

基于 Web 的开放领域问答系统研究

付鸿鹄^{1,2}

¹(中国科学院文献情报中心 北京 100080) ²(中国科学院研究生院 北京 100049)

【摘要】 主要对基于 Web 的开放领域问答系统及其涉及的信息检索、信息抽取、自然语言处理等相关技术内容进行了分析研究。提出了系统的一般结构,并对系统中问题分析、信息检索、答案生成各部分功能步骤进行了总结和分析。

【关键词】 信息检索 问答系统 自然语言处理 **【分类号】** TP311

Research of Web - Based Open Domain Question Answering System

Fu Honghu^{1,2}

¹ (Library of Chinese Academy of Sciences, Beijing 100080, China)

² (Graduate School of the Chinese Academy of Sciences, Beijing 100049, China)

【Abstract】 This article discusses Web - Based Open Domain Question Answering System and the relevant technologies, such as information retrieval, information extraction, natural language process and so on. It gives a general system architecture of Open Domain Question Answering System and analyzes the component of the it such as question analysis, information retrieval, answer generation.

【Keywords】 Information retrieval Question answering Natural Language process

1 引言

随着互联网的高速发展,如何快速准确地获取信息是信息技术领域研究的核心问题。目前网络搜索引擎已取得了巨大的成就,如 Google。但这些搜索引擎采用的仍然是基于关键词的检索方式,提供给用户的是与用户查询请求相关的网页列表,用户需要自己从中找出需要的信息。而用户更需要一种采用自然语言描述问题并获取问题答案而非相关文档的检索方式。顺应这种需求,开放领域的问答系统(Open Domain Question Answering, QA)成为近年来国际上研究的一个热点。

开放领域的问答系统(QA)是一个复杂的研究领域,它由信息检索、信息抽取和自然语言处理等共同构成。其基本任务是:给定一个问题和一个大文档集,返回问题的答案而非相关文档^[1]。

TREC(Text Retrieval Conference)即文本检索会议,是信息检索领域国际著名的评测会议^[2]。在 1999 年的 TREC8 中第一次设立 QA 任务,指出 QA 的目标是检索包含问题的真实答案的文档片断,而不是像传统的文本检索系统那样返回文档的列表^[3]。QA 任务的提出吸引了

很多大公司和科研机构参加,在其后的几届会议中 QA 成为最受关注的项目之一。

本文主要结合 TREC 以及其它相关会议中对于问答系统的讨论进行了基于 Web 的开放领域问答系统分析和研究。

2 QA 系统结构概述

2.1 QA 系统体系结构

QA 系统一般主要由问题分析、信息检索、答案生成(抽取/合并)三部分组成。如图 1 所示 QA 系统结构图。

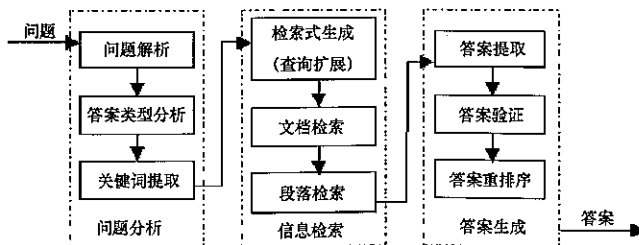


图 1 QA 系统一般结构图

问题分析即通过对问题的语法、语义结构等进行解析,获取该问题的关键词,提取问题的焦点、确定问题类型和答案类型等等。主要涉及自然语言处理技术。

信息检索主要包括文档检索和段落检索两步处理。

首
问
前
成
落
查

合
2.

中
法

型
式
的
(1
(1
tic

包
直
ar
文
多
问
类

理
一
也
问
自
友

是
是
这
些
问
题
自
身

首先采用传统的检索方式获取相关文档。在基于 Web 的问答系统中,通常使用成熟的搜索引擎,如 Google 等作为前期的检索工具,获取网上相关文档。然后,将文档划分成段落或者句子进行段落检索以获得问题相关的具体段落或句子。这一部分通常需要采用查询扩展技术以改善查询效果。主要涉及信息检索和信息抽取等技术。

答案生成,需要对候选答案进行去重和排序、抽取/合并等处理,组织成对问题的最终答案。

2.2 不同类型问题的处理

图1只是一个一般性系统结构,在具体实现的系统中,由于处理的问题类型、采用的建模方式、实现策略、算法等方面的不同,系统的结构也各不相同。

用户用自然语言提出的问题包括各种类型。不同类型问题需要的答案有很大差别,如果采用相同的处理方式,有些类型的问题得到的答案就有可能不能满足用户的要求。因此,在 TREC2003 和 2004 中,QA 主任务(Main Task)把问题分为三类分别处理,即,陈述类问题(Factoid)、列表类问题(List)和定义类问题(Definitions)^[4],要求返回问题的精确答案。

陈述类问题要求系统返回一个确切的答案,而不是包含答案的文本。答案字符串并不要求是从某个文档中直接抽取出来的,而是由 query-id、run-tag、doc-id、answer-string 构成。列表类问题要求系统抽取位于多个文档中的信息来组织答案,给出符合查询要求的信息的多个实例,如:“List the names of chewing gums.”。定义类问题是诸如“Who is Colin Powell?”或“What is mold?”之类的问题。

这三类问题对答案的要求各不相同,因此系统在处理上也要区别对待。列表类问题可以看作是多次查询同一个陈述类问题的简略版。即,能满足陈述问题的答案也是列表问题的恰当答案。由于在陈述类问题和列表类问题之间存在如此的相似性,许多参与单位或组织建立的系统对这两种问题都采用了相同或相似的处理办法。处理这两类问题通常可以采用类似图1的体系结构。

在对 Web 搜索引擎日志进行统计时发现,定义类问题在 Web 搜索中经常出现,因此这是一个非常重要的问题类型。然而,评价一个系统处理定义问题的能力要比评价一个系统处理陈述问题的能力要难得多。因为对系统给出的回答仅判断是对或错是不够的,还需要一种对预期的答案和系统给出的答案进行概念上的匹配的机制。因此,许多系统在处理定义问题时采用了模式匹配的技术。这种情况下,采用如图1的体系结构来处理就显得不够,需要增加模式库以及模式匹配的算法和模块,

建设模式库,并在问题分析和答案生成阶段分别进行模式匹配^[5-7]。

3 QA 系统功能实现

实现一个 QA 系统通常涉及多种技术和方法。目前有许多成熟的独立的工具,在构建一个问答系统时可以借用这些外部工具或方法以减少系统实现的复杂度,提高效率。

3.1 问题分析

在问题分析阶段,主要是对用户输入的自然语言表达的问题进行分析,确定问题的类型、问题的焦点、预期答案的类型并将问题转化为合理的符合逻辑的查询表达式,以备在信息检索和寻找答案时使用。这部分工作主要涉及自然语言处理技术。

自然语言处理可以包括两部分内容:词法解析和句法解析。由于研究是在一个开放领域中进行,输入的提问是开放领域的问题,因此大部分词典不能够对可能的提问进行充分的覆盖。在此阶段,可以使用一些外部工具来进行改善^[8]。如:Brill tagger,进行词性标注;BBN Identifinder,进行命名实体标注;WordNet,用于发现语义分类扩展关系、抽取词干等等。

问题分析部分经过细化需要完成以下几部分工作:确定问题的类型(Question Types)、确定问题的焦点(Question Focus)、确定答案的类型(Answer Types)、提取问题的关键词(Keywords)等。在这些工作中,无论是确定问题焦点、问题类型还是答案类型目的都是为了可以帮助用户缩小可能答案的范围。

在构建中文的问答系统时,在问题分析阶段首先需要对汉语句子进行分词和词性标注,并且在确定问题类型和预期答案类型时要根据汉语的语言特点和语法规则创建分析处理规则^[9]。

(1) 问题类型

确定问题的类型,即对问题进行分类。通过对大量问题的统计分析发现,用户提出的问题大概可以分为若干种类型,如询问人(Who)、询问时间(When)、询问地点和位置(Where)、询问数量(How much)、询问原因(Why)、询问定义(What)、其他等。对问题这样分类的好处是便于以后分析答案的类型。

问题分类可以使用机器学习方法自动完成^[10,11],也可以采用多种算法。如:基于语言建模的分类算法^[12-14]、简化贝叶斯算法等等。

语言建模的目的是建立一个可以捕获自然语言 L 的统计特征的语言模型 M_L 。给定一个词串 $S = w_1 w_2 \dots w_l$,使用 M_L 预测 S 在 L 中出现的概率 $\Pr[S|M_L] = \Pr_L[S]$ 。在进行语言

建模时,最常用的是 N-Gram 模型。N-Gram 模型原理简单但在实际应用中的效果良好。N-Gram 模型算法是基于如下假设:语言可以假设成是由一个 Markov 信源产生,该信源的符号集就是语言的最小单位词,信源向外源源不断地发出符号,这些符号串形成句子,然后由句成文。整句的概率就是各个词出现概率的乘积。

(2)问题的焦点

问题焦点是问题中的一个名词或名词短语,说明了问题的主要内容,使问题意义清楚,并突出强调了答案的类型。问题焦点对找到答案非常重要,系统通过一定的规则找到问题焦点,对问题焦点进行分析、最终确定出问题的答案类型,问题焦点最主要的目的是通过分析问题得到希望的答案。

(3)答案类型

答案的类型通常由对问题类型的分析结果确定。确定答案类型在后期答案生成阶段有重要作用。如,可以作为答案验证(Answer Justification)的依据,提高答案的准确率。

在对 TREC-8 的 QA 提供的问题进行分析时得出这样的结果:其中 80% 的问题需要的答案是一个命名实体(NE)^[15]。如:who(PERSON),when(TIME|DATE),where(LOCATION),how long(DURATION|LENGTH),how far(LENGTH)。

通常,特定的问题类型对应特定的答案类型。

(4)关键词提取

问题关键词的提取直接影响到后面的检索效果。一般来说关键词主要由名词、动词、形容词、限定性副词等组成。但在实际应用中可以把问题中除了疑问词和停用词以外的大部分词都作为关键词以提高检索的精确度。

关键词可能分为两种:一般性关键词、“必须含有”性关键词。所谓“必须含有”性关键词指的是这些关键词必须在答案段落中包含。而一般性关键词可以不被答案段落所包含。

关键词被赋予不同的权重,在检索段落时这些权重用来计算段落的权重。通常名词、“必须含有”性关键词、具有限定性作用的副词会有比较高的权重。“必须含有”性关键词主要由专用名词、限定性副词(如最大、最高、最远、最快、第一等等)、时间、数词等组成,问题焦点也是“必须含有”性关键词。之所以要制定“必须含有”性关键词原则是因为它们对问题有极强的限制性作用,不含有它们的段落几乎不可能是正确的答案,因此通过这些关键词的作用可以极大的提高检索的准确性^[16]。

关键词通常有三种类型:单词、短语或适当的名称^[8]。

3.2 信息检索

(1)查询扩展

由于在问题空间和文档空间之间存在着天然的语义鸿沟,因此,使用基于问题分析直接提取的关键词进行查询,常常不能获得满意的效果。如在答案段落中某些词常常不是原

来问题的关键词而是这些关键词的扩展,为了提高检索系统的精确度和召回率,必须对查询进行适当的扩展。

可以采用的扩展方式有:名词同义词扩展和语义蕴涵扩展、动词同义词扩展(意义用法相同的词)、根据问题类型扩展。对于这种关键词的扩展不仅能提高答案的召回率,而且还可以提高正确率。通常采用的工具有 WordNet、HowNet、Ontology 等等。在中文问答系统中通常可以采用《同义词词林》作为进行查询扩展的工具^[17]。

这部分工作在有些系统中由一个叫做“检索提问生成器”的组件完成,它采用不同的方法生成扩展后的检索式。如把词的变形态和导出词作为检索词^[5]。

(2)文档检索

在基于 Web 的问答系统中,文档检索主要是对 Web 信息进行检索,多采用 Google 作为前端的检索工具^[5,17]。这是由于 Google 拥有世界最大的网页索引并且它用于计算相关度的 PageRank 算法具有很高的准确率使它在同类工具中脱颖而出。查询序列提交到搜索引擎后,将返回列表中的网页下载下来作为相关文档。由于搜索引擎返回的页面非常多,通常选择排序靠前的几十个相关度最大页面下载。

(3)段落检索

与问题相关的答案通常存在于文档的某个段落或者句子中,因此需要把文档检索部分获得的文档拆分成段落或句子,从中选择最相关的部分。可以采用多种方法如段落过滤或信息抽取等,以获得尽可能少的摘录覆盖尽可能多的问题内容,缩小候选文档集。

如 IBM 的段落检索算法^[5,18-19]采用段落评分的方法。该方法对每个段落计算一系列的匹配度,包括:词匹配度(Matching Words Measure),同时出现在查询语句和段落中的词的 idf 值之和;词典匹配度(Thesaurus Match Measure),对于出现在查询语句中的词,它在 WordNet 中的同义词出现在段落中,计算它们的 idf 值之和;词不匹配度(Mis-Match Words Measure),出现在查询语句中但未出现在段落中的词的 idf 值之和;词群匹配度(Cluster Words Measure),计算同时出现在查询语句和段落中的邻近词的数量。这些值组合在一起给出某个段落最后的得分。

(4)段落权重与排序

检索出来的段落很多,需要人为制定一些权重机制,如:计算该段落或句子与检索式相似度等,然后计算出每个段落的权重并进行排序,找出相关性最强的那些段落(通常为前 10 个或 50 个等)作为最后抽取答案的候选答案集。相似度的计算可以采用 N-Gram 模型、传统的向量空间模型等方法进行^[17]。

向量空间模型是一个广泛采用的文本计算模型,可以将给定的文本转换成一个维数很高的向量,其最突出的特点是可以方便的计算出两个向量的相似度,即向量所对应的文本

的相似性。

3.3 答案生成

从检索阶段获得候选答案集中抽取问题的答案,并对抽取的答案进行验证、去重、重新排序、组合,生成符合提问需求的答案。

(1) 答案抽取

采用特定的方法从候选答案中选择问题的正确答案。可以使用信息抽取中命名实体的抽取办法。命名实体识别可以被看作是一个分类问题,每个词被划分到某实体类型或者不属于任何类型。相关算法如 *IdentiFider*,它基于隐马尔科夫模型,使用一个改进的算法,通过机器学习对人名、日期、时间、数字等进行正确地识别和分类。在抽取答案时,先用 *IdentiFider* 对候选段落或句子进行处理,找出所有的命名实体并识别其类型,然后将这些命名实体中与查询类型相同的作为问题的候选答案。

也可以采取多种方法、模型和策略。如 *TextMap*^[6] 系统采用了基于知识、模式、统计三种不同的方法进行答案的选择抽取。

(2) 答案验证

对于抽取出来的答案需要进行进一步的验证,判断其是否是真正符合问题要求的答案。答案验证试图建立起问题和候选答案之间的词汇—语义联系,并通过对这些联系的逻辑验证来判断答案正确与否^[1]。逻辑验证采用两种类型的公理。第一种从文本答案的事实中获得;第二种由人工建造的 *Ontology* 生成。

例如,类型检查就是进行答案验证的一种重要方法。有些答案是问题相关答案,但不符合问题的确切要求。如:对于问题“*What continent is Egypt on?*”。可能检索出的候选答案有:“*Middle East*”、“*Arab*”、“*Islam*”、“*Africa*”等等。而需要的答案类型是 *LOCATION(Continent)*,所以,只有“*Africa*”可以作为问题的最终答案^[8]。

4 结语

目前基于 *Web* 的问答系统研究还不是很多。华盛顿大学的 *MULDER*^[20] 是第一个完全基于网络的自动问答系统,它将检索到的网络文档下载到本地,并对这些文档进行详细的语法分析从中抽取答案。但现有系统的性能大部分没有达到理想的结果,还需要进一步的改进以达到更好的检索性能,基于 *Web* 的问答系统研究也越来越受到关注。

现有大部分问答系统将问题划分为不同的类型,都是相对简单的问题类型。但现实中存在很多复杂的问题,不能用某一个特定的问题类型来标识,需要辨识问题和答案的语法、句法、语义和逻辑结构。答案的选择也显

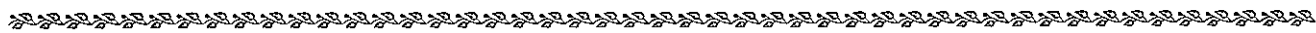
得非常复杂。需要确定问题的领域知识、把复杂问题分解成简单问题等。同时,对问题答案的结构组织也需要参照问题的语义和逻辑结构将简单问题的答案组织成员是复杂问题的答案。

近年来,语音识别技术取得了很大的进展。将语音识别技术与 *QA* 系统相结合,可以实现问题的语音输入。再对由语音输入的问题进行相应的分析处理,实现对语音输入问题的回答。即由语音驱动的问答系统。这项研究的关键还是语音识别技术和对语音输入问题的分析。

参考文献:

- 1 H. Yang, H. Cui, M. Maslennikov, L. Qiu, M. - Y. Kan, T. Chua. *QUALIFIER In TREC - 12 QA Main Task*. In: *Proceeding of the 12th Text Retrieval Conference (TREC)*. NIST. Gaithersburg. MD. 2003. 480 - 488. <http://trec.nist.gov/pubs/trec12/papers/nus-yang.qa.pdf> (Accessed May 23, 2005)
- 2 <http://trec.nist.gov/> (Accessed May 23, 2005)
- 3 E. M. Voorhees. *The TREC - 8 Question Answering Track Report*. In: *Proceeding of the 8th Text Retrieval Conference (TREC)*, NIST, Gaithersburg, MD, 1999. 77 - 82. http://trec.nist.gov/pubs/trec8/papers/qa_report.pdf (Accessed May 23, 2005)
- 4 E. M. Voorhees. *Overview of the TREC2003 Question Answering Track*. In: *Proceeding of the 12th Text Retrieval Conference (TREC)*. NIST. Gaithersburg. MD. 2003. 54 - 68. <http://trec.nist.gov/pubs/trec12/papers/QA.OVERVIEW.pdf> (Accessed May 23, 2005)
- 5 B. Katz, J. Lin, D. Loreto, W. Hildebrandt, M. Bilotti, S. Felshin, A. Fernandes, G. Marton, F. Mora. *Integrating Web - based and Corpus - based Techniques for Question Answering*. *Proceeding of the 12th Text Retrieval Conference (TREC)*. NIST. Gaithersburg. MD. 2003. 426 - 435 <http://trec.nist.gov/pubs/trec12/papers/mit.qa.pdf> (Accessed May 23, 2005)
- 6 A. Echihabi, U. Hermjakob, E. Hovy, D. Marcu, E. Melz, D. Ravichandran. *Multiple - Engine Question Answering in TextMap*. *Proceeding of the 12th Text Retrieval Conference (TREC)*. NIST. Gaithersburg. MD. 2003, 772 - 781. <http://trec.nist.gov/pubs/trec12/papers/usc-isi.qa.pdf> (Accessed May 23, 2005)
- 7 S. Harabagiu, D. Moldovan, C. Clark, M. Bowden, J. Williams, J. Benseley. *Answer Mining by Combining Extraction Techniques with Abductive Reasoning*. *Proceeding of the 12th Text Retrieval Conference (TREC)*. NIST. Gaithersburg. MD. 2003, 375 - 383. <http://trec.nist.gov/pubs/trec12/papers/loc.qa.pdf> (Accessed May 23, 2005)
- 8 E. Nyberg, T. Mitamura, J. Callan, J. Carbonnell, R. Frederking, K. Collins - Thompson, L. Hiyakumoto, Y. Huang, C. Huttenhower, S. Judy, J. Ko, A. Kupsc, L. V. Lita, V. Pedro, D. Svoboda, and B. Van Durme. *The JAVELIN Question - Answering System at TREC 2003; A Multi - Strategh Approach with Dynamic Planning*. *Proceeding of the 12th Text Retrieval Conference (TREC)*. NIST.

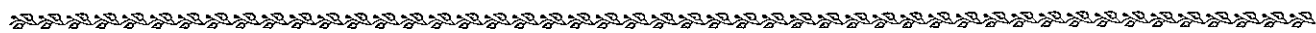
- Gaithersburg, MD. 2003. <http://trec.nist.gov/pubs/trec12/papers/cmu.javelin.qa.pdf> (Accessed May 23, 2005)
- 9 崔桓,蔡东风,苗雪雷.基于网络的中文问答系统及信息抽取算法研究.中文信息学报,2004(3):24-31
- 10 X. Li, D. Roth. Learning Question Classifiers. Proceedings of the 19th International Conference on Computational Linguistics (COLING). Taipei, Taiwan, 2002, 556 - 562. <http://acl.ldc.upenn.edu/C/C02/C02-1150.pdf> (Accessed May 23, 2005)
- 11 D. Zhang, W. S. Lee. Question Classification using Support Vector Machines. Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR). Toronto, Canada, 2003, 26 - 32. <http://www.comp.nus.edu.sg/~leews/publications/p31189-zhang.pdf> (Accessed May 23, 2005)
- 12 D. Zhang, W. S. Lee, National University of Singapore - MIT Alliance. A Language Modeling Approach to Passage Question Answering. Proceeding of the 12th Text Retrieval Conference (TREC). NIST. Gaithersburg, MD. 2003, 489 - 495. <http://trec.nist.gov/pubs/trec12/papers/nus-zhang.qa.pdf> (Accessed May 23, 2005)
- 13 W. Li, Question Classification Using Language Modeling. CHIR Technical Report, University of Massachusetts, Amherst, 2002. <http://ci-ir.cs.umass.edu/~weili/papers/qc.pdf> (Accessed May 23, 2005)
- 14 F. Peng, D. Schuurmans, Combing Na? ve Bayes and n - Gram Language Models for Text Classification. Proceedings of the 25th European Conference on Information Retrieval Research (ECIR), Pisa, Italy, 2002, 335 - 350. <http://net.pku.edu.cn/~wbia/Readings/classification/Combining%20Naive%20Bayes%20and%20n-Gram%20Language%20Models%20for%20Text%20Classification.pdf> (Accessed May 23, 2005)
- 15 R. Srihari, W. Li. Information Extraction Supported Question Answering. In: Proceeding of the 8th Text Retrieval Conference (TREC), NIST, Gaithersburg, MD, 185 - 197. <http://trec.nist.gov/pubs/trec8/papers/cymfony.pdf> (Accessed May 23, 2005)
- 16 李季,迟呈英.中文问答系统的研究.鞍山科技大学学报,2003(12):437-440
- 17 张永奎,赵辄谦,白丽君,陈鑫卿.基于互联网的中文问答系统.计算机工程,2003(9):84-85
- 18 Abraham Ittycheriah, Martin Franz, Weijing Zhu, Adwait Ratnaparkhi, IBM's Statistical question answering system. In: Proceeding of the 9th Text Retrieval Conference (TREC). NIST. Gaithersburg, MD. 2000. http://trec.nist.gov/pubs/trec9/papers/ibm_qa.pdf (Accessed May 23, 2005)
- 19 Abrahm Ittycheriah, Martin Franz, Salim Roukos, IBM's Statistical question answering system - TREC - 10. In: Proceeding of the 10th Text Retrieval Conference (TREC). NIST. Gaithersburg, MD. 2001. <http://trec.nist.gov/pubs/trec10/papers/trec2001.pdf> (Accessed May 23, 2005)
- 20 Kwok C, Etzioni O, Weld D S. Scaling Question Answering to the Web. In: Proceedings of the 10th World Wide Web Conference (WWW2001), HongKong, 2001. <http://www.cs.washington.edu/homes/etzioni/papers/acm-mulder.pdf> (Accessed May 23, 2005)
- (作者 E-mail: fuhh@mail.las.ac.cn)



(上接第71页)

参考文献:

- 1 安美荣,姜仁珍,张翠英. Web 引文的可查证性分析.情报学报,2004(8):480-483
- 2 杜云祥.三种图书情报学期刊网络参考文献著录分析研究.图书馆论坛,2004,24(6):96-98,130
- 3 CNNIC(中国互联网络信息中心)中国互联网络发展状况统计报告(2005/1). <http://www.cnnic.net.cn/index/0E/00/11/index.htm> (Accessed Mar. 6, 2005)
- 4 学科半衰期. <http://ei.lib.hit.edu.cn/edu/chapter2.files/frame.htm> (Accessed Mar. 6, 2005)
- (作者 E-mail: CooledCoffee@gmail.com)



(上接第91页)

- <http://www.w3.org/TR/2004/WD-xquery> (Accessed Oct. 29, 2004)
- 5 XML Path Language (XPath) 2.0 [EB/OL]. <http://www.w3.org/TR/2004/WD-xpath> (Accessed Oct. 29, 2004)
- 6 Prosis, J. 王铁等译. Microsoft. NET 程序设计技术内幕. 北京:清华大学出版社, 2003
- 7 Microsoft. XQuery over Documents with XQueryCommand [EB/OL]. http://whidbey.msdn.microsoft.com/library/default.asp?url=/library/en-us/wd_xquery/html/19d51739-5f82-43bd-97e1-42b0e05d2dd8.asp (Accessed Oct. 29, 2004)
- 8 Esposito, D. 宁建平译. Microsoft. NET XML 程序设计. 北京:机械工业出版社, 2003
- (作者 E-mail: ouhg@sohu.com)