



(21) 申请号 202610644648.5

G06N 5/025 (2023.01)

(22) 申请日 2026.05.11

G06N 5/04 (2023.01)

(71) 申请人 中煤科工开采研究院有限公司

地址 101399 北京市顺义区中关村科技园
区顺义园临空二路1号

(72) 发明人 吕依濛

(74) 专利代理机构 北京清亦华知识产权代理事

务所(普通合伙) 11201

专利代理师 金星玉

(51) Int. Cl.

G06F 18/20 (2023.01)

G06F 18/25 (2023.01)

G06F 40/194 (2020.01)

G06F 40/30 (2020.01)

G06F 40/205 (2020.01)

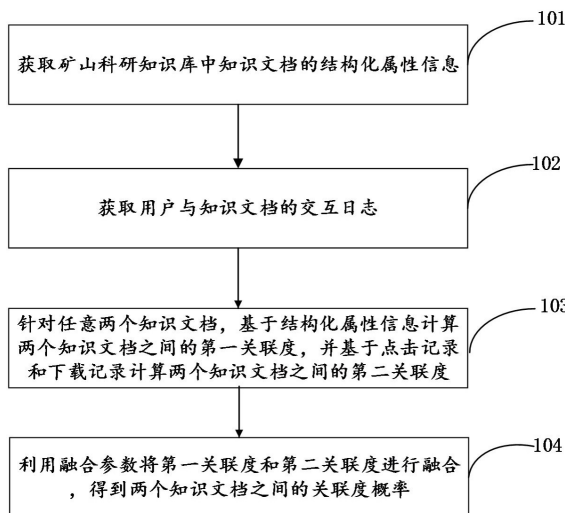
权利要求书2页 说明书10页 附图2页

(54) 发明名称

应用于矿山科研知识服务平台的知识文档
关联度确定方法

(57) 摘要

本公开是关于一种应用于矿山科研知识服务平台的知识文档关联度确定方法。其中,方法包括:获取矿山科研知识库中知识文档的结构化属性信息,以及用户对知识文档的点击记录和下载记录,针对任意两个知识文档,分别基于结构化属性信息计算两个知识文档之间的第一关联度,并基于点击记录和下载记录计算两个知识文档之间的第二关联度,进而利用融合参数将第一关联度和第二关联度进行融合,得到两个知识文档之间的关联度概率。本方案可以实现更全面、准确地反映矿山科研知识文档之间的实际关联程度。



1. 一种应用于矿山科研知识服务平台的知识文档关联度确定方法,其特征在于,包括:
 获取矿山科研知识库中知识文档的结构化属性信息;
 获取用户与所述知识文档的交互日志;所述交互日志包括用户对知识文档的点击记录和下载记录;

针对任意两个知识文档,基于所述结构化属性信息计算两个知识文档之间的第一关联度,并基于所述点击记录和下载记录计算两个知识文档之间的第二关联度;

利用融合参数将所述第一关联度和所述第二关联度进行融合,得到两个知识文档之间的关联度概率。

2. 根据权利要求1所述的应用于矿山科研知识服务平台的知识文档关联度确定方法,其特征在于,所述基于所述结构化属性信息计算两个知识文档之间的第一关联度,包括:

针对每一种属性 n ,计算两个知识文档 a 和 b 在该属性上的余弦相似度 $\text{similarity}(\text{attribute}[a,n], \text{attribute}[b,n], n)$;

获取各属性的重要度权重 $Q(n)$,且 $\sum Q(n)=1$;

采用以下公式计算第一关联度 $\text{similarity_compare}(a,b)$:

$$\text{similarity_compare}(a,b) = \frac{\sum_n \text{similarity}(\text{attribute}[a,n], \text{attribute}[b,n], n) * Q(n) * 1_{\{\text{similarity}(\text{attribute}[a,n], \text{attribute}[b,n], n) > \beta\}}}{\sum_n Q(n)}$$

其中, β 为预设相似度阈值。

3. 根据权利要求1所述的应用于矿山科研知识服务平台的知识文档关联度确定方法,其特征在于,所述基于所述点击记录和下载记录计算两个知识文档之间的第二关联度,包括:

设定多个时间窗口 t_i ,每个时间窗口对应一个时段权重 $S(i)$,且 $\sum S(i)=1$;

在每个时间窗口 t_i 内:

统计同一用户连续点击文档 a 和 b 的次数 $\text{scan}(a,b,t_i)$,以及同一用户任意连续点击两个知识文档的总次数 $\text{scan}(t_i)$;

统计同一用户连续下载文档 a 和 b 的次数 $\text{download}(a,b,t_i)$,以及同一用户任意连续下载两个知识文档的总次数 $\text{download}(t_i)$;

统计同一用户连续点击一个文档且下载另一个文档的次数 $\text{mix}(a,b,t_i)$,以及同一用户任意连续点击和下载两个知识文档的总次数 $\text{mix}(t_i)$;

按照下式计算两个知识文档 a 和 b 的点击下载关联度 $\text{mix}(a,b)$:

$$\text{mix}(a,b) = \frac{\sum_i \frac{\text{scan}(a,b,t_i) * R_1 + \text{download}(a,b,t_i) * R_2 + \text{mix}(a,b,t_i) * R_3}{\text{scan}(t_i) + \text{download}(t_i) + \text{mix}(t_i)} * S(i)}{\sum_i S(i)}$$

其中, R_1 为 $\text{scan}(t_i)$ 对应的权重, R_2 为 $\text{download}(t_i)$ 对应的权重, R_3 为 $\text{mix}(t_i)$ 对应的权重, $R_1+R_2+R_3=1$ 。

4. 根据权利要求1所述的应用于矿山科研知识服务平台的知识文档关联度确定方法,其特征在于,所述利用融合参数将所述第一关联度和所述第二关联度进行融合,得到两个

知识文档之间的关联度概率,包括:

采用以下公式计算两个知识文档之间的关联度概率 $P_{pre}(a,b)$:

$$P_{pre}(a,b) = \alpha * similarity_compare(a,b) + (1-\alpha) * mix(a,b)$$

其中, α 为预设的融合参数, $0 \leq \alpha \leq 1$, $similarity_compare(a,b)$ 为第一关联度, $mix(a,b)$ 为第二关联度。

5. 根据权利要求3所述的应用于矿山科研知识服务平台的知识文档关联度确定方法, 其特征在于, 所述时间窗口包括多个不同时间粒度的窗口, 所述时段权重 $S(i)$ 为归一化权重, 其中, 同一时间窗口对应的时间粒度与时段权重取值成反比。

6. 一种应用于矿山科研知识服务平台的知识文档关联度预测装置, 其特征在于, 包括:

第一获取单元, 用于获取矿山科研知识库中知识文档的结构化属性信息;

第二获取单元, 用于获取用户与所述知识文档的交互日志; 所述交互日志包括用户对知识文档的点击记录和下载记录;

计算单元, 用于针对任意两个知识文档, 基于所述结构化属性信息计算两个知识文档之间的第一关联度, 并基于所述点击记录和下载记录计算两个知识文档之间的第二关联度;

融合单元, 用于利用融合参数将所述第一关联度和所述第二关联度进行融合, 得到两个知识文档之间的关联度概率。

7. 根据权利要求6所述的应用于矿山科研知识服务平台的知识文档关联度确定装置, 其特征在于, 所述计算单元具体用于:

针对每一种属性 n , 计算两个知识文档 a 和 b 在该属性上的余弦相似度 $similarity(attribute[a,n], attribute[b,n], n)$;

获取各属性的重要度权重 $Q(n)$, 且 $\sum Q(n)=1$;

采用以下公式计算第一关联度 $similarity_compare(a,b)$:

$$similarity_compare(a,b) = \frac{\sum_n similarity(attribute[a,n], attribute[b,n], n) * Q(n) * 1_{\{similarity(attribute[a,n], attribute[b,n], n) > \beta\}}}{\sum_n Q(n)}$$

其中, β 为预设相似度阈值。

8. 一种电子设备, 其特征在于, 包括: 存储器、处理器及存储在所述存储器上并可在所述处理器上运行的计算机程序, 所述处理器执行所述计算机程序时, 实现如权利要求1至5中任一项所述的方法。

9. 一种计算机可读存储介质, 其上存储有计算机程序, 其特征在于, 所述计算机程序被处理器执行时实现如权利要求1至5中任一项所述的方法。

10. 一种计算机程序产品, 包括计算机程序, 其特征在于, 所述计算机程序在被处理器执行时实现如权利要求1至5中任一项所述的方法。

应用于矿山科研知识服务平台的知识文档关联度确定方法

技术领域

[0001] 本公开涉及数据处理技术领域,尤其涉及一种应用于矿山科研知识服务平台的知识文档关联度确定方法。

背景技术

[0002] 相关技术中,随着企业及科研机构内部知识库规模的持续扩大,其中存储的技术报告、专利、论文等知识文档数量激增,且多为非结构化数据,文档之间难以建立有效的语义关联。传统方法通常基于文档标题、摘要、关键词等内容特征进行相似度分析,或利用文档自身包含的作者、发表时间、来源渠道等属性信息进行匹配,以识别文档间的潜在联系。然而,这些方法主要依赖文档的静态特征,难以充分挖掘文档之间的深层次关系。

发明内容

[0003] 为克服相关技术中存在的问题,本公开提供一种应用于矿山科研知识服务平台的知识文档关联度确定方法。

[0004] 根据本公开实施例的第一方面,提供一种应用于矿山科研知识服务平台的知识文档关联度确定方法,包括:

[0005] 获取矿山科研知识库中知识文档的结构化属性信息;

获取用户与所述知识文档的交互日志;所述交互日志包括用户对知识文档的点击记录和下载记录;

针对任意两个知识文档,基于所述结构化属性信息计算两个知识文档之间的第一关联度,并基于所述点击记录和下载记录计算两个知识文档之间的第二关联度;

利用融合参数将所述第一关联度和所述第二关联度进行融合,得到两个知识文档之间的关联度概率。

[0006] 根据本公开实施例的第二方面,提供一种应用于矿山科研知识服务平台的知识文档关联度预测装置,包括:

第一获取单元,用于获取矿山科研知识库中知识文档的结构化属性信息;

第二获取单元,用于获取用户与所述知识文档的交互日志;所述交互日志包括用户对知识文档的点击记录和下载记录;

计算单元,用于针对任意两个知识文档,基于所述结构化属性信息计算两个知识文档之间的第一关联度,并基于所述点击记录和下载记录计算两个知识文档之间的第二关联度;

融合单元,用于利用融合参数将所述第一关联度和所述第二关联度进行融合,得到两个知识文档之间的关联度概率。

[0007] 根据本公开实施例的第三方面,一种电子设备,包括:存储器、处理器及存储在所述存储器上并可在所述处理器上运行的计算机程序,所述处理器执行所述计算机程序时,实现如第一方面中任一项所述的方法。

[0008] 根据本公开实施例的第四方面,提供一种计算机可读存储介质,其上存储有计算机程序,所述计算机程序被处理器执行时实现如第一方面中任一项所述的方法。

[0009] 根据本公开实施例的第五方面,提供一种计算机程序产品,包括计算机程序,所述计算机程序在被处理器执行时实现如第一方面中任一项所述的方法。

[0010] 本公开的实施例提供的技术方案可以包括以下有益效果:通过获取矿山科研知识库中知识文档的结构化属性信息,以及用户对知识文档的点击记录和下载记录,针对任意两个知识文档,分别基于结构化属性信息计算两个知识文档之间的第一关联度,并基于点击记录和下载记录计算两个知识文档之间的第二关联度,进而利用融合参数将第一关联度和第二关联度进行融合,得到两个知识文档之间的关联度概率;其有益效果在于:一方面,通过结构化属性信息计算第一关联度,能够从文档自身特征维度挖掘文档间的内在联系,另一方面,通过点击记录和下载记录计算第二关联度,能够从用户行为维度捕捉文档间的潜在关联,二者融合得到的关联度概率综合了静态属性与动态行为的多维信息,从而更全面、准确地反映矿山科研知识文档之间的实际关联程度,为后续知识图谱构建、智能推荐或科研价值评估提供可靠的关联数据基础。

[0011] 应当理解的是,以上的一般描述和后文的细节描述仅是示例性和解释性的,并不能限制本公开。

附图说明

[0012] 此处的附图被并入说明书中并构成本说明书的一部分,示出了符合本发明的实施例,并与说明书一起用于解释本发明的原理。

[0013] 图1是根据一示例性实施例示出的一种应用于矿山科研知识服务平台的知识文档关联度确定方法的流程图。

[0014] 图2是根据一示例性实施例示出的一种应用于矿山科研知识服务平台的知识文档关联度预测装置的框图。

[0015] 图3是根据一示例性实施例示出的一种用于应用于矿山科研知识服务平台的知识文档关联度确定方法的装置的框图。

具体实施方式

[0016] 这里将详细地对示例性实施例进行说明,其示例表示在附图中。下面的描述涉及附图时,除非另有表示,不同附图中的相同数字表示相同或相似的要素。以下示例性实施例中所描述的实施方式并不代表与本发明相一致的所有实施方式。相反,它们仅是与如所附权利要求书中所详述的、本发明的一些方面相一致的装置和方法的例子。

[0017] 在本公开实施例使用的术语是仅仅出于描述特定实施例的目的,而非旨在限制本公开实施例。在本公开实施例和所附权利要求书中所使用的单数形式的“一种”和“该”也旨在包括多数形式,除非上下文清楚地表示其他含义。

[0018] 应当理解,尽管在本公开实施例可能采用术语第一、第二、第三等来描述各种信息,但这些信息不应限于这些术语。这些术语仅用来将同一类型的信息彼此区分开。例如,在不脱离本公开实施例范围的情况下,第一信息也可以被称为第二信息,类似地,第二信息也可以被称为第一信息。取决于语境,如在此所使用的词语“如果”及“若”可以被解释成为

“在……时”或“当……时”或“响应于确定”。

[0019] 此外,可以使用本公开实施例所示的各种形式的流程,重新排序、增加或删除步骤。例如,本公开中记载的各步骤可以并行地执行也可以顺序地执行也可以不同的次序执行,只要能够实现本公开公开的技术方案所期望的结果,本文在此不进行限制。

[0020] 需要说明的是,本公开的技术方案中,所涉及的用户个人信息的收集、存储、使用、加工、传输、提供和公开等处理,均符合相关法律法规的规定,且不违背公序良俗。

[0021] 图1是根据一示例性实施例示出的一种应用于矿山科研知识服务平台的知识文档关联度确定方法的流程图,如图1所示,需要说明的是,本公开实施例的应用于矿山科研知识服务平台的知识文档关联度确定方法应用于应用于矿山科研知识服务平台的知识文档关联度预测装置中。如图1所示,该方法可以包括以下步骤:

步骤101,获取矿山科研知识库中知识文档的结构化属性信息。

[0022] 在本公开一些实施例中,矿山科研知识库中存储有大量的知识文档,包括但不限于技术报告、专利、论文、软件著作权等。这些知识文档最初可能以非结构化的形式存在,例如PDF文件、Word文档等。

[0023] 作为一种示例,可以对这些知识文档进行预处理,例如删除冗余文本、错误字符等,并从中提取结构化的属性信息。

[0024] 具体地,可以提取的属性信息可以包括作者、文件名、创建时间、渠道来源等,将这些信息作为知识文档的结构化数据存储起来,形成文档的属性集合。

[0025] 可以理解的是,这些结构化属性信息为后续计算文档之间的关联度提供了基础数据。通过提取结构化属性信息,可以将非结构化的文档转化为结构化的数据表示,便于计算机进行分析和处理。

[0026] 步骤102,获取用户与知识文档的交互日志。

[0027] 其中,所述交互日志包括用户对知识文档的点击记录和下载记录。

[0028] 作为一种示例,知识服务平台会记录用户与知识文档之间的交互行为,形成交互日志。

[0029] 在本公开实施例中,获取这些交互日志,其中至少包括用户对知识文档的点击记录和下载记录。点击记录反映了用户对文档的关注或浏览行为,下载记录则反映了用户对文档的更深层次需求,可以表明该文档对用户具有较高的价值。

[0030] 可以理解的是,用户的交互行为隐含着文档之间的潜在关联,例如同一用户在短时间内连续点击或下载两篇文档,可能意味着这两篇文档在内容或主题上具有相关性。通过获取点击记录和下载记录,可以从用户行为维度挖掘文档间的联系。

[0031] 步骤103,针对任意两个知识文档,基于结构化属性信息计算两个知识文档之间的第一关联度,并基于点击记录和下载记录计算两个知识文档之间的第二关联度。

[0032] 在本公开实施例中,针对任意两个知识文档a和b,一方面基于结构化属性信息计算第一关联度,即通过比较文档在各属性维度上的相似性,并综合考虑不同属性的重要程度以及预设的相似度阈值,从文档自身特征层面量化两者之间的内在联系;另一方面基于点击记录和下载记录计算第二关联度,即通过统计用户在不同时间窗口内的连续点击、连续下载以及点击-下载混合行为,并综合考虑不同行为类型的重要程度以及不同时间粒度的权重,从用户行为层面挖掘两者之间的潜在关联。通过上述两个维度的并行计算,为后续

融合得到最终的关联度概率提供了基础。

[0033] 在本公开一些实施例中,步骤103具体可以包括以下子步骤:

步骤a1,针对每一种属性 n ,计算两个知识文档 a 和 b 在该属性上的余弦相似度 $\text{similarity}(\text{attribute}[a,n], \text{attribute}[b,n], n)$ 。

[0034] 在本公开实施例中,对于任意两个知识文档 a 和 b ,需要逐一比较它们在各属性上的相似程度。余弦相似度是一种常用的向量相似度度量方法,通过计算两个向量夹角的余弦值来度量它们的相似程度,值域为 $[-1,1]$,1表示完全相同,0表示正交(不相关),-1表示完全相反。对于每个属性 n ,将文档 a 和 b 在该属性上的取值转化为向量形式,计算两个知识文档之间的余弦相似度。通过这种方式,可以得到文档 a 和 b 在各个属性维度上的相似度得分。

[0035] 步骤a2,获取各属性的重要度权重 $Q(n)$,且 $\sum Q(n)=1$ 。

[0036] 在本公开一些实施例中,不同的属性对于判断文档间关联的重要程度可能不同。例如,在矿山科研领域,作者属性可能比文件名属性更能反映文档的专业相关性。因此,本公开根据知识服务平台的实际使用数据统计,为每个属性 n 赋予一个重要度权重 $Q(n)$,且所有属性的权重之和为1。权重 $Q(n)$ 的值越大,代表该属性对于判断文档间关联的重要度越大。通过设置权重,可以体现不同属性在关联度计算中的差异化贡献。

[0037] 步骤a3,采用以下公式计算第一关联度 $\text{similarity_compare}(a,b)$:

$$\text{similarity_compare}(a,b) = \frac{\sum_n \text{similarity}(\text{attribute}[a,n], \text{attribute}[b,n], n) * Q(n) * 1_{\{\text{similarity}(\text{attribute}[a,n], \text{attribute}[b,n], n) > \beta\}}}{\sum_n Q(n)}$$

其中, β 为预设相似度阈值。

[0038] 在本公开一些实施例中,为了确保只有具备一定意义关联的属性才被纳入计算,设定阈值 β 。当某个属性的余弦相似度大于 β 时,该属性的相似度被计入第一关联度的计算中。计算公式中的分子是对所有有效属性(即相似度大于 β 的属性)的余弦相似度进行加权求和,分母是所有权重之和(由于 $\sum Q(n)=1$,分母实际为1,但保留分母形式更一般化)。通过这种方式,第一关联度 $\text{similarity_compare}(a,b)$ 综合了多个属性的相似度信息,并考虑了属性的重要程度和有效性阈值,从而更准确地反映文档 a 和 b 在静态属性层面的关联程度。

[0039] 可以理解的是,第一关联度从文档自身特征维度挖掘文档间的内在联系,为后续融合提供了基础。

[0040] 步骤b1,设定多个时间窗口 t_i ,每个时间窗口对应一个时段权重 $S(i)$,且 $\sum S(i)=1$ 。

[0041] 在本公开实施例中,用户交互行为的发生时间对于判断文档关联具有重要意义。例如,用户在极短时间内连续点击两篇文档,其关联强度远大于在较长时间内分别点击。因此,本公开设定多个时间窗口 t_i ,如1秒内、1分钟内、1天内、一个月内等。每个时间窗口对应一个时段权重 $S(i)$,且所有时段权重之和为1。时段权重的设定原则是:时间窗口越小,表明用户行为在越短的时间内发生,关联性越强,因此时段权重取值越大。通过时段权重,可以体现不同时间粒度下的行为关联强度差异。

[0042] 步骤b2,在每个时间窗口 t_i 内:

统计同一用户连续点击文档 a 和 b 的次数 $\text{scan}(a,b,t_i)$,以及同一用户任意连续点

击两个知识文档的总次数 $\text{scan}(ti)$;

统计同一用户连续下载文档a和b的次数 $\text{download}(a,b,ti)$,以及同一用户任意连续下载两个知识文档的总次数 $\text{download}(ti)$;

统计同一用户连续点击一个文档且下载另一个文档的次数 $\text{mix}(a,b,ti)$,以及同一用户任意连续点击和下载两个知识文档的总次数 $\text{mix}(ti)$ 。

[0043] 在本公开一些实施例中,对于每个时间窗口 ti ,分别统计三种类型的用户行为:

点击-点击行为:同一用户连续点击文档a和b的次数记为 $\text{scan}(a,b,ti)$;同一用户任意连续点击两个知识文档的总次数(即所有文档对被同一用户连续点击的总次数)记为 $\text{scan}(ti)$ 。

[0044] 下载-下载行为:同一用户连续下载文档a和b的次数记为 $\text{download}(a,b,ti)$;同一用户任意连续下载两个知识文档的总次数记为 $\text{download}(ti)$ 。

[0045] 点击-下载混合行为:同一用户连续点击一个文档且下载另一个文档的次数(顺序不限)记为 $\text{mix}(a,b,ti)$;同一用户任意连续点击和下载两个知识文档的总次数(即所有文档对被同一用户连续点击和下载的总次数)记为 $\text{mix}(ti)$ 。

[0046] 可以理解的是,这些统计量为后续计算条件概率提供了基础。 $\text{scan}(a,b,ti)/\text{scan}(ti)$ 反映了在点击行为中,文档a和b被连续点击的概率;类似地, $\text{download}(a,b,ti)/\text{download}(ti)$ 反映了在下载行为中,文档a和b被连续下载的概率; $\text{mix}(a,b,ti)/\text{mix}(ti)$ 反映了在混合行为中,文档a和b被连续点击和下载的概率。

[0047] 步骤b3,按照下式计算两个知识文档a和b的点击下载关联度 $\text{mix}(a,b)$:

$$\text{mix}(a,b) = \frac{\sum_i \frac{\text{scan}(a,b,ti) * R_1 + \text{download}(a,b,ti) * R_2 + \text{mix}(a,b,ti) * R_3}{\text{scan}(ti) + \text{download}(ti) + \text{mix}(ti)} * S(i)}{R_1 + R_2 + R_3}$$

其中, R_1 为 $\text{scan}(ti)$ 对应的权重, R_2 为 $\text{download}(ti)$ 对应的权重, R_3 为 $\text{mix}(ti)$ 对应的权重, $R_1+R_2+R_3=1$ 。

[0048] 在本公开一些实施例中,不同类型的用户行为对于判断文档关联的重要程度可能不同。例如,下载行为通常比点击行为更能反映用户对文档的认可,因此权重可以设置得更高。本公开设定行为类型权重 R_1 、 R_2 、 R_3 ,分别对应点击-点击、下载-下载、点击-下载三种行为,且三者之和为1。对于每个时间窗口 ti ,先计算该窗口内的行为关联度分量:将三种行为的条件概率分别乘以对应的行为权重,再除以行为权重之和(由于 $R_1+R_2+R_3=1$,分母实际为1,但保留分母形式更一般化)。然后,对各时间窗口的关联度分量按时段权重 $S(i)$ 进行加权平均,得到最终的第二关联度 $\text{mix}(a,b)$ 。

[0049] 可以理解的是,第二关联度从用户行为维度捕捉文档间的潜在关联,通过多层加权(行为类型权重+时段权重)的设计,能够精细地反映不同行为模式和时间粒度下的关联强度。

[0050] 通过步骤a1-a3和步骤b1-b3,本公开分别得到了第一关联度 $\text{similarity_compare}(a,b)$ 和第二关联度 $\text{mix}(a,b)$ 。第一关联度基于文档的静态属性,第二关联度基于用户的动态行为,二者从不同维度刻画了文档之间的关联程度。

[0051] 步骤104,利用融合参数将第一关联度和第二关联度进行融合,得到两个知识文档

之间的关联度概率。

[0052] 在本公开一些实施例中,作为一种示例,步骤104具体可以包括以下子步骤:
步骤c1,采用以下公式计算两个知识文档之间的关联度概率 $P_{pre}(a,b)$:

$$P_{pre}(a,b) = \alpha * similarity_compare(a,b) + (1-\alpha) * mix(a,b)$$

[0053] 其中, α 为预设的融合参数,且 $0 \leq \alpha \leq 1$, $similarity_compare(a,b)$ 为第一关联度, $mix(a,b)$ 为第二关联度。

[0054] 在本公开一些实施例中,融合参数 α 用于平衡第一关联度和第二关联度在最终关联度概率中的贡献。当 α 接近1时,最终关联度更侧重于文档的静态属性;当 α 接近0时,更侧重于用户的动态行为;当 $\alpha=0.5$ 时,两者贡献相等。 α 的具体取值可以根据知识服务平台的实际运行情况通过统计或机器学习方法确定。

[0055] 需要说明的是,第一关联度 $similarity_compare(a,b)$ 的取值范围为 $[-1,1]$,第二关联度 $mix(a,b)$ 的取值范围为 $[0,1]$ 。在融合之前,可以对第一关联度进行归一化处理,将其映射到 $[0,1]$ 区间,以确保两者在相同的量纲下进行线性融合。例如,可以通过 $(similarity_compare+1)/2$ 的方式将其映射到 $[0,1]$ 。

[0056] 通过步骤104,本公开将文档的静态属性特征和用户的动态行为特征有机融合,得到最终的两个知识文档之间的关联度概率。这一概率值综合了多维信息,能够更全面、准确地反映矿山科研知识文档之间的实际关联程度。

[0057] 作为一种示例,本公开还可以对融合后的关联度概率进行进一步处理,例如设定阈值以判断是否建立文档间的引用关系,或作为后续知识图谱构建、智能推荐、科研价值评估等应用的基础数据。可以理解的是,通过上述方法得到的关联度概率具有较高的可信度和实用性。

[0058] 在本公开一些实施例中,作为一种示例,所述时间窗口包括多个不同时间粒度的窗口,所述时段权重 $S(i)$ 为归一化权重,其中,同一时间窗口对应的时间粒度与时段权重取值成反比。也就是说,时间窗口越小(时间粒度越细),时段权重取值越大;时间窗口越大(时间粒度越粗),时段权重取值越小。这种设定符合直观认知:用户在极短时间内连续操作两篇文档,其关联性更显著,因此应赋予更高的权重。

[0059] 根据本公开实施例提出的应用于矿山科研知识服务平台的知识文档关联度确定方法,通过获取矿山科研知识库中知识文档的结构化属性信息,以及用户对知识文档的点击记录和下载记录,针对任意两个知识文档,分别基于结构化属性信息计算两个知识文档之间的第一关联度,并基于点击记录和下载记录计算两个知识文档之间的第二关联度,进而利用融合参数将第一关联度和第二关联度进行融合,得到两个知识文档之间的关联度概率;其有益效果在于:一方面,通过结构化属性信息计算第一关联度,能够从文档自身特征维度挖掘文档间的内在联系,另一方面,通过点击记录和下载记录计算第二关联度,能够从用户行为维度捕捉文档间的潜在关联,二者融合得到的关联度概率综合了静态属性与动态行为的多维信息,从而更全面、准确地反映矿山科研知识文档之间的实际关联程度,为后续知识图谱构建、智能推荐或科研价值评估提供可靠的关联数据基础。

[0060] 图2是根据一示例性实施例示出的一种应用于矿山科研知识服务平台的知识文档关联度预测装置框图。参照图2,该装置包括第一获取单元201,第二获取单元202,计算单元203和融合单元204。

[0061] 其中,第一获取单元201,用于获取矿山科研知识库中知识文档的结构化属性信息;

第二获取单元202,用于获取用户与所述知识文档的交互日志;所述交互日志包括用户对知识文档的点击记录和下载记录;

计算单元203,用于针对任意两个知识文档,基于所述结构化属性信息计算两个知识文档之间的第一关联度,并基于所述点击记录和下载记录计算两个知识文档之间的第二关联度;

融合单元204,用于利用融合参数将所述第一关联度和所述第二关联度进行融合,得到两个知识文档之间的关联度概率。

[0062] 在本申请一些实施例中,计算单元203具体可以用于:

针对每一种属性 n ,计算两个知识文档 a 和 b 在该属性上的余弦相似度 $\text{similarity}(\text{attribute}[a,n], \text{attribute}[b,n], n)$;

获取各属性的重要度权重 $Q(n)$,且 $\sum Q(n)=1$;

采用以下公式计算第一关联度 $\text{similarity_compare}(a,b)$:

$$\text{similarity_compare}(a,b) = \frac{\sum_n \text{similarity}(\text{attribute}[a,n], \text{attribute}[b,n], n) * Q(n) * 1_{\{\text{similarity}(\text{attribute}[a,n], \text{attribute}[b,n], n) > \beta\}}}{\sum_n Q(n)}$$

[0063] 其中, β 为预设相似度阈值。

[0064] 在本申请一些实施例中,计算单元203具体可以用于:

设定多个时间窗口 t_i ,每个时间窗口对应一个时段权重 $S(i)$,且 $\sum S(i)=1$;

在每个时间窗口 t_i 内:

统计同一用户连续点击文档 a 和 b 的次数 $\text{scan}(a,b,t_i)$,以及同一用户任意连续点击两个知识文档的总次数 $\text{scan}(t_i)$;

统计同一用户连续下载文档 a 和 b 的次数 $\text{download}(a,b,t_i)$,以及同一用户任意连续下载两个知识文档的总次数 $\text{download}(t_i)$;

统计同一用户连续点击一个文档且下载另一个文档的次数 $\text{mix}(a,b,t_i)$,以及同一用户任意连续点击和下载两个知识文档的总次数 $\text{mix}(t_i)$;

按照下式计算两个知识文档 a 和 b 的点击下载关联度 $\text{mix}(a,b)$:

$$\text{mix}(a,b) = \frac{\sum_i \frac{\frac{\text{scan}(a,b,t_i)}{\text{scan}(t_i)} * R_1 + \frac{\text{download}(a,b,t_i)}{\text{download}(t_i)} * R_2 + \frac{\text{mix}(a,b,t_i)}{\text{mix}(t_i)} * R_3}{R_1 + R_2 + R_3} * S(i)}{\sum_i S(i)}$$

[0065] 其中, R_1 为 $\text{scan}(t_i)$ 对应的权重, R_2 为 $\text{download}(t_i)$ 对应的权重, R_3 为 $\text{mix}(t_i)$ 对应的权重, $R_1+R_2+R_3=1$ 。

[0066] 在本申请一些实施例中,融合单元204具体可以用于:

采用以下公式计算两个知识文档之间的关联度概率 $P_pre(a,b)$:

$$P_pre(a,b) = \alpha * \text{similarity_compare}(a,b) + (1-\alpha) * \text{mix}(a,b)$$

[0067] 其中, α 为预设的融合参数, $0 \leq \alpha \leq 1$, $\text{similarity_compare}(a,b)$ 为第一关联

度, $\text{mix}(a, b)$ 为第二关联度。

[0068] 在本申请一些实施例中, 所述时间窗口包括多个不同时间粒度的窗口, 所述时段权重 $S(i)$ 为归一化权重, 其中, 同一时间窗口对应的时间粒度与时段权重取值成反比。

[0069] 关于上述实施例中的装置, 其中各个模块执行操作的具体方式已经在有关该方法的实施例中进行了详细描述, 此处将不做详细阐述说明。

[0070] 根据本公开实施例提出的应用于矿山科研知识服务平台的知识文档关联度预测装置, 通过获取矿山科研知识库中知识文档的结构化属性信息, 以及用户对知识文档的点击记录和下载记录, 针对任意两个知识文档, 分别基于结构化属性信息计算两个知识文档之间的第一关联度, 并基于点击记录和下载记录计算两个知识文档之间的第二关联度, 进而利用融合参数将第一关联度和第二关联度进行融合, 得到两个知识文档之间的关联度概率; 其有益效果在于: 一方面, 通过结构化属性信息计算第一关联度, 能够从文档自身特征维度挖掘文档间的内在联系, 另一方面, 通过点击记录和下载记录计算第二关联度, 能够从用户行为维度捕捉文档间的潜在关联, 二者融合得到的关联度概率综合了静态属性与动态行为的多维信息, 从而更全面、准确地反映矿山科研知识文档之间的实际关联程度, 为后续知识图谱构建、智能推荐或科研价值评估提供可靠的关联数据基础。

[0071] 图3是根据一示例性实施例示出的一种用于应用于矿山科研知识服务平台的知识文档关联度确定方法的装置的框图。例如, 装置300可以是电子设备, 例如可以是移动电话, 计算机, 数字广播终端, 消息收发设备, 平板设备, 个人数字助理等。

[0072] 参照图3, 装置300可以包括以下一个或多个组件: 处理组件302, 存储器304, 电力组件306, 多媒体组件308, 音频组件310, 输入/输出(I/O)的接口312, 传感器组件314, 以及通信组件316。

[0073] 处理组件302通常控制装置300的整体操作, 诸如与显示, 电话呼叫, 数据通信, 相机操作和记录操作相关联的操作。处理组件302可以包括一个或多个处理器320来执行指令, 以完成上述的方法的全部或部分步骤。此外, 处理组件302可以包括一个或多个模块, 便于处理组件302和其他组件之间的交互。例如, 处理组件302可以包括多媒体模块, 以方便多媒体组件308和处理组件302之间的交互。

[0074] 存储器304被配置为存储各种类型的数据以支持在设备300的操作。这些数据的示例包括用于在装置300上操作的任何应用程序或方法的指令, 联系人数据, 电话簿数据, 消息, 图片, 视频等。存储器304可以由任何类型的易失性或非易失性存储设备或者它们的组合实现, 如静态随机存取存储器(SRAM), 电可擦除可编程只读存储器(EEPROM), 可擦除可编程只读存储器(EPROM), 可编程只读存储器(PROM), 只读存储器(ROM), 磁存储器, 快闪存储器, 磁盘或光盘。

[0075] 电力组件306为装置300的各种组件提供电力。电力组件306可以包括电源管理系统, 一个或多个电源, 及其他与为装置300生成、管理和分配电力相关联的组件。

[0076] 多媒体组件308包括在所述装置300和用户之间的提供一个输出接口的屏幕。在一些实施例中, 屏幕可以包括液晶显示器(LCD)和触摸面板(TP)。如果屏幕包括触摸面板, 屏幕可以被实现为触摸屏, 以接收来自用户的输入信号。触摸面板包括一个或多个触摸传感器以感测触摸、滑动和触摸面板上的手势。所述触摸传感器可以不仅感测触摸或滑动动作的边界, 而且还检测与所述触摸或滑动操作相关的持续时间和压力。在一些实施例中, 多媒

体组件308包括一个前置摄像头和/或后置摄像头。当设备300处于操作模式,如拍摄模式或视频模式时,前置摄像头和/或后置摄像头可以接收外部的多媒体数据。每个前置摄像头和后置摄像头可以是一个固定的光学透镜系统或具有焦距和光学变焦能力。

[0077] 音频组件310被配置为输出和/或输入音频信号。例如,音频组件310包括一个麦克风(MIC),当装置300处于操作模式,如呼叫模式、记录模式和语音识别模式时,麦克风被配置为接收外部音频信号。所接收的音频信号可以被进一步存储在存储器304或经由通信组件316发送。在一些实施例中,音频组件310还包括一个扬声器,用于输出音频信号。

[0078] I/O接口312为处理组件302和外围接口模块之间提供接口,上述外围接口模块可以是键盘,点击轮,按钮等。这些按钮可包括但不限于:主页按钮、音量按钮、启动按钮和锁定按钮。

[0079] 传感器组件314包括一个或多个传感器,用于为装置300提供各个方面的状态评估。例如,传感器组件314可以检测到设备300的打开/关闭状态,组件的相对定位,例如所述组件为装置300的显示器和小键盘,传感器组件314还可以检测装置300或装置300一个组件的位置改变,用户与装置300接触的存在或不存在,装置300方位或加速/减速和装置300的温度变化。传感器组件314可以包括接近传感器,被配置用来在没有任何的物理接触时检测附近物体的存在。传感器组件314还可以包括光传感器,如CMOS或CCD图像传感器,用于在成像应用中使用。在一些实施例中,该传感器组件314还可以包括加速度传感器,陀螺仪传感器,磁传感器,压力传感器或温度传感器。

[0080] 通信组件316被配置为便于装置300和其他设备之间有线或无线方式的通信。装置300可以接入基于通信标准的无线网络,如WiFi,2G或3G,或它们的组合。在一个示例性实施例中,通信组件316经由广播信道接收来自外部广播管理系统的广播信号或广播相关信息。在一个示例性实施例中,所述通信组件316还包括近场通信(NFC)模块,以促进短程通信。例如,在NFC模块可基于射频识别(RFID)技术,红外数据协会(IrDA)技术,超宽带(UWB)技术,蓝牙(BT)技术和其他技术来实现。

[0081] 在示例性实施例中,装置300可以被一个或多个应用专用集成电路(ASIC)、数字信号处理器(DSP)、数字信号处理设备(DSPD)、可编程逻辑器件(PLD)、现场可编程门阵列(FPGA)、控制器、微控制器、微处理器或其他电子元件实现,用于执行上述方法。

[0082] 在示例性实施例中,还提供了一种包括指令的非临时性计算机可读存储介质,例如包括指令的存储器304,上述指令可由装置300的处理器320执行以完成上述方法。例如,所述非临时性计算机可读存储介质可以是ROM、随机存取存储器(RAM)、CD-ROM、磁带、软盘和光数据存储设备等。

[0083] 在示例性实施例中,还提供了一种计算机程序产品,包括计算机程序,所述计算机程序在被装置300的处理器320执行时实现上述方法。

[0084] 本领域技术人员在考虑说明书及实践这里公开的发明后,将容易想到本发明的其它实施方案。本公开旨在涵盖本发明的任何变型、用途或者适应性变化,这些变型、用途或者适应性变化遵循本发明的一般性原理并包括本公开未公开的本技术领域中的公知常识或惯用技术手段。说明书和实施例仅被视为示例性的,本发明的真正范围和精神由下面的权利要求指出。

[0085] 应当理解的是,本发明并不局限于上面已经描述并在附图中示出的精确结构,并

且可以在不脱离其范围进行各种修改和改变。本发明的范围仅由所附的权利要求来限制。

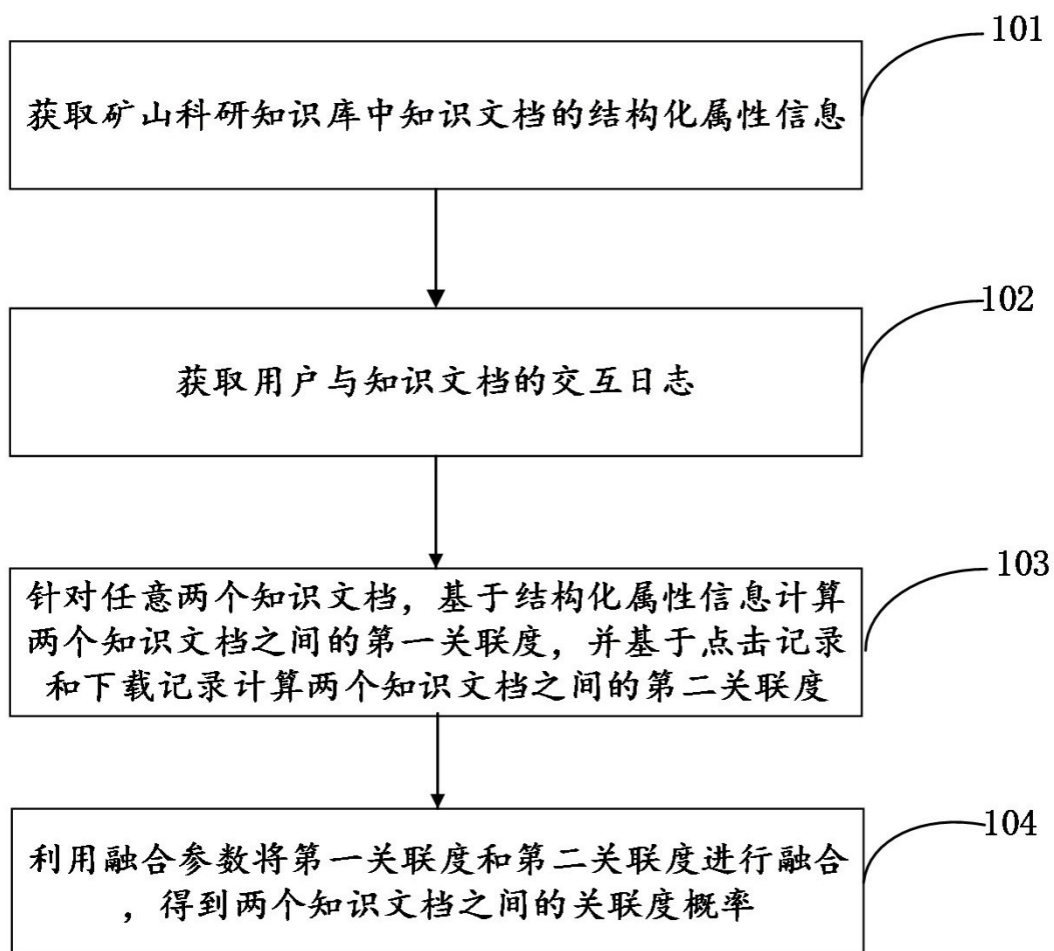


图1

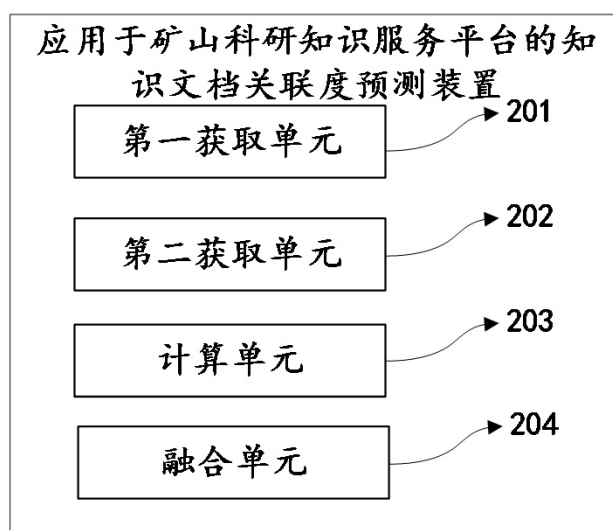


图2

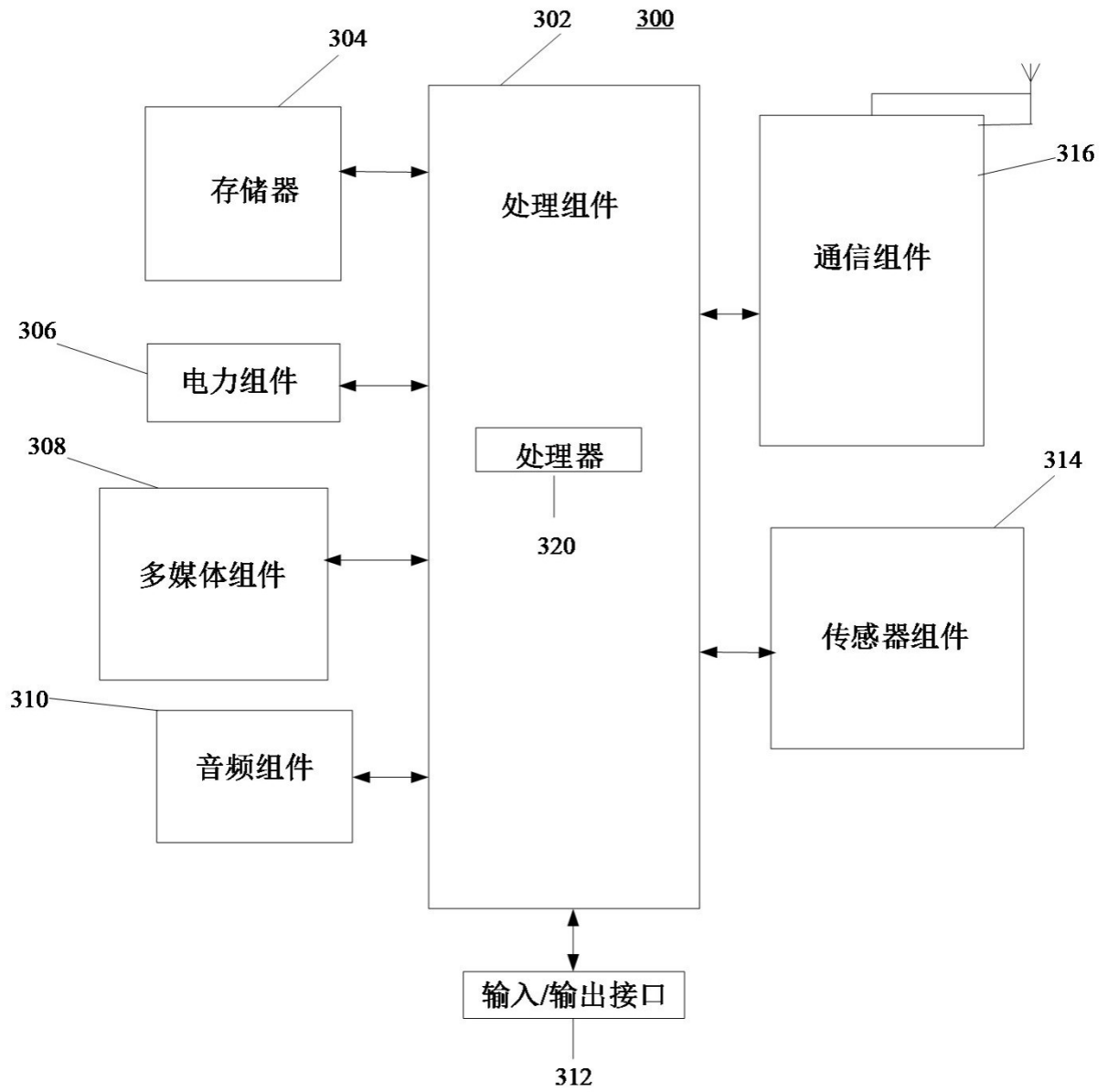


图3