WEB community mining and WEB log mining:Commodity Cluster based Execution

Masaru Kitsuregawa **Masashi Toyoda** **Iko Pramudiono**

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

Abstract

The emergence of WWW has drawn new frontiers for database research. Web mining has become a hot topic since WWW rapid expansion rate and chaotic nature have exposed some technical challenges as well as interesting discoveries. In general web mining can be classified into web structure mining and web usage mining. Here we introduce two applications of web mining, first from mining the web structure we identify web communities, and the second we mine web usage of mobile internet users on location aware search engine. Those applications require heavy computational power as well as good scalability. Cluster of commodity PCs is suitable as the platform to handle such applications. Here we also report some approaches for optimal parallel execution of mining algorithms on PC cluster.

Keywords: web mining, web community, PC cluster, parallel mining.

1 Introduction

The exponentially growing WWW implies changes in how people interact each other, how they look for information as well as how they do business. Web mining is quickly gained popularity among researchers and business players because its potential to reveal the behavior of web users.

Here we introduce two applications of web mining, the first is the application of link analysis to extract web communities, and the second is the application of web usage mining to understand the behavior of mobile internet users on location aware search engine.

A web community is a collection of web pages created by individuals or any kind of associations that have a common interest on a specific topic. Such community usually shares some "authority" and "hub" pages, valuable pages that heavily connected by hyperlinks from the members of the community. Some researches on link analysis have been conducted to identify such communities (Kleinberg 1998, Dean & Henzinger 1999, Toyoda & Kitsuregawa 2001). A considerable number of web pages and links is required to perform such analysis. Since nowadays the WWW has exceeded a billion web pages and continues expanding, the task to identify web communities and the relations between those communities is getting more challenging.

We proposed a technique to create a web community chart, that connects related web communities. This allows the user to navigate through related web communities, and can be used for a 'What's Related Community' service that provides not only the web community including a given page but also related web communities.

Our technique is based on a related page algorithm(RPA) that gives related pages to a given page using only link analysis. We show that the algorithm can be used for creating the chart by applying the algorithm to each seed, then using similarities of the results to classify seeds into clusters and to deduce their relationships.

The second application examine the importance of location specific information on the web. Rapid growth of internet access from mobile users puts much importance on such information. An unique web service called Mobile Info Search (MIS) from NTT Laboratories gathers the information and provide location aware search facilities. We performed association rule mining and sequence pattern mining against the access logs which were accumulated at the MIS site in order to get some insight into the behavior of mobile users regarding the spatial information on the web.

Despite its popularity, the real applications of web mining so far are limited. Because of the scale of data from web and its growth rate, significant computational power is required. Many websites are overwhelmed by the accumulation of the access logs only, and could not afford to extract valuable knowledge from the data.

PC cluster is recently regarded as one of the most promising platforms for heavy data intensive applications, such as web mining. We proposed some new parallel algorithms to mine association rule and generalized association rule with taxonomy and showed that PC cluster can handle large scale mining with them.

Section 2 is devoted to the Web community extraction. Section 3 explains the mining of mobile internet user behavior from location aware portal site access logs. Section 4 describes parallel algorithms for Apriori and generalized association rules mining on PC cluster. Section 5 concludes the paper.

2 Web Community Mining

Most research on web communities (Gibson, Kleinberg & Raghavan 1998, Kumar, Raghavan, Rajagopalan & Tomkins 1999) is based on the notion of *authorities* and *hubs* proposed by Kleinberg (Kleinberg 1998). An authority is a page with good contents on a topic, and is pointed to by many good hub pages. A hub is a page with a list of hyperlinks to valuable pages on the topic, that is, points to many good authorities. HITS (Kleinberg 1998) is an algorithm that extracts authorities and hubs from a given subgraph of the Web with efficient iterative calculation.

A set of authorities and hubs was regarded as a community core(Gibson et al. 1998, Kumar et al.

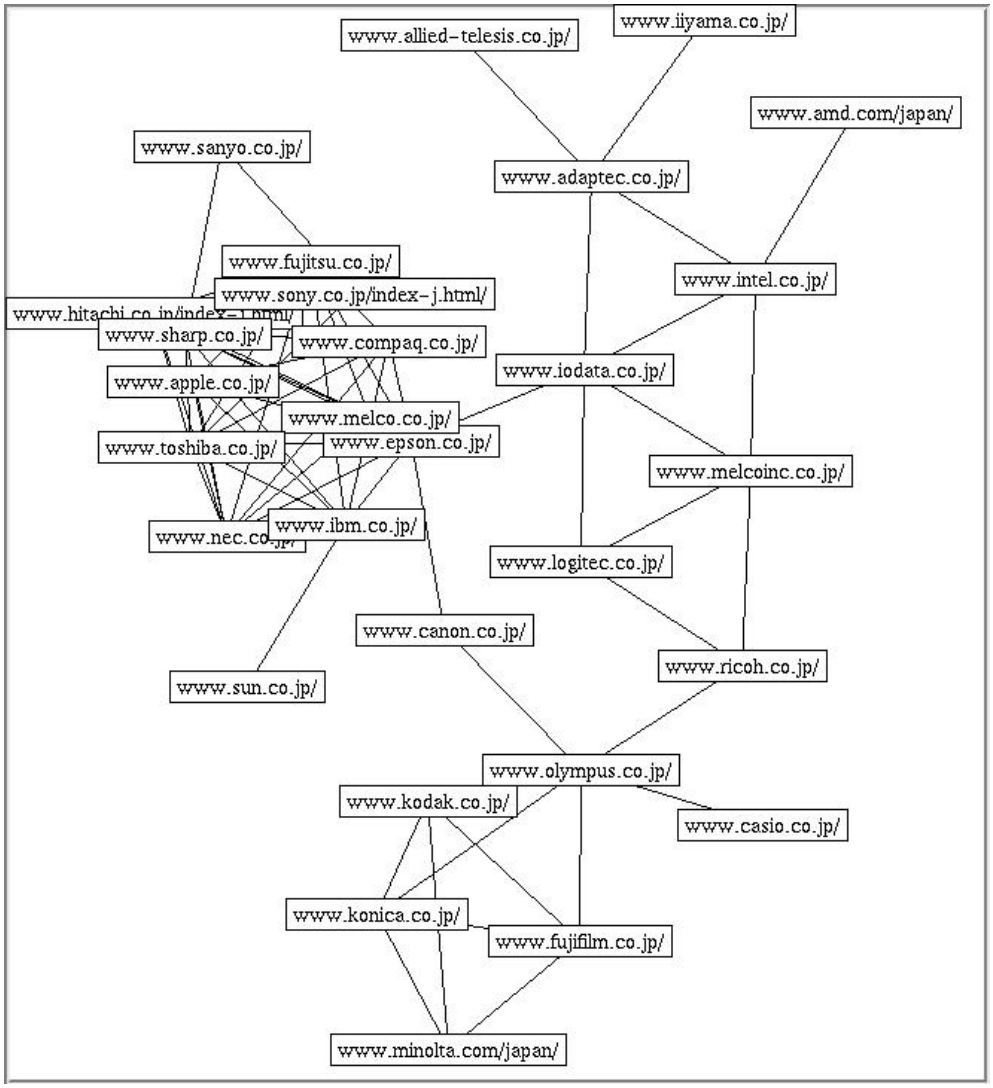

Figure 1: A connected component in the symmetric derivation graph

1999). Given a set of seed pages we can mine web communities from their neighborhood. As the result of web community mining, we generate web community chart. The web community chart is a graph that consists of communities as nodes, and weighted edges between relevant communities. The weight of each edge represents the relevance of communities at both ends.

Since existing RPAs, such as HITS (Kleinberg 1998) and Companion (Dean & Henzinger 1999), provide insufficient precision, we use an improved algorithm, Companion– (Toyoda & Kitsuregawa 2001). We have gained a better precision by using a smaller subgraph than HITS and Companion by ignoring error prone parts.

In this section, we briefly describe our technique to mine web communities. Refer to (Toyoda & Kitsuregawa 2001), for more detailed descriptions.

2.1 Web Community Mining

Our algorithm mine web communities from a given seed set. The main idea is applying a related page algorithm (RPA) to each seed, then investigate how each seed derives other seeds as related pages. RPA first builds a subgraph of the Web around the seed, and extracts authorities and hubs in the graph using HITS. Then authorities are returned as related pages.

To identify web communities and to deduce their relationships, we first put focus on the relationship between a seed page and derived related pages by the algorithm.

Consider that a page s derives a page t as a related page, and t also derives s as a related page. This often means that the both pages s and t are pointed to by similar sets of hubs. For example, a fan page of a baseball team derives other fan pages as related pages. When we apply the related page algorithm to one of the other fans, the page derives the original fan, because those fan pages are mutually linked by each other, that is, pointed to by similar sets of hubs. If each fan derives other fans as related pages, we can consider that these fans form a fan community.

Then, consider that a page s derives a page t as a related page, but t does not derive s as a related page. This means that t is pointed to by many different hubs, so that t derives a different set of related pages excluding s. For example, a fan page of a baseball team often derives an official page of the team as one of related pages. However, when we apply the algorithm to the official page, it derives official pages of other teams as related pages instead of the fan page. This is due to the fact that the official page of the team is often linked together with official pages of other teams in a number of more generic hubs, and the number of such hubs is greater than the number of hubs for the fans. In this case, we can consider that the official page is related to the fan community,

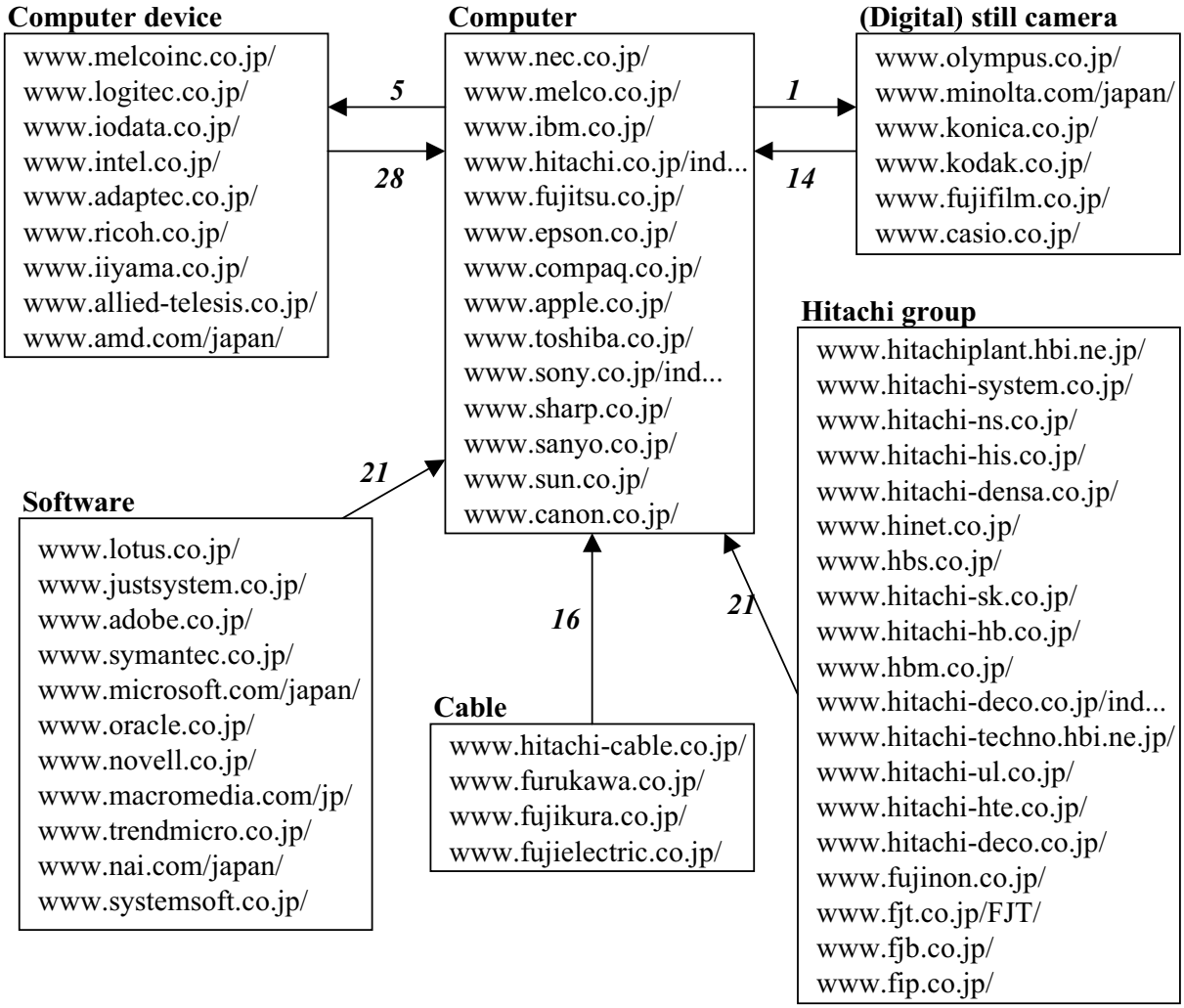

Figure 2: A part of the web community chart

but the page itself is a member of the baseball team community. This is the mechanism by which we find related communities.

Under these observations, we put focus on the former *symmetric derivation relationship* for identifying communities. Using this symmetric relationship, we refine the definition of communities and their relationships. We define that a community is a set of pages strongly connected by the symmetric relationships, and that two communities are related when a member of one community derives a member of an another community.

We depict one of the connected components that includes 29 nodes on multiple categories in Figure 1. In total, this component can be regarded as a community of companies related to computer hardware. However, further observation of this component reveals that it includes three communities. There are computer vendors (NEC, TOSHIBA, SONY, etc.) on the top-left, companies of computer devices (Adaptec, Intel, Logitec, etc.) on the top-right, and companies of digital still camera (OLYMPUS, Minolta, etc.) at the bottom. In this case, we can partition the component into these three communities, by cutting the edge between any two communities.

Since the number of web communities is so big, we also need a tool to browse them, we generate web community chart to visualize the results of web community mining.

2.2 Web Community Chart

Our data set for experiments is an archive of Japanese web pages. The archive includes about 17 million pages in the 'jp' domain, or ones in other domains but written in Japanese characters. We collected these pages from July to September 1999 by running web crawler that collects web pages from given seed pages.

From the archive, we built a connectivity database that can search outgoing and incoming links of a given page. Our database indexed about 120 million hyperlinks between about 30 million pages (17 million pages of pages in the archive, and 13 million pages pointed to by pages in the archive).

In Figure 2, we show a part of the web community chart that consists of communities connected by highly weighted edges. Each box represents a community that includes list of URLs. Note that the category label on each box is attached manually. In a community, each node is assigned a connectivity score that is a number of derivation relationships from the node to other nodes in the community. URLs in the box are sorted by the connectivity score in the descending order. The number attached to each directed edge denotes the weight.

In Figure 2, we chose the 'Computer' community as a center, since it has most edges in the chart. We selected only communities that have more than 15 edges between the 'Computer'. Therefore, there are more communities, that are not shown in Figure 2,

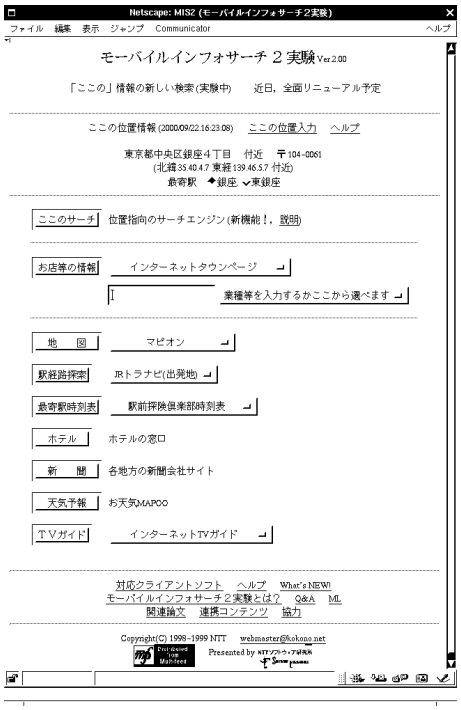

Figure 3: Index page of Mobile Info Search

connected by lower weighted edges. For example, there are communities of computer shops and audio-visual equipment companies around 'Computer'.

As shown in Figure 2, these communities are clearly classified and actually related to the 'Computer' community. On the top of Figure 2, there are three communities that are also in Figure 1. At the bottom, there are three more communities that have only outgoing edges to the 'Computer'. The 'Software' community includes Lotus, Microsoft, Oracle, etc., and obviously related to the 'Computer'. The companies in the 'Cable' community provides cables and optical fibers. The 'Hitachi group' community is slightly different from other communities. Although Hitachi is famous as a computer company, it is also one of the largest conglomerate in Japan. Since, all the companies in the 'Hitachi group' derive 'www.hitachi.co.jp' as one of authorities, the community has a highly weighted edge to the 'Computer'.

3 Web Access Log Mining of Mobile Internet User Portal Site

Here we will report some user behavior analysis results on Mobile Info Search access logs, with association rule mining and sequential rule mining.

3.1 Mobile Info Search(MIS)

Mobile Info Search (MIS) is a research project conducted by NTT Laboratories whose goal to provide location aware information from the internet by collecting, structuring, organizing, and filtering in a practicable form(Takahashi, Yokoji & Miura 2000). MIS employs a mediator architecture. Between users and information sources, MIS mediates database-type resources such as online maps, internet "yellow-pages" etc. using *Location-Oriented Meta Search* and static files using *Location Oriented Robot-based Search*.

The site is available to the public since 1997. Its URL is *http://www.kokono.net*. In average 500 searches are performed on the site daily. A snapshot

Mobile Info Search 2 Ver.2.00
Location Information(2000/09/22 16:23:08) Tokyo, Chuo-ku, Ginza, 4-chome ZIP 104-0061 (NL 35.40.4.7 EL 139.46.5.7) Nearest station : Ginza, Higashi-Ginza
Kokono (nearby area) Search
Shops Information Internet-Townpage Keywords [! type]
Maps Train Route Train Timetable Hotels Newspapers Weather Report TV Guide

Figure 4: Description of MIS index page

of this site is shown in Figure 3 and its description in Figure 4. [1]

3.1.1 User Location Acquisition

Users input their location using address, nearest station, latitude-longitude or postal number. If the user has a Personal Handy Phone(PHS) or Geo Positioning System(GPS) unit, the user location is automatically obtained. The PHS service was launched in Japan in July 1995. Unlike conventional cellular telephone systems, PHS use many small base stations.

Users can also input their location using some map softwares such as "ProAtlas", or softwares to search train routes and schedule such as "Japan Railway(JR) Travel Navigator".

3.1.2 MIS Functionalities

MIS has two main functionalities :

1. Location Oriented Meta Search

   Many local information on the web are database-type, that is the information is stored in backbone database. In contrast to static pages, they are accessed through CGI program of WWW server. MIS provides a simple interface for local information services which have various search interfaces. It converts the location information and picks the suitable *wrapper* for the requested service.

2. Location-Oriented Robot-Based Search "kokono Search"

   kokono Search provides the spatial search that searches the document close to a location. "kokono" is a Japanese word means *here*. *kokono Search* also employs "robot" to collects static documents from internet. While other search engines provide a keyword-based search, *kokono Search* do a location-based spatial search. It displays documents in the order of the distance between the location written in the document and the user's location. For example, since Institute of Industrial Science (IIS), The University of Tokyo is located in Komaba, when a user's location is at the IIS, *kokono Search* will return home pages that contain the word "Komaba" and other addresses in IIS vicinity.

[1] The page is also shown in English at *http://www.kokono.net/english/*

Not so many good restaurants in Akihabara ?
[keyword=][address=Tokyo,][station=Akihabara] ⇒ [shop_cond=restaurant]
In Hokkaido, people looks for gasoline stand at night from its address
[access_hour=20][address=Hokkaido,][from=address] [shop_web=townpage] ⇒ [shop_cond=gasoline]
People from Gifu-ken quite often searches for restaurants
[address=Gifu-ken,][shop_web=townpage] ⇒ [shop_cond=restaurant]
However people from Gifu-ken search for hotels on Saturday
[access_week=Sat][address=Gifu-ken,] [shop_web=townpage] ⇒ [shop_cond=hotel]
People from Gifu-ken must search for hotel around stations
[address=Gifu-ken,][shop_web=townpage] [station=Kouyama] ⇒ [shop_cond=hotel]

Table 1: Some results of MIS log mining regarding search condition

Most frequent searches for restaurants around 16:00 if they start from address on Friday
[access_week=Fri][from=address][shop_cond=restaurant] ⇒ [access_hour=16]
Most frequent searches for department store stand at 20:00 if start from address.
[from=address][shop_cond=department] ⇒ [access_hour=20]
Looking for gasoline stand on Sunday ?
[from=address][shop_cond=gasoline][shop_web=townpage] ⇒ [access_week=Sun]
Search for hotels often from station if at Kanagawa-ken
[address=Kanagawa-ken,][shop_cond=hotel] ⇒ [from=station]
People at Osaka start searching convenience stores from ZIP number !
[address=Osaka,][shop_cond=conveni] ⇒ [from=zip]
People at Hokkaido always search convenience stores from address
[address=Hokkaido,][shop_cond=conveni] [shop_web=townpage] ⇒ [from=address]

Table 2: Some results of MIS log mining regarding time and location acquisition method

3.2 Association Rule Mining

Agrawal et. al.(Agrawal, Imielinski & Swami 1993, Agrawal & Srikant 1994) first suggested the problem of finding association rule from large database. An example of association rule mining is finding "if a customer buys A and B then 90% of them buy also C" in transaction databases of large retail organizations. This 90% value is called confidence of the rule. Another important parameter is support of an itemset, such as {A,B,C}, which is defined as the percentage of the itemset contained in the entire transactions. For above example, confidence can also be measured as support({A,B,C}) divided by support({A,B}).

In most cases, items can be classified according to some kind of "is a" hierarchies (Srikant & Agrawal 1995). When such hierarchy exist, association rule mining is often cited as generalized association rule mining with taxonomy. Since names of places follow some kind of hierarchy, such as "city is a part of prefecture" or "a town is a part of a city", we introduce taxonomy between them. The introduction of the hierarchy allows us to find not only rules specific to a location but also wider area that covers that location. This is useful since in many case the locations specified as search condition are sparsely distributed. We show some results in Table 1 and 2. Derived association rules can be used to improve the value of web site. We can identify from the rules some access patterns of users that access this web site.

For example, from the first rule we know that though Akihabara is a well known place in Tokyo for electronic appliances/parts shopping, user that searches around Akihabara station will probably looks for restaurant. From this unexpected result, we can prefetch information of restaurant around Akihabara station to reduce access time, we can also provide links to this kind of user to make his navigation easier or offer proper advertisement banner. In addition, learning users behavior provides hint for business chance for example the first rule tell us the shortcoming of restaurants in Akihabara area.

Other search results show how location affects the search conditions. The second rule in Table 1 shows that people in Hokkaido, the largest and the most sparse prefecture in Japan, has particular problem to find gasoline stand at night. The rest of the rules show how people at Gifu-ken, a modest prefecture in the middle of Japan, often looks for restaurants. However more people, some of them might be travellers, looks for hotel around the station in the weekend.

Some results in Table 2 show that in addition to the location, time and location acquisition method might affect search conditions. For example, the third rule indicates that hotels in Kanagawa-ken are more likely searched from their nearest station since Kanagawa-ken, being the suburb of Tokyo, has extensive railway. In contrast, the last rule shows that people at Hokkaido are more comfortable to find convenience stores from the address.

After finding a shop, check how to go there and the weather
[submit_shop=Shop Info] → [submit_rail=Search Train] → [submit_newspaper=Newspaper] ⇒ [submit_weather=Weather Forecast]
Or decide the plan after checking the weather first
[submit_weather=Weather Forecast] → [submit_shop=Shop Info] [shop_web=townpage] → [submit_kokono=Kokono Search] ⇒ [submit_map=Map]
Looking for shops after closing time
[submit_shop=Shop Info] [access_hour=22] [access_week=Fri] ⇒ [submit_map=Map] [access_hour=22] [access_week=Fri]

Table 3: Some results of sequential pattern mining

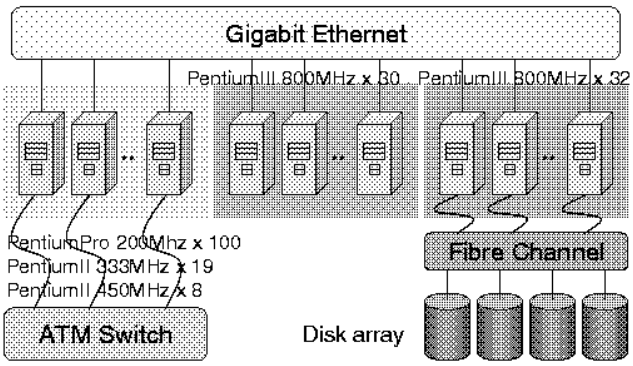

Figure 5: PC cluster

We found that the spatial information is highly valuable to derive users' preferences, in particular mobile users. Our rules show that items with location information such as *address* and *station* increase the confidence of the rules significantly.

3.3 Sequential Rule Mining

The problem of mining sequential patterns in a large database of customer transactions was also introduced by Agrawal et. al.(Agrawal & Srikant 1995).

The transactions are ordered by the transaction time. A sequential pattern is an ordered list (sequence) of itemsets such as "5% of customers who buy both A and B also buy C in the next transaction".

We show some sequential patterns that might be interesting in Table 3. Some patterns indicate the behavior of users that might be planning to do shopping. We can derive from second pattern that significant part of users check the weather forecast first, then they look for the shops in the yellow-pages service called "Townpage" then look again for additional information in the vicinity with *kokono Search* and finally they confirm the exact location in the map.

From sequential rules we can also derive how different services are used together. In particular, we are interested in how users act to information in their vicinity gathered by the *kokono Search*.

4 Parallel Data Mining on Large Scale PC cluster

We propose PC cluster as the platform for web mining. Here we explain our PC cluster system and some researches on large scale parallel data mining on the PC cluster.

4.1 PC Cluster

We have developed a large scale PC cluster consists of 128 PCs interconnected with 155 Mbps ATM and 10 Mbps Ethernet networks(Tamura et al. 1997). The project was launched in 1995 and the equipments came at the end of 1996. The system started at February 1997.

Initially the PC cluster was made up of 100 PCs with 200 MHz Pentium Pro only and then we have added another 19 nodes but with more powerful 333 MHz Pentium II, 8 nodes with 450 MHz Pentium II, and 62 nodes with 800 MHz Pentium III since the performance of PC hardware had improved dramatically. Additional 32 nodes with 1.5 GHz Pentium 4 are coming soon.

Those PCs are interconnected with 128 ports 100 Mbps ATM switch, 192 ports Gigabit Ethernet and Fibre Channel. The configuration of the PC cluster is depicted in Figure 5. The details of the development of this system has been written in (Tamura et al. 1997).

4.2 Highly Parallel Data Mining Algorithms

J.S.Park, et.al proposed bit vector filtering for association rule mining and naive parallelization of Apriori (Agrawal & Shafer 1996, Park, Chen & Yu 1995), where every node keeps the whole candidate itemsets and scans the database independently. Communication is necessary only at the end of each pass. Although this method is very simple and communication overhead is very small, memory utilization efficiency is terribly bad. Since all the nodes have the copy of all the candidate itemsets, it wastes memory space a lot.

Hash Partitioned Apriori(HPA) was proposed in (Shintani & Kitsuregawa 1996). The candidate itemsets are not copied over all the nodes but are partitioned using hash function. Then each node builds hash table of candidate itemsets. The number of itemsets at second pass is usually extremely high, sometimes three orders of magnitude larger than the first pass in a certain retail transaction database which we examined. When the user-specified support is low, the candidate itemsets overflow the memory space and incur a lot of disk I/O.

While reading transaction data for support counting, HPA applies the same hash function to decide where to send the transactions and then probe the hash table of candidate itemsets to increase the count. Although it has to exchange transaction data among nodes, utilization whole memory space through partitioning the candidates over nodes instead of duplication results in better parallelization gain.

Hybrid approach between candidate duplication and candidate partitioning is proposed at (Han, Karypis & Kumar 1997) at 1997. The processors are divided into some number of groups. Within each group, all the candidates are duplicated and among groups, candidates are partitioned.

4.2.1 Parallel Algorithms for Generalized Association Rule Mining

Here, we describe our parallel algorithms for finding all large itemsets on shared-nothing environment proposed in (Shintani Kitsuregawa 1998).

Non Partitioned Generalized association rule Mining : NPGM

NPGM copies the candidate itemsets over all the nodes. Each node can work independently.

Hash Partitioned Generalized association rule Mining : HPGM

HPGM partitions the candidate itemsets among the nodes using a hash function like in the hash join, which eliminate broadcasting.

Hierarchical HPGM : H-HPGM

H-HPGM partitions the candidate itemsets among the nodes taking the classification hierarchy into account so that all the candidate itemsets whose root items are identical be allocated to the identical node, which eliminates communication of the ancestor items. Thus the communication overhead can be reduced significantly compared with original HPGM.

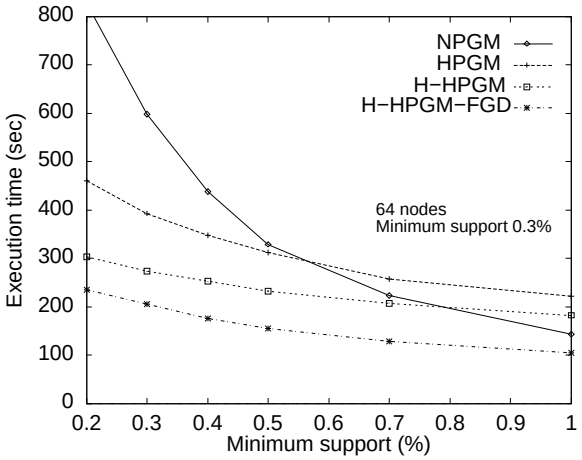

Figure 6 : Execution time

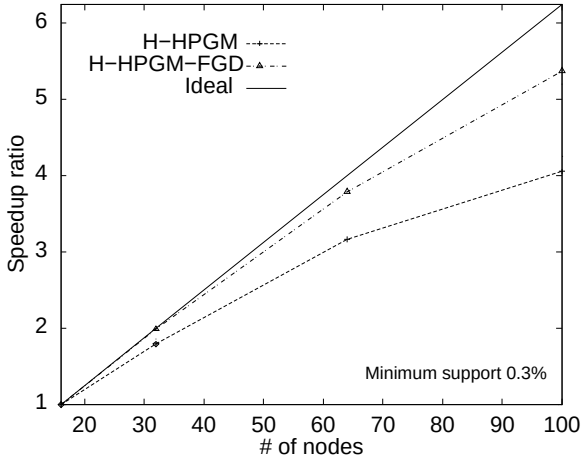

Figure 7 : Speedup ratio

H-HPGM with Fine Grain Duplicate: H-HPGM-FGD

In the case the size of the candidate itemsets is smaller than available system memory, H-HPGM-FGD utilizes the remaining free space. H-HPGM-FGD detects the frequently occurring itemsets which consists of the any level items. It duplicates them and their all ancestor itemsets over all the nodes and counts the support count locally for those itemsets like in NPGM.

4.2.2 Performance Evaluation Results

We use synthetic transaction data generated using procedure in (Agrawal & Srikant 1994). For large scale experiments of generalized association rules we use the following parameters : (1)the number of items is 50,000, the number of roots is 100, the number of levels is 4–5, fanout is 5, (2)the total number of transactions is 20,000,000(1GBytes), the average size of transactions is 5, and (3)the number of potentially large itemsets is 10,000. We use 100 nodes with PentiumPro CPU for this experiment.

We show the execution time at pass 2 of all parallel algorithms varying the minimum support in Figure 6. The execution time of all the algorithms increases when the minimum support becomes small. When the minimum support is small, the candidate partitioned methods can attain good performance. H-HPGM-FGD significantly outperforms other algorithms.

Figure 7 shows the speedup ratio with varying the number of nodes used 16, 32, 64 and 100. The curves are normalized by the execution time of 16 nodes system. H-HPGM-FGD attains higher linearity than H-HPGM. Since H-HPGM duplicates no candidate itemsets, the workload skew degrades the linearity. The skew handling methods detect the frequently occurring candidate itemsets and duplicate them so that the remaining free memory space can be utilized as much as possible.

5 Conclusion

Real web mining applications require high performance computing system. Here we presented two examples of such applications, web community extraction from a significant set of WWW in Japan domain and web usage mining of a portal site of mobile internet users.

PC cluster can offer sufficient performance with reasonable cost. It is also scalable since the system configuration can be easily expanded when needed.

We examined the effectiveness of parallel algorithms on large scale parallel PC cluster. Our experiments have showed that PC cluster, with its scalable performance and high cost performance is a promising platform for data intensive applications such as web mining.

References

Agrawal, R., Imielinski, T. & Swami, A. (1993), Mining association rules between sets of items in large databases, *in* 'ACM SIGMOD International Conference on Management of Data', Vol. 22, ACM Press, Washington DC, USA, pp. 207–216.

Agrawal, R. & Shafer, J.C. (1996), Parallel Mining of Associaton Rules *in* IEEE TKDE, Vol. 8, No. 6, pp. 962–969

Agrawal, R. & Srikant, R. (1994), Fast Algorithms for Mining Association Rules. *in* 'Proceedings of VLDB', pp. 487–499.

Agrawal, R. & Srikant, R. (1994), Mining Sequential Patterns, *in* 'Proceedings of Int. Conf. on Data Engineering'.

Dean, J. & Henzinger, M. R. (1999), Finding related pages in the World Wide Web *in* 'Proceedings of the 8th World-Wide Web Conference'

Gibson, D. , Kleinberg, J. M. & Raghavan, P. (1998), Inferring Web Communities from Link Topology *in* 'Proceedings of HyperText 1998'

Han, E.-H., Karypis, G. & Kumar, V. (1997), Scalable Parallel Data Mining for Association Rules, *in* 'ACM SIGMOD International Conference on Management of Data', pp. 277–288.

Kleinberg, J. M. (1998), Authoritative Sources in a Hyperlinked Environment *in* 'Proceedings of the ACM-SIAM Symposium on Discrete Algorithms'

Kumar, R. , Raghavan, P., Rajagopalan, S. & Tomkins, A. (1999), Trawling the Web for emerging cyber-communities *in* 'Proceedings of the 8th World-Wide Web Conference'

Park, J.S., Chen, M.-S. & Yu, P.S. (1995), Efficient Parallel Algorithms for Mining Association Rules *in* 'Proceedings of CIKM', pp. 31–36

Shintani, T. & Kitsuregawa, M. (1996), Hash Based Parallel Algorithms for Mining Association Rules. *in* 'Proceedings of PDIS', pp. 19–30

Shintani, T. & Kitsuregawa, M. (1998), Parallel Mining Algorithms for Generalized Association Rules with Classification Hierarchy. *in* 'ACM SIGMOD International Conference on Management of Data',pp. 25–36

Srikant, R. & Agrawal, R. (1995), Mining Generalized Association Rules *in* 'Proceedings of VLDB'

Takahashi, K., Yokoji, S. & Miura, N. (2000), Location Oriented Integration of Internet Information - Mobile Info Search *in* ' Designing the Digital City' Springer-Verlag

Tamura, T., Oguchi, M. & Kitsuregawa, M. (1997), Parallel Database Processing on a 100 Node PC Cluster: Cases for Decision Support Query Processing and Data Mining. *in* 'Super Computing 97::High Performance Networking and Computing'

Toyoda, M. & Kitsuregawa, M. (2001), Finding related pages in the World Wide Web *in* 'Proceedings of Hypertext 2001' pp. 103–112.