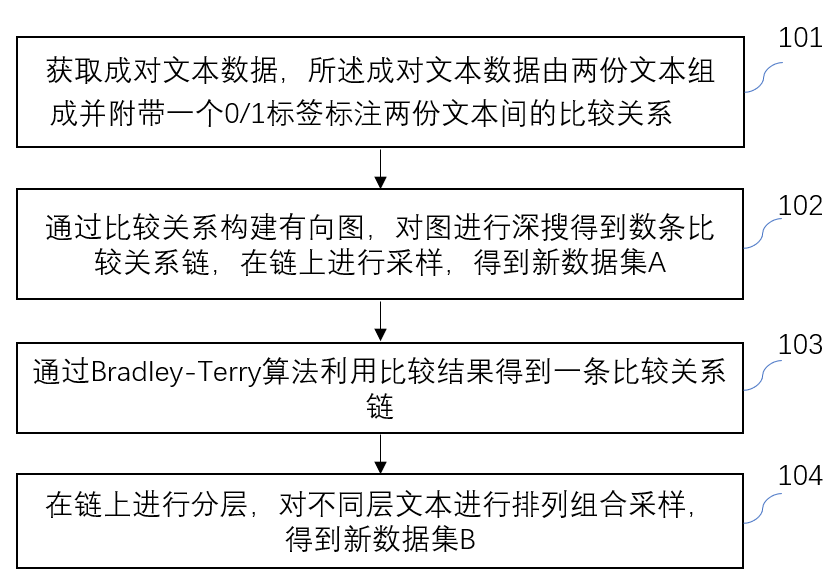
**说 明 书 摘 要**

本申请提供了一种对成对文本的数据增强方法、装置，包括：获取成对文本数据，所述成对文本数据由两份文本组成并附带一个0/1标签标注两份文本间的比较关系；通过比较关系构建有向图，对图进行深搜得到数条比较关系链，在链上进行采样，得到新数据集A；通过Bradley-Terry算法利用比较结果得到一条比较关系链，在链上根据分层和排列思想采样得到新数据集B， 所述成对文本标签是多次比较的多数结果。本方法提供了一套完整的数据增强方法流程，使用Bradley-Terry算法能利用成对文本标注时的进行的多次标注情况，即多次标注的一致性信息；在原始采样方法的基础上，结合分层策略和排列组合思想，缓解了原方法中存在的采样结果受随机性影响较大和采样不公平情况。

**摘 要 附 图**



**权 利 要 求 书**

1. 一种对成对文本的数据增强方法，其特征在于，包括：

取成对文本数据，所述成对文本数据由两份文本组成并附带一个0/1标签标注两份文本间的比较关系；

通过比较关系构建有向图，对图进行深搜后得到数条比较关系链，在链上进行采样，得到新数据集A；

通过Bradley-Terry算法利用比较结果得到一条比较关系链， 所述成对文本标签是多次比较的多数结果；

在链上进行分层，对不同层文本进行排列组合采样，得到新数据集B

2.如权力要求2所述的0/1标签用于文本间的比较关系，比较内容可以是文本积极、悲伤、冒犯程度等。成对标注主要适用于情感极性标注，标注人员难以直接给出具体情感极性程度分数，所述方法可以扩展到任何相似场景下。

3. 如权力要求1所述的采样，指的是让文本按在链中显示的大小关系，从小到大排列组合获得文本对，标签默认为1表示采样得到文本对中左边文本值小于右边。

4. 如权力要求1所述的标签，标签结果并非单次标注结果，为了保证标注结果一致性，标注次数取n次（n为奇数且大于1）。最终结果取多次标注中多次结果

5. 如权力要求1所述Bradley-Terry 模型是一个体育比赛的统计模型，用几个参赛队（或运动员）两两竞技的胜负场次来估计每个参赛队的实力。本专利使用该方法来利用进行标注时的多次标注信息来获得比较关系链。

6. 如权力要求1所述利用Bradley-Terry的得到的链过长，采样结果高达。且，为了尽可能使得每个句子都被采样到，我们可以对每个句子按比较关系链往后采样数次。

7．如权力要求1所述分级方法在链上分层，是对链上结果从小到大均等分层，每层包含相邻的文本，在分层基础上进行排列组合采样。Bradley–Terry是一个近似方法，容易在相邻比较关系中产生错误标注。相邻比较关系分到同一层内，互相之间不进行采样，能减少错误标注产生的噪声。

8. 如权力要求6所述，按比较关系链往后采样数次时，采样次数k，某文本在链后有n条文本，当k<<n时，采样结果受随机种子影响较大。通过进行分层，采样只在不同层的文本间进行，每次采样范围缩小到了区间大小，缓解了上述的采样受随机种子影响的情况。

9. 如权力要求7所述在分层基础上进行排列组合采样，有m层，排名为m的层中文本（即最低的区间）采样m-1次，对排名为1的区间不进行采样。即在层上进行排列组合，每个文本在排名更高的层中都采样一次，尽可能让不同层文本等概率在采样结果中出现。

**说 明 书**

**成对文本数据增强方法、装置**

### 技术领域

本申请涉及情感分析技术领域，尤其涉及一种成对文本数据增强方法、装置。

### 背景技术

情感分析是对人们对产品、服务、组织、个人、问题、事件、话题及其属性的观点、情感、情绪、评价和态度的计算研究。

该领域的开始和快速发展离不开社交媒体的高速发展，论坛、博客、微博等社交平台的快速增长，为此带来了海量的数字形式的数据。

早在2000年，情感分析就成为NLP中最活跃的研究领域之一。它在数据挖掘、Web挖掘、文本挖掘和信息检索方面得到了广泛的研究。现有的许多研究已经被应用到情绪分析的各种任务中产生了，包括有监督和无监督的方法。近年来，将深度学习应用于情绪分析也变得非常流行。

在有监督情况下，想要进行情绪分析，最基本的条件就是一份高质量的标注数据集。标注人员一般能较以较高质量完成一些粗粒度的标注，如一句话表达情绪是否积极，但是当标注人员被要求进行更细粒度的实值分数标注（如积极程度分数），这个任务变得相当困难。当被问及文本的实值情绪分数时，相比于简单地将情绪分为正面或负面，给出具体分数让受访者面临更高的认知负荷。此外，不同注释者很难保持一致，每个人都有自己的评判标准，相同的文本在不同身份的注释者眼中的情绪程度可能有较大差异。

通过向注释者提供成对的文本并询问哪一个更积极，可以克服这些问题，这缓解了标准评分量表中存在的几种注释偏差，如量表区域偏差，并提高了注释的一致性。

然而，现在面临的情况从注释n个实例变为需要注释个实例对，这可能是不现实的。出于人工成本考虑，实际标注的数量要远远少于 。事实上也不需要完全标注如此多句子对。举一个简单的例子，我们标注了A>B,B>C,显然我们可以知道一个隐藏的句子对关系就是A>C。我们希望能根据标注结果挖掘出更多这样的隐藏信息作为数据集的补充，帮助有监督情况下的情感分析。

### 发明内容

有鉴于此，本申请的目的在于提出一种成对文本数据增强方法、装置，用以解决或部分解决上述技术问题。

本文以数据增强为目的，提出了一种基于Bradley–Terry方法结果进行采样的数据增强方法，并在原始方法上利用分层和排列的思想，改善了原始方法采样结果容易受随机种子影响和不同句子采样不公平的缺陷。

与现有的方案相比，本发明所获得的技术效果：

1.根据比较数据集，使用了Bradley–Terry方法的结果来进行采样策略来挖掘更多数据信息。

2.在原始采样方法的基础上，结合分层策略和排列组合思想，缓解了原方法中存在的采样结果受随机性影响较大和采样不公平情况。

### 附图说明

此处所说明的附图用来提供对本发明的进一步理解，构成本申请的一部分，本发明的示意性实施例及其说明用于解释本发明，并不构成对本发明的不当限定。在附图中：

图1为本申请实施例的漏洞检测方法的流程示意图；

### 具体实施方式

为使本申请的目的、技术方案和优点更加清楚明白，以下结合具体实施例，并参照附图，对本申请进一步详细说明。

如图1所示，为本发明实例提供的一种一种成对文本数据增强方法流程图，该方法包括：

步骤101，获取成对文本数据，所述成对文本数据由两份文本组成并附带一个0/1标签标注两份文本间的比较关系；

步骤102，通过比较关系构建有向图，对图进行深搜得到数条比较关系链，在链上进行采样，得到新数据集A；

步骤103，通过Bradley-Terry算法利用比较结果得到一条比较关系链；

步骤104，在链上进行分层，对不同层文本进行排列组合采样，得到新数据集B；

对于步骤101来说，数据为成对句子对用(A,B)表示，根据比较关系可以来进行数据增强。即已知(A,B)代表A>B,又知道(B,C)代表B>C，我们可以得出一个新的句子对关系(A,C)。如果将单个句子作为点，比较关系(A,B)构建一条从B到A的有向边，由此定义可由数据集构建一个大的有向图，代表了句子之间的比较关系。

在该有向图中可能存在一个或多个弱连通子图，子图是一个弱连通图（如果一个有向图将所有的有向边替换为无向边后是连通图，则有向图是弱连通图），子图之间没有边连接。子图如图2所示：

对于步骤102来说，对于每个子图而言，想要根据比较关系进行采样，我们要从每个入度为0的节点出发，如图2中A节点，沿着有向边深搜整个图，可以得到数个比较关系链。如图3所示。对比较关系链去重后，可以按分数从低到高进行排列组合采样。如D->C->B->A，可以采样得到(C,D)、(B,D)、(A,D)、(B,C)、(A,C)、(A,B)。

对于步骤103而言，在数据标注时为了减少单个注释者可能的标注错误，一般会选择两个及以上的注释者，当注释者之间的标注结果存在冲突时，根据少数服从多数原则，选择多数结果。步骤101、102中的比较关系采样方法，只使用了最终的标注结果，而没利用标注中可能存在的冲突信息，事实上我们可以认为当注释者之间标注存在冲突时，往往意味着文本对之间文本分数比较接近，而Bradley-Terry算法则能使用到这部分信息。

己知有n条文本以及其中任意两个文本组成的文本对的标注结果，通过这些数据就可以给出每个文本的强度值（在具体场景下可能时冒犯程度、消极程度、正面程度等）。假设每个文本的强度为 ，标注结果中文本高于的次数为 ，便有对数最大似然:

关于上式的求解问题有着一个非常常见的迭代算法，令是一个随机选择的初始解，那么就有：

用归一化使。在该模型条件下，极大似然估计的存在性和唯一性。

对于步骤104而言，特别说明Bradley–Terry是一个近似方法，与步骤101、102直接构建比较链不同，该方法并不能得到一定正确的比较关系链。相当于做了错误标注为数据带来了噪声，对于一条比较关系链a->b->c->d->e，将b->c标注错误为c->b的情况要比a->e错标为e->a的情况更有可能出现。简而言之，噪声发生在相邻的比较关系之间概率要远大于在间隔的比较关系之间。同时步骤103在具体实施中存在两个问题。

首先是采样结果受随机种子影响较大。如步骤103比较关系链往后采样数次时，采样次数k，某文本在链后有n条文本，当k<<n时，采样结果受随机种子影响较大。

其次是采样不公平，沿图3中链进行采样，作为比较关系中分数较小的文本A，与一个分值更大的文本B，构成（A,B）对。显然，一个排名高的文本，被采样的次数的期望要高于一个排名低的文本，因为该文本至少被采样k次，同时有可能被排名低的文本采样到。

具体来说，对于步骤104，将原来的比较关系划分为数个连续区间，对于区间内的文本互相之间不进行采样，减少上述的可能出现的Bradley-Terry方法带来的噪声影响，采样只在不同区间的文本间进行，如图4所示。图中的区间1文本对区间2内文本进行采样，此时每次采样范围缩小到了区间大小，缓解了上述的采样受随机种子影响的情况。其次对于排名低的区间增大采样次数，排名高的区间减少或者不采样，我们采用排列思想，对总数为m的区间，排名为m的区间（即最低的区间）采样m-1次，依次类推，对排名为1的区间不进行采样，也能有效缓解采样不公平的问题。

**说 明 书 附 图**

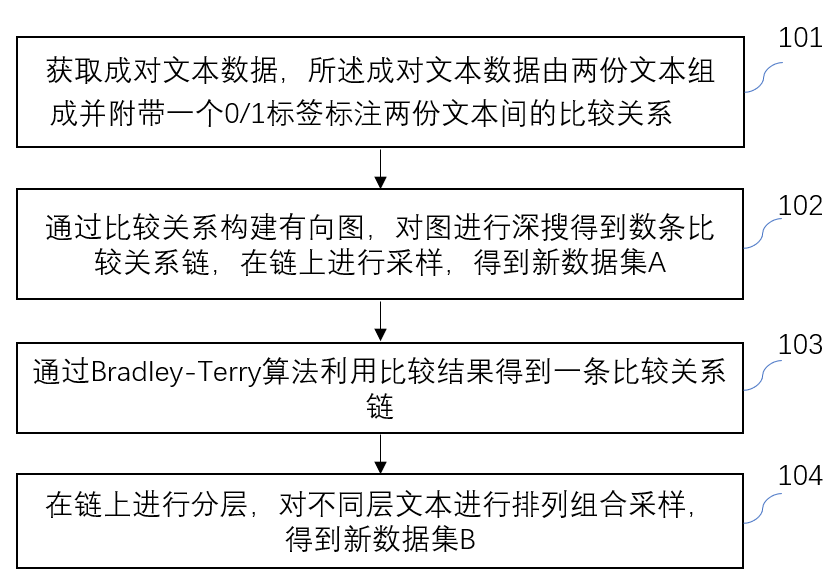


图1

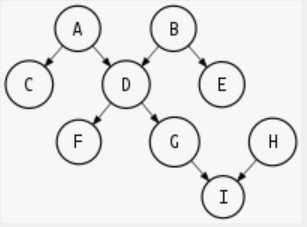


图2

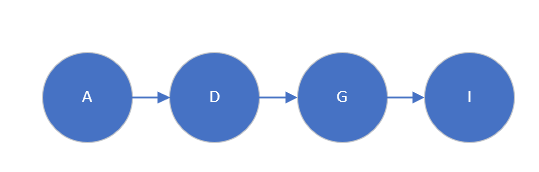


图3

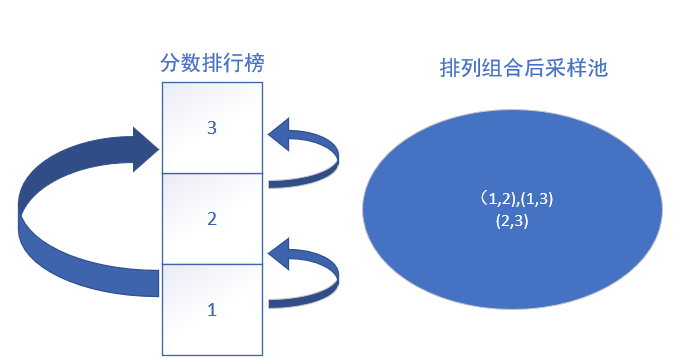


图4