



(12) 发明专利申请

(10) 申请公布号 CN 103870565 A

(43) 申请公布日 2014. 06. 18

(21) 申请号 201410086557. 1

(22) 申请日 2014. 03. 11

(71) 申请人 华东师范大学

地址 200241 上海市闵行区东川路 500 号

(72) 发明人 何格 顾君忠 兰小敏

(74) 专利代理机构 上海蓝迪专利事务所 31215

代理人 徐筱梅 张翔

(51) Int. Cl.

G06F 17/30(2006. 01)

G06F 17/27(2006. 01)

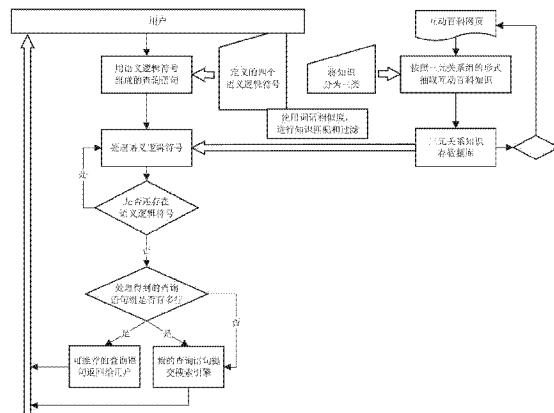
权利要求书1页 说明书5页 附图6页

(54) 发明名称

一种基于互动百科知识的语义逻辑引导的搜索方法

(57) 摘要

本发明公开了一种基于互动百科知识的语义逻辑引导的搜索方法，该方法将互动百科在线描述的知识数据分为概念、关系和实体三类，根据三类知识之间的关系，形成三元关系组，以三元关系组的形式抽取并存储互动百科的知识；设定四个语义逻辑符号，用户使用四个语义逻辑符号构建查询语句，然后利用三元关系组来处理该查询语句，得到的新的查询语句提交搜索引擎，并将处理得到的其他查询语句推荐给用户；本发明使用在一般搜索引擎不能处理的相似、相关逻辑，并将此查询语句处理为一般搜索引擎能有效处理的新查询语句，提高了搜索引擎的搜索精度。同时可以推荐给用户更精确的查询语句。从而提高了搜索引擎的效率，提高用户查询满意度。



1. 一种基于互动百科知识的语义逻辑引导的搜索方法,其特征在于该方法包括以下具体步骤:

a) 将互动百科在线描述的知识数据分为概念、关系和实体三类,根据三类知识之间的关系,形成三元关系组,以三元关系组的形式抽取并存储互动百科的知识;

b) 设定四个语义逻辑符号,用户使用四个语义逻辑符号构建查询语句,然后利用三元关系组来处理该查询语句,得到的新的查询语句提交搜索引擎,并将处理得到的其他查询语句推荐给用户;其中,

所述概念具有唯一的语义,为互动百科的标题;关系是概念的属性以及所有与概念有联系的关系的描述;实体不具有唯一的语义,概念的某种关系下对应的是实体或是某个概念;

所述三类知识之间的关系是:概念、实体及两者之间的关系和概念、概念及两者之间的关系;概念、关系及实体形成三元关系组或者概念、概念及两者之间形成三元关系组;

所述以三元关系组的形式抽取互动百科知识是:互动百科网页标题定为该页主概念,该互动百科网页内的信息对 - 属性、人物关系、从属关系,有冒号两边、从标题及其下属文本规律;信息对分别对应关系和实体或者关系和概念;

所述四个语义逻辑符号是:

“.” 从属、构成相关逻辑:根据概念和其相关的关系获取对应的实体或概念集;

“:” 相关、定义逻辑:根据概念描述或与概念相关的词来明确该概念;

“^” 相关逻辑:根据概念、关系及实体或概念之间的三元关系组的特点,由其中的两元推出第三元;

“~” 相似逻辑:使用相似逻辑来筛选数据;

所述使用四个语义逻辑符号构建查询语句,利用三元关系组处理查询语句是:用户使用符号代替自然语言中的逻辑,然后利用三元关系组的特点来处理该逻辑。

一种基于互动百科知识的语义逻辑引导的搜索方法

技术领域

[0001] 本发明涉及信息抽取、查询处理、信息检索，具体地说是基于互动百科海量的三元关系组知识集，结合知识匹配算法来引导用户的搜索。

背景技术

[0002] 信息检索技术包括搜索引擎是人们日常生活中寻找信息的常用工具，尤其是互联网使用的普及化后，搜索引擎成为必不可少的网络应用。

[0003] 随着互联网信息的海量增长和人们对信息检索的高准确性要求，单单依靠倒排索引和词匹配技术的搜索引擎，在用户输入的查询语句普遍比较短，不能详细而准确描述查询意图的情况下，搜索效果不够理想。尤其是词语的多义问题让搜索效果更加恶化。目前有很多研究从搜索引擎方面如语义消歧、查询扩展、查询日志分析、基于概念的搜索等以及从用户使用角度的公式化查询、用户实时交互、查询处理等在一定程度上提高了搜索引擎的搜索效果。而从用户角度上如何让用户以更直接的更自然的语言输入搜索引擎能有效处理的查询语句的研究还需很大程度上的投入。

发明内容

[0004] 本发明的目的是从用户角度帮助用户更准确的描述搜索意图，尤其是帮助用户将搜索引擎不能处理的带有语义逻辑的查询语句处理为依靠倒排和词匹配的搜索引擎可以处理的一般逻辑而提供的一种基于互动百科知识的语义逻辑引导的搜索方法，该方法定义了四个语义逻辑符号，用户可以使用这些符号来编辑查询语句，编辑好的查询语句在提交给搜索引擎之前，会先使用从互动百科中抽取的海量三元关系组结合词语相似度的知识匹配和筛选来解析符号逻辑，生成搜索引擎能有效处理的只包含一般的与或非逻辑的查询语句。同时可以返回推荐给用户的其他查询语句。

[0005] 实现本发明目的的具体技术方案是：

[0006] 一种基于互动百科知识的语义逻辑引导的搜索方法，其特征在于该方法包括以下具体步骤：

[0007] a) 将互动百科在线描述的知识数据分为概念、关系和实体三类，根据三类知识之间的关系，形成三元关系组，以三元关系组的形式抽取并存储互动百科的知识；

[0008] b) 设定四个语义逻辑符号，用户使用四个语义逻辑符号构建查询语句，然后利用三元关系组的特点来处理该查询语句，得到的新的查询语句提交搜索引擎，并将处理得到的其他查询语句推荐给用户；其中，

[0009] 所述概念具有唯一的语义，为互动百科的标题；关系是概念的属性以及所有与概念有联系的关系的描述；实体不具有唯一的语义，概念的某种关系下对应的是实体或是某个概念；

[0010] 所述三类知识之间的关系是：概念、实体及两者之间的关系和概念、概念及两者之间的关系；概念、关系及实体形成三元关系组或者概念、概念及两者之间形成三元关系组；

[0011] 所述以三元关系组的形式抽取互动百科知识是：互动百科网页标题定为该页主概念，该互动百科网页内的信息对 - 属性、人物关系、从属关系，有冒号两边、从标题及其下属文本规律；信息对分别对应关系和实体或者关系和概念；

[0012] 所述四个语义逻辑符号是：

[0013] “.” 从属、构成相关逻辑：根据概念和其相关的关系获取对应的实体或概念集；

[0014] “：“ 相关、定义逻辑：根据概念描述或与概念相关的词来明确该概念；

[0015] “^” 相关逻辑：根据概念、关系及实体或概念之间的三元关系组的特点，由其中的两元推出第三元；

[0016] “~” 相似逻辑：使用相似逻辑来筛选数据；

[0017] 所述使用四个语义逻辑符号构建查询语句，利用三元关系组处理查询语句是：用户使用符号代替自然语言中的逻辑，然后利用三元关系组的特点来处理该逻辑。

[0018] 本发明将百科知识分为三类数据，通过定义的四个语义逻辑符号，利用从互动百科海量的百科知识中抽取的三元关系组知识集，结合词语相似度的知识匹配，来帮助用户以更自然的语言逻辑构建查询语句，尤其是使用在一般搜索引擎不能处理的相似、相关逻辑，并将此查询语句处理为一般搜索引擎能有效处理的新的查询语句，提高了搜索引擎的搜索精度。同时可以推荐给用户更精确的查询语句，甚至直接返回给用户想要的知识。从而提高了搜索引擎的效率，提高用户查询满意度。

附图说明

[0019] 图 1 为本发明的流程图；

[0020] 图 2 为本发明中互动百科知识抽取的详细流程。

[0021] 图 3-6 为本发明在搜索中四个语义逻辑符号的处理流程图；

[0022] 图 7、图 8、图 9 为实施本发明实施例图。

具体实施方式

[0023] 语义逻辑符号引导的搜索需要使用语义逻辑符号编辑查询语句，同时需要互动百科里的三元关系组知识集。参阅图 1，其中互动百科知识的抽取和语义逻辑符号的定义和处理是重要部分，分别如下所述。

[0024] 1. 互动百科知识的抽取

[0025] 和维基百科一样，互动百科包含了大量的人工编辑、判断等参与工作，是一个非常开放、自增长且不断进步的百科全书。一般情况下，可以在任意一篇互动百科的网页文档上看到各种三元关系，所谓三元关系是指两个事物加上两者之间的关系共同构成一个三元关系组。针对互动百科的网页文档来看，其中的一元是网页的标题，另外的两元会分布在冒号的两边或者粗体字和它下面的叙述文本之间。

[0026] 在这里，将一篇互动百科网页的标题定义为概念，上述的另外两元分别为关系和实体(实体也可以是一个概念)。也就是说三元关系组是概念与概念或者概念与实体加上它们之间的关系构成的。那么将互动百科的知识分为三大类：概念、关系、实体。

[0027] 实体由一组短语或术语组成，可以是具体的人名、组织名、书名、具体的事件等。不过很多时候，不能单凭一个人名及确定他是谁，因为有很多同名同姓却不同的人。例如互动

百科上,针对名字“王菲”除了“王菲 [中国女歌手]”还有“王菲 [大同大学教授]”等。也就是说,不能通过一个实体确定它是什么。

[0028] 概念则表示一个确定的事物。例如在互动百科上只有一个“王菲 [中国女歌手]”。

[0029] 概念之间或者概念与实体之间的关系有很多种,例如“王菲 [中国女歌手]”的“血型”、“专辑”、“电影作品”等等。

[0030] 依据互动百科网网页文档的 DOM 框架及其规律,通过解析其 HTML 网页来抽取三元知识,这个过程可以实时进行。

[0031] 其中的 DOM 规律主要有(以“王菲”的互动百科页面为例):

[0032] 1) 通过“div.content-h1h1”得到标题“王菲 [中国女歌手]”,此为概念。

[0033] 2) “p#polysemyPart a”得到关于“王菲”的多个概念:“王菲 [中国女歌手]”、“王菲 [四川省广元市代市长]”和“王菲 [大同大学教授]”。

[0034] 3) “div.summary p”得到摘要。

[0035] 4) “div#figurerelation li a”得到人物关系:李亚鹏前夫;夏桂影母亲;王佑林父亲等。

[0036] 5) “div.module.zoom td”得到“王菲”的属性:中文名:王菲;英文名:Faye Wong 等。

[0037] 6) “div#content p”获取概况信息:“出生医院:北京协和医院,就读小学:北京地坛小学”等。

[0038] 7) “div#content h3”和获取其他相关关系。如:专辑

[0039] 1985 年:《风从哪里来》《迷人的卡勒》《迷人小姐》

[0040] 1986 年:《分手》《邓丽君故乡情》

[0041] 1989 年:《王靖雯》等。

[0042] 具体抽取过程如图 2。

[0043] 2、语义逻辑符号的定义和处理

[0044] 为了充分利用互动百科海量的三元关系组知识来帮助信息检索,定义四个语义逻辑符号。其基本原理是根据三元关系组其中的两元来确定第三元。

[0045] 1) 根据概念和关系词获取对应的实体,使用符号“.”。其格式为:实体. 关系词

[0046] 例如“王菲. 女儿”,从互动百科的三元关系组知识库里可以得到“李嫣、窦靖童”。

[0047] 通过此符号获取的实体可以是多个,例如“王菲. 专辑”得到的是一系列的王菲歌曲专辑。

[0048] 第一个“实体”为多语义的情况下,例如“王菲”有多个概念“中国女歌手”、“大同大学教授”等。可以根据是否同时存在两元分别为给定的两个实体,来确定为哪个概念。例如“王菲. 专辑”则可以确定为获取“中国女歌手王菲”的专辑,因为其他两个概念“王菲”——“王菲 [四川省广元市市长]”和“王菲 [大同大学教授]”没有“专辑”。

[0049] 注意:处理后的结果为给定的两个“实体”加上获取的第三元。如“王菲. 女儿”处理后变为“王菲女儿李嫣”。

[0050] 2) 根据三元中的任意两元获取第三元,使用符号“^”。

[0051] 实体 ^ 词语

[0052] 这里的“词语”可以为关系词或者实体。

[0053] 处理后的结果即为第三元。例如：“美国 ^ 奥巴马”可以从三元关系组知识库里得到“总统”。处理得到的结果可能不止一个，可能还有“国籍”等。

[0054] 3) 根据概念词或相关词确定明确概念，使用符号“：“。

[0055] 实体：词语

[0056] 通过描述具体概念的词或者相关词可以确定具体的概念。这里是通过两个词的共线性来确定具体概念词的，类似前面的例子“王菲 . 专辑”。

[0057] 例如“王菲 : 歌手”或“王菲 : 专辑”可以处理得到“中国女歌手王菲”。

[0058] 4) 从词集中选取与某词最相似的词，使用符号“~”。

[0059] 词集 ~ 词语

[0060] 相似逻辑通常配合获取实体集的符号“.”来使用。例如“古天乐 . 电影 ~ 甜蜜”其意义是指从古天乐的电影列表中找到与“甜蜜”最相似的电影名字。其最后的处理结果为“古天乐电影甜言蜜语”。

[0061] 这里的相似逻辑就需要用到词语的相似度计算。考虑到用户提交搜索时，对于未知或模糊的部分一般输入具有相同字或词。这里就使用了统计相同字和词的方法来计算词语相似度，其中词相比字更重要，所以两者应该有不同的权重计算：

$$[0062] TS(p, p_i) = \sigma \cdot \left(\alpha \cdot \frac{Sch(p, p_i)}{Lch(p)} + \beta \cdot \frac{Ste(p, p_i)}{Lte(p)} \right) + \tau \cdot \left(\alpha \cdot \frac{Sch(p, p_i)}{Lch(p_i)} + \beta \cdot \frac{Ste(p, p_i)}{Lte(p_i)} \right)$$

[0063] 其中， p 是要对照的语句，而 p_i 是被筛选的语句集中的其中一个， $Sch(p, p_i)$ 是两个语句 p 和 p_i 之间相同字的数目， $Lch(p)$ 是语句 p 的字数， $Ste(p, p_i)$ 是两个语句之间相同词的数目， $Lte(p)$ 是语句 p 的词数， α 和 β 为加权参数，满足 $\alpha + \beta = 1$ ；同样， σ 和 τ 为加权参数，满足 $\sigma + \tau = 1$ 。

[0064] 另外，确定使用符号“：“通过输入概念词确定概念时，由于用户不一定输入完整的概念描述，例如“王菲 : 歌手”，这种情况下也需要使用词语相似度计算来从“王菲”的多个概念中选取概念词“中国女歌手”。同样，符号“.”和“^”的应用，其关系词的输入也存在同样的问题，也需要通过词语相似度计算来进行知识匹配，其中每种语义逻辑对应的词语相似度计算时所用的调整参数不一定相同。

[0065] 总的来说，四个语义逻辑符号的使用中，都需要使用词语相似度将用户输入的词语与互动百科的知识相匹配。四个符号的处理逻辑分别见图 3-6。

[0066] 用户使用四个语义逻辑符号编辑的查询语句在处理过程中会遵守从左至右的优先处理原则。

[0067] 3、语义逻辑引导的搜索

[0068] 基于互动百科知识的语义逻辑引导的搜索流程见图 1，主要步骤如下：

[0069] 第一步，用户使用语义逻辑符号输入查询语句；例如输入查询语句“古天乐 . 电影 ~ 甜蜜女主角”。

[0070] 第二步，按从左至右的优先原则，根据每个逻辑符号的使用规则，解析查询语句中的逻辑符号。并根据原查询语句的顺序，依次将解析后的部分查询词加入到新的查询语句序列中。没有使用语义逻辑符号的部分直接按序加入。考虑到语义逻辑的处理结果会不止一个，将新的查询语句变成词语组的序列（可以二维数组表示）。例如“古天乐 . 电影 ~ 甜蜜

女主角”查询语句依次处理为：

[0071] 1) 处理第一个逻辑符号“.”后新的查询语句组变为：

[0072]

古天乐	电影	0 记重案实录
——	——	男儿当入樽
——	——	玩火
——	——	...

[0073] 2) 处理第二个逻辑符号“~”后变为：

[0074]

古天乐	电影	甜言蜜语
-----	----	------

[0075] 3) 处理完后新的查询语句组：

[0076]

古天乐	电影	甜言蜜语	女主角
-----	----	------	-----

[0077] 第三步,将处理得到的新查询语句组的第一行提取出来作为处理后的查询语句。例子中新的查询语句为“古天乐电影甜言蜜语女主角”。

[0078] 第四步,将新的查询语句提交给搜索引擎。

[0079] 第五步,将查询得到的网页结果返回给用户。同时,如果得到的查询语句组不只一行,则将除了第一行的其它行中非空列的数据推荐给用户。

[0080] 根据以上描述,基于 Google 搜索引擎实现了一个基于互动百科知识的语义逻辑引导的搜索方法,图 7、图 8 及图 9 显示了实施该方法后返回的结果。实施例 1 中,当用户不清楚电影名称,只隐约记得“甜蜜”一词时,通过相似逻辑处理可以直接得到其想要的电影名称“甜言蜜语”,从而搜索得到其想要的网页结果。如果没有该处理,返回的网页结果无法满足其需求。实施例 2 中,除了“醉拳”外,还给用户推荐了其他带“拳”的电影,可以提供用户更多信息。实施例 3 中,通过“歌手”确定概念“王菲〔中国女歌手〕”,虽然仅仅通过“王菲歌手”可以查询得到用户想要的结果,但是通过语义逻辑符号确定概念可以帮助消歧,尤其是在基于概念的搜索中会有很大帮助。

[0081] 综上所述,利用本发明可以一定程度上帮助用户按照其自然的语言逻辑构建查询语句,同时将其处理为以倒排和词语匹配为基础的搜索引擎能有效处理的查询语句,可以给用户推荐精确的信息,减少获取结果所需的搜索次数,甚至直接返回给用户所需的确精确信息,从而提高用户满意度。

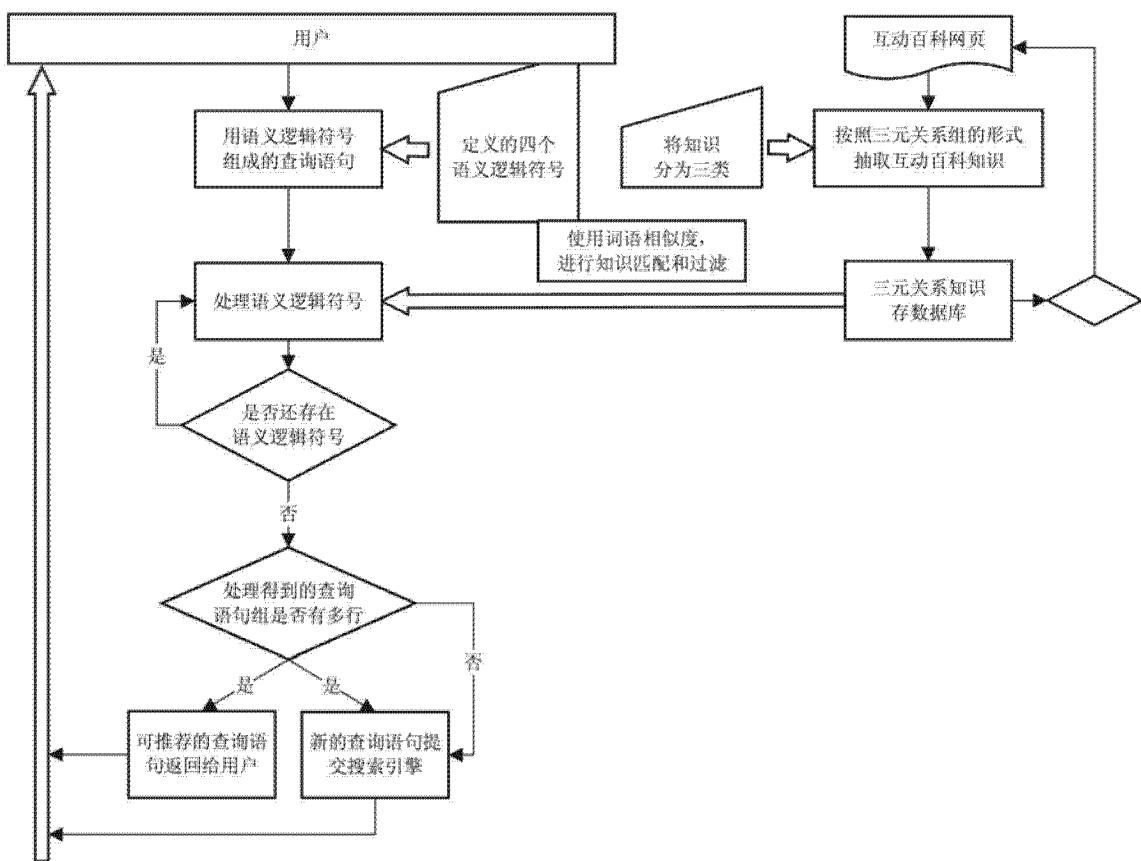


图 1

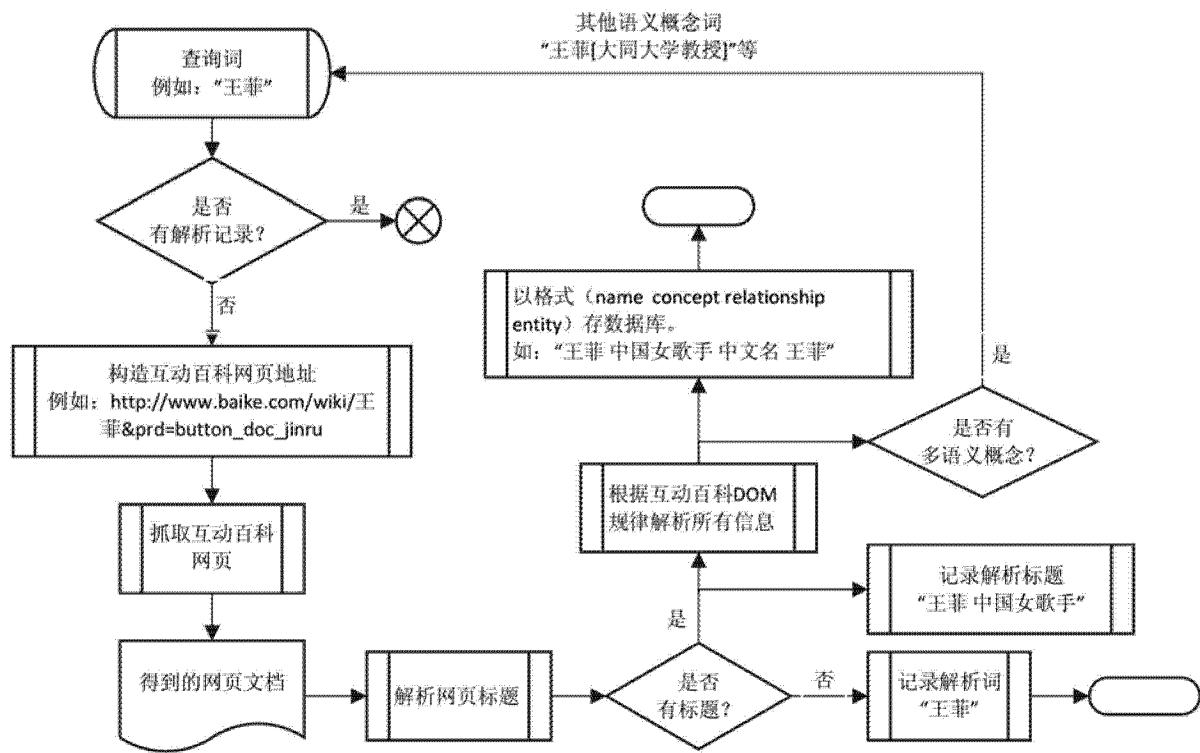


图 2

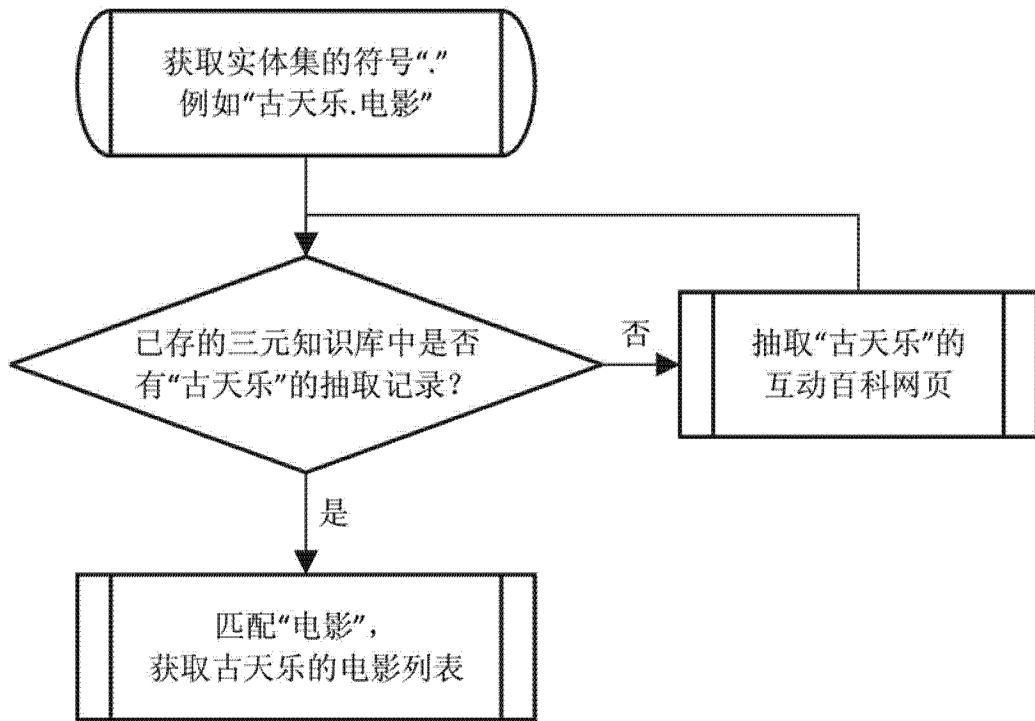


图 3

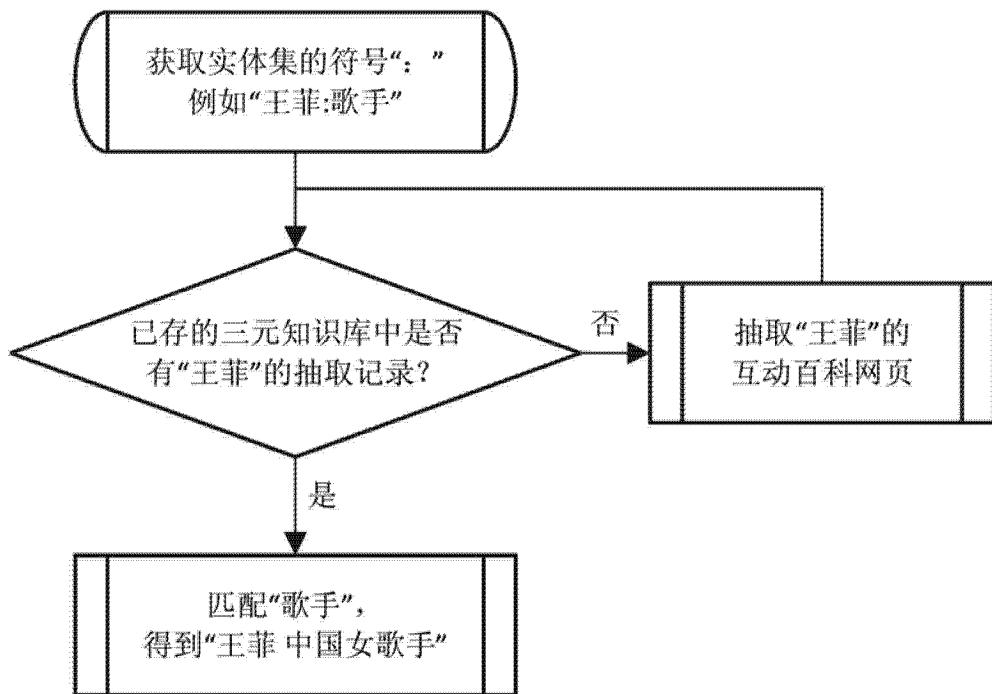


图 4

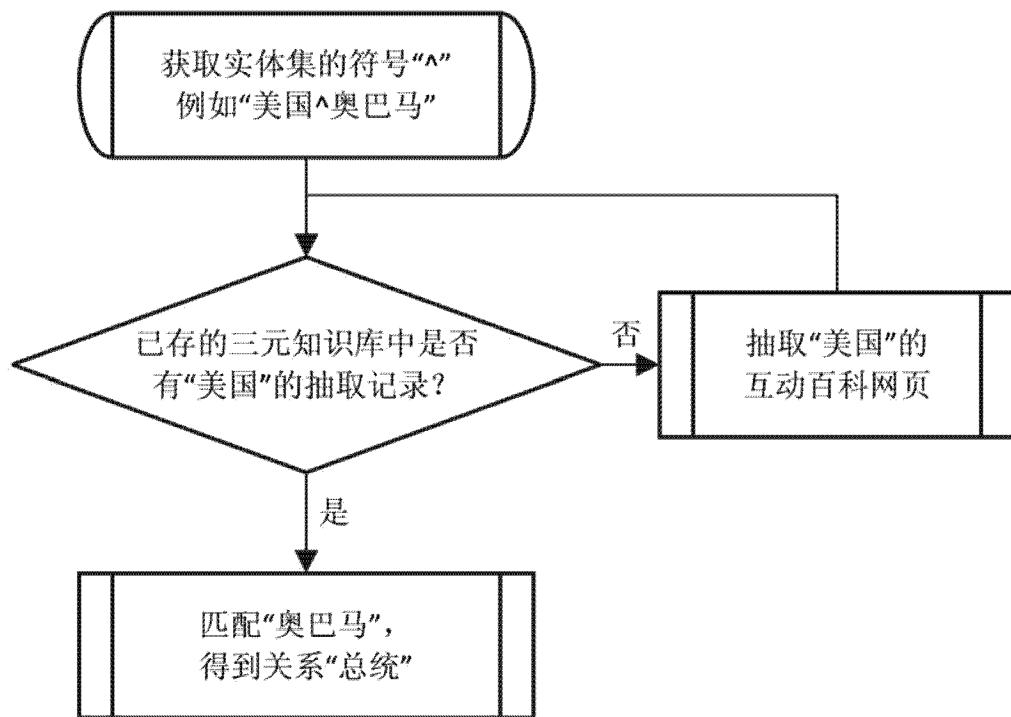


图 5

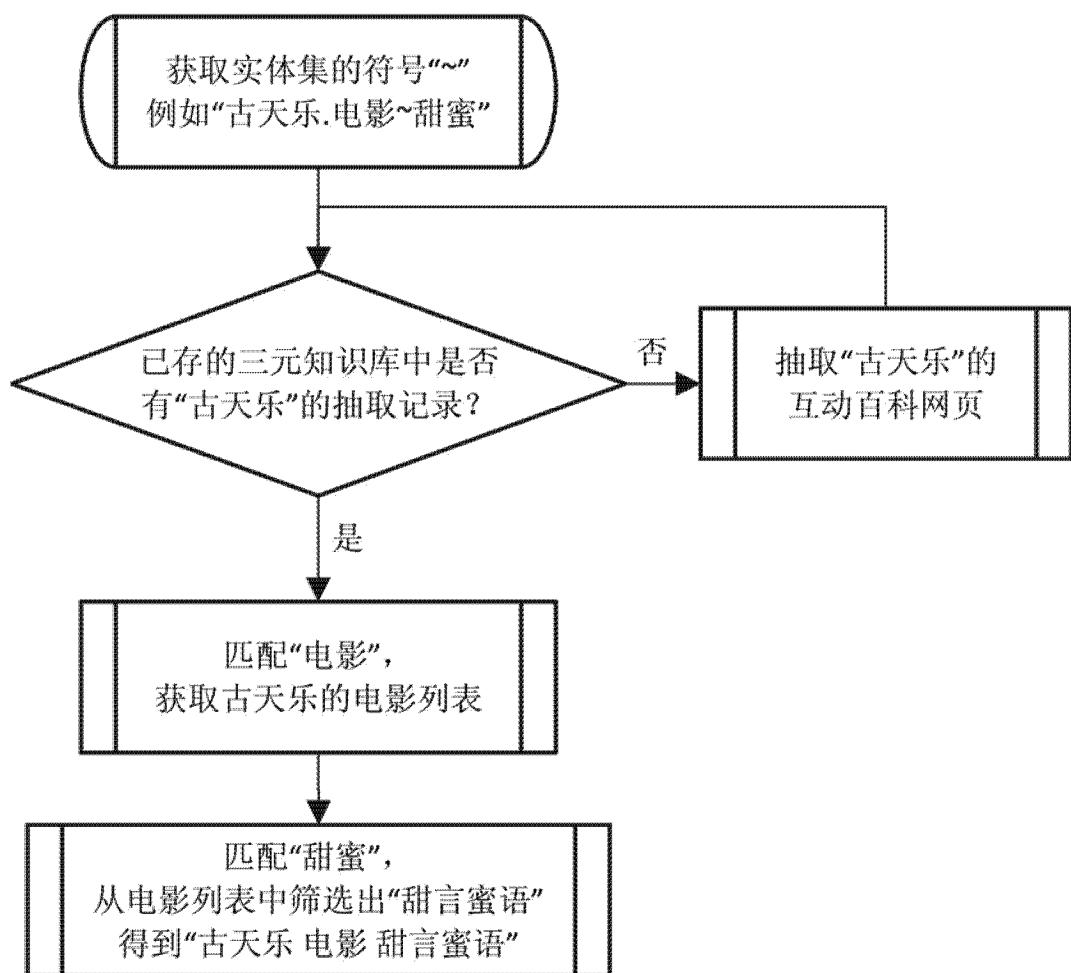


图 6

Logical Semantic Oriented

当前搜索的是：古天乐 电影~甜蜜 女主角

Google

当前搜索的是：古天乐 电影 甜蜜 女主角

古天乐电影(22部) (2012年2月12日更新7部) - 360Doc个人
2012年2月12日 ... 古天乐电影(22部) (2012年2月12日更新7部) ... 神奇侠侣；古天乐吴君如最佳主演的“魔”喜剧古天乐吴君如联袂主演的“碧”喜剧... 豆瓣评分5.0；导演:谷德昭 ...
<http://www.360doc.com/content...>

单身男女(单身男女影评) - 豆瓣电影
2011年2月31日 ... 单身男女(单身男女影评) - 豆瓣电影，我只能遗憾的对自己说：“你怎么能跟同性... 女主角的迷失，更加的突显出了古天乐对喜剧圈的一首甜蜜、浪漫的...
<http://movie.douban.com/review...>

古天乐-百度视频
128687条评论 ... 古天乐，ta的微博 ta的贴吧，演员/制片，别名：古仔；性别：男；生日：1970-10-21；出生地：... 李冰冰甜蜜牵手古天乐喜呼亲爱的，爱奇艺，20:47；杨幂古天乐罗志祥温情演绎百事微电影... 56, 01:17；曝古天乐暗恋新欢女主角遭男友怒吼。
http://v.baidu.com/person_intro/

生日快樂-维基百科，自由的百科全书
《生日快樂》(英文：Happy Birthday)是一部香港電影，由馬楚成執導，故事是改編自劉若英... 女主角：古天樂 (饰演小南)... 年香港影帝最佳女主角、2005年中國華語電影金雞獎最佳女主角及最近的2006年中國金雞百花獎最佳女主演... 他更憑《川島芳子》榮膺金馬獎以及金像獎最佳影后；《玻璃之城》及《甜蜜蜜》則分別獲得...
<http://zh.wikipedia.org/zh/%E2%80%9C生日快樂%E2%80%9D>

图 7

Logical Semantic Oriented

当前搜索的是：成龙 作品 拳精
您要搜索的是：成龙 作品 醉拳 search

Google

当前搜索的是：成龙 作品 季

醉拳-在线播放-《醉拳》-电影-优酷网，视频高清在线观看
醉拳广东武术界名人黄麒英自儿子黄飞鸿(成龙饰)小时侯起跟芦柏训练其习武；可惜黄飞鸿少年时生性轻浮；经常用三脚猫功夫在外惹事生非，武学上一直无大的...
http://v.youku.com/v_showid_X...

醉拳I成龙作品-在线播放-优酷网，视频高清在线观看
醉拳1成龙作品太励志了... 简介：醉拳1成龙作品。00:00, 00:00/00:00, 屏幕, 报告, 老板！第5集；小四代, 10:37, 《报告老板！》第5集；小四代, 报告老板, 9,319,423 ...
http://v.youku.com/v_showid_X...

醉拳I(成龙)-在线观看-PPS爱频道-PPS网络电视
醉拳I(成龙)-国语描述广东于虎之一黄飞鸿的少年时代,一个不屈不挠,终获成功的江湖 小人物。
http://v.pps.tv/play_34NR35.html

6.2.2. 醉拳-在线观看-360影视
电影《拳精》在线观看，... 你或许也喜欢该片主演作品该片导演作品、《警察故事2013》海报 2013年6.0分 - 警察故事2013、成龙/动作，《一招半式闯江湖》海报 1980年6.5 ...
<http://v.360.cn/mgkPqYUuQh...>

图 8

Logical Semantic Oriented

当前搜索的是：王菲 中国女歌手

Google

当前搜索的是：王菲 歌手

【王菲】最新王菲好听的歌曲 MV 资料图片 新歌 王菲空...
王菲简介：王菲，中国著名女歌手、影视演员，其演唱的《红豆》《但愿人长久》等歌曲至今依然拥有广泛的传唱度... 歌曲名称：专辑；歌手；审批文号；试听；分享；下载。
<http://www.kuwo.cn/mingxing/>

王菲-维基百科，自由的百科全书
王菲 (英文名: Faye Wong, 1969年8月8日 -)，曾用艺名王靖雯 (英文名: Shirley Wong)，华语著名女歌手、演员。王菲出生于中国北京市，1987年随父亲举家移居 ...
<http://zh.wikipedia.org/zh/%E2%80%9C王菲%E2%80%9D>

王菲-最新王菲好听的歌曲 MV 资料图片 新歌 王菲空...
王菲简介：王菲，中国著名女歌手、影视演员，其演唱的《红豆》《但愿人长久》等歌曲至今依然拥有广泛的传唱度... 歌曲名称：专辑；歌手；审批文号；试听；分享；下载。
<http://www.kuwo.cn/mingxing/>

王菲歌曲mp3试听下载 王菲最新专辑 王菲下载-搜狗音乐
搜狗音乐大本营王菲为您提供：王菲最新歌曲mp3试听下载，王菲最新专辑、王菲 演唱 ... 歌手介绍：姓名：王菲出生日期：1969年8月8日出生地点：北京地区：中国香港 ... 音乐、风...
<http://music.sogou.com/singer/>

传李亚鹏反对麦当劳当歌手王菲见法王助女圆梦-搜狐娱乐
2013年7月13日 ... 王菲丈夫李亚鹏已将歌唱事业放一边，专心修佛及处理嫣然基金慈善工作，虽然令不少乐迷大感失望。麦当劳之后继有入，现年十六岁，跟...
<http://music.sogou.com/singer/>

图 9