(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 111581928 B
(45)授权公告日 2022.03.01

(21)申请号 202010362633.2

(22)申请日 2020.04.30

(65)同一申请的已公布的文献号
申请公布号 CN 111581928 A

(43)申请公布日 2020.08.25

(73)专利权人 北京理工大学
地址 100081 北京市海淀区中关村南大街5号

(72)发明人 汪雪锋 刘玉琴 刘佳

(74)专利代理机构 北京智行阳光知识产权代理事务所(普通合伙) 11738
代理人 黄锦阳

(51)Int.Cl.
G06F 40/177(2020.01)
G06F 16/28(2019.01)

(56)对比文件
CN 109800397 A,2019.05.24
CN 110400101 A,2019.11.01
US 2004158578 A1,2004.08.12

审查员 蒋娜

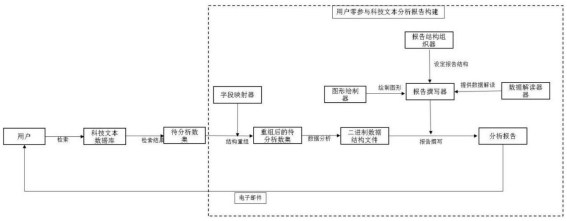
权利要求书4页 说明书10页 附图1页

(54)发明名称

用户零参与的科技文本分析报告自动构建系统及方法

(57)摘要

本发明公开了一种用户零参与的科技文本分析报告自动构建系统及方法,系统包括用于对待分析科技文本进行结构重组的字段映射器、用于存储分析结果的二进制数据结构以及对应的二进制分析结果存储文件、数据解读器、图形绘制器、报告结构组织器和报告撰写器。利用本发明,用户使用科技文本服务商的数据,不需要与分析软件进行交互,不需要进行与分析软件相关的操作,即可获取经过解读的科技文本分析报告。



CN 111581928 B

1.用户零参与的科技文本分析报告自动构建系统,其特征在于,包括用于对待分析科技文本进行结构重组的字段映射器、用于存储分析结果的二进制数据结构以及对应的二进制分析结果存储文件、数据解读器、图形绘制器、报告结构组织器和报告撰写器;

用于对待分析科技文本进行结构重组的字段映射器:所述字段映射器对待分析科技文本按照序号、作者、机构、国家、省份、时间、类别1、类别2、出版物、资助项目、关键词、题目、摘要、全文这11个维度进行重组;

用于存储分析结果的二进制数据结构以及对应的二进制分析结果存储文件;所述用于存储分析结果的二进制数据结构包括:

11个字段数据结构:用于存储作者、机构、国家、省份、时间、类别1、类别2、出版物、资助项目、关键词、主题词对应的科技文本数量,这11个字段定义为基本维度;所述11个字段数据结构存储的内容定义为一维统计结果;

11*11个数据表结构:用于存储11个基本维度两两组合后的科技文本数量,11*11个数据表结构存储的内容定义为二维统计结果;

8个图形数据结构:每个图形数据结构由节点和连线构成,分别用于存储作者、机构、国家、省份、类别1、类别2、关键词、主题词8个维度的共现网络图;每个维度的共现网络图G(V,E)中,节点集合V存储的是对应维度的内容集合,E存储的是各个维度内容在同一个科技文本中出现的次数,作为图形连线;

数据解读器:所述数据解读器用于读入分析结果并对分析结果进行自动解读,然后输出一段文字描述;

一维统计结果为表格类份额数据,数据解读器对于表格类份额数据按照数量和占比进行解读,对于数量和占比排序前N位的数据输出文字描述;

带有时间维度的二维统计结果为表格类趋势数据,数据解读器对于表格类趋势数据具体按照数据的整体趋势、最大值、最小值、增值率进行解读并输出文字描述;

数据解读器对于共现网络图按照关系的强弱进行解读并输出文字描述;

图形绘制器:所述图形绘制器用于读入分析结果并绘制图形;

对于表格类份额数据,采用柱形图进行图形绘制;

对于表格类趋势数据,采用折线图进行图形绘制;

对于共现网络图,采用带文字的球形节点和粗细的连线进行绘制;

报告结构组织器:所述报告结构组织器用于对输出的分析报告的内容和结构进行限定和组织;报告结构组织器定义有描述符,所述描述符用于对分析报告的内容和结构进行组织,对报告中需要连接的二进制数据结构;

报告撰写器:报告撰写器用于按照报告结构组织器的描述符对分析报告进行撰写,遇到对应的描述符,调取所需的二进制数据结构数据,按照描述符的描述进行输出。

2.根据权利要求1所述的系统,其特征在于,主题词是从科技文本的题目、摘要、正文中进行计算机分词的词组。

3.根据权利要求1所述的系统,其特征在于,数据解读器对于表格类份额数据的解读所输出文字描述为“排序前N位的{0}分别为{1}{2}{3}{4}{5},其数据量分别为{6}{7}{8}{9}{10},数量占比分别为{11}{12}{13}{14}{15}”,其中{0}为基本维度中任意一个,{1}-{5}为对应的科技文本数量,{6}-{10}为对应的科技文本数量占比。

4.根据权利要求1所述的系统，其特征在于，数据解读器对于表格类趋势数据的解读所输出文字描述为“整体趋势递增/递减/趋势不明显，XXXX年达到最多，数量为X，XXXX年为最少，为X，增长较为显著的年份包括XXXX、YYYY、ZZZZ”；其中，整体趋势的判断通过计算不同时间段的斜率进行判断，如果斜率为正的情况多于斜率为负数的情况，则为整体趋势递增，反之则为整体趋势递减，如果相等，则为整体趋势不明显；最大值和最小值的判断通过两两数值比较；增长较为显著的年份通过增长率排序，增长率为正且排序前三的年份为增长较为显著的年份。

5.根据权利要求1所述的系统，其特征在于，数据解读器对共现网络图的解读所输出文字描述为“{0}主要分为以下几组：X1、X2、X3…；Y1、Y2、Y3…；Z1、Z2、Z3…，关系较强的分组是第i、j、k…组”；其中{0}为作者、机构、国家、省份、类别1、类别2、关键词、主题词8个维度，分组的判断采用Kmeans聚类将节点划分分组，关系大于关系中位数的分组为关系较强的分组。

6.根据权利要求1所述的系统，其特征在于，描述符分为7个类别，具体包括：

1)参数描述符，基本格式为：

param|id＝;data＝;where＝;type＝;

指明分析报告应该在此输出参数，包括编号，二进制数据结构数据、字段和类型；

2)段落描述符，基本格式为：

paragraph|level＝;linesbefore＝;linespace＝;charactersbefore＝;fontsize＝;fontfamily＝;italic＝;bold＝;align＝;content＝;

指明分析报告应该在此输出一段文字，包括由content决定的段落文字内容及其在大纲级别、段前、断后、行间距、字体、字号、斜体、粗体、对齐上的格式设置；

3)表格描述符，基本格式为：

tablestatic|name＝;row＝;column＝;style＝;data＝;

指明分析报告应在此输出一个表格，包括表格名称、行数、列数、样式、二进制数据结构数据；

4)动态图形描述符，基本格式为：

figuredynamic|name＝;data＝;param＝;

指明分析报告应在此输出一个非网络的图形，包括图形名称、对应的二进制数据结构数据、临时存储路径、参数；

5)网络图描述符，基本格式为：

figurememory|name＝;func＝;params＝;save＝;

指明分析报告应在此输出一个共现网络图，包括名称、步骤S4中的数据解读器、绘制使用的参数、临时存储的路径；

6)横向纵向排版描述符，基本格式为：

segmentpage|orientation＝;

指明分析报告在当前页面排版的纸张方向；

7)目录描述符，基本格式为：

content|type＝;

指明分析报告应在此输出一个目录，目录的类型是全文目录、图目录和/或表目录。

7.一种利用上述任一权利要求所述的系统进行科技文本分析报告自动构建的方法,其特征在于,包括如下步骤:

S1、用户在科技文本数据库中进行检索,将检索得到的对待分析科技文本输入用户零参与的科技文本分析报告自动构建系统中;

S2、字段映射器对待分析科技文本进行结构重组:

字段映射器待分析的科技文本按照序号、作者、机构、国家、省份、时间、类别1、类别2、出版物、资助项目、关键词、题目、摘要、全文进行重组;

S3、按照二进制数据结构中的字段数据结构统计结构重组后的待分析科技文本中,作者、机构、国家、省份、时间、类别1、类别2、出版物、资助项目、关键词、主题词11个基本维度对应的科技文本数量得到一维统计结果;

按照二进制数据结构中的字段数据结构统计结构重组后的待分析科技文本中,作者、机构、国家、省份、时间、类别1、类别2、出版物、资助项目、关键词、主题词11个基本维度两两组合后的科技文本数量,得到二维统计结果;

按照二进制数据结构中的图形数据结构统计结构重组后的待分析科技文本中作者、机构、国家、省份、类别1、类别2、关键词、主题词8个维度的共现网络图;

一维统计结果、二维统计结果和共现网络图存入二进制分析结果存储文件中;

S4、图形绘制器读入步骤S3得到的一维统计结果、二维统计结果和共现网络图并绘制图形,其中:

所述一维统计结果为表格类份额数据,对于表格类份额数据的图形,采用柱形图进行绘制并存储到临时目录;

带有时间维度的二维统计结果为表格类趋势数据,对于表格类趋势数据的图形,采用折线图绘制并存储到临时目录;

对于共现网络图G(V、E)的图形,采用带文字的球形节点和粗细的连线进行绘制并存储到临时目录;

另外,数据解读器读入步骤S3中得到的一维统计结果、二维统计结果和共现网络图并进行自动解读,然后输出一段文字描述;

一维统计结果为表格类份额数据,数据解读器对于表格类份额数据按照数量和占比进行解读,对于数量和占比排序前N位的数据输出文字描述;

带有时间维度的二维统计结果为表格类趋势数据,数据解读器对于表格类趋势数据,按照数据的整体趋势、最大值、最小值、增值率进行解读并输出文字描述;

数据解读器对于共现网络图,按照关系的强弱进行解读并输出文字描述;

S5、报告结构组织器对输出的分析报告的内容和结构进行限定和组织,其中,报告结构组织器中定义的描述符对分析报告的内容和结构进行组织;

报告撰写器按照报告结构组织器的描述符对分析报告进行撰写,遇到对应的描述符,调取相应的二进制数据结构数据,按照描述符的描述进行输出,最终生成所需的科技文本分析报告。

8.根据权利要求7所述的方法,其特征在于,步骤S4中,数据解读器的解读过程如下:

一维统计结果为表格类份额数据,对于表格类份额数据的解读按照数量和占比进行解读,对于排序前N位的数据输出文字描述“排序前N位的{0}分别为{1}{2}{3}{4}{5},其数据

量分别为{6}{7}{8}{9}{10},数量占比分别为{11}{12}{13}{14}{15}”,其中{0}为基本维度中任意一个,{1}-{5}为对应的科技文本数量,{6}-{10}为对应的科技文本数量占比;

带有时间维度的二维统计结果为表格类趋势数据,对于表格类趋势数据,数据解读器按照数据的整体趋势、最大值、最小值、增值率进行解读,输出文字描述为“整体趋势递增/递减/趋势不明显,XXXX年达到最多,数量为X,XXXX年为最少,为X,增长较为显著的年份包括XXXX、YYYY、ZZZZ”;其中,整体趋势的判断通过计算不同时间段的斜率进行判断,如果斜率为正的情况多于斜率为负数的情况,则为整体趋势递增,反之则为整体趋势递减,如果相等,则为整体趋势不明显;最大值和最小值的判断通过两两数值比较;增长较为显著的年份通过增长率排序,增长率为正且排序前三的年份为增长较为显著的年份;

共现网络图的解读输出文字描述为“{0}主要分为以下几组:X1、X2、X3…;Y1、Y2、Y3…;Z1、Z2、Z3…,关系较强的分组是第i、j、k…组”;其中{0}为作者、机构、国家、省份、类别1、类别2、关键词、主题词8个维度,分组的判断采用Kmeans聚类将节点划分分组,关系大于关系中位数的分组为关系较强的分组。

9.根据权利要求7所述的方法,其特征在于,步骤S5中,所述报告结构组织器中:

参数描述符指明分析报告应该在此输出参数,包括编号、二进制数据结构数据、字段和类型;

段落描述符指明分析报告应该在此输出一段文字,包括段落文字内容及其在大纲级别、段前、断后、行间距、字体、字号、斜体、粗体、对齐上的格式设置;

表格描述符指明分析报告应在此输出一个表格,包括表格名称、行数、列数、样式、二进制数据结构数据;

动态图形描述符指明分析报告应在此输出一个非网络的图形,包括图形名称、二进制数据结构中的数据、临时存储路径、参数;

网络图描述符指明分析报告应在此输出一个共现网络图,包括名称、数据解读器、绘制使用的参数、临时存储的路径;

横向纵向排版描述符指明分析报告在当前页面排版的纸张方向;

目录描述符指明分析报告应在此输出一个目录,目录的类型是全文目录、图目录和表目录。

用户零参与的科技文本分析报告自动构建系统及方法

技术领域

[0001] 本发明涉及计算机辅助撰写科技文稿领域，具体涉及一种用户零 参与的科技文本分析报告自动构建系统及方法。

背景技术

[0002] 科学技术迅猛发展、科研难度日趋加大、学科间渗透交叉、研究 者之间协作竞争、文本格式的科技文本资源(如科研论文、专利、科 技报告等)呈爆炸式增长，这些对于科技数据分析人员提出了新的挑 战，即在日新月异的海量科技文本资源中如何迅速提取有价值的科技 信息并尽快做出反应。随着信息技术的飞速发展，一些科技资源提供 服务商提供了软件工具对其科技文本进行分析利用，这些软件工具常 常将分析结果全部或部分呈现给用户，构建一个分析报告，便于用户 快速的了解科技文本所蕴含的深层次信息。

[0003] 这些分析报告的构建方式有两大类：

[0004] 一是用户主动式。科技资源服务商提供科技文本数据和配套的软 件工具，这些软件工具提供固定的分析方法，用户选择某个分析方法 和数据后，形成一个分析结果，再由若干分析结果构成一个分析报告。

[0005] 二是用户被动式。科技资源服务商提供科技文本数据和配套的软 件工具，软件工具基于固定模板构建统计表格和统计图形，配合简单 的文字说明表格和图形中的数据是什么，一次性将其能够提供的分析 内容呈现给用户。

[0006] 其中，上述方式一构建的分析报告的缺点在于，用户需要不断地 与科技资源服务商提供的软件工具进行交互操作，用户需要对服务商 的软件工具较为熟悉，并耗费一定的学习时间和软件工具操作时间。

[0007] 上述方式二构建的分析报告的缺点则在于，每个统计表格和图形 缺少深入的文字解读，仅仅是对统计表格和统计图形内容是什么进行 说明，没有深入解读，特别是一些复杂的分析图形，如机构之间的合 作关系图、技术主题之间的关联关系图等，往往只是包含图形的展示 和图形的标题。

[0008] 以上两种方式构建的分析报告都需要用户进行二次解读、撰写文 字表述和对报告进行组织排版，费时、费力，且深入程度不足。

发明内容

[0009] 针对现有技术的不足，本发明旨在提供一种用户零参与的科技文 本分析报告自动构建系统及方法，用户使用科技文本服务商的数据， 不需要与分析软件进行交互，不需要进行与分析软件相关的操作，即 可获取经过解读的科技文本分析报告。

[0010] 为了实现上述目的，本发明采用如下技术方案：

[0011] 本发明提供一种用户零参与的科技文本分析报告自动构建系统， 包括用于对待分析科技文本进行结构重组的字段映射器、用于存储 分析结果的二进制数据结构以及对应的二进制分析结果存储文件、 数据解读器、图形绘制器、报告结构组织器和报告撰写器；

[0012] 用于对待分析科技文本进行结构重组的字段映射器:所述字段映 射器对待分析科技文本按照序号、作者、机构、国家、省份、时间、 类别1、类别2、出版物、资助项目、关键词、题目、摘要、全文这 11个维度进行重组;

[0013] 用于存储分析结果的二进制数据结构以及对应的二进制分析结 果存储文件;所述用于存储分析结果的二进制数据结构包括:

[0014] 11个字段数据结构:用于存储作者、机构、国家、省份、时间、 类别1、类别2、出版物、资助项目、关键词、主题词对应的科技文 本数量,这11个字段定义为基本维度;所述11个字段数据结构存储 的内容定义为一维统计结果;

[0015] 11*11个数据表结构:用于存储11个基本维度两两组合后的科 技文本数量,11*11个数据表结构存储的内容定义为二维统计结果;

[0016] 8个图形数据结构:每个图形数据结构由节点和连线构成,分别 用于存储作者、机构、国家、省份、类别1、类别2、关键词、主题 词8个维度的共现网络图;每个维度的共现网络图G(V,E)中,节 点集合V存储的是对应维度的内容集合,E存储的是各个维度内容在 同一个科技文本中出现的次数,作为图形连线;

[0017] 数据解读器:所述数据解读器用于读入分析结果并对分析结果进 行自动解读,然后输出一段文字描述;

[0018] 一维统计结果为表格类份额数据,数据解读器对于表格类份额数 据按照数量和占比进行解读,对于数量和占比排序前N位的数据输出 文字描述;

[0019] 带有时间维度的二维统计结果为表格类趋势数据,数据解读器对 于表格类趋势数据具体按照数据的整体趋势、最大值、最小值、增值 率进行解读并输出文字描述;

[0020] 数据解读器对于共现网络图按照关系的强弱进行解读并输出文 字描述;

[0021] 图形绘制器:所述图形绘制器用于读入分析结果并绘制图形;

[0022] 对于表格类份额数据,采用柱形图进行图形绘制;

[0023] 对于表格类趋势数据,采用折线图进行图形绘制;

[0024] 对于共现网络图,采用带文字的球形节点和粗细的连线进行绘制;

[0025] 报告结构组织器:所述报告结构组织器用于对输出的分析报告的 内容和结构进行限定和组织;报告结构组织器定义有描述符,所述描 述符用于对分析报告的内容和结构进行组织,对报告中需要连接的二 进制数据结构;

[0026] 报告撰写器:报告撰写器用于按照报告结构组织器的描述符对分 析报告进行撰写,遇到对应的描述符,调取所需的二进制数据结构数 据,按照描述符的描述进行输出。

[0027] 进一步地,上述系统中,主题词是从科技文本的题目、摘要、正 文中进行计算机分词的词组。

[0028] 进一步地,上述系统中,数据解读器对于表格类份额数据的解读 所输出文字描述为“排序前N位的{0}分别为{1}{2}{3}{4}{5},其数 据量分别为{6}{7}{8}{9}{10},数量占比分别为 {11}{12}{13}{14}{15}”,其中{0}为基本维度中任意一个,{1}-{5} 为对应的科技文本数量,{6}-{10}为对应的科技文本数量占比。

[0029] 进一步地,上述系统中,数据解读器对于表格类趋势数据的解读 所输出文字描述为“整体趋势递增/递减/趋势不明显,XXXX年达到 最多,数量为X,XXXX年为最少,为X,增长较为显著的年份包括XXXX、YYYY、ZZZZ”;其中,整体趋势的判断通过计算不同时间段的斜

率进 行判断,如果斜率为正的情况多于斜率为负数的情况,则为整体趋势 递增,反之则为整体趋势递减,如果相等,则为整体趋势不明显;最 大值和最小值的判断通过两两数值比较;增长较为显著的年份通过增 长率排序,增长率为正且排序前三的年份为增长较为显著的年份。

[0030] 进一步地,上述系统中,数据解读器对共现网络图的解读所输出 文字描述为“{0}主要分为以下几组:X1、X2、X3…;Y1、Y2、Y3…;Z1、 Z2、Z3…,关系较强的分组是第i、j、k…组”;其中{0}为作者、机 构、国家、省份、类别1、类别2、关键词、主题词8个维度,分组 的判断采用Kmeans聚类将节点划分分组,关系大于关系中位数的分 组为关系较强的分组。

[0031] 进一步地,上述系统中,描述符分为7个类别,具体包括:

[0032] 1)参数描述符,基本格式为:

[0033] param|id=;data=;where=;type=;

[0034] 指明分析报告应该在此输出参数,包括编号,二进制数据结构数 据、字段和类型;

[0035] 2)段落描述符,基本格式为:

[0036] paragraph|level=;linesbefore=;linespace=;charactersbefo re=;fontsize=;fontfamily=;italic=;bold=;align=;content=;

[0037] 指明分析报告应该在此输出一段文字,包括由content决定的段 落文字内容及其在大纲级别、段前、断后、行间距、字体、字号、斜 体、粗体、对齐上的格式设置;

[0038] 3)表格描述符,基本格式为:

[0039] tablestatic|name=;row=;column=;style=;data=;

[0040] 指明分析报告应在此输出一个表格,包括表格名称、行数、列数、 样式、二进制数据结构数据;

[0041] 4)动态图形描述符,基本格式为:

[0042] figuredynamic|name=;data=;param=;

[0043] 指明分析报告应在此输出一个非网络的图形,包括图形名称、对 应的二进制数据结构数据、临时存储路径、参数;

[0044] 5)网络图描述符,基本格式为:

[0045] figurememory|name=;func=;params=;save=;

[0046] 指明分析报告应在此输出一个共现网络图,包括名称、步骤S4 中的数据解读器、绘制使用的参数、临时存储的路径;

[0047] 6)横向纵向排版描述符,基本格式为:

[0048] segmentpage|orientation=;

[0049] 指明分析报告在当前页面排版的纸张方向;

[0050] 7)目录描述符,基本格式为:

[0051] content|type=;

[0052] 指明分析报告应在此输出一个目录,目录的类型是全文目录、图 目录和/或表目录。

[0053] 本发明还提供一种利用上述系统进行科技文本分析报告自动构 建的方法,包括如下步骤:

[0054] S1、用户在科技文本数据库中进行检索,将检索得到的对待分析 科技文本输入用

户零参与的科技文本分析报告自动构建系统中；

[0055] S2、字段映射器对待分析科技文本进行结构重组：

[0056] 字段映射器待分析的科技文本按照序号、作者、机构、国家、省 份、时间、类别1、类别2、出版物、资助项目、关键词、题目、摘 要、全文进行重组；

[0057] S3、按照二进制数据结构中的字段数据结构统计结构重组后的待 分析科技文本中，作者、机构、国家、省份、时间、类别1、类别2、 出版物、资助项目、关键词、主题词11个基本维度对应的科技文本 数量得到一维统计结果；

[0058] 按照二进制数据结构中的字段数据结构统计结构重组后的待分 析科技文本中，作者、机构、国家、省份、时间、类别1、类别2、 出版物、资助项目、关键词、主题词11个基本维度两两组合后的科 技文本数量，得到二维统计结果；

[0059] 按照二进制数据结构中的图形数据结构统计结构重组后的待分 析科技文本中作者、机构、国家、省份、类别1、类别2、关键词、 主题词8个维度的共现网络图；

[0060] 一维统计结果、二维统计结果和共现网络图存入二进制分析结果 存储文件中；

[0061] S4、图形绘制器读入步骤S3得到的一维统计结果、二维统计结 果和共现网络图并绘制图形，其中：

[0062] 所述一维统计结果为表格类份额数据，对于表格类份额数据的图 形，采用柱形图进行绘制并存储到临时目录；

[0063] 带有时间维度的二维统计结果为表格类趋势数据，对于表格类趋 势数据的图形，采用折线图绘制并存储到临时目录；

[0064] 对于共现网络图G(V、E)的图形，采用带文字的球形节点和粗 细的连线进行绘制并存储到临时目录；

[0065] 另外，数据解读器读入步骤S3中得到的一维统计结果、二维统 计结果和共现网络图并进行自动解读，然后输出一段文字描述；

[0066] 一维统计结果为表格类份额数据，数据解读器对于表格类份额数 据按照数量和占比进行解读，对于数量和占比排序前N位的数据输出 文字描述；

[0067] 带有时间维度的二维统计结果为表格类趋势数据，数据解读器对 于表格类趋势数据，按照数据的整体趋势、最大值、最小值、增值率 进行解读并输出文字描述；

[0068] 数据解读器对于共现网络图，按照关系的强弱进行解读并输出文 字描述；

[0069] S5、报告结构组织器对输出的分析报告的内容和结构进行限定和 组织，其中，报告结构组织器中定义的描述符对分析报告的内容和结 构进行组织；

[0070] 报告撰写器按照报告结构组织器的描述符对分析报告进行撰写，遇到对应的描述符，调取相应的二进制数据结构数据，按照描述符的 描述进行输出，最终生成所需的科技文本分析报告。

[0071] 进一步地，上述方法的步骤S4中，数据解读器的解读过程如下：

[0072] 一维统计结果为表格类份额数据，对于表格类份额数据的解读按 照数量和占比进行解读，对于排序前N位的数据输出文字描述“排序 前N位的{0}分别为{1}{2}{3}{4}{5}，其数据量分别为 {6}{7}{8}{9}{10}，数量占比分别为{11}{12}{13}{14}{15}”，其中{0} 为基本维度中任意一个，{1}-{5}为对应的科技文本数量，{6}-{10} 为对应的科技文本数量占比；

[0073] 带有时间维度的二维统计结果为表格类趋势数据,对于表格类趋 势数据,数据解读器按照数据的整体趋势、最大值、最小值、增值率 进行解读,输出文字描述为“整体趋势递增/递减/趋势不明显,XXXX 年达到最多,数量为X,XXXX年为最少,为X,增长较为显著的年份 包括XXXX、YYYY、ZZZZ”;其中,整体趋势的判断通过计算不同时间 段的斜率进行判断,如果斜率为正的情况多于斜率为负数的情况,则 为整体趋势递增,反之则为整体趋势递减,如果相等,则为整体趋势 不明显;最大值和最小值的判断通过两两数值比较;增长较为显著的 年份通过增长率排序,增长率为正且排序前三的年份为增长较为显著 的年份;

[0074] 共现网络图的解读输出文字描述为“{0}主要分为以下几组:X1、X2、X3…;Y1、Y2、Y3…;Z1、Z2、Z3…,关系较强的分组是第i、j、k…组”;其中{0}为作者、机构、国家、省份、类别1、类别2、关键 词、主题词8个维度,分组的判断采用Kmeans聚类将节点划分分组,关系大于关系中位数的分组为关系较强的分组。

[0075] 进一步地,上述方法的步骤S5中,所述报告结构组织器中:

[0076] 参数描述符指明分析报告应该在此输出参数,包括编号、二进制 数据结构数据、字段和类型;

[0077] 段落描述符指明分析报告应该在此输出一段文字,包括段落文字 内容及其在大纲级别、段前、断后、行间距、字体、字号、斜体、粗 体、对齐上的格式设置;

[0078] 表格描述符指明分析报告应在此输出一个表格,包括表格名称、 行数、列数、样式、二进制数据结构数据;

[0079] 动态图形描述符指明分析报告应在此输出一个非网络的图形,包 括图形名称、二进制数据结构中的数据、临时存储路径、参数;

[0080] 网络图描述符指明分析报告应在此输出一个共现网络图,包括名 称、数据解读器、绘制使用的参数、临时存储的路径;

[0081] 横向纵向排版描述符指明分析报告在当前页面排版的纸张方向;

[0082] 目录描述符指明分析报告应在此输出一个目录,目录的类型是全 文目录、图目录和表目录。

附图说明

[0083] 图1为本发明实施例2的流程示意图。

具体实施方式

[0084] 以下将结合附图对本发明作进一步的描述,需要说明的是,本实 施例以本技术方案为前提,给出了详细的实施方式和具体的操作过 程,但本发明的保护范围并不限于本实施例。

[0085] 实施例1

[0086] 本实施例提供一种用户零参与的科技文本分析报告自动构建系 统,包括:

[0087] 用于对待分析科技文本进行结构重组的字段映射器:所述字段映 射器对待分析科技文本按照序号、作者、机构、国家、省份、时间、 类别1、类别2、出版物、资助项目、关键词、题目、摘要、全文这 11维度进行重组;所述字段映射器如下所示:

```
<?xml version="1.0" encoding="utf-8"?>
<Config>
  <FieldMap>
    <Source>X</Source>
    <ID>X</ID>
    <Keyword>X</Keyword>
[0088]
    <Abstract>X</Abstract>
    <Authors>X</Authors>
    <Affiliation>X</Affiliation>
    <Class1></Class1>
    <Class2>X</Class2>
    <Countries>X</Countries>
    <Provinces>X</Provinces>
    <Founders>X</Founders>
    <Publication>X</Publication>
    <Description>X</Description>
[0089]
    <Time>X</Time>
    <Title>X</Title>
  </FieldMap>
</Config>
```

[0090] 通过字段映射器对待分析科技文本进行结构重组的目的是规范 化科技文本数据,方便分析。

[0091] 用于存储分析结果的二进制数据结构以及对应的二进制分析结 果存储文件:

[0092] 所述用于存储分析结果的二进制数据结构包括:

[0093] 11个字段数据结构:用于存储作者、机构、国家、省份、时间、 类别1、类别2、出版

物、资助项目、关键词、主题词对应的科技文 本数量，这11个字段定义为基本维度，其中主题词是从题目、摘要、 正文中进行计算机分词的词组；所述11个字段数据结构存储的内容定义为一维统计结果；

[0094] 11*11个数据表结构：用于存储11个基本维度两两组合后的科 技文本数量，11*11个数据表结构存储的内容定义为二维统计结果；

[0095] 8个图形数据结构：每个图形数据结构由节点和连线构成，分别 用于存储作者、机构、国家、省份、类别1、类别2、关键词、主题 词8个维度的共现网络图；每个维度的共现网络图G(V,E)中，节 点集合V存储的是对应维度的内容集合，E存储的是各个维度内容在 同一个科技文本中出现的次数，作为图形连线；

[0096] 以作者这个维度为例，该维度的共现网络图G(V,E)中，节点 集合V存储的是作者姓名的集合，作为图形节点，E存储的是作者之 间共同在同一个科技文本中出现的次数，作为图形连线。

[0097] 数据解读器：所述数据解读器用于读入分析结果并对分析结果进 行自动解读，然后输出一段文字描述；从而实现分析结果解读的用户“零参与”。

[0098] 所述数据解读器解读表格类份额数据、表格类趋势数据和网络数 据：

[0099] 1)表格类份额数据的解读：

[0100] 所述一维统计结果为表格类份额数据，对于表格类份额数据的解 读按照数量和占比进行解读，对于排序前N位的数据输出文字描述 “排序前N位的{0}分别为{1}{2}{3}{4}{5}，其数据量分别为 {6}{7}{8}{9}{10}，数量占比分别为{11}{12}{13}{14}{15}”，其中{0} 为基本维度中任意一个，{1}-{5}为对应的科技文本数量，{6}-{10} 为对应的科技文本数量占比；

[0101] 2)表格类趋势数据的解读：

[0102] 本实施例中，带有时间维度的二维统计结果为表格类趋势数据， 对于表格类趋势数据的解读，按照整体趋势、最大值、最小值、增值 率进行解读，输出文字描述为“整体趋势递增/递减/趋势不明显，XXXX 年达到最多，数量为X，XXXX年为最少，为X，增长较为显著的年份 包括XXXX、YYYY、ZZZZ”；其中，整体趋势的判断通过计算不同时间 段的斜率进行判断，如果斜率为正的情况多于斜率为负数的情况，则 为整体趋势递增，反之则为整体趋势递减，如果相等，则为整体趋势 不明显；最大值和最小值的判断通过两两数值比较；增长较为显著的 年份通过增长率排序，增长率为正且排序前三的年份为增长较为显著 的年份。

[0103] 3)共现网络图G(V,E)的解读：

[0104] 共现网络图的解读输出文字描述为“{0}主要分为以下几组：X1、X2、X3…；Y1、Y2、Y3…；Z1、Z2、Z3…，关系较强的分组是第i、j、k…组”；其中{0}为作者、机构、国家、省份、类别1、类别2、关键 词、主题词8个维度，分组的判断采用Kmeans聚类将节点划分分组， 关系大于关系中位数的分组为关系较强的分组。

[0105] 图形绘制器：所述图形绘制器用于读入分析结果并绘制图形，实 现图形绘制的用户“零参与”；

[0106] 图形绘制器绘制的图形分为三类，具体为：

[0107] 1)表格类份额数据图形：表格类份额数据采用柱形图进行绘制， 存储到临时目录，用于报告结构组织器使用。

[0108] 2)表格类趋势数据图形:表格类趋势数据采用折线图绘制,存 储到临时目录,用于报告结构组织器使用。

[0109] 3)共现网络图G(V、E)的图形绘制:共现网络图的绘制采用 带文字的球形节点和粗细的连线进行绘制,节点的布局采用弹性模型Spring-EmbeddedModel算法[1],和改进弹性模型 Fruchterman-Reingoldlayout算法[2]进行,每种算法分别循环设定次 数,输出共现网络图,存储到临时目录,用于报告结构组织器使用。

[0110] 报告结构组织器:所述报告结构组织器用于对需要输出的分析报 告的结构进行限定和组织;实现分析报告结构组织的用户“零参与”。

[0111] 报告结构组织器中定义有描述符,所述描述符用于对分析报告的 结构进行组织。

[0112] 描述符基本构成为“描述类型|前置条件、描述内容格式限定、 参数设定、数据设定”,描述符分为7个类别,具体包括:

[0113] 1)参数描述符,基本格式为:

[0114] param|id=;data=;where=;type=;

[0115] 指明分析报告应该在此输出参数,包括编号,二进制数据结构数 据、字段和类型;

[0116] 所述序号按照数字排列的序号,1,2,3…。

[0117] 2)段落描述符,基本格式为:

[0118] paragraph|level=;linesbefore=;linespace=;charactersbefo re=;fontsize=;fontfamily=;italic=;bold=;align=;content=;

[0119] 指明分析报告应该在此输出一段文字,包括由content决定的段 落文字内容及其在大纲级别、段前、断后、行间距、字体、字号、斜 体、粗体、对齐上的格式设置;

[0120] 3)表格描述符,基本格式为:

[0121] tablestatic|name=;row=;column=;style=;data=;

[0122] 指明分析报告应在此输出一个表格,包括名称、行数、列数、样 式、对应的二进制数据结构数据。

[0123] 4)动态图形描述符,基本格式为:

[0124] figuredynamic|name=;data=;param=;save=;

[0125] 指明分析报告应在此输出一个非网络的图形,包括名称、对应的 二进制数据结构中的数据、参数、临时存储路径。

[0126] 5)网络图描述符,基本格式为:

[0127] figurememory|name=;func=;params=;save=;

[0128] 指明分析报告应在此输出一个共现网络图,包括名称、步骤S4 中的数据解读器、绘制使用的参数、临时存储的路径。

[0129] 6)横向纵向排版描述符,基本格式为:

[0130] segmentpage|orientation=;

[0131] 指明分析报告在当前页面排版的纸张方向(横向或纵向);

[0132] 7)目录描述符,基本格式为:

[0133] content|type=;

[0134] 指明分析报告应在此输出一个目录,目录的类型是全文目录、图 目录和表目录;

[0135] 报告撰写器:报告撰写器用于按照报告结构组织器的描述符对分 析报告进行撰

写,遇到对应的描述符,调取二进制数据结构数据,按 照描述符的描述进行输出。

[0136] 实施例2

[0137] 本实施例提供一种利用实施例1所述系统进行科技文本分析报告 自动构建的方法,如图1所示,包括如下步骤:

[0138] S1、用户在科技文本数据库中进行检索,将检索得到的对待分析 科技文本输入用户零参与的科技文本分析报告自动构建系统中;

[0139] S2、字段映射器对待分析科技文本进行结构重组:

[0140] 字段映射器待分析的科技文本按照序号、作者、机构、国家、省 份、时间、类别1、类别2、出版物、资助项目、关键词、题目、摘 要、全文进行重组;

[0141] S3、按照二进制数据结构中的字段数据结构统计结构重组后的待 分析科技文本中,作者、机构、国家、省份、时间、类别1、类别2、 出版物、资助项目、关键词、主题词11个基本维度对应的科技文本 数量得到一维统计结果;

[0142] 按照二进制数据结构中的字段数据结构统计结构重组后的待分 析科技文本中,作者、机构、国家、省份、时间、类别1、类别2、 出版物、资助项目、关键词、主题词11个基本维度两两组合后的科 技文本数量,得到二维统计结果;

[0143] 按照二进制数据结构中的图形数据结构统计结构重组后的待分 析科技文本中作者、机构、国家、省份、类别1、类别2、关键词、 主题词8个维度的共现网络图;

[0144] 一维统计结果、二维统计结果和共现网络图存入二进制分析结果 存储文件中;

[0145] S4、图形绘制器读入步骤S3得到的一维统计结果、二维统计结 果和共现网络图并绘制图形,其中:

[0146] 所述一维统计结果为表格类份额数据,对于表格类份额数据的图 形,采用柱形图进行绘制并存储到临时目录;

[0147] 带有时间维度的二维统计结果为表格类趋势数据,对于表格类趋 势数据的图形,采用折线图绘制并存储到临时目录;

[0148] 对于共现网络图G(V、E)的图形,采用带文字的球形节点和粗 细的连线进行绘制,节点的布局采用弹性模型Spring-Embedded Model算法[1],和改进弹性模型Fruchterman-Reingoldlayout算法[2]进行,每种算法分别循环设定次数,输出共现网络图并存储到临时目 录;

[0149] 另外,数据解读器读入步骤S3中得到的一维统计结果、二维统 计结果和共现网络图的数据并进行自动解读,然后输出一段文字描述; 其中:

[0150] 所述一维统计结果为表格类份额数据,对于表格类份额数据的解 读按照数量和占比进行解读,对于排序前N位的数据输出文字描述 “排序前N位的{0}分别为{1}{2}{3}{4}{5},其数据量分别为 {6}{7}{8}{9}{10},数量占比分别为{11}{12}{13}{14}{15}”,其中{0} 为基本维度中任意一个,{1}-{5}为对应的科技文本数量,{6}-{10} 为对应的科技文本数量占比;

[0151] 带有时间维度的二维统计结果为表格类趋势数据,对于表格类趋 势数据的解读,数据解读器按照整体趋势、最大值、最小值、增值率 进行解读,输出文字描述为“整体趋势递增/递减/趋势不明显,XXXX 年达到最多,数量为X,XXXX年为最少,为X,增长较为显著的年份 包括XXXX、YYYY、ZZZZ”;其中,整体趋势的判断通过计算不同时间 段的斜率进行判

断,斜率计算公式为:

[0152] 斜率＝(Y(t)-Y(t-1))/(X(t)-X(t-1));

[0153] t为时间中的年份数,Y为t下的数据值,(X(t)-X(t-1))为 时间跨度,可限定为1年。

[0154] 如果斜率为正的情况多于斜率为负数的情况,则为整体趋势递增,反之则为整体趋势递减,如果相等,则为整体趋势不明显;最大值和 最小值的判断通过两两数值比较;增长较为显著的年份通过增长率排 序,增长率为正且排序前三的年份为增长较为显著的年份。

[0155] 共现网络图的解读输出文字描述为“{0}主要分为以下几组:X1、X2、X3…;Y1、Y2、Y3…;Z1、Z2、Z3…,关系较强的分组是第i、j、k…组”;其中{0}为作者、机构、国家、省份、类别1、类别2、关键 词、主题词8个维度,分组的判断采用Kmeans聚类将节点划分分组,关系大于关系中位数的分组为关系较强的分组;

[0156] S5、报告结构组织器对输出的分析报告的内容和结构进行限定和 组织,其中,报告结构组织器中定义的描述符对分析报告的结构进行 组织,对报告中需要连接的步骤S2中的数据结构。

[0157] 参数描述符指明分析报告应该在此输出参数,包括参数的编号是 什么,参与应用二进制数据结构中的是哪个数据、哪个字段,类型是 什么;

[0158] 段落描述符指明分析报告应该在此输出一段文字,包括由 content决定的段落文字内容及其在大纲级别、段前、断后、行间距、 字体、字号、斜体、粗体、对齐上的格式设置;

[0159] 表格描述符指明分析报告应在此输出一个表格,包括名称、行数、 列数、样式、对应的二进制数据结构数据。

[0160] 动态图形描述符指明分析报告应在此输出一个非网络的图形,包 括名称、对应的二进制数据结构中的数据、临时存储路径、参数。

[0161] 网络图描述符指明分析报告应在此输出一个共现网络图,包括名 称、数据解读器、绘制使用的参数、临时存储的路径;

[0162] 横向纵向排版描述符指明分析报告在当前页面排版的纸张方向 (横向或纵向);

[0163] 目录描述符指明分析报告应在此输出一个目录,目录的类型是全 文目录、图目录和表目录;

[0164] 报告撰写器按照报告结构组织器的描述符对分析报告进行撰写,遇到对应的描述符,调取二进制数据结构数据,按照描述符的描述进 行输出,最终生成所需的科技文本分析报告。

[0165] 具体地,可以设置在科技文本分析报告撰写完成后邮件发送到设 定好的电子邮箱。

[0166] 对于本领域的技术人员来说,可以根据以上的技术方案和构思,给出各种相应的改变和变形,而所有的这些改变和变形,都应该包括 在本发明权利要求的保护范围之内。

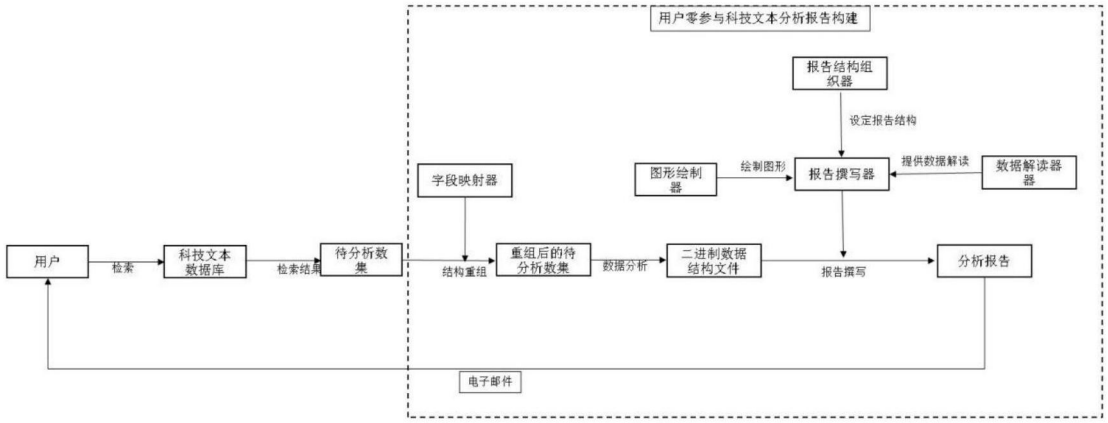

用户零参与科技文本分析报告构建
报告结构组织器
设定报告结构
字段映射器
图形绘制器
绘制图形
报告撰写器
提供数据解读
数据解读器
用户
检索
科技文本数据库
检索结果
待分析数集
结构重组
重组后的待分析数集
数据分析
二进制数据结构文件
报告撰写
分析报告
电子邮件

图1