文章编号: 1671-9352(2010)05-0042-06

针对 Web论坛的一种结构化数据自动抽取方法

关冕, 马军
(山东大学计算机科学与技术学院, 山东 济南 250101)

摘要: 由于网页布局设计的复杂性和用户发表帖子的灵活性, 从论坛网页中抽取结构化的数据是一项未能很好解决并非常具有挑战性的任务。本文提出了一种从任意的论坛站点中自动抽取结构化数据的通用解决方案, 通过分析网页结构发现列表页和帖子页中的数据记录, 并利用一组产生式规则从发现的数据记录中抽取结构化的数据。实验结果表明该方法在抽取论坛数据记录方面明显优于已有的方法, 对论坛帖子的标题、作者、发表时间和内容文本块等元数据的抽取达到了较高的准确率。

关键词: 论坛; 结构化数据; 信息抽取; Web挖掘
中图分类号: TP391 文献标志码: A

Automatic structured data extraction from Web forums

GUAN Mian, MA Jun
(School of Computer Science and Technology, Shandong University, Jinan 250101, Shandong, China)

Abstract: Because of both complex page layout designs and unrestricted user created posts, extracting structured data from Web forum pages is a very challenging task and not easily solved. A general solution to automatically extract structured data from any forum site was proposed. By analyzing page structure, a group of data records were found from both list page and post page, and then a set of production rules was used to extract structured data from these data records. Experimental results showed that the proposed approach significantly outperformed some existing methods in extracting data records and achieved high accuracy in extracting some metadata of Web forums such as title, author, time and content.

Key words: Web forums; structured data; information extraction; Web mining

0 引言

目前论坛正成为网络上一个重要的数据源, 越来越多的研究工作利用从论坛数据中抽取的信息建立各种应用, 如提供问答服务[1]、获得商业智能[2]和发现专家网络[3]等。大部分应用都是首先从论坛网页中抽取结构化的数据, 再进一步利用这些数据实现各种功能。论坛的结构化数据抽取是对论坛中帖子的标题、作者、发表时间和内容文本块等关于论坛帖子元数据的抽取。它是处理论坛数据的基础, 也是我们在开发“驻济高校校园网信息搜集与智能分析系统”的过程中所遇到的必须解决的问题。然而由于不同的论坛站点通常使用不同的模板, 论坛中的帖子在布局、格式等方面存在很大差异。因此, 从任意的论坛站点中自动抽取结构化的数据是一个十分复杂的问题, 这一问题已经成为有效利用论坛数据的一个主要障碍, 目前似乎缺乏一种通用的 Web论坛结构化数据自动抽取的有效方法。

与 Web论坛结构化数据抽取最相关的工作是 Web信息抽取[4-12]。一般而言, Web信息抽取的方法可以分为两类: 依赖模板的方法[4-8]和不依赖模板的方法[9-12]。依赖模板的方法利用包装器(wrapper)作为一组由相同布局模板生成的网页的抽取

收稿日期: 2009-09-26
基金项目: 国家自然科学基金资助项目(60970047); 山东省自然科学基金资助项目(Y2008G19); 山东省科技攻关资助项目(2008GG10001026, 2007GG10001002)
作者简介: 关冕(1985-), 男, 硕士研究生, 主要研究方向为信息检索. Email: guanmian@hotmail.com

器, 通常关注于有限数量网站的信息抽取。大多数这些方法利用网页 DOM① 树的结构信息构造包装器, 构造方法包括手工构造、通过交互式学习半自动生成[7]和全自动发现[6]等。但是, 因为 DOM树的结构通常十分复杂, 构造健壮有效的包装器并不是一项简单的任务[4]。即使只关注有限数量的网站, 包装器的维护仍然是一个困难的问题[5]。对于论坛来说, 不同的论坛站点通常使用不同的模板。即使对于那些使用相同论坛软件构建的论坛, 如 "Discuz!" 和 "vBulletin"等, 它们仍然具有各种各样的定制化模板②。而且为了提供更好的用户体验, 大部分论坛站点周期性地更新它们的模板。如此多的论坛模板使构造和维护包装器的代价变得非常高, 因此依赖模板的方法不适合从论坛站点中抽取数据。

不依赖模板的方法为 Web信息抽取提供了一种不依赖于具体模板的更加通用的解决方案, 可用于处理具有不同布局特征的网页。其中一些方法基于概率模型, 如文献 [9] 利用关系马尔可夫网络从生物医学文本中抽取蛋白质的名字, 文献 [10] 利用条件随机场从纯文本的政府统计报告中抽取表格等。还有一些方法如 MDR(mining data records)算法是一种自动的通用方法[12], 首先利用网页的 HTML③ 标签对网页进行分割, 然后运用启发式知识从网页中识别数据记录。这些方法取得了较好的效果并具有一定的适应性。但是, 目前大多数不依赖模板的方法只利用单个页面内部的特征, 而对于论坛数据抽取, 只利用单个页面的特征不能够处理复杂的网页布局设计和灵活多变的帖子。

本文试图给出一种尽可能通用的解决方案实现从任意的论坛站点中自动抽取结构化的数据。论坛站点不同于一般站点, 它具有自己独特的结构。文中通过分析论坛中列表页和帖子页的网页结构发现其中的数据记录, 并利用一组产生式规则从识别的数据记录中抽取论坛帖子的元数据。在 20个论坛上的实验结果验证了本文提出方法的有效性。

1　基本术语

站点图 (sitemap)　站点图是一个有向图, 图中每个顶点代表一类具有相同布局结构的网页, 每条边代表两个顶点之间一种特定的链接关系[13]。本文利用文献 [14] 中构造站点图的方法构造论坛的站点图, 通过删除无用的顶点和边, 论坛的站点图可以简化为树的形式。

列表页 (list page)　论坛站点图内部结点中的网页称为列表页。列表页包含了论坛中一组帖子的导航信息。

帖子页 (post page)　论坛站点图叶子结点中的网页称为帖子页。帖子页包含了论坛中某一特定帖子的详细信息。

数据记录 (data record)　数据库中的记录以规则的模式在网页上形成的结构化数据对象称为数据记录。数据记录是网络上一种非常重要的信息类型。

图 1为济南大学 (bbs.ujn.edu.cn)论坛中列表页和帖子页的一个示例。图中左侧的列表页包含了 20个数据记录, 右侧的帖子页包含了 3个数据记录。

状态	主题	作者	回复/人气	最后更新/回复人
	[置顶]一句话手册 [D1234...]	君忘忧	142/10579	2009-09-16 21:30 | 五月雪
	[置顶]毕业了 [D1234...]	风一样的男子	559/67070	2009-09-12 21:33 | allkf
	[置顶]我们的似水流年 [D1234...]	丰主流	380/46498	2009-06-21 15:34 | 可爱水晶
	你敢回答校上的问题吗？？？ [D1234...]	生而为意	128/5260	2009-09-19 10:36 | 五月的雨
	[投票]单纯向左，济大向右(完整整理版) [D1234...]	zhangqinujn	62/6891	2009-09-18 20:07 | langzinhai
	每人说一件在济大最喜欢做的事 [D1234...]	到痛的快感	181/8270	2009-09-18 19:54 | langzinhai
	[投票]毕业后，有多少人会怀念曾经的大学生活？ [D12]	天亮就走	25/1040	2009-09-18 19:50 | langzinhai
	谈谈我毕业三年的人生所遇 [D123]	nan3113	33/1295	2009-09-18 19:48 | langzinhai
	说一个济南大学最让你怀念的地方 [D12]	云冰冰	19/560	2009-09-18 19:41 | langzinhai
	梁山人在济大	水泊猴	0/8	2009-09-18 17:48 | 水泊猴
	[原创] 活着的尸体..... [D1234...]	allkf	86/3161	2009-09-18 16:20 | 丰主流
	在鲁闽南人的群	stsimen	0/7	2009-09-18 15:11 | stsimen
	回来看看那 一瞬间	梦知边缘	1/26	2009-09-18 13:17 | 羽叶草
	人活着为了什么？ [D1234]	assimilate	54/2198	2009-09-18 13:16 | 羽叶草
	我知道，有些人，有些事，我再也不能回头。	lotus887	2/81	2009-09-12 21:37 | allkf
	寻找记忆，寻找望冰	行风	6/290	2009-09-12 21:13 | allkf
	怀念济大	心云	0/26	2009-09-10 10:22 | 心云
	林A降红网络为读者"洗心改命"	achao0539	0/19	2009-09-10 00:24 | achao0539
	5月份坐过201路公交车的女生一定要看啊！	kingtiger	2/186	2009-09-08 12:45 | shivuye
	突然想想，一年没来了。	风一样的男子	0/32	2009-09-07 12:29 | 风一样的男子

图 1　济南大学论坛中的列表页和帖子页
Fig. 1　List page and post page in bbs.ujn.edu.cn

① http://zh.wikipedia.org/wiki/DOM
② http://www.vbulletin.faq.com/skins.styles.htm
③ http://zh.wikipedia.org/wiki/HTML

2 论坛数据记录抽取

基于对目前 Web论坛数据的调研, 发现绝大部分论坛数据满足如下两点假设:

(1)一组数据记录通常出现在网页的一个连续区域中并由相似的 HTML标签生成。这样的区域称为数据记录区(简称数据区)。如果把网页的 HTML标签看作一个长字符串, 可以使用字符串匹配的方法比较不同的子串, 从而发现那些可能代表相似数据记录的相似子串。但是, 由于一个数据记录可以开始、结束于任何标签, 这种方法的计算量十分巨大, 下一假设有助于解决这一问题。

(2)网页中 HTML标签的嵌套结构自然地形成了一个标签树[15]。一组相似的数据记录对应于标签树中同一父结点的一些子树。一个数据记录不会开始于一个子树的中间, 结束于另一子树的中间。相反, 它通常开始于一个子树的开始, 结束于这一子树的结束。基于这一假设, 可以设计出一个非常有效的算法来识别数据记录, 因为它限制了一个数据记录在标签树中开始和结束的标签位置。

在假设 Web论坛数据满足上述条件时, 提出如下一种论坛数据记录抽取算法 FDRE(forum data records extraction)。

论坛数据记录抽取算法 FDRE

输入: 要处理的论坛网页 P。

输出: P中的数据记录。

步骤:

(1)建立网页 P的 HTML标签树;

(2)使用标签树和字符串比较挖掘网页 P中的数据区;

(3)从步骤 2得到的数据区中识别数据记录。

以下给出算法每一步的具体实现方法。

步骤 **1** 建立网页的 HTML标签树。

网页中的大部分 HTML标签都是成对出现的。每对 HTML标签由一个开始标签和一个结束标签构成, 它们之间可以包含其他 HTML标签对, 从而形成了网页 HTML代码的嵌套结构。使用网页的 HTML代码建立网页的标签树, 树中每个结点对应一对 HTML标签。图 2为图 1中列表页对应的标签树。

步骤 **2** 使用标签树和字符串比较挖掘网页中的数据区。

通过比较标签树中结点(包括它们的子孙结点)的标签字符串可以发现网页中的数据区。使用的字符串比较算法基于标准编辑距离[16]。假设要进行比较的两个字符串分别为 s_1和 s_2, 算法的时间复杂性为 $O(|s_1||s_2|)$[16]。在实际应用中, 由于只关注非常相似的字符串, 如果两个字符串明显不同(如一个字符串的长度大于另一个长度的 2倍), 则这两个字符串无需比较, 因此算法的时间复杂性远小于 $O(|s_1||s_2|)$。以图 2中的标签树为例: 基于结点标签字符串的编辑距离(本算法中编辑距离的阈值取 0.3), 树中最下方 table结点的所有孩子结点 tr都相似, 它们形成了一个数据区。识别数据区的字符串比较的数量并不是很大, 只需要在一个父结点的孩子结点之间进行比较。

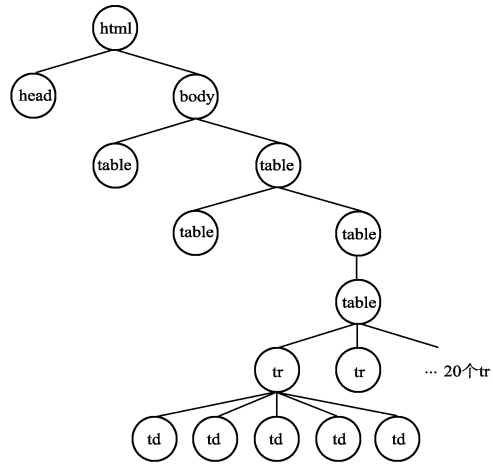

图 2 图 1中列表页对应的标签树

Fig. 2 Tag tree of the list page in Figure 1

步骤 **3** 从数据区中识别数据记录。

在找到数据区之后, 从数据区中识别数据记录。将数据区中的每个元素即每个相似的结点(包括它们的子孙结点)作为一个数据记录。图 2中的每个 tr结点都对应一个数据记录。对于论坛中的帖子页, 如果一个帖子页只包含一个帖子, 那么该帖子页中不存在数据区, 步骤 2无法产生输出。这时利用论坛的站点图信息找到与这一帖子页位于同一顶点的其他某一帖子页中的第一个数据记录, 并利用它在对应帖子页中的 DOM树路径和它自身的 DOM树结构识别数据记录。

3 基于 FDRE的论坛结构化数据抽取

基于 FDRE算法, 提出了一组产生式规则抽取论坛帖子中的元数据。FDRE抽取的每个数据记录都对应于论坛网页 DOM树中的一个子树。该树中的结点根据它们的属性可以分为 3类: 文本结点(t)、超链接结点(h)和内部结点(i)。以帖子标题、

作者、时间和内容的抽取为例, 给出相应的产生式规则如下:

$$\forall i,h\ IsDataRecord(i) \wedge HasPostLink(i,h) \Rightarrow IsTitleNode(h) \quad (1)$$

$$\forall i,h\ IsDataRecord(i) \wedge HasAuthorLink(i,h) \Rightarrow IsAuthorNode(h) \quad (2)$$

$$\forall t\ IsTimeFormat(t) \wedge UnderSameOrder(t) \Rightarrow IsTimeNode(t) \quad (3)$$

$$\forall i,i'\ IsDataRecord(i) \wedge HasDescendant(i,i') \wedge ContainTextNodes(i') \Rightarrow IsContentNode(i'). \quad (4)$$

产生式规则中的谓词及相关说明见表 1。

表 1　产生式规则中的谓词描述

Table 1　Descriptions of predicates in the production rules

谓词	描述
IsDataRecord(i)	内部结点 i是一个数据记录
IsTimeFormat(t)	文本结点 t包含时间格式的字符串
HasPostLink(i,h)	内部结点 i包含指向帖子页的超链接结点 h
HasAuthorLink(i,h)	内部结点 i包含指向用户信息页的超链接结点 h
HasDescendant(i,i′)	内部结点 i是内部结点 i的一个后裔结点
UnderSameOrder(t)	文本结点 t形成的时间集合以升序或降序排列
ContainTextNodes(i′)	内部结点 i′包含多个文本结点
IsTitleNode(h)	超链接结点 h是标题结点
IsAuthorNode(h)	超链接结点 h是作者结点
IsTimeNode(t)	文本结点 t是时间结点
IsContentNode(i′)	内部结点 i′是内容结点

以下是对上述产生式规则的自然语言解释。

规则 (1): 帖子标题的抽取

从列表页的每个数据记录中抽取帖子的标题。对于每个数据记录, 利用论坛的站点图信息找到其中的 postlink(指向帖子页的链接), 其锚文本就是该数据记录对应帖子的标题。

规则 (2): 帖子作者的抽取

从帖子页的每个数据记录中抽取帖子的作者。对于每个数据记录, 利用论坛的站点图信息找到其中的 authorlink(指向用户信息页的链接), 其锚文本就是该数据记录对应帖子的作者。

规则 (3): 帖子时间的抽取

从帖子页的每个数据记录中抽取帖子的时间。对于每个数据记录, 如果其中某个文本结点包含时间格式的字符串, 并且这些字符串形成的集合以升序或降序排列, 那么这些时间格式的字符串就是其对应数据记录的帖子时间。

规则 (4): 帖子内容的抽取

从帖子页的每个数据记录中抽取帖子的内容。帖子内容包含在每个数据记录的一个内部结点中, 以该内部结点为根的 DOM树包含多个文本结点。

4　实验

4.1　数据集和评价方法

不同的论坛通常具有不同的布局设计。本文选择了 20个不同的论坛站点, 它们分别从属于 4个类别: 高校、书籍、数码产品和工作, 如表 2所示。其中"高校"包含了 12个论坛站点, 它们来自于我们研发的系统"驻济高校校园网信息搜集与智能分析系统"。该系统通过对 40多个高校站点的信息搜集与分析, 提供信息检索、信息分类、信息简报和热点信息等服务。另外, 为了评价本文的方法在各种情况下的性能, 又从"书籍"、"数码产品"、"工作"3个类别中选择了其余的 8个论坛站点。一些论坛站点, 例如"bbs.ujn.edu.cn"和"www.sdadz.edu.cn/bbs"是由流行的论坛软件"vBulletin"生成的; "bbs.iepub.net"和"bbs.imobile.com.cn"是由另一个工具"Discuz!"生成的; 还有一些论坛是定制的。显然这些论坛中的网页在布局设计上有很大不同, 可以用于评价本文方法的一般化能力。

表 2　论坛站点

Table 2　Web forum sites

Id	论坛站点	描述	Id	论坛站点	描述
1	bbs.sdu.edu.cn	山东大学	11	www.52szy.net	山东中医药大学
2	bbs.ujn.edu.cn	济南大学	12	guofeng9day.org	山东体育学院
3	bbs.sdjzu.edu.cn	山东建筑大学	13	bbs.iepub.net	电子图书
4	www.sdadz.edu.cn/bbs	山东工艺美术学院	14	www.chinesepdf.com	中文 PDF读书
5	www.sdfibbs.net/bbs	山东财政学院	15	bbs.imobile.com.cn	手机
6	210.44.144.24/bbs	山东轻工业学院	16	nbbbs.zol.com.cn	笔记本电脑
7	www.shanyibbs.com	山东艺术学院	17	www.qqdc.com.cn	数码相机
8	www.shanshiren.com	山东师范大学	18	bbs.jobuu.com	工作论坛
9	www.sdufe.com	山东经济学院	19	bbs.ojob.com.cn	赢才论坛
10	www.jyday.cn	山东交通学院	20	bbs.yingjiesheng.com	应届生 BBS

对表 2中的 20个论坛站点,从每个站点中随机选择 50个列表页和 50个帖子页作为实验的数据集。对于每个站点,手工构造一个包装器来抽取所有的目标数据作为实验的基准。在实验中,采用精确率(precision)、召回率(recall)和它们的调和平均值(F1)作为衡量标准,它们定义为:precision= Ec/E,recall= Ec/N,F1=(2×precision×recall)/(precision+recall)。其中 Ec为正确抽取的目标数据的个数,E为抽取的目标数据的总数,N为列表页或帖子页中目标数据的个数。

4.2 实验结果

4.2.1 与 MDR的比较

实验中使用的 MDR程序由其作者提供。MDR算法存在一个相似性阈值,根据其作者的建议将其设置为 60%。实验结果如表 3所示。表 3的数据结果反映出,相对于 MDR算法,FDRE算法的精确率提高了 1.2%,召回率提高了 13.1%,F1值提高了 7.6%。导致 MDR算法召回率较低的原因为:论坛中存在大量的只包含一个数据记录的帖子页,即这种帖子页中只包含一个帖子,该帖子在帖子作者发表以后还没有任何回复。MDR方法要求网页中存在 2个或更多的数据记录,因此无法处理帖子页只包含一个帖子的情况。

表 3 FDRE与 MDR的比较结果
Table 3 Comparison results with MDR

算法	Precision	Recall	F1
MDR	0.965	0.842	0.899
FDRE	0.977	0.973	0.975

4.2.2 论坛结构化数据抽取

以下分别考察了本文算法对不同元数据的抽取能力。实验结果如表 4所示。表 4的数据结果反映出,本文的算法在表 2中的 20个论坛站点上取得了较高的精确率、召回率和 F1值。应当指出的是:帖子标题的精确率、召回率和 F1值均好于帖子作者、帖子时间和帖子内容。这是因为论坛站点中绝大部分列表页的数据区是由 HTML标签“table”生成的。相对于布局设计复杂的帖子页来说,FDRE方法更擅长从“table”标签中抽取数据记录。

表 4 对表 2中 20个论坛站点的实验结果
Table 4 Experimental results on 20 web forum sites in Table 2

Label	Precision	Recall	F1
Title	0.980	0.983	0.981
Author	0.925	0.901	0.913
Time	0.899	0.878	0.888
Content	0.894	0.872	0.883

5 结论

讨论了如何从论坛站点中自动抽取结构化数据的问题。实验表明,本文提出的论坛数据记录抽取算法 FDRE明显改善了已知算法 MDR(召回率提高了 13.1%,F1值提高了 7.6%)。基于 FDRE算法,又提出了一组产生式规则抽取论坛帖子中的元数据。在 20个论坛上的实验结果取得了较高的精确率、召回率和它们的调和平均值,说明了本文的方法处理不同论坛的一般化能力。然而由于论坛中的帖子在布局、格式等方面存在很大差异,对帖子时间、帖子内容等元数据的抽取还有待改善,这也是进一步的研究方向。

参考文献:

[1] CONG G, WANG L, LIN CY, et al. Finding question-answer pairs from online forums[C]//Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM Press, 2008: 467-474.

[2] GLANCE N, HURST M, NIGAM K, et al. Deriving marketing intelligence from online discussion[C]//Proceedings of the 11th Annual International ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York, USA: ACM Press, 2005: 419-428.

[3] ZHANG J, ACKERMAN MS, ADAMIC L. Expertise networks in online communities: structure and algorithms[C]//Proceedings of the 16th International Conference on World Wide Web. New York, USA: ACM Press, 2007: 221-230.

[4] KUSHMERICK N. Wrapper induction: efficiency and expressiveness[J]. Artificial Intelligence, 2000, 118: 15-68.

[5] LERMAN K, MINTON S, KNOBLOCK C. Wrapper maintenance: a machine learning approach[J]. Journal of Artificial Intelligence Research, 2003, 18: 149-181.

[6] ZHAI Y, LIU B. Web data extraction based on partial tree alignment[C]//Proceedings of the 14th International Conference on World Wide Web. New York, USA: ACM Press, 2005: 76-85.

[7] ZHENG S, WU D, SONG R, WEN JR. Joint optimization of wrapper generation and template detection[C]//Proceedings of the 13th Annual International ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York, USA: ACM Press, 2007: 894-902.

[8] 杨少华, 林海略, 韩燕波. 针对模板生成网页的一种数据自动抽取方法[J]. 软件学报, 2008, 19(2): 209-

223.

[9] BUNESCU R, MOONEY R J. Collective information extraction with relational Markov networks[C]//Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics. San Francisco, USA: Morgan Kaufmann Publishers, 2004: 439-446.

[10] PINTO D, MCCALLUM A, WEI X, et al. Table extraction using conditional random fields[C]//Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM Press, 2003: 235-242.

[11] 胡仁龙，袁春风，武落山，等. 基于重复模式的自动 Web信息抽取[J]. 计算机工程，2008，34(22)：73-76.

[12] LIU B, GROSSMAN R, ZHAI Y. Mining data records from Web pages[C]//Proceedings of the 9th Annual International ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York, USA: ACM Press, 2003: 601-606.

[13] YANG J M, CAI R, WANG Y, et al. Incorporating site-level knowledge to extract structured data from Web forums[C]//Proceedings of the 18th International Conference on World Wide Web. New York, USA: ACM Press, 2009: 181-190.

[14] CAI R, YANG J M, LAI W, et al. iRobot: An intelligent crawler for Web forums[C]//Proceedings of the 17th International Conference on World Wide Web. New York, USA: ACM Press, 2008: 447-456.

[15] CHAKRABARTI S. Mining the Web: discovering knowledge from hypertext data[M]. San Francisco, USA: Morgan Kauffman Publishers, 2002: 228-235.

[16] BAEZA-YATES R. Algorithms for string matching: a survey[J]. ACM SIGIR Forum, 1989, 23(3-4): 24-58.

(编辑：许力琴)

(上接第 41页)

[7] CHU C H, GU J H, HOU X D, et al. A heuristic ant algorithm for solving QoS multicast routing problem[C]//IEEE Proceedings of Evolutionary Computation. CEC'02. Piscataway(N.J. USA): IEEE, 2002, 2: 1630-1635.

[8] 石钊，葛连升. 一种解多 QoS约束组播问题的改进蚁群算法[J]. 山东大学学报：理学版，2007，42(9)：41-45.

[9] 杨云，徐佳，高飞，等. 基于蚁群系统的多 QoS约束组播路由算法[J]. 小型微型计算机系统，2006，27(11)：2031-2035.

[10] HUANG Lin, HAN Haishan, HOU Jian. Multicast routing based on the ant system[J]. Applied Mathematical Science, 2007, 1(57): 2827-2838.

[11] 王兴伟，邹荣珠，黄敏. 基于蚂蚁算法的 ABC支持型 QoS组播路由机制[J]. 东北大学学报：自然科学版，2009，30(7)：959-963.

[12] RUBINSTEIN R Y. The cross-entropy method for combinatorial and continuous optimization[J]. Methodology and Computing in Applied Probability, 1999, 2: 127-190.

[13] SCHOONDERWOERD R, BRUTEN J, HOLLAND O, et al. Ant-based load balancing in telecommunications networks[J]. Adaptive Behavior, 1996, 5(2): 169-207.

[14] HELVIK B E, WITTNER O. Using the cross entropy method to guide/govern mobile agent's path finding in networks[C]//Proceedings of 3rd International Workshop on Mobile Agents for Telecommunication Applications. Heidelberg: Springer-Verlag, 2001: 255-268.

[15] NS-2[EB/OL]. [2010-01-25]. http://www.isi.edu/nsnam/ns.

[16] JAMIN S, WINICK J. Inet-3.0: internet topology generator[R]. Ann Arbor: University of Michigan, 2002.

(编辑：许力琴)