Proceedings of the 5th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/133/2025.20607

Exploring the Commercial Value of China National Knowledge Infrastructure (CNKI) Based on Data Analysis Technology in the Context of the Big Data Era

Zhengwei Li[1,a,*]

[1]Xidian University, Xi'an, Shaanxi Province, 710000, China
a. 2460152527@qq.com
**corresponding author*

Abstract: In the current big data era, where information technology is advancing quickly, the Internet is becoming more and more popular, data analysis methods are becoming more sophisticated, and data value is being gradually mined in depth, this paper makes use of data analysis technology to thoroughly explore the economic and commercial value of CNKI. The primary research question is what CNKI's business strategy or operational procedures are used to make money and what kind of commercial value CNKI possesses. This article applies techniques such as linear regression analysis, data screening, significance testing, and data visualization in conjunction with R language and tools such as Jupyter Notebook to study the problem. Finally, it was concluded that CNKI has taken a leading position in the market through intermediary models, proportional fees based on the number of literature pages, and a closed-loop monopoly of knowledge data. CNKI has also gained a competitive advantage in the market through its extensive literature and user base, as well as strong search technology capabilities.

Keywords: CNKI, Data Analysis, Data Commercialization, Business Value, Data Visualization

1. Introduction

This paper explains the phenomenon of data commercialization in the modern era, analyzes the business data of CNKI, and evaluates the commercial market potential of CNKI against the backdrop of the big data era, which is marked by the rapid advancement of science and technology, the widespread use of the Internet, and the development of increasingly sophisticated data analysis techniques. The main purpose of these contents is to deeply explore the commercial value behind CNKI. The main research question is what business model or operating methods CNKI uses to generate profits and what commercial value CNKI has. The analysis method adopts techniques such as linear regression analysis, data screening, significance testing, and data visualization, combined with R language and tools such as Jupyter Notebook, to study the problem. Its importance stems from examining CNKI's potential worth, determining whether it can lead the sector, create an industry model, support the commercialization of data, create more jobs, address certain employment issues in society, ease capital flow, and raise a portion of the nation's income. In addition, it can also enhance the efficiency of academic research and knowledge innovation and promote interdisciplinary

DOI: 10.54254/2755-2721/133/2025.20607

integration. Finally, resource allocation can be optimized and service quality can be improved to expand market potential and commercial applications.

2. Commercialization of data in today's era

2.1. Commercialization of Big Data on Network Platforms

Due to its practicality and logical purpose, big data is now routinely commercialized on online platforms. The rationality of the purpose is reflected in the commercialization of big data, which fills the gap between people's information and has fairness. It enables people to exchange their needs through transactions, leveraging the unique advantages of different groups while fostering professions and positions and promoting the circulation of the social economy. And practicality is reflected in the core of big data commercialization, which lies in data mining, data cleaning, data integration, and other data analysis processes. This can be achieved through logical analysis combined with algorithm calculation and statistical estimation today, and its rationality has also been confirmed by current reality.

Statistics reveal that as early as 2014, the world's average annual data production had soared to approximately 40 trillion bytes, dwarfing the total data generated over the past 5000 years [1]. And in this massive amount of data, there are countless commercial values that can be mined. For instance, the China National Knowledge Infrastructure (CNKI) platform contains over 57.5 million literature materials, and by downloading the literature for a fee, one can obtain extremely high commercial profits. According to existing data, in a single online platform of China National Knowledge Infrastructure, regular digital publishing journals charge a download fee of 0.5 yuan per page for full-text articles and conference papers, 7.5 yuan per master's thesis, and 9.5 yuan per doctoral thesis. The huge profits obtained from this are the benefits brought by the commercialization of data today.

2.1.1. Commercialization of Data Elements

The commercialization of data elements is a growing trend in today's society, with data trading volumes exceeding 100 million yuan in August 2023 [2]. This demonstrates the vastness of the data trading market, which includes non-personal data, government data, and enterprise data, as well as some personal data. The commercialization of personal data is crucial for the commercial utilization process of big data, as it provides guidance and achieves utilization effects [3]. In the digital economy, data is considered a new factor of production, similar to labor and money. It forms the foundation of digital industrial chains, driving production, predicting future trends, driving innovation, and reducing costs for decision-making in various fields. The commercialization of data elements can promote society's development and construction. Senior expert in digital transformation Cai Jianjun states that when resources are digitized and data is integrated into production, traditional production will be comprehensively upgraded, reducing costs and increasing efficiency. Data also endows society with intelligent capabilities, making commercializing data elements valuable [4]. For example, the China National Knowledge Infrastructure website now sells papers at a price of 0.5 yuan per page and earns knowledge protection fees from publishers.

3. Business Data Analysis Technology and Process

3.1. What is Business Data Analysis

Business data analysis begins with the collection of business data, followed by the application of descriptive data analysis, predictive data analysis, and normative data analysis, with the aim of supporting and demonstrating the business decision-making process and organizational performance.

DOI: 10.54254/2755-2721/133/2025.20607

It is a process of utilizing large amounts of structured or unstructured data for business insights, forecasting, and decision support. The technologies include data warehousing, data mining, machine learning, and other methods used to extract business knowledge and insights.

3.2. Business data analysis process

3.2.1. Data collection

Data collection is the front-end process of business data analysis. In essence, data production is data collection, focusing on the action of "collecting," which refers to the behavior of recording specific facts or phenomena themselves and their formation or development processes in the form of data and forming a machine-recognizable combination of "0" and "1" [5]. Simply put, data collection is to search for the conditions required to solve a problem before analyzing it. And the collection methods of these conditions are constantly developing with the progress of the times. Traditional data collection is mainly done through manual search, using access methods, which are inefficient, time-consuming, and laborious. However, today's data collection methods are mainly machine retrieval, collecting data from various sources such as social media, online platforms, mobile devices, etc. This greatly improves the efficiency of data collection and reduces labor costs.

3.2.2. Data filtering

After using commercial data collection, information on certain papers was gathered from the China National Knowledge Infrastructure (CNKI), with the majority being graduate theses, academic journals, and newspapers. Due to the large amount of paper data in CNKI, data screening is necessary. Among the various filtering methods, Python was chosen for practical implementation in this article. This is because Python, being the predominant programming language in the field of data science, is widely utilized. Its data filtering and processing techniques are quite flexible and can be applied to various professions and fields. In Python, numerous standard or third-party libraries are available for data processing and filtering, and pandas is one of the most widely used libraries [6]. Using pandas, 200 articles were planned to be selected for a sampling survey from an already edited Excel data table. The files had been categorized in the data table, and the categories were journals and newspapers (1 represents journals, 2 represents newspapers). This is because much of the literature stored in CNKI comes from academic journals and newspapers, and the prices of these two are very different. In this way, 200 articles from both journals and newspapers were obtained for further exploration of data visualization and the establishment of data models to explore the commercial value of China National Knowledge Infrastructure (CNKI).

3.2.3. Data Exploration and Visualization

Data exploration and visualization hold a high position in bus. Infrastructure, we also use data visualization methods. The first thing to establish is a correlation heatmap, which can be written in Jupyter Notebook using pandas, numpy, matplotlib. pyplot, and matplotlib. In this way, a correlation analysis can be obtained for the four factors in the table (Figure1).

DOI: 10.54254/2755-2721/133/2025.20607

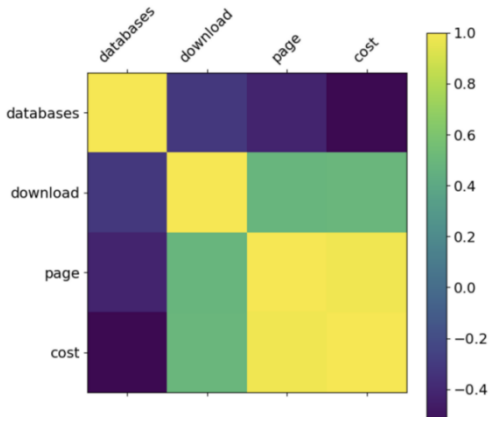

Figure 1: Heat map of four factors related to the correlation analysis

3.2.4. Model building

The construction of models offers substantial support for business decision-making, as it can assist businesses or individuals in establishing a unified standard among numerous data. In this article, the construction of a linear regression model is adopted. From the correlation heatmap, it can be intuitively observed that there is a very high correlation between the page and cost of papers belonging to journals. Through the use of R Studio to conduct a univariate linear regression analysis on this, it is found that the p-value is less than 2.2e-16, and the R^2 value is close to 1, indicating an absolute and significant correlation between these two factors. In addition, when establishing a multiple regression between download, page, and cost, it is discovered that the p-value was 1.439e-11 and the R^2 value was approximately 0.25. Thus, these three factors also exhibit a high correlation and significance. This might be because the amount of downloads can affect people's desire to download this article, and when the amount is not proportional to the demand, people will refrain from purchasing. After the model construction, in order to more intuitively reflect the correlation between data, we also created a visualization of the data (Figure 2). By combining the results analyzed from the graph and model construction, a highly reliable conclusion can be obtained.

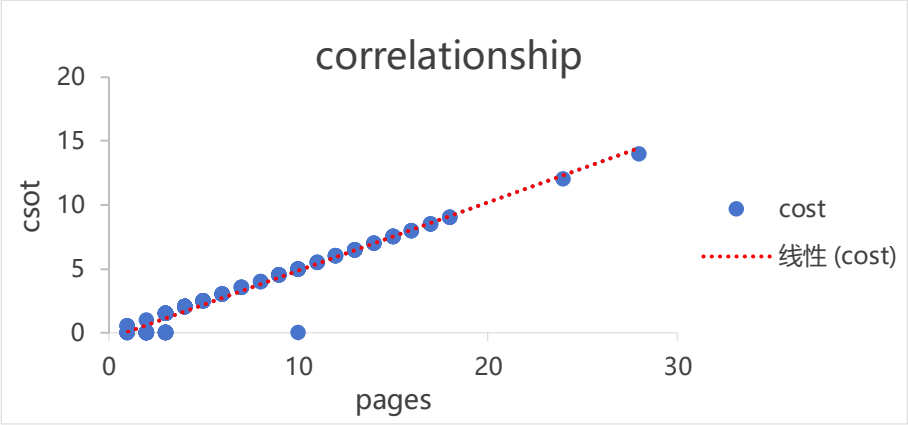

Figure 2: Data visualization view of the linear regression equation between cost and page

DOI: 10.54254/2755-2721/133/2025.20607

4. The commercial market potential of CNKI

4.1. Market position of CNKI

China National Knowledge Infrastructure (CNKI) is the largest CNKI digital library in the world in terms of the volume of academic papers and information. It collects over 95% of officially published Chinese academic resources and has a huge resource volume. At the same time, CNKI's market share is also far ahead, with its market share increasing year by year, reflecting its increasingly frequent data trading and sharing. The influx of more users has put CNKI's market share in a leading position in the academic knowledge service market.

4.2. The core competitive advantage of CNKI

China National Knowledge Infrastructure (CNKI) has several core competitive advantages, including vast database resources, technological advantages, and brand advantages. Its extensive database covers multiple disciplines and sources, offering efficient and convenient services. The brand advantage is due to its high visibility and influence in the academic community. Data analysis of CNKI's financial statements reveals a continuous revenue growth, with a nearly threefold increase in revenue due to its vast customer base and rich literature resources. The gross profit margin of CNKI is consistently between 50% and 70%, making it beyond the reach of many enterprises with large data industry chains. The user base of CNKI is also a competitive advantage. Its early establishment, excellent search technology, and large literature storage have attracted many users, leading to the digitization and consolidation of national journals and papers on a large platform. Although this approach has caused significant losses and impacts in the future, it remains the core competitive advantage of CNKI.

4.3. The Business Model of CNKI

In brief, the business model of CNKI can be described as an intermediary model. The authors of papers send submitted work to CNKI, which provides storage and protection services while providing a certain profit to the authors. Those in need of paper materials will obtain the corresponding papers through queries and are required to pay corresponding fees according to the number of articles or pages. The price differential thus earned constitutes the main source of income for CNKI. In addition, CNKI provides different levels of membership services to diverse user groups, such as institutions and individuals, through membership subscription services. Members can enjoy more privileges such as literature search, download, and citation, which makes users who need to cite and search literature for a long time more inclined to purchase. In addition, China National Knowledge Infrastructure (CNKI) collaborates with other academic institutions to promote its own advertisements on different platforms to attract users, thereby increasing its brand awareness and attracting a large number of customers to subscribe as members. Finally, CNKI also earns fees by collaborating with university libraries. According to the regional comparison report on the development of literature resources in university libraries by the Ministry of Education, the purchase rate of China National Knowledge Infrastructure ranks first among various Chinese electronic resources such as Chaoxing, Wanfang, and VIP data platforms [7]. This indicates that most universities today cannot do without the information services of CNKI, which has formed a certain scale of literature monopoly and earned a considerable income.

DOI: 10.54254/2755-2721/133/2025.20607

4.4. User satisfaction of CNKI

User satisfaction is derived from customer satisfaction. In the 1960s, the concept of user satisfaction first appeared in the literature related to scholar Cardozo. It refers to whether customers' evaluations of products and services are lower or higher than satisfaction, and it is a psychological state [8]. For the user satisfaction survey of CNKI in this article, a literature survey was used. We found through literature review that before using knowledge services, CNKI had already made users very satisfied with its external image and trusted its services through external promotion and other means. At the same time, users are also satisfied with CNKI's convenient search methods and fast query speed. According to the survey, a large number of CNKI users are students and teachers. During the user satisfaction survey, it was found that teachers had higher average satisfaction with seven variables: image, expected quality, perceived value, perceived quality, knowledge service process, quality, and user satisfaction than student users. Most teachers use CNKI knowledge services for a longer period of time than students and have a deeper understanding of knowledge services. They are often used to assist scientific research and consult professional knowledge. And the average satisfaction of teachers with each variable is higher than that of students, indicating that the knowledge services provided by CNKI are indeed beneficial to them [9]. However, some literature also shows that a large number of universities refuse to renew contracts with CNKI, and the underlying reasons indicate that CNKI has a monopoly on knowledge and information, as well as bad behaviors such as closed-loop "money making." High-priced literature materials make people hesitate. Indeed, the high-priced literature on CNKI has resulted in poor user experience for some student and teacher groups who have been using CNKI for a long time, leading to a decrease in user satisfaction.

5. Current Challenges and Future Development

The most conspicuous dilemma confronting CNKI at present is that its exorbitant paper fees have given rise to a high level of negative sentiment among users, leading to a certain degree of user attrition and a continuous decline in its online reputation in recent years. This is attributable to the fact that some universities have declined to continue collaborating with CNKI due to the high renewal fees. Some professionals, upon analyzing CNKI's financial reports and certain profit practices, have discovered that CNKI attempts to achieve zero-cost introduction of resources while having a super high gross profit margin and engages in inappropriate behaviors such as "blood-sucking" loops. This caused dissatisfaction among a large number of users and the public and even later led to the State Administration for Market Regulation imposing high penalties on it. So the most important thing for CNKI today is to improve its own reputation, reduce the cost of browsing and downloading literature, and meet the search needs of the public. Furthermore, CNKI is also facing the dilemma of innovation. With the rapid development of digitization and intelligence, users' demand for knowledge services is constantly changing and upgrading. And society's awareness of data protection is constantly strengthening, which will have an impact on CNKI's business methods (CNKI is suspected of monopolizing the knowledge literature database) [10]. The United States' measures against super-large platforms. The review is constantly deepening, mainly examining whether the platform has engaged in data abuse, monopolistic leverage, predatory pricing, core facilities and refusal to trade, bundling, self-preferential treatment, and stifling acquisitions [11]. This serves as a cautionary signal that CNKI needs to institute changes to mitigate restrictions on platform users and relax its control over literature data.

6. Conclusion

Based on the above analysis, it can be concluded that CNKI primarily generates profits through multiple channels. It charges a proportional fee based on the number of pages to those interested in

DOI: 10.54254/2755-2721/133/2025.20607

protecting and storing their literature and those with a certain demand for literature. Additionally, it offers different levels of membership and continuously renews cooperation with universities. Meanwhile, according to the analysis of its commercial market potential, it is evident that CNKI wields significant influence and possesses a vast file and user base in China. Combined with its fast and convenient search method, it has cultivated strong market competitiveness. In this manner, CNKI is able to facilitate the flow of funds and generate more transactions. At the same time, if the managers of CNKI are willing, more job positions can also be established to share the pressure of social employment. In addition, CNKI's vast database and resource aggregation capabilities can help scholars collect resources faster and provide comprehensive and accurate retrieval and download services. Similarly, its membership subscription service and advertising cooperation have both achieved growth in social and national income. This is its commercial value. Nevertheless, it should be noted that this article has only examined some of the domestic business information of Tongfangzhiwang, Wanfang, and Chaoxing, and thus is not exhaustive. The established model is only a significant model for price and page count and price and people's purchasing desire. These studies are all aimed at exploring CNKI's current business model to reflect its commercial value. In the future, the research focus can be on how CNKI can profit from external operations and how CNKI should improve its pricing system to achieve a true win-win situation with users. Concentrating research in these areas could endow CNKI with greater influence and a better reputation, enabling it to progress further.

References

[1] *Liu Zhihui and Zhang Quanling: "A Review of Big Data Technology Research", Journal of Zhejiang University (Engineering Edition), June 2014, Volume 48, Issue 6.*

[2] *Jiang Muyun and He Shasha: "Preliminary landing of commercialization of data element market by increasing the activity of data exchanges in various regions", China Business News/September 25, 2023/B07 page:1-3 edition.*

[3] *Wang Lei: "Conflicts of Interest and Their Solutions in the Commercial Use of Personal Data", Legal Science (Journal of Northwest University of Political Science and Law), May2021, page:3*

[4] *Jiang Muyun and He Shasha: "Preliminary landing of commercialization of data element market by increasing the activity of data exchanges in various regions", China Business News/September 25, 2023/B07 page:1-3 edition.*

[5] *Wang Zhenping, Peng Xiaojian: The Rule of Law Path for Government Data Collection 2021 Phase 1 Sun Yat sen University Law School Guangzhou 511400, page:1-2*

[6] *Cao Haiying, Yuan Yuan: Research on Data Filtering and Filtering Based on Python Language, Issue 14, 2024 (Department of Mathematics and Computer Science, Hetao University, Bayannur, Inner Mongolia page:2*

[7] *Tong Yanhua from Henan University Library, Fully utilizing China National Knowledge Infrastructure to deepen information services in university libraries, page:1*

[8] *Li Jianxia, Liang Ru, Liu Ying Empirical Study on User Satisfaction Evaluation of University Libraries Based on Improved LibQUAL+: A Case Study of East China University of Science and Technology Library [J] New Century Library, 2017 (12): 48-54*

[9] *A Study on the Satisfaction of Teachers and Students of Gao Xueqian University with Knowledge Services of China National Knowledge Infrastructure (CNKI) and the Full Text Database of Excellent Master's Thesis in China, May 2021 page:44-52*

[10] *State Administration for Market Regulation: "State Administration for Market Regulation issues administrative penalty decision on CNKI's abuse of market dominance case" 2022-12-26 16:00, https://www.samr.gov.cn/fldes/tzgg/xzcf/art/2023/art_27cab7312a424e0ea46c6fa9e5044371.html*

[11] *INVESTIGATION OF COMPETITION IN DIGITAL MARKETS, https: / /judiciary.house.gov/uploadedfiles/competition_in_digital_markets. pdf? utm_campaign = 4493 - 519, last visited on January 31, 2021.*