



(12) 发明专利申请

(10) 申请公布号 CN 105335456 A

(43) 申请公布日 2016. 02. 17

(21) 申请号 201510610831. 5

(22) 申请日 2015. 09. 23

(71) 申请人 昆明理工大学

地址 650093 云南省昆明市五华区学府路
253 号

(72) 发明人 邵玉斌 井妍 王晨歌 杜庆治

(51) Int. Cl.

G06F 17/30(2006. 01)

权利要求书1页 说明书11页 附图1页

(54) 发明名称

一种用于环境保护法规检索的关联优先排序方法

(57) 摘要

本发明涉及一种用于环境保护法规检索的关联优先排序方法,属于知识发现领域。本发明首先对环保法律法规检索系统构建一个关键词表、关键字表;然后清洗用户输入的数据并提炼候选词;最后根据候选词的个数,计算距离并排序输出。本发明采用索引的方式,将庞大的信息源提炼成一个关键词表,作为整个信息源的目录索引,只要与索引匹配查询便能快速的在庞大的信息源中找到有意义的信息,进一步提高检索效率;采用计算整个关键词库中各个独立汉字之间的距离,将其距离值存储在关键字表中,因此在查询匹配的时候就只需要去寻找距离值最小的元素就能找到关联度很高的词语或词组;在提高检索效率的同时,也提高了检索结果与搜索意图之间关联度的准确性。

1. 一种用于环境保护法规检索的关联优先排序方法,其特征在于:首先对环保法律法规检索系统构建一个关键词表 A、关键字表 B;然后清洗用户输入的数据并提炼候选词;最后根据候选词的个数,计算距离并排序输出。

2. 根据权利要求 1 所述的用于环境保护法规检索的关联优先排序方法,其特征在于:所述用于环境保护法规检索的关联优先排序方法的具体步骤如下:

Step1、首先建立系统模型:

对环保法律法规检索系统构建一个关键词表 A、关键字表 B;其中,关键词表 A:存储着法规名称及法规中抽取出来的 t 组关键词;关键字表 B:存储着关键词表 A 中每个关键词拆分成的不同字 m 个及各字之间的特征值 A_{ij} ; A_{ij} 表示角标为 i 和 j 所代表的字的组合出现在关键词表 A 中的频数,角标 i、j 为关键词表 A 中每个关键词拆分成的不同字在关键字表 B 中的标记;

Step2、清洗用户输入的数据并提炼候选词:

针对用户输入的数据进行分词并去除停用词,将剩余的分词作为候选词;

Step3、根据候选词的个数,计算距离并排序输出:

Step3. 1、若候选词个数为 1 时:

从关键字表 B 中获取与候选词的首字 x 联结的字、首字 x 之间的特征值 A_{ix} 、获取尾字 y、与尾字 y 联结的字之间的特征值 A_{yj} ;计算 $A_{ix} \neq 0$ 情况下首字与关键字表 B 中字的距离 d_{ix} 且得到 ixy 对应的词组合,计算 $A_{yj} \neq 0$ 情况下尾字与关键字表 B 中字的距离 d_{yj} 且得到 xyj 对应的词组合;根据 d_{ix} 、 d_{yj} 从小到大的顺序排列其对应的词组合;根据词组合的顺序,将词组合与关键词表 A 中的关键词进行匹配获取对应的法规名称,将匹配的结果去除重复后按照顺序显示;其中,当出现 $d_{ix} = d_{yj}$,则 d_{ix} 、 d_{yj} 对应的词组合进行随机排序;

Step3. 2、若候选词个数不为 1 时:

将多个候选词按输入顺序排列,分别计算相邻两个候选词中先输入的候选词的尾字 u 与后输入的候选词的首字 v 的距离 d_{uv} 及对应的两个候选词构成的词组合;从关键字表 B 中获取与各个候选词的首字 x 联结的字、首字 x 之间的特征值 A_{ix} 、获取尾字 y、与尾字 y 联结的字之间的特征值 A_{yj} ;计算 $A_{ix} \neq 0$ 情况下首字与关键字表 B 中字的距离 d_{ix} 且得到 ixy 对应的词组合,计算 $A_{yj} \neq 0$ 情况下尾字与关键字表 B 中字的距离 d_{yj} 且得到 xyj 对应的词组合;根据 d_{uv} 、 d_{ix} 、 d_{yj} 从小到大的顺序排列其对应的词组合;根据词组合的顺序,将词组合与关键词表 A 中的关键词进行匹配获取对应的法规名称,将匹配的结果去除重复后按照顺序显示;其中,当出现 $d_{uv} = d_{ix} = d_{yj}$,则仅仅保留 d_{uv} 对应的词组合进行排序,当出现 $d_{ix} = d_{yj}$,则 d_{ix} 、 d_{yj} 对应的词组合进行随机排序;

所述 u、v、x、y 为字在关键字表 B 中的标记。

3. 根据权利要求 2 所述的用于环境保护法规检索的关联优先排序方法,其特征在于:

所述 $d_{uv} = \frac{1+e^{-A_{uv}}}{t}$ 、 $d_{ix} = \frac{1+e^{-A_{ix}}}{t}$ 、 $d_{yj} = \frac{1+e^{-A_{yj}}}{t}$;其中 A_{uv} 、 A_{ix} 、 A_{yj} 分别表示角标为 u、v 所

代表的字的组合,角标为 i、x 所代表的字的组合,角标为 y、j 所代表的字的组合出现在关键词表 A 中的频数; d_{uv} 、 d_{ix} 、 d_{yj} 分别表示角标为 u、v 所代表的字,角标为 i、x 所代表的字,角标为 y、j 所代表的字的距离。

一种用于环境保护法规检索的关联优先排序方法

技术领域

[0001] 本发明涉及一种用于环境保护法规检索的关联优先排序方法,属于知识发现领域。

背景技术

[0002] 信息爆炸是当今信息社会的一大特点,从 web 上进行搜索会查询到大量冗余繁琐信息,需要我们再逐一去筛选来获得我们想要的信息。因而如何快速找到一种方法,给用户更简洁的呈现出更有意义的信息成为了一个关键的问题。因此,为解决这一问题,提出知识发现,知识发现是从数据集中识别出有效的、新颖的、潜在有用的,以及最终可理解的模式非平凡过程。目的是向使用者屏蔽原始数据的繁琐细节,从原始数据中提炼出有意义的、简洁的知识,直接向使用者报告。为了向使用者提供更有意义的信息,本方法被提出来,它通过计算元素与元素之间的距离,即关联度,以最快的方式寻找到与使用者想搜索的信息的距离最优的词语组合,然后对应索引目录快速准确查找出更有意义的信息,即用户所需要信息。

发明内容

[0003] 本发明提供了一种用于环境保护法规检索的关联优先排序方法,以用于解决快速查找用户所需要信息的问题。

[0004] 本发明的技术方案是:一种用于环境保护法规检索的关联优先排序方法,首先对环保法律法规检索系统构建一个关键词表 A、关键字表 B;然后清洗用户输入的数据并提炼候选词;最后根据候选词的个数,计算距离并排序输出。

[0005] 所述用于环境保护法规检索的关联优先排序方法的具体步骤如下:

[0006] Step1、首先建立系统模型:

[0007] 对环保法律法规检索系统构建一个关键词表 A、关键字表 B;其中,关键词表 A:存储着法规名称及法规中抽取出来的 t 组关键词;关键字表 B:存储着关键词表 A 中每个关键词拆分成的不同字 m 个及各各个字之间的特征值 A_{ij} ; A_{ij} 表示角标为 i 和 j 所代表的字的组合出现在关键词表 A 中的频数,角标 i、j 为关键词表 A 中每个关键词拆分成的不同字在关键字表 B 中的标记;

[0008] Step2、清洗用户输入的数据并提炼候选词:

[0009] 针对用户输入的数据进行分词并去除停用词,将剩余的分词作为候选词;

[0010] Step3、根据候选词的个数,计算距离并排序输出:

[0011] Step3. 1、若候选词个数为 1 时:

[0012] 从关键字表 B 中获取与候选词的首字 x 联结的字、首字 x 之间的特征值 A_{ix} 、获取尾字 y、与尾字 y 联结的字之间的特征值 A_{yj} ;计算 $A_{ix} \neq 0$ 情况下首字与关键字表 B 中字的距离 d_{ix} 且得到 ixy 对应的词组合,计算 $A_{yj} \neq 0$ 情况下尾字与关键字表 B 中字的距离 d_{yj} 且得到 xyj 对应的词组合;根据 d_{ix} 、 d_{yj} 从小到大的顺序排列其对应的词组合;根据词

组合的顺序,将词组合与关键词表 A 中的关键词进行匹配获取对应的法规名称,将匹配的结果去除重复后按照顺序显示;其中,当出现 $dix = dyj$,则 dix 、 dyj 对应的词组合进行随机排序;

[0013] Step3. 2、若候选词个数不为 1 时:

[0014] 将多个候选词按输入顺序排列,分别计算相邻两个候选词中先输入的候选词的尾字 u 与后输入的候选词的首字 v 的距离 duv 及对应的两个候选词构成的词组合;从关键字表 B 中获取与各个候选词的首字 x 联结的字、首字 x 之间的特征值 Aix 、获取尾字 y 、与尾字 y 联结的字之间的特征值 Ayj ;计算 $Aix \neq 0$ 情况下首字与关键字表 B 中字的距离 dix 且得到 ixy 对应的词组合,计算 $Ayj \neq 0$ 情况下尾字与关键字表 B 中字的距离 dyj 且得到 xyj 对应的词组合;根据 duv 、 dix 、 dyj 从小到大的顺序排列其对应的词组合;根据词组合的顺序,将词组合与关键词表 A 中的关键词进行匹配获取对应的法规名称,将匹配的结果去除重复后按照顺序显示;其中,当出现 $duv = dix = dyj$,则仅仅保留 duv 对应的词组合进行排序,当出现 $dix = dyj$,则 dix 、 dyj 对应的词组合进行随机排序;

[0015] 所述 u 、 v 、 x 、 y 为字在关键字表 B 中的标记。

[0016] 所述 $duv = \frac{1+e^{-Auv}}{t}$ 、 $dix = \frac{1+e^{-Aix}}{t}$ 、 $dyj = \frac{1+e^{-Ayj}}{t}$;其中 Auv 、 Aix 、 Ayj 分别表示角标

为 u 、 v 所代表的字的组合,角标为 i 、 x 所代表的字的组合,角标为 y 、 j 所代表的字的组合出现在关键词表 A 中的频数; duv 、 dix 、 dyj 分别表示角标为 u 、 v 所代表的字,角标为 i 、 x 所代表的字,角标为 y 、 j 所代表的字的距离。

[0017] 本发明的有益效果是:

[0018] 采用索引的方式,将庞大的信息源提炼成一个关键词表,作为整个信息源的目录索引。因此,只要与索引匹配查询便能快速的在庞大的信息源中找到有意义的信息,进一步提高检索效率。

[0019] 采用计算整个关键词库中各个独立汉字之间的距离,将其距离值存储在关键字表中。因此在查询匹配的时候就只需要去寻找距离值最小的元素就能找到关联度很高的词语或词组。在提高检索效率的同时,也提高了检索结果与搜索意图之间关联度的准确性。

附图说明

[0020] 图 1 为本发明元素间距离网状示意图;

[0021] 图 2 为本发明元素间距离网状实例示意图。

具体实施方式

[0022] 实施例 1:如图 1-2 所示,一种用于环境保护法规检索的关联优先排序方法,首先对环保法律法规检索系统构建一个关键词表 A、关键字表 B;然后清洗用户输入的数据并提炼候选词;最后根据候选词的个数,计算距离并排序输出。

[0023] 所述用于环境保护法规检索的关联优先排序方法的具体步骤如下:

[0024] Step1、首先建立系统模型:

[0025] 对环保法律法规检索系统构建一个关键词表 A、关键字表 B;其中,关键词表 A:存储着法规名称及法规中抽取出来的 t 组关键词;关键字表 B:存储着关键词表 A 中每个关键

词拆分成的不同字 m 个及各个字之间的特征值 A_{ij} ; A_{ij} 表示角标为 i 和 j 所代表的字的组合出现在关键词表 A 中的频数, 角标 i 、 j 为关键词表 A 中每个关键词拆分成的不同字在关键字表 B 中的标记;

[0026] Step2、清洗用户输入的数据并提炼候选词:

[0027] 针对用户输入的数据进行分词并去除停用词, 将剩余的分词作为候选词;

[0028] Step3、根据候选词的个数, 计算距离并排序输出:

[0029] Step3. 1、若候选词个数为 1 时:

[0030] 从关键字表 B 中获取与候选词的首字 x 联结的字、首字 x 之间的特征值 A_{ix} 、获取尾字 y 、与尾字 y 联结的字之间的特征值 A_{yj} ; 计算 $A_{ix} \neq 0$ 情况下首字与关键字表 B 中字的距离 d_{ix} 且得到 ixy 对应的词组合, 计算 $A_{yj} \neq 0$ 情况下尾字与关键字表 B 中字的距离 d_{yj} 且得到 xyj 对应的词组合; 根据 d_{ix} 、 d_{yj} 从小到大的顺序排列其对应的词组合; 根据词组合的顺序, 将词组合与关键词表 A 中的关键词进行匹配获取对应的法规名称, 将匹配的结果去除重复后按照顺序显示; 其中, 当出现 $d_{ix} = d_{yj}$, 则 d_{ix} 、 d_{yj} 对应的词组合进行随机排序;

[0031] Step3. 2、若候选词个数不为 1 时:

[0032] 将多个候选词按输入顺序排列, 分别计算相邻两个候选词中先输入的候选词的尾字 u 与后输入的候选词的首字 v 的距离 d_{uv} 及对应的两个候选词构成的词组合; 从关键字表 B 中获取与各个候选词的首字 x 联结的字、首字 x 之间的特征值 A_{ix} 、获取尾字 y 、与尾字 y 联结的字之间的特征值 A_{yj} ; 计算 $A_{ix} \neq 0$ 情况下首字与关键字表 B 中字的距离 d_{ix} 且得到 ixy 对应的词组合, 计算 $A_{yj} \neq 0$ 情况下尾字与关键字表 B 中字的距离 d_{yj} 且得到 xyj 对应的词组合; 根据 d_{uv} 、 d_{ix} 、 d_{yj} 从小到大的顺序排列其对应的词组合; 根据词组合的顺序, 将词组合与关键词表 A 中的关键词进行匹配获取对应的法规名称, 将匹配的结果去除重复后按照顺序显示; 其中, 当出现 $d_{uv} = d_{ix} = d_{yj}$, 则仅仅保留 d_{uv} 对应的词组合进行排序, 当出现 $d_{ix} = d_{yj}$, 则 d_{ix} 、 d_{yj} 对应的词组合进行随机排序;

[0033] 所述 u 、 v 、 x 、 y 为字在关键字表 B 中的标记。

[0034] 所述 $d_{uv} = \frac{1+e^{-A_{uv}}}{t}$ 、 $d_{ix} = \frac{1+e^{-A_{ix}}}{t}$ 、 $d_{yj} = \frac{1+e^{-A_{yj}}}{t}$; 其中 A_{uv} 、 A_{ix} 、 A_{yj} 分别表示角标

为 u 、 v 所代表的字的组合, 角标为 i 、 x 所代表的字的组合, 角标为 y 、 j 所代表的字的组合出现在关键词表 A 中的频数; d_{uv} 、 d_{ix} 、 d_{yj} 分别表示角标为 u 、 v 所代表的字, 角标为 i 、 x 所代表的字, 角标为 y 、 j 所代表的字的距离。

[0035] 实施例 2: 如图 1-2 所示, 一种用于环境保护法规检索的关联优先排序方法, 首先对环保法律法规检索系统构建一个关键词表 A 、关键字表 B ; 然后清洗用户输入的数据并提炼候选词; 最后根据候选词的个数, 计算距离并排序输出。

[0036] 所述用于环境保护法规检索的关联优先排序方法的具体步骤如下:

[0037] Step1、首先建立系统模型:

[0038] 对环保法律法规检索系统构建一个关键词表 A 、关键字表 B ; 其中, 关键词表 A : 存储着法规名称及法规中抽取出来的 t 组关键词; 关键字表 B : 存储着关键词表 A 中每个关键词拆分成的不同字 m 个及各个字之间的特征值 A_{ij} ; A_{ij} 表示角标为 i 和 j 所代表的字的组合出现在关键词表 A 中的频数, 角标 i 、 j 为关键词表 A 中每个关键词拆分成的不同字在关

键字表 B 中的标记；

[0039] Step2、清洗用户输入的数据并提炼候选词；

[0040] 针对用户输入的数据进行分词并去除停用词，将剩余的分词作为候选词；

[0041] Step3、根据候选词的个数，计算距离并排序输出；

[0042] Step3. 1、若候选词个数为 1 时：

[0043] 从关键字表 B 中获取与候选词的首字 x 联结的字、首字 x 之间的特征值 A_{ix} 、获取尾字 y、与尾字 y 联结的字之间的特征值 A_{yj} ；计算 $A_{ix} \neq 0$ 情况下首字与关键字表 B 中字的距离 d_{ix} 且得到 ixy 对应的词组合，计算 $A_{yj} \neq 0$ 情况下尾字与关键字表 B 中字的距离 d_{yj} 且得到 xyj 对应的词组合；根据 d_{ix} 、 d_{yj} 从小到大的顺序排列其对应的词组合；根据词组合的顺序，将词组合与关键词表 A 中的关键词进行匹配获取对应的法规名称，将匹配的结果去除重复后按照顺序显示；其中，当出现 $d_{ix} = d_{yj}$ ，则 d_{ix} 、 d_{yj} 对应的词组合进行随机排序；

[0044] Step3. 2、若候选词个数不为 1 时：

[0045] 将多个候选词按输入顺序排列，分别计算相邻两个候选词中先输入的候选词的尾字 u 与后输入的候选词的首字 v 的距离 d_{uv} 及对应的两个候选词构成的词组合；从关键字表 B 中获取与各个候选词的首字 x 联结的字、首字 x 之间的特征值 A_{ix} 、获取尾字 y、与尾字 y 联结的字之间的特征值 A_{yj} ；计算 $A_{ix} \neq 0$ 情况下首字与关键字表 B 中字的距离 d_{ix} 且得到 ixy 对应的词组合，计算 $A_{yj} \neq 0$ 情况下尾字与关键字表 B 中字的距离 d_{yj} 且得到 xyj 对应的词组合；根据 d_{uv} 、 d_{ix} 、 d_{yj} 从小到大的顺序排列其对应的词组合；根据词组合的顺序，将词组合与关键词表 A 中的关键词进行匹配获取对应的法规名称，将匹配的结果去除重复后按照顺序显示；其中，当出现 $d_{uv} = d_{ix} = d_{yj}$ ，则仅仅保留 d_{uv} 对应的词组合进行排序，当出现 $d_{ix} = d_{yj}$ ，则 d_{ix} 、 d_{yj} 对应的词组合进行随机排序；

[0046] 所述 u、v、x、y 为字在关键字表 B 中的标记。

[0047] 实施例 3：如图 1-2 所示，一种用于环境保护法规检索的关联优先排序方法，首先对环保法律法规检索系统构建一个关键词表 A、关键字表 B；然后清洗用户输入的数据并提炼候选词；最后根据候选词的个数，计算距离并排序输出。

[0048] 实施例 4：如图 1-2 所示，一种用于环境保护法规检索的关联优先排序方法，首先对环保法律法规检索系统构建一个关键词表 A、关键字表 B；然后清洗用户输入的数据并提炼候选词；最后根据候选词的个数，计算距离并排序输出。

[0049] 所述方法的具体步骤如下：

[0050] 对环保法律法规检索系统构建一个关键词表 A、关键字表 B；其中，关键词表 A：存储着法规名称及法规中抽取出来的 t 组关键词；关键字表 B：存储着关键词表 A 中每个关键词拆分成的不同字 m 个及各各个字之间的特征值 A_{ij} ； A_{ij} 表示角标为 i 和 j 所代表的字的组合出现在关键词表 A 中的频数，角标 i、j 为关键词表 A 中每个关键词拆分成的不同字在关键字表 B 中的标记；

[0051] 表 A- 关键词表

[0052]

法规名称	关键词
中华人民共和国水污染防治法	水污染防治、城镇污水、饮用水水源、水污染、污染物、污染物排放
中华人民共和国大气污染防治法	大气污染防治、大气污染、大气污染物、大气环境质量、污染物排放
中华人民共和国环境噪声	环境噪声污染、噪声污染防治、环境

[0053]

污染防治法	噪声、环境噪声监测、噪声污染
中华人民共和国固体废物污染环境防治法	固体废物污染、生活垃圾、固体废物、污染环境防治
中华人民共和国海洋环境保护法	海洋环境、污染损害、陆源污染物
污染源监测管理办法	污染源监测、污染物排放、污染源
饮用水水源保护区污染防治管理规定	饮用水水源、水源保护区、污染防治、水污染防治、排放污水
中华人民共和国水法	水资源、生活用水、水量分配
中华人民共和国环境保护法	污染物排放、防治污染、环境影响
中华人民共和国水污染防治法实施细则	生活污水、水污染防治、生活饮用水

[0054] 关键字表 B : 存储着关键词表 A 中每个关键词拆分成的不同字 (称之为元素) m 个及各个元素之间的特征值 A_{ij} , A_{ij} 表示角标为 i 和 j 所代表的字的组合出现在关键词表 A 中的频数。

[0055] 表 B- 关键字表

[0056]

i	j	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
		水	污	染	源	物	防	治	海	洋	饮	用	气	环	境	质	量	排	放	噪	声
1	水	1	3	0	3	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	
2	源	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
3	放	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
4	境	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	3	0	
5	声	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
6	活	0	1	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	
7	治	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
8	气	0	3	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	
9	染	0	0	0	2	7	6	0	0	0	0	0	0	0	0	0	0	0	0	0	
10	污	4	0	23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
11	防	0	0	0	0	0	0	7	0	0	0	0	0	0	0	0	0	0	0	0	
12	城	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
13	镇	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
14	饮	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	
15	用	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
16	物	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	
17	大	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	
18	环	0	0	0	0	0	0	0	0	0	0	0	0	7	0	0	0	0	0	0	
19	质	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	
20	量	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
21	排	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	
22	噪	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	
23	监	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	
24	测	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
25	固	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
26	体	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
27	废	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
28	生	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

[0057]

29	垃	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
30	圾	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
31	海	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
32	洋	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
33	损	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
34	害	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
35	陆	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
36	保	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
37	护	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
38	区	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
39	资	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
40	分	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

[0058]

i	j	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
		监	测	固	体	废	垃	圾	损	害	陆	保	护	区	资	分	城	镇	大	生	活
1	水	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	源	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
3	放	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	境	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	声	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	活	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	治	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	气	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	染	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	污	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	防	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	城	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
13	镇	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	饮	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	用	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	物	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	大	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	环	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	质	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	量	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	排	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	噪	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	监	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
24	测	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
25	固	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
26	体	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
27	废	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
28	生	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
29	垃	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
30	圾	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
31	海	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
32	洋	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
33	损	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
34	害	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
35	陆	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
36	保	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
37	护	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
38	区	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
39	资	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
40	分	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

[0059] 从表 A 中,可以看出, $t = 40$; 从表 B 中,可知, $i = j = 1, 2, \dots, 40$ 。

[0060] Step2、清洗数据并提炼候选词：

[0061] 查询的本质,就是对词语词组的匹配或查询,则需要获取能表示替代用户查询意图的词语或词组,为方便表示,我们称之为候选词。

[0062] Step2. 1、假设用户输入“污染”,则：

[0063] 清洗用户输入的数据,将其分词并去除停用词,抽取出候选词;即提炼出候选词——“污染”一个候选词。

[0064] Step3、判断候选词类型并计算距离并排序输出：

[0065] 判断候选词个数为 1,则计算候选词“污染”的首字与尾字与关键字表中各个字的距离;读取关键字表,易知:与首字“污”联结的字与其之间的特征值如下: $A_{12} = 3, A_{22} =$

1, A32 = 1, A52 = 3, A62 = 1, A72 = 1, A82 = 3 ;与尾字“染”联结的字与其之间的特征值如下 :A94 = 2, A95 = 6, A96 = 6 ;在这里,联结方式为 :若是找到与词语的首字的联结组合,则找到形为“**+ 首字”的组合,若是找到与词语的首字的联结组合,则找到形为“尾字+**”的组合,下文不再赘述 ;

[0066] 则本实例中该词的第一个字与关键字表中字的距离为 :

$$[0067] \quad d_{12} = \frac{1+e^{-A_{12}}}{t} = \frac{1+e^{-3}}{40};$$

$$[0068] \quad d_{22} = \frac{1+e^{-A_{22}}}{t} = \frac{1+e^{-1}}{40};$$

$$[0069] \quad d_{32} = \frac{1+e^{-A_{32}}}{t} = \frac{1+e^{-1}}{40};$$

$$[0070] \quad d_{52} = \frac{1+e^{-A_{52}}}{t} = \frac{1+e^{-3}}{40};$$

$$[0071] \quad d_{62} = \frac{1+e^{-A_{62}}}{t} = \frac{1+e^{-1}}{40};$$

$$[0072] \quad d_{72} = \frac{1+e^{-A_{72}}}{t} = \frac{1+e^{-1}}{40};$$

$$[0073] \quad d_{82} = \frac{1+e^{-A_{82}}}{t} = \frac{1+e^{-3}}{40};$$

[0074] 本实例中该词的第二个字与关键字表中字的距离为 :

$$[0075] \quad d_{94} = \frac{1+e^{-A_{94}}}{t} = \frac{1+e^{-2}}{40};$$

$$[0076] \quad d_{95} = \frac{1+e^{-A_{95}}}{t} = \frac{1+e^{-6}}{40};$$

$$[0077] \quad d_{96} = \frac{1+e^{-A_{96}}}{t} = \frac{1+e^{-6}}{40};$$

[0078] 根据字之间距离越小相关性越大,将距离 d 从小到大排列,若值相等则随机排列,则其顺序为 :d95, d96, d12, d52, d82, d94, d22, d32, d62, d72 ;将关键字表里面字的与输入词中的字组合起来,回到关键词表中去匹配 :

[0079] 首先,由以上实例计算结果按距离从小到大(距离值并列则随机排列)可以得到的组合为 :“污染物”、“污染防”、“水污染”、“声污染”、“气污染”、“污染源”、“源污染”、“放污染”、“活污染”、“治污染”。

[0080] 然后,将得到的上列组合与关键词表 A 中的关键词匹配,看其是否存在于关键词表中,若存在,则将该关键词所对应的法规优先显示输出,若匹配不存在则进行下一组合的匹配。根据上列组合 :“污染物”能够和关键词表 A 中的“污染物”、“大气污染物”、“污染物排放”、“陆源污染物”匹配,可以索引到如下法规 :

[0081] 《中华人民共和国水污染防治法》

[0082] 《中华人民共和国大气污染防治法》

[0083] 《中华人民共和国海洋环境保护法》

[0084] 《中华人民共和国环境保护法》

[0085] 所以将这些法规优先显示,后序组合中如:“污染防”能够和关键词表 A 中的“水污染防治”、“大气污染防治”、“噪声污染防治”、“污染防治”匹配,可以索引到如下法规:

[0086] 《中华人民共和国水污染防治法》、《中华人民共和国大气污染防治法》、《中华人民共和国环境噪声污染防治法》、《饮用水水源保护区污染防治管理规定》、《中华人民共和国水污染防治法实施细则》,去重复,将本条得到的法规与之前组合索引得到的法规去重复,得到如下法规:《中华人民共和国环境噪声污染防治法》、《饮用水水源保护区污染防治管理规定》、《中华人民共和国水污染防治法实施细则》。

[0087] 后序组合依次类推。

[0088] Step4、假设用户输入“污染与防治”,则:

[0089] 清洗用户输入的数据,将其分词并去除停用词,抽取出候选词;即分词,去除停用词“与”提炼出候选词——“污染”和“防治”两个候选词。

[0090] 判断候选词类型并计算距离并排序输出:

[0091] 判断候选词个数不为 1,则:

[0092] 首先计算候选词“污染”和“防治”这两个候选词之间的距离,即计算出“染”与“防”这两个字的距离。读取关键字表 B,易知:“染”与“防”这两个关键字的特征值为 $A_{96} = 6$;则本实例中这两个候选词之间的距离为:

$$[0093] \quad d_{96} = \frac{1+e^{-A_{96}}}{t} = \frac{1+e^{-6}}{40};$$

[0094] 然后,计算各个候选词的首字与尾字与关键字表中各个字的距离;读取关键字表,易知:

[0095] 与第一个候选词“污染”的首字“污”联结的字与其之间的特征值如下: $A_{12} = 3$, $A_{22} = 1$, $A_{32} = 1$, $A_{52} = 3$, $A_{62} = 1$, $A_{72} = 1$, $A_{82} = 3$;与尾字“染”联结的字与其之间的特征值如下: $A_{94} = 2$, $A_{95} = 6$, $A_{96} = 6$;

[0096] 则本实例中该词的第一个字与关键字表中字的距离为:

$$[0097] \quad d_{12} = \frac{1+e^{-A_{12}}}{t} = \frac{1+e^{-3}}{40};$$

$$[0098] \quad d_{22} = \frac{1+e^{-A_{22}}}{t} = \frac{1+e^{-1}}{40};$$

$$[0099] \quad d_{32} = \frac{1+e^{-A_{32}}}{t} = \frac{1+e^{-1}}{40};$$

$$[0100] \quad d_{52} = \frac{1+e^{-A_{52}}}{t} = \frac{1+e^{-3}}{40};$$

$$[0101] \quad d_{62} = \frac{1+e^{-A_{62}}}{t} = \frac{1+e^{-1}}{40};$$

$$[0102] \quad d_{72} = \frac{1+e^{-A_{72}}}{t} = \frac{1+e^{-1}}{40};$$

$$[0103] \quad d_{82} = \frac{1+e^{-A_{82}}}{t} = \frac{1+e^{-3}}{40};$$

[0104] 本实例中该词的第二个字与关键字表中字的距离为：

$$[0105] \quad d_{94} = \frac{1+e^{-A_{94}}}{t} = \frac{1+e^{-2}}{40};$$

$$[0106] \quad d_{95} = \frac{1+e^{-A_{95}}}{t} = \frac{1+e^{-6}}{40};$$

$$[0107] \quad d_{96} = \frac{1+e^{-A_{96}}}{t} = \frac{1+e^{-6}}{40};$$

[0108] 计算第二个候选词“防治”的首字与尾字和关键字表中各个字的距离；读取关键字表 B，易知：与首字“防”联结的字与其之间的特征值如下： $A_{46} = 1$ ， $A_{96} = 6$ 与尾字“治”联结的字与其之间的特征值皆为 $A_{72} = 1$ ；

[0109] 则本实例中“防治”这一词的首字与关键字表中的距离为：

$$[0110] \quad d_{46} = \frac{1+e^{-A_{46}}}{t} = \frac{1+e^{-1}}{40};$$

$$[0111] \quad d_{96} = \frac{1+e^{-A_{96}}}{t} = \frac{1+e^{-6}}{40};$$

[0112] 本实例中该词尾字与关键字表中字的距离为：

$$[0113] \quad d_{72} = \frac{1+e^{-A_{72}}}{t} = \frac{1+e^{-1}}{40};$$

[0114] 根据字之间距离越小相关性越大，将距离 d 从小到大排列，若值相等则随机排列，其顺序为： d_{96} ， d_{95} ， d_{12} ， d_{52} ， d_{82} ， d_{94} ， d_{22} ， d_{32} ， d_{62} ， d_{72} ， d_{46} ；将关键字表里面字的与输入词中的字组合起来，回到关键词表中去匹配：

[0115] 首先，由以上实例计算结果按距离从小到大（距离值并列则随机排列）可以得到的组合为：“污染防治”、“污染物”、“水污染”、“声污染”、“气污染”、“污染源”“源污染”、“放污染”、“活污染”、“治污染”、“境防治”；（“污染防治”为计算两个候选词之间的距离所得到的组合，由于两个候选词之间的距离所得到的 d_{96} 与后续的词的首尾字与关键字表中的字的组合之间的距离出现同一个值，即 d_{96} ，所以看起来有问题，实际上当两个词之间的距离与其他字的组合的距离出现重复距离值时，选择两个词之间的距离组合）

[0116] 然后，将得到的上列组合与关键词表 A 中的关键词匹配，看其是否存在于关键词表中，若存在，则将该关键词所对应的法规优先显示输出，若匹配不存在则进行下一组合的匹配。

[0117] 根据上列组合：

[0118] “污染防治”能够和关键词表 A 中的“水污染防治”、“大气污染防治”、“噪声污染防治”、“污染防治”匹配，索引得到如下法规结果：

[0119] 《中华人民共和国水污染防治法》

[0120] 《中华人民共和国大气污染防治法》

[0121] 《中华人民共和国环境噪声污染防治法》

[0122] 《饮用水水源保护区污染防治管理规定》

[0123] 《中华人民共和国水污染防治法实施细则》

[0124] “境防治”能够和关键词表 A 中的“水污染防治”、“大气污染防治”匹配,索引得到如下法规:

[0125] 《中华人民共和国固体废物污染环境防治法》;

[0126] 所以这些法规优先显示,其他组合依次类推。

[0127] 上面结合附图对本发明的具体实施方式作了详细说明,但是本发明并不限于上述实施方式,在本领域普通技术人员所具备的知识范围内,还可以在不脱离本发明宗旨的前提下作出各种变化。

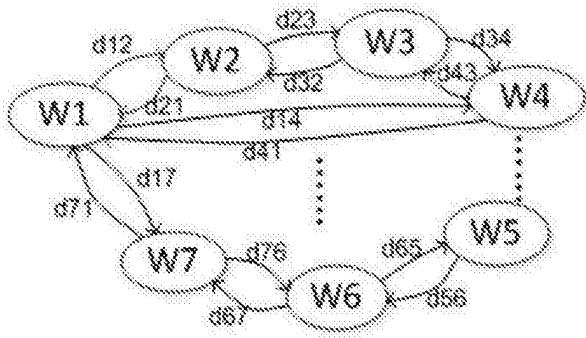


图 1

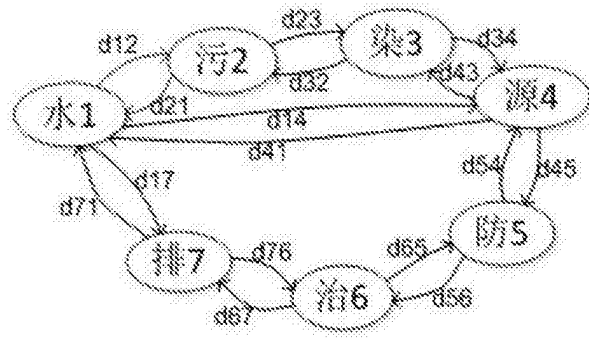


图 2