Encoding Detection in Internet Search Engines

Sivaraj D <sivaraj@intelligroup.com>

Tamil content in the internet uses a wide variety of character encodings which are mutually incompatible. Thus developing a search engine for Tamil content faces serious challenges, in (1) identifying the Tamil content on the internet, (2) identifying the encoding used in such web sites, and (3) presenting the result in in a format that user can use.

This paper explains a method by which automated search engines can identify Tamil content on the net and encoding used in that content. It uses an algorithm to match predetermined patterns. It also explains methods by which such content can be stored and presented to the user. This algorithm does not have any presumptions about the design of the web site. Any current Tamil web site on the net, which has sufficient sample Tamil content can be identified. However it can be used only to determine encodings that are previously known to the software.

Introduction

Tamil is leading the internet usage among Indian languages, with a good number of quality web sites. 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. Net users are forced to download fonts for each web site separately.

There were free fonts developed by Tamil computing enthusiasts or software developers, which encouraged users to be use them instead of creating own fonts. Examples are Mylai, Murasu Anjal, and Tamilnet. Many amateur web authors started building web sites using these fonts.

Later many of these enthusiasts and software developers joined together to create a singular standard encoding, namely TSCII. During the Tamilnet'99 conference in Chennai, then Tamilnadu Government also announced two standard encodings, namely TAB and TAM. TAB is a bilingual encoding and TAM is monolingual. TSCII and TAM seem to have considerable usage on the net.

Unfortunately these standards do not seem to have many converts among the popular Tamil websites. Todate, none of the popular print magazines present on the Net use one of these standard encodings.

While it is important for all web sites on the net to use the same encoding, that seems to be a difficult task to accomplish in the short term. Future popularity of Unicode and easy availability of quality Unicode software might force the web sites move towards it, but in the interim, the situation seems bleak to make the web sites use a standard encoding.

So we need to search for alternatives to catalog web sites and to implement various web based technologies. This paper discusses a method by which these web sites with different kinds of encodings can be searched and indexed by a search engine.

Method of identifying Tamil content

To index the web pages, a search engine should first identify which web sites have Tamil content, and the encoding used in each of these web sites. Once the web sites and encodings are identified, the search engine should translate those web sites into a standard internal encoding, preferably Unicode. Optionally the search engine can also have a display encoding that can be set in the user's preferences.

To identify Tamil content and the encoding, we need to proceed in several steps. We will first follow the steps that will tell us explicitly the encoding used in the web page. If the particular web page cannot be identified by this means, we will use statistical sampling to determine if the web site might fall one of the known encodings.

Step 1: Explicit information in the document tags

Web pages may specify their natural language using HTML META tag. HTTP header Content-Language defines the natural language of the content. This can be specified in a web page using the META tag attribute HTTP-EQUIV. In practice, not many web sites use this construct.

The META tag Content-Encoding [FIXME] specifies the encoding used in the web page. For web pages using standardized encodings or de facto standards this tag may be specified. For example, web pages using TSCII encoding might specify this tag as "x-tscii". However specifying this tag with an encoding unknown to the web browser usually ends up with the page not being visible in that web page. So web authors usually specify this tag as "x-user-defined", to instruct the web page that this is a custom encoding.

Step 2: Looking for FONT FACE tag

Many Tamil language web sites use FONT FACE tag to display text in a Tamil language font that mimics the behavior of ISO-8859-1 (Latin-1) encoding, or as a user-defined encoding. By building a list of font names used on the net we can determine the encoding used in those web pages.

Often fonts of the same encoding with have different names. A list of font names and their respective encodings need to be maintained. These font names can be compared to the font name in the FONT FACE tag. Font in certain encodings may have a prefix or suffix to identify the encoding. For example, all the fonts in TSCII 1.7 encoding are supposed to start with TSC in the beginning of their names. This also can be used to identify the encoding used in the font.

Step 3: Matching popular patterns

Certain Tamil words occurs frequently in Tamil web pages, the word 'Tamil' itself being one of them. By searching for such words in the content of web pages in various encodings, it can determined whether they contain any Tamil content and the encoding of that content. On the net there are many fonts that have the same or slightly differing encoding (Traditional type writer based fonts on a English keyboard). These can be grouped into a handful of encodings using this method.

First many generic Tamil web sites need to be index for the words contained in those pages. Words in these indexes are then sorted by their number of occurances, and a list of words with most occurances is prepared.

This word list is converted to all the supported encodings, and stored as a list. These are the patterns that need to be searched in the pages that need to be identified.

In each page that need to be identified, search for the occurance of patterns in our list. Note that occurance of one of these patterns in the page does not guarantee that the page is in the encoding of the pattern. We need to count the occurances of different patterns with same encoding. Based on experiments with various web pages, we can determine the probability of that page in the given encoding. For example, to be sure that we find the correct encoding, we might need to make sure that at least 3 words in our pattern list appear on the web page 3 times each. This logic can be fine tuned and more word patterns added as more web pages are indexed. Each word can also be given probability weightage. Occurance of Words that have more weightage will increase the probability of the web page being in that encoding.

Step 4: Storing in a common encoding

Once the content language and encoding are determined, the result can be converted into a standard encoding (say, Unicode) and stored in a cache database (as in Google).

Further the search engine may also allow the users to store their favorite encoding, and can have a convert option, which will convert the page into the encoding of their preference and display it.

Step 5: Feedback loop

A option can be included in the search engine result, which can check the validity of the result using user feedback. This can be used to further refine the search phrases in Step 3. We can ask the user whether the result provided is correct. If it is correct the search engine can harvest new words from the web page for its pattern list. If the result is wrong, then the words that we used to identify the page is suspicious, or we need decrease the probability weight for these words.

Conclusion

This search method is valid for all Indian languages and other languages that have the problem of proliferation of encodings. Once basic algorithm is developed it can be readily applied for other languages. If such technology is developed in a open manner it may readily be adopted by popular internet search engines.