SCHOOL OF
ELECTRICAL AND
ELECTRONIC ENGINEERING

THE UNIVERSITY of ADELAIDE

Who Wrote the Letter to Hebrews?

Data Mining for Detection of Text Authorship

Stage 2 Progress Report

Zhaokun Wang (1167141)

Supervisor:

Professor Derek Abbott

Dr Brian Ng

Partners:

Kai He (1179552)

Yan Xie (1164699)

Date Submit:

28th Oct 2011

Acknowledgments

This project is supervise and supported by Professor Derek Abbott who has provided the project group with valuable information. The project is also supervised by Dr Brian Ng who has provided us with reasonable guidance and clear instructions.

Executive Summary

The aim of this project is to detect text authorship using data mining. The project group has proposed two approaches to perform the method of detection about the text authorship.

This progress report is focus on designing, coding and programming with different algorithms. Over the past eight weeks, the project group has made significant progression, with the project on the schedule and under budget.

In the future, the key missions have been addressed for the upcoming semester, we will concentrate on the algorithm comparison. Up to now, the project is mostly likely to be delivered on time and under budget.

Table of Contents

List of Figures

List of Tables

1 Introduction

1.1 Aim

The aim of this project is solve the controversy “Who Wrote the Letter to Hebrews?” There are two extraction algorithms to perform text of features. They are Common N-Grams (CNG) and Word Sequences. Three text classifiers which are Dissimilarity Calculation, Naïve Bayes and Support Vector Machine (SVM), are for purpose of category. The detection of authorship contributes the problem of identifying the author of text whose authorship is still in doubt.

1.2 Objectives

The objectives of the project are listed as below:

- Implement two different algorithms in authorship attribution based on character-level and word-level.
- Using to approaches to compare performance.
- Apply our method to identify the authorship detection.

1.3 Proposed Algorithms

Up to previous studies, proposed extraction algorithms have been identified. There are common n-gram and work sequence.

- Common n-gram is based on character level, which is independent of language. Varying of n value would improve the accuracy of the featured text.
- Maximum Work sequence is to extract the sequence of words in grouped sentences. Variable attempts are based on lexical level and extract the function words and content words. From performance with two difference type of words to figure out the author.

1.4 Technical Challenge Identified

In the process of this project, the following technical challenge has been identified and explained more details in section 2.

- Text pre-processing extraction and Design the dominant N-value
- Extraction algorithms of Common N-Gram programming
- Text classifier design

2 Current Project Progress Status

This section reports the current status of this project from the early development. According to work breakdown structure, this report focuses on development of extraction algorithms for Common N-Grams (CNG). The following section presents algorithms development progress over the past eight weeks.

2.1 Further Research and Proposed Approach

Based on character level, CNG has been widely used in authorship contribution due to advantage of language-independent. The aim of CNG is to build byte-level character n-grams author profile of their work. Hence the factor about space, lowercase and uppercase letter does not effect the result.

Before beginning the CNG algorithm, text pre-processing for the preparation plays an important role, which could remove redundancy of digital characters [1]. The information represented by the digitals corresponds to dates, values, telephone numbers and so on. These digitals are mainly associated with text-genre rather than authorship. But CNG will still extract from text. Hence if all digitals are replaced with a special symbol (e.g., '@'), the redundancy would much lower. For example, extracted |123| for tri-gram could be replaced by |@@@|.

Figure 1 presents the examination of the effect of this text pre-processing procedure on the authorship [2]. Compared with raw text, the accuracy is increased by using text pre-processing method. As well as, the performance of the text pre-processing method is better when features are more than 2000. This indicates that simple text transformations can yield considerable improvement in accuracy.

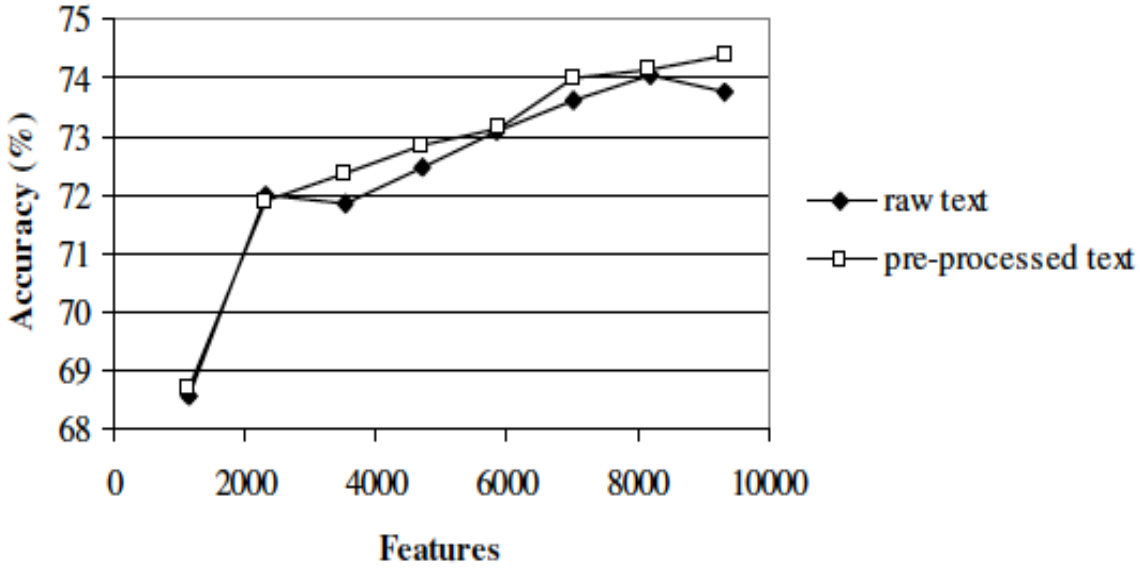


Figure 1 Effect of Text Pre-processing

2.2 Selecting the dominant N-Values

The original algorithm induces variable N. In the extract processing, it is not necessary to evaluate large numbers of N. keeping N is dominant to optimize the result. Therefore, LocalMaxs is introduced. This algorithm would compute local maxima comparing each N-Gram with similar N-Grams. The rules show following:

$$if\ (C.length\ > 3)$$

$$g(C) \geq g\big(ant(C)\big) \cap g(C) > g(succ(C), \forall ant\ (C), succ(c))$$

$$if\ (C.length\ = 3)$$

$$g(C)\ > g\big(succ(C)\big), \forall\ succ(C)$$

- g (c) is the glue of N-Gram C, that is the power holding its characters together.
- ant (C) is an antecedent of an N-Gram C, which is a shorter string having size (n - 1).
- $succ\ (C)$ Is a successor of C, that is, a longer string of size (n + 1), i.e., having one extra character either on the left or right side of C.

According to previous studies, 3-grams, 4-grams and 5-grams could provide the best result in the framework of the authorship identification. It indicates that n varies from 3 to 5. Followed the rules are explained above, 3-grams are only compared with

successor n-grams and 5-grams are only compared with antecedent n-grams. Consequently, the proposed N-Grams are 3-grams and 5-grams against 4-grams.

2.3 Extraction Algorithm Programming and Text Classifier

Implement of CNG algorithm is using JAVA and text classifier is using MatLab. The framework of processing is as following table.

Text Version	Algorithms Description	Implement Environment
V0: Raw Text		
V1.0: Pre-processed Text	Text pre-processing removes the redundancy of digital characters	MatLab
V2.0: First-Extracted Text	Calculate appeared probabilities for only 3-grams and 5-grams using given text	JAVA
V3.0: Second-Extracted Text	Calculate effected probabilities for 3-gram and 5-gram effected by previous two words	JAVA
V4.0: Classified Text	According to the results from V2, to calculate probabilities using dissimilarity calculation method	MatLab

Table 1 Version Description

Based on pre-processed text, appeared probabilities could be implemented for 3-grams and 5-grams about each feature. The conception of algorithm of second-extracted text assumes that the previous two words would affect the first-extracted word. As a result, effected probabilities are calculated by:

$$P_e\left(W_{i\,,}\, W_{i-2}\right) = \frac{N\,(W_i)}{N\,(W_{i-2})}$$

$N\,(W_i)$ Indicates featured 3-grams and 5-grams appeared times.

$N\,(W_{i-2})$ Indicates previous two words appeared times corresponding to above featured 3-grams and 5-grams.

Consequently, the ration of these two terms indicates that features 3-grams and 5-grams appear possibilities, which are effected by previous two words. If the ration is

large, this high-possibility would be considered. In opposite, the low ration condition would not be considered.

The dissimilarity calculation is a simple algorithm to calculate between two-extracted texts [3]. For two identical texts A and B, the occurrences of each feature figure out. Then A and B with different features sets are:

$A = \{(x_1, f_{1A}), (x_2, f_{2A}) \dots (x_n, f_{nA})\}$

$B = \{(x_1, f_{1B}), (x_2, f_{2B}) \dots (x_n, f_{nB})\}$

$f_{nA}\ and\ f_{nB}$ is probability about feature x_n.

- D is dissimilarity set.
- Normalized difference of A and B with feature x_n is

  $d = (\frac{2(f_{nA} - f_{nB})}{f_{nA} + f_{nB}})^2$

- Add the calculated result to D: D = D + d
- Integral dissimilarity $\sum D\ (x_n)$

3 Future Approaches

In this section, some of the main missions for the upcoming semester have been address.

3.1 General Missions

Currently, coding and programming had been done. Outputs with each algorithm will be delivered after testing.

3.2 Algorithm Comparison

The mainly of next stage is comparison results with the different algorithms. The

Total Number of Input Data	Type of Algorithm	Text Classifier	Feature Keywords	Number of Disputed Text	Accuracy

4 Project Progress Control and Monitoring

This section will describe the project management tasks need to be done the general idea of the milestone updates, tasks allocation and Gantt chart updates.

4.1 Milestone Updates

The table lists the finished tasks and tasks for the future.

Important Even	Date	Responsibility
Finalize Algorithms Option (finished)	26th Aug 2011	All Group Members
Design, Coding and Programming (finished)	23th Oct 2011	All Group Members
Testing (in process)	23th Jan 2012	All Group Members
Comparisons	27th Feb 2012	All Group Members
Final Delivery of Project	11th Mar 2012	All Group Members

Table 3 Milestone Updates

According to milestone updates, the project tasks have been finished on time. The following progression will begin in the next milestone.

4.2 Tasks Allocation

The table list is constructed based the work breakdown structure from the early stage, some of allocation has been modified.

Task	Responsibility
Researching and Planning	All Group Members
Project Scoping	All Group Members
Testing Past Algorithms	All Group Members
Developing New Algorithms	All Group Members
Common N-Gram	Yan Xie, Zhaokun Wang
Word Sequence	Kai He
SVM	Yan Xie
Dissimilarity Calculation	Zhaokun Wang
Naïve Bayes	Kai He
Comparing Algorithms	All Group Members
Apply to Controversies	All Group Members

5 Conclusions

From the discussion of the tasks have been done, in the first semester, the project currently is on schedule and on budget. The key technical challenges are identified and have been taken care of accordingly. The current progress of the project is considered to be satisfactory. The detailed plan of the future tasks will ensure the on-time delivery of the project.

Reference

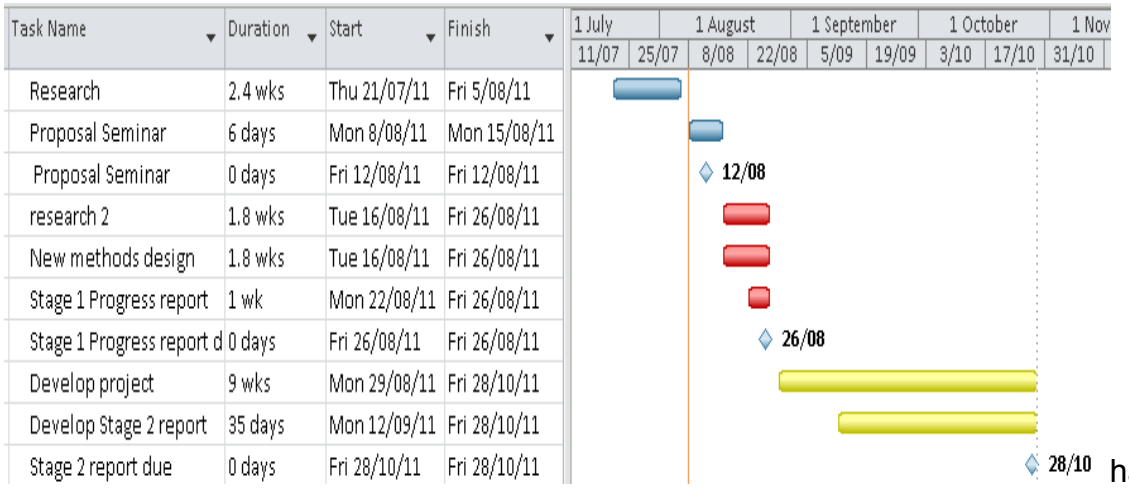
[1] Matron, Y., Wu, N., Hellerstein, “On Compression-Based Text Classification. Advances in Information Retrieval: 27th European Conference on IR Research,” Springer LNCS – 3408, pp. 300-314 (2005).

[2] Jhon H, Efstathios S, “N-Gram Feature Selection for Authorship Identification,” Dept. of Information and Communication System Eng, University of the Aegean (2005).

[3] Vlado Keseij, Fuchun Peng, Nick Cerconry, Calvin Thomas. 2003. “N-Gram-Based Author Profiles for Authorship Attribution.” Dalhousic University and University of Waterloo, Canada.

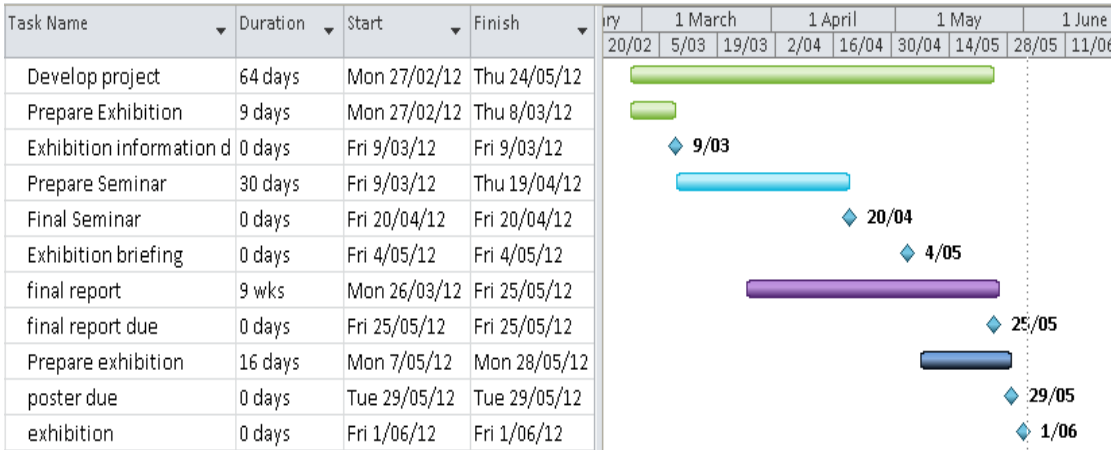
Appendix:

Appendix A: Gantt Chart



Task Name	Duration	Start	Finish
Research	2.4 wks	Thu 21/07/11	Fri 5/08/11
Proposal Seminar	6 days	Mon 8/08/11	Mon 15/08/11
Proposal Seminar	0 days	Fri 12/08/11	Fri 12/08/11
research 2	1.8 wks	Tue 16/08/11	Fri 26/08/11
New methods design	1.8 wks	Tue 16/08/11	Fri 26/08/11
Stage 1 Progress report	1 wk	Mon 22/08/11	Fri 26/08/11
Stage 1 Progress report d	0 days	Fri 26/08/11	Fri 26/08/11
Develop project	9 wks	Mon 29/08/11	Fri 28/10/11
Develop Stage 2 report	35 days	Mon 12/09/11	Fri 28/10/11
Stage 2 report due	0 days	Fri 28/10/11	Fri 28/10/11

hart



Task Name	Duration	Start	Finish
Develop project	64 days	Mon 27/02/12	Thu 24/05/12
Prepare Exhibition	9 days	Mon 27/02/12	Thu 8/03/12
Exhibition information d	0 days	Fri 9/03/12	Fri 9/03/12
Prepare Seminar	30 days	Fri 9/03/12	Thu 19/04/12
Final Seminar	0 days	Fri 20/04/12	Fri 20/04/12
Exhibition briefing	0 days	Fri 4/05/12	Fri 4/05/12
final report	9 wks	Mon 26/03/12	Fri 25/05/12
final report due	0 days	Fri 25/05/12	Fri 25/05/12
Prepare exhibition	16 days	Mon 7/05/12	Mon 28/05/12
poster due	0 days	Tue 29/05/12	Tue 29/05/12
exhibition	0 days	Fri 1/06/12	Fri 1/06/12

Appendix B: Risk Assessment

Risk	Preventive Measures	Probability Rating (/10)	Impact (/10)	Priority (/100)
Behind Schedule	Monitor the project progress regularly	6	8	48
Unclear of given tasks	Divide tasks into simple sub-tasks and discuss thoroughly	5	8	40
Absence of team members	Have backup plans in advance	4	7	28
Data Lost	Backup files regularly	2	9	18