Experiments on Query Expansion for Internet Yellow Page Services Using Web Log Mining

Yusuke Ohura †, Katsumi Takahashi †‡, Iko Pramudiono †, Masaru Kitsuregawa †

† Institute of Industrial Science, University of Tokyo
4-6-1 Komaba, Meguro-ku, Tokyo
153-8505, Japan
{ohura,katsumi,iko,kitsure}@tkl.iis.u-tokyo.ac.jp

‡ NTT Information Sharing Platform Laboratories
Nippon Telegraph and Telephone Corporation
3-9-11 Midori-cho, Musashino-shi, Tokyo
180-8585, Japan
takahashi.katsumi@lab.ntt.co.jp

Abstract

Tremendous amount of access log data is accumulated at many web sites. Several efforts to mine the data and apply the results to support end-users or to re-design the Web site's structure have been proposed. This paper describes our trial on access logs utilization from commercial yellow page service called "iTOWNPAGE". Our initial statistical analysis reveals that many users search various categories - even non-sibling ones in the provided hierarchy - together, or finish their search without any results that match their queries. To solve these problems, we first cluster user requests from the access logs using enhanced K-means clustering algorithm and then apply them for query expansion. Our method includes two-steps expansion that 1) recommends similar categories to the request, and 2) suggests related categories although they are non-similar in existing category hierarchy. We also report some evaluations that show the effectiveness of the prototype system.

Proceedings of the 28th VLDB Conference, Hong Kong, China, 2002

1 Introduction

1.1 Introduction

The rapid progress on storage capacity and processor performance brought us a chance to analyze huge log data left on Web servers. With early success of "Click-stream" analysis, many research groups and industries are paying more attention to Web log mining techniques. By utilizing those techniques, several proposals have been made to support end-users or to re-design web site. But as far as we know, no technical report on huge log data mining is available to public.

This paper reports results of log data mining and query expansion experiments on the huge commercial Web service called "iTOWNPAGE", an online Japanese telephone directory system. We analyze 450 million lines of iTOWNPAGE log data and create session clusters from 24 million lines of selected log data. Our initial statistical analysis founds that many categories that are not sibling in the given yellow pages hierarchy are searched together in one user session. We also found that many queries fail that no result matched to the user requests. To cope with these problems, we propose a query expansion method using user requests clusters obtained by our enhanced K-means clustering on log data. Our method includes two-steps expansion that 1) recommends similar categories to the request, and 2) suggests non-similar categories in existing hierarchy but found to be related in log-analysis as well. Here we also report the implementation details of our prototype and also some evaluations that prove its effectiveness.

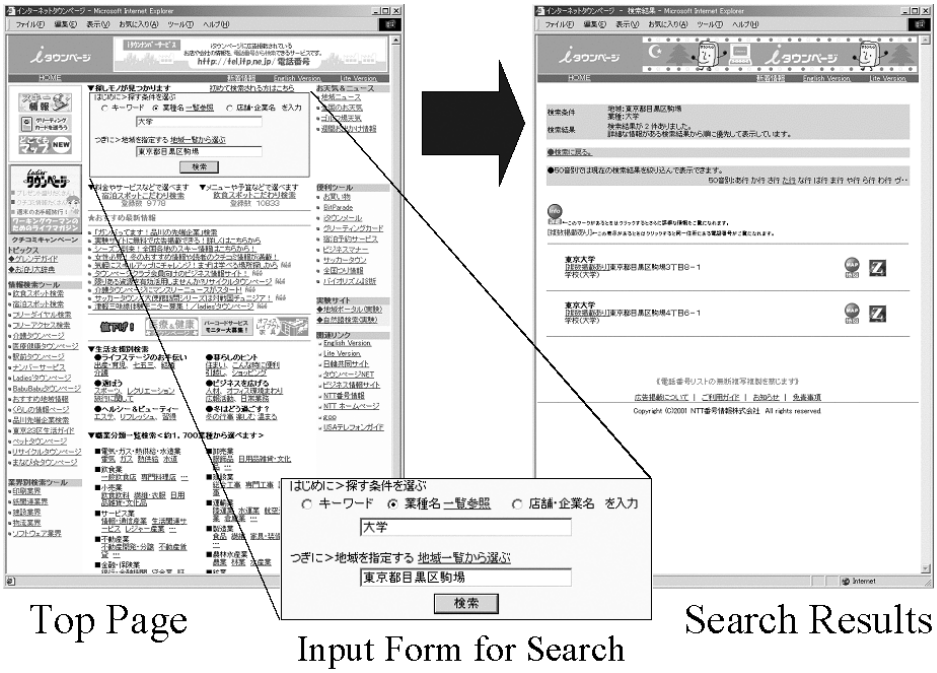

Figure 1: Top Page of iTOWNPAGE and The Result Page

1.2 Related Works

Although many experiments on Web log analysis have been made [1, 2, 3, 4, 5, 6], there are not so many reports for complete process of producing end-user assistant application through Web log mining. Some of those proposals include Web page recommendation and automatic link generation through small log analysis. Yan et al. [7] developed a system for automatic link generation using user-session clusters based on accessed Web pages and their frequency. Mobasher et al. [8] proposed a Web page recommendation system by mining association rules from Web page accesses.

Collaborative filtering[9, 10, 11, 12] is another well-known method for item recommendation applications. But, as it requires user profiles that record user preferences for recommendation, it is not suitable for applications like yellow pages which do not identify the individuals that use it. In this paper, we discuss recommendation methods suitable for anonymous users with no user profiles available.

2 Internet Yellow Page Service and its Problems

This section describes the overview of the target service iTOWNPAGE and problems found after statistical examination.

2.1 iTOWNPAGE

"TOWNPAGE" is known as the national yellow page provided by NTT Directory Services Co. Its Internet version iTOWNPAGE has been available since 1995. Now it is available even on mobile phone services including i-mode. Its 11 million telephone listings cover all Japanese shops, services, and companies. These listings are classified under 2,000 categories. The service records 50 million page-views per month at the end of 2001.

An example of Web iTOWNPAGE is illustrated on Figure 1. The left window is the top page of iTOWNPAGE. Users can search the directory by inputting free key words, categories, or company names, paired with a location. The right window of Figure 1 is the result page that displays listings; company names, addresses, phone numbers with advertisements, URLs, and maps of their location.

If a user wishes to search for a certain category, he/she can input the keywords directly, or browse the category in an alphabetical list or from a category hierarchy. An example of category selection is displayed in Figure 2. The category hierarchy in iTOWNPAGE has about 15 top level nodes, 80 second level nodes, 500 third nodes, and 3,000 leaf category nodes. Note that since there are some alias categories, number of leafs are more than 2,000 actual categories. For example, a category "Hotels" is found by selecting "Leisure Industries" from the top level, and then "Accommodations" from the second level.

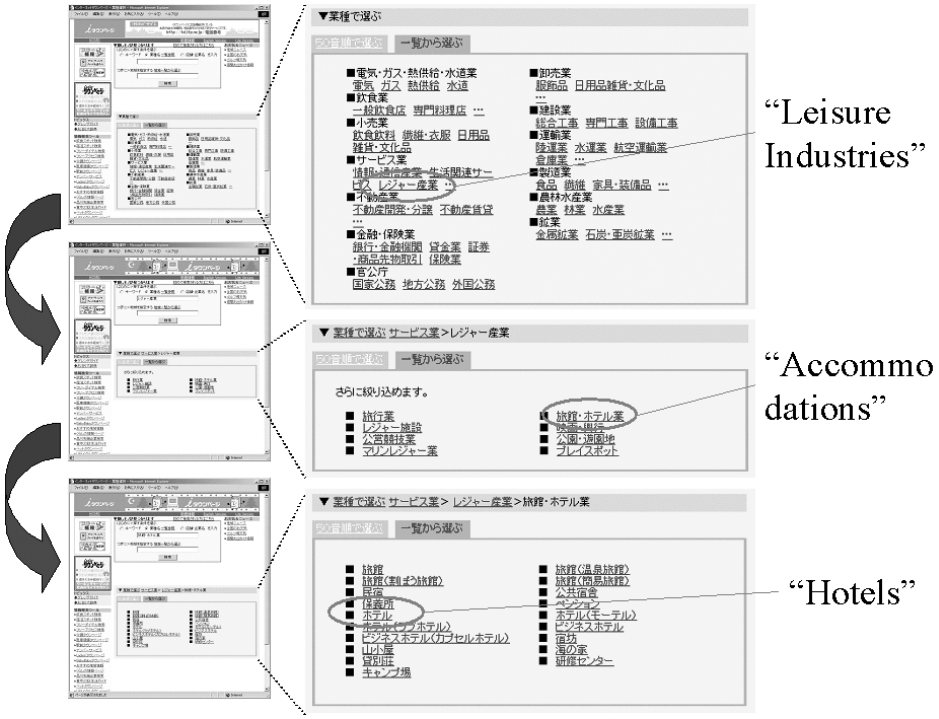

Figure 2: An Example of Category Selection

2.2 Problems Found Through Statistical Analysis

We analyze log data on iTOWNPAGE recorded from 1st February to 30th June 2000. Size of analyzed data was 450 million lines, 200 GB.

Problems found in this examination are reported in this section.

An overview of the search request types on iTOWNPAGE is shown in Figure 3. As is usual with yellow page services, most requests (about 60 %) are searches by a category and an address. Our first research target is to support this kind of searches by helping category selection.

First issue for this target is regarding sessions with multiple categories. A session is a series of search requests from a user (defined in later section). We found 27.2 % of search sessions with category as their variable input are multiple category sessions; that includes more than two different categories in sequence. These multiple category sessions are typical in yellow pages; users might try to look for the category they wish to search (users in trouble) or they are willing to search thoroughly in the yellow pages (good customers of the service provider).

Looking further into these multi category sessions, 75.2 % of them used non sibling categories which do not share the parent in the category hierarchy iTOWNPAGE provides. This may indicate a gap between the category structure expected by users and the one supplied by the designer of the service provider. It can also indicate many users have multiple purposes at the same search session.

It is obvious if a given category hierarchy differs from a users search trend, users may feel inconvenient when they use the service. That is why category hierarchy should be updated frequently. But it is difficult to maintain an ideal category structure at all the time. Therefore it is important to analyze users search trends in a periodic basis, and to have a mechanism that can reflect the trend in how the hierarchy is provided to the users.

Second issue is the case when users can not get any results for their search requests. From the Figure 3, notice that in category-and-address searches, 25 % of the requests return no listings. At this moment, iTOWNPAGE only displays the help message asking the users to modify the address. Since this situation might frustrate the users, a more precise hint that match their need is required.

3 Log Analysis of iTOWNPAGE

We have found that many users request non sibling categories (categories placed in the different node of the hierarchy given by the yellow page) together in the same session. If the category hierarchy does not work well for users to select categories, users might be unhappy. We will examine in detail that trend and discuss our method to overcome the problems. This section describes our mining method for the log data, clustering algorithms, and mining results.

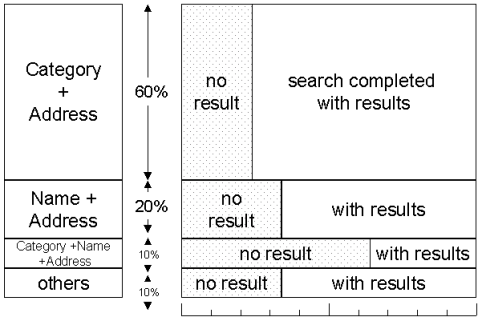

Figure 3: Overview of Search Requests on iTOWNPAGE

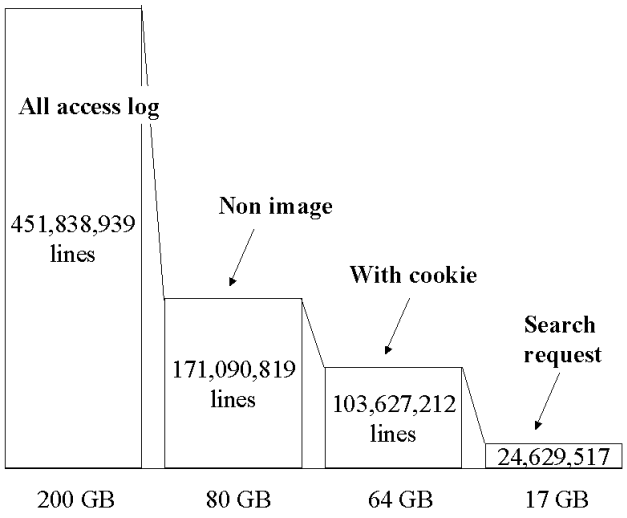

Figure 4: The Size of Log Data

3.1 Log Preprocessing Outline

iTOWNPAGE has standard web access log (apache web server format) and application server log. Access log includes time, remote host name, user agent name, request URI, referrer URL, and cookie ID. Application server log includes time, remote host name, cookie ID, requests for database and the status of search result (number of results). These two logs are joined using cookie ID and some other heuristics. The size of log data is illustrated on Figure 4. The access log for our test is five months log collected from February to June 2000. It consists of about 450 million lines, about 200 GB data in size. In addition, we also have another 20 GB application log. About 62 % of total log are requests for image files and about 60 % have cookie ID. Even after deleting image requests and non-search requests within requests with cookie, data for our mining experiments consists of 24,629,517 lines, about 17 GB in size.

These logs are divided into "sessions"; the sequence of requests from a user. As we employed well-known 30 minutes threshold for the maximum interval [13], two continuous requests within 30 minutes interval are regarded as the same session. For clustering the users sessions, we define user session vector $\vec{s}$. When total category number is N_c, a session vector $\vec{s_i}$ in the i-th session is defined in formula 1 as N_c-dimension vector. Note that we do not take the hierarchy structure of the categories into account.

$$s_{ij} = \begin{cases} 1 & \text{when category j is requested.} \\ 0 & \text{when category j is not requested.} \end{cases} \quad (1)$$

3.2 Clustering by Enhanced K-means

We use the well known K-means algorithm for clustering. However since we can not predict the number of clusters in advance, we improve the algorithm so that it can dynamically decide the number of clusters to be generated. Instead of setting the initial number of clusters K, we define a similarity threshold. Here we describe the algorithm in detail.

When the number of sessions as the input is N, and input vectors $\vec{s_1}$ ... $\vec{s_n}$.

- Initially, the first input vector $\vec{s_1}$ becomes the centroid vector $\vec{c_1}$ of the first cluster C_1 and S_1 becomes the member of the cluster C_1.
- Then for each successive input vector $\vec{s_i}$, the similarity with existing clusters C_1 ... C_k is calculated with formula 2. If the similarity with any cluster is below the similarity threshold TH_{sim}, new cluster is generated and the input vector becomes the centroid cluster of the new cluster. Otherwise when the similarity for some existing clusters is higher than the similarity threshold, the input vector becomes a member of the cluster with the highest similarity. Centroid vector is used when calculating the similarity and it is incrementally recalculated with formula 3 when new members are introduced to the cluster.
- The process is iteratively executed until it converges.

$$SIM(i,j) = \frac{\vec{s_i} \cdot \vec{c_j}}{|\vec{s_i}||\vec{c_j}|} \quad (2)$$

$$\vec{c_j} = \frac{\sum_{S_i \in C_j} \vec{s_i}}{M_j} \quad (3)$$

where M_j is number of members of the cluster C_j.

3.3 Results of Log Clustering

We only use user sessions whose more than two different categories as input. The number of sessions with

Table 1: The distribution of cluster size

Number of Sessions	Number of Clusters
[57 , 1000)	667
[1000 , 2000)	106
[2000 , 3000)	34
[3000 , 4000)	8
[5000 ,)	11

this criterion is 564,355. Some of the generated clusters contains only very few members (sessions). Since those small clusters are insignificant for the result analysis, clusters whose number of members is smaller than minimum cluster size MIN_{cl1} are not counted. In this experiments we set this value to 0.01 % of the number of sessions ($MIN_{cl1} = 56.4$), so we ignore clusters whose user sessions is less than 56. For the similarity threshold TH_{sim} we use the value of 0.1. The number of generated clusters is 826, those clusters ranged in size from as small cluster with 57 members to big cluster with 21,029 members. The size distribution is depicted in Table 1. The average size is 678.8 and the median is 330.

Some clustering results that contain "Hotels" are shown in Table 2. Please refer to the appendix for some other examples. The clusters are shown with the total number of sessions(members), the categories chosen during those user sessions and the number of sessions for each member category. Now we introduce new threshold TH_{cat} and only display categories whose number of sessions are more than TH_{cat} of total sessions for that cluster. For example, cluster 1 in Table 2 consists of 15318 sessions, among them 15318 sessions input "Hotels" and 13654 sessions also input "Business hotels". Other categories are not shown since they only have less than 1531 sessions (10 % of 15318) since TH_{cat} is set to 10 % (0.10).

When we compare the results of clustering and the category hierarchy used at iTOWNPAGE, many non-sibling categories in the category hierarchy appear in the clusters. Figure 5 gives an example, the left side shows a part of category hierarchy while a part of clustering results is shown at the right side. Notice cluster 1 at the bottom right is composed from "Hotels" and "Business hotels". Both are sibling categories, members of "Accommodations" in iTOWNPAGE category hierarchy. On the other hand, cluster 5 at the top right also contains "Hotel bookings" that is a member of "Travel service" in the category hierarchy, differs from the rest of the cluster members such as "Hotels", Business hotels", "Inns", "Spa Inns" that are members of "Accommodations".

Some examples in Table 2 show that searches for "Hotels" are often accompanied by far related non-sibling categories such as "Wedding Halls", "Rent-a-car", "Golf course", "Rental Meeting Rooms" etc.

From the results, we can infer that the search session with the same input such as "Hotels" are performed on various demands and contexts. Some users indeed look for place to stay, while some others look for wedding halls, rent-a-car or meeting rooms. The clustering of web access logs is effective to understand the user behavior. The clustering also confirms that many user sessions input search requests for different categories in the hierarchy. Clusters that only composed of sibling categories are merely 144 out of 826 (16.2 %), while clusters with non-sibling categories are up to 692 (83.8 %).

The phenomenon that the searches stretch over the category hierarchy is the indication of the heterogeneity of user requests or the defect in the site design. If the categories are clearly similar, they can be reflected in next site redesigning. However when the reason behind it is not so obvious, the redesigning will only confuse the users. If we combine "Rent-a-car" and "Hotels" into the same hierarchy, since cluster 7 at Table 2 whose top category is "Rent-a-car" with 1158 sessions and about 10 % of them (120 sessions) also consists of "Hotels", the users will have difficulties to understand the categorization policy of the site.

Here we propose the expansion of the query requested by users to improve the interface. We will give the detail in the following section.

4 Query Expansion Using Web Log Mining

4.1 Motivation

We have described that there are many requests end with no result in section 2. To support these requests, one solution is to recommend another address or to expand requested location to broader area. When we have coordinate information for addresses such as longitude-latitude, location recommendation / expansion is not a difficult task. However we do not have such data, so we concentrate on recommending categories. If we want to support a user query whose request has no result, we need an analysis on the similarity between categories. We extract that information by clustering the user access logs.

We also mentioned that there are many sessions consist of non-sibling categories. These user variations can be another target for our approach. We propose another expansion method for recommending categories, not similar but having some relation to the input category. For example, if "Hotels" is requested, we recommend other categories of accommodations first, then expand to non-sibling but related categories like "Wedding Halls", "Conference Rooms", or "Rent-a-car".

4.2 Strategies for Query Expansion

As mentioned above, we propose a two-step query expansion that recommends categories from user re-

Table 2: Result examples

#	Cluster Size (# of Members)	Category	# of Input
1	15318	Hotels	15318
		Business Hotels	13654
2	3293	Spa Inns	3293
		Spas	1153
		Hotels	1145
		Hot Spring Supply	549
3	2847	Bed and Breakfast	2847
		Inns	2448
		Hotels	899
		Business Hotels	700
		Spa Inns	295
4	1805	Assembly Halls	1805
		Rental Meeting Rooms	719
		Hotels	331
		Auditorium and Assembly Halls	211
5	1628	Hotel Bookings	1628
		Hotels	1387
		Business Hotels	733
		Inns	721
		Spa Inns	193
6	1258	Wedding Halls	1258
		Hotels	609
		Assembly Halls	303
7	1158	Rent-a-car	1158
		Hotels	120
8	811	Golf Course	811
		Hotels	155
9	799	Ski Resort	799
		Hotels	126
10	732	Rental Meeting Rooms	732
		Hotels	215
		Community Halls	94
11	346	Wedding Reception Presentation	346
		Wedding Halls	300
		Assembly Halls	104
		Congratulatory Gifts	51
		Hotels	42

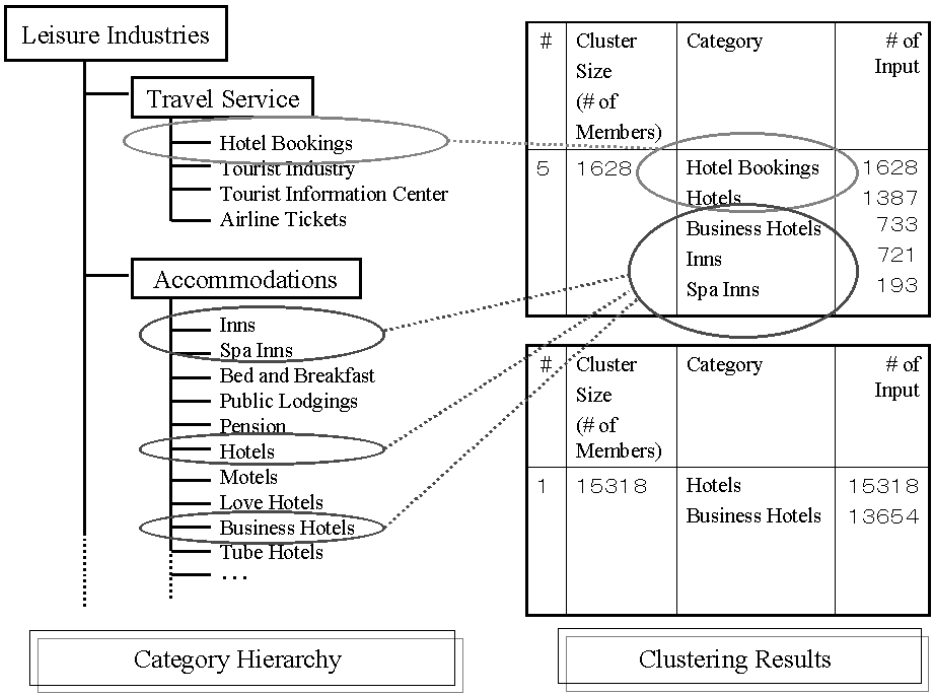

Figure 5: A part of Category Hierarchy and A Part of Clustering Results

quests clusters. The first step is to recommend the sibling categories of the user input category (CAT_{input}). Here we define siblings as categories that are classified in the same sub-category with CAT_{input} in the iTOWNPAGE yellow-page directory. The second step is to recommend non-sibling categories. We name the first one "Intra-Category Recommendation", the second one "Inter Category Recommendation".

Intra-Category Recommendation selects sibling categories that appear in major clusters of CAT_{input}.

1. Find clusters that have CAT_{input} as a member. Calculate the appearance ratio of CAT_{input} to the size of each cluster. Then sort these clusters in the order of the appearance ratio.
2. Choose a sibling category that has the most count from each cluster until the number of sibling categories reaches MAX_{sibl} (=10). If the number of sibling categories is still less than MAX_{sibl}, next sibling categories with most count in each cluster are also displayed until the number reaches MAX_{sibl}.
   Note that we directly refer to the hierarchy defined in the iTOWNPAGE yellow-page directory to decide whether a category is a sibling or not.
3. Clusters whose size is larger than MIN_{cl1} are used for this step. We used 56.4 as the MIN_{cl1} value which equals to 0.01

Ex) When "Hotel" is requested, sibling categories such as "Business Hotel", "Spa Inns", "Bed and Breakfast", and "Inns", are recommended

Inter-Category Recommendation selects non-sibling categories that appear in major clusters of CAT_{input}.

1. Choose the maximum non-sibling category of CAT_{input} from each clusters up to $MAX_{non-sibl}$ (=10) in the same way of "Intra-Category" step.
2. Clusters whose size is larger than MIN_{cl2} are used for this step. We used 564.4 as the MIN_{cl2} value which equals to 0.1 MIN_{cl1}.

Ex) When "Hotels" is requested, non-sibling categories such as "Assembly Halls", "Hotel Bookings", "Wedding Halls", and "Rent-a-car", are recommended.

5 Implementation and Evaluation

5.1 Implementation

We develope a query expansion prototype system using the method proposed in section 4. This system

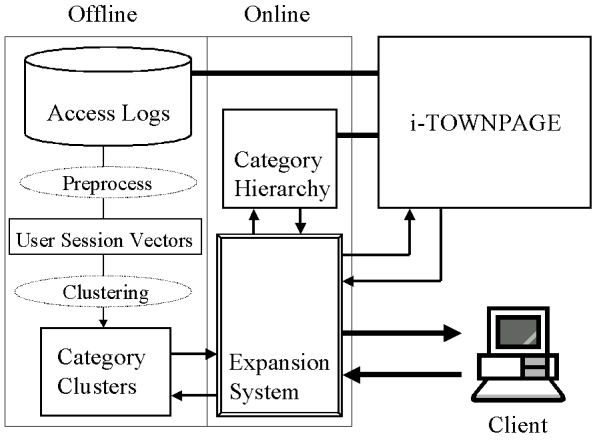

Figure 6: Overview of System Architecture

accepts a category and an address just like original yellow pages system does and displays search results from iTOWNPAGE together with our recommendations. The architecture of the system is illustrated in Figure 6. The system is composed of two modules, online one and offline one. The offline module generates user session vectors first, then creates clusters from these vectors. The online module receives the queries from the users and delegates actual searches to iTOWNPAGE. At the same time, it looks up and displays related categories from the mined clusters to expand the queries.

Search example is illustrated in Figure 7. The left frame displays the expanded categories from our method and the right one is for the query answers from iTOWNPAGE. In the left frame, Intra-Category Recommendation results are displayed on the upper part while results from Inter-Category Recommendation are displayed on the lower part.

5.2 Some Examples

Expanded categories displayed with the search results are anchor strings that users can click. By clicking these links, users can search iTOWNPAGE under the category displayed as links, modify their unsuccessful queries or go through another related search.

Examples of expansion are listed on Table 3. In this experiment, we set the threshold value MIN_{cl2} for Inter-Category Recommendation to 564.4. The value 564.4 equals to 0.1 % of all sessions in this test and clusters that have less than 564 sessions are not considered in this Inter-Category Recommendation. Another threshold TH_{cat} is set to 0.10, only categories requested in more than 10 % of sessions in a cluster are employed for expansion. In the Table 3, categories displayed from Intra-Category Recommendation and Inter-Category Recommendation are shown. For example, Intra-Category Recommendation of "Hotels" are "Business Hotels", "Inns", etc. in other words

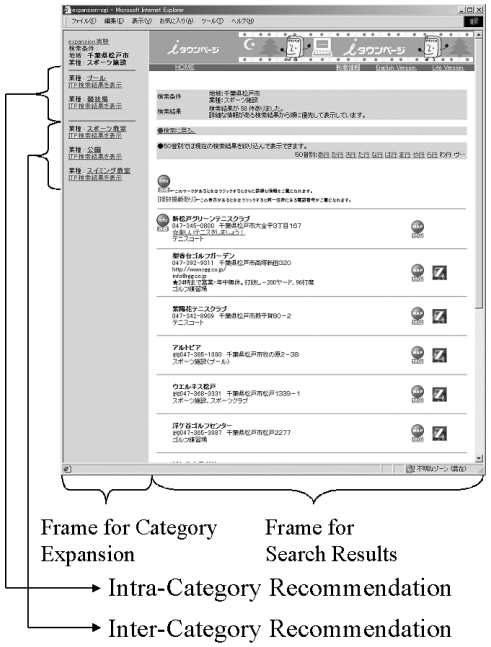

Figure 7: Result Page of Prototype System

sibling categories that share the same upper category "Accommodations". On the other hand, the results show that Inter-Category Recommendation can provide non-sibling but related categories such as "Hotel Bookings", Wedding Halls", "Rent-a-car". These categories could not be recommended using only the category hierarchy in the yellow page services.

5.3 Evaluation

We used another log data from 1st July to 20th July 2000 to test our expansion method. Firstly, test data is converted into sessions such as, "Category A → Category B → Category C". Then transition relations like "Category A → Category B", "Category B → Category C" are extracted from the sessions. Our expansion method is evaluated using these test transition relations. If the right entry of the test relation appeared in expanded categories when the left entry is requested, we call this a successful expansion. When the number of test relations is N, the number of successful expansions after the expansion test is S, expansion success rate is defined in equation below.

$$ExpansionSuccessfulRate = S/N \quad (4)$$

And when the number of expanded categories displayed for i-th test request is C_i, average expanded category number is defined in this way.

$$AverageExpandedCategoryNumber = \sum C_i/N \quad (5)$$

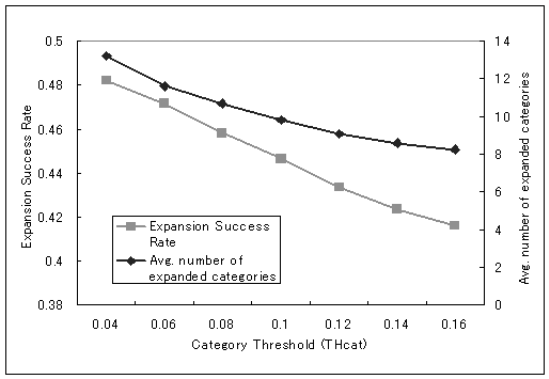

Figure 8: Expansion Success Rate

We obtained 318,899 test relations (N=318,899). For these test data, "Expansion Success Rate" and "Average Number of Expanded Categories" are calculated for various cluster thresholds (TH_{cat}). From the result (Figure 8), our expansion works for more than 40 % of requests. These expansions support users to find categories that they really want to search, or they may have interests in. And this will reduce users burden so that users may not have to browse category hierarchy again to find the suitable category. Users submit search requests 4.21 times during a search session in average. Proposed expansion is expected to reduce some unnecessary accesses.

Our expansion also works well for the searches with no result, since it displays alternatives to the original request. An example is shown in Figure 9. In this example, search on "Tube Hotels" in a city is requested but no result is returned. Our Intra-Category Recommendation displays alternatives such as, "Inns" with 2 results, "Business Hotels" with 3 results, "Hotels" with 10 results. In addition, Inter-category Recommendation gives the user "Sauna" with 4 results, "Bathhouse" with 1 result and "Hotel Bookings" with 2 results.

To evaluate the effect of the expansion for requests with no result, we extracted no-result requests from the test data, tested these requests in our prototype, and calculated the potential number of results brought by recommended categories from our query expansion. In Figure 10, The potential number of results are evaluated for various thresholds TH_{cat}. From the graph (Figure 10), more than 2.5 results can be provided to users that have received no result with their initial queries.

6 Conclusion

This paper reported the experimental results of mining access log from a huge commercial site of Japanese yellow pages, iTOWNPAGE, and proposed a method of query expansion based on user requests clusters.

From the statistical analysis, many users request

Table 3: Examples of Expansion

#	Input Category	Intra-Category Recommendation	Inter-Category Recommendation
1	Hotels	Business Hotels Inns Love Hotels Public Lodgings Pension Spa Inns Bed and Breakfast Simple Hotels Motels Rest Centers	Hotel Bookings Accommodations Wedding Halls Rental Meeting Rooms Golf Course Assembly Halls Ski Resort Rent-a-car Spas Community Halls
2	Swimming Class	Flower Design Class Tennis Class Ceramic Art Class Sports Class Painting Class Cooking Class Golf Class Handicraft Class Dancing Class Knitting Class	Sports Clubs Pools Sports Facilities

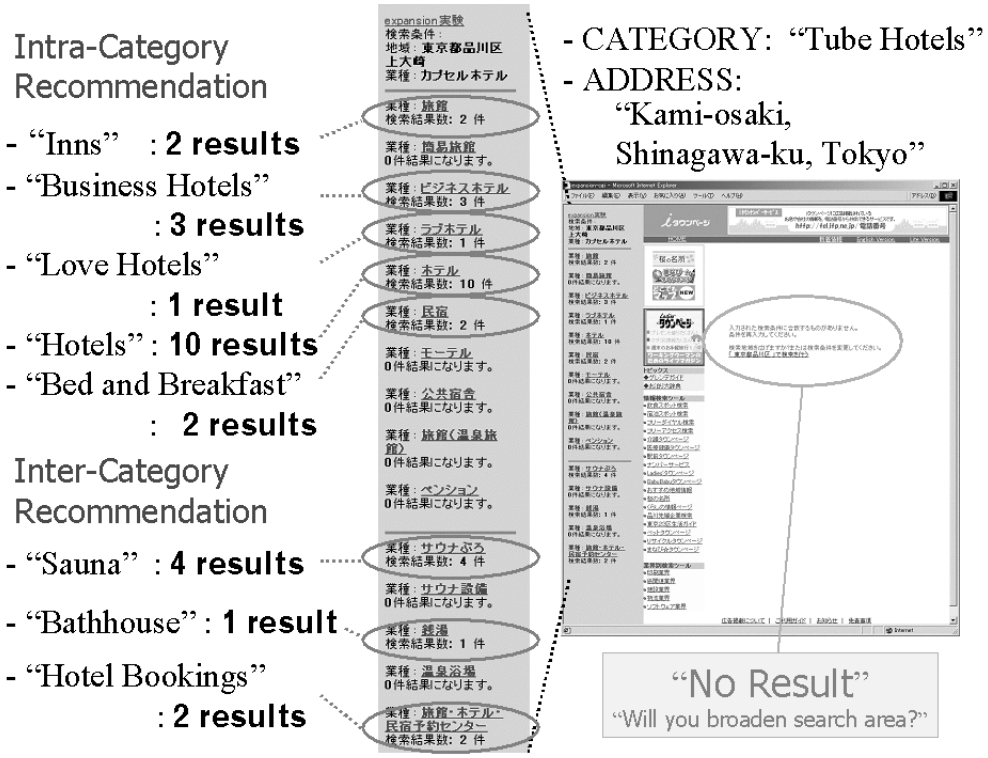

Figure 9: A Case of No Result

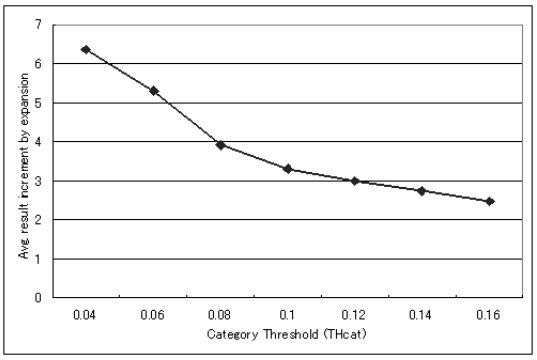

Figure 10: Effectiveness of Expansion against No Result Searches

non-sibling categories together within a session, or fail the searches with no result. To cope with these problems, we proposed a query expansion method based on clustering of user requests. We enhanced K-means clustering algorithm for this puspose so it can dynamically decide the number of clusters. Our expansion method has two-step expansion, enables recommendation for similar categories to the user requests, and recommendation for related categories although they are non-similar in category hierarchy.

We implemented the system of our proposed expansion method and evaluated it. From the results, our expansion works for more than 40 % of requests. These expansions support users to find categories that they could not find or took too much time to find, and also help users find unexpectedly interesting related categories.

Acknowledgment

This work is supported by the joint research between University of Tokyo and NTT. We would like to express our gratitude to NTT Directory Services Co. for providing us with web access log data of NTT iTOWNPAGE. We also would like to thank Mr. Masashi Toyoda of our laboratory for his support and good advice.

References

[1] R. Cooley, J. Srivastava and B. Mobasher: "Web mining: Information and pattern discovery on the world wide web", in *Proc.the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97)* (1997).

[2] C. Shahabi, A. M. Zarkesh, J. Adibi and V. Shah: "Knowledge Discovery from Users Web-Page Navigation", in *Proc. the 7th IEEE Intl. Workshop on Research Issues in Data Engineering (RIDE)*, pp. 20–31 (1997).

[3] O. R. Zaiane, M. Xin and J. Han: "Discovering web access patterns and trends by applying OLAP and data mining technology on web logs", in *Advances in Digital Libraries*, pp. 19–29 (1998).

[4] M. Perkowitz and O. Etzioni: "Adaptive web sites : Automatically synthesizing web pages", *AAAI/IAAI*, pp. 727–732 (1998).

[5] O. Nasraoui, H. Frigui, A. Joshi and R. Krishnapuram: "Mining web access logs using relational competitive fuzzy clustering" (1999). *The Eight International Fuzzy Systems Association World Congress - IFSA 99.*

[6] T. Nakayama., H. Kato and Y. Yamane: "Discovering the Gap Between Web Site Designers' Expectations and Users' Behavior", in *Proc. the 9th International World Wide Web Conference* (2000).

[7] T. W. Yan, M. Jacobsen, H. Garcia-Molina and U. Dayal: "From User Access Patterns to Dynamic Hypertext Linking", in *Proc. the 5th International World Wide Web Conference*, Vol. 28, pp. 1007–1014 (1996).

[8] B. Mobasher, H. Dai, T. Luo and M. Nakagawa: "Effective Personalization Based on Association Rule Discovery from Web Usage Data", in *Proc. the 3rd ACM Workshop on Web Information and Data Management (WIDM01)* (2001).

[9] D. Goldberg, D. Nichols, B. Oki and D. Terry: "Using Collaborative Filtering to Weave an Information Tapestry.", *Comunications of the ACM*, **35**, 12, pp. 61–70 (1992).

[10] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom and J. Riedl: "GroupLens: An Open Architecture for Collaborative Filtering of Netnews", in *Proc. ACM CSCW'94*, pp. 175–186 (1994).

[11] U. Shardanand and P. Maes: "Social Information Filtering: Algorithms for Automating "Word of Mouth"", in *Proc. ACM CHI'95*, pp. 210–217 (1995).

[12] L. H. Ungar and D. P. Foster: "A Formal Statistical Approach to Collaborative Filtering", in *Proc. CONALD'98* (1998).

[13] L. D. Catledge and J. E. Pitkow: "Characterizing Browsing Behaviors on the World Wide Web", *Computer Networks and ISDN Systems*, **27**, 6 (1995).

Appendix

1. Some Results of Clusters (Table 4)
2. Some Examples of Expansion (Table 5)

Table 4: Some Results of Clusters

#	Cluster Size (# of Members)	Category	Number of Input	#	Cluster Size (# of Members)	Category	Number of Input
12	15192	Catering Establishment	15192	19	2400	Convenience Store	2400
		Restaurant	2320			Super Market	1046
13	3522	Beauty Salon	3522	20	2154	Building Industry	2154
		Barbers	1528			Property Deal	250
		Heir Designer	415	21	2131	Sports Club	2131
14	2158	General Hospital	2158			Sports Facilities	1530
		Internal Medicine	254			Pool	502
15	2965	Department Store	2965	22	1713	Drug Shop	1713
		Super Market	1152			Dental Surgery	208
16	2675	Toyshop	2675			Doctor's Office	186
		Game Softs	375	23	1355	Temple	1338
17	2604	Bank	2604			Shrine	396
		Post Office	297			Funeral Industry	238
18	2410	Chinese Noodle	2410	24	1610	Tourist Industry	1610
		Chinese Restaurant	1048			Airline Ticket	367

Table 5: Some Examples of Expansion

#	Input Category	Intra-Category Recommendation	Inter-Category Recommendation
3	Tourist Industry	Airline Tickets	Sight-seeing Bus
		Tourist Information Center	Tourist Hotels
4	Aesthetic Salon	Beauty Salon	Massage
		Nail Salon	Chiropractic
		Beauty Adviser	Cosmetics
		Barbers	
		Laundries	
		Spas	
		Bathhouses	
5	Parks	Amusement Parks	Sports Facilities
6	Bus	Sight-seeing Bus	Railway Industries
		Salon Bus	Traffic Information Service
7	Parks	Amusement Parks	Sports Facilities
8	Do-it-yourself Stores	Hardware Stores	Supermarkets
		Cutting Tool Shops	Carden Supply Shops
		Daily Necessities Shops	Discount Department Stores
		Keys	Flower Shops
			Instruments for Gardening
			Seed and Plant Suppliers
9	Car Parkings	Car Parkings (Monthly Contract)	Property Deal
10	Funeral Industry	Pet Cemeteries	Temple
		Cemeteries	Shrine