

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局

(43) 国际公布日
2012 年 8 月 23 日 (23.08.2012)



(10) 国际公布号
WO 2012/109959 A1

- (51) 国际专利分类号:
G06F 17/30 (2006.01)
- (21) 国际申请号: PCT/CN2012/070824
- (22) 国际申请日: 2012 年 2 月 1 日 (01.02.2012)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
201110043030.7 2011 年 2 月 18 日 (18.02.2011) CN
- (71) 申请人 (对除美国外的所有指定国): **腾讯科技 (深圳) 有限公司 (TENCENT TECHNOLOGY (SHENZHEN) COMPANY LIMITED)** [CN/CN]; 中国广东省深圳市福田区振兴路赛格科技园 2 栋东 4 楼, Guangdong 518044 (CN)。
- (72) 发明人; 及
- (75) 发明人/申请人 (仅对美国): **赫南 (HE, Nan)** [CN/CN]; 中国广东省深圳市福田区振兴路赛格科技园 2 栋东 4 楼, Guangdong 518044 (CN)。 **王迪 (WANG, Di)** [CN/CN]; 中国广东省深圳市福田区振兴路赛格科技园 2 栋东 4 楼, Guangdong 518044 (CN)。 **郭阳 (GUO, Yang)** [CN/CN]; 中国广东省深圳市福田区振兴路赛格科技园 2 栋东 4 楼, Guangdong 518044 (CN)。 **胡立新 (HU, Lixin)** [CN/CN]; 中

国广东省深圳市福田区振兴路赛格科技园 2 栋东 4 楼, Guangdong 518044 (CN)。 **王艳敏 (WANG, Yan-min)** [CN/CN]; 中国广东省深圳市福田区振兴路赛格科技园 2 栋东 4 楼, Guangdong 518044 (CN)。 **朱建朋 (ZHU, Jianpeng)** [CN/CN]; 中国广东省深圳市福田区振兴路赛格科技园 2 栋东 4 楼, Guangdong 518044 (CN)。

(74) 代理人: **北京德琦知识产权代理有限公司 (DEQI INTELLECTUAL PROPERTY LAW CORPORATION)**; 中国北京市海淀区知春路 1 号学院国际大厦 7 层, Beijing 100083 (CN)。

(81) 指定国 (除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。

(84) 指定国 (除另有指明, 要求每一种可提供的地区保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ,

[见续页]

(54) Title: CLUSTERING METHOD AND DEVICE FOR SEARCH TERMS

(54) 发明名称: 检索词的聚类方法和装置

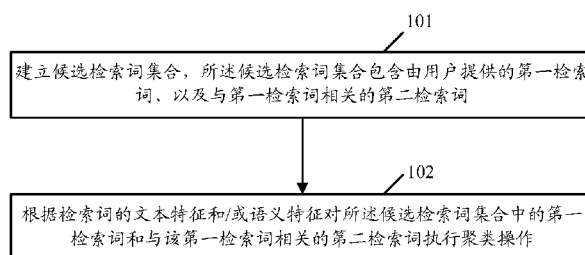


图 1 / Fig. 1

101 ESTABLISH A CANDIDATE SEARCH TERM SET CONTAINING A FIRST SEARCH TERM PROVIDED BY A USER AND A SECOND SEARCH TERM RELATED TO THE FIRST SEARCH TERM
102 CLUSTER THE FIRST SEARCH TERM IN THE CANDIDATE SEARCH TERM SET AND THE SECOND SEARCH TERM RELATED TO THE FIRST SEARCH TERM ACCORDING TO THE TEXT FEATURES AND/OR SEMANTIC FEATURES OF THE SEARCH TERMS

(57) Abstract: Provided are a clustering method and device for search terms, wherein the method includes: A. establishing a candidate search term set containing a search term provided by a user and a search term related to the search term provided by the user; and B. clustering the search terms in the candidate search term set according to the text features and/or semantic features of the search terms. The application of the present invention can improve the accuracy and relevance of search term clustering.

[见续页]



WO 2012/109959 A1



BY, KG, KZ, MD, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG)。

本国际公布:

— 包括国际检索报告(条约第 21 条(3))。

(57) 摘要:

本发明提供了检索词的聚类方法和装置, 其中, 该方法包括: A, 建立候选检索词集合, 所述候选检索词集合包含由用户提供的检索词、以及与该用户提供的检索词相关的检索词; B, 根据检索词的文本特征和/或语义特征对所述候选检索词集合中的检索词执行聚类操作。采用本发明, 能够提高检索词聚类的准确度和相关度。

检索词的聚类方法和装置

本申请要求于 2011 年 02 月 18 日提交中国专利局、申请号为
201110043030.7、发明名称为“检索词的聚类方法和装置”的中国专利
5 申请的优先权，其全部内容通过引用结合在本申请中。

技术领域

本发明涉及网络搜索技术，特别涉及检索词的聚类方法和装置。

发明背景

在网络搜索技术中，用户都是通过检索词搜索到相应的结果。应用
10 于竞价广告系统中，该检索词具体实现时可为广告商提供的广告的标识，也可称为购买词，目的是便于用户通过该检索词搜索到相应的广告。

在竞价广告系统中，为了提高广告展现质量，需要对广告商提供的检索词进行聚类。其中，对检索词进行聚类的过程可以抽象为对一个短文本串的集合进行聚类的过程。

15 目前，最常用的聚类方法为：针对一广告商提供的检索词，仅从目前已存在的所有广告商提供的检索词中找出字面上相似度较大的检索词，将该广告商提供的检索词和找出的检索词聚类在一起。如此，当搜索引擎用户通过一检索词检索相应的广告时，将与该检索词对应的广告，以及与该检索词聚类在一起的检索词对应的广告展示给用户。

20 然而，有一些检索词，尽管广告商没有提供，但其实质上与广告商提供的检索词对应的广告相关，而前述的聚类方法是仅将广告商提供的检索词进行字面相关的聚类，没有考虑到这些与广告商提供的检索词语义相关的、且目前还未被广告商提供的其他检索词，这降低了检索词聚类的准确度。

发明内容

本发明提供了检索词的聚类方法和装置，以提高检索词聚类的准确度和相关度。

本发明提供的技术方案包括：

5 一种检索词的聚类方法，包括：

建立候选检索词集合，所述候选检索词集合包含由用户提供的第一检索词、以及与第一检索词相关的第二检索词；

根据检索词的文本特征和/或语义特征对所述候选检索词集合中的第一检索词和与该第一检索词相关的第二检索词执行聚类操作。

10 一种检索词的聚类装置，包括：

建立单元，用于建立候选检索词集合，所述候选检索词集合包含由用户提供的第一检索词、以及与第一检索词相关的第二检索词；

15 聚类单元，用于根据检索词的文本特征和/或语义特征对所述候选检索词集合中的第一检索词和与该第一检索词相关的第二检索词执行聚类操作。

由以上技术方案可以看出，本发明提供的检索词的聚类方法和装置在进行检索词聚类时，并非像现有技术那样仅将用户提供的检索词进行字面关系的聚类，而是同时考虑到用户提供的检索词，以及与该用户提供的检索词相关的其他检索词，并根据检索词的文本特征和/或语义特征
20 对用户提供的检索词，以及与该用户提供的检索词相关的其他检索词进行聚类，因而能够增加检索词聚类的准确度和相关度。

附图简要说明

图 1 为本发明实施例提供的基本流程图；

图 2a 为本发明实施例提供的步骤 102 的流程图；

图 2b 为本发明实施例提供的潜在聚类关系挖掘流程图；

图 3a 为本发明实施例提供的检索词之间的拓扑图结构第一示意图；

图 3b 为本发明实施例提供的检索词之间的拓扑图结构第二示意图；

图 3c 为本发明实施例提供的检索词之间潜在的聚类关系示意图；

5 图 3d 为本发明实施例提供的增加检索词时拓扑图结构第三示意图；

图 4 为本发明实施例提供的新增加检索词时的流程图；

图 5 为本发明实施例提供的装置的基本结构图；

图 6 为本发明实施例提供的装置的详细结构图。

实施本发明的方式

10 为了使本发明的目的、技术方案和优点更加清楚，下面结合附图和具体实施例对本发明进行详细描述。

本发明在进行检索词聚类时，并非像现有技术那样仅将用户比如广告商提供的检索词进行字面关系的聚类，而是根据检索词的文本特征和/或语义特征将用户提供的检索词，以及与该检索词相关的检索词聚类，
15 以便增加检索词聚类的准确度，下面对本发明提供的方法进行描述。

参见图 1，图 1 为本发明实施例提供的基本流程图。如图 1 所示，该流程可包括以下步骤：

步骤 101，建立候选检索词集合，所述候选检索词集合包含由用户提供的第二检索词、以及与第二检索词相关的第三检索词。

20 本步骤 101 中，与用户提供的第二检索词相关的第三检索词具体可包括以下两种方式或任一方式确定：方式 1，确定与该用户提供的第二检索词匹配的检索词，将该确定的检索词确定为与用户提供的第二检索词相关的第三检索词；方式 2，以该用户提供的第二检索词为关键词搜索，将搜索结果中的检索词确定为与用户提供的第二检索词相关的第三

检索词。

其中，方式 1 得到的检索词，具体可为：通过该用户提供的
第一检索词进行简单的字符串变换处理所得到的检索词；或者根据实际经验
确定出的与第一检索词经常在一起使用的检索词，比如，如果用户提供
5 的第一检索词为咖啡壶，则根据经验可以知道咖啡壶通常与咖啡杯等经
常使用，基于此，可确定与该用户提供的咖啡壶匹配的检索词可为咖啡
杯等。

其中，方式 2 得到的检索词，具体可为：以用户提供的
第一检索词为关键词进行搜索，所得到的搜索结果中的检索词。其中，该搜索具体
10 可通过用户搜索串与广告检索词映射整合系统（QBM: Query Bidterm
Merge）实现，其中，QBM 具体实现时可为：以用户提供的
第一检索词作为输入进行搜索，从搜索到的搜索结果中获取检索词，
将该获取的检索词作为与该用户提供的
第一检索词相关的检索词。

至此，通过步骤 101 即可得到候选检索词集合。需要说明的是，本
15 实施例需要保证步骤 101 得到的候选检索词集合中没有重复的检索词。

步骤 102，根据检索词的文本特征和/或语义特征对所述候选检索词
集合中的第一检索词和与该第一检索词相关的第二检索词执行聚类操
作。

本步骤 102 具体实现时，可根据第一检索词的文本特征和/或语义特
20 征计算该第一检索词和所述候选检索词集合中与该第一检索词相关的
第二检索词之间的相似度值，将该第一检索词和与该第一检索词具有较
高相似度值的第二检索词聚类在一起。具体地，该步骤 102 可通过图 2a
所示的流程体现。

参见图 2a，图 2a 为本发明实施例提供的步骤 102 的流程图。该流
25 程示出了基本聚类关系的具体实现原理，如图 2a 所示，该流程可包括以

下步骤:

步骤 201a, 根据第一检索词的文本特征和/或语义特征计算该第一检索词和其相关的每一个第二检索词之间的相似度值。

步骤 202a, 如果该第一检索词和第二检索词之间的相似度值大于或
5 等于第一预设阈值, 则将该第一检索词和该第二检索词聚类在一起。

通过步骤 202a, 能够将第一检索词和其相关的、且与该第一检索词之间的相似度值大于或等于第一预设阈值的第二检索词聚类在一起, 即实现了本发明实施例的基本聚类。

优选地, 为保证更加完整的聚类关系, 本实施例还提供了潜在聚类
10 关系的挖掘流程, 具体可通过图 2b 所示的流程体现。

参见图 2b, 图 2b 为本发明实施例提供的潜在聚类关系挖掘流程图。
如图 2b 所示, 该流程可包括以下步骤:

步骤 201b, 从与第一检索词相关的各个第二检索词中选取与该第一检索词之间的相似度值大于或等于第二预设阈值的第二检索词。

15 作为本发明实施例的一种扩展, 为降低潜在聚类关系挖掘的复杂度, 本步骤 201b 还可替换为: 从与第一检索词聚类在一起的各个第二检索词中选取与该第一检索词之间的相似度值大于或等于第二预设阈值的第二检索词。

其中, 本步骤 201b 中的第二预设阈值与步骤 202a 中的第一预设阈
20 值无关, 两者可相等, 也可不等。

步骤 202b, 计算该选取的任意两个第二检索词之间的相似度值, 如果该计算的相似度值大于或等于所述第一预设阈值, 则将该两个第二检索词聚类在一起。

通过步骤 201b 至步骤 202b, 能够实现潜在聚类关系的挖掘。

25 如此, 本发明实施例将步骤 202a 中聚类在一起的第一检索词和第二

检索词（即该第一检索词和第二检索词之间具有聚类关系），以及步骤 202b 中聚类在一起的第二检索词合并在一起，即可形成了本发明实施例的全量聚类结果。优选地，本实施例中，步骤 202a 的聚类 and 步骤 202b 的聚类均可按照类似现有的机器学习模型实现，这里并不具体限定。

5 为使图 2 所示的流程更加清楚，下面通过一个具体实施例对本发明提供的流程进行描述。

假如用户提供的第二检索词分别为 b1, b3、b4 和 b5，其中，通过步骤 101，可以得到：与 b1 相关的第二检索词为 b2, b3 和 b4，与 b3 相关的第二检索词为 b5、b6 和 b4，与 b4 相关配的第二检索词为 b7、
10 b8 和 b9，与 b5 相关的第二检索词为 b3。将所有检索词通过图 3a 所示的图数据结构表示。参见图 3a，图 3a 为本发明实施例提供的检索词之间的拓扑图结构第一示意图。在图 3a 中，将每个检索词作为节点 bi（i 取值为 1 至 9），从节点 bi 至节点 bj（j 取值为 1 至 9）的箭头，表示 bi 可扩展出 bj，也即，与 bi 的相关检索词为 bj。从图 3a 可以看出，图 3a
15 示出的拓扑图是一个有向无环图，也就是说，两个检索词之间的相关关系并非保证是双向相关，具体为：从 bi 可以扩展出与 bi 相关的检索词为检索词 bj，但从检索词 bj 并非一定扩展出与检索词 bj 相关的检索词为检索词 bi。

之后，基于步骤 201a，则可得到：针对 b1，根据 b1 的文本特征和/
20 或语义特征计算 b1 和 b2 之间的相似度值 w12，b1 和 b3 之间的相似度值 w13，b1 和 b4 之间的相似度值 w14；针对 b3，根据 b3 的文本特征和/
或语义特征计算 b3 和 b4 之间的相似度值 w14，b3 和 b5 之间的相似度值 w35，b3 和 b6 之间的相似度值 w36；针对 b4，根据 b4 的文本特征和/
或语义特征计算 b4 和 b7 之间的相似度值 w47，b4 和 b8 之间的相似度
25 值 w48，b4 和 b9 之间的相似度值 w49；针对 b5，根据 b5 的文本特征和

/或语义特征计算 b5 和 b3 之间的相似度值 w53。

之后，针对图 3a 中用户提供的每一个第一检索词执行步骤 202a，当执行完步骤 202a 时，图 3a 即可变成图 3b。参见图 3b，图 3b 为本发明实施例提供的检索词之间的拓扑图结构第二示意图。图 3b 示出了相互连接的检索词之间的聚类关系，其中，实线连接的两个检索词表示该两个检索词具有的聚类关系为：两者被认为等价，可聚类在一起；虚线连接的两个检索词具有的聚类关系为：两者不等价，不可聚类在一起，后续可去掉该虚线。

由于在图 3a 所示的拓扑图中，与同一个第一检索词相关的各个第二检索词之间也可能具有潜在的聚类关系。这种聚类关系可能已经在步骤 203 找到（比如，b3 和 b4 之间的聚类关系），也可能没有（比如，b2 和 b3 之间的聚类关系）。为使检索词聚类更加精确，依据图 2b 所示的潜在聚类关系挖掘流程，可得到其中，与用户提供的相关的之间潜在的聚类关系可通过图 3c 中的虚线表示的潜在的聚类关系。以图 3c 中用户提供的的第一检索词 b1 为例进行描述，用户提供的其他检索词原理类似。如此，根据上面图 3a 的描述可以知道，与 b1 的第二检索词为：b2、b3 和 b4，如此，基于步骤 201b，当 b2、b3 和 b4 与 b1 之间的相似度值均大于或等于第二预设阈值时，本发明可以补充挖掘出 3 条潜在的聚类关系：b2 与 b3 之间的聚类关系，b2 与 b4 之间的聚类关系，以及 b3 与 b4 之间的聚类关系。其中，b3 与 b4 之间的聚类关系已经在上述步骤 202a 被确定，因此，作为本发明实施例的一种扩展，本发明可省略执行确定 b3 与 b4 之间的聚类关系的操作，只需增加 b2 与 b3 之间的聚类关系和 b2 与 b4 之间的聚类关系。之后计算出 b2 与 b3 之间的相似度值，以及 b2 与 b4 之间的相似度值，判断 b2 与 b3 之间的聚类关系和 b2 与 b4 之间的聚类关系是否符合聚类的标准，具体为：基于上述步骤 202b，判断该

b2 与 b3 之间的相似度值是否大于或等于第一预设阈值，如果是，则确定 b2 与 b3 之间的聚类关系为：b2 和 b3 等价，可聚类在一起，否则，确定 b2 与 b3 之间的聚类关系为：不将 b2 和 b3 聚类在一起。同理，b2 与 b4 之间的相似度值也是执行类似方法。

5 当通过上面描述验证出图 3c 中虚线连接的两个检索词等价，可以聚类在一起时，将该虚线变为实线；否则，保持该虚线不变，即认为虚线连接的两个检索词不等价，不可聚类在一起，后续可将该虚线去掉。之后，将最终由实线连接的所有检索词作为本发明实施例最终的聚类结果。

10 由于本发明实施例中，检索词之间的聚类关系通过检索词之间的实线（也称为边关系）表示，因此，本发明实施例可仅遍历边关系，这样，可以使本发明实施例的复杂度降低为 $O(n+e)$ ，其中 n 表示检索词的个数， e 表示边关系的个数。

 需要说明的是，作为本发明实施例的一种扩展，本发明实施例中，
15 还可进一步挖掘图 3a 中与用户提供的第二检索词相关的第二检索词，以及该第二检索词在 N （比如 N 为 3）跳以内各“子孙”节点之间潜在的聚类关系，具体实现方式参见图 2b 所示流程，这里不再详述。

 另外，在竞价广告系统中，候选检索词集合并非固定不变，其可随着时间的推移递增检索词。比如，在某一个时间点，候选检索词集合又
20 新增加了用户提供的第二检索词，该新增加的第二检索词，相对之前的所有检索词是新出现的。对该新增加的第二检索词，也需要对其执行类似图 2a 和图 2b 所示的聚类操作，同时，将执行聚类操作后得到的结果与之前的聚类结果整合到一起。具体见图 4 所示的流程。

 参见图 4，图 4 为本发明实施例提供的新增加第二检索词时的流程
25 （记为增量更新流程）图。如图 4 所示，该流程可包括以下步骤：

步骤 401, 确定与该增加的第一检索词相关的第二检索词, 并将该增加的第一检索词和该确定的与该增加的第一检索词相关的第二检索词中与所述候选检索词集合中任一检索词不同的第二检索词添加到所述候选检索词集合中。

5 比如, 候选检索词集合在执行步骤 401 之前存放的检索词为图 3a 所示的 b1 至 b9, 当执行到本步骤 401 时, 如果新增加以下两个第一检索词: n1 和 n2。其中, 与 n1 相关的第二检索词为 b5 和 b6, 与 n2 相关的第二检索词为 b1、b2、b3、b4、b8 和 n3, 具体可参见图 3d 所示。由于与 n1 相关的 b5 和 b6, 以及与 n2 相关的 b1、b2、b3、b4、b8 都已存
10 放在候选检索词集合中, 因此, 本步骤 401 仅可将 n1、n2, 以及与 n2 相关的 n3 添加至候选检索词集合。

步骤 402, 根据检索词的文本特征和/或语义特征对所述候选检索词集合中新增加的第一检索词和与该第一检索词相关的第二检索词执行聚类操作。

15 该聚类操作与图 2a 所示的流程类似。下面仅以新增加的第一检索词为 n1 为例对本步骤 402 进行描述, 增加的其他检索词原理类似。

 则, 针对 n1, 基于步骤 401, 确定出与该 n1 相关的第二检索词为 b5 和 b6。如此, 执行到本步骤 402 时, 基于图 2a 所示的流程, 则根据 n1 的文本特征和/或语义特征计算 n1 与 b5 之间的相似度值, 以及计算
20 n1 与 b6 之间的相似度值, 之后判断 n1 与 b5 之间的相似度值是否大于或等于第一预设阈值, 如果是, 则确定 n1 与 b5 等价, 两者可以聚类在一起, 否则, 不将 n1 与 b5 聚类在一起。针对 n1 与 b6 之间的相似度值也执行同样操作。

 步骤 403, 对候选检索词集合中与增加的第一检索词相关的第二检索词进行潜在聚类关系的挖掘。
25

本步骤 403 可采用图 2b 所示的流程进行潜在聚类关系的挖掘，简单描述为：从候选检索词集合中与增加的第一检索词相关的各个第二检索词，或者从与增加的第一检索词聚类在一起的各个第二检索词中选取与该第一检索词之间的相似度值大于或等于第二预设阈值的第二检索词；计算该选取的任意两个第二检索词之间的相似度值，如果该计算的相似度值大于或等于所述第一预设阈值，则将该两个第二检索词聚类在一起。

仍以新增加的第一检索词为检索词 n1 为例，由于在步骤 401 确定出与该 n1 相关的第二检索词为 b5 和 b6，因此，执行到本步骤 403 时，如果 b5 和 b6 分别与 n1 之间的相似度值均大于第二预设阈值，则可计算 b5 和 b6 之间的相似度值，如果该计算的相似度值大于或等于第一预设阈值，则将该两个检索词 b5 和 b6 聚类在一起，否则，不将 b5 和 b6 聚类在一起。

至此，通过上述步骤 401 至步骤 403 实现了新增加的第一检索词（记为增量检索词）与原来已存在的检索词（记为旧检索词）之间的聚类关系（下文记为增量聚类结果）。该增量聚类结果与之前存在的全量聚类结果统称为本发明最终的聚类结果。

需要说明的是，本实施例中，与第一检索词相关的第二检索词并非固定不变，其也会根据用户增删检索词发生变化，基于此，本发明实施例提供的方法还应能够反映该变化。该变化通过定期更新候选检索词集合（记为全量更新）实现，具体实现为：在设定的全量更新时间到达时，针对所述候选检索词集合中的第一检索词，确定与该第一检索词相关的第二检索词，将该第一检索词和确定的与该第一检索词相关的第二检索词均放入一个新的候选检索词集合中，之后按照图 2a 和图 2 所示的流程对该新的候选检索词集合中的检索词进行聚类，得到全量聚类结果。这

可通过表 1 形象描述。

假设第一天用户提供的的第一检索词为： B_1 ，该第一检索词相应的 QBM 扩展结果为 $Q_1=Q(B_1)$ 中，该扩展结果主要为与该第一检索词相关的第二检索词的集合。基于图 2a 和图 2b 所示的流程对第一检索词和
5 第二检索词进行聚类得到的聚类结果为： $C_1=C(Q(B_1))$ ；如此，随着时间推移需要增加检索词时，参见表 1：

	增量更新	全量更新	备注
第 2 天	截至当天全量检索词： B_2 增加的检索词： $B_{21}=B_2-B_1$ 增加的检索词对应的 QBM 扩展结果： $Q(B_{21})$ 增量聚类结果： $C(Q(B_{21}))$ 最终聚类结果： $C_2=C(Q(B_{21}))$ $\cup C_1$		只做增量更新，不做全量更新。
第 3 天	截至当天全量检索词： B_3 增加的检索词： $B_{32}=B_3-B_2$ 增加的检索词对应的 QBM 扩展结果： $Q(B_{32})$ 增量聚类结果： $C(Q(B_{32}))$ 最终聚类结果： $C_3=C(Q(B_{32}))$ $\cup C_2$		只做增量更新，不做全量更新。
			只做增量更新，不做全量更新。
第 i 天	截至当天全量检索词： B_i 增加的检索词： $B_{i3}=B_i-B_{i-1}$ 增加的检索词对应的 QBM 扩展结果： $Q(B_{i(i-1)})$ 增量聚类结果： $C(Q(B_{i(i-1)}))$ 最终聚类结果： $C_i=C(Q(B_i$	基于截止第 i 天的全量检索词数据，全量更新准备中...	开始做基于第 i 天全量检索词的全量更新，这个过程可能持续几天。

	$(_{i-1}) \cup C_{i-1}$		
		全量更新准备 中...	
第 j 天		全量更新准备 中.....	
第 k 天	截至当天全量检索词: B_k 增加的检索词: $B_{kj}=B_k-B_j$ 增加的检索词对应的增量 QBM 扩展结果: $Q(B_{kj})$ 增量聚类结果: $C(Q(B_{kj}))$ 最终聚类结果: $C_k=C(Q(B_{kj}))$ $\cup C_j$	最新的全量 QBM 扩展结果: $total_Q_k=Q(B_i)$ 相应的全量聚类 结 果 : $total_C_k=C(Q(B_i))$	到第 k 天的时候, 基 于第 i 天全量检索词数 据的全量聚类结果已经 算好了。
第 L 天	截至当天全量检索词: B_L 增加的检索词: $B_{Li}=B_L-B_i$ 增加的检索词对应的 QBM 扩展结果: $Q(B_{Li})$ 增量聚类结果: $C(Q(B_{Li}))$ 最终聚类结果: $C_L=C(Q(B_{Li}))$ $\cup total_C_k$		到第 L 天, 用第 k 天 算好的全量检索词聚类 结果做同步, 让增量更 新流程中基于这个最新 的全量做增量扩展。因 此, 增量检索词是相对 第 i 天的。
第 m 天	截至当天全量检索词: B_m 增量检索词: $B_{mL}=B_m-B_L$ 增量 QBM 扩展结果: $Q(B_{mL})$ 增量聚类结果: $C(Q(B_{mL}))$ 最 终 聚 类 结 果 : $C_m=C(Q(B_{mL})) \cup C_L$		只做增量更新, 不做 全量更新, 重复最开始 的流程周期。

从表 1 可以看出, 全量更新在第 i 天开始, 第 k 天结束, 在第 k+1 (也即 L) 天, 做全量数据与增量数据的同步操作, 即, 将截止到第 k+1 (也即 L) 天候选检索词集合中的所有第一检索词执行图 4 所示的流程。

下面对本发明实施例提供的装置进行描述。

参见图 5，图 5 为本发明实施例提供的装置的基本结构图。如图 5 所示，该装置可包括：

建立单元 501，用于建立候选检索词集合，所述候选检索词集合包含由用户提供的第二检索词、以及与第二检索词相关的第三检索词；

聚类单元 502，用于根据检索词的文本特征和/或语义特征对所述候选检索词集合中的第二检索词和与该第二检索词相关的第三检索词执行聚类操作。

在具体实现时，图 5 所示的装置具体可参见图 6。

参见图 6，图 6 为本发明实施例提供的装置的详细结构图。如图 6 所示，该装置可包括建立单元 601 和聚类单元 602，其中，建立单元 601 和聚类单元 602 具有的功能分别与图 5 所示的建立单元 501 和聚类单元 502 类似，这里不再赘述。

优选地，如图 6 所示，该装置可进一步包括：

添加单元 603，用于当用户增加新的第二检索词时，确定与该增加的第二检索词相关的第三检索词，并将该增加的第二检索词和该确定的与该增加的第二检索词相关的第三检索词中与所述候选检索词集合中任一检索词不同的第三检索词添加到所述候选检索词集合中；

基于此，聚类单元 602 还用于根据检索词的文本特征和/或语义特征对所述候选检索词集合中新增加的第二检索词和与该第二检索词相关的第三检索词执行聚类操作。

优选地，如图 6 所示，该装置进一步包括：

更新单元 604，用于在设定的全量更新时间到达时，针对所述候选检索词集合中的第二检索词，确定与该第二检索词相关的第三检索词，将该第二检索词和确定的与该第二检索词相关的第三检索词均放入一

个新的候选检索词集合中。

基于此，聚类单元 602 还用于根据检索词的文本特征和/或语义特征对该新的候选检索词集合中第一检索词和与该第一检索词相关的第二检索词执行聚类操作。

5 具体地，聚类单元 602 通过以下子单元执行聚类操作：

计算子单元 6021，用于根据第一检索词的文本特征和/或语义特征分别计算该第一检索词和与该第一检索词相关的各个第二检索词之间的相似度值；

10 聚类子单元 6022，用于在第一检索词和第二检索词之间的相似度值大于或等于第一预设阈值时，将该第一检索词和该第二检索词聚类在一起。

优选地，聚类子单元 6022 还用于从与第一检索词相关的各个第二检索词，或者从与第一检索词聚类在一起的各个第二检索词中选取与该第一检索词之间的相似度值大于或等于第二预设阈值的第二检索词；以及
15 计算该选取的任意两个第二检索词之间的相似度值，如果该计算的相似度值大于或等于所述第一预设阈值，则将该两个第二检索词聚类在一起，所述第一预设阈值与第二预设阈值无关。

以上对本发明实施例提供的装置进行了描述。

由以上技术方案可以看出，本发明提供的检索词的聚类方法和装置
20 在进行检索词聚类时，并非像现有技术那样仅将用户提供的检索词进行字面关系的聚类，而是同时考虑到用户提供的检索词，以及与该用户提供的检索词相关的其他检索词，并根据检索词的文本特征和/或语义特征对用户提供的检索词，以及与该用户提供的检索词相关的其他检索词进行聚类，这显然大大增加检索词聚类的准确度；

25 进一步地，本发明还挖掘出与用户提供的第一检索词相关的各个第

二检索词之间的聚类关系，这相比于现有技术，可深度挖掘出检索词之间的聚类关系，使检索词的聚类更加精确。

以上所述仅为本发明的较佳实施例而已，并不用以限制本发明，凡在本发明的精神和原则之内，所做的任何修改、等同替换、改进等，均
5 应包含在本发明保护的范围之内。

权利要求书

1、一种检索词的聚类方法，其特征在于，该方法包括：

建立候选检索词集合，所述候选检索词集合包含由用户提供的第一检索词、以及与第一检索词相关的第二检索词；

5 根据检索词的文本特征和/或语义特征对所述候选检索词集合中的第一检索词和与该第一检索词相关的第二检索词执行聚类操作。

2、根据权利要求1所述的方法，其特征在于，当用户增加第一检索词时，该方法进一步包括：

10 确定与该增加的第一检索词相关的第二检索词，并将该增加的第一检索词和该确定的与该增加的第一检索词相关的第二检索词中与所述候选检索词集合中任一检索词不同的第二检索词添加到所述候选检索词集合中；

根据检索词的文本特征和/或语义特征对所述候选检索词集合中新增加的第一检索词和与该第一检索词相关的第二检索词执行聚类操作。

15 3、根据权利要求1所述的方法，其特征在于，该方法进一步包括：

在设定的全量更新时间到达时，针对所述候选检索词集合中的第一检索词，确定与该第一检索词相关的第二检索词，将该第一检索词和确定的与该第一检索词相关的第二检索词均放入一个新的候选检索词集合中，根据检索词的文本特征和/或语义特征对该新的候选检索词集合中
20 第一检索词和与该第一检索词相关的第二检索词执行聚类操作。

4、根据权利要求1至3任一所述的方法，其特征在于，根据检索词的文本特征和/或语义特征对第一检索词和与该第一检索词相关的第二检索词执行聚类操作包括：

根据第一检索词的文本特征和/或语义特征分别计算该第一检索词

和与该第一检索词相关的各个第二检索词之间的相似度值，如果第一检索词和第二检索词之间的相似度值大于或等于第一预设阈值，则将该第一检索词和该第二检索词聚类在一起。

5、根据权利要求4所述的方法，其特征在于，该方法进一步包括：

5 从与第一检索词相关的各个第二检索词，或者从与第一检索词聚类在一起的各个第二检索词中选取与该第一检索词之间的相似度值大于或等于第二预设阈值的第二检索词；

计算该选取的任意两个第二检索词之间的相似度值，如果该计算的相似度值大于或等于所述第一预设阈值，则将该两个第二检索词聚类在一起。
10

6、根据权利要求1所述的方法，其特征在于，与第一检索词相关的第二检索词包括：

与第一检索词匹配的检索词，和/或，以第一检索词为关键词进行搜索所得到的搜索结果中的检索词。

15 7、一种检索词的聚类装置，其特征在于，该装置包括：

建立单元，用于建立候选检索词集合，所述候选检索词集合包含由用户提供的第一检索词、以及与第一检索词相关的第二检索词；

聚类单元，用于根据检索词的文本特征和/或语义特征对所述候选检索词集合中的第一检索词和与该第一检索词相关的第二检索词执行聚类操作。
20

8、根据权利要求7所述的装置，其特征在于，所述装置进一步包括：

添加单元，用于当用户增加第一检索词时，确定与该增加的第一检索词相关的第二检索词，并将该增加的第一检索词和该确定的与该增加的第一检索词相关的第二检索词中与所述候选检索词集合中任一检索词不同的第二检索词添加到所述候选检索词集合中；
25

所述聚类单元,还用于根据检索词的文本特征和/或语义特征对所述候选检索词集合中新增加的第一检索词和与该第一检索词相关的第二检索词执行聚类操作。

- 9、根据权利要求7所述的装置,其特征在于,所述装置进一步包括:
- 5 更新单元,用于在设定的全量更新时间到达时,针对所述候选检索词集合中的第一检索词,确定与该第一检索词相关的第二检索词,将该第一检索词和确定的与该第一检索词相关的第二检索词均放入一个新的候选检索词集合中;

所述聚类单元还用于根据检索词的文本特征和/或语义特征对该新的候选检索词集合中第一检索词和与该第一检索词相关的第二检索词执行聚类操作。

10、根据权利要求7至9任一所述的装置,其特征在于,所述聚类单元通过以下子单元执行聚类操作:

- 计算子单元,用于根据第一检索词的文本特征和/或语义特征分别计
- 15 算该第一检索词和与该第一检索词相关的各个第二检索词之间的相似度值;

聚类子单元,用于在第一检索词和第二检索词之间的相似度值大于或等于第一预设阈值时,将该第一检索词和该第二检索词聚类在一起。

- 11、根据权利要求10所述的装置,其特征在于,所述聚类子单元还
- 20 用于从与第一检索词相关的各个第二检索词,或者从与第一检索词聚类在一起的各个第二检索词中选取与该第一检索词之间的相似度值大于或等于第二预设阈值的第二检索词;以及计算该选取的任意两个第二检索词之间的相似度值,如果该计算的相似度值大于或等于所述第一预设阈值,则将该两个第二检索词聚类在一起。

1/7

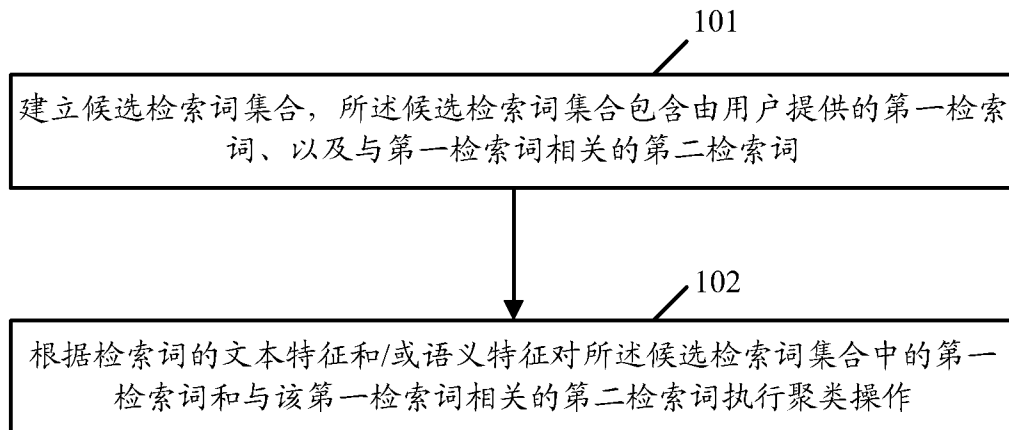


图 1

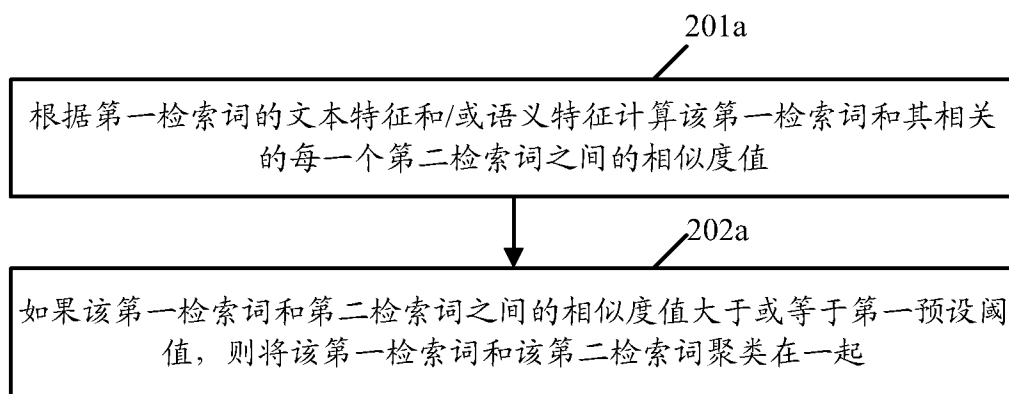


图 2a

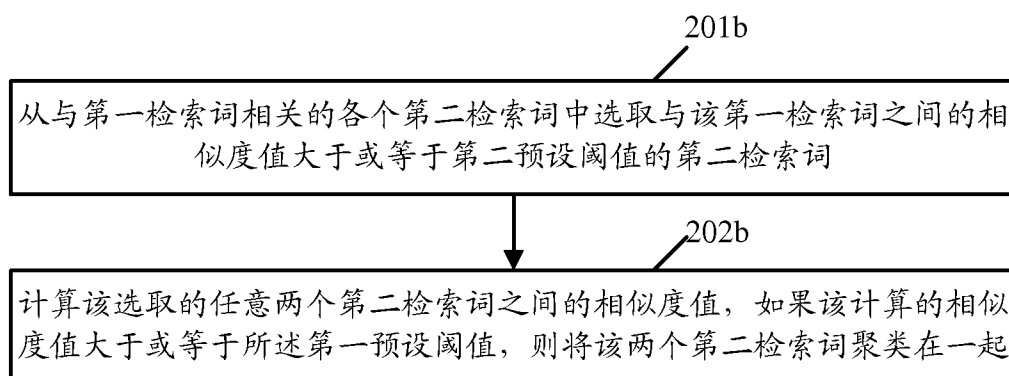


图 2b

2/7

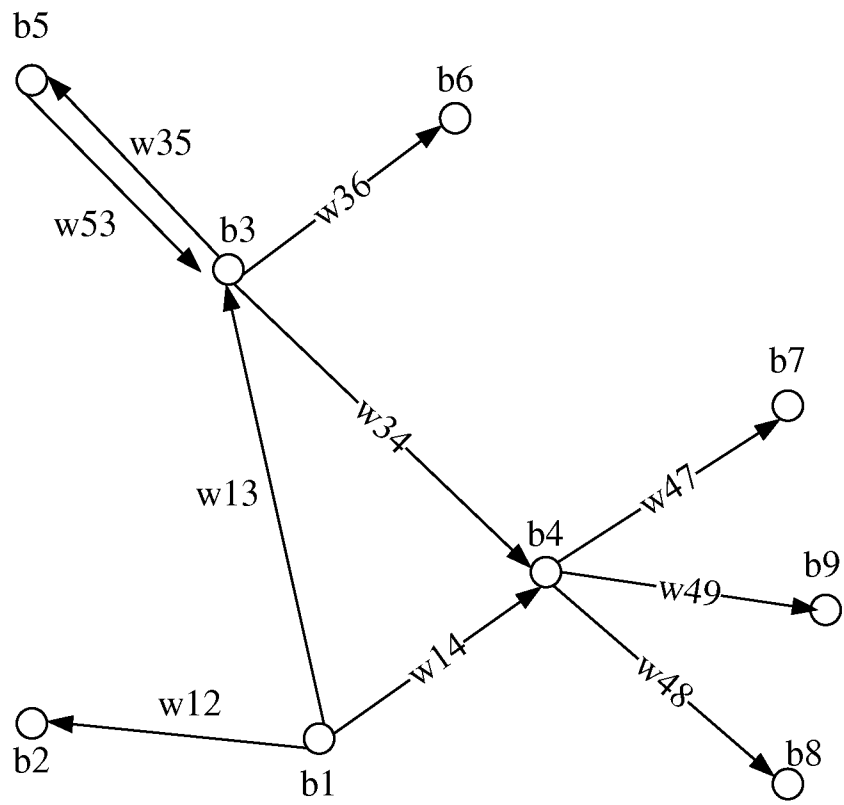


图 3a

3/7

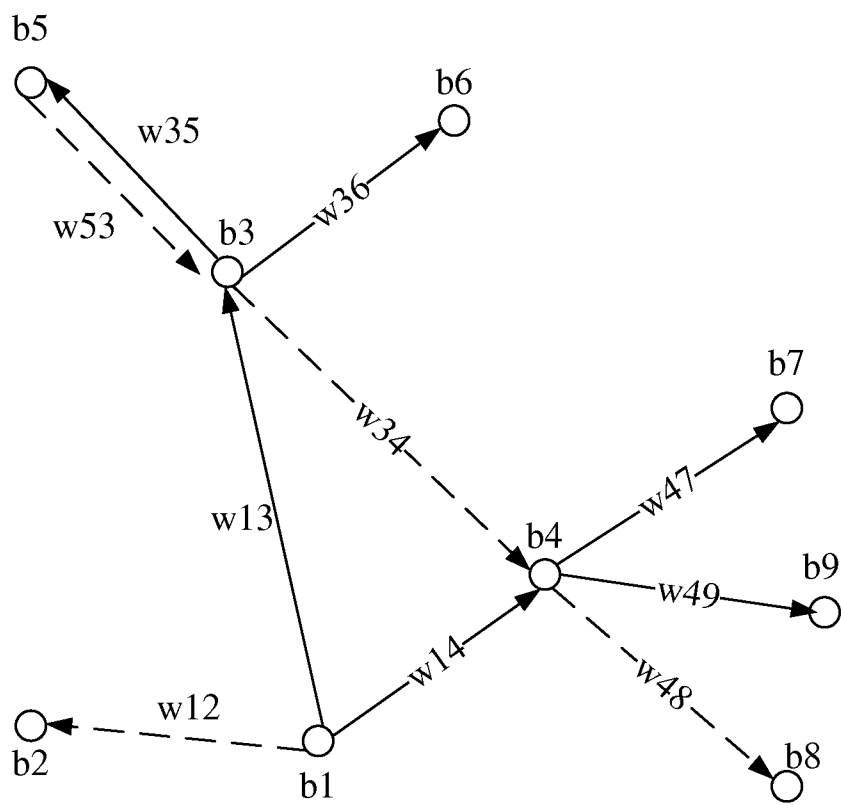


图 3b

4/7

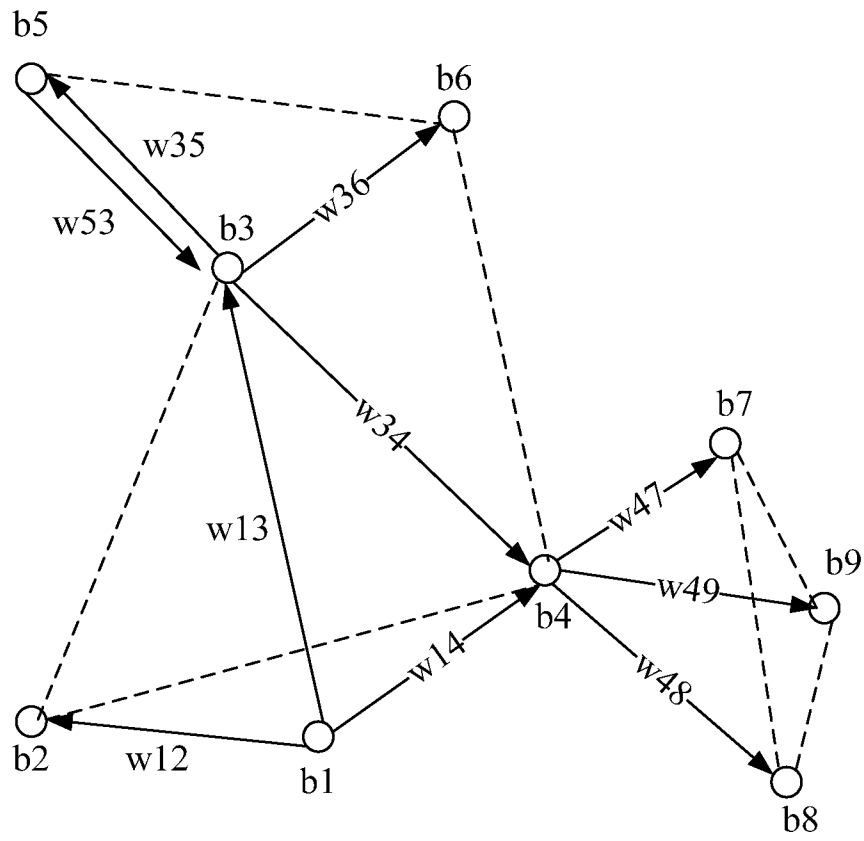


图 3c

5/7

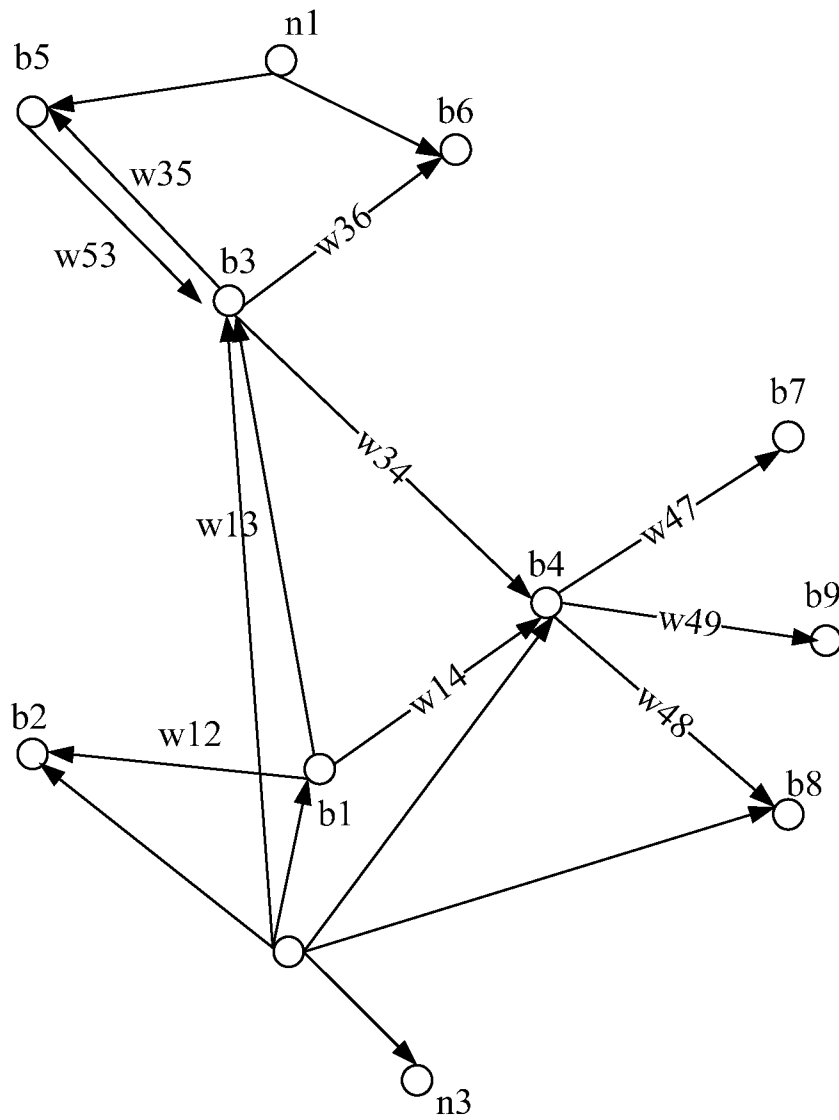


图 3d

6/7

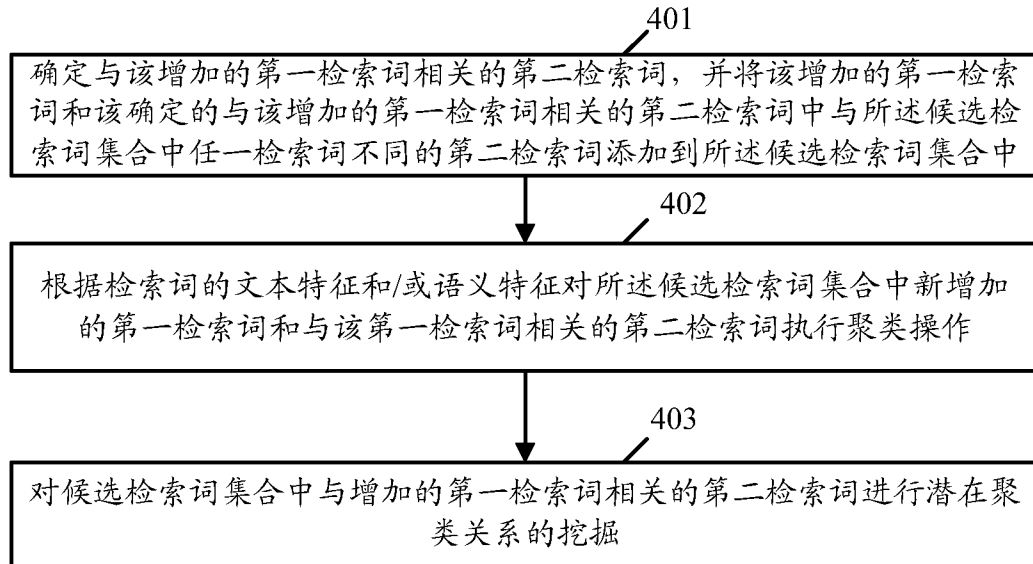


图 4

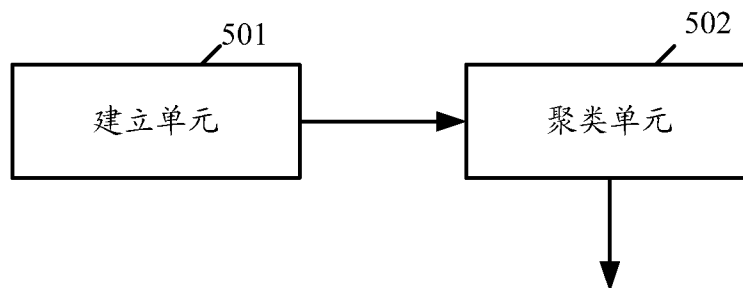


图 5

7/7

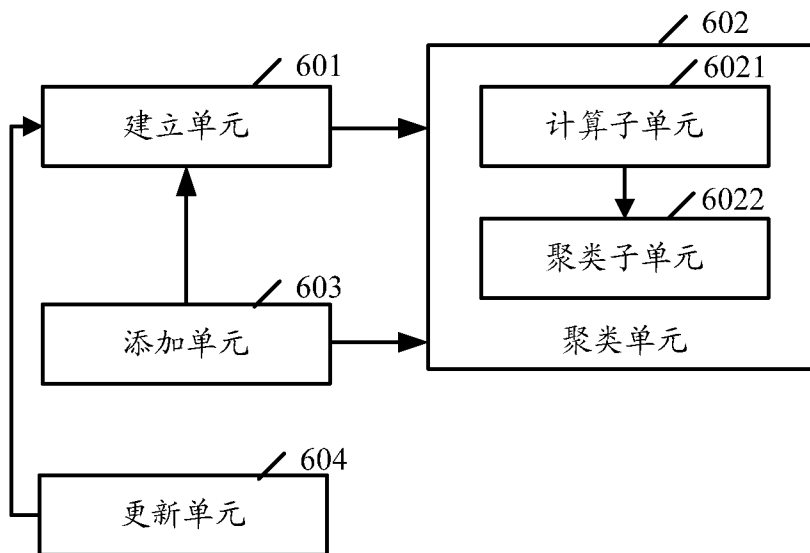


图 6

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2012/070824

A. CLASSIFICATION OF SUBJECT MATTER

G06F 17/30(2006.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC: G06F 17/-

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNKI; CNABS; WPI; EPODOC; keyword?, key w word?, cluster+, recommend+

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	KR 20100106718 A (NHN CORP) 04 Oct.2010(04.10.2010) abstract, description, paragraphs 4, 16-85, figures 1-7	1-4,6-10
A		5-11
A	US 2010131563 A1(YEBOL CORP)27 May 2010(27.05.2010) the whole document	1-11

☐ Further documents are listed in the continuation of Box C.

☒ See patent family annex.

* Special categories of cited documents:	“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
“A” document defining the general state of the art which is not considered to be of particular relevance	“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
“E” earlier application or patent but published on or after the international filing date	“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	“&”document member of the same patent family
“O” document referring to an oral disclosure, use, exhibition or other means	
“P” document published prior to the international filing date but later than the priority date claimed	

Date of the actual completion of the international search 27 Mar.2012(27.03.2012)	Date of mailing of the international search report 19 Apr.2012(19.04.2012)
Name and mailing address of the ISA State Intellectual Property Office of the P. R. China No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088, China Facsimile No. 86-10-62019451	Authorized officer LIU, Lin Telephone No. (86-10)62411953

International application No.
PCT/CN2012/070824

Form PCT/ISA/210 (patent family annex) (July 2009)

A. 主题的分类

G06F 17/30(2006.01)i

按照国际专利分类(IPC)或者同时按照国家分类和 IPC 两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

IPC: G06F17/-

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称, 和使用的检索词(如使用))

CNKI;CNABS;WPI;EPODOC;检索词,关键词,聚类,推荐,keyword?, key w word?, cluster+, recommend+

C. 相关文件

类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求
X	KR 20100106718 A(NHN CORP) 04.10 月 2010(04.10.2010) 摘要、说明书第 4,16-85 段、图 1-7	1-4,6-10
A		5,11
A	US 2010131563 A1(YEBOL CORP) 27.5 月 2010(27.05.2010) 全文	1-11



其余文件在 C 栏的续页中列出。



见同族专利附件。

* 引用文件的具体类型:

“A” 认为不特别相关的表示了现有技术一般状态的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇
引用文件的公布日而引用的或者因其他特殊理由而引
用的文件(如具体说明的)

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了
理解发明之理论或原理的在后文件“X” 特别相关的文件, 单独考虑该文件, 认定要求保护的
发明不是新颖的或不具有创造性“Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件
结合并且这种结合对于本领域技术人员为显而易见时,
要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期

27.3 月 2012(27.03.2012)

国际检索报告邮寄日期

19.4 月 2012 (19.04.2012)

ISA/CN 的名称和邮寄地址:

中华人民共和国国家知识产权局

中国北京市海淀区蓟门桥西土城路 6 号 100088

传真号: (86-10)62019451

受权官员

刘琳

电话号码: (86-10) 62411953

关于同族专利的信息

PCT/CN2012/070824

10.06.2010