Sindh Univ. Res. Jour. (Sci. Ser.) Vol. 48 (4) 945-948 (2016)

SINDH UNIVERSITY RESEARCH JOURNAL (SCIENCE SERIES)

Design and Development of Unicode based Sindhi Language Thesaurus

Z. BHATTI++, I. A. ISMAILI++, S. ZARDARI*, W. J. SOOMRO

Institute of Information and Communication Technology, University of Sindh, Jamshoro,

Received 20th June 2016 and Revised 17th November 2016

Abstract: Recent advancements in Computer Technologies have rapidly revolutionized the world. These advancements have immensely increased the need of localization of computer technologies in regional languages and for convenient natural language processing. In this paper, the problem of design and development of Unicode based digital thesaurus is discussed for Sindhi language. Sindhi is one of the oldest and richest languages of the world with a very rich linguistics and literary text. The development of digital Sindhi Thesaurus application is done on Java platform, using hash table structure to act as a database for storing word repository. The hash table structure provides a convenient and easy to implement data structure with multiple advantages of speed and ease of use. The words data is saved as a java bean object in the hash table element with the primary Sindhi word as key. The object is then retrieved and displayed on a user interface of thesaurus.

Keywords: Sindhi Language, Sindhi Thesaurus, Sindhi Language, Sindhi computing, Unicode

1. INTRODUCTION

In this ever changing modern world with new tendencies, people are always connected with the world of computer and internet. The literary work requires various linguistics based sources such as dictionaries, thesaurus, online database and digital libraries. Language has always been the source of development for a country as well as a nation. Sindhi is one of the oldest and richest languages with very rich language resources.

Thesaurus is similar to dictionary; however it contains words and their synonyms or similar meaning words rather than definition like in dictionaries. A book that is like a dictionary, but in which the word are arranged in groups that has similar meanings (Wikipedia, 2014).Thesauruses are very commonly used by all sorts of people working in varying fields. Thesaurus provides a vocabulary of words, to a word in query, with similarly meaning words. Every common and popular linguistics based language has their thesaurus especially digital thesaurus, in this technological era. Nowadays, it is considered old fashioned and hectic process to visit the libraries and look for new books and new words because everyone is connected with internet. Therefore, a digital Sindhi thesaurus is the best source to save the time and energy, in this fast flourishing world.

The research and development is Sindhi Computing is progressing slowly and gradually making its way into digital age by means of research conducted in Sindhi dictionary(Soomro, *et al,* 2004) and repository design and usage; from developing Sindhi Typing Tutor (Bhatti *et al,* 2013a), Sindhi Academic portal (Bhatti, *et al.,* 2013b), Sindhi Spell Checker (Bhatti, *et al.,* 2014a,Bhatti, *et al.,* 2015) to developing GUI's in Sindhi (Ismaili, *et al.,* 2011). This Sindhi computing field is growing. Yet, there are further areas and problems that need to be addressed, and this paper addresses one such problem of designing and developing a digital Unicode based Sindhi thesaurus.

2. DESIGN AND DEVELOPMENT METHODOLOGY

There are few dictionaries available like Sindhi to Sindhi and Sindhi to English in a computer form(Bhatti, *et al.,* 2014b , Hakro, *et al.,* 2014 and Shah, *et al.,* 2011).We extended their methodology and developed this digital thesaurus project. The basic architecture of application is based on Java technology. The use of java, as an underlying tool, allows the system to work platform independently and render Unicode characters on screen correctly. This project also uses similar hashtable structure as discussed in (Bhatti, *et al.,* 2014b) (Ismaili, *et al.,* 2012) to store the Sindhi words in a repository.

2.1 Word Sources

Various sources and people were contacted and approached, in-order to gather Sindhi words and resources regarding their synonyms. We visited Sindhi Language Authority (SLA) and met Chairperson Fahmeeda Hussain, and Taj Joyoto discuss this project. We also met some writers like Sir Atta Mohammad Bhambhroo to seek and collect resources for Sindhi synonyms. We used Sindhi Lugat Dictionary (By Dr: Nabi Bux Balouch) to compile initial list of words, and their synonyms were also extracted from it. A comprehensive English-Sindhi Dictionary (By Memon) was also used as per suggestion of SLA, along with few other dictionary sources such as Colour Oxford Dictionary and

++Correspondence Author: Email: zeeshan.bhatti@usindh.edu.pk , iai_a@yahoo.com
*Department of Computer Science and Information Technology, NED University, Karachi.
shehnilaz@neduet.edu.pk, waseem.soomro@usindh.edu.pk,

thesaurus, English Oxford advance learners Dictionary (Shah, *et al.,* 2011), for the purpose of developing authentic Sindhi Words repository (Abbasi, *et al.,* 2015 and Mughal, 2016).

2.2Keyboard Preferences

The most essential and crucial part for developing a regional level based application, is to have an easiest and simplest means of entering the regional script. For Sindhi Language, which is an Arabic style script, we developed a custom On Screen keyboard, so that the user can easily enter Sindhi text in search text field. The Sindhi keyboard design layout has been adopted from approved keyboard layout for Sindhi Language by SLA. The snapshot custom built on-screen Sindhi keyboard for this project is shown in **(Fig. 1).**

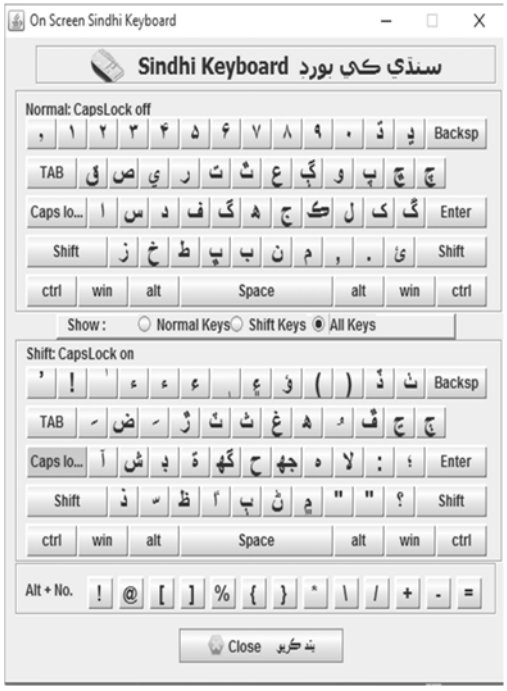

Fig. 1: Custom built On-Screen Sindhi Keyboard

2.3 Hashtable

The system uses a Hash table structure as its primary means of storing and retrieving words repository. Hash table is a very simple and basic data structure that allows only two columns. First column is called a Key that usually stores any value based on string data type and second column is known as element that stores any java object. The Figure 2 shows the basic structure of hashtable and its implementation logic used in this project. The single Sindhi word is used as a key to store and search its relevant similar words from the element column. The element column holds the java bean class object that contains the multiple similar words or synonyms for that particular key (which is a Sindhi Word). This hashtable object is then saved onto a file using a java object output stream, and the file is written and saved using File output stream classes.

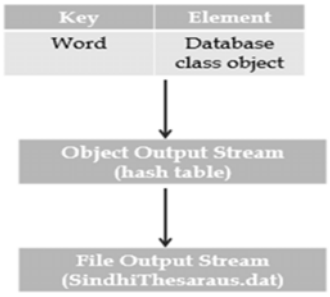

Fig. 2: Hashtable structure for thesaurus

This process is then reversed, for the retrieval and loading of the dictionary data onto the user interface. This retrieval process is illustrated in **(Fig. 3).**

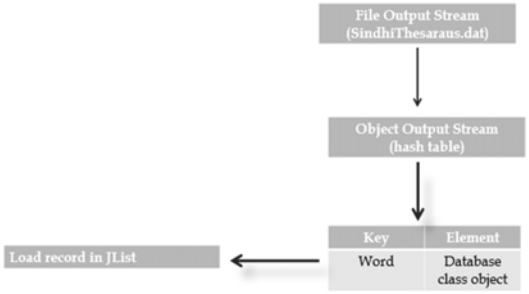

Fig. 3: Hashtable retrieval process

2.4 Sindhi Thesaurus bean class

Based on hash table principle, the database for Sindhi thesaurus is designed as such that all the data is inside a java bean object. Then that object is saved as an element inside the hashtable. The class structure of this bean object is shown in **(Fig. 4).** The bean class is a simple in structure with two variables of string data type. There are four methods declared, to store and retrieve the values stored in this object.

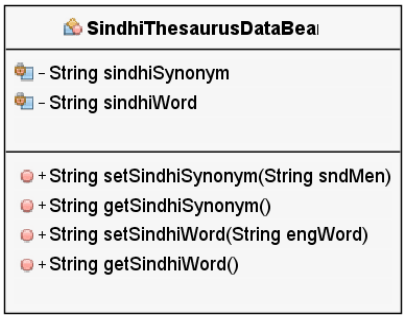

Fig. 4: Sindhi Thesaurus Database

The basic process of data retrieval used in this project is illustrated in **(Fig. 5)**. The retrieval processisinitiated by user action, which is searching for any particular word. The word is used as a key. The 'hashtable. get (key)' method is used to fetch the element form the hashtable, which returns the bean object for the corresponding word. Then the system reads the object using the get Sindhi Synonym() method and the results are loaded into the relevant textbox.

Fig. 5: Process of retrieving word and its synonym from bean object

2.5Activity diagram

This activity diagram of the Sindhi thesaurus is shown in Figure 6. The basic process flow illustrates the step by step activities performed by user. When that user clicks the java based executable '.jar' file; the welcome screen appears and is displayed to the user. Meanwhile, at the backend, the system loads and retrieves the dictionary class and hashtable object into the memory. After this the GUI class initializes with swing package, the system opens the dictionary file, initializes the object input stream class, and load the 'sindhithesaurusdictionary.dat' file. The system then loads the Sindhi thesaurus words using java.util.hash table class, sorts the words and finally loads the Sindhi words in Java list interface object.

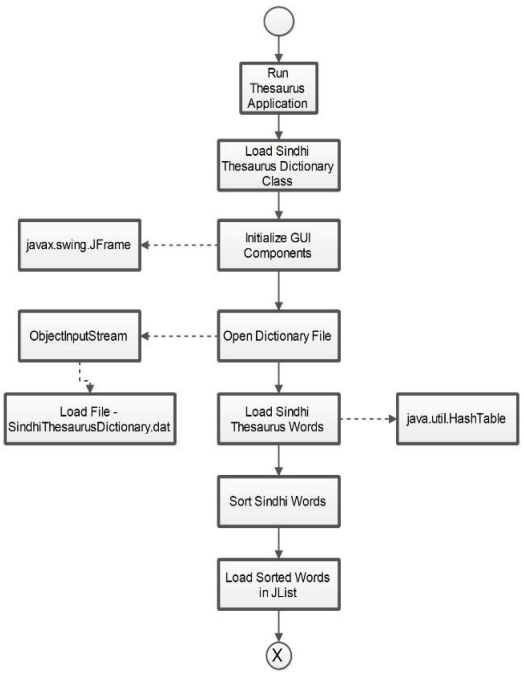

Fig. 6: Activity Diagram for Sindhi Thesaurus

3. RESULTS

The final application was developed and designed using Java and netbeans IDE. Java's swing package was used for designing the graphical user interface. Since Sindhi language is written from right to left direction, thus the GUI has to be designed to facilitate the Sindhi users' approach of reading from Right to Left. Therefore, the interface and its elements are designed with Sindhi text, and organized to provide best user experience.

The project has two main user sections. First is the Administrator tool and the second is General User App. Each is discussed in subsequent sections.

3.2 User Interface

The second section of the project is the main User Interface for the general Sindhi people. Using this software, any user can easily type and search the Sindhi word and its relevant synonyms would appear in the Sindhi meaning textbox. The user input functionality involves a fast and character by character searching mechanism. This allows the user to see the words search result as they type each Sindhi character. Figures 11 and 12 show various snapshots of the user interface for digital Sindhi thesaurus.

Fig. 11: (left) Welcome splash screen. (right) Main user interface of Sindhi Thesaurus

Fig.12: Various user interactions with GUI of Thesaurus

4. CONCLUSION

In this paper, the design and development process of digital Sindhi thesaurus has been discussed. The digital Sindhi thesaurus is a real need and necessity of current Sindhi community and for further research and development in Sindhi Computing. Initially, Sindhi words and their synonyms were gathered from various sources. Then a Sindhi user interface was designed to facilitate the Sindhi user. The word repository was saved using hashtable based data structure. The basic word and their synonyms were put inside a custom java bean object and then in a hashtable element with word as key. This allows the system to search and retrieve data quickly and efficiently. Overall, the project has been successfully completed. The thesaurus contains approximately 16000 (SixteenThousand) words having same meaning and some opposites. In future the system needs to expand and increase the word repository, as well as authenticate the accuracy of Sindhi words and their synonyms.

ACKNOWLEDGMENT

In this project I would like to courteously thank my students who helped me during this project in gathering and inserting the Sindhi words data into the system. Namely, Mazhar Hussain, Touseef Iqbal, Rashi Samo and Tagio Khan of batch 2K7 at IICT, University of Sindh, Jamshoro.

REFERENCES:

Abbasi, A. M., and S. Hussain, (2015). Phonetic Analysis of Lexical Stress in Sindhi. Sindh University Research Journal-SURJ (Science Series), 47(4).

Bhatti, Z., D. N. Hakro, and A. A. Jarwar, (2013b). "Sindhi Academic Informatic Portal". American Journal of Information Systems, 1(1), 21-25. DOI: 10.12691/ajis-1-1-3.

Bhatti, Z., I. A. Ismaili, D. N. Hakro, and W. Javaid, (2015) “Phonetic based Sindhi Spell Checker System Using a Hybrid Model” DSH: Digital Scholarship in the Humanities, Oxford Journals,
DOI 10.1093/llc/fqv005

Bhatti, Z., I. A. Ismaili, D. N. Hakro, and A. Waqas, (2014b). Unicode Based Bilingual Sindhi-English Pictorial Dictionary for Children. American Journal of Software Engineering, 2(1), 1-7.
DOI: 10.12691/ajse-2-1-1

Bhatti, Z., I. A. Ismaili, W. I. Khan, A. S. Nizamani, (2013a) "Development of Unicode based Sindhi Typing System", Journal of Emerging Trends in Computing and Information Sciences, Volume 4, Issue 3, 309-314, ISSN 2079-8407

Bhatti, Z., A. Waqas, I. A. Ismaili, D. N. Hakro, and W. J. Soomro, (2014a). Phonetic based SoundEx and ShapeEx algorithm for Sindhi Spell Checker System. Adv. Environ. Biol., 8(4), 1147-1155, AENSI Publisher, 2014 ISSN:1995-0756 EISSN: 1998-1066.

Hakro, D. N., I. A. Ismaili, A. Z. Talib, Z. Bhatti, and G. N. Mojai, (2014) Issues and Challenges in Sindhi OCR. Sindh University Research Journal (Science Series). Vol. 46 (2). 143-152.

Ismaili, I. A., Bhatti, Z. and A. A. Shah, (2014). Design and Development of the Graphical User Interface for Sindhi Language. Mehran University Research Journal of Engineering and Technology, 30(4). arXiv preprint arXiv:1401.1486.

Ismaili, I. A., Z. Bhatti, and A. A. Shah, (2014). Towards a generic framework for the development of Unicode based digital Sindhi dictionaries. Mehran University Research Journal of Engineering and Technology Volume 31, No. 1, January 2012. arXiv preprint arXiv:1401.2641.

Mughal, M. U. (2016). Sindhi Spelling Error Detection And Correction-A Hybrid Approach (Doctoral dissertation).

Shah, Z. A., and G. M. Mashori, (2011). Oxford English-Sindhi Dictionary: A Critical Study in Lexicography. ELF Annual Research Journal, 13, 37-46.

Soomro, H. K., A.A. Shah, and A. A. Shaikh, (2004) “Development of Computerized Sindhi to English and English to Sindh Dictionary”, Mehran University Research Journal of Engineering and Technology [ISSN 0254-7821], Volume 23, No. 4, 289-296, Jamshoro, Pakistan, October,.

Wikipedia, (2014) “Thesaurus” Online. Retrieved on 04/05/2014.
URL:http://en.wikipedia.org/wiki/Thesaurus