Contents

Chapter 1 Introduction

1.1 Research Background

In 21 centuries, there is a tremendous change with the development of internet. More and more people begin to participate in the internet and the ways to distribute information is getting much more convenient. As a matter of fact, the quantity of information in the internet is increasing rapidly and exponentially. A large amount of text and rich text information has undoubtedly enriched people's lives. However, it takes people much more time to keep focus on searching the certain data within a large amount of text information. The emergence of search engines has meet the requirement of people to find certain information during a large amount of data. On the other hand, people also has the requirement to search for those breaking news and the generalized information based on that specific breaking news. In order to meet people's requirements, some corresponding services and applications have appeared in the network, such as internet subscription service. This requires users to subscribe to some certain information sources to the get the latest information.

The emergence of breaking news mining and analysis is to meet those requirements of internet uses. It could not only help us to find out the latest breaking news that have occurred in the internet recently, but also divide these breaking news events into various categories or say different groups. The news events that are grouped together will then be arranged and counted in a certain order. This way, we could derive a series of categories of the breaking news event. This helps us to understand the information trend in the internet more easily and quickly and also guides internet users to focus on the breaking news in social life. In other words, the emergence of breaking news analysis has greatly facilitates people's grasp of breaking news.

Over the years, with the emergence of a large number of online media and the emergence of Web 2.0, various kinds of information has been greatly enriched. It is because of this phenomenon that people and increasingly demanding the service s of mining and analyzing breaking news events. As a result, this task has received more and more attentions and concerns. At the same time, the analysis of breaking news event is also an important aspect of the evolution of the internet. It is of great significance to the study of the evolution of the internet:

From the user's point of view, the analysis of breaking news event will undoubtedly provide users with much more convenient and efficient services. It can enable users to obtain the most comprehensive content at a lowest cost, and it can help users to understand the breaking news events more accurately and quickly. At the sometime, breaking news event also enables users to keep track of new information while keeping abreast of the latest and most recent breaking events in their own areas of interest. The breaking news analysis system can also enable users to participate more in the discussion of breaking events and can make themselves not only the recipients of information but also the initiators of information. On the other hand, the information providers and internet's controllers could study from the breaking news events of the internet and get the internet users' behaviors and the public opinion of the internet. In this way, the website builders and designers can better design website content according to the new preferences of web users, so that the quality of the website will be greatly improved. Eventually, the mining analysis of breaking news can also help the internet supervisors to control and guide the content typed of the internet. Therefore, the analysis of breaking news has a very important significance. This task is extremely challenging due to the extensiveness and complexity of this task itself.

Breaking news analysis mainly includes: First, we need to obtain news information from the internet. This part of information mainly related to web page information. We need to set up a time and topic to get the certain information from the internet. After getting the original web page information, we need to analyze the web page to extract the important information that helps the breaking news analysis. Such as the title of the news page, date, main text, publisher and so on. After extracting the original web page information, we need to process the information of those collected texts. These processes mainly include:

The classification of web content. There are different fields of the news, such as politics, economy, education, entertainment, etc. News classification can not only help users to search for certain data they want, but also has the following advantages: it can organize a large amount of news data effectively, so as to discover the laws of some new trends. However, there is a certain degree of overlap between various news on the internet. For example, for a breaking event(such as presidential election, natural disaster) at a certain time, there will be various media and websites that report the news in different forms. However, the content is very similar. If we could categorize news by topic and divide similar news to similar category,

people could then read the related news according to corresponding topic.

Classifying news information is also more conducive to grouping news under the same topic, because there are obvious differences between some news in the internet. These news can be easily identified by the classification method. If we simply cluster them together without discrimination, the result of clustering may be influenced by the noise data. What's more, this will also affect the efficiency of clustering.

Next is to cluster information under the same category. The so-called clustering refers to the aggregation of the same or similar news content under the category, so that it is more convenient for users to search for information. Finally, after analyzing the news information, we will show the large categories of information and the news topics under each category. This could help users to grasp the trends of the breaking news.

1.2 Research Status

With the abundance of internet resources and the upcoming people's dependence on the internet. It is not only the text processing technologies have been greatly developed, it has also received more and more attention and concerns in the field of breaking news information analysis. First of all, the task of mining web page breaking news requires a lot of text processing technology. Researches related to this technology has become more and more well-developed. For example, the parsing of HTML text after obtaining text, extraction of the body of web page, natural language processing (including analysis of word segmentation and grammar and syntax), classification and clustering of text, etc. The classification of breaking news information is an organic integration of the above-mentioned related technologies. When classifying the breaking news, we should consider the information processing flow, related technologies used in each specific step and the related optimization and customization of technologies.

1.2.1 Web Crawling

Based on the search objects, the mainstream web crawler technologies include the following two types:

The first is based on link analysis. In the 1990s, foreign search engine developers have begun to model social networks. Experts have designed and developed hyperlinks between pages through a network of people-to-people relationships. At the same time, they were

surprised to find that the highest degree of similarity was in traditional citations. In this way, it is possible to analyze the conclusions through comparison. Starting from the perspective of the relational network, it is possible to classify a large number of web pages on the Internet. As early as 2002, the most primitive link-based search system appeared in Europe and the United States.

The second is content-based search. Compared to search methods based on link analysis, this is a breakthrough in search technology. They have adopted a new way of thinking and established a theme-oriented thesaurus. When the user searches in a professional field, the thesaurus and the crawler can be combined for information retrieval. Due to the change of search aspect, this new technology has gradually begun to attract people's attention. In the 1990s, the Fish Search System was developed as the first content-based search system. Later, in 1998 and 1999, Shark Search System and Focused Crawler emerged one after another.

Up to now, web crawler technology has achieved a considerable development and progress. Foreign typical systems include CORA and IBM Focused Crawler is known all over the world. CORA is a topical search engine designed for computer science in 1999 by A.K. McCallum and M. Nigam of Carnegie Mellon University. CORA adopts the method of mechanical cognition. Its main target is the content associated with the subject of the computer. The content of the user's needs is classified by the principle of implicit Malf. Although the ability of CORA of analyzing addresses and topics is still inadequate, and it does not have the ability to analyze web pages, it still made a significant achievements in automatic collecting resources.

S. Chakrabarti proposed the IBM Focused Crawler at the beginning of this century, which is a brand-new crawling system. From the point of view of current technology, IBM Focused Crawler adopted a new dual-module system which are classifiers and selectors. The classifier is mainly used to calculate the relevance, and the selector is used to determine the main pages. In the subsequent development process, S. Chakrabarti further improved the entire system, resulting in a significant increase in the accuracy and relevance of the system[1].

The American Diligenti designed focused crawler in a way that creates contextual maps. They named them Context Graphs Focused Crawler. This method of learning citations from web pages was later proved to be inefficient, but it was also an important innovation at that time. The system will use the backlink service to find the webpage that points to the page,

establish a reference relationship between the two webpages, thereby establishing a crawl path that focuses on the crawler. Through the change of user search parameters, each page will establish a corresponding reference relationship. In this process, the classifier will determine their hierarchical relationship. After the determination, the page's link will be added to the queue, which will extract all the web pages that need to be crawled[2].

1.2.2 Chinese Word Segmentation

Depending on whether machine-readable lexicons and statistical information are used or not, automatic Chinese word segmentation methods can be classified into three categories: dictionary-based methods, statistics-based methods, and hybrid methods.

The three elements of the dictionary-based word segmentation method are word segmentation dictionary, text scanning order and matching rules. The scanning sequence of the text has forward order, reverse order and bidirectional scanning order. The forward scanning refers to scan from the beginning of the segmentation statement, and the reverse scanning refers to scan from the end of the segmentation statement. The bidirectional scan is a combination of forward scanning and reverse scanning. The matching rules mainly include maximum matching, minimum matching, word-by-word matching and best matching.

The statistical models applied by the statistical word segmentation method are: mutual information, N-gram grammar model, neural network model, hidden Markov model and maximum entropy model. These statistical models mainly use the joint occurrence probability of words and words as part of word segmentation information. The advantages of the statistical word segmentation method are not limited by the field of the text to be processed and not requiring a machine-readable dictionary. The disadvantages is that a large amount of training text is required to establish the parameters of the model and the accuracy of word segmentation is related to the choice of training texts.

1.2.3 TF-IDF

In the field of text classification, the dominant text representation is the space vector model. To represent text with a space vector model, we must first segment the text, then perform feature selection and weight calculation, and finally form an N-dimensional space vector. There are many different ways to calculate the weights. The weights of the feature items will affect the overall performance of the text classification algorithm. Among them,

TF-IDF has been favored by related researchers and many application fields because it is relatively simple and has high accuracy and recall rate. Since the concept of IDF has been proposed, the TF-IDF algorithm has undergone many improvements. These improvements are made in order to adapt to different text classification fields and different applications.

The concept of IDF (Inverse Document Frequency) was first proposed in [3], pointing out that in a set of documents, feature items (words) characterizing a document can be assigned weights according to the frequency of appearance in the group of documents. The more specific words that appear in only a few documents, the more important than the weight of the words appearing in multiple documents. Shannon's information theory explains the meaning of IDF for us: if the feature item appears more frequently in all documents, it contains less information entropy; if the feature items appear more concentrated, only in a few documents with a higher frequency of occurrence, it has a higher information entropy. Therefore, IDF can be understood as the cross-entropy of the probability distribution of keywords under a specific condition.

Salton proposed the TF-IDF algorithm in [4]. Since then, Salton has repeatedly demonstrated the validity of the TF-IDF formula in information retrieval [5]. In 1988, he elaborated on the application of multi-word weight calculation methods in document retrieval [6]. TF-IDF mainly embodies the following idea: The higher the frequency of occurrence of a word in a particular document, the stronger its ability to distinguish the content of the document (TF); the wider the scope of a word appears in the document, It distinguishes the document content with the lower ability(IDF).

In the 1990s, both domestic and foreign researchers began to pay attention to the application of TF-IDF in text categorization. Many scholars analyzed the defects of TF-IDF, improved it, and verified the effectiveness of the improvement through experiments. TF- IDF was also expanded with the development of research and has been applied to multiple new areas.

1.2.4 K-means Clustering

Clustering is an important technology in data mining and information retrieval. It can effectively analyze the data and find useful information from a large number of data. Clustering will devide data objects into several classes or clusters, so that there is a high

degree of similarity between objects in the same cluster, and objects in different clusters will vary greatly. Through clustering, people can both identify dense and sparse regions and find interesting relationships between global distribution patterns.

Cluster analysis divides large amounts of data into sub-categories of the same feature to facilitate the understanding of the distribution of data. Unlike other data mining methods, users generally do not know the characteristics of the data set before performing cluster analysis. Therefore, from a certain perspective, cluster analysis is an unsupervised learning process that is based on observation rather than case-based learning.

As a branch of statistics, cluster analysis has been widely studied for many years, mainly focusing on distance-based cluster analysis. What's more, the k-means clustering information tool has been added to many statistical analysis software packages suck as scikit-learn.

In the field of machine learning, clustering is an example of unsupervised learning. Unlike classification, clustering and unsupervised learning do not rely on predefined classes and training instances with class labels. Clustering is observational learning, not example learning due to this reason. In conceptual clustering, a group of objects can be formed as a cluster only when they can be described by a certain concept. This is different from the traditional clustering based on similarity of geometric distance metrics.

In order to improve the performance of the clustering method, the methods in other fields are combined with the clustering method to make up for some of the deficiencies in the clustering method in the data mining field. In this way, the goal of realize theoptimal performance of the clustering method will be achieved. The well-known methods often used are: genetic algorithm, immune algorithm, ant algorithm and so on.

At present, the research direction of clustering algorithm mainly consists of the following directions: First, the choice of initial value and influence of the input order. The measures that can be taken in the field of data mining can use multiple sets of different initial values and perform multiple iterations. The best one is selected as the calculation result at end of the process. However, it cannot be guaranteed that the global optimum solution will definitely be achieved. The essence of the optimal solution clustering process is an optimization process. Through an iterative operation, the object function of the system will work out an optimal solution. However, this object function is a non-convex function in the state space. It has many minimum values, of which only one is the global minimum value, and the others are

local minimum values. The goal of optimization is to achieve the goal of deriving global optimization. Therefore, the optimization of a non-convex function is a research topic to be solved. What's more, the efficiency of the algorithm. To improve the efficiency of the algorithm is also an important issue in the field of clustering. By improving the existing clustering algorithm to increment the ability of clustering and improve the scalability of the algorithm.

K-means clustering algorithm is a widely used partitioning method in cluster analysis, which is simple and rapid. However, the K-means clustering algorithm is sensitive to the initial value. This means different initial values will often lead to different clustering results. Instead of deriving a global optimal result, it will derive a local optimal result sometimes.

1.3 Research Content

At present, in the field of information processing, various technologies have made considerable progress. This helped us to achieve this task. At this stage, there have also been many work related to the analysis of breaking news. But in general, these related tasks are generally focused on only one aspect of the whole process. This article is based on this point, effectively using the existing Web-related technologies, such as web crawling, words cutting, TF-IDF and K-means clustering. By combining the four steps naturally, I implement the analysis and mining of the breaking news.

1.4 Paper Structure

Chapter 1 is an introduction. This chapter briefly introduces the backgrounds of Web crawling, words segmentation, TF-IDF algorithm and K-means clustering. What's more, this chapter also indicates the goal and meaning of these technologies.

Chapter 2 is related work. This chapter introduces the related knowledge and rules of web crawling under different coding set, principles of words cutting, TF-IDF and K-means. In this chapter, some possible defects of the algorithms has also been proposed.

Chapter 3 is the analysis and design of using web crawlers to obtain the data set from breaking news, the words cutting of the extracted data set, the TF-IDF to process the cut data set and to perform K-means clustering on matrix generated by TF-IDF.

Chapter 4 is the implementation of web crawling of baidu breaking news, the words

cutting of crawed baidu news, TF-IDF performed on the cut data and K-means clustering on the matrix which TF-IDF generated. I give proof of the correctness of the algorithm, and analyze the performance of the algorithm theoretically, including the time consumed by the algorithm and the complexity of the algorithm. At the same time, we combine the characteristics of different algorithms to improve the accuracy of the classification.

Chapter 5 concludes with the conclusions and future work of this article.

Chapter 2 Related Work and Theoretical Basis

2.1 Web Crawling

Web crawling refers to obtaining web page information required by a user. According to the content of the information that needs to be obtained, it can be roughly divided into two types, general web page crawlers and customizable web crawlers. For the second one, we only need some specific web page information. This requires to establish some filtering devices.
After understanding the basic concepts and classifications of the web page crawlers, we should also understand the basic flow of web page access technology and some important parts when we crawl data. This flow diagram of web crawler is as Figure 2.1:

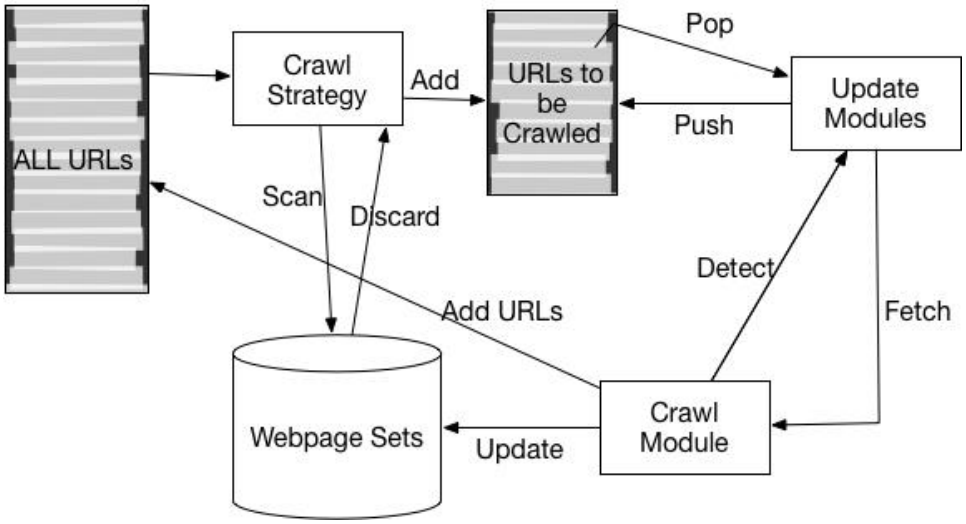

Figure 2.1 Web crawler flow diagram

Before we crawl the webpage, we need a batch of known URLs. These URLs are called seeds in the web crawling process. These seeds are generally sites that have been selected by the user which have high-quality. When the seed URL is selected, we add it to the list of URLs to visit, so that each time we pop up the next page to visit from the list of URLs to visit. We then visit the corresponding page and add the URL to the URL list we have already visited. This queue is used to determine if a page has been accessed in order to avoid repeated visits. Next, all URLs contained in the web page are parsed from the web page using page parsing technology, and the URLs of the non-visited web pages will be added to the non-visited URL queue by depth-first and width-first methods.

At the same time, we will also store the information we need for the next step. In general, the web crawler will continue to repeat the above procedure until the list of URLs to be

visited is empty. However, in actual situations, this kind of stop condition is difficult to achieve. Therefore, according to the disk capacity, the crawl depth or crawl time is used as a web crawling stop condition. The whole process described above can be described as a spider crawling on a spider web. Therefore, web crawlers are also called web crawlers or web spiders.

In general, the actual web crawling process is much more complicated than the description mentioned above. In order to ensure the speed and quality of web crawling, multiple threads are used to accomplish a specific step in the crawling process. In order to ensure the quality of web crawling, we need to use limited resources to crawl important web pages first. This involves the issue of the web page fetching order selection strategy. In addition to the above two issues, the strategy of web crawling is also a very important issue in web crawling. Regarding web page renewal strategies, Junghoo also have detailed descriptions in his research[7].

2.2 Words Segmentation

After extracting the information in the webpage, another important part of the Chinese webpage is the words segmentation of the webpage. Words segmentation technology is the key to the semantic analysis of web page information. Chinese word segmentation technology is currently a relatively mature technology. According to different methods, it can be roughly divided into three categories: segmentation method based on string matching, segmentation method based on understanding, and segmentation based on statistics.

The so-called string matching method is based on a certain strategy to match the Chinese character string with a dictionary. If you find a string in the dictionary, the match is successful. The comprehension-based method refers to the effect of recognizing a word by letting a computer simulate a person's understanding of a sentence. The basic idea is to perform syntactic and semantic analysis at the same time when it did the job of word segmentation. And to handle ambiguity using syntactic information and semantic information. What’s more, the basic statistics method is based on the co-occurrence of words as the basis for word segmentation. Because words are a combination of stable characters, the more often the adjacent characters appear in the context, the more likely it is to form a word. Therefore, the frequency or probability of co-occurrence of characters and characters can reflect the

credibility of words. The frequency of the combinations of adjacent co-occurring words in the corpus may be counted to calculate their co-occurrence information. Mutual information reflects the closeness of the relationship between Chinese characters. When the level of closeness is high, characters can be considered as a word. The above three methods are popular among the years. For any mature word segmentation system, it is impossible to rely on a single algorithm to implement it, and it is necessary to synthesize different algorithms together

2.3 Data Mining.

Data mining is a powerful technology that helps people find the most important information in data set. Data mining tools can predict future behaviors and make knowledge-driven decisions. The automated prospective analysis provided by data mining has gone far beyond the retrospective analysis of past events. Data mining tools can answer questions that traditionally require a lot of time to answer. It can search through the hidden patterns, categories and some rules information in a large mass of data. Data mining is essentially the discovery process of discovering the relationship between data essence and data, identifying the trends and trends that are potential in the data, which could guide us to understand things and help me to make better decision. Data mining can be roughly divided into three categories: relationship discovery, pattern discovery, and trend behavior discovery. In the following studies, the relevant technologies we need to use for data mining are mainly found by using the relationship between thousands of objects. Specifically, we hope to analyze the various news information objects and find the category relations between them through certain methods, so as to divide and gather similar information into the same category. Here we introduce the techniques of text classification and text clustering.

2.3.1 Text Classification

Text classification technology is based on some basic characteristics of the observed data to establish a prediction function for the target value, so as to classify the new unknown data instance. Text classification is a two-step process. The first step is to create a model that describes the predefined set of data types. It is constructed by the characteristics of a predetermined set of data classes. This predetermined set of data classes is also called a training set. The second step is to use the model obtained in the first step to classify, and the

obtained model classification function is applied to the characteristics of the new data instance, and the new data can be divided into corresponding predefined data categories.

Classification algorithm is a mature branch of data mining technology. Up to now, there are many classification algorithms and they have been widely used. Among the common classification algorithms are Naive Bayes, Bayes Network, K-Nearest Neighbor, Decision Tree, Support Vector Machine and Neural Network.

2.3.2 Text Clustering

Text clustering has similar goals as text classifications. That is, a data set is divided into different categories according to the relationship between samples. The biggest difference between them is that the categories of thousands of clusters are not known in advance. We may not know how many categories we need to divide, nor do we know the specific meaning of the clustered categories. Cluster analysis requires that the data objects in the same class have higher similarity in the class into which the data objects are divided, and the differences in the different classes are as large as possible. Therefore, when there comes text clustering, the main problems that need to be faced are: the way of data division, how the distance between data individuals is defined, and how the data category labels are generated.

Cluster analysis algorithms can be roughly divided into the following categories: partition method, hierarchical method, density-based method, grid-based method and model-based method. In my work, I used the partition method to cluster. The partition method refers to a data set with N tuples or records. The partition method will construct K groups, each of which represents a cluster. K<N. Furthermore, the K group satisfies the following conditions: (1) Each group contains at least one data record (2) Each data record belongs to and belongs to only one group. For a given K, the algorithm first gives an initial grouping method. Afterwards, the grouping is changed through repeated iterative methods, so that the grouping scheme after each improvement is better than the previous one. A good criterion is that the closer the records in the same group are better, the further the records in different groups are better. Algorithms using this basic idea are: K-MEANS algorithm, K-MEDOIDS algorithm, CLARANS algorithm. The partition method is much more convenient to implement, and it is also very accurate. Therefore, this type of method has a more extensive use.

Chapter 3 Algorithm Analysis and Design

3.1 Application Background and System Framework

Before introducing the specific workflow of the system, it is necessary to elaborate on the issues which we want to study and explain some basic concepts. At the same time, I will explain the methods in my work in detail and finally introduce the framework of the system.

3.1.1 Application Background

News information: News information refers to the web information with certain timeliness. It is generally a description of the specific time of action that takes place at a specific time and place.

Breaking news information: For a certain piece of news information *I*, within a time interval *T*, the number *N* of news information about this piece of news information exceeds a certain fixed threshold θ, and we will call this piece of news information as breaking news information.

Breaking news information mining: Breaking news information mining is the use of automated text mining analysis methods. It is a process to discover all aspects of the network of breaking news information in a timely and accurate manner.

The main problems of this process are: how to obtain breaking news information; how to obtain the detail of each breaking news item; how to better analyze the news information effectively and clustering these pieces of breaking news accurately. What's more, I will give a detailed description of every step.

Firstly, considering that there are various forms of existence and expressions of news information in the internet, it is impossible to obtain all types of news information. For the sake of research convenience, the text information used in this article are Chinese breaking news at the real time. The system uses web crawlers to crawl every breaking news item on Baidu Breaking News and its corresponding websites. In this case, we could obtain the original data sets. Admittedly, there are two main parts that are difficult to deal with in this section: HTML page parsing and the way to get the body content of a web page. In this article, I used some existing condition and made some improvements to complete this part of work.

In the end, after obtaining these basic information, we need to do some research and

analysis to process the data set. Through these processes, the data sets can be better reorganized and merged, which could enable users to track the detail of every breaking news on Baidu Breaking News Page. The tasks that need to be completed in this section are: adopting some methods of text mining and processing for the whole obtained news information and clustering news information. Eventually, we could check the result of clustering and its accuracy.

The information in the internet is endless and varied. In this case, the first step is to get the breaking news information in the targeted webpage. While avoiding the waste of resources, this article has mainly adopted a customizable web crawling strategy. In general, we believe that the web pages for large sites and high-visibility websites has relatively high quality, which can reflect the direction of public opinion and can become a piece of breaking news in the internet. In this case, we choose Baidu breaking news as our target crawling webpage. Using the boundary delimitation, the breaking news page is obtained. In this way, the maximum benefit of limited resources can be better utilized, and powerful data support can be provided for reducing the noise data in the subsequent web page analysis phase.

The ways to obtain key information in a web page is an important issue in this article. The most difficult part is how to automatically obtain the body part of each web page. In general, text extraction of web pages is more difficult. Here, we use the basic template to perform semi-automatic extraction. It is a compromise between accuracy and flexibility. This requires a pre-determined algorithm to analyze each web page from a different site to obtain a text extraction template for each site, and then use the resulting template to extract new web pages. This also requires to update the template and it also needs to refine the granularity of the template as much as possible. The finer the granularity has, the more accurate the resulting extraction results will be.

The last problem has to be solved is to extract breaking news from a large number of specific web page data. To solve this problem, we must figure out the content of each piece of news and number of pages have the same topic. In fact, these two problems are both web page classification problems and web page clustering problems. To understand the topic of each piece of news, we have to know the type of news first, such as sports news, entertainment news, etc. Only after knowing the category the news belongs to can you further analyze its content. The number of pages having the same topic is actually a clustering problem. With

clustering, we can distinguish web pages in the same category according to the topics they belong to.

3.1.2 System Framework

In this section, we will focus on the framework of the entire system. Figure 3.1 shows the main modules of the system which this article will implement. From Figure 3.1, we can see the operation flow of the entire system and the interaction process between the various system modules. It can be seen that during the entire system operation, all major modules of the system are coupled through data. The advantage of this design is that the division of responsibilities of each function module is relatively clear, and each module only depends on the output data of the previous module. In this way, it will be convenient for us to parallelize the original work flow. This type of architecture design can greatly improve the efficiency of data processing.

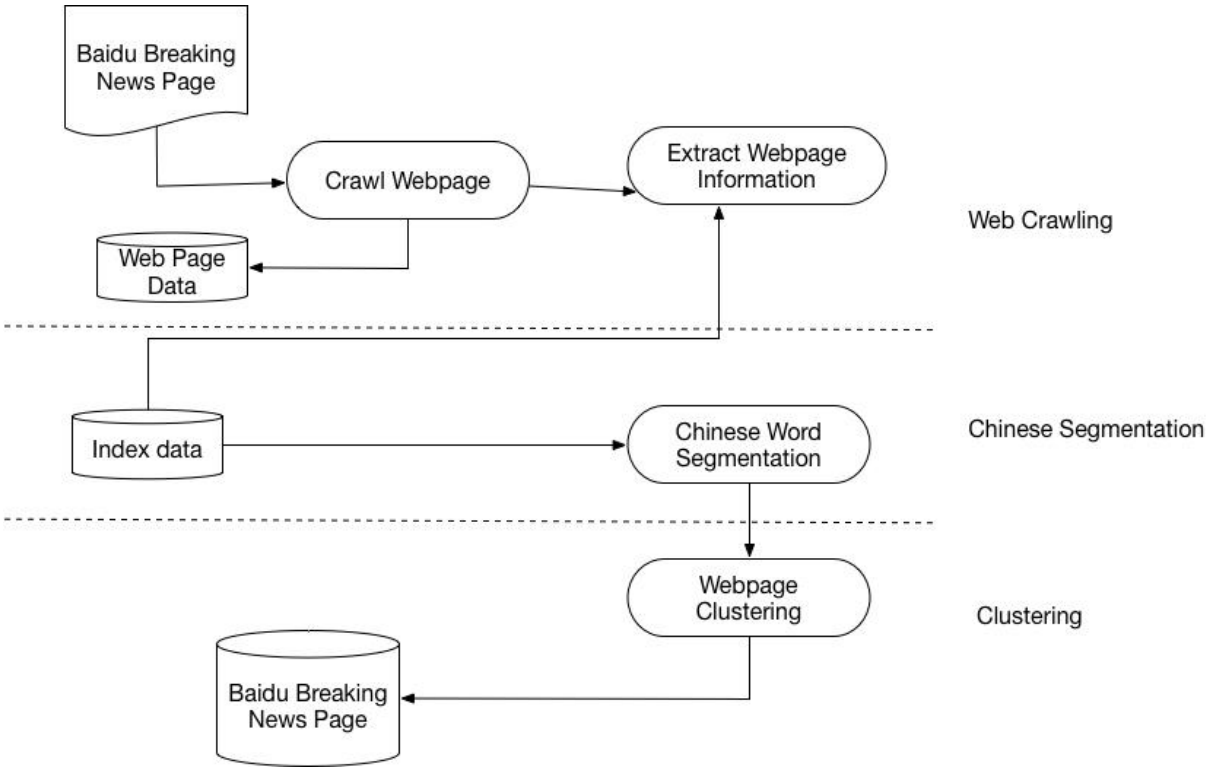

Figure 3.1 System framework

Next, the flow of the entire system will be described in detail in accordance with Figure 3.1. First, the system needs to obtain a large amount of information resources，which is the news page. In order to ensure the quantity and quality of the acquired pages and the timeliness of the news pages, the system will use a customizable web crawler module to start the depth-first crawl from the selected Baidu breaking news site. During the process of crawling, some specific filtering conditions are set. These filtering conditions can be set according to the requirements, so as to ensure that we can use limited resources to obtain good quality pages. As we can see, the crawling module is the basis of this system. The main problem should be solved in this module is to determine the correct web page crawling strategy, which is to use a limited resource to maximize the quality of the crawled web page collection. Another problem is to determine the update strategy for web pages. Only by solving these two problems can we ensure that the work we have done is not useless. At the same time, it can also reduce the interference in the accuracy of information analysis in the following process.

After getting a collection of news pages, there are two main steps to be done. The first step is to store the original webpages which are crawled. The webpages are stored using a database. The has two purposes: On the one hand, it facilitates the collection of information and can provide historical information to facilitate the analysis of news information. On the other hand, the stored original web page can provide a snapshot of the webpage. This could solve the problem which some webpages may not be accessible for some reason. What's more, users can access snapshots when resources in the network are inaccessible with storage backup. The second step is to use the web page information extraction module to remove noise and structure information from web pages. In this step, the noise information in the web page needs to be removed and get the useful information in the original web page structured. The removal of webpages noise mainly refers to the removal of some information on the webpage that is of little value to the user, such as advertisement information on the top and right side of the webpage, registration and copyright information at the bottom of the webpage, etc. Due to the heterogeneous feature of web pages in the web, it is not easy to handle this task. In general, there are template-based methods and learning-based methods to solve this problem. In the system, a learning-based method is used to de-noise the webpage. The algorithm for de-noising is described in detail in the next section. After removing the noise information, it is necessary to make the useful information structured and extract information

from some key fields, such as the title of the web page, link information and the text of the web page. After these process, it is necessary to perform segmentation of the web page.

Now, we have completed the extraction and structuring of web pages. The next step is the core module of the system, which analyzes and mines information of web pages. This part is done through the clustering of web pages. In order to accomplish this task better, it is necessary to implement the function of data reorganization. This reorganization is to establish an inverted index on the web page and build an index, mainly for the sake of subsequent calculation of the similarity between documents. In the process of clustering web pages, one of the main metrics is to calculate the similarity between web pages. The calculation of similarity is based on the inner product of word vectors. The advantage of building an inverted index is that by querying the inverted index table, you can clearly know that only the document list contained in the index list needs to be calculated between each other. This can greatly reduce some invalid calculations, thus ensuring the efficiency of classification and clustering algorithms. The indexing process first standardizes the structured web page information in a specific format, and these documents are then passed to the index building component for indexing. The created index is stored in a file. In the subsequent page clustering module, an index file is used to contribute to the calculation of the similarity between documents. At the same time, structured web page information is stored in the web page information database.

After this, we have to face the web page information mining and analysis module, which is also the core module of the system. This is mainly divided into two steps. First, the segmented text will be processed by TF-IDF algorithm, which could generate a two-dimensional matrix. Then the clustering analysis of web pages is used to cluster together some similar or similar web pages. The clustering of webpages uses the K-Means algorithm. I have also done some optimization on the implementation of K-means to make it to serve the system better. The clustering categories for web pages are stored as csv. At this point, all data processing has been completed. It started from the seed nodes of some news sites and was processed by a series of data processing modules. Finally, the clustering results were worked out.

3.2 Algorithm Definition

3.2.1 Modules

(1) Web Crawling Module

The crawling of news web pages is the first step in obtaining news information. It is the basis for obtaining data. Only by obtaining enough news information can we better complete the mining and analysis of later news information. Considering limited resources, we have to ensure the quality and speed of news web crawling. In the system, only Baidu breaking news webpage and its corresponding webpages are crawled. In this way, noise in the data can be effectively reduced, which could guarantee the quantity and quality. At the same time, it can avoid excessive repetition of web pages in the data and this could save resources.

The main task of the web crawling module is to crawl the news webpage of the corresponding webpage of the portal website and provide data for analyzing breaking news. Since the captured webpage serves the next series of analysis work, the webpage capture has some features that are not the same as general web crawling. The web page crawling module requires that web pages that meet certain conditions be captured from specific web sites on a regular basis. In this case, the crawling process has more customization features and is a custom task crawler. In this module, it is mainly to crawl those specified webpages and store these web pages in a file for later use.

(2) Word Segmentation Module

After the information extraction process, the next step is to process the web page and create an index for storage. The purpose of the follow-up news page for dicing the word is needed for mining and analysis. Only by decomposing the web page into a collection of term can we analyze the similarity relationship between web page objects. Subsequent analysis is based on the web page as a Term vector. Then the word processing means that the main body of the webpage and the headline should be used by punctuation, English letters and spaces to break sentences first. Then use a word segmentation system for each sentence and use the interface provided by it to cut sentences into words and output them.

(3) TF-IDF Module

TF-IDF, short for term frequency–inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus[8]. It is often used as a weighting factor in searches of information retrieval, text mining, and user modeling. The TF-IDF value increases proportionally to the number of times a word appears in the document and is offset by the frequency of the word in the corpus, which helps to adjust

for the fact that some words appear more frequently in general.

(4) K-means Module

K-means clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. k-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells.

The problem is computationally difficult (NP-hard). However, there are efficient heuristic algorithms that are commonly employed and converge quickly to a local optimum. These are usually similar to the expectation-maximization algorithm for mixtures of Gaussian distributions via an iterative refinement approach employed by both k-means and Gaussian mixture modeling. Additionally, they both use cluster centers to model the data; however, k-means clustering tends to find clusters of comparable spatial extent, while the expectation-maximization mechanism allows clusters to have different shapes.

3.2.2 Problem Definition

(1) Web crawling

The link and these two importance values are used to evaluate the importance of an element node. The purpose of calculating the degree of presentational style is to detect noise by the number of presentation styles. At the same time, the purpose of calculating the importance of content is to identify the main content of these web pages presented in a similar style of presentation. Finally, the importance of the element nodes is given by combining the importance of the presentation and the importance of the content. The more important the combined importance of an element node, the more likely it is to be the main web content.

In Table 1, T_n is the weight of *n*-th node. *n* is the depth of node tree. δ is the weight factor.

Table 1 Node Weight Table

Node Name	Weight	Rules
UL	ξ_{ul}	V_2
LI	ξ_{li}	V_3
DIV	ξ_{div}	V_1
H	ξ_w	V_4

SPAN	ξ_{other}	V_n
IMG	ξ_{other}	V_n
P	ξ_w	V_4
DD	ξ_{ul}	V_2
P;	ξ_{li}	V_3
DL	ξ_{ul}	V_2

In rulemaking, the smaller the weight level subscript is, the higher the weight is. The table shows that the div represents the highest weight, indicating that its child nodes may form a list cluster. A list cluster can form a list of node sets through fixed assembly nodes. The set forms a list area. The weight of the list cluster can be defined as $(\xi_{ul} + n\xi_{li})$。

In the HTML language, DIV represents a position block that is strictly set and can clearly show the sub-node's display form and block under the DIV. In the HTML language, the UL tag defines an unordered list. In the sample model, the lists are combined by the UL and LI tags which could present the list. Therefore, the tag clusters that match the list combination are extracted and used as the candidate list module. In the path from the root node to the leaf node, the influence of the style tree of the webpage is reduced. In the tree edit distance algorithm, the weights of the nodes and the depths of the nodes are reduced by increasing according to the Formula 3.1.

$$T_n = \frac{1}{(1+\delta)^n} \tag{3.1}$$

Since the influence of the web page style tree on the layout of the web page decreases as the depth of the tree layer increases, a decrement factor α is set, where $\alpha < 1$. The div weight value of each node is $\xi_{div} * \alpha^i$, where i is the number of tree layers. The weighting method for calculating a DIV node is T_n according to Formula 3.2.

$$T_n = \sum_{i=0}^{n} \alpha^i \left(\beta\left(\xi_{u_i} + i\xi_{li}\right) + \xi_{other} + \sum_i^n \xi_{div}\right) \tag{3.2}$$

β is the influence factor of the list cluster under the div element and could control the weight calculation of the div list cluster. The ξ_{other} is the weight contribution of other nodes to the div node. $\sum_i^n \xi_{div}$. is the weight contribution of the child div node to its parent node. The higher the weight of the div node is, the higher the complexity of the div tree is. The process of recursively traversing from the leaf node to the root node div determines the list frame candidates based on the complexity of the div. Through the calculation of the node style tree, the correlation between tree nodes can be highlighted. Based on the complexity of the

child nodes, the weights can be contributed from the bottom of the tree to the root layer of the tree. This could effectively extract the candidate of the block node model.

(2) Word Segmentation Module

From a statistical point of view, the input of the word segmentation problem is a string *C=C1, C2… Cn*. The output is a word string *S=W1, W2…Wm*, where *m<=n*. For a string C, there will be multiple segmentation schemes *S*. The task of segmentation is to find the most probable segmentation scheme among these *S*, that is, to segment the input string into the most likely words sequence (Formula 3.3).

$$Seg(c)_{s\in G}^{argmaxP(S|C)} = \underset{s\in G}{argmax}\frac{P(C|S)P(S)}{P(C)} \tag{3.3}$$

The task of the word segmentation of the probabilistic language model is to find a segmentation scheme *S* in all the results obtained by the full segmentation so that the *P(S)* is maximized. So how do we express *P(S)*? For ease of implementation, we assume that the probability between each word is context-independent, we could derive Formula 3.4.

$$P(S) = P(W_1, W_2, \dots, W_m) \approx P(W_1) \times \dots \times P(W_n) \propto log(W_1) + \cdots + log(W_m) \tag{3.4}$$

Among them, for different *S*, the value of *m* is not the same. In general, the larger the *m* is, the smaller *P(S)* will be. In other words, the more words we split, the smaller the probability it is. This is in line with practical observations, such as the maximum length matching segmentation tends to make *m* smaller. The probability of calculating any word appears as Formula 3.5.

$$log(W_i) = log(Freq_w) - log\, N \tag{3.5}$$

From another point of view, the maximum probability of calculation is equal to the shortest path of the segmentation graph. However, the Dijkstra algorithm cannot be used here, and the shortest path is solved by dynamic programming.

In order to find the most probable word string as soon as possible, we assume that the probability between each word is context-independent. This could satisfy the optimal sub-structure properties and non-post-effects required for solving with dynamic programming. In the process of solving the dynamic programming, instead of creating all the possible segmentation paths Si, the maximum value of *P(Si)* is obtained and *Si* is directly output by the backtracking method. The maximum probability up to node *Node i* is called the probability of node *Node i*. If the ending node of *Wj* is *Node i*, it is said that *Wj* is the precursor of *Node i*. Here *prev(Node i)* is the set of predicates for node *i*. Therefore, the maximum probability of

segmentation max*(P(S))* is *P(Node m)=P(the best precursor of node m) P(the best precursor of node m).*

(3) TF-IDF Module

①Vector Space model

The selection and extraction of feature items is the basis of the entire Web document clustering system. It maps the system from the concept space to the computable space, thus making the entire system possible. The basic concept of vector space model can be described as follows：

Document: refers to a general text or a fragment of a text (paragraph, sentence, or sentence), usually an article. Although the document can be a multimedia object, it is assumed to be a text object in our discussion and no distinction is made between text and documents.

Item (Feature Item): The content of the text is expressed by some feature items, which are generally represented by the basic language units (words, words, phrases or phrases, etc.) In other words, a vector space is formed by these items, each item representing one dimension.

The weight of items: In the text, each feature item is given a weight *W* to indicate the importance of this feature item in the text. Weights are generally calculated based on the frequency of feature items.

Similarity measure: The degree of correlation between two texts d_1 and d_2 is often measured by their similarity $Sim(d_1, d_2)$. In the vector space model, we can use some distance between vectors to represent the similarity between texts.

Vector Space Model (VSM): Given a natural language text, there are still some difficulties in analysis due to the repetitive and sequential relationships in the text. In order to simplify the analysis, the order of precedence in the text may be taken into account and the requirements may be different. After discarding the sequence information between each feature item, a text is represented as a vector, which is a point in the feature item space and a text set can be represented as a matrix, that is, some of the feature item space.

②Selection of Feature Items

Text mining feature attributes are flexible and changeable, which is different from data mining using fixed attribute features. Abstract concepts are hard to express and hard to formalize, and text features are often high-dimensional. On the other hand, many of the

information in the document is highly redundant, so the extraction of text features is very important, which often determines the efficiency of text mining.

The selection of terms *t* in the target representation is called feature extraction. There are two main types of methods: independent evaluation methods and comprehensive evaluation methods. The basic idea of the former is to perform independent evaluation of each feature in the feature set, so that each feature gets a weight and then sorts by weight size. The number of weights or predetermined features selects the best feature subset as the result of the feature extraction. The latter is to find fewer comprehensive indicators that describe these features from the high-dimensional set of original features. These integrated indicators are independent of each other. Then use the resulting composite indicators to characterize the feature set.

The advantages and disadvantages of the expression of vector space model (VSM) directly depend on the selection of feature items and the calculation of weights. Here we briefly introduce some definitions:

Word features: Word features have a better ability to represent text and vocabulary can express semantic information more completely. However, not all words are suitable for feature items. Studies have shown that high-frequency words and low-frequency words represent texts that are smaller than mid-frequency words. Because high-frequency words have similarly high frequencies in all articles, ow-frequency words appear less often in texts. Intermediate-frequency words and texts are more related to themes and represent the strongest.

Word features: The ability of words to represent text is relatively poor compared to word features and cannot completely express semantic information independently. The feature extraction process using word features is relatively simple. Since the number of commonly used Chinese characters is small, the time and space overhead of the extraction process will not be too great.

③Weight Calculation Module

After the feature extraction process, the extracted text is used to represent the text as the dimension of the vector. The initial vector representation is completely in the form of 0, 1. If the word appears in the text, then the dimension of the text vector is 1. Otherwise it will be 0. This method cannot reflect the role of the word in the text, so gradually 0, 1 is replaced with a

more precise weight.

The vector $d = (w_1, w_2, \dots, w_m)$ represents the characteristic terms and corresponding weights of the document *d*. *m* is the number of entries in the document set, and $w_i (i = 1, \dots, m)$ represents the weight of entry t_i in document *d*. The calculation of feature weights w_i is usually based on the classic TF-IDF algorithm and is standardized as Formula 3.6.

$$w_i = -\frac{(TF \times log_2(d_{DFi}^N))}{\sqrt{\sum_{i=1}^{m} [TF \times log_2(d_{DFi}^N)]^2}} \tag{3.6}$$

TF represents the term t_i frequency in the document *d*, DF_i represents the number of documents in the document set containing the term t_i and N represents the number of documents in the document set.

(4) K-means module

Clustering belongs to unsupervised learning. Previous regressions, naive Bayes and SVM all have class labels y, that is, the examples have been given examples of classification. However, in the clustering sample, there is no given y, only the feature x. Assuming that the stars in the universe can be represented as a set of points in a three-dimensional space. The purpose of clustering is to find the potential category y for each sample x and put the samples x of the same category y together. For example, for the above stars, after clustering, the results are clusters of stars. The points in the cluster are relatively close to each other, and the distance between stars in the cluster is relatively long.

K-means algorithm clusters samples into k clusters. Given a set of observations $\{x_1, x_2, \dots, x_n)$, where each observation is a *d*-dimensional real vector, *k*-means clustering aims to partition the *n* observations into *k* sets $S = \{S_1, S_2, \dots, S_k\}$ so as to minimize the within-cluster sum of squares. Formally, the objective is to find Formula 3.7.

$$arg\,min \sum_{i=1}^{k} \sum_{x \in S_i} \|x - \mu_i\|^2 = arg\,min \sum_{i=1}^{k} |s_i| v_{ar} s_i \tag{3.7}$$

Where μ_i is the mean of points in S_i. This is equivalent to minimizing the pairwise squared deviations of points in the same cluster (Formula 3.8).

$$arg\,min \sum_{i=1}^{k} \frac{1}{2|s_i|} \sum_{x,y \in S_i} \|x - y\|^2 \tag{3.8}$$

The Equivalence can be deduced from identity. Because the total variance is constant, this is also equivalent to maximizing the squared deviations between points in different clusters

3.2.3 Problem Explanations

(1) Web Crawling Module

In order to ensure the timeliness of the news, the crawler module needs to go to the Baidu breaking news webpage every day to crawl a new page. This requires the crawler to automatically go to the specified address to crawl the corresponding page. The program's timing execution function can be implemented using the system's mission plan. Such a strategy can balance the number of web pages crawled and reduce the pressure on the target web site.

Directed crawling is also an important process. Since we only need some specific web pages, we need to make judgments on the web pages we crawled. After researching some portal website news webpages, we know that the webpage's links have certain characteristics, so the judgment of webpages is achieved through the regular expression matching of URLs. In the specific implementation, it is mainly based on the current Baidu Breaking news webpage to crawl its respective news pages. In order to ensure that the capture of news webpages is not too separated, crawlers will mainly crawl those specific webpages corresponding to the Baidu breaking news site. This will not only ensure the relative concentration of news pages, but also guarantee the quality of the web pages. In addition, in order to capture more and better web pages on the premise of a certain time and space, the home page is used as a seed, and web pages that do not exceed 2 levels are generally processed.

Since the news web pages need to be clustered by timeline, the webpage time must be recorded. Since the links of the web pages of the portal site include the order of the web pages, the numbers are extracted from the connections before being stored, and then stored in the files according to the sequential order of the web pages. This step stores the original web page for the convenience of processing information in the next module.

The link extraction module is an important module in the crawler strategy. It provides the crawler with a seed link to the page depth of the Tong countryside layer. Due to the need to match the template, in this step you need to first extract the links in all web pages as a sample set.

(2) Word Segmentation Module

In the Chinese word segmentation subsystem, the N-shortest path rough cuts produce *N*

best coarse split results quickly and the split result set can cover ambiguity cuts as much as possible. Unregistered words refer to words that are not included in the dictionary, such as names of people, place names and names of institutions. The class-based Segmentation is a segmentation after the identification of unregistered words. In this segmentation, unlisted words participate in the competition just as ordinary words. Binary segmentation graph is a key intermediate data structure. It plays a pivotal role in the identification, disambiguation and segmentation of unlisted words.

(3) TF-IDF module

Suppose you want to summarize a document or a paragraph using few keywords. One technique is to pick the most frequently occurring terms (words with high term frequency or tf). However, the most frequent word is a less useful metric since some words like 'this', 'a' occur very frequently across all documents. Hence, we also want a measure of how unique a word is i.e. how infrequently the word occurs across all documents (inverse document frequency or idf). Hence, the product of tf times idf (tfidf) of a word gives a product of how frequent this word is in the document multiplied by how unique the word is w.r.t. the entire corpus of documents. Words in the document with a high tfidf score occur frequently in the document and provide the most information about that specific document.

(4) K-means module

The k-means clustering algorithm attempts to split a given anonymous data set (a set containing no information as to class identity) into a fixed number (k) of clusters. Initially k number of so called centroids are chosen. A centroid is a data point (imaginary or real) at the center of a cluster. In Praat each centroid is an existing data point in the given input data set, picked at random, such that all centroids are unique (that is, for all centroids ci and cj, ci $\neq$ cj). These centroids are used to train a kNN classifier. The resulting classifier is used to classify (using k = 1) the data and thereby produce an initial randomized set of clusters. Each centroid is thereafter set to the arithmetic mean of the cluster it defines. The process of classification and centroid adjustment is repeated until the values of the centroids stabilize. The final centroids will be used to produce the final classification/clustering of the input data, effectively turning the set of initially anonymous data points into a set of data points, each with a class identity.

3.3 Profile Design of Algorithm

(1) Web Crawling Module

According to the design of the web crawler based on the content framework, the crawling system is mainly divided into two layers.

The role of the application layer is to collect the topics of the webpages and artificially filter every breaking news item in the specified webpage for use by subsequent modules. It is also responsible for the implementation of crawler crawling web pages, including the storage of web resources, processing links, seed screening control, and so on.

The function of the link extraction layer is to match the stored web resource with the existing model, and the web page is modeled by calculating the model degree and stored as test data.

(2) Word Segmentation Module

This module adopts the Chinese lexical analysis of Cascading Hidden Model. This method mainly includes the process of N-shortest path coarse classification, the identification of unregistered words, elimination of ambiguity and vocabulary tagging.

(3) TF-IDF Module

After we processed the document, we find a corpus that can be used to fit the Tfidf Transformer. Then we need to create a count vector. To create a count vector we'll need to implement sklearn's CountVectorizer and fit it with the corpus and the new document we're looking to summarize. After this, we have to build the tf-idf matrix. What's more, to rank each sentence we need to score each sentence using the tf-idf values calculated.

(4) K-means Module

There are two approaches to carrying out the *K*-Means procedure. The approaches vary as to how the procedure begins the partitioning. The approach is to do this randomly, that is to start out with a random partitioning of subjects into groups and go from there. The alternative is to start with an additional set of starting point to form the centers of the clusters.

3.4 Detailed Design of Algorithm

(1) Web Crawling Module

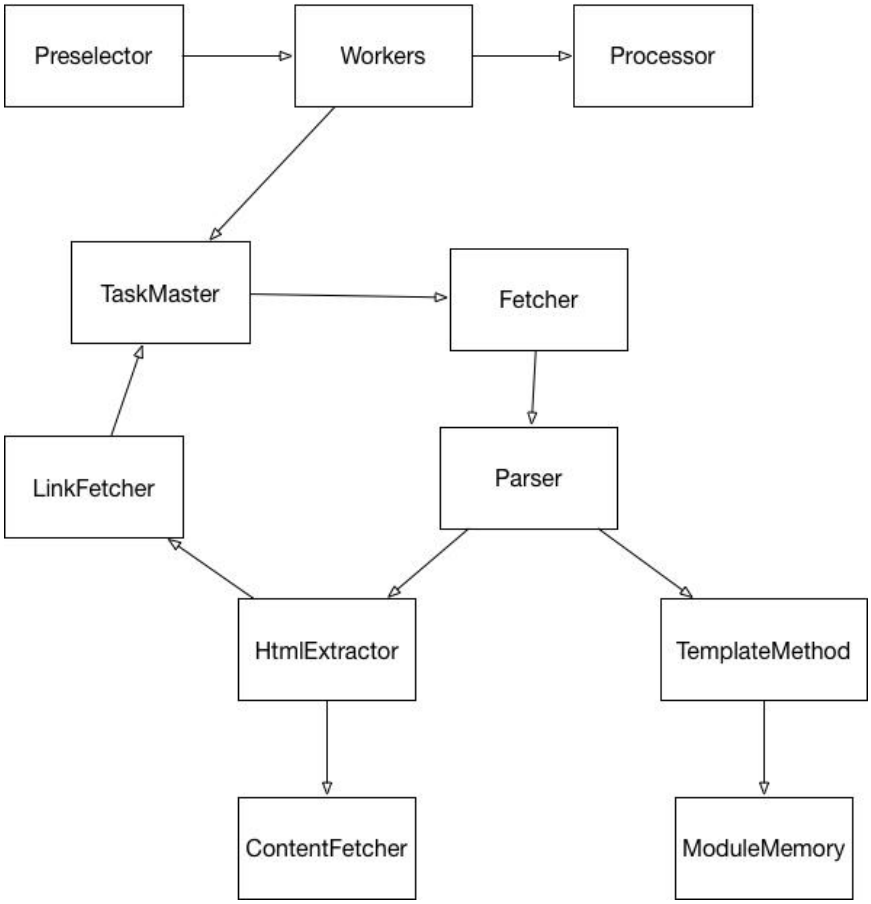

Figure 3.2 Application Layer Module

The task of the application layer is based on the implementation of crawler function. The program for automatically extracting the webpage starts from the URL of one or several initial webpages, obtains the URL on the initial webpage, and continuously extracts new information from the current webpage during the webpage crawling process. The URL is put into the queue, which meets certain stopping conditions. The workflow of this system's crawler is more complex. It needs to filter links according to certain webpage analysis algorithms, keep useful links and put them into URL queues waiting to be crawled. Then, it selects the URL of the web page to be crawled from the queue according to a certain search strategy. After this, it repeats the above process until it reaches a certain condition. In addition, all webpages crawled by crawlers will be stored in the system, subjected to certain analysis, filtering, and indexing for later query and retrieval. The analysis results obtained in this process may also be applied to the subsequent crawling process. The Application layer module is as Figure 3.2.

The pre-selector mainly does some preparatory work, includes adding the incident seed link to the URL queue and starting crawling. The seed is a link and is including seed links, parent links, child links and so on.

After this, we could complete the process of splitting seed link domain names. Domain

names can be divided into different levels, including top-level domain names and second-level domain names. According to the theme of the website, the names of the third-level or second-level domain names will be different according to different topics, but the second-level domain names will remain the same. In this case, we need to get the corresponding information. There is another way to distinguish the subject areas through the resource location path method on the website and this method is added to the link processing policy according to the domain name.

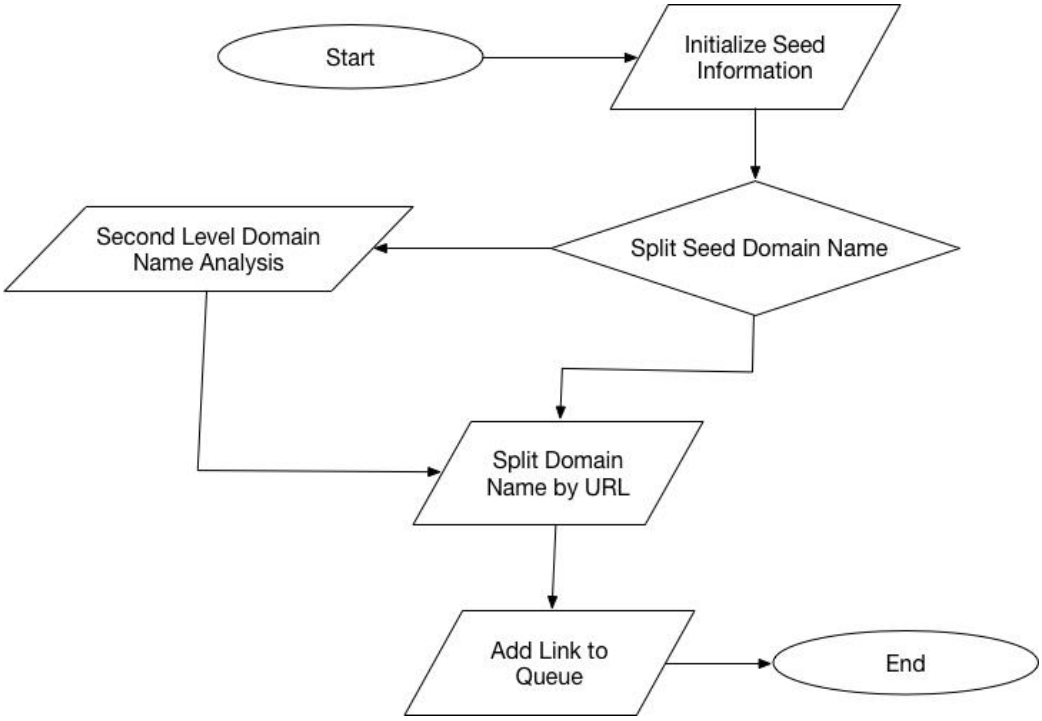

Figure 3.3 Pre-fetch Chain Model Flow

Then, what we have to do is to link the file to the record. Since the Internet is actually a huge map and each page can be viewed as a node. The links in the page are the directed edges of the graph. Therefore, the traversal of the graph can be accessed through the traversal of the

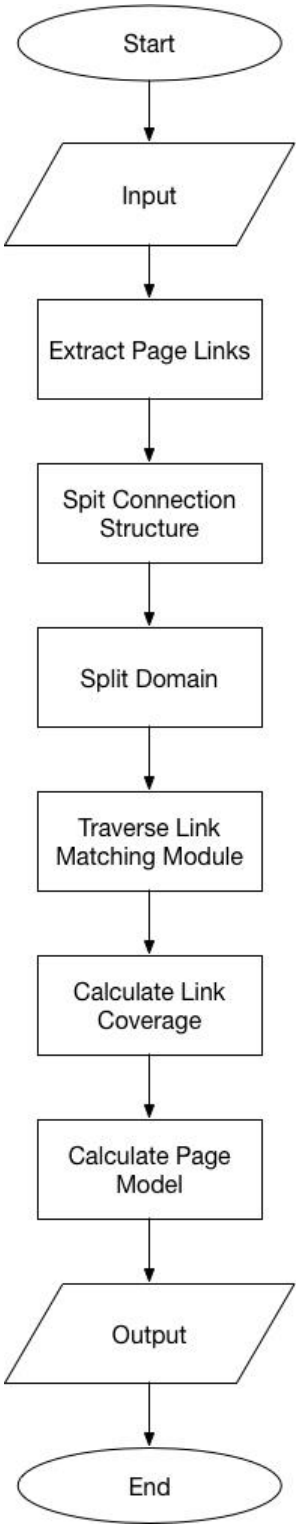

Figure 3.4 Link Extraction Module Flow

graph. The traversal of the graph can be divided into width-first traversal and depth-first traversal, but the depth-first traversal may be much too deep in depth. Most crawlers do not

use this approach and there may be a callback between the simultaneous retrenchment of the breadth-first traversal if we have not recorded the link of the site has been crawled. This will lead to the result of an infinite loop. In this case, storing the link in the file can effectively avoid the possibility of duplicate links crawling again. Finally, we add the link to the queue. The seed URL is added to the URL queue for the next module to seed it. This whole process is shown as Figure 3.3.

The link extraction module is an important module in the crawler strategy. It provides the crawler with a seed link to the page depth of the Tong countryside layer. Due to the need to match the template, in this step we need to extract the links in all web pages as a sample set. The link extraction module workflow is shown in Figure 3.4.

Suppose there is a link sample set $\{S_1, S_2, S_3, S_4, S_5 \dots S_n\}$ in the page *S*, and *S* represents the object with the link and link name. Based on the linked sample set, the structure of each connection object needs to be extracted. Link is generally composed of these parts, top-level domain, second-level domain, third-level domain name, port number (default 80), resource path. The domain name consists of two or more words, separated by periods and the rightmost word is called the top-level domain name. Generally, com, net and cn are used as top-level domain names. The distinction between different blocks of a website is generally reflected in third-level domain names. The main steps of the link extraction module are:

After processing the seed link, the link in the list candidate module is taken out and the module identifier is set. The candidate module set is $H = \{H_1, H_2, H_3, H_4, H_5 \dots H_n\}$, and the link set in the candidate module is marked as Formula 3.9.

$$\begin{aligned} H_1 &= \{H_1S_1, H_1S_2 \dots H_1S_{n-a}\} \\ H_1 &= \{H_2S_{n-a+1}, H_2S_{n-a+2} \dots H_2S_{n-a+b}\} \\ &\dots \\ H_n &= \{H_nS_{n-n_1}, H_nS_{n-n_1+1} \dots H_nS_n\} \end{aligned} \tag{3.9}$$

The link weight of each candidate module is denoted as δ_h, which is the number of module links divided by the number of full-page links. This could lead to the link weights in the module can be obtained. Calling the domain name division to sort the number of domain names could get the result of $num(Dn)$, the link of the candidate modules consists of a set of links of different third-level domain names.

(2) Word Segmentation

The structure of Chinese word segmentation module framework is as Figure 3.5.

The role is the kind of identity the words look forward to in a context. For a given initial partitioning result $W=(w_0 w_1 w_2 \ldots w_n)$. In the role category, assuming that $R=(r_0 r_1 r_2 .. r_n)$ is the character sequence corresponding to the word sequence. We take the character sequence with the greatest probability as the final role of the word sequence. We treat the words as observations and treat the characters as state values. According to the Markov chain we can get Formula 3.10.

$$P(R) = \prod_{i=1}^{m} p(r_i|r_{i-1}) p(w_i|r_i) \tag{3.10}$$

Through the algorithm described later, the system obtains a sequence of maximal probabilistic characters. A specific type of unnamed word recognition is achieved through the

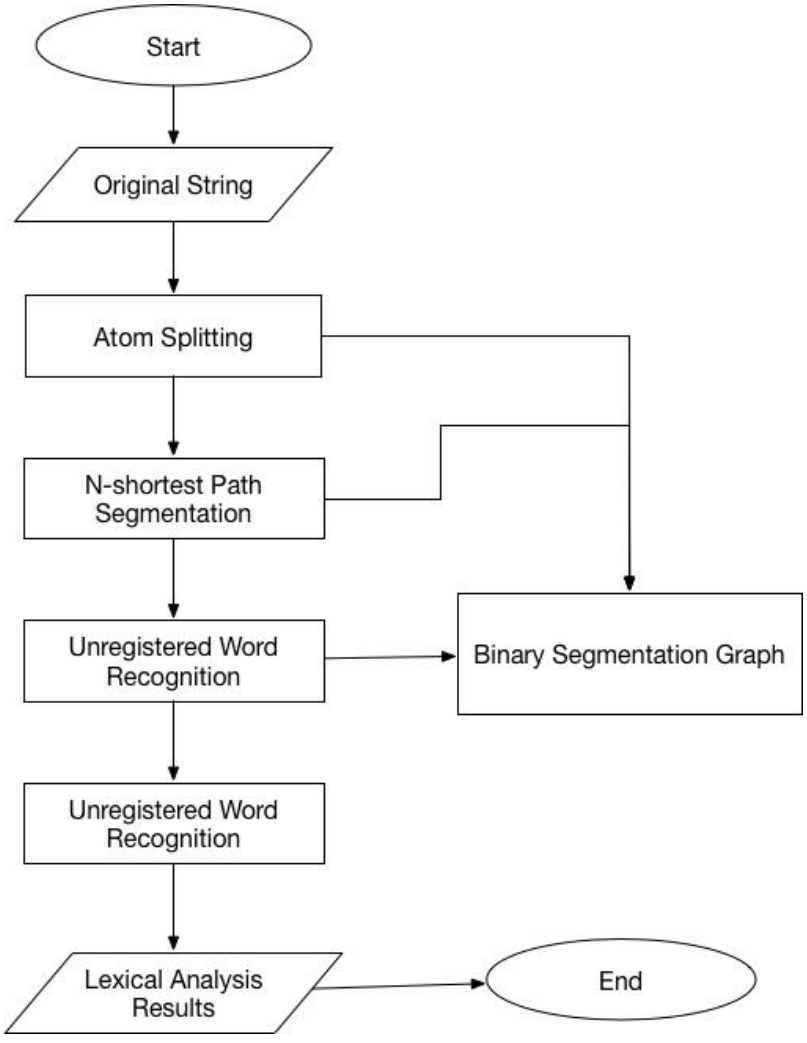

Figure 3.5 Flow Diagram of Word Segmentation

matching of a specific person name template. For example, if the sequence of characters obtained by the post-algorithm sequence is CDEAEF, then the corresponding words of CDE are combined together because the name combination of Chinese names is CDE (C: Chinese surname, D: Chinese first name, E: Last word of name). Solving the largest role sequence: As

shown in Figure 3.6, in order to facilitate the calculation, it is assumed that the arc length from $r_{i-1,k}$ to $r_{i,j}$ is $\left(-log_P(r_i|r_{i-1}) - p(w_i|r_i)\right)$, then the problem will be converted into

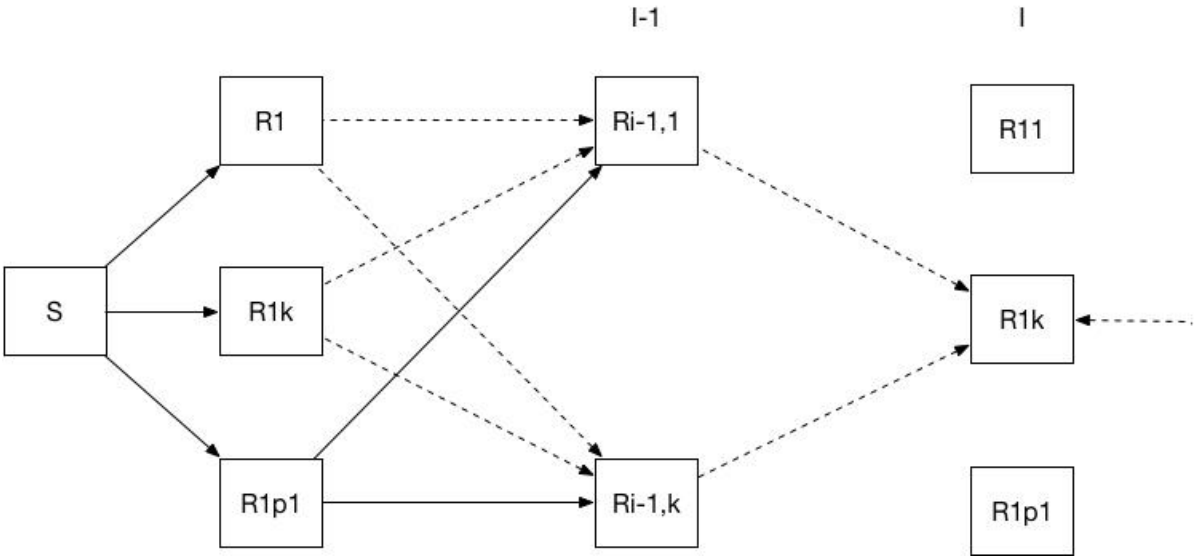

Figure 3.6 Find the maximum sequence diagram

the shortest path from the *S* node to the endpoint *T*. Assume that *D[i][j]* represents the shortest path of the *j*-th character from S to the *i*-th word, then *D[i][j]=min{*$\left(-log_P(r_i|r_{i-1}) - p(w_i|r_i)\right)$*}*, where *n* is the number of characters the word w_{i-1} exists. Therefore, this paper uses dynamic programming algorithm to solve the shortest path from *S* node to *T* node. In order to record k for *D[i][j]*, the system introduces the path recording matrix *BestPrev[m][n]*, where *m* is the number of words and *n* is the number of characters with the most words.

Finds the fill matrix BestPrev when finding the shortest path from S to T could be described as Figure 3.7.

```
ShortestPath()
{
1.  For(int i = 0; i < Words[0].nPos; i++)
a)  BestPrev[0][i] = -1;
2.  For(int i = 1; i < nWords; i ++)
a)  For(int i = 1; i < nWords[i].nPos; j++)
i.  temp = maxno
ii. For(int k = 0; k < Words[i-1].nPos; k++)
1.  If(D[i-1][k] + (-logp) - logp()) < temp)
a)  temp = D[i-1][k] + (-logp) - logp());
b)  BestPrev[i][j] = k;
2.  D[i][j] = Temp;
}
```

Figure 3.7 Pseudocode of Find the Shortest Path

Get the best role and save it on the stack PosStack, which is Figure 3.8.

```
GetBestPos()
{
    int k = BestPrev[nWords][0];
    While(k != -1)
    {
        PosStack.Push(Words[nWords-1].Pos[k]);
        k = BestPrev[nWords-1][k];
        nWords --;
    }
}
```

Figure 3.8 Pseudocode of Get Best Roles

At this point, the best role of a given word division has been decided. The system will match this character string with the pattern of a specific unregistered word, and if it matches, it will be grouped together as a new word.

The weight of new unlisted words: In order for the newly-identified unlisted words and ordinary words to participate in the next round of competition, we need to calculate the weight of their competition. The specific calculation method is $P(w|C) = p(w_{p+j}|r_{p+j})$; where k is the length of the identified unregistered word, and r_{p+j} is the optimal role value obtained above.

After the system performs the operation as shown in the figure, it needs to sort the entire N paths and use the best result as our result. In addition, the system uses a core dictionary, an unregistered word dictionary, and a character frequency statistics dictionary.

(3) TF-IDF module

Figure 3.9 is the general process we'll follow to summarize a document.

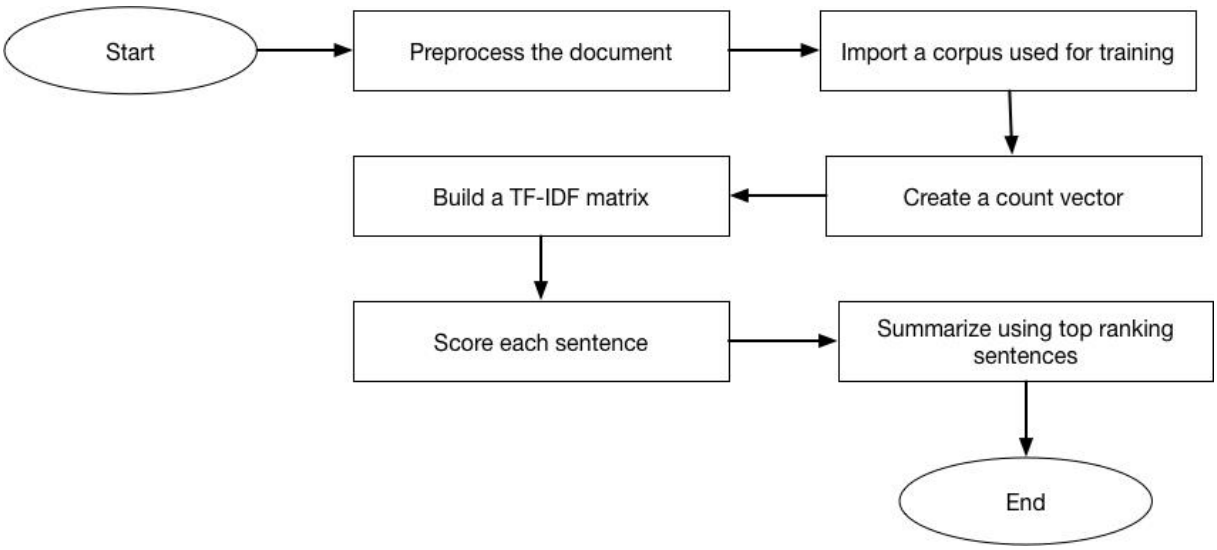

Figure 3.9 General Process to Summarize a Document

When doing NLP, we often find ourselves working with tokenized sentences. Sentences

are often separated (tokenized) where ever there is a period. If there are any extra periods within a sentence, for example those used in acronyms we'll incorrectly split the sentence. Thus it's good practice to explore the types of documents you'll be working to identify these nuances so that they can be properly addressed.

Then we have to gather corpus for training. We used the corpus which generated for the baidu breaking news. To create a count vector we'll need to implement sklearn's CountVectorizer and fit it with the corpus and the new document we're looking to summarize.

Next is to create the tf-idf matrix, which is to pass the freq_term_matrix we defined above into TfidfTransformer's fit method. Once we have the tf-idf transformer fitted, we can take the original document, vectorize it, and transform it into a tf-idf matrix.

To rank each sentence we need to score each sentence using the tf-idf values calculated above. Rather than simply taking the summation of all the values for a given sentence, we'll be using some additional techniques outlined in this paper. These include:

①Only summing tf-idf values where the underling word is a noun. This total is then divided by the summation of all the document tf-idf values.

②Add an additional value to a given sentence if it has any words that are included in the title of the document. This value is equal to the count of all words in a sentence found in the title divided by the total number of words in the title. This "heading similarity score" is then multiplied by an arbitrary constant (0.1) and added to the tf-idf value.

③Apply a position weighting. Order each sentence from 0 to 1 equally based on the sentence number in the document. For example if there are 10 sentences in a document, sentence nine's "position weighting" would be 0.9. This weighting is then multiplied by the value calculated in point 2.

After we've tagged the sentences, it's as simple as looking up the index value (bag of word mapping) for each word in a sentence and finding the tf-idf score in doc_tfidf_matrix.

After applying the above we can finally sort our sentences in descending order and choose the top 3 (or 4, 5, 6 …). And boom, we have a summary based on the most important sentences found in a document.

(4) K-means

The principle of the k-means algorithm is shown in Figure 3.10[9].

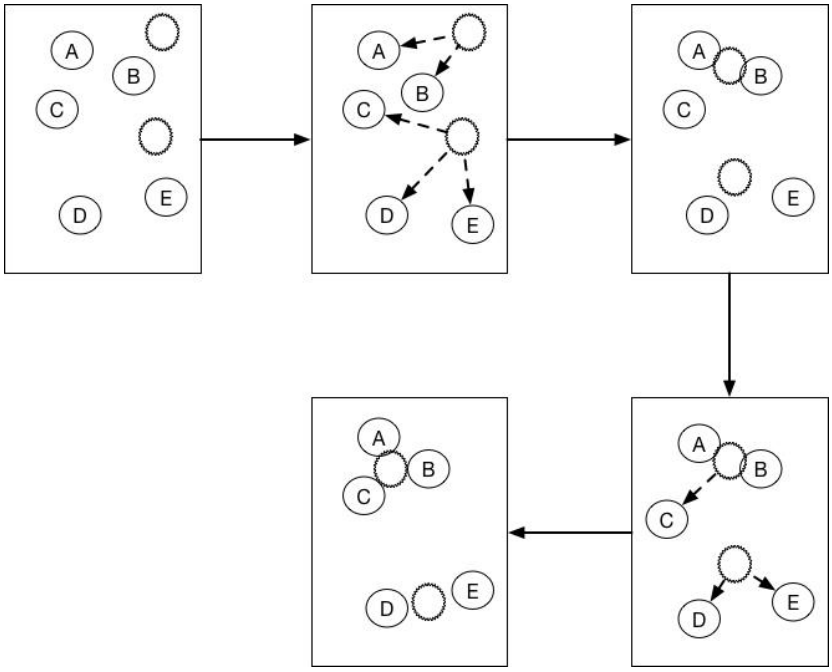

Figure 3.10 K-means Algorithm Principle

From the above figure, we can see that A, B, C, D, E are five po0nts in the figure. The grey point is our seed point, which is the point we use to find the point group. There are two seed points, so K=2. Then K-Means' algorithm is as follows:

② Take K (here K=2) seed points randomly in the graph.

②Then find the distance from the K seed points for all the points in the graph. If point A is closest to the seed point K1, then A belongs to the K1 point group. (In Figure 3.10, we can see that A, B belongs to the above seed point, and C, D, E belong to the seed point below the middle)

③Next, we want to move the seed point to the center of his “point group”.

④Then repeat steps 2 and 3 until the seed points have not moved (we can see that the seed points above the fourth step in the figure converged on A, B, C, and the seed points below aggregated D , E).

According to the above steps, draw the algorithm execution flow as shown in Figure 3.11.

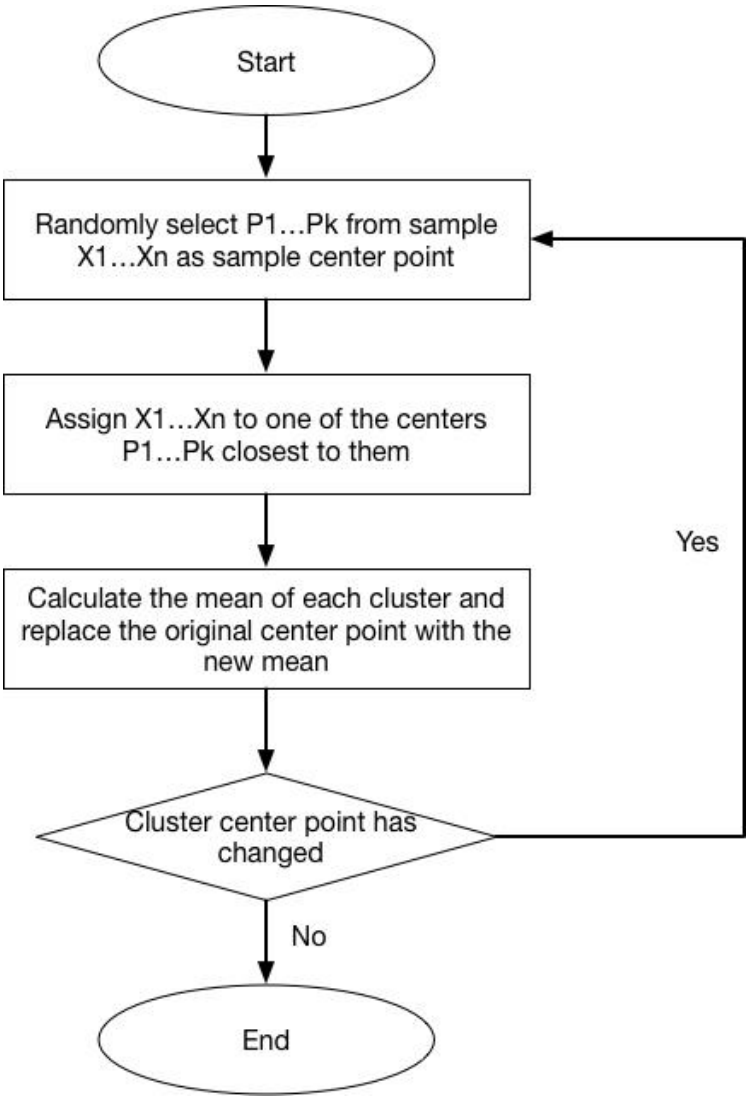

Figure 3.11 K-means Execution Flow

Chapter 4 Algorithm Implementation and Evaluation

4.1 Algorithm Implementation

4.1.1 Web Crawling

(1) Implementation of web crawling

First, we use the pyquery library to locate td with a class attribute of 'keyword' as :

```
#百度风云榜页面网址(含有 50 个热门新闻的关键词)
fengyunbang_url = 'http://top.baidu.com/buzz?b=1'
#从 html 文件中解析出事件字段和网址字段
doc = PyQuery(resp.text)
for item in doc.items('.keyword'):
    keyword = item('a').text().split(' ')[0]
    keyword_link = item('a').attr.href
    news_links = get_keywords_news_links(keyword_link)
    for news_link in news_links:
        try:
            content = get_news_content(news_link)
            if content:
                print(keyword, content[0:20])
                writer.writerow((content, keyword))
        except:
            print(news_link)
```

Run, the result keyword is all garbled, there is no trace of Chinese. This is the problem we need to overcome，which is the html coding problem. With this type of problem, we may first look for the charset character set in the html tag. The general charset values are utf-8, gbk, gb2312, ascii and so on. To solve the problem, we could use the chardect library.

Above we have got the keywords and their links. We want to get the news content. To get news content, we need to know the news corresponding links. First of all, we have to locate the link. Here we use regular expressions to locate the link, which is:

```
def get_keywords_news_links(keyword_link):
    """
    访问关键词百度网址，得到相关新闻的 link
    :param keyword_link:
    :return:
    """
    headers = {'User-Agent': 'Mozilla/5.0 (Macintosh; Intel Mac OS X 10_13_1) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/66.0.3359.139 Safari/537.36'}
    resp = requests.get(keyword_link, headers=headers)
    bsObj = BeautifulSoup(resp.text, 'html.parser')
    news_items = bsObj.find_all('div', {'class': 'result c-container '})
    news_links = []
    for item in news_items:
        links = re.findall('href="(.*?)"', str(item))
        news_links.extend(links)

    news_links = set([link for link in news_links if 'http://cache.baiducontent' not in link])
return news_links
```

With the link obtained above, we write a simple code to get the text content. Since hundreds of web sites are sourced from a web site, if you want to accurately obtain news content, you need to position each web site one-on-one. This is too much trouble. For quickness and convenience, we use regular matches to match all Chinese content, which is:

```
    news_text = ''.join(re.findall('[\u4e00-\u9fa5]+', resp.text))
    if not news_text:
        chaset = chardet.detect(resp.content)['encoding']
        resp.encoding = chaset
        news_text = ''.join(re.findall('[\u4e00-\u9fa5]+', resp.text))
        return news_text
return news_text
```

However, during operation, it often returns null when matches Chinese. The most likely reason is that there is no Chinese in the page, but we have detected that these pages are in

Chinese, it is very likely that the page is garbled, resulting in regular [\u4e00-\u9fa5]+ cannot match to Chinese. After testing, it is garbled. Here I use the new method, chardect library to detect the encoding used in binary data.

Finally, write the main crawler function and save the data in csv:

```
csvf = open('data.csv', 'a+', encoding='gbk', newline='')
writer = csv.writer(csvf)
writer.writerow(('news_content', 'keyword'))
```

(2) Result

Run the web crawler, I have collected a total of 439 news articles with 50 keywords.

4.1.2 Text Processing

(1) Implementation of words segmentation

Firstly, we import corpora which is collection of documents to form a document list of documents. For each document in the documents, we have to cut words and set space interval to generate new new_documents. Python sklearn machine learning can only deal with such space-separated text data in English, so here we have to translate into Chinese structure.

(2) Implementation of TF-IDF

We use the TF-IDF to characterize new_documents and generate a data matrix. In this process, we eliminated some words that can appear in more than 50% of documents. These frequently appearing words often have no information and are therefore excluded.

(3) Implementation of K-means Clustering

From [1,100], the number of clusters of the K-means algorithm is used to test the error of matrix in K-means. The best point of error improvement is selected as the clustering k value. Learn the taxonomy of groups in matrix based on the best k-value. Save the predicted category along with the original text in csv.

4.2 Algorithm Evaluation

(1) Experimental environment

For the above constraint conditions and models, the experimental environment is in the PyCharm environment. The operating environment is 2.3 GHz Intel Core i5, 8 GB 2133 MHz LPDDR3, macOS High Sierra 10.13.1.

(2) Experimental data

First of all, we need to find the best k value as the number of clusters, and then learn the characterization data to discover the taxonomy. The general method to find the best k value is to perform a wide range of investigations on the range of k values. For example, k is searched in [1,100] as Figure 4.1.

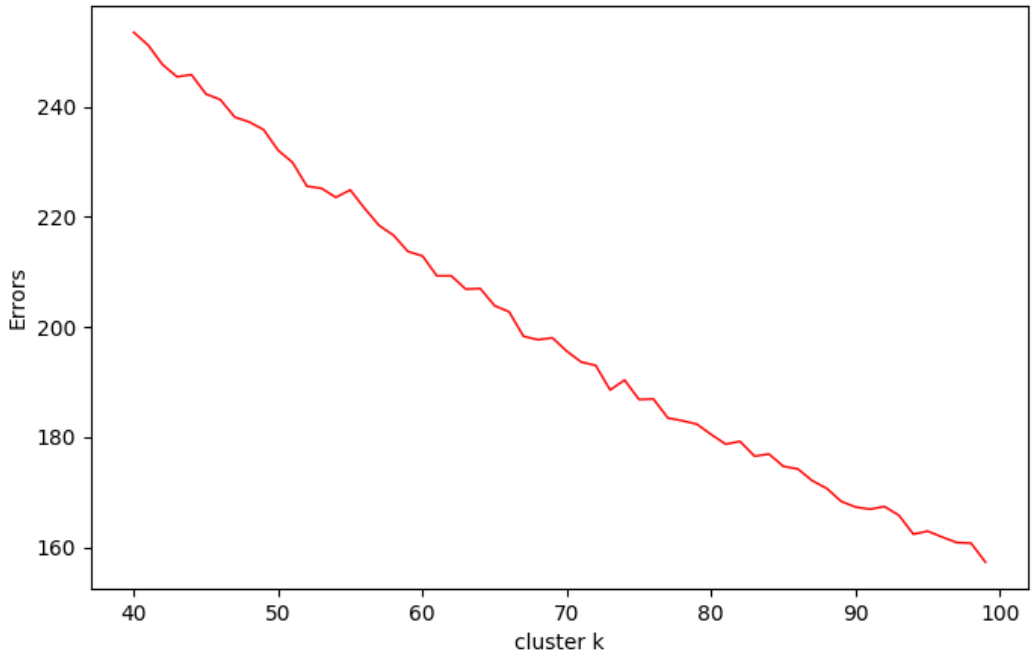

Figure 4.1 Cluster k with errors

We could see that the trend of the curve between 50-55 has slowed down, and there should be a better value of k nearby. We performed the classification effects of 51, 52 and 53 respectively.

Part of the 51 clusters k-means clustering is as Table 2.

Table 2 Part of K-means clustering with 51 clusters

Real Label	**id**	**Cluster**
情侣遭敲门威胁	1-10	42, 29, 8, 42, 42, 24, 49, 42,8
暴漫创始陵园道歉	11-19	17, 17, 17, 17, 17, 17, 17, 49, 17
重庆面馆天价面条	20-29	43, 4, 50,50, 49, 23, 50, 47, 49, 7
一学习就流鼻血	30-40	43, 23, 12, 16, 8, 36, 47, 19, 47, 47, 16
借宝马当婚车被烧	41-48	34, 16, 14, 16, 16, 16, 16, 50
王毅回应美撤邀请	49-57	4, 13, 14, 9, 14, 15, 14, 0, 4

Part of the 52 clusters k-means clustering is as Table 3.

Table 3 Part of K-means clustering with 51 clusters

Real Label	**id**	**Cluster**
情侣遭敲门威胁	1-10	47, 27, 8, 27, 27, 6, 37, 27, 8
暴漫创始陵园道歉	11-19	34, 34, 34, 34, 34, 34, 34, 37
重庆面馆天价面条	20-29	6, 6, 6, 6, 37, 6, 6, 1, 37, 24
一学习就流鼻血	30-40	40, 12, 25, 2, 8, 40, 1, 20, 1, 1, 17
借宝马当婚车被烧	41-48	13, 42, 25, 42, 42, 42,17, 25
王毅回应美撤邀请	49-57	40, 40, 39, 21, 40, 40, 0, 40, 40
张艺谋吐槽低头族	58-66	29, 21, 8, 47, 6, 24, 2, 6, 7

Part of the 53 clusters k-means clustering is as Table 4.

Table 4 Part of K-means clustering with 51 clusters

Real Label	**id**	**Cluster**
情侣遭敲门威胁	1-10	33, 15, 6, 33, 30, 33, 2, 33, 6
暴漫创始陵园道歉	11-19	16, 16, 16, 16, 16, 16, 16, 2, 16
重庆面馆天价面条	20-29	33, 18, 12, 27, 2, 0, 33, 3, 2, 9
一学习就流鼻血	30-40	33, 0, 22, 44, 6, 25, 3, 20, 3, 3, 12
借宝马当婚车被烧	41-48	5, 20, 32, 22, 39, 39, 12, 14
王毅回应美撤邀请	49-57	18, 18, 1, 4, 18, 46, 18, 18, 18
张艺谋吐槽低头族	58-66	29, 50, 6, 50, 50, 9, 21, 18, 30

(3) Evaluation result analysis

It can be seen from the clustering results that the clustering results obtained by the K-means algorithm have good stability. In every set of real label, the cluster result is centered, this means the clustering results is accurate. The clustering results are basically consistent with the original ids, indicating that the clustering effect is good.

Chapter 5 Conclusion and Future Work

This paper studies a Web document clustering system based on K-means algorithm and develops a web document clustering system composed of web crawler, Chinese word segmentation, TF-IDF, cluster analysis and other modules. The article summarizes data mining, cluster analysis, web mining and introduces the entire system architecture. It also makes deep research on modules such as web crawler, Chinese word segmentation, TF-IDF and cluster analysis. Finally, a web document clustering system composed of web crawlers, Chinese word segmentation, TF-IDF and cluster analysis was developed to conduct a comparative experiment. The system has improved both in accuracy and stability. At the end of the article, the main work content of this paper is introduced and further researches and issues that need to be improved are put forward.

The research of this thesis has only involved a small part of this field. Therefore, we will continue to conduct in-depth research based on existing research in the future, including:

1. Perform more experimental analysis and comparison of the effects of traditional clustering algorithms to deepen the understanding of the characteristics of the cluster analysis.

2, While improving the algorithm, improving the accuracy and stability of the algorithm.

3. Visualize the clustering results and display the clustering results more vividly.

4. Improve K-means algorithm and apply it in more fields of application.

References

1. Punam B, Anjali T. A Multi-Threaded Semantic Focused Crawler [J], Journal of Computer Science and Technology, 2012, 2:16.

2. Subhendu K P, Deepak M, Bikram K R. Integration of Web mining and web crawler: Relevance and State of Art [J], International Journal on Computer Science and Engineering, 2010, 772.

3. Jones K S. A statistical interpretation of term specificity and its application in retrieval [J], Journal of Documentation, 1972, 28 (1): 11- 21.

4. Salton G, Clement T Y. On the construction of effective vocabularies for information retrieval[C], Proceedings of the 1973 Meeting on Programming Languages and Information Retrieval, New York: ACM, 1973: 11.

5. Salton G, Fox E A, Wu H. Extended boolean information retrieval[J], Communications of the ACM, 1983, 26 (11): 1022 -1036.

6. Salton G, Buckley C. Term weighting approaches in automatic text retrieval[J], Information Processing and Management, 1988:513 - 523.

7. Junghoo C, Hector G M, Lawrence P. Efficient Crawling Through URL Ordering, Computer Networks and ISDN Systems, 30(1-7):161-172, 1998.

8. Rajaraman A, Ullman J D. Mining of Massive Datasets[J], Data Mining, 2003: 1–17.

9. Stuart P, Least squares quantization in PCM Information Theory, IEEE Transactions on 28.2 (1982): 129-137.

10. http://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

Acknowledgement

. This research was supported by my parents. I am thankful to Lecturer Dongqi Wang who provided expertise that greatly assisted the research. I have to express out appreciation to Lily for sharing her pearls of wisdom with me during the course of this research. I am also immensely grateful to Dongqi Wang for his comments on an earlier versions of the manuscript, although any errors are my own and should not tarnish the reputations of these esteemed professionals.