Glossary

**Abandonment Rate** is a KPI that measures the percentage of visitors who got to that point on the site but decided not to perform the target action.

**Action** is an action is a specific utterance of the user.

**Action Object \((a, o)\) Pair** stands for action and \(o\) stands for object.

**Action-Object Pair Approach** is where one \((a, o)\) pair is an interaction between the user and the system. A series of \((a, o)\) pairs or \(a\)-\(o\) matrix can represent the interaction session, which is defined as a series of interactions between the user and the system to fulfill the user’s certain information need.

**Adaptive Hypermedia** is an approach to automatic personalization.

**Adaptive Prompting** is a context sensitive method of issuing diagnostics based on patterns of actions as well as individual actions by the user.

**Alignment-Centric Performance Management** is a method of defining a site’s business goals by choosing only a few key performance indicators.

**Analysis – First Order** is an analysis of transaction patterns in which state pairs are evaluated and the immediately previous state is used to predict the current state.

**Analysis – Higher Order** is an analysis of transaction patterns in which a sequence of states greater than two are evaluated and the current state is predicted on the basis of previous states (for example, a second-order process analysis would look at two previous states to predict the current state, a third order would look at three previous states, and so forth).

**Analysis of Variance** is a procedure, where the total variation in the dependent factor is partitioned into meaningful components.

**Analysis – Zero Order** is an analysis of transactions in which only the current state is evaluated. This is usually characterized by studies in which frequency counts of particular states are reported irrespective of their context.

**Anonymized Data** has been collected with identifying information, but has had subsequent removal of any links between the data and identifying information so that the researcher can no longer discern the specific owner of the data. Also, Anonymous data is the data that is collected without any associated identifying information.

**Argumentative Mining** is the detection of an argumentative structure in a discourse and the
recognition of its composing components such as the premises and conclusions of the argument; possibly the integration of the found arguments into a knowledge structure used for reasoning.

**Average Order Value** is a KPI that measures the total revenue to the total number of orders.

**Average Time on Site (ATOS):** See visit length.

**Background Knowledge** is the body of text, images, databases, or other data that is related to a particular machine learning classification task. The background knowledge may contain information about the classes; it may contain further examples; it may contain data about both examples and classes.

**Behavioral Targeting** is a technique used by online publishers and advertisers to increase the effectiveness of their campaigns. The idea is to observe a user’s online behavior anonymously and then serve the most relevant advertisement based on their behavior. Theoretically, this helps advertisers deliver their online advertisement to the users who are most likely to be influenced by them.

**Behaviorism** is a research approach that emphasizes the outward behavioral aspects of thought. For transaction log analysis, we take a more open view of behaviorism. In this more encompassing view, behaviorism emphasizes the observed behaviors without discounting the inner aspects that may accompany these outward behaviors.

**Bigram Analysis** is the analysis of all sequences of two adjacent words in each query.

**Bigram** is a group of two words or characters.

**Bilingual query** is a search query that contains words in more than one language.

**Biomedicine** is the broad domain of biology and health care including research and practice related to living organisms often focused on, but not limited to, human health and disease.

**Blog (short for Web Log)** is a Web based publication consisting primarily of periodic content. Also, Blog is shorthand for Weblog. A frequent and chronological publication of comments and thoughts on the Internet. It is a journal that may be instantly published to a host web site.

**Blog-Circuit** is the communicational circuit established among participants of different social networking sites, accomplished through links exchange and reciprocal visits.

**Blogger** is a person who creates and maintains a Weblog.

**Cache Busting** is a techniques used to prevent browsers or proxy servers from serving content from their cache, in order to force the browser or proxy server to fetch a fresh copy for each user request. Cache busting is used to provide a more accurate count of the number of requests from users.

**Cached Files** are some files that are saved and retrieved by browsers or proxy servers to save network resources.

**Chat** is also known as instant messaging. Allows people to communicate online by broadcasting messages to people in real time, often as one-on-one channel, but also in a group forum sometimes called a chat room.

**Checkout Conversion Rate** is a KPI that measures the percent of total visitors who begin the checkout process.

**Chinese Search Logs** contain the Chinese queries are often received in different character encodings. GB-2312, GBK, and BIG 5 are the three most popular Chinese language encoding schemes. They are used in different Chinese speaking regions with different popularity. For
example, Traditional Chinese, usually encoded in BIG 5, is widely used in Hong Kong and Taiwan, while Simplified Chinese, usually encoded in GB-2312, is more commonly used in mainland China and Singapore.

**Classification or Taxonomy** is a terminology where terms are arranged by “is_a” or “is_member_of” relationships into classes.

**Clickstream Data/Clicktrail** is the recording of Web pages that a computer user clicks on while Web browsing or using a personal computer.

**Clickstream Tracking** is the passive collection of data that computer users generated when they click the mouse on a Web site. A computer user’s “clickstream” is the list of events they have initiated by clicking their mouse.

**Client-Side Log** are all users’ computer activities saved in a client’s computer as a computer file.

**Client-Side Logging** is the software that records Web browsing behavior at the user’s computer. This is generally achieved either through a custom web browser or through browser plug-ins such as tool bars or browser helper objects.

**Commerce Website** is a type of Website where the goal is to get visitors to purchase goods or services directly from the site.

**Committed Visitor Index** is a KPI that measures the percentage of visitors that view more than one page or spend more than 1 minute on a site (these measurements should be adjusted according to site type).

**Computer-Mediated Communication (CMC)** is communications that occur via computer-mediated formats (i.e., Weblogs, instant messages, e-mails, chat rooms) between two or more individuals.

**Concepts** are cognitive constructs based on entities in the real world such as “nose” or “anatomy.”

**Conceptual Framework/Inquiry** is a methodology to build and use conceptual framework as a plan and direction for research.

**Conditional Random Field (CRF)** is a learning system for classification often used for labeling sequential data (such as natural language data); as a type of Markov random field, it is an undirected graphical model in which each vertex represents a random variable, whose distribution is to be inferred, and each edge represents a dependency between two variables.

**Conditional Random Fields** are a probabilistic framework for labeling and segmenting sequential data, based on conditional probabilities (Wallach, 2004)

**Connector Website** is a Website providing a relatively simple means of interaction for users who seek to offer or obtain goods, services, or information. It is an intermediary offering peer-to-peer web applications that collectively make up an infrastructure for social exchange, networking, and diffusion processes. Over time, user-to-user interactions gradually generate a majority portion of the website content and the regulation of which is governed jointly between the host organization and the online community of users. Depending on the surrounding social and economic conditions, as well as site design and development, the connector website should excel in facilitating the discovery and coordination of context-based communications and transactions.

**Consumer (of Healthcare)** is assumed to be a member of the lay public, as opposed to a researcher or clinician. Therefore, a consumer is not an expert in biomedical science or terminology.

**Content Analysis** is a methodical and replicable methodology to determine, quantify and analyze presence of research objects within large data sets.

**Content/Media Website** is a type of Website focused on advertising.
**Contextual Privacy** concerns: privacy concerns vary in any given instance according to the inherent privacy concerns of the user and the situational factors at play. These include the viewer of the information, level of control retained over the information, and the type of information. Furthermore, these factors can vary according to the device in use and the location.

**Conversion Rate** is a KPI that measures the percentage of total visitors to a Website that perform a specific action.

**Cookies (HTTP cookies or Web cookies)** are parcels of text left by a Website on the computer user’s hard disk drive; these data are then accessed by the Website’s computer server each time the user re-visits the Website. Cookies are used to authenticate, track, and maintain specific information about users, such as site preferences and the contents of their electronic shopping carts.

**Cost Per Lead (CPL)** is a KPI that measures the ratio of marketing expenses to total leads and shows how much it costs a company to generate a lead.

**Customer Loyalty** is a KPI that measures the ratio of new to existing customers.

**Customer Satisfaction Metrics** are KPI items that measures how the users rate their experience on a site.

**Demographics and System Statistics** is a metric that measures the physical location and information of the system used to access the Website.

**Density** describes the general level of linkage among the actors in a social network.

**Depth of Visit** is a KPI that measures the ratio between page views and visitors.

**Digital Environment** is the social setting produced through computer technology.

**Discourse Analysis** is a scientific argument evaluation method.

**Discussion Board** is also known as forum, message board, and bulletin board, for the purpose of exchanging information only. A Website location where users may post text communication for one another. Not sensitive to time constraints or structures.

**Domain Knowledge** is the knowledge possessed or required of a person or system within a specific topical area.

**Domain Ontologies** are the description of concepts and relations regarding some knowledge field.

**Electronic Commerce Research** is all forms of investigation of online selling of goods or services.

**Electronic Survey** is one in which a computer plays a major role in both the delivery of a survey to potential respondents and the collection of survey data from actual respondents.

**Entropy** is the measurement that can be used in machine learning on a set of data that is to be classified. In this setting it can be defined as the amount of uncertainty or randomness (or noise) in the data. If all data is classified with the same class, the entropy of that set would be 0. The entropy of a set \( T \) that has a probability distribution of classes \( \{p_1, p_2, \ldots, p_n\} \) can be defined as:

\[
-(p_1 \times \log(p_1) + p_2 \times \log(p_2) + \ldots + p_n \times \log(p_n))
\]

**Episode of Use** is a time frame used to measure a specific occasion of use.

**Ethnography** is a methodological technique for examining and understanding community life. Also, Ethnography is a research technique used traditionally in Anthropology, in which long term permanence of the researcher in the field and systematic description of social situations
provides data for analyzing the culture of a given group or society.

**Ethnomethodology** is a term coined by Harold Garfinkel in the 1960s, that refers to a branch of sociology that examines the ways in which people make sense of their world, share their understandings and produce collectively the social order in which they live.

**Ethogram** is an index of the behavioral patterns of a unit. An ethogram details the different forms of behavior that an actor displays. In most cases, it is desirable to create an ethogram in which the categories of behavior are objective, discrete, not overlapping with each other. The definitions of each behavior should be clear, detailed and distinguishable from each other. Ethograms can be as specific or general as the study or field warrants.

**Feedback** is Website “currency” that builds or detracts reputation for users or specific content. Within a website’s feedback system, for example, a user may give positive or negative point(s) to another user or that user’s posted content based on some interaction.

**Fieldwork Diary** is the research technique in which the ethnographer take systematic notes describing fieldwork situations.

**Fieldwork** is the ethnographic activity held in a given period and place in which the researcher collects data through direct contact with the group being studied.

**Flash Cookies** are similar to “cookies” (above), but coded with Macromedia Flash software; Flash cookies are more difficult to remove than traditional cookies, and as a result, they tend to be more reliable.

**Folksonomy** is a word combining “folk” and “taxonomy,” meaning the “people’s classification management”. Refers to the collaborative but unsophisticated way in which information is being categorized on the web. Instead of using a centralized form of classification, users are encouraged to assign freely chosen keywords (called tags) to pieces of information or data, a process known as tagging.

**Geo-Mapping** is a visual representation of the geographical location of Website visitors layered on top of map or satellite imagery.

**Group-Level Cohesion** can be used to identify who was communicating with whom in a discussion forum.

**Hidden Markov Model** is a stochastic process, where the underlying process or parameters are not observable, but can only be monitored through another stochastic process with observable parameters.

**Historical Method** collects and examines, and interprets facts about events, people and environment of the past.

**Identical Query** is a query within a session that is a copy of a previous query within that session.

**Individual Differences** are the demographic and psychological characteristics of people that distinguish one person from another.

**Information Extraction** is the identification, and consequent or concurrent classification and structuring into semantic classes, of specific information found in unstructured data sources providing additional aids to access and interpret the unstructured data by information systems.

**Information Gain** is the amount of information in a given set of data can be defined as \((1 - \text{entropy})\). If any observation about the given data is made, new information can then be recomputed. The difference between the two information values is the “information gain”. In other words, the change of entropy is the information that is gained by the observation.
Information Seeking Behavior refers to how people seek information in different contexts.

Informational Query is a query intended to satisfy a general information need, as opposed to an attempt to locate a specific article or group of articles (navigational query).

Inherent Privacy Concerns: An individual’s general privacy concerns; their disposition to privacy. Factors which may impact a person’s disposition to privacy include their age and computer experience.

Initial Query is the first query submitted in a session by a given user.

Instant Messenger is an online service that alerts users when friends or colleagues are online and allows them to communicate with each other in real time on a private online chat window.

Interaction Design is designing interactive systems that support certain functionality and a range a user behaviors.

Interaction Schema/model is a formalized description of interaction rules and actions allowed in specific contexts.

Interactional (in the field of Social Sciences) related to or property of the social interaction.

Interactions are the physical expressions of communication exchanges between the searcher and the system.

Interactive Information Systems (IIS) are capable of converging on a person’s information need by stages.

Internal Search is a metric that measures information on keywords and results pages viewed using a search engine embedded in the Website.

Interpretivism is a tradition in social and humanities research that assumes findings are to be interpreted by the researcher. This contrasts with positivism, which assumes the researcher “finds” or simply “observes” findings.

IS Research Methodologies refers to the common research methods used by information scientists.

Key Performance Indicator (KPI) is a combination of metrics tied to a business strategy.

Labeled Set is a set of item-label pairs. The item consists of an actual example that can be classified, and the label is the classification. In a supervised learning paradigm this set is sometimes referred to as the “training set”.

Lead Generation Website is a type of Website that is used to obtain user contact information in order to inform them of a company’s new products and developments, and to gather data for market research.

Lifestyle Time Frame is the general media use during the lifetime.

Log Analysis is the analysis of user behavior based on the actions recorded during interaction. In addition, log file analysis is a method of gathering metrics that uses information gathered from a log file to gather Website statistics. Also, Log File Analysis is the analyze of log files (Web server logs) to review the aggregate results.

Log File is a log kept by a Web server of information about requests made to the Website including (but not limited to) visitor IP address, date and time of the request, request page, referrer, and information on the visitor’s Web browser and operating system. Also, Logfile is a log kept by a Web server regarding registers left on a Website. Also, Log Files or Web Server Logs are files automatically created and maintained by a computer server on which a Website is hosted of the activity on that Website (traffic, hits, etc.). A typical example is a Web server log which maintains a history of page requests.
Logging Module/System is the component of an interactive system that logs/records relevant interaction between the user and the system (events, user actions, system responses).

Machine Learning is the area of artificial intelligence that studies the algorithms and processes that allow machines to learn. These algorithms use a combination of techniques to learn from examples, from prior knowledge, or from experience.

Markov Models or Chains are a stochastic process that considers a finite number of values and states.

Markov Process is a stochastic process in which the transition probabilities can be estimated on the basis of first order data. Such a process is also stationary in that probability estimates do not change across the sample (generally across time).

Maximum Entropy Model is a learning system used for classification that computes the probability distributions corresponding to an object and its class based on training examples, and that selects the one with maximum entropy, where the computed probability distributions satisfy the constraints set by the training examples.

Maximum Entropy Modeling is a methodology aiming to model random and stochastic events, that is motivated by the principle of generating probability distributions from a training dataset, and calculating the conditional probability that event y occurs given that event x has occurred.

Mediated Information Retrieval is a model of IR interaction in which the systems supports the user’s exploration of the information space and the formulation of queries.

MEDLINE is a database of biomedical literature created and maintained by the US National Library of Medicine (NLM, a unit of the National Institutes of Health). MEDLINE is a bibliographic database, meaning that it contains the reference information needed to find articles, but not the actual full-text articles.

MeSH stands for Medical Subject Headings.

Metrics is a statistical data collected from a Website such as number of unique visitors, most popular pages, etc.

Monte-Carlo Simulation is a static simulation scheme that employs random numbers, and is used for solving stochastic or deterministic problems, where time plays no substantial role.

Multi-Episode Segment of Time is the media use during the particular segments of the lifetime such as hours, weeks, or months.

Named Entity Recognition classifies named expressions in text (such as person, company, location or protein names).

Naturalistic Perspective is an empiricist approach of the Social Sciences based on the premise of collecting data essentially from “natural” situations, those that happen despite the presence or participation of the researcher.

Navigational Query is a query intended to locate a particular article or group of articles, as opposed to a general information need (informational query).

Nethnography refers to ethnographic research concerning digital environments.

Networked Communities are those support network-based communities that have few, if any, geographic and/or temporal boundaries, which VLSCs support.

Neural Networks are massively parallel distributed processor that has a natural propensity for storing experiential knowledge and making it available for use.
New Topic Identification is discovering when the user has switched from one topic to another during a single search session to group sequential log entries that are related to a common topic or session identification.

New Visitor is a user who is accessing a Website for the first time.

New Visitor Percentage is a KPI that measures the ratio of new visitors to unique visitors.

N-Gram Analysis is the analysis of all sequences of n adjacent words in each query.

N-Gram is a group of n words or characters.

Noun Phrase Coreferent is when two or more noun phrases are coreferent, when they refer to the same situation described in the text.

Object is an object is a self-contained information object, the receipt of the action.

Online Business Performance Management (OBPM) is a method of defining a site’s business goals that emphasizes the integration of business tools and Web analytics to make better decisions quickly in an ever-changing online environment.

Online Community (also known as virtual community) is a group of people communicating or interacting with each other by means of information technologies, typically the Internet, rather than face to face. Online communities can be used loosely for a variety of social groups interacting via the Internet. The concept does not necessarily mean that there is a strong bond among the members. The term virtual community is attributed to the book of the same title by Howard Rheingold in 1993. Also, Online or virtual communities are sets of people that interact primarily using information communication technology (e.g., listserv, email, social networking applications) instead of face to face.

Online Interaction is social interaction held by co-participants of a digital environment.

Open Coding involves reading and comparing individual data units so as to label similar units into categories.

Opinion Mining is the detection of the opinion or subjective assessment in a certain medium (mostly text) where the opinion is usually expressed towards a certain entity or an entity’s attribute; possibly the aggregation of the found opinions into a score that reflects the opinion of a community.

Order Conversion Rate is a KPI that measures the percent of total visitors who place an order on a Website.

Page Access is users’ one screen access to the Web server content.

Page Depth is a KPI that measures the ratio of page views for a specific page and the number of unique visitors to that page.

Page Requests are users’ requests to the Web server to send files to the users’ browser.

Page Tagging (Web Bug/Beacon) is an object that is embedded in a Web page or e-mail and is usually invisible to the user but allows checking that a user has viewed the page or e-mail. Also, Page tagging is a method of gathering metrics that uses an invisible image to detect when a page has been successfully loaded and then uses JavaScript to send information about the page and the visitor back to a remote server.

Parser is a software program which analyses the grammatical structure of a sentence according to the grammar of the language; a parser is often automatically trained from annotated examples; it captures the implied hierarchy of the input sentence and transforms it into a form suitable for further processing (e.g., a dependency tree).
**Glossary**

**Participant Observations** are the research strategies based on a close and intimate familiarity with a group and its practices in their natural environment, usually over an extended period of time.

**Part-of-Speech** is a word class or category (also called lexical class) which is generally defined by the syntactic or morphological behaviour of the word in question; common classes are noun, verb and adjective among others.

**Personalization** is the process that adjusts the results obtained by users when accessing Web systems.

**Phenomenology** is an interpretive methodology that examines users’ behavior.

**Poisson Sampling** is a useful random sampling process as it includes the properties of (1) Unbiased Sampling (2) Proportional Sampling (3) Comparability of Heterogeneous Poisson sampling Arrivals, and (4) Flexibility on the Stochastic Arrival Process from Which the Sample is Selected.

**Positivist Epistemology** (also referred to as positivism) refers to the school of research thought that sees observable evidence as the only form of defensible scientific findings. Positivist epistemology, therefore, assumes that only “facts” derived from the scientific method can make legitimate knowledge claims. It also assumes the researcher is separate from and not affecting the outcomes of research.

**Privacy** is the claim of an individual to determine what information about himself or herself should be known to others.

**Prospect Rate** is a KPI that measures the percentage of visitors who get to the point in a site where they can perform the target action (even if they do not actually complete it).

**Protocol Analysis** is the systematic evaluation of protocols using automated or manual content analysis tools.

**Protocol** is in this domain, a protocol is the “verbatim” record of user/system interaction for the entire user session (or selected portions) generally with time stamps on each action and perhaps some indication of system resources in use at the time.

**Proxy Logging** is the software that serves as an intermediary between the user’s web browser and the web site servers. Users generally have to log-in to the proxy and the proxy server can be used to augment retrieved web pages.

**Public Conversations** are those that are open and accessible to anyone. Conversations can occur between many individuals behind closed walls, e.g., a major company’s employees working on a major project.

**PubMed** is a freely-available interface onto MEDLINE created and maintained by the NLM.

**Query Clustering** is grouping the sequential log entries into different clusters in terms of topics or users.

**Query** is a string of terms submitted by a searcher in a given instance.

**Query Length** is the number of terms in the query. May or may not include stop words.

**Query Modification** is the modification by a search of a previous query.

**Question Answering (QA) Systems** go one step further than a typical information retrieval system that takes as input a user query and returns a ranked list of documents considered relevant to it. QA) systems return an exact answer extracted from one of the documents.

**Rating or Net Feedback** is an indicator of reputation on a particular Website.

**Referrers and Keyword Analysis** is a metric that measures which sites have directed traffic to the Website and which keywords visitors are using to find the Website.
**Regression** is an approach that generates a model characterizing the relationship between independent and dependent factors of a system from sample data representing a certain observable fact.

**Repeat Query** is a query submitted more than once during the data collection period, irrespective of the user.

**Repeat Visitor** is a user who has been to a Website before and is now returning.

**Research Methodology** is the general knowledge approaches to conducting and designing research.

**Research Methods** are specific approaches employed in research that are typically derived from the research questions or aims.

**Research Setting** is the social situations in which an ethnographer develops his/her fieldwork.

**Returning Visitors** is a KPI that measures the ratio of unique visitors to total visits.

**Review, also known as Testimonial, Bulletin, and Wall** is a structured discussion board that allows users to submit critical text about an idea, user, product, or message. Often supplements ratings. See Amazon.com.

**RSS Feed** is shorthand for Real Simple Syndication. A family of XML file formats for Web syndication used by news websites and blogs.

**Search Engine** is a software program that searches one or more databases and gathers the results related to the search query.

**Search Engine Referrals** are KPI metrics that measures the ratio of referrals to a site from specific search engines compared to the industry average.

**Search log Analysis (SLA)** is the use of data collected in a search log to investigate particular research questions concerning interactions among Web users, the Web search engine, or the Web content during searching episodes. Also, SLA is defined as the use of data collected in a search log to investigate particular research questions concerning interactions among Web users, the Web search engine, or the Web content during searching episodes.

**Search Log Analysis (SLA) Process** is a three stage process of collection, preparation and analysis.

**Search Log** is an electronic record of interactions that have occurred during a searching episode between a Web search engine and users searching for information on that Web search engine. Also, a search log is an electronic record of interactions that have occurred during a searching episode between a Web search engine and users searching for information on that Web search engine.

**Searching Episode** is a set of interactions between a user and a search engine. An searching episode is composed on one or more sessions.

**Semantic** is of or relating to meaning in language.

**Semantic Web** is the set of resources intended to improve the actual possibilities of Web applications.

**Server Logs:** See log files.

**Server-Side Logging** is the software that records Web browsing behavior at the server. Data collection is generally limited to navigation information.

**Server-Side Logs** are all users’ Web access activities on a Web server saved in a Web server as a computer file.

**Session Duration** is the period from the time of the first interaction to the time of the
Glossary

last interaction for a searcher interacting with a search engine.

**Session Identification** is discovering the group of sequential log entries that are related to a common user or topic or new topic identification.

**Session** is a set of sequentially or semantically related clicks. Also, Session is the series of queries submitted by a user during one interaction with the Web search engine.

**Session Length** is the number of queries submitted by a searcher during a defined period of interaction with the search engine.

**Single Access Ratio** is a KPI that measures the ratio of total single access pages (or pages where the visitor enters the site and exits immediately from the same page) to total entry pages.

**Site Overlay** is any type of content that is superimposed over a Web page; for the purpose of Web analytics, the site overlay typically shows click and conversion data superimposed over the links on a Web page.

**Social Interaction** is the mutual action and/or influence among co-participants of the same social situation.

**Social Network Analysis (SNA)** is a technique used to study the interactions between individuals in a community.

**Social Network Site** is a Web-based services that allow individuals to (1) construct a public or semi-public profile within a bounded system, (2) articulate a list of other users with whom they share a connection, and (3) view and traverse their list of connections and those made by others within the system. The nature and nomenclature of these connections may vary from site to site.

**Social Networking** is a term describing an online process. It is a website technology that allows users to search, identify, and communicate with other people as contacts, fitting closest to their specified preferences and criteria.

**Sociograms** visually convey relationships between actors. These sociograms make network structure explicit as collections of nodes with links that portray directionality and connection strength.

**Sociology of Computing** is a stream in sociology that researches the interactions between humans and computers as well as the social effects of using computers.

**State Diagram** (state charts) is the model of an interactive system that describes (i) a finite number of existence conditions, called states; (ii) the events accepted by the system in each state; (iii) the transitions from one state to another, triggered by an event; (iv) the actions associated with an event and/or state transition.

**Stickiness** is a KPI that measures how many people arrive at a homepage and proceed to traverse the rest of the site. Also, stickiness is a popular term for marketing a message. Short-term stickiness describes a website’s ability to keep a user on the website for as long as possible. Long-term stickiness refers to a website’s ability to motivate a user to return to that particular website.

**Stochastic Process** is a process that is probabilistic rather than deterministic in behavior. In the current context, a user state can be estimated but not determined with certainty when a sequence of previous states is available (e.g. a partial transaction log).

**Support Vector Machine (SVM)** is a learning system used for classification and regression that uses a hypothesis space of linear functions in a high dimensional feature space, trained with a learning algorithm from optimisation theory; special property of an SVM is that it simultaneously minimizes the empirical classification error and maximizes the geometric margin that
separates two classes; hence SVMs are known as maximum margin classifiers. Also, SVM is a methodology of statistical learning theory, which is based on generating functions from a set of labeled training data.

**Support/Self Service Website** is a type of Website that focuses on helping users find specialized answers for their particular problems.

**Survey Instruments** are a data collection procedure that one can use in a variety of research designs.

**Survey Research** is a method for gathering information by directly asking respondents about some aspect of themselves, others, objects, or their environment.

**Tag:** In the practice of collaborative categorization using freely chosen keywords, these are descriptors that individuals assign to objects. Tags can be used to specify properties of an object that are not obvious from the object itself. They can then be used to find objects with some desired set of properties, or to organize objects.

**Task** refers to the information seeking task an individual user experiences that instills a need for information and motivates the user to satisfy this information need through some sort of information seeking behavior. Task is the context surrounding a person’s information need.

**Term** is a linguistic label for concepts. Also: a series of characters separated by white space or other separator.

**Term Pair** is two terms that occur within the same query.

**Terminology** is a set of terms.

**Text classification** is the process of assigning classes (or labels) to textual data. Textual data can range from short phrases to much longer documents. Sometimes referred to as “text categorization”, a text classification task can be defined as follows: Given a set of documents \( D = \{d_1, d_2, \ldots, d_n\} \) and a set of classes \( C = \{c_1, c_2, \ldots, c_m\} \) assign a label from the set \( C \) to each element of set \( D \).

**TLA** is the study of electronically recorded interactions between on-line information retrieval systems and the persons who search for information found in those systems.

**Tokenization** breaks a text into tokens or words. It distinguishes words, components of multipart words and multiword expressions.

**Top Pages** is a metric that measures the pages in a Website that receive the most traffic.

**Topic Analysis** is analysis aiming to identify the topic of search engine queries.

**Topic Identification** is automatically identifying or estimating the topic of search engine queries without human intervention.

**Total Bounce Rate** is a KPI that measures the percentage of visitors who scan the site and then leave.

**Trace Data** or measures offer a sharp contrast to directly collected data. The greatest strength of trace data is that it is unobtrusive. The collection of the data does not interfere with the natural flow of behavior and events in the given context. Since the data is not directly collected, there is no observer present in the situation where the behaviors occur to affect the participants’ actions. Trace data is unique; as unobtrusive and nonreactive data it can make a very valuable research course of action. In the past, trace data was often time consuming to gather and process, making such data costly. With the advent of transaction logging software, trace data for the studying of behaviors of users and systems has really taken off.

**Traffic Concentration** is a KPI that measures the ratio of number of visitors to a certain area in a Website to total visitors.
**Glossary**

**Transaction** is a two-item set consisting of a query and a response, in which the IR system contributes either the query or the response and in which the response may be null. This definition allows human-to-machine, machine-to-human, and machine-to-machine transactions. It also allows for unanswered queries.

**Transaction Log Analysis** is a broad categorization of methods that covers several subcategorizations, including Web log analysis (i.e., analysis of Web system logs), blog analysis and search log analysis (analysis of search engine logs). In addition, transaction log analysis is the study of electronically recorded interactions between online information retrieval systems and the persons who search for information found in those systems.

**Transaction Log** is an autonomous file (or log) containing records of the individual transactions processed by a computerized IR system. Also, transaction logs are an electronic record of interactions that have occurred between a system and users of that system. These log files can come from a variety of computers and systems (Websites, OPAC, user computers, blogs, listserv, online newspapers, etc.), basically any application that can record the user – system – information interactions.

**Treebank** is a syntactically processed corpus that contains annotations of natural language data at various linguistic levels (word, phrase, clause and sentence levels). A treebank provides mainly the morphosyntactic and syntactic structure of the utterances within the corpus and consists of a bank of linguistic trees, thereby its name.

**UMLS** stands for Unified Medical Language System.

**Unique Term** is a term submitted one or more times in the data set.

**Unique Visit** is a single visit to a Website (regardless of if the user has previously visited the site); an alternative to unique visitors.

**Unique Visitor** is a specific user who accesses a Website.

**Unlabeled Set** is a set of examples whose labels or classes are unknown. If the class of an unlabeled example is learned, it can then be added to a “labeled set”.

**Unobtrusive Methods** are research practices that do not require the researcher to intrude in the context of the actors. Unobtrusive methods do not involve direct elicitation of data from the research participants or actors. This approach is in contrast to obtrusive methods such as laboratory experiments and surveys requiring that the researchers physically interject themselves into the environment being studied.

**User Behavior** is the set of actions taken by a user interacting with the system in order to reach a goal or complete a task.

**User Experience** refers to the immersive character of technology use and is typically evoked by designers of technology. The “user experience” is assumed to be architected by interaction designers.

**User** is a person who uses a computer system, software application, or website. Users may need to identify themselves for the purposes of accounting, security, logging and resource management. In order to identify oneself, a user has a user account and a user name, and in most cases also a password. Users employ the user interface for access to a system or website, and the process of identification is often referred to as log in.

**User profiles** are the set of information regarding user preferences, necessities and knowledge.

**Very Large-Scale Conversations (VLSCs)** are those that involve interchanges between hundreds and thousands of people. Newsgroups, chat forums, and Weblogs are examples of spaces where the volume of messages posted can range in the tens and hundreds of thousands.
**Visit Length** is a metric that measures total amount of time a visitor spends on the Website.

**Visit Value** is a KPI that measures the total number of visits to total revenue.

**Visitor Path** is a metric that measures the route a visitor uses to navigate through the Website.

**Visitor Type** is a metric that measures users who access a Website. Each user who visits the Website is a unique user. If it is a user’s first time to the Website, that visitor is a new visitor, and if it is not the user’s first time, that visitor is a repeat visitor.

**Web 2.0** is a second generation of Web-based communities and hosted services, such as social-networking sites, wikis and blogs, which facilitate collaboration and sharing between users.

**Web Analyst** is a job title used by private-sector practitioners, which typically involves analyzing Web-traffic data.

**Web Analytics** is the measurement of visitor behavior on a Website. Also, Web Analytics is the study of the behavior of Website visitors; the use of data collected from a Website to determine which aspects of the Website work towards the business objectives (for example, which landing pages encourage people to make a purchase).

**Web Browsing Behaviors** are user behaviors on the Web including their browsing activities and Web browser interactions. Privacy concerns have been found to impact Web browsing behaviors.

**Web Browsing Environment** is the context within which Web browsing occurs. For studies of Web usage this includes the Web browser and its associated tools (e.g., history, specialized toolbars), the task, and the motivation for conducting the browsing.

**Web Information Seeking** behavior refers to information seeking behaviors that occur over the Web. There are four main modes of information seeking on the Web ranging from wayward browsing to goal-directed search (undirected viewing, conditioned viewing, informal search, and formal search) where each mode is characterized by predominant information seeking moves or activities (undirected viewing: starting and chaining; conditioned viewing: browsing and differentiating; informal search: differentiating, monitoring, and extracting; and formal search: monitoring and extracting).

**Web Metrics** are a generic term for the many types of measurements that can be made about a Website and its visitors. Also, Web metrics pertains to the measures by which to assess a person’s Web information seeking behavior or to assess and monitor activity on a Website. Examples of commonly used Web metrics include page views, page transitions, and session times.

**Web Systems** are any application designed to be used on the Web.

**Web Tracking** refers to the automated collection of Web information seeking behavioral data.

**Web Usage Mining** is the set of techniques to generate patterns and discover knowledge from the web usage data.

**Weblog** is also known as blog and is a Web page in which the author(s) publish constantly updated contents. Entries are written in chronological order and commonly displayed in reverse chronological order, in the form of posts, usually describing personal experiences. A typical Weblog combines text, images, links to other Weblogs, Webpages and other media, and provides means of interaction with readers, using e-mail, guestbook or comments linked to single posts.

**Webmail** is email received and sent only locally on a particular website. The user’s other email accounts remain unaffected.
**Glossary**

**Widget** is a Web widget is a portable chunk of code that can be installed and executed within any separate HTML-based web page by an end user without requiring additional compilation. They are derived from the idea of reusable code that has existed for years. Nowadays other terms used to describe web widgets including: gadget, badge, module, capsule, snippet, mini and flake. Web widgets often but not always use DHTML, Adobe Flash or JavaScript programming languages.

**Wiki** is a series of Web pages that allows users to generate content, but also allows others (often unrestricted) to edit the content. A tool for online collaboration and without constraints of time.

**Zipf Distribution** is a distribution in which the frequency of any object is inversely proportional to its frequency rank. It has been observed in text corpora, database contents, and other natural phenomena.