ADAMA SCIENCE & TECHNOLOGY UNIVERSITY

SCHOOL OF ELECTRICAL ENGINEERING & COMPUTING

DEPARTMENT OF COMPUTING

A MASTER'S THESIS
ON
INTERACTIVE AMHRARIC VOISE BASED WEB SEARCH WITH ENHANCED PAGE RANKING ALGORITHM FOR VISUALLY IMPAIERD

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENT FOR THE DEGREE OF MASTER OF SCIENCE IN SOFTWARE ENGINEERING

By:

Awel Negash

January 2018

Adama, Ethiopia

ADAMA SCIENCE & TECHNOLOGY UNIVERSITY

SCHOOL OF ELECTRICAL ENGINEERING & COMPUTING

DEPARTMENT OF COMPUTING

INTERACTIVE AMHRARIC VOISE BASED WEB SEARCH WITH ENHANCED PAGE RANKING ALGORITHM FOR VISUALLY IMPAIERD

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENT FOR THE DEGREE OF MASTER OF SCIENCE IN SOFTWARE ENGINEERING

By: Awel Negash

Advisor: Dr. Wazih Ahmed (Ph.D.)

January 2018

Adama, Ethiopia

ADAMA SCIENCE & TECHNOLOGY UNIVERSITY
SCHOOL OF GRADUATE STUDIES
DEPARTMENT OF COMPUTING

INTERACTIVE AMHRARIC VOISE BASED WEB SEARCH WITH ENHANCED PAGE RANKING ALGORITHM FOR VISUALLY IMPAIERD

Name and signature of members of the examining board

Name	Signature	Date
_____	_____	_____
Advisor		
_____	_____	_____
External Examiner		
_____	_____	_____
Internal Examiner		

Declaration

I declare that this study is my original work submitted in partial fulfillment of the requirement for the degree of Master of Science in software engineering and has not been submitted for any Degree or Diploma in any University or conference. To the best of my knowledge, all source of materials used for the study has been duly acknowledged. I have undertaken the study with the guidance and support of the research advisor.

Awel Negash

Signature:_____

Advisor: Dr. Wazih Ahmed

Signature:_____

January 26, 2018

Acknowledgement

This thesis represents the peak point of my studies, so far, and it incorporates knowledge gathered from different sources, academics and professionals alike. First of all, I would like to express my deepest gratitude to the almighty God for his mercy, guidance, support; Glory to God. Next, I would like to express my sincere gratitude to my research advisor Dr. Wazih Ahmed for his advice, patience and constructive criticisms of this work. Thirdly, my thanks goes to all interviewees from Bahirdar University.

Finally, I would like to express my thanks to my friend Filipo Engidashet for his support in developing the plug-in.

Table of Contents

List of Tables

List of Figures

List of Acronyms

IR: Information Retrieval

EWPR: Enhanced Weighted Page Rank

WF: Weight factor

HITS: Hyperlink-Induced Topic Search

SALSA: Stochastic Approach for. Link Structure Analysis

CSA: Central Statistical Agency

WWW: World Wide Web

WAI: Web Accessibility Initiative

SEO: Search engine optimization

W3C: World Wide Web Consortium

XHTML: Extensible HyperText Markup Language

HTML: HyperText Markup Language

URL: Uniform Resource Locator

ICT: Information and communication technology

PHP: PHP

JAWS: Job Access With Speech

IT: Information Technology

SSP: Speech to Text

API: Application Program Interface

JSON: JavaScript Object Notation

IDE: Integrated development environment

JEE: Java Enterprise Edition:

AJAX: Asynchronous JavaScript And XML

TCP: Transmission Control Protocol

TP: True positive value

FP: False positive value

HMM: Hidden Markov Models

Abstract

Nowadays, web-based technologies influence in every ones daily life such as in education, business, entertainment and others. For the purpose of making web application impressive web developers are too eager to develop their web applications with fully animation graphics and forgetting its accessibility to its users. Unfortunately, most web sites are inaccessible to many disabled people and fail to satisfy the most basic standards for accessibility. Thus, this paper would highlight on the usability and accessibility of Amharic voice recognition browser plug-in as a tool to facilitate the visually impaired and blind learners in accessing virtual learning environment. . More importantly to enable visually impaired user access the relevant web page by enhancing the existing content based relevancy algorithm so that web pages that are relevant and which can be easily synthetized to voice will be ranked in priority. Specifically, the objectives of the study are (i) to explore the challenges faced by the visually impaired learners in accessing virtual learning environment (ii) to determine the suitable guidelines for developing Amharic voice recognition browser plug-in that is accessible to the visually impaired. Furthermore, this study was prepared based on an observation conducted with the Bahirdar University visually impaired learners. Finally, the result of this study would underline on the development of an accessible Amharic voice recognition browser plug-in for the visually impaired.

Keywords—Accessibility, Usability, Virtual Learning, Visually Impaired, Voice Recognition.

CHAPTER ONE

1. INTRODUCTION

In this chapter the background of the research and the rationale initiated to do this research has been presented. It presents the overall objective, methodology, significance of the study, scope, and limitation of the thesis and problem that has to be addressed and finally the organization of the thesis is presented.

1.1. Background

Since the Internet was introduced in 1969, it has grown rapidly and is expected to continue to do so over years. The World Wide Web is a very popular and interactive information resource acting as a storehouse for image, text, audio, video, and metadata. The amount of information available on Web is very huge and noisy. The noise comes from two major sources. First, an emblematic Web page contains many pieces of information, e.g., the main content of the page, advertisements, copyright notices, routing links, privacy policies, etc. Second, due to the fact that the Web does not have quality control of information, i.e., one can write anything that one likes. The popularity of WWW is largely dependent on the search engines. A web search engine is a software system that is designed to search for information on the World Wide Web. Now anyone can quickly search for helpful direction tips, personal information, recipes, vacancy, pictures, organizational websites and more with search engines. Search engines consist of four discrete software components: Crawler or Spider: This is the part of the search engine which combs through the pages on the internet and gathers the information for the search engine.; Indexer: a blender like program that dissects web pages that are downloaded by spiders; The database: The search engine's database is what you are actually searching. All of the information that a web crawler retrieves is stored in a database. Every time you use a search engine, it is this database you are searching, not the live internet

This vast amount of data is making search more and more difficult with traditional search engine as they return huge data for a given query which is consisting of relevant as well as irrelevant data. [1][2] This not only results in wastage of user time but also leads to data overload problem. So, users are not satisfied with searching the information by traditional search engine. So the problem of re-ranking search pages or results has become one of the main problems in IR field.

Exactly what information the user wants is unpredictable. So the web page ranking algorithms are designed to anticipate the user requirements from various static (e.g., number of hyperlinks, textual content) and dynamic (e.g., popularity) features.

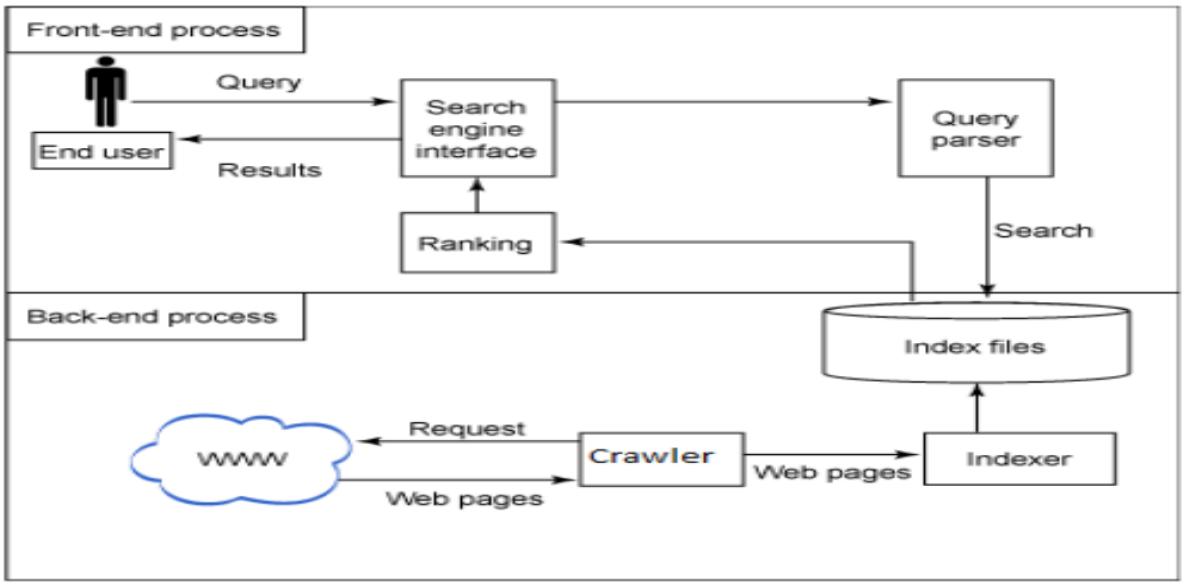

Figure 1: Simplified Search Engine Architecture

The Various ranking algorithms developed for ranking web pages are Page Rank, Weighted Page rank, Page Content Rank, HITS, SALSA, SUBSPACE HITS, SIMRANK etc. Most of these algorithms are either based on web structure mining or web content mining. Web content mining extracts useful information from the content of web documents whereas web structure mining is used to set links between references and referents in the web [3]. In this paper, Enhanced Weighted Page Rank (EWPR) is being proposed for search engines that works on the basis of Weighted Page Rank algorithm and takes into account Weight factor (WF). The important purpose of this proposed algorithm is to find more relevant information as per the queries of the user [4].

There is one group of users, however, that have been largely ignored in the rush to use the WWW. These are blind or visually impaired users, who have particular problems in accessing web material. Whereas accessibility issues to sighted users may be a matter of response time, getting lost in cyberspace, or less performance of search engine retrieval, these are issues that may be inconvenient, but they are not challenging problems. The considerations are not the same for a blind or visually impaired user trying to access the same information. Problems of accessibility to a person who is visually impaired covers all those for sighted users, plus a number that are unique

to this group of users. These include the issue of screen design, the use of font size, color, the use of these features, designed to be appealing to the sighted user may make Internet pages inaccessible to a visually impaired user

1.2. Statement of the problem

According to Central Statistical Agency (CSA) of Ethiopia there are about 800,000 blinds and more than 1.4 low visions learners which are deprived of the Internet. The visually impaired learners are left to learn using the conventional method of "talk and Braille" by the teacher. They cannot use the functionalities like "linking" of web pages through the available browsers with their local language to acquire the various information and knowledge from the web. The visually impaired learners are also deprived of enjoying services in the Internet like sending and receiving e-mails unlike their other normal friends.

Most importantly due to rapid development of the internet and exponential growth of information amount it has been difficult for the visually impaired to search the relevant information from search engines. The web has become difficult for users to extract and filter the information that is more relevant. Since the visually impaired users try to access the internet through voice recognition mechanism and search results are provided to them by converting the contents of the pages to voice and reading the result to them. Because of this result from search engines should have to be re-ranked by giving a priority to those relevant pages and page that are not full of graphic contents.

Discussing on the capabilities of the World Wide Web (WWW) to everyone's daily activities especially in online education, there is still some limitation to people with disabilities, especially the visually impaired learners to access information. To avoid social exclusion for people with disability, web accessibility is a requirement for websites. Today, there are large numbers of websites which fail to meet the requirements of web accessibility. Therefore, this research would focus on the accessibility related to the development of Amharic voice browser plug-in that facilitates the visually impaired in seeking information via internet.

1.3. The Objectives of the Study

1.3.1. General objective

The objective of this paper is to focus on a study of the usability and accessibility topic for developing Amharic voice-based browser plug-in as an assistive tool that facilitates the visually

impaired in seeking information via Internet. More importantly to enable visually impaired user access the relevant web page by enhancing the existing content based relevancy algorithm so that web pages that are relevant and which can be easily synthetized to voice will be ranked in priority. By developing this interactive Amharic voice recognizer browser Plug-in and integrating the enhanced algorithm with the Plug-in to facilitates the visually impaired learners accessing information through the Internet as a part of accessing to virtual learning

1.3.2. Specific objectives

- ✓ To explore the challenges faced by the visually impaired learners in accessing virtual learning environment.
- ✓ To design and develop a browser plug-in that enables them to browse the Internet through Amharic and English voice recognition system.
- ✓ To enhance existing content based web page re-ranking algorithm to re-rank search results from search engines which are accessible and relevant to the visually impaired users.
- ✓ This browser plug-in is an accessible browser plug-in that allows them to navigate the Internet with less complexity by using a medium of speech for alternative input and output.

1.4. Research Specific Questions

The proposed research shall implicitly discuss the following exploratory questions like:

- ✓ Whether the application of Amharic voice-based web searching concept can be replacement to the conventional method of learn using “talk and Braille”.
- ✓ How accurate the system recognizes Amharic words?

The research questions for the proposed research are:

(i) How can we evaluate the performance of proposed enhanced algorithm in terms of search results?

(ii) Whether the enhanced Content Based Web Page Re-Ranking Using Relevancy Algorithm provide the relevant page to the visually impaired users.

(iii) Whether Amharic voice-based web searching is useful in filling the gap between the visually impaired and normal web users.

1.5. Theoretical / Conceptual Framework

The theoretical framework for the study was constituted by quite a large amount of experience reports on web page re-ranking algorithms. Reviewing the literature, it is found that there was a mismatch between the page rank provided by search engines and the interests of the visually impaired users. The companies work with general page ranking algorithms to rank web page in the World Wide Web their ranking, which are aimed with manual user quires rather than interest of specific user. First much of existing search engines focus on manual keyboard query insertion for searching and the results from searching are not relevant for visually impaired users since this users access the result as voice output using voice synthetize. Secondly even if there are some search engines with voice searching, the results or the ranking of the search engines are the same with manual key board searching which are intended for normal users.

1.6. Scope and Limitation of the Study

The scope of this research is enhancing the existing content based relevancy based web page re-ranking algorithm and developing interactive Amharic voice-based browser Plug-in by integrating with the algorithm as an assistive tool that facilitates the visually impaired in seeking information via Internet, which comprises of recognizing voice in Amharic, process the recognized token or commands, searching from Google search engine and re-rank search results by the re-ranking application, get questions and answer as artificial intelligence and finally read the search result for impaired users

This browser plug-in is only compatible with Google chrome browser and run only on windows operating system. The plug-in supports only Amharic and English languages. This research will contribute new way of web page re-ranking for visually impaired internet users with Amharic voice search.

1.7. Significance and Contribution of the Study

Page ranking algorithm has a potential to extract useful pages or documents from huge collection of data from the web by fulfilling the need of web users. The beneficiaries of this research are visual impaired and low vision internet users. This research enable the visually impaired users to access the internet by their voice and get relevant information from the internet by using the proposed page re-ranking algorithm and it will fill the gap between visual impaired and normal

internet users in digital learning environment by getting relevant information with new way of re-ranking search results from search engines. Local visual impaired Amharic voice speakers will access the internet through interactive Amharic voice recognizer plugin to surf the internet.

The proposed web page re-ranking algorithm can be used by other researches for further researching with other language across the world and more for multi lingual voice based searching.

1.8. Organization of paper

The rest of the study is organized as follows: Chapter two covered related literatures on the basic concepts of page ranking algorithm: different types of page ranking algorithms, comparison of existing algorithms using different metrics , advantages and disadvantages of existing algorithms ,issues of webpage accessibility and literature review on existing speech recognition technology's related works done on page ranking algorithms. Chapter three describes the research methodology and strategy that aims to identify the potential problem the visually impaired users face to access the internet. It begins by describing exploratory research approaches. Primary data collection and analysis technique were described as information acquisition method. Validity and reliability requirements also identified. In Chapter four we have discussed backgrounds of system framework and architectural design of the proposed algorithm. Chapter five presents the results of the interview with techniques .In chapter six the implementation of the prototype and tools used to develop the browser plug-in have been discussed. Finally, in chapter seven conclusions about the research and suggestions for future research direction were presented.

CHAPTER TWO

2. LITERATURE REVIEW

To understand the internet and its accessibility for the visually impaired internet users this study conducted a literature review on three basic areas which are web accessibility, existing page ranking algorithm and voice recognition based technologies.

2.1. Web Accessibility

The Web is fundamentally designed to work for all people, beside their software, hardware, culture, location, language, or physical or mental ability. When the Web meets this goal, it is accessible to people with a diverse range of hearing, movement, sight, and cognitive ability. [15]More specifically, Web accessibility means that people with disabilities can explore, perceive, understand, navigate, and interact with the Web, and that they can contribute to the Web. Web accessibility also benefits others, including older people with changing abilities due to aging.

[16]Web Accessibility in its most basic definition is about making sure websites work for the widest possible audience. For most people, it is easy to browse the web, they can point and click, visually skip over content they don't want to read, listen and watch a video clip, and skim for what they are looking for. For those with disabilities, all of these things can be barriers to access if we don't use the code, the methods, inherently provided by the creators of the web to ensure that it would be usable by all. We have the obligation to make sure our web presence works in the most inclusive and equal way possible. There is a slowly dying myth that accessible websites will be boring, dull and void of interesting features. This couldn't be further from the truth. In fact, almost all of the features of an accessible website are underneath, in the code, meaning that you can make your website accessible with no to very little changes to the look of your site.

The web is providing unprecedented access to information, interaction, goods and services. With so much of our daily lives now being spend interacting with technology and web content through things such as; watching movies, purchasing goods, registering for classes, buying event tickets, having class discussions, and communicating with others, having internet access is no longer a privilege but a right. We must be diligent, and everyone must have access in order for this right to be achieved. Without it, how will students with disabilities compete with other normal students?

The mission of the Web Accessibility Initiative (WAI) is to lead the Web to its full potential to be accessible, enabling people with disabilities to participate equally on the Web.

Web accessibility encompasses all disabilities that affect access to the Web, including visual, auditory, and physical, speech, cognitive, and neurological disabilities. Millions of people have disabilities that affect their use of the Web. Nowadays most Web sites and Web software have accessibility issues that make it difficult or impossible for many people with disabilities to use the Web. As more accessible Web sites and software become available, people with disabilities are able to use and contribute to the Web more effectively.

Web accessibility also benefits people without disabilities. For example, a key principle of Web accessibility is designing Web sites and software that are flexible to meet different user needs, preferences, and situations. This flexibility also benefits people *without* disabilities in certain situations, such as people using a slow Internet connection, people with "temporary disabilities" such as a broken arm, and people with changing abilities due to aging.

I. Who is affected by Inaccessibility?

All of us can be affected by inaccessible websites, however, individuals with disabilities are most affected. Most of the literature on accessible web design refer to how to make content accessible for individuals who are blind. We can understand that these individuals are very sensitive to think about, but are not the only ones affected by inaccessible web design. Web accessibility encompasses all disabilities.

There are four main types of disabilities that can be used to explain how we should be thinking about accessibility.

2. Individuals with Visual Disabilities

Many individuals who are blind interact with computers using screen reader software. Screen readers turn content into an easier to follow linear format. This is an important concept. Since individuals who are blind cannot browse content the way sighted individuals do - by visually scanning and finding the relevant

information - there needs to be a way to express the content from point to point. Screen readers do this, but they cannot do it alone, using the policy and guidance provided, you will learn how to create content in a way that included individuals who are blind.

Individuals who have low-vision, who are colorblind or who have photosensitivity issues also can be affected by inaccessible content. The use of appropriate colors, good contrast, and consistent layout will help many of these individuals. Many of these individuals interact with computers and websites without any assistive technology, however, some use magnifiers, speech recognition software and/or increased contrast and zoom.

3. Individuals with Mobility Disabilities

Keyboard access is the most important concept when thinking about the accessibility of content for individuals with mobility disabilities. Some people do not have use of, or do not have arms, hands or fingers. Additionally, many other individuals have limited control of their arms, others have diminishing fine motor controls. All of these individuals might have difficulty using a mouse, many only use the keyboard to navigate. Others use items such as trackballs, adaptive keyboards, headwands, mouthsticks, and speech recognition software.

4. Individuals with Hearing Disabilities

An often overlooked area of web accessibility are the needs of individuals who are deaf or hard of hearing. With the amount of multimedia, audio, and video content on websites growing daily, these individuals are often left out of full participation because of inaccessible materials.

Providing captions and transcripts are most important for these individuals, and must be provided on all publicly available content. To not do so will leave out a valued and large group within the OSU community.

5. Individuals with Cognitive or Mental Disabilities

For individuals with cognitive disabilities such as learning disabilities, distractibility, comprehension, and dyslexia, content is the most important barrier to accessibility and their ability to interact with websites.

These individuals benefit more from well structured, semantically organized pages that provide instructions, illustrations, diagrams or any process that helps make the content easier to understand and navigate.

II. Why Web Accessibility is Important

The Web is an increasingly important resource in many aspects of life: education, employment, government, commerce, health care, recreation, and more. It is essential that the Web be accessible in order to provide equal access and equal opportunity to people with diverse abilities. Indeed, the UN Convention on the Rights of Persons with Disabilities recognizes access to information and communications technologies, including the Web, as a basic human right.

Accessibility supports social inclusion for people with disabilities as well as others, such as older people, people in rural areas, and people in developing countries. There is also a strong business case for accessibility**.** Accessibility overlaps with other best practices such as mobile web design, device independence, multi-modal interaction, usability, design for older users, and search engine optimization (SEO). Case studies show that accessible websites have better search results, reduced maintenance costs, and increased audience reach, among other benefits. Developing a Web Accessibility Business Case for Your Organization details the social, technical, financial, and legal benefits of web accessibility.

III. Making Your Web Site Accessible

Much of the focus on Web accessibility has been on the responsibilities of Web developers. However, Web software also has a vital role in Web accessibility. Software needs to help developers produce and evaluate accessible Web sites, and be usable by people with disabilities.

The W3C Web Accessibility Initiative (WAI) brings together people from industry, disability organizations, government, and research labs from around the world to develop guidelines and resources to help make the Web accessible to people with disabilities including auditory, cognitive, neurological, physical, speech, and visual disabilities. WAI's coverage of web accessibility includes 'web content' (websites and web applications), authoring tools (such as content management systems (CMS) and blog software), browsers and other 'user agents', and W3C technical specifications, including WAI-ARIA for accessible rich Internet applications. These WAI guidelines are considered the international standard for Web accessibility.

Making a Web site accessible can be simple or complex, depending on many factors such as the type of content, the size and complexity of the site, and the development tools and environment. Many accessibility features are easily implemented if they are planned from the beginning of Web site development or redesign. Fixing inaccessible Web sites can require significant effort, especially sites that were not originally "coded" properly with standard XHTML markup, and sites with certain types of content such as multimedia. The Web Content Accessibility Guidelines and techniques documents provide detailed information for developers. Accessibility is essential for developers and organizations that want to create high quality websites and web tools, and not exclude people from using their products and services.

IV. Examples of Web Accessibility

Properly designed websites and web tools can be used by people with disabilities. However, currently many sites and tools are developed with accessibility barriers that make it difficult or impossible for some people to use them. Below are just a few examples.

Alternative Text for Images

Alt text is the classic example. Images should include *equivalent alternative text* in the markup/code.

If alt text isn't provided for images, the image information is inaccessible, for example, to people who cannot see and use a screen reader that reads aloud the information on a page, including the alt text for the visual image.

When equivalent alt text is provided, the information is available to everyone to people who are blind, as well as to people who turned off images on their mobile phone to lower bandwidth charges, people in a rural area with low bandwidth who turned off images to speed download, and others. It's also available to technologies that cannot see the image, such as search engines.

Keyboard Input

Some people cannot use a mouse, including many older users with limited fine motor control. An accessible website does not rely on the mouse; it provides all functionality via a keyboard. Then people with disabilities can use assistive technologies that mimic the keyboard, such as speech input.

V. Evaluating the Accessibility of a Web Site

When developing or redesigning a site, evaluating accessibility early and throughout the development process can identify accessibility problems early when it is easier to address them. Simple techniques such as changing settings in a browser can determine if a Web page meets some accessibility guidelines. A comprehensive evaluation to determine if a site meets all accessibility guidelines is much more complex.

There are evaluation tools including tools and guides created by WebAIM (Web Accessibility in Mind), these tools can only provide you with basics, like, do your images all have alternative text (alt attributes) for images. These tools cannot fully analyze if the alt text is appropriately written for the image, as it cannot fully understand what the image is for. These tools also cannot check for everything, so don't be lured into the false assumption that if an automatic tool comes back with no issues found, that your website is accessible that help with evaluation. However, no tool alone can determine if a site meets accessibility guidelines. Knowledgeable human evaluation is required to determine if a site is accessible.

2.2. Page Ranking Algorithm

Since the early stages of the World Wide Web, search engines have developed different methods to rank web pages. Until today, the occurrence of a search query phrase within a document is one major factor within ranking techniques of virtually any search engine. The occurrence of a search phrase can thereby be weighted by the length of a document (ranking by keyword density) or by its accentuation within a document by HTML tags.

For the purpose of better search results and especially to make search engines resistant against automatically generated web pages based upon the analysis of content specific ranking criteria (doorway pages), the concept of link popularity was developed. Following this concept, the number of inbound links for a document measures its general importance. Hence, a web page is generally more important, if many other web pages link to it. The concept of link popularity often avoids good rankings for pages which are only created to deceive search engines and which don't have any significance within the web, but numerous webmasters elude it by creating masses of inbound links for doorway pages from just as insignificant other web pages.

[17]Contrary to the concept of link popularity, PageRank is not simply based upon the total number of inbound links. The basic approach of PageRank is that a document is in fact considered the more important the more other documents link to it, but those inbound links do not count equally. First of all, a document ranks high in terms of PageRank, if other high ranking documents link to it.

So, within the PageRank concept, the rank of a document is given by the rank of those documents which link to it. Their rank again is given by the rank of documents which link to them. Hence, the PageRank of a document is always determined recursively by the PageRank of other documents. Since - even if marginal and via many links - the rank of any document influences the rank of any other, PageRank is, in the end, based on the linking structure of the whole web. Although this approach seems to be very broad and complex, Page and Brin were able to put it into practice by a relatively trivial algorithm.

2.2.1. Overview of web Page Ranking Algorithms

Page Rank Algorithms

One of the most known and influential algorithms for computing the relevance of web pages is the Page Rank algorithm used by the Google search engine. It was invented by Larry Page and Sergey Brin while they were graduate students at Stanford, and it became a Google trademark in 1998. The Page Rank algorithm is based on the idea if a page contains important links towards it then the links of this page towards the other page are also to be considered as important pages. The importance of any web page can be judged by looking at the pages that link to it. If we create a web page *i* and include a hyperlink to the web page *j*, this means that we consider *j* important and relevant for our topic. If there are a lot of pages that link to *j*, this means that the common belief is that page *j* is important. If on the other hand, *j* has only one backlink, but that comes from an authoritative site *k*, (like www.google.com, www.cnn.com, www.cornell.edu) we say that *k* transfers its authority to *j*; in other words, *k* asserts that *j* is important. Whether we talk about popularity or authority, we can iteratively assign a rank to each web page, based on the ranks of the pages that point to it. A simplified version of PageRank is given by:

$$PR(u) = \sum_{v \in B_u} \frac{PR(v)}{L(v)}$$

Where the PageRank value for a web page u is dependent on the PageRank values for each web page v out of the set Bu (this set contains all pages linking to web page u), divided by the number L(v) of links from page v.An example of back link is shown in figure 2 below. U is the back link of V & W and V & W are the back links of X.

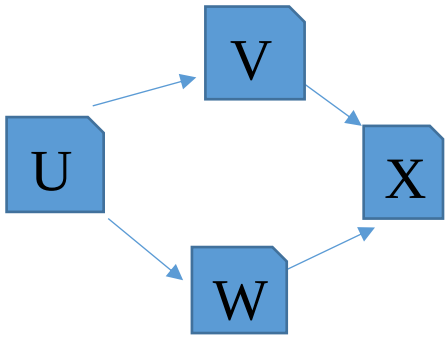

Figure 2 Illustration of back links

Weighted Page Rank Algorithm

Weighted Page Rank [5] Algorithm is proposed by Wenpu Xing and Ali Ghorbani. Weighted page rank algorithm (WPR) is the modification of the original page rank algorithm. WPR decides the rank score based on the popularity of the pages by taking into consideration the importance of both the in-links and out-links of the pages. This algorithm provides high value of rank to the more popular pages and does not equally divide the rank of a page among it's out-link pages. Every out-link page is given a rank value based on its popularity. Popularity of a page is decided by observing its number of in links and out links. Simulation of WPR is done using the Website of Saint Thomas University and simulation results show that WPR algorithm finds larger number of relevant pages compared to standard page rank algorithm. As suggested by the author, the performance of WPR is to be tested by using different websites and future work include to calculate the rank score by utilizing more than one level of reference page list and increasing the number of human user to classify the web pages.

Weighted Links Rank Algorithm

A modification of the standard page rank algorithm is given by Ricardo Baeza-Yates and Emilio Davis named[18] as weighted links rank (WLRank). This algorithm provides weight value to the link based on three parameters i.e. length of the anchor text, tag in which the link is contained and relative position in the page. Simulation results show that the results of the search engine are improved using weighted links. The length of anchor text seems to be the best attributes in this algorithm. Relative position, which reveal that physical position does not always in synchronism with logical position is not so result oriented. Future work in this algorithm includes, tuning of the weight factor of every term for further evolution.

EigenRumor Algorithm

As the number of blogging sites is increasing day by day, there is a challenge for service provider to provide good blogs to the users. Page rank and HITS are very promising in providing the rank value to the blogs but some limitations arise, if these two algorithms are applied directly to the blogs The rank scores of blog entries as decided by the page rank algorithm is often very low so it cannot allow blog entries to be provided by rank score according to their importance. To resolve these limitations, a EigenRumor algorithm [19] is proposed for ranking the blogs. This algorithm

provides a rank score to every blog by weighting the scores of the hub and authority of the bloggers depending on the calculation of Eigen vector.

Distance Rank Algorithm

An intelligent ranking algorithm named as distance rank is proposed by Ali Mohammad Zareh Bidoki and Nasser Yazdani [20]. It is based on reinforcement learning algorithm. In this algorithm, the distance between pages is considered as a punishment factor. In this algorithm the ranking is done on the basis of the shortest logarithmic distance between two pages and ranked according to them. The Advantage of this algorithm is that it can find pages with high quality and more quickly with the use of distance based solution. The Limitation of this algorithm is that the crawler should perform a large calculation to calculate the distance vector, if new page is inserted between the two pages.

Time Rank Algorithm

An algorithm named as TimeRank, for improving the rank score by using the visit time of the web page is proposed by H Jiang et al [21]. Authors have measured the visit time of the page after applying original and improved methods of web page rank algorithm to know about the degree of importance to the users. This algorithm utilizes the time factor to increase the accuracy of the web page ranking. Due to the methodology used in this algorithm, it can be assumed to be a combination of content and link structure. The results of this algorithm are very satisfactory and in agreement with the applied theory for developing the algorithm.

TagRank Algorithm

A novel algorithm named as TagRank [22] or ranking the web page based on social annotations is proposed by Shen Jie,Chen Chen,Zhang Hui,Sun Rong-Shuang,Zhu Yan and He Kun. This algorithm calculates the heat of the tags by using time factor of the new data source tag and the annotations behavior of the web users. This algorithm provides a better authentication method for ranking the web pages. The results of this algorithm are very accurate and this algorithm index new information resources in a better way. Future work in this direction can be to utilize co-occurrence factor of the tag to determine weight of the tag and this algorithm can also be improved by using semantic relationship among the co-occurrence tags.

Relation Based Algorithm

Fabrizio Lamberti, Andrea Sanna and Claudio Demartini [23] proposed a relation based algorithm for the ranking the web page for semantic web search engine. Various search engines are presented for better information extraction by using relations of the semantic web. This algorithm proposes a relation based page rank algorithm for semantic web search engine that depends on information extracted from the queries of the users and annotated resources. Results are very encouraging on the parameter of time complexity and accuracy. Further improvement in this algorithm can be the increased use of scalability into future semantic web repositories. Query Dependent Ranking Algorithm Lian- Wang Lee, Jung- Yi Jiang, ChunDer Wu and Shie-Jue Lee have presented a query dependent raking algorithm for search engine. In this approach a simple similarity measure algorithm is used to measure the similarities between the queries. A single model for ranking is made for every training query with corresponding document. Whenever a query arises, then documents are extracted and ranked depending on the rank scores calculated by the ranking model. The ranking model in this algorithm is the combination of various models of the similar training queries. Experimental results show that query dependent ranking algorithm is better than other algorithms.

2.2.2. Summary of various web page ranking algorithms

By going through the literature analysis of some of the important web page ranking algorithms, it is concluded that each algorithm has some relative strengths and limitations. A tabular summary is given below in table 1, which summarizes the techniques, advantages and limitations of some of important web page ranking algorithms.

Author/Year	Technique	Advantages	Limitations
S. Brin et al. 1998	Graph based algorithm based on link structure of web pages. Consider the back links in the rank calculations.	Rank is calculated on the basis of the importance of pages.	Results are computed at the indexing time not at the query time.
Jon Kleinberg, 1998	Rank is calculated by computing hub and authorities score of the	Returned pages have high relevancy and importance.	With less efficiency and problem of topic drift.

	pages in order of their relevance.		
Sung Jin Kim et al. 2002	This algorithm probabilistically estimates that clear semantics and the identified authoritative documents correspond better to human intuition.	Well defined semantics with clear interpretation. Efficiently provide answer to quantitative bibliometric questions.	Priori should be decided on the number of factors to model. Trades computational expense for the risk of getting stuck in local maxima.
Wenpu Xing et al. 2004	Based on the calculation of the weight of the page with the consideration of the outgoing links, incoming links and title tag of the page at the time of searching.	It gives higher accuracy in terms of ranking because it uses the content of the pages.	It is based only on the popularity of the web page.
Ricardo BaezaYates et al. 2004	This algorithm ranks the page by providing different weights based on three attributes i.e. relative position in page, tag where link is contained & length of anchor text.	It has less efficiency with reference to precision of the search engine.	Relative position was not so effective, indicating that the logical position not always matches the physical position
Ko Fujimura et al. 2005	Use of the adjacency matrix, constructed from agent to object link not by page to page link. Three vectors i.e. hub, authority and reputation are needed for score calculation of the blog.	Useful for ranking of blog as well as web pages because input and output links are not considered in the algorithm.	Specifically suited for blog ranking.
Ali Mohammad Zareh Bidoki et al. 2007	Based on reinforcement learning which consider the logarithmic distance between the pages.	Algorithm consider real user by which pages can be found very quickly with high quality.	A large calculation for distance vector is needed, if new page inserted between the two pages.

Hua Jiang et al. 2008	Visitor time is used for ranking. Use of sequential clicking for sequence vector calculation with the uses of random surfing model.	Useful when two pages have the same link structure but different contents.	It does work efficiently when the server log is not present.
Shen Jie et al. 2008	The algorithm is based on the analysis of tag heat on social annotation web.	Ranking results are very exact and new information resources are indexed more effectively.	Co-occurrence factor of tag is not considered which may influence the weight of the tag.
Fabrizio Lamberti et al. 2009	Ranking of web pages for semantic search engine. It uses the information extracted from the queries of the user and annotated resources.	Effectively manage the search page. Ranking task is less complex.	In this ranking algorithm every page is to be annotated with respect to some ontology, which is the very tough task.
Lian-Wang Lee et al. 2009	Individual models are generated from training queries. A new query ranked according to the combined weighted score of these models.	It gives the results for user's query as well as results for similar type of query.	Limited numbers of characteristics are used to calculate the similarity.
Milan Vojnovic et al. 2009	Suggest the popular items for tagging. It uses the three randomized algorithms i.e. frequency proportional sampling, move-to-set and frequency move-to-set.	Tag popularity boost up because large number of tag are suggested by this method.	Does not consider alternative user choice model, alternative rules for ranking and alternative suggestive rules.
Narayan L Bhamidipati et al. 2009	Based on the score fusion technique.	It is used when two pages have same ranking.	Does not consider the case when score vector T generated from specific distribution.
Xiang Lian et al. 2010	Retrieval of moving object in the uncertain databases. It uses the PRank (probabilistic	It is fast because it uses the R-Tree.	Experimental results are very promising only with limited number of parameters such that. wall

	ranked query) and J-Prank (probabilistic ranked query on join).		clock time and number of Prank candidate

Table 1 Summery of techniques, advantages and limitations of some of important web page ranking algorithms

2.2.3. Comparison of various web page ranking algorithms

Based on the literature analysis, a comparison of some of various web page ranking algorithms is shown in table 2. Comparison is done on the basis of some parameters such as main technique use, methodology, input parameter, and relevancy, quality of results, importance and limitations.

Algorithm	Page Rank	HITS	Weighted Page Rank	Eigen Rumor	Web Page Ranking using Link Attributes
Main Technique	Web Structure Mining	Web Structure Mining, Web Content Mining	Web Structure Mining	Web Content Mining	Web Structure Mining, Web Content Mining
Methodology	This algorithm computes the score for pages at the time of indexing of the pages.	It computes the hubs and authority of the relevant pages. It relevant as well as important page as the result.	Weight of web page is calculated on the basis of input and outgoing links and on the basis of weight the importance of page is decided.	Eigen rumor use the adjacency matrix, which is constructed from agent to object link not page to page link.	it gives different weight to web links based on 3 attributes: Relative position in page, tag where link is contained, length of anchor text.
Input Parameter	Back links Content, Back and Forward links	Back links and Forward links.	Agent/Object Content, Back and Forward links		
Relevancy	Less (this algo. rank the pages	More (this algo. Uses the hyperlinks so	Less as ranking is based on the calculation of	High for Blog so it is mainly	more (it consider the

	on the indexing time)	according to Henzinger, 2001 it will give good results and also consider the content of the page)	weight of the web page at the time of indexing.	used for blog ranking.	relative position of the pages)
Quality of results	Medium	Less than PR	Higher than PR	Higher than PR and HITS	Medium
Importance	High. Back links are considered.	Moderate. Hub & authorities scores are utilized.	High. The pages are sorted according to the importance.	High for blog ranking.	Not specifically quoted.
Limitation	Results come at the time of indexing and not at the query time.	Topic drift and efficiency problem	Relevancy is ignored.	It is most specifically used for blog ranking not for web page ranking as other ranking like page rank, HITS.	Relative position was not so effective, indicating that the logical position not always matches the physical position.

Table 2 Comparison of some important web page ranking algorithms

2.3. Speech recognition

Basic concepts of speech recognition

Speech recognition is the process by which a computer (or other type of machine) identifies spoken words. Basically, it means talking to your computer, AND having it correctly recognize what you are saying. Speech is a complex phenomenon. People rarely understand how is it produced and perceived. The wrong insight is often that speech is built with words and each word comprises of phones. The reality is unluckily very different. Speech is a dynamic process without clearly distinguished parts. It's always useful to get a sound editor and look into the recording of the speech and listen to it. Here is for example the speech recording in an audio editor.

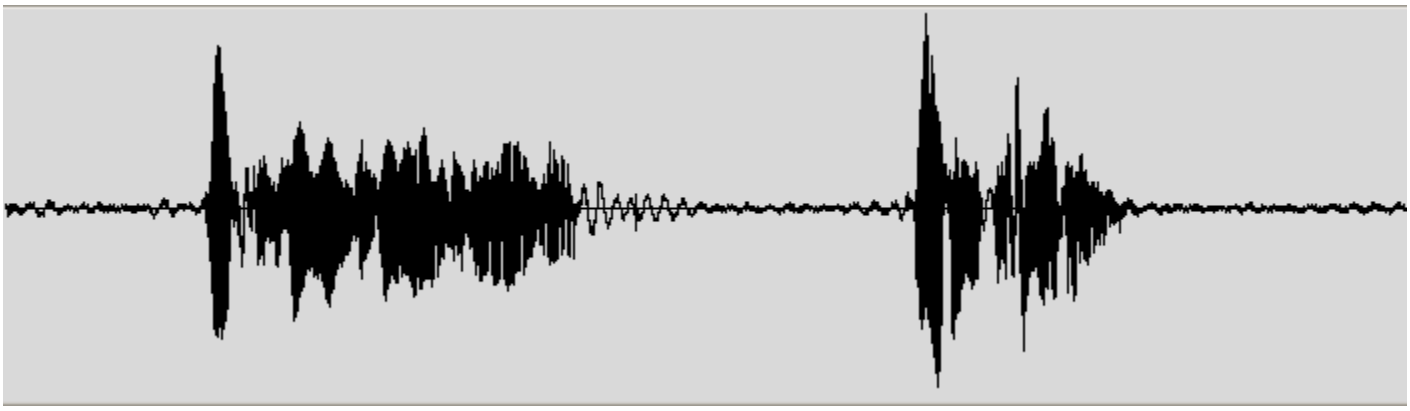

Figure 3 Speech recording an audio editor

All modern descriptions of speech are to some degree probabilistic. That means that there are no certain boundaries between units, or between words. Even in the modern technology, Speech to text translation and other applications of speech are never 100% correct. That idea is rather unusual for software developers, who usually work with deterministic systems. And it creates a lot of issues specific only to speech technology.

Structure of speech

In current practice, speech structure is understood as follows: Speech is a continuous audio stream where rather stable states mix with dynamically changed states. In this sequence of states, one can define more or less similar classes of sounds, or **phones**. Words are understood to be built of phones, but this is certainly not true. The acoustic properties of a waveform corresponding to a phone can vary greatly depending on many factors - phone context, speaker, style of speech and so on. The so-called coarticulation makes phones sound very different from their "canonical" representation. Next, since transitions between words are more informative than stable regions, developers often talk about **diphones** - parts of phones between two consecutive phones. Sometimes developers talk about subphonetic units - different substates of a phone. Often three or more regions of a different nature can be found. The number three can easily be explained: The first part of the phone depends on its preceding phone, the middle part is stable and the next part depends on the subsequent phone. That's why there are often three states in a phone selected for speech recognition.

Sometimes phones are considered in context. Such phones in context are called **triphones** or even **quinphones**. For example "u" with left phone "b" and right phone "d" in the word "bad" sounds a bit different than the same phone "u" with left phone "b" and right phone "n" in word "ban".

Please note that unlike diphones, they are matched with the same range in waveform as just phones. They just differ by name because they describe slightly different sounds.
For computational purpose it is helpful to detect parts of triphones instead of triphones as a whole, for example if you want to create a detector for the beginning of a triphone and share it across many triphones. The whole variety of sound detectors can be represented by a small amount of distinct short sound detectors. Usually we use 4000 distinct short sound detectors to compose detectors for triphones. We call those detectors **senones**. A senone's dependence on context can be more complex than just the left and right context. It can be a rather complex function defined by a decision tree, or in some other ways.

Next, phones build subword units, like syllables. Sometimes, syllables are defined as "reduction-stable entities". For instance, when speech becomes fast, phones often change, but syllables remain the same. Also, syllables are related to an intonational contour. There are other ways to build subwords - morphologically-based (in morphology-rich languages) or phonetically-based. Subwords are often used in open vocabulary speech recognition.Subwords form words. Words are important in speech recognition because they restrict combinations of phones significantly. If there are 40 phones and an average word has 7 phones, there must be 40^7 words. Luckily, even people with a rich vocabulary rarely use more than 20k words in practice, which makes recognition way more feasible.Words and other non-linguistic sounds, which we call **fillers** (breath, um, uh, cough), form **utterances**. They are separate chunks of audio between pauses. They don't necessary match sentences, which are more semantic concepts.

The following definitions are the basics needed for understanding speech recognition technology. Utterance: An utterance is the vocalization (speaking) of a word or words that represent a single meaning to the computer. Utterances can be a single word, a few words, a sentence, or even multiple sentences. Speaker Dependence: Speaker dependent systems are designed around a specific speaker. They generally are more accurate for the correct speaker, but much less accurate for other speakers. They assume the speaker will speak in a consistent voice and tempo. Speaker independent systems are designed for a variety of speakers. Adaptive systems usually start as speaker independent systems and utilize training techniques to adapt to the speaker to increase their recognition accuracy.

Vocabularies: Vocabularies (or dictionaries) are lists of words or utterances that can be recognized by the SR system. Generally, smaller vocabularies are easier for a computer to recognize, while larger vocabularies are more difficult. Unlike normal dictionaries, each entry doesn't have to be a single word. They can be as long as a sentence or two. Smaller vocabularies can have as few as 1 or 2 recognized utterances (e.g."Wake Up"), while very large vocabularies can have a hundred thousand or more!

Accuracy: The ability of a recognizer can be examined by measuring its accuracy - or how well it recognizes utterances. This includes not only correctly identifying an utterance but also identifying if the spoken utterance is not in its vocabulary. Good ASR systems have an accuracy of 98% or more! The acceptable accuracy of a system really depends on the application.

Training: Some speech recognizers have the ability to adapt to a speaker. When the system has this ability, it may allow training to take place. An ASR system is trained by having the speaker repeat standard or common phrases and adjusting its comparison algorithms to match that particular speaker. Training a recognizer usually improves its accuracy. Training can also be used by speakers that have difficulty speaking, or pronouncing certain words. As long as the speaker can consistently repeat an utterance, ASR systems with training should be able to adapt.

Types of Speech Recognition

Speech recognition systems can be separated in several different classes by describing what types of utterances they have the ability to recognize. These classes are based on the fact that one of the difficulties of ASR is the ability to determine when a speaker starts and finishes an utterance. Most packages can fit into more than one class, depending on which mode they're using.

1. **Isolated Words:** Isolated word recognizers usually require each utterance to have quiet (lack of an audio signal) on BOTH sides of the sample window. It doesn't mean that it accepts single words, but does require a single utterance at a time. Often, these systems have "Listen/Not-Listen" states, where they require the speaker to wait between utterances (usually doing processing during the pauses). Isolated Utterance might be a better name for this class.
2. **Connected Words:** Connect word systems (or more correctly 'connected utterances') are similar to Isolated words, but allow separate utterances to be 'run-together' with a minimal pause between them.

3. **Continuous Speech:** Continuous recognition is the next step. Recognizers with continuous speech capabilities are some of the most difficult to create because they must utilize special methods to determine utterance boundaries. Continuous speech recognizers allow users to speak almost naturally, while the computer determines the content. Basically, it's computer dictation.
4. **Spontaneous Speech:** There appears to be a variety of definitions for what spontaneous speech actually is. At a basic level, it can be thought of as speech that is natural sounding and not rehearsed. An ASR system with spontaneous speech ability should be able to handle a variety of natural speech features such as words being run together, "ums" and "ahs", and even slight stutters.
5. **Voice Verification/Identification:** Some ASR systems have the ability to identify specific users. This document doesn't cover verification or security systems.

Recognition process

How Recognizers Work

Recognition systems can be broken down into two main types. Pattern Recognition systems compare patterns to known/trained patterns to determine a match. Acoustic Phonetic systems use knowledge of the human body (speech production, and hearing) to compare speech features (phonetics such as vowel sounds). Most modern systems focus on the pattern recognition approach because it combines nicely with current computing techniques and tends to have higher accuracy. Most recognizers can be broken down into the following steps:

1. Audio recording and Utterance detection
2. Pre-Filtering (pre-emphasis, normalization, banding, etc.)
3. Framing and Windowing (chopping the data into a usable format)
4. Filtering (further filtering of each window/frame/freq. band)
5. Comparison and Matching (recognizing the utterance)
6. Action (Perform function associated with the recognized pattern)

Although each step seems simple, each one can involve a multitude of different (and sometimes completely opposite) techniques.

(1) Audio/Utterance Recording: can be accomplished in a number of ways. Starting points can be found by comparing ambient audio levels (acoustic energy in some cases) with the sample just

recorded. Endpoint detection is harder because speakers tend to leave "artifacts" including breathing/sighing, teeth chatters, and echoes.

(2) Pre-Filtering: is accomplished in a variety of ways, depending on other features of the recognition system. The most common methods are the "Bank-of-Filters" method which utilizes a series of audio filters to prepare the sample, and the Linear Predictive Coding method which uses a prediction function to calculate differences (errors). Different forms of spectral analysis are also used.

(3) Framing/Windowing involves separating the sample data into specific sizes. This is often rolled into step 2 or step 4. This step also involves preparing the sample boundaries for analysis (removing edge clicks, etc.)

(4) Additional Filtering is not always present. It is the final preparation for each window before comparison and matching. Often this consists of time alignment and normalization.

There are a huge number of techniques available for (5), Comparison and Matching. Most involve comparing the current window with known samples. There are methods that use Hidden Markov Models (HMM), frequency analysis, differential analysis, linear algebra techniques/shortcuts, spectral distortion, and time distortion methods. All these methods are used to generate a probability and accuracy match.

(6) Actions can be just about anything the developer wants. *GRIN*

Digital Audio Basics

Audio is inherently an analog phenomenon. Recording a digital sample is done by converting the analog signal from the microphone to a digital signal through the A/D converter in the sound card. When a microphone is operating, sound waves vibrate the magnetic element in the microphone, causing an electrical current to the sound card (think of a speaker working in reverse). Basically, the A/D converter records the value of the electrical voltage at specific intervals.

There are two important factors during this process. First is the "sample rate", or how often to record the voltage values. Second, is the "bits per sample", or how accurate the value is recorded? A third item is the number of channels (mono or stereo), but for most ASR applications mono is sufficient. Most applications use pre-set values for these parameters and users shouldn't change them unless the documentation suggests it. Developers should experiment with different values to determine what works best with their algorithms.

So what is a good sample rate for ASR? Because speech is relatively low bandwidth (mostly between 100Hz-8kHz), 8000 samples/sec (8kHz) is sufficient for most basic ASR. But, some people prefer 16000 samples/sec (16kHz) because it provides more accurate high frequency information. If you have the processing power, use 16kHz. For most ASR applications, sampling rates higher than about 22kHz is a waste.

And what is a good value for "bits per sample"? 8 bits per sample will record values between 0 and 255, which means that the position of the microphone element is in one of 256 positions. 16 bits per sample divides the element position into 65536 possible values. Similar to sample rate, if you have enough processing power and memory, go with 16 bits per sample. For comparison, an audio Compact Disc is encoded with 16 bits per sample at about 44 kHz.

The encoding format used should be simple - linear signed or unsigned. Using a U-Law/A-Law algorithm or some other compression scheme is usually not worth it, as it will cost you in computing power, and not gain you much. The following definitions are the basics needed for understanding speech recognition technology.

Utterance: An utterance is the vocalization (speaking) of a word or words that represent a single meaning to the computer. Utterances can be a single word, a few words, a sentence, or even multiple sentences.

Speaker Dependence: Speaker dependent systems are designed around a specific speaker. They generally are more accurate for the correct speaker, but much less accurate for other speakers. They assume the speaker will speak in a consistent voice and tempo. Speaker independent systems are designed for a variety of speakers. Adaptive systems usually start as speaker independent systems and utilize training techniques to adapt to the speaker to increase their recognition accuracy.

Vocabularies: Vocabularies (or dictionaries) are lists of words or utterances that can be recognized by the SR system. Generally, smaller vocabularies are easier for a computer to recognize, while larger vocabularies are more difficult. Unlike normal dictionaries, each entry doesn't have to be a single word. They can be as long as a sentence or two. Smaller vocabularies can have as few as 1 or 2 recognized utterances (e.g."Wake Up"), while very large vocabularies can have a hundred thousand or more!

Accuracy: The ability of a recognizer can be examined by measuring its accuracy - or how well it recognizes utterances. This includes not only correctly identifying an utterance but also identifying

if the spoken utterance is not in its vocabulary. Good ASR systems have an accuracy of 98% or more! The acceptable accuracy of a system really depends on the application.

Training: Some speech recognizers have the ability to adapt to a speaker. When the system has this ability, it may allow training to take place. An ASR system is trained by having the speaker repeat standard or common phrases and adjusting its comparison algorithms to match that particular speaker. Training a recognizer usually improves its accuracy. Training can also be used by speakers that have difficulty speaking, or pronouncing certain words. As long as the speaker can consistently repeat an utterance, ASR systems with training should be able to adapt.

Models

According to the speech structure, three models are used in speech recognition to do the match:

An **acoustic model** contains acoustic properties for each senone. There are context-independent models that contain properties (the most probable feature vectors for each phone) and context-dependent ones (built from senones with context).

A **phonetic dictionary** contains a mapping from words to phones. This mapping is not very effective. For example, only two to three pronunciation variants are noted in it. However, it's practical enough most of the time. The dictionary is not the only method for mapping words to phones.

A **language model** is used to restrict word search. It defines which word could follow previously recognized words (remember that matching is a sequential process) and helps to significantly restrict the matching process by stripping words that are not probable. The most common language models are n-gram language models–these contain statistics of word sequences–and finite state language models–these define speech sequences by finite state automation, sometimes with weights. To reach a good accuracy rate, your language model must be very successful in search space restriction. This means it should be very good at predicting the next word. A language model usually restricts the vocabulary that is considered to the words it contains. That's an issue for name recognition. To deal with this, a language model can contain smaller chunks like subwords or even phones. It should be noted that the search space restriction in this case is usually worse and the corresponding recognition accuracies are lower than with a word-based language model.

Speech optimization

When speech recognition is being developed, the most complex problem is to make search precise (consider as many variants to match as possible) and to make it fast enough to not run for ages. Since models aren't perfect, another challenge is to make the model match the speech.

Usually the system is tested on a test database that is meant to represent the target task correctly. The following characteristics are used:

Word error rate. Let's assume we have an original text and a recognition text with a length of *N* words. *I* is the number of inserted words, *D* is the number of deleted words and *S* represent the number of substituted words. With this, the word error rate can be calculated as

WER = (I + D + S) / N

The WER is usually measured in percent.

Accuracy. It is almost the same as the word error rate, but it doesn't take insertions into account.

Accuracy = (N - D - S) / N

For most tasks, the accuracy is a worse measure than the WER, since insertions are also important in the final results. However, for some tasks, the accuracy is a reasonable measure of the decoder performance.

Speed. Suppose an audio file has a recording time (RT) of 2 hours and the decoding took 6 hours. Then the speed is counted as 3xRT.

ROC curves. When we talk about detection tasks, there are false alarms and hits/misses. To illustrate these, ROC curves are used. Such a curve is a diagram that describes the number of false alarms versus the number of hits. It tries to find the optimal point where the number of false alarms is small and the number of hits matches 100%.

Hardware

1. **Sound Cards**

Because speech requires a relatively low bandwidth, just about any medium-high quality 16 bit sound card will get the job done. You must have sound enabled in your kernel, and you must have correct drivers installed. Sound cards with the 'cleanest' A/D (analog to digital) conversions are recommended, but most often the clarity of the digital sample is more dependent on the microphone quality and even more dependent on the environmental noise. Electrical "noise" from monitors, pci slots, hard-drives, etc. are usually nothing compared to audible noise from the computer fans, squeaking chairs, or heavy breathing.

2. **Microphones**

A quality microphone is key when utilizing ASR. In most cases, a desktop microphone just won't do the job. They tend to pick up more ambient noise that gives ASR programs a hard time.

Hand held microphones are also not the best choice as they can be cumbersome to pick up all the time. While they do limit the amount of ambient noise, they are most useful in applications that require changing speakers often, or when speaking to the recognizer isn't done frequently (when wearing a headset isn't an option).

The best choice, and by far the most common is the headset style. It allows the ambient noise to be minimized, while allowing you to have the microphone at the tip of your tongue all the time. Headsets are available without earphones and with earphones (mono or stereo).

Computers/Processors

ASR applications can be heavily dependent on processing speed. This is because a large amount of digital filtering and signal processing can take place in ASR.

As with just about any CPU intensive software, the faster the better. Also, the more memory the better. It's possible to do some SR with 100MHz and 16M RAM, but for fast processing (large dictionaries, complex recognition schemes, or high sample rates), you should shoot for a minimum of a 400MHz and 128M RAM. Because of the processing required, most software packages list their minimum requirements.

2.4. Related work

As [5] there are two basic concepts or procedures which are query expansion and re-ranking algorithm. When user search for any query, it request is sent to query expansion. Query expansion attaches useful keywords from user profile based on users search query and it also provides some other keywords that is utilized for general search for same. According to this paper keywords retrieved for general search is helpful to recognizing the current interest of user on individual query. The expanded query is given as an input to the search engine and the result from search engine is passed to re-ranking algorithm which evaluates rank of each returned links and reorder according to rank.

A. Query expansion

The author used Co-occurrence based query expansion approach for query expansion. The system arranges each word in query into three levels according to their term frequency value.

Term frequency is calculated by

$$tfw = \sum_{k=1}^{n} \frac{|wk|}{|w|}$$

Where *tfw* is term-frequency of keyword w. |*Wk*|is no. of time keyword *Wk* presents in set and |*W* | total occurrence of all the keyword in the set. Whenever a user click on a particular link by his search, that link, its heading, query searched, actions performed on that link, snippets, how much time user spend on particular link and other information are stored as a history of user. This is stored as a three level semi-structured format. From this title, snippet and URL is used for finding user interest list. Then, term frequency of each keyword is calculated. If this keyword is exists both in the three level semi-structured formats created earlier and the user interest list its frequency value will be higher, then this keyword is denoted as a valuable keyword for further search. Each time a user search a query, the system locates the level of particular query in the user interest list stored as a three level semi-structured format. After that the weight of each lower level term is computed and the weight is calculated by its number of Child. The highest number of child containing term is added with query word and this is forward to search engine for find personalized result and the remaining terms are used for general search

B. Re-ranking Algorithm

I this stage the search result is further arranged by the re-ranking algorithm by considering the rank of each links. The author of this paper used Vector space model for document representation and Co-sign similarity function get matched terms related to particular query from web log three level semi-structured data. Co-sign similarity of two vectors query and link terms is calculated using this equation.

$$cos\theta = \frac{d.q}{||d||.||q||}$$

After obtain the related terms about query from web log, bring out the link, the time spent on that link, action performed and number of times the particular user clicked that link. This link list along with above details and search result from search engine is send to re-ranking algorithm. Re-ranking algorithm first take the rank by page rank algorithm then calculate the Perform Rank(li) by Action Value(li), Click Value(li) and Dwell Time(li) of particular link. So the actual rank of the link is the sum of these two ranks. Then arrange the search result from search engine by this rank.

Re-ranking Algorithm

Input: O;

Output: O';

O= {o_1, o_2, o_3,…….. o_m} Rank from Search Engine

O'= {o_1, o_2, o_3,…….. o_m} Rank from Search Engine

Re-ranking module ()

{

For each document in O, do

{

Rank=Page rank ();

Perform rank = Click value(*li*) + *Dwell Time*(*li*)+*Action value*(*li*);

Actual rank= Rank+ *Perform rank*

}

Arrange each document in O by rank

Return O' to the Search Engine interface

}

Perform Rank(li) is the summation of Action Value(li), Click Value(li) and Dwell Time(li).

Perform Rank(*li*) = *Click value*(*li*) + *Dwell T ime*(*li*)+*Action value*(*li*)

where

• ***Click value*(*li*)** denotes the proportion of number of clicks made by user u on the page *li* for query 'q' with respect to total number of clicks on all the pages made by user u for query 'q' given by

$$Click\ value(li) = \frac{No.of\ clicksonpagelibyuseruforquery\ \boldsymbol{q}}{Totalno.ofclicksonallthepagesmadebyuserfor\ query\ \boldsymbol{q}}$$

• ***Dwell T ime*(*li*)** denotes the time spent by user 'u' on the page li given by

$$Dwell\ T\ ime(li) = \left(\frac{No.of\ clT\ imespentbyuseruonpageli}{Maximumtimespentbyuseruonanypageliinpast}\right)^2$$

• ***Action value*(*li*)** denotes the action performed on the page *li* like saving, printing etc. given by

Action value(*li*) = *P rint*(*li*) + *Save*(*li*) + *Bookmark*(*li*) + *Send*(*li*)

[8]The queries provided by the users act as an input to the Google API which are available to construct the custom search. Root words are extracted from the user provided query by eliminating the stop words and synonyms are identified for root words and dictionary is constructed.

Google API produces a list of 'N' URLs. Hash table is used to store key words that denote words that are within the HTML title tag and content words that denote words that are within the HTML body tag. After comparing the keywords with the root word and its synonym on the dictionary, if they match a weight of teen is assigned to the keyword. Similarly by comparing the content word with the dictionary, if they match a weight of one is assigned to the content word. Finally total relevancy is computed by summing all the key word weights and content word weights. This procedure is done for all URLs and based on this total relevancy value URLs are re-ranked by re-arranging them in decreasing order

The objectives of this algorithm are:

- Reducing search time

- Satisfying user with relevant information.

[9]According to the author the main goal of the paper is to fulfill the requirement of the user with page ranking algorithms which ranks the document in the better way and holds the capability to re-rank search results effectively and tray the best to arrange the web results which are the most relevant for the user using the user time attention algorithm. As the author stated the algorithm is based on result of semantic web in which it will rank the page based on term frequency factor which. The author explained term frequency factor as how many time the same key word is repeating in the web page. According to this paper there are two major tasks which are semantic information retrieval and ranking of search results.

The algorithm depends on the user attention time on a particular document it looks users reading time for every document and if the time duration of reading that document passes a threshold value the system assumes that document is according to the user interest. The system will filter results and give a high rank to this documents that are similar to the document that was having more attention earlier and it will arrange this document from high value to low value

[10]Web mining is an application of data mining which has become an important area of research due to vast amount of World Wide Web services in recent years. The author tried to explain different categories of web mining and explained them briefly in the paper.

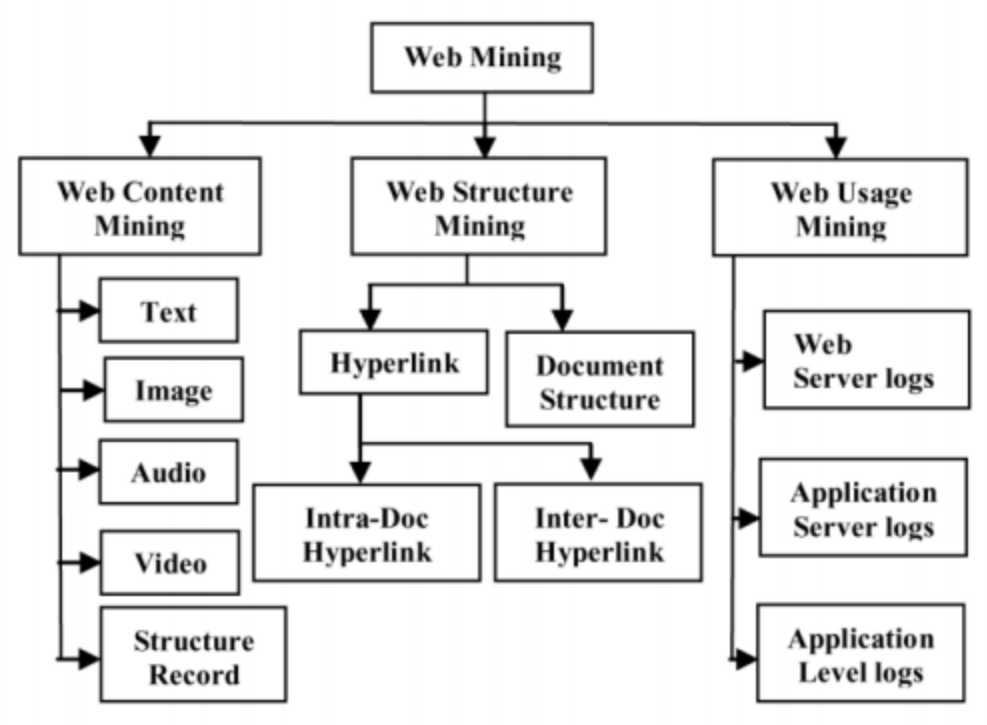

Figure 4 Type of web mining

Web structure mining: According to the author Web structure mining discovers useful knowledge from hyperlinks (or links for short), which represent the structure of the Web. In the example the author provided important web page are discovered from links, which is incidentally a key technology used in search engines. Users who share common interests can also be discovered. Traditional data mining does not perform such tasks because there is usually no link structure in a relational table.

Web content mining: Web content mining extracts or mines useful information or knowledge from Web page contents. For example, we can automatically classify and cluster Web pages according to their topics. These tasks are similar to those in traditional data mining. However, we can also discover patterns in Web pages to extract useful data such as descriptions of products, postings of forums, etc., for many purposes.

Web usage mining: Web usage mining refers to the discovery of user access patterns from Web usage logs, which record every click made by each user. Web usage mining applies many data mining algorithms. One of the key issues in Web usage mining is the pre-processing of click stream data in usage logs in order to produce the right data for mining

2.5. Conclusion

Based on the algorithm used, the ranking algorithm provides a definite rank to resultant web pages. A typical search engine should use web page ranking techniques based on the specific needs of the users. After going through exhaustive analysis of algorithms for ranking of web pages against the various parameters such as methodology, input parameters, relevancy of results and importance of the results, it is concluded that existing techniques have limitations particularly in terms of time response, accuracy of results, importance of the results and relevancy of results. An efficient web page ranking algorithm should meet out these challenges efficiently with compatibility with global standards of web technology.

CHAPTER 3

3. METHODOLOGY

In order to achieve the specific and general objectives of the study and to answer the research questions, different data gathering and analysis strategies have been used.

3.1. Research design

This research is an empirical study involving exploratory and in-depth investigation. A mixed method research approach was applied, with a combination of Quantitative and Qualitative research methods. Since this research intended to enhance the existing content based webpage re-ranking algorithm for visually impaired internet users, it is necessary to develop a Amharic voice recognizer plugin prototype to examine the enhanced re-ranking algorithm. A qualitative approach was used in this study to conduct the subjective assessment of attitudes, behaviors, and opinions of target data sources (respondents).

In exploratory nature survey is used representing the quantitative research method. An in-depth study using interview approach and participant observation was done representing the qualitative research method.

3.2. Process Model

In this research we follow the method of Problem identification, Define the objectives for a solution, Design and prototype development and Evaluation

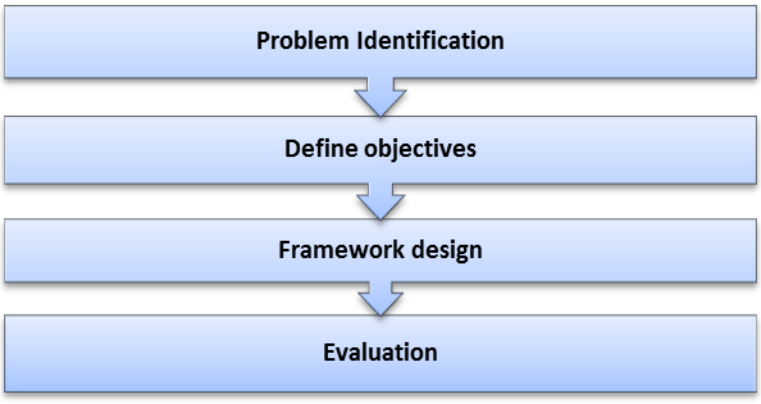

Figure 5: Research process model.

3.2.1. Problem identification and motivation

This phase includes defining the knowledge of the state of the problem and the importance of its solution. We have begun the research by defining the specific research problem and justify the value of a solution by conducting an interview and rigorous analysis of the literature on internet usage and page ranking of search results among visually impaired Bahirdar University students .Since based on the problem definition the researchers atomize the problem conceptually to capture its complexity for developing the proposed page ranking algorithm that can effectively provide a solution. The value of the solution also justified motivating the researcher and the audience of the research to pursue the solution and to accept the results and to understand the reasoning associated with the researcher's understanding of the problem.

3.2.2. Define the objectives for a solution

This phase aims at to identify or drive the requirements for developing the proposed page re-ranking algorithm from knowledge of the state of problems and current solutions. The researchers infer the objectives of a solution from the problem definition and knowledge of what is possible and feasible to align with the visually impaired internet users requirements. The qualitative objectives are used, which described how a proposed algorithm is expected to support solutions to problems. The researcher used the conceptual framework to inferred objectives rationally from the problem specification. In order to develop the study conceptual architecture, researchers started with the literature review to gather relevant requirements and aspects of existing page ranking algorithm.

3.2.3. Architectural Design

Conceptually, a design research artifact can be any designed object (models, architecture or instantiations) in which a research contribution is embedded in the design. This phase includes determining the framework's desired functionality and its architecture and then creating the actual artifact. The knowledge of theory that can be brought to bear in a solution is required to move from the objective of a solution to design and development of the proposed algorithm. Based on the required knowledge of a solution identified in phase one and two with the literature review and expert interviews analysis to develop the proposed algorithm. Finally, the architectural design of the proposed algorithm has been designed.

3.2.4. Evaluation

This phase aims to Observe and measure how well the proposed algorithm supports a solution to the problem. Depending on the nature of the problem venue and the solution proposed, this study took the qualitative evaluation analysis techniques which include: developing a prototype, a comparison of the proposed algorithm functionality with the solution objectives and user satisfaction.

3.3. Data sources

The research focused on both secondary and primary data. For the secondary data, thorough reviews were done and for the primary data, interview techniques and participant observation were used. Survey was conducted with the intention of capturing respondents' feelings and opinions about voice based web searching. Interview was a process to gain respondents' thoughts on existing voiced based web searching in case of Google search engine, whereas participant observation is the process of observing while visually impaired students conduct their learning process.

3.4. Data collection techniques

In the previous section, it's mentioned that different data sources which have been used to achieve the objective of the research. In this section, we have explained that the way how and where to gather this information and why we choose those techniques.

Semi-structure interview method was selected for this study and disabled students (Visual and Motor) from Bahirdar university Peda campus where interviewed. Semi-structured interviews give the researcher flexibility to establish the own style of conversation depending on the direction of the interview. We undertook a study using ten blind or visually impaired students to assess their reactions about the existing learning methods in the University; interview questions were organized in the way that enables to get information related to the reaction of the disabled students about the current learning environment, availability of digital learning(digital library), availability of Internet and voice based web searching mechanism, gap between normal and disabled students on learning environment, and reaction about the proposed solution. Interviews were conducted between April 15 and May 25, 2017, The Subjects chosen for this research were attending a

university course within Bahirdar university Peda campus that includes civics, English and history. They represent a wide range of visual impairments. Students were identified by their lecturer as having a range of difficulties.

3.4.1. Designing Interview Questions:

In investigating voice based web searching and digital learning environment for visually impaired internet users; this research used a standard open-ended type of interview. This means that the questions used by the researcher were open-ended that provided respondents the freedom to express their opinions. The interview questions tried to investigate the gap between the visually impaired and other normal friends in digital learning environment and web searching. The standardized interview facilitated the data analysis process for this research. The interview questions were divided into four main parts as shown in Figure 6 and explained further following the figure:

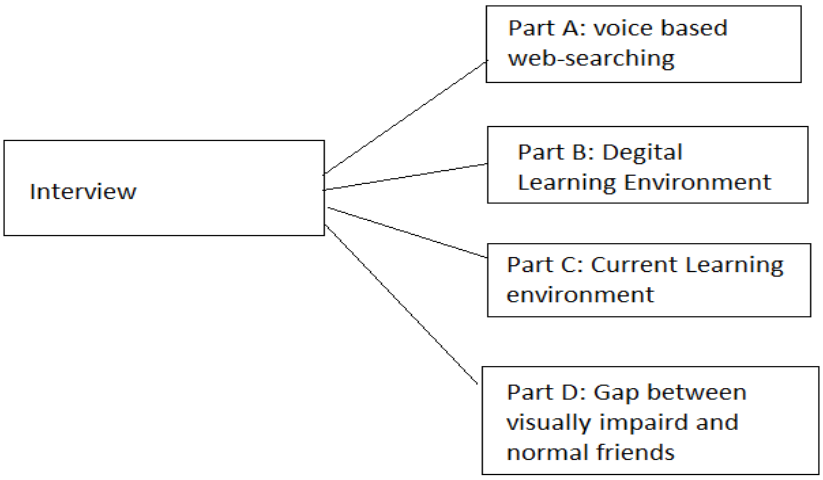

Figure 6: Interview question design

Part A: Voice-based web searching – In Voice-based web searching, the respondents were asked about their experience in using the internet, how they accessed, how the search results were interpreted for them and whether the search results were satisfactory, for example:

- ✓ Do you use the internet?
- ✓ How do you interact with the internet?
- ✓ How did the search results were interpreted for you?
- ✓ Did the search results were satisfactory?

Part B: Digital learning environment – Digital learning environment, the respondents' experience on digital learning environment like using digital library to conduct higher education and how they use this tool, for example:

- ✓ Availability of digital library.
- ✓ Do you think using digital learning is supportive in learning environment?

Part C: Current learning environment – in this section the interview focused on how do the teaching and learning process is generally conducted. The respondents' answer about how they are conducting their education, for example:

- ✓ How did the teacher deliver lecture class?
- ✓ How did you take exams?

Part D: Gap between visually impaired and normal friends - In Gap between visually impaired and normal friends, the respondents were asked about gaps between them and their other normal friends during lecture class, exam and reference materials, for example:

- ✓ Do fill there is gap between you and normal class mates in the teaching and learning environment?

3.4.2. Participatory observation

An observation on how Bahirdar University visually impaired students conduct their class has been observed being in their class. During exam times observation on the exam are delivered has been viewed. How the visually impaired students access the digital library of Bahirdar university Peda campus to prepared for exam and to work their assignment have been observed

We undertook a study using ten blind or visually impaired students to assess their reactions to a number of Internet search engines. The ten Subjects chosen for this research were attending a university course within Bahirdar University that includes Civics, English History Departments. They represent a wide range of visual impairments. Students were identified by their lecturer as having a range of difficulties.

The students were all mature students. They were all very willing to co-operate with the study and enthusiastic to convey their views on particular points that arose. They were all relatively new to computing and using the WWW.The research was carried out using the Bahirdar University facilities.

3.5. The Development of the Prototype

The programming language used in the implementation is Object Oriented JavaScript. To implement the prototype the component of java NetBeans 8.0.2 IDE called JEE (Java Enterprise Edition) platform which is built on the top of java standard edition have been used. The java platform enterprise edition defines a sample standard that applies to all aspects of architecting, developing, and managing multitier server based application. The aim of the JEE platform is to provide developers with a powerful set of APIs while shortening development time, reducing application complexity, and improving application performance [25]. In this prototype, we especially used web technologies like; web socket API, WebSocket provides a welcomed alternative to the AJAX technologies we've been making use of over the past few years. This new API provides a method to push messages from client to server efficiently and with a simple syntax. WebSocket API is the next generation method of asynchronous communication from client to server. Communication takes place over single TCP socket using the ws (unsecure) or wss (secure) protocol and can be used by any client or server application. WebSocket is currently being standardized by the W3C. WebSocket is currently implemented in Firefox 4, Chrome 4, Opera 10.70, and Safari 5. The overall implementation of custom Websocket for our system server is explained below in detail. The have used JavaScript Web Speech API which makes it easy to add speech recognition to our system. This API allows fine control and flexibility over the speech recognition capabilities in Chrome version 25 and later. To Scraping URLs we have used different java libraries which are:

- ✓ Websocket: A barebones WebSocket client and server implementation written 100% in Java.
- ✓ Boilerpipe: The boilerpipe library provides algorithms to detect and remove the surplus "clutter" (boilerplate, templates) around the main textual content of a website.
- ✓ Gson is a Java library that can be used to convert Java Objects into their JSON representation. It can also be used to convert a JSON string to an equivalent Java object.

Gson can work with arbitrary Java objects including pre-existing objects that you do not have source-code of.

- ✓ NekoHTML is a simple HTML scanner and tag balancer that enables application programmers to parse HTML documents and access the information using standard XML interfaces. The parser can scan HTML files and "fix up" many common mistakes that human (and computer) authors make in writing HTML documents. NekoHTML adds missing parent elements; automatically closes elements with optional end tags; and can handle mismatched inline element tags.
- ✓ Okhttp: is an open source project designed to be an efficient HTTP client. It supports the SPDY protocol. SPDY is the basis for HTTP 2.0 and allows multiple HTTP requests to be multiplexed over one socket connection.
- ✓ Okio: You'll also need Okio, which OkHttp uses for fast I/O and resizable buffers,, Okio is a library that complements java.io and java.nio to make it much easier to access, store, and process your data.
- ✓ Reactive-streams: Reactive Streams is an initiative to provide a standard for asynchronous stream processing with non-blocking back pressure. This encompasses efforts aimed at runtime environments (JVM and JavaScript) as well as network protocols.
- ✓ Rxjava: Reactive Streams is an initiative to provide a standard for asynchronous stream processing with non-blocking back pressure. This encompasses efforts aimed at runtime environments (JVM and JavaScript) as well as network protocols.

3.6. Issues of Reliability and Validity

Interview questions were designed in the way helps to answer the research questions and respondents who are visually impaired were chosen for an interview. The result was organized in a group and compared with other researcher’s result for normal friend peoples. Readily available literature was carefully selected from the authenticated sources. The Scientific research method was applied to the interpretation and explanation of findings. Finally, the proposed algorithm has been evaluated by giving a search query and search result of the re-ranking application is compared with Google results..

CHAPTER 4

4. PROPOSED PAGE RE-RANKING ALGORITHM

This research has aimed to develop accessible Amharic voice based web searching browser plug-in and enhance the existing content based page re-ranking algorithm for the visually impaired internet users to access the relevant web pages. To do this, Firstly, we have studied how the current page re-ranking algorithm operates and which search results are relevant for the visually impaired in search engines results. Secondly, the voice based search browser plug-in which enable us to search using voice are presented and finally integration of the algorithm with the voice has be identified.

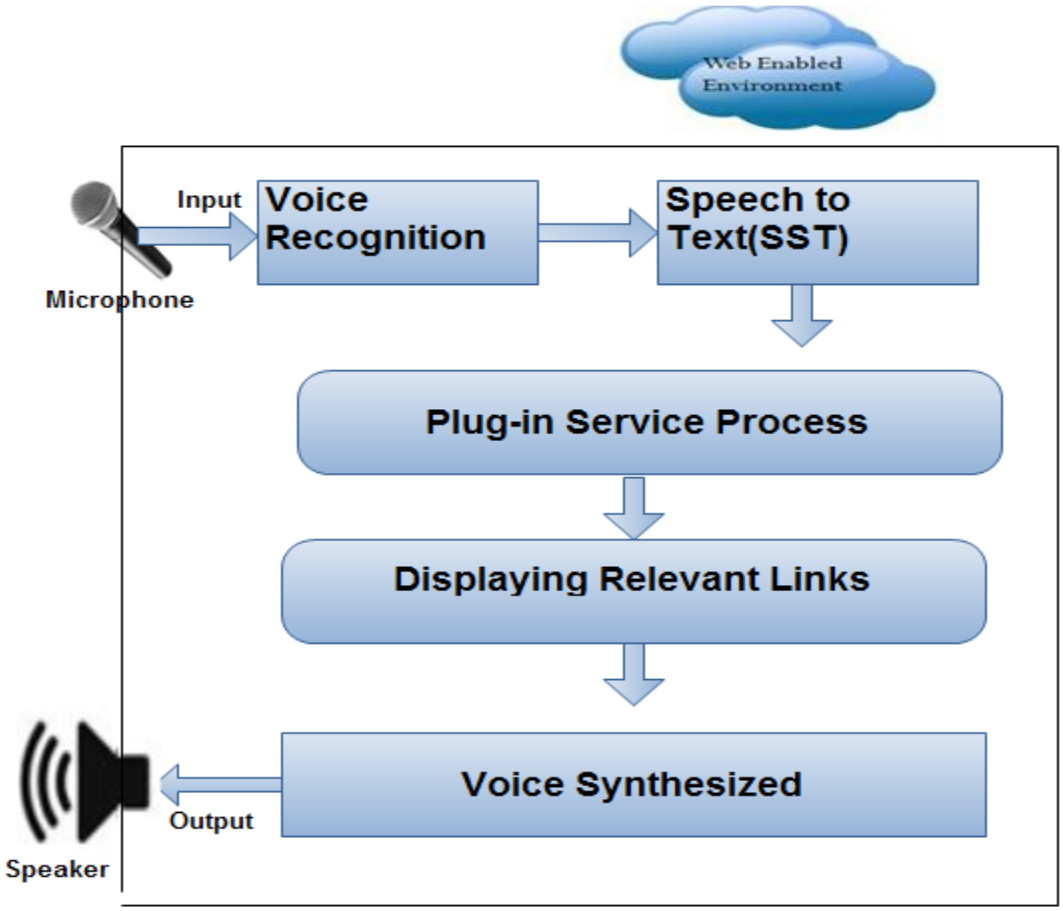

Figure 7: General frame work of the system

4.1. Voice Recognition

Voice or speech recognition is the ability of a machine or program to receive and interpret dictation, or to understand and carry out spoken commands. The moment of recognizing a phrase is an event, which happens when the recognizer detects an input. An event handler is assigned to this event so there is access to the property of the event, such as the text of the result of the recognition. It is important to notice that, there is no access to all of the properties of a recognition event. By that, it is meant; this event always returns a phrase, regardless of it being what was actually said by the user. For example, consider a user saying “life” and the system detects “like”. That is because these two words have almost the same pronunciation. In such case, the system carries on the actions as the value being “like”, which not something that the user inputs. The whole process of detection of words, that is, the process of translating spoken input (voice patterns) to text-words is handled by Google Cloud Speech API, where voice patterns are matched and from those phrases are created.

Google Cloud Speech API enables developers to convert audio to text by applying powerful neural network models in an easy to use API. The API recognizes over 110 languages and variants, to support a global user base. Amharic is one of this languages. You can transcribe the text of users dictating to an application’s microphone, enable command-and-control through voice, or transcribe audio files, among many other use cases. Recognize audio uploaded in the request, and integrate with your audio storage on Google Cloud Storage, by using the same technology Google uses to power its own products. Speech API can stream text results, returning partial recognition results as they become available, with the recognized text appearing immediately while speaking. Alternatively, Speech API can return recognized text from audio stored in a file.

4.2. Speech to Text (SSP)

After recognizing voice it should have to be changed to text which is done in this stage. Google has a dep neural network based voice recognition and speech to text algorithm and we have used the Google voice recognition API to overcome the recognition process. The accuracy of the recognition is quite impressive which helped us to overcome the noise problem while detecting spoken words.

4.3. Plug-in Service (Re-ranking)

The proposed system helps the visually impaired learners to acquire and search for information required for their learning process. The plug-in allows the visually impaired to say keywords into the microphone and the recognized voice will be converted to text. The converted text will be sent to search for the information required to Google search engine.

Google Custom Search enables you to create a search engine for your website, your blog, or a collection of websites. You can configure your search engine to search both web pages and images. You can fine-tune the ranking, customize the look and feel of the search results, and invite your friends or trusted users to help you build your custom search engine. There are two main use cases for Custom Search - you can create a search engine that searches only the contents of one website (site search), or you can create one that focuses on a particular topic from multiple sites. You can use your expertise about a subject to tell Custom Search which websites to search, prioritize, or ignore. Because you know your users well, you can tailor the search engine to their interests.

With Google Custom Search, you can:

- Create custom search engines that search across a specified collection of sites or pages
- Enable image search for your site
- Customize the look and feel of search results, including adding search-as-you-type auto completions
- Add promotions to your search results
- Leverage structured data on your site to customize search results

Having this all functions our main aim of using Google custom search for this research purpose is to retrieve Google search engine results programmatically. After we retrieve the search results from Google search engine programmatically we can apply our algorithm application to re-rank the search results

The Custom Search API with JSON / Atom lets you develop applications to retrieve and display results from the Google Custom Search programmatically. With this API, we used use RESTful requests to get search results in JSON and save the document in a local server.

JSON, or JavaScript Object Notation is a lightweight data-interchange format. It is used primarily to transmit data between a server and web application, as an alternative to XML. Squarespace uses JSON to store and organize site content created with the CMS. It is easy for humans to read and write. It is easy for machines to parse and generate. It is based on a subset of the JavaScript Programming Language, Standard ECMA-262 3rd Edition - December 1999. JSON is a text format that is completely language independent but uses conventions that are familiar to programmers of the C-family of languages, including C, C++, C#, Java, JavaScript, Perl, Python, and many others. These properties make JSON an ideal data-interchange language.

JSON is built on two structures:

- A collection of name/value pairs. In various languages, this is realized as an *object*, record, struct, dictionary, hash table, keyed list, or associative array.
- An ordered list of values. In most languages, this is realized as an *array*, vector, list, or sequence.

After retrieving search results in JSON object format we will download each URLs to our local data base using the okhttp java library for scraping process. Scraping each and every URL we will extract total number of images with no description, total number of key words and total number of content words from each link. The scraping process can be observed in Figure 3.Using information extracted from each page and comparing the results from each link the re-ranking algorithm will re-rank and displays search results on the browser. The search results found from the re-ranking will be listed and the headings of the URLs are read out based on no. 1, 2, 3...10. When the visually impaired learner requires a desired URL, the number is mentioned through the microphone and the system opens the URL and contents of the URL will be read out. If the user wants to stop the system from reading the headings or entire pages, the can say „stop reading“ and the system will halt the reading and expects other commands.

Architecture design of the algorithm

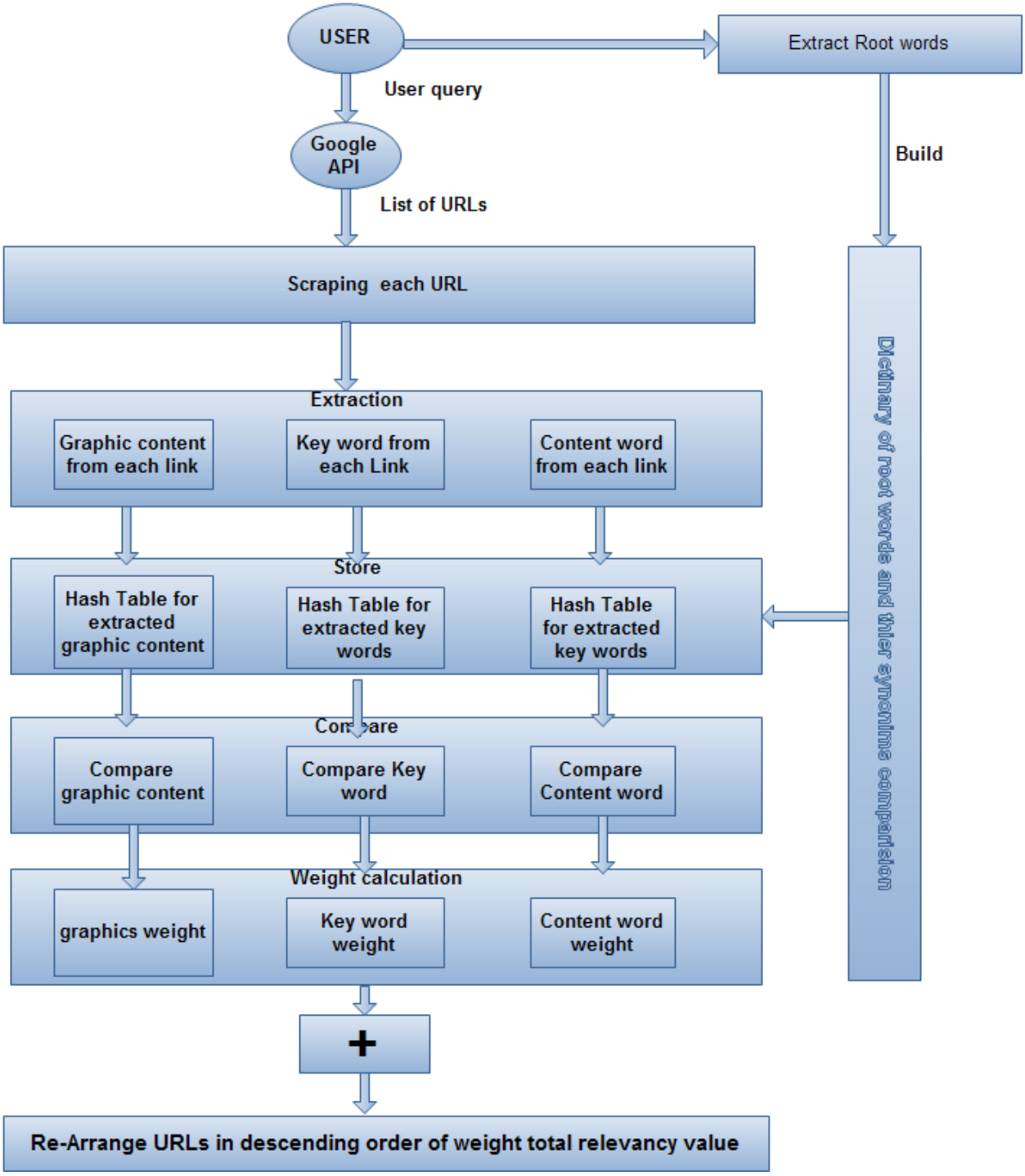

Figure 8: The general architecture of the proposed re-ranking Algorithm.

The scenario about the architectural design of the algorithm

1. The query put forth by users act as input to the Google APIs which are available to build the custom search. Root words are extracted from the user query by removing the stop words and synonyms are identified for each root word and a dictionary is constructed.

2. Google API produces a list of 'N' URLs. All URLs are scraped, and Keywords, content words and total number of graphics content are extracted from each URL and stored in the form of a hash table.

3. Keywords are those words that are within the HTML title tag. Content words are those words that are within the HTML body tag. Graphics contents are images and videos within HTML body tag .Keywords are compared with the root words and its synonyms in the dictionary, if they match a weight of ten is assigned to the keyword. Similarly content words are compared with the dictionary, if they match a weight of one is assigned to the content word.

4. Finally total relevancy is computed by summing all the keyword weights, the content word weights. If the total relevancy of URLs is equal the graphics content value of each URLs is compared and the one with less number of graphics content will be prioritized. This is repeated for all the URLs and based on the total relevancy values URLs are re-ranked by rearranging them in decreasing order.

4.4 Display relevant links on browser

After the re-ranking is done the URLs will be re-ranked based on weight calculation and they will be displayed. When the re-ranked results are displayed on the browser Text to Speech process will be activated. The text to speech process will read loud the title of the first ten search results for the user, but if the user want to halt the reading process at any time he can use the key word to halt the reading process. After the system read loud the titles of the search results the user will be asked to choose which link to open for e.g. If they are interested in the 4th search result they say "open four". The fourth search result will be opened in the new tab and content of the page will be read loud for the user.

4.5 Speech Synthesis

Speech synthesis refers to a computer's ability to produce sound that resembles human speech. Although they can't imitate the full spectrum of human cadences and intonations, speech synthesis systems can read text files and output them in a very intelligible. Many systems even allow the user to choose the type of voice [12] -for example, male or female, reading volume and speed can be controlled, emotions can be expressed [13],[14]. Speech synthesis systems are particularly valuable for seeing-impaired individuals.

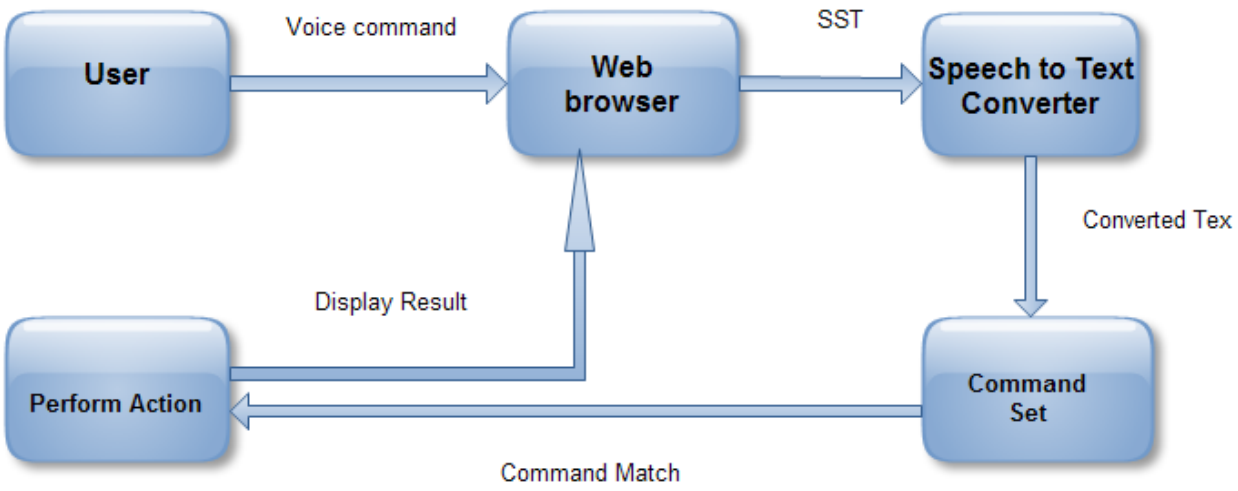

Figure 9: Speech synthesis model

Description

1. User: The User interacts with the browser by giving voice commands to the browser plug-in.

2. Web browser: The web browser takes the commands from the user and sends it to the Speech to Text Converter for text conversion.

3. Speech to Text (STT) Converter: This section converts the obtained voice commands to text and matches with the command set.

4. Command Set: This section contains all the commands the user can give to the browser and the given commands are matched here. If the command matches the specified action is performed else nothing happens.

5. Perform action: This section performs the specified action and displays it on the browser

These are the list sample commands in the browser:

Website name	It will be used to open websites stored in command set Eg. Google.com, Yahoo.com, etc.
Connect websocket	Connects the websocket.
Disconnect (አሰናክል)websocket	Disconnects websocket.
Enable(አንቃ)speech recognition	Enable speech recognition.
Enable (አንቃ)text to speech	Enable text to speech.
Forward	It will open next webpage.
Open(ክፈት) 1,2,3..10	It will open the specified link
Close (ዝጋ) 1,2,3,4,..10	It will close the specified link
Language Amharic	Change keyboard to Amharic.
ቋንቃ አንጊሊዝንያ	Change keyboard to English.
Search (ፈልግ)	Search's query when keyboard is in English.

Figure 10: Description of system key words

CHAPTER 5

5. RESULT AND DISCUSSION

The results obtained from Interview, observation, literature review were presented in this chapter. Interview questions were organized in the way that enables to get information related to the reaction of the disabled students about the current learning environment, availability of digital learning, availability of Internet and voice based web searching, gap between normal and disabled students, and reaction about the proposed solution A total of 8 responses were received from the targeted 10 potential respondents, which constitutes a 80% response rate for the interview. This chapter will bring in the presentation of the findings and analysis derived from structured interview, focused group discussion and personal observations.

Tables and diagrams have been used to facilitate a simplistic reader-friendly writing. Finally, the summary of this chapter is provided.

5.1. Presentation of the Interview Result

Background of the Respondents

The survey has been targeted for 10 Bahirdar University main campus (Peda) disabled (visually impaired) students. However, only 8 participants responded to the invitation to participate within the defined time frame. The following is the response rate according to the target versus actually received:

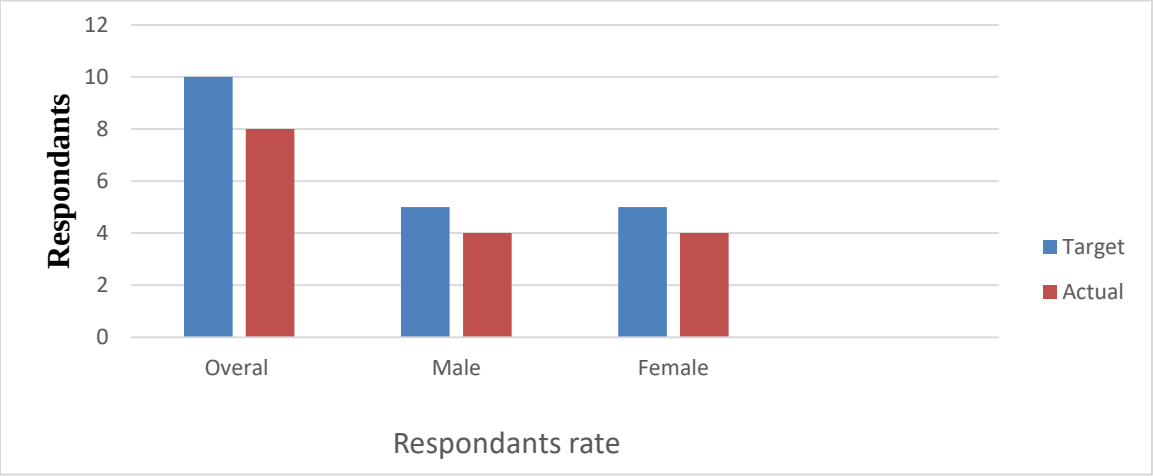

Figure 11: Respondents rate

The survey targeted 5 male and 5 female respondents to constitute 10 responses. Out of the total survey participants, 30% or 3 students are studying civics and ethical education and among this group the first one is first year first degree (BA) student, the second one is second year first degree (BA) student and the third one is third year first degree (BA) students. The other 30% or 3 students of the participants are studying history and among this group the first one is first year first degree (BA) student, the second one is second year first degree (BA) student and the third one is third year first degree (BA) student. The last 40% or 4 students are studying English and among this group the first two are first year first degree (BA) students, the third one is second year first degree (BA) student and the third one is third year first degree (BA) student.

5.2. Descriptive Analysis

A. Current learning environment:

All respondents explained the current learning environment. As the respondents explained there is no special class reserved for them, they take lecture class with other normal friends. They try to follow the lecture class and they use a digital recorder to record the voice of the teacher during lecture class incase if they miss anything and they listen to this record when they get home to recoup the lecture class. During exam time as they explained someone will be appointed to read the exam loudly and they will conduct the exam with the help of someone else.

B. Special service offered to help them in education area:

As the respondents explained there are limited services for them. The services the university provide are explained in the following table.

Disability type	Service type	Availability
Visually impaired	Braille	yes
	Tape recorder	yes
	Digital voice recorder	yes
	Audio Cassettes	no
	Battery for tape recorder	no
	Audio Books	no
	Computer with JAWS	no

	Internet services	no
	Braille printing/Embosser	no
	Exam reader/Scribe	no
	Computer usage skill	no
	Volunteer reading service	yes
	Volunteer recording services	no

Table 3: Special service offered to help them in education area:

C. Access to Internet:

All of the respondents explained as there is no special internet service for them. They have been questioned about the factors that affect their accessibility to internet. This has been clarified using a predefined set of factors from which respondents can choose one or more factors that influence this. Each factor is attended by all respondents. The responses have been as follows:

What could be the reasons you would claim for not having access to Internet?		
Answer Options	Response Percent	Response Count
I do not possess a computer	100%	8
IT equipments (computers and modems) are expensive	0%	8
Internet services are unavailable in my area	0%	8
There is no tools to access using voice search	100%	8
I do not know how to use internet	100%	8
There is no public internet access centers	0.0%	8
Internet charges are very high at Internet access centers	0.0%	8
Total Responses to the question		**8**

Table 4: Respondents reason for inaccessibility of internet

D. Digital learning environment

All of the respondents replied as they are not familiar with digital learning environment. Even if there is a digital library in the university, there are no audio books and there are no screen reader software installed on those computers to help them use this digital library.

What could be the reasons you would claim for not having access to Internet?		
Answer Options	**Response Percent**	**Response Count**
There is no digital library in the university	0	0
I don't know how to use digital library	8	8
Total Responses to the question		**8**

Table 5: Respondents reason for inaccessibility of digital learning

E. Gap between visually impaired and normal friends

Since there are no special supporting class and special facilities for the disabled persons in the education area, there is a big gap between this disabled and their normal friend. Both normal and disabled students take class in the same room but normal students do have many option for reference material to help them conduct their study in the university like digital library, internet service, and others which are inaccessible for the disabled students.

5.3. Participatory observation

An observation on how Bahirdar University visually impaired students conduct their class has been observed being in their class. During exam times observation on the exam are delivered has been viewed. How the visually impaired students access the digital library of Bahirdar university Peda campus to prepared for exam and to work their assignment have been observed.

CHAPTER SIX

6. PROTOTYPE DEVELOPMENT FOR THE PROPOSED FRAMEWORK AND VALIDATION OF RESULTS

6.1. Overview of the Prototype

In the previous chapter the general framework of the proposed system which contain five different layers and the architectural design of the algorithm is presented. In this chapter, we have discussed how the proposed browser plug-in is implemented and become usable to the visually impaired in seeking information via web. We put here, the detailed implementation process and the programming tools used to implement the framework.

6.2. The Development Environment

To implement the prototype the component of java NetBeans 8.0.2 IDE called JEE (Java Enterprise Edition) platform which is built on the top of java standard edition have been used. The java platform enterprise edition defines a sample standard that applies to all aspects of architecting, developing, and managing multitier server based application. The aim of the JEE platform is to provide developers with a powerful set of APIs while shortening development time, reducing application complexity, and improving application performance [11]. In this prototype, we especially used web technologies like; web socket API, WebSocket provides a welcomed alternative to the AJAX technologies we've been making use of over the past few years. This new API provides a method to push messages from client to server efficiently and with a simple syntax. WebSocket API is the next generation method of asynchronous communication from client to server. Communication takes place over single TCP socket using the ws (unsecure) or wss (secure) protocol and can be used by any client or server application. WebSocket is currently being standardized by the W3C. WebSocket is currently implemented in Firefox 4, Chrome 4, Opera 10.70, and Safari 5. The overall implementation of custom Websocket for our system server is explained below in detail. The have used JavaScript Web Speech API which makes it easy to add speech recognition to our system. This API allows fine control and flexibility over the speech recognition capabilities in Chrome version 25 and later. For Scraping URLs we have used different java libraries which are:

Websocket-1.3.4.jar: A barebones WebSocket client and server implementation written 100% in Java.

Boilerpipe-1.1.1.jar: The boilerpipe library provides algorithms to detect and remove the surplus "clutter" (boilerplate, templates) around the main textual content of a website.

Gson-2.6.2.jar: Gson is a Java library that can be used to convert Java Objects into their JSON representation. It can also be used to convert a JSON string to an equivalent Java object. Gson can work with arbitrary Java objects including pre-existing objects that you do not have source-code of.

Nekohtml-1.9.20.jar: NekoHTML is a simple HTML scanner and tag balancer that enables application programmers to parse HTML documents and access the information using standard XML interfaces. The parser can scan HTML files and "fix up" many common mistakes that human (and computer) authors make in writing HTML documents. NekoHTML adds missing parent elements; automatically closes elements with optional end tags; and can handle mismatched inline element tags.

Okhttp-3.2.0.jar: is an open source project designed to be an efficient HTTP client. It supports the SPDY protocol. SPDY is the basis for HTTP 2.0 and allows multiple HTTP requests to be multiplexed over one socket connection.

Okio-1.1.0.jar== You'll also need Okio, which OkHttp uses for fast I/O and resizable buffers,, Okio is a library that complements java.io and java.nio to make it much easier to access, store, and process your data.

Reactive-streams-1.0.1-RC2.jar: Reactive Streams is an initiative to provide a standard for asynchronous stream processing with non-blocking back pressure. This encompasses efforts aimed at runtime environments (JVM and JavaScript) as well as network protocols.

Rxjava-2.1.2.jar: Reactive Streams is an initiative to provide a standard for asynchronous stream processing with non-blocking back pressure. This encompasses efforts aimed at runtime environments (JVM and JavaScript) as well as network protocols.

6.3. Components of the system

6.3.1. WebSocket server

A WebSocket server is a TCP application listening on any port of a server that follows a specific protocol, simple as that. We have implemented custom WebSocket server using java script code. The screenshots of the server and explanation of it components will be as follows

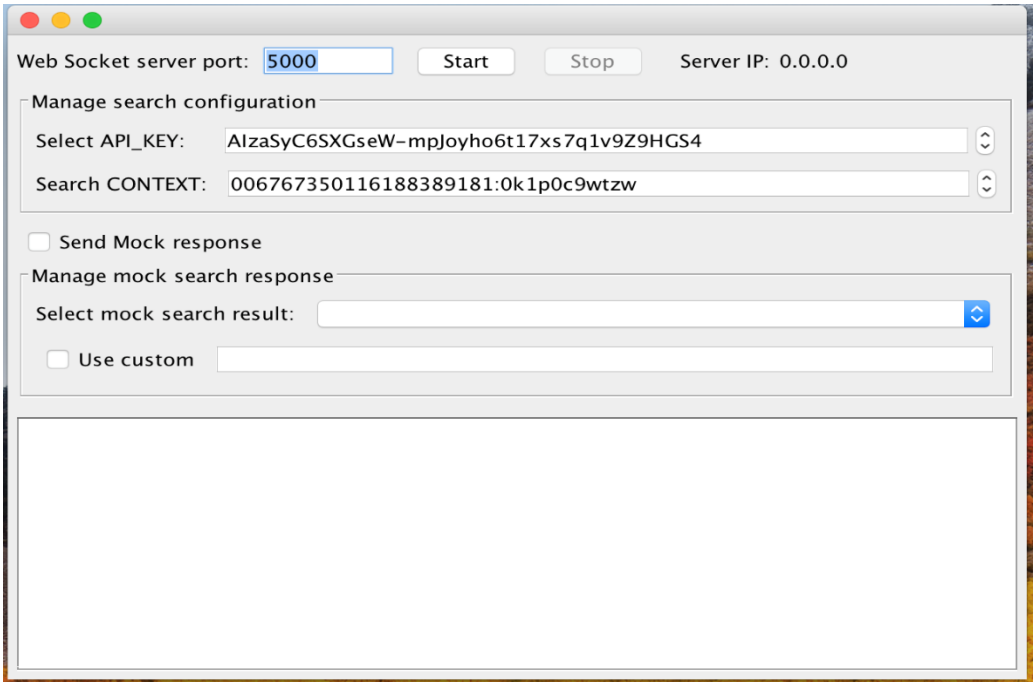

Figure 12: WebSocket server screen shoot

WebSockets provide a persistent connection between a client and server that both parties can use to start sending data at any time. The client establishes a WebSocket connection through a process known as the WebSocket handshake. This process starts with the client sending a regular HTTP request to the server. An Upgrade header is included in this request that informs the server that the client wishes to establish a WebSocket connection.

WebSocket URLs use the ws scheme. There is also wss for secure WebSocket connections which is the equivalent of HTT. If the server supports the WebSocket protocol, it agrees to the upgrade and communicates this through an Upgrade header in the response.

Now that the handshake is complete the initial HTTP connection is replaced by a WebSocket connection that uses the same underlying TCP/IP connection. At this point either party can starting sending data.

With WebSockets you can transfer as much data as you like without incurring the overhead associated with traditional HTTP requests. Data is transferred through a WebSocket as *messages*, each of which consists of one or more *frames* containing the data you are sending (the payload). In order to ensure the message can be properly reconstructed when it reaches the client each frame is prefixed with 4-12 bytes of data about the payload. Using this frame-based messaging system helps to reduce the amount of non-payload data that is transferred, leading to significant reductions in latency.

Opening Connections

To open or start the a connection we have created a new web sockeck connection.

```
ws = new WebSocket('ws://localhost:5000');
```

Once the connection has been established the start event will be fired on the WebSocket instance.

Handling Errors

You can handle any errors that occur by listening out for the error event.

For our prototype application added some code that will log any errors to the console.

```
socket.onerror = function(error) {
  console.log(' Lost connection: ' + error);
};
```

Sending Messages

To send a message through the WebSocket connection you call the send()method on your WebSocket instance; passing in the data you want to transfer.

```
function sendMessage() { var text = document.getElementById('userInput').value; var request = {
"requestType": 0, "query": text }; ws.send(JSON.stringify(request)); }
```

Receiving Messages

When a message is received the message event is fired. This event includes a property called data that can be used to access the contents of the message.

```
function showMessage(text) { console.log(text); document.getElementById('message').innerHTML = text; } var ws = new WebSocket('ws://localhost:5000'); showMessage('Connecting...'); ws.onopen = function () { showMessage('Connected!'); };

};
```

Closing Connections

Once you're done with your WebSocket you can terminate the connection using the close() method.

```
ws.onclose = function () { showMessage('Lost connection'); };
```

After the connection has been closed the browser will fire a close event. Attaching an event listener to the close event allows you to perform any clean up that you might need to do. The other important parts on the screenshot is managing the search configuration. In this part there are two basic things which are Select API_KEY and Search CONTEXT. The API_KEY used to access the Google search engine results programmatically whereas the CONTEXT is contexts the search process uses.

6.3.2. Google custom search

The aim of using Google custom search for this research purpose is to retrieve Google search engine results programmatically. After we retrieve the search results from Google search engine programmatically we can apply our algorithm to re-rank the search results. With this API, we used RESTful requests to get search results in JSON and save the document in a local server. JSON, or JavaScript Object Notation is a lightweight data-interchange format.

Figure 3 shows sample JSON object format for the query **"የ አትዮጲያ ታሪክ** ".

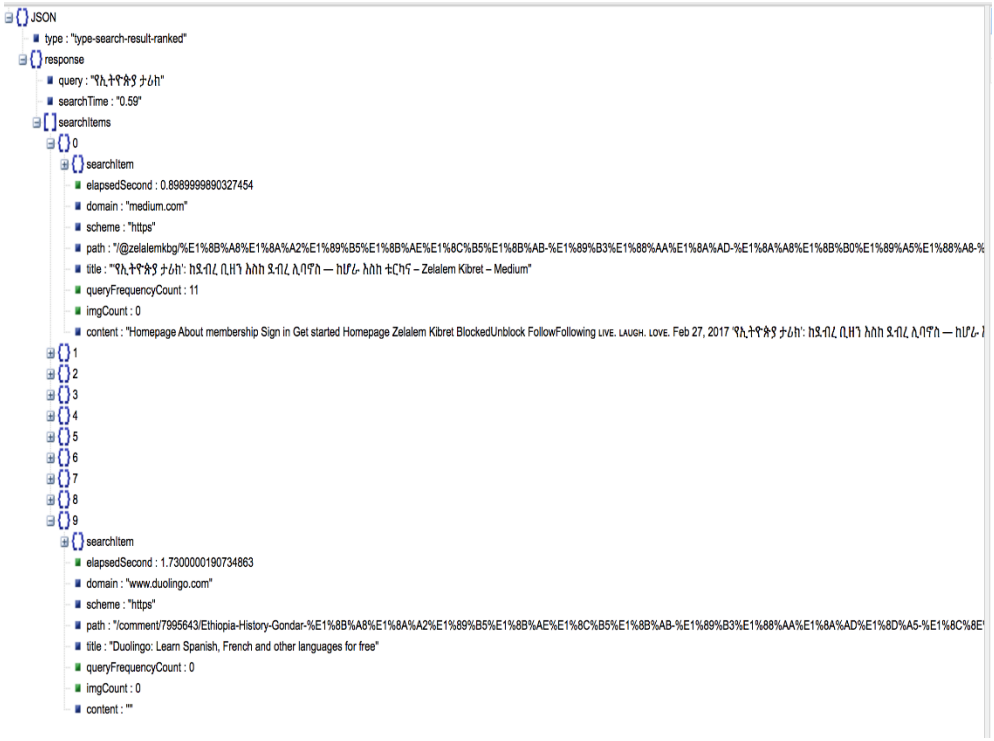

Figure 13: Sample JSON object for query "የ አትዮጲያ ታሪክ ".

6.3.3. Web scraping

Web scraping (**web harvesting** or **web data extraction**) is data scraping used for extracting data from websites. The World Wide Web can be directly accessed by web scraping software using Hypertext Transfer Protocol or trough web browser. While web scraping can be done manually by a software user, the term typically refers to automated processes implemented using a bot or web crawler. It is a form of copying, in which specific data is gathered and copied from the web, typically into a central local database or spreadsheet, for later retrieval or analysis.

Web scraping a web page involves fetching it and extracting from it. Fetching is the downloading of a page (which a browser does when you view the page). Therefore, web crawling is a main component of web scraping, to fetch pages for later processing. Once fetched, then extraction can take place. The content of a page may be parsed, searched, reformatted, its data copied into a spreadsheet, and so on. Web scrapers typically take something out of a page, to make use of it for another purpose somewhere else. Web scraping involve listening to data feeds from web servers. We have used JSON as a transport storage mechanism between the client and the web server. After

scraping each URL from Google custom search results we have extracted total number of graphics contents, Key word and content word from each links.

After retrieving values from the scarping results the algorithm will compare total weight of each links and it will re-arrange search results based on the calculated total weight of each links

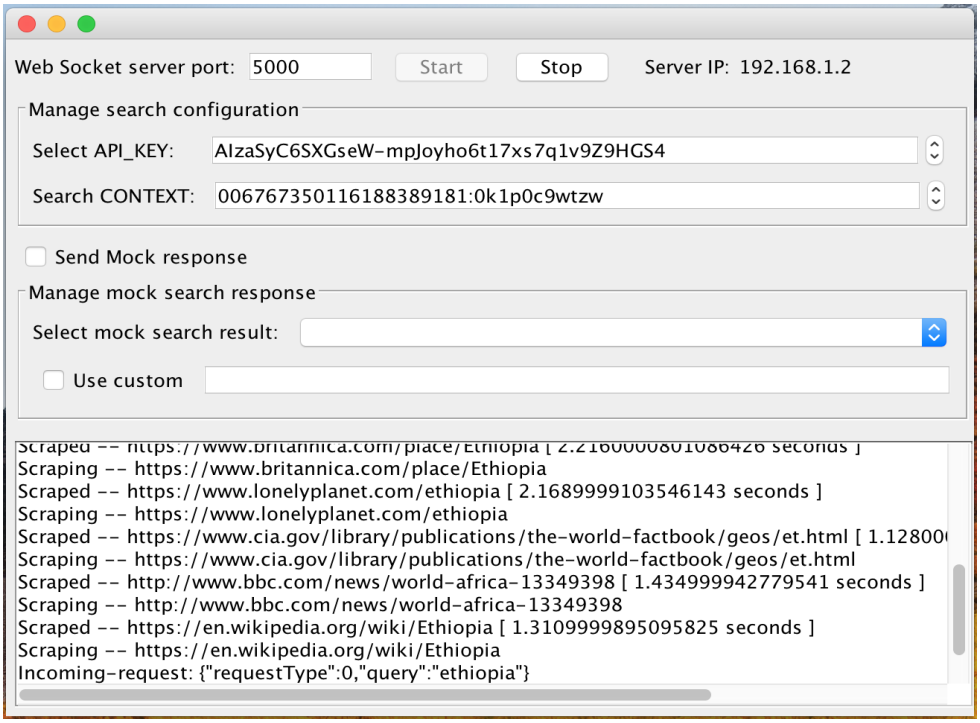

Figure 14: Web scraping process

After retrieving values from the scarping results the algorithm will compare total weight of each links and it will re-arrange search results based on the calculated total weight of each link

The original search result part on the right side of the screenshot displays Google search results and the re-ranked search part on the left side of the screenshot displays the re-ranked search results.

6.4. Validation of Results

Proposed system is tested by giving the voice query ““የ አትዮጲያ ታሪክ ”.”. Figure 1 shows the search result for the query. The original search result part on the right side of the screenshot displays Google search results and the re-ranked search part on the left side of the screenshot displays the re-ranked search results.

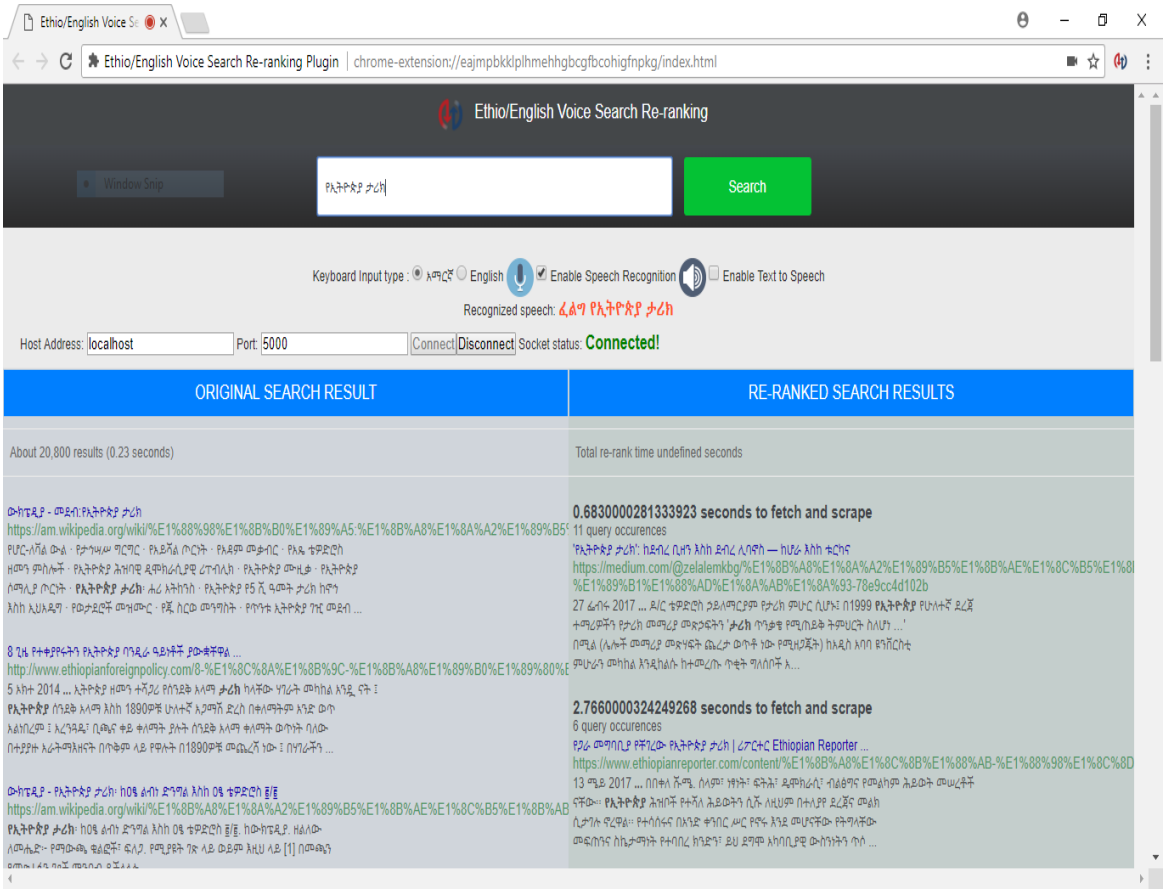

Figure 15: Proposed system search result for query "የ አትዮጲያ ታሪክ ".

Performance evaluation

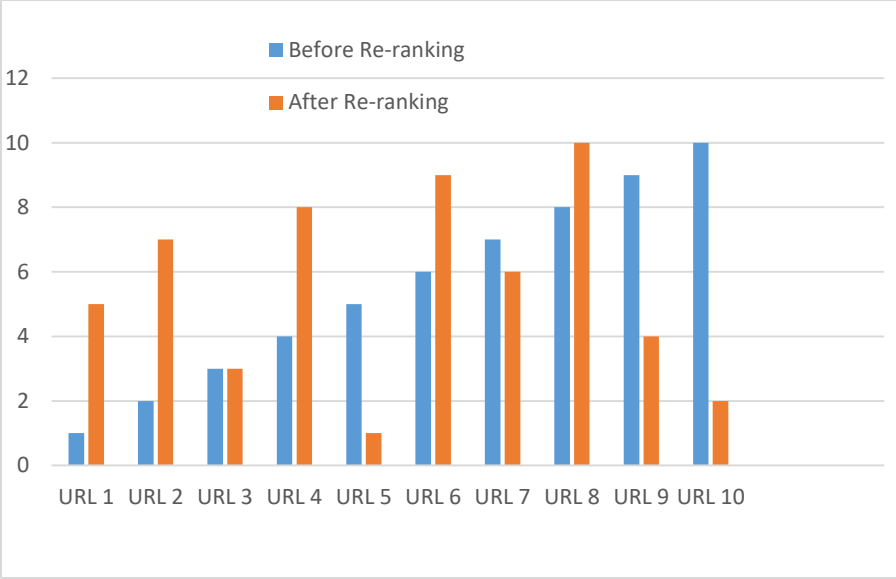

Figure 16: before and after re-rank performance evaluation.

Figure 16 shows that the URL1 which was ranked first by the search engine is now ranked second, URL2 which was ranked second now ranks first. Similarly, it is observed that the other ranking variation before and after re-ranking. The performance of the proposed system is evaluated by computing the precession using the true positive value (TP) and false positive value (FP). Formula to find the precession is as given below.

Precession= TP / (TP + FP) ------>Eq (1)

TP = Correct Identification

FP =Incorrect Identification

The table 2 shows the comparison of Google rank, proposed rank and manual rank by two different users.

URLs	Google Rank	Manual Rank	Proposed Rank
URL1	1	4	5
URL2	2	7	7
URL3	3	3	3
URL4	4	8	8
URL5	5	1	1
URL6	6	10	9

URL7	7	6	6
URL8	8	9	10
URL9	9	5	4
URL10	10	2	2

Table 6: Rank comparison

Google has true positive value of 1 for URL3 and false positive value of 9 for URLs 1, 2, 4,5, 6,7,8,9,10. Proposed system has true positive value of 6 for URLs 1, 2, 3, 6, 7, 8 and false positive value of 4 for URLs 4,5, 9,10. The following graph shows the precession for Google rank and the proposed rank.

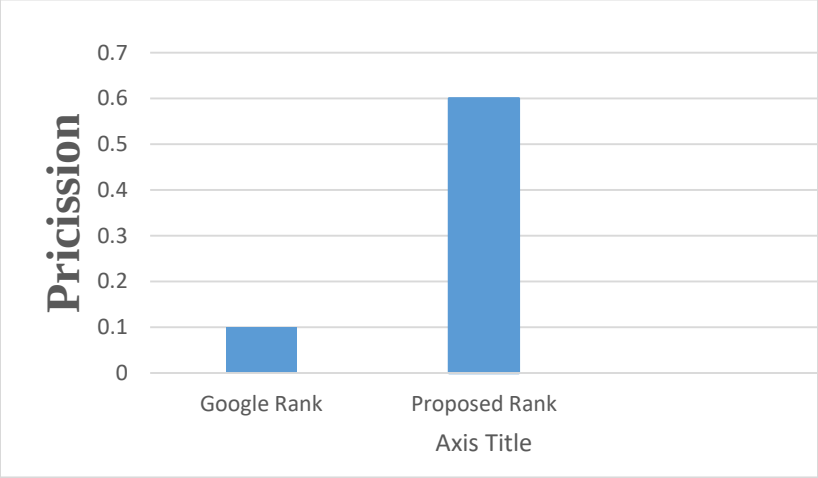

Table 7: Precision of Google rank and proposed rank.

CHAPTER SEVEN

7. CONCLUSION AND FUTURE WORKS

7.1. Conclusion

This research has been begun with the intent to seek a solution to the problem that we have observed from the environment. The problem that initiates to do this research is the inaccessibility of webpage for the visually impaired users. We have inspired by this problem and followed a scientific research method by asking three research question which enables to understand how the how the visually impaired access the internet, how relevancy of search results from visually impaired perspective, and also what techniques and technologies are there to solve the existing problem. Interview questions are designed in the way that enables to capture the problems visually impaired students face in accessing information from the web ,After formulating the problem in digital learning we have reviewed a number of papers to give a solution to the existing problem. We have seen that many developers are eager in designing and developing their application fully with animation graphics. Their main purpose is only to make their applications look fantastic and impressive which made difficult for visually impaired to access the net.

This research can contribute on creating awareness of accessibility issues in designing and developing the speech browser plug-in for visually impaired as a tool for seeking information effectively. The findings from this research support the idea on enhancing the development of application in a form that is accessible to everybody. This can be referred to the accessibility implications for the websites' developers. First, it is necessary for the websites' developer to concentrate their design comply with the accessibility guidelines developed by W3C. Second, Accessibility issues need to be addressed by all web developers to enhance the accessibility and usability among the visually impaired learners to enable them in seeking information via Internet Finally, the speech synthesizer developed by the developers must consider the accuracy of the synthetic pronunciation in order to bring the correct information that is meaningful to the visually impaired. Adapting the accessibility technology in web development can also give benefits and opportunities for the sighted learners in accessing information. Therefore, an accessible Amharic voice based search browser plug-in is an essential technology that greatly benefits the disabled especially the visually impaired in seeking information. Finally, it is hopeful that all web

developers are aware on the accessibility matter when developing their web applications in order to give equal access and opportunity to people with disabilities, specifically the visually impaired. This can help them to be more active in their participation in the society.

The results produced by the search engine are enormous and irrelevant to the user context. The proposed approach makes use of the text scraping on the web pages listed by the search engine to rank and increase the relevancy of pages and provide visually impaired users with quality search results. The experimental results show that the results obtained by this approach have more relevant web pages displayed at the higher positions to the user.

7.2. Future Works

In this thesis we have seen how much the visually impaired are ignored in the internet zone. We have managed to propose a solution to the problem which is developing a prototype of browser plugin which works with two languages Amharic and English language but There are millions of visually impairers around the world who speak different language so in the future there should have to be a multi lingual system which help the visually impaired users surf the net easily.

References

1. Guan-yu LI, Sui-ming YU and Sha- ional conference on network and parallel computing, pp.1010 -1015, 2007.
2. Jerome Euzenat, Pavel Shvaiko, "Ontology Matching", Springer-Verlag, Berlin Heidelberg(DE),2007 , isbn:3-540- 49611-4
3. S. Kadry, A. Kalakesh, "On the Improvement of Weighted Page Content Rank", *Journal of Advances in Computer Networks,* vol.1,no.2, pp. 110-114, June 2013.
4. R. Kosala , H. Blockeel, "Web Mining Research: A Survey", *SIGKDD Explorations, Newsletter of the ACM Special Interest Group on Knowledge Discovery and Data Mining, vol.2,* no.1, pp. 1-15,Feb 2000.
5. Wenpu Xing and Ali Ghorbani, "Weighted PageRank Algorithm", In proceedings of the 2rd Annual Conference on Communication Networks & Services Research, PP. 305-314, 2004.
6. Sung Jin Kim and Sang Ho Lee, "An Improved Computation of the PageRank Algorithm", In proceedings of the European Conference on Information Retrieval (ECIR), 2002.
7. Renjini L, Prof. Ratheesh T K: "Re-ranking algorithm for better personalized web search"
8. Harish Kumar B T, Vibha Lakshmikantha, Venugopal K R: "Content Based Web Page Re-Ranking Using Relevancy Algorithm", *Quest Journals Journal of Electronics and Communication Engineering ResearchVolume 2 ~ Issue 7(2014) pp: 01-08*
9. Juhi Agrawal, Nishkarsh Sharma, Pratik Kumar, Vishesh Parshav, R H Goudar: "Ranking of Searched Documents Using Semantic Technology" nternational Conference On DESIGN AND MANUFACTURING, IConDM 2013, Procedia Engineering 64 (2013) 1 – 7)
10. Dharmarajan Kaliyaperumal, "Current literature review-web mining" Research Gates, *Issue (*September 2014*) pp: 31-42.*
11. Jendrock, E., Navarro, R. C., Evans, I., Haase, K. and Markito, W. (2014). Java platform enterprise edition. Oracle.
12. Arnaud Ramey, Javier F. Gorostiza, Miguel A. Salichs, "A Social Robot as an Aloud Reader: Putting together Recognition and Synthesis of Voice and Gestures for HRI Experimentation" International conference on Human-Robot Interaction, March 2012.
13. Ramin Halavati, Saeed Bagheri shouraki and Saman Harati Zadeh, "Recognition of human speech phonemes using a novel fuzzy approach" ,Journal appliedto soft computing, Volume 7, Issue 3, June 2007
14. Heiga Zen and Tomoki Toda " An Overview of Nitech HMM-based Speech Synthesis System for Blizzard challenge", Blizzard Workshop-2005 .

15. Sergio Luján-Mora, Firas Masri. Integration of Web Accessibility into Agile Methods. Proceedings of the 14th International Conference on Enterprise Information Systems (ICEIS 2012), Volume 3, p. 123-127, Wroclaw (Poland), June 28 - July 1 2012. ISBN: 978-989-8565-12-9.
16. Web Accessibility. Retrieved June 15, 2017, from http://www. accessibility@oregonstate.edu
17. A Survey of Google's PageRank Retrieved June 25, 2017, from http://pr.efactory.de/e-references.shtml
18. Ricardo Baeza-Yates and Emilio Davis ,"Web page ranking using link attributes" , In proceedings of the 13th international World Wide Web conference on Alternate track papers & posters, PP.328-329, 2004.
19. Ko Fujimura, Takafumi Inoue and Masayuki Sugisaki,, "The EigenRumor Algorithm for Ranking Blogs", In WWW 2005 2nd Annual Workshop on the Weblogging Ecosystem, 2005.
20. Ali Mohammad Zareh Bidoki and Nasser Yazdani, "DistanceRank: An Iintelligent Ranking Algorithm for Web Pages", Information Processing and Management, 2007.
21. H Jiang et al., "TIMERANK: A Method of Improving Ranking Scores by Visited Time", In proceedings of the Seventh International Conference on Machine Learning and Cybernetics, Kunming, 12-15 July 2008
22. Shen Jie,Chen Chen,Zhang Hui,Sun Rong-Shuang,Zhu Yan and He Kun, "TagRank: A New Rank Algorithm for Webpage Based on Social Web" In proceedings of the International Conference on Computer Science and Information Technology,2008.
23. Fabrizio Lamberti, Andrea Sanna and Claudio Demartini , "A Relation-Based Page Rank Algorithm for. Semantic Web Search Engines", In IEEE Transaction of KDE, Vol. 21, No. 1, Jan 2009.
24. Dilip Kumar Sharma et al. A Comparative Analysis of Web Page Ranking Algorithms / (IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 08, 2010, 2670-2676.
25. AlJahdali, H., Albatli, A., Garraghan, p., Townend, p., Lau, l., & Xu, j. (2014). Multi-Tenancy in Cloud Computing. Leeds, United Kingdom: White rose.

Appendix

Sample codes

Core service implementation

```
public Single<RankedSearchResult> search(String query) {
    String key = this.callback.getSelectedSearchApiKey();
    String context = this.callback.getSelectedSearchContext();
    return searchApi.fetch(key, context, query, 1, 10)
            .map(new Function<String, SearchResponse>() {
                @Override
                public SearchResponse apply(String t) throws Exception {
                    return new Gson().fromJson(t, SearchResponse.class);
                }
            })
            .map(new Function<SearchResponse, SearchResult>() {
                @Override
                public SearchResult apply(SearchResponse response) throws Exception {
                    return new SearchResultMapper().map(response);
                }
            })
            .doOnSuccess(new Consumer<SearchResult>() {
                @Override
                public void accept(SearchResult t) throws Exception {
                    if (callback != null) {
                        callback.onLoadSearchResult(t);
                    }
                }
            })
```

```
        .flatMap(new Function<SearchResult, SingleSource<RankedSearchResult>>() {
          @Override
          public SingleSource<RankedSearchResult> apply(SearchResult result) throws Exception {
            return Flowable.fromIterable(result.getSearchItems())
                .flatMap(new Function<SearchItem, Publisher<ScrapeResponse>>() {
                  @Override
                  public Publisher<ScrapeResponse> apply(SearchItem t) throws Exception {
                    if (callback != null) {
                      callback.onNotifyScrapeStarted(t.link);
                    }
                    return scraperApi.fetch(new URL(t.link)).toFlowable();
                  }
                })
                .doOnNext(new Consumer<ScrapeResponse>() {
                  @Override
                  public void accept(ScrapeResponse t) throws Exception {
                    if (callback != null) {
                      callback.onNotifyScrapedContent(t)      }
                          })
                .toList()
                .map(new AbstractFunction<List<ScrapeResponse>, RankedSearchResult, SearchResult>(result) {

                  @Override
                  protected RankedSearchResult apply(List<ScrapeResponse> input, SearchResult subject) {
                    List<SearchItem> searchItems = subject.getSearchItems();
                    String query = subject.getQuery();
```

```
                    RankedSearchResult result = new RankedSearchResult(query,
subject.getFormattedSearchTime());
                    List<RankedSearchItem> rankedSearchItem = new LinkedList<>();

                    for (int i = 0; i < searchItems.size(); i++) {
                       RankedSearchItem rankItem = new RankedSearchItem();

                       SearchItem searchItem = searchItems.get(i);

                       rankItem.setSearchItem(searchItem);

                       ScrapeResponse sr = input.get(i);
                       if (sr != null) {
                          ScrapedContent sc = sr.getResponse();
                          if (sc != null) {
                             String content = sc.getContent();
                             int queryOccurencesCount =
StringUtils.getQueryOccurenceCount(content.toLowerCase(), query.toLowerCase());
                             List<Image> imgUrls = sc.getImgUrls();

                             try {
                                rankItem.setContent(new String(content.getBytes("UTF-8"),
"UTF-8"));
                             } catch (UnsupportedEncodingException ex) {

Logger.getLogger(CoreServiceImpl.class.getName()).log(Level.SEVERE, null, ex);
                             }
                             rankItem.setQueryFrequencyCount(queryOccurencesCount);
                             rankItem.setImgCount(imgUrls != null ? imgUrls.size() : 0);
                             rankItem.setElapsedSecond(sc.getElapsedSecond());
                             rankItem.setTitle(sc.getTitle());
```

```
                rankItem.setPath(sc.getPath());
                rankItem.setDomain(sc.getDomain());
                rankItem.setScheme(sc.getScheme());
            }
        }
        rankedSearchItem.add(rankItem);
    }

    Collections.sort(rankedSearchItem, new Comparator<RankedSearchItem>() {
        @Override
        public int compare(RankedSearchItem o1, RankedSearchItem o2) {
            if (o1 != null && o2 != null) {
                Integer queryFrequencyCount1 = o1.getQueryFrequencyCount();
                Integer queryFrequencyCount2 = o2.getQueryFrequencyCount();
                Integer imgCount1 = o1.getImgCount();
                Integer imgCount2 = o2.getImgCount();

                int diff = queryFrequencyCount2.compareTo(queryFrequencyCount1);
                if (diff != 0) {
                    return diff;
                }
                return imgCount2.compareTo(imgCount1);
            } else if (o1 == null && o2 == null) {
                return 0;
            } else {
                return 1;
            }
        }
```

```
                    });

                    result.setSearchItems(rankedSearchItem);
                    if (callback != null) {
                        callback.onSearchContentsRanked(result);
                    }
                    return result;
                }
            });
        }
    });
  }
}
```