(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 109255127 A
(43)申请公布日 2019.01.22

(21)申请号 201811132214.9

(22)申请日 2018.09.27

(71)申请人 华东师范大学
地址 200062 上海市普陀区中山北路3663号
申请人 上海博预网络科技有限公司

(72)发明人 史建琦 李志辉 黄滟鸿 鲍钰 战云龙 孙文圣

(74)专利代理机构 北京辰权知识产权代理有限公司 11619
代理人 刘广达

(51)Int.Cl.
G06F 17/27(2006.01)

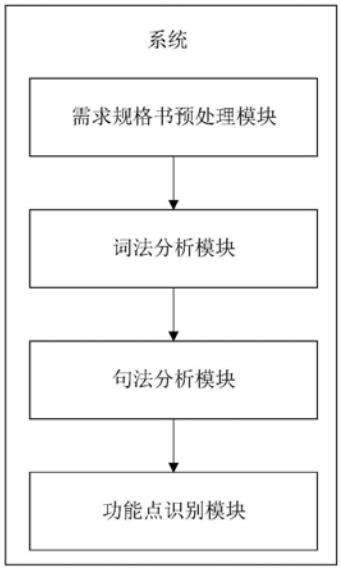
权利要求书1页 说明书4页 附图1页

(54)发明名称

一种需求功能点智能识别系统

(57)摘要

本发明公开了一种需求功能点识别系统，包括：需求规格书预处理模块，用于将需求规格书进行段落拆分，将需求规格书的多级标题剔除，得到初级规格书；词法分析模块，用于将初级规格书进行分词、词性标注、命名实体识别，产生词法分析结果；句法分析模块，用于将词法分析结果进行句法分析，产生句法分析结果；功能点识别模块，用于根据句法分析结果进行功能点识别，并统计功能点类别和数量。本发明通过结合自然语言处理(NLP)技术对需求规格书进行智能分析，实现对需求规格书所含功能的准确快速统计，代替人工分析，提高效率，降低成本。



CN 109255127 A

1.一种需求功能点识别系统,其特征在于,包括:

需求规格书预处理模块,用于将需求规格书进行段落拆分,将所述需求规格书的多级标题剔除,得到初级规格书;

词法分析模块,用于将所述初级规格书进行分词、词性标注、命名实体识别,产生词法分析结果;

句法分析模块,用于将所述词法分析结果进行句法分析,产生句法分析结果;

功能点识别模块,用于根据所述句法分析结果进行功能点识别,并统计功能点类别和数量。

2.如权利要求1所述的识别系统,其特征在于,所述需求规格书为需要分析的中文版的需求规格书,所述需求规格书预处理模块利用Lucene框架将所述需求规格书进行段落拆分。

3.如权利要求1所述的识别系统,其特征在于,所述词法分析模块包括:

分词单元,采用基于最大熵分词方法,将字符串频率统计和字符串匹配结合;

词性标注单元,采用基于最大熵的词性标注方法,以高频词性为依据进行标注;

命名实体识别单元,采用条件随机场算法作为判别式概率模型。

4.如权利要求3所述的识别系统,其特征在于,所述词法分析模块采用神经网络模型,进行分词、词性标注、命名实体识别。

5.如权利要求4所述的识别系统,其特征在于,所述神经网络模型词法分析模块利用AC多模式匹配算法实现分词,或者利用基于所述需求规格书中的自定义词典进行分词,所述分词策略采用字典词汇最长匹配原则。

6.如权利要求1所述的识别系统,其特征在于,所述句法分析模块进行句法分析包括:句子中词语的依赖关系分析、搭配关系分析。

7.如权利要求6所述的识别系统,其特征在于,所述句法分析模块利用基于最大熵模型的最大生成树算法进行所述需求规格书的中文依存句法的分析,最大熵依存利用条件概率模型,将所有依存关系概率的累积作为目标函数的打分,取打分最大的依存关系树作为输出。

8.如权利要求1所述的识别系统,其特征在于,所述句法分析结果以CoNLL格式输出,所述句法分析结果包括:当前词语在句子中的序号、当前词的词性、当前词语的句法特征、前词语的中心词、当前词语与中心词语的依存关系。

9.如权利要求1所述的识别系统,其特征在于,所述功能点识别模块根据所述句法分析结果、自定义的功能点关键词、自定义的目标匹配关系,精细匹配功能点,最终统计并分类输出。

10.如权利要求9所述的识别系统,其特征在于,所述自定义的目标匹配关系包括:主谓关系、定中关系、动宾关系。

一种需求功能点智能识别系统

技术领域

[0001] 本发明涉及自然语言处理和软件工程秀分析领域，特别涉及一种需求功能点智能识别系统。

背景技术

[0002] 在传统的需求功能点评估领域，功能点的判断识别有两种处理情况，其一，不将功能点识别纳入考虑范围，不去评估这些功能所代表的工作量与价值，导致软件的外包定制存在不统一的价格要求与时间要求。其二，软件功能评估依靠人工分析来识别。因为需求规格书往往多达百页，甚至更多。所以，这种以人工为主的做法，不仅需要资深的领域专家介入，而且非常的耗费时间和精力。间接的增加了软件工程的环节，增加了软件开发的费用。同时人工分析存在不确定性，不稳定性。这些原因导致需求功能点识别不够智能和高效。

[0003] 随着社会生产领域大量的软件需求的出现，需求规格书也急速增加，而针对软件功能评估人才还很缺乏。大量的需求功能识别评估工作，亟须一种自动而又智能的方法来处理。

发明内容

[0004] 本发明的目的是通过以下技术方案实现的。为处理上述问题，本发明构造一种基于自然语言处理(NLP)技术的功能点智能识别系统。本发明构造的智能系统可有效挖掘需求规格书中的功能点，通过结合NLP技术对需求规格书逐段，逐句，逐词的拆解分析。挖掘出每一个词的词性，识别句子中的命名实体和提取出句子中各个部分之间的依存关系，通过句法关系匹配和关键词匹配，最终实现功能点的识别并归类。减少人类分析提取的低效和不稳定性。

[0005] 一种需求功能点识别系统，包括：

[0006] 需求规格书预处理模块，用于将需求规格书进行段落拆分，将所述需求规格书的多级标题剔除，得到初级规格书；

[0007] 词法分析模块，用于将所述初级规格书进行分词、词性标注、命名实体识别，产生词法分析结果；

[0008] 句法分析模块，用于将所述词法分析结果进行句法分析，产生句法分析结果；

[0009] 功能点识别模块，用于根据所述句法分析结果进行功能点识别，并统计功能点类别和数量。

[0010] 优选地，需求规格书为需要分析的中文版的需求规格书，所述需求规格书预处理模块利用Lucene框架将所述需求规格书进行段落拆分。

[0011] 优选地，词法分析模块包括：

[0012] 分词单元，采用基于最大熵分词方法，将字符串频率统计和字符串匹配结合；

[0013] 词性标注单元，采用基于最大熵的词性标注方法，以高频词性为依据进行标注；

[0014] 命名实体识别单元，采用条件随机场算法作为判别式概率模型。

[0015] 优选地，词法分析模块采用神经网络模型，进行分词、词性标注、命名实体识别。

[0016] 优选地，所述神经网络模型词法分析模块利用AC多模式匹配算法实现分词，或者利用基于所述需求规格书中的自定义词典进行分词，所述分词策略采用字典词汇最长匹配原则。

[0017] 优选地，句法分析模块进行句法分析包括：句子中词语的依赖关系分析、搭配关系分析。

[0018] 优选地，句法分析模块利用基于最大熵模型的最大生成树算法进行所述需求规格书的中文依存句法的分析，最大熵依存利用条件概率模型，将所有依存关系概率的累积作为目标函数的打分，取打分最大的依存关系树作为输出。

[0019] 优选地，句法分析结果以CoNLL格式输出，所述句法分析结果包括：当前词语在句子中的序号、当前词的词性、当前词语的句法特征、前词语的中心词、当前词语与中心词语的依存关系。

[0020] 优选地，功能点识别模块根据所述句法分析结果、自定义的功能点关键词、自定义的目标匹配关系，精细匹配功能点，最终统计并分类输出。

[0021] 优选地，自定义的目标匹配关系包括：主谓关系、定中关系、动宾关系。

[0022] 本发明的优点在于：基于自然语言处理，其中的词法分析，句法分析所依据的神经网络模型，可以不断学习分析过的文档，具有自主进化能力，不断提高处理的准确率。该发明有效提高文档功能点挖掘的效率，降低成本。将人类从文档分析挖掘中解放出来。

附图说明

[0023] 通过阅读下文优选实施方式的详细描述，各种其他的优点和益处对于本领域普通技术人员将变得清楚明了。附图仅用于示出优选实施方式的目的，而并不认为是对本发明的限制。而且在整个附图中，用相同的参考符号表示相同的部件。在附图中：

[0024] 附图1示出了根据本发明实施方式的功能点识别系统的模块图；

[0025] 附图2示出了根据本发明实施方式的功能点识别系统的结构示意图。

具体实施方式

[0026] 下面将参照附图更详细地描述本公开的示例性实施方式。虽然附图中显示了本公开的示例性实施方式，然而应当理解，可以以各种形式实现本公开而不应被这里阐述的实施方式所限制。相反，提供这些实施方式是为了能够更透彻地理解本公开，并且能够将本公开的范围完整的传达给本领域的技术人员。

[0027] 根据本发明的实施方式，提出一种需求功能点识别系统，如图1所示，包括：需求规格书预处理模块，用于将需求规格书进行段落拆分，将所述需求规格书的多级标题剔除，得到初级规格书。词法分析模块，用于将所述初级规格书进行分词、词性标注、命名实体识别，产生词法分析结果。句法分析模块，基于神经网络训练的模型，将词法分析结果进行句法分析，句产生句法分析结果，法分析包括：句子中词语的依赖关系分析、搭配关系分析。功能点识别模块，利用句法分析结果，根据规定的句子成分搭配关系和功能词库，进行功能点识别，并统计功能点类别和数量。

[0028] 本发明提出的功能点智能识别系统基于自然语言处理(NLP)，所述需求规格书预

处理模块基于Lucene框架将整个需求文本提取为结构化的段落，剔除图表和多级标题。实现对原生需求规格书的初步信息提取和处理。需求规格书为需要分析的中文版的需求规格书。

[0029] 需求规格书预处理模块，自动的读入文档数据，采用SVM的理论打分判断文本相似性，基于TF-IDF理论评价词语的重要性，不仅考虑词在文档中的频率，也考虑词在整个文档中的区分度；这些方法有效提高预处理的速度和准确度，当文档页数多的时候，处理用时间明显缩短。

[0030] 基于自然语言处理(NLP)的功能点智能识别系统中，词法分析模块，利用神经网络模型，如图2所示，对句子进行拆分，词性标注，命名实体识别。在分词方面，采用基于最大熵分词方法，该方法将字符串频率统计和字符串匹配结合起来，提高匹配分词的切分速度。在词性标注上，采用基于最大熵的词性标注方法，以高频词性为依据实现标注的准确性；在命名实体识别上，采用条件随机场(CRF)算法，作为判别式概率模型有很强的特征融入能力，该方法可以有效提高命名实体识别的准确率。

[0031] 词法分析模块包括：分词单元，采用基于最大熵分词方法，将字符串频率统计和字符串匹配结合，提高匹配分词的切分速度；词性标注单元，采用基于最大熵的词性标注方法，以高频词性为依据实现标注的准确性；命名实体识别单元，采用条件随机场算法作为判别式概率模型有很强的特征融入能力，提高命名实体识别的准确率。词法分析模块采用神经网络模型进行分词、词性标注、命名实体识别。所述神经网络模型词法分析模块利用AC多模式匹配算法实现分词，或者利用基于所述需求规格书中的自定义词典进行分词，所述分词策略采用字典词汇最长匹配原则。

[0032] 基于自然语言处理(NLP)的功能点智能识别系统中，词法分析模块，利用AC多模式匹配算法将自定义字典中词语和文档中待切分句子进行匹配分词，该算法时间复杂度低，可以有效减少匹配分词时间，提高分词速度。

[0033] 基于自然语言处理(NLP)的功能点智能识别系统中，句法分析模块，利用最大熵模型估计任意两个单词之间最可能的依存关系以及概率，最大熵中的约束通过特征函数来实现，特征函数的使用解决长距离依存问题，提高句法分析的准确率。最大生成树算法在解析时，使用最大生成树搜索整句的最优依存树，具有全局性，能有效提高句法依存分析的准确率。

[0034] 基于自然语言处理(NLP)的功能点智能识别系统中，句法分析模块，通过输出CoNLL格式的分析结果，这种结构化的数据形式可以方便功能点分析模块根据功能分析的不同侧重点进行方便的匹配。句法分析结果包括：当前词语在句子中的序号、当前词的词性、当前词语的句法特征、前词语的中心词、当前词语与中心词语的依存关系。

[0035] 功能点识别模块，利用句法分析结果，根据所述句法分析结果、自定义的功能点关键词、自定义的目标匹配关系，精细匹配功能点，最终统计并分类输出。自定义的目标匹配关系包括：主谓关系、定中关系、动宾关系。自定义的功能点关键词，将含有指定关系和指定关键词的句子匹配成不同的功能，以实现功能的精细化匹配；处理不同领域的需求规格书时，通过调整匹配词，匹配关系进行柔性定制分析。

[0036] 本发明提出的基于自然语言处理(NLP)的功能点智能识别系统，实现将自然语言处理技术应用到对需求规格书中含有功能点的识别中，可在数分钟内对长达百页的需求规

格书智能分析挖掘其中语句的词法关系和句法依存关系。在句法/词法分析的基础上，智能快速挖掘需求中所含有的多种功能点。极大降低需求功能点挖掘的人工成本，使需求功能分析更加智能化，无人化。

[0037] 以上所述，仅为本发明较佳的具体实施方式，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内，可轻易想到的变化或替换，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。

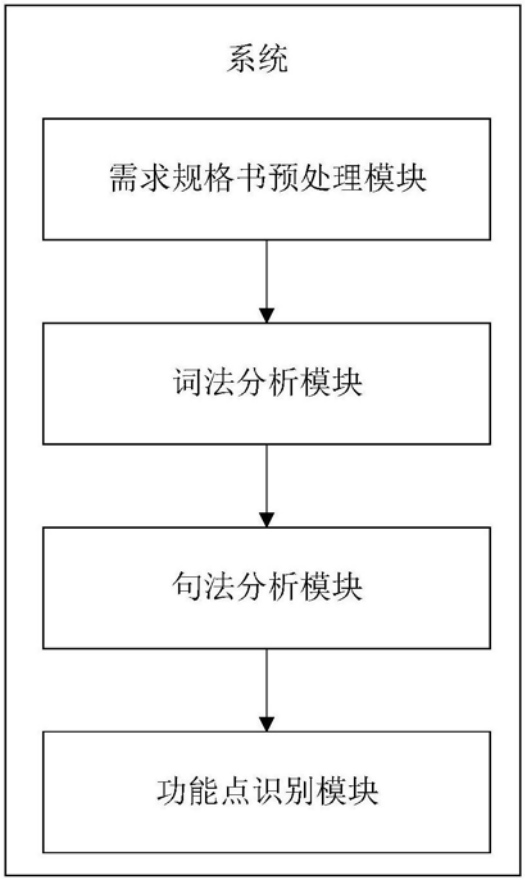

系统
需求规格书预处理模块
词法分析模块
句法分析模块
功能点识别模块

图1

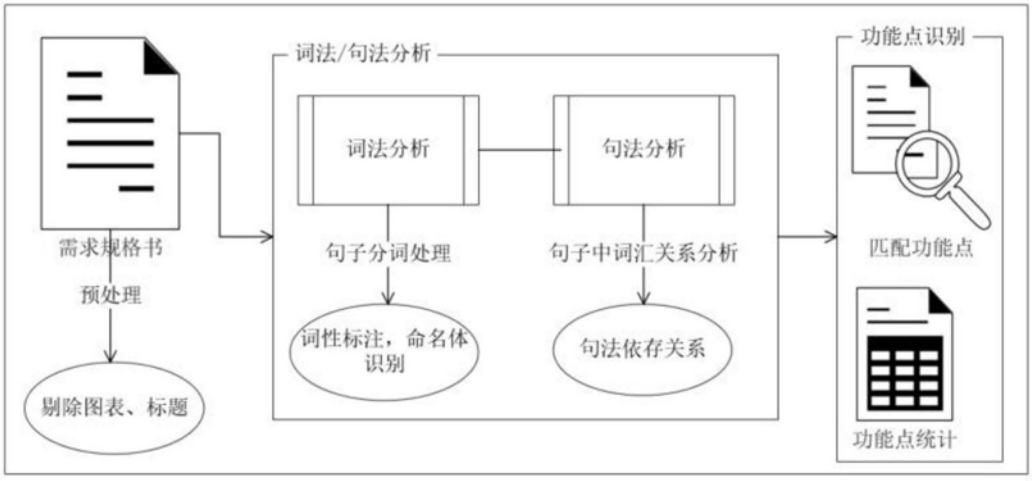

需求规格书
预处理
剔除图表、标题
词法/句法分析
词法分析
句法分析
句子分词处理
句子中词汇关系分析
词性标注，命名体识别
句法依存关系
功能点识别
匹配功能点
功能点统计

图2