(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 103631859 B
(45)授权公告日 2017.01.11

(21)申请号 201310509358.2

(22)申请日 2013.10.24

(65)同一申请的已公布的文献号
申请公布号 CN 103631859 A

(43)申请公布日 2014.03.12

(73)专利权人 杭州电子科技大学
地址 310018 浙江省杭州市下沙高教园区2号大街

(72)发明人 徐小良 吴仁克 林建海 陈秋

(74)专利代理机构 杭州君度专利代理事务所(特殊普通合伙) 33240
代理人 杜军

(51)Int.Cl.
G06F 19/00(2011.01)

(56)对比文件
CN 101075942 A,2007.11.21,
CN 102880657 A,2013.01.16,
CN 102495860 A,2012.06.13,
CN 102855241 A,2013.01.02,
胡斌.科技项目评审专家推荐系统的研究与实现.《中国优秀硕士学位论文全文数据库(信息科技辑)》.2013,(第7期),全文.

审查员 梁静静

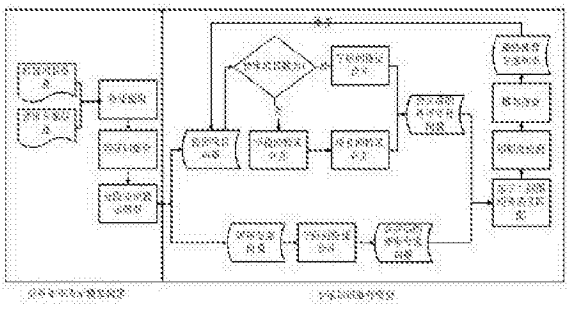
权利要求书5页 说明书13页 附图2页

(54)发明名称

一种面向科技项目的评审专家智能推荐方法

(57)摘要

本发明提供一种面向科技项目的评审专家智能推荐方法。本发明包括如下步骤:1).将待审科技项目、专家信息主要文本切分成子串序列并进行中科院ICTCLAS分词,对分词结果进行停用词过滤得到词语集合;2).构建项目信息的词语网络,基于统计特征和聚集特征提取特征词;专家信息较精简,直接将步骤1所得词语集合作为特征词;3).基于特征词所在字段和权值构建知识表示模型;建立相关信息索引;4).分组推荐专家对知识表示模型做字段间、项目间特征合并操作;5).基于语义计算专家与待审科技项目(组)的相似度,设定阀值截断产生最终推荐专家列表。上述方法的实现可极大地缓解推荐存在工作量大、评审决策缺乏科学性等问题。



CN 103631859 B

1.一种面向科技项目的评审专家智能推荐方法，其特征在于该方法包括以下步骤：

步骤1、把科技项目和专家信息中的通用词和惯用词作为专业停用词库；把标点符号、非汉字作为切分标记库；

步骤2、对科技项目信息、专家信息进行分词：根据科技项目信息中切分标记，将项目名称、主要研究内容、技术指标切分成子串序列；根据评审专家信息中切分标记，抽取专家信息、获奖情况、发明情况、发表论文情况、课题承担过的项目及完成情况、研究方向切分成子串序列，一个子串序列即一个字段信息；利用中科院ICTCLAS对子串序列进行分词；

步骤3、科技项目特征词语提取：利用通用停用词库和专业停用词库对分词进行停用词过滤，所述的通用停用词库采用哈工大停用词表，把去除停用词的分词结果作为一个词语集合；

专业停用词库的构建是一个自学习不断完善的过程，在信息分词过程中不断统计词语的词频，词语在文本出现的概率大于一定阈值，将它纳入到停用词库；

科技项目信息量较大，对词语集合进行词语间语义相似度计算，根据词的语义关系和词的共现关系构建词语网络，计算网络中的词语聚集特征值；然后结合词语的统计特征值，计算词语的关键度来提取出科技项目特征词语；科技项目的特征词语就是提取综合文本的统计特征信息和语义特征信息，更加准确地提取出特征词语；

步骤4、评审专家特征词语提取：根据通用停用词库和专业停用词库进行停用词过滤，提取每个专家的特征词集合；

步骤5、构建科技项目、评审专家的分字段知识表示模型：通过对空间向量模型和物元知识集模型进行扩展，依据科技项目中的不同字段信息建立文本表示模型PRO=(id,F,WF,T,V)，其中id表示在项目库中的标识字段；F表示科技项目中字段类别集合；WF为字段的权重；T为特征词语；V表示字段所对应的词语及其权重集合即$V_i=\{v_{i1},f(v_{i1}),v_{i2},f(v_{i2}),\ldots,v_{in},f(v_{in})\}$，$v_{ij}$表示第i个字段中的第j个特征词语，$f(v_{ij})$表示$v_{ij}$关键词所对应的频数；科技项目信息的知识表示如下：

$$PRO=\begin{pmatrix} \text{科技项目682} & \text{项目名称} & 0.4 & \text{中间件} & 0.25 \\ & \text{主要研究内容} & 0.3 & \text{项目} & 0.21 \\ & \text{技术指标} & 0.3 & \text{云计算} & 0.12 \end{pmatrix};$$

同理，根据专家中的不同字段信息建立知识表示模型TM=(id,F,WF,T,V)；其中，id表示在专家库中的标识字段；F表示评审专家中字段类别集合；WF为字段的权重集合；T为特征词语；V表示字段所对应的特征词语及其权重集合即$V_i=\{v_{i1},f(v_{i1}),v_{i2},f(v_{i2}),\ldots,v_{in},f(v_{in})\}$，$v_{ij}$表示第i个字段中的第j个特征词语，$f(v_{ij})$表示$v_{ij}$特征词语在所对应的字段内的出现频率；评审专家信息的知识表示为：

$$TM=\begin{pmatrix} \text{评审专家2851} & \text{专家简历} & 0.1 & \text{中间件} & 0.5 \\ & \text{获奖情况} & 0.2 & \text{数字化} & 0.22 \\ & \text{发明专利情况} & 0.2 & \text{嵌入式} & 0.12 \\ & \text{承担项目及完成情况} & 0.1 & \text{平台} & 0.42 \\ & \text{发表论文情况} & 0.2 & \text{验收} & 0.42 \\ & \text{研究方向} & 0.2 & \text{组态} & 0.12 \end{pmatrix};$$

评审专家信息索引库构建:待评审专家知识表示模型构建完成后,将信息索引入库:首先从专家库中读取一个评审专家的内容项信息;基于分词结果建立词语语义网络并提取评审专家所包含的特征词;依据知识表示模型并利用Apache Lucene对其建立索引;将建立好的索引按所属类别加至对应的索引库中,直到所有的评审专家索引入库;

步骤6、根据项目的个数,推荐方式分为单一待审项目推荐专家和分组待审项目推荐专家;分组推荐专家对步骤5的待审项目知识表示模型做相应的字段间和项目间的特征合并操作,单一待审专家推荐只做相应的字段间特征合并操作;同时,对步骤5的评审专家的知识表示模型进行字段间特征合并;依据知识表示模型并利用Apache Lucene对合并后的特征信息建立索引;其中,科技项目索引构建在进行项目推荐时进行;

科技项目申报管理系统中待审项目往往是需要分组推荐的,上述特征合并操作,确保不会消除步骤5中知识表示模型设置不同字段权重对相似度计算产生推荐的贡献差异;

步骤7、经过步骤6的评审专家和科技项目的知识表示模型的字段间特征进行合并,假设评审专家信息向量若表示为$P=\{s_1,f(s_1),s_2,f(s_2),\ldots,s_n,f(s_n)\}$,科技项目信息向量表示为$Q=\{t_1,f(t_1),t_2,f(t_2),\ldots,t_n,f(t_n)\}$,基于最大匹配算法计算待审科技项目向量与评审专家的语义相似度;

步骤8、设置相似度截断,依据相似度的大小产生推荐指数,产生最终的推荐评审专家列表。

2.根据权利要求1所述的一种面向科技项目的评审专家智能推荐方法,其特征在于:步骤3中所述的语义相似度计算过程如下:

在知网语义词典中,如果对于两个词语W_1和W_2,W_1有n个概念:S11,S12,...,S1n,W_2有m个概念:S21,S22,...,S2m;词语W_1和W_2的相似度SimSEM(W1,W2)等于各个概念的相似度之最大值:

$$SimSEM(W1,W2)=\max_{i=1,\ldots n,j=1\ldots m} Sim(S_{1i},S_{2j});$$

实词和虚词具有不同的描述语言,需要计算其对应的句法义原或关系义原之间的相似度;实词概念包括第一基本义原、其他基本义原、关系义原描述、关系符号描述,相似度分别记为$Sim1(p_1,p_2)$、$Sim2(p_1,p_2)$、$Sim3(p_1,p_2)$、$Sim4(p_1,p_2)$;两个特征结构的相似度计算最终还原到基本义原或具体词的相似度计算;

$$Sim_4(S_1,S_2)=\sum_{i=1}^{4}\beta_i Sim_i(S_1,S_2);$$

$\beta_i(1\leqslant i\leqslant 4)$是可调节的参数,且有:$\beta_1+\beta_2+\beta_3+\beta_4=1$,$\beta_1\geqslant\beta_2\geqslant\beta_3\geqslant\beta_4$;

设CW={C1,C2,...,Cm}为处理后得到的词语集合,其对应的语义相似度邻接矩阵S_m定义为:

$$S_m=\begin{pmatrix} Sim(C_1,C_1) & \cdots & Sim(C_1,C_m) \\ \vdots & \ddots & \vdots \\ Sim(C_m,C_1) & \cdots & Sim(C_m,C_m) \end{pmatrix};$$

其中,$Sim(C_1,C_2)$为词C_1与词C_2的语义相似度,$Sim(C_i,C_i)$为1,$Sim(C_i,C_j)=Sim(C_j,C_i)$;

词语集合CW={C1,C2,...,Cm}经过词语语义相似度计算得到$m\times(1+m)/2$个词语间相似度的值;

所述的词的共现关系计算过程如下:

词共现模型是基于统计方法的自然语言处理研究领域的重要模型之一;根据词共现模型,若两个词经常共现在文档的同一窗口单元,这两个词在意义上是相互关联的,它们在一定程度上表达该文本的语义信息;利用滑动窗口对词语序列中的词语进行词语共现度计算:

首先,对词语序列进行词语提取,即去除空格,null以及合并相同的词,得到词语集合CW={C1,C2,...,Cm},其中m≤n;

词语集合CW对应的词语共现度矩阵Cm定义为:

$$C_m=\begin{pmatrix} Coo(C_1,C_1) & \cdots & Coo(C_1,C_m) \\ \vdots & \ddots & \vdots \\ Coo(C_m,C_1) & \cdots & Coo(C_m,C_m) \end{pmatrix};$$

Cm初始时,Coo(Ci,Cj)为01(1≤i,j≤m);

借助滑动窗口对词语序列进行词语共现度计算,滑动窗口中的词为$T_{i-1}T_iT_{i+1}$(1<i<n):

1)若i=n-1,转4);若T_{i-1}是空格或null,滑动窗口滑向下一个词,i++;否则,转2);

2)若T_i为中文,则Coo(T_{i-1},T_i)++,转3);若T_i为null,转3);否则转1);

3)若T_i是中文,则Coo(T_{i-1},T_{i+1})++,i++,转1);否则,转1);

4)若T_{n-2}是中文,转5);否则,转7)

5)若T_{n-1}是中文,Coo(T_{n-2},T_{n-1})++,转6);若T_{n-1}是空格,转6);否则结束;

6)若T_n是中文,Coo(T_{n-2},T_n)++,结束;否则结束;

7)若T_{n-1}是中文,且T_n也是中文,则Coo(T_{n-1},T_n)++,结束;否则结束;

经过上面步骤的计算,得到词语共现度矩阵Cm,并对Cm的每一个元素进行归一化处理,也就是每一个元素除以矩阵中所有元素的最大值,即max{Coo(C_i,C_j)|1≤i,j≤m};

所述的词语网络如下:

在构建带权词语网络时,首先要得到词语网络的权值矩阵,定义权值矩阵Wm为:

$$W_m=\alpha*C_m+\beta*S_m=\begin{pmatrix} W(C_1,C_1) & \cdots & W(C_1,C_m) \\ \vdots & \ddots & \vdots \\ W(C_m,C_1) & \cdots & W(C_m,C_m) \end{pmatrix};$$

其中,α为0.3,β为0.7,强化词语之间的语义关系,弱化词语之间的共现关系;

W_m作为输入的词语网络对应的邻接矩阵,则其对应的网络图定义为:G={V,E};其中图G为无向加权图,V表示图G中的顶点集,E表示G中的边集,v_i表示V中第i个顶点;

所述的词语聚集特征值的计算过程如下:

词语网络的重要特征有度分布、平均最短路径、聚集度与聚集系数;节点的度体现该节点与其它节点的关联情况;节点的聚集度和聚集系数体现在此节点局部范围内的节点相互连接密度;节点的度和聚集系数体现该节点在局部范围内的重要性;通过节点的加权度、聚集系数和节点介数来计算节点的聚集特征值,既能让重要的词语赋予较高的权值,又保证与许多重要的词语有关联的词也有较高的评分;

在词语语义相似度网络图中,无序偶对(v_i,v_j)表示节点v_i与v_j之间的边,则节点v_i的加

权度的定义为：

$$WD_i=\sum_{j=1}^{n}w_{ij}/n;$$

其中，w_{ij}为节点v_i与v_j间边上的权值，n为节点的总个数；

在词语语义相似度网络图中，无序偶对(v_i,v_j)表示节点v_i与v_j之间的边，节点v_i的非加权度D_i为$D_i=|\{(v_i,v_j):(v_i,v_j)\in E,v_i,v_j\in V\}|$；节点$v_i$的聚集度$K_i$为邻居节点间存在的实际边数：$T_i=|\{(v_j,v_k):(v_i,v_k)\in E,(v_j,v_k)\in E,v_i,v_j\in V\}|$，则节点$v_j$的聚集系数$C_i$的定义为：

$$C_i=\frac{T_i}{\binom{D_i}{2}}=2T_i/D_i(D_i-1);$$

在词语语义相似度网络图中，节点介数Betweenness是节点x和w间且最短路径通过节点v_i的可能性概率；两个非相邻节点间的联系度依赖于连接两点间的最短路径上的节点，这些节点潜在扮演控制节点间交互信息流的角色，B_i体现节点v_i在局部环境下的互连接度，则节点介数Betweenness的定义为：

$$B_i=\sum_{w\in G,x\in G}\frac{r_{vi}(w,x)}{d(w,x)};$$

d(w,x)表示带权词语语义相似度网络图中任意两节点w和x间最短路径数目，$r_{v_i}(w,x)$表示任意两节点w和x且经过v_i的最短路径数目$v_i\in G$；

将节点v_i的平均加权度、聚集系数和介数Betweenness进行加权综合衡量节点的聚集特征值，节点v_i的聚集特征值Z_i的定义为：

$$Z_i=a\times WD_i+b\times C_i/\sum_{j=1}^{n}c_j+c\times B_i;$$

其中，a+b+c＝1；

所述的词语的统计特征值的计算过程如下：

采用非线性函数对词频进行归一化处理；词语W_i在文本中的词频权重TFi定义为：

$$TF_i=\frac{f(W_i)}{\sum_{j=1}^{n}f(p_j)};$$

其中，TFi表示词语W_i的词频权重，p_j表示文本中的某个词语，f为词频统计函数；

词语W_i在文本中的词性权重posi定义为：

$$pos_i=\begin{cases}1,\ \text{if词为名词}\\0.6,\ \text{if词为动词或形容词}\\0.2,\ \text{otherwise}\end{cases};$$

词越长越能反映具体的信息，反之，较短的词的所表示意义通常较抽象；尤其在文档中的特征词语多是一些专业学术组合词汇，长度较长，其含义更明确，更能反映文本主题；增加长词的权重，有利于对词汇进行分割，从而更准确地反映出词在文档中的重要程度；

词语W_i在文本中的词长权重leni定义为：

$$len_i=\begin{cases}1-1/L,\ \text{if词的长度L大于2}\\ 0,\ otherwise\end{cases};$$

对于词语序列中的每个词,其统计特征值为

$stats_i=A*TF_i+B*pos_i+C*len_i$;

其中,A+B+C=1;

所述的词语W_i关键度的计算过程如下:

对应于加权词语网络中的每个节点,它的关键度值Imp_i定义为:

$Imp_i=\beta*stats_i+(1-\beta)*Z_i$;

其中,$0<\beta<1$;

通过计算将得到关键度的值,从大到小排序,设定一个阈值γ,$0<\gamma<1$,取出前q个的值,则这些词语将作为科技项目的特征词语,这些词语充分反映主题,而且是比较重要的词语。

3.根据权利要求1所述的一种面向科技项目的评审专家智能推荐方法,其特征在于:步骤6中所述的特征合并通过逻辑异或操作进行过程如下:

(1)一个待审项目、一个评审专家的字段间特征合并

假设字段特征词集合W'_1和W'_2合并,则定义W'_1和W'_2合并规则$W'_1\oplus W'_2$为:

$$W'_1\oplus W'_2=\{\forall i,j,\{word_{1i},\frac{f(word_{1i})+f(word_{2i})}{2}\}\mid word_{1i}=word_{2j}\};$$

其中,$word_{1i}$,$word_{2j}$为特征词;

加入字段权重改进并扩展上述定义,对评审专家、科技项目的字段间特征进行合并,合并规则为:

$$W'_1\oplus W'_2=\{\forall i,j,\{word_{1i},\frac{w1*f(word_{1i})+w2*f(word_{2i})}{\sqrt{w1^2+w2^2}}\}\mid word_{1i}=word_{2j}\};$$

(2)分组待审项目的项目间特征合并

这一合并过程操作只针对待审科技项目的特征向量,不针对评审专家特征向量,专家特征向量只需要做字段间特征合并操作;若$V(d_1)$和$V(d_2)$分别是两个科技项目经过字段间特征合并后的向量模型,对任意$t_{1j}\in V(d_1)$,$t_{2j}\in V(d_2)$,若存在t_{1j}与t_{2j}相同则合并;$V(d_1)\oplus V(d_2)$定义为:

$$V(d_1)\oplus V(d_2)=\{<t_k,w_k(p)=\frac{w_i(d_1)+w_j(d_2)}{2}>\};$$

其中,k=1,…,n,t_k为特征词条项,$w_k(p)$为t_k的权重;

知识表示模型产生的基本过程如下:

a).合并科技项目字段间特征,得到每个项目的向量模型V(d);

b).将所有科技项目向量模型集合采用合并策略$V(p)=V(d_1)\oplus V(d_2)\ldots\oplus V(d_n)$;通过上述的方法,对科技项目组建立基于向量空间的知识表示模型;

$V(p)=\{<t_1,w_1(p)>,<t_2,w_2(p)>,\ldots,<t_n,w_n(p)>\}$;

其中,k=1,…,n,t_k为项目组特征词词条项,$w_k(p)$为t_k的权重。

一种面向科技项目的评审专家智能推荐方法

技术领域

[0001] 本发明属于专家推荐技术领域，尤其涉及一种基于网络服务的科技项目评审专家智能推荐方法，它是一种辅助科技项目立项决策的智能方法。

背景技术

[0002] 随着科技项目管理系统在我国各职能部门迅速普及，科技项目的评审工作从以往的集中会议模式发展到当前的网络模式，打破了评审工作中专家地域的限制。评审专家根据领域知识和资助机构的资助标准，对项目申请书进行评议，资助机构依据专家的评议情况决定是否资助。

[0003] 目前面向科技项目的专家推荐大多仅凭项目管理人员的主观意识推荐专家对待审项目进行评审，一个待审项目往往需要多个专家进行评审，人工推荐专家势必存在效率不高、工作量大、缺乏科学性等问题，所遴选出的专家并非是最合适的。因此，对科技项目评审专家智能推荐的研究是非常关键的，可以有效地缓解专家与所评项目内容不匹配等问题，大大提升科技项目评审工作的社会服务能力。

[0004] 现今智能推荐技术，如协同过滤推荐、基于内容的推荐等，大多应用在影视推荐网站、商品推荐网站，鲜有在科技项目评审专家信息库中的研究与应用，由于特定领域的限制，为科技项目智能推荐专家技术与一般的推荐技术还是有区别的：首先，科技项目管理系统的推荐涉及各行各业，领域知识非常复杂；其次，科技项目评审专家的推荐涉及到科技项目的资助基金，对专家推荐的客观性、公正性和精准性的要求是非常高的。目前在这方面，我国还缺乏系统化的方法指导和成熟的技术支持。而信息文本具有“半结构化”等特征，专家信息和待审科技项目信息的内容是可以进行匹配的，本发明充分利用结构特征以及词语语义信息计算项目与专家的信息相似度。若相似度较高，则表示专家对该项目熟悉，产生推荐专家列表对项目进行评审。本发明同时提供一种为科技项目推荐评审专家的决策支持系统(Decision Support System,DSS)，将评审专家分配到领域知识相匹配的项目进行科学评审，使得辅助专家(决策用户)实现科学的决策，帮助决策用户提高决策水平和质量，使评审更具科学性和客观性。

发明内容

[0005] 本发明针对现有技术的不足，提供了一种面向科技项目的评审专家智能推荐方法。

[0006] 本发明面向科技项目的评审专家推荐过程包括如下步骤：

[0007] 步骤1.把科技项目和专家信息中的通用词和惯用词作为专业停用词库；把标点符号、非汉字作为切分标记库。

[0008] 步骤2.对科技项目信息、专家信息进行分词：根据科技项目信息中切分标记，将项目名称、主要研究内容、技术指标等信息切分成子串序列；根据评审专家信息中切分标记，抽取专家信息、获奖情况、发明情况、发表论文情况、课题承担过的项目及完成情况、研究方

向等信息切分成子串序列，一个子串序列即一个字段信息；利用中科院ICTCLAS对子串序列进行分词。

[0009] 步骤3.科技项目特征词语提取：利用通用停用词库和专业停用词库对分词进行停用词过滤，通用停用词库采用哈工大停用词表，把去除停用词的分词结果作为一个词语集合。

[0010] 专业停用词库的构建是一个自学习不断完善的过程，在信息分词过程中不断统计词语的词频，词语在文本出现的概率大于一定阈值，将它纳入到停用词库。

[0011] 科技项目信息量较大，对词语集合进行词语间语义相似度计算，根据词的语义关系和词的共现关系构建词语网络，计算网络中的词语聚集特征值；然后结合词语的统计特征值，计算词语的关键度来提取出科技项目特征词语；科技项目的特征词语就是提取综合文本的统计特征信息和语义特征信息，更加准确地提取出特征词语。

[0012] 所述的语义相似度计算过程如下：

[0013] 在知网语义词典中，如果对于两个词语W_1和W_2，W_1有n个概念：S11，S12，...，S1n，W_2有m个概念：S21，S22，...，S2m。词语W_1和W_2的相似度SimSEM(W1，W2)等于各个概念的相似度之最大值：

[0014] $$SimSEM(W1,W2)=\max_{i=1,\ldots n,j=1\ldots m} Sim(S_{1i},S_{2j})$$

[0015] 实词和虚词具有不同的描述语言，需要计算其对应的句法义原或关系义原之间的相似度。实词概念包括第一基本义原、其他基本义原、关系义原描述、关系符号描述，相似度分别记为$Sim1(p_1,p_2)$、$Sim2(p_1,p_2)$、$Sim3(p_1,p_2)$、$Sim4(p_1,p_2)$。两个特征结构的相似度计算最终还原到基本义原或具体词的相似度计算。

[0016] $$Sim_4(S_1,S_2)=\sum_{i=1}^{4}\beta_i Sim_i(S_1,S_2)$$

[0017] $\beta_i(1\leqslant i\leqslant 4)$是可调节的参数，且有：$\beta_1+\beta_2+\beta_3+\beta_4=1$，$\beta_1\geqslant\beta_2\geqslant\beta_3\geqslant\beta_4$。

[0018] 设CW＝{C1，C2，...，Cm}为处理后得到的词语集合，其对应的语义相似度邻接矩阵S_m定义为：

[0019] $$S_m=\begin{pmatrix} Sim(C_1,C_1) & \cdots & Sim(C_1,C_m) \\ \vdots & \ddots & \vdots \\ Sim(C_m,C_1) & \cdots & Sim(C_m,C_m) \end{pmatrix}$$

[0020] 其中，$Sim(C_1,C_2)$为词C_1与词C_2的语义相似度，$Sim(C_i,C_i)$为1，$Sim(C_i,C_j)=Sim(C_j,C_i)$。

[0021] 词语集合CW＝{C1，C2，...，Cm}经过词语语义相似度计算得到m×(1+m)/2个词语间相似度的值。

[0022] 所述的词的共现关系计算过程如下：

[0023] 词共现模型是基于统计方法的自然语言处理研究领域的重要模型之一。根据词共现模型，若两个词经常共现在文档的同一窗口单元(如一句话、一个自然段等)，这两个词在意义上是相互关联的，它们在一定程度上表达该文本的语义信息。利用滑动窗口(滑动窗口长度为3)对词语序列中的词语进行词语共现度计算，滑动窗口如图1所示：

[0024] 首先，对词语序列进行词语提取，即去除空格，null以及合并相同的词，得到词语

集合CW＝{C1,C2,...,Cm},其中m≤n。

[0025] 词语集合CW对应的词语共现度矩阵Cm定义为：

[0026] $$C_m=\begin{pmatrix} Coo(C1,C1) & \cdots & Coo(C1,Cm) \\ \vdots & \ddots & \vdots \\ Coo(Cm,C1) & \cdots & Coo(Cm,Cm) \end{pmatrix}$$

[0027] Cm初始时,Coo(Ci,Cj)为01(1≤i,j≤m)。

[0028] 借助滑动窗口对词语序列进行词语共现度计算,滑动窗口中的词为$T_{i-1}T_iT_{i+1}$(1<i<n)：

[0029] 1)若i＝n-1,转4);若T_{i-1}是空格或null,滑动窗口滑向下一个词,i++;否则,转2)。

[0030] 2)若T_i为中文,则Coo(T_{i-1},T_i)++,转3);若T_i为null,转3);否则转1)。

[0031] 3)若T_i是中文,则Coo(T_{i-1},T_{i+1})++,i++,转1);否则,转1)。

[0032] 4)若T_{n-2}是中文,转5);否则,转7)

[0033] 5)若T_{n-1}是中文,Coo(T_{n-2},T_{n-1})++,转6);若T_{n-1}是空格,转6);否则结束。

[0034] 6)若T_n是中文,Coo(T_{n-2},T_n)++,结束;否则结束。

[0035] 7)若T_{n-1}是中文,且T_n也是中文,则Coo(T_{n-1},T_n)++,结束;否则结束。

[0036] 经过上面步骤的计算,得到词语共现度矩阵Cm,并对Cm的每一个元素进行归一化处理,也就是每一个元素除以矩阵中所有元素的最大值,即max{Coo(C_i,C_j)|1≤i,j≤m}。

[0037] 所述的词语网络如下：

[0038] 在构建带权词语网络时,首先要得到词语网络的权值矩阵,定义权值矩阵Wm为：

[0039] $$W_m=\alpha*C_m+\beta*S_m=\begin{pmatrix} W(C1,C1) & \cdots & W(C1,Cm) \\ \vdots & \ddots & \vdots \\ W(Cm,C1) & \cdots & W(Cm,Cm) \end{pmatrix}$$

[0040] 其中,α为0.3,β为0.7,强化词语之间的语义关系,弱化词语之间的共现关系。

[0041] W_m作为输入的词语网络对应的邻接矩阵,则其对应的网络图定义为:G＝{V,E};其中图G为无向加权图,V表示图G中的顶点集,E表示G中的边集,v_i表示V中第i个顶点(词)。

[0042] 所述的词语聚集特征值的计算过程如下：

[0043] 词语网络的重要特征有度分布、平均最短路径、聚集度与聚集系数。节点的度体现该节点与其它节点的关联情况。节点的聚集度和聚集系数体现在此节点局部范围内的节点相互连接密度。节点的度和聚集系数体现该节点在局部范围内的重要性。本发明通过节点的加权度、聚集系数和节点介数来计算节点的聚集特征值,既能让重要的词语赋予较高的权值,又保证与许多重要的词语有关联的词也有较高的评分。

[0044] 在词语语义相似度网络图中,无序偶对(v_i,v_j)表示节点v_i与v_j之间的边,则节点v_i的加权度的定义为：

[0045] $$WD_i=\sum_{j=1}^{n} w_{ij}/n$$

[0046] 其中,w_{ij}为节点v_i与v_j间边上的权值,n为节点的总个数。

[0047] 在词语语义相似度网络图中，无序偶对(v_i,v_j)表示节点v_i与v_j之间的边，节点v_i的非加权度D_i为$D_i=|\{(v_i,v_j):(v_i,v_j)\in E,v_i,v_j\in V\}|$；节点$v_i$的聚集度$K_i$为邻居节点间存在的实际边数：$T_i=|\{(v_j,v_k):(v_i,v_k)\in E,(v_j,v_k)\in E,v_i,v_j\in V\}|$，则节点$v_j$的聚集系数$C_i$的定义为：

[0048] $$C_i=\frac{T_i}{\binom{D_i}{2}}=2T_i/D_i(D_i-1)$$

[0049] 在词语语义相似度网络图中，节点介数Betweenness是节点x和w间且最短路径通过节点v_i的可能性概率。两个非相邻节点间的联系度依赖于连接两点间的最短路径上的节点，这些节点潜在扮演控制节点间交互信息流的角色，B_i体现节点v_i在局部环境下的互连接度，则节点介数Betweenness的定义为：

[0050] $$B_i=\sum_{w\in G,x\in G}\frac{r_{vi}(w,x)}{d(w,x)}$$

[0051] d(w,x)表示带权词语语义相似度网络图中任意两节点w和x间最短路径数目，$r_{v_i}(w,x)$表示任意两节点w和x且经过$v_i(v_i\in G)$的最短路径数目。

[0052] 将节点v_i的平均加权度、聚集系数和介数Betweenness进行加权综合衡量节点的聚集特征值，节点v_i的聚集特征值Z_i的定义为：

[0053] $$Z_i=a\times WD_i+b\times C_i/\sum_{j=1}^{n}c_j+c\times B_i$$

[0054] 其中，a+b+c=1。

[0055] 所述的词语的统计特征值的计算过程如下：

[0056] 采用非线性函数对词频进行归一化处理。词语W_i在文本中的词频权重TFi定义为：

[0057] $$TF_i=\frac{f(W_i)}{\sum_{j=1}^{n}f(p_j)}$$

[0058] 其中，TFi表示词语W_i的词频权重，p_j表示文本中的某个词语，f为词频统计函数。

[0059] 中文文本中能标识文本特性的一般是实词，如名词、动词、形容词等。而感叹词、介词、连词等虚词对确定文本类别基本没有意义，会对特征词语提取带来很大干扰。词语W_i在文本中的词性权重posi定义为：

[0060] $$pos_i=\begin{cases}1,\ \text{if词为名词}\\0.6,\ \text{if词为动词或形容词}\\0.2,\ \text{otherwise}\end{cases}$$

[0061] 词越长越能反映具体的信息，反之，较短的词的所表示意义通常较抽象。尤其在文档中的特征词语多是一些专业学术组合词汇，长度较长，其含义更明确，更能反映文本主题。增加长词的权重，有利于对词汇进行分割，从而更准确地反映出词在文档中的重要程度。

[0062] 词语W_i在文本中的词长权重leni定义为：

[0063] $len_i = \begin{cases} 1-1/L, & \text{if词的长度L大于2} \\ 0, & otherwise \end{cases}$

[0064] 对于词语序列中的每个词，其统计特征值为

[0065] $stats_i = A*TF_i + B*pos_i + C*len_i$

[0066] 其中，A+B+C=1。

[0067] 所述的词语W_i关键度的计算过程如下：

[0068] 对应于加权词语网络中的每个节点，它的关键度值Imp_i定义为：

[0069] $Imp_i = \beta*stats_i + (1-\beta)*Z_i$

[0070] 其中，$0<\beta<1$。

[0071] 通过计算将得到关键度的值，从大到小排序，设定一个阈值γ（$0<\gamma<1$），取出前q个的值，则这些词语将作为科技项目的特征词语，这些词语充分反映主题，而且是比较重要的词语。

[0072] 步骤4.评审专家特征词语提取：评审专家信息量较科技项目信息少，科技项目的特征词构建网络并基于统计特征和语义特征的提取技术，不适合评审专家信息的特征词语提取，直接根据通用停用词库和专业停用词库进行停用词过滤，提取每个专家的特征词集合，通用停用词库是也是采用哈工大停用词表，专业停用词库需要人员进行不断地维护。

[0073] 步骤5.构建科技项目、评审专家的分字段知识表示模型：通过对空间向量模型和物元知识集模型进行扩展，依据科技项目中的不同字段信息建立文本表示模型PRO=(id,F,WF,T,V)，其中id表示在项目库中的标识字段；F表示科技项目中字段类别集合；WF为字段的权重；T为特征词语；V表示字段所对应的词语及其权重集合即$V_i=\{v_{i1},f(v_{i1}),v_{i2},f(v_{i2}),\ldots,v_{in},f(v_{in})\}$，$v_{ij}$表示第i个字段中的第j个特征词语，$f(v_{ij})$表示$v_{ij}$关键词所对应的频数。科技项目信息的知识表示如下：

[0074] $$PRO=\begin{pmatrix} 科技项目682 & 项目名称 & 0.4 & 中间件 & 0.25 \\ & 主要研究内容 & 0.3 & 项目 & 0.21 \\ & 技术指标 & 0.3 & 云计算 & 0.12 \end{pmatrix}$$

[0075] 同理，根据专家中的不同字段信息建立知识表示模型TM=(id,F,WF,T,V)。其中，id表示在专家库中的标识字段；F表示评审专家中字段类别集合；WF为字段的权重集合；T为特征词语；V表示字段所对应的特征词语及其权重集合即$V_i=\{v_{i1},f(v_{i1}),v_{i2},f(v_{i2}),\ldots,v_{in},f(v_{in})\}$，$v_{ij}$表示第i个字段中的第j个特征词语，$f(v_{ij})$表示$v_{ij}$特征词语在所对应的字段内的出现频率。评审专家信息的知识表示为：

[0076] $$TM=\begin{pmatrix} 评审专家2851 & 专家简历 & 0.1 & 中间件 & 0.5 \\ & 获奖情况 & 0.2 & 数字化 & 0.22 \\ & 发明专利情况 & 0.2 & 嵌入式 & 0.12 \\ & 承担项目及完成情况 & 0.1 & 平台 & 0.42 \\ & 发表论文情况 & 0.2 & 验收 & 0.42 \\ & 研究方向 & 0.2 & 组态 & 0.12 \end{pmatrix}$$

[0077] 评审专家信息索引库构建：待评审专家知识表示模型构建完成后，将信息索引入库：首先从专家库中读取一个评审专家的内容项信息；基于分词结果建立词语语义网络并

提取评审专家所包含的特征词;依据知识表示模型并利用Apache Lucene对其建立索引;将建立好的索引按所属类别加至对应的索引库中,直到所有的评审专家索引入库。

[0078] 步骤6:根据项目的个数,推荐方式分为单一待审项目推荐专家和分组(多个)待审项目推荐专家。分组推荐专家对步骤5的待审项目知识表示模型做相应的字段间和项目间的特征合并操作,单一待审专家推荐只做相应的字段间特征合并操作。同时,对步骤5的评审专家的知识表示模型进行字段间特征合并。依据知识表示模型并利用Apache Lucene对合并后的特征信息建立索引。其中,科技项目索引构建在进行项目推荐时进行。

[0079] 科技项目申报管理系统中待审项目往往是需要分组推荐的,上述特征合并操作,确保不会消除步骤5中知识表示模型设置不同字段权重对相似度计算产生推荐的贡献差异。

[0080] 所述的待审项目、评审专家的特征合并通过逻辑异或操作进行过程如下:

[0081] (1)一个待审项目、一个评审专家的字段间特征合并

[0082] 假设字段特征词集合W'1和W'2合并,则定义W'1和W'2合并规则$W'_1 \oplus W'_2$为:

[0083] $$W'_1 \oplus W'_2 = \{\forall i,j,\{word_{1i}, \frac{f(word_{1i}) + f(word_{2i})}{2}\} \mid word_{1i} = word_{2j}\}$$

[0084] 其中,$word_{1i}$,$word_{2j}$为特征词。

[0085] 加入字段权重改进并扩展上述定义,对评审专家、科技项目的字段间特征进行合并,合并规则为:

[0086] $$W'_1 \oplus W'_2 = \{\forall i,j,\{word_{1i}, \frac{w1 * f(word_{1i}) + w2 * f(word_{2i})}{\sqrt{w1^2 + w2^2}}\} \mid word_{1i} = word_{2j}\}$$

[0087] (2)分组待审项目的项目间特征合并

[0088] 这一合并过程操作只针对待审科技项目的特征向量,不针对评审专家特征向量,专家特征向量只需要做字段间特征合并操作。若$V(d_1)$和$V(d_2)$分别是两个科技项目经过字段间特征合并后的向量模型,对任意$t_{1j} \in V(d_1)$,$t_{2j} \in V(d_2)$,若存在t_{1j}与t_{2j}相同则合并。$V(d_1) \oplus V(d_2)$定义为:

[0089] $$V(d_1) \oplus V(d_2) = \{< t_k, w_k(p) = \frac{w_i(d_1) + w_j(d_2)}{2} >\}$$

[0090] 其中,k=1,…,n,t_k为特征词条项,$w_k(p)$为t_k的权重。

[0091] 科技项目组的知识表示模型产生的基本过程如下:

[0092] a).合并科技项目字段间特征,得到每个项目的向量模型V(d);

[0093] b).将所有科技项目向量模型集合采用合并策略$V(p) = V(d_1) \oplus V(d_2) \ldots \oplus V(d_n)$。通过上述的方法,对科技项目组建立基于向量空间的知识表示模型。

[0094] $V(p) = \{< t_1, w_1(p) >, < t_2, w_2(p) >, \ldots, < t_n, w_n(p) >\}$

[0095] 其中,k=1,…,n,t_k为项目组特征词词条项,$w_k(p)$为t_k的权重。

[0096] 步骤7.经过步骤6的评审专家和科技项目的知识表示模型的字段间特征进行合并,假设评审专家信息向量若表示为$P = \{s_1, f(s_1), s_2, f(s_2), \ldots, s_n, f(s_n)\}$,科技项目信息(组)向量表示为$Q = \{t_1, f(t_1), t_2, f(t_2), \ldots, t_n, f(t_n)\}$,基于最大匹配算法计算待审科技项目(组)向量与评审专家的语义相似度。

[0097] 步骤8.设置相似度截断,依据相似度的大小产生推荐指数,产生最终的推荐评审

专家列表。

[0098] 本发明有益效果如下：

[0099] 能够更加便捷地、智能地、精准地推荐出科技项目评审专家；能够大大减轻科技项目申报管理系统科技工作者对评审专家的分配任务，减少管理的成本费用；能够保证评审专家与待审科技项目具有较高的领域匹配度，保证评审专家对项目的评审做到客观性、公正性和科学性，提供自动的、高效的、公正的决策支持，避免科技项目审批出现人情关系网、“马太效应”等审批不端的问题。

附图说明

[0100] 图1是本发明中进行词语共现度计算滑动窗口。

[0101] 图2是本发明中基于二部图的最大匹配算法原理示意图。

[0102] 图3是本发明中面向科技项目的评审专家智能推荐方法流程图。

[0103] 图4是本发明中科技项目和评审专家信息的特征词的提取流程图。

[0104] 图5是本发明中评审专家知识索引库构建流程图。

具体实施方式

[0105] 下面结合附图对本发明作进一步说明，应该强调的是下述说明仅仅是示例性的，而不是为了限制本发明的范围及其应用。以下对本发明的具体实施方式作进一步详述，基于发明中的实施例，本领域普通技术人员在没有创造性劳动前提下所获得的所有其他实施例，都属于本发明的保护范围。

[0106] 如图3所示，本发明的推荐方法的主要思路是：(1)针对科技项目申报管理系统中的专家信息和待审科技项目信息，将主要文本切分成子串序列并进行中科院ICTCLAS分词，对分词结果进行停用词过滤得到词语集合；(2)科技项目信息包括主要研究内容、技术指标等信息，信息量较大，发明根据词的语义关系和词的共现关系构建词语网络，并计算词语网络的节点聚集特征值，与统计特征值加权计算词语关键度，提取每个科技项目的特征词；(3)专家信息比科技项目信息精简，信息量较少，直接将每个专家信息经过滤得到的词语集合作为特征词；(4)根据科技项目、专家字段信息的重要性不同设置字段权重，依据(2)和(3)得到的特征词，分别构建针对项目和专家的知识表示模型，构建专家索引库；(5)分组推荐专家模型待审项目知识表示模型做字段间和项目间的特征合并操作，单一待审项目专家推荐只做字段间特征合并操作。同时对专家知识表示模型做字段间特征合并。(6)综合考虑词语具有语义模糊匹配的特征，计算专家信息与待审科技项目信息的相似度，通过设定阈值截断产生最终推荐专家列表。

[0107] 步骤1.把科技项目和专家信息中的通用词和惯用词作为专业停用词库；把标点符号、非汉字作为切分标记库。

[0108] 步骤2.对科技项目信息、专家信息进行分词：根据科技项目信息中切分标记，将项目名称、主要研究内容、技术指标等信息切分成子串序列；根据评审专家信息中切分标记，抽取专家信息、获奖情况、发明情况、发表论文情况、课题承担过的项目及完成情况、研究方向等信息切分成子串序列，一个子串序列即一个字段信息；利用中科院ICTCLAS对子串序列进行分词。

[0109] 步骤3.科技项目特征词语提取:利用通用停用词库和专业停用词库对分词进行停用词过滤,通用停用词库采用哈工大停用词表,把去除停用词的分词结果作为一个词语集合,参见图4。

[0110] 专业停用词库的构建是一个自学习不断完善的过程,在信息分词过程中不断统计词语的词频,词语在文本出现的概率大于一定阈值,将它纳入到停用词库。

[0111] 科技项目信息量较大,对词语集合进行词语间语义相似度计算,根据词的语义关系和词的共现关系构建词语网络,计算网络中的词语聚集特征值;然后结合词语的统计特征值,计算词语的关键度来提取出科技项目特征词语;科技项目的特征词语就是提取综合文本的统计特征信息和语义特征信息,更加准确地提取出特征词语。

[0112] 所述的语义相似度计算过程如下:

[0113] 在知网语义词典中,如果对于两个词语W_1和W_2,W_1有n个概念:S11,S12,...,S1n,W_2有m个概念:S21,S22,...,S2m。词语W_1和W_2的相似度SimSEM(W1,W2)等于各个概念的相似度之最大值:

[0114] $$SimSEM(W1,W2)=\max_{i=1,\ldots n;j=1\ldots m} Sim(S_{1i},S_{2j})$$

[0115] 实词和虚词具有不同的描述语言,需要计算其对应的句法义原或关系义原之间的相似度。实词概念包括第一基本义原、其他基本义原、关系义原描述、关系符号描述,相似度分别记为Sim1(p_1,p_2)、Sim2(p_1,p_2)、Sim3(p_1,p_2)、Sim4(p_1,p_2)。两个特征结构的相似度计算最终还原到基本义原或具体词的相似度计算。

[0116] $$Sim_4(S_1,S_2)=\sum_{i=1}^{4}\beta_i Sim_i(S_1,S_2)$$

[0117] β_i(1≤i≤4)是可调节的参数,且有:$\beta_1+\beta_2+\beta_3+\beta_4=1$,$\beta_1\geq\beta_2\geq\beta_3\geq\beta_4$。

[0118] 设CW={C1,C2,...,Cm}为处理后得到的词语集合,其对应的语义相似度邻接矩阵S_m定义为:

[0119] $$S_m=\begin{pmatrix} Sim(C_1,C_1) & \cdots & Sim(C_1,C_m) \\ \vdots & \ddots & \vdots \\ Sim(C_m,C_1) & \cdots & Sim(C_m,C_m) \end{pmatrix}$$

[0120] 其中,Sim(C_1,C_2)为词C_1与词C_2的语义相似度,Sim(C_i,C_i)为1,Sim(C_i,C_j)=Sim(C_j,C_i)。

[0121] 词语集合CW={C1,C2,...,Cm}经过词语语义相似度计算得到m×(1+m)/2个词语间相似度的值。

[0122] 所述的词的共现关系计算过程如下:

[0123] 词共现模型是基于统计方法的自然语言处理研究领域的重要模型之一。根据词共现模型,若两个词经常共现在文档的同一窗口单元(如一句话、一个自然段等),这两个词在意义上是相互关联的,它们在一定程度上表达该文本的语义信息。利用滑动窗口(滑动窗口长度为3)对词语序列中的词语进行词语共现度计算,滑动窗口如图1所示:

[0124] 首先,对词语序列进行词语提取,即去除空格,null以及合并相同的词,得到词语集合CW={C1,C2,...,Cm},其中m≤n。

[0125] 词语集合CW对应的词语共现度矩阵Cm定义为:

[0126] $$C_m=\begin{pmatrix} Coo(C1,C1) & \cdots & Coo(C1,Cm) \\ \vdots & \ddots & \vdots \\ Coo(Cm,C1) & \cdots & Coo(Cm,Cm) \end{pmatrix}$$

[0127] Cm初始时，Coo(Ci,Cj)为01(1≤i,j≤m)。

[0128] 借助滑动窗口对词语序列进行词语共现度计算，滑动窗口中的词为$T_{i-1}T_iT_{i+1}$(1<i<n)：

[0129] 1)若i=n-1，转4)；若T_{i-1}是空格或null，滑动窗口滑向下一个词，i++；否则，转2)。

[0130] 2)若T_i为中文，则Coo(T_{i-1},T_i)++，转3)；若T_i为null，转3)；否则转1)。

[0131] 3)若T_i是中文，则Coo(T_{i-1},T_{i+1})++，i++，转1)；否则，转1)。

[0132] 4)若T_{n-2}是中文，转5)；否则，转7)

[0133] 5)若T_{n-1}是中文，Coo(T_{n-2},T_{n-1})++，转6)；若T_{n-1}是空格，转6)；否则结束。

[0134] 6)若T_n是中文，Coo(T_{n-2},T_n)++，结束；否则结束。

[0135] 7)若T_{n-1}是中文，且T_n也是中文，则Coo(T_{n-1},T_n)++，结束；否则结束。

[0136] 经过上面步骤的计算，得到词语共现度矩阵Cm，并对Cm的每一个元素进行归一化处理，也就是每一个元素除以矩阵中所有元素的最大值，即max{Coo(C_i,C_j)|1≤i,j≤m}。

[0137] 所述的词语网络如下：

[0138] 在构建带权词语网络时，首先要得到词语网络的权值矩阵，定义权值矩阵Wm为：

[0139] $$W_m=\alpha*C_m+\beta*S_m=\begin{pmatrix} W(C1,C1) & \cdots & W(C1,Cm) \\ \vdots & \ddots & \vdots \\ W(Cm,C1) & \cdots & W(Cm,Cm) \end{pmatrix}$$

[0140] 其中，α为0.3，β为0.7，强化词语之间的语义关系，弱化词语之间的共现关系。

[0141] W_m作为输入的词语网络对应的邻接矩阵，则其对应的网络图定义为：G={V,E}；其中图G为无向加权图，V表示图G中的顶点集，E表示G中的边集，v_i表示V中第i个顶点(词)。

[0142] 所述的词语聚集特征值的计算过程如下：

[0143] 词语网络的重要特征有度分布、平均最短路径、聚集度与聚集系数。节点的度体现该节点与其它节点的关联情况。节点的聚集度和聚集系数体现在此节点局部范围内的节点相互连接密度。节点的度和聚集系数体现该节点在局部范围内的重要性。本发明通过节点的加权度、聚集系数和节点介数来计算节点的聚集特征值，既能让重要的词语赋予较高的权值，又保证与许多重要的词语有关联的词也有较高的评分。

[0144] 在词语语义相似度网络图中，无序偶对(v_i,v_j)表示节点v_i与v_j之间的边，则节点v_i的加权度的定义为：

[0145] $$WD_i=\sum_{j=1}^{n} w_{ij}/n$$

[0146] 其中，w_{ij}为节点v_i与v_j间边上的权值，n为节点的总个数。

[0147] 在词语语义相似度网络图中，无序偶对(v_i,v_j)表示节点v_i与v_j之间的边，节点v_i的非加权度D_i为D_i=|{(v_i,v_j):(v_i,v_j)∈E,v_i,v_j∈V}|；节点v_i的聚集度K_i为邻居节点间存在

的实际边数：$T_i=|\{(v_j,v_k):(v_i,v_k)\in E,(v_j,v_k)\in E,v_i,v_j\in V\}|$，则节点$v_j$的聚集系数$C_i$的定义为：

[0148] $$C_i=\frac{T_i}{\binom{D_i}{2}}=2T_i/D_i(D_i-1)$$

[0149] 在词语语义相似度网络图中，节点介数Betweenness是节点x和w间且最短路径通过节点v_i的可能性概率。两个非相邻节点间的联系度依赖于连接两点间的最短路径上的节点，这些节点潜在扮演控制节点间交互信息流的角色，B_i体现节点v_i在局部环境下的互连接度，则节点介数Betweenness的定义为：

[0150] $$B_i=\sum_{w\in G,x\in G}\frac{r_{vi}(w,x)}{d(w,x)}$$

[0151] d(w,x)表示带权词语语义相似度网络图中任意两节点w和x间最短路径数目，$r_{v_i}(w,x)$表示任意两节点w和x且经过v_i($v_i\in G$)的最短路径数目。

[0152] 将节点v_i的平均加权度、聚集系数和介数Betweenness进行加权综合衡量节点的聚集特征值，节点vi的聚集特征值Z_i的定义为：

[0153] $$Z_i=a\times WD_i+b\times C_i/\sum_{j=1}^{n}c_j+c\times B_i$$

[0154] 其中，a+b+c＝1。

[0155] 所述的词语的统计特征值的计算过程如下：

[0156] 采用非线性函数对词频进行归一化处理。词语W_i在文本中的词频权重TFi定义为：

[0157] $$TFi=\frac{f(Wi)}{\sum_{j=1}^{n}f(p_j)}$$

[0158] 其中，TFi表示词语W_i的词频权重，p_j表示文本中的某个词语，f为词频统计函数。

[0159] 中文文本中能标识文本特性的一般是实词，如名词、动词、形容词等。而感叹词、介词、连词等虚词对确定文本类别基本没有意义，会对特征词语提取带来很大干扰。词语W_i在文本中的词性权重posi定义为：

[0160] $$pos_i=\begin{cases}1,\ \text{if词为名词}\\0.6,\ \text{if词为动词或形容词}\\0.2,\ \text{otherwise}\end{cases}$$

[0161] 词越长越能反映具体的信息，反之，较短的词的所表示意义通常较抽象。尤其在文档中的特征词语多是一些专业学术组合词汇，长度较长，其含义更明确，更能反映文本主题。增加长词的权重，有利于对词汇进行分割，从而更准确地反映出词在文档中的重要程度。

[0162] 词语W_i在文本中的词长权重leni定义为：

[0163] $$len_i=\begin{cases}1-1/L,\ \text{if词的长度L大于2}\\0,\ otherwise\end{cases}$$

[0164] 对于词语序列中的每个词，其统计特征值为

[0165] $stats_i = A*TF_i + B*pos_i + C*len_i$

[0166] 其中，A+B+C＝1。

[0167] 所述的词语W_i关键度的计算过程如下：

[0168] 对应于加权词语网络中的每个节点，它的关键度值Imp_i定义为：

[0169] $Imp_i = \beta*stats_i + (1-\beta)*Z_i$

[0170] 其中，$0<\beta<1$。

[0171] 通过计算将得到关键度的值，从大到小排序，设定一个阈值γ（$0<\gamma<1$），取出前q个的值，则这些词语将作为科技项目的特征词语，这些词语充分反映主题，而且是比较重要的词语。

[0172] 步骤4.评审专家特征词语提取：评审专家信息量较科技项目信息少，科技项目的特征词构建网络并基于统计特征和语义特征的提取技术，不适合评审专家信息的特征词语提取，直接根据通用停用词库和专业停用词库进行停用词过滤，提取每个专家的特征词集合，通用停用词库是也是采用哈工大停用词表，专业停用词库需要人员进行不断地维护。

[0173] 步骤5.构建科技项目、评审专家的分字段知识表示模型：通过对空间向量模型和物元知识集模型进行扩展，依据科技项目中的不同字段信息建立文本表示模型PRO＝(id,F,WF,T,V)，其中id表示在项目库中的标识字段；F表示科技项目中字段类别集合；WF为字段的权重；T为特征词语；V表示字段所对应的词语及其权重集合即$V_i = \{v_{i1}, f(v_{i1}), v_{i2}, f(v_{i2}), \ldots, v_{in}, f(v_{in})\}$，$v_{ij}$表示第i个字段中的第j个特征词语，$f(v_{ij})$表示$v_{ij}$关键词所对应的频数。科技项目信息的知识表示如下：

[0174] $$PRO=\begin{pmatrix} 科技项目682 & 项目名称 & 0.4 & 中间件 & 0.25 \\ & 主要研究内容 & 0.3 & 项目 & 0.21 \\ & 技术指标 & 0.3 & 云计算 & 0.12 \end{pmatrix}$$

[0175] 同理，根据专家中的不同字段信息建立知识表示模型TM＝(id,F,WF,T,V)。其中，id表示在专家库中的标识字段；F表示评审专家中字段类别集合；WF为字段的权重集合；T为特征词语；V表示字段所对应的特征词语及其权重集合即$V_i = \{v_{i1}, f(v_{i1}), v_{i2}, f(v_{i2}), \ldots, v_{in}, f(v_{in})\}$，$v_{ij}$表示第i个字段中的第j个特征词语，$f(v_{ij})$表示$v_{ij}$特征词语在所对应的字段内的出现频率。评审专家信息的知识表示为：

[0176] $$TM=\begin{pmatrix} 评审专家2851 & 专家简历 & 0.1 & 中间件 & 0.5 \\ & 获奖情况 & 0.2 & 数字化 & 0.22 \\ & 发明专利情况 & 0.2 & 嵌入式 & 0.12 \\ & 承担项目及完成情况 & 0.1 & 平台 & 0.42 \\ & 发表论文情况 & 0.2 & 验收 & 0.42 \\ & 研究方向 & 0.2 & 组态 & 0.12 \end{pmatrix}$$

[0177] 评审专家信息索引库构建：待评审专家知识表示模型构建完成后，将信息索引入库：首先从专家库中读取一个评审专家的内容项信息；基于分词结果建立词语语义网络并提取评审专家所包含的特征词；依据知识表示模型并利用Apache Lucene对其建立索引；将建立好的索引按所属类别加至对应的索引库中，直到所有的评审专家索引入库，参见图5。

[0178] 步骤6：根据项目的个数，推荐方式分为单一待审项目推荐专家和分组(多个)待审

项目推荐专家。分组推荐专家对步骤5的待审项目知识表示模型做相应的字段间和项目间的特征合并操作，单一待审专家推荐只做相应的字段间特征合并操作。同时，对步骤5的评审专家的知识表示模型进行字段间特征合并。依据知识表示模型并利用Apache Lucene对合并后的特征信息建立索引。其中，科技项目索引构建在进行项目推荐时进行。

[0179] 科技项目申报管理系统中待审项目往往是需要分组推荐的，上述特征合并操作，确保不会消除步骤5中知识表示模型设置不同字段权重对相似度计算产生推荐的贡献差异。

[0180] 所述的待审项目、评审专家的特征合并通过逻辑异或操作进行过程如下：

[0181] (1)一个待审项目、一个评审专家的字段间特征合并

[0182] 假设字段特征词集合W'_1和W'_2合并，则定义W'_1和W'_2合并规则$W'_1 \oplus W'_2$为：

[0183] $$W'_1 \oplus W'_2 = \{\forall i,j,\{word_{1i}, \frac{f(word_{1i}) + f(word_{2i})}{2}\} \mid word_{1i} = word_{2j}\}$$

[0184] 其中，$word_{1i}$，$word_{2j}$为特征词。

[0185] 加入字段权重改进并扩展上述定义，对评审专家、科技项目的字段间特征进行合并，合并规则为：

[0186] $$W'_1 \oplus W'_2 = \{\forall i,j,\{word_{1i}, \frac{w1 * f(word_{1i}) + w2 * f(word_{2i})}{\sqrt{w1^2 + w2^2}}\} \mid word_{1i} = word_{2j}\}$$

[0187] (2)分组待审项目的项目间特征合并

[0188] 这一合并过程操作只针对待审科技项目的特征向量，不针对评审专家特征向量，专家特征向量只需要做字段间特征合并操作。若$V(d_1)$和$V(d_2)$分别是两个科技项目经过字段间特征合并后的向量模型，对任意$t_{1j} \in V(d_1)$，$t_{2j} \in V(d_2)$，若存在t_{1j}与t_{2j}相同则合并。$V(d_1) \oplus V(d_2)$定义为：

[0189] $$V(d_1) \oplus V(d_2) = \{< t_k, w_k(p) = \frac{w_i(d_1) + w_j(d_2)}{2} >\}$$

[0190] 其中，$k=1,\cdots,n$，t_k为特征词条项，$w_k(p)$为t_k的权重。

[0191] 科技项目组的知识模型表示产生的基本过程如下：

[0192] a).合并科技项目字段间特征，得到每个项目的向量模型V(d)；

[0193] b).将所有科技项目向量模型集合采用合并策略$V(p) = V(d_1) \oplus V(d_2) \ldots \oplus V(d_n)$。通过上述的方法，对科技项目组建立基于向量空间的知识表示模型。

[0194] $V(p) = \{< t_1, w_1(p) >, < t_2, w_2(p) >, \ldots, < t_n, w_n(p) >\}$

[0195] 其中，$k=1,\cdots,n$，t_k为项目组特征词词条项，$w_k(p)$为t_k的权重。

[0196] 步骤7.经过步骤6的评审专家和科技项目的知识表示模型的字段间特征进行合并，假设评审专家信息向量若表示为$P=\{s_1, f(s_1), s_2, f(s_2), \ldots, s_n, f(s_n)\}$，科技项目信息(组)向量表示为$Q=\{t_1, f(t_1), t_2, f(t_2), \ldots, t_n, f(t_n)\}$，基于最大匹配算法计算待审科技项目(组)向量与评审专家的语义相似度。

[0197] 所述待审科技项目(组)向量与评审专家向量的基于二部图最大匹配算法计算语义相似度计算过程如下：

[0198] 基于最大匹配算法计算语义相似度，就是获得两个文本的采用基于二部图的最大匹配算法相似度。如图2所示，基于二部图的最大匹配算法计算特征项的相似度，其原理就

是把科技项目(组)向量的每个特征词作为X部的一个顶点,评审专家向量的每个特征词作为Y部的一个顶点,等效为求一个完备二部图的最大权匹配,附图2中粗线部分就是X部特征词语与某个Y部特征词最大的语义相似度。

[0199] 所谓语义相似度,就是基于知网的相似度计算获得的。本发明借助知网语义词典和最大匹配算法计算待审项目(组)和评审专家间的语义相似度,则计算公式为:

[0200] $$SimSEM(P,Q)=(\sum_{k=1}^{p}\sqrt{f(s_i)*f(t_j)}*SimSEM(s_i,t_j))/\min(m,n)$$

[0201] 其中,s_i,t_j为语义相似度最大值SimSEM(s_i,t_j)的边(图2中粗线)所对应的两个词语节点,m,n分别为科技项目向量表示的特征词个数和评审专家向量表示的特征词个数。p为语义相似度最大的边(图2中粗线)的数目。

[0202] 上述待审项目(组)与评审专家信息的语义相似度涉及到语言、词语语义、词语结构等多种因素,它表示两者的匹配程度,相似度大,说明两者匹配度高,评审专家适合评审该项目(组)。

[0203] 步骤8.设置相似度截断,依据相似度的大小产生推荐指数,产生最终的推荐评审专家列表。

[0204] 以上所述仅是本发明的优选实施方式,应当指出,对于科技项目评审专家领域的智能机器推荐技术,在不脱离本发明技术原理的前提下,还可以做出若干改进和变形,这些改进和变形也应该视为本发明的法律保护范围。

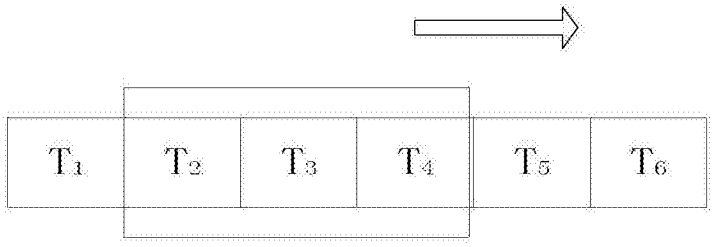

T1
T2
T3
T4
T5
T6

图1

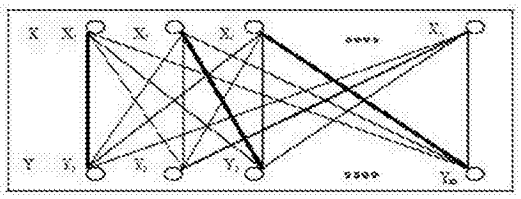


图2

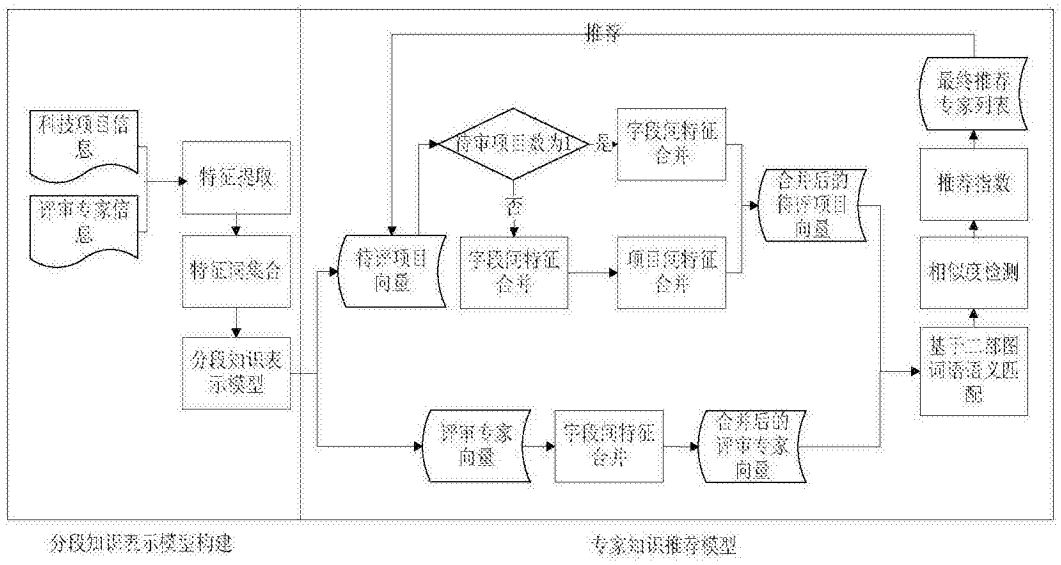

科技项目信息
评审专家信息
特征提取
特征向量合
分段知识表示模型
待评项目向量
待审项目数为1
是
否
字段间特征合并
项目间特征合并
合并后的待评项目向量
推荐
评审专家向量
字段间特征合并
合并后的评审专家向量
最终推荐专家列表
推荐指数
相似度检测
基于二部图词语语义匹配
分段知识表示模型构建
专家知识推荐模型

图3

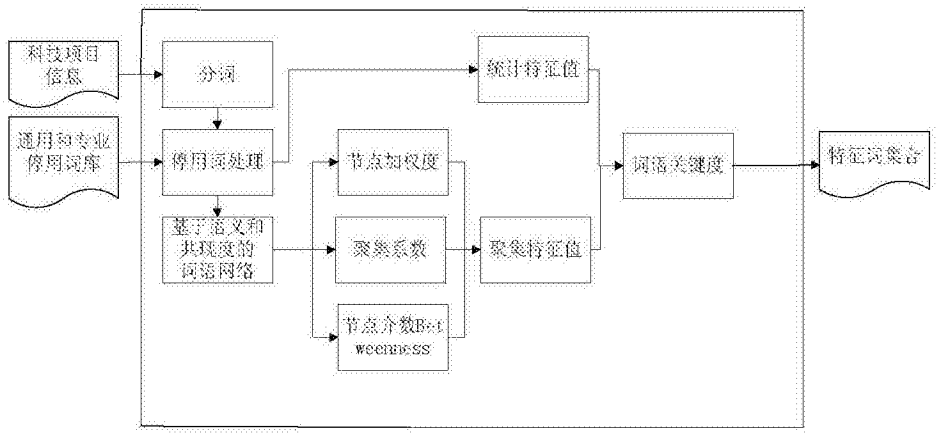

科技项目信息
通用和专业停用词库
分词
停用词处理
基于语义和共现度的词语网络
统计特征值
节点加权度
聚集系数
节点介数Betweenness
聚集特征值
词语关键度
特征词集合

图4

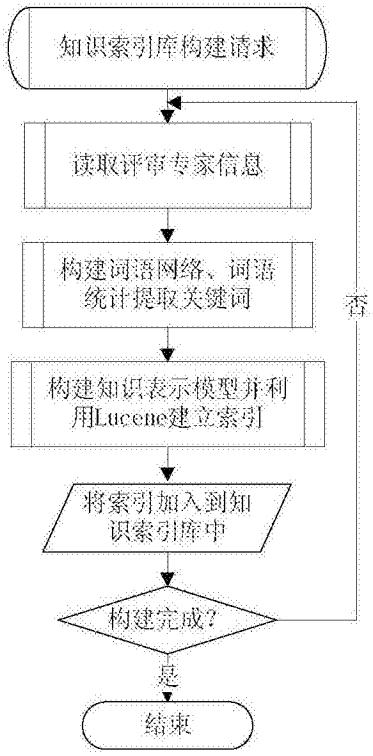

知识索引库构建请求
读取评审专家信息
构建词语网络、词语统计提取关键词
构建知识表示模型并利用Lucene建立索引
将索引加入到知识索引库中
构建完成?
否
是
结束

图5