Problems Encountered in Searching in Tamil on the Internet : A Suggestion for a Versatile Search Engine

Mr.Rangarajan (Sujatha) & Miss. N. Jayaradha
Dishnet DSL Ltd , Chennai- 600034

ABSTRACT

This paper describes the need for a Tamil search engine, which will search across the web and a simple architecture for a search engine, which can search for Tamil text documents irrespective of encoding. It also states the various problems encountered while designing such an engine.

INTRODUCTION

Search engines are the most important means of navigating the web. The most popular search engines like google are supporting search in different languages partially (Including Tamil). These types of search engines are indexing the whole web document and throwing results not only in Tamil but also in Korean and French languages. This is because of explosive growth of dynamic web information, it is necessary to develop systems that can efficiently search, index and retrieve information in Tamil and English.

OVERVIEW

Individual search engines exchange information about their own local indexes these individual search techniques can be implemented with the help of simple Microsoft index server technology. (Ref: http://search.ambalam.com). Some search engines like yahoo, google etc can search their own index and these can also refer queries to multiple search engines.

Generally search engines are constructed by making use of programs called robots. These robots use depth first search or breadth first search algorithms to index the document. While indexing it is necessary for the Tamil search engine to index the Tamil document separately, so that the robot does language detection (like INKTOMI Japanese search engine) to segregate only Tamil documents. This language detection can be done by tag extraction. where the document publisher should mention the character set and the language they are using in META tag. The documents and web addresses should be collected and sent to the Search engine's indexing software. The indexing software extracts information from the documents, storing it in a database. The kind of information indexed depends on the particular search engine. Some index every word in a document and some index title only. When a search is performed by entering a keyword, the database is searched for a document that matches the web pages that result as hypertext.

In this type of engines the number of resulting matches thrown by the search is very large and mostly irrelevant. For example if you are searching for thirukkural in Tamil you can get nearly 7,00,000 matches. And most of the matches are irrelevant. The matched links have contents "thi", "thirumurai", "thirupur" and such irrelevant links. So it is necessary to select

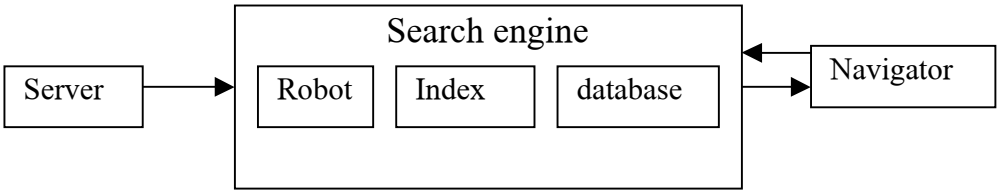
the keyword from a document by comparing most commonly occurring words for each and every document. This can be done with the help of canonical form generation.

Since the Tamil web pages support various encoding and different fonts it is very difficult to combine everything into a single search engine. If we are taking TAB or TSCII as standard (since they are bilingual and they support dynamic fonts) all the key words in documents should be converted to one of the standard encoding during indexing. The encoding conversion should be only for keyword matching and throwing of results. After going inside the page the searcher should have the required fonts to view the sites.

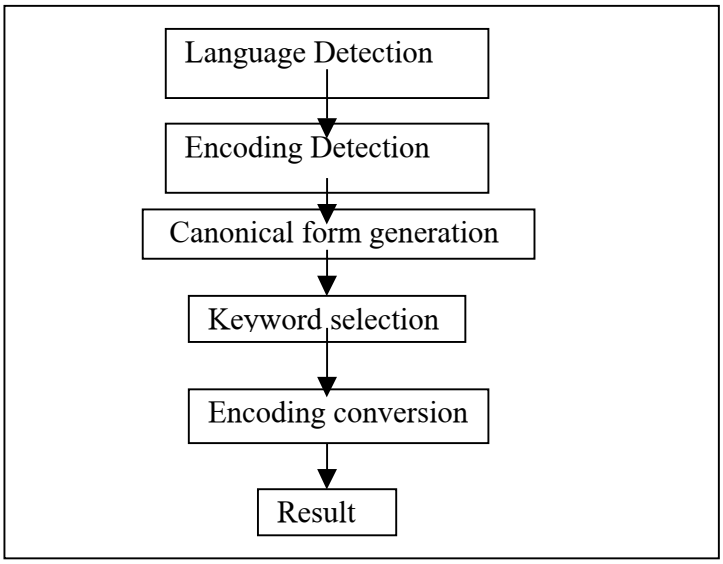
At present the source encoding is frequently TAM, ISCII, Anjal, TAB and TSCII and the target should be TAB or TSCII, which are accepted encoding of Tamilnadu and Indian governments.

This encoding conversion and canonical form generation and the keyword selection can be done with the help of algorithms like MURASU and PADAMI.

Simple architecture of Multiple encoding search engine



Search engine:



Architecture:

Document publishing plays an important role in this search engine. The document publisher should specify the language and the encoding used in the document. So that the indexing can be based on that. Even if the document is in different character encoding the search result should be in single encoding

So the search engine should work in the following sequence

1. Document Detection
2. Language Detection
3. Encoding Detection
4. Canonical form generation
5. Keyword selection
6. Encoding Conversion

Document detection:

The Document detection phase should be capable of parsing and selecting the keyword for the document with arbitrary contents in various file formats. Sequence of operations involved in this phase is as follows.

1. Document type detection
2. Extraction of terms and tags
3. Keyword selection

The document type detection involves selecting .html, .asp, .htm, .txt, .doc files. Most of the robots won't accept the .js files, java files, automatically generated dynamic web pages and some private documents.

Language Detection:

It is necessary to understand the language and character encoding. The language and character type must be done during tag extraction. It is however often difficult to, extract useful information from the document. So it needs well-formatted Meta tag such as date, author, format, and language information. During tag extraction the Meta tags which specify language and character set should be detected.

Encoding Deduction:

The encoding detection should be done with the help of separate encoding detection software. The most commonly used encoding are TAB, TAM, TSCII, Anjal, ISCII. Some Tamil sites use their own encoding .If possible they can also be included in search.

Canonical form generation:

Canonical form is defined as the simplest form a term can take in a document. For example the term routing, router, routable can be considered as single term route

Keyword generation:

The keyword generation is achieved by comparing frequently occurring words in the web document.

Encoding conversion:

Encoding conversion can be done before canonical form generation and keyword generation so as to index the search pages in single encoding.

Difficulties in implementation:

The language detection phase is considered as an important phase to define a scope. It mainly depends on Meta tags and tag extraction principles. In most of the web documents the Meta tag is automatically generated. And the document publishers are publishing the documents without language information. The Meta information needs some standardization.

For example
Title: Title of the page
Keyword: term1,term2, term3, ...
Mime-version:1.0
Content . type : text/plain ; character set = iso-2022-jp
Language = Japanese English
Like Japanese search engine.

The TAM to TAB conversion will give problems in some characters like ni, no etc.. this is mainly because of monolingual to bilingual conversion. Likewise same the TSCII to TAB and TAB to TSCII are also having some problems.

Conclusion:

Tamil Internet community should form some rules and strictly follow the rules for search engines particularly for Meta tags used in web pages and also encoding restrictions to utilize the web efficiently.

Reference:

www.Inktomi.com - Japanese search engine
www.isoc.org: chinese search engine
www.ultraseek.com: Architecture of robot, spider, and crawler
www.lub.lu.se
www.dlib.org: A multilingual federated search engine