Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

School of Computer Science and Statistics

The Search for the Searcher: Features influencing the Knowledge Gain by the User

Yash Mundra
16338461

April 25, 2019

A Dissertation submitted in partial fulfilment
Of the requirements for the degree of
Master in Arte Ingeniaria at
Trinity College Dublin
The University of Dublin
With the supervision of
Prof. Seamus Lawless and Dr. Annalina Caputo

Submitted to the University of Dublin, Trinity College, April 2019

Declaration

I, Yash Mundra, declare that the following dissertation, except where otherwise stated, is entirely my own work; that it has not previously been submitted as an exercise for a degree, either in Trinity College Dublin, or in any other University; and that the library may lend or copy it or any part thereof on request.

I have read and I understand the plagiarism provisions in the General Regulations of the University Calendar for the current year, found at http://www.tcd.ie/calendar.

I have also completed the Online Tutorial on avoiding plagiarism 'Ready Steady Write', located at http://tcd-ie.libguides.com/plagiarism/ready-steady-write.

Signed:_____ Date: _____

Summary

The digital presence and accessibility made the internet connection viable to all irrespective of ages. The purpose of exploring internet is the need for information. People use search engine for getting knowledge in a particular area or new topic. It may even be for upgrading their knowledge on a specific topic. They use to search by posing some questions in the search engine and these queries are informational queries. Though the search machine presents many results, there is no appropriate quantification about the gaining of knowledge by the searchers in a particular search session. The search engine can be tailor made once this factor is known. This paper intends to design such prediction model and aims to analyse the impact of features on the users' knowledge gain using search log analysis. The feature selection for selecting only the imperative features are performed using three models of regression. Two different sets of feature are considered for analysis to have a comparative analysis and to ensure the effect of features on the user's knowledge gain by an efficient prediction model. The set 1 will carry more features that are considered important and have impact on the model and the other set will carry the features selected by an appropriate regression model. These set of features are fed to various models of classification and the performance of models are evaluated using various performance indices. The prediction models are premeditated to have three different labels (low, medium, and high) with regard to the knowledge gained by the user. The model is validated using openly available crowdsourcing dataset. Majority of the designed models show promising results.

Acknowledgements

I would like to thank my project supervisors, Seamus Lawless and Annalina Caputo for lending me their support and expertise throughout the course of the project.

I would also like to thank my Dad, Mom and brother for their constant support and for always believing in me.

Contents

List of Figures

List of Tables

1 Introduction

One of the most often online actions nowadays is browsing the web for want of information. An internet search query is nothing, but a query related to a search term entered by the users in a particular search engine to fulfil their needs. The typical search queries can be fit into 3 various categories viz. navigational, transactional and informational queries (Andrei Broder, 2002). A query of the navigational kind is intended to get a particular website or webpage. That is, instead of using URL in the navigation bar, the user types the name of the website in the search bar. A transaction sort of query focuses on the completion of the transaction (payment) as in the case of buying a product. The informational queries are rather special and comprise of queries enclosing comprehensive topics to which multiple appropriate results could be fetched. And user usually prefers to get the wanted results within the first two pages.

The research on “Search as Learning” (SAL) has emphasized the significance of learning possibilities and intended on identifying the needs of learning while searching the web. With permitted internet connections as well as due to the affordability and accessibility of internet connections, people of current world generation prefer to search online for information and associated achievement of knowledge (Eickhoff et al., 2014). Earlier works have proven the prominence of learning as an inherent outcome of web search. A current assessment demonstrates that 80% of the general public will have a digital existence online by the year 2023. It is inferred therefore that many people will be spending their time on the Internet than ever to gain or expand their knowledge.

In a survey conducted by Salesforce involving about 7000 consumers, it is seen that 57% of them are ready to share their private data in exchange for custom made online experiences. A better experience can be made with personalization of their search sessions. Anyhow, the plan of understanding the degree of learning throughout the search session is not explored greatly. Modern search engines on the web is optimized for relevancy. If the extent of knowledge of a user with respect to a particular topic is also appended in the model, it would be an essential advancing phase of knowledge gain. In the current period, the systems for searching have incorporated the preferences of users, present context and latest behavior to design the interests and purposes of users with higher accuracy (Teevan et al., 2018).

1.1 Motivation

Some of the previous works have analysed the search log for enhancing the learning experience and to ease the learners who are not proficient in the techniques of searching to retrieve the information at a faster rate. The search log analysis, which is otherwise known to be transaction log analysis, is an electronic record of interfaces/communications that have happened all through the searching incident among an internet search engine and users probing for info and evidence on that particular search machine. The web server record and stock these communications on the server with the help of a software application based on the format of the file (Jiang et al., 2013a). The various format of the log is the access log, log of referrer, extended log etc. The usual format for search log is the extended format and holds regarding Internet Protocol (IP) address of client's system, query by the user, accessing time of search engine, and sites referred (Joshila Grace et al., 2011). The recorded data during a search in the server can be analysed for any useful information. This kind of investigation is generally called a search log analysis. The objective is to have better insight into the searcher interactions, details/essence, etc. This would assist in the point of achieving improved system design and the knowledge gained by the user by recognizing the searching behaviour of the user towards getting information.

There is no much focus on the extent of the knowledge improvement by the users. Researching with this focus would definitely help to design the search machine as per the knowledge requirement of distinct users. The basic idea behind is that the search engines should act in a different way for different user's subject to their level of knowledge on a specific topic. There exists a need for finding the impact of requirements for information on the search actions and knowledge enhancement of users. The search analysis, if serves as an indicator of quantifying the knowledge acquired by the user during their search, it could customize the search engine as per the knowledge requirements of the distinct users.

1.2 Objective and Research Questions

The main objective of this work is to examine the possibility of how the transaction (search) log of informational search could offer enough data to validate that the user is trying to enhance their knowledge or any insight could be derived through the log analysis during search session about how distinct users expand their knowledge of a topic/title during the period of search?

The meaning or definition of knowledge in the context of "information search" is different from conventional meaning. The focus of this work is not lying on the estimation of log analysis at signifying the level of knowledge. Also, there is no survey involved to judge the level of knowledge of the users. The information of transaction log is utilized just to acquire

an implication of the expertise of the users and the way it influences their behavior of search. The emphasis will rather be on the difficulty/ intricacy of the query. Moreover, it is assumed that an experienced/proficient, knowledgeable (about the topic) user will issue specific queries with higher complexity levels.

The following research questions will be examined to accomplish the objective of this work.

- Q1: Does query complication intensify during a search period and is there an affirmative association amongst complexity of the query and explicit topic?
- If it can be revealed that a user keeps on modifying the query with higher complexity, then it can also be proved that the user gains knowledge on that topic.
- Q2: Do the queries of a higher level of complexity increase the number of Clicks?
- It is assumed that the more the number of clicks, more will be the learning of appropriate documents by the user. Thus, a rise in a number of clicks relates to an adequate search for the user.
- Q3: Do the subjects of queries or clicked Uniform Resource Locators unite to one definite topic during a session?

A single session for the search of information is expected to be connected with a particular data need. The convergence of the topic within a search period implies that the user directs his/her queries on a definite topic. The behaviour of a particular user can be understood by investigating the recurrent reformulation of queries by the user.

- Q4: To what extent can a supervised machine learning model be trained to predict knowledge gain in a search session using features captured during those sessions?

The prediction would be better if the supervised model is trained with the significant features mined through the search session.

1.3 Choosing a dataset

The dissertation will be structured as follows: firstly, the background literature section, focussing on previous research discoveries that will aid in answering the research questions, including past research that was done with the AOL dataset. Then the data preparation techniques are outlined, as well as an in-depth look at the AOL dataset. This section presents some general statistics about the available data, and shows how the data was reduced for the purposes of this study. Following from this, the data analysis is discussed; how the results were obtained and what specific formulae were used. Finally, the results are presented and discussed.

For evaluating the knowledge of the user during search sessions, a dataset has been collected (Yu et al., 2018b) using crowdsourcing, in which, the users involved themselves in real practical search sessions. The dataset encompasses the search history of 468 individual users and the total number of topics searched during this process is 11. This dataset is publicized by the authors and available in the public domain. The same dataset is utilized for the proposed work.

1.4 Dissertation Structure

Chapter 1 outlines the background relevance and motivation to conduct the study to examine the extent to which user's expertise level of effect their knowledge gain for a topic on informational search sessions. Highlighting implications of few key studies conducted in this context, this chapter establishes research question to be answered to find out the effect of expertise level on search results. Goal of the dissertation and data collection source were also outlined along with design of the study. Chapter 2 discusses theoretical and conceptual work in the area of web search activity log and their impact on knowledge gain. A discussion has been developed to evaluate potential impact of different capability level to make effective informational web search and ability to obtain required knowledge. Range of recent of prior studies has been reviewed in this chapter. Chapter 3 presents research methods and outlines the design proposed to make quantifiable investigation on dataset collected from CrowdFlower platform. Detail about the sample size of the dataset and design to analyse data have been discussed in this chapter. Chapter 4 discusses about the dataset used in this study in detail exploring the impact of search sessions of 11 topics on knowledge gain of users. This chapter establishes an empirical model to evaluate the impact of user's knowledge state or expertise level on knowledge gain of a topic. Different activities and informational search sessions are aggregated and filtered to produce relevant results to be analysed and tested. Chapter 5 holds a key place in the study as it analyses the data collected from similar dataset used in the study by Yu et al (2018) through statistical analysis using liner regression methods. Different factors have been identified as variables to investigate the extent of user's knowledge level on their knowledge gain during web search sessions. Chapter 6 summarizes the dissertation presenting final thoughts on the topic and establishing key outcomes in relation to research objectives that led the study to answer the research question. In addition, limitations encountered during the study have been highlighted. More importantly, future scope of further work in related area has also been discussed.

2 Literature Review

Searching through the World Wide Web by public is a common phenomenon in current age to attain knowledge for fulfilling the objective concerned with learning. Anyhow, the progress of learning through search process is almost not known. It is exceedingly helpful to recognize the web user, the reason for their searching and the kind of interaction they do with the site for searching for supporting web search users (Li et al., 2017). If the user is new to the field of search, the following questions should be raised for guiding and helping the user leading to considerable knowledge gain after the search (Hendahewa & Shah, 2015). Does the entry level user learn about the topic of interest gradually during search? Is there any refinement for making precise questions with the new information? In case of professional scholar, the question set may be of the following format: Does he or she present targeted and specific queries while he/she collects sources for study? How does the change reflect while changing the expertise and dealings of the user?

Analysing the behaviour of the user on the web has been an eminent area of research since the inception of internet (Sanchiz et al., 2019). In fact, even previously, it has been examined how users inquired a database, when reformulation of queries was carried out by manually referring the words in a lexicon (Hagen et al., 2016). But in present age and time, while issuing queries, the search engines itself are providing suggestions to complete the queries. For instance, the moment the user types a query, the next word/words in sequence are provided to him as suggestions. Once the query is offered, the engine expands the query by suggesting suitable words that may come subsequently based on the previous searches even without the knowledge of the user. Then the results are displayed to the user based on the ranking decided by the history of search. The preferred pages with highest ranking order are displayed first (Ren et al., 2018).

As search engines are becoming smarter, the Users are adapting themselves to the changes and they also change their style of interacting with computer. The queries are more and more presented as language phrases instead of keywords, nowadays. The reason behind this is the technology called voice to search, where the spoken words are translated to query of search by the engine. This implies that the punctuation and stop words are added to keywords of searching quires. The algorithms for search engine are incorporated with appropriate technologies and concepts. The user's expertise and knowledge gained after the search will fetch correct prediction during search. The researches associated to knowledge gain by the users during the search are discussed in this section. Most of the recent researches focus on search / transaction log analysis to distinguish between experts and

entry level users while the field level of expertise is well-known.

2.1 Search Log Analysis

Transaction or search log is nothing but a file/ folder (log) of the transactions among computer and its user. Also, the time wise record of communications between searcher and the search engine of the web maintained in electronic format is called as Search log. It holds the details of searchers such as session ID and IP. In addition, the search information that covers the stream of search, operators, terminologies etc. will also be available in the log. The behaviour of s of a particular person while searching for information can be assessed through the information available in the transaction log (Choo et al., 2000).

Enormous amount of data with regard to search log has mounted up in web search engines. A widely held search engine, at present, can accept countless queries and gather records of terabytes size regarding day-to-day search behaviours of users. The authors has collected search log data as well as browse log data for analysis. The browse log data have been acquired through browser plugins of client. Even though these mammoth data of search and browse logs offer boundless chances of mining the knowledge of masses and enhancing web search, cleansing, processing and modelling log data are quite challenging and makes a call for competent and effective methodologies (Jiang et al., 2013b).

As per the opinion of (Wang et al., 2000), search log are truthful, inconspicuous, longitudinal, transactional, time-based and can be gathered and treated automatically. Few of the demerits identified are the data though accurate or not is available for interpretation, difficulty in managing the huge amount of data, and the privacy issues. The shortcomings could be overridden by adopting a suitable methodology. It is the skill of the searcher to investigate the log and ascertain the correctness. In addition to these drawbacks, the access to the log is restricted and obtaining the same is expensive. Thus, researchers can create search log by themselves.

The understanding about searching for information during an online search can well be perceived from the data stockpiled in transaction logs of search engines, web sites, and intranets. This knowledge in turn uplifts the model of information system and provides leads for developing and creating the structural design of information for searching contents (Jansen, 2006).
There are various factors are to be analysed in a search log for assessing the knowledge gain of the user. The factors helping to judge the knowledge gain are generally called as features. The search session is one of the features that influences the users' knowledge gain. The sessions for user search are designed as a procedure containing four phases viz., the formulation of the query, choice of result, reformulation and ending (Mat Hassan & Levene,

2005).

2.2.1 Importance

The understanding about searching for information during an online search can well be perceived from the data stockpiled in transaction logs of search engines of internet, web sites, and intranets. This knowledge in turn uplifts the model of information system, provides leads to developing and creating the structural design of information for collecting contents (Jansen, 2006).

2.2.2 Behaviour of User towards searching Information

The sessions for user search is designed as a procedure containing four phases viz., the formulation of Query, choice, reformulation and ending. The arrangements of the phases are stowed in a multi path digital tree named trie (Mat Hassan & Levene, 2005). A trie is utilized for stowing data provided with key and values. The key is used to recognize the data and the value keeps extra info associated.

2.2 Studies Related to Expertise of the User

Knowledge, as per the meaning given in thesaurus, is stated as “facts, information, and skills acquired through experience or education; the theoretical or practical understanding of a subject”. Expertise is otherwise indicated as “expert skill or knowledge in a particular field”. This description is somewhat extensive and indicates that the phrase ‘knowledge’ can differ in sense based on the usage in the context. In the context of search machines, knowledge refers to the awareness and understanding of the subject/topic/area/field/domain in which, the operator is searching, however it may even mean experience by means of a searching machine. Thus, it becomes significant to state openly what knowledge is with respect to the included data of the logs of searching machine.

The web search can be improved by examining the sort of knowledge that fetches correct information, the structures and stratagems of knowledge involved. Hölscher and Strube, (2000) presented 2 different experiments with various methods and various standpoints. In one of the tests, 12 candidates those who have strong knowledge in using internets, are interrogated about the stratagems involved in searching and they are asked to perform practical tasks of searching on the web. An information quest model is derived out of this test and validated in the next experimentation. A comparison has been done directly with two hypothetical knowledge types here. The impacts of internet expertise and subject expert background are examined with continuous tasks of searching in the field of economics. A differential and joint influence of both type of knowledge (web and subject) has been revealed through the test. The success of search performance relies on the collective expertise both on the subject and web. Anyhow, the stratagems pertaining to the

above knowledge could be identified individually.

Malik and Mahmood, (2009) considered various aspects with regard to knowledge gain and success of retrieval of relevant information for understanding the searching behaviour of Punjab university students, Lahore. The aspects on which, the results attained are, the background of the user, user's web know-how (expertise), skills towards searching, purpose behind using, formulation of queries, regularity of use, preferred searching machine etc. The data pertaining to the experimentation have been collected from the students in the discipline of Economics and Management Sciences in the university by questionnaire. The results revealed that most of the students used web for academic related work and their preferred search engine is Google. The search success was high with query reformulation, advance search features, surfing with initial ten hits. The searching speed also affects the information search. Query reformulation helped the students to get appropriate results at a faster rate.

A similar study has been adopted with respect to online learning platform. A learning folder can be created and the teaching sessions along with learners' response will be recorded and maintained in the folder online. This enables the teacher to analyse and assess the performance of learning of students. This stands as a feedback to the teacher and the student, student specific assistance in teaching can be appended then and there by the teacher. Hwang et al.,(2008) suggested a Meta-Analysers model for aiding the teachers for student analysis on the basis of search machine usage to solve the given problem by students. In a real time experiments, 220 students and fifty-four teachers have been considered.

The search performance and gratification of entry level and established users has been found with Scopus and Web of Science interfaces. Okhovati et al.,(2016) conducted an extensive investigative study with snowball sampling. The testing platform for the aimed study included equal number of beginners and experts (15 each) and Camstia has been applied to record the performance of searching. The satisfaction/gratification level of users has been found with Questionnaire for User Interface Satisfaction (QUIS). The expert users were found to be with higher level of satisfaction than beginners. Same kind of results had been experienced for both Scopus and Web of Science. It is implied that even a minimum experience in searching would provide more assistances in terms of locating the information. Better outcomes could be attained with appropriately modelled interface for the users.

Another study introduces an expert support exercising system for searching sessions of beginners in the aspect of analysing the knowledge gain. The ability of this system has been investigated on how the scheme assists the beginners, that is, entry level post graduate students and nurtures the capacity of searching towards information. The scheme involves a

quasi-experimental model prepared through 18 months in which, 8 entry level doctoral candidates partaken in five sessions at a stretch, with qualified expert. The differences in the search of information pertaining to beginners and expert have been examined to get the impact of training sessions through which the beginners are guided by the expert. The formulation of complex queries, keywords selection and the use of operators are missing with beginners and this is the reason why the users are not gaining as much as knowledge on par with the searching time. Thus, additional library/ archive programmes can support the beginners to improve the literacy expertise on information (Theng et al., 2015).

One more literature also emphasized about the expertise needed for knowledge gain during search sessions. White et al., (2009) described a big-scale, record-based and longitudinal analysis on the influence of field proficiency on searching activities pertaining to medicinal, legal, financial, and computer science areas. The domain specialists and amateurs are presented with characterized/customized tests on the basis of the nature of questions, sessions for searching, visited sites, success of search etc. The analysis centres on the vocabulary, search job, usage of resource under realistic circumstances. The authors carried out the process of segregation on the basis of one or more particular websites related to the area of subject. The characterization of experts is done through various attributes such as the number of visited pages, number of questions posted, time span of the session, revisited pages, number of unique domains the proportion of querying with respect to surfing etc. The users with domain expertise were able to use advance features and could obtain proper results than the users without domain expertise.

2.3 Studies Related to Query Analysis

Query logs will contain a log of queries posed by the users. More precisely, Query log gives a clear picture about what and what not the users are finding.

A study is focused on the query log analysis particularly for the needs of information relevant to children, Duarte Torres et al., (2010) have utilized a big scale query log for the purpose of fulfilling the info needs of kids by the way of investigating the sessions and queries. It has been carried out with two main analyses. The first one intends to recognize the variations amongst such type of questions & sessions from common questions & sessions. The next one aims to improve the query log by means of including annotations of questions, actions and sessions for children's info retrieval in future. The results yielded statistical differences among the general set and a dedicated set of children's requests.

Appropriate retrieving tools help to capture pertinent documents according to the users' need for information and to use web data commendably. . Web search engines though matched well with documents the designed is not much compatible with regard to data. To address this, Kacprzak et al., (2017) have suggested an analysis of query log for the search

database, depending on the logs of 4 public data portal at national level. It is focused to gain improved understanding of standard users of said portals as well as the queries offered by them and propose a model for improved data search from the findings. The structure and dimension of data provided through data portal is different from that of searching machines. The results infer that the portal is utilized in an explorative manner than response attentive queries. This necessitates powerful technology for retrieval to enhance the data portal knowledge of users.

2.4 Studies Related to Topic Analysis

In search information analysis, one of the distinct factors is the query topic. Chau et al., (2005) have considered the hunt logs of searching engine of web site of Utah state government. The general purpose search engines and websites are the same in the aspect of amount of search terms but are different as per the indicators like topics of search and the terminologies used. This is probably due to the different needs of information with respect to website and general search machines. The result findings could improve the web services and can enhance the research aspects of scholars. This sort of analysis can be applicable to e-government research through which, the delivery of info to the users with government websites could be examined.

By analysing the public browsing, it is understood that many people use only few topics/terms to search, very few have altered the query, they have viewed only few pages in website and very few in the public utilized advanced features in the search (Spink et al., 2001). Though the number of terms is less, the results reveal that certain terms have been used very often and most of the used terms are distinct. Coming to language usage, a variety could be found. The queries related to entertaining and recreation is ranking top the list.

2.5 Studies Related to Classification

Classification plays a vital role in search log analysis, especially in the query classification. In fact, the success of the system relies on the type of the classification adopted. (Trevisan et al., 2012) suggested a query log analysis with “GALATEAS LangLog”. This system mines the hits and queries from log files and performs pre-processing techniques like identification of language, tokenization, lemmatization etc. It is then followed by the important process like classification and clustering. Clustering and Classification permit the content providers for recognizing the type of the users and the queries target the information. The queries are classified into various types as per the plan for classification. When the plan of content provider and the classifier are different, hints of non-matching items of website are provided during classification. The analysis by the use of cluster could be applied to recognize fresh segments of market or current trends identified in the behaviour of the

user. The flexibility yielded through cluster is higher than classification. Anyhow, the output or the information is more precise in the case of classification. The authors have used hierarchical involving k-means clustering and soft clustering involving Latent Dirichlet Allocation (LDA). The classification methods used comprises of both unsupervised and supervised classification. The Naïve Bayes classifier was used a set of training data through supervised classification, while, the unsupervised method neither required training data nor the corpus related to a particular domain. The classification methods have been evaluated using the metrics F-Score.

(Bennett et al., 2010) suggested a query log analysis by the way of ranking them with effective classification methods. The classifiers applied for ranking is the Logistic Regression Classifier along with L2 regularizer. The ratio used for training and validation is 70:30 respectively. The metric named Normalized Discounted Cumulative Gain has assessed the ranking performance. The User Query classification is a perplexing task because the queries presented are small in general and most of the time; they are not clear and distinct. They contain noise and to be enriched. Few literatures have suggested a feedback mechanism with pseudo-relevance. Though it yields superior performance, the classification happens after the retrieval because the feedback should be derived from the output. There are certain applications that necessitate the classification process either simultaneously or before the retrieval. Zhu et al., (2009), to address this issue, performed a hybrid plan through which the training phase utilizes the relevance feedback and the testing phase is carried on the basis of secondary sources. In other words, the classifier that is built offline through data for training utilizes the top results for training and the same could not be utilized in testing phase. Obviously, there exists an asymmetry between training and testing. Moreover, for classifying queries, Support Vector Machines (SVM) using Liblinear2 toolkit is adapted. The toolkit is an open to public package to unravel big scale linear standardized problems of classification.

Other study for analysis employs automatic classification for web search involving naive bayes classifier, decision trees, multi-layer perceptron classifier, SVM, and logit model classifiers. Though the performances of the selected classifiers are comparable, the authors claim that the logit model is superior to others (Bhatia et al., 2012).

2.6 Other Related Studies

Gadiraju et al., (2018) have offered a study to know about the knowledge gained by the users after the sessions of searching for information in the internet. The authors hired 500 individual users with the help of crowdsourcing and arranged practical sessions for searching with 10 topics covering various domains and info requirements. Crowdsourcing is a popular platform through which, opinion, facts and figures are abstracted from people of huge group, who present data by means of several social platforms including internet. The

knowledge of net operators is calibrated/ standardized prior to and after the sessions of searching using knowledge experiments that are designed systematically. This process helps to quantify the gain in knowledge. The influence of info requirements on the pattern of search and gain in knowledge of the user has been examined. It just disclosed a considerable correlation between the search pattern and info requirements though no connection is established on the gain in knowledge. Also, it has been observed that high level of knowledge increase is attained only when the users searched in the area in which they are not much familiar. Through the result, it is found that there is a kind of bias in presenting the query during search by the users because of the pre-session test. The analysis has been carried out only on few topics. The same kind of results may probably not be attained when diverse topics with large users are involved.

A similar study has been noticed in the literature by Yu et al., (2018a). They covered a work using a survey for detecting the essentials of learning, user's thought with regard to the status of the knowledge, tasks of learning and the growth of learning all through the searching sittings. The original concern of requirements of learning is also included in the entire course of retrieval and ranking. While other studies on same topic focused on intent of detection using query-based approaches, the authors made it automatic during search sessions with supervised classification models. The search considered involved multiple dimensions through which 22 features were extracted with 3 different groups, viz. "Query related features" like the query count and the resemblance amongst the queries, "Session related features" like number of issued queries, duration of session and session intervals, and, "Browsing manner related features" like number of clicks, pages visited again and similarity among the query and URL. The classification model for the same comprised of Logistic Regression, Random Forest, Decision Tree and Support Vector Machine, which was decided considering the count and characters of features. The designed model had been applied to a dataset of true logs of query having 6860 questions attained from 124 users in more than 900 sessions. Every session was annotated with minimum of 2 users to assign one class from the above said categories. The results produced were on the basis of F1 score, precision and recall. Though the results are better on an average, there exists some ambiguity when a particular session is concerned. The nature of trial dataset is also important and the query logs are to be updated for making the supervised model effective. This necessitates for real-life data with updated query record.

A detailed session analysis has been made on session learning leading to enhancement of knowledge. The logs of search made by people in explicit way through standard search engines are set as base for this work. A inside session and cross session progress of knowledge, by concentrating on how the influence of language and pattern of search of user on particular topic grows over time has been inspected. It has been recognized that the pages visited during sessions have got high significance and lifts the process of learning. It has been established through demonstration that there exists a robust link between clicks

and expertise metrics. Depending on the user model and context, a potential predicting method which is capable of performing automatically to predict such click that enables improved learning with great accurateness (Eickhoff et al., 2014).

The prevailing researches were attentive in what way user interacts with search engines for getting information and related learning. The degree of search engine being a viable medium of learning in comparison to conventional or video guided learning is still the area to be explored. To analyze this, a pit study has been made for learning assisted by a guide against three search occurrences like, search through single user, searching as a tool for supporting conventional learning and a collective search. The gain through learning by 151 candidates has been measured for the task of vocabulary learning and the following outcomes have been listed: Video lecturing improves the knowledge gain by 24% in comparison to search by single user, the collective search has not shown any significant enhancement of learning and video lecture aided with search engine improved the gain of learning by 41% in comparison to video lecture alone (Moraes et al., 2018).

Though it is difficult to mention specific gaps in the existing researches, it is understood from the existing research that the model supporting the knowledge gain of the user after the search relies on the relevant queries presented and the type of classification methods or clustering methods adopted. Also, an expertise user can attain the require information through the search well before the beginner. Any library program assisting the entry level user would serve problem better. Based on this, a model is intended to be made to assist to improve the knowledge gain of the user after the search, in this work.

2.7 Studies on Knowledge Gain in Search Sessions

Gadiraju et al., (2018) have offered a study to know about the knowledge gained by the users after the sessions of searching for information in the Internet. The authors hired 500 individuals with the help of crowdsourcing and arranged practical sessions for searching with 10 topics covering various domains and info requirements. The knowledge of internet users is calibrated/standardized prior to and after the sessions of searching using knowledge experiments that are systematically designed. This process helps to quantify the gain in knowledge. The influence of info requirements on the pattern of search and gain in knowledge of the user has been examined. It just disclosed a considerable correlation between the search pattern and info requirements though no connection is established on the gain in knowledge. Also, it has been observed that high level of knowledge increase is attained only when the users searched in the area in which they are not much familiar. Through the result, it is found that there is a kind of bias in presenting the query during search by the users because of the pre-session test. The analysis has been carried out only

on few topics. The same kind of results may probably not be attained when diverse topics with large users are involved.

A similar study has been noticed in the literature by Yu et al., (2018a). They covered a work using a survey for detecting the essentials of learning, user's thought with regard to the status of the knowledge, tasks of learning and the growth of learning all through the searching sessions. While other studies on same topic focused on intent of detection using query-based approaches, the authors made it automatic during search sessions with supervised classification models. The search considered involved multiple dimensions through which 22 features were extracted with 3 different groups, viz. "Query related features" like the query count and the resemblance amongst the queries, "Session related features" like number of issued queries, duration of session and session intervals, and, "Browsing behaviour related features" like number of clicks, pages visited again and similarity among the query and URL. The classification model for the same comprised of Logistic Regression, Random Forest, Decision Tree and Support Vector Machine, which was decided considering the number and types of features. The designed model had been applied to a dataset of true logs of query having 6860 questions attained from 124 users in more than 900 sessions. Every session was annotated with minimum of 2 users to assign one class from the above said categories. The results produced were on the basis of F1 score, precision and recall. Though the results are better on an average, there exists some ambiguity when a particular session is concerned. The nature of trial dataset is also important, and the query logs are to be updated for making the supervised model effective. This necessitates for real-life data with updated query record.

A detailed session analysis has been made on session learning leading to enhancement of knowledge (Eickhoff et al., 2014). The logs of search made by people in explicit way through standard search engines are set as base for this work. The authors inspect the inside session and cross session progress of knowledge, by concentrating on how the influence of language and pattern of search of user on particular topic grows over time. It has been recognized that the pages visited during sessions have got high significance and lifts the process of learning. It has been established through demonstration that there exists a robust link between clicks and expertise metrics. Depending on the user model and context, a potential predicting method which is capable of performing automatically to predict such click that enables improved learning with great accurateness.

The prevailing researches were attentive in what way user interacts with search engines for getting information and related learning. The degree of search engine being a viable medium of learning in comparison to conventional or video guided learning is still an area to be explored. To analyse this, a pit study has been made for learning assisted by a guide against three search occurrences like, search through single user, searching as a tool for supporting conventional learning and a collective search. The gain through learning by 151

candidates has been measured for the task of vocabulary learning and the following outcomes have been listed: Video lecturing improves the knowledge gain by 24% in comparison to search by single user, the collective search has not shown any significant enhancement of learning and video lecture aided with search engine improved the gain of learning by 41% in comparison to video lecture alone (Moraes et al., 2018).

Though it is difficult to mention specific gaps in the existing researches, it is understood from the existing research that the model supporting the knowledge gain of the user after the search relies on the relevant queries presented and the type of classification methods or clustering methods adopted. Also, an expert user can attain the required information through the search well before the beginner. Any library program assisting the entry level user would serve the problem better. Regardless of the expertise level, the knowledge gain assessment is important for all users. The level of knowledge gain can be assessed through the features used in the search log. Anyhow, the search log will carry many features. If the important features are identified, which have real impact on the output, then improving knowledge gain becomes easier. Hence, the effect of features on knowledge gain is considered in this dissertation. The feature selection will be made using an appropriate model. These features will be used for classification of knowledge gain of users under various categories.

2.8 Common Evaluation Metrics (Retrieval Measure)

They are used to quantify the performance of retrieval schemes of information and to relate them between schemes. The common measures utilized are the metrics called recall and precision. Another metric that is popularly used is the F1 score.

2.8.1 Precision

Precision quantifies the capability of the retrieval scheme to yield only appropriate results. “Precision is the ratio between the number of relevant information/documents retrieved by the system and the total number of information/documents retrieved”.
A perfect model would yield a precision mark of 1, that is, every document retrieved by the system is adjudged pertinent. The calculation procedure for precision is comparatively simple. But a complication faced with precision calculation with regard to search is the quantity of results generally specified back in reply to usual queries. In most of the circumstances, hunt engines return innumerable results. In an assessment state, it is not practicable to review so much of results. So, cut-off rates (e.g. 30 for the initial 30 hits) are applied in tests of retrieval.

2.8.2 Recall

Recall, quantifies the potential of any retrieval method to determine the whole set of

related outcomes from the assemblage of documents. “Recall is the ratio of the number of relevant documents retrieved by the system to the total number of relevant documents for the given query”. In the context of searching machine, the overall number of appropriate documents refers to complete pertinent documents on the Web.

2.8.3 F1 Score

F1 score or measure is measured on the basis of both the above metrics. It is stated as “the ratio of twice the product of precision and recall to the sum of precision and recall”.

2.9 Summary

To understand the importance of knowledge gain of the user after the search through search engine various literatures have been explored on the areas of search log analysis, Influence of user’s expertise, query analysis, topic analysis, classification techniques with its advantages and disadvantages and other studies pertaining to the selected topic. Also, the common evaluation metrics to assess either the performance of the system or the classification performance are studied.

3 Data Preparation

It is very much necessary to redesign and sharpen up the raw dataset into usable/operational dataset for leveraging the analysis. This is needed because the dataset may have discrepancies/ bad data and sometimes it lacks the desired traits needed for the point of interest. If the quality of the data is poor, it would definitely affect the accuracy. Thus, the dataset analysis invariably starts with the preparation of data. The main steps involved in this process entail the cleaning of data, assimilation of data, transformation of data, data reduction and discretization. Thus the unreliable data are corrected and the data noises are smoothened out. The preferred methods for noise reduction are regression, Bayesian, decision trees etc

3.1 The Dataset

For evaluating the knowledge of the user during search sessions, a dataset has been collected (Yu et al., 2018b) using crowdsourcing, in which, the users involved themselves in real practical search sessions. The dataset encompasses the search history of 468 individual users and the total number of topics searched during this process is 11. This dataset is publicized by the authors and available in the public domain. The same dataset is utilized for the proposed work. The fields covered in the dataset are listed in Table 1.

Table 1: Dataset Fields

User ID
Session ID
Query String
Query Timestamp
Link Click Timestamp
Link Rank
Link URL

Link Header
Link Excerpt
Pre-Session Knowledge Score
Post Session Knowledge Score

3.2 Data pre-processing

The preparation stage is generally the data pre-processing stage, involves the data cleaning. The techniques adapted in this work for pre-processing are the pruning/trimming. This removes the extent of unwanted or superfluous data, wherein which, the white spaces, stop words, HTML tags, punctuation marks, etc. are removed. Also, the uppercase letters are changed to lowercase. In addition, the tokenization is carried out in order to break the string sequences into symbols, key/clue-words, words, and other components referred to as tokens.

3.3 Sampling Users

In order to achieve objectives of the proposed work, a study conducted by Yu et al (2018) has been considered and the same data set has been used. Yu et al (2018) have recruited 468 distinct users from Crowdflower which is a crowdsourcing platform for web search. These users have made web search on 11 different topics and information area on internet. Using similar data, this study presents a different examination and investigation model to evaluate the impact of the features on the knowledge gain during web search by users. User's knowledge state before and after their informational search sessions has been tested using statistical tools and models such as liner regression where different factors as variables are used to establish relationship between factors and impact on knowledge gain. Linear equation was formed to observed data.

3.4 Features

The amount of knowledge of a particular user can be decided through the features of the search sessions. These features can then be utilized to assess the knowledge level of a user and accordingly the search engine can be customized. The probable features that may be considered for analysis may include, Query related features (queries & unique queries, time per query, query terms, unique query terms & unique query terms ratio and query complexity), Session related features (search clicks & search clicks rank), Title/Topic related features (title/topic relevance) and Excerpt related features (excerpt relevance).

All the group of data cannot be accounted because involving every data will increase the size and ambiguity. The group of data with a minimum of one search result click is considered; for the rest of the group, in case if there is any knowledge gain, it can be credited to exterior factors or the knowledge is gained without any metrics (like reading search extract outcome without ticking the link for the result). In a nutshell, the weightage or impact of features is calculated based on some metrics and suitability of models assumed. Only the significant features that influence the knowledge gain are considered for analysis.

3.5 Search Sessions

A session from a circumstantial lookout can be stated as a succession of interactions/communications by the user for the purpose of addressing a single information requirement. Intrinsically, the session is the investigation at critical level to find the victory or disaster of an information system of web. If the user's need of information is fulfilled from the session, then the system (or the team of user and the system) is said to be successful (Jansen, et al. 2007).

A deliberate searching session connected to learning, will have a sequence of activities by users with regard to fulfilling the learning quest of them by the way of informational queries (Yu, et al. 2018). The user, in fact, start the search by posing queries to the web and the sequence of activity proceeds with surfing through the results of search, clicking and scrolling deeds, hyperlink navigations, reforming the queries based on the information acquired through the said activities and the like.

The knowledge gain is supposed to be predicted within a search session. This is hence decided by the period of activity as well as inactivity by the users. A definite long period of inactivity ends the session and any queries issued by the user even on the same topic or similar queries are considered as the subsequent session. The inactivity period considered here is 20 minutes.

The search duration, query details (length of the query, query term, complexity of query) and search engine result page related information (duration of search, minimum, maximum and average clicks made during the search along with their ranking) are contributing as the indicative or decisive features in deciding the extent of knowledge acquired by the user.

3.6 Query Classification

The last phase of data preparation is the query classification. The queries with noise are eliminated and the categories of queries or the query topics are confirmed at this stage. The queries of search are categorized based on their purpose into three types namely,

navigational search queries, informational search queries and transactional search queries ((Andrei Broder, 2002). This work is focusing on the queries issued to web for getting information for the purpose of gaining knowledge by the user. Thus, the informational queries are defined as such queries that presented by the user to obtain details/information, expected to be available on few of the web pages.

3.7 Categorization of queries and URL

Having finalized the queries after removing the queries of navigational type, the grouping could be done using the topic of interest. As per the dataset used for this work, the topics and related queries have been classified. The prime intention is to relate the features of navigational queries with the knowledge enhancement of user after the search. The features considered for this work include the following:

- Queries
- Unique Queries
- Session Duration
- Time per Query
- Query Terms
- Unique Query Terms
- Unique Query Terms Ratio
- Query Complexity - Average
- Search Clicks - Average
- Search Clicks Rank - Average
- Title Relevance - Average
- Excerpt relevance - Average
- Query Complexity - First
- Query Complexity - Last
- Query Complexity - Max
- Query Complexity - Min
- Search Clicks - Total
- Search Clicks Rank - Max
- Search Clicks Rank - Min
- Title Relevance - Max
- Title Relevance - Min
- Excerpt relevance - Max
- Excerpt relevance – Min

3.8 Summary

Having known the significance of data preparation, the pre-processing work covering the

cleaning techniques have been carried out. The dataset relevant to the work has been chosen by analysing various appropriate literatures. The various features have been decided based on the search session details entailing the session duration, query details, search engine result page or click details, excerpt level details. The idea is to extract the significant features that are affecting the knowledge enhancement of the users within the search session. For accomplishing the same, regression and classification models are used, which would be discussed in detail in the subsequent sections.

4 Data Analysis

This section will provide the framework of analysis of data by giving much attention to research question 4, that is, machine learning model to guess the knowledge gain by using the features extracted during the search session and analysing the addition/deletion of significant features. Also, the various models used for finding the important features and various classification models are presented.

4.1 Model for Analysis

It is envisioned to calibrate/standardize the user's knowledge prior to and next to the sessions of searching and to calculate the knowledge growth (Refer to Table 1). The prediction will be on the basis of the features derived through the search sessions, and the prediction aims to gauge the state of knowledge of the user and the increase in knowledge after the information search.

For accomplishing the same an efficient model to genuinely predict the results are essential. The predictive models are having two different areas of classification viz. Regression and Pattern Classification. The former model is made on the basis of investigating the linkages among variables and tendencies for making a guess about uninterrupted variables, whereas the latter assigns distinct labels of class to some specific observation, which is a predictable outcome.

Statistical and machine learning models are much used in prediction models. Machine learning, more precisely the arena of predictive sculpting is principally aimed with lessening the error of a model or building the most truthful probable predictions. As such, regression has been established in the statistical field and is premeditated as a model for knowing the association between input and output numerical variables. It is an algorithm that can be thought both under statistics and machine learning. The other original machine learning methodologies are the supervised and unsupervised models. It is intended to judge the percentage of knowledge gain by the users through web search using either statistical or machine learning models.

4.1.1 Regression Model:

It is an imperative method for predictive modelling and examines the linkage among dependent and independent variables of a particular problem. In this method, a curve or line is fit to the data points in such a way that the variances among data point distance and the curve / line are diminished (Feng et al., 2014). The taxonomy of regression techniques is based on the number of independent variables, dependent variables' type and the regression line type. Out of the various types, Linear, Lasso and Ridge type regression models are majorly applied in the field of search log analysis and associated knowledge gain.

4.1.2 Machine Learning Model:

The supervised machine learning works with known class labels that are utilized to classification model building (Li, 2013). The unsupervised model deals with unlabelled cases. That is, the classes should be derived through unstructured data (Mishra et al., 2011). Precisely, supervised learning does classification while the unsupervised model does the clustering. The problem of concern in the proposed work should deal with classification. Hence, the regression and supervised models could be applied to accomplish the said objective.

4.2 Methodology

The analysis methodology has been tailored to the nature of data available for the study. The fields that are found in the "test-score files" of the dataset are the user ID, pre-score (i.e. knowledge level of the user prior to search), and post-score (i.e. knowledge level of the user after searched for information in the web by informational queries). Likewise, in the "query-and-click files" of the dataset, various fields like, user ID, session ID, query string, query timestamp (for all queries), as well as link click timestamp, link rank on the page, link URL, link header, and link excerpt for the cases where a result link has been clicked are available. It has been decided to go through every pair of files topic-wise and then link the data records by user ID because that is the only common field between the two files mentioned. The score gain (indicator of knowledge gain) is calculated by subtracting pre-score from post-score. As and when the score gain is either negative or zero, such outliers are ignored, because the deviation may have been triggered because of the external factors. The various metrics are calculated from the data available in query-and-click files, by grouping data as per the following two cases:

I. Based on user id
II. Based on user id and session id

Also, for the analysis, only those data groups where at least one search result has been

clicked are considered. For all other data groups, knowledge gain (if any) can be endorsed to external aspects, or to knowledge gained in a way for which metrics are not available (such as, reading the excerpt in search results, without clicking on the result link). Three different cross validation regression models are chosen for analysing the impact of features on knowledge gain. They are,

I. Linear Regression
II. Ridge Regression
III. Lasso Regression

For experimentation, Scikit-learn library (used with machine learning) of Python language of programming is utilized. It supports the regression and classification models used in this work.

4.2.1 Linear Regression

It tries to model the association amongst two variables by correct a linear equation from the detected data. One of the two variables is thought out to be an independent variable, which is also referred as predictor or explanatory variable, while the next variable is thought as a dependent variable that is known to be either response variable or outcome variable. Prior to create a linear model, it is necessary to find whether there is any relation existing between the concerned variables or not. A graph is acting as a tool for visual identification to identify the relationship. If at all there exists no linking between the variable, it could be easily seen through the shape or trend of graph (neither increasing nor decreasing). Mathematically, the linear model of regression can be expressed by straight line equation, which is,

$$Y = aX + b \tag{1}$$

Here X denotes the predictor variable and Y is the response variable. The constant “a” is the slope or gradient of the straight line and the constant “b” denotes the intercept, obtained by making the value of X as zero. This model is very simple but it is sensitive to outliers and it is giving better results, only when there is linearity between the input and output. Moreover, it suffers through “noise” in case of large number of parameters.

4.2.2 Ridge Regression

Ridge regression is basically a method used to analyze multi regression statistics/information suffering through collinearity which is otherwise known to be multi-collinearity. It is nothing but the presence of close by- linear associations between the independent variables (Abram, et al. 2016). During such conditions, the estimated least

squares become unbiased nevertheless they have big variances. Thus, the value obtained or the derived results might be far from the real values. This could be addressed by adding a percentage of bias to the estimates of regression. The standard errors are then reduced and this model is the ridge regression model.

This is an extension of linear regression method. The linear regression suffers from the problem of overfitting when dealing which large parameters. This could be overcome by properly adding the weightage factors (penalty factor). This method is also called as linear regression with L2 regularization.
The mathematical expression of the loss function of this model can be written as

$$L = \sum (\hat{Y}_i - Y_i)2 + \lambda \sum \beta^2 \quad (2)$$

The first term indicates the summation of distance amongst the prediction and the reality, whereas the second term is a summation square of β and multiplies with λ. The multiplier λ is called the penalty factor. "λ" represented at this juncture, is in fact designated by parameter called alpha in the ridge regression function. Thus, by altering the alpha values, the penalty term are essentially controlled. More the values of alpha, higher are the penalty and consequently the scales of coefficients are getting reduced.

4.2.3 Lasso Regression

This is also very much similar to Ridge regression. Although Ridge and Lasso may seem to work for the same purpose (avoiding overfitting), the characteristic properties and real-world application cases vary considerably. It is known that they both perform by penalizing the scale of feature coefficients in addition to minimize the error amongst predicted and definite observations (Hara & Maehara, 2017). These methods are generally named as 'regularization' procedures. The important difference is in what way they allocate penalty to the coefficients. While Ridge performs L2 regularization, Lasso performs L1 regularization. L2 increases penalty equal to the square of the magnitude of coefficients while Lasso (L1) adds penalty one and the same as absolute value of the coefficients' magnitude, which could be understood from the mathematical equation of Lasso loss function

$$L = \sum (\hat{Y}_i - Y_i)2 + \lambda \sum \|\beta\| \quad (3)$$

All the terms in the equation have the same meaning as that the one presented for Ridge regression.

For each of the three models described above, the coefficients of every factor, percentage of model fitment and the root mean square of residuals have been calculated. The selection

of variables is automatic in Lasso CV and it makes the coefficients of the all the variables that are non-relevant as zero. Thus Lasso CV is considered superior to other two models. Also, it is preferable to opt for the model with least RMSE (Root Mean Square Error).

RMSE is nothing but the Standard Deviation (SD) of the prediction errors (residuals), whereas, the residuals are the portion (measure) of the distance between data points and regression line. RMSE measures the range (spread) of residuals or it indicates how much data are concentrated around the best fit line. The results of all the models are presented in the next chapter.

4.3 Analysis

When the analysis has been started, by grouping the data just with user ID, for all the three models, it has been observed that Lasso CV has given high value of RMSE and very low fitment in comparison to other models. This might be because of the cross-linkage between the factors. The only appropriate factor found by Lasso CV has been session duration (Grouping by User ID alone). On eyeball analysis of data files, it has been realized that refactoring of the data are required to group them by user ID as well as session ID, because some user session are disjoint and measuring session duration across session makes no sense. Hence, the analysis has been continued with user session queries. Now going through the output, it has been viewed that Lasso CV has got the least RMSE and it has done variable selection to find five relevant factors (features) as Pre-Score, Session Duration, Time per Query, Number of Query Terms, and Search Click Rank (max and min). The pre-score is however the previous knowledge level of the user, which could not be considered as a feature. The other features (four) identified by Lasso CV are the features that would impact the knowledge gain of the user at a higher rate. A re-run with only these selected factors gives nearly identical results for all models. Lasso gives the best fit as it regularises the model by removing dependency between the features. Since, the prediction of the knowledge gain being difficult with regression models, the problem could be approached with classification model. The supervised model has been taken for analysis. The idea is that the various models can be inputted with more features and also with features suggested by Lasso CV. This would be helpful in comparing the prediction accuracy of various classification models considered in this study.

4.4 Classification Models

Classification models stand as a good prediction models. The classification models have labels than scores. Thus, the knowledge gain is evaluated using various labels instead of knowledge scores. The labels considered for this study are “low”, “medium”, and “high”. Obviously the label “high” indicates the highest knowledge gain. The coding is done to obtain the mentioned labels using Standard Deviation. With 0.5 SD on either side of “mean”

is considered “medium”. The values lower than “medium” are considered “low” and higher are considered “high”.

The various classification models used for this work include, Gaussian Naive Bayes, Multinomial Naive Bayes, Complement Naive Bayes, SVM, K-Nearest Neighbors Classifier, Bagging Classifier, Ada Boost Classifier, Decision Tree, and Extra Trees. The brief notes on all the considered models are provided in this section.

4.4.1 Naive Bayes Classifiers

Naive Bayes classifiers are group of algorithms used for classification that work on the principle of Bayes’ Theorem. Here, every one pair of features that need to be classified is independent (Li, et al. 2018). In machine learning, it is aimed to select the pre-eminent hypothesis (h) for given data (d). In the problem of classification, the hypothesis (h) possibly will be the class to allocate for a fresh data instance (d).

Bayes’ Theorem delivers a method of calculating the probability of a hypothesis with given (prior) knowledge. It is represented as,

$$P\langle h|d\rangle = \left(P\langle d|h\rangle * P(h)\right) \Big/ P(d) \quad (4)$$

P⟨h|d⟩ is the probability of hypothesis h for the given data d. It is named as posterior probability.

P⟨d|h⟩ is the probability of data d when the given hypothesis h is true.

P(h) is the probability of hypothesis h being true irrespective of data. It is referred to as prior probability of h.
P(d) is the probability of the data irrespective of the hypothesis.

The Naive Bayes classifiers considered for this work are, Gaussian Naive Bayes, Multinomial Naive Bayes, and Complement Naive Bayes.

In Gaussian Naive Bayes, continuous values related to every feature are assumed to be disseminated as per Gaussian distribution, which is referred as Normal distribution. The final plot is bell shaped and is symmetric about the mean value of feature values. The vectors containing the features will represent the frequencies with which some events have been made by a multinomial distribution. This is the event model normally utilized for classifying the documents. The Complement Naive Bayes type of classifier is intended to spot-on the “severe assumptions” created through standard Multinomial Naive Bayes model and is best

fit for excessive (imbalance) data sets.

4.4.2 Support Vector Machine (SVM) Classifier

For any dataset comprising of set of features and set of labels, the SVM classifier forms a model to guess classes for new-fangled examples. It allocates fresh data points to one of the classes. It takes two forms namely, the linear and non-linear model (Tajiri, et al. 2010). The training model in linear classification is plotted in space and the data points are segregated by a gap and hyperplane (straight line for this case) is used for separating the classes. It is usually aimed at maximizing the distant between all classes and hyper plane. In non-linear model, the plotting of data points will be in space of higher dimensions. In such cases, the separation is not on straight line basis but using a kernel trick.

4.4.3 K-Nearest Neighbors Classifier

KNN algorithm is the simplest and most used learning algorithms and it is one of the most used learning algorithms (Begum, et al. 2015). KNN is a non-parametric, lazy learning algorithm. Non-parametric means, it does not work on any assumption. It is referred as lazy because, there is no explicit training phase or it is very minimal. This model is based on the similarity, i.e. the closeness between the features out of samples and the training set. The nearest one is assigned with a label.

4.4.4 Bagging Classifier

This classifier works under ensemble method of learning. Bagging put efforts to implement related learners on small sample populaces and takes the mean value of the entire predictions (Tran, et al. 2017). Different learners of different population can be utilized in bagging. This method is helpful in decreasing the error of variance.

4.4.5 Ada Boost Classifier

Boosting algorithms syndicate numerous weak models or models with low to make a new strong model with high value of accuracy. It is a type of ensemble (composite) approach of machine learning. The fundamental idea behind Adaboost is to fix the weights of classifiers and training the data sample, at all iterations, in such a way that uncommon observations are predicted with maximum accuracy (Silapachote, et al. 2005). Any type of machine learning method that is capable of receiving weights on dataset could be utilized as base model.

4.4.6 Decision Tree

A Decision Tree is a type of directed graph in the form of tree. The nodes of tree signify decisions, and the ends or branches are binary (in the pattern “yes/no”, “true/false”) demonstrating potential paths between nodes (Xie, et al. 2018). The explicit type of decision

tree utilized for machine learning has no arbitrary transitions. For using the decision tree for classification, one takes a set of features and starting at the root, travelling via every succeeding decision node to the terminal node. The process is very spontaneous and simple to understand, which permits trained decision trees to be applied for selection of variables or features.

4.4.7 Extra Trees

It is an extremely randomized tree classifier. They differ from regular decision tree by the way they have been constructed. The best split method has been adapted here, for which, random splits are drawn for all randomly chosen features (Pliakos & Vens, 2016).

The supervised models of classification, work by training the dataset or they have two functions viz. training and testing/validation. The ratio of training and testing for this work is in the ratio 80:20. The accuracy and fitment are calculated for all the classification models considered. The feature important score has been included in the model wherever available.

4.5 Evaluation

Once the prediction has been done/ classified, the various models are evaluated on some metrics. The statistical evaluation metrics such as recall, precision and accuracy are computed for each class. In a multi class or multi label problem, the average values of precision, recall and accuracy is required. This is performed in three different manners and accordingly we have micro average, macro average and weighted average.

The "True Positive (TP)", "True Negative (TN)", "False Positive (FP)" and "False Negative (FN)" are to be known for each class to calculate the values of the basic evaluation metrics. The mathematical expressions for them are described as:

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (5)$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (6)$$

$$F1 - Score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (7)$$

F1-Score is the weighted harmonic mean of precision and recall.
It is sufficient to calculate the above three measures in case of single class problem. For multi-class analysis, the average value of the entire three measures is required. It may be by macro average and micro average method. Sometimes weighted average technique is also employed.

Micro:

It calculates the average of metrics generally by totalling the true positives, false negatives and false positives.
For example, the micro-average precision is calculates as,

$$Micro - Average\ Precision = \frac{T_{P1} + T_{P2} + \cdots . + T_{Pn}}{T_{P1} + T_{P2} + \cdots + T_{Pn} + F_{P1} + F_{P2} + \cdots + F_{Pn}} \quad (8)$$

$$Micro - Average\ Recall = \frac{T_{P1} + T_{P2} + \cdots . + T_{Pn}}{T_{P1} + T_{P2} + \cdots + T_{Pn} + F_{N1} + F_{N2} + \cdots + F_{Nn}} \quad (9)$$

The subscripts 1, 2 and n represent the respective classes.

Macro:

It calculates the average of basic metrics for each label and determines the mean of them without any weights added. It does not consider the imbalance in the label.

For example, the macro-average precision is calculates as,

$$Macro - Average\ Precision = \frac{P_1 + P_2 + \cdots + P_n}{n} \quad (10)$$

$$Macro - Average\ Recall = \frac{R_1 + R_2 + \cdots + R_n}{n} \quad (11)$$

The Macro-average method could be applied in case to know the overall system performance across the dataset. However, micro-average could be a beneficial measure when the size of dataset differs.

Weighted:

It calculates metrics for each label as macro does but in contrast to macro it computes their average weighted by support (the quantity of true cases for every label). The label imbalance is hence rightly accounted.

For example, the precision average with weights for two classes could be computed as,

$$Weighted_P = \frac{(P_{C1} * |C_1|) + (P_{C2} * |C_2|)}{|C_1| + \lceil C_2 \rceil} \quad (12)$$

Where, P_{C1}, P_{C2} are the precision of class 1 and 2 respectively and C_1 and C_2 denote the number of instances of respective classes. The weights are assigned in terms of support calculated through the model.

Accuracy:

The important classification metric is the accuracy. It is actually the percentage of right predictions obtained by the considered model. It could be expressed as,

$$Accuracy = \frac{Number\ of\ right\ predictions}{Total\ Number\ of\ predictions} \quad (13)$$

Also,

$$Accuracy = \frac{T_P + T_N}{T_P + T_N + F_P + F_N} \quad (14)$$

The more the value of accuracy, the more good is the classification model.

The evaluation metrics have been calculated for all the classification model is computed with the features in the dataset as well as with the important features extracted by the regression model to show the weightage of knowledge gain as an impact through the selected features. Also, it is possible to compare the various classification models based on their accuracy for each case.

4.6 Summary

The process involved in data analysis had been detailed in this chapter. The model analysis using regression and classification has been discussed. The three different model of regression with the preferred one has been elaborated. This modelling has been done for extracting only the features impacting the knowledge gain of the user. Also, the analysis with user ID alone and including Session ID has been discussed. Having extracted the features, it has been aimed to model and test the various classification models. The ratio for supervised learning phases of training and testing has been set as 80:20. A brief note about the several models of classification has been provided. After testing the models to obtain the results, the different evaluation metrics used has been discussed. Since, the problem of consideration involves, multi class/label, the various averaging methods also have been detailed. The results of all the models used for the study will be presented in the next chapter.

5 Results

5.1 Results of Cross Validation Regression Models

The effect of various features on the enhancement of knowledge of the user has been investigated by the use of three models of regression with cross validation models, namely the linear type, ridge type and lasso type. In all the models, the feature coefficients are computed. The model fitment and the Root Mean Square Error (RMSE) are also calculated. The model that yields low value of RMSE is considered as the better model.

Root Mean Square Error

"RMS error", or else "RMS deviation", is a regularly preferred measure of the differences/variances among values foreseen through a model or an estimator and the values found in reality (Chai & Draxler, 2014). These distinct variances are known as "residuals" if the computations are done within the sample data and named as "prediction errors" if the estimations are for out of samples. RMSE helps to sum up the "prediction error" magnitudes for different intervals into an only one collective measure of power predicted. It is a decent accurateness measure, but merely to associate anticipating errors of diverse models for a specific variable and not amongst variables.

The errors are squared for "n" samples and their mean gives the value of MSE. The square root of MSE provides RMSE (Jachner, et al. 2007). Mathematically,

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left[\widehat{Y}_i - Y_i\right]^2 \quad (15)$$

Y_i - Vector on predictions
$\widehat{Y}_i$ - True value vector
n - Number of samples
MSE – Mean Square Error
RMSE is just square root value of MSE.

For any regression line, the error difference is nothing but the distance between the line and

the point of concern (lying either below or above the line). The overall error size is computed through the RMSE (Wang & Lu, 2018).

Scikit-learn library of Python language of programming is exploited for conducting the experiment with three models of regression. The results of all the three models are presented in Tables 2-4.

Table 2: Results of Linear Regression Model

Evaluation Metric	Root Mean Square Error (RMSE)	0.16831017
Fitness	Fitment %	50.67024515
Prior knowledge level of the User	PreScore	-0.59213730
Query Related Features	Queries	-0.00363711
	Unique Queries	-0.000000000000037
	Time per Query	-0.00095873
	Query Terms	-0.00702392
	Unique Query Terms	0.00075016
	Unique Query Terms Ratio	-0.10501150
	Query Complexity - Average	0.00075016
	Query Complexity - First	0.00075016
	Query Complexity - Last	0.00075016
	Query Complexity - Max	0.00075016
	Query Complexity - Min	0.00075016
Session Related Feature	Session Duration	0.00018339
Search Click Related Features	Search Clicks - Average	0.10501150
	Search Clicks Rank - Average	0.02173992
	Search Clicks - Total	-0.00363711
	Search Clicks Rank - Max	-0.01171261

	Search Clicks Rank - Min	-0.00609011
Title Related Features	Title Relevance - Average	1.09175800
	Title Relevance - Max	-0.41488350
	Title Relevance - Min	-0.42532410
Excerpt Related Features	Excerpt relevance - Average	-1.01703600
	Excerpt relevance - Max	0.50819960
	Excerpt relevance - Min	0.73724550

Table 3: Results of Ridge Regression Model

Evaluation Metric	Root Mean Square Error (RMSE)	0.16291774
Fitness	Fitment %	50.33011000
Prior knowledge level of the User	PreScore	-0.57276300
Query Related Features	Queries	-0.00437400
	Unique Queries	0.00000000
	Time per Query	-0.00094900
	Query Terms	-0.00658900
	Unique Query Terms	0.00072200
	Unique Query Terms Ratio	-0.09454400
	Query Complexity - Average	0.00072200
	Query Complexity - First	0.00072200
	Query Complexity - Last	0.00072200
	Query Complexity - Max	0.00072200
	Query Complexity - Min	0.00072200

Session Related Feature	Session Duration	0.00018300
Search Click Related Features	Search Clicks - Average	0.09454400
	Search Clicks Rank - Average	0.01852200
	Search Clicks - Total	-0.00437400
	Search Clicks Rank - Max	-0.01015200
	Search Clicks Rank - Min	-0.00502900
Title Related Features	Title Relevance - Average	0.07185200
	Title Relevance - Max	-0.00590300
	Title Relevance - Min	0.06696500
Excerpt Related Features	Excerpt relevance - Average	0.04689000
	Excerpt relevance - Max	0.05939100
	Excerpt relevance - Min	0.06889400

Table 4: Results of Lasso Regression Model

Evaluation Metric	Root Mean Square Error (RMSE)	0.15829993
Fitness	Fitment %	49.21725375
Prior knowledge level of the User	PreScore	-0.54365600
Query Related Features	Queries	0.00000000
	Unique Queries	0.00000000
	Time per Query	-0.00068100
	Query Terms	-0.00346500
	Unique Query Terms	0.00000000
	Unique Query Terms Ratio	0.00000000

	Query Complexity - Average	0.00000000
	Query Complexity - First	0.00000000
	Query Complexity - Last	0.00000000
	Query Complexity - Max	0.00000000
	Query Complexity - Min	0.00000000
Session Related Feature	Session Duration	0.00012500
Search Click Related Features	Search Clicks - Average	0.00000000
	Search Clicks Rank - Average	0.00000000
	Search Clicks - Total	0.00000000
	Search Clicks Rank - Max	-0.00016900
	Search Clicks Rank - Min	0.00124900
Title Related Features	Title Relevance - Average	0.00000000
	Title Relevance - Max	0.00000000
	Title Relevance - Min	0.00000000
Excerpt Related Features	Excerpt relevance - Average	0.00000000
	Excerpt relevance - Max	0.00000000
	Excerpt relevance - Min	0.00000000

The features having zero values are considered irrelevant/redundant. It is seen through Tables 2 and 3 that all considered features have some value in Table 2, least importance given to unique queries, while in Table 3, there is no value for “unique queries”. Hence, it could be understood that only the feature “unique queries” has been identified as irrelevant features by linear and ridge regression methods. In Table 4, the analysis is found to be fairly well because, more irrelevant features are rejected by this model. The very purpose of feature selection is to discard those features, which do not have any impact on the model output. Also, the evaluation metric for the regression models is the Root Mean Square Error

(RMSE). Comparing the three regression models considered, it is found that the Lasso model has the minimum value of 0.158 (Refer Tables 2-4) and therefore, the selection of features by this model is considered superior.

The graphical representation of results of various regression models are given in Figure 1.

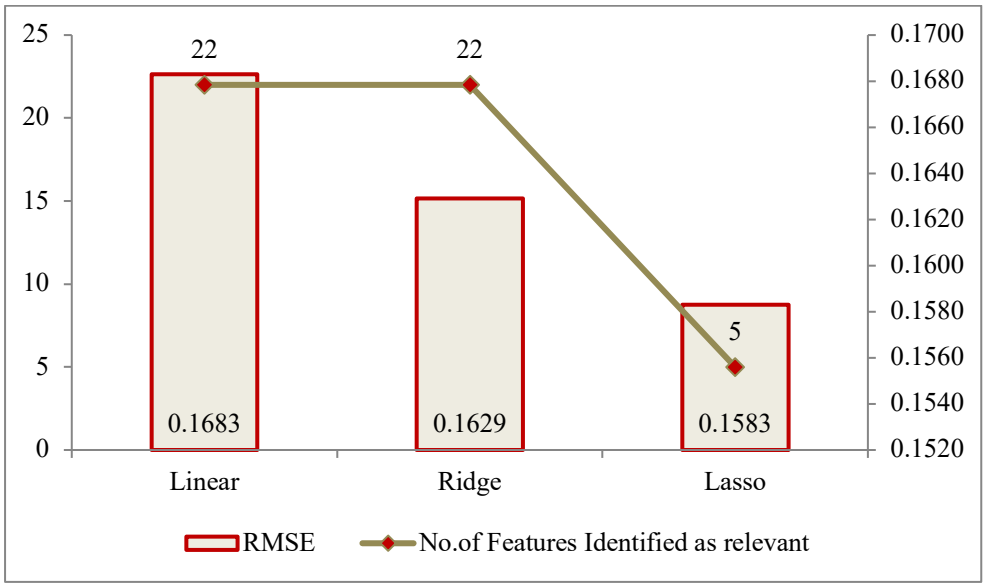

Figure 1: Results of Regression Models

The selected features by Lasso regression are Time per Query and Query terms from Query related features, Session Duration, Search Click Rank – Max and Search Click Rank – Min from Search Click related features.

5.2 Results of Classification Models

The Dissertation will proceed to relevant classification model for assessing the weightage of features or their impact in the knowledge gain of the user. Every model is tested with "including many features" and with "including just the features suggested by Lasso". The efficacy of classification model is calculated in terms of accuracy and other evaluation metrics. The classification models considered in the Dissertation involves, Gaussian Naïve Bayes, Multinomial Naïve Bayes, Complement Naïve Bayes, SVC, K- Nearest Neighbors, Bagging, Ada boost, Decision Tree, Extra Trees, and Random Forest Classification Models. For each model various evaluation metrics are calculated and compared. The results are analyzed as two cases; Case 1 considers some features but they need not be necessarily the features selected by lasso regression cross-validation model and Case 2 considering only those features that are suggested by lasso regression.

Case 1: Features selected randomly and the number of features is more than that selected by cross-validation lasso model of regression.

The features given to the various classification models comprise of Queries, Query Terms, Query Unique Terms, Session Duration, Time per Query, Average Query Complexity, Total Search Clicks and Maximum Search Clicks Rank. The report of classification for various classifiers is displayed in Table 5.

Table 5: Classification Report for Case 1

Knowledge Gain	Type of Classifier	Precision	Recall	f1-score	Support
Low	Gaussian Naïve Bayes	0.29	0.11	0.16	37
	Multinomial Naïve Bayes	0.33	0.16	0.22	37
	Complement Naïve Bayes	0.33	0.22	0.26	37
	SVM/SVC	0.46	0.86	0.6	37
	K- Nearest Neighbors	0.42	0.41	0.41	37
	Bagging	0.4	0.46	0.43	37
	Ada Boost	0.46	0.3	0.36	37
	Decision Tree	0.46	0.46	0.46	37
	Extra Trees	0.43	0.43	0.43	37
	Random Forest	0.41	0.49	0.44	37
Medium	Gaussian Naïve Bayes	0.36	0.22	0.27	37
	Multinomial Naïve Bayes	0.31	0.14	0.19	37
	Complement Naïve Bayes	0.25	0.05	0.09	37
	SVM/SVC	0.63	0.32	0.43	37
	K- Nearest Neighbors	0.42	0.46	0.44	37
	Bagging	0.44	0.38	0.41	37
	Ada Boost	0.38	0.57	0.45	37
	Decision Tree	0.4	0.38	0.39	37
	Extra Trees	0.44	0.41	0.42	37
	Random Forest	0.4	0.38	0.39	37
High	Gaussian Naïve Bayes	0.25	0.64	0.36	25
	Multinomial Naïve Bayes	0.23	0.6	0.33	25
	Complement Naïve Bayes	0.25	0.68	0.37	25
	SVM/SVC	0.6	0.24	0.34	25
	K- Nearest Neighbors	0.48	0.44	0.46	25
	Bagging	0.36	0.36	0.36	25
	Ada Boost	0.37	0.28	0.32	25
	Decision Tree	0.37	0.4	0.38	25
	Extra Trees	0.32	0.36	0.34	25
	Random Forest	0.32	0.36	0.34	25

As already described in Chapter 4, The labels considered for judging the knowledge gain are

“low”, “medium”, and “high”. Apparently the label “high” specifies the maximum knowledge gain. The coding is done to attain the stated labels using Standard Deviation (SD). With 0.5 SD on either side of “mean” is considered “medium”. The values lesser than “medium” are labelled as “low” and higher than “medium” are branded as “high”. The classifiers classify the knowledge enhancement level effectively as per the label. Their performances are actually estimated through some metrics.
The recall metric is the sensitivity but realizing 100 percent sensitivity is not possible; hence this evaluation measure is assessed along with precision. The f1-score is nothing but the subcontrary mean of both these indices. It is obtained by calculating the reciprocal of arithmetic mean. Thus, the classification efficiency is based on all the three metrics stated. The support values shown in the Table 5 are the quantity of samples of the factual responses existing in a particular class. The “weights” are actually the values provided as support.

It is seen from Table 5, that the highest precision for knowledge gain with label “low” is 0.46 and is accomplished by SVC, Ada Boost and Decision tree. The highest recall value is 0.86 for SVC. The f1-score is higher for SVC again with a value 0.6. It is easy here to arrive at a conclusion that SVC performs better with high values for all the three metrics.

Coming to the label “medium” indicating reasonable gain of knowledge by the user, though SVC presents more precision of 0.63, the recall and f1-score are high for Ada Boost Classifier with the values, 0.57 and 0.45. It is difficult to arrive at a conclusion. Likewise, in highest knowledge gain of user with label “high” the higher values of precision, recall and f1-score are 0.6, 0.68 and 0.46 exhibited by three different classifiers. Thus, the precision, recall and f1-score as a measure of average in multiple class condition could be analyzed.

The recall as well as precision values could be stated in the multiple class situations. In this case, the metrics are "averaged" through the entire classes in several conceivable techniques. Micro, macro and weighted are few of such kinds (Lipton, et al. 2014, Hodo et al. 2017).

micro: metrics are computed universally by totaling the overall number of times all classes are predicted “correctly” and “incorrectly”.

macro: metrics are computed for every "class" individually, and their mean is found without accounting their weights. The label imbalance is ignored here.

weighted: metrics are estimated for every label and their average is found by taking into account the support weights, that is, the true instance counts for individual labels. The label imbalance is taken in this case.

The metrics of evaluation for various type of average measure of classifiers for Case 1 are given in Table 6.

Table 6: Classification Report for Case 1 with different average metrics

Type of Average measure	Type of Classifier	Precision	Recall	f1-score	Support
micro-average	Gaussian Naïve Bayes	0.28	0.28	0.28	99
	Multinomial Naïve Bayes	0.26	0.26	0.26	99
	Complement Naïve Bayes	0.27	0.27	0.27	99
	SVM/SVC	0.51	0.51	0.51	99
	K- Nearest Neighbors	0.43	0.43	0.43	99
	Bagging	0.40	0.40	0.40	99
	Ada Boost	0.39	0.39	0.39	99
	Decision Tree	0.41	0.41	0.41	99
	Extra Trees	0.40	0.40	0.40	99
	Random Forest	0.39	0.39	0.39	99
macro-average	Gaussian Naïve Bayes	0.30	0.32	0.26	99
	Multinomial Naïve Bayes	0.29	0.30	0.25	99
	Complement Naïve Bayes	0.28	0.32	0.24	99
	SVM/SVC	0.56	0.48	0.46	99
	K- Nearest Neighbors	0.44	0.43	0.44	99
	Bagging	0.40	0.40	0.40	99
	Ada Boost	0.4	0.38	0.38	99
	Decision Tree	0.41	0.41	0.41	99
	Extra Trees	0.40	0.40	0.40	99
	Random Forest	0.39	0.38	0.38	99
weighted-average	Gaussian Naïve Bayes	0.31	0.28	0.25	99
	Multinomial Naïve Bayes	0.30	0.26	0.24	99
	Complement Naïve Bayes	0.28	0.27	0.22	99
	SVM/SVC	0.56	0.51	0.47	99
	K- Nearest Neighbors	0.44	0.43	0.43	99
	Bagging	0.41	0.40	0.40	99
	Ada Boost	0.40	0.39	0.38	99
	Decision Tree	0.41	0.41	0.41	99
	Extra Trees	0.41	0.40	0.41	99
	Random Forest	0.39	0.39	0.39	99

From Table 6, it is evident that the metrics measure for multiple class settings by various averaging methods are superior to the previous case used with binary classifier. It is also

revealed the values of “precision”, “recall” and “f1-score” are higher for SVC/ SVM classifier in all cases of average measures (micro-, macro- and weighted- average methods).

Another important parameter for evaluating the classification performance is the accuracy. The fitment of all the classifiers is also evaluated. The Table 7 will provide these values for various classification models.

Table 7: Classification Accuracy and Training Fitment for Case 1

Type of Classifier	Accuracy	Training Fitment
Gaussian Naïve Bayes	0.28283	0.355329949
Multinomial Naïve Bayes	0.26263	0.360406091
Complement Naïve Bayes	0.27273	0.360406091
SVM/SVC	0.50505	0.870558376
K- Nearest Neighbors	0.43434	0.840101523
Bagging	0.40404	0.852791878
Ada Boost	0.39394	0.598984772
Decision Tree	0.41414	0.870558376
Extra Trees	0.40404	0.870558376
Random Forest	0.39394	0.857868020

The graphical representation of Table 7 is presented in Figure 2.

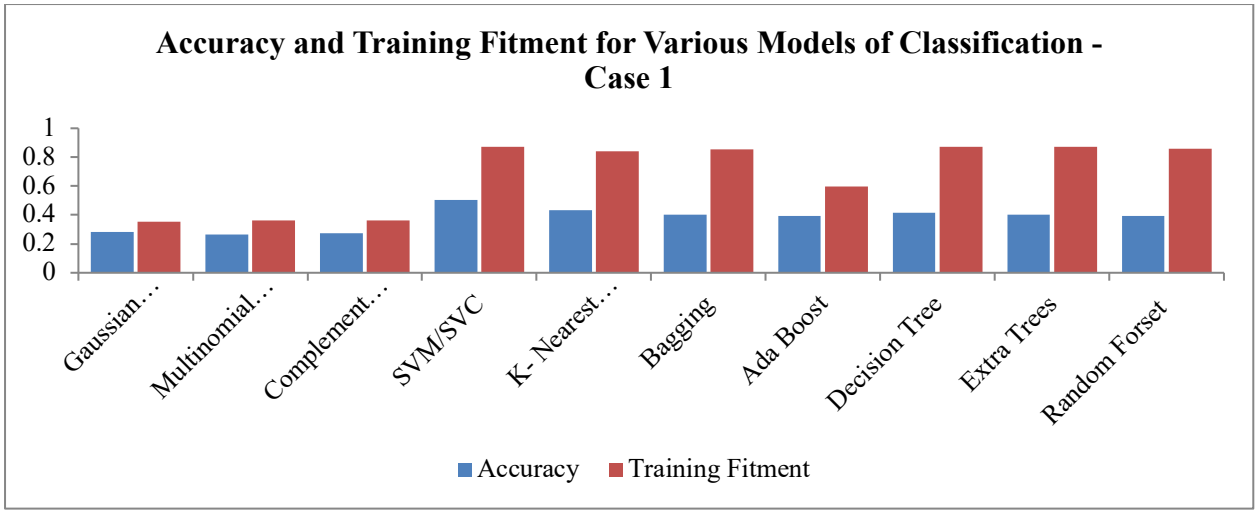

Figure 2: Accuracy and Training Fitment for Case 1

It is observed from Table 7 and Figure 2 that the SVC provides highest accuracy of 0.505 along with a fitment of 87.05 %. Though the same value of fitment is achieved through other few classifiers namely, Decision Tree, Extra Trees, and Random Forest, the accuracy is high only for SVC. Hence, SVC is considered to have superior performance for the case of identifying the knowledge gain by the users and classify under the three labels such as, "low", "medium", and "high".

Case 2: Features selected as per lasso regression.

The features given to the various classification models comprise of Query Terms, Session Duration, Time per Query, and Maximum Search Clicks Rank. The report of classification for various classifiers is displayed in Table 8.

Table 8: Classification Report for Case 2

Knowledge Gain	Type of Classifier	Precision	Recall	f1-score	Support
Low	Gaussian Naïve Bayes	0.50	0.11	0.18	37
	Multinomial Naïve Bayes	0.41	0.30	0.34	37
	Complement Naïve Bayes	0.38	0.22	0.28	37
	SVM/SVC	0.45	0.84	0.58	37
	K- Nearest Neighbors	0.43	0.41	0.42	37
	Bagging	0.38	0.46	0.41	37
	Ada Boost	0.42	0.30	0.35	37
	Decision Tree	0.43	0.43	0.43	37
	Extra Trees	0.47	0.41	0.43	37

	Random Forest	0.44	0.46	0.45	37
Medium	Gaussian Naïve Bayes	0.35	0.22	0.27	37
	Multinomial Naïve Bayes	0.33	0.14	0.19	37
	Complement Naïve Bayes	0.29	0.11	0.16	37
	SVM/SVC	0.58	0.30	0.39	37
	K- Nearest Neighbors	0.43	0.49	0.46	37
	Bagging	0.35	0.32	0.34	37
	Ada Boost	0.31	0.32	0.32	37
	Decision Tree	0.39	0.35	0.37	37
	Extra Trees	0.43	0.41	0.42	37
	Random Forest	0.33	0.30	0.31	37
High	Gaussian Naïve Bayes	0.26	0.72	0.39	25
	Multinomial Naïve Bayes	0.21	0.48	0.29	25
	Complement Naïve Bayes	0.23	0.60	0.34	25
	SVM/SVC	0.64	0.28	0.39	25
	K- Nearest Neighbors	0.50	0.44	0.47	25
	Bagging	0.40	0.32	0.36	25
	Ada Boost	0.26	0.36	0.31	25
	Decision Tree	0.34	0.40	0.37	25
	Extra Trees	0.31	0.40	0.35	25
	Random Forest	0.44	0.48	0.46	25

This case can also be discussed as that of Case 1. From Table 8, it is seen that for the label “low”, the highest values of “precision”, “recall “and “f1-score” are obtained by different classifiers. That is, Gaussian Naïve and SVC for other two metrics with values of 0.5, 0.84 and 0.58 respectively. Similarly, for label “medium” the values of metrics in the given order are 0.58, 0.49 and 0.46, which are acquired through SVM for first metric and by K – nearest neighbor for the remaining metrics. For label “high” the values are 0.64, 0.60 and 0.47 for precision, recall and f1-score correspondingly. The classifiers yielding theses values are again different. It is ambiguous to reach to a conclusion. Thus, the average measure type is adapted for measuring “precision”, “recall” and “f1-score” and for further analysis. The results are displayed in the Table 9.

Table 9: Classification Report for Case2 with different average metrics

Type of Average measure	Type of Classifier	Precision	Recall	f1-score	Support
micro-average	Gaussian Naïve Bayes	0.30	0.30	0.30	99
	Multinomial Naïve Bayes	0.28	0.28	0.28	99
	Complement Naïve Bayes	0.27	0.27	0.27	99

	SVM/SVC	0.49	0.49	0.49	99
	K- Nearest Neighbors	0.44	0.44	0.44	99
	Bagging	0.37	0.37	0.37	99
	Ada Boost	0.32	0.32	0.32	99
	Decision Tree	0.39	0.39	0.39	99
	Extra Trees	0.40	0.40	0.40	99
	Random Forest	0.40	0.40	0.40	99
macro-average	Gaussian Naïve Bayes	0.37	0.35	0.28	99
	Multinomial Naïve Bayes	0.32	0.30	0.28	99
	Complement Naïve Bayes	0.30	0.31	0.26	99
	SVM/SVC	0.55	0.47	0.46	99
	K- Nearest Neighbors	0.45	0.44	0.45	99
	Bagging	0.38	0.37	0.37	99
	Ada Boost	0.33	0.33	0.32	99
	Decision Tree	0.39	0.39	0.39	99
	Extra Trees	0.40	0.40	0.40	99
	Random Forest	0.41	0.40	0.40	99
weighted-average	Gaussian Naïve Bayes	0.38	0.30	0.26	99
	Multinomial Naïve Bayes	0.33	0.28	0.27	99
	Complement Naïve Bayes	0.31	0.27	0.25	99
	SVM/SVC	0.54	0.49	0.46	99
	K- Nearest Neighbors	0.45	0.44	0.44	99
	Bagging	0.37	0.37	0.37	99
	Ada Boost	0.34	0.32	0.33	99
	Decision Tree	0.40	0.39	0.39	99
	Extra Trees	0.41	0.40	0.41	99
	Random Forest	0.40	0.40	0.40	99

From Table 9, it is identified that for all the labels, “low”, “medium”, and “high”, the same classifier, that is, SVC provides highest values for all the three metrics. Thus, it is concluded that SVM has better performance in classification for the case of “knowledge gain” by the user.

The other significant factor for assessing the classifier performance is the accuracy. It is calculated along with the fitment of classifiers. The Table 10 will present these values for different classification models.

Table 10: Classification Accuracy and Training Fitment for Case 2

Type of Classifier	Accuracy	Training Fitment
Gaussian Naïve Bayes	0.30303	0.360406091
Multinomial Naïve Bayes	0.28283	0.35786802
Complement Naïve Bayes	0.27273	0.373096447
SVM/SVC	0.49495	0.860406091
K- Nearest Neighbors	0.44444	0.83248731
Bagging	0.37374	0.847715736
Ada Boost	0.32323	0.609137056
Decision Tree	0.39394	0.860406091
Extra Trees	0.40404	0.860406091
Random Forest	0.40404	0.845177665

The graphical representation of Table 10 is presented in Figure 3.

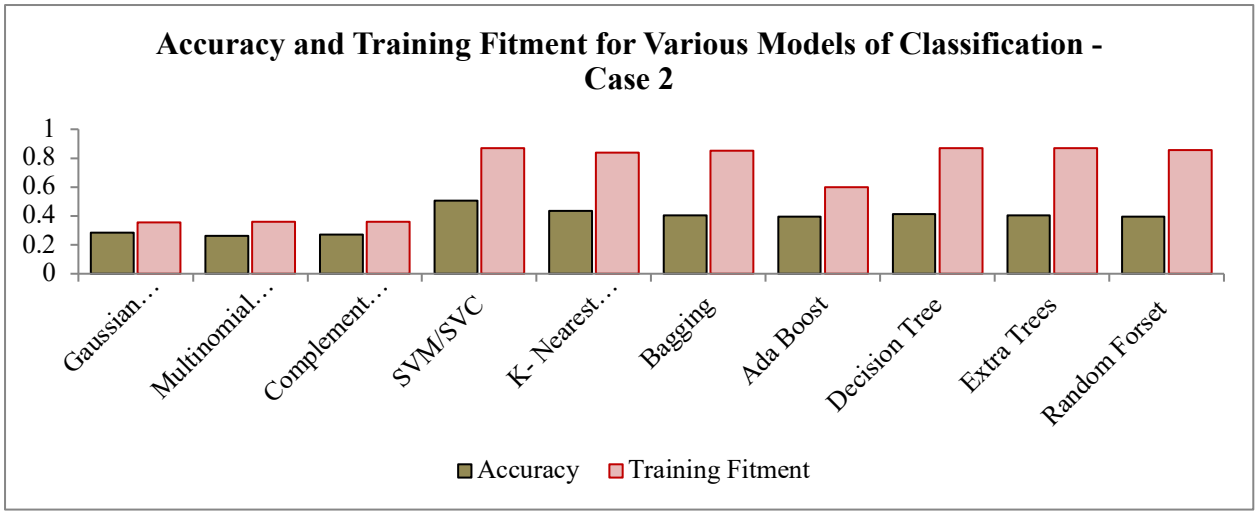

Figure 3: Accuracy and Training Fitment for Case 2

It is observed from Table 10 and Figure 3 that the SVC delivers uppermost accuracy of 0.4949 in addition to a fitment of 86.04 %. Although other few classifiers specifically,

Decision Tree, and Extra Trees accomplish the same value of fitment, the accuracy is more exclusively for SVC. Therefore, SVC is preferred as it shows competent performance for the case of identifying the knowledge gain by the users and classify under the three labels such as, “low”, “medium”, and “high”.

5.3 Comparative Analysis with Feature Set 1 and Feature Set 2

The various metrics have been used for the analysis. The considered problem belongs to multiple class. Thus, the metrics that are to be taken for comparative analysis are the Precision, Recall, f1-score through average methods (micro-, macro-, weighted- average methods) and the accuracy.

The classification accuracy of a model is mathematically formulated as,

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{16}$$

It is simply the amount of correct prediction made out of all predictions or total predictions made. Accuracy is an instinctive measure for performance, and for getting reasonable accuracy the dataset should be symmetrical. In such cases, where accuracy is considered to be a main metric the number of samples considered for false positive and false negative should be almost same to get highest accuracy. Hence, other parameters are to be taken care to estimate the model’s performance. Therefore, the comparative analysis not only analyzes the accuracy of the model but also the other metrics like Precision, Recall and f1-score.

“The precision is simply the amount (number) of positive predictions made out of the positive class values”. It is also called by the name “Positive Predicted Value”. Low value of precision is an indication of more false positives. Recall is the sensitivity and it is the true positive rate. The f1-score delivers the balance between Precision and recall.

The comparison is done based on Feature set1 and Feature set2. The Feature set1 comprises of Queries, Query Terms, Query Unique Terms, Session Duration, Time per Query, Average Query Complexity, Total Search Clicks, and Maximum Search Clicks Rank, whereas, the Feature set2 contains Session Duration, Time per Query, Query Terms and Maximum Search Clicks Rank. The Feature set2 is the one that selected through Lasso Regression and considered as important features. The classification models are analyzed to know impact of Feature set2 on the users’ knowledge gain.

The evaluation metrics with regard to Feature set1 and Feature set2 are listed in Tables 11 and 12.

Table 11: Accuracy and Classification with Feature sets 1 and 2

Type of Classifier	Accuracy		Training Fitment	
	Feature Set 1	Feature Set 2	Feature Set 1	Feature Set 2
Gaussian Naïve Bayes	0.28283	0.30303	0.35533	0.36041
Multinomial Naïve Bayes	0.26263	0.28283	0.36041	0.35787
Complement Naïve Bayes	0.27273	0.27273	0.36041	0.37310
SVM/SVC	0.50505	0.49495	0.87056	0.86041
K- Nearest Neighbors	0.43434	0.44444	0.84010	0.83249
Bagging	0.40404	0.37374	0.85279	0.84772
Ada Boost	0.39394	0.32323	0.59898	0.60914
Decision Tree	0.41414	0.39394	0.87056	0.86041
Extra Trees	0.40404	0.40404	0.87056	0.86041
Random Forest	0.39394	0.40404	0.85787	0.84518

Table 12: Other Metrics with Feature sets 1 and 2

Type of Average measure	Type of Classifier	Feature set1			Feature set 2		
		Precision	Recall	f1-score	Precision	Recall	f1-score
micro-average	Gaussian Naïve Bayes	0.28	0.28	0.28	0.30	0.30	0.30
	Multinomial Naïve Bayes	0.26	0.26	0.26	0.28	0.28	0.28
	Complement Naïve Bayes	0.27	0.27	0.27	0.27	0.27	0.27

	SVM/SVC	0.51	0.51	0.51	0.49	0.49	0.49
	K- Nearest Neighbors	0.43	0.43	0.43	0.44	0.44	0.44
	Bagging	0.40	0.40	0.40	0.37	0.37	0.37
	Ada Boost	0.39	0.39	0.39	0.32	0.32	0.32
	Decision Tree	0.41	0.41	0.41	0.39	0.39	0.39
	Extra Trees	0.40	0.40	0.40	0.40	0.40	0.40
	Random Forest	0.39	0.39	0.39	0.40	0.40	0.40
macro-average	Gaussian Naïve Bayes	0.30	0.32	0.26	0.37	0.35	0.28
	Multinomial Naïve Bayes	0.29	0.30	0.25	0.32	0.30	0.28
	Complement Naïve Bayes	0.28	0.32	0.24	0.30	0.31	0.26
	SVM/SVC	0.56	0.48	0.46	0.55	0.47	0.46
	K- Nearest Neighbors	0.44	0.43	0.44	0.45	0.44	0.45
	Bagging	0.40	0.40	0.40	0.38	0.37	0.37
	Ada Boost	0.4	0.38	0.38	0.33	0.33	0.32
	Decision Tree	0.41	0.41	0.41	0.39	0.39	0.39
	Extra Trees	0.40	0.40	0.40	0.40	0.40	0.40
	Random Forest	0.39	0.38	0.38	0.41	0.40	0.40
weighted-average	Gaussian Naïve Bayes	0.31	0.28	0.25	0.38	0.30	0.26
	Multinomial Naïve Bayes	0.30	0.26	0.24	0.33	0.28	0.27
	Complement Naïve Bayes	0.28	0.27	0.22	0.31	0.27	0.25
	SVM/SVC	0.56	0.51	0.47	0.54	0.49	0.46
	K- Nearest Neighbors	0.44	0.43	0.43	0.45	0.44	0.44
	Bagging	0.41	0.40	0.40	0.37	0.37	0.37
	Ada Boost	0.40	0.39	0.38	0.34	0.32	0.33
	Decision Tree	0.41	0.41	0.41	0.40	0.39	0.39
	Extra Trees	0.41	0.40	0.41	0.41	0.40	0.41
	Random Forest	0.39	0.39	0.39	0.40	0.40	0.40

It is evident from the Tables 11, 12 that there is an improvement of evaluation metrics for four classifiers and the values are same for two classifier models, and there is slight

decrement in the values of the metrics for the four classification models. The percentage of improvement alone is projected in Tables 13 and 14.

Table 13: Percentage Improvement in Accuracy and Training Fitment using Feature Set2

Type of Classifier	Improvement Percentage
Accuracy	
Gaussian Naïve Bayes	7.14
Multinomial Naïve Bayes	7.69
K- Nearest Neighbors	2.33
Random Forest	2.56
Training Fitment	
Gaussian Naïve Bayes	1.41
Complement Naïve Bayes	3.52
Ada Boost	1.69

Table 14: Percentage Improvement in Other Metrics using Feature Set2

Type of Average measure	Type of Classifier	Percentage Improvement		
		Precision	Recall	f1-score
micro-average	Gaussian Naïve Bayes	7.1	7.1	7.1
	Multinomial Naïve Bayes	7.7	7.7	7.7
	K- Nearest Neighbors	2.3	2.3	2.3
	Random Forest	2.6	2.6	2.6
macro-average	Gaussian Naïve Bayes	23.3	9.4	7.7
	Multinomial Naïve Bayes	10.3	0.0	12.0
	K- Nearest Neighbors	2.3	2.3	2.3
	Random Forest	5.1	5.3	5.3
weighted-average	Gaussian Naïve Bayes	22.6	7.1	4.0
	Multinomial Naïve Bayes	10.0	7.7	12.5
	Complement Naïve Bayes	10.7	0.0	13.6
	Random Forest	2.6	2.6	2.6

The graphical representation of Tables 13 and 14 are presented in Figures 4-8.

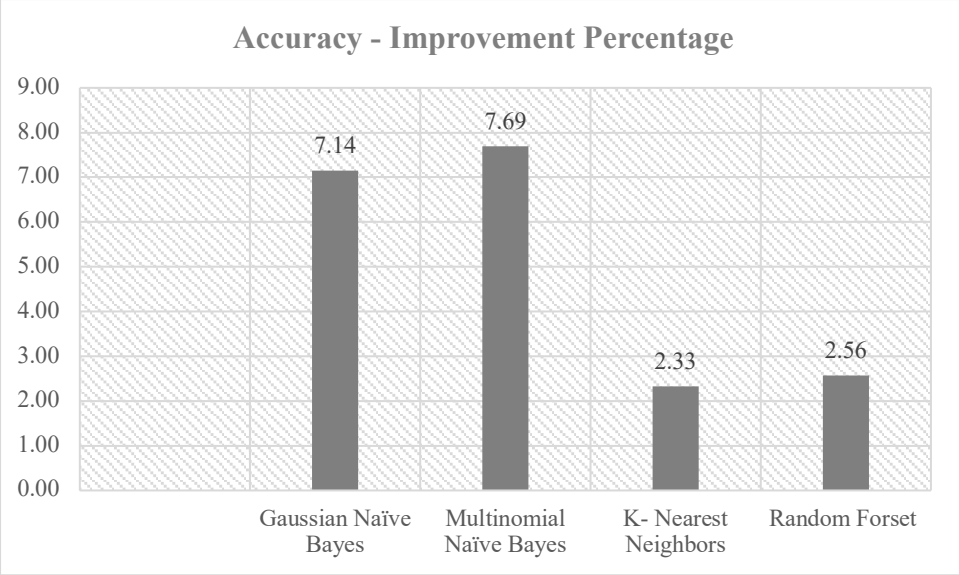

Figure 4: Improvement in Accuracy with Feature Set 2

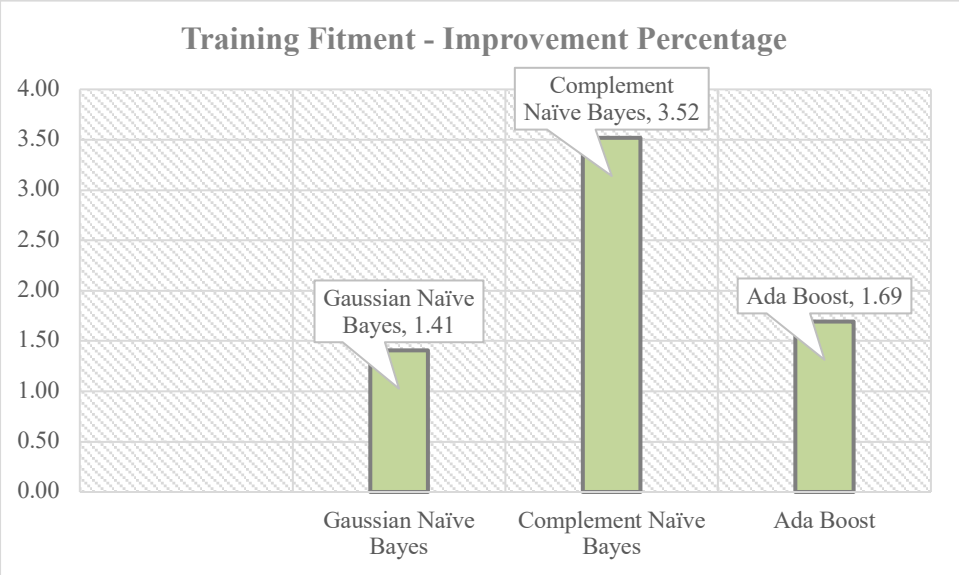

Figure 5: Improvement in Training Fitment with Feature Set 2

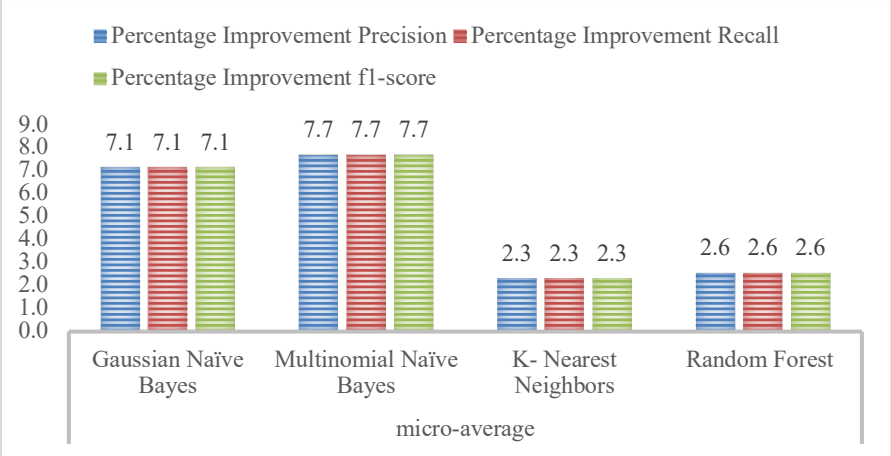

Figure 6: Improvement in Other Metrics with Feature Set 2 (micro-average)

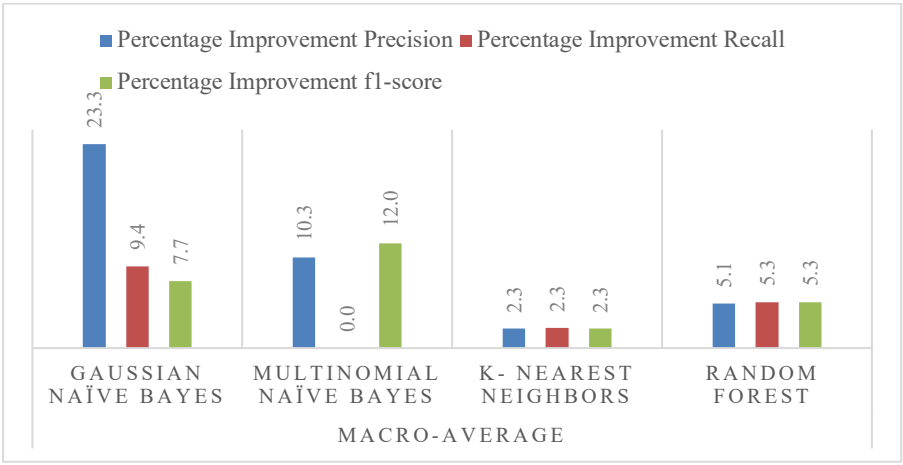

Figure 7: Improvement in Other Metrics with Feature Set 2 (macro-average)

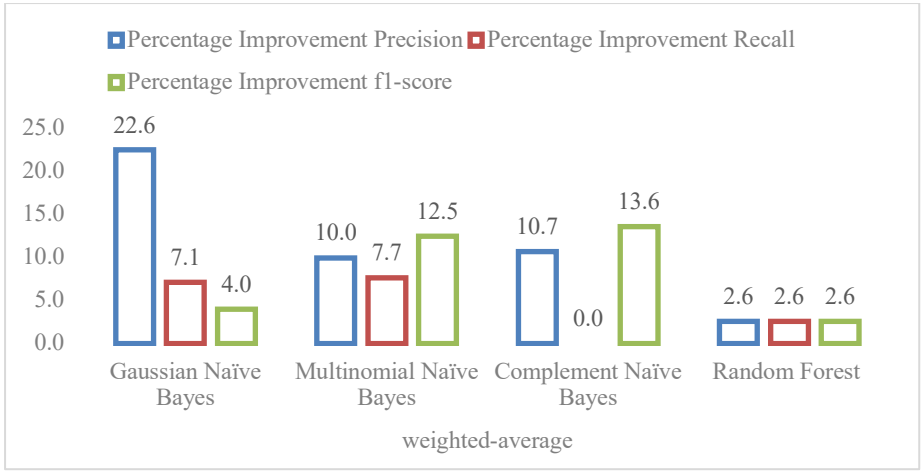

Figure 8: Improvement in Other Metrics with Feature Set 2 (weighted-average)

From Figures 4-8, it is seen that four of the eight numbers of classifier models with feature

set2 are with superior performance. All the classifier models have categorized knowledge gain of the user with labels, “low”, “medium” and “high”. The evaluations metrics are identical for two classifier models with both feature set1 and feature set2.

Accuracy and Training Fitment

The Gaussian Naïve Bayes, Multinomial Naïve Bayes, K-Nearest Neighbor and the Random Forest Models are exhibiting improved accuracy with features set2. The models Gaussian Naïve Bayes, Complement Naïve Bayes and Ada boost are working with enhanced training fitment with regard to feature set2. The highest percentage of improvement of 7.69 has been attained with Multinomial Naïve Bayes and the minimum percentage of improvement is with classifier model K-Nearest Neighbor with a value of 2.33. The range of training fitment improvement is 1.41 to 3.52. Maximum fitness is observed for Complement Naïve Bayes classifier model.

Precision, Recall and f1-Score (micro-average)

The average precision, average recall and average f1-score using micro average method are improved with feature set2, for four models. The range of improvement is same for all the three metrics and is 2.3 to 7.7. The highest improvement is attained with Multinomial Naïve Bayes classifier model.

Precision, Recall and f1-Score (macro-average)

The average precision, average recall and average f1-score using macro average method are improved with feature set2, for four models. The range of improvement for average precision is 2.3 to 23.33. The highest improvement is attained with Gaussian Naïve Bayes classifier model. The maximum improvement percentage in average recall is 9.4 and is achieved again using Gaussian Naïve Bayes classifier model. The range of improvement for average recall is 2.3 to 9.4. The range of improvement in f1-score with feature set2 is 2.3 to 12.0. The highest improvement is established with Multinomial Naïve Bayes classifier model.

Precision, Recall and f1-Score (weighted-average)

The average precision, average recall and average f1-score using macro average method are improved with feature set2, for four models. The range of improvement for average precision is 2.6 to 22.60. The highest improvement is attained with Gaussian Naïve Bayes classifier model. The maximum improvement percentage in average recall is 7.7 and is achieved using Multinomial Naïve Bayes classifier model. The range of improvement for average recall is 2.3 to 9.4. The range of improvement in f1-score with feature set2 is 2.6 to 12.5. The highest improvement is obtained by means of Multinomial Naïve Bayes classifier model.

5.4 Summary

The crowdsource dataset publicly available has been considered for testing the classification model intended for assessing the knowledge gain of the internet user with informational search queries with 11topics involving 468 distinct searchers. The queries presented by them have been analyzed during session including the search clicks, session duration, etc. There are many such features for analyzing the search log. However, including all feature will increase the feature vector size and in turn increases the time of computation. Thus, only selective features are included. The feature selection is carried out with regression models with three type. The number of features is minimum in Lasso Regression with minimum RMSE. Two different feature sets are taken for the analysis. The set1 of features are 8 random features. The set2 of features are just 4 and they have been decided by the Lasso Regression Model. These features are fed to 10 classification models. Three different categories of classification have been designed. All the models performed well. The evaluation metrics for all the multiple class classifier has been computed. The impact of feature set2 has been analyzed using the improvement in the evaluation metrics from the classifier models using feature set1 and feature set2.

6 Conclusion

In the present digital world, almost everyone has internet connection and people almost accustomed to use it regularly. They explore the internet by placing some queries using a search engine. By statistical reports, it is found that many are using the internet for getting some information about new topics or to update their knowledge on a specific topic. Based on the queries, the search engine provides the user many search results. There is no specific measure about whether or not the searchers gain knowledge during a search session. In order to help the users and customize the search engine based on the knowledge requirements of users, a search log analysis is performed. The expertise of the users has been analyzed in existing literatures but the features and their impact on knowledge gain is not explored fully. To analyze the same, few imperative features have been chosen with suitable regression models and these features are given as input to the classifier. The classifiers are designed to classify the knowledge of the searcher during a search session into three various labels such as, “low”, “medium” and “high”. Then the performance of the classifiers is measured using evaluation indices. Then a feature set involving more features than the one suggested by Lasso regression is taken and the performance metrics of the same classifiers have been computed with these feature set. A thorough comparative analysis has been done.

6.1 Regression Models

The three models of regression are the linear, ridge and lasso regression models. They are used for selecting important features and discarding the redundant data. Out of the given 23 features, only one feature has been eliminated with linear and ridge regression models. While, the lasso model identified only 5 features as imperative/ significant features and the remaining features are ignored. The evaluation metrics for regression models are RMSE. This value should be minimum for superior models. The RMSE obtained through lasso is minimum and hence that model is considered as suitable and those features alone are chosen.

6.2 Classification Models

This is main part of this dissertation. The objective of this dissertation is to find a suitable

classification model to classify the knowledge gain by the internet searchers in an effective manner. Supervised classification model has been considered and ten different classifiers are tested. The efficiency has been assessed using the performance metrics of the model. The evaluation indices are calculated for every considered model with two features sets. Set1 will have more features and set2 will have the features selected through regression model. A comparison is carried out and the suitable classifiers are identified. In terms of accuracy the Multinomial Naïve Bayes model performs well though SVC has got highest accuracy with feature set2. This is because, while changing the features set with more features, the accuracy has been increased with this model. However, Multinomial model showed improved accuracy with features set2 and the percentage improvement is 7.69. In terms of training fitment, Complement Naïve Bayes showed better improvement. The classifier model considered is of multiple class. Hence, the performance metrics like Precision, Recall and f1-score are to be calculate as an average. Three averaging methods namely, micro-, macro-, and weighted- average methods are used. With regard to micro-average, the precision, recall and f1-score showed improvement in case of Gaussian Naïve Bayes, Multinomial naïve Bayes, K- Nearest Neighbor and Random Forest classifier models. With respect to macro-average, the better classifiers identified with feature set2 are the same four classifiers as in the case of micr0-average.
In the weighted- average case, the better performing classifiers identified with feature set2 are Gaussian Naïve Bayes, Multinomial naïve Bayes, Complement Naïve Bayes and Random Forest classifier models.

6.3 Conclusion

Finally, it is concluded that the feature set2 contains only important feature because the model with these features provided least value of RMSE. Lasso Regression works well in comparison to the other two regression models. The classifiers working well with these features are Gaussian Naïve Bayes, Multinomial Naïve Bayes, K- Nearest Neighbor, Complement Naïve Bayes and Random Forest classifier models. These 5 models are decided as better models because of the improvement they have shown in the values of accuracy, precision, recall and f1-score. The other five models either have the same value of evaluation metrics (two models) or reduced values with feature set2. From the analysis, it is seen that the features that are chosen have impact on the classification (prediction) models designed for labelling the knowledge gain of the users during the search session. When the features change, the accuracy and other indices change. With more features, the time of computation and the dimension of the feature set increase. The efficiency of the model is thus affected. The classification (prediction) models have labels than scores. Therefore, the “knowledge gain” is judged using different labels instead of knowledge scores. The three labels of the prediction models are “low”, “medium”, and “high”. Apparently, the label “high” specifies the maximum knowledge gain. The coding is done to get the stated labels using Standard Deviation. With 0.5 SD on either side of “mean” is considered “medium”.

The values lower than “medium” are “low” and higher are “high”. Finally, a good prediction model with significant features can genuinely predict the users’ knowledge gain with a good model for selecting the desirable features and eliminating the unwanted features.

Bibliography

1. Andrei Broder (2002). A taxonomy of web search. *SIGIR Forum*. [Online]. 36 (2). pp. 3–10. Available from: https://www.cis.upenn.edu/~nenkova/Courses/cis430/p3-broder.pdf.

2. Bennett, P.N., Svore, K. & Dumais, S.T. (2010). *Classification-Enhanced Ranking*. [Online]. pp. 1–10. Available from: https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/classEnhancedRanking.pdf.

3. Bhatia, S., Brunk, C. & Mitra, P. (2012). Analysis and automatic classification of web search queries for diversification requirements. *Proceedings of the American Society for Information Science and Technology*. [Online]. 49 (1). pp. 1–10. Available from: http://doi.wiley.com/10.1002/meet.14504901188.

4. Chau, M., Fang, X. & Liu Sheng, O.R. (2005). Analysis of the query logs of a Web site search engine. *Journal of the American Society for Information Science and Technology*. [Online]. 56 (13). pp. 1363–1376. Available from: http://doi.wiley.com/10.1002/asi.20210.

5. Choo, C.W., Detlor, B. & Turnbull, D. (2000). Information seeking on the Web: An integrated model of browsing and searching. *First Monday*. [Online]. 5 (2). Available from: http://journals.uic.edu/ojs/index.php/fm/article/view/729.

6. Duarte Torres, S., Hiemstra, D. & Serdyukov, P. (2010). *Query log analysis in the context of information retrieval for children*. [Online]. Available from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.182.1222&rep=rep1&type=pdf.

7. Eickhoff, C., Teevan, J., White, R. & Dumais, S. (2014). Lessons from the Journey: A Query Log Analysis of Within-Session Learning. In: *Proceedings of the 7th ACM international conference on Web search and data mining - WSDM '14*. [Online]. 2014, New York, New York, USA: ACM Press, pp. 223–232. Available from: https://www.microsoft.com/en-us/research/wp-content/uploads/2016/04/wsdm14.pdf.

8. Feng, C., Wang, H., Lu, N., Chen, T., He, H., Lu, Y. & Tu, X.M. (2014). Log-transformation and its implications for data analysis. *Shanghai archives of psychiatry*. [Online]. 26 (2). pp. 105–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25092958.

9. Gadiraju, U., Yu, R., Dietze, S. & Holtz, P. (2018). Analyzing Knowledge Gain of Users in Informational Search Sessions on the Web. In: *Proceedings of the 2018 Conference on Human Information Interaction&Retrieval - CHIIR '18*. [Online]. 2018, New York, New York, USA: ACM Press, pp. 2–11. Available from: http://dl.acm.org/citation.cfm?doid=3176349.3176381.

10. Hagen, M., Potthast, M., Völske, M., Gomoll, J. & Stein, B. (2016). How Writers Search. In: *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval - CHIIR '16*. [Online]. 2016, New York, New York, USA: ACM Press, pp. 193–202. Available from: http://dl.acm.org/citation.cfm?doid=2854946.2854969.

11. Hendahewa, C. & Shah, C. (2015). Implicit search feature based approach to assist users in exploratory search tasks. *Information Processing & Management*. [Online]. 51 (5). pp. 643–661. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0306457315000795.

12. Hölscher, C. & Strube, G. (2000). Web search behavior of Internet experts and newbies. *Computer Networks*. [Online]. 33 (1–6). pp. 337–346. Available from: http://linkinghub.elsevier.com/retrieve/pii/S1389128600000311.

13. Hwang, G.-J., Tsai, P.-S., Tsai, C.-C. & Tseng, J.C.R. (2008). A novel approach for assisting teachers in analyzing student web-searching behaviors. *Computers & Education*. [Online]. 51 (2). pp. 926–938. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0360131507001121.

14. Jansen, B.J. (2006). Search log analysis: What it is, what's been done, how to do it. *Library & Information Science Research*. [Online]. 28 (3). pp. 407–432. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0740818806000673.

15. Jiang, D., Pei, J. & Li, H. (2013a). Mining search and browse logs for web search. *ACM Transactions on Intelligent Systems and Technology*. [Online]. 4 (4). pp. 1–37. Available from: http://www.hangli-hl.com/uploads/3/1/6/8/3168008/logminingsurvey.pdf.

16. Jiang, D., Pei, J. & Li, H. (2013b). Mining search and browse logs for web search. *ACM Transactions on Intelligent Systems and Technology*. [Online]. 4 (4). pp. 1–37. Available from: http://dl.acm.org/citation.cfm?doid=2508037.2508038.

17. Joshila Grace, L.K., Maheswari, V. & Nagamalai, D. (2011). Analysis of Web Logs And Web User In Web Mining. *International Journal of Network Security & Its Applications*. [Online]. 3 (1). pp. 99–110. Available from: https://arxiv.org/ftp/arxiv/papers/1101/1101.5668.pdf.

18. Kacprzak, E., Koesten, L.M., Ibáñez, L.-D., Simperl, E. & Tennison, J. (2017). *A Query Log Analysis of Dataset Search*. In: [Online]. pp. 429–436. Available from: http://link.springer.com/10.1007/978-3-319-60131-1_29.

19. Li, W. (2013). *Automatic Log Analysis using Machine Learning*. [Online]. (November). Available from: http://www.diva-portal.org/smash/get/diva2:667650/FULLTEXT01.pdf.

20. Li, X., Liu, Y., Cai, R. & Ma, S. (2017). Investigation of User Search Behavior While Facing Heterogeneous Search Services. In: *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining - WSDM '17*. [Online].

2017, New York, New York, USA: ACM Press, pp. 161–170. Available from: http://dl.acm.org/citation.cfm?doid=3018661.3018673.

21. Malik, A. & Mahmood, K. (2009). *Web search behavior of university students: A case study at University of the Punjab*. [Online]. Available from: https://www.researchgate.net/publication/45223728_Web_search_behavior_of_university_students_A_case_study_at_University_of_the_Punjab.

22. Mat Hassan, M. & Levene, M. (2005). Associating search and navigation behavior through log analysis. *Journal of the American Society for Information Science and Technology*. [Online]. 56 (9). pp. 913–934. Available from: http://doi.wiley.com/10.1002/asi.20185.

23. Mishra, N., Saha Roy, R., Ganguly, N., Laxman, S. & Choudhury, M. (2011). Unsupervised query segmentation using only query logs. In: *Proceedings of the 20th international conference companion on World wide web - WWW '11*. [Online]. 2011, New York, New York, USA: ACM Press, pp. 91. Available from: http://portal.acm.org/citation.cfm?doid=1963192.1963239.

24. Moraes, F., Putra, S.R. & Hauff, C. (2018). Contrasting Search as a Learning Activity with Instructor-designed Learning. In: *Proceedings of the 27th ACM International Conference on Information and Knowledge Management - CIKM '18*. [Online]. 2018, New York, New York, USA: ACM Press, pp. 167–176. Available from: http://dl.acm.org/citation.cfm?doid=3269206.3271676.

25. Okhovati, M., Sharifpoor, E., Aazami, M., Zolala, F. & Hamzehzadeh, M. (2016). Novice and experienced users' search performance and satisfaction with Web of Science and Scopus. *Journal of Librarianship and Information Science*. [Online]. 49 (4). pp. 359–367. Available from: http://journals.sagepub.com/doi/10.1177/0961000616656234.

26. Ren, P., Chen, Z., Ma, J., Wang, S., Zhang, Z., Ren, Z. & Ma, T. (2018). User session level diverse reranking of search results. *Neurocomputing*. [Online]. 274. pp. 66–79. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0925231216305537.

27. Sanchiz, M., Amadieu, F., Fu, W.T. & Chevalier, A. (2019). Does pre-activating domain knowledge foster elaborated online information search strategies? Comparisons between young and old web user adults. *Applied Ergonomics*. [Online]. 75. pp. 201–213. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0003687018305738.

28. Spink, A., Wolfram, D., Jansen, M.B.J. & Saracevic, T. (2001). Searching the Web: The Public and Their Queries. *Journal of the American Society for Information Science and Technology*. [Online]. 52 (3). pp. 226–234. Available from: https://tefkos.comminfo.rutgers.edu/JASIST2001.pdf.

29. Teevan, J., Dumais, S.T. & Horvitz, E. (2018). Personalizing Search via Automated Analysis of Interests and Activities. *ACM SIGIR Forum*. [Online]. 51 (3). pp. 10–17. Available from: http://susandumais.com/sigir2005-personalizedsearch.pdf.

30. Theng, Y.-L., Lee, E.A., Chu, S.K.-W., Lee, C.W.Y., Chiu, M.M.-L. & Chan, R.C.H. (2015). Scaffolding in information search: Effects on less experienced searchers. *Journal of Librarianship and Information Science*. [Online]. 48 (2). pp. 177–190. Available from: http://journals.sagepub.com/doi/10.1177/0961000615595455.

31. Trevisan, M., Dini, L., Barbu, E., Barsanti, I., Lagos, N., Segond, F., Rhulmann, M. & Vald, E. (2012). Query log analysis with GALATEAS LangLog. In: *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*. [Online]. 2012, pp. 87–91. Available from: https://www.aclweb.org/anthology/E12-2018.

32. Wang, P., Hawk, W.B. & Tenopir, C. (2000). Users' interaction with World Wide Web resources: an exploratory study using a holistic approach. *Information Processing & Management*. [Online]. 36 (2). pp. 229–251. Available from: http://linkinghub.elsevier.com/retrieve/pii/S030645739900059X.

33. White, R.W., Dumais, S.T. & Teevan, J. (2009). Characterizing the influence of domain expertise on web search behavior. In: *Proceedings of the Second ACM International Conference on Web Search and Data Mining - WSDM '09*. [Online]. 2009, New York, New York, USA: ACM Press, pp. 132. Available from: http://portal.acm.org/citation.cfm?doid=1498759.1498819.

34. Yu, R., Gadiraju, U. & Dietze, S. (2018a). Detecting, Understanding and Supporting Everyday Learning in Web Search. *Engineering Letters*. [Online]. pp. 1–6. Available from: https://arxiv.org/pdf/1806.11046.pdf.

35. Yu, R., Gadiraju, U., Holtz, P., Rokicki, M., Kemkes, P. & Dietze, S. (2018b). Predicting User Knowledge Gain in Informational Search Sessions. In: *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval - SIGIR '18*. [Online]. 2018, New York, New York, USA: ACM Press, pp. 75–84. Available from: http://dl.acm.org/citation.cfm?doid=3209978.3210064.

36. Zhu, Z., Levene, M. & Cox, I.J. (2009). Query classification using asymmetric learning. In: *2009 Second International Conference on the Applications of Digital Information and Web Technologies*. [Online]. August 2009, IEEE, pp. 518–524. Available from: http://ieeexplore.ieee.org/document/5273856/.