Research on the Motif of Chinese Science Fiction Literature Based on Big Data Mining

Sixin Zhu(✉)

College of Literature and Journalism, Sichuan University, Chengdu, China
sigrid_zsx@163.com

Abstract. Data mining is an important technology in the field of big data. In addition to being applied to artificial intelligence, finance and other fields, data mining can also be applied to the research of humanities and social sciences, and has high application value. China has entered the era of big data. In the field of humanities and social sciences, the state has also actively introduced relevant policies to promote the development of digital humanities. In this paper, the text clustering technology in data mining technology is adopted to conduct cluster analysis on a wide range of Chinese sci-fi literature works that are difficult to be fully analyzed, and then the motif is summarized, which helps to promote quantitative research into the field of traditional literature qualitative research, provide objective data support for literature research, and promote the process of "literature digitization".

Keywords: Big data mining · LDA topic model · Motif · Semantic cohort · Chinese science fiction

1 Introduction

Subject extraction is a research hotspot in the field of big data and NLP, which includes LDA (Latent Dirichlet Allocation) model and its variants ATM (Author-Topic Model) model and DTM (Dynamic Topic Models) model, etc. LDA model is mainly used for text topic analysis, which can express the content preference of articles and feature words. ATM model adds the "author" factor to LDA model, and the output includes new content such as the author's topic preference and similar author recommendations. DTM model adds a "time" factor, which can show the trend of topic change of articles over time, and also predict the topic preference of new articles. Each of the three models has its own advantages and is applicable to different research occasions.

"Digital Humanities" has become a new trend in the field of literary research in recent years, and China is also introducing policies to promote digital literature, so literature research and big data also intersect. Some scholars have attempted to use big data technology to research literary issues, mostly for texts with a high degree of objectivity, such as news reviews and book evaluations. In recent years, a few scholars have also used LDA model in big data technology to research literary "theme" and have obtained convincing data. However, the meaning of "topic" in LDA model differs greatly

G. Guan et al. (Eds.): ICBDSS 2022, AHCS 8, pp. 99–109, 2023.
https://doi.org/10.2991/978-94-6463-064-0_12

from “theme” in literature research, and it is inappropriate to directly equate the two. However, it doesn’t mean LDA model isn’t suitable for literature research. It is valuable to combine traditional literary research with big data technology to bring quantitative research into the field of literary research and to provide scientific and objective data to support literary research.

I want to use LDA model to find and summarize the motifs of Chinese science fiction literature, and then excavate the development process of Chinese science fiction literature behind these motifs.

2 Concepts and Analysis of Research Rationality

2.1 Motif and Theme

Motif is an imported word, derived from the Latin word “motivere”, meaning movement, motive. Before its use in literary studies, it was used in the fields of painting and musical art to denote a characteristic element of content or the smallest melodic unit. Motif was later borrowed from the field of literature to meet the needs of studies related to the taxonomy of folk tales. It is generally accepted that J. Kohler, a German scholar, first introduced the concept of the motif.

The American scholar S. Thompson, in his The Folktale, has been widely influential in academic circles in defining and classifying motif as follows.

A motif is the smallest element of a story that can last in tradition. To do so it must have some unusual and moving power. Most motifs are divided into three categories. The first is the role in a story. Or even traditional characters. The second kind of motif involves a certain background of the plot. The third kind of motifs are those single events [1].

Through Thompson’s definition of a motif, it is clear that the motif as the smallest unit in a literary work recurs in different works, and that the expression of the same motif in different works may change with the work and time, but the core meaning of each motif doesn’t change. At the same time, Thompson’s classification of motifs: character, environment, story, has been followed by many scholars.

Wenxian Sun summed up the characteristics of motif as follows: “motif must appear repeatedly in different texts in a typological structure or stylized speech form. The structural form or linguistic form that is constant and can be recognized is an important feature of the motif.” [2].

Based on the viewpoints of two scholars, the author defines the motif as follows.

The motif is the smallest unit of meaning that can recur in a text, is structured, and contains a certain plot or background. It is objective in nature and does not contain the author’s subjective emotions.

Zhaoyi Meng defines “theme” as follows.

The theme is the highly concentrated ideological crystallization obtained by refining the theme of the work and shaping the image… Themes are subjective to authors and readers [3].

It is clear from this that the author, makes the abstract theme matter visible in his or her thinking, implying it in the text. It is impossible for the reader to know the author’s subjective intentions, and even the reader's grasp of the theme of the same work may vary considerably from reader to reader. It is therefore difficult to define the scope of a subjective theme and to prove that a work falls within it, so a purely quantitative approach to the study of literary themes is in fact ill-conceived.

2.2 Reasonableness of LDA Model Application

The LDA model will output n "topics" and m feature words under each topic when extracting topics. Look at the definition of "topic", "A subject that you talk, write or learn about". Rethinking the definition of "subject", "A thing or person that is being discussed, described or dealt with". The "topic" output of the model is in fact a "subject" in a broad sense, which can refer to something central to a conversation, a person, etc. But such a "topic" is not the same as a "theme" in literary studies. So we can think of topics as semantic cohorts, which contain a certain plot or context. Thus, the feature words included in each semantic cohort reflect more the content of the text than the abstraction of the author's thoughts.

The theme itself is subjective in nature, so the generalization of the theme must also involve subjective thoughts, and the LDA model is difficult to uncover the author's thoughts hidden under the text, so it is inappropriate to use this model to study literary themes only. However, since the LDA model can output feature words that are highly relevant to the content of the text, and each topic can reflect a certain plot or background, the LDA model is suitable for use in the study of the motif. It can generalize the content of each article and help the researcher to better identify multiple plots in the text, as well as help the researcher not to drift into a certain direction when analyzing the feature words to grasp the content of the article. It can also provide objective support for the final summary of the motif.

Although Chinese science fiction literature has a relatively short history, it has its own unique elements that allow it to exist as a new genre of literature. By using the LDA model to analyze specific works and identify the unique motifs, we can fill the gaps in the study of science fiction literature.

3 Process of Research

The author has tentatively selected some representative writers' works written from 1976 to 1989 for analysis.

After the collection of works selected, the author first carries on the OCR processing, and respectively stored as text and form. Then use jieba library to segment the text and remove common stop words. After the preprocessing, the text is input into the LDA topic model for subject extraction, and the code parameters are repeatedly adjusted, the stop words are added, and the required words are added. Finally, six semantic cohorts are selected to output, and each semantic cohort has 10 feature words.

The corresponding semantic cohorts and feature words lists of science fiction works from 1976 to 1989 are as follows (Table 1).

According to the visualization of the results of the LDA model, it can be seen that there is no overlap between each semantic cohort, proving that the subject extraction is effective (Fig. 1).

Table 1. Semantic cohorts and feature words

Sematic cohorts	Feature words
Topic #0:	Robot, Father, Computer, Teacher, Grandpa, Artificiality, Uncle, Space, Machine, Screen
Topic #1:	Master, Moonlight, Archaeological, Experiment, Diary, Teacher, Survey, Marine, Hotel
Topic #2:	Continents, Ships, Tents, Expeditions, Crews, Sea, Water, Motherland, Ice fields, Docks, Earthquakes
Topic #3:	War, Stellar, Spaceship, Engineer, Star, Mars, Laser, Energy, Base, Commander-in-chief
Topic #4:	Spaceship, Submarine, System, Doctor, Seawater, Engineer, Language, Captain, Check, Sea
Topic #5:	Alien, Rocket, Spaceship, Archaeology, Signal, Information, Planet, Code, Space, Microwave

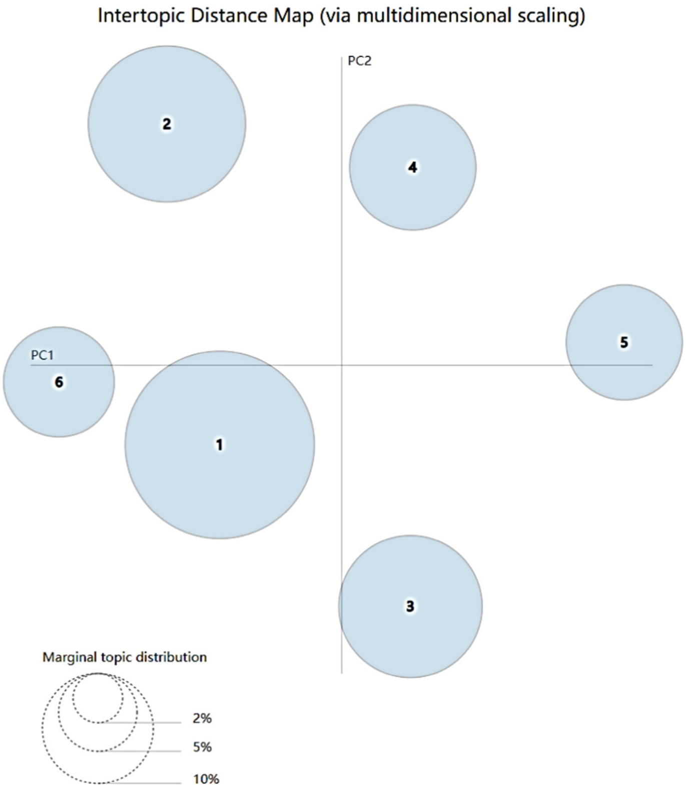

Fig. 1. Visualization diagram of LDA model results

3.1 Semantic Cohort Analysis

In semantic cohort one, words such as robot, computer, artificiality and space are all related to technology. In the context of the time period, these words all represent the high technology of the time and are closely related to new technology and inventions. Robots are mostly found in works of futuristic imagination, such as robots that can carry goods. However, in Yahua Wei's *The Dream of a Soft Country*, it is a rare exploration ahead of its time to discuss the ethical issues between robots and humans. Computers appear in similar contexts to robots, often representing technology from the future, such as computers that can control the devices of an entire house. The same is true of artificiality, which often appears in episodes depicting new technologies, such as "artificial skies". The description of space in this period is part of a simple imagination of the future, describing the image of space in imagination, such as Yonglie Ye's *Xiaolingtong Roaming the Future.* There are also some works in space as the background, the description of war, the truth between people. Such as Wenguang Zheng's *Fly to Horse Constellation.*

The semantic cohort also includes words such as father and teacher, which are often used in the context of explaining new inventions to children. Overall, this semantic cohort describes scenes in which knowledgeable elders introduce new technologies to children, and lead them to experience new inventions. Therefore, the works under this semantic cohort are still biased towards the purpose of popular science, which also reflects the diversity of works in the "exploration period" of science fiction literature in China. Therefore, the core of semantic cohort one should take the person as the carrier to show high technology.

The first feature word of semantic cohort two is captain, which also contains words like archaeology and experiment, pointing to works that have a certain ideology while popularizing scientific knowledge in the context of exploration and archaeology. For example, Jiazi Cheng's *The Mystery of Ancient Star Map*, in the use of high-tech to break the copper ball left by aliens, it also conveys the idea of protecting natural ecology through the description of alien civilization. Semantic cohort two also contains words such as moonlight and experiment, and is linked to Tao Jin's *The Moonlight Island*, which tells the story of a scientific research team of advanced alien civilization, the Sirius people, who descend on the moonlight island and pretend to be humans to observe Earth's civilization. This kind of work has formed the bright contrast with the Republic early advocated "the human determination wins the heaven" thought work, reflected the humanity in this time regarding the self-reflection. Therefore, the core of semantic cohort two is human introspection besides science and technology.

The feature words in semantic cohort three are related to science research, adventure and nature. The works of science fiction in the new period aiming at popularization of science, excluding those works that use narration, demonstration and visit of high-tech inventions to popularize science, are mostly set in the context of adventure and section, in which new technologies are applied to solve problems in the process of adventure, often enabling the discovery of new things and achieving the purpose of popularization of science in a slightly tense atmosphere. For example, Tao Jin's *Iceland Missing* and Ke Lu's *BB-1.*

Semantic cohort five can be analyzed in conjunction with semantic cohort three. The two semantic cohorts are similar in that both reflect the context of the section, but differ in that semantic cohort five is more oriented towards the equipment and staffing of the expedition, whereas semantic cohort three is more oriented towards the process of the expedition and the external environment. A more subtle difference is the presence of words such as “spaceship” and “engineer” in semantic cohort five, which are mainly found in space-related works.

Thus, the core of semantic cohort three is science and exploration, while the core of semantic cohort five is science and the high technology used in going to outer space.

Semantic cohort four can be analyzed in conjunction with semantic cohort six. Semantic cohort four takes “war” as the background and describes more about the war between people or between people and aliens. For example, words such as star, Mars and base often appear in *Fly to Horse Constellation*, which depicts several young men who are unfortunate enough to wander into outer space, but instead of giving upon themselves, they study and research hard, and are eventually picked up by another group of brave astronauts after the war on Earth is won. Yichang Song’s *After the Box of Woe is Opened*, on the other hand, describes the story of the Simes who want to invade Earth, and humans unite different races of aliens and eventually defeat the invaders. In semantic cohort six, alien is the most probable feature word, so it is more inclined to describe the process of human-alien communication or exploration of outer space. Words such as archaeology, microwave, signal and information can be linked to the discovery of the mysterious copper ball in *The Mystery of Ancient Star Map*, and then the use of microwave technology to find the signal left by the aliens and eventually discover them.

So semantic cohort four is centered on war, space and technology, while semantic cohort six is centered on communication between humans and aliens and the discovery of extraterrestrial civilizations with the help of high technology.

3.2 Motif Induction and Analysis

Based on the above analysis, the author concludes seven motifs.

(1) Discovering new things with the help of technological inventions
(2) Adventures completed with the help of technology
(3) Technology for a “modern" future
(4) Non-human groups present new standards of moral evaluation of humans
(5) Nuclear war brings disaster to mankind
(6) The risks associated with high-tech war
(7) Humanity wins the interstellar war

These motifs more or less reflect the background of the times, the progress of science and technology of all mankind, and the change of the overall view of the times. Specific analysis is as follows.

After Comrade Deng Xiaoping's important speech *Emancipating the Mind, Seeking Truth from Facts*, Looking Forward to Unity, the status of science was restored. Science fiction literature also flourished as a result. [4] In response to the call of the times, a large part of the works of science fiction literature in the new era depict new technologies, the convenient life created by technology for people, and the new things created by technology within a reasonable range of imagination, all of which can be grouped under the motifs one and two.

At the same time, Comrade Deng Xiaoping also put forward the major strategic goal of Building Four Modernizations of Chinese Style. As a result, many of the works of this period depict a better future after the development of science and technology to a certain level. Among them is Yonglie Ye 's *Xiaolingtong Roaming the Future*. With the help of high-tech products, people can carry out more modern scientific research and adventure.

But in works where technology creates a bright future of "modernity" or where technology benefits mankind in every way, individuality comes to the fore, unlike the popular science works of the early Republic, where the characters appear only as a background to the popular science, presenting a symbolic character that serves only to popularize scientific knowledge. Although Xingsi Liu's *The Legend of Mountain Fog* popularizes a lot of knowledge, it portrays Zhong'an Cao as a brave and knowledgeable archaeologist. In the works of popular science of this period, there is always a desire for "literary" quality. For example, although Jiazi Cheng's *The Mystery of Ancient Star Map* explains the knowledge and equipment related to "microwaves", the more important purpose of these works is not to popularize science, but to depict the sincere feelings of human beings, to think about and reflect on war itself, and to advocate for the ecological environment. These innovations reflect the "emancipatory" atmosphere of the time and the unconscious exploration of the independence of science fiction literature in terms of genre.

The core of motif four is the question of why humans need new standards of moral evaluation. In Tao Jin's *The Moonlight Island*, the Sirius bluntly name a dozen human "evils" such as "greed and selfishness", and in *After the Box of Woe is Opened*, The Simes say that man is arrogant and "learning" man invades the earth without hesitation. How much behind this can reflect a decade after the "catastrophe", human pain began to reflect on their own. Scar Literature, which was flourishing at that time, affected the creation of science fiction literature to some extent.

In *Fly to Horse Constellation*, there is a nuclear war on Earth, and much of the human perspective on nuclear war is described in the text. [5] In *After the Box of Woe is Opened*, there is a clear description of the "green war" that prevents the destruction of the ecological environment that nuclear war may cause, leading to the destruction of humanity. Nuclear war is a direct reflection of our country's fear of nuclear weapons in the political context of the "Cold War", which is also expressed in different ways in science fiction.

Due to the revival of science, high technology is also being used in warfare. For example, the use of antennas in *The Dream of Peace*, the modern form of warfare in *The Dead Light on Coral Island*, where scientists are attacked on their return home and countries compete with each other for high-tech weapons in a ploy, and the use of nuclear weapons and spaceships in *Fly to Horse Constellation*, and *After the Box of Woe is Opened*, all reflect the social context of "emancipation" at the time.

Many of the early works of the Republic can also be summarized in motifs of technology. The difference between this period motif and the new period motif lies in the status of the "human being" is very much enhanced, and the richness and integrity of the plot of the work is improved. So, although there are certain similarities between the two periods in terms of technological motifs, the works of the new period are more literary in nature.

For motif seven, humans always won in the interstellar wars during this period, compared with the works of human failure in the "turn of the century" (1990–2000), which can reflect the increasingly common motif change of "interstellar".

On the whole, although the exploration period of science fiction literature is only six years, the motif of this period has changed greatly compared with before, more attention to the human itself. The deficiency is that due to the short exploration period and the limitation of the technological level at that time, the depth of the work reflected by the motif does not reach a very profound level. However, the works of this period expanded the scope of science fiction literature in various forms and explored more possibilities of Chinese science fiction. Therefore, this period can indeed be used as the "exploration period" of science fiction literature.

3.3 Classification of Motifs of Works

By combining the semantic cohort and the work itself most likely to correspond to each work output by LDA model, each work can be classified into one or more motifs. The results are as follows (Table 2).

Table 2. The motif of the work

Motif	Work
Discovering new things with the help of technological inventions	'*The Mystery of Ancient Star Map* by Jiazi Cheng *Xiao Ping Pong Has Changed* by Ye Guo *Magic Shoes, Magic Box, Man and Beast* by Tao Jin *The Eye of the Sea, The Falling Dust of Life, The Legend of the Dead City* by Xingshi Liu *BB-1* by Ke Lu *The Dead Light on the Coral Island* by Enzheng Tong *Xiaolingtong Roaming the Future* by Yonglie Ye *Fly to Horse Constellation* by Wenguang Zheng

(*continued*)

Table 2. (*continued*)

Motif	Work
Adventures completed with the help of technology	*The Mystery of Ancient Star Map* by Jiazi Cheng *The Dream of Peace* by Junzheng Gu *Summer in the Blizzard, Iceland Missing* by Tao Jin *The Eye of the Sea, The Legend of the Dead City, The Legend of Mountain Fog, Columbus from America* by Xingshi Liu *The Mist of the Ancient Gorge* by Enzheng Tong *Dream Under the Ice* by Xiaoda Wang
Technology for a "modern" future	*Xiao Ping Pong Has Changed* by Ye Guo *Magic Shoes, Magic Box* by Tao Jin *Xiaolingtong Roaming the Future* by Yonglie Ye
Non-human groups present new standards of moral evaluation of humans	*The Moonlight Island, Chinese New Year's Eve* by Tao Jin *After the Box of Woe is Opened* by Yichang Song *The Dead Light on the Coral Island* by Enzheng Tong *The Dream of a Soft Country* by Yahua Wei *Professor Shalom's Mistake* by Jianheng Xiao
Nuclear war brings disaster to mankind	*After the Box of Woe is Opened* by Yichang Song *Fly to Horse Constellation* by Wenguang Zheng
The risks associated with high-tech war	*After the Box of Woe is Opened* by Yichang Song *Fly to Horse Constellation* by Wenguang Zheng
Humanity wins the interstellar war	*The Dream of Peace* by Junzheng Gu *The Mist of the Ancient Gorge* by Enzheng Tong *After the Box of Woe is Opened* by Yichang Song *Fly to Horse Constellation* by Wenguang Zheng *Dream Under the Ice* by Xiaoda Wang

4 Conclusion and Future Work

This study is not perfect. The defect is that the number of texts is small, which leads to a small number of motifs that can be summarized, and the level between motif and motif is not clear. It is like that the motif "discovering new things with the help of technological inventions" contains more content than the motif "humanity wins the interstellar war". The author will increase the amount of text in subsequent studies, and finally summarize the motif in different levels, and use the motif to write a brief history of science fiction literature. At the same time, the big data technology will draw a knowledge map containing all the motifs of Chinese science fiction literature, so as to visualize the results. This map is expected to be used for the World Science Fiction Conference held in Chengdu in 2023.

Although the existing research has shortcomings, it has been able to reflect the advantages of big data technology in the field of humanities and social sciences. Researchers can quickly grasp the text framework and content range of massive texts through big data mining technology, so as to avoid biased results analysis in a certain direction. The result of LDA model used by the author itself is the key content of the text summary, which also provides a broad range for the summary of the motif. These data themselves are powerful data that support the rationality of the summarized motif. At the same time, the result of text clustering can classify the works orderly, which makes the result analysis more organized.

In short, big data technology has an urgent application value in the field of humanities and social science research, which can promote social science research into a new level.

Acknowledgments. This work was supported by the National Social Science Foundation of China for Young Scholars (No. 21CZW055), the College Students' Innovation and Entrepreneurship Training Program of China (No. S202210610297).

References

1. S. Thompson., The Folktale, Shanghai Literature and Art Publishing House, 1991.
2. Wenxian Sun, As a structural form of motif analysis-language criticism methodology II, Journal of Central China Normal University (Humanities and Social Sciences) (06) (2001), 68 – 76.
3. Shunqing Cao, Jingyao Sun, Xudong Gao, Introduction to Comparative Literature, Higher Education Press, 2015.
4. Yan Wu, History of Chinese Science Fiction in the 20th Century, Peking University Press, 2022.
5. Zhenyu Jiang, Contributions and Mistakes: Zheng Wenguang and "Science Fiction Realism", China Modern Literature Series (08) (2017), 78–92. DOI: 10.16287 / j.cnki.cn11–2589 / i.2017.08.008.