

(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
2014 年 4 月 17 日 (17.04.2014)



(10) 国际公布号
WO 2014/056397 A1

- (51) 国际专利分类号:
G06F 17/30 (2006.01)
- (21) 国际申请号: PCT/CN2013/084096
- (22) 国际申请日: 2013 年 9 月 24 日 (24.09.2013)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
201210385630.6 2012 年 10 月 12 日 (12.10.2012) CN
- (71) 申请人: 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 403 室, Guangdong 518000 (CN)。
- (72) 发明人: 贺翔 (HE, Xiang); 中国广东省深圳市福田区振兴路赛格科技园 2 栋东 403 室, Guangdong 518000 (CN)。王业 (WANG, Ye); 中国广东省深圳市福田区振兴路赛格科技园 2 栋东 403 室, Guangdong 518000 (CN)。元超 (YI, Chao); 中国广东省深圳市福田区振兴路赛格科技园 2 栋东 403 室, Guangdong 518000 (CN)。

(74) 代理人: 北京市柳沈律师事务所 (LIU, SHEN & ASSOCIATES); 中国北京市朝阳区北辰东路 8 号汇宾大厦 A0601, Beijing 100101 (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, IS, JP, KE, KG, 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, 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, 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)。

[见续页]

(54) Title: LABEL OF INTEREST RECOMMENDATION METHOD, SYSTEM AND COMPUTER READABLE MEDIUM

(54) 发明名称: 兴趣标签推荐方法、系统及计算机可读介质

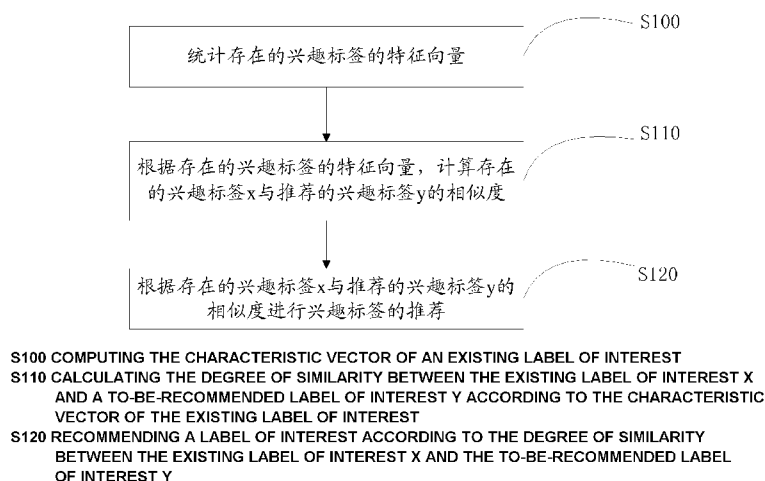


图 1 / Fig. 1

(57) Abstract: Provided are a label of interest recommendation method and system, the label of interest recommendation method comprising: computing the characteristic vector of an existing label of interest; calculating the similarity between the existing label of interest and a to-be-recommended label of interest according to the characteristic vector of the existing label of interest; and recommending a label of interest according to the similarity between the existing label of interest and the to-be-recommended label of interest. The present invention calculates the similarity between a specific label of interest and other labels, and recommends a label of interest according to the similarity, thus improving recommendation relevancy, and reducing interference.

(57) 摘要: 提供兴趣标签推荐方法及系统。该兴趣标签推荐方法包括: 统计存在的兴趣标签的特征向量; 根据存在的兴趣标签的特征向量, 计算存在的兴趣标签与待推荐的兴趣标签的相似度; 根据存在的兴趣标签与待推荐的兴趣标签的相似度进行兴趣标签的推荐。本发明通过计算某个兴趣标签与其他标签的相似度, 并根据该相似度推荐兴趣标签, 从而提高了推荐的相关性, 减少了干扰。



WO 2014/056397 A1



本国际公布:

— 包括国际检索报告(条约第 21 条(3))。

兴趣标签推荐方法、系统及计算机可读介质

技术领域

- 5 本发明涉及兴趣标签推荐方法、系统及计算机可读介质。具体地说，本发明涉及利用与存在的兴趣标签相关的互联网文章的特征词的统计特性来推荐兴趣标签的兴趣标签推荐方法、系统及计算机可读介质。

背景技术

- 10 随着互联网技术的不断发展,越来越多的用户通过互联网进行相关活动,例如,浏览新闻、购物等。但是随着互联网信息的不断增多,用户需要关注的内容随着增多。为帮助用户关注其感兴趣的内容,很多互联网应用中增加了兴趣标签,根据用户的兴趣标签向用户推荐相应的内容。兴趣标签是用户用来描述自己兴趣的词语,例如,用户可以用“篮球”、“NBA”、“林书豪”等
- 15 词语作为兴趣标签,来描述自己的兴趣。现有兴趣标签的推荐方法主要有随机推荐和热点推荐两种方法。随机推荐就是给用户随机推荐几个兴趣标签,而热点推荐是根据当前的热点事件,给用户推荐兴趣标签按类别推荐。但是,现有的推荐方式都存在明显的缺点,例如,随机推荐方式准确性不高,推荐效果差,可能会推荐很多用户不感兴趣的标签,而热点推荐方式只能覆盖热
- 20 门兴趣标签,不能根据用户的选择和喜好进行相关推荐。

因此,存在提高兴趣标签的推荐方式的准确性的必要。

发明内容

- 25 本发明通过统计与存在的兴趣标签相关的互联网文章的特征词来推荐兴趣标签,以提高推荐的兴趣标签的准确性。

- 根据本发明的一个方面,提供了一种兴趣标签推荐方法,包括步骤:统计与存在的兴趣标签相关的互联网文章的特征词以产生所述存在的兴趣标签的特征向量;根据存在的兴趣标签的特征向量,计算存在的兴趣标签与待推荐的兴趣标签的相似度;以及根据存在的兴趣标签与待推荐的兴趣标签的相
- 30 似度进行兴趣标签的推荐。

根据本发明的另一个方面，本发明还提供了一种兴趣标签推荐系统，包括：特征向量统计模块，用于统计与存在的兴趣标签相关的互联网文章的特征词以产生所述存在的兴趣标签的特征向量；相似度计算模块，用于根据存在的兴趣标签的特征向量，计算存在的兴趣标签与待推荐的兴趣标签的相似度；以及兴趣标签推荐模块，用于根据存在的兴趣标签与待推荐的兴趣标签的相似度进行兴趣标签的推荐。

根据本发明的另一个方面，本发明还提供了一种计算机可读介质，其存储用于实施向用户推荐兴趣标签的方法的计算机可读程序，所述方法包括步骤：统计与存在的兴趣标签相关的互联网文章的特征词以产生所述存在的兴趣标签的特征向量；根据存在的兴趣标签的特征向量，计算存在的兴趣标签与待推荐的兴趣标签的相似度；以及根据存在的兴趣标签与待推荐的兴趣标签的相似度进行兴趣标签的推荐。

由此可见，利用本发明上述的技术方案，根据某个存在的兴趣标签，统计该存在的兴趣标签与其他兴趣标签的相似度以及存在的兴趣标签与其他兴趣标签之间的共现关系，并根据相似度和共现关系从其他兴趣标签中推荐合适的兴趣标签，从而提高推荐的兴趣标签的准确性。

附图说明

为了使本发明的技术方案及优点更加清楚明白，以下结合附图描述本发明的实施例，在这些附图中，

图 1 是根据本发明第一实施例的兴趣标签推荐方法的流程图；

图 2 是根据本发明第二实施例的兴趣标签推荐方法的流程图；

图 3 是根据本发明第一实施例的兴趣标签推荐系统的结构示意图；以及

图 4 是根据本发明第二实施例的兴趣标签推荐系统的结构示意图。

具体实施方式

以下结合附图对本发明的实施例进行详细说明。应当理解，此处所描述的具体实施例仅仅用以解释本发明，并不用于限定本发明。

图 1 是根据本发明第一实施例的兴趣标签推荐方法的流程图。参考图 1，根据本发明第一实施例的兴趣标签推荐方法在步骤 S100 中，统计与存在的

兴趣标签相关的互联网文章的特征词以产生所述存在的兴趣标签的特征向量。这里，存在的兴趣标签为用户预先设定的兴趣标签，或者通过统计用户的浏览历史或者操作等事先计算出来的兴趣标签。

具体来说，根据本发明的该实施例，在步骤 S100 中，为了产生存在的
5 兴趣标签的特征向量，根据本发明第一实施例的兴趣标签推荐系统（下文将给出其详细描述）提取互联网相关文件中标题或正文包含该存在的兴趣标签的所有文章，对包含该存在的兴趣标签的所有文章切词、提取特征词，统计特征词的词频（Tf, term frequency），并且对特征词按照词频进行相关排序，截取所排序的特征词列中的一定数量的特征词，作为存在的兴趣标签的特征
10 向量 V_x 。例如，根据本发明的一个实施例，对特征词根据 $Tf * Idf$ （inverse document frequency，文档频率倒数）排序，截取所排序列的前 50 个特征词，作为这个兴趣标签 x 的特征向量 V_x 。可以理解，取词的数量可以根据需求进行设定。

接下来，在图 1 的步骤 S110 中，根据本发明第一实施例的兴趣标签推
15 荐系统根据存在的兴趣标签的特征向量，计算存在的兴趣标签 x 与待推荐的兴趣标签 y 的相似度。这里，存在的兴趣标签 x 与待推荐的兴趣标签 y 的相似度定义为： $R(x, y) = \cos(V_x, V_y)$ ，其中， $R(x, y)$ 表示存在的兴趣标签 x 与待推荐的兴趣标签 y 的相似度， $\cos$ 表示余弦函数， V_x 及 V_y 分别表示存在的兴趣标签 x 及待推荐的兴趣标签 y 的特征向量。

20 接下来，在步骤 S120 中，本发明第一实施例的兴趣标签推荐系统根据存在的兴趣标签 x 与待推荐的兴趣标签 y 的相似度进行兴趣标签的推荐。在步骤 S120 中，可以设定相似度阈值，在存在的兴趣标签 x 与待推荐的兴趣标签 y 的相似度大于设定的阈值时才进行推荐。另外，可以理解，相似度阈值可以根据需要进行设定和变更。

25 图 2 是根据本发明第二实施例的兴趣标签推荐方法的流程图。在该流程图中，步骤 S230 和步骤 S240 分别与图 1 中的步骤 S100 和 S110 相同，因此将省略其详细描述。以下参考图 2，来描述根据本发明第二实施例的兴趣标签推荐方法中的其他步骤。

在步骤 S200 中，根据本发明第二实施例的兴趣标签推荐系统（下文将进行详细描述）可以事先准备兴趣标签集，并对其进行分类。在这个步骤中，
30 比如，根据本发明第二实施例的兴趣标签推荐系统可以根据需求获得一个兴

趣标签候选集合，常见的兴趣标签集包括影视明星、体育赛事、文学艺术等方面的词条。

接下来，在步骤 S210 中，根据本发明第二实施例的兴趣标签推荐系统收集互联网相关本文作为语料，用来统计兴趣标签的关系。例如，根据本发
5 明的实施例，可以采用用户发表在博客或微博中的博文作为语料，也可以采用用户日志、心情或者文章等其他文本作为语料。

接下来，在步骤 S220 中，根据本发明第二实施例的兴趣标签推荐系统确定兴趣标签的特征词词集。根据本发明的实施例，为了确定兴趣标签的特征词词集，首先对语料切词、统计每个词的 Idf (Idf : inverse document
10 frequency, 文档频率倒数) 值，去掉高频词、停用词，去掉低频词，然后将剩下的词作为特征词。这里，文档频率指出现某个词的文档个数。

在步骤 S250 中，根据本发明第二实施例的兴趣标签推荐系统统计存在的兴趣标签 x 与待推荐的兴趣标签 y 之间的共现关系。这里，存在的兴趣标签 x 与待推荐的兴趣标签 y 之间的共现关系是指存在的兴趣标签 x 与待推荐
15 的兴趣标签 y 在同一篇文档中出现的次数 $P(x,y)$ ，即 $P(x,y)$ 为语料中同时包含标签词 x 和标签词 y 的文档数目， $P(x)$ 为语料中包含标签词 x 的文档数目。

然后，在步骤 S260 中，根据本发明第二实施例的兴趣标签推荐系统根据存在的兴趣标签 x 与待推荐的兴趣标签 y 的相似度和共现关系进行兴趣标
20 签的推荐。

例如，根据本发明的实施例，根据下列方式进行标签推荐。

假定 $K1$ 、 $K2$ 和 $K3$ 为预定阈值。若待推荐的兴趣标签 y 满足

存在的兴趣标签 x 与待推荐的兴趣标签 y 的相似度 $R(x,y) > K1$ 和

$P(x,y) = \text{存在的兴趣标签 } x、\text{待推荐的兴趣标签 } y \text{ 在语料中的共现次}$
25 $\text{数} > K3$ 且

$$M(x,y) = P(x,y)/(P(x)*P(y)) > K2$$

则将向用户推荐兴趣标签 y 。

注意，若已知用户对多个存在的兴趣标签 $x_1, x_2, \dots, x_n$ 感兴趣，则根据本发明的一个实施例，可以将存在的兴趣标签 x_i ($i=1, \dots, n$) 各自的相关待推荐的兴趣标签 y_i 都推荐给用户，某个待推荐的兴趣标签 y_i 在用户的相关待推荐的兴趣标签列表中出现的次数越多，表示用户对该待推荐的兴趣
30

标签 y_i 有兴趣的概率越大。

图 3 是根据本发明第一实施例的兴趣标签推荐系统的结构示意图。参考图 3, 本发明第一实施例的兴趣标签推荐系统包括特征向量统计模块 310、相似度计算模块 320 和兴趣标签推荐模块 330。

5 具体来说, 本发明第一实施例的兴趣标签推荐系统的特征向量统计模块 310 用于统计与存在的兴趣标签相关的互联网文章的特征词以产生所述存在的兴趣标签的特征向量。这里, 存在的兴趣标签为用户预先设定的兴趣标签, 或者通过统计用户的浏览历史或者操作等事先计算出来的兴趣标签。

10 具体来说, 本发明第一实施例的兴趣标签推荐系统的特征向量统计模块 310 可以提取互联网相关文件中标题或正文包含该存在的兴趣标签的所有文章, 对包含该存在的兴趣标签的所有文章切词、提取特征词、统计特征词的词频 (Tf, term frequency), 并且对特征词按照词频进行相关排序, 截取所排序的特征词列中的一定数量的特征词, 作为存在的兴趣标签的特征向量 V_x 。

15 此外, 本发明第一实施例的兴趣标签推荐系统的相似度计算模块 320 用于根据存在的兴趣标签的特征向量, 计算存在的兴趣标签 x 与待推荐的兴趣标签 y 的相似度。存在的兴趣标签 x 与待推荐的兴趣标签 y 的相似度定义为: $R(x, y) = \cos(V_x, V_y)$, 其中, $R(x, y)$ 表示存在的兴趣标签 x 与待推荐的兴趣标签 y 的相似度, $\cos$ 表示余弦函数, V_x 及 V_y 分别表示存在的兴趣标签 x 及待推荐的兴趣标签 y 的特征向量。

20 此外, 兴趣标签推荐模块 330 用于根据存在的兴趣标签 x 与待推荐的兴趣标签 y 的相似度进行兴趣标签的推荐。兴趣标签推荐模块 330 根据存在的兴趣标签 x 与待推荐的兴趣标签 y 的相似度是否大于设定的阈值, 来确定是否进行兴趣标签 y 的推荐, 另外, 可以理解, 相似度阈值可以根据需要进行设定和变更。

图 4 是根据本发明第二实施例的兴趣标签推荐系统的结构示意图。本发明第二实施例的兴趣标签推荐系统包括兴趣标签集分类模块 410、语料收集模块 420、特征词词集确定模块 430、特征向量统计模块 440、相似度计算模块 450、共现关系统计模块 460 和兴趣标签推荐模块 470。

30 在本发明第二实施例的兴趣标签推荐系统中, 兴趣标签集分类模块 410 用于事先准备兴趣标签集, 并对其进行分类。例如, 兴趣标签集分类模块 410

可以根据需求获得一个兴趣标签候选集合,常见的兴趣标签集包括影视明星、体育赛事、文学艺术等方面的词条。

此外,语料收集模块 420 用于收集互联网相关本文作为语料。例如,语料收集模块 420 用于从互联网上收集相关的文本作为语料,用来统计兴趣标签的关系。例如,可以采用用户发表在博客或微博中的博文作为语料,也可以采用用户日志、心情或者文章等其他文本作为语料。

此外,特征词词集确定模块 430 用于确定兴趣标签的特征词词集。例如,特征词词集确定模块 430 可以对语料进行切词、统计每个词的 Idf (Idf : inverse document frequency, 文档频率倒数) 值、去掉高频词、停用词以及去掉低频词之类的操作,然后将剩下的词作为特征词。

此外,特征向量统计模块 440 用于统计与存在的兴趣标签相关的互联网文章的特征词以产生所述存在的兴趣标签的特征向量。这里,存在的兴趣标签为用户预先设定的兴趣标签,或者通过统计用户的浏览历史或者操作等事先计算出来的兴趣标签。具体来说,特征向量统计模块 440 可以提取互联网相关文件中标题或正文包含该存在的兴趣标签的所有文章,对包含该存在的兴趣标签的所有文章切词、提取特征词、统计特征词的词频 (Tf , term frequency), 并且对特征词按照词频进行相关排序,截取所排序的特征词列中的一些数量的特征词,作为存在的兴趣标签的特征向量 V_x 。

此外,相似度计算模块 450 用于根据存在的兴趣标签的特征向量,计算存在的兴趣标签 x 与待推荐的兴趣标签 y 的相似度。

此外,共现关系统计模块 460 用于统计存在的兴趣标签 x 与待推荐的兴趣标签 y 之间的共现关系。存在的兴趣标签 x 与待推荐的兴趣标签 y 之间的共现关系是指存在的兴趣标签 x 与待推荐的兴趣标签 y 在同一篇文档中出现的次数 $P(x,y)$, 即 $P(x,y)$ 为语料中同时包含标签词 x 和标签词 y 的文档数目, $P(x)$ 为语料中包含标签词 x 的文档数目。

兴趣标签推荐模块 470 用于根据存在的兴趣标签 x 与待推荐的兴趣标签 y 的相似度和共现关系进行兴趣标签的推荐。例如,兴趣标签推荐模块还可以根据下列方式进行标签推荐。

假定 $K1$ 、 $K2$ 和 $K3$ 为预定阈值。若待推荐的兴趣标签 y 满足

$$R(x,y) > K1$$
$$P(x,y) = \text{兴趣标签 } x, y \text{ 在语料中的共现次数} > K3$$

$$M(x, y) = P(x, y) / (P(x) * P(y)) > K2$$

则将向用户推荐兴趣标签 y 。

另外，若已知用户对多个存在的兴趣标签 $x_1, x_2, \dots, x_n$ 感兴趣，则根据本发明的一个实施例，可以将存在的兴趣标签 x_i ($i=1, \dots, n$) 各自的相关待推荐的兴趣标签 y_i 都推荐给用户，某个待推荐的兴趣标签 y_i 在用户的相关待推荐的兴趣标签列表中出现的次数越多，表示用户对该待推荐的兴趣标签 y_i 有兴趣的概率越大。

通过本发明的上述实施例，可以根据某个兴趣标签，计算该兴趣标签与其他标签的相似度，进行兴趣标签推荐，从而提高推荐的相关性，并减少对用户的干扰。而且，本发明的上述实施例还在兴趣标签推荐时参考了标签间的共现关系，从而使推荐更加准确。

此外，根据本发明的一个实施例，根据图 1 和图 2 所示的兴趣推荐方法可以分别是由图 3 和图 4 所示的兴趣推荐系统中的相应模块来执行的兴趣推荐方法。而且，根据本发明的另一个实施例，图 3 和图 4 所示的兴趣推荐系统中的各个模块可以分别或全部合并为一个或若干个另外的模块，或者其中的某个（些）模块还可以再拆分为功能上更小的多个模块。

根据本发明的另一个实施例，可以通过在包括中央处理模块（CPU）、随机存取存储器（RAM）、只读存储器（ROM）等处理元件和存储元件的例如计算机的通用计算设备上运行能够执行如图 1 和图 2 中所示的兴趣推荐方法的计算机程序（包括程序代码），来分别构造如图 3 和图 4 中所示的兴趣推荐系统，以及来实现根据本发明的实施例的兴趣推荐方法。所述计算机程序可以记载于例如计算机可读记录介质上，并通过计算机可读记录介质装载于上述计算设备中，并在其中运行。以上描述了本发明的各个实施例，但是它们不用以限制本发明的范围，本发明的范围由后附的权利要求来定义。在不脱离本发明的精神和原则的情况下，所做的任何修改、等同替换、改进等，均应包含在权利要求的保护范围之内。

权利要求书

1、一种兴趣标签推荐方法，包括步骤：

统计与存在的兴趣标签相关的互联网文章的特征词以产生所述存在的兴趣标签的特征向量；

根据存在的兴趣标签的特征向量，计算存在的兴趣标签与待推荐的兴趣标签的相似度；以及

根据存在的兴趣标签与待推荐的兴趣标签的相似度进行兴趣标签的推荐。

2、根据权利要求1所述的兴趣标签推荐方法，还包括步骤：

统计存在的兴趣标签与待推荐的兴趣标签之间的共现关系，该共现关系是指存在的兴趣标签与待推荐的兴趣标签在同一篇文档中出现的次数，

根据存在的兴趣标签与待推荐的兴趣标签的相似度，和存在的兴趣标签与待推荐的兴趣标签的所述共现关系进行兴趣标签的推荐。

3、根据权利要求1或2所述的兴趣标签推荐方法，其中，所述统计与存在的兴趣标签相关的互联网文章的特征词以产生所述存在的兴趣标签的特征向量的步骤包括子步骤：

提取互联网相关文件中标题或正文包含所述存在的兴趣标签的所有文章；

对包含所述存在的兴趣标签的所有文章切词、提取特征词并统计特征词的词频；并且

根据所述词频对特征词进行相关排序，并取出所排序列中的预定数量的特征词作为兴趣标签的特征向量。

4、根据权利要求1或2所述的兴趣标签推荐方法，其中，所述统计与存在的兴趣标签相关的互联网文章的特征词以产生所述存在的兴趣标签的特征向量的步骤前还包括步骤：

准备兴趣标签集并将其进行分类；

收集互联网相关本文作为语料；并且

根据语料确定兴趣标签的特征词词集，

其中，所述特征词包括在特征词词集中，并且所述兴趣标签包括在已分

类的所述兴趣标签集中。

5、根据权利要求4所述的兴趣标签推荐方法，其中，所述确定兴趣标签的特征词词集的步骤包括：

对语料切词，统计每个词的文档频率倒数值，去掉高频词、停用词及低频词，并将剩下的词作为特征词。

6、一种兴趣标签推荐系统，包括：

特征向量统计模块，用于统计与存在的兴趣标签相关的互联网文章的特征词以产生所述存在的兴趣标签的特征向量；

相似度计算模块，用于根据存在的兴趣标签的特征向量，计算存在的兴趣标签与待推荐的兴趣标签的相似度；以及

兴趣标签推荐模块，用于根据存在的兴趣标签与待推荐的兴趣标签的相似度进行兴趣标签的推荐。

7、根据权利要求6所述的兴趣标签推荐系统，还包括：

共现关系统计模块，用于统计存在的兴趣标签与待推荐的兴趣标签之间的共现关系，该共现关系是指存在的兴趣标签与待推荐的兴趣标签在同一篇文档中出现的次数，

其中，所述兴趣标签推荐模块用于根据存在的兴趣标签与待推荐的兴趣标签的相似度以及存在的兴趣标签与待推荐的兴趣标签的所述共现关系进行兴趣标签的推荐。

8、根据权利要求6或7所述的兴趣标签推荐系统，其中，所述特征向量统计模块具体用于：

提取互联网相关文件中标题或正文包含所述存在的兴趣标签的所有文章；

对包含所述存在的兴趣标签的所有文章切词，提取特征词，并统计特征词的词频；并且

根据所述词频对特征词进行相关排序，并取出所排序列中的预定数量的特征词作为兴趣标签的特征向量。

9、根据权利要求6或7所述的兴趣标签推荐系统，还包括：

兴趣标签集分类模块，用于准备兴趣标签集，并将其进行分类；

语料收集模块，用于收集互联网相关本文作为语料；

特征词词集确定模块，用于根据语料确定兴趣标签的特征词词集，

其中，所述特征词包括在特征词词集中，并且所述兴趣标签包括在已分类的所述兴趣标签集中。

10、根据权利要求 9 所述的兴趣标签推荐系统，其中，所述特征词词集
5 确定模块具体用于对语料切词，统计每个词的文档频率倒数值，去掉高频词、停用词及低频词，并将剩下的词作为特征词。

11、一种计算机可读介质，其存储用于实施向用户推荐兴趣标签的方法的计算机可读程序，所述方法包括步骤：

10 统计与存在的兴趣标签相关的互联网文章的特征词以产生所述存在的兴趣标签的特征向量；

根据存在的兴趣标签的特征向量，计算存在的兴趣标签与待推荐的兴趣标签的相似度；以及

根据存在的兴趣标签与待推荐的兴趣标签的相似度进行兴趣标签的推荐。

15 12、根据权利要求 11 所述的计算机可读介质，所述方法还包括步骤：

统计存在的兴趣标签与待推荐的兴趣标签之间的共现关系，该共现关系是指存在的兴趣标签与待推荐的兴趣标签在同一篇文档中出现的次数，

根据存在的兴趣标签与待推荐的兴趣标签的相似度，和存在的兴趣标签与待推荐的兴趣标签的所述共现关系进行兴趣标签的推荐。

20 13、根据权利要求 11 或 12 所述的计算机可读介质，其中，所述统计与存在的兴趣标签相关的互联网文章的特征词以产生所述存在的兴趣标签的特征向量的步骤包括子步骤：

提取互联网相关文件中标题或正文包含所述存在的兴趣标签的所有文章；

25 对包含所述存在的兴趣标签的所有文章切词、提取特征词并统计特征词的词频；并且

根据所述词频对特征词进行相关排序，并取出所排序列中的预定数量的特征词作为兴趣标签的特征向量。

30 14、根据权利要求 11 或 12 所述的计算机可读介质，其中，所述统计与存在的兴趣标签相关的互联网文章的特征词以产生所述存在的兴趣标签的特

征向量的步骤前还包括步骤:

准备兴趣标签集并将进行分类;

收集互联网相关本文作为语料; 并且

根据语料确定兴趣标签的特征词词集,

- 5 其中, 所述特征词包括在特征词词集中, 并且所述兴趣标签包括在已分类的所述兴趣标签集中。

15、根据权利要求 14 所述的计算机可读介质, 其中, 所述确定兴趣标签的特征词词集的步骤包括:

- 10 对语料切词, 统计每个词的文档频率倒数值, 去掉高频词、停用词及低频词, 并将剩下的词作为特征词。

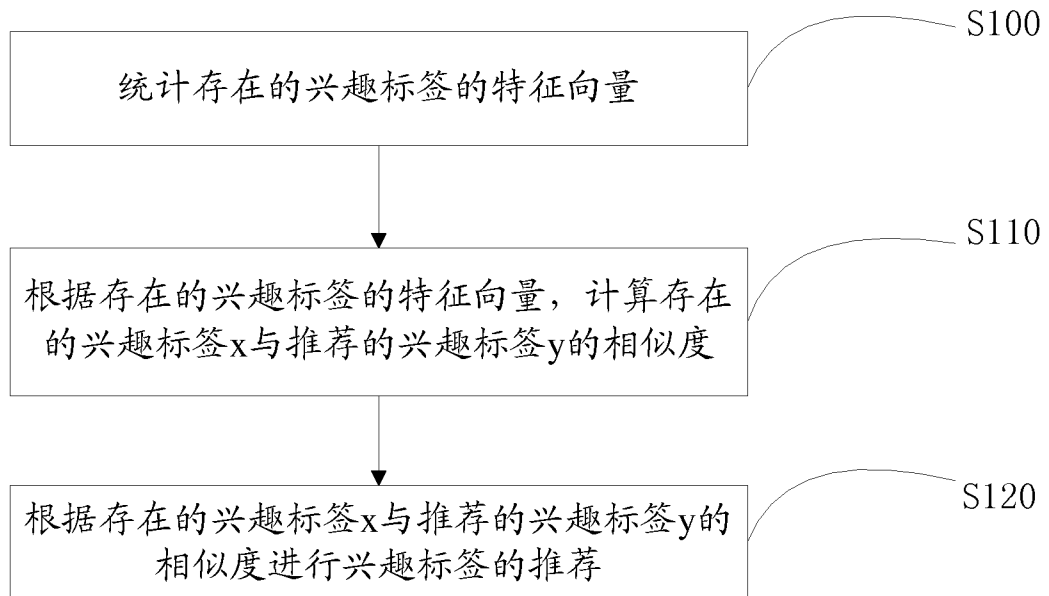


图 1

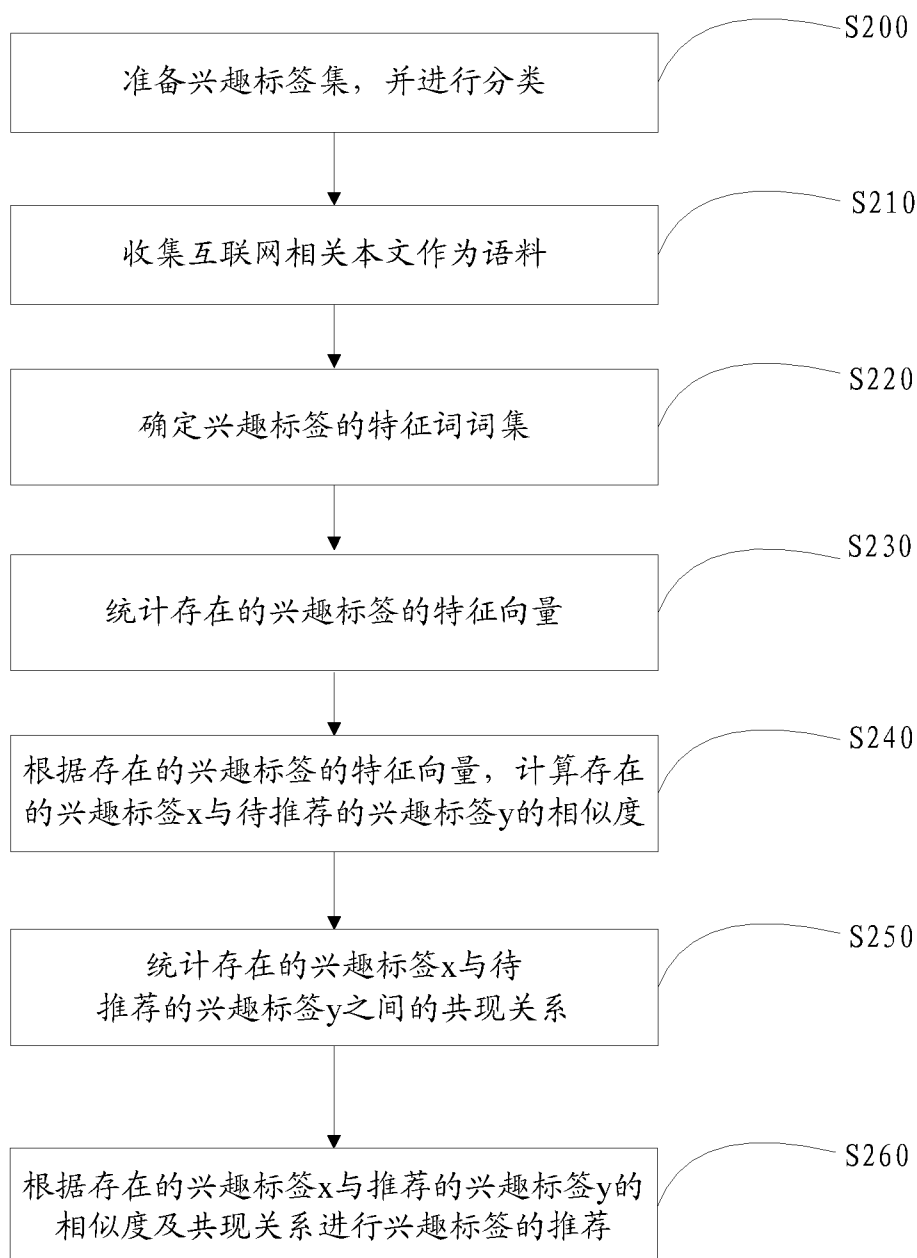


图 2



图 3

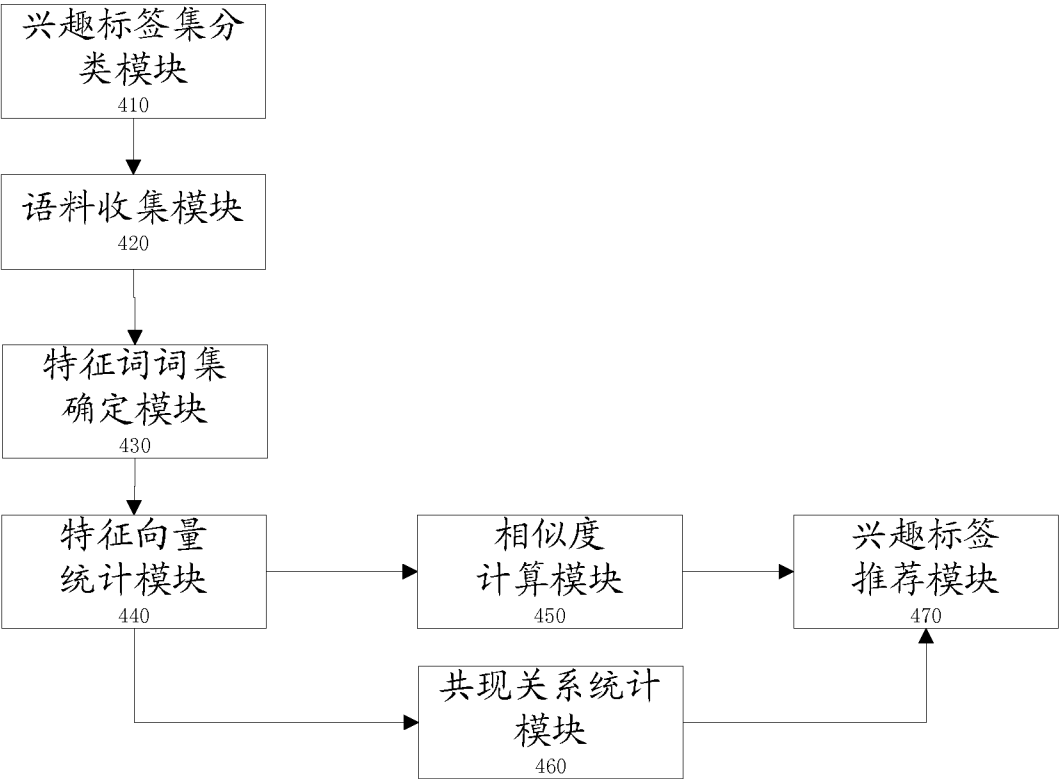


图 4

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2013/084096

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

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, EPODOC, WPI, CNKI: recommend???, suggest???, interest???, character+, extract???, stat+, similar+

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 101546341 A (TOSHIBA CORP.), 30 September 2009 (30.09.2009), description, page 2, lines 4-15, and page 6, line 11 to page 7, line 6	1-15
A	CN 101782920 A (INSTITUTE OF AUTOMATION, CHINESE ACADEMY OF SCIENCES), 21 July 2010 (21.07.2010), the whole document	1-15
A	CN 101923545 A (BEIJING LMOBILE MEDIA TECHNOLOGY CO., LTD.), 22 December 2010 (22.12.2010), the whole document	1-15
A	JP 2001-142899 A (NIPPON TELEGRAPH AND TELEPHONE CORP.), 25 May 2001 (25.05.2001), the whole document	1-15

☐ 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 06 December 2013 (06.12.2013)	Date of mailing of the international search report 26 December 2013 (26.12.2013)
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 KANG, Kai Telephone No.: (86-10) 62413998

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2013/084096

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 101546341 A	30.09.2009	US 2009/0248678 A1	01.10.2009
		JP 2009-238115 A	15.10.2009
CN 101782920 A	21.07.2010	None	
CN 101923545 A	22.12.2010	None	
JP 2001-142899 A	25.05.2001	None	

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: 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, EPODOC, WPI, CNKI

推荐, 兴趣, 特征, 提取, 相似, 统计

recommend???, suggest???, interest???, character+, extract???, stat+, similar+

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 101546341 A (株式会社东芝) 30.9 月 2009 (30.09.2009) 说明书第 2 页第 4-15 行、说明书第 6 页第 11 行-第 7 页第 6 行	1-15
A	CN 101782920 A (中国科学院自动化研究所) 21.7 月 2010 (21.07.2010) 全文	1-15
A	CN 101923545 A (北京百分通联传媒技术有限公司) 22.12 月 2010 (22.12.2010) 全文	1-15
A	JP 特开 2001-142899 A (日本电信电话株式会社) 25.5 月 2001 (25.05.2001) 全文	1-15



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

06.12 月 2013 (06.12.2013)

国际检索报告邮寄日期

26.12 月 2013 (26.12.2013)

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

康凯

电话号码: (86-10) 62413998

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2013/084096

检索报告中引用的 专利文件	公布日期	同族专利	公布日期
CN 101546341 A	30.09.2009	US 2009/0248678 A1 JP 特开 2009-238115 A	01.10.2009 15.10.2009
CN 101782920 A	21.07.2010	无	
CN 101923545 A	22.12.2010	无	
JP 特开 2001-142899 A	25.05.2001	无	