北京大学

本科生毕业论文

论文题目：英文学术论文基本规范检查系统的改进

学　　院：信息科学技术学院

年　　级：2012 级

专　　业：智能科学与技术

姓　　名：孙海洋

学　　号：1200012989

指导教师：林宙辰

2016 年 5 月 17 日

摘要

在学生完成了一篇英文学术论文之后，导师和学生通常都要花费较多时间来检查论文是否符合英文学术论文的基本规范。这些需要重复检查的项目之中，有些具有十分明确的规则，可以由计算机代为检查。为了节省导师和学生人工检查英文学术论文所耗费的时间，前人设计了一个英文学术论文基本规范检查系统。

由于是第一个版本，原系统存在一些错误和不足。本毕业设计修正了原系统的一些错误，提高了原系统某些功能的精度，加入了一些新的功能，并对用户界面做出了一定的改进。改进之后，系统的实用性大大增加。

关键词

格式检测；英语语法；正则表达式；功能改进

The Development of A System for Automatically Checking the Basic Rules of English Paper Writing

Haiyang Sun (Intelligence Science and Technology)

Directed by Prof. Zhouchen Lin

Abstract

Usually, when a student hands on an English research paper, both the mentor and the student have to spend lots of time on checking it to see whether it obeys the basic rules in English paper writing. However, some rules are very clear. So people can check them using a computer. A former schoolmate built a system to inspect the basic rules in writing English paper writing in order to save time.

There are some errors and failures in the previous system and I improved it, including improving the existing functions, adding new functions and improving of the interface. After this, the system has become more practical.

Key words

Format detection; English grammar; regular expression; improvement of functions

目录

第一章 序言

1.1 问题的背景

在大学里的学习、科研工作中，需要学生撰写英文学术论文的时候越来越多。但是，由于很多学生对英文学术论文写作规范的不熟悉，在学生完成了一篇论文之后，导师和学生通常都要花很多时间来检查论文是否符合英文学术论文的基本规范[1]。在这些规范之中，有的规范是英语拼写或者语法的问题：比如 e.g.的正确拼写，冠词的正确搭配等等。有的规范则是口头语言与书面语言混淆的问题：比如在英文学术写作中应该尽量少地使用被动语态，isn’t、doesn’t 应该写作 is not、does not 等等。在导师检查完一遍论文之后，如果学生稍作修改，那么他们之后可能又需要重新检查一次。这些情况使得英文学术论文的写作、修改过程极为费时费力。值得注意的是，在这些需要重复检查的基本规范之中，有些规范具有十分明确的规则，实际上是可以由计算机代为检查的。在这个背景下，为了节省导师和学生人工检查英文学术论文所耗费的时间，前人便开发了一个英文学术论文基本规范检查系统[2]。该检查系统能够针对 Portable Document Format(PDF) 格式的英文论文进行检查。

1.2 原系统的架构

原英文学术论文基本规范检查系统由两个部分组成：文本读取部分和文本检查部分。如图 1-1 所示。

先用一个文本读取部分来将 PDF 格式论文中的文本读取出来。PDF 是一种采用了 PostScript 技术的树形结构文档，文本、图片、表格都可以排版于其中 [3]。原系统作者调研了 Karim Hadjar 团队开发的 Xed 工具 [4] 和 Herve Dejean 团队开发的系统 [5] 等一些可以从 PDF 文档中读取信息的工具。最终在这个部分使用了 Python 语言的 PDF 处理工具 PDFMiner [6]。该处理工具能够读取 PDF 中所有的文本信息。

文本检查部分逐项检查文本读取部分读取出的论文文本，读取方式为按行读取。对读取出的论文文本针对基本规范进行检查。根据调研的计算机英语语法检查领域的相关工作[7-10]，采用了多种方法，如正则表达式、字符串查找及参考单词表，其中最主要使用的方法是正则表达式。

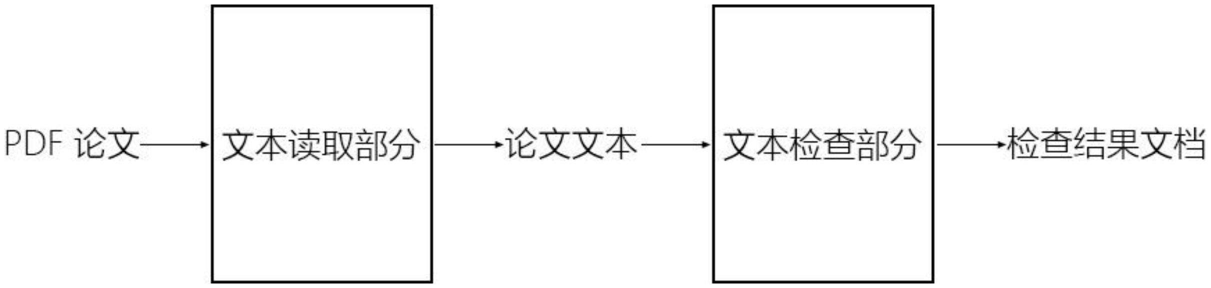


图 1-1 原系统的架构

1.3 原系统实现的规范检查功能

原系统对十余项英文学术论文基本规范进行了检查，主要内容如下[2]：

1. Use passive tense as few as possible.

2. When using an acronym, give its full name when it first appears (except in title).

3. Make clear the usage of “a” and “an”. Use “an” before a word or a math entity whose pronunciation leads with a vowel.

4. If there are only two objects, A and B, write “A and B”. Do not write “A, B”.

5. If there are more than two objects, A, B, …, Y, and Z, normally you can write both “A, B, …, Y, and Z” and “A, B, …, Y and Z”. However, the former is recommended.

6. Do not write “A, B, and C, etc.” Write “A, B, C, etc.” instead.

7. Do not write “isn’t”, “aren’t”, “don’t”, “doesn’t”, etc. Write “is not”, “are not”, “do not”, “does not”, etc., instead.

8. Do not write “can not”. Write “cannot” instead.

9. Notice the correct dots in “e.g.”, “etc.”, “et al.”, etc.

10. Put a comma before “respectively”.

11. Leave a space between texts and left parenthesis, left bracket, or citation. Also leave a space between comma or period and the successive texts.

12. If you refer to multiple figures or tables simultaneously, write “Figures A and B” or “Tables A and B”, rather than “Figure A and Figure B” or “Table A and Table B.”

13. Every reference must be cited in the body text. Naturally achieve this by using a .BIB file.

第二章 对原系统规范检查功能的改进

原系统对十余项英文学术论文基本规范进行了检查，但一些功能精度不够，经常将本是正确的内容报告为错误，现定义这种情况为“误报”，误报情况较多就会浪费学生和导师的时间，与系统本要起到的作用相违背；另外，有一些基于规则的基本规范原系统并未实现。所以，原系统的规范检查功能十分有改进的必要。

2.1 提升原有功能的精度

原系统中有 4 项精度过低，出现了较多的误报现象。针对不同的功能，本系统作出了不同方面的改进，从而提高了这几项的精度。源代码请查看附录一。

2.1.1 原有功能 1

即“Use passive tense as few as possible”。原系统在这一项里检测过去分词来判断被动语态，对于不规则过去分词根据词典[11]进行查找。但是实际上使用了过去分词并不代表使用了被动语态，有可能只是使用了现在完成时。也就是说，原系统对所有返回情况中实际错误情况的准确率作出了妥协，会返回一些并不是被动语态的句子。

针对这种情况，本系统加入了对系动词 be 的检查。因为在使用被动语态时，过去分词前面一定会有系动词 be 出现。而且，be 和过去分词之间只能介入副词。根据原系统转换的英语论文格式，其正文文本一行大约有 7~13 个英文单词。我们可以假定，be 和被动分词的距离不会超过 1 行，同时对系动词 be 和过去分词进行检测。

在已有的检测过去分词功能上，加一个标记，用来表示上一行有没有 be 的几种形式，包括 am, is, are, was, were, been 等，并在本行进行检测。如果存在，系统再进行报错处理。

流程图如图 2-1(A)所示。

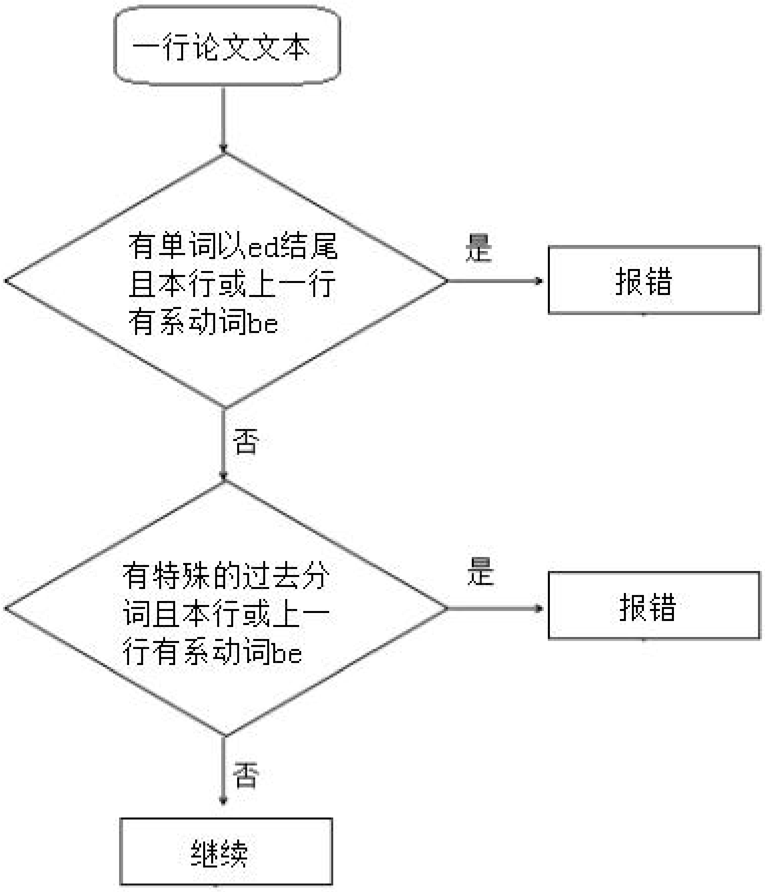

一行论文文本
有单词以ed结尾且本行或上一行有系动词be
是
报错
否
有特殊的过去分词且本行或上一行有系动词be
是
报错
否
继续

图 2-1(A)

2.1.2 原有功能 2

即“When using an acronym, give its full name when it first appears (except in title), e.g.,‘Principal Component Analysis (PCA).’”。

原系统使用正则表达式 ^.*\b[A-Z]{2,}\b.*$ 找到了所有的缩写词并排除了单个字母的情况。这样有很多缩写重复出现，没有必要，增加了学生或是教师的复查难度；另外,参考文献中是允许出现未提前说明的缩写的，如 IEEE。

新版本只检查参考文献之前的内容。由于文本转换功能有一些瑕疵，在读入两列并存的 PDF 时有可能出现错位现象。但是其来自于第三方的库，出现这种情况无法修改，所以检查到“reference”出现后的第一篇参考文献。如果作者确实没写参考文献，仍会检查全部内容。建立一个字典，每检测到一个大写单词, 就到字典中扫描，如果没有，写进该文件；如果有，则跳过该单词。

这样改动之后，系统对每个大写单词只会报错一遍。在这写大写单词中，有一

些是在小标题中出现的单词，只是每个字母都大写，并不是缩写词。所以本系统加入了对“INTRODUCTION”，“REPRESENTATION”，“CONCLUSION”，“REFERENCES”，“I”，“II”，“III”等词的特殊判断。

除此之外，还有一些缩写词是经过提前标注的，主要有两种形式。一是先给出一个词组，在之后用括号表示成缩写词的样子，如“Principal Component Analysis (PCA)”；第二种是先写出缩写，然后在后面括号中表示出是那些单词的缩写，如“CML (Subspace Clustering and Multi-label Learning)”。所以本系统加入了对括号的检查，以便对这两种情况进行判断。其中第一种情况直接判定为书写正确，第二种情况则对括号后面单词的首字母进行检查，如果其与缩写词对应在判定为书写正确。

由于系统是逐行检查，会碰上一行中出现多个缩写词的情况，如“and tag-based image retrieval[3, 4] (TBIR). CBIR takes an”中，“TBIR”本是一个书写正确的缩写词，但由于遇到了“CBIR”，它不会被记录进字典。解决这个问题需要在检查出错误时检查本行有没有可以记录进字典的，如果有进行记录。而本系统正是这样做的。

流程图如图 2-1(B)所示。

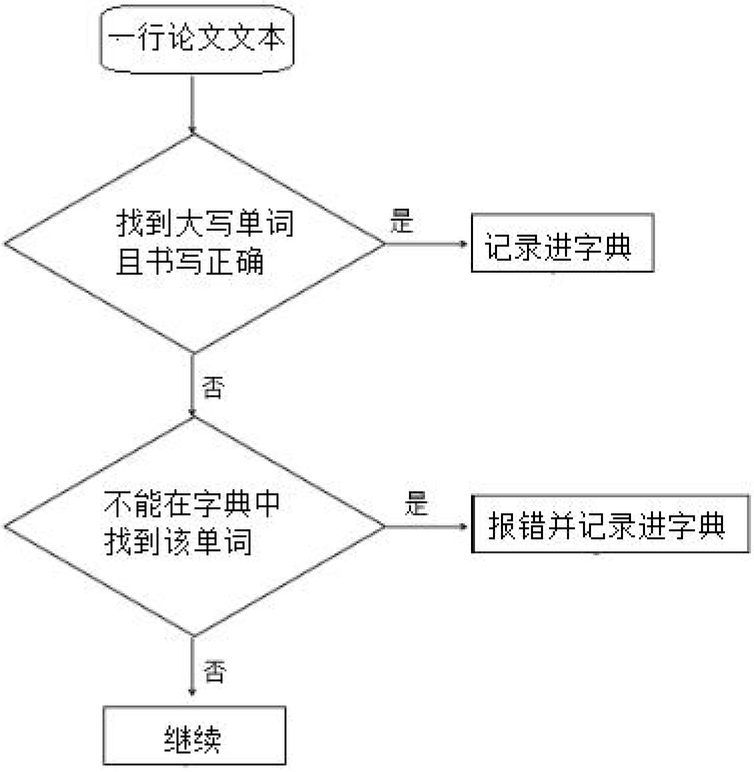


图 2-1(B)

2.1.3 原有功能 4

即“If there are only two objects, A and B, write ‘A and B’. Do not write ‘A, B’.”。

原系统使用正则表达式 ^.*\b\w*\b, \b[^(etc)(respectively)]\w*\b\..*$ 找到“A, B”形式，并排除掉了 etc，respectively 等正常搭配的干扰。更改之后的版本对前面单词是副词或后面单词是“otherwise”，“and”等正常搭配也不会报错。在参考文献中遇到后面是数字也不会报错。为什么要强调在参考文献中呢？因为在正文里是会出现“the number 1,2”这种情况出现的，而在参考文献中，这种情况后面是日期，是合理的。

流程图请见图 2-1(C)。

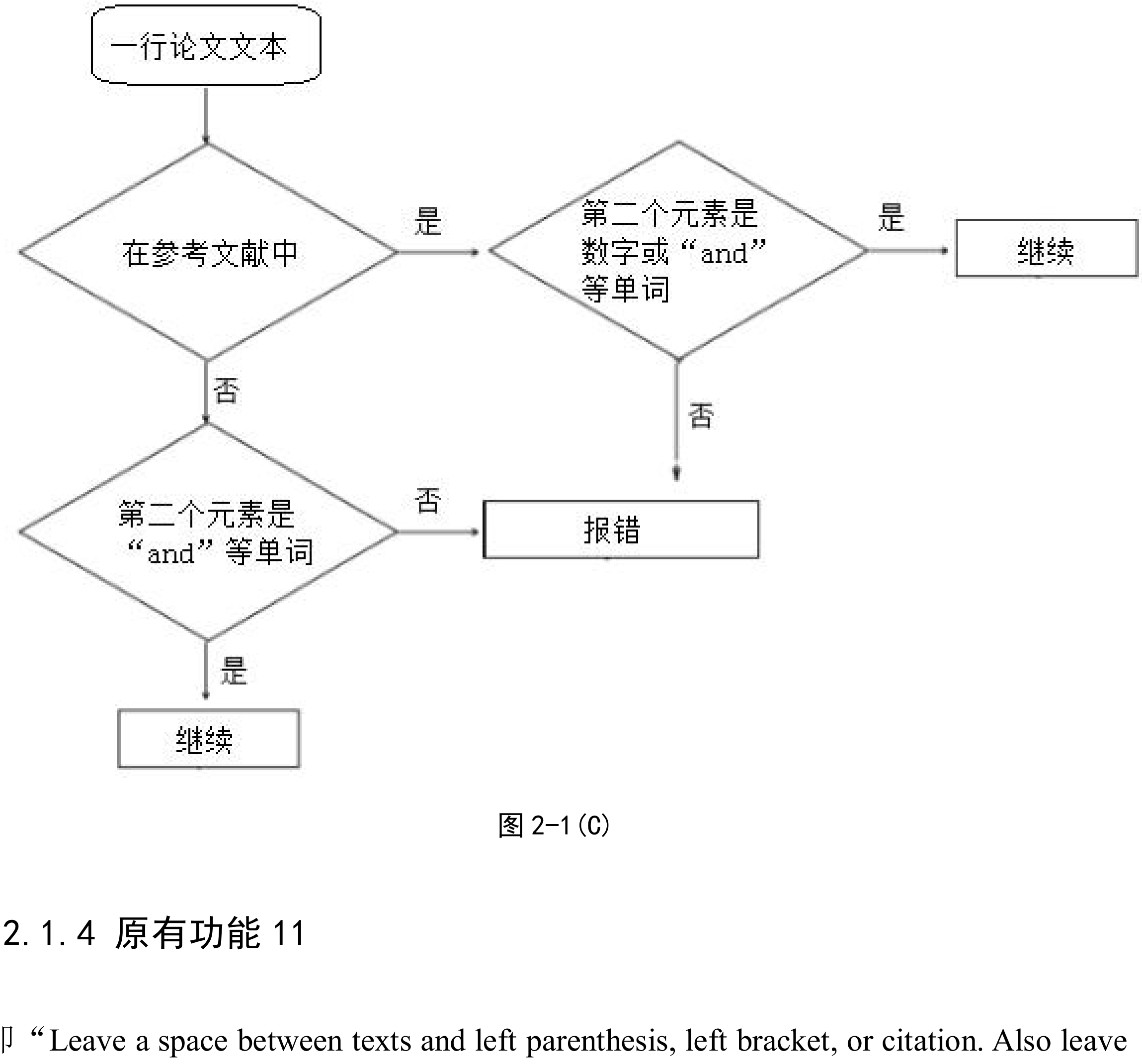


图 2-1(C)

2.1.4 原有功能 11

即“Leave a space between texts and left parenthesis, left bracket, or citation. Also leave a space between comma or period and the successive texts.”。

原系统。使用正则表达式^.*((\S(\(|\[|<))|((,|\.)\w)).*$ 找到所有没按规范留下空格的地方，在寻找句号后未留空格的情况时，会把小数点的情况也判断在内，如 3.14。英文

中不允许有数字作为句子开头的情况，所以只要“.”两边都是数字，即可认为是小数点。但是，还需要考虑的一点是，参考文献里标明页数时也有可能出现“.”两边都是数字的情况，但这是前面会有冒号，利用[^:]去除掉这种情况。参考文献中表示位置的冒号与后面的括号之间是不必有空格的，也要考虑进去。另外，由于读入公式为文本格式，所以进行了特殊处理。

流程图请见图 2-1(D)。

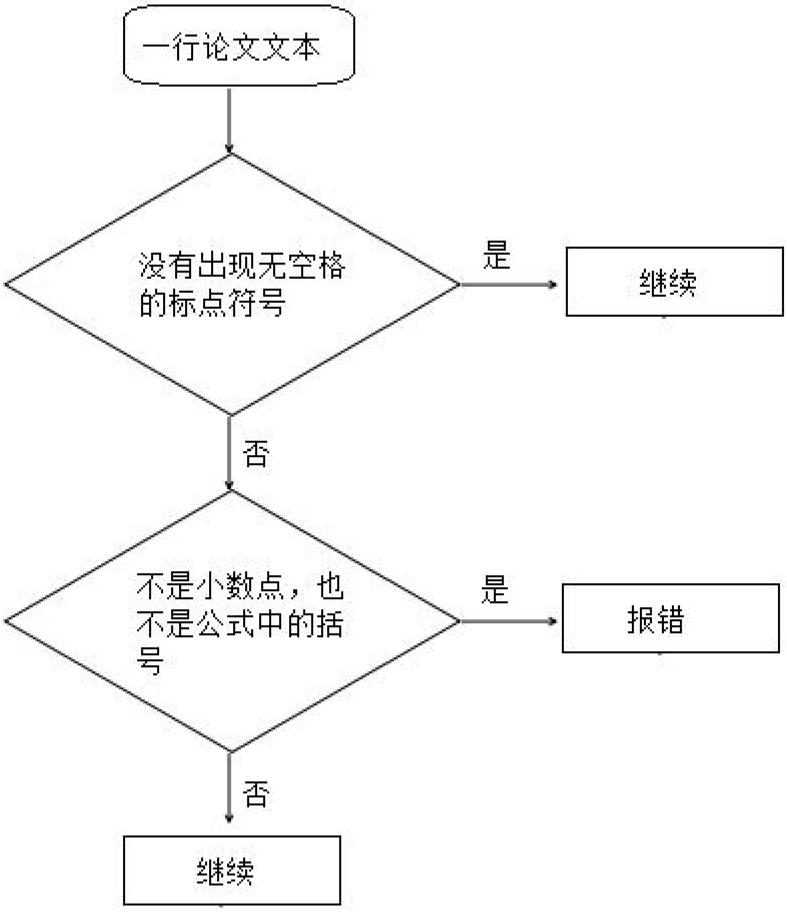


图 2-1(D)

2.2 新增功能

本系统总共在英文文本，数学公式和参考文献三个方面增加了五项新功能。

2.2.1 英文文本方面

在英文文本方面增加了一项新功能，为“Use ‘``’ in the .tex file for the left quotation mark.”。

在使用 Latex 等工具撰写英文学术论文时，转换格式后经常会遇到将右引号误写为左引号的事情发生，如“are ” four” and ” her virginity” respectively. List pricing is”。

本系统利用正则表达式 r'^.* “.*$'和 r'^.*” .*$'找到所有出现右引号却没有出现左引号的行数。

源代码请见附录二代码 1，流程图请见图 2-2(A)。

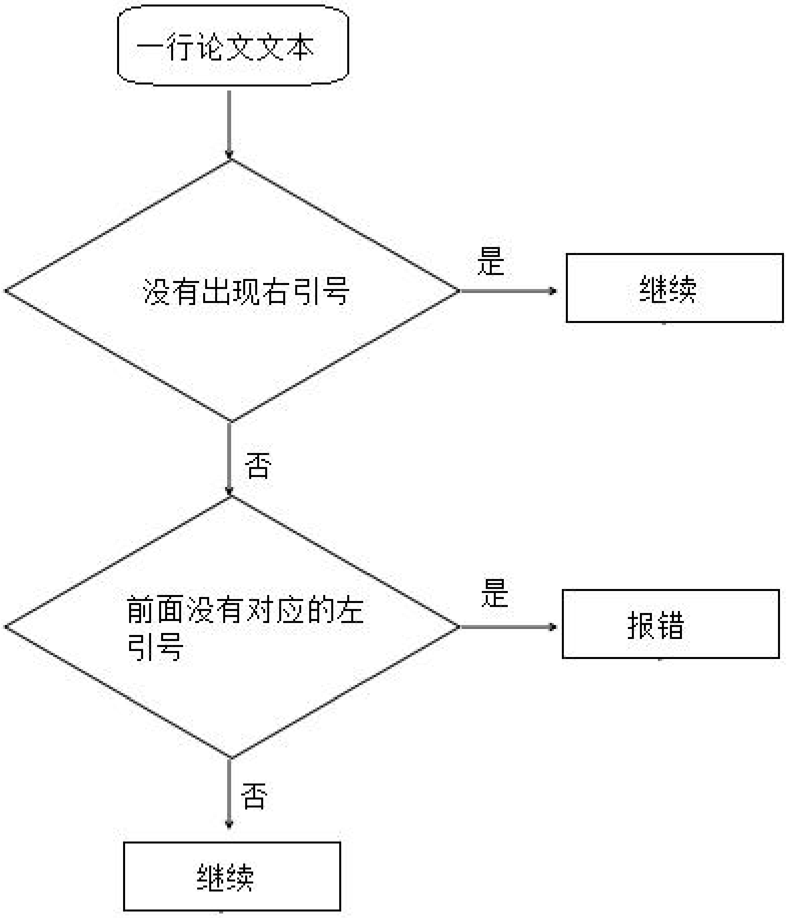


图 2-2(A)

2.2.2 数学公式方面

在数学公式方面增加了两项新的功能，分别是：

1. There must be a punctuation after every standalone math expression.

每个单独的数学公式后面都应该有标点符号，逗号或句号均可。考虑到学术论文中书写公式后面基本都有标号而且在行末。所以可以考虑检测每一个位于行末的标号前面是否有标点符号。但是，其实有些在后文中不会提到的数学表达式是没有标号的。而且，由于在 PDF 中，数学公式的排版是打乱的，打乱顺序根据不同的数学公式也有所不同。所以这种检测方式效果很差，系统最后没有使用这种检查方式。

因而需要找到一种新的检查方式。考虑到一个比较显然的数学表达式的共同点，在数学公式中，一定会有等号或者不等号，包括大于等于号，小于等于号等。这就是一个检测的标识，系统会检测出现这些符号的行末是否有标点符号。

源代码请见附录二代码 2，流程图请见图 2-2(B)。

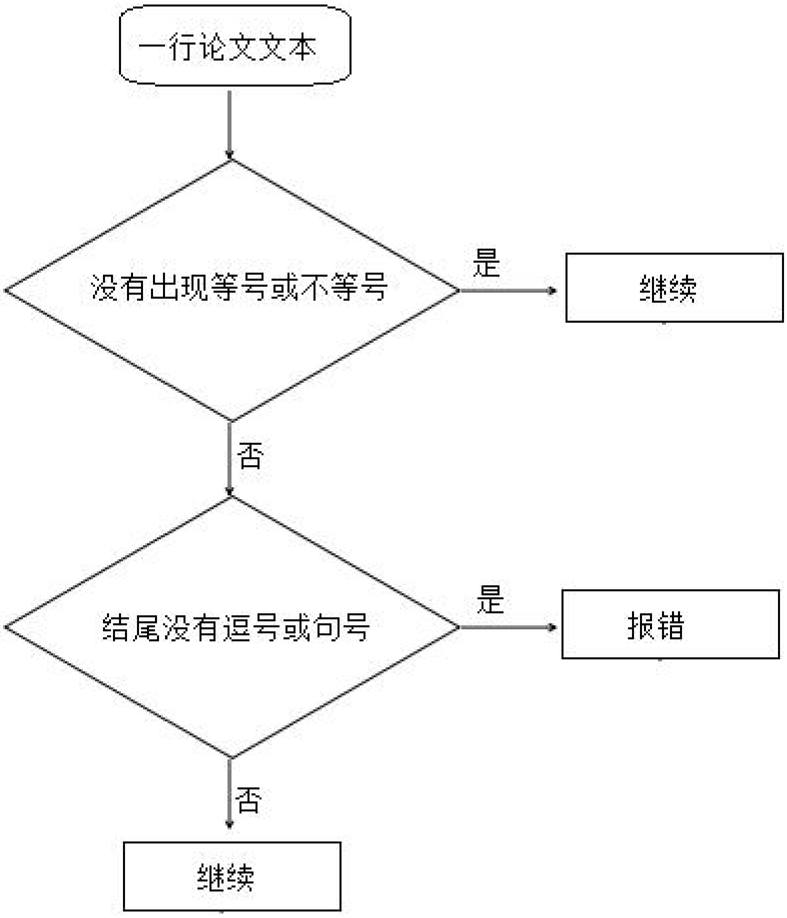


图 2-2(B)

2. For norms, use “\|” instead of “||”

在英文学术论文中，出现“||”的情况绝大多数为数学公式，所以利用正则表达式

r'^.*\|\|.*$'，只要出现“||”就会报错。

源代码请见附录二代码 3，流程图请见图 2-2(C)。

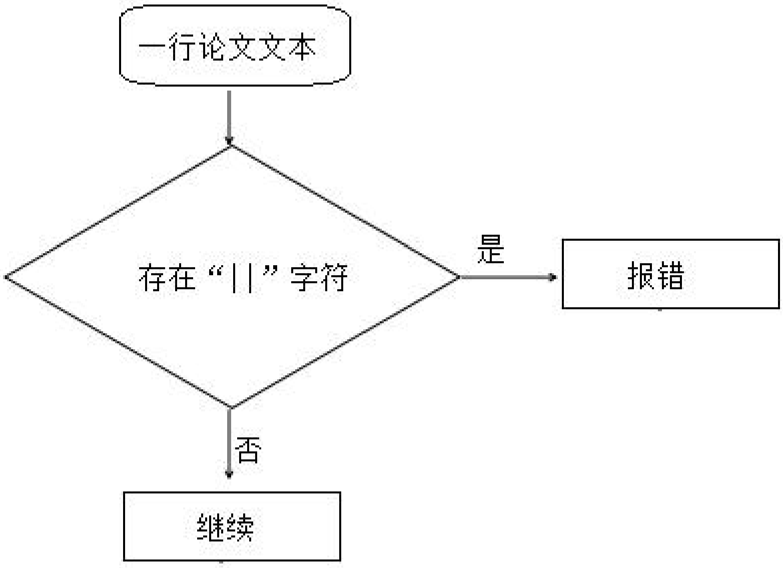


图 2-2(C)

2.2.3 参考文献方面

在参考文献方面增加了两项新的功能，分别是：

1. Every reference must be cited in the body text.

对于参考文献是以数字表示，并且像“[1][3]”这样单独出现的情况，原系统已经实现了该功能。但是出现了对文献的计数错误，其原因在于原版本通过对“referrence”后出现的所有“[”进行计数。但是对于分成两边的 PDF 格式，第三方库有可能出现错位的现象。现有的版本则对于第一篇文献之后的所有位于行首的“[”，从而得到了改进。

但是，参考文献的实际情况要复杂得多。就以纯数字表示的参考文献来说，在文中有可能一段引用了多篇参考文献，可能会以[1-5]或[1,3,5]这样的形式表现出来，而不会麻烦地表示成[1][2][3][4][5]或[1][3][5]的形式。由于系统是逐行检测，检查这种引用多篇参考文献的还需要考虑换行的问题。考虑到行首不回忆标点符号开始，所以可以利用标记找到如“5]”对应的前半部分。这类表示不会跨行，所以不会出现分辨不出是[1-5]或[1,3,5]哪种形式。

另外，有的参考文献不是以数字表示的，而是以[会议，页数，日期]表示的，对于这种类型的表示方法，本系统有兼容性，即不会出错，并且同样可以进行检测。

源代码请见附录二代码 4，流程图请见图 2-2(D)。

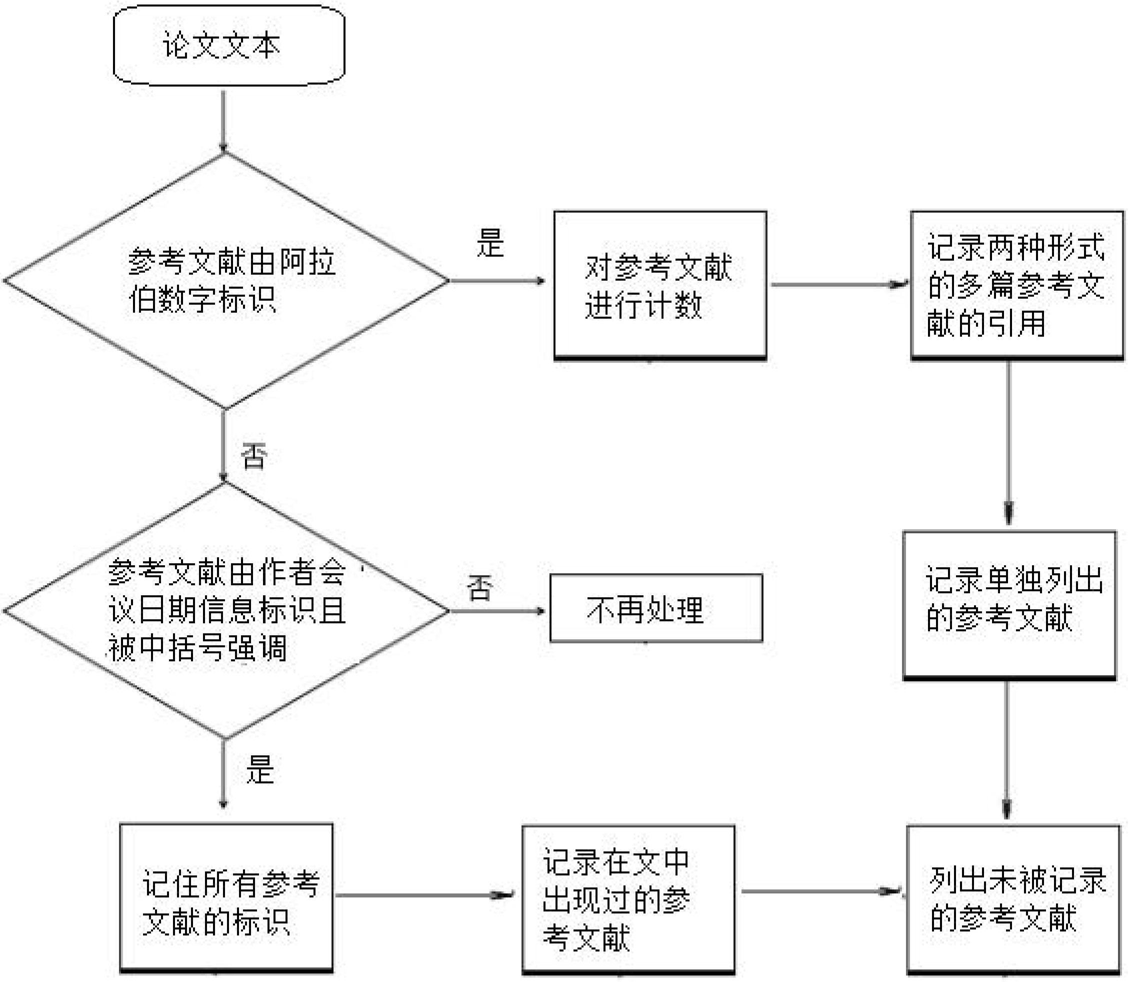


图 2-2(D)

2. The reference information should be complete.

参考文献的书写格式不是固定的，对网址，期刊，图书的书写要求也各有不同。但有些东西是固定的，比如说对期刊和图书，一定有年份和页码，成分个数不会低于四个。基于这些特点进行检测。对于网址则直接认为正确，跳过处理，网址的正确与否不在检查之列。

另外，对于 LaTeX 初级使用者来说，在标注作者姓名时有可能忘记书写逗号，出现省略的情况，遇到这种情形系统会直接报错。

源代码请见附录二代码 5，流程图请见图 2-2(E)。

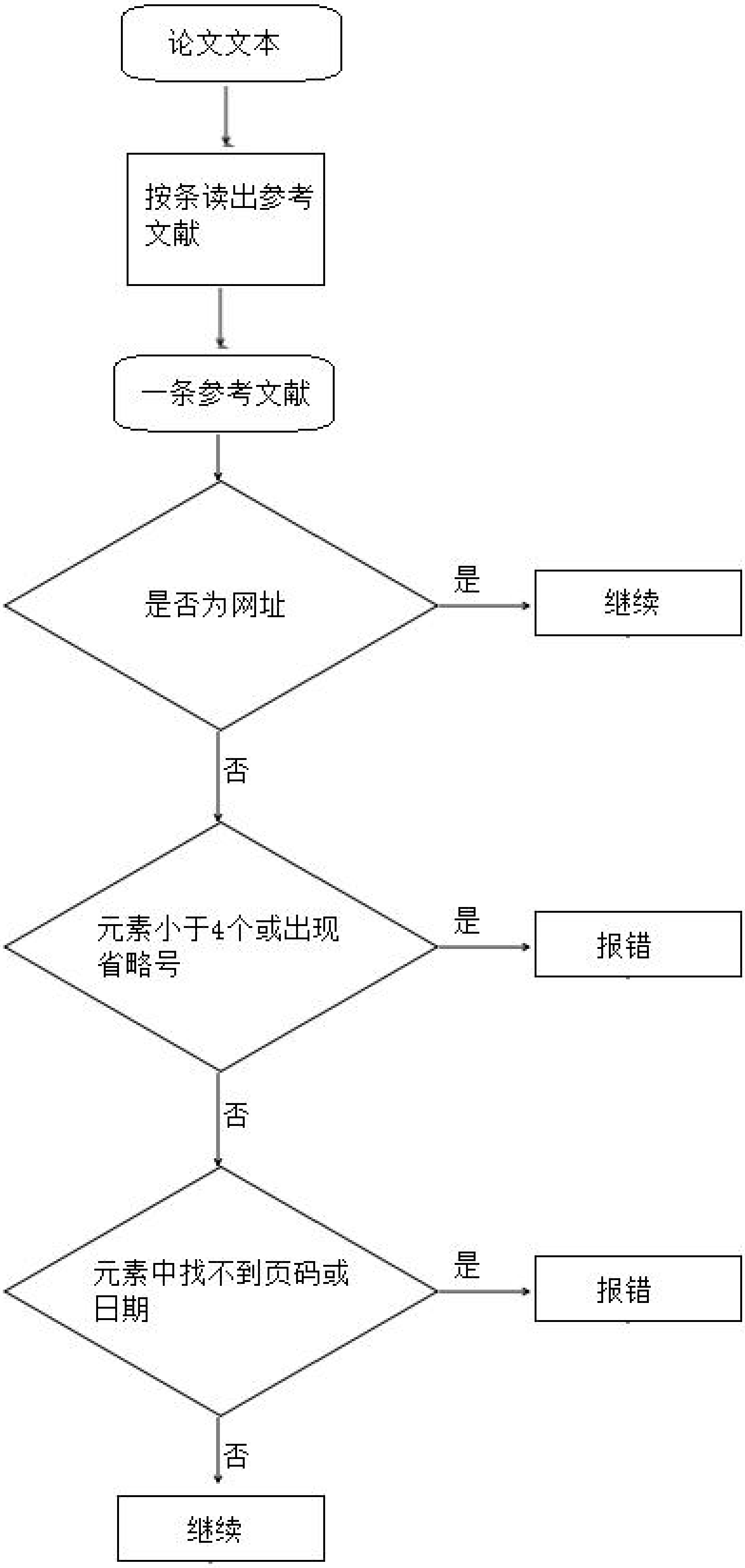

论文文本
按条读出参考文献
一条参考文献
是否为网址
是
继续
否
元素小于4个或出现省略号
是
报错
否
元素中找不到页码或日期
是
报错
否
继续

图 2-2(E)

第三章 用户界面的改进

原系统的输出格式为文本文档，查阅起来很不方便，各种类型的错误也不够醒目。作为检查系统，交互界面可以朴素，但必须要清晰。原系统的输出显得有些杂乱，没有起到一目了然的作用。所以，用户界面也很有改进的必要。

3.1 改进后的系统架构

更改后将结果文档输出为HTML格式，可以将错误类型更好地分门别类。在错误门类后面标注出错误个数，方便使用者查看。每个具体的错误都会标注出位置。清楚具体在哪一行，可以跳回原文进行查看。这样整个系统的架构就如图3-1所示。

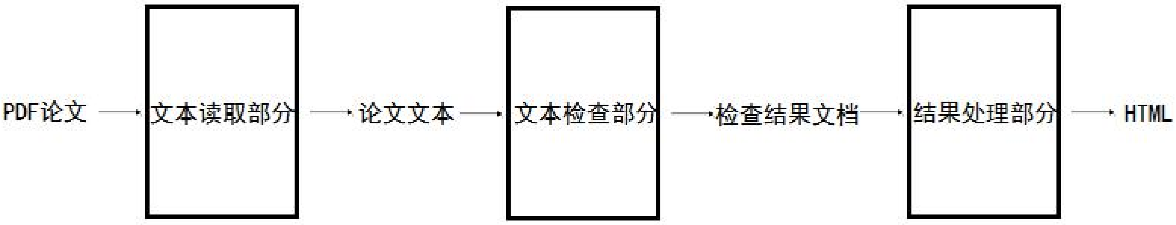
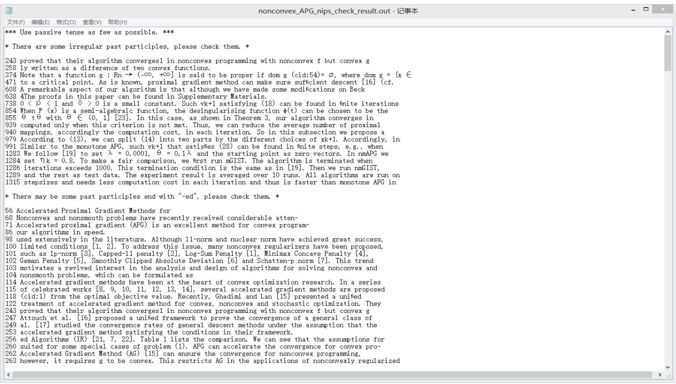


图 3-1

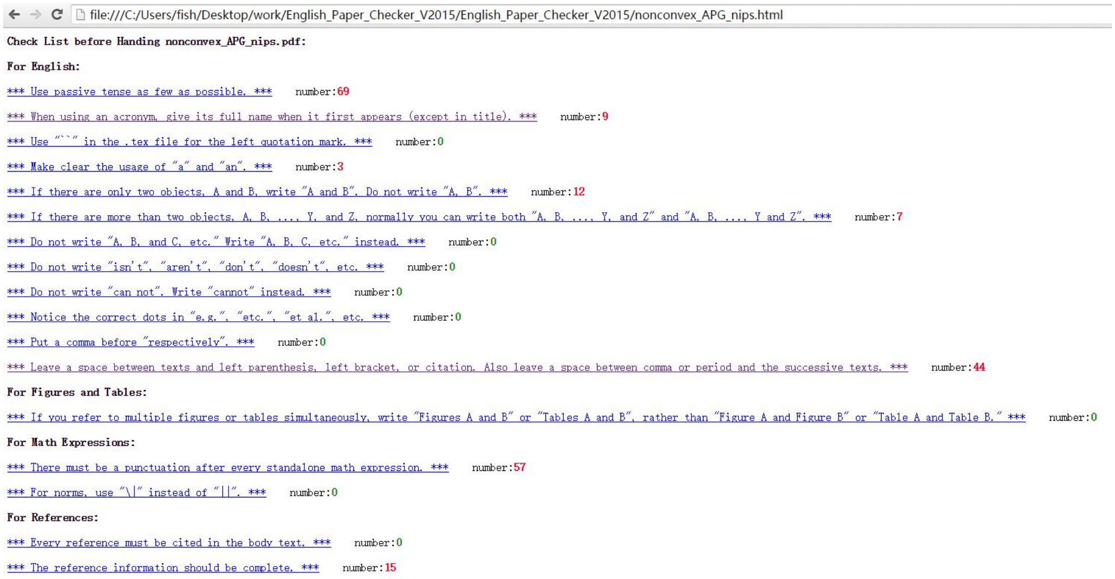
3.2 界面比对

原界面如图3-2(A)所示，为文本文档格式；新界面如图3-2(B)，图3-2(C)和图3-2(D)所示，为HTML格式。其中图3-2(B)是主界面，显示了错误种类以及个数，错误个数为0的，用绿色表示，有错误的则用红色表示；图3-2(C)是具体的某种错误中的显示，可以通过点击左侧的行数查看处理后的论文文本；图3-2(D)是处理后的论文文本的HTML格式。



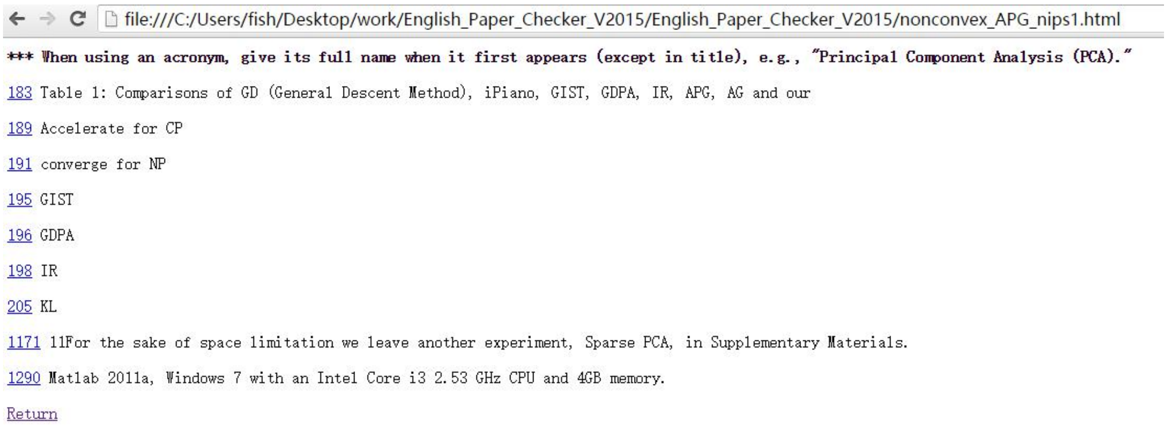
```
*** Use passive tense as few as possible. ***

* There are some irregular past participles, please check them. *

243 proved that their algorithm converges1 in nonconvex programming with nonconvex f but convex g
258 1y written as a difference of two convex functions.
374 Note that a function g : Rn → (-∞, +∞] is said to be proper if dom g (cid:54)= ∅, where dom g = {x ∈
471 to a critical point. As is known, proximal gradient method can make sure sufficient descent [16] (cf.
608 A remarkable aspect of our algorithm is that although we have made some modifications on Beck
638 4The proofs in this paper can be found in Supplementary Materials.
738 0 < ρ < 1 and δ > 0 is a small constant. Such vk+1 satisfying (18) can be found in finite iterations
854 When F (x) is a semi-algebraic function, the desingularising function ϕ(t) can be chosen to be the
855 θ tθ with θ ∈ (0, 1] [23]. In this case, as shown in Theorem 3, our algorithm converges in
939 computed only when this criterion is not met. Thus, we can reduce the average number of proximal
940 mappings, accordingly the computation cost, in each iteration. So in this subsection we propose a
979 According to (13), we can split (14) into two parts by the different choices of xk+1. Accordingly, in
991 Similar to the monotone APG, such vk+1 that satisfies (28) can be found in finite steps, e.g., when
1283 We follow [19] to set λ = 0.0001, θ = 0.1λ and the starting point as zero vectors. In nmAPG we
1284 set ηk = 0.8. To make a fair comparison, we first run mGIST. The algorithm is terminated when
1286 iterations exceeds 1000. This termination condition is the same as in [19]. Then we run nmGIST,
1289 and the rest as test data. The experiment result is averaged over 10 runs. All algorithms are run on
1315 stepsizes and needs less computation cost in each iteration and thus is faster than monotone APG in

* There may be some past participles end with "-ed", please check them. *

56 Accelerated Proximal Gradient Methods for
68 Nonconvex and nonsmooth problems have recently received considerable atten-
71 Accelerated proximal gradient (APG) is an excellent method for convex program-
86 our algorithms in speed.
98 used extensively in the literature. Although l1-norm and nuclear norm have achieved great success,
100 limited conditions [1, 2]. To address this issue, many nonconvex regularizers have been proposed,
101 such as lp-norm [3], Capped-l1 penalty [2], Log-Sum Penalty [1], Minimax Concave Penalty [4],
102 Geman Penalty [5], Smoothly Clipped Absolute Deviation [6] and Schatten-p norm [7]. This trend
103 motivates a revived interest in the analysis and design of algorithms for solving nonconvex and
104 nonsmooth problems, which can be formulated as
114 Accelerated gradient methods have been at the heart of convex optimization research. In a series
115 of celebrated works [8, 9, 10, 11, 12, 13, 14], several accelerated gradient methods are proposed
118 (cid:1) from the optimal objective value. Recently, Ghadimi and Lan [15] presented a unified
122 treatment of accelerated gradient method for convex, nonconvex and stochastic optimization. They
243 proved that their algorithm converges1 in nonconvex programming with nonconvex f but convex g
247 Attouch et al. [16] proposed a unified framework to prove the convergence of a general class of
249 al. [17] studied the convergence rates of general descent methods under the assumption that the
253 accelerated gradient method satisfying the conditions in their framework.
256 ed Algorithms (IR) [21, 7, 22]. Table 1 lists the comparison. We can see that the assumptions for
260 suited for some special cases of problem (1). APG can accelerate the convergence for convex pro-
262 Accelerated Gradient Method (AG) [15] can ensure the convergence for nonconvex programming,
263 however, it requires g to be convex. This restricts AG in the applications of nonconvexly regularized
```

图 3-2(A)



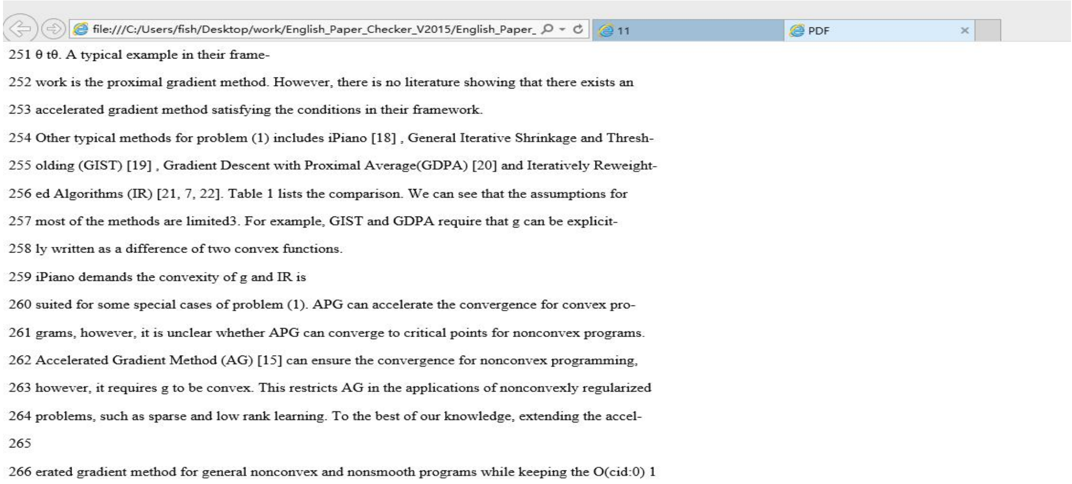
```
file:///C:/Users/fish/Desktop/work/English_Paper_Checker_V2015/English_Paper_Checker_V2015/nonconvex_APG_nips.html

Check List before Handing nonconvex_APG_nips.pdf:

For English:

*** Use passive tense as few as possible. ***    number:69

*** When using an acronym, give its full name when it first appears (except in title). ***    number:9

*** Use "``" in the .tex file for the left quotation mark. ***    number:0

*** Make clear the usage of "a" and "an". ***    number:3

*** If there are only two objects, A and B, write "A and B". Do not write "A, B". ***    number:12

*** If there are more than two objects, A, B, ..., Y, and Z, normally you can write both "A, B, ..., Y, and Z" and "A, B, ..., Y and Z". ***    number:7

*** Do not write "A, B, and C, etc." Write "A, B, C, etc." instead. ***    number:0

*** Do not write "isn't", "aren't", "don't", "doesn't", etc. ***    number:0

*** Do not write "can not". Write "cannot" instead. ***    number:0

*** Notice the correct dots in "e.g.", "etc.", "et al.", etc. ***    number:0

*** Put a comma before "respectively". ***    number:0

*** Leave a space between texts and left parenthesis, left bracket, or citation. Also leave a space between comma or period and the successive texts. ***    number:44

For Figures and Tables:

*** If you refer to multiple figures or tables simultaneously, write "Figures A and B" or "Tables A and B", rather than "Figure A and Figure B" or "Table A and Table B." ***    number:0

For Math Expressions:

*** There must be a punctuation after every standalone math expression. ***    number:57

*** For norms, use "\|" instead of "||". ***    number:0

For References:

*** Every reference must be cited in the body text. ***    number:0

*** The reference information should be complete. ***    number:15
```

图 3-2(B)



```
file:///C:/Users/fish/Desktop/work/English_Paper_Checker_V2015/English_Paper_Checker_V2015/nonconvex_APG_nips1.html

*** When using an acronym, give its full name when it first appears (except in title), e.g., "Principal Component Analysis (PCA)."

183 Table 1: Comparisons of GD (General Descent Method), iPiano, GIST, GDPA, IR, APG, AG and our

189 Accelerate for CP

191 converge for NP

195 GIST

196 GDPA

198 IR

205 KL

1171 11For the sake of space limitation we leave another experiment, Sparse PCA, in Supplementary Materials.

1290 Matlab 2011a, Windows 7 with an Intel Core i3 2.53 GHz CPU and 4GB memory.

Return
```

图 3-2(C)



```
file:///C:/Users/fish/Desktop/work/English_Paper_Checker_V2015/English_Paper_

251 θ tθ. A typical example in their frame-
252 work is the proximal gradient method. However, there is no literature showing that there exists an
253 accelerated gradient method satisfying the conditions in their framework.
254 Other typical methods for problem (1) includes iPiano [18] , General Iterative Shrinkage and Thresh-
255 olding (GIST) [19] , Gradient Descent with Proximal Average(GDPA) [20] and Iteratively Reweight-
256 ed Algorithms (IR) [21, 7, 22]. Table 1 lists the comparison. We can see that the assumptions for
257 most of the methods are limited3. For example, GIST and GDPA require that g can be explicit-
258 ly written as a difference of two convex functions.
259 iPiano demands the convexity of g and IR is
260 suited for some special cases of problem (1). APG can accelerate the convergence for convex pro-
261 grams, however, it is unclear whether APG can converge to critical points for nonconvex programs.
262 Accelerated Gradient Method (AG) [15] can ensure the convergence for nonconvex programming,
263 however, it requires g to be convex. This restricts AG in the applications of nonconvexly regularized
264 problems, such as sparse and low rank learning. To the best of our knowledge, extending the accel-
265
266 erated gradient method for general nonconvex and nonsmooth programs while keeping the O(cid:0) 1
```

图 3-2(D)

第四章 检查结果

本章中我们使用 2 篇刻意构造的英文学术论文和 8 篇真实的英文学术论文，对改进后的英文学术论文基本规范检查系统进行实验。两篇刻意构造的分别是 SPL.pdf(2 页)和 test.pdf(2 页)。在几篇真实的英文学术论文中有 6 篇是尚未发表的，确实需要检查的，分别是 RMF-MM3.pdf(13 页)，paper.pdf(12 页)，CFA.pdf(14 页)，LADM.pdf(9 页)，LADM-supp.pdf(25 页)和 P2.pdf(4 页)；另外 2 篇是人工检查过，已经对外公布的分别是 APG_nips.pdf(9 页)和 0503.pdf(5 页)。这 10 篇论文均由林宙辰老师提供。

4.1 原有功能的改进实例

在 SPL.pdf 和 APG_nips.pdf 中选取几个具体的实例对精度的提升进行验证。在检查系统的输出结果中，错误句子开头的数字是它在 PDF 文件中的行号。因为文章不全存在需要改良的问题，所以以几篇典型文章的典型问题为例。

4.1.1 原有功能 1

即“Use passive tense as few as possible”。示例文章为 SPL.pdf。

原实验结果：

```
*** Use passive tense as few as possible. ***
* There are some irregular past participles, please check them. *


12 gland tissue samples from 46 dolphins that were found dead
87 cannot written as two words, both forms are acceptable usage.
89 up can not in the online unabridged, you will be sent to a list
92 not The historical illustrations given for the negative in the
143 set of LatLRR with denoised data.
148 years away respectively. The two meals cost us 50 and 80
161 But that logic has not won favour with the NSF,which in an
162 unusual 7 May statement said the House measure contradicts
175 shot! Some small(amount of it is likely entirely useless, but a
```

* There may be some past participles end with "-ed", please check them. *

11 In the latest study, researchers analysed lung and adrenal-

13 in Louisiana, Mississippi and Alabama areas that experienced

14 significantly elevated levels of petroleum compounds.

74 wasn’t just handed to him.

78 Well, just because the general population isn’t as fascinated

89 up can not in the online unabridged, you will be sent to a list

90 of suggestions headed by cannot. According to the entry in

95 And ?ou ?at he deed fore cannot sorus be. 1451 Paston Lett.

143 set of LatLRR with denoised data.

149 respectively. He earned an M.A. and a Ph.D. from Chicago

151 are aged 10 and 13 respectively. Her two daughters, Jo and

157 tively shall hereinafter be referred to as Building A and

164 The legislation has also displeased groups such as the Ameri-

169 lodged letters of protest with Smith.

170 That person would never make it to personhood,(or indeed

182 Some of the non-coding DNA is translated¡ into RNA that is

183 not directly translated into the production of proteins, but may

232 low rank and sparse graph for semi-supervised learning,” in CVPR, 2012,

239 [12] P. Favaro, R. Vidal, and A. Ravichandran, “A closed form solution for

243 [13] R. Liu, Z. Lin, F. D. L. Torre, and Z. Su, “Fixed-rank representation for

245 unsupervised visual learning,” in CVPR, 2012, pp. 598–605.

247 [14] G. Bull and J. Gao, “Transposed low rank representation for image clas-

267 [19] J. Shi and J. Malik, “Normalized cuts and image segmentation,” IEEE

现有结果：

*** Use passive tense as few as possible. ***

There are some irregular past participles, please check them:

12 gland tissue samples from 46 dolphins that were found dead

87 cannot written as two words, both forms are acceptable usage.

175 shot! Some small(amount of it is likely entirely useless, but a

There may be some past participles end with "-ed", please check them:

13 in Louisiana, Mississippi and Alabama areas that experienced

14 significantly elevated levels of petroleum compounds.

74 wasn’t just handed to him.

78 Well, just because the general population isn’t as fascinated

143 set of LatLRR with denoised data.

151 are aged 10 and 13 respectively. Her two daughters, Jo and

169 lodged letters of protest with Smith.

182 Some of the non-coding DNA is translated¡ into RNA that is

183 not directly translated into the production of proteins, but may

通过以上结果可以看出对没有系动词 be 出现的过去分词进行处理之后，确实起到了提高精度的作用。

4.1.2 原有功能 2

即“When using an acronym, give its full name when it first appears (except in title)”。实例文档为 SPL.pdf。

原实验结果：

*** When using an acronym, give its full name when it first appears (except in title), e.g., "Principal Component Analysis (PCA)." ***

5 Hongyang Zhang, Zhouchen Lin, Senior Member, IEEE, Chao Zhang, Member, IEEE, Junbin Gao

9 I. INTRODUCTION

16 ABC

18 BCD

19 CDE

83 II. ROBUST LATENT LOW RANK REPRESENTATION

91 the OED, cannot is the ordinary modern way of writing can

93 OED shows cannot, can not, and even canot, as well as the

99 II. 256 The House cant agree to this. 1741 RICHARDSON

100 Pamela I. 56 If he, as you say cant help it. 1742 YOUNG

102 Legions of angels cant confine me there. 1827 KEBLE Chr.

136 III. CONCLUSION

141 LatLRR, which ?rst denoises the data by robust PCA and then

145 REFERENCES

161 But that logic has not won favour with the NSF,which in an

163 the goal of increasing US economic competitiveness.

167 Consortium of Social Science Associations (COSSA).They are

172 The idea that we only ” use” [4Most of our DNA is¡ non-coding

173 DNA¿－that is, DNA that does not code for the production of

176 great deal of it is anything but DNA does more than just code

182 Some of the non-coding DNA is translated¡ into RNA that is

185 RNA or transfer RNA.

193 [1] R. Vidal, “Subspace clustering,” IEEE Signal Process. Mag., vol. 28,

198 structures by low-rank representation,” IEEE Trans. Patt. Anal. Mach.

203 pursuit for image segmentation,” in ICCV, 2011, pp. 2439－2446.

207 representation,” in ICML, vol. 3, 2010, pp. 663－670.

213 segmentation and feature extraction,” in ICCV, 2011, pp. 1615－1622.

219 [7] E. Elhamifar and R. Vidal, “Sparse subspace clustering,” in CVPR, 2009,

224 sparsity pursuit,” IEEE Trans. Image Process., vol. 21, no. 3, pp. 1327－

228 representation,” in IEEE Int’l Conf. on Acoustics, Speech, and Signal

232 low rank and sparse graph for semi-supervised learning,” in CVPR, 2012,

237 analysis?” J. ACM, vol. 58, no. 3, pp. 1－37, 2011.

240 robust subspace estimation and clustering,” in CVPR, 2011, pp. 1801 –

245 unsupervised visual learning,” in CVPR, 2012, pp. 598 – 605.

252 using nuclear norm as a convex surrogate of rank,” in ECML PKDD,

257 for robust visual tracking,” in ECCV, 2012, pp. 470 – 484.

260 subspace segmentation with semidefinite guarantees,” in IEEE Int’l Conf.

264 representation for computer vision and pattern recognition,” Proc. IEEE,

267 [19] J. Shi and J. Malik, “Normalized cuts and image segmentation,” IEEE

273 lighting and pose,” IEEE Trans. Patt. Anal. Mach. Intell., vol. 23, no. 6,

277 Fisherfaces: Recognition using class specific linear projection,” IEEE

现有结果：

*** When using an acronym, give its full name when it first appears (except in title), e.g., "Principal Component Analysis (PCA)."

5 Hongyang Zhang, Zhouchen Lin, Senior Member, IEEE, Chao Zhang, Member, IEEE, Junbin Gao

16 ABC

18 BCD

19 CDE

91 the OED, cannot is the ordinary modern way of writing can

99 II. 256 The House cant agree to this. 1741 RICHARDSON

100 Pamela I. 56 If he, as you say cant help it. 1742 YOUNG

102 Legions of angels cant confine me there. 1827 KEBLE Chr.

141 LatLRR, which ?rst denoises the data by robust PCA and then

161 But that logic has not won favour with the NSF,which in an

163 the goal of increasing US economic competitiveness.

172 The idea that we only ” use” [4Most of our DNA is¡ non-coding

182 Some of the non-coding DNA is translated¡ into RNA that is

从以上结果可以看出去掉了重复报错，可以识别出“(COSSA)”这样的正确格式以

及部分小标题中的大写单词，并且不再检查参考文献中的缩写词。

4.1.3 原有功能 4

即“If there are only two objects, A and B, write ‘A and B’. Do not write ‘A, B’.”。实例文档为 APG_nips.pdf。

原实验结果：

```
42 I like A, B.
44 I like apple, banana.
45 I like Amy, Bob.
46 I like Google, Apple.
195 no. 2, pp. 52–68, 2011.
197 [2] G. Liu, Z. Lin, S. Yan, J. Sun, and Y. Ma, “Robust recovery of subspace
199 Intell., vol. 35, no. 1, pp. 171–184, 2013.
201 [3] B. Cheng, G. Liu, Z. Huang, and S. Yan, “Multi-task low-rank affinities
205 [4] G. Liu, Z. Lin, and Y. Yu, “Robust subspace segmentation by low-rank
217 tion for subspace segmentation,” arXiv:1107.1561, 2010.
223 [8] C. Lang, G. Liu, J. Yu, and S. Yan, “Saliency detection by multi-task
224 sparsity pursuit,” IEEE Trans. Image Process., vol. 21, no. 3, pp. 1327–
225 1338, 2012.
231 [10] L. Zhuang, H. Gao, Z. Lin, Y. Ma, X. Zhang, and N. Yu, “Non-negative
235 [11] E. Cand`es, X. Li, Y. Ma, and J. Wright, “Robust principal component
237 analysis?” J. ACM, vol. 58, no. 3, pp. 1–37, 2011.
239 [12] P. Favaro, R. Vidal, and A. Ravichandran, “A closed form solution for
243 [13] R. Liu, Z. Lin, F. D. L. Torre, and Z. Su, “Fixed-rank representation for
251 [15] H. Zhang, Z. Lin, and C. Zhang, “A counterexample for the validity of
255 [16] T. Zhang, B. Ghanem, S. Liu, and N. Ahuja, “Low-rank sparse learning
259 [17] Y. Ni, J. Sun, X. Yuan, S. Yan, and L. Cheong, “Robust low-rank
263 [18] J. Wright, Y. Ma, J. Mairal, G. Sapiro, and T. S. Huang, “Sparse
265 vol. 98, no. 6, pp. 1031–1044, 2010.
```

```
269 Trans. Patt. Anal. Mach. Intell., vol. 22, no. 8, pp. 888 – 905, 2000.
271 [20] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman, “From few
273 lighting and pose,” IEEE Trans. Patt. Anal. Mach. Intell., vol. 23, no. 6,
274 pp. 643 – 660, 2001.
276 [21] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, “Eigenfaces vs.
278 Trans. Patt. Anal. Mach. Intell., vol. 19, no. 7, pp. 711 – 720, 1997.
```

现有结果：

```
*** If there are only two objects, A and B, write "A and B". Do not write "A, B". ***
42 I like A, B.
44 I like apple, banana.
45 I like Amy, Bob.
46 I like Google, Apple.
235 [11] E. Cand`es, X. Li, Y. Ma, and J. Wright, “Robust principal component
```

从以上结果可以看出，经过新的处理后确实提高了一部分精度，但对于“A,B,and C”这种情况某些仍然无法识别。

4.1.4 原有功能 11

即“Leave a space between texts and left parenthesis, left bracket, or citation. Also leave a space between comma or period and the successive texts.”。

实例文档为 APG_nips.pdf。

原实验结果：

```
** Leave a space between texts and left parenthesis, left bracket, or citation. Also leave a space between comma or period and the successive texts. ***
60 Anonymous Author(s)
81 point, and the convergence rates remain O(cid:0) 1
109 F (x) = f (x) + g(x),
120 with O(cid:0) 1
239 and accelerates with an O(cid:0) 1
```

255 olding (GIST) [19] , Gradient Descent with Proximal Average(GDPA) [20] and Iteratively Reweight-

266 erated gradient method for general nonconvex and nonsmooth programs while keeping the O(cid:0) 1

360 2. For our APGs, the convergence rates maintain O(cid:0) 1

373 2.1 Basic Assumptions

375 R : g(x) < +∞}. g is lower semicontinuous at point x0 if lim inf x→x0 g(x) ⩾ g(x0).

377 proper and lower semicontinuous. We assume that F (x) is coercive, i.e., F is bounded from below

378 and F (x) → ∞ when (cid:107)x(cid:107) → ∞, where (cid:107) • (cid:107) is the l2-norm.

380 2.2 Review of APG in the Convex Case

390 xk+1 = prox α kg(yk − α k∇f (yk)),

392 (cid:112)4(tk)2 + 1 + 1

427 zk+1 = prox α kg(yk − α k∇f (yk)),

432 (cid:112)4(tk)2 + 1 + 1

434 Both the two algorithms enjoy the O(cid:0) 1

455 We establish the convergence in the nonconvex case and the O(cid:0) 1

465 3.1 Monotone APG

538 update yk, zk+1, vk+1, tk+1 and xk+1 by (9)-(13).

556 zk+1 = prox α yg(yk − α y∇f (yk)),

557 vk+1 = prox α xg(xk − α x∇f (xk)),

559 (cid:112)4(tk)2 + 1 + 1

593 F (xk+1) ⩽ F (xk) − δ (cid:107)vk+1 − xk(cid:107)2,

601 {vk} generated by (9)-(13) are bounded. Let x* be any accumulation point of {xk}, we have

602 0 ∈ ∂F (x*), i.e., x* is a critical point.

604 and Teboulle [12]’s algorithm, the O(cid:0) 1

610 Theorem 5.1 in [12], we have the following theorem on the accelerated convergence

in the convex

613 (9)-(13) satisfies

624 2 means that APG can ensure to have an O(cid:0) 1

626 α y(N + 1)2(cid:107)x0 − x∗ (cid:107)2.

709 α x = α x,0 ρ h,

710 vk+1 = prox α xg(xk − α x∇f (xk)),

716 F (vk+1) ⩽ F (xk) − δ (cid:107)vk+1 − xk(cid:107)2,

718 (cid:12)(cid:12)(cid:12)(cid:12) (sk)T sk

720 (cid:12)(cid:12)(cid:12)(cid:12)

722 (cid:12)(cid:12)(cid:12)(cid:12) (sk)T rk

724 (cid:12)(cid:12)(cid:12)(cid:12) ,

726 α x,0 =

733 or α x,0 =

746 f (zk+1) ⩽ f (yk) + (cid:104)∇f (yk), zk+1 − yk(cid:105) +

750 (cid:107)zk+1 − yk(cid:107)2.

754 3.2 Convergence Rate under the KL Assumption

767 assume that f and g satisfy the KL property and the desingularising function ϕ(t) = C

806 (k − k3)d2(1 − 2 θ)

854 When F (x) is a semi-algebraic function, the desingularising function ϕ(t) can be chosen to be the

860 rate (at least O(1

929 3.3 Nonmonotone APG

941 nonmonotone APG to speed up convergence. Although the usual APG (2)-(4) is also a nonmonotone

982 F (zk+1) ⩽ ck − δ (cid:107)zk+1 − yk(cid:107)2,

983 F (vk+1) ⩽ ck − δ (cid:107)vk+1 − xk(cid:107)2.

991 Similar to the monotone APG, such vk+1 that satisfies (28) can be found in finite steps, e.g., when

999 ck+1 ⩽ ck − δ (cid:107)xk+1 − yk(cid:107)2

1011 ck+1 ⩽ ck － δ(cid:107)xk+1 － xk(cid:107)2

1103 APG also enjoys the convergence property in the nonconvex case and the O(cid:0) 1

1128 Remark 2. Different from (20), we use the following sk and rk to initialize α x,0 when line search

1134 This is because in nonmonotone APG, vk is not computed in every iteration. So α x,0 should be

1135 initialized only by the recent and existing information. Similarly, α y,0 is initialized by the following

1142 F (yk) － δ(cid:107)zk+1 － yk(cid:107)2 holds.

1154 log(1 + exp(−yixT

1156 i w)) + r(w).

1160 n(cid:88)

1247 2.20

1248 1.70

1249 3.07

1250 1.01

1253 306.03

1254 228.41

1255 146.05

1256 41.79

1259 3.05%

1260 3.05%

1261 3.02%

1262 3.07%

1264 We choose r(w) as the capped l1 penalty [2], defined as

1266 d(cid:88)

1268 r(w) = λ

1270 min(|wi|, θ), θ > 0.

1283 We follow [19] to set $\lambda = 0.0001$, $\theta = 0.1\lambda$ and the starting point as zero vectors. In nmAPG we

1284 set $\eta k = 0.8$. To make a fair comparison, we first run mGIST. The algorithm is terminated when

1290 Matlab 2011a, Windows 7 with an Intel Core i3 2.53 GHz CPU and 4GB memory.

1310 programs and the convergence rate is maintained at O(cid:0) 1

1321 12http://www.public.asu.edu/ yje02/Software/GIST

1322 13http://www.csie.ntu.tw/cjlin/libsvmtools/datasets

1382 [1] E.J. Candes, M.B. Wakin, and S.P. Boyd. Enhancing sparsity by reweighted l1 minimization. Journal of

1384 Fourier Analysis and Applications, 14(5):877 – 905, 2008. 1

1388 [3] S. Foucart and M.J. Lai. Sparsest solutions of underdeterminied linear systems via lq minimization for

1391 $0 < q \leqslant 1$. Applied and Computational Harmonic Analysis, 26(3):395 – 407, 2009. 1

1393 [4] C.H. Zhang. Nearly unbiased variable selection under minimax concave penalty. The Annals of Statistics,

1395 38(2):894 – 942, 2010. 1

1399 on Image Processing, 4(7):932 – 946, 1995. 1

1403 of the American Statistical Association, 96(456):1348 – 1360, 2001. 1

1407 Machine Learning Research, 13(1):3441 – 3473, 2012. 1, 2

1409 [8] Y.E. Nesterov. A method for unconstrained convex minimization problem with the rate of convergence

1411 O(1/k2). Soviet Mathematics Doklady, 27(2):372 – 376, 1983. 1

1413 [9] Y.E. Nesterov. On an approach to the construction of optimal methods of minimization of smooth convex

1417 [10] Y.E. Nesterov. Smooth minimization of nonsmooth functions. Mathematical programming, 103(1):127 –

1421 [11] Y.E. Nesterov. Gradient methods for minimizing composite objective functions. Technical report, Center

1423 for Operations Research and Econometrics(CORE), Catholie University of Louvain, 2007. 1

1426 and deblurring problems. IEEE Transactions on Image Processing, 18(11):2419 – 2434, 2009. 1, 2, 3, 4, 5

1429 SIAM J. Imaging Sciences, 2(1):183 – 202, 2009. 1, 2, 3, 5

1437 ming. arXiv preprint arXiv:1310.3787, 2013. 1, 2

1439 [16] H. Attouch, J. Bolte, and B.F. Svaier. Convergence of descent methods for semi-algebraic and tame prob-

1449 SIAM J. Image Sciences, 7(2):1388 – 1419, 2014. 2

1469 nonsmooth problems. Mathematical Programming, 146(1-2):459 – 494, 2014. 5

1471 [24] H. Zhang and W.W. Hager. A nonmonotone line search technique and its application to unconstrained

1475 [25] S.K. Shevade and S.S. Keerthi. A simple and efficient algorithm for gene selection using sparse logistic

1477 regression. Bioinformatics, 19(17):2246 – 2253, 2003. 7

1479 [26] A. Genkin, D.D. Lewis, and D. Madigan. Large-scale bayesian logistic regression for text categorization.

1481 Technometrics, 49(14):291 – 304, 2007. 7

现有结果：

60 Anonymous Author(s)

109 F (x) = f (x) + g(x),

255 olding (GIST) [19] , Gradient Descent with Proximal Average(GDPA) [20] and Iteratively Reweight-

390 xk+1 = prox ɑ kg(yk − ɑ k∇f (yk)),

427 zk+1 = prox ɑ kg(yk − ɑ k∇f (yk)),

556 zk+1 = prox ɑ yg(yk − ɑ y∇f (yk)),

557 vk+1 = prox ɑ xg(xk − ɑ x∇f (xk)),

613 (9)-(13) satisfies

709 αx = αx,0ρh,

710 vk+1 = proxαxg(xk − αx∇f (xk)),

726 αx,0 =

733 or αx,0 =

767 assume that f and g satisfy the KL property and the desingularising function ϕ(t) = C

806 (k − k3)d2(1 − 2θ)

854 When F (x) is a semi-algebraic function, the desingularising function ϕ(t) can be chosen to be the

860 rate (at least O(1

1154 log(1 + exp(−yixT

1264 We choose r(w) as the capped l1 penalty [2], defined as

1268 r(w) = λ

从上述结果可以看出去除掉了小数点的情况，参考文献和公式中的情况也得到了修正。

4.2 原有功能的改进结果

定义：误报率=（报错数-实错误数）/报错数

减错率=（原系统报错数-现有系统报错数）/原系统报错数

分子分母均为 0 的，定义为 0。

采用前文中 10 篇英文学术论文进行实验，以下是各个功能的实验结果。原有功能 1 请见表 4-2(A)，原有功能 2 请见表 4-2(B)，原有功能请见表 4-2(C)，原有功能请见表 4-2(D)

	原系统		现有系统			
PDF 标题	报错数	误报率	报错数	误报率	实错误数	减错率
SPL	32	0.7813	12	0.4167	7	0.6250
test	21	0.6190	10	0.2000	8	0.5238
APG_nips	125	0.6400	69	0.3478	45	0.4480
0503	119	0.5966	77	0.3766	48	0.3529
RMF-MM3	240	0.7375	115	0.4522	63	0.5208
paper	277	0.7798	103	0.4078	61	0.6282
CFA	299	0.6856	161	0.4161	94	0.4615
LADM	115	0.7565	53	0.4717	28	0.5391
LADM-supp	31	0.8387	7	0.2857	5	0.7742
P2	87	0.8046	34	0.5	17	0.6091
平均		0.7240		0.3875		0.5483

表 4-2(A)

	原系统		现有系统		实错误数	减错率
PDF 标题	报错数	误报率	报错数	误报率		
SPL	41	0.7561	13	0.2308	10	0.6829
test	27	0.9630	2	0.5000	1	0.9259
APG_nips	97	0.9072	9	0.0000	9	0.9072
0503	37	0.8649	18	0.7222	5	0.5135
RMF-MM3	171	0.9649	17	0.6471	6	0.9006
paper	405	0.9654	56	0.7500	14	0.8617
CFA	337	0.9792	38	0.8158	7	0.8872
LADM	135	0.9556	10	0.4000	6	0.9259
LADM-supp	142	0.9507	15	0.5333	7	0.8944
P2	90	0.8000	33	0.4545	18	0.6333
平均		0.9107		0.5054		0.8133

表 4-2(B)

	原系统		现有系统			
PDF 标题	报错数	误报率	报错数	误报率	实错误数	减错率
SPL	29	0.8621	5	0.2000	4	0.8276
test	28	0.8214	4	0.2500	3	0.8571
APG_nips	35	1.0000	1	1.0000	0	0.9714
0503	6	1.0000	0	0.0000	0	1.0000
RMF-MM3	45	1.0000	7	1.0000	0	0.8444
paper	128	0.9844	5	0.6000	2	0.9609
CFA	72	1.0000	21	1.0000	0	0.7083
LADM	30	1.0000	0	0.0000	0	1.0000
LADM-supp	3	1.0000	1	1.0000	0	0.6667
P2	42	0.9762	17	1.0000	0	0.5952
平均		0.9644		0.6050		0.8431

表 4-2(C)

	原系统		现有系统			
PDF 标题	报错数	误报率	报错数	误报率	实错误数	减错率
SPL	16	0.5625	7	0.0000	7	0.5625
test	12	0.7500	3	0.0000	3	0.7500
APG_nips	111	0.9910	19	0.9474	1	0.8288
0503	128	1.0000	34	1.0000	0	0.7344
RMF-MM3	977	0.9949	55	0.9091	5	0.9437
paper	372	0.9785	31	0.7419	8	0.9167
CFA	759	0.9842	52	0.7692	12	0.9315
LADM	227	0.9780	57	0.9123	5	0.7489
LADM-supp	869	0.9413	226	0.7743	51	0.7399
P2	218	0.9541	22	0.5455	10	0.8991
平均		0.9135		0.6600		0.8056

表 4-2(D)

4.3 新增功能的实验结果

4.3.1 英文文本方面

新增功能“Use ‘``’ in the .tex file for the left quotation mark.”。

只在 SPL.pdf 中发现此类错误，且全部找到。结果如下：

```
*** Use "``" in the .tex file for the left quotation mark."
155 are ”four” and ”her virginity” respectively. List pricing is
172 The idea that we only ”use”[4Most of our DNA is¡ non-coding
```

4.3.2 数学公式方面

1. There must be a punctuation after every standalone math expression.

能找到所有的错误情况，对在语句中出现的公式有误报情况。以 P2.pdf 为例，结果如下：

```
*** There must be a punctuation after every standalone math expression. ***
172 vectors, as V = [v1, v2, . . . , vn]. Assuming that they are
188 s.t. V = VZ
193 where Z(cid:62) = [z1, z2, . . . , zn] is the coeffi cient matrix with
199 define the affi nity matrix as A = |Z| + |Z(cid:62)|. Subspaces are
235 where M = PQ(cid:62) [13] and P ∈ Rfi×r, Q ∈ Rr×ft are of rank
260 singular values. Minimizing the trace-norm of M = PQ(cid:62) is
293 We want to get the refined tag matrix ˆO = VPQ(cid:62)T(cid:62)
314 i=1
316 j=1
326 similarity, i.e. gij = cos(vi, vj). The formulation forces tag
338 ) = min
342 where Lv = diag(G1)− G is the Graph Laplacian [3] of the
396 Ls ˆO) = min
```

406 where Ls = diag(H1)−H is the Graph Laplacian of the sim-

626 N = 2

630 N = 5

632 N = 10

735 N = 2

761 N = 5

763 N = 10

2. For norms, use “\|” instead of “||”.

只在 SPL.pdf 中构造了此类问题，且全部找到。结果如下：

*** For norms, use “\|” instead of “||”***

41 ||kx-4||

4.3.3 参考文献方面

0503.pdf 和 CFA.pdf 因书写格式不兼容，无法检查。

1. Every reference must be cited in the body text.

只有 SPL.pdf 和 test.pdf 出现该问题，且全部找到。以 test.pdf 为例，结果如下：

*** Every reference must be cited in the body text. ***

The reference [2] is not cited in the body text.

The reference [4] is not cited in the body text.

The reference [6] is not cited in the body text.

The reference [8] is not cited in the body text.

The reference [10] is not cited in the body text.

The reference [12] is not cited in the body text.

The reference [14] is not cited in the body text.

The reference [16] is not cited in the body text.

The reference [18] is not cited in the body text.

The reference [20] is not cited in the body text.

2. The reference information should be complete.

能发现所有的错误，但对只有 1 页的参考文献一定有误报情况出现。

实验结果如表 4-3 所示

PDF 标题	报错数	误报率	实错误数
SPL	1	0.0000	1
test	2	1.0000	0
APG_nips	15	0.7333	4
paper	58	0.4310	33
RMF-MM3	15	0.6000	6
LADM	22	0.7727	5
LADM-supp	5	0.2000	4
P2	0	0.0000	0
平均		0.4671	

表 4-3

结论

综合减错率和误报率来看，此英文学术论文基本规范检查系统对原有功能的改进是成功的。大部分功能的减错率都超过了 0.8，唯一没有超过的一项误报率降低到了 0.4 以下。就新增功能来说，本系统也能找出所有的错误，但对数学公式后的标点问题和参考文献信息的完整性问题有误报的情况发生。就输出界面来说，本系统较原系统更加友好，提高了师生查阅的效率。

综上所述，本毕业设计较好地完成了原先计划的英文学术论文基本规范检查系统的改进工作。本系统较原系统大大减少了误报的情况，检查结果界面增加了交互，比原系统更能为导师和学生节省时间，实用性也更强。

经本毕业设计改进后的系统仍有改进的余地。比如可以利用英文分词的知识对英文学术论文中不规范使用逗号的情况进行检查；可以和数据库相结合判断参考文献的信息是否有误；可以用其他方式读入 PDF，对图表和公式的处理作出改良。这些都可以作为今后的工作继续提高本系统的实用性。

参考文献

[1]林宙辰，http://www.cis.pku.edu.cn/faculty/vision/zlin/Check%20List%20before%20Handing%20in%20Your%20Paper.htm

[2]骆启明，英文学术论文基本规范检查系统，北京大学本科生毕业论文，2015

[3] Lovegrove, William S., and David F. Brailsford. "Document analysis of PDF files: methods, results and implications." Electronic Publishing--Origination, Dissemination and Design 8.3 (1995): 207-220.

[4] Hadjar, Karim, et al. "Xed: a new tool for extracting hidden structures from electronic documents." Document Image Analysis for Libraries, 2004. Proceedings. First International Workshop on. IEEE, 2004.

[5] Dejean, Herve, and Jean-Luc Meunier. "A system for converting PDF documents into structured XML format." Document Analysis Systems VII. Springer Berlin Heidelberg, 2006. 129-140.

[6] Shinyama, Y. "PDFMiner: Python PDF parser and analyzer (2010)."

[7] Liou, Hsien-Chin. "Development of an English grammar checker: A progress report." CALICO journal 9.1 (2013): 57-70.

[8] Park, Jong C., Martha Stone Palmer, and Clay Washburn. "An English Grammar Checker as a Writing Aid for Students of English as a Second Language." ANLP. 1997.

[9] Naber D. A rule-based style and grammar checker[J]. 2003.

[10] Bredenkamp A, Crysmann B, Petrea M. Looking for Errors: A Declarative Formalism for Resource-adaptive Language Checking[C]//LREC. 2000.

[11] Cobuild C, University of Birmingham (GB). Collins COBUILD advanced learner's English dictionary[M]. HarperCollinsPublishers, 2006.

附录一

原有功能 1 代码：

```
f3 = open('past_participle.txt', 'r')
    plist = f3.readlines()
    a = 1
    t = 0
    for line in checklist:
        b = 1
        for word in plist:
            rule = r"^.*\b" + word.strip() + r"\b.*$"
            if line != '\n' and re.match(rule, line):
                 if t==1 or re.match(r'^.*\bam.*$', line) or\
                re.match(r'^.*\bis.*$', line) or re.match(r'^.*\bare.*$', line) or\
                re.match(r'^.*\bwas.*$', line) or\
                re.match(r'^.*\bwere.*$', line) or\
                re.match(r'^.*\bbeen.*$', line):
                    f2.write(str(a) + ' ' + line)
                    break
            b+=1
            t=0
              if    re.match(r'^.*\bam.*$', line) or\
              re.match(r'^.*\bis.*$', line) or re.match(r'^.*\bare.*$', line) or\
              re.match(r'^.*\bwas.*$', line) or\
              re.match(r'^.*\bwere.*$', line) or\
              re.match(r'^.*\bbeen.*$', line):
                    f2.write(str(a) + ' ' + line)
               t=1
        a+=1
    f3.close()
```

```
a=1
    t=0
    for line in checklist:
        if line != '\n' and re.match(r'^.*ed\b.*$', line):
            if t==1 or re.match(r'^.*\bam.*$', line) or\
            re.match(r'^.*\bis.*$', line) or re.match(r'^.*\bare.*$', line) or\
            re.match(r'^.*\bwas.*$', line) or\
            re.match(r'^.*\bwere.*$', line) or\
            re.match(r'^.*\bbeen.*$', line):
                f2.write(str(a) + ' ' + line)
        t=0
        If   re.match(r'^.*\bam.*$', line) or\
        re.match(r'^.*\bis.*$', line) or re.match(r'^.*\bare.*$', line) or\
        re.match(r'^.*\bwas.*$', line) or\
        re.match(r'^.*\bwere.*$', line) or\
        re.match(r'^.*\bbeen.*$', line):
            t=1
        a+=1
```

原有功能 2 代码：

```
T={'INTRODUCTION':1,'REPRESENTATION':1,'CONCLUSION':1,\
    'REFERENCES':1,'I':1,'II':1,'III':1}
    d = 1
    t = 0
    for line in checklist:
        if line != '\n' and re.match(r'^REFERENCES$', line):
            t=1
        if line != '\n' and re.match(r'^\[.*\].*$', line) and t==1:
            break
        d+=1
```

```
            a = 1
            for line in checklist:
                if line != '\n' and re.match(r'^.*\b[A-Z]{2,}\b.*$', line):
                    if re.match(r'^.*\b[A-Z]{2,}\b\s\([A-Z].*$', line):
                        m=re.match(r'^.*\b([A-Z]{2,})\b\s\(([A-Z]).*$', line)
                        if m.group(1)[0]==m.group(2):
                            T[m.group(1)]=1
                    if re.match(r'^.*\(\b[A-Z]{2,}\b.*$', line):
                        m=re.match(r'^.*\(\b([A-Z]{2,})\b.*$', line)
                        T[m.group(1)]=1
                    m=re.match(r'^.*\b([A-Z]{2,})\b.*$', line)
                    if not m.group(1) in T:
                       f2.write(str(a) + ' ' + line)
                       T[m.group(1)]=1
                a+=1
                if a == d:
                    break
```

原有功能 4 代码：

```
    a = 1
        for line in checklist:
            if line != '\n' and re.match(r'^.*\b[^\w*ly]\w*\b, \
             \b[^(etc)(respectively)(otherwise)(and)]\w*\b\..*$', line) and a<=d:
                 f2.write(str(a) + ' ' + line)
             if line != '\n' and re.match(r'^.*\b[^\w*ly]\w*\b, \
             \b[^(etc)(respectively)(otherwise)(and)]\D*\b\..*$', line) and a>d:
                 f2.write(str(a) + ' ' + line)
                 n0+=1
             a+=1
```

其中 d 是指参考文献开始的行数。

原有功能 11 代码：

```
a = 1
    for line in checklist:
        if line != '\n':
            if    re.match(r'^.*((\S(\(|\[))|((,|\.)\w)).*$', line):
                if re.match(r'^.*[^:]\d*\.\d*.*$',line) or re.match(r'^.*\)\(.*$',line) or\
                re.match(r'^.*\(cid.*$',line):
                     a+=1
                     continue
                f2.write(str(a) + ' ' + line)
        a+=1
```

附录二

代码 1：

```
a=1
    t=0
    for line in checklist:
        if re.match(r'^.*“.*$', line):
            t=1
        if line != '\n' and re.match(r'^.*”.*$', line) and t==0:
            f2.write(str(a) + ' ' + line)
            t=0
        a+=1
```

代码 2：

```
a = 1
   for line in checklist:
       if line != '\n':
           if re.match(r'^.*<.*$', line) or re.match(r'^.*>.*$',line) or \
           re.match(r'^.*=.*$',line) or re.match(r'^.*≤.*$',line) or \
           re.match(r'^.*≥.*$',line):
               if not re.match(r'^.*\.$',line) and not re.match(r'^.*,$',line):
                     f2.write(str(a) + ' ' + line)
       a+=1
```

代码 3：

```
a = 1
   for line in checklist:
       if line != '\n' and re.match(r'^.*\|\|.*$', line):
           f2.write(str(a) + ' ' + line)
       a+=1
```

代码 4：

```
f3 = open(checkfile, 'r')
    p = f3.read()
    divide = p.find("REFERENCES")
    flagdid = p.find("[1]")
    T={}
    d = 1
    x = 0
    t=0
    for line in checklist:
        if line != '\n' and re.match(r'^REFERENCES$', line):
            t = 1
        if line != '\n' and re.match(r'^\[.*\].*$', line) and t==1:
            break
        d+=1

    if flagdid==-1:
        a=0
        for line in checklist:
            a+=1
            if a>=d:
                if line != '\n' and re.match(r'^\[.*\].*$', line):
                    m=re.match(r'^(\[.*\]).*$', line)
                    T[m.group(1)]=0
        a=1
        for line in checklist:
            a+=1
            b = 1
            if a==d:
                break
        for t in T:
            rule = r"^.*"+t+r".*$"
            if line != '\n' and re.match(rule, line):
                T[t]=1
                break
            b+=1
        flag=1
        for t in T:
            if T[t]==0:
                f2.write("The reference [" + str(t) + "] is not cited in the body \
                text.\n")
                flag=0
        if flag==1:
            f2.write("All the references are cited in the body text.\n")
            h0.write("<p>All the references are cited in the body text.</p>")
    f2.close()

else:
    a=0
    x=1
    num=0
    for line in checklist:
         a+=1
         if a>=d:
            if line != '\n' and re.match(r'^\[.*$', line):
                T[x]=0
                num=x
                x+=1

    flag=0
```

```
head=0
tail=0
a=1
for line in checklist:
    a+=1
    if a==d:
        break
    if line != '\n' and re.match(r'^.*\d\].*$', line) and flag==1:
        m=re.match(r'^(.*\d)].*$', line)
        tail=int(m.group(1))
        flag=0
        x = int(head)
        while x<=tail:
            T[x]=1
            x+=1

    if line != '\n' and re.match(r'^.*\[\d*-\d*\].*$', line):
        m=re.match(r'^.*\[(\d*)-(\d*)].*$', line)
        x=int(m.group(1))
        f2.write(m.group(2)+'\n')
        while x<=int(m.group(2)):
            T[x]=1
            x+=1

    if line != '\n' and re.match(r'^.*\[.*\d-$', line):
        flag=1
        m=re.match(r'^.*\[(.*\d)-$', line)
        head=m.group(1)

flag=0
```

```
head=' '
tail=' '
a=1
for line in checklist:
    a+=1
    if a==d:
        break
    if line != '\n' and re.match(r'^\d.*\].*$', line) and flag==1:
        m=re.match(r'^(\d.*)\].*$', line)
        tail=m.group(1)
        flag=0
        x = head+tail
        L=re.split(r'\,+',x)
        for l in L:
            T[int(l)]=1

    if line != '\n' and re.match(r'^.*\[.*\d,\d*\].*$', line):
        m=re.match(r'^.*\[(.*\d,\d*)\].*$', line)
        x=m.group(1)
        L=re.split(r'\,+',x)
        for l in L:
            T[int(l)]=1

    if line != '\n' and re.match(r'^.*\[.*\d,$', line):
        flag=1
        m=re.match(r'^.*\[(.*\d,)$', line)
        head=m.group(1)

for x in range(num):
    x += 1
```

```
        if p[:divide].find("[" + str(x) + "]") > 0:
            T[x]=1

flag=1
for x in T:
    if T[x]==0:
        flag=0
        f2.write("The reference [" + str(x) + "] is not cited in the body \
        text.\n")

if flag==1:
    f2.write("All the references are cited in the body text.\n")
    h0.write("<p>All the references are cited in the body text.</p>")

f3.close()
```

代码 5：

```
d = 1
s = ' '
for line in checklist:
    if line != '\n' and re.match(r'^REFERENCES$', line):
        t=1
    if line != '\n' and re.match(r'^\[.*\].*$', line) and t==1:
        s=line[:-1]
        break
    d+=1
T={}
x=1
a=0
num=0
for line in checklist:
        a+=1
        if a>=d:
            if line != '\n' and re.match(r'^\[.*$', line):
                T[x]=0
                num=x
                x+=1
x=1
sumline=x
while x<=num:
    T[x]=0
    x+=1

a = 1
x=1
for line in checklist:
        a+=1
        if a>d+1:
            if line != '\n' and re.match(r'^\[.*$', line):
                pageflag=0
                timeflag=0
                if re.match(r'^.*http.*$',s):
                    T[x]=1

                else:
                    elements=re.split(r'\,+',s)
                    if re.match(r'^.*\.\.\..*$',s) or len(elements)<4:
                        T[x]=0
                    else:
                        for e in elements:
                            if re.match(r'^.*\d{4}.*$',e) and not \
                            re.match(r'^.*\d－\d.*$',e):
                                timeflag=1

                            if re.match(r'^.*\d－\d.*$',e)or \
                            re.match(r'^.*:\d*\.\d.*$',e):
                                pageflag=1
                        if pageflag==1 and timeflag==1:
                            T[x]=1
                s = line[:-1]
                x+=1
            else:
                s+=line[:-1]
                if x==num and a>=sumline:
                    pageflag=0
                    timeflag=0
                    if re.match(r'^.*http.*$',s):
                        T[x]=1
                    else:
                        elements=re.split(r'\,+',s)
                        if re.match(r'^.*\.\.\..*$',s) or len(elements)<4:
                            T[x]=0
                        else:
                            for e in elements:
                                if re.match(r'^.*\d{4}.*$',e) and not \
                                re.match(r'^.*\d－\d.*$',e):
                                    timeflag=1
                                if re.match(r'^.*\d－\d.*$',e)or \
                                re.match(r'^.*:\d*\.\d.*$',e):
                                    pageflag=1
                            if pageflag==1 and timeflag==1:
```

```
                T[x]=1
flag = 1
for x in T:
    if T[x]==0:
        flag=0
        f2.write("Something may be missing in Reference " + str(x) + ".\n")
if flag==1:
    f2.write("All the references are completed.\n")
```

致谢

林宙辰老师对毕业设计、论文写作给予了悉心指导，在我感觉困惑的时候提供了帮助。每次出现困难时，林老师总能及时地为我解答。实验所用的论文也均由林宙辰老师提供。借此机会，向林宙辰老师表达衷心的感谢。

计算机科学与技术系的肖倾城同学在工具安装方面提供了无私的帮助。智能科学与技术系的郑雅伦同学分享了一些关于英文学术论文参考文献的知识。在此，感谢他们和另外一些朋友的支持。

环境可以影响人，也可以塑造人。认真地做毕业设计和一个好的环境是分不开的。所以在这里表达对舍友王立巍，李广亥和朱奎鑫三位同学的感谢，我们共同营造了一个良好的宿舍环境。

四年的本科时光如白驹过隙，但这四年学到的知识不仅丰富了我的知识体系，还提高了我的能力。而且，四年里我也学到了许多文化知识以外的东西，使我能更好地适应社会，为祖国的建设贡献一份自己的力量。所以，在这离别之际，对母校北京大学有更多的感激与不舍。