Visualized Representation of Data Using Regex

Shilpa[1], Tarandeep Kaur[2]

School of Computer Science and Engineering, Lovely Professional University,
Phagwara Punjab

Abstract

With the advent of various technologies, the need of data is increasing day by day. Data is a collection of complex datasets and massive data that becomes difficult to make the decisions from the complex datasets. To make the complex data easy to understand various types of analysis techniques are applied which are defined as Visual Analysis (VA) and Big Data Analytics (BDA). The challenging issues are the processing and analyzing of the large amount of data. In any business the data generates in large size which is in terabytes. As a result it becomes difficult to make the decisions. For the decisions making data is transmitted into graphical representation termed as Visual Analytic. Visual analytics is related to the visualization of the data and data sets that make the results more interactive.

Keywords: Big Data; Regular Expression; Variety; Visual Analysis; visualization

Introduction

"Data visualization is an art and science and there are many methods which are used to apply on data and to understand what about the information."

In today generation, data is arising in very complex and large in size. Due to the big size and complexity of the data it becomes very difficult to manage in current automotive world. For better progress of structure and unstructured data there is a need of interactive visual analysis techniques for the visualization.

Visualization helps in the understanding of the data in graphical form. The visualization mainly concerned with three important terms: data, information and knowledge. In various existing tools there are many challenges regarding VA from which one is human perception that defines that user is not able to understand the data properly and second is limited screen space that define that objects are not properly visible. These challenges make the data difficult to understand and make aggressive to make the decisions.

Structured and unstructured data collected for the visual analysis that converts the data into visualization form that helps in gaining knowledge and making choices.

Visualization is one of the most powerful techniques for exploration of data.

- Visualization allows the graphical representation of the data that defines the complexity of data.
- Human can analyze large amount of data in visualized form easily.
- Human can easily detect the patterns which are hidden and visualization defines the relationship between datasets.

For the visualization of the datasets various VA tools are used to make it interactive.

Visualization tools

Visualization tools help in the graphical demonstration of the data or concepts to increase the understanding of data. Visualization tools have various forms to exhibit the data in graphical representations for making it interactive.

There exist various visualization tools that are used to illustrate the data in graphical depiction generated from large datasets. Some tools are freely available to represent the data in graphical form.

Data visualization tools are becoming more popular with the affixed requirement of the Big Data. Data visualization is only an art by which it is possible to reach at desired conclusion. For better understanding of the data visualization tools help the companies to create a cohesive understanding of the data which is impossible from the raw data.

ManyEyes

ManyEyes, the visualization tool which is developed by IBM Cognos software group. ManyEyes is an online community where users upload their datasets and visualized. But this community is not secure because the data sets which are uploaded online become publically available for the other users. ManyEyes tool is used to upload the datasets and create the visualization such as Wordtree, Treemap and Geographic Map Tools. ManyEyes is Web-based data visualization that is used to combine the graphical analysis with community. ManyEyes provide the suggestions to use the different types of visual data representation which are defined in the tool. It provides more than a dozen output forms. With the help of ManyEyes the complex data which is render for the analysis becomes high level interactive visualization. IBM's ManyEyes is used when anyone wants to make the decisions at any time and anyone can make the interactive picture for the proper understanding by visualizing the data.

Objectives

To gain the useful information from the data it is necessary to represent it in meaningful manner for the decision making and to discover the hidden patterns. There is more preferment in big data representation through the various computational methods and various algorithms which are complex to understand for all the analysts. Many Eyes is a gamble on the power of human visual intelligence to find patterns. Our goal is to democratize visualization and to enable a new social kind of data analysis.

Main objectives of visualization are:

- Understanding of data properly which is recorded.
- Accurate graphical representation of data
- To run the search query to locate the location of the text
- For discovering the hidden patterns
- Visibility of all the data items

.The challenges include the unstructured data, real time analytics, fault tolerance, processing and storage of the data and many more.

Problems in Visualization Tool

ManyEyes tool is web-based data visualization tool used for the graphical analysis of the data. ManyEyes tool provide the dozens of output format for the visualization of the data.The main drawback in this tool is that the dataset or the visualization form of that dataset, which is uploaded by the user for the visual analysis make available publically that can be easily downloaded, shared and used or commented upon by other users. There can be great for the certain types of the users including government agencies, nonprofits, and the organizations for sharing on server budget but not for some people.

Methodology

The proposed system is implemented with the help of ASP.NET where the proper GUI interface is provided where the user creates the profile and upload the datasets as per their needs and the main operation which is performed is searching the text location from the pdf file. The data set uploaded by the user is not available publically. Only the owner has authentication to see that datasets and the visualization of that data sets.

Proposed system

In this system the proper interface is provided for the creation of the profile. The interface recommended for the registrations without the registration user can not makes the VA. This proposed system having following Phase:

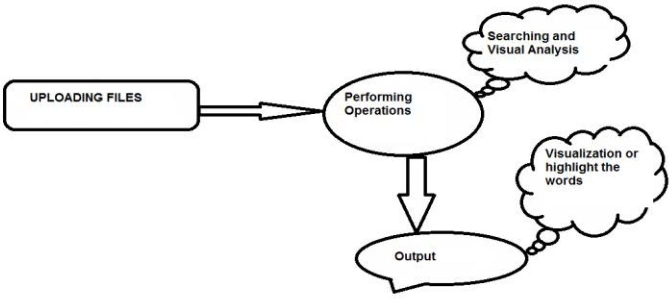

UPLOADING FILES
Performing Operations
Searching and Visual Analysis
Visualization or highlight the words
Output

Figure 1: Processing operation

For the implementation process various libraries concepts are used which are as follow:

itextsharp

itext is library which is freely available and open source library used for creating and manipulation the PDF files. itext is used as itextsharp in .NET Framework. itextsharp is written in c# or java. With the itextsharp concept, the PDF document from XML file as well as from databases can be easily developed dynamically. PDF files can be easily saved to a browser, adding bookmarks, page numbers, watermarks etc. Anyone can easily split the PDF file pages and can manipulate them.

Anyone can search and extract the pages from PDF File. itextsharp provides following feature which are:

- Searching of specified string with page number.
- Use of Regular expressions for the text searching.
- Pages extraction based on even, odd, range specification.

For performing the operation on PDF files there is a need to add the Library file which is itexhsharp.dll from which the following namespace are used:

```
using iTextSharp;
using iTextSharp.text;
using iTextSharp.text.pdf;
```

With the help of these classes, it is easy to take the text alignment and its position that makes easy to read the PDF file contents.

From these classes the concept of chunk, annotation PDFWriter, PDFReader is performed. By this one can extract the extension of the file which is being used by the function GetExtension (Filename).

The chunk is used to represent the smallest text object that contains StringBuffer used to store the selected text or characters in the same format in which it is taken in the document.

using iTextSharp.text.pdf.parser;

This namespace is used to extract the text. It is used to keep the track of the position of the string. If there is any change then it will represent the PDF contents at that position. The function used for this is as follow:

ITextExtractionStrategy strategy= new SimpleTextExtractionStrategy ();

It is used to extract the contents from multiple lines of PDF document and used as string builder

TextExtractor extract = newTextExtractor ();

This function is used to extracting the contents based on search items

string ext = System.IO.Path.GetExtension (FileUpload1.FileName)

This function is used to get the extension of the file which is uploaded for the analysis.

PdfTextExtractor.GetTextFromPage (pdfReader, page, strategy);

It is used to extract the text from the page of PDF file

Regular Expressions

Regular expression is an independent language which is supported by many languages which are Java, JavaScript, C# etc.

Regular expressions are patterns which are used to extracting the text against the defined input. For the regular expressions the concept of Regex class is used. The name space for the regular expressions having set of classes to utilize the power of regular expressions is:

using System.Text.RegularExpressions;

In this namespace many methods are defined from which are used for matching the expression from the given input string. From all those methods following are most used for matching the expression based on given string and as a result it returns the bool values whether the string finds the match or not and number of matches can be build using Matchcollection method. The Matchcollection method is used to make the list of all the matches occurring during Regex match from the substring.

Security

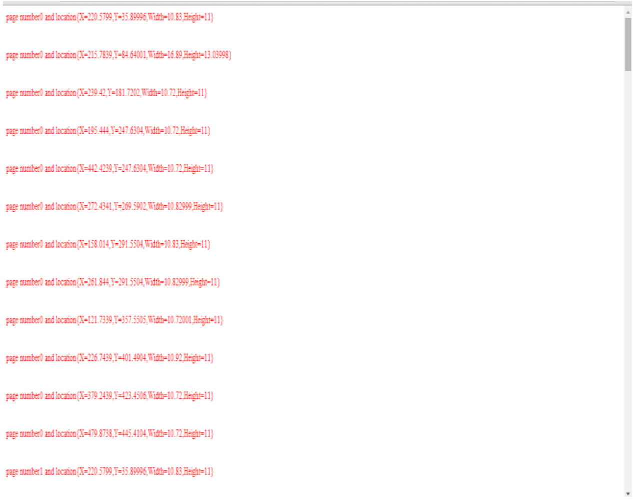
In ManyEyes tool there is issue regarding the security of the profiles to make the profiles secure web security concept is used for this the following namespace are used.

```
using System.Security;
using System.Security.Cryptography;
using System.Web.Security;
using System.Web.UI.HtmlControls;
```

RESULTS

This proposed tool is implemented in ASP.NET with C# and for the database management SQL SERVER is used where all the records are maintained properly. ASP.Net is server side Web application framework which is mostly used in the development of Web Applicatios, Window Applications and Web Sites etc. The implemented results are given below:

Page Number Result



```
page number0 and location{X=220.5799,Y=35.89996,Width=10.83,Height=11}

page number0 and location{X=215.7839,Y=84.64001,Width=16.89,Height=13.03998}

page number0 and location{X=239.42,Y=181.7202,Width=10.72,Height=11}

page number0 and location{X=195.444,Y=247.6304,Width=10.72,Height=11}

page number0 and location{X=442.4239,Y=247.6304,Width=10.72,Height=11}

page number0 and location{X=272.4341,Y=269.5902,Width=10.82999,Height=11}

page number0 and location{X=158.014,Y=291.5504,Width=10.83,Height=11}

page number0 and location{X=261.844,Y=291.5504,Width=10.82999,Height=11}

page number0 and location{X=121.7339,Y=357.5505,Width=10.72001,Height=11}

page number0 and location{X=226.7439,Y=401.4904,Width=10.92,Height=11}

page number0 and location{X=379.2439,Y=423.4506,Width=10.72,Height=11}

page number0 and location{X=479.8738,Y=445.4104,Width=10.72,Height=11}

page number1 and location{X=220.5799,Y=35.89996,Width=10.83,Height=11}
```

Figure 2: Page Numbers of Searched word

Conclusion and Future Scope

This research provides the GUI to the user for Visual Analysis of the datasets which are complex and massive in collection. The main objective of this research is to provide the security and privacy to the user data. The Visual analysis tool architecture is an integrated system with GUI interface. The challenging issues in Existing ManyEyes tool for searching data item with location is resolved. In the implementation of this interface the issues of publically data is taken care off. The Whole system is secure and data is hidden from public no one can access the others datasets and can't share that. This interface basically provides a tool for the VA on datasets that helps the users to understand the information and can search the data from the Big amount of data. The result contains the information about the location of the data existing in the document file.

In this proposed system the main concerned is with the privacy and security of the data and VA is applied on the data.

REFERENCES

[1] Abousalh-Neto, Nascif A., Kazgan, Sumeyye, "Big data exploration through visual analytics", Visual Analytics Science and Technology (VAST), IEEE Conference, ISBN: 978-1-4673-4752-5, (2012) Page(s): 285 – 286.

[2] Ben Shneiderman, "The Big Picture for Big Data: Visualization", international science & engineering visualization challenge, ISSN 0036-8075, Vol.no. 6171, (2013) pp. 600-610.

[3] Danial Keim, "Big-Data Visualization", Computer Graphics and Applications, IEEE ISSN: 0272-1716, Volume: 33, Issue: 4, (2013) Page(s): 20 – 21.

[4] Freiler, Matkovic, K.; Hauser, H., "Interactive Visual Analysis of Set-Typed Data", Visualization and Computer Graphics, IEEE Transactions on (Volume: 14, Issue: 6) , ISSN: 1077-2626, (2008) Page(s): 1340 – 1347.

[5] Garlasu, D.; Sandulescu, V.; Halcu, I.; Neculoiu, G., "A Big Data implementation based on Grid Computing", Roedunet International Conference (RoEduNet), 2013, ISSN: 2068-1038, ISBN: 978-1-4673-6114-9, Page(s): 1 – 4.

[6] Ghanbari, M., "Visualization Overview", System Theory, SSST '07. Thirty-Ninth Southeastern Symposium, ISSN: 0094-2898, E-ISBN: 1-4244-1126-2, (2007) Page(s): 115 – 119.

[7] Ghanbari, M., "Scalability of visualization's evaluation:" Southeastcon,. IEEE, E-ISBN: 978-1-4244-1884-8,(2008) Page(s):318 – 322.

[8] Hansen, C., "Big Data: A Scientific Visualization Perspective", SCI Institute Professor of Computer Science, University of Utah, ISSN: 2165-8765,(2013) Page(s): xi.

[9] Hassan, S., Sanger, J.; Pernul, G., "SoDA: Dynamic visual analytics of big social data", Big Data and Smart Computing (BIGCOMP), 2014, Page(s): 183 – 188.

[10] Jaegul Choo and Haesun Park "Customizing Computational Methods for Visual Analytics with Big Data", Computer Graphics and Applications, IEEE (Volume: 33, Issue: 4), ISSN: 0272-1716 , (2013) Page(s):22 – 28.

[11] Manashvi Birla, Aditya B. Patel, Ushma Nair, "Addressing Big Data Problem Using Hadoop and Map Reduce", Engineering (NUiCONE), Nirma University International Conference, ISBN: 978-1-4673-1720-7,2012 Page(s): 1 – 5.

[12] Memon, B.R., Wiil, U.K., "Visual Analysis of Heterogeneous Networks", Intelligence and Security Informatics Conference (EISIC), (2013) Page(s): 129 – 134.

[13] Mukherjee, A.; Datta, J.; Jorapur, R.; Singhvi, R.; Haloi, S.; Akram, W., (18-22 Dec.,2012) , "Shared disk big data analytics with Apache Hadoop" High Performance Computing (HiPC), 2012 , E-ISBN :978-1-4673-2370-3 , Page(s):1 – 6.

[14] Nasridinov, A., Young-Ho Park, "Visual Analytics for Big Data Using R", Cloud and Green Computing (CGC), Third International Conference, INSPEC Accession Number: 13991521, (2013) Page(s): 564 – 565.

[15] Sagiroglu, S., Sinanc, D., "Big data: A review", Collaboration Technologies and Systems (CTS), 2013 International Conference, ISBN: 978-1-4673-6403-4,Page(s):42–4 7.