Extended Link Analysis for Extracting Spatial Information Hubs

Jianwei Zhang* Yoshiharu Ishikawa*† Hiroyuki Kitagawa*†
*Department of Computer Science, Graduate School of Systems and Information Engineering
†Center for Computational Sciences
University of Tsukuba
Tsukuba, Ibaraki, Japan
zjw@kde.cs.tsukuba.ac.jp, {ishikawa,kitagawa}@cs.tsukuba.ac.jp

Abstract

Recently, web mining that tries to find useful knowledge from the vast amount of web pages has attracted a lot of research interests. Besides, it is becoming an essential task to provide web pages related to a user-specified geographic area. In this paper, we propose an approach to extract spatial information hubs from the web. A spatial information hub *is a web page which is related to a specified geographic area and has much local information and/or many hyperlinks to local web pages. In the traditional approach of web link analysis, the importance and quality of pages are judged only by their contents and hyperlink structures. However, we take their geographic localities into consideration. In our approach, we first extract geographic information from web pages to create* spatial nodes *and* spatial links*, then conduct a link analysis based on the extended link structures. We also show our approach works well based on the experiments.*

1. Introduction

Since the rise of the World Wide Web, finding useful information efficiently from the vast amount of web pages is becoming more and more important. Recently the research of *web mining* [5, 7] has attracted a lot of research interests. In particular, *link analysis* which uses link information between web pages is becoming an important technology for finding web pages with high popularities.

Besides, with the progress of mobile devices and GPS technologies, it is also becoming an essential task to provide information related to a specified geographic area. For example, several researches consider geographic localities of web pages to be important and try to provide information on a specified geographic area to users [1, 8].

In this paper, we propose a method to extract *spatial information hubs* from the web. We informally define a *spatial information hub* as:

> A web page which is related to a specified geographic area and contains much local information and/or many hyperlinks to local web pages.

Generally, two types of web pages can be considered as spatial information hubs. First, a type of pages that refer to good pages related to the specified geographic area. For example, a navigation page for National Research and Educational Institutions in Tsukuba[1] is a good spatial information hub for Tsukuba city. It links the homepages of some universities and institutes in Tsukuba. More detailed information can be acquired by following the links in this page. Second, a type of pages that refer to good locations. For example, a list page of hotels in Tsukuba[2] is also a good spatial information hub for Tsukuba city. It is a good information source for the Tsukuba area since the page contains useful local information on Tsukuba. Note that this page has no hyperlinks so that it can not be ranked higher if we use a traditional web link analysis. However, in our approach, it can be identified as a good spatial information hub by the extended link analysis method described below. If such two types of pages are extracted from the web, they can be used to construct a local portal site related to the specified geographic area.

One of the features of our approach is that we extend the web graph structures using *spatial nodes* and *spatial links*. Spatial nodes represent locations with local interests in the geographic space. Spatial links are used for two purposes: 1) to connect spatial nodes and web pages containing the corresponding geographic information, and 2) to connect two spatial nodes whose corresponding locations are close. In the traditional approach of web link analysis, the importance and quality of pages are judged only by their contents and hyperlink structures. However, we employ geographic information to create spatial nodes and spatial links and conduct a link analysis based on the extended link structures. In this way, we evaluate web pages, also taking into consideration their geographic localities.

The remaining part of this paper is organized as follows. Section 2 reviews the related work. Section 3 describes the proposed method. Section 4 shows the experimental results based on our proposed approach. Finally, Section 5 concludes this paper and discusses the future work.

2. Related work

2.1. Link analysis

In this paper, we focus on *a web link analysis* and extend its basic approach to extract spatial information hubs. Among existing approaches, Kleinberg's *HITS* [2] and Google's *PageRank* [6] are the most representative algorithms on link analysis. PageRank enforces a recursive idea that pages are important if important pages link to them. HITS takes a subset of

1 http://www.info-tsukuba.org/educate/index.html
2 http://www.cbrc.jp/etc/hotel.eng.html

Proceedings of the 2005 International Workshop on Challenges in Web Information Retrieval and Integration (WIRI'05)
0-7695-2414-1/05 $20.00 © 2005 IEEE

a web graph and generates hub and authority scores for each page in the subset based on an iterative procedure. A good authority is linked by many good hubs and a good hub links to many good authorities.

In our approach, we propose an extension of HITS [2] so that here we introduce the original HITS algorithm. HITS is performed in two phases. The first phase selects a subset of the web graph and the second phase calculates hub and authority scores based on an iterative procedure. The first phase proceeds as follows:

1) Send a query given from a user to a search engine and take top-ranked pages from the search engine. This step generates a *root set* of web pages.
2) Identify pages which are linked by and link to the pages in the root set.

The set of pages generated in these two steps is called a *base set*. Let $G = (V, E)$ be a graph constructed from the base set, where V and E denote the sets of nodes and edges, respectively.

Algorithm 1 HITS

```
1:  1 := [1, ..., 1] ∈ R^|V|
2:  a_0 := h_0 := 1
3:  t := 1
4:  repeat
5:     for all v ∈ V do
6:        a_t(v) := Σ_{w∈parent[v]} h_{t-1}(w)
7:        h_t(v) := Σ_{w∈child[v]} a_{t-1}(w)
8:     end for
9:     a_t := a_t/||a_t||
10:    h_t := h_t/||h_t||
11:    t := t + 1
12: until ||a_t − a_{t-1}|| + ||h_t − h_{t-1}|| < ε
13: return (a_t, h_t)
```

Algorithm 1 shows the HITS algorithm. The vector $\boldsymbol{h}$ represents the hub scores of the base set pages, while the vector $\boldsymbol{a}$ represents the authority scores. Two vectors are initially set as unit vectors. Line 6 says that "let the authority score of page v be equal to the sum of the hub scores of the pages that link to v", while line 7 says that "let the hub score of page v be equal to the sum of the authority scores of the pages that v links to." Line 9 and 10 ensure that $\boldsymbol{a}$ and $\boldsymbol{h}$ remain to be unit length vectors. After the iterative process, the authority pages are those with the largest corresponding values in $\boldsymbol{a}$, and the hub pages are those with the largest corresponding values in $\boldsymbol{h}$.

2.2. Web and spatial locality

There exist many approaches for extracting pages related to a specified geographic area from the web. For example, a location-based search engine is presented in [8].

[1] proposes the notion of a *localness degree* to discover local information from the web. A localness degree is used to decide whether a page has local information or not. It is estimated based on the content of a target page such as geographic content words, correlation between the target page and its link target pages. We extend the idea of [1] and consider a more sophisticated use of link information. In our approach, we give higher scores to the pages with high localness and to the pages which have links to the ones with high localness.

An *augmented web space* is proposed in [4]. It consists of web pages, hyperlinks, and *generic links* that represent geographic relations between web pages. An example of a generic link is: "If a department store with homepage *A* is near to another department store with homepage *B*, a generic link is created between *A* and *B*." While their approach does not apply generic links to link analysis, we use spatial nodes and spatial links to extend the web space and conduct a link analysis on the extended link structures so as to extract spatial information hubs.

3. Proposed method

3.1. Extracting spatial information from web pages

In our method, web pages are collected initially for the analysis. Then we extract spatial information from these collected web pages and map them to the corresponding coordinates. For the extraction of spatial information, we use zip codes contained in the target web pages. Of course other geographic information may be contained in a web page such as addresses and phone numbers, but accurate extraction of them is quite difficult so that we do not extract other geographic information except for zip codes for the sake of precision and simplicity. In this way, each web page is related to zero or more coordinates.

3.2. Constructing a base set

A link analysis starts when a geographic area is specified from a user. We select pages that contain at least one coordinate (zip code) which is inside of the specified geographic area. Let a *root set* be the set of these selected pages. Like HITS, we add the pages related to the ones in the root set to construct a *base set*. That is to say, we find the pages that are linked by or link to the ones in the root set then consider these pages in addition to the root set pages as the base set. In this way, a subgraph of the entire web space, which consists of web pages related to a specified geographic area, is constructed. The left part of Figure 1 shows an example of a root set and a base set. Nodes 1 to 5 have geographic descriptions related to the specified area and the root set consists of these five pages. Nodes 6 to 11 are the ones that are linked by and link to the root set. These eleven pages compose the base set.

In the original HITS, a root set is constructed based on keywords given by a user. Our approach is different from it since we construct a root set considering localities of web pages.

3.3. Creating spatial nodes and spatial links

In this section, we describe a creation method of spatial nodes and spatial links. Remind that a *spatial node* is a node in the *geographic space* and corresponds to a geographic description (e.g., an address and a zip code) contained in a web page in the base set. Note that for a web page, zero or more spatial nodes may be created depending on its content, but we create only one spatial node instance for the multiple appearance of a same geographic description in multiple web pages; namely, two or more pages "share" a spatial node in such a case.

Next, for each spatial node, bidirectional links are created between the spatial node and the web pages containing the corresponding geographic descriptions. We call this kind of link a *spatial link* because it is based on a spatial reference. In addition to this, in case that the distance between two spatial nodes is below a threshold value, we also create a bidirectional spatial link between them. This kind of spatial link represents their geographic closeness.

A base set is extended by adding spatial nodes and spatial links to the original base set. Below, we call this set an *extended base set*. An example of an extended base set is shown in Figure 1. In this way, if two web pages contain geographic descriptions that correspond to close points in a spatial sense, a spatial relationship will arise between each other even if there is no hyperlink relationship between them.

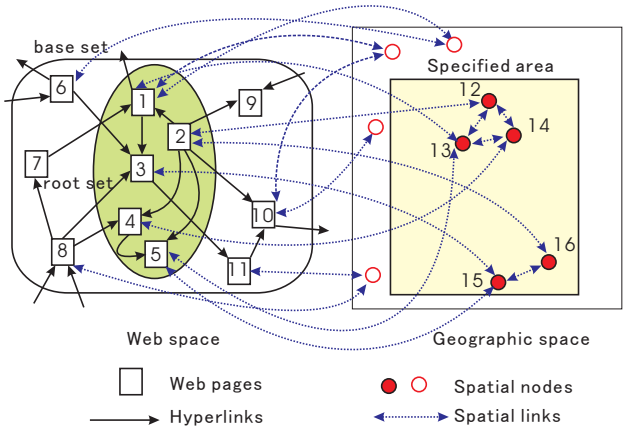

Figure 1. Example of extended base set

The graph structure of an extended base set constructed as mentioned above combines the relationship in the web space reflected by the hyperlinks, and the closeness in the geographic space reflected by the spatial links. In our approach, we conduct a link analysis based on the extended base set. Therefore, the importance and quality of web pages are evaluated according to link structures and geographic contents of web pages.

3.4. Link analysis

3.4.1. Introduction of in_ratio and out_ratio In our approach, spatial nodes and spatial links are created before the process of link analysis. Note that the entire set of spatial nodes contains spatial nodes that are outside of the user-specified geographic area. However, only the spatial nodes in the specified geographic area and the spatial links related to them are considered to be important while conducting an actual link analysis. In addition, we have noticed that it is common that a web page contains two or more spatial descriptions. The distance among these geographic positions may be very far and some of them may be outside of the specified area.

To put high importance on the pages containing much information on the specified area (i.e. having many spatial links to the specified area) while put low importance on the ones containing much information on other areas (i.e. having many spatial links to the areas outside of the specified area), we extend the idea of attaching weights to nodes or links [3] and define the following two factors.

Let $outlinks(v)$ be the total number of links (including hyperlinks and spatial links) which go out from a node v and let $effective\_outlinks(v)$ be the sum of the number of hyperlinks to the base set and that of spatial links to the spatial nodes in the specified area. We define out_ratio as

$$out\_ratio(v) = \frac{effective\_outlinks(v) + 1}{outlinks(v) + 1}. \quad (1)$$

This value has a characteristic that it becomes high when most of links in node v point to the nodes inside of the extended base set, and it becomes low when most of links point to the nodes outside of the extended base set. In order to make this formula calculable even when $outlinks(v) = 0$, we add one's to the numerator and the denominator, respectively.

In the same way, let $inlinks(v)$ be the total number of links that come into a node v and let $effective\_inlinks(v)$ be the sum of the number of hyperlinks which come into node v from the pages in the base set and that of spatial links from the spatial nodes in the specified area. We define in_ratio as

$$in\_ratio(v) = \frac{effective\_inlinks(v) + 1}{inlinks(v) + 1}. \quad (2)$$

Example: We illustrate our approach using Fig. 1. For example, within five links going out from node 1, three links point to the nodes outside of the extended base set. Two links point to the nodes in the extended base set. One (hyper)link points to a base set node (node 3) in the web space. The other (spatial) link points to a spatial node (node 13) in the specified geographic area.

Therefore, for node 1, we get

$$out\_ratio(1) = \frac{2+1}{5+1} = \frac{1}{2}. \quad (3)$$

Similarly, we get

$$in\_ratio(1) = \frac{3+1}{5+1} = \frac{2}{3}. \quad (4)$$

3.4.2. Extension of HITS We extend the HITS algorithm using out_ratio and in_ratio proposed in the above section. The hub and authority scores of web pages that are more relevant to the extended base set are evaluated higher. On the contrary, the web pages that are more relevant to the nodes outside of the extended base set are evaluated lower.

Algorithm 2 shows the extended HITS algorithm. Line 6 and 7 are different from the original HITS in that we attach weights using the two factors mentioned above.

3.4.3. Example of score calculation Table 1 shows the calculation result of hub and authority scores based on the extended base set in Fig. 1. Node 2 is the one whose hub score is the highest. Node 4 and 5 is the second and the third highest ones. See node 5 as an example: the node is impossible to become a hub in the sense of the original HITS. The reason is that the original HITS considers only the web space and this page has no hyperlinks which go out from it. However, in our approach that introduces spatial nodes and spatial links, node 5 is evaluated as a good hub with a high hub score because it has much spatial information on the specified geographic area (i.e. it has many spatial links to the nodes in the specified geographic area.)

Algorithm 2 extended HITS algorithm

$\mathbf{1} := [1, \ldots, 1] \in \mathcal{R}^{|V|}$
$\boldsymbol{a}_0 := \boldsymbol{h}_0 := \mathbf{1}$
$t := 1$
repeat
  for all $v \in V$ **do**
    $\boldsymbol{a}_t(v) := in\_ratio(v) \times \sum_{w \in parent[v]} \boldsymbol{h}_{t-1}(w)$
    $\boldsymbol{h}_t(v) := out\_ratio(v) \times \sum_{w \in child[v]} \boldsymbol{a}_{t-1}(w)$
  end for
  $\boldsymbol{a}_t := \boldsymbol{a}_t / \|\boldsymbol{a}_t\|$
  $\boldsymbol{h}_t := \boldsymbol{h}_t / \|\boldsymbol{h}_t\|$
  $t := t + 1$
until $\|\boldsymbol{a}_t - \boldsymbol{a}_{t-1}\| + \|\boldsymbol{h}_t - \boldsymbol{h}_{t-1}\| < \varepsilon$
return $(\boldsymbol{a}_t, \boldsymbol{h}_t)$

NodeID	Hub Score	Authority Score
1	8	19
2	67	7
3	2	14
4	27	33
5	8	56
6	2	0
7	6	1
8	4	0
9	0	22
10	0	6
11	1	0
12	18	47
13	48	21
14	32	30
15	32	4
16	4	32

Table 1. Hub/authority scores (our method)

In contrast with Table 1, Table 2 shows the result by the original HITS. It is calculated based on the base set in the left part of Fig. 1 without extension. Node 2 is the one which has the highest hub score, which is similar to the result of our approach. The hub score of node 5, however, is evaluated lower (actually the hub score is zero) and the hub score of node 8 is higher than that of ours. In our approach, node 8 is evaluated lower because it does not have related geographic information on the specified area.

NodeID	Hub Score	Authority Score
1	9	43
2	88	0
3	0	22
4	17	50
5	0	43
6	9	0
7	17	15
8	35	0
9	0	36
10	0	43
11	17	0

Table 2. Hub/authority scores (original HITS)

4. Experiments and their analysis

4.1. Preliminary experiment

We use a document dataset `NW100G-01` offered by National Institute of Informatics, Japan. The dataset consists of web pages crawled from the ".jp" domain. It includes $11,038,720$ web pages and $79,699,256$ links between them.

In the preprocessing, we extracted the meta-information of each page and the link information between pages then stored them into a database so as to conduct link analysis more efficiently. For spatial information extraction, we focused on seven-digit Japanese zip code patterns. We translated these extracted zip codes to corresponding coordinates and stored them with corresponding page ids into the database. Actually, $86,315$ zip codes were extracted from the dataset.

doc_id	page content
13039139	A list of travel agencies in Toshima, Tokyo
13513770	A list of associations of real estate appraiser in Toshima, Tokyo
7160587	A list of home care support centers in Toshima, Tokyo
6641959	A list of medical institutions in Tokyo
10565146	A list of goods agencies in Toshima, Tokyo
10585207	A list of NICOS gift cards selling windows in Tokyo

Table 3. Contents of top-ranked hubs in our method (Toshima)

doc_id	s_link	eff_s_link	w_link	eff_w_link	out_ratio	eval
13039139	8	6	0	0	0.7777	Y
13513770	11	6	0	0	0.5833	Y
7160587	12	5	0	0	0.4615	B
6641959	8	2	8	8	0.6470	N
10565146	8	2	0	0	0.3333	N
10585207	20	3	0	0	0.1904	N

Table 4. Classification of links in top-ranked hubs in our method (Toshima)

4.2. Experimental result

4.2.1. Effect of in_ratio and out_ratio First, we examine the difference with/without the use of in/out_ratio factors. Suppose as an example the specified circle area centered by the location "Honchou, Ikebukuro, Toshima, Tokyo (zip code `170-0011`)" with a radius 0.015 [3]. Note that we calculate the distance between two coordinate points approximately by using the latitude and the longitude values and the Euclidean distance. We have tried several settings of τ, the threshold value whether or not to create a spatial link between two spatial nodes. Here we show the experimental result when the threshold value τ is 0.002. In this case, 157 spatial nodes are created. In addition, 296 spatial links between web pages and spa-

3 Tokyo, the capital of Japan, has twenty-three wards. Toshima is one of the wards and Ikebukuro is a crowded area in Toshima ward. Honchou is located in Ikebukuro area.

tial nodes and 896 spatial links between spatial nodes are created.

Table 3 shows the contents of the top-ranked six hub pages using in/out_ratio factors. Classification of links contained in each page is shown in Table 4, where "doc_id" denotes the ID of a document, "s_link" is the total number of the spatial links, while "eff_s_link" means the number of the spatial links which point to the specified area. The label "w_link" is the total number of the hyperlinks, while "eff_w_link" represents the number of the hyperlinks which point to the web pages in the base set. The label "out_ratio" denotes the ratio of the "good" links among all the links (including hyperlinks and spatial links) and "eval" shows the evaluation by the authors whether the page is a good hub page or not. We have evaluated each page by looking at its content and links, where "Y" denotes a page is evaluated as a "good" hub page, while "N" denotes a bad one. The symbol "B" means a border page and we could not judge clearly whether the page is a good hub page or not.

For the comparison purpose, we have also performed experiments with the same specified area and threshold value but in/out_ratio are not used. The experimental result is shown in Table 5 and Table 6. As we can observe from Table 6, if in_ratio and out_ratio are not used, there are only a few spatial links to the specified area in the top-ranked pages although these pages contain numerous hyperlinks and spatial links. They are not the desired ones because they have much more geographic information outside of the specified area than that inside of the specified area.

doc_id	page content
10822643	A list of diving shops and services in 23 wards of Tokyo
9321218	A list of public libraries in Tokyo
11724304	A list of recycle associations and shops over Japan
13039139	A list of travel agencies in Toshima, Tokyo
13513770	A list of associations of real estate appraiser in Toshima, Tokyo
7160587	A list of home care support centers in Toshima, Tokyo

Table 5. Contents of top-ranked hubs without in/out_ratios (Toshima)

doc_id	s_link	eff_s_link	w_link	eff_w_link	eval
10822643	62	4	26	26	N
9321218	371	3	22	22	N
11724304	209	4	31	31	N
13039139	8	6	0	0	Y
13513770	11	6	0	0	Y
7160587	12	5	0	0	B

Table 6. Classification of links in top-ranked hubs without in/out_ratios (Toshima)

For example, the top web page `10822643` in Table 5 and Table 6 has a content of diving shops and services in twenty-three ward areas of Tokyo. It contains only four spatial descriptions on the specified area (the area around the zip code `170-0011`), while sixty-two spatial descriptions appear in the page. It cannot be considered as a good hub page because its ratio of spatial information related to the specified area is low. As shown in Table 3 and Table 4, such pages do not come to the top place in our case. That is to say, they are evaluated lower by our approach with the in/out_ratio factors. Page `13039139` in Table 3 and Table 4 is the top hub page extracted by our approach. This page is a list of travel agencies in the Toshima ward of Tokyo. There are eight spatial descriptions in this page and among them six ones are relevant to the specified area. This page shows relevance on the specified area. Consequently, we consider it is a "good" hub page.

4.2.2. Influence of threshold value τ We examine the influence of changing the threshold value τ that is used to determine whether to create a spatial link between two spatial nodes. As an example, now we consider an area centered by the location "Tennoudai, Tsukuba, Ibaraki, Japan (zip code `300-1253`)" with a radius 0.05[4]. In this case, 179 spatial nodes and 329 spatial links between web pages and spatial nodes are created. We first set the threshold value τ to 0.005 then we get 822 spatial links between two spatial nodes.

Table 7 shows the contents of the top-ranked hub pages when $\tau = 0.005$. The link classification result is shown in Table 8. Page `6608658` and page `6191997` become the first and the second top-ranked ones just because they have so many hyperlinks. However, page `9795990`, the third one, has a different feature; it is a link collection which contains URLs of research organizations, nonprofit foundations, and schools in Tsukuba city. Although there is no geographic information in this page, it has many "good" hyperlinks. Namely, we can find geographic information on the specified area by visiting pages linked from this page.

doc_id	page content
6608658	A list of links to homepages of all ministries over Japan
6191997	A list of links to schools and research institutes over Japan
9795990	A list of links to many websites in Tsukuba
7511229	A List of links to research institutes over Japan
13941808	A list of links to business and industry organizations in Tsukuba
8254577	An introduction of a survey association

Table 7. Contents of top-ranked hubs (Tsukuba, $\tau = 0.005$)

doc_id	s_link	eff_s_link	w_link	eff_w_link	out_ratio	eval
6608658	0	0	630	45	0.0729	N
6191997	0	0	154	24	0.1612	N
9795990	0	0	92	17	0.1935	Y
7511229	0	0	30	9	0.3225	B
13941808	1	1	10	10	1	Y
8254577	0	0	43	1	0.0454	N

Table 8. Classification of links in top-ranked hubs (Tsukuba, $\tau = 0.005$)

In order to compare the influence of changing the threshold value τ, we also show another experimental result with the

4 Ibaraki is a prefecture of Japan. Tsukuba is a city located in this prefecture and is much more desolate than Tokyo. We use Tsukuba city here as the target area, but we observed a similar tendency when we used Toshima ward of Tokyo. The main difference is that we should use a larger radius for Tsukuba city since it is located in the suburban area of Tokyo and more "sparse" than Toshima.

same target area and a larger threshold value $\tau = 0.007$. The number of spatial links between spatial nodes is increased to 1086.

Table 9 and Table 10 show the contents of the top-ranked hub pages when $\tau = 0.007$ and their link classification. For example, page `4968622` in Table 9 and Table 10 with the highest hub score is a summary page of public halls in Tsukuba city. There are 15 spatial links in this page and among them 12 spatial links point to the spatial nodes in the specified area. We consider this page a good hub page because it has much spatial information and the ratio of spatial links pointing to the specified area is high.

Comparing Table 8 and Table 10, we can observe that the pages which have no spatial links but many hyperlinks come to the top place when the threshold value τ is relatively small and as τ increases, the ones which have no hyperlinks but many spatial links become higher. We can adjust the tradeoff of importance between web space and geographic space by changing the threshold value τ.

doc_id	page content
4968622	A list of public halls in Tsukuba
4968829	A list of government offices in Tsukuba
9921832	A list of ATM locations of a bank in Ibaraki
12469677	A list of research institute over Japan
12343866	An introduction of a support organization
7204947	A list of business and industry organizations in Tsukuba

Table 9. Contents of top-ranked hubs (Tsukuba, $\tau = 0.007$)

doc_id	s_link	eff_s_link	w_link	eff_w_link	out_ratio	eval
4968622	15	12	0	0	0.8125	Y
4968829	5	3	0	0	0.6667	Y
9921832	9	3	0	0	0.4	B
12469677	52	10	44	44	0.5670	N
12343866	8	4	5	5	0.7143	N
7204947	9	3	5	5	0.6	B

Table 10. Classification of links in top-ranked hubs (Tsukuba, $\tau = 0.007$)

5. Conclusions and future work

In this paper, we have proposed a method to extract spatial information hub pages, which contain "good" local information and/or "good" hyperlinks to local web pages. We have extended the web link analysis method HITS considering geographic features. In the experiment, we have examined the effectiveness of our approach. We have shown the usefulness of the in/out_ratio factors and observed the influence of changing the threshold value for spatial link creation between spatial nodes.

Our future work is as follows. First, we should develop more sophisticated formulas and parameter tuning methods, for example,

- improved definition of in/out_ratio: In the current approach, the ratio of effective links (hyperlinks and spatial links) is calculated in one factor in a unified manner. We may be able to improve the calculation method by distinguishing two types of links.
- weighting on links: We have simply treated hyperlinks and spatial links equally, but we can assign different weights on hyperlinks and spatial links to improve the effectiveness.
- tuning method of the threshold value τ: We need to develop a (semi-)automatic scheme to determine the threshold value τ.

Second, we need to develop a method based on a different link analysis algorithm. We have used HITS as the basis of our approach, but HITS is not the only candidate. We should consider the possibility to use other algorithms and their comparisons.

Third, we need to develop an effective data collection method. We have performed the experiments using `NW100G-01` dataset, but there was a problem in the dataset. Since the number of collected web pages is limited, many link target pages are not contained in the dataset. Consequently, a link analysis may tend to be unreliable for an area where a limited number of web pages are collected. Therefore, we are now developing a method to efficiently crawl web pages focusing on the ones related to the specified area.

Acknowledgements

This research is partly supported by the Grant-in-Aid for Scientific Research (16500048, 15300027) from Japan Society for the Promotion of Science (JSPS), Japan and the Grant-in-Aid for Scientific Research on Priority Areas (16016205) from the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan, and the program in the Core Research for Evolutional Science and Technology from Japan Science and Technology Agency. In addition, this work is supported by the grants from the Asahi Glass Foundation and the Inamori Foundation.

References

[1] C. Matsumoto, Q. Ma, and K. Tanaka, Web Information Retrieval Based on the Localness Degree, *Proc. DEXA 2002*, LNCS 2453, pp. 172-181, 2002.

[2] J.M. Kleinberg, Authoritative Sources in a Hyperlinked Environment, *JACM*, Vol. 46, No. 5, pp. 604-632, 1999.

[3] K. Bharat and M.R. Henzinger, Improved Algorithms for Topic Distillation in a Hyperlinked Environment, *Proc. SIGIR*, pp. 104-111, 1998.

[4] Kaoru Hiramatsu and Toru Ishida, An Augmented Web Space for Digital Cities. *Proc. SAINT 2001*, pp. 105-112, 2001.

[5] P. Baldi, P. Frasconi, and P. Smyth, *Modeling the Internet and the Web: Probabilistic Methods and Algorithms*, Wiley, 2003.

[6] S. Brin and L. Page, The Anatomy of a Large-Scale Hypertextual Web Search Engine, *Computer Networks and ISDN Systems*, Vol. 30, pp. 1-7, 1998.

[7] S. Chakrabarti, *Mining the Web: Analysis of Hypertext and Semi Structured Data*, Morgan Kaufmann, 2002.

[8] Seiji Yokoji, Katsumi Takahashi, and Nobuyuki Miura, Kokono Search: A Location Based Search Engine, *Proc. WWW 2001*