Function Words for Chinese Authorship Attribution

Bei Yu
School of Information Studies
Syracuse University
byu@syr.edu

Abstract

This study explores the use of function words for authorship attribution in modern Chinese (C-FWAA). This study consists of three tasks: (1) examine the C-FWAA effectiveness in three genres: novel, essay, and blog; (2) compare the strength of function words as both genre and authorship indicators, and explore the genre interference on C-FWAA; (3) examine whether C-FWAA is sensitive to the time periods when the texts were written.

1 Introduction

Function words are an important feature set for Authorship Attribution (hereafter "AA") because they are considered *topic-independent* or *context-free*, and that they are largely used in an unconscious manner (Holmes, 1994; Stamatatos, 2009; Koppel et al., 2009). The *Federalist Papers* (Mostellar and Wallace, 1964) may be the most famous example of AA in English. Mostellar and Wallace (1964) conducted a detailed study of searching and testing function words to distinguish Hamilton and Madison as the authors of the disputed Federalist Papers.

Although Function Word based Authorship Attribution (hereafter "FWAA") has been successful in many studies (Stamatatos, 2009), Juola (2008) argued that FWAA are mainly applied in English texts, and it may not be appropriate for other highly inflected languages, like Finnish and Turkish. This may not be the case in that it is the content words, not the function words, that are inflected in those languages. However, function words are indeed rarely used for AA in non-English texts. It was left out in the comprehensive authorship analysis of *The Quiet Don* (in Russian) by Kjetsaa et al. (1984). The literature review for this study found several examples of FWAA in Modern Greek (Mikros and Argiri, 2003) and Arabic (Abbasi and Chen, 2005). Overall, the effectiveness of FWAA has not been tested on many languages.

Some studies on FWAA also reported negative results. Holmes (1994), in his comprehensive survey on authorship attribution, cited doubts given by (Damerau, 1975) and (Oakman, 1980), and called for further investigation on the stability of function word use within an author's work and between works by the same author.

Another problem for FWAA is to explain exactly what authorial characteristics are captured by function words, since function words may also characterize other textual properties like genre, author gender, and even topic, although function words are generally considered *topic-independent* or *context-free* (Stamatatos, 2009; Herring and Paolillo, 2006; Clement and Sharp 2003; Mikros and Argiri, 2007).

Clement and Sharp (2003) found that function words worked as well as content words in identifying document topics. Their further investigation showed that author and topic are not arbitrarily orthogonal to each other. Using the significance level of two-way ANOVA test as measure, Mikros and Argiri (2007) found that some function words in Modern Greek can distinguish both topic and author, providing further evidence for possible topic-author correlation based on function word dimensions.

Function words are also used as indicators for author gender (Argamon et al., 2002; Koppel et al., 2003) and text genre (Biber, 1993). Koppel et al. (2003) found gender preference on certain personal

Workshop on Computational Linguistics for Literature, pages 45–53,
Montréal, Canada, June 8, 2012.

pronouns and prepositions. Herring and Paolillo (2006) repeated Argamon and Koppel's experiment by mixing genre and gender in the data set, and discovered that the same gender indicators actually captured genre characteristics.

In summary, related work has shown that function words may contribute to distinguishing topic, authorship, author gender, and genre. A question soon emerges: which dimension do function words characterize the most saliently? In other words, given a document set of mixed author, topic, and genre, would they interfere with each other in classification tasks? Answer to this question would help guide experiment design for AA tasks, and explain the real authorial characteristics captured by function words.

This paper aims to study the use of function words for Chinese authorship attribution (C-FWAA), since FWAA has not been well-studied in Chinese. Existing studies of C-FWAA are limited to the analysis of famous authorship dispute cases like whether Gao E or Cao Xueqin wrote the last 40 chapters of *the Dream of the Red Chamber*, and no consensus was reached among these C-FWAA studies (Zeng and Zhu, 2006). Therefore no baseline was available yet for general-purpose C-FWAA studies.

This study consists of three tasks. First, examine the effectiveness of C-FWAA in three genres of creative writing: novel, essay, and blog. Second, compare the strength of function words as both genre and authorship indicators, and explore the genre interference on C-FWAA. Third, examine whether C-FWAA is sensitive to the time periods when the texts were written.

The third task is proposed for a unique reason that the influence of ancient Chinese (文言文) on modern Chinese (白话文) may affect function word use. For example, "also" corresponds to "亦" in ancient Chinese, and "也" in modern Chinese. "的" ("'s" or "of"), "地" ("-ly"), and "得" ("so") are only used in modern Chinese. The government of Republic of China (RoC, 1912-1949) and the government of People's Republic of China (PRC, 1949-) both made changes to the Chinese language. Hence the hypothesis is that Chinese function word use may also reflect the time period of literary works.

2 Experiment set up

2.1 Constructing Chinese function word list

Various function word lists have been used in AA tasks in English, and the selection process usually follows arbitrary criteria (Stamatatos, 2009). To construct the Chinese function word list, this study chose 300 most frequent characters from Jun Da's Modern Chinese Character Frequency List (Du, 2005), removing the characters that contain solid meaning, e.g. "来" ("to come"), and removing all personal pronouns, e.g. "我" ("myself") in that they have been known as genre/register indicators (Biber, 1993). This screening process resulted in 35 function words (see Table 1). Detailed English translation can be found in (Du, 2005).

Every text document was then converted to a vector of 35 dimensions, each corresponding to one function word. The value for each dimension is the corresponding function word's number of occurrences per thousand words.

的 / of	是 / be,yes	不 / no	了/*
在 / at/in	有 / exist	这 / this	为 / for
地 / -ly	也 / also	得 / so	就 / then
那 / that	以 /**	着 / ***	之 / of
可 / can	么 / question	而 / but	然 / so
没 / no	于 / at	还 / also	只 / only
无 / no	又 / also	如 / if	但 / but
其 / it	此 / this	与 / and	把 / hold
全 / all	被 / passive	却 / but	

Note: * completion mark; ** according to; *** on-going status mark

Table 1: Chinese function word list

2.2 EM clustering algorithm

This study chose EM clustering algorithm as the main method to evaluate the effectiveness of C-FWAA. Most AA studies use supervised learning methods in that AA is a natural text categorization problem. However, training data may not be available in many AA tasks, and unsupervised learning methods are particularly useful in such cases. In addition, this study aims to examine the clusters emerging from the data and explain whether they represent authors, genres, or time periods.

This study uses Weka's Simple EM algorithm for all experiments. This algorithm first runs k-Means 10 times with different random seeds, and then chooses the partition with minimal squared error to start the expectation maximization iteration. Weka calculates the clustering accuracy as follows: after clustering the data, Weka determines the majority class in each cluster and prints a confusion matrix showing how many errors there would be if the clusters were used instead of the true class (Witten et al., 2011).

2.3 Selecting writers and their works

To exclude gender's affect, all writers chosen in this study are males. Parallel analysis for female writers will be conducted in future work.

Representative writers from three different time periods were selected to examine the relationship between time period and function word use. The first time period (TP1) is the 1930-40s, when modern Chinese replaced ancient Chinese to be the main form of writing in China, and before the PRC was founded. The second time period (TP2) is the 1980-90s, after the Cultural Revolution was over. The third time period (TP3) is the 2000s, when the publishing business has been strongly affected by the free-market economy. Three representative writers were chosen for each time period. The time period from the foundation of PRC (1949) to the end of the Cultural Revolution was excluded from this study because during that time most literary works were written under strong political guidelines. Tables 2 and 3 listed the representative writers and their selected works. Two long novels are separated into chapters in order to test whether C-FWAA is able to assign all chapters in a book to one cluster. Common English translations of the titles are found through Google Search. Chinese Pin Yin was provided for hard-to-translate titles.

All writers have to meet the requirements that their works cross at least two genres: fiction (novel) and non-fiction (essay). The TP3 (2000s) writers should have well-maintained blogs as well. Therefore this study will examine C-FWAA effectiveness in three genres: novel, essay, and blog.

All electronic copies of the selected works were downloaded from online literature repositories such as YiFan Public Library[1] and TianYa Book[2].

[1] URL http://www.shuku.net:8082/novels/cnovel.html

Time period	Authors
TP1 (1930-40s)	沈从文(Shen CongWen, SCW) 钱钟书(Qian ZhongShu, QZS) 汪曾祺(Wang ZengQi, WZQ)
TP2 (1980-90s)	王朔(Wang Shuo, WS) 王小波(Wang XiaoBo, WXB) 贾平凹(Jia PingWa, JPW)
TP3 (2000s)	郭敬明(Guo JingMing, GJM) 韩寒(Han Han, HH) 石康(Shi Kang, SK)

Table 2: selected writers in three time periods

TP	Writer	#Novels	essays	blogs
1	汪曾祺[3] WZQ	5	6	
	钱钟书 QZS	14*	10	
	沈从文 SCW	11**	7	
2	王朔 WS	5	16	30
	王小波 WSB	3	10	
	贾平凹 JPW	3	10	
3	郭敬明 GJM	8	6	
	韩寒 HH	5	11	92
	石康 SK	4	14	30

Note: *one long novel 围城(*Fortress Besieged*) is separated into 10 chapters. **one long novel 边城 (*Border Town*) is separated into 7 chapters.

Table 3: statistics of selected works

3 Experiment and result

3.1 Test the effectiveness of EM algorithm for FWAA

The first experiment was to test the effectiveness of the EM algorithm for FWAA. The famous *Federalist Papers* data set was used as the test case. The *Federalist Papers* experiment was repeated using the function words provided in (Mostellar and Wallace, 1964). The original *Ferderalist Papers* and their author identifications were downloaded from the Library of Congress website[4]. Function words were extracted using a Perl script and the word frequencies (per thousand words) were calculated. The 85 essays consist of 51 by Hamilton, 15 by Madison, 3 jointly by Hamilton

[2] URL http://www.tianyabook.com/

[3]汪曾祺(Wang Zengqi) is an exception in that his writing career started in the 1930s but peaked in the 1980s.

[4] URL: http://thomas.loc.gov/home/histdox/fedpapers.html

and Madison, 5 by Jay, and 11 with disputed authorship. Mosteller and Wallace (1964) supported the opinion that Madison wrote all 11 disputed essays, which is also the mainstream opinion among historians.

In the first round of experiment, Jay's five essays and the three jointly-written ones were excluded, making the task easier. The cluster number was set to two. EM returned results similar to that in (Mostellar and Wallace, 1964) by assigning all disputed papers to Madison (Table 4). However it did make several mistakes by assigning 3 Hamilton's essays to Madison and one Madison's essay to Hamilton, resulting in an overall accuracy of (66-4)/66=94% in the not-disputed subset.

	C0 (Hamilton)	C1 (Madison)
Hamilton	48	**3**
Madison	**1**	14
Disputed	0	11

Table 4: Hamilton vs. Madison (clustering errors in bold)

In the second round Jay's five essays were added to the test data. The cluster number was then changed to three. The EM algorithm successfully attributed the essays to their real authors with only one error (assigning one Madison's essay to Jay, see the confusion matrix in Table 5). It also assigned all disputed essays to Madison. The 3-author AA result in Table 4 seems even better than the 2-author AA result, but the difference is small.

	C 0	C1	C 2
Hamilton	51	0	0
Madison	0	14	**1**
Jay	0	0	5
Disputed	0	11	0

Table 5: Hamilton vs. Madison vs. Jay

In the third round the three jointly-written essays were added to the test data. These jointly-written essays may resemble either Hamilton or Madison, which would result in 3 clusters still, or they may exhibit a unique style and thus form a new cluster. The test result shows that these three jointly-authored essays did confuse the algorithm no matter if the cluster number is set to three or four. When setting the cluster number to three (Table 6), all three joint essays were assigned to C2, which also attracted 11 Hamilton's, 2 Madison's, 2 Jay's, and 1 disputed essays. Increasing the cluster number to 4 does not reduce the confusion: Hamilton still dominated Cluster 0 with 40 out of 51 essays in it; C1 is still dominated by Madison (13 out of 15) and the disputed essays (9 out 11). Jay's essays were split into C1 and C2. This result actually shows that function words are highly sensitive to noise like the jointly-written essays.

	C0	C1	C2
H-M	0	0	3
Hamilton	40	0	11
Madison	0	13	2
Jay	0	3	2
disputed	1	9	1

Table 6: impact of the jointly-written essays

3.2 Chinese FWAA with genre and time period controlled

This section describes the experiments and results for task 1: evaluating the effectiveness of C-FWAA using EM and the 35 Chinese function words as features. Controlling the time period and genre, the same experiment was repeated for each genre and each TP.

In the first round, the authors within each TP were paired up in the novel genre to distinguish them, which is expected to be easier than distinguishing multiple authors. The results in Table 7 show that the authors of TP1 and TP2 novels are perfectly distinguishable, but those in TP3 are not.

Compared to the writers of TP1 and TP2, writers in TP3 face a new market-driven economy. Writing-for-profit becomes acceptable and even necessary for many writers. TP3 writers like Han Han (HH) and Guo JingMing (GJM) obtained huge financial success from the publication market. Both of them also received doubts regarding the authenticity of their works.

Guo Jingming was found to plagiarize in his book *Meng Li Hua Luo Zhi Duo Shao*, which was also not assigned to his main cluster by C-FWAA. Guo JingMing founded a writing studio and hired

employees to publish and market his books. He publicly admits the existence of "group writing" practice in his studio because his name is used more as a brand than as an author.

C-FWAA also encountered difficulty in distinguishing Han Han and Shi Kang's novels. This finding is consistent with the fact that Han Han publicly acknowledged that his *Xiang Shao Nian La Fei Chi* mimicked Shi Kang's style. Since the beginning of 2012, a huge debate surged in Chinese social media over whether Han Han's books and blogs were ghost-penned by his father and others. In this striking "crowd-sourcing Shelock Holmes" movement, numerous doubts were raised based on netizens' amateur content analysis on contradicting statements in Han Han's public videos and different book versions. A separate study is undergoing to analyze the stylistic similarity between Han Han and the candidate pens.

As described in Section 3.1, FWAA is highly sensitive to noise like joint authorship. This may explain the low performance of C-FWAA in TP3 when plagiarism, group writing, and ghostwriting are involved.

After C-FWAA on the novel genre, the same experiment was then repeated on the other two genres: essay and blog. The results in Table 7 show an average accuracy .87 for essays and .83 for blogs. Overall, this round of experiment demonstrates that C-FWAA is effective in distinguishing two authors in all genres and time periods.

	Author pair	Novel	Essay	Blog
TP1	WZQ-SCW	1	.77	
	SCW-QZS	1	.94	
	WZQ-QZS	1	.81	
TP2	WS-JPW	1	1.00	
	WS-WXB	1	.96	
	WXB-JPW	1	.85	
TP3	GJM-HH	.77	1	
	GJM-SK	.75	.65	
	HH-SK	.56	.84	.84
TP2-3	HH-WS			.77
	SK-WS			.88
avg		.90	.87	.83

Table 7: pair-wise C-FWAA

In the second round C-FWAA was tested on the task of distinguishing three authors, also starting from the novel genre and TP1. In the 3-cluster result (Table 8), C0 is devoted to SCW's novel 边城 (*Border Town*), a masterpiece in Chinese literature, C1 captured all other SCW novels, and WZQ and QZS remain in C2 together. WZQ and QZS were further separated after increasing the cluster number to four (with only two errors, highlighted in Table 8, of assigning QZS's two works *God's Dream* and the Foreword of *Fortress Besieged* to SCW). Two long novels that are separated into chapters are also successfully assigned into same clusters except for the Foreword of *Fortress Besieged*.

The 3-author experiment was then repeated on TP2 and obtained 100% accurate results.

The 3-author AA result for TP3 is similar to its 2-author result: HH and SK remain in one cluster. When increasing the cluster number to 4, GJM still dominated C0 and C1, but now HH and SK were separated into C2 and C3 respectively.

The C-FWAA accuracy was then calculated by choosing the better result from 3-cluster and 4-cluster experiments (Table 8). Overall, C-FWAA is able to distinguish three authors in the novel genre effectively.

30s-40s	cluster num = 3			cluster num = 4			
	C0	C1	C2	C0	C1	C2	C3
SCW	7	4	0	7	4	0	0
WZQ	0	0	5	0	0	5	0
QZS	0	0	14	0	**2**	0	12

2000s	cluster num = 3			cluster num = 4			
	C0	C1	C2	C0	C1	C2	C3
GJM	4	3	1	4	3	**1**	**0**
HH	1	0	4	**1**	0	3	**1**
SK	0	1	3	0	**1**	0	3

TP	Accuracy
TP1	28/30=.93
TP2	11/11=1.00
TP3	13/17=.76
Avg	.90

Table 8: 3-author C-FWAA on Chinese novels

The above experiment was then repeated on the essay and blog genres. In the essay genre, the

average 3-author C-FWAA accuracy is .83, .89, .84 for TP1, TP2, and TP3 respectively (Table 9), average accuracy .85. For blogs the accuracy is .68 (Table 10).

30s-40s	TP1				2000s	TP2		
	C0	C1	C2			C0	C1	C2
SCW	5	**2**	0		GJM	6	0	0
WZQ	0	6	0		HH	0	11	0
QZS	0	**2**	8		SK	**1**	**4**	9

80s-90s	cluster num = 3			cluster num = 4			
	C0	C1	C2	C0	C1	C2	C3
WS	16	0	0	15	0	**1**	0
WXB	0	10	0	0	8	**1**	**1**
JPW	2	4	4	0	0	**1**	9

Time period	Accuracy
1930s-1940s	19/23=.83
1980s-1990s	32/36=.89
2000s	26/31=.84
Average	.85

Table 9: 3-author C-FWAA on Chinese essays

Acc=104/152=.68	C0	C1	C2	C3	C4
HH	63	**7**	**8**	**2**	**12**
WS	**11**	13	**1**	0	**5**
SK	**1**	**1**	28	0	0

Table 10: 3-author C-FWAA on Chinese blogs

Comparing the C-FWAA accuracy on three genres, we can see that function words are quite effective in distinguish writers in all three genres. It is the most effective in novels, then essays, and blogs are the hardest. One possible explanation is that novels are the longest, essays are shorter, and blogs are the shortest. Hence novels provide the largest amount of data for precise measure of authorial characteristics. Further examination is needed to test this hypothesis. Another possible explanation is that blogs pose less constraint on the writers with regard to the writing format, and thus writers may write in much freer and more informal style. Overall, C-FWAA reached over 80% accuracy in distinguishing two or three authors in all three genres. This concludes the task #1.

3.3 Function words as genre indicators with author and time period controlled

This section reports a series of experiments that aim to evaluate the effectiveness of function words as genre indicators and the genre interference on C-FWAA. The first round of experiment examines whether the function words can distinguish novels from essays in each TP. The cluster number was set to two and the clustering result was compared against the genre labels. The error analysis also reveals which genre is less cohesive (failing to hold all of its instances in one cluster).

TP	Author	Accuracy	Which genre is less cohesive?
TP1	WZQ	.73	Essay (3->novel)
	SCW	.78	Essay (3->novel)
	QZS	1	
TP2	JPW	.54	Essay (7->novel)
	WS	1	
	WXB	.85	Essay (2->novel)
TP3	GJM	.71	Novel (4->essay)
	HH	.63	Both (5 essay->novel; 1 novel->essay)
	SK	.66	Essay (2->novel)
	avg	.77	

Table 11: function words as genre indicator (novel vs. essay)

The results in Table 11 show that the average accuracy (over 9 authors) is .77 to distinguish an author's novels and essays, demonstrating that function words are also strong genre indicators. For some authors QZS, WS, and WXB, their novels and essays are highly separable based on function word use. Interestingly, for all writers, their novels hold together perfectly except for GJM, but the essays often spread across two clusters. Again, the explanation may still be that novels are longer than essays, and thus provide more precise style estimation. If so, novels and essays may not be a fair comparison. However, the lengths of essays and blogs are similar. Therefore, the above experiment was repeated to distinguish essays and blogs from same authors. The results in Table 12 show that this task is not easier. The average accuracy is .71, which is a little worse than .77 in distinguishing novels and essays. Once again, one genre, this time it is the essay, that hold together very well, and blogs spread across clusters.

Combining the results in Section 3.2 and this section, we can see that function words are indicators of both authorship and genre, and the C-FWAA performance is affected by genre: it is the easiest for novel, then essay, and hardest for blogs.

Author	Acc	#E->B	#B->E
WS	.80	0/16	9/30
HH	.56	0/11	58/92
SK	.78	5/14	5/31
Avg	.71	.12	.36

Table 12: function words as genre indicator (essay vs. blog)

3.4 Which one do function words characterize more saliently, genre or authorship?

In the experiments reported in this section TP was still controlled, but in each TP the three authors and two genres are mixed together. The experiment was repeated for each TP. Each experiment consists of two steps. First, the cluster number was set to two, and the clustering result was compared against the genre labels. Second, the cluster number was set to three, and the result was compared against the author labels. If genre plays stronger impact on function word use, we should see high accuracy in the 2-cluster result, and if authorship is more salient, the 3-cluster result should be better. The results show that for all three TPs, the author-genre mix decreased the performance of authorship clustering (column #3 "AA in mixed genres" vs. column #4 "AA in novel" and column #5 "AA in essay"), indicating clear genre interference to authorship attribution. In comparison, the genre clustering in mixed authors (column #1) was worse than genre clustering in single author (column #6) in TP1 only. In TP2 and TP3 genre clustering in mixed-authors yielded higher accuracy than that in single-author, showing that mixing authors may increase or decrease genre identification performance.

To better understand the interference between authorship and genre, the 3-cluster result for each TP was visualized in Figures 1-3. The clusters in TP1 (Figure 1) include authorship cluster C0 (bottom row: SCW), genre cluster C2 (top: essay), and mixed cluster C1 (middle: WZQ, QZS, novels, and essays), demonstrating competing influence of authorship and genre on function words. The clusters in TP2 (Figure 2) are more genre-oriented, with C0 dominated by novels and C1 and C2 by essays. The clusters in TP3 (Figure 3) are also as mixed as in TP1, but more authorship-oriented, with C0 dominated by Shi Kang, C1 by Guo JingMing, and C2 by Han Han. In summary, function words characterize authors more saliently in TP1 and TP3, and genres more saliently in TP2. Therefore, we conclude for task #2 that the level of genre interference on authorship attribution is not arbitrary but is actually dependent on individual data set.

	2-genre clustering	3-author clustering	Novel AA	Essay AA	N-E genre
TP1	.51	.64	.93	.83	.84
TP2	.89	.70	1.00	.89	.80
TP3	.70	.75	.76	.84	.67

Table 13: genre vs. authorship

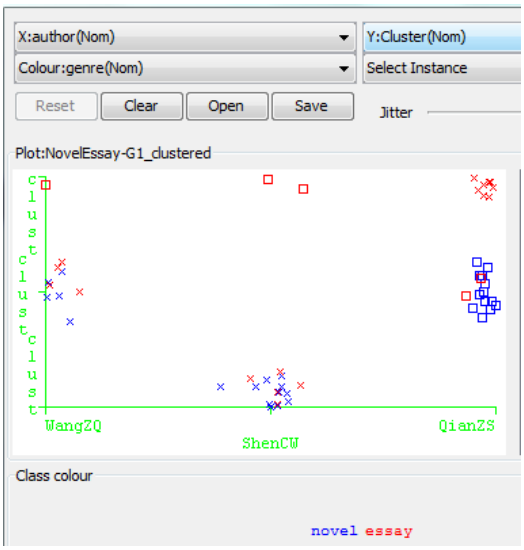

Figure 1: mixing authorship and genre in TP1

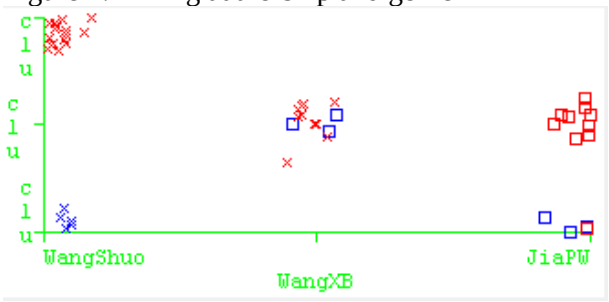

Figure 2: mixing authorship and genre in TP2

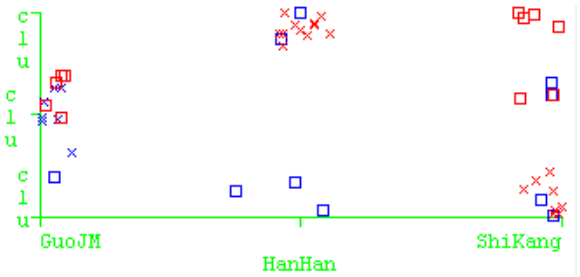

Figure 3: mixing authorship and genre in TP3

3.5 Is C-FWAA dependent on time period?

The task #3 is to examine whether C-FWAA is dependent on time period. The hypothesis is that writers of different times may use the function words differently because of the drastic change in Mandarin Chinese throughout the 20th century. When mixing the novels written in TP1, TP2, and TP3, the algorithm may be more sensitive to the time period than individual authorship. If the hypothesis is true, we should see the clustering result aligns with the time period, not authorship or genre. This time the cluster number is set to -1, which allows EM to use cross validation to automatically determine the optimal number of clusters (Smyth, 1996; McGregor et al., 2004).

EM returns 4 clusters: C0 is dominated by QZS (1940s), C1 by WZQ, WS, and JPW (1980-90s), C2 by SCW (1930s) and WXB (1980-90s), C3 by GJM (2000s). Therefore no obvious relationship is observed between the clusters and the time periods. Further, all TP1 and TP2 writers share one thing in common – their works stay in one cluster, but TP3 writers' works spread across multiple clusters: GJM 2, SK 3, and HH 4. This result is consistent with two facts that Han Han publicly acknowledged that (1) his *Xiang Shao Nian La Fei Chi* mimicked Shi Kang's style, and (2) his *San Chong Men* mimicked Qian ZhongShu's *Wei Cheng*.

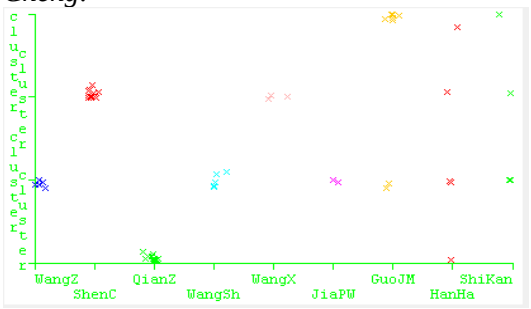

Figure 4: clustering all novels from 9 authors

Repeating the experiment on essays resulted in only two clusters. Most writers' essays remain in one cluster with few exceptions (e.g. SCW, QZS, WXB and JPW in C0, and WZQ, WS and GJM in C1), while HH and SK's essays spread across the two clusters. The clusters do not seem to relate to the time periods either. What do these two clusters mean then? An examination of the cluster assignment of HH's essays reveals that his essay books *Du, Jiu Zhe Yang Piao Lai Piao Qu*, and *Ke Ai De Hong Shui Meng Shou* belong to C1, all written in casual and conversational style, and the more formal essays like *Qiu Yi*, *Shu Dian*, *Bei Zhong Kui Ren*, and *Yi Qi Chen Mo* belong to C1. Interestingly, most essays in C1 are doubted to be penned by his father. This result suggests that the clustering result actually captured two sub-genres in essays. However, further analysis is needed to test this hypothesis. In summary, no solid relationship was found between time period and Chinese function word use.

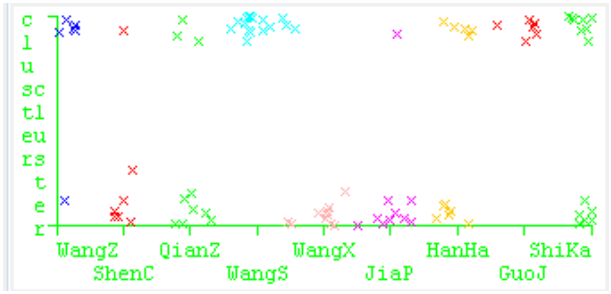

Figure 5: clustering all essays from 9 authors

4 Conclusion and limitations

This study made three contributions. First, it examined the effectiveness of using function words for Chinese authorship attribution (C-FWAA) in three different genres: novel, essay, and blog. Overall C-FWAA is able to distinguish three authors in each genre with various level of success. C-FWAA is the most effective in distinguishing authors of novels (averaged accuracy 90%), followed by essay (85%), and blog is the hardest (68%). Second, this study confirmed that Chinese function words are strong indicators of both genre and authorship. When the data set mixed authors and genres, these two factors may interfere with each other, and in such cases it depends on the data set which factor do function words characterize more saliently. Third, this study examined the hypothesized relationship between time period and

Chinese function word use in novels and essays between 1930s and 2000s, but did not find evidence to support this hypothesis.

This study has several limitations that need to be improved in future work. First, the data set is small and not quite balanced. More authors and works will be added in the future. Second, the random seed for EM is set to the default value 100 in Weka. However, EM clustering result may vary to some extent with different random seeds. More rigorous design is needed for robust performance comparison. One design is to run each clustering experiment multiple times, each time with a different random seed. The clustering accuracy will be averaged over all runs. This new design will allow for performance comparison based on paired-sample t-test significance. Third, the Cultural Revolution time period is excluded from this study due to strong political influence on writers. One reviewer pointed out that this time period should be valuable for examining the relationship between authorship, genre, and time period. Relevant data will be collected in future study.

5 Acknowledgment

Sincere thanks to Peiyuan Sun for his assistance in data collection and the anonymous reviewers for the insightful comments.

References

Ahmed Abbasi & Hsinchun Chen. 2005. Applying Authorship Analysis to Extremist-Group Web Forum Messages. *IEEE Intelligent Systems*, September/October 2005, 67-76.

Shlomo Argamon, Moshe Koppel, Jonathan Fine and Anat Rachel Shimoni. 2003.Gender, genre, and writing style in formal written texts. *Text* 23: 321–346.

Ross Clement and David Sharp. 2003. Ngram and Bayesian Classification of Documents for Topic and Authorship. Literary and Linguistic Computing, 18(4):423-447

Jun Da. Modern Chinese Character Frequency List. 2005. http://lingua.mtsu.edu/chinese-computing/statistics/char/list.php?Which=MO

Fred J. Damerau. 1975. The use of function word frequencies as indicators of style. *Computers and Humanities*, 9:271-280

Susan C. Herring and John C. Paolillo. 2006. Gender and Genre Variation in Weblogs. Journal of Sociolinguistics, 10(4):439-459.

David I. Holmes. 1994. Authorship Attribution. Computers and Humanities, 28:87-106.

Patrick Juola. 2008. Authorship Attribution. Foundations and Trends in Information Retrieval, 1(3):233-334.

Geir Kjetsaa, Sven Gustavsson, Bengt Beckman, and Steinar Gil. 1984. *The Authorship of The Quiet Don*. Solum Forlag A.S.: Oslo; Humanities Press: New Jersey.

Moshe Koppel, Shlomo Argomon, and Anat Rachel Shimoni. 2002. Automatically Categorizing Written Texts by Author Gender. *Literary and Linguistic Computing* 17:401–412.

Moshe Koppel, Jonathan Schler, and Shlomo Argamon. 2009. Computational Methods for Authorship Attribution. *JASIST*, 60(1):9-26.

Anthony McGregor, Mark Hall, Perry Lorier, and James Brunskill. 2004. Flow Clustering Using Machine Learning Techniques. *PAM 2004, LNCS 3015*, 205-214. Springer-Verlag: Berlin.

George K. Mikros & Eleni K. Argiri. 2007. Investigating topic influence in authorship attribution. *Proceedings of the SIGIR'07 Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection*, 29–35.

Frederick Mosteller and David L. Wallace. 1964. *Inference and Disputed Authorship: The Federalist*. CSLI Publications.

Robert L. Oakman. 1980. *Computer Methods for Literary Research*. University of South Carolina Press, Columbia, SC.

Efstathios Stamatatos. 2009. A Survey of Modern Authorship Attribution Methods. *JASIST*, 60(3):538-556.

Padhraic Smyth. 1996. Clustering Using Monte Carlo Cross-Validation. *Proceedings of KDD'96*, 126-133.

Ian Witten, Eibe Frank, and Mark A. Hall. 2011. *Data Mining: Practical Machine Learning Tools and Techniques*. 3rd edition. Morgan-Kaufmann.

Yi-ping Zeng and Xiao-wen Zhu. 2006. Application of computational methods to the Study of Stylistics in China. *Journal of Fujian Normal University (Philosophy and Social Science Edition)*, 136 (1): 14-17.