ARRAY Logo

icon

# A suitable research methodology for analyzing online banking behaviour

**By Geoffrey J. L. van Meer and W. Fred van Raaij**
**Department of Economic and Social Psychology, University of Tilburg, The Netherlands**

**e-mail:** geoffrey.van.meer@mail.ing.nl

BRIEF BIOGRAPHICAL DESCRIPTION

Geoffrey van Meer is working on issues as on line financial behaviour and data mining as a researcher at a research department of a Dutch bank. He is also working on a doctoral thesis in the school of Economic and Social Psychology at the University of Tilburg (The Netherlands).

Fred van Raaij is a Professor in the school of Economic and Social Psychology at the University of Tilburg (The Netherlands). He is the author of more than 30 books and over hundreds of articles about consumer behaviour specifically on topics as financial planning and media use.

## Abstract

**Banks and financial institutions still perceive the Internet as a "black box" in which little insight is provided about individual-level online behaviour. The authors state that clickstream analyses open up the black box and illuminate online banking behaviour. In this article, examples are presented that show that clickstream analysis is a suitable research methodology for integrating the Internet in the marketing strategy of a bank or financial institution.**

## 1 Introduction

Since the inception of the Internet, the ability of websites to track the behaviour of their visitors has been considered one of the most promising facets of the new medium. The detailed records of web usage behaviour provide researchers and practitioners with the opportunity to study how users browse or navigate websites and to assess the performance of these sites in a variety of ways (Bucklin & Sismeiro, 2002). Many online retailers monitor visitor traffic as a measure of their stores' success. However, summary measures, such as the total number of visits per month, provide little insight about individual-level shopping behaviour. Behaviour may evolve over time, especially in a changing environment like the Internet (Moe & Fader, 2002). As a consequence they are unable to integrate the Internet in their direct marketing strategy. Walsh & Godfrey (2000) mentioned that the Internet offers a

number of advantages. They stated that e-tailors are able to collect and analyse an extensive amount of information continuously, in a very short time and at relatively low costs. E-tailors collect three types of data: (1) basic, personal information provided on registration or via self-completion questionnaires; (2) purchasing habits; and (3) clickstream or site navigation. We are interested in the third type of data: data from clickstream navigation. Kimball & Merz (2000) see the clickstream as a new and exiting data source: "The clickstream contains a record for every page request from each visitor to our site. In many ways, we can imagine that the clickstream is a record for every gesture each visitor makes, and we are beginning to realize that these gestures add up to descriptions of behaviour we have never been able to see before". Clickstream analyses have a few important benefits in analysing online banking behaviour. Based on our experiences over the past two years these benefits are elaborated with examples from practice.

# 2 Clickstream data

How can we analyse visitors' clicking pattern? Therefor we need to collect data from clickstream or site navigation. We decided to use clickstream data from an Application Server Provider (ASP). An ASP is a third party "observer" that collects clickstream data from a website. A pixel gif is a small snippet of HTML code that is inserted into every page the bank wishes to be tracked. Each requested page sends a signal to the ASP. The ASP logs each interaction between the visitor and the website. One of the most compelling reasons to use ASP-data is that it enables the ability to identify unique users across website domains. Another reason is that additional information about screen resolution, browser plug-ins and other type of technical information can be collected. Also additional custom analytic variables (e.g., keywords) can be inserted into the pixel gif. Moreover, ASP-data generates a single data file to import for analyses. The pre-processing phase with ASP data is, comparing with the other types of data collection (e.g., web-server based or user-based) less time consuming. A file with clickstream data is like a maritime log: a detailed description of all the peregrinations on the site. There are different levels of granularity of clickstream data. What types of components are logged depends on the configuration of the server. The following types of components are most essential: Transfer log, Referral log, Agent log. For a more detailed description of different types of log components and types of collecting clickstream data see Mattison (1999); Mena (1999); Kimball & Merz (2000) and Linoff & Berry (2002).

### 2.1 Pre-processing clickstream data

To guarantee the reliability of research findings, data files have to be of high quality. Unfortunately, the quality of "raw clickstream data" is most of the time poor. Depending on the granularity of the data in the imported file, data preparation is therefor necessary and worthwhile. Pre-processing consists of converting usage information contained in the various data sources into the data abstractions necessary for pattern discovery. The extract, transformation, and load (ETL) mechanism is appropriate to get data from where it is created to the data warehouse where it will be used. Data extraction is concerned with extraction from the data sources on an ongoing basis. The data transformation programs do basic translations such as character set conversion or translating codes into human-readable equivalents. The data loading process involves that new or changed dimension data must be processed before data can be loaded into the data webhouse (Sweiger et al., 2002). In the end a few levels c.q., dimensions of the clickstream data might be loaded into the data webhouse. Analyzing activity at the lowest level of granularity, the site hit, is quite useful for generating site traffic statistics and calculating many different metrics associated with site administration. But most businesses analyse user activity from a page or session perspective, requiring two new aggregate tables at the Page Activity and Session Activity levels (Sweiger et al., 2002).

The treatment of outliers, errors, and incomplete data that can easily occur due to reasons inherent to web browsing (Batista & Silva, 2001), is a relevant part of data pre-processing. We like to discuss a few difficulties with collecting and analyzing clickstream data to evaluate users' navigation behaviour. The first problem is caching. Web clients save previous visited pages in the cache to reduce the traffic on the server and the costs. As a result, when a user hits the "back button", the cached page is

displayed and the web server is not aware of the repeat page access (Mobasher et al., 2000). Cooley et al. (1999) use knowledge of the structure of the site to solve the problem of caching: if two pages are not connected but directly visited after each other, this means that a cache-page connected to both pages is revisited. The second problem with collecting clickstream data is proxy servers. In a web server log, all requests potentially represent more than one user. Proxy servers provide an intermediate level of caching and create even more problems with identifying site usage (Cooley, 1999). Also, due to proxy server level caching, a single request from the server could actually be viewed by multiple users throughout an extended period of time (Cooley, 1999). Another impediments to collecting reliable usage data are web spiders. Web spiders are searching the site, but do not understand the site. Web spiders are robots and they leave non-human tracks behind that are thus misleading for understanding visitors' navigation behaviour (Spiliopoulou, 2000).

Aside from counting visitors to your website, the most important information you want to obtain is the number of unique sessions, determined by the different and unique IP addresses in your access log file (Mena, 1999). However, this can be misleading due to the fact that dial-up users are usually assigned a 'dynamic' IP address by their Internet Service Provider (ISP) from a reserved lot. And, corporate users typically connect through proxy servers, which means that groups of individuals use the same address. This means that it is very difficult to identify an individual with a unique IP address. This fact leads to two conclusions: (1) sessions are more important than individual addresses, and (2) tracking by other means is necessary whether by cookies or forms (Mena 1999). Cookies are small text files created by servers on visiting browsers' hard disk that contain an identification code. They are passed by servers to browsers so that each time a website visitor returns to a server that passed the cookie, it can recognize it and read what it wrote before, such as what pages were visited during the prior session. The absence of one unique user key is a problem in analyzing visitor's behaviour on a corporate website of a bank.

The problem in the online realm is that visitors are invisible without clear definitions of a user. People come to your site, leave footprints and move on. However, those footprints are merely an indication that they were there and tell you nothing about the people who made those marks (Cutler & Sterne, 2000). Nowadays, most banks have an online banking application, where customers are able to perform transactions such as fund transfer or bill payment. A customer has a unique identification number, e.g., account number, and a password to gain access to the online applications. It is necessary to log a unique user identification key (e.g., account number or relation number) of the visitors to recognize online banking behaviour of customers.

Standard definitions of core measurements must be defined. Many e-businesses feel they have a good grasp on the basic concepts of hits, page views, and impressions. However, 'what is and what is not a page view'? Does a page merely have to requested or actually delivered? In our research different types of statistical analyses are applied on clickstream data, which contain page requests. There is more discussion about a standard definition of a session. Ideally each user session gives an extract accounting of who accessed the website, which pages were requested and in which order, and how long each page was viewed. We follow previous research and assume that a page request starts a new session if it requested after an idle period of at least 30 minutes (Catledge & Pitkow, 1995).

## 3 Analyzing clickstream data in order to measure online banking behaviour

Information that reflects online banking behaviour can be extracted from a file with clickstream data by applying data mining techniques. Data mining can be defined as the iterative process of detecting and extracting these patterns from large databases: it is a kind of pattern-recognition. Data mining lets us identify 'signatures' hidden in large databases. (Mena, 1999). To analyse clickstream data of a visitor's session it is possible to have a clear insight in the visitor's behaviour in detail. Not only descriptive analyses, but also explorative analyses on clickstream data are possible to get a better understanding of customers' online behaviour. Mining clickstream data is also defined as the process of applying data mining techniques to the discovery of usage patterns from web logs data, to identify web users behaviour (Batista and Silva, 2001). The concept of applying data mining techniques to web server logs was first proposed by Mannila and Toivonen (1996); Yan, Jacobsen, Garcia-Molina and

Dayal (1996); Chen, Park and Yu (1998) and Cooley et al., (1999). Since then many studies were conducted in this field, but none in the field of financial services. Down here two applications of analysing clickstream data are described: (1) website usage and (2) online banking behaviour. Regarding the latter we provide three examples to underline what types of techniques can be used to analyse online banking behaviour. These research activities show that analysing clickstream data is a suitable research methodology to measure online banking behaviour.

### 3.1 Website usage

It is necessary for e-marketers to monitor the usage of the site. Monitoring means to follow and to evaluate actual number of page views, sessions and visitors. Therefor time series are needed to visualise the site's usage per month, per day or even per second. Analyzing clickstream data during a past period provides an exact overview of the visitors' usage of the website.

Clickstream analyses function as a market barometer. Whatever happens is registered very precisely. Time series are very sensitive for any small change in website usage. Based on this type of information, marketers are able to keep up with the latest developments in financial markets. With the proper infrastructure, monitoring website usage is possible from day-to-day or even in real time.

Clickstream analyses can also function as an early warning system for the performance of a site. When the number of page requests suddenly drops or the number of errors generated by the server increases, it signals that the website faces performance difficulties. The customers cannot enter the online store. Clickstream analyses may not only contribute to commercial, but also to operational benefits.

### 3.2 Online banking behaviour

You cannot manage what you do not measure, and you cannot measure what you do not define (Cutler & Sterne, 2000). There are many different types of behaviour on a website. How could we define online banking behaviour? Depending on the strategy of the website, only a limited number of visitors' actions is desirable. These actions are measured by 'key performance indicators (KPIs)'. KPIs are important determinants of a successful website, and basically follow from the business process that is being supported by the site. It is possible to specify different types of desirable online banking behaviour, i.e. KPIs. E-marketers, together with researchers, must make a list with KPIs for each website. Also from explorative clickstream analyses useful and interesting KPIs can be found. KPIs should be monitored continuously.

From explorative clickstream analyses also other useful and interesting KPIs can be found. Again, it is necessary for e-marketers to monitor usage of the site. Therefor time series of the specified KPIs are needed to visualize the site's usage in a time perspective.

Based on this type of information marketers are capable to monitor the total number of online payments on a detailed level. It is important to monitor the key performance indicators of a website. Time series are suitable to visualise the usage of a website, and the different types of online banking behaviour on a website. Time series are ideal evaluation tools. E-marketers must set themselves explicit targets for the KPIs. Alterations in the design of the website are also measured. With the proper infrastructure, time series could be reported automatically and displayed in every desired format for the end-users.

*Relationship between visitors' surfing behaviour and KPIs*

Every company strives to reach the KPI targets. How can a bank improve and adjust the website to reach or even exceed the targets? We must analyse how the design of the website contributes to realize the KPI targets. Detailed analyses of visitors' surfing behaviour detect which parts of the website are more or less useful or absolutely important for the KPIs. With this type of analysis unknown

relationships are extracted. This proves that clickstream analyses is data-driven. Perhaps unknown important factors are extracted.

Clickstream analyses offer ideas for new content. The consequences of the various alternations on the websites are evaluated. Clickstream analyses functions as a monitor for a system of KPIs as well as a tool for optimising the structure, design and content of a website.

*Segmentation of visitors' sessions*
To gain insight in visitors' online behaviour we need to analyse what groups of pages are requested within a session. If there are distinct clusters of pages, there is also a segmentation possible of surfing behaviour of visitors. We need 'correspondence analysis' to cluster pages and 'hierarchical cluster analysis' to find a segmentation of sessions based on clusters of pages. Correspondence analysis is needed to visualize the hidden structure in the datamatrix (Greenacre, 1984). Correspondence Analyse is a multivariate representation of contingency table containing information about combinations of the requested pages during one session. The 'principal inertion' of the algorithm qualifies if the representation is acceptable, and what is the importance of the axes (proportional variance) for the total solution (Hoffman & Franke, 1986). Pages in the figure that are requested relatively often within the same session are lying close to each other. And, pages that are not combined in the same session are lying far apart.

Groups of pages provide the right information about how to steer customers directly to the KPIs. It is also clear that this information helps content managers to decide which part of the website needs more development and what part does not.

We assume that a visitor begins a session with a specific need for information. The design of the website should match visitors' needs. If a customer wants something from his or her bank, it is just one mouse-click away on the Internet. Therefor a segmentation of visitors' sessions - different types of information needs - is absolutely necessary for optimising the website. The segmentation is based on different typologies of sessions.

Typology of sessions helps to differ in communication with different types of visitors. Clickstream analysis operates here as a tool for optimising the website as well as directing customers through the website.

## 4 Clickstream analysis versus OLAP

Online Analytical Processing (henceforth, OLAP) is way of presenting relational data to end-users to facilitate understanding the data and important patterns inside it. Like visualization tool in the arsenal of weapons used for extracting and presenting information (Berry & Linoff, 1997). The basic difference between OLAP and data mining is that former is about aggregates, while data mining is about ratios (Mena, 1999). Another difference between OLAP and data mining is how they operate on the data. Similar to the direction of statistics, OLAP is a top-down approach to data analysis. Manual OLAP may be based on need-to-know facts, such as regional sales reports stratified by type of business, while automatic data mining is based on the need-to-discover what factors are influencing these sales (Mena, 1999). OLAP tools are powerful and fast tools for reporting on data, in contrast to data mining tools that focus on findings patterns in data. Most areas where there is interest in data mining can also benefit from OLAP (Berry & Linoff, 1997). Besides all differences mentioned above, an important benefit of clickstream analyses is that data mining provides bottom-up analysis and requires no assumptions (Mena, 1999). Besides this, when a unique user identity key, e.g. account number, is logged, clickstream data can be enriched with customer information from the back office. So, not only online customer behaviour is viewed, but customer behaviour through multi channels.

## 5 Clickstream analysis versus market research

Every year a company invests a lot of resources in market research. The outcomes of market research contribute to the development and the optimisation of products and services of the company. Market research asks customers what they think, need or wish. Unfortunately, there is a gap between what people say and what they actually do. Comparing with market research clickstream analysis has a few advantages: (i) clickstream data are collected unobtrusively and based on observed behaviour, (ii) they are free from confounds of researcher interaction, and (iii) time pattern and order of activity is recorded.

## 6 Conclusions

Banks and financial institutions still face the Internet as a "black box" in which little insight is provided about online behaviour at the individual level. As a consequence they are unable to integrate the Internet in their direct marketing strategy. Clickstream analyses open up the black box and illuminate online banking behaviour. The detailed records of web usage behaviour provide researchers and practitioners with the opportunity to study how users browse or navigate websites and to assess the performance of these sites in a variety of ways. Pre-processing the clickstream data consists of converting usage information into the data abstractions necessary for pattern discovery. It is a hard job, but is worthy. The opportunities for clickstream analyses are far-reaching. We described the following applications of analysing clickstream data in order to measure online banking behaviour:

- Website usage: time series are accurate and easily show trends in visitors' behaviour. Clickstream analyses provide insight in customers' behaviour on a very detail level. Monitoring website usage with basic statistics foster decision-making to counteract on a possible decrease. E-marketers and content managers may decide whether or not to change the content of the website. As a result clickstream analyses integrates client focus into the daily operation.
- Online banking behaviour: depending on the strategy of the website, only a limited number of visitors' actions is profitable. These actions are defined and measured by "key performance indicators (KPI)". KPIs are important determinants of a successful website. KPIs should be measured continuously with time series. Based on the past figures marketers or managers can put future targets. Clickstream analyses may help to optimise the profit of a website.
- Relationship between visitor's surfing behaviour and KPIs: banks face problems in a lack of knowledge how to construct website to optimise KPIs. How can content managers improve and adjust the website to reach or even exceed the KPI targets? A logistic regression, with the KPI as a dependent variable and other types of page visits as independent variables, shows which pages are increasing the probability of reaching the KPI target. Clickstream analyses functions as a tool for optimizing the structure, design and content of a website.
- Segmentation of visitors' sessions: the design of the website should match visitors' needs. We assume that a visitor begins a session with a specific need for information. If a customer wants something from his or her bank, it is just one mouse-click away on the Internet. Therefor a segmentation of visitors' sessions is absolutely necessary for optimising the website.

All these applications prove that clickstream analyses functions as (1) a monitor for a system of KPIs as well as (2) a tool for optimizing the structure, design and content of a website. Based on experiences with clickstream analyses the benefits, in contrast to other tools, are:

- Different types of OLAP-tools measure site usage, but are incapable to understand online behaviour. Clickstream analyses operate as need-to-discover on the clickstream data. Perhaps unknown important factors are extracted. Hidden relationships between certain types of surfing behaviours and KPIs are found. Clickstream analyses are able to explain the mechanism of the measured behaviour at the individual level.

- Clickstream data are collected unobtrusively and based on observed behaviour, rather than self-reports, and time pattern and order of activity is recorded. Clickstream data displays customers' online behaviour in detail; every (mili-)second customers' online banking behaviour is recorded. This type of data source can be enriched with customer information from the back office. So, not only online customer behaviour can be viewed, but also customer behaviour through multi-channels like mail or telephone.

Just the presence of a new data source is not necessarily sufficient to call for a new program of research (Bucklin et al., 2002). Future research should find out whether or not online banking behaviour is any different from off line banking behaviour. If it is not, then we already have a suitable research paradigm and a wide range of research techniques for analyzing online behaviour. Studying banking behaviour is interesting because financial services are non-tangible and digitally stored, and most consumers have already bought financial service products, resulting in a high penetration of financial services among consumers. Clickstream analyses are not only applicable for online banking behaviour, but of course for any other kind of online behaviour as well. Banks or financial institutions still face the Internet as a 'black box'. Summary measures, such as the total number of visits per month, provide little insight about individual-level shopping behaviour. Clickstream analyses open up the black box and it illuminates online banking behaviour and provides opportunities to integrate the Internet in the direct marketing strategy of a bank.

**References**


Batista, P. & Silva, M.J. (2001). Mining web access logs of an online newspaper. 12th International Meeting of the Euro Working Group on Decision Support Systems.

Berry, M.J.A. & Linoff, G. (1997). Data Mining Techniques for Marketing, Sales, and Customer Support. John Wiley & Sons Inc.

Bucklin, R.E., Lattin, J.M., Ansari, A., Gupta, S., Bell, D., Coupey, E., Little, J.D.C., Mela, C., Montgomery, A. & Steckel, J. (2002). Choice and the Internet: From clickstream to research stream. Marketing Letters 13 (3) pp. 245-258.

Bucklin, R.E., & Sismeiro, C. (2002). A model of website browsing behavior. Journal of Marketing Research, (August) pp. 249-267.

Catledge, L.D. & Pitkow, J.E. (1995). Characterizing browsing behaviors on the world wide web. Computer Networks and ISDN Systems, 27 (6) pp. 1065-1073.

Chen, M.S., Park, J.S., & Yu, P.S. (1998). Efficient data mining for path traversal patterns. IEEE Transactions on Knowledge and Data Engineering, 10 (2) pp. 209-221.

Cooley, R., Mobasher, B., & Srivastava, J. (1999). Data preparation for mining World Wide Web browsing patterns. Journal of Knowledge and Information Systems, (1) 1, pp. 11-15.

Cutler, M., & Sterne, J. (2000). E-Metrics - Business Metrics for the New Economy. NetGenesis corp. - White Paper.

Greenacre, M.J. (1984). Theory and Applications of Correspondance Analysis. London: Academic Press.

Hoffman, D. & Franke, G. (1986). Correspondence analysis: graphical representation of categorical data in market research. Journal of Marketing Research, 23, pp. 213-227.

Kimball, R., & Merz, R. (2000). The Data Webhouse Toolkit. John Wiley & Sons Inc.

Linoff, G.S. & Berry, M.J.A. (2002). Mining the Web: Transforming Customer Data into Customer Value. John Wiley & Sons Inc.

Mannila, H., & Toivonen, H. (1996). "Discovering generalized episodes using minimal ocurrences." In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, pp. 210-215. Montreal, Quebec.

Mattison, R. (1999). Web Warehousing and Knowledge Management. McGraw-Hill.

Mena, J. (1999). Data Mining your Website. Digital Press.

Mobasher,B., Cooley, R. & Srivastava, J. (2000). Automatic personalization based on web usage mining. Communications of the Association for Computing Machinery. 43 (8) pp.142-151.

Moe, W.W. & Fader, P. (2002). Capturing Evolving Visit Behavior in Clickstream Data, "under 2nd review at Journal of Interactive Marketing".

Spiliopoulou, M. (2000). Web usage mining for websites evaluation. Communications of the Association for Computing Machinery. 43 (8), pp.127-134.

Sweiger, M., Madsen, M.R., Langston, J. & Lombard, H. (2002). Clickstream Data Warehousing. Wiley Computer Publishing.

Walsh, J. & Godfrey, S. (2000). The Internet: A new era in customer service. European Management Journal. 18 (1), pp. 85-92.

Yan, T., Jacobsen, M., Garcia-Molina, H., & Dayal, U. (1996). "From user patterns to dynamic hypertext linking." In: Fifth International World Wide Web Conference, Paris, France.