Innovative Predictive Search using Real Time Session based NLP of Search Query

K.Murali Krishna Teja
B.Tech, Final year, Department of Computer Science
Pondicherry Engineering College

ABSTRACT
Information Retrieval (IR) has come a long way in the recent years with giant strides of research and development in the field. One of the Recent (from 1994), widely renown and hugely used application of IR is a Search Engine. This paper focuses on improvisation of existing predictive search mechanism by using a real time session based NLP of a search query thus resulting in a more diversified but related suggestions being provided to the user. The main intention is to provide appropriate, accurate suggestions to dedicated users of web search engines such as researchers who mostly concentrate on a particular topic to search in a session providing an optimal balance between response rate and accuracy.

General Terms
Real-time session based search query processing model, Information retrieval (IR).

Keywords
Session based query processing, Natural Language Processing(NLP),Intelligent Information Retrieval(IIR), Prediction Optimization, Semantic database, Session based Query processing.

1. INTRODUCTION
Since the evolution of Internet from ARPANET which was a defense project of the US government, WORLD WIDE WEB (WWW) played and continues to play a vital role as one of its fundamental and most important feature. Searching for information on the internet has been a way of life for researchers, students etc., due to the availability of vast amount of information, ease of use and various other factors. Even though trivial Search mechanisms have been existent from long past, Search engines which implement Information Retrieval(IR) in a feasible and structured way came into existence from 1994 and have undergone several frequent but noticeable changes ever since. The implementation of predictive search can be considered one of the major over hauls in the history of web search engines. .

Natural Language Processing (NLP) [1] and Machine Learning (ML) [10] which are the concepts of Artificial Intelligence can be mapped with the Information Retrieval (IR) [4]concepts of search engines to provide appropriate as well as accurate predictions to the user of the search engine by performing keyword segmentation and grouping, Identifying Parts of speech in a sentence, Use of semantic web database to provide context based related suggestions using existing spelling correction mechanisms.

1.1 Natural Language Processing (NLP):
Natural language processing (**NLP**) is a "*field of computer science, artificial intelligence, and linguistics concerned with the interactions between computers and human (natural) languages"*. As such, NLP is related to the area of human–computer interaction. Many challenges in NLP involve natural language understanding -- that is, enabling computers to derive meaning from human or natural language input.

1.2 Machine Learning:
Machine learning is a "*type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed"*. Machine learning focuses on the development of computer programs that can teach themselves to grow and change when exposed to new data.

The process of machine learning is similar to that of data mining. Both systems search through data to look for patterns. However, instead of extracting data for human comprehension as is the case of data mining applications, machine learning uses that data to improve the program's own understanding

2. SESSION BASED SEARCH QUERY PROCESSING
This paper presents a novel approach to provide suggestions to the user by making use of his/her previous searches in the session. This method is mainly intended towards researchers because the efficiency of the mechanism increases when the cohesiveness of the queries being searched increases. Any researcher looking for information in the internet will be focusing on a specific point or a domain rather than generic terms. The objective of this paper is to propose a model which makes use of the users previous search queries to perform NLP on them and with the help of semantic database containing the dictionary, synonyms, antonyms and relations among the words used in the query. This can be incorporated using 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 box again.

This method is highly dependent on the compactness, correctness of the backend database and the performance of the web servers and the network communication speeds between the user and Search engine. With the advent of high performance web servers which can execute distributed and parallel computations, the performance factors of the existing web servers for are highly sufficient for this type of query processing. The network communication speeds which are highly required for an optimal search experience to the user are painfully dependent on the local ISP(Internet Service Provider)*.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*. This

session based search query processing model 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.

This model realistically assumes that the user will spend some amount of time to read a web page related to his first or basic query and this model will take some amount of time to perform NLP and to return appropriate suggestions or predictions for the user when he/she tries to query for a second time. The advantage of this model is that these computations occur in background during the time the user is reading his intended webpage out of a list of webpages returned as a result to his first query by the search engine.

2.1 Division of Web Search Queries

The web search queries can be broadly divided into the following 3 categories.

1. Navigational Search queries
2. Informational Search queries
3. Transaction Search queries

2.1.1 Navigational Search Queries

A navigational query is a search query entered with the intent of finding a particular website or webpage. For example, a user might enter *"YouTube*" into Google's search bar to find the YouTube site rather than entering the URL into a browser's navigation bar or using a bookmark. In fact, *"Facebook*" and *"YouTube"* are the top two searches on Google, and these are both navigational queries.

2.1.2 Informational Search Queries

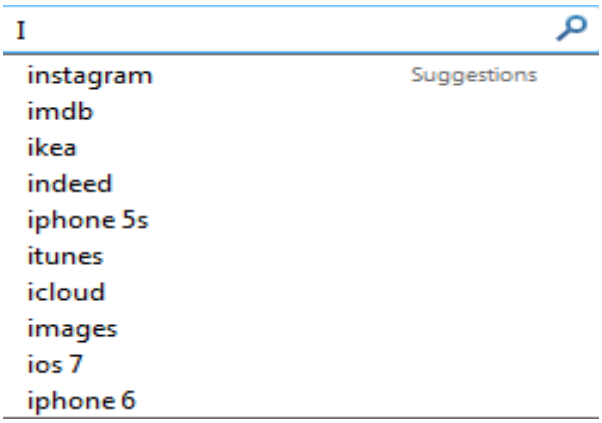
Wikipedia defines informational search queries as "Queries that cover a broad topic (e.g. *Chennai* or *truck*s) for which there may be thousands of relevant results." When someone enters an informational search query into Google or another search engine, they're looking for information – hence the name. They are probably not looking for a specific site, as in a navigational query, and they are not looking to make a commercial transaction. They just want to answer a question or learn how to do something.

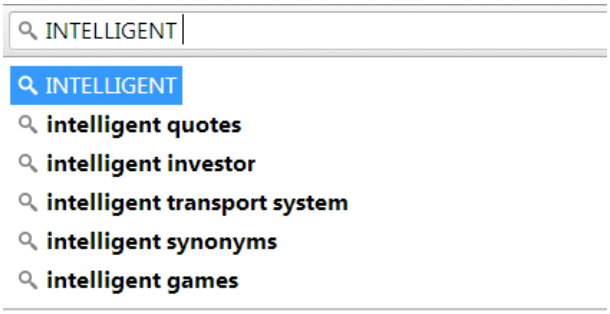
2.1.3 Transaction Search Queries

A transactional search query is a query that indicates intent to complete a transaction, such as making a purchase. Transactional search queries may include exact brand and product names (like *"iPhone 5s")* or be generic (like "*iced coffee maker*") or actually include terms like "buy," "purchase," or "order." In all of these examples, you can infer that the searcher is considering making a purchase in the near future, if they're not already pulling out their credit card. In other words, they're at the business end of the conversion funnel. Many local searches (such as *"pizza hut*") are transactional as well.

3. EXISTING MODEL [6]

The existing model for prediction based search is shown below. As seen below the predictions or suggestions that are offered to the user are provided by using the information obtained from millions of users and are not session specific. These suggestions are simply a reflection of the most frequently searched terms starting with the predictions initial letter and progressing iteratively to the predictions of words and sentences.





3.1 Disadvantages of Existing Model

Though the predictions are accurate, 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. In the above context the user may be researching about Information Retrieval (IR) for a significant amount of time but the next time he tries to search about IR he/she is provided with the same suggestions that are provided for his first query related to Information Retrieval (IR) thus wasting valuable time for waiting the user to type the required specific query.

4. PROPOSED MODEL

The proposed model overcome the above disadvantage and also includes additional advantages like providing suggestions with words not present in the earlier queries rather than various meaningful forms of the words present in the previous query. These additional words are commonly found in the web search queries some of them are *"How To"," Steps for"* etc. These words not only include the above two but are also extracted from the most frequently shown suggestions to provide a mix of suggestions which are both specific and generic to the users previous query but appropriate and related to his future ones.

This model receives the queries that the 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 is of high importance as predictions which are accurate but late are less than useless because late predictions are of no use to the user and the server and database resources are wasted for NLP in the background.

Let T_N be the time required for the NLP of a search query, and T_A be the average time taken by the user to type another query after typing the previous one. Since this model is predominantly focused on researchers, T_A will not be

negligible. T_N on the other hand is dependent on various factors like the Similarity Factor (SF) of the keywords. The SF is calculated by using the following formula

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

The Number of distinct keywords can be found out by using semantic database which contains the synonyms, antonyms of a word as well as relations among the words.

T_N is also dependent on the length of the search query (L).If the length of the search query is more, it indicates that the query contains more terms thus taking more amount of time for the completion of NLP.

T_N also varies with the number of distinct and independent prepositions present in the query(N_p).In addition to the above factors, T_N is also effected by Communication latencies(C) and much less frequently by backend database problems(D).By taking all the above factors into consideration, T_N can be expressed as

$$T_N \alpha \frac{SF}{L * N_p * C * D}$$

By ignoring insignificant factors such as Communication latencies(C) and Database problems (D), the above expression of T_N can be refined as

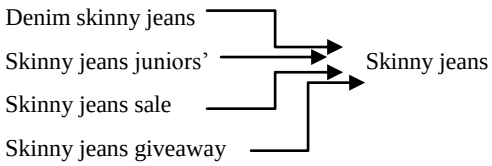
$$T_N \alpha \frac{SF}{L * N_p}$$

4.1 NLP of the Search Query

Even though NLP is required of the full search query, in order to generate additional relevant predictions only the keywords in the search query are taken into consideration to perform 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.

SEARCH QUERIES KEYWORDS



4.1.1 Major tasks in NLP

The NLP of the search query or keywords of the proposed model proceeds in 4 steps each in their order of listing in an iterative way

4.1.1.1 Parts of Speech labeling [5] [7] [8]

Table 1. Major parts of speech used in web search queries

Type	Frequency
Proper noun	40%
Noun	32%
Adjective	7%
Preposition	3.7%
Remaining	17.3%

The above table shows the parts of speech which are used majorly 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

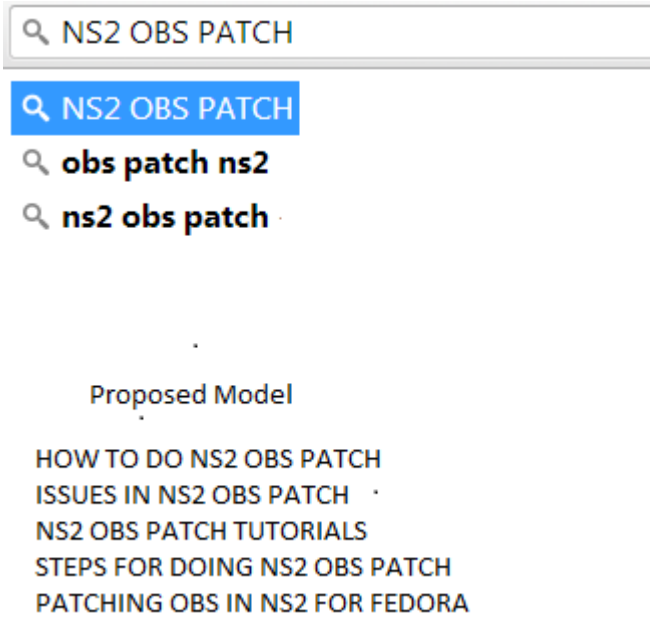
1. A Verb always succeeds a Noun or a Pronoun or another Verb

2. A Conjunction can never occur at the start of the sentence.

3. A preposition precedes and succeeds only a noun, pronoun, verb and adverb. (90% of the cases)

4. A noun always succeeds and precedes a verb or a preposition or a conjunction

Fig 1: Working of Existing vs. proposed model for the query *"NS2 OBS PATCH"*



5. An Adjective can succeed every parts of speech except another adjective.

6. An Adjective precedes preposition, adverb, verb and a Noun.

7. An Interjection and Conjunction can never occur consecutively in a query or a statement.

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. *"In order to achieve the above the dictionary and*

thesaurus implemented in the database have to support automatic text analysis and AI applications".

4.1.1.2 Tense Conversion [2][5]

The existing tenses of the keywords are changed and all the combinations of the keywords are checked in the thesaurus and dictionary which contain the relationship among the keywords thus evaluating and processing only meaningful queries. This method is useful to return the predictions in a tense different than the original query.

4.1.1.3 Semantic based word prediction [3][6]

The In this task, the contextual meaning of the word is inferred using the relations among the keywords or the Entire search query terms. This method requires huge amount of knowledge to be populated in the thesaurus which is to be 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: NS2 OBS PATCH

The word "*PATCH"* can have several meanings such as,

1. A small piece of material affixed to another, larger piece to conceal, reinforce, or repair a worn area, hole, or tear.
2. A small cloth badge affixed to a garment as a decoration or an insignia, as of a military unit.
3. A small piece, part, or section, especially that which differs from or contrasts with the whole
4. A piece of code added to software in order to fix a bug, especially as a temporary correction between two releases.
5. An indefinite period of time; a spell.

But in the above context the meaning of the word PATCH is evaluated using the knowledge about NS2 & OBS. The relationships can be

- computer science/ns2/optical/patches
- time/indefinite period
- clothing/decoration/patch
- clothing/repair/patch

Out of the above only the First relationship is appropriate to the query and hence only that is processed. The synonyms of the word PATCH are also processed using the same method described above, thus resulting in a diversified but related predictions

4.1.1.4 Prepositional Merging and combinatorial processing

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 decided 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.

4.1.1.5 Unification of keywords and Specific search terms

This method implements the merging of keywords, prepositions, SSTs *(Specific Search Terms*) which are derived by careful analysis of the terms which are used most frequently by the researchers and which are independent of the domain. Some of them are *"How", Steps", "issues", "and solutions"* and their synonyms .This list of SSTs given above are just a subset of the original SSTs which can be derived using logs of search engines by analyzing thousands of queries. These SSTs are unified with the combinations of keywords and prepositions obtained in the previous step to draw meaningful conclusions..

5. 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

1. Removal of duplicate suggestions if they exist,

2. The *Similarity Factor(SF*) of the predictions is compared with the SF of previous query if only one previous query exists in the session or with the average of the Similarity Factors of the queries in the session and a *Correlation Factor* (CF) is generated which indicates the affinity between the predictions and search queries of that session. Higher the CF, higher the Similarity and vice versa.

3. The Time required to generate a particular prediction is also calculated for each prediction. This is denoted by $T_{C.}$

4. 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.

The predictions are displayed in the order of increasing priority to the user .Predictions of priority 4 are neglected as they are no way near accurate and they utilize huge amount of resources for NLP. But, the implementation of Iterative approach will reduce the probability of predictions with a priority of 4 to minimum.

6. CONCLUSION

This model implements the session based NLP of the search query to produce the relevant and accurate results to the users within a minimum time frame. This model is aimed at the researchers' fraternity where the possibility and probability of two consequent searches being related is high. The time taken for NLP is also dependent on various factors but most importantly on the Network communication speeds and the processing power of the web servers.

This model aims to provide accurate predictions to the user based on his/her previous queries in the session by making use of the knowledge embedded in the database and by performing NLP on the previous queries and their synonyms in an iterative way. The accuracy of the result is directly proportional to the completeness of knowledge in the database.

7. REFERENCES

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

[2] Zhong, N., J.Liu and Y.Yao, 2002. "In search of the Wisdom web". IEEE Computer, 35: 27-31.

[3] P.C.Reghu Raj and S.Raman. "Applied Artificial Intelligence", in Taylor and Francis Inc. 19:559-599 2005.

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

[5] M.Mitra. B.B.Chaudhuri., "Multilingual IR" in Information Retrieval2, 141-163(2000).

[6] H.Chu., M. Rosenthal., "Search engines for WWW: A comparative study and evaluation methodology" 2007.

[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.