electronics

MDPI

Article

Key Information Extraction and Talk Pattern Analysis Based on Big Data Technology: A Case Study on YiXi Talks

Hao Xu [1,2], Chengzhi Jiang [1,2], Chuanfeng Huang [2], Yiyang Chen [3], Mengxue Yi [2] and Zhentao Zhu [2,*]

[1] School of Information Management, Nanjing University, Nanjing 210023, China; j00000003045@njit.edu.cn (H.X.); jcz@njit.edu.cn (C.J.)
[2] School of Economics & Management, Nanjing Institute of Technology, Nanjing 211167, China; hcfnjit@njit.edu.cn (C.H.); x00209170926@njit.edu.cn (M.Y.)
[3] School of Computer Science, University of St. Andrews, St. Andrews KY16 9AJ, UK; chenyiyang2019@udirecter.com
* Correspondence: j00000000832@njit.edu.cn; Tel.: +86-137-7667-8250

Abstract: In the attempt to extract key information and talk patterns from YiXi talks in China to realize "strategic reading" for readers and newcomers of the speaking field, text mining methods are used by this work. The extraction of key information is realized by keyword extraction using the TF-IDF algorithm to show key information of one talk or one category of talks. Talk pattern recognition is realized by manual labeling (100 transcripts) and rule-based automatic programs (590 transcripts). The labeling accuracy rate of "main narrative angle" recognition is the highest (70.34%), followed by "opening form" (65.25%) and "main narrative object", and the "ending form" is around 50%, with the overall accuracy of the rule-based automatic recognition program for talk patterns at approximately 60%. The obtained results show that the proposed keyword extraction technology for transcripts can provide "strategic reading" to a certain extent. Mature speech mode can be summarized as follows: speakers tend to adopt a self-introducing opening format. They tell stories and experiences through a first-person narrative angle and express expectations and prospects for the future. This pattern is reasonable and can be referenced by new speakers.

Keywords: text mining; keyword extraction; talk pattern recognition; transcripts feature extraction; YiXi

Citation: Xu, H.; Jiang, C.; Huang, C.; Chen, Y.; Yi, M.; Zhu, Z. Key Information Extraction and Talk Pattern Analysis Based on Big Data Technology: A Case Study on YiXi Talks. *Electronics* **2022**, *11*, 640. https://doi.org/10.3390/electronics11040640

Academic Editor: Stefanos Kollias

Received: 27 December 2021
Accepted: 15 February 2022
Published: 18 February 2022

1. Introduction

Technology, entertainment, design (TED) talks, which started in the United States in 1984, has received widespread attention all over the world due to its dedication to "change the world with the power of thought" and its resounding, straightforward, widely varied, and novel views. In China, in 2012, YiXi talks became a representative of new media, a form of dissemination that uses digital technology to provide users with information and services through computer networks, wireless communication networks, satellites, and terminals. YiXi talks commit to the spread of novel and interesting ideas with profound topics and interesting story expressions, and it has good social influence and brand value in China [1]. YiXi talks has formed a variety of multiple communication media (video, audio, text, WeChat, audio/video app, website, etc.), and up to the end of 2019, approximately 700 speakers from multiple areas were invited to share stories from several fields (humanities, science and technology, and natural history, etc.). In the era of big data, people are in an "ocean" of information, and Internet industry has proposed higher requirements for text/image/video information processing [2]. How can people not be "submerged" by complicated information and how to make the vivid, unique, and profound stories of YiXi be perceived by audiences to inspire resonance and thought? Do successful speakers' talk style and skills have specific patterns that can be imitated?

This research applies text mining to the corpus of YiXi transcripts. The key technology is text feature extraction, mainly used for keyword extraction and talk pattern recognition. Transcripts of YiXi can be regarded as a multidimensional vector combination composed of many words or sentences and key contents are several words or phrases that most fit the subject of transcripts [3,4]. Talk patterns are the disclosure of specific features of a particular paragraph in transcripts (primarily focuses on opening or ending form, narrative angle, and narrative object). Therefore, a key task of text feature extraction is to filter keywords or patterns that can represent the main connotation of transcripts to achieve strategic reading, which means reading the brief (key content extraction by text mining) to see if it is worth reading or not. This paper explores key content and the talk pattern for each of these YiXi talks, through application of text mining. The former can be used for "strategic reading", while the latter can be referenced by speakers (especially newcomers), to promote multidimensional communication of knowledge.

2. Related Work

This research applies technologies in the field of big data to feature recognition, which mainly involves the extraction of text characteristics and transcripts analysis.

2.1. Text Feature Extraction

(1) Keyword Extraction Technology. In 1957, Luhn first proposed an automatic keyword extraction method based on word frequency. The automatic method for keywords extraction has derived a variety of categories, and scholars have systematically combined them in practical applications [5–8]. Some scholars have used visual methods to reveal the co-occurrence relationship between extracted keywords with the help of CiteSpace [9]. In terms of classification, keyword extraction algorithms can be divided into two categories: supervised and unsupervised learning algorithms. The former extracts the candidate words in the text in advance, delimits labels in accordance with whether the candidate words are keywords or not, and trains classifier as a training set. When extracting keywords for a new document, the classifier first extracts all candidate words, matches the words in keyword vocabulary, and uses candidate words where their tags are designated as keywords to be keywords from a document. This algorithm is fast but requires many annotated data to improve the accuracy. Compared with a supervised learning method, an unsupervised algorithm has the characteristic of early appearance and multiple types. Its key steps mainly include text preprocessing, determining the set of candidate words, sorting the list of candidate words, and evaluating effect of keyword extraction. Unsupervised learning methods can be divided into simple statistics-based methods, graph-based methods, and language model-based methods [10]. The basic principle of its work is to statistically classify some specific indicators of candidate words that characterize the text and then sort in accordance with the statistical results, such as N-gram [11], term frequency-inverse document frequency (TF-IDF) [12], word frequency [13], word co-occurrence [14], and other algorithms to evaluate the importance of extracted candidate words in a document. The unsupervised algorithm does not need to maintain word lists, nor does it rely on manually labeled corpus to train the classifier. Among unsupervised learning algorithms, TF-IDF can consider word frequency and freshness to filter words that can represent key information of documents and always with simpler and more convenient operation process and better accuracy. Therefore, our work uses TF-IDF algorithm to extract key information from documents (namely YiXi transcripts).

(2) Text Feature Recognition Technology. Another key technique is feature recognition of talk transcripts, which is a form of extractive automatic abstract of text mining [15]. Machine learning techniques are used in several areas of practice, such as disease prediction [16] and detection [17], stock trend prediction [18], judicial case decisions [19], etc. Different from automatic abstract, our research focuses on the start and ending forms, narrative angle, and the object of YiXi talks rather than summarizing documents

by sentences from documents. In our work, classical generative methods need to combine linguistic and domain knowledge for reasoning and judgment. High-quality abstracts are usually generated after text analysis, representation transformation, and abstract synthesis. However, the shortcoming is that they are domain-constrained and rely on large-scale real corpora and professional domain knowledge. Traditional methods and technologies in automatic summarization have been developing slowly, while the use deep learning methods in automatic summarization has shown promising [20].

2.2. Analysis for Talk Transcripts

Our research uses transcripts of YiXi talks to conduct direct text analysis. Text analysis methods have been applied to the feature recognition of transcripts. Kavita S. [21] used text analysis methods to identify degrees of lecture users' favor for lectures. Zhou HY [22] applied text analysis methods to analyze and study the form, audiovisual system, narrative, and expression structure of transcripts with the name of "Under the Dome". (a speech made by Chai Jing in China). Using text mining, Min Y [23] analyzed Konosuke Matsushita's transcripts and concluded data extraction analysis, and visual presentation of analysis results constitute the key content in the process. Important words and phrases in the text can represent the subject of literature. Supported by this view, Liu H [24] used 11 words identified by word frequency analysis to analyze the generation mechanism of diplomatic discourse with transcripts of May's speech at the 71–73 session of the UN General Conference taken as corpus.

Documentary transcripts can be the basis for identifying key information, though they rarely involve large-scale, talk-type corpus construction. Further, the technology adopted does not rely on big data analysis technology. However, the large scale and complex structures of real networks [25] make it difficult to identity key information and talk patterns in a large-scale corpus. Relevant studies of "YiXi" are limited to a qualitative perspective, such as demonstrating the innovation and improvement measures, communication characteristics of new media [26], and program style and dissemination effect [27]. Big data analysis technology has not been conducted; therefore, our work intends to expand on the above two aspects.

3. Data Sources and Methods

Big data management and analysis has become more and more important in academia and industry [28]. Therefore, its collection, organization, and analysis is a systematic project. YiXi talks are currently distributed through the official website, as well as Sina Microblog, Tencent Video, NetEase Cloud Music, and other platforms. Our research collects data from the official website of YiXi (https://yixi.tv/ (accessed on 26 December 2021)), specifically, the transcripts of YiXi talks, since speaker information and user commentary information for each talk are also included and available for further research.

The information collection program is developed by Python, and the collected data are structured and stored in SQL Server 2012. Data are structured in terms of speakers' information, transcripts information, and video comments. The relevant data acquisition process is shown in Figure 1. On 12 December 2019, 690 transcripts and the corresponding speaker information as well as 23,685 comments were collected, and 690 transcripts were analyzed through text mining to identify key information features of YiXi talks.

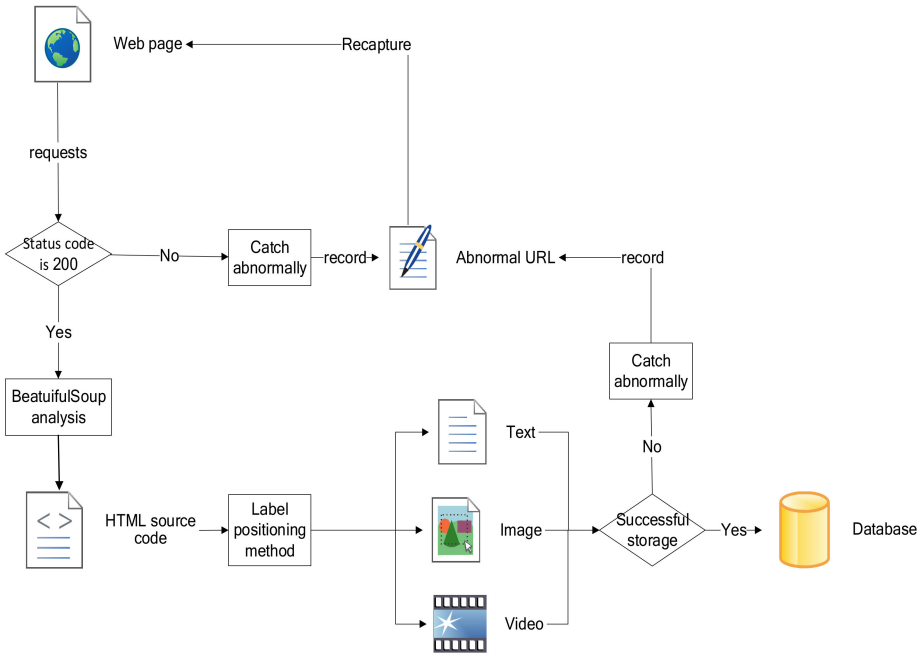

Figure 1. Data collection process of YiXi talks.

4. Keyword Extraction of YiXi Transcripts

In the era of big data, conducting "strategic reading" is particularly important to avoid being "submerged" by information. Recognizing key information and patterns of talks by big data analysis technology is exploratory research on the characteristics of massive texts. This process can provide users with a "recommended reading" method.

Keyword extraction is of great importance for text clustering, automatic summarization, and information retrieval. Many related explorations based on various corpora are reported [29–31]. However, in practical applications, few talks have keywords marked by the speaker, and most of them are recommended or extracted automatically by platforms or search engines automatically. Our work uses word segmentation, deactivation word removal, keyword extraction, and other steps to identify key content of the transcripts and recommend to readers to save readers' reading costs and realize "recommended reading".

4.1. Word Segmentation of Transcripts

Different from English, no clear separation marks exist between words in Chinese, which are always presented as a continuous string. Thus, Chinese natural language processing tasks must solve problems of Chinese sequence segmentation and word segmentation [32]. At present, word segmentation for Chinese text is relatively mature, with the help of general Chinese word segmentation tools. The problem of ambiguity in Chinese word segmentation has been widely studied by scholars. In accordance with the implementation principles and characteristics, word segmentation is mainly divided into dictionary-based and statistics-based ones. In practice, common tokenizers adopt a combination of these two approaches to improve the accuracy of word segmentation and the applicability of the tokenizer to texts in different fields. This paper uses the jieba tokenizer (with precise tokenization mode) to conduct word segmentation [33]. To improve the efficiency of word segmentation, authors continuously enrich the underlying vocabulary of the jieba tokenizer.

4.2. Removal of Stop Words

A large number of "stop words" (such as "you", "thereby", and "in general", etc.) in the text cause interference to the extraction of key information and affect the efficiency and accuracy. Therefore, "eliminating noise" in text after Chinese word segmentation is necessary to improve the accuracy [34]. Our work used a stop word dictionary constructed

by merging the Baidu stop word list (which contains 1395 stop words), the Harbin Institute of Technology stop word list (767 stop words), and the Sichuan University Machine Intelligence Laboratory stop word list (976 stop words). After deduplication, the resulting stop word list includes 3136 different stop words. On the basis of 4.1, the new and merged stop word dictionary is used to eliminate noise and reduce the dimensionality of vectors.

4.3. Keyword Extraction from Yixi Transcripts

Supervised learning algorithms and unsupervised learning algorithms are two methods used for keyword extraction. The former extracts candidate words in advance, delimits labels in accordance with whether the candidate words are keywords or not, and uses them as a training set to train a classifier. When extracting keywords from a new document, the classifier extracts all candidate words in advance, matches words in keyword vocabulary, and uses candidate words classified as keywords. This algorithm always requires labeled data to improve accuracy. The latter extracts candidate words in advance, scores each candidate, and then outputs the top N candidate words with the highest score as keywords. The core of such algorithms is scoring strategy, which include TF-IDF, PageRank, and so on. Transcripts of talks can be regard as a multidimensional vector of words and has the characteristics of high-dimensional vectors. While less time consumption is proposed, TF-IDF is suitable for this task, which is widely used in search engines, document classification, and related fields [35,36], Besides, TF-IDF also considers word frequency, the freshness of candidate words which were extracted from text. Key information from transcripts is independent of position, and the extracted words can express the main content of the document. For these reasons we chose to use the TF-IDF to extract keywords from transcripts initially, and the results were checked by researchers.

TF indicates the frequency of keywords appearing in selected documents. The higher the TF, the more important the keyword t is to the document D, which can be calculated as

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{1}$$

where $n_{i,j}$ represents the number of occurrences of keyword t in document d_j, and $\sum_k n_{k,j}$ represents the sum of number of occurrences of all words in document d_j.

IDF is the reverse file frequency. If fewer documents are containing keyword t, then the larger the IDF. This condition indicates that keyword t has a "good" classification ability at the level of entire document set. The formula is expressed as

$$idf_i = \log \frac{|D|}{|\{j : t_i \in d_j\}|} \tag{2}$$

where $|D|$ represents total number of all documents in the corpus, and $|\{j : t_i \in d_j\}|$ represents the number of documents containing the word t_i. If t_i is not in corpus, the dividend is zero. $1 + |\{j : t_i \in d_j\}|$ is used as the dividend to avoid this situation.

If one word appears more frequently in a document (that is, TF value is high) and the frequency of occurrence in other documents in the corpus is lower (that is, the IDF value is high), then the word is considered to have better classification ability. The TF-IDF algorithm multiplies the two in the actual calculation, and its calculation formula is

$$tfidf_{i,j} = tf_{i,j} \times idf_i \tag{3}$$

Hu Qi [37] proposed a keyword extraction method that uses the ACE2005 corpus as a data set and which calculates keyword weights from four parts: word frequency, spacing, part of speech, and importance. Accuracy, recall rate, and F-measure evaluation index are used to measure the extraction effect, verifying the effectiveness of one new method. When the number of keywords extracted is seven, the accuracy is the highest. Hu Xuegang et al. [38] used the KEUD algorithm to construct a vocabulary chain and combined

the word frequency feature, location feature, and cluster feature to extract keywords from NetEase News web pages. The accuracy and recall rate of this method are better than those based on TF-IDF only. Choosing nouns and verbs as candidate words improves the accuracy of keyword extraction, and these candidate keywords extracted by TF-IDF are checked manually.

Python's jieba library has implemented keyword extraction technology based on TF-IDF algorithm, and the process follows procedure of Algorithm 1.

Algorithm 1. Keyword extraction algorithm.

Input: original speech draft set SD
Output: keyword list KL
Procedure:

1. set up stop word dictionary $SDict$
2. read speech draft set SD from database, $SD = \{d_1, d_2 \ldots d_i, \ldots d_n\}$ where $\boldsymbol{n}$ = 690, $sd_i = \{id_i, speechtext_i\}$,
3. set $KL = \{\}$, $WL = \{\}$
   for all $d_i \in SD$
4. use Python "jieba" package (accurate pattern) to split $speechtext_i$ to candidate words $wl_i = \{word_1, word_2 \ldots word_i, \ldots\}$, add wl_i to WL
5. endfor
6. for all $wl_j \in WL$
7. for all $word_i \in wl_i$
8. calculate $\boldsymbol{tf_{i,j}}$ in accordance with Equation (1)
9. end for
10. endfor
11. for all $wl_j \in WL$
12. for all $word_i \in wl_i$
13. for all $wl_j \in WL$
14. calculate $\boldsymbol{idf_i}$ in accordance with Equation (2)
15. endfor
16. calculate $\boldsymbol{tfidf_{i,j}}$ according to Equation (3)
17. endfor
18. select 7 keywords that have top 7 $\boldsymbol{tfidf_{i,j}}$ values. create $kl_j = \{word_1, word_2 \ldots word_7\}$
19. add kl_j to KL
20 endfor
21. return KL

One talk with the title of "Philosophy in the Trash" (https://yixi.tv/speech/034 (accessed on 26 December 2021)) is taken as an example. On the basis of previous research results [35], the seven words with the highest $TF - IDF$ value of nouns, noun verbs, and noun form words are shown in Table 1.

Table 1. Keyword extraction from "Philosophy in the Trash".

No.	Keywords	TF-IDF Value
1	rubbish	1.201
2	government	0.199
3	classification	0.178
4	trash can	0.134
5	waste incineration	0.121
6	house	0.114
7	culture	0.111

The $TF - IDF$ value of the word "rubbish" is higher than that of other words, which shows that it is one of the core keywords for this talk. The remaining words are related to the topic of talks, and the core word is expanded. As shown in Table 1, the talk appears to be

mainly about "rubbish", involving government functions and behavior, waste incineration and classification, and other garbage disposal issues. This process may be raised to the cultural level for discussion. The keyword extraction results of transcripts are highly consistent with the actual content by checking the original video and transcripts. This result indicates that keyword extraction progress in this research has a certain accuracy and is suitable for YiXi transcripts, and additional test results can draw similar conclusions.

4.4. Word Cloud Drawing and Interpretation

The procedure of keyword extraction for one transcript can be expanded to a category so that the mining of transcript characteristics for a specific subject category can save readers' time and cost for a specific category. YiXi talks divides 690 talks in 12 categories according to the main content of talk. Keyword extraction expands to categories to check whether the procedure is suitable for categories. Our work takes the "Natural history", which consists of 31 talks, as an example. The result is shown in Table 2, and the "AI" figure is used as the shape of the word cloud to visualize the result. The top 200 words in the "natural history" category from the word cloud are filtered, and the top 7 are shown in Figure 2 (which reveal key information of one category).

Table 2. Keywords of "Natural history" category from YiXi talks.

No.	Keywords	TF-IDF Value
1	animal	0.097
2	zoo	0.085
3	universe	0.054
4	plant	0.050
5	local	0.048
6	ant	0.047
7	first name	0.046

Figure 2. Word cloud diagram for "Natural History category"of YiXi talks.

Combining the results of Table 2 and Figure 2 (Table 2 is the top 7, while Figure 2 is the top 200) we can see that the "natural history" category of YiXi talks mainly focuses on

animals and plants in nature, ranging from zoos to large universes, and species involving ants, elephants, humans, and shells. Research observations show the living environment, habits, characteristics, and other aspects of living things in stories. Thus, keyword extraction and word cloud drawing can be used for other classifications of talks and can make contribution to strategic reading.

5. Talk Pattern Recognition and Analysis

The most popular talk in YiXi has been watched by over 9 million people, which indicates that its content has received widespread public attention. Some users think that YiXi is a door that allows them to see how big the world is. On the basis of good speakers, discovering the common characteristics of their talk patterns can help others to refine the presentation skills (especially newcomers). This paper assumes the opening, ending form, the main narrative objects, and the main narrative angle of YiXi talks as characteristics of transcripts (namely patterns of talks) and conduct experiments by machine learning and manual annotation.

5.1. YiXi Talk Patterns Recognition

A total of 690 transcripts are labeled, manually or automatically. During labeling, manual and feature-based automatic labeling are implemented separately. During the process of manual annotation, the rules are summarized to improve the accuracy of automatic labeling of the program.

One hundred transcripts are randomly selected and manually labeled by three annotators who were trained in advance. Two annotators used back-to-back to annotate fifty transcripts each, and the third was responsible for verifying results of the first two annotators. The manually labeled results were accepted if they are consistent; otherwise, they were removed. The annotated results were given after discussing with each other. The general format is shown in Table 3.

Table 3. Examples of manual labeling format for YiXi talks.

Speech_ID	Opening Form	Feature Words, e.g.	Ending Form	Feature Words, e.g.	Narrative Object	Narrative Angle
S_031_20121201	Self-introduction	I'm	Thoughts	Feel	Thing	First person
S_032_20130112	Memory	More than 70 years ago, when	Hope	Hope	People	First person
S_034_20130330	Small talk	just saw	Quote	: ""	Thing	First person
S_035_20130330	Questioning	?	Outlook	Future	Matter	Third person
S_042_20130818	Self-introduction	My name is	Thoughts	Feel	Thing	First person
S_055_20170219	Self-introduction	My name	Hope	If, I want	Matter	Third person
S_056_20160520	Point out the theme	Title	Outlook	Future, today	People	Third person
S_072_20120826	Small talk	Today, originally	Thoughts	Think, dream	Thing	First person
S_079_20160306	Memory	1998, at that time	Quote	There is a sentence, ""	Thing	First person
S_094_20121028	Point out the theme	Title	Summarize	So, today	People	Third person

The next step is feature extraction of manually labeled information. In accordance with results of manual annotation (Table 3 for example), we summarized rules of opening or ending form (focusing on the first sentence and last three sentences of one transcript), main narrative object, and angle (focusing on full text). The pattern category and feature information are shown in Table 4.

Table 4. Characteristic rules of YiXi talk pattern.

Analysis Perspective	Analysis Scope	Pattern Category e.g.,	Feature Information e.g.,
Opening form	First sentence	Point out the theme Memory Questioning Gossip/small talk Quotation/Exhibition Self-introduction	The title of one talk That year, then, some certain year, some certain time ? Today, just now, now “”, :, photos, poems I am, I come from, my name is
Main Narrative Object	Full text	People Thing Matter	count(nt+nr+he+her)>=count(ns+t+he+her) count(nt+nr+him+her)>=count(an+t+ng+ns+nz+vn+it) count(ns+t+he+her)>=count(nt+nr+he+her) count(ns+t+him+her)>=count (an+t+ng+ns+nz+vn+it) count(an+t+ng+ns+nz+vn+it)>=count(nt+nr+he+her) count(an+t+ng+ns+nz+vn+it)>=count(ns+t+he+her)
MainNarrative angle	Full text	First person Second person Third person	count(me)>=count(you+everyone) count(me)>=count(he+she+it) count(you+everyone)>=count(me) count(you+everyone)>=count(he/she/it) count(he+she+it)>=count(me) count(he+she+it)>=count(you+everyone)
Ending form	last three sentences	Summary Hope Feel/Thoughts Quote Questioning	Title, fact, is, reason, meaning Hope, future, if Feel, think, understand One sentence, one song,: , “”, “” ?, What, ask, how

Feature information of Narrative Object and Narrative angle references People's Daily Annotated Corpus, for example: nt denotes noun and tuan (the initial of “tuan” is t), and nr denotes noun and ren (the initial of “ren” is r). Refer to GitHub and reference for details.

5.2. YiXi Talk Pattern Automatic Recognition

In accordance with the feature rules extracted in Table 3, the remaining 590 records were labeled automatically with help of a program (coded by python). The procedure and core algorithms are shown in Figure 3. In Figure 3, Algorithm 2 gives the overall process of YiXi talk pattern automatic recognition and consists of Algorithms 3–6. Research data mainly come from YiXi title and transcript from database, which was constructed in Section 3.

Algorithm 2. Talk pattern recognition algorithm.

Input: original speech data set *DataSet*
Output: opening form set *OpenForm*, ending form set *EndForm*, main narrative objects set *NaObj*, main narrative angle set *NaAng* of speech set
Procedure:

1. read *DataSet* from database
2. extract the title set of speeches as *TSet*, extract draft set of the speeches as *DSet*
3. for all $title_i$ in *TSet*
4. segment $title_i$ to word set as $wset_i$ by using jieba package in Python
5. add $wset_i$ to *WSet*
6. endfor
7. for all $draft_i$ in *DSet*
8. segment $draft_i$ to sentence set as $sset_i$ by periods and exclamation points
9. add $sset_i$ to *SSet*
10. endfor
11. run Algorithm 3 to produce *OpenForm* and *EndForm*
12. run Algorithm 4 to produce *NaObj* and *NaAng*
13. return *OpenForm*, *EndForm*, *NaObj*, *NaAng*

In the test, all transcripts are extracted from the database, then traverse the text collection, and segment the transcript by period and exclamation. Next, Chinese word segmentation is performed for the title to form a word list, and the next steps are as follows:

1. Algorithm 3 provides a method to identify the opening from of YiXi talks. First extract the first sentence of transcript and perform Chinese word segmentation. This process is performed to match the title segment word list for determining whether the word segmentation result contains the opening feature word or the word in talk name. If included, mark the corresponding opening form.
2. Extract the last three sentences in the clause set, perform Chinese word segmentation, and determine whether the word segmentation result contains the ending feature word or words from title. If included, mark the corresponding ending form. The details of this procedure are shown in Algorithm 4.
3. Perform Chinese word segmentation for all clauses, and count the number of pronouns "he/them", "she/them", and "it/them", and mark the main narrative angle in accordance with the numerical comparison results. This procedure is shown in Algorithm 5.
4. Perform Chinese word segmentation for all clauses, identify part of speech of the word, and calculate the amount with part of speech. The part of word includes NT (institutional groups noun and the initial of "tuan" is t), NR (name, noun, and the initial of "ren" is r), NG (nominal morpheme), NS (place name), NZ (other proper names), T (time), VN (verbal nouns), and AN (adjectival nouns). The main narrative angle is marked in accordance with the numerical comparison results. This procedure is shown in Algorithm 6.

The details of Algorithms 3–6 are discussed in the following sections.

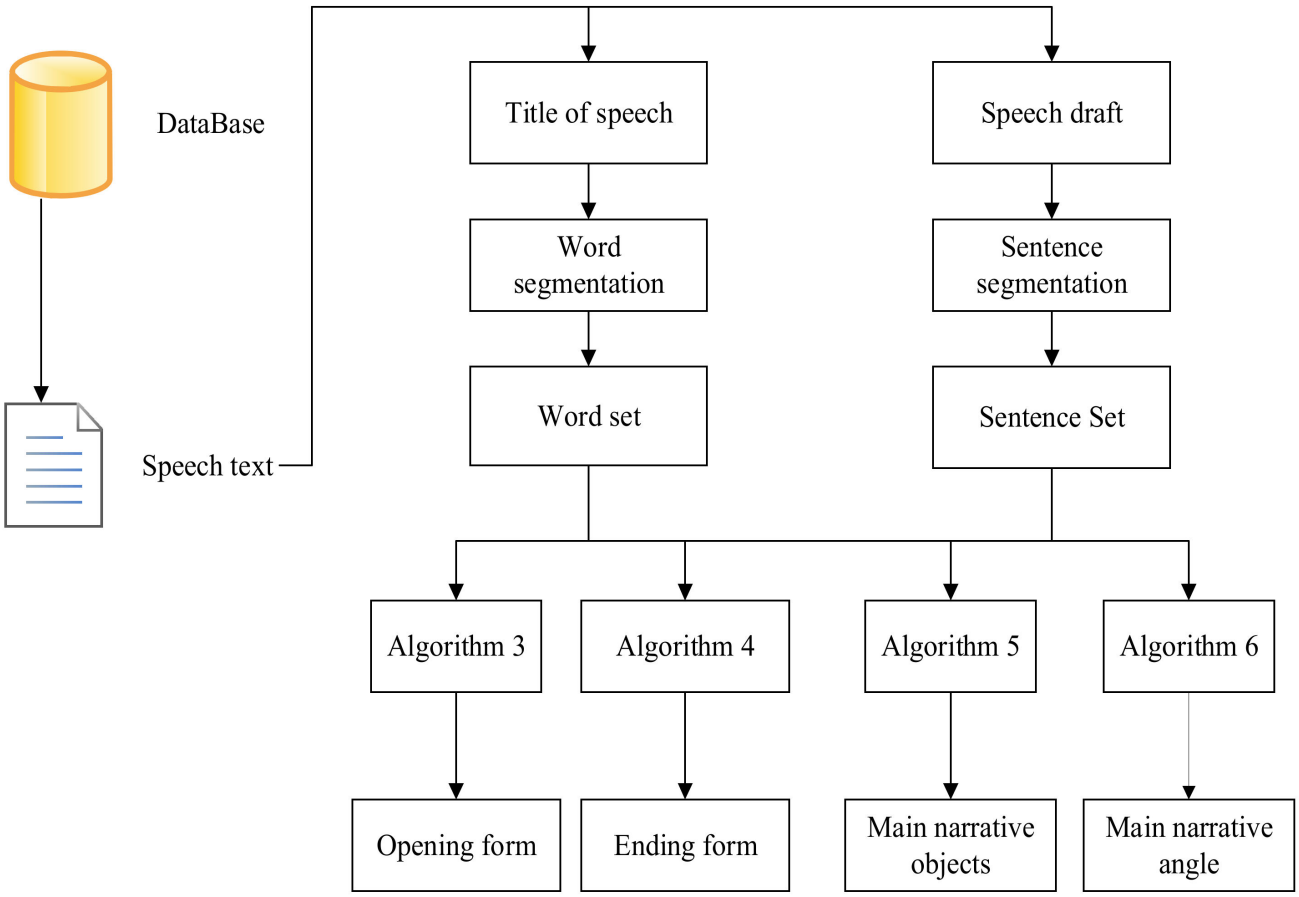

Figure 3. Program flowchart of YiXi talk pattern automatic labeling.

Algorithm 3. Talk opening form analysis.

Input: talk title word set *WSet*, sentence set *SSet*
Output: opening form set *OpenForm*
Procedure:

read *WSet* and *SSet*
construct feature rules set for opening form as *OfSet* according to Table 3.
$OfSet = \{'Pattern'_1 : feature_1 \ldots' Pattern'_i : feature_i \ldots\}$ includes the title words in *WSet*
set *OpenForm* = {}
for all $sset_i$ in *SSet*
  set $OpenForm_i$ = ''
  extract first sentence in $sset_i$ as $sset_{i0}$
  segment $sset_{i0}$ to a word set *SSetW*
  for all $word_i$ in *SSetW*
    if $OpenForm_i$ == ''
      for all $feature_i$ in *OfSet*
        if $word_i$ in $feature_i$, $OpenForm_i =' Pattern'_i$, end of search loop of $sset_i$, go to the next searchloop for $sset_{i+1}$
        endif
      endfor
    endif
  endfor
  add $OpenForm_i$ to *OpenForm*
endfor
return *OpenForm*

Algorithm 4. Talk ending form analysis.

Input: talk title word set *WSet*, talk sentence set *SSet*
Output: ending form set *EndForm*
Procedure:

read *WSet* and *SSet*
construct feature rules set for ending form as *EfSet* according to Table 3.
$EfSet = \{'Pattern'_1 : feature_1 \ldots' Pattern'_i : feature_i \ldots\}$ includes the title words in *WSet*
set *EndForm* = {}
for all $sset_i$ in *SSet*
  set $EndForm_i$ = ''
  extract last three sentences in $sset_i$ as $last_i = \{sset_{in-2}, sset_{in-1}, sset_{in}\}$
  segment each element in $last_i$ to word list as $lastw_i = \{ssetw_{in-2}, ssetw_{in-1}, ssetw_{in}\}$
  for all $ssetw_i$ in $last_i$
    for all $word_i$ in $ssetw_i$
      if $EndForm_i$ == ''
        for all $feature_i$ in *EfSet*
          if $word_i$ in $feature_i$, $EndForm_i =' Pattern'_i$, end of search loop of $sset_i$ go to the next search loop for $sset_{i+1}$
          endif
        endfor
      endif
    endfor
    add $EndForm_i$ to *EndForm*
endfor
return *EndForm*

5.3. Analysis of YiXi Talk Pattern

Systematic sampling is adopted, and one out of five records was selected for manual labeling after talk patterns were automatically recognized. This process is performed to validate results of automatic labeling. A total of 118 records are randomly selected and manually labeled to verify the accuracy of automatic recognition by matching the results of manual recognition and automatic program determination, and statistical results are shown in Table 5.

Algorithm 5. Main narrative object analysis.

Input: Transcript sentence set *SSet*
Output: main narrative objects set *NaObj*
Procedure:

read *SSet*
construct feature rules set for narrative objects as *NbSet* according to Table 3.

$$NbSet = \{'People' : [taglist_1, wordlist_1], 'Thing' : [taglist_2, wordlist_2], \\ 'Matter' : [taglist_3, wordlist_3]\}$$

for all $sset_i$ in *SSet*
  set *narraObj* = ''
  segment $sset_i$ to word list with tags $SSetWt = \{(word_1, tag_1) \ldots (word_i, tag_i) \ldots\}$
  for all $word_i, tag_i$ in *SSetWt*
    counterP = counterP + 1 if $word_i$ in $worlist_1$ or tag_i in $taglist_1$
    counterT = counterT + 1 if $word_i$ in $worlist_2$ or tag_i in $taglist_2$
    counterM = counterM + 1 if $word_i$ in $worlist_3$ or tag_i in $taglist_3$
  endfor
  if counterP >= counterT and counterP >= counterM, *narraObj* = 'People'
  endif
  if counterT >= counterP and counterT >= counterM, *narraObj* = 'Thing'
  endif
  if counterM >= counterP and counterM >= counterT, *narraObj* = 'Matter'
  endif
  add *narraObj* to *NaObj*
endfor
return *NaObj*

Algorithm 6. Main narrative angle analysis.

Input: Transcript sentence set *SSet*
Output: main narrative angles set *NaAng*
Procedure:

read *SSet*
construct feature rules set for narrative objects as *NaSet* according to Table 3

$$NaSet = \{'FirstPerson' : wordlist_1, 'SecondPerson' : wordlist_2, \\ 'ThirdPerson' : wordlist_3\}$$

for all $sset_i$ in *SSet*
  set *narraAng* = ''
  segment $sset_i$ to word list $SSetW = \{word_1, word_2 \ldots word_i, \ldots\}$
  for all $word_i$ in *SSetW*
    counter1 = counter1 + 1 if $word_i$ in $worlist_1$
    counter2 = counter2 + 1 if $word_i$ in $worlist_2$
    counter3 = counter3 + 1 if $word_i$ in $worlist_3$
  enfor
  if counter1 >= counter2 and counter1 >= counter3, *narraAng* = 'FirstPerson'
  endif
  if counter2 >= counter1 and counter2 >= counter3, *narraAng* = 'SecondPerson'
  endif
  if counter3 >= counter1 and counter2 >= counter2, *narraAng* = 'ThirdPerson'
  endif
  add *narraAng* to *NaAng*
endfor
return *NaAng*

Table 5. Accuracy of program labeling in talk pattern.

Labeled Content	Number of Matches	Total Number of Samples	Accuracy (%)
Opening form	77	118	65.25
Main narrative angle	83	118	70.34
Main narrative object	63	118	53.39
Ending form	58	118	49.15
Total	281	472	59.53

As shown in Table 5, the labeling accuracy rate of "main narrative angle" recognition is the highest (70.34%), followed by "opening form" (65.25%), with "main narrative object" and "ending form" being around 50%. Therefore, the overall accuracy of the rule-based automatic recognition program for talk patterns of our research is approximately 60%.

This paper identifies 690 talks patterns by manual labeling and a rule-based program. The recognition angle includes perspectives of opening form, narrative angle, narrative object, and ending form to summarize the talk pattern of YiXi, which may be referenced by related fields or speakers.

(1) Clustered histogram of opening form of YiXi talk is shown in Figure 4. Figure 4 shows that opening of "self-introduction" is more popular than other types. YiXi speakers tend to introduce themselves briefly at the beginning. "Point out the theme", "memory", and "gossip" opening form are secondary choices for openings. From a linguistic and psychological point of view in China, "Point out the theme" helps the audience understand the main content of a talk, while "memory" and "gossip" may ease speakers' tension through simple communication and interaction. The "questioning" and "citation" openings are the forms speakers rarely used.

(2) Clustered histograms of main narrative angles and objects of YiXi are shown in Figures 5 and 7. YiXi encourages sharing of insights, experiences, and imaginations about the future. Speakers usually use specific examples to share unique experiences, knowledge, and opinions. Therefore, the main content is mainly about things and contains fewer descriptions of people and of matter. Compared with the second and third person, the first person may be better which is more convenient for the elaboration and expression of the content.

(3) Clustered histogram of the ending form of YiXi talk is shown in Figure 6.

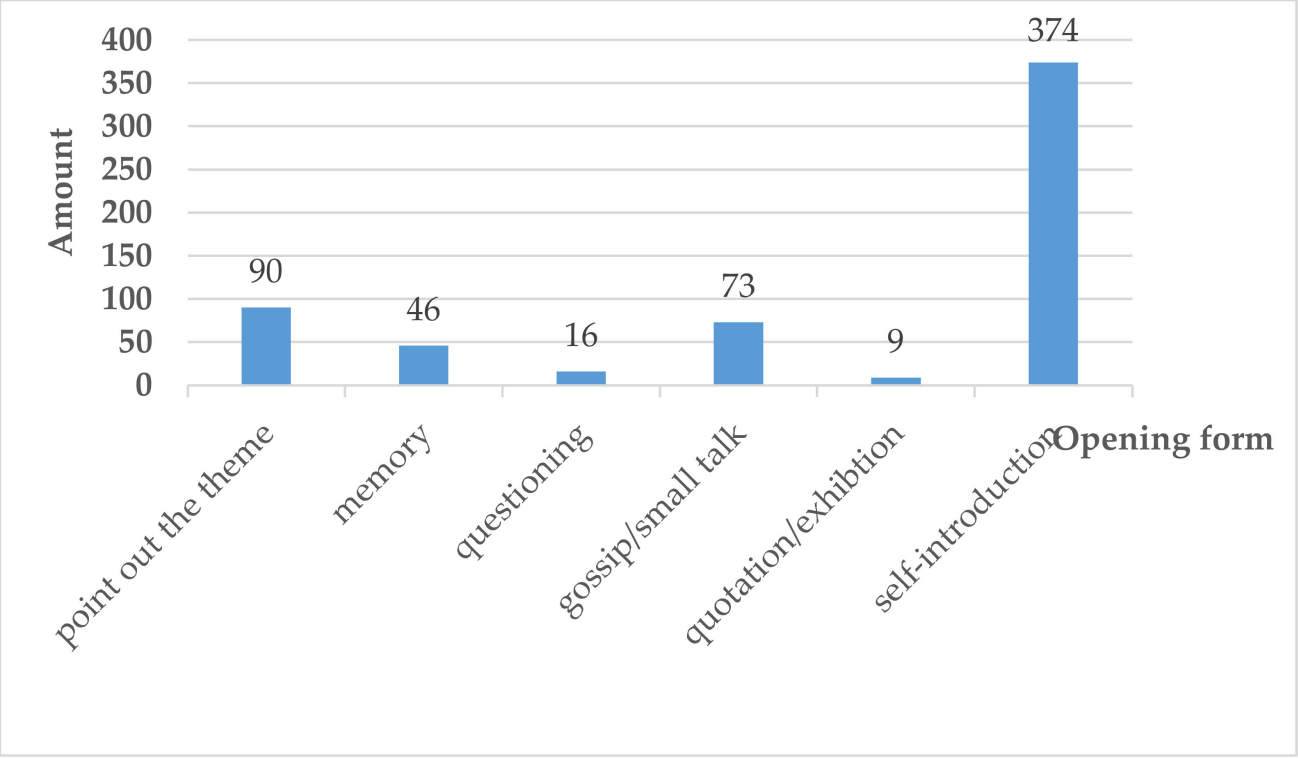

Figure 4. Clustered histogram of opening form of YiXi talk.

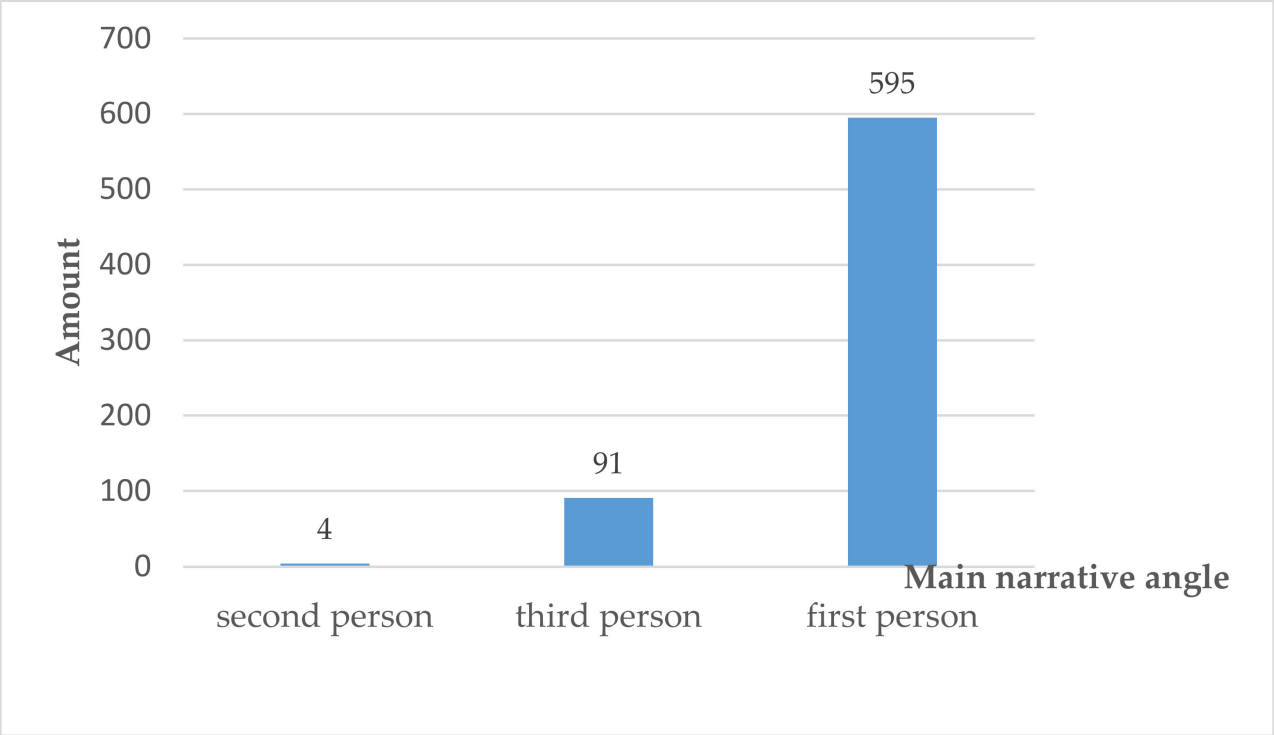

Figure 5. Clustered histogram of the main narrative angle of YiXi talk.

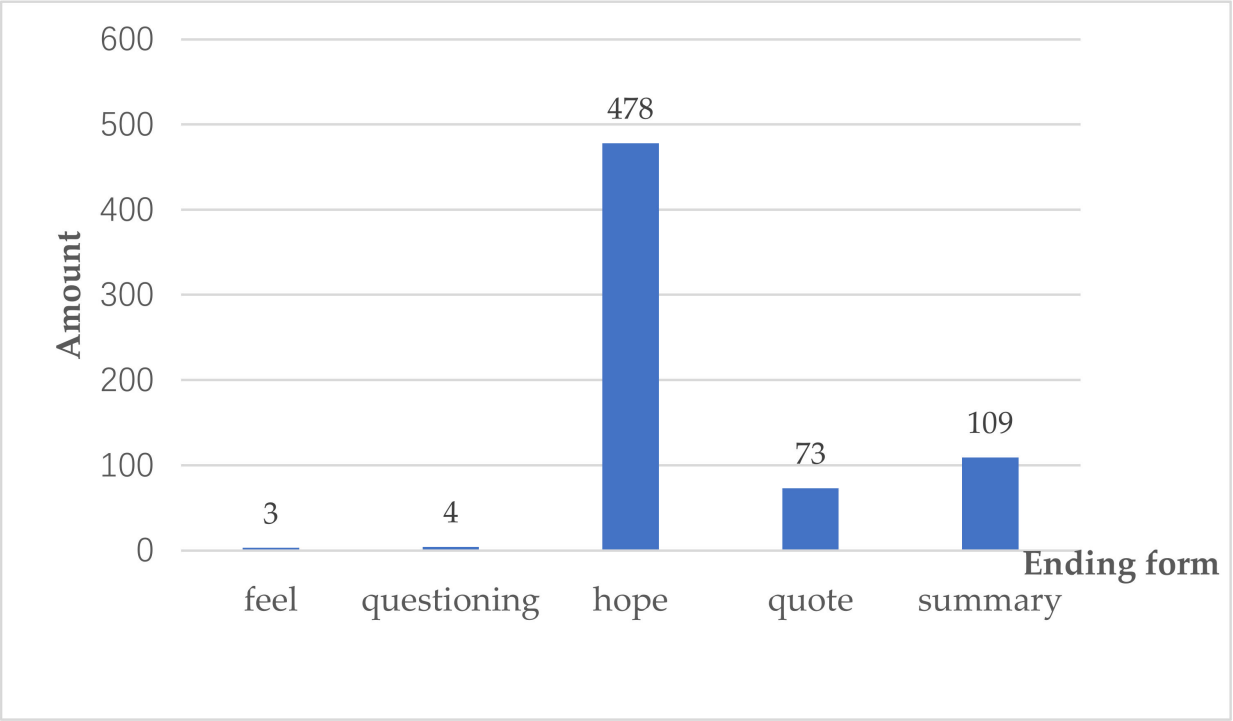

Figure 6. Clustered histogram of the ending form of YiXi talk.

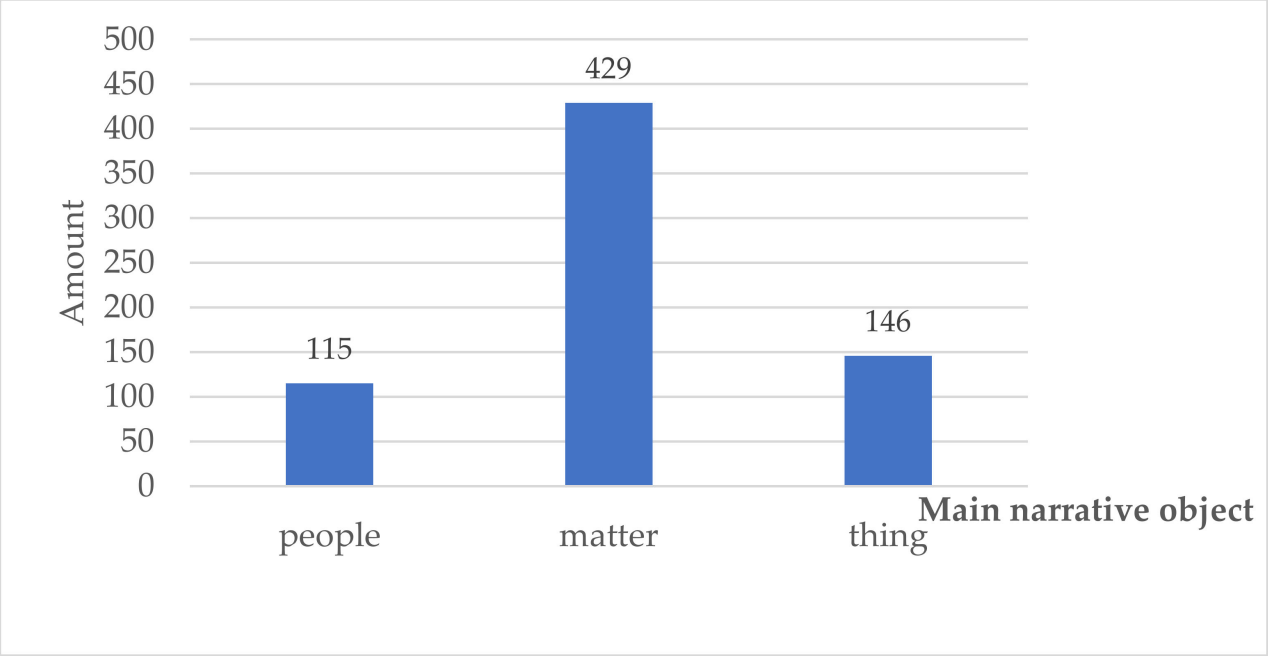

Figure 7. Clustered histogram of the main narrative object of YiXi talk.

"Hope" is the most common form of endings, followed by "quote" and "summary," "questioning" or "expressing feelings". A hopeful expression at the end may reveal the speaker's ardent expectations for things, looking forward to the future, and a highly generalized sublimate of the talk. A relatively pattern of a successful talk can be summarized as follows: firstly, the speaker introduces himself to audience, then uses the first-person narrative perspective, clear logic, and concise content to produce a lively and interesting storyline. Finally, the speaker summarizes by expressing expectations, and focusing on a two-way interaction with the audience and motivating them to obtain more information beyond the content are equally important.

6. Conclusions

This paper uses YiXi talk transcripts as data source and utilizes text mining and feature recognition methods. The textual characteristics of YiXi transcripts are identified from key information extraction and talk pattern recognition. The main research conclusions are as follows:

(1) The key information of YiXi talk transcripts can be revealed by text mining methods, which were followed by Chinese word segmentation, stop word removal, and keyword extraction. Certain category characteristics are reflected by extraction results. The technology of keyword extraction can make contributions to "strategic reading" to some degree. The research results can solve the general problem that keywords missing of transcripts, and we advocate that speakers give a summary and recommendation on the topic of talks to ensure the effect of knowledge diffusion.

(2) YiXi talks has fixed talk patterns and characteristics, that is, speakers tend to adopt a self-introducing opening form, tell the story and experience through the first person narrative angle, and express expectations and prospects for the future at the end. This pattern can provide a reference for newcomers. However, the talk pattern analysis implemented in our work has certain exploratory characteristics, and the angle only includes the opening mode, narrative angle, narrative object, and end form categories. The analysis perspectives are few, and the characteristic refinement of manual data labeling and the logic of program labeling still need to be optimized.

Subsequent related research will be further studied, and more languages and categories, such as TED Talks and other language talks, will be continuously expanded and studied in the future. Expanding keyword extraction algorithms and comparing the results of keyword extraction with precision and recall instead of a manual check to give an automated suggestion. Expanding the angle and depth of talk pattern analysis and the process will reduce manual involvement. The application of machine learning methods in talk pattern recognition and the key information extraction process will be also explored. At the same time, the expression of Chinese talk has a certain randomness, and the characteristics of Chinese are ideographic will be considered in the future.

Author Contributions: Conceptualization, H.X. and Z.Z.; formal analysis, C.J. and C.H.; data curation C.J. and M.Y.; visualization, Y.C.; methodology, Y.C., C.J. and H.X.; software, Y.C. and H.X.; writing—original draft, H.X. and Z.Z; writing—review and editing, H.X., Y.C., C.J. and Z.Z. All authors have read and agreed to the published version of the manuscript.

Funding: The research was supported by JIANGSU PROVINCIAL SOCIAL SCIENCE FOUNDATION YOUTH PROJECT: Grant number 21TQC003; the INNOVATION FUND GENERAL PROJECT I OF NANJING INSTITUTE OF TECHNOLOGY, Grant number CKJB202003; JIANGSU PROVINCE EDUCATION SCIENCE "14TH FIVE-YEAR PLAN" 2021 ANNUAL PROJECT, Grant number C-c/2021/01/62; MAJOR PROJECT OF PHILOSOPHY AND SOCIAL SCIENCE RESEARCH IN UNIVERSITIES OF JIANGSU PROVINCIAL DEPARTMENT OF EDUCATION, grant number 2020SJZDA069; and the TEACHING REFORM AND CONSTRUCTION PROJECT OF NANJING INSTITUTE OF TECHNOLOGY, Grant number JXGG2021031.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Hu, P.P. Innovation and Improvement of the Speech Program "Yi Xi". *Radio TV J.* **2018**, *11*, 16–18.
2. Hou, Z.; Huang, C.; Wu, J.; Liu, L. Distributed Image Retrieval Base on LSH Indexing on Spark. In *Big Data and Security. ICBDS 2019. Communications in Computer and Information Science*; Tian, Y., Ma, T., Khan, M., Eds.; Springer: Singapore, 2020; Volume 1210, pp. 429–441.
3. Turney, P.D. Learning Algorithms for Key phrase Extraction. *Inf. Retr.* **2000**, *2*, 303–336. [CrossRef]
4. Liu, Z.; Huang, W.; Zheng, Y. Automatic Key Phrase Extraction via Topic Decomposition[C]. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, Cambridge, MA, USA, 9–11 October 2010; Association for Computational Linguistics: Stroudsburg, PA, USA, 2010; pp. 366–376.
5. Zhang, C.Z. Review and Prospect of Automatic Indexing Research. *New Technol. Libr. Inf. Serv.* **2007**, *2*, 33–39.
6. Zhao, J.S.; Zhu, Q.M.; Zhou, G.D.; Zhang, L. Review of Research in Automatic Keyword Extraction. *J. Softw.* **2017**, *28*, 2431–2449.
7. Merrouni, Z.A.; Frikh, B.; Ouhbi, B. Automatic Keyphrase Extraction: A Survey and Trends. *J. Intell. Inf. Syst.* **2020**, *54*, 391–424. [CrossRef]
8. Papagiannopoulou, E.; Tsoumakas, G. A Review of Keyphrase Extraction. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2020**, *10*, 1339. [CrossRef]
9. Chen, Y.; Yang, Z.; Ye, Z.; Liu, H. Research Character Analyzation of Urban Security Based on Urban Resilience Using Big Data Method. In *Big Data and Security. ICBDS 2019. Communications in Computer and Information Science*; Tian, Y., Ma, T., Khan, M., Eds.; Springer: Singapore, 2020; Volume 1210, pp. 371–381.
10. Hu, S.H.; Zhang, Y.Y.; Zhang, C.Z. Overview of Keyword Extraction Research. *Data Anal. Knowl. Discov.* **2021**, *3*, 45–59.
11. Cohen, J.D. Highlights: Language and Domain-Independent Automatic Indexing Terms for Abstracting. *J. Am. Soc. Inf. Sci.* **1995**, *46*, 162–174. [CrossRef]
12. Salton, G.; Yang, C.S.; Yu, C.T. A Theory of Term Importance in Automatic Text Analysis. *J. Am. Soc. Inf. Sci.* **1975**, *26*, 33–44. [CrossRef]
13. Luhn, H.P. A Statistical Approach to Mechanized Encoding and Searching of Literary Information. *IBM J. Res. Dev.* **1957**, *1*, 309–317. [CrossRef]
14. Matsuo, Y.; Ishizuka, M. Keyword Extracyion from a Single Document Using Word Co-occurrence Statistical Information. *Int. J. Artif. Intell. Tools* **2008**, *13*, 157–169. [CrossRef]
15. Zhao, H. A Survey of Deep Learning Methods for Abstractive Text Summarization. *J. China Soc. Sci. Tech. Inf.* **2020**, *39*, 330–344.
16. Kourou, K.; Exarchos, T.P.; Exarchos, K.P.; Karamouzis, M.V.; Fotiadis, D.I. Machine learning applications in cancer prognosis and prediction. *Comput. Struct. Biotechnol. J.* **2015**, *13*, 8–17. [CrossRef]

17. Srivastava, A.; Jain, S.; Miranda, R.; Patil, S.; Pandya, S.; Kotecha, K. Deep learning based respiratory sound analysis for detection of chronic obstructive pulmonary disease. *PeerJ Comput. Sci.* **2021**, *5*, 1–22. [CrossRef]
18. Asghar, M.Z.; Rahman, F.; Kundi, F.M.; Ahmad, S. Development of stock market trend prediction system using multiple regression. *Comput. Math. Organ. Theory* **2019**, *25*, 271–301. [CrossRef]
19. Ullah, A.; Asghar, M.Z.; Habib, A.; Aleem, S.; Kundi, F.M.; Khattak, A.M. Optimizing the Efficiency of Machine Learning Techniques. In *Big Data and Security. ICBDS 2019. Communications in Computer and Information Science*; Tian, Y., Ma, T., Khan, M., Eds.; Springer: Singapore, 2020; Volume 1210, pp. 429–441.
20. Zhao, H.; Wang, F.; Wang, X.Y.; Zhang, W.C.; Yang, J. Research on Construction and Application of a Knowledge Discovery System Based on Intelligent Processing of Large-scale Governmental Documents. *J. China Soc. Sci. Tech. Inf.* **2018**, *37*, 805–812.
21. Kavita, S.O.; Poornima, G.N. Prediction of Online Lectures Popularity: A Text Mining Approach. *Procedia Comput. Sci.* **2016**, *92*, 468–474.
22. Zhou, H.Y. Finely crafted live-speaking documentaries -text analysis of Under the Dome[C]. In *A Collection of Papers from the 7th Graduate Forum on Journalism and Communication in Anhui Province*; School of Journalism and Communication, Anhui University: Hefei, China, 2015; pp. 346–365.
23. Min, Y.; Lao, Q.; Ding, X.J. Enlightenment of Text Data mining Technology on Shorthand Teaching and Corpus Construction-Taking Konosuke Matsushita's Speech Data Analysis as an Example. *J. Shaoguan Univ.* **2015**, *7*, 170–174.
24. Liu, H. Word Frequency Analysis on the Construction of British National Image-Taking the Speech of Former British Prime Minister Theresa May as an Example. *Overseas Engl.* **2020**, *1*, 215–216, 245.
25. Chen, Y.; Wang, L.; Qi, D.; Zhang, W. Community Detection Based on DeepWalk in Large Scale Networks. In *Big Data and Security: First International Conference, ICBDS 2019 [C]*; Tian, Y., Ma, T.H., Khan, M.K., Eds.; Springer: New York, NY, USA, 2020; pp. 568–583.
26. Shan, X.Q. Analysis of New Media Communication Characteristics of YiXi Speech. *Res. Transm. Competence* **2020**, *4*, 79–80.
27. Zhao, Z.Q. Let "Hero Dreams" Come into the Eyes of More People-Analysis of the Market Operation of the Network Speech Program "YiXi". *J. News Res.* **2017**, *8*, 229.
28. Alwabel, A. Privacy Issues in Big Data from Collection to Use. In *Big Data and Security. ICBDS 2019. Communications in Computer and Information Science*; Tian, Y., Ma, T., Khan, M., Eds.; Springer: Singapore, 2020; Volume 1210, pp. 382–391.
29. Monali, B.; Saroj, K.B. *Keyword Extraction from Micro-Blogs Using Collective Weight*; Springer: Vienna, Austria, 2018; Volume 8, p. 429.
30. Saroj, K.B.; Monali, B.B.; Jacob, S. A graph based keyword extraction model using collective node weight. *Expert Syst. Appl.* **2018**, *97*, 51–59.
31. Ou, Y.Y.; Ta, W.K.; Anand, P.; Wang, J.F.; Tsai, A.C. Spoken dialog summarization system with happiness/suffering factor recognition. *Front. Comput. Sci.* **2017**, *11*, 15. [CrossRef]
32. Tang, L.; Guo, C.H.; Chen, J.F. Summary of Research on Chinese Word Segmentation Technology. *Data Anal. Knowl. Discov.* **2020**, *4*, 1–17.
33. Yu, H.; Li, Z.X.; Jiang, Y. Using GitHub Open Sources and Database Methods Designed to Auto-Generate Chinese Tang Dynasty Poetry. In *Big Data and Security. ICBDS 2019. Communications in Computer and Information Science*; Tian, Y., Ma, T., Khan, M., Eds.; Springer: Singapore, 2020; Volume 1210, pp. 417–428.
34. Zheng, J.H.; Zhang, J.F.; Tan, H.Y. Strategies for Segmentation of Ambiguity in Chinese Word Segmentation. *J. Shanxi Univ. (Nat. Sci. Ed.)* **2016**, *47*, 228–232.
35. Passing, F.; Moehrle, M.G. Measuring technological convergence in the field of smart grids: A semantic patent analysis approach using textual corpora of technologies. In Proceedings of the 2015 Portland International Conference on Management of Engineering and Technology (PICMET), Portland, OR, USA, 2–6 August 2015; pp. 559–570.
36. Qaiser, S.; Ali, R. Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents. *Int. J. Comput. Appl.* **2018**, *181*, 25–29. [CrossRef]
37. Hu, Q.; Hao, X.Y.; Zhang, X.Z.; Chen, Y.W. Keyword Extraction Strategy Research. *J. Taiyuan Univ. Technol.* **2016**, *47*, 228–232.
38. Hu, X.G.; Li, X.H.; Xie, F.; Wu, X.D. Keyword Extraction Method of Chinese News Web Pages based on Vocabulary Chain. *Pattern Recognit. Artif. Intell.* **2010**, *23*, 45–51.