

(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
2023 年 10 月 26 日 (26.10.2023)



(10) 国际公布号
WO 2023/202170 A1

(51) 国际专利分类号:

G06F 40/284 (2020.01)

(21) 国际申请号:

PCT/CN2023/071668

(22) 国际申请日:

2023 年 1 月 10 日 (10.01.2023)

(25) 申请语言:

中文

(26) 公布语言:

中文

(30) 优先权:

202210421185.8 2022 年 4 月 21 日 (21.04.2022) CN

(71) 申请人: 北京沃东天骏信息技术有限公司 (**BEIJING WODONG TIANJUN INFORMATION TECHNOLOGY CO., LTD.**) [CN/CN]; 中国北京市北京经济技术开发区科创十一街 18 号院 2 号楼 4 层 A402 室, Beijing 100176 (CN)。北京京东世纪贸易有限公司 (**BEIJING JINGDONG CENTURY TRADING CO., LTD.**) [CN/CN]; 中国北

京市北京经济技术开发区科创十一街 18 号 C 座 2 层 201 室, Beijing 100176 (CN)。

(72) 发明人: 李鹏 (**LI, Peng**); 中国北京市北京经济技术开发区科创十一街 18 号院 2 号楼 4 层 A402 室, Beijing 100176 (CN)。李勇 (**LI, Yong**); 中国北京市北京经济技术开发区科创十一街 18 号院 2 号楼 4 层 A402 室, Beijing 100176 (CN)。包勇军 (**BAO, Yongjun**); 中国北京市北京经济技术开发区科创十一街 18 号院 2 号楼 4 层 A402 室, Beijing 100176 (CN)。颜伟鹏 (**YAN, Weipeng**); 中国北京市北京经济技术开发区科创十一街 18 号院 2 号楼 4 层 A402 室, Beijing 100176 (CN)。

(74) 代理人: 中原信达知识产权代理有限责任公司 (**CHINA SINDA INTELLECTUAL PROPERTY LIMITED**); 中国北京市西城区金融街 19 号富凯大厦 B 座 11 层, Beijing 100033 (CN)。

(54) **Title:** PRODUCT WORD DISAMBIGUATION METHOD AND APPARATUS

(54) 发明名称: 产品词的消歧方法和装置

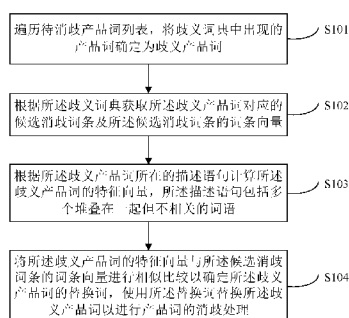


图 1

- S101 Traverse a product word list to be disambiguated, and determine a product word appearing in an ambiguity dictionary as an ambiguous product word
- S102 Acquire, according to the ambiguity dictionary, a candidate disambiguation entry corresponding to the ambiguous product word and an entry vector of the candidate disambiguation entry
- S103 Calculate a feature vector of the ambiguous product word according to a description statement where the ambiguous product word is located, wherein the description statement comprises a plurality of unrelated words which are stacked together
- S104 Perform similarity comparison on the feature vector of the ambiguous product word and the entry vector of the candidate disambiguation entry to determine a substitute word of the ambiguous product word, and substitute the substitute word for the ambiguous product word for disambiguation of the product word

(57) **Abstract:** The present application relates to the technical field of computers. Disclosed are a product word disambiguation method and apparatus. A specific embodiment of the method comprises: traversing a product word list to be disambiguated, and determining a product word appearing in an ambiguity dictionary as an ambiguous product word; acquiring, according to the ambiguity dictionary, a candidate disambiguation entry corresponding to the ambiguous product word and an entry vector of the candidate disambiguation entry; calculating a feature vector of the ambiguous product word according to a description statement where the ambiguous product word is located, wherein the description statement comprises a plurality of unrelated words which are stacked together; and performing similarity comparison on the feature vector of the ambiguous product word and the entry vector of the candidate disambiguation entry to determine a substitute word of the ambiguous product word, and substituting the substitute word for the ambiguous product word for disambiguation of the product word.

(57) 摘要: 本申请公开了产品词的消歧方法和装置, 涉及计算机技术领域。该方法的一具体实施方式包括: 遍历待消歧产品词列表, 将歧义词典中出现的词确定为歧义产品词; 根据歧义词典获取歧义产品词对应的候选消歧词条及所述候选消歧词条的词条向量; 根据歧义产品词所在的描述语句计算歧义产品词的特征向量, 描述语句包括多个堆叠在一起但不相关的词语; 将歧义产品词的特征向量与候选消歧词条的词条向量进行相似比较以确定歧义产品词的替换词, 使用替换词替换歧义产品词以进行产品词的消歧处理。

(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, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, 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, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, 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, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, 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, ME, 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))。

产品词的消歧方法和装置

相关申请的交叉引用

5 本申请要求享有 2022 年 4 月 21 日提交的发明名称为“一种产品词的消歧方法和装置”的中国专利申请 No. 202210421185.8 的优先权，在此全文引用上述中国专利申请公开的内容以作为本申请的一部分或全部。

10 技术领域

 本公开涉及计算机技术领域，尤其涉及一种产品词的消歧方法和装置。

背景技术

15 现有的词语消歧的方法，句法分析一般针对普通自然语言句子，并不适用于由多个词语堆叠、且不存在相关词的场景，例如电商平台的商品标题、网页描述标签词等场景，在这些场景下的词语消歧效果差、精确度低，使得商品匹配结果和网页信息的检索结果不够准确。

20 发明内容

 有鉴于此，本公开实施例提供一种产品词的消歧方法和装置。

 根据公开明实施例的一个方面，提供了一种产品词的消歧方法，包括：

25 遍历待消歧产品词列表，将歧义词典中出现的 product 词确定为歧义 product 词；

 根据所述歧义词典获取所述歧义 product 词对应的候选消歧词条及所述候选消歧词条的词条向量；

30 根据所述歧义 product 词所在的描述语句计算所述歧义 product 词的特征向量，所述描述语句包括多个堆叠在一起但不相关的词语；

将所述歧义产品词的特征向量与所述候选消歧词条的词条向量进行相似比较以确定所述歧义产品词的替换词，使用所述替换词替换所述歧义产品词以进行产品词的消歧处理。

5 根据本公开的一个或多个实施例，所述歧义词典是通过以下方式
来构建的：对训练集包括的描述语句进行训练以构建产品词编码模型；
使用所述产品词编码模型提取同义词典中每个产品词的特征向量；根
据所述产品词的特征向量判断所述同义词典中的词条是否为歧义词，
其中，每个词条包括不少于一个产品词；对判定为歧义词的词条进行
10 切分构建歧义词典。

 根据本公开的一个或多个实施例，对训练集中的描述语句进行训
练以构建产品词编码模型，包括：对训练集中的每个描述语句，从所
述描述语句中抽取产品词，得到产品词列表；基于所述产品词列表对
15 语义表示模型进行训练，构建产品词编码模型，用于对所述产品词列
表中的每个产品词进行特征提取。

 根据本公开的一个或多个实施例，使用所述产品词编码模型提取
同义词典中每个产品词的特征向量，包括：对同义词典中的每个产品
20 词，从训练集包括的描述语句中，提取包含所述产品词的描述语句集
合；对所述描述语句集合中的每个描述语句，利用所述产品词编码模
型计算<描述语句，所述产品词>的特征向量，作为所述描述语句对应
的特征向量；对得到的所述产品词的所有描述语句对应的特征向量取
平均值，得到所述产品词的特征向量。

25 根据本公开的一个或多个实施例，根据所述产品词的特征向量判
断所述同义词典中的词条是否为歧义词，包括：对所述同义词典中的
每个词条，获取所述词条包括的产品词及产品词的特征向量；根据所
述产品词的特征向量进行产品词聚类，并获取聚类类簇的个数；在所
30 述聚类类簇的个数大于 1 的情况下，判定所述词条为歧义词；并且，

对判定为歧义词的词条进行切分构建歧义词典，包括：将判定为歧义词的词条的每个聚类类簇作为一个新词条，构建歧义词典。

5 根据本公开的一个或多个实施例，根据所述产品词的特征向量进行产品词聚类，包括：构建产品词的全连通图，所述全连通图的顶点为产品词，边为两个产品词之间的相似度，所述相似度根据产品词的特征向量来计算；使用图切分方法对所述全连通图进行切分，得到类簇以进行产品词聚类。

10 根据本公开的一个或多个实施例，还包括：将所述新词条对应的聚类类簇的类簇向量作为所述新词条的词条向量，所述类簇向量为所述聚类类簇中包括的产品词的特征向量的平均值。

15 根据本公开实施例的另一方面，提供了一种产品词的消歧装置，包括：

歧义产品词确定模块，用于遍历待消歧产品词列表，将歧义词典中出现的产品词确定为歧义产品词；

候选消歧词条确定模块，用于根据所述歧义词典获取所述歧义产品词对应的候选消歧词条及所述候选消歧词条的词条向量；

20 特征向量计算模块，用于根据所述歧义产品词所在的描述语句计算所述歧义产品词的特征向量，所述描述语句包括多个堆叠在一起但不相关的词语；

25 歧义产品词替换模块，用于将所述歧义产品词的特征向量与所述候选消歧词条的词条向量进行相似比较以确定所述歧义产品词的替换词，使用所述替换词替换所述歧义产品词以进行产品词的消歧处理。

根据本公开实施例的又一方面，提供了一种产品词消歧的电子设
备。

30 一种产品词消歧的电子设备，包括：一个或多个处理器；存储装

置，用于存储一个或多个程序，当所述一个或多个程序被所述一个或多个处理器执行，使得所述一个或多个处理器实现本公开实施例所提供的产品词的消歧方法。

5 根据本公开实施例的再一方面，提供了一种计算机可读介质。

一种计算机可读介质，其上存储有计算机程序，所述程序被处理器执行时实现本公开实施例所提供的产品词的消歧方法。

10 上述的非惯用的可选方式所具有的进一步效果将在下文中结合具体实施方式加以说明。

附图说明

附图用于更好地理解本公开，不构成对本公开的不当限定。其中：
15 图 1 是根据本公开实施例的产品词的消歧方法的主要步骤示意图；
图 2 是本公开实施例的同义词典的结构示意图；
图 3 是本公开实施例的歧义词典的构建流程示意图；
图 4 是本公开实施例的歧义词典的构建流程示意图；
图 5 是本公开实施例的歧义词典的词条结构示意图；
20 图 6 是根据本公开实施例的产品词的消歧装置的主要模块示意图；
图 7 是本公开实施例可以应用于其中的示例性系统架构图；
图 8 是适于用来实现本公开实施例的终端设备或服务器的计算机系统的结构示意图。

具体实施方式

25 以下结合附图对本公开的示范性实施例做出说明，其中包括本公开实施例的各种细节以助于理解，应当将它们认为仅仅是示范性的。因此，本领域普通技术人员应当认识到，可以对这里描述的实施例做出各种改变和修改，而不会背离本公开的范围和精神。同样，为了清楚和简明，以下的描述中省略了对公知功能和结构的描述。
30

现有技术的词语消歧的方法，多是针对逻辑性较强的文本来进行的，对于由多个堆叠在一起但不相关的词语所构成的描述语句（例如，电商平台的商品标题）的词语消歧，效果并不好，会存在以下几个缺陷：

1）现有句法分析对电商平台的商品标题不太适用，电商平台的商品标题是一些词的堆砌，现有句法分析模型在电商平台的商品标题上产出结果可用性不高；

2）消歧使用相关词以及相关词的上下位词来表示词义，但相关词以及上下位词对标题中产品词的描述可能并不精确，消歧结果对于用户理解存在难度；

3）构建上下位词典的成本较高，词典新词扩充速度较慢。

为了解决上述技术问题，本公开提供了一种产品词的消歧方法，面向电商平台的商品标题等由多个堆叠在一起但不相关的词语所构成的描述语句进行产品词消歧。消歧包括两部分处理：一是进行歧义词标记，二是对歧义词进行词义消歧。其中，歧义词标记：对电商平台的商品标题的每个产品词，输出产品词否是歧义词的标签；词义消歧：使用同义替换词来表示歧义词，替换词代表歧义产品词在标题中的含义。本公开的技术方案具有以下优点：

1、本公开不需要做句法分析，适合电商平台的商品标题这类结构性不强的文本；

2、消歧使用产品词在同义词典中的同义词作为替代词，不使用上下位词典。同义产品词词典构建成本（包括标注成本）要低于上下位词典，低成本有利于快速收录新产品词；

3、使用同义替换词进行消歧相比使用上下位词典进行消歧，输出结果在表意上更具体明确，输出结果与原始标题产品词在同一空间，更容易适配下游应用。

以对商品标题进行词语消歧为例，本公开实施例的技术方案主要

包括两部分：

1、歧义产品词词典构建。使用产品词同义词典以及商品标题库作为输入，遍历同义词典中的词条，判断词条是否为歧义词，如果词条是歧义词，那么对词条进行拆分，放入到歧义词典中，同时计算拆分后的词条的向量表示；

2、商品标题产品词消歧。主要包括：

（1）歧义词标记：遍历商品标题中的产品词，如果出现在歧义词典中，那么该词被标记为歧义产品词；

（2）词义消歧：使用商品标题计算产品词的向量表示，与歧义词典中的词条向量进行相似度计算，选择相似度最大的词条作为替代词。

图 1 是根据本公开实施例的产品词的消歧方法的主要步骤示意图。如图 1 所示，本公开实施例的产品词的消歧方法主要包括如下的步骤 S101 至步骤 S104。

步骤 S101：遍历待消歧产品词列表，将歧义词典中出现的产品词确定为歧义产品词。其中，待消歧产品词列表可以是基于描述语句进行分词后筛选确定的，描述语句包括多个堆叠在一起但不相关的词语，例如电商领域的商品标题等。在本公开的一个实施例中，商品标题例如“04 不锈钢 牵引 宠物用品 大型犬 金毛德牧 防咬栓狗绳 3mm 粗-长 1.6 米”，则对其进行分词后得到的词列表如下表 1 中左边列所示。

表 1

分词后的词列表	词标签（人工标注）
04	O
不锈钢	O
牵引	O
宠物	B-Product
用品	I-Product
大型犬	O
金毛	O
德牧	O
防咬	O

拴狗绳	B-Product
3mm	O
粗	O
-	O
长	O
1.6	O
米	O

在表 1 中，右边列示出了分词后得到的每个词对应的词标签，根据词标签可以筛选出具体的产品词。在表 1 中，标签 B-Product 表示产品词的开始，标签 I-Product 表示产品词的中间，标签 O 表示非产品词。通过标签可以提取产品词：以标签 B-Product 对应的词作为开始，不断拼接标签 I-Product 对应的词，直到遇到 O 标签为止。根据表 1 可以确定，该商品标题对应的产品词包括“宠物用品”和“拴狗绳”。

之后，通过将待消歧产品词列表中的产品词与预先构建的歧义词典进行比较即可确定歧义产品词。歧义词典是基于产品词的同义词典构建的。其中，同义词典的构建方法例如是挖掘产品词的候选同义词，然后人工对候选同义词进行标注，以构建同义词典。图 2 是本公开实施例的同义词典的结构示意图，如图 2 中所示，同义词典中每个词条包括词条键，以及词条的同义词列表两部分。词条可以表示为键值对的形式，词条键是单个词，词条值是词列表，词列表的每个词都是词条键的同义词。比如词条“宠物用品：狗绳、狗链、胸背带、窝垫、狗垫子”，词条键是“宠物用品”，词条值是“狗绳、狗链、胸背带、窝垫、狗垫子”，表示“狗绳、狗链、胸背带、窝垫、狗垫子”是“宠物用品”的同义词。词条键在同义词典中是唯一的。

图 3 是本公开实施例的歧义词典的构建流程示意图。在本公开的一个实施例中，歧义词典是通过以下方式构建的：

步骤 S301：对训练集包括的描述语句进行训练以构建产品词编码模型；

步骤 S302：使用所述产品词编码模型提取同义词典中每个产品词

的特征向量；

步骤 S303：根据所述产品词的特征向量判断所述同义词典中的词条是否为歧义词，其中，每个词条包括不少于一个产品词；

步骤 S304：对判定为歧义词的词条进行切分构建歧义词典。

5

图 4 是本公开实施例的产品词编码模型的构建流程示意图。在本公开的一个实施例中，步骤 S301 中对训练集中的描述语句进行训练以构建产品词编码模型，具体可以包括：

10 步骤 S401：对训练集中的每个描述语句，从所述描述语句中抽取产品词，得到产品词列表。以商品标题作为描述语句为例，从商品标题中抽取产品词时，可以通过预先训练的实体识别模型来进行抽取，其中，实体识别模型是基于训练集中的每个商品标题进行分词后进行标注来进行产品词抽取的；

15 步骤 S402：基于所述产品词列表对语义表示模型进行训练，构建产品词编码模型，用于对所述产品词列表中的每个产品词进行特征提取。使用商品标题的产品词列表构建产品词编码模型，编码模型可以使用预训练模型（比如语义表示模型 ELMo），输入是商品标题的产品词列表，输出是每个产品词的向量表示。

20 根据本公开的技术方案，在构建产品词编码模型时，也可以使用描述语句中除产品词外的其他信息，只要对于判断产品词词义有帮助的信息都可以使用。本公开对此不作限定。

25 在构建产品词编码模型之后，即可对每个产品词进行特征提取以得到特征向量。具体地，在本公开的另一个实施例中，步骤 S302 中使用所述产品词编码模型提取同义词典中每个产品词的特征向量，具体可以包括：

对同义词典中的每个产品词，从训练集包括的描述语句中，提取包含所述产品词的描述语句集合；

30 对所述描述语句集合中的每个描述语句，利用所述产品词编码模

型计算<描述语句, 所述产品词>的特征向量, 作为所述描述语句对应的特征向量;

对得到的所述产品词的所有描述语句对应的特征向量取平均值, 得到所述产品词的特征向量。

5

根据本公开的又一个实施例, 步骤 S303 中根据所述产品词的特征向量判断所述同义词典中的词条是否为歧义词, 具体可以包括:

对所述同义词典中的每个词条, 获取所述词条包括的产品词及产品词的特征向量;

10

根据所述产品词的特征向量进行产品词聚类, 并获取聚类类簇的个数; 其中, 在进行聚类时, 可以基于图进行聚类;

在所述聚类类簇的个数大于 1 的情况下, 判定所述词条为歧义词;

并且, 步骤 S304 中对判定为歧义词的词条进行切分构建歧义词典, 包括:

15

将判定为歧义词的词条的每个聚类类簇作为一个新词条, 构建歧义词典。

在本公开的一个实施例中, 在判定所述词条为歧义词之后, 还可以对判定的歧义词进行人工校验, 以进一步验证判定结果的正确性。

20

在本公开的又一个实施例中, 在根据所述产品词的特征向量进行产品词聚类时, 具体可以包括:

构建产品词的全连通图, 所述全连通图的顶点为产品词, 边为两个产品词之间的相似度, 所述相似度根据产品词的特征向量来计算。

25

具体地, 可以通过计算两个产品词的特征向量的余弦距离来计算产品词之间的相似度;

使用图切分方法对所述全连通图进行切分, 得到类簇以进行产品词聚类。其中, 图切分方法例如社区发现算法等, 通过图切分方法进行全连通图的切分, 即可将相似度高的产品词作为一个类簇进行切分, 从而实现了产品词聚类。

30

根据本公开的技术方案，在进行聚类时，也可选用其他的聚类方法，本公开对此不作限定。

5 在产品词聚类之后，即可得到相应的类簇，对于每个聚类类簇，将所述聚类类簇中包括的产品词的特征向量的平均值作为类簇向量，以用于后续在构建歧义词典时作为词条向量。之后，即可对判定为歧义词的词条进行切分，将每个聚类类簇包括的产品词作为一个新词条，并将所述新词条对应的聚类类簇的类簇向量作为所述新词条的词条向
10 量，结合词条向量来构建歧义词典。

图 5 是本公开实施例的歧义词典的词条结构示意图。如图 5 所示，歧义词典中每个词条包括词条键、词条的同义词列表、旧词条键、词条向量 4 个字段。其中，词条键使用旧词条键和聚类类簇的编号组合得到；旧词条键使用拆分前的同义词词典的词条键；词条向量使用类簇向量。例如，旧词条键“宠物用品”，其对应的词条键包括“宠物用品-0”和“宠物用品-1”，且词条键“宠物用品-0”与词条键“宠物用品-1”中包括的词条的同义词列表是不同的，是根据聚类结果得到的。相应地，每个词条键对应的词条向量也就不同。

20 在本公开的实施例中，通过遍历待消歧产品词列表，将每个产品词与歧义词典中的词条的旧词条键进行比较，如果产品词与歧义词典中词条的旧词条键相同，那么该产品词即被标记为歧义产品词。

25 步骤 S102：根据所述歧义词典获取所述歧义产品词对应的候选消歧词条及所述候选消歧词条的词条向量。在将产品词与歧义词典中的词条的旧词条键进行比较时，如果产品词与歧义词典中词条的旧词条键相同，那么该产品词即被标记为歧义产品词，相应地，歧义词典中对应的词条即作为候选消歧词条。例如，商品标题分词结果为“304 不
30 锈钢，牵引，宠物用品，大型犬，金毛德牧，防咬，栓狗绳，3mm 粗-

长 1.6 米”，抽取的待消歧产品词列表为“宠物用品，栓狗绳”。遍历待消歧产品词列表，发现“宠物用品”出现在歧义词典的旧词条键字段，那么“宠物用品”被标记为歧义产品词。歧义词典中相关词条“宠物用品-0”、“宠物用品-1”作为候选消歧词条。

5

步骤 S103：根据所述歧义产品词所在的描述语句计算所述歧义产品词的特征向量，所述描述语句包括多个堆叠在一起但不相关的词语。根据步骤 401 和步骤 S402 中构建的产品词编码模型，即可根据所述歧义产品词所在的描述语句计算所述歧义产品词的特征向量。

10

步骤 S104：将所述歧义产品词的特征向量与所述候选消歧词条的词条向量进行相似比较以确定所述歧义产品词的替换词，使用所述替换词替换所述歧义产品词以进行产品词的消歧处理。根据歧义产品词的特征向量，和候选消歧词条的词条向量，通过计算两个的向量相似度，选择相似度最大的候选消歧词条作为该歧义产品词的替换词。例如，“宠物用品”为歧义产品词，歧义词典中相关词条“宠物用品-0”、“宠物用品-1”为候选消歧词条，那么，分别计算“宠物用品”的特征向量与“宠物用品-0”、“宠物用品-1”的词条向量之间的相似度，从而从中筛选出“宠物用品”的替换词。假设“宠物用品”与“宠物用品-0”的相似度更高，那么替换词即为“宠物用品-0”。“宠物用品-0”这个词条键对应的词条的同义词列表包括了“狗绳、狗链、胸背带”等信息，可以很好的反映产品词“宠物用品”的含义。

15

20

图 6 是根据本公开实施例的产品词的消歧装置的主要模块示意图。如图 6 所示，本公开实施例的产品词的消歧装置 600 主要包括歧义产品词确定模块 601、候选消歧词条确定模块 602、特征向量计算模块 603 和歧义产品词替换模块 604。

25

歧义产品词确定模块 601，用于遍历待消歧产品词列表，将歧义词典中出现的 product 词确定为歧义产品词；

30

候选消歧词条确定模块 602, 用于根据所述歧义词典获取所述歧义产品词对应的候选消歧词条及所述候选消歧词条的词条向量;

特征向量计算模块 603, 用于根据所述歧义产品词所在的描述语句计算所述歧义产品词的特征向量, 所述描述语句包括多个堆叠在一起但不相关的词语;

歧义产品词替换模块 604, 用于将所述歧义产品词的特征向量与所述候选消歧词条的词条向量进行相似比较以确定所述歧义产品词的替换词, 使用所述替换词替换所述歧义产品词以进行产品词的消歧处理。

根据本公开的一个实施例, 产品词的消歧装置 600 包括歧义词典构建模块 (图中未示出), 用于构建歧义词典, 并且, 所述歧义词典是通过以下方式来构建的: 对训练集包括的描述语句进行训练以构建产品词编码模型; 使用所述产品词编码模型提取同义词典中每个产品词的特征向量; 根据所述产品词的特征向量判断所述同义词典中的词条是否为歧义词, 其中, 每个词条包括不少于一个产品词; 对判定为歧义词的词条进行切分构建歧义词典。

根据本公开的另一个实施例, 歧义词典构建模块 (图中未示出) 在对训练集中的描述语句进行训练以构建产品词编码模型时, 还用于: 对训练集中的每个描述语句, 从所述描述语句中抽取产品词, 得到产品词列表; 基于所述产品词列表对语义表示模型进行训练, 构建产品词编码模型, 用于对所述产品词列表中的每个产品词进行特征提取。

根据本公开的又一个实施例, 歧义词典构建模块 (图中未示出) 在使用所述产品词编码模型提取同义词典中每个产品词的特征向量时, 还用于: 对同义词典中的每个产品词, 从训练集包括的描述语句中, 提取包含所述产品词的描述语句集合; 对所述描述语句集合中的每个描述语句, 利用所述产品词编码模型计算<描述语句, 所述产品词>的特征向量, 作为所述描述语句对应的特征向量; 对得到的所述产品词的所有描述语句对应的特征向量取平均值, 得到所述产品词的特征

向量。

根据本公开的又一个实施例，歧义词典构建模块（图中未示出）在根据所述产品词的特征向量判断所述同义词典中的词条是否为歧义词时，还用于：对所述同义词典中的每个词条，获取所述词条包括的产品词及产品词的特征向量；根据所述产品词的特征向量进行产品词聚类，并获取聚类类簇的个数；在所述聚类类簇的个数大于 1 的情况下，判定所述词条为歧义词；

并且，在对判定为歧义词的词条进行切分构建歧义词典时，还用于：将判定为歧义词的词条的每个聚类类簇作为一个新词条，构建歧义词典。

根据本公开的又一个实施例，歧义词典构建模块（图中未示出）在根据所述产品词的特征向量进行产品词聚类时，还用于：构建产品词的全连通图，所述全连通图的顶点为产品词，边为两个产品词之间的相似度，所述相似度根据产品词的特征向量来计算；使用图切分方法对所述全连通图进行切分，得到类簇以进行产品词聚类。

根据本公开的又一个实施例，歧义词典构建模块（图中未示出）还用于：将所述新词条对应的聚类类簇的类簇向量作为所述新词条的词条向量，所述类簇向量为所述聚类类簇中包括的产品词的特征向量的平均值。

根据本公开实施例的技术方案，通过遍历待消歧产品词列表，将歧义词典中出现的 product 词确定为歧义 product 词；根据歧义词典获取歧义 product 词对应的候选消歧词条及候选消歧词条的词条向量；根据歧义 product 词所在的描述语句计算歧义 product 词的特征向量，描述语句包括多个堆叠在一起但不相关的词语；将歧义 product 词的特征向量与候选消歧词条的词条向量进行相似比较以确定歧义 product 词的替换词，使用替换词替换歧义 product 词以进行 product 词的消歧处理，可以基于构建的歧义词典

来确定歧义产品词，无需进行句法分析即可确定歧义产品词，更适合结构性不强的文本，词语消歧效果好、精确度高，提高了商品匹配结果和网页信息的检索结果的准确度；基于产品词的同义词典来构建歧义词典，使用同义词典中的同义词作为替换词来进行消歧，避免了使用上下位词来进行消歧，使得消歧后的文本理解变得简单，输出结果在表意上更具体明确，输出结果与原始标题产品词在同一空间，更容易适配下游应用；同时由于产品词歧义词典的构建成本低，故而利于快速收录新产品词，词典扩充速度快。

图 7 示出了可以应用本公开实施例的产品词的消歧方法或产品词的消歧装置的示例性系统架构 700。

如图 7 所示，系统架构 700 可以包括终端设备 701、702、703，网络 704 和服务器 705。网络 704 用以在终端设备 701、702、703 和服务器 705 之间提供通信链路的介质。网络 704 可以包括各种连接类型，例如有线、无线通信链路或者光纤电缆等等。

用户可以使用终端设备 701、702、703 通过网络 704 与服务器 705 交互，以接收或发送消息等。终端设备 701、702、703 上可以安装有各种通讯客户端应用，例如购物类应用、网页浏览器应用、搜索类应用、电商平台、文本消歧类应用等（仅为示例）。

终端设备 701、702、703 可以是具有显示屏并且支持网页浏览的各种电子设备，包括但不限于智能手机、平板电脑、膝上型便携计算机和台式计算机等等。

服务器 705 可以是提供各种服务的服务器，例如对用户利用终端设备 701、702、703 所发来的文本消歧请求提供支持的后台管理服务器（仅为示例）。后台管理服务器可以对接收到的标签词消歧请求等数据进行遍历待消歧产品词列表，将歧义词典中出现的 product 词确定为

歧义产品词；根据所述歧义词典获取所述歧义产品词对应的候选消歧
词条及所述候选消歧词条的词条向量；根据所述歧义产品词所在的描
述语句计算所述歧义产品词的特征向量，所述描述语句包括多个堆叠
在一起但不相关的词语；将所述歧义产品词的特征向量与所述候选消
歧词条的词条向量进行相似比较以确定所述歧义产品词的替换词，使
用所述替换词替换所述歧义产品词等处理，并将处理结果（例如消歧
处理结果--仅为示例）反馈给终端设备。

需要说明的是，本公开实施例所提供的产品词的消歧方法一般由
服务器 705 执行，相应地，产品词的消歧装置一般设置于服务器 705
中。

应该理解，图 7 中的终端设备、网络和服务器的数目仅仅是示意
性的。根据实现需要，可以具有任意数目的终端设备、网络和服务器。

下面参考图 8，其示出了适于用来实现本公开实施例的终端设备或
服务器的计算机系统 800 的结构示意图。图 8 示出的终端设备或服务
器仅仅是一个示例，不应对本公开实施例的功能和使用范围带来任何
限制。

如图 8 所示，计算机系统 800 包括中央处理单元（CPU）801，其
可以根据存储在只读存储器（ROM）802 中的程序或者从存储部分 808
加载到随机访问存储器（RAM）803 中的程序而执行各种适当的动作
和处理。在 RAM 803 中，还存储有系统 800 操作所需的各种程序和数
据。CPU 801、ROM 802 以及 RAM 803 通过总线 804 彼此相连。输入/
输出（I/O）接口 805 也连接至总线 804。

以下部件连接至 I/O 接口 805：包括键盘、鼠标等的输入部分 806；
包括诸如阴极射线管（CRT）、液晶显示器（LCD）等以及扬声器等的
输出部分 807；包括硬盘等的存储部分 808；以及包括诸如 LAN 卡、

调制解调器等的网络接口卡的通信部分 809。通信部分 809 经由诸如因特网的网络执行通信处理。驱动器 810 也根据需要连接至 I/O 接口 805。可拆卸介质 811，诸如磁盘、光盘、磁光盘、半导体存储器等等，根据需要安装在驱动器 810 上，以便于从其上读出的计算机程序根据需要被安装入存储部分 808。

特别地，根据本公开公开的实施例，上文参考流程图描述的过程可以被实现为计算机软件程序。例如，本公开公开的实施例包括一种计算机程序产品，其包括承载在计算机可读介质上的计算机程序，该计算机程序包含用于执行流程图所示的方法的程序代码。在这样的实施例中，该计算机程序可以通过通信部分 809 从网络上被下载和安装，和/或从可拆卸介质 811 被安装。在该计算机程序被中央处理单元（CPU）801 执行时，执行本公开的系统限定上述功能。

需要说明的是，本公开所示的计算机可读介质可以是计算机可读信号介质或者计算机可读存储介质或者是上述两者的任意组合。计算机可读存储介质例如可以是一——但不限于——电、磁、光、电磁、红外线、或半导体的系统、装置或器件，或者任意以上的组合。计算机可读存储介质的更具体的例子可以包括但不限于：具有一个或多个导线的电连接、便携式计算机磁盘、硬盘、随机访问存储器（RAM）、只读存储器（ROM）、可擦式可编程只读存储器（EPROM 或闪存）、光纤、便携式紧凑磁盘只读存储器（CD-ROM）、光存储器件、磁存储器件、或者上述的任意合适的组合。在本公开中，计算机可读存储介质可以是任何包含或存储程序的有形介质，该程序可以被指令执行系统、装置或者器件使用或者与其结合使用。而在本公开中，计算机可读的信号介质可以包括在基带中或者作为载波一部分传播的数据信号，其中承载了计算机可读的程序代码。这种传播的数据信号可以采用多种形式，包括但不限于电磁信号、光信号或上述的任意合适的组合。计算机可读的信号介质还可以是计算机可读存储介质以外的任何计算机可读介质，该计算机可读介质可以发送、传播或者传输用于由

指令执行系统、装置或者器件使用或者与其结合使用的程序。计算机可读介质上包含的程序代码可以用任何适当的介质传输，包括但不限于：无线、电线、光缆、RF 等等，或者上述的任意合适的组合。

5 附图中的流程图和框图，图示了按照本公开各种实施例的系统、方法和计算机程序产品的可能实现的体系架构、功能和操作。在这点上，流程图或框图中的每个方框可以代表一个模块、程序段、或代码的一部分，上述模块、程序段、或代码的一部分包含一个或多个用于实现规定的逻辑功能的可执行指令。也应当注意，在有些作为替换的实现中，方框中所标注的功能也可以以不同于附图中所标注的顺序发生。例如，两个接连地表示的方框实际上可以基本并行地执行，它们有时也可以按相反的顺序执行，这依所涉及的功能而定。也要注意的是，框图或流程图中的每个方框、以及框图或流程图中的方框的组合，可以用执行规定的功能或操作的专用的基于硬件的系统来实现，或者
10 可以用专用硬件与计算机指令的组合来实现。
15

描述于本公开实施例中所涉及到的单元或模块可以通过软件的方式实现，也可以通过硬件的方式来实现。所描述的单元或模块也可以设置在处理器中，例如，可以描述为：一种处理器包括歧义产品词确定模块、候选消歧词条确定模块、特征向量计算模块和歧义产品词替换模块。其中，这些单元或模块的名称在某种情况下并不构成对该单元或模块本身的限定，例如，歧义产品词确定模块还可以被描述为“用于遍历待消歧产品词列表，将歧义词典中出现的产品词确定为歧义产品词的模块”。
20

25 作为另一方面，本公开还提供了一种计算机可读介质，该计算机可读介质可以是上述实施例中描述的设备中所包含的；也可以是单独存在，而未装配入该设备中。上述计算机可读介质承载有一个或者多个程序，当上述一个或者多个程序被一个该设备执行时，使得该设备
30 包括：遍历待消歧产品词列表，将歧义词典中出现的产品词确定为歧

义产品词；根据所述歧义词典获取所述歧义产品词对应的候选消歧词条及所述候选消歧词条的词条向量；根据所述歧义产品词所在的描述语句计算所述歧义产品词的特征向量，所述描述语句包括多个堆叠在一起但不相关的词语；将所述歧义产品词的特征向量与所述候选消歧词条的词条向量进行相似比较以确定所述歧义产品词的替换词，使用所述替换词替换所述歧义产品词以进行产品词的消歧处理。

根据本公开实施例的技术方案，通过遍历待消歧产品词列表，将歧义词典中出现的产品词确定为歧义产品词；根据歧义词典获取歧义产品词对应的候选消歧词条及候选消歧词条的词条向量；根据歧义产品词所在的描述语句计算歧义产品词的特征向量，描述语句包括多个堆叠在一起但不相关的词语；将歧义产品词的特征向量与候选消歧词条的词条向量进行相似比较以确定歧义产品词的替换词，使用替换词替换歧义产品词以进行产品词的消歧处理，可以基于构建的歧义词典来确定歧义产品词，无需进行句法分析即可确定歧义产品词，更适合结构性不强的文本，词语消歧效果好、精确度高，提高了商品匹配结果和网页信息的检索结果的准确度；基于产品词的同义词典来构建歧义词典，使用同义词典中的同义词作为替换词来进行消歧，避免了使用上下位词来进行消歧，使得消歧后的文本理解变得简单，输出结果在表意上更具体明确，输出结果与原始标题产品词在同一空间，更容易适配下游应用；同时由于产品词歧义词典的构建成本低，故而利于快速收录新产品词，词典扩充速度快。

上述具体实施方式，并不构成对本公开保护范围的限制。本领域技术人员应该明白的是，取决于设计要求和因素，可以发生各种各样的修改、组合、子组合和替代。任何在本公开的精神和原则之内所作的修改、等同替换和改进等，均应包含在本公开保护范围之内。

权利要求

1. 一种产品词的消歧方法，包括：

5 遍历待消歧产品词列表，将歧义词典中出现的 product 词确定为歧义 product 词；

根据所述歧义词典获取所述歧义 product 词对应的候选消歧词条及所述候选消歧词条的词条向量；

10 根据所述歧义 product 词所在的描述语句计算所述歧义 product 词的特征向量，所述描述语句包括多个堆叠在一起但不相关的词语；

将所述歧义 product 词的特征向量与所述候选消歧词条的词条向量进行相似比较以确定所述歧义 product 词的替换词，使用所述替换词替换所述歧义 product 词以进行 product 词的消歧处理。

15 2. 根据权利要求 1 所述的方法，其中，所述歧义词典是通过以下方式构建的：

对训练集包括的描述语句进行训练以构建 product 词编码模型；

使用所述 product 词编码模型提取同义词典中每个 product 词的特征向量；

20 根据所述 product 词的特征向量判断所述同义词典中的词条是否为歧义词，其中，每个词条包括不少于一个 product 词；

对判定为歧义词的词条进行切分构建歧义词典。

25 3. 根据权利要求 2 所述的方法，其中，对训练集中的描述语句进行训练以构建 product 词编码模型，包括：

对训练集中的每个描述语句，从所述描述语句中抽取 product 词，得到 product 词列表；

基于所述 product 词列表对语义表示模型进行训练，构建 product 词编码模型，用于对所述 product 词列表中的每个 product 词进行特征提取。

30

4. 根据权利要求 2 所述的方法，其中，使用所述产品词编码模型提取同义词典中每个产品词的特征向量，包括：

对同义词典中的每个产品词，从训练集包括的描述语句中，提取包含所述产品词的描述语句集合；

5 对所述描述语句集合中的每个描述语句，利用所述产品词编码模型计算<描述语句，所述产品词>的特征向量，作为所述描述语句对应的特征向量；

对得到的所述产品词的所有描述语句对应的特征向量取平均值，得到所述产品词的特征向量。

10

5. 根据权利要求 2 所述的方法，其中，根据所述产品词的特征向量判断所述同义词典中的词条是否为歧义词，包括：

对所述同义词典中的每个词条，获取所述词条包括的产品词及产品词的特征向量；

15 根据所述产品词的特征向量进行产品词聚类，并获取聚类类簇的个数；

在所述聚类类簇的个数大于 1 的情况下，判定所述词条为歧义词；并且，对判定为歧义词的词条进行切分构建歧义词典，包括：

20 将判定为歧义词的词条的每个聚类类簇作为一个新词条，构建歧义词典。

6. 根据权利要求 5 所述的方法，其中，根据所述产品词的特征向量进行产品词聚类，包括：

25 构建产品词的全连通图，所述全连通图的顶点为产品词，边为两个产品词之间的相似度，所述相似度根据产品词的特征向量来计算；

使用图切分方法对所述全连通图进行切分，得到类簇以进行产品词聚类。

7. 根据权利要求 5 所述的方法，进一步包括：

30 将所述新词条对应的聚类类簇的类簇向量作为所述新词条的词条

向量，所述类簇向量为所述聚类类簇中包括的产品词的特征向量的平均值。

8. 一种产品词的消歧装置，包括：

5 歧义产品词确定模块，用于遍历待消歧产品词列表，将歧义词典中出现的 product 词确定为歧义产品词；

候选消歧词条确定模块，用于根据所述歧义词典获取所述歧义产品词对应的候选消歧词条及所述候选消歧词条的词条向量；

10 特征向量计算模块，用于根据所述歧义产品词所在的描述语句计算所述歧义产品词的特征向量，所述描述语句包括多个堆叠在一起但不相关的词语；

歧义产品词替换模块，用于将所述歧义产品词的特征向量与所述候选消歧词条的词条向量进行相似比较以确定所述歧义产品词的替换词，使用所述替换词替换所述歧义产品词以进行产品词的消歧处理。

15

9. 一种产品词消歧的电子设备，包括：

一个或多个处理器；

存储装置，用于存储一个或多个程序，

20 当所述一个或多个程序被所述一个或多个处理器执行，使得所述一个或多个处理器实现如权利要求 1-7 中任一所述的方法。

20

10. 一种计算机可读介质，其上存储有计算机程序，所述程序被处理器执行时实现如权利要求 1-7 中任一所述的方法。

25

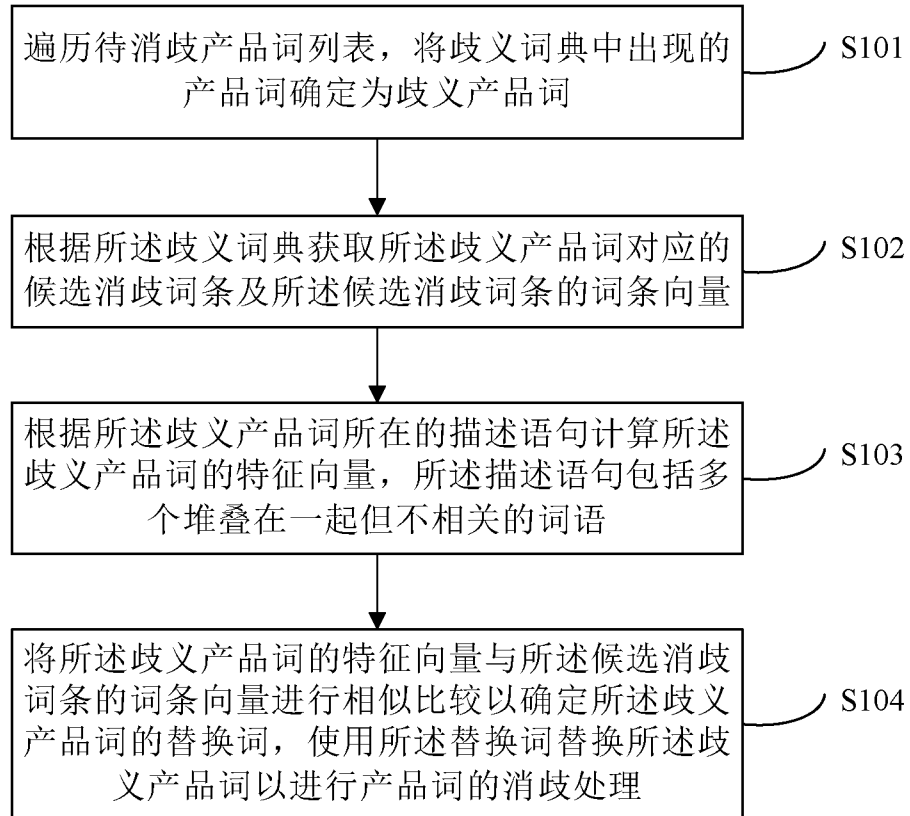
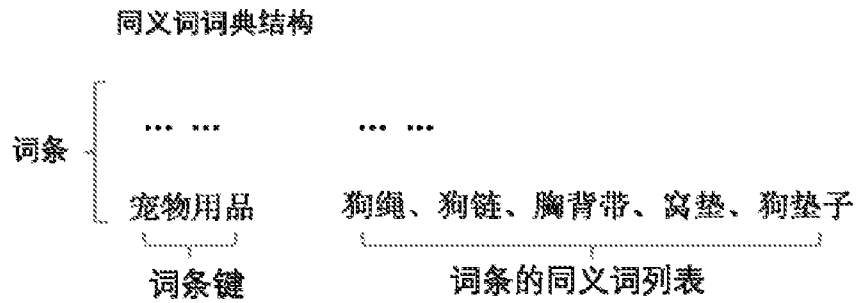
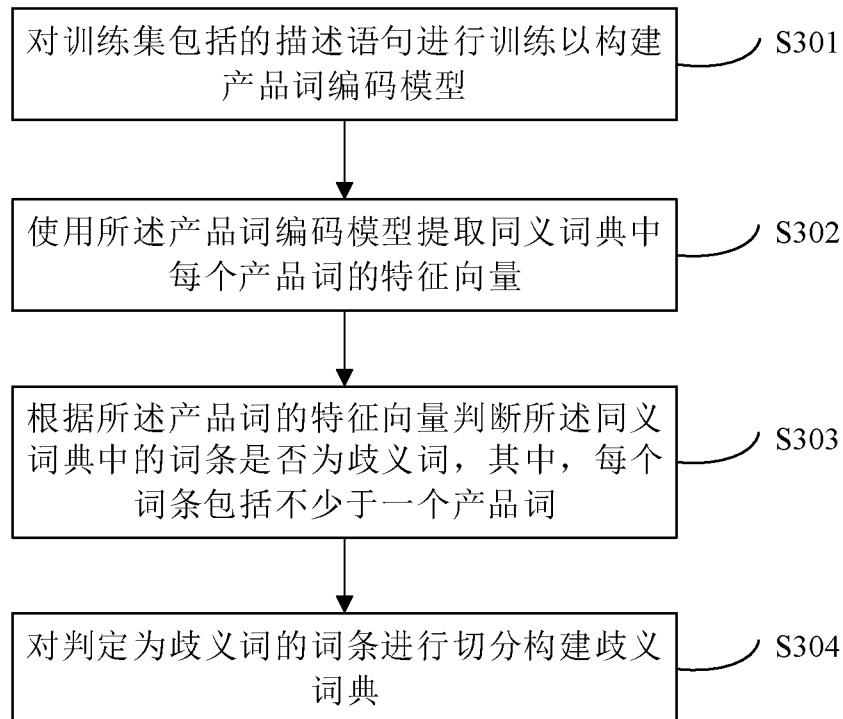
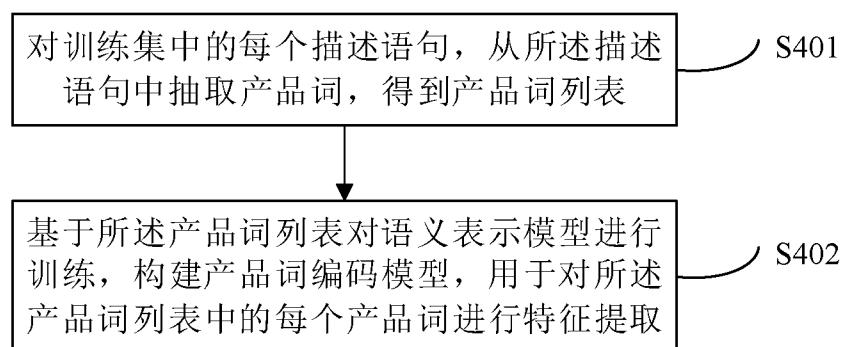


图 1

**图 2****图 3****图 4**

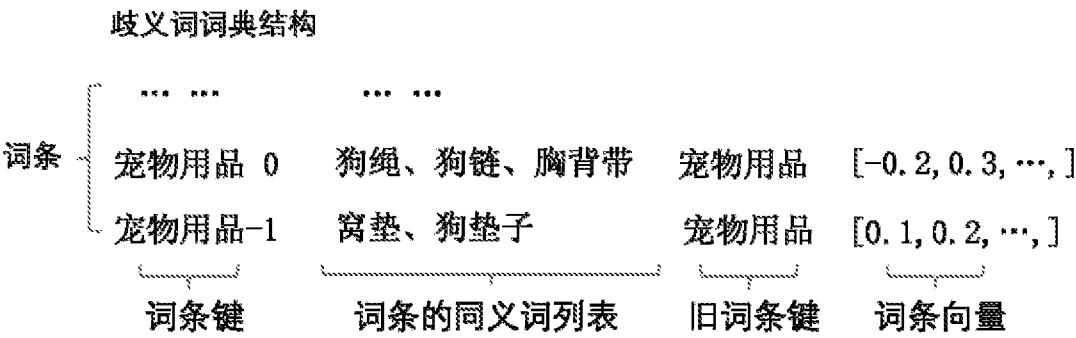


图 5

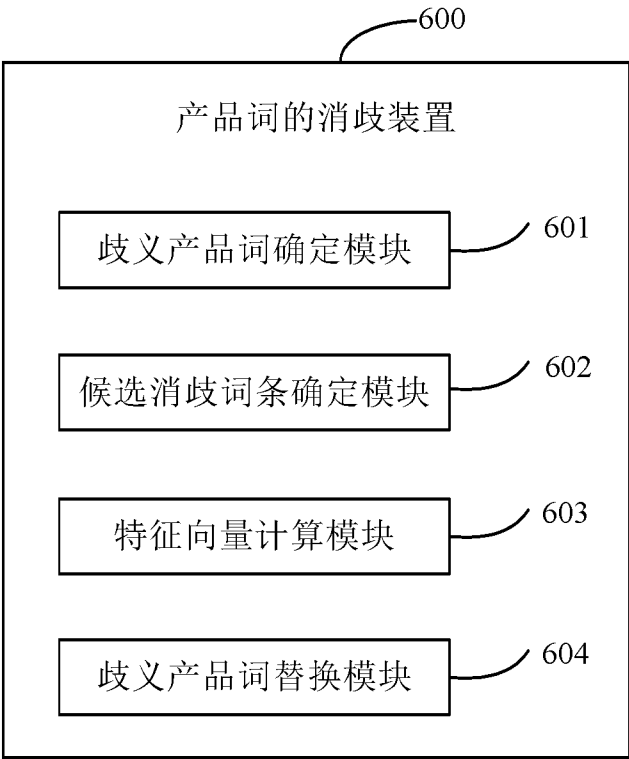


图 6

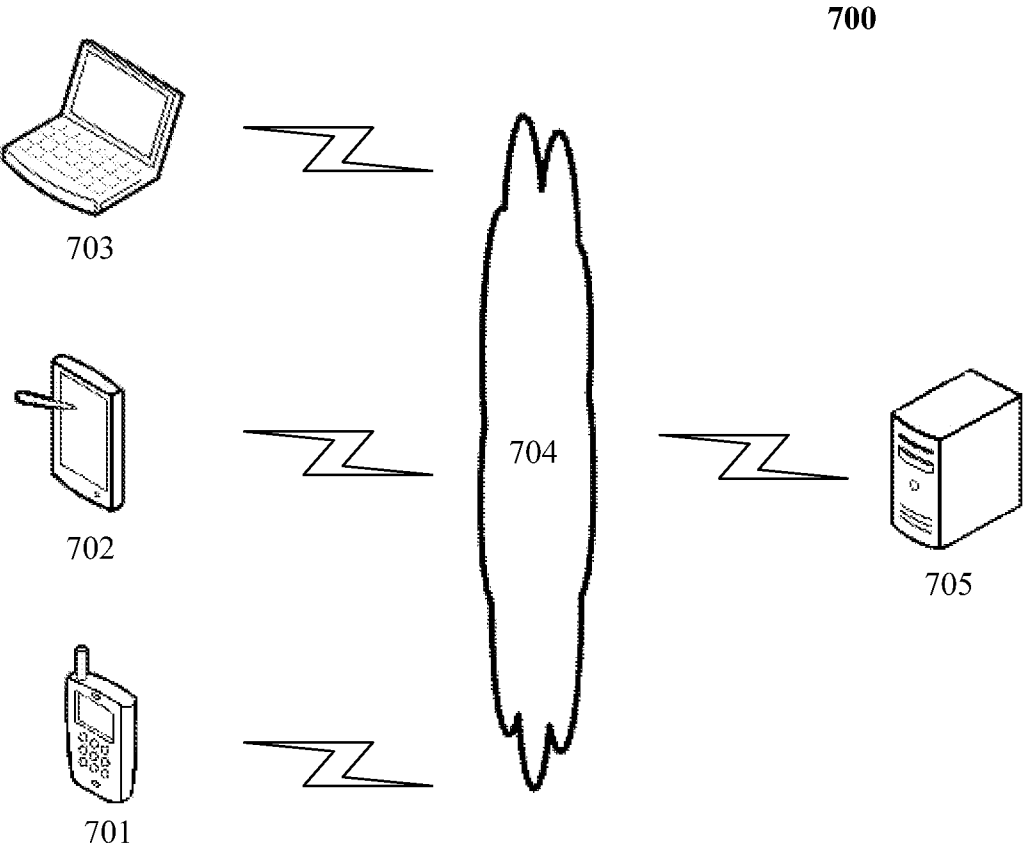


图 7

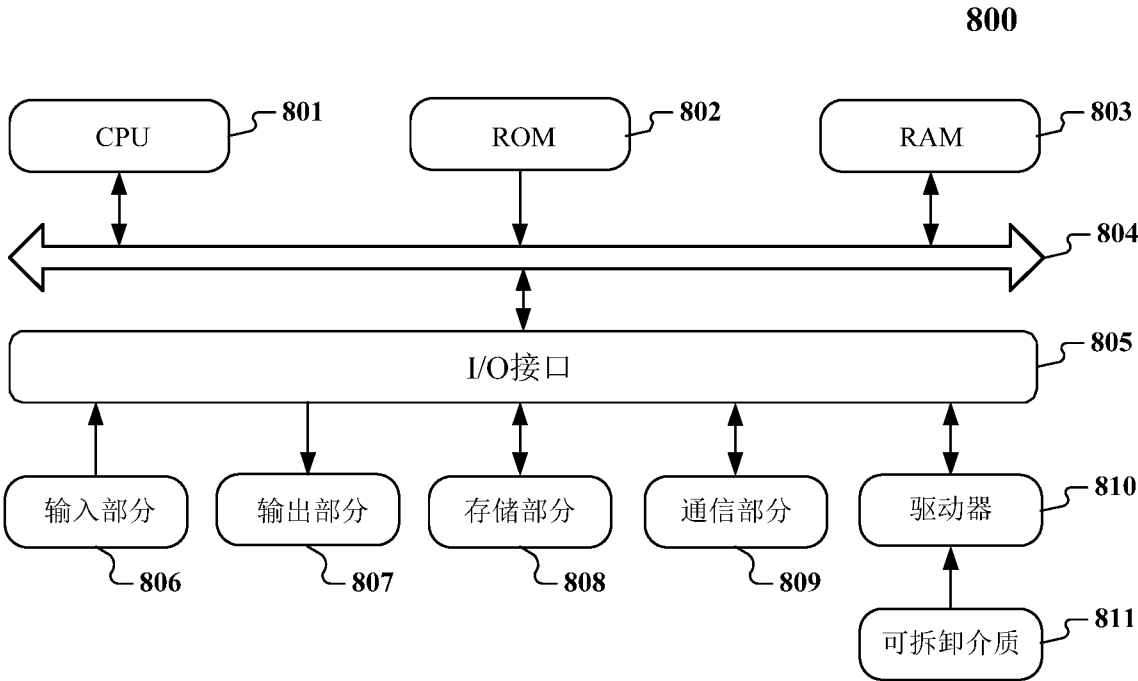


图 8

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2023/071668

A. CLASSIFICATION OF SUBJECT MATTER

G06F 40/284(2020.01)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)

VEN, CNABS, CNTXT, WOTXT, EPTXT, USTXT, IEEE, ISI, CNKI: 消歧, 产品词, 词典, 候选, 词条, 向量, 商品, 标题, 相似, 替换, disambiguation, product word, dictionary, candidate, entries, vector, title, similar, replace

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	CN 107102989 A (NANJING UNIVERSITY) 29 August 2017 (2017-08-29) description, paragraphs [0008]-[0013]	1-10
Y	CN 111428486 A (BEIJING WODONG TIANJUN INFORMATION TECHNOLOGY CO., LTD.) 17 July 2020 (2020-07-17) description, paragraphs [0035]-[0062]	1-10
A	CN 108073570 A (FOCUS TECHNOLOGY CO., LTD.) 25 May 2018 (2018-05-25) entire document	1-10
A	CN 111986007 A (BEIJING ZHIDEMAI TECHNOLOGY CO., LTD.) 24 November 2020 (2020-11-24) entire document	1-10



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance

“D” document cited by the applicant in the international application

“E” earlier application or patent but published on or after the international filing date

“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)

“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

“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

“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

“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

“&” document member of the same patent family

Date of the actual completion of the international search

17 March 2023

Date of mailing of the international search report

22 March 2023

Name and mailing address of the ISA/CN

China National Intellectual Property Administration (ISA/
CN)
China No. 6, Xitucheng Road, Jimenqiao, Haidian District,
Beijing 100088

Facsimile No. (86-10)62019451

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2023/071668

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN	107102989	A	29 August 2017	None	
CN	111428486	A	17 July 2020	None	
CN	108073570	A	25 May 2018	None	
CN	111986007	A	24 November 2020	None	

A. 主题的分类

G06F 40/284(2020.01) 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))

VEN, CNABS, CNTXT, WOTXT, EPTXT, USTXT, IEEE, ISI, CNKI: 消歧, 产品词, 词典, 候选, 词条, 向量, 商品, 标题, 相似, 替换, disambiguation, product word, dictionary, candidate, entries, vector, title, similar, replace

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
Y	CN 107102989 A (南京大学) 2017年8月29日 (2017 - 08 - 29) 说明书第[0008]-[0013]段	1-10
Y	CN 111428486 A (北京沃东天骏信息技术有限公司) 2020年7月17日 (2020 - 07 - 17) 说明书第[0035]-[0062]段	1-10
A	CN 108073570 A (焦点科技股份有限公司) 2018年5月25日 (2018 - 05 - 25) 全文	1-10
A	CN 111986007 A (北京值得买科技股份有限公司) 2020年11月24日 (2020 - 11 - 24) 全文	1-10

☐ 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

“D” 申请人在国际申请中引证的文件

“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

2023年3月17日

国际检索报告邮寄日期

2023年3月22日

ISA/CN的名称和邮寄地址

中国国家知识产权局
中国北京市海淀区蓟门桥西土城路6号 100088

传真号 (86-10)62019451

授权官员

李娜

电话号码 (+86) 010-53961346

PCT/ISA/210 表(第2页) (2022年7月)

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2023/071668

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	107102989	A	2017年8月29日	无	
CN	111428486	A	2020年7月17日	无	
CN	108073570	A	2018年5月25日	无	
CN	111986007	A	2020年11月24日	无	