一种基于 LCS 的微博相似页面检测方法

张宗福

（广东省江门职业技术学院 广东 529090）

摘 要 微博是基于关系的信息分享、传播以及获取的平台，是网络舆情发起的源头、信息传播的重要阵地。微博便捷的转发操作，使得大量相同或相似的微博页面在微博空间内迅速传播。对微博相似页面进行检测，对于减轻用户浏览负担和提高网络舆情分析的效率有着重要的意义。本文针对微博相似页面提出了一种基于 LCS 的微博相似页面检测方法：首先计算可能相似的微博页面文档子集，其次计算其 LCS 并提取可信部分，最终检测出微博相似页面。实验表明，这一方法能准确、高效地检测出微博数据中的相似页面。

关键词 LCS；相似性检测；相似性度量；微博页面

A Method Based on LCS for Detecting Similar Microblog Pages

ZHANG Zong-fu

(*Jiangmen Polytechnic., Jiangmen* 529090, *China*)

Abstract Microblog is a relation-based platform for sharing, spreading and acquiring information, and also the source of internet public opinion and the important battlefield of information transmission. The convenient forwarding operations of microblog result in the rapid spread of plenty of identical or similar microblog pages in the microblog space. Therefore, the detection of similar microblog pages is of great importance to lighten the client's burden of browsing and improve the analytic efficiency of internet public opinion. A method based on LCS is introduced to detect similar microblog page: First is to calculate the files' subset of the possibly similar microblog pages, and the next is to calculate its LCS and extract the reliable parts so as to ultimately detect the similar microblog pages. Experiments show that this method can detect the similar pages from the microblog data accurately and efficiently.

Keywords Longest Common Subsequence; near-duplicate detection; similarity measurement; microblog page

1 引 言

微博是一个基于关系的信息分享、传播以及获取的平台，用户可以通过 WEB、WAP 以及各种客户端组件，针对人或者事件等，以 140 字左右的文字作出评论或者转发，并实现即时分享[1]，因此微博往往是网络舆情的发起源头、信息传播的重要阵地。微博信息已成为了网络舆情浏览与分析的重要数据源。

在使用微博的时候，便捷的转发操作，使得大量相同或相似的微博页面在微博空间内迅速传播。对于微博用户浏览而言，虽然看到的微博信息很多，可是有意义的信息量却很有限。对于舆情浏览与分析而言，在海量的微博信息中，相同或相似的微博页面仅仅具有一定的统计意义[2]。因此对微博相似页面进行检测，对于减轻用户浏览负担和提高网络舆情分析的效率有着重要的意义。

对于微博相似性的检测，到目前为止有许多研究人员提出了很多方法，具有代表意义的方法是“基于内容计算的相似微博双重检测与过滤”，它认为大量的转发微博更多地出现在相近时间里，重复率随着发表时间差距的增大而减小，利用 VSM 模型表达、向量夹角的余弦计算两条微博的相似值，该值高于所设阈值则定义为重复，重复率即为重复微博占段内总微

基金项目：国家自然科学基金项目(项目批准号：61272013)和广东省教育科学“十二五”规划 2012 年度研究项目(项目批准号：2010TJK311)。
作者简介：张宗福，CCF 会员，讲师，研究方向为计算机应用，E-mail：jmptzhang@126.com。

博数的比例，在短时间内微博的重复率较高，随着相隔时间的加大，微博重复率迅速减少到很小值。在此基础上，提出基于内容相似性计算的双重过滤法：首先对抓取的一个时间段内微博进行第一重过滤——分段过滤，再对相近时间发表的微博进行第二重过滤——索引过滤，达到微博文本流整体上的过滤。这样对发表时间相隔较短的微博去重，既能保证准确率，同时极大地减少处理时间，提高可用性。

考虑到微博相似页面大部分是由用户转发所造成的，我们尝试通过计算可能相似的微博页面文档子集，然后在此基础上提取可信部分，最终检测出微博相似页面。本文提出了一种基于 LCS 的微博相似页面检测算法，它可以很好地度量微博页面之间的相似度和包含关系，并获得高的准确度。本文在从新浪微博中随机选取一些数据进行了综合的实验，结果表明，本文提出的方法是可行的有效的。

2 基于 LCS 的相似性度量方法

LCS是 Longest Common Subsequence 的缩写，即最长公共子序列[3]。一个序列，如果是两个或多个已知序列的子序列，且是所有子序列中最长的，则为最长公共子序列[4]。

求解两个序列 A 和 B 的 LCS 有很多的算法，典型的、表现最好的是 Myers 提出的 O(ND)算法。在 Myers 的算法中，它认为求解 LCS 的问题其实就是在编辑图中求解包含最多斜线的路径[5]。如两个序列分别为：A=abcabba，B=cbabac。求解从 A 转变到 B 的最短编辑脚本(SES)是“1D 2D 3Ib 6D 7Ic”，从而可方便地得到 A 和 B 的 LCS 为：caba。

基于 Myers 的求解 LCS 的算法，我们可以提出判断两篇文档相似性的度量方法：假如有两个文档 A 和 B，|LCS|为它们的 LCS 的长度，|SES|为最短编辑脚本的长度。假设 A=abcabba，B=cbabac，A 和 B 的长度分别表示为|A|和|B|，则有：

$$|A|+|B|=2|LCS|+|SES|$$

定义 resemble rate 为：

$$r(A,B)=|LCS| / (|A|+|B|-|LCS|)$$

定义 contain rate 为：

$$c(A,B)=|LCS| / |B|$$

这两参数分别表示了两篇文档 A 和 B 的相似度和包含率。所以，判断两篇文档相似性的标准是：如果 r(A,B)或者 c(A,B)超过了一定阈值[5]。

从上面的分析可知，基于 LCS 的相似性检测的一般过程为：首先，按一定的方法对需要检测的文档进行排序，尽量使形成最大子集的文档排在前面；其次，读取第一个文档，放入第一个子集；最后，按顺序逐步读取其他文档，与已有子集的第一个文档进行比较，如果不相似就生成一个新的子集。

在微博页面的相似性比较中，词的顺序是一个重要的信息，LCS 能很好地体现词的顺序。微博的转发操作容易产生大量的具有包含关系的页面，通过 LCS 进行相似性检测，仍然认为他们是相似的。微博页面中存在很多诸如广告信息的噪音信息，它们主要是存在于页面的头尾部分，通过对不同位置的文字赋予相应的权值，LCS 检测可以将这些信息和页面的正文信息区分出来，从而提高检测的精度。

所以，基于 LCS 的相似性检测算法主要解决两个问题，一是在计算 LCS 之前如何尽可能排除掉那些完全不需要进行判断的文档，只有在两篇文档有可能相似的时候才需要再次进行 LCS 计算；二是在进行检测时，如何排除微博页面的干扰，在 LCS 中提取真正可信的部分来计算文档相似性。

由此可见，使用基于 LCS 的检测算法和在此基础上的相似性度量方法不仅仅可以获得较高的检测准确度，也可以获得比较高的效率。

3 算法与实现

为有效地计算微博页面的相似性，根据上述基于 LCS 的相似性检测的度量方法和需要解决的问题，首先要将微博页面文档集合分割为可能相似文档子集，只在可能相似的文档子集中才计算 LCS；其次计算 LCS 并选取它的可信部分来计算相似度。

3.1 计算可能相似的微博页面文档子集

根据前面提到的方法，首先我们要将微博页面文档的集合分割为可能相似的文档子集，然后对可能相似的文档进行相似性比较。在这里我们可以使用 Myers 提出的 O(ND)算法。其步骤如下：

(1)选取需要进行相似性检测的微博页面，去掉所有的 HTML 标签和页面格式化的信息，并将其页面文档按照标点符号分割为句子的集合。

(2)计算微博页面文档的指纹，其方法是：首先，选择长度大于一定值 X 的句子，然后，对计算其 MD5 摘要，对 k 取模，不同值分为 k 个组；其次，从每个组选取 MD5 值最小的分布在不同位置的 Y 个

不同句子，计算一个 128 位的哈希值，称为指纹，这样共得到 k 个指纹；最后，对 MD5 值循环移位一个字节，重复上述两个步骤，因为 MD5 值为 16 个字节，所以最多可以得到 16×k 个指纹。

对于每个微博页面文档，根据前面的步骤所知，最多可以得到 16×k 个指纹。调节 X、Y、Z 三个参数(Z 为 Y 个不同句子覆盖的文字长度)，当参数值越大时，表明条件越严格，当参数值一定的时候，只要两个微博页面文档存在同一个指纹，可以认为它们是相似页面的概率也很高。所以，只要存在一个指纹相同，我们可以认为它们是相似页面。仅仅通过一个指纹来判断是否相似，可能会遗漏很多可能相似的微博页面，因此我们可以使用多个指纹，比如设 k=5，最多可以有 80 个指纹。由于微博页面比较简单，这样的取值是完全合理的。

根据前面得出的指纹，我们可以建立指纹的索引，将它们聚集到可能相似的子集里。由于微博“转发”操作的缘故，微博页面的相似性存在着明显的传递性，所以我们可以对微博页面文档进行排序，相似性只能从大到小单向传递，然后把他们都聚集到相似子集中[6]。

综上所述，我们对能相似的微博页面文档子集进行聚集计算的时候，先对文档按它可生成的指纹数从大到小排序，指纹越多就越可能找到越多的相似页面文档，然后按顺序从微博页面文档集合中逐个读取，根据指纹索引从剩下的微博页面文档集合中读取和它可能相似的以及根据传递性得到的可能相似页面文档聚集在一起，形成可能相似子集。

3.2　计算 LCS 并提取可信部分

微博页面的噪音信息(例如广告信息等)通常都出现在微博页面正文内容的头部和尾部，两篇不同用户的微博页面，其模板是相同的，而正文内容内部一般是连续的。所以一个微博页面的可信部分有两个特征：一是可信部分正文内容是连续的；二是可信部分在微博页面中是在中心部位的。

计算微博页面文档的相似度需要计算 LCS 并提取出它的可信部分，在这里，我们仍然使用 Myers 的 O(ND)算法。实际上我们在这里计算得到的两篇微博文档的 LCS 和 SES 都不是严格意义的，而是它们的近似值。

接下来重点分析如何提取 LCS 的可信部分，图 1 描述的是如何计算基础区域，图 2 描述的是如何从基础区域中提取可信部分。

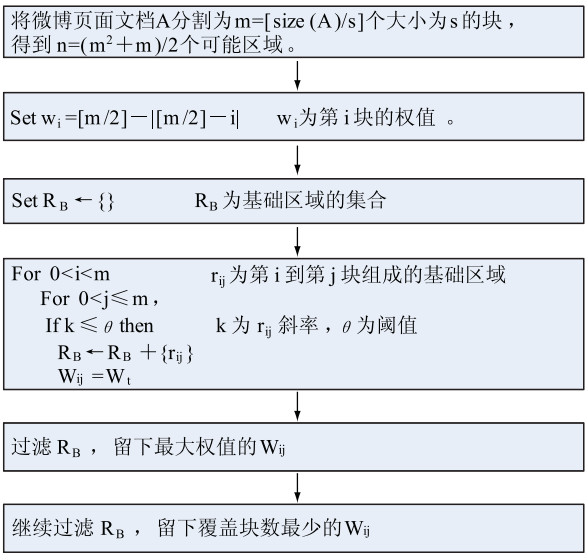

图 1　计算基础区域

算法中通过将微博页面文档分割为 m=[size(A)/s]个大小为 s 的块，得到 n=(m²+m)/2 个可能区域。然后给每个块赋予一定的权值，最后在 R_B 中可得到两个权值最大而覆盖块数最小的基础区域 r_{ij}。

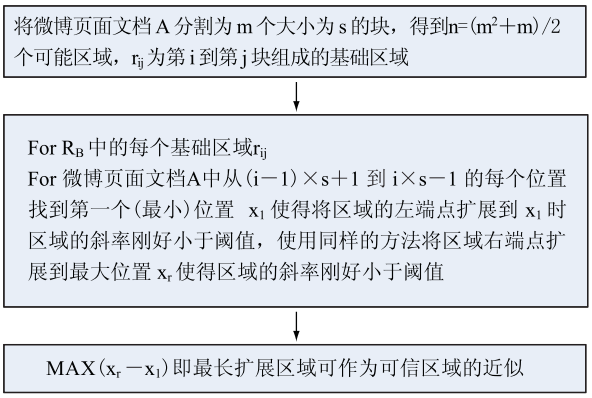

图 2　计算并提取可信区域

从图 2 可以知道，从微博页面文档的基础区域 r_{ij} 中，从 (i−1)× s + 1 到 i × s −1 的每个位置找到第一个(最小)位置 x_1 使得将区域的左端点扩展到 x_1 时区域的斜率刚好小于阈值 θ，使用同样的方法将区域右端点扩展到最大位置 x_r 使得区域的斜率刚好小于阈值 θ。求出 (x_r-x_1) 的最大值(即最长扩展区域)作为可信区域的近似。

4　实　验

微博页面都是由用户创作完成的，一般具有一段连续的正文，所以我们将微博页面的相似关系分

为三种情况：相似、不相似、待定(一般是检测时无法判别的情况)。我们可以通过三个实验来组织：实验 1，评估指纹数量对生成可能相似微博页面文档子集的影响；实验 2，评估可信 LCS 区域斜率阈值的影响；实验 3，评估算法的性能。

4.1 实验 1：评估指纹数量对生成可能相似微博页面文档子集的影响

随机从公共大厅里下载了 1000 条微博页面文档进行评估。首先，对每个选中微博页面文档，我们使用上述的方法对整个集合运行一次相似性检测，通过计算它们的 LCS，并提取可信部分，判断是否有相似，然后我们通过人工检查来判断是否相似。其次，对于每个选中的微博页面文档，我们在生成后的可能相似的子集中，再次使用上述的方法进行检测。

从上述算法中，我们知道，设置指纹数量为不同的值会得到不同的可能相似子集。所以我们把指纹数量从 5 增加到 80 进行实验，以 15 为增加值，共得到 6 组数据。实验结果表明，随着指纹数量的增加，覆盖率不断上升，而可信率不断下降，也就是说可信率越低，每个微博页面需要比较的平均次数也越多，计算代价就越高，所以，可以通过增加计算资源来获得更高的覆盖率。覆盖率和可信率在指纹数为 25 时取得平衡。

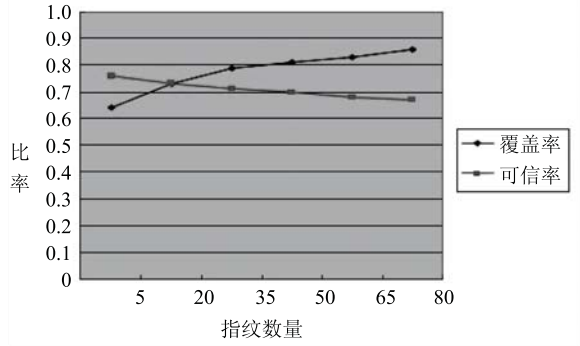

图 3 指纹数量对生成可能相似微博页面文档子集的影响

4.2 实验 2：评估可信 LCS 区域斜率阈值的影响

随机从公共大厅里下载了 1000 个微博页面文档，取得它们对应的可能相似子集来进行评估，对每一个选中微博页面文档，在它的可能相似子集中使用算法去寻找相似微博页面文档，并计算准确率和近似的召回率，最后取平均值。通过改变改变斜率阈值从 0.02 到 0.2，在此过程中调节相似度量参数，使近似的召回率保持在 0.9，然后计算相应的准确率。当斜率阈值取一定值时，准确率能达到一个最高值。

实验表明通过计算 LCS 并提取可信部分在一定程度上提高了准确度。如果不采取计算 LCS 并提取可信部分的方法，则相当于取斜率阈值为无限大值，准确率会有一定的影响，所以采取计算 LCS 并提取可信部分的算法是有一定意义的。

4.3 实验 3：评估算法的性能

微博页面数据属于海量数据，如果运用本算法对整个数据集进行计算，时间效率低下，难于达到实时应用的目的。因此我们采取抓取新浪微博一个时间段内的共 3000 个微博页面进行相似性检测，耗时大概是 120 秒，共检测出 15% 的相似微博页面。实验说明，本文提出的基于 LCS 的微博相似页面检测算法不仅能检测出微博的相似页面，同时它所需要的计算代价是可以接受的。

5 结束语

本文通过分析微博页面数据的特点，针对微博转发操作产生的相似页面提出了一种基于 LCS 的微博相似页面检测方法，首先计算可能相似的微博页面文档子集，其次计算 LCS 并提取可信部分，最终检测出微博相似页面，效果理想，效率较高。实验表明，这一方法能有效地检测出微博数据中相似的页面，对于减轻用户浏览负担和提高网络舆情分析的效率有着重要的意义。

然而，这些工作还需要我们进一步深入与完善，比如随着微博的不断发展，数据量不断地增加，在海量数据中检测相似页面，效率的提高是一个永恒的命题；另外，微博相似页面的检测还有其他的方法，值得我们不断对比与探究，找到最合适的最高效的方法。

参 考 文 献

[1] 王琳, 冯时, 徐伟丽, 等. 一种面向微博客文本流的噪音判别与内容相似性双重检测的过滤方法 [J]. 计算机应用与软件, 2012, 29(8): 25-29.

[2] 周刚, 邹鸿程, 熊小兵, 等. MB-SinglePass: 基于组合相似度的微博话题检测 [J]. 计算机科学, 2012, 39(10): 198-202.

[3] 王映龙, 杨炳儒, 宋泽锋, 等. 基因序列相似程度的 LCS 算法研究 [J]. 计算机工程与应用, 2007, 43(31): 45-47.

[4] 曾波, 潘少彬, 陆璐. 改进的 LCS 方法在测试脚本序列比对中的应用 [J]. 计算机工程与应用, 2011, 47(35): 71-76.

[5] 于海英. 字符串相似度度量中 LCS 和 GST 算法比较 [J]. 电子科技, 2011, 24(3): 101-104.

［6］ 宋鳌, 支睁, 周军, 等. 基于 LCS 的特征树最大相似性匹配网页去噪算法［J］. 电视技术, 2011, 35(13): 44-47.

［7］ 黄连恩. 历史网页的持续收藏及其再访问的关键技术研究［D］. 北京: 北京大学, 2008: 7-9.

［8］ Stein B. Principles of hash-based text retrieval［C］// Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval, 2007: 527-534.

［9］ Fetterly D, Manasse M, Najork M. On the evolution of clusters of near-duplicate web pages［C］// In Proceedings of First Latin American Web Congress, 2003: 37-45.

［10］ Manku G S, Jain A, Sarma A D. Detecting near-duplicates for web crawling［C］// Proceedings of the 16th International Conference on World Wide Web, 2007: 141-149.

［11］ Henzinger M. Finding near-duplicate web pages: A large-scale evaluation of algorithms［C］// Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval, 2006: 284-291.

［12］ Nakatsu N, Kambayashi Y, Yajima S. The string-to-string correction problem［j］. Journal of ACM, 1974, 21(1): 168-173.