23

A Tool for Document Visualization

S. Chithirapoovizhi, Ranjani Parthasarathi, T.V. Geetha

Resource Centre for Indian Language Technology Solutions – Tamil
School of Computer Science and Engineering
Anna University, Chennai, India.
E Mail : {scpoovizhi@yahoo.com, rp@annauniv.edu, tvgee@cs.anauniv.edu}

Abstract

This paper presents the details of a tool developed to visualize any tagged document. It aims at presenting abstracted information extracted from tagged text files visually. Instead of forcing a user to read a whole document just to grasp certain information, document visualization offers very important clues about the document.

1. Introduction

Visualization is a cognitive process performed by humans in forming a mental image of a domain space. In computer and information science it is, more specifically, the visual representation of a domain space using graphics, images, animated sequences and sound augmentation to present the data, structure and dynamic behavior of large, complex data sets that represent systems, events, processes, objects and concepts.

Visualization relies on the fact that users can distinguish positions, colors, textures, and relationships. Relationships can be shown in displays by proximity, by containment, by connected lines, by color-coding, etc. Fields containing hundreds or thousands of points can be scanned rapidly and efficiently for clusters, outliers, trends, and gaps. Attention can be drawn to salient items using a variety of techniques including highlighting, blinking, motion, and size. Direct manipulation of visualizations can be accom plished with a variety of methods, such as pointing to select, dragging, and zooming. Feedback is immediate and intuitive in such environments. "The eye, the hand and the mind seem to work smoothly and rapidly as users perform actions on visual displays" [1].

This paper presents a visualization tool that can be used to visualize text or documents. It is organized as follows**.** The rest of the section presents in detail the idea behind different types of visualization. Section 2 discusses about Need for Document Visualization, Section 3 gives details about the Proposed Tool, final section gives the conclusion

1.1 What is visualization?

Visualization techniques transform associated data chunks into a cohesive visual representation that can be understood easily by the user. The given data could be of anything like scientific data, gene informa tion, protein structure, library, documents, text, etc. Visualization enables the

human mind to process in parallel vast quantities of information. It enables present-ation/discover/reuse of information.

Classification of visualization techniques is often based on the dimension of the domain of the quantity that is visualized, i.e. the number of independent variables of the domain on which the quantity acts, and on the type of the quantity, i.e. scalar, vector, or tensor. They are also classified based on application domain.

Visualization techniques can also be divided into surface rendering techniques, and (direct) volume rendering techniques. Surface rendering is an indirect geometry based technique which is used to visualize structures in 3D scalar or vector fields by converting these structures into surface representations first and then using conventional computer graphics techniques to render these surfaces. Direct volume rendering is a technique for the visualization of 3D scalar data sets without a conversion to surface representations [3]

1.2 Kinds of Visualization

Scientific Visualization [3]

Scientific visualization is the representation of scientific data graphically as a means of gaining understanding and insight into the data. It is sometimes referred to as visual data analysis. This allows the researcher to gain insight into the system that is studied in ways previously impossible. This involves research in computer graphics, image processing, high performance computing and other areas.

As a science, scientific visualization is the study concerned with the interactive display and analysis of data. Often one would like the ability to do real-time visualization of data from any source. Thus our purview is information, scientific, or engineering visualization and closely related problems such as computational steering or multivariate analysis, applicable to datasets of any size, while still retaining high interactivity. As an emerging science, its strategy is to develop fundamental ideas leading to general tools for real applications.

Scientific visualization is applied on domains such as Airplane cockpit, Weather conditions, Temperature variations, Protein structure and Biological phenomena.

Business Visualization [4]

Business data visualization allows users to quickly & accurately analyze and present most complex business data in tangible form so that the business managers can recognize the patterns and trends hidden in large data. This visualization helps many people to make accurate and informed decisions faster and cost effectively.

Data visualization comprises of visual components like bar, pie and line charts, parabox, scatter plot, histogram and data constellation These components are linked into one view of business data to answer targeted business questions. Multiple perspectives are linked together under a structured workflow to create analytical dashboards. Perspectives and dashboards allow business users to instantaneously view their data in multiple dimensions.

Document Visualization

The idea behind document visualization is as follows: Any online document cannot be shown entirely to the reader due to the constraints of the size of the visual display unit. Instead many visual clues about the document may be presented to the reader in order to abstract him/her with information that is easily understood. Such visual clues could be statistical information or visual perceptions.

A reader may not read a huge document completely. Instead (s)he may just need certain information that is relevant to him/her. Filtering such information plays a vital role. The next step is to present to the user, the filtered information in a way as to keep the context in view. Thus, visualization helps in extracting information from a huge collection of data and presenting the result of such data extraction in a user-interactive way. The success of the visualization lies in the simplicity of the presented view. It also presents different views and allows the user to select the view of his/her choice for various purposes. Thus visualization has two major benefits,

- It provides visual abstraction, for speedy pattern detection in a collection of data.
- It communicates large amount of data effectively.

2. Need for Document Visualization Tool

2.1 Existing tools for Visualization

The following are some of the tools available for the visualization.

Pad++ [1] - A Zoomable interface which presents the given document to the user in the form of small thumbnails in elliptical manner. The user has to click on the thumbnail he wishes to view.

One difficulty with the elliptical view is that it is only practical for a relatively small numbers of pages (less than 30). Also, for the current implementation, achieving pleasing layouts for documents of different lengths requires manual change of parameters.

Data Explorer (DX) [5] -- It is a general-purpose software package for data visualization and analysis for scientific data. It employs a data-flow driven client-server execution model and provides a graphical program editor that allows the user to create a visualization using a point and click interface.

Visual Insights Explorer [6] - A tool that is available commercially for visualization of business related data.

The features that are widely used to view documents in standard text viewers include Scrollbars and Range slider.

Scrollbars keep in view the current page. But they allow the user to focus only on certain part of the page. They do not even provide the user with a facility to know the number of pages in the document

Range Slider is useful for selecting range for integer attributes and other attributes with an order relation. The range slider provides a natural representation of the query it is representing i.e. range query. In creasing and decreasing the range of an attribute allows for powerful exploration of trends and anomalies.

2.2 Limitation in the existing tools

Most of the tools are available only for data visualization. The availability of text or document visualization tools are few and primitive too. They do not offer much information on content to the user. They offer very less possibilities to explore the information. Thus there is a need to go for document visualization.

3. The Proposed Tool

3.1 Desirable features of the Proposed tool

- The entire document should be presented, with graphical representation, in one display.

- The display should be presented from the user's perspective and the user should be able to change the display interactively.
- Must be able to perform independent of language and document formats.
- Abstracting the given data and presenting the document to the user in the form of different views.
- Should provide the focus of necessary part of the document and also keep in view the context in the document.
- A provision for specifying mutual relationship between tags in the document should exist.
- The most important document attributes, as defined by the user, should be retained in the display.
- Options to decide tag delimiters should exist.

3.2 Overview

This tool takes tagged document as its input. Tags can be in markup languages like XML or of any proprietary format. Tags can be in nested format. The tool will essentially process the tags and generate intermediate file along with tag properties. It has been designed keeping in mind the various views so that it provides various levels of information about the document. Four kinds of views presented are scatter plot view, zoomer view, statistical information view and synchronous depth view. The scatter plot view shows relative tag positions in a two-dimensional plane and gives information about the selected tag. Document zoomer view shows thumb views of pages. The statistical information view shows statistics about the tags in the input document and synchronous depth view gives an insight of how the tags are spread over the document.

3.3 Design Details

This tool comprises of three major modules viz., Input module, Tag builder module and Data mapper comes output module.

3.3.1 Input Module

This module determines the format of the input file and format of tags in the input file. It analyses the input file and extracts details of the given input document.

3.3.2 Tag Builder

This module searches for suitable tags from the input file. The tag delimiters are also taken as input, in order to take care of proprietary tag types. These tags are stored as a n*n matrix. A color is associated with each tag. The color information is generated by using random integer generator function. This color information is used in all the views. The name of the tags and their respective file positions, are stored in the internal data structure. An intermediate file is generated wherein the properties of tags like color, position of occurrence, name and type of tag are included.

3.3.3 Data Mapper and Output Module

The input to the module is the intermediate file generated by the tag builder module. The data object builder retrieves tags and its related information and builds data objects. These data objects have information related to the different views namely Scatter plot, Zoom, Statistical information and Synchronous depth view.

The views like zoom and scatter plot, displays the data objects on a two dimensional plane. In case of bar chart and synchronous depth views, the transform group takes the data objects and applies the necessary transformations and converts them into viewable objects on the logical canvas. The transformations implemented are color rendering, translation and rotation on the data objects. The logical canvas takes care of three-dimensional viewing of the objects. The details of each view are given below.

Document zoomer view

This view is based on Graphical Fisheye and Lens Views. The document zoom view takes given document as a input file and prepares a new file. This new file has proper alignment of 80 characters per line and 24 lines per page. Each page is identified by a page marker and is shown in small thumbnail. On clicking the mouse at a thumbnail, the corresponding page is shown in full view in the right side pane. It also has a search facility. It highlights all the occurrences of a given search string in that document, both in thumbnail view and in the right pane. Fig 3.1 shows the sample output for Document Zoomer View.

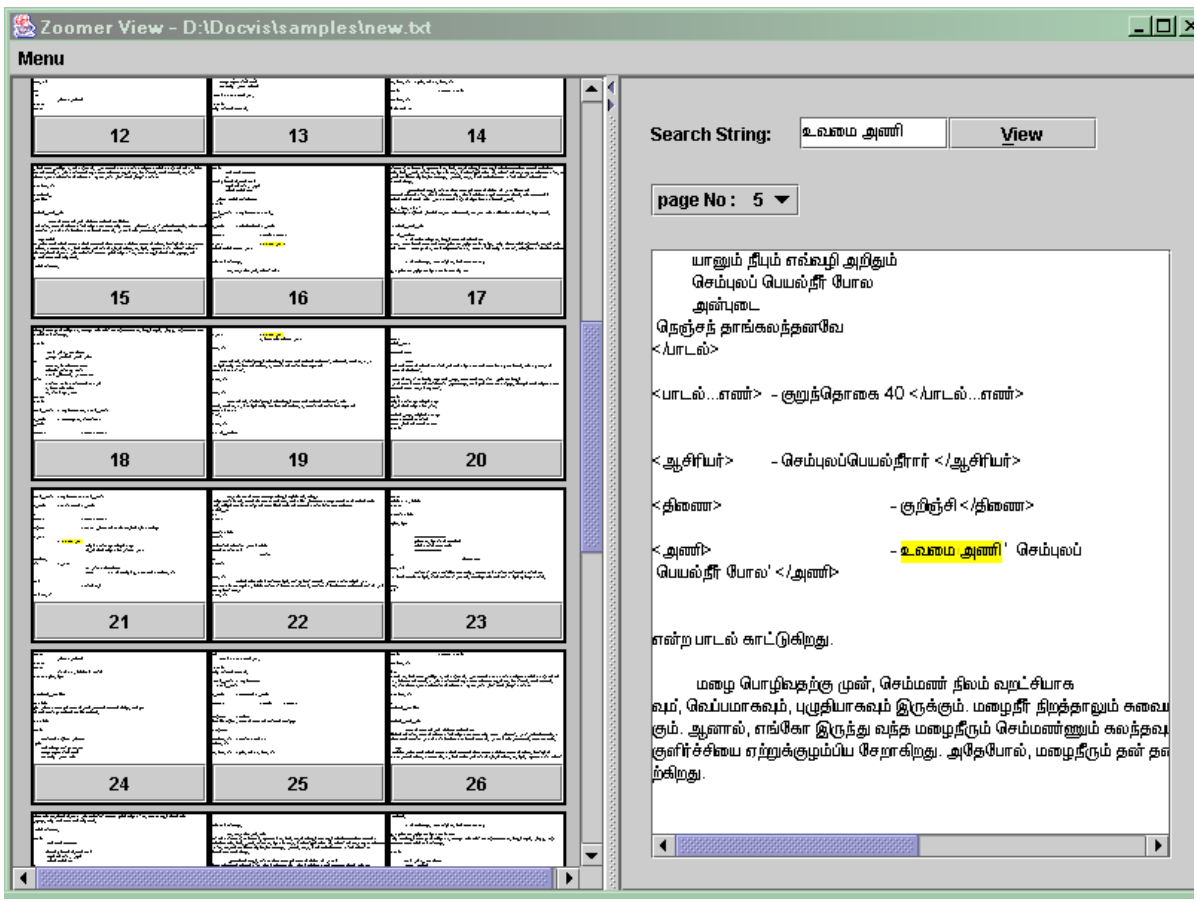

Fig 3.1 Document Zoomer View

Scatter Plot view

Here the document is mapped on a graph, x-axis marks the tag position & y- axis marks the pages. This view gives the relevant position of the tag in a particular page. First it will parse the intermediate file and get information about the tags, its positions, color etc. Then it normalizes

the tag position according to the display. The position of the tags in that pixel lines are marked with circle of pre-defined size with the color that was specified in the tagbuilder. Here the nested and normal tags are distinguished by its size. The color of the tag is used to highlight the circular region. All the related tags are highlighted using the same color. A separate window shows all the tags along with its corresponding color, in the source document and allow the user to select multiple tags from them. Only those tags selected by the user are shown in the display.

Apart from displaying the tags, it gives the relevant information about a particular tag in a separate screen when the user clicks on it and, if the selected tag is of nested type, it also drops a line from the starting position to the end of the tag. User can also view preceding and succeeding tag. Option is available to navigate from this window to zoom view display. Fig 3.2 shows a sample view of Scatter plot of the select tag.

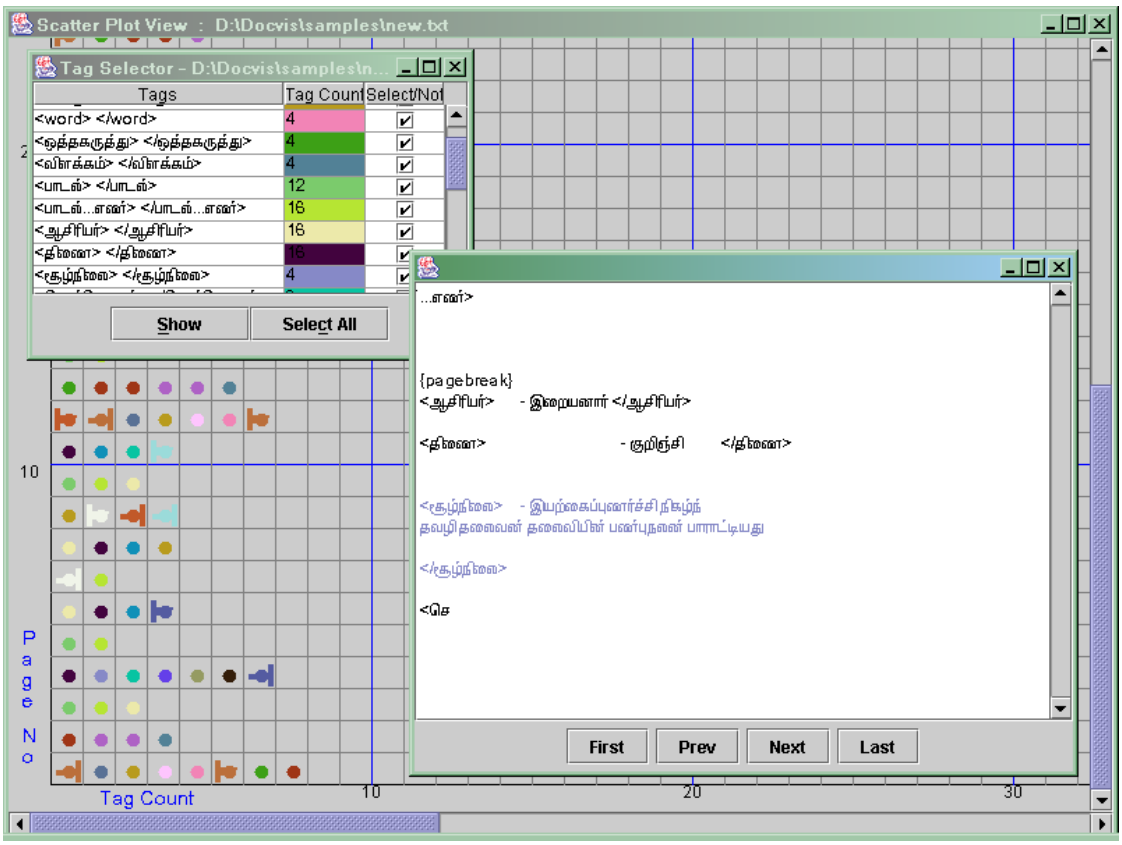

Fig 3.2 Scatter Plot View

Statistical Information view

This view presents the frequencies of the tags identified in the input file in a bar chart form. It parses the intermediate file and gets frequencies and color of related tags. It normalized this frequency information. In this view, on the left pane the frequency of the tags is shown by means of the height of cylinder, which varies as the frequency varies. These cylinders are drawn in 3D plane. The user has an option to translate these cylinders along x-y axis. The pane on the right side shows the color and tag strings. Fig 3.3 shows a sample screen of Statistical Information view.

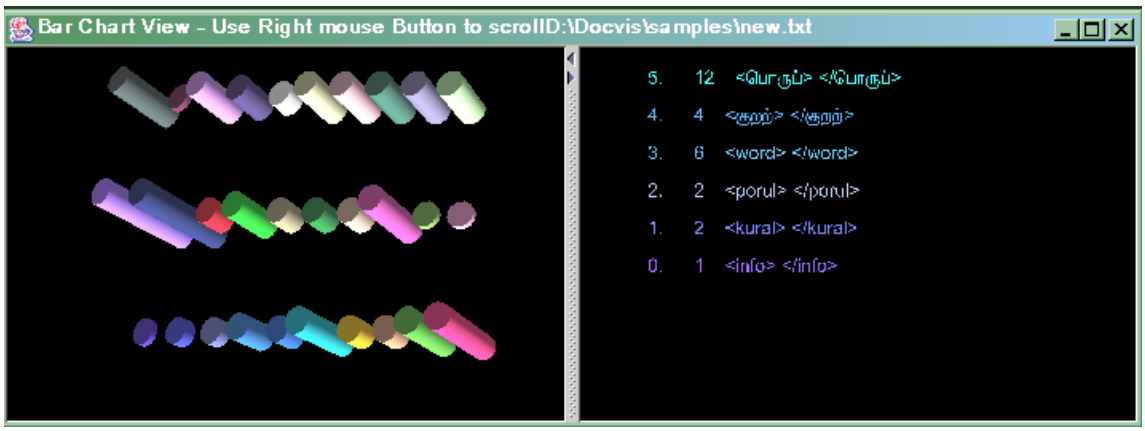

Fig 3.3 Statistical Information View

Synchronous Depth view

This view projects the input document onto the 3-dimensional plane. The whole docu-ment is considered to be an object in 3-dimensional plane. The tag and other information are extracted from the intermediate file. This view helps to get the picture of how the tags are spread over the document. This view is similar to the scatter plot view, but on 3-dimensional plane. Fig 3.4 shows the output screen of Synchronous Depth view.

Fig 3.4 Synchronous Depth view.

Testing

The tool has been tested for various documents of sizes around 600 KB and it works effectively on documents in Tamil, English and Sanskrit.

4. Conclusion

This tool presents a new technique for document presentation. The developed tool can effectively visualize any generic tagged document. It provides simple interface, which is easy to use and understand. It is a language independent tool.

A number of features need to be added to make the tool more effective. The Document Zoomer view can be improved by enabling it to show selective tags, which distinguish certain text from others. For example, chapter tag could be given preference than page breaks. Currently it handles only bilingual documents (i.e. Single font documents). It should be enhanced to handle multilingual documents.

References

[1] Lars Erik Holmquist , “Zoom Browser - Presenting a Focus+Context View of World Wide Web Documents", 1999.

[2] Document Visualization by Emile Morse in the year Submitted: December 15, 1997 Revised: January 15, 1998 - http://www.itl.nist.gov/iaui//vvrg/emorse/papers/soa/Document Visualization.htm#_Toc409515488

[3] Scientific Visualization Tutorial - http://www.cc.gatech.edu/scivis/tutorial/linked/ classification.html

[4] http://www.advizorsolutions.com/products/software_products.asp

[5] http://www.opendx.org/

[6] www.visualinsights.com/products/eBizInsights_XL.asp - 26k - 26 Jun 2003