A VERTICAL META SEARCH ENGINE WITH QUERY EXPANSION USING ARTIFICIAL RELEVANCE FEEDBACK MECHANISM

Sandeep Joshi¹, Satpal Singh Kushwaha²

¹Associate Professor, Dept. of Computer Science, SEC, Rajasthan, India, sjoshinew@yahoo.com
²Research Scholar, Dept. of Computer Science, SEC, Sikar, Rajasthan, India, singh_satpal25@rediffmail.com

Abstract

World Wide Web is growing day by day so with this rapid development in the size of internet, Information extraction[1] on Internet is gaining its importance day by day. At present there are millions of Websites and billions of homepages available on the Internet. A large amount of on-line information resides on the invisible web[2] – web pages generated dynamically from databases and other data sources hidden from the user. They are not indexed by a static URL but is generated when queries are asked via a search engine (we denote them as specialized search engines or vertical search engine). OpenFind states that it indexes 3.5 billion Web pages; Google claims 2.4 billion, AlltheWeb - 2.1 billion, Inktomi - a little more than 2 billion, WiseNut - 1.5 billion and AltaVista - 1 billion Web pages. No search engine index more than one third of the total size[3] of the web. So from this big collection of web pages information retrieval is a very crucial task. The user query plays a vital role in the information retrieval process. So for the better information retrieval results several methods have been devised which assists the user in the query expansion task. In the proposed system we present a Vertical Meta Search Engine with query expansion using Artificial Relevance feedback mechanism. The proposed system provides a simple way of query expansion based on relevance feedback and reduces the user’s searching time with less no of hits to get the accurate results.

Index Terms: Vertical Search engines, invisible web, query expansion.

1. INTRODUCTION

A Meta search engine (also known as multi-threaded engine) is a search tool that sends user’s search query[4] in parallel to several search engines, Web directories and sometimes to the so-called Invisible (Deep) Web. Invisible web is a collection of online information that is not indexed by traditional search engines. After retrieving the results, the Meta search engine will remove the duplicate links and, according to its algorithm, combine/rank the results into a single merged list which is finally presented to the user. As compare to individual search engines and directories, the Meta search engines do not have their own database of web pages and also do not accept URL submissions.

Query expansion is one of the important parts of the whole searching process. In the proposed system the query expansion is performed with relevance feedback technique. Here we receives the user query from the user and find the unique terms or information terms from the web, then high frequency terms from among these terms are presented in front of the user to choose suitable word to expand his query.

In this paper some real time experiments are also carried out and the produced results show that the proposed system can reduce the overall searching process. Our experiments show that 70 percent cases the suggested terms for query expansion are correct and the system presents suitable vertical search engine to the user.

2. MOTIVATION

1. Increased search coverage because a Meta search engine return results from multiple search engines.
2. Reduces the no of hits to get the relevant document using single query submission.
3. Getting the top best results from the ranked list of the vertical search engine.
4. Query expansion becomes less laborious work, because user doesn’t have to maintain vast size of dictionaries and their maintenance.
3. OPERATIONS

The operations of the proposed Vertical Meta search engine can be divided in the following two parts:

3.1 Offline Operation
3.2 Online Operation

3.1 Offline Operation

This operation basically contains the creation of the database of the vertical search engines. The vertical search engine database contains the relevant terms which describe the search engine, URL of the search engine, connection parameters and domain.

For getting relevant terms which describe a vertical search engine [6] neighborhood based topic identification method is used. In this process relevant terms are obtained from the neighbor web pages, which contain weblink to that vertical search engine. Following method of neighborhood [7] based topic identification is used to collect terms relevant to the vertical search engine.

Back link method

In this method the relevant terms are collected from those pages which have a weblink to the home page of the vertical search engine. we have to send the url of the vertical search engine (http://www.monsterindia.com the job search engine) to a general purpose search engine like http://www.google.com , when the results returned then we can extract unique words which describe the vertical search engine in a better way. Like in case of the http://www.monsterindia.com the relevant terms will be job, career, resume, biodata, interview, etc. The above unique words are describing that the above search engine is a job search engine. The algorithmic steps are as follows.

Algorithm

1. Web pages that have links to the home page of the vertical search engine are identified. This is done simply by sending the vertical search engine’s URL to a general purpose search engine here we uses Google.
2. The identified web pages are downloaded.
3. Distinct terms from the web documents are extracted.
4. Extracted distinct terms are stored in the search engine selection index with their normalized frequencies.

Normalized frequency of a distinct word is calculated as follows:

Let \(HW\) is the frequency of a word that has highest occurring frequencies in the list of the extracted word list.

So normalized frequency of a word \(w\) is

\[\text{Occurring frequency of word } w/ HW\]

We have taken a limit on the no of words to be stored in the search engine selection database to 15. The distinct words with their frequencies are stored in the search engine selection index.

3.2 Online Operation

Online operation can be divided into the following two parts:

3.2.1 Query expansion
3.2.2 Search engine ranking

3.2.1 Query expansion

In 1960, Maron and Kuhn’s mentioned query modification first time by suggesting that terms closely related to the original query terms can be added to the query and thus retrieve more relevant documents.

There are three approaches for the user query expansion[8]

- a. Manual Query Expansion
- b. Interactive Query Expansion
- c. Automatic Query Expansion

Manual and interactive query expansion requires user’s involvement in the query expansion process. In the proposed system we uses the approach of automatic query expansion which is the process of supplementing additional term or phrases to the original query to improve the retrieval performance without users intervention.

Several methods for query expansion have been developed and used. Some methods get relevant terms from thesaurus[9], but the creation and the maintenance of the thesaurus is a laborious work. Other methods are using Latent semantic analysis where user query terms are semantically analyzed and relationship in between them is used to expand the user query but that approach also uses large dictionaries.
Recent methods for query expansion are mining user logs[10] and construction of user profile[11]. Web based query expansion is the process where the information terms to the user query are retrieved from web directly by sending the user query to the general purpose search engine then the unique or distinct words with high information are retrieved from the returned results. Because we are using automatic[12] query expansion so the top three unique words with high frequency are automatically selected to expand the user query.

The proposed algorithm proceeds as follows:

1. The user submits the query to the proposed system.
2. Sending the user query to a general purpose search engine.
3. Extraction of the distinct words from the results returned from search engine.
4. Calculating the frequency of the extracted distinct words.
5. The unique words are then arranged in the decreasing order of their frequencies.
6. At last, top three words are chooses for the query expansion.

3.2.2 Vertical Search Engine Ranking

In this process the best vertical search engine is picked out from the list of the vertical search engine. Here the relevant terms which were retrieved in the query expansion step are compared with the representative terms of the vertical search engine terms stored in the search engine database. After the matching process is finished top two or three best search engines are picked out from the search engine database.

Search engines are selected on the basis of the relevance factor. The relevance factor of a search engine $e$ for a given set $W= (w_1, w_2 \ldots)$ of query expansion terms is as follows:

\[
\text{Relevance Factor} \ (e, W) = \sum (f_i \cdot c_i)
\]

Where $c_i$ is the no of occurrences of $w_i$ counted in the query expansion process and $f_i$ is the frequency of term $w_i$ in the search engine selection index for $e$.

After the search engine ranking [14] the results are presented in front of the user. The results contain the most suitable vertical search engine for the expanded user query.

4. EXPERIMENTAL RESULTS:

Real time queries are supplied to the proposed system. We have devided the experimental results in three parts which are as follows:

4.1 Results during creation of search engine selection index
4.2 Query Expansion Results
4.3 Search Engine Ranking Results

4.1 Results during creation of search engine selection index

We have presented some results which were received during creation of search engine selection index. For the creation of search engine selection index total 69 vertical search engines were selected from different different areas like medical, news, blogs, people search, source code etc. In the search engine selection index candidate terms or representative terms related to a vertical search engine are stored with their frequencies. These terms are stored in the decreasing order of their frequencies. This search engine selection index is used at the time of the search engine ranking. Following is the table containing URL of the vertical search engine and representative terms.

<table>
<thead>
<tr>
<th>Website URL</th>
<th>Distinct words with high frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://www.animalsearch.net">www.animalsearch.net</a></td>
<td>Animal, wallpaper, Pets, desktop, originals, wildlife, include, pet, website, insects, Dogs, cats, birds, categories, breed, Enter, charolais</td>
</tr>
<tr>
<td><a href="http://www.musicgama.com">www.musicgama.com</a></td>
<td>Gama, flash, media, music, song, lyrics, opacity, plain, volar, videos, haces, open, bomb, absolute, Indian, Pictures, license, mp3</td>
</tr>
<tr>
<td><a href="http://www.artcyclopedia.com">http://www.artcyclopedia.com</a></td>
<td>Art, artcyclopedia, online, fine, database, viewed, deals, john, Canadian, founded, site, malyon, museum, quality, artist, website, names, artists, museums</td>
</tr>
<tr>
<td><a href="http://www.medhecker.com/">http://www.medhecker.com/</a></td>
<td>Medical, information, checker, engine, drug, check, med, info, health, Lookup, photos, WebMD, seo, conditions, Analysis, online, medications, interaction, advice</td>
</tr>
</tbody>
</table>

Table-1: Representative terms with URL of Vertical Search Engine
4.2 Query Expansion Results

<table>
<thead>
<tr>
<th>User Query</th>
<th>Terms for query expansion with their frequencies</th>
</tr>
</thead>
</table>

Table-2: Terms for query expansion with their frequencies

4.3 Search Engine Ranking Results

The expanded user query terms are compared with representative terms of the vertical search engine which are stored in the search engine selection index. So on the basis of the relevance factor the top three vertical search engines are selected which are presented to the user with the relevance factor and expanded query.

<table>
<thead>
<tr>
<th>Suggested Term</th>
<th>Url</th>
<th>Relevance Factor</th>
<th>Fire Query</th>
</tr>
</thead>
<tbody>
<tr>
<td>cheap hotels hostels</td>
<td><a href="http://www.aardvar">http://www.aardvar</a> ktravel.net/</td>
<td>12</td>
<td>book+london +cheap book+london +hotels book+london +hostels</td>
</tr>
</tbody>
</table>

Table-3: Results for user query “Book London”

4.3.1 User Query: “Book London” Related Results

4.3.2 User Query: “Video” Related Results

<table>
<thead>
<tr>
<th>Suggested Term</th>
<th>Url</th>
<th>Relevance Factor</th>
<th>Fire Query</th>
</tr>
</thead>
<tbody>
<tr>
<td>videos funny online</td>
<td><a href="http://www.videosur">http://www.videosur</a> f.com/</td>
<td>12</td>
<td>video+videos video+funny video+online</td>
</tr>
<tr>
<td>videos funny online</td>
<td><a href="http://www.youtube.com/">http://www.youtube.com/</a></td>
<td>9</td>
<td>video+videos video+funny video+online</td>
</tr>
<tr>
<td>videos funny online</td>
<td><a href="http://video.google.com/">http://video.google.com/</a></td>
<td>7</td>
<td>video+videos video+funny video+online</td>
</tr>
</tbody>
</table>
4.3.3 User Query : “Art” Related Results,

<table>
<thead>
<tr>
<th>Suggested Term</th>
<th>Url</th>
<th>Relevance Factor</th>
<th>Fire Query</th>
</tr>
</thead>
<tbody>
<tr>
<td>gallery cultural artist</td>
<td><a href="http://www.searchgov.com/">http://www.searchgov.com/</a></td>
<td>6</td>
<td>art+gallery art+cultural art+artist</td>
</tr>
<tr>
<td>gallery cultural artist</td>
<td><a href="http://search.nic.in/">http://search.nic.in/</a></td>
<td>4</td>
<td>art+gallery art+cultural art+artist</td>
</tr>
<tr>
<td>gallery cultural artist</td>
<td><a href="http://www.artcyclopedia.com/">http://www.artcyclopedia.com/</a></td>
<td>3</td>
<td>art+gallery art+cultural art+artist</td>
</tr>
</tbody>
</table>

We are especially thankful to Google Search Engine from the bottom of our heart for returning useful reference pages, articles and free eBooks.

We shall be failing in our duties, if we don’t express our indebtedness to blessing of our parents.

REFERENCES


5. CONCLUSION

Query expansion is a very important task in the search process. We have presented the results of the web query expansion approach. Our approach reduces the human efforts to maintain the large dictionaries, which is a very laborious work. In our approach the user need not to store these types of dictionaries and the maintenance of the same. We have applied this approach to the vertical Meta search engine where the results shows that the user query is expanded appropriately and suitable search engine is selected every time. In the future scope user feedback can be used to improve the search engine selection index, where the frequency of the representative terms which are stored in the search engine selection index, can be increased on the basis of the user feedback to improve the relevance of the vertical search engine.

ACKNOWLEDGEMENT

First and foremost, we thank Lord Almighty for the grace, strength and hope to make our endeavor a success.


BIOGRAPHIES

Sandeep Joshi is M. Tech. in Computer Science & Engineering and having vast experience in teaching.

Satpal Singh Kushwaha is pursuing his M.Tech from Rajasthan Technical University, Kota(Rajasthan).