

(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
2016 年 12 月 15 日 (15.12.2016) WIPO | PCT



(10) 国际公布号
WO 2016/197577 A1

- (51) 国际专利分类号:
G06F 17/30 (2006.01)
- (21) 国际申请号: PCT/CN2015/097774
- (22) 国际申请日: 2015 年 12 月 17 日 (17.12.2015)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
201510325108.2 2015 年 6 月 12 日 (12.06.2015) CN
- (71) 申请人: 百度在线网络技术(北京)有限公司 (BAIDU ONLINE NETWORK TECHNOLOGY (BEIJING) CO., LTD.) [CN/CN]; 中国北京市海淀区上地十街 10 号百度大厦三层, Beijing 100085 (CN)。
- (72) 发明人: 李炫 (LI, Xuan); 中国北京市海淀区上地十街 10 号百度大厦三层, Beijing 100085 (CN)。 沈剑平 (SHEN, Jianping); 中国北京市海淀区上地十街 10 号百度大厦三层, Beijing 100085 (CN)。 莫洋 (MO, Yang); 中国北京市海淀区上地十街 10 号百度大厦三层, Beijing 100085 (CN)。 宋元峰 (SONG, Yuanfeng); 中国北京市海淀区上地十街 10 号百度大厦三层, Beijing 100085 (CN)。 郑楚煜 (ZHENG, Chuyu); 中国北京市海淀区上地十街 10 号百度大

厦三层, Beijing 100085 (CN)。 车丽美 (CHE, Limei); 中国北京市海淀区上地十街 10 号百度大厦三层, Beijing 100085 (CN)。 齐沁芳 (QI, Qinfang); 中国北京市海淀区上地十街 10 号百度大厦三层, Beijing 100085 (CN)。

(74) 代理人: 北京金律言科知识产权代理事务所(普通合伙) (INGENUITY INTELLECTUAL PROPERTY LAW FIRM); 中国北京市海淀区上地十街 1 号辉煌国际大厦 1 号楼 1004 室, Beijing 100085 (CN)。

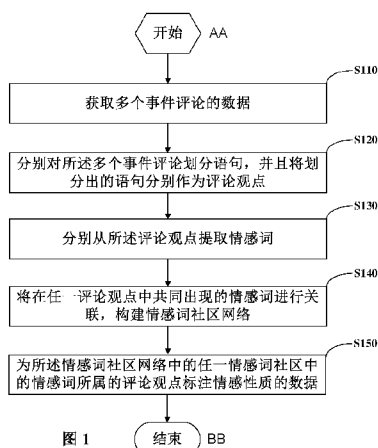
(81) 指定国 (除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, 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,

[见续页]

(54) Title: METHOD AND APPARATUS FOR LABELLING COMMENT INFORMATION AND COMPUTER DEVICE

(54) 发明名称: 评论信息的标注方法、装置和计算机设备



S110 Acquiring data of comments on a plurality of events
S120 Respectively dividing sentences of the comments on the plurality of events, and respectively taking the divided sentences as comment viewpoints
S130 Respectively extracting sentiment words from the comment viewpoints
S140 Associating sentiment words commonly appearing in any one of the comment viewpoints to construct a sentiment word community network
S150 Labelling data of sentiment properties for the comment viewpoints to which the sentiment words in any one of the sentiment word communities of the sentiment word community network belong
AA Start
BB End

(57) Abstract: A method and apparatus for labelling comment information and a computer device. The method comprises: acquiring data of comments on a plurality of events (S110); respectively dividing sentences of the comments on the plurality of events, and respectively taking the divided sentences as comment viewpoints (S120); respectively extracting sentiment words from the comment viewpoints (S130); associating sentiment words commonly appearing in any one of the comment viewpoints to construct a sentiment word community network (S140); and labelling data of sentiment properties for the comment viewpoints to which the sentiment words in any one of the sentiment word communities of the sentiment word community network belong (S150), wherein the sentiment properties may be positive, negative or neutral, and the sentiment word communities comprise a group of directly or indirectly associated sentiment words. The labelling method and apparatus and the computer device can automatically perform sentiment clustering on event comments and label sentiment properties therefor.

(57) 摘要:

[见续页]

WO 2016/197577 A1



RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, 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, KM, 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))。

一种评论信息的标注方法、装置和计算机设备，所述方法包括：获取多个事件评论的数据（S110）；分别对所述多个事件评论划分语句，并且将划分出的语句分别作为评论观点（S120）；分别从所述评论观点提取情感词（S130）；将在任一评论观点中共同出现的情感词进行关联，构建情感词社区网络（S140）；为所述情感词社区网络中的任一情感词社区中的情感词所属的评论观点标注情感性质的数据（S150），所述情感性质是正面、负面或中立，所述情感词社区包括一组直接或间接关联的情感词。该标注方法、装置和计算机设备能够自动地对事件评论进行情感聚类并为其标注情感性质。

评论信息的标注方法、装置和计算机设备

本申请要求于 2015 年 6 月 12 日提交中国专利局、申请号为
5 201510325108.2、发明名称为“评论信息的标注方法及装置”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

技术领域

10 本发明涉及网络信息处理技术领域，尤其涉及一种评论信息的标注方法、装置和计算机设备。

背景技术

通常，现有的评论聚类方法多是关于商品评论的，例如天猫，亚马逊等网站中的用户评价。其主要是围绕商品的不同属性对评论进行聚
15 类。一般来说，首先构建属性的类别，其次从评论中挖掘包含的属性，然后把评论归到所包含属性的类别。其中，从评论中挖掘属性常采用例如基于词典、基于机器学习等方法。

然而，事件评论与商品评论有着本质的不同，事件评论大多都没有属性。因此，无法简单地把商品聚类方法照搬到事件评论上。其次，事
20 件评论内容较广，使得常用的基于词典预先构建类别的方法无法适用于事件评论。

发明内容

本发明实施例的目的在于，提供一种评论信息的标注方法、装置和
25 计算机设备，以实现自动地对事件评论进行情感聚类，并为其标注情感性质，提升用户网络体验。

为实现上述发明目的，本发明的实施例提供了一种评论信息的标注方法，包括：获取多个事件评论的数据；分别对所述多个事件评论划分
30 语句，并且将划分出的语句分别作为评论观点；分别从所述评论观点提取情感词；将在任一评论观点中共同出现的情感词进行关联，构建情感词社区网络；为所述情感词社区网络中的任一情感词社区中的情感词所

属的评论观点标注情感性质的数据，所述情感性质是正面、负面或中立，所述情感词社区包括一组直接或间接关联的情感词。

5 优选地，所述分别从所述评论观点提取情感词的处理包括：分别对所述评论观点的语句进行切词，并且通过将切出的分词与预先构建的情感词词典匹配来选取所述情感词，所述情感词词典包括多个情感词以及其情感性质的数据。

10 优选地，所述将在任一评论观点中共同出现的情感词进行关联，构建情感词社区网络的处理还包括：对于所述情感词社区网络中的任意两个关联的情感词，计算所述两个情感词共同出现在同一评论观点中的共现频度，并且如果计算的共现频度的值低于预定的共现频度阈值，则去除所述两个情感词之间的关联关系。

优选地，所述为所述情感词社区网络中的任一情感词社区中的情感词所属的评论观点标注情感性质的数据的处理包括：根据所述情感词社区中的情感词的情感性质为其所属的评论观点标注情感性质的数据。

15 优选地，所述情感词词典还包括所述多个情感词的情感强度的数据，所述为所述情感词社区网络中的任一情感词社区中的情感词所属的评论观点标注情感性质的数据的处理还包括：如果任一所述评论观点包括不同情感性质的情感词，则为所述评论观点标注情感强度最强的情感词对应的情感性质的数据。

20 优选地，所述分别对所述多个事件评论划分语句的处理还包括：去除字数超过预定句长的语句，和/或去除广告性质的语句。

25 本发明的实施例还提供了一种评论信息的标注装置，包括：评论获取模块，用于获取多个事件评论的数据；评论分句模块，用于分别对所述多个事件评论划分语句，并且将划分出的语句分别作为评论观点；情感词提取模块，用于分别从所述评论观点提取情感词；情感网络构建模块，用于将在任一评论观点中共同出现的情感词进行关联，构建情感词社区网络；情感标注模块，用于为所述情感词社区网络中的任一情感词社区中的情感词所属的评论观点标注情感性质的数据，所述情感性质是正面、负面或中立，所述情感词社区包括一组直接或间接关联的情感词。

优选地，所述情感词提取模块用于分别对所述评论观点的语句进行切词，并且通过将切出的分词与预先构建的情感词词典匹配来选取所述情感词，所述情感词词典包括多个情感词以及其情感性质的数据。

5 优选地，所述情感网络构建模块用于对于所述情感词社区网络中的任意两个关联的情感词，计算所述两个情感词共同出现在同一评论观点中的共现频度，并且如果计算的共现频度的值低于预定的共现频度阈值，则去除所述两个情感词之间的关联关系。

优选地，所述情感标注模块用于根据所述情感词社区中的情感词的情感性质为其所属的评论观点标注情感性质的数据。

10 优选地，所述情感词词典还包括所述多个情感词的情感强度的数据，所述情感标注模块还用于如果任一所述评论观点包括不同情感性质的情感词，则为所述评论观点标注情感强度最强的情感词对应的情感性质的数据。

15 优选地，所述评论分句模块还用于去除字数超过预定句长的语句，和/或去除广告性质的语句。

本发明的实施例还提供了一种计算机设备，所述计算机设备包括：一个或多个处理器；存储器；一个或多个程序，所述一个或多个程序存储在所述存储器中，且经配置以由所述一个或者多个处理器执行所述一个或者多个程序包含的用于执行所述评论信息的标注方法的指令：获取
20 多个事件评论的数据；分别对所述多个事件评论划分语句，并且将划分出的语句分别作为评论观点；分别从所述评论观点提取情感词；将在任一评论观点中共同出现的情感词进行关联，构建情感词社区网络；为所述情感词社区网络中的任一情感词社区中的情感词所属的评论观点标注情感性质的数据，所述情感性质是正面、负面或中立，所述情感词社区
25 区包括一组直接或间接关联的情感词。

本发明实施例提供的评论信息的标注方法、装置和计算机设备，通过对获取到的多个事件评论进行分句得到多个评论观点，再以评论观点中包含的情感词为聚类依据，从而自动地对事件评论进行情感聚类，并为其标注情感性质，使得用户能够快速知晓舆情的大体，方便用户阅读，

极大丰富了用户体验。

此外，标注了情感性质的数据的评论观点使得用户能够方便了解其他用户关注事件热点的看法。

5 附图说明

图 1 是示出本发明实施例一的评论信息的标注方法的流程图；

图 2 是示出本发明实施例一的评论信息的标注方法的情感词词典的示例图；

图 3 是示出本发明实施例一的评论信息的标注方法的情感词社区网络的示例图；

图 4 是示出本发明实施例一的评论信息的标注方法的情感聚类结果的示例图；

图 5 是示出本发明实施例二的评论信息的标注装置的逻辑框图；

图 6 是示出本发明实施例三的的计算机设备的逻辑框图。

具体实施方式

本发明的基本构思是，在获取到多个事件评论的数据之后，进一步地对事件评论划分语句，并将划分出的语句作为评论观点，采用“情感词”作为评论观点聚类的关键，从而自动地对事件评论进行情感聚类，并为其标注情感性质。所述情感性质可以是，但不限于，正面、负面或中立，最终生成例如正面言论、负面言论以及中立言论的效果，使得用户能够方便、快速地知道其他用户关注事件的各个方面，提升用户体验。

此外，本发明适用范围广，类似于新闻资讯类的用户评论尤其适用于所述方法，同时还可为舆情监控提供大量的数据来源。

下面结合附图对本发明实施例一种评论信息的标注方法、装置和计算机设备进行详细描述。

实施例一

图 1 是示出本发明实施例一的评论信息的标注方法的流程图。可在

例如微博服务器上执行所述方法。

参照图 1，在步骤 S110，获取多个事件评论的数据。

这里，事件评论的数据可以是例如但不限于，微博、贴吧、新闻、论坛等来源用户发表的评论文本。

- 5 在步骤 S120，分别对所述多个事件评论划分语句，并且将划分出的语句分别作为评论观点。

也就是说，对每个事件评论进行分句，划分出的语句作为一个独立的评论观点。为了更为准确地获得评论观点，根据本发明的可选实施例，步骤 S120 包括：去除字数超过预定句长的语句，和/或去除广告性质的
10 语句。

在步骤 S130，分别从所述评论观点提取情感词。

根据本发明的示例性实施例，步骤 S130 包括：分别对所述评论观点的语句进行切词，并且通过将切出的分词与预先构建的情感词词典匹配来选取所述情感词。

- 15 需要说明的是，现有技术关于情感词抽取一般采用直接匹配的方式，但是直接匹配的方式无法保证抽取效果，有可能会忽略一些情感词，然而，本实施例采用先切词再匹配的方式，也就是说，将一个汉字序列切分成一个一个单独的词语，从而将切出的分词与预先构建的情感词词典进行匹配，匹配到的词语选取为情感词。

- 20 这里，所述情感词词典包括多个情感词以及其情感性质的数据。所述情感性质是，但不限于，正面、负面或中立。

- 图 2 是示出本发明实施例一的评论信息的标注方法的情感词词典的示例图，其中列出了多个情感词样例，例如博学多才、毒谋、猛然等等。这里仅是示意性得列出部分情感词，实际上情感词词典可包括上万甚至
25 更多情感词以及其情感性质的数据。例如，“博学多才”这个情感词表达的是一种正面的情感，其情感性质的数据为“1”代表正面，同理，“毒谋”这个情感词表达的是负面的情感，其情感性质的数据为“-1”代表负面，而“猛然”这个情感词没有明显的情感倾向，其情感性质的数据为“0”，则代表中立。

- 30 在步骤 S140，将在任一评论观点中共同出现的情感词进行关联，构

建情感词社区网络。

具体地，将情感词是否共同出现作为依据，对任一评论观点中的情感词进行关联，那么事件评论的数据中所有的情感词即可构成一个情感词社区网络。这里，所述情感词社区包括一组直接或间接关联的情感词。

5 图 3 是示出本发明实施例一的评论信息的标注方法的情感词社区网络的示例图，其中的每个小圆圈代表一个情感词，两个小圆圈之间的连线代表情感词之间的关联关系，连线长度越短表明情感词之间的关联关系越强，例如情感词 a 和情感词 b 是相关联的，三个虚线大圆圈分别代表三个情感词社区。

10 以关于《亚太城市房地产研究院院长：楼市仍有 20 年黄金期》的事件评论为例，经过前述步骤 S110~S140 的处理得到如下的四个情感词社区：

情感词社区 1：改革、危机、确实、单纯、计较、不行、你以为自己、遥远、奇怪、差的、什么东西、烦恼、问题、快乐、不知道、健康、
15 应该、经济、发展、负荷、牛逼、想、理解、垃圾、无法、其实；

情感词社区 2：皑皑、守候、天真、忠义、敬、痴、硬汉、两肋插刀、解语花、诚、愿、坎坷；

情感词社区 3：自然、清新、标准、最高、精装；

情感词社区 4：太棒了、最爱、感谢。

20 为了更为准确地构建情感词社区网络，根据本发明的示例性实施例，步骤 S140 还包括：对于所述情感词社区网络中的任意两个关联的情感词，计算所述两个情感词共同出现在同一评论观点中的共现频度，并且如果计算的共现频度的值低于预定的共现频度阈值，则去除所述两个情感词之间的关联关系。

25 在具体的实现方式中，可根据所述两个情感词各自的出现次数及其共现次数，计算所述两个情感词的共现频度。这里，共现次数通常是指两个情感词在同一评论观点中出现的次数，比如“博学多才”和“前程似锦”在某一评论观点中同时出现，那么二者的共现次数为 1，倘若“博学多才”和“前程似锦”又在另一评论观点中再次同时出现，它们的共
30 现次数就为 2。前述共现频度的计算可通过以下公式执行：

$$N_{(word1,word2)} = \frac{1}{2} \left[\left(\frac{(word1, word2)}{(word1)} + \frac{(word1, word2)}{(word2)} \right) \right]$$

其中，word1 为情感词 a，word2 为情感词 b，(word1,word2) 为情感词 a 和情感词 b 的共现次数，(word1) 为情感词 a 的出现次数，(word2) 为情感词 b 的出现次数， $N_{(word1,word2)}$ 为情感词 a 与情感词 b 的共现频度。

在获得各个情感词之间的共现频度之后，进一步选取低于预定的共现频度阈值的两个情感词，将符合上述条件的情感词之间的关联关系去除。此外，为了避免个别低频情感词与其他情感词的共现频度过高，还可去除出现次数低于设定出现次数阈值的情感词与其他情感词之间的关联关系。

在步骤 S150，为所述情感词社区网络中的任一情感词社区中的情感词所属的评论观点标注情感性质的数据。

具体地，根据所述情感词社区中的情感词的情感性质为其所属的评论观点标注情感性质的数据。也就是说，如果某个情感词的情感性质为正面，那么将所述情感词所属的评论观点的情感性质也标注为正面。

根据本发明的优选实施例，所述情感词词典还可以包括所述多个情感词的情感强度的数据，所述情感强度的数据是对情感性质的量化。

同样参照图 2，图中最后一竖列数据代表情感词的情感强度的数据，其数值越大则代表所述情感词的情感越强烈，例如“博学多才”这个情感词，其情感强度数据是“7”，表明“博学多才”体现的正面的情感性质比较重，再例如“前程似锦”这个情感词，其情感强度数据是“9”，根据前述“博学多才”的情感强度数据“7”要大，则表明“前程似锦”比“博学多才”表达正面的情感性质更重。

此外，还会出现一些特殊的情况，例如在同一评论观点中出现不同情感性质的情感词时，该如何为所述评论观点标注情感性质，相应地，根据本发明的示例性实施例，步骤 S150 还包括：如果任一所述评论观点包括不同情感性质的情感词，则为所述评论观点标注情感强度最强的情感词对应的情感性质的数据。

举例来说，假设某一评论观点中包括了“前程似锦”和“怨气”两个情感词，恰好“前程似锦”的情感性质是正面，而“怨气”的情感性

质是负面，那就比较二者的情感强度的大小，同样参照图 2 可知，“前程似锦”的情感强度数据是“9”，“怨气”的情感强度数据是“7”，由此可见，“前程似锦”的情感强度数据比“怨气”的情感强度数据要大，则表明所述评论观点表达正面的情感倾向最强，那么就将所述评论观点的情感性质也标注为正面，其情感性质的数据为“1”。

经过前述步骤 S110~S150 的处理，最终得到标注了情感性质的数据的评论观点。图 4 是示出本发明实施例一的评论信息的标注方法的情感聚类结果的示例图，参照图 4，同样以关于《亚太城市房地产研究院院长：楼市仍有 20 年黄金期》的事件评论为例，情感性质为正面的评论观点为一类，即图中所示的正面言论，同理，情感性质为负面的评论观点为一类，即图中所示的负面言论，情感性质为中立的评论观点为一类，即图中所示的中立言论。从而达到了一个类似于正方反方的效果，使得用户方便快速地了解舆情的大体，极大程度地提升了用户网络体验。

本发明实施例提供的评论信息的标注方法，首先，对获取到的多个事件评论进行分句得到多个评论观点，其次，以评论观点中包含的更准确反映用户意见倾向的情感词作为聚类依据，从而自动地对事件评论进行情感聚类，并为其标注情感性质，标注了情感性质的评论观点使得用户更为直观、便捷地了解其他用户关注事件热点的看法，提升了用户体验。本发明实施例适用范围广，适用于任何用户评论的归类，尤其是事件资讯类的用户评论。

实施例二

图 5 是示出本发明实施例二的评论信息的标注装置的逻辑框图。可用于执行如图 1 所示实施例的方法步骤。

参照图 5，所述评论信息的标注装置包括评论获取模块 510、评论分句模块 520、情感词提取模块 530、情感网络构建模块 540 和情感标注模块 550。

评论获取模块 510 用于获取多个事件评论的数据。

评论分句模块 520 用于分别对所述多个事件评论划分语句，并且将划分出的语句分别作为评论观点。

可选地，所述评论分句模块 520 还用于去除字数超过预定句长的语

句，和/或去除广告性质的语句。

情感词提取模块 530 用于分别从所述评论观点提取情感词。

具体地，所述情感词提取模块 530 用于分别对所述评论观点的语句进行切词，并且通过将切出的分词与预先构建的情感词词典匹配来选取
5 所述情感词，所述情感词词典包括多个情感词以及其情感性质的数据。

情感网络构建模块 540 用于将在任一评论观点中共同出现的情感词进行关联，构建情感词社区网络。

优选地，所述情感网络构建模块 540 用于对于所述情感词社区网络中的任意两个关联的情感词，计算所述两个情感词共同出现在同一评论
10 观点中的共现频度，并且如果计算的共现频度的值低于预定的共现频度阈值，则去除所述两个情感词之间的关联关系。

情感标注模块 550 用于为所述情感词社区网络中的任一情感词社区中的情感词所属的评论观点标注情感性质的数据，所述情感性质是正面、负面或中立，所述情感词社区包括一组直接或间接关联的情感词。

15 优选地，所述情感标注模块 550 用于根据所述情感词社区中的情感词的情感性质为其所属的评论观点标注情感性质的数据。

进一步地，所述情感词词典还包括所述多个情感词的情感强度的数据，所述情感标注模块 550 还用于如果任一所述评论观点包括不同情感性质的情感词，则为所述评论观点标注情感强度最强的情感词对应的情感
20 性质的数据。

本发明实施例提供的评论信息的标注装置，先对获取到的多个事件评论进行分句得到多个评论观点。进一步以评论观点中包含的情感词作为聚类依据，从而自动地对事件评论进行情感聚类，并为其标注情感性质，使得用户能够快速知晓舆情的大体，方便用户阅读，极大丰富了用
25 户体验。

实施例三

图 6 是示出本发明实施例三的计算机设备的逻辑框图。

参照图 6，所述计算机设备可用于实施上述实施例中提供的评论信息的标注方法。具体来讲：

30 计算机设备可包括输入单元 610、包括有一个或一个以上计算机可

读存储介质的存储器 620、包括有一个或者一个以上处理核心的处理器 630、显示单元 640、通信单元 650、以及电源 660 等部件。本领域技术人员可以理解，图中示出的计算机设备结构并不构成对计算机设备的限定，可以包括比图示更多或更少的部件，或者组合某些部件，或者不同的部件布置。其中：

输入单元 610 可用于接收输入的数字或字符信息，以及产生与用户设置以及功能控制有关的键盘、鼠标、操作杆、光学或者轨迹球信号输入。优选地，输入单元 610 可包括触敏表面 611 以及其他输入设备 612。触敏表面 611 也称为触摸显示屏或者触控板，可收集用户在其上或附近的触摸操作（比如用户使用手指、触笔等任何适合的物体或附件在触敏表面 611 上或在触敏表面 611 附近的操作），并根据预先设定的程式驱动相应的连接装置。可选的，触敏表面 611 可包括触摸检测装置和触摸控制器两个部分。其中，触摸检测装置检测用户的触摸方位，并检测触摸操作带来的信号，将信号传送给触摸控制器；触摸控制器从触摸检测装置上接收触摸信息，并将它转换成触点坐标，再送给处理器 630，并能接收处理器 630 发来的命令并加以执行。除了触敏表面 611，输入单元 610 还可以包括其他输入设备 612。优选地，其他输入设备 612 可以包括但不限于物理键盘、功能键（比如音量控制按键、开关按键等）、轨迹球、鼠标、操作杆等中的一种或多种。

存储器 620 可用于存储软件程序以及数据，处理器 630 通过运行存储在存储器 620 的软件程序以及数据，从而执行各种功能应用以及数据处理。存储器 620 可主要包括存储程序区和存储数据区，其中，存储程序区可存储操作系统、至少一个功能所需的应用程序（比如声音播放功能、图像播放功能等）等；存储数据区可存储根据计算机设备的使用所创建的数据（比如音频数据、电话本等）等。此外，存储器 620 可包括高速随机存取存储器，还可包括非易失性存储器，例如至少一个磁盘存储器件、闪存器件、或其他易失性固态存储器件。相应地，存储器 620 还可包括存储器控制器，以提供处理器 630 和输入单元 610 对存储器 620 的访问。

处理器 630 是计算机设备的控制中心，利用各种接口和线路连接整

个计算机设备的各个部分，通过运行或执行存储在存储器 620 内的软件程序和/或模块，以及调用存储在存储器 620 内的数据，执行计算机设备的各种功能和处理数据，从而对计算机设备进行整体监控。

显示单元 640 可用于显示由用户输入的信息或提供给用户的信息以及计算机设备的各种图形用户接口，这些图形用户接口可以由图形、文本、图标、视频和其任意组合来构成。显示单元 640 可包括显示面板，可选的，可以采用 LCD (Liquid Crystal Display, 液晶显示器)、OLED (Organic Light-Emitting Diode, 有机发光二极管) 等形式来配置显示面板。虽然在图示中，输入单元 610 与显示单元 640 是作为两个独立的部件来实现输入和输出功能，但是在某些实施例中，可将输入单元 610 与显示单元 640 集成而实现输入和输出功能。

通信单元 650 可用于收发信息或通话过程中，信号的接收和发送，通信单元 650 可以为 RF (Radio Frequency, 射频) 电路、路由器、调制解调器、等网络通信设备。此外，通信单元 650 还可通过无线通信与网络和其他设备通信。无线通信可使用任一通信标准或协议，包括但不限于 GSM (Global System of Mobile communication, 全球移动通讯系统)、GPRS (General Packet Radio Service, 通用分组无线服务)、CDMA (Code Division Multiple Access, 码分多址)、WCDMA (Wideband Code Division Multiple Access, 宽带码分多址)、LTE (Long Term Evolution, 长期演进)、电子邮件、SMS (Short Messaging Service, 短消息服务) 等。

计算机设备还可包括给各个部件供电的电源 660 (比如电池)，优选的，电源可通过电源管理系统与处理器 630 逻辑相连，从而通过电源管理系统实现管理充电、放电、以及功耗管理等功能。电源 660 还可包括一个或一个以上的直流或交流电源、再充电系统、电源故障检测电路、电源转换器或者逆变器、电源状态指示器等任意组件。

尽管未示出，计算机设备还可以包括摄像头、蓝牙模块、传感器 (比如光传感器、运动传感器以及其他传感器等)、音频电路和无线通信单元等，在此不再赘述。

在本实施例中，计算机设备包括一个或者多个处理器 630、存储器

620, 以及一个或者多个程序, 一个或者多个程序存储于存储器 620 中, 且经配置以由一个或者多个处理器 630 执行一个或者多个程序包含的用于执行评论信息的标注方法的指令: 获取多个事件评论的数据; 分别对所述多个事件评论划分语句, 并且将划分出的语句分别作为评论观点;

5 分别从所述评论观点提取情感词; 将在任一评论观点中共同出现的情感词进行关联, 构建情感词社区网络; 为所述情感词社区网络中的任一情感词社区中的情感词所属的评论观点标注情感性质的数据, 所述情感性质是正面、负面或中立, 所述情感词社区包括一组直接或间接关联的情感词。

10 此外, 所述分别从所述评论观点提取情感词的处理包括: 分别对所述评论观点的语句进行切词, 并且通过将切出的分词与预先构建的情感词词典匹配来选取所述情感词, 所述情感词词典包括多个情感词以及其情感性质的数据。

15 另外, 所述将在任一评论观点中共同出现的情感词进行关联, 构建情感词社区网络的处理还包括: 对于所述情感词社区网络中的任意两个关联的情感词, 计算所述两个情感词共同出现在同一评论观点中的共现频度, 并且如果计算的共现频度的值低于预定的共现频度阈值, 则去除所述两个情感词之间的关联关系。

20 此外, 所述为所述情感词社区网络中的任一情感词社区中的情感词所属的评论观点标注情感性质的数据的处理包括: 根据所述情感词社区中的情感词的情感性质为其所属的评论观点标注情感性质的数据。

25 进一步地, 所述情感词词典还包括所述多个情感词的情感强度的数据, 所述为所述情感词社区网络中的任一情感词社区中的情感词所属的评论观点标注情感性质的数据的处理还包括: 如果任一所述评论观点包括不同情感性质的情感词, 则为所述评论观点标注情感强度最强的情感词对应的情感性质的数据。

另外, 所述分别对所述多个事件评论划分语句的处理还包括: 去除字数超过预定句长的语句, 和/或去除广告性质的语句。

30 本发明实施例提供的计算机设备, 先对获取到的多个事件评论进行分句得到多个评论观点, 进一步以评论观点中包含的情感词作为聚类依

据，从而自动地对事件评论进行情感聚类，并为其标注情感性质，使得用户能够快速知晓舆情的大体，方便用户阅读，极大丰富了用户体验。

在本发明所提供的几个实施例中，应该理解到，所公开的装置和方法，可以通过其它的方式实现。例如，以上所描述的装置实施例仅仅是
5 示意性的，例如，所述模块的划分，仅仅为一种逻辑功能划分，实际实现时可以有另外的划分方式。

另外，在本发明各个实施例中的各功能模块可以集成在一个处理模块中，也可以是各个模块单独物理存在，也可以两个或两个以上模块集成在一个模块中。上述集成的模块既可以采用硬件的形式实现，也可以
10 采用硬件加软件功能模块的形式实现。

上述以软件功能模块的形式实现的集成的模块，可以存储在一个计算机可读取存储介质中。上述软件功能模块存储在一个存储介质中，包括若干指令用以使得一台计算机设备（可以是个人计算机，服务器，或者网络设备等）或处理器（processor）执行本发明各个实施例所述方法
15 的部分步骤。而前述的存储介质包括：U 盘、移动硬盘、只读存储器（Read-Only Memory, ROM）、随机存取存储器（Random Access Memory, RAM）、磁碟或者光盘等各种可以存储程序代码的介质。

以上所述，仅为本发明的具体实施方式，但本发明的保护范围并不局限于此，任何熟悉本技术领域的技术人员在本发明揭露的技术范围
20 内，可轻易想到变化或替换，都应涵盖在本发明的保护范围之内。因此，本发明的保护范围应以所述权利要求的保护范围为准。

权 利 要 求 书

1、一种评论信息的标注方法，其特征在于，所述方法包括：
获取多个事件评论的数据；

5 分别对所述多个事件评论划分语句，并且将划分出的语句分别作为
评论观点；

分别从所述评论观点提取情感词；

将在任一评论观点中共同出现的情感词进行关联，构建情感词社区
网络；

10 为所述情感词社区网络中的任一情感词社区中的情感词所属的评
论观点标注情感性质的数据，所述情感性质是正面、负面或中立，所述
情感词社区包括一组直接或间接关联的情感词。

2、根据权利要求 1 所述的方法，其特征在于，所述分别从所述评
论观点提取情感词的处理包括：

15 分别对所述评论观点的语句进行切词，并且通过将切出的分词与预
先构建的情感词词典匹配来选取所述情感词，所述情感词词典包括多个
情感词以及其情感性质的数据。

3、根据权利要求 1 或 2 所述的方法，其特征在于，所述将在任一
评论观点中共同出现的情感词进行关联，构建情感词社区网络的处理还
包括：

20 对于所述情感词社区网络中的任意两个关联的情感词，计算所述两
个情感词共同出现在同一评论观点中的共现频度，并且如果计算的共现
频度的值低于预定的共现频度阈值，则去除所述两个情感词之间的关联
关系。

25 4、根据权利要求 1~3 中任一项所述的方法，其特征在于，所述为
所述情感词社区网络中的任一情感词社区中的情感词所属的评论观点
标注情感性质的数据的处理包括：

根据所述情感词社区中的情感词的情感性质为其所属的评论观点
标注情感性质的数据。

30 5、根据权利要求 2~4 中任一项所述的方法，其特征在于，所述情
感词词典还包括所述多个情感词的情感强度的数据，

所述为所述情感词社区网络中的任一情感词社区中的情感词所属的评论观点标注情感性质的数据的处理还包括：

如果任一所述评论观点包括不同情感性质的情感词，则为所述评论观点标注情感强度最强的情感词对应的情感性质的数据。

- 5 6、根据权利要求 1~5 中任一项所述的方法，其特征在于，所述分别对所述多个事件评论划分语句的处理还包括：

去除字数超过预定句长的语句，和/或

去除广告性质的语句。

7、一种评论信息的标注装置，其特征在于，所述装置包括：

- 10 评论获取模块，用于获取多个事件评论的数据；

评论分句模块，用于分别对所述多个事件评论划分语句，并且将划分出的语句分别作为评论观点；

情感词提取模块，用于分别从所述评论观点提取情感词；

- 15 情感网络构建模块，用于将在任一评论观点中共同出现的情感词进行关联，构建情感词社区网络；

情感标注模块，用于为所述情感词社区网络中的任一情感词社区中的情感词所属的评论观点标注情感性质的数据，所述情感性质是正面、负面或中立，所述情感词社区包括一组直接或间接关联的情感词。

- 20 8、根据权利要求 7 所述的装置，其特征在于，所述情感词提取模块用于分别对所述评论观点的语句进行切词，并且通过将切出的分词与预先构建的情感词词典匹配来选取所述情感词，所述情感词词典包括多个情感词以及其情感性质的数据。

- 25 9、根据权利要求 7 或 8 所述的装置，其特征在于，所述情感网络构建模块还用于对于所述情感词社区网络中的任意两个关联的情感词，计算所述两个情感词共同出现在同一评论观点中的共现频度，并且如果计算的共现频度的值低于预定的共现频度阈值，则去除所述两个情感词之间的关联关系。

- 30 10、根据权利要求 7~9 中任一项所述的装置，其特征在于，所述情感标注模块用于根据所述情感词社区中的情感词的情感性质为其所属的评论观点标注情感性质的数据。

11、根据权利要求 8~10 中任一项所述的装置，其特征在于，所述情感词词典还包括所述多个情感词的情感强度的数据，所述情感标注模块还用于如果任一所述评论观点包括不同情感性质的情感词，则为所述评论观点标注情感强度最强的情感词对应的情感性质的数据。

5 12、根据权利要求 7~11 中任一项所述的装置，其特征在于，所述评论分句模块还用于去除字数超过预定句长的语句，和/或去除广告性质的语句。

13、一种计算机设备，其特征在于，所述计算机设备包括：

一个或多个处理器；

10 存储器；

一个或多个程序，所述一个或多个程序存储在所述存储器中，且经配置以由所述一个或者多个处理器执行所述一个或者多个程序包含的用于执行所述评论信息的标注方法的指令：

获取多个事件评论的数据；

15 分别对所述多个事件评论划分语句，并且将划分出的语句分别作为评论观点；

分别从所述评论观点提取情感词；

将在任一评论观点中共同出现的情感词进行关联，构建情感词社区网络；

20 为所述情感词社区网络中的任一情感词社区中的情感词所属的评论观点标注情感性质的数据，所述情感性质是正面、负面或中立，所述情感词社区包括一组直接或间接关联的情感词。

1/5

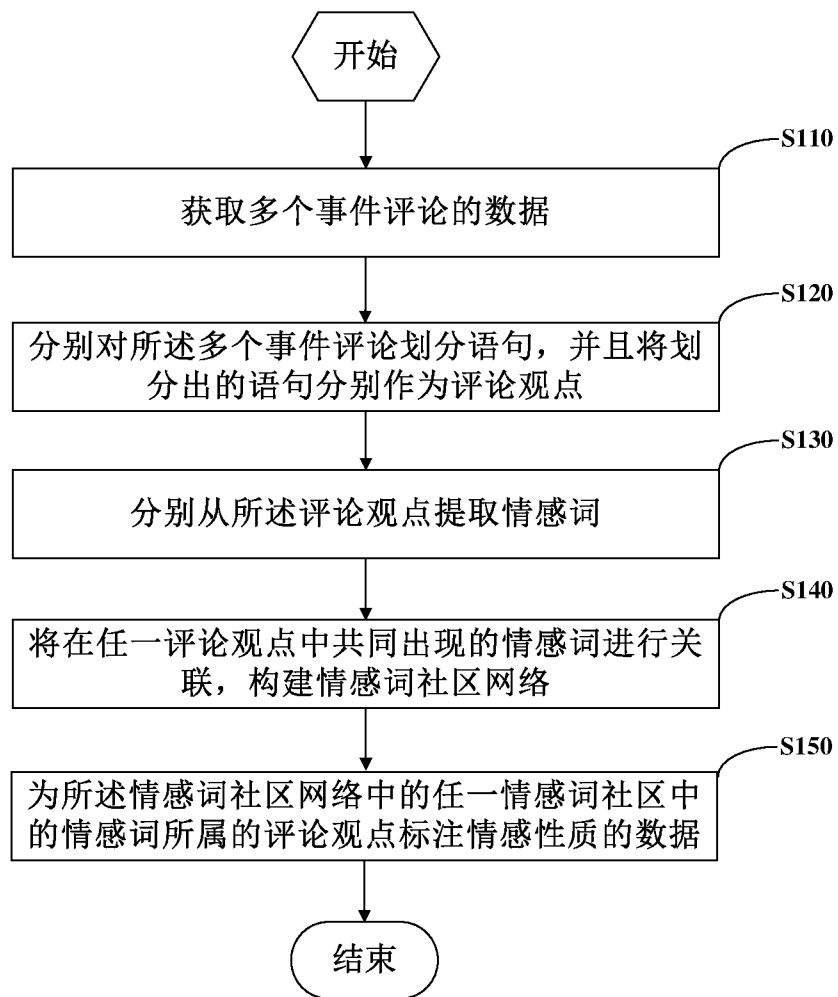


图 1

情感词	情感性质	情感强度
博学多才	1	7
毒谋	-1	7
乱丢垃圾	-1	
进取	1	
不思进取	-1	
进取心	1	
冤气	-1	7
猛然	0	5
前程似锦	1	9
乌尽弓藏	-1	5

图 2

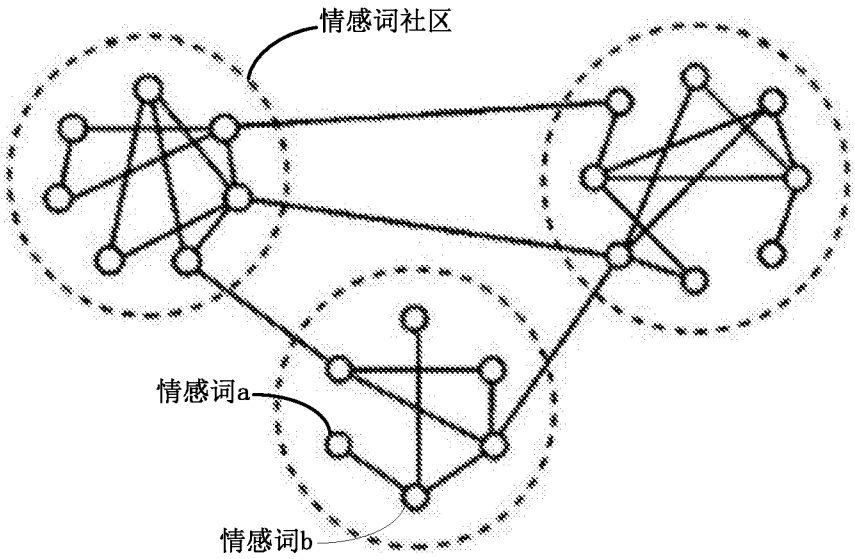


图 3



图 4

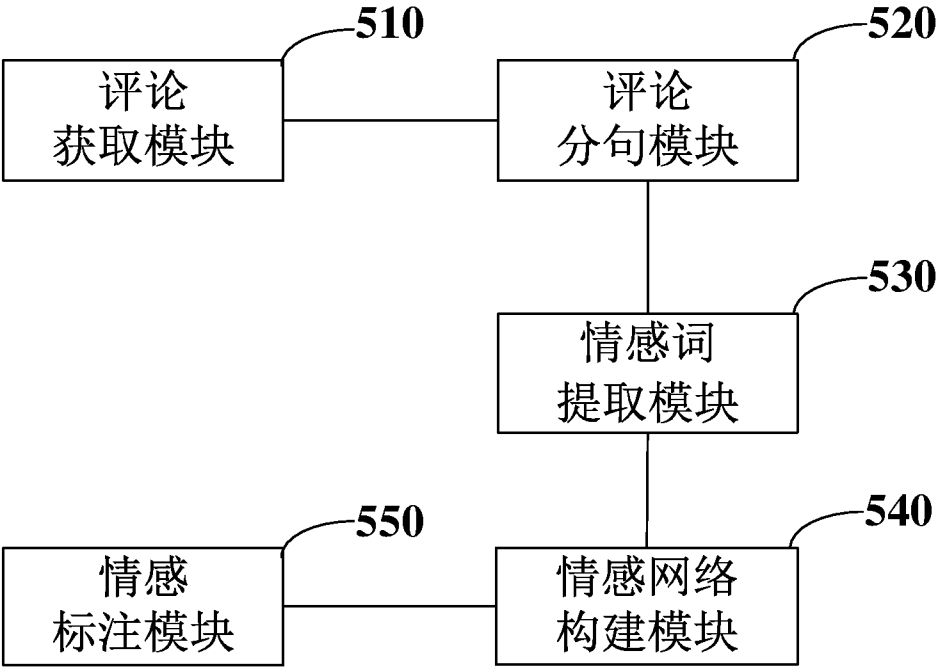


图 5

5/5

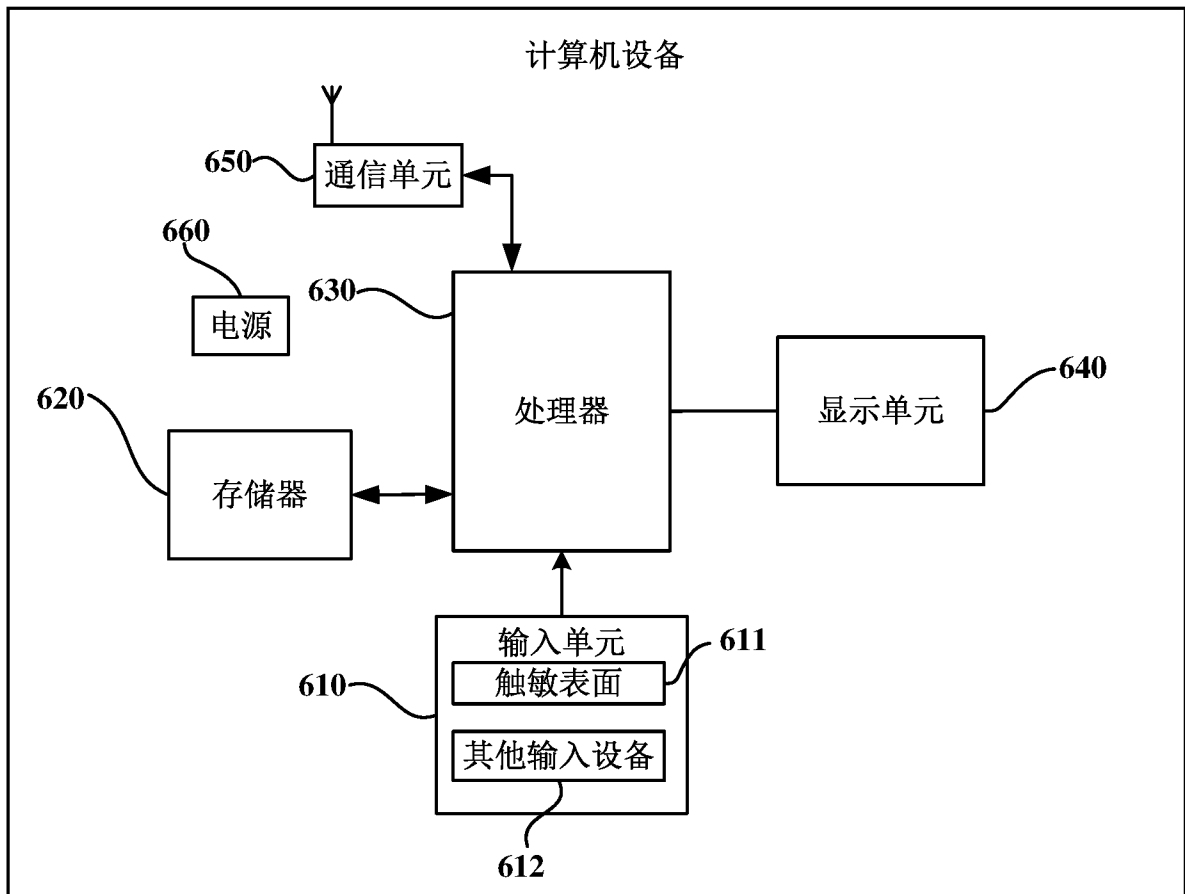


图 6

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2015/097774

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)

G06F

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)

CNPAT, WPI, EPODOC, CNKI, GOOGLE: emotional word, mark, comment information, extract, relevance, network, emotional property, propensity

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 104933130 A (BAIDU ONLINE NETWORK TECHNOLOGY (BEIJING) CO., LTD.), 23 September 2015 (23.09.2015), claims 1-12, description, paragraphs [0074]-[0075], and figure 1	1-13
Y	CN 101609459 A (PEKING UNIVERSITY), 23 December 2009 (23.12.2009), description, page 4, paragraph 2 to page 7, paragraph 3, and figures 1-3	1-13
Y	CN 102375838 A (FUJITSU LIMITED), 14 March 2012 (14.03.2012), description, paragraphs [0048]-[0071]	1-13
Y	CN 102999485 A (BEIJING UNIVERSITY OF POSTS AND TELECOMMUNICATIONS), 27 March 2013 (27.03.2013), description, paragraphs [0018]-[0019]	2-6, 8-12
A	CN 101408883 A (UNIVERSITY OF ELECTRONIC SCIENCE AND TECHNOLOGY OF CHINA), 15 April 2009 (15.04.2009), the whole document	1-13
A	CN 103699626 A (SOUTH CHINA UNIVERSITY OF TECHNOLOGY), 02 April 2014 (02.04.2014), the whole document	1-13
A	US 2011131485 A1 (INTERNATIONAL BUSINESS MACHINES CORPORATION), 02 June 2011 (02.06.2011), the whole document	1-13

☒ 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 24 February 2016 (24.02.2016)	Date of mailing of the international search report 15 March 2016 (15.03.2016)
Name and mailing address of the ISA/CN: 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 ZHANG, Xia Telephone No.: (86-10) 61648105

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2015/097774

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 2007294230 A1 (SINEL, J. et al.), 20 December 2007 (20.12.2007), the whole document	1-13

International application No.
PCT/CN2015/097774

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

国际检索报告

国际申请号

PCT/CN2015/097774

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)

G06F

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))

CNPAT, WPI, EPDOC, CNKI, GOOGLE: 情感词, 标注, 评论信息, 抽取, 相关性, 网络, 情感性质, 倾向, emotional word, mark, comment information, extract, relevance, network, emotional property, propensity

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
PX	CN 104933130 A (百度在线网络技术北京有限公司) 2015年 9月 23日 (2015 - 09 - 23) 权利要求1-12, 说明书第[0074]-[0075]段, 图1	1-13
Y	CN 101609459 A (北京大学) 2009年 12月 23日 (2009 - 12 - 23) 说明书第4页第2段-第7页第3段, 图1-3	1-13
Y	CN 102375838 A (富士通株式会社) 2012年 3月 14日 (2012 - 03 - 14) 说明书第[0048]-[0071]段	1-13
Y	CN 102999485 A (北京邮电大学) 2013年 3月 27日 (2013 - 03 - 27) 说明书第[0018]-[0019]段	2-6, 8-12
A	CN 101408883 A (电子科技大学) 2009年 4月 15日 (2009 - 04 - 15) 全文	1-13
A	CN 103699626 A (华南理工大学) 2014年 4月 2日 (2014 - 04 - 02) 全文	1-13
A	US 2011131485 A1 (INTERNATIONAL BUSINESS MACHINES CORPORATION) 2011年 6月 2日 (2011 - 06 - 02) 全文	1-13

☒ 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

2016年 2月 24日

国际检索报告邮寄日期

2016年 3月 15日

ISA/CN的名称和邮寄地址

中华人民共和国国家知识产权局(ISA/CN)
中国北京市海淀区蓟门桥西土城路6号 100088

授权官员

张霞

传真号 (86-10)62019451

电话号码 (86-10)61648105

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
A	US 2007294230 A1 (SINEL, JOSHUA 等) 2007年 12月 20日 (2007 - 12 - 20) 全文	1-13

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2015/097774

检索报告引用的专利文件			公布日 (年/月/日)	同族专利			公布日 (年/月/日)
CN	104933130	A	2015年 9月 23日	无			
CN	101609459	A	2009年 12月 23日	无			
CN	102375838	A	2012年 3月 14日	无			
CN	102999485	A	2013年 3月 27日	无			
CN	101408883	A	2009年 4月 15日	无			
CN	103699626	A	2014年 4月 2日	无			
US	2011131485	A1	2011年 6月 2日	CN	102486771	A	2012年 6月 6日
US	2007294230	A1	2007年 12月 20日	WO	2007142998	A2	2007年 12月 13日

表 PCT/ISA/210 (同族专利附件) (2009年7月)