CHAPTER 1

INTRODUCTION

1.1 Overview

With an increasingly globalized economy, the ability to find information in other languages is becoming a necessity. Though initially the World Wide Web was dominated by English, now less than half of existing web pages are in English. Thus the problem is that much Internet content is effectively inaccessible as it is stored in languages that are not searchable without specific technical know-how.

To overcome the above problem and to make communication global, the access to cross lingual information is more and more widespread within this society.

The demand for fully multilingual multimedia retrieval systems first formulated in 1997, is increasingly relevant today as the global networks become vital sources of information for both professional and leisure activities.

Information retrieval (IR) is a process that determines the relationship between document and query. It also defines the strength between them. Cross language information retrieval (CLIR) is the subfield of information retrieval. The basic idea behind the cross language information retrieval (CLIR) system is to retrieve documents in a language different from a query language made in the user's own language. This may be desirable even when the user is not a speaker of the language used in the retrieved documents. Once it is known that the information exists and is relevant, the retrieved documents can be translated by a translator.

Research in the area of cross-language information retrieval (CLIR) has focused mainly on methods for translating queries. Full document translation for large collections is impractical, thus query translation is a viable alternative.

For this project, we have used Google APIs for translation, transliteration and for searching. Google API is a set of JavaScript APIs developed by Google that allows interaction with Google Services and integration of rich, multimedia, search or feed-based Internet content into web applications. They extensively use AJAX scripting and can be easily loaded using Google Loader.

In the later part, we have added a feature that provides the user with the summery of each URL displayed on mouse hover.

1.2 Problem Definition

To develop an extension in the Google app store that provides an interface to the user to retrieve and access information in one or more specific languages. It allows the user to enter the query in one specific language and retrieve the document in the same or any other language with English being one of the default output languages. The application reduces the number of irrelevant documents for manual transactions and outputs only four top result in each language. The application provides a feature that generates the summary/snippet for each of the displayed result on mouse hover.

1.3 Motivation

There is lot of information available on internet in different languages mainly in English. Due to language barrier most of the information is not accessible to the common people of India. They are unable to form a query in English or in language other than their native language. The main motivation of developing this system is to allow people to access the information available in English and Hindi by forming a query in Marathi or other Indian language they are comfortable in.

1.4 Benefits

1. This app is simple and hassle free to use.
2. Allows users to retrieve results in one or more language at a time.
3. Allows user to check the summary of each resulted url.
4. The app is open source and easily accessible.

5. The interface is user friendly and provides optimal functionality.
6. It’s safe and secure.

CHAPTER 2

LITERATURE SURVEY

CLIA IIT Bombay group created an application named “Sandhan” [4]. This is a mission mode project executed by a consortium of academic and research institutions and industry partners. It focuses on Tourism domain. It supports 6 Indian languages which are Hindi, Marathi, Bengali, Panjabi, Tamil and Telugu and English. This system is developed on top of Nutch and Lucene Architecture.

One of the biggest issues with CLIR studied by Erbuğ Çelebi, Baturman Şen, Burak Günel to access the bi-lingual parallel corpus [5]. So, the first step of this study was to construct a parallel Turkish- English corpus. They have constructed a corpus that has 1801 parallel documents. The corpus has been divided in to two parts, first one for training the system and second one for testing the system. Latent semantic indexing (LSI) techniques applied to the training set to obtain the language relations. After the training, they have performed set of tests (queries) to measure the effectiveness of LSI based retrieval on Turkish-English parallel corpus. The experimental results show that, LSI based CLIR outperforms the non-LSI based retrieval where their retrieval successes are %69 and %26 respectively [5].

Wessel Kraaij and Ren´ee Pohlmann have conducted two experiments in the domain of Cross Language Information Retrieval[6]. The basic approach is to translate queries word by word using machine readable dictionaries. The first experiment compared different strategies to deal with word sense ambiguity: i) keeping all translations and integrate translation probabilities in the model, ii) a single translation is selected on the basis of the number of occurrences in the dictionary iii) word by word translation after word sense disambiguation in the source language. In a second experiment we constructed parallel corpora from web documents in order to construct bilingual dictionaries or improve translation probability estimates. They concluded that the best dictionary based CLIR approach is based on keeping all possible translations, not by simple substitution of a query term by its translations but by including reverse translation probabilities in the retrieval model.

Statistical language model estimation requires large amounts of domain-specific text, which is difficult to obtain in many languages. Woosung Kim and Sanjeev Khudanpur proposed the techniques which exploit domain specific text in a resource-rich language to adapt a language model in a resource-deficient language [7]. A primary advantage of this technique is that in the process of cross-lingual language model adaptation and do not rely on the availability of any machine translation capability. Instead, we assume that only a modest-sized collection of story-aligned document-pairs in the two languages is available. It uses ideas from cross-lingual latent semantic analysis to develop a single low-dimensional representation shared by words and documents in both languages, which enables us to (i) find documents in the resource-rich language pertaining to a specific story in the resource-deficient language, and (ii) extract statistics from the pertinent documents to adapt a language model to the story of interest.

Gareth J. F. Jones, Fabio Fantino, Eamonn Newman and Ying Zhang developed the syatem for Query translation for CLIA [8]. The main focus was on cultural heritage domain. The author have used Dictionary based approach for query translation and used Wikipedia for retrieving results.

Fatiha Sadat, Masatoshi Yoshikawa, Shunsuke Uemura have presented an approach to bilingual lexicon extraction from non-aligned comparable corpora, phrasal translation as well as evaluations on Cross-Language Information Retrieval in the model they developed [10]. A two stages translation model is proposed for the acquisition of bilingual terminology from comparable corpora, disambiguation and selection of best translation alternatives according to their linguistics-based knowledge. Different rescoring techniques are proposed and evaluated in order to select best phrasal translation alternatives. Results demonstrate that the proposed translation model yields better translations and retrieval effectiveness could be achieved across Japanese-English language pair.

N.Swapna, Padmaja Rani ,Kiran Kumar described some of the most important areas of information retrieval. In particular, Cross-lingual Information Retrieval (CLIR) and Multilingual Information Retrieval(MLIR) [11]. CLIR deals with asking questions in one language and retrieving documents in different language. MLIR deals with asking questions in one or more languages and retrieving documents in one or more different languages. With an increasingly globalized economy, the ability to find information in other languages is becoming a necessity. The authors also present the evaluation initiatives of information retrieval domain.

Ari Pirkola and Turid Hedlund have done literature on dictionary-based cross-language information retrieval (CLIR) and presents CLIR research done at the University of Tampere (UTA). The main problems associated with dictionary based CLIR, as well as appropriate methods to deal with the problems are discussed. The authors presented the structured query model by Pirkola and report findings for four different language pairs concerning the effectiveness of query structuring. The architecture of automatic query translation and construction system is presented [12].

Xabier Saralegi and Maddalen López de Lacalle discused two main problems in Cross-language Information Retrieval are translation selection and the treatment of out-of vocabulary terms [13]. The authors have focused on the problem concerning the translation selection. Structured queries and target co-occurrence-based methods seem to be the most appropriate approaches when parallel corpora are not available. However, there is no comparative study. Authors also compared the results obtained using each of the aforementioned methods, we specify the weaknesses of each method, and finally we propose a hybrid method to combine both. In terms of mean average precision, results for Basque-English cross-lingual retrieval show that structured queries are the best approach both with long queries and short queries.

The problem of transliterating English names using Chinese orthography in support of cross-lingual speech and text processing applications is described by Paola Virga, Sanjeev Khudanpur [14]. It demonstrates the application of statistical machine translation techniques to translate the phonemic representation of an English name, obtained by using an automatic text-to-speech system, to a sequence of initials and finals, commonly used sub-word units of pronunciation for Chinese. Statistical translation model is used to map the initial/final sequence to Chinese characters. It also presents an evaluation of this module in retrieval of Mandarin spoken documents from the TDT corpus using English text queries.

CHAPTER 3

TECHNOLOGIES USED

3.1 Python

Python is a widely used general-purpose, high-level programming language. Its design philosophy emphasizes code readability, and its syntax allows programmers to express concepts in fewer lines of code. The language provides constructs intended to enable clear programs on both a small and large scale.

Python supports multiple programming paradigms, including object-oriented, imperative and functional programming or procedural styles. It features a dynamic type system and automatic memory management and has a large and comprehensive standard library [1].

- **Indentation:** Python uses whitespace indentation, rather than curly braces or keywords, to delimit blocks; a feature also termed the off-side rule. An increase in indentation comes after certain statements; a decrease in indentation signifies the end of the current block. It is considered beneficial by Python programmers, but others have criticized it.
- **Libraries:** Python has a large standard library, commonly cited as one of Python's greatest strengths, providing tools suited to many tasks.
- **Development environments:** Most Python implementations can function as a command line interpreter, for which the user enters statements sequentially and receives the results immediately. In short, Python acts as a shell.

Other shells add capabilities beyond those in the basic interpreter, including IDLE and IPython. While generally following the visual style of the Python shell, they implement features like auto-completion, retention of session state, and syntax highlighting.

3.2 ligHTTP Server:

lighttpd (pronounced "lighty")is an open-source web server optimized for speed-critical environments while remaining standards-compliant, secure and flexible.It was originally written by Jan Kneschke as a proof-of-concept of the c10k problem - how to handle 10,000 connections in parallel on one server, but has gained worldwide popularity.

It runs natively on Unix-like operating systems as well as Microsoft Windows.Following are some of its features:

1. Load balancing FastCGI, SCGI and HTTP proxy support
2. select()-/poll()-/epoll() based web server
3. TLS/SSL with SNI support, via OpenSSL.
4. Server Side Includes support (but not server-side CGI)
5. Flexible virtual hosting
6. Modules support
7. Cache Meta Language (currently being replaced by mod_magnet) using the Lua programming language
8. Light-weight (less than 1 MB)
9. Single-process design with only several threads. No processes or threads started per connection.

3.3 JavaScript:

JavaScript (JS) is a dynamic computer programming language. It is most commonly used as part of web browsers, whose implementations allow client-side scripts to interact with the user, control the browser, communicate asynchronously, and alter the document content that is displayed. It is also being used in server-side programming, game development and the creation of desktop and mobile applications.

JavaScript is a prototype-based scripting language with dynamic typing and has first-class functions. Its syntax was influenced by C. JavaScript copies many names and naming conventions from Java, but the two languages are otherwise unrelated and have very different semantics. The key design principles within JavaScript are taken from the Self and Scheme programming languages. It is a multi-paradigm language, supporting object-oriented, imperative, and functional programming styles [16].

Features:

1. Imperative and structured
2. Dynamic typing
3. Object-based
4. Run-time evaluation

3.4 PHP:

PHP is a server-side scripting language designed for web development but also used as a general-purpose programming language. As of January 2013, PHP was installed on more than 240 million websites (39% of those sampled) and 2.1 million web servers. Originally created by Rasmus Lerdorf in 1995, the reference implementation of PHP is now produced by The PHP Group [16]. While PHP originally stood for Personal Home Page, it now stands for PHP: Hypertext Preprocessor, a recursive backronym.

PHP code is originally designed to be interpreted by a web server with a PHP processor module, which generates the resulting web page. PHP commands can be embedded directly into a HTML source document rather than calling an external file to process data. It has also evolved to include command-line interface capability and can be used in standalone graphical applications.

PHP is free software released under the PHP License. PHP has been widely ported and can be deployed on most web servers on almost every operating system and platform, free of charge.

3.5 Ajax

Ajax (acronym for Asynchronous JavaScript and XML) is a group of interrelated Web development techniques used on the client-side to create asynchronous Web applications. With Ajax, Web applications can send data to, and retrieve data from, a server asynchronously (in the background) without interfering with the display and behavior of the existing page. Data can be retrieved using the XMLHttpRequets object [16]. Despite the name, the use of XML is not required; JSON is often used instead (see AJAJ), and the requests do not need to be asynchronous.

Ajax is not a single technology, but a group of technologies. HTML and CSS can be used in combination to mark up and style information. The DOM is accessed with JavaScript to dynamically display, and allow the user to interact with, the information presented. JavaScript and the XMLHttpRequest object provide a method for exchanging data asynchronously between browser and server to avoid full page reloads.

3.6 jQuery

jQuery is a cross-platform JavaScript library designed to simplify the client-side scripting of HTML. It was released in January 2006 at BarCamp NYC by John Resig. It is currently developed by a team of developers led by Dave Methvin. Used by over 80% of the 10,000 most visited websites, jQuery is the most popular JavaScript library in use today.

jQuery is free, open source software, licensed under the MIT License. jQuery's syntax is designed to make it easier to navigate a document, select DOM elements, create animations, handle events, and develop Ajax applications. jQuery also provides capabilities for developers to create plug-ins on top of the JavaScript library [16]. This enables developers to create abstractions for low-level interaction and animation, advanced effects and high-level, theme-able widgets. The modular approach to the jQuery library allows the creation of powerful dynamic web pages and web applications.

Tooltip:

The tooltip or infotip or a hint is a common graphical user interface element. It is used in conjunction with a cursor, usually a pointer. When user hovers the pointer over an item, without clicking it, and a tooltip may appear—a small "hover box" with information about the item being hovered over. Tooltips do not appear on mobile operating systems, because there is no cursor.

The term tooltip originally came from older Microsoft applications (like Microsoft Word 95), which had a toolbar where moving the mouse over the buttons (the Toolbar icons) displayed these tooltips, a short description of the function of the tool in the toolbar. More recently, these tooltips are used in various parts of an interface, not only on toolbars.

Features:

1. DOM element selections using the multi-browser open source selector engine Sizzle, a spin-off of the jQuery project

2. DOM traversal and modification (including support for CSS 1–3)

3. DOM manipulation based on CSS selectors that uses node elements name and node elements attributes (id and class) as criteria to build selectors

4. Events

5. Effects and animations

6. AJAX

7. JSON parsing

8. Extensibility through plug-ins

9. Utilities - such as user agent information, feature detection

10. Compatibility methods that are natively available in modern browsers but need fall backs for older ones - For example the inArray() and each() functions.

11. Multi-browser support.

Browser support

Both version 1.x and 2.x of jQuery support "current-1 versions" (meaning the current stable version of the browser and the version that preceded it) of Firefox, Google Chrome, Safari, and Opera. The version 1.x also supports Internet Explorer 6 or higher. However, JQuery version 2.x dropped Internet Explorer 6–8 support (which represents less than 28% of all browsers in use) and can run only with IE 9 or higher [16].

3.7 Google API

Google APIs (or Google AJAX APIs) is a set of JavaScript APIs developed by Google that allows interaction with Google Services and integration of rich, multimedia, search or feed-based Internet content into web applications. They extensively use AJAX scripting and can be easily loaded using Google Loader [15].

Google Loader (or Google AJAX APIs Loader) is a JavaScript API which allows web developers to easily load other JavaScript APIs provided by Google and other developers of popular libraries. Google Loader provides a JavaScript method for loading a specific API (also called module), in which additional settings can be specified such as API version, language, location, selected packages, load callback and other parameters specific to a particular API. Dynamic loading or auto-loading is also supported to enhance the performance of the application using the loaded APIs.

3.7.1 Google transliteration API

The Transliterate API supports Firefox 1.5+, IE6+, Safari, and Chrome. The Transliterate API can be loaded without errors in almost every browser.

google.language.transliterate(wordsArray, srcLang, destLang, callback) is a global method that transliterates given text from the source language to the destination language [15]. The API returns the result asynchronously to the given callback function as the resultobject. Parameters for this method are:

- wordsArray: provides the text to be transliterated as an array.
- srcLang: provides the source language as a language code.
- destLang: provides the destination language as a language code.
- Callback: is the callback function that receives the result.

google.language.transliterate() has no return value. google.language.transliterate() outputs the result object.

3.7.2 Google translation API

Google Translate is a free, multilingual statistical machine-translation service provided by Google Inc. It automatically translates text from one language to another language. The source text is the text to be translated. The source language is the language that the source text is written in. The target language is language that the source text is translated into.

The specific format for the single Google Translate API URI is:

http://www.googleapis.com/language/translate/v2?*parameters*

where *parameters* are any parameters to apply to the query [15].

There are three methods to invoke in the Google Translate API:

- Translate: Translates source text from source language to target language
- languages: List the source and target languages supported by the translate methods
- detect: Detect language of source text

3.7.3 Google search API

The Google Web Search API allows us to add Google Search in the web pages with JavaScript. We can embed a simple, dynamic search box and display search results in your own web pages or use the results in innovative, programmatic ways.

The Google Web Search API currently supports Firefox 1.5+, IE 6, Safari, Opera 9+, and Chrome.

To use the Web Search API within web site, it is needed to include the URL for the Google APIs loader (https://www.google.com/jsapi). This library allows us to load various APIs via google.load('api', 'version'). Searcher objects determine which search services the search control operates over. The different types of objects are WebSearch, VideoSearch, BlogSearch, NewsSearch, ImageSearch, PatentSearch and BookSearch [15].

CHAPTER 4

SYSTEM REQUIREMENT SPECIFICATION

Software requirement Specification is a fundamental document, which form the foundation of the software development process. It not only lists the requirements of a system but also has a description of its major feature. An SRS is a basically an understanding (in writing) of a customer or potential client's system requirements and dependency at a particular point in time (usually) prior to any actual design or development work. It's a two way insurance policy that assures that both the client and organization understand the other's requirements from that perspective at a given point in time. The SRS also functions as a blueprint for completing the project with as little cost growth as possible. The SRS is often referred to as the "parent" document because all subsequent project management documents, such as design specifications, statements of work, software architecture specifications, testing and validation plans and documentation plans are related to it. It is important to note that an SRS contains functional and non-functional requirements only. It does not offer design suggestions, possible solutions to technology or business issues, or any other than what the development team understands customer's system requirements to be.

4.1 Functional Requirements

Functional requirements is a list of functionality that the application or software is supposed to provide. Thus the functional requirements of the application are as follows:

1. To design an extension to provide CLIR mechanism.
2. Project will be developed and deployed under Linux platform.
3. The user interface will be web based.
4. Summarizes the information for each of the resulted links

5. Provides an option to the user for selecting the language in which he/she wants the result to be displayed.

4.2 Non-Functional Requirements

Non-functional requirements are those requirements that are not directly concerned with the specific functions delivered by the system. They may relate to emergent system properties such as reliability, response time and store occupancy. Alternatively, they may define constraints on the system such as capability of the input/output devices and data representations used in system interfaces. Many non-functional requirements relate to the system as whole rather than to individual system features. This means they are often critical than the individual functional requirements. The following non-functional requirements are worthy of attention:

1. Performance: The performance depends upon the network connections.
2. Reliability: Searches should be reliable and relevant
3. Availability: The application has to be available 24 X 7 and should provide functionality at all times.
4. Portable: Can be used in any of the flavor of Linux and windows.

4.3 Hardware Requirements

1. **Processor :** Intel® core(m) 2 Duo CPU
2. **RAM** : 256 MB
3. **Hard Disk** : 500 GB

4.4 Software Requirements

1. **Operating System** : Ubuntu Linux(Elementary OS)
2. **Browser :** Google Chrome

3. **Server :** Lighttpd
4. **Languages Used :** Python, PHP, JavaScript

CHAPTER 5

SYSTEM DESIGN

Our project is based on client server architecture. Client-server network is a network in which certain computers have special dedicated tasks, providing services to other computers (in the network). The client–server model of computing is a distributed application structure that partitions tasks or workloads between the providers of a resource or service, called servers, and service requesters, called clients. Often clients and servers communicate over a computer network on separate hardware, but both client and server may reside in the same system. A server host runs one or more server programs which share their resources with clients. A client does not share any of its resources, but requests a server's content or service function. Clients therefore initiate communication sessions with servers which await incoming requests.

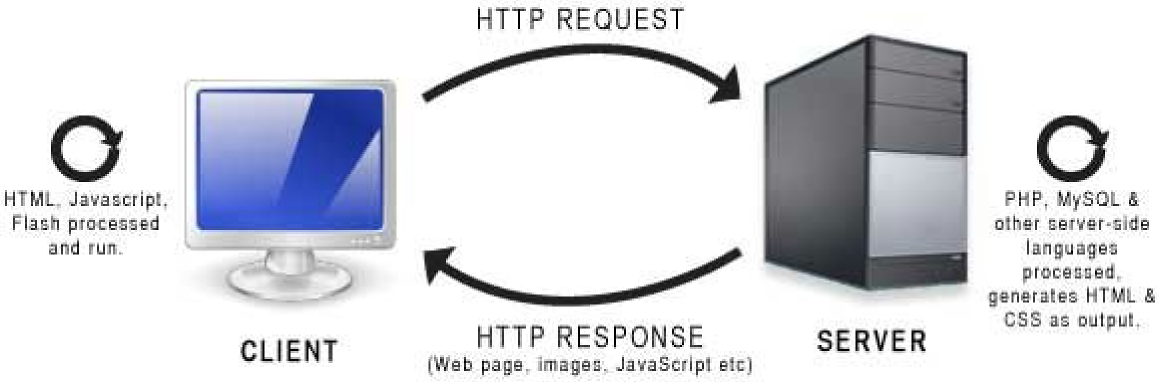

HTTP REQUEST
HTML, Javascript, Flash processed and run.
CLIENT
HTTP RESPONSE
(Web page, images, JavaScript etc)
SERVER
PHP, MySQL & other server-side languages processed, generates HTML & CSS as output.

Figure 5.1: System Design

5.1 System Architecture:

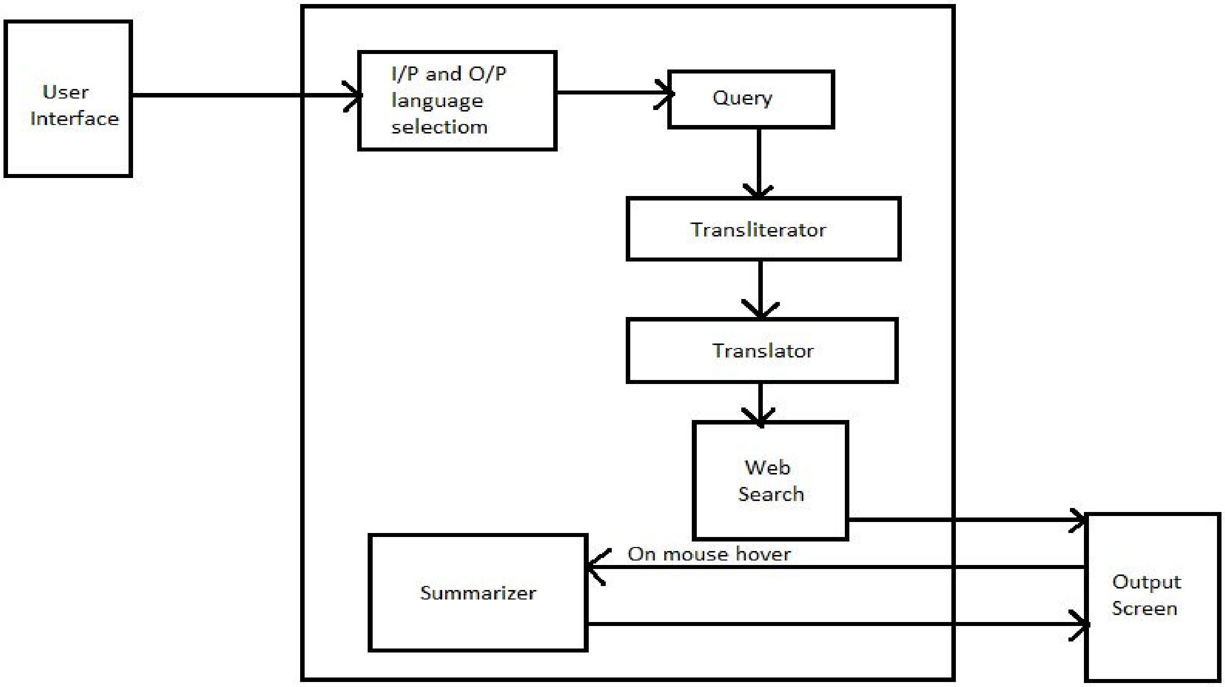


Figure 5.2: System Architecture

The diagram describes the system architecture design of Cross lingual search engine. The user needs to add extension or plug-ins to avail CLIR functionality to the browser. The user needs to input the query language. If the input language selected is any other than English, then the query will get transliterated into the input language. The user can select one or more output language. The query will get translated in the output language which would then be searched by the web. Top four results of each output language selected will be displayed. The user can read the summary of the each resulted URL by mouse hover over the URL, which would display the summery in a pop-up box.

4.2 Use-Case Diagram

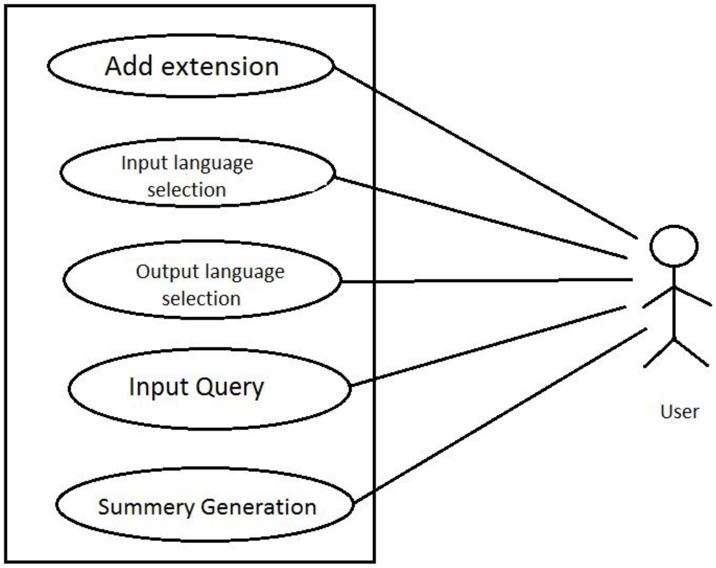


Figure 5.3: Use-case diagram

5.2.1 Actor:

User: User, any person who wants to perform a cross lingual search.

5.2.2 Use-cases:

1. Add extension: The user must add the chrome extensions to access the facility of cross lingual information retrieval.
2. Select input language: By default the input language is English. The user may opt any one of the desired language specified.
3. Select output language: By default the output language is English. The user may opt one or more desired language from the specified language list.
4. Input query: The user can enter the query in any of the specified language.

5. Summary generation: The user may check the summary generation check box in order to generate summary of each result on mouse hover.

5.3 Activity Diagram

5.3.1 Add Extension

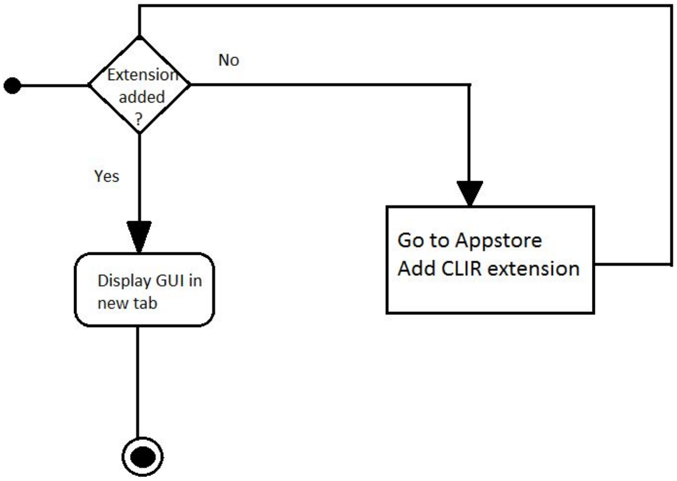


Figure 5.4: Add extension

Basic flow:

1. Add the extension
2. Click on extension
3. User interface will be displayed.

Pre-condition:

Extension should be added to the browser.

Post condition:

When clicked on the extension, the interface should be provided to the user in a new tab.

5.3.2 Input and Output Language Selection

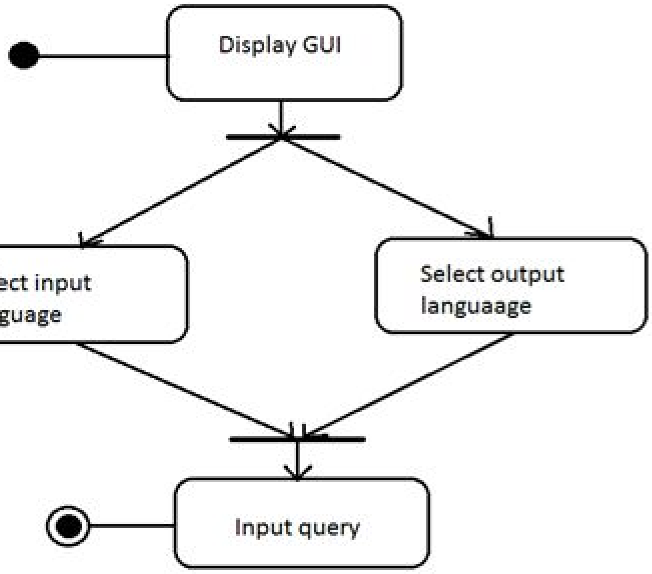


Figure 5.5: Input and output language selection

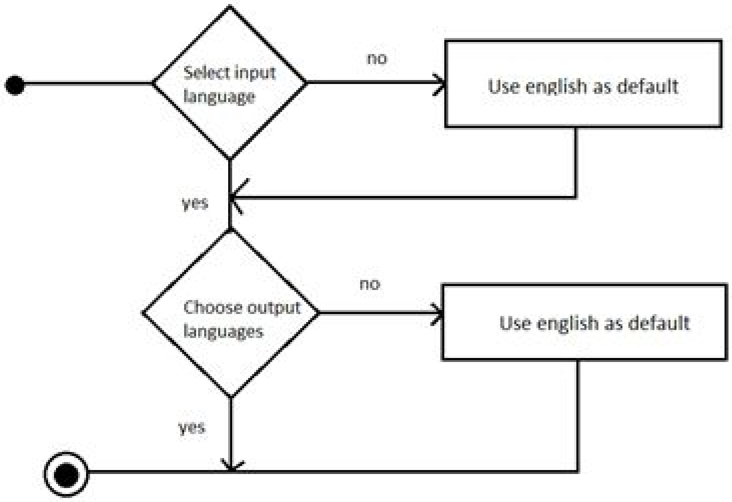


Figure 5.6: Default input and output language selection

Basic flow:

1. Select Input and Output Language.
2. Input the Query.

Pre-condition:

English is selected as the default input and output language

Post-condition:

Minimum one input language and one or more output language should be selected.

5.3.3 Transliteration

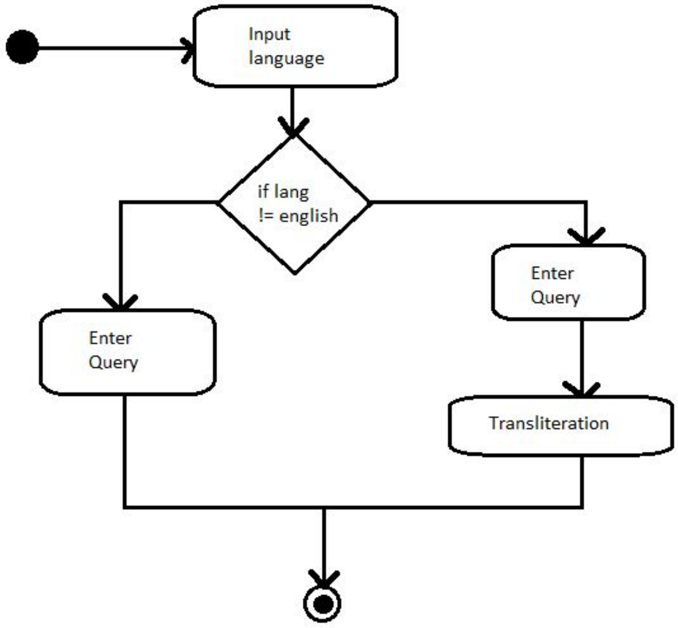


Figure 5.7: Transliteration

Basic flow:

1. Select the input language.

2. Input a one word of query and hit space for transliteration.

Pre-condition:

Input language should be any other than English.

Post-condition:

The query should be in transliterated form.

5.3.4 Search

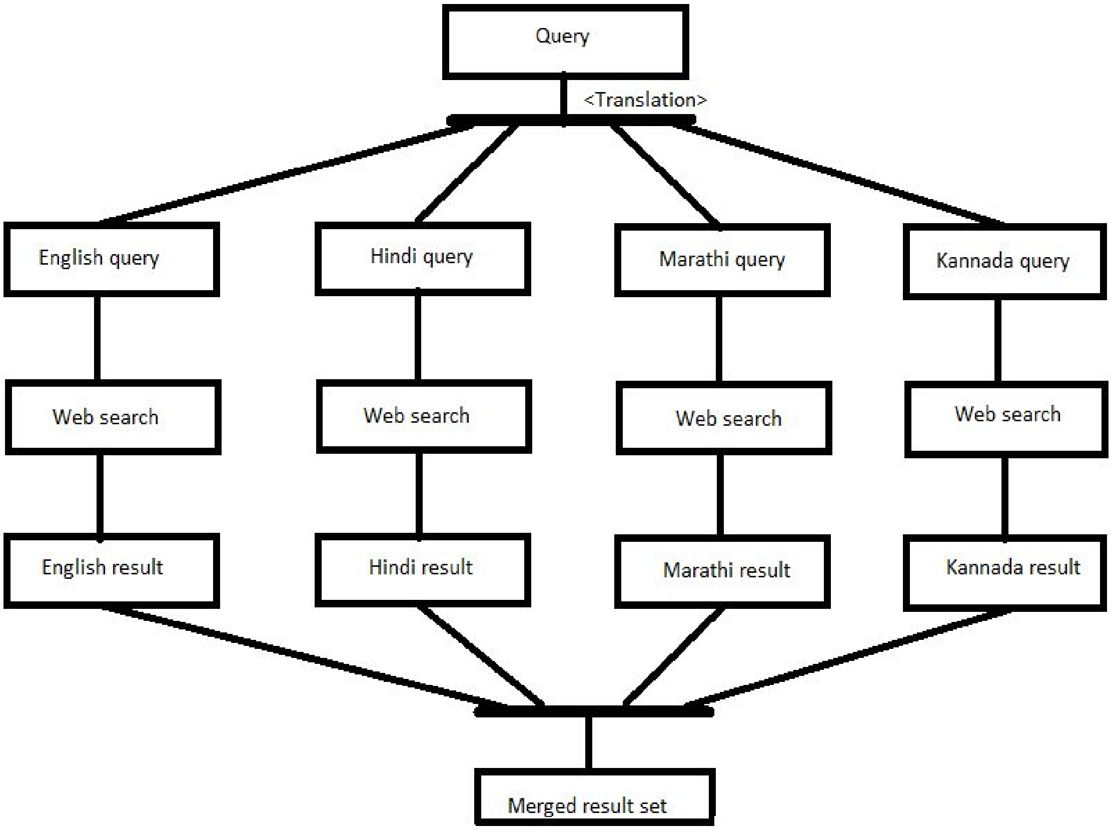


Figure 5.8: Translation and search

Basic flow:

1. The query will be translated to all the user selected languages.
2. Each translated query will be searched.

3. Results in different languages will be merged.

Pre-condition:

The query has to be entered

Post-condition:

The query should get translated and results in all selected languages should be displayed

5.3.5 Summary Generation

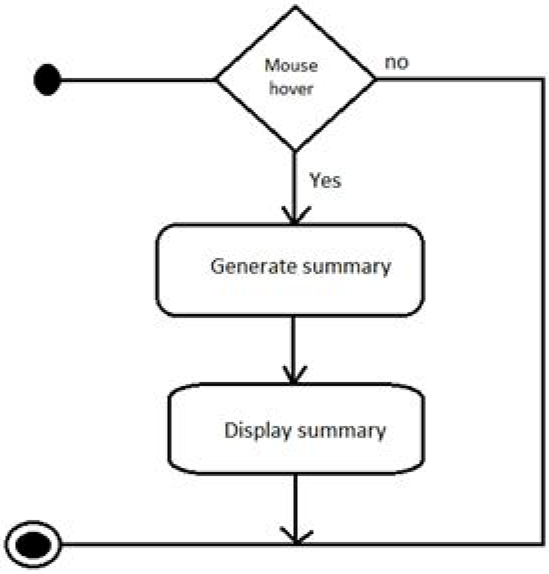


Figure 5.9: Summary generation

Basic flow:

1. Place cursor on any result displayed.
2. Summary is generated for that result.
3. Generated summary is displayed in the popup box.

Pre-condition:

The URLS should be displayed and cursor should be brought to the url of which the summary has to be checked.

Post-condition:

Summary will e generated in popup box.

5.4 Software Model

Software is an abstract representation of software process. It represents a process from a particular perspective so only provides partial information about the process. The software paradigm applied in the proposed system is the Iterative Development Model.

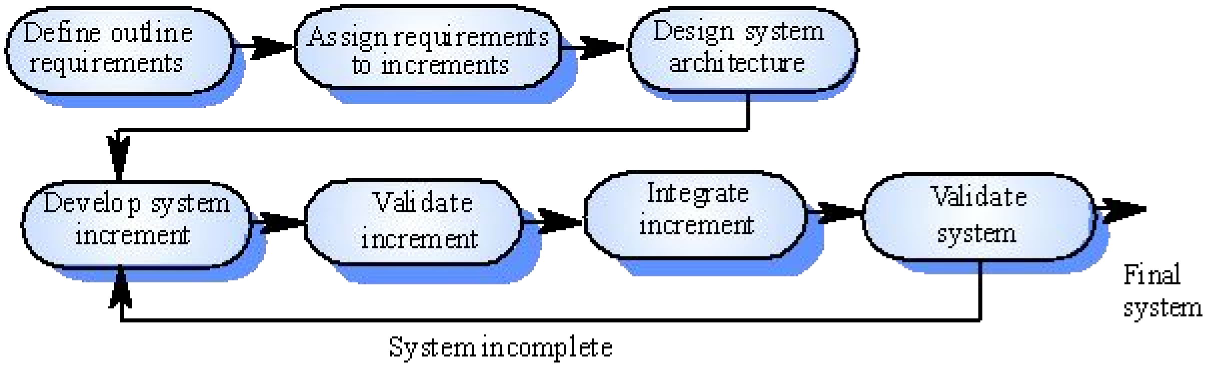


Figure 5.10: Iterative Development Model

Incremental development deals with delivering the system broken down into increments with each increment delivering part of the required functionality, rather than a single delivery date. User requirements are prioritized and the highest priority requirements are included in early increments.

Once the development of an increment is underway, the requirements are frozen though requirements so that later increments can continue to evolve. Customer value can be delivered with each increment so system functionality is available earlier in the development process.

Early increments act as a prototype to help draw out requirements for later increments. This therefore results in a lower risk of overall project failure.

CHAPTER 6

IMPLEMENTATION

6.1 Introduction

This section deals with the implementation details of the project. The basic data structures used in the construction of the application are given below. The algorithm form implementation is also shown in this section.

6.2 Algorithm

6.2.1 Algorithm: Create extension

Step 1: start

Step 2: Create a *manifest file* named manifest.json. The manifest file contains details like the name, version, and permissions of the extension.

Step 3: Visit chrome://extensions in your browser .

Step 4: Ensure that the checkbox in the top right-hand corner is checked.

Step 5: Click Load unpacked extension… to pop up a file-selection dialog.

Step 6: Navigate to the directory in which your extension files live, and select it.

Step 7: stop

6.2.2 Algorithm: Transliterate

Step 1: start

Step 2: check for input language selection if not equal to English

Step 3: call to Google transliterate API

Step 4: display the transliterated query

Step 5: stop

6.2.3 Algorithm: Translate

```
Step 1: start
Step 2: check for language selection
Step 3: call to Google translate API
Step 4: store translated query into file
Step 5: go to step 2 if more languages are selected
Step 6: stop
```

6.2.4 Algorithm: Search

```
Step 1: start
Step 2: check for language selection
Step 3: call to Google search API
Step 4: pass translated query to search API
Step 5: retrieve the search results
Step 6: Write title and URL of retrieved results into file for display
Step 7: go to step 2 if more language are selected
Step 8: stop
```

6.2.5 Algorithm: Summary generation

```
Step 1: start
Step 2: eliminate html tags
Step 3: find stop words and eliminate them
Step 4:  calculate the frequency of key words.
Step 5: rank the sentences
Step 6: display top ranked sentences.
Step 7: stop
```

7.1 Testing Methodologies

A strategy for system testing integrates system test cases and design techniques into well planned series that results in the successful construction of software. The testing strategy must co-operate test planning, test case design, test execution, and the resultant data collection and evaluation. A strategy for software testing must accommodate low-level tests that are necessary to verify that a small source code segment has been correctly implemented as well as high level tests that validate major system functions against user requirements. Software testing is a critical element of software quality assurance and represent the ultimate review of specification design and coding .Testing represents an interested anomaly for the software, thus ,a series of testing are performed for the proposed system before the system is ready for user acceptance testing.

7.1.1 Unit testing

Unit testing focuses verification effort on the smallest unit of software design that is the module. Unit testing exercises specific paths in a modules controls structure to ensure complete coverage and maximum error detection. This test focuses on each module individually, ensuring that it functions properly as a unit, hence, the naming is unit testing. During the testing, each module is testing individually and the module interfaces are verified for the consistency with design specification. All important processing paths are tested for the expected results. All error handling paths are also tested.

7.1.2 Integration testing

Integration testing addresses the issues associated with the dual problems of verification and problem construction. After software has been integrated a set of high order tests are conducted. The main objective of this testing is to take unit tested modules and builds a program structure that has been dictated by design.

The final module complete UI integration mainly covers up integration testing and high level testing is done .In our application category, subcategory , product, shop, offer are all tested as a single unit to see whether the apps working fine as a whole or not.

7.1.3 User Acceptance Testing

User acceptance of the system is the key factor for the success of any system. The system under consideration is tested for user acceptance by constantly keeping in touch with the prospective system users at the time of developing and making changes wherever required. The system developed provides a friendly user interface that can easily be understood even by a person who is new to the system.

7.1.4 Output testing

Output testing involves checking the application output to check whether the expected output and observed output are same for the given input. The application is installed on a test machine. There are no issues that are detected during installation. The application accurately provides all functionality.

7.2 Test Plan Contents

The testing process involves the description of major phases of the testing process. Requirements tractability where all the requirements described earlier is tested individually. Tested items where the products of the software process to be tested and specified. This involves testing all the executable codes through a single interface.

Testing schedule, which is linked to more general project development schedule? Test recording procedures where the results of the tests performed under tested items are systematically recorded hardware and software requirements.

7.3 Test Plan

System testing is very expensive. For some large system with complex non functional requirements, half of the system development budget may be spent on testing. Careful planning is needed to get the most out of the testing and to control testing costs. Test planning is concerned with setting out of testing process rather that description product test. As far as possible, system has been tested so, that programs do not contain any bugs and run properly. However 100% bug-free software is not attainable, so some discrepancies might have crept in.

7.4 Test Environment

Table 7.1: Test Environment

Requirements	Specification
Processor	Intel® core(m) 2 Duo CPU
RAM	4GB
Hard disk	500GB
Operating system	Ubuntu Linux(elmentray OS)
Browser	Google Chrome
Server	Lighttpd
Languages	Python, html, php, javascript

7.5 Test Cases

7.5.1 Add Extension

Table 7.2: Add Extension

Test condition	Test data	Expected Result	Actual Result
On click <extension>	If extension is being opened	The UI of the web application	Display the UI of the application

7.5.2 User Interface

Table 7.3: User interface

Test condition	**Test data**	**Expected Result**	**Actual Result**
On button <radio> On click	Input language selection	Only one language should be selected	Only one language should be selected
On button <checkbox> On click	Output language selection	one or more language should be selected	one or more language should be selected
On type in <textbox> On click	Single word query	Transliterated form of the input language query	Transliterated form of the input language query
On type in <textbox> On click	Copied query	Sentence copied from anywhere will not be transliterated	Sentence copied from anywhere will not be transliterated

7.5.3 Result

Table 7.4: Result table

Test condition	**Test data**	**Expected Result**	**Actual Result**
On click <Go>	Query	Top four Results in all selected languages	Top four Results in all selected languages

7.5.4 Summary Generation

Table 7.5: Summary generation

Test condition	Test data	Expected Result	Actual Result
On mouse hover <url>	Web document	Summary of the web page	Summary of the web page

CHAPTER 8

SNAPSHOTS

8.1 Chrome Extension

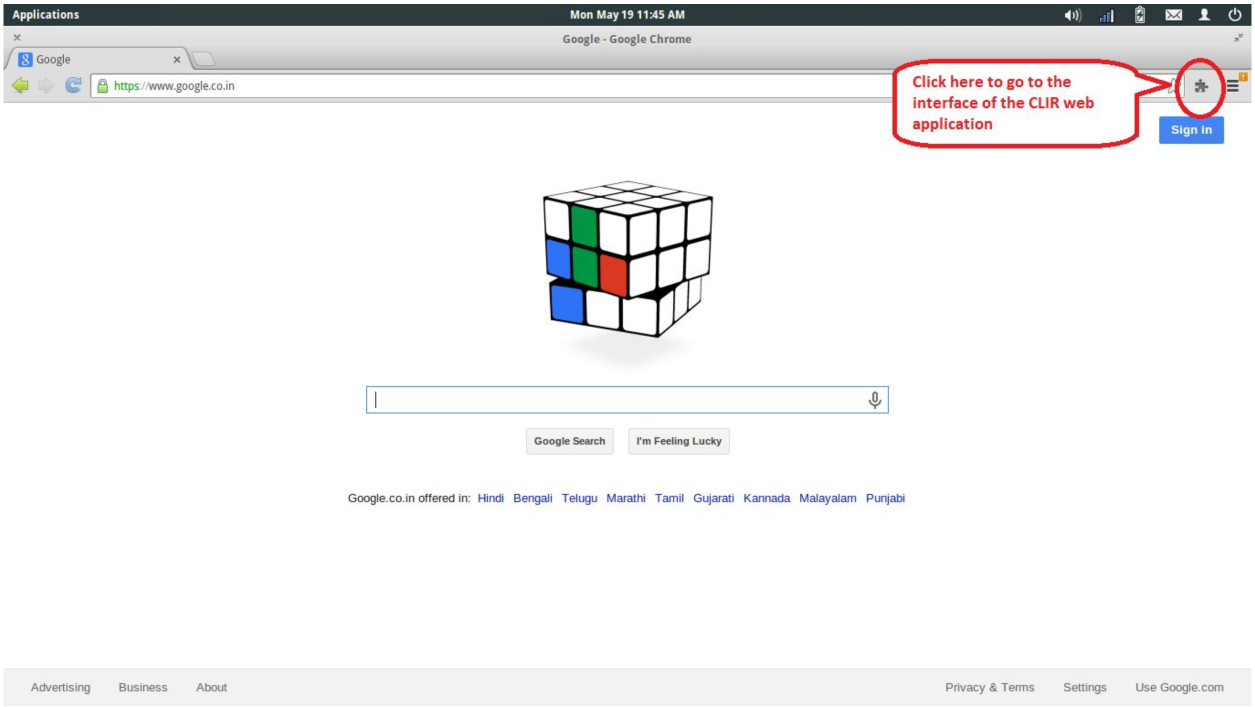
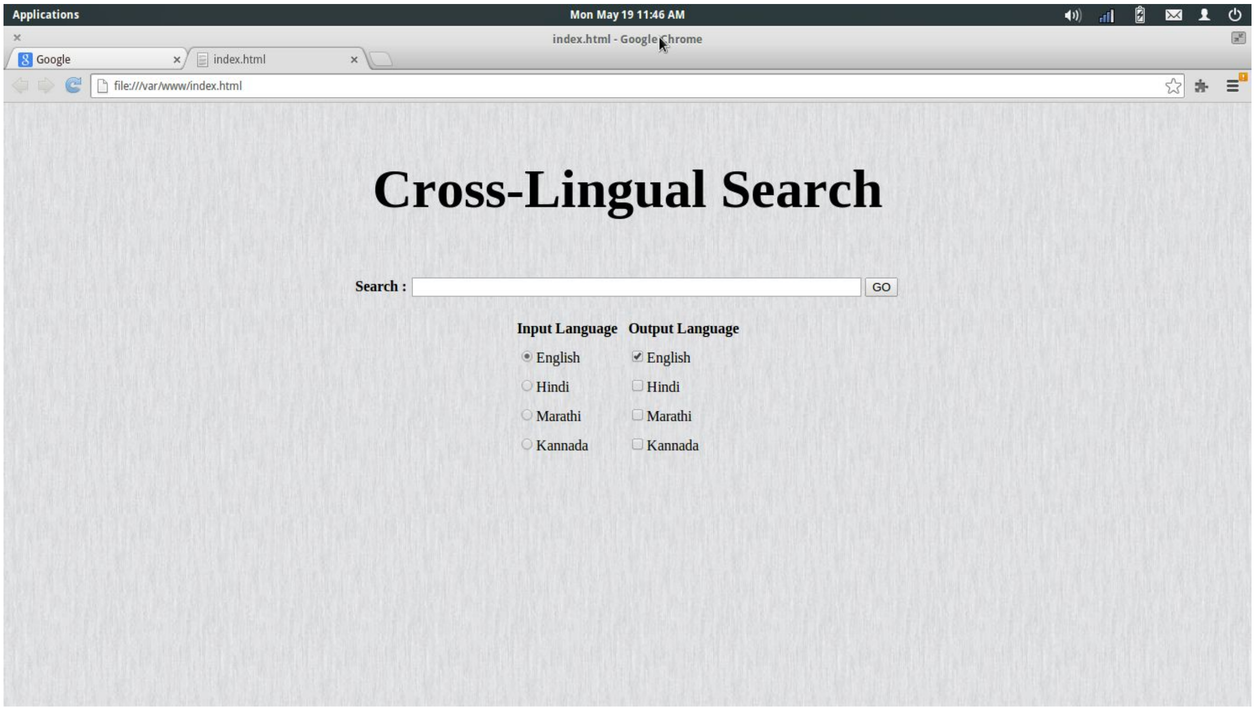


Figure 8.1: Chrome extension

Add the extension to the chrome browser to avail the facility of Cross Lingual Search & Retrieval. Clicking on the extension (as shown in the red box) will take you to the application's user interface.

8.2 User Interface



Cross-Lingual Search

Search : GO

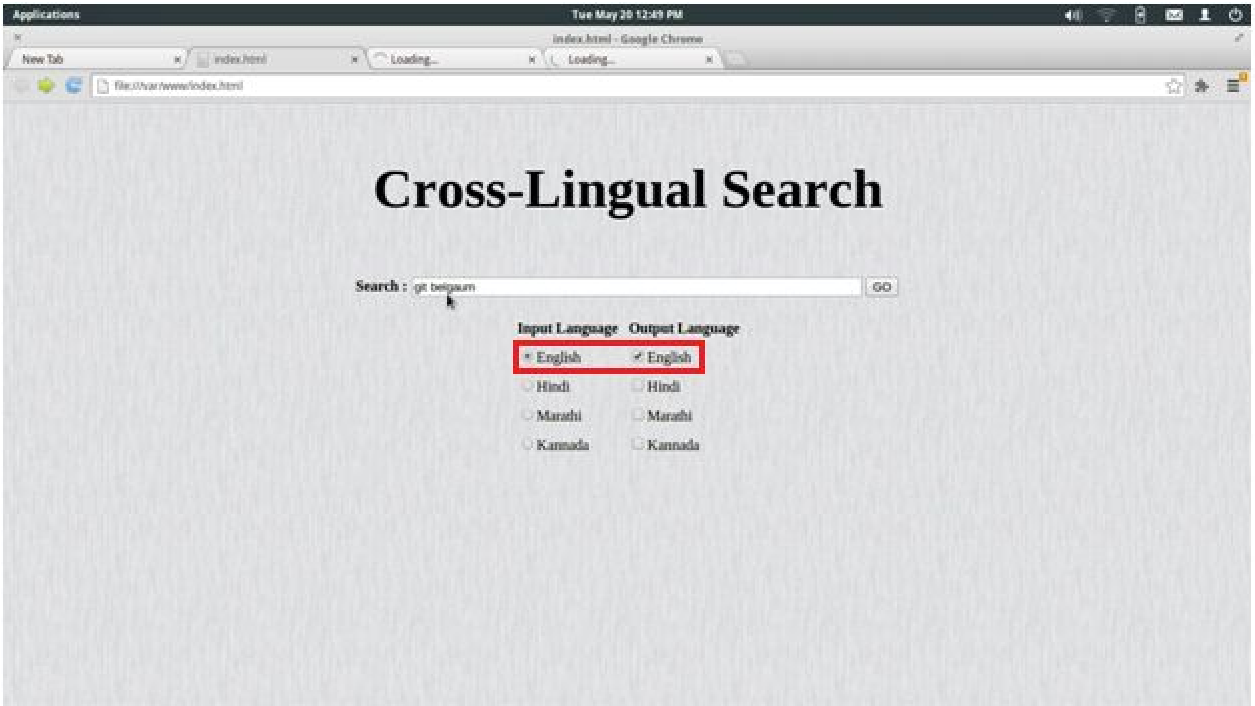
Input Language	Output Language
● English	☑ English
○ Hindi	□ Hindi
○ Marathi	□ Marathi
○ Kannada	□ Kannada

Figure 8.2: User interface

This snap shot shows the user interface of the application. The interface consists of a search text box, a set of radio buttons for input language selection and a set of check boxes for output language selection.

8.3 Case 1

Input language: English Output language: English

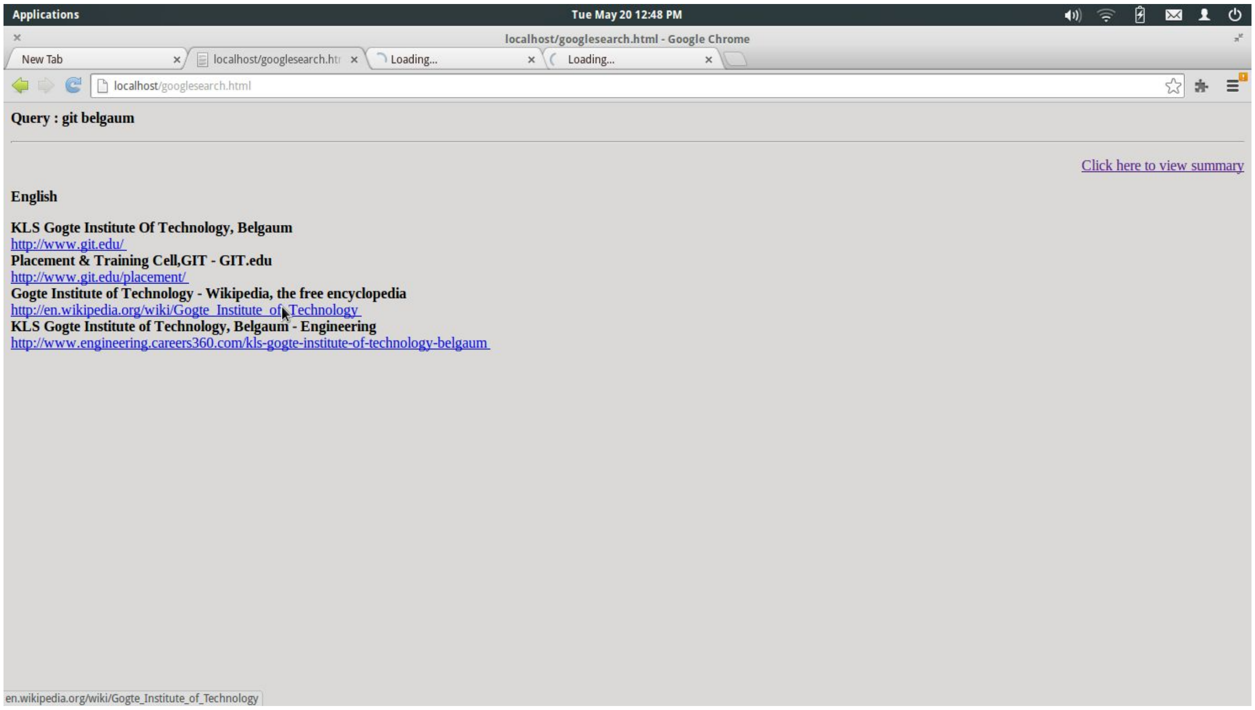


Cross-Lingual Search

Search : git belgaum GO

Input Language	Output Language
● English	☑ English
○ Hindi	□ Hindi
○ Marathi	□ Marathi
○ Kannada	□ Kannada

Figure 8.3: English Query


Applications
Tue May 20 12:48 PM
localhost/googlesearch.html - Google Chrome
New Tab
localhost/googlesearch.htr
Loading...
Loading...
localhost/googlesearch.html
en.wikipedia.org/wiki/Gogte_Institute_of_Technology

Query : git belgaum

Click here to view summary

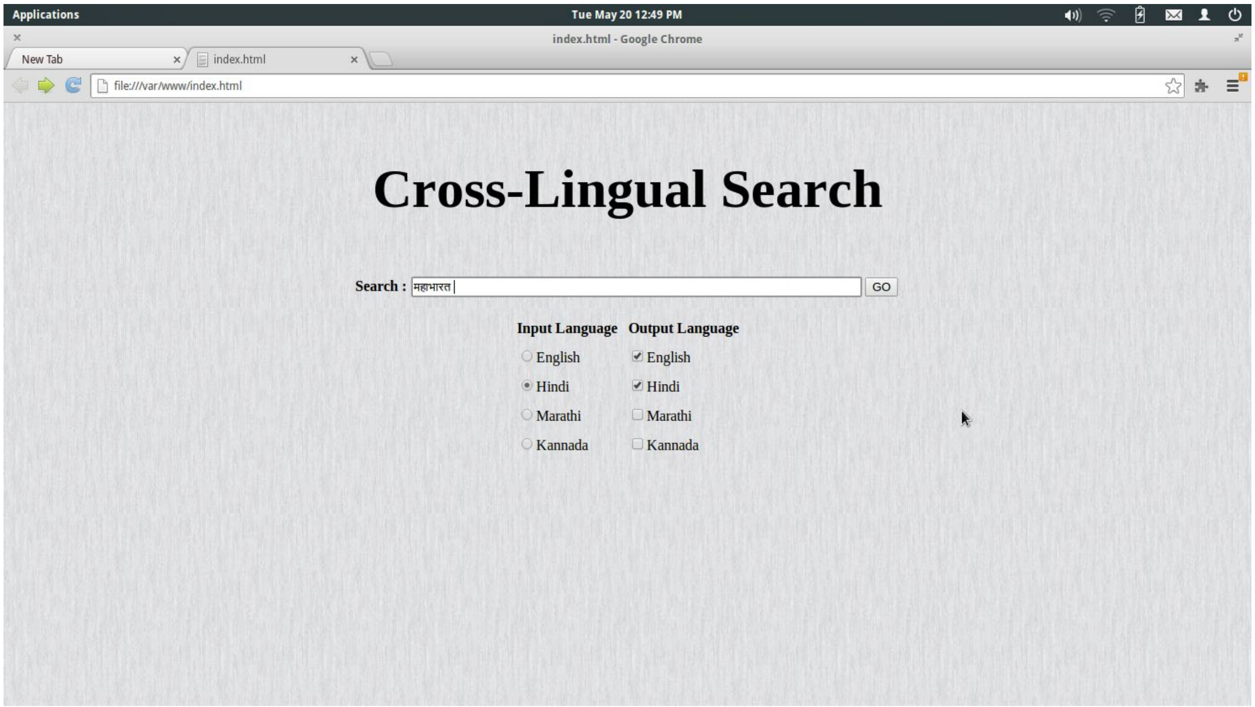
English

KLS Gogte Institute Of Technology, Belgaum
http://www.git.edu/
Placement & Training Cell,GIT - GIT.edu
http://www.git.edu/placement/
Gogte Institute of Technology - Wikipedia, the free encyclopedia
http://en.wikipedia.org/wiki/Gogte_Institute_of_Technology
KLS Gogte Institute of Technology, Belgaum - Engineering
http://www.engineering.careers360.com/kls-gogte-institute-of-technology-belgaum

Figure 8.4: English results

8.4 Case 2

Input language: Hindi Output language: English and Hindi

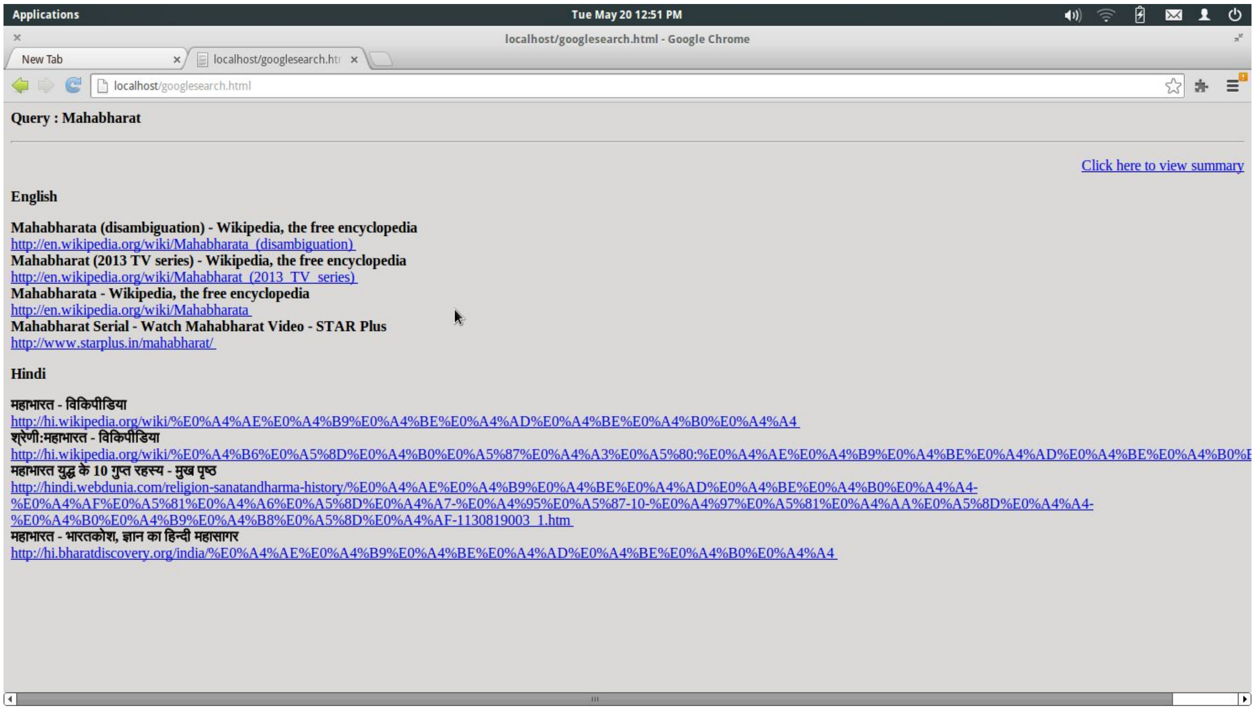


Cross-Lingual Search

Search : महाभारत

Input Language	Output Language
○ English	☑ English
● Hindi	☑ Hindi
○ Marathi	□ Marathi
○ Kannada	□ Kannada

Figure 8.5: Hindi Query


Applications
Tue May 20 12:51 PM
localhost/googlesearch.html - Google Chrome
New Tab
localhost/googlesearch.htr
localhost/googlesearch.html

Query : Mahabharat

Click here to view summary

English

Mahabharata (disambiguation) - Wikipedia, the free encyclopedia
http://en.wikipedia.org/wiki/Mahabharata_(disambiguation)
Mahabharat (2013 TV series) - Wikipedia, the free encyclopedia
http://en.wikipedia.org/wiki/Mahabharat_(2013_TV_series)
Mahabharata - Wikipedia, the free encyclopedia
http://en.wikipedia.org/wiki/Mahabharata
Mahabharat Serial - Watch Mahabharat Video - STAR Plus
http://www.starplus.in/mahabharat/

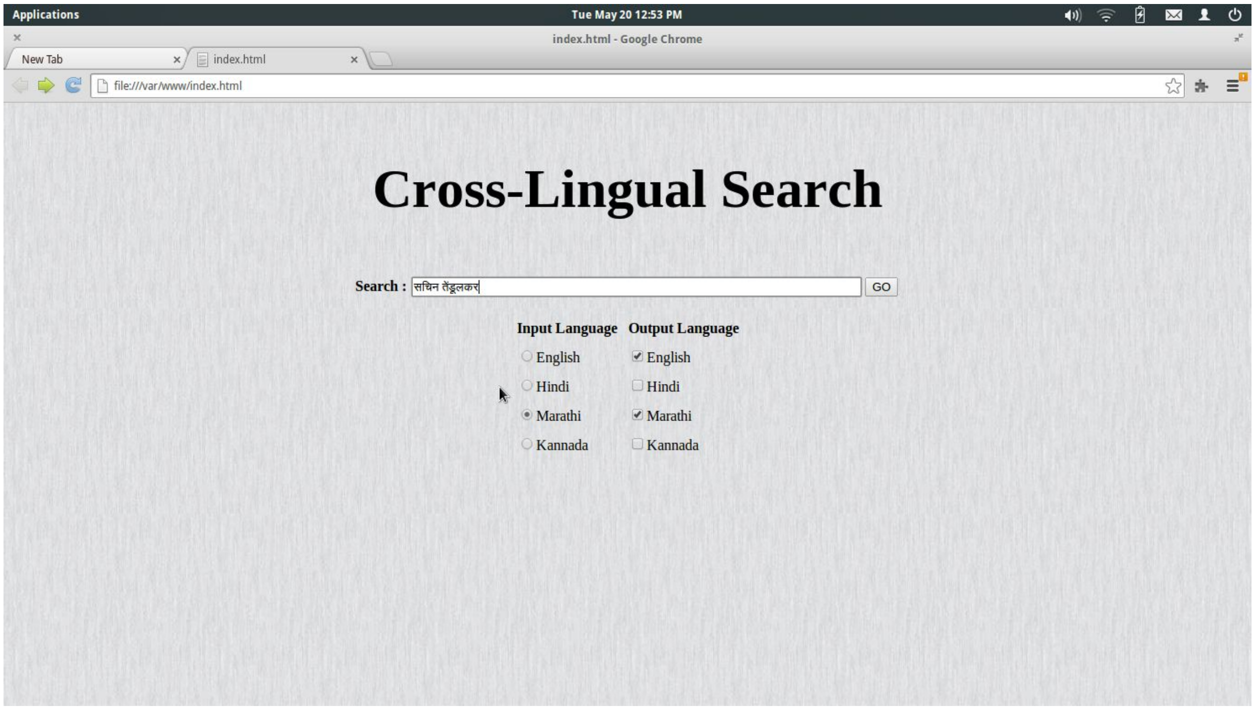
Hindi

महाभारत - विकिपीडिया
http://hi.wikipedia.org/wiki/%E0%A4%AE%E0%A4%B9%E0%A4%BE%E0%A4%AD%E0%A4%BE%E0%A4%B0%E0%A4%A4
श्रेणी:महाभारत - विकिपीडिया
http://hi.wikipedia.org/wiki/%E0%A4%B6%E0%A5%8D%E0%A4%B0%E0%A5%87%E0%A4%A3%E0%A5%80:%E0%A4%AE%E0%A4%B9%E0%A4%BE%E0%A4%AD%E0%A4%BE%E0%A4%B0%E0%E
महाभारत युद्ध के 10 गुप्त रहस्य - मुख पृष्ठ
http://hindi.webdunia.com/religion-sanatandharma-history/%E0%A4%AE%E0%A4%B9%E0%A4%BE%E0%A4%AD%E0%A4%BE%E0%A4%B0%E0%A4%A4-
%E0%A4%AF%E0%A5%81%E0%A4%A6%E0%A5%8D%E0%A4%A7-%E0%A4%95%E0%A5%87-10-%E0%A4%97%E0%A5%81%E0%A4%AA%E0%A5%8D%E0%A4%A4-
%E0%A4%B0%E0%A4%B9%E0%A4%B8%E0%A5%8D%E0%A4%AF-1130819003_1.htm
महाभारत - भारतकोश, ज्ञान का हिन्दी महासागर
http://hi.bharatdiscovery.org/india/%E0%A4%AE%E0%A4%B9%E0%A4%BE%E0%A4%AD%E0%A4%BE%E0%A4%B0%E0%A4%A4

Figure 8.6: Hindi results

8.5 Case 3

Input language: Marathi Output language: English and Marathi

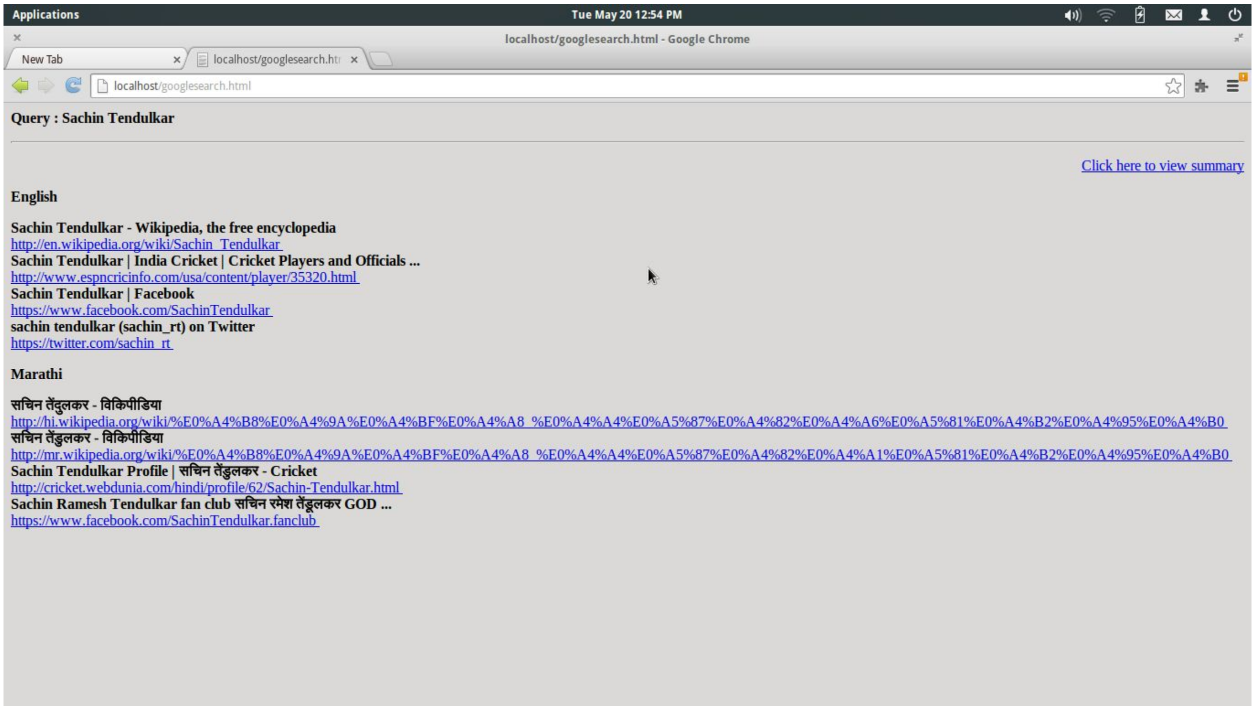


Cross-Lingual Search

Search : सचिन तेंडूलकर

Input Language	**Output Language**
○ English	☑ English
○ Hindi	□ Hindi
● Marathi	☑ Marathi
○ Kannada	□ Kannada

Figure 8.7: Marathi query


Applications
Tue May 20 12:54 PM
localhost/googlesearch.html - Google Chrome
New Tab
localhost/googlesearch.htr
localhost/googlesearch.html

Query : Sachin Tendulkar

Click here to view summary

English

Sachin Tendulkar - Wikipedia, the free encyclopedia
http://en.wikipedia.org/wiki/Sachin_Tendulkar
Sachin Tendulkar | India Cricket | Cricket Players and Officials ...
http://www.espncricinfo.com/usa/content/player/35320.html
Sachin Tendulkar | Facebook
https://www.facebook.com/SachinTendulkar
sachin tendulkar (sachin_rt) on Twitter
https://twitter.com/sachin_rt

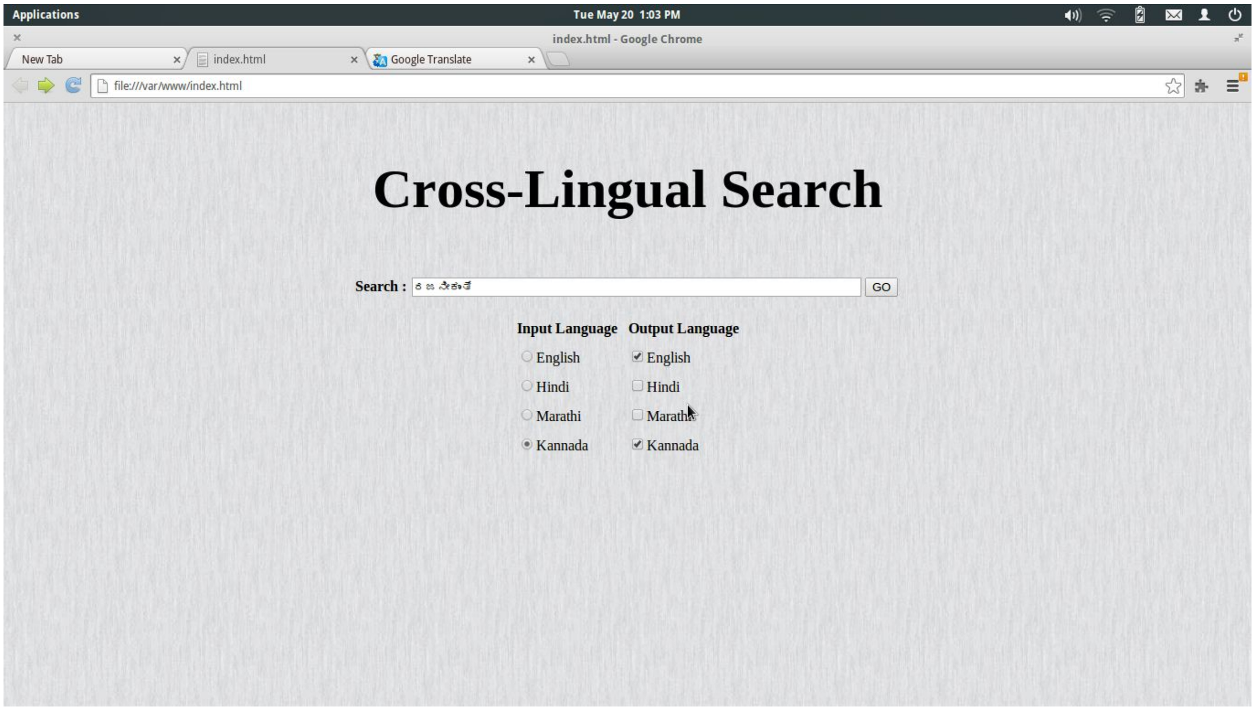
Marathi

सचिन तेंदुलकर - विकिपीडिया
http://hi.wikipedia.org/wiki/%E0%A4%B8%E0%A4%9A%E0%A4%BF%E0%A4%A8_%E0%A4%A4%E0%A5%87%E0%A4%82%E0%A4%A6%E0%A5%81%E0%A4%B2%E0%A4%95%E0%A4%B0
सचिन तेंडुलकर - विकिपीडिया
http://mr.wikipedia.org/wiki/%E0%A4%B8%E0%A4%9A%E0%A4%BF%E0%A4%A8_%E0%A4%A4%E0%A5%87%E0%A4%82%E0%A4%A1%E0%A5%81%E0%A4%B2%E0%A4%95%E0%A4%B0
Sachin Tendulkar Profile | सचिन तेंडुलकर - Cricket
http://cricket.webdunia.com/hindi/profile/62/Sachin-Tendulkar.html
Sachin Ramesh Tendulkar fan club सचिन रमेश तेंडूलकर GOD ...
https://www.facebook.com/SachinTendulkar.fanclub

Figure 8.8: Marathi results

8.6 Case 4

Input language: Kannada Output language: English and Kannada

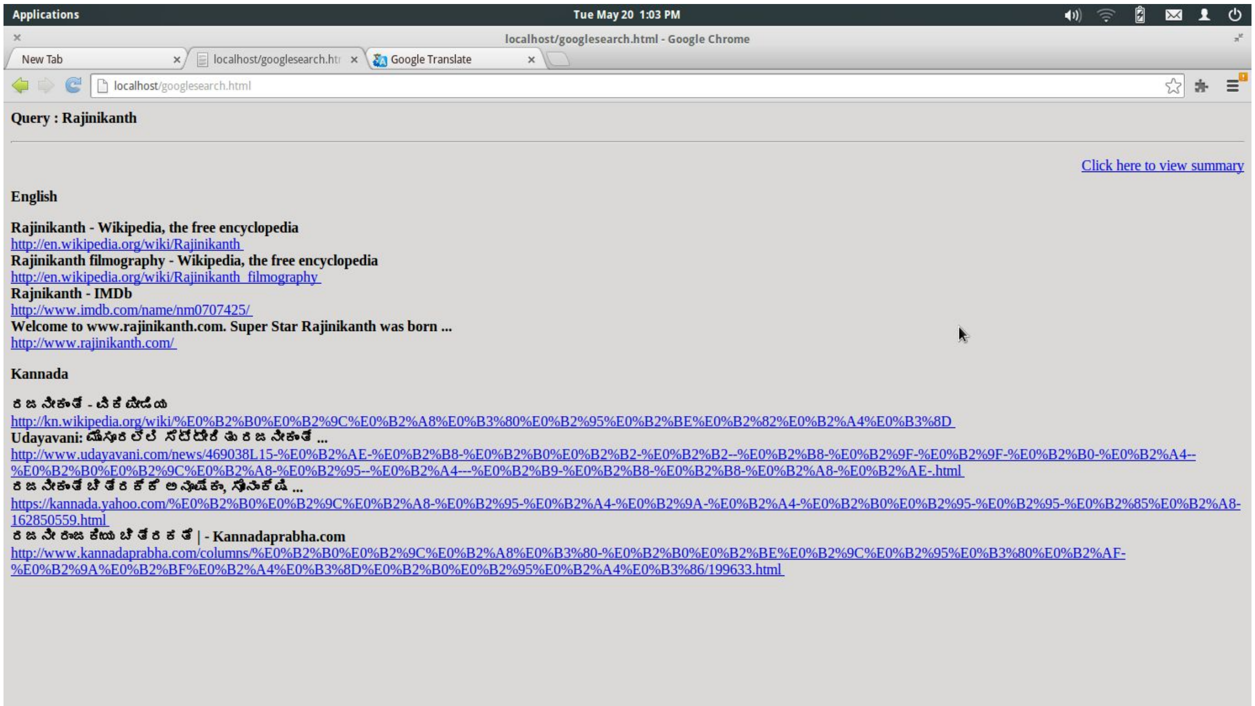


Cross-Lingual Search

Search : ರ ಜ ನೀಕಾಂತ್ GO

Input Language	Output Language
○ English	☑ English
○ Hindi	□ Hindi
○ Marathi	□ Marathi
● Kannada	☑ Kannada

Figure 8.9: Kannada query


Applications
Tue May 20 1:03 PM
localhost/googlesearch.html - Google Chrome
New Tab
localhost/googlesearch.ht
Google Translate
localhost/googlesearch.html

Query : Rajinikanth

Click here to view summary

English

Rajinikanth - Wikipedia, the free encyclopedia
http://en.wikipedia.org/wiki/Rajinikanth
Rajinikanth filmography - Wikipedia, the free encyclopedia
http://en.wikipedia.org/wiki/Rajinikanth_filmography
Rajnikanth - IMDb
http://www.imdb.com/name/nm0707425/
Welcome to www.rajinikanth.com. Super Star Rajinikanth was born ...
http://www.rajinikanth.com/

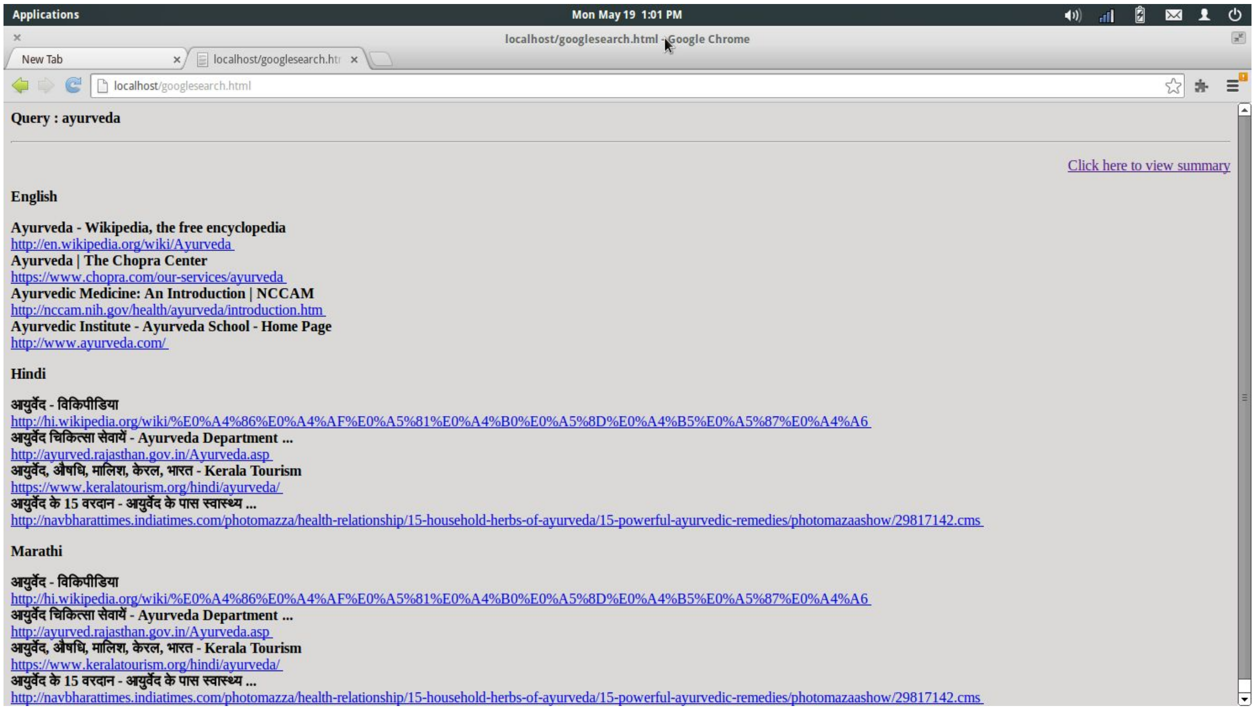
Kannada

ರ ಜ ನೀಕಾಂತ್ - ವಿಕಿಪೀಡಿಯ
http://kn.wikipedia.org/wiki/%E0%B2%B0%E0%B2%9C%E0%B2%A8%E0%B3%80%E0%B2%95%E0%B2%BE%E0%B2%82%E0%B2%A4%E0%B3%8D
Udayavani: ಮೈಸೂರ ಲ್ಲಿ ಸೆಟ್ಟೇರಿ ತು ರ ಜ ನೀಕಾಂತ್ ...
http://www.udayavani.com/news/469038L15-%E0%B2%AE-%E0%B2%B8-%E0%B2%B0%E0%B2%B2-%E0%B2%B2--%E0%B2%B8-%E0%B2%9F-%E0%B2%9F-%E0%B2%B0-%E0%B2%A4--%E0%B2%B0%E0%B2%9C%E0%B2%A8-%E0%B2%95--%E0%B2%A4---%E0%B2%B9-%E0%B2%B8-%E0%B2%B8-%E0%B2%A8-%E0%B2%AE-.html
ರ ಜ ನೀಕಾಂತ್ ಚಿ ತ್ರ ಕ್ಕೆ ಅ ನುಷ್ಕಾ, ಸೊನಾಕ್ಷಿ ...
https://kannada.yahoo.com/%E0%B2%B0%E0%B2%9C%E0%B2%A8-%E0%B2%95-%E0%B2%A4-%E0%B2%9A-%E0%B2%A4-%E0%B2%B0%E0%B2%95-%E0%B2%95-%E0%B2%85%E0%B2%A8-162850559.html
ರ ಜ ನೀ ರಾಜ ಕೀಯ ಚಿ ತ್ರ ಕ ತೆ | - Kannadaprabha.com
http://www.kannadaprabha.com/columns/%E0%B2%B0%E0%B2%9C%E0%B2%A8%E0%B3%80-%E0%B2%B0%E0%B2%BE%E0%B2%9C%E0%B2%95%E0%B3%80%E0%B2%AF-%E0%B2%9A%E0%B2%BF%E0%B2%A4%E0%B3%8D%E0%B2%B0%E0%B2%95%E0%B2%A4%E0%B3%86/199633.html

Figure 8.10: Kannada results

8.7 Case 5

Input language: English Output language: All four languages


Applications
Mon May 19 1:01 PM
localhost/googlesearch.html - Google Chrome
New Tab
localhost/googlesearch.htr
localhost/googlesearch.html

Query : ayurveda

Click here to view summary

English

Ayurveda - Wikipedia, the free encyclopedia
http://en.wikipedia.org/wiki/Ayurveda
Ayurveda | The Chopra Center
https://www.chopra.com/our-services/ayurveda
Ayurvedic Medicine: An Introduction | NCCAM
http://nccam.nih.gov/health/ayurveda/introduction.htm
Ayurvedic Institute - Ayurveda School - Home Page
http://www.ayurveda.com/

Hindi

आयुर्वेद - विकिपीडिया
http://hi.wikipedia.org/wiki/%E0%A4%86%E0%A4%AF%E0%A5%81%E0%A4%B0%E0%A5%8D%E0%A4%B5%E0%A5%87%E0%A4%A6
आयुर्वेद चिकित्सा सेवायें - Ayurveda Department ...
http://ayurved.rajasthan.gov.in/Ayurveda.asp
आयुर्वेद, औषधि, मालिश, केरल, भारत - Kerala Tourism
https://www.keralatourism.org/hindi/ayurveda/
आयुर्वेद के 15 वरदान - आयुर्वेद के पास स्वास्थ्य ...
http://navbharattimes.indiatimes.com/photomazza/health-relationship/15-household-herbs-of-ayurveda/15-powerful-ayurvedic-remedies/photomazaashow/29817142.cms

Marathi

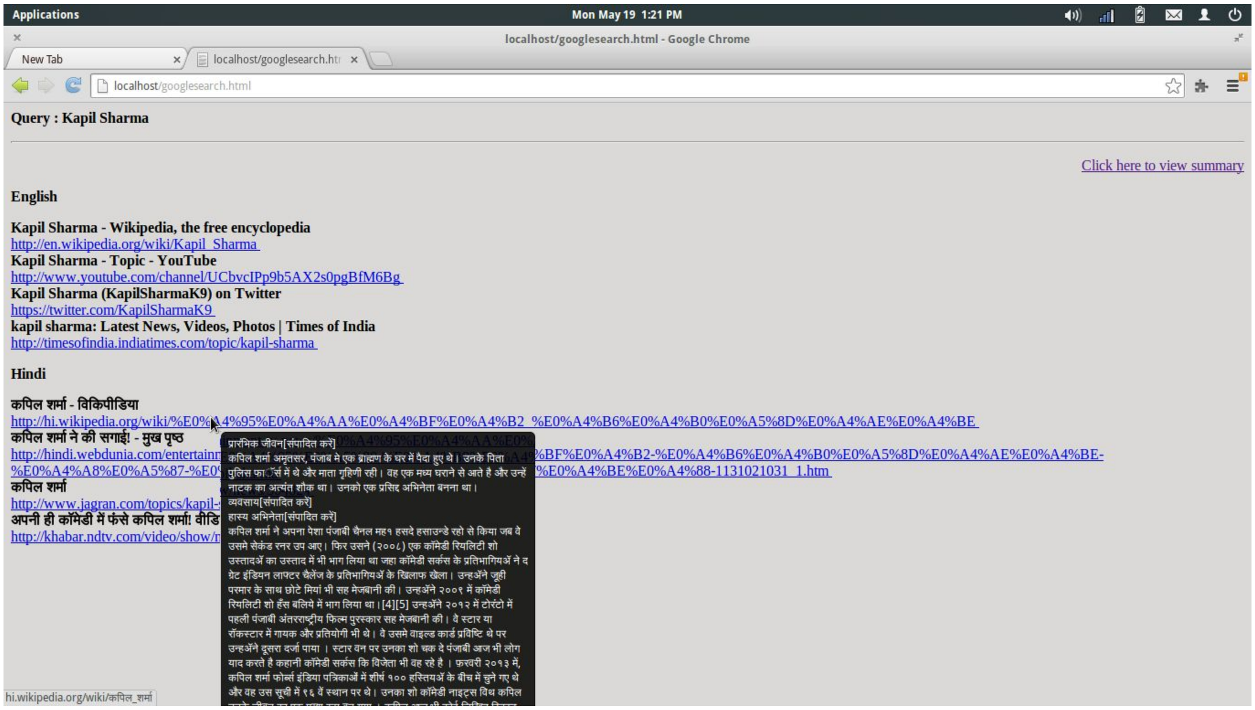
आयुर्वेद - विकिपीडिया
http://hi.wikipedia.org/wiki/%E0%A4%86%E0%A4%AF%E0%A5%81%E0%A4%B0%E0%A5%8D%E0%A4%B5%E0%A5%87%E0%A4%A6
आयुर्वेद चिकित्सा सेवायें - Ayurveda Department ...
http://ayurved.rajasthan.gov.in/Ayurveda.asp
आयुर्वेद, औषधि, मालिश, केरल, भारत - Kerala Tourism
https://www.keralatourism.org/hindi/ayurveda/
आयुर्वेद के 15 वरदान - आयुर्वेद के पास स्वास्थ्य ...
http://navbharattimes.indiatimes.com/photomazza/health-relationship/15-household-herbs-of-ayurveda/15-powerful-ayurvedic-remedies/photomazaashow/29817142.cms

Figure 8.11: Result page

The snapshot shows the result page. The output languages selected were English, Hindi, Marathi and Kannada

8.8 Summary

Figure 8.12: Summary in English



Applications
Mon May 19 1:21 PM
localhost/googlesearch.html - Google Chrome
New Tab
localhost/googlesearch.htr
localhost/googlesearch.html

Query : Kapil Sharma

Click here to view summary

English

Kapil Sharma - Wikipedia, the free encyclopedia
http://en.wikipedia.org/wiki/Kapil_Sharma
Kapil Sharma - Topic - YouTube
http://www.youtube.com/channel/UCbvcIPp9b5AX2s0pgBfM6Bg
Kapil Sharma (KapilSharmaK9) on Twitter
https://twitter.com/KapilSharmaK9
kapil sharma: Latest News, Videos, Photos | Times of India
http://timesofindia.indiatimes.com/topic/kapil-sharma

Hindi

कपिल शर्मा - विकिपीडिया
http://hi.wikipedia.org/wiki/%E0%A4%95%E0%A4%AA%E0%A4%BF%E0%A4%B2_%E0%A4%B6%E0%A4%B0%E0%A5%8D%E0%A4%AE%E0%A4%BE
कपिल शर्मा ने की सगाई! - मुख पृष्ठ
http://hindi.webdunia.com/entertainm … A4%BF%E0%A4%B2-%E0%A4%B6%E0%A4%B0%E0%A5%8D%E0%A4%AE%E0%A4%BE-
%E0%A4%A8%E0%A5%87-%E0 … %E0%A4%BE%E0%A4%88-1131021031_1.htm
कपिल शर्मा
http://www.jagran.com/topics/kapil-
अपनी ही कॉमेडी में फंसे कपिल शर्मा! वीडि
http://khabar.ndtv.com/video/show/r

प्रारंभिक जीवन[संपादित करें]
कपिल शर्मा अमृतसर, पंजाब मे एक ब्राह्मण के घर में पैदा हुए थे। उनके पिता पुलिस फा र्स में थे और माता गृहिणी रही। वह एक मध्य घराने से आते है और उन्हें नाटक का अत्यंत शौक था। उनको एक प्रसिद्द अभिनेता बनना था।
व्यवसाय[संपादित करें]
हास्य अभिनेता[संपादित करें]
कपिल शर्मा ने अपना पेशा पंजाबी चैनल मह१ हसदे हसाउन्डे रहो से किया जब वे उसमे सेकंड रनर उप आए। फिर उसने (२००८) एक कॉमेडी रियलिटी शो उस्तादअें का उस्ताद में भी भाग लिया था जहा कॉमेडी सर्कस के प्रतिभागियअें ने द ग्रेट इंडियन लाफ्टर चैलेंज के प्रतिभागियअें के खिलाफ खेला। उन्हअेंने जूही परमार के साथ छोटे मियां भी सह मेजबानी की। उन्हअेंने २००९ में कॉमेडी रियलिटी शो हँस बलिये में भाग लिया था।[4][5] उन्हअेंने २०१२ में टोरंटो में पहली पंजाबी अंतरराष्ट्रीय फिल्म पुरस्कार सह मेजबानी की। वे स्टार या रॉकस्टार में गायक और प्रतियोगी भी थे। वे उसमे वाइल्ड कार्ड प्रविष्टि थे पर उन्हअेंने दूसरा दर्जा पाया । स्टार वन पर उनका शो चक दे पंजाबी आज भी लोग याद करते है कहानी कॉमेडी सर्कस कि विजेता भी वह रहे है । फ़रवरी २०१३ में, कपिल शर्मा फोर्ब्स इंडिया पत्रिकाओं में शीर्ष १०० हस्तियअें के बीच में चुने गए थे और वह उस सूची में १६ वें स्थान पर थे। उनका शो कॉमेडी नाइट्स विथ कपिल

Figure 8.13: Summary in Hindi

CONCLUSION

This project provides Cross lingual information retrieval feature to the users. This feature is provided to the user through an extension which has to be added to the browser through the Google App store. The extension provides an interface where the user can enter the query in one of the four specified languages, that are English, Hindi. Kannada and Marathi and can retrieve the output in any of these languages with English being the default output language. The results of the search are prioritized and the top four results in each selected output language are displayed. The application also provides the user with the summery of each search result.

This project can further be extended to support various Indian and foreign languages. The search can be enhanced by performing meta search and can also be built on platforms like iOS ,android etc.

REFERENCES

[1] Herbert Schildt,“*Python-The Complete Reference*”, 8th Edition, Pearson Education.

[2] Learning Python, 5th Edition, O’Reilly Media.

[3] Steven Holzner,“*A Beginner’s Guide Ajax*”, Mc Graw Hill 2009.

[4] “*Cross lingual information access from marathi to english”,* CLIA IIT Bombay group, Mumbai.2012

[5] “*Turkish – English Cross Language Information Retrieval using LSI*”, Erbuğ Çelebi, Baturman Şen, Burak Günel.

[6] “*Different approaches to Cross Language Information Retrieval*”, Wessel Kraaij and Ren´ee Pohlmann.

[7] *“Cross – lingual latent semantic analysis for language maodelling”* Woosung Kim and Sanjeev Khudanpur

[8] “*Domain-Specific Query Translation for Multilingual Information Access using Machine Translation Augmented with Dictionaries Mined from Wikipedia*”, Gareth J. F. Jones, Fabio Fantino, Eamonn Newman, Ying Zhang

[9] “*Bilingual terminology acquisition from comparable corpora and phrasal translation to CLIR*” Fatiha Sadat, Masatoshi Yoshikawa, Shunsuke Uemura

[10] “*A Survey on the Cross and Multilingual Information Retrieval*” N.Swapna , Padmaja Rani ,Kiran Kumar

[11] “*Dictionary-Based Cross-Language Information Retrieval: Problems, Methods, and Research Findings*”. Ari Pirkola, Turid Hedlund

[13] “*Comparing different approaches to treat Translation Ambiguity in CLIR: Structured Queries vs. Target Co-occurrence Based Selection*” Xabier Saralegi and Maddalen López de Lacalle

[14] “*Transliteration of Proper Names in Cross-Lingual Information Retrieval*” Paola Virga, Sanjeev Khudanpur

[15] https://developers.google.com/ - Google developer guide for translation, transliteration and serach.