An Approach to Increase the Accuracy of Aboutness of Web Documents for Search Engines using OGP

Amit Soni, Deepak Gupta, Ajay Dadhich

Abstract - This paper describes the potential use of the Open Graph Protocol (OGP) developed by the Facebook, to increase the accuracy of “Aboutness” of web documents available on World Wide Web, so that search engines can display highly accurate searches to end-users. This paper emphasis about the subjective analysis, the term “Aboutness”, information organization by search engines on the web. This paper concludes in-depth understanding of the OGP and how it can help search engines to give optimized search results. Moreover OGP can help search engines to do better sentimental analysis rather than automated sentimental analysis and also helps to index in a better way. In the end it is argued that how social networking can help understand the author’s views towards the article (web document) and the conclusion.

Keywords: Aboutness, crowd sourcing, open graph protocol, subject analysis.

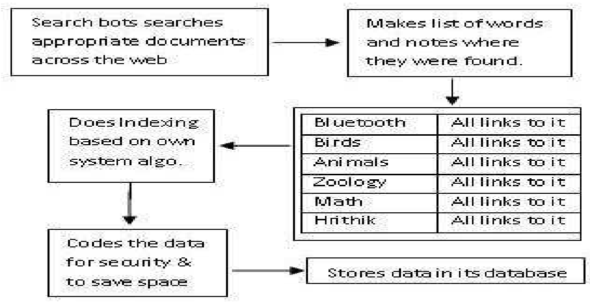
1 Search Engine Working

On the internet, millions of web pages are available which presents information on any topic. When we need to know about any particular topic, we search on the search engines and how they know which page to display? It is the mechanism which they previously follow and stores data in their database. There are different ways of various search engines but all perform three basic tasks:-

- The search about the topic in advance on the Internet and stores results in their databases.
- They keep index of words they find and from where they find them.
- When request comes from users, they match request words with their index words or a combination of these.

To find where the information is, search engines share their special software bots which are called web spiders which builds the list of words found across millions of web pages. This process is called Web Crawling. How any software bots does starts it’s searching? The usual starting points are list of heavily used servers and very popular pages. The bots will begin with a very popular websites, then indexes the words on its pages and follows every link that is available within the site. Now how does the search engine gather information about the web page? These are actually the Meta tags which allow the owner of a page to specify key words and the concept behind writing the page so that search engine indexes it under particular category. This is very helpful for engines especially in cases in which the words of the page might have ambiguous/double meanings, meta tags can guide further to get the meanings of these words because sometimes authors unknowingly or intentionally confuses the bots about the aboutness of pages. So this over dependency is not good. Hence, bots correlates Meta tags with page contents but this too is not enough which produces wrong results. So search engines developed automated sentimental analysis software bots which gives the most accurate results.

Building the index: - Once search engine bots have completed the task of finding relevant information across web (we should note that this task never actually completes the constantly changing nature of web means that they should work always). Search engines stores it in a way that when there is a need to search it would be very convenient to access it from its database. There are two ways involved in getting data accessible to users firstly, the information associated with the data and secondly, the method by which information is stored in database. This is the simplest way of storing information in the databases of search engines.


Search bots searches appropriate documents across the web
Makes list of words and notes where they were found.
Does Indexing based on own system algo.
Codes the data for security & to save space
Stores data in its database

Bluetooth	All links to it
Birds	All links to it
Animals	All links to it
Zoology	All links to it
Math	All links to it
Hrithik	All links to it

Fig 1.

2 Aboutness In Web Documents

With the massive increment of information humans curiosity to organize it increases. We have developed several techniques that easily allow us to index it and retrieve it on need. For example, one such method is to meticulously attribute the ideas and thoughts to the ones who have developed them. Another simpler strategy is to cluster the same ideas and create association among them. Undoubtedly, such methods have proven to be very successful and continue to be so, as majority of end users gets what they are looking for. The basic reason is that the subject analysis [1] is easily undertaken, because it tactically balances between what the author wants to say and audiences believes it to be, allowing for effective classification and presentation. So subject analysis means to relate ones thought with another and then analyzing on context point of view whether it properly signifies the objective or not. It is very much important to analyze this as one may give wrong ideas and can distract the end user who seeks the information into wrong path. Traditionally, librarians perform subjective analysis on neutral perspective allowing the ideas to speak on behalf of their authors. Perhaps librarians understand better where to put similar thoughts or ideas and it's placed with greater intellectual context which is far beyond the proposed limited scope of author. With the increment of uncountable information of greater magnitude, the concept of librarian manages the information lay behind. Today's human depends on computers which evaluates information and stores immense information. Recent development in techniques such as OGP [2] signifies that how end users can approach to large information accurately without being falsely trapped into wrong information.

3 Analysis Of Www

To understand the wide acceptance of potential OGP, it is important to analyze some basic facts that one can't expect anyone to perform analysis from an objective standpoint as ones understanding towards one subject may dominate the author's original ideas towards the document. In 1975 W.J.Hutchin [3] identified an inherent problem with the linguistic use of the process itself, judging the "subject" to be too confusing to give correctly the credit to author's ideas. Instead Hutchins suggested the term "Aboutness" from R.A. Fairthorne (1969) who uses the term intentional Aboutness and extensional Aboutness. So intentional Aboutness is defined as the pure concept i.e. pure notion of the total document and the purpose behind creating it. Extensional Aboutness – it shows the individual elements of a document such as paragraphs, headings and its style (as cited in Hutchins 1977 page 24). In 1922, B. Hjorland [4] stated that the term "Aboutness" could not be viewed as neutral because neither the authors, readers, librarians nor any others point of view have certain objective knowledge about the subject of a document. Following with this argument, it lights the need of a method which presents the unique solution and which deals with the problem related to subject analysis and one such possible solution available is Open Graph Protocol. Web engine initially dependent on the "extensional Aboutness" of a web document to perform Aboutness of the document. Search engines semantically analyze the web documents' using their own search algorithms could use keyword matching to determine best approximation of documents subject i.e. what is it about? But this is not that much helpful as it can't judge the overall worthiness of the document as a whole. As a solution to this, Dublin Core Meta data [6] initiative suggested the use of Meta data element set to emphasize on "intentional Aboutness" of documents. As with the increase of information availability across the globe finding appropriate information on the WWW is very difficult task due to the existence of large number of network resources. Also, current web indexing strategies is not enough for richer varieties of resources description. Therefore, more emphasis is done on the purpose of the document i.e. what is it about? Unfortunately, such system bears critical flaws as was so aptly revealed by Google as illustrated by S.Brin and L. Page in 1998. There was no control on what the people put on web and their meta data effort was largely failed because any text on the page which is not directly represented to the user is abused to manipulate search engines (1998 page 6), as stated in there, "Page Rank" algorithm. They used Page Rank algorithms to judge the Aboutness of algorithms through the number of anchor links to the site. However, this is not successful. The success of Google proves them right but still there are flaws to this. Google can't do precisely sentimental analysis. This can be solved by crowd sourcing [7]. However, there are two major categories of criticisms about crowd sourcing, first the value and impact of work received from the crowd and second the ethical implications of low wages paid to crowd workers. Most of these criticisms are directed towards crowd sourcing systems that provide money in terms of the credit to their work .Hence the possible solution to these problems is Open graph Protocol.

4 OPEN Graph Protocol

Using OGP one can integrate their web page to the face books social graph. It is designed for web pages which represents profile of real world things like movies, sports, persons etc. The Open Graph Protocol was developed by Face book engineers and is inspired by Dublin Core Meta data, Micro formats and RDFa. It allows its users to place Meta data information into web documents that allow standardized categorization and classification of its contents (The OGP 2010 Footer). In this aspect, it works similar to RDFa. This solves the problem of false delivering of information in documents Meta data through the use of 'like' button that the user can click on it and link it in his profile so that his social network can share this with him. Hence, it works similar to Google's popular and relevant web documents. However, the contents of the citation are determined by the Meta data specified by the author of the document. Thus on Facebook this Meta data is used to display information about the document in the user's profile page. This means that if user wants to add false Meta data in their document either in some wrong intention or by mistake the user who previously liked it can in turn 'unlike' the document and ultimately disconnects itself from that web document. Moreover if the information is malicious i.e. not appropriate for others as well we can inform the company against those who are involved in provoking people for riots etc. by stating harmful content in their documents. An analyst could possibly suggest that the verb 'like' signifies certain amount of likeliness of a person towards a subject (or topic) shows the positiveness of a person means we can assume that the user has understood that subject but this may lead to trouble when the content of document is of negative attitude e.g., against religion etc. , suppose if your boss is against any religion and he likes that, you too have to like it (there may be any reason) or your friend likes something which is negative, you too does it without knowing the context or author views, likes it. So "recommend" option is better than like as it shows neutrality and can be liked at your own will but this too has flaws . Studies show that some of the companies/ media publishing use OGP functionality to make a preliminary assessment of the 'Aboutness' of the document. So they take people opinion before finalizing an article, document, even products, etc. therefore in terms of organization and sharing of information, the OGP represents a fundamental advancement in how subjective analysis is performed on web document.

4 OGP For Search Engine Results

Search Engines especially the largest one Google can take help of facebook (via its OGP) and takes millions of reviews of any web document across the globe within seconds but this should be kept in mind that relevant documents should be shown to relevant users based on the designation attribute of users profile. Otherwise say for e.g. an artist can't give his views on topics of engineering.

Facebook user's sees it, likes it and that information of the most likes should be used by the Search engines to prioritize their results based on the number of likes. So it has two benefits First sentimental analysis has already been done by large number of persons, second search engines gets the accurate "Aboutness" of documents automatically most of the time. Now as a Matter of fact that there may be a possibility that certain likes may be done without knowing what the Document is all about, there may any reason but now what to do? Then for such, search engines conform that whether the information is relevant or not they should do two types of analysis first of all the most liked documents must be psychologically analyzed and secondly manually sentimental analyzed by the expert. So these measures can drastically improve accuracy of the "Aboutness" of documents and search engines get the easy approach of accuracy of information widespread across the globe because people through social media sentimentally analyze by themselves. So Search engines must take help of social sites (like Facebook) in getting the analysis of information "Aboutness", to present the correct and satisfying search results.

Subject	DNA Computing			
S.no	Document Name	No. of Likes	PRIORITY	VERIFIED
1.	DNA	3456000	HIGH	YES
2.	DNA 1	55620	HIGH	YES
3.	DNA 2	35122	MED	YES
4.	DNA 3	2000	LOW	YES

5 Search Engines Indexing On Their Data Based On Number Of Likes

Search Engines can use simple data structures or any advance data structure which they use. Only it requires is an extra column of number of likes plus final verification by the company analysts. So this way might be convenient for companies to access their database on request.

6 Proposed Model

Search Engines with authentication publishes on Facebook its articles, and Facebook OGP spreads it across the globe. Relevant articles published on appropriate designation users e.g. artist can't give his review on engineering topic etc. Also, it will not be displayed to users with no designation, only general topics will be displayed to the user. Then search engines gathers most liked documents data from facebook's database for indexing its own database. Facebook delivers it to search engine simply in tabular format of number of likes for the relevant document and search engines indexes their databases. When request comes, it will display the most liked documents first according to what the user has requested.

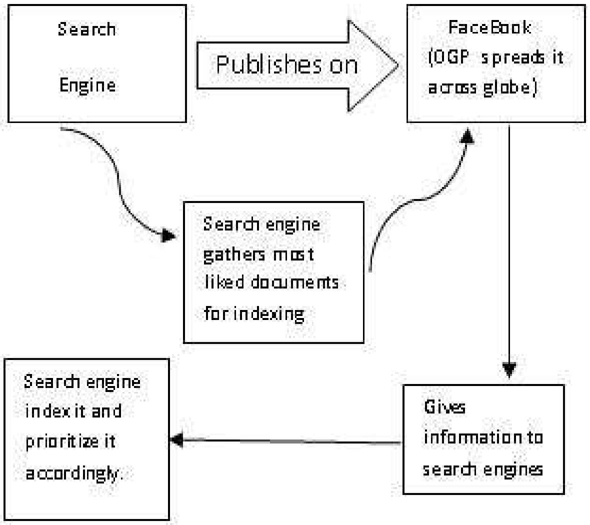

Search Engine
Publishes on
FaceBook (OGP spreads it across globe)
Search engine gathers most liked documents for indexing
Search engine index it and prioritize it accordingly.
Gives information to search engines

Fig 2.

7 Conclusion

If Search engine uses OGP it can increase their accuracy in the outcome of search result because it covers all aspects of determining the "Aboutness" of a web document. People from different nations, views, ideas give their reviews about the document simply by like even if 50% of them reads and tries to understand what the author wants to say, the whole battle is won. They have done sentimental analysis and search engines gets the by far the most accurate reviews of the documents. Important fact is that any user voluntarily sees the document and likes it according to the understanding of the subject. We can say it is neutrality approach of user towards any topic because the user is not paid for this that brings positiveness and can give its reviews without any pressure and likes the document on Facebook. Hence, with this idea we conclude that OGP is definitely a good choice among any other searching techniques on the web.

REFERENCES

[1] Defining intention aboutness in web documents by Max Neuvians.
[2] The Open Graph Protocol. (2010). Retrieved from http://opengraphprotocol.org/
[3] Hutchins, W. J. (1975). Languages of indexing and classification; a linguistic study of structures and functions
[4] Hjorland, B. (1992). The concept of 'subject' in information science. Journal of Documentation, 48(2), 172-200.
[5] Brin, S. & Page, L. (1998). The anatomy of a large-scale hyper textual web search engine. Paperpresented at the Seventh International World-Wide Web Conference (WWW1998).Retrievedfromhttp://ilpubs.stanford.edu:8090/361/1/1998-8.pdf
[6] Dublin Core metadata retrieved from Dublincore.org
[7] http://en.wikipedia.org/wiki/Crowdsourcing

- *Amit Soni,B.Tech.(I.T.), GEC Ajmer*
  amitsonieca7@gmail.com
- *Deepak Gupta,Asst.Prof.,Dept.of CSE,GEC Ajmer*
  gupta_de@rediffmail.com
- *Ajay Dadhich,Asst. Prof.,Dept. of EIC,GEC Ajmer*
  ajaydadhich@gmail.com