A study on Digital Book Classification and Quick Search in University Libraries

Deepak Ramesh Patil

Assistant Professor, Department of Library and Information Science, Pratap College Amalner (Autonomous)

Abstract

With the advancement of society and the accelerated pace of people's lives, people cannot spend too much time classifying and finding books, so the study of book classification and quick finding in university libraries is very important. A digital library is a digital information resource system supported by modern high technology, a next-generation information resource management model on the Internet, and the result of the digitization of library collections. The primary focus of this study is to improve existing algorithms for the categorization and fast search of books in academic libraries using digital information technology. In this work, we primarily perform tests on automated text and support vector machine (one-to-many and global optimization) approaches, and then we compare the results in terms of classification accuracy, classification time, search time, and other metrics. Experimental findings suggest that these three classification approaches have a classification accuracy of between 86% and 94%. On the other hand, the global optimization classification achieves the maximum accuracy in the sample size of each interval when compared to the other two approaches (automatic text classification and one-to-many classification). Their average fitness is between 24% and 27%, with automated text classification having a classification time of less than 30s and one-to-many taking the greatest time.

Keywords:- Digital Information Technology, University Libraries, Quick Search, digital library.

1. Introduction

Rapid advancements in computer and Internet technology have had far-reaching effects on many facets of human civilization, with certain areas undergoing profound transformations. Because of this, the digital library is also created in the digital era, expanding its reach beyond the bounds of the conventional library. Transitioning from physical to digital libraries is just part of what's needed to build a thriving publishing sector; what's also needed is the research and practise to develop the Internet's next generation of digital libraries. Unfortunately, the digital library is not currently seeing a great deal of scholarly investigation. They pay attention to the HCI part of the research but disregard the book categorization and rapid search parts of the digital library. If you have a decent algorithm, you can drastically cut down on the time it takes to categorise books and do a quick search. Thus, it is crucial to develop a more effective algorithm and implement it in the digital library, since this is the primary focus of modern academic inquiry.

Since a digital library is a novel area of computer application that makes use of numerous technologies (such as the Web, multimedia, data storage, data mining, and intellectual property protection), it has enormous potential for growth in both the marketplace and the advancement of education in the modern era of digital information. Time spent searching for books and shelving them by librarians should be minimised. The studies of book organisation and easy retrieval are therefore of crucial importance.

In this research, we focus on understanding the inner workings of automated text categorization and the support vector machine so that we may build a system that uses these two approaches in tandem. Specifically, we conduct our experiments on data taken from Sogou Lab's Chinese Text Classification corpus for Sogou

News. We initiate the experiment, capture the data, and evaluate the benefits and drawbacks of the strategy on a total of 2,000 text data samples.

This study makes new contributions in the following ways: (1) This article provides an overview of automated text categorization and its associated algorithms, as well as a number of support vector machine techniques and ideas. This document highlights the benefits and drawbacks of the three approaches discussed here—automatic text, one-to-many, and global optimization. (3) This study also covers basic optimization concerns for interactive user interfaces and gives a detailed introduction to the theory, functions, and components of human-computer interaction.

2. Related Work

Research on digital library organisation and search efficiency has increased recently. Through a case study of China's Digital Library, Li S was able to examine the current state of the art in terms of big data information technology, content, and relationships, and draw the conclusion that the blockchain can improve data accuracy, security, and efficiency in data collection, storage, and dissemination. Using these premises, we develop a comprehensive scenario for the use of cutting-edge IT in the virtual library [1]. Paletta looked at the IT life cycle management behind digital libraries, as well as IT's dynamic potential to provide innovations that have an impact on service quality. Digital libraries may benefit from making better use of these new technologies [2]. In contrast, there is no widespread experimental confirmation. In order to sort out the wheat from the chaff, Sonkar researches all of the unanswered questions that occur in the process of creating a digital library of clippings, including concerns of metadata selection, preservation, technological obsolescence, and copyright violations [3]. Yadav tackled the challenging challenge for libraries to access these materials in the future [4] by studying the categorization and preservation procedures employed by chosen libraries in New Delhi, India. However, the price is prohibitive. Kato's study focuses on how digital library (DL) materials are created, promoted, adopted, and used in higher education. By highlighting these key characteristics of DL services, he demonstrates how effective DL utilities are and how easy it is to have access to information over the web [5]. Linlin researched how academic libraries are using data processing technologies. Virtual worlds that were formerly text- and image-based are evolving into rich, photorealistic 3D worlds. His plan for creating a digital library has three stages: planning, pilot testing, and implementation [6]. Umeozor looked at how content-based image retrieval (CBIR) and reverse image search exploit picture reuse in digital libraries (RIL). Four published case studies on evaluating the usage of images in digital libraries were also briefly discussed [7]. However, comparing this approach with others is inadequate.

3. Theoretical Knowledge and Methods

3.1. Digital Libraries

The evolution of computing, networking, and communications has considerably boosted the speed and efficiency with which digital information can be created, processed, and disseminated [8]. The convenience of digital information storage, transmission, and processing has led to its widespread adoption [9]. To keep up with the rapid changes in current social technology, classic information management systems like library administration and audio-visual file management are becoming obsolete. Unfortunately, there are still many unanswered questions about the Internet and computers. The first issues to be resolved at the moment are how to successfully organise, extract, acquire, and intelligently and efficiently use all types of vast digital information, and how to make the most of the benefits offered by the "Internet" [10]. Science has proposed the idea of a digital library as a solution to these issues.

For China's Internet, the future generation of information resource management looks like the digital library, which is a system for storing and retrieving digital information resources. The university library is the document and information centre of higher education, and the library that serves the teaching and scientific research of higher education is the library. The digital library is a significant accomplishment of the

comprehensive digitization of Chinese library collections [11, 12]. Traditional libraries, which cannot keep up with the ever-increasing demands of modern education, are giving way to digital library systems as a result of the exponential growth of scientific and technological knowledge. Electronic books, magazines, electronic newspapers, and research reports now make up the bulk of the library's collection. Through the Internet, they may also connect to other digital resource websites, expanding the controlled set of resources beyond only the library's own publications. Digital libraries allow for study that is not limited to the confines of traditional libraries.

3.2. Automatic Text Classification Methods

A digital library has a wealth of information. Unlike traditional libraries, the digital library's service centre focuses on patrons rather than collections. The data library's qualities take its company to a new level, one that relies on information, and through the intelligent integration and administration of information, the library's resources become an information system [14].

Classifying vast volumes of text into different groups according to their shared characteristics or characteristics that may be extracted from the text is called text classification. One kind of supervised learning method is a text categorization algorithm. Manually organised training materials and distinct document types are required. Our goal is to build a classifier from this trained model and use it to assign labels to fresh documents. As a result, standard procedures for processing data will not work. Attribute-representing information must be extracted from the text during preprocessing. Metadata, also known as intermediate data or relay data, is data about data that describes the data itself and serves various purposes, such as indicating where data is stored, providing context for current data, facilitating the discovery of relevant resources, and recording changes to existing files. Metadata is an electronic catalogue used for cataloguing. The first step in dealing with difficult-to-represent items is identifying a representation that can be processed by a computer. To develop a mining model is to develop a target representation. Similarly, Chinese papers need to be tokenized in advance. Different target representation models exist. Commonly used types include Boolean, vector space, probability, and so on. Under the Boolean paradigm, a document's existence or absence may be indicated by a single digit value between 0 and 1, with 1 denoting presence and 0 indicating absence. While its simplicity is a benefit, this representation is seldom utilised since it fails to adequately communicate information about the relative significance of distinct traits. The probabilistic model determines which tokens belong to which categories based on the likelihood that the given text belongs to each. The model suffers from the fact that it does not account for the occurrence of index terms in the body of text. Vector space modelling techniques have emerged as the gold standard for accurate and efficient object representation in recent years.

In order to prevent overfitting, feature extraction plays a crucial role in text analysis by decreasing the size of the vector space and streamlining the algorithm. The exponential nature of the relationship between the number of feature subsets and the total number of features makes it very impossible to count them all; as a result, it is often assumed that features are unrelated to one another. This is why we have the feature subset extraction approach; it adjusts the original feature extraction procedure. Thousands of words with the highest scores are chosen as feature words via the scoring function of each individual feature, the score of each feature and the division method of each digital library feature may be tallied and organised according to the score, and so on. Using the Gini coefficient, we may extract useful features for text categorization. In economics, a number between 0 and 1 is used to describe inequality using the Gini coefficient. Any number below (typically) 0.2 indicates a lack of authority, while any value beyond (usually) 0.4 suggests an inappropriate distribution of wealth. A value of 1 implies that the wealth of a nation is in the hands of a specific individual.

3.3. SVM Support Vector Machine Classification Method

When it comes to solving the issue of pattern recognition, the support vector approach presented by Vapnik provides the theoretical foundation for support vector machines, which are the most practically relevant part of the statistical learning theory. SVM is a methodical procedure that produces consistent outcomes.

Optimizing a quadratic objective function on a convex set, which does not experience local optima errors, is a lengthy procedure in SVM training. In data mining, SVM excels at solving binary classification issues, such as those encountered in text classification applications.

4. Application Experiment of Digital Information Technology in University Library Book Classification and Quick Search

4.1. Quick Search of Digital Library Based on Human-Computer Interaction

The term "human-computer interaction" refers to the use of various computer input and output technologies to facilitate efficient communication between people and computers. It entails computers providing a great deal of relevant information and requesting instructions through output or display devices, and people inputting relevant information and requesting actions via input devices. In recent years, the topic of interactive interfaces has emerged as a distinct and vital area of study, attracting the interest of computer manufacturers from all over the globe and opening up yet another arena of rivalry in the IT sector. Human-computer interaction technology is an integral aspect of the evolution of computer science, and it helps to establish the parameters for the development of related software and hardware. For the next generation of computing systems to be really revolutionary, this technology must be further refined. Technology for human-computer interaction and user interface is now trending toward becoming more natural and harmonious. The human-machine interface is quickly emerging as the information interface of the future, because to the proliferation of computers and the growth of data in many industries. This is an issue with any system, not just digital libraries. As a result, it's not easy to take advantage of every method for transforming data for computer presentation.

The human-computer interface (also known as the UI) is a crucial component of any digital library since it provides the means through which users access the system and conduct operations such as searching for and retrieving content. Human-computer interaction is made possible by the digital library's ability to search for the user's query in the background and display the results on the library's user interface. We succeed in making the UI invisible to the end user. The user will be less apathetic and more engaged in their pursuit of answers. It is important to keep the following in mind while creating an interactive user interface: it should be accessible to a wide range of people, easy enough for even novices to use, empathetic without being patronising, and cerebral without being too abstract. Get the user's attention and make it simple for them to utilise. The next generation of human-machine interfaces will likely include the merging of virtual reality technology with data visualisation. When compared to other forms of human-computer interaction technologies, VR stands as the most promising for achieving a "people-oriented" and harmonious HMI.

4.2. Application Experiment and Book Classification and Quick Search in University Library

In this experiment, we use information from the Sogou Lab Chinese text classification corpus for real-world use. The bulk of what makes up the Chinese text categorization corpus are actual news stories archived by media outlets like Sohu. Selecting just news extracts, this article has manually sorted this subset of data, then labelled it all using classification to guarantee proper labelling. There are seventeen distinct labels for sections within the system, most of which are chosen in relation to the subject matter of the report. You'll find mostly macroeconomic news, sports headlines, and tech sector analysis.

5. Discussion

Automatic text categorization and support vector machine are the primary classification techniques used in this work. Following the development of these two approaches, this work implements the relevant system. The Sogou News Chinese Text Classification corpus on Sogou Labs is used as data for experiments. There are a total of 2,000 text samples spread over 8 categories in the sample dataset. After that, it gets to work organising tests, locating them, collecting data, and evaluating the method's efficacy. There are, however, still flaws in this experiment due to the fact that no university library could possibly contain just 2000 books; in

fact, many university libraries have tens of thousands more volumes. There will be dozens of distinct types of books, which is substantially different from the overall sample size of the experiment. In addition, this experiment did not optimise the appropriate procedure, which did not increase the system accuracy on the original premise. It also doesn't investigate the search algorithm any further. However, this experiment's findings have a certain degree of dependability in the grand scheme of things. It may serve as a benchmark against which future attempts at improvement might be evaluated.

6. Conclusion

One-to-many classification, global optimization classification, and automated text classification are primarily compared throughout this study. Experiments have shown that these three techniques of categorization have an accuracy of between 86% and 94%. On the other hand, the global optimization classification achieves the maximum accuracy in the sample size of each interval when compared to the other two approaches (automatic text classification and one-to-many classification). Classification times for these items are all under 30 seconds, making them the quickest of all automated text classifiers. Most time-consuming are one-to-many classification samples, with an average fitness of 24%-27%. Classification methods based on autonomous text analysis and global optimization may work better for digital libraries. Materials can be located quickly in both of these approaches' digital libraries, with searches taking less than 5 seconds on average. The material presented demonstrates the breadth of potential future research applications for digital information technology. Particularly when used for digital library categorization and search, accuracy and efficiency would be improved.

References

1. Li S., Hao Z., Ding L., Xu X. Research on the application of information technology of Big Data in Chinese digital library. Library Management . 2019;40(8/9):518–531. doi: 10.1108/lm-04-2019-0021. [CrossRef] [Google Scholar]
2. Sonkar S. K., Makhija V., Ashok Kumar A. K., Singh M. Application of greenstone digital library (GSDL) software in newspapers clippings. DESIDOC Bulletin of Information Technology . 2005;25(3):9–17. doi: 10.14429/dbit.25.3.3655. [CrossRef] [Google Scholar]
3. Linlin Z. Analysis of the information processing technology of university libraries in the big data era. Agro Food Industry Hi-Tech . 2017;28(1):2036–2040. [Google Scholar]
4. Umeozor S. N. Information networking and its application in the digital era with illustration from the university of port harcourt library. International Journal of Knowledge Content Development & Technology . 2019;9(2):33–44. [Google Scholar]
5. Liu X., Li Y., Wang Q. Multi-view hierarchical bidirectional recurrent neural network for depth video sequence based action recognition. International Journal of Pattern Recognition and Artificial Intelligence . 2018;32(10) doi: 10.1142/s0218001418500337.1850033 [CrossRef] [Google Scholar]
6. Thompson S., Reilly M. A picture is worth a thousand words": Reverse image lookup and digital library assessment. Journal of the Association for Information Science and Technology . 2017;68(9):2264–2266. doi: 10.1002/asi.23847. [CrossRef] [Google Scholar]
7. Xi Q., Zhang Y., Li X., Wu W. Application of WeChat in university libraries in China. Education for Information . 2018;33(4):217–230. doi: 10.3233/efi-170135. [CrossRef] [Google Scholar]
8. doi: 10.1177/0266666920983393. [CrossRef] [Google Scholar]
9. Matthew O U., Nwamaka U O. Application of Internet of things for multimedia utility in the digital libraries in the south eastern Nigeria: issues and development. International Journal of Scientific Engineering and Research . 2019;10(7):383–392. doi: 10.14299/ijser.2019.07.02. [CrossRef] [Google Scholar]
10. Shadadeh F., Samadbeik M., Amiri F., Hajipourtalebi A. The digital gap in patients' use of health information technology and effective factors and strategies; a systematic review. Health Research Journal . 2019;4(3):181–188. doi: 10.29252/hrjbaq.4.3.181. [CrossRef] [Google Scholar]

11. Molinillo S., Japutra A. Organizational adoption of digital information and technology: a theoretical review. Bottom Line . 2017;30(01):33–46. doi: 10.1108/bl-01-2017-0002. [CrossRef] [Google Scholar]
12. Gijsbert W. Algorithm driven care%artificial intelligence%competencies%digital age%information technology%leadership%specialist in laboratory medicine. Clinical Chemistry and Laboratory Medicine . 2018;57(3):398–402. [Google Scholar]