(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 101833579 A
(43)申请公布日 2010.09.15

(21)申请号 201010168294.0

(22)申请日 2010.05.11

(71)申请人 同方知网（北京）技术有限公司
地址 100084 北京市海淀区清华园清华大学36区华业大厦B1410、1412、1414室

(72)发明人 张振海 孙雄勇

(74)专利代理机构 北京捷诚信通专利事务所 11221
代理人 魏殿绅

(51)Int.Cl.
G06F 17/30(2006.01)

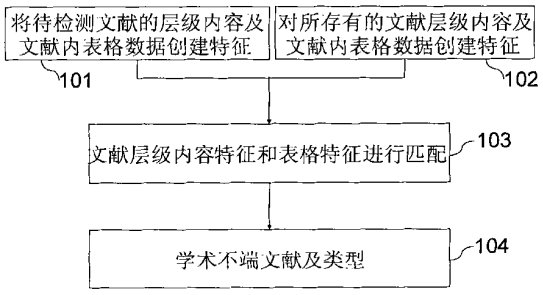
权利要求书 2 页 说明书 4 页 附图 3 页

(54)发明名称

一种自动检测学术不端文献的方法及系统

(57)摘要

本发明公开了一种自动检测学术不端文献的方法及系统，所述方法将待检测文献的层级内容及文献内表格数据创建特征；对所存有的文献层级内容及文献内的表格数据创建特征；将待检测文献的层级内容特征和待检测文献内的表格数据特征与所存有文献的层级内容特征及所存有文献内的表格特征进行匹配；判断待检测文献中是否含有学术不端内容、不端表格数据及不端内容的类型。所述系统包括待检测文献特征区、待检测文献比对资源区、分层内容特征匹配区及不端学术文献及类型判定区。本发明通过分层多阶特征结构，不仅可以对超长文献进行快速检测，而且，也满足了文献最小特征粒度短句的检测，提高了检准率和检全率；而且该发明还支持文献内表格数据特征的创建及匹配全部文献的一次性匹配。



CN 101833579 A

1. 一种自动检测学术不端文献的方法，其特征在于，待检测文献特征与所存全部文献特征同时进行匹配，所述方法具体包括：

将待检测文献的层级内容及文献内表格数据创建特征；

对所存有的文献层级内容及文献内的表格数据创建特征；

将待检测文献的层级内容特征和待检测文献内的表格数据特征与所存有文献的层级内容特征及所存有文献内的表格特征进行匹配；

判断待检测文献中是否含有学术不端内容、不端表格数据及不端内容的类型。

2. 根据权利要求 1 所述的自动检测学术不端文献的方法，其特征在于，所述待检测文献层级内容与所存有文献层级内容创建的特征为唯一特征。

3. 根据权利要求 1 所述的自动检测学术不端文献的方法，其特征在于，所述不端学术文献及类型的判断是根据待检测文献与所存文献中的相似阈值、文献时间属性及文献作者属性，其不端内容的类型包括抄袭与剽窃、篡改及一稿多投。

4. 根据权利要求 1 所述的自动检测学术不端文献的方法，其特征在于，所述待测文献与所存有文献特征间的匹配是进行篇章级别的特征匹配、段落级别的特征匹配及句子级别的特征匹配。

5. 根据权利要求 1 所述的自动检测学术不端文献的方法，其特征在于，所述文献内表格数据特征是通过提取表格的属性信息、表格内容的文字处理及表格的行数和列数进行创建。

6. 根据权利要求 4 所述的自动检测学术不端文献的方法，其特征在于，所述

文献全文篇幅的特征创建，是利用关键词词典对全文分词，保留关键词词典中的词，将全部词排序并统计词频，按照词频比例排序，得到次序列表，依据该列表生成文献全文级别的一个特征；

或

按照词拼写排序得到一个词序列表，依据该列表生成文献全文级别的一个特征；

段落特征创建，是利用单元词词典对段落分词，只保留单元词词典中的词，并统计频率，按照词频比例排序，依据该列表生成段落级别的一个特征；

所述句子特征创建，是利用单元词词典对句子分词，只保留单元词词典中的词，利用同义词词典更新该列表中词，针对更新之后的列表按照词形排序，依据排序好的列表生成句子级别的一个特征。

7. 一种自动检测学术不端文献的系统，其特征在于，包括：

待检测文献特征区，用于对接收到的待检测文献的层级内容及文献内表格数据创建特征；

待检测文献比对资源区，用于对所存有的文献层级内容及文献内的表格数据创建特征；

分层特征匹配区，用于接收待检测文献特征区创建的文献的层级内容特征及文献内的表格数据特征，并将所述待检测文献的层级内容特征和文献内的表格数据特征与待测文献比对资源区所存有的文献的层级内容特征及文献内的表格特征进行匹配；

不端学术文献及类型判定区，用于判断待检测文献中是否含有学术不端内容、不端表格数据及不端学术内容的类型。

8. 根据权利要求 7 所述的自动检测学术不端文献的系统，其特征在于，所述待检测文献层级内容与所存有文献层级创建的特征为唯一特征。

9. 根据权利要求 7 所述的自动检测学术不端文献的系统，其特征在于，所述不端学术文献及类型判定区的判断是根据待检测文献与所存有文献中的相似阈值、文献时间属性及文献作者属性；所述文献内表格数据特征是通过提取表格的属性信息、表格内容的文字处理及表格的行数和列数进行创建；其不端内容的类型包括抄袭与剽窃、篡改及一稿多投。

10. 根据权利要求 7 所述的自动检测学术不端文献的系统，其特征在于，所述文献层级内容是按照文献篇幅、段落、句子进行划分。

一种自动检测学术不端文献的方法及系统

技术领域

[0001] 本发明涉及智能信息处理和计算机技术领域，尤其涉及一种自动检测学术不端文献及文献内表格数据的方法及系统。

背景技术

[0002] 随着网络的快速发展和迅速普及，目前在互联网上发布的电子文本成为当前知识产权保护的一个重点。由于电子文本易于复制和下载，已成为许多人研究、引用的对象，一些电子文本被大版面的复制而被认为抄袭的案例时有发生。而目前网络上的电子文本保护措施主要通过阻止和检测法。

[0003] 目前，也出现了电子文本内容剽窃的方法，如专利申请号为“200810232309.8 一种检测及定位电子文本内容剽窃的方法”与专利申请号为“03134562.X 一种利用计算机程序检测电子文本剽窃的方法”该现有专利主要是通过向计算机系统提交被检测文本，提取特征模块生成项序列，然后从项序列中依次取出每个项映射到已知项表上，生成疑似剽窃队列，获取剽窃证据表；最后计算文本的雷同度，判断被检测文本中是否含有剽窃的内容。上述检测过程只是单层特征的检测，不能针对文献内的表格创建特征；其匹配也不是一篇文献针对多篇文献同时进行匹配，只是一篇针对一篇，两篇文献之间的匹配（如图 1 所示）浪费了匹配的时间；而且对内容也只是检测抄袭的过程。

发明内容

[0004] 为解决上述中存在的问题与缺陷，本发明提供了一种不仅支持一篇文献针对多篇文献间的一次性匹配，而且还可检测文献内表格数据、判断不端文献抄袭、篡改、一稿多投类型的自动检测学术不端文献的方法及系统。所述技术方案如下：

[0005] 一种自动检测学术不端文献的方法，包括：

[0006] 将待检测文献的层级内容及文献内表格数据创建特征；

[0007] 对所存有的文献层级内容及文献内的表格数据创建特征；

[0008] 将待检测文献的层级内容特征和待检测文献内的表格数据特征与所存有文献的层级内容特征及所存有文献内的表格特征进行匹配；

[0009] 判断待检测文献中是否含有学术不端内容、不端表格数据及不端内容的类型。

[0010] 一种自动检测学术不端文献的系统，包括：

[0011] 待检测文献特征区，用于对接收到的待检测文献的层级内容及文献内表格数据创建特征；

[0012] 待检测文献比对资源区，用于对所存有的文献层级内容及文献内的表格数据创建特征；

[0013] 分层特征匹配区，用于接收待检测文献特征区创建的文献的层级内容特征及文献内的表格数据特征，并将所述待检测文献的层级内容特征和文献内的表格数据特征与待测文献比对资源区所存有的文献的层级内容特征及文献内的表格特征进行匹配；

[0014] 不端学术文献及类型判定区，用于判断待检测文献中是否含有学术不端内容、不端表格数据及不端学术内容的类型。

[0015] 本发明提供的技术方案的有益效果是：

[0016] 通过分层多阶特征结构，不仅可以对超长文献进行快速检测，而且，也满足了文献最小特征粒度短句的检测，提高了检准率和检全率；而且该发明还支持文献内表格数据特征的创建及匹配全部文献的一次性匹配。

附图说明

[0017] 图 1 是现有技术文本内容匹配方法结构图；

[0018] 图 2 是检测学术不端文献方法流程图；

[0019] 图 3 是学术不端文献匹配方法结构图；

[0020] 图 4 是文献多层特征生成方法结构图；

[0021] 图 5 是文献内表格数据特征生成方法结构图；

[0022] 图 6 是检测学术不端文献系统结构图。

具体实施方式

[0023] 为使本发明的目的、技术方案和优点更加清楚，下面将结合附图对本发明实施方式作进一步地详细描述：

[0024] 实施例 1

[0025] 本实施例提供了一种自动检测学术不端文献的方法如图 2 所示，该方法包括：

[0026] 步骤 101 将待检测文献的层级内容及文献内表格数据创建特征。

[0027] 步骤 102 对所存有的文献层级内容及文献内的表格数据创建特征；

[0028] 上述待检测文献与所存有的文献是指任意文献，对其文献进行分层处理，按照篇章、段落、句子等层级分别创建唯一特征。

[0029] 步骤 103 将待检测文献的层级内容特征和待检测文献内的表格数据特征与所存有文献的层级内容特征及所存有文献内的表格特征进行匹配；

[0030] 首先是进行篇章级别的特征匹配，如果整个篇章级别匹配成功，则不再对其段落级别进行匹配，如果整个篇章级别匹配不成功的话，则继续对其段落级别进行匹配。如果整个段落级别匹配成功，则不再对其句子级别进行匹配；如果整个段落匹配不成功的话，则继续对其句子级别进行匹配，总之，各层如果匹配成功，则不再进行该特征下更小粒度层的匹配。其对待测文献特征与所存有文献特征库的匹配方法如图 3 所示，待测文献多层特征库中的特征所存有全部文献特征库集成倒排索引中的特征 ID 进行相匹配，如果匹配成功则判断出所述文献的 ID、特征匹配的数量及特征原始文本的长度。

[0031] 步骤 104 判定学术不端文献及类型；

[0032] 不端文献的类型为抄袭与剽窃、或者篡改、或者一稿多投等学术不端文献类型。

[0033] 如图 4 所示，为全文特征、章节特征、段落特征及句子特征的生成方法结构图，其中全文特征提取方法、章节特征提取方法是利用关键词词典对全文分词（对表征文献主题内容具有实质意义的词），只保留关键词词典中的词，全部词排序并统计词频，照词频比例排序，得到词序列表，依据该列表生成全文级别的一个特征，或按照词拼写排序得到一个词

序列表，依据该列表生成全文级别的一个特征。段落特征生成方法，利用单元词词典对段落分词，（从文献内容中抽出的最基本的、字面上不能再分的词。如“经济、美国、鲁迅”等无定语的词都是单元词，单元词不包括虚词、介词、连词、助词等无实质表征的词汇）只保留单元词词典中的词，并统计频率，按照词频比例排序，依据该列表生成段落级别的一个特征。句子特征生成方法，利用单元词词典对句子分词，（从文献内容中抽出的最基本的、字面上不能再分的代表实质意义的词。如“经济、美国、鲁迅”等无定语的词都是单元词。单元词不包括虚词、介词、连词、助词等无实质表征的词汇）只保留单元词词典中的词，利用同义词词典更新该列表中词，例如“电脑”一词全部替换为“计算机”“ontology”全部替换为“本体”，针对替换之后的列表按照词形排序。依据排序好的列表生成句子级别的一个特征。

[0034] 如图 5 所示，文献内表格数据生成方法结构图，首先是根据文献表格内容提取表格的属性信息，特征库根据表格内容中的标题信息、行数据信息、列数据信息、多行组合信息及多列组合信息来提取表格特征的特征。在提取时，其全部表格内容作为文字处理，数字如果有小数点则循环乘 10 至转化为整数为止。根据表格列数、行数及列数的多少分别组合多行、多列表格，参见表 1 和表 2。

[0035] 表 1

[0036]

行数	组合粒度
＜ 9 行	不组合，单行为一个特征单位
＞ 8 行 ＜ 20 行	2 行组合为一个特征单位
＞ 19 行	3 行组合为一个特征单位

[0037] 表 2

[0038]

列数	组合粒度
＜ 9 列	不组合，单列为一个特征单位
＞ 8 列 ＜ 20 列	2 列组合为一个特征单位
＞ 19 列	3 列组合为一个特征单位

[0039] 上述检测学术不端的方法适用于任何语言文献，在检测其它语言特征库时，其特征库的生成方法过程与所用词词典内容有所区别。

[0040] 实施例 2

[0041] 如图 6 所示，为检测学术不端文献系统结构图，包括待检测文献特征区、待测文献比对资源区、分层特征匹配区及不端学术文献及类型判定区，其中待测文献特征区，对接收到的待检测文献的层级内容及文献内表格数据创建特征；待测文献比对资源区，用于对所存有的文献层级内容及文献内的表格数据创建特征；要检测的文献来源可以是用户自由指定，实时生成文献多层内容特征加入到文献特征库中；待测文献比对资源区的文献可以是中国学术文献网络出版总库中的文献，也可以来源用户自由指定的文献。分层特征匹配区，用于接收待检测文献特征区创建的文献的层级内容特征及文献内的表格数据特征，并将所述待检测文献的层级内容特征和文献内的表格数据特征与待测文献比对资源区所存有的文献的层级内容特征及文献内的表格特征进行匹配；不端学术文献及类型判定区，用于判断待检测文献中是否含有学术不端内容、不端表格数据及不端学术内容的类型。

[0042] 所述待检测文献层级内容与所存有文献层级内容创建的特征为唯一特征，其文献层级是按照文献篇幅、段落、句子进行划分，这种分层多阶特征结构，不仅可以满足对超长文献的快速检测，而且也满足了对文献的最小特征粒度的短句。上述文献内表格特征的生成方法是在特征库中通过提取表格的属性信息，即文献的标题信息、行数据信息、列数据信息、多行组合信息及多列组合信息进行提取表格特征。

[0043] 上述不端学术文献及类型判定区的判断是根据待检测文献与所存有文献中的相似阈值、文献时间属性及文献作者属性，其不端内容的类型包括抄袭与剽窃、篡改及一稿多投。

[0044] 当然，本发明还可有其他多种实施例，在不背离本发明精神及其实质的情况下，本领域技术人员当可根据本发明作出各种相应的改变和变形，但这些相应的改变和变形都应属于本发明所附的权利要求的保护范围。

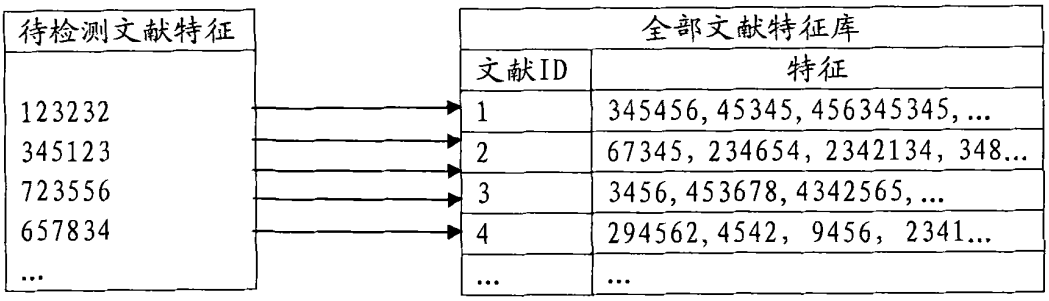


图 1

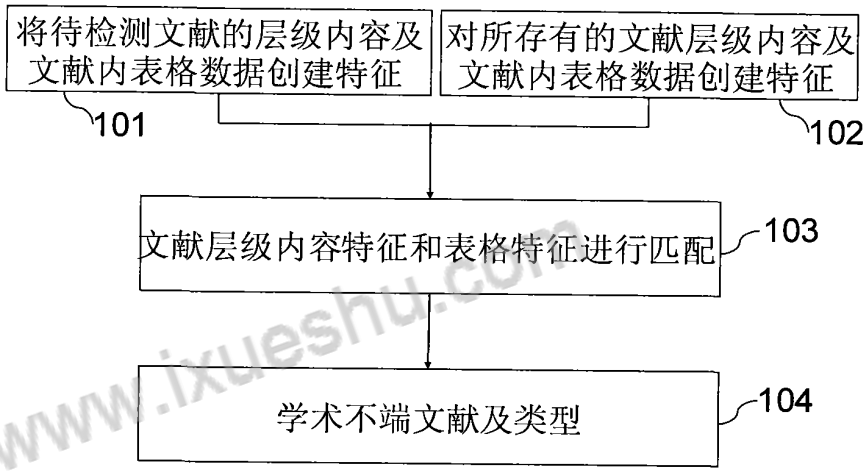


图 2

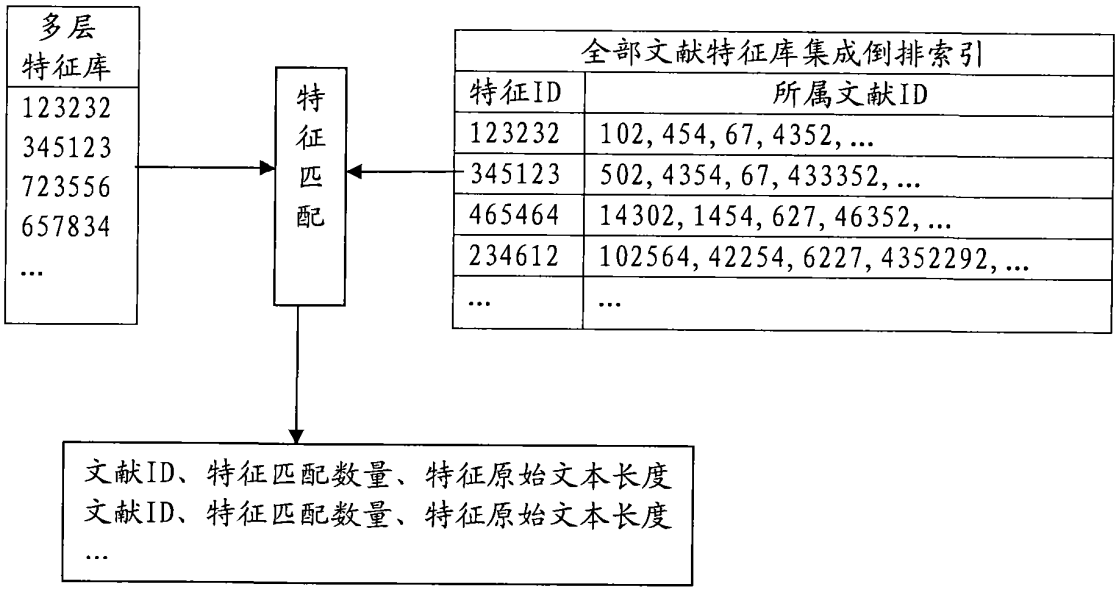


图 3

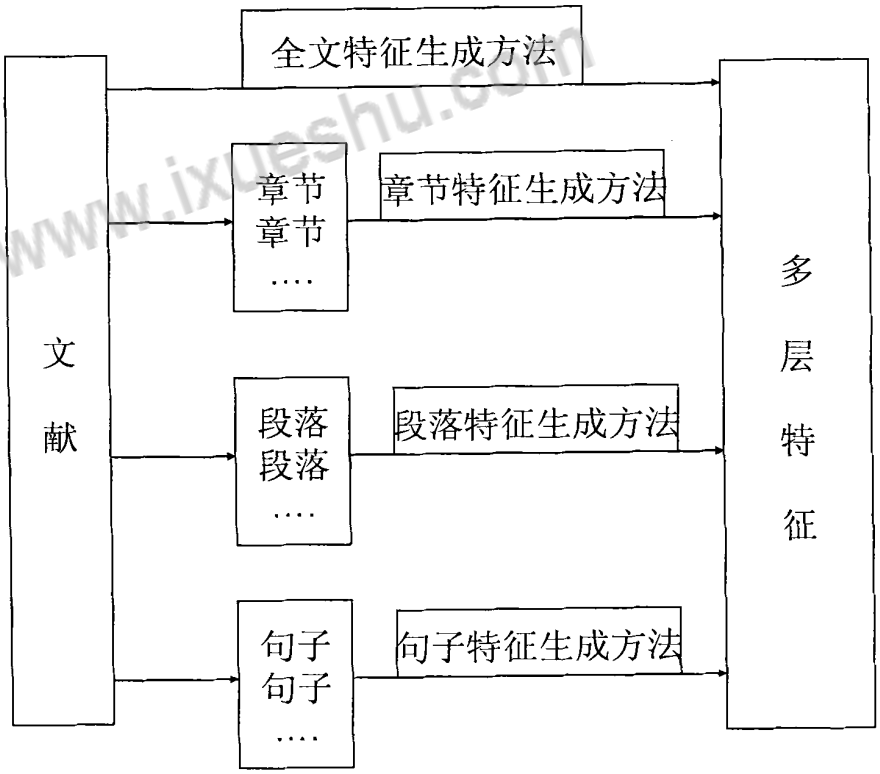


图 4

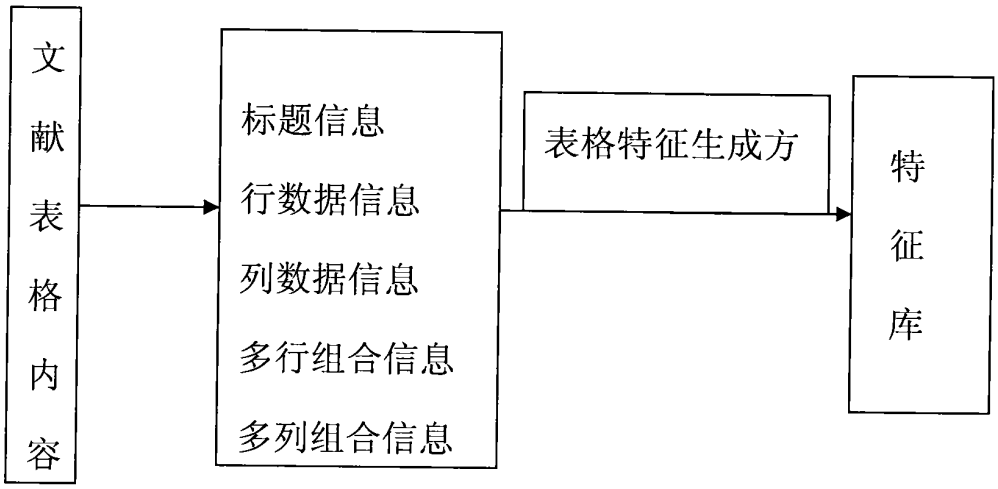


图 5

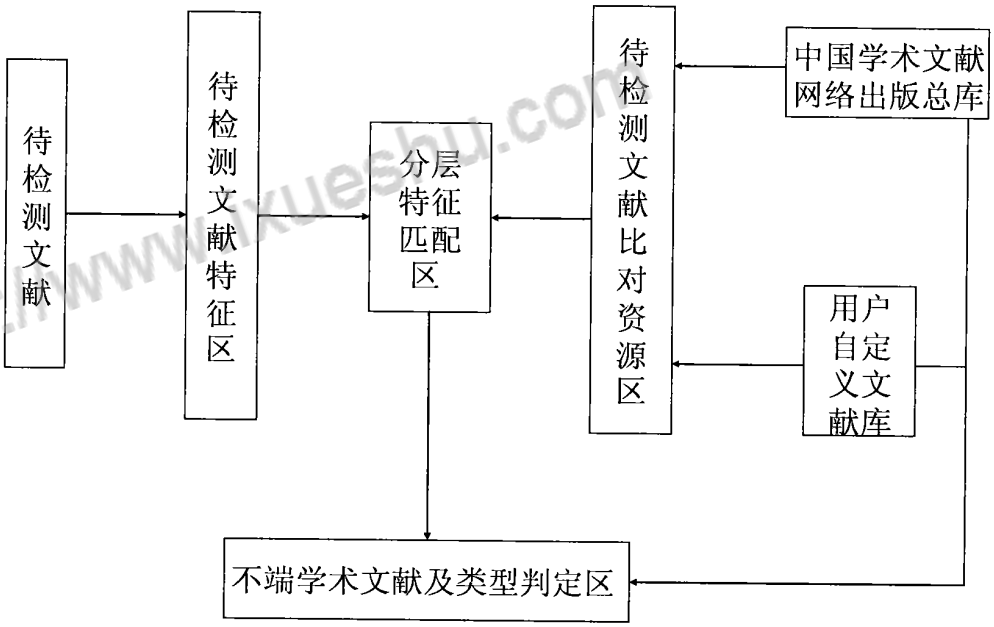


图 6

阅读此文的还阅读了：

26. 本刊启用学术不端文献检测系统(AMLC)

27. 本刊启用学术不端文献检测系统（AMLC）

28. 本刊启用学术不端文献检测系统（AMLC）

29. “学术不端文献检测系统”是治理学术不端的有效手段

30. 本刊启用学术不端文献检测系统

31. 本刊启用“学术不端文献检测系统(SMLC)”

32. 启用学术不端文献(期刊)检测系统公告

33. 启用学术不端文献（期刊）检测系统公告

34. 学术不端文献检测系统数据分析

35. 本刊启用学术不端文献检测系统（AMLC）

36. 学报初审采用“学术不端文献检测系统”

37. 关于加入“学术不端文献检测系统”的启事

38. 关于本刊启用“学术不端文献检测系统”的公告

39. 学术不端文献检测系统

40. 本刊启用学术不端文献检测系统（AMLC）

41. 期刊学术不端文献检测系统误检分析

42. 本刊启用学术不端文献检测系统（AMLC）

43. 对“学术不端文献检测系统”的审稿功能探析

44. 关于本刊启用“学术不端文献检测系统”的公告

45. 启用学术不端文献（期刊）检测系统公告

46. 学术不端文献检测系统应用现状评析

47. 关于加入“学术不端文献检测系统”的启事

48. 学术不端文献检测系统

49. 本刊启用学术不端文献检测系统（AMLC）

50. 本刊启用学术不端文献检测系统(AMLC)