Tamil Search Engine

Baskaran Sankaran
AU-KBC Research Centre, MIT campus of Anna University,
Chromepet, Chennai - 600 044. India.
E-mail: baskaran@au-kbc.org

Abstract

The Internet marks the era of Information revolution. The Internet has been largely dominated by English till recently. The importance of reaching out to non-English speakers around the globe has been felt increasingly and this has lead to the spread of other languages on the Internet. Tamil is the fastest growing language in Internet among the Indian languages. With the number of Tamil websites crossing the two thousand mark, the amount of information available in these websites grows exponentially with the time. Searching for the required Information in the Tamil websites become increasingly difficult if not impossible. Search directories and Search engines help to locate the required information from the Internet. Search directory is a manually created database where websites are categorized according to their subject content. Categorizing websites is a time consuming and a laborious task and therefore these search directories often lag behind what is currently available on the Internet. Search Engines circumvent the need of manual categorization by crawling the entire web by traversing through the hyperlinks. This paper presents the details about developing a highly efficient, full-fledged Tamil Search Engine. This search engine is first such attempt for Tamil or for any Indian language, though search directories are available for many Indian languages.

1. Introduction

After many revolutions that took the world by storm, the history is now witnessing the Information revolution, which is more intense that the previous ones. Ever increasing computer usage and the high growth of the Internet are the important reasons behind the Information revolution. This has resulted in a great amount of information available over the web, which can be easily accessed by people. As the Web is flooded with contents created by varied sources, it is unreasonable to expect that all the information will be relevant to a particular subject and pertinent to the needs of millions of web users. With the continuous growth of the web, the proliferation of the information will pose a great problem for web users unless some solution is found to extract the material from the web, which is relevant to their needs. This problem has been addressed successfully by the development of search engines.

Further expansion of web to other regional languages creates the same problem of quickly finding relevant information for the information seekers in these languages. Tamil is the fastest growing language in the Internet among Indian languages and as a result it is increasingly becoming difficult to locate required information from these Tamil websites. Thus it became necessary to develop a Tamil search engine which can look into Tamil web pages and retrieve relevant pages for the user.

This paper explains in detail the technology behind the Tamil search engine and the features of the search engine. The next section explains the general architecture of a search engine. Section 3 explains the general features of the Tamil search engine. The features that are specific to the Tamil search engine are discussed in section 4. Section 5 chalks out the future extensions and this is followed by the conclusion in section 6.

2. Architecture of a Search Engine

A *search engine* is software that searches for documents dealing with a specific topic in the Internet. The basic architecture of a search engine is shown in fig.1. It consists of two parts, viz. a back-end database and a front-end graphical user interface (GUI), to facilitate the user to type the search term.

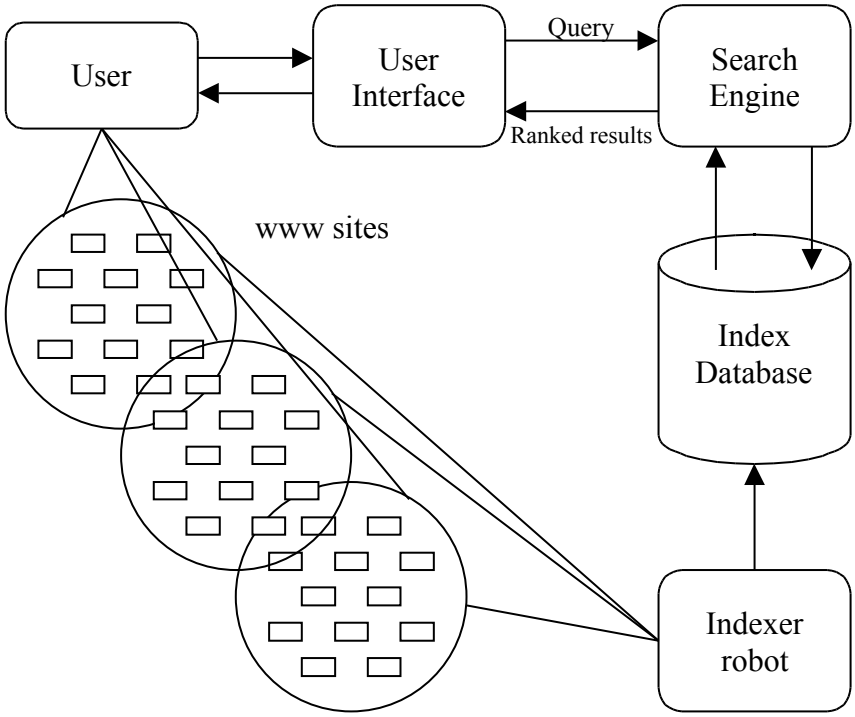

Figure 1 - Architecture of a Search Engine

On the server side, the process involves creation of a database and its periodic updating done by software called *spider*. The spider also called as *crawler*, *robot* or *indexer*, crawls the webpages periodically and indexes these crawled webpages in the database. Before indexing the keywords, these keywords are morphologically analysed to remove the inflections suffixed to the word and only the roots of the words are stored in the database. The hyperlinked nature of the Internet makes it possible for the spider to traverse the web. The front-end of the search engine is the client side having a graphical user interface, which prompts the user to type in the search query. The interface between the client and server side consists of matching the user query with the entries in the database and retrieving the matched webpages to the user's machine.

As stated earlier, the spider crawls the web pages through the hyperlinks. In this

process it extracts the 'title', 'keywords', and any other related information needed for providing complete search results from the HTML document. Sometimes, the entire content of the HTML document except for the stop words[1] is extracted and indexed in the database. The idea of indexing the whole page except the stop words is based on the fact that a page dealing with a particular issue will have relevant words throughout its page. Thus indexing all the words in a document increases the probability of getting the relevant webpages to a query. One point is worth noting here: before the query words are processed they are removed of the morphological inflections before they are searched for in the database.

The database consists of a number of tables that are arranged so as to facilitate faster retrieval of the data. This database is housed in a database server, which is connected to the search engine. The typical English search engines will have more than one database server due to the huge number of English websites. The Tamil search engine uses a single database server, because of the small number of Tamil websites.

When the user types in the query it is taken to the server housing the search engine. The search engine translates this query into the structured query language (SQL) which is understandable to the database and passes this SQL query to the database server. The database server identifies the database entries that match with the query given and sends to the search engine server these entries along with other information related to these entries such as the URL and the matching portion from the content of the corresponding entry. The search engine sorts these database entries using a ranking algorithm. The ranking algorithm determines the relevancy of a retrieved webpage to the user query. The retrieved sites are then displayed along with links to these sites and a small portion of text from the matched content. This text gives an idea to the user about the page before the user goes to that particular page.

3. General Features of the Tamil Search Engine

3.1. Search options

Search options allows the user to search for various combinations of the query terms. Some of the search options include Boolean search and phrasal search.

Boolean Search

Several options should be available to the user to refine the query. This is important because the search should return only the relevant pages to the user. Boolean search option includes OR, AND and NOT logic, the default being AND followed by OR. Boolean search can be illustrated by the following example.

Consider a query containing two words. The search results for the OR logic will retrieve the pages containing either of the two terms and the search results for the AND logic retrieve the pages containing both the query terms. The NOT search returns the webpages that does not contain the NOT term. To search for query A and query B, the user should type *A +B* in the query field. The third option NOT typed as *!A*, returns the pages that does not contain the term A. The example illustrates the Boolean search for two terms in the query and it is straightaway to extend this to more terms.

[1] Frequently occurring words not having any topical specification are called stop words. Examples includes words such as a, the, an, is, was, for, and, or etc. These are also referred to as functional or common words.

The current implementation of the search engine combines the AND & OR search options effectively and has no provision for NOT search. The search engine automatically searches for both the AND & OR logic. The results of AND search is displayed at the beginning followed by the results of the OR search.

Phrase search

Phrasal search looks for a phrase instead of a word in the database. To include phrase search in the query the user should type the phrase between two quotes. The corresponding phrase will be searched *as is* in the database. This option is particularly useful if the user knows a phrase in the domain of his search. However, this option requires huge processing power and bigger memory in the database.

3.2. Ranking

Ranking is done primarily to provide highly relevant documents to the user at the first place followed by the documents that are not so relevant. Ranking algorithms rank the retrieved webpages to determine the relevancy of these pages with the query terms. Webpages are ranked based on multiple factors ranging from simple factor – the presence of any of the query terms in the page, to computationally complex criterion – the proximity of the query terms with each other. The factors determining the rank of a webpage are listed below in the decreasing priority order.

- the proximity of the query terms with each other, when the query contains more than one term
- frequency of the query term (term frequency) in the page normalised by the total number of pages having the query term (document frequency) in the database
- the presence of keywords i) at the beginning of the text ii) in bold or italics
- the presence of any of the query terms in the content

The efficiency of the ranking process can be improved by using the *hypertext vector method* of ranking where a site is ranked based on the number of other sites having a link that points to the current site. This algorithm gives higher rank to a particular site if many other sites refer it. Such an algorithm is complex and needs high computing power.

4. Specifics of the Tamil Search Engine

This section describes the features of the Tamil search engine that are different from the existing English search engines.

Morphological analysis vs. Word stemming

Tamil is inflectionally a rich language and because of this the query term may be inflected sometimes. The inflections attached to the root word, as suffixes should be removed so as to retrieve all the relevant webpages. The *word stemming* module identifies the word stems (root words) by chopping off the inflections in an ordered way. Majority of English search engines removes the standard inflections by using a dictionary and the process is simply referred as word stemming.

Tamil search engine uses a morphological analyser[2] for this purpose. Morphological analysis is a complex process than the simple word stemming. Morphological analysis is preferred here because of i) the inflectional nature of Tamil and ii) its high accuracy to simple word stemming. The accuracy of the morphological analyser used here is around 95%, which is much higher than any other morphological analyser for Tamil.

User Interface

Two types of user interface have been provided in this Tamil search engine, keeping in mind two different types of users. This allows the user to choose the suitable mode of typing the query.

- Tamil keyboard – This feature is provided to users who know to type in Tamil. The current version of the search engine supports the TamilNet99 keyboard and it is further planned to provide support to TamilNet97, Mylai, Old and New typewriter layouts.
- Phonetic keyboard – This feature is provided for those who knows English and are not familiar with any of the Tamil keyboards. On choosing this option, the user has to type the query in transliterated English. A client side script will automatically transliterate this to Tamil and will display it in the screen.

Encoding Schemes

A major stumbling block for Tamil to be widely used in computers is the lack of a standard and *used-by-all* encoding scheme. Though TamilNadu government has standardised TAB and TAM schemes, majority of the sites uses different schemes with the most popular being the TSCII standard. The Tamil search engine supports the three most important and widely used encoding schemes, viz. TAB, TAM and TSCII.

5. Future Extensions

- The English to Tamil machine translation system can be added to the search engine so that the query can also be searched in English webpages. The search results will be displayed after being translated to Tamil. Such translation services are already available for European and some East Asian languages in altavista, google etc.
- The efficiency of the ranking algorithm can be improved by using the *hypertext vector* method in addition to the existing criterion for ranking.
- The retrieval efficiency of the search engine can be improved by looking for the synonym of the query term, in addition to the query term itself. The synonym can be obtained using either a dictionary or a thesaurus.
- The meta-data information in HTML pages can be used to index the database effectively. For this, the content developers should use the meta-data in their webpages.

6. Conclusion

The need for a Tamil search engine is necessitated by the enormous growth of the

[2] Morphological analyser splits a word into its constituent morphemes by identifying and cutting the inflections in an iterative and ordered fashion

Tamil content in the web. In the first place, general architecture of the search engine was discussed, followed by the features such as query options and ranking. The issues that are specific to Tamil search engines such as morphological analysis, user interface, encoding schemes etc. are also elaborated. Features that can be added to improve the efficiency of the search engine were also listed.

Acknowledgments

This project was partly funded by ELCOT, Chennai – an agency of TamilNadu government. The author appreciates the duo K Swarna and P Divya for their help in implementing the ranking algorithm. The author is grateful to S V Ramanan and Vijay K Shanker for their inestimable ideas. The author would like to thank the NLP team for their constant encouragement and suggestions.

References

1. Baeza-Yates Ricardo, Ribeiro-Neto Berthier, 1999, *Modern Information Retrieval*. Addison-Wesley, USA.

2. Belew K. Richard, 2000, *Finding Out About – A Cognitive Perspective on Search Engine Technology and the WWW*. Cambridge University Press, Cambridge, UK.

3. Jurafsky Daniel, Martin H. James, 2000, *Speech and Language Processing - An introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. Prentice Hall, NJ, USA.

4. Korfhage R. Robert, 1997, *Information Storage and Retrieval*. John Wiley & Sons, NY, USA

5. Manning D. Christopher, Schutze Hinrich, 1999, *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, MA, UK.

6. http://www.searchenginewatch.com