國立中山大學九十學年度博士班招生考試試題

科目：論文評述（資訊科技）【資管所】第一節　　共　頁第　頁

Please read the paper "Correlation-based Document Clustering using Web Logs" and answer the following questions.

1. Please identify the research contributions of this paper. (10%)

2. What are the main differences between the DBSCAN algorithm (as shown in page 3) and the RDBC algorithm (as shown in page 4)? (15%)

3. What are the limitations or drawbacks of the proposed approach? (10%)

4. Is the evaluation discussed in the paper appropriate? Why? Please propose an evaluation plan for evaluating the approach (or technique) proposed in the paper. (15%)

第一节

Correlation-based Document Clustering using Web Logs*

Zhong Su[1], Qiang Yang[2], Hongjiang Zhang[3], Xiaowei Xu[4], Yuhen Hu[5]
[1]Department of Computing Science, Tsinghua University, Beijing 100084, China
[2] School of Computing Science, Simon Fraser University, Burnaby, BC Canada V5A 1S6
[3] Microsoft Research China, 5F, Beijing Sigma Center, Beijing 100080 P.R. China
[4] Siemens AG, Information and Communications Corporate Technology, D-81730 Munich, Germany
[5] Department of E&E, University of Wisconsin-Madison, Madison, WI 53706 USA
suzhong_bj@hotmail.com[1], qyang@cs.sfu.ca[2], hjzhang@microsoft.com[3],
xiaowei.xu@mchp.siemens.de[4], Hu@engr.wisc.edu[5]

Abstract

A problem facing information retrieval on the web is how to effectively cluster large amounts of web documents. One approach is to cluster the documents based on information provided only by users' usage logs and not by the content of the documents. A major advantage of this approach is that the relevancy information is objectively reflected by the usage logs; frequent simultaneous visits to two seemingly unrelated documents should indicate that they are in fact closely related. In this paper, we present a recursive density based clustering algorithm that can adaptively change its parameters intelligently. Our clustering algorithm RDBC (Recursive Density Based Clustering algorithm) is based on DBSCAN, a density based algorithm that has been proven in its ability in processing very large datasets. The fact that DBSCAN does not require the pre-determination of the number of clusters and is linear in time complexity makes it particularly attractive in web page clustering. It can be shown that RDBC require the same time complexity as that of the DBSCAN algorithm. In addition, we prove both analytically and experimentally that our method yields clustering results that are superior to that of DBSCAN.

1. Introduction

A problem facing information retrieval on the web is how to effectively cluster large amounts of web documents. One approach is to cluster the documents based on information provided only by users' usage logs and not by the content of the documents. A major advantage of this approach is that the relevancy information is objectively reflected by the usage logs; frequent simultaneous visits to two seemingly unrelated documents should indicate that they are in fact closely related. In this paper, we present an efficient algorithm for clustering large sets of web documents based on distance measures that are provided by only the server log data.

There is a great deal of work done previously in clustering, including K-means [6], HAC[3][12][1], CLANRNS [11] etc. In the IR community, the Scatter/Gather algorithm [5] is aimed at re-organizing document search results by examining document contents. It is similar to K-means in that it requires pre-set cluster number, which is a requirement that we do not assume in our paper. Suffix-Tree [14] is another closely related clustering method. Its input is also portions of the document contents and thus is different from the problem we face.

Because we only have server log information, we can build a distance metric similar to that by [9]. Based on this distance information, we choose to extend DBSCAN [7], an algorithm to group neighboring objects of the database into clusters based on local distance information. It is very efficient because only one scan through the database is required. Moreover, it does not require a predetermined cluster number to operate.

DBSCAN constructs clusters using distance transitivity based on a density measure defined by the user. Documents that have many co-visited documents around them are considered dense. DBSCAN performs this clustering using a fixed threshold value to determine "dense" regions in the document space. Because this threshold value is constant across all points in the space, the algorithm often cannot distinguish between dense and loose points, and as a consequence, often the entire document space is lumped into a single cluster.

* This work was performed in Microsoft Research China

In this paper we present a recursive density-based clustering algorithm for web document clustering. Our only source of information is web log data that records users' document access behavior. We wish to use this information to construct clusters that represent closely related documents where the relevancy information cannot be observed by simply examining the documents themselves. One requirement is that we must not predetermine the number of clusters and that we must use as little initial information as possible.

To meet this need, our algorithm can adaptively change its parameters intelligently. The algorithm is based on DBSCAN and is applicable to any database containing data from a metric space, e.g., to a web log database. Our clustering algorithm calculates a density measure based on the distance metrics that is computed from the web logs according to our distance definition. It then selects the points that are dense enough in the space of distance metrics and constructs an abstract space based on these points. It does this recursively until no more abstraction space can be built. Because it can change the parameters intelligently during the recursively process, RDBC can yield results superior than that of DBSCAN. It can be shown that RDBC requires the same time-complexity as that of the DBSCAN algorithm. In addition, we show experimentally that our method yields clusters that are more superior than that of DBSCAN on the same web logs.

The remainder of this paper is organized as follows. We discuss previous work in this area and provide background on the clustering work before briefly introducing the clustering algorithm DBSCAN. In section 3, we present RDBC, our recursive density based clustering algorithm. In section 4, we describe our web-document clustering algorithm based on the web logs. In section 5, we experimentally evaluate variants of RDBC on three realistic web server logs, and compare the performance of RDBC to DBSCAN. We conclude with a discussion of future work and a summary of our contributions.

2. Clustering Background

There is a great deal of work done previously in clustering. Typical of the clustering work are the K-means clustering and hierarchical agglomerative clustering (HAC). K-means constructs a partition of a database of n objects into a set of k clusters where k is an input parameter. Each cluster is represented by the center of gravity of the cluster (k-means) or by one of the objects of the cluster located near its center (k-medoid) and each object is assigned to the cluster with its representative closest to the considered object. Typically, this algorithm starts with an initial partition of database and then uses an iterative control strategy to optimize the clustering quality. However, K-means requires that the user provide the number K of clusters as initial input.

HAC creates a hierarchical decomposition of a database. The hierarchical decomposition is represented by a dendro-gram, a tree that iteratively splits database into smaller subsets consists of only one object. In such a hierarchy, each level of the tree represents a clustering of database. It works as follows. Initially, each object is placed in a unique cluster. For each pair of clusters, some value of dissimilarity or distance is computed. In every step, the clusters with the minimum distance in the current clustering are merged until all points are contained in one cluster.

The density-based method DBSCAN [7] is very efficient to execute and does not require the user to pre-specify the number of clusters. The latter is a major advantage in our application. Density-based methods are based on the idea that it is likely that in a space of objects, dense objects should be grouped together into one cluster. Thus, a cluster is a region that has a higher density of points than its surrounding region. For any points in a space, where a point corresponds to a web page, the more web pages that co-occur with it, the higher its density is.

DBSCAN, as introduced in [7], is a type of single scan clustering algorithm. The basic idea of this algorithm is to group neighboring objects of the database into clusters based on a local cluster condition, thus performing only one scan through the database. It is very efficient if the retrieval of the neighborhood of an object is efficiently supported by the DBMS. So it is the most efficient algorithm on large database. It just assumes a distance function. It can deal with any arbitrary shapes of data distribution.

More specifically, DBSCAN accepts a radius value ε based on a user defined distance measure, and a value **Mpts** for the number of minimal points that should occur in around a dense object; the latter is used to determine, out of many points in a space, which region is considered dense. DBSCAN then iteratively computes the density of points in an N-dimensional space, and groups the points into clusters. Next, we provide more precise definitions for the definitions of clusters.

First we define the ε-neighborhood of a point as the set of points that are within ε distance from the point.

Definition 1: (ε-neighborhood of a point) [7]

the ε-*neighborhood* of a point p, denoted by $N_\varepsilon(p)$, is defined by $N_\varepsilon(p) = \{q \in D \mid dist(p,q) \le \varepsilon\}$

Given a value for minimal points Mitpts, a point q within the ε neighborhood of a point p is said to be directly density-reachable from q.

Definition 2: (directly density-reachable) [7]

A point p is *directly density-reachable* from a point q with respect to ε, Minpts if

1) $p \in N_\varepsilon(q)$ And

2) $|N_\varepsilon(q)| \ge MinPts$ (Core-point condition).

In this case, q is known as a *core point* because it is a dense point where there are enough other points surrounding it.

Armed with the notion of directly density reachable, we can define density-reachable by transitivity.

Definition 3: (density-reachable) [7]

A point p is *density-reachable* from a point q with respect to ε and MinPts if there is a chain of points, $p_1,\ldots,p_n, p_1 = q, p_n = p$ such that p_{i+1} is directly density-reachable from p_i.

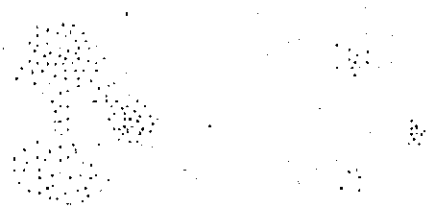


Figure 1. Original datapoint distribution (left) and core point abstraction of the same distribtuion.(right)

Definition 4: (density-connected) [7]

A point p is *density-connected* to a point q with respect to ε and MinPts if there is a point o such that both, p and q are density-reachable from o with respect to with respect to ε and MinPts.

Definition 5: (cluster) [7]

Let D be a database of points with a distance definition upon it. A *cluster* C with respect to ε and MinPts is a non-empty subset of D satisfying the following conditions:

1) $\forall p,q$: if $p \in C$ and q is density-reachable from p with respect to ε and MinPts, then $q \in C$. (Maximality)

2) $\forall p,q \in C$: p is density-connected to q with respect to ε and MinPts.(Connectivity)

Given *fixed* ε and MinPts values, the DBSCAN algorithm looks for a core point to start. It recursively expands a cluster using definition 5.

To support disk-based processing of very large scale database, once a point that is assigned to a cluster, it will no longer be reassigned again in the remaining computation. Therefore this algorithm incurs a very efficient N*log (N) time complexity, where N is number of points. This algorithm is listed below.

Algorithm DBSCAN(DB, ε, MinPts)

```
1   for each o ∈ DB do
2       if o is not yet assigned to a cluster then
3           if o is a core-object then
4               collect all objects density-reachable from o
5                   according to ε and MinPts;
                assign them to a new cluster;
```

DBSCAN in nature is connectivity based clustering algorithm. It focuses on local connectivity (density). Consequently, it is less sensitive to the global cluster formation. While DBSCAN is applied to web page clustering where each point corresponds to a web page, we observe that fixed values of ε and MinPts often leads to a single, giant cluster which is not useful at all. This is illustrated in Figure 1 below. To the left of that figure is the original data point distribution. After apply DBSCAN, only a single cluster emerges because each pair of points within this set of points are reachable according to above definitions. To remedy this problem, we propose a PageCluster algorithm called RDBC (recursive density based clustering algorithm) that attempts to solve this problem by varying ε and MinPts whenever necessary.

3. The RDBC Algorithm

RDBC is an improvement of DBSCAN for the web page clustering application. In RDBC, it calls DBSCAN with different distance thresholds ε and density threshold MinPts, and returns the result when the number of clusters is appropriate. The key difference between RDBC and DBSCAN is that in RDBC, the identification of core points are performed separately from that of clustering each individual data points. We call this an abstraction because these core points can be regarded as clustering centers that are representative of the data points. For this purpose, different values of ε and Mpts are used in RDBC to identify this core point set, Cset. Only after appropriate Cset is determined, the core points are clustered, and the remaining data points are then assigned to clusters according to their proximity to a particular cluster.

The algorithm can be summarized below:

```
uplherc.upl.com - - [01/Aug/1995:00:08:52 -0400] "GET /shuttle/resources/orbiters/endeavour-logo.gif HTTP/1.0" 200 5052
pm9.j51.com - - [01/Aug/1995:00:08:52 -0400] "GET /images/WORLD-logosmall.gif HTTP/1.0" 200 669
139.230.35.135 - - [01/Aug/1995:00:08:52 -0400] "GET /images/NASA-logosmall.gif HTTP/1.0" 200 786
uplherc.upl.com - - [01/Aug/1995:00:08:52 -0400] "GET /shuttle/resources/orbiters/endeavour-logo.html HTTP/1.0" 200 5052
pm9.j51.com - - [01/Aug/1995:00:08:52 -0400] "GET /images/WORLD-logosmall.html HTTP/1.0" 200 669
139.230.35.135 - - [01/Aug/1995:00:08:52 -0400] "GET /images/NASA-logosmall.html HTTP/1.0" 200 786
```

Figure 2. A sample server log from NASA

Algorithm RDBC ((ε, Mpts, WebPageSet)

```
Set initial values ε = ε1, and Mpts =Mpts1
      according to [7];
WebPageSet=Web_Log;

RDBC (ε, Mpts , WebPageSet)
{
   Use ε and Mpts to get the core points set
   CSet
   if size(CSet) > size ( WebPageSet ) / 2
   { // Stopping criterion is met.
      DBSCAN( WebPageSet, ε, Mpts);
   }
   else
    // Continue to abstract core points:
      ε= ε/2; Mpts =Mpts/4;
      RDBC(ε, Mpts , CSet);
      Collect all other points in (WebPageSet-
      CSet) around clusters found in last step
      according to ε²
   }
}
```

Intuitively, the algorithm goes into a cycle in which the core points themselves are taken as the points in a space, and clustering is done on those core points with more strict requirement on a core-point (with smaller radius around a core point.) This process stops when nearly half the points that remain are core points. Then, the algorithm will begin a gathering process to gather the rest of the points around the core points found into clusters. This is done with a larger radius value ε^2. Intuitively, this process can avoid connecting too many clusters via "bridges".

In our preliminary implementation, only one recursion is realized to achieve satisfactory results. However, we believe there are applications that may require more levels of recursion in order to identify appropriate clustering centers. In particular, we execute the following steps:

1) Use pre-defined values of ε and MinPts to compute core points and place them into Cset. This is illustrate in the figure to the right in Figure 1.
2) Perform DBSCAN on Cset to cluster core points only;
3) Assign remaining data points not in Cset to the clusters formed by core points.

The time complexity of DBSCAN is O(N * log N), where N is the number of distinct web pages. We keep the recursive time limited to a constant (such as two). Thus, we just run the DBSCAN algorithm just once. So the time complexity of our algorithm is O(N * log N).

Compared to traditional clustering algorithms such as K-means algorithm and the Scatter/Gather algorithm, our proposed RDBC algorithm has several potential advantages: (a) RBDC does not require number of clusters and clustering distance threshold (ε) to be pre-specified. Instead, these parameters are computed during the execution of the algorithm. (b) Because the algorithm uses density-based (connectivity) criterion, it may discover clusters of arbitrary shape. (c) In addition, it has a log-linear time in complexity and hence is very efficient in processing large-scale real world data.

4. Steps of Clustering Web Documents based on Web Logs

Depicted in Figure 2 is an example of NASA server log (url). In this section, we plan to illustrate the use of RDBC to cluster this web server log. We need to first build correlation information for distance measures. Our algorithm is described as the follow four steps:

Step 1. Pre-process access log into sessions.

1. We remove requests made to access image files (.gif, .jpg) in the log. Since most of them are accompanying figures to a specific web page, these image files are not requested explicitly by the users. We achieve this by executing the SQL statement as follows:

```
Insert into DiscardedData select * from Raw_Log
Where PageID like '%.jpg'
Delete from Raw_Log  where PageID like '%.jpg'
```

3. The upper commands can filter out image files that have 'jpg' as their extension .We repeatedly do this on all files that have extensions as 'jpg', 'jpeg', 'gif', 'bmp', and 'xbm'.
   1) Extract sessions from the data. A natural boundary for sessions is when users make unusually long pauses between browsing activities. These can be detected by observing the density of activities as a function of time.

We have found that it is often the case that for a given web logs, one can obtain a threshold value on the time interval between two adjacent page visits. If the time interval between the visits is greater than a time threshold T, then these visits are considered to belong to two different sessions. For example, we have observed that it is safe to set T at two hours for NASA data that we present later, and 24 hours for MSN data.

Step 2. Compute the co-occurrence frequencies between pages within a window size W (W is given as input), and create a distance matrix.

1) Determine the size of a moving window within which URL requests will be regarded as co-occurrence. Note that here we implicitly define a temporal locality between successive web page access. Since we are not using the content of each web page as feature vectors for clustering web pages, temporal proximity is used instead to indicate two web pages are relevant (close) in the data space. While this distance measure is not always satisfactory, it is the best information we can extract from the web server log alone. Any pair of URLs (P_i, P_j) outside the window are considered irrelevant and thus have a co-occurrence frequency of zero.
2) Calculate the co-occurrence times $N_{i,j}$ of each pair of URL's (P_i, P_j) based on the W. Also, calculate the request occurrence N_i , N_j of this pair of URL's.
3) P (P_i, P_j) = $N_{i,j} / N_j$.
4) We can select any of the following three distance functions for our applications; the first distance definition is the same as that by Perkowitz and Etzioni.

$$Dis1(A,B) = Max(1/P(A \mid B), 1/P(B \mid A))$$
$$Dis2(A,B) = 0.5(1/P(A \mid B) + 1/P(B \mid A))$$
$$Dis3(A,B) = \sqrt{(1/P(A \mid B) \cdot 1/P(B \mid A))}$$

In all, we spend W * L in distance calculation in worse case time. Because W is a constant, the time complexity for this step is O(L).

The first distance definition is the same as that by [9]. In our application, we found this definition to be too restrictive because in our application it often yields infinite distances between many URL's. This gives very skewed results where many web pages are considered a single cluster by themselves. So frequently we cannot get desired results for our clustering using this distance definition. The second definition is the arithmetic mean whereas the third is geometric mean. We find that the third definition gives the best result in all three domains where we test our algorithm.

Step 3. Run RDBC on the distance matrix

Step 4. Output the clusters generated above.

5. Experimental Validation

In this section, we present our experimental results that test the performance of our algorithm. We test the clustering and index-page construction algorithm on three data sets. We compare our algorithm's performance with that of DBSCAN.

We first analyze the data set under consideration. Our experiments draw on data collected from three web sites: Monash University of Australia, NASA and MSN. The first data set is used in Zukerman et al.'s work on predicting user's requests [2]. It consists of server log data collected during a 50-day period of time. It includes 525,378 total user requests of 6727 unique URL's (clicks) by 52,455 different IP's, consisting of 268,125 sessions. The NASA data set contains two months worth of all HTTP requests to the NASA Kennedy Space Center WWW server in Florida. The log was collected from 00:00:00 August 1, 1995 through 23:59:59 August 31, 1995. In this period there were 1,569,898 requests. Timestamps have 1-second resolution. There are a total of 18,688 unique IP's requesting pages, having a total of 171,529 sessions. A total of 15,429 unique pages are requested. The MSN.com log is obtained from the server log of **msn.com**, with all identity of users stripped away. It consists of data collected from Jan 27, 1999 to Mar 26, 1999, with a total of 417,783 user requests. This log contains 722 unique IP's requesting 14,048 unique pages. The MSN.com log is unique in that some requests are from groups of users submitted by Proxies or ISP's. Therefore the lengths of some sessions are long. For example, the long sessions range from 8,384 consecutive requests to 166,073 requests.

We compare the clustering quality between out algorithm RDBC and DBSCAN on these three data sets. We also measure the efficiency for index page construction using the different clustering results. The following tables and figures are our experimental results.

Cluser1	/shuttle/missions/41-c/news/ /shuttle/missions/61-b/ /shuttle/missions/sts-34/ /shuttle/missions/41-c/images/ ...
Cluster2	/history/apollo/sa-2/news/ /history/apollo/sa-2/images/ /history/apollo/sa-1/sounds/ /history/apollo/sa-9/sa-9-info.html ...
Cluster3	/software/winvn/userguide/3_3_2.htm /software/winvn/userguide/3_3_3.htm /software/winvn/userguide/3_8_1.htm /software/winvn/userguide/3_8_2.htm ...
...	...

Table 1: some clustering results using RDBC. If use DBSCAN all these pages are belong to the same cluster.

Table 2 and Figure 3 show the clustering result and efficiency comparison between the two clustering algorithms. We see that using RDBC, while having about the same time complexity as DBSCAN, we obtain more clusters for the data set that is more reasonable and will generate clusters with more even distribution than DBSCAN. The same result applies to both NASA and MSN data (see Tables 3—4 and Figures 4 – 5). By examining the contents of the logs, we have see that the clusters we construct are more reasonable since similar topics are indeed grouped together and different topics are separated.

	RDBC	DBSCAN
Number of Web Pages	6727	6727
Run Time (Sec)	20	22
ε/Mpts	10/20 5/5	10/20
Number of Clusters	125	6

Table 2. Comparing clusters obtained by RDBC and DBSCAN on Monash University Data

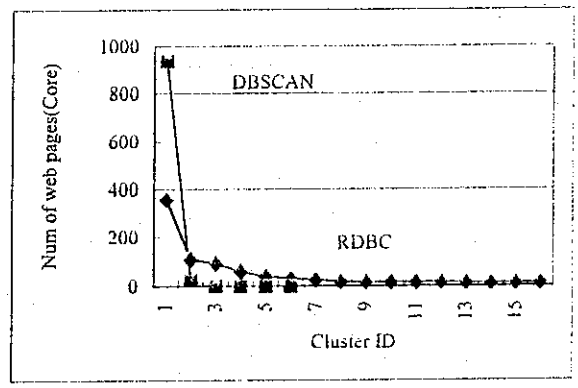


Figure 3: Compare RDBC and DBSCAN on Monash University's log.

	RDBC	DBSCAN
Number of Pages	15,429	15,429
Run Time (Sec)	21	25
ε/Mpts	10/20 5/5	10/20
Number of Clusters	44	4

Table 3. Comparing clusters obtained by RDBC and DBSCAN on NASA's Data

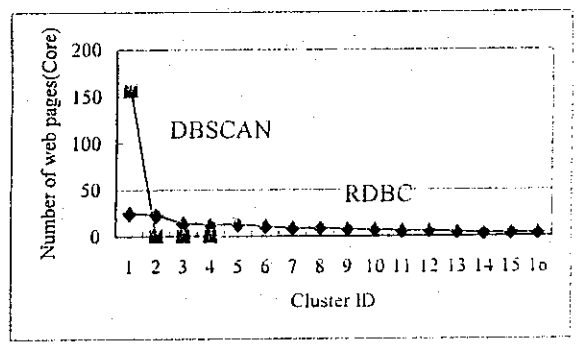


Figure 4: Compare RDBC and DBSCAN on NASA's log.

	RDBC	DBSCAN
Number of Web Pages	14,048	14,048
Run Time (Sec)	21	24
ε/Mpts	5/25 3/9	5/25
Number of Clusters	125	3

Table 4. Compare RDBC and DBSCAN on MSN's log.

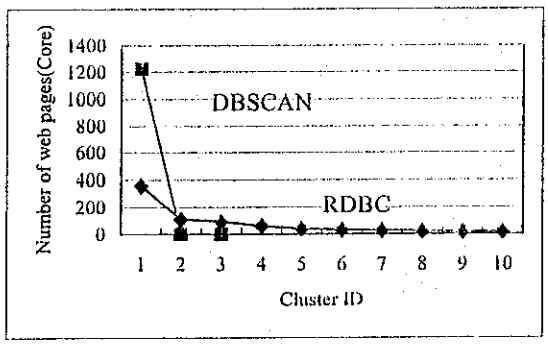


Figure 5: Compare RDBC and DBSCAN on MSN's log.

6. Conclusions and Future Work

In this paper we present an algorithm for clustering web documents based only on the log data. We do this by using a recursive density based clustering algorithm that can adaptively change its parameters intelligently. Our clustering algorithm calculates a density measure based on the distance metrics that is mined from the web logs according to our distance definition. It then selects the dense-enough points in the space of documents and constructs an abstract space based on these points. It does this recursively until no more abstraction space can be built. Because it can change the parameters intelligently during the recursively process RDBC can yield clustering results more superior than that of DBSCAN. It can be shown that RDBC goes as fast as that of the DBSCAN algorithm.

The work reported in this paper is part of our ongoing effort in utilizing the user information for re-organization of web pages.

7. References

[1] A. Bouguettaya 1996. On-Line Clustering. IEEE Transactions on Knowledge and Data Engineering, Vol.8, No.2, 1996, pp.333-339.

[2] D.W. Albrecht, I. Zukerman, and A. E. Nicholson, (1999). Pre-sending documents on the WWW: A comparative study. IJCAI99 – Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence.

[3] E.M. Voorhees. Implementing agglomerative hierarchical clustering algorithms for use in document retrieval. *Information Processing & Management*, 22:465-476, 1986.

[4] E. Rasmussen. Clustering algorithms. In W.B. Frakes and R. Baeza-Yates, editors, *Information Retrieval*, pages 419-442. Prentice Hall, Eaglewood Cliffs, N.J., 1992.

[5] Hearst M. A. and Pedersen J. O. 1996. Reexamining the Cluster Hypothesis : Scatter/Gather on Retrieval Results. In proceedings of the 19th Annual International ACM SIGIR Conference, Zurich, June 1996.

[6] L. Kaufman and P. J. Rousseeuw 1990.Finding Groups in Data: An Introduction to Cluster Analysis, John Wiley & Sons, 1990

[7] M. Ester, H.P Kriegal, J. Sander , and X. Xu 1996. A Denity-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. KDD 96 – Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining. 1996

[8] M. Ester, H.P Kriegal, J. Sander , M. Wimmer and X. Xu 1998. Incremental Clustering for Mining in a Data Warehousing Environment. VLDB 1998 – Proceedings of 24rd International Conference on Very Large Data Bases. 323-333 .1998

[9] M. Perkowitz and O. Etzioni. Adaptive web sites: an AI challenge. In *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence*, 1997.

[10] P. Willet. Recent trends in hierarchical document clustering: a critical review. *Information Processing and Management*, 24:577-97, 1988.

[11] R. Ng and J. Han 1994. Efficient and Effective Clustering Methods for Data Mining. Proc. Of 1994 Int'l Conf. On Very Large Data Bases(VLDB'94), Santiago, Chile, September 1994, pp 144-155.

[12] R. Sibson 1973. SLINK: an optimally efficient algorithm for the single-link cluster method. The Computer Journal, Vol.16, No. 1, 1973, pp. 20-34.

[13] X. Xu, M. Ester, H.P. Kriegel and J. Sander 1998. A Distribution-Based Clustering Algorithm for Mining in Large Spatial Databases. ICDE 1998 – Proceedings of the Fourteenth International Conference on Data Engineering. 324-331. 1998

[14] Zamir O. and Etzioni O. 1998. Web Document Clustering : A Feasibility Demonstration. Proceedings of the 21nd International Conference on Research and Development in Information Retrieval (SIGIR'98).

1. Can you explain why Alipes adapts to the dynamic nature of users' interests, which can change from slowly to suddenly, from on domain to another, over a very short to very long time? (15%)

2. In page 407, the authors compute the final document score with Score ($Proile_p$, fv_d) = max(w_{logn}, w_{pos}) + min(w_{logn}, w_{neg}). Can you explain why the authors do this? (10%)

3. In learning short-term interest weights in Section 3.3.3, equations 11 and 12 express update rules for learning positive feedback. Can you compare the major difference between these equations and equation 13 which learns long-term interest weights? (10%)

4. In a professional cyber community, knowledge in terms of documents is shared among community residents. An automated recommendation function based on *Alipes* can be used for introducing articles to individual users along with the addition of new articles. Can you foresee any limitations of this application? For those limitations you specified, please propose your solutions to remove them? (15%)

第二節

An Adaptive Algorithm for Learning Changes in User Interests

Dwi H. Widyantoro, Thomas R. Ioerger, John Yen
Department of Computer Science
Texas A&M University
College Station, TX 77844-3112
dhw7942,ioerger,yen@cs.tamu.edu

Abstract

In this paper, we describe a new scheme to learn dynamic users' interests in an automated information filtering and gathering system running on the Internet. Our scheme is aimed to handle multiple domains of long-term and short-term user's interests simultaneously, which is learned through positive and negative user's relevance feedback. We developed a 3-descriptor approach to represent the user's interest categories. Using a learning algorithm derived for this representation, our scheme adapts quickly to significant changes in user interest, and is also able to learn exceptions to interest categories.

Keywords
Information Filtering, Intelligent Agents.

1 Introduction

The spread of the World Wide Web and online news sources on the Internet recently has changed the way people locate information and their news reading habits. As more online news sources become available on the Internet, people have more options to read news articles that they think are interesting. However, selecting relevant articles from a group of news articles on various topics and online sources is still considered a time consuming process. Although search engines can help finding relevant news articles, it still requires the user to describe interests each time the user wishes to pull the news. Recent efforts have been devoted to overcome this problem by personalizing an information filtering system. This system takes into account the user profile information to present relevant information to its user effectively.

We have developed *Alipes*, a personalized news agent that gathers articles periodically from various online news sources and filters them on behalf of its users [16]. *Alipes* maintains the profiles of its users based on which a set of relevant news articles from the World Wide Web is recommended to its users. Moreover, *Alipes* adapts to the dynamics of the user's interests by learning from the user's feedback. This paper describes a new scheme for learning user interest that has been incorporated in *Alipes*. Our scheme is able to adapt to the dynamic nature of users' interests, which can change from slowly to suddenly, from one domain to another, over a very short to very long period.

Most previous information filtering systems on the Internet, for example *WebMate* [5] use a keyword vector to represent categories of user interest. Incremental learning algorithms with such a representation have trouble adapting in an appropriate time frame as interests slowly or quickly shift focus. Our approach uses a 3-descriptor scheme to represent a category of interest in a profile and its learning algorithm. In this scheme, an interest category consists of three descriptors: one long-term descriptor to maintain long-term interests, and other two descriptors, positive and negative, to keep up with short-term interests. This approach is similar to an incremental method for learning in domains with concept drift, where multiple concept representations that generalize examples over different window sizes are maintained simultaneously [14, 15]. Compared to systems that mainly use a single-descriptor model for interest category representation, the 3-descriptor scheme has several advantages. The 3-descriptor scheme allows learning of long-term and short-term interests simultaneously, and also handles exceptions of interests within an interest category. This capability cannot be achieved using the single-descriptor representation.

The rest of this paper will be organized as follows. Related work and its limitations will be described briefly

in the following section. The third section describes our approach for modeling and learning of user profiles. Then, a brief description of the evaluation methodology and results is presented in the fourth section, followed by a conclusion.

2 Previous Work

There are many systems that have recently been described for news and information filtering. *Webmate* keeps track of user interest through multiple TF-IDF vectors [5]. *Fab* is an adaptive system for Web page recommendation which represents user profiles as a single feature vector [2] and handles multi-topic interests [3]. *Syskil & Webert* is an intelligent agent that represents a profile as Boolean features and uses a Naive Bayesian classifier to determine whether a Web page is relevant or not [10]. Lang compared various alternatives to learn a static user profile in his *NewsWeeder*, a newsnet filtering system [7]. Neural networks have also been explored to learn user profiles for topic spotting [18] and for filtering news articles on the Internet [13]. In *NewT* [12] and *Amalthea* [9], genetic algorithm is employed to learn user interests. Incremental relevance feedback is a common method used to learn user profile for information filtering in these systems. Allan explored the effectiveness of this method and demonstrated that good results can be obtained using only a few judgments [1].

Although the performance of these systems improves after learning a user profile, most of them do not address the effectiveness of their approaches to adapting to changing interests and handling exceptions of interests within an interest category. Except in works by [4, 8, 9, 12], their evaluation assumes that the user's interests do not change during the evaluation process. In real life, however, both the user's long-term and short-term interests usually change over time. Long-term interests are interests that result from an accumulation of experiences over a long time-span. Meanwhile, short-term interests are interests in events on a day-to-day basis which change over a short period. Therefore, the capability to adapt to these changes effectively and to handle exceptions to categories is still an open problem, and these issues will be addressed in this paper.

3 Modeling and Learning User Profile

The capability to model and learn a user profile is at the heart of a personalized information filtering system. We will describe in this section our approach to designing a profile representation, how to use the representation for information filtering, and how to develop a learning algorithm that adapts to the dynamics of the user's interests.

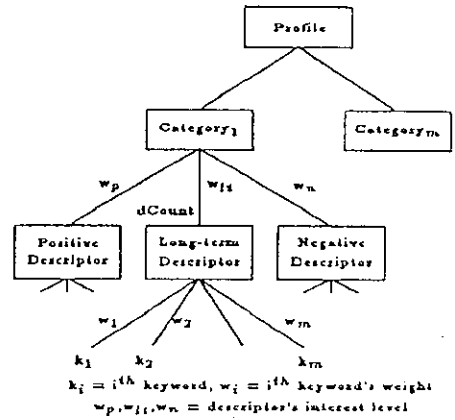


Figure 1: A 3-descriptor Representation

3.1 Profile Representation

The basic structure of an interest category representation is a feature vector. It contains a list of keywords and each keyword is weighted according to its degree of importance. There are several keyword weighting schemes that have been developed such as TF-IDF [11] and LSI [6]. In this work, we use the TF-IDF weighting scheme to assign the keywords' weights due to its appropriateness for use in an online learning algorithm. Based on TF-IDF, the keyword importance is proportional to the frequency of occurrence of each term in each document and inversely proportional to the total number of documents in a document collection in which the term occurs [11]. It assumes that keywords appearing in fewer documents discriminate better than the ones appearing in more documents. The weight of keyword i in a document d is then defined as follows:

$$w_i^d = tf_i^d . \log \frac{N}{df_i} \qquad (1)$$

where N is the number of documents in the document collection, df_i is the document frequency of term i, and tf_i^d is the frequency of term i in document d. The m highest weighted terms and bigrams (pairs of adjacent words) are then kept and normalized such that $|w_j^d| = 1$ for j = [1..m]. The terms are extracted from a document that has been pre-processed by: removing HTML tags and links, Java scripts and stop words [1], stemming words [2].

In a 3-descriptor representation, an interest category C is composed of three descriptors: a positive d_p^c, a negative d_n^c, and a long-term d_{lt}^c descriptor. Each of the descriptors consists of a list of pairs of keywords and their

[1] The stop list consists of 293 common words, like "a", "the", "although", etc.

[2] We use Porter's stemming algorithm to find the root of words and thus reduce the number of terms.

weights. Figure 1 illustrates the structure of a profile using this scheme. The *positive* and *negative* descriptors maintain a feature vector learned from documents with positive and negative feedback respectively, while the *long-term* descriptor maintains the feature vector of a document from both types of feedback. Each descriptor also has an interest weight to represent the interest level of the corresponding interest category's descriptor. Interest weights w_p^c, w_n^c and w_{lt}^c are used to describe the level of interest in positive, negative and long-term descriptors of interest category C, respectively. The range of w_p^c and w_n^c is [0,1], while the range of w_{lt}^c is (-1,1). Negative and positive values of w_{lt}^c describe the uninterestingness and interestingness in the domain of interest represented by the feature vector of the long-term descriptor. In addition to the long-term descriptor, a document counter *dCount* is maintained to keep the total number of documents that have been observed. Formally, the representation of interest category C can be written as follows.

$$Cat_c = \langle (w_p^c, d_p^c); (w_n^c, d_n^c); (dCount, w_{lt}^c, d_{lt}^c) \rangle \quad (2)$$

The user may have multiple interest categories, and the profile of a user P having n interest categories is represented as:

$$Profile_p = \{Cat_1^p, Cat_2^p, ..., Cat_n^p\} \quad (3)$$

Based on the above representation, the document-filtering process is performed and a learning algorithm to model user profiles is developed to accommodate intuitively defined behavior of changing user's interests.

3.2 Information Filtering

The information-filtering process is performed by selecting the n most relevant documents from a set of documents. For each document in the set, the interestingness of the document is assessed according to the match to an interest category of the profile and the degree of interest in that category. The assessment is calculated as a numeric value ranging from -1 to 1 that is assigned to a document as the score of the document with respect to the profile being considered. A positive value of the score indicates that the document is interesting to some degree. Conversely, a document with a negative score is uninteresting. Based on these scores, the documents are ranked, and the n most relevant documents are obtained from the n top ranked documents.

Given a document feature vector fv_d, the score of fv_d for a profile P is computed as follows.

1. Calculate the relevance of each category C in profile P with the document being examined. The category relevance is defined by equation 4 as the maximum similarity between a document feature vector and either the feature vector of positive, negative or long-term descriptor, where the cosine similarity is employed to measure the similarity between the two vectors.

$$\begin{aligned} \mathrm{Rel}(Cat_i, fv_d) = \max\{&\mathrm{Sim}(d_p^i, fv_d), \\ &\mathrm{Sim}(d_n^i, fv_d), \mathrm{Sim}(d_{lt}^i, fv_d)\} \end{aligned} \quad (4)$$

where

$$\mathrm{Sim}(d^i, fv_d) = \frac{d^i \cdot fv_d}{|d^i| \cdot |fv_d|} \quad (5)$$

2. Calculate the score of each descriptor in the category C with the greatest relevance:

$$\begin{aligned} w_{pos} &= w_p^c * \mathrm{Sim}(d_p^c, fv_d) \\ w_{neg} &= w_n^c * \mathrm{Sim}(d_n^c, fv_d) \\ w_{long} &= w_{lt}^c * \mathrm{Sim}(d_{lt}^c, fv_d) \end{aligned} \quad (6)$$

where

$$c = \arg\max_i \{\mathrm{Rel}(Cat_i, fv_d)\}. \quad (7)$$

3. Compute the final document score as follows.

$$\begin{aligned} \mathrm{Score}(Profile_p, fv_d) = &\max(w_{long}, w_{pos}) + \\ &\min(w_{long}, -w_{neg}) \end{aligned} \quad (8)$$

The final value of the document score is a fusion between the score of positive w_{pos} and negative w_{neg} interest. The score of long-term interest w_{long} contributes to either the positive or negative interest depending on the sign of its value.

3.3 Learning User Profiles

The learning algorithm in *Alipes* allows incremental and online learning, which enables reactive learning as well as long-run learning. For the clarity of presentation, the learning algorithm will be presented in a high-level description prior to explaining the details of the algorithm that follows.

3.3.1 *Learning Algorithm*

The learning process in a personalized information filtering system relies on a user's feedback. Using the feedback information, the profile is modified such that it will be incorporated in future information-filtering tasks. The feedback consists of feedback type *fbType*, document to be learned fv_d and learning rate α. The feedback type can be positive or negative to represent that the user likes or dislikes the document's content. The learning rate represents the strength of the user's preference (e.g. very interesting, interesting, not bad, uninteresting etc.) and its range is (0,1]. In general, the algorithm to modify a user profile P is defined as follows.

Input: *fbType*, fv_d and α
Output: *modified P*

1. Find the most relevant category C in profile P
2. **If** Rel(Cat_c, fv_d) $\geq \theta$ **then**
3. LearnUserFeedback (P, *fbType*, fv_d, α)
4. **Else**
5. CreateNewCategory (P, *fbType*, fv_d, α)
6. **End if**

Finding the most relevant category in the above algorithm is the same process of finding the greatest category relevance in document scoring described earlier. A *threshold* constant θ is defined to determine when the highest similarity to an existing category is low enough to justify creating a new interest category. This process is used to learn various categories of interests based on the category relevance measure and the threshold constant. How to set this value will be addressed later in the evaluation section. The learning process in step 3 includes updating the descriptor feature vectors and modifying the long-term and short-term descriptors' interest weights.

3.3.2 *Updating the Feature Vectors of Descriptors*

The modification of a descriptor's feature vector with the feature vector of a sample document should accommodate the learning of short-term and long-term interests. Short-term interests tend to be reactive so that feedback will be incorporated immediately in future information-filtering. On the contrary, long-term interests change gradually. The modification of a long-term interest area should be sufficiently small that it will preserve the feature vector of documents from past feedbacks while still considering the contribution of document feature vector from the most recent one. Taking all these into account, the updating of a descriptor feature vector in category C is as follows:

$$d^c_{(new)} = d^c_{(old)} * (1 - \alpha) + fv_d * \alpha \qquad (9)$$

where d^c is either d^c_p for positive feedback, d^c_n for negative feedback, or d^c_{lt} for both positive and negative feedback. The learning rate α is used to adjust the contribution of the learned document. For the short-term descriptors (e.g. d^c_p and d^c_n), the value of α is obtained directly from the user's preference when giving feedback. A high learning rate results in a significant contribution of the learned document to the positive or negative descriptor. Therefore, the modification of these descriptors will be in line with the user's preference, and will determine the reactive behavior of short-term interests. However, the learning rate for the long-term descriptor d^c_{lt} is determined inversely by *dCount*, the number of example documents that have been learned so far. The value of α in equation 9 is derived by equation 10 to modify the feature vector of the long-term descriptor.

$$\alpha = \frac{1}{dCount + 1} + 0.05 \qquad (10)$$

As more feedback is learned, the contribution of the most recently learned document becomes smaller and therefore the previously learned interests are still preserved. The constant 0.05 is used to prevent a complete stoppage of learning, since α would otherwise converge to 0 in the limit (after learning many documents). Thus, it allows the long-term descriptor to keep learning regardless of the number of previously learned examples.

3.3.3 *Learning Short-term Interest Weights*

As mentioned earlier, the descriptor feature vector represents the interest area, and the degree of interest in the area is denoted by the descriptor's interest weight. The learning of short-term interests is performed by modifying the positive and negative descriptors' interest weights, w^c_p and w^c_n. These weights are updated to reflect the user's short-term interests so that any feedback (positive or negative) will be incorporated immediately in future information filtering. Specifically, the update of these interest weights is performed by increasing the corresponding interest weight according to the level of confidence obtained from the relevance feedback, and by decreasing the interest weight of the opposite descriptor. The amount of reduction in interest weight of the opposite descriptor is proportional to the learning rate and the similarity between the feature vector of the learned document and the one of the opposite descriptor. Equations 11 and 12 express these update rules for learning positive feedback.

$$w^c_{p(new)} = w^c_{p(old)} + (1 - w^c_{p(old)}) * \alpha \qquad (11)$$
$$w^c_{n(new)} = w^c_{n(old)} * (1 - \alpha * \mathrm{Sim}(d^c_n, fv_d)) \qquad (12)$$

For learning negative feedback, the same formulas are used by changing w_p with w_n and d_n with d_p, and vice versa.

3.3.4 *Learning Long-term Interest Weights*

In a long-term descriptor, the modification of the descriptor's interest weight w^c_{lt} should capture a reluctance of the interest to change after learning in the long run. For this motivation, a bipolar sigmoid function is used to govern the change of the long term descriptor's interest weight so that the change of w^c_{lt} will be more gradual. The function ranges from -1 to 1 where the lower and upper limit can be approached using argument values $-\infty$ and $+\infty$ respectively. By defining the ordinate (y-axis) of the function to be the value of w^c_{lt}, the use of this function is expressed in equation 13.

$$w^c_{lt(new)} = f(f^{-1}(w^c_{lt(old)}) \pm \alpha) \qquad (13)$$

where $f(x)$ is a bipolar sigmoid function.

$$f(x) = \frac{2}{1 + \exp^{-x}} - 1 \qquad (14)$$

First, the current value $w^c_{lt(old)}$, the ordinate, is projected to its abscissa using the inverse of bipolar sigmoid function. Second, the learning rate α is then added to the abscissa value for positive feedback or subtracted from the abscissa value for negative feedback. Finally, the new abscissa value is projected back to its ordinate as the new value of $w^c_{lt(new)}$. The input α to update $w^c_{lt(new)}$ is obtained from the user's preference rather than the one derived in equation 10. So the same amount of effort to change the level of long-term interests is required as to build them.

3.3.5 *Creating New Interest Categories*

The learning of new interests from positive feedback is initialized as follows:

$$d^c_{lt} = fv_d \qquad w^c_{lt} = f(\alpha)$$
$$d^c_p = fv_d \qquad w^c_p = \alpha \qquad (15)$$
$$d^c_n = \{\} \qquad w^c_n = 0 \qquad (16)$$

where $f(\alpha)$ is the bipolar sigmoid function. The assignment of w^c_{lt} and w^c_p uses equations 13 and 11 respectively by setting their initial values to zero. For negative feedback, $w^c_{lt} = f(-\alpha)$ and the assignment of equations 15 and 16 are swapped one of another so that $d^c_n = fv_d, w^c_n = \alpha, d^c_p = \{\}$ and $w^c_p = 0$.

4 Evaluation

Experimental evaluation has been conducted to measure the performance of *Alipes* to learn user's interests from user feedback. The main objectives are to evaluate the adaptability of the 3-descriptor scheme to the changing interests of the user and the ability of the scheme to handle exceptions to categories.

4.1 Method

4.1.1 *Data*

Documents used in our experiment are news articles in HTML format collected from 12 different online newspapers and magazines (Yahoo and Excite's Sport News, UsaToday, USNews, Fortune, PCWeek, PCMagazine, BusinessWeek, Windows, People, Time and Internet World), at different times. The collection contains 1427 documents with six different general topics: world, financial, health, weather, technology and sport news. The length of each processed document varies with an average number of distinct terms of 228.

4.1.2 *Procedure*

The experimental procedure to evaluate the adaptability of the 3-descriptor model to the changing interest of the user is designed to simulate the application of the scheme in a news agent. A detailed description of this procedure is described in [17]. Starting with an empty profile, the system provides a recommendation of 10 articles to the user. The user (real or simulated) examines the articles and gives feedback on whether each article is interesting or uninteresting with a degree of confidence. The system then learns from the user's feedback and modifies its profile. At this point, one cycle of evaluation ends. The next cycle starts using the most recently modified profile. At each cycle, the system's performance is measured, and a different set of 200 documents is selected to be filtered. This simulates the changing of news articles in online daily newspapers or weekly magazines that may have overlapping topics. To observe how well the system adapts to the changes of interest, the user's interest is *inverted* at the twentieth cycle by swapping the interesting and uninteresting domains of interest. In these experiments, the user is simulated by a target profile containing a list of interesting and uninteresting domains of interest. Following the experiment by Moukas and Zacharia [9], the positive or negative feedback is given based on the similarity between the examined article and the target profile. We use *accuracy* as the measure of system's performance. Accuracy is defined as the percentage of the n highest ranked documents (in this experiment n=10) recommended by the system that agree with those selected by the target profile, assuming they use the same document set.

To measure the 3-descriptor scheme's performance to handle exceptions of interest, a different experimental procedure was used, which will be described later.

4.1.3 *Performance Comparison*

To compare the performance between our 3-descriptor model and a single-descriptor scheme, we used the algorithm for learning user profiles employed in the *WebMate* system [5]. The algorithm was chosen due to its similarity in profile representation and in interest-domain clustering [3]. To make the algorithm comparable with the one developed in our work, bigram identification was added, and reward for keywords appearing in a document's title and header, applied in the original algorithm, was eliminated in the modified algorithm. An important difference between *WebMate* and *Alipes* is that *WebMate* was not originally designed to learn from negative feedback. To make the comparison fair, we implemented a version of the system that could do

[3] Experiments in comparing Alipes with the Rocchio algorithm are currently in progress.

this (referred to *Webmate-neg* below). The learning of negative feedback is performed by subtracting the feature vector of a learned document from the matching category in the profile.

4.2 Results

We conducted initial experiments to determine the optimal threshold value θ for creating new categories and the optimal number of keywords for representing feature vectors. By varying the threshold values from 0.05 to 0.65, we found that the optimal setting, where it gives the highest average accuracy, was 0.25. Having lower or higher threshold values, which leads to fewer or more categories, degrades the system performance. Similarly, by varying the number of keywords obtained from 20 to 220 highest weighted keywords, the optimal value for this parameter was found to be 90. Fewer keywords tend to remove important keywords while too many keywords will add more noise. The experimental results described in this sub section use these parameter values.

4.2.1 *Accuracy*

Figure 2 shows that the performance of the system employing a 3-descriptor model(*Alipes*) in general outperforms the one using a single-descriptor model(*Webmate* or *Webmate-neg*). The performance in both the 3-descriptor and single-descriptor models increases rapidly after the first iteration. However, the system's performance is erratic afterwards because a different set of documents is used in each round. In the subsequent iterations, therefore, the new document set may not provide documents representing a previously learned interest category, and may introduce other documents that match the target profile but have not yet been learned by the system. After the target profile is inverted at the twentieth iteration, the system's performance drops to its lowest value. It takes a short time for our 3-descriptor approach to adapt to this sudden change before the system regains its performance. The recovery process is worse in the single-descriptor case. It takes much more time for the single-descriptor scheme to re-stabilize after inversion.

We observe that *Alipes* takes only slightly longer to reach its highest accuracy after profile-inversion (about 5 iterations) than when starting from the scratch. The fact that learning new interests from an empty profile is faster than learning the inverted interest is an effect of long-term learning. To increase the interest level of a long-term descriptor from a negative value after the profile is inverted requires more effort (e.g. more feedback from the user)than starting from an empty profile. *WebMate*, however, recovers much more slowly,

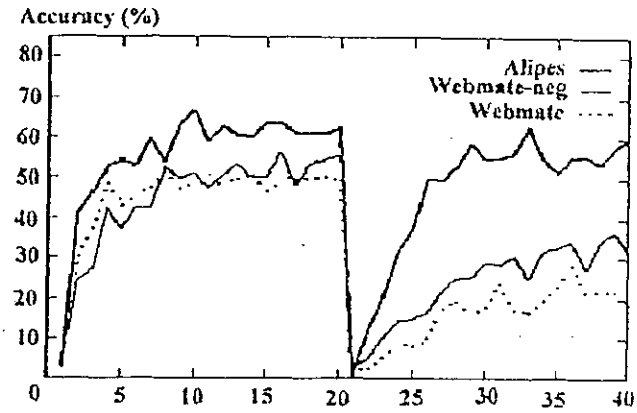


Figure 2: The System's Accuracy (percent of top matches that are relevant). At iteration 20, the target profile was inverted, simulating a dramatic change in interests. *Webmate-neg* is a version of WebMate that learns from both positive and negative feedback, *Webmate* learns only from positive feedback.

and never completely restores its performance from before profile-inversion by the end of the experiment. This is because *WebMate*'s use of single-descriptors to represent category interests is easily confused by a sudden change in interests. Instead of mixing all the feedback, *Alipes* effectively maintains long-term interests, while using its short-term descriptors to provide more reactive behavior.

The role of long-term descriptor is to maintain the past learned interests, and this descriptor basically performs the same function as the single-descriptor scheme in *WebMate*. By making a slight change in this descriptor, most feature vectors of the past learned documents are retained while still incorporating the new learned document. This enables the system to provide recommendations according to the current and the previous learned interests. Therefore, it is not surprising that by adding short-term descriptors as in the 3-descriptor model, its accuracy of prediction is better than the single-descriptor model during the first twenty evaluations.

The role of short-term descriptors becomes apparent in the presence of changing interests of the user. The feature vector update rule in these descriptors allows the system to be responsive to the most recently learned interest, with respect to the confidence level of the user's relevance feedback. On a strong positive feedback, which is given by the user as he/she reacts due to the low performance of the system on the change of user's interests, the learning process in the positive descriptor enables the system to adapt quickly to the user's new interest. The system's responsiveness is strengthened by the negative descriptor when learn-

ing a strong negative feedback, which allows the system to quickly exclude uninteresting documents. Additionally, the interaction between the positive and the negative descriptors (e.g. increasing the level of positive interest will reduce the level of negative interest according to the similarity between both interests, and vice versa) makes the adaptation even faster. As a result, the effectiveness of the 3-descriptor scheme over the single-descriptor model to adapt to the drastic change is evident as shown in Figure 2.

Thus, the short-term descriptors explain why the performance improves over a single-descriptor model. The most recently learned interest is significantly taken into account during the information-filtering process. In the single-descriptor model, however, this is not the case. As more documents are learned by the single-descriptor system, the contribution of the new learned interest becomes insignificant. Because the representation of the 3-descriptor scheme is more expressive than the single-descriptor model to capture the user's interests, its accuracy of prediction to recommend interesting documents with respect to the target profile is also better.

We have also conducted experiments that change the target profile more gradually. From these experiments we found that in a setting where the information to be filtered changes slowly over time (e.g. the content of news articles in newspapers or magazines), the difference of performance that is due to the target profile change from the diversity of information sources is less apparent.

4.2.2 *Learning Exceptions to Categories*

In this experiment, our objective was to evaluate the other potential advantage of our 3-descriptor scheme over single-descriptor models: that it can learn interest categories with exceptions. Specifically, the negative descriptor of a category in a user profile allows the system to distinguish (with a unique set of keywords) documents that are related to the overall category but given negative feedback by the user. In this experiment, we attempted to train both *Alipes* (using the 3-descriptor model) and *Webmate* (using a single-descriptor scheme) on a set of Sports documents taken from an online news source, excluding articles about Golf, and then test each system to determine how the use of a negative descriptor affects the ability to rank documents correctly according to this specialized interest area.

The sources of the documents were from the sites described above, which provide documents in a predetermined hierarchy of categories. Three groups of documents were selected: 397 Non-Sports articles, 20 articles about Golf, and 118 articles about Other-Sports. In a given run, 20 Other-Sports articles and 10 Golf articles were chosen at random as a training set. *Alipes* was trained by giving the 20 Other-Sports articles with positive feedback and the 10 Golf articles with negative feedback, while *Webmate* was only trained on the 20 Other-Sports articles (since the original system could only use positive feedback). We also tested a version of *WebMate* (referred to as *Webmate-neg*) that was modified to accept negative feedback, and we gave it both the positive feedback (Other-Sports articles) and negative feedback (Golf articles). Then a separate test set, consisting of 10 randomly-selected other-Sports articles and 5 random Golf articles not used during training, along with the 397 Non-Sports articles, was ranked in terms of user interest by all three systems.

Sports	Average Ranking		
	Alipes	*Webmate*	*Webmate-neg*
Other-Sports	6.1%	5.8%	5.4%
Golf	95.3%	14.3%	10.4%

Table 1: Learning Exceptions to Categories. All three systems were trained on articles about all Sports except Golf. *Alipes* and *Webmate-neg* were also given explicit negative feedback about Golf articles. Average rankings of test articles in these categories relative to a large set of Non-Sports articles are shown. Top of ranking = 0%; bottom of ranking = 100%.

This training and testing procedure was repeated 10 times. In each run, we calculated the average ranking of the Other-Sports test documents (the target category) and the Golf test documents (the exception category), and divided these rankings by the total number of documents ranked (412) to get a percentile score (0%=highest interest, top of list; 100%=lowest interest, bottom of list).

Table 1 shows the results of this experiment. In the case of *Alipes*, the Golf documents were consistently ranked at the bottom of the list (95.3%, i.e. within top 393 out of 412 document, on average), and the Other-Sports documents were highly recognized articles in the target category (average ranking of Other-Sports was 6.1%). In contrast, both Golf and Other-Sports documents in *Webmate* were ranked high (14.3% for Golf and 5.8% for Other-Sports). The explanation for this behavior is that the single-descriptor model, when given a wide range of documents about Other-Sports, generalizes this by identifying keywords that are associated with Sports in general. Hence this category covers Golf documents, which unintentionally get ranked high by *Webmate*. In *Webmate-neg*, Golf documents are still ranked high, eventhough they were given explicit negative feedback. The negative feedback causes the keywords that are unique to the exception category to be dropped from the feature vector, but others

are retained, which reflects the inadequacy of a single-descriptor representation. Hence both single-descriptor schemes fail to discriminate between the two categories. In contrast, the negative descriptor in *Alipes* can recognize the Golf documents as a negative interest and ranks them very low. This helps avoid the over-generalization of Other-Sports made by the positive descriptor. So the 3-descriptor scheme with negative feedback enables *Alipes* to learn user interest categories with exceptions more accurately.

5 Conclusion

Changing interests are an undeniable fact in real life. The time scale may vary from hours to years long and the degree from slight to extreme change. This paper has described a 3-descriptor scheme and learning algorithm, in an intelligent news filtering system called *Alipes*, to tackle this very important issue. By treating separately the long-term and short-term interests, and handling carefully the interaction between positive and negative interests in short-term interest, the scheme is able to adapt quickly to large changes of interest, and handle exceptions of interests within the broader scope of an interest category. Our experimental evaluation demonstrated the effectiveness of this scheme, which outperforms that of a single-descriptor model.

6 Acknowledgement

We would like to thank Dr. James Wall, LTC Robert J. Hammell, Jianwen Yin, Magy Seif El-Nasr, Linyu Yang, Anna Zacchi and Chris Standish for fruitful discussion. This research was in part supported by ARL project number DAAL01-97-M-0235.

References

[1] Allan, J. 1996. Incremental Relevance Feedback for Information Filtering. In *Proc of the 19th Int'l ACM-SIGIR Conf on Research and Development in Information Retrieval*, 270-278.

[2] Balabanović, M. 1997. An Adaptive Web Page Recommendation Service. In *Proc of the 1st Int'l Conf on Autonomous Agents*, 378-385.

[3] Balabanović, M. 1998. Learning to Surf: MultiAgent Systems for Adaptive Web Page Recommendation. *Ph.D. dissertation*, Dept. of Computer Science, Stanford University.

[4] Billsus, D. and Pazzani, M. 1999. A Personal News Agent that Talks, Learns and Explains. In *Proc of the 3rd Int'l Conf on Autonomous Agents*, Seattle, WA.

[5] Chen, L., and Sycara, K. 1998. WebMate: Personal Agent for Browsing and Searching. In *Proc of the 2nd Int'l Conf on Autonomous Agents*, 132-139.

[6] Deerwester, S. et.al 1990. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41:391-407.

[7] Lang, K. 1995. NewsWeeder: Learning to Filter Netnews. In *Proceedings of Machine Learning Conference*, 331-339.

[8] Lam, W., Mukhopadhyay, S., Mostafa J. and Palakal, M. 1996. Detection of Shifts in User Interests for Personalized Information Filtering. In *Proc of the 19th Int'l ACM-SIGIR Conf on Research and Development in Information Retrieval*, 317-325.

[9] Moukas, A. and Zacharia G. 1997. Evolving a Multi-agent Information Filtering Solution in Amalthea. In *Proc of the 1st Int'l Conf on Autonomous Agents*, 394-403.

[10] Pazzani, M. and Billsus, D. 1997. Learning and Revising User Profiles: The Identification of Interesting Web Sites. *Machine Learning*, 27:313-331.

[11] Salton, G., and McGill, M. J. 1983. *Intr to Modern Information Retrieval*. New York: McGraw-Hill.

[12] Sheth, B. D. 1993. A learning Approach to Personalized Information Filtering. *Master thesis*, Dept. of Electrical Eng. and Computer Science, MIT.

[13] Tan, A., and Teo, C. 1998. Learing User Profile for Personalized Information Dissemination. In *Proc of Int'l Joint Conf on Neural Network 1998*, 183-188.

[14] Widmer, G., Kubat, M. 1996. Learning in the Presence of Concept Drift and Hidden Contexts. *Machine Learning*, 23(1):69-101, Kluwer Acad. Publ.

[15] Widmer, G. 1997 Tracking Context Changes through Meta-Learning. *Machine Learning*, 27(3):259-286, Kluwer Academic Publisher.

[16] Widyantoro, D.H, Yin, J., Seif El-Nasr, M., Yang, L., Zacchi, A. and Yen, J. 1999. Alipes: A Swift Messenger in Cyberspace. In *AAAI'99 Spring Symp on Intelligent Agent in Cyberspace*, 62-67.

[17] Widyantoro, D.H. 1999 Learning User Profile in Personalized News Agent. *Master Thesis*, Dept. of Computer Science, Texas A&M University.

[18] Wiener, E., Pederson, J. and Weigend, A. 1995. A NN Approach to Topic Spotting. In *4th Symp on Doc Analysis and Inf Retrievel*, Las Vegas, NV.

國立中山大學九十學年度博士班招生考試試題

科目：論文評述（資訊管理）【資管所】(二)　　共 7 頁 第 1 頁

請仔細閱讀論文 "Diffusion of Innovations in the Cyberorganization"，回答下列各問題。注意，各問題可能有回答字數限制，請仔細構思你的回答並在字數限制內清楚完整表達。（超過限制字數將扣分）

1、請列出最能代表本文的十個名詞鑰字（keywords）5 分

2、本文共分七段，請簡單摘述各段之主旨（請限制於二百字以內）　10 分

3、就本論文的研究主題和寫作架構而言，　15 分

甲、研究目的是什麼？

乙、研究結果有哪些？是否呼應其研究目的？在第七節（364 頁）的第一句話說：

"The results of this study provide empirical evidence that supports the need for a more context based model of innovation adoption."您同意這個結論嗎？請解釋。

丙、您認為可以如何改進本論文的研究主題，使本研究更有價值或是更具前瞻性？

4、本文最後的兩句話說：

"While many believe that technology will dictate future economic strategy, others argue that economic realities will ultimately influence technological initiatives."

這段文字敘述了兩種觀點，您同意哪一種？請就這段文字，發表你的論述，您必須選擇一個觀點發表您的論述。請注意，你的論述可以根據本論文，也可以和此論文無關。下筆之前，請花一點時間構思並整理你的想法，寫作完成並檢視您的文字是否修要修改潤飾。將依據您如何組織和編排文章段落，如何敘述您的觀點、舉證、或分析，並得到結論而作評分（請以 1000 字為限）。20 分

Diffusion of Innovations in the Cyberorganization

Mahesh S. Raisinghani, Ph.D

Assistant Professor, University of Dallas
Graduate School of Management

And

Lawrence L. Schkade, Ph.D., CCP, FAAAS, FDSI

Ashbel Smith Professor
University of Texas at Arlington
Department of Information Systems and Management Sciences

Introduction

The global information infrastructure serves as the foundation for new modes of personal interaction and business transactions in a collection of activities known as Electronic Commerce (EC). EC defined as a variety of market transactions that are enabled by information technology (Applegate et al., 1996), represents the entire collection of actions that support commercial activities on a network. EC represents one of the most promising directions for generating competitive advantage at the micro level of the organization and for increasing productivity at the macro level of the economy. In addition, the wealth of online information and the potential to facilitate diverse business transactions highlights the need for developing original research in this area. **The main research question for this study is: How well does innovation diffusion theory explain the adoption of Internet/Intranet/Extranet (NET) technologies for electronic commerce applications?**

1. Research model and hypotheses

Diffusion is the process by which an innovation is communicated through certain channels over time among the members of a social system (Rogers, 1983). Innovation has been described as an idea, a product, a technology, or a program that is new to the adopting unit (Zaltman et al., 1973; Rogers, 1983; Cooper and Zmud, 1990; Grover and Goslar, 1993). The innovation-decision process in the minds of the adopters can be partitioned into the following five steps: **1. Knowledge 2. Persuasion 3. Decision 4. Implementation 5. Confirmation.**

Studies have shown that the adoption of innovation follows an S-curve as depicted in figure 1.

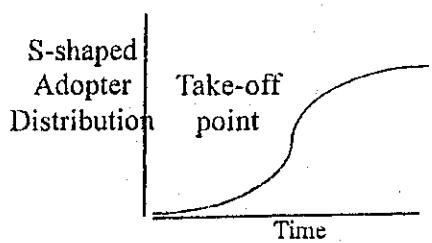


Figure 1. Individual Adoption Process

The fundamental question is whether Rogers' framework is applicable to the diffusion of EC technology. The secondary question explored in this study is how can the framework above be effectively applied to EC technology diffusion? The strategy employed in this study is to identify the critical elements in Rogers' framework that enable the members of the top management team initiate/adopt/implement an EC

0-7803-5082-0/98/$10.00 ©1998 IEEE.

application. The adoption process is not studied in this research since there is a temporal variable involved in studying its five steps. This necessitates a longitudinal research design which is outside the scope of the present research due to time and resource constraints.

In their meta-analysis on twenty-seven adoption and diffusion articles to identify common variables that motivate implementation, Tornatzky and Klein (1982) concluded that relative advantage, compatibility and complexity are the only three innovation characteristics possess consistent associations with innovation behaviors.

Relative advantage is a measure of the degree to which an innovation is perceived as better than the idea it supersedes. Positive perception of the benefits of EC should provide significant impetus for use of the technology, thereby leading to greater implementation success, the following three hypotheses are proposed:

H1a: Relative advantages will be positively related to the extent of initiation.

H1b: Relative advantages will be positively related to the extent of adoption.

H1c: Relative advantages will be positively related to the extent of implementation.

Compatibility is a measure of the degree to which an innovation is perceived as being consistent with the existing values, past experiences and needs of the potential adopters. In the context of EC applications, it is proposed that:

H2a: Compatibility will be positively associated with the extent of initiation.

H2b: Compatibility will be positively associated with the extent of adoption.

H2c: Compatibility will be positively associated with the extent of implementation.

Complexity is a measure of the degree to which an innovation is perceived as difficult to learn and use. It is proposed that:

H3a: Complexity will be negatively related to the extent of initiation.

H3b: Complexity will be negatively related to the extent of adoption.

H3c: Complexity will be negatively related to the extent of implementation.

One additional construct was identified beyond Rogers' classification that was thought important in the decision to adopt an EC innovation. This was *Image*, defined as the degree to which use of an innovation is perceived to enhance one's image or social status. It is proposed that:

H4a: Image will be positively related to the extent of initiation.

H4b: Image will be positively related to the extent of adoption.

H4c: Image will be positively related to the extent of implementation.

Outside of Roger's framework, the innovation literature has consistently recognized that environmental contingencies such as environmental uncertainty and heterogeneity facilitate innovation (Schroeder and Benbasat, 1975; Pierce and Delbecq, 1977; Dimaggio and Powell, 1983). In the context of EC, it is proposed that:

H5a: Environmental uncertainty will be positively related to the extent of initiation.

H5b: Environmental uncertainty will be positively related to the extent of adoption.

H5c: Environmental uncertainty will be positively related to the extent of implementation.

Among the problems with innovation research pointed out by Tornatzky and Klein (1982) are the need to focus on both adoption and implementation as dependent variables (possibly including a scale that measures degree of implementation) and the need to use replicable and reliable measures.

2. Research methodology and data collection

A live, web-based survey for those respondents with Internet access as well as an off-line, hard copy questionnaire for those respondents who either did not provide an email address or preferred to fill it out on paper utilizing a multi-item survey instrument for data collection.

Twelve hundred and fifty firms from a wide variety of industries that were involved in EC were randomly selected from the Internet Commerce Directory (1997) and the Dallas/Fort Worth EDI Forum, a Texas non-profit organization for EC's Member Directory (1997). Two follow-up e-mails and phone calls after two weeks and four weeks respectively were necessary to obtain completed questionnaires.

Instrument Validation. Content validation was achieved by using multi-item measures of a variable to specify the construct domain in order to average out the uniqueness of individual items, make finer distinctions between people, and provide greater reliability.

Criterion-related validity or predictive validity was not considered in this research due to the difficulty in establishing strong theoretically based criterion variables in the field of EC which is still in its infancy.

Construct validation (i.e., the ability of homogeneous items to converge together on a factor and away from other factors (Nunnally, 1978)-), and its two components, convergent and discriminant validity, were assessed by using factor analysis (one of the most powerful methods to test construct validity (Kerlinger,1986)).

Internal consistency and measurement reliability of the items was verified by computing the Cronbach's alpha (Nunnally 1978). Items with low inter-item correlations were dropped from the study, since it indicates that the items were not drawn from the same domain.

Operationalization of Constructs. A fairly comprehensive set of items from pre-validated instruments were used to measure the three dependent and five independent variables. Each of the independent variables were measured on a seven point Likert scale.

The validity analysis checks whether the instrument measures the attribute of interest. Reliability analysis checks whether the instrument produces identical results in repeated applications.

Validity of the Scales. Validity of a measure is the extent to which it measures what it is supposed to measure. Since the very definition of a construct implies a domain of content (Pedhazur and Schmelkin, 1991), this study assessed the content validity (i.e., theoretical support/face validity) and the construct validity (i.e., convergent and discriminant validity using factor analysis) of the scales.

Content validity. A logical analysis of the electronic commerce domain revealed that its closest link was to the telecommunications technologies. First, a critical evaluation of the definition of each construct was conducted by reviewing theories and research findings relevant to the construct under consideration. Second, the item content for each construct was adapted either from existing scales reported in the literature or from a panel of experts in the electronic commerce domain.

Construct Validity. Construct validity was determined using factor analysis of the multiple items comprising each construct.

Reliability of Constructs. Reliability refers to the extent to which the constructs are free from error and therefore, yield consistent results. Cronbach's alpha was used to measure the internal consistency of the multi-item scales used in this study.

3. Research results

Response Analysis. The final sample consisted of 154 firms, representing a 12.32 percent overall response rate. Of the firms that did respond, 36.4 percent responded by

completing the questionnaire on the hard copy (coded as dummy variable 0), and 63.6 percent responded by completing the survey on the website (coded as dummy variable 1). Six percent of the respondents were CEO's, 27 percent were one level from the CEO, 37 percent were two levels from the CEO, and 30 percent were three levels from the CEO. Since this survey was targeted at the members of the TMT, there were no respondents that were four or more levels below the CEO.

60 percent of the respondents were IS executives and 40 percent were non-IS business executives. Titles such as CIO, Executive VP, senior VP, IS Director, Web Administrator, and senior manager were found among those respondents that reported that they were one level from the CEO. There were 103 male respondents and 46 female respondents to this survey, showing that the senior IS levels are still possibly dominated by males.

With respect to the EC business model adopted by their firms, approximately 20 percent of the respondents reported a business-to-consumer/retail model, 72 percent reported a business-to-business model, and 8 percent reported an enterprise network model.

It is interesting to note that exactly half the respondents reported that their EC groups were independently funded and empowered to make decisions. Ideally, EC initiatives need to have more autonomy than the rest of the organization, since traditional thinking and traditional measures are not applicable to the novel characteristics and rules of EC.

The main intent of the majority of the organization's EC endeavor reported was either transaction processing with electronic payment (38.3 percent) or information distribution (20.1 percent). More than a third of the respondents felt that the results that their company had been able to measure were steadily increasing traffic on their company's website, higher percentage of business partners online, more satisfied customers, streamlined business processes, and better collaboration with business partners. Customer service is among the top objectives of companies that are implementing EC to capture the demographic profile of their customers and share and analyze account information.

50 percent of the respondents reported that they had hard proof that EC "works"; whereas 38 percent believed that EC could make them more competitive, but were waiting for hard proof, and 12 percent reported that they were still in development and unsure if there was a payoff.

Many companies (approximately 48 percent) reported that MIS and other departments are being trained on Internet concepts and technologies. Approximately 12 percent had outsourced EC work and 15 percent had hired a number of new EC specialists. The EC readiness from a human resources perspective of the remaining 25 percent reflects that there are several new technical and non-technical jobs that are created for content management and creative development of EC.

4. Discussion

As Bryan Plug, President and CEO of Pandesic, the Intel-SAP joint venture, told the audience at Networld+Interop 1997 trade show, "Doing business on the Web is not as hard as doing business in a traditional way. It's harder." (www.pandesic.com). IDC/Link (www.idc.com) predicts that the number of people with access to the WWW will rise from 55 million in 1997 to 550 million by the year 2000. The business -to-business Internet market in the U.S. has been predicted to reach three billion U.S. dollars in 1999, according to the EDI Group (www.edigroup.cpm/ research/718-intr.html).

Moreover, EC users seek information through personal networks of colleagues and custom tailor the support and training for the innovation to their own specific

needs, as they begin to develop the knowledge, expertise, and skills to use it effectively. As a result, EC innovations spread in a decentralized and horizontal communication network among peers.

5. Limitatations

Even though the Internet and the WWW are global applications, this study is limited to U.S. based organizations. This limitation leads to a more homogeneous sample, with respect to basic cultural differences in the strategic evaluation of EC technology. Internal validity of this study may be affected because there is no control over independent variables and the interaction between the subjects/Top Management Team (TMT) members. In addition, the strength and range of variables studied are limited due to the need for reasonably fast and easily understood communication in the interview and the time constraints faced by the members of the TMT.

An additional threat to internal validity concerns the lack of cross references between written business and information technology plans. Due to the lack of written long term plans for EC applications and the impracticality of obtaining written short term plans. Besides, asking people to comment on their written plans in a survey instrument might overly challenge their ability to answer correctly, since many people interviewed may not have looked at the written plans in the last few months. Finally, the lack of respondent anonymity may cause the responses to be biased by the methodological artifact that respondents are known to the researcher.

Since most EC applications are less than a year old, reliable measures of return on investment, internal rate of return, net present value, and other financial measures are not available. Development costs of EC applications are difficult to quantify, since these application costs are often hidden in other budgets.

Besides the research strategy, the theoretical approach utilized also affects a study's validity. The two major approaches to the study of implementation (Benbasat 1984, Rogers 1983) are the factors approach and the process approach. The factors approach attempts to identify static forces which lead to successful IT implementation. The process approach is longitudinal in nature, since it examines the behavior of stakeholders over time and focuses on the dynamics of the implementation.

Finally, within the vast and diffuse IS and strategic management literature, this theoretically grounded exploratory study may be regarded as an early attempt to reconcile the complex array of existing models, cases, and opinions in diffusion of innovations literature that apply to EC technologies.

6. Implications for practice

The study results demonstrate that EC is becoming critical in three interrelated dimensions evaluated in the questionnaire, (i.e., business-to-business interactions, business-to-consumer interactions, and the internal organizational functioning/Intranet dimension. Improved knowledge of the evolution of EC should be helpful to practicing managers for understanding the circumstances under which EC is appropriate for their organizations.

However, in order to take a strategic approach to using the NET, the entire organization needs to be re-engineered with the NET as a major business objective. This re-engineering includes developing a new business plan, a technology strategy and a total business process redesign.

The website should be designed not only to draw traffic but to ensure repeat traffic by keeping information current. Users need to be allowed to customize their shopping experience by choosing screen colors and resolution, selecting product notification options, and storing electronic signatures for gift cards.

The motivation for using EC differs

between the consumers and the organization. While the business-to-consumer model focuses on the desire to obtain current/ "dynamic" information, to buy and sell products and/or services, and to communicate with others; the business-to-business model focuses on the need to "do more with less," to develop new electronic applications that streamline current business practices such as the supply-chain management applications, collaborative applications/workgroup computing, and information sharing /electronic publishing while keeping an eye on standards and bandwidth availability (Kalakota and Whinston, 1997).

7. Conclusions

The results of this study provide empirical evidence that supports the need for a more context based model of innovation adoption. A parallel could be drawn between the findings of this research and the research on contribution of IS to productivity (Brynjolfsson, 1993; Brynjolfsson and Hitt, 1996). Earlier studies on the business value of computers reported a "productivity paradox" of IS, where despite enormous improvements in the underlying technology, the benefits of IS spending were not found in aggregate output statistics (Strassman, 1990; Barua et al., 1991; Brynjolfsson, 1993). However, a recent study by Brynjolfsson and Hitt (1996), found that IS spending has made a substantial and statistically significant contribution to firm output. The difference in the findings is attributed to "mismeasurement", "lags", "redistribution", and "mismanagement" by Brynjolfsson's (1993) review that concludes that the "shortfall of evidence is not necessarily evidence of a shortfall."

Similarly, in this research on the strategic evaluation of EC, perhaps a longitudinal study with a larger sample size and additional independent variables such as size of a firm, number of years of experience a firm has with EC technologies, and so forth may yield different results. At the present time it is not known whether the lack of support for the proposed hypotheses was due to the relative newness of the EC field or the noise in the current research model.

An enhanced understanding of the relationships between technology infrastructure components and technology applications can have a major impact on future IS technology management. The impact of global networks on economic activity and market structures can play a major role in the evolution of ubiquitous communication. While many believe that technology will dictate future economic strategy, others argue that economic realities will ultimately influence technological initiatives.

References:

References are available from Dr. Mahesh S. Raisinghani upon request.

國立中山大學九十學年度博士班招生考試試題

科目：論文評述（資訊管理）【資管所】(一)　　共 1 頁 第 1 頁

論文評述：資訊管理

請閱讀所附的論文「網路行銷決策研究－以線上購物行為分析為基礎」，並回答下面的問題：

一、 請詳閱該論文後，以英文撰寫一份八百字以內的論文的摘要。該摘要應包括研究的問題、研究方法、研究發現、結論與建議等重要內容。（２０％）

二、 資訊管理的研究以研究問題及研究方法而言，可以分哪些類別？並請說明該論文在您所界定的資訊管理研究中的定位與價值。（１０％）

三、 請評論該研究論文的優缺點，並針對您指出的缺點說明該研究應該如何修改才更為優秀？（２０％）

2/15

網路行銷決策研究 -以線上購物行為分析為基礎

一、研究背景

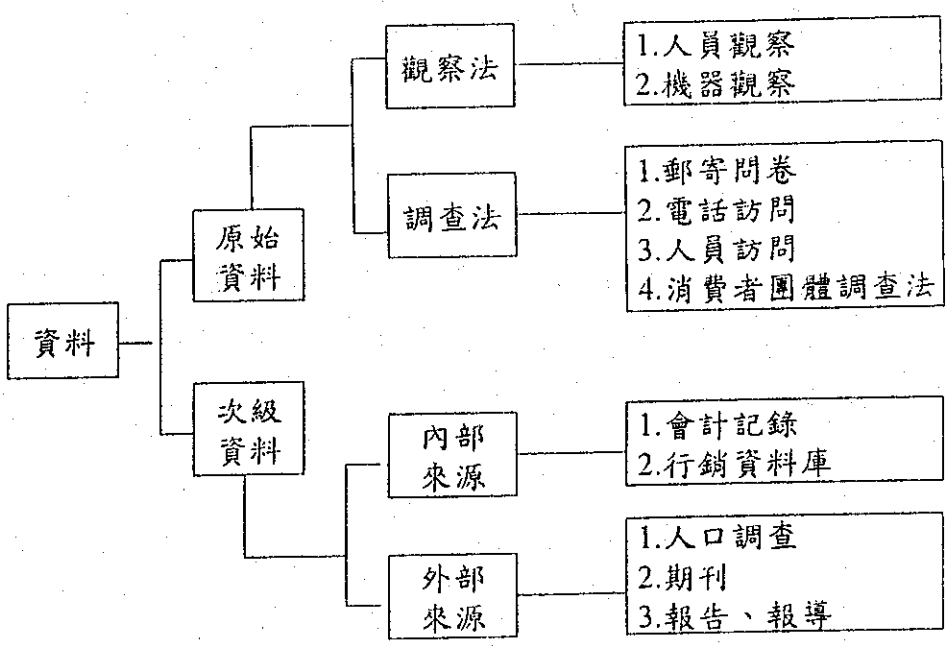
近年來，隨著網際網路的蓬勃發展所帶來的商機正不斷地快速成長中，其主因在於以全球資訊網為環境所進行的電子商務能同時為行銷者與消費者帶來傳統媒體或通路所不及的優點。就行銷者而言，網站的建置成本(如租賃、實體設備費用)遠低於實體商店 (Berthon, et al., 1996)，並具有容易維護、快速變更等特性，能因應市場環境的變化進行快速的調整 (Kotler, 1997)；網際網路的多媒體與互動性，能直接與消費者進行一對一的行銷溝通，建立穩固的關係；此外，網際網路尚擁有全球展示、一天二十四小時及全年無休等優勢。

在消費者方面，Janal (1995)曾指出網際網路為消費者帶來了三個主要利益：(1)購物便利性：消費者能在任何時間、任何可上網的地方訂購產品，其購物過程不再需要搭乘交通工具、尋覓停車場，也不需要親自走訪數家商店才能檢視商品。(2)快速蒐集豐富資訊：消費者可在網際網路中同時蒐集相關企業與其產品，以及不同競爭廠商間的資訊，因而較能在價格、品質、功能等方面進行客觀的評估。(3)購物干擾較少：消費者不再需要面對銷售人員不斷地說服與強迫推銷，或是暴露在嘈雜的購物環境中，可平心靜氣地根據自己的需要進行商品選購。

由於全球資訊網能同時為行銷者與消費者帶來上述傳統媒體或通路所不及的優點，許多業者已體認到其商機無限，不僅已架設好網站，並把網路行銷之預算編列在一般的媒體廣告費用內，蓄勢待發，例如國內的中時電子報、SeedNet、奇摩、TVBS 等皆已成為刊登廣告的熱門網站，而 acerMall、年代資訊、中華生協、博客來網路書店等亦試圖成為人群聚集的購物網站。然而，與傳統企業相較，網站架設的成本、進入障礙以及消費者的忠誠度均較低，使得網站間的競爭加劇，再加上網路世界的消費者行為仍然不甚明朗，行銷研究因而顯得頗為重要。

過去行銷研究中，主要的資料類型及研究方法如圖一所示。觀察法是消費者店內行為的記錄，例如要決定店內商品展示的排列方式，可觀察消費者的行動流

程來安排 (余朝權，民 80)。一般而言，要在傳統的零售店透過人員或機器進行觀察，其成本頗高。調查法是店內購物經驗如態度與意見的再蒐集，與消費購物的接近性較低 (Otnes, et al., 1995)，且費時費力。至於在次級資料的內部來源中，會計資料提供了訂單、銷售、價格、存貨水準、應收帳款、應付帳款等提供「結果」之資料，而行銷資料庫則是提供「正在發生」的資料 (Kotler, 1997)。要能獲得這些資料，必需藉助人員將購買資料輸入，或是產品需先貼條碼，而在結帳時將資料掃描進入 POS 系統中，亦需投入大量的人力物力。外部來源方面，包括人口調查、期刊、報告等，這些資料多半是產業或環境面的分析，且其時效性較差。



圖一：行銷研究之資料類別及研究方法

資料來源：余朝權，*現代行銷管理*，五南圖書出版公司，民國 80 年 10 月。

全球資訊網所提供的環境，則提供了行銷者蒐集消費者行為資料、了解消費者的新的機會。由於 Web 所具有互動能力，能將消費者的註冊資料及購買資料直接存入資料庫中，而且網站主機的日誌檔(log files)，可自動將消費者的瀏覽行為記錄下來，因此與上述資料蒐集的方法相較，Web 能以較低的成本取得消費者的資料、即時記錄消費者進入網站的所有動作，並與內部資料庫相結合。本研究之研究背景，即基於電子商務蓬勃發展，網站競爭加劇，而網路技術提供了蒐

集消費者資料更豐富、更快速且成本較低的方法。

二、研究動機

行銷決策向來便是企業的重要營運決策，常常藉助於各項行銷研究結果來擬定。根據 Kinnear 與 Root (1994)對於 435 家公司調查的結果，歸納出一般企業主要的行銷研究活動包括產業與經濟、訂價、產品、通路、銷售促進，以及購買行為六大類。基於研究背景已提及的網路商店競爭激烈，且網站修改與建置所需的時間與成本均較傳統商店少，使得網站管理者必需快速修正行銷決策，才能滿足消費者的需求，進而掌握商機。此外，網站不僅可記錄消費者的購買結果，還可記錄線上瀏覽與購買過程，亦即相較於傳統市場研究而言，網站可以蒐集到更多的資料來進行市場研究，進而對於顧客及市場有更多的了解。但是網站若未能在架站之初即已規劃需要蒐集、儲存哪些資料，將會發生累積了大量的資料，卻不知如何處理的問題，或是欠缺了部份的資料，導致日後的分析不夠周延。因此，若能彙總網站管理者最常面臨的是哪些決策問題，要分析這些問題需要蒐集哪些資料，以及所使用的分析方法，將能對網站管理者有所助益，此為本研究的第一個研究動機。

至於目前網站經營者所進行的分析，以日誌檔與線上訂單為主 (Lai & Yang, 2000)。由於網站伺服器均會自動將瀏覽者的存取記錄寫入日誌檔中，因此日誌檔是網站管理者最容易取得的一種資料來源。日誌檔所儲存的資訊欄位包括了 IP 位址、使用者識別(AuthUser)欄位、格林威治標準時間、資料請求格式、檔案名稱、狀態或錯誤碼、檔案大小等 (Allen, et al., 1997)，及其他可記錄送出與接收之位元組、花費時間、Cookie 等擴充欄位。針對日誌檔進行分析可獲得網頁流量及瀏覽狀況等之參考資訊，例如 WebTrends 即是一個能針對日誌檔提供多種分析功能的軟體產品，並以統計圖表顯示其分析結果。

就目前日誌檔分析之軟體所提供之功能來看，有以下兩點不足之處：

(1)僅能粗略分析出網頁流量與瀏覽行為：由於日誌檔所能分析的欄位有限，無法再深入分析使用者人口統計資料與瀏覽行為、網頁性質以及行銷做法間的關係。本研究將配合會員制的做法，提出其他如產品特性、人口背景、行銷做

法、網頁安排等擴充資料與日誌檔資料進行參照，俾能進行更有意義的行銷決策分析。此外，由於日誌檔所記錄的資料僅限於線上瀏覽，無法反映出消費者的購買行為，這部份的分析必需藉助於線上訂單。

(2)多侷限於一般統計彙總之功能：事實上，在資料探勘(data mining)領域或是智慧型系統中，已發展出一些相關技術用來分析網頁結構，例如 Chen 等人(1998)所提出的路徑旅行型態(path traversal patterns)問題，由使用者一系列的瀏覽路徑中，找出最大參考序列(maximal reference sequences)，以了解使用者最常走訪的網頁路徑，進而輔助系統之設計與制定更好的行銷決策，如決定更佳的超鏈結、較佳的廣告放置網頁等；Wasfi (1999)則根據多位使用者瀏覽網頁的型態，計算每個網頁被擷取的條件機率，藉以做為瀏覽過程之推薦；而 IBM 所研發的 WBI (Web Browser Intelligence)代理人則會監控使用者的瀏覽路徑，提供捷徑連結(shortcut links)之線上輔助 (Barrett, et al., 2000)。分析工具若能納入這些資料探勘技術，對於網站管理者在設計與安排網頁上將能有所助益。

網站上的交易方式，是消費者在網頁上訂購其有興趣的產品，再透過線上或離線付款的機制進行付款。由於這些線上訂單均可輕易地被網路商店所記錄下，因此這些線上訂單成了網路商店分析消費者購買行為的另一個主要的資料來源。至於線上訂單所能進行的分析方式，則可以參考零售業者在購物籃分析(basket analysis)的做法。由於 POS 系統中，零售商店只要透過掃描機等設備即可將消費者的購買資料存入電腦中，因此零售業者在八〇年代即開始發展出一些購物籃為分析單位的方法，做為制定產品與商店佈置等決策之參考。

就目前線上訂單所能進行之分析來看，有以下兩點不足之處：

(1)欠缺結合會員制、線上購物及行銷做法等之線上訂單分析工具：目前雖已有可進行關聯法則(association rules)與集群分析(clustering analysis)等資料探勘的相關工具，但是這部份必需配合會員制與線上購物功能才能蒐集到會員背景、線上購物及所享之行銷優惠等相關資料，而且資料必需經過轉換才能供現有軟體進行分析。此外，目前軟體亦無法深入探討特定行銷做法或產品特性對於線上購買的影響。

(2)無法分析消費者之購物過程行為：消費者可能在購買過程中，對於許多

產品均有興趣，基於互補或替代產品的考量、預算限制、折扣誘因等因素，而決定其最後之購買決策。由於線上訂單僅是最後訂購結果的記錄，因此無法進一步針對消費者的購買過程進行分析。

由於全球資訊網具有很好的互動特性，能追蹤並記錄使用者的線上行為，因此，除了上述所提的日誌檔以及線上訂單外，目前在全球資訊網中還能記錄購物車(shopping cart)的資料，進而分析消費者在購物時的決策過程。所謂的購物車，是提供消費者在瀏覽產品型錄時，若是看到其所欲購買的產品，將產品暫存的一個空間；當消費者欲完成訂購時，會先確定購物車內的資料無誤，再將線上訂單送出以完成交易。一般購物車的基本設計，提供了增加、刪除產品，以及修改訂購數量的功能；若是進階的設計，則可以再選擇訂購產品的屬性，例如顏色、尺寸等。

目前的分析工具，僅就日誌檔或是線上訂單分別進行分析。事實上，若能把日誌檔的線上瀏覽、購物車的購物過程，以及線上訂單的購物結果資料進行整合分析，將能更精確地掌握消費者的興趣與偏好，例如找出哪些產品消費者可能瀏覽過、或是放入購物車中而最後卻未購買，再深入探究其潛在的原因。又如購物車中的某些產品可能會因為其他產品的加入而被移出，如此將能更深入分析出存在於產品間的替代性，此為本研究的第二個研究動機。

綜合言之，與傳統商店相較，網路商店的確因著許多的線上行為很容易地被記錄下來而累積了大量的資料，但是由於目前的分析工具所提供的功能有限，無法提供經營者了解產品特性、人口背景、行銷做法、網頁安排之間的關係，而且亦未能提供一個能整合線上瀏覽、購買過程及購買結果的分析架構，徒使經營者面對大量的資料卻又苦無適當的工具輔助其進行網路行銷決策之制定。

三、研究目的

根據上述之研究背景與動機，本研究之主要研究目的即在於提出支援網路行銷決策之線上行為分析架構，並探討如何加入購物車資料於日誌檔及線上訂單分析中，以實驗方法驗證整合日誌檔、訂單及購物車之分析方法能改進網站運作之績效。本研究的主要研究目的可以歸納為下列三點：

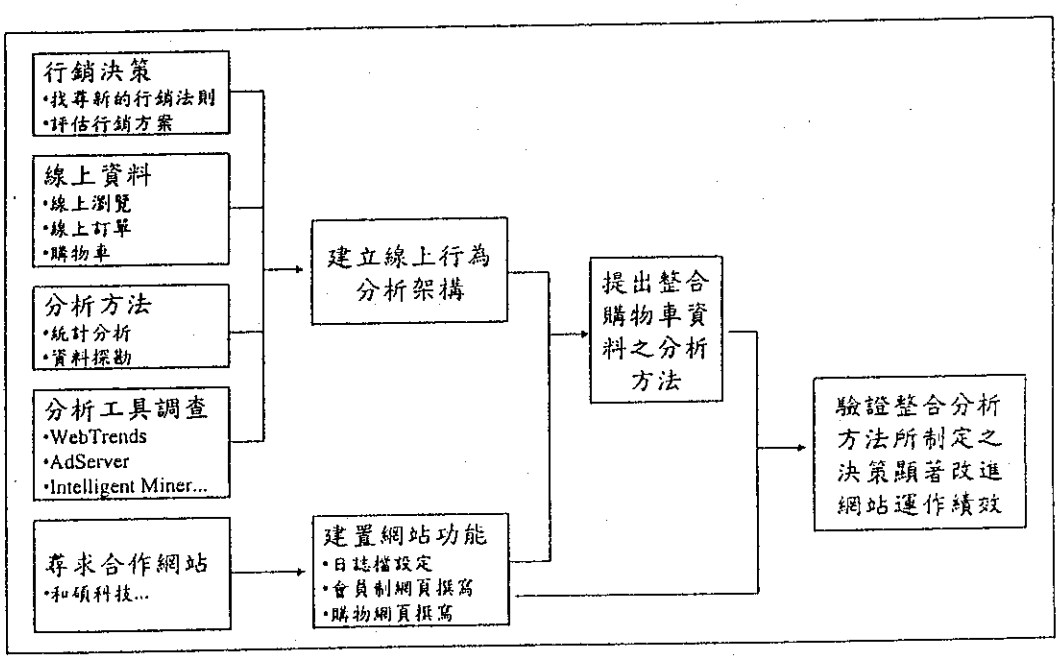
(一)根據相關文獻，整理出網站經營者在網路商店有哪些重要的行銷決策，並針對各個重要網路行銷決策，提供所需蒐集之資料以及分析方法，以提出支援網路行銷決策之線上行為分析架構。

(二)探討如何整合購物車資料以支援日誌檔及線上訂單資料分析，並提出購物車資料有哪些新的資料探勘問題。

(三)與實際運作中之網站合作，以實驗方法驗證整合日誌檔、線上訂單及購物車之分析方法所制定之決策能改進日誌檔分析及線上訂單分析所制定決策之網站運作績效。

四、研究內容與進行步驟

本研究之進行步驟如圖二所示，可分成四個主要階段進行，各個階段所進行的內容敘述如下：



圖二：研究進行步驟

階段一：建立線上行為分析架構

本研究將針對行銷者所需制定的行銷決策、可能的線上決策資料來源以及分析方法之相關文獻，並調查現有分析線上行為之工具，以提出一個支援網路行銷決策之線上行為分析架構，如表一所示。

表一：線上行為分析架構

分析方法 / 資料分類		統計分析		資料探勘技術		
線上行為	原始資料	相關資料				
		產品屬性	個人背景	行銷做法	時機	網頁屬性	敘述統計	歸納統計	限制基礎	自動啟發	最佳化法則
線上瀏覽											
線上訂單								行銷決策法則			
購物車											

在線上行為相關的行銷決策方面，根據過去的相關文獻 (Julander, 1992; McCann & Gallagher, 1991; Russell & Kamakura, 1997)，可分為找尋新的行銷法則以及評估行銷方案兩大類。其中找尋新的行銷法則是透過資料的蒐集與分析，輔助行銷者尋找出有意義的行銷法則，包含了：(1)市場區隔、產品區隔、消費者分群等決策；(2)熱門網頁、產品等決策；(3)聯合促銷、組合產品、替代品及互補品等決策。而在評估行銷方案方面，則是針對既有的行銷方案，透過線上資料的蒐集與分析，評估進行各個方案時之行銷效果，以供行銷者調整其行銷決策之參考。包括：(1)優惠專案、廣告、折扣等行銷方案之效果之評估；以及(2)網頁安排、產品型錄等之評估。

在線上資料分類上，可分為原始資料與相關資料兩大類，其中前者是指網站伺服器所能記錄下的行為資料，有關線上瀏覽、線上訂單及購物車之資料來源與原始資料格式，整理於表二。其中線上瀏覽行為可直接透過網站主機的日誌檔進行記錄，而線上訂單與購物車資料則需另外撰寫 CGI 程式將資料寫入資料庫中，若是希望能直接寫入日誌檔，則購物網頁需配合 Cookies 的撰寫以及日誌檔格式的擴充設定。至於相關資料方面，則是將原始資料針對下列屬性進行對應，供後續行銷分析之用：(1)產品屬性：將產品編號對應至其所代表的產品名稱、分類、價格、與哪些產品是替代品或互補品等 (郭興恩、許中川，民 88)；(2)個人背景：將使用者編號對應至該使用者之人口統計、地理統計或生活型態等資料 (Adomavicius & Tuzhilin, 1999)；(3)行銷做法：將產品編號及日期時間對應至其所在活動期間的行銷做法 (Julander, 1992; Kahn & Schmittlein,1992)；(4)時機：與時間相關的情況或情節，例如時辰、星期、假日與否等 (McCann & Gallagher, 1991; Walters & Bergiel, 1989)；(5)網頁屬性：描述網頁性質之資料，例如網頁結構、

網頁位置、網頁深度及網頁屬性分類等。

表二：線上行為之資料來源與原始資料格式

線上行為	資料來源	原始資料格式
線上瀏覽	日誌檔	(瀏覽編號，使用者編號，瀏覽網頁，日期時間)
線上訂單	CGI 程式/日誌檔	(訂單編號，使用者編號，(產品編號，單價，訂購數量)，運費，訂購金額，日期時間)
購物車	CGI 程式/日誌檔	(瀏覽編號，(產品編號，單價，數量變動)，運費，訂購金額，日期時間)

而目前應用於行銷研究的主要分析方法，主要是統計分析的方式。根據顏月珠 (民 80)的整理，統計分析可分為敘述統計學、歸納統計學、實驗設計三大類。其中的敘述統計學包括統計方法中關於統計資料的蒐集、整理、陳示、分析、解釋等部份，亦即僅針對統計資料本身進行討論，並不將其意義推廣至更大範圍者；而歸納統計學又稱為推論統計學，探討如何由樣本的統計量(statistics)推論母體的母數(parameter)；至於實驗設計，由於其主要目的在於實驗計畫的考量，並不是單純針對資料進行分析，因此不納入本計畫的分析架構中。

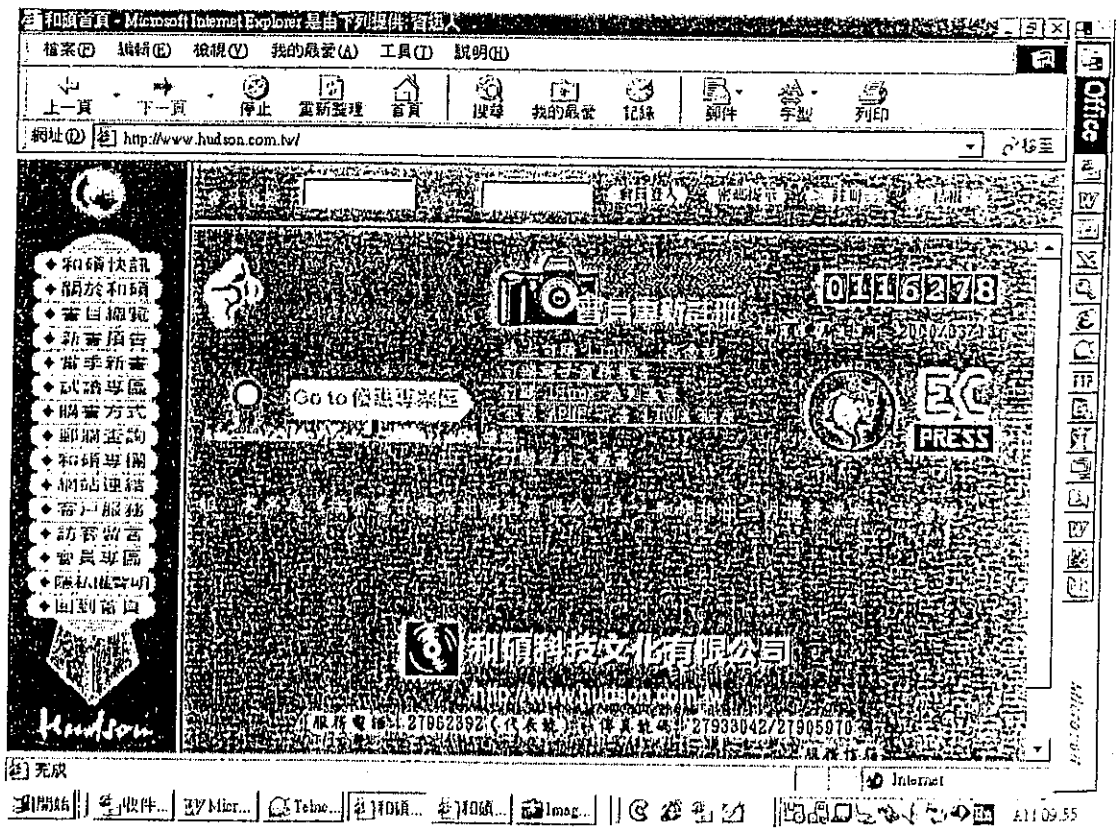
此外，由於科技的進步，可以蒐集到大量資料，因此目前已發展出許多資料探勘的技術，用來輔助由大量資料中快速找出有價值的行銷法則。根據 Bayardo 與 Agrawal (1999)的歸納，資料探勘可分為三類：(1)限制基礎(constraint-based)：探勘出一群符合限制條件的所有法則，例如關聯法則即是由大筆交易中，找出超過最小支持水準(support，亦即項目同時出現次數)之法則。(2)經驗法則(heuristic)：試圖由資料特性中自動學習出法則，但並不保證找出的法則完整，如決策樹等機器學習方法。(3)最佳化法則：找出最有價值的法則，此方式在以限制基礎找出太多法則或是花費太多時間時特別有用。

本研究在第一階段，將提出上述線上行為分析架構，並以此架構為基礎，探討重要的網路行銷決策，需要蒐集哪些線上資料，及其可用之分析方法。

階段二：建置網站功能

本研究在第二階段的主要工作，則是尋求合作之網路商店，建置所需的網站功能，以蒐集消費者之線上資料。本研究目前已與和碩科技網站(www.hudson.com.tw)合作，其網站伺服器是架設於 Windows NT 作業系統下的 IIS (Internet Information Server)，本研究以 ASP (Active Server Pages)開發其會員制及購物網頁，資料庫則使用 SQL Server 7.0，並將日誌檔設為 W3C Extension 的擴充格式，每日記錄線上所有的瀏覽行為。

網站進站首頁如圖三所示，當使用者連至網頁時，系統透過 global.asa 自動給予一個新的編號 SessionNo 存入 cookie 中，由於本研究在 W3C Extension 的日誌檔擴充格式中增加了 cookie 欄位，因此透過 SessionNo 即可區分出不同的瀏覽行為。日誌檔實例如表三所示。若是使用者以會員身份登入，則資料庫將貯存該 SessionNo 所對應之會員編號，藉以辨別出不同會員的瀏覽行為。



圖三：實驗網站登入畫面

表三：日誌檔實例

```
#Software: Microsoft Internet Information Server 4.0
#Version: 1.0
#Date: 1999-10-20 00:04:33
#Fields: date time c-ip s-sitename cs-method cs-uri-stem sc-status sc-bytes cs-bytes time-taken cs(Cookie)
1999-10-20 00:04:33 198.3.103.66 W3SVC5 GET /HudsonColumns/Column59c++.html 200 15769 126 1187 -
1999-10-20 00:06:21 24.0.196.202 W3SVC5 GET /BooksDir/cbuss.htm 200 18249 344 1094
    SessionNo=11331;+ASPSESSIONIDGGQQGQPA=INAIMLEBGPBOHMEKFAFPBPJN
1999-10-20 00:06:30 24.0.196.202 W3SVC5 GET /cbu/new.gif 200 1257 334 9406
    SessionNo=11331;+ASPSESSIONIDGGQQGQPA=INAIMLEBGPBOHMEKFAFPBPJN
1999-10-20 00:06:30 24.0.196.202 W3SVC5 GET /cbu/booksdir01.gif 200 2175 341 6766
    SessionNo=11331;+ASPSESSIONIDGGQQGQPA=INAIMLEBGPBOHMEKFAFPBPJN
```

在會員制部份，系統目前提供了會員登入、密碼提示、註冊、資料修改及會員登出等功能。會員在註冊時，系統所蒐集的會員資料包括會員帳號(身份證號碼)、姓名、E-Mail 信箱、性別、出生年份、血型、教育程度、職業、年收入、居住地區、每週上網時間、經常上網地點、上網主要目的等。至於在購物功能部份，提供了購物車與結帳兩大功能。使用者可透過購物車功能將喜歡的產品置入、取出，以及更改訂購數量；而在確定結帳時，透過結帳功能輸入地址、連絡電話、付款方式即可完成交易。由於購物功能亦將 SessionNo 存入，因此亦可藉以區分出不同使用者的購買行為。

階段三：提出整合購物車資料之分析方法

在第三階段將針對線上行為分析架構的主要行銷決策問題，以及分析實驗網站所蒐集的資料，提出如何整合購物車資料，輔助原先之日誌檔分析及線上訂單分析，並根據購物車資料，提出有哪些新的資料探勘問題。

本研究以民國 88 年 8 月至 89 年 1 月為止，於和碩科技所蒐集的半年資料來舉例。表四與表五是以會員為單位，分別購買、瀏覽 2 項產品以上之關聯法則前三名之分析結果。比較此二表，可發現二者所產生的關聯產品差異甚大，推究其因，可能在於網站中試讀部份並未直接鏈結至購物車，且瀏覽與實際訂購間亦存在差距，因此本研究將僅有試讀而未瀏覽圖書簡介的資料剔除，找出線上瀏覽

中，超過 support=0.05 的關聯法則，再檢視這些關聯產品同時出現於購物車的比例，其結果如表六所示。本研究將於第四階段以實驗方法驗證整合購物車之分析方法是否顯著優於其他方法。

表四：線上訂單之關聯法則分析結果

產品項目一	產品項目二	Support
行銷Any Time--1對1網際網路行銷(B15)	行銷DIY--網際網路行銷計劃(B16)	0.137
電子商務概論(B18)	電子商務管理概論(B21)	0.105
行銷Any Time--1對1網際網路行銷(B15)	電子商務概論(B18)	0.095

表五：線上瀏覽之關聯法則分析結果

產品項目一	產品項目二	Support
Linux萬萬歲(L001)	Visual Basic 控制項妙錦囊(S07)	0.180
LINUX 核心研究篇(L06)	Visual Basic 控制項妙錦囊(S07)	0.168
電子商務管理概論(B21)	Visual Basic 控制項妙錦囊(S07)	0.116

表六：整合購物車資料與線上瀏覽資料之關聯法則分析結果

Item1	Item2	Support	購物車二者均無出現	購物車二者出現其一	購物車二者同時出現
電子商務概論(B18)	電子商務管理概論(B21)	0.082	0.784	0.147	0.069
商機Any Time-打造虛擬商店(B19)	電子商務管理概論(B21)	0.056	0.862	0.087	0.050
電子商務入門(B20)	電子商務管理概論(B21)	0.116	0.839	0.133	0.028

購物車亦可能存在新的資料探勘問題。例如本研究針對同一位會員之購物車資料進行分析，找出曾經被放入購物車又被取出的產品，與其他最後留在購物車產品所進行之關聯分析，如表七所示。而表八則是以其中一個產品被移去，另一個產品最後留在購物車之分析結果。此兩種分析方法將有助於行銷者了解哪些產品可能具有替代性。

表七：購物車產品替代分析法一

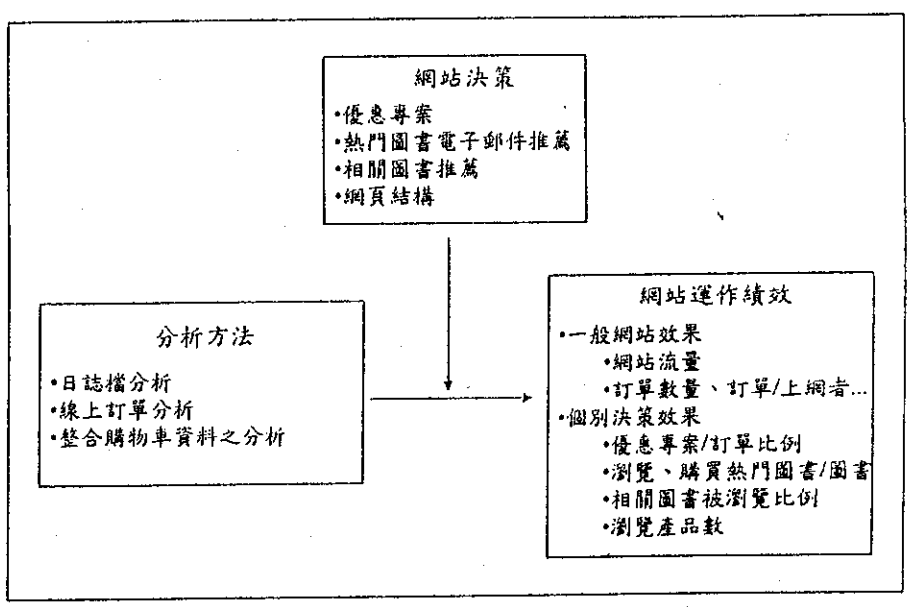
曾置入購物車，最後被移去之產品	最後仍留在購物車之產品	出現會員數
廣告Any Time--網際網路廣告(B13)	行銷Any Time--1對1網際網路行銷(B15)	4
舉登 Linux 高峰(L04)	LINUX 核心研究篇(L06)	4
舉登 Linux 高峰(L04)	Linux 網路應用篇(L307)	4
廣告Any Time--網際網路廣告(B13)	網際網路客戶服務(B09)	4

表八：購物車產品替代分析法二

置入購物車產品中，其一最後留下，另一被移去之互相替代產品	出現會員數
廣告Any Time--網際網路廣告(B13), 行銷Any Time--1對1網際網路行銷(B15)	7
舉登 Linux 高峰(L04), LINUX 核心研究篇(L06)	7
廣告Any Time--網際網路廣告(B13), 網際網路客戶服務(B09)	5

階段四：驗證網站運作績效

本研究在第四階段將以實驗方法驗證整合購物車資料之分析方法所制定的網站決策，其網站運作績效是否顯著優於日誌檔分析及線上訂單分析方法，實驗架構如圖四所示。



圖四：實驗架構

在不同網站決策的比較上，本研究擬進行優惠專案、熱門圖書電子郵件推薦、相關圖書推薦以及網頁結構調整等實驗。其中優惠專案將以關聯法則分析，找出產品組合進行優惠促銷；熱門圖書電子郵件推薦則是以集群分析將消費者分群，找出各群最熱門的十本圖書，以電子郵件寄給會員進行推薦；相關圖書推薦是指圖書介紹網頁有一個超鏈結指到相關書籍，本研究將以個別圖書為單位，分別找出與其最常一起出現之圖書，更改其相關圖書之鏈結；網頁結構調整則將計算進入圖書介紹頁主要是由哪些相關鏈結(和碩首頁、書目總覽、試讀專區、當季新書等)，及各本圖書被點選數與網頁位置被點選數，以修正網頁結構。

本研究將分別以日誌檔分析、線上訂單分析及整合購物車資料之分析三種方法找出上述行銷決策適當的做法，以實驗方法驗證不同方法所制訂定決策之績效。在網站運作績效上，本研究擬以一般網站效果及個別決策效果來衡量。其中前者包括網站流量、訂單數量及訂單佔上網者比例等；而個別決策效果包括優惠專案佔訂單比例、瀏覽及購買熱門圖書佔所有圖書比例、相關圖書之超連結被點選比例、瀏覽產品數等。

五、參考文獻

余朝權，*現代行銷管理*，五南圖書出版公司，民國 80 年 10 月。

郭興恩、許中川，「會員消費資料分析與探勘」，第十屆國際資訊管理研討會論文集，中央警察大學，民國 88 年 6 月 4、5 日，pp. 991-998。

顏月珠，*商用統計學*，三民書局，民國 80 年 8 月。

Adomavicius, G., and Tuzhilin, A., "User Profiling in Personalization Applications Through Rule Discovery and Validation," *Proceedings of the fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Diego, CA USA, 1999, pp.377-381.

Allen, C., Kania, D., and Yaeckel, B., *Internet World Guide to One-to-One Web Marketing*, John Wiley & Sons, 1997.

Barrett, R., Maglio, P. P., and Kellem, D. C., "How to Personalize the Web," *http://www.almaden.ibm.com/cs/wbi/papers/chi97/wbipaper.html*, Jan 19, 2000.

Bayardo, R. J., and Agrawal, R., "Mining the Most Interesting Rules," *Proceedings of the fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Diego, CA USA, 1999, pp.145-154.

Berthon, P., Pitt, L. F., and Watson, R. T., "The World Wide Web as an Advertising Medium: Toward an Understanding of Conversion Efficiency," *Journal of Advertising Research*, Vol.36, No.1, 1996, pp.43-54.

Chen, M.-S., Park, J. S., and Yu, P. S., "Efficient Data Mining for Path Traversal Patterns," *IEEE Transactions on Knowledge and Data Engineering*, Vol.10, No.2, 1998, pp.209-221.

Janal, D. S., *Online Marketing Handbook*, Van Nostrand Reinhold, New York, 1995.

Julander, C.-R., "Basket Analysis: A New Way of Analysing Scanner Data," *International Journal of Retail & Distribution Management*, Vol.20, No.7, 1992, pp.10-18.

Kahn, B. E., and Schmittlein, D. C., "The Relationship Between Purchases Made on

Promotion and Shopping Trip Behavior," *Journal of Retailing*, Vol.68, No.3, 1992, pp.294-315.

Kinnear, T. C., and Root, A. R., *1994 Survey of Marketing Research: Organization, Functions, Budget Compensation*, Chicago: American Marketing Association, 1994.

Kotler, P., *Marketing Management - Analysis, Planning, Implementation, and Control*, Prentice-Hall, N.J., 1997.

Lai, H., and Yang, T.-C., "A Group-based Inference Approach to Customized Marketing on the Web -- Integrating Clustering and Association Rules Techniques," *Proceedings of the Thirty-Third Annual Hawaii International Conference on System Science*, 2000.

McCann, J. M., and Gallagher, J. P., *Databases and Knowledge Systems in Merchandising*, Van Nostrand Reinhold, New York, 1991.

Otnes, C., McGrath, M. A., and Lowrey, T. M., "Shopping with Consumers," *Journal of Retailing and Consumer Services*, Vol.2, No.2, 1995, pp.97-110.

Russell, G. J., and Kamakura, W. A., "Modeling Multiple Category Brand Preference with Household Basket Data," *Journal of Retailing*, Vol.73, No.4, 1997, pp.439-461.

Walters, C. G., and Bergiel, B. J., *Consumer Behavior: A Decision-Making Approach*, South-Western Publishing Co, Cincinnati, Ohio, 1989.

Wasfi, A. M. A., "Collecting User Access Patterns for Building User Profiles and Collaborative Filtering," *Proceedings of the 1999 International Conference on Intelligent User Interfaces*, Redondo, CA, 1999, pp.57-64.