International Journal of Latest Trends in Engineering and Technology
Special Issue SACAIM 2016, pp. 188-193
e-ISSN:2278-621X

AN ANALYSIS AND THE STUDY OF DIFFERENT QUERY REFORMULATION STRATEGIES

Ruban S[1], Marel Neill Moras[2] and Soujanya Nayak[3]

Abstract- Information Retrieval involves the process of retrieving information from various sites, depending on user's needs or benefits. It can either be a Full-Text or content based search. Various developments have been made to improve the retrieval process. One of the ways to do it is through Query Expansion. It is a process of adding more relevant words to the original seed query. Many ways have been suggested over the period of time. Thesaurus is one such concept in Information Retrieval. Some works suggest adding more terms from Thesaurus increases the precision of the results retrieved. Word net is similar to thesaurus and it finds similar words based on proximity of the words in the lexical database and groups them based on their meaning. Some works have been done in this area as well. They show a mixed response for different queries. In this work we study the query expansion based on thesaurus and Word Net. We illustrate our experimental results by evaluating them in the most used search Engine Google, then comparing the results.

Keywords: Information Retrieval, Query, Synonym, Lexical database, Synsets.

I.INTRODUCTION

Information Retrieval involves the process of retrieving information from various sites, depending on user's needs or benefits. It can either be a Full-Text or context based search. Various developments have been made to improve the retrieval process. One of the ways to do it is through Query Expansion. It is a process of adding more relevant words to the original seed query. Many ways have been suggested over the period of time. Thesaurus is one such concept in Information Retrieval. Some works suggest adding more terms from Thesaurus increases the precision of the results retrieved. Word net is similar to thesaurus and it finds similar words based on proximity of the words in the lexical database and groups them based on their meaning. Thesaurus lists all the words together grouped depending on the similarities which contain synonym and sometimes antonyms and related words. It helps in finding words which are related but having different shades of meaning. Thesaurus makes use of the synonyms that are grouped into unordered sets. Thesaurus also consists of POS (Parts of Speech) tagger.

WordNet groups the words into Synonyms called synsets, it also specifies the relationship between these synonyms. It interlinks words which are in close proximity to one another. WordNet also gives the semantic relation among words. We make use of Thesaurus Api

[1] *St Aloysius Institute of Management and Information Technology Beeri, Mangalore, Karnataka, India*
[2] *St Aloysius Institute of Management and Information Technology Beeri, Mangalore, Karnataka, India*
[3] *St Aloysius Institute of Management and Information Technology Beeri, Mangalore, Karnataka, India*

which suggests similar meanings and helps in query refinement. The old manual Thesaurus methods would just provide better, richer vocabulary to a writer. But now many improvement has taken place and it aims on: Indexing by giving a common precise and controlled vocabulary, Coordinate document indexing and document retrieval, it selects the most relevant term and it also refines query by reformulating and query contraction. WordNet makes use of semantic databases and is used for various purposes in information systems. Various techniques have been used such as close proximity, finding the distance between two words. The closer the distance between the words the better their meaning. WordNet is developed and maintained by Princeton University.In this experiment we make use of the output from Thesaurus and the output from WordNet, later we will compare these two results and find the best way to retrieve the results among these two methods that will increase the precision of the results.

II.RELATED WORK

In the past few systems have been proposed that is expected to increase the query expansion. Thesaurus is iIn Literature Query expansion are studied in different ways for instance query expansion methods are classified into Automatic Query Expansion (AQE) methods and Interactive Query Expansion Methods (IQE) [1] Earlier studies reveal that many of the automatic query expansion methods rely on the Relevance Feedback techniques [2] proposed by salton and Buckley, in which the terms featuring prominently in documents marked relevant by the user are automatically added to the query.

Later Srinivasan came up with a Retrieval Feedback technique [3] that adds terms from the top relevant documents to the query. This technique has shown considerable improvement in many retrieval tasks. Query logs was used as a means of query expansion by Hangs et al [4]. Later Huang et al [5] proposed a query expansion algorithm of pseudo relevance feedback based on matrix-weighted association rule mining.

However in the year 2001 Aronson [6] proved that query refinement that is based on ontology is much more efficient than the other methods that were available. Using ontology for query expansion goes back to 1994 where Voorhees [7] attempted using the Domain independent ontology WordNet for query expansion. Since then there has been some works done in this area. The word sense information and the ontology was used for query expansion by Navigli and Velardi[8]. They succeeded in using ontology to extract the semantic domain of a word and then the query is expanded further using co-occurring words. Further Query refinement techniques based on domain and geographical ontology was studied by Fu,G et al [9]. The Domain ontology was modeled after tourism which consists of some non-spatial terms such as “near” whereas the geographical ontology consists of some spatial terms such as place names. A domain specific ontology based on Stockholm University Information systems (SUiS) was developed by Nilsson et al10].

III. EXPERIMENT AND RESULTS

In this analysis we have followed certain procedures to get the relevant result from the query which is given by the users. The input query which is also called as seed query is entered by the user into the browser. The query is passed to the thesaurus API. The browser returns the result, the result contains noun, adjectives, verb Synonyms and sometimes antonyms. The results obtained are in different forms which are grouped as syn(synonym), user(user suggestions), ant(antonyms), sim (similar terms), rel(related).

Now we enter the same seed query into WordNet, we got the Synonyms, antonyms, similar terms, definition of the entered query and also meaningful sentences for the query. We have found a few similar words but WordNet even provided meaningful sentences, the described

the query. A few queries such as 'mobile phone', 'animal kingdom' returned equal number of results in Thesaurus and WordNet whereas queries such as 'Types of Indian Spices', 'Home Appliances' returned results in WordNet but not in Thesaurus and finally some queries WordNet returned more number of results than that of Thesaurus.

These are the following queries:

Table 1: Sample Experimental Queries

Slno	Original Query	Query using Thesaurus	Query using WordNet
1	Fountain pen	Fountain pen or pen	Fountain pen or pen or ball point or quill pen
2	Animal Kingdom	Animalia or Kingdom or Animalia Kingdom	Kingdom or Animalia or Monera or Protoctista or Plantae or Fungi
3	Solar System	Scheme or System	System, scheme or language system or judiciary or eco system or social organization or dragnet or machinery or network or nonlinear system or body
4	Home Appliances	Home Appliances	Household appliances or dryer or drier or home appliances
5	Water tank	Cistern or storage tank or tank	Tank or storage or aquarium or cistern or gas holder or gas tank or reservoir or water heater
6	Computer Hardware	Hardware or component or consistent or element or software	Computer Hardware
7	Mobile Phone	Cellular telephone or cellular phone or cell phone or radiophone or radio telephone or wireless phone	Radio telephone or cellular telephone or cell or cell phone or mobile phone
8	World Wide	World Wide	Global or cosmopolitan or world
9	Common sense	Good sense or gumption or horse sense or sense or mother wit or discernment or judgment or sagacity	Common sensible or common sense
10	Types of Indian Spices	Types of Indian Spices	Types of Indian Spices
11	Story teller	Story teller	Speaker or all iterator or caller or chatter or growler or lecturer or lisper
12	Parts of speech	Parts of speech	Grammatical category or case or number or person or gender or tense
13	Social Service	Welfare work or work	Work or wash or action or job or service or shining or paperwork or timework or coursework or care or attention or duty or investigation
14	Present Tense	Present tense	Tense or present or aorist or past or future or progressive or perfective
15	Self-made	Self-made	Self-made

The following table illustrates the results derived based on the query expansion using Thesaurus. In our example we used the widely used Thesaurus called BigHugeLabs and also the domain independent ontology called WordNet. The values that were generated using google is given below.

Table 2: Precision results of Google for Refined Queries using Thesaurus and WordNet

Slno	Without Refinement	Thesaurus Refinement Results	Word Net Refinement Results
1	19	20	25
2	13	13	17
3	1	1	19
4	19	19	42
5	32	32	36
6	29	29	43
7	23	23	19
8	28	28	28
9	10	10	43
10	20	20	20
11	34	34	34
12	29	29	39
13	10	10	23
14	22	22	21
15	24	24	24

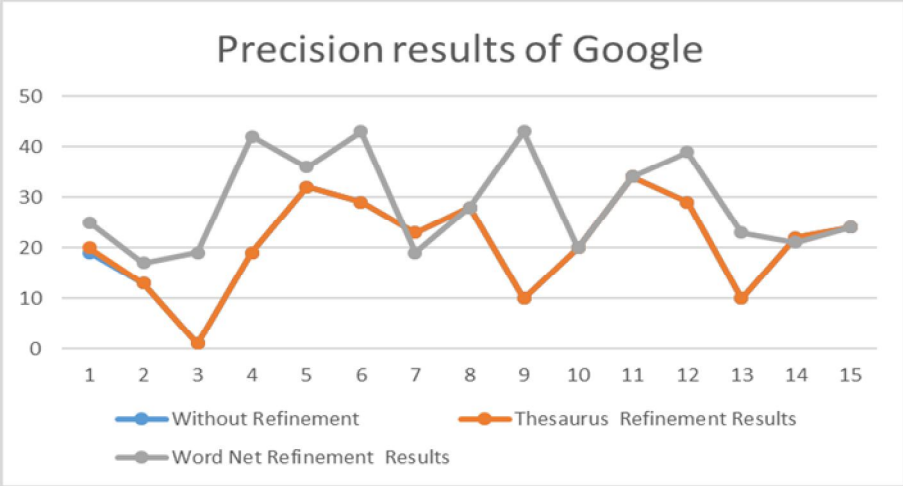

Figure 1

The values that were generated using Bing is given below:

Table 3: Precision results of Bing for Refined Queries using Thesaurus and WordNet

Slno	Without Refinement	Thesaurus Refinement Results	Word Net Refinement Results
1	30	34	30
2	44	47	33
3	1	7	3
4	14	18	26
5	27	29	16
6	46	49	38

7	17	19	17
8	28	35	28
9	3	9	35
10	41	41	41
11	41	58	27
12	43	45	33
13	17	17	35
14	42	42	49
15	37	37	37

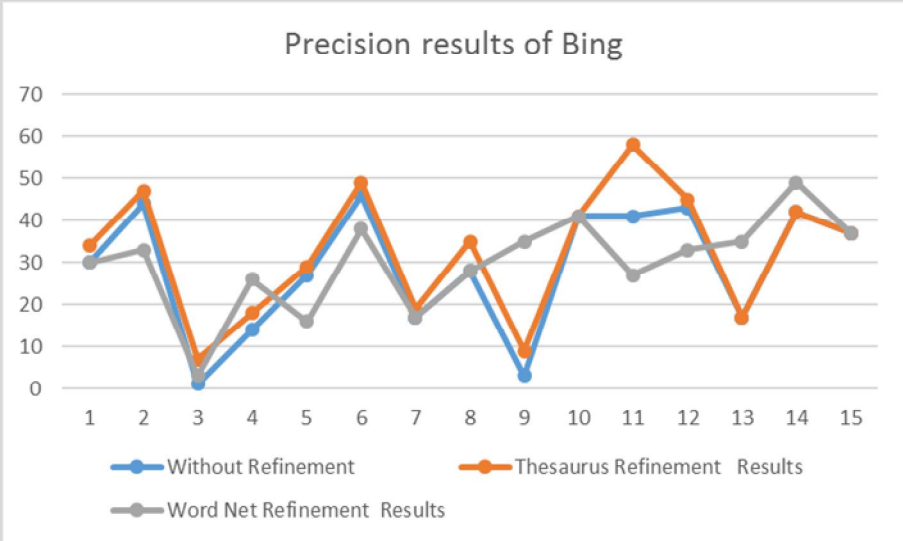

Figure 2

V. CONCLUSION

The results we got may differ based on the time of execution, but we conclude that as per the conducted experiment WordNet has returned more number of results compared to that of Thesaurus. WordNet also gives examples in the form of meaningful sentences for the query given by the user which helps the user in understanding the query better compared to Thesaurus. This helps to remove disambiguation of words and helps understand the meaning of the user's word thus providing results that are more optimised. So therefore we conclude that for query expansion WordNet is a better system than the Thesaurus.

REFERENCES

[1] Query Expansion, Efthimiadis, E.N(1996), Annual Review of information science and Technology.
[2] Salton G & Buckley C (1990). Improving Retrieval performance by relevance feedback, Journal of the American society for Information Science.
[3] Padmini Srinivasan, "Retrieval Feedback in MEDLINE", Journal of the American Medical Informatics Association, 3(2):157-167, 1996c, doi: 10.1136/jamia.1996.96236284
[4] C.Hang, W. Ji-Rong and N. Jian-Yun, "Probabilistic query expansion using query logs", proceedings of the eleventh international conference on World wide Web(2002)
[5] M.Huang, x.Yan and S.Zhang, "Query expansion of pseudo Relevance feedback based on matrix-weighted association rules mining", Journal of Software , 20(7):1854-1865 (2009)
[6] Padmini Srinivasan, "Retrieval Feedback in MEDLINE", Journal of the American Medical Informatics Association, 3(2):157-167, 1996c, doi: 10.1136/jamia.1996.96236284

[7] Alan R. Aronson, "Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program." Proceedings of AMIA, Annual Symposium, pages 17-21, 2001.
[8] E.M Voorhees, "Query Expansion using lexical-semantic relations", proceedings of the 17th annual international ACM SIGIR conference on Research and development in information Retrieval(1994)
[9] R.Navigli and P.Velardi, "An analysis of ontology-based query expansion strategies", workshop on Adaptive Text Extraction and Mining(2003).
[10] Lin Fu,Dion Hoe-Lian Goh and Schubert shouo-Boon Foo, "Evaluating the effectiveness of a collaborative querying environment", proceedings of the 8th international conference on Asian digital libraries [2005].