

(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 104820713 A

(43) 申请公布日 2015. 08. 05

(21) 申请号 201510256349. 6

(22) 申请日 2015. 05. 19

(71) 申请人 苏州工讯科技有限公司

地址 215400 江苏省苏州市太仓市娄东街道
北京东路 88 号东 P 楼

(72) 发明人 张晶晶

(74) 专利代理机构 苏州慧通知识产权代理事务
所（普通合伙）32239

代理人 安纪平

(51) Int. Cl.

G06F 17/30(2006. 01)

G06F 17/27(2006. 01)

权利要求书1页 说明书4页

(54) 发明名称

一种基于用户历史数据获得工业产品名称同义词的方法

(57) 摘要

本发明涉及一种基于用户历史数据获得工业产品名称同义词的方法，通过对用户使用工业产品搜索引擎产生的历史数据进行分析，挖掘具有相同含义的工业产品名称，包括以下几步：对工业产品搜索词条进行分词；对工业产品名称意图挖掘；对工业产品名称同义词匹配，获得工业产品名称同义词。本发明的基于用户历史数据获得工业产品名称同义词的方法，该方法构建同义词库的覆盖范围广，不易出现遗漏，并且该方法是基于用户数据挖掘工业产品名称的同义词，数据基数大。

1. 一种基于用户历史数据获得工业产品名称同义词的方法,其特征在于,通过对用户使用工业产品搜索引擎产生的历史数据进行分析,挖掘具有相同含义的工业产品名称,包括以下几步:

第一步:对工业产品搜索词条进行分词;

第二步:对工业产品名称意图挖掘;

第三步:对工业产品名称同义词匹配,获得工业产品名称同义词。

2. 根据权利要求 1 所述的一种基于用户历史数据获得工业产品名称同义词的方法,其特征在于,第一步中对工业产品搜索词条进行切割,将所述搜索词条切割成至少一个汉语单词,汉语单词中包含工业产品名称。

3. 根据权利要求 2 所述的一种基于用户历史数据获得工业产品名称同义词的方法,其特征在于,第二步中对工业产品名称意图进行挖掘,包括以下几步:

首先,计算工业产品名称的搜索倾向:通过用户历史数据,计算每种工业产品名称的每个被点击的搜索结果和相应被点击的次数,对于任意一个工业产品名称 W 与任意一个搜索结果 D , 以 $\text{Count}(W, D)$ 表示通过搜索包含 W 的词条而点击 D 的次数;对于任意一个搜索词条 Q 与任意一个搜索结果 D , 以 $\text{Count}(Q, D)$ 表示所有用户搜索词条 Q 点击结果 D 的次数总和;

其次,对工业产品的搜索倾向进行表征:对于任意一个搜索名词 W_j , 其对于每个搜索结果的搜索倾向: $\text{Count}(W_j, D_k)$, 对于 W_j 做如下处理:

去除 $\text{Count}(W_j, D_k) = 0$ 的文档 D_k , 只保留 $\text{Count}(W_j, D_k) \neq 0$ 的文档 D_k ;

将所有保留的 D_k 按照 $\text{Count}(W_j, D_k)$ 进行降序排序,取排名靠前 10% 的文档 $D_k (k = 1, 2, 3, \dots, N)$ 组成的集合为名称 W_j 的意图表征。

4. 根据权利要求 3 所述的一种基于用户历史数据获得工业产品名称同义词的方法,其特征在于, $\text{Count}(W, D) = \text{SUM}(\text{Count}(Q_i, D))$, 其中, Q_i 表示所有通过分词后包含 W 的搜索词条, i 为自然数, SUM 为求和函数。

5. 根据权利要求 4 所述的一种基于用户历史数据获得工业产品名称同义词的方法,其特征在于,两个工业产品的意图表征相同,则两个工业产品名称互为同义词。

6. 根据权利要求 1 所述的一种基于用户历史数据获得工业产品名称同义词的方法,其特征在于,所述历史数据包括用户输入的搜索词条和用户在搜索该词条后,在搜索结果里点击的结果。

7. 根据权利要求 1 所述的一种基于用户历史数据获得工业产品名称同义词的方法,其特征在于,用户在工业产品搜索平台输入搜索词条发起搜索请求,工业产品搜索引擎直接搜索用户输入的搜索词条,工业产品搜索引擎还降搜索词条中的工业产品名称替换为其同义词,并且重新向工业产品搜索引擎发起搜索请求。

一种基于用户历史数据获得工业产品名称同义词的方法

技术领域

[0001] 本发明属于 B2B 领域,尤其是基于同义词搜索方法,具体涉及一种基于用户历史数据获得工业产品名称同义词的方法。

背景技术

[0002] B2B 是指企业对企业之间的营销关系,将企业内部网通过 B2B 网站与客户紧密结合起来,通过网络的快速反应,为客户提供更好的服务,从而促进企业的业务发展。

[0003] 在 B2B 领域中,一项核心技术为“基于互联网的工业产品搜索功能”,该功能为企业之间的产品贸易提供枢纽和入口。产品需求方为购买所需的工业产品,首先该企业需要在 B2B 互联网平台上,通过工业产品的搜索功能在互联网上搜索到其所需的工业产品,以获得其所需要的产品信息(供给方、价格、物流等等)。然后,在线下单订购,并开展后期的线下贸易行为。

[0004] 然而,工业产品 B2B 平台的搜索技术面临一项重要的实际使用问题,即:搜索用户常常无法准确输入其所期望的产品名称,或搜索用户输入的搜索词条与数据库中存储的工业产品名称不匹配。原因在于:(1)工业产品种类极为繁杂,工业产品的名称经常极为生僻、拗口,难以记忆和掌握;(2)用户素质参差不齐、行业背景不同,导致用户往往不具有足够的相关知识,无法准确输入其所需的工业产品名称;(3)由于工业产品往往具有很多“别名”,搜索用户所输入的产品名称可能与数据库内存储的工业产品名称不一致,即使两种名称所表示同一种工业产品。

[0005] 例如,“工业酒精”又称“变性酒精”,或“工业火酒”。如果供应商 A 发布的供货名称为“工业火酒”,而需求企业 B 由于先验知识不足,仅仅搜索了“工业酒精”,在传统的字符串搜索平台中,B 将无法直接搜索到 A 所发布的供货信息。可见,在用户输入的搜索词条不够准确时,用户往往无法搜索到其所需要的产品结果。

[0006] 针对工业产品搜索的这一问题,现有技术主要为“基于同义词搜索的解决方案”,主要分为两类:基于工业产品搜索服务提供商的方法和基于工业产品发布者的方法。

[0007] 第一类为基于工业产品搜索服务提供商的方法,即搜索服务提供商(B2B 平台搜索技术人员)在搜索引擎平台内部规定每种工业产品名称的同义词,即搜索服务提供商首先利用先验知识,构建工业产品同义词库。在工业产品同义词库中,规定了每种工业产品名称的同义词列表,例如上面例子中的:工业酒精=变性酒精=工业火酒。在用户搜索某一工业产品名称时,搜索引擎将对其同义词列表中的每个名词发起搜索,并将所有搜索结果进行整合,统一返回给搜索者。

[0008] 第二类方案为基于工业产品发布者的方法。即工业产品发布者在其发布的产品名称中(或数据库的其他字段中,或网页的其他部分中)罗列多个产品名称(SEO 技术)。例如,将发布的产品名称从“工业酒精”改为“工业酒精变性酒精工业火酒”,以提高其被搜索到的概率。

[0009] 现在技术中“工业产品名称同义词识别”技术主要具有如下缺点:

[0010] 1. 容易出现漏识别,由于工业产品名称过于冗繁、数目庞大,构建同义词库的过程极为耗时耗力,效率低下;切构建同义词库覆盖所有同义词的可能性较低,往往会出现漏识别的情况。2. 第一类方案对人力消耗大,构建时间长,不易与技术升级,时效性差。3. 第二类方案会破坏工业产品搜索结果美感,降低了结果可读性。4. 第二类方案对产品发布者技能要求高,不利于搜索公平性。

发明内容

[0011] 为解决上述技术问题,本发明提供了一种基于用户历史数据获得工业产品名称同义词的方法,该方法不易出现漏识别,构建同义词库的过程效率高,并且容易构建同义词库,时效性好。

[0012] 为达到上述目的,本发明的技术方案如下:

[0013] 一种基于用户历史数据获得工业产品名称同义词的方法,通过对用户使用工业产品搜索引擎产生的历史数据进行分析,挖掘具有相同含义的工业产品名称,包括以下几步:

[0014] 第一步:对工业产品搜索词条进行分词;

[0015] 第二步:对工业产品名称意图挖掘;

[0016] 第三步:对工业产品名称同义词匹配,获得工业产品名称同义词。

[0017] 本发明的一个较佳实施例中,进一步包括,第一步中对工业产品搜索词条进行切割,将所述搜索词条切割成至少一个汉语单词,汉语单词中包含工业产品名称。

[0018] 本发明的一个较佳实施例中,进一步包括,第二步中对工业产品名称意图进行挖掘,包括以下几步:

[0019] 算工业产品名称的搜索倾向:通过用户历史数据,计算每种工业产品名称的每个被点击的搜索结果和相应被点击的次数,对于任意一个工业产品名称 W 与任意一个搜索结果 D ,以 $\text{Count}(W, D)$ 表示通过搜索包含 W 的词条而点击 D 的次数;对于任意一个搜索词条 Q 与任意一个搜索结果 D ,以 $\text{Count}(Q, D)$ 表示所有用户搜索词条 Q 点击结果 D 的次数总和;

[0020] 其次,对工业产品的搜索倾向进行表征:对于任意一个搜索名词 W_j ,其对于每个搜索结果的搜索倾向: $\text{Count}(W_j, D_k)$,对于 W_j 做如下处理:

[0021] 去除 $\text{Count}(W_j, D_k) = 0$ 的文档 D_k ,只保留 $\text{Count}(W_j, D_k) \neq 0$ 的文档 D_k ;

[0022] 将所有保留的 D_k 按照 $\text{Count}(W_j, D_k)$ 进行降序排序,取排名靠前 10% 的文档 D_k ($k = 1, 2, 3, \dots, N$) 组成的集合为名称 W_j 的意图表征。

[0023] 本发明的一个较佳实施例中,进一步包括, $\text{Count}(W, D) = \text{SUM}(\text{Count}(Q_i, D))$,其中, Q_i 表示所有通过分词后包含 W 的搜索词条, i 为自然数, SUM 为求和函数。

[0024] 本发明的一个较佳实施例中,进一步包括,两个工业产品的意图表征相同,则两个工业产品名称互为同义词。

[0025] 本发明的一个较佳实施例中,进一步包括,所述历史数据包括用户输入的搜索词条和用户在搜索该词条后,在搜索结果里点击的结果。

[0026] 本发明的一个较佳实施例中,进一步包括,用户在工业产品搜索平台输入搜索词条发起搜索请求,工业产品搜索引擎直接搜索用户输入的搜索词条,工业产品搜索引擎还降搜索词条中的工业产品名称替换为其同义词,并且重新向工业产品搜索引擎发起搜索请

求。

[0027] 本发明的有益效果是：

[0028] 其一、本发明的基于用户历史数据获得工业产品名称同义词的方法，该方法构建同义词库的覆盖范围广，不易出现遗漏，并且该方法是基于用户数据挖掘工业产品名称的同义词，数据基数大。

[0029] 其二、本发明的方法数据来源于用户的真实操作行为，数据反映了用户真正的搜索意图。

[0030] 其三、本发明的方法缓解了用户先前经验知识不足的缺陷。

具体实施方式

[0031] 下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例仅仅是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其他实施例，都属于本发明保护的范围。

[0032] 实施例

[0033] 本实施例中公开了一种基于用户历史数据获得工业产品名称同义词的方法，通过对用户使用工业产品搜索引擎产生的历史数据进行分析，挖掘具有相同含义的工业产品名称，所述历史数据包括用户输入的搜索词条和用户在搜索该词条后，在搜索结果里点击的结果。

[0034] 包括以下几步：

[0035] 第一步：对工业产品搜索词条进行分词；

[0036] 第二步：对工业产品名称意图挖掘；

[0037] 第三步：对工业产品名称同义词匹配，获得工业产品名称同义词。

[0038] 具体的，第一步中对工业产品搜索词条进行切割，将所述搜索词条切割成至少一个汉语单词，汉语单词中包含工业产品名称。搜索词条为搜索用户输入的原始文字，而“工业产品名称”为某种工业产品的名称。例如，如果用户搜索“工业酒精如何购买”，则“工业酒精如何购买”为其“搜索词条”，对该词条进行分词，得到“工业酒精”、“如何”、“购买”，三个汉语单词，其中“工业酒精”为工业产品名称。

[0039] 本发明的一个较佳实施例中，进一步包括，第二步中对工业产品名称意图进行挖掘，包括以下几步：

[0040] 首先，计算工业产品名称的搜索倾向：通过用户历史数据，计算每种工业产品名称的每个被点击的搜索结果和相应次数，对于任意一个工业产品名称 W 与任意一个搜索结果 D ，以 $\text{Count}(W, D)$ 表示通过搜索包含 W 的词条而点击 D 的次数；对于任意一个搜索词条 Q 与任意一个搜索结果 D ，以 $\text{Count}(Q, D)$ 表示通过搜索词条 Q 而点击 D 的次数；用户历史数据包含了 Q 与 D 的一一对应关系，计算 $Q-D$ 对即可以得到 $\text{Count}(Q, D)$ 。而计算 $\text{Count}(W, D)$ 的方法为： $\text{Count}(W, D) = \text{SUM}(\text{Count}(Q_i, D))$ 。其中， Q_i 表示所有通过分词后包含 W 的搜索词条， i 为自然数， SUM 为求和函数。

[0041] 其次，对工业产品的搜索倾向进行表征：对于任意一个搜索名词 W_j ，其对于每个搜索结果的搜索倾向： $\text{Count}(W_j, D_k)$ ，对于 W_j 做如下处理：

- [0042] 去除 $\text{Count}(W_j, D_k) = 0$ 的文档 D_k , 只保留 $\text{Count}(W_j, D_k) \neq 0$ 的文档 D_k ;
- [0043] 将所有保留的 D_k 按照 $\text{Count}(W_j, D_k)$ 进行降序排序, 取排名靠前 10% 的文档 D_k ($k = 1, 2, 3, \dots, N$) 组成的集合为名称 W_j 的意图表征。
- [0044] 如果, 任意两个工业产品的意图表征相同, 则两个工业产品名称互为同义词。
- [0045] 用户在工业产品搜索平台输入搜索词条发起搜索请求, 工业产品搜索引擎直接搜索用户输入的搜索词条, 工业产品搜索引擎还降搜索词条中的工业产品名称替换为其同义词, 并且重新向工业产品搜索引擎发起搜索请求。
- [0046] 以本实施例上述公开的基于用户历史数据获得工业产品名称同义词的方法, 其过程如下:
- [0047] 1. 在工业产品搜索引擎中, 记录用户输入的每一条搜索词条, 同时记录其随后点击的每一个搜索结果, 并积累该数据一段时间, 形成 100 万条以上的数据源。
- [0048] 2. 获得“搜索词条”-“被点击的搜索结果”的对应关系, 对数据中的“搜索词条”进行分词, 得到“工业产品名称”-“被点击的搜索结果”的对应关系。
- [0049] 3. 对于每个“工业产品名称”, 抛弃其未点击的搜索结果, 计算被点击的搜索结果的点击次数, 并取出点击次数最多的 10% 的点击结果, 以该“结果集合”作为该“工业产品名称”的意图表征。
- [0050] 4. 将意图表征相同的“工业产品名称”归集在一起, 被归集在一起的“工业产品名称”互为同义词。
- [0051] 5. 基于该同义词关系, 用户在工业产品搜索平台中发起搜索请求时, 引擎不但直接搜索用户输入的搜索词条, 同时, 引擎还将词条中的工业产品名称替换为其任意的同义词, 并重新发起搜索请求。
- [0052] 本发明的基于用户历史数据获得工业产品名称同义词的方法, 该方法构建同义词库的覆盖范围广, 不易出现遗漏, 并且该方法是基于用户数据挖掘工业产品名称的同义词, 数据基数大; 数据来源于用户的真实操作行为, 数据反映了用户真正的搜索意图。
- [0053] 用户在工业产品搜索平台输入搜索词条发起搜索请求, 工业产品搜索引擎直接搜索用户输入的搜索词条, 工业产品搜索引擎还降搜索词条中的工业产品名称替换为其同义词, 并且重新向工业产品搜索引擎发起搜索请求, 缓解了用户先前经验知识不足的缺陷。
- [0054] 对所公开的实施例的上述说明, 使本领域专业技术人员能够实现或使用本发明。对这些实施例的多种修改对本领域的专业技术人员来说将是显而易见的, 本文中所定义的一般原理可以在不脱离本发明的精神或范围的情况下, 在其它实施例中实现。因此, 本发明将不会被限制于本文所示的这些实施例, 而是要符合与本文所公开的原理和新颖特点相一致的最宽的范围。