5.5 Indian Language Search Engine

5.5.1

- **Name of the Technology/Product** : ***Malayalam Web based Search Engine (ANWESHANAM™)***
- **Nature of Technology** : Knowledge Tool
- **Level : Product**
- **Technical Description :** "Anweshanam" is a directory based search engine, which searches for Malayalam content and information in web pages. This solution provides a Malayalam interface that helps the user to search for information quickly and easily on the web. It searches for specific Malayalam keywords and generates a list of links to web pages containing the searched information. It searches for specific content in the pages on the server. This solution was developed using Java Server pages.

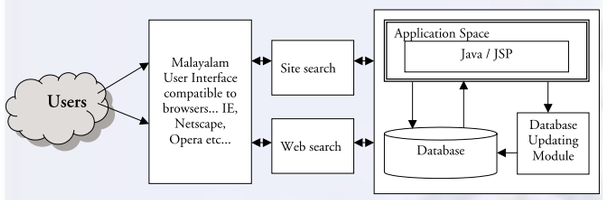


Figure 7.1 Block diagram of Search Engine

The interface is in Malayalam and includes easy input facilities with keyboard driver and floating character map. The keyboard driver supports the INSCRIPT keyboard with automatic formation of conjuncts. The application provides facilities for searching either English or Malayalam keywords. The result pages contain links to the web pages containing the searched information along with their brief description in Malayalam. A form is also provided whereby new pages and sites can be added to the database with details of the content.

- **Scalability/Portability/Expandability:** The application can be expanded to function as multilingual search engine and also as a meaning based search engine.
- **Readiness of Transfer of Technology (ToT):** Ready for TOT
- **Availability of documentation :** Documentation available.
- **Testing of the Product/Technology:** Development testing done.
- **Potential beneficiaries:** Beneficial to Multilingual portal developers and is useful for any one looking for some particular Malayalam content on the web.
- **User-agency tie-up:** Solicited
- **Name and address of the Resource Person:**
  R. Ravindra Kumar
  Additional Director & Coordinator
  RCILTS- Malayalam
  C-DAC, Thiruvananthapuram.
  Email: ravi@erdcitvm.org.
  Web : http://www.malayalamresourcentre.org

5.5.2

- **Name of the Technology** : ***Bavani (Tamil Search Engine)***
- **Nature of Technology** : Knowledge Tool
- **Level** : Product
- **Technical Description :** Tamil Search Engine is a software package that searches for documents dealing on a particular topic in the Internet. The Tamil Search Engine aims at looking up for information sought by a user in Tamil sites. It searches for Tamil words in popular font encoding schemes in Tamil web sites. An important feature of the search engine is the integration of the Tamil Morphological Analyser. With the help of Analyser, the keywords given by the user are reduced to the root form in search.

  The following are the important modules handled in the search engine:

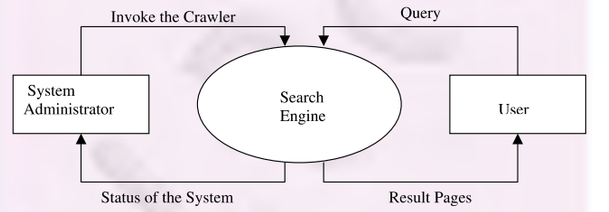
  Crawler : This is the kernel part of the design. Crawler is responsible for retrieving pages from the web and handing them to the database management module. A list of seed URLs is taken as its input and sequentially traces through all the links in the list. Each document is downloaded and then the links contained in each document are extracted. These extracted links are added to the URL list if they are not already present. This process is repeated periodically to obtain the updated status of the web documents.

  Database: A database is maintained to store each word and its corresponding appearance in all the documents. The position and the number of occurrences of the root word, the type of proprietary HTML tags are maintained and used for ranking analysis. The database module maintains all the words and their appearance in each document.

  Search : In the search module, when a user queries the search engine for a particular topic, the database is looked up and queried based on the root word, and the records(pages) which satisfy the query are retrieved. Ranking is then performed on these pages and the results are sent back to the user. The ranking is based on the order of number of outgoing links, number of occurrences, description (meta), head and title. The results are displayed in the order of AND, OR operations. The document, which has the maximum number of query words, gets a higher rank. The results are displayed in lists of ten pages. The user has the facility to navigate through this list.

  Block Diagram

Block Diagram



CRAWLING	SEARCHING
Parser	Searcher
Crawler (invoked)	Database (retrieved)
file converters (used)	BTree (built)
generic font converter (used)	Search (request)
BTree (built)	Ranking
Database (stored)	Display (response)

User Interface : The user interface is provided to enter the query by the user. The query is analyzed with the information from the database and returns the matched information to the user.

- **Specifications of the Technology :** The following environment is used for the development of the Tamil Search Engine:

  Platform

  Windows ME/2000

  Software Components

  Java 1.4 & above

  Microsoft Access 98

  JRun Server

  Microsoft Internet Explorer 5.0

Hardware Configuration

Processor Intel Pentium PIII

RAM 128 MB

Hard Disk 20.0 GB

Requirements

J2RE 1.4

Operating Systems: Windows/Linux

- **Representative Snapshots:**

- **Scalability** : Yes
- **Portablility** : Yes
- **Readiness of Transfer of Technology** : Yes
- **Availability of documentation** : User and Technical Manuals are available
- **Testing of the Product / Technology :** 1600 pages are crawled and tested with many words and the results retrieved are available
- **Open-source /IPR** : Open Source
- **Potential Beneficiaries** : Public and students.
- **User-agency tie-up** : No
- **Name and address of the Resource Person** :

Dr. T.V.Geetha,
Dr.Ranjani Parthasarathi,
RCILTS -Tamil,
School of Computer Science and Engineering,
Anna University, Chennai-600 025.