Relevance of Time Spent on Web Pages

Peter I. Hofgesang
Free University Amsterdam
Department of Computer Science, Faculty of Sciences
De Boelelaan 1081a, 1081 HV Amsterdam
The Netherlands
hpi@few.vu.nl

ABSTRACT

What is the intention of an online visitor? This is one of the most valuable pieces of information for an online service provider. In the case of a real-world shop, customers have the ability to explicitly express what they are looking for. However, on the web, the intention is hidden and can only be partially revealed from implicit indicators in the traces users leave behind while they browse through a website.

The vast majority of researchers in web usage mining exploit only two types of information: the order of visited pages and their popularity - i.e. the number of times they were visited. However, several studies in information retrieval and human-computer interaction have suggested a third factor, the time spent on web pages (TSP), as an important measure of intention and relevance. The key contributions of this paper are: (1) an extensive survey of possible factors that influence the TSP measure, and (2) a similarity measure that applies to TSP and can be used to cluster users based on their assumed intentions. Our experiments are based on log files generated by several commercial websites.

Key Words: web usage mining, time spent on web pages, clustering

1. INTRODUCTION

In the real world, the customer has the ability to express himself naturally; he can use his (native) language for assistance and he can describe his purpose. As a consequence of this, he either gets the desired information or product, or he gets redirected somewhere else. In general, one party has a concept of the purpose of the visit and the concern of the other party is to resolve this concept as accurately as possible.

In contrast to the real-world situation, when a user – a potential client – enters the website of an online service provider he is usually left to explore on his own. His explicit goal and intention are hidden (from the website owner) and they only manifest partially in the form of implicit interest indicators trapped in so-called server- or client-side (web access) log files. Therefore, a conventional interaction that would mediate the supply and demand between two parties is not possible. The interest indicators in the log files include the objects – most often web pages – visited by the user, and the order and time stamp of these visits. If the application also allows the tracing and measurement of client-side behaviour, we obtain a source of additional indicators. These include a more elaborate measurement of page view time, and the tracking of mouse activity and page scrolling behaviour; in special applications, tracking of eye movements on the page is also possible.

In web usage mining (WUM), these indicators have been employed in several models and with different methods to discover the possible interest of users, clustering them based on their traced behaviour, etc.

However, the vast majority of WUM researchers usually apply only one or two types of information: the list of web pages visited by the users and the order in which these pages are visited. These are considered to be the most important types of information for characterising user behaviour. In addition, their importance also results from being able to be recorded accurately and automatically on the server-side web access log.

The web access log data contain yet another measure – the time spent on pages (TSP) – that is a well-recognised relevance and interest indicator in other fields such as information retrieval (IR), human-computer interaction (HCI) and E-Learning (see section 2 for related research). It could easily be assumed that TSP would also be a clear and natural indicator of importance of a page in WUM: the more time users spend on a web page, the more important the page is assumed to be for them. However, only very few articles in WUM present models applying the TSP measure. Why is TSP not interesting to researchers? Why do researchers believe that the frequency measure is a much more relevant indicator of user interest?

In our paper we present related literature concerning TSP in different fields. Our contributions are as follow: we give an extensive overview of the factors influencing the TSP measure. In particular, we draw attention to the most important aspects that must be considered when using TSP. Specifically, we present the statistical properties of TSP using several real-world data sets. Finally, we present a similarity measure that applies the TSP measure and can be used to cluster users based on their assumed intentions.

WEBKDD'06, August 20, 2006, Philadelphia, Pennsylvania, USA.

Paper Organization Section 2 reviews the literature related to TSP and its applications. Section 3 presents a comparative analysis of frequency and TSP measures. Issues concerning preprocessing and statistical properties of TSP are discussed in sections 4 and 5, respectively. A TSP-based similarity measure and the results of the clustering experiment are presented in section 6. Lastly, we summarise our conclusions in section 7.

2. RELATED RESEARCH

Much work has been done on the implicit measures of user preference in the field of IR, HCI and E-Learning. Kelly and Teevan (2003)[13] give an overview of related work on implicit feedback techniques in IR.

One of the earliest evaluations of time aspects was presented by Morita and Shinoda (1994)[16]. Their experiments showed a positive correlation between user interest and the reading time of articles. In addition, they found a low correlation between reading time and the length and readability of an article. Konstan et al. (1997)[14] applied collaborative filtering on Usenet news data to facilitate the selection of relevant information. Their study showed that users' explicit ratings were strongly correlated with the implicit TSP measure. Furthermore, predictions based on this implicit measure were nearly as accurate as those based on explicit numerical ratings. Ding et al. (2002)[7] presented a usage-based ranking algorithm for web IR systems that applies the TSP against standard selection-frequency-based ranking. The study of Kellar et al. (2004)[12] focused on the relation between web search tasks and the time spent on reading results. Their results support the correlation and show that it is even stronger as the complexity of a given task increases.

Despite the proven relevance of the time spent on (web) pages to user interest, only a relatively small number of researches apply this measure in WUM. Some of this research uses TSP to classify web pages into link and content pages. The idea behind this is that link pages are used only as traversal paths in order to reach the content pages that are of the interest to the user. In their paper, Cooley et al. (1999)[6] distinguished auxiliary and content pages based on the reference length of the pages – the amount of time a user spends viewing a page – for finding association rules. Xie and Phoha (2001)[24] used the same definition to get rid of auxiliary pages and to run their clustering algorithm only on the content pages. In the paper by Fu et al. (2001)[9], the two categories, index and content pages, are also distinguished by time spent on pages, among several other heuristics for website reorganisation. Xing and Shen (2004)[25] partitioned navigation time into four discrete intervals: passing, simple viewing, normal viewing and preferred viewing. They defined a preference measure from these categories together with the page counts and used it for mining preferred navigation patterns.

In addition, some researchers applied the TSP measure in order to cluster web users. Heer and Chi (2002)[11] used multi-modal clustering and reported that weighting navigation paths by view time improves accuracy. Shahabi et al. (1997)[19] defined a similarity angle through the feature space based on the total time spent on all common sub-sequences of each session pairs. Banerjee and Ghosh (2001)[3] also proposed a similarity measure similar to [19], but with the main difference of introducing an importance factor to weight the difference of time pairs within the common sub-sequence. Xiao et al. (2001)[23] defined four similarity based measures, among them a viewing-time- and a matrix-based method to cluster web users based on their common interest. Gunduz and Ozsu (2003)[10] combined the order of visited pages and visit time into a similarity metric for page prediction.

Several papers apply TSP for special tasks. Agyemang et al. (2004)[1] defined the concept of a web usage outlier based on the combination of page frequency and the TSP measure. Srikant and Yang (2001)[20] presented an algorithm to discover the expected location of pages to support website reorganisation. Their algorithm locates the backtrack point of a user by evaluating the time spent on pages.

Finally, to illustrate that TSP is a non-trivial measure that strongly depends on preprocessing and the domain, we refer to contradicting issues in related work. The first issue relates to the collection of TSP. Shahabi et al. (1997)[19] proposed a client-side solution to measure TSP more accurately. However, in their survey Srivastava et al. (2000) [21] claimed that client- and server-side measure of TSP is equal or that client-side measures are even worse because of the performance overhead. Mobasher et al. (2000)[15] claimed that the time spent on a page (together with the frequency of occurrence of a page) may not be a good indication of user interest, yet numerous researchers have claimed the opposite (e.g. [16]). They proposed binary weights for clustering instead of the TSP or frequency measures. In their research, Morita and Shinoda (1994)[16] found no correlation between reading time and message length or reading difficulty level. However, a normalisation of TSP by document length was proposed by many researchers, e.g. White et al. (2002)[22] normalised the measured reading time by the length of the documents.

3. INFLUENTIAL FACTORS: FREQUENCY VS. TSP

Users' click information forms the basis of WUM. This information is presumably available for all web server configurations in the form of standardised web access log data. The data allows us to reconstruct the actual visits of users. These reconstructed sessions implicitly include the order and occurrence of visited pages. Since the web access log was originally designed to trace server breakdowns, it is a very limited source of client information. However, the data do also include another measure: the page view time. In theory, both measures are good indicators of user interest: the more frequently a page was visited and/or the more time was spent on it, the more interesting the page is supposed to be to the user. However, on the one hand, the frequency measure is a widely accepted indicator of user interest and, on the other hand, TSP seems to be excluded from WUM research. In this section, we give an overview of the possible influential factors and compare their effects on the frequency and TSP measures.

3.1 Website Hierarchy

The hierarchy of a website has a strong influence on the frequency measure [25]. Pages at the top of the hierarchy get traversed more often as intermediate nodes to reach the desired pages. For example, the root of a website (the home page) often gets the most hits. The position of a web page

within the hierarchy has no direct effect on TSP.

3.2 Data preprocessing: filtering out robot transactions and session identification

We considered the effects of two important data preparation steps: the removal of robot transactions and session identification. Robots are automated programs that systematically fetch information from websites. Therefore, to avoid distortion of usage patterns, transactions generated by robots should be eliminated during preprocessing. However, due to malicious robot transactions, which do not identify themselves, the complete removal of such sessions can be rather difficult if not impossible. The problem is that even a few remaining robot transactions can result in dramatic changes in both measures. A single robot session can result in hundreds or thousands of "artificial" clicks. As a result of these systematic visits, the effect in the case of TSP is a more or less equal page view time (usually very short) throughout such sessions. Setting up thresholds to eliminate robot transactions is not straightforward.

Session identification is another non-trivial step in web data preparation. Because there are several possible session identification methods, the page frequencies within sessions may differ accordingly. The most common method of session identification is called the time frame identification [6]. This method sets a threshold for maximal page view time to form sessions. Therefore, TSP has an upper limit in this threshold (in practice the threshold is around 30 minutes). Other methods, e.g. session identification by cookie information, may lead to extreme values of several days or weeks of page view time.

3.3 Distraction

The effect of distraction is quite obvious and one of the most important issues for TSP. Chatting, answering a telephone call, having a coffee break, etc. all result in a longer TSP although the user is not actively looking at the page. Distraction has no clear influence on the frequency measure; however, interruption of a user may result in revisits and, in the case of larger websites, may lead to the user getting lost in the structure.

3.4 Page type

In addition, the type (information page, contact form, etc.), quality (density, layout, complexity, etc.) and other parameters (length, etc.) of a web page may also influence the TSP measure. Nowadays, websites generate thousands of pages dynamically. Identification of granularity strongly depends on the application and on human labelling criteria.

Furthermore, two other issues may also influence TSP. The speed of reading differs between individuals resulting in unequal TSP measurement. Network traffic (bandwidth) and server load may also alter page view times considerably.

4. TSP DATA PREPARATION

In the previous section, we gave an overview of the most important factors that influence the frequency and TSP measures. Here, we review the essential aspects to consider when preparing web data and, in particular, TSP information. Let us first present a notation for observed session data. We augment session information with the time stamp and TSP.

NOTATION 1. *Let us denote our observed data set as* $D = \{s_1, ..., s_i, ..., s_N\}$*, where* s_i *is the ith session. Each* s_i *consists of an ordered sequence of one or more triplets of page identifiers (of the visited page), time stamps of visits and page view times:* $s_i = (< p_{i1}, T_{i1}, t_{i1} >, ..., < p_{ij}, T_{ij}, t_{ij} >, ..., < p_{ni}, T_{ni}, t_{ni} >)$*, where each* $p_{ij} \in P = \{p_1, ..., p_m, ..., p_M\}$ *(the set of all page identifiers),* T_{ij} *is the actual time stamp of the transaction, and* t_{ij} *is the time spent on page* p_{ij}*.*

A natural way to calculate TSP (t_{ij}) for a given page (p_{ij}) is to subtract the time stamp of the page from the time stamp of the following page:

$$t_{ij} = T_{ij+1} - T_{ij} \ (j < ni, \ ni > 1).$$

Note that web access log data do not contain enough information to calculate TSP for the last visited pages (t_{ni}). This is due to the stateless status of the HTTP protocol. After the information of the last click is registered, there is no further communication with the server and, as a consequence of this, no information about when the user stopped the session. However, TSP of the last page can be rather informative. Did the user find the information he wanted or did he just leave the browser idle? Again, in the case of special applications, client-side measurement may provide a solution. In the case of internal websites (requiring login), we can also obtain this information, and extending the web pages by special scripts (which requires some modification of the pages) would also provide a solution for publicly available sites.

An ideal page view time is the pure time spent on actively reading or interacting (scrolling, filling in a form, etc.) with the given page. The above calculation simplifies this ideal measure at several points, as described in the previous section. To calculate the ideal TSP, we would have to consider the time spent on network traffic, server-side page generation time and distraction, which is the most difficult factor and impossible even to approximate based on pure server-side data. We can summarise the (server-side) calculation of the ideal TSP using the following formula: $t_{ijideal} = T_{ij+1} - T_{ij} - T_{networkTraffic} - T_{serverPageGeneration} - T_{distraction}$

Network traffic and server-side influence on TSP
The first issues to mention are network traffic and server load or server page generation time ($T_{networkTraffic}$ and $T_{serverPageGeneration}$). Both server- and client-side solutions are possible to measure the overhead. On the server side, the time required to generate the requested page can be measured exactly and an estimation can be calculated for the network relay based on request-response times. Client-side measures (e.g.,[19]) based on specialised browsers or scripts can measure the page view time directly.

We believe that the largest problem with client-side measures is that they are limited to special applications (for which it is possible to apply client-side changes, and the data can be collected and merged). In contrast, server-side-based approximations are suitable for most tasks. Furthermore, in practice, page generation time is negligible in many cases and network traffic is also small compared with page view times. However, unexpected high volume traffic or a sudden performance drop would lead to abrupt growth in the "interest" of users in the form of misleading longer page view times.

Article	Min.	Max.	Calc.	Status	Domain
[5]	1 sec	20 min	E	D	General
[18]	-	96 sec	F	R	Job service
[8]	5 sec	10 min	E	D	E-learning

Table 1: Thresholds for TSP outlier detection. Even in similar domains, thresholds can vary widely in range. (Calculation: E = exact threshold, F = given by a formula; Status: D = removal, R = replaced by "normal" view time)

Controlling the distraction factor ($T_{distraction}$) After normalising the technical factors, we must still identify user distraction to obtain the real page view time. As we described in previous sections, coffee breaks, parallel browsing activity, etc. all distract the users' attention. It is impossible to identify such activities on the server side. Client-side measures (such as mouse movement, page scrolling, lost-focus attribute of the current browser, eye-tracking, etc.) can be used to approximate distraction. In practice, however, researchers tend to depend on heuristics gained from observing the real data. It is common to set a threshold for maximal (reasonable) page view time and replace the extreme values by some standard view time calculated from the observed data. There is no golden rule for this threshold in WUM. It depends on the domain (e.g. a news portal has long articles, while a retail shop has images and short descriptions) as well as the users' reading capacity. Standard statistical outlier detection algorithms do not work because they would identify most of the important page view times, where users spent relatively more time, as outliers.

In the literature, several different criteria and thresholds are used to eliminate TSP outliers. Claypool et al. (2001)[5] defined the maximal page view time at 20 minutes and removed outliers (general web browsing). Rafter and Smyth (2001)[18], in a job recruitment web service environment, defined normal reading time using the median of median reading time values per individual job access for both users and jobs. The extreme values were then identified as the values larger than twice the normal reading time and were replaced by this value. In their experiment these definitions resulted in 48 seconds for normal view time and about 1.5 minutes for outlier threshold. Farzan and Brusilovsky(2005)[8] used TSP to weight page visitation frequency. They drew the maximum TSP threshold at 10 minutes and, in the case of an outlier, they left the frequency count intact.

Some of these papers also define a minimal threshold for page viewing time outliers. The idea behind a minimal threshold, as [5] suggests, is that users cannot accurately assess interest in a page in less than 1 second [5] (or 5 seconds in [8]). As a guideline, we refer to an eye-tracking study stating that important information is processed during the first few seconds of a visit [17]. Very small TSP values can also be caused by incidental or "anxious" double clicks. However, we believe that sometimes fast visual scanning of the page, that may take a fraction of a second, is sufficient to decide whether to click further (e.g. viewing images of clothes and clicking on the image itself to forward) or backtrack. In practice, a huge volume of very small TSP values indicates robot transactions or re-direct pages; however, these transactions should be identified and eliminated during preprocessing.

Domain	Mean	SD	Median
Bank	37.7 sec	44.4	27.7 sec
Retail 1	46.4 sec	136.5	13 sec
Retail 2	53.9 sec	186.3	12.4 sec

Table 2: TSP statistics of different domains.

Granularity of pages Nowadays, most websites provide dynamically generated pages, thus the concept of a web page is no longer well defined. In the case of a retail shop, there could be tens of thousands of articles and other parameters that can define a single web page. We can set the granularity on the article level (e.g. a specific pair of shoes) or on the level of article categories (e.g. shoes), etc. The selection of "interest" entities indirectly influences TSP values.

Furthermore, TSP values can be normalised by some page parameters (e.g. length, density, readability). In the paper by White et al.(2002)[22], for instance, reading time was normalised by document length. In contrast to this, Morita and Shinoda (1994)[16] concluded that message length or reading difficulty level are not correlated with reading time. Additionally, the heterogeneity of most websites (text, images of different resolutions, etc.) – unlike the homogenous collections of news digests, "uniform" documents, etc. – aggravate the normalisation process.

Client-side measures As we described earlier, client-side measures may provide extra usage information and more elaborate measurements. However, most of the applications do not allow modifications on the client-side (e.g., use of special browsers). In their work Atterer et al. (2006)[2] presented an approach to track client-side user activity transparently. Since the method does not need client-side modifications it broadens the scope of client-side measures.

5. PROPERTIES OF TSP

Here we give an overview of the most important characteristics of TSP using real-world data sets. The data include two retail shop data sets and clickstream data of a bank's online service. During data preprocessing we identified and removed robot transactions and needless or noisy traffic. We identified unique users based on their cookie information or, when absent, their IP address, and collected their sessions using the time frame identification method [6] with a maximal gap of 30 minutes. Note that further preprocessing of TSP, as described in Section 4, is performed for the evaluation of clustering in the following section.

Table 2 contains standard statistics of the three data sets. The results show the domain influence on TSP. Figure 1 shows the TSP distributions of the three data sets. Surprisingly, the Bank and Retail 1 data have similar characteristics. In general, all three distributions have the common characteristics of a significant peak of around 5-10 seconds, and a very long and tapering tail.

Yan et al. (1996)[26] observed that the distribution of time spent on pages is roughly Zipfian. We ranked our data sets in descending order by frequency values and plotted them on a graph (Figure 2) using logarithmic scales on both axes. Zipf curves follow a straight line when plotted on a log-log scale. Our graphical analysis shows that none of the three distributions follow the shape of the classic Zipf curve.

Both the largest and smaller sizes appear to differ from the simple power laws.

The bank data contained categorical information (e.g., age groups) of clients that we used for evaluation purposes. One of our hypotheses was that the age of users has a high influence on the TSP. We calculated the empirical distributions of TSPs over the different age groups (Figure 5). To our surprise, the distributions were very much alike.

Our experiments (Figure 6) showed an inverse relation between the frequency of pages and the mean time spent on them. The more users visit a page the less time they tend to spend on it and the more time indicates less visitations.

6. CLUSTERING USERS BASED ON ASSUMPTIONS OF THEIR INTENTIONS

In this section we present the clustering of sessions based on a similarity metric together with our experimental results. The goal was to group users based on assumptions of their intentions. It is a rather ambitious goal given that web access log data not only lack explicit descriptive measures, but also are an incomplete record of available implicit indicators.

As stated previously, TSP is considered by prior research in IR and HCI to be a good indicator of user interest. In our approach, we combined TSP and the visitation frequency of pages, since this latter measure is the most accepted indicator of user interest within the WUM community.

Let us first define the notion of the composite session that forms the base of our clustering.

DEFINITION 1. *We combined several aggregates of interest indicators to measure the relevance of a page for a user within a session. The* ***composite session*** *of the ith session is*

$$CS_i = (cs_{i1}, ..., cs_{im}, ..., cs_M),$$

where the components are normalised and combined interest values for all page types.

$$cs_{im} = w_i \prod_{k=1}^{K} measure_k(i, m),$$

where $measure_k(i, m)$ *is the kth measure calculated for the mth page type within the ith session, and* w_i *is a normalisation factor for the ith session.*

In our experiments we used two measures ($K = 2$),

- $measure_1(i, m) = \sum_{j=1}^{ni} \begin{cases} 1 & if \quad p_{ij} = p_{im} \\ 0 & otherwise \end{cases}$
  the frequency component
- $measure_2(i, m) = \sum_{j=1}^{ni} \begin{cases} t_{ij} & if \quad j < ni \quad and \\ & \quad p_{ij} = p_{im} \\ 0 & otherwise \end{cases}$
  the time component

with $w_i = 1/max_m(measure_1(i, m))max_m(measure_2(i, m))$ normalisation factor.

The idea behind this measure is that it biases toward pages that occurred frequently and that more time was spent on. The assumption is that the intention of a user is better reflected by popular pages where users also spend more time.

When measuring the similarity of two sessions, we defined an ideal baseline session based on the two composite sessions in a way that the components sum up to exactly 1.

DEFINITION 2. *The* ***baseline session*** *is defined as the pairwise mean values of the two composite sessions* (s_α, s_β) *normalised by their sum:*

$$BS_{\alpha\beta} = (bs_{\alpha\beta 1}, ..., bs_{\alpha\beta m}, ..., bs_{\alpha\beta M}),$$

where

$$bs_{\alpha\beta m} = \frac{cs_{\alpha m} + cs_{\beta m}}{\sum_{m=1}^{M} cs_{\alpha m} + \sum_{m=1}^{M} cs_{\beta m}}.$$

In addition, we introduced a penalty function that degrades the components of the ideal baseline session.

DEFINITION 3. *Our* ***similarity measure*** *is defined by the sum of pairwise differences between baseline session components and the minimum of the baseline components and the standard deviation of the pairwise composite session components:*

$$SM(\alpha, \beta) = \sum_{m=1}^{M} \widehat{bs}_{\alpha\beta m},$$

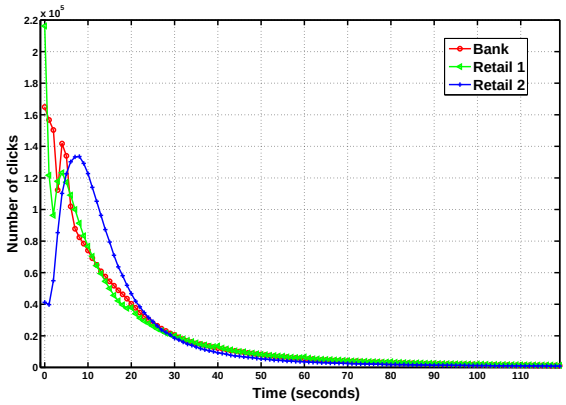

Figure 1: TSP distributions of the three data sets (< 2 minutes, rounded TSP values)

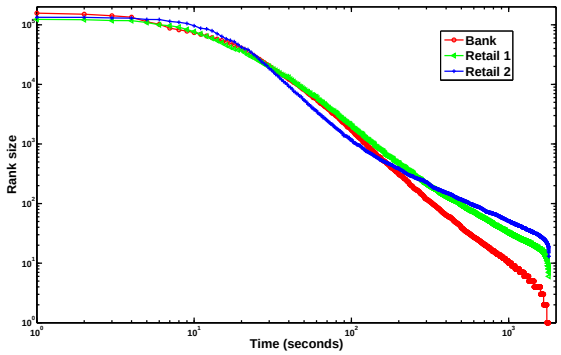

Figure 2: Rank size of the TSP distributions over the three data sets (complete 30-minute time range); logarithmic scales on both axes

where

$$\widehat{bs}_{\alpha\beta m} = bs_{\alpha\beta m} - min(bs_{\alpha\beta m}, std(cs_{\alpha m}, cs_{\beta m})),$$

where $std(cs_{\alpha m}, cs_{\beta m})$ *is the standard deviation of values* $cs_{\alpha m}, cs_{\beta m}$.

Note that this measure is defined to sum up to 1 in cases where the original sessions were identical, and 0 if they were completely different. In other cases, a value between 0 and 1 reflects the similarity of the two sessions.
In addition, note that, in practice, most sessions include only a few visits to distinct pages; therefore, the composite sessions are very sparse and calculations can be simplified accordingly.

6.1 Clustering experiment

In our experiments we analysed web usage data of a retail company's website (Retail 2). We performed the data preprocessing steps described in section 5. In addition, a server-side approximation was available for network latency and server page generation time, which we used to normalise the time spent on page values. After extensive exploration of the data we chose a fixed, 5-minute threshold to detect page view outliers. We replaced these values by the mean page view time (without the outliers) within the sessions.

Our retail shop data set contained 1 month of traffic, which we transformed into 2.7 million sessions. In addition, we removed sessions that contained fewer than 3 clicks, assuming that most of them were accidental visits. From the remaining set, we randomly selected 5,000 sessions. We calculated the pairwise similarity matrix (only the upper triangle) based on our similarity measure, transformed the values into 2D space using multidimensional scaling and, finally, we used k-means clustering with different k values (1-10).

We analysed the k-clusters and evaluated them based on analysis of the sessions belonging to each cluster. We found that sessions within their clusters are similar and that clusters reflect our assumptions of initial intentions. Figure 3 shows 5 clusters based on average histograms of the composite sessions.

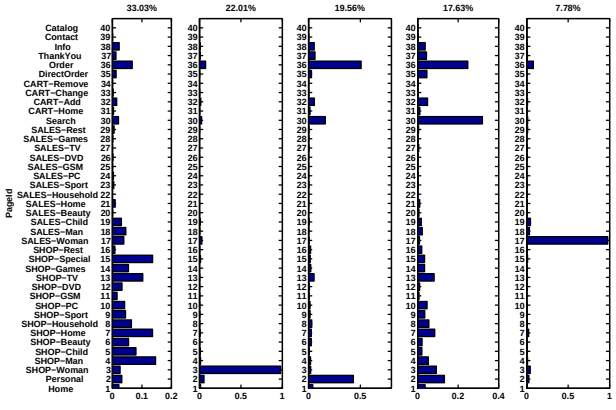

Figure 3: 5 clusters from composite sessions based on the similarity measure

We compared our results to the results of a frequency-based clustering model (our implementation is based on [4])

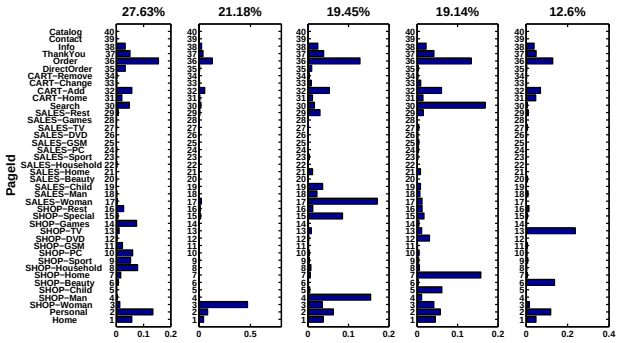

Figure 4: 5 frequency clusters based on [4] for comparison

(Figure 4). In contrast to the frequency-based model, our results show two "focused" groups with interests only in single categories (women's fashion and sales of women's fashion). In the results using the other model, these categories were also dominant within their clusters but they were also mixed together with other categories – which probably had short page view times. Our model clustered men's fashion together with other categories, such as household appliances and televisions, while the frequency-based model formed a focused, separate cluster for this category.

7. CONCLUSION

In this paper we investigated possible reasons for the relative disregard of a presumably relevant interest indicator in WUM, the time spent on pages. We gave an extensive overview of the literature concerning TSP in different fields. We outlined the most likely influential factors of the TSP measure in comparison with the more widely applied frequency measure. We described the problems and gave a methodology for TSP preprocessing. We gave an overview of the statistical properties of page view time for several real-world data sets. Lastly, we defined a similarity measure based on the combination of frequency and TSP measures and evaluated it by clustering retail web shop data.

Both frequency and TSP measures are influenced by several factors. While frequency seems to be a more solid, "plug-and-play" measure, the strong influence of distraction and hardware performance makes TSP more vulnerable. However, we believe that after careful preprocessing of web data, the TSP measure is of great value in identifying user intention. Special attention should be paid to robot filtering and session and page identification. The effects of network and server overhead should be normalised. TSP values that possibly comprise distraction factors should be identified and replaced by normalised values. Thresholds should be chosen with special attention to the application domain. For many applications, the combination of TSP and frequency measures can be the optimal choice.

Furthermore, we presented a similarity-based clustering that grouped sessions according to assumptions of users' intentions.

8. REFERENCES

[1] Malik Agyemang, Ken Barker, and Reda Alhajj. Framework for mining web content outliers. In *SAC '04: Proceedings of the 2004 ACM symposium on Applied computing*, pages 590–594, New York, NY, USA, 2004. ACM Press.

[2] Richard Atterer, Monika Wnuk, and Albrecht Schmidt. Knowing the user's every move: user activity tracking for website usability evaluation and implicit interaction. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 203–212, New York, NY, USA, 2006. ACM Press.

[3] A. Banerjee and J. Ghosh. Clickstream clustering using weighted longest common subsequences. In *Proceedings of the Web Mining Workshop at the 1st SIAM Conference on Data Mining*, 2001.

[4] I. V. Cadez, P. Smyth, E. Ip, and H. Mannila. Predictive profiles for transaction data using finite mixture models. In *Technical Report UCI-ICS*, pages 01–67, 2001.

[5] Mark Claypool, Phong Le, Makoto Wased, and David Brown. Implicit interest indicators. In *IUI '01: Proceedings of the 6th international conference on Intelligent user interfaces*, pages 33–40, New York, NY, USA, 2001. ACM Press.

[6] Robert Cooley, Bamshad Mobasher, and Jaideep Srivastava. Data preparation for mining world wide web browsing patterns. *Knowledge and Information Systems*, 1(1):5–32, 1999.

[7] Chen Ding, Chi-Hung Chi, and Tiejian Luo. An improved usage-based ranking. In *WAIM '02: Proceedings of the Third International Conference on Advances in Web-Age Information Management*, pages 346–353, London, UK, 2002. Springer-Verlag.

[8] Rosta Farzan and Peter Brusilovsky. Social navigation support in e-learning: What are real footprints? In *Proceedings of IJCAI'05 Workshop on Intelligent Techniques for Web Personalization*, pages 49–56, 2005.

[9] Yongjian Fu, Mario Creado, and Chunhua Ju. Reorganizing web sites based on user access patterns. In *CIKM '01: Proceedings of the tenth international conference on Information and knowledge management*, pages 583–585, New York, NY, USA, 2001. ACM Press.

[10] S. Gunduz and M. Ozsu. A web page prediction model based on click-stream tree representation of user behavior. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 535–540, 2003.

[11] Jeffrey Heer and Ed H. Chi. Separating the swarm: categorization methods for user sessions on the web. In *CHI '02: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 243–250, New York, NY, USA, 2002. ACM Press.

[12] Melanie Kellar, Carolyn Watters, Jack Duffy, and Michael Shepherd. Effect of task on time spent reading as an implicit measure of interest. *ASIST 2004*, pages 168–175, 2004.

[13] Diane Kelly and Jaime Teevan. Implicit feedback for inferring user preference: a bibliography. *SIGIR Forum*, 37(2):18–28, 2003.

[14] Joseph A. Konstan, Bradley N. Miller, David Maltz, Jonathan L. Herlocker, Lee R. Gordon, and John Riedl. GroupLens: Applying collaborative filtering to Usenet news. *Communications of the ACM*, 40(3):77–87, 1997.

[15] Bamshad Mobasher, Robert Cooley, and Jaideep Srivastava. Automatic personalization based on Web usage mining. *Communications of the ACM*, 43(8):142–151, 2000.

[16] Masahiro Morita and Yoichi Shinoda. Information filtering based on user behavior analysis and best match text retrieval. In *Proceedings of the 17th annual international ACM SIGIR conference (SIGIR '94)*, pages 272–281, New York, NY, USA, 1994.

[17] Bing Pan, Helene Hembrooke, Geri Gay, Laura A. Granka, Matthew K. Feusner, and Jill K. Newman. The determinants of web page viewing behavior: an eye-tracking study. In *ETRA*, pages 147–154, 2004.

[18] Rachael Rafter and Barry Smyth. Passive profiling from server logs in an online recruitment environment. In *IJCAI Workshop on Intelligent Techniques for Web Personalization (ITWP 2001)*, pages 35–41, 2001.

[19] C. Shahabi, A. M. Zarkesh, J. Adibi, and V. Shah. Knowledge discovery from users web-page navigation. In *Proceedings of the 7th International Workshop on Research Issues in Data Engineering (RIDE '97)*, page 20, Birmingham, England, 1997.

[20] Ramakrishnan Srikant and Yinghui Yang. Mining web logs to improve website organization. In *WWW '01: Proceedings of the 10th international conference on World Wide Web*, pages 430–437, New York, NY, USA, 2001. ACM Press.

[21] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, and Pang-Ning Tan. Web usage mining: discovery and applications of usage patterns from web data. *SIGKDD Explor. Newsl.*, 1(2):12–23, 2000.

[22] Ryen W. White, Ian Ruthven, and Joemon M. Jose. Finding relevant documents using top ranking sentences: an evaluation of two alternative schemes. In *SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 57–64, New York, NY, USA, 2002. ACM Press.

[23] Jitian Xiao, Yanchun Zhang, Xiaohua Jia, and Tianzhu Li. Measuring similarity of interests for clustering web-users. In *ADC '01: Proceedings of the 12th Australasian conference on Database technologies*, pages 107–114, Washington, DC, USA, 2001.

[24] Yunjuan Xie and Vir V. Phoha. Web user clustering from access log using belief function. In *K-CAP '01: Proceedings of the 1st international conference on Knowledge capture*, pages 202–208, New York, NY, USA, 2001. ACM Press.

[25] Dongshan Xing and Junyi Shen. Efficient data mining for web navigation patterns. *Information & Software Technology*, 46(1):55–63, 2004.

[26] Tak Woon Yan, Matthew Jacobsen, Hector Garcia-Molina, and Umeshwar Dayal. From user access patterns to dynamic hypertext linking. In *Proceedings of the fifth international World Wide Web conference on Computer networks and ISDN systems*, pages 1007–1014, Amsterdam, The Netherlands, 1996.

APPENDIX

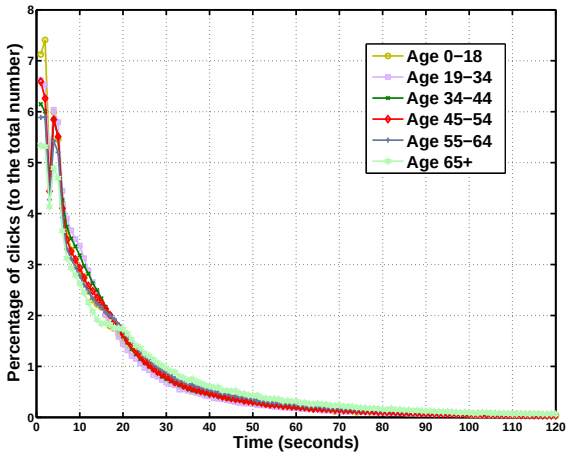

Figure 5: TSP distributions over different age groups

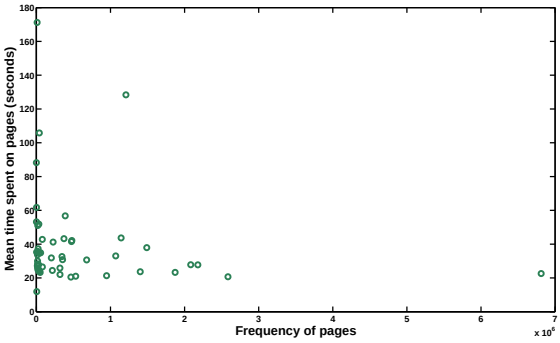

Figure 6: TSP vs frequency (over all page categories)