(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 106407316 B
(45)授权公告日 2020.05.15

(21)申请号 201610785149.4

(22)申请日 2016.08.30

(65)同一申请的已公布的文献号
申请公布号 CN 106407316 A

(43)申请公布日 2017.02.15

(73)专利权人 北京航空航天大学
地址 100191 北京市海淀区学院路37号北京航空航天大学新主楼G506

(72)发明人 刘旭东 孙海龙 孙富民 王旭

(74)专利代理机构 北京同立钧成知识产权代理有限公司 11205
代理人 杨泽 刘芳

(51)Int.Cl.
G06F 16/903(2019.01)

(56)对比文件
CN 105653706 A,2016.06.08,
CN 105069143 A,2015.11.18,
CN 104298776 A,2015.01.21,
CN 105843795 A,2016.08.10,

审查员 解欣

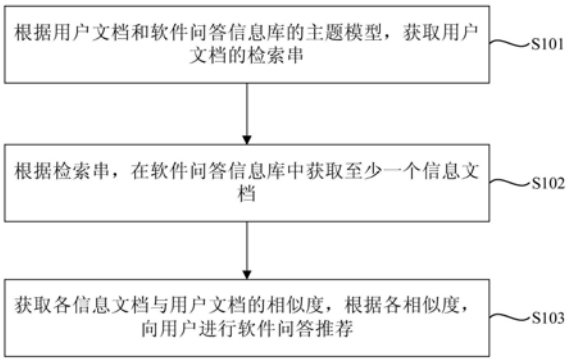
权利要求书3页 说明书9页 附图3页

(54)发明名称

基于主题模型的软件问答推荐方法和装置

(57)摘要

本发明提供一种基于主题模型的软件问答推荐方法和装置，该方法包括：根据软件问答信息库的主题模型和用户文档，获取用户文档的检索串；根据检索串，在软件问答信息库中获取至少一个信息文档；获取各信息文档与用户文档的相似度，根据各相似度，向用户进行软件问答推荐。本发明提供的基于主题模型的软件问答推荐方法，通过采用基于主题模型的检索串获取方法，可自动根据开发人员的当前操作文档中的内容确定准确的检索串，不仅简化了检索操作而且结合充分利用了用户文档的上下文信息，使得提高了检索串的准确性，提高了软件问答的推荐效果，并为各信息文档与用户文档计算相似度，确定推荐顺序，进一步提高了软件问答的推荐效果。



CN 106407316 B

1.一种基于主题模型的软件问答推荐方法,其特征在于,包括:

根据用户文档和软件问答信息库的主题模型,获取所述用户文档的检索串;

根据所述检索串,在所述软件问答信息库中获取至少一个信息文档;

获取各所述信息文档与所述用户文档的相似度,根据各所述相似度,向所述用户进行软件问答推荐;

所述获取各所述信息文档与所述用户文档的相似度,根据各所述相似度,向所述用户进行软件问答推荐,包括:

针对任一信息文档,获取所述信息文档与所述用户文档的相似度;

根据所述信息文档的至少一项元信息特征,对所述信息文档的相似度进行修正,得到修正后的相似度;

根据各所述修正后的相似度,向所述用户进行软件问答推荐;

所述根据用户文档和软件问答信息库的主题模型,获取所述用户文档的检索串之前,还包括:

采用隐含狄利克雷分布算法,获取所述软件问答信息库的主题模型;

确定所述用户文档的变化量达到预设变化量,则确定执行获取所述用户文档的检索串的操作。

2.根据权利要求1所述的方法,其特征在于,所述根据用户文档和软件问答信息库的主题模型,获取所述用户文档的检索串,包括:

根据所述主题模型,获取所述用户文档的主题结构;

根据所述用户文档的主题结构,获取所述用户文档中出现概率最高的预设数量个单词,作为所述用户文档的检索串。

3.根据权利要求2所述的方法,其特征在于,所述根据所述用户文档的主题结构,获取所述用户文档中出现概率最高的预设数量个单词,包括:

根据所述用户文档的主题结构,采用如下的公式一获取所述用户文档中出现概率最高的L个单词,作为所述用户文档的检索串Q;

$$Q = \arg\max P(Q|D) = \arg\max \prod_{i=1}^{L} P(q_i|D) \qquad \text{公式一;}$$

其中,$P(q_i|D)$是所述用户文档D中单词q_i出现的概率,L为所述预设数量,i的取值范围为从1至L的正整数;$P(q_i|D) = \sum_{j=1}^{K} P(q_i|z_j,\varphi_j)P(z_j|\theta,D)$,$P(z_j|\theta,D)$是所述用户文档D中主题$z_j$出现的概率,θ是所述用户文档D的主题分布,$P(q_i|z_j,\varphi_j)$是主题$z_j$中单词$q_i$出现的概率,$\varphi_j$是所述主题$z_j$的单词分布,K为所述主题模型中主题的总数量,j的取值范围为从1至K的正整数。

4.根据权利要求3所述的方法,其特征在于,所述获取各所述信息文档与所述用户文档的相似度,包括:

采用如下的公式二确定所述用户文档与各所述信息文档的相似度Sim;

$$\mathrm{Sim}=\frac{\sum_{j=1}^{K}P(z_j\mid\theta,D)\times P(z_j\mid\theta',D')}{\sqrt{\sum_{j=1}^{K}P(z_j\mid\theta,D)^2}\times\sqrt{\sum_{j=1}^{K}P(z_j\mid\theta',D')^2}}$$ 公式二；

其中，$P(z_j|\theta',D')$表示任一信息文档D′中主题z_j出现的概率，θ'为所述信息文档D′的主题分布。

5.根据权利要求1所述的方法，其特征在于，所述根据所述信息文档的至少一项元信息特征，对所述信息文档的相似度Sim进行修正，得到修正后的相似度Sim′，包括：

根据所述信息文档的至少一项元信息特征t_m，采用如下的公式三对所述信息文档的相似度Sim进行修正，得到修正后的相似度Sim′；

$$Sim'=\alpha\times Sim+(1-\alpha)\sum_{m=1}^{M}\beta_m\times t_m$$ 公式三；

其中，α为所述信息文档的相似度Sim的权重，M为所述信息文档的元信息特征的总数量，m的取值为从1至M的正整数，β_m为元信息特征t_m的权重，α和β_m的取值为从0至1的实数。

6.根据权利要求1所述的方法，其特征在于，所述软件问答信息库中包括如下至少一种文档：

百科文档、网络贴吧文档、网络社区文档、问答网站中的文档。

7.根据权利要求6所述的方法，其特征在于，当所述信息文档为问答网站中的文档时，所述元信息特征包括如下中的至少一项：

信息文档中的问题质量、答案质量、提问用户声望、回答用户声望、应用程序接口相似度、文本相似度。

8.一种基于主题模型的软件问答推荐装置，其特征在于，包括：

检索串获取模块，用于根据软件问答信息库的主题模型和用户文档，获取所述用户文档的检索串；

信息文档获取模块，用于根据所述检索串，在所述软件问答信息库中获取至少一个信息文档；

推荐模块，用于获取各所述信息文档与所述用户文档的相似度，根据各所述相似度，向所述用户进行软件问答推荐；

所述推荐模块，还用于：

针对任一信息文档，获取所述信息文档与所述用户文档的相似度；

根据所述信息文档的至少一项元信息特征，对所述信息文档的相似度进行修正，得到修正后的相似度；

根据各所述修正后的相似度，向所述用户进行软件问答推荐；

所述装置还包括：

主题模型获取模块，用于采用隐含狄利克雷分布算法，获取软件问答信息库的主题模型；

变化检测模块，用于在确定用户文档的变化量达到预设变化量时，控制所述检索串获

取模块执行获取用户文档的检索串的操作。

基于主题模型的软件问答推荐方法和装置

技术领域

[0001] 本发明涉及信息技术，尤其涉及一种基于主题模型的软件问答推荐方法和装置。

背景技术

[0002] 在软件开发、代码编写等各类工作中，当技术人员遇到难以解决的问题时，通常会采用检索的方式在网络中搜索相关资料，以作为参考。

[0003] 技术人员在检索之前，需根据遇到的问题自行设定检索关键词，在搜索引擎或相关论坛、资料共享网站中输入检索关键词，得到检索结果。技术人员需在所有检索结果中，进行人工筛选，确定出相关度较高的信息后再进一步进行仔细分析，确定是否能够真正解决问题。

[0004] 由于人工筛选消耗时间和精力较多，且可能因为关键词设置不合理，导致技术人员即使耗时较长也无法检索得到对问题有帮助的信息，人工检索效率较低。现有技术提供一种自动推荐相关信息的方法，采用提前训练得到的关键词与网页的对应模型，将技术人员实时输入的内容作为关键词，自动为技术人员提供可能相关的网页。但是由于模型过于简单，导致推荐的信息不准确、推荐效果较差。

发明内容

[0005] 本发明提供一种基于主题模型的软件问答推荐方法和装置，用以解决现有软件问答推荐中推荐效果较差的问题。

[0006] 本发明一方面提供一种基于主题模型的软件问答推荐方法，包括：

[0007] 根据用户文档和软件问答信息库的主题模型，获取所述用户文档的检索串；

[0008] 根据所述检索串，在所述软件问答信息库中获取至少一个信息文档；

[0009] 获取各所述信息文档与所述用户文档的相似度，根据各所述相似度，向所述用户进行软件问答推荐。

[0010] 如上所述的基于主题模型的软件问答推荐方法，所述根据用户文档和软件问答信息库的主题模型，获取所述用户文档的检索串，包括：

[0011] 根据所述主题模型，获取所述用户文档的主题结构；

[0012] 根据所述用户文档的主题结构，获取所述用户文档中出现概率最高的预设数量个单词，作为所述用户文档的检索串。

[0013] 如上所述的基于主题模型的软件问答推荐方法，所述根据所述用户文档的主题结构，获取所述用户文档中出现概率最高的预设数量个单词，包括：

[0014] 根据所述用户文档的主题结构，采用如下的公式一获取所述用户文档中出现概率最高的L个单词，作为所述用户文档的检索串Q；

[0015] $$Q = \arg\max P(Q|D) = \arg\max \prod_{i=1}^{L} P(q_i|D)$$ 公式一；

[0016] 其中，$P(q_i|D)$是所述用户文档D中单词q_i出现的概率，L为所述预设数量，i的取值

范围为从1至L的正整数；$P(q_i|D)=\sum_{j=1}^{K}P(q_i|z_j,\varphi_j)P(z_j|\theta,D)$，$P(z_j|\theta,D)$是所述用户文档D中主题$z_j$出现的概率，θ是所述用户文档D的主题分布，$P(q_i|z_j,\varphi_j)$是主题$z_j$中单词$q_i$出现的概率，$\varphi_j$是所述主题$z_j$的单词分布，K为所述主题模型中主题的总数量，j的取值范围为从1至K的正整数。

[0017] 如上所述的基于主题模型的软件问答推荐方法，所述获取各所述信息文档与所述用户文档的相似度，包括：

[0018] 采用如下的公式二确定所述用户文档与各所述信息文档的相似度Sim；

[0019] $$Sim=\frac{\sum_{j=1}^{K}P(z_j|\theta,D)\bullet P(z_j|\theta',D')}{\sqrt{\sum_{j=1}^{K}P(z_j|\theta,D)^2}\bullet\sqrt{\sum_{j=1}^{K}P(z_j|\theta',D')^2}}$$ 公式二；

[0020] 其中，$P(z_j|\theta',D')$表示任一信息文档D′中主题z_j出现的概率，θ′为所述信息文档D′的主题分布。

[0021] 如上所述的基于主题模型的软件问答推荐方法，所述根据用户文档和软件问答信息库的主题模型，获取所述用户文档的检索串之前，还包括：

[0022] 采用隐含狄利克雷分布算法，获取所述软件问答信息库的主题模型；

[0023] 确定所述用户文档的变化量达到预设变化量，则确定执行获取所述用户文档的检索串的操作。

[0024] 如上所述的基于主题模型的软件问答推荐方法，所述获取各所述信息文档与所述用户文档的相似度Sim，根据各所述相似度，向所述用户进行软件问答推荐，包括：

[0025] 针对任一信息文档，获取所述信息文档与所述用户文档的相似度Sim；

[0026] 根据所述信息文档的至少一项元信息特征，对所述信息文档的相似度Sim进行修正，得到修正后的相似度Sim′；

[0027] 根据各所述修正后的相似度Sim′，向所述用户进行软件问答推荐。

[0028] 如上所述的基于主题模型的软件问答推荐方法，所述根据所述信息文档的至少一项元信息特征，对所述信息文档的相似度Sim进行修正，得到修正后的相似度Sim′，包括：

[0029] 根据所述信息文档的至少一项元信息特征t_m，采用如下的公式三对所述信息文档的相似度Sim进行修正，得到修正后的相似度Sim′；

[0030] $$Sim'=\alpha\bullet Sim+(1-\alpha)\sum_{m=1}^{M}\beta_m\bullet t_m$$ 公式三；

[0031] 其中，α为所述信息文档的相似度Sim的权重，M为所述信息文档的元信息特征的总数量，m的取值为从1至M的正整数，β_m为元信息特征t_m的权重，α和β_m的取值为从0至1的实数。

[0032] 如上所述的基于主题模型的软件问答推荐方法，所述软件问答信息库中包括如下至少一种文档：

[0033] 百科文档、网络贴吧文档、网络社区文档、问答网站中的文档。

[0034] 如上所述的基于主题模型的软件问答推荐方法，当所述信息文档为问答网站中的文档时，所述元信息特征包括如下中的至少一项：

[0035] 信息文档中的问题质量、答案质量、提问用户声望、回答用户声望、应用程序接口相似度、文本相似度。

[0036] 本发明另一方面提供一种基于主题模型的软件问答推荐装置，包括：

[0037] 检索串获取模块，用于根据用户文档和软件问答信息库的主题模型，获取所述用户文档的检索串；

[0038] 信息文档获取模块，用于根据所述检索串，在所述软件问答信息库中获取至少一个信息文档；

[0039] 推荐模块，用于获取各所述信息文档与所述用户文档的相似度，根据各所述相似度，向所述用户进行软件问答推荐。

[0040] 本发明提供的基于主题模型的软件问答推荐方法和装置，基于软件问答信息库的主题模型，获取用户文档的检索串，再根据检索串在软件问答信息库中获取至少一个信息文档，最后获取各信息文档与用户文档的相似度，根据获取到的各相似度向用户进行软件问答推荐，通过采用基于主题模型的检索串获取方法，可自动根据开发人员的当前操作文档中的内容确定准确的检索串，不仅简化了检索操作而且结合充分利用了用户文档的上下文信息，使得提高了检索串的准确性，提高了软件问答的推荐效果，并为各信息文档与用户文档计算相似度，确定推荐顺序，进一步提高了软件问答的推荐效果。

附图说明

[0041] 为了更清楚地说明本发明实施例的技术方案，下面将对实施例描述中所需要使用的附图作一简单地介绍，显而易见地，下面描述中的附图是本发明的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动性的前提下，还可以根据这些附图获得其他的附图。

[0042] 图1为本发明提供的基于主题模型的软件问答推荐方法实施例一的流程示意图；

[0043] 图2为本发明提供的基于主题模型的软件问答推荐方法实施例二的流程示意图；

[0044] 图3为本发明提供的基于主题模型的软件问答推荐方法实施例三的流程示意图；

[0045] 图4为本发明提供的基于主题模型的软件问答推荐方法实施例四的流程示意图；

[0046] 图5本发明提供的基于主题模型的软件问答推荐装置实施例一的结构示意图。

具体实施方式

[0047] 为使本发明实施例的目的、技术方案和优点更加清楚，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作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。

[0048] 在软件开发领域中，软件开发人员在代码编写过程中，若遇到技术难题，需从开发环境切换至浏览器，再自行设定检索关键词在搜索引擎或软件开发问答网站中进行搜索，在搜索得到的相关文档中人工筛选具有参考性的文档。为简化软件开发人员的操作、提高效率，现有技术在开发环境中集成软件问答推荐工具，用于自动根据软件开发人员输入的

内容，生产检索关键词，在网页中检索相关的信息文档，并按相似度向软件开发人员推荐，节省了软件开发人员在开发环境和浏览器之间切换的时间，但是在生成检索关键词时，仅在开发人员的开发环境中检测是否存在预设关键词或根据开发人员当前输入的单词生成检索关键词，导致检索效果较差，而且，在确定检索关键词后，采用提前训练得到的关键词与网页的对应模型，自动为技术人员提供可能相关的网页，因此推荐效果较差。

[0049] 为解决上述问题，本发明实施例提供一种基于主题模型的软件问答推荐方法和装置，根据软件问答信息库中的信息文档的主题模型，自动为开发人员当前的编写的代码文档提取多个检索关键词作为检索串，并根据检索串在软件问答信息库中进行检索，得到用于参考的信息文档并向开发人员进行推荐，提高了软件问答的推荐效果。

[0050] 下面以具体地实施例对本发明的技术方案以及本发明的技术方案如何实现进行详细说明。

[0051] 本发明实施例提供一种基于主题模型的软件问答推荐方法，该方法的执行主体可以为基于主题模型的软件问答推荐装置，该装置可以由软件和/或硬件实现，集成在软件开发人员使用的开发环境中。图1为本发明提供的基于主题模型的软件问答推荐方法实施例一的流程示意图。如图1所示，该方法包括：

[0052] 步骤101、根据用户文档和软件问答信息库的主题模型，获取用户文档的检索串；

[0053] 步骤102、根据检索串，在软件问答信息库中获取至少一个信息文档；

[0054] 步骤103、获取各信息文档与用户文档的相似度，根据各相似度，向用户进行软件问答推荐。

[0055] 具体的，在步骤101中，该装置实时监测开发人员在开发环境中进行的代码编写，代码文档即为用户文档，当用户编写的代码不同，用户可能遇到的技术问题也不同，需根据用户文档中的内容选择至少一个检索关键词作为检索串。示例性的，可以根据软件问答信息库的主题模型，生成用户文档的检索串。可选的，软件问答信息库中包括如下至少一种文档：百科文档、网络贴吧文档、网络社区文档、问答网站中的文档。软件问答信息库还可以为开发人员预设的其他网络文档，本发明对此不做限定。上述文档中包括任意开发人员分享的开发过程中所遇到的问题的解决办法。为便于准确确定各文档内容，便于检索，需获取软件问答信息库的主题模型，主题模型包括上述各文档的主题分布，以及各文档的各主题下的单词分布，即表示一个单词、一个主题在某一文档中的出现概率。根据该主题模型，对用户文档进行训练，可以得到用户文档的主题结构，用户文档的主题结构中包括用户文档的主题分布，以及各主题下的单词分布。通过采用基于主题模型的方法，自动获取用户文档的检索串，可以准确的确定出能够代表用户文档的检索关键词。

[0056] 具体的，在步骤102中，根据步骤101中获取的检索串，在软件问答信息库中进行搜索，确定出至少一个信息文档。示例性的，在软件问答信息库中进行检索时，利用软件问答信息库自身的搜索引擎进行搜索，得到搜索引擎提供的至少一个信息文档。示例性的，当搜索得到的信息文档过多时，可仅选用搜索引擎推荐的相关度较高的N个信息文档，N为开发人员预设的正整数。可以示例性的认为相关度排名在N之后的信息文档与检索串代表的用户文档相关度较低。

[0057] 具体的，由于步骤102中获取到的各信息文档的排序是根据其与检索串的吻合度进行的排序，并不能完全代表各信息文档与用户文档的相似度。因此在步骤103中，对步骤

102中获取到的各信息文档分别与用户文档进行相似度计算，得到各信息文档与用户文档的相似度，再根据各相似度的值的大小，按照从大到小的顺序，向用户进行推荐，以提高推荐效果。具体的，在进行相似度计算时，可采用资讯检索资讯探勘的常用加权技术(term frequency-inverse document frequency，简称TF-IDF)、浅层语义分析(Latent semantic analysis，简称LSA)等算法。

[0058] 本发明提供的基于主题模型的软件问答推荐方法，基于软件问答信息库的主题模型，获取用户文档的检索串，再根据检索串在软件问答信息库中获取至少一个信息文档，最后获取各信息文档与用户文档的相似度，根据获取到的各相似度向用户进行软件问答推荐，通过采用基于主题模型的检索串获取方法，可自动根据开发人员的当前操作文档中的内容确定准确的检索串，不仅简化了检索操作而且结合充分利用了用户文档的上下文信息，使得提高了检索串的准确性，提高了软件问答的推荐效果，并为各信息文档与用户文档计算相似度，确定推荐顺序，进一步提高了软件问答的推荐效果。

[0059] 下面在图1所示实施例的基础上，以具体地实施例对本发明的获取检索串的方法进行详细说明。

[0060] 图2为本发明提供的基于主题模型的软件问答推荐方法实施例二的流程示意图，如图2所示，获取检索串具体包括：

[0061] 步骤201、根据主题模型，获取用户文档的主题结构；

[0062] 步骤202、根据用户文档的主题结构，获取用户文档中出现概率最高的预设数量个单词，作为用户文档的检索串。

[0063] 具体的，采用隐含狄利克雷分布算法，根据软件问答信息库的主题模型对用户文档进行训练，即可确定用户文档在该主题模型上的主题结构。然后，根据用户文档的主题结构，即可得到用户文档中包括的每个单词出现的概率，概率越高的单词在用户文档中出现的次数越多，越能代表用户文档，因此，可筛选出用户文档中出现概率最高的预设数量个单词，作为用户文档的检索串。

[0064] 进一步的，在上述实施例的基础上，具体获取检索串的方法包括：

[0065] 根据用户文档的主题结构，采用如下的公式一获取用户文档中出现概率最高的L个单词，作为用户文档的检索串Q；

[0066] $$Q = \arg\max P(Q|D) = \arg\max \prod_{i=1}^{L} P(q_i|D)$$ 公式一；

[0067] 其中，$P(q_i|D)$是用户文档D中单词q_i出现的概率，L为预设数量，i的取值范围为从1至L的正整数；$P(q_i|D) = \sum_{j=1}^{K} P(q_i|z_j,\varphi_j)P(z_j|\theta,D)$，$P(z_j|\theta,D)$是用户文档D中主题$z_j$出现的概率，θ是用户文档D的主题分布，$P(q_i|z_j,\varphi_j)$是主题$z_j$中单词$q_i$出现的概率，$\varphi_j$是主题$z_j$的单词分布，K为主题模型中主题的总数量，j的取值范围为从1至K的正整数。

[0068] 具体的，每次取L个概率$P(q_i|D)$，代入公式$\prod_{i=1}^{L} P(q_i|D)$中，即可得到多个P(Q|D)，再在多个P(Q|D)中获取$\max \prod_{i=1}^{L} P(q_i|D)$，当获取到最大的P(Q|D)时，用于生产最大的P(Q|D)的

L个单词，组成用户文档的检索串Q。

[0069] 进一步的，在上述实施例的基础上，采用基于主题模型的相似度计算方法，获取各信息文档与用户文档的相似度。

[0070] 示例性的，采用如下的公式二确定用户文档与任一信息文档D′的相似度Sim；

[0071] $$\mathrm{Sim}=\frac{\sum_{j=1}^{K}P(z_j\mid\theta,D)\cdot P(z_j\mid\theta',D')}{\sqrt{\sum_{j=1}^{K}P(z_j\mid\theta,D)^2}\cdot\sqrt{\sum_{j=1}^{K}P(z_j\mid\theta',D')^2}}$$ 公式二；

[0072] 其中，$P(z_j|\theta',D')$表示任一信息文档D′中主题z_j出现的概率，θ′为信息文档D′的主题分布。

[0073] 具体的，在获取任一信息文档D′与用户文档D的相似度时，也可基于主题模型，利用信息文档D′的主题结构与用户文档D的主题结构，计算两文档的相似度。示例性的，当两篇文档中的主题分布相同，且单词分布相同，则认为两篇文章在内容上越相似。如公式二所示，当两篇文档中的各主题的分布越相似，则Sim的值越趋近于1。

[0074] 可选的，在上述任一实施例的基础上，结合图1或图2所示实施例对本发明实施例的获取检索串之前的准备步骤进行详细说明。图3为本发明提供的基于主题模型的软件问答推荐方法实施例三的流程示意图，如图3所示，本发明提供的基于主题模型的软件问答推荐方法，包括：

[0075] 步骤301、采用隐含狄利克雷分布算法，获取软件问答信息库的主题模型；

[0076] 步骤302、确定用户文档的变化量达到预设变化量；

[0077] 步骤303、根据用户文档和软件问答信息库的主题模型，获取用户文档的检索串；

[0078] 步骤304、根据检索串，在软件问答信息库中获取至少一个信息文档；

[0079] 步骤305、获取各信息文档与用户文档的相似度，根据各相似度，向用户进行软件问答推荐。

[0080] 具体的，在获取用户文档的检索串之前，需采用隐含狄利克雷分布算法，对开发人员确定的软件问答信息库进行训练，得到软件问答信息库的主题模型，并定期对主题模型进行更新。当检测到开发人员在用户文档中进行代码编写时，对开发人员的操作进行实时监控，当检测到开发人员对用户文档进行了修改，且确定用户文档的变化量达到预设变化量时，开始确定检索串，为开发人员提供信息文档，以方便开发人员在遇到问题时，无需再手动检索。示例性的，可以为在检测到用户文档中的代码变化量达到两行时，开始执行获取检索串的操作。

[0081] 进一步的，在上述任一实施例的基础上，针对获取相似度的过程进行详细说明。图4为本发明提供的基于主题模型的软件问答推荐方法实施例四的流程示意图。如图4所示，获取相似度的过程具体包括：

[0082] 步骤401、针对任一信息文档，获取信息文档与用户文档的相似度Sim；

[0083] 步骤402、根据信息文档的至少一项元信息特征，对信息文档的相似度Sim进行修正，得到修正后的相似度Sim′；

[0084] 步骤403、根据各修正后的相似度Sim′,向用户进行软件问答推荐。

[0085] 具体的,针对任一信息文档,可采用上述实施例所述的基于主题模型的方法,获取信息文档与用户文档的相似度Sim。考虑到软件问答信息库中的各信息文档除了包含开发人员分享的用于解决技术问题的内容外,还包括文档发布者的信息,该文档被查阅、引用次数的信息等。该些元信息特征也可用于评价信息文档的可参考性。因此,在获取到相似度Sim后,进一步根据信息文档的至少一项元信息特征,对信息文档的相似度Sim进行修正,得到修正后的相似度Sim′。最后根据各修正后的相似度Sim′,向用户进行软件问答推荐。

[0086] 例如,当获取到两篇相似度分别为S1和S2的信息文档D1和D2,且S1>S2时,考虑到信息文档S1的发布者的声望F1远大于信息文档S2的发布者的声望F2,则根据F1和F2对原本的相似度S1和S2进行修订,得到修订后的相似度S11、S22,且修订后的S11<S22,则按照先信息文档D2、后信息文档D1的顺序进行软件问答推荐。

[0087] 示例性的,在上述实施例的基础上,对相似度进行修正的过程,具体包括:

[0088] 根据信息文档的至少一项元信息特征t_m,采用如下的公式三对信息文档的相似度Sim进行修正,得到修正后的相似度Sim′;

[0089] $$Sim' = \alpha \cdot Sim + (1-\alpha)\sum_{m=1}^{M} \beta_m \cdot t_m$$ 公式三;

[0090] 其中,α为信息文档的相似度Sim的权重,M为信息文档的元信息特征的总数量,m的取值为从1至M的正整数,β_m为元信息特征t_m的权重,α和β_m的取值为从0至1的实数。

[0091] 可选的,当信息文档为问答网站中的文档时,元信息特征包括如下中的至少一项:

[0092] 信息文档中的问题质量、答案质量、提问用户声望、回答用户声望、应用程序接口(Application Programming Interface,简称API)相似度、文本相似度。

[0093] 其中,问题质量指问答网站的用户对于问题的评分,答案质量指问答网站的用户对于答案的评分,提问用户声望指问答网站的其他用户对提出问题的用户的评分,回答用户声望指问答网站的其他用户对提供答案的用户的评分,API相似度指信息文档中若包含代码时,信息文档所包含的代码与用户文档中的代码的相似度,文本相似度指信息文档的词向量与用户文档的词向量的相似度。

[0094] 示例性的,可先对上述元信息特征进行归一化,然后根据归一化后的元信息特征对相似度进行修正。

[0095] 本发明另一方面提供一种基于主题模型的软件问答推荐装置,用于执行如上述实施例所述的基于主题模型的软件问答推荐方法,具有相同的技术特征和技术效果,本发明不再赘述。

[0096] 图5本发明提供的基于主题模型的软件问答推荐装置实施例一的结构示意图。如图5所示,包括:

[0097] 检索串获取模块501,用于根据用户文档和软件问答信息库的主题模型,获取用户文档的检索串;

[0098] 信息文档获取模块502,用于根据检索串,在软件问答信息库中获取至少一个信息文档;

[0099] 推荐模块503,用于获取各信息文档与用户文档的相似度,根据各相似度,向用户进行软件问答推荐。

[0100] 可选的，检索串获取模块501具体用于：

[0101] 根据主题模型，获取用户文档的主题结构；

[0102] 根据用户文档的主题结构，获取用户文档中出现概率最高的预设数量个单词，作为用户文档的检索串。

[0103] 可选的，检索串获取模块501具体用于：

[0104] 根据用户文档的主题结构，采用如下的公式一获取用户文档中出现概率最高的L个单词，作为用户文档的检索串Q；

[0105] $$Q = \arg\max P(Q|D) = \arg\max \prod_{i=1}^{L} P(q_i|D)$$ 公式一；

[0106] 其中，$P(q_i|D)$是用户文档D中单词q_i出现的概率，L为预设数量，i的取值范围为从1至L的正整数；$P(q_i|D) = \sum_{j=1}^{K} P(q_i|z_j,\varphi_j)P(z_j|\theta,D)$，$P(z_j|\theta,D)$是用户文档D中主题$z_j$出现的概率，θ是用户文档D的主题分布，$P(q_i|z_j,\varphi_j)$是主题$z_j$中单词$q_i$出现的概率，$\varphi_j$是主题$z_j$的单词分布，K为主题模型中主题的总数量，j的取值范围为从1至K的正整数。

[0107] 可选的，推荐模块503具体用于：

[0108] 采用如下的公式二确定用户文档与各信息文档的相似度Sim；

[0109] $$Sim = \frac{\sum_{j=1}^{K} P(z_j|\theta,D) \cdot P(z_j|\theta',D')}{\sqrt{\sum_{j=1}^{K} P(z_j|\theta,D)^2} \cdot \sqrt{\sum_{j=1}^{K} P(z_j|\theta',D')^2}}$$ 公式二；

[0110] 其中，$P(z_j|\theta',D')$表示任一信息文档D′中主题z_j出现的概率，θ′为信息文档D′的主题分布。

[0111] 可选的，该装置还包括：

[0112] 主题模型获取模块，用于采用隐含狄利克雷分布算法，获取软件问答信息库的主题模型；

[0113] 变化检测模块，用于在确定用户文档的变化量达到预设变化量时，控制检索串获取模块501执行获取用户文档的检索串的操作。

[0114] 可选的，推荐模块503具体用于：

[0115] 针对任一信息文档，获取信息文档与用户文档的相似度Sim；

[0116] 根据信息文档的至少一项元信息特征，对信息文档的相似度Sim进行修正，得到修正后的相似度Sim′；

[0117] 根据各修正后的相似度Sim′，向用户进行软件问答推荐。

[0118] 可选的，推荐模块503具体用于：

[0119] 根据信息文档的至少一项元信息特征t_m，采用如下的公式三对信息文档的相似度Sim进行修正，得到修正后的相似度Sim′；

[0120] $$Sim' = \alpha \cdot Sim + (1-\alpha)\sum_{m=1}^{M} \beta_m \cdot t_m$$ 公式三；

[0121] 其中，α为信息文档的相似度Sim的权重，M为信息文档的元信息特征的总数量，m的取值为从1至M的正整数，β_m为元信息特征t_m的权重，α和β_m的取值为从0至1的实数。

[0122] 可选的，软件问答信息库中包括如下至少一种文档：百科文档、网络贴吧文档、网络社区文档、问答网站中的文档。

[0123] 可选的，当信息文档为问答网站中的文档时，元信息特征包括如下中的至少一项：

[0124] 信息文档中的问题质量、答案质量、提问用户声望、回答用户声望、应用程序接口相似度、文本相似度。

[0125] 在本申请所提供的几个实施例中，应该理解到，所揭露的系统、装置和方法，可以通过其它的方式实现。例如，以上所描述的装置实施例仅仅是示意性的，例如，所述单元的划分，仅仅为一种逻辑功能划分，实际实现时可以有另外的划分方式，例如多个单元或组件可以结合或者可以集成到另一个系统，或一些特征可以忽略，或不执行。另一点，所显示或讨论的相互之间的耦合或直接耦合或通信连接可以是通过一些接口，装置或单元的间接耦合或通信连接，可以是电性，机械或其它的形式。

[0126] 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单元来实现本实施例方案的目的。

[0127] 另外，在本发明各个实施例中的各功能单元可以集成在一个处理单元中，也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在一个单元中。

[0128] 所述功能如果以软件功能单元的形式实现并作为独立的产品销售或使用时，可以存储在一个计算机可读取存储介质中。基于这样的理解，本发明的技术方案本质上或者说对现有技术做出贡献的部分或者该技术方案的部分可以以软件产品的形式体现出来，该计算机软件产品存储在一个存储介质中，包括若干指令用以使得一台计算机设备(可以是个人计算机，服务器，或者网络设备等)执行本发明各个实施例所述方法的全部或部分步骤。而前述的存储介质包括：U盘、移动硬盘、只读存储器(Read-Only Memory，简称ROM)、随机存取存储器(Random Access Memory，简称RAM)、磁碟或者光盘等各种可以存储程序代码的介质。

[0129] 最后应说明的是：以上各实施例仅用以说明本发明的技术方案，而非对其限制；尽管参照前述各实施例对本发明进行了详细的说明，本领域的普通技术人员应当理解：其依然可以对前述各实施例所记载的技术方案进行修改，或者对其中部分或者全部技术特征进行等同替换；而这些修改或者替换，并不使相应技术方案的本质脱离本发明各实施例技术方案的范围。

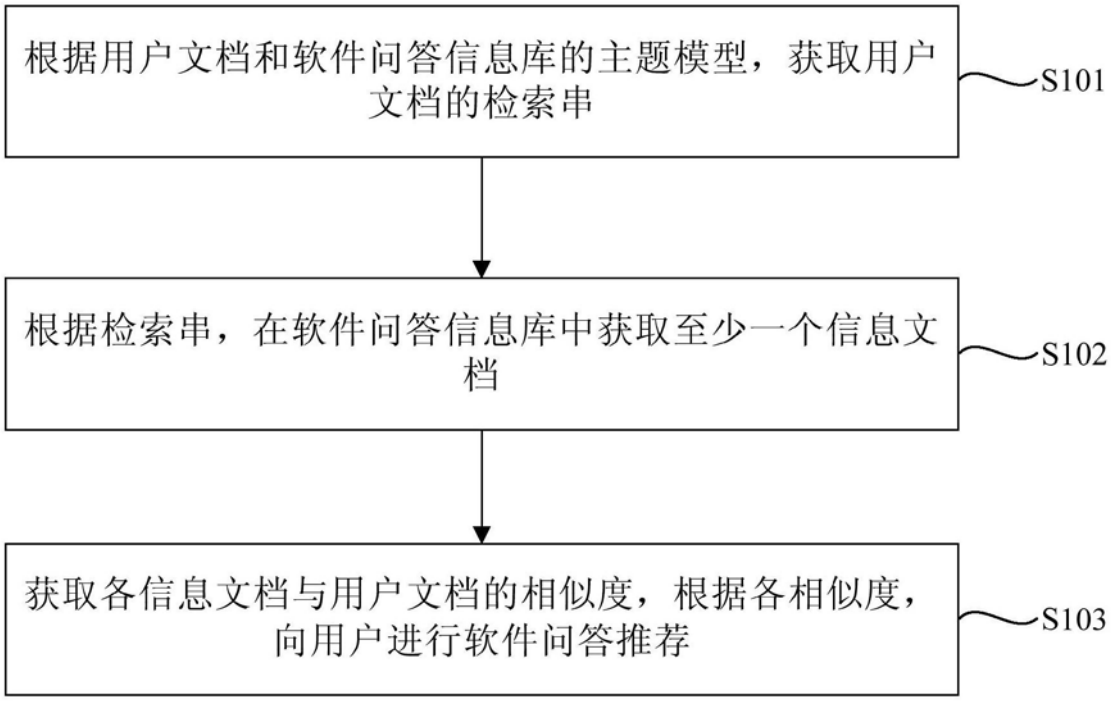

根据用户文档和软件问答信息库的主题模型，获取用户文档的检索串
S101
根据检索串，在软件问答信息库中获取至少一个信息文档
S102
获取各信息文档与用户文档的相似度，根据各相似度，向用户进行软件问答推荐
S103

图1

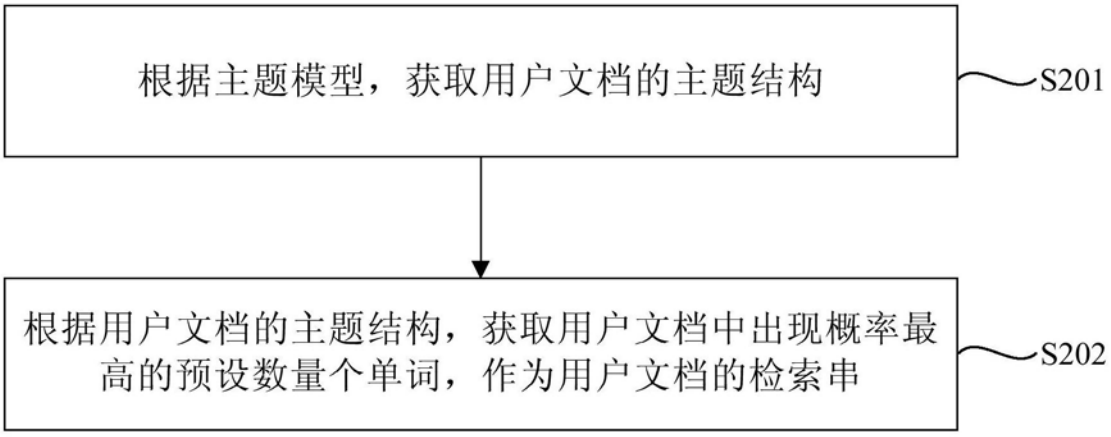

根据主题模型，获取用户文档的主题结构
S201
根据用户文档的主题结构，获取用户文档中出现概率最高的预设数量个单词，作为用户文档的检索串
S202

图2

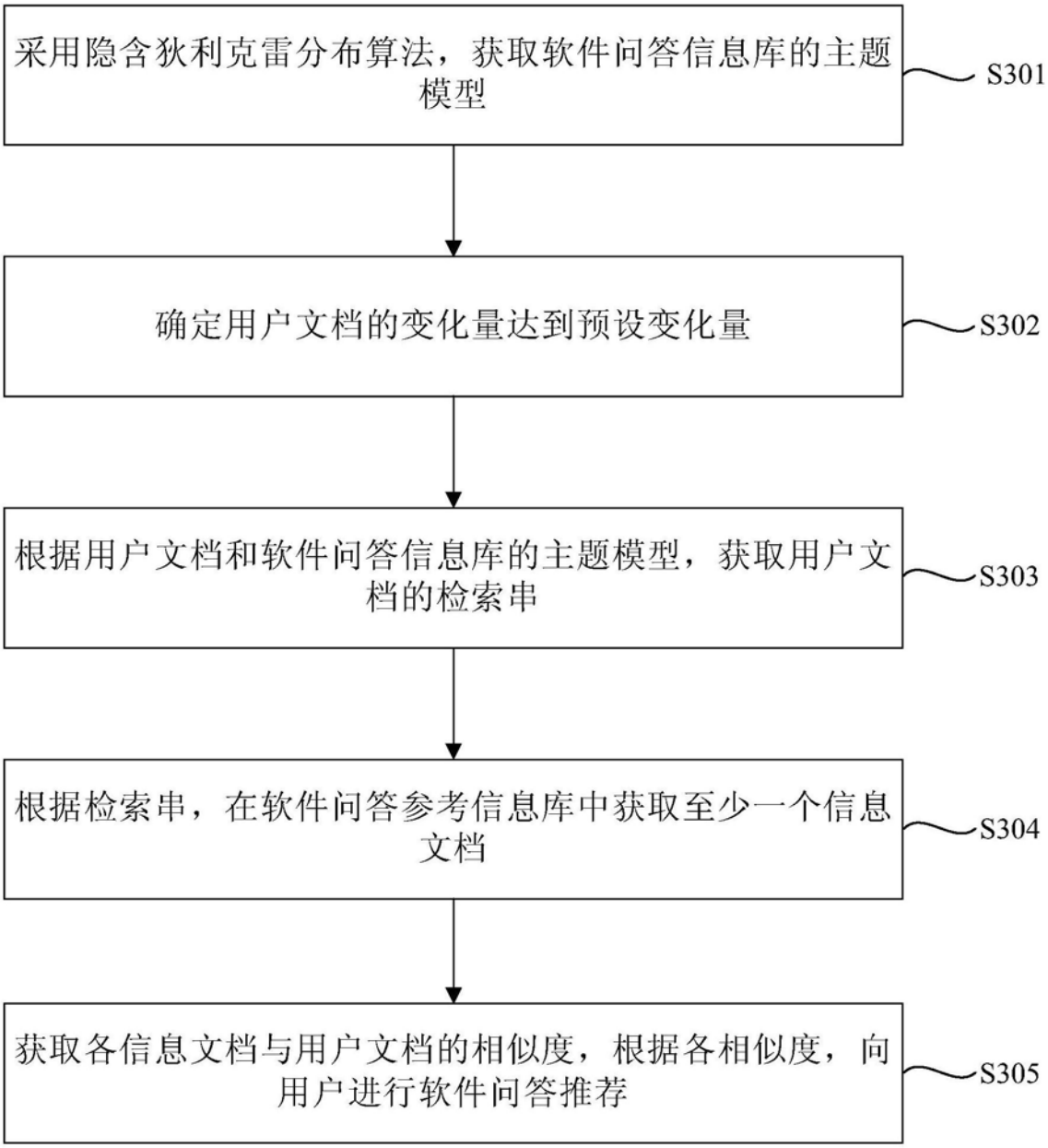

采用隐含狄利克雷分布算法，获取软件问答信息库的主题模型
S301
确定用户文档的变化量达到预设变化量
S302
根据用户文档和软件问答信息库的主题模型，获取用户文档的检索串
S303
根据检索串，在软件问答参考信息库中获取至少一个信息文档
S304
获取各信息文档与用户文档的相似度，根据各相似度，向用户进行软件问答推荐
S305

图3

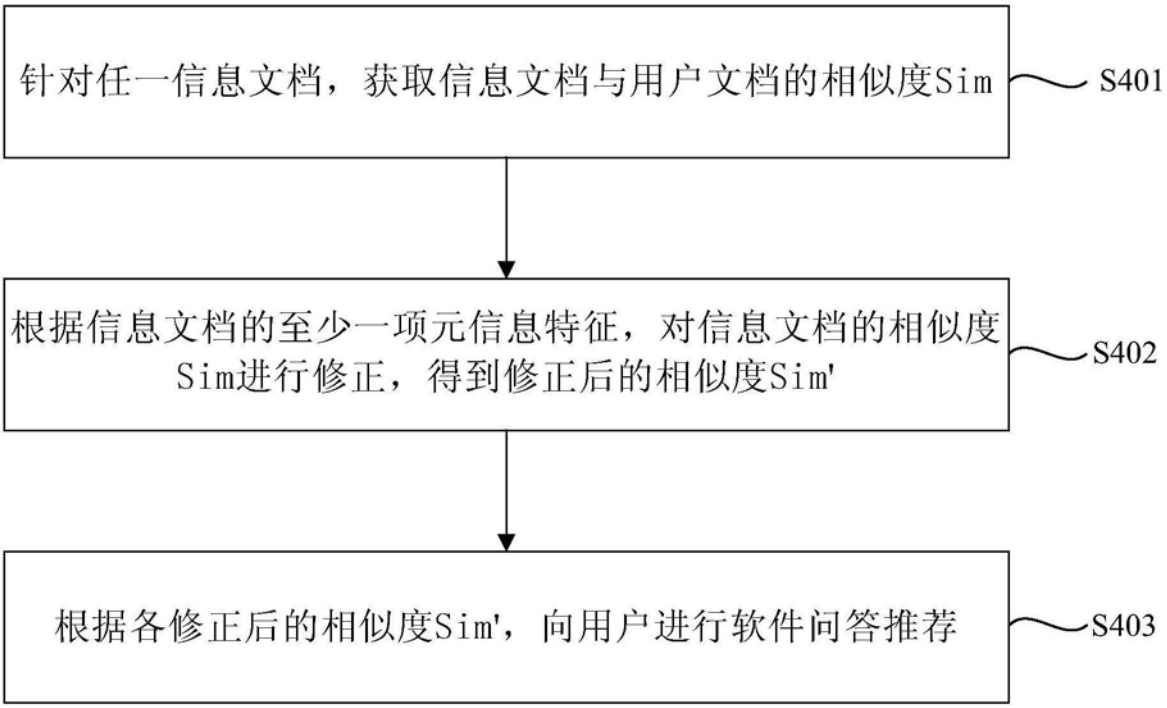

针对任一信息文档，获取信息文档与用户文档的相似度Sim
S401
根据信息文档的至少一项元信息特征，对信息文档的相似度Sim进行修正，得到修正后的相似度Sim'
S402
根据各修正后的相似度Sim'，向用户进行软件问答推荐
S403

图4

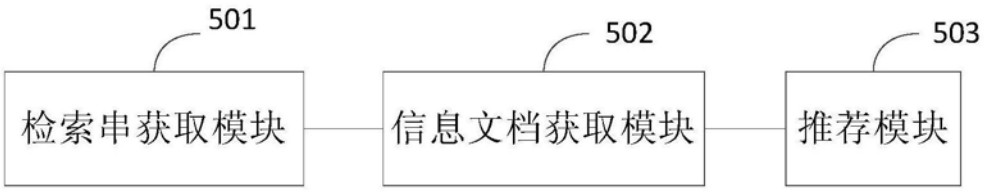

501
检索串获取模块
502
信息文档获取模块
503
推荐模块

图5