NDVS: A System of News Data Visualize

Han Yuping, Zhu Ligu, Zhang Lei, Zhang Ruisong
Communication University of China
Beijing Key Laboratory of Big Data in Security & Protection Industry
Beijing, China
yuping0713@126.com

Abstract—**This article leads the concept of data visualization into the huge amounts of news data, design and implements a system of News Data Visual System (NDVS). It is a news gathering, analysis, structured storage and visual analysis system. NDVS collects news data in the schedule time from seeds website by a web spider that written by java every day. The structured documents that processed and analyzed by detached parameters from URL and classify news will be saved in distributed database of NOSQL – MongoDB. WEB will visually and interactively animate the documents in the means of D3.js. Design and realization of NDVS will be emphasized in the article, and we find that NDVS has a good capability when lots of clients concurrent access. We believe NDVS will become a useful tool between news and media works, because system follows the usability policy that helps people to find the potential tendency of news' development and guide the public opinion in the stage of visualization design.**

Keywords-Data visualization; News data; News gathering; Interactive; Visualization design

I. Introduction

With the rapid development and broad popularization of the new media platform in China, more and more people conveniently acquire information via the Internet. The advent of new-media era brings unprecedented increase of information. That brings troubles when user search the key information due to data in the state of exponential growth. ProPublica, is an online news website founded in 2008, has opened up a channel called Data to display news data visual works created by journalists. Now it has become the most active website in the users about data news in America. In developed countries, the business of news data analysis has tended to be mature by data visualization. And we can find the preliminary signs in Cultural and Creative Industry of China. Baidu also has news visualization case -- Baidu Index which is placed in the cloud platform. It visualized the data from several news websites and supplied analysis result to users. Not only the less of dimensions but also the singleness of style by contrast news data visualization of China with ProPublica. And domestic news data visualization cases are less than developed countries. NDVS fully integrate the technology of a large number of news data mining, repeat information filter, redundant information condense and key details display. That exactly solves the problems which users of news focus be confronted with. NDVS provides an intuitionistic and highlighting visual analysis interface to catch the key information of news for users.

II. Method and Implementation

NDVS is a web-based news data crawl and visualization analysis system, which can display every day's changes of news data and specific of news for the user, and guide users to find the contents they interested in quickly. NDVS provides rich and flexible interactivity, also guides users to explore the internal essence of news data continually. Furthermore, the user could find the news' change trend of specific period time under the different classification according to individual interests. That can help user find news' regular pattern [2]. NDVS includes four modules: data crawling, structured storage, data analysis and data visual.

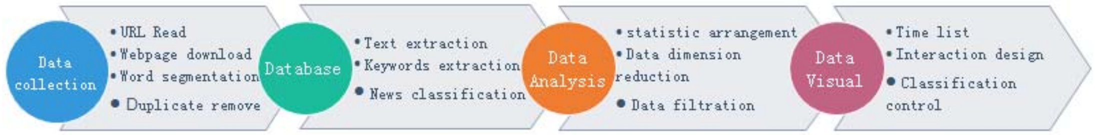

Fig. 1. System Framework of NDVS

A. Data Collection

Web spider adapts to large-scale data crawl, it visits webpages and relevant links automatically and grabs the needful information at the same time. It independents of users' operation when the web spider is running, so it can save many expenditures for us [3]. NDVS crawls news from seeds websites, such as news media sites of Sohu and Sina, by a web spider written by java. The process of data crawling follows the strategy of breadth-first. NDVS' work of duplicate webpages removing is not only delete the similar webpages that similarity beyond 80% by compare their title and text, also avoid climbing the same URL multiply. NDVS created a repository of URL for removing duplicate webpage efficiently and saving CPU resources. NDVS always search if the link has been downloaded once every time before web pages crawl, and load download URL into memory.

B. Structured Storage

Before NDVS save news data structurally, we have to classify the news data by decompose URL parameter, such as

This work was supported by the National Natural Science Foundation of China. Project no.61370063.

978-1-4673-7723-2/15 $31.00 © 2015 IEEE
DOI 10.1109/IMCCC.2015.106

the URL which is released by Sina on April 2, 2015: http://sports.sina.com.cn/nba/2015-04-02/00007561841.shtml. From ‘sports.sina.com.cn’ we know that is about sports, and the release time can also be obtained in the URL. Through the news that is grabbed from Sohu news media website, we can get the publish time of news by the same method, but getting the news classification is extracted by news tags [4]. Then we stored data in the distributed document storage database -- MongoDB. MongoDB is a database based on object storage, so we choose it. We use BSON as the data format which is similar to JSON. Furthermore, we must use a liberal and flexible database when facing with increasingly large amounts of news data [5]. Database has two collections to save news data: Quantity of Different Classifications of News and Details of News. The first one saves the quantity of daily the news which includes eight different classifications and two different data sources, and one day's news data are stored as a data document. The second one save specific detailed information about every news: release time, grab time, news category, news link, news source, news headline, link of image and text, such as a piece of news is stored as a data document.

C. Data Analysis

Statistician William Cleveland pointed out that a good chart is not only to be understood quickly but also how about the displayed content even helps you to catch sight of something that has not seen before [6]. NCVS has a flexible interactivity and could display the variation tendency of different news data classification according to time series. NDVS uses three views to analyze news data of different time period. The three views are YDV (Data View of Year), MDV (Data View of Month) and WDV (Data View of Week). YDV controls and analyzes the news data within a year by line charts and stacked bar chart which is implemented by elements of SVG and HTML5. MDV controls and analyzes the news data and general situation of news within a month by line charts. WDV is similar to MDV, but it is more detailed about the analysis of daily the news content. Web visual analysis based on AngularJS + D3.js, then we will introduce the modular concept and data visualization of NDVS.

1) Modular Analysis View：In front of the introduction, NDVS presents visual analysis of news data with three views, each view consists of three modules: Context Module, Focus Module and Classification Control Module [7].

Context Module is a time control and data preview area, and it can present the news data’s variation tendency of everyday with the line chart drew by D3.js. Besides that, Context Module configuration a brush for users to select subsets of prescriptive cycle time. Users can click-and-drag the brush, then the change trend and key details of the news will show in Focus Module. Focus Module shows achievements of visual analysis about news data. D3.js draw stacked bar chart to show a month's worth of news data in Focus Module of YDV. That data of news are selected by brush and classified by user. Focus Module gets news data which are user selected used to bind HTML tags by D3, and shows the news of the key information in MDV and WDV. Classification Control Module controls classifications of news which are user selected by a directive that written with AngularJS. The first two modules is realized by controller which is written with AngularJS. Communication between directive and controller is through the bubbling mechanism, the bubble will messaging parameters of news classification from directive to corresponding controller.

AppCtrl is the main controller of three views about NDVS, and each view has its own controller and directive. AppCtrl defines the global variable, supplied for three children controllers. Controllers are responsible for the interface rendering, including initialized page, responsive page after brush moved and classification control. Directives control news classification and pass parameters to controllers.

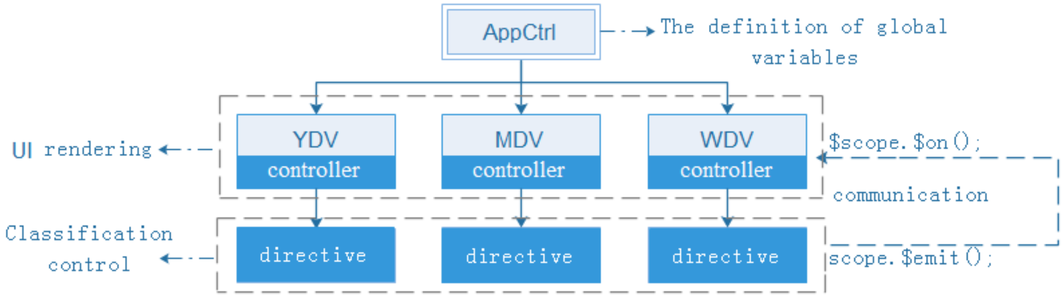

Fig. 2. Web Technical Architecture of NDVS

2) Data Filtration: NDVS intercepts the news data of different cycle time in different views. YDV fetches news data of one year by the first collection. MDV and WDV fetch news data by the whole collections. Custom brush of Context Module can be used to click-and-drag date range freely for users. That provides flexibly interactive experience. NDVS uses one day as unit when user drag brush. Date range of brush remains the same when moving and initialized. YDV’s date range of brush is the number of days of the current month. In Context Module, the domain of time is a continuous time scale on its well-defined intervals, however the domain of time is discrete ordinal scale on its well-defined intervals in Focus Module. So the time domain of brush is just obtained two points of time. Therefore we can split date by ‘d3.time.days’, and then display daily the news data with stacked bar chart in Focus Module. MDV and YDV’s date range of brush are 7 days and 1 day respectively. NDVS keeps the time of current day at 23:59:59 in Context Module for

keeping the date range of brush fixed and slip general.

III. Achievements of News Data Visual Analysis

NDVS shows the charm of news data fully by time list and interaction design. YDV is the first one.

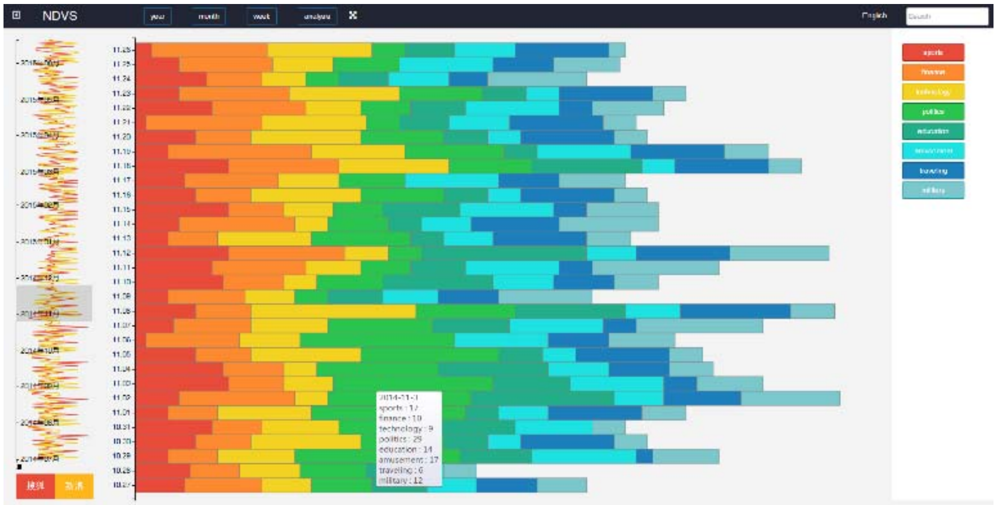

Fig. 3. Visual Analysis of YDV

Pictured above, Context Module of YDV displays the situation of news which is published from Sohu and Sina. D3.js draw the line chart to show one year's data. The line in the limited area is relatively concentrated. However we can look over the quantity of different classifications of daily the news from Focus Module by click-and-drag brush. The maximum quantity of news publish is political news on November 3rd, 2014 during the period of APEC as in Fig.3. That proved visual analysis of news data has higher directivity. For the visual data smoothly over Focus Module when we carry out the operations to update data, such as clicking the button of classification or reset, drag-and-drop brush, click to relocate the time range of brush whenever and wherever possible. NDVS added animation effects in each module of data binding. That gives user more comfortable experience in the visual sense.

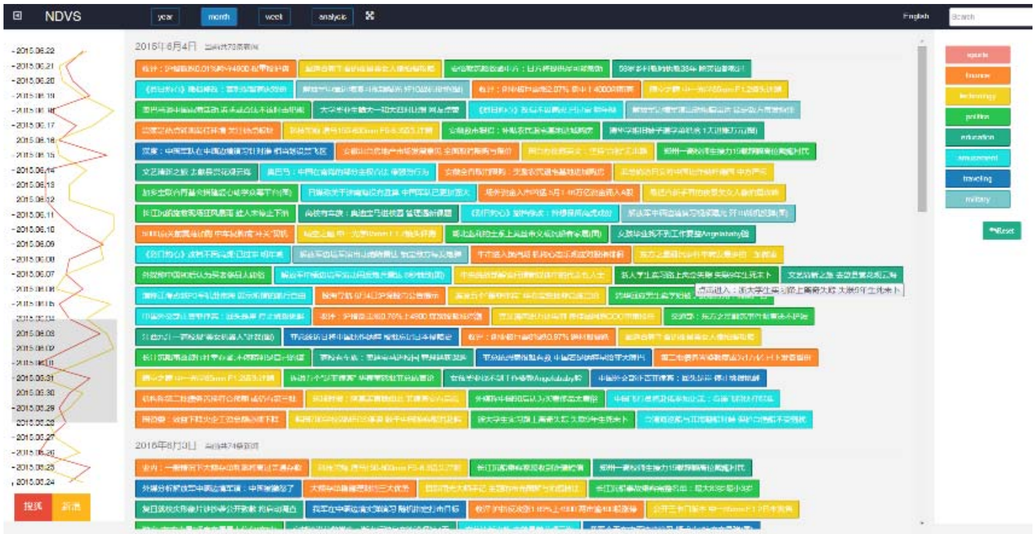

Fig. 4. Visual Analysis of MDV

MDV reveals the situation of 7 days of the news which selected by brush except sports news as in Fig.4. The floating range of brush is a month in Context Module. Congruent relationship between classifications of news and colors is realized by creating a discrete ordinal scale of style. As mentioned previously, AppCtrl define the global variable: classification array of object named $scope.categories, which is including classification name of news, the corresponding color with classification, the corresponding name of CSS property class with classification. We can obtain the operations of clicking the button of classification or reset, then messaging parameters from directive to corresponding controller through the bubbling mechanism to update data and interface rendering. WDV has screened and simplified news, as a result, view mitigated the burden of news filter for users.

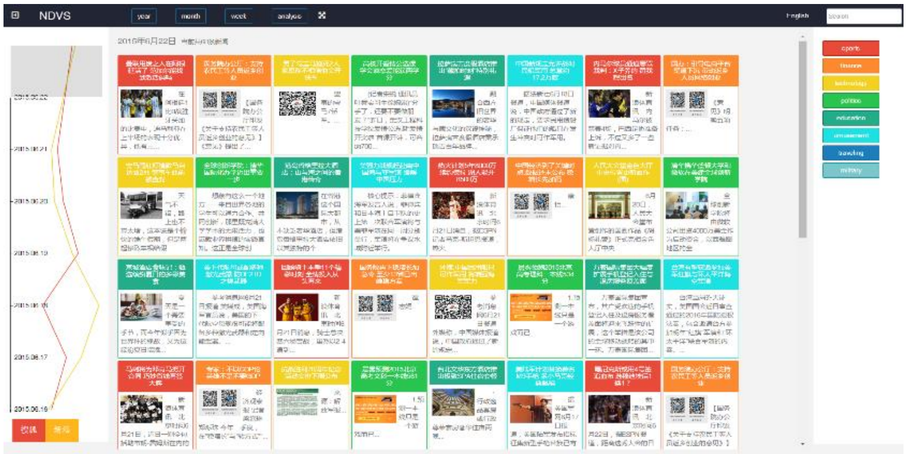

Fig. 5. Visual Analysis of WDV

Analysis view of WDV is illustrated on Fig.5. WDV shows important details of every news on Focus Module, including news title, news abstract, and some pictures. Every news headlines have added hyperlink, so that user can click to look over news website. Since every news was saved as a data document, news data which is selected have been filtered will come into being three-dimensional array. NDVS must carry on the operation of reducing the dimension of news data array, in order to provide convenience for the data binding and some interrelated operations. Users can view and catch the news that they interested in quickly in a day by WDV, in addition, user can also fast comprehend the current situation of public opinion.

IV. System Performance Test

NDVS is an applied system on the web, including browser, web server, application server, database server and basic network system[8]. NDVS uses Nodejs server which supports high concurrency. We test the performance of NDVS by WebBench at the server side when large number of users concurrent access to the system. Line chart displays the change of throughput capacity of system with the scale of users keeps growing. We used this command 'WebBench -c 1000 -t 60 http://222.31.81.3/', that has tested 1000 clients visited 59856 pages triumphantly in one minute, and the throughput capacity of the system is 3.52M per second. After tested many times, we find the phenomenon of failed request appearing when the number of users up to 3600, this means that the system is overload. The test results as shown in Fig. 6.

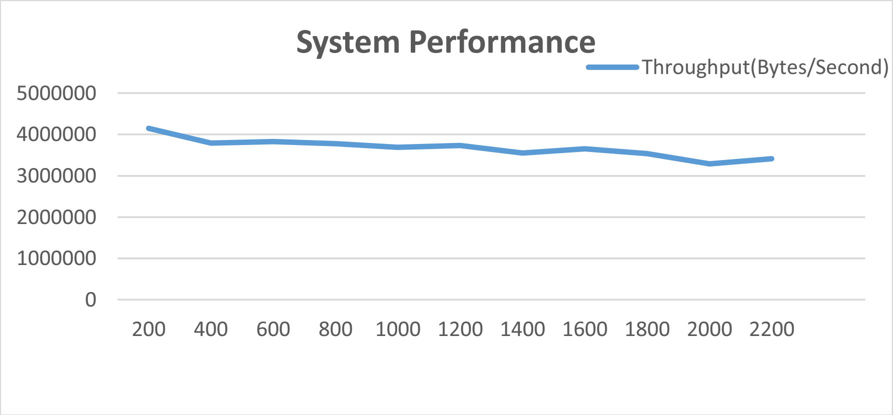

Fig. 6. Line Chart of System Performance with the Number of Clients

NDVS supports high concurrent access within the LAN as in pictured above, and picture testified system has a good performance of working. In addition, although the scale of all the news data keeps expanding with data mining, shown on the 3 views of the data size is changeless. Analysis view shown information of news no more than 31 days, so the size of news data is from 2700 to 3200. From the statistic of page load time, we can find that Chrome and IE browser just need 1.2 seconds when thousands of news data rendering. That explains the scale of news data will not affect the effect of D3 rendering, so NDVS can keep its performance efficiently in the process of later use.

V. Conclusion

We are facing such a problem as massive users can't fast catch the key details and public opinion when they read multifarious and mass of news, as a result, we design and implement NDVS. System improves experience of read news for users preferably. Test of system performance proved NDVS has feasibility. NDVS shows change trend and key details of news which have eight different classifications. NDVS displays news of a period of time and provides users with rich flexible and interactivity function. The user interface is visualized and readable, can guide users constantly explore the inner nature of news data, and help users decision-making. NDVS implemented data filtration and transition animations of view by clicking or drop-and-down brush, and permitted switch views between different types of visual based on the same subset of data. Users can browse change trend about news of a period of time depends on individual interests, and NDVS help users find the regular pattern of news. Although NDVS has made some progress in the visual analysis of news data, system still needs further more improvement. The next work will add analysis and evaluation for hot news or the news which is searched by keywords. NDVS will improve the technology of natural language processing and carry on more method of mathematical statistics used to process acquired information. Finally, NDVS displays daily the news data by more flexible and plentiful visual interactivity view.

References

[1] Julie Steele, Noah Iliinsky. Designing Data Visualizations [M]. Canada: O'Reilly Media, 2011.

[2] Lloyd L, Kechiagas D, Skiena S. Lydia: A system for large scale news analysis [J] .Volume Lecture Notes in Computer Science, 2005, 3772 :161-166.

[3] Menczer F. Complementing search engines with online web mining agents [J]. Decision Support System, 2003.

[4] Ian H. Witten, Gordon W. Paynter, Eibe Frank, et al. KEA: Practical automatic keyphrase extraction [C] .In :Proceedings of the fourth ACM conference on Digital libraries .ACM, 1999 :254 -256

[5] Chodorow, Kristina. MongoDB: The Definitive Guide [M]. Canada: O'Reilly Media, 2013.

[6] Toby Segaran, Jeff Hammerbacher. Beautiful Data: The Stories Behind Elegant Data Solutions [M]. Canada: O'Reilly Media, 2009.

[7] Manuel Lima. Visual Complexity: Mapping Patterns of Information [M]. New York: Princeton Architectural Press, 2011.

[8] Peng Yufeng, Zhang Zhan, Wang Zhige, Gao Chuanshan. Transformation of the Web Server Performance Benchmark Webstone to Windows Platform [J]. Computer Engineering. 2003(5)

[9] Zhu Jing, Shen Meiming, Wang Dongsheng. Performance Analysis and Test of Web Service System [J]. Computer Engineering and Applications, 2001(15)

[10] Yan HF, Wang JY, Li XM, et al. Architectual design and evaluation of an efficient Web-crawling system [J].Journal of Systems and Software, 2002.

[11] Pirolli P. Information Foraging Theory: Adaptive Interaction with Information. New York: Oxford University Press, 2007. 31-35.