doi:10.3772/j.issn.2095-915x.2017.03.014

一种采用 SpotSigs 算法的中文新闻网页相似性检测方法

1. 中国科学技术信息研究所 北京 100038；
2. 万方数据股份有限公司 北京 100038;
3. 古联 (北京) 数字传媒科技有限公司 北京 100049

李岩 [1] 徐硕 [1] 吴广印 [1,2] 干生洪 [3]

摘 要 互联网的高速发展使得新闻网页成为了网民了解国内外大事的首要选择，然而中国互联网存在着大量重复新闻网页的现象，对于提高用户体验以及新闻情报的分析造成了一定的困难。本文以 SpotSigs 算法为基础提出了一种中文新闻网页相似性检测方法，在先行词选取阶段使用基础先行词与优化先行词相结合的选择策略，从而降低了网页中的导航栏、广告等噪音对中文新闻网页相似性检测的影响。以实际的中文新闻网页为实验数据集，通过准确率、召回率两项指标验证了基于 SpotSigs 算法的中文新闻网页相似性检测方法的有效性和可行性，特别在相似度阈值较低的情况下具有较好的性能。

关键词： SpotSigs 算法 , 新闻网页 , 相似性检测 , 先行词选取
中图分类号： G35

A Chinese News Webpages Similarity Detection Approach Using SpotSigs Algorithm

1. Institute of Scientific and Technical Information of China, Beijing 100038, China;
2. Wanfang Data, Beijing 100038, China;
3. Gulian (Beijing) Media Tech Co., Ltd. Beijing 100049, China

LI Yan[1] XU Shuo[1] WU GuangYin[1,2] GAN ShengHong[3]

Abstract With the rapid development of Internet, news webpages become the primary choice for Internet users to learn about what’s happening. However, there are a lot of repetitive Chinese internet news

基金项目： 本文受国家自然科学基金项目“基于论文和专利资源的技术机会发现研究”（71403255），“十二五”国家科技支撑计划项目“面向科技情报分析的信息服务资源开发与支撑技术研究”（2015BAH25F01）的资助。

作者简介： 李岩（1994–），硕士研究生，研究方向：数据挖掘、专家发现，Email：liyan2015@istic.ac.cn；徐硕（1979–），博士，副研究员，研究方向：智能情报分析，数据挖掘和大数据；吴广印（1965–），研究员，总工程师，研究方向：云计算、知识组织、大数据挖掘与分析；干生洪（1982–），图书馆学，馆员，研究方向：知识组织与知识服务。

webpages, thus causing poor user experience and difficulties of data mining on news information. This paper proposed a Chinese news webpages similarity detection approach on the basis of SpotSigs algorithm, which combines basic and optimized antecedents in order to reduce the noise of navigation bar or advertisement. Experimental results on real-world Chinese news webpages indicated that our approach can effectively detect similar Chinese news webpages in terms of precision and recall, especially for the case of low similarity threshold.

Keywords：SpotSigs algorithm, news web pages, similarity detection, antecedents selection

自 1994 年 4 月我国全面接入互联网以来，中国互联网经历了 20 多年的高速发展。截至 2016 年 12 月，我国网民规模达 6.95 亿，互联网普及率达 53.2%[1]。随着网民数量的不断增加以及移动终端设备的快速发展，越来越多的公众把网络新闻作为了解国内外大事的首要选择，网络新闻的作用越来越重要，影响也越来越广泛。网络新闻也成为了政府引导社会舆论、与公众沟通的良好平台，对其研究将会促进舆情监测发展 [2]。

目前我国有多家大型网络新闻运营发布平台，比如新浪、搜狐、腾讯、网易、新华网等。为了提高点击率和增加流量等各种原因，每当有重大事件发生时，各大新闻网站将争相报道 [3]。虽然报道用语稍有差异，但事件内容从字面上看基本一样，鲜有字面表达差别较大但意义表达接近的情形发生。比如：网易新闻中一篇题为《蔡英文对“九二共识”表态》的新闻与环球网《蔡英文首对“九二共识”清晰表态》其字面意思稍有不同，但是其新闻内容事件相同。据不完全统计，中国的重复网页比例高达 25% [4]。

转载以及重复报道对于网络舆情监测、资源共享、扩大传播范围等是有益的，可以从中揭示热点事件，更好地了解民意 [5]。但对于网络新闻聚合器的用户来说，将大大增加新闻阅读的负担 [6]。为了提高对于网络新闻类开源情报的分析，经常需要将网络新闻加工成熟语料，但转载以及重复报道的新闻不利于熟语料的加工。

经仔细分析，典型网络新闻页面一般由 4 部分组成：导航栏、新闻正文、广告栏、相关新闻推荐等，大致布局结构如图 1 所示。

目前有大量相似性检测的算法被提出，甚至被产品化，如论文抄袭检测、镜像页面识别等。但由于导航栏、广告栏以及相关新闻推荐等部分的存在，直接将原有算法用于新闻网页相似性检测的效果通常并不理想 [7]，因此需要专门研发针对新闻网页的相似性检测方法。Theobald 等人 [8] 于 2008 年提出了一种针对英文新闻网页相似性检测的 SpotSigs 算法，考虑了英文中词的位置对于相似性检测的影响，从论文的实验描述来看效果还不错。由于中文文本相较于英文文本没有天然的词语边界，中文先行词的选取则有些麻烦。因此本文尝试以 SpotSig 算法为基础，提出基础先行词与优化先行词相结合的先行词选择策略，开展中文新闻网页相似性检测研究工作，并洞察未来研究的努力方向。

doi:10.3772/j.issn.2095-915x.2017.03.014

本文的后续章节安排如下：第 1 节从相似性检测的三大步骤分别介绍相关研究工作。第 2 节介绍了 SpotSigs 算法的流程，第 3 节讨论了利用 SpotSigs 进行中文新闻网页相似性检测的方法，第 4 节给出算法在实际数据集上的性能评测结果，最后是本文的总结与下一步的工作。

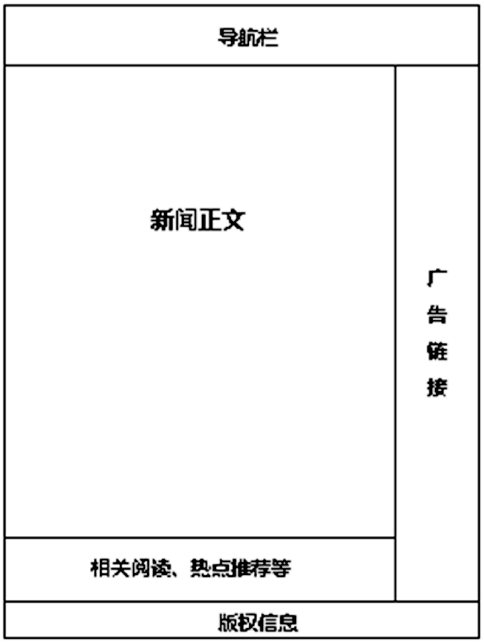

图1 新闻网页的典型结构示意图

1 相关研究

近年来已经有大量的网页相似性检测算法被提出，根据原理不同将其分为两类：一类采用基于字符串的比较方法。Border 等人 [9] 最早提出 “Shingling” 为特征进行相似度检测，将文档使用固定长度的特征序列来表示，基于 “Shingling”、“super-Shingling” 提出 DSC,DSC-SS 算法。随后的 COPS[10]、KOALA[11] 等文本检测系统也都使用了这种方法。2002 年 Chowdhury 等人 [12] 从过滤词语方面入手提出了 I-Match 算法，使用特征词典的方式来提取特征，通过 IDF 技术对网页的噪音内容进行过滤。但由于其不考虑低频词与高频词带来的影响，所以精度不高。针对此问题，文献 [13] 利用随机化的方式对其进行改进。

第二类基于词频统计。1995 年 Shivakumar 和 Garcia-Molina 提取单词在文章中出现的频率为特征，建立了基于向量空间模型的的 SCAM 算法 [14]，随后又提出了 dSCAM 算法 [15]。Gionis A 等人 [16] 提出了采用 LSH 技术，在高维空间快速的进行相似性检测。随后 LSH 的改进算法、Hamming-LSH[17] 和 LSH-Tree[18] 等算法被相继提出。Theobald 等人 [8] 提出了 SpotSigs 算法，通过提取英文中的停用词特征来对网页的噪声内容进行过滤。之后 Mao 等人 [19] 提出 AF-SpotSigs、SizeSpotSigs 两种算法应用于短文本网页的相似性检测。Emre Varol 等人 [20] 提出 CoDet 算法，在检测网页之间的包含关系方面取得较好效果。

由此可见，目前网页相似性检测技术已经有了长远的发展。但是就目前来看，中文网页相似性检测还主要是借鉴已有的英文网页相似性检测方法，将其应用于中文网页 [21]。针对目前中文新闻网页相似性检测问题，就目前查阅资料看，国内在此领域的研究还并不完善 [22]。由于中文文本相较于英文文本其没有天然语言边界，分词可能会破坏语义结构 [23]，因而直接将英文相似度检测算法应用于中文新闻网页的相似性检测可能不能取得很好的效果。

目前关于网页相似性检测算法，关键的不

同在于特征提取与特征比较两个环节。本文通过对相关研究进行全面的调研和总结，认为基于 Spotsigs 算法对中文新闻网页进行相似性检测是可行的，通过具有语义性的特征提取，在一定阈值范围内可以更加准确地检测出相似网页。

2 SpotSigs算法

SpotSigs 算法的流程如图 2 所示，该算法在文档表示阶段使用先行词后一定距离与链长的词语作为 Shingle，利用长度分区筛选候选对，同时建立倒排索引来加快比较的速度。

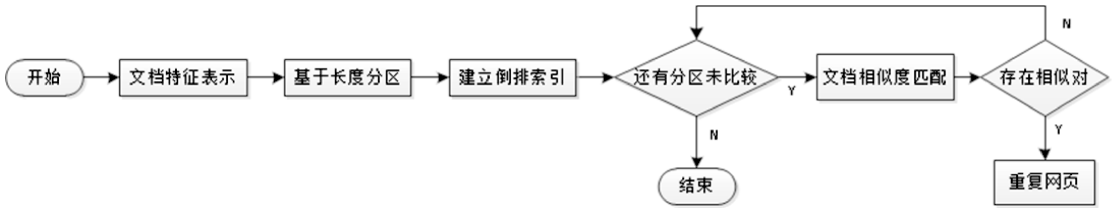

图2 Spotsigs算法流程图

分区中文档相似度匹配部分伪代码如下：

1: Input: document vectors d_i with weighted spot signatures s_{ij}; partitions P with boundaries$[p_k,p_{k+1}]$ and inverted lists $list_{kj}$

2: $pairs \leftarrow \emptyset$

3：for all d_i in random order of $|d_i|$ using t threads in parallel do

4: $partition_k \leftarrow P.get(|d_i|)$

5: sort all $s_{ij} \in d_i$ by asc. document frequency in $partition_k$

6: $\delta_1 \leftarrow 0$

7: $checkedd_i \leftarrow \emptyset$

8: for all $s_{ij} \in d_i$ do

9: $list_{kj} \leftarrow partition_k.get(s_{ij})$

10： $\delta_2 \leftarrow 0$

11: for all $d_{i'} \in list_{kj}$ sorted by descending $|d_{i'}|$ do

12: $\delta_2 \leftarrow |d_i|-|d_{i'}|$

13: if $d_i=d_{i'}$ or $d_{i'} \in checked_i$ then

14: continue

15: else if $\delta_2<0$ and $\delta_1-\delta_2>(1-\tau)|d_{i'}|$ then

16: continue

17: else if $\delta_2 \geqslant 0$ and $\delta_1+\delta_2>(1-\tau)|d_i|$ then

18: break

19: else if $sim(d_i,d_{i'}) \geqslant \tau$ then

20: $pairs \leftarrow pairs \cup \{<d_i,d_{i'}>\}$

21: $checked_i \leftarrow checked_i \cup \{d_{i'}\}$

22: end if

23: end for

24: $\delta_1 \leftarrow \delta_1+freq_{di}(s_{ij})$

25: if $\delta_1 \geqslant (1-\tau)max\{|d_{i'}|\}_{di' \in partitionk-checkedi}$ then

26: break

27: end if

28: end for

doi:10.3772/j.issn.2095-915x.2017.03.014

29: if $p_{k+1}-|d_i| \leqslant (1-\tau)\, p_{k+1}$ then
30: $partition_k \leftarrow P.get(p_{k+1})$
31: goto 6
32: end if
33: end for
34: return *pairs*

2.1 新闻网页特征表示

新闻网页所包含大量噪声不利于相似性检测，有效的特征表示要尽可能多地滤掉噪声部分，保留新闻正文部分的内容。经仔细分析，Theobald 等人发现广告和导航栏等部分很少出现停用词，而为了表达通顺，新闻正文会使用大量的停用词，因此提出将停用词作为先行词，在其后取一定距离与链长的词语作为 Shingle。

具体来说，首先定义一个先行词集合 C，对网页从头开始扫描，每遇到一个先行词 $c \in C$，便从该先行词后面的第一个非先行词开始取相对距离为 d 的非先行词，直到取到规定个数 n，重复这个过程直到结束，得到该网页的特征表示集合，如例 1 所示。需要说明的是，同一个文档集中的所有文档需要取相同的链长 n 和距离 d。

例 1：对于如下一篇文章：The unified loan and return system in Chengdu was rolled out in 2014 and gives local residents free access to all public library resources at all branches.

定义先行词集合 C={and, in, out, was, to, at}，取距离为 1，链长为 2 提取特征，得到特征集合。S={and:return:system,in:Chengdu:rolled, was:rolled:2014,out:2014:gives,in:2014:gives,and: gives:local,to:all:public,at:all:branches}

2.2 基于长度分区的候选对筛选

给定相似度阈值，如果两篇新闻网页特征表示集合的大小之比 $|d_1|/|d_2|<\tau$（假设 $|d_1|\leq|d_2|$），则这两篇新闻网页一定不相似。不难发现，基于长度分区的候选对筛选方法会漏掉那些内容相似但长度差别较大的新闻网页，比如简讯和一般新闻。正是基于这种直观想法，Theobald 等人提出了一种最优分区方法，将所有文档按长度进行分区，使得所有相似新闻网页只出现在同一个或相邻的分区。这种分区方法与文档本身无关，只与相似度阈值有关，如例 2 所示。

例 2：当网页集合中包含最多先行词的网页含 261 个先行词时，τ=0.5，共划分为 7 个区间：[1,3], [4,9], [10,21], [22,45], [46,93], [94,189], [190,+∞]; τ=0.6，共划分为 9 个区间：[1,2], [3,6], [7,12], [13,22], [23,39], [40,67], [68,114], [115,192], [193,+∞]。

3 中文新闻网页相似度检测方法

考虑到中文新闻网页的特殊性，图 3 给出了中文新闻网页相似度检测流程。

3.1 数据噪声

通过对中文新闻网页的深入分析发现（如图 1 所示），中文新闻网页存在两类对结果影响较大的噪声：（a）导航栏、广告和其他新闻链接标题等，这类噪声可以通过特征表示有效去除；（b）极少部分新闻网页中包含相关阅读推荐的长文本，此类文本与新闻正文看起来极像，对中文新闻网页相似性检测的结果影响较大，尤其是对于新闻简讯。为了屏蔽第二类噪声的不利影响，本文在爬取网页时剔除相关阅读层。

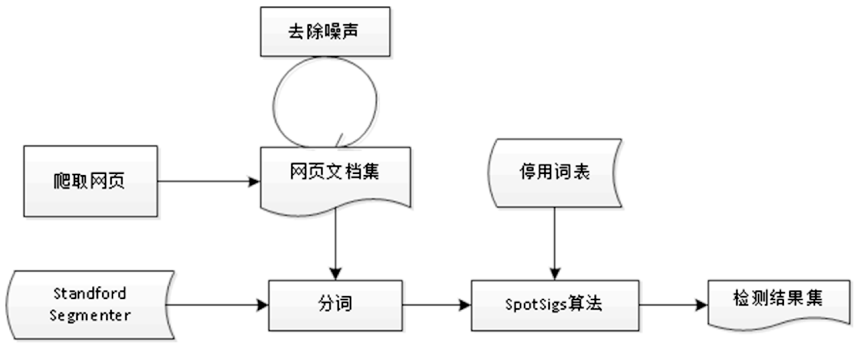

图3 利用SpotSigs进行中文新闻网页相似度检测流程

3.2 中文分词

与英文新闻网页不同，中文新闻网页没有天然的词语边界，因此相似性检测之前还需要对中文新闻网页进行分词处理。本文采用的分词工具为斯坦福大学自然语言处理团队开发的 Stanford segmenter[24]，版本为 3.6.0。

3.3 停用词选择策略

为了提取到足够多的有效特征，需要先确定先行词集合。鉴于其他新闻链接标题仍然含有不少停用词，本文采用以下过程进行先行词的选择：

（1）准备足够的适当网页，手工或定制模板精确提取出网页中的文本层（即忽略了导航栏、其余新闻标题链接等）。

（2）将网页中的词与通用停用词表 [25] 求交集，计算在网页中出现的停用词及其频次。

（3）将网页中的文本层与通用停用词表求交集，计算在纯文本中出现的停用词及其频次。

（4）对于每一个文本层中出现的停用词，计算其频次 t_1 与出现在整个网页中的频次 t_2 的商。选取 $t_1/t_2>n_1$（$n_1>0.6$）且出现频次较高的停用词作为基础先行词，使用这些先行词可在相似度阈值较低的情况下检测出相似对。选取 $t_1/t_2>n_2$（$n_2>0.8$）的词作为“优化”先行词，使用这些先行词提升相似网页之间的相似度。

4 实验结果及讨论

4.1 实验数据及预处理

实验数据来自于搜狐、腾讯、新浪和网易四大门户网站，按政治、体育、军事、财经和生活五个频道随机选取 125 个新闻网页。其中搜狐新闻网页 34 个，腾讯新闻网页 28 个，新浪新闻网页 35 个，网易新闻网页 28 个。经人工标注发现，不与其他网页重复的网页有 49 个，两个为一组的重复网页 16 组，3 个一组的有 12 组，4 个一组的有 2 组。只有新浪新闻中包含推荐新闻部分，本文在相似性检测之前去除了其中的推荐新闻层。

4.2 停用词选择

对数据集的网页进行处理，选取在文本层出现频次大于 900 次且在网页中出现频次占比例大于 0.65 的 4 个停用词“的”、“是”、“在”、“了”作为基础先行词。出现频次以及比例见表 1。

doi:10.3772/j.issn.2095-915x.2017.03.014

表1 选取的基础先行词在文本层、网页中的频次及其比例

基础先行词	在文本层中出现频次 t_1	在网页中出现频次 t_2	t_1/t_2
的	4066	6038	0.67
在	1127	1652	0.68
了	1081	1471	0.73
是	897	1340	0.67

使用基础先行词可以在网页集合中检索出重复网页，但是由于网页中广告、推荐阅读等栏目的存在，其召回率较低，且检出的相似对的 Jaccard 相似度较低。故选取 $t_1/t_2>0.9$ 的停用词作为“优化”先行词，降低噪音影响。一共有如“仍然”、“据”、“此外”等先行词 207 个。最终使用的先行词一共 211 个。

4.3 实验过程及结果

本文令链长 $n = 1$，距离 $d = 1$，将新闻网页表示成 Shingle 集合，网页中最少包含 5 个先行词，最多有 261 个先行词，在时共划分得到 9 个区间，分别为：[1,2], [3,6], [7,12], [13,22], [23,39], [40,67], [68,114], [115,192], [193,+∞]，网页之间相似度采用 Jaccard 多重集合相似度来计算，如例 3 所示。

例 3：实验数据中有搜狐、腾讯、网易、新浪的四个网页《中国农村：“80 后”不会种地“90 后”不提种地》《中国农村现状：老人妇女是主力 不再指望种地致富》《真实农村：老人妇女成种地主力 流转成本大涨》《农业大县种地调查：粮食价格走低 土地流转遇冷》，其内容皆为反映河南延津县农村土地使用现状，其中分别提取出 72、69、63、146 个特征，均出现{是：主力}、{的：司寨乡}、{据：了解}等特征。在阈值 τ=0.6 时，四篇网页位于 [68,114]、[115、192] 两个相邻分区。其 Jaccard 相似度分别为（按前文出现顺序编号，括号内为相似度，均保留了三位小数）1-2（0.903）、1-3（0.887）、1-4（0.747）、2-3（0.955）、2-4（0.773）、3-4（0.806）。

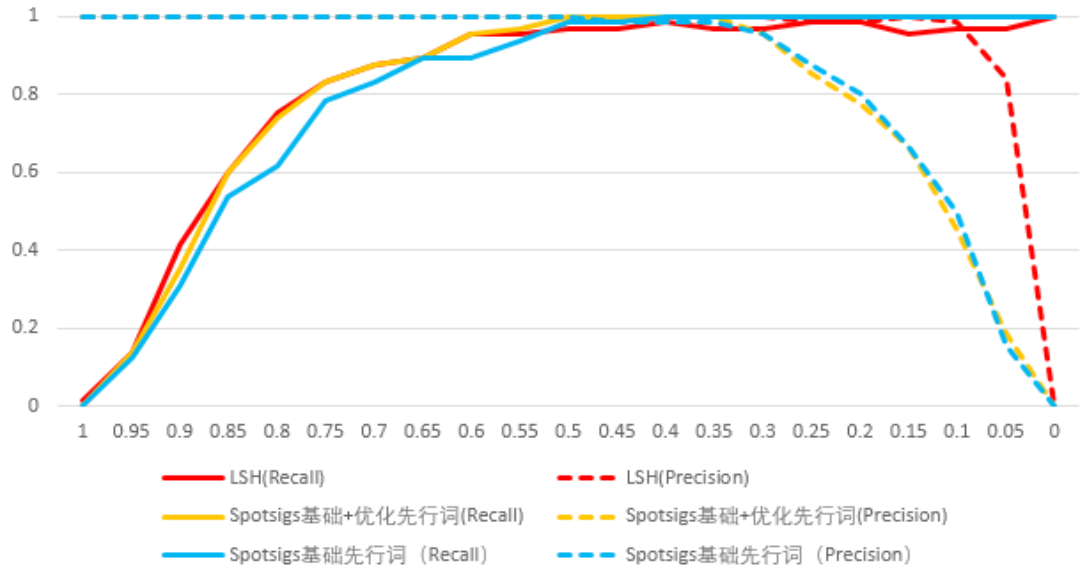

图4 LSH及使用两种先行词选择策略的SpotSigs算法准确率，召回率随相似度阈值变化曲线

通过实验发现，阈值 τ=0.6 时有 3 对假阴性结果，在阈值 τ=0.5 时可检测出所有相似新闻对。具体来说，共有 65 对相似新闻网页对被检测出来，其中相似度高于 0.8 的新闻网页对占 78.6%。其中，军事类平均相似度为 0.786，体育类 0.823，政治类 0.921，生活类 0.839，财经类 0.837。在政治类新闻检测效果最好，军事类新闻评价相似度较低，与数据集中军事类新闻网页大部分为军事简讯，文本较短有关。

为了更好的观察 SpotSigs 的性能，实验使用 LSH 算法（k=6, l=32）进行比对。图 4 给出 LSH 与使用两种先行词选择策略的 SpotSigs 算法的准确率、召回率随相似度阈值变化的曲线。不难发现，使用优化先行词之后 SpotSigs 算法的召回率会有稍许提升。当相似度阈值在 0.35–0.6 时，两个算法都显示较好的效果。当相似度阈值在 0.35-0.5 时，Spotsigs 可以准确检测出所有的真阳性中文新闻网页对，而且不包含假阳性的结果 ,LSH 则不能完全检测出所有真阳性中文新闻网页对。在相似度阈值为 0.1–0.35 时，LSH 则有更高的准确率。结果证明，在相似度阈值较低的情况下，基于 SpotSigs 的中文新闻相似性检测方法可以检测出所有相似网页。

5 结语

移动互联网的高速发展使网络新闻的发布和阅读变得方便、快捷，网络新闻也成了政府引导舆论并与公众交流的重要平台。但是网络新闻中存在着大量的转载、重复现象 [26]，对新闻聚合器以及开源情报分析等工作带来不利影响。本文的研究重点主要针对日益冗余的中文新闻网页相似性检测工作。对于中文新闻网页中存在广告、导航栏等噪声影响以及中文区别于英文的语言特性 [27]，在原有 SpotSigs 算法的基础上，提出针对中文新闻网页的相似度检测方法，并对方法进行了实例研究验证。验证结果表明，本文所设计方法可以有效检测相似中文新闻网页，特别是在相似度阈值较低的情况下具有极好的性能，且对长篇新闻网页的检测效果较之短篇新闻网页要好。

当然，由于本文所使用实验数据的规模较小，尚难以判断本文所提方法是否适应大数据时代的要求，未来将尝试以众包（crowdsourcing）的方式构建一个大规模的相似中文新闻网页的标注语料，以验证本文所提方法的伸缩性。另外，如何对简讯与一般新闻网页进行相似检测也是下一步的研究方法之一。

参考文献

[1] 中国互联网络信息中心（CNNIC）. 第 39 次《中国互联网发展状况统计报告》[EB/OL]. [2016-07-13]. http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/hlwtjbg/201701/P020170123364672657408.pdf.

[2] 甄书秀 . 网络新闻传播的舆论引导功能 [J]. 军事记者 , 2004(1): 52-53.

[3] 周涛 . 网络新闻的作用与影响 [J]. 中国传媒科技 , 2012(18): 107-108.

[4] 张利国 , 王恩海 . 2004 年中国互联网络信息资源发展的特点 [J]. 中国信息导报 , 2005(7): 23-24.

[5] 王葆慧 . 新闻转载的正面效应探析 [J]. 赤峰学院学报 : 汉文哲学社会科学版 , 2012(9): 214-215.

[6] 李东平 . 搜索引擎对新闻业的影响 [J]. 西南民族大学学报 : 人文社科版 , 2010, 31(6): 131-134.

doi:10.3772/j.issn.2095-915x.2017.03.014

[7] 郑鹏 . 搜索引擎中的相似网页探测算法研究 [D]. 武汉 : 华中科技大学 , 2008.

[8] Theobald M, Siddharth J, Paepcke A. SpotSigs: Robust and Efficient near Duplicate Detection in Large Web Collections[C]// International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2008: 563-570.

[9] Broder A. On the Resemblance and Containment of Documents[C]// Compression and Complexity of Sequences 1997. Proceedings. IEEE, 1997: 687-690.

[10] Brin S, Davis J, Garcia-Molina H. Copy Detection Mechanisms for Digital Documents[C]// ACM SIGMOD Record. ACM, 1995, 24(2): 398-409.

[11] Heintze N. Scalable Document Fingerprinting[C]// 1996 USENIX Workshop on Electronic Commerce. 1996.

[12] Chowdhury A, Frieder O, Grossman D, et al. Collection Statistics for Fast Duplicate Document Detection[J]. ACM Transactions on Information Systems, 2002, 20(2): 171-191.

[13] Kołcz A, Chowdhury A. Lexicon Randomization for Near-duplicate Detection with I-Match[J]. The Journal of Supercomputing, 2008, 45(3): 255-276.

[14] Shivakumar N, Garcia-Molina H. SCAM: A Copy Detection Mechanism for Digital Documents[C]// Internetional Conference in Theory and Practice of Digital Libraries. 1995.

[15] Garcia-Molina H, Gravano L, Shivakumar N. Dscam: Finding Document Copies across Multiple Databases[C]// Parallel and Distributed Information Systems, 1996., Fourth International Conference on. IEEE, 1996: 68-79.

[16] Gionis A, Indyk P, Motwani R. Similarity Search in High Dimensions via Hashing[C]// International Conference on Very Large Data Bases. Morgan Kaufmann Publishers Inc. 2000: 518-529.

[17] Cohen E, Datar M, Fujiwara S, et al. Finding Interesting Associations without Support Pruning[J]. IEEE Transactions on Knowledge & Data Engineering, 2001, 13(1): 64-78.

[18] Bawa M, Condie T, Ganesan P. LSH Forest: Self-tuning Indexes for Similarity Search[C]// International Conference on World Wide Web. ACM, 2005: 651-660.

[19] Mao X, Liu X, Di N, et al. SizeSpotSigs: An Effective Deduplicate Algorithm Considering the Size of Page Content[C]// Advances in Knowledge Discovery and Data Mining Pacific-Asia Conference, PAKDD 2011, Shenzhen, China, May 24-27, 2011, Proceedings. 2011: 537-548.

[20] Varol E, Can F, Aykanat C, et al. CoDet: Sentence-based Containment Detection in News Corpora[C]// Proceedings of the 20th ACM international conference on Information and Knowledge Management. ACM, 2011: 2049-2052.

[21] 唐蓉 . 搜索引擎重复网页检测技术研究 [D]. 重庆 : 重庆理工大学 , 2011.

[22] 梁浩 . 网络新闻相似度检测系统 [D]. 长春 : 吉林大学 , 2011.

[23] 徐德玉 . 中文文档内容相似度检测方法研究 [D]. 长春 : 长春工业大学 , 2010.

[24] 项炜 , 金澎 . 大规模语料库上的 Stanford 和 Berkeley 句法分析器性能对比分析 [J]. 电脑知识与技术 , 2013(8): 1984-1986.

[25] 顾益军 , 樊孝忠 , 王建华 , 等 . 中文停用词表的自动选取 [J]. 北京理工大学学报 , 2005, 25(4): 337-340.

[26] 李雪 . 面向网络新闻的舆情检测与分析系统设计与实现 [D]. 济南 : 山东师范大学 , 2014.

[27] 韦永壮 . 中文新闻重复网页检测研究 [D]. 南京 : 南京大学 , 2014.