(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 105608201 A
(43)申请公布日 2016.05.25

(21)申请号 201510995759.2

(22)申请日 2015.12.28

(71)申请人 湖南蚁坊软件有限公司
地址 410003 湖南省长沙高新区麓谷企业广场A4栋607室

(72)发明人 舒琦

(51)Int.Cl.
G06F 17/30(2006.01)

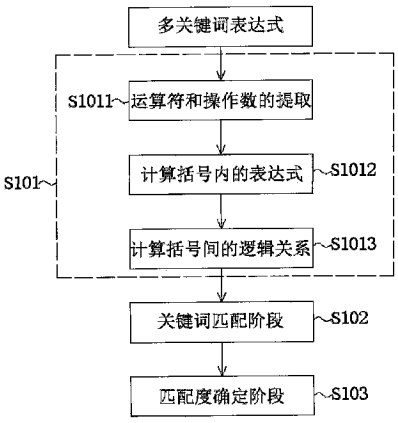
权利要求书1页 说明书2页 附图1页

(54)发明名称

一种支持多关键词表达式的文本匹配方法

(57)摘要

本发明涉及网络检索技术领域，特别是一种支持多关键词表达式的文本匹配方法，包括以下步骤，步骤S101：语法转换阶段，将多关键词表达式转换为多组关键词；步骤S102：关键词匹配阶段，以语法转换阶段输出的多组关键词作为输入，采用关键词匹配算法完成，获得文本中出现过的关键词；步骤S103：匹配度确定阶段，以关键词匹配阶段输出的出现过关键词的文本作为输入，将关键词匹配阶段出现过的关键词与语法转换阶段获得的多组关键词进行匹配程度确定。采用上述方法后，本发明多关键词表达式的文本匹配方法，支持多关键词表达式进行文本匹配，能够在一个表达式中表达复杂的匹配逻辑，提供了更为强大的匹配能力。



CN 105608201 A

1.一种支持多关键词表达式的文本匹配方法，其特征在于，包括以下步骤，

步骤S101：语法转换阶段，将多关键词表达式转换为多组关键词；

步骤S102：关键词匹配阶段，以语法转换阶段输出的多组关键词作为输入，采用关键词匹配算法完成，获得文本中出现过的关键词；

步骤S103：匹配度确定阶段，以关键词匹配阶段输出的出现过关键词的文本作为输入，将关键词匹配阶段出现过的关键词与语法转换阶段获得的多组关键词进行匹配程度确定。

2.按照权利要求1所述的一种支持多关键词表达式的文本匹配方法，其特征在于，所述步骤S101语法转换阶段具体包括以下步骤，

步骤S1011：运算符和操作数提取，提取多关键词表达式中的运算符和操作数；

步骤S1012：括号内表达式计算，优先计算括号内的表达式；

步骤S1013：括号间逻辑关系，计算各括号间的逻辑关系。

3.按照权利要求1所述的一种支持多关键词表达式的文本匹配方法，其特征在于：所述步骤S101中任意一组中的关键词必须同时出现，组之间任意出现一组，表示文本匹配成功。

4.按照权利要求3所述的一种支持多关键词表达式的文本匹配方法，其特征在于：步骤S103中将关键词匹配阶段出现过的关键词与语法转换阶段获得的多组关键词进行匹配程度确定是指判断语法转换阶段获得的多组关键词中是否存在任意一组关键词是关键词匹配阶段出现的关键词的子集；如果存在，则确定该文本匹配成功，否则匹配失败。

一种支持多关键词表达式的文本匹配方法

技术领域

[0001] 本发明涉及网络检索技术领域，特别是一种支持多关键词表达式的文本匹配方法。

背景技术

[0002] 针对文本数据，存在多个关键词需要匹配的情况下，已有较多经典算法，这些算法解决的问题都是如何在文本数据中精确匹配关键词，主要差别在算法的复杂度上，即给定n个关键词，针对一段文本数据，当计算结束时，会给出在文本中出现过的关键词。

[0003] 在实际运用中，可以借助逻辑运算符将多个关键词连接在一起，形成一个关键词表达式，从而能够表述关键词间更复杂的逻辑关系，继而获得更强大的匹配能力，这是目前的算法所不支持的。

[0004] 举个例子，给定3个关键词：中国、足球、2015，传统算法只会给出匹配到了哪些关键词；假如用户想关心的是中国足球在2015年的相关信息，那么可以将3个关键词表达为“中国&&足球&&2015”，意为该3个关键词必须在同一个文本数据中同时出现，才算命中。

[0005] 中国发明专利申请CN 101398820 A公开了一种大规模关键词匹配方法，包括预处理阶段和模式匹配阶段，预处理阶段包括关键词特征串裁剪、基于关键词特征串集合的多个简单布隆过滤器的构造，基于关键词特征串集合的哈希表构造：模式匹配阶段包括：利用先前构造的简单布隆过滤器序列实现当前窗口中文本串不与任何关键词特征串匹配的快速判定；在判定失败情况下执行与候选关键词的精确匹配；文本扫描过程中，可以利用递归算法快速计算出当前文本相对于各简单布隆过滤器的当前散列值。虽然，此发明利用里递归散列算法高效的特点，可实现大规模关键词场景下的高速匹配，但是此发明无法对关键词表达式进行文本匹配。

发明内容

[0006] 本发明需要解决的技术问题提供一种能够基于多关键词表达式进行文本匹配的方法。

[0007] 为解决上述的技术问题，本发明的一种支持多关键词表达式的文本匹配方法，包括以下步骤，

[0008] 步骤S101：语法转换阶段，将多关键词表达式转换为多组关键词；

[0009] 步骤S102：关键词匹配阶段，以语法转换阶段输出的多组关键词作为输入，采用关键词匹配算法完成，获得文本中出现过的关键词；

[0010] 步骤S103：匹配度确定阶段，以关键词匹配阶段输出的出现过关键词的文本作为输入，将关键词匹配阶段出现过的关键词与语法转换阶段获得的多组关键词进行匹配程度确定。

[0011] 进一步的，所述步骤S101语法转换阶段具体包括以下步骤，

[0012] 步骤S1011：运算符和操作数提取，提取多关键词表达式中的运算符和操作数；

[0013] 步骤S1012:括号内表达式计算,优先计算括号内的表达式;

[0014] 步骤S1013:括号间逻辑关系,计算各括号间的逻辑关系。

[0015] 进一步的,所述步骤S101中任意一组中的关键词必须同时出现,组之间任意出现一组,表示文本匹配成功。

[0016] 更进一步的,步骤S103中将关键词匹配阶段出现过的关键词与语法转换阶段获得的多组关键词进行匹配程度确定是指判断语法转换阶段获得的多组关键词中是否存在任意一组关键词是关键词匹配阶段出现的关键词的子集;如果存在,则确定该文本匹配成功,否则匹配失败。

[0017] 采用上述方法后,本发明多关键词表达式的文本匹配方法,支持多关键词表达式进行文本匹配,能够在一个表达式中表达复杂的匹配逻辑,提供了更为强大的匹配能力。

附图说明

[0018] 下面将结合附图和具体实施方式对本作进一步详细的说明。

[0019] 图1为本发明一种支持多关键词表达式的文本匹配方法的流程图。

具体实施方式

[0020] 如图1所示,本发明的一种支持多关键词表达式的文本匹配方法,包括以下步骤,

[0021] 步骤S101:语法转换阶段,将多关键词表达式转换为多组关键词。

[0022] 所述步骤S101语法转换阶段具体包括以下步骤,

[0023] 步骤S1011:运算符和操作数提取,提取多关键词表达式中的运算符和操作数;

[0024] 步骤S1012:括号内表达式计算,优先计算括号内的表达式;

[0025] 步骤S1013:括号间逻辑关系,计算各括号间的逻辑关系。

[0026] 语法转换阶段是将多关键词表达式转换为另一种表达形式,即转换为多组关键词,一组中的关键词必须是同时出现,组之间任意出现一组,就表示文本匹配成功。以“(西游记之大圣归来||捉妖记)&&影评”为例,转换后的表达形式为2组关键词:“西游记之大圣归来影评”、“捉妖记影评”,待匹配文本只要出现上述2组关键词中的任一组即匹配成功。

[0027] 步骤S102:关键词匹配阶段,以语法转换阶段输出的多组关键词作为输入,采用关键词匹配算法完成,获得文本中出现过的关键词。基于经典的多关键词匹配算法完成,算法有多种,可根据实际需求进行选择,在此不再累述,该阶段完成后,获得文本中出现过的关键词。

[0028] 步骤S103:匹配度确定阶段,以关键词匹配阶段输出的出现过关键词的文本作为输入,将关键词匹配阶段出现过的关键词与语法转换阶段获得的多组关键词进行匹配程度确定。步骤S103中将关键词匹配阶段出现过的关键词与语法转换阶段获得的多组关键词进行匹配程度确定是指判断语法转换阶段获得的多组关键词中是否存在任意一组关键词是关键词匹配阶段出现的关键词的子集;如果存在,则确定该文本匹配成功,否则匹配失败。

[0029] 虽然以上描述了本发明的具体实施方式,但是本领域熟练技术人员应当理解,这些仅是举例说明,可以对本实施方式作出多种变更或修改,而不背离发明的原理和实质,本发明的保护范围仅由所附权利要求书限定。

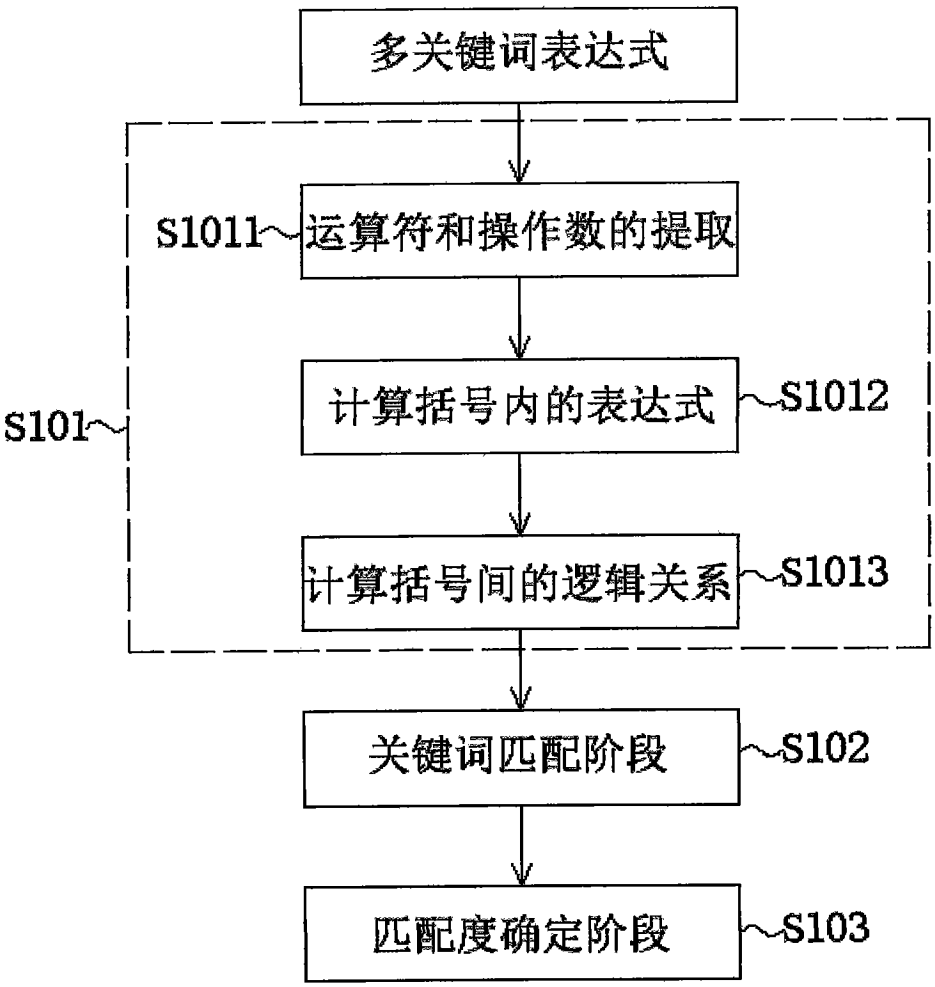

多关键词表达式
S1011
运算符和操作数的提取
计算括号内的表达式
S1012
S101
计算括号间的逻辑关系
S1013
关键词匹配阶段
S102
匹配度确定阶段
S103

图1