

A Text-Mining Approach to the Authorship Attribution Problem of *Dream of the Red Chamber*

Hsieh-Chang Tu^{*} and Jieh Hsiang^{**}
National Taiwan University

Abstract

Dream of the Red Chamber (DRC), written in the 18th century, is among the greatest Chinese classic novels. Indeed, so many studies have been devoted to this work that the term *Redology* was created to designate this field of research (Pan 1974). In 1921, Hu Shi (胡適) provided solid evidence to show that, of the 120 chapters, the first 80 were written by Cao Xueqin (曹雪芹) based on his life. He also attributed the remaining 40 chapters to Gao E (高鶚) (Hu 1921). While the first conclusion is commonly accepted, the second is not settled.

Researchers have also used statistical methods to study this problem. The most common approaches use pre-defined linguistic features, usually *function words*, to check frequencies of words. Interestingly, however, people came to different conclusions when choosing different features.

We propose a text-mining approach to the DRC author attribution problem. We define a mining function to find terms that clearly show discrepancies between the two corpuses. Some of the terms are semantic in nature, thus avoiding the pitfalls with the more syntactic function words approach. In addition to supporting the claim that the first 80 chapters and the last 40 were written by different authors, a somewhat surprising side result is the evidences that show Chapters 64 and 67, two chapters missing from the oldest existing edition, may also have been written by someone else.

1. Introduction

Authorship attribution is a well-researched subject. Brinegar (1963) used *word length* as the text feature to conclude that the 10 Quintus Curtius Snodgrass letters were not written by Mark Twain. Recent approaches, that usually assume the

^{*} tu@turing.csie.ntu.edu.tw

^{**} jhsiang@ntu.edu.tw

contextual independence of the texts being compared, make use of most frequent words and clustering analysis to identify the most likely author (Peng and Hengartner 2001, Burrows 2002, Hoover 2004, Malyutov 2006, Stamatatos 2009, Jockers and Witten 2010).

A well-known author attribution problem in Chinese literature is the author of the last 40 chapters of the novel *Dream of the Red Chamber*. Past stylistic studies lead to contradictory claims due to different feature selections and experiment designs. Karlgren (1952), Chan (1986), and He (2002) concluded that the entire DRC was written by the same person, while Zhao and Chen (1975), Yu (1998), and Yang (2003) observed significant differences between the first 80 and the last 40 chapters.

Most of these works started from choosing certain linguistic features (usually *function words*). A hypothesis testing method is then deployed to check whether the frequency distributions of features in the first 80 chapters are significantly different from those in the last 40. Yang (2003) used a different approach. They first partitioned DRC into 12 documents, each with 10 chapters. Instead of using pre-defined words, they designed a simple function that used the frequencies of unigrams to associate similarities between each pair of the 12 documents. They found strong similarities in the first 2 documents (containing Chapters 1-20), the next 6 documents (Chapters 21-80), and the final 4 (Chapters 81-120), and thus concluded that the final 40 chapters were written by a different author. However, following the same reasoning, one should also conclude that the first 20 chapters and the middle 60 were written by different authors.

2. Our text-mining approach

We propose a text-mining approach to the DRC author attribution problem. Instead of choosing pre-defined words, we design a mining function to generate candidate words. In addition to term frequencies, we also consider the number of chapters in which a term appears.

2.1 The edition question

The first question is to choose a proper edition. The earliest of DRC (1754) contains merely 16 chapters, and the second, *gengchen* edition (庚辰本) of 1760, has 78 chapters (1-80 except 64 and 67). The earliest existing version with 120 chapters, edited by Cheng Weiyuan and Gao E, appeared in 1791. The full text we chose was

	t	f(t)	A _t	B _t
1	嬖	22.3	34	0
2	裡	22.3	34	0
3	嗎	22.2	1	28
4	展	17.3	26	0
5	疆	14.8	0	11

	t	f(t)	A _t	B _t
1	豈知	31.0	0	24
2	知端	27.9	43	0
3	未知	24.5	1	31
4	一語	22.9	35	0
5	嬖嬖	22.3	34	0

	t	f(t)	A _t	B _t
6	當下	22.3	34	0
7	皆是	21.0	32	0
8	語未	20.4	31	0
9	取笑	19.8	30	0
10	愜記	18.5	0	14

Table 1. The top 5 high-scored unigrams and top 10 bigrams computed by the mining function $f(t)$.

the one provided by YuanZe University,¹ which is the closest to the earliest editions.

2.2 The text-mining function

Regarding each chapter as a document, we use A and B to denote the corpuses of the first 80 chapters and the last 40 respectively. Thus $|A|=80$ and $|B|=40$. We use $t \in d$ if the term t occurs in document d . Let $D_t = \{d: t \in d, d \in D\}$ be the subset of D which contains term t . We call $|D_t|$ the *document frequency* of t in D . We define the *average document frequency* of t , a term, in D , a document set, to be $p_t(D) = |D_t|/|D|$. $p_t(D)$ indicates the average probability for any document in D to contain t .

We define the text-mining function to be

$$f(t) = \frac{\max(p_t(A), p_t(B)) + k}{\min(p_t(A), p_t(B)) + k},$$

where a constant k is added to avoid the case $f(t) = \infty$ when $p_t(A)$ or $p_t(B)$ equals 0. We set $k=0.02$ in our experiments. We assume $p_t(A) \geq p_t(B)$ and use $k=0$ to illustrate how the function works. Then $f(t) = p_t(A)/p_t(B)$ and a big $f(t)$ means a high ratio of $p_t(A)$ to $p_t(B)$. Thus a high-scored $f(t)$ means that the average document frequency of term t in A is significantly different from that in B .

The top 5 unigrams and top 10 bigrams obtained through $f(t)$ are given in Table 1. We have studied the top 30 unigrams and bigrams, which all showed similar behavior.

2.3 Some findings

We now briefly discuss some of our findings. The top-scored unigram *ma* (嬖) occurs only in the form of *mama* (嬖嬖) which we shall discuss later. The second unigram *li* (裡) is interchangeable with another *li* (裏), thus could have been replaced during transcribing and should not be considered. However, we remark that among the 109 appearances of *li* (裡) in the first 80 chapters, 54 of which are in Chapter 67

¹ <http://cls.hs.yzu.edu.tw/hlm/>

alone! This strongly suggests that the current Chapter 67 (missing in the *gengchen* edition) was later added by another person.

The bigrams reveal even more insight. The top scored *qizhi* (豈知), which occurs in 24 chapters in corpus *B* but none in corpus *A*, does not have a clear semantics in itself. The third bigram *weizhi* (未知) appears in 31 chapters of the last 40, and the *only* chapter of the first 80 in which it appears is Chapter 64, another chapter missing in the *gengchen* edition. This provides another evident that both chapters (64 and 67) were added later by someone else. The third example, the fifth bigram *mama* (嬷嬷, a respectful title given to an elder wet nurse) occurs in 34 of the first 80 chapters but none in the last 40. There are many *mama*'s in DRC. They all conspicuously disappeared after Chapter 80.

2.4 The three-author question?

Recall that Yang (2003) also did not chose function words *a priori* and found strong discrepancies between Chapter 1-20, 21-80, and 81-120. Thus, if one is to conclude from their studies that the last 40 chapters were written by a different author, one may also need to declare that the first 80 chapters were also written by two different authors.

To make sure that our method does not pose similar problems, we ran the same experiment between the texts of Chapters 1-20 and Chapters 21-80. Not surprisingly, we found some unigrams and bigrams that appear in one corpus but not in the other. A careful analysis, however, shows that they are mostly event-dependent, involving persons or places that appeared later in the story or died. Considering that there are more than 400 characters in DRC, such event-dependent differences are expected.

3. Discussions

Our studies support the thesis that the last 40 chapters of DRC were written by someone other than Cao Xueqin. It also shows that Chapters 64 and 67 may also have been written by another person. Furthermore, the text-mining method we used offers a different approach to the author attribution problem.

A common textual analysis approach is to use function words to detect discrepancies in different texts. For instance, in Chinese *ma* (嗎) as a function word has the equivalent *me* (麼). Suppose one uses *ma* as a proof that an article is not written by a certain person, can the verdict be overturned if one uniformly replaces all the function word occurrences of *ma* by *me*?

The text-mining approach proposed here is different. Although it is also based

on differences in word style, the words are generated by the method itself. Take the term *mama* as an example. *Mama* appeared in 34 of the first 80 chapters, and was used to address quite a few minor characters in the book (last appearance in Chapter 80). However, not only was the term completely missing from the last 40 chapters, so did the concept and the characters! Such "semantic" differences seem to provide more solid evidence than purely syntactic ones.

References

- Claude S. Brinegar (1963), *Mark Twain and the Quintus Curtius Snodgrass Letters: A Statistical Test of Authorship*, Journal of the American Statistical Association, Vol. 58, No. 301, pp.85-96.
- John Burrows (2002), '*Delta*': a Measure of Stylistic Difference and a Guide to Likely Authorship, Literary and Linguistic Computing, Vol. 17, No. 3, 2002.
- Bing-Cho Chan (1986), *The authorship of the Dream of the red chamber based on a computerized statistical study of its vocabulary*, Joint Publishing Co Ltd., Hong Kong.
- Guang-Gu He (2002), *From Chinese function words to the characteristics of authors – also the author attribution problem of Dream of the Red Chamber*, Traditional Chinese Literature e-Journal, Hualian.
- David L. Hoover (2004), *Testing Burrows's Delta*, Literary and Linguistic Computing, Vol. 19, No.4, 2004.
- Shi Hu (1921), *Textual Research on the Dream of the Red Chamber*, reprinted by Yuandong Publishing, 1985.
- Matthew L. Jockers and Daniela M. Witten (2010), *A comparative study of machine learning methods for authorship attribution*, Literary and Linguistic Computing, Vol. 5, No. 2.
- Bernhard Karlgren (1952), *New Excursions in Chinese Grammar*, Bulletin of the Museum of Far Eastern Antiquities (Stockholm), No. 24, pp. 51-80.
- M.B. Malyutov (2006), *Authorship attribution of texts: a review*. General Theory of Information Transfer and Combinatorics, Springer-Verlag, pp. 362-380.
- Chong-Gui Pan (1974). *Sixty years of Redology*. 2,226 pages. Wen-shi-je Publishing, Taipei.

Roger Peng and Nicolas Hengartner (2001), *Quantitative Analysis of Literacy Styles*.
Department of Statistics Papers, Department of Statistics, UCLA.

Efstathios Stamatatos (2009), *A Survey of Modern Authorship Attribution Methods*,
Journal of the American Society for Information Science and Technology,
Volume 60 Issue 3, March 2009, pp. 538-556.

Albert C.-C. Yang, C.-K Peng, H.-W. Yien, Ary L. Goldberger (2003), *Information
categorization approach to literary authorship disputes*, Physica A 329,
pp. 473-483.

Qing-Xiang Yu (1998), *Applications of Statistical methods to Dream of the Red
Chamber*, Journal of National Cheng-Chi University, vol 76, pp. 303-327.

Gang Zhao and Zhongyi Chen (1975),《紅樓夢研究新編》 *A New Compilation on the
research of The Dream of The Red Chamber*, Linking Publishing, Taipei.