社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2003－ＩＣＳ－133（19）
2003／9／15

An Automatic Collection System for Official Accommodation Websites

Kohichiro TAKAGI†, Masahito YAMAMOTO†, Masashi NAKATSUGAWA††, Hidenori KAWAMURA†, and Azuma OHUCHI†

† Graduate School of Engineering, Hokkaido University North 13, West 8, Sapporo, 060-8628, Japan
†† Japan Science and Technology Corporation Honmachi 4-1-8, Kawaguchi 332-0012, Japan
E-mail: †{takagi,masahito,masashi,kawamura,ohuchi}@complex.eng.hokudai.ac.jp

Abstract It is becoming increasingly difficult to find desired information from the World Wide Web (WWW) due to its rapid growth, and especially, to find web pages belonging to a certain category. For example, it is difficult for tourists to search for official accommodation websites such as hotels and pensions in the area that they will visit. Existing search engines are not suitable for this purpose because these search engines utilize some keywords to extract the information, and accommodation websites do not always contain the most common words. In this paper, we develop an automatic collection system that can extract some accommodation websites in a certain region and detect whether a website is an official one. Our main idea is the utilization of telephone numbers and link-structure analysis. We have applied our proposed system to Hokkaido local accommodation websites in some areas, although the proposed system can also be applied to other objects by changing the extraction rule. This is the first step of our trial to make a dynamic Internet Directory.

Key words WWW, official accommodation websites, dynamic Internet Directory

宿泊施設の公式ホームページ収集システム

髙木耕一郎† 山本 雅人† 中津川雅史†† 川村 秀憲† 大内 東†

† 北海道大学大学院工学研究科 〒060-8628 北海道札幌市北区北13条西8丁目
†† 科学技術振興事業団 〒332-0012 埼玉県川口市本町4丁目1番8号
E-mail: †{takagi,masahito,masashi,kawamura,ohuchi}@complex.eng.hokudai.ac.jp

あらまし WWWの急速な発達により、WWW上での情報検索はますます困難となってきている。特にあるカテゴリーに属するウェブページを発見することは難しい。例えば旅行者が訪れようとしている地域の宿泊施設の公式ホームページを見つけることは困難である。そのような目的は既存の検索エンジンに適さない。なぜならば既存の検索エンジンはキーワードを用いており、宿泊施設の公式ホームページは共通のキーワードを含まないからである。本稿で我々は宿泊施設の公式ホームページを収集するシステムを提案する。それは任意の地域を対象とし、さらに公式ホームページであるか否かも判断することができる。我々は北海道のいくつかの地域に本システムを適用した。また、ルールを変更し他の対象に対するシステムを作成することもできる。これは動的インターネットディレクトリ作成の第一歩である。

キーワード WWW, 宿泊施設の公式ホームページ, 動的インターネットディレクトリ

1. Introduction

The World Wide Web (WWW) creates new challenges for information retrieval because of its amount of information. The search engines are the WWW information retrieval systems that most people are familiar with. However, they give poor results, because of a great deal of information and many variations of the WWW.

Mainly, the search engine consists of two kinds. Here, we call them the term based search engine and the internet directory.

The term based search engine is made by the robot which generally is called crower or spider. There are "Google" [1], "altavista" [2], and so on as its example. In advance, the

crower collects many web pages by using link structures and keeps it as the data base. For example, Google's crower collects three billion web pages. As the user sends a query string, web pages which contain the query string are shown to the user. At that time, the result is lined up by the ranking algorithm.

However, when we do not know what kind of word is contained in advance, we can not find the page by using term based search engines. For example, we consider official accommodation websites. Some websites contains the word "hotel", although there are many websites which do not contain the word "hotel". It is difficult to expect what kind of word is contained in it. Therefore, it is difficult to find official accommodation websites by using the term based search engines.

Moreover, when we send some common query strings like "hotel" to the term based search engine, hundreds of thousands of results are often shown. Though the result is lined up by the ranking algorithm, and many related works about the ranking algorithm have been done. There is a room of the improvement in it.

The internet directories are made by the human works. There are "Yahoo! Japan" [3], "ODP" [4], and so on as its examples. Web pages which are registered in the internet directory are classified by the human works. Therefore, its quality is far higher than the quality of result which is made by the term based search engine. However, while the number of web page increases, the amount of the human works is limited. The page registered in the internet directory is just a part of the whole.

It is required to make a dynamic internet directory, which is the internet directory made by the robot, automatically. It overcomes the problems of both the term based search engine and the internet directory. The problem of the term based search engine is the extraction of the information from the data base. The problem of the internet directory is the number of the registered pages.

In order to make a dynamic internet directory, we have to build some of the rules in each category. In this paper, we propose the rule about official accommodation websites. Our main idea is the utilization of telephone numbers and the link-structure analysis. The idea is based upon the fact that almost all accommodation websites have its telephone number, and the fact that the official one is often linked from many other websites.

Our idea can also be applied to other objects, especially the official websites about some companies or enterprises. The websites which have description about some companies or enterprises often have its telephone number. The official websites are often linked from many other websites.

2. The Related Work

In this section, we describe various link-based webpage analyses. Two popular link-based webpage ranking algorithms are HITS and PageRank. These algorithms use the link topology in order to capture the notion of some average opinion of the webpage creator. The hyperlinks of these webpages form a directed graph $G = (V, E)$, where V is the set of nodes representing webpages, and E is the set of hyperlinks. The hyperlink topology of the web graph is contained in the asymmetric adjacency matrix $L = (L_{ij})$, where $L_{ij} = 1$ if $p_i \rightarrow p_j$ and $L_{ij} = 0$ otherwise.

Kleinberg [5] presented HITS algorithm which can identify "hub" and "authority" web pages. The hub pages link to many authority pages. The authority pages are linked to by many hub pages. The definition of the two page type is recursive and mutually reinforcing. In its algorithm, each webpage p_i has both a hub score y_i and an authority score x_i. The L^{OP} represents the idea that a good authority is pointed to by many good hubs. The O^{OP} represents the idea that a good hub points to many good authorities.

$$X = L^{OP}(Y) = L^T Y \tag{1}$$

$$Y = O^{OP}(X) = LX \tag{2}$$

Vectors $X = (x_1, x_2, \cdots, x_n)^T$ and $Y = (y_1, y_2, \cdots, y_n)^T$ are the authority score and the hub score of each webpage. The final authority and hub scores of every webpage can be obtained through the iterative processes which can represent next expression.

$$cX^{(t+1)} = L^T L X^{(t)} \tag{3}$$

$$cY^{(t+1)} = LL^T Y^{(t)} \tag{4}$$

c is a normalization constant such that $||x|| = ||y|| = 1$ and $x^{(t)}$, $y^{(t)}$ represent the authority and the hub scores at the t^{th} iteration.

There are many improved algorithms which compute authority and hub scores. The ARC algorithm of Chakrabarti [6] extends Kleinberg's algorithm with textual analysis. ARC computes a distance-2 neighborhood graph and weights edges. The weight of each edge is based on the match between the query terms and the text surrounding the hyperlink in the source document. This algorithm is aim at making resource lists similar to those provided by Internet Directory Yahoo! or Infoseek. Their aim is similar to ours, but ARC depends on a textual analysis. Therefore, ARC cannot find web pages belonging to a certain category such as official accommodation websites. The official accommodation websites do not always contain the most common words.

Bharat [7] identified three problems with Kleinberg's algorithm and devised algorithms to tackle them. The essence of

their approach is to augment a previous connectivity analysis based algorithm with content analysis. Chang [8] presented a study about customized authority lists. They presented a technique that incorporates user feedback by adjusting the measure of authority to match an individual's internal notion of what sources are important. Gibson [9] presented a study about robustness of HITS algorithm.

Brin and Page [10] [11] presented PageRank algorithm which is used in the search engine Google. PageRank uses an idea similar to HITS that a good webpage should connect to or be pointed to by other good web pages. However, instead of mutual reinforcement, it adopts a web surfing model based on a Markov process in determining the scores. Each webpage p_i has PageRank scores z_i, and we define $Z = (z_1, z_2, \cdots, z_n)^T$. The final PageRank scores of every webpage can be obtained through the iterative processes which can represent next expression.

$$Z^{(t+1)} = P^T Z^{(t)} \quad (5)$$

$$P^T = \alpha L^T D_{out}^{-1} + (1-\alpha)(1/n)ee^T \quad (6)$$

Where, the out-degree of a webpage p_i is defined as $o_i = \Sigma_k L_{ik}$.

$$d_{out} = (o_1, o_2, \cdots, o_n)^T \quad (7)$$

$$D_{out} = diag(dout) \quad (8)$$

PageRank models two types of random jumps on the Internet. First, link-tracking jump: a surfer often follows the hyperlinks of web pages by simply clicking on them. This is modeled by $L^T D_{out}^{-1}$. Next, link-interrupt jump: a surfer sometimes jumps to another webpage not hyperlinked by the current webpage. Page Rank models such link-interrupt jump with a simple uniform distribution $(1-a)/n$, where $a = 0.8$ to 0.9 and $e = (1, 1, \cdots, 1)^T$.

Richardson [12] presented the text based expansion of PageRank. Haveliwala [13] presented the topic sensitive PageRank algorithm. A. Y. Ng's study [14] is a stability of the HITS algorithm and the PageRank algorithm. C. Ding [15] [16] presented the analysis about HITS and PageRank and their unified algorithms. Other link-based webpage analysis is a definition of a web community. Flake [17] define a web community to be a set of web pages that link in either direction to more web pages in the community than to pages outside of the community. Members of such a community can be efficiently identified in a maximum flow / minimum cut framework.

3. How to collect the official accommodation websites

The collection of the official accommodation websites consists of two steps. First, many candidates of the official accommodation websites are collected. This step is performed by using the term based search engine and the link structure of web pages. The system can gather most official accommodation websites as a candidate.

However, the system collects many web pages except official accommodation websites, too. For example, it is websites of a travel agency or websites of a self-governing body. Such websites contain information of many hotels, pensions and so on. As a time, the page of these websites looks like the official accommodation website. It is necessary to distinguish the official websites from non-official one.

Next, the system extracts the official accommodation websites from candidates. This step is performed by the analysis of the link structures and the text data of the candidate web pages. The system can extracts most official accommodation websites.

3.1 Collection of the candidate web pages

We mention how to collect the candidate web pages by using the term based search engine and the link structures. We adopt the most popular search engine "Google" as the term based search engine.

First, the system collects web pages by using the search engine. We call this result R. We use two kinds of words as the query string of the term based search engine. The one is the word which shows the geographical position. We define the set of this word as G. The other is the word which is often contained in the official accommodation websites. We define the set of this word as A.

We use the area code of the telephone numbers as the geographical term. This is based on the fact that the area code of the telephone numbers is the geographical term which is contained most often in the official accommodation websites. We checked 358 websites which are Hokkaido local accommodation websites registered in the Internet Directory "So-net" [18]. The telephone number is written in all of the websites. Sorry to say, the telephone number may be written by the image. The system ignores such a website. The probability which such a website exists in is 4.19%. Our idea to use telephone numbers can be applied to other objects as well. The telephone number is usually written in the websites of many companies or enterprises, such as a restaurant or a golf course.

We adopt the following twelve words as the elements of A. A ={"hoteru", "pensyon", "yu-suhosuteru", "rojji", "roddi", "minsyuku", "ryokan", "syukuhaku", "yoyaku", "ryoukin", "ryouri", "onsen"}. We checked 358 official accommodation websites, and the either twelve words are contained in all websites. When we apply our system to other objects, we have only to change these words.

We use "AND reference" of two words. It is necessary

for websites to contain the webpage which include both the element of G and the element of A. The probability that websites do not contain such web pages is 2.23%.

We adopt t high ranks of each result. This is because R doesn't grow big too much. We exclude duplication, and its result is R.

Next, the system collects the web pages which R is point to on the link structure. We call its result R^+. The system also collects the web pages which is point to R on the link structure. We call its result R^-. For getting R^-, we use the special function of Google called "Link Page".

This method is the part of the Kleinberg's HITS. There are two reasons in doing this operation. First, the candidate pages which are collected by using the term based search engine may not contain all official accommodation websites. Therefore, it is necessary to expand R. In fact, R contains about 70% websites of the official accommodation websites which the system finds. About 30% websites is collected in this step. Because the link-connected web pages often have the same topics, we expand R by using the link structure. Next, as we use the link structure for extracting the official websites in the next step, we need the set of pages which is connected thickly.

Union of R, R^+ and R^- is defined as S. S contains most official accommodation websites. We can define the following procedure :

Candidate(G,A,t,ϵ)

G : area code of a telephone number
A : words which are often contained official accommodation websites
t : a natural number
ϵ : a term based search engine
Let R denote the top t results of ϵ on G and A.
Set $S := R$

For each page $p \in R$
Let $R^+(p)$ denote the set of all pages p points to.
Let $R^-(p)$ denote the set of all pages pointing to p.
Add all pages in $R^+(p)$ to S.
Add all pages in $R^-(p)$ to S.
End
Return S

3.2 The candidate web pages

We mention what kind of page there is in S. S includes the web pages which do not include the information about the accommodation. We have to exclude these web pages. This can be done by the simple text analysis. We have only to check whether the web pages include a certain words or not.

The web pages which include the information about the accommodation can be divided three types.
Type 1 : Page of the official accommodation websites : This is made by the hotels or pensions itself.
Type 2 : List page such as a telephone number : This page often includes the information of many accommodations.
Type 3 : Pages except for type 1 and type 2.
We want to extract type 1 pages. It is easy to distinguish type 2 pages, for type 2 pages usually contain many telephone numbers. However, it is difficult to distinguish type 3 from type 1.

The popular example of type 3 is websites of a travel agency or websites of a self-governing body. These websites have many pages which contain the information about various accommodations. These pages include the necessary information such as hotel charges, address, telephone number, check-in time, check-out time, and so on. Those contents are almost the same as the official websites. Therefore, we cannot distinguish type 3 from type 1 by using the text data. However, we find the fact that there are many links from type 2 and type 3 to type 1. We can extract type 1 by counting the link from other pages.

3.3 Extraction of the official websites

We mention how to extract the page of the official accommodation websites from S. Here, we classify the web pages of S in list pages and non-list pages. List page is the page that has telephone numbers more than l. The web pages except for list pages are non-list pages, and we define these pages as S'. We adopt $l = 5$. Most official accommodation websites have one or two telephone numbers. One out of two is FAX number. And, there are some websites which have three or four telephone numbers just only. Therefore, the web pages which have more than five telephone numbers are not surely the official accommodation websites.

The system extracts telephone numbers from S'. We define these telephone numbers as t_n. The set of pages in which t_n is written is $S_n \subseteqq S$ and $S'_n \subseteqq S'$. Two manipulations are necessary about each telephone number t_n. We have to extract the official websites from S_n, and we have to detect whether t_n is the accommodation telephone number or not.

Here, we define the directed graph $D = (S, E)$. The nodes correspond to the pages. The directed edge $(p, q) \in E$ indicates the presence of a link from $p \in S$ to $q \in S$. From the graph D, we can isolate sub graphs. If $V \subseteqq S$ is a subset of the pages, we use $D[V]$ to denote the graph induced on V. Its nodes are the pages in V. Its edges correspond to all the links between pages in V.

In order to extract the official websites from S_n, the system manipulates the followings in every telephone number t_n. All S'_n are ranked by in-degree of graph $D[S_n]$. The page which in-degree is the biggest is the official website in that graph. If two or more pages have the same in-degree, the system uses the number of characters of URL. The page which the number of characters of URL is smallest is the official website.

In order to detect whether t_n is the accommodation telephone number or not, the system analysis the TITLE Tag of S'_n. If more than one page of S'_n has TITLE Tag which contains the following words A', t_n is the accommodation telephone number. A' are "hoteru", "pensyon", "yado", "ryokan", "rojji", "roddi", "hosuteru", "sansou", "kote-ji", and these Japanese "hiragana", "katakana", "kanji", and these English capital letter and small letter and its "zenkaku-moji", and "hankaku-moji". We can define the following procedure :

Extract(S,A',l)

A' : words which are often contained the TITLE Tag of the official accommodation websites

l : a natural number

Let $S' \subset S$ denote the set of page which contain less than l telephone numbers.

Let T denote the telephone numbers written in S'.

Let n denote the number of element of T.

Let $t_i \in T$ denote the element of T.

Let $S_i \subset S$ denote the set of page which contains t_i.

Let $S'_i \subset S'$ denote the set of page which contains t_i.

Set T' :=the empty set

For $i = 1, 2, \cdots, n$

If more than one page of S'_i contain TITLE Tag which contain A',

then Add t_i to T'.

End

Let m denote the number of elements of T'.

Set O :=the empty set

For $i = 1, 2, \cdots, m$

Let $p \in S'_i$ denote each page of S'_i.

Let $In-degree(p)$ denote the number of pages in S_i which points to p.

Let $URL-length(p)$ denote the number of characters of page p URL.

Let $P \subset S'_i$ denote the page which $In-degree(p)$ is biggest.

Let $P' \subset P$ denote the page which $URL-length(p)$ is smallest.

Add P' to O.

End

Return O

4. Experiment and Evaluation

We applied our proposed system to Kuttyan town, Hokkaido local accommodation websites. The telephone numbers of Kuttyan town is 0136 ($2a$) $bcde$. a is 1 or 2 or 3. b, c, d, e are the optional numbers.

(1) The system made the set R. G is "0136" and its Japanese "zenkaku moji". We adopt $t = 200$. The system collected 1302 web pages as the set R.

(2) The system made the set S. The system collected 19454 web pages as the set S.

(3) From S', the system extracted all telephone numbers. The system found 189 telephone numbers. Some accommodations may have two or more telephone numbers. Such a case is 62. Our system doesn't exclude such a case, so two or more results about one accommodation may be made. Here, we ignore such a case, and evaluate 127 telephone numbers. We define these telephone numbers $t_1, t_2, \cdots, t_{127}$.

(4) From $S_1, S_2, \cdots, S_{127}$, the number of official websites is 71. The system extracted 69 official websites, and the probability that it succeeded was 95.8%.

(5) From $t_1, t_2, \cdots, t_{127}$, the number of accommodation telephone numbers is 71, and the number of non accommodation telephone numbers is 56. The system classified 127 telephone numbers into accommodation telephone numbers and non-accommodation telephone numbers. The system found that 59 telephone numbers are accommodation telephone numbers, and 56 is right and 3 is wrong. The probability it succeeded is 94.9%. The system found that 68 telephone numbers are non-accommodation telephone numbers, and 53 is right and 15 is wrong. The probability it succeeded is 77.9%.

Table 1 is an example of the results about a certain accommodation. These web pages have common telephone numbers, so have description about common accommodation. Page 1 is the page of the official website. The system succeeded in discovering that it is the official one. Page 2, 3, 4, 5, 8, 9, 10, 11, 12, 13, 15, 16 have link to page 1. There is no link between these pages except for that.

table1 An example of the result

Page	URL
1	http://plaza16.mbn.or.jp/ sirotaya/
2	http://www.staff-f.co.jp/htl002.htm
3	http://www.petpet.ne.jp/hotel/yado_info.asp?id=357
4	http://www.asahi-net.or.jp/ LT6H-KBYS /tomaru/yado/to016.html
5	http://www.step-life.com/pet.html
6	http://www.tourism.ne.jp/1_1.html
7	http://www.petyado.com/hokkaido1_2.html
8	http://www.1101.com/noren/shop/hokkaido-0.html
9	http://www.jpinfo.ne.jp/yado/pref/hokkaido.html
10	http://www.kitakita.ne.jp/syukuhakusisetu /syu-ou.html
11	http://www.yomi21.co.jp/pension/hokaido /ph-kuccyan.htm
12	http://www.cybercrea.net/special/pet_0203 /yado_01_hokk.htm
13	http://www.wombat.zaq.ne.jp/sena/yado /hokaidou/hokaidou.htm
14	http://www.nisekoexpress.net/activity /taiken/taiken_winter.html
15	http://members.jcom.home.ne.jp/dogstravel /Data/hokaido/hoka2.html
16	http://www.neko-tsushin.com/catnavi /accommodation/acm-hokkaido01.html
17	http://www.dogoo.com/database/hotel /hotelsearch.cgi?category=01hokkaidou

5. Conclusion

We showed that it was possible to find web pages belonging to a certain category which was difficult to find so far. We proposed the extraction rule for official accommodation websites. It is possible to collect other objects by an easy change of the extraction rule. The telephone number is mentioned in the websites of almost all the companies or enterprises, and there are many link collection pages about each category. Therefore, our main idea (the utilization of telephone number and the analysis of link-structure) can be applied to official websites of many companies or enterprises. In this case, the rules A and A' are changed.

As a future work, we will find the equivalent to A and A', automatically. If this goes well, we can build many rules for other objects, and we can make the dynamic internet directory about many official websites of companies or enterprises.

Reference

[1] http://www.google.com/

[2] http://www.altavista.com/

[3] http://www.yahoo.co.jp/

[4] http://dmoz.org/

[5] J. Kleinberg, "Authoritative Sources in a Hyperlinked Environment", In Proceedings of the ACM-SIAM Symposium on Discrete Algorithms, pp.668-677, New York, May 1998.

[6] S. Chakrabarti, B. Dom, P. Raghavan, S. Rajagopalan, D. Gibson, and J. Kleinberg, "Automatic Resource Compilation by Analyzing Hyperlink Structure and Associated Text", In Proceedings of 7th International World Wide Web Conference, May 1998.

[7] K. Bharat and M. Henzinger, "Improved algorithms for topic distillation in a hyperlinked environment", In Proceedings of ACM-SIGIR Conference, 1998.

[8] H. Chang, D. Cohn, and A. McCallum, "Creating Customized Authority Lists", In Proceedings of 17th International Conference of Machine Learning, 2000.

[9] D. Gibson, J. Kleinberg, and P. Raghavan, "Inferring Web Community from link Topology", In Proceedings of the 9th ACM Conference on Hypertext and Hypermedia (HYPER-98), pp.225-234, New York, June 1998.

[10] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine", In Proceedings of 7th World Wide Web Conference, 1998.

[11] L. Page, S. Brin, R. Motowani, and T. Winograd, "PageRank citation ranking: bring order to the web", Stanford Digital Library working paper 1997-0072, 1997.

[12] M. Richardson and P. Domingos, "The Intelligent Surfer: Probabilistic Combination of Link and Content Information in PageRank", MIT Press volume 14, 2002.

[13] T. Haveliwala, "Topic-Sensitive PageRank", In Proceedings of the 11th International World Wide Web Conference, May 2002.

[14] A. Y. Ng, A. X. Zheng, and M. I. Jordan, "Stable algorithms for link analysis", In Proceedings of the 24th International Conference on Research and Development in Information Retrieval (SIGIR2001), 2001.

[15] C. Ding, X. He, P. Husbands, H. Zha, and H. Simon, "PageRank, HITS and a Unified Framework for Link Analysis", LBNL Tech Report 49372, 2001.

[16] C. Ding, H. Zha, X. He, P. Husbands, and H. Simon, "Link Analysis: Hubs and Authorities on the World Wide Web", LBNL Tech Report 47847, 2001.

[17] G. Flake, S. Lawrence, and C. L. Giles, "Efficient Identification of Web Communities", In Proceedings of 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2000.

[18] http://so-net.excite.co.jp/