Research on the Data Mining and Image Retrieval Framework of Classical Literature in the View of Big Data

Weiguang Zhang

Baicheng Normal University, Baicheng 137000, China

Abstract

It’s been one of the great challenges before the humane and social science to explore the valuable information and knowledge from the massive classical literature treasury of China. Along with the development of the digitalization of ancient books and database technologies, the challenge has become less daunting now. But compared with the information technologies and network technologies that advance with each passing day, the ancient book digitalization and related applications are still in a primary stage. The importance people attached to the value of literature data and their application is not even adequate. The recently rising Big Data theory and its application have opened a door for applying the network and other IT technologies in the traditional humanities and probably will produce a great impact on future literature study and particularly the classical literature research that requires great time devoted in processing huge amounts of literatures and documents.

Keyword: The View of Big Data, Classical Literatures, Data Mining, Image Retrieval.

1. BACKGROUND

1.1 Overview

In about recent twenty years, the network-oriented information technologies have achieved a rapid growth. But, to an extent, such development is more of a technological one. The data in itself seldom attracted the attention they deserve. According to the modern knowledge system formation theories, the knowledge, the primary source of human Wisdoms and Thoughts, is systemized information originated from properly processed data. This conclusion is from the famous DIKW Theory (Data-to-Information-to-Knowledge-to-Wisdom). More systematic discussion on the subject may be found in the Management Support Systems: Towards Integrated Knowledge Management, published in 1987 by Milan Zeleny, a famous US educationist, and in the “From Data to Wisdom”, Human Systems Management, published in 1989 by Russell L. Ackoff, a management thinker (Zheng, 2014). As is shown in the DIKW chain, the human wisdom is borne in a hierarchical progress from data to information and then wisdom. The data is the foundation of the chain. Without properly processed data, there can be no information or knowledge coming, let alone the wisdoms and thoughts standing above them. In an era of big data, the importance of data has never been so persuasively proven. Based on these theories, it’s obvious for us to review the role of the classical literature digitalization from a different perspective and discuss the necessity and urgency of enhancing database construction in the humanity disciplines and using the big data concepts and methods to facilitate our researches (Li and Gong, 2015).

1.2 Objectives

It has become a common view in the circle that the IT has given a great push to the development of classical literatures. Specifically speaking, the boost was felt mainly in the aspects of data mining and retrieval. In data retrieval aspect, there have emerged a few excellent achievements, e.g. the publication of the digital version of Wenyuange Si Ku Quan Shu (Complete Library of the Four Branches of Literature) and the digital version of Si Bu Cong Kan, and the Database of Chinese Classic Ancient Books. In data analysis area, the stage after the retrieval, people have made some progress, too. The Anthology of Poems of the Song Dynasty Analysis System is an exemplary work for late-comers. Since such accomplishments have been widely discussed by many experts, no more details will be given here (Li and Liu, 2015). The impact of IT development to traditional way of living is great, but to the traditional way of academic study, even greater. In today’s text research area, the big-data-based data-mining is the most noteworthy thing. There are two technical directions about applying the data mining in the text research: manually develop sufficient rules to break down the non-structural texts into semi-structured or
structured texts that the machine can understand, or, keep intelligentizing the machine with deep-learning algorithms until it can understand the non-structural texts without breaking them down, which of course represents the mainstream in current IT world. It’s sensible for researchers facing this general trend to take a welcoming and positive attitude and carry their text researches to a deeper level with the help of the new technologies. For example, it’s feasible to develop a special analysis platform for Yuefu poetry research on the basis of such key factors as titles, allusions, tune modes and form styles. Also, it’s possible to study gravestone epitaphs or letters with the help of data mining to understand the unique writing approach of them. It’s foreseeable that the knowledge system so developed will benefit the researches on other literary forms (Zhao and Yao, 2015).

2. DATA MINING OF CLASSICAL LITERATURE IN THE VIEW OF BIG DATA

2.1 Extract ideas and thoughts in poems

Our forefathers have left behind numerous ideas and thoughts in their insightful notes of poem and poetry. Let’s use some technical tools to verify them.

A frequently-used typical data mining system usually has a structure shown in Figure 1.

![Figure 1. The Structure of a Typical Data Mining System](image)

Let’s take the study on the Buddhist thoughts in poems as an example. The earliest representative figure who tried to combine Buddhist thoughts with poetry was Xie Lingyun (A.D. 385 - 433). He had profound attainments in the translation of Buddhist books, poetry writing and the enrichment of philosophical thoughts. He learned Buddhism from an eminent monk of his time, the Monk Hui Yuan. Deeply attracted to the thoughts and anecdotes of this prestigious monk, the young poet in whose mind the ideas of Buddhism had just begun to germinate travelled over a thousand Li to visit him personally. The meeting had a profound influence over the Buddhist thoughts from which many of Xie’s poems originated. According to Xie, the Origination of everything was “something like fine dust when it disperses, something like the world when it gathers together, when the properties of it alter, there is no dust or world, when there are dust and world, it’s nothing but a name”. And about the Emptiness, he described that “it’s bad to have it if it’s against the dharma, it’s good to have it if it’s not against the dharma. At the extremity of dharma, there is no difference between having and not having it.” (Liu, 2015) In Xie’s mind, therefore, the Highest Wisdom (Prajna) was “there are lines not to be crossed, but it’s all right to cross it if it’s so taught by the Buddha and not a disturbance to the peace of mind.” He did a lot of work to embed such thoughts into his scenic poems and, sometimes, even directly used the Buddhist terms in his poems, particularly the word “Emptiness”. For example,

*To return to seaside I’m fit; I lie sick, facing empty wood.* (On Poolside Tower)
To the empty the Dongting Lake billows; For no one I snapped this Osmanthus twig. (On Stone Gate New Barrack)

Gone with the flows the past of three rivers; Standing in emptiness were the records of saga’s thoughts. (At the Entrance of Pengli Lake)

In addition, he wrote a lot of poems that contains the concept of Prajna (the Highest Wisdom) with a profound understanding of his own. For example, in the Lonely Islet in River, he wrote:

Getting tired of the South of Yangtze River, I found it’s been long since I left the North. The road is so tortuous to a heat eager to greet something new; The time flies so quickly to an eye for spectacular scenes. Cross turbulent currents I sailed to the islet that lies so beautifully in the middle of the river. In this fresh world between the clean sky and the river, the clouds and the sun shine. As obvious as the beauty and delicacy are, no man is here to enjoy; As plain as the truths are, who’d be the one to tell?

In this poem, Xie Lingyun thought the dharma lies in the nature of this beautiful world itself. The “spectacle” in “the time flies so quickly to an eye for spectacular scenes” in fact represents “emptiness”. The changing world and the flying time are emptiness, just as said in The Heart Sutra, “Form is emptiness and the very emptiness is form; Emptiness does not differ from form; Form does not differ from emptiness; Whatever is form, that is emptiness; Whatever is emptiness, that is form”. Essentially, the scenes, the mountains, the waters, the stars, and the moving of stars are emptiness. (Guo and Xu, 2012)

2.2 Study the poetic diction and subject matters

As we know, the collage poems were made of lines written by poets of different generations in the same era. By identifying the reappearance ratio of the same line in different poems, it’s much easier for researchers to pinpoint the sources of the lines in a collage poem. The data mining process could be used here is shown in Figure 2.

![Figure 2. The Basic Process of Data Mining](image)

For example, from the perspective of subject matters, we could see that, before the Wutai Poem Case, except for those poems written for his friends and families, the political concerns and propositions were the most important subject matters in his poetry works. But after the Wutai Poem Case, it’s rare to see him writing poems promoting his political ideas. Instead, he devoted a lot of his poems to breathing out his thoughts on misfortune and expressing the feelings he captured from people’s day-to-day life. Before the Case, the life to Su, as a student or an official, was but an easy sailing without hitting the rough waters or head winds, which developed in Su an attitude of “abhor evils as deadly foes and spit in the face of evil as if it’s a fly on the table”. In a poem addressed to Kong Zhongwen, he wrote “An elk in nature, never have I foreseen me learning the skills of a draft horse (Huang, 2014). Escaping at the faintest voice of a groom, never would I bear the restraint of a yoke.” This was a clear statement of
his scornful attitude towards the glamorous official circle. In the Joke to Ziyou, he wrote “Now I am used to what I felt shamed for; ruling the poorest with sticks and whips”, “Called upon at the road side by Yanghu, knowing all his wrongdoings, I greeted him with yeses and ayes”, “Now in high positions, I’ve completely forgotten the sorrows of the poor. I don’t know what’s good about me anymore, since there is little integrity left in me”. It’s clear that he’d like to criticize political events in a satirical way. But because of exactly the same poems, he was trapped in a big trouble. Suffering a heavy blow from the Wutai Poem Case, he woke up to the fact that the satire style would bring him nothing but more misfortunes. After being degraded to a low position in Huangzhou, he felt completely lost about his life, lamenting “The world is but a dream and how many times can a man feel the cold freshness of the autumn”. His Tune: A Southern Song to Xu Junyou in Hanhui Tower on the Double Ninth Day made a very clear disclosure of a man torn apart in the conflict between the fortune and misfortune, the optimism and melancholy, and the magnanimity and helplessness. In this period of time, he turned to Buddhism and Daoism to console himself and to accept what had happened, i.e. the Wutai Poem Case, and his degraded official life in Huangzhou. As a result, he no longer touched the topics of exposing the cruel ruling or pointing out social problems. Instead, he began writing more about the little things in life, which brought out another side of his personality, a generous and magnanimous man. He’s not paid for his service in the Huangzhou government. So, he had to go to the fields and do all farming works. Living a life as hard as average farmers, he wrote in a poem to his friend, Sun Pingzhong.

Removing debris on the eastern slope last year, I planted a small mulberry bush.

Now building a thatched cottage for this winter, I had my face tanned in the sunshine and winds.

When it’s been too long without a rain, he’s as worried as every farmer, and as happy as them when the rain finally came.

Copiously is the rain falling, it’s hard to predict the next change in the nature as it has no heart.

Finally getting some rest, I lie in bed, listening to the happy footsteps outside the east wall.

In his Huangzhou times, he took Zhaoyun as his concubine. In the sixth year of Yuanfeng, she gave birth to a son, for whom he wrote a poem in remembrance of this happy moment:

People wish their kids be wise; Wisdom spoils me in my whole life.

I wish my son be foolish but robust, making his way to the high ranks without rougs.

2.3 Give emotions a body

Since the Dian (a form recording state affairs), Mo (a form recording plans and strategies), Zhaoling (a form of imperial decrees) and other forms appeared in the Shangshu, the ancient Chinese have had an initial view of literary forms. As dynasties replaced one another, different emotions began influencing each other (Li, 2016). Although the poets before Li Yu had their way of describing and expressing emotions, none of these attempts was genuinely deep-going. When wanting to express the happiness, they directly describe the happiness, and when the sadness, the same direct description of sadness. They simply couldn’t move the emotions from the lines into the hearts. But Li Yu found a unique way to express emotion, i.e. give emotions a body. In his Yu Mei Ren, he wrote “If you ask me how my sorrow has increased, just see the over-brimming river flowing east.” The melancholy and grief here go deeply in heart and carry the mind away farther and farther just as the river does in spring. This is a line that will go down with history with Chinese culture forever. In his Xiang Jian Huan, he wrote “Shear constant tangled is sadness, do the general feeling in their hearts.” The sadness then has a body, a tangled mess, showing an image of tangled and intertwined things. The sadness then has a body, a tangled mess, showing an image of tangled and intertwined things, and triggering off in a reader’s heart the anxiety and agitation. From the perspective of modern style theories, he used analogy to bring out an indescribable emotion with physical objects which in themselves resonate with the emotions in readers (Tang and Shi, 2017). Then the awareness of the tangled objects in return pulls the reader back to the emotions that Li tried to convey, letting them have a genuine experience of the sadness between the lines. His skill was learnt by late-coming poets. For example, in her Wu Lingchun, Li Qingzhao said “I’m afraid the locust canoes in the Shuangxi Brook couldn’t bear all my woe.” Such follow suit is a clear demonstration of Li Yu’s position in Chinese poetry history.

The Page Rank is a classical algorithm in web structure mining and the most representative one of all mining algorithms. Assume the probability is \((I-Q)\) and randomly open a webpage \((w)\). Randomly assign the current page as the page \(w\) at a probability of \(q\). Then, the Page Rank iterative formula shall be:

\[
PR(W) = (1 + q) + q \sum_{e=B} \frac{PR(e)}{|B|} \tag{1}
\]

In which, the parameter \(q\) is the attenuation factor having a scope of \((0, 1)\).

If the TH-Page-Rank algorithm is taken, first of all, we need to classify all pages into different categories in line with the ODP classification rules. Then, based on the category of each webpage, we need to figure out the sensitivity value of a subject and then add the value into the algorithm to obtain the Page Rank Value of the web-pages under the ODP method. Then each web page has a Page Rank value in different ODP category. The equation is shown as follows:

\[
PR_e(a) = (1 - q) \times S_e + q \times \sum_{d=B} \frac{PR(d)}{|B(d)|} \tag{2}
\]

In which, the \(e\) stands for any of the 16 basic classification standards (\(e=1, 2, 3, \ldots, 16\)). In the \(S_e = \frac{1}{|G_e|}\), the \(G_e\) is the out-bound chain web-page set of the current page \(p\) under a category; The \(B(O)\) is the web-page set of in-bound chain to web-page \(P\); The \(Z(d)\) is the out-bound chain web-page set of web-page \(d\). When a query is entered, the algorithm shall use the equation, \(PR_e(p) = \sum v \times PR_e(q), e = 1, 2, 3, \ldots, 16\), to figure out every \(p\) web-page’s final Page Rank Value. In the equation, the \(v\) stands for the probability of the page \(p\) belonging to an ODP classification.

To modify the Page Rank algorithm, we need the help of MP-Page Rank. The equation is shown as follows:

\[
PR_e(p) = (1 + q) \times P_e(e) + q \times \sum_{d=B} PR_e(e) \times G_e(1, e) \tag{3}
\]

In which, the \(G_e(1, e)\) refers to the probability of page transition from page \(i\) to \(e\) under the query subject; And the \(G_e(1, e)\) and \(p_e(e)\) shall be figured out with of the correlation function \((R_v(e))\) of a query \(q\) and a page \(e\).

\[
P_e(e) = \frac{R_v(e)}{\sum_{s=m} R_v(s)}
\]

\[
G_e(1, e) = \frac{R_v(e)}{\sum_{s=f} R_v(s)} \tag{4}
\]

In which, \(M\) stands for the web-page set of the network and the correlation function is indicated with the appearance times of the key words of the enquiry \(q\) in the text in the page \(e\). Also, we need to carry out the similarity calculation of the documents with the formula shown as follows:

\[
sim(x_1, x_2) = \frac{\sum_{i=1}^{m} p_i \times q_i}{\sum_{i=1}^{m} p_i \sum_{i=1}^{m} q_i} \tag{5}
\]

Below is the equation used to obtain the Page Rank values for people adjust the algorithm for key words serving as the corresponding vectors of text.

\[
PR(u) = (1 - q) + q \times \sum_{v=B} PR(v) \times \left[ \partial \times t + (1 - \partial) \times \text{sim}(u, v) \right] \tag{6}
\]

3. IMAGE RETRIEVAL OF CLASSICAL LITERATURE IN THE VIEW OF BIG DATA

3.1 OCR and image retrieval
The first thing about the image retrieval is OCR, i.e. Optical Character Recognition, a process of using electronic equipments or software to translate characters into a language that the computers understand. The image retrieval algorithm is shown in Figure 3:

![Image Retrieval Algorithm](image.png)

Figure 3. Image Retrieval Algorithm

Generally speaking, the problem for the OCR to resolve is the character recognition. Let take Sun Wukong, the hero of *The Journey to the West*, as an example of the character analysis. He has all the virtues and weakness of a monkey, a man and a god. Or, he is a perfect combination of the three. He can cover a distance of 108,000 Li in a blink of eye, in fact, “The time from dawn to dusk was enough for him to travel from the North Sea to Cangwu Mountain in the South.” He can grow very big, as big as a sky-atlas pillar, or very small, as small as mustard seed. And above all, he has 72 metamorphoses. He can shift into the shape of a dragon, snake, tiger or leopard, or a ghost, monster, human or god, and looks very like it in the flesh and spirit (Liu, 2017). Sun has extraordinary talents. When learning from Master Bodhi, he learned way faster than all of his fellow students. He cracked the riddle of the Master Bodhi and went to his bedroom to learn the secret of immortal life at the midnight. During his learning, he achieved a profound understanding over the Buddhist teachings, just as the Tang Monk did. He fears nothing and has the power to overcome all difficulties. He believes in freedom and would fight back as hard as he can in the face of any oppression. He is in himself a monkey that “not governed by Qilin or Feng Huang or any emperor of human society”. Unwilling to swallow insults and humiliation silently, he’d simply rebel the heaven. Unwilling to put his life and fortune in other’s hand, he’d directly go down the nether world and write off his name from the Book of Life and Death. Having no respect to divine power, he’d challenge the authority of the Jade Emperor and the Lord Lao Zi, leaving a heroic saying “I can be the emperor of all kingdoms as long as I want to” (Wang, 2016). Even in front of the Buddha himself, he’s not daunted at all and even made fun of him, saying he’s a relative of a demon. In his eyes, there is no place for gods. At the slightest disappointment or displeasure, he’d say to a god “Stretch your leg and let me give you five sticks as welcome gift. So, I’d be happy.”

3.2 The image retrieval of classical literatures

The second thing about the image is the image retrieval in its real sense. Simply put. The Image Retrieval is to find and extract the image meeting customer requirements from a database storing massive data (Ouyang, 2016). There are a few types of image retrieval. First, the TBIR (Text-based Image Retrieval) defines an image using factors like text description, image name, image size, creator’s era, etc (Zhou and Wang, 2005). Second, the Content-based image retrieval is shown in Figure 4:
For example, the Moonlight on the Spring River is a poem depicting a series of scenes or objects located nearby, faraway, in water, underground and in heaven, giving a vivid description of the spring, river, flowers, the moon and the evening full of the charm of Yangtze Delta in a water-ink scroll painting. In addition to the natural scenes, the poet gave a precise description of the sadness and parting grief of a married woman with several different figures of speech, e.g. extending, exaggerating and implying, to achieve the effects of reviving emotions at the sight of a familiar scene and fusing feelings with the natural scene, and then trigger off a sign of short life in an infinite world and the grief parting with the beloved one for too long while even the moon could put itself together at least once a month, which is very effective in extending the artistic conception into the infinite universe in a man’s heart. The symbols abovementioned frequently appeared in many poems throughout the history. In fact, up to now, there have appeared too many poems about the Moon. It’s self-evident that how important the Moon is for poets to create their artistic conceptions. Even if Zhang Ruoxu chose not to refer the Moon as many times as he did, a few simple reference of the Moon would be good enough for reader to relive everything, because the Moon-based conception has gone too deep into the imagination of Chinese people, the Moon in itself has already become a symbol of beauty. Of course, in this poem, the art presentation of the Moon is too wonderful for words (Yu and Zhao, 2005).

4. CONCLUSION

The development of big data technologies is a milestone marking another huge step that mankind has taken forward in exploring the world. But more importantly, it offered us a great opportunity to take a closer look into the great achievements in our civilization from a data-analysis perspective. Since the world entered the modern times, China either ignored the revolution in human thoughts driven by the technological progress or played a role of learner or follower of such transformation only. This time, in the revolution set off by the development of the big data and cloud computing technologies, China should and can do better since now is the time it is closest to the frontline (Zhou, 2009). It’s true we should remain humble to the traditional reading and writing forms and principles. But it doesn’t mean we should give up the opportunities coming along with the technological advance. In fact, on the basis of the results we have achieved in the digitalization of classical literatures, we should pace up the transformation from the digitalizing ancient cultural heritages to datalizing them with the help of new technologies. It’s time for us to embrace new ideas and roll out a new landscape in the research of humanities and particularly the research of classic literature.

REFERENCES


Li L., Liu Y. (2015). The content of the image retrieval method, Wavelet transform based on computer science, 42 (02), 306-310.


Liu J.C. (2017). The study of literature and geography data in the perspective of "rushuji" - a record for northbound recorded notes as the center, Literature review, (01), 159-170.


