

(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 11 月 17 日 (17.11.2016)

WIPO | PCT

(10) 国际公布号
WO 2016/180268 A1

- (51) 国际专利分类号:
G06F 17/30 (2006.01)
- (21) 国际申请号: PCT/CN2016/081090
- (22) 国际申请日: 2016 年 5 月 5 日 (05.05.2016)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
201510242860.0 2015 年 5 月 13 日 (13.05.2015) CN
- (71) 申请人: 阿里巴巴集团控股有限公司 (ALIBABA GROUP HOLDING LIMITED) [—/CN]; 英属开曼群岛大开曼资本大厦一座四层 847 号邮箱, Grand Cayman (KY)。
- (72) 发明人: 冯文镛 (FENG, Wenyong); 中国浙江省杭州市余杭区文一西路 969 号 3 号楼 5 楼阿里巴巴集团法务部, Zhejiang 311121 (CN)。
- (74) 代理人: 北京三友知识产权代理有限公司 (BEIJING SANYOU INTELLECTUAL PROPERTY AGENCY LTD.); 中国北京市金融街 35 号国际企业大厦 A 座 16 层, Beijing 100033 (CN)。

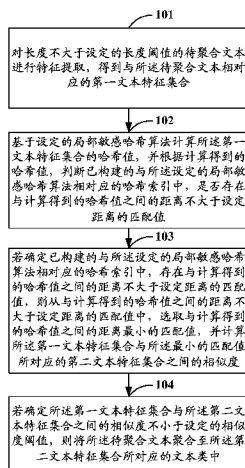
- (81) 指定国 (除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, 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, 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))。

(54) Title: TEXT AGGREGATE METHOD AND DEVICE

(54) 发明名称: 一种文本聚合方法及装置



101 PERFORM FEATURE EXTRACTION ON A TEXT TO BE AGGREGATED HAVING A LENGTH LESS THAN A PRESET LENGTH THRESHOLD, AND ACQUIRE A FIRST TEXT FEATURE SET CORRESPONDING TO THE TEXT TO BE AGGREGATED

102 ON THE BASIS OF A PRESET LOCALITY SENSITIVE HASHING ALGORITHM, CALCULATE A HASH VALUE OF THE FIRST TEXT FEATURE SET, AND ACCORDING TO THE CALCULATED HASH VALUE, DETERMINE WHETHER A MATCHING VALUE HAVING A DISTANCE FROM THE CALCULATED HASH VALUE NOT GREATER THAN A PRESET DISTANCE EXISTS IN DETERMINED HASH INDEXES CORRESPONDING TO THE PRESET LOCALITY SENSITIVE HASHING ALGORITHM

103 IF SO, SELECT A MATCHING VALUE HAVING A SHORTEST DISTANCE FROM THE CALCULATED HASH VALUE FROM MATCHING VALUES HAVING DISTANCES FROM THE CALCULATED HASH VALUE NOT GREATER THAN THE PRESET DISTANCE, AND CALCULATE SIMILARITY BETWEEN THE FIRST TEXT FEATURE SET AND A SECOND TEXT FEATURE SET CORRESPONDING TO THE MINIMUM MATCHING VALUE

104 IF THE SIMILARITY BETWEEN THE FIRST TEXT FEATURE SET AND THE SECOND TEXT FEATURE SET IS DETERMINED TO BE NOT LESS THAN A PRESET SIMILARITY THRESHOLD, AGGREGATE THE TEXT TO BE AGGREGATED TO A TEXT TYPE CORRESPONDING TO THE SECOND TEXT FEATURE SET

图 1

(57) Abstract: A text aggregate method and device. The method comprises: when a text feature set corresponding to a text to be aggregated is acquired, employing a determination method combining a locality sensitive hashing algorithm and a similarity check to perform a similarity analysis on the text to be aggregated so as to aggregate the text to be aggregated. Therefore, the present invention can address the problem of a low-accuracy and high-latency text aggregate result due to performing a short-text similarity analysis on the basis of a vector space model or a probabilistic model, thus accurately and quickly aggregating a short text.

(57) 摘要: 一种文本聚合方法及装置, 在得到与待聚合文本相对应的文本特征集合之后, 可采用局部敏感哈希算法结合相似度校验的判定方法, 对所述待聚合文本进行相似性分析以实现待聚合文本的聚合, 从而可解决基于向量空间模型或概率模型进行短文本相似性分析时所导致的文本聚合结果准确性较低、实时性较低的问题, 达到准确且又快速地对短文本进行聚合的效果。

WO 2016/180268 A1

一种文本聚合方法及装置

本申请要求 2015 年 05 月 13 日递交的申请号为 201510242860.0、发明名称为“一种文本聚合方法及装置”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。

5 技术领域

本申请涉及互联网技术领域，尤其涉及一种文本聚合方法及装置。

背景技术

在传统的通信应用（如短信、邮件等）以及新型的互联网社交应用（如微信、微博、论坛等）等场景中，时刻都会产生大量的短文本数据，如，长度不大于设定的长度阈值（如 150~200 个字等，其中，英文单词或者连续数字按一个汉字计算）的中文文本数据。这些文本数据中存在大量有价值的信息，通过对其进行聚合可以发现信息中潜在的热点或者规律。

具体地，文本聚合是一种在给定的相似性度量之下对文本集合进行分组，使彼此相近的文本分到同一个组内的技术。文本聚合具体可包括文本特征提取以及文本相似性分析等步骤。

具体地，由于目前，在对文本进行相似性分析以实现文本的聚合时，主要基于向量空间模型或概率模型进行。而在向量空间模型中，是采用文本中的字或者词作为特征表示文本，用特征向量之间的相似度来度量文本的相关性。因而，对于长度过短的文本，会存在特征向量过于稀疏，导致计算结果无法满足相似性分析的要求，进而导致最终所得到的文本聚合结果并不准确的问题。另外，在概率模型中，若使用过短的文本，则大部分特征都会是概率平滑的结果，不能反映真实数据的信息，因而，也会存在聚合结果并不准确、无法满足用户需求的问题。再有，由于上述两类传统的文本相似度算法计算量巨大，因而，还会存在难以满足通常可以达到千万级甚至亿级的短文本数据的实时分析的问题，使得文本聚合的效果并不佳。

也就是说，目前，在对短文本数据进行文本聚合时，存在文本相似性分析的方式较差所导致的文本聚合的准确性较低、实时性较低的问题，因此，亟需提供一种新的文本聚合方法以解决上述问题。

30 发明内容

本申请实施例提供了一种文本聚合方法及装置，用以解决目前的文本聚合方式存在文本相似性分析的方式较差所导致的文本聚合的准确性较低、实时性较低的问题。

本申请实施例提供了一种文本聚合方法，包括：

对长度不大于设定的长度阈值的待聚合文本进行特征提取，得到与所述待聚合文本
5 相对应的第一文本特征集合；

基于设定的局部敏感哈希算法计算所述第一文本特征集合的哈希值，并根据计算得到的哈希值，判断已构建的与所述设定的局部敏感哈希算法相对应的哈希索引中，是否存在与计算得到的哈希值之间的距离不大于设定距离的匹配值；

若是，则从与计算得到的哈希值之间的距离不大于设定距离的匹配值中，选取与计
10 算得到的哈希值之间的距离最小的匹配值，并计算所述第一文本特征集合与所述最小的匹配值所对应的第二文本特征集合之间的相似度；

若确定所述第一文本特征集合与所述第二文本特征集合之间的相似度不小于设定的相似度阈值，则将所述待聚合文本聚合至所述第二文本特征集合所对应的文本类中。

相应地，本申请实施例还提供了一种文本聚合装置，包括：

15 特征提取单元，用于对长度不大于设定的长度阈值的待聚合文本进行特征提取，得到与所述待聚合文本相对应的第一文本特征集合；

文本聚合单元，用于基于设定的局部敏感哈希算法计算所述第一文本特征集合的哈希值，并根据计算得到的哈希值，判断已构建的与所述设定的局部敏感哈希算法相对应的哈希索引中，是否存在与计算得到的哈希值之间的距离不大于设定距离的匹配值；若是，则从与计算得到的哈希值之间的距离不大于设定距离的匹配值中，选取与计算得到的哈希值之间的距离最小的匹配值，并计算所述第一文本特征集合与所述最小的匹配值所对应的第二文本特征集合之间的相似度；以及，若确定所述第一文本特征集合与所述第二文本特征集合之间的相似度不小于设定的相似度阈值，则将所述待聚合文本聚合至所述第二文本特征集合所对应的文本类中。
20

25 本申请有益效果如下：

本申请实施例提供了一种文本聚合方法及装置，在本申请实施例所述技术方案中，可在得到与待聚合文本相对应的文本特征集合之后，采用局部敏感哈希算法结合相似度校验的判定方法，对所述待聚合文本进行相似性分析以实现待聚合文本的聚合，从而可解决基于向量空间模型或概率模型进行短文本相似性分析时所导致的文本聚合结果准确性
30 性较低、实时性较低的问题，达到准确且又快速地对短文本进行聚合的效果。

附图说明

为了更清楚地说明本申请实施例中的技术方案，下面将对实施例描述中所需要使用的附图作简要介绍，显而易见地，下面描述中的附图仅仅是本申请的一些实施例，对于本领域的普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

图 1 所示为本申请实施例一中所述文本聚合方法的流程示意图；

图 2 所示为本申请实施例二中所述文本聚合装置的结构示意图。

10 具体实施方式

为了使本申请的目的、技术方案和优点更加清楚，下面将结合附图对本申请作进一步地详细描述，显然，所描述的实施例仅仅是本申请一部分实施例，而不是全部的实施例。基于本申请中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其它实施例，都属于本申请保护的范围。

15 实施例一：

本申请实施例一提供了一种文本聚合方法，如图 1 所示，其为本申请实施例一中所述文本聚合方法的流程示意图，所述文本聚合方法可包括以下步骤：

步骤 101：对长度不大于设定的长度阈值的待聚合文本进行特征提取，得到与所述待聚合文本相对应的第一文本特征集合。

20 可选地，所述待聚合文本具体可为长度不大于设定的长度阈值（如 150~200 个字等，其中，英文单词或者连续数字按一个汉字计算）的中文文本数据，本申请实施例对此不作赘述。

进一步地，由于互联网上的大量短文本数据存在用词不规范、存在各种变形等特点，因而使得，在使用传统的分词方法对其进行特征提取（如利用普通的分词器进行分词，并将相应的分词结果作为文本的特征描述）时，可能存在无法获得较好的特征提取结果，25 进而导致最终所得到的文本聚合结果并不准确的问题。

因而，为了提高文本特征的提取效果，在本申请所述实施例中，可采用以下方式对长度不大于设定的长度阈值的待聚合文本进行特征提取，得到与所述待聚合文本相对应的文本特征集合：

30 基于机械分词结合 N 元模型（N-gram）的特征提取方式对长度不大于设定的长度阈

值的待聚合文本进行特征提取，得到与所述待聚合文本相对应的第一文本特征集合，所述 N 为大于 1 的自然数。

需要说明的是，相对于采用传统的分词方法对短文本数据进行特征提取来说，采用机械分词结合 N 元模型的特征提取方式可达到较好的文本特征提取效果。这是因为，机械分词是忽略语意对文本进行机械地分割，而 N 元模型则是给孤立的特征之间建立了一定的依赖性，从而能够提供更大的特征集合，丰富了特征集合的信息，这对本身信息就较少的短文本来说起到了很好的补充作用，因而，可在不规范的短文本特征提取中取得良好的效果，进而提高文本聚合的准确性。

可选地，基于机械分词结合 N 元模型的特征提取方式对长度不大于设定的长度阈值的待聚合文本进行特征提取，得到与所述待聚合文本相对应的文本特征集合，可包括：

以中文汉字以及连续的字符串（如连续的拉丁文字符串、连续的数字串、或连续的拉丁文数字字符串等）为最小切分单元，对所述待聚合文本进行分词，得到多个分词；例如，以待聚合文本为“我的生日是 1989-01-22”为例，可将所述待聚合文本分词为“我 / 的 / 生 / 日 / 是 / 1989-01-22”；

基于 N 元模型，将得到的多个分词中的任意 N 个连续的分词组合为一文本特征，得到与所述待聚合文本相对应的文本特征集合。例如，以所述 N 的取值为 2（即所述 N 元模型为 Bi-gram），且待聚合文本为“我的生日是 1989-01-22”为例，最终所得到的与所述待聚合文本相对应的文本特征集合可表示为{我的，的生，生日，日是，是 1989-01-22}。

进一步地，为了提高文本质量，进而提高文本聚合的准确性，在对长度不大于设定的长度阈值的待聚合文本进行特征提取之前，所述方法还可包括以下步骤：

对所述待聚合文本进行预处理，以便后续可依据预处理后的待聚合文本进行相应的文本特征提取；其中，所述预处理至少可包括以下操作中的任意一种或多种，本申请实施例对此不作任何限定：

去除待聚合文本中的特殊标签（如 html 标签等）、去除待聚合文本中的非文字特殊符号（如&、*等）、对待聚合文本进行繁简字体转换（如将待聚合文本中的繁体字转换为简体字等）、以及将待聚合文本中的连续性的拉丁文和/或数字归一化为设定的字符串（如，将“Abc1234”或“1989-01-22”归一化成“xxxxxxx”等）等。

步骤 102：基于设定的局部敏感哈希算法计算所述第一文本特征集合的哈希值，并根据计算得到的哈希值，判断已构建的与所述设定的局部敏感哈希算法相对应的哈希索引中，是否存在与计算得到的哈希值之间的距离不大于设定距离的匹配值。

具体地，所述设定的局部敏感哈希算法不限于为 Simhash 算法或 Minhash 算法等。其中，Simhash 算法是一种用来对网页去重的常用方法，其通过对网页的内容生成一个数字签名，然后通过计算数字签名之间的差异来判定网页内容的相似程度。另外，与 Simhash 算法一样，Minhash 算法也是局部敏感哈希算法的一种，可以用来快速估算两个集合的相似度，最初用于在搜索引擎中检测重复网页，当然也可以应用于大规模聚类问题等，本申请实施例对此均不作赘述。

优选地，由于 Simhash 算法的速度较快，因此，在本申请所述实施例中，可优先选用所述 Simhash 算法来计算第一文本特征集合的哈希值。相应地，以所述设定的局部敏感哈希算法为 Simhash 算法为例，步骤 102 可具体执行为：基于 Simhash 算法计算所述第一文本特征集合的 Simhash 值，并根据计算得到的 Simhash 值，判断已构建的 Simhash 索引中，是否存在与计算得到的 Simhash 值之间的距离（具体可为海明距离，即 Hamming 距离）不大于设定距离的匹配值。

其中，所述设定距离可根据实际情况灵活设定，如以海明距离为例，可设置为 3~5 等，本申请实施例对此不作赘述。另外，需要说明的是，在信息论中，两个等长字符串之间的海明距离是指两个字符串对应位置的不同字符的个数，即，将一个字符串变换成另外一个字符串所需要替换的字符个数，本申请实施例对此也不作赘述。

步骤 103：若确定已构建的与所述设定的局部敏感哈希算法相对应的哈希索引中，存在与计算得到的哈希值之间的距离不大于设定距离的匹配值，则从与计算得到的哈希值之间的距离不大于设定距离的匹配值中，选取与计算得到的哈希值之间的距离最小的匹配值，并计算所述第一文本特征集合与所述最小的匹配值所对应的第二文本特征集合之间的相似度。

可选地，所述第一文本特征集合与所述第二文本特征集合之间的相似度至少可通过以下任意一种或多种相似度度量参数来表示：Jaccard 相似度、欧式距离以及海明距离等。也就是说，在计算所述第一文本特征集合与所述最小的匹配值所对应的第二文本特征集合之间的相似度时，可计算所述第一文本特征集合与所述第二文本特征集合之间的 Jaccard 相似度、欧式距离以及海明距离等，本申请实施例对此不作赘述。

步骤 104：若确定所述第一文本特征集合与所述第二文本特征集合之间的相似度不小于设定的相似度阈值，则将所述待聚合文本聚合至所述第二文本特征集合所对应的文本类中。

其中，所述设定的相似度阈值可根据实际情况灵活设定，如，当对文本聚合的准确

性要求较高时，可将所述相似度阈值设置为一个相对较高的数值，当对文本聚合的准确性要求较低时，可将所述相似度阈值设置为一个相对较低的数值等，本申请实施例对此不作赘述。

需要说明的是，在本申请所述实施例中，之所以对所述第一文本特征集合与所述第二文本特征集合之间的相似度进行校验，主要是为了消除将局部敏感哈希算法应用于短文本数据的聚合时，局部敏感哈希算法的碰撞概率所导致的误判现象，以提高文本聚合的准确性。

例如，以采用 Simhash 算法计算第一文本特征集合的哈希值、进而选取相应的匹配值为例，在采用 Simhash 算法计算第一文本特征集合的哈希值、进而选取相应的匹配值之后，可进一步对所述第一文本特征集合与选取的匹配值所对应的第二文本特征集合之间的相似度（如 Jaccard 相似度等）进行校验，以消除 Simhash 碰撞导致的误判问题。

需要说明的是，Jaccard 相似度是最常见的衡量两个集合相似性的一种方法，其也很适合用于衡量短文本的相似性，但由于计算量过大，所以无法直接用于大数据量的文本聚合。但是，通过 Jaccard 相似度校验，却可以完全解决 Simhash 算法的碰撞问题，消除了 Simhash 碰撞导致的误判问题。因而，采用 Simhash 算法结合 Jaccard 相似度校验的判定方法对待聚合文本进行相似性分析时，可达到准确且又快速地对短文本进行聚合的效果。

进一步地，在本申请所述实施例中，所述方法还可包括以下步骤：

若确定已构建的与所述设定的局部敏感哈希算法相对应的哈希索引中，不存在与计算得到的哈希值之间的距离不大于设定距离的匹配值；或者，确定已构建的与所述设定的局部敏感哈希算法相对应的哈希索引中，存在与计算得到的哈希值之间的距离不大于设定距离的匹配值、且确定所述第一文本特征集合与所述第二文本特征集合之间的相似度小于设定的相似度阈值；则将计算得到的哈希值更新至（即添加至）已构建的与所述设定的局部敏感哈希算法相对应的哈希索引中，并基于所述待聚合文本创建一个新的文本类，以及将所述待聚合文本归至创建的所述新的文本类中。

也就是说，若确定待聚合文本不归属于任何一个已创建的文本类时，可将所述待聚合文本对应的哈希值添加至相应的哈希索引中，并将所述待聚合文本归至一个新创建的文本类中，本申请实施例对此不作赘述。

进一步地，需要说明的是，本申请实施例所述方案无语言、软件或者硬件的限制。但是，为了提高文本聚合的效率，可优先选用性能高的编程语言（如 C++ 或者 Java 等）

和性能高的硬件等来实现，本申请实施例对此不作赘述。

本申请实施例一提供了一种文本聚合方法，在本申请实施例一所述技术方案中，可对长度不大于设定的长度阈值的待聚合文本进行特征提取，并在得到与所述待聚合文本相对应的文本特征集合之后，可采用局部敏感哈希算法结合相似度校验的判定方法，对所述待聚合文本进行相似性分析以实现待聚合文本的聚合，从而可在解决基于向量空间模型或概率模型进行短文本相似性分析时所导致的文本聚合结果准确性较低、实时性较低的问题，达到准确且又快速地对短文本进行聚合的效果，如可实现大数据流量（如大于 1 万条 / 秒等）下的短文本的实时聚合，以支持对数据流的实时分析。

实施例二：

基于同一发明构思，本申请实施例二提供了一种文本聚合装置，该文本聚合装置的具体实施可参见上述方法实施例一中的相关描述，重复之处不再赘述，如图 2 所示，该文本聚合装置主要可包括：

特征提取单元 21，可用于对长度不大于设定的长度阈值的待聚合文本进行特征提取，得到与所述待聚合文本相对应的第一文本特征集合；

文本聚合单元 22，可用于基于设定的局部敏感哈希算法计算所述第一文本特征集合的哈希值，并根据计算得到的哈希值，判断已构建的与所述设定的局部敏感哈希算法相对应的哈希索引中，是否存在与计算得到的哈希值之间的距离不大于设定距离的匹配值；若是，则从与计算得到的哈希值之间的距离不大于设定距离的匹配值中，选取与计算得到的哈希值之间的距离最小的匹配值，并计算所述第一文本特征集合与所述最小的匹配值所对应的第二文本特征集合之间的相似度；以及，若确定所述第一文本特征集合与所述第二文本特征集合之间的相似度不小于设定的相似度阈值，则将所述待聚合文本聚合至所述第二文本特征集合所对应的文本类中。

其中，所述设定的局部敏感哈希算法不限于为 Simhash 算法或 Minhash 算法等。且，所述第一文本特征集合与所述第二文本特征集合之间的相似度至少可通过以下任意一种或多种相似度度量参数来表示：Jaccard 相似度、欧式距离以及海明距离等。

进一步地，所述文本聚合单元 22，还可用于若确定已构建的与所述设定的局部敏感哈希算法相对应的哈希索引中，不存在与计算得到的哈希值之间的距离不大于设定距离的匹配值；或者，确定已构建的与所述设定的局部敏感哈希算法相对应的哈希索引中，存在与计算得到的哈希值之间的距离不大于设定距离的匹配值、且确定所述第一文本特征集合与所述第二文本特征集合之间的相似度小于设定的相似度阈值；则将计算得到的

哈希值更新至已构建的与所述设定的局部敏感哈希算法相对应的哈希索引中，并基于所述待聚合文本创建一个新的文本类，以及将所述待聚合文本归至创建的所述新的文本类中。

进一步地，为了提高文本特征的提取效果，在本申请所述实施例中，所述特征提取单元 21 具体可用于基于机械分词结合 N 元模型的特征提取方式对长度不大于设定的长度阈值的待聚合文本进行特征提取，得到与所述待聚合文本相对应的第一文本特征集合，所述 N 为大于 1 的自然数。

可选地，所述特征提取单元 21 具体可用于以中文汉字以及连续的字符串为最小切分单元，对所述待聚合文本进行分词，得到多个分词；并基于 N 元模型，将得到的多个分词中的任意 N 个连续的分词组合为一文本特征，得到与所述待聚合文本相对应的文本特征集合。

进一步地，所述装置还可包括预处理单元 23：

所述预处理单元 23，可用于在对长度不大于设定的长度阈值的待聚合文本进行特征提取之前，对所述待聚合文本进行预处理；其中，所述预处理至少可包括：去除待聚合文本中的特殊标签、去除待聚合文本中的非文字特殊符号、对待聚合文本进行繁简字体转换、以及将待聚合文本中的连续性的拉丁文和/或数字归一化为设定的字符串等中的一种或多种。

本领域技术人员应明白，本申请的实施例可提供为方法、装置（设备）、或计算机程序产品。因此，本申请可采用完全硬件实施例、完全软件实施例、或结合软件和硬件方面的实施例的形式。而且，本申请可采用在一个或多个其中包含有计算机可用程序代码的计算机可用存储介质（包括但不限于磁盘存储器、CD-ROM、光学存储器等）上实施的计算机程序产品的形式。

本申请是参照根据本申请实施例的方法、装置（设备）和计算机程序产品的流程图和/或方框图来描述的。应理解可由计算机程序指令实现流程图和/或方框图中的每一流程和/或方框、以及流程图和/或方框图中的流程和/或方框的结合。可提供这些计算机程序指令到通用计算机、专用计算机、嵌入式处理机或其他可编程数据处理设备的处理器以产生一个机器，使得通过计算机或其他可编程数据处理设备的处理器执行的指令产生用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的装置。

这些计算机程序指令也可存储在能引导计算机或其他可编程数据处理设备以特定方

式工作的计算机可读存储器中，使得存储在该计算机可读存储器中的指令产生包括指令装置的制造品，该指令装置实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能。

5 这些计算机程序指令也可装载到计算机或其他可编程数据处理设备上，使得在计算机或其他可编程设备上执行一系列操作步骤以产生计算机实现的处理，从而在计算机或其他可编程设备上执行的指令提供用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的步骤。

10 尽管已描述了本申请的优选实施例，但本领域内的技术人员一旦得知了基本创造性概念，则可对这些实施例作出另外的变更和修改。所以，所附权利要求意欲解释为包括优选实施例以及落入本申请范围的所有变更和修改。

显然，本领域的技术人员可以对本申请进行各种改动和变型而不脱离本申请的精神和范围。这样，倘若本申请的这些修改和变型属于本申请权利要求及其等同技术的范围之内，则本申请也意图包含这些改动和变型在内。

权利要求书

1、一种文本聚合方法，其特征在于，包括：

对长度不大于设定的长度阈值的待聚合文本进行特征提取，得到与所述待聚合文本相对应的第一文本特征集合；

5 基于设定的局部敏感哈希算法计算所述第一文本特征集合的哈希值，并根据计算得到的哈希值，判断已构建的与所述设定的局部敏感哈希算法相对应的哈希索引中，是否存在与计算得到的哈希值之间的距离不大于设定距离的匹配值；

若是，则从与计算得到的哈希值之间的距离不大于设定距离的匹配值中，选取与计算得到的哈希值之间的距离最小的匹配值，并计算所述第一文本特征集合与所述最小的
10 匹配值所对应的第二文本特征集合之间的相似度；

若确定所述第一文本特征集合与所述第二文本特征集合之间的相似度不小于设定的相似度阈值，则将所述待聚合文本聚合至所述第二文本特征集合所对应的文本类中。

2、如权利要求 1 所述的方法，其特征在于，所述方法还包括：

若确定已构建的与所述设定的局部敏感哈希算法相对应的哈希索引中，不存在与计
15 算得到的哈希值之间的距离不大于设定距离的匹配值；或者，确定已构建的与所述设定的局部敏感哈希算法相对应的哈希索引中，存在与计算得到的哈希值之间的距离不大于设定距离的匹配值、且确定所述第一文本特征集合与所述第二文本特征集合之间的相似度小于设定的相似度阈值，则

将计算得到的哈希值更新至已构建的与所述设定的局部敏感哈希算法相对应的哈
20 希索引中，并基于所述待聚合文本创建一个新的文本类，以及将所述待聚合文本归至创建的所述新的文本类中。

3、如权利要求 1 或 2 所述的方法，其特征在于，所述对长度不大于设定的长度阈值的待聚合文本进行特征提取，得到与所述待聚合文本相对应的文本特征集合，包括：

基于机械分词结合 N 元模型的特征提取方式对长度不大于设定的长度阈值的待聚
25 合文本进行特征提取，得到与所述待聚合文本相对应的第一文本特征集合，所述 N 为大于 1 的自然数。

4、如权利要求 3 所述的方法，其特征在于，所述基于机械分词结合 N 元模型的特征提取方式对长度不大于设定的长度阈值的待聚合文本进行特征提取，得到与所述待聚合文本相对应的第一文本特征集合，包括：

30 以中文汉字以及连续的字符串为最小切分单元，对所述待聚合文本进行分词，得到

多个分词；

基于 N 元模型，将得到的多个分词中的任意 N 个连续的分词组合为一文本特征，得到与所述待聚合文本相对应的文本特征集合。

5 5、如权利要求 1 或 2 所述的方法，其特征在于，所述设定的局部敏感哈希算法不
限于为 Simhash 算法或 Minhash 算法。

6、如权利要求 1 或 2 所述的方法，其特征在于，所述第一文本特征集合与所述第二文本特征集合之间的相似度至少通过 Jaccard 相似度、欧式距离以及海明距离中的任意一种或多种相似度度量参数来表示。

10 7、如权利要求 1 或 2 所述的方法，其特征在于，在对长度不大于设定的长度阈值的待聚合文本进行特征提取之前，所述方法还包括：

对所述待聚合文本进行预处理；其中，所述预处理至少包括：去除待聚合文本中的特殊标签、去除待聚合文本中的非文字特殊符号、对待聚合文本进行繁简字体转换、以及将待聚合文本中的连续性的拉丁文和/或数字归一化为设定的字符串中的一种或多种。

8、一种文本聚合装置，其特征在于，包括：

15 特征提取单元，用于对长度不大于设定的长度阈值的待聚合文本进行特征提取，得到与
所述待聚合文本相对应的第一文本特征集合；

文本聚合单元，用于基于设定的局部敏感哈希算法计算所述第一文本特征集合的哈希值，并根据计算得到的哈希值，判断已构建的与所述设定的局部敏感哈希算法相对应的哈希索引中，是否存在与计算得到的哈希值之间的距离不大于设定距离的匹配值；若是，则从与计算得到的哈希值之间的距离不大于设定距离的匹配值中，选取与计算得到的哈希值之间的距离最小的匹配值，并计算所述第一文本特征集合与所述最小的匹配值所对应的第二文本特征集合之间的相似度；以及，若确定所述第一文本特征集合与所述第二文本特征集合之间的相似度不小于设定的相似度阈值，则将所述待聚合文本聚合至所述第二文本特征集合所对应的文本类中。

25 9、如权利要求 8 所述的装置，其特征在于，

所述文本聚合单元，还用于若确定已构建的与所述设定的局部敏感哈希算法相对应的哈希索引中，不存在与计算得到的哈希值之间的距离不大于设定距离的匹配值；或者，确定已构建的与所述设定的局部敏感哈希算法相对应的哈希索引中，存在与计算得到的哈希值之间的距离不大于设定距离的匹配值、且确定所述第一文本特征集合与所述第二
30 文本特征集合之间的相似度小于设定的相似度阈值，则

将计算得到的哈希值更新至已构建的与所述设定的局部敏感哈希算法相对应的哈希索引中，并基于所述待聚合文本创建一个新的文本类，以及将所述待聚合文本归至创建的所述新的文本类中。

10、如权利要求 8 或 9 所述的装置，其特征在于，

5 所述特征提取单元，具体用于基于机械分词结合 N 元模型的特征提取方式对长度不大于设定的长度阈值的待聚合文本进行特征提取，得到与所述待聚合文本相对应的第一文本特征集合，所述 N 为大于 1 的自然数。

11、如权利要求 10 所述的装置，其特征在于，

10 所述特征提取单元，具体用于以中文汉字以及连续的字符串为最小切分单元，对所述待聚合文本进行分词，得到多个分词；并基于 N 元模型，将得到的多个分词中的任意 N 个连续的分词组合为一文本特征，得到与所述待聚合文本相对应的文本特征集合。

12、如权利要求 8 或 9 所述的装置，其特征在于，所述设定的局部敏感哈希算法不限于为 Simhash 算法或 Minhash 算法。

15 13、如权利要求 8 或 9 所述的装置，其特征在于，所述第一文本特征集合与所述第二文本特征集合之间的相似度至少通过 Jaccard 相似度、欧式距离以及海明距离中的任意一种或多种相似度度量参数来表示。

14、如权利要求 8 或 9 所述的装置，其特征在于，所述装置还包括预处理单元：

所述预处理单元，用于在对长度不大于设定的长度阈值的待聚合文本进行特征提取之前，对所述待聚合文本进行预处理；

20 其中，所述预处理至少包括：去除待聚合文本中的特殊标签、去除待聚合文本中的非文字特殊符号、对待聚合文本进行繁简字体转换、以及将待聚合文本中的连续性的拉丁文和/或数字归一化为设定的字符串中的一种或多种。

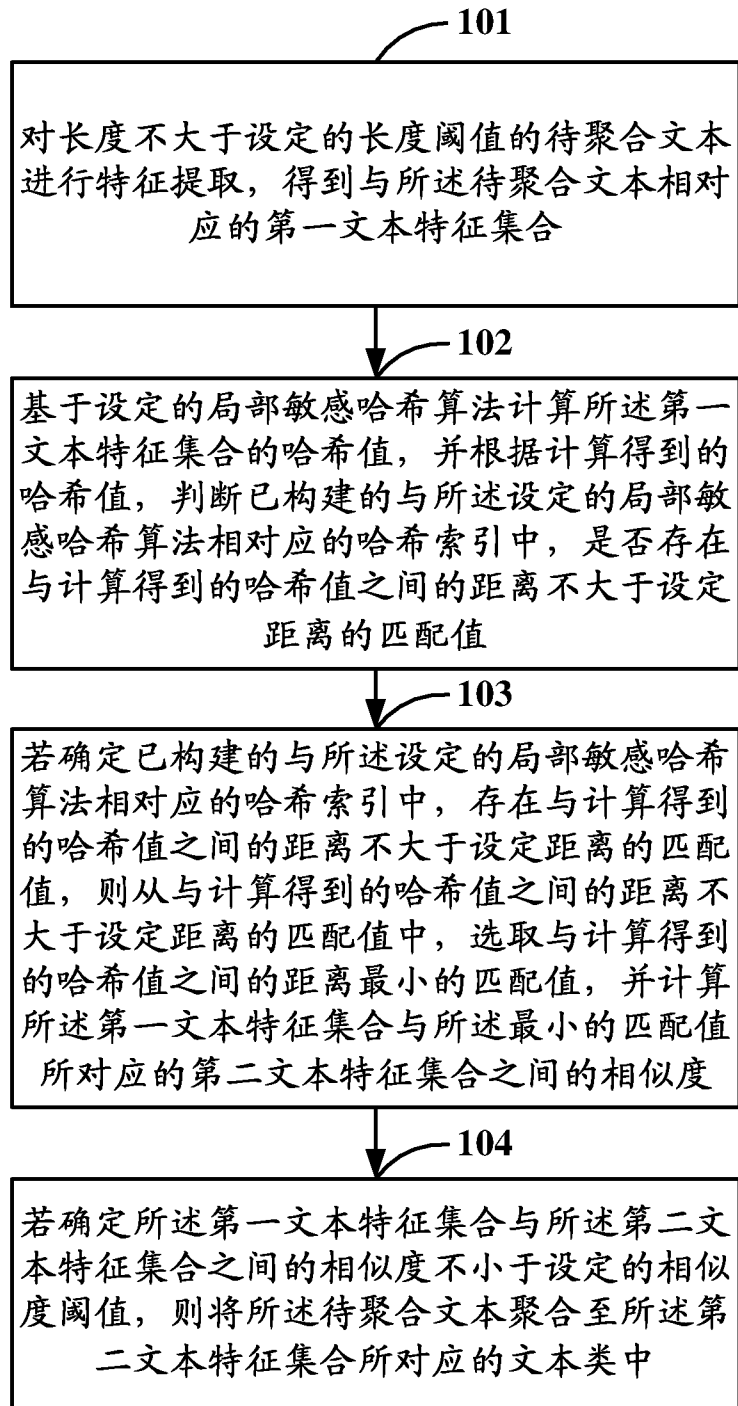


图 1

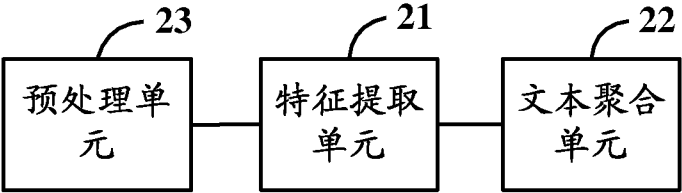


图 2

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2016/081090

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)

WPI, EPODOC, CNPAT, CNKI, IEEE, GOOGLE: short message, microblog, Hamming distance, feature, text, short, message, blog, aggregat+, hash+, Simhash, Minhash, distance, similar+, segment +, signature

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 103646080 A (BEIJING JINGDONG SHANGKE INFORMATION TECHNOLOGY CO., LTD. et al.), 19 March 2014 (19.03.2014), description, paragraphs [0039]-[0046]	1-14
A	CN 104391963 A (BEIJING ZHONGKE CHUANGYI TECHNOLOGY CO., LTD.), 04 March 2015 (04.03.2015), the whole document	1-14
A	CN 103744964 A (TONGJI UNIVERSITY), 23 April 2014 (23.04.2014), the whole document	1-14
A	CN 103064887 A (HUAWEI TECHNOLOGIES CO., LTD.), 24 April 2013 (24.04.2013), the whole document	1-14
A	CN 103914463 A (BEIJING FEINNO COMMUNICATION TECH CO., LTD.), 09 July 2014 (09.07.2014), the whole document	1-14
A	JP 2012118659 A (NIPPON TELEGRAPH & TELEPHONE CORPORATION), 21 June 2012 (21.06.2012), the whole document	1-14

☐ 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 21 July 2016 (21.07.2016)	Date of mailing of the international search report 16 August 2016 (16.08.2016)
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 ZHAO, Ting Telephone No.: (86-10) 62414434

International application No.
PCT/CN2016/081090

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

国际检索报告

国际申请号

PCT/CN2016/081090

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))

WPI, EPODOC, CNPAT, CNKI, IEEE, GOOGLE: 文本, 短, 短信, 微博, 聚合, 哈希, 距离, 海明距离, 相似, 分词, 特征, text, short, message, blog, aggregat+, hash+, Simhash, Minhash, distance, similar+, segment+, signature

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	CN 103646080 A (北京京东尚科信息技术有限公司 等) 2014年 3月 19日 (2014 - 03 - 19) 说明书第[0039]-[0046]段	1-14
A	CN 104391963 A (北京中科创益科技有限公司) 2015年 3月 4日 (2015 - 03 - 04) 全文	1-14
A	CN 103744964 A (同济大学) 2014年 4月 23日 (2014 - 04 - 23) 全文	1-14
A	CN 103064887 A (华为技术有限公司) 2013年 4月 24日 (2013 - 04 - 24) 全文	1-14
A	CN 103914463 A (北京新媒传信科技有限公司) 2014年 7月 9日 (2014 - 07 - 09) 全文	1-14
A	JP 2012118659 A (NIPPON TELEGRAPH & TELEPHONE CORPORATION) 2012年 6月 21日 (2012 - 06 - 21) 全文	1-14

☐ 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 7月 21日

国际检索报告邮寄日期

2016年 8月 16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

赵婷

传真号 (86-10)62019451

电话号码 (86-10)62414434

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2016/081090

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	103646080	A	2014年 3月 19日	无	
CN	104391963	A	2015年 3月 4日	无	
CN	103744964	A	2014年 4月 23日	无	
CN	103064887	A	2013年 4月 24日	无	
CN	103914463	A	2014年 7月 9日	无	
JP	2012118659	A	2012年 6月 21日	无	

表 PCT/ISA/210 (同族专利附件) (2009年7月)