TAMIL SEARCH ENGINE FOR UNICODE STANDARD

A PROJECT REPORT PRESENTED BY

CASSIM MOHAMED MOHAMED MANSOOR

to the Board of Study in Statistics & Computer Science of the

POSTGRADUATE INSTITUTE OF SCIENCE

in partial fulfilment of the requirement

for the award of the degree of

MASTER OF SCIENCE IN COMPUTER SCIENCE

of the

UNIVERSITY OF PERADENIYA

SRI LANKA

2009

TAMIL SEARCH ENGINE FOR UNICODE STANDARD

C. M. M. Mansoor
Post Graduate Institute of Science
University of Peradeniya
Peradeniya
Sri Lanka

ABSTRACT

The web creates new challenges for information retrieval. The amount of information on the web, as well as the number of new users is growing rapidly. Search engine technology has to scale dramatically to keep up with the grow of the web. Web search engines have emerged as one of the central applications on the Internet. In fact, search has become one of the most important activities that people perform engage on the Internet. The internet has become number one source of information. A growing number of businesses are depending on web search engines for customer acquisition.

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. A search engine capable of searching the web documents written in languages other than English is highly needed, especially when more and more sites are coming up with localized content in multiple languages.

Tamil is one of the fastest growing languages in the 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.

Further expansion of web to other regional languages creates the same problem of quickly finding the relevant information for information seekers in these languages. Since the retrieval of information written in Tamil is very difficult it is necessary to develop a Tamil search engine which can look into Tamil web pages and retrieve relevant pages for the user.

Each language has some encoding. For Tamil language, there are more than 50 encoding standards including TAM, TAB, TSCII, or ISCII as their encoding scheme. However due to lack of usable standards in the past, authors of web sites proceeded to create their own font encoding. This created a proliferation of incompatible encodings on the net. As a result, Net users are forced to download fonts for each web site separately.

It is very difficult to search their relative information using different search engines. This report proposes a Search engine for Tamil language based on Unicode encoding scheme. Unicode is a universal character encoding scheme designed for cover all world languages. Unicode is two bytes encoding to cover all of the world's common writing systems. When the Unicode is used, it is not necessary to consider the platform, the language and the program used.

The proposed Search Engine software allows full-text indexing and searching of a database of documents written in any Unicode encoding Tamil Language. Like any other Search engine, it consists of a crawler, Search techniques and Data Base System to store the information. The engine developed was tested for its validity. The test results indicated that it facilitate the searching of information written on Tamil quickly and consistently.