The Pennsylvania State University

The Graduate School

Department of Computer Science and Engineering

TEMPORAL IDENTIFICATION OF USAGE PATTERNS AND OUTLIERS IN WEB SEARCHING USING TENSOR ANALYSIS

A Thesis in

Computer Science and Engineering
by

Chandrika Gopalakrishna

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science

May 2008

The thesis of Chandrika Gopalakrishna was reviewed and approved* by the following:

Bernard J. Jansen
Assistant Professor of Information Sciences and Technology
Thesis Advisor

Trent Jaeger
Associate Professor of Computer Science and Engineering

C. Lee Giles
David Reese Professor of Information Sciences and Technology
Affiliate Professor of Computer Science and Engineering

Mahmut Taylan Kandemir
Director of Graduate Affairs
Department of Computer Science and Engineering

*Signatures are on file in the Graduate School

ABSTRACT

This paper attempts to recognize patterns and outliers in the data stream from huge search engine transaction files incorporating tensor analysis. The aim is to analyze the correlation between different attributes of data recorded in a search engine transaction file. From this, one can study the trends in variation of attributes over a period among a set of selected search engine in order to summarize the online search activity. This thesis presents a proof-of-concept that tensor analysis is a valid methodology for mining search engine logs to study correlation of characteristics, identify patterns, and isolate outliers. One of the main challenges involved in analyzing search engine transaction logs is the huge volume of data that is continuously evolving with time, which tensor analysis resolves.

The experimental design consisted of two main scenarios aimed at studying trends and attribute correlation in five log files from well-known search engines. The trend analysis presents the variation of a set of attributes over a period of 24 hours. The correlation analysis detects two kinds of patterns occurring over this 24-hour period. One of these patterns is recognized as the normal or main trend, while the other as an abnormal trend that is deviating from this main trend. The results show that three of the four search attributes (Search Pattern, Number of Queries and Query length) are positively correlated with each other and anti-correlated with the fourth attribute (User Intent) in the main trend analysis. In the abnormal trend, first and third attributes (Search Pattern and Query length) are anti-correlated with other two attributes. This type of analysis allows us to identify the outliers as those log entries that contribute towards

occurrence of an abnormal pattern. A time window of high search engine usage was identified during a 24-hour period.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I wish to acknowledge and thank those people who contributed to this thesis:

Dr. Bernard J. Jansen, for his invaluable guidance and encouragement in carrying out the studies presented in this thesis. His supervision in different stages of this project such as the formulation of problem, discussions of results and preparation of this document have been highly valuable.

Dr. Trent Jaeger, for his friendly interface all through my Master's program.

Dr. Lee C. Giles, for having graced me with his consent to serve on my committee. The course I took with him has given me the gist of how search engines function and important aspects associated with information theory.

Dr. Jimeng Sun, for having generously offered his knowledge and guidance on Tensor analysis to carry out experiments. His expertise in Pattern Discovery on Streams is undoubtedly worthy of appreciation and I am obliged to have him verify our findings.

Betty Blair is another beautiful human being whose name has to be stated here. Her amiable nature and unending dynamic energy to help out in all possible ways during my stay at Penn State was comforting.

Vasant, Nari, Sanjeev and all my friends and collegues for their perennial reservoir of help and moral support.

My parents have been the pillars of my existence and have supported me to the utmost in everything that I have done. I would have been nothing but a heap of dust without the grace and blessings of my beloved parents. I owe every part of my existence to them. It is their firm conviction and dream that I am in a position to write this acknowledgement for my research that culminates into my Masters' Thesis.

Dr. Shivprakash Iyer, my husband, best friend, wonderful companion who has always been at my side for guidance. Without his support and love this work would not have been complete. His dynamic personality has set an example for me. His thrust for knowledge and intelligent thoughts made me see new avenues to problem solving. His remarkable patience and outlook to life has changed my attitude towards life as a person and a professional.

I also take this opportunity to thank my brother **Mukund** and sister **Prabha** and their families for their moral support. Their guidance in my educational journey was great importance.

This work is dedicated to Sai my divine Guru.

Chandrika Gopalakrishna

May 2008 The Pennsylvania State University

Chapter 1

Introduction and Problem Motivation

Information retrieval (IR) deals with the representation, storage, organization of, and access to information items. The representation and organization of the information items should provide the user with easy access to the information in which the user is interested. IR has changed considerably with the expansion of the Web (World Wide Web) and the advent of modern and inexpensive graphical user interfaces and mass storage devices. There are several definitions of IR but the one that pertains more to academic pursuits is by Christopher, Prabhakar and Hinrich [1].They defines IR as,

> " *finding materials (usually documents) of an unstructured nature (usually text) that satisfy information need from within large collections (usually stored on computers)"*

Past few decades witnessed a rapid pile up of databases in response to inexpensive and efficient storage devices. Hence, the process of retrieving relevant materials involves significant effort [2]. In general, most IR systems adopt either Boolean matching or probabilistic methods, with most Web systems using some version of page ranking techniques

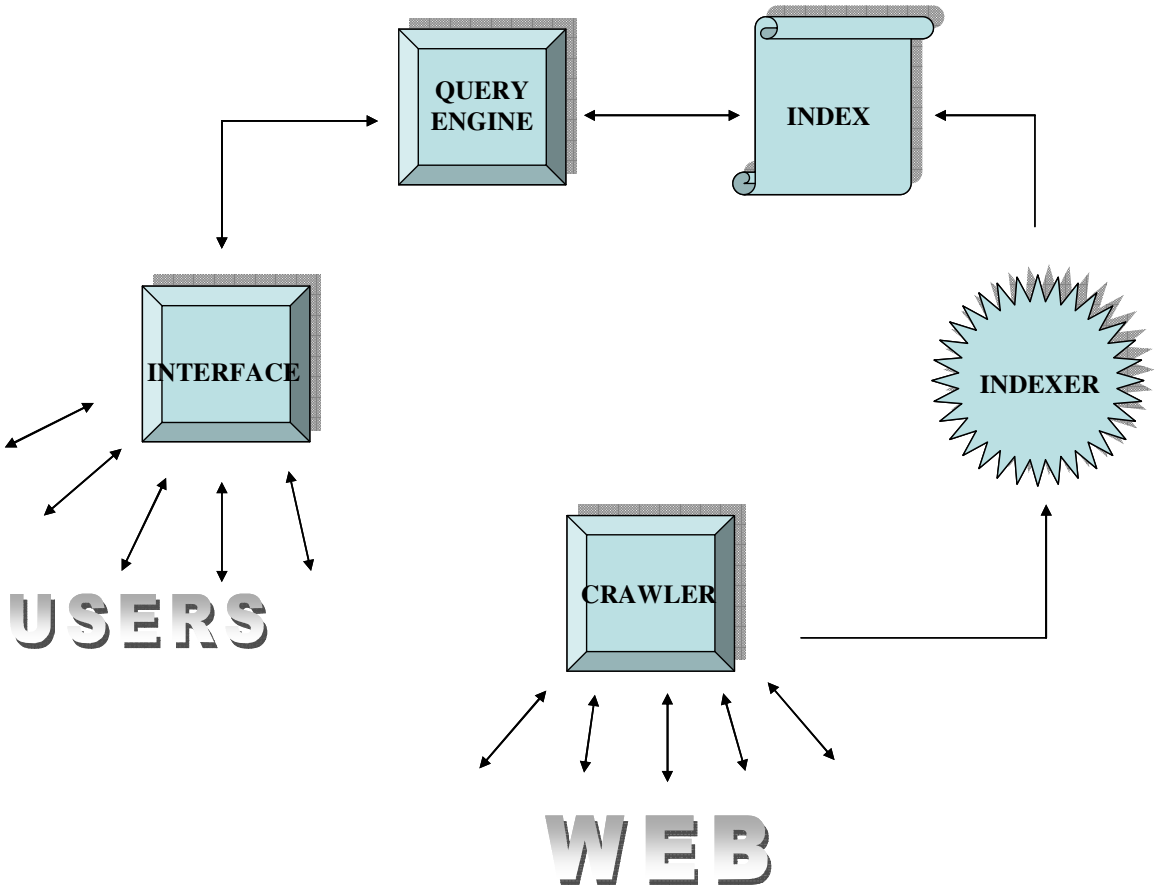


Figure 1-1: **Pictorial representation of a typical search engine**

.

1.1 Search Engine as Web IR System

The Web is a dominate information source as the primary target for informational and other needs [3] [4]. According to Nielsen Media, Web search engines are extensively used to access information on Web. It is evident from Nielsen Media data that 71 percent

of the Web users select search engines to access Websites of their interest [5] (Commerce Net/Nielsen Media, 1997). An outline of a typical search engine structure is as shown in Figure 1-1.

The major components of a typical search engine are Interface, Query Engine, Index, Indexer and Crawler. The functionalities of these components are listed below:

1. The Interface component, as the name suggests, provides an interface between the user and the search engine. The input to the Interface component is a query string (search terms). The output comprises of the retrieved documents in response to an entered query.
2. Query Engine takes the query string entered by the user and parses it and passes it on to the indexer. The top documents are matched corresponding to the query entered.
3. Index generated by the indexer is changed depending on the algorithm incorporated for matching documents corresponding to the query entered by the user.
4. Indexer automatically generates the index for the documents collected by the Web crawler.
5. Crawler is the dynamic part of the search engine, which browses the Web, collecting URLs on its path. It indexes the keywords and text of each Web page it encounters. It is sometimes referred to as a spider or a robot.

1.2 Importance of Search Engine Transaction Logs

The widespread use of search engines has made the search engine logs a valuable resource for providing insights into the interaction between users and systems. These log files are build incrementally and continuously, with new interactions recorded at every clock tick on almost every single search engine server. Modeling and analyzing this temporal search data requires novel methods to extract patterns of user search behavior or identify possible outliers or abnormal behavior. Transaction log analysis (TLA), as it is popularly known in the field of information retrieval, leads to better understanding of Web search behavior and redesigning the existing engines to suit user's need [6].

1.3 Nature of Search Engine Transaction Logs

Search engine log data is multi-aspect in that each data entered has multiple attributes. On a broad scenario, at every instant someone out there is using a search engine for his/her informational need. Thus, a record can be logged onto the search engine server that maintains attributes of the search in a systematic way. Each interaction with a search engine creates a record in the log file.. A single log file from a search engine server itself is two dimensional in nature, with list of attributes forming one dimension and time as the other. Figure 1-2 provides a snapshot of a typical search engine transaction log.

SEARCH ENGINE TRANSACTION LOG : Table

QUERY IDENTIFIER	USER IDENTIFIER	COOKIE	TIME STAMP	QUERY STRING	VERTICAL	QUERY LENGTH	LEVEL ONE USER INTENT
533322	12.104.119.239	215PNM9A4BW8SPW	11:38:57	pitchers of dodge viper	Images	4	informational
1450174	12.149.13.1	S4WJ47A3UOD1Y9	09:10:18	basel ii	Web	2	informational
2795904	12.156.151.226	2W5KSM6A4BOJZPW	10:54:02	pba architects	Web	2	informational
2656434	12.160.149.10	SF964FA4BL4IPW	18:22:37	hersheys chocolate	Web	2	informational
2010420	12.183.100.194	TP1JF6A4D74PHI	09:32:15	helicopters for sale	Web	3	informational
2176906	12.214.253.116	262TS2WA49045GD	11:26:18	temporal bone	Web	2	informational
2902648	12.221.153.62	22G13SA4BRJF8V	11:20:22	madonna	Images	1	informational
1385976	12.227.128.114	JK7I4NA3TYFVGG	16:00:34	dalene kurtis	Images	2	informational
1275755	12.25.176.211	L091FMA35F3PG7	17:46:53	basic lighting calculations	Web	3	informational
447487	12.32.58.227	C7O9V5A4BTD6ZS	00:34:12	kathryn sullivan	Images	2	informational
3726049	128.146.58.164	PHDRY4A4BW3PPK	16:40:02	ahdmah	Web	1	informational
430469	129.44.21.43	BUP1ZIA4BUEEHL	19:15:09	chloroform damsels	Images	2	informational
735384	131.9.254.125	BUP1ZIA4BUEEHL	11:16:11	american idol	Web	2	informational
2611377	137.197.119.15	2277JL9A4BJS9PR	12:39:04	msn hotmail	Web	2	informational
1844373	138.16.20.231	WKP6E1A4CCM48N	14:00:11	lomi lomi	Web	2	informational
420982	139.168.246.115	M8UJRMA4BR9W8W	20:25:53	pre teen underwear model	Images	4	transactional
2020766	139.55.26.118	K41MNUA4DB4VPI	10:21:02	capitalone	Web	1	navigational

Record: 7 of 400

Figure 1-2: **Snapshot of a typical Search Engine log**

Another important aspect of search engine log data is that a user might interact several times with a search engine until his/her need is satisfied. This episode of continuous interactions termed as a session shows gradual evolution or changes in the attributes of search [7]. This also explains the temporal dependence or influence of previous search attributes on the current and future interactions. Studies have been done for such single file temporal data [8].

However, a challenge is to analyze multiple such log files from several search engines to determine the trends in various attributes and their interactions. From this analysis, one can determine trends across the entire user population. The number of search engines of interest adds the third dimension to this interesting problem in the temporal analysis of search engine logs.

Therefore, for effective and efficient analysis, methods are needed that can handle search engine logs that are (1) huge, (2) multi-dimensional, and (3) temporal.

1.4 Overview of Methodology and Experiments

In this study, we apply tensor analysis as a tool to identify significant temporal patterns in the time-evolving data stream of Web search engine logs. Tensor analysis presents a suitable way of modeling and analyzing the huge data stream with equal justice to each attribute and dimension under consideration. This is evident by the working of tensor decompositions, as explained later in the methodology section. Tensor analysis is also extensible, in that, it can be extended to any number of dimensions that are of interest and is not restricted to the three dimensional descriptions as explained in section

The widespread use of search engines has made the search engine logs a valuable resource for providing insights into the interaction between users and systems. These log files are build incrementally and continuously, with new interactions recorded at every clock tick on almost every single search engine server. Modeling and analyzing this temporal search data requires novel methods to extract patterns of user search behavior or identify possible outliers or abnormal behavior. Transaction log analysis (TLA), as it is popularly known in the field of information retrieval, leads to better understanding of Web search behavior and redesigning the existing engines to suit user's need [6].

1.3. Since the application of tensors is relatively unexplored in the field of search engine log analysis, basics of tensor analysis methods are briefly presented. This study also explains how search engine log data are preprocessed before application of tensor analysis methods.

1.5 Scope

This study implements tensor analysis on five transaction logs from four different search engines. Results of the tensor decomposition show that, few attributes are positively correlated and few others are anti-correlated with one another. There is a peak usage time of search engines at particular part of the 24-hour period. Findings also show that tensor decomposition is a new and novel way to model and analyze search engine log data. In conclusion, this research discusses the implication of these findings and avenues for future research using the tensor methodology.

The remaining sections of this thesis are organized as follows: In Chapter 2, a brief overview of the related work in this field is presented. Chapter 3 outlines the scope of this work and poses the research questions. Chapter 4 describes the methodology used for data analysis in detail. Chapter 5 discusses the results of tensor analysis and its implications. Chapter 6 states the conclusions derived from this transaction log analysis study.

Chapter 2

Related Work

Search engine logs or transaction logs are storehouses of wide variety of information related to human computer information interaction. Transaction log analysis forms one of the major subsets of research interests in area of Information retrieval (IR). Peters defines transaction log analysis in simple terms as "*...the study of electronically recorded interactions between on-line information retrieval systems and the persons who search for the information found in those systems*"[6]. The retrieval systems of interest in this work are search engines. Traditionally the information-search process has five elements associated with it, and TLA deals with the fourth element which is the search user's actions [10] [11]. Thus, transaction log analysis has its own limitations [12] [13]. A good knowledge of these issues help in making the most out of available information through these electronic recordings. Sandore (1993) has reviewed methods of applying results of TLA [14]. Banks (2000) discusses the usefulness of TLA [15].

There are several classes of studies that use transaction logs, few to name are:

1. Studies related to performance of retrieval systems in terms of precision and recall metrics,
2. Studies dealing with understanding of online user-intent and user-behavior, and

3. Those focused on temporal variations of various user-search attributes and statistical summarization of the search engine transaction logs [16] [17].

There are experiments to investigate relevance using the click-through data [20]. Personalization is another aspect that has been explored for increasing the relevance of system results [21]. Ma has examined the concept of object level search [22]. Joachims has utilized click-through data to evaluate the performance of the retrieval systems in an automated way without using manual relevance judgments that are by far slower.

Identification of informational needs of the search engine user forms an important aspect of TLA. The user's intent is not limited to informational but can be navigational or transactional in nature [24]. There are several studies experimenting on different ways of identifying a user's intent [25] [26]. Several efforts have been carried out to better understand the user's behavior during an online searching [29]. Qiu and Cho attempted to automate the process of recognizing the user interest during an online activity [30].

Trend analysis studies use either server side logs or client side logs depending on the research goals and needs [31][32]. Kammenhuber, Feldmann and Weikum have used client side logs for their trend analysis and have incorporated finite state Markov models [33]. The Markov model was used to identify the navigation of a user through different Web pages in terms of state definitions. Fenstermacher and Ginsburg and Fenstermacher have used script-based client side monitoring methodology that is more comprehensive than a Web-browser [34].

Park, Bae and Lee has dealt with both client side and server side transaction logs. Their study focused on session length, query length, query complexity, and content viewed on the Korean Web search engine [57]. There are studies done to model the relationship between certain search attributes. Research by Heckerman and Horvitz is such an example. They incorporated Bayesian approach to model the relationship between words in a user's query for assistance and the informational goals of the user. They proposed several extensions that centre on integrating additional distinctions and structure about language usage and user goals into Bayesian models [58].

Beitzel, Jensen, Chowdhury, Grossman and Frieder (2004) reviewed query logs that constitute total query traffic of a general purpose commercial Web search engine. The study shows that query traffic from particular topical category differs from the query stream and other categories. Such results provide valuable data for improvising retrieval efficiency [59].

Özmutlu, Spink and Seda used artificial neural networks for topic identification. Excite transaction log was used to train the neural network, which then is engaged to identify topic changes in the transaction log. Their report observes that topic shifts were estimated correctly with 77.8 percent precision [60]. In a follow on research Özmutlu, Spink and Huseyin provide results from a time-based Web study of US - based Excite and Norwegian-based Fast Web search logs [61]. The study focused on variations in the search's behavior with time of the day. Such results help in intelligent reallocation of resources and reconstructing the search structure.

Montgomery and Faloutsos deal with identifying Web browsing trends and have confirmed with empirical research that browsing trends are surprisingly stable over time [35]. Perkio, Perttu and Buntine used Topic-based search engines for trend analysis [31]. .Chau, Olivia and Fang deal with trend analysis in a single Web site search engine [36]. Two studies by Jansen, Spink and Booth and Jansen and Spink and are examples of trend analysis on a single search engine log [25] [37].

These previous studies have nearly exclusively experimented on a single Website log or single search engine log or a topic based analysis. The research presented in this thesis is a unique application of tensor analysis for summarization, attribute correlation and trend analysis of multiple search engine logs over a time window of 24 hours. The uniqueness of this research is the attempt to summarize the user search behavior for a framework of multiple search engine logs taken together. Using tensor analysis, the correlation of search attributes is not limited by time, to a single user, a single session, or even a single search engine, but all these aspects taken together for analysis.

It should be noted that topic-based sampling was not used in this analysis. The aim was to analyze entire log files of all search engines. This study involves deriving new search attributes based on the existing data in a transaction log, although the approach could be used for specific topic analysis.

Thus, this requires a convenient model such as a tensor for the data analysis to cater to the nature of data (voluminous, temporal and multi-aspect) for the analysis of transaction logs.

There are several other methodologies that deal with high dimensional data analysis. Research paper by Bouveyron, Girard and Schmid discusses high-dimensional

data clustering (HDDC). The approach to data analysis in this work was that high-dimensional data usually exist in different low-dimensional subspaces hidden in the original space [**65**]. Lawrence in his technical report on visualization of high dimensional data explains about principal component analysis [**66**]. Apart from HDDC and principal component analysis as modes of high dimensional data analysis, Vector space models and Graph Based models also are adopted. In this work we have chosen Tensor analysis because it is relatively unexplored in the field of transaction log analysis and especially for trend analysis and summarization purposes.

2.1 Tensor Related Study

In this research, tensor decompositions are applied to recognize trends in five data logs from four search engines. Tensors are extensions of matrices to accommodate additional dimensions. Tensors are N-dimensional-arrays, which have their origin from multi-linear algebra [38]. Tensors have been used in many fields where the data to be analyzed is multi-dimensional in nature and voluminous [39][40][41]. Powerful tools have been proposed like Tucker and PARAFAC/Canonical decomposition [42][43]. Shashua and Levin adopted tensors in machine vision research for linear image coding [44]. Vasilescu and Terzopoulos used tensor decompositions for face recognition [45]. Researchers in the graphics field have to deal with efficient representation techniques and compression methods for enormous amount of visual effects data [41].

Sun, Philip and Yu introduced window-based tensor analysis on high-dimensional and multi-aspect streams evolving from an environmental sensor-monitoring network. This study used periodic data generated from 52 sensors measuring environmental variables [46].

In the only IR research located that used tensor analysis, Sun, Liu,Yuchang Lu, and Chen attempted to improve Web search on a single search engine using CubeSVD (Cube Singular Vector Decomposition). However, this study dealt with using tensor decomposition to model user Web search data from a single search engine only (the MSN search engine) [47]. In addition, the attributes experimented by Sun, Liu,Yuchang Lu, and Chen are different from those in this study. This research uses several derived attributes from the available search engine log data.

The work presented in this thesis is a unique addition to the vast TLA literature wherein a new data model for analyzing multi-search engine data simultaneously is experimented. Chapter 3 discusses the scope of research and states several questions that help in understanding the need for a tensor based analysis methodology.

Chapter 3

Research Scope

3.1 Research Questions

The goal of this research is to experiment a tensor-based method for recognizing temporal patterns and outliers in multi-search engine transaction logs. This study builds on prior work on tensor decompositions to demonstrate the promise of tensor-based methods in mining time evolving voluminous and multi-dimensional data. This thesis is one of the first works in the IR field that uses tenor analysis.

The research questions addressed in this study are:

1. Is the tensor data model a viable way to represent the time evolving data stream with high dimensionality and multiple aspects of Web searching?
2. Are there recognizable patterns (normal and abnormal) in user-system-information interactions, and what are the characteristics of these patterns?
3. How are search attributes correlated with each other over a fixed period?
4. What are the outliers from the normal data stream, and is there any meaningful reasoning for their abnormal characteristic?

3.2 Research Data

The data used in this study comprises of five transaction logs from three different well-known search engines. Each of these logs has data that spread over a 24-hour period (i.e. 12 AM to 12 PM):

- Excite log collected in 1997 with 1,025,907 transaction entries.
- Excite log collected in 2001 with 594,940 transaction entries.
- AltaVista log collected in 2002 with 208,154 transaction entries.
- Dog pile log collected in 2004 with 1,523,793 transaction entries.
- Dog pile log collected in 2006 with 4,201,071 transaction entries.

Each transaction log had eight user-search attributes of which three were numerical values (namely, *Query length*, *Time* and *Number of queries*). Five other attributes are categorical in nature (namely, *User-Intent, Search-Pattern, Vertical* and *Rank*). A brief description of these attributes is given in Table **3-1**

Table 3-1: **Search attribute descriptions**

Field	Description
Record Identification Number	An integer that uniquely identifies each transaction entry.
IP Address	Information that indentifies the computer from which the search process is carried out.
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 users, such as site preferences and the contents of their electronic shopping carts.
Time	The particular time at which the interaction was electronically recorded by the search engine server measure in hours, minutes and seconds
Query	Series of terms as typed by the Web searcher into the search box of a search engine
Vertical	The different content collection types that make the search experience more convenient to obtain information in required format. Basically there are five types of verticals namely Web, Audio, Image, Video and News.
Search Pattern	Variations or evolution of search terms in a session categorized based on certain criteria
User Intent	The ultimate goal of a Web search engine user during his interaction with a search engine.
Rank	Ranking of the result page chosen by the searcher.
Browser	Web browser used by the searcher.
Query Length	Number of words or terms in a query

Chapter 4

Methodology

This section is a brief review of fundamental definitions of tenors and their decompositions that are considered in this study. Advanced information on tensor analysis can be found in [51] [52] and [46]. Tensor analysis has similarities to matrices and Markov models. Matrices are a powerful option for mathematical analysis of given data. They have only two dimensions, and one often runs into situations where data is multi-dimensional in nature. Examples of such data are, data centric monitoring, environmental sensors, social networks, network forensics, and Web mining [48] [49][50]. Tensors can aptly represent such multi-dimensional data.. Using tensors, it is possible to deal with a much wider span of problems than what can be solved using solely matrix representation [48].

Markov models adopted by Kammenhuber, Feldmann and Weikum [33] are probability-based prediction models. The assumption made in such statistical approaches is that the system being modeled is a Markov process. Markov process is a stochastic process in which there is conditional dependence between the states of a system. Another aspect of Markov models is that it is a static analysis. The data has to be completely available before the analysis is started. It is to be noted that tensor model is a mathematical approach with strong roots from multi-linear algebra. There is no assumed system process and no defined states. It is a mathematical entity or structure on which certain operations are defined. Tensor analysis accommodates for dynamic data analysis.

Data need not be present at the time of analysis, but can arrive as it is generated and can be added on to existing results for current analysis phase. This is the main difference between the Markov model approach and tensor analysis used in this study.

4.1 Tensor

A tensor is defined as a multi-dimensional or N-way array, where N ≥3 [53]. In linear algebraic terms, tensors are multi-linear mappings over a set of vector spaces. For example, a scalar is a 0^{th} order tensor, vector is a first order tensor, and a matrix is a second order tensor. Figure 4-1 represents this notion pictorially. Table 4-1 explains the notations used in definitions related to tensors throughout this paper

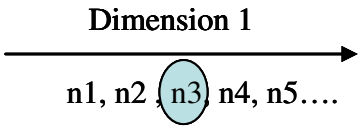
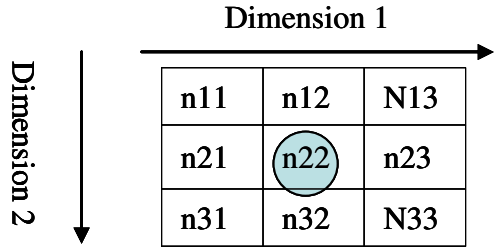
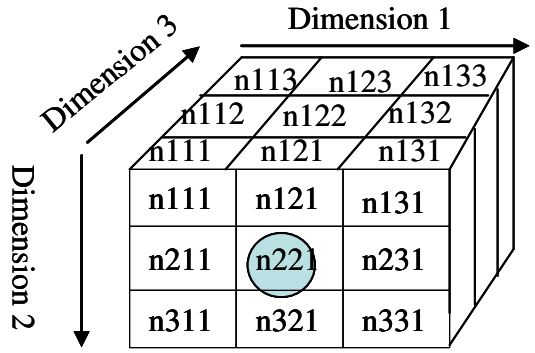
Order	Type	Example
1	Vector	
2	Matrix	
3	3D Array/ MDA	

Figure 4-1: **Pictorial representation of tensors.**

Table 4-1: **Tensor notations.**

Notation	Description
B	Lower case bold letter represents a vector
b(i)	The i- th element of vector **b**
M	Upper case bold letter represents a matrix
$\mathbf{M}^T$	The transpose of a matrix
$\mathbf{M}_i \Big\vert_{i=1}^{n}$	Sequence of N matrices $\mathbf{M}_1, \mathbf{M}_2, \mathbf{M}_3, \mathbf{M}_4, \ldots\ldots\ldots, \mathbf{M}_n$
M (i , j)	The entry (i ,j) of **M**
M (i , :) **or M** (: , i)	i-th row or column of **M**
$\mathcal{T}$	Calligraphic style denotes a Tensor
$\mathcal{T}$ $(i_{1,\ldots\ldots,}i_M)$	The element of $\mathcal{T}$ with index $(i_{1,\ldots\ldots,}i_M)$
M	Capital and italicized denotes order of a tensor
N	Denotes the dimensionality of the i th mode ($1 \le i \le M$)

The order of a tensor $\mathcal{T} \in \Re^{I_1 \times I_2 \times \ldots\ldots\ldots I_M}$ *is* M. An element of $\mathcal{T}$ is denoted as $\mathcal{T}_{i_1 \ldots\ldots i_n \ldots\ldots i_N}$ or $t_{i_1 \ldots\ldots i_n \ldots\ldots i_N}$ where $1 \le i_n \le I_n$. Thus, an n[th] order tensor is accessed via 'n' vertices.

Rank of a tensor $\mathcal{T}$ is denoted as rank($\mathcal{T}$). It is the smallest number of rank-one tensors that can add up to $\mathcal{T}$ as their sum. Tensor $\mathcal{T} \in \Re^{I_1 \times I_2 \times \ldots\ldots\ldots I_M}$ is said to be a rank-one tensor if it can be expressed as the outer product of *M* vectors. This also refers to smallest number of components in PARAFAC decomposition.

The definition of tensor rank is similar to matrix rank, but it is to be noted that the properties of matrix rank and tensor are quite different. The rank of a matrix is well defined with no special cases attached to it. The Tensor rank has the complications attached to it because there is no straightforward algorithm to determine the rank of the given tensor. There are different kinds of ranks of a tensor namely, *maximum* and *typical* ranks. *Maximum* rank is referred to as the largest attainable rank for a tensor. *Typical* rank refers to any rank of a tensor, which occurs with a probability greater than zero. Thus, tensors can have more than one ranks associated with them. They can have more than one typical rank in addition to one maximum rank.

4.2 Matricization

Transforming a tensor into set of matrices is an important step in decomposing a tensor. The process of unfolding the tensor along each of its dimension (mode) is known

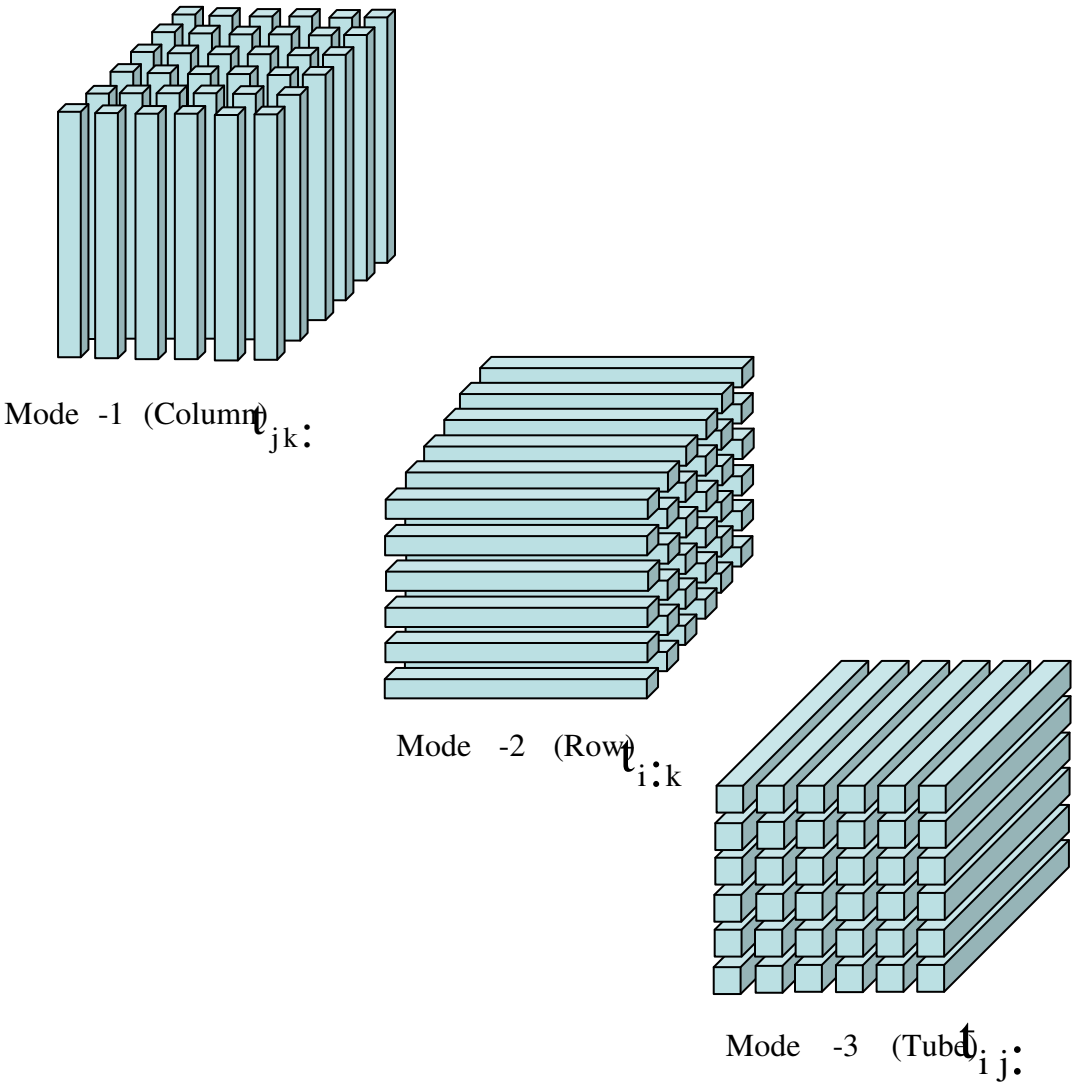


Figure 4-2: **Fibers in a 3rd order Tensor**

as *Matricizing*. The n[th]-mode Matricization of a tensor $\mathcal{T} \in \Re^{I_1 \times I_2 \times \ldots\ldots I_M}$ is denoted by $\mathbf{A}_{(n)}$ which arranges n[th]-mode fibers into columns of a matrix.

Fibers are higher order analogues of matrix rows and columns. Figure 4-2 illustrates three kinds of fibers possible in a 3-dimensional tensor.

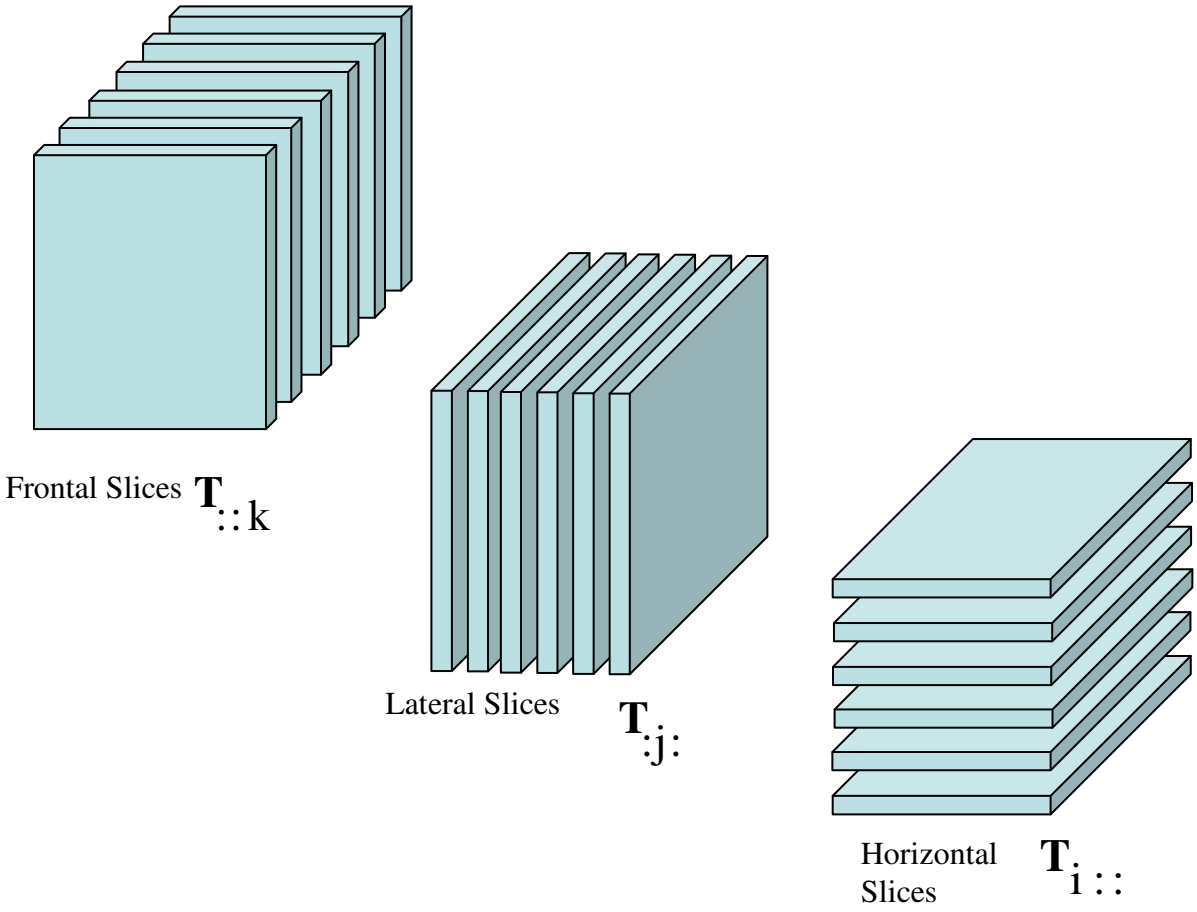
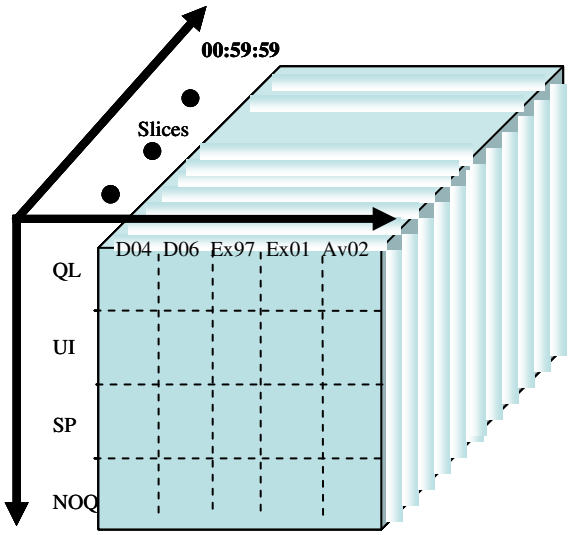


Figure 4-3: **Slices of a 3rd order Tensor**

Slices are two-dimensional sections of a tensor obtained by setting all but two indices. Figure 4-3 illustrates three kinds of slices in a three dimensional tensor.



QL	Query Length
UI	User Intent
SP	Search Pattern
NOQ	Number of queries

Figure 4-4: **Tensor streams.**

4.3 Tensor Streams

A sequence of equal ordered tensors is termed as a tensor stream. Each tensor can be referred to as an individual slice of the sequence or the stream. Figure 4-4 shows tensors of equal dimensions stacked up in time that were used in this research. For each slice, the rows constitute attributes whereas the columns are search engine transaction logs from which the combined data is assembled into the tensors.

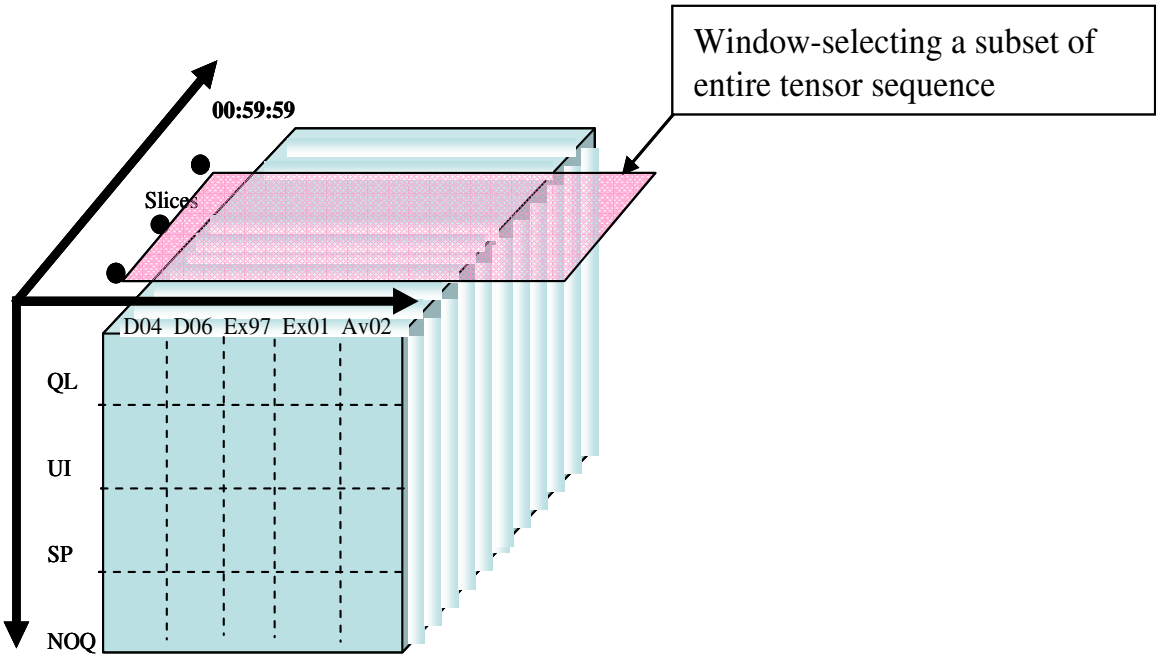


Figure 4-5: **Tensor window.**

4.4 Tensor Window

A tensor window is subset of a tensor sequence based on certain criteria for example, a tensor stream ending at a particular time. The number of individual tensors or slices that make up this subset denotes the size of the window (see Figure 4-5).

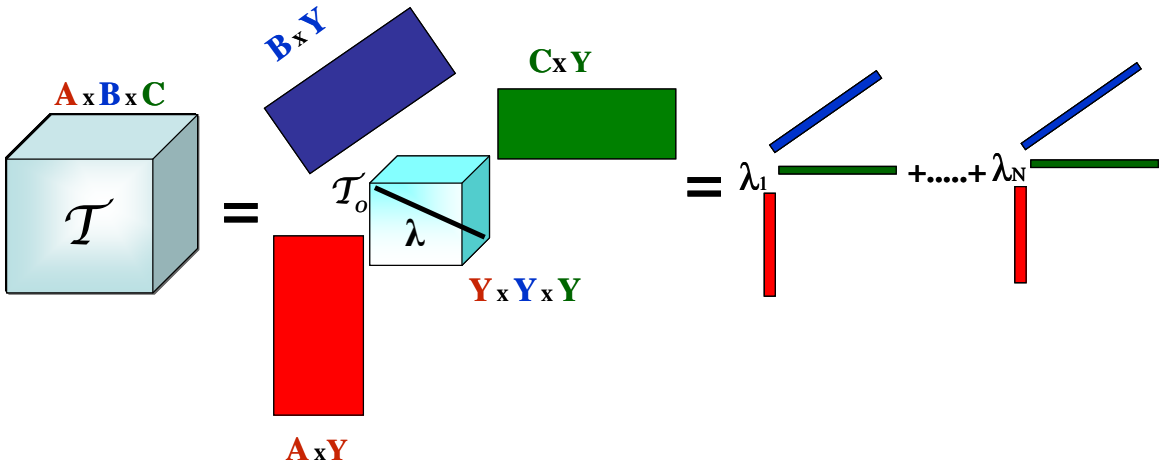


Figure 4-6: **Schematic representation of PARAFAC decomposition.**

4.5 Tensor Analysis

Given a sequence of tensors of equal order, the process of finding the orthogonal matrices $\mathbf{U}_i \in \Re^{N_i \times R_i}$ one for each mode given that the reconstruction error is at minimum is called tensor analysis. Subscript '*i*' refers to the i^{th} mode of an N-mode tensor.

4.6 Tensor Decomposition

Higher order tensor decompositions are analogous to the more familiar Singular Value decompositions (SVD), but they overcome the limitations of the matrices. Tensor

analysis is carried out by tensor decompositions. There are several varieties of tensor decompositions of which the following are most commonly used in many applications:

a) Carroll and Chang in 1970 proposed CANDECOMP decomposition (CANonical DECOMPosition) and Harshman in 1970 proposed PARAFAC decomposition (PARAlell FACtors). The two decompositions are jointly also known as the CANDECOMP-PARAFAC (CP) model or decomposition.

b) TUCKER model and was proposed by Tucker in the year 1966.

This research adopts PARAFAC decomposition as a means to analyze transaction logs. Figure **4-6** provides a schematic representation of PARAFAC decomposition. A tensor toolbox designed by Bader and Tamara was adapted for this implementation into MATLAB software. It may be noted that this research deals with the

application of existing tools that implement tensor decomposition algorithms to analyze transaction logs instead of proposing new decomposition techniques.

The tensor toolbox has a function, “parafac_als” that computes an estimate of the best rank-R PARAFAC model of a tensor using an alternating least-squares algorithm.

4.7 Algorithm Outline for Tensor Decomposition

Input to the algorithm is a tensor window $\mathcal{T} \in \Re^{W \times A \times S}$ ($\mathcal{T} \in \Re^{Window \times Attributes \times SearchEngine}$) where *window* refers to the time window that selects a subset of tensor stream based on certain criteria. As an example, a one-minute tensor

slices that belong to the 0^{th} hour are all transactions that occurred between 00:00:00 to 00:59:59. Each tensor in the stream is of the form $\mathcal{T} \in \Re^{A \times S}$ ($\mathcal{T} \in \Re^{Attributes \times SearchEngine}$) occurring at the time hh:mm:ss. Output of the algorithm is a core tensor of the form $\mathcal{T}_o \in \Re^{W_0 \times A_0 \times S_0}$ and the three projection matrices for each dimension namely,

$\mathbf{U_0} \in \Re^{W \times W_0}$, $\mathbf{U}_1 \in \Re^{A \times A_0}$ and $\mathbf{U}_2 \in \Re^{S \times S_0}$ .The first column of each projection matrix has dominating (main) pattern values whereas the second column has pattern values that deviate from this main pattern. Figure **4-7** presents an outline of the iterative algorithm for tensor analysis used in this research.

Initialization

Choose T_0 , A_0, S_0

Calculate U_0, U_1, U_2 as follows:

U_0 = Leading T_0 left singular vectors of

U_1 = Leading A_0 left singular vectors of

U_2 = Leading S_0 left singular vectors of

Until converge do…..

U_0= T_0 leading left singular vectors of $X_{(1)}$(S A)

U_1 = A_0 leading left singular vectors of $X_{(2)}$ (S W)

U_2 = S_0R leading left singular vectors of $X_{(3)}$(A W)

Solve for core tensor

$$\mathcal{T}_0 = W \prod_{i=0}^{2} \times_i \mathbf{U}_i$$

Figure 4-7: **Outline of Iterative Tensor Decomposition.**

4.8 Software Support for Tensor Calculations

There are several toolboxes and math soft wares available for tensor-related calculations. Mathematica, Maple and MATLAB have extended their functionalities to

enable tensor related calculations. Mathematica has a separate package to implement working tensors.

N-way, CuBatch, PLS and Tensor Toolbox are some of the toolboxes developed to support tensors. N-way Toolbox for MATLAB by Anderson and Bro [**54**] has support for several algorithms for different kinds of decompositions. CuBatch [**55**] provides a graphical interface for analysis of data portrayed in as N-mode and build using N-way Toolbox. PLS is a commercial toolbox for MATLAB but with an added importance to Chemometrics related data analysis.

The toolbox designed by Kolda and Bader [**56**] has a general-purpose set of function calls to support tensor decompositions. It also supports structured tensors for storage and manipulation.

Chapter 5

Experimental Design

5.1 Pre-processing

As mentioned earlier in Section 3.2 , the data is comprised of five transaction logs from three different search engines. Each of these logs has data spread over a 24 hour period (i.e. 12 AM to 12 PM) totaling approximately 8 million queries. Each transaction log has eight user-search attributes of which three were numerical values namely, *Query length*, *Time* and *Number of queries.* Five other attributes are categorical in nature namely, *User-Intent, Search-Pattern, Vertical* and *Rank.*

The *Time* data on all logs was converted into serial time, which is a common integer ranging from 0 to 235,959 (24 hours). The categorical data, namely, *User-Intent*, *Search-Pattern and Vertical* were given integer values as per research by Jansen, Spink ,Blakely, and Koshman. Table 5-2 and Table 5-3 describe the assignment of integer values to categorical variables.

5.2 Note on User Intent

A Web search engine user interacts with the search interface in response to some need. The user paraphrases his need through a string of terms called query. The search engine takes the query as input and retrieves relevant documents that possibly satisfy the

users need. The need that motivated all this process is referred to as *user goal* or *user intent*. Thus, user intent is the purpose with the user approaches the search engine to satisfy the need. It is to be noted that a *user* in this work refers to a Web search engine user and *searcher* and *user* are used interchangeably throughout this work depending the emphasis.

Traditional IR systems have operated on the idea that the user's need is always *informational* in nature. Recent studies have argued and indicated that, a user's need is not restricted to *informational* but extends to other categories [**26**]. On a broad scenario the other categories to which the user intent extends are namely,

1. **Navigational**: The intention of the user is to reach or locate a specific Website.
2. **Transactional**: The intention of the user is to carry out Web mediated transactional activity.
3. **Informational**: The intention of the user is to amass information or knowledge from multiple sources.

Thus, the intention behind a query submitted by the user is not entirely informational in nature. Studies have identified subcategories within this broad classification [**27**][**28**]. A brief description of these sub divisions is given in the Table **5-1**.

Table 5-1: **Description of User Intent.**

User Intention or Goal	Description	Examples
1.**Navigational**	The intention of the user is to navigate or reach to a specific Website.	Costco Sam's club
2. **Informational**	The intention of the user to gain knowledge by browsing several Web pages.	
2.1 *Directed*	The intention of the user is to learn about a specific subject.	What is a search engine?
2.1.1 Closed	The intention of the user is to obtain a single unambiguous answer.	Who is the president of United States of America?
2.1.2 Open	The intention of the user is to obtain answer to an open ended question.	Cricket
2.2 *Undirected*	The intention of the user is to know everything about a subject or topic.	Cancer
2.3 *Advice*	The intention of the user is to get some advice on the subject.	Help reducing hair loss.
2.4 *Locate*	The intention of the user is to locate a particular place.	Original Waffle Shop
2.5 *List*	The intention of the user is to obtain a list of other Websites that are useful in satisfying unspecified interest.	Cricket Clubs
3.**Transactional**	The user intention is to perform a transaction or get hold of a resource.	
3.1 Download	The user intention is to download materials from online source.	Beethoven music Files
3.2 *Entertainment*	The user goal is to get entertained by viewing the items available on the Webpage.	YouTube videos
3.3 *Interact*	The user intention is to interact with a resource using another program or service.	Temperature Converter
3.4 *Obtain*	The user intention is to get hold of a resource that does not need future use of a computer.	Citizenship application documents

Table 5-2: **Representation of 'User Intent' categorical data as integer.**

User Intent	Description	Value Assigned
Informational	User intents to acquire information assumed to be present on one or more Web pages	1
Transactional	User intends to perform some Web mediated transaction	2
Navigational	User intends to reach a particular Web site	3

Table 5-3: **Representation of 'Vertical' categorical data as integer.**

Vertical	Value Assigned
Web	1
Image	2
Audio	3
Video	4
News	5

The fields *Time, Query, Query-Length, Vertical and Rank* were logged and sorted as data columns in the transactions log. *Search-Pattern, User-Intent, Number-Of-Queries* were derived from the existing data using techniques outlined in [**27**]. Table **5-2** and Table **5-3** show the values assigned to the User Intent and Search engine verticals. Description of Search Pattern attribute is explained in Table **5-4**.

5.3 Note on Search Pattern

The interactions among the user, the search engine and the content provided by the search engine is temporal in nature. During a search process, several levels of interaction may occur. Thus, there might be several sessions within one searching process. Identifying and characterizing these sub sessions are an interesting filed of study. Studies by He, Aye and Harper in 2001 and Shneiderman, Byrd and Croft in 1998 are example for session pattern identification [**12**] [**11**].

Based on the work by Jansen, Spink, Blakely and Koshman in 2007 , which discusses three methods to identify sessions, the algorithm to derive the search pattern data was adopted [**7**]. Table **5-4** provides a brief description of the search patterns incorporated for data analysis.

As a common rule, a null entry in any data column was assigned an integer value of zero for uniformity. All of the above pre-processing was carried out using Microsoft Access application. The resulting database files were exported to text files. A MATLAB

Table 5-4: **Description of Search Pattern.**

Search Pattern	Description	Value assigned
Assistance	Query generated by the searcher based on some assistance from the search engine. Example "Are you looking for?"	1
Content change	Query identical to previous one but executed on different vertical or content collection.	2
Generalization	Current query is on same topic as previous one but user seeks more general information.	3
Generalization with Reformulation	Generalization + reformulation	4
Specialization	Current query same as previous one but more specific in nature	5
Specialization with Reformulation	Specialization + reformulation	6
Reformulation	Current query on same topic as previously entered query and both have common terms	7
New	The query is on new topic	8

script imported these text files and performed initial normalization (zero mean and unit variance) operations. After normalizing every attribute by data column, the rows of the matrices were sorted based on the *Time* attribute (0-235959). Refer to Figure 5-1 for the various processing steps in tensor analysis of transaction logs.

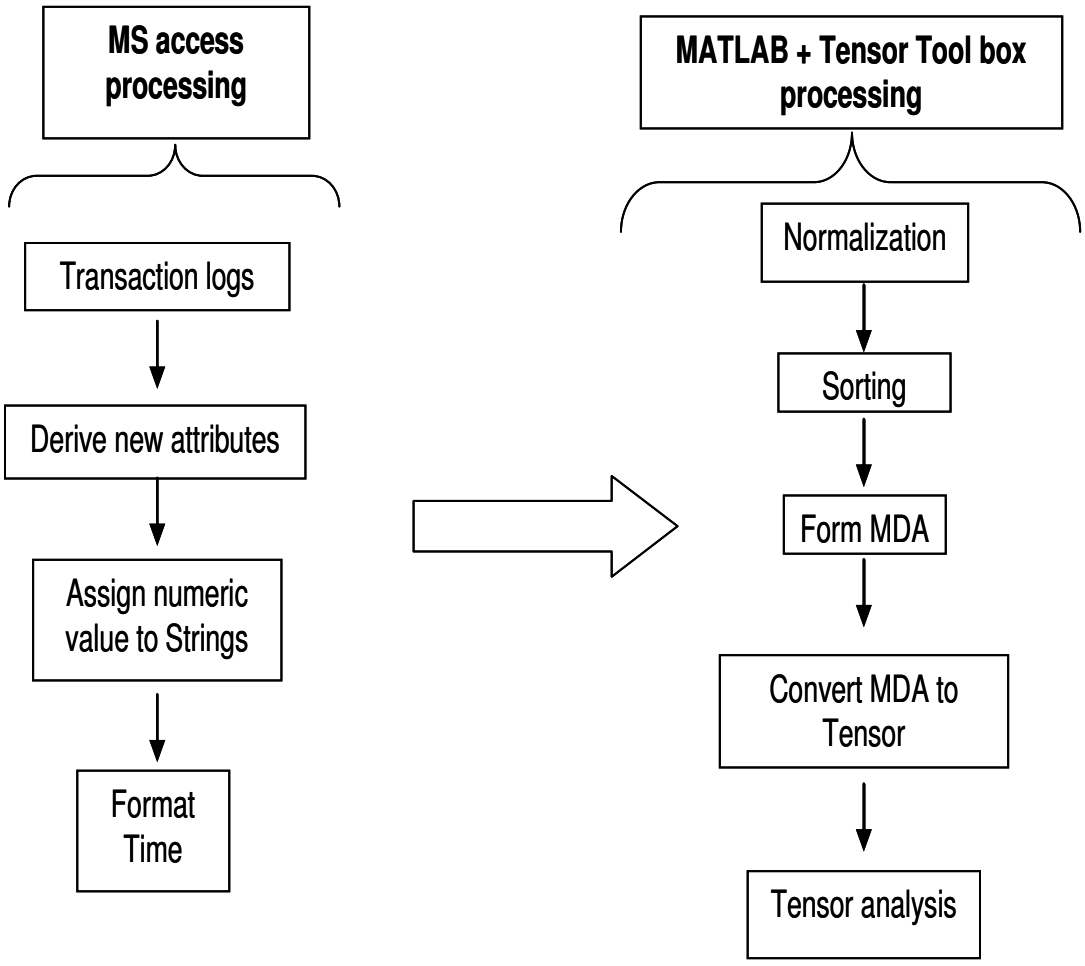


Figure 5-1: **Processing Stages.**

5.4 Tensor Construction & Decomposition

At the end of preprocessing stage, the transaction logs are formatted such that numerical data can be exported to build a tensor. Figure **5-2, portion a,** shows a sample format of the logs after preprocessing.

In a preprocessed log file, every column refers to a unique search attribute and each row corresponds to a unique transaction entry (or query). This data is converted into a matrix with similar structure.

A separate tensor was built for every hour of the day consisting of several slices. For a tensor spanning a time stamp of 1 hour, the idea is to pool up transactions from all search engines that occur at a specific second in that hour (see Figure **5-2**, **portion b**).

If we assume that there is at least one transaction in every second of that hour, there will be at least 3,600 slices in the tensor. Each slice consists of attributes of a transaction corresponding to a particular time stamp across all search engine logs as shown in Figure **5-2(portion b)**. Accordingly, there are 24 tensors for an analysis period of 24 hours each having a variable number of slices depending on number of transactions recorded in the log for that time span.

Once the tensors have been assembled, the next step in the analysis procedure is to decompose them into their constituent orthogonal projection matrices corresponding to each dimension of the tensor. In this research , the tensor is decomposed into orthogonal projections identifying the first and second factors for three modes, namely, attributes, search engines, and time as shown in Figure **5-2 (portion c)**.

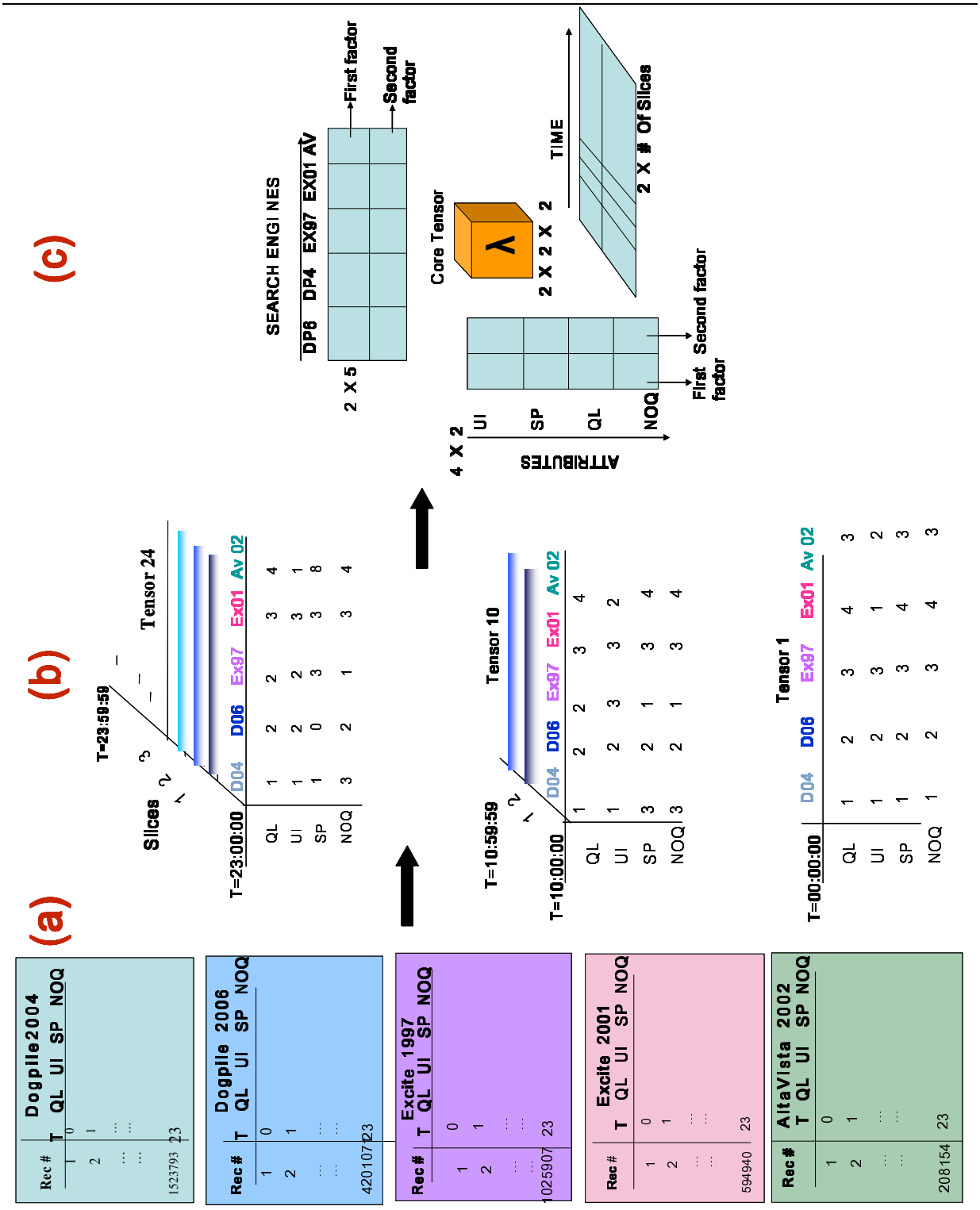


Figure 5-2: **Tensor construction and decomposition.**

All experiments were designed to analyze the data in the *Hourly Multiple-tensor Analysis* approach. In this approach, the tensors built for every single hour for a 24 hour period where used to study variation of search attributes. Each of this hourly tensors, were decomposed to obtain the normal (1^{st} factor) and abnormal (2^{nd} factor) trend values. Interesting time slots are selected from the results of the *Hourly Multiple-tensor Analysis* and analyzed for attribute correlation at those time slots.

Chapter 6

Results and Discussion

6.1 Hourly Multiple-Tensor Analysis

In this experimental set up, as explained earlier in section **5.4,** separate tensors with varying size along the time dimension were constructed. The decomposition values were plotted on a single graph to observe the variation of search attributes over the 24-hour period. The following sections in this chapter discuss the obtained results.

6.1.1 *High Usage Time window*

The following are some preliminary results and not as part of tensor decomposition. These are basic statistical observations made. The number of transactions per clock tick varies for each search engine.Figure **6-1** lays out the total number of transactions that occurred at each hour (0^{th} to 23^{rd}) for all three-search engines It can also be inferred that between 10^{th} and 17^{th} hours, all had substantial increase in the number of transactions that denotes this period as a peak period in search engine usage.

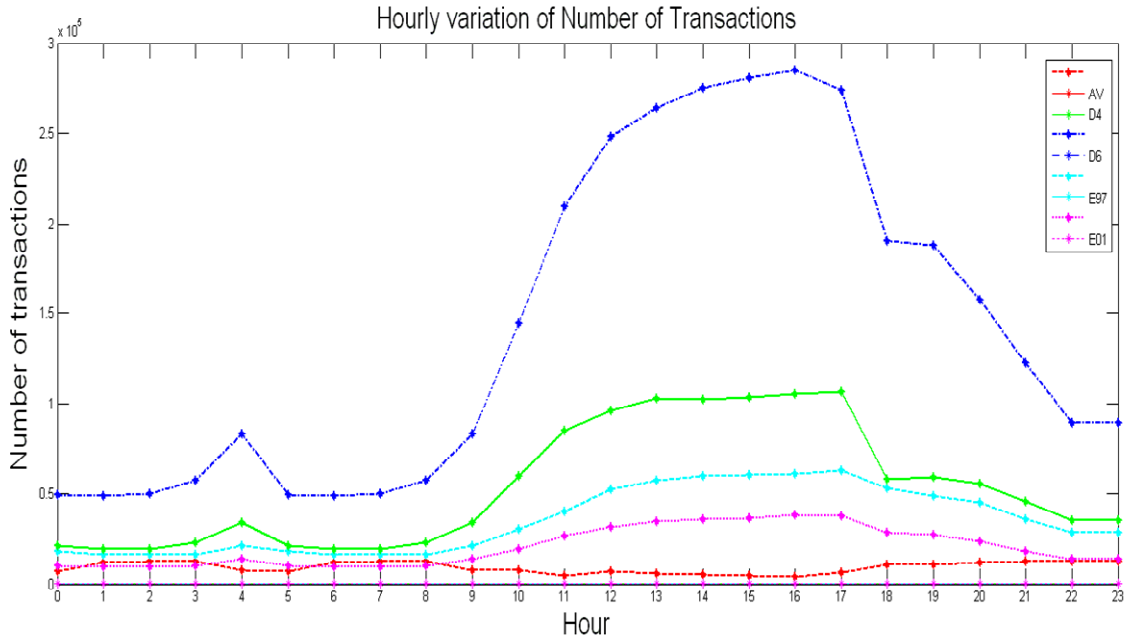


Figure **6-1**: **Number of transactions per hour.**

Analogous to the above observation, tensors formed for each hour have similar metrics when it comes to number of slices per tensor. As shown in Figure **6-2**, between 10th and 17th hour, the tensor slices were consistently increasing in number and gradually decrease towards 23rd hour that marks the end of the day and hence indicates lesser interaction. The largest tensor was formed for the 15th hour that falls in the peak hour. Peak hour referes to the time window where the number of transactions recorded are substantially higher than other time slots. This indicated high search engine usage.

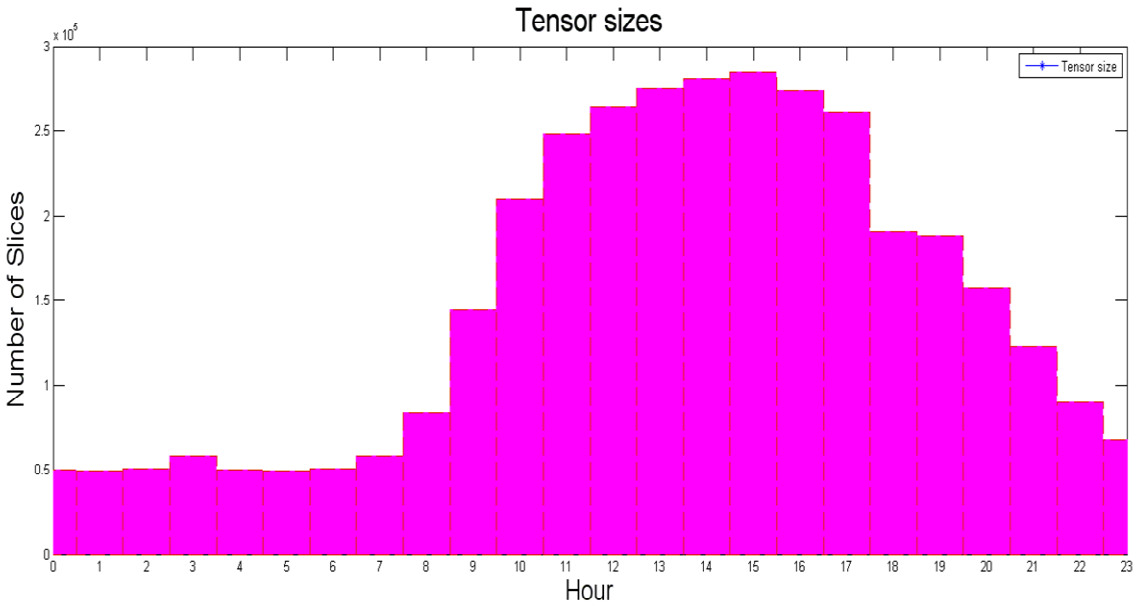


Figure **6-2**: **Variation of tensor size with time.**

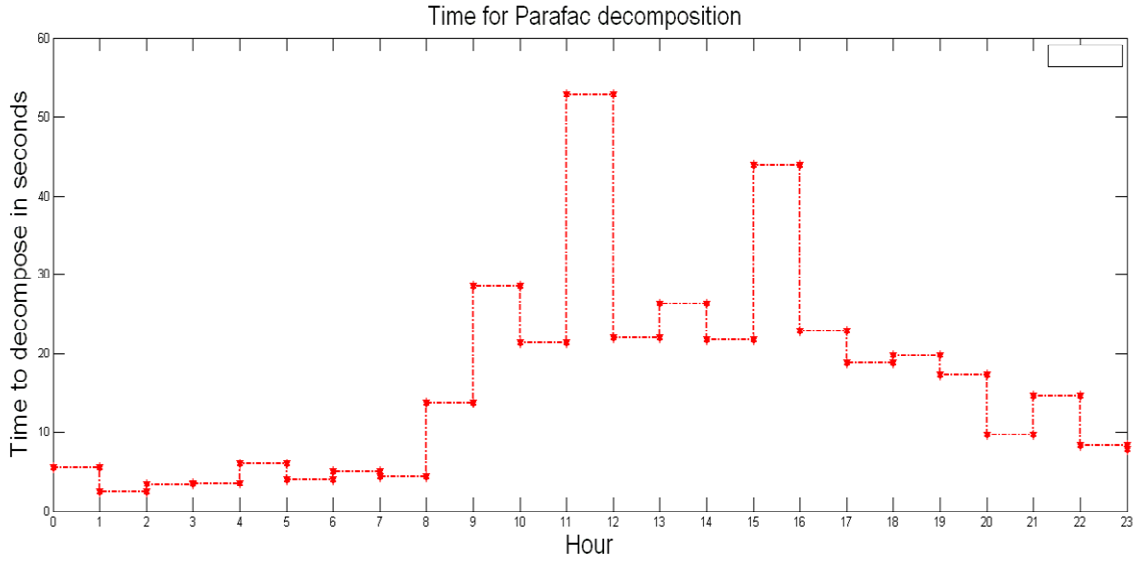


Figure **6-3**: **Decomposition time.**

6.1.2 *Time to Decomposition*

Time to decompose a tensor depends on the size of the tensor as well as the time required for the convergence part of the decomposition algorithm. Figure **6-3** shows two peaks at 11^{th} hour and 15^{th} hour. Although tensor sizes were large during the peak hour ($10\text{-}17^{th}$ hour), time to decompose some of the peak time tensors did not follow a linear relationship with their size. Only the 11^{th} and 15^{th} hour tensors took observably higher time than others did. This may be due to a quicker convergence in the iterative least squares algorithm used to decompose these peak time tensors.

6.1.3 *Variation of search attributes over 24 hour window*

Normal and abnormal trends of four search attributes were studied in this work namely *Search Pattern, Number of queries, Query length and User Intent.* Figure **6-4,** Figure **6-5,** Figure **6-6** and Figure **6-7** , present the trend analysis on all four attributes. In the plot, a green line represents normal trend and a red line represents the abnormal trend.

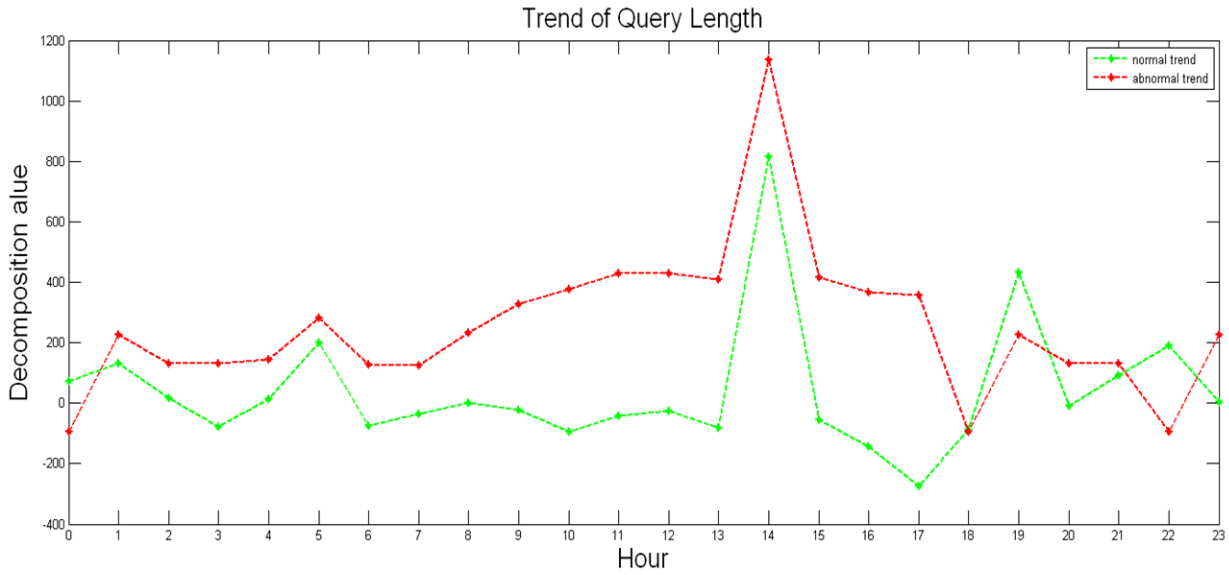


Figure 6-4: **Trend analysis for Query Length (QL) attribute.**

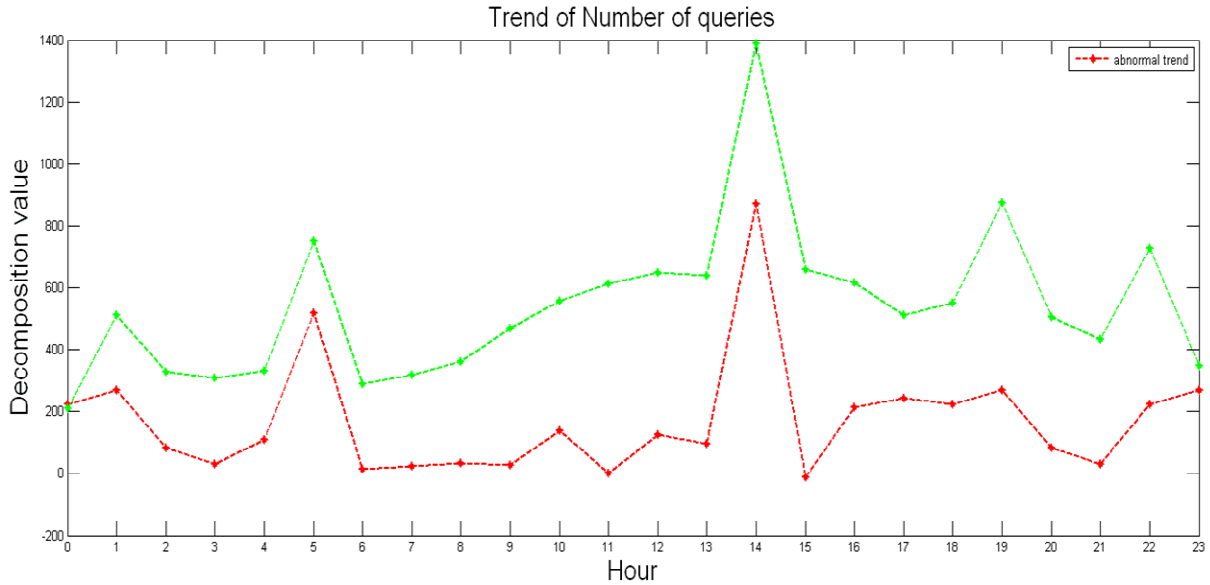


Figure 6-5: **Trend analysis for No. of Queries (NOQ) attribute.**

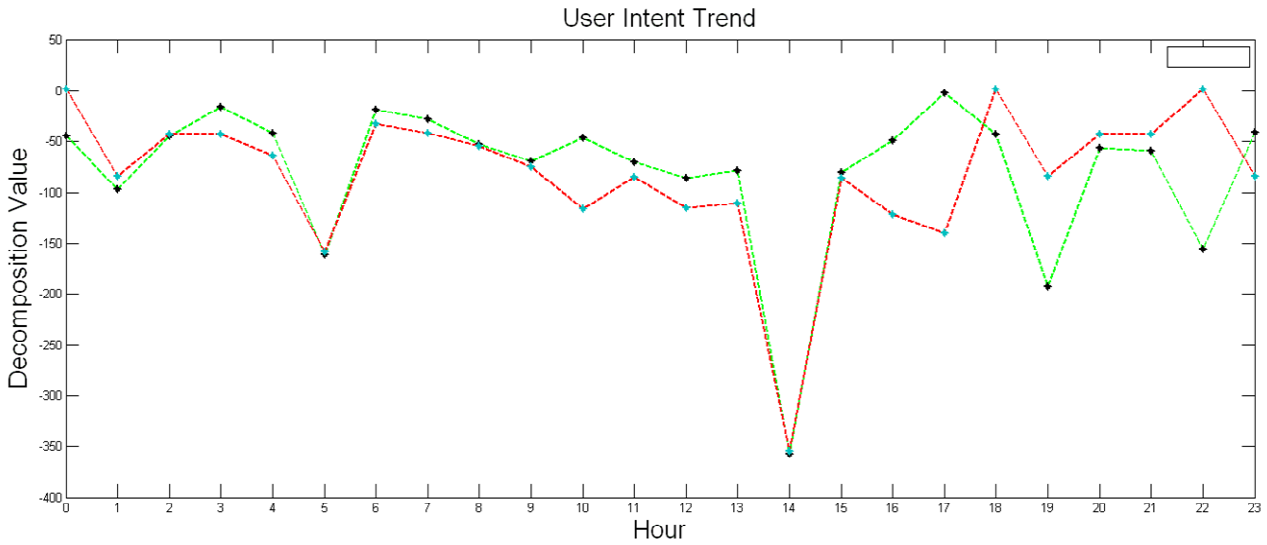


Figure 6-6: **Trend analysis for User Intent (UI) attribute.**

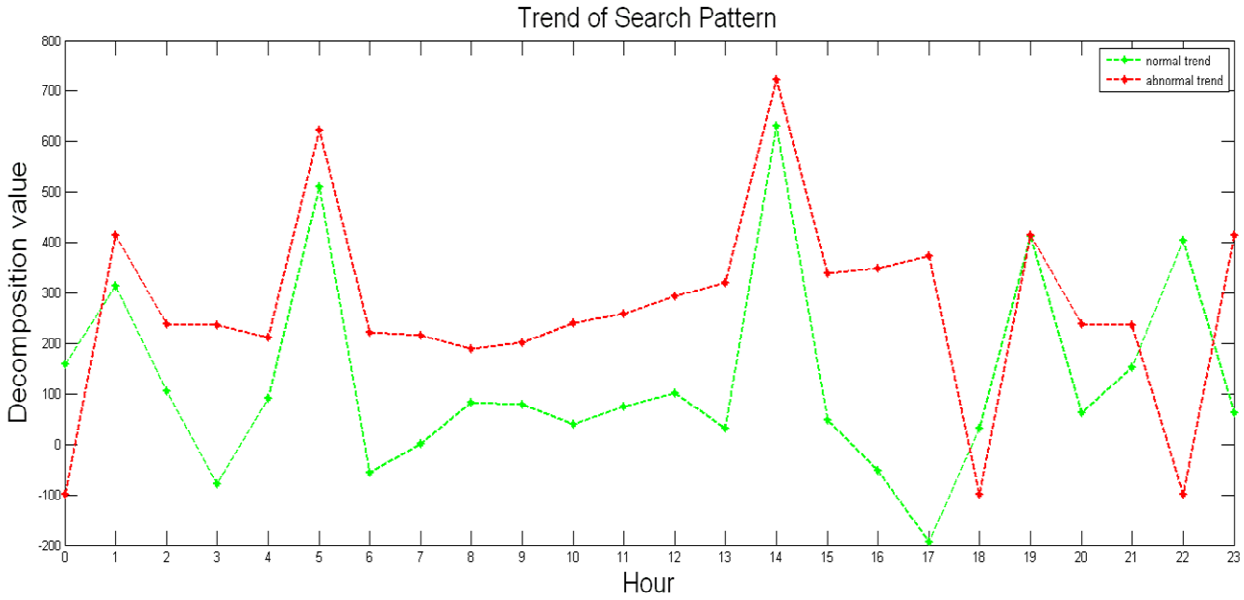


Figure 6-7: **Trend analysis for Search Pattern (SP) attribute.**

All the plots (except *User Intent* plot) have a common observably high value (both normal and abnormal trend lines) at 14th hour, which falls in the peak hour window.

The 14th hour also had the second largest tensor as shown in Figure **6-2**. It is observed that at 18th hour the abnormal trend line for *Search Pattern* attribute shows a downward peak which also coincides with downward peak in the *Query Length* (abnormal trend) and shows a noticeable increased value for *User Intent* (abnormal trend). It can be inferred that after the peak phase of the day, transactions that had abnormal trend had smaller *query length*, were lower in *search pattern* values and higher in *User Intent* values. This implies that these interactions were *transactional* or *navigational* in nature, used assistance and had smaller query lengths.

Figure **6-8** and Figure **6-9** summarize the participation of the search engines towards trends recognized during tensor analysis. Participation is defined as the proportion of influence of the data from search engine transaction logs that have contributed towards these trends. It is evident that Dogpile 2006 has consistently overshadowed other log files in the trend analysis. This is consistent with observations made from Figure **6-1** dealing with *Number of transactions per hour* for search engine logs. Figure **6-8** and Figure **6-9** both show a high value at the 14th hour for normal and abnormal trends.

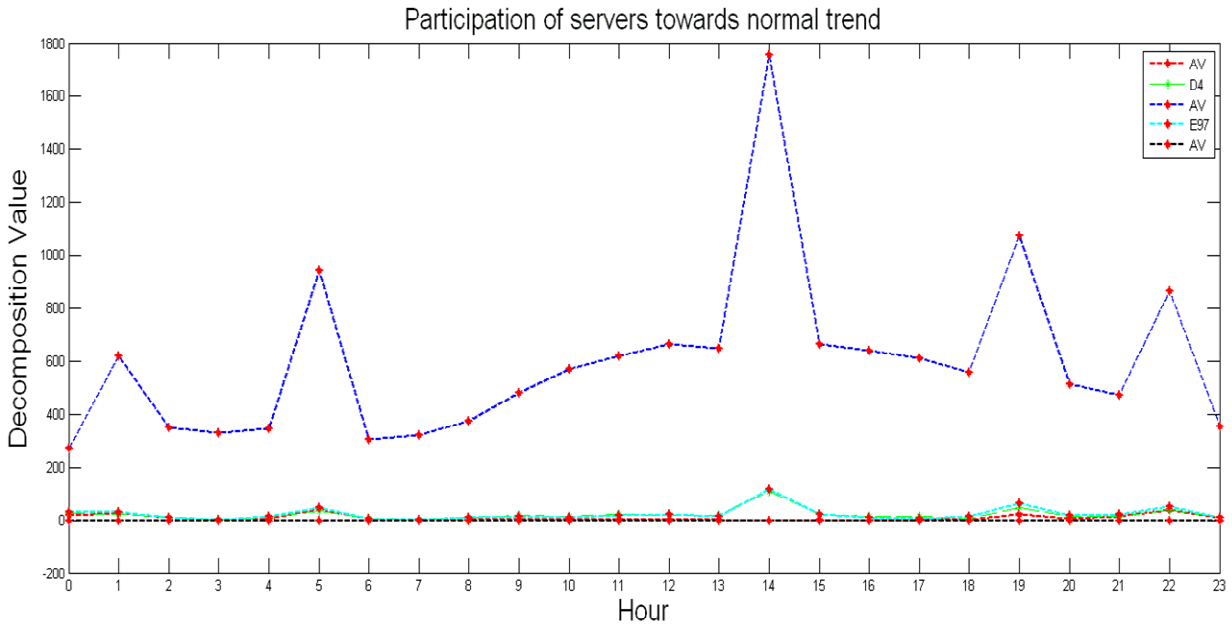


Figure **6-8**: **Participation of search engines in normal trend analysis.**

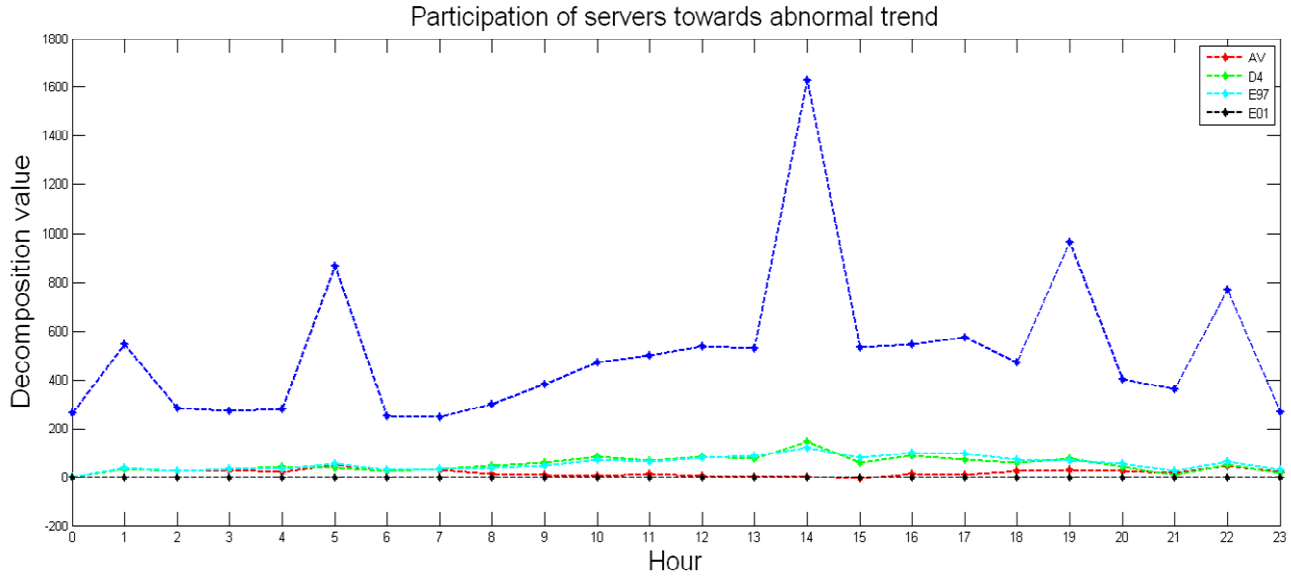


Figure **6-9**: **Participation of search engines in abnormal trend analysis.**

6.2 Implications of Results

Although our approach is limited by the datasets available, we can draw some implication of the tensor analysis method with an ideal dataset. Assuming that the search engine logs were taken from the same year, given region, and standard sampling method, we can summarize the above observations as follows:

There is a peak phase between 10^{th} and 17^{th} hour where the transactions recorded for all search engines are higher than other cases. This corresponds to times 10AM and 5PM.

There are five hours that are noticeable for their deviation from the normal trend, namely, 1^{st}, 5^{th}, 14^{th}, 19th and 22^{nd} hours. This corresponds to 1AM, 5AM, 2PM, 7PM and 10PM. Since only one of these falls in the peak hour window, it would be interesting to study the contributing factors towards high activity outside the peak window.

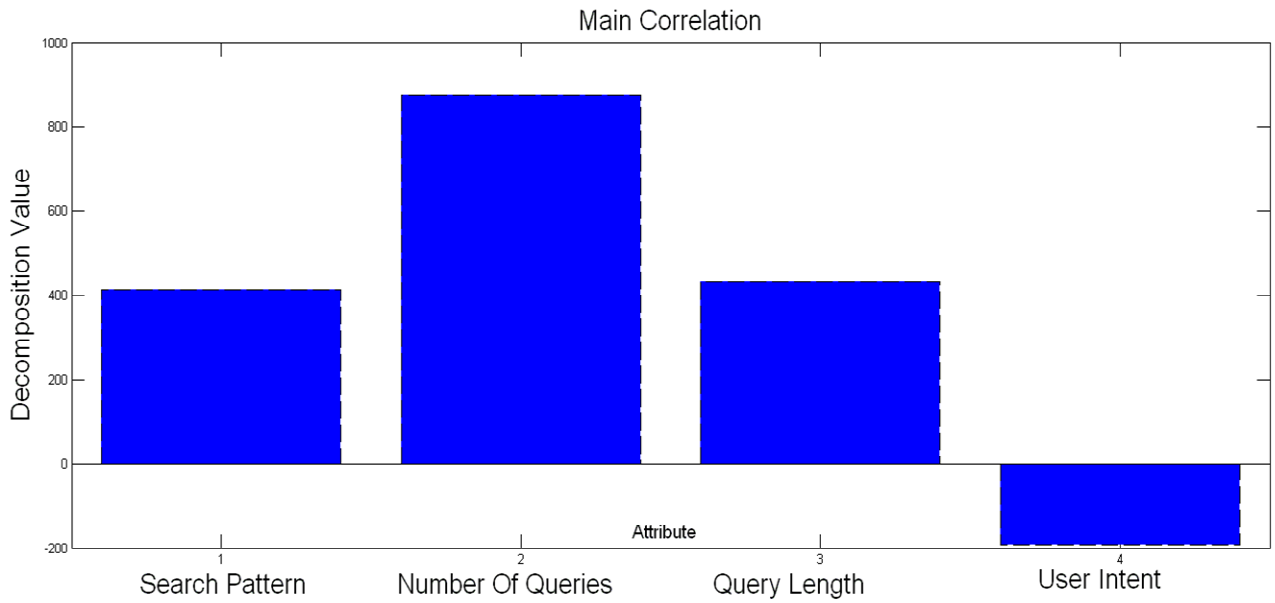


Figure **6-10**: **Main correlation between search attributes.**

For example, the 22nd hour (non-peak hour) was further analyzed to reason out the trend that appears in the plots. As mentioned earlier, the first factor of each dimension in the decomposed tensor corresponds to normal pattern and the second factor corresponds to some major anomalies Figure **6-10** and Figure **6-11** highlight the normal and abnormal correlations in that hour.

The first three attributes are positively correlated with each other and anti-correlated with *User Intent*. A higher value on *Number of queries* implies that searchers were manipulating the previous query in different ways rather than starting new queries. A lower value on *Query length* means these query modifications were such that query length did not increase much.

The positive correlation between *Search Pattern* and *Number of queries* is indeed meaningful. An anti-correlation with *User-Intent* gives us the indication that queries with

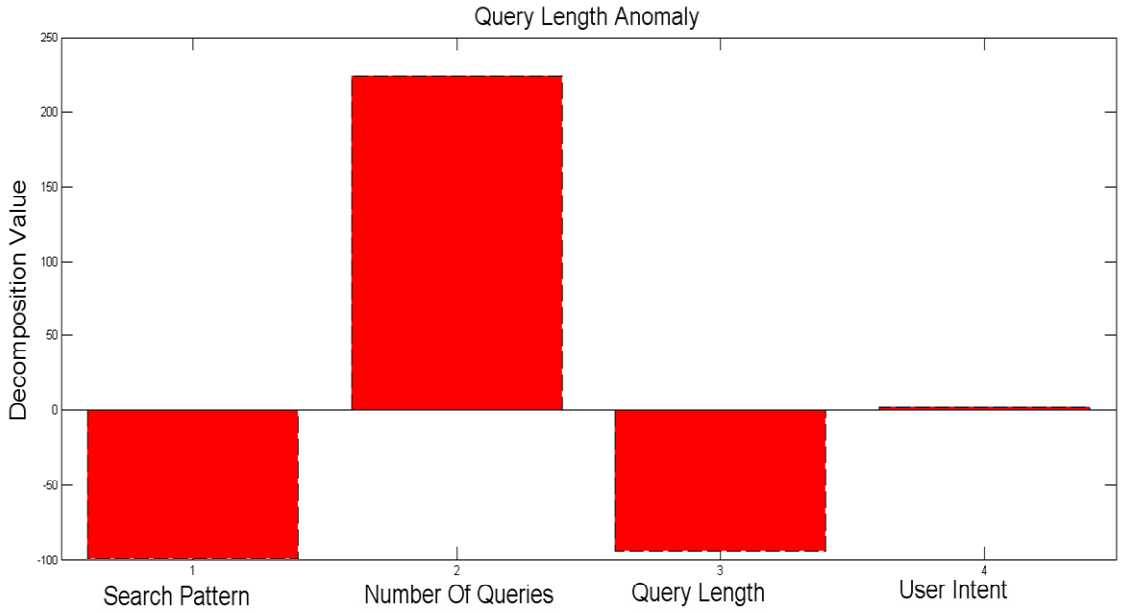


Figure **6-11**: **Anomalous correlation between search attributes.**

the above explained normal trend characteristics were more *informational* in nature than *transactional* or *navigational*. Figure **6-11** shows anomalous correlation between Search Pattern and Query Length with other attributes. Hence, all transactions with these characteristics can be considered as outliers in a transaction log data stream.

Peaks in the trend line for Search Pattern, Query Length and Number of Queries correspond to respective dips in User Intent. This suggests that higher the former three attributes, lower is the User Intent and the Web search is more informational in nature than transactional or navigational.

The experimentation and results prove that tensors are a convenient way to model the search engine transaction logs. A single main or dominating pattern and another

pattern that is deviating from main pattern was identified. Thus, there are recognizable patterns in the user-system interactions. The characteristics of the patterns are discussed in section 6.1.3. The characteristics of the data points that stand out from the rest of the data stream were recognized. Thus, the results from this experimentation have answered a large part of our research questions.

Chapter 7

Conclusions

Search engine transaction logs are potential sources for valuable information that promote effectiveness and efficiency of search engines. Mining this data requires novel methods. Statistical approach, probabilistic methods, and database methodologies have being applied in earlier studies; however, with limited results in dealing with massive, temporal datasets from multiple sources. This research aims at incorporating tensors as a data model and tensor analysis as a tool to mine this huge, temporal and multi-aspect data.

It may be noted that the transaction logs are taken from different time stamps (years) since this was the only data available at the time the study began. Although transaction logs from multiple-search engines for the same time window would be ideal, such data was difficult to acquire. Moreover, this study aims at showing that tensor is an intuitive data model and tensor decomposition is a viable method for TLA. Further, it is shown that meaningful conclusions can be mined from transaction logs.

Tensor is a promising data model for transaction log data representation. Prominent patterns in the transaction logs have been recognized. The variation of search attributes over the 24-hour period shows five peak values at five different hours and shows a consistent high search engine usage between 11Am and 6PM.

The experimentation was carried at two stages case

i) Hourly multiple tensor analysis and case

ii) Further correlation analysis at interesting time windows in non-peak activity hours. The hourly multiple tensor analysis provides hourly variations of search attributes with respect to each other for 24 hour period. The correlation analysis for selected time slots shows the normal and abnormal correlation between different search attributes.

It is observed that *User Intent* is anti-correlated with other attributes. Analyzing the 22nd hour tensor, which is one of the prominent peaks along the day, shows anomalous correlation between search attributes. Thus, this study was successful in identifying the characteristics of outlier data points in search engine transaction logs spanning several servers totaling around eight million queries. It would need further study and research to pin point transaction entries with these characteristics.

Although the data that was available for the study came from different time stamps, the work presented in this thesis has provided a new avenue for the application of tensor analysis in the temporal analysis of search engine transaction logs.

7.1 Future research

The results found during this experimentation are useful for commercial search engines companies to identify peak search engine usage, usage patterns, and outliers.

Pattern analysis can lead to better usage and searching systems. The characterization of outliers helps in identifying any abnormal activity occurring online. Outliers can lead to a variety of benefits, including possible harbingers of future trends. These efforts can lead to intelligent resource allocation.

As for methods, this research can be adopted for online monitoring system with functionalities to cater to dynamically generated data. The present research work is a static tensor analysis on transaction logs. This means the data was already available completely before the experiment was started. The dynamic real time tensor analysis would require a novel way of decomposing data as and when they are created. This would need the preprocessing stages to be real time as well. This kind of a set up will produce real time results to promote immediate actions as necessary.

Another avenue for future work is to construct and decompose the tensor data for whole 24 hours. This would involve all the transaction entries from all the search engine logs taken together for analysis. The advantage of such an experiment is that there is an influence of previous transactions entries on all the future ones .This kind of analysis gives an overall picture of the day's attribute correlation. The trend of variations are not important in such analysis but attribute correlation for the whole 24 hour period becomes the centre of experimentation.

The tensor analysis can be applied at different level of data analysis. The present work applied static tensor analysis on a very global level. Tensors can be constructed per session basis and attribute characteristics on a session level can be known. After recording the characteristics of an outlier point, it would be interesting to identify the

transactions with in the log that are outliers. This can eventually be developed to identify users for any illegal or criminal activities through a search engine.

Tensor analysis provides way for numerous possibilities with respect to mining a search engine transaction log. This methodology has to be tapped to maximum for leveraging Web search engine functionalities and associated industries (online advertisements). This would lead to a completely new outlook towards online monitoring systems for search engines.

References

1) Christopher D. Manning, P. R. a. H. S. (2008). Introduction to Information Retrieval. Cambridge University Press.

2) H. Chen and V. Dhar (1990). User misconceptions of online information retrieval systems. International journal of man-machine studies, 32(6), 673-692.

3) Cole, J. I., Suman, M., Schramm, P., Lunn, R., & Aquino, J. S. (2003). The ucla internet report : Surveying the digital future year three. Retrieved 1.2.2003.

4) Fox, S. (July 2002). The Pew Internet & American Life Project. Retrieved 15.10.2002.

5) Sullivan, D. (2006). Nielsen NetRatings Search Engine Ratings. http://searchenginewatch.com/showPage.html?page=2156451.

6) T.A., P. (1993). The history and development of transaction log analysis. Library hi tech,11(2), 41-66.

7) Bernard J. Jansen, A. S., Chris Blakely, Sherry Koshman (2007). Defining a Session on Web Search Engines. Journal of the American Society for Information Science and Technology 58(6), 862 - 871

8) Bernard J. Jansen, A. S., Jan Pedersen (2005). A temporal comparison of AltaVista Web searching: Research Articles. Journal of the American Society for Information Science and Technology 56(6), 559 - 570

9) Bader, T. G. K. a. B. W. (November 2007). Tensor decompositions and applications. Sandia National Laboratories. SAND2007-6702.

10) Saracevic.T, K. P., Chamis.A, Trivison.D (1988). A study of information seeking and retrieving. I. Background and methodology. American Society for Information Science, 39(3), 175-190.

11) Bernard J. Jansen, U. P. (2001). A review of Web searching studies and a framework for future research. Journal of the American Society for Information Science and Technology, 52(3), 235-246.

12) Kaske.N (1993). Research methodologies and transaction log analysis: Issues, questions, and a proposed model. Library Hi Tech, 11(2), 79-86.

13) Kurt.M (1993). The limits and limitations of transaction log analysis. Library Hi Tech, 11(2), 98-104.

14) Sandore.B. (1993). Applying the results of transaction log analysis. Library Hi Tech, 11(2), 87-97.

15) Banks.J (2000). Are transaction logs useful? A ten year study. Journal of Southern Academic and Special Librarianship 1(3).

16) Blecic, D., Bangalore, N.S., Dorsch, J.L., Henderson, C.L., Koenig, M.H. and Weller A.C (1998). Using transaction log analysis to improve OPAC retrieval results. College and Research Libraries 59(1), 11,39-50.

17) Hao-Ren Ke, R. K., Yu-Min Tai and Li-Chun Chen (2002). Exploring behavior of E-journal users in science and technology: Transaction log analysis of Elsevier's Science Direct OnSite in Taiwan. Library & Information Science Research 24(3), 265-291.

18) Shneiderman, B., Byrd, D., & Croft, W.B. (1998). Sorting out searching: a user-interface framework for text searches. Communications of the ACM.41(4),95-98.

19) He, D., Göker, A., & Harper, D.J. (2002). Combining evidence for automatic Web session identification. Information Processing & Management, 38(5), 727–742.

20) En Cheng, F. J., Lei Zhang, Hai Jin (2006). Scalable relevance feedback using click-through data for Web image retrieval. Proceedings of the 14th annual ACM international conference on Multimedia.173-176. http://portal.acm.org/toc.cfm?id=1180639&type=proceeding&coll=GUIDE&dl=GUIDE&CFID=57302368&CFTOKEN=28614962

21) Jaime Teevan, S. T. D., Eric Horvitz (2005). Personalizing search via automated analysis of interests and activities. Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, pp 449-456.

22) Ma, W.-Y. (2005). From relevance to intelligence: toward next generation Web search. International Multimedia Conference; Proceedings of the 7th ACM SIGMM international workshop on Multimedia information retrieval, pp 1-1.

23) Joachims, T. (2002). Evaluating Retrieval Performance using Clickthrough Data. Proceedings of the SIGIR Workshop on Mathematical/Formal Methods in Information Retrieval.

24) Broder, A. (2002). A taxonomy of Web search. ACM SIGIR Forum, 36 (2), 3-10.

25) Bernard J. Jansen, D. L. B., Amanda Spink - forthcoming (2007, July). Determining the informational, navigational and transactional intent of Web queries. Information Processing and Management: an International Journal.

26) Uichin Lee, Z. L., Junghoo Cho (2005). Automatic identification of user goals in Web search. Proceedings of the 14th international conference on World Wide Web, pp 391-400.

27) Jansen, B. J., Booth, D., and Spink, A. (Forthcoming) Determining the informational, navigational, and transactional intent of Web queries. Information Processing & Management

28) D. E. Rose and D. Levinson (2004). In Understanding user goals in Web search. Proceedings of the 13th international conference on World Wide Web. pp 13-19.

29) Quint, B. (1991). Inside a Searcher's Mind: The Seven Stages of an Online Search—part 1. Online Inc, 15(3),13-18.

30) Feng Qiu, J. C. (2006). Automatic identification of user interest for personalized search. Proceedings of the 15th international conference on World Wide Web, pp 727-736.

31) Jukka Perkio, W. B., Sami Perttu (2004). Exploring Independent Trends in a Topic-Based Search Engine. Web Intelligence archive WI '04: Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence, pp 664-668.

32) Alexandros Nanopoulos, Y. M., Maciej Zakrzewicz, Tadeusz Morzy (2002). Indexing Web access-logs for pattern queries. Proceedings of the 4th international workshop on Web information and data management, pp 63-68.

33) Nils Kammenhuber, J. L., Anja Feldmann, Gerhard Weikum (2006). Web search clickstreams. Proceedings of the 6th ACM SIGCOMM conference on Internet measurement, pp 245-250.

34) Kurt D. Fenstermacher , M. G. (2003, May). Client-side monitoring for Web mining. Journal of the American Society for Information Science and Technology, 54(7), 625-637.

35) Alan L. Montgomery , C. F. (2001, July). Identifying Web Browsing Trends and Patterns. Computer archive, 34(7), 94-95.

36) Michael Chau , X. F., Olivia R. Liu Sheng (2005 , August 31). Analysis of the query logs of a Web site search engine. Journal of the American Society for Information Science and Technology, 56(13),1363-1376.

37) Bernard J. Jansen, A. S. (2005, March). An analysis of Web searching by European AlltheWeb.com users. Information Processing and Management, 41(2), 361-381.

38) Bader, T. G. K. a. B. W. (November 2007). Tensor decompositions and applications. Sandia National Laboratories, technical report number - SAND2007-6702.

39) Lathauwer., L. d. (1997). Signal processing based on Multilinear Algebra. University of Leuven, Belgium. PhD thesis.

40) Bro, R. (2004, July). Practical Problems and Solutions in Applied Multiway Analysis. AIM Research Conference Center, ARCC Tensor Decomposition Workshop. Invited talk.

41) Hongcheng Wang, Q. W., Lin Shi ,Yizhou Yu,Narendra Ahuja (2005). Out-of-core tensor approximation of multi-dimensional matrices of visual data. ACM Transactions on Graphics (TOG), 24(3),527-535.

42) Kroonenberg, P. M. (1986). Three-mode Principal Component Analysis. Biometrics,42(1), 224-225.

43) R.Boque, A. K. S. (1999). Monitoring and diagnosing batch processes with multiway regression models. AIChE 45(7),1504-1520.

44) Levin, A. S. a. A. (2001). Linear image coding for regression and classification using tensor-rank principle. CVPR, 1(1), 42-49.

45) M. A. O. Vasilescu , D. T. (2004, August). Tensor Textures: Multilinear Image-Based Rendering. ACM Transactions on Graphics (TOG), 23(3), 336-342.

46) Jimeng Sun, S. P., Philip S. Yu (2006). Window-based Tensor Analysis on High-dimensional and Multi-aspect Streams. ICDM;Proceedings of the Sixth International Conference on Data Mining, pp 1076-1080.

47) Jian-Tao Sun, H.-J. Z., Huan Liu,Yuchang Lu,Zheng Chen (2005). CubeSVD: a novel approach to personalized Web search. Proceedings of the 14th international conference on World Wide Web, pp 382-390.

48 Jimeng Sun, S. P., Christos Faloutsos (2005). Online Latent variable detection in sensor networks. In Proceedings of the IEEE International Conference on Data engineering (ICDE), 1(2), 1126-1127.

49) Jimeng Sun, D. T., Christos Faloutsos (2006, August). Beyond Streams and Graphs: Dynamic Tensor Analysis. Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pp 374-383.

50) Faloitsos, K., Sun (2007). Mining Large time_evloving data using Matrix and Tensor tools. Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining tutorial.

51) Kolda, B. W. B. a. T. G. (2006). *MATLAB tensor classes for fast algorithm prototyping*. ACM Transactions on Mathematical Software, 32(4).

52) Kolda, T. G. (2006). Multilinear operators for higher-order decompositions. Technical Report Number SAND2006-2081, Sandia National Laboratories, Albuquerque, NM and Livermore, CA. *http://www.prod.sandia.gov/cgi-bin/techlib/access-control.pl/2006/062081.pdf*.

53) Tamara G. Kolda , B. W. B. (2007 , November). Tensor decompositions and applications. Technical Report Number SAND2007-6702, Sandia National Laboratories, Albuquerque, NM and Livermore, CA.

54) C.A. Andersson and R. Bro (2000). The N-way toolbox for MATLAB. Chemometrics and Intelligent Laboratory Systems,52(1),1-4.

55) S.Gourvenec, G.Tomsi, C.Durville, E.Di rescenzo, C.Saby, D.Massart, R.Bro, G.Oppenheim(2004). CuBatch, a MATLAB Interface for n-mode data Analysis. Chemometrics and Intelligent Laboratory Systems, 77(2005),122-130.

56) G. Kolda, B.W.B (2007 January).MATLAB tensor toolbox version 2.2. *http://csmr.ca.sandia.gov/~tgkolda/TensorToolbox*.

57) 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.

58) Heckerman, D. and Horvitz,E. (1998). In Infering Informational Goals from Free-Text Queries: A Bayesian Approach. Paper presented at Fourteenth Conference of Uncertainty in Artificial Intelligence, San Francisco, CA, USA, pp 230-237.

59) Steven M. Beitzel, E. C. J., Abdur Chowdhury , David Grossman ,Ophir Frieder (2004). Hourly analysis of a very large topically categorized Web query log. Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, pp 321-328.

60) Özmutlu, H. C., Spink, A., & Özmutlu, S. (2002). Analysis of large data logs: An application of poisson sampling on excite Web queries. information processing and management. Information Processing & Management, 38(4), 473-490.

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

62) J. Carroll , J. Chang (1970). Analysis of individual differences in multidimensional scaling via an *n*-way generalization of the ``eckhard-young'' composition. *Psychometrika,* 35, pp 283-319.

63) Harshman, R. (1970). Foundations of the PARAFAC procedure: Models and conditions for an "exploratory" multimodal factor analysis. *UCLA Working Papers in Phonetics*, 16, pp 1-84.

64) Ledyard Tucker (1966). Some mathematical notes on three-mode factor analysis. Psychometrika, 31(3), 279-311.

65) C. Bouveyron, S. Girard and C. Schmid (2007). High-dimensional data clustering. Computational Statistics & Data Analysis 52(1), 502-519.

66) Neil D. Lawrence (2003). Gaussian Process Latent Variable Models for Visualization of High Dimensional Data. Technical report.