# 基于 Deep Web Search 技术的主题式爬虫模块研究与设计

孟 敬[1],刘寿强[2,3]

1. 广东交通职业技术学院,广州 510650
2. 华南师范大学物理与电信工程学院,广州 510006
3. 华南理工大学计算机科学与工程学院,广州 510040

**摘要** 随着 Web 技术的飞速发展，海量数据的管理与搜索变得尤为重要。海量信息的异构性和动态性特点要求信息集成需要 Web 爬虫来自动获取这些页面,以便进一步处理数据。而一些企业内部的资料既要保密又要供不同的内部职员使用,这种既开放又保守的特点成为企业发展的瓶颈。为了帮助用户完成这样的任务,本文改变传统的资源共享形式,为企业提供了一个高效便利保密的资源共享管理平台——企业搜索引擎(ESE),提出了一种基于主题式爬虫的 Deep Web 页面的企业搜索引擎(ESE)的和基于开源 Java Lucene 的索引企业搜索系统设计与实现方法。通过在电信行业 Deep Web 站点部署实验,经运行检验,结果达到了设计指标要求,为电信行业搜索发挥了作用。并对搜索的精度、速度,以及垃圾网页反舞弊等方面研究进行了展望。

**关键词** 主题式爬虫;企业搜索引擎;Deep Web 搜索技术;电信;设计与实施

**中图分类号** TP311　　**文献标识码** A　　**doi** 10.3981/j.issn.1000-7857.2011.21.004

# Research and Design of Topical Crawl Module Based on Deep Web Search Technology

MENG Jing[1], LIU Shouqiang[2,3]

1. *Guangdong Communication Polytechnic, Guangzhou 510650, China*
2. *School of Physics & Telecommunication Engineering, South China Normal University, Guangzhou 510006, China*
3. *School of Computer Science & Engineering, South China University of Technology, Guangzhou 510040, China*

**Abstract** As the web rapidly grows, massive data management and search becomes particularly important. Heterogeneous mass information and dynamic characteristics of information integration require Web crawlers to automatically access these Web pages in order to further process the data, the internal confidential information of enterprises must be only used by different internal staffs, the openness and conservative features become the major bottleneck for the enterprise development. To help out this task, some forms of the traditional resource sharing are changed, an efficient, convenient, and confidential resource sharing management platform-Enterprise Search Engine (ESE) is provided, and the design and implementation method for Deep Web ESE based on topical crawl and indexed enterprise search systems based on open source Java Lucene is proposed. After the deployment and experiment of Deep Web site in the telecommunications industry, the results are proved to meet the design target. It plays an important role in the telecommunications industry. Finally, the studies on the search accuracy and speed, anti-spam pages and fraud, etc are looked forward.

**Keywords** topical crawl; Enterprise Search Engine; Deep Web Search technology; telecommunications; design and implementation

## 0 引言

随着 World Wide Web 的飞速发展,Deep Web 蕴含了日益增长的大量各种各样的数据，并且在以惊人的速度增长。如何对 Deep Web 数据进行有效管理，以满足用户不断增长

收稿日期:2011-06-29;修回日期:2011-07-08
基金项目:广东省交通科技项目(2007-27);国家自然科学基金项目(61072028)
作者简介:孟敬,副教授,研究方向为计算机网络通信及其应用,电子信箱:menjin@163.com

的信息需求,成为新的研究课题。海量的信息往往表现为异构动态的,尤其是行业数据库,变得越来越专业和庞大。其中小部分存储在表层网络(Surface Web,即静态网页),通常包括小型网络站点和个人网页,如 *.htm、*.html、*.shtml、*.txt、*.xml 等;大部分存储在深层网络(Deep Web,即动态网页,URL 中含"?"或输入参数的网页),包括 ASP、PHP、PERL、CGI 等在 Server 方进行处理的网页,主要由大型站点构成。后者无论在质量还是数量上都是前者无法比拟的。通用的搜索引擎如 Google、Yahoo 等,使用 spider 或 crawler 根据网页间的超链接关系对静态网页(主要是 Surface Web)进行抓取、存储、分类和索引。但是这类传统搜索引擎无法抓取动态产生的 Web 页面(主要是 Deep Web),因而无法获取数据库中的信息。因此,目前通用搜索引擎能够索引到的信息只是表层网络中的。解决以上问题的方法是对表层网页和深层网页分别进行采集和保存:Surface Web 由传统搜索引擎进行收集;Deep Web 由 Deep Web 搜索引擎进行采集[1-3]。

Deep Web Search 技术不以抓取和检索动态网页为目的,而是通过向可搜索的信息数据库或专用搜索引擎发出查询请求,对得到的结果进行分析,把动态网页中重复的规律找出来,从而把原始的结构化数据提取出来。Deep Web Search 技术搜索流程见图 1[1,3]。

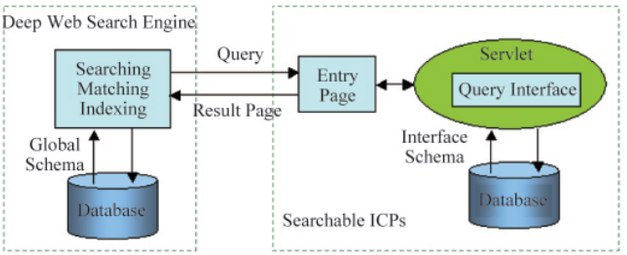


图 1　Deep Web Search 技术搜索流程

Fig. 1　Search process for deep Web technology

以 Deep Web Search 技术为基础可以构造出通用的数据搜索平台,该平台允许用户根据自己对信息的需求,将政府站点、搜索引擎、新闻站点、电信站点、教育科研等站点添加为数据源,并根据用户的查询内容,同时向多个数据源发出请求,对查询结果提供分类检索,从而使用户得到更加专业,内容更为丰富的信息查询服务[4]。

当前,如何利用搜索技术整合企业和组织内外部信息资源,提供专业和定向的搜索服务,建立企业搜索引擎(Enterprise Search Engine,ESE)已经成为信息化建设的焦点。这里的"企业"并非指单纯的企业,可以理解为"企业级",如政府、教育、科研、媒体、医疗、军队、安全部门都有类似的应用需求,与普通的互联网搜索引擎相比,企业级搜索有更高的要求。因此,针对企业网中不同的用户,对不同的资源其使用权限都可能不一样,需要企业搜索引擎能够对用户、资源、权限分级管理和控制,确保系统的安全。从查全率来看,互联网搜索引擎无从谈起查全率,而在企业级的某些应用中,是不允许有所遗漏的检索,必须对企业内部每个需要提供服务的信息进行索引。在检索机制上,必须在保障效率的前提下达到全面搜索的要求。

在互联网上,由于信息自由等特点,决定了搜索只能通过"关键词匹配"这种核心检索手段去实现。而在企业内部,信息的组织更复杂,要求企业级搜索引擎有完善的信息分类体系和元数据、对象数据多层逻辑的组织形式,在查询上满足基于对象数据内容和元数据标引体系的精确查询要求。企业内部的搜索服务具备业务特性,需要将搜索结果运用到企业的运营和决策中。所以通过搜索引擎提供的服务,必须能动态地反应实际情况,即当内部的信息发生变化时,必须能够实时反应。因此,企业搜索引擎应该具有复杂结构数据的搜索、严格的安全搜索、高可靠的查全和查准和实时的信息搜索服务等特性[3]。

## 1　电信行业高级搜索引擎设计框架

本系统是利用 Deep Web Search 技术开发电信行业高级搜索引擎的原型系统。该系统的结构见图 2。

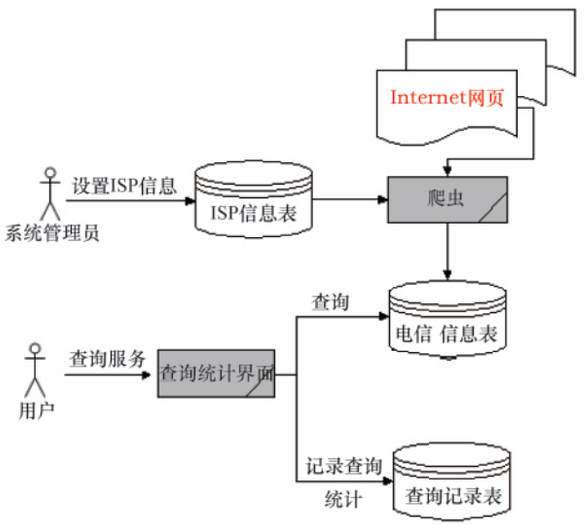


图 2　电信 Deep Web Search 搜索流程

Fig. 2　Search process for telecom Deep Web Search

电信行业高级搜索引擎原型系统主要由两个模块构成:爬虫模块和查询统计模块。查询统计模块采用 B/S 结构,B/S 结构是网络计算时代的软件技术架构,与传统的 C/S 结构的软件相比具有以下优点:客户端软件实现简单,易于操作和管理,容易实现跨操作系统平台、跨数据库平台的分布式应用,能够有效提高整个系统的运行效率,降低应用系统部署和管理的难度[1-3]。

## 2　主题式爬虫模块设计

主题式爬虫模块设计方案是整个系统的一个整体规划,如模块组成、体系结构和应用部署等,是对系统需求进一步理解之后,进行一定程度的优化而得出的整个设计,它应用的对象是项目分析、设计、编码和测试及项目实施人员。

根据需求,爬虫系统可归纳成 5 大组成部分:爬虫种群初始化、爬虫搜索模块、爬虫分析处理模块、爬虫数据存储与索引模块和主题相关查询模块[1-3]。

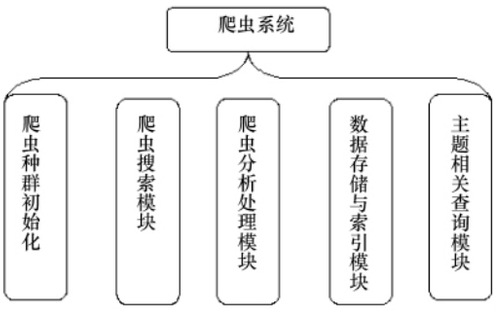


图 3　爬虫系统模块组成

Fig. 3　Crawl system modules

各个模块的功能如下。

爬虫种群初始化:实现爬虫系统对初始种群 url 的设定和订制,包括自动和人工修改。

爬虫搜索模块:实现爬虫系统对 url 的页面信息采集。

爬虫分析处理模块:实现爬虫系统对采集的页面信息进行分析,评定是否为主题相关。

爬虫数据存储与索引模块:实现爬虫系统对评定合格的页面信息进行存储,并生成相应的索引。

主题相关查询模块:实现爬虫前台页面进行主题相关的查询。

## 3　功能指标

### 3.1　爬虫与信息抽取

(1) 维护人员可方便地指定采集的目标站点或页面。

(2) 维护人员可方便地设定信息监控的时间周期,包括指定 1d 之内的多个定点执行时间, 或者设定两次更新之间的时间间隔,以 min 为单位,并可设置为不间断运行。

(3) 爬虫模块负责按照设定的 ISP 信息下载网页。根据行业需求可以确定信息来源网站,例如 c114,cnii 等行业信息网站,爬虫模块应能够根据配置信息从上述指定网站抓取网页,网页抓取率应达到 100%;对于需要登陆才能查看内容的网站,模块应该能过自动登录。

(4) 爬虫模块应该是可配置的。由于目标网站(信息源)和网页信息的格式是变化的,因此,爬虫模块应该是可配置的,而且项目成果要包括一个客户端配置工具,避免手工修改配置文件;配置工具应该是友好的,非专业人员经过适当的培训也可以使用配置工具; 配置文件应该是容易测试的,以保证对配置文件的修改是正确的。

(5) 论坛搜索模块——针对论坛的爬虫。针对论坛的特殊性(登陆需要输入验证码;对同一 IP 发出连接请求有时间间隔限制;论坛内容需要回复或购买才可以浏览,或者需要特殊条件才可以浏览等),爬虫应可以对论坛内容进行搜索。

(6) 提供有效的更新手段,已经采集过的信息不会重复采集,更新时只获取前次采集后更新的网页。

(7) 程序运行可靠,容错性强,完成初始设定后可长时间稳定运行。

(8) 程序能够从 Word、PDF、PPT、Excel 等各类型文档中抽取信息。

### 3.2　信息处理

(1) 支持基于网页内容的自动分类。

(2) 支持基于网页内容的自动排重,可将重复的网页进行标注,只发布不重复信息;可根据文档内容的匹配程度确定是否重复(应该比利用网页标题和大小等规则判断据有更强的准确性、实用性以及运行效率)。

(3) 能够统计相同或相似信息的引用次数,并且该值应该作为查询结果的排序依据之一。

(4) 自动过滤网页中新闻的正文内容,剔除垃圾信息。

(5) 对带有表格内容的网页保留表格原格式内容。

### 3.3　应用指标

(1) 提供高级查询:可以根据事件发生时间、发生地区等信息交叉查询。

(2) 提供二次查询功能:如果查询到的信息过多,可以在“信息查询结果”页面进行二次查询,提高查询准确率。

(3) 相关关键字:根据用户输入的关键字,系统可以向用户提供查询次数最多的几个相关关键字。

(4) 搜索效率:系统的响应速度应足够快,初期目标应达到每秒处理 100 个并发请求[1,5-7]。

## 4　总体设计与实施

### 4.1　系统的运行环境

考虑到系统的安全性和成本问题。爬虫系统采用高安全性和低成本的 J2EE 多层架构开发,数据库为 Mysql,采用的操作系统为 Windows 2003 以上,硬件配置如下。

(1) 硬件环境(应用服务器)

硬盘:160GB;CPU:2 pcs;内存:2.0GB;网卡:2pcs;RAID:1;OS:Windows。

(2) 系统软件环境

数据库服务器采用 Mysql,Web 服务器采用 Tomcat,操作系统为 Windows 2003,客户端采用 Internet Explorer 6.0。

### 4.2　软件体系结构

根据系统的特点,构架见图 4。

前台查询采用 3 层 B/S 结构,是将应用功能分成表示层、

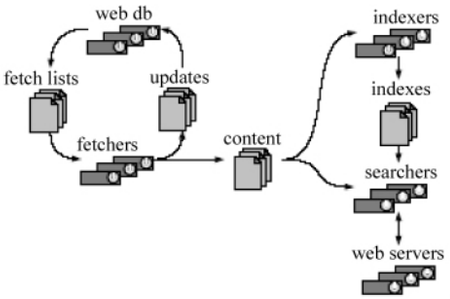


图 4　系统软件体系结构

Fig.4　System software architecture

应用层和数据层 3 部分。其解决方案是:对这 3 层进行明确分割,并在逻辑上使其独立。

表示层负责与客户交互，并将客户的请求向应用层提交。

应用层主要是响应表示层的请求,执行任务,从数据层提取数据,并将数据回传给表示层。

爬虫后台管理采用的 C/S 结构，封装了应用系统与远程控制代理的通信。考虑到本系统中需要跨平台进行数据同步和交互操作。如果采用基于 TCP 的访问方式,就需要开放数据库的 IP 端口访问及受限的用户名和密码给 Intranet 才能实现。但这种方法存在两方面的问题:一是在开放数据库 IP 端口访问的用户名和密码的同时,也给该数据库所在部门带来了安全隐患,一旦 Web 服务器受到攻击而使用户名和密码被窃取,开放的数据库也就处在被攻击的风险之中;二是有的使用的是第三方开发的基于数据库的工具,其部门本身并没有数据库的管理权限，因而也就无法添加用户名和分配权限。因此,基于安全和通用性的问题,在开发过程中,采用基于 Socket 通信的跨平台数据同步和交互的方法。

### 4.3　系统功能模块

系统包含 5 大模块:爬虫种群初始化、爬虫搜索模块、爬虫分析处理模块、爬虫数据存储与索引模块、主题相关查询模块。各模块的设计如下。

(1) 爬虫种群初始化

由于主题爬虫是面向选定主题的,初始种子的赋予应来自本领域,否则爬虫无法展开爬行工作。具体做法是,采用元搜索引擎搜索出的网页,从中选取质量较高的主题 URL。通常由人工完成筛选,这样可信度会更高。

(2) 爬虫搜索模块

根据初始种群以及相应的设置,包括深度、线层数等条件,通过“注射器(injector)”把网址“注入”到临时文件中。执行状况如下:

060915 032503 parsing file:/J:/nutch/conf/nutch-default.xml
060915 032503 parsing file:/J:/nutch/conf/crawl-tool.xml
060915 032503 parsing file:/J:/nutch/conf/nutch-site.xml
060915 032503 No FS indicated,using default:local
060915 032504 No FS indicated,using default:local
060915 032504 爬虫存放目录: crawltest
060915 032504 rootUrlFile = urls
060915 032504 线程 = 1
060915 032504 爬行深度 = 1

(3) 爬虫分析处理模块

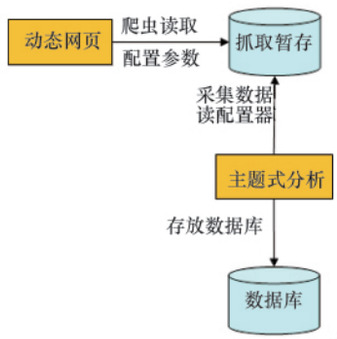


图 5　爬虫分析处理模块

Fig. 5　Crawl analysis processing module

对于爬虫获取的页面,为了保证爬虫获取的网页能够尽量向主题靠拢,必须对网页进行过滤,将主题相关度较低的网页(小于设定的阈值)剔除,这样就不会在下一步爬行中处理该页面中的链接。因为一个页面的主题相关度如果很低,就说明该网页很可能只是偶尔出现某些关键词,而页面的主题可能和指定主题几乎没有什么关系,处理其中的链接意义很小,这是主题爬虫和普通爬虫的根本区别。普通爬虫是根据设定的搜索深度,对所有链接进行处理,结果返回了大量无用的网页,而且进一步增加了工作量。主题相关度的计算是采用向量空间模型算法,具体算法待续。

其中,关键的技术就是如何确立主题。主题确立是主题爬虫工作的基础，采用的方法是用关键词集来确立主题,其中每个关键词拥有指定的不同权值。权值的设置有两种方法:手工设置和特征提取。特征提取是指给定一个与主题有关的网页集合，由程序自动提取这些网页里面共同的特征,并根据频率确定权值。手工设置的好处是实现简单,同时,人的经验一般比较准确,跟实际情况不会出现大的偏差,缺点是可能有缺漏,权值的量化定义不够精确;特征提取的优点是权值量化定义精确,但要求选取用来提取特征的网页集合必须是很有代表性和全面概括性的,否则就可能出现很大的偏差。最佳的方法是综合二者的优点。执行状况如下

060915 032513 fetching http://www.wx800.com/index.php
060915 032514 fetching http://www.c114.net/
060915 032515 fetching http://comm.ccidnet.com/
060915 032532 fetching http://www.ccidnet.com/
060915 032533 fetching http://www.tele.hc360.com/
060915 032534 fetching http://www.cttl.cn/
060915 032535 fetching http://www.handsite.com/

060915 032536 fetching http://www.enet.com.cn/enews/

(4) 爬虫数据存储与索引模块

该模块包括爬虫数据存储与爬虫数据库索引建立两部分。将通过分析的网页存储到临时数据中,爬虫在爬行结束前,将所有暂存的页面存储到 mysql 中,并对其建立索引。执行状况如下:

060915 032544 Updating J:\nutch\Twebdb\segments from J:\nutch\Twebdb\db

060915 032544 reading J:\nutch\Twebdb\segments\20060910012353

060915 032544 reading

060915 032544 Sorting pages by url...

060915 032544 Getting updated scores and anchors from db...

060915 032545 Sorting updates by segment...

060915 032545 Updating segments...

(5) 主题相关查询模块

爬虫获取的数据需要通过 Web 页面进行查询,主题查询包括的功能有中文分词、查询排列、数据分析、数据库更新。

其中,中文分词用来拆分查询条件,拆分的好坏直接影响到查询结果的精确度;排序模块的作用是对网页的重要程度进行排序,把价值高的网页排到前面,以便它们更容易地被选择到。影响网页排序的因素很多,通过主题相关度的计算已经能够得到一个比较合理的排序,但为了得到一个更加合理的排序,必须辅助考虑其他因素。在对网页进行排序时,可综合考虑主题相关度和链接分析两个关键因素,链接分析主要模仿改进的 PageRank 算法, 因为 Google 的成功实践证明,PageRank 算法的结果还是比较令人满意的。具体的算法实现中,应该对主题相关度和 PageRank 赋予不同的权重。

数据分析是给出当前查询出的页面的具体信息,包括大小、访问次数等,数据库更新是指每次被查询的页面都会更新其查询率[1,3-4,6,8]。

## 5 应用

应用部署如图 6 所示,所有的组件均部署在应用服务器上,同时,应用程序部署在受监控的设备上。爬虫后台管理启动一个代理服务,负责接受和执行应用程序发来的消息和命令,并定时向系统应用监控程序发送收集的信息。爬虫后台搜索应用程序负责处理业务逻辑和接受代理的返回结果,并负责写入相应的数据库中。经运行检验,结果达到了设计指标要求,正在为电信行业搜索发挥作用[4,6,9]。

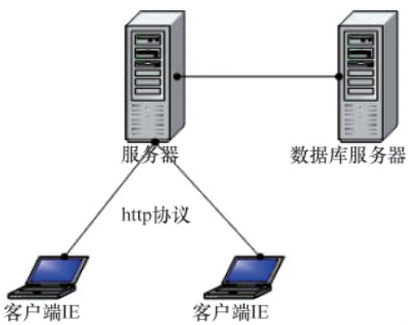


图 6 应用部署

Fig. 6 Application deployment

## 6 结论与展望

海量数据的深度挖掘是当今信息共享与应用的热点与难点,而企业内部资源管理是企业信息化进程中的一个重要环节,实际上需要考虑的因素还很多,尤其是今后 ESE 要面向大中型企业内部信息网全面实时安全分级的信息检索与更新等发展,对图形界面的优化、索引建立的时间、查询的效率等方面提出了更高的要求[5-8]。本文把 Deep Web Search 与 ESE 垂直搜索的结合这一重要课题进行了有益的探索,并在电信行业进行了应用,取得了一定的效果。但是对于检索的中文分词处理, 结果的排列方式和反馈方式等涉及比较少,今后有待加强和改善。下一步考虑搜索的精度、速度,以及垃圾网页反舞弊等方面研究,把链接分析和内容分析这两种学术界主流的研究方法结合起来, 更好地防治垃圾网页的泛滥;还可以从体系结构等方面入手提高服务器性能,如采用分布式技术,如 P2P、网格、云计算等进一步提高爬虫系统的性能,以适应更大规模和未来发展的需要。

## 参考文献(References)

[1] 郑冬冬, 赵朋朋, 崔志明. Deep Web 爬虫研究与设计 [J]. 清华大学学报: 自然科学版, 2005, 45(S1): 1896-1902.
Zheng Dongdong, Zhao Pengpeng, Cui Zhiming. *Journal of Tsinghua University: Science and Technology*, 2005, 45(S1): 1896-1902.

[2] 李晓明, 刘建. 搜索引擎技术及趋势[EB/OL]. [2003-04]. http://www.se-express.com/se/se07.htm.
Li Xiaoming, Liu Jian. Search engine technology and trends [EB/OL]. [2003-04]. http://www.se-express.com/se/se07.htm.

[3] 王春永, 尹祥贵, 邵乐, 等. 基于 Nutch 构建主题搜索引擎的研究[OL]. 中国科技论文在线, 2006.
Wang Chunyong, Yin Xianggui, Shao Le, *et al*. Research of topic specific search engine based on nutch[OL]. http://www.paper.cn, 2006.

[4] 高琰, 谷士文, 谭立球, 等. 基于 Lucene 的搜索引擎的设计与实现[J]. 微型计算机发展, 2004, 14(10): 27-30.
Gao Yan, Gu Shiwen, Tan Liqiu, *et al*. *Microcomputer Development*, 2004, 14(10): 27-30.

[5] 张岩. 一场无休止的竞赛——搜索引擎与垃圾信息制造者之间的战争[J]. 中国计算机学会通讯, 2007, 3(4): 3-9.
Zhang Yan. *Communications of the CCF*, 2007, 3(4): 3-9.

[6] Zhang Z G, Chen J, Li X M. A preprocessing framework and approach for web applications[J]. *Journal of Web Engineering*, 2004, 2(3): 176-192.

[7] Josepb S K, Bebnam A R, Nima S, *et al*. Collaborative spam filtering using e-mail networks[J]. *IEEE Computer Society*, 2006, 39(8): 67-73.

[8] Page L, Brin S, Motwani R, *et al*. The pagerank citation ranking: Bringing order to the web [D]. Standford: Standford Digital Library Technologies Project, 2001.

[9] 马晓玲, 吴永和. 对于搜索引擎优化(SEO)的研究[J]. 情报杂志, 2005(12): 119-121.
Ma Xiaoling, Wu Yonghe. *Intelligence magazine*, 2005(12): 119-121.

(责任编辑 林祥磊)