ISSN 2250-3153

Optimizing the relevancy of Predictions using Machine Learning and NLP of Search Query

Kilari Murali krishna Teja

Abstract- One of the most important and promising branches of Artificial Intelligence (AI) is Machine Learning (ML),which strive to make a machine intelligent by "learning" from the data. Information Retrieval is also a popular and predominant technique having as one of its application, the ubiquitous Search Engine. Search Engine optimization (SEO) has seen remarkable advancements during the recent years. The objective of this paper is to optimize the existing predictive search mechanism by incorporating pattern based Machine Learning techniques, the association with a Semantic Database, Natural Language Processing of search query to produce more relevant predictions to the user. The main intention is to provide diversified but apt, intelligent predictions for both the diversified set of users whose domain of search queries is not constrained, as well as for the dedicated researchers whose domain will be confined, coupled with an optimal balance between the Response times, Relevancy of predictions.

Index Terms- Intelligent Information Retrieval (IIR); Machine Learning (ML); Prediction Optimization; Semantic database; Natural Language Processing (NLP).

I. INTRODUCTION

What started as a packet switching network by the name of ARPANET in 1963 has undergone evolution into the present day INTERNET and has revolutionized the way of living, the way of working of humans. The impact of Internet is so huge that around 70% of world's population uses the Internet. If it was not for the passion of the researchers, this huge development would not have been possible. In modern days, Researchers can find the helpful information in the Internet; share their thoughts, ideas with their peers over the Internet.

Information Retrieval (IR) is the means to effectively find the information in the Internet and Search Engine is the primary application of Information Retrieval. Even though the foundations for Search Engines has been laid with "Archie" from 1990, they have undergone noticeable changes from the year 1994.In 1996 ,in the paper of Larry page and Sergey Brin they have discussed about a search engine mechanism that uses page rank algorithm to rank the documents in a relevant manner. This Search Engine was named "GOOGLE" and this has been one of the most widely used search engine from the past 14 years because of its distinctive features such as the Autocorrect, Prediction mechanisms, Response Rate, Relevancy as well as Reliability.

The fundamental branches of Artificial Intelligence like Machine Learning (ML), Natural Language Processing (NPL) can be applied to the Search Engines in order to receive relevant, accurate suggestions at appropriate location and time, by making the system intelligent which is achieved by the association of a Dynamic semantic web database which is populated with huge amount of knowledge and also by performing Keyword tagging, Labeling parts of Speech, Prepositional Merging to provide the Predictions which are accurate depending on the history, cache of the user as well as his queries in the current session.

A. MACHINE LEARNING

Machine learning [2] is about "Learning to do better in the future based on what was experienced in the past without the need of explicit programming*"*. The learning that is being done is always based on some sort of observations or data, such as examples, direct experience or instruction. Machine learning focuses on the development of computer programs that can teach themselves to grow and change when exposed to new data.

Machine learning studies computer algorithms for learning to do stuff .The goal is to devise learning algorithms that do the learning automatically without human intervention or assistance. A machine learning system search through data to look for patterns and uses that data to improve the program's own understanding. It is a core subarea of Artificial Intelligence.

B. NATURAL LANGUAGE PROCESSING (NLP)

Natural language processing (NLP) [1] is a "field of artificial intelligence and linguistics concerned with analyzing and representing naturally occurring texts at one or more levels of linguistic analysis to achieve human-like language processing".

As such, NLP is related to the area of Human–Computer Interaction (HCI).The main challenge lies in enabling the computers to derive meaning from human or natural language input. In order to derive meaning, various operations like Segmentation, Parts of Speech labeling or Tagging, tense conversion have to be performed on the language. NLP systems are most widely used in Artificial Intelligence applications.

II. QUERY PROCESSING USING MACHINE LEARNING AND NLP

This paper presents a novel approach to provide accurate predictions to the user by making use of his/her previous search queries which are present in the history, cache, search log and the queries of the current session. The efficiency of this method increases with the cohesiveness of the pattern of the queries that the user search.

For Example, Consider a user having a habit of searching for a particular stock price at a specific time in the morning say between 7.30AM to 8.30AM. Every day the user may not search for a particular stock but will most probably search for stocks that were present in the web pages he visited previously and have had profit. The existing search engine mechanism does not provide any predictions based on this data, but with the concept

of supervised machine learning, one can train the search engine to return more relevant, appropriate predictions.

The objective of this paper is to propose a model which learn and analyses the pattern of the queries depending on the time, location, the frequency of a particular query at particular time and location, the system then compares the pattern with the other similar patterns which it may have obtained from other users by using a semantic web database and then performs NLP by using Parts of Speech tagging, Tense conversion, Prepositional merging to increase the probability of relevant prediction results .This can be incorporated with an extension in a web browser which when enabled carries the above operations and returns the results to the user when he/she tries to enter a query in the search field again.

This method is hugely relies on the efficiency, reliability, correctness of the Semantic Web database and the communication latencies between the user and Search engine. Since, the modern day high performance web can execute distributed and parallel computations, the performance factors of the existing web servers will suffice for this type of Query processing. The Sematic Web database is a dynamic database having huge collection of knowledge populated into it. It consists of relationships among various members, the various meaning of a particular word out of which, only one will be fitting for a particular context in a query. In order to obtain satisfactory performance, the Network communication rates must be sufficiently high, meaning that an Internet connection having a transfer rate less than a minimum threshold value will not result in a timely response. As per the report of Akamai, a global content delivery network the average internet download speed of the world is 3.1 Mbps (raising 4% from the previous quarter. The Query processing model discussed in this paper would require at least a 1Mbps internet connection so as to accept a query, process it and provide appropriate suggestions in a minimum time frame.

When concerned with the predictions within a session, the system realistically assumes that the user takes some time to find the required information in a web page, which is one of the results of his first query in the session and this model will take some amount of time to perform NLP and to obtain relevant predictions and to compare them with the existing patterns and returns the eventual predictions to the user when he starts to type his next search query. The advantage of this model is that these computations occur in background and the probability that the time taken to compute NLP and to return the end results being more than the time between consecutive searches is very less.

A. TAXONOMY OF MACHINE LEARNING ALGORITHMS

Depending on the desired output of each algorithm, the Machine Learning algorithms are commonly classified into the following types,

1) SUPERVISED LEARNING:

In this type of learning, the algorithm generates a function that maps inputs to desired outputs. For some examples, the correct results are known and are given in input to the algorithmic model during the learning process. This type of learning is fast and accurate.

2) UNSUPERVISED LEARNING:

Unsupervised Learning seems much harder, the goal is to have the computer learn how to do something but we don't tell it how to do. Clustering is the most common method in Unsupervised Learning

3) REINFORCEMENT LEARNING:

Reinforcement learning algorithm learns a policy of how to act given an observation of the world. Every action has some impact in the environment, and the environment provides feedback that guides the learning algorithm.

There are also other classes of Machine Learning such as Semi-Supervised Learning, Transduction, Learning to learn but the above mentioned models are the important ones and widely used algorithms

III. EXISTING MODEL

The Predictions given by an existing Search Engine model are shown below. As shown below, the predictions offered to the user are provided by using the information obtained from millions of users of search engine and are not relevant to the customized needs of the user. These suggestions are simply the most frequently searched terms that start with the initial letter the user types and they transform dynamically into words and sentences changing with each letter or alphabet typed before them.

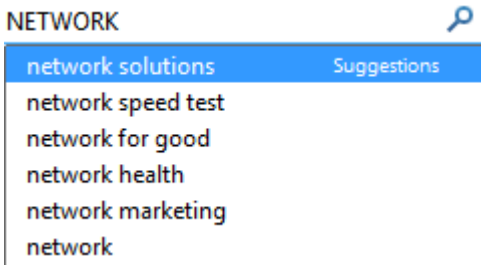

Figure 1: Predictions for the word "NETWORK" irrespective of the context, pattern of the search queries

A. DISADVANTAGES OF EXISTING MODEL

The existing system provides predictions which are generated by consolidating the most frequently searched terms, and giving the results that start with the letter user types, iteratively. The main disadvantage is that, they are not appropriate to the needs of the user in most of the cases and in almost all of the cases related to researchers whose domain of search will be a specific field.

For Example, In the above context, the user may be researching about Network Simulator for a significant amount of time but the next time he tries to search about that, he is provided with the same suggestions that are provided to every other user, whose queries starts with "NETWORK", thus wasting valuable time due to waiting by the user to type the required specific query.

IV. PROPOSED MODEL

The proposed model overcomes the above disadvantage by providing relevant predictions[10] [3] with respect to Time, Location or both Time and Location, it also includes additional features such as providing predictions with words that were not present in the previous queries and provide different meaningful

forms of the words found in the previous queries. These additional words are extracted and analyzed from the previous Search queries made by the user, as well as from the logs present with the web server. However, In a Session based approach, these words may be "*what is*", *"How to"," Steps for"* and other predominantly used words for informational queries. These words not only include the above three but are also extracted from the most frequently shown suggestions to provide a mix of suggestions which are both specific to the users previous query but appropriate and related to his future ones.

This model makes use of the queries made by the user which are present in the history, cache or Search Logs stored by the Web Server to analyze the pattern with the help of a semantic web database, Perform NLP on the query and calculate the Probability of a particular prediction, and then rank the predictions in the order of decreasing probability. Each prediction is associated with certain unique fields. The fields include Time required to calculate that prediction, Relevance Factor (RF), Probability, and Knowledge Factor (KF).

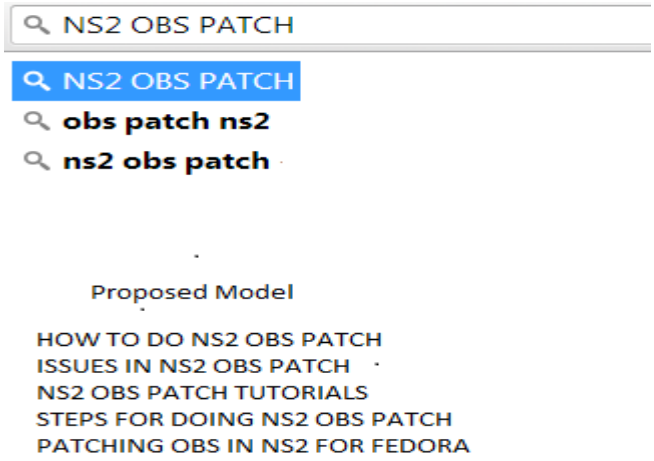

Figure 2: Predictions for the query" NS2 OBS PATCH" with relevance to previous user queries by Machine Learning

When concerned with the case of Session-based predictions, to predict the appropriate results, this model utilizes the queries users has previously typed during that particular session, separates keywords from the query, perform NLP to return appropriate and accurate predictions to the user before he/she types another query.

The time constraint on the NLP of the query is stringent and of high importance because producing late predictions but accurate ones is of little or no use to the user and furthermore, the Server, Database resources utilized for the computation are considered waste. Therefore, the challenge is to overcome the existing latencies in the network, other issues to provide optimized predictions to the user

A. PREDICTION AND LEARNING BASED ON TIME

This method is helpful for users whose probability of search for a specific domain at a particular time of a day is high. For example, an employee who checks the stock prices every day morning 7.30 to 8.30 AM or a sports enthusiast who checks for scores daily at 9.00 to 10.00 PM. The algorithm proceeds in the following order,

- In First step, this method makes use of the periodicity of the queries made by the user being pre-assigned the duration of the Time frame it checks daily,
- Then calculates the probability of the Query during the time frame every day, until the probability reaches a minimum value,
- Next creates a pattern of the queries in the decreasing order of their probabilities,
- Analyzes the pattern by comparing the pattern with the existing patterns, the system learnt from other users i.e. through Supervised Machine learning.
- Computes the common predictions and assigns relevancy factor to each prediction.
- Extracts the keywords that were not present in the history or cache from the observed pattern,
- Perform the NLP by using Thesaurus populated with different contextual meanings, Synonyms of words.
- Compute and provide meaningful, relevant predictions to the user during the specific time.
- Optimizes the predictions by using the relevancy factor

After the computation of the results, the resulting outputs with their respective relevance factors are associated with the observed pattern and are embedded into the system, to enable the system to learn from the data it produced to achieve further optimization of Predictions.

B. PREDICTION AND LEARNING BASED ON LOCATION

In modern days, every browser has an option to detect the location of the user in order to give the priority to results found in accordance with the country domains like "*.in"* for India, *".uk*" for United Kingdom etc. The needs of the user can be highly relying on his/her location, depending on the frequency with which it changes. The algorithm proceeds with the following steps.

- At the first step, the system maintains the location of the user.
- It keeps track of search queries containing words such as "tickets", "journey", flights", "time" when occurred in a combination with a location name other than the current location.
- If a change is noticed in the location of the user, without waiting for any prompt on the part of the user, the following steps are performed,
- The system analyzes the history, log data of several other similar users who queried similar search terms, to create a pattern of queries the users searched when their respective location has changed to the location in the present context.
- Computes the predictions based on the Information present about the users, and the keywords present in the history, by performing NLP on both of them and merging them to form relevant predictions.
- The system also assigns a relevant factor for each prediction that is generated eventually.

After the generation of predictions, the system is updated with the present pattern so as to make it more efficient when used for the next time.

C. NLP USING SEMANTIC WEB DATABASE

Even though NLP [6] [7] for the full search query seems plausible, the method will not be efficient, considering the Computational time, resources required to achieve that within the specified time frame, so, in order to generate relevant predictions within the minimum time frame, only the keywords of the appropriate search queries present in the history, cache, Logs and that particular Session are considered for NLP. A key word can be differentiated from the other search terms in that it is defined as an abstraction which is extrapolated from multiple search queries. Example is shown in the adjacent page.

For Example, consider the below Search queries

- What is Cristiano Ronaldo biography?
- How many goals Cristiano Ronaldo scored?
- What are Cristiano Ronaldo records?

In the above three cases, there exist more than one keyword for each query and they are called keyword phrases.

TABLE I. SEGMENTATION OF KEYWORDS

Search Query	Keywords or Keyword Phrases
What is Cristiano Ronaldo biography?	Cristiano Ronaldo, biography
How many goals Cristiano Ronaldo scored?	Cristiano Ronaldo, goals
What are Cristiano Ronaldo records?	Cristiano Ronaldo, records

The Segmentation of keywords is necessary for the Parts of Speech tagging and Tense conversion in the process of NLP. The above mentioned operations require a reliable, efficient Semantic web database, the role and functioning of which is described below.

D. SEMANTIC WEB DATABASE

The primary purpose of the Semantic Web Database [4] [9] is to infer the contextual meaning of the word is using the relations among the keywords or the Entire search query terms. In order for efficient functioning, huge amount of knowledge has to be populated in the thesaurus which is stored in the Database. A particular search query term which can be either a single word or a phrase can have more than one general meaning *("Polysemous"*),by using the relations in the database the meaning which is used in that particular context. For Example, consider the below query

Query: Bugs in Windows 8.1

The total possibilities of the meaning of the word "BUG" present in the thesaurus are given below. This example is given in order to show the importance of context for processing a query,

- An Insect or an Organism, small piece of material affixed to another, larger piece to conceal, reinforce, or repair a worn area, hole, or tear.
- A Defect in the design of a machine,
- A Defect in the coding of a program small piece, part, or section, especially that which differs from or contrasts with the whole
- To Bug means to annoy, pester.
- To equip the room with a concealed electronic device.

But in the above context the meaning of the word BUG is found out using the relationships and the knowledge embedded in the Semantic web database about "Bugs" and "Windows". The relationships processed for the above operation can be,

- Computer Science/ Operating Systems/Bug /Error
- Fauna/ Invertebrates/ Organisms/ Insects/ Bugs
- Electronic Devices/ Spying/ Concealment/ Bug

Out of the above three, only the first relationship is relevant to the query because Windows 8.1 is found under Operating Systems category, and hence only it is analyzed further to generate Synonyms. The Synonyms of the word "BUG" is also processed using the same method described above; the total synonyms possible are given below in a categorized manner

- Bug/ Glitch/ Error/ Flaw/ Defect.
- Bug/ Flea/ Defect.
- Bug/ Record/ Tap/ Overhear.

The NLP of the Search query or more specifically the keyword phrase of the Search query proceeds by selecting the appropriate combinations of the meaning and the available synonyms in the Semantic Web Database.

The NLP of the keyword phrases consists of 3 major tasks as elucidated below, with each of them executed iteratively,

E. PARTS OF SPEECH TAGGING AND TENSE CONVERSION

TABLE II. MAJOR PARTS OF SPEECH USED IN WEB SEARCH QUERIES [8]

Type	**Frequency**
Proper noun	40%
Noun	32%
Adjective	7%
Preposition	3.7%

The above table gives information about the most frequently used parts of speech in web search queries. Certain rules can be specified in the database which helps in the simplification of the identification and labeling of parts of speech in the search query

by the web server. This is done only for the keywords present in the query to reduce the computational time. Some of the rules which can be applied are given below

- A preposition precedes and succeeds only a noun, pronoun, verb and adverb. (90% of the cases)
- A Conjunction can never occur at the start of the sentence.
- A Verb always succeeds a Noun or a Pronoun or another Verb
- An Interjection and Conjunction can never occur consecutively in a query or a statement.
- *An Adjective precedes preposition, adverb, verb and a Noun.*
- *An Adjective can succeed every part of speech except another adjective.*
- A noun always succeeds and precedes a verb or a preposition or a conjunction

The list of rules given above only make a subset of the actual rules that are to be populated into the database for accurate labeling.

After the above operation, the tenses of the keywords are changed [2] [5] and all the possible meaningful combinations are checked in the thesaurus which contains the relationship among the keywords, thus resulting in the evaluation of only meaningful queries. This method is used to obtain the predictions in a tense different than the previous queries present in the history, log as well as session.

F. PREPOSITIONAL UNION

Since the whole preposition set is exhaustive, only certain prepositions which are used predominantly in most of the web based search queries are taken into consideration for NLP of the query. They are "For, to, of, in". Since the rules of grammar are already populated into the database, the meaningful combinations of the keywords and the prepositions mentioned above are evaluated in an iterative way.

This method of combinatorial processing even though sounds like a colossal task, the iterative method of approach simplifies it significantly. At every point the rule set of grammar is checked with the current combination of keywords and prepositions and a decision whether meaningful or not will be made. If a decision is computed to be not meaningful at any point during checking, then the all the further combinations of the keyword and queries need not be processed as they in turn will not be meaningful and hence useless as predictions to the users.

G. OPTIMIZING PREDICTIONS

After the NLP of the search query is finished, the resultant predictions are not displayed to the user as they are. The resultant predictions are further optimized by the following steps

- Remove duplicated Predictions if they exist,
- The "Relevance Factor" (RF) of each of the predictions is calculated and compared with the RF of previous query if only one previous query exists in the session or with the average of the Similarity Factors of the related queries present in the history, cache or session and a "Correlation Factor" (CF) is generated which indicates the affinity between the predictions and search queries of that session. It must be noted that Higher the CF, higher the Similarity and vice versa.

$$RF = \frac{\text{Number of similar or related keywords}}{\text{Total Number of keywords}}$$

- The Time required for computing NLP on a particular query is calculated for each prediction. This is denoted by $T_{N.}$ It varies with Relevance Factor (RF), the number of distinct and independent prepositions present in the query (N_P), Communication latencies (C) and much less frequently by backend database problems (D)

$$T_N \ \alpha \ \frac{RF}{L * N_P * C * D}$$

- The Probability of a Search query being made by the user at a specific time in a day is calculated by specifying the minimum periodic time frame and designated by "P_T".

$$P_T = \frac{\text{Number of queries in the time frame with one or more than one keyword phrase in common}}{\text{Total Number of Queries in that time frame}}$$

- The eventual probability of a prediction "P" will be the sum of Knowledge Factor (KF) and $P_{T.}$ i.e.

$P = P_T + KF$ (KF is newly learnt data from Database)

- The predictions are displayed to the user in the order of their priority.
  - If SF is high and T_C is high ,then priority is 1
  - If SF is high and T_C is low, then priority is 2.
  - If SF is low and T_C is low, then priority is 3.
  - If SF is low and T_C is high then priority is 4.

V. CONCLUSION

This model implements the Machine learning techniques and NLP of the search query to optimize the predictive search experience for the users, by providing relevant, apt predictions within a reasonable time frame. This model is aimed towards almost every category of users, may it be Researchers, Periodic viewers and with the existing machine learning techniques can further improvise the prediction relevancy. In particular, The Session based approach is helpful to researchers in whose case the probability of two consequent searches being related is high.

The proposed model aims to provide accurate predictions to the user based on his/her previous queries, by making use of the knowledge embedded in the database, by creating and analyzing patterns based on the existing information, then performing NLP on the previous queries and their synonyms in an iterative way, and finally, updating the System with the resulting patterns to improvise accuracy. The accuracy of the result is directly proportional to the completeness of knowledge in the database, where as the response rate of the system is directly proportional to the Network Communication speeds.

REFERENCES

[1] A.Geetha," A Note on NLP based Search Engines" in International Journal of Wisdom Based Computing, Vol. 1 (2), August 2011.

[2] McCallum, Andrew, Kamal Nigam, Jason Rennie, and Kristie Seymore. "A machine learning approach to building domain-specific search engines." In IJCAI, vol. 99, pp. 662-667. 1999.

[3] Blum, Avrim L., and Pat Langley. "Selection of relevant features and examples in machine learning." Artificial intelligence 97, no. 1 (1997): 245-271.

[4] "Information Retrieval and Semantic Web" in Proceedings of the 38th Hawaii International Conference on System Sciences – 2005.

[5] Joachims, Thorsten. Text categorization with support vector machines: Learning with many relevant features. Springer Berlin Heidelberg, 1998.

[6] Lawrence, Steve. "Context in web search." IEEE Data Eng. Bull. 23, no. 3 (2000): 25-32.

[7] P. F. Brown, P. V. deSouza, R. L. Mercer, V. J. D. Pietra, and J C. Lai." Class-based n-gram models of natural language. Computational Linguistics", 18(4):467–479, 1992a.

[8] Cory Barr, Rosie Jones and Moira Regelson,"The Linguistic Structure of English Web-Search Queries".

[9] Hendler, J., T.B-Lee and E.Miller. "Integrating Applications on the Semantic web". J. Institute Elec. Eng. Japan, 2002. 122: 676-680.

[10] R. K. Ando and T. Zhang. "A framework for learning predictive structures from multiple tasks and unlabeled data". Journal of Machine Learning Research (JMLR), 6:1817–1953, 2005.

AUTHORS

First Author – Kilari Murali Krishna Teja, krishnateja@pec.edu.