

B2C 电子商务中的信息抽取技术^{*})

Information Extraction Techniques Applied to B2C E-Commerce

于 琨¹ 蔡 智² 糜仲春¹ 蔡庆生²

(中国科学技术大学信息管理与决策科学系¹ 计算机科学技术系² 合肥230026)

Abstract After stepping out the valley of net economy, B2C e-commerce is about to come into a climax of development. Information extraction techniques are going to be one of the most important factors to promote B2C e-commerce. In this paper, we present a review on the recent progress of information extraction techniques applied to B2C e-commerce. The characteristics of each technique are also appraised.

Keywords Information extraction, B2C e-commerce, Internet

1 引言

B2C(商业机构对消费者)电子商务主要指借助于 Internet 开展的在线销售活动^[1]。eMarketer 于2001年发布的对于2000年至2004年全球 B2C 电子商务营业额的统计预测数据(图1)^[2]表明,在未来的三年中 B2C 电子商务将进入一个快速发展的时期。

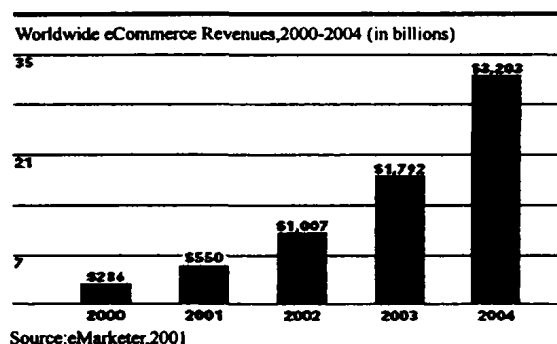


图1 2000-2004年全球 B2C 电子商务营业额

1999年于美国公布的第一个 Internet 商务标准明确规定从事 B2C 电子商务的销售商应该提供对整个站点的信息和产品进行检索的能力^[3]。而在 eMarketer 于2001年统计的用户不参与网络购物的原因(图2)^[2]中,难以获取商品信息位居第五,成为除了服务、隐私、价格和安全之外阻碍消费者参与 B2C 电子商务的主要因素之一。因此,随着 B2C 电子商务的迅速发展,能否从浩如烟海的电子商务网站中快速、准确地获取有用的商品信息不仅是 B2C 电子商务发展的必要条件,还将成为吸引消费者参与 B2C 电子商务的一个重要因素。

为了满足 B2C 电子商务中信息获取准确、便捷的要求,应用于 B2C 电子商务的信息获取技术应具有信息查全率与查准率高、用户负担轻、信息获取速度快和容易扩充等特点。B2C 电子商务中的信息获取起初由传统的信息检索技术实现。但是,由于传统的信息检索技术只能对已收录的网站信息进行查询,信息查全率低,并且需要专人维护数据库,难以扩充,因此随着 B2C 电子商务网站不断增加,这类技术难以满足 B2C 电子商务中消费者对信息获取的要求。而于20世纪90年代出现的应用于 Internet 中的信息抽取技术可以迅速地从 Internet 中自动抽取包含特定内容或具有特定结构的信息,信息查全率高,并且信息抽取范围不受限制。因此,随着信

息服务在 B2C 电子商务中的地位不断提高,方便、高效的信息抽取技术很有希望代替传统的信息检索技术,成为促进 B2C 电子商务发展的主要力量之一。

Main Reasons Internet Users Do Not Buy Online, 2001	
1	Poor, slow or non-existent customer service options
2	Privacy--fear or reluctance to disclose personal info and data
3	Higher total order prices than offline (including high shipping charges)
4	Security potential for credit card fraud, hackers, etc
5	Confusing site structure/ Cannot find product info
6	Not able to see/feel/test merchandise (where applicable)
7	General lack of trust in vendor
8	Inability to return goods
9	Site pages take too long to download
10	Out of stock merchandise

source: eMarketer, 2001

图2 用户不参与网络购物的原因

本文对 B2C 电子商务中的信息抽取技术进行了系统的分析与比较,并介绍了国内外信息抽取技术研究的最新进展。

2 B2C 电子商务中的信息抽取技术

2.1 利用网站查询表格的信息抽取技术

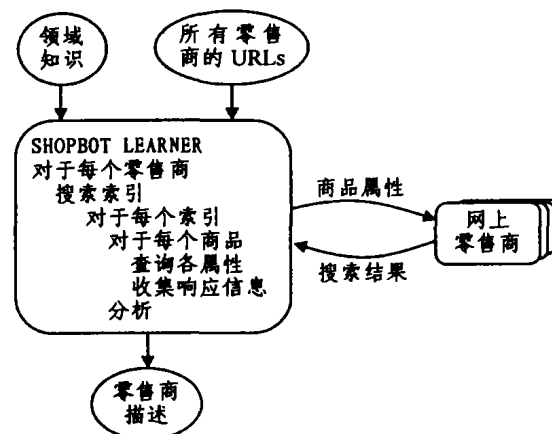


图3 ShopBot 的学习过程

利用网站查询表格的信息抽取技术是最早应用于 B2C

^{*}) 本文得到国家自然科学基金项目资助 (No.70171052, No.60075015)。于 琨 硕士生, 目前主要研究方向为 Internet 信息智能获取。

电子商务商品信息获取中的信息抽取技术。该技术首先对待抽取商品在网站中的描述进行离线学习,再利用归纳学习出的商品描述寻找合适的查询表格,最后利用网站原有的查询机制完成信息抽取^[4]。这种技术避免了复杂的自然语言理解,只需要用户提供少量的原始信息,用户负担较轻。

美国华盛顿大学1997年研制的 ShopBot 是这种技术的典型代表(图3)^[4],由于其优越的性能,该技术已经被著名网站 Excite.com 收购。然而,这种技术只能实现对具有查询表格的网站的信息抽取,信息查全率较低。

2.2 基于归纳学习的的信息抽取技术

该技术对若干个待抽取网页实例进行学习,自动分析出待抽取信息在网页中的结构特征并实现信息抽取(图4)^[6]。这种技术的信息查准率高,并且容易扩充,是信息抽取中常用的一种方法。

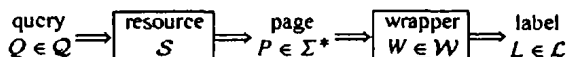


图4 基于归纳学习的的信息抽取过程

基于归纳学习的的信息抽取技术由美国华盛顿大学的 N. Kushmerick 提出,实验证明当用户提供2-44个学习实例时该技术可在10秒内对于70%的网站实现信息抽取,并且信息查准率可达100%^[5]。然而,这种技术需要用户提供大量实例,用户负担过重,并且当待抽取网页书写不规范或者待抽取信息的结构不唯一时难以抽取正确的信息。因此,C. Hgu 等人在归纳学习过程中增加了上下文规则以增加信息抽取的准确性^[6],而 D. Freitag 通过给学习实例反复赋予不同权值来提高对不规则信息的查准率^[7]。为了减轻用户负担,I. Muslea 等人在归纳学习过程中加入启发式规则,实验表明当用户提供10个训练实例时信息抽取查准率为85%~100%^[8]。随后,N. Kushmerick 在归纳学习过程中增加了对用户实例的上下文分析,可在仅需用户提供2-3个实例的情况下对70%的网站实现100%的信息查准率^[9]。

但是,由于归纳学习方法的特殊要求,基于归纳学习的信息抽取技术仍然需要用户提供学习实例,给用户增加了负担,并且对用户实例的学习也影响了信息抽取速度。

2.3 基于网页结构分析的信息抽取技术

该技术利用电子商务网站商品信息结构类似和 Html 语言半结构化的特点,对待抽取网页进行结构分析,并使用启发式规则找出可以有效划分信息边界的分割标记,再根据分割标记抽取在网页中重复出现的信息^[10]。这种技术具有抽取速度快、查全率与查准率高的优点,也是信息抽取中常用的一种方法。

基于网页结构分析的信息抽取技术由 D. W. Embley 等人于1999年提出,采用了领域描述等5条启发式规则进行标记分割,标记分割的正确率为100%,并且标记分割时间与网页规模呈线性关系^[10]。然而,这种技术需要用户提供待抽取信息所属领域的描述,而用户描述的正确性对信息查全率影响很大。为了解决这个问题,台湾的 C. H. Chang 等人用 PAT 树代替普通树结构存储网页标记信息以提高信息抽取质量,从而获得接近90%的信息查全率^[11]。而 D. Buttler 等人利用相同类型信息通常具有相同标志结构的特点替换了领域描述规则,实验表明替换后的信息查全率为93-98%,而标记分割算法的正确率同样可达100%^[12]。

但是,基于网页结构分析的信息抽取技术采用的启发式

规则仅适用于结构规范的网站,因此难以扩充。

2.4 基于自定义查询语言的信息抽取技术

该技术同样利用电子商务网站商品信息结构类似和 Html 语言半结构化的特点对网页结构进行分析,并将用户提供的自定义查询语言作为信息抽取时的启发式规则进行信息抽取^[13]。这种技术与单纯分析网页结构的技术相比,具有容易扩充的优点。

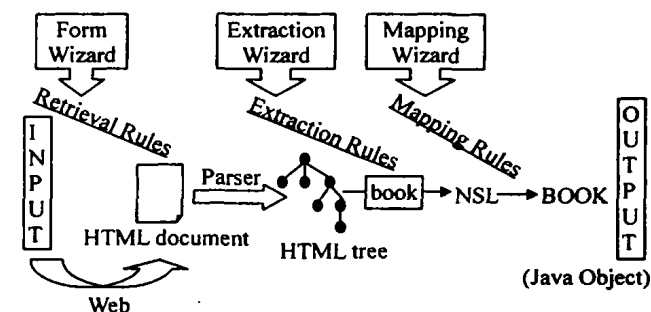


图5 W4F 工作流程

由 A. Saguhuet 等人研制的信息抽取器制造工具 W4F 是这种技术的代表之一(图5),它可以迅速生成针对不同内容、不同结构信息的信息抽取器,可扩充性强^[13]。为了减轻用户分析网页的困难,M. Bauer 等人又在系统中加入了动态提示信息,并使用用户示范(programming by demonstration)方法来弥补系统对网页结构分析的不足^[14]。然而,这种技术需要用户书写查询语言,对用户的个人能力要求较高,增加了用户负担。

2.5 基于模式匹配的信息抽取技术

该技术根据大量学习实例,归纳学习出待抽取信息的语法结构模式,并根据这些模式从待抽取网页中抽取相匹配的信息^[15],尤其适用于复杂结构信息的抽取。

M. Califf 等人于1997年开发了一个信息抽取系统 Rapi-er,将基于模式匹配的信息抽取技术应用于网页信息抽取,实验表明该系统可以在对90个实例进行学习的基础上实现86%的信息查准率与60%的信息查全率^[15]。然而,这种技术的信息查准率较低,此外学习过程中的语法分析也降低了系统的工作效率。为了提高信息查准率,S. Soderland 等人在归纳学习中加入了对领域规则的学习^[16],而南京大学的黄豫清等人提出由用户指定抽取信息结构的方法^[17]。随后 S. Soderland 又通过确定语段分割符来避免学习过程中复杂的语法语义分析^[18]。

但是,基于模式匹配的信息抽取技术只能实现单一领域的信息抽取,信息查全率低,并且需要用户提供大量学习实例,用户负担较重。

2.6 基于隐式马尔科夫模型的信息抽取技术

该技术将待抽取信息的属性作为马尔科夫模型中的一个状态,利用隐式马尔科夫模型(Hidden Markov Model)进行信息抽取(图6)^[19]。这种技术适用于结构化信息的抽取并且信息查准率较高。

基于隐式马尔科夫模型的信息抽取技术由 A. McCallum 等人于1999年提出,实验证明该技术的信息查准率为91.1%^[19],但是状态模型结构对信息查准率有很大影响。为了减轻这一影响,D. Freitag 等人引入了随机优化学习方法来确定合适的状态模型结构^[20]。但是,这种技术需要人工监测训练实例以建立模型,难以满足 B2C 电子商务中用户负担轻

的要求。

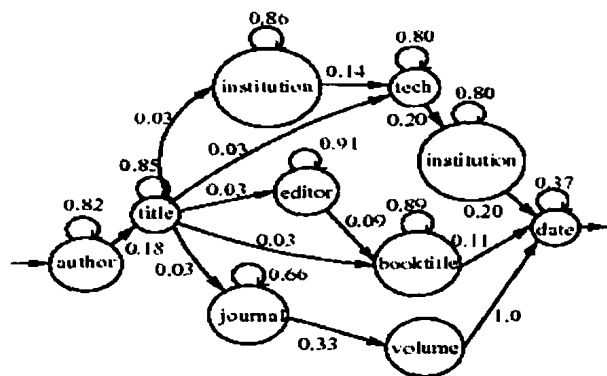


图6 基于HMM的信息抽取实例

总结与展望 在走出了网络经济的低谷之后, B2C 电子商务即将进入一个快速发展的时期, 而信息抽取技术很有希望成为推动 B2C 电子商务发展的主要力量之一。在目前存在的信息抽取技术中, 利用网站查询表格的信息抽取技术的信息查全率低; 基于归纳学习的信息抽取技术和基于自定义查询语言的信息抽取技术的信息查全率与查准率较高, 但是用户负担重; 基于网页结构分析的信息抽取技术不需要用户参与, 但是难以扩充; 基于模式匹配的信息抽取技术和基于隐式马尔科夫模型的信息抽取技术则增加了用户负担。由此可见, 目前存在的信息抽取技术均难以满足 B2C 电子商务的特殊要求。因此, 研制一种对个人能力要求低、用户负担轻、信息查全率与查准率高、抽取速度快并且容易扩充的信息抽取技术不仅可以帮助 B2C 电子商务企业在网络经济市场中取得竞争优势, 还可以吸引更多的消费者参与 B2C 电子商务, 从而促进 B2C 电子商务的发展。

参考文献

- 1 楼德宏. 国际互联网和企业电子商务. 轴承, 2001(5): 33~37
- 2 <http://www.the-dma.org/library/whitepapers/eCommerce-b2c-Exec-Summ-SS.pdf>
- 3 <http://www.linan.net/dzsw/b2c.htm>
- 4 Doorenbos R, Etzioni O, Weld D. A Scalable Comparison-Shopping Agent for the World-Wide Web. In: Proc. First International Conference Autonomous Agents, Marina del Rey, CA, 1997.

(上接第88页)

结束语 本文对移动 IP 环境下的通信匿名和路由问题作了探讨,提出了一种具有匿名特性的动态地址分配和路由策略。新的通信匿名策略具有以下特点:

1)移动节点在外地链路使用临时 IP 地址标志其身份,数据报文通过匿名和安全的连接传送,实现了通信匿名。

2)临时 IP 地址利用移动用户和家乡代理的共享密钥加密后传给移动用户,保证了移动用户身份信息的端到端保密性。

3)临时 IP 地址动态分配和更新,防止了对移动用户位置移动性进行跟踪。

4)能够有效地支持移动 IP 下的虚拟私用网。

同时我们对移动主机和移动网络的实现情况分别进行论述。新的通信匿名策略不需要对原有的认证与加密算法、密钥

39~48

- 5 Kushmerick N, Weld D, Doorenbos R. Wrapper induction for information extraction. In: Proc. IJCAI-97, Nagoya, Japan, 1997
- 6 Hgu C, Dung M. Generating finite-state transducers for semi-structured data extraction from the Web. J. Information Systems, 1998, 23(8)
- 7 Freitag D, Kushmerick N. Boosted wrapper induction. In: Proc. of the Seventeenth National Conference on Artificial Intelligence (AAAI2000), 2000. 577~583
- 8 Muslea I, Minton S, Knoblock C. A hierarchical approach to Wrapper induction. In: Proc. 3rd International Conference Autonomous Agents, 1999
- 9 Kushmerick N. Wrapper induction: Efficiency and expressiveness. Artificial Intelligence, 2000 (118): 15~68 (this issue)
- 10 Embley D W, Jiang Y, Ng Y K. Record-boundary discovery in web-documents. In: Proc. of the 1999 ACM SIGMOD, Philadelphia, Pennsylvania, USA, 1999
- 11 Chang C H, Hsu C N. Automatic extraction of information blocks using PAT trees. In: Proc. of the National Computer Symposium, Taipei, Taiwan, 1999
- 12 Buttler D, Liu L, Pu C. A Fully Automated Object Extraction System for the World Wide Web. Intel. Conf. on Distributed Computing Systems, 2001
- 13 Sahguet A, Azavant F. W4F: a WysiWyg web wrapper factory. Technical report, 1998
- 14 Bauer M, Dengler D, Paul G. Instructible information agents for Web mining. In: Proc. 2000 Conference on Intelligent User Interfaces, 2000
- 15 Califf M, Mooney R. Relational learning of pattern-match rules for information extraction. In: Workshop in Natural Language Learning, Conference Assoc. Computational Linguistics, 1997
- 16 Soderland S. Learning to extract text-based information from the World Wide Web. In: Proc. 3rd Intl. Conf. Knowledge Discovery and Data Mining, 1997
- 17 黄豫清, 戚广志, 张福炎. 从 WEB 文档中构造半结构化信息的抽取器. 软件学报, 2000, 11(1): 73~78
- 18 Soderland S. Learning information extraction rules for semi-structured and free text. Machine Learning, 1999, 34(1-3): 233~272
- 19 McCallum A, et al. A machine learning approach to building domain-specific search engines. In: Proc. IJCAI-99, Stockholm, Sweden, Morgan Kaufmann, San Francisco, CA, 1999. 662~667
- 20 Freitag D, McCallum A. Information extraction with HMM structures learned by stochastic optimization. In: Proc. of the Seventeenth National Conference on Artificial Intelligence (AAAI-2000), 2000

分配协议和路由协议进行修改,在实际应用中易于实现。

参考文献

- 1 Perkins. Mobile IP. IEEE communication magazine, May, 1997. 84~99
- 2 Solomon. 移动 IP. 机械工业出版社, 2000. 1
- 3 Doraswamy. IP Sec: the new security standard for the Internet, intranet, and virtual private networks. 机械工业出版社, 2000
- 4 Murhammer. 虚拟私有网络技术. 清华大学出版社, 2000
- 5 Fasbender. Variable and Scalable Security: Protection of Location Information in Mobile IP. IEEE VTC, 1996
- 6 Reed. Anonymous Connections and Onion Routing. IEEE JSAC, 1998, 10(4)
- 7 Distributed Registration to Mobile IP. draft-chuahli-mobileip-dremip-00.txt