OmSE: Tamil aearch Engine
- using open source software Greenstone

Anandh Jayaraman[1] , Srinivas Sangani [1] , Madhavi Ganapathiraju [2] ,
N. Balakrishnan[1] and Raj Reddy[2]
[1]Indian Institute of Science, Bangalore, 560012, India; [2] Carnegie Mellon University, Pittsburgh, 15213, USA
<jayaraman.anandh@accenture.com, srini@serc.iisc.ernet.in, madhavi+@cs.cmu.edu, balki@serc.iisc.ernet.in, rr+@cmu.edu>

Abstract

There are more than 3000 websites with content in Tamil and their number is growing.. All the websites in the Indian languages are not searched or indexed by any of the major search engines like Google ,Yahoo etc. This content is presently untapped for the users looking for content in Tamil. In this paper we present an implementation of Tamil Search Engine using open source software Greenstone for indexing the content collected by a specially written crawler which crawls all the major Tamil portals across the web and use freely available Om transliteration package which uses case insensitive mapping of characters for parsing. The prototype with a full text search on Tamil web content is available at
http://revati.dli.ernet.in/SearchTamil.html

By using a Transliteration scheme called OM, developed jointly by the Indian Institute of Science and the Carnegie Mellon University, we have been able to separate the storage from rendering. The storage is in ASCII and reflects the phonetic and is language independent. The rendering is in the language chosen by the viewer. This has made the search engine that has been described here applicable all Indian Languages and not limited to Tamil. Unicaode to ITRANS mapping is also available.

1. Introduction

The Internet revolution is taking place at an astonishing speed. In the last decade itself the number of websites has grown from 21,000 in 1994 to 39,363,493 in July, 2004. This has resulted in a great amount of information available over the web, which can be easily accessed by people. More than 29% of Internet users worldwide use search engines to find information---for most of these users, “if it is not retrievable by a search engine, it is not there at all”. Most of today’s text content is born digital. For languages like English, where a large portion of the target audience is likely to have computer and internet access, the content is not only created digital but is also ready for digital delivery, access and use. However, for Indian languages, content is not created for a digital delivery; because of this, formats and tools for Tamil Internet 2004, Singapore 2 by word of mouth, and in palm leaves. In order not to miss out in a similar way in the time of Digital Revolution, Indian Institute of Science, in partnership with Carnegie Mellon University in the US, and a host of other institutes in India, has undertaken the creation of Digital Library of India. A large number of printed books that are out of copyright or are deposited by the authors and publishers are scanned and converted

to digital form. The goal of this project is also to create a test bed for research in language technologies for Indian languages.

This paper explains in detail the technology behind the Tamil search engine developed from using open source software called Greenstone, and the features of this search engine.

2. Open Source Software Suite - Greenstone

Greenstone is a open source suite of software for building and distributing digital library collections. It provides a new way of organizing information and publishing it on the Internet. The Greenstone software also supports the indexing of each word in the full document text, as well as providing indexing on document metadata (if present). It has a search engine (Boolean) and allow to quickly find what one seeks

3. Om Transliteration

Om is a transliteration scheme for typing Indian languages using the standard keyboard. It has been designed with phonetic mappings such that it is easy to remember. Om transliteration is largely based on mapping scheme developed for Indian Language Transliteration (ITRANS) package. ITRANS is a carefully designed scheme that has been in use for many years now. Om mapping is meant to add many more features to enhance the usability and readability, and has been designed on the following principles: (i) easy readability (ii) case-insensitive mapping: while preserving readability, this feature allows the use of standard natural language processing tools for parsing and information retrieval to be directly applied to the Indian language Texts and (iii) phonetic mapping, as much as possible. An important feature is that the case insensitive phonetic mapping is also highly readable by itself in the English script. This is of particular importance since often people can only speak fluently and understand the language but cannot read the script. India being a multi-lingual country, and inter-mixed population, often the people can speak and understand more than one Indian language and also English. Hence even in the absence of Om to native font converters, people around the globe can type and publish texts in Om scheme which can be read and understood by many even when they cannot read native script.

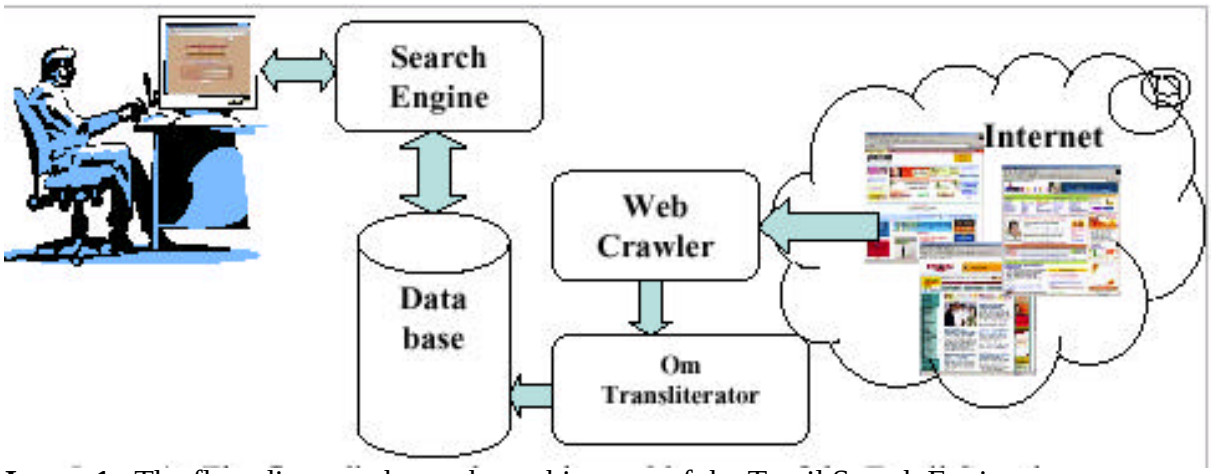
Om is a freely available Indian language transliteration package for multiple Indian languages (http://swati.dli.ernet.in/om/). The advantage of using Om is that it uses case-insensitive mapping; while preserving readability, this feature allows the use of standard natural language processing tools for parsing and information retrieval to be directly applied to the Indian language texts.

4. Design of Tamil Search Engine

4.1 Architecture

The basic architecture of the Tamil search engine has a server which contains a database, a web crawler which crawls and downloads the Tamil Language web content from various Tamil Web portals and the Om transliterator which converts from ASCII to Om transliteration

format .



Insert 1 . The flow diagram shows the architecture of the Tamil Search Engine

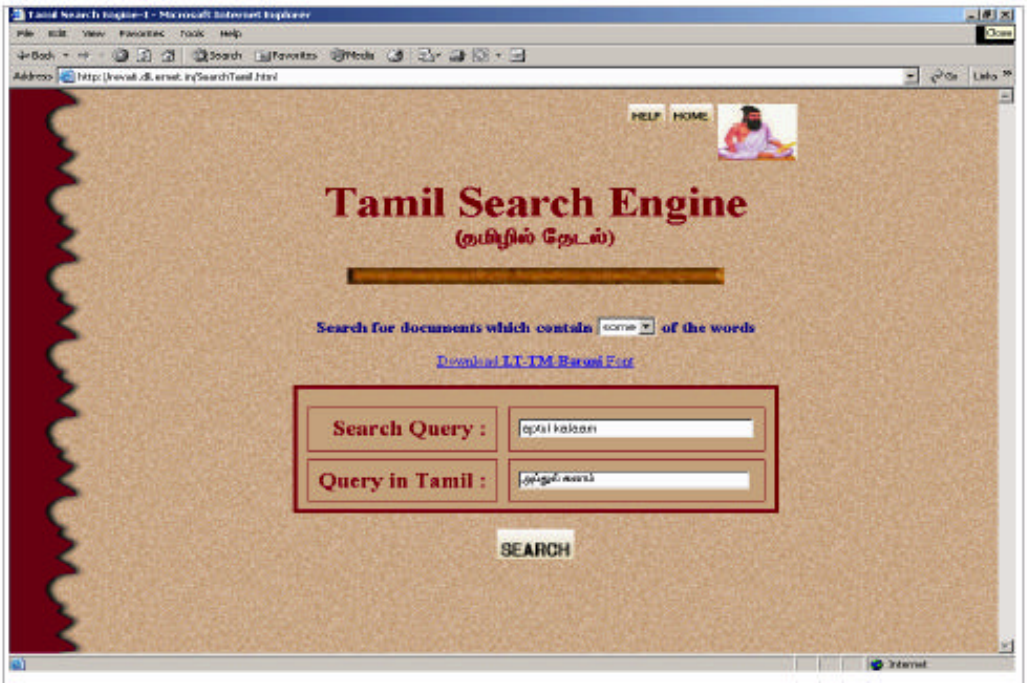
On the server a database of Tamil language WebPages is created. This is done by a web crawler program written specially for this purpose. The web crawler crawls the Tamil web portals like www.dailythanthi.com, www.webulagam.com, etc, periodically. The web pages collected from these portals are converted from the native font representation to Om transliteration format and are stored in the database. The converted web pages are stored in plain text format.

Using the Open Source software Greenstone, the collected files are indexed by first importing the text files into Greenstone's database; the primary task here is to convert documents from their native format into the Greenstone Archive Format used within Greenstone, and write a summary file which will be used when the collection is built.

During the building process the text is compressed, and the full-text indexes that are specified in the database configuration file are created. Furthermore, information about how the Internet database contents are to appear on the web, for example information about icons and titles, and information produced by classifiers are precalculated and incorporated into the databases.

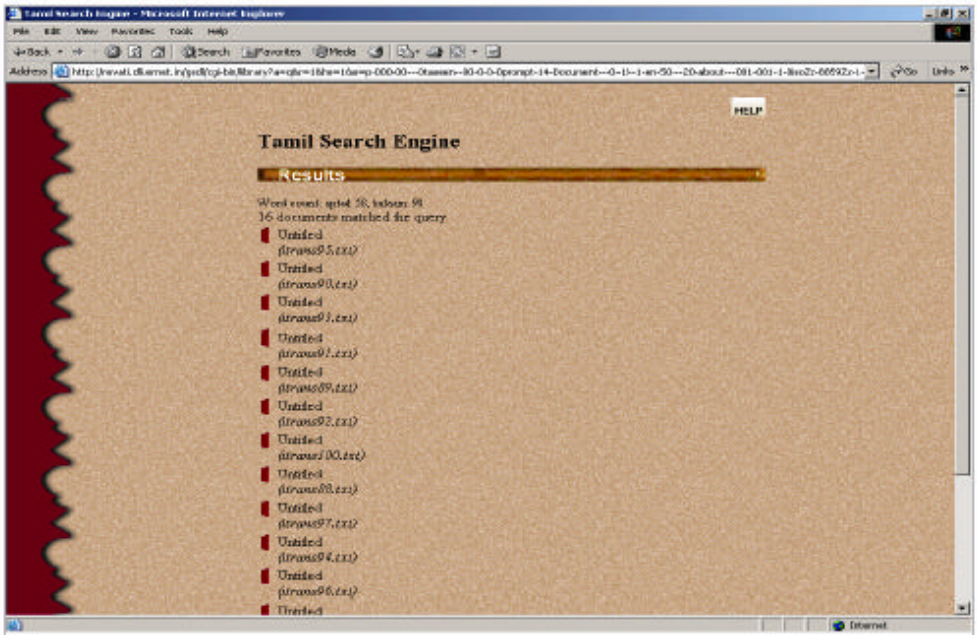
The front-end of the search engine is the client side having a graphical user interface, which prompts the user to type in the search query in Om transliteration format. The query typed by the user is also displayed in Tamil font for the user to make corrections if required while entering the keyword in Om Transliteration format.

E.g. the keyword *Abdul Kalam*, in Om format it is keyed in as *Aptul kalaam*.



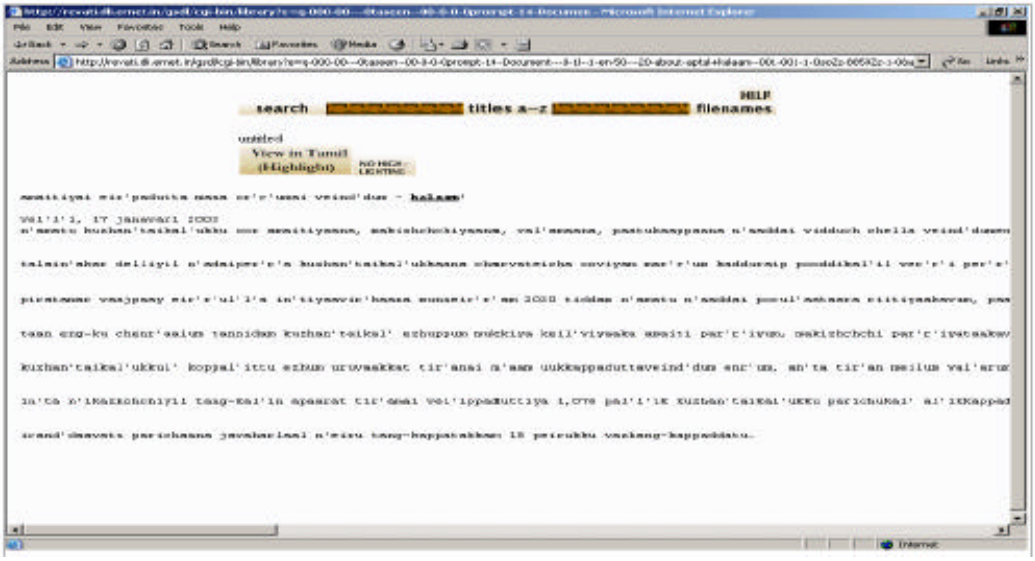
Insert 2. **User Interface for Tamil Search Engine.** The search query is entered in Om transliteration format , the exact transliteration of the search query appears in Tamil. The search can be performed for either "some" of the words or "all" the words

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. The search engine takes the query to the database and looks at the matches as per ranking. The search engine then sorts these database entries using a ranking algorithm. The ranking algorithm of Greenstone determines the relevancy of a retrieved webpage to the user query. The retrieved sites are then displayed along with links to these sites in text format.

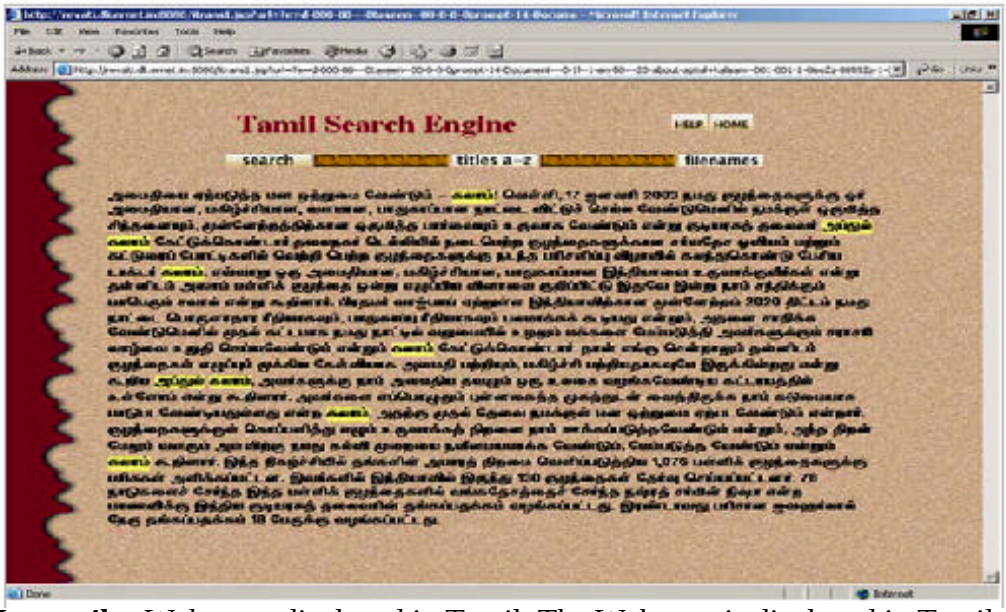


Insert . 3. Results of the Search for the given keyword. The results for the given keyword are displayed with number of hits and the number of the instances of keyword found

The resultant web pages are then displayed to the user in both Om transliteration in English as shown in **Insert 4a** and in Tamil font, with the keywords highlighted as shown in **Insert 4b.**



Insert 4a:. Web page displayed in Om transliterated format.



Insert 4b:. Web page displayed in Tamil. The Webpage is displayed in Tamil using Om transliterated text and the search keywords are highlighted.

4.2 User Interface

The user interface provided in this Tamil search engine, has a Search box in which the query to be searched has to be entered in Om transliteration format. The transliterated output is displayed in Tamil for users help.

Search terms

Query should be keyed in Om transliteration format. Whatever you type into the query box is interpreted as a list of words called "search terms". Each term contains nothing but alphabetic characters and digits. Terms are separated by white space. If any other characters such as punctuation appear, then they will be treated as phrases that is search term covered by double quote. If characters like "(" appears then that will be replaced by white space. For example, the query

janaatipati (aptul kalaam) kujaraat chenr'aar

will be treated the same as

"janaatipati aptul kalaam kujaraat chenr'aar"

Query type

There are two different kinds of query.
Queries for all of the words. These look for documents (or chapters, or titles) that contain all the words you have specified. Documents that satisfy the query are displayed, in alphabetical order. Queries for some of the words. Just list some terms that are likely to appear in the documents you are looking for. Documents are displayed in order of how closely they match the query. When determining the degree of match, the more search terms a document contains, the closer it matches. The rare terms are more important than common ones and short documents match better than long ones.

Use as many search terms as you like--a whole sentence, or even a whole paragraph. If you specify only one term, documents will be ordered by its frequency of occurrence.

5. Conclusion and Future Work

We have made a prototype of the Tamil Search Engine using a standard search engine called Greenstone, by integrating and interfacing to it the Indian language processing through Om Transliteration; web content from three Tamil news portals namely www.dailythanthi.com, www.webulagam.com and www.dinamalar.com is deposited into the database by the crawler. The Tamil search engine is hosted at http://revati.dli.ernet.in/SearchTamil.html . The mapping scheme used by Om transliteration is consistent across the many Indian languages. Thus, the same text is transliterated into any of the other Indian languages.

Acknowledgments

The search engine is a part of the integrated efforts towards developing language technologies for Indian languages. This work has benefited from the many resources developed by the entire team of the Digital Library at the Indian Institute of Science. Especially, we would like to acknowledge the contributions of Mini Balakrishnan and Sravan Kumar towards the development of the Om transliteration tools.

References

1. Ian H. Witten, 2001, How to build a Digital Library using Open-source Software, 4th International Conference of Asian Digital Libraries, ICADL 2001 Tutorials, Bangalore,

India
2. Baskaran Sankaran, 2002, Tamil Search Engine, Tamil Internet 2002, California, USA
3. http://www.searchenginewatch.com
4. http://www.sims.berkeley.edu/research/projects/how-much-info-2003/internet.htm
5. http://www.zooknic.com/Domains/counts.html