Extraction and Comparison of Tourism Information on the Web

Xiaobin Wu[1)], Sachio Hirokawa[2)], Chengjiu Yin[2)], Tetsuya Nakatoh[2)], Yoshiyuki Tabata[2)]

[1)]*Graduate School of Information Science and Electrical Engineering,*
[2)]*Research Institute for Information Technology, Kyushu University*
Hakozaki 6101,Fukuoka,
8128581,JAPAN
2ie10056y@s.kyushu-u.ac.jp, {hirokawa,yin,nakatoh,tabata}@cc.kyushu-u.ac.jp

Abstract: The number of tourists to Japan from foreign countries is drastically increased in recent years. However, there is a scene where the traveler is made uneasy by differences between the word, the custom and the culture in traveling abroad. We are aiming at the construction the horizontal search engine intended for tourism information in the Kyushu region as a test case of a special vertical search engine. As the first step for the research, we extracted 312 events from a public tourism portal site and compared the ranking of each event in the site with that in the general search engine. We confirmed a weak correlation of the ranking. Moreover, we found that the big difference of rankings for the events with strong regionality.The number of tourists to Japan from foreign countries is drastically increased in recent years. However, there is a scene where the traveler is made uneasy by differences between the word, the custom and the culture in traveling abroad. We are aiming at the construction the horizontal search engine intended for tourism information in the Kyushu region as a test case of a special vertical search engine. As the first step for these the research, we extracted 312 events from a public tourism portal site and compared the ranking of each event in the site with that in the general search engine. We analyzed the correlation of the rankings and confirmed a weak correlation. In addition, it was also confirmed that there was a large gap for the events with strong regionality.

Keywords: Search engine, information extraction, rank correlation coefficient ,data collection, area tourism

I. INTRODUCTION

According to the development of the Internet, many information came to exist on WWW. By using a general search engine, we can get Web pages which contain the keywords we send to the search engine as query. However, it is not easy to reach the information unless we know appropriate keywords. Even if we know some of suitable keywords, the target pages, we are searching for, might be buried under the major pages displayed with higher ranked position of the search results.

One of the reasons of this problem is in that the coverage of general search engine is too wide and too general for particular purpose. We are studying the search engine focused on specific fields and objects on WWW. By limiting the search targets, we can expect to realize a search engine excellent in precision and recall for the targets. The present paper focuses on the tourism information on WWW, and compares the feature of information found in a tourism portal site and in a general search engine.

The number of foreign tourists visiting Japan are increasing continuously in recent years. Thanks to the simplification of the visa to the well-to-do population of China, more and more Chinese tourists are expected to come to Japan. However, there are many scenes where a traveler feels uneasy by the difference in language, in a custom, or in culture. Of course, guide books have been used as helpful tools. Information on the Internet is much useful and reliable with respect to newest and up-to-date tourism information. This paper chooses touristic event information in Kyushu area in Japan, as the target of the analysis.

We collected the tourism information from the portal site of "Kyushu Tourism Promotion Organization." We analyzed the ranking of each event in the portal and the ranking of that in the general search engine Google. To evaluate the ranking, we used the estimated number of Web pages that match the name of the events which we send as a query.

Fig.1. The information of Welcomekyushu

II. Collection of Tourism Event Information

The event information on the Kyushu area was chosen as tourism information dealt with by this

research. We prepare the list of events as follows. Firstly, we obtained the 906 information from the "event list of Kyushu and Okinawa" in the web site of the "Kyushu Tourism Promotion Organization", which we refer as "welcomekyushu" in the sequel of the present paper. This list contains not only the names of event but also the names of food, such as "colander tofu", and the names of places, such as "alpine rose park". We removed these non-event names and compiled 316 event names. Some of the events are held every year with the name of the year, such as "X festival 2009" and "X festival 2010". The list of 312 events was obtained by removing these duplication. Some of the list is shown in Table 1.

Table 1. 20 Events of Event Name list of Kyushu

十五夜ソラヨイ，鏡山スカイスポーツフェスティバル，串木野浜競馬，「小城」ホタルの里ウォーク，とす弥生まつり，べっぷ鶴見岳一気登山大会，宮崎神宮大祭，美山窯元祭り，小倉祇園太鼓，四十九所神社「やぶさめ祭り」，白地楽，幸若舞，うすき竹宵，筑前いいづか雛のまつり，牛深ハイヤ祭り，武雄の荒踊り，門司みなと祭，鹿児島カップ火山めぐりヨットレース，南大隅町ねじめドラゴンボートフェスティバル，山鹿灯籠浪漫・百華百彩

Next, in order to compare the amount of information which exists in the general Web page on WWW, and the amount of information which exists in a tourism special site, the number of the search results with the event names was obtained. Google was chosen as a general search engine and the Kyushu Tourism Promotion Organization was chosen as a tourism special site. First, each event name was searched with Google and the number of the obtained search results was made into the amount of information on a general Web page. Next, the search range of Google was limited to www.welcomekyushu.jp using the "site:" operator. The number of the search results obtained from welcomekyushu was made into the amount of information of a tourism special site. The number of the information in a general Web page was from 29 to 241,000,000 by the event name. by the event name, The number of the information on a tourism special site was from 2 to 603 similarly. Fig. 2 plotted the number of search results in descending order of it.

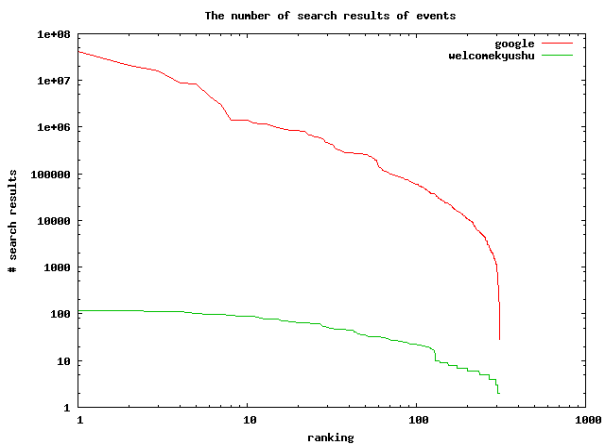

Fig.2.The Number of Search Results of Each Event

III. Analysis by Rank Correlation

1. Kendall tau rank correlation coefficient

We compared event information with Kendall tau rank correlation coefficient [1].A computational procedure consists of four following Steps.The paper title text is 15 point bold font in Times New Roman.

Step 1. The number of measured value is set to n about two variables *X* and *Y*.That is, there are measured values of X_k and Y_k (k= 1 .. n).

Step 2. Ranking is attached to an ascending order at Variable *X* and Variable *Y*. When there is equal order, average of the ranks is given.They are arranged in an ascending order about Variable *X*.

Step 3. The number of the equal order in Variable *X* and Variable *Y* is set to n_x and n_y.The size of equal order is set to t_i and t_j *(i=1,2,...,*n_x*; j=1,2,...,*n_y*)*.Then, T_x and T_y can be found as follows.

$$Tx = \sum_{i=1}^{n_x} \frac{t_i(t_i-1)}{2} \qquad T\_y = \sum_{i=1}^{n_y} \frac{t_j(t_j-1)}{2}$$

Step 4. Number in the case of being *Yi* < *Yj* about Y_i *(i=1, ..., n-1)* and Y_j *(j=i+1, i+2, ..., n)* is set to P_i. Number in the case of being $Y_i > Y_j$ about Y_i and Y_j is set to Q_i. $\sum P_i$ is the number of times whose direction of the ranking of two variables corresponds. $\sum Q_i$ is the number of times whose direction of the ranking of two variables corresponds with an opposite direction. At this time, a rank correlation coefficient is calculated for by the following formula.

$$\tau_k = \frac{\sum_{i=1}^{n} P_i - \sum_{i=1}^{n} Q_i}{\sqrt{\frac{n(n-1)}{2} - T\_x}\sqrt{\frac{n(n-1)}{2} - T\_y}}$$

2. Examination

To compare the ranking of each event from 312 events, in the portal site "welcomekyushu" and with that in general search engine, we used the estimated number of Web pages that contain the name of the event. The obtained rank correlation coefficient was 0.20744. The correlation coefficient is not so strong to confirm a clear correlation. However, the p-value we obtained is 4.616*10-8, which is less than the required significance level 5% (α= 0.05). Thus, we can reject the null hypothesis at the given level of significance. Eventually, we see that the correlation is true.

IV. Events with Ranking Gaps

In Fig.3, each point represents to an event, where the x-axis is the ranking of the event in Google, and the y-axis is the ranking of the event in welcomekyushu.

There is a large gap between the ranking by Google and by welcomekyushu,when an event is displayed in distance from the line A, which represents the strong correlation.By analyzing those events, we observed the following four facts.

1)The points in Domain X support the correlation of the two rankings.The more points are in X, the higher is the correlation coefficient.We can find out weak correlation from Fig. 3.

2) The points in Domain Y have a high ranking in Google, and a low ranking in welcomekyushu.The names of these events seems to appear anywhere all over Japan. "Ohito Kabuki" and "Amazake Festival"(Fig.2) are samples of such names.

3)The points in Domain Z have a low ranking in Google, and a high ranking in Welcomekyushu. Those events are peculiar to Kyushu area.Many event titles contained a name of a place."Hosenji firefly festival" and the "Sakurajima enjoying -the-evening-cool tour ship" (Fig.3) are samples of these names.

4)The area Y contains more events than that of Z.The region Z displays the regional strength of the site.

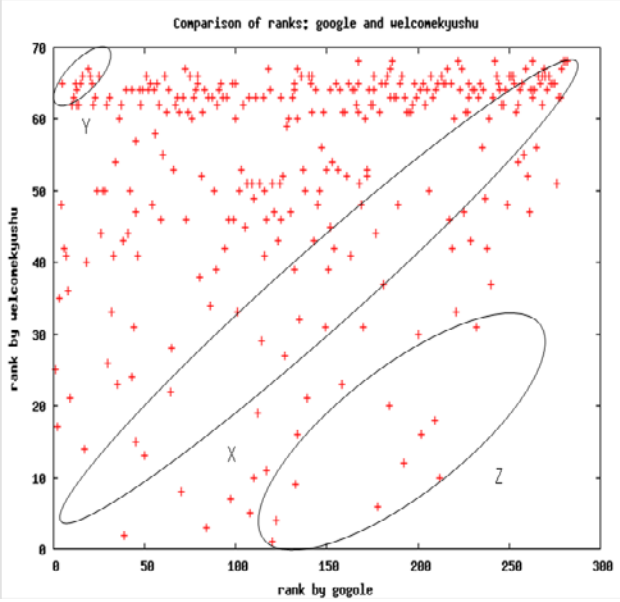

Fig.3 Ranking of Events

We compared the ranking of the number of the search results for each event name, in order to confirm the difference between the amount of information in a general Web page, and the amount of information of a tourism special site. Table 2 shows the list of events with high ranking with Google search engine. Table 3 shows the list of events with high ranking in welcomekyushu. From these two tables, we can say that regional events are much easier to find by welcomekyushu than by Google.

Table 2. 10 Events with High Ranking in Google, Low Ranking in WelcomeKyushu

Ranking	*Rk−Rg**	Rg*	Rg	Rk	G	K	Event
1	63.79	1.21	5	65	8710000	5	竹ん芸
2	62.42	4.58	19	67	858000	3	おしろい祭
3	62.14	3.86	16	66	935000	4	みそ五郎まつり
4	62.11	2.89	12	65	1240000	5	甘酒まつり
5	61.38	3.62	15	65	965000	5	鮒市
6	61.18	4.82	20	66	836000	4	八天神社例大祭
7	60.87	3.13	13	64	1180000	6	大人歌舞伎
8	60.35	2.65	11	63	1400000	7	出の山ホタル恋まつり
9	59.97	6.03	25	66	625000	4	サン・サン・さんわフェスティバル
10	59.94	5.06	21	65	819000	5	くも合戦

Table 3. 10 Events with High Ranking in WelcomeKyushu, Low Ranking in Google

Ranking	*Rg*−Rk*	Rg*	Rg	Rk	G	K	Event
303	23.07	32.07	133	9	25700	93	皿山まつり
304	24.37	44.37	184	20	10200	65	早岐茶市
305	24.94	55.94	232	31	3970	47	長串山つつじまつり
306	25.42	29.42	122	4	32700	112	薩摩のひなまつり
307	27.94	28.94	120	1	36300	603	山鹿温泉祭
308	32.40	50.40	209	18	6100	67	仙酔峡つつじ祭り
309	32.71	48.71	202	16	6790	72	筑後吉井おひなさまめぐり
310	34.30	46.30	192	12	8980	83	宝泉寺ホタル祭り
311	36.92	42.92	178	6	10900	101	桜島納涼観光船
312	41.12	51.12	212	10	5950	90	人吉球磨は、ひなまつり

Each item shown in Table 2 and Table 3 is described as below

Rg*: normalization ranking of Google search results

Rg: ranking of Google search results

Rk: ranking of Welcomekyushu search results

G : number of Google search Results

K : number of Welcomekyushu search results

V. Related Work

The tourism industry is one of the industries that suffered big influence of internet. Sightseeing information was unavailable only in a special travel agent before. Now, everyone can easily obtain it thanks to the internet. We can find sightseeing information on Web in (a) tourism portal sites, (b) general web pages, and (c) blog sites.

There are several systems and researches intended for each targets. [2] proposes a recommendation and a clustering system and shows their effectiveness for tourism portals.[6]proposes a natural language interface for tourism search engine. [5] analyses 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. [3] studies the clue words that can be used to extract tourism related NES. [4] reports the characteristic keywords that distinguish tourism blogs from other general blogs.

We are aiming at the construction of the vertical search engine for tourism information. The difference of the characteristic of (a) and (b) was examined in the present paper.

VI. CONCLUSION

This paper analyzed tourism information available on a public tourism portal site that covers 7 prefectures in Kyushu area. 312 events were extracted from the site and were used to compare the portal site with a general search engine. We analyzed the correlation of the rankings of each event in the portal site and in the general search engine. It turned out that there exists a weak correlation between the rankings. It is also confirmed that there is a large gap for the events with strong regionality, which indicates a characteristic of the portal site. One of the lessons we learned is the usefulness of the lists of names, such as the names of events, locations, shrines and souvenirs. We used 312 events for comparison of portal and general search engine. We extracted them from the lists in the portal site. Therefore, the development of the technique of the list discovery is an important problem. It is also necessary to devise a method to identify the same events with different names. As the next step of the research, we are considering to analyze other regional portal sites to compare with each other.

REFERENCES

[1] H. Abdi, Kendall rank correlation, In N.J. Salkind (Ed.): Encyclopedia of Measurement and Statistics. Thousand Oaks (CA): Sage,2007.

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

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

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

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

[6] 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, R. Martinez-Bejar, Accessing Touristic Knowledge Bases through a Natural Language Interface, Proc. PKAW2008, Springer LNAI 5465, pp.147-160, 2009.