网络信息时效技术①

陈 默[1,2], 杨小平[1], 柳 增[1], 孙丹雯[2]

[1](中国人民大学 信息学院, 北京 100872)
[2](北京联合大学 商务学院, 北京 100025)

摘 要: 随着大数据时代的到来, 对网络信息的时效性进行评价已成为当今研究的热点. 将以 Web 新闻作为研究对象, 对大数据环境下的 Web 信息提取和中文分词处理等技术进行研究, 并在此基础上, 提出一种基于 Web 语义信息提取的网络信息时效性评价算法.实验结果将充分体现算法实现的有效性, 既可引导网络用户关注更有价值的 Web 信息, 也可帮助网站管理者构建一个时效性更高的网站.

关键词: Web 语义提取; 网络信息时效性; 语义相似度; 语义距离

Network Information Currency Technology Based on Web Semantic Extraction Method

CHEN Mo[1,2], YANG Xiao-Ping[1], LIU Zeng[1], SUN Dan-Wen[2]

[1](School of Information, Renmin University of China, Beijing 100872, China)
[2](School of Business, Beijing Union University, Beijing 100025, China)

Abstract: With the arrival of the big data era, the currency evaluation of network information has become a spot for today's research. This paper will take Web news as the object of study and study the technology of Web information extraction and Chinese word segmentation in big data environment. On the basis of the above, this paper proposes an algorithm of network information currency evaluation based on Web semantic extraction method. The experimental results fully reflect the validity of the algorithm implementation. The study of technology plays a very important role in leading network users pay attention to more valuable Web information and helping Web site managers build a higher currency network.

Key words: Web semantic extraction; network information currency; semantic similarity; semantic distance

1 引言

当今的互联网与信息技术领域已步入了大数据时代, 新浪微博的注册用户已超过 3 亿, 其帖子内容的日更新量已超过 1 亿条[1]; 全球最大视频网站 YouTube 的日访问量已超过 10 亿次; 全球知名的社交网站 Facebook 日新增评论量已超过 32 亿条[2], 网络数据正在呈现爆炸式增长的态势[3].

在这种大数据的背景下, Web 信息仍存在着更新不及时、垃圾信息较多、信息来源不可靠等现象, 这已使得网络用户对 Web 时效性提出了质疑, 并已成为当前迫切需要解决的问题. 若要保证网络信息具有高时效性, 若要快速有效地引导网络用户关注更有价值的 Web 信息, 若要帮助网站管理者构建一个时效性更高的网站, 就应综合考虑 Web 信息规模性(Volume)、多样性(Variety)、高速性(Velocity)和价值性(Value)等具体要求. 本文将以网络信息中 Web 新闻作为研究对象, 提出一种基于 Web 语义提取方法的网络信息时效性评价算法, 以探究如何对 Web 时效性进行准确的评价, 并利用实验结果表明该算法的有效性.

2 相关工作

近年来, 许多国内外的专家和学者运用不同的理论和方法对 Web 语义进行了大量的研究, 并取得了一定的成果. 比如, 朱旭东在文献[4]中对 Web 语义标注进行了深入的研究, 并在此基础上, 还构建了一个面

①基金项目:国家自然科学基金(71271209);北京市优秀人才培养项目(2012D005022000013);北京市教育委员会社科计划面上项目(SM201311417008);北京联合大学人才强校计划人才资助项目(BPHR2012A02)

收稿时间:2013-12-09;收到修改稿时间:2014-01-23

向 Deep Web 的搜索引擎原型系统; 赵良等人在文献[5]中对基于语义层次的 Web 个性化信息推荐方法进行了深入的研究, 并在此基础之上, 还导出了语义层次的 Web 使用文档和生成了个性化推荐的 Web 页面集; CELIK Duygu 等人在文献[6]中对 Web 服务中的语义匹配方法进行了深入的研究, 并在此基础上, 还提出了一个 Web 语义匹配算法, 以通过服务质量参数标准的匹配过程挖掘出相似的 Web 服务; 李坤等人在文献[7]中对基于带有约束性提取和结构分析的 Web 语义服务发现算法进行了深入的研究, 提出了基于约束提取的概念性语义匹配方法和在匹配失败情况下的基于结构分析的算法; 黄辉等人在文献[8]中对 Web 语义自动提取服务的组合模型进行了深入的研究, 在分析了当前分布式模型管理发展现状的基础上, 提出了模型多重组合的方法, 上述工作是对 Web 语义应用方向的一些相关研究.

对于 Web 时效技术, 一些专家和学者对其也进行了一定的研究. 比如, 袁敏等人在文献[9]中对主动式有状态的 Web 时效索引机制进行了深入的研究, 并在分析现有 Web 服务机制的基础上, 还提出了一种能够适应时效索引机制的 Web 服务单元; 沈云斐等人在文献[10]中对基于时效性的 Web 页面个性化推荐模型进行了深入的研究, 并在此基础上, 还提出了一种基于时效价值系数的增量挖据算法; 杨朝军在文献[11]中对基于一种信息时效自动链接方法的 Web 用户界面导航可用性进行了改进研究, 并提出了一种设计方法, 以支持信息时效自动链接和用户界面的导航作用; 马崔常等人在文献[12]中对基于信息系统时效评估的用户行为分析方法进行了深入的研究, 提出了如何进一步提取通用信息时效行为的模型结构, 以替代 Ellis 模型, 并又提出如何利用其它理论系统的关键元素去整合通用信息时效行为的结构模型, 以构建一个相对完整的理论系统; 白志斌等人在文献[13]中对 Web 可用性自动评估系统设计与实现进行了深入的研究, 提出了面向基于大量真实用户行为数据的 Web 自动评估系统, 上述工作是对 Web 时效技术应用方向的一些相关研究.

从上述已完成的相关工作中可看出, 专家学者分别对 Web 语义和时效技术进行的研究成果已成规模, 但利用 Web 语义提取方法对网络信息时效性评价进行深入研究的成果还较少. 因此, 本文将重点提出一种基于 Web 语义提取方法的网络信息时效性评价算法, 以探究如何对 Web 可用性的时效进行准确的评价.

3 理论基础

在对基于 Web 语义提取方法的网络信息时效技术进行研究之前, 针对本文的研究对象, 本节将首先简述 Web 新闻时效性所产生的社会作用, 其次分析 Web 新闻语义的提取在时效性评价中所起到的作用, 这将为本文的研究重点奠定理论基础.

3.1 Web 新闻时效性

Web 新闻时效性是指 Web 新闻是否能够及时报道最新发生的事件, 是否能够激发广大网民的共同讨论兴趣, 是否能够反映当前舆论所关注的焦点, Web 新闻内容中的核心事件发生在何时, 而此新闻又在何时发布到网上, 新闻发布后又会引起何种社会舆论效果, Web 新闻时效性评价的结果必将对新闻网的社会影响力产生重要的作用.

从 Web 新闻传播的过程分析, Web 新闻时效性主要受三个时间距离的影响, 其一是从核心事件发生到 Web 新闻发布之间的时距; 其二是从 Web 新闻发布到网络用户关注新闻之间的时距; 其三是从网络用户关注新闻到 Web 新闻所产生的社会舆论之间的时距.若要求 Web 新闻具有较高的时效性, 就需要保证 Web 新闻的传播速度应与其价值成正比, 尽可能地缩短 Web 新闻传播过程中各环节间的时距, 并保证 Web 新闻要在合适的时机发布, 以吸引网络用户的情感注意力, 更大限度地增强 Web 新闻报道的效果.

综合考虑影响 Web 新闻时效性的三个因素, 本文将通过两个基础性的指标、一个计算性的指标和一个参考性的时效区间等定量特征来定性评价 Web 新闻时效性. 两个基础性的指标分别为文本距离和时间距离, 一个计算性指标是语义距离, 本文将利用 Web 新闻元素的提取技术对文本数据进行结构化处理; 将利用 Web 新闻内容的语义分析技术对 Web 新闻内容进行分词处理、对时间词进行精确标注、对时间词所对应的事件进行定位、对二元组语义信息进行提取、对待评价新闻与其它新闻之间的文本距离和时间距离进行计算; 将利用 Web 新闻噪声记录过滤技术对 Web 新闻语义相似度较低、时间差较大的噪声记录进行处理; 利用 Web 新闻时效性评价技术计算 Web 新闻的语义距离、分析 Web 新闻的时效性区间和待评价 Web 新闻的

时效性区间、分析 Web 新闻的时效性评价结果和待评价 Web 新闻的时效性评价结果.

3.2 Web 新闻语义

Web 新闻群是对网络信息进行时效技术研究的对象, 其实例可以是任何一篇 Web 新闻. 不同于报刊和媒体上所传播的新闻, Web 新闻需要利用半结构化的超文本标识语言(Hypertext Markup Language, HTML)或非结构化的文档描述, 并借助于互联网供用户阅读、评议和传播[14]. 互联网虽然能够显示网页中的 Web 新闻, 但不能真正理解 Web 新闻所要表达的含义, 更不能对 Web 新闻进行自动化处理, 这就为网络信息时效技术研究带来了困难, 然而, Web 语义可将新闻所要表达的信息转换成计算机能够理解和处理的形式[15], 这又将为网络信息时效技术的研究提供了有效的数据分析方法.

在 Web 新闻实例中, 其结构主要包括了新闻 URL、标题、发布时间、导语、主体和结语等多个元素, 而在 Web 新闻导语、主体或结语中, 所报道的核心事件和其发生时间可以组合成一个二元关系, 而由核心事件所引发的外延事件和其所对应的发生时间又可以组合成多个二元关系, 这为网络用户获取 Web 新闻的主题信息提供了重要的引导作用.因此, 除了提取 Web 新闻中的众多元素以外, 如何精准地提取 Web 新闻内容中时间和事件的语义信息, 并将其进行有效的关联, 这将成为本文研究的创新点.

4 基于Web语义提取方法的网络信息时效性评价算法

在大数据时代发展的背景下, 如何通过数据的提取、语义分析、噪声过滤等过程研究网络信息时效技术, 这已成为 Web 文本挖掘的一个重要研究方向.基于对大数据进行流处理(Stream Processing)和批处理(Batch Processing)两种模式的研究[16], 依据 Web 新闻所蕴含的价值将会随着时间的流逝而不断减少的特点, 基于 Web 语义提取方法的网络信息时效性评价算法将对 Web 新闻采用流处理的模式, 尽可能快地对 Web 新闻做出时效性分析, 并得出时效性评价结果, 为进一步评价网络信息时效性提供可借鉴的方案.

4.1 算法设计思想

该算法采用了分层设计思想, Web 新闻元素的提取层主要负责将 Web 新闻群或 Web 新闻实例中包含的新闻 URL、标题、发布时间和内容等文本数据进行结构化处理, 并将提取的结果组织成 Web 新闻语料库, 以供 Web 新闻内容的语义分析层使用.

作为算法的核心层, Web 新闻内容的语义分析层主要完成如下多个任务, 首先对 Web 新闻内容进行分词处理; 其次对时间词进行精确标注; 再次定位时间词所对应的事件; 再次提取二元组语义信息; 最后计算待评价新闻与其它新闻之间的文本距离和时间距离, 并将分析的结果组织成 Web 新闻语料库, 以供 Web 新闻噪声记录的过滤层使用.

Web 新闻噪声记录的过滤层主要负责的任务是过滤掉 Web 新闻语义相似度较低、时间差较大的噪声记录, 以降低噪声数据对 Web 新闻时效性评价的干扰.

作为算法的核心层, Web 新闻时效性的评价层主要负责如下多个任务, 首先计算 Web 新闻的语义距离; 其次分析出 Web 新闻的时效性区间和待评价 Web 新闻的时效性区间; 最后分析出 Web 新闻的时效性评价结果和待评价 Web 新闻的时效性评价结果, 并将分析的结果组织成 Web 新闻语料库, 该库可为 Web 新闻站点的时效性评价提供数据支持.

在上述分层的算法设计思想中, 如图 1 所示, Web 新闻内容的语义分析层和 Web 新闻时效性的评价层是本文研究的重点, 其算法的实施过程也是本文研究的创新点, 它可为网络信息的 Web 语义提取和时效性研究奠定真实有效的实践基础.

4.2 算法过程描述

基于 Web 语义提取方法的网络信息时效性评价算法流程图, 本节将详细研究算法的流程, 输入内容为 Web 新闻实例的 URL, 输出内容为 Web 新闻的标题、发布时间、发布内容、内容分词结果、内容特征词、<时间, 事件>二元组语义信息、语义距离的分布、时效性区间、时效性结果等.在整个算法中, 需要用于分析的数据均从网络资源中直接获取, 以保证这些数据均与本文的研究对象密切相关, 并且数据量较大, 数据质量真实有效, 算法的具体过程如下所示.

Algorithm Step 1:

根据输入的 Web 新闻实例的 URL, 提取 Web 新闻的标题、发布时间、发布内容等元素, 并将其存储在 Web 新闻语料库中. 为了提高 Web 新闻元素提取的准确率和效率, 本步骤将利用能够解析 HTML 页面的开

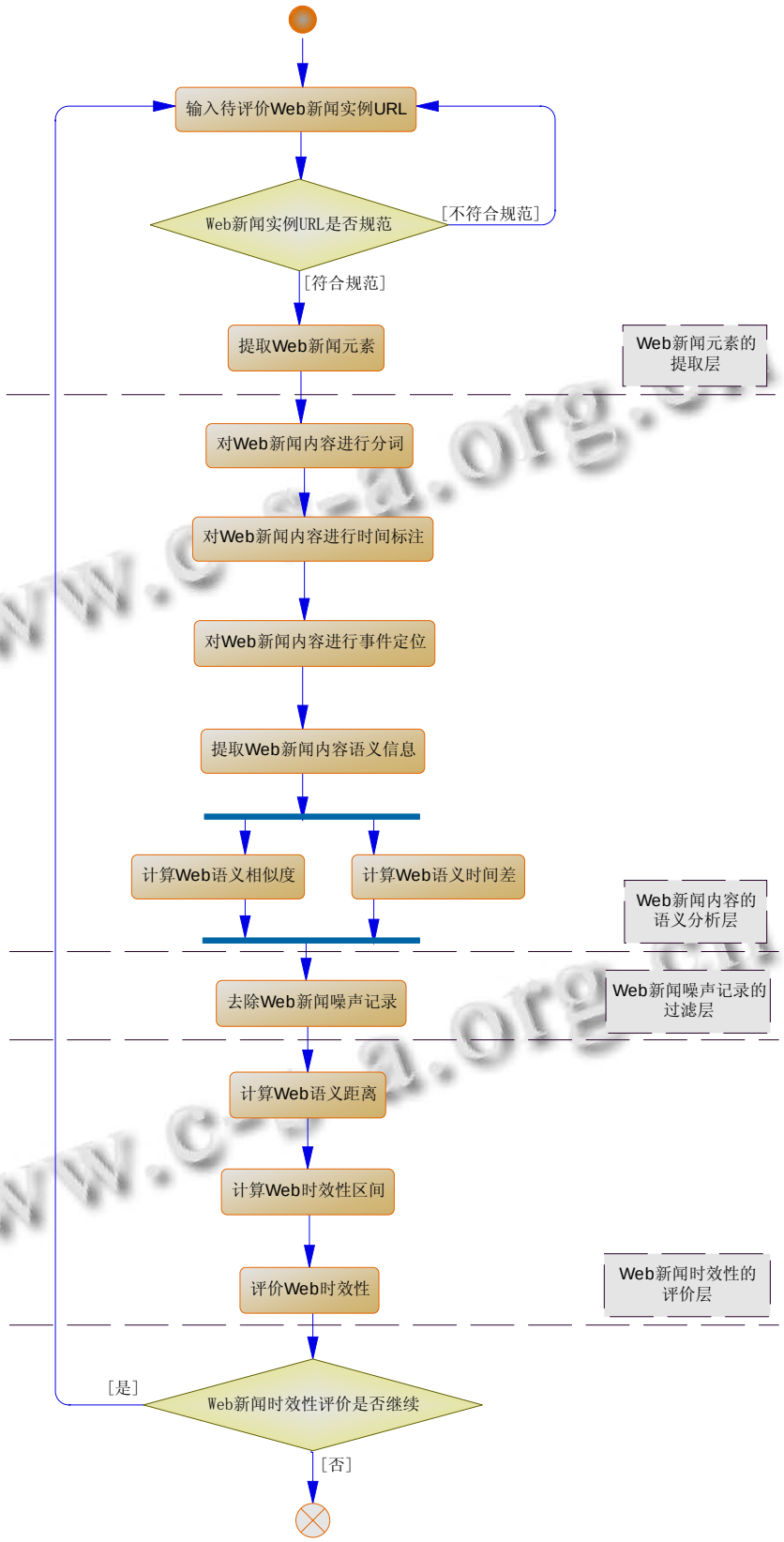

图 1 基于 Web 语义提取方法的网络信息时效性评价算法流程图

源库NekoHtml, 先将Web新闻页面中的数据转换成纯文本格式, 再通过分析Web新闻元素的组织结构特征, 有针对性地在<title>标签中定位 Web 新闻标题, 并在Web新闻标题的下一行定位其发布时间. 由于Web新闻的发布时间和内容具有一定的规范格式, 因此, 本步骤将设计相应正则表达式, 以判断提取的文本是否为时间信息, 以判断提取的内容是否为Web新闻内容.

Algorithm Step 2:

利用基于Lucene的中文分词处理技术[17], 为了提高 Web 新闻内容的划分准确率和效率, 本步骤将已存储在Web新闻语料库中的Web新闻内容进行分词处理, 并将其存储在 Web 新闻语料库中.

Algorithm Step 3:

本步骤是本文对算法研究的创新点, 为了提高从 Web 新闻中提取二元组语义信息的准确率和效率, 首先, 基于常用时间词库、时间方位词库、时间副词库和时间介词库, 本步骤将对 Web 新闻内容分词结果中的时间词进行精确标注, 并基于时间和事件的逻辑关系, 进一步定位时间词与其对应的事件; 其次, 通过整理已标定的时间和事件关系, 本步骤将从 Web 新闻中提取一系列<Time, Event>二元组语义信息, 并将其存储到 Web 新闻语料库中. Algorithm1 主要描述了中文时间词标注和Web语义信息存储的过程, Algorithm2 主要描述了事件定位和 Web 语义信息提取的过程.

```
Algorithm1: Tagging_Chinese_time
Input: strUrl,
WebNewsDB={r1, r2, …, ri, ri+1, …, rn}.
Output: pubTime, webContent, position.
Tagging_Chinese_time(strUrl)
//strUrl 对象存储了 Web 新闻实例的 URL
Begin
    setpubTime(selectPubTime(strUrl));
    //提取 Web 新闻语料库中的 strUrl 新闻实例的发布时间, 并存储在 pubTime 属性中.
    //此属性值将作为待评价新闻实例的待评价时间
    setwebContent(selectContent(strUrl));
    //提取 strUrl 新闻实例的分词内容, 并存储在 webContent 属性中.
    convert_ArrayList(webContent);
    //将带有词性标注的strUrl新闻实例的分词内容转存到列表 list 对象中
    for i=0 to list.size()-1 do
    //遍历list列表, 标注时间词出现的位置, 并将其添加到列表position对象中.
        if (timeWord(list.get(i))) then
                                position.add(i);
    event=Locating_event(position,list);
    //事件定位和 Web 语义信息提取, 该过程将在 Algorithm2 中描述.
    for n=0 to event.size()-1 do
    Begin
        keyWord(event.get(n));
        //统计<Time, Event>二元组中事件特征词出现的频率
        timeevent.add(event.get(n));
        //将能够表示时间和事件关系的<Time, Event>二元组添加到列表timeevent对象中
    End
    //将列表timeevent对象中的<Time, Event>二元组语义信息存储到Web新闻语料库
    for i=0 to timeevent.size()-1 do
    Begin
        str=str+(String)timeevent.get(i)+"\n";
        DBConnection.update(con,"update
                newsinfo set
    keyOfEvent=\'"+str+"\'where url=\'"+strUrl+"\'");
    End
End
Algorithm2: Locating_event
Input: position, list.
Output: event
Locating_event(position,list)
Begin
    for j=0 to position.size()-1 do
    //遍历已标注的时间位置, 定位时间所对应的事件.
    Begin
        int k=(Integer)position.get(j);
        String time_word=(String)list.get(k);
        //提取标注的时间词
```

```
boolean is=false;
while
(!((String)list.get(k+1)).endsWith("/t")
                    && k<list.size()-2) do
//在相邻时间词之间, 搜索前驱时间对应的事件.
Begin
    if
    (((String)list.get(k+1)).endsWith("/n"
 ) && testChinese((String)list.get(k+1)))
    then
    //将搜索到的前驱时间所对应的事件词加入到event_word对象中
    Begin
        event_word.append((String)list.
                        get(k+1)+" ");
        is=true;
    End
End
if (!is) then continue;
else
Begin
    String
final_event_word=event_word.toString();
    final_event_word=final_event_word.
                        replace("/n", "");
    event.add(time_word+" "+
                    final_event_word);
    //将时间所对应的事件整理成二元组, 添加到列表event对象中.
    End
  End
End
```

Algorithm Step 4:

本步骤是本文对算法研究的创新点, 为了提高 Web 文本距离计算的准确率和效率, 利用已提取的<Time, Event>二元组集合, 根据代表事件的特征词个数, 本步骤将采用向量空间模型(Vector Space Model, VSM)或字符串匹配方式计算 Web 新闻语义相似度, 即 Web 文本距离.若代表事件的特征词大于 10 个, 则采用 VSM 计算 Web 新闻语义相似度; 若代表事件的特征词小于或等于 10 个, 则无法提供足够的维度去区分事件间的关系, 利用 VSM 计算出的 Web 新闻语义相似度的准确性也较差, 因此, 将选择采用字符串匹配方式计算 Web 新闻语义相似度.

Algorithm Step 5:

本步骤是本文对算法研究的创新点, Web 新闻时间差是指待评价 Web 新闻发布时间和其它与之相关的 Web 新闻发布时间的差值[18], 即 Web 时间距离, 由于时间划分粒度的不同, 为了提高 Web 时间距离计算的准确率和效率, 本步骤将利用如下公式计算 Web 新闻时间差.

$$T = \frac{t_i - \min(t_j)}{\max(t_j) - \min(t_j)} \tag{1}$$

如公式(1)所示, ti 表示待评价 Web 新闻的发布时间, min(tj) 表示与待评价 Web 新闻相关的所有 Web 新闻发布时间的最小值, max(tj) 表示与待评价 Web 新闻相关的所有 Web 新闻发布时间的最大值, T 表示 Web 新闻时间差.

Algorithm Step 6:

基于训练样本集判别式, 本步骤首先将对 Web 新闻信息进行分类, 并在得到噪声集之后, 再采用 Fisher 判别方法过滤噪声数据[19], 并对语义相似度较低和时间差较大的 Web 新闻记录进行过滤处理.

Algorithm Step 7:

本步骤是本文对算法研究的创新点, 为了提高 Web 语义距离计算的准确率和效率, 将采用 Euclid Distance 公式计算 Web 新闻的语义距离, 如公式 2 所示, S 表示 Web 新闻的文本距离, T 表示 Web 新闻的时间距离, 表示权重系数, D 表示 Web 新闻的语义距离, 即 Web 语义距离.从公式中可看出, D 值越小, 说明 Web 新闻的语义距离越小, Web 新闻实例间的关联程度也就越大.

$$D = \sqrt{(1-S)^2 + \omega T^2} \tag{2}$$

Algorithm Step 8:

本步骤是本文对算法研究的创新点, 在获得了待评价 Web 新闻与其它 Web 新闻的语义距离分布关系之后, 本步骤将推断出最短的 Web 新闻时效性区间, 此区间内包含的 Web 新闻实例数将占总新闻数的 80%以上, 并且位于时效性区间内的 Web 新闻将与待评价

Web 新闻的相似度之和达到最大. 若待评价的 Web 新闻发布时间落在该时效性区间内, 则说明该 Web 新闻所报道的信息时效性较高, 反之则较低.

参考上述算法的描述过程和已得出的时效评估结果, 不仅能够判断 Web 信息是否反映了当前网络信息传播的焦点内容, 而且还能够从更细的粒度上进一步确定 Web 信息被关注的程度, 还可向网络用户或网站管理者提供简短易懂的 Web 时效性评估报告, 以说明网络用户所浏览的 Web 信息的时效性, 帮助网站管理者掌握所发布的 Web 信息的时效性, 这必将有助于高时效性网站的商业运营.

5 实验结果

基于 Web 语义提取方法的网络信息时效性评价算法, 以网络数据中的 Web 新闻作为实验对象, 笔者设计了模拟实验的窗体.该窗体利用 MyEclipse 平台中的 Matisse Form Class 作为顶层容器[20], 其包括了获取、筛选、存储、分析、过滤和评价 Web 新闻群和实例的模块; 爬取待评价 Web 新闻源数据、标题、发布时间、内容的模块; 将待评价 Web 新闻内容进行分词, 并完成特征词统计的模块; 提取<Time, Event>二元组集合, 并将上述已爬取、统计和分析待评价 Web 新闻的结果存入 Web 新闻语料库的模块; 生成待评价 Web 新闻时效区间和时效结论的模块.

以 http://news.163.com/13/0819/21/96M1AV6S00014JB6.html 为待评估 Web 新闻对象, 首先, 在 Web 新闻实例内容爬取界面中, 笔者输入 Web 新闻实例 URL, 单击 “Web News Crawling”按钮, 实验控制器将接收到的爬取 Web 新闻实例内容的请求转交给模型处理, Web 新闻实例内容爬取界面上将高效且精确地显示爬取到的 Web 新闻源数据、标题、发布时间和内容, 如图 2 所示.

其次, 在 Web 新闻实例内容分析界面中, 单击 “Web News Divided Words”按钮, 根据 Web 新闻实例内容爬取界面中已爬取的 Web 新闻正文内容, 实验控制器将接收到的分词请求转交给模型处理, Web 新闻实例内容分析界面上将高效且精确地显示分词结果; 单击“Web News Words Frequences”按钮, 根据已分词的结果, 实验控制器将接收到的特征词统计请求转交给模型处理, Web 新闻实例内容分析界面上还将高效且精确地显示特征词统计的结果, 如图 3 所示.

图 2 Web 新闻实例内容爬取界面

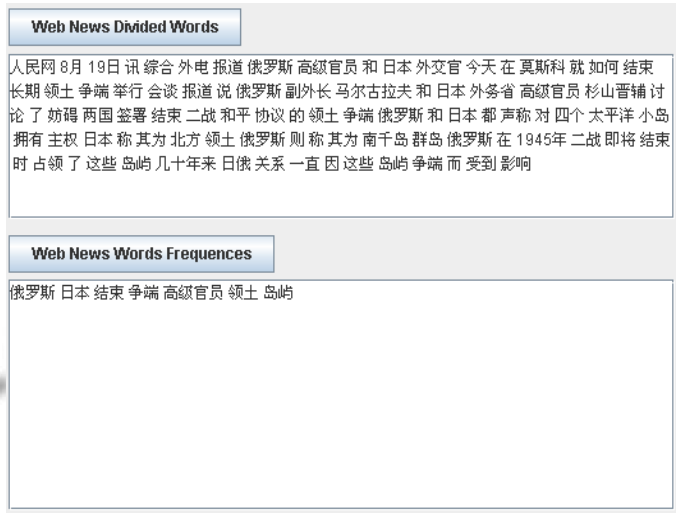

图 3 Web 新闻实例内容分析界面

再次, 在 Web 新闻实例内容<Time, Event>提取界面中, 单击 “Web News Time And Event”按钮, 根据上述实验步骤的结果, 实验控制器将接收到的<Time, Event>提取请求转交给模型处理, Web 新闻实例内容<Time, Event>提取界面上将高效且精确地显示<Time, Event>提取结果, 这是本文应用 Web 语义提取方法的特色, 如图 4 所示.

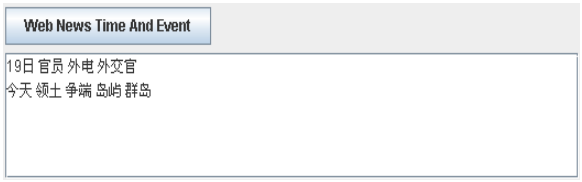

图 4　Web 新闻实例内容<Time，Event>提取界面

最后，在 Web 新闻实例内容评价界面中，单击“Web News Time Range and Effect”按钮，根据上述实验步骤的结果，实验控制器将接收到的评估请求转交给模型处理，Web 新闻实例内容评价界面上将高效且精确地显示 Web 新闻时效性评价结果，如图 5 所示. 在该图中，其它 Web 新闻实例是指与待评价 Web 新闻实例主题相关的 Web 新闻，其分析过程与待评价 Web 新闻相同，通过计算待评价 Web 新闻分别与 18 个 Web 新闻实例之间的文本距离和时间距离，即可得出它们之间的语义距离，进而统计出待评价 Web 新闻的时效性区间，本实验待评价的 Web 新闻发布时间落在了该时效性区间内，则说明该 Web 新闻所报道的信息时效性较高，这是本文应用 Web 语义提取方法进行网络信息时效评价的特色.

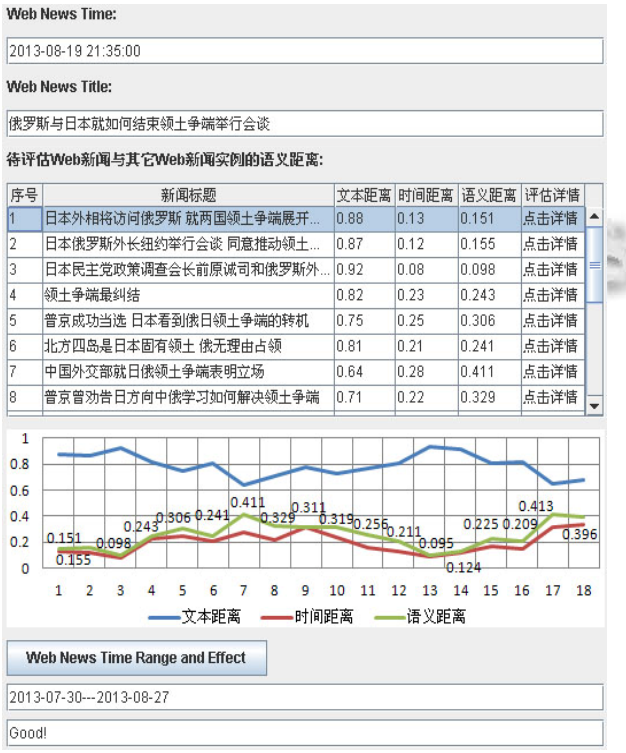

图 5　Web 新闻实例内容评价界面

本实验已利用搜狐新闻、新浪新闻、网易新闻、腾讯新闻、新华网、人民网、中国新闻网中 2013 年 1 月至 10 月的 Web 新闻作为真实数据集，从表 1 中可看出，该算法在各层次处理每个 Web 新闻的平均时间和准确率均较高，达到了算法设计的预期效果，但在平均时间上仍有提升的空间.

由于评价网络信息时效的其它算法还未成熟，因此，笔者只将本文所实验的算法结果同网络用户的调查结果进行了对比，在 Web 新闻时效性的评价结果中，有 97.472%同网络用户的调查结果相同，达到了算法预期的应用目标，但在准确率上仍有提升的空间.

表 1　本算法在各层次处理每个 Web 新闻的平均时间和准确率

	Web 新闻元素提取层	Web 新闻内容语义分析层	Web 新闻噪声记录过滤层	Web 新闻时效评价层
平均时间	0.178s	0.241s	0.054s	0.209s
准确率	99.152%	98.028%	99.985%	97.472%

6　结论

以 Web 语义提取方法为核心，通过 Web 数据的获取、数据的分析、数据的过滤、时效性评价等流程，本文完成了一种基于 Web 语义提取方法的网络信息时效技术研究.该研究实验证明了 Web 时效性评价需求的可行性，对于提高用户检索 Web 信息效率、提升 Web 可用性和科学建设、改进 Web 站点服务性能、提高 Web 商务运营效率和点击率，该研究过程必将具有一定的实际应用价值.在后续的研究过程中，笔者还可在时间效率和准确率等方向上，对已提出的基于 Web 语义提取方法的网络信息时效性评价算法进行再改进，还可对网络信息时效性区间的再细分和如何再划定多个时效性评价等级进行深入的探究.

参考文献

1 Zhou X, Li F. Mining aspects and opinions from microblog events. Journal of Computational Information Systems, 2013, 9(6): 2399–2400.

2 Yan Q, Wu LR, Zheng L. Social network based microblog user behavior analysis. Physica A-statistical Mechanics and its Application, 2013, 392(7): 1712–1713.

3 Suraj Pandey Surya N. Cloud computing and scientific

applications-big data, scalable analytics, and beyond. Future Generation Computer Systems, 2013,29(7):1774−1775.

4 朱旭东.基于本体学习的 DeepWeb 语义标注关键问题研究[学位论文].苏州:苏州大学,2012.

5 赵良,张云婧.一种以 Web 语义挖掘的个性化信息推荐设计.电脑知识与技术,2011,7(8):1731−1732.

6 Celik D, Elci A. A broker-based semantic agent for discovering Semantic Web services through process similarity matching and equivalence considering quality of service. Science China Information Sciences, 2013, 56: 2−3.

7 Li K, Jiang LL. Research of semantic web service discovery algorithm based on constraint extraction and structure analysis. Computer Engineering & Science, 2013, 35(8): 144−146.

8 Huang H, Chen XG, Wang ZW. Survey of automatic model composition based on semantic web service. Computer Science, 2013, 40(7): 9−11.

9 袁敏,黄志球,马潇潇.一种主动式有状态的 Web 时效索引机制.计算机应用研究,2004,24(7):224−225.

10 沈云斐,沈国强,蒋丽华,覃征.基于时效性的 Web 页面个性化推荐模型的研究.计算机工程,2006,32(13):80−81.

11 Yang CJ. Improvement of web usability with UI navigation based on an information auto-linking method. Journal of Modern Information, 2009, 29(9): 55−57.

12 Ma CC, Cao SJ. Information behavior method based on usability evaluation of information system. Information Studies: Theory & Application, 2013, 36(5): 98−100.

13 Bai ZB, Yang D, Li J. Design and implementation of automated usability evaluation system for web system. Computer Engineering and Design, 2013, 34(10):3649−3652.

14 Luo ZC, Tang JT, Wang T. Improving key phrase extraction from web news by exploiting comments information. Lecture Notes in Computer Science, 2013, 7808: 141−142.

15 Jiang B, Luo ZY. A new algorithm for semantic web service matching. Journal of Software, 2013, 8(2): 352−353.

16 Guo YK. Big data, big science, big collaboration: Delivering connected R&D for better value. Scientific Computing, 2013, 30(2): 5−6.

17 Xiu Chi. An analysis of key issues in Chinese word segmentation. Journal of Computational Information Systems, 2013, 9(3): 891−893.

18 Bai ST, Ning Y, Yuan S, Zhu TS. Predicting reader's emotion on Chinese web news articles. Lecture Notes in Computer Science, 2013, 7719: 18−19.

19 Li SH, Huang SM, Yen DC, Sun JC. Semantic-based transaction model for web service. Information Systems Frontiers, 2013, 15(2): 394−396.

20 Hao LC. Application of MVC platform in bank E-CRM. International Journal of Service, Science and Technology, 2013, 6(2): 34−36.