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

学　　院：信息科学技术学院

年　　级：2011 级

专　　业：计算机科学与技术

姓　　名：骆启明

学　　号：1100012955

指导教师：林宙辰

2015 年　5 月　22 日

版权声明

任何收存和保管本论文各种版本的单位和个人，未经本论文作者同意，不得将本论文转借他人，亦不得随意复制、抄录、拍照或以任何方式传播，否则一旦引起有碍作者著作权之问题，将可能承担法律责任。

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很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，有些项目具有十分明确的规则，实际上是可以由计算机代为检查的。

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，本毕业设计开发了一个英文学术论文基本规范检查系统。提交一篇 PDF 格式的英文学术论文之后，该系统能自动参照检查项目检查论文是否符合基本规范，并分门别类地指出有错误或需留意的句子及其在文中的位置。学生可以很快地发现错误或需留意之处。

本系统能大大节省导师和学生的检查时间，对提高学生英文学术论文写作也有很大的帮助。

关键词

格式检测；英语语法；正则表达式

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

Qiming Luo (Computer Science and Technology)

Directed by Prof. Zhouchen Lin

Abstract

Usually, both mentors and students have to spend a lot of time checking English research papers after a student have completed one, to see if the paper obeys the basic rules of English paper writing. Among those needing to be checked repeatly, some rules are so clear that they can be checked by computer programs.

In order to save mentors' and students' time wasted on checking papers manually, I developed a system for automatically checking the basic rules of English academic paper writing. When one submits an English research paper in PDF format, the system can help check the paper to see whether it meets the basic rules. It works by locating sentences with certain kinds of mistakes and sentences suspected to be problematic in the submitted paper. Students will be able to find them out quickly.

This system would not only save mentors' and students' time greatly but also assist a lot in improving students' academic English writing.

Key words

Format detection; English grammar; Regular expression

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。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，便有了开发一个英文学术论文基本规范检查系统的需求。由于学生提交的英文学术论文通常都是Portable Document Format（PDF）格式，因此这个检查系统需要特别针对PDF格式的英文论文进行检查。

在提交一篇PDF格式的英文学术论文之后，该英文学术论文检查系统应该能够自动参照检查项目检查论文是否符合基本规范，并对每一项基本规范分门别类地指出有错误或需留意的句子及其在文中的位置。这样学就生可以通过检查结果文档中给出的行号或者通过在PDF文件中搜索错误句子很快地发现文中的错误或需留意之处。

1.2 检查系统的工作内容

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。

由于学生提交的英文学术论文通常都是PDF格式，而算法检查的是纯文本内容。所以需要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。有一些Python语言的PDF处理工具可以使用到这个部分之中。

检查部分是英文学术论文基本规范检查系统的主要部分，它能自动参照英文学术论文基本规范的检查项目，逐项检查文本读取部分读取出的论文文本，并分门别类地指出有错误或需留意的句子及其在文中的位置。

本系统所要检查的基本规范内容包括过多使用被动语态、不定冠词误用、英文标点误用、口头用语与书面语言混淆等十余项英文学术论文基本规范。为了针对这些基本规范进行检查，本系统采用了多种方法，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。

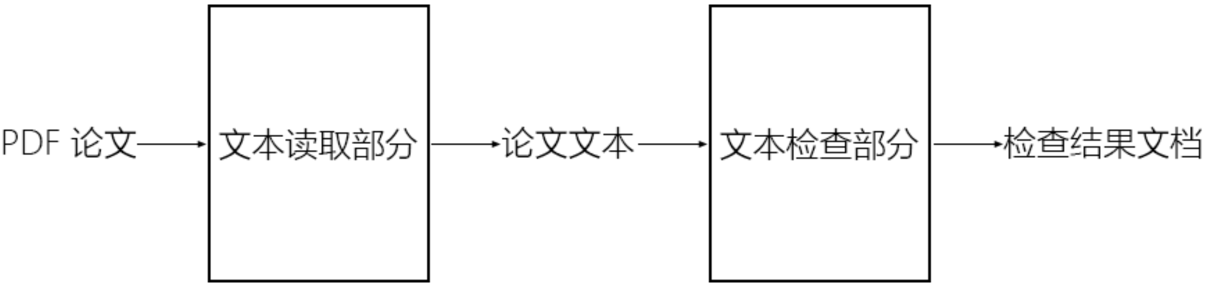


图 1–1 英文学术论文基本规范检查系统的处理流程

1.3 检查系统的优势与意义

在没有使用英文学术论文基本规范检查系统的时候，我们检查英文学术论文通常费时费力，而且由于需要顾及的基本规范过多，往往还会产生一些遗漏的错误。在学生获得论文修改建议，并对论文进行修改之后，导师又需要将论文重新检查一遍，会做很多耗费时间的工作。

在采用了英文学术论文基本规范检查系统之后，与纯人工检查相比的优势有：

（1）节省大量重复检查的时间。原先耗时的检查过程现在可以被计算机程序的快速自动运行取代。导师和学生的检查时间都能大大节省下来。

（2）不会有遗漏的错误。人工检查中会犯下的疏漏在计算机程序中不会出现。

（3）学生可以得到即时的反馈。由于不需要导师花时间检查，学生每次修改完可以自行检查并立刻得到错误反馈，对提高学生英文学术论文的写作水平有很大的帮助。

第二章 从 PDF 格式论文中读取文本

现在有大量的文件都以 PDF 格式存储，在学习、科研活动之中 PDF 格式更是人们的首选。所以在我们检查英文学术论文的时候，我们首先得到的原始文件几乎都会是 PDF 格式的。

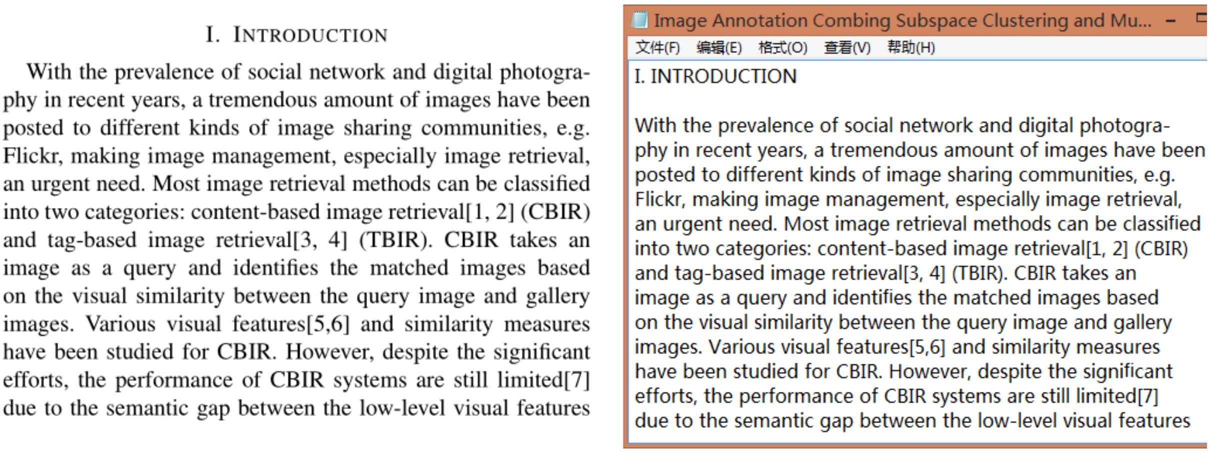
在基本规范检查系统中的文本读取部分，便是负责读取一个 PDF 格式的英文学术论文文档，将 PDF 格式论文中的文本读取出来。

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 [1]。使用 Karim Hadjar 团队开发的 Xed 工具 [2] 或者 Herve Dejean 团队开发的系统 [3] 都可以从 PDF 文档中读取到结构化的 XML 文档。

考虑到基本规范检查系统只会检查从论文中读取出的纯文本内容，并且 Python 语言在处理文本时非常方便，效果也较好，所以最后文本读取部分采用了一个 Python 语言的 PDF 处理工具 PDFMiner [4]。经测试，利用 PDFMiner 中提供的库函数能够成功读取出 PDF 格式文档中的文本内容。

文本读取部分的输入文件是一个 PDF 格式的英文学术论文文档，若其名字为 ABC.pdf，则其读取之后的输出文件是读取出的一个文本格式的中间文档 ABC.out。

在文本格式的中间文档 ABC.out 中，每一行都对应 PDF 文档中的一行。如果 PDF 文档的页面是分为左右两栏的形式，那么文本读取部分会先读取左栏再读取右栏，就像正常人的阅读顺序一样。



I. INTRODUCTION

With the prevalence of social network and digital photography in recent years, a tremendous amount of images have been posted to different kinds of image sharing communities, e.g. Flickr, making image management, especially image retrieval, an urgent need. Most image retrieval methods can be classified into two categories: content-based image retrieval[1, 2] (CBIR) and tag-based image retrieval[3, 4] (TBIR). CBIR takes an image as a query and identifies the matched images based on the visual similarity between the query image and gallery images. Various visual features[5,6] and similarity measures have been studied for CBIR. However, despite the significant efforts, the performance of CBIR systems are still limited[7] due to the semantic gap between the low-level visual features

I. INTRODUCTION

With the prevalence of social network and digital photogra-
phy in recent years, a tremendous amount of images have been
posted to different kinds of image sharing communities, e.g.
Flickr, making image management, especially image retrieval,
an urgent need. Most image retrieval methods can be classified
into two categories: content-based image retrieval[1, 2] (CBIR)
and tag-based image retrieval[3, 4] (TBIR). CBIR takes an
image as a query and identifies the matched images based
on the visual similarity between the query image and gallery
images. Various visual features[5,6] and similarity measures
have been studied for CBIR. However, despite the significant
efforts, the performance of CBIR systems are still limited[7]
due to the semantic gap between the low-level visual features

图 2-1 文本读取部分的原文与运行结果对比

第三章 根据基本规范检查论文文本

文本检查部分是英文学术论文基本规范检查系统的主要部分，在系统中负责检查文本读取部分输出的论文文本内容。

3.1 文本检查部分的工作内容

文本检查部分的输入文件是文本读取部分读取出的中间文档 ABC.out，输出文件是指出有错误或需留意的句子及其在文中的位置的检查结果文档 ABC_check_result.out。

文本检查部分参照的英文学术论文基本规范如下，这也是检查系统所要做的主要工作内容：

1. Use passive tense as few as possible.
2. When using an acronym, give its full name when it first appears (except in title), e.g., “Principal Component Analysis (PCA).”
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. Those are for spoken communication. 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.

在检查结果文档中，各个检查出的错误按照基本规范的项目顺序排列。每一项先输出此项目所检查的基本规范是什么，然后再输出存在问题的行号及该行的句子内容。学生可根据行号或者通过在 PDF 格式论文中搜索句子内容找到需要修改的地方。

3.2 相关工作

英语写作的纠错是一个很多人都关心的议题，有一部分致力于改善英语教学的人士已经做过了不少计算机英语语法检查领域的相关工作。

为了辅助 English-as-a-Foreign-Language（ESL）的英文写作教学，减轻教师的压力，Hsien-Chin Liou 建立了一个包含 1402 个单词词干的小型词典并设计了一个后缀处理器，最终能够成功识别七种不同的英文错误 [5]。

在一个更大的规模上，Park, Jong C.团队利用了一个 317K 大小的语法搭配词汇表和在线电子词典 NOAH，开发了一个具有网页界面的英语检查程序，能够检查出六大类型的英文错误，用户输入一句话，计算机就能检查出他是否在其中犯下了错误 [6]。

检查是否符合英语语法，很多人都会选用正则表达式来作为检查的工具。如致力于开发语法检查工具而不关心单词拼写错误的 Naber D.，他便是使用正则表达式完成了自己的工作 [7]。此外，Bredenkamp A 团队也是通过用正则表达式描述错误的方法检查出了英文文本里的语法错误 [8]。

3.3 功能实现细节

由于英语语法的复杂性，在检查过程中可能会出现一些模棱两可的地方，即无法准确判断是不是属于错误情况。如果出现这种情况，本系统倾向于同样指出这些地方。也就是说，本系统的做法是尽量提高错误情况的召回率，在此基础上可以对所有返回情况

中实际错误情况的准确率作出适当的妥协。

这种做法在实际应用中是有意义的，因为学生很容易就能从返回结果中发现正确情况，而想要发现返回结果中遗漏的错误之处则困难许多。

依照英文学术论文基本规范每一项的顺序，文本检查部分的功能实现细节列举如下：

1. Use passive tense as few as possible.

被动语态在一个口头语言中经常出现，但在论文中过多出现是不妥当的。被动语态可以由动词的过去分词检测到，但过去分词并不能代表一定使用了被动语态。所以在这一项里检测过去分词来判断被动语态，实际上对所有返回情况中实际错误情况的准确率作出了适当的妥协。也就是说，会返回一些并不是被动语态的句子，但学生在看到检查结果文档之后也可以很容易地判断出来。

不规则过去分词较难判断，所以先需要根据词典 [9] 得到一个记录不规则过去分词的词表，然后在程序中通过词表中的词汇生成正则表达式来发现文本中的不规则过去分词：

```
f3 = open('past_participle.txt', 'r')
plist = f3.readlines()
a = 1
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):
            f2.write(str(a) + ' ' + line),
            break
        b+=1
    a+=1
f3.close()
```

规则的过去分词则通过检测词语的-ed 后缀作出判断，这里也牺牲了一部分准确率：

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

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

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，学生可以据此找到是否在每个词第一次出现时给出了完整形式：

```
a = 1
for line in checklist:
    if line != '\n' and re.match(r'^.*\b[A-Z]{2,}\b.*$', line):
        f2.write(str(a) + ' ' + line)
    a+=1
```

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

a 和 an 这两个不定冠词的常见搭配，除了通常基于元音辅音判断的规则之外，还存在着不少特例 [10]。对于它们的特殊搭配，本系统参考了两个整理出的词表文件 a_.txt 和 an_.txt，词表中同时考虑到了单词与字母这两种情况：

```
f_a = open('a_.txt', 'r')
f_an = open('an_.txt', 'r')
check_a = f_a.readlines()
check_an = f_an.readlines()
rule = r"((^(A|.*\W A|.*\ba)\b (?!(one"
for line in check_a:
    rule += "|" + line.strip() + "\W"
rule += "))([aeiou]"
for line in check_an:
    rule += "|" + line.strip()
rule += r"))|(^(An|.*\W An|.*\ban)\b (?!(hour"
for line in check_an:
    rule += "|" + line.strip() + "\W"
rule += "))([b-df-hj-np-tv-z]"
for line in check_a:
    rule += "|" + line.strip()
rule += r"))).*$"
a = 1
for line in checklist:
    if line != '\n' and re.match(rule, line):
        f2.write(str(a) + ' ' + line)
    a+=1
```

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

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：

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

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.

本规范属于标点误用。使用正则表达式 ^.*\b\w*\b, \b\w*\b and \b\w*\b.*$ 找到“A, B, …, Y and Z”形式：

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

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

使用正则表达式 ^.*\b, and \b\w*\b, etc\..*$ 找到“A, B, and C, etc”形式：

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

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

本规范属于口头语言和书面语言的易混淆情况。使用正则表达式 ^.*n\’t.*$ 找到“n’t”形式：

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

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

本规范属于口头语言和书面语言的易混淆情况。使用正则表达式 ^.*can not.*$ 找到“can not”：

```
a = 1
```

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

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

使用几个正则表达式找到这些词的所有错误搭配形式：

```
a = 1
for line in checklist:
    if line != '\n' and re.match(r'^.*\be\.g\b[^\.].*$', line) \
                  or re.match(r'^.*\be[^\.]?g\b.*$', line) \
                  or re.match(r'^.*\betc\b[^\.].*$', line) \
                  or re.match(r'^.*\bet\.?al\b.*$', line) \
                  or re.match(r'^.*\bet al\b[^\.].*$', line):
        f2.write(str(a) + ' ' + line)
    a+=1
```

10. Put a comma before “respectively”.

本规范属于标点误用。使用正则表达式 ^.*[^,] respectively.*$ 找到前面没有逗号的“respectively”：

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

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.

本规范属于写作中格式错误。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。考虑到实际应用中很有可能出现句末是数字，接下来的句首也是数字的情况，便没有对其进行排除：

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

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.”

使用正则表达式 ^.*((Figure \b\w*\b,?)+and Figure \b\w*\b)|((Table \b\w*\b,?)+and Table \b\w*\b).*$ 找到使用“Figure A and Figure B”或者 “Table A and Table B”形式的地方，对于两个以上的情况，也能找到“Figure A, Figure B, and Figure C”或者 “Table A, Table B, and Table C”等形式：

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

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

先数出 REFERENCES 后面列了多少条参考文献，再依次去正文中查找，没有出现的就指出来哪些条没有出现，如果全部出现了则出现“All the references are cited in the body text.”。

```
f3 = open(checkfile, 'r')
p = f3.read()
```

```
        divide = p.find("REFERENCES")
        num = p[divide:].count("[")
        flag = 1
        for x in range(num):
            x += 1
            if p[:divide].find("[" + str(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")
        f3.close()
```

第四章 检查系统的实验

本章中我们使用一篇刻意构造的英文学术论文和一篇真实的英文学术论文，对英文学术论文基本规范检查系统进行实验。

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，用户可以方便地根据行号或者搜索句子内容来找到错误的地方。

4.1 一篇刻意构造的英文学术论文

首先根据英文学术论文基本规范检查系统所检查的基本规范刻意构造出一篇充满错误的英文学术论文（见附录一），再使用本系统检查这篇论文。结果运行良好，构造出的错误均被发现。

按照检查结果文档的格式，即在检查项目下标出行号和原句内容，对成功发现的一些错误列举如下：

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


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


    11 The world record was broken again recently, too.


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


    20 But we know the program has been abused.


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


    171 Most of our DNA is (non-coding DNA)–that is, DNA that
```

*** Make clear the usage of "a" and "an". ***

32 An university is good.

34 A honest man is good.

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

60 I like apple, banana.

*** 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. ***

15 in Louisiana, Mississippi and Alabama areas that experienced

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

78 I like apple, banana, and totato, etc.

*** Do not write "isn't", "aren't", "don't", "doesn't", etc. Those are for spoken communication. Write "is not", "are not", "do not", "does not", etc., instead. ***

86 Because he didn’t have much money in college and it

*** Do not write "can not". Write "cannot" instead. ***

104 A new study argues that birds can not be both masterful

```
*** Notice the correct dots in "e.g.", "etc.", "et al.", etc. ***

125 Marsupials (e.g, kangaroos, opossums) have a pouch.
137 The question is whether Russia et al really mean it.

*** Put a comma before "respectively". ***

146 Steven and James are aged 10 and 13 respectively.

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

161 Most of our DNA is(non-coding DNA)-that is, DNA that
166 For decades,parties at Witanhurst attracted potentates and
169 aspiring professors.Know how to make the best out of them.

*** 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." ***

181 Figure A and Figure B show something important.

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

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

4.2 一篇真实的英文学术论文

接下来选取一篇学生提交的真实英文学术论文（见附录二），使用本系统检查这篇

实际的英文学术论文。结果运行良好，论文中出现的错误均被发现。所用时间远少于人工检查时间，且不会忽略人可能漏掉的错误。

按照检查结果文档的格式，即在检查项目下标出行号和原句内容，对成功发现的一些错误列举如下：

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

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

120 description of experiments is given in Section IV. And finally

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

27 based multi-label learning method is introduced to refine the

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

11 Among these tasks tag-based-image-retrieval (TBIR) plays a cen-

*** Make clear the usage of "a" and "an". ***

106 • We formulate the annotation task as an subspace cluster-

*** Do not write "isn't", "aren't", "don't", "doesn't", etc. Those are for spoken communication. Write "is not", "are not", "do not", "does not", etc., instead. ***

202 rather than de-noising. The reason why we don’t use tag
209 completed and we don’t have to fuse them together.
```

*** Put a comma before "respectively". ***

375 datasets, 25K and 270K respectively, well demonstrated the

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

220 diseases(tags) featuring the genes(images) in high level. The

240 spaces of their own,for example,images belonging to different

第五章 总结及展望

5.1 工作总结

此英文学术论文基本规范检查系统可以达到节省导师和学生人工检查英文学术论文所耗费时间的目的。提交一篇 PDF 格式的英文学术论文之后，该系统能自动参照检查项目检查论文是否符合基本规范，并分门别类地指出有错误或需留意的句子及其在文中的位置。学生可以很快地发现错误或需留意之处。

本系统使用了一些复杂的正则表达式，较好地完成了原先计划的基本规范检查功能。使用本系统能大大节省导师和学生大量重复检查的时间，人工检查中会犯下的疏漏在计算机程序中也不会出现。此外，由于学生可以得到即时的反馈，学生每次修改完可以自行检查并立刻得到错误反馈，本系统对提高学生英文学术论文写作也有很大的帮助。

5.2 进一步工作展望

在本系统中，文本读取部分无法读取 PDF 格式论文中的图表与公式。这些内容都有待进一步的人工检查，能够自动检查出其中的错误也是今后工作中可以继续考虑的。

现在的检查结果文档是纯文本格式，学生尚需要根据行号或者搜索来找到错误内容，未来如果能做到直接在原 PDF 格式论文上做出标记，那么学生发现错误的地方会更为方便。

如功能细节中所述，为了保证检查中错误结果的召回率，本系统的准确率有所妥协，会返回一些计算机无法准确判断的句子。提高返回结果的准确率也是今后工作的一个重要方向，可以作出进一步的优化。

参考文献

[1] 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.

[2] 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.

[3] 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.

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

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

[6] 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.

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

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

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

[10] Gabrielatos C, Torgersen E N, Hoffmann S, et al. A corpus-based sociolinguistic study of indefinite article forms in London English[J]. Journal of English Linguistics, 2010.

附录一：刻意构造的英文学术论文

Robust Latent Low Rank Representation for Subspace Clustering

Hongyang Zhang, Zhouchen Lin, Chao Zhang, Junbin Gao

I. Passage 1

The world record was broken again recently, too.
In the latest study, researchers analysed lung and adrenal-gland tissue samples from 46 dolphins that were found dead in Louisiana, Mississippi and Alabama areas that experienced significantly elevated levels of petroleum compounds.
Jindal knows creationism is a farce, and he knows its a tremendous stain on Louisianas reputation to allow it to be taught in science classes.
But we know the program has been abused.

I have an dog.
I have a ant.
I have a elephant.
An European is good.
An one is good.
An uniform is good.
An union is good.
An unique man is good.
An unit is good.
An university is good.
An user is good.
A honest man is good.
A honor is good.
A hour is good.
A E is good.
A F is good.
A H is good.

I have a dog.
I have an ant.
I have an elephant.
A European is good.
A one is good.
A uniform is good.
A union is good.
A unique man is good.
A unit is good.
A university is good.
A user is good.
An honest man is good.
An honor is good.
An hour is good.
An E is good.
An F is good.
An H is good.

I like apple, banana.
I like Amy, Bob.
I like Google, Apple.

I like apple and banana.
I like Amy and Bob.
I like Google and Apple.

I like apple, banana, and totato.
I like Amy, Bob, and Cherry.
I like Google, Apple, and Microsoft.
I like apple, banana and totato.
I like Amy, Bob and Cherry.
I like Google, Apple and Microsoft.

I like apple, banana, and totato, etc.
I like Amy, Bob, and Cherry, etc.
I like Google, Apple, and Microsoft, etc.
I like apple, banana, totato, etc.
I like Amy, Bob, Cherry, etc.
I like Google, Apple, Microsoft, etc.

Because he didn't have much money in college and it wasn't just handed to him.
Well, just because the general population isn't as fascinated with him as they are with Jobs or Zukerberg doesn't mean some people aren't.
Honestly, I don't get it.

Because he did not have much money in college and it was not just handed to him.
Well, just because the general population is not as fascinated with him as they are with Jobs or Zukerberg does not mean some people are not.
Honestly, I do not get it.

II. Passage 2

A new study argues that birds can not be both masterful divers and flyers, because flying abilities must weaken as the animals adapt to diving.
Why some interfaces can not be sharp.
But we can not win this war by killing them.
I can not access my subscription on Nytimes.com via Firefox on my MacBook Pro.

A new study argues that birds cannot be both masterful divers and flyers, because flying abilities must weaken as the

animals adapt to diving.
Why some interfaces cannot be sharp.
But we cannot win this war by killing them.
I cannot access my subscription on Nytimes.com via Firefox on my MacBook Pro.

Marsupials (e.g., kangaroos, opossums) have a pouch.
Marsupials (e.g, kangaroos, opossums) have a pouch.
Marsupials (e g, kangaroos, opossums) have a pouch.
Marsupials (e g., kangaroos, opossums) have a pouch.
Marsupials (eg, kangaroos, opossums) have a pouch.
Marsupials (eg., kangaroos, opossums) have a pouch.

Sure, there are the textbook examples: Hitler, Stalin, etc.
Sure, there are the textbook examples: Hitler, Stalin, etc

The question is whether Russia et al. really mean it.
The question is whether Russia et al really mean it.
The question is whether Russia et.al really mean it.
The question is whether Russia et.al. really mean it.
The question is whether Russia etal really mean it.
The question is whether Russia etal. really mean it.

Both release new albums this month after eight and 11 years away respectively.
Steven and James are aged 10 and 13 respectively.
Her two daughters, Jo and Fiona, were born in 1968 and 1975 respectively.
Dialog and Robi are the company's Sri Lanka and Bangladesh operations respectively.

Both release new albums this month after eight and 11 years away, respectively.
Steven and James are aged 10 and 13, respectively.
Her two daughters, Jo and Fiona, were born in 1968 and 1975, respectively.
Dialog and Robi are the company's Sri Lanka and Bangladesh operations, respectively.

Most of our DNA is(non-coding DNA)–that is, DNA that does not code for the production of proteins.
Most of our DNA is[non-coding DNA]–that is, DNA that does not code for the production of proteins.
For decades,parties at Witanhurst attracted potentates and royalsincluding,in 1951,Elizabeth,the future Queen.
ontract teaching positions are becoming the norm for many aspiring professors.Know how to make the best out of them.

Most of our DNA is (non-coding DNA)–that is, DNA that does not code for the production of proteins.
Most of our DNA is [non-coding DNA]–that is, DNA that does not code for the production of proteins.
For decades, parties at Witanhurst attracted potentates and royalsincluding, in 1951, Elizabeth, the future Queen.
ontract teaching positions are becoming the norm for many aspiring professors. Know how to make the best out of them.

Figure A and Figure B show something important.
Figure A, Figure B, and Figure C show something important.
Table A and Table B show something important.
Table A, Table B, and Table C show something important.

Figure A and B show something important.
Figure A, B, and C show something important.
Table A and B show something important.
Table A, B, and C show something important.

III. Passage 3

This paper aims at addressing the non-unique-solution issue of LatLRR. On the basis of the theoretical analysis in [1] [3] [5] [7] [9] [11] [13] [15] [17] [19] [21].

References

[1] R. Vidal, "Subspace clustering," *IEEE Signal Process. Mag.*, vol. 28, no. 2, pp. 52–68, 2011.

[2] G. Liu, Z. Lin, S. Yan, J. Sun, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 35, no. 1, pp. 171–184, 2013.

[3] B. Cheng, G. Liu, Z. Huang, and S. Yan, "Multi-task low-rank affinities pursuit for image segmentation," in *ICCV*, 2011, pp. 2439–2446.

[4] G. Liu, Z. Lin, and Y. Yu, "Robust subspace segmentation by low-rank representation," in *ICML*, vol. 3, 2010, pp. 663–670.

[5] G. Liu and S. Yan, "Latent low-rank representation for subspace segmentation and feature extraction," in *ICCV*, 2011, pp. 1615–1622.

[6] S. Wei and Z. Lin, "Analysis and improvement of low rank representation for subspace segmentation," *arXiv:1107.1561*, 2010.

[7] E. Elhamifar and R. Vidal, "Sparse subspace clustering," in *CVPR*, 2009, pp. 2790–2797.

[8] C. Lang, G. Liu, J. Yu, and S. Yan, "Saliency detection by multi-task sparsity pursuit," *IEEE Trans. Image Process.*, vol. 21, no. 3, pp. 1327–1338, 2012.

[9] E. Elhamifar and R. Vidal, "Clustering disjoint subspaces via sparse representation," in *IEEE Int'l Conf. on Acoustics, Speech, and Signal Process.*, 2010, pp. 1926–1929.

[10] L. Zhuang, H. Gao, Z. Lin, Y. Ma, X. Zhang, and N. Yu, "Non-negative low rank and sparse graph for semi-supervised learning," in *CVPR*, 2012, pp. 2328–2335.

[11] E. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *J. ACM*, vol. 58, no. 3, pp. 1–37, 2011.

[12] P. Favaro, R. Vidal, and A. Ravichandran, "A closed form solution for robust subspace estimation and clustering," in *CVPR*, 2011, pp. 1801–1807.

[13] R. Liu, Z. Lin, F. D. L. Torre, and Z. Su, "Fixed-rank representation for unsupervised visual learning," in *CVPR*, 2012, pp. 598–605.

[14] G. Bull and J. Gao, "Transposed low rank representation for image classification," in *2012 Int'l Conf. Digital Image Computing – Techniques and Applications*, 2012, pp. 1–7.

[15] H. Zhang, Z. Lin, and C. Zhang, "A counterexample for the validity of using nuclear norm as a convex surrogate of rank," in *ECML PKDD*, 2013.

[16] T. Zhang, B. Ghanem, S. Liu, and N. Ahuja, "Low-rank sparse learning for robust visual tracking," in *ECCV*, 2012, pp. 470–484.

[17] Y. Ni, J. Sun, X. Yuan, S. Yan, and L. Cheong, "Robust low-rank subspace segmentation with semidefinite guarantees," in *IEEE Int'l Conf. Data Mining Workshops*, 2010, pp. 1179–1188.

[18] J. Wright, Y. Ma, J. Mairal, G. Sapiro, and T. S. Huang, "Sparse representation for computer vision and pattern recognition," *Proc. IEEE*, vol. 98, no. 6, pp. 1031–1044, 2010.

[19] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, 2000.

[20] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 23, no. 6, pp. 643–660, 2001.

[21] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, 1997.

Image Annotation Combing Subspace Clustering and Multi-label Learning

Anonymous VCIP Submission
Paper ID:

***Abstract*—Tag based image management methods, especially image retrieval, has gained much attention during recent years. Among these tasks tag-based-image-retrieval (TBIR) plays a central role. However, tag-depended applications still suffer severely from the degradation of the missing and inaccuracy of user-provided tags. Although there have been a variety of research conducting on this field, their performance is still not satisfying. In this study, we focus on the problem of image annotation that aims to recover the missing tags and correct the noisy ones. The novel component of our study is that we formulate the image annotation task as a subspace clustering framework. It assumes that the observed images are sampled from a union of multiple linear (or affine) subspaces. Images sampled from the same subspace, as well as its corresponding tags, should form a compatible image-tag sub-matrix. We segment the subspaces by the Low Rank Representation method using only visual features and share tags in each subspace subsequently. Thus we can completes the tags and classifies images simultaneously. In addition, a regression based multi-label learning method is introduced to refine the image-tag matrix, exploiting both visual and semantic features. Our empirical study with multiple benchmark datasets for image annotation shows that the proposed algorithm outperforms state-of-the-art approaches when handling missing and noisy tags.**

***Index Terms*—Image Annotation, Subspace Clustering, Low Rank Representation, Voting, Multi-label Learning**

I. Introduction

With the prevalence of social network and digital photography in recent years, a tremendous amount of images have been posted to different kinds of image sharing communities, e.g. Flickr, making image management, especially image retrieval, an urgent need. Most image retrieval methods can be classified into two categories: content-based image retrieval[1, 2] (CBIR) and tag-based image retrieval[3, 4] (TBIR). CBIR takes an image as a query and identifies the matched images based on the visual similarity between the query image and gallery images. Various visual features[5,6] and similarity measures have been studied for CBIR. However, despite the significant efforts, the performance of CBIR systems are still limited[7] due to the semantic gap between the low-level visual features used to represent images and the high-level semantic meaning behind images, as shown in Fig.1 , giving rise to research on TBIR.

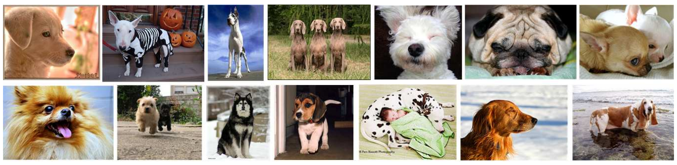


Fig. 1. The problem of semantic gap

Generally, large-scale social images are associated with user-provided tags for describing their semantic content, which could be utilized to overcome the semantic gap. Compare to CBIR, TBIR is usually more accurate in identifying relevant images[8] and more efficient in retrieving relevant images[9].

However, due to the time-consuming tagging process and the arbitrariness of user tagging behaviors, the user-provided tags usually contain imprecise ones, and they are usually incomplete, as also revealed in [10, 11], which leads to performance degradations of TBIR. Therefore, tag refinement, including de-noising and completion, has become an attractive subject of many ongoing researches. However, previous work on tag refinement focused more on de-noising but less on completion, and the results are not very satisfying.

In this paper we propose a novel tag completion scheme denoted as SCML (Subspace Clustering and Multi-label Learning) and tackle the problem from the perspective of subspace clustering. Specifically, we assume that images are sampled from a union of multiple linear (or affine) subspaces. Images sampled from the same subspace, as well as their corresponding tags, should form a compatible (low rank) submatrix. since the tags are too noisy and incomplete, we perform clustering only in the visual feature space, using the state-of-the-art method Low Rank Representation[12]. Then we utilize a simple tag transfer algorithm[29] to share tags in each submatrix. After that, a regression based multi-label learning method is used to tackle the sparsity of the image-tag matrix and refine the annotation result. We utilized both the image features and tag features during the refine procedure to overcome the semantic gap. The main contributions of our research are summarized as follows.

- We formulate the annotation task as an subspace clustering framework, which, in our knowledge, is the first time to exploit the subspace property of the datasets in this field.
- We refine the tag matrix using a regression based multi-label learning method to overcome the sparsity of the the image-tag matrix. It is the first time to using the formulation to perform refinement in the image annotation community.

The remainder of this paper is organized as follows. Section II gives an overview of related work. Section III presents the formulation details of the proposed SC. Then detailed description of experiments is given in Section IV. And finally we conclude the paper in Section V.

II. Previous Work

Many machine learning methods have been proposed for image annotation tasks(see [13] and references therein). They can roughly be grouped into 3 major categories, depending on the way they learn.

The majority of the methods learned the joint probability of image regions and words, which belongs to the generative learning approach. Images are typically represented by properties of each of their segments, or blobs. Once all the images were segmented, quantization can be used to obtain a finite vocabulary of blobs. Thus, the images are treated as bags of words and blobs, each of which are assumed generated by hidden variables which spawn a multivariate distribution over blobs and a multinomial distribution over words. Once the joint word-blob probabilities were learned, the annotation problem for a given image is reduced to a likelihood problem relating blobs to words.

Barnard et al.[15] and Wang C et al.[16] extended the LDA model and proposed a Correlation LDA to relate words and images. Other work[17] considers the annotation task as a machine translation problem. Jeon J et al.[18] explore the Mixture Model approach and propose the famous CMRM model. Other researchers further extended CRM to CRM[19], MBRM[20], DCMRM[21].For a comprehensive review of the Mixture Model method, see [22].

Different from the generative methods, discriminative methods use the classified images and tags to train a dictionary of concept models and formulate image annotation as a supervised learning problem. They annotate new images using the likelihood between images and tags.

In particular, Subspace techniques are used for learning shared subspace[23] or for clustering[24, 25]. However, our SCML model is absolutely different from all these works. We will explain in detail in Sec III.

In recent years, search-based method, or local learning, have gained more and more attention owing to their efficiency in time. The elegant search-based method JEC[29] shows that simple baseline techniques can also get high performance. TagProp[7] applied metric learning in the neighborhood of the feature space to annotation query images. Search-based methods always search in the neighborhood to find the most relevant images to the query image, and passing tags to the query image[31]. Tag Relevance[32] design a Neighbor-Voting algorithm to transfer tags, which is similar to our tag propagation method.

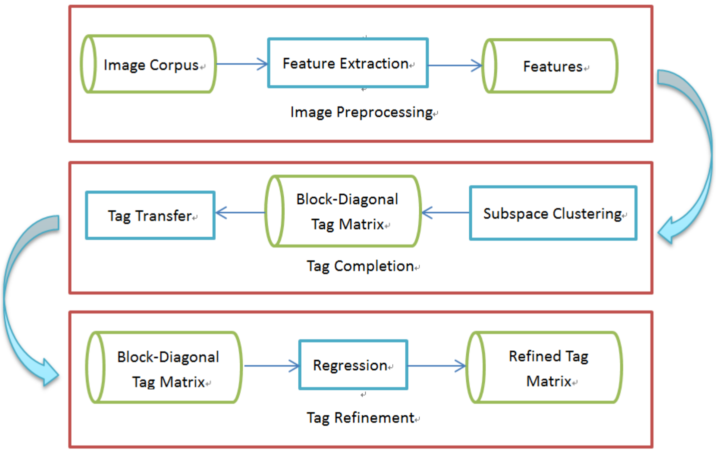


Fig. 2. The flowchart of the proposed SCML

III. Annotation based on Subspace Clustering and Multi-label Learning

A. Framework Overview

In this section, we introduce the proposed annotation framework as illustrated by Figure 2. We summarize the flowchart as follows:

1) Image Preprocessing: For simplicity, we just choose part of the image features employed in Tagprop[7], namely, 1 Gist descriptor and 8 bag-of-features (2 features types x 2 descriptors x 2 layouts). These features include global image descriptors such as Gist[39] and local descriptors such as SIFT and robust hue descriptor[40], which are commonly used for image search and annotation. Here we just concatenate all the features together to form the unique feature vector for each image. We further follow the way they compute the distance from the descriptors, using L2 as the base metric for Gist, and χ^2 for the others.

2) Tag Completion: Here We apply the Low Rank Representation[12] method to cluster the image feature vectors into different subspaces. We can get a block-diagonal matrix using the result of the algorithm, each submatrix of which represents an independent subspace. Thus we can identify the membership of each image from the block-diagonal matrix. Then we can complete the image-tag matrix by sharing tags in each subspaces separately. To transfer tags in the subspace, we just adopt the algorithm proposed in JEC[29] for efficiency. Note that the tag completion procedure focuses on completing rather than de-noising. The reason why we don't use tag feature is that the user-provided tags are too noisy and too incomplete, and it is difficult to fuse tag feature with image feature together to improve the clustering performance. We will utilize tags to overcome the semantic gap problem in the following Tag Refinement procedure, where tags are more completed and we don't have to fuse them together.

3) Tag Refinement: To refine the tag-image matrix, we adopt a regression based multi-label learning method introduced in the bioinformatics field[33], which was developed for exploring the gene-disease associations. This method is first used in the annotation field. The original problem is very similar to the image annotation problem in that the nonlinear associations between images and tags are almost the same as the links between genes and diseases, where diseases(tags) featuring the genes(images) in high level. The gene-disease matrix is very sparse, making most of the traditional method,e.g. RPCA model[8], unsuitable. In their work, the authors solve the problem using regression model, which, in essential, is a multi-label learning framework. We reformulate the object function in an equivalent way and solve the problem using LADM[35]. We use WordNet[41,42] to compute tag feature, which could be improved in further work.

B. Subspace Clustering

In the image annotation field, one usually needs a parametric model to characterize the user-provided data. Researchers have exploit the RPCA model [8], in which the tag refinement problem is formulated as a decomposition of the user-provided tag matrix into a low-rank refined matrix and a sparse error matrix. However, a given data set is seldom well described by a single subspace. In practice, the data points could be drawn from multiple subspaces and the membership of the data points to the subspaces might be unknown[41]. Images corresponding to different categories may lie in different subspaces of their own,for example,images belonging to different categories should lies in different subspaces. Thus it is more reasonable to assume the data as samples approximately drawn from a mixture of several low-rank subspaces, as shown in Fig. 3[34], leading to the challenging problem of subspace segmentation. Here, the goal is to segment (or cluster) data into clusters with each cluster corresponding to a subspace.

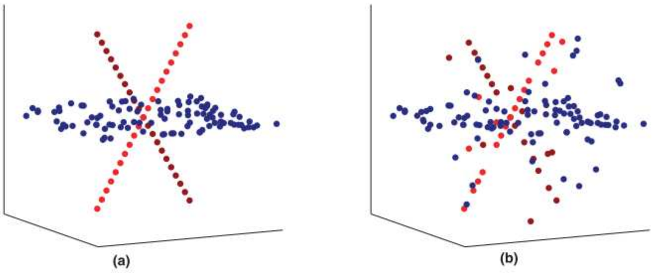


Fig. 3. A mixture of subspaces consisting of a 2D plane and two 1D lines. (a) The samples are strictly drawn from the underlying subspaces. (b) The samples are approximately drawn from the underlying subspaces.

Subspace clustering has found numerous applications in computer vision (e.g., image segmentation [44], motion segmentation [45] and face clustering [46]), image processing and so on[41].

1) Low Rank Representation: A number of approaches to subspace clustering have been proposed in the past two decades. A review of methods from the data mining community can be found in [47].

One of the state-of-the-art subspace clustering method it the Low Rank Representation [34] algorithm, which seeks the lowest rank representation among all the candidates that can represent the data samples as linear combinations of the bases in a given dictionary. Low Rank Representation can perform robust subspace clustering and error correction in an efficient and effective way.

Here,we denote the set of image feature vectors as $X = [x_1, x_2, \ldots, x_n]$, drawn from a union of k subspaces $\{S_i\}_{i=1}^{k}$.Each column of X is a feature vector in R^D and can be represented by the linear combination of the basis in a "dictionary" $A = [a_1, a_2, \ldots, a_m]$:

$$X = AZ,$$

where $Z = [z_1, z_2, \ldots, z_n]$ is the coefficient matrix with each z_i being the representation of x_i.In order to segment the data into their respective subspaces, we need to compute an affinity matrix that encodes the pairwise affinities between data vectors.So we just use the data X itself as the dictionary,i.e. represented by the linear combination of the basis in a "dictionary" $A = [a_1, a_2, \ldots, a_m]$:

$$\min_{Z} = \|Z\|_*,$$
$$s.t., X = XZ.$$

This is the Low Rank Representation formulation.We can solve the problem using LADM method[35] efficiently.After we got the solution Z, we can perform subspace segmentation according to Algorithm 2[34] and identify the subspaces.

2) Tag Transfer: Since we have identified the subspaces and membership, we can refine all the image-tag subspaces. For each subspace, we rank the tags of all the images in the subspace, and complete the tags of each image in the same subspace, taking tag frequency, tag co-occurrence and local frequency in to consideration[29]. The transfer algorithm can be improved using more sophisticated strategies [31, 32].

C. Image Tag Refinement

Let $x_i \in R^D$ denote the feature vector of image i, and $y_j \in R^T$ denote the feature vector of tag j, which could be computed from pre-trained word2vec[38], the refinement problem is to recover a low-rank matrix $N \in R^{D \times T}$ using the observed entries from the image-tag matrix M, where $M = X * Z = X - Z$. Denote the set of observed entries by Ω,i.e. $\Omega = \{(i,j)|M_{i,j} > 0\}$.The entry $M_{i,j}$ of the matrix is modeled as $M_{i,j} = x_i^T N y_j$ and the goal is to learn N using the observed entries Ω.We formulate the refinement procedure as a regression based multi-label learning method

$$\min \sum_{(i,j)\in\Omega} loss(M_{i,j}, x_i^T N y_j) + \lambda rank(N),$$

Using the standard relaxation of the rank constraint is the trace norm,thus we get:

$$\min \sum_{(i,j)\in\Omega} loss(M_{i,j}, x_i^T N y_j) + \lambda \|N\|_*,$$

A common choice for loss function is the squared loss function. Note that the object function is convex, so we just adopt the LADM[35] method to solve the problem instead of the original method[33].

IV. Experimental Evaluation

The proposed algorithm was evaluated on two large volume benchmark datasets: MIRFlickr- 25K [36] and NUS-WIDE-270K [37].

A. Datasets and Experimental Setup

The MIRFlickr-25K and NUS-WIDE-270K datasets are both collected from Flickr website. The MIRFlickr-25K dataset contains 25, 000 images with 1, 386 tags. The second dataset, NUS-WIDE- 270K, comprises a total of 269, 648 images with 5, 018 unique tags. Note that the tags in the above two collections are rather noisy and many of them are misspelling or meaningless words.Hence, a pre-processing was performed to filter out these tags. We matched each tag with entries in a Wikipedia thesaurus and only the tags with coordinates in Wikipedia were retained. We use the pre-trained word and phrase vectors [38] to extract tag vectors. As to the parameters of the Low Rank Representation algorithm, we just adopt the suggested settings.

Table I summarizes the basic statistics of the two datasets.

TABLE I
Statistics of Two Datasets

Statistics	MIR-Flickr	NUS-WIDE
No. of images	25,000	269,648
Vocabulary Size	1,386	5,018
No. of Tags per Image(mean/max)	12.7/76	14.4/82
No. of Images per Tag(mean/max)	416.5/76,890	575.5/87,120

B. Comparison to state-of-the-art Annotation Methods

To evaluate the performance of the proposed SCML, we compare the proposed SCML algorithm to the following remarkable state-of-the-art algorithms for automatic image annotation:

1. Multiple Bernoulli relevance models (MBRM) [20] that models the joint distribution of annotation tags and visual features by a mixture distribution.

2. Joint equal contribution method (JEC) [29] that finds appropriate annotation words for a test image by a k nearest neighbor classifier that combines multiple distance measures derived from different visual features.

3. TagProp[7] which perform discriminative metric learning in nearest neighbor models for image auto-annotation.Tags of test images are predicted using a weighted nearest-neighbor model to exploit labeled training images. Neighbor weights are based on neighbor rank or distance.

The experimental results of tag completion are measured with average precision@5(i.e. AP@5) and coverage@5(i.e.C@5). Precision@5 is to measure the ratio of correct tags in the top 5 competed tags, coverage@5 is to measure the ratio of test images with at least one correct completed tag.

Table II concludes the performance of SCML comparing to another two state-of-the-art methods.

TABLE II
Annotation Performance Comparisions

	SCML	MBRM	JEC	TagProp
AP@5	0.236	0.139	0.144	0.146
C@5	0.50	0.47	0.49	0.51

V. Conclusion

Motivated by the fact that the existing user-provided image tags in public photo sharing platforms are imprecise and incomplete, we proposed an efficient approach for image tag refinement by formulating the problem in a subspace clustering framework combing with a regression based multi-label learning refinement procedure. Experiments on large-scale image datasets, 25K and 270K respectively, well demonstrated the effectiveness of the proposed algorithm. Our future work shall focus on three directions. First, more effective feature engineering techniques shall be integrated in the current framework for representing the image content more precisely. For example, recent study on Image Captioning shows that pre-trained deep features,such as fc7, and deep features fine-tuned for the specific task , such as fc7-fine, are very effective at representing high-level semantic information comparing to traditional features such as Gist[48].We could utilize deep features for better representing images. Second,the tag features are computed using pre-trained word2vec tools, whose train data has a different distribution with our specific task, leading the word vectors not so precise to our specific task. We may exploit word embedding techniques to train our own word vector so as to better present the tags.Third, kernel method should be developed so that the proposed algorithm can better exploit the nonlinear associations between images and tags

致谢

本科阶段的最后一件工作终于完成了，谨向所有关心我的老师、同学表示衷心的感谢。

感谢林宙辰老师对毕业设计、论文写作给予的悉心指导，在我感觉困惑的时候提供了宝贵的建议。

感谢家人与朋友们一直以来的支持。正是有了你们的支持我才能够战胜一个又一个的困难。

最后，感谢我最爱的北京大学，陪伴我在这四年中不断成长。

北京大学学位论文原创性声明和使用授权说明

原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品或成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本声明的法律结果由本人承担。

论文作者签名：　　　　　　日期：　　年　　月　　日

学位论文使用授权说明

（必须装订在提交学校图书馆的印刷本）

本人完全了解北京大学关于收集、保存、使用学位论文的规定，即：

- 按照学校要求提交学位论文的印刷本和电子版本；
- 学校有权保存学位论文的印刷本和电子版，并提供目录检索与阅览服务，在校园网上提供服务；
- 学校可以采用影印、缩印、数字化或其它复制手段保存论文；
- 因某种特殊原因需要延迟发布学位论文电子版，授权学校在 □ 一年 / □ 两年 / □ 三年以后在校园网上全文发布。

（保密论文在解密后遵守此规定）

论文作者签名：　　　　　　导师签名：　　　　　　日期：　　年　　月　　日