

一种为文本添加超级链接的方法, 装置以及计算机存储介质

本申请要求于 2012 年 3 月 29 日提交中国专利局、申请号为 201210087642.0、发明名称为“一种为文本添加超级链接的方法和装置”的中国专利申请的优先权, 其全部内容通过引用结合在本申请中。

5 技术领域

本发明涉及文本处理技术, 特别涉及一种为文本添加超级链接的方法和装置。

发明背景

超级链接, 是指互联网中指向某个特定目标的连接, 点击超级链接
10 后, 将自动跳转到指定目标。

通过在文本中添加超级链接, 可将不同的文本联系在一起。如图 1 所示, 图 1 为现有添加了超级链接的文本示意图。其中的“太和殿”、“中和殿”、“保和殿”等词即添加了超级链接。在实际应用中, 通常将添加了超级链接的词称为超链词, 用带下划线的蓝色字体表示。

15 现有技术中, 通常采用以下方式来为文本添加超级链接:

1) 针对待添加超级链接的文本, 人工确定对其中的哪些词添加超级链接;

2) 预先生成一个超链词列表, 针对待添加超级链接的文本, 通过与超链词列表进行匹配, 为在超链词列表中出现的词添加超级链接。

20 但是, 上述两种方式在实际应用中均会存在的问题:

对于方式 1), 由于需要人工进行操作, 因此实现起来很不方便, 尤其是当需要对大规模文本添加超级链接时;

对于方式 2), 虽然实现起来比较方便, 但只是简单地为匹配上的词添加超级链接, 可能导致所添加的超级链接与文本的相关性较低。

发明内容

有鉴于此, 本发明提供了一种为文本添加超级链接的方法以及一种
5 为文本添加超级链接的装置, 能够提高所添加的超级链接与文本的相关性, 且实现起来简单方便。

为达到上述目的, 本发明的技术方案是这样实现的:

一种为文本添加超级链接的方法, 包括:

预先生成一个超链词列表, 并收集各种文本, 通过对各文本进行切
10 词处理, 生成一个特征词列表, 针对每个特征词, 分别确定其与每个超链词的共现频率;

针对每个待添加超级链接的文本 X, 分别进行如下处理:

对文本 X 进行切词处理;

从切词结果中提取出在超链词列表中出现的超链词以及在特征词
15 列表中出现的特征词, 并确定每个提取出的超链词和每个提取出的特征词的权值;

根据每个提取出的特征词与每个提取出的超链词的共现频率以及每个确定出的权值, 分别得到每个提取出的超链词的最终权值;

按照最终权值由大到小的顺序对提取出的各超链词进行排序, 为排
20 序后处于前 K 位的超链词添加超级链接, K 为正整数。

一种为文本添加超级链接的装置, 包括:

预处理模块, 用于预先生成一个超链词列表, 并收集各种文本, 通过对各文本进行切词处理, 生成一个特征词列表, 针对每个特征词, 分别确定其与每个超链词的共现频率;

添加模块，用于针对每个待添加超级链接的文本 X，分别进行如下处理：

对文本 X 进行切词处理；

从切词结果中提取出在超链词列表中出现的超链词以及在特征词
5 列表中的出现的特征词，并确定每个提取出的超链词和每个提取出的特征词的权值；

根据每个提取出的特征词与每个提取出的超链词的共现频率以及每个确定出的权值，分别得到每个提取出的超链词的最终权值；

按照最终权值由大到小的顺序对提取出的各超链词进行排序，为排
10 序后处于前 K 位的超链词添加超级链接，K 为正整数。

可见，采用本发明所述方案，通过统计收集到的文本中的词与词之间的共现关系，得到词与词之间的相关性，进而根据从待添加超级链接的文本中提取出的各超链词与提取出的各特征词的共现频率等得到提取出的各超链词的最终权重，并为最终权值较大的超链词添加超级链
15 接，从而提高了所添加的超级链接与文本的相关性；而且，采用本发明所述方案后，可自动确定出为哪些词添加超级链接，无需人工操作，实现起来简单方便。

附图简要说明

图 1 为现有添加了超级链接的文本示意图。

20 图 2 为本发明为文本添加超级链接的方法实施例的流程图。

图 3 为本发明为文本添加超级链接的装置实施例的组成结构示意图。

实施本发明的方式

针对现有技术中存在的问题，本发明中提出一种为文本添加超级链接的方案，能够提高所添加的超级链接与文本的相关性，且实现起来简单方便。

- 5 为使本发明的技术方案更加清楚、明白，以下参照附图并举实施例，对本发明所述方案作进一步地详细说明。

图 2 为本发明为文本添加超级链接的方法实施例的流程图。如图 2 所示，包括：

- 步骤 21：预先生成一个超链词列表，并收集各种文本，通过对各文
10 本进行切词处理，生成一个特征词列表，针对每个特征词，分别确定其与每个超链词的共现频率。

本步骤中，首先生成一个超链词列表，其中具体包括哪些超链词可根据实际需要而定，该列表可由人工编辑生成，也可由机器通过某种方式自动生成，如何生成不作限制。

- 15 之后，可从互联网中收集各种文本，用来生成特征词列表以及统计词与词之间的共现关系。如何收集文本为现有技术，理论上来说，收集的文本数越多越好，具体数目可根据实际需要而定。

上述生成特征词列表以及统计词与词之间的共现关系的具体过程可包括：

- 20 1) 针对收集到的每个文本，分别对其进行切词处理；
 2) 将切词得到的所有不重复词均作为特征词，组成一个特征词列表，

或者，为减少后续的处理工作量，也可先从切词得到的所有不重复词中去除高频词、停用词和低频词等，将剩余词作为特征词，组成一个特征词列表；

如何进行切词以及如何区分哪些词为高频词、停用词和低频词均为
5 现有技术；

另外，在得到特征词列表之后，还需要针对每个特征词，分别确定其逆文本频率（IDF，Inverse Document Frequency）值，IDF值通过用收集到的所有文本的个数除以出现了该特征词的文本个数，再将得到的商取对数得到；

10 3) 针对每个特征词，分别确定其与每个超链词的共现频率：

针对每个特征词 y 和每个超链词 x ，分别计算两者的共现频率 $P(x|y)$ ：

$$P(x|y) = xy \text{ 共现次数} / y \text{ 出现次数}; \quad (1)$$

其中， xy 共现次数表示在收集到的所有文本中，同时出现了特征词 y 和超链词 x 的文本的个数， y 出现次数表示在收集到的所有文本中，

15 出现了特征词 y 的文本的个数；

或者，

针对每个特征词 y 和每个超链词 x ，分别计算两者的共现频率 $P(x|y)$ ：

$$P(x/y) = H(x,y) / I(x,y) = H(x,y) / (H(x) + H(y) - H(x,y)); \quad (2)$$

其中， H 表示信息熵， I 表示互信息， H 和 I 的具体计算方式均为本
20 领域公知；

在实际应用中，可根据实际需要选择使用以上两种方式中的一种。

步骤 22: 针对每个待添加超级链接的文本 X, 分别按照步骤 23 ~ 26 所示过程进行处理。

为便于表述, 用文本 X 来代表任意一个需要添加超级链接的文本。

步骤 23: 对文本 X 进行切词处理。

- 5 步骤 24: 从切词结果中提取出在超链词列表中出现的超链词以及在特征词列表中出现的特征词, 并确定每个提取出的超链词和每个提取出的特征词的权值。

10 将切词结果与步骤 21 中生成的超链词列表以及特征词列表进行匹配, 从切词结果中提取出在超链词列表中出现的超链词以及在特征词列表中出现的特征词。

并且, 针对每个提取出的超链词 H, 分别计算其权值 WH:

$$WH = TFH * IDFH; \quad (3)$$

其中, TFH 表示超链词 H 的词频 (TF, Term Frequency) 值, 即超链词 H 在文本 X 中的出现次数, IDFH 表示超链词 H 的 IDF 值;

- 15 针对每个提取出的特征词 F, 分别计算其权值 WF:

$$WF = TFF * IDFF; \quad (4)$$

其中, TFF 表示特征词 F 的 TF 值, IDFF 表示特征词 F 的 IDF 值。

各 IDF 值已在步骤 21 中计算得到。

步骤 25: 根据每个提取出的特征词与每个提取出的超链词的共现频

率以及每个确定出的权值，分别得到每个提取出的超链词的最终权值。

本步骤中，针对每个提取出的超链词 H，分别计算其最终权值 W_H' ：

$$W_H' = W_H * \sum_{i=1}^n P(H/F_i) * W_{F_i}; \quad (5)$$

其中，n 表示提取出的特征词的个数。

5 $P(H/F_i)$ 值已在步骤 21 中计算得到。

步骤 26：按照最终权值由大到小的顺序对提取出的各超链词进行排序，为排序后处于前 K 位的超链词添加超级链接，K 为正整数。

K 的具体取值可根据实际需要而定。另外，如何为超链词添加超级链接为现有技术。

10 至此，即完成了关于本发明方法实施例的介绍。

基于上述介绍，图 3 为本发明为文本添加超级链接的装置实施例的组成结构示意图。如图 3 所示，包括：

预处理模块，用于预先生成一个超链词列表，并收集各种文本，通过对各文本进行切词处理，生成一个特征词列表，针对每个特征词，
15 分别确定其与每个超链词的共现频率；

添加模块，用于针对每个待添加超级链接的文本 X，分别进行如下处理：

对文本 X 进行切词处理；

从切词结果中提取出在超链词列表中出现的超链词以及在特征词列

表中出现的特征词，并确定每个提取出的超链词和每个提取出的特征词的权值；

根据每个提取出的特征词与每个提取出的超链词的共现频率以及每个确定出的权值，分别得到每个提取出的超链词的最终权值；

- 5 按照最终权值由大到小的顺序对提取出的各超链词进行排序，为排序后处于前 K 位的超链词添加超级链接，K 为正整数。

其中，预处理模块中可具体包括：

第一处理单元，用于生成一个超链词列表；

- 10 第二处理单元，用于收集各种文本，通过对各文本进行切词处理，生成一个特征词列表，针对每个特征词，分别确定其与每个超链词的共现频率。

第二处理单元中又可具体包括（为简化附图，未图示）：

第一处理子单元，用于收集各种文本；

- 15 第二处理子单元，用于对各文本进行切词处理，将切词得到的所有不重复词均作为特征词，组成一个特征词列表，或者，从切词得到的所有不重复词中去除高频词、停用词和低频词，将剩余词作为特征词，组成一个特征词列表；

并且，针对每个特征词 y 和每个超链词 x，分别计算两者的共现频率 $P(x|y)$ ：

- 20
$$P(x|y) = \text{xy 共现次数} / y \text{ 出现次数}; \quad (1)$$

其中, xy 共现次数表示在收集到的所有文本中, 同时出现了特征词 y 和超链词 x 的文本的个数, y 出现次数表示在收集到的所有文本中, 出现了特征词 y 的文本的个数;

或者, 针对每个特征词 y 和每个超链词 x , 分别计算两者的共现频率 $P(x|y)$:

$$P(x/y) = H(x,y) / I(x,y); \quad (2)$$

其中, H 表示信息熵, I 表示互信息。

添加模块中可具体包括:

第三处理单元, 用于对文本 X 进行切词处理;

10 第四处理单元, 用于从切词结果中提取出在超链词列表中出现的超链词以及在特征词列表中出现的特征词, 并确定每个提取出的超链词和每个提取出的特征词的权值; 根据每个提取出的特征词与每个提取出的超链词的共现频率以及每个确定出的权值, 分别得到每个提取出的超链词的最终权值;

15 第五处理单元, 用于按照最终权值由大到小的顺序对提取出的各超链词进行排序, 为排序后处于前 K 位的超链词添加超级链接, K 为正整数。

另外, 第二处理子单元还可进一步用于, 针对每个特征词, 分别确定其 IDF 值, IDF 值通过用收集到的所有文本的个数除以出现了该特征词 20 词的文本个数, 再将得到的商取对数得到;

第四处理单元中又可进一步包括 (为简化附图, 未图示):

第三处理子单元，用于从切词结果中提取出在超链词列表中出现的超链词以及在特征词列表中出现的特征词；

并针对每个提取出的超链词 H，分别计算其权值 WH:

$$WH = TFH * IDFH; \quad (3)$$

- 5 其中，TFH 表示超链词 H 的 TF 值，即超链词 H 在文本 X 中的出现次数，IDFH 表示超链词 H 的 IDF 值；

针对每个提取出的特征词 F，分别计算其权值 WF:

$$WF = TFF * IDFF; \quad (4)$$

其中，TFF 表示特征词 F 的 TF 值，IDFF 表示特征词 F 的 IDF 值；

- 10 第四处理子单元，用于针对每个提取出的超链词 H，分别计算其最终权值 WH':

$$W_H' = W_H * \sum_{i=1}^n P(H/F_i) * W_{Fi}; \quad (5)$$

其中，n 表示提取出的特征词的个数。

- 图 3 所示装置实施例的具体工作流程请参照图 2 所示方法实施例中的相应说明，此处不再赘述。

- 15 本发明实施例还提供了一种机器可读的存储介质，存储用于使一机器执行如本文所述的为文本添加超级链接的方法的指令。具体地，可以提供配有存储介质的系统或者装置，在该存储介质上存储着实现上述实施例中任一实施例的功能的软件程序代码，且使该系统或者装置的计算机
- 20 机（或 CPU 或 MPU）读出并执行存储在存储介质中的程序代码。

在这种情况下，从存储介质读取的程序代码本身可实现上述实施例中任何一项实施例的功能，因此程序代码和存储程序代码的存储介质构成了本发明的一部分。

用于提供程序代码的存储介质实施例包括软盘、硬盘、磁光盘、光盘（如 CD-ROM、CD-R、CD-RW、DVD-ROM、DVD-RAM、DVD-RW、DVD+RW）、磁带、非易失性存储卡和 ROM。可选择地，可以由通信网络从服务器计算机上下载程序代码。

此外，应该清楚的是，不仅可以通过执行计算机所读出的程序代码，而且可以通过基于程序代码的指令使计算机上操作的操作系统等来完成部分或者全部的实际操作，从而实现上述实施例中任意一项实施例的功能。

此外，可以理解的是，将由存储介质读出的程序代码写到插入计算机内的扩展板中所设置的存储器中或者写到与计算机相连接的扩展单元中设置的存储器中，随后基于程序代码的指令使安装在扩展板或者扩展单元上的 CPU 等来执行部分和全部实际操作，从而实现上述实施例中任一实施例的功能。

以上所述仅为本发明的较佳实施例而已，并不用以限制本发明，凡在本发明的精神和原则之内，所做的任何修改、等同替换、改进等，均应包含在本发明保护的范围之内。

权利要求书

1、一种为文本添加超级链接的方法，其特征在于，包括：

预先生成一个超链词列表，并收集各种文本，通过对各文本进行切词处理，生成一个特征词列表，针对每个特征词，分别确定其与每个超链词的共现频率；

针对每个待添加超级链接的文本 X，分别进行如下处理：

对文本 X 进行切词处理；

从切词结果中提取出在超链词列表中出现的超链词以及在特征词列表出现的特征词，并确定每个提取出的超链词和每个提取出的特征词的权值；

根据每个提取出的特征词与每个提取出的超链词的共现频率以及每个确定出的权值，分别得到每个提取出的超链词的最终权值；

按照最终权值由大到小的顺序对提取出的各超链词进行排序，为排序后处于前 K 位的超链词添加超级链接，K 为正整数。

2、根据权利要求 1 所述的方法，其特征在于，所述通过对各文本进行切词处理，生成一个特征词列表包括：

将切词得到的所有不重复词均作为特征词；

或者，从切词得到的所有不重复词中去除高频词、停用词和低频词，将剩余词作为特征词。

3、根据权利要求 1 所述的方法，其特征在于，所述针对每个特征词，分别确定其与每个超链词的共现频率包括：

针对每个特征词 y 和每个超链词 x，分别计算两者的共现频率 $P(x|y)$ ：

$P(x|y) = xy \text{ 共现次数} / y \text{ 出现次数}$ ；

其中，xy 共现次数表示在收集到的所有文本中，同时出现了特征词

y 和超链词 x 的文本的个数, y 出现次数表示在收集到的所有文本中, 出现了特征词 y 的文本的个数;

或者,

针对每个特征词 y 和每个超链词 x, 分别计算两者的共现频率 $P(x|y)$:

$$5 \quad P(x/y) = H(x,y) / I(x,y);$$

其中, H 表示信息熵, I 表示互信息。

4、根据权利要求 3 所述的方法, 其特征不在于,

所述生成一个特征词列表之后, 进一步包括: 针对每个特征词, 分别确定其逆文本频率 (IDF, Inverse Document Frequency) 值, IDF 值通过用收集到的所有文本的个数除以出现了该特征词的文本个数, 再将得到的商取对数得到;

所述确定每个提取出的超链词和每个提取出的特征词的权值包括:

针对每个提取出的超链词 H, 分别计算其权值 W_H :

$$W_H = TF_H * IDF_H;$$

15 其中, TF_H 表示超链词 H 的词频 (TF, Term Frequency) 值, 即超链词 H 在文本 X 中的出现次数, IDF_H 表示超链词 H 的 IDF 值;

针对每个提取出的特征词 F, 分别计算其权值 W_F :

$$W_F = TF_F * IDF_F;$$

其中, TF_F 表示特征词 F 的 TF 值, IDF_F 表示特征词 F 的 IDF 值。

20 5、根据权利要求 4 所述的方法, 其特征不在于, 所述根据每个提取出的特征词与每个提取出的超链词的共现频率以及每个确定出的权值, 分别得到每个提取出的超链词的最终权值包括:

针对每个提取出的超链词 H, 分别计算其最终权值 W_H' :

$$W_H' = W_H * \sum_{i=1}^n P(H/F_i) * W_{F_i};$$

其中，n表示提取出的特征词的个数。

6、一种为文本添加超级链接的装置，其特征在于，包括：

预处理模块，用于预先生成一个超链词列表，并收集各种文本，通过对各文本进行切词处理，生成一个特征词列表，针对每个特征词，分别确定其与每个超链词的共现频率；

添加模块，用于针对每个待添加超级链接的文本 X，分别进行如下处理：

对文本 X 进行切词处理；

从切词结果中提取出在超链词列表中出现的超链词以及在特征词列表
10 表中出现的特征词，并确定每个提取出的超链词和每个提取出的特征词的权值；

根据每个提取出的特征词与每个提取出的超链词的共现频率以及每个确定出的权值，分别得到每个提取出的超链词的最终权值；

按照最终权值由大到小的顺序对提取出的各超链词进行排序，为排
15 序后处于前 K 位的超链词添加超级链接，K 为正整数。

7、根据权利要求 6 所述的装置，其特征在于，所述预处理模块中包括：

第一处理单元，用于生成一个超链词列表；

第二处理单元，用于收集各种文本，通过对各文本进行切词处理，
20 生成一个特征词列表，针对每个特征词，分别确定其与每个超链词的共现频率。

8、根据权利要求 7 所述的装置，其特征在于，所述第二处理单元中包括：

第一处理子单元，用于收集各种文本；

25 第二处理子单元，用于对各文本进行切词处理，将切词得到的所有

不重复词均作为特征词，组成一个特征词列表，或者，从切词得到的所有不重复词中去除高频词、停用词和低频词，将剩余词作为特征词，组成一个特征词列表；

并且，针对每个特征词 y 和每个超链词 x ，分别计算两者的共现频率 $P(x|y)$ ： $P(x|y)=xy$ 共现次数 / y 出现次数，其中， xy 共现次数表示在收集到的所有文本中，同时出现了特征词 y 和超链词 x 的文本的个数， y 出现次数表示在收集到的所有文本中，出现了特征词 y 的文本的个数；或者，针对每个特征词 y 和每个超链词 x ，分别计算两者的共现频率 $P(x|y)$ ： $P(x|y)=H(x,y)/I(x,y)$ ，其中， H 表示信息熵， I 表示互信息。

9、根据权利要求 8 所述的装置，其特征在于，所述添加模块中包括：
第三处理单元，用于对文本 X 进行切词处理；

第四处理单元，用于从切词结果中提取出在超链词列表中出现的超链词以及在特征词列表中出现的特征词，并确定每个提取出的超链词和每个提取出的特征词的权值；根据每个提取出的特征词与每个提取出的超链词的共现频率以及每个确定出的权值，分别得到每个提取出的超链词的最终权值；

第五处理单元，用于按照最终权值由大到小的顺序对提取出的各超链词进行排序，为排序后处于前 K 位的超链词添加超级链接， K 为正整数。

10、根据权利要求 9 所述的装置，其特征在于，

所述第二处理子单元进一步用于，针对每个特征词，分别确定其逆文本频率 (IDF, Inverse Document Frequency) 值，IDF 值通过用收集到的所有文本的个数除以出现了该特征词的文本个数，再将得到的商取对数得到；

所述第四处理单元中包括：

第三处理子单元，用于从切词结果中提取出在超链词列表中出现的超链词以及在特征词列表中出现的特征词；并针对每个提取出的超链词 H，分别计算其权值 W_H ： $W_H = TF_H * IDF_H$ ，其中， TF_H 表示超链词 H 的词频(TF, Term Frequency)值，即超链词 H 在文本 X 中的出现次数， IDF_H 表示超链词 H 的 IDF 值；针对每个提取出的特征词 F，分别计算其权值 W_F ： $W_F = TF_F * IDF_F$ ，其中， TF_F 表示特征词 F 的 TF 值， IDF_F 表示特征词 F 的 IDF 值；

第四处理子单元，用于针对每个提取出的超链词 H，分别计算其最终权值 W_H' ： $W_H' = W_H * \sum_{i=1}^n P(H/F_i) * W_{Fi}$ ，其中，n 表示提取出的特征词的个数。

11、一种计算机存储介质，其特征在于，其中存储有计算机程序，该计算机程序用于执行所述权利要求 1 至 5 任一项所述的方法。