International Journal of Latest Trends in Engineering and Technology
Special Issue SACAIM 2016, pp. 537-543
e-ISSN:2278-621X

A STUDY AND ANALYSIS OF THE IMPACT OF STEMMING OVER NAVIGATIONAL QUERIES

Shekar[1], Sukesh, Vishal[2], Vishwas[3] and Ruban S[4]

Abstract:-The information space that is created because of the impact of web is very vast and of unimaginable size. Though it has offered lot of advantages towards the seeking of information, it has also resulted in lot of challenges, which has been of great concern in the research community. Information Retrieval is an area in computer science that deals with search of information in a given domain or information repository. In the context of Information Retrieval, stemming has also been one of the sub processes that is used for query reformulation, which is aimed at increasing the relevancy of the search. A study and the analysis on the queries has classified queries into three categories such as i) Informational queries ii) Navigational queries and iii) Transactional queries. Identifying the type of query also plays a vital role in increasing the precision of search. In this paper, we have studied the impact of stemming over the Navigational queries. Navigational queries are aimed to reach a site through the search engine. Not every query expansion is profitable. Hence, in our experimental study we analyze the impact of Stemming over the Navigational queries. We used the well-known stemming algorithm called Porter's stemming algorithm for our experiment, and the implementation was done using Java. We also studied the impact in three well-known search engines such as Google, Bing and Ask.

Keywords : Information Retrieval, Stemming, Navigational queries, Query Reformulation.

I. INTRODUCTION

The information space that is created because of the impact of web is very vast and of unimaginable size. Though it has offered lot of advantages towards the seeking of information, it has also resulted in lot of challenges which has been of great concern in the research community. Information Retrieval is an area in computer science that deals with search of information in a given domain or information repository. After the advent of world wide web, the scope of this field has grown a lot, in particular after the success of searching applications, there has been vibrant growth in Information Retrieval. The information space shared in the online information retrieval systems such as search engines are very larger compared with that of traditional retrieval systems whose scope was very less or narrow. Today because of the information explosion that has happened, the data that is in the web is very dynamic in nature and of highly unstructured type. Hence the traditional way of searching the data will never be helpful to handle this mountain of data. In our research we will focus on using ontologies with the standard web. Actually, we are going to focus on using ontologies for query refinement. In

[1] *Aloysius Institute of Management and Information Technology (AIMIT) St Aloysius College(Autonomous) Mangalore, Karnataka, India*
[2] *Aloysius Institute of Management and Information Technology (AIMIT) St Aloysius College(Autonomous) Mangalore, Karnataka, India*
[3] *Aloysius Institute of Management and Information Technology (AIMIT) St Aloysius College(Autonomous) Mangalore, Karnataka, India*
[4] *Aloysius Institute of Management and Information Technology (AIMIT) St Aloysius College(Autonomous) Mangalore, Karnataka, India*

this proposed work we would like to refine the queries using the stemming algorithm and will study the impact of it in navigational queries.

II.QUERY

A query is something a user uses to enter into a web search engine to satisfy his or her information needs. Search queries are very distinctive in that they are often hypertext or plain with optional search-directives. They change or vary greatly from standard query languages, which are followed by the strict syntax rules as command languages withpositional parametersor keyword. The average length or size of a search query is 2.4 terms. About half of the users entered a single query while a little less than a third of users entered three or more unique queries. Close to half of the users examined only the first one or two pages of results. Less than 5% of users uses advanced search features (e.g., Boolean operators like AND, OR, and NOT). The top four most frequently used terms were, *(empty search), and, of,* and *sex.*Queriescan be mapped to pages during search sessions based upon information in the query longs and click logs based upon user actions, such as the selections of pages, and whether or not clicks to them were long clicks or short clicks.

.Navigational Query:

A navigational query is a web search query entered with the intent of finding a particular webpage. For example, a user might enter “voot” into Google's search bar to find the Voot 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.A quality score used to determine whether a query is a navigational query can include an information retrieval score for the results that appear for the query as well as characteristics of the URLs that show up for it as well. For example, the shorter the length of the URL and lower the directory depth of the URL (the closer to the root directory) the higher the quality score of the query.

III.STEMMING ALGORITHM

Stemming is the process of removing affixes (prefixes and suffixes) from words. In the information retrieval, stemming is used to conflate word forms to avoid mismatches that may affect a recall. As a simple example, consider searching for a document entitled “How to singing”. If the user issues the query “singing” there will be no match with the title. However, if the query is stemmed, so that “sing” becomes “sing”, then retrieval will be successful. In many variouslanguages stemming is effective for retrieval performance. For instance, in Hebrew, stemming increase the number of documents retrieved by between 10 and 40 times. In English profit tend to less dramatic. Nonetheless, stemming has resulted to produce reliable retrieval improvement. Furthermore, affixes often carry information such as part of speech, plurality, and/or tense that is crucial for the development of more sophisticated question/answer information systems. For instance, consider the sentence “The stugynitfelsuntrooled the drutablejupan.” From the word order and affixes, alone, you know that more than one thing did something to one thing. You know this happened in the past and that what was done was the opposite of trooling. Question/answer systems will rely on such structural cues and hence will require a high precision stemmer as a preprocessing step. The most widely cited stemming algorithm was introduced by Porter (1980).

Algorithm:

Step 1: deals with plurals and past participles. The subsequent steps are much more straightforward.
Step 2:

(m>0) ATION ->	ATE	predication	->	predicate
(m>0) ATOR ->	ATE	operator	->	operate
(m>0) ALISM ->	AL	feudalism	->	feudal

The test for the string S1 can be made fast by doing a program switch on the penultimate letter of the word being tested. This gives a fairly even breakdown of the possible values of the string S1. It will be seen in fact that the S1-strings in step 2 are presented here in the alphabetical order of their penultimate letter. Similar techniques may be applied in the other steps.

Step 3:

(m>0) ATIVE ->	formative	->	form	
(m>0) ALIZE ->	formalize	->	formal	
(m>0) ICITI ->	electriciti	->	electric	

Step 4:

(m>1) AL ->	revival	->	reviv	
(m>1) ANCE ->	allowance	->	allow	
(m>1) ENCE ->	inference	->	infer	
(m>1 and (*S or *T)) ION ->	adoption	->		adopt

The suffixes are now removed. All that remains is a little tidying up.
Step 5a:

(m>1) E ->	probate	->	probat			
	rate	->	rate			
(m=1 and not *o) E ->	cease	->	ceas			

Step 5b:

(m> 1 and *d and *L) ->	single letter controll	->		control	
	roll	->	roll		

IV.RELATED WORK

There are various approaches quoted in the literature, which will be explained in the related work section. General lexical ontologies, namely WordNet, have also been used. [Voorhees 1994] describes several experiments onusing WordNet for query expansion. In a more recent study, on which we will build upon, several alternativestrategies to use ontologies for query expansion are presented ([Navigli et al 2003]). In their work a standard webengine, namely *Google.* was used. WordNetis augmented with some additional relations; apart from the standardrelations, namely synonyms, hypernyms (vehicle in “car is-a vehicle”) and meronyms (“car has an engine”) otherrelations such as gloss were considered. The gloss relation can be defined as “appearing in the definition of the word inWordNet”. Initially, a Word Sense Disambiguation algorithm is used and terms are selected based on the type of therelationships. All relationships but gloss caused only a small increase in performance. Only gloss based expansioncaused a significant 26% improvement. However, we should note that the successful gloss relation is not actually apure ontological relation but is very close to statistical co-occurrence relations. Another thing that we should also noteis that the use of WordNet has been criticized on the grounds of low coverage and the fact that the words are selectedfor maximal generality [Schütze et al. 1995] [Kruschwitz 2003]. (Andrei Broder, 2002) at IBM Research [3] proposed the widely known classification based on user intent as navigational web queries, Informational web queries and Transactional

web queries. His classification was however based on some search logs that he collected when he was working with AltaVista Corporation and later was supported by the surveys. His experiments revealed that based on query log analysis 30% are transactional web queries, whereas 48% are informational and 20% are navigational web queries. Out of the user survey he conducted he reveals that an estimate of 39% is informational whereas an estimate of 36% is transactional and 24.5% are navigational.

(Jansen, Spink and Pedersen, 2005) based on their experiment conducted with AltaVista states that there is an increased use of search engines for navigation. Because of the hypermedia capability of the web, it provides a very unique way of browsing where we use search engines to move to another site of interest or of need. Another classification of queries was made by (Rose and Levinson, 2004) based on the query user used to search, the results which the user viewed. They sub categorized as informational queries, navigational queries and then resource queries.

The studies that were quoted above were based on query logs from the Search Engines and many of the classifications were manually done. Some of the works that followed automatic classification are as follows. (Lee, Liu and Cho, 2005) in their work based on the automatic identification of user goals classified the queries into informational and navigational queries. They did this experiment with 50 queries collected from a group of students in a university. They could experience a success rate of over 50%.

(Dai et al, 2006) in their work on detecting commercial intention studied whether the user information need given by user has any commercial motivation or not. They later revealed that 38% of the queries that were taken for the study have commercial intention. (Baeza-Yates, Calderon_Benavides and Gonzalez, 2006) in their work on the intention behind the web queries used supervised and unsupervised learning to divide the queries as informational or not informational or ambiguous. They declared that they were able to achieve 50% of precision. (Bernard J. Jansen, Danielle L. Booth and Amanda Spink, 2008) in their work on determining the user intent of queries defined and presented three levels of hierarchical classification of user intent for searching web.

Ruban and Behin Sam in their work on Navigational queries[10] studied the impact of Wordnet in processing Navigational queries.

In this work, we study the impact of stemming while processing Navigational queries.

V.EXPERIMENTAL EVALUATION

Navigational queries are aimed to reach a site through the search engine. Not every query expansion is profitable. Hence, in our experimental study we analyze the impact of Stemming over the Navigational queries. We used the well-known stemming algorithm called Porter's stemming algorithm for our experiment, and the implementation was done using Java. We also studied the impact in three well-known search engines such as Google, Bing and Ask.

The following table lists out the different Navigational queries and their relevant results generated in different search Engines at different time intervals.

Table 1: List of Navigational Queries and their precision values

Sl_No	Navigational Query	Google	Bing	ASK
1	VRL Tourist and Travels Udupi	49	37	25
2	How to solve disconnected network in laptops	67	63	19
3	Lenovo service centers around udupi	39	17	23
4	Ocean of games address	56	38	10

5	Meenakshi Textiles Udupi	24	10	02
6	Samsung Mobiles Sales and Service centre	82	51	24
7	Flipkart Big Billion Day Offers	91	78	30
8	Amazon Diwali Offers	81	69	29
9	Weather updates in karnataka	56	48	23
10	Kerala Tourism Places for monsoon season	66	73	24
11	Kalyan Jewelers Kerala	57	23	25
12	Malabar Gold and Daimonds	47	61	26
13	Different types of Malls in City Centre Mangalore	36	14	16
14	Flagship Smartphones under 15000 in India	81	68	38
15	Jos Alukas Jewelers Mangalore	42	23	27
16	Club Laptops Service Centre Udupi	50	1	29
17	Different Types Of Connections	64	30	16
18	Top Ten Social Networking Sites	80	62	24
19	Top Ten Haunted Places in India	79	54	28
20	Harry Potter Series E-Books	93	39	26

The following table (Table 2) lists out the queries that were generated once the original queries go through the process of Stem ming.

Table 2: RefinedNavigational Queries using Stemming Algorithm and their precision values

Sl_No	Refined Query After Stemming	Google	Bing	ASK
1	vrl tourist and travel udupi	36	34	27
2	how to solv disconnect network in laptop	51	58	19
3	lenovoservic center around udupi	47	17	06
4	ocean of game address	61	39	10
5	meenakshitextiludupi	19	10	02
6	samsungmobil sale and serviccentr	71	45	23
7	flipkart big billion dai offer	79	52	30
8	amazon diwali offer	90	58	29
9	weather updat in karnataka	39	42	23
10	kerala tourism place for monsoon season	64	64	24
11	kalyan jewel kerala	54	15	25
12	malabar gold and daimond	61	56	26
13	differ type of mall in citicentrmangalor	34	9	16
14	flagship smartphon under 15000 in india	73	48	38
15	joaluka jewel mangalor	34	14	26

16	club laptop serviccentrudupi	44	1	29
17	differ type of connect	49	9	14
18	top ten social network site	77	54	24
19	top ten haunt place in india	71	28	28
20	harri potter seri e-book	24	18	26

The following figures (Fig 1, fig 2, Fig 3) illustrated the comparative study between the values that were generated between the search Engines. Though it generated mixed response, calculating the average precision, there is a marginal increase in the precicion values after refinement using stemming.

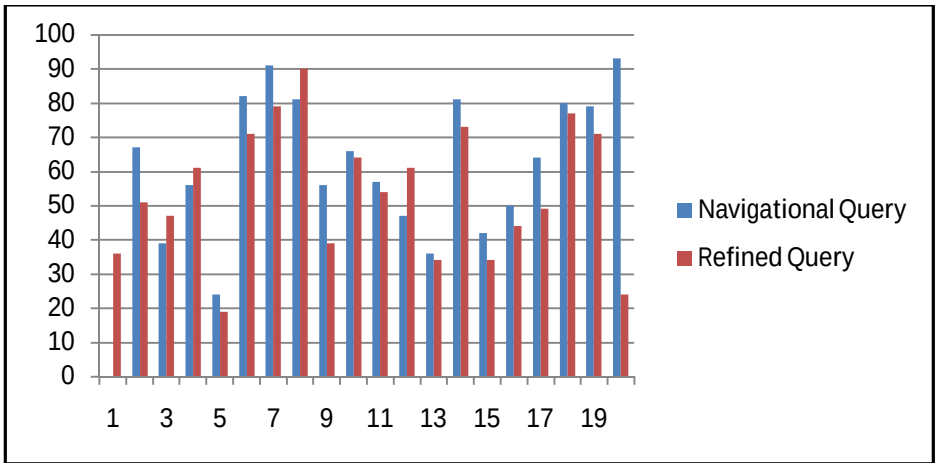

Fig 1: Navigational Queries Vs Refined Navigational queries in GOOGLE

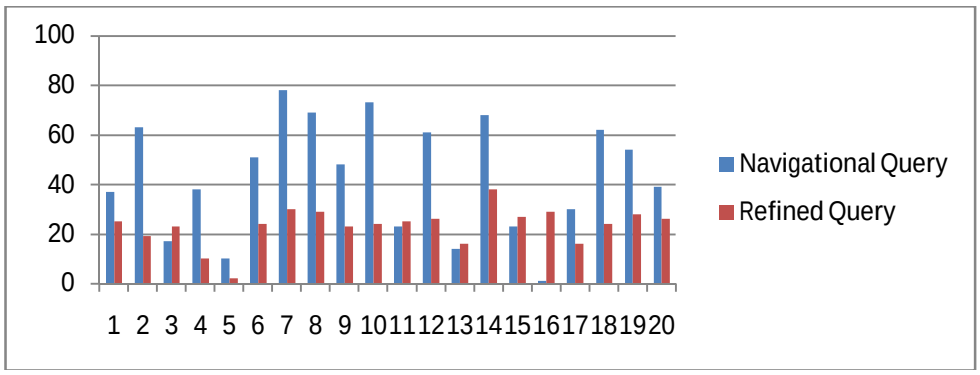

Fig2: Navigational Queries Vs Refined Navigational queries in BING

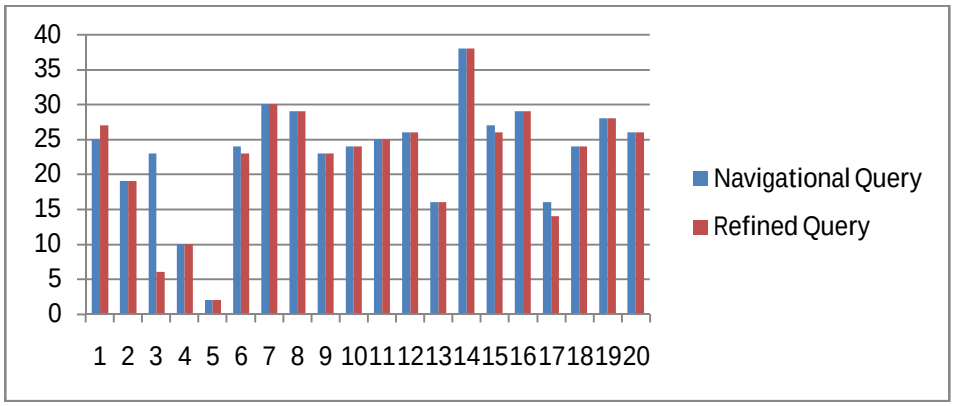

Fig 3: Navigational Queries Vs Refined Navigational queries in ASK

Hence, we conclude that, though stemming is a important process in query processing, in case of navigational queries it will not benefit the query expansion, rather it decreases or does not impact the query expansion in a positive way. But, however the values we got were during the month of October 2016. Hence, the timing of execution will also have an impact in the results obtained.

VI.CONCLUSION

This paper has presented a study and an analysis on the process of stemming over navigational query and, we study the differences between the navigational query and refined query after stemming. The results are however based on the time of execution.

REFERENCES

[1]. Ellen M. Voorhees, 1994, Query Expansion using Lexical-semantic relations, In proceedings of the 17th ACM-SIGIR Conference, pages 61-69.

[2]. R.Navigli and P.Velardi, “An analysis of ontology-based query expansion strategies”, workshop on Adaptive Text Extraction and Mining(2003).

[3]. Kruschwitz ,” Automatically acquires domain knowledgefor ad hoc search: evaluation results. Kruschwitz U.In Proceedings of the 2003 International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE'03), Beijing, 2003.IEEE.

[4]. Baeza-Yates, R., Calder´on-Benavides, L., &Gonz´alez, C. (2006). In The intention behind Web queries (pp. 98–109). Paper presented at the string processing and information retrieval (SPIRE 2006), 11–13 October, Glasgow, Scotland.

[5]. Bernard J. Jansen, Danielle L. Booth, Amanda Spink. (2008). Determining the informational, navigational and transactional intent of web queries, Information processing and Management 44(2008) 1251 – 1266.

[6]. Broder, A. (2002).A taxonomy of Web search. SIGIR Forum, 36(2), 3–10 Carmel, E., Crawford, S., & Chen, H. (1992).In Browsing in hypertext: A cognitive study (pp. 865–884). Paper presented at the IEEE transactions on systems, man and cybernetics, 5–10 October, Chicago IL.

[7]. Dai, H. K., Nie, Z., Wang, L., Zhao, L., Wen, J. -R., & Li, Y. (2006). In Detecting online commercial intention (OCI) (pp. 829–837). Paper presented at the World Wide Web conference (WWW2006), 23–26 May, Edinburgh, Scotland.

[8]. Jansen, B. J., Spink, A., & Pedersen, J. (2005). Trend analysis of AltaVista Web searching. Journal of the American Society for Information Science and Technology, 56(6), 559–570 Lee, U., Liu, Z., & Cho, J. (2005). In Automatic identification of user goals in Web search (pp. 391–401). Paper presented at the World Wide Web conference, 10–14 May, Chiba, Japan.

[9]. Rose, D. E., & Levinson, D. (2004). In Understanding user goals in Web search (pp. 13–19). Paper presented at the World Wide Web conference (WWW 2004), 17–22 May, New York, NY, USA.

[10]. Ruban S and Behin Sam., An Experimental analysis of The Impact of Domain Independent Ontology Over Navigational Queries. ***Aust. J. Basic & Appl. Sci.,*** *9(16): 244-248, 2015*