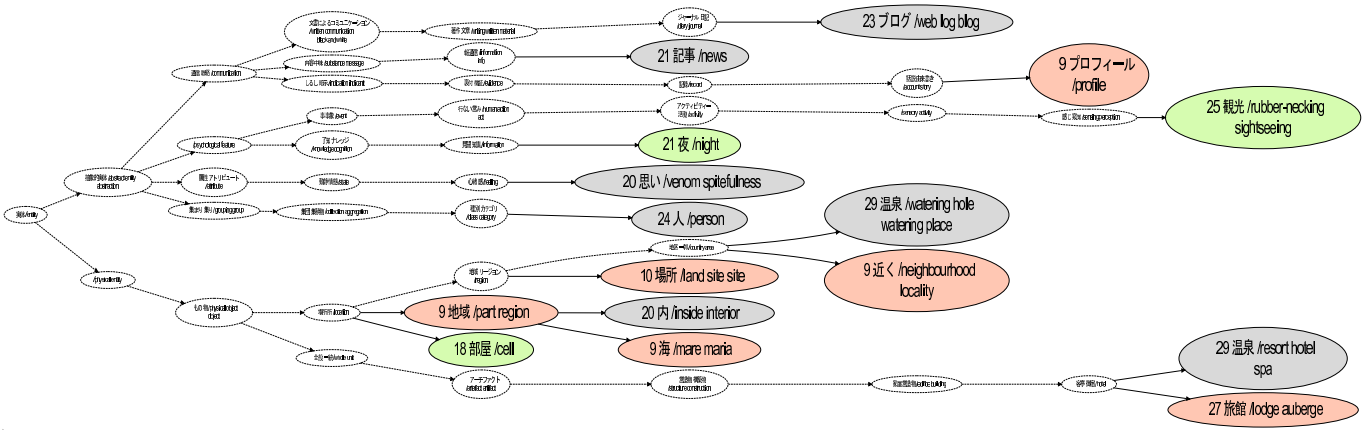
In this paper, we proposed the analysis system which visualizes the feature word of document groups using WordNet as a thesaurus. By actually visualizing search results for the blog articles about the tourism of Kyushu, we confirmed that global analysis and detailed analysis could be simultaneously conducted by a proposal system.

The present system has stopped at displaying the relation of the feature words. For more detailed analysis, we are planning construction of the interactive re-search engine based on a display result.

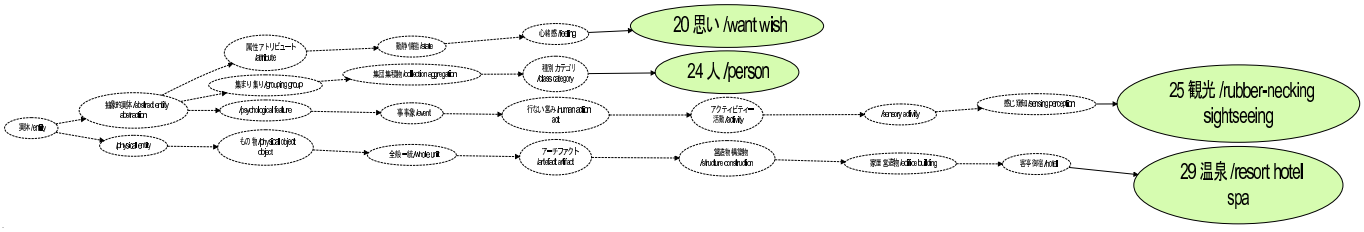
Moreover, although this system used WordNet as a conceptual thesaurus, we think that the thesaurus which specialized



(a) Ryokan Only



(b) Ryokan and Hotel



(c) Hotel Only

Fig. 4. Comparison of Tourism Information about Ryokan and Hotel

in tourism is also required. The development is also a future work.

REFERENCES

[1] X. Wu, S. Hirokawa, C. Yin, T. Nakatoh and Y. Tabata, "Extraction and Comparison of Tourism Information on the Web," in *Proc. of AROB2011*, 2011.

[2] T. Baba, L. Liu and S. Hirokawa, "Formal Concept Analysis of Medical Incident Reports," *Lecture Notes in Computer Science, Volume 6278/2010*, pp.207-214, 2010.

[3] T. Baba, T. Nakatoh and S. Hirokawa, "Text Mining of Bankruptcy Information using Formal Concept Analysis," *2nd World Congress on Computer Science and Information Engineering (CSIE 2011)*, Changchun, China. 2011.

[4] S. Esparcia, V. Sanchez-Anguix, E. Argente, A. Garcia-Fornes and V. Julian, "Integrating Information Extraction Agents into a Tourism Recommender System," in *Proc. HAIS2010, Springer LNAI 6077*, pp.193-200, 2010.

[5] J. M. Ruiz-Martinez, D. Castellanos-Nieves, R. Valencia-Garcia, J. T. Fernandez-Brieis, F. Garcia- Sanchez, P. J. Vivancos-Vincente, J. S. Castejon-Garrido, J. B. Camon and R. Martinez-Bejar, "Accessing Touristic Knowledge Bases through a Natural Language Interface," *Springer LNAI 5465*, pp.147-160, 2009.

[6] H. Saito and A. Ohuchi, "A Study of Visualizing Method of WWW Documents to Construct the Concept on Sightseeing Information," *IEICE Tech. Report, DE2001-07*, pp.261-267, 2001. (in Japanese)

[7] I. Kinjo and A. Ohuchi, "Web data analysis for Hokkaido tourism information," *IEICE Tech. Report, DE2001-07*, pp.99-104, 2001. (in Japanese)

[8] Q. Hao, R. Cai, Ch. Wang, R. Xiao, J.-M. Yang, Y. Pang and L. Zhang, "Equip Tourist with Knowledge Mined from Travelogues," in *Proc. WWW2010*, pp.401-410, 2010.

[9] H. Ozaku, M. Utiyama and M. Kidode, "An Event Information Retrieval Method Using Features of Keyword Appearance in Newspaper Corpora," *Trans. JSAI, AI19*, pp.225-233, 2004. (in Japanese)

[10] A. Ishino, H. Nanba, H. Gaguma, T. Ozaki, D. Kobayashi and T. Takezawa, "Automatic Compilation of Travel Information from Automatically Identified Travel Blogs," *IEICE Tech Report, WI2-2009*, pp.19-23, 2009.(in Japanese)

[11] H. Okumura, M. Tokuhisa, J. Murakami and M. Murata, "Trial of extracting and classifying strong points in sightseeing area," *IEICE technical report. Natural language understanding and models of communication 110(245)*, pp.25-30, 2010. (in Japanese)

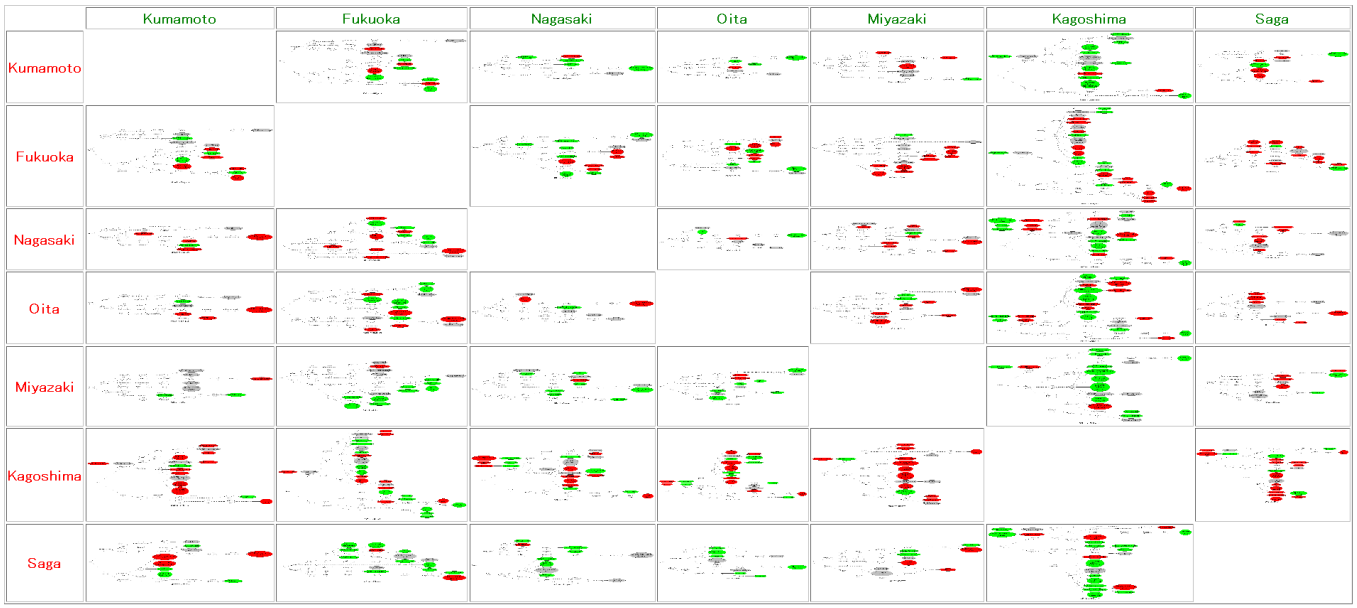


Fig. 5. Comparison of the Tourism Information on Seven Prefectures in Kyushu

[12] F. Bond, H. Isahara, K. Kanzaki and K. Uchimoto, "Boot-strapping a WordNet using Multiple Existing WordNets," *In LREC-2008*, Marrakech. 2008.

[13] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross and K. J. Miller, "Introduction to WordNet: An on-line lexical database," *International Journal of Lexicography*, vol. 3, no. 4, pp. 235-312, (1990).

[14] Manuel de Buenaga Rodríguez, José María Gómez Hidalgo and Belén Díaz-Agudo, "Using Wordnet to complement Training Information in Text Categorization," *In Journal of CoRR, Vol.cmp-lg/9709007*, 1997.

[15] S. Scott and S. Matwin, "Text classification using WordNet hypernyms," in *Proc. of the Conference: Use of WordNet in Natural Language Processing Systems*, pp.38-44, 1998.

[16] E. M. Voorhees, "Query expansion using lexical-semantic relations," in *Proc. of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pp.61-69, 1994.

[17] R. Navigli and P. Velardi, "An analysis of ontology-based query expansion strategies," in *Proc. of Workshop on Adaptive Text Extraction and Mining*, pp.42-49, 2003.

[18] T. Iwata, T. Yamada and N. Ueda, "Visualizing Documents Based on Topic Models," *Transactions of Information Processing Society of Japan*, vol. 50, no. 6, pp.1649-1659, 2009.

[19] S. Hirokawa, Y. Shimoji and T. Wada, "Construction of Concept Graph from Documents," *IPSJ SIG Notes 2005(94)*, pp.79-84, 2005. (in Japanese)

[20] C. Yin, X. Wu, S. Hirokawa and T. Nakatoh, "A Proposal of 'TOIEBA' Search Engine for Tourism Event," *IEICE technical report. Artificial intelligence and knowledge-based processing* vol. 110, no. 301, pp.43-47, 2010. (in Japanese)

[21] C. Yin, T. Nakatoh, S. Hirokawa, X. Wu and J. Zeng, "A proposal of search engine " XYZ " for tourism events," *Second IITA International Joint Conference on Artificial Intelligence (IITA-JCAI)*, 2010.

[22] William W. Cohen, "Web-based information system that reasons with structured collections of text," in *Proc. of the 1998 2nd International Conference on Autonomous Agents*, pp.400-407, 1998.