2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)
Amity University, Noida, India. June 4-5, 2020

Integrating "Random Forest" with Indexing and Query Processing for Personalized Search

Hussain Naeem Nawazish
Student, Bachelor of Science
Information Technology
Amity University, Dubai, UAE
hussainN@amitydubai.ae

Vinod Kumar Shukla
Department of Engineering and Architecture
Amity University, Dubai, UAE
vshukla@amityuniversity.ae

***Abstract*: The internet has become an integral part of at least 4.4 billion lives. An average person looks at their device at least 20 times a day. One can only imagine the amount of queries a search engine gets on a daily basis. With the help of all the data acquired over the years, the internet updates us with all the biggest trends and live events happening all over the world. A search engine is able to provide query suggestions based on the number of times a keyword has been searched for or the current query relates to a certain trend. All these trends are updated to every device internationally or locally. This concept is generalized throughout all devices that use any kind of search engine on any application. Through this paper we intend to propose to use Random Forest as a predictive model to be integrated with the indexing process of the search engine to produce query suggestions that a user would want to search, contrary to the query suggestions that are usually displayed based on hyped trends and fashion.**

Keywords:-Random forest, Search Engine, Predictive Analysis, Query Processing, Personalize search, Indexing

I. INTRODUCTION

Different Internet search engines have different algorithms to determine which web pages are most relevant to a search engine, and which web pages should appear at the top of the results page of the search engine. Relevance is the key for online search engines–users obviously prefer a search engine that provides the best and best results. Search engines often have their search algorithms well covered, as their particular algorithm aims to produce the most relevant results. The most important search engines are the best and often the most popular search engines.

Over the past decade, search engines have developed in order to autocomplete[1] a search query using the most optimal query suggestions[2] in order to provide the most relevant result possible. User experience personalization creates huge opportunities to present the user with information which is much more useful and handier in contrary to query suggestion that may relate to mainly what is in trend or fashion.

A web search usually returns thousands of results in response to a query. This makes it difficult for the user to find relevant information needed. To produce such effective result pages, clustering is a great data-mining [3] tool to group and evaluate documents based on meaningful categorization using keywords and their correlating document and web clicks.

This paper aims to integrate personalization for search engines using random forest as the predictive model to help in enhancing user experience by providing effective query suggestions, hence, more relevant, useful, and accurate search results.

II. LITERATURE REVIEW

Predictive Modeling has been used by Robert Ficcaglia et al., [4] as an apparatus for location based commercial services so as to gain user location user information which can then be used as a predictor of a consumer demographic profile.

Vrushali Kulkarni et al., [5] discuss the future research directions and improvements for random forest algorithm in terms of accuracy, performance, online data, etc.

It is important to note that queries and documents have varying size of information. It is usually hard for predictive models to work on short text with lack of contextual information while data enrichment creating noise. Ameni Bouaziz et al., [6] proposed to combine data enrichment with the introduction of semantics in Random Forests where each short text is enriched with data semantically similar to its words.

The Ontology-Based, Multi-Facet (OMF) personalization framework as proposed by Kenneth Leung et al., [7] for automatically extracting user's content and location preferences based on the user's click through history.

Kenneth Leung et al., also introduce the notion of content and location entropies to measure the diversity of content and location associated with a query and click content to know the user's interest types [7].

In an attempt to improve web search personalization, Rakesh Kumar et al., [8] proposed the system to continuously update the user profile in order to build an enhanced user profile which would be used suggest relevant search results instead of the generic user profile.

The system used by Thomas Hofmann et al. [9] to integrate query to be searched with user profile, query logs from the search history and user reviews for the items and user expectations for the unrated items and documents.

With the use predictive query completion predictive search results invented by Othar Hansson et al. [10] to provide the

978-1-7281-7016-9/20/$31.00 ©2020 IEEE

continuous rendering of query suggestion based on the user's input.

III. PREDICTIVE MODELING

Predictive Modeling is a technique that uses statistics for prediction. Each model consists of several predictors, which re-variables that will affect future outcomes. This also helps to reduce the cost for companies which incurred in business outcomes, environmental factors, competitive intelligence, and market conditions.

It aims to work upon the provided information to reach an end conclusion after an event has been triggered. Both historical and existing data are triggered to find patterns and behaviors.

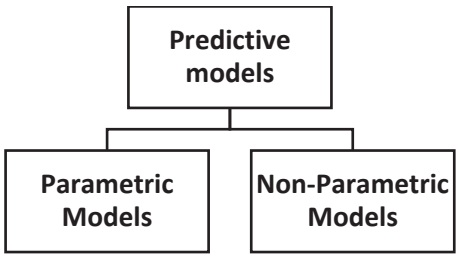

Fig. 1. Types of Predictive models

Predictive models fall into two camps: parametric and non-parametric models (Fig. 1). Parametric models do not require much data to train an algorithm which in practice is unlikely to produce an appropriate mapping function and are highly constrained to a specified function.

Non-Parametric models on the other hand are good when you have a lot of data and no prior knowledge, and you don't want to worry about choosing the right features [11]. This model is capable of exercising a greater number of functional forms and aims for higher performance and accuracy.

Specifically, some of the different types of predictive models are as follows:

Random Forests: Based on model aggregation ideas, for both classification and regression problems, Breiman's Random Foresting is a popular algorithm. The Random Forests principle is the combined use of several samples from the learning sample in many binary decision trees, and the selection of an explanatory variable at each node is random.

It offers greater precision. The random forest classifier handles missing values and keeps a large proportion of data accurate. If more trees exist, they will not allow the model to over fit trees. It can handle a large data set with a higher dimension.

Online Forest Random Algorithms produce online decisions based on online bagging principles and highly randomized trees.

Regression: Regression is a collection of statistical methods to test relations between a dependent variable and one or more independent variables. It is commonly used in modeling and forecasting, where its use is growing significantly with machine learning. This is done by establishing a comprehensive dataset between the relation of dependent and independent variables. This enables to predict definitive true and false values using the confusion matrix.

Classification: A model of classification tries to draw a conclusion from the values observed. Given one or more inputs, one or more results will be predicted by a classification model. The results are labels to which a dataset can be applied. A training dataset is fed into the algorithm. This lets the algorithm know what is, for example, "relevant". The data is compared with other documents to determine whether they are "irrelevant".

Decision Trees: The algorithm for Decision Trees is used to solve regression and problem classification. The primary objective of the decision-making tree is to establish a model for training. This training model is used to predict the target variable values or classes. The algorithm for decision-making trees is much easier to understand than other classification algorithms.

IV. INDEXING PROCESS

The process of arranging and recording all the words of a webpage along with its location and address is referred as Indexing. Search engine also records the number of time each work in occurring in each page. Very quick and related example can be taken as the index of book.

This involves text acquisition from multiple sources which is used for document data storage and text transformation. Text transformation transforms document into index terms or features [12]. Following terms and feature helps to create indexes supporting fast searching and document ranking.

- *Web crawlers* to follow link to find documents for site search and vertical search and document crawlers for system and desktop search.
- *Live feeds* such as real-time streams of document and RSS reader
- *Document Data Store* that stores text, metadata, and other related content document, and also providing fast access to document contents for search engines components such as result list generation.

Text Transformation: Text transformation involves the classification as heading, links, and titles of series of structural elements in the documents. Remove common words to increase search efficiency and effectiveness. And the radiation of the words of the questions varies from language to language. Using links and anchor texts in web sites, it describes reputation and public data such as page rank.

V. QUERY PROCESS

The query must be converted to numbers, to allow your request to be processed by the engine. However, the search engine will remove several terms before it converts into numbers. Most search engines have a stop word list, which is not searched for words. Most search engines won't look for, and, it, it, will, and so on. These brief words are only computer filling. You must include these words in quotation marks, or you must add a plus sign prior to the term in Google. The

engine can then mathematically calculate what indexed terms are closest to what you requested when the terms are converted to numbers. The algorithms are complex, but they give back items based on the closeness to your query mathematically. Those closer to the list of results are listed above. Some motors are even relevant for a percentage.

The higher relevance ratings are: if the word is in the title instead of in the text, the word occurs in bold or italics on the page, how often the word is written on a page, the number and the quality of links on the page, and if the words occur in the header.

VI. Query Suggestion

Search systems can provide search suggestions to users to help users satisfy their informational needs [10]. As used herein, the term 'Query Suggestion' is a Suggested data for a query that can be used to refine a search or refine a search strategy [10].

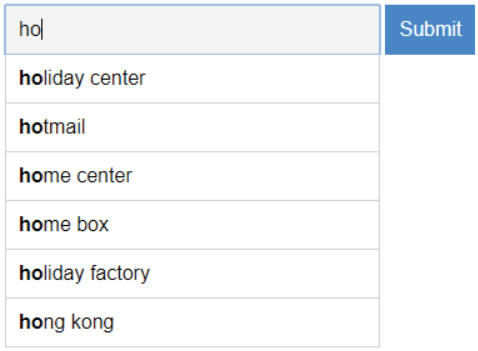

Fig. 2. Example for Query Suggestion

Query Suggestion is a well-known UI feature that most of the highly effective websites provide. When a user provides a query, a dropdown menu appears with a lot of specific – and more widespread – queries (Fig.2).

Normally, modern search engines and tools such as browser toolbars provide query recommendations for request prefixes. For instance, if a user enters the non-prefix, the search engine may suggest one or more queries with the 'ho' prefix, such as holiday center, 'Hotmail', 'home center', 'home box', 'Hong Kong', etc. Typically, by referring to a prefix table, the search engine generates the Suggested query terms. For example, the table can contain a row entry for the 'ho' prefix with possible queries with the 'hom' prefix.

VII. Personalized Search

Personalized search is the customization of a search result through a filter that includes the user's search history, address, click through history, and other preferences. The user's interests may be generated from multiple systems, including those which are not connected to the search system. The search requirement may be narrowed or otherwise changed as a result of the actual application obtained, on the basis of the searcher's interests. The search results may then be sent to the search engine as the search results sought. Personalizing search result through user's location show significant growth on search relevance and user experience.

VIII. Flow Chart for Predictive Analysis

The following figure (Fig.3) implies that the Random forest is used as the predictive model which is integrated with index and query processing.

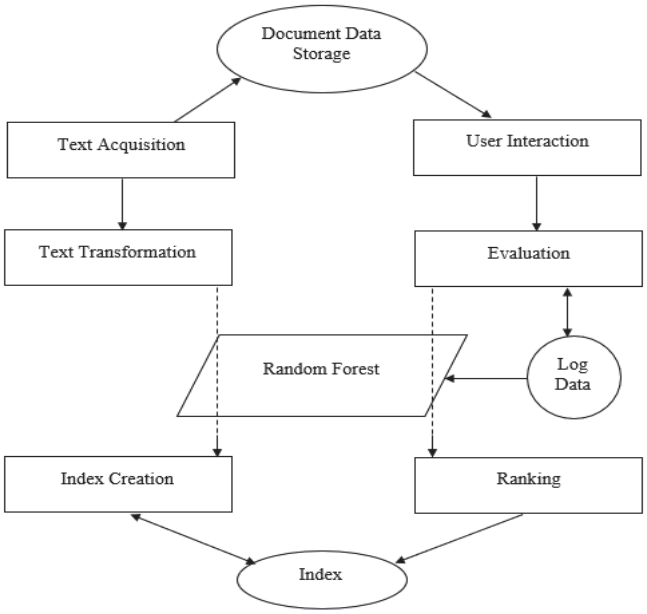

Fig. 3. Incorporation of Indexing and Query Process with Random Forest

The text transformation and evaluation are the processes that go the algorithm to create an index and ranking of documents based on the user's preferences and log data.

To implement the procedure given in the above figure, a dataset regarding user profile and search history was gathered (Fig.4).

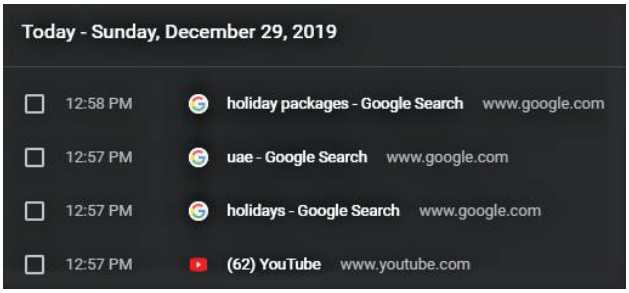

Fig. 4. Recent Search History

The query suggestions shown in Fig. 5 are attained by the search history shown in Fig. 4 which are the keywords that are processed through the indexing process as shown in Fig. 4.

The decisions taken by the search engine are based on the factors that determine personalized search results i.e. Keywords, Address, and History. The algorithm applied is Breiman's Random Forest to classify the text acquired relating to the keywords attained through user interaction.

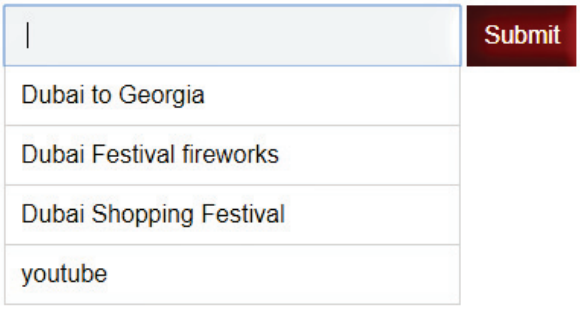

Fig. 5. Query Suggestion Based on Search History

The log data that includes the user behavior with past search history, address, and click through to enhance the accuracy of the algorithm. Any query completion suggested is purely based on the user's search history and preferences as *pre-query suggestion* i.e. before the user's initial input in search interface.

IX. Conclusion

Through this paper I hereby propose solution to personalize the query suggestions by integrating Random forest model with index and query processing data mining techniques to improve search result retrieval that are generally based on latest trends and fashion. Integrating a predictive model with personalized searching allows increasing accuracy and discovering aspects in order to enhance user experience while also providing motivation to display or create more engaging content for the user rather than to enhance superficial ranking in the search engine. The future scope in implementing predictive modeling to personalized searching can allow to focus on forecasting documents that will be released based on the user's search and click-through history.

Reference

[1] Weininger, N.B., Cornea, R.C., Markovich, Y., Zinenko, D. and Fey, N.G., Google LLC, 2013. *Presenting autocomplete suggestions*. U.S. Patent 8,601,019.

[2] Harman, M. (2010). The relationship between search based software engineering and predictive modeling. *Proceedings of the 6th International Conference on Predictive Models in Software Engineering – PROMISE 10*. doi: 10.1145/1868328.1868330

[3] Aggarwal, C. C., & Zhai, C. (2012). An Introduction to Text Mining. *Mining Text Data*, 1–10. doi: 10.1007/978-1-4614-3223-4_1

[4] Ficcaglia, R. and Zapata, D., MotivePath Inc, 2009. System, method and apparatus for predictive modeling of spatially distributed data for location based commercial services.

[5] Kulkarni, V.Y. and Sinha, P.K., 2012, July. Pruning of random forest classifiers: A survey and future directions. In *2012 International Conference on Data Science & Engineering (ICDSE)* (pp. 64-68). IEEE.

[6] Bouaziz, A., Dartigues-Pallez, C., da Costa Pereira, C., Precioso, F. and Lloret, P., 2014, September. Short text classification using semantic random forest. In *International Conference on Data Warehousing and Knowledge Discovery* (pp. 288-299). Springer, Cham.

[7] Leung, K.W.T., Lee, D.L. and Lee, W.C., 2010, March. Personalized web search with location preferences. In *2010 IEEE 26th International Conference on Data Engineering (ICDE 2010)* (pp. 701-712). IEEE.

[8] Kumar, R. and Sharan, A., 2014, February. Personalized web search using browsing history and domain knowledge. In *2014 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT)* (pp. 493-497). IEEE.

[9] Hofmann, T. and Puzicha, J.C., Recommind Inc, 2008. *System and method for personalized search, information filtering, and for generating recommendations utilizing statistical latent class models*. U.S. Patent 7,328,216.

[10] Hansson, O., Black, D., Wiley, J.M., Tungare, M., Mahkovec, Z., McMahan, B.J., Gomes, B.A., Effrat, J.J., Wright, J.R. and Wichary, M.K., Google LLC, 2014. *Predictive query completion and predictive search results*. U.S. Patent 8,706,750.

[11] Russell, S.J. and Norvig, P., 2016. *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited,.

[12] Wesley, Addison, Metzler, Donald, and Leuski, Anton, "Applications of Natural Language Processing – Information Retrieval", USCICT, 2012