Design of Document Viewer with Integrated Web-Assistant for New File Format

Anil Singh[1] Gaurav Solanki[2] Rajat Patodia[3] Vicky Prasad[4] Dr. K. Rajeswari[5]
[1,2,3,4,5]Department of Computer Engineering
[1,2,3,4,5]Pune University PCCOE, Pune,2017

Abstract— This paper proposes the design of a document viewer which supports a new file format with file extension (.andx). The .andx file format supports features like multimedia support, integrating YouTube videos and animation support. The document viewer has built-in web-assistant that enables user to search for more content from web and provide them with summarized results using text summarization.

Key words: Document Viewer, Web Assistant, File format, andx

I. INTRODUCTION

Digital Presentation is an effective way of expressing thoughts by visualization techniques. Visualization helps a person to remember better and be more attentive towards the context. In the current system, due to absence of interactive contents and responsive mechanisms, presentations are losing its effectiveness. So, this paper discusses about designing a document viewer which will provide an interactive web assistant and gives live responses to user queries. It also allows embedding videos, animations and gifs within the presentation file.

The document viewer supports a newly designed file format which overcomes the limitations of the current system with the addition of real time responsive web support, integrated YouTube support and built-in screen reader.

II. BASIC DEFINITION

This section introduces several concepts and notations that are used in subsequent section to describe how the system works.

A. Document Viewer

A document viewer is application software which presents the data stored in computer file in user-friendly form. The contents of file are displayed on the screen. Alternately they can be read out loud using speech synthesis [3].

Document viewer cannot edit these files, but it can export data in a different file format or print them. A document viewer is limited-functionality software in the sense that it does not have the capability to create a file, or modify the content of an existing one. File viewers must have knowledge about the file format to be viewed in order to handle different byte orders, code pages or newline styles.

B. Web Assistant

A web assistant is a software agent that performs tasks for an individual. They work via text or voice. Web assistants use natural language processing (NLP) to match user text or voice input to executable commands. They learn using artificial intelligence techniques including machine learning. To activate a web assistant using the voice, a specific keyword must be spoken.

Web assistant can provide wide variety of services which includes providing information about facts, definitions, meanings and other useful information from web.

C. File Format

A file format is a standard way in which information is encoded for storage in a computer file. It specifies how bits are used to encode information in a computer's. File formats may be either proprietary or free and may be either unpublished or open. File formats often have a published specification describing the encoding method and enabling testing of program intended functionality.

D. Text Summarization

Before going to the Text summarization, first we, have to know that what a summary is. A summary is a text that is produced from one or more texts, that conveys important information in the original text, and it is of a shorter form. The goal of automatic text summarization is presenting the source text into a shorter version with semantics. The most important advantage of using a summary is ,it reduces the reading time. Text Summarization methods can be classified into extractive and abstractive summarization. An extractive summarization method consists of selecting important sentences, paragraphs etc. from the original document and concatenating them into shorter form. An Abstractive summarization is an understanding of the main concepts in a document and then express those concepts in clear natural language [1].

The text summarization technique used in the proposed system is TextRank based on paper TextRank: Bringing Order into Texts.

TextRank is a graph based ranking model for text processing [2].TextRank works as follows:

- Pre-process the text: remove stop words and stem the remaining words.
- Create a graph where vertices are sentences.
- Connect every sentence to every other sentence by an edge. The weight of the edge is how similar the two sentences are.
- Run the PageRank algorithm on the graph [4].
- Pick the vertices (sentences) with the highest PageRank score.

III. PROPOSED SYSTEM

A. Proposed Document Viewer

The document viewer application sends a .andx file to ANDX Decoder (refer section 3.4) which decodes the file into frame objects (here frame refers to single page). The decoder then, return set of all frames within the files. This set of frames is given to frame handler which in parallel renders each frame onto renderer surface using renderer

with help of surface manager. It also tells web assistant to download attachment files. User actions on surface like zooming, touch or click, scroll events are handled by surface manager which executes predefined actions and update device screen accordingly. Surface manager also decides current page to be displayed on device screen.

The application also controls speech inputs from user and sends it to speech to text engine. This text is passed to web assistant which queries server. After getting query result from server the web assistance gives the query result to application which tells surface manager to update the device display. The application also provides feature to read the search query result with the help of text to speech engine.

B. Proposed File Format

The new developed file format has a file extension .andx. The general structure of andx file format is discussed in the following topics.

1) ANDX File Format:

The structure of andx file format file contains a header, sequence of data stream of each slide. The figure 3.2.1 shows general file format of andx file.

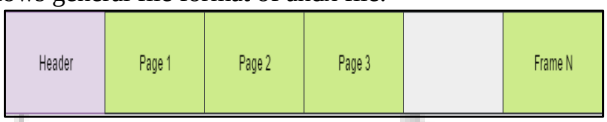

Fig. 3.2.1: ANDX File Format

2) ANDX file header:

The header of andx contains following fields:

- Magic Number: This is a 5 byte field used to identify a andx file.
- Version Number: This is a 3-byte field which tells the version number.
- Encoding: This is a 10-byte field which stores encoding used.
- Password: This is 8-byte optional field to set password to open the file.
- Frame Count: This is a variable length field which tells number of pages.

The header format is shown in fig below.

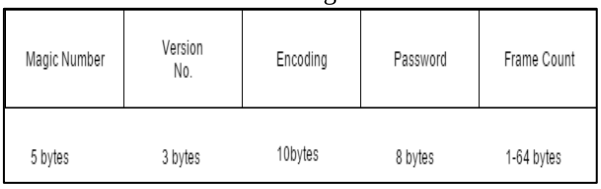

Fig. 3.2.2: ANDX Frame format

a) ANDX Page Format:

The frame format contains fields as follows:

- Start of Frame: indicates start of page/frame.
- Compression code: indicates compression algorithm used.
- Frame length: Actual data length in bytes.
- Start of frame data: Indicates start of frame data.
- Frame Data: The actual data is stored here.
- End of frame data: This indicate end of actual data.
- End of frame: This indicates end of frame/page.

The Fig. 3.1.2.1 shows ANDX Page Format.

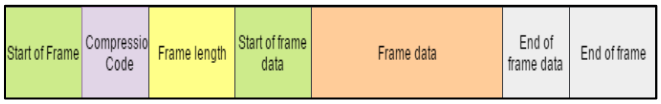

Fig. 3.2.2.1: ANDX Frame header

3) Specifications for ANDX file:

- The number of pages in a single andx file is 18e18.
- The maximum file size can be up to 18 EB (1EB=1000 Petabytes)
- The header can be from 27-90 bytes.
- The maximum video size can be up to 18,000PB

C. ANDX Encoder Architecture

The ANDX Encoder encodes raw input data into andx compatible data streams. Figure 3.3 shows architecture of andx encoder.

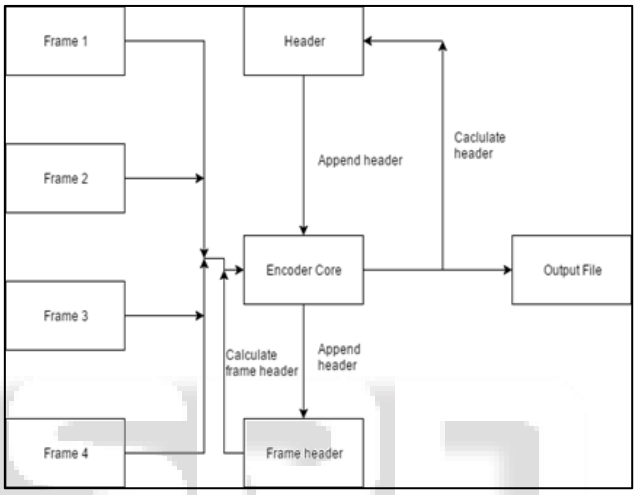

Fig. 3.3: ANDX Encoder Architecture

D. ANDX Decoder Architecture

The ANDX Decoder decodes andx data streams into andx compatible reader application. Figure 3.4 shows architecture of andx decoder.

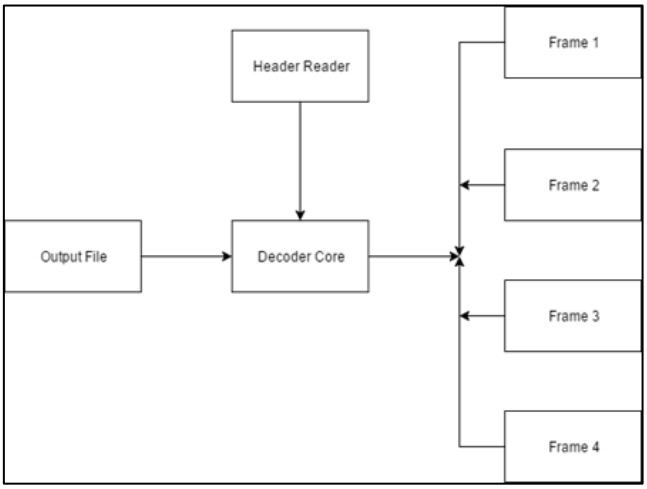

Fig. 3.4: ANDX Decoder Architecture

E. Web Assistant

The web assistant will help the user to search more information, images, videos, etc. from the web for given keyword and download attached external files. The user's search query is sent to server. The server will forward this query to the query optimizer which autocorrect search

keyword. It also optimise search query (if applicable). The query optimiser then return optimised query to server which queries the database system if query results are found in database they are returned to web assistant. If no results were found the query gets queried to Google using the Google Search API alternatively, a web-crawler can be implemented and list of links relevant to given query are retrieved from fetched HTML page using JSOUP. Each of the retrieved link is visited and the content from each link are extracted. A summary of result is made using text summarization (refer section 3.6). The results are then send in batches of 10 result at a time to the web-assistant and also saved in database. The query results gets displayed on user's device screen providing.

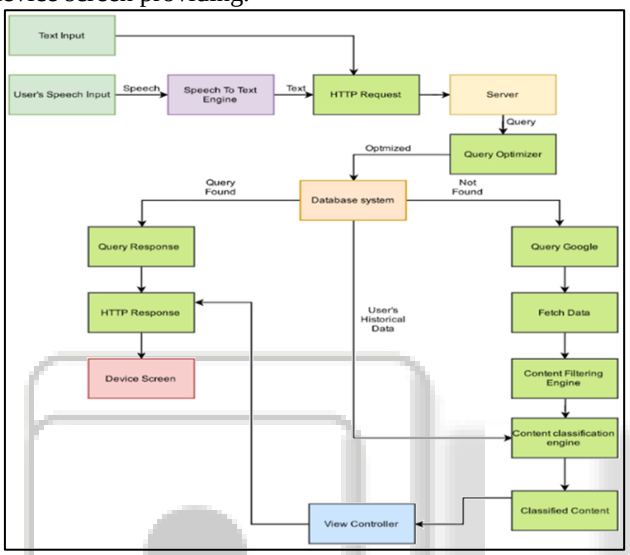

Fig. 3.5: Web Assistant architecture

IV. CONCLUSION

In this paper, we propose a new document viewer and a new file format (.andx). The document viewer has integrated web-assistant which helps user to read more content from the internet. For this purpose the system relies on Google Search API to return links of websites. Each link is then visited and meaningful contents are extracted. The extracted contents are summarized using TextRank algorithm. The summarized results are then returned to user.

REFERENCES

[1] Babar, Samrat & Tech-Cse, M &, Rit. (2013). Text Summarization: An Overview.
[2] TextRank: Bringing Order into Texts In Conference on Empirical Methods in Natural Language Processing (2004) by Rada Mihalcea, Paul Tarau.
[3] Iwai, I., Doi, M., Yamaguchi, K., Fukui, M., and Y. Takebayashi. A document Layout System Using Automatic Document Architecture Extraction. In the Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI'89), 1989.
[4] S. Brin and L. Page. 1998. The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems, 30(1–7).
[5] E. Frank, G. W. Paynter, I. H. Witten, C. Gutwin, and C. G. Nevill-Manning. 1999. Domain-specific keyphrase extraction. In Proceedings of the 16th International Joint Conference on Artificial Intelligence.
[6] Toward the Next Generation of Recommender Systems: A Survey of the State-of-the Art and Possible Extensions. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 17, NO. 6, JUNE 2000
[7] Rajni Jindal,Ruchika Malhotra. "Techniques for text classification".2nd International Conference on Systems and Computer Science (ICSCS), 2013.
[8] Buyukkokten, O., Garcia-Molina, H.,Paepcke, A. Seeing the Whole in Parts: Text Summarization for Web Browsing on Handheld Devices. In the Proceedings of the Tenths International World Wide Web Conference (WWW 10), 2001.
[9] R.J Keeble, R.D Macredie; Assistant agents for the world wide web intelligent interface design challenges, Interacting with Computers, Volume 12, Issue 4, 1 February 2000, Pages 357–38.
[10] R. Mihalcea. 2004. Graph-based ranking algorithms for sentence extraction, applied to text summarization. In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004) (companion volume), Barcelona, Spain. P. Turney. 1999.
[11] Learning to extract keyphrases from text. Technical report, National Research Council, Institute for Information Technology.
[12] Wilcox-O'Hearn, L. Amber. 2008. Applying trigram models to real-word spelling correction. MSc thesis, Department of Computer Science, University of Toronto [forthcoming] ,2001
[13] Church, Kenneth W. and William A. Gale. 1991. Probability scoring for spelling correction. Statistics and Computing, 1, 93–103, 1997
[14] Item Based Collaborative Filtering Recommendation Algorithms.
[15] Content-based Recommender Systems: State of the Art and Trends Pasquale Lops, Marco de Gemmis and Giovanni

