



(12) 发明专利

(10) 授权公告号 CN 112908487 B

(45) 授权公告日 2023. 09. 22

(21) 申请号 202110418664.X

EP 0747836 A1, 1996.12.11

(22) 申请日 2021.04.19

US 2019286741 A1, 2019.09.19

(65) 同一申请的已公布的文献号
申请公布号 CN 112908487 A

RU 2592396 C1, 2016.07.20

US 2005010863 A1, 2005.01.13

US 2020176098 A1, 2020.06.04

(43) 申请公布日 2021.06.04

US 2016125169 A1, 2016.05.05

(73) 专利权人 中国医学科学院医学信息研究所
地址 100020 北京市朝阳区雅宝路3号

US 2018121618 A1, 2018.05.03

KR 20210040862 A, 2021.04.14

US 2020152336 A1, 2020.05.14

(72) 发明人 吴思竹 崔佳伟 修晓蕾 钱庆

US 2020027545 A1, 2020.01.23

(74) 专利代理机构 北京集佳知识产权代理有限公司 11227

US 2008071519 A1, 2008.03.20

US 2020226164 A1, 2020.07.16

专利代理师 周初冬

US 2020394365 A1, 2020.12.17

(51) Int. Cl.

EP 3719805 A1, 2020.10.07

G16H 70/00 (2018.01)

US 2016148096 A1, 2016.05.26

G06F 40/211 (2020.01)

CN 108491487 A, 2018.09.04

G06F 40/284 (2020.01)

WO 2017163346 A1, 2017.09.28

G06F 40/289 (2020.01)

US 2018365228 A1, 2018.12.20

G06F 40/30 (2020.01)

US 2014129246 A1, 2014.05.08

G06F 18/22 (2023.01)

CN 111145052 A, 2020.05.12

(续)

(56) 对比文件

审查员 高海琪

US 2019236102 A1, 2019.08.01

权利要求书4页 说明书22页 附图7页

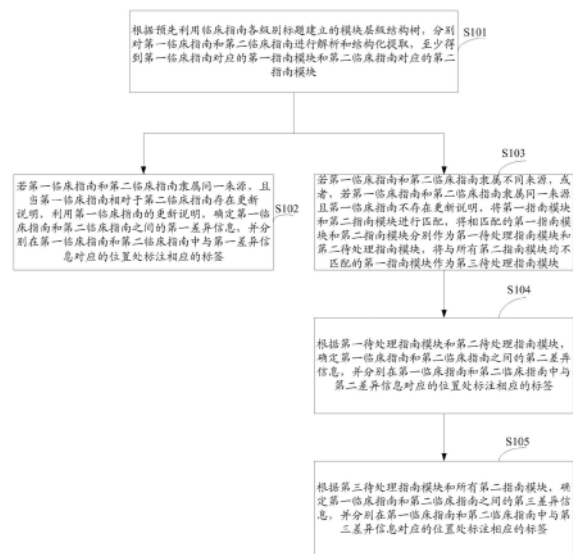
(54) 发明名称

一种临床指南更新内容的自动识别方法及系统

(57) 摘要

本发明提供了一种临床指南更新内容的自动识别方法及系统,该方法为:根据预先利用临床指南各级别标题建立的模块层级结构树,分别对第一临床指南和第二临床指南进行解析和结构化提取,至少得到第一临床指南对应的第一指南模块和第二临床指南对应的第二指南模块;确定第一指南模块和第二指南模块之间的差异信息,并分别在第一临床指南和第二临床指南中与该差异信息对应的位置处标注相应的标签,不需要人工查阅需要比对的两个临床指南来找到不同临床指南间的差异和变化,提高确定不同临床指南间差异和变化情况的效率和准确度。

CN 112908487 B



[转续页]

[接上页]

(56) 对比文件

US 2021012896 A1,2021.01.14

WO 2020084734 A1,2020.04.30

CN 110008304 A,2019.07.12

US 2013054612 A1,2013.02.28

JP 2009211480 A,2009.09.17

WO 2011107893 A1,2011.09.09

陶兴.多源学术新媒体用户生成内容的知识聚合研究.《信息科技》.2020,(第08期),I141-6.

robin wm vermooij,et al.guidance for updating clinical practice guidelines:

a systematic review of methodological handbooks.《implementation science》.2014,(第3期),1-9.

qian qing ,et al.structuring the chinese disjointed literature-based knowledge discovery system :the key technologies to success.《journal of information science》.2012,第38卷(第6期),532-539.

崔佳伟等.临床指南结构化研究.《中华医学图书情报杂志》.2020,第29卷(第1期),35-41.

1. 一种临床指南更新内容的自动识别方法,其特征在于,所述方法包括:

根据预先利用临床指南各级别标题建立的模块层级结构树,分别对第一临床指南和第二临床指南进行解析和结构化提取,至少得到所述第一临床指南对应的第一指南模块和所述第二临床指南对应的第二指南模块,所述第一指南模块为所述第一临床指南中最小级别标题所包含的文本内容,所述第二指南模块为所述第二临床指南中最小级别标题所包含的文本内容;

若所述第一临床指南和所述第二临床指南隶属同一来源,且当所述第一临床指南相对于所述第二临床指南存在更新说明,利用所述第一临床指南的所述更新说明,确定所述第一临床指南和所述第二临床指南之间的第一差异信息,并分别在所述第一临床指南和所述第二临床指南中与所述第一差异信息对应的位置处标注相应的标签,所述标签为新增标签、删除标签或修改标签;若所述第一差异信息为所述第一临床指南相对于所述第二临床指南中的修改部分,则分别在所述第一临床指南和所述第二临床指南中与修改部分对应的位置处标注修改标签;若所述第一差异信息为所述第一临床指南相对于所述第二临床指南中的新增部分,则分别在所述第一临床指南和所述第二临床指南中与新增部分对应的位置处标注新增标签;若所述第一差异信息为所述第一临床指南相对于所述第二临床指南中的删除部分,则分别在所述第一临床指南和所述第二临床指南中与修改部分对应的位置处标注删除标签;

若所述第一临床指南和所述第二临床指南隶属不同来源,或者,若所述第一临床指南和所述第二临床指南隶属同一来源且所述第一临床指南不存在所述更新说明,将所述第一指南模块和所述第二指南模块进行匹配,将相匹配的所述第一指南模块和所述第二指南模块分别作为第一待处理指南模块和第二待处理指南模块,将与所有所述第二指南模块均不匹配的所述第一指南模块作为第三待处理指南模块;

根据所述第一待处理指南模块和所述第二待处理指南模块,确定所述第一临床指南和所述第二临床指南之间的第二差异信息,并分别在所述第一临床指南和所述第二临床指南中与所述第二差异信息对应的位置处标注相应的标签;

其中,根据所述第三待处理指南模块和所有所述第二指南模块,确定所述第一临床指南和所述第二临床指南之间的第三差异信息,并分别在所述第一临床指南和所述第二临床指南中与所述第三差异信息对应的位置处标注相应的标签;

将所述第一指南模块和所述第二指南模块进行匹配的过程,包括:

针对每一个第一指南模块,利用预设的深度语义匹配模型,确定所述第一指南模块的标题和每个所述第二指南模块的标题之间的标题相似度;

针对每一个第一指南模块,若所有所述标题相似度均小于标题相似度阈值,确定所述第一指南模块与所有所述第二指南模块均不匹配,若至少一个所述标题相似度大于等于所述标题相似度阈值,确定所述第一指南模块与最大所述标题相似度对应的所述第二指南模块相匹配;

所述根据所述第一待处理指南模块和所述第二待处理指南模块,确定所述第一临床指南和所述第二临床指南之间的第二差异信息,并分别在所述第一临床指南和所述第二临床指南中与所述第二差异信息对应的位置处标注相应的标签,包括:

分别对所述第一待处理指南模块和所述第二待处理指南模块进行分句处理,得到所述

第一待处理指南模块对应的多个第一句子,及得到所述第二待处理指南模块对应的多个第二句子;

对于所述第一待处理指南模块的第 m 个第一句子,计算所述第 m 个第一句子与所述第二待处理指南模块的 H 个第二句子之间的句子相似度, m 为大于等于1小于等于 x 的整数, x 为所述第一待处理指南模块所包含的第一句子的总个数, m 从1开始且递增1, H 为大于等于1小于等于 y 的整数, y 为所述第二待处理指南模块所包含的第二句子的总个数;

若所述第 m 个第一句子和第 n 个第二句子之间的句子相似度等于1,确定所述第 m 个第一句子和所述第 n 个第二句子相同,不执行标注处理, n 为大于等于1小于等于 y 的整数;

若所述第 m 个第一句子和所述第 n 个第二句子之间的句子相似度大于等于句子相似度阈值且小于1,在所述第一临床指南中与所述第 m 个第一句子对应的位置处标注修改标签,及在所述第二临床指南中与所述第 n 个第二句子对应的位置处标注修改标签,当 n 大于 m 时,确定所述第二待处理指南模块中位于所述第 n 个第二句子之前的,与前 m 个所述第一句子的句子相似度均小于所述句子相似度阈值且未进行标注处理的第三句子,在所述第二临床指南中与所述第三句子对应的位置处标注删除标签;

若所述第 m 个第一句子与 H 个所述第二句子之间的句子相似度均小于所述句子相似度阈值,在所述第一临床指南中与所述第 m 个第一句子对应的位置处标注新增标签;

所述根据所述第三待处理指南模块和所有所述第二指南模块,确定所述第一临床指南和所述第二临床指南之间的第三差异信息,并分别在所述第一临床指南和所述第二临床指南中与所述第三差异信息对应的位置处标注相应的标签,包括:

计算所述第三待处理指南模块中首个句子的前 $P\%$ 内容与每个所述第二指南模块的多个第二句子之间的首句相似度;

若至少一个所述首句相似度大于首句相似度阈值,确定所述第三待处理指南模块与最大所述首句相似度对应的所述第二指南模块相匹配;

从所述第三待处理指南模块的首个句子的前 $P\%$ 内容处开始,将位于其之后的所述第一临床指南中已有的标签变更为修改标签;

从与所述第三待处理指南模块相匹配的第二指南模块中最大所述首句相似度对应的第二句子处开始,将位于其之后的所述第二临床指南中已有的标签变更为修改标签。

2. 根据权利要求1所述的方法,其特征在于,所述分别对第一临床指南和第二临床指南进行解析和结构化提取,至少得到所述第一临床指南对应的第一指南模块和所述第二临床指南对应的第二指南模块之后,还包括:

分别对所述第一指南模块和所述第二指南模块进行预处理,并分别提取进行预处理后的所述第一指南模块和所述第二指南模块中的知识特征。

3. 根据权利要求2所述的方法,其特征在于,所述在所述第一临床指南中与所述第 m 个第一句子对应的位置处标注修改标签,及在所述第二临床指南中与所述第 n 个第二句子对应的位置处标注修改标签之后,还包括:

比对所述第 m 个第一句子和所述第 n 个第二句子内的所述知识特征之间的差异,得到知识特征差异信息;

分别在所述第 m 个第一句子中和所述第 n 个第二句子中与所述知识特征差异信息对应的位置处标注相应的标签。

4. 根据权利要求1~2中任一所述的方法,其特征在于,还包括:
分别利用不同的展示形式,展示不同类别的所述标签。

5. 根据权利要求2所述的方法,其特征在于,分别对第一临床指南和第二临床指南进行解析和结构化提取,至少得到所述第一临床指南对应的第一指南模块和所述第二临床指南对应的第二指南模块之后,还包括:

对所述第一指南模块和所述第二指南模块进行规范化处理。

6. 根据权利要求1所述的方法,其特征在于,分别对第一临床指南和第二临床指南进行解析和结构化提取,至少得到所述第一临床指南对应的第一指南模块和所述第二临床指南对应的第二指南模块之后,还包括:

将所述第一指南模块和所述第二指南模块存储至数据库中,及将所有所述第一指南模块之间的层级关系存储至所述数据库中,及将所有所述第二指南模块之间的层级关系存储至所述数据库中。

7. 一种临床指南更新内容的自动识别系统,其特征在于,所述自动识别系统用于执行如权利要求1~6任一项所述的临床指南更新内容自动识别方法,所述系统包括:

解析单元,用于根据预先利用临床指南各级别标题建立的模块层级结构树,分别对第一临床指南和第二临床指南进行解析和结构化提取,至少得到所述第一临床指南对应的第一指南模块和所述第二临床指南对应的第二指南模块,所述第一指南模块为所述第一临床指南中最小级别标题所包含的文本内容,所述第二指南模块为所述第二临床指南中最小级别标题所包含的文本内容;

第一处理单元,用于若所述第一临床指南和所述第二临床指南隶属同一来源,且当所述第一临床指南相对于所述第二临床指南存在更新说明,利用所述第一临床指南的所述更新说明,确定所述第一临床指南和所述第二临床指南之间的第一差异信息,并分别在所述第一临床指南和所述第二临床指南中与所述第一差异信息对应的位置处标注相应的标签,所述标签为新增标签、删除标签或修改标签;若所述第一差异信息为所述第一临床指南相对于所述第二临床指南中的修改部分,则分别在所述第一临床指南和所述第二临床指南中与修改部分对应的位置处标注修改标签;若所述第一差异信息为所述第一临床指南相对于所述第二临床指南中的新增部分,则分别在所述第一临床指南和所述第二临床指南中与新增部分对应的位置处标注新增标签;若所述第一差异信息为所述第一临床指南相对于所述第二临床指南中的删除部分,则分别在所述第一临床指南和所述第二临床指南中与修改部分对应的位置处标注删除标签;

第二处理单元,用于若所述第一临床指南和所述第二临床指南隶属不同来源,或者,若所述第一临床指南和所述第二临床指南隶属同一来源且所述第一临床指南不存在所述更新说明,将所述第一指南模块和所述第二指南模块进行匹配,将相匹配的所述第一指南模块和所述第二指南模块分别作为第一待处理指南模块和第二待处理指南模块,将与所有所述第二指南模块均不匹配的所述第一指南模块作为第三待处理指南模块;

第三处理单元,用于根据所述第一待处理指南模块和所述第二待处理指南模块,确定所述第一临床指南和所述第二临床指南之间的第二差异信息,并分别在所述第一临床指南和所述第二临床指南中与所述第二差异信息对应的位置处标注相应的标签;

第四处理单元,用于根据所述第三待处理指南模块和所有所述第二指南模块,确定所

述第一临床指南和所述第二临床指南之间的第三差异信息,并分别在所述第一临床指南和所述第二临床指南中与所述第三差异信息对应的位置处标注相应的标签。

一种临床指南更新内容的自动识别方法及系统

技术领域

[0001] 本发明涉及数据处理技术领域,具体涉及一种临床指南更新内容的自动识别方法及系统。

背景技术

[0002] 随着临床研究范围(例如肿瘤临床研究范围)的扩展和临床诊疗技术的创新,新的医学证据也在不断迭代,此种情况也就加快了临床指南的更新频率。

[0003] 目前较多的新版本临床指南中不会给出更新说明,即使给出更新说明,更新说明中也不能直观的给出新旧两个版本的临床指南之间的知识差异及知识差异的精准位置。需要临床医生人工查阅新旧版本的临床指南来找到不同版本临床指南间的差异和变化,但此种方式需要耗费大量的人力和时间,并且人工查阅容易出现漏查等情况,确定不同版本临床指南间的差异和变化情况的效率较低和准确度较低。

发明内容

[0004] 有鉴于此,本发明实施例提供一种临床指南更新内容的自动识别方法及系统,以解决目前确定不同临床指南间的差异和变化情况的方式存在的效率低和准确度低等问题。

[0005] 为实现上述目的,本发明实施例提供如下技术方案:

[0006] 本发明实施例第一方面公开一种临床指南更新内容的自动识别方法,所述方法包括:

[0007] 根据预先利用临床指南各级别标题建立的模块层级结构树,分别对第一临床指南和第二临床指南进行解析和结构化提取,至少得到所述第一临床指南对应的第一指南模块和所述第二临床指南对应的第二指南模块,所述第一指南模块为所述第一临床指南中最小级别标题所包含的文本内容,所述第二指南模块为所述第二临床指南中最小级别标题所包含的文本内容;

[0008] 若所述第一临床指南和所述第二临床指南隶属同一来源,且当所述第一临床指南相对于所述第二临床指南存在更新说明,利用所述第一临床指南的所述更新说明,确定所述第一临床指南和所述第二临床指南之间的第一差异信息,并分别在所述第一临床指南和所述第二临床指南中与所述第一差异信息对应的位置处标注相应的标签,所述标签为新增标签、删除标签或修改标签;

[0009] 若所述第一临床指南和所述第二临床指南隶属不同来源,或者,若所述第一临床指南和所述第二临床指南隶属同一来源且所述第一临床指南不存在所述更新说明,将所述第一指南模块和所述第二指南模块进行匹配,将相匹配的所述第一指南模块和所述第二指南模块分别作为第一待处理指南模块和第二待处理指南模块,将与所有所述第二指南模块均不匹配的所述第一指南模块作为第三待处理指南模块;

[0010] 根据所述第一待处理指南模块和所述第二待处理指南模块,确定所述第一临床指南和所述第二临床指南之间的第二差异信息,并分别在所述第一临床指南和所述第二临床

指南中与所述第二差异信息对应的位置处标注相应的标签；

[0011] 根据所述第三待处理指南模块和所有所述第二指南模块，确定所述第一临床指南和所述第二临床指南之间的第三差异信息，并分别在所述第一临床指南和所述第二临床指南中与所述第三差异信息对应的位置处标注相应的标签。

[0012] 优选的，将所述第一指南模块和所述第二指南模块进行匹配的过程，包括：

[0013] 针对每一个第一指南模块，利用预设的深度语义匹配模型，确定所述第一指南模块的标题和每个所述第二指南模块的标题之间的标题相似度；

[0014] 针对每一个第一指南模块，若所有所述标题相似度均小于标题相似度阈值，确定所述第一指南模块与所有所述第二指南模块均不匹配，若至少一个所述标题相似度大于等于所述标题相似度阈值，确定所述第一指南模块与最大所述标题相似度对应的所述第二指南模块相匹配。

[0015] 优选的，所述根据所述第一待处理指南模块和所述第二待处理指南模块，确定所述第一临床指南和所述第二临床指南之间的第二差异信息，并分别在所述第一临床指南和所述第二临床指南中与所述第二差异信息对应的位置处标注相应的标签，包括：

[0016] 分别对所述第一待处理指南模块和所述第二待处理指南模块进行分句处理，得到所述第一待处理指南模块对应的多个第一句子，及得到所述第二待处理指南模块对应的多个第二句子；

[0017] 对于所述第一待处理指南模块的第 m 个第一句子，计算所述第 m 个第一句子与所述第二待处理指南模块的 H 个第二句子之间的句子相似度， m 为大于等于1小于等于 x 的整数， x 为所述第一待处理指南模块所包含的第一句子的总个数， m 从1开始且递增1， H 为大于等于1小于等于 y 的整数， y 为所述第二待处理指南模块所包含的第二句子的总个数；

[0018] 若所述第 m 个第一句子和第 n 个第二句子之间的句子相似度等于1，确定所述第 m 个第一句子和所述第 n 个第二句子相同，不执行标注处理， n 为大于等于1小于等于 y 的整数；

[0019] 若所述第 m 个第一句子和所述第 n 个第二句子之间的句子相似度大于等于句子相似度阈值且小于1，在所述第一临床指南中与所述第 m 个第一句子对应的位置处标注修改标签，及在所述第二临床指南中与所述第 n 个第二句子对应的位置处标注修改标签，当 n 大于 m 时，确定所述第二待处理指南模块中位于所述第 n 个第二句子之前的，与前 m 个所述第一句子的句子相似度均小于所述句子相似度阈值且未进行标注处理的第三句子，在所述第二临床指南中与所述第三句子对应的位置处标注删除标签；

[0020] 若所述第 m 个第一句子与 H 个所述第二句子之间的句子相似度均小于所述句子相似度阈值，在所述第一临床指南中与所述第 m 个第一句子对应的位置处标注新增标签。

[0021] 优选的，所述根据所述第三待处理指南模块和所有所述第二指南模块，确定所述第一临床指南和所述第二临床指南之间的第三差异信息，并分别在所述第一临床指南和所述第二临床指南中与所述第三差异信息对应的位置处标注相应的标签，包括：

[0022] 计算所述第三待处理指南模块中首个句子的前 $P\%$ 内容与每个所述第二指南模块的多个第二句子之间的首句相似度；

[0023] 若至少一个所述首句相似度大于首句相似度阈值，确定所述第三待处理指南模块与最大所述首句相似度对应的所述第二指南模块相匹配；

[0024] 从所述第三待处理指南模块的首个句子的前 $P\%$ 内容处开始，将位于其之后的所

述第一临床指南中已有的标签变更为修改标签；

[0025] 从与所述第三待处理指南模块相匹配的所述第二指南模块中最大所述首句相似程度对应的第二句子处开始，将位于其之后的所述第二临床指南中已有的标签变更为修改标签。

[0026] 优选的，所述分别对第一临床指南和第二临床指南进行解析和结构化提取，至少得到所述第一临床指南对应的第一指南模块和所述第二临床指南对应的第二指南模块之后，还包括：

[0027] 分别对所述第一指南模块和所述第二指南模块进行预处理，并分别提取进行预处理后的所述第一指南模块和所述第二指南模块中的知识特征。

[0028] 优选的，所述在所述第一临床指南中与所述第m个第一句子对应的位置处标注修改标签，及在所述第二临床指南中与所述第n个第二句子对应的位置处标注修改标签之后，还包括：

[0029] 比对所述第m个第一句子和所述第n个第二句子内的所述知识特征之间的差异，得到知识特征差异信息；

[0030] 分别在所述第m个第一句子中和所述第n个第二句子中与所述知识特征差异信息对应的位置处标注相应的标签。

[0031] 优选的，还包括：

[0032] 分别利用不同的展示形式，展示不同类别的所述标签。

[0033] 优选的，分别对第一临床指南和第二临床指南进行解析和结构化提取，至少得到所述第一临床指南对应的第一指南模块和所述第二临床指南对应的第二指南模块之后，还包括：

[0034] 对所述第一指南模块和所述第二指南模块进行规范化处理。

[0035] 优选的，分别对第一临床指南和第二临床指南进行解析和结构化提取，至少得到所述第一临床指南对应的第一指南模块和所述第二临床指南对应的第二指南模块之后，还包括：

[0036] 将所述第一指南模块和所述第二指南模块存储至数据库中，及将所有所述第一指南模块之间的层级关系存储至所述数据库中，及将所有所述第二指南模块之间的层级关系存储至所述数据库中。

[0037] 本发明实施例第二方面公开一种临床指南更新内容的自动识别系统，所述系统包括：

[0038] 解析单元，用于根据预先利用临床指南各级别标题建立的模块层级结构树，分别对第一临床指南和第二临床指南进行解析和结构化提取，至少得到所述第一临床指南对应的第一指南模块和所述第二临床指南对应的第二指南模块，所述第一指南模块为所述第一临床指南中最小级别标题所包含的文本内容，所述第二指南模块为所述第二临床指南中最小级别标题所包含的文本内容；

[0039] 第一处理单元，用于若所述第一临床指南和所述第二临床指南隶属同一来源，且当所述第一临床指南相对于所述第二临床指南存在更新说明，利用所述第一临床指南的所述更新说明，确定所述第一临床指南和所述第二临床指南之间的第一差异信息，并分别在所述第一临床指南和所述第二临床指南中与所述第一差异信息对应的位置处标注相应的

标签,所述标签为新增标签、删除标签或修改标签;

[0040] 第二处理单元,用于若所述第一临床指南和所述第二临床指南隶属不同来源,或者,若所述第一临床指南和所述第二临床指南隶属同一来源且所述第一临床指南不存在所述更新说明,将所述第一指南模块和所述第二指南模块进行匹配,将相匹配的所述第一指南模块和所述第二指南模块分别作为第一待处理指南模块和第二待处理指南模块,将与所有所述第二指南模块均不匹配的所述第一指南模块作为第三待处理指南模块;

[0041] 第三处理单元,用于根据所述第一待处理指南模块和所述第二待处理指南模块,确定所述第一临床指南和所述第二临床指南之间的第二差异信息,并分别在所述第一临床指南和所述第二临床指南中与所述第二差异信息对应的位置处标注相应的标签;

[0042] 第四处理单元,用于根据所述第三待处理指南模块和所有所述第二指南模块,确定所述第一临床指南和所述第二临床指南之间的第三差异信息,并分别在所述第一临床指南和所述第二临床指南中与所述第三差异信息对应的位置处标注相应的标签。

[0043] 基于上述本发明实施例提供的一种临床指南更新内容的自动识别方法及系统,该方法为:根据预先利用临床指南各级别标题建立的模块层级结构树,分别对第一临床指南和第二临床指南进行解析和结构化提取,至少得到第一临床指南对应的第一指南模块和第二临床指南对应的第二指南模块;若第一临床指南和第二临床指南隶属同一来源,且当第一临床指南相对于第二临床指南存在更新说明,利用第一临床指南的所述更新说明,确定第一临床指南和第二临床指南之间的第一差异信息,并分别在第一临床指南和第二临床指南中与第一差异信息对应的位置处标注相应的标签;若第一临床指南和第二临床指南隶属不同来源,或者,若第一临床指南和第二临床指南隶属同一来源且第一临床指南不存在更新说明,将第一指南模块和第二指南模块进行匹配,将相匹配的第一指南模块和第二指南模块分别作为第一待处理指南模块和第二待处理指南模块,将与所有第二指南模块均不匹配的第一指南模块作为第三待处理指南模块;根据第一待处理指南模块和第二待处理指南模块,确定第一临床指南和第二临床指南之间的第二差异信息,并分别在第一临床指南和第二临床指南中与第二差异信息对应的位置处标注相应的标签;根据第三待处理指南模块和所有第二指南模块,确定第一临床指南和第二临床指南之间的第三差异信息,并分别在第一临床指南和第二临床指南中与第三差异信息对应的位置处标注相应的标签。不需要人工查阅需要比对的两个临床指南来找到不同临床指南间的差异和变化,提高确定不同临床指南间差异和变化情况的效率和准确度。

附图说明

[0044] 为了更清楚地说明本发明实施例或现有技术中的技术方案,下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍,显而易见地,下面描述中的附图仅仅是本发明的实施例,对于本领域普通技术人员来讲,在不付出创造性劳动的前提下,还可以根据提供的附图获得其他的附图。

[0045] 图1为本发明实施例提供的一种临床指南更新内容的自动识别方法的流程图;

[0046] 图2为本发明实施例提供的在第一临床指南和第二临床指南中进行标签标注的流程图;

[0047] 图3为本发明实施例提供的一种临床指南更新内容的自动识别方法的另一流程

图；

[0048] 图4为本发明实施例提供的EAU肾细胞癌临床指南章节要点总结的示意图；

[0049] 图5为本发明实施例提供的标注更新标签的示意图；

[0050] 图6为本发明实施例提供的标注知识特征差异信息对应的标签的示意图；

[0051] 图7为本发明实施例提供的标注知识特征差异信息对应的标签的另一示意图；

[0052] 图8为本发明实施例提供的标注知识特征差异信息对应的标签的又一示意图；

[0053] 图9为本发明实施例提供的更新时序的展示示意图；

[0054] 图10为本发明实施例提供的一种临床指南更新内容的自动识别系统的结构框图。

具体实施方式

[0055] 下面将结合本发明实施例中的附图,对本发明实施例中的技术方案进行清楚、完整地描述,显然,所描述的实施例仅仅是本发明一部分实施例,而不是全部的实施例。基于本发明中的实施例,本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例,都属于本发明保护的范围。

[0056] 在本申请中,术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含,从而使得包括一系列要素的过程、方法、物品或者设备不仅包括那些要素,而且还包括没有明确列出的其他要素,或者是还包括为这种过程、方法、物品或者设备所固有的要素。在没有更多限制的情况下,由语句“包括一个……”限定的要素,并不排除在包括所述要素的过程、方法、物品或者设备中还存在另外的相同要素。

[0057] 由背景技术可知,目前在比对不同临床指南的差异和变化情况时,通常需要从临床医生查阅该不同临床指南来找到不同临床指南的差异和变化,但是人工比对的方式需要耗费大量的人力和物力,并且在比对过程中容易出现漏查等情况,这就导致了比对不同临床指南的效率较低和准确度较低。

[0058] 因此,本发明实施例提供一种临床指南更新内容的自动识别方法及系统,根据预先利用临床指南各级别标题建立的模块层级结构树,分别对第一临床指南和第二临床指南进行解析和结构化提取,至少得到第一临床指南对应的第一指南模块和第二临床指南对应的第二指南模块;确定第一指南模块和第二指南模块之间的差异信息,并分别在第一临床指南和第二临床指南中与该差异信息对应的位置处标注相应的标签,不需要人工查阅需要比对的两个临床指南来找到不同临床指南间的差异和变化,以提高确定不同临床指南间差异和变化情况的效率和准确度。

[0059] 需要说明的是,临床指南中存在多个级别的标题,本发明实施例中所涉及的指南模块(例如下文中涉及的第一指南模块和第二指南模块)具体是指:临床指南中最小级别标题下所包含的文本内容。由前述内容可知,各指南模块也存在相应的标题。

[0060] 可以理解的是,在本发明实施例中所涉及的临床指南(例如肿瘤临床指南等),即为医学指南或临床实践指南等关于特定医疗保健领域的指导文件。

[0061] 在本发明实施例中,需要比对第一临床指南和第二临床指南之间的差异情况,此时第一临床指南可被称为匹配临床指南,第二临床指南可被称为待匹配临床指南。

[0062] 参见图1,示出了本发明实施例提供的一种临床指南更新内容的自动识别方法的流程图,该自动识别方法包括:

[0063] 步骤S101:根据预先利用临床指南各级别标题建立的模块层级结构树,分别对第一临床指南和第二临床指南进行解析和结构化提取,至少得到第一临床指南对应的第一指南模块和第二临床指南对应的第二指南模块。

[0064] 需要说明的是,第一指南模块为第一临床指南中最小级别标题所包含的文本内容,第二指南模块为第二临床指南中最小级别标题所包含的文本内容。

[0065] 在具体实现步骤S101的过程中,预先根据临床指南的各级别标题构建模块层级结构树,制定模块层级结构树与该临床指南中各个指南模块之间的映射规则,从而实现基于模块层级结构树的指南模块映射。

[0066] 利用模块层级结构树,对非结构化的第一临床指南进行解析和结构化提取,至少得到第一临床指南对应的第一指南模块,以及,利用模块层级结构树,对非结构化的第二临床指南进行解析和结构化提取,至少得到第二临床指南对应的第二指南模块。

[0067] 可以理解的是,对第一临床指南进行解析和结构化提取后,得到指定格式(比如CSV格式)的数据,该指定格式的数据的列数据为第一指南模块的模块名称,该指定格式的数据的行数据为第一指南模块的模块内容。同理,对第二临床指南进行解析和结构化提取后,也得到指定格式的数据,其中列数据为第二指南模块的模块名称,行数据为第二指南模块的模块内容。

[0068] 优选的,在提取得到第一指南模块和第二指南模块之后,将第一指南模块(也就是上述提及的指定格式的数据)和第二指南模块存储至数据库(例如关系数据库)中,及将所有第一指南模块之间的层级关系存储至数据库中,及将所有第二指南模块之间的层级关系存储至数据库中。可以理解的是,所有第一指南模块之间的层级关系具体是指:所有第一指南模块的模块名称之间的层级关系,所有第二指南模块之间的层级关系具体是指:所有第二指南模块的模块名称之间的层级关系。

[0069] 也就是说,数据库中存储两部分数据,一部分为指南模块(例如上述提及的CSV格式数据),另一部分为指南模块的模块名称,及模块名称之间的层级关系,该模块名称之间的层级关系即为上述提及的模块层级结构树。在数据库中,根据模块名称的层级关系(也就是模块层级结构树),可将指南模块与它们之间的结构(即指南模块之间的结构)自动映射起来。

[0070] 需要说明的是,所提取得到的第一指南模块和第二指南模块对应的模块内容中可能存在不规范的内容,因此,优选的,对第一指南模块和第二指南模块进行规范化处理。

[0071] 规范化处理的具体方式为(此处以指南模块表示第一指南模块和第二指南模块):由于指南模块对应的模块内容(也就是文本内容)中穿插较多参考文献引用标识,因此通过制定正则表达式匹配规则,去除指南模块中所有参考文献引用标识,同时对指南模块进行文本清洗、去停用词、词干化、将英文数字转换为阿拉伯数字和将缩略语全称扩展等降噪处理,从而将指南模块对应的模块内容转换成规范化文本。

[0072] 利用进行规范化处理后的第一指南模块和第二指南模块,进行以下各步骤的相关处理,相关处理的内容详见以下各个步骤示出的内容。

[0073] 步骤S102:若第一临床指南和第二临床指南隶属同一来源,且当第一临床指南相对于第二临床指南存在更新说明,利用第一临床指南的更新说明,确定第一临床指南和第二临床指南之间的第一差异信息,并分别在第一临床指南和第二临床指南中与第一差异信

息对应的位置处标注相应的标签。

[0074] 需要说明的是,标签为新增标签、删除标签或修改标签,新增标签为指示第一临床指南相较于第二临床指南的新增内容的标签,即新增内容只存在第一临床指南中,在第二临床指南中不存在该新增内容;删除标签为指示第一临床指南相较于第二临床指南的删除内容的标签,即删除内容只存在第二临床指南中,第一临床指南中不存在该删除内容;修改标签为指示第一临床指南相较于第二临床指南的修改内容的标签,即修改内容为第一临床指南和第二临床指南中相似但不相同的内容。

[0075] 需要说明的是,临床指南所附带的更新说明中,会标明临床指南中的哪一页的哪部分内容做了哪些改变,或者,在PDF中附带更新内容的超链接,点击该超链接可定位到更新部分。

[0076] 可以理解的是,第一临床指南和第二临床指南可能隶属不同来源,也可能隶属相同来源。当第一临床指南和第二临床指南隶属相同来源时,第一临床指南为相较于第二临床指南的更新版本,此时如果第一临床指南存在更新说明(相对于第二临床指南的更新说明),则利用更新说明确定第一临床指南和第二临床指南之间的差异信息,如果第一临床指南不存在更新说明,则需要比对第一指南模块和第二指南模块来确定第一临床指南和第二临床指南之间的差异信息。

[0077] 同时,如果第一临床指南和第二临床指南隶属于不同来源,则也需要比对第一指南模块和第二指南模块来确定第一临床指南和第二临床指南之间的差异信息。

[0078] 比如:假设有两个来源的肾肿瘤指南,一个为NCCN(也就是临床指南的来源)编写的,NCCN编写的肾肿瘤指南附带更新说明,另一个为EAU编写的。如果第一临床指南为NCCN编写的2020年第2版本V2,第二临床指南为NCCN编写的2020年第1版本V1,即第一临床指南与第二临床指南隶属同一来源且第一临床指南附带更新说明,则直接利用更新说明确定第一临床指南和第二临床指南之间的差异信息。如果第一临床指南为EAU编写的2020年第2版本V2,第二临床指南为EAU编写的2020年第1版本V1,即第一临床指南与第二临床指南隶属同一来源但第一临床指南没有附带更新说明,则需要比对第一指南模块和第二指南模块来确定第一临床指南和第二临床指南之间的差异信息。如果第一临床指南为EAU编写的2020年第2版本V2,第二临床指南为NCCN编写的2020年第2版本V2,即第一临床指南与第二临床指南隶属不同来源,则需要比对第一指南模块和第二指南模块来确定第一临床指南和第二临床指南之间的差异信息。

[0079] 在具体实现步骤S102的过程中,若第一临床指南和第二临床指南隶属同一来源(此时第一临床指南为相对于第二临床指南的新版本临床指南),且当第一临床指南相对于第二临床指南存在更新说明时,利用预设的指南更新识别和标注规则,对第一临床指南和第二临床指南之间更新的指南模块和更新的句子进行自动定位和识别,从而确定第一临床指南和第二临床指南之间的第一差异信息,并分别在第一临床指南和第二临床指南中与第一差异信息对应的位置处标注相应的标签。

[0080] 如果第一差异信息为第一临床指南相对于第二临床指南中的修改部分(指示发生修改的指南模块和该指南模块中发生修改的句子),则分别在第一临床指南和第二临床指南中与修改部分对应的位置处标注修改标签。

[0081] 如果第一差异信息为第一临床指南相对于第二临床指南中的新增部分,则在第一

临床指南中与新增部分对应的位置处标注新增标签。

[0082] 如果第一差异信息为第一临床指南相对于第二临床指南中的删除部分,则在第二临床指南中与删除部分对应的位置处标注删除标签。

[0083] 需要说明的是,设置指南更新识别和标注规则的过程为:总结和归纳临床指南的更新说明的书写规律,并按照该书写规律设置指南更新识别和标注规则,例如:部分临床指南中用带“remove”的句式阐述旧版本临床指南中删除的内容,用带“modified”的句式阐述新版本临床指南中修改的内容,或者,用单独的表格将新版本临床指南中新增的内容单独列举出来。

[0084] 可以理解的是,由上述内容可知,差异信息可以指示第一临床指南相较于第二临床指南的修改部分、新增部分或删除部分,优选的,在第一临床指南和第二临床指南中与第一差异信息对应的位置处标注相应的标签时,可分别利用不同的展示形式,展示不同类别的标签。

[0085] 也就是说,通过不同的展示形式,在第一临床指南和第二临床指南中,分别展示新增标签、删除标签和修改标签,从而区分新增标签、删除标签和修改标签。

[0086] 例如:通过利用不同颜色突出显示文本的方式,展示新增标签、删除标签和修改标签,其中,以黄色突出显示文本的方式,展示第一临床指南和第二临床指南中发生修改的部分,即以黄色突出显示文本的方式展示修改标签;以蓝色突出显示文本的方式,展示第一临床指南中相较于第二临床指南的新增部分,即以蓝色突出显示文本的方式展示新增标签;以红色突出显示文本的方式,展示第二临床指南中相较于第一临床指南的删除部分,即以红色突出显示文本的方式展示删除标签。

[0087] 上述关于展示新增标签、删除标签和修改标签的方式仅用于举例说明,在实际应用中,也可采用其它不同的方式展示新增标签、删除标签和修改标签,在本发明实施例中关于如何展示新增标签、删除标签和修改标签的方式,不做具体限定。

[0088] 步骤S103:若第一临床指南和第二临床指南隶属不同来源,或者,若第一临床指南和第二临床指南隶属同一来源且第一临床指南不存在更新说明,将第一指南模块和第二指南模块进行匹配,将相匹配的第一指南模块和第二指南模块分别作为第一待处理指南模块和第二待处理指南模块,将与所有第二指南模块均不匹配的第一指南模块作为第三待处理指南模块。

[0089] 在具体实现步骤S103的过程中,若第一临床指南和第二临床指南隶属不同来源,或者,若第一临床指南和第二临床指南隶属同一来源但第一临床指南不存在更新说明(相较于第二临床指南的更新说明),将第一指南模块和第二指南模块进行匹配,将相匹配的第一指南模块和第二指南模块分别作为第一待处理指南模块和第二待处理指南模块,将与所有第二指南模块均不匹配的第一指南模块作为第三待处理指南模块,并执行步骤S104和步骤S105。

[0090] 在具体实现中,针对每一第一指南模块,将该第一指南模块与所有第二指南模块进行匹配,若能确定得到与该第一指南模块相匹配的第二指南模块,则将该第一指南模块作为第一待处理指南模块,将与该第一指南模块(也就是第一待处理指南模块)相匹配的第二指南模块作为第二指南模块;若该第一指南模块与所有第二指南模块均不匹配,则将第一指南模块作为第三待处理指南模块。

[0091] 每对第一待处理指南模块和第二待处理指南模块即为：相匹配的第一指南模块和第二指南模块。

[0092] 需要说明的是，发明人经研究发现，指南模块的标题是对指南模块的模块内容的高度概括，指南模块的标题具有简短和高度明确性的特点，并且随着临床指南版本修订与更新，各个指南模块的标题名称大多数不会发生巨大变化。

[0093] 因此，将第一指南模块和第二指南模块进行匹配的具体方式为：针对每一个第一指南模块，利用预设的深度语义匹配模型(DSSM语义匹配模型)，确定该第一指南模块的标题和每个第二指南模块的标题之间的标题相似度。

[0094] 针对每一个第一指南模块，若该第一指南模块和所有第二指南模块之间的所有标题相似度均小于标题相似度阈值，确定该第一指南模块与所有第二指南模块均不匹配；若该第一指南模块和至少一个第二指南模块之间的标题相似度大于等于标题相似度阈值，确定该第一指南模块与最大标题相似度对应的第二指南模块相匹配。

[0095] 也就是说，如果第一指南模块与所有第二指南模块之间的标题相似度均小于标题相似度阈值，则该第一指南模块与所有第二指南模块均不匹配，该第一指南模块即为第三待处理指南模块。

[0096] 如果第一指南模块与一个及以上的第二指南模块之间的标题相似度大于等于标题相似度阈值，确定该第一指南模块与最大标题相似度对应的第二指南模块相匹配，该第一指南模块为第一待处理指南模块，最大标题相似度对应的第二指南模块为第二待处理指南模块。

[0097] 先利用所有确定得到的第一待处理指南模块和与之对应的第二待处理指南模块，确定第一临床指南和第二临床指南之间的第二差异信息，并在第一临床指南和第二临床指南中与第二差异信息对应的位置处标注相应的标签；之后再利用第三待处理指南模块确定第一临床指南和第二临床指南之间的第三差异信息，最后并分别在第一临床指南和第二临床指南中与第三差异信息对应的位置处标注相应的标签；具体实施内容详见以下说明。

[0098] 步骤S104：根据第一待处理指南模块和第二待处理指南模块，确定第一临床指南和第二临床指南之间的第二差异信息，并分别在第一临床指南和第二临床指南中与第二差异信息对应的位置处标注相应的标签。

[0099] 在具体实现步骤S104的过程中，针对每一对相匹配的第一待处理指南模块和第二待处理指南模块，利用该第一待处理指南模块和该第二待处理指南模块，确定第一临床指南和第二临床指南之间的第二差异信息，并分别在第一临床指南和第二临床指南中与第二差异信息对应的位置处标注相应的标签，该第二差异信息即指示相匹配的第一待处理指南模块和第二待处理指南模块中的差异部分，例如：第一待处理指南模块相较于第二待处理指南模块中的修改部分，第一待处理指南模块相较于第二待处理指南模块中的新增部分，第一待处理指南模块相较于第二待处理指南模块中的删除部分。

[0100] 关于在第一临床指南和第二临床指南中与第二差异信息对应的位置处标注相应的标签的具体内容，可参见上述步骤S102中的内容，在此不再赘述。

[0101] 可以理解的是，在利用相匹配的第一待处理指南模块和第二待处理指南模块确定第二差异信息时，是以句子为单位确定第一待处理指南模块和第二待处理指南模块之间的差异(即第二差异信息)的。因此，需要先将第一待处理指南模块和第二待处理指南模块进

行分句处理,再计算第一待处理指南模块和第二待处理指南模块的句子之间的句子相似度,最后利用计算得到的句子相似度,来确定第一待处理指南模块和第二待处理指南模块之间的差异(即第二差异信息)。

[0102] 需要说明的是,利用指定算法计算句子间的句子相似度,比如利用通用语句编码器(Universal Sentence Encoder)计算句子相似度,在本发明实施例中对于计算句子相似度的方式不做具体限定。

[0103] 步骤S105:根据第三待处理指南模块和所有第二指南模块,确定第一临床指南和第二临床指南之间的第三差异信息,并分别在第一临床指南和第二临床指南中与第三差异信息对应的位置处标注相应的标签。

[0104] 由上述内容可知,第三待处理指南模块为与所有第二指南模块均不匹配的第一指南模块,需要说明的是,第一指南模块和第二指南模块的标题之间的标题相似度小于标题相似度阈值(即该第一指南模块和该第二指南模块不相似)存在3种情况。

[0105] 第1种情况是:该第二指南模块对应的模块内容为第一临床指南相较于第二临床指南的删除内容,即该第二指南模块对应的模块内容在第二临床指南中存在,但在第一临床指南中不存在。

[0106] 第2种情况是:该第一指南模块对应的模块内容为第一临床指南相较于第二临床指南的新增内容,即该第一指南模块对应的模块内容在第一临床指南中存在,而在第二临床指南中不存在。

[0107] 第3种情况是:如果该第一指南模块为第三待处理指南模块,该第三待处理指南模块对应的模块内容合并另一版临床指南的其它指南模块中。“另一版临床指南”具体为:如果第三待处理指南模块对应的内容在第一临床指南中,则另一版临床指南为第二临床指南,反之亦然。

[0108] 针对上述第3种情况,在具体实现步骤S105的过程中,针对每一第三待处理指南模块,利用该第三待处理指南模块和所有第二指南模块,确定第一临床指南和第二临床指南之间的第三差异信息,并分别在第一临床指南和第二临床指南中与第三差异信息对应的位置处标注相应的标签,关于在第一临床指南和第二临床指南中与第三差异信息对应的位置处标注相应的标签的具体内容,可参见上述步骤S102中的内容,在此不再赘述。

[0109] 具体利用第三待处理指南模块和所有第二指南模块确定第三差异信息的过程为:计算第三待处理指南模块中首个句子的前P%(比如前20%)内容与每个第二指南模块的多个(例如5个)第二句子之间的首句相似度;若至少一个首句相似度大于首句相似度阈值,确定第三待处理指南模块与最大首句相似度对应的第二指南模块相匹配,也就是说,如果只有一个首句相似度大于首句相似度阈值,确定第三待处理指南模块与首句相似度大于首句相似度阈值的第二指南模块相匹配,如果有多个首句相似度大于首句相似度阈值,确定第三待处理指南模块与最大首句相似度对应的第二指南模块相匹配。

[0110] 需要说明的是,计算第三待处理指南模块中首个句子的前P%(比如前20%)内容与每个第二指南模块的多个(例如5个)第二句子之间的首句相似度具体是指:最多计算第三待处理指南模块中首个句子的前P%内容与第二指南模块的多个第二句子之间的首句相似度,例如:最多计算第三待处理指南模块中首个句子的前P%内容与第二指南模块的5个第二句子之间的首句相似度。

[0111] 由上述步骤S104可知,已经在第一临床指南和第二临床指南中与第二差异信息对应的位置处标注相应的标签,即第一临床指南和第二临床指南中可能已经存在部分标签,在确定第三待处理指南模块和与之相匹配的第二指南模块之后,从第三待处理指南模块的首个句子的前P%内容处开始,将位于其之后的第一临床指南中已有的标签变更为修改标签;并且,从与第三待处理指南模块相匹配的第二指南模块中最大首句相似度对应的第二句子处开始,将位于其之后的第二临床指南中已有的标签变更为修改标签。

[0112] 也就是说,如果根据第三待处理指南模块中首个句子的前P%内容,能匹配到第二指南模块的句子(首句相似度大于首句相似度阈值且最大),则在第一临床指南中,从第三待处理指南模块的首个句子的前P%内容处开始,利用修改标签覆盖前P%内容处之后的内容中已有的标签(步骤S104中已标注的新增标签);以及,在第二临床指南中,从与第三待处理指南模块匹配的第二指南模块中最大首句相似度对应的第二句子处开始,利用修改标签覆盖该句子之后的内容中已有的标签(骤S104中已标注的删除标签)。

[0113] 优选的,在执行步骤S104和步骤S105后,在前端展示步骤S104和步骤S105所标注的标签,其中利用不同的展示形式,展示不同类别的标签。

[0114] 在本发明实施例中,根据预先利用临床指南各级别标题建立的模块层级结构树,分别对第一临床指南和第二临床指南进行解析和结构化提取,至少得到第一临床指南对应的第一指南模块和第二临床指南对应的第二指南模块;确定第一指南模块和第二指南模块之间的差异信息,并分别在第一临床指南和第二临床指南中与该差异信息对应的位置处标注相应的标签,不需要人工查阅需要比对的两个临床指南来找到不同临床指南间的差异和变化,提高确定不同临床指南间差异和变化情况的效率和准确度。

[0115] 上述本发明实施例图1步骤S104中涉及的分别在第一临床指南和第二临床指南中与第二差异信息对应的位置处标注相应的标签的过程,参见图2,示出了本发明实施例提供的在第一临床指南和第二临床指南中进行标签标注的流程图,包括以下步骤:

[0116] 步骤S201:分别对第一待处理指南模块和第二待处理指南模块进行分句处理,得到第一待处理指南模块对应的多个第一句子,及得到第二待处理指南模块对应的多个第二句子。

[0117] 需要说明的是,指南模块的句子中可能存在相应的知识特征,优选的,在获取第一待处理指南模块和第二待处理指南模块之后,分别对第一指南模块和第二指南模块进行预处理,并分别提取进行预处理后的第一指南模块和第二指南模块中的知识特征,具体提取指南模块(以此表征第一指南模块和第二指南模块)的知识特征的方式为:利用预先定义的知识特征类型,采用基于规则或机器学习的方法对指南模块中的知识特征进行识别和抽取。

[0118] 指南模块中的知识特征包括但不限于:医学实体特征、等级特征、数量特征和时间特征,各个知识特征的详细解释详见以下说明。

[0119] 医学实体特征:医学实体特征指疾病诊疗过程中涉及临床症状、药物和检查方法等各类医学术语,在临床指南中主要用于说明各类医疗活动和行为涉及的对象,是临床医生学习临床指南时主要关注的知识。

[0120] 疾病等级特征:是指临床指南中提及的某一医疗干预措施,临床指南编委基于证据来源、获取途径及专家共识情况确定的此干预措施的证据或推荐等级,级别越高说明证

据越充分,也就更推荐应用于临床实践。例如在各种肿瘤临床指南文本中通常有固定的级别分类和表达形式,表达形式诸如“category+等级”和“LE:+等级”等,并直接位于说明某一医疗干预措施的句子中或以括号的形式附注在句子末尾。

[0121] 数量特征:用于描述诊疗相关的各类统计量,除包含样本数、P值和生存率等基本实验统计数据外,还对用药剂量、体内物质含量、肿瘤大小等临床诊疗数据进行说明,主要存在基础数词和复合数词两种形式。

[0122] 其中,基础数词的表现形式包括三种,分别为英文基数词、阿拉伯数字和百分比数。英文基数词为诸如“sixty-four”等形式,阿拉伯数字这一表现形式可包含小数点、逗号和空格,阿拉伯数字这一表现形式也可不带小数点、逗号和空格,百分比数可以是:英文基础词+“percent”(例如seventypercent),百分比数也可以是:阿拉伯数字+“%”(例如70%)。

[0123] 复合数词指利用数词修饰符,对临床指南中基础数词描述的内容进行说明或限定,例如表示用药剂量的“mg/dL”、表示肿瘤尺寸的“cm”和表示倍数的“times”等。

[0124] 时间特征:用于表达临床指南中某一医疗事件、操作发生的时间或持续时长,主要包括时刻和时段两类。

[0125] 其中,时刻是指某一具体时间点,在临床指南中主要用于阐述某一发现和实验所在具体年份时间,基本组成形式为四位阿拉伯数字。时段是指持续的一段时间,通常用于表述某一医疗操作的持续时间或某一年份范围,例如“6months”和“2011to 2013”等,其基本组成形式除常见的数字(阿拉伯数字和英文)结合时间量词以为,还可通过连接词或连接符号将不同表示时刻的词连接到一起(例如“2006through 2015”或“2006-2015”),或者将表示时刻的词与时间量词整合(例如“5-year”等)。

[0126] 上述内容即为部分知识特征的详细介绍,对于其它类型的知识特征就不一一举例并详细说明,本领域技术人员可根据实际情况确定需要提取的知识特征,在此不做具体限定。

[0127] 在具体实现步骤S201的过程中,对第一待处理指南模块和第二待处理指南模块进行分句处理并对各句子中的知识特征进行语义扩展和规范化处理,得到第一待处理指南模块对应的多个第一句子,及得到第二待处理指南模块对应的多个第二句子。

[0128] 具体对指南模块(以此表征第一待处理指南模块和第二待处理指南模块)各句子中的知识特征进行语义扩展和规范化处理的方式为:先对指南模块进行分句,再利用基于指定疾病编码(例如ICD-10)和受控词表对指南模块的句子中的知识特征进行语义扩展和规范化处理,解决该指南模块的句子中的知识特征的词法和结构上的异构问题,从而得到指南模块的经过语义扩展和规范化处理后的句子。其中,词法异构包括同义词、缩写、首字母缩写、大小写敏感度和变形等,句法问题包括排序、分隔符和缺失等。

[0129] 步骤S202:对于第一待处理指南模块的第m个第一句子,计算第m个第一句子与第二待处理指南模块的H个第二句子之间的句子相似度。

[0130] 需要说明的是,m为大于等于1小于等于x的整数,x为第一待处理指南模块所包含的第一句子的总个数,m从1开始且递增1,H为大于等于1小于等于y的整数,y为第二待处理指南模块所包含的第二句子的总个数。

[0131] 由上述本发明实施例图1中的内容可知,第一待处理指南模块和第二待处理指南

模块互相对应,在具体实现步骤S202的过程中,基于上述步骤S201进行语义扩展和规范化处理后的句子的知识特征和知识特征的属性(例如知识特征的类别和位置等),对于第一待处理指南模块的第 m 个第一句子,将第 m 个第一句子与第二待处理指南模块中的第二句子进行遍历相似计算,即计算第 m 个第一句子与第二待处理指南模块的 H 个第二句子之间的句子相似度。

[0132] 也就是说,从第一待处理指南模块的第1个第一句子开始,计算第1($m=1$)个第一句子与第二待处理指南模块的 H 个第二句子间的句子相似度,然后到第一待处理指南模块的第2个第一句子,计算第2($m=2$)个第一句子与第二待处理指南模块的 H 个第二句子之间的句子相似度,以此类推,直至计算第 x ($m=x$)个第一句子与第二待处理指南模块的 H 个第二句子之间的句子相似度。

[0133] 可以理解的是,为节约计算资源和提高效率,设置句子迭代步长 H ,即只计算第 m 个第一句子与第二待处理指南模块的部分(H 个)第二句子之间的句子相似度。

[0134] 也就是说,最多计算第 m 个第一句子与第二待处理指南模块的 H 个第二句子之间的句子相似度,该 H 个第二句子的序号范围为:与第 $m-1$ 个第一句子的句子相似度大于等于句子相似度阈值的第二句子的序号往后推 H 。

[0135] 例如:假设句子迭代步长为5($H=5$),从第一待处理指南模块的第1个第一句子开始,最多计算第1($m=1$)个第一句子与第二待处理指南模块的前5个(包括第5个)第二句子的句子相似度,假设此时第1个第一句子与第1个第二句子的句子相似度大于等于句子相似度阈值。然后到第一待处理指南模块的第2个第一句子,最多计算第2($m=2$)个第一句子与第二待处理指南模块的5个第二句子(此时为计算第2个第二句子至第6个第二句子)的句子相似度,假设此第2个第一句子与第3个第二句子的句子相似度大于等于句子相似度阈值。然后到第一待处理指南模块的第3个第一句子,最多计算第3($m=3$)个第一句子与第二待处理指南模块的5个第二句子(此时为计算第4个第二句子至第8个第二句子)的句子相似度,以此类推进行计算。

[0136] 需要说明的是,句子迭代步长 H 的取值,可根据实际情况进行设置,在此不做具体限定。

[0137] 步骤S203:若第 m 个第一句子和第 n 个第二句子之间的句子相似度等于1,确定第 m 个第一句子和第 n 个第二句子相同,不执行标注处理。

[0138] 需要说明的是, n 为大于等于1小于等于 y 的整数,且 n 位于与第 m 个第一句子计算句子相似度的 H 个第二句子的序号范围内。

[0139] 在具体实现步骤S203的过程中,如果第一待处理指南模块中的第 m 个第一句子与第二待处理指南模块中的第 n 个第二句子之间的句子相似度等于1,则说明第 m 个第一句子和第 n 个第二句子相同,此时在第一待处理指南模块中的第 m 个第一句子对应的位置处不做标注,以及在第二待处理指南模块中的第 n 个第二句子对应的位置处不做标注。

[0140] 步骤S204:若第 m 个第一句子和第 n 个第二句子之间的句子相似度大于等于句子相似度阈值且小于1,在第一临床指南中与第 m 个第一句子对应的位置处标注修改标签,及在第二临床指南中与第 n 个第二句子对应的位置处标注修改标签,当 n 大于 m 时,确定第二待处理指南模块中位于第 n 个第二句子之前的,与前 m 个第一句子的句子相似度均小于句子相似度阈值且未进行标注处理的第三句子,在第二临床指南中与第三句子对应的位置处标注删

除标签。

[0141] 在具体实现步骤S204的过程中,若第 m 个第一句子和第 n 个第二句子之间的句子相似度大于等于句子相似度阈值且小于1,说明第一待处理指南模块中的第 m 个第一句子相较于第二待处理指南模块中的第 n 个第二句子存在修改内容,此时在第一临床指南的第一待处理指南模块中与第 m 个第一句子对应的位置处标注修改标签,及在第二临床指南的第二待处理指南模块中与第 n 个第二句子对应的位置处标注修改标签。

[0142] 与此同时,当 n 大于 m 时,确定第二待处理指南模块中位于第 n 个第二句子之前的、与前 m 个第一句子的句子相似度均小于句子相似度阈值且未进行标注处理的第三句子,在第二临床指南的第二待处理指南模块中与第三句子对应的位置处标注删除标签。

[0143] 例如:假设第4个第一句子和第5个第二句子之间的句子相似度大于等于句子相似度阈值且小于1,此时 n ($n=5$) 大于 m ($m=4$),确定第二待处理指南模块中位于第5个第二句子之前的、与前4个(包括第4个)第一句子的句子相似度均小于句子相似度阈值且未进行标注处理的第三句子(假设为第2个第二句子),在第二临床指南中与第2个第二句子对应的位置处标注删除标签。

[0144] 可以理解的是,当确定第一待处理指南模块中的第 m 个第一句子相较于第二待处理指南模块中的第 n 个第二句子存在修改内容时,还可能指示第 m 个第一句子的知识特征相较于第 n 个第二句子的知识特征存在更新情况,此时需要确定第 m 个第一句子和第 n 个第二句子内的知识特征之间的差异。

[0145] 优选的,在执行步骤S204之后,基于上述步骤S201进行语义扩展和规范化处理后的句子的知识特征和知识特征的属性(例如知识特征的类别和位置等),比对第 m 个第一句子和第 n 个第二句子内的知识特征之间的差异,得到知识特征差异信息,分别在第 m 个第一句子中和第 n 个第二句子中与知识特征差异信息对应的位置处标注相应的标签。

[0146] 也就是说,在第 m 个第一句子和第 n 个第二句子中发生更新变化的知识特征的位置处标注相应的标签,指示该知识特征在第一临床指南的第一待处理指南模块和第二临床指南的第二待处理指南模块中发生变化,

[0147] 同时,以不同的展示形式,展示不同类别的用于指示知识特征差异信息的标签,比如:以红色下划线及字体加粗表示实体知识特征差异信息,以青色下划线及字体加粗展示等级知识特征差异信息,以蓝色下划线及字体加粗展示数量知识特征差异信息,以绿色下划线展示时间知识特征差异信息。具体展示形式可根据实际情况进行设置,在此不做具体限定。

[0148] 优选的,还可基于时间序列(临床指南发布的时间)统计并展示知识特征差异信息。与此同时,在前端展示以不同的展示形式,展示不同类别的用于指示知识特征差异信息的标签。

[0149] 步骤S205:若第 m 个第一句子与 H 个第二句子之间的句子相似度均小于句子相似度阈值,在第一临床指南中与第 m 个第一句子对应的位置处标注新增标签。

[0150] 在具体实现步骤S205的过程中,若第 m 个第一句子与 H 个第二句子之间的句子相似度均小于句子相似度阈值,表示第一待处理指南模块中的第 m 个第一句子为相较于第二待处理指南模块中的新增内容,在第一临床指南的第一待处理指南模块中与第 m 个第一句子对应的位置处标注新增标签。

[0151] 通过上述步骤S201至步骤S205的内容,依次处理(m从1至x)所有第一待处理指南模块中的每个第一句子,从而得到第一临床指南和第二临床指南之间的第二差异信息,并分别在第一临床指南和第二临床指南中与第二差异信息对应的位置处标注相应的标签。

[0152] 在本发明实施例中,从指南模块、句子和知识特征3个维度,确定第一临床指南和所述第二临床指南之间的第二差异信息,并分别在第一临床指南和第二临床指南中与第二差异信息对应的位置处标注相应的标签,不需要人工查阅需要比对的两个临床指南来找到不同临床指南间的差异和变化,提高确定不同临床指南间差异和变化情况的效率和准确度。以及基于时间序列进行知识特征差异信息的统计,挖掘临床指南中知识特征的时序变化,将静态和非结构化的临床指南转化为结构化、知识特征化和可视化的表示形式,实现对临床指南的更新内容的多层次和多维度揭示,以辅助临床医生直观了解临床指南之间的差异和变化情况,提高临床医生的学习效率。

[0153] 为更好解释说明上述本发明实施例图1和图2中的内容,通过图3示出的内容进行举例说明,其中,图3示出的第一临床指南和第二临床指南隶属同一来源,参见图3,示出了本发明实施例提供的一种临床指南更新内容的自动识别方法的另一流程图,该自动识别方法包括:

[0154] 步骤S301:对第一临床指南和第二临床指南进行解析和结构化提取,至少得到第一临床指南对应的第一指南模块和第二临床指南对应的第二指南模块。

[0155] 步骤S302:对第一指南模块和第二指南模块进行规范化处理。

[0156] 在具体实现步骤S302的过程中,对指南模块进行规范化处理的具体内容,参见上述本发明实施例图1步骤S101中的内容,在此不再赘述。

[0157] 步骤S303:提取第一指南模块和第二指南模块中的知识特征。

[0158] 步骤S304:第一临床指南是否存在更新说明,若存在,执行步骤S305,若不存在,执行步骤S306。

[0159] 步骤S305:利用第一临床指南的更新说明,确定第一临床指南和第二临床指南之间的第一差异信息,并分别在第一临床指南和第二临床指南中与第一差异信息对应的位置处标注相应的标签,并执行步骤S309。

[0160] 步骤S306:将第一指南模块和第二指南模块进行匹配,将相匹配的第一指南模块和第二指南模块分别作为第一待处理指南模块和第二待处理指南模块,将与所有第二指南模块均不匹配的第一指南模块作为第三待处理指南模块。

[0161] 步骤S307:根据第一待处理指南模块和第二待处理指南模块,确定第一临床指南和第二临床指南之间的第二差异信息,并分别在第一临床指南和第二临床指南中与第二差异信息对应的位置处标注相应的标签。

[0162] 步骤S308:根据第三待处理指南模块和所有第二指南模块,确定第一临床指南和第二临床指南之间的第三差异信息,并分别在第一临床指南和第二临床指南中与第三差异信息对应的位置处标注相应的标签。

[0163] 步骤S309:在前端利用不同的展示形式,展示步骤S305或步骤S307至步骤S308所标注的标签。

[0164] 需要说明的是,上述步骤S301至步骤S309的执行原理,参见本发明实施例图1和图2中的内容,在此不再赘述。

[0165] 为更好解释说明上述本发明实施例图1至图3的内容,以具体的临床指南为例,通过过程A1至A7,对上述本发明实施例图1至图3的内容进行解释说明,其中过程A1至A7的内容仅用于举例说明。

[0166] 利用肾肿瘤临床指南的各级别标题,建立肾肿瘤临床指南对应的模块层级结构树,并制定模块层级结构树与各个指南模块的内容之间的映射规则。

[0167] A1、根据肾肿瘤临床指南对应的模块层级结构树,对非结构化的第一临床指南和第二临床指南进行解析和结构化提取,至少得到第一临床指南对应的第一指南模块和第二临床指南对应的第二指南模块。

[0168] 其中,针对肾细胞癌临床指南文档中的文字和图片内容,选择调用java开源工具Spire.PDF、Spire.Doc和PDFBox进行处理,对于肾细胞癌临床指南文档中的表格,选择调用Python开源库中的PDFPlumber进行处理。

[0169] 对所提取得到的临床指南(以此表征第一临床指南和第二临床指南)的文字内容以TXT和DOCX格式进行存储,图片内容以PNG格式存储,表格内容以CSV格式存储。同时,为便于将临床指南中各级别标题与正文内容有所区分和标识,将临床指南的各级别标题转存储入单独的TXT文件,对于临床指南中引用的参考文献也单独存储。将所提取的第一指南模块和第二指南模块存储至数据库中,及将所有第一指南模块之间的层级关系存储至数据库中,及将所有第二指南模块之间的层级关系存储至数据库中。

[0170] A2、对各个第一指南模块和第二指南模块进行文本清洗、去停用词、词干化、将英文数字转换为阿拉伯数字和将缩略语全称扩展等处理,将第一指南模块和第二指南模块转换成规范化文本。

[0171] 在具体实现中,文本清洗的具体处理方式为:基于正则表达式匹配规则,去除第一指南模块和第二指南模块中所有参考文献引用标识。

[0172] 停用词是指在文本中出现频率很高但不含有实际意义的单词,例如the、of、is、are和各种标点符号等,去除停用词的方式为通过词典匹配方法去除第一指南模块和第二指南模块中的停用词。

[0173] 由于肾细胞癌临床指南的书写语言为英语,故第一临床指南和第二临床指南中的单词存在复杂的词形变化,例如动词的各种时态、名词单复数和形容词的比较级,因此对第一临床指南和第二临床指南进行词干化处理,将第一临床指南和第二临床指南中单词的各种变体进行统一的标准化的,仅保留词干部分以提高第一临床指南和第二临床指南相似度计算结果的有效性。

[0174] 利用基于规则和由右至左的逆序扫描方法的缩略语识别方式,构建缩略语-全称映射表,并按照缩略语-全称映射表将第一临床指南和第二临床指南中的缩略语进行替换,将缩略语扩展为在第一临床指南和第二临床指南中意图指代的全称,如将缩略语“RCC”替换成“Renal Cell Carcinoma”这一全称。

[0175] A3、随机抽取一个版本的第一临床指南和第二临床指南,并对第一临床指南和第二临床指南进行知识特征的识别和提取、人工校对和效果评估。识别肾细胞癌临床指南(即第一临床指南和第二临床指南的类别)中的5类知识特征,该5类知识特征分别为:“临床表现”、“治疗方法”、“治疗药物(除二级类目“联合用药”外)”、“检查方法”和“疾病(除肾肿瘤)”,采用基于词典和规则相结合的方式进行知识特征的识别和提取。评估知识特征的识

别和提取的标准为:准确率 (precision,P),召回率 (Recall,R),准确率和召回率的调和平均数 (F-measure,F1),其中 $P = \text{正确识别并提取的知识特征数量} / \text{总识别并提取的知识特征数量} * 100\%$, $R = \text{正确识别并提取的知识特征数量} / \text{应识别并提取的知识特征数量} * 100\%$, $F1 = 2RP / (R+P) * 100\%$ 。

[0176] A4、可以理解的是,NCCN肾癌临床指南所附带的更新说明仅针对要点总结部分,分页阐述要点总结不同页面中图片和表格的变化情况,未直接涉及正文部分,因此仅基于要点总结页面对该更新说明进行拆分即可,不做详细后续处理。

[0177] 对于EAU肾细胞癌临床指南,该EAU肾细胞癌临床指南的正文肾细胞癌临床知识部分,除了分章节对肾细胞癌病理和诊疗相关知识进行说明以外,还有部分章节以表格的形式对相关重点证据和推荐进行梳理,并标注证据等级和推荐强度,形成章节要点总结,该章节要点总结的具体内容参见图4示出的EAU肾细胞癌临床指南章节要点总结的示意图。EAU肾细胞癌临床指南中的更新说明的形式为表格,分章节列出相较于上一版本EAU肾细胞癌临床指南各章节要点总结中新增的证据和推荐。

[0178] 上述内容为NCCN肾癌临床指南和EAU肾细胞癌临床指南的更新说明的相关内容。

[0179] 需要说明的是,在临床指南PDF文档的解析部分,已经将临床指南中更新说明内容和各章节要点总结分别进行抽取和转储为CSV格式表格,故可通过遍历临床指南的更新说明,在各章节要点总结中进行检索和定位,从而确定第一临床指南(此处为新版本临床指南)相较于第二临床指南(此处为旧版本临床指南)的新增内容,并在第一临床指南中与新增内容对应的位置处标注新增标签。对于第一临床指南相较于第二临床指南的删除内容或修改内容,则需要利用更新说明内容发现方法,基于相似度计算结果确定删除内容或修改内容,并进行相应的标注,具体如何标注参见上述本发明实施例图1步骤S102示出的内容,在此不再赘述。

[0180] 上述内容为根据临床指南的更新说明确定并标注第一临床指南和第二临床指南之间的差异信息,以下内容为不利用更新说明确定并标注第一临床指南和第二临床指南之间的差异信息(相当于上述本发明实施例图1步骤S103至步骤S105,及图2的内容)。

[0181] 可以理解的是,选取两个版本的NCCN肾癌临床指南,以及选取两个版本的EAU肾细胞癌临床指南,进行临床指南之间的比对(不利用更新说明确定两版本临床指南之间的差异和变化),其中句子相似度阈值设置为0.51。

[0182] A5、由上述过程A4可知,句子相似度阈值设置为0.51,根据上述本发明实施例图1步骤S103至步骤S105,及结合上述本发明实施例图2各步骤示出的内容,在不利用更新说明的情况下,采用正序限定遍历的方式,比对两个版本的NCCN肾癌临床指南之间的差异信息并进行相应标注,以及比对两个版本的EAU肾细胞癌临床指南之间的差异信息并进行相应标注。

[0183] 可以理解的是,可通过准确率、召回率和F1值等指标评估两个版本的NCCN肾癌临床指南之间的比对结果,以及评估两个版本的EAU肾细胞癌临床指南之间比对结果。

[0184] A6、上述过程A5在比对两版本的临床指南的过程中,对于上述本发明实施例图1步骤S103中提及的第三待处理指南模块(与所有第二指南模块均不匹配的第一指南模块),通过上述本发明实施例图1步骤S105示出的内容对该第三待处理指南模块进行处理和进行相应的标注,在此不再赘述。

[0185] A7、比对两版本的临床指南之后,以黄色突出显示文本的方式,标注出两版本临床指南中发生修改的部分(即修改标签),两版本临床指南中的黄色突出内容一一互相对应;以蓝色突出显示文本的方式,标注出新版本临床指南中新增的部分(即新增标签);以红色突出显示文本的方式,标注出旧版本临床指南中删除的部分(即删除标签)。

[0186] 为更好解释说明如何突出显示两版本的临床指南之间的差异信息,通过图5示出的标注更新标签的示意图进行举例说明。

[0187] 可以理解的是,图5中所示出的临床指南为两个版本的EAU肾细胞癌临床指南“流行病学”这一指南模块的正文之间的差异信息,其中,左侧为2016年版本的EAU肾细胞癌临床指南“流行病学”这一指南模块的正文,右侧为2018年版本的EAU肾细胞癌临床指南“流行病学”这一指南模块的正文。

[0188] 在图5中,以黄色突出显示文本的方式标注出的内容,为两版本EAU肾细胞癌临床指南中相似但不完全一致的内容;左侧2016年版本的EAU肾细胞癌临床指南中,以红色突出显示文本的方式标注出的内容,表示该部分内容为2016年版本的EAU肾细胞癌临床指南中删除的内容,此部分内容(2016年版本的EAU肾细胞癌临床指南中红色突出显示的内容)在2018年版本的EAU肾细胞癌临床指南中无相关描述;右侧2018年版本的EAU肾细胞癌临床指南中,以蓝色突出显示文本的方式标注出的内容,表示该部分内容为2018年版本的EAU肾细胞癌临床指南中新增的内容,此部分内容(2018年版本的EAU肾细胞癌临床指南中蓝色突出显示的内容)在2016年版本的EAU肾细胞癌临床指南中未出现。

[0189] 由上述本发明实施例图2步骤S204示出的内容可知,对于两版本临床指南(即步骤S204中的第一临床指南和第二临床指南)之间的修改内容,即对于两版本临床指南之间对应的黄色突出显示内容,两版本的修改内容之间的知识特征可能会出现变化,此时需要确定修改内容之间的知识特征差异信息并进行相应的标注,同时以不同的展示形式,展示不同类别的用于指示知识特征差异信息的标签。

[0190] 需要说明的是,由前述内容可知,肾细胞癌临床指南的书写语言为英语,故上述图5示出的临床指南的内容,以及以下图6至图9示出的临床指南的正文内容,均采用英文书写,

[0191] 进一步需要说明的是,图6至图8中的左侧为2016年版本的EAU肾细胞癌临床指南的部分内容,右侧为2018年版本的EAU肾细胞癌临床指南的部分内容。

[0192] 如图6示出的内容,在两版本临床指南的黄色突出显示内容中,以红色下划线及字体加粗表示实体知识特征的差异,即两版本临床指南中以红色下划线及字体加粗标注出的实体知识特征存在变化。

[0193] 如图7示出的内容,在两版本临床指南的黄色突出显示内容中,以青色下划线及字体加粗表示等级知识特征的差异,即两版本临床指南中以青色下划线及字体加粗标注出的等级知识特征存在变化。

[0194] 如图8示出的内容,在两版本临床指南的黄色突出显示内容中,以蓝色下划线及字体加粗表示数量知识特征的差异,以绿色下划线及字体加粗表示时间知识特征的差异,即两版本临床指南中以蓝色下划线及字体加粗标注出的数量知识特征存在变化,两版本临床指南中以绿色下划线及字体加粗标注出的时间知识特征存在变化。

[0195] 可以理解的是,展示更新时序可以从时间维度对不同版本临床指南间的实体知识

特征差异进行梳理,通过浏览各版本临床指南实体知识特征相较于上一版本或下一版本所发生的变化,能辅助阅读人员快速理清临床指南的知识更新脉络。以“复发、进展或转移性肾细胞癌的靶向治疗(Target Therapy of Relapsed or Advanced or Metastatic RCC)”这一指南模块内容为例,通过图9示出的更新时序的展示示意图对展示更新时序进行举例说明。

[0196] 在图9示出的内容中,相较于2015年版本的临床指南,2016年版本的临床指南的该“复发、进展或转移性肾细胞癌的靶向治疗”指南模块新增药物三种;相较于2016年版本的临床指南,2017年版本的临床指南的该“复发、进展或转移性肾细胞癌的靶向治疗”指南模块新增药物一种,删除药物五种;相较于2017年版本的临床指南,2018年版本的临床指南的该“复发、进展或转移性肾细胞癌的靶向治疗”指南模块新增药物一种。

[0197] 需要说明的是,图9中加粗显示的内容表示来源为临床指南的各级标题。

[0198] 与上述本发明实施例提供一种临床指南更新内容的自动识别方法相对应,参见图10,本发明实施例还提供了一种临床指南更新内容的自动识别系统的结构框图,该自动识别系统包括:解析单元100、第一处理单元110、第二处理单元120、第三处理单元130和第四处理单元140;

[0199] 解析单元100,用于根据预先利用临床指南各级别标题建立的模块层级结构树,分别对第一临床指南和第二临床指南进行解析和结构化提取,至少得到第一临床指南对应的第一指南模块和第二临床指南对应的第二指南模块,第一指南模块为第一临床指南中最小级别标题所包含的文本内容,第二指南模块为第二临床指南中最小级别标题所包含的文本内容。

[0200] 第一处理单元110,用于若第一临床指南和第二临床指南隶属同一来源,且当第一临床指南相对于第二临床指南存在更新说明,利用第一临床指南的更新说明,确定第一临床指南和第二临床指南之间的第一差异信息,并分别在第一临床指南和第二临床指南中与第一差异信息对应的位置处标注相应的标签,标签为新增标签、删除标签或修改标签。

[0201] 第二处理单元120,用于若第一临床指南和第二临床指南隶属不同来源,或者,若第一临床指南和第二临床指南隶属同一来源且第一临床指南不存在更新说明,将第一指南模块和第二指南模块进行匹配,将相匹配的第一指南模块和第二指南模块分别作为第一待处理指南模块和第二待处理指南模块,将与所有第二指南模块均不匹配的第一指南模块作为第三待处理指南模块。

[0202] 在具体实现中,第二处理单元120具体用于:针对每一个第一指南模块,利用预设的深度语义匹配模型,确定第一指南模块的标题和每个第二指南模块的标题之间的标题相似度;针对每一个第一指南模块,若所有标题相似度均小于标题相似度阈值,确定第一指南模块与所有第二指南模块均不匹配,若至少一个标题相似度大于等于标题相似度阈值,确定第一指南模块与最大标题相似度对应的第二指南模块相匹配。

[0203] 第三处理单元130,用于根据第一待处理指南模块和第二待处理指南模块,确定第一临床指南和第二临床指南之间的第二差异信息,并分别在第一临床指南和第二临床指南中与第二差异信息对应的位置处标注相应的标签。

[0204] 第四处理单元140,用于根据第三待处理指南模块和所有第二指南模块,确定第一临床指南和第二临床指南之间的第三差异信息,并分别在第一临床指南和第二临床指南中

与第三差异信息对应的位置处标注相应的标签。

[0205] 在具体实现中,第四处理单元140具体用于:计算第三待处理指南模块中首个句子的前P%内容与每个第二指南模块的多个第二句子之间的首句相似度;若至少一个首句相似度大于首句相似度阈值,确定第三待处理指南模块与最大首句相似度对应的第二指南模块相匹配;从第三待处理指南模块的首个句子的前P%内容处开始,将位于其之后的第一临床指南中已有的标签变更为修改标签;从与第三待处理指南模块相匹配的第二指南模块中最大首句相似度对应的第二句子处开始,将位于其之后的第二临床指南中已有的标签变更为修改标签。

[0206] 在本发明实施例中,根据预先利用临床指南各级别标题建立的模块层级结构树,分别对第一临床指南和第二临床指南进行解析和结构化提取,至少得到第一临床指南对应的第一指南模块和第二临床指南对应的第二指南模块;确定第一指南模块和第二指南模块之间的差异信息,并分别在第一临床指南和第二临床指南中与该差异信息对应的位置处标注相应的标签,不需要人工查阅需要比对的两个临床指南来找到不同临床指南间的差异和变化,提高确定不同临床指南间差异和变化情况的效率和准确度。

[0207] 优选的,结合图10示出的内容,第三处理单元130包括:分句子单元、计算子单元、第一标注子单元、第二标注子单元和第三标注子单元,各个子单元的执行原理如下:

[0208] 分句子单元,用于分别对第一待处理指南模块和第二待处理指南模块进行分句处理,得到第一待处理指南模块对应的多个第一句子,及得到第二待处理指南模块对应的多个第二句子。

[0209] 计算子单元,用于对于第一待处理指南模块的第m个第一句子,计算第m个第一句子与第二待处理指南模块的H个第二句子之间的句子相似度,m为大于等于1小于等于x的整数,x为第一待处理指南模块所包含的第一句子的总个数,m从1开始且递增1,H为大于等于1小于等于y的整数,y为第二待处理指南模块所包含的第二句子的总个数。

[0210] 第一标注子单元,用于若第m个第一句子和第n个第二句子之间的句子相似度等于1,确定第m个第一句子和第n个第二句子相同,不执行标注处理,n为大于等于1小于等于y的整数。

[0211] 第二标注子单元,用于若第m个第一句子和第n个第二句子之间的句子相似度大于等于句子相似度阈值且小于1,在第一临床指南中与第m个第一句子对应的位置处标注修改标签,及在第二临床指南中与第n个第二句子对应的位置处标注修改标签,当n大于m时,确定第二待处理指南模块中位于第n个第二句子之前的,与前m个第一句子的句子相似度均小于句子相似度阈值且未进行标注处理的第三句子,在第二临床指南中与第三句子对应的位置处标注删除标签。

[0212] 第三标注子单元,用于若第m个第一句子与H个第二句子之间的句子相似度均小于句子相似度阈值,在第一临床指南中与第m个第一句子对应的位置处标注新增标签。

[0213] 优选的,结合图10示出的内容,该自动识别系统还包括:

[0214] 预处理单元,用于分别对第一指南模块和第二指南模块进行预处理,并分别提取进行预处理后的第一指南模块和所述第二指南模块中的知识特征。

[0215] 相应的,第二标注子单元还用于:比对第m个第一句子和第n个第二句子内的知识特征之间的差异,得到知识特征差异信息;分别在第m个第一句子中和第n个第二句子中与

知识特征差异信息对应的位置处标注相应的标签。

[0216] 在本发明实施例中,从指南模块、句子和知识特征3个维度,确定第一临床指南和所述第二临床指南之间的第二差异信息,并分别在第一临床指南和第二临床指南中与第二差异信息对应的位置处标注相应的标签,不需要人工查阅需要比对的两个临床指南来找到不同临床指南间的差异和变化,提高确定不同临床指南间差异和变化情况的效率和准确度。以及基于时间序列进行知识特征差异信息的统计,挖掘临床指南中知识特征的时序变化,将静态和非结构化的临床指南转化为结构化、知识特征化和可视化的表示形式,实现对临床指南的更新内容的多层次和多维度揭示,以辅助临床医生直观了解临床指南之间的差异和变化情况,提高临床医生的学习效率。

[0217] 优选的,结合图10示出的内容,该自动识别系统还包括:

[0218] 展示单元,用于分别利用不同的展示形式,展示不同类别的所述标签。

[0219] 优选的,结合图10示出的内容,该自动识别系统还包括:

[0220] 规范化单元,用于对第一指南模块和第二指南模块进行规范化处理。

[0221] 优选的,结合图10示出的内容,该自动识别系统还包括:

[0222] 存储单元,用于将第一指南模块和第二指南模块存储至数据库中,及将所有第一指南模块之间的层级关系存储至数据库中,及将所有第二指南模块之间的层级关系存储至数据库中。

[0223] 综上所述,本发明实施例提供一种临床指南更新内容的自动识别方法及系统,根据预先利用临床指南各级别标题建立的模块层级结构树,分别对第一临床指南和第二临床指南进行解析和结构化提取,至少得到第一临床指南对应的第一指南模块和第二临床指南对应的第二指南模块;确定第一指南模块和第二指南模块之间的差异信息,并分别在第一临床指南和第二临床指南中与该差异信息对应的位置处标注相应的标签,不需要人工查阅需要比对的两个临床指南来找到不同临床指南间的差异和变化,提高确定不同临床指南间差异和变化情况的效率和准确度。

[0224] 本说明书中的各个实施例均采用递进的方式描述,各个实施例之间相同相似的部分互相参见即可,每个实施例重点说明的都是与其他实施例的不同之处。尤其,对于系统或系统实施例而言,由于其基本相似于方法实施例,所以描述得比较简单,相关之处参见方法实施例的部分说明即可。以上所描述的系统及系统实施例仅仅是示意性的,其中所述作为分离部件说明的单元可以是或者也可以不是物理上分开的,作为单元显示的部件可以是或者也可以不是物理单元,即可以位于一个地方,或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部模块来实现本实施例方案的目的。本领域普通技术人员在不付出创造性劳动的情况下,即可以理解并实施。

[0225] 专业人员还可以进一步意识到,结合本文中所公开的实施例描述的各示例的单元及算法步骤,能够以电子硬件、计算机软件或者二者的结合来实现,为了清楚地说明硬件和软件的可互换性,在上述说明中已经按照功能一般性地描述了各示例的组成及步骤。这些功能究竟以硬件还是软件方式来执行,取决于技术方案的特定应用和设计约束条件。专业技术人员可以对每个特定的应用来使用不同方法来实现所描述的功能,但是这种实现不应认为超出本发明的范围。

[0226] 对所公开的实施例的上述说明,使本领域专业技术人员能够实现或使用本发明。

对这些实施例的多种修改对本领域的专业技术人员来说将是显而易见的,本文中所定义的一般原理可以在不脱离本发明的精神或范围的情况下,在其它实施例中实现。因此,本发明将不会被限制于本文所示的这些实施例,而是要符合与本文所公开的原理和新颖特点相一致的最宽的范围。

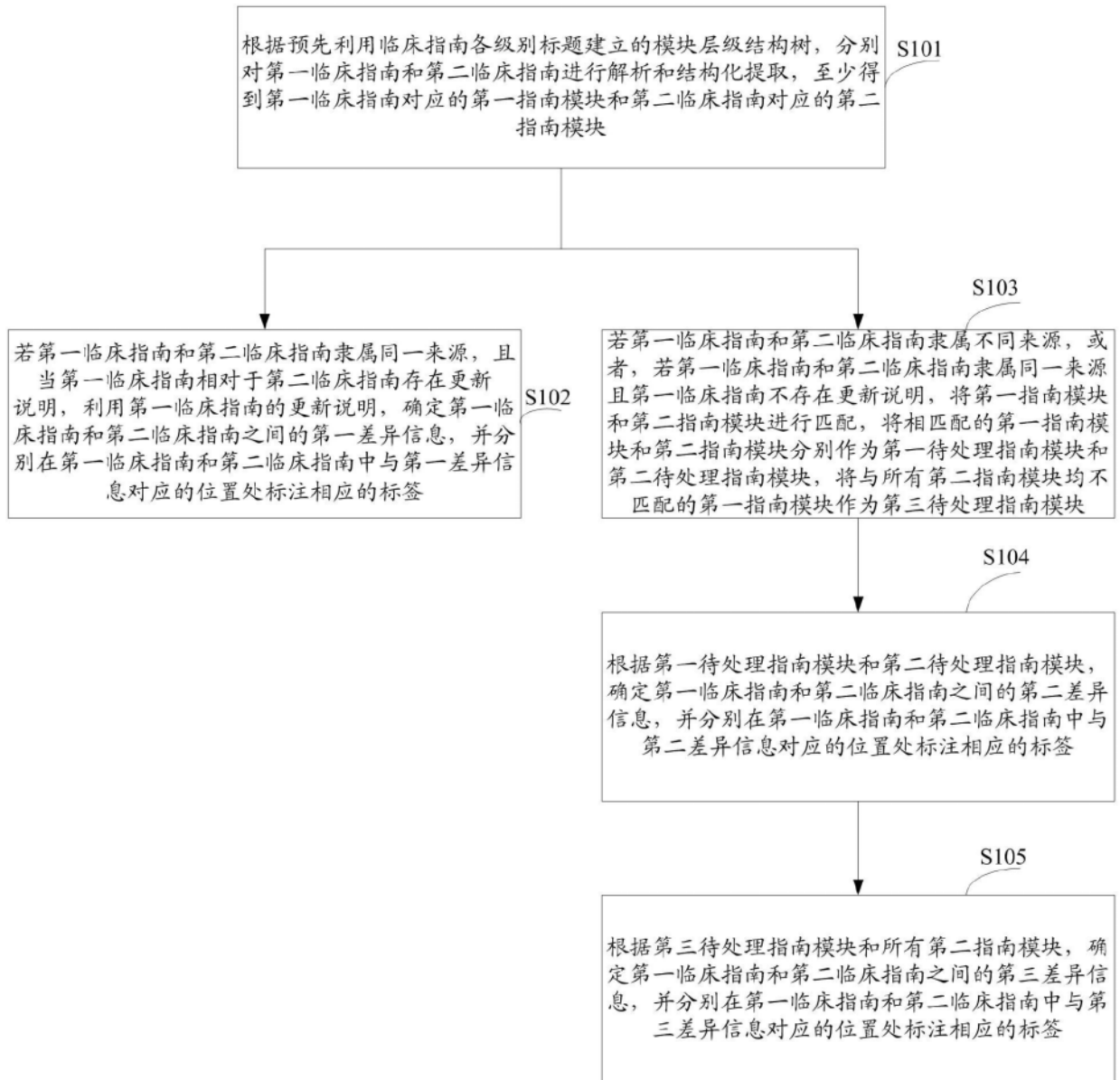


图1

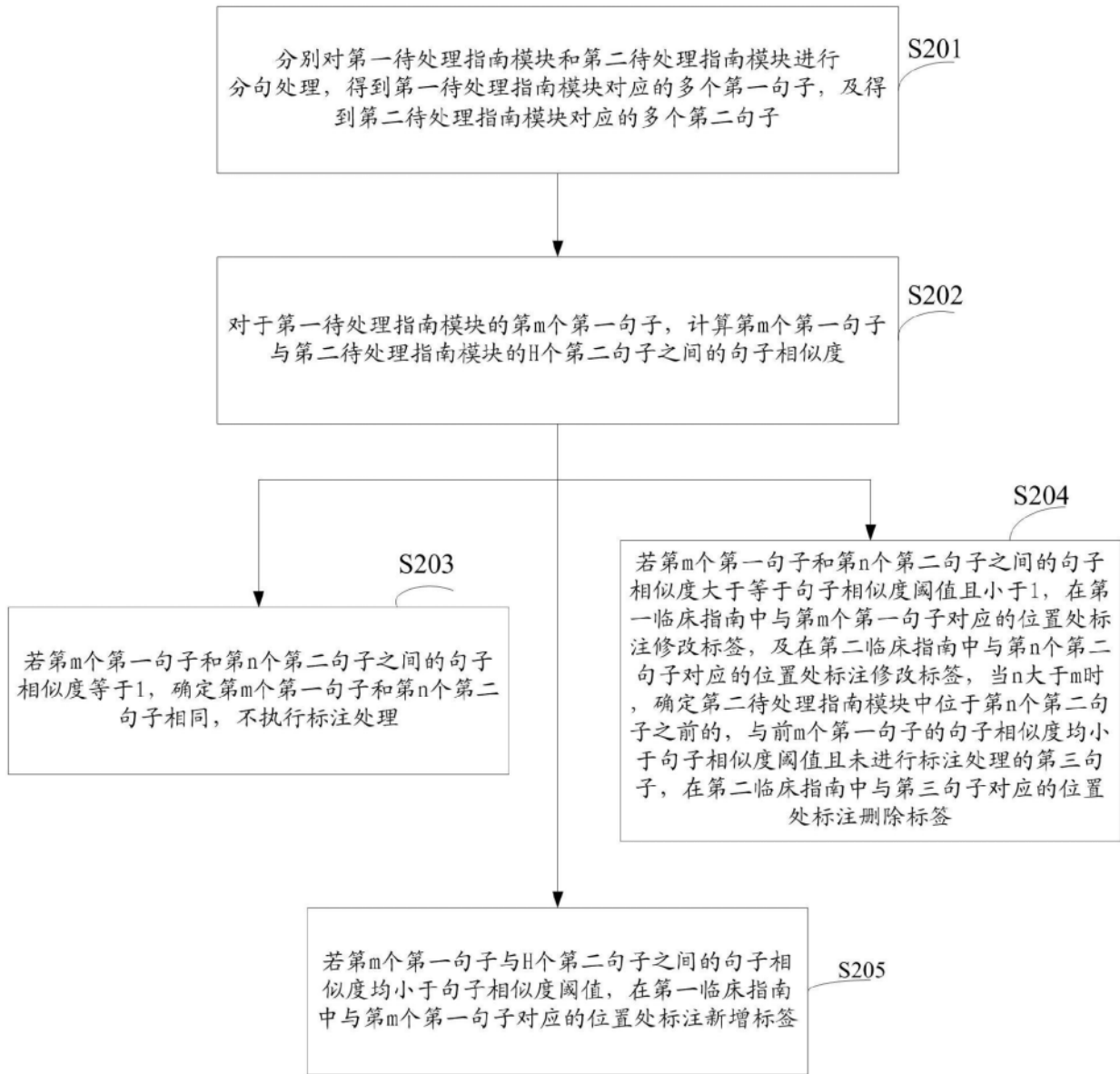


图2

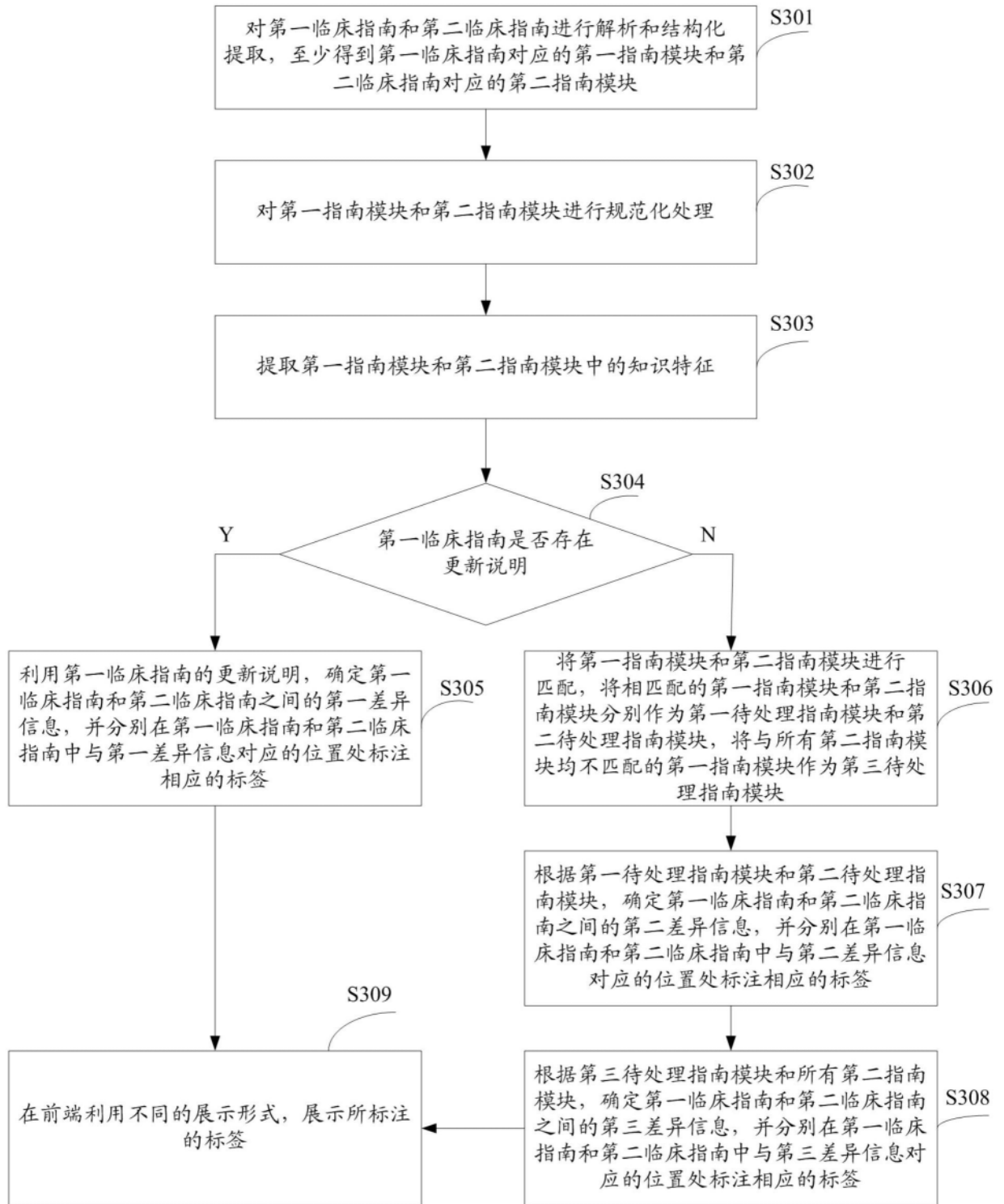


图3

5.4 Summary of evidence and recommendations for the diagnostic assessment of renal cell cancer	
Summary of evidence	LE
Contrast enhanced multi-phasic computed tomography (CT) has a high sensitivity and specificity for characterisation and detection of RCC, invasion, tumour thrombus and metastatic RCC.	2
Magnetic resonance imaging (MRI) has a slightly higher sensitivity and specificity for small cystic renal masses and tumour thrombus as compared to CT.	2
Contrast enhanced ultrasound (CEUS) has a high sensitivity and specificity for characterisation of renal masses.	2
Ultrasound, power-doppler US and positron-emission tomography (PET) CT have a low sensitivity and specificity for detection and characterisation of RCC.	2
Recommendations	Strength rating
Use multi-phasic contrast-enhanced computed tomography (CT) of abdomen and chest for the diagnosis and staging of renal tumours.	Strong
Use MRI to better evaluate venous involvement, reduce radiation or avoid intravenous CT contrast medium.	Weak
Use non-ionising modalities, mainly contrast enhanced ultrasound (CEUS), for further characterisation of small renal masses, tumour thrombus and differentiation of unclear renal masses.	Weak
Do not routinely use bone scan and/or positron-emission tomography (PET) CT for staging of RCC.	Weak
Perform a renal tumour biopsy before ablative therapy and systemic therapy without previous pathology.	Strong
Perform a percutaneous biopsy in select patients who are considered for active surveillance.	Weak
Use a coaxial technique when performing a renal tumour biopsy.	Strong
Do not perform a renal tumour biopsy of cystic renal masses.	Strong
Use a core biopsy technique rather than fine needle aspiration for histological characterisation for solid renal tumours.	Strong

图4

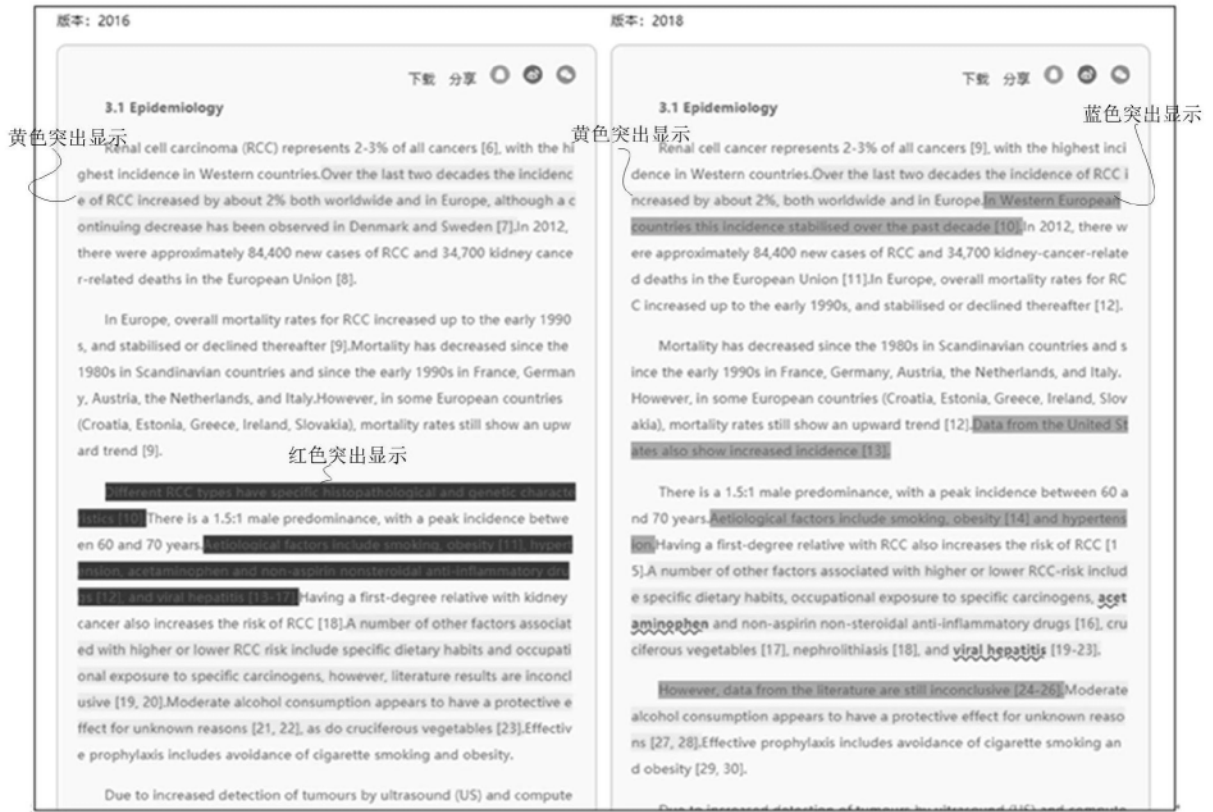


图5

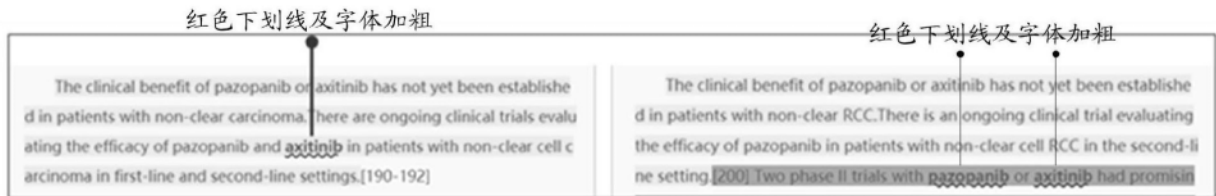


图6



图7



图8

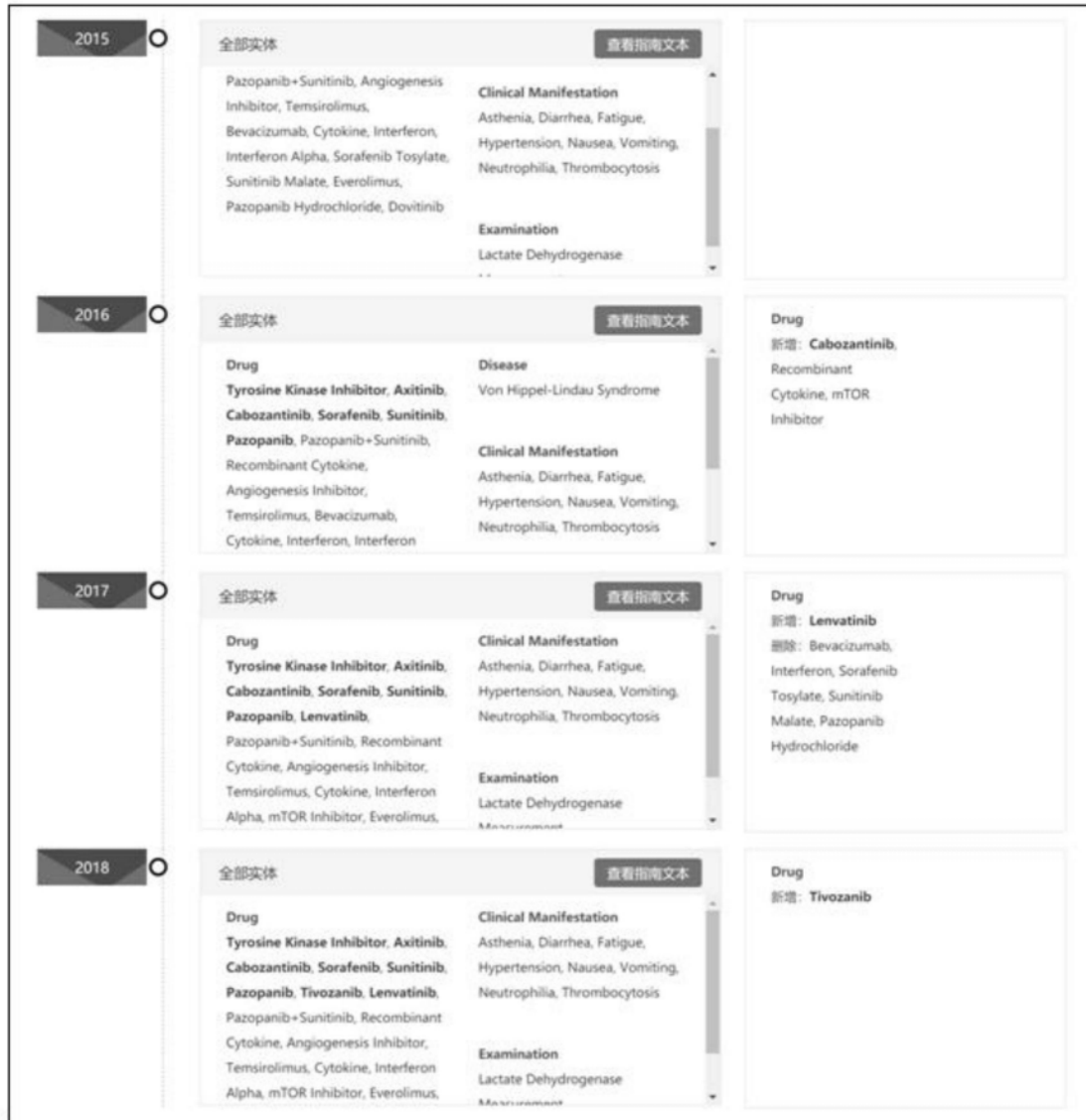


图9

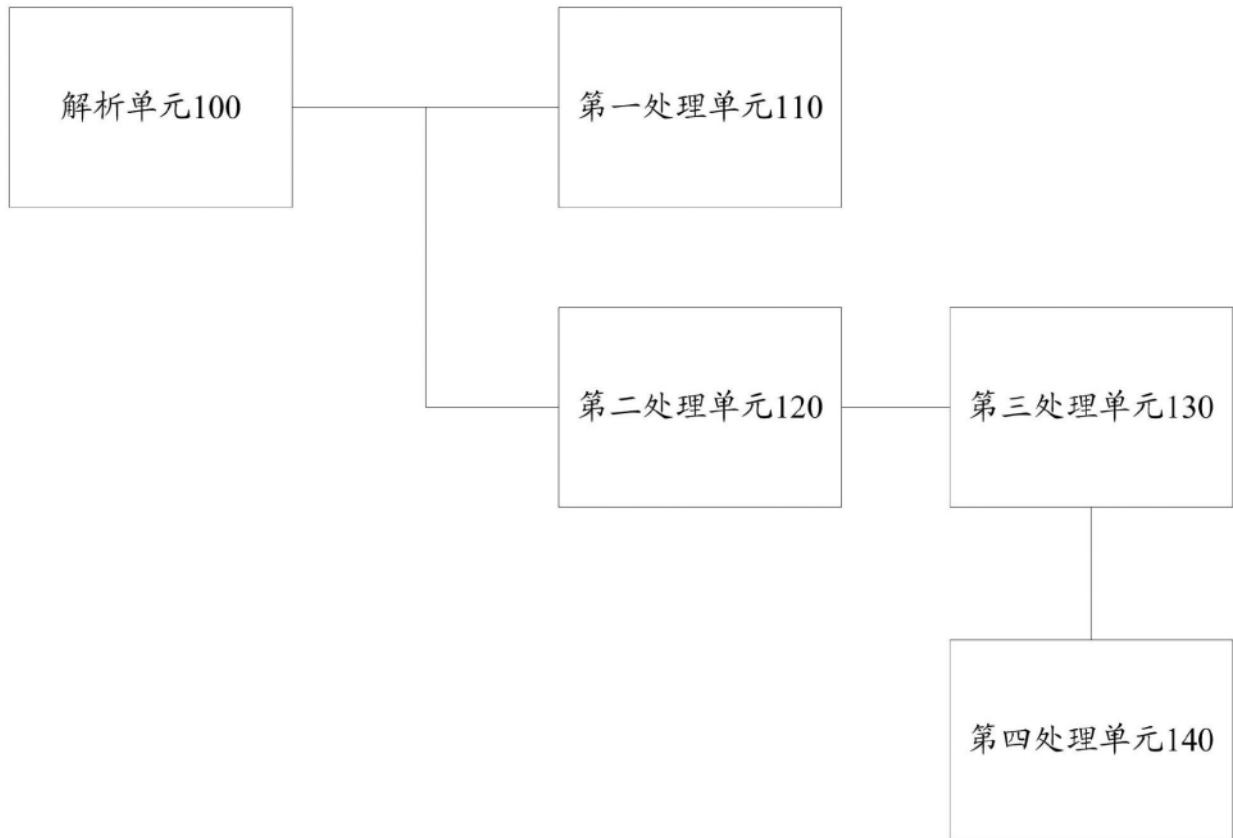


图10