

一种提供搜索引擎标签的方法和装置

5 技术领域

本发明涉及一种提供搜索引擎标签的方法和装置。

背景技术

10 目前，用户在电子商务网站上搜索商品时只能根据商品的客观属性来搜索筛选，例如颜色、尺码等，而对于一些带有主观色彩的搜索，例如搜索词为“性价比好的相机”，通常都是无结果。对于主观语义搜索，当前用户一般需要通过通用搜索引擎先查询到一些商品型号再去电商网站上搜索具体商品。这无疑会增加用户操作，并且分析可知通用搜索引擎给出的搜索结果大多基于用户在 BBS 等网站上的评价。

15

电子商务网站本身就拥有相当丰富的用户评论数据，因此现有技术中也基于电子商务网站中的用户评论数据来获取搜索引擎的标签。其中的主要技术路线是从评论文本中自动识别观点信息并分析观点，用以获得用户对商品的各个属性特征的评价，然后将挖掘出的评价和商品关联形成搜索引擎标签。在得到搜索引擎标签之后即可以使用已有的搜索引擎技术向用户提供包含评价数据的搜索服务，该搜索引擎标签能够表达用户的主观意图，于是采用这种搜索引擎标签能够支持向用户提供主观意图搜索服务。

20

25

30

现有技术中的一种得到上述搜索引擎标签的方法是，先根据语义词典识别出评论文本中的观点词，例如：好、棒、不错等，然后通过对观点词上下文的截取得到一个包含观点词的具有合适长度并且语义相对完整的短句，再利用语义分析工具例如斯坦福大学分析器分析该短句，得到一系列依存关系，最后对这些依存关系进行分析，抽取出观点词所修饰的对象即属性词，例如性价比、外观等。属性词也称“非

谓形容词”、“区别词”，是从传统语法中名词、动词、形容词中脱离出来的一类新兴词类。属性词只表示人、事物的属性或特征，具有区别或分类的作用。属性词一般只能作定语，不能作谓语。

5 上述方法中，对观点词的提取依赖于词典，对于词典未收录的词则难以提取观点词，从而提供标签的范围有限。另外上述方法中，在提取属性词之前需要对文本基于观点词作上下文截取，处理方式效率较低。

10 发明内容

有鉴于此，本发明提供一种提供搜索引擎标签的方法和装置，能够在更广的范围内提供搜索引擎标签并且有较高的处理速度。

15 为实现上述目的，根据本发明的一个方面，提供了一种提供搜索引擎标签的方法。

20 本发明的提供搜索引擎标签的方法包括：提取句子中的一个或多个属性词；对所述句子进行依存关系分析，得出针对各个属性词的从含有该属性词到含有观点词的依存关系路径；根据所述依存关系路径提取所述句子中各个所述属性词分别对应的观点词；用所述属性词和所述观点词组成搜索引擎标签。

25 可选地，所述提取句子中的一个或多个属性词的步骤之前还包括：按预设的规则对文本数据进行过滤；从所述文本数据中获取句子。

30 可选地，从所述文本数据中获取句子的步骤包括：对文本数据按标点符号进行子句划分得到短子句；获取所述短子句作为所述句子。

30 可选地，对所述句子进行依存关系分析，得出针对各个属性词的从含有该属性词到含有观点词的依存关系路径的步骤包括：对所述句

子进行依存关系分析得出该句子的一系列依存关系；根据所述属性词和所述一系列依存关系得出针对各个属性词的从含有该属性词经由所述一系列依存关系到含有观点词的依存关系；遍历所述含有观点词的依存关系从而得出所述依存关系路径。

5

可选地，根据所述依存关系路径提取所述句子中各个所述属性词分别对应的观点词的步骤包括：从所述依存关系路径中选择出现频率较高的依存关系路径；根据选择出的依存关系路径得出依存关系规则；根据所述依存关系规则提取所述句子中各个属性词对应的观点词。

10

可选地，用所述属性词和所述观点词组成搜索引擎标签的步骤之后，还包括：根据同义词表，将包含同义的观点词的多个标签合并成一个标签。

15

根据本发明的另一方面，提供了一种提供搜索引擎标签的装置。

20

本发明的提供搜索引擎标签的装置包括：属性词提取模块，用于提取句子中的一个或多个属性词；依存关系分析模块，用于对所述句子进行依存关系分析，得出针对各个属性词的从含有该属性词到含有观点词的依存关系路径；观点词提取模块，用于根据所述依存关系路径提取所述句子中各个所述属性词分别对应的观点词；搜索引擎标签模块，用于用所述属性词和所述观点词组成搜索引擎标签。

25

可选地，还包括预处理模块，用于按预设的规则对文本数据进行过滤，然后从所述文本数据中获取句子。

30

可选地，所述预处理模块还用于对文本数据按标点符号进行子句划分得到短子句，然后获取所述短子句作为所述句子。

可选地，所述依存关系分析模块还用于：对所述句子进行依存关

系分析得出该句子的一系列依存关系；根据所述属性词和所述一系列依存关系得出针对各个属性词的从含有该属性词经由所述一系列依存关系到含有观点词的依存关系；遍历所述含有观点词的依存关系从而得出所述依存关系路径。

5

可选地，所述观点词提取模块还用于：从所述依存关系路径中选择出现频率较高的依存关系路径；根据选择出的依存关系路径得出依存关系规则；根据所述依存关系规则提取所述句子中各个属性词对应的观点词。

10

可选地，还包括归一化模块，用于根据同义词表，将包含同义的观点词的多个标签合并成一个标签。

15

根据本发明的技术方案，挖掘出属性词并根据依存关系挖掘对应的观点词，同时也可以在没有对应的观点词的情况下过滤挖掘的属性词。本实施例的技术方案不依赖于词典，所以有助于在更广的范围内提供搜索引擎标签，而且不需要对语句做上下文截取，有助于提高处理速度。

20

附图说明

附图用于更好地理解本发明，不构成对本发明的不当限定。其中：

图 1 是根据本发明实施例的提供搜索引擎标签的方法的示意图；

图 2 是根据本发明实施例的提供搜索引擎标签的装置的基本结构的示意图。

25

具体实施方式

以下结合附图对本发明的示范性实施例做出说明，其中包括本发明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，而不会背离本发明的范围和精神。同样，为了清

30

楚和简明，以下的描述中省略了对公知功能和结构的描述。

图 1 是根据本发明实施例的提供搜索引擎标签的方法的示意图，如图 1 所示，该方法主要包括步骤 S11 至步骤 S14。

5

10

步骤 S11：提取句子中的一个或多个属性词。可以采用词性模式匹配的方式提取评论语句中的名词（NN）、动词（NN）及名词+动词（NN+VV）等复合形式作为候选属性词。这里的句子是从文本数据中获取，可以先按预设的规则对文本数据进行过滤，然后对文本数据按标点符号进行子句划分得到短子句，将该短子句作为本步骤中的句子。以上述的文本数据是电子商务网站的商品评论信息为例，上述的过滤即为对网站上抓取的原始评论进行预处理，按规则过滤掉评论数据中营销广告、停用词及默认评论等无意义语句，再去除同条评论中大量重复的字段或语句。

15

20

步骤 S12：对步骤 S11 中的句子进行依存关系分析，得出针对各个属性词的从含有该属性词到含有观点词的依存关系路径。在本步骤中，具体可以是先对上述句子进行依存关系分析得出该句子的一系列依存关系，然后根据属性词和这一系列依存关系得出针对各个属性词的从含有该属性词经由这一系列依存关系到含有观点词的依存关系，最后遍历该含有观点词的依存关系从而得出依存关系路径。可以看出本步骤中采用了多个传递依存关系来形成依存关系路径，有助于深入地挖掘出观点词。

25

30

步骤 S13：根据步骤 S12 中的依存关系路径提取句子中各个属性词分别对应的观点词。如果针对某个属性词没有提取到观点词，则将该属性词从步骤 S11 中获得的属性词集合中删除。本步骤中，具体可以先从上述依存关系路径中选择出现频率较高的依存关系路径，然后根据选择出的依存关系路径得出依存关系规则，再根据该依存关系规则提取句子中各个属性词对应的观点词。

步骤 S14：用属性词和观点词组成搜索引擎标签。这里的属性词是步骤 S13 之后的属性词集合。在本步骤之后，可以按搜索引擎标签中的观点词的同义词作一个合并，即根据同义词表，将包含同义的观点词的多个标签合并成一个标签。例如，将“性价比好”、“性价比高”、和“性价比无敌”合并成“性价比高”。

标签可以给商品建索引供用户搜索。但在有些情况下用户自己输入的搜索词可能不是按图 1 所示的步骤得出的属性词，因此可以继续执行步骤 S15。

步骤 S15：输出步骤 S14 中的搜索引擎标签。根据本步骤，搜索引擎标签被呈现在用户使用的终端设备的人机界面中，例如网页上，用户点击这种搜索引擎标签时就将该搜索引擎标签提供给搜索引擎从而启动搜索，使用户能够根据页面中展示的各种属性词来实现商品的筛选。

图 2 是根据本发明实施例的提供搜索引擎标签的装置的基本结构的示意图。如图 2 所示，提供搜索引擎标签的装置 20 基本地包括属性词提取模块 21、依存关系分析模块 22、观点词提取模块 23、和搜索引擎标签模块 24。属性词提取模块 21 用于提取句子中的一个或多个属性词。依存关系分析模块 22 对所述句子进行依存关系分析，得出针对各个属性词的从含有该属性词到含有观点词的依存关系路径。观点词提取模块 23 根据所述依存关系路径提取所述句子中各个所述属性词分别对应的观点词。搜索引擎标签模块 24 用所述属性词和所述观点词组成搜索引擎标签。

提供搜索引擎标签的装置 20 还可以包括预处理模块（图中未示出），用于按预设的规则对文本数据进行过滤，然后从该文本数据中获取句子。预处理模块还可用于对文本数据按标点符号进行子句划分

得到短子句，然后获取该短子句作为所述的句子。

提供搜索引擎标签的装置 20 还可以包括归一化模块（图中未示出），用于根据同义词表，将包含同义的观点词的多个标签合并成一个标签。

依存关系分析模块 22 还可用于对所述句子进行依存关系分析得出该句子的一系列依存关系；根据所述属性词和所述一系列依存关系得出针对各个属性词的从含有该属性词经由所述一系列依存关系到含有观点词的依存关系；遍历所述含有观点词的依存关系从而得出所述依存关系路径。

观点词提取模块 23 还可用于从所述依存关系路径中选择出现频率较高的依存关系路径；根据选择出的依存关系路径得出依存关系规则；根据所述依存关系规则提取所述句子中各个属性词对应的观点词。

根据本发明实施例的技术方案，挖掘出属性词并根据依存关系挖掘对应的观点词，同时也可以在没有对应的观点词的情况下过滤挖掘的属性词。本实施例的技术方案不依赖于词典，所以有助于在更广的范围内提供搜索引擎标签，而且不需要对语句做上下文截取，有助于提高处理速度。

以上结合具体实施例描述了本发明的基本原理，但是，需要指出的是，对本领域的普通技术人员而言，能够理解本发明的方法和设备的全部或者任何步骤或者部件，可以在任何计算装置（包括处理器、存储介质等）或者计算装置的网络中，以硬件、固件、软件或者它们的组合加以实现，这是本领域普通技术人员在阅读了本发明的说明的情况下运用他们的基本编程技能就能实现的。

因此，本发明的目的还可以通过在任何计算装置上运行一个程序

或者一组程序来实现。所述计算装置可以是公知的通用装置。因此，本发明的目的也可以仅仅通过提供包含实现所述方法或者装置的程序代码的程序产品来实现。也就是说，这样的程序产品也构成本发明，并且存储有这样的程序产品的存储介质也构成本发明。显然，所述存储介质可以是任何公知的存储介质或者将来开发出的任何存储介质。

还需要指出的是，在本发明的装置和方法中，显然，各部件或各步骤是可以分解和/或重新组合的。这些分解和/或重新组合应视为本发明的等效方案。并且，执行上述系列处理的步骤可以自然地按照说明的顺序按时间顺序执行，但是并不需要一定按照时间顺序执行。某些步骤可以并行或彼此独立地执行。

上述具体实施方式，并不构成对本发明保护范围的限制。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。任何在本发明的精神和原则之内所作的修改、等同替换和改进等，均应包含在本发明保护范围之内。

权 利 要 求 书

1. 一种提供搜索引擎标签的方法，其特征在于，包括：

提取句子中的一个或多个属性词；

5 对所述句子进行依存关系分析，得出针对各个属性词的从含有该属性词到含有观点词的依存关系路径；

 根据所述依存关系路径提取所述句子中各个所述属性词分别对应的观点词；

 用所述属性词和所述观点词组成搜索引擎标签。

10

2. 根据权利要求 1 所述的方法，其特征在于，

 所述提取句子中的一个或多个属性词的步骤之前还包括：按预设的规则对文本数据进行过滤；

 从所述文本数据中获取句子。

15

3. 根据权利要求 2 所述的方法，其特征在于，从所述文本数据中获取句子的步骤包括：

 对文本数据按标点符号进行子句划分得到短子句；

 获取所述短子句作为所述句子。

20

4. 根据权利要求 1 所述的方法，其特征在于，对所述句子进行依存关系分析，得出针对各个属性词的从含有该属性词到含有观点词的依存关系路径的步骤包括：

 对所述句子进行依存关系分析得出该句子的一系列依存关系；

25 根据所述属性词和所述一系列依存关系得出针对各个属性词的从含有该属性词经由所述一系列依存关系到含有观点词的依存关系；

 遍历所述含有观点词的依存关系从而得出所述依存关系路径。

5. 根据权利要求 1 或 4 所述的方法，其特征在于，根据所述依存关系路径提取所述句子中各个所述属性词分别对应的观点词的步骤包

30

括：

从所述依存关系路径中选择出现频率较高的依存关系路径；

根据选择出的依存关系路径得出依存关系规则；

根据所述依存关系规则提取所述句子中各个属性词对应的观点词。

6. 根据权利要求 1 至 4 中任一项所述的方法，其特征在于，用所述属性词和所述观点词组成搜索引擎标签的步骤之后，还包括：根据同义词表，将包含同义的观点词的多个标签合并成一个标签。

7. 一种提供搜索引擎标签的装置，其特征在于，包括：
属性词提取模块，用于提取句子中的一个或多个属性词；
依存关系分析模块，用于对所述句子进行依存关系分析，得出针对各个属性词的从含有该属性词到含有观点词的依存关系路径；

观点词提取模块，用于根据所述依存关系路径提取所述句子中各个所述属性词分别对应的观点词；

搜索引擎标签模块，用于用所述属性词和所述观点词组成搜索引擎标签。

8. 根据权利要求 7 所述的装置，其特征在于，还包括预处理模块，用于按预设的规则对文本数据进行过滤，然后从所述文本数据中获取句子。

9. 根据权利要求 8 所述的装置，其特征在于，所述预处理模块还用于对文本数据按标点符号进行子句划分得到短子句，然后获取所述短子句作为所述句子。

10. 根据权利要求 7 所述的装置，其特征在于，所述依存关系分析模块还用于：

对所述句子进行依存关系分析得出该句子的一系列依存关系；

根据所述属性词和所述一系列依存关系得出针对各个属性词的从含有该属性词经由所述一系列依存关系到含有观点词的依存关系；

遍历所述含有观点词的依存关系从而得出所述依存关系路径。

5 11. 根据权利要求 7 或 10 所述的装置，其特征在于，所述观点词提取模块还用于：

从所述依存关系路径中选择出现频率较高的依存关系路径；

根据选择出的依存关系路径得出依存关系规则；

10 根据所述依存关系规则提取所述句子中各个属性词对应的观点词。

12. 根据权利要求 7 至 10 中任一项所述的装置，还包括归一化模块，用于根据同义词表，将包含同义的观点词的多个标签合并成一个标签。

15