http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2016.0785

面向地震应急响应的互联网信息处理

宋建功[1], 王之欣[2], 李勤勇[1], 李雅婷[3], 马世龙[1], 吕江花[1,*]

(1. 北京航空航天大学 计算机学院, 北京 100083; 2. 中国科学院 软件研究所, 北京 100190;
3. 百度时代网络技术(北京)有限公司, 北京 100193)

摘 要: 近年来国内外地震灾害给人类的生命财产造成了巨大损失。海量的互联网地震灾情信息可为应急响应和及时救援提供决策支持的依据,因此互联网灾情信息高效快速处理需求迫切。针对地震互联网灾情信息处理的研究,定义了地震事件模型、网页对象模型等,用极限方法定义 Web 信息收敛性,刻画了互联网灾情信息传播特点。根据灾情信息时效性特点,提出了一种支持动态收敛性的 Web 信息抽取算法,完成互联网灾情信息提取。提出一种针对灾害信息随时间的变化进行时序统计的方法,形成信息统计报告,为制定救援决策提供依据。设计并实现了面向地震应急响应的互联网信息智能处理系统,并进行了工程实践验证。

关 键 词: 地震灾害; 应急响应; 信息提取; 数据清洗; 多文档摘要
中图分类号: V221+.3; TB553
文献标识码: A **文章编号**: 1001-5965(2017)06-1155-10

根据中国地震局官网统计得出,从 2007 年 1 月1 日至 2016 年 5 月 31 日近 10 年间灾害性的中国及周边地区 5 级以上地震 352 起,世界 7 级以上地震 197 起[1],其中多数地震都给人类的生命财产造成巨大损失。特别是 2008 年 5.12 汶川 8.0 级地震,受灾面积十万多平方公里,造成遇难和失踪人口 87 703 人,直接经济损失 8 451.4 亿元[2-3]。而应急响应信息系统建设滞后,应急救援人员对突发地震事件信息处理工作量大[4],阻碍了应急响应的效率。对于海量的互联网地震灾情信息及时有效地采集,进行灾情数据提取、分析,为及时救援决策提供可靠依据,最大限度地降低损失。

由于互联网应用的普及,网上资源具有海量性特点,互联网已经成为各行业数据重要的信息源。互联网信息提取[5-6]从传统的信息提取中衍生出来,它将整个互联网空间作为信息源,即从格式复杂、无结构的 Web 页面中识别出用户感兴趣或有价值的数据[7],并表示成一种语义清晰的结构化形式,使用户和程序能够更方便地查询和使用数据。信息提取技术,依据其自动化程度由低到高,依次分为手工信息提取、半自动信息提取和全自动信息提取[8]。文献[9]以形式化的方法定义了数据的一致性(consistency)、正确性(correctness)、完整性(completeness)和最小性(minimality),而数据质量被定义为这 4 个指标在信息系统中得到满足的程度。Inmon[10]指出,数据质量是指信息和数据可作为规定应用的可信资料来源的程度,就是在正确的时间、正确的地点,将正确信息的正确集合提供给正确人员,用来支撑决策。

面向应急响应的互联网地震灾情信息处理特点体现为:①应急响应紧迫,需要系统立刻生成告

收稿日期: 2016-10-10; **录用日期**: 2017-01-09; **网络出版时间**: 2017-02-24 09:31
网络出版地址: www.cnki.net/kcms/detail/11.2625.V.20170224.0931.001.html
基金项目: 国家自然科学基金(61300007, 61305054); 软件开发环境国家重点实验室自主探索基金(SKLSDE-2012ZX-28, SKLSDE-2014ZX-06)

* **通讯作者**: E-mail: jhlv@nlsde.buaa.edu.cn

引用格式: *宋建功, 王之欣, 李勤勇, 等. 面向地震应急响应的互联网信息处理[J]. 北京航空航天大学学报, 2017, 43(6): 1155-1164. SONG J G, WANG Z X, LI Q Y, et al. Internet information processing for earthquake emergency response [J]. Journal of Beijing University of Aeronautics and Astronautics, 2017, 43(6): 1155-1164 (in Chinese).*

警信息,并启动互联网信息处理;②地震事件相关的网页数据具有生命周期性特点,信息提取具有时效性;③互联网信息存在大量重复和冗余,需要进行有效的数据筛选。因此,海量互联网信息手动处理几乎不可能,要求处理流程必须自动化。基于此,面向强震应急响应的互联网信息处理包括信息提取、网页信息清洗、灾情数据时序统计、Web 灾情信息摘要生成等几部分。

对于互联网上的灾情信息经过提取、清洗后生成可供决策支持的摘要,摘要生成方法研究取得了非常广泛的进展。基于查询的聚类和摘要生成方法对每类文档聚类,生成一个简洁且综合的符合用户查询的摘要[11];用机器学习方法通过检测用户-对象关系来建模,不同用户模型可以独立使用或以一个集成方式使用,对于相同信息源和查询,不同用户能够生成不同的摘要[12]。对于单文档和多文档摘要,预处理把文档分解成句子,选择出突出的句子,随后连接所有选中句子生成摘要[13]。通过语义角色标注假定变量结构来表示源文档的内容,基于优化的特性对假定变量结构分类来生成摘要内容[14]。以事件作为基本语义单元的生成式摘要方法,通过对事件聚类反应篇章的主题分布,并利用事件指导多语句压缩生成自然语句来构建摘要[15]。

对于互联网信息处理的研究涉及到广泛的领域,各种方法都做出相应的贡献,但对于地震应急响应领域研究尚显不足。本文将针对地震应急响应领域以地震事件相关互联网信息提取、灾情相关网页数据清洗、灾害信息时序统计和 Web 灾情数据摘要生成几部分加以论述。并在笔者所在课题组研发的应用软件系统“国际强震应用处置系统之互联网信息智能处理子系统”中应用上述方法,进行了工程实践与验证。

1 地震事件相关互联网信息提取

地震事件相关的互联网信息具有动态性特点,信息提取贯穿整个地震应急响应过程中,并最终收敛于相对稳定的信息集合。

1.1 相关概念定义

互联网信息提取所涉及的相关概念如下。

定义 1 地震事件 EthqkEvt = (ID, Time, Location, ThemeSet, Content)。对于一个具体地震事件 ethqkevt,可定义为 ethqkevt = (id, t_0, l, themeset, content) ∈ EthqkEvt。

其中:id ∈ ID 为事件对象标识, ID 为事件对象标识集合;t_0 ∈ Time 为事件发生时间,Time 为事件发生时间集合;l ∈ Location 为事件发生的地点,Location 为事件发生地点集合;themeset = $\{\text{theme}_i \mid i = 1,2,\cdots,m\}$ ∈ ThemeSet 为某次具体地震事件的互联网灾情信息主题词集合,ThemeSet 为多个地震事件的嵌套集合;content ∈ Content 为提取出的主题相关内容,Content 为从互联网提取的地震事件相关内容集合。为了表示网页信息提取的动态性特点,用 content^t ∈ Content 表示地震事件后的 t 时刻($t > t_0$)提取的主题相关内容,Content 内容随时间变化动态更新,在某一时刻最终收敛,内容相对不变。

定义 2 互联网信息源 WIS = $\{\text{WebPage}_i \mid i = 1,2,\cdots,m\}$。

其中:WebPage_i($1 \leqslant i \leqslant m$)为互联网信息源中的某类网页。一个特定的信息源 wis = $\{\text{webpage}_i^t \mid i = 1,2,\cdots,m\} \subseteq$ WIS, webpage_i^t ∈ WebPage_i 为 t 时刻的一个特定网页。

通常某个灾情相关的互联网信息源包含若干类网页,本文所研究提取的网页通常是指地震行业专业网站网页和地震相关的新闻报道。

定义 3 灾情信息网页对象 webpage = {(ID, WebURL, URI, Title, PageContent, Time)}。

其中:WebURL 为灾情相关网站地址;URI 为网页地址;Title 为网页标题;PageContent 为网页中的正文内容。

1.2 地震事件相关互联网信息提取的收敛性

互联网上地震事件相关的灾情信息具有生命周期,所有信息源中新增网页随时间增加而减少,直至没有新增网页出现。

设某个地震事件产生时刻为 t_0,相关信息源 wis 中灾情相关网页更新时间序列为 t_0, t_1, …, t_i, t_{i+1}, …, $t_{i+1} = t_i + \Delta t$, Δt 为更新网页时间间隔。对应该时间序列提取的地震相关正文内容信息集合为 $\text{content}^{t_0}, \text{content}^{t_1}, \cdots, \text{content}^{t_i}, \text{content}^{t_{i+1}}, \cdots$, content^{t_0}为∅。

分别计算此集合序列的上、下极限,公式分别为式(1)和式(2)。

$$\overline{\lim_{i\to\infty}}\ \text{content}^{t_i} = \bigcap_{i=1}^{\infty}\bigcup_{j=i}^{\infty}\text{content}^{t_j} \tag{1}$$

$$\underline{\lim_{i\to\infty}}\ \text{content}^{t_i} = \bigcup_{i=1}^{\infty}\bigcap_{j=i}^{\infty}\text{content}^{t_j} \tag{2}$$

如果上极限和下极限都存在并且相等,即 $\overline{\lim_{i\to\infty}}\ \text{content}^{t_i} = \underline{\lim_{i\to\infty}}\ \text{content}^{t_i} = \psi$,则说明该集合序列的极限也存在,即 $\lim_{i\to\infty}\ \text{content}^{t_i} = \psi$,则必有某一时刻 t_{end},使得 $\text{content}^{t_{\text{end}}} = \psi$。或称为存在某一时刻

t_{end}使得地震相关正文内容信息集合收敛于 ψ。

1.3 地震事件相关互联网信息提取算法

对于互联网信息提取算法,给定地震灾害事件 ethqkevt,主要的算法描述为

算法 1 EERWE(Earthquake Event Related Webpages' Extraction)

输入:地震事件 ethqkevt = (id, t_0, l, themeset, content);地震灾情信息源集合 WIS;地震相关信息递减时刻经验性界定估值 N。

输出:灾情信息 $content^{t_{end}}$,分时刻的灾情信息统计报告 $report^{t_0}$, $report^{t_1}$, …, $report^{t_{end}}$。

```
1  content^{t_0} ← ∅; i ← 0
2  for each t_i do
3     content_1^{t_i} ← ISExtraction(wis_1) ‖
      content_2^{t_i} ← ISExtraction(wis_2) ‖ …
      ‖ content_m^{t_i} ← ISExtraction(wis_m)
4     content^{t_i} ← content_1^{t_i} ∪ content_2^{t_i} ∪ … ∪ content_m^{t_i}
5  content^{t_i} ← EESRC(content^{t_i})
6  report^{t_i} ← EEDITS(content^{t_i})
7  if t_i ≥ t_N then
      dis^{t_{i+1}} ← ContentDistance(content^{t_{i-1}}, content^{t_i})
8     if dis^{t_{i+1}} = -1 then
9        t_end ← t_i
10 content^{t_end} ← content^{t_i}
11    return content^{t_end}
12    end if
13 end if
14 i ← i + 1
15 end for
```

其中:$content_1^{t_i} \leftarrow ISExtraction(wis_1) \| content_2^{t_i} \leftarrow ISExtraction(wis_2) \| \cdots \| content_m^{t_i} \leftarrow ISExtraction(wis_m)$ 为每个信息源中的相关网页信息提取;EESRC($content^{t_i}$)为对 t_i 时刻提取出来的网页对象做重复记录的检测及去重;EEDITS($content^{t_i}$)为对文本内容集合 $content^{t_i}$进行灾情数据时序统计;ContentDistance($content^{t_{i-1}}$, $content^{t_i}$)为计算相邻网页内容集合 $content^{t_{i-1}}$和 $content^{t_i}$的距离 $dis^{t_{i+1}}$,若距离 $dis^{t_{i+1}}$为 -1 则表示相邻时刻网页内容集合之间的距离数列极限存在,信息提取收敛。

2 灾情相关网页数据清洗

用传统算法对互联网中的海量数据查重,其算法的时间复杂度和空间复杂度偏高。在地震应急响应领域应考虑到数据与地震事件相关度应具有不同优先级。因此本文提出一种改进的数据记录去重算法,具体表现为:定义数据属性优先级并计算相应的权值;对数据记录聚类,然后对所有类单独完成多次匹配。

定义 4 权重。

对于某次地震相关数据集 $D = \{d_i \mid 0 \leqslant i \leqslant n\}$,其属性向量 $\boldsymbol{P} = (P_1, P_2, \cdots, P_k, \cdots, P_m)$, P_k 为第 k 个属性,每个属性赋予不同值代表该属性的重要程度,称为属性的权重。对应于属性向量 $\boldsymbol{P}$ 的权重向量可表示为 $\boldsymbol{W} = (W_1, W_2, \cdots, W_k, \cdots, W_m)$,则任一数据 d_i 可表示为(d_{i1}, d_{i2}, …, d_{ik}, …, d_{im})。记录的相同属性值越多,并且相同属性的权重值越大,则记录的相似度越高。

2.1 属性字段权重计算

计算地震数据集属性字段权重的算法分 3 步:①行业专家为各个属性字段指定优先级;②取各专家指定属性字段优先级的均值作为该属性的最终优先级;③采用 RC(Rank-Centroid)算法[16]计算各属性字段的权重,属性 P_k 的权重 W_k 算法见式(3)。

$$W_k = W_k(\mathrm{RC}) = \frac{1}{T}\sum_{t=T}^{T_k}\frac{1}{t} \tag{3}$$

式中:T 为专家指定最低优先级,数值最小;T_k 为属性 P_k 的最终优先级。

2.2 数据集聚类

针对某次地震从互联网提取的数据集,按权重最高的属性字段进行聚类,从而形成若干小数据集,采用 LP 算法[17]聚类过程为:①将元素 d_1, d_2, …, d_n 分别插入初始为空的集合 $S_1, S_2, \cdots, S_n$,分别计算集合间彼此的相似度 sim;②当 sim ≥ γ(阈值),则合并为一个集合;③对未执行聚类的集合进行聚类,直至不存在未执行过聚类的单元素集合。

2.3 字段匹配

字段匹配算法用于计算不同记录同一属性字段之间的相似度。先把各属性中的字符串分解为多个具有独立意义的原子串,再分别比较不同记录同一属性的每个原子串的相似度,最后计算出不同记录的字段相似度。

在地震应急响应领域,把文本串和数字作为独立的原子串处理,如"死亡 25 人"。字段相似度的算法如式(4)所示。

$$\mathrm{SimField}(d_{ik}, d_{jk}) = \frac{\sum_{t=1}^{q}\max(\mathrm{score}(a, d_{jk}))}{|d_{ik}|} \tag{4}$$

式中:a 为记录 d_{ik}中的原子串,与 d_{jk}中的所有原

http://bhxb.buaa.edu.cn

子串匹配后，$score(a,d_{jk})$为计算所得相似度，且$0 \leqslant \text{score}(a,d_{jk}) \leqslant 1$；$|d_{ik}|$为记录$d_{ik}$的长度；$q$为记录$d_{ik}$的原子串个数。

2.4 清洗重复记录

对于不同组的重复记录，选择包含与主题词相关度高的属性值的记录，从中选取网页更新时间最近的予以保留，其余删除。生成完整的无重复的数据集。

灾情相关网页数据检测和清洗描述如下。

算法 2 EESRC(Earthquake Event Similarity Record Clearing)

输入：数据集 content，属性向量 $\boldsymbol{P}=(P_1,P_2,\cdots,P_k,\cdots,P_m)$。

输出：处理掉重复记录后的数据集。

$W_1,W_2,\cdots,W_m \leftarrow$ Weight($\boldsymbol{P}$)
$S_1,S_2,\cdots,S_n \leftarrow$ Clustering(content)
$\text{flag}_1,\text{flag}_2,\cdots,\text{flag}_n \leftarrow 0$
for each $S \leftarrow S_1,S_2,\cdots,S_n$ do
  选取排序关键字 key
  $S' \leftarrow$ sort (S,key)
  for each $p_i \in S'$, $1 \leqslant i \leqslant |S'|$ do
    for each $p_j \in S'$, $i \leqslant j \leqslant |S'|$ do
      if SimRecord(p_i,p_j) $>\gamma$ then
        if $\text{flag}_i=0$ and $\text{flag}_j=0$ then
          $\text{flag}_i \leftarrow i$; $\text{flag}_j \leftarrow i$
        else
          if $\text{flag}_i=0$ then $\text{flag}_i \leftarrow \text{flag}_j$
          else $\text{flag}_j \leftarrow \text{flag}_i$
          end if
        end if
      end if
    end for
    if 满足多趟查找结束条件 then return
    end if
  end for
  依据 $\text{flag}_1,\text{flag}_2,\cdots,\text{flag}_n$ 处理重复记录
end for

3 地震灾害数据时序统计

3.1 地震应急响应领域的词典定义

在地震应急响应领域，涉及3个方面的数据内容：灾情方面、灾区方面、处置需求。每一方面由多个相关关键词组成。本文中采用XML定义词典。

定义 5 词典。

```
<Dictionary>
  <Word type="灾情">
  <Word> <MainWord>死亡</MainWord>
  <SameWord>丧生</SameWord>
  </Word>
  …
  <Word>…</Word>
<Word type="灾区方面">
  <Word> <MainWord>山体滑坡</MainWord>
  </Word>
  …
</Dictionary>
```

其中：MainWord标签为主词；SameWord标签为同义词。把主词与同义词权重之和作为内容(content)中主题词的权重。

3.2 数据集中权重高的生词提取

在互联网信息提取过程中，把高权重的生词做为新的主题词处理。

处理的主要步骤为

1）把词典中的主词(mainword)添加到主题词集合中，计算各个主题词在内容(content)中的权重。

2）计算内容(content)中分词处理后的词项权重。

3）若其权重大于主题词权重平均值，则取出该词项作为生词加入到主题词集合中。

具体算法如下所示：

算法 3 HighWeightWordExtracting (Dictionary, content)

输入：词典 Dictionary，地震时间相关互联网信息提取内容 content。

输出：主题词集合 Word。

Word $\leftarrow \varnothing$
for each $\text{word}_i, 1 \leqslant i \leqslant N, \text{word}_i \in$ Dictionary do
  $\text{weight}_i \leftarrow$ TF(mainword) * IDF(mainword)
  $\text{weight}_i \leftarrow \text{weight}_i + \sum$ TF(sameword) * IDF(sameword)
  Word $\leftarrow$ Word $\cup \{\text{word}_i\}$
end for
$j \leftarrow 1$
for each w do
  weight $\leftarrow$ TF(w) * IDF(w)
  if weight > average($\text{weight}_1,\text{weight}_2,\cdots,\text{weight}_N$) then
    $\text{word}_{N+j} \leftarrow w$, $\text{weight}_{N+j} \leftarrow$ weight
    Word $\leftarrow$ Word $\cup \{\text{word}_{N+j}\}$
    $j \leftarrow j+1$
  end if

http://bhxb.buaa.edu.cn

15　end for

其中:TF(mainword)为 mainword 在地震事件相关提取内容 content 中出现的频率;IDF(mainword)为 content 中包含 mainword 的记录数目的反频率;TF(mainword) * IDF(mainword)计算主词 mainword 在 content 中的权重;∑TF(sameword) * IDF(sameword)计算该主词的所有同义词 sameword在 content 中的权重之和,将该主词与其所有同义词的权重相加作为该主题词在 content 中的权重。average($weight_1$, $weight_2$, …, $weight_N$)为主题词集合中属于词典中的主题词在 content 中的权重的平均值,N 为上述主题词的个数。

3.3　地震灾害主题词时序统计

使用二分法对地震灾害主题词进行时序统计,取数据集中网页更新时间最早的时间点与当前系统时间中间的时间段,平均切分成 2 个时间段,然后分别再次递归使用二分法针对各个小时间段内主题词进行时序统计,直至该时间段时长小于最小时间间隔或者该时间段内主题词的频数小于最小密度,则作为叶子节点结束切分。最终生成一棵以切分结果为节点的二叉树,并返回该二叉树的根节点。

在描述具体的算法步骤前,需要先定义作为时序统计切分结果的时序对象 TimeTrace,定义为七元组。

定义 6　时序对象 TimeTrace = (Word, PageSet, T_L, T_R, LeftSon, RightSon, WPAGE),对一个具体的时序对象 $timetrace_{t_l}^{t_r}$,可定义为 $timetrace_{t_l}^{t_r}$ = (word, $content_{t_l}^{t_r}$, t_l, t_r, leftson, rightson, wpage) ∈ TimeTrace。

其中:$timetrace_{t_l}^{t_r}$为一个具体的 t_l 到 t_r 时间段内的时序对象,$t_l \in T_L$,T_L 为统计时间段开始时间集合,$t_r \in T_R$,T_R 为统计时间段结束时间集合;word ∈ Word 为一个主题词,Word 为主题词典 Dictionary 中主题词集合;$content_{t_l}^{t_r} \subset$ PageSet 为网页更新时间在 t_l 到 t_r 时间段内的互联网信息提取内容集合,PageSet 为全部互联网信息提取内容的嵌套集合;leftson ∈ LeftSon 为该时序对象在时序统计中时间划分的左子节点,LeftSon 为左子节点集合;rightson ∈ RightSon 为该时序对象在时序统计中时间划分的右子节点,RightSon 为右子节点集合;wpage ∈ WPAGE 为该时间段内该主题词出现频率最高的代表网页对象,WPAGE 为互联网提取的网页集合。主要的算法步骤描述为

1)取起始时间和结束时间,若该时间段长小于最小时间间隔或者其主题词出现的频率小于最小密度,则取出该主题词出现频率最高的网页对象作为该时序对象的代表文本并返回。

2)否则,取从起始时间到结束时间这一时间段的中间时刻,把该时间段平均分为 2 段时间,并分别对 2 段时间递归进行该主题词的时序统计分析。

3)合并 2 个时间段的时序对象。

详细的算法步骤如下所示:

算法 4　TimeAnalysis(word, t_l, t_r, $content_{t_l}^{t_r}$)

输入:主题词 word,起始时间 t_l,结束时间 t_r,t_l 到 t_r 时间段之内的地震事件相关内容集合 $content_{t_l}^{t_r}$。

输出:该段时间内的时序对象 $timetrace_{t_l}^{t_r}$。

1　if　$t_r - t_l <$ 最小时间间隔 or count < 最小密度 then

2　　wpage←word　// 出现频率最高的网页对象

3　　$timetrace_{t_l}^{t_r}$←(word, $content_{t_l}^{t_r}$, t_l, t_r, null, null, wpage)

4　else

5　　$m \leftarrow (t_l + t_r)/2$

6　　$timetrace_{t_l}^{m}$←TimeAnalysis(word, t_l, m, $content_{t_l}^{m}$)

7　　$timetrace_{m}^{t_r}$←TimeAnalysis(word, m, t_r, $content_{m}^{t_r}$)

8　　$timetrace_{t_l}^{t_r}$←merge($timetrace_{t_l}^{m}$, $timetrace_{m}^{t_r}$)

9　end if

其中:merge($timetrace_{t_l}^{m}$, $timetrace_{m}^{t_r}$)将 2 个具体的时序对象 $timetrace_{t_l}^{m}$ 和 $timetrace_{m}^{t_r}$ 进行合并,生成的时序对象 $timetrace_{t_l}^{t_r}$为(word, $content_{t_l}^{m} \cup content_{m}^{t_r}$, t_l, t_r, leftson, rightson, wpage),leftson 指向 $timetrace_{t_l}^{m}$,rightson 指向 $timetrace_{m}^{t_r}$,wpage 为 word 在 t_l 到 t_r 时间段内出现频率最高的网页对象。

3.4　灾害信息时序统计算法描述

算法 5　EEDITS(Earthquake Event Disaster Information Timing Statistics)

输入:数据集 content。

输出:时序统计图。

1　定义主题词典 Dictionary

2　Word←HighWeightWordExtracting (Dictionary, content)

3　M←Word 中的主题词数目, starttime←after{地震发生时间 t_0, content 中最早的网页更新时间}, endtime←content 中最晚的网页更新时间

4　for each $word_i$, $1 \leqslant i \leqslant M$ do

5　　$timetrace_{t_l}^{t_r}$←TimeAnalysis($word_i$, starttime,

http://bhzb.buaa.edu.cn

endtime, content)

6 end for

7 生成时序统计图

其中:HighWeightWordExtracting (Dictionary, content)将词典中的主题词和 content 中权重大的生词组成统计的主题词集合,after{地震发生时间 t_0, content 中最早的网页更新时间}表示统计开始时间取 2 个时间中更新的时间,TimeAnalysis ($word_i$, starttime, endtime, content)针对于主题词在 starttime 到 endtime 时间段内的 content 集合进行时序统计。

4 Web 灾情数据摘要生成

对提取的网页摘要生成前首先需要进行预处理。首先,将 HTML 文档通过浏览器解析成 DOM 树,然后通过 HtmlParser 解析提取出网页正文,从 DOM 树中通过网页去噪得到摘要分析的正文,最后将正文划分成句子的组合。接下来采用摘要生成算法进行摘要生成。

多文档摘要生成算法用于地震相关文档生成的主要算法之一,使用 TextRank 算法提取出每篇文档的摘要句。先将文本句子拆分,过滤掉文档中"的、了"等常见的停用词,将此句子的集合按关键词所在的句子位置进行权重计算,一般网页段落的首尾句为关键句,加入摘要句集合,例如句子若包含地震地点名称,则说明该句子包含描述地震的重要信息,可以提取出该句子中的数字作为死亡人数、建筑物倒塌等数据,生成随时间发布的网页信息走势图,最终可以插入到文档中。对于一个文档 $\boldsymbol{F}$,可定义为一个 n 维向量。

定义 7 文档 $\boldsymbol{F}=\boldsymbol{F}(w_1,w_2,\cdots,w_i,\cdots,w_n)$。

其中:n 为文档内容的词语总数;w_i 为每一个词语的权重。

对于用户趣向内容的可定义为一个向量。

定义 8 用户趣向向量 $\boldsymbol{U}=\boldsymbol{U}((y_1,u_1),(y_2,u_2),\cdots,(y_i,u_i),\cdots,(y_n,u_n))$。

其中:y_i 为用户感兴趣的词语;u_i 为 y_i 的权重。

通过计算上述 2 个向量的夹角的余弦值,得出文档与用户趣向相似度。相似度越大,表示该文档满足用户的需求程度越大,相似度如式(5)所示。

$$\mathrm{sim}(\boldsymbol{F},\boldsymbol{U})=\frac{\boldsymbol{F}\times\boldsymbol{U}}{|\boldsymbol{F}||\boldsymbol{U}|} \tag{5}$$

多文档摘要生成是根据文档中的主题词,先用余弦相似度计算出哪些文档可以比较接近用户的需求。对于向量 $\boldsymbol{U}$ 用地震应急响应主题词,如"地震、死亡、伤亡、建筑物倒塌"等,加上地点,如"北川、雅安"等及相应权重表示用户兴趣度高的向量,计算出比较相似的文档集合。

如用 p_1 进行余弦度计算表示想找出"汶川地震"关键词比较相似的文档集合,p_2 进行余弦度计算表示想找出"北川建筑物倒塌"关键词比较相似的文档集合,p_3 进行余弦度计算表示想找出"捐助"关键词比较相似的文档集合,其中 $0\leqslant p_1, p_2, p_3\leqslant 1$,为余弦相似度值。根据用户对用户趣向向量 $\boldsymbol{U}$ 的不同定义,在网页信息处理时对数据库中关于该地震事件的搜索结果进行分类。

分类后得到 A 类、B 类等,如果对 A 类中包含较多主题词的文档,其中句子权重最高,则选择该网页为代表 A 类的文档,再对文档中的句子进行提取。

5 应用系统案例

针对笔者所在课题组研发的应用软件系统"国际强震应用处置系统之互联网信息智能处理子系统",应用上述方法进行了研发工作。其系统架构如图 1 所示。

互联网信息智能处理子系统主要包括搜索层、处理层、存储层、服务层等几个层次构成。其中,搜索层主要利用搜索引擎对专业地震信息网站、综合信息网站、核心网站实现互联网搜索、内容维护、采集源维护以及搜索模板管理等工作。然后提交处理层进行信息的后续处理工作;处理层主要实现对采集的网页内容进行信息提取、过滤、分拣以及统计等工作及主题字典管理。利用分类结果更新主题队列、有关词频统计以及提取生词,最终完成互联网信息的搜索与处理;存储层主要对样本分析模板库、本体词典库、信息库以及快照库实现统一的存储与访问管理;服务层面向用户及其他子系统提供信息访问服务,包括信息自定义查询、信息统计、话题追踪以及任务管理等几个功能。

5.1 信息提取算法收敛性实验验证

实验选取 2014 年 10 月 7 日 21 时 49 分云南景谷地震,对本节中提出的地震应急救援领域互联网信息提取算法的收敛性进行验证。选用中国地震局、美国地质调查局 (USGS)、凤凰网、人民网、新华网、新浪网、BBC 几个网站作为信息源,从地震发生时刻起,采用本文提出的互联网信息提取算法对云南景谷地震的信息进行提取。同

http://bhxb.buaa.edu.cn

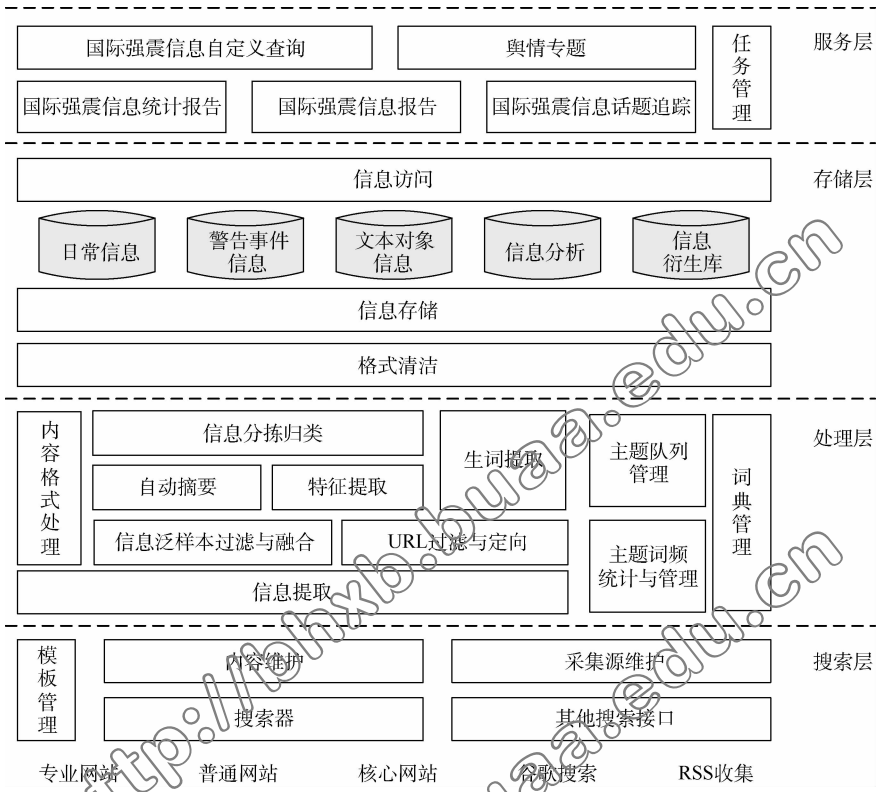


图 1　互联网信息智能处理子系统组织架构

Fig. 1　Architecture of subsystem of Internet information intelligent processing

时,为评估信息提取效果,在地震发生时刻起,在各个信息源中一直进行爬虫抓取,提取了 34 970 张网页作为实验样本。设定充分大的 N,取一周时间与信息源更新网页时间间隔的比值,经统计,相关信息网页更新时间间隔设为 5 min,则取 N 值为 2 016,地震发生一周的时间后计算相邻时刻信息提取内容的距离,当 N 为 2 541 时收敛,其 content 内容变化如图 2 所示。

从图 2 中的实验结果表明,地震应急响应领域互联网信息提取在大约地震发生一周之后逐渐开始收敛。地震应急响应领域专家主要包括科学研究专家、管理人员和现场救援人员。在进行震后应急响应的相关研究时,重点研究震后一周左右时间内的地震发展概况来辅助地震救援。现场救援人员在地震领域的黄金 72 h 之内争取尽量多地救出伤员,72 h 之后继续营救,在一周左右的时间之后现场救援工作也逐渐收尾。管理人员在地震发生之后制定救援方案,并在之后的一周左右的时间内,调整各项救援工作安排。实验结果与领域专家的经验正好吻合,验证了互联网信息提取过程的收敛性,即验证了其时效性。

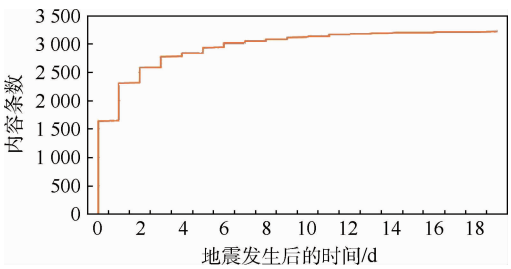


图 2　内容提取收敛性

Fig. 2　Content extraction convergence

5.2　相似重复记录清洗实验

查找相似重复记录算法的效性指标有查全率和查准率。X_a 为原数据集真实存在的重复记录集合,X_b 为识别出来的重复记录集合,则查准率表示为 $|X_a \cap X_b|/|X_b|$,查全率表示为 $|X_a \cap X_b|/|X_a|$。分别选取云南景谷地震、四川康定地震真实地震案例数据,对其内容进行插入、删除、替换和交换等操作,并按照 100%、200%、300% 的比例引入重复元素形成 3 个重复度逐渐增大的数据集,然后计算其平均查全率和查准率。本文提出的算法与近邻记录排序方法(Sorted Neighborhood Method,SNM)和多趟近邻排序算法(Muli-Pass sorted Neighborhood,MPN)[18-19] 算法相比较,在查全率、查准率上进行实验,针对不同冗余度的数据进行清洗的实验结果如表 1 所示。本文提出的相似重复记录清洗算法的查全率查准率保持在 95% 左右,表现较好。

5.3　灾害信息时序统计

选取云南景谷真实地震案例作为地震事件样

http://bhxb.buaa.edu.cn

本,选取其经过数据清洗之后的数据集作为本部分实验样本,为便于分析,词典中仅定义少量主题词,即泥石流、海啸 2 个主题词,提取出生词余震、消防、泥石流、救灾、灾区、救援、受伤、慈善、记者、得主、广播,最终的时序统计分析图如图 3 所示。

表 1 3 种算法查全率、查准率对比

Table 1 Comparison of recall ratio and precision ratio among three algorithms %

冗余度	查全率			查准率		
	SNM	MPN	本文	SNM	MPN	本文
100	80	90	94	83	88	93
200	79	91	96	88	89	95
300	76	90	95	90	90	95

5.4 摘要生成

选取 2015 年 4 月 25 日尼泊尔 8.1 级地震为例,系统生成的 Office Word 版的摘要文档如图 4 所示;相关舆情简报如图 5 所示。

对于面向地震应急响应的互联网信息智能处理,本文提出了一套完整的方法,并得出基于特

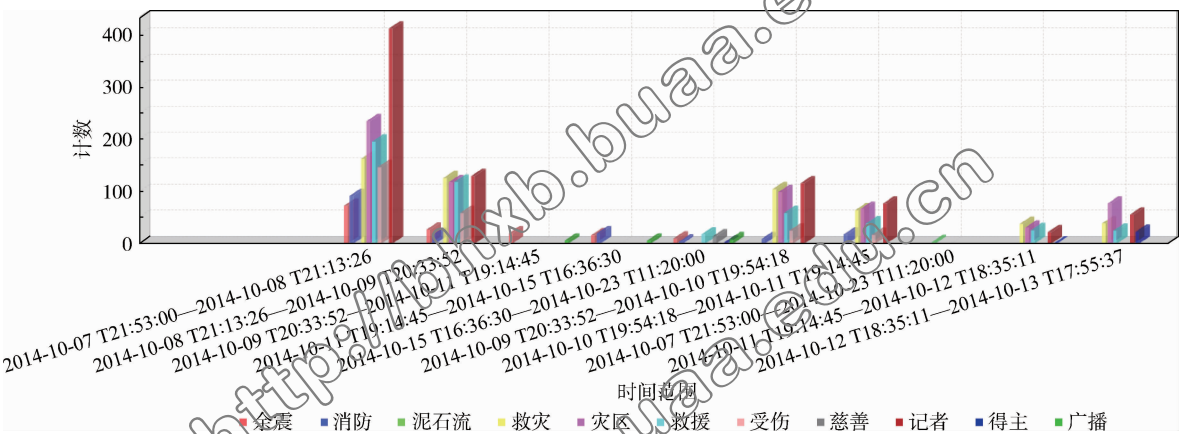


图 3 地震事件时序统计

Fig. 3 Time series statistics of earthquake events

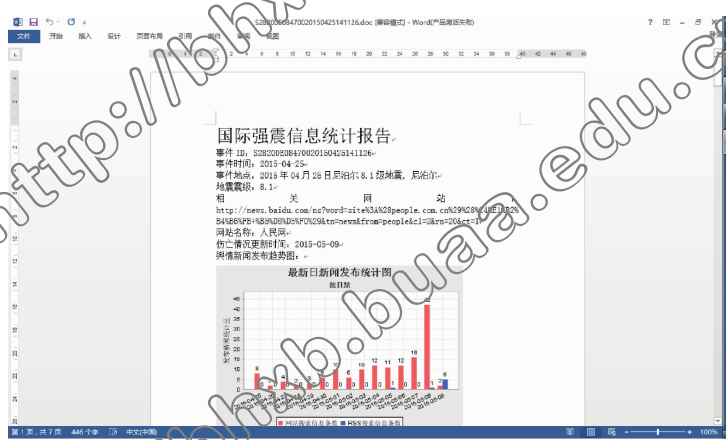


图 4 摘要文档

Fig. 4 Summary document



图 5 舆情简报

Fig. 5 Public opinion briefing

http://bhxb.buaa.edu.cn

定地震事件的摘要生成及舆情简报,为地震应急响应提供了决策支持的依据。

6　结　论

本文针对地震应急响应互联网信息智能处理展开了研究,研究结果如下:

1) 基于对地震应急救援领域应用需求与相关技术的全面调研和深入分析,首先定义了事件模型、网页对象模型,然后针对应急需求,在分析网页更新的规律和生命周期后,提出了互联网信息提取过程的收敛性定义,通过引入极限,定义收敛性来刻画互联网信息提取过程的时效性。

2) 分析了地震事件特点,提出了一种支持动态收敛性的互联网信息提取算法 EERWE,并对其中的关键步骤进行研究,将标准向量空间模型改进为多向量空间模型,并扩展定义了互联网信息时效性,使其具有动态性,能够准确评估具有时效性特点的互联网信息提取效果。

3) 分析了地震救援领域的网页信息较容易出现的数据质量问题,提出了网页相似重复记录的检测和清洗算法 EESRC,并进一步在清洗后的数据基础上,提出一种针对灾害信息随时间的变化进行时序统计的算法 EEDITS,形成信息统计报告。

4) 提出了面向地震应急响应的摘要生成和舆情简报算法,可以生成针对某次地震的摘要以及舆情简报,为制定救援方案决策提供依据。

最后,通过应用于国际强震应用处置系统之互联网信息智能处理子系统,为地震应急响应提供了保障。

参考文献 (References)

[1] 中国地震局. 历史地震目录[EB/OL]. 北京:中国地震局,2016(2016-06-14)[2016-09-10]. http://www.cea.gov.cn/publish/dizhenj/468/496/index.html.
China Earthquake Administration. The record of earthquake history [EB/OL]. Beijing: China Earthquake Administration, 2016(2016-06-14)[2016-09-10]. http://www.cea.gov.cn/publish/dizhenj/468/496/index.html(in Chinese).

[2] 新华网. 四川汶川地震抗震救灾进展情况[EB/OL]. 北京:新华网,2008(2008-06-22)[2016-09-10]. http://news.xinhuanet.com/newscenter/2008-06/22/content_8417853.htm.
Xinhuanet. Report on the earthquake rescue in Wenchan, Sichuan [EB/OL]. Beijing: Xinhuanet, 2008 (2008-06-22)[2016-09-10]. http://news.xinhuanet.com/newscenter/2008-06/22/content_8417853.htm(in Chinese).

[3] 赵亚辉. 汶川地震直接经济损失 8 451 亿元[EB/OL]. 北京:人民网,2008(2008-09-05)[2016-09-10]. http://society.people.com.cn/GB/41158/7805669.html.
ZHAO Y H. Direct economic losses of 845 billion 100 million yuan in Wenchuan earthquake[EB/OL]. Beijing: People, 2008 (2008-09-05)[2016-09-10]. http://society.people.com.cn/GB/41158/7805669.html(in Chinese).

[4] 霍娜. 突发事件追踪报道信息提取的研究[D]. 太原:山西大学, 2012:19-25.
HUO N. Research of sudden event information extraction of tracking reports[D]. Taiyuan: Shanxi University, 2012: 19-25 (in Chinese).

[5] HE J, GU Y Q, LIU H Y, et al. Scalable and noise tolerant web knowledge extraction for search task simplification[J]. Decision Support Systems, 2013, 56(5):156-167.

[6] SLEIMAN H A, CORCHUELO R. A class of neural-network-based transducers for web information extraction[J]. Neuro computing, 2013, 135(5):61-68.

[7] 侯明燕. 基于网页信息定位的数据抽取技术的研究[D]. 广州:暨南大学,2011:32-37.
HOU M Y. Data extraction technology research based on the location of Web information[D]. Guangzhou: Jinan University, 2011:32-37(in Chinese).

[8] AOL C, ZHANG P, CAO Y N. Estimating the locations of emergency events from Twitter streams[J]. Procedia Computer Science, 2014, 31:731-739.

[9] AEBI D, PERROCHON L. Towards improving data quality [C]// Proceedings of the International Conference on Information Systems and Management of Data. Delhi: Sarda, 1999: 273-281.

[10] INMON W H. DW2.0:下一代数据仓库的架构[M]. 王志海,王建林,译. 北京:机械工业出版社,2010:174-180.
INMON W H. DW2.0: The architecture for the next generation of data warehousing[M]. WANG Z H, WANG J L, translated. Beijing: China Machine Press, 2010:174-180(in Chinese).

[11] OUMSIYEH R, NG Y K. Enhancing web search by using query-based clusters and multi-document summaries[J]. Knowledge and Information Systems, 2016, 47(2):355-380.

[12] VALIZADEH M, BRAZDIL P. Exploring actor-object relationships for query-focused multi-document summarization[J]. Soft Computing, 2015, 19(11):3109-3121.

[13] ALGULIYEV R M, ALIGULIYEV R M, ISAZADE N R. An unsupervised approach to generating generic summaries of documents[J]. Applied Soft Computing, 2015, 34(9):236-250.

[14] KHAN A, SALIM N, KUMAR Y J. A framework for multi- document abstractive summarization based on semantic role labeling[J]. Applied Soft Computing, 2015, 30(5):737-747.

[15] 王振超,孙锐,姬东鸿. 基于事件指导的多文档生成式摘要方法[J]. 计算机应用研究,2016,34(2):343-346.
WANG Z C, SUN R, JI D H. Event-guided method for abstractive multi-document summarization[J]. Application Research of Computers, 2016, 34(2):343-346(in Chinese).

[16] DEY D, SARKAR S, DE P. A distance-based approach to entity reconciliation in heterogeneous databases[J]. IEEE Transactions on Knowledge and Data Engineering, 2002, 14(3): 567-582.

[17] 姚清耘,刘功申,李翔. 基于向量空间模型的文本聚类算法

http://bhxb.buaa.edu.cn

[J]. 计算机工程,2008,34(18):39-44.
YAO Q Y,LIU G S,LI X. VSM-based text clustering algorithm [J]. Computer Engineering, 2008, 34 (18): 39-44 (in Chinese).
[18] HERNÁNDEZ M A,STOLFO S J. Real-world data is dirty:Data cleansing and the merge/purge problem[J]. Data Mining and Knowledge Discovery,1998,2(1):9-37.
[19] HERNÁNDEZ M A,STOLFO S J. The merge/purge problem for large databases[C]// Proceedings of International Conference on Management of Data. New York: ACM SIGMOD, 1995: 127-138.

作者简介:

宋建功 男,博士研究生。主要研究方向:软件形式化方法与应用、互联网信息处理。

马世龙 男,博士,教授,博士生导师。主要研究方向:软件理论、大数据分析、云计算。

吕江花 女,博士,讲师。主要研究方向:计算机软件形式化、自动化测试、大数据分析。

Internet information processing for earthquake emergency response

SONG Jiangong[1], WANG Zhixin[2], LI Qinyong[1], LI Yating[3], MA Shilong[1], LYU Jianghua[1,*]

(1. School of Computer Science and Engineering, Beijing University of Aeronautics and Astronautics, Beijing 100083, China;
2. Institute of Software, Chinese Academy of Sciences, Beijing 100190, China;
3. Baidu Online Network Technology (Beijing) Co., Ltd., Beijing 100193, China)

Abstract: In recent years, the domestic and international earthquake disaster has caused huge losses to human life and property. Massive earthquake disaster information on the Internet can provide the basis for decision support for emergency response and timely rescue, so there is an urgent need for efficient and rapid processing of the disaster information. To study the information processing of the earthquake disaster in the Internet, the seismic event model, webpage object model, etc. were defined, and the convergence of Web information was defined with the limit method. The characteristics of the Internet disaster information dissemination were described. According to the characteristics of the timeliness of disaster information, the Web information extraction algorithm was proposed, which supports dynamic convergence. The method of time series statistics for the change of the disaster information with time was presented, and the information statistic report was formed, which provides the basis for rescue decision making. The intelligent information processing system for earthquake emergency response was designed and implemented. The models and methods were verified in a practical engineering project.

Keywords: earthquake disaster; emergency response; information extraction; data cleaning; multi-document summary

http://bhxb.buaa.edu.cn

Received: 2016-10-10; **Accepted**: 2017-01-09; **Published online**: 2017-02-24 09:31
URL: www.cnki.net/kcms/detail/11.2625.V.20170224.0931.001.html
Foundation items: National Natural Science Foundation of China (61300007, 61305054); Foundation of the State Key Lab of Software Development Environment (SKLSDE-2012ZX-28, SKLSDE-2014ZX-06)

* **Corresponding author**. E-mail: jhlv@nlsde.buaa.edu.cn