基于信任模式验证的论述性Web文本可信性判定方法

黄帅彪[1] 曾国荪[2] 王 伟[2]
(同济大学计算机科学与技术系 上海 201804)[1]
(国家高性能计算机工程技术中心同济分中心 上海 201804)[2]

摘 要 互联网中,海量Web信息自由开放,真假有益危害信息混在一起,许多Web文本的内容不可信。如何正确判断Web文本内容的可信性,从而在海量的Web信息中选择有用可信的资源信息,是当前急需解决的问题。提出了一种基于信任模式验证的论述性Web文本可信性判定方法,首先定义论述性Web文本的信任模式并对信任模式进行形式化的描述,然后构建论述性Web文本阅读自动机,最后进行模型检测以判断论述性Web文本是否满足信任模式,并根据检测结果求解Web文本可信度。实验表明该方法具有良好的效果。

关键词 论述性Web文本,信任模式,模型检测,文本可信度

Trust Pattern Verification Based Trustworthy Judgment Method of Explanatory Web Text

HUANG Shuai-biao[1] ZENG Guo-sun[2] WANG Wei[2]
(Department of Computer Science and Engineering, Tongji University, Shanghai 201804, China)[1]
(Tongji Branch, National Engineering & Technology Center of High Performance Computer, Shanghai 201804, China)[2]

Abstract There are much open Web information, which are composed of true and false information and useful and harmful information, and much information are untrustworthy. Therefore, it is more and more important how to judge the text credibility to choose the trustworthy and useful information from much Web information. This article introduced a kind of judgment method of Web text content trust based on trust pattern verification. First, the author defined the trust pattern of Web text and described trust pattern by the formal method, then the author modeled Web text, and finally checked the Web text model to judge whether Web verification model satisfies the trust pattern, and designed the algorithm of calculating Web text trustworthy degree according to the result of model checking. Experiment proves this method is effective.

Keywords Discoursive Web text, Trust pattern, Model checking, Web trustworthy degree

1 引言

互联网已经成为当今社会最重要的信息来源。然而在开放的互联网中,信息来源广泛、内容不一、质量良莠不齐等因素使得爆炸的信息资源越来越不可信。如何安全有效地利用网络信息资源,成为当前急需解决的问题。传统的信息安全研究,如认证和信誉[1],本质上是解决实体的信任问题,只能反映实体身份的合法与否,不足以保证互联网中信息交换的内容本身可信与否。因此,根据信息内容本身评估信息资源的信任度,即内容信任问题[2],已成为近年的研究热点。2004年美国Brigham Young大学的David Embley在研究异构信息系统的互操作时,提出了一种语义理解的本体信息抽取方法;2006年德国Bamberg大学的Claudia Hess, Klaus Stein和Christoph Schlieder[2]通过整合论文本身及参考文献的信任信息,开发了一个基于信任网络的论文推荐系统。同一年,日本的NICT(National Institute of Information and Communications Technology)启动了"信息可信评估"项目,主要是通过Web信息的发送者、Web文档的外观等方面来分析Web信息的可信性。这些研究虽然与内容信任有关,但都只是提及内容信任的某一方面。美国南加州大学的Gil[2]在2006年的国际互联网大会上,首次提出了Web资源的内容信任的概念,列举了19个影响内容信任度的因素,但这些信任因素较抽象,实现起来比较困难。

本文从论述性Web文本的结构和内容的规范性考虑,归纳反映文本质量好坏的信任模式,认为符合行文规范的论述性Web文本在很大程度上具有较好的信息质量,是可信的,反之则Web文本内容在很大程度上不可信;利用模型检测的方法,通过验证反映文本行文规范的信任模式,从而提出了一种基于信任模式验证的论述性Web文本可信判断方法。

2 问题描述

在目前的互联网中,大量的Web信息资源是以半结构化

到稿日期:2010-03-05 返修日期:2010-05-25 本文受863项目(2007AA01Z425,2009AA012201),973计划课题(2007CB316502),国家自然基金项目(90718015),NSFC-微软亚洲研究院联合项目(60970155),教育部博士点基金项目(20090072110035),上海市优秀学科带头人计划项目(10XD1404400),高效能服务器和存储技术国家重点实验室开放基金项目(2009HSSA06)资助。

黄帅彪(1987--),男,硕士生,主要研究方向为信息检索、内容信任,E-mail:huangshuaibiao123@126.com;曾国荪(1964--),男,博士,教授,博士生导师,主要研究方向为并行处理、可信计算。

文档的形式存在，网络的自由开放使得许多这样的半结构化文档在结构和内容上出现了各种形式的不规范。如许多论述性 Web 文档中，存在一些行文上的不规范问题，包括：

(1)概念术语在使用之前没有定义。

(2)主题只是在文档的开始部分出现，之后没有任何阐述。

(3)没有标题或标题出现在叙述部分的后面。

(4)定义中出现的概念在文中的其它部分没有相应的例子。

(5)核心内容中的定义在之后的部分没有用到。

(6)文档中没有出现对于部分内容或全文的总结。

这些行文上的不规范问题使得读者难以理解文档中的内容信息，从而使得文档信息质量低下、不可信。图 1 以有向图的形式描述一个半结构化文档[3]的基本结构，该文档来自维基百科[3]。

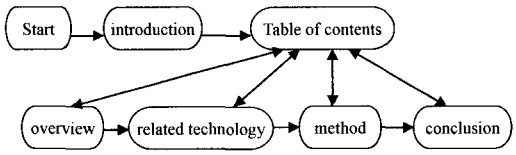

图 1　示例文档

其中顶点代表一 html 文档，有向边代表各个 html 文档之间的链接。该文档以 start 开始，紧接着是链接文档引言 introduction，然后是所有 html 文档链接的目录 Table of contents，以及该文档链接的各个 html 文档，首先是总揽 overview，其次是相关技术 related technology，然后是方法 method，最后是结论 conclusion。

该文档中包含的问题有：首先，文档中的总揽部分出现的主题概念在之后的部分都没有提及；其次，文档中出现的一些概念在之前没有相关定义；最后，在文档方法部分中，出现的定义在之后的部分没有用到，而且定义中出现的概念在之后也没有相应的例子加以阐述。这些行文上的不规范性在很大程度上影响该文档的信息质量，使得该 Web 文本的可信性低下。

3　论述性 Web 文本信任模式

在 Web 信息文本中，存在一些判断文本内容可信性的固有模式，例如可以根据电子邮件中的链接、图片等判断邮件的可信性；可以根据新闻报道的来源判断新闻的可信性；可以根据 Web 文本的外观形式判断其可信性等。我们从这些模式中归纳出信任模式的定义。

定义 1(信任模式)　所谓信任模式是指判断研究对象是否可信的标准或依据。

例如，我们可以根据文章作者的身份来判断该篇文章是否可信，这里的作者身份就是判断文章是否可信的一个信任模式。

定义 2(论述性 Web 文本信任模式)　所谓论述性 Web 文本信任模式是指对于论述性 Web 文本的结构和内容规范性的描述，其作用是判断文本的信息质量和可信度。

本文对论述性 Web 文本结构和内容的规范性进行了归纳和总结，提取了论述性 Web 文本的信任模式如下：

规范 1　专用概念和术语在使用之前一定要先定义或解释。

规范 2　关键词一定要出现在摘要正文里。

规范 3　主题在之后的部分一定要有所阐述。

规范 4　定义中出现的概念在文中的其它部分要有相应的例子。

规范 5　核心内容中的定义在之后的部分要有所阐述。

规范 6　标题一定要在叙述部分的前面。

规范 7　文中要出现对于部分内容或全文的总结。

规范 8　总结部分不应再出现新的概念、定义。

这些规范在很大程度上决定论述性 Web 文本的内容信息质量及可信性。例如，如果一篇论述性 Web 文本中出现的专用概念或术语在之前没有进行相关的定义或解释，当读者看到这些概念或术语时就很难理解其含义，这样就使得该文本内容的可信性低下；反之，假使该论述性 Web 文本满足以上的规范，那么读者就会因为该论述性 Web 文本在行文上是规范的而更倾向于认为该文本的内容是可信的。

4　基于 ALCCTL 的文本信任模式描述

4.1　ALCCTL 描述逻辑

ALCCTL[8]是结合了描述逻辑 ALC 和分支时序逻辑 CTL 的一种时序描述逻辑。ALC 能将一个领域中基于对象的知识表示进行形式化描述，以用其描述 Web 文本中的内容信息；CTL 在普通逻辑的基础上增加了表示时序的修饰符，用其描述文本内容块之间的时序关系。因此 ALCCTL 既可以描述 Web 文本内容块之间的时序关系，也可以描述 Web 文本中的内容信息。

ALCCTL 拥有命题连接词 $\rightarrow$，$\subseteq$，$\wedge$，$\vee$，$\forall$，$\exists$ 等，时态连接词 AG，AF，EF，EX，B，U 等。其中，AG 表示在所有路径上的所有状态；AF 表示在所有路径上都存在一个状态；EF 表示某条路径上存在一个状态；EX 表示在某条路径上的下一个状态，这些时态连接词后面可以跟一个 ALCCTL 概念或是一个 ALCCTL 公式；B，U 的前后一般都是 ALCCTL 概念或 ALCCTL 公式。$x\,B\,y$ 表示 x 在 y 之前出现，$x\,U\,y$ 表示在 y 出现之前一直是 x。如 $AG(\rightarrow termUdefition)$ 表示在所有路径上的所有状态都要满足专用术语在使用之前一定要先定义。

4.2　信任模式的 ALCCTL 描述

为了能对本文第 3 节中定义的信任模式进行验证，我们需要对信任模式进行形式化的描述。由于定义的信任模式不仅涉及到 Web 文本结构上的时序关系，还涉及 Web 文本的内容，因此我们用 ALCCTL 进行信任模式的描述。结合 ALCCTL 的描述能力，用 $majortopic$ 表示文本的主题；$Keywords$，$definition$，$title$，$term$，$majorparagraph$，$summary$ 分别表示文本中出现的关键词、定义、概念、标题、专用术语、核心段落、总结；$\exists addressedBy.X$ 表示会在 X 中有所阐述，其中 X 可以是摘要 $Abstract$、后面的叙述段落 $narrativeunit$ 等，如 $\exists addressedBy.Abstract$ 表示在文本的摘要中会有所阐述；$\exists followed.X$ 表示在之后会出现 X，如 $\exists followed.summary$ 表示之后要出现总结部分。用 ALCCTL 时序描述逻辑对第 3 节中定义的规范依次进行描述如下：

规范 1　$AG(\rightarrow termUdefition)$

规范 2　$Keywords \subseteq AF \exists addressedBy.Abstract$

规范 3　$majortopic \subseteq AF\exists addressedBy.narrativeunit$

规范 4　$AG(Definition \subseteq \forall defines.EX\exists illustratedBy.Example)$

规范 5　$majorparagraph.definition \subseteq EF\exists addressedBy.narrativeunit$

规范 6　$AG(titleBnarrativeunit)$

规范 7　$textUmajorparagraph \subseteq EF\exists followed.summary$

规范 8　$AG(summary \subseteq \neg\exists definition)$

5　论述性 Web 文本阅读自动机

5.1　论述性 Web 文本阅读自动机

为了对论述性 Web 文本进行信任模式的验证，需要对 Web 文本的结构和内容进行形式化描述。由于判断一篇文本是否满足某些行文规范是在读者阅读文本的过程中进行的，因此为了对论述性 Web 文本进行自动验证，以检测其是否满足信任模式，本文构建了论述性 Web 文本阅读自动机。

定义 3(论述性 Web 文本阅读自动机)　所谓论述性 Web 文本阅读自动机是指描述人们阅读论述性 Web 文本内容过程的一种特殊自动机，主要用于论述性 Web 文本信任模式的验证。

通常情况下，人们在阅读论述性 Web 文本时总是按照自上而下的顺序进行的。但在某些情况下，读者也会返回到先前读过的部分。如在读者阅读过程中，当遇到一些不常用的概念或术语时，读者就会很自然返回先前读过的部分，查找概念或术语的定义或解释；当文本中出现“正如某一节所述”这些陈述时，读者也会很自然地返回到所指示的部分进行阅读。因此，论述性 Web 文本阅读自动机的构建与文本中具体的概念术语等内容相关；而本文第 3 节定义的论述性 Web 文本信任模式也与文本中具体存在的主题概念等内容相关，故本文首先提取 Web 文本中的主题以及概念术语等语义信息，然后结合这些语义信息构建论述性 Web 文本阅读自动机。

5.2　论述性 Web 文本阅读自动机的构建

5.2.1　论述性 Web 文本中的相关内容提取与描述

由于本文第 3 节中定义的论述性 Web 文本信任模式涉及到文本内容的类型、主题、关键词，各个内容块的主题、概念、定义等，因此为了生成 Web 文本阅读自动机，以便进行信任模式的验证，我们首先利用 JTidy 将 HTML 文件转化为 XML 文件，然后结合相关本体库提取 Web 文本的元数据，包括 Web 文本内容的类型、主题、各个内容块的主题、关键词、概念、定义等。然后再利用 XQuery 程序[8]将 XML 代码转化为 RDF 元数据。图 2 是对本文文第 2 节中的示例文档 RDF 描述的语义信息。

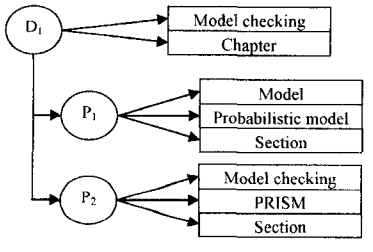

图 2　示例文档的 RDF 描述

提取的这些文本内容最终是要进行论述性 Web 信任模式的验证，因此需要将这些内容用描述逻辑进行描述。我们借助 XQuery 程序[8]将 RDF 元数据描述转化为描述逻辑的描述。对于图 2 中的 RDF 信息，经过 XQuery 转化为描述逻辑的描述如下：

```
D1:hasTopic={(D1,"Model cheking")}
front:hasTopic={P1,"model"}
      term={Probabilistic model}
back:hasTopic={P2,"Model checking"}
      term={PRISM}
```

5.2.2　论述性 Web 文本阅读自动机的构建

正如本文 5.1 节所述，正常阅读论述性 Web 文本的顺序是自上而下。当在一个段落中遇到专用概念术语或者是出现“正如某一节所述”这些陈述时，人们通常会返回到上面的段落重新阅读，以便更好地理解文本内容。因此，我们将 Web 文本的各个段落作为状态，人们阅读论述性 Web 文本过程中当前阅读段落的变化作为变迁，结合本文 5.2.1 节中用描述逻辑描述的各个段落的相关语义信息，构建论述性 Web 文本阅读自动机。具体的构建算法如下：

算法 1　阅读自动机生成算法 GARA(HTML)

```
输入:论述性 Web 文本
输出:论述性 Web 文本阅读自动机
GARA(HTML)
{
1 Stack L;String g←"as section stated";int flag[10];
2 L.push( sign "<p");
3 construct the state d;
4 get info by the method in secton 5.2.1;// info 表示 5.2.1 节中 ALCCTL 描述的文本主题、概念等信息
5 while(L! =null)
6   while(u! =/p>)
7       if(u="<p")
8         L.push( sign "<p");
9       if(u="/p>")
10          L.push( sign "/p>");
11          Construct the state s;
12          put info into s;
13    if(info contains concept || term || g)
14      flag(s)=1;
15    if("<p" and "/p>" are adjacent)
16      L.pop("<p");
17      L.pop("/p>");
18    Add a direct line between d and s;
19  d←s;
20 for(each s)
21  if(flag(s)=1)
22        Add a direct line between s and state(s.concept || s.term);
23    if(s.g=1)
24      Add a direct line between s and state(s.g);
}
```

第 1 行定义了一个栈和标记数组；第 2－3 行主要是当遇到一个段落标记时新建一个状态；第 4 行通过上述 5.2.1 节相关内容提取与描述得到语义信息集合 info；第 5－19 行遍历 HTML 文档，当遇到段落结束标记时将标记压入栈中，并新建一个状态，再将对应的 info 信息加入状态之中；当栈中

段落的开始标记和结束标记相邻时，将标记弹出栈，并在已建状态和新建状态之间添加一条有向边；第 20－24 行遍历每一个新建的状态，如果某个状态的 info 信息中包含专用概念或术语，则在该状态和 info 信息中包含专用术语对应定义的状态之间添加一条有向边；如果某个状态 info 信息中包含 g，则在该状态和 g 所指示的章节段落状态之间添加一条有向边。

通过算法 GARA(HTML)，就可以将论述性 Web 文本转化为论述性 Web 文本阅读自动机。图 3 是第 2 节中示例文档的部分阅读自动机，其中状态 S_0，S_1，S_2 表示论述性 Web 文本的各个段落；S_0，S_1，S_2 中需要验证的信息是用 ALCCTL 描述逻辑描述的，它们之间的变迁表示人们阅读该论述性 Web 文本中阅读状态的转变。

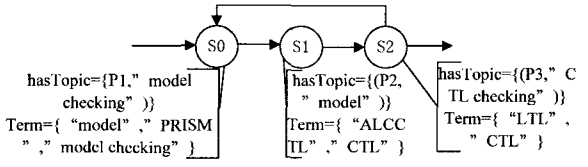

图 3　Web 文本阅读自动机示例

6　Web 文本信任模式验证及可信度计算

6.1　基于模型检测的信任模式验证算法

模型检测是一种基于算法的性质验证方法。该方法主要是对于表示系统行为的模型和表示系统性质(或规范)的某种时序逻辑公式，根据一定的模型检测算法，判定这个系统的模型是否满足定义的时序逻辑公式。常用的模型检测技术有基于证明的方法、形式化验证方法，其中形式化验证方法中的标记算法[10]适合自动机模型的时序逻辑公式的验证。

以本文第 3 节定义的 Web 文本信任模式和第 5 节构建的文本阅读自动机为基础，我们通过以下这个改进的标记算法对 Web 文本阅读自动机进行检测，以确定其是否满足本文 4.2 节中的 ALCCTL 公式。具体算法如下：

算法 2　改进的标记算法 Mark(M,ϕ)

输入：论述性 Web 文本阅读自动机 M 和 4.2 节中的 ALCCTL 公式集合 ϕ

输出：M 中满足 ϕ 中公式 φ 的状态集 S(φ)组成的集合{S(φ)}

```
Mark(M,φ)
{
1. for(each φ in φ)
2.   min←φ. minimal_subformula;
3.   while(each s in M)
4.     if(s. min=1)
5.         mark(s)← mark(s)+{min};
6.   while(ψ! =φ)
7.     while(each s in M)
8.         if(ψ∉mark(s))
9.           mark(s) ← mark(s)+{¬ψ};
10.         if(α∈mark(s) &&β∈mark(s))
11.           mark(s) ← mark(s)+{α∧β};
12.         if(ψ∈mark(s)) // mark(s)包含 ψ
13.           mark(s) ← mark(s)+{EF[ψ∨β],AFψ,EFψ};
14.           while(each p)
15.             if(ψ∈mark(p) && EF[ψ∨β]∈mark(p1))
16.               mark(p) ← mark(p)+ EF[ψ∨β];
17.               for(each p1 of p)
18.                 if(AF ψ∈mark(p1))t++;
19.               if(t=p. num)
20.                 mark(p) ← mark(p)+{AF ψ};
21.       if(ψ∈mark(q))
22.         mark(s) ← mark(s)+EX ψ;
23.   ψ←φ. next_subformula;
24.   for(each s in M)
25.   if(φ∈mark(s));
26.     S(φ)←S(φ)+{s};
27 {S(φ)}←{S(φ)}+S(φ);
}
```

第 1 行遍历公式集 ϕ 的每一个公式 φ，对 φ 进行 2－26 行的处理。第 2－5 行取公式 φ 的最小子公式 min，然后对于 M 的每个状态 s 是否满足 min 进行判断，如果满足则将公式 min 添加到状态 s 的标记集合中。第 6－23 行由小到大依次扩展 φ 的子式，直至 φ 结束，每扩展一次子式 ψ，对 M 中的每个状态 s 作如下处理：如果 s 的标记集合不包含 ψ，则将 $\neg\psi$ 添加到 s 的标记集合中，否则就将 EF[$\psi\vee\beta$]，AFψ，EFψ 都添加到 s 的标记集合中；如果 s 同时包含子公式 α 和 β，则将 $\alpha\wedge\beta$ 添加到状态 s 的标记集合中；对于 s 的所有前驱状态 p，如果满足公式 ψ 并且 p 的后继状态中存在一个状态满足公式 EF[$\psi\vee\beta$]，则将公式 EF[$\psi\vee\beta$]添加到状态 p 的标记集合中；如果 p 的所有后继状态都满足 AFψ，则将 AFψ 添加到 p 的标记集合中；如果 s 的一个后继状态 q 满足 ψ，则将 ψ 添加到 s 的标记集合中。第 24－27 行遍历 M 中所有的状态，如果状态 s 的标记集合中包含公式 φ，则将 s 添加到集合 $S(\varphi)$中，最后将每个公式 φ 对应的 $S(\varphi)$都添加到集合{$S(\varphi)$}中。

6.2　基于信任模式验证的 Web 文本可信度计算

本文第 4.2 节中定义的每个信任模式对 Web 文本的可信度影响程度是不同的。若存在两篇论述性 Web 文本 T_1，T_2，其中 T_1 不满足专用概念术语之前一定要定义这个信任模式，T_2 不满足文中出现的定义之后没有相应定义这个信任模式，那么 T_1 的可信度要低于 T_2 的可信度。因为相对 T_2 来说，T_1 由于专用术语没有定义致使文本的可理解性更加低下。因此，我们结合人们认识新事物的一般规律，通过对各个信任模式对于人们理解文本内容的影响程度进行客观分析，设定上述 4.2 节中 8 个信任模式对于文本可信度的影响权重的比例 $w_1:w_2:w_3:w_4:w_5:w_6:w_7:w_8$ 为 3:3:1:1:2:1:3:3。

然后，我们求解由单个信任模式影响而得的文本可信度。设 4.2 节中描述信任模式的 ALCCTL 公式集合为 $\phi=\{f_i\mid i=1,\cdots,T\}$，由 6.1 节中的算法 Mark(M,$\phi$)，可以得到 M 中分别满足$\{f_i\mid i=1,\cdots,N\}$的状态集合 $S(f_i)$。Web 文本阅读自动机中的状态代表 Web 文本的段落，设状态个数为 n，$S(f_i)$中的状态个数为 $m_i(m_i\leqslant n)$，则由信任模式 f_i 影响而得到的 Web 文本的可信度为$\frac{m_i}{n}$。

最后，我们结合之前设定的各个 f_i 对于文本可信度的影响权重，得到由 8 个信任模式共同影响所得的 Web 文本内容可信度计算公式如下：

$$trustdgree(web)=\frac{\sum_{i=1}^{T}w_i\times\frac{m_i}{n}}{N} \tag{1}$$

式中，$w_i(i=1,2,\cdots,T)$表示公式 f_i 代表的信任模式对 Web 文本可信度的影响权值。

(下转第 206 页)

[9] Tomita E, Kameda T. An efficient branch-and-bound algorithm for finding a maximum clique with computational experiments [J]. Journal of Global Optimization, 2007, 37: 95-111

[10] Pullan W, Hoos H H. Dynamic local search for the maximum clique problem[J]. Journal of Artificial Intelligence Research, 2006, 25: 159-185

[11] Pullan W. Phased local search for the maximum clique problem [J]. Journal of Combinatorial Optimization, 2006, 12: 303-323

[12] 吕强,柏战华,夏晓燕. 一种求解最大团问题的并行交叉熵算法[J]. 软件学报, 2008, 19(11): 2899-2907

[13] Ouyang Q, Kaplan P D, Liu S, et al. DNA solution of the maximal clique problem[J]. Science, 1997, 278: 446-449

[14] 李肯立,周旭,周舒婷. 一种改进的最大团问题 DNA 计算机算法[J]. 计算机学报, 2008, 31(12): 2173-2181

[15] Cui G, Li C, Li H, et al. Application of DNA self-assembly on maximum clique problem[C]// Yu W, Sanchez E N, eds. Advances in Computational Intell., AISC 61. Springer, 2009: 359-368

[16] Singh A, Gupta A K. A hybrid heuristic for the maximum clique problem[J]. Journal of Heuristics, 2006, 12: 5-22

[17] Geng X, Xua J, Xiao J, et al. A simple simulated annealing algorithm for the maximum clique problem[J]. Information Sciences, 2007, 177(22): 5064-5071

[18] Yang G, Yi J, Zhang Z, et al. A TCNN filter algorithm to maximum clique problem[J]. Neuro Computing, 2009, 72(4-6): 1312-1318

[19] Newman M E J. Finding community structure in networks using the eigenvectors of matrices[J]. Physical Review E, 2006, 74: 036104

[20] Palla G, Farkas I J, Pollner P, et al. Fundamental statistical features and self-similar properties of tagged networks[J]. New Journal of Physics, 2008, 10: 23-26

(上接第 180 页)

7 应用分析举例

我们利用上述方法对第 2 节中示例文本进行评估。首先提取文本中一些与信任模式相关的信息,包括主题、各个内容块中的定义、概念和术语、标题以及总结,再利用第 5.2.2 中的论述性 Web 文本阅读自动机构建算法 GARA(HTML),构建文本阅读自动机,生成的阅读自动机如图 4 所示。

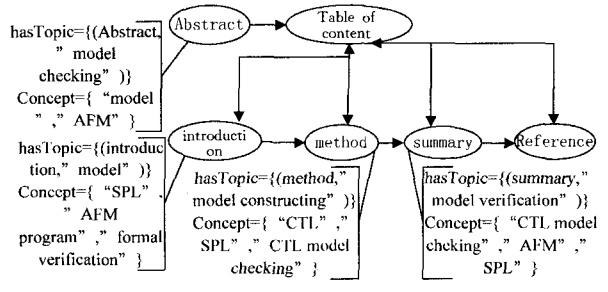

图 4 示例文本的阅读自动机

然后利用第 6.1 节中改进的标记算法 Mark(M, ϕ)对该文本验证模型进行检测,检测结果如表 1 所列。

表 1 示例文本的模型检测结果

序号	ALCCTL 描述的信任模式	文本是否满足
1	AG(→termUdefition)	仅在 Method 中满足
2	Keywords⊑AF∃ addressedBy. Abstract	否
3	majortopic⊑AF∃ addressedBy. narrativeunit	否
4	AG(Definition⊑∀ defines. EX ∃ illustratedBy. Example)	仅在 method 中满足
5	majorparagraph. definition⊑EF∃ addressedBy. narrativeunit	否
6	AG(titleBnarrativeunit)	仅在 summary 中不满足
7	textUmajorparagraph⊑EF ∃ followed. summary	是
8	AG(summary⊑→∃ definition)	是

该检测结果和实际情况是相符的。最后,利用式(1)计算该论述性 Web 文本的可信度。计算结果为文本可信度是 0.25,表明该文本由于行文不规范致使可信度低下,这与实际情况也是相符的。

结束语 本文提出了基于信任模式验证的 Web 文本内容信任判断这一新方法,其主要通过定义 Web 文本的信任模式并用形式化方法描述,然后对 Web 文本阅读自动机,最后利用模型检测算法对 Web 文本内容进行信任模式的验证,从而判断 Web 文本的可信性。这种方法可以在已有的模型检测技术基础上通过设计实现对 Web 文本的信任模式的验证,进而实现 Web 文本可信性的判定。实验证明这种方法是有效的。

在今后的工作中,我们将研究 Web 文本的元数据的自动提取及向描述逻辑的自动转化,并努力挖掘 Web 文本的其它信任模式,再结合这些信任模式对 Web 文本进行建模,然后进行模型检测,判断文本可信性,从而提高 Web 文本可信判断的准确性。

参 考 文 献

[1] Golbeck J, Hendler J. Inferring reputation on the semantic Web [C]// Proceedings of the 13th International World Wide Web Confe-rence. May 2004: 265-275

[2] Gil Y, Artz D. Towards content trust of Web resources[C]// Proceedings of the 15th International World Wide Web Conference. May 2006: 345-357

[3] wiki[EB/OL]. http://en.wikipedia.org/wiki

[4] Weitl F, Jaksic M, Freitag B. Towards the automated verification of semi-structured documents[J]. February 2009: 292-317

[5] 张东启,曾国荪,王伟. 基于信任事实的信任文本信任度评估方法[J]. 计算机科学, 2008, 35(8A): 202-205, 240

[6] Liu P, Chetal A. Trust-based Secure Information Sharing Between Federal Government Agencies[J]. J. of the American for Information Science and Technology, 2005, 56(3): 283-298

[7] 刘群,张华平,俞鸿魁,等. 基于层叠隐马模型的汉语词法分析[J]. 计算机研究与发展, 2004, 41(8): 1421-1429

[8] Weitl F. Document Verification with Temporal Description Logics[D]. Fakultat fur Informatik and Mathematic, University Passau, 2007: 114-145

[9] Schonberg C, Jaksic M, Weitl F, et al. Veirfication of Web-Content: A Case Study on Technical Documentation[C]// Proceedings of the 5th International Worskshop on Automated Specification and Verification of Web System(WWV09). Linz, Austria, February 2009