



(12) 发明专利申请

(10) 申请公布号 CN 102681979 A

(43) 申请公布日 2012.09.19

(21) 申请号 201210150808.9

(22) 申请日 2012.05.15

(71) 申请人 北京师范大学

地址 100875 北京市海淀区新街口外大街
19号

(72) 发明人 杨现民 余胜泉

(74) 专利代理机构 北京科迪生专利代理有限责
任公司 11251

代理人 许玉明

(51) Int. Cl.

G06F 17/24(2006.01)

G06F 17/27(2006.01)

G06F 17/30(2006.01)

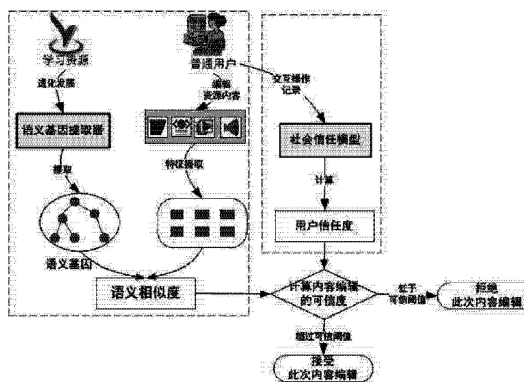
权利要求书 1 页 说明书 6 页 附图 3 页

(54) 发明名称

一种面向开放知识社区的内容编辑智能审核方法

(57) 摘要

一种面向开放知识社区的内容编辑智能审核方法,包括以下步骤:(1)抽取资源内容的语义特征信息,计算新添加内容的特征项集与语义基因的语义相似度;(2)基于用户在开放知识社区中的交互操作数据,应用信任评估模型计算用户的信任度;(3)综合语义相似度和用户信任度两个指标,判断是否自动接受或拒绝此次内容编辑。本发明的内容编辑智能审核方法,可以有效减轻开放知识社区中人工审核内容编辑的负担,且具有较高的准确率。



1. 一种面向开放知识社区的内容编辑智能审核方法,其特征包括以下步骤:

(1) 抽取资源内容的语义特征信息,计算新添加内容的特征项与语义基因的语义相似度;

(2) 基于用户的交互操作数据,应用信任评估模型计算用户的信任度;

(3) 综合语义相似度和用户信任度两个指标,判断是否接受此次内容编辑。

2. 根据权利要求1所述的一种面向开放知识社区的内容编辑智能审核方法,其特征在在于:在所述步骤(1)中的语义基因是指学习资源背后的内在知识结构,能够反映资源所要表达的核心内容,形式化表示为基于本体描述的带有权重的概念集合以及概念间的语义关系。

3. 根据权利要求1所述的一种面向开放知识社区的内容编辑智能审核方法,其特征在在于:在所述步骤(1)中的语义基因提取方法为:基于领域本体提取资源内容的特征项;根据特征评价函数计算特征项的权重;将特征项映射为领域本体中的概念;基于语义网开源框架—JENA 框架提取概念在本体库中存在的语义关系。

4. 根据权利要求1所述的一种面向开放知识社区的内容编辑智能审核方法,其特征在在于:在所述步骤(1)中的新添加内容的特征项集与语义基因的语义相似度计算方法为:应用同义词词林将特征项集和语义基因中的同义词进行替换;采用余弦系数法计算语义相似度。

5. 根据权利要求1所述的一种面向开放知识社区的内容编辑智能审核方法,其特征在在于:在所述步骤(2)中的信任评估模型包含资源信任度和用户信任度两个核心部件;资源信任度依据直接信任度评价数据和间接信任度评价数据采用加权求和的方法计算得出,所述直接信任度评价数据是用户直接对资源信任度进行打分,所述间接信任度评价数据是用户对资源的交互操作信息;用户信任度依据用户之间的交互信息及其创作知识的可信度采用加权求和的方法计算得出;资源信任度和用户信任度之间相互影响,采用迭代逼近法交叉计算二者的信任度。

6. 根据权利要求5所述的一种面向开放知识社区的内容编辑智能审核方法,其特征在在于:所述的迭代逼近法是通过设定一个最大误差值 \max_error ,通过多次的迭代计算,至到前后两次计算结果中对应的所有信任度之差的绝对值都小于 \max_error ,才结束运算,生成逼近真实的资源信任度和用户信任度。

7. 根据权利要求1所述的一种面向开放知识社区的内容编辑智能审核方法,其特征在在于:在所述步骤(3)中的判断是否接受此次内容编辑的方法为:设定一个内容编辑能够被接受的最低阈值 AT ;对语义相似度和用户信任度两个指标采用加权求和法;若计算结果大于或等于 AT ,则接受此次内容编辑;否则,拒绝此次内容编辑。

8. 根据权利要求4所述的一种面向开放知识社区的内容编辑智能审核方法,其特征在在于:所述同义词词林为哈工大扩展版的同义词词林。

一种面向开放知识社区的内容编辑智能审核方法

技术领域

[0001] 本发明属于计算机科学与技术领域,具体地说,一种面向开放知识社区的内容编辑智能审核方法,应用于各种开放知识社区的资源建设与管理。

背景技术

[0002] 近年来,以维基百科(Wikipedia)为代表的开放知识社区越来越流行,国外的有Google Knol、Cloudworks、Freebase等,国内的有百度知道、百度百科、互动百科、学习元平台等。开放的知识平台固然有其独特的优势,可以吸引大量用户参与、发挥集体智慧、促进知识流通与分享。然而,开放的同时也带来了不可回避的麻烦。维基百科完全开放式的编辑与组织方式使开放知识社区中资源的质量和可靠性问题成为关注的焦点。Web2.0时代人人可以生产、消费、传播资源,用户群体的复杂性和生产的自由化直接导致开放知识社区中资源数量的爆炸式增长和资源质量的良莠不齐。

[0003] 为了保证开放知识社区中的资源质量的可靠性,当前各种开放知识社区主要采用人工决策与手动内容编辑审核的方式来实现对资源内容质量的控制。维基百科采用的是一套基于人工协作的协调机制和一系列的约束规则(3R规则、事实校验和实时同级评审规则等)来实现信息的有序进化。其它各种开放知识社区(百度百科、Google Knol、Freebase等)也基本上沿用维基百科的控制模式,在反馈、交流的基础上最终通过人工审核的方式实现内容版本的不断更迭和进化。近年来,有研究者(Javanmardi et al., 2010; Moturu&Liu, 2009)开始尝试构建维基百科中的信任评估模型,来帮助用户筛选高质量的资源。但是,这些信任评估模型大都是基于编辑历史数据构建的,忽视了其他更加丰富的有助于判断用户和资源信任度的交互性信息,比如收藏、订阅、分享等操作数据。

[0004] 随着资源群体和用户群体规模的不断扩大,内容编辑审核的工作量将急速增加,必然给知识的创作者、管理者带来巨大的工作负担。当前的人工决策与内容审核技术已经无法适应开放知识社区发展的需要,因此,急需探索一种可以有效减轻开放知识社区中用户进行内容编辑审核负担和压力的智能化方法。

发明内容

[0005] 本发明要解决的技术问题是:克服现有人工审核内容编辑的不足,提供一种智能化的内容编辑审核方法,该方法可以有效减轻开放知识社区中人工审核内容编辑的负担,且具有较高的准确率。

[0006] 本发明解决其技术问题所采用的技术方案是:一种面向开放知识社区的内容编辑智能审核方法,其特征在于包括以下步骤:

[0007] (1) 抽取资源内容的语义特征信息(语义基因),计算新添加内容的特征项与语义基因的语义相似度;

[0008] (2) 基于用户的交互操作数据,应用信任评估模型计算用户的信任度;

[0009] (3) 综合语义相似度和用户信任度两个指标,判断是否接受此次内容编辑。

[0010] 上述步骤(1)中的语义基因是指学习资源背后的内在知识结构,能够反映资源所要表达的核心内容,可以形式化表示为基于本体描述的带有权重的概念集合以及概念间的语义关系。语义基因提取方法为:基于领域本体提取资源内容的特征项;根据特征评价函数计算特征项的权重;将特征项映射为领域本体中的概念;基于 JENA 框架提取概念在本体库中存在的语义关系。所述的特征评价函数采用词频统计 TF(Term Frequency)法。

[0011] 上述步骤(1)中的新添加内容的特征项集与语义基因的语义相似度计算方法为:应用哈工大扩展版的同义词词林将特征项集和语义基因中的同义词进行替换;采用余弦系数法计算语义相似度。

[0012] 上述步骤(2)中的信任评估模型核心设计思路是:包含资源信任度和用户信任度两个核心部件;资源信任度依据直接信任度评价数据(用户直接对资源信任度进行打分)和间接信任度评价数据(用户对资源的交互操作信息,如订阅、收藏等)采用加权平均的方法计算得出;用户信任度依据用户之间的交互信息和其创作知识的可信度采用加权平均的方法计算得出;资源信任度和用户信任度之间相互影响,采用迭代逼近法交叉计算二者的信任度。迭代逼近法核心思想是通过设定一个最大误差值 max_error,通过多次的迭代计算,至到前后两次计算结果中对应的所有信任度之差的绝对值都小于 max_error,才结束运算,生成逼近真实的资源信任度和用户信任度。

[0013] 上述步骤(3)中的内容编辑是否接受的判断方法为:设定一个内容编辑可以被接受的最低阈值 AT;对语义相似度和用户信任度两个指标采用加权求和法;若计算结果大于或等于 AT,则接受此次内容编辑;否则,拒绝此次内容编辑。

[0014] 本发明与现有技术相比的有益效果:

[0015] (1)本发明方法通过综合应用语义基因和信任评估模型,可以实现对开放知识社区中内容编辑的自动审核,大大减轻人工审核的负担和压力,同时具有较高的准确性。

[0016] (2)本发明基于领域本体提取资源内容的语义特征信息(语义基因),较之常规的文本特征提取技术,能达到更好的提取效果。

[0017] (3)本发明提出的信任评估模型,可以有效评价开放知识社区中用户信任度和资源信任度,具有计算数据丰富、考虑到用户信任和资源信任间的联动关系、更加接近现实社会中的信任关系等特点。

附图说明

[0018] 图 1 为本发明的语义基因的提取过程;

[0019] 图 2 为本发明的基于领域本体的特征项提取过程;

[0020] 图 3 为本发明提出的信任评估模型;

[0021] 图 4 为本发明提出的迭代逼近法解决交叉计算问题的过程;

[0022] 图 5 为本发明的内容编辑智能审核过程。

具体实施方式

[0023] 下面结合附图及具体实施方式详细介绍本发明。

[0024] 本发明的一种面向开放知识社区的内容编辑智能审核方法,采用如下步骤:

[0025] (1)抽取语义基因,计算语义相似度

[0026] 语义基因的概念是指学习资源背后的内在知识结构,能够反映资源所要表达的核心内容。区别于文本相似度比较中的文档特征项,语义基因不是简单的关键词集合,而是资源背后所隐藏的语义概念网络。

[0027] 语义基因可以被形式化地表示为有序三元组,即 $SG = \langle CS, WS, RS \rangle$, 其中 CS 是核心概念集合,集合大小不超过 10, $CS = \{C_1, C_2, C_3, \dots, C_n\}$, $1 \leq n \leq 10$, WS 是概念项的权重集合, $WS = \{W_1, W_2, W_3, \dots, W_n\}$, 其中 W_i 为 C_i 的权重, $1 \leq i \leq n$, 且所有权重之和为 1; RS 为核心概念间的关系集, $RS = \{R_1, R_2, R_3, \dots, R_n\}$, 每个关系采用领域本体中的 RDF 三元组 $\langle \text{Subject}, \text{Predicate}, \text{Object} \rangle$ 表示, $R_1 = \langle \text{Concept}_1, \text{Relationship}, \text{Concept}_2 \rangle$, 这里的 Concept_1 和 Concept_2 不一定包含在 CS 中, 可以是领域本体库的其他概念, Relationship 是从领域本体库中提取的概念关系。

[0028] 语义基因提取的过程如图 1 所示,包含四个关键性步骤:基于领域本体的特征项提取;根据特征评价函数计算特征项的权重;特征词到本体概念的映射;基于语义网开源框架—JENA 提取特征项(概念)在本体库中存在的语义关系。基于领域本体的特征项提取流程见图 2,特征评价函数采用词频统计 TF(Term Frequency) 法。

[0029] 算法 1 基于 JENA 的特征词到本体概念的映射算法(Term Mapping to Concept, TM2C)

[0030] 输入:资源的特征项集合 $TS = \{t_i | i = 1, 2, 3, \dots, n\}$

[0031] 输出:资源的概念集合 $CS = \{C_j | j = 1, 2, 3, \dots, n\}$

[0032] 算法的伪代码如下:

[0033]

```
List<OntClass> CS; //定义返回的本体概念集合
```

```
String uri; //定义临时本体类的 uri 字符串
```

```
OntClass concept; //定义临时本体类
```

```
OntModel model = getModelFromDB ( ontology ); //获取 ontology 的 JENA 本体模型, 用于
```

[0034]

操作本体

```

    for( int i=1 to n ){
        uri = getClassURI ( TS[ i ], model ); //本体库中查找名字为 TS[i]的类是否存在, 获取
uri
        if(uri is null) { //不存在, 自动生成本体概念, 加入本体库
            concept = model.createClass ( uri );
        }else{ //存在则获取本体概念
            concept = model.getOntClass ( uri );
        }
        CS.add ( concept ); //添加到 CS 集合
    }

    return CS; //返回本体概念集合

```

[0035] 算法 1 使用语义 Web 领域流行的 JENA 框架, 将资源内容中提取的特征项逐个映射为领域本体中对应的概念。

[0036] 算法 2 基于 JENA 的概念关系抽取 (Concept Relationship Extraction, CRE)

[0037] 输入: 资源的概念词集合 $CS = \{C_j | j = 1, 2, 3, \dots, n\}$

[0038] 输出: 概念关系集合 $RS = \{R_j | j = 1, 2, 3, \dots, m\}$

[0039] 算法的伪代码如下:

[0040]

List<Statement> RS; //定义返回的关系集合, Statement 为 JENA 中的内置对象, 表示一个三元组描述的陈述

List<Statement> tmpList; //定义临时 RDF 三元组集合

OntModel model = getModelFromDB (ontology); //获取 ontology 的 Jena 本体模型, 用于操作本体

while (c in CS){ //循环读取 CS 中的元素

 If (c exists in OntModel) { //如果 c 在本体库中存在

[0041]

```

tmpList = OntModel.listStatements( c ); //列出 Subject 为 c 的所有 RDF 三元组

    While (Statement s in tmpList) {
        if ( s.object is concept ){ //将三元组中 Object 是概念的 Statement 加入到 RS
中
            RS.add (s);
        }
    }
}

return RS; //返回概念关系集合

```

[0042] 算法 2 使用语义 Web 领域流行的 JENA 框架,将算法 1 获取到的概念集合在领域本体中存在的概念关系提取出来。

[0043] 算法 3 语义基因提取算法

[0044] 输入:资源的 Title、Tag、Content 和 SemanticData

[0045] 输出:资源的语义基因 SG=<CS, WS, RS>

[0046] 关键步骤:

[0047] 步骤 1 调用 ICTCLAS 将 Title 进行分词处理和噪音过滤

[0048] 步骤 2 调用 ICTCLAS 将 Tag 进行切割和噪音过滤

[0049] 步骤 3 调用 ICTCLAS 对 Content 进行 html 标签过滤,分词处理,噪音过滤(去除虚词)

[0050] 步骤 4 获取语义描述信息中的本体类

[0051] 步骤 5 对 Step2 到 Step5 中得到的特征词集合,结合领域本体进行词语组合,识别新的特征词

[0052] 步骤 6 调用停用词表,将 Step6 得到的词语集合进行停用词过滤

[0053] 步骤 7 结合哈工大的扩展版同义词词林和领域本体进行同义词替换,得到特征词集 TS

[0054] 步骤 8 应用特征评价函数计算各特征词的权重,得到特征词的权重集合 WS

[0055] 步骤 9 应用 TM2C 算法得到概念集合 CS

[0056] 步骤 10 应用 CRE 算法提取概念关系集合 RS

[0057] 步骤 11 算法结束,输出 CS、WS 和 RS

[0058] 采用余弦系数法计算当前学习资源的语义基因和用户新添加内容的文本特征项集的语义相似度。用 X 表示资源的语义基因: $X=(C_1, WC_1; C_2, WC_2; \dots; C_n, WC_n)$, 其中 C_k 是语义基因中的概念项, WC_k 是 C_k 的权重, $1 \leq k \leq n$; 用 Y 表示新内容的文本特征向量: $Y=(T_1, WT_1; T_2, WT_2; \dots; T_m, WT_m)$, 其中 T_k 是特征向量中的特征词, WT_k 是 T_k 的权重, $1 \leq k \leq m$ 。进行余弦相似度计算之前,需要将 X 和 Y 中的元素进行同义词替换。同义词替换一方面基于 LCS 中内置的哈工大扩展版的同义词词林,另一方面,可以从资源语义基因中包含的概念关

系三元组中查找 X 和 Y 中的同义词, 并进行替换。语义相似度计算公式如下:

$$[0059] \quad SIM(X, Y) = \text{syn_replace}(X, Y) = \cos(\theta) = \frac{\vec{X} \cdot \vec{Y}}{\|X\| \cdot \|Y\|}$$

[0060] (2) 应用信任评估模型计算用户的信任度

[0061] 本发明提出一种面向开放知识社区的信任评估模型——双向互动反馈模型 (Tway Interactive Feedback Model, TIFM), 见图 3。TIFM 包括资源信任度和用户信任度两个核心部件, 二者相互影响; 两侧是信任度的各项影响因素; 中心是有关信任评估的四条假设。需要说明的是, 这里的信任指的是全局信任, 而非 P2P 网络中两个对等节点间的信任关系。资源信任度表示所有社区用户对资源节点的整体信任评价, 用户信任度表示社区中的所有其他用户对当前用户的整体信任评价。

[0062] 资源信任度 (Resource Trust, RT) 包含直接信任度和间接信任度两部分: 直接信任度 (Direct Resource Trust, DRT) 根据用户进行的显性信任评价计算得出; 间接信任度评价 (Indirect Resource Trust, IRT) 根据用户与资源的交互记录计算得出。在资源创建初期, 由于用户参与直接信任评价的次数较少, 因此 DRT 在资源的总体信任度评价中所占权重 w 偏低。权重 w 是以直接信任评价次数为自变量的增函数, 将根据直接评价次数的变化动态调整, 随着直接信任评价次数的增多, DRT 将越来越能够代表 RT, w 值也将逐渐提高。资源信任度的计算公式可以表示为: $RT = w \times DRT + (1-w) \times IRT$ 。DRT 采用加权平均法; IRT 先采用加权求和法, 然后进行归一化处理。

[0063] 开放知识社区中的用户信任度表征为四元组: $UT = \{UT_{res}, UT_{col}, UT_{fri}, UT_{rev}\}$, UT_{res} 表示由用户所创建资源的信任度计算得出的用户信任度, UT_{col} 表示由用户之间协作关系计算得出的用户信任度, UT_{fri} 表示由用户之间的好友关系计算得出的用户信任度, UT_{rev} 表示由编辑历史计算得出的用户信任度。用户信任度的计算公式采用加权求和法, 可以表示为: $UT = UW1 \times UT_{res} + UW2 \times UT_{col} + UW3 \times UT_{fri} + UW4 \times UT_{rev}$ 。

[0064] TIFM 存在交叉计算的问题, 即资源信任度的计算用到了用户信任度, 而用户信任度的计算又用到了资源的信任度。本发明采用迭代逼近法解决此问题, 流程见图 4, 核心思路是通过多次迭代计算系统中所有资源和所有用户的信任度, 直到前后两次计算结果中各项信任值之差的绝对值都小于设定的最大误差值, 表明所有信任值的计算结果趋于稳定, 接近真实值。

[0065] (3) 内容编辑智能审核

[0066] 内容编辑智能审核的流程见 5。内容编辑的可信度通过语义相似度和用户信任度两个指标进行加权求和得出, 公式可以表示为: $CT = W1 \times SS + W2 \times UT$, SS 表示资源语义基因与新添加内容文本特征项集的语义相似度, UT 表示用户的信任度, $W1$ 为语义相似度所占权重, $W2$ 为用户信任度所占权重, 其中 $W1 + W2 = 1$, $CT \in [0, 1]$ 。如果 CT 大于或等于 AT , 则此次内容编辑将被系统自动审核通过; 否则, 将被系统自动拒绝。

[0067] 其中 AT 为预设的内容编辑可接受阈值。 AT 可以根据实际应用效果进行调整。 AT 可用于调控内容编辑智能审核的严厉程度, AT 越高, 用户进行的内容编辑越不容易被自动接受; AT 越低, 用户进行的内容编辑将越不容易被自动拒绝。

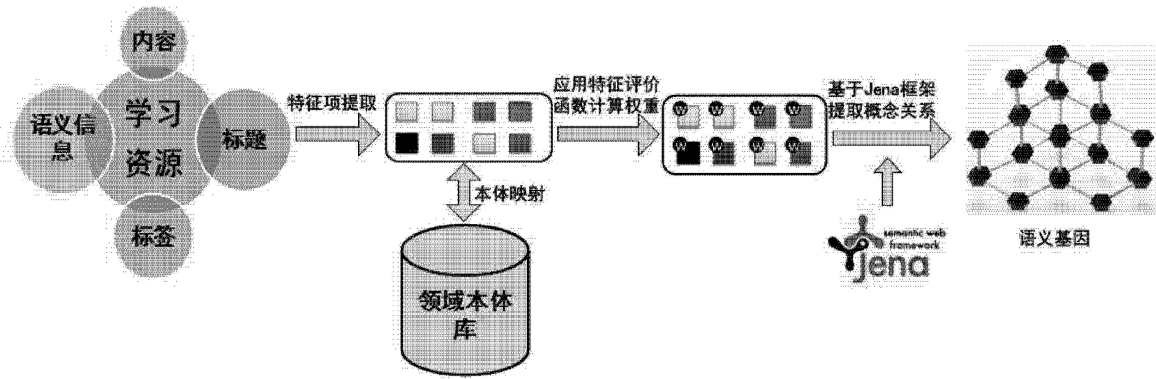


图 1

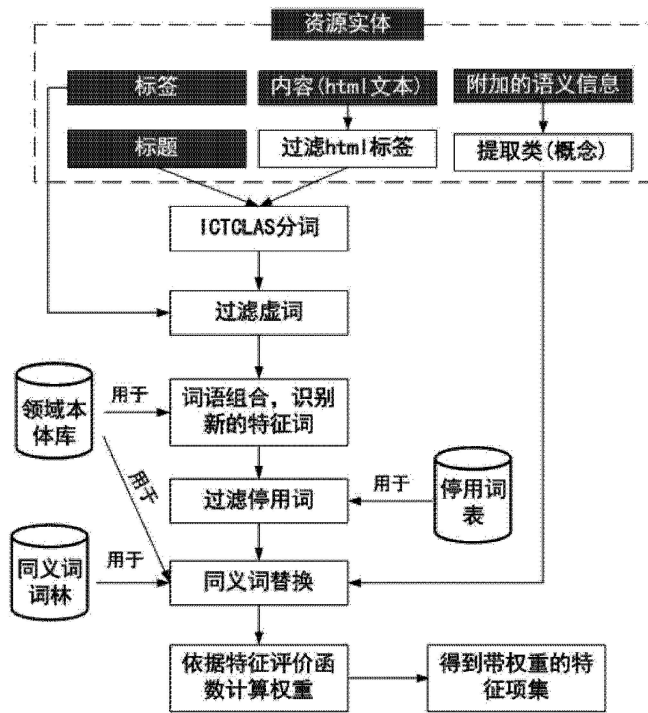


图 2

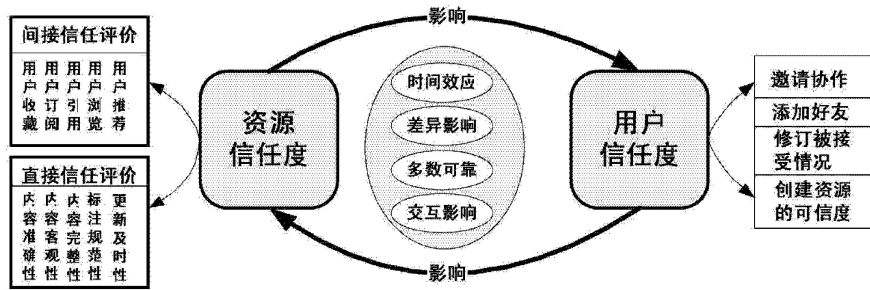


图 3

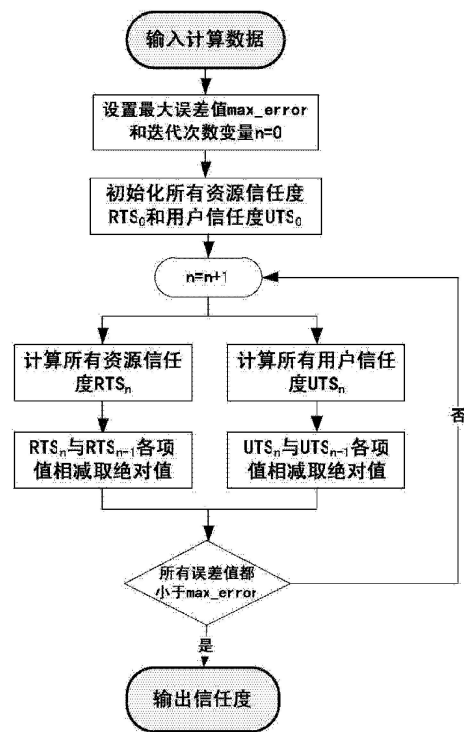


图 4

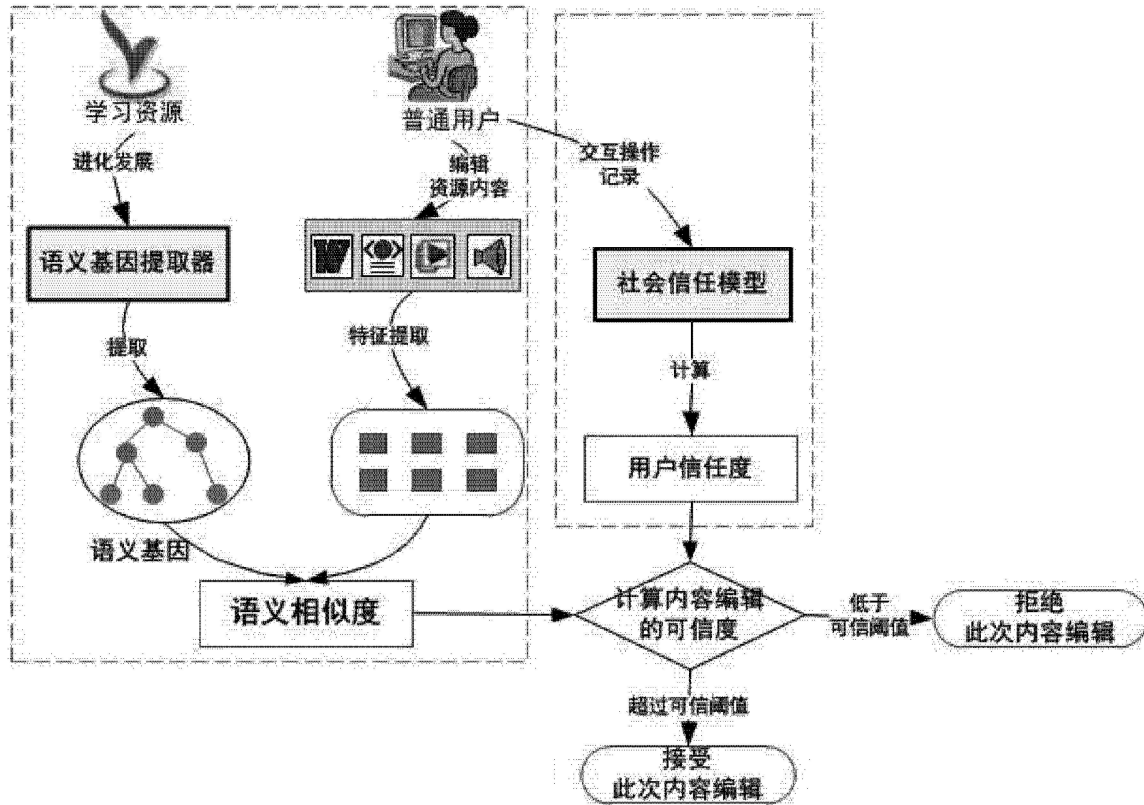


图 5