ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, *Vol 8(24), DOI: 10.17485/ijst/2015/v8i24/80883, September 2015* ISSN (Online) : 0974-5645

A Hybrid Framework to Refine Queries using Ontology

Ruban[1], Behin Sam[2]and Annapoorna Shetty[3]

[1]Department of Computer Science, Bharathiar University, Coimbatore - 641046, Tamil Nadu, India
[2]Department of Computer Science, Rajeswari Vedachalam Arts and Science College, Chengalpattu – 603001, TamilNadu, India
[3]Department of MCA, AIMIT, St Aloysius College, Mangalore – 575022, karnataka, India;

Abstract

The growth of the World Wide Web in the last two decades has posed a lot of challenges to the field of Information Retrieval. The way in which the information is collected, shared and searched is changed drastically. Searching information has never been so easy because of the search engines. Any Information Retrieval application has different components such as query handling where the user enters the user information need, Indexing part where the document representation is stored and maintained, the Ranking part which arranges the documents based on the relevance and the matching part which compares the query representation with the document representation. Most of the time since the user information need is not specified correctly, the documents that are retrieved may not be relevant or the relevant links may be less. Hence it is the challenge to be addressed by the search applications which can transform the original query into another representation which will be more responsive for the information retrieval. In this work we propose a hybrid framework which can be used to transform the original query representation to another representation which helps to retrieve more relevant results than the original representation. We further validate our point with an experiment we conducted.

Keywords: Information Retrieval, Ontology, Query Expansion, Query Refinement, Search Engine

1. Introduction

Search Engines has become an integral part of our life. The field of Information Retrieval has grown well beyond imagination in the last couple of decades with the advent of World Wide Web. Though the World Wide Web offers lot of opportunities for searching and finding data, it also poses many challenges such as minimizing the irrelevant results, query handling etc. The Information Retrieval system helps to find relevant information for a given user's query. Any Information Retrieval application has different components such as query handling where the user enters the user information need, Indexing part where the document representation is stored and maintained, the Ranking part which arranges the documents based on the relevance and the matching part which compares the query representation with the document representation. Most of the time since the user information need is not specified correctly, the documents that are retrieved may not be relevant or the relevant links may be less. Hence it is the challenge to be addressed by the search applications which can transform the original query into another representation which will be more responsive for the information retrieval. In this context many studies have been carried out to analyse how the user information need is represented and handled. Hence Query expansion or query refinement or Query Improvement has become a vibrant field of Research in the domain IR.

There are various approaches quoted in the literature, which will be explained in the related work section. Recently Researchers have started to use ontology to refine the query. Ontologies have been categorized as Domain independent ontology and Domain dependent ontology. One of the widely used domain independent

* *Author for correspondence*

ontology is Wordnet which was developed by G.A. Miller in Princeton University[1] Though there has been lot of study on using the Wordnet for Information Retrieval; Ellen M. Voorhees[2] in her study using Wordnet for TREC Collection concluded that less well developed queries can be significantly improved by refinement. The next Section presents the related research concerning the different query refinement approaches that were used for query refinement.

2. Related Studies

In Literature different approaches towards Query Refinement has been used and they are classified into Local analysis and global Analysis. One of the widely used query refinement method is called as Relevance Feedback techniques[3] which was 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[4] that adds terms from the top relevant documents to the query. This technique has shown considerable improvement in many retrieval tasks. Query logs were used as a means of query expansion by Hangs et al[5]. Later Huang,et al[6]. proposed a query expansion algorithm of pseudo relevance feedback based on matrix-weighted association rule mining.

However in the year 2001 Aronson[7] 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[2] 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 modelled 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 al[10].

Initially the relevant terms from the Domain independent ontology WordNet is displayed to the user, and the option is given to the user to select more terms which he thinks is more relevant, once the refinement is done then the Refined query is passed to the Domain dependent ontology say in our case the plants ontology that we have developed, the IR system then lists the different options that is available from the Ontology and user picks up the terms that he considers as relevant. After concatenating the terms both from the Domain Independent ontology and the Domain dependent ontology, the refined query is then passed on to the search API to search the web. Our experiment reveals that this optimized representation of the user information need will result in retrieving more relevant results than giving the queries directly.

3. Experimental Methodology

3.1 Research Design for the Query Expansion Methodology

For research objective one we analyzed prior work in this area with an analysis of numerous actual query refinement approaches that has been developed and proposed over the period of time. The user query that is entered by the user is subjected to the query refinement process, where we refine the query to make it into a format that will be more responsive for information retrieval. In our case we refine using ontology. There are two types of ontology such as Domain independent ontology such as WordNet or some domain dependent ontology that can be developed for a specific domain. In this case we expand using WordNet.

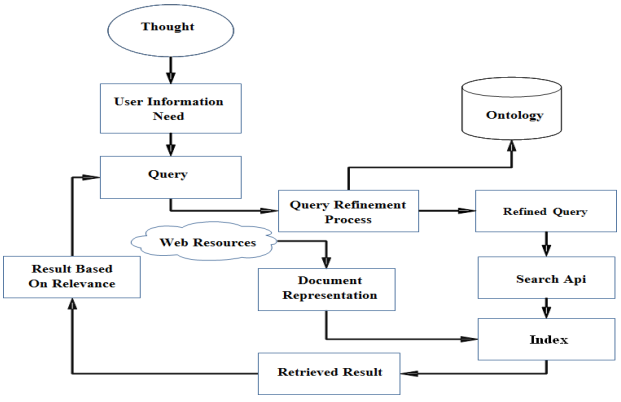

Figure 1. Proposed information retrieval framework that uses ontology for query expansion.

The Research framework that we would use in this experimental study is being shown below. Every user information need that is being given by the user is exactly a thought that is out in verbal form. Query Refinement

process is the process of adding more terms to the original seed terms that makes it more responsive for Information Retrieval. The refined query is later passed to the search API.

3.2 Properties of Web Queries

As already analysed in the related work, based on user intent the queries are being classified based on the task they are deployed for. For eg a query that is aimed to search some information comes under the category of informational queries, whereas a query that is aimed to lead to some website or to specific organization or a specific individual is called as navigational queries and finally the query whose aim is to point to a website with the intention of buying some product or executing some task or execute a transaction fall into the category of Transactional queries.

3.3 Selection of Queries

To achieve the research objective three, we selected 15 queries that belong to the Domesticated plants domain, but on seeing the queries and analyzing them, the aim is to get some information about the topic, hence they come under the category of Informational queries. The following table lists out the queries that we used for our experiment.

Table 1. Experimental Queries

Sl No	Queries
1.	Palm Trees
2.	Usage of herbs
3.	Kinds of FruitTrees
4.	Most popular crop grown by farmers
5.	Examples of ornamentalTrees
6.	Types of gardenPants
7.	Pesticides used for crops
8.	uses for MedicinalPlants
9.	List of CookingPlants
10.	Fast growing vegetables
11.	Some of MedicinalPlant
12.	Types of Grasses
13.	Types of HybridFruits
14.	Which plants are used for vines
15.	vines plants example

4. Results

The Experiment that was conducted was done in the time interval of 4 months from Dec 2014 to Mar 2015. The user's initial query was directly given to the Search API, in our case it was given to the Google search engine. The precision value is recorded, for the first 100 results that were retrieved. The values are recorded in the table and are used to evaluation.

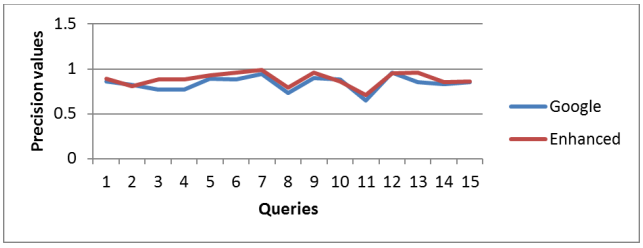

Figure 2. Search API vs after hybrid query expansion.

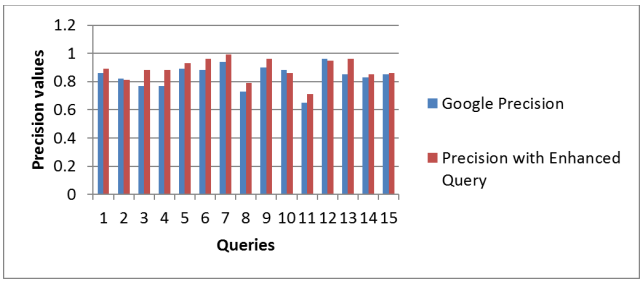

Figure 3. Precision of search API vs precision after hybrid query expansion.

The other set of values are generated using our proposed framework, the initial queries are refined by adding more relevant terms from the domain independent ontology in our case word net, and then refined further by adding some relevant terms from the domain dependent ontology, in our case it was the domesticated plants ontology that was developed in protege 4.0.

Both the set of queries are being listed below and the relevant results are being mentioned below.

The following Figures depicts the pictorial representation between the list of queries that are executed in the normal way, and the other list of queries that were refined using our proposed framework. Figure 2 and Figure 3 represents the comparative results got when executing the experiments in the traditional setup and the hybrid proposed framework. The values are the precision results that were obtained when the experiments were conducted.

Table 2. Queries and their Precision values

Sl No	Sample Queries	Search API Precision	Refined Queries	With Enhanced Query Precision
1	Palm Trees	0.86	Palm Plants or Palm Trees or Edible plants	0.89
2	Usage of herbs	0.82	Uses or Usage of herbs or herbal tea	0.81
3	Kinds of FruitTrees	0.77	Variety or Kinds of FruitTrees or NutTrees	0.88
4	Most popular crop grown by farmers	0.77	Craw or Most popular crop grown by farmers or shrubs	0.88
5	Examples of ornamentalTrees	0.89	Good example or Examples of ornamentalTrees or Palms	0.93
6	Types of gardenPants	0.88	Type or Types of gardenPants or HousePlants	0.96
7	Pesticides used for crops	0.94	Veggieor Pesticides used for crops or coffee	0.99
8	uses for MedicinalPlants	0.73	Usage or uses for MedicinalPlants or Shrubs	0.79
9	List of CookingPlants	0.90	Name or List of CookingPlants or Cereals	0.96
10	Fast growing vegetables	0.88	Rise or Fast growing vegetables or Vegetables oil	0.86
11	Some of MedicinalPlant	0.65	Variety or Some of MedicinalPlant or climber	0.71
12	Types of Grasses	0.96	Type or Types of Grasses or Cyprus	0.95
13	Types of HybridFruits	0.85	Type or Types of HybridFruits or CitrusHybrid	0.96
14	Which plants are used for vines	0.83	Plant or Which plants are used for vines or Hedera	0.85
15	vines plants example	0.85	illustration or vines plants example or Mandevilla	0.86

In Figure 2 the Blue line represents the values got using the search API in our case we used Google, whereas the Red line represents the values got using the Hybrid framework. In Figure 3, the blue bar represents the values got using the Google API, whereas the Red bar represents the values got using the Proposed Hybrid Framework. It is clearly visible that the precision of proposed framework is more than the Traditional searching API. Figure 4 represents the comparison between the average precision value obtained for the 15 query results using hybrid query expansion and the query results in the normal traditional environment.

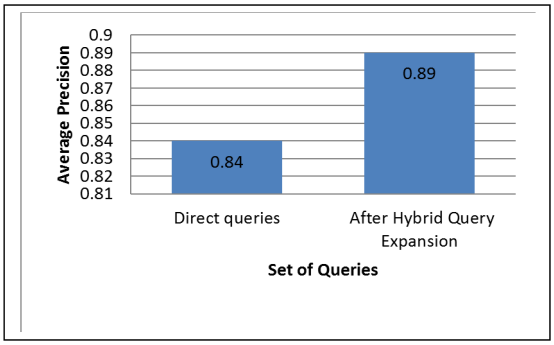

Figure 4. Average precision of direct queries vs after hybrid query expansion.

5. Conclusion

The values that we got may vary based on the time of execution and the date of execution, but our experiment revealed that the precision of the results that we got through the refinement that happened through our proposed hybrid framework is better than the results that we got through the traditional way of executing the queries. Since we have used the Domesticated plants ontology in our work, the better results can be obtained only if the queries related to the ontology are given.

6. References

1. Miller GA. Wordnet: An on-line lexical database. International Journal of Lexicography.1990; 3(4):235–44.
2. Voorhees EM. Query expansion using lexical-semantic relations. Proceedings of the 17th ACM-SIGIR Conference; 1994; p. 61–9.
3. Salton G, Buckley C. Improving retrieval performance by relevance feedback. Journal of the American Society for Information Science. 1990:355–63.
4. Srinivasan P. Retrieval Feedback in MEDLINE. Journal of the American Medical Informatics Association. 1996; 3(2):157–67. doi: 10.1136/jamia. 1996.96236284.
5. Hang C, Ji-Rong W, Jian-Yun N. Probabilistic query expansion using query logs. Proceedings of the 11th International Conference on World Wide Web; 2002.

6. Huang M, Yan X, Zhang S.Query expansion of pseudo relevance feedback based on matrix-weighted association rules mining. Journal of Software. 2009; 20(7):1854–65.
7. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: The metamap program. Proceedings of AMIA, Annual Symposium; 2001. p.17–21.
8. Navigli R, Velardi P. An analysis of ontology-based query expansion strategies. Workshop on Adaptive Text Extraction and Mining; 2003.
9. Fu L, GohDHoe-Lian, Foo SS-B. Evaluating the effectiveness of a collaborative querying environment. Proceedings of the 8th International Conference on Asian Digital Libraries; 2005.
10. Nilsson K, Hjelm H, Oxhammar H. SuiS – cross-language ontology-driven information retrieval in a restricted domain. Proceedings of the 15th NODALIDA Conference; 2005.