Interactive Visualization of People's Daily

Xiaohui Wang[1](✉), Jingyan Qin[1](✉), and Dawei Li[2](✉)

[1] School of Mechanical Engineering,
University of Science and Technology Beijing, Beijing, China
wangxh14@ustb.edu.cn, qinjingyanking@foxmail.com

[2] School of History and Civilization, Shaanxi Normal University, Xi'an, China
dw-li11@outlook.com

Abstract. As the number of documents grows larger and larger, it becomes more and more difficult for people to make sense of it all. People's Daily is the official newspaper of Chinese Communist Party Central Committee, which has a decisive guiding role for the Chinese mainland politics in different periods. In this paper, we develop an interactive visual analytic system to represent 1,365,802 documents of People's Daily from 1946 to 2003, in order to help analysts examine them more quickly and dig out potential information more efficiently. It is an easy-to-use system, which provides four distinct views of document visualization, including document view, calendar view, storyline view and query view. Besides, abundant human-centered interactions and text visualization techniques are adopted to improve user experiences. Experiments verify the usability of the system. Some discoveries about the change and development in Chinese society are found by using the system.

Keywords: Interaction · Text visualization · People's Daily
Data mining

1 Introduction

Information visualization is the study of transforming data, information and knowledge into interactive visual representations [1]. For large collection of data, information visualization is a good method for data overview and mining. The visualization system has been widely used in many fields, such as chemistry [2], engineering [3], public service [4].

People's Daily is the official newspaper of Chinese Communist Party (CCP) Central Committee. On July 1 in 1946, Chairman Mao personally wrote the Chinese headline for People's Daily. In 1992, People's Daily was named as one of the top ten newspapers in the world by UNESCO. As the mouthpiece of CCP and Chinese government, People's Daily does not only actively promote the policy advocacy of CCP and Chinese government and disseminate information in all fields at home and abroad, but also record the change and development in Chinese society. In addition to providing the outside world with direct information

A. Marcus and W. Wang (Eds.): DUXU 2018, LNCS 10919, pp. 389–400, 2018.
https://doi.org/10.1007/978-3-319-91803-7_29

about the CPC's policies and opinions, the editorial of People's Daily reflects the views of the CPC Central Committee on the handling of the incidents.

Because People's Daily has a decisive guiding role for the Chinese mainland politics in different periods, many political watchers at home and abroad usually go through the essays in People's Daily to find the true meaning of the CPC Central Committee's wishes and some of Chinese unique political messages. So in this paper, we download the documents of People's Daily from 1946 to 2003 and dig out more useful information about CCP, Chinese government and society through the visualization research.

The objective of our research is to build an interactive visual analytic system to help investigators better review, analyze, understand and explore data. The challenge is how to clearly show the huge document collection and effectively explore the data to find the insightful patterns. We design the system from four distinct perspectives, called views, including document view, calendar view, storyline view and query view. The four views cooperate to provide various analytic tasks in different levels of data, and combine with convenient interactions to offer a more easy-to-use tool. In these views, some text visualization techniques are adopted to the system, such as summarizing a single document, showing the words and topics, creating storylines.

The main contributions include:

- An interactive visual analytic system is developed to investigate the data of People's Daily from four distinct views. Human-centered interactions and text visualization techniques are appropriately adopted to increase the usability of the system and improve user experiences.
- Some discoveries about the change and development in Chinese society from 1946 to 2003 are found by using the proposed system.

The rest of the paper is organized as follows. Section 2 gives related work. Section 3 shows the data collection and processing. Section 4 describes the design of four visualization views in the system. Section 5 illustrates the experimental results. Section 6 finally draws the conclusions and future work planned for the system.

2 Related Work

2.1 Text Visualization

Text visualization techniques regarding their design goals can be largely divided into five categories: revealing content, exploring document corpus, visualizing document similarity, visualizing sentiments and emotions of the text, and analyzing various domain-specific rich-text corpus [5]. Visualizing the content of a text document, a few documents, and even hundreds of thousands of documents is essential for overview of large text data, which is one of the most important tasks in text visualization. From the different levels of details, showing content of documents can be from the following aspects: summarizing a single document, showing the words and topics, detecting events, and creating storylines [5]. Many studies on exploring document corpus are query-based techniques so that users can retrieve the data based on their interests [6].

In this paper, some of these visualization techniques are adopted for our proposed interactive system, such as summarizing a single document, showing the words and topics, creating storylines, query-based document exploring.

2.2 Interactive Visualization System

The interactive visualization system has been used to many different fields, such as chemistry [2], engineering [3]. A visual analytic system Jigsaw represents documents and their entities visually from multiple coordinated views [13]. Shi et al. presents an interactive visual system for exploring complex flow patterns of Public Bicycle System [4]. A system architecture called Reactive Vega provides the robust and comprehensive treatment of declarative visual and interaction design for data visualization [14].

In this paper, we present an interactive visual analytic system to explore the change and development in Chinese society from People's Daily data.

3 Data Collection and Processing

We download 1,365,802 documents in People's Daily from May 1946 to December 2003 from the website [7]. Each document is textual, in Chinese natural language, and in loose narrative format. News, stories and reports are the main types of the documents with a few paragraphs.

We organize these documents in Json files according to the published time, and each Json file uses published month as its name, such as '194605.json', includes all the documents published in this month. Each data item in the Json file contains four parts: url, title, published time and original content. There are 692 Json files in total.

The number of documents in each year is shown in Table 1. From the statistics, we can see that there are tens of thousands of documents a year on average. In other words, there are dozens of documents or more than a hundred documents a day on average. So the data is a very comprehensive source on historical events.

For Chinese texts, word segmentation is the first step for text analysis. We use Jieba (Chinese for 'to stutter') toolkit for Chinese text segmentation [8]. The algorithm is to generate a directed acyclic graph (DAG) composed of all possible Chinese words in the sentence, then by using the dynamic programming to find the maximum probability path to find the maximum segmentation based on word frequency combination. For unregistered words, an HMM model based model and Viterbi algorithm are built based on Chinese word formation.

Besides, the keywords are automatically extracted, and the documents to be extracted can be any combinations of the downloaded documents, such documents in one day or one year. The TF-IDF based keyword extraction algorithm [15] is adopted in this system.

Table 1. The number of documents in each year.

Year	#	Year	#	Year	#	Year	#	Year	#	Year	#
1946	5954	1956	23821	1966	12538	1976	12583	1986	34345	1996	38688
1947	10773	1957	25374	1967	9461	1977	12991	1987	33397	1997	35645
1948	8247	1958	28074	1968	9364	1978	13603	1988	34298	1998	35799
1949	19023	1959	23993	1969	10342	1979	17779	1989	31074	1999	36463
1950	20169	1960	23929	1970	11762	1980	27708	1990	32783	2000	36399
1951	14504	1961	22608	1971	10908	1981	28898	1991	34774	2001	37248
1952	13397	1962	17887	1972	14537	1982	28934	1992	38446	2002	36431
1953	12820	1963	16047	1973	15029	1983	31889	1993	37562	2003	44955
1954	14035	1964	17028	1974	12984	1984	32440	1994	35321		
1955	15659	1965	16724	1975	13543	1985	35024	1995	39791		

4 System Design and Implementation

4.1 Overview

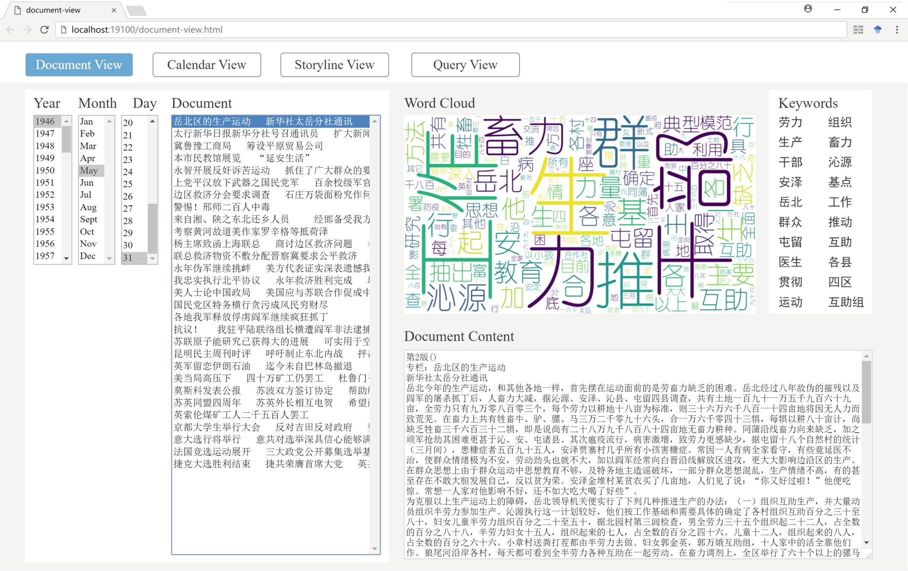
The interactive visual analytic system is a web based system to represent documents visually in order to help analysts learn about the content more effectively. The system visualizes the data from different aspects through various distinct views as follows.

- **Document view.** Document view is a single document summarization. Users can select one document, the system gives the word cloud and keywords automatically extracted from the document.
- **Calendar view.** Calendar view presents a quick overview of documents in one year or one month.
- **Storyline view.** Storyline view demonstrates the keywords in time series to explore the process of the events.
- **Query view.** Query view provides a word-based document search function.

Specially, the four views focus on the data in different levels and cooperate to provide various analytic tasks. Convenient interactions are combined with these four perspectives to provide a more easy-to-use tool for obtaining a more comprehensive understanding of data. The detail for each view and interactions will be described in the following subsections.

4.2 Document View

The document view, shown in Fig. 1, is a single document summarization. There are four parts in the document view interface. The left part is the selection area, users select one day, then the document list shows all the documents published in that day. After users select one document, three visualization perspectives of



岳北区的生产运动　新华社太岳分社通讯
太行新华日报新华分社号召通讯员　扩大新闻
冀鲁豫工商局　筹设平原贸易公司
本市民教馆展览　“延安生活”
永智开展反奸诉苦运动　抓住了广大群众的要
上党平汉放下武器之国民党军　百余校级军官
边区救济分会要求调查　石庄万袋面粉究作何
警惕！邢师二百人中毒
来自湘、陕之东北还乡人员　经邯备受我方
考察黄河故道美作家罗辛格等抵荷泽
杨主席致函上海联总　商讨边区救济问题
联总救济物资不敷分配晋察冀要求公平救济
永年伪军继续挑衅　美方代表证实深表遗憾我
我忠实执行北平协议　永年救济胜利完成
美人士论中国政局　美国应与苏联合作促成中
国民党区特务横行贪污成风民穷财尽
各地我军释放俘虏阎军继续疯狂抓丁
抗议！　我驻平陆联络组长横遭阎军非法逮捕
苏联原子能研究已获得大的进展　可实用于空
昆明民主周刊时评　呼吁制止东北内战　抨
英军留恋伊朗石油　迄今未自巴林岛撤退
美当局高压下　四十万矿工仍罢工　杜鲁门
莫斯科发表公报　苏波双方签订协定　帮助
苏英同盟四周年　苏英外长相互电贺　希望
英索伦煤矿工人二千五百人罢工
京都大学生举行大会　反对吉田反对政府
意大选行将举行　意共对选举深具信心能够满
法国竞选运动展开　三大政党公开募集选举基
捷克大选胜利结束　捷共荣膺首席大党　英

劳力	组织
生产	畜力
干部	沁源
安泽	基点
岳北	工作
群众	推动
屯留	互助
医生	各县
贯彻	四区
运动	互助组

第2版()
专栏：岳北区的生产运动
新华社太岳分社通讯
岳北今年的生产运动，和其他各地一样，首先摆在运动面前的是劳畜力缺乏的困难。岳北经过八年敌伪的摧残以及阎军的屠杀抓丁后，人畜力大减，据沁源、安泽、沁县、屯留四县调查，共有土地一百九十一万五千九百六十九亩，全劳力只有九万零八百零三个，每个劳力以耕地十八亩为标准，则三十六万六千八百一十四亩地将因无人力而致荒芜。在畜力上共有牲畜牛、驴、骡、马三万二千零九十六头，合一万六千零四十三犋，每犋以耕八十亩计，尚缺乏牲畜三千六百三十二犋，即是说尚有二十八万九千八百八十四亩地无畜力耕种。同蒲沿线畜力向来缺乏，加之顽军抢劫其困难更甚于沁、安、屯诸县。其次瘟疫流行，病害激增，致劳力更感缺少，据屯留十八个自然村的统计（三月间），患糠症者五百九十五人，安泽贾寨村几乎所有小孩害糠症。常因一人有病全家看守，有些竟延医不治，使群众情绪极为不安，劳动劲头也就不大，加以阎军经常向白晋沿线解放区进攻，更大大影响边沿区的生产。在群众思想上由于群众运动中思想教育不够，及特务地主造谣破坏，一部分群众思想混乱，生产情绪不高，有的甚至存在不敢大胆发展自己，反以贫为荣。安泽金堆村某贫农买了几亩地，人们见了说：“你又好过啦！”他便吃惊。常想一人家对他影响不好，还不如大吃大喝了好些”。
为克服以上生产运动上的障碍，岳北领导机关便实行了下列几种推进生产的办法：（一）组织互助生产，并大量动员组织半劳力参加生产。沁源执行这一计划较好，他们按工作基础和需要具体的确定了各村组织互助百分之三十至八十，妇女儿童半劳力组织百分之二十至五十，据北园村第三闾检查，男全劳力三十五个组织起二十二人，占全数的百分之八十八，半劳力妇女十五人，组织起来的七人，占全数的百分之四十六。儿童十二人，组织起来的八人，占全数的百分之六十六。小章村送粪打茬都由半劳力去做。妇女郭金英，郭万娥互助组，十人家中的活全靠他们作。狼尾河沿岸各村，每天都可看到全半劳力各种互助在一起劳动。在畜力调剂上，全区举行了六十个以上的骡马

Fig. 1. The document view.

a document (word cloud, keywords and original content) are illustrated on the right parts.

Word cloud [9] is one of the most popular and intuitive techniques for word visualization, which shows a bag of words with different sizes and colors that summarize the content of the input document and packed together without any overlap. The higher the frequency of a word is, the larger its font size is. Color is used to distinguish different words for easy recognition. There are different types of word cloud based on various packing methods, such as Wordle [10], Radcloud [11]. In this paper, we use the word cloud toolkit [12], which is a little word cloud generator in Python. The input of the word cloud toolkit is the segmentation results by the Jieba toolkit [8].

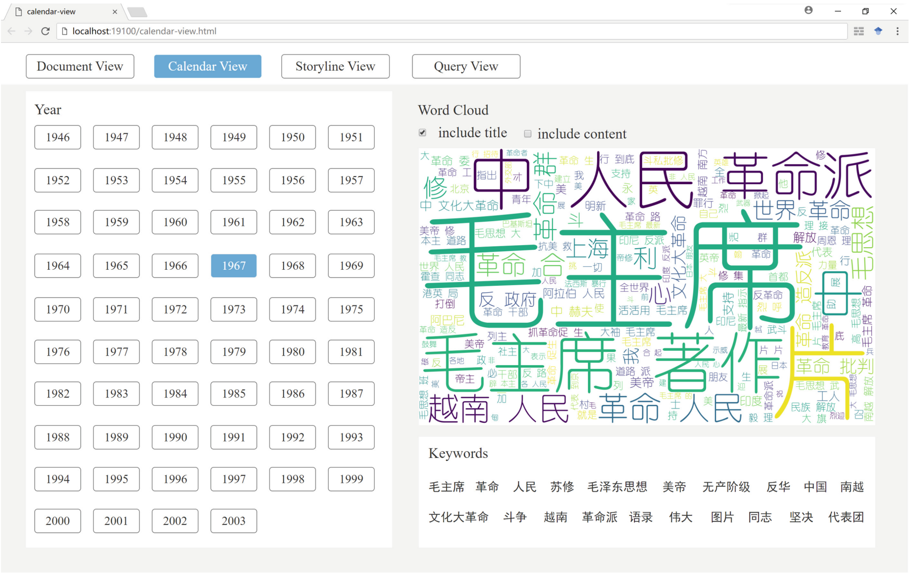
The part of keywords provides a more intuitive view of the document. 20 Keywords are automatically extracted from the document content by the TF-IDF based keyword extraction algorithm [15].

The visualization results cannot replace the documents. If the users find the interests from Word Cloud and keywords, in most cases they want to read the original document content carefully to learn more about it. So we present the document content in the right corner.

In the word cloud part, users can zoom in or zoom out to review the details. If users hover over a keyword, all of this keyword in the original content are highlighted to give users an intuitive view.

4.3 Calendar View

The calendar view, illustrated in Fig. 2, presents a quick overview of documents in one year or one month. The style of the selector on the interface is a familiar calendar showing years and month. The users can click the year, then the selected year is highlighted and the right part shows the word cloud and keywords generated by all the documents in this year. Besides, if the year on the left selection area is double clicked, the month selector is shown. Then users can select one month, the right part shows the word cloud and keywords generated by all the documents in this month. Besides, users can select more than one year and one month, and combine years and months anyway. As same as in the document view, users can zoom in or zoom out to review the details in Word Cloud part.



毛主席 革命 人民 苏修 毛泽东思想 美帝 无产阶级 反华 中国 南越
文化大革命 斗争 越南 革命派 语录 伟大 图片 同志 坚决 代表团

Fig. 2. The calendar view.

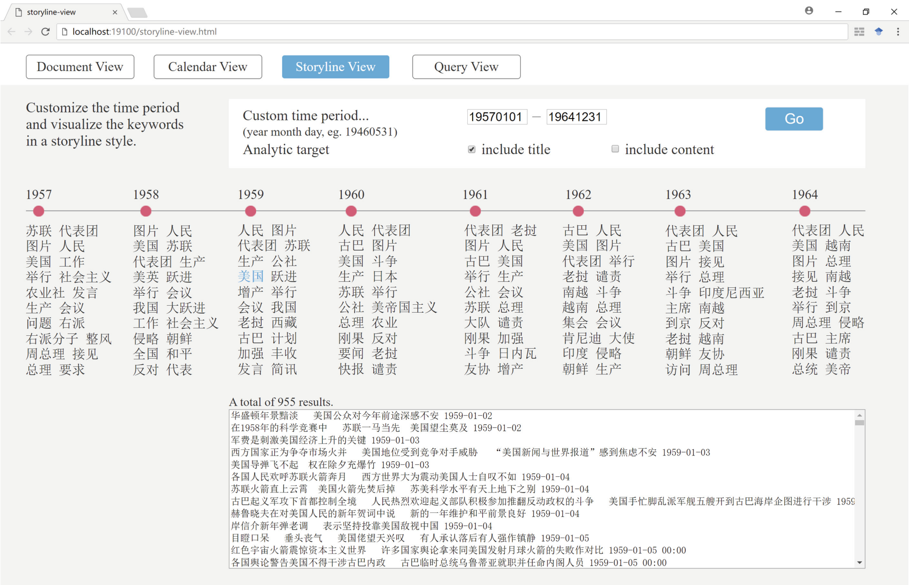
In Fig. 2, the year 1967 is selected, the word cloud and keywords are generated by all the documents published in 1967. The word 'Chairman Mao' is in the biggest size and also the top one keyword. Besides, from 'Zedong Mao', 'Revolutionaries' and 'the Great Proletarian Cultural Revolution' in Word Cloud and Keywords, we can infer to the ear of Great Proletarian Cultural Revolution in China. Chairman Mao got started the Great Proletarian Cultural Revolution from May, 1996 to October, 1976, which was the most volatile and disastrous stage in China. In People's Daily in 1967, the words 'Zedong Mao' and 'Revolutionaries' occurred frequently, which illustrated that the Great Proletarian Cultural Revolution becames the very important political event in China.

The calendar view provides a very quick and easy-to-understand method to find potential interests. To achieve this purpose, users can only choose by years and month as the unit in this view. If users are interested with keywords of one year or one month, they can jump to the document view to review the detail. Furthermore, if users want to customize any time period, the storyline view provides this function.

4.4 Storyline View

The document view and calendar view give the whole view for any time periods. Sometimes, users want to explore the process of the events. For this purpose, the storyline view is designed, shown in Fig. 3, which demonstrates the keywords in time series. Users can customize the time period and select the analytic target. The analytic target can be the title, the content or the both. This view supports different scales of keyword visualization, such as year, month or day. The users can zoom in or zoom out to switch to different scales. Under the time point in the storyline visualization, top 20 keywords extracted from the document titles or contents based on the selected analytic target in the given time period are shown.

One keyword can be selected by hovering over it, then the bottom box immediately shows all the document titles in which the analytic target contains the keyword. Users can double click the item in the bottom box, then the word cloud, the keywords and the original document content are popped up in a floating window. When the mouse clicks the storyline view window, the floating window disappears.



Customize the time period
and visualize the keywords
in a storyline style.

Custom time period... 19570101 – 19641231
(year month day, eg. 19460531)
Analytic target ☑ include title □ include content

1957	1958	1959	1960	1961	1962	1963	1964
苏联 代表团	图片 人民	人民 图片	人民 代表团	代表团 老挝	古巴 人民	代表团 人民	代表团 人民
图片 人民	美国 苏联	代表团 苏联	古巴 图片	图片 人民	美国 图片	古巴 美国	美国 越南
美国 工作	代表团 生产	生产 公社	美国 斗争	古巴 美国	代表团 举行	图片 接见	图片 总理
举行 社会主义	美英 跃进	美国 跃进	生产 日本	举行 生产	老挝 谴责	举行 总理	接见 南越
农业社 发言	举行 会议	增产 举行	苏联 举行	公社 会议	南越 斗争	斗争 印度尼西亚	老挝 斗争
生产 会议	我国 大跃进	会议 我国	公社 美帝国主义	苏联 总理	越南 总理	主席 南越	举行 到京
问题 右派	工作 社会主义	老挝 西藏	总理 农业	大队 谴责	集会 会议	到京 反对	周总理 侵略
右派分子 整风	侵略 朝鲜	古巴 计划	刚果 反对	刚果 加强	肯尼迪 大使	老挝 越南	古巴 主席
周总理 接见	全国 和平	加强 丰收	要闻 老挝	斗争 日内瓦	印度 侵略	朝鲜 友协	刚果 谴责
总理 要求	反对 代表	发言 简讯	快报 谴责	友协 增产	朝鲜 生产	访问 周总理	总统 美帝

A total of 955 results.

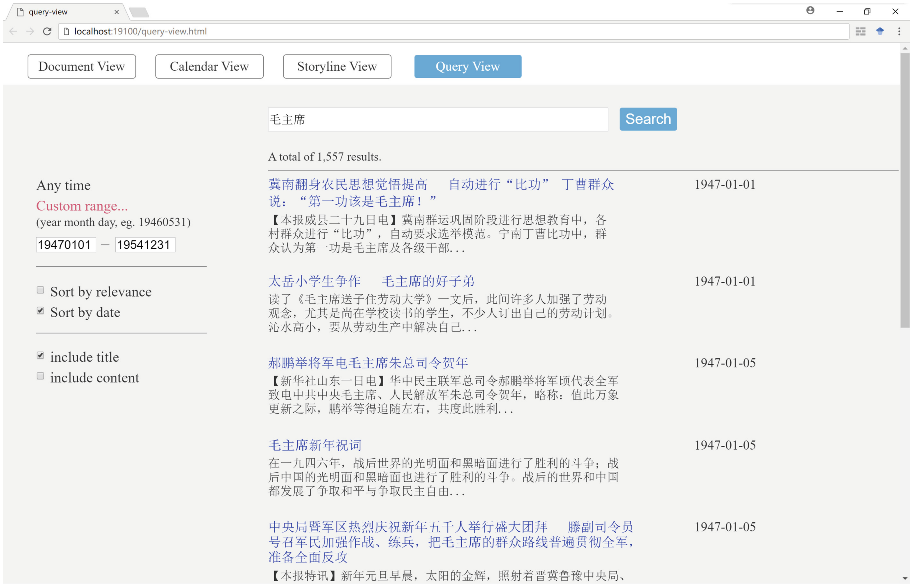
```
华盛顿年景黯淡　　美国公众对今年前途深感不安 1959-01-02
在1958年的科学竞赛中　　苏联一马当先　美国望尘莫及 1959-01-02
军费是刺激美国经济上升的关键 1959-01-03
西方国家正为争夺市场火并　　美国地位受到竞争对手威胁　　“美国新闻与世界报道”感到焦虑不安 1959-01-03
美国导弹飞不起　权在除夕充爆竹 1959-01-03
各国人民欢呼苏联火箭奔月　　西方世界大为震动美国人士自叹不如 1959-01-04
苏联火箭直上云霄　美国火箭先焚后掉　　苏美科学水平有天上地下之别 1959-01-04
古巴起义军攻下首都控制全境　　人民热烈欢迎起义部队积极参加推翻反动政权的斗争　　美国手忙脚乱派军舰五艘开到古巴海岸企图进行干涉 1959
赫鲁晓夫在对美国人民的新年贺词中说　　新的一年维护和平前景良好 1959-01-04
岸信介新年弹老调　　表示坚持投靠美国敌视中国 1959-01-04
目瞪口呆　　垂头丧气　　美国佬望天兴叹　　有人承认落后有人强作镇静 1959-01-05
红色宇宙火箭震惊资本主义世界　　许多国家舆论拿来同美国发射月球火箭的失败作对比 1959-01-05 00:00
各国舆论警告美国不得干涉古巴内政　　古巴临时总统乌鲁蒂亚就职并任命内阁人员 1959-01-05 00:00
```

Fig. 3. The storyline view.

For example, in Fig. 3, the time period is set from Jan. 1, 1957 to Dec. 31, 1964. The analytic target only includes titles. The keywords of the eight years are shown in the storyline. The storyline results show the keywords by unit 'year' as the visualization scale first based on the selected time period and the space on the interface. Users can zoom in the storyline area to review the details. By hovering over the keyword 'United States' under the year 1959, there are a total of 955 documents which contain the keyword. The bottom box shows the titles and published dates of all the 955 documents.

4.5 Query View

The query view, shown in Fig. 4, provides a word-based document search function. The search options contain time period, sort type and search target. The time period can be anytime and custom range. Any time is from May 1946 to December 2003. The search results are sorted by relevance or by date specified by the sort type. The search target sets the search range, including titles or contents. The search result demonstrates the document title, the published date and part of the content. For each result, users can view the original document content by clicking the title whose font color is a little dark blue.



毛主席

Any time
Custom range...
(year month day, eg. 19460531)
19470101 – 19541231

□ Sort by relevance
☑ Sort by date

☑ include title
□ include content

A total of 1,557 results.

冀南翻身农民思想觉悟提高　自动进行“比功”　丁曹群众说：“第一功该是毛主席！” 1947-01-01
【本报威县二十九日电】冀南群运巩固阶段进行思想教育中，各村群众进行“比功”，自动要求选举模范。宁南丁曹比功中，群众认为第一功是毛主席及各级干部...

太岳小学生争作　毛主席的好子弟 1947-01-01
读了《毛主席送子住劳动大学》一文后，此间许多人加强了劳动观念，尤其是尚在学校读书的学生，不少人订出自己的劳动计划。沁水高小，要从劳动生产中解决自己...

郝鹏举将军电毛主席朱总司令贺年 1947-01-05
【新华社山东一日电】华中民主联军总司令郝鹏举将军顷代表全军致电中共中央毛主席、人民解放军朱总司令贺年，略称：值此万象更新之际，鹏举等得追随左右，共度此胜利...

毛主席新年祝词 1947-01-05
在一九四六年，战后世界的光明面和黑暗面进行了胜利的斗争；战后中国的光明面和黑暗面也进行了胜利的斗争。战后的世界和中国都发展了争取和平与争取民主自由...

中央局暨军区热烈庆祝新年五千人举行盛大团拜　滕副司令员号召军民加强作战、练兵，把毛主席的群众路线普遍贯彻全军，准备全面反攻 1947-01-05
【本报特讯】新年元旦早晨，太阳的金辉，照射着晋冀鲁豫中央局、

Fig. 4. The query view.

For example, in Fig. 4, the search item is 'Chairman Mao', the time period is from Jan. 1, 1947 to Dec. 31, 1954, the search results are sorted by date and the search target only includes titles. There are 1,557 search results, which are listed in the main area of the interface.

5 Experiments

In the experiments, we invited a few teachers and students in China School of History and Civilization to use the system. We just tell them the link of the system and let them to find the interest point. In addition to testing the functionality of the system, we want to test the ease use of the system, so we do not demonstrate the function and operations. Using the proposed system, many CCP and Chinese government policies toward the political, economy, culture and so on from 1946 to 2003 are found.

5.1 Case Study 1: Chinese Economic Policy Analysis

The first case is an example of Chinese economic policy analysis by using the document view of the visualization system.

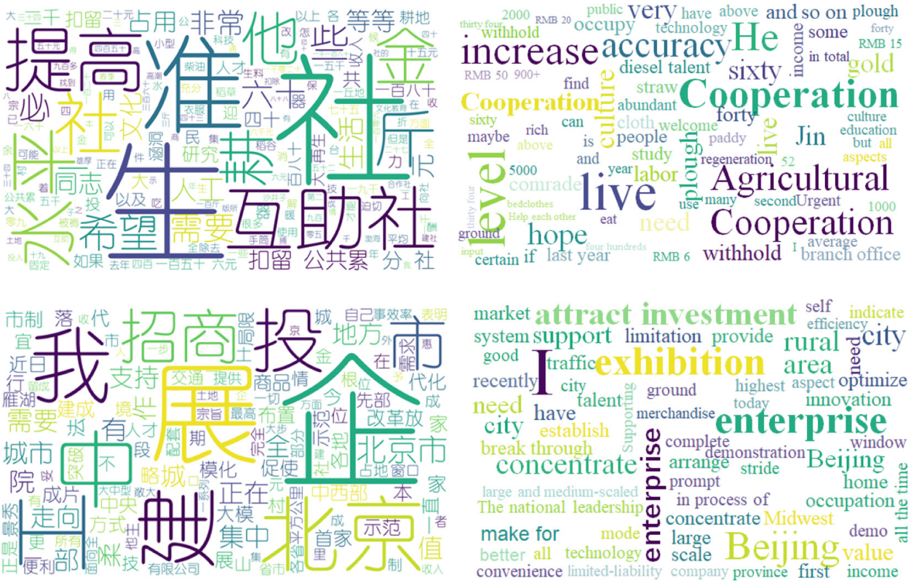


Fig. 5. The word clouds of two documents. The top one corresponds to the document published in Feb. 2, 1958. The bottom one corresponds to the document published in Nov. 26, 1992.

Two typical documents are chosen. The first one is 'new problems arising from the great leap forward in agricultural production' published in Feb. 2, 1958. The second one is 'the operation of the market mechanism to all parts of the country merchants' published in Nov. 26, 1992. The word clouds of the two documents are shown in Fig. 5.

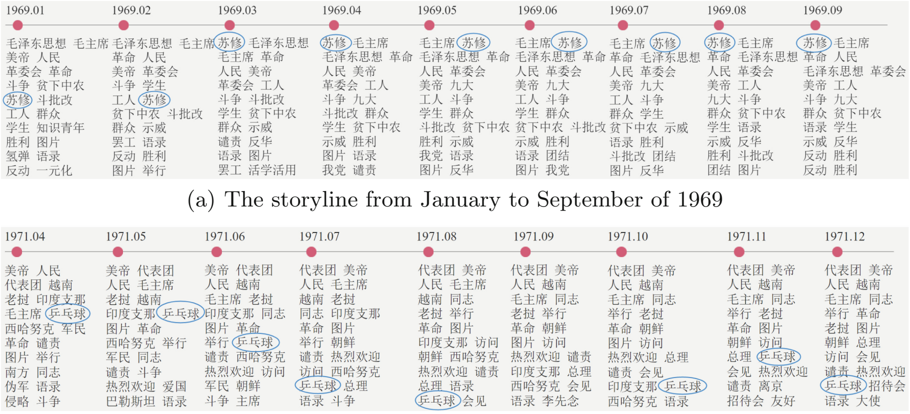
From the top word cloud in Fig. 5, we find that the word 'Agricultural Cooperation' is the interest point. The Agricultural Cooperation from 1956 to 1958 was a part of Chinese socialist system of public ownership and the Planned Economy, in which agricultural lands, machines and other production materials belong to the agricultural collectives. Most of the Agricultural Unions profits is taken away by government. The personal consumption of farms are allocated according to his work.

From the bottom word cloud in Fig. 5, we find that there were many enterprises from the rural areas of China which sought to attract the capital investment in Beijing in 1992. This phenomenon tells us that the rural areas were included into the system of Socialist Market Economy at this time. But, in 1950 s there was not any enterprises in the rural areas, because CCP and Chinese government implemented the Agricultural Cooperation Policy and prohibited the Market Economy. So, we can see the changing of CCP and Chinese government in economic policies and the rural areas from 1950 s to 1990s by the two documents.

5.2 Case Study 2: Foreign Policy Analysis

The second case is an example of foreign policy analysis by using the storyline view.

From the storyline view, we find the keyword 'Suxiu' from January to December of 1969. Figure 6(a) is a screenshot of the storyline from January to September of 1969, users can zoom in and zoom out to switch to different visualization scales. 'Suxiu' literally is Soviet Revisionism, which refers to that Nikita Khrushchev criticized Stalins doctrine and put pressure on China in 1960 s after he became the Premier of the Soviet Union. Since then, the relationship between Soviet Union and China has deteriorated, and the Socialist Alliance has ruptured. So Soviet Union was called 'Suxiu' in People's Daily in 1960s, by which we can find the change of relationship between Soviet Union and China.


1969.01
毛泽东思想 毛主席 美帝 人民 革委会 革命 斗争 贫下中农 苏修 斗批改 工人 群众 学生 知识青年 胜利 图片 氢弹 语录 反动 一元化
1969.02
毛泽东思想 毛主席 革命 人民 美帝 革委会 斗争 学生 工人 苏修 贫下中农 斗批改 群众 示威 罢工 语录 反动 胜利 图片 举行
1969.03
苏修 毛泽东思想 毛主席 革命 人民 美帝 革委会 工人 斗争 斗批改 学生 贫下中农 群众 示威 谴责 反华 语录 图片 罢工 活学活用
1969.04
苏修 毛主席 毛泽东思想 革命 人民 美帝 革委会 工人 斗争 九大 斗批改 群众 学生 贫下中农 示威 胜利 图片 语录 我党 谴责
1969.05
毛主席 苏修 毛泽东思想 革命 人民 革委会 美帝 九大 工人 斗争 学生 群众 斗批改 贫下中农 胜利 示威 我党 语录 图片 反华
1969.06
毛主席 苏修 毛泽东思想 革命 人民 革委会 美帝 九大 工人 斗争 学生 群众 贫下中农 斗批改 示威 胜利 语录 团结 图片 我党
1969.07
毛主席 苏修 革命 毛泽东思想 人民 革委会 美帝 九大 工人 斗争 群众 学生 贫下中农 示威 语录 胜利 斗批改 团结 图片 反华
1969.08
苏修 毛主席 革命 毛泽东思想 人民 革委会 美帝 工人 九大 斗争 群众 贫下中农 学生 语录 示威 反华 胜利 斗批改 团结 图片
1969.09
苏修 毛主席 革命 人民 毛泽东思想 革委会 美帝 工人 斗争 九大 群众 贫下中农 语录 学生 示威 反华 反动 胜利 反动 胜利

(a) The storyline from January to September of 1969

1971.04
美帝 人民 代表团 越南 老挝 印度支那 毛主席 乒乓球 西哈努克 军民 革命 谴责 图片 举行 南方 同志 伪军 语录 侵略 斗争
1971.05
美帝 代表团 人民 毛主席 老挝 越南 印度支那 乒乓球 图片 革命 西哈努克 举行 军民 同志 谴责 斗争 热烈欢迎 爱国 巴勒斯坦 语录
1971.06
美帝 代表团 人民 越南 毛主席 老挝 印度支那 同志 图片 革命 举行 乒乓球 谴责 西哈努克 热烈欢迎 访问 军民 朝鲜 斗争 主席
1971.07
代表团 美帝 人民 毛主席 越南 老挝 同志 印度支那 革命 图片 举行 朝鲜 谴责 热烈欢迎 访问 西哈努克 乒乓球 总理 语录 斗争
1971.08
代表团 美帝 人民 毛主席 越南 同志 老挝 举行 革命 图片 印度支那 访问 朝鲜 西哈努克 热烈欢迎 谴责 总理 语录 乒乓球 会见
1971.09
代表团 美帝 人民 越南 毛主席 同志 老挝 举行 革命 朝鲜 图片 访问 热烈欢迎 谴责 印度支那 总理 西哈努克 会见 语录 李先念
1971.10
代表团 美帝 人民 越南 毛主席 同志 举行 老挝 革命 朝鲜 图片 访问 热烈欢迎 总理 谴责 会见 印度支那 乒乓球 西哈努克 语录
1971.11
代表团 美帝 人民 越南 毛主席 同志 举行 老挝 革命 图片 朝鲜 访问 总理 乒乓球 会见 热烈欢迎 谴责 离京 招待会 友好
1971.12
代表团 美帝 人民 越南 毛主席 同志 举行 革命 老挝 图片 朝鲜 总理 访问 会见 谴责 热烈欢迎 乒乓球 招待会 语录 大使

(b) The storyline from April to December of 1971

Fig. 6. The storylines.

From the storyline view shown in (b) of Fig. 6, we find the keyword 'Pingpong' from April to December in 1971. From April 10 to 17 in 1971, American Pingpong team was invited to China, which was the first official activity between America and China after the establishment of People's Republic of China and considered as symbol of ice breaker. After then, America and China have established the formal diplomatic relation. So we can find the important events of international relations and the change of international relations by the keywords in these storylines.

6 Conclusion

In this paper, we build an interactive visualization system for exploring the text data of People's Daily. First, we download 1,365,802 textural documents in People's Daily from May 1946 to December 2003. Then, we design the visualization system from different aspects through four distinct views. Convenient interactions are combined with these four perspectives to provide a more easy-to-use tool. Finally, experiments verify the usability of the system. By using the system, some history events about the change and development in Chinese society from 1946 to 2003 are shown clearly.

In the future research, more computational linguistics analysis techniques can be applied to the system. For example, the topics of the documents are automatically extracted and analysts can digest the development from different perspectives such as politics, economics, military and culture. Besides, the pictures in these documents can be collected and image processing techniques are adopted to enrich the analysis. All of the above are fruitful fields for future work.

Acknowledgment. This work was supported by the National Natural and Science Foundation of China (61602033), Science and Technology Plan Project of Beijing (Z171100001217009) and the Social Science Fund of Beijing (16YTC027). This work was also supported by the State Scholarship Fund of China.

References

1. Liu, S., Cui, W., Wu, Y., Liu, M.: A survey on information visualization: recent advances and challenges. Vis. Comput. **30**, 1373–1393 (2014)
2. Pettersen, E.F., Goddard, T.D., Huang, C.C., et al.: UCSF Chimera a visualization system for exploratory research and analysis. J. Comput. Chem. **25**(13), 1605–1612 (2004)
3. Upson, C., Faulhaber, T.A., Kamins, D., et al.: The application visualization system: a computational environment for scientific visualization. IEEE Comput. Graph. Appl. **9**(4), 30–42 (1989)
4. Shi, X., Yu, Z., Chen, J., Xu, H., Lin, F.: The visual analysis of flow pattern for public bicycle system. J. Vis. Lang. Comput. **45**, 51–60 (2018)
5. Cao, N., Cui, W.: Overview of text visualization techniques. Introduction to Text Visualization. ABAI, vol. 1, pp. 11–40. Atlantis Press, Paris (2016). https://doi.org/10.2991/978-94-6239-186-4_2

6. Kucher, K., Kerren, A.: Text visualization techniques: taxonomy, visual survey, and community insights. In: 2015 IEEE Pacific Visualization Symposium (PacificVis), Hangzhou, China, pp. 117–121 (2015)
7. People's Daily. http://paper.people.com.cn
8. Jieba: Chinese text segmentation toolkit. https://github.com/fxsjy/jieba/
9. Kaser, O., Lemire, D.: Tag-cloud drawing: algorithms for cloud visualization. arXiv preprint cs/0703109 (2007)
10. Viegas, F.B., Wattenberg, M., Feinberg, J.: Participatory visualization with wordle. IEEE Trans. Visual. Comput. Graph. **15**(6), 1137–1144 (2009)
11. Burch, M., Lohmann, S., Beck, F., Rodriguez, N., Di Silvestro, L., Weiskopf, D.: Radcloud: visualizing multiple texts with merged word clouds. In: 2014 18th IEEE International Conference on Information Visualisation (IV), pp. 108–113 (2014)
12. Word cloud toolkit. https://github.com/amueller/word_cloud
13. Stasko, J., Gorg, C., Liu, Z.: Jigsaw: supporting investigative analysis through interactive visualization. Inf. Visual. **7**, 118–132 (2008)
14. Satyanarayan, A., Russell, R., Hoffswell, J., Heer, J.: Reactive vega: a streaming dataflow architecture for declarative interactive visualization. IEEE Trans. Visual. Comput. Graph. **22**(1), 659–668 (2016)
15. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)