Machine Learning in Search Engines

[1]Neenu Ann Sunny
[1]Assistant Teacher
[1]Saintgits College of Applied Sciences

***Abstract* - The relevance of a web page is an innately biased matter and based on readers knowledge, interests and attitudes, web page is depended. To say justly about the relative importance of web pages, there is still much. One factor which makes it difficult for search engines to give relevant results to the users within a stipulated time is the explosive growth of internet. Classified directories are used by search engines for storing the webpages and for this process, some search engines even depend on human expertise. Automated methods are used by most of the web pages for classification of web pages. We can infer from experimental results that machine learning techniques for automated classification of the web pages proves to be the best and more relevant method for search engines.**

***keywords* - for search engines. Keywords— Search Engines, expertise, machine learning, web pages, automated.**

I. INTRODUCTION

Search engines are used for searching webpages with accurate searched results in microseconds of time. Finding required information on web was unfeasible before search engines were introduced. We can say that, a search engine is a software program that searches for sites based on the words that users designate as search terms. Search Engine Optimization plays a great role in the area of internet and search engines. The successful company to launch search engine is Google which had made searching web pages in simple and precise way. Nowadays, most of the search engines use machine learning techniques for the automated classification of web pages as well as for the web page ranking. Machine learning can be applied in various fields or areas related to search engines and web page ranking.

II. RELATED WORKS

Webpage ranking algorithm, a well known approach to rank the web pages available on cyber world. It helps us to know how the search engine exactly works and how a machine learn itself while giving priority to the page that which page is important to successfully fulfills the user query need and which page is worth less. Machine Learning approach also helps us in understanding the complex part of page priority criteria in most popular search engines like Google, Yahoo, AltaVista, Dog pile and many more search engines like that. Page ranking mainly unrevealed the structure of web[1].

The world wide web has immense resources for all kind of people for their specific needs. Searching on the web using search engines such as Google, Bing, Ask have become an extremely common way of locating information. Searches are factorized by using either term or keyword sequentially or through short sentences. The challenge for the user is to come up with a set of search terms/keywords/sentence which is neither too larger nor too small to get the desired result. No matter, how the user specifies the search query, the results retrieved, organized and presented by the search engines are in terms of millions of linked pages of which many of them might not be useful to the user fully. In fact, the end user never knows that which pages are exactly matching the query and which are not, till one check the pages individually. This task is quite tedious and a kind of drudgery. This is because of lack of refinement and any meaningful classification of search result. Providing the accurate and precise result to the end users has become Holy Grail for many search engines like Google, Bing, Ask, etc. There are number of implementations arrived on web in order to provide better result to the users in the form of many search engines like Yippy, Dog pile, etc. This paper proposes development of a meta search engine called SEReleC[2] that will provide an interface for refining and classifying the search engines so as to narrow down the search results in a sequentially linked manner resulting in a huge reduction of number of pages[2].

Since the use of internet has incredibly increased, it becomes an important source of knowledge about anything for everyone. Therefore, the role of search engine as an effective approach to find information is critical for internet's users. The study of search engine users behaviour has attracted considerable research attention. These studies are helpful in developing more effective search engine and are useful in three points of view: for users at the personal level, and for government and marketing at social society level. These kinds of studies can be done through analysing the log file of search engine where in the interactions between search engine and the users are captured[3].

Meta search engine is an effective tool for searching information online. In comparison with independent search engine like Google, Bing, and etc., meta search engine has a wider coverage and can meet the requirements of information retrieval in a better manner. In particular, when a query is received from the user, the meta search engine sends it to some proper candidate member engines, collects results from them, and then replies to the user. An important issue here is how to better select the underlying member engines, collects results from them, and then replies to the user[4].

III. MACHINE LEARNING IN SEARCH ENGINES

A. Introduction to Search Engines

Search Engine[1] is a service that allows internet users to search for contents via the world wide web(www). Search engine is a software program that searches for sites based on the words that you designate as search terms. They look through their own

databases of information inorder to find what it is that you are looking for. There are mainly three components for search engines. They are:

- web crawler
- database
- search interfaces

Web crawlers are also known as spiders or bots. It is a software component that traverses the web to gather information. All the information on the web are stored in a database. It consists of huge web resources. Search interface acts as an interface between user and the database. It helps the users to search through the database.

B Classification of Search Engines

There are many search engines on web based on the usage and features, users can use them. Every search engine has many web pages stored on their database but search engines with large number of pages on web are not top search engines. Search engines which will provide accurate information based on requested keyword will be the top search engines. Search engines are classified as follows:

- Crawler based search engines[3]
- Human powered directories
- Meta search engines[4]
- Hybrid search engines
- Speciality search engines

B.1 Crawler based Search Engines

Crawler based search engines[3] such as Google create their listings automatically. They crawl or spider the web, then people search through what they have found. If you change your webpages, crawler based search engine will find these changes and that can affect how you are listed. Three elements in crawler based search engines are:

- Crawler or spider
- index or catalog
- search engine software

Crawler or spider visits webpages and reads it and index or catalog is like a giant book containing a copy of every webpage that crawler or spider finds. If a webpage changes, then this book is updated with a new one.

B.2 Human Powered Directories

A human powered directory such as the open directory depends on humans for its listings. In this type of search engine, site owner submits a short description of the site to the directory along with category it is to be listed. Submitted site is then manually reviewed and added in the appropriate category or rejected for listing. Keywords entered in a search box will be matched with the description of the sites. This means the changes made to the content of web pages are not taken into consideration as it is only the description that matters. A good site with good content is more likely to be reviewed for free compared to a site with poor content.

B.3 Meta Search Engines

Meta search engines[4] gives results based on a combination of results from other search engine databases. It uses complex algorithms and virtual databases. A search engine that queries other search engines and then combines the results that are received from all. In effect, the user is not using just one search engine but a combination of many search engines at once to optimize web searching. For example, Dog pile is a meta search engine.

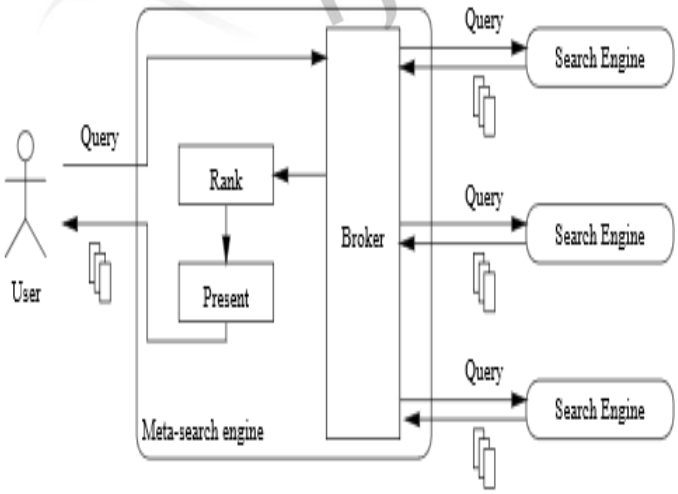

Figure B.3: Meta Search Engine

B.4 Hybrid Search Engines

Hybrid search engines present either crawler based results or human powered listings. Nowadays, it uses combination of both results. Most of the crawler based search engines like Google basically uses crawlers as a primary mechanism and human powered

directories as secondary mechanism. For example, google may take the description of a webpage from human powered directories and show in the search results. As human powered directories are disappearing, hybrid types are becoming more and more crawler based search engines. But still there are manual filtering of search result happens to remove the copied and spammy sites. When a site is being identified for spammy activities, the website owner needs to take corrective action and resubmit the site to search engines. The experts do manual review of the submitted site before including it again in the search results. In this manner though the crawlers control the processes, the control is manual to monitor and show the search results naturally.

B.5 Speciality Search Engines

Speciality search engines search a specially created database which is limited to a particular subject. A speciality search engine, sometimes called a topican or vertical search engine, searches a specially-created database limited to a particular subject. Speciality search engines fall into two main categories:

- service
- subject-specific

Speciality service search engines provide services that are often not available from larger general search engines. Subject-specific search engines search a database tailored to a particular subject. Depending on your area of interest and the type of information you are seeking, speciality search engines can provide more relevant results more quickly than a general purpose search engine such as Google or Yahoo. Speciality search engines are also an excellent source for typical research. Because of this it would be wise to also submit your blog or website to some of the speciality search engines that cater for your niche.

C Search Engine Working

While you should always create website content geared to your customers rather than search engines, it is important to understand how a search engine works. Most search engines build an index based on crawling, which is the process through which engines like Google, Yahoo and others find new pages to index. Mechanisms known as bots or spiders crawl the web looking for new pages. The bots typically start with a list of website. URL's determined from previous crawls. When they detects new links on these pages, through tags like HREF and SRC, they add theses to the list of sites to index. Then, search engine use their algorithms to provide you with a ranked list from their index of what pages you should be most interested in based on the search terms you used. Then, the engine will return a list of web results ranked using its specific algorithm. On Google, other elements like personalized and universal results may also change your page ranking. In personalized results, the search engine utilizes additional information it knows about the user to return results that are directly catered to their interests. Universal search results combine video, images and Google news to create a bigger picture result, which can mean greater competition from other websites for the same keywords.

Search engine optimization is a set of rules that can be followed by website owners to optimize their websites for search engines and thus improve their search engine ranking. In addition, it is a great way to increase the quality of your website by making it userfriendly, faster and easier to navigate. Steps in search engine optimization are as follows:

- Website analysis
- Client requirements
- Keyword research
- Content writing
- Website optimization
- SEO submission
- Link building
- Reporting

D Introduction to Machine Learning

Machine Learning is a branch of artificial intelligence(AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Some applications of machine learning are: email spam and filtering, online fraud detection, product recommendations. There are mainly three types of learning. They are as follows:

- supervised learning
- unsupervised learning
- reinforcement learning

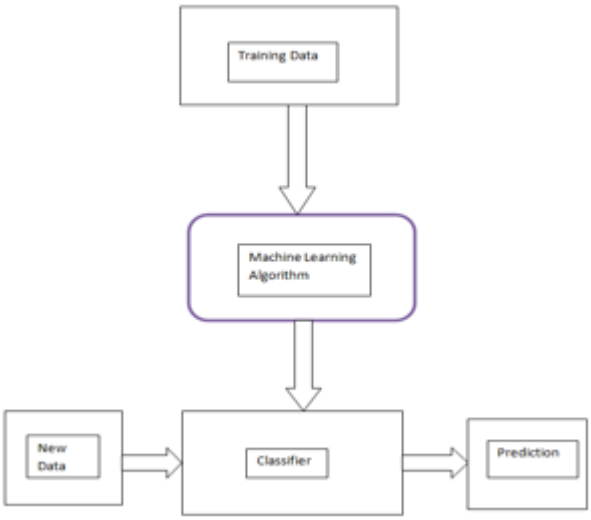

Figure D: Machine Learning

E Applications of Machine Learning in Search Engines

Machine Learning can be applied in various areas realted to search engines. They are:

- Pattern Detection
- Identifying new signals
- Custom signals based on specific query
- Image search to understand photos
- Identifying similarities between words in a search query
- Improve ad quality
- Query understanding
- URL/Document understanding
- Search features[2]
- Crawling[1]
- User classification
- Search Ranking
- Synonyms Identification/Query Expansion
- Intent Disambiguation

Image search to understand photos: Users can upload photos on google image search and get information about the image, similar looking images, etc. There is a lot of data in the form of images on the internet. Hence, search engines can use machine learning on the huge number of images and power their feature of image searching.

Query Understanding: Machine Learning is used for understanding the search queries typed by the users. Query classification is one of the problem that is solved by using machine learning. Search engines run different classifiers on the search query. They are:

- Navigational search queries
- Informational search queries
- Transactional search queries

URL/Document Understanding: This includes everything that is done to understand a URL(Uniform Resource Locator). For example, spam detection, page classification, etc.

Intent Disambiguation: Consider an example, when you search for eagles, is it eagles the band or Philadelphia eagles or the bird or all of them together. Machine Learning is applied in these types of scenarios.

Improve Ad quality: A lot of revenue of the search engines comes from the advertisements that they display on their websites. Suggesting the advertisements that are relevant to the users increases the chance of the user actually purchasing the advertised product or service, which would, in turn be beneficial for the company providing the product or service and the search engine that advertised the product. Both will take money. Machine Learning is used by the search engines to identify the correct target audience for the showing of the various advertisements. Depending on the queries the user asks on the search engine, relevant advertisements are shown to him/her.

Identifying the meaning of words based on their usage: The number of words in the English language is constantly growing. Around ten years back, no one knew words and phrases like “selfie”, “it’s lit”, “muggle”, etc. When a word or phrase is very new

and is not used widely yet, the search engines may not be able to give its exact meaning. However, as more an more people start using it and it is written at several places on the internet, the machine learning algorithms of the search engines will gather the data and try to decipher the meaning of the word or phrase. Overtime, the search engines are able to exactly understand it.

Pattern Detection: Search engines are using Machine Learning for pattern detections that help identify spam or duplicate content. Eventhough there are still human quality raters, Machine Learning has helped Google automatically to move through web pages to weed out low quality pages without an actual having to look at it first. Machine Learning is an unevolving technology, so the more pages that are analysed, the more accurate it is.

Eliminating spam and low quality contents from search results: Machine Learning is used by the search engines to identify the spam, duplicate, or low quality content. Some common attributes of such content are the presence of several outbound links that actually link to unrelated pages, usage of stop words and synonyms in abundance, etc. Search engines try to weed out such contents from their search results so as to provide more relevant contents to the users, thereby increasing the user experience. Machine Learning has drastically reduced the human effort required to identify the low quality content. Although there are human quality raters still, the human involvement, overall, has reduced tremendously.

User classification: User Classification means figuring out what kind of a user you are. This is especially useful for personalised search.

Search Features: Machine Learning is used for generating search features like site links, related searches, knowledge graph data, etc.

Understanding User Queries: Whenever you write your question in a search engine, for example, Google, Bing, etc., the most important thing for the search engine becomes to understand what you are trying to ask. If a search engine is not able to understand what you are trying to ask or if a search engine is not able to understand your query well, it will not be able to give you appropriate answers, which would make the search engine useless for you. This is where machine learning comes into the picture. Users can make spelling errors while typing their queries in the search engines. We cannot assume a user to write all the spellings correctly. In fact, many people use search engines to verify their spellings. If you write a wrong spelling in a search engine, it shows you the correct spelling of the same word. The search engine would be smart enough to identify the word that you are typing to write, even if you have made some spelling mistake. Hence, Machine Learning is used for spelling correction in the search engines.

Synonyms Identification: If you use synonyms, even the most rarely used ones, the good search engines are able to answer your queries appropriately. The search engines are able to identify what you want to search. Machine Learning is used here too. Sometimes, users may ask queries which are a bit ambiguous. For example, suppose a user types the query "The Indian Ocean". Now, the query can imply the actual Indian Ocean band. A good search engine should be able to identify the ambiguity and work a way around it. Use of machine learning is here as well. A search engine can also classify a query into one of the various categories, for example, whether a query is navigational or transactional or information or belongs to any other category. Machine Learning is used for identification of the category of the queries. Depending on the category of the queries, the search engine may give the appropriate additional information. For example, if you search for "Westminster Abbey", you will get its information and also a Google Maps location for the same.

F Search Ranking[1]

Ranking refers to where a website or page is ranked within search engine results. A webpage rank within a search engine is commonly called as a position. Search engine ranking is the position at which a particular site appears in the results of a search engine query. Each page of the search results typically lists 10 websites, although they are sometimes augmented with local listings, videos and images. Ranking higher in the search results actually corresponds to a lower number, while ranking lower corresponds to a higher number. Many site owners engage in SEO campaigns in order to improve their search engine ranking and move their website closer to the top of the results because websites that are ranked higher typically get a larger percentage of click-throughs and attract more visitors than lower ranked websites. Search engine ranking is influenced by a multitude of factors including age of site, the quality of a site's link portfolio, relevancy of the page, social signals and level of competition, among others. Search engines rank individual pages of a website, not the entire site. This means that the homepage might rank #1 for certain keywords, while a deep internal page might be listed on the third page.

G Google RankBrain

RankBrain is a machine learning algorithm that Google uses to sort the search results, It also helps Google process and understand search queries. Google recently announced that RankBrain is Google's third most important ranking signals and it is becoming more important every day. Before RankBrain, Google would scan pages to see if they contained the exact keywords someone searched for. Today, RankBrain understands what user's are asking and it provides hundred percent accurate set of results. RankBrain tries to actually figure out what users like a human would. For instance, Google may have noticed that lots of people who search for "grey console developed by nintendo" and they have learned that people who search for "grey console developed by Nintendo" want to see a set of results about gaming consoles. So when someone searches for that search query, RankBrain brings up similar results to the keyword.

If RankBrain sees a word or phrase it isn't familiar with, the machine can make a guess as to what words or phrases might have a similar meaning and filter the result accordingly, making it more effective at handling never-before-seen search queries or

keywords. Search queries are sorted into word vectors, also known as "distributed representations," which are close to each other in terms of linguistic similarity. RankBrain attempts to map this query into words(entities) or clusters of words that have the best chance of matching it. Therefore, RankBrain attempts to guess what people mean and records the results, which adapts the results to provide better user satisfaction.

H Working of Google RankBrain

Google RankBrain goes beyond simple keyword matching. It turns your search term into concepts and tries to find pages that cover that concepts. RankBrain performs two main jobs. They are:

- Understanding search queries.
- Measuring how people interact with the results.

As the new age,google search algorithm RankBrain has induced massive quality improvement to SERP as it looks beyond mere keyword matching. This is done by looking at the big picture in the search by transforming searching terms into concepts than scouting for pages covering that concept. The rankbrain also tracks user satisfaction by understanding new keywords thanks to the ability to tweak the algorithm on its own. In other words, RankBrain as the google algorithm for search engine optimization can show search results that will get maximum liking by the users. In the results, pages liked more by users for the information can expect a better google search ranking as a fall out of the RankBrain perspective.

RankBrain takes a serious look at how a user interacts with the search results. It keeps a tab on the pogo sticking effect where unsatisfied users hit the back button out of sheer frustration at the search results. If a web page has people leaving it quickly the message to Google is the page stinks and needs remedial action. Google RankBrain will diminish pogo-sticking on a specific result and a popular page will be made easier to find. RankBrain spams websites that carry too many different topics and is without a focal area because it can't understand who will use such multiple and diverse content.

The methods Google already uses to refine queries generally all flow back to some human being somewhere doing work, either having created stemming lists or synonym lists or making database connections between things. Sure, there's some automation involved. But largely, it depends on human work. The problem is that Google processes three billion searches per day. RankBrain is designed to help better interpret those queries and effectively transalate them, behind the scenes in a way, to find the best pages for the searcher. RankBrain is paying very close attention to how users interact with the search results.

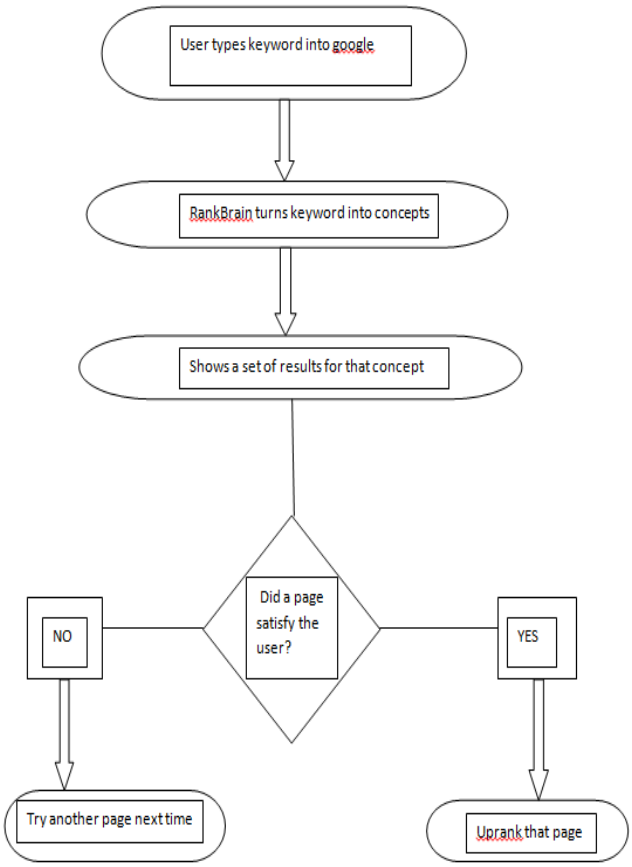

Figure H: Working of Google Rank Brain

RankBrain considers the following UX (User Experience) Signals:

- Dwell time
- Pogo sticking
- Bounce rate
- Organic click-through rate

Three essential components in the RankBrain environment are:

- Different rankings signals apply to different queries.
- Signals apply to your website's reputation.
- One keyword-one page is really,really dead.

IV. CONCLUSION

Web page ranking is a global ranking of all web pages, regardless of their content, based solely on their location in the web's graph structure. Using these web page ranking techniques, we are able to order search results so that more important and central web pages are given preference. All things considered, search engine optimization will become more resourceful in the upcoming years, but also more complex, forcing marketers to develop more elaborate strategies that bring more types of content, devices and tools into the equation. But no matter which combination of elements you see, the focus should stay on the user and their needs, as machine learning and artificial intelligence technologies will transform ranking factors that can better reflect the needs and expectations of searchers.

REFERENCES

[1] Vishwas Ravall and Padam Kumar, "SEReleC (Search Engine Result Refinement and Classification) – A meta search engine based on combinatorial search and search keyword based link classification," in IEEE-International Conference on advances in Engineering, science and management(ICAESM-2012), March 30,31,2012.Saad ALBAWI, Tareq Abed MOHAMMED, Saad AL-ZAWI, "Understanding of a convolutional neural network," ICET 2017.

[2] Vijay Chauhan, Arunima Jaiswal, Junaid Khalid khan, "Web page ranking using machine learning approach," in *Fifth International Conference on Advanced Computing and Communication Technologies*, 2015.T. Yamunarani, G.Kanimozhi, "Hand gesture recognition system for disabled people using arduino," vol. 4, 2018.

[3] Farzaneh Shoeleh, Mohammad Sadegh Zahedi, Moigan Farhoodi, "Search Engine Pictures: Empirical analysis of a web search engine query log," in Third International Conference on Web Research(ICWR), 19,20 April 2017.

[4] Donghong Liu, Xan Xu, Yu Long, "On member search engine selection using artificial neural network(ann) in meta search engine," in 2017..