Information Processing and Management 45 (2009) 230–245

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Time series analysis of a Web search engine transaction log

Ying Zhang [a], Bernard J. Jansen [b,*], Amanda Spink [c]

[a] The Harold and Inge Marcus Department of Industrial and Manufacturing Engineering, College of Engineering, The Pennsylvania State University, University Park, PA 16802, United States
[b] 329F Information Sciences and Technology Building, College of Information Sciences and Technology, The Pennsylvania State University, University Park, PA 16802, United States
[c] Faculty of Information Technology, Queensland University of Technology, Gardens Point Campus, GPO Box 2434, Brisbane, QLD 4001, Australia

ARTICLE INFO

Article history:
Received 21 October 2007
Received in revised form 17 June 2008
Accepted 16 July 2008
Available online 31 August 2008

Keywords:
ARIMA
Box–Jenkins model
Search engine
Time series analysis
Transactional log

ABSTRACT

In this paper, we use time series analysis to evaluate predictive scenarios using search engine transactional logs. Our goal is to develop models for the analysis of searchers' behaviors over time and investigate if time series analysis is a valid method for predicting relationships between searcher actions. Time series analysis is a method often used to understand the underlying characteristics of temporal data in order to make forecasts. In this study, we used a Web search engine transactional log and time series analysis to investigate users' actions. We conducted our analysis in two phases. In the initial phase, we employed a basic analysis and found that 10% of searchers clicked on sponsored links. However, from 22:00 to 24:00, searchers almost exclusively clicked on the organic links, with almost no clicks on sponsored links. In the second and more extensive phase, we used a one-step prediction time series analysis method along with a transfer function method. The period rarely affects navigational and transactional queries, while rates for transactional queries vary during different periods. Our results show that the average length of a searcher session is approximately 2.9 interactions and that this average is consistent across time periods. Most importantly, our findings shows that searchers who submit the shortest queries (i.e., in number of terms) click on highest ranked results. We discuss implications, including predictive value, and future research.

1. Introduction

Search engines, such as Ask, Google, Microsoft Live, and Yahoo!, use Internet spiders to index Web pages that users can search (Chau, Zeng, & Chen, 2003). However, different search engines have different searcher interfaces, interpret queries in different ways (Beitzel, Jensen, Lewis, Chowdhury, & Frieder, 2007), support different types of advanced search functionalities, and employ different search algorithms (Chu, Fang, Olivia, & Liu, 2005). In order to identify and build better Web search engines, we must develop better ways to understand searchers' behaviors.

To analyze searchers' behaviors, some researchers use specific software on searchers' computers to investigate searching interactions (Fenstermacher & Ginsburg, 2003; Montgomery & Faloutsos, 2001). More recently, other researchers have used laboratory studies or client-side monitoring techniques (e.g., Hotchkiss, 2004). However, these methods typically involve subjective elements, small number of participants, and have large variance in the results due to sampling bias.

Using the large amounts of searcher behavior information now available on Websites and search engines can address these shortcomings. For example, search engine transactional logs contain significant amounts of useful information concerning Web search engine customers as well as customers of other commercial organizations.

* Corresponding author. Tel.: +1 814 865 6459.
E-mail addresses: yzz114@psu.edu (Y. Zhang), jjansen@ist.psu.edu (B.J. Jansen), ah.spink@qut.edu.au (A. Spink).

0306-4573/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2008.07.003

A search engine's transactional log is an electronic record of interactions that have occurred during a searching episode between a Web search engine and searchers who are seeking information on that Web search engine (Jansen, 2006, p. 408). Generally, transactional logs contain data such as the client computer's Internet Protocol (IP) address, submission time, searcher query, and vertical searched, among other fields. These huge datasets of recorded Web searchers' behaviors could contain information of significant commercial value permitting companies to detect the trends of their customers. It can be also useful to leverage the log file analytical results to fine tune the setting of search engine searching parameters. Efforts in pattern detection of customers, in combination with statistical analysis, may yield correlations between seemingly unrelated events. This potential motivates our research.

Alternatively, the server-side search engine transactional log data, like that used in this study, provides objective criteria and the quantity of data needed to analyze searchers' behaviors. However, one needs to subject these logs to rigorous data analysis in order to make sense of them. Standard transaction log analysis provides descriptive characteristics of specific interactions (Jansen, 2006); however, search logs are temporal streams of information from which one can possibly make predictive declarations. Therefore, we use time series analysis, which is good at detecting dynamic patterns in large datasets within a particular time range, to develop a predictive model for Web searching.

Given that transaction logs record the interactions between searcher and search engine, they are an excellent record of customer behaviors and are therefore very applicable to the research aim of detecting customer trends. We use a transaction log from the search engine Dogpile (www.dogpile.com), which is a top 10 ranked Web search engine. Jansen and Spink (2005) have shown that searcher behavior across search engines is very similar, so we believe that the methodology and results of this research will be applicable to a wide range of search engines.

In the next sections, we first review related studies on the use of Web transactional logs to investigate searcher behavior and the methodology employed. Then, we present our results using transaction log analysis (TLA) reporting on query length, frequency of log on records, and online behaviors of different searchers. After that, we use Autoregressive Integrated Moving Average (ARIMA) time series analysis method to do a one-period-ahead prediction on the log data. We then explore relationships between different fields in the dataset using the Box–Jenkins model approach. Finally, we provide comprehensive conclusions and discussion of findings.

2. Literature review

Companies interested in Web searcher behavior face the daunting task of scrutinizing enormous amounts of data. For example, Nielsen/NetRatings measures the search behavior of approximately 500,000 people worldwide (Sullivan, 2006), and transitional logs of this size are significant challenges.

Some researchers have investigated TLA as an approach for examining both server and client-side interactions (e.g., Park, Bae, & Lee, 2005; Wang, Berry, & Yang, 2003). For example, Wang et al. report on the use of a search engine for an academic Website, reporting statistics similar to general purpose search engines except for term usage. Park et al. (2005) report results from research that examines characteristics and interactions with a Korean-based search engine from the perspectives of session length, query length, query complexity, and content viewed among the Web search engines. Results were similar to other search engines, with allowances for searching in Hangul. However, this line of research is primarily descriptive. Exploring different methodologies that are more inferential, such as developing predictive models using automated data analysis techniques, could take TLA to a more meaningful level. Prior work has already reported some of this type of research.

Heckerman and Horvitz (1998) describe a Bayesian approach to modeling the relationship between words in a searcher's query for assistance and the information goals of the searcher. After reviewing the general method, the researchers describe several extensions that center on integrating additional distinctions and structure about language usage and searcher goals into the Bayesian models.

Özmutlu, Spink, and Özmutlu (2002) analyze contextual information in search engine query logs to develop a topic identification algorithm using artificial neural networks. A sample from the Excite data log was selected to train the neural network, and then the neural network was used to identify topic changes in the data log. The researchers report that topics shifts were estimated correctly, with a 77.8% precision in the overall database.

Özmutlu, Spink, and Özmutlu (2004) report results from a comprehensive, time-based Web study of US-based Excite and Norwegian-based Fast Web search logs, exploring variations in searcher behavior related to time. Their findings indicate that the analysis of datasets is very useful to Web search engines in reconstructing the search structure and reallocating the resources with respect to times.

Beitzel, Jensen, Chowdhury, Grossman, and Frieder (2004) review a query log of hundreds of millions of queries that constitute the total query traffic of a general purpose commercial Web search service. The researchers show that query traffic from particular topical categories differs both from the query stream as a whole and from other categories. This analysis provides valuable insight for improving retrieval effectiveness and efficiency. It is also relevant to the development of enhanced query disambiguation, routing, and caching algorithms.

The research from Beitzel et al. (2004) offer valuable insights on the temporal aspects of the data, especially in regards to automatic indexing. Automatic indexing models assume the uniformity of query characteristics; however, this approach might work less effectively in cases where the query characteristics vary with respect to time. It could be beneficial for the development of a new indexing method where the prioritization of the results is altered with respect to when a query is submitted.

Yates, Benavides, and Gonźalez (2006) present a framework for using query logs to identify searcher interest. The researchers found that, given certain established goals and categories, supervised learning could identify searcher interest. However, with unsupervised learning, one can validate the goals and categories used, refine them, and then select the most appropriate for the searcher's needs.

Although these approaches provide greater insight into the searching characteristics of Web searchers, they still focus largely on a particular slice of the log, offering no predictive value. Our research furthers this line of inquiry by examining the temporal trends inherent in search logs.

3. Research objectives

Our research objectives follow:

1. Verify that time series analysis, based on search engine logs, is a viable method for investigating Web searching behavior. If we can detect the dynamic behaviors of searchers through a large amount of online activities, we can force the response into the designed target by changing the Web searching system design factors. Therefore, statistical time series analysis may be a useful technique to analyze transaction logs, provide predictions on searcher behavior, and give guidance for examining searchers' interactions. Time series analysis could extend existing research to find the relationship among the different aspects of searchers' behaviors.
2. Identify characteristics of trends in Web search engine usage based on attributes recorded in search engine logs. Time-based transactional log analysis could be an efficient way to understand searchers' behaviors. Investigating patterns of Web searchers' behaviors based on a time series analysis method could provide worthy results on searchers' daily searching behaviors, which can promote the efficiency of search engines advertisement and information retrieval functions. Current studies do not clearly describe how to use time series analysis methods to predict the internal relationships among different fields of information.
3. Isolate predictive relationships among Web search engine usage attributes recorded in search engine logs. Given that transaction log data is temporal in nature, time series analysis can isolate trends that occur within the data across set periods. Such results can highlight critical or interesting descriptive data that analysis of the entire dataset or isolated sub-set could not. As stated, much of the prior work on transaction log analysis has been descriptive in nature. Times series analysis permits the development of predictive models among attributes in transaction logs.

4. Methodology

In this research, we use statistical time series analysis to examine and develop a prediction model using a Dogpile search engine transactional log. Our goal is to model and analyze searchers' behaviors on Web search engines over time. Additionally, we extend the time series analysis using the ideas of a transfer function and closed loop control theory to identify relationships among the different aspects of searchers' behaviors.

We first discuss the use of a proportional sampling method to reduce billions of records into over a thousand equidistance groups in order to ensure that varying group periods do not impact the analysis. Then, we conduct a TLA focusing on different fields of the dataset, including query length, frequency of submissions, online behaviors of different searchers, etc. After that, we use the ARIMA method to do a one-period-ahead prediction on the data and discover relationships among the different fields in the dataset using the Box–Jenkin transfer function model. Finally, we present results, implications for Web search engine design, and future research.

Time series analysis detects the internal structure (such as autocorrelation, trend, or seasonal variation) of data points taken over time that should be accounted for (Box & Jenkins, 1976). Time series analysis can be a useful model to predict future events based on known past events or to predict future data points before they are measured. A classic example of time series analysis is the opening price of a share of stock based on its past performance.

The basic requirement for the use of temporal data in a time series analysis is that the sequence of data points, measured typically at successive times, be spaced at uniform or equal time intervals. The essence of the analysis is to use the dynamic characteristics of the data to predict trends in the data stream. Many variations of time series analysis already exist, such as analysis of rhythmic variance, seasonal adjustment, and abnormal detection. Non-linear dependence on previous data points is also of interest because of the possibility of producing a chaotic time series.

Models for time series data can have many forms. Three broad classes of practical importance are the autoregressive (AR) models, the integrated (I) models, and the moving average (MA) models. These three classes depend linearly on previous data points (Box & Jenkins, 1976; Hosking, 1984; Lee, 1991; Tiao & Tsay, 1994). Autoregressive moving average models (ARMA) and autoregressive integrated moving average (ARIMA) are common combinations of these three models.

An ARIMA model can be expressed as $A_p(B)(1-B)^d Y_t = C_q(B)\varepsilon_t$, where p is the order of AR model; q is the order of MA order; and d is the degree of differencing needed to achieve a stationary process.

In manufacturing industry, time series analysis is used in the control area. Control is the engineering discipline that focuses on mathematical modeling of systems of a diverse nature and analyzing their dynamic behavior. Using control theory provides a reference to create a controller that will cause the systems to behave in a desired manner. Time series analysis is

the foundation for control theory by providing a method to detect the dynamic behavior among the data, which is indispensable to discovering the relationship between interested fields.

This exploration of relationships is also applicable for Web searching. To do this, we use transfer function theory. A transfer function describes the dynamic relationship, which is based on a linear, time-invariant, discrete difference equation. The transfer function can be expressed as $A_r(B)Y_t = B_s(B)B^kX_t$, where $A_r(B)$ is a polynomial in B of order r with first element equal to 1, and B_s (B) is a polynomial of order s with first element equal tob_0, and time k is called the input–output delay.

However, disturbances normally enter into the process and usually are not controllable, in contrast with X_t, which usually is controllable. To model the disturbances, Box and Jenkins (1976) proposed to add a noise term N_t that follows an ARIMA(p,d,q) of the transfer function model:

$$Y_t = \frac{B_s(B)}{A_r(B)}X_{t-k} + \frac{C_q(B)}{D_p(B)B^d}\varepsilon_t$$

One can code this model in statistical software packages, such as SAS and SPSS. In our study, we use the SAS software package version 9 to implement the ARIMA time series analysis.

5. Research design

5.1. Data collection

In this study, we used a Dogpile (www.dogpile.com) transaction log collected on 15 May 2006 and representing a snapshot of the searches executed on that server. We imported the transaction log into a relational database for initial pre-processing and cleaning (Jansen, 2006). Table 1 shows the fields included in this log.

Additionally, we calculated three extra attributes for each record, presented in Table 2.

Concerning the searcher intent field, Jansen, Booth, and Spink (2008) define and present a comprehensive classification and automatic identification of searcher intent for Web searching with an accuracy of 74%. They categorized searches based on intent in terms of the type of content specified by the query and other searcher expressions and operationalized these classifications with defining characteristics. Then, they implemented these categories in a program that automatically classified Web search engine queries. Using the same implementation approach, we also automatically classified the searcher intent of queries.

As reported in (Jansen et al., 2008), the algorithm used to classify each query was:

Algorithm: Web Query Classification based on User Intent

Assumptions:

1. Transaction log is sorted by IP address, cookie, and time (ascending order by time).
2. Search engine result page requested are removed.
3. Null queries are removed.
4. Queries are primarily English terms.

Table 1
Fields in Dogpile transactional log

Field	Description
Record identification number	The unique key to distinguish the records
IP address	The IP address recording the computer on which the searchers log on
Cookie	Parcels of text sent by a server to a Web browser and then sent back unchanged by the browser each time it accesses that server. Cookies are used for authenticating, tracking, and maintaining specific information about searchers, such as site preferences and the contents of their electronic shopping carts
Time	The time when the interaction was recorded by the search engine server
Query	The terms of the queries that the searcher typed into the search engine text box when searching for information
Vertical	There are five types of vertical (Web, Audio, Image, Video, News) representing different content collections. They are automatically divided by the search engines and provide a convenience for the searchers to find different information in different formats
Sponsored	One of two possible types of links retrieved and presented on the search engine results page (SERP). Sponsored links appear because a company, organization, or individual purchased the keywords that the searcher used in the search query. If the searcher entered a sponsored link, then this record will show 1
Organic	The other type of the link retrieved and presented on the SERP. The organic links are retrieved by a search engine using its proprietary matching algorithm. If the searcher entered an organic link, then this record will show 1
Browser	The browser used by the searchers
Location	The place/country where the searcher used the search engine

Table 2
Additional calculated fields in Dogpile transactional log

Field	Description
Searcher intent	There are three categories of searcher intent, which are *informational*, *transactional*, and *navigational*. They reflect the information type retrieved by the search engine. The calculation method is explained below
Query length	The number of terms contained in a particular query
Results pages	A number representing the SERP viewed (blank is first page, 1 is second page, etc.)

Input:
Record R_i with IP address (IP_i), cookies (K_i), query Q_i, source S_i, and query length QL_i.
Record R_{i+1} with IP address (IP_{i+1}), cookies (K_{i+1}), query Q_{i+1}, source S_{i+1}, and query length QL_{i+1}.
I: conditions of information query characteristics
N: conditions of navigational query characteristics
T: conditions of transactional query characteristics
Variable: B: Boolean//(if query matches conditions, 'yes' else 'no')
Output: Classification of User Intent, C

begin
Move to R_i (this module establishes the initial boundary condition)
Store values for IP*i*, K_i, Q_i, F_i, and QL_i
Compare (IP_i, K_i, Q_i, F_i, and QL_i) to *N*
If B then C = *N*
Elseif Compare (IP_i, K_i, Q_i, F_i, and QL_i) to *T*
If B then C = *T*
Elseif Compare (IP_i, K_i, Q_i, F_i, and QL_i) to *I*
If B then C = *I*
While not end of file
Move to R_{i+1}
Compare (IP_i, K_i, Q_i, F_i, and QL_i) to *N*
If B then C = *N*
Elseif Compare (IP_i, K_i, Q_i, F_i, and QL_i) to *T*
If B then C = *T*
Elseif Compare (IP_i, K_i, Q_i, $F_{i,\text{ and QL}i)\text{ to}}$ I
If B then C = *I*
(*Ri*+1 now becomes R_i)
Store values for $R_{i+1\text{ as IP}i,}$ $K_{i,}$ $Q_{i,}$ $S_{i,\text{ and QL}i}$
end loop

5.2. Sampling strategy

This processing left us with a dataset of 4,193,956 records, with 13 fields per record. However, even some of the most powerful statistical software packages (i.e., SPSS and SAS) could only handle approximately one million records. Therefore, we used a sample to represent the whole dataset. In prior work, researchers either used context-wise interpretation of data that requires manual analysis (Spink, Jansen, & Özmutlu, 2000) or designed a particular sampling strategy to represent the whole dataset effectively (Özmutlu et al., 2002). In our study, in order to ensure the accuracy of our results using partial data to reflect the characteristics of the whole dataset, we could not use randomly selected records. Instead, we selected the records whose unique record identification number's last digit was a one, which is a proportional method to get 10% of the whole data to make the analysis feasible. This method resulted in 419,395 records (10% of the original dataset) for analysis.

After getting the proportional sample, we used Matlab connected to a SQL server to do the statistical calculation based on the fields of information type, the rank of link clicked, the type of vertical, the type of the link (*sponsored* or *organic*), *query length* and the SERP number that the link clicked belongs.

5.3. Selection of time intervals

With time series analysis, we had to ensure that each period we were analyzing had an equal period so that the length of time slot would not affect the statistical data. In our study, we divided the 419,395 records into 1080 equidistance groups. The data are based on the 24-h daily transitional log (from 00:00 to 24:00), which means each time slot contained 80 s.

From this part of the data preparation, we could tell that only the *average grouped query length* and *average rank* for each time slot were numeric and stationary and thus had the potential to find their initial relationship. We did not use symbolic or

qualitative data, such as vertical or browser type, because they are not continuous data, even if we subjectively transferred them into numerical indicators. Moreover, we do not use the non-stationary data because we could not predict whether the data would drift away over time.

6. Results

6.1. Basic data analysis

6.1.1. Interactions with the search engine

Fig. 1 shows the number of interactions between the searchers and the search engine within each group used for the time analysis. Namely, the population flow goes up during the daytime and goes down during the nighttime. There were extreme data fluctuations that happened at about the 70th and 270th periods (i.e., a superabundant of submissions at these time spots). We regarded it as abnormal behaviors that could be caused by a number of factors (e.g., software agent submissions, start of work period, etc.).

Our results are similar to those in the time-based analysis of the Excite and AlltheWeb search engines studied by Özmutlu et al. (2004), and the decrease in the queries per session indicates that Web search engine searchers spend less time retrieving their information needs later in the day relative to earlier in the day.

6.1.2. Browser usage

We also analyzed the popularity of different browsers through calculating the number of browsers used among the time slots. From Fig. 2, we can tell that most of the people used the Internet Explorer (IE) browser, compared with Firefox, Mozilla, and other browsers. The rate of utilization is rarely affected by the time of day. The rate of usage remained at a stable level through the day.

6.1.3. Vertical accessed

Fig. 3 shows the results from the analysis of the vertical types assessed by the searchers. People use the Web vertical rather than image or audio verticals, and the rate of searching images and audios is not affected by time. Video and news searching was relatively almost non-existent.

6.1.4. Searcher intent

Fig. 4 shows the condition for searcher intent, namely which type of result searchers are looking for during different periods. There are three broad categories that cover most Web search queries (Jansen et al., 2008; Rose & Levinson, 2004), which include *informational* queries, *navigational* queries and *transactional* queries. Informational queries are searches that cover a broad type (e.g., *Colorado* or *trucks*) for which there may be millions of relevant results. Navigational queries are searches that seek a single Website or Web page of a single entity (e.g., *youtube* or *delta airlines*). Transactional queries are searches that reflect the intent of the searcher to perform a particular action like purchasing a car or downloading a screen saver.

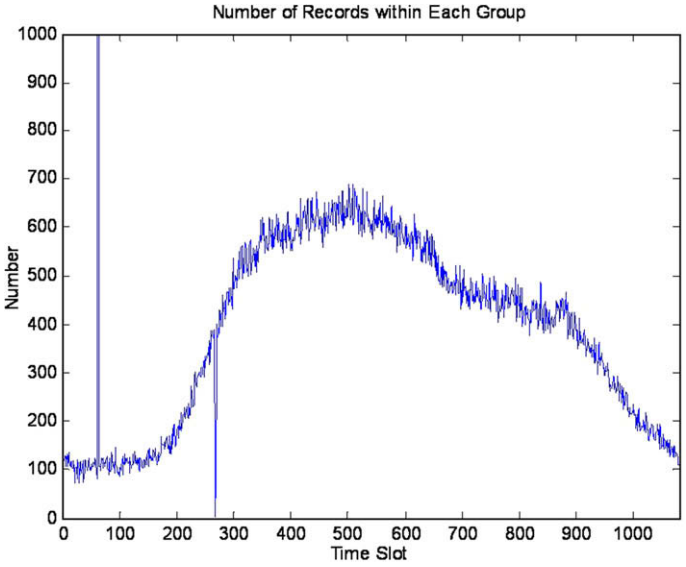

Fig. 1. Number of records within each group (population flow).

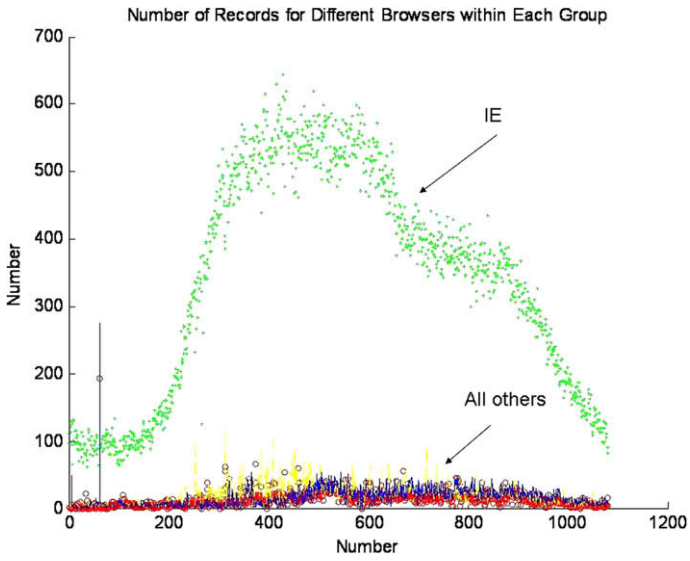

Fig. 2. Number of records for different browsers within each group (blue – Firefox, green – IE, red – Mozilla, yellow – other browsers). (For the interpretation of color in this figure legend, the reader is referred to the Web version of this article.)

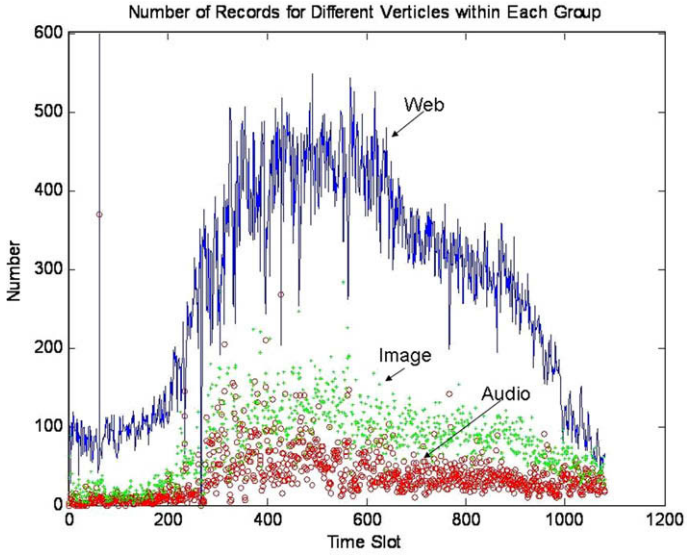

Fig. 3. Number of records for different verticals within each group (blue – Web, green – Image, red – Audio). (For the interpretation of color in this figure legend, the reader is referred to the Web version of this article.)

Most Web queries belong to the *informational* level, and *transactional* and *navigational* information are a relative small proportion within all the queries. In addition, we found *navigational* and *transactional* queries are rarely affected by the period, while rates for *transactional* queries vary during different periods.

6.1.5. Rank of clicked link

Figs. 5 and 6 present the average rank entered by the searchers based on the total number of records and the number of links clicked within each time group. To calculate the data about the number of opened Web pages, we gathered from all the records the rank of clicked links. To calculate the data about the average rank entered by the searchers, we summed up all the recorded rank of the clicked hyperlinks and divided it by the number of entered links within each time slot (i.e., the total number of records within each time slot). If we used the total number of records as the denominator, the average rank for the links entered was six. If we only considered the links that had been opened, the average rank for the links entered was approximately 10.

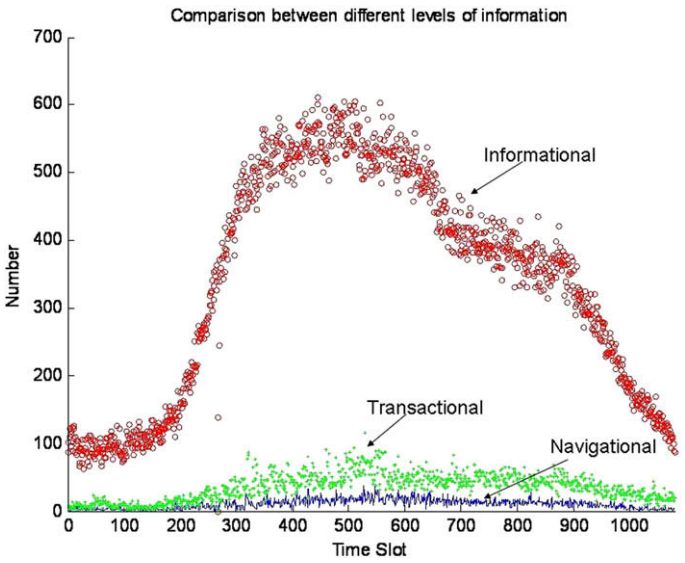

Fig. 4. Number of records for different levels within each group (blue – navigational, green – transactional, red – informational). (For the interpretation of color in this figure legend, the reader is referred to the Web version of this article.)

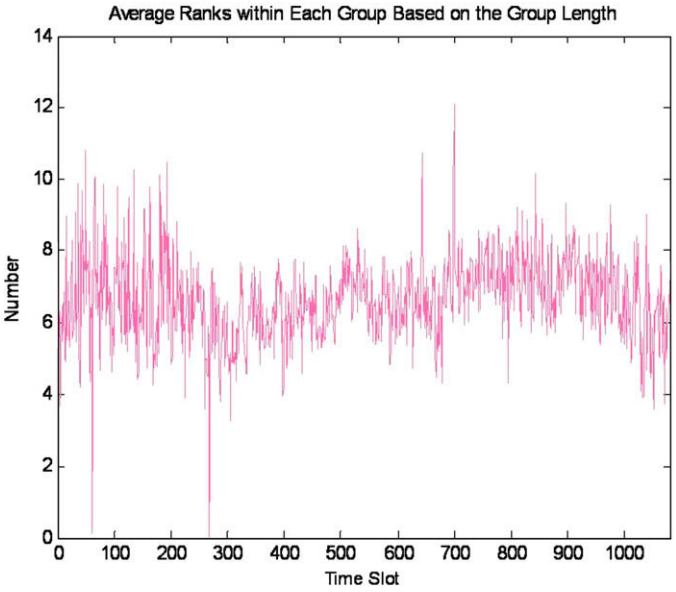

Fig. 5. Average rank based on the group length.

This result shows that people generally just consider the information shown on the first SERP and usually spend some time to browse many of links listed on the first SERP. In addition, due to little fluctuation over the whole time line, we could tell that searchers' behavior on clicking the links with different rank is not affected by time.

We analyzed the navigation of result pages, which is shown in Fig. 7, and derived similar findings as from the rank analysis, namely that most people simply look at the first page and do not view SERP after the second page. Moreover, we could tell that this trait was constant through the whole time line because the data does not go through dynamic vicissitudes during the different periods.

6.1.6. Query length

Fig. 8 shows that the average length of query per time slot is about 2.9 terms, and the length does not change with the changing of periods. Based on different datasets, the average length of query length reported by Özmutlu, Çavdur, Spink, and Özmutlu (2004) is 2.9, and the number reported by Spink, Jansen, Wolfram, and Saracevic (2002) is 2.4. The change in the

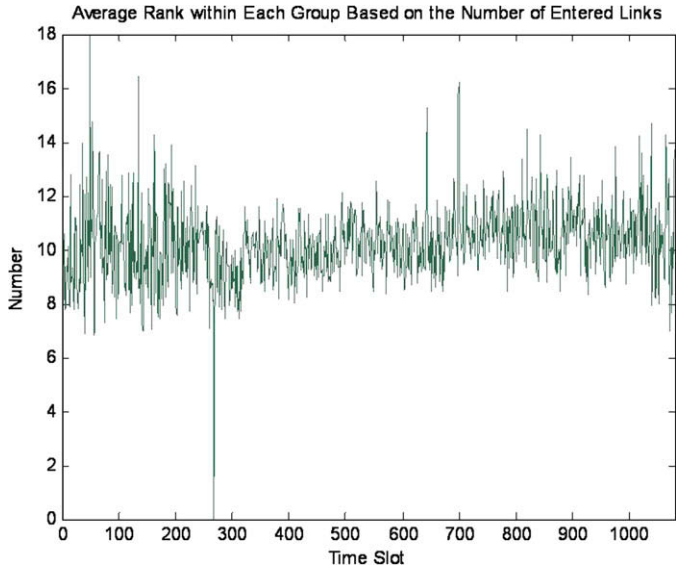

Fig. 6. Average rank based on the number of entered links.

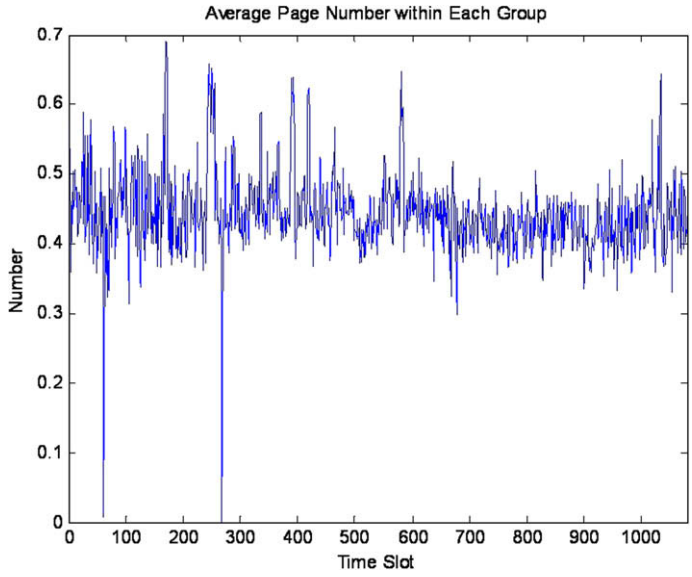

Fig. 7. Average page number within each group.

number of terms in the consecutive queries is also consistent from the beginning of the day to the end of the day, as reported in prior work (Özmutlu et al., 2004). However, we note extreme outliers at various timeslots.

6.1.7. Clicks on sponsored and organic links

Now, we look at the findings concerning the clicking on sponsored links and organic links. From the time series plots (see Figs. 9 and 10), we can see that 10% of searchers looked at sponsored links while 90% only looked at organic links. Interestingly, from 22:00 to 24:00, people nearly always clicked on organic links instead of sponsored links. During some periods, however, the click thru rate on sponsored links significantly increased, to more than 30%.

6.2. Extended data analysis (Predictive behavior of searchers)

Extending the basic data analysis, we used the time series analysis methodology to calculate one-step predictions as to what would happen from one time slot to the next. As discussed in Section 4, we used Matlab and SAS together to conduct

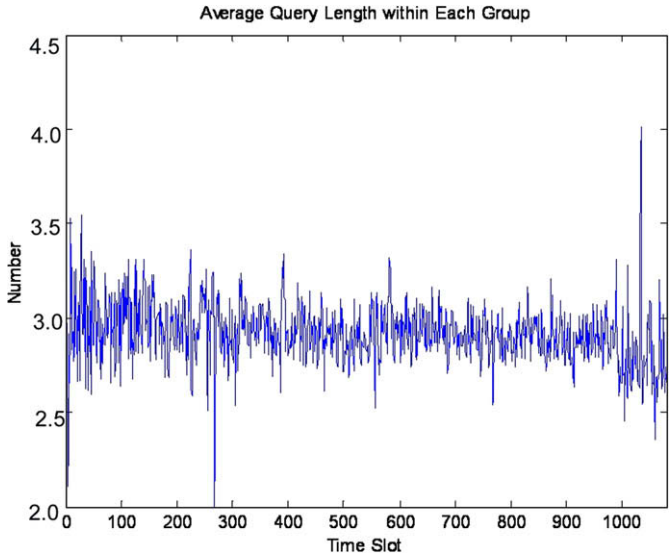

Fig. 8. Average query length within each time slot.

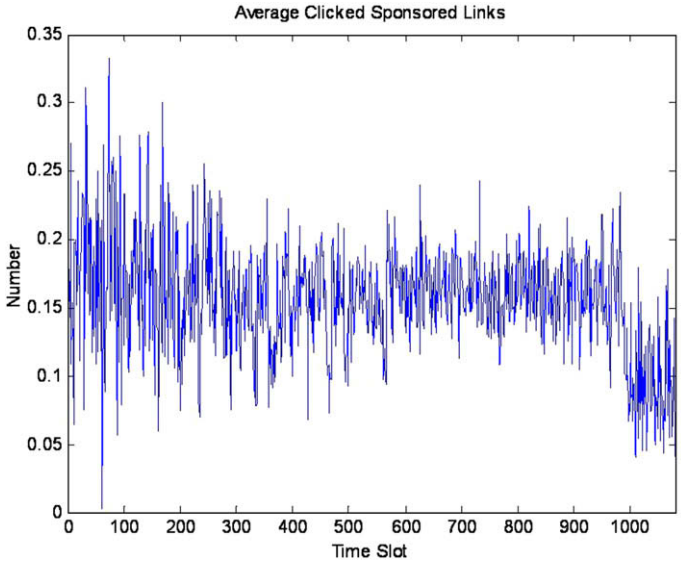

Fig. 9. Average clicked sponsored links.

this analysis. Our intent was to find a one-step prediction model based on the prepared search log data for various searching characteristics.

6.2.1. Keywords in the queries

From autocorrelation function (ACF) in Fig. 11 and partial autocorrelation function (PACF) in Fig. 12, as well as the later diagnostic tests about the autocorrelation existing in the residuals, we could tell that the model for grouped query length follows the ARIMA(1,2) time series model with AR model at lag 1 and MA model at lag 2. Using the minimum mean square error (MMSE) estimation, we get the final model, as shown in Eq. (1). Obviously, our model is simplified because of the elimination of the abnormal data:

$$Y_t = 2.36437 - 0.3967Y_{t-1} + \varepsilon_t - 0.0415\varepsilon_{t-1} + 0.0045\varepsilon_{t-2} \quad (1)$$

6.2.2. Average rank of clicked result

From Figs. 13 and 14 as well as from the later diagnostics, we could tell the model for average rank follows the AIMA (1, 1, 1) with one order of differencing, and AR model at lag 1 and MA model at lag 1($A_1(B)(1-B)^1 Y_t = C_1(B)\varepsilon_t$), time series model,

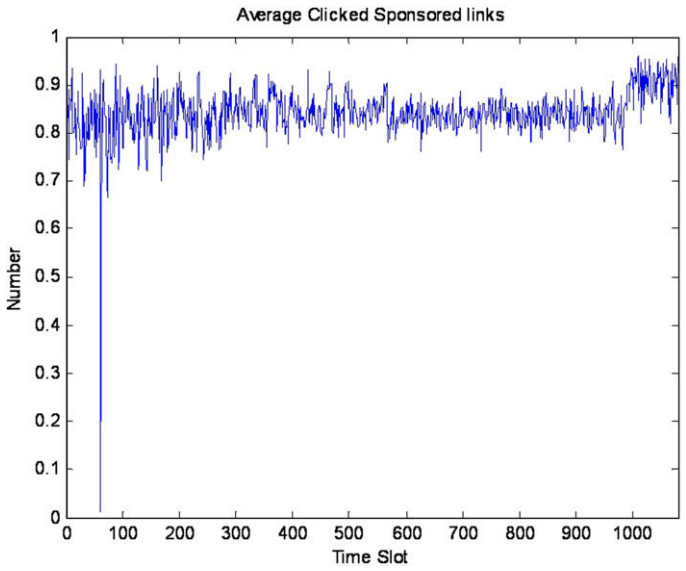

Fig. 10. Average clicked organic links.

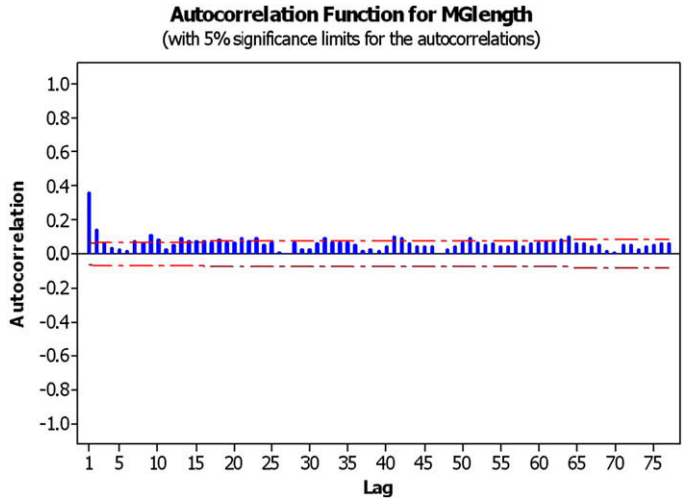

Fig. 11. Autocorrelation function (ACF) for the grouped keywords.

which has AR model with lag 1, integrate model with lag 1 and MA model with lag 1. Using the MMSE estimation, we get the final model as shown in

$$Y_t = -0.000132 + 1.2023Y_{t-1} - 0.2023Y_{t-2} + \varepsilon_t - 0.9539\varepsilon_{t-1} \quad (2)$$

6.2.3. Predictive model

The ARIMA analysis generally will have significant meaning if we find the relationship among different fields of data. From the basic analysis, we could tell that only average grouped query length (see Fig. 8) and average rank (see Fig. 7) are numeric and stationary, and thus have the potential to reveal their initial relationship. Therefore, we proposed to find the impact of these two fields on each other. Since Matlab assumes the expected value of all the data is zero when it copied with the transfer function, we first normalized the original data to ensure a zero mean (see Figs. 15 and 16).

After that, we used Matlab's ident toolbox, which is a module to do time series analysis, to find the transfer function. From V weight chart (Fig. 17), we could identify the value of r, k, and s according to various patterns based on impulse response function. We could tell that at lag 1, the v weight is significant, and it does not show damped sinusoid pattern (i.e., a sinusoid pattern would show a second-order dynamical process).

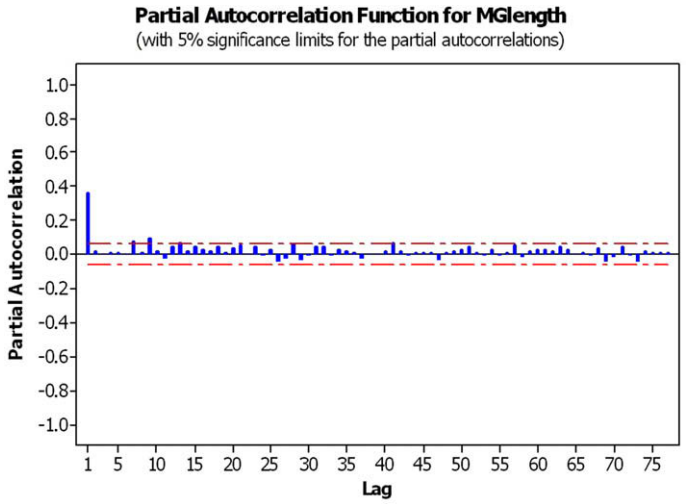

Fig. 12. Partial autocorrelation function (PACF) for the grouped keywords.

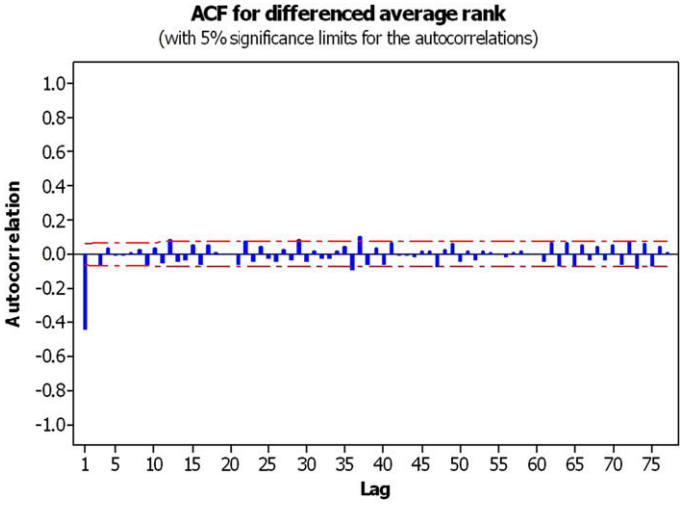

Fig. 13. ACF for the average rank.

Therefore, we fitted the transfer function model as $s = 0$, $r = 1$, and $k = 1$. After that, we looked at the autocorrelation of the residuals after fitting the transfer function model. We kept fitting the disturbance model until no dynamics showed in the residuals. The disturbance model we fit was ARMA (2,1). Finally, the diagnostic tests showed our model is well fitted (Fig. 18).

Fig. 19, which is a measure of the simulated output, shows that our model is well fitted and reflects the trends of the real data. Simulated output is the output calculated by the transfer function generated by the Matlab's ident toolbox as calculated by formula (3).

Now, we present the fitted model, which is

$$\tilde{y}(t) = \frac{-0.8093}{1 + 0.1845B}\tilde{x}(t-1) + \frac{1 - 0.9463B}{1 - 1.079B - 0.09298B^2}\varepsilon_t \tag{3}$$

From the above model, we see that the output (*average rank*) has a one-period-ahead relationship with the input (*average query length*). Namely, the searchers who typed in the fewest query terms one period ahead (i.e., less than the *average query length*) were more likely to click higher ranked links (i.e., top ranked results) in the following period.

This could explain some people's potential searching behaviors. More importantly, it shows the potential of the time series analysis approach to investigating and developing predictive models of other behaviors of Web searchers.

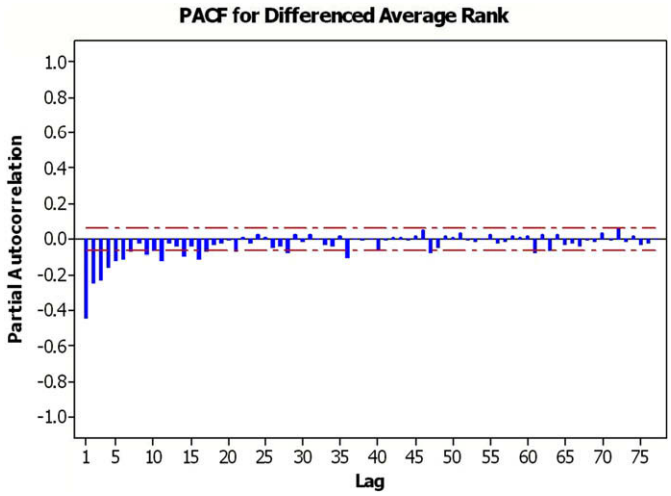

Fig. 14. PACF for the average rank.

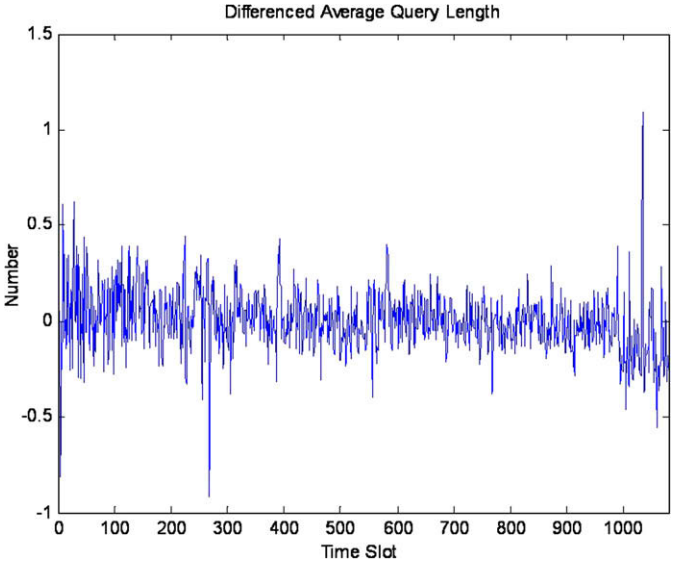

Fig. 15. Differenced average query length.

7. Discussion and implications

This study analyzed the transaction log from the Dogpile Web search engine in order to explore the characteristics of searchers' behaviors, validate the use of time series analysis for Web log analysis, and develop a predictive model of Web searchers interacting with a Web search engine. Our study is unique in that we used a large dataset, thereby achieving more reliable results than can be obtained by using a small sample. After basic and extended data analysis, we found results with potential commercial value and demonstrated that time series analysis is both worthwhile and feasible for detecting the trends of search engine customers.

For the basic analysis of this log, we found or confirmed that people use search engines more frequently in the daytime than they do very late at night. Given that the search engine server logs the time, this would imply a geographical concentration of searchers for the search engine within nearby time zones of the Web server.

We also found that people prefer to use the IE browser more than other browsers, and the proportion among different browsers is about 6:1:1:1 (IE:Firefox:Mozilla: other browsers). Therefore, despite the publicity concerning alternate browsers, IE continues to dominate the market. There is no variance in usage based on time. People prefer to search for information on the *Web* vertical, rather than via *Audio* or *Images*. The proportion is about 6:2:1 (Web:Image:Audio). Whether or not this finding reflects searcher needs goals or the fact that the Web vertical is the default is unknown, but we suspect a combination of the two. Again, there was no variance due to a temporal factor.

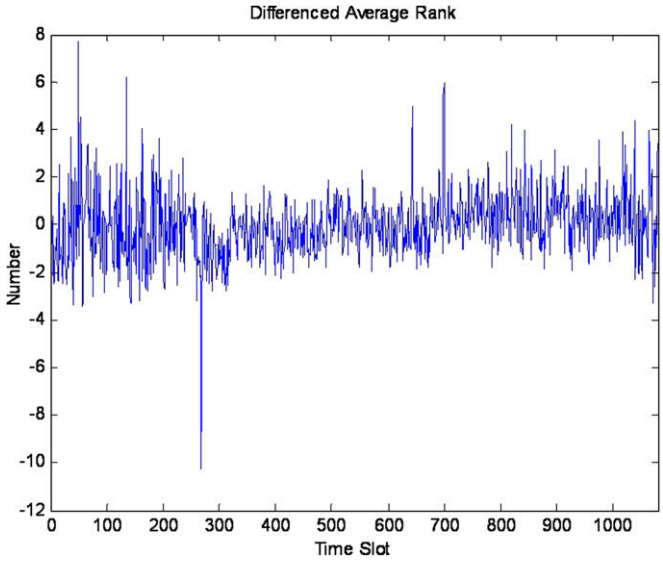

Fig. 16. Differenced average rank.

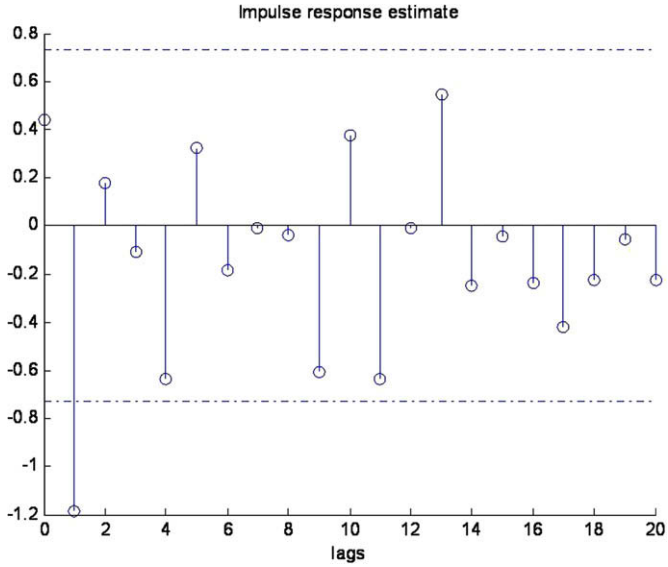

Fig. 17. v Weight chart of input (average query length).

People search for *informational* needs much more often than they do for *transactional* and *navigational* needs. The proportion among these three classifications of searcher intent is about 12:2:1 (*Informational*:*Transactional*:*Navigational*), with no variance due to time.

Ten percent of searchers look at sponsored links while 90% of searchers look only at organic links. From 22:00 to 24:00, people usually click on the organic links, ignoring the sponsored links, thereby indicating that this period may have reduced commercial value. Additionally, this result may point to a need to advertisers to incorporate time-based impressions of their advertising campaigns.

Based on our research results, Web search engine companies can design the information retrieval techniques more efficiently and arrange information storage spaces to support these behaviors.

For the extended analysis, we focused on the predictive aspects of two fields, which were average query length and average clicked ranks. We found that the average length of the query is about 2.9, and the query length does not change with the changing of time. If we used the total number of records as the denominator, the average rank for the links entered is 6, while if we only considered the links that searchers opened, the average rank for the links entered is about 10. Therefore, approximately 40% of the time, searchers entered queries but did not click on any links in the SERP. In addition, these two fields of data followed the MA (2) and ARIMA (1,1,1) models for the one-period-step-ahead prediction.

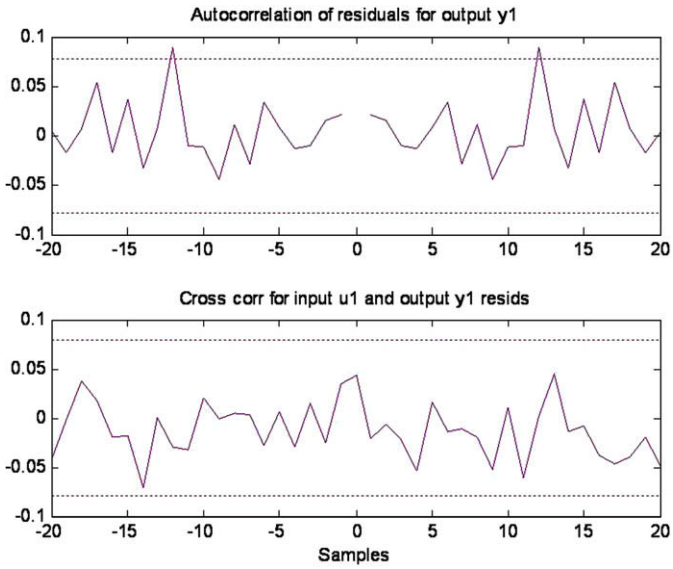

Fig. 18. Diagnostics of the transfer function.

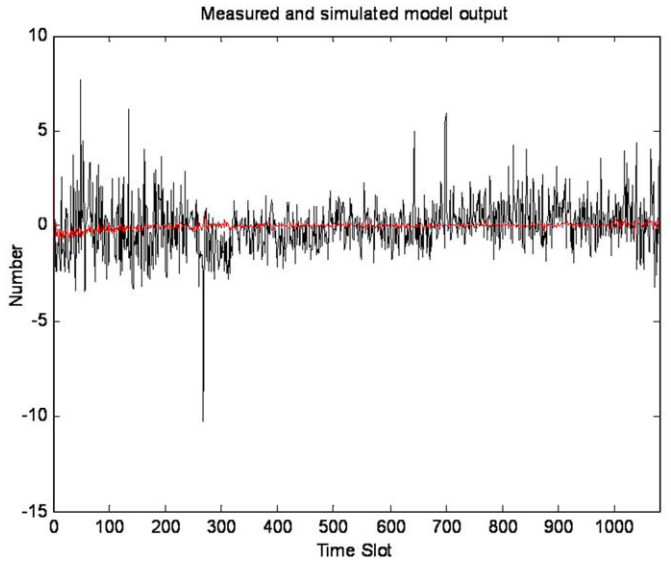

Fig. 19. Simulated model output.

The Box–Jenkin transfer function model shows that these two numeric fields of data have a close, predictive relationship between each other. Namely, if people in a given period typed in fewer query terms, then they were more likely to click on the top ranked results in the following period. This could be one indication of a particular searcher need, most probably *navigational* searcher intent. Jansen et al. (2008) found that searchers with *navigational* queries clicked on higher ranked results than did searchers with *informational* and *transactional* needs. With this predictive information, search engines could tailor results based on the characteristics of the searcher query.

Finally, the times series analysis approach, in combination with basic and advanced analysis of search logs, yields significant and insightful results that can impact the development of future Web search engines and can lead to greater understanding of Web searching.

8. Conclusion and future research

Although we have derived many interesting results from this dataset, our study is limited to the single system of the dataset. However, given that prior work has shown that Web searching is similar across search engines (Jansen & Spink, 2005),

we believe that our results are applicable to other search engines. The study is also limited by the period of data collection (i.e., one day), and therefore, we could not explore other interesting phenomena using the time series analysis technique due to data limitations. However, given the limited number of published works using time series analysis in the search log area, this study is valuable for both its results and methodology.

This research has several strengths and important implications. This study extends the existing work in TLA by incorporating a methodology that recognizes the temporal-basis of search logs. Based on our results, time series analysis appears to be a viable approach for predictive analysis of search logs. Used in conjunction with the Box–Jenkins transfer function model, one can identify the relationship among searching attributes across time periods, permitting search engines to generate probabilities of what content searchers desire and permitting system time to service this content. As an example, our results show that users who enter the shortest queries are more likely to click the top most ranked results.

In future studies, we would like to examine transaction logs that are of a longer period (i.e., more than one day) with cyclical trends instead of just having one day or one cycle of the data. We would also like to evaluate data using multiple sources, such as a search logs in conjunction with searcher studies or multiple logs across many search engines.

References

Beitzel, S. M., Jensen, E. C., Chowdhury, A., Grossman, D., & Frieder, O. (2004). In M. Sanderson, K. Järvelin, J. Allan, & P. Bruza (Eds.), *Hourly analysis of a very large topically categorized Web query log* (pp. 321–328). Paper presented at the 27th Annual international conference on research and development in information retrieval, Sheffield, UK, 25–29 July.

Beitzel, S. M., Jensen, E. C., Lewis, D. D., Chowdhury, A., & Frieder, O. (2007). Automatic classification of Web queries using very large unlabeled query logs. *ACM Transactions on Information Systems*, *25*(2), Article No. 9.

Box, G. E. P., & Jenkins, G. M. (1976). *Time series analysis forecasting and control* (2nd ed.). San Francisco: Holden-Day.

Chau, M., Zeng, D., & Chen, H. (2003). In E. A. Fox, & C. L. Borgman (Eds.), *Personalized and focused web spiders* (pp. 79–87). Paper presented at the 1st ACM/IEEE-CS joint conference on digital libraries, 24–28 June, Roanoke, VI. New York: ACM/IEEE.

Chu, M., Fang, X., Olivia, R., & Liu, S. (2005). Analysis of the query logs of a Website search engine. *Journal of the American Society for Information Science and Technology, 56*(13), 1363–1376.

Fenstermacher, K. D., & Ginsburg, M. (2003). Client-side monitoring for Web mining. *Journal of the American Society for Information Science and Technology, 54*(7), 625–637.

Heckerman, D., & Horvitz, E. (1998). In G. Cooper, & S. Moral (Eds.), *Inferring informational goals from free-text queries: A Bayesian approach* (pp. 230–237). Paper presented at the fourteenth conference on uncertainty in artificial intelligence, 24–26 July, Madison, WI, USA. Los Altos, CA: Morgan Kaufmann.

Hosking, J. R. M. (1984). Modeling persistence in hydrological time series using fractional differencing. *Water Resources Research, 20*(12), 1898–1908.

Hotchkiss, G. (2004). Inside the mind of the searcher. <http://www.enquiro.com/research.asp> Retrieved 15.03.05.

Jansen, B. J. (2006). Search log analysis: What is it; what's been done; how to do it. *Library and Information Science Research, 28*(3), 407–432.

Jansen, B. J., Booth, D., & Spink, A. (2008). Determining the informational, navigational, and transactional intent of Web queries. *Information Processing & Management, 44*(3), 1251–1266.

Jansen, B. J., & Spink, A. (2005). How are we searching the World Wide Web? A comparison of nine search engine transaction logs. *Information Processing & Management, 42*(1), 248–263.

Lee, S. (1991). Prediction of long-memory time series. Unpublished Dissertation, University of Chicago, Chicago, IL, USA.

Montgomery, A., & Faloutsos, C. (2001). Identifying Web browsing trends and patterns. *IEEE Computer, 34*(7), 94–95.

Özmutlu, H. C., Çavdur, F., Spink, A., & Özmutlu, S. (2004). In A. Grove (Ed.), *Neural network applications for automatic new topic identification on excite web search engine data logs* (Vol. 42, pp. 1–10). Paper presented at the 68th annual meeting of the American Society for Information Science and Technology (ASIST), Providence, RI, 12–17 November.

Özmutlu, H. C., Spink, A., & Özmutlu, S. (2002). Analysis of large data logs: An application of Poisson sampling on excite Web queries. *Information Processing & Management, 38*(4), 473–490.

Özmutlu, S., Spink, A., & Özmutlu, H. C. (2004). A day in the life of Web searching: An exploratory study. *Information Processing & Management, 40*(2), 319–345.

Park, S., Bae, H., & Lee, J. (2005). End user searching: A Web log analysis of NAVER, a Korean Web search engine. *Library & Information Science Research, 27*(2), 203–221.

Rose, D. E., & Levinson, D. (2004). In S. Feldman, M. Uretsky, M. Najork, & C. Wills (Eds.), *Understanding user goals in Web search* (pp. 13–19). Paper presented at the World Wide Web conference (WWW 2004), New York, USA, 17–22 May.

Spink, A., Jansen, B. J., & Özmutlu, H. C. (2000). Use of query reformulation and relevance feedback by Web users. *Internet Research – Electronic Networking Applications and Policy, 10*(4), 317–328.

Spink, A., Jansen, B. J., Wolfram, D., & Saracevic, T. (2002). From E-sex to E-commerce: Web search changes. *IEEE Computer, 35*(3), 107–111.

Sullivan, D. (2006). Nielsen/NetRatings search engine ratings, 23 February. <http://searchenginewatch.com/showPage.html?page=2156451> Retrieved 1.06.06.

Tiao, G. C., & Tsay, R. S. (1994). Some advances in non-linear and adaptive modeling in time series. *Journal of Forecasting, 13*, 109–131.

Wang, P., Berry, M., & Yang, Y. (2003). Mining longitudinal Web queries: Trends and patterns. *Journal of the American Society for Information Science and Technology, 54*(8), 743–758.

Yates, R. B., Benavides, L. C., & Gonźalez, C. (2006). The intention behind Web queries. In F. Crestani, P. Ferragina, & M. Sanderson (Eds.), *Lecture notes in computer science. String processing and information retrieval (SPIRE 2006)* (Vol. 4209/2006, pp. 98–109). Berlin: Springer.