海量网络学术文献自动分类系统*

王效岳 白如江 王晓笛 祝娜

[摘 要] 随着Internet的发展,互联网上的学术文献数量呈指数增长,很难为科研工作者所利用,因此亟需一种方法对海量的网络学术文献进行自动的搜集、整理、分类。在前期充分的实验论证后,设计实现一个海量网络学术文献自动分类系统,该系统使用模块化设计,包括学术文献自动抓取模块、学术文献词-文档矩阵处理模块、本体集成模块以及基于语义驱动的分类模块。实验证明,该系统可以有效地完成海量学术文献的自动抓取、处理和分类工作。

[关键词] 学术文献 自动分类 并行处理
[分类号] G350
DOI:10.7536/j.issn.0252-3116.2013.16.022

1 引 言

目前,网络学术资源无论是在广度上还是深度上都呈现上升趋势,日益受到学术界的关注。海量网络学术文献规模巨大且更新速度快,对其进行充分挖掘有着重要的学术价值,然而,这些特点也成为了阻碍科研工作者对其进行利用的绊脚石。如何获取及处理海量学术文献,对计算机处理和吞吐能力都是严峻的考验,不论从处理速度、存储空间、容错性还是从访问速度而言,单台计算机平台架构及处理能力都难以圆满地完成这一任务。

并行处理技术则可以较好地解决这一问题。基于MapReduce计算范式,可以将所有处理任务分解成独立运行的任务,并将任务在同一集群的不同节点上运行,节点之间可以互相通信。在存储方面,通过分布式存储系统,数据也被分解成块,使用冗余的方式进行存储。这样,通过任务和数据的分解,可以较好地解决单机时内存消耗大、处理速度慢以及特征向量维度高等问题。

由于网络学术文献数量巨大,有效利用较为困难,对其进行基于学科的自动分类有着现实意义。文档自动分类在信息检索、数据挖掘、垃圾邮件过滤、数字图书馆等领域具有广泛的应用。常见的分类方法有两类:一类是基于规则的,通常需要大量的领域专家对文本进行规则提取,费时费力且分类效果较差;另一类方法基于统计学的机器学习方法,包括最近邻法、支持向量机、朴素贝叶斯、决策树、神经网络[1]等,这类方法通常采用特征向量空间进行文档分类模型的训练。然而词特征向量忽略词间语义关系,对于同义词、多义词以及词间上下位关系不能有所体现,导致向量空间维度过高,在对海量文档分类时会出现内存不足、分类速度慢、分类性能低等问题,无法将文档自动分类技术与方法更广泛地应用于具体领域的实践中。

为解决传统基于词向量空间的文档自动分类过程中存在的问题,国内外学者提出了一系列语义驱动的文档自动分类方法,如潜在语义分析法、本体语义映射法、概念格构建法、规范化概念分析法等。语义驱动的文本自动分类方法虽然能够极大地降低文档向量空间维度,但也有很多缺陷,如语义推理能力要求高、计算复杂度也高,无法快速有效地对网络上海量文档进行语义分类等。

这些问题的出现为海量网络学术文献自动分类的研究提供了新的视角,也使本研究更加必要和有意义。本文拟基于Heritrix与Hadoop[2],提出一种海量网络学术文献自动获取及并行处理模型。使用Heritrix平台指定规则对种子站点数据进行抓取,对于抓取到的

* 本文系2010年国家社会科学基金项目"海量网络学术文献自动分类研究"(项目编号:10BTQ47)和文化部科技创新项目"大规模网络学术文献并行处理与自动分类研究"研究成果之一。
[作者简介] 王效岳,山东理工大学科技信息研究所教授,博士,E-mail:wangxy@sdut.edu.cn;白如江,山东理工大学科技信息研究所馆员;王晓笛,山东理工大学科技信息研究所硕士研究生;祝娜,山东理工大学科技信息研究所硕士研究生。
收稿日期:2013-06-18 修回日期:2013-07-30 本文起止页码:117-122 本文责任编辑:易飞

文件资源，根据设定的学术文献特征规则，对其进行判定，然后选择其中的部分文献邀请领域专家进行类别标引，训练机器学习分类算法，最终实现所有文献的分类。在实现上述步骤的过程中，提取出处理过程中存在的可并行工作的分量，用分布式模型来实现这些并行分量的并行执行过程。另外，基于 WordNet[3]、DBpedia、Cyc、HowNet 本体集成的语义分类系统实现了多本体之间的概念一一映射关系的集成本体库；然后基于该集成本体库，将传统高维词向量空间转换成低维的概念向量空间，从而实现获取的海量学术文献文档的自动分类。

2　海量网络学术文献自动分类系统

本系统以自动获取海量文献以及自动对文献进行分类为目标，其框架如图 1 所示：

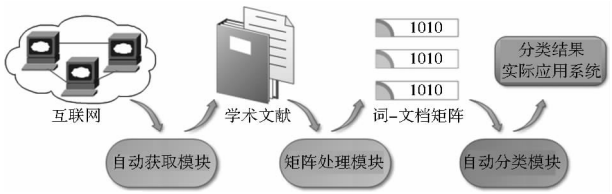

图 1　海量网络学术文献自动分类系统框架

图 1 中，文献自动获取模块首先从互联网上按照预定的规则和条件抓取并判定学术文献，从而过滤无关文件；然后通过矩阵处理模块将学术文献转换为词 - 文档矩阵，以便后续处理；最终将词 - 文档矩阵导入经过训练和本体集成的自动分类模块，得到分类结果。

2.1　海量网络学术文献自动获取

在海量网络学术文献自动分类系统中，需要实现海量学术文献的获取。首先使用 Heritrix 从特定网站上抓取域名下所有的 PDF 文件，并用 CheckPDF 读取所有 PDF 文件，通过基于规则的判定方法识别学术文献，如图 2 所示：

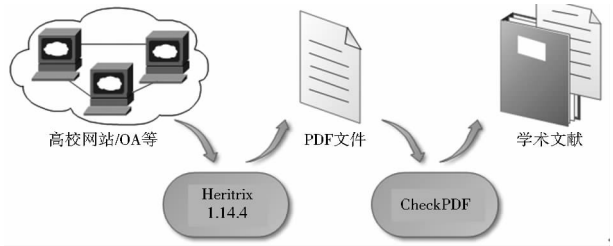

图 2　海量网络学术文献自动获取

在抓取工具的选择上，笔者从抓取效率、可扩展性等方面研究分析了 Nutch、Heritrix、Jspider、Web-Harvest 等网络资源抓取平台，最终选择 heritrix 作为抓取平台[4]。Heritrix 具有高度的可扩展性，可以保留文件原始结构和目录，并且具有一个 Web 用户界面，运行在 Linux 系统上，可以保证较高的抓取速度。

在文件格式上，考虑到后续处理的便捷性以及各种文件类型所占比例，选定以 PDF 作为主要抓取文件类型。

PDF 文件抓取完成后，需要对其进行筛选，保留其中的学术文献。使用基于规则的判定方法，即通过关键词进行判定。通过分析大量的学术文献，发现其特有的特征词包括 Abstract、Keywords、Introduction、Discussion、Conclusion 和 Acknowledgement 等词，不同的文献可能分别含有其中的几个词，因此可设阈值，根据上述词出现的数量进行判定。通过分析，笔者认为阈值设为 2 时即可实现判定。

2.2　海量网络学术文献词 - 文档矩阵处理

鉴于需要处理的文献数量较大，因此采取分布式处理方式进行词频矩阵生成。本部分使用 Hadoop[5] 实现，包括 HadoopNamenode 和 HadoopDatanode，其中 Namenode 负责并行处理的调度，Datanode 负责实际的并行处理工作。学术文献首先被读入 Hadoop 平台，在 Namenode 保存一份全部文件的索引，实际文件以冗余的形式保存在至少 2 部 Datanode 上，最后通过 Namenode 调用并行处理程序完成学术文献的词 - 文档矩阵生成，如图 3 所示：

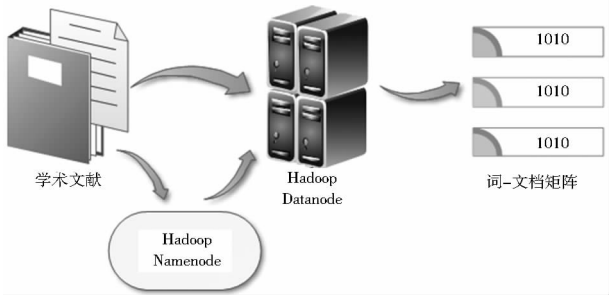

图 3　海量网络学术文献词 - 文档矩阵处理

在 Hadoop 的 Map 阶段，使用 StringTokenizer 依次提取文献中的词并生成一个 Key\Value 对 < 词，文献 ID > 。在 Hadoop 的 Reduce 阶段，同一个词使用一个 Reducer 进行处理，创建一个长度为文献数的数组保存当前词在相应文献中的词频，然后依次接受 Key\Value 对并对数组进行更新。在所有的 Reducer 工作完成之后输出矩阵。鉴于此矩阵为稀疏型矩阵，因此可以删除 0 位后输出 Sparse Matrix，减少存储空间。

2.3　本体集成

为了对自然语言进行理解，通行的方法是使用本体库对文本进行标注和集成。本部分主要使用

PROMPT，PROMPT 首先读入本体，然后分析概念之间的关系，将相同的概念进行映射，对于某一本体库中出现的特殊概念则予以保留，最终生成一个集成的综合本体，如图 4 所示：

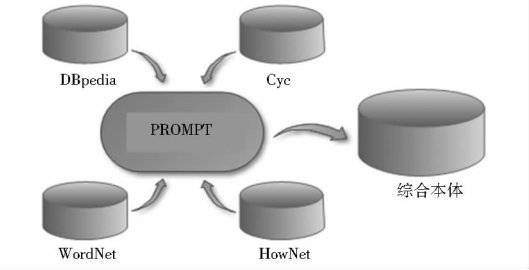

图 4 本体集成

在选用本体库方面，笔者做了大量工作，对比了 WordNet、DBpedia、Cyc、HowNet、TMO、UMLS Semantic Network、Gene Ontology 以及 Enterprise Ontoloy[6]，最终选择了 WordNet、DBpedia、Cyc 以及 HowNet，这些本体库都有官方（或第三方）开发的 Java API，便于集成。

鉴于本系统需要对文件进行分类，并且无法预知目标文献的类型，因此采用本体集成的方法将上述本体库集成为综合本体库，此综合本体库包含了上述所有本体的信息。在本体集成工具的选择上，考察了 PROMPT、OntoMerge、GLUE、OntoMap、Falcon - AO、OnMerge 等[7]，最终选择使用 PROMPT 作为本体集成工具。通过集成上述本体库，共得概念 12 万个，词汇数超过 20 万个，使用 RDF 描述语言进行表达，并设计了一套基于 JAVA 的 API。

2.4 基于语义驱动的分类

基于语义驱动的分类部分是本系统的核心。该部分主要包括映射系统 Projector 和 SVM 分类器。首先需要使用经过映射后的训练集对 SVM 分类器进行训练，训练完成后，则将 SVM 分类器应用于已被映射过的真实数据，最终通过分类器的判断输出分类结果（见图 5）。

在分类时，首先将词 - 文档矩阵映射至语义空间，本系统在此前研究的基础上[8-9]开发了基于 Java 的映射系统 Projector。通过调用本体库的 synonymy、antonymy、hyponymy、meronymy、troponymy、entrailment 关系对原空间进行映射并减少冗余，最终得到矩阵。

本部分选择 SVM（支持向量机）作为分类算法，选用较成熟的 LIBSVM 库作为分类的 Java 实现。通过对比得出基于 CVS（Concept Vector Space）的 SVM 明显优于基于 WVS（Word Vector Space）的 SVM 进行对比，鉴

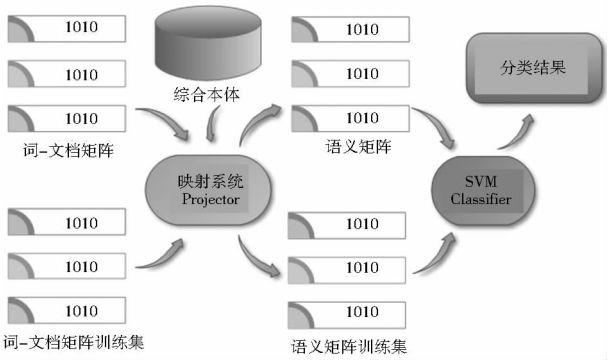

图 5 基于语义驱动的分类

于测试数据集上的结果中基于 CVS 的 SVM 的 F1 值高于 WVS 的 SVM，因此选用基于 CVS 的 SVM 的分类算法。

人工预设分类后，训练相应数量的 SVM 分类器，通过分类结果判定文献的类别从属关系。

在具体实现上，使用 LIBSVM 的 nu - SVR（回归型）模型，该模型训练后在预测时不会直接输出 1、0 结果，而是输出一个 0 到 1 之间的实数，对于可能出现的属于多个分类的文献，通过对比该实数大小可判定类别从属关系。使用线性核函数，Cost 属性设为 10，Cross - Validation 设为 5，其他属性使用默认值。

3 实 验

3.1 开发运行环境

硬件环境：x86 计算机 6 台，配置为 Intel 奔腾 D 3.0Ghz CPU、DDR2 1G 主内存、7200 转 1T 硬盘。

系统环境：6 台计算机全部配置 Ubuntu10.04.3，其中一台使用 Windows XP SP3 组成双系统。

软件环境：

Apache httpd 2.0.64（Linux）：Web 服务器；

Eclipse Juno（Windows）：Java 开发工具；

Hadoop 1.0.0（Linux）：并行处理平台；

Heritrix - 1.14.4（Linux）：数据抓取平台；

Java Runtime Environment 1.6.29（Windowns & Linux）：必要的 Java 运行环境；

LIBSVM（Java）：分类器；

MySQL 5.0（Linux）：数据库；

MySQL-connector 5.1.6（Java）：MySQL API；

PDFbox1.6.0（Java）：PDF 工具包；

PHP 5.3（Linux）：网页解释引擎；

PROMPT 2.3.2 for Protégé 3.0（Windows）：本体集成工具；

SQLite 3.7.9(Windows):文献元数据数据库;
Sqlitejdbc - v056(JAVA):SQLite API。

3.2　PDF 抓取与学术文献识别

在数据源方面,通过分析不同的目标源后发现,著名高校网站以及部分学科门户和 OA 仓储中含有大量公开发表的学术文献,并且可以不受限制地抓取,因此确定以高校网站、OA 仓储、学科门户作为目标源,为了使结果更具代表性,还加入了会议网站和研究者个人主页。本实验选择的目标站点如表 1 所示:

表 1　文献抓取目标站点

编号	站点	简介	类型
1	http://www.stanford.edu/	斯坦福大学网站	高校网站
2	http://www.omicsonline.org/	Omnics 集团网站	OA 仓储
3	http://www.acm.org/	美国计算机学会网站	学科门户
4	http://www.webis.de/research/events/pan-11/	国际会议 PAN 2011 年网站	会议网站
5	http://www.cs.columbia.edu/~mcollins/	Michael Collins 个人主页	个人主页

在文献获取过程中,目标站点网络环境的不同,造成了不同站点的采集耗费的时间相差较大,由于网络原因,可能造成实际抓取值与 PDF 的实际存在数量有差距。在获取完 PDF 文件后,使用 DirReader 读取全部 PDF,并调用 CheckPDF 进行学术文献判定。最终结果统计如表 2 所示:

表 2　海量网络学术文献自动获取结果

站点编号	PDF 数量(个)	有效文件数量(个)	学术文献数量(个)	采集耗时	采集时间
1	472 730	260 506	71 793	74 小时	2012-03-03
2	3 245	1 875	1 654	2 小时	2012-03-06
3	156 874	87 453	34 512	49 小时	2012-03-07
4	35	31	31	5 分钟	2012-03-10
5	69	67	59	10 分钟	2012-03-10
总计	632 953	349 932	108 049	125 小时	

3.3　词-文档矩阵生成

本部分使用 6 台计算机组成分布式处理平台,其中 Namenode 使用一台普通配置计算机,Datanode 使用 5 台部署了 Hadoop 的 x86 计算机。

在前期进行数据试验时,考虑到大学网站的文献类型具有多样性,因此选定斯坦福大学网站数据作为实验数据,分别使用 71 个、717 个、7 179 个斯坦福学术文献 PDF 进行矩阵生成,选择的方式为对全部学术文献 PDF 进行编号,然后分别对 1 000、100、10 进行求模运算,模为 0 者为实验文献。实验 1 共使用了 71 个文献 PDF 生成的词-文档矩阵,大小为 5.8MB,用时 3 秒;实验 2 共使用 717 个文献 PDF,生成的词-文档矩阵,大小为 333.8MB,用时 15 秒;实验 3 共使用 7 179 个文献 PDF,生成的词-文档矩阵,大小为 18.8GB,用时 137 秒;而使用传统的单机处理方法所用时间分别为 10 秒、60 秒和 1 716 秒。通过对比,可以看出分布式处理平台大大提高了运算效率,在真正的海量处理问题上可以发挥巨大的作用。

实验 4 对全部 108 049 个文献 PDF 生成的词-文档矩阵大小为 196.7GB,用时 958 秒。具体结果如表 3 所示:

表 3　词-文档矩阵生成结果

实验编号	文献数量(个)	矩阵行数	矩阵列数	矩阵大小	运算耗时
1	71	71	29 874	5.8MB	3 秒
2	717	717	176 982	333.8MB	15 秒
3	7 179	7 179	745 148	18.8GB	28 分 36 秒
4	108 049	108 049	925 479	196.7GB	6 小时 35 分 23 秒

3.4　矩阵映射和分类

本实验分类标准选用中华人民共和国教育部专业分类目录预设定的 12 个大类,分别是哲学、经济学、法学、教育学、文学、历史学、理学、工学、农学、医学、管理学、军事学。

在选定了分类标准后,随即进行训练集的分类标注。我们从斯坦福大学学术文献中随机选择了 1 000 个文献进行人工标注,并进行了矩阵映射,作为分类算法的训练集,人工标注邀请了 12 个学科领域的专家,共 10 人,其中哲学、法学为同一位专家,文学和历史学为同一位专家。对全部 1 000 个文献,每个专家进行了独立的 1/0 判定,即判定该文献是否属于特定的领域,属于为 1,不属于则为 0,人工标注结果见表 4。表 4 中的人工标注结果求和后大于 1 000 的原因是其中有部分文献被判定为同时属于两个或多个学科,最终实际结果也有此现象。

1 000 个训练学术文献 PDF 和 108 049 个学术文献 PDF 最终映射至语义空间的大小分别为 156M 和 87G 的未分类矩阵。

对训练语义矩阵进行标注后,共训练了 12 个基于 CVS 的 SVM 分类器,然后使用这 12 个分类器对全部学术文献 PDF 进行分类,最终结果见表 4。

4　海量网络学术文献自动分类系统的实现

4.1　系统特点

系统数据驱动部分采用 PHP + MySQL 技术,静态

表 4 分类结果

分类	人工标注结果(个)	最终结果(个)
哲学	36	3 255
经济学	49	4 352
法学	138	12 533
教育学	82	7 746
文学	104	9 748
历史学	114	11 315
理学	170	15 667
工学	205	26 111
农学	14	1 654
医学	48	3 830
管理学	118	9 574
军事学	21	2 263

展示部分采用 DIV + CSS 技术。使用 PHP + MySQL 可以在最大化节省成本的基础上实现最高的运算效率，对于海量的数据展示有着直接的效益。而静态展示部分使用 DIV + CSS 技术，既方便了系统的二次重构，同时也有利于商业搜索引擎的索引，从而能够更容易地实现本系统的应用价值。

系统所有元数据格式采用 DC 元数据标准，使用了 Title(标题)、Creator(创作者)、Date(日期)、Format(格式)、Source(来源)。存储格式采用 RDFS。鉴于没有做文本级别的深度挖掘，因此元数据不够充足，这是今后的工作重点。

所有的学术文献的分类结果都可以通过本系统查看。年份数据并未做文本级别的挖掘，而直接使用 PDF 的文件元数据(创建日期)。在文献展示页面上可以查看文献的标题、类别、原始 URL 以及正文的摘录。各界面都配有分面搜索功能，方便用户二次检索。

4.2 系统功能

海量网络学术文献自动分类的结果只有通过展示应用系统才可以被科研工作者利用，本应用系统的结构见图 6。该系统包括三种导航模式：

4.2.1 分类浏览 本系统可按照任一预设分类进行浏览，进入后可显示某一分类的全部文献。系统分类浏览(即首页)如图 7 所示。

4.2.2 分面检索 本系统提供了分面检索功能，可以依照论文多维属性进行检索，其中包括分类和年份，分面检索可以在首页、分类首页和搜索页面使用，提高检索效率。界面如图 7 右侧所示。

4.2.3 搜索 本系统可以进行标题搜索和全文搜索，检索页面可使用分面检索。

如需要查找 2000 年的管理学文献，则先从首页选

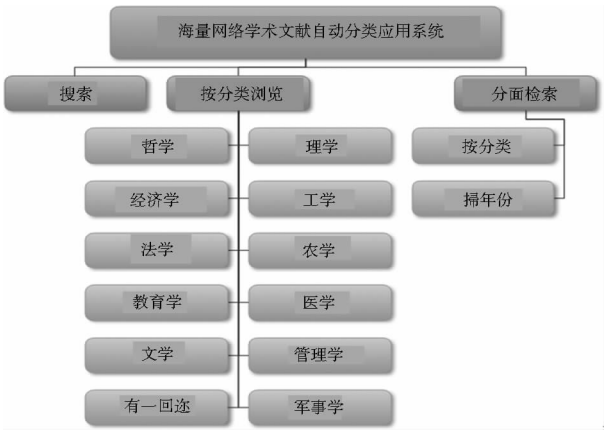

图 6 海量网络学术文献自动分类应用系统结构

图 7 海量网络学术文献自动分类系统展示平台首页

择管理学，进入管理学分类首页，然后使用分面检索将年份限定为 2000 年。

如需要查找含有特定关键词的文献，例如含有“3D Particle Path Integration”的文献，则只需要在首页的检索入口选择全文，然后在文本框中输入“3D Particle Path Integration”后检索即可，检索到的文献界面见图 8：

图 8 海量网络学术文献自动分类系统展示平台内容页

5　结　语

在充分调研了海量网络学术文献从自动抓取、处理到分类的整个过程的研究现状后，本文设计并实现了海量网络学术文献自动分类系统。为解决海量数据的处理问题，该系统使用了并行处理方法、本体集成方法以及基于语义驱动的分类方法，既有理论上的创新，也为实际解决海量数据处理问题提供了可行性方案。

该系统的成功设计与实现，既可以解决海量文献处理过程中面临的内存消耗大、处理速度慢、特征向量维度高等问题，让科研工作者有效获取并利用文献，同时也解决了两大异构本体库的集成及在具体领域如何应用的问题以及传统词向量空间因维度过高、缺乏语义而无法满足新一代语义 Web 环境下人们对海量网络信息资源语义分类、语义导航与语义检索的需求问题，因此具有学术价值和实践意义。该系统的设计思想和构架可以直接应用于电子政务系统、门户网站、垂直搜索引擎、数字图书馆网站等。

参考文献：

[1] Manning C D，Schuetze H. Foundations of statistical natural language processing [M]. Cambridge：The MIT Press，1999.
[2] Hadoop W T. The definitive guide [M]. US：Yahoo Press，2010.
[3] Miller G A. WordNet：A lexical database for English [J]. Communications of the ACM，1995，38(11)：39－41.
[4] 白如江，王效岳，亢丽芸. 基于 Heritrix 的网络学术文献获取研究 [J]. 图书情报工作，2012，56(11)：99－104.
[5] 亢丽芸，王效岳，白如江. MapReduce 原理及其主要实现平台分析 [J]. 现代图书情报技术，2012(2)：60－67.
[6] 白如江，于晓繁，王效岳. 国内外主要本体库比较分析研究 [J]. 现代图书情报技术，2011(1)：3－13.
[7] 于晓繁，王效岳，白如江. 本体集成方法和工具综述 [J]. 现代图书情报技术，2011(1)：14－21.
[8] 马范玲，胡泽文. 基于 SUMO 本体的图书自动分类模型研究 [J]. 情报杂志，2011(1)：168－173.
[9] 胡泽文，王效岳，白如江. 基于 SUMO 和 WordNet 本体集成的文本分类模型研究 [J]. 现代图书情报技术，2011(1)：31－38.

An Automatic Classification System of Mass Online Academic Literatures

Wang Xiaoyue　Bai Rujiang　Wang Xiaodi　Zhu Na
Institute of Scientific & Technical Information，Shandong University of Technology，Zibo 255049

[Abstract] With the development of the Internet，the amount of online academic literatures has been increasing exponentially，and it is difficult for science researchers to harness the power of the literature. It is necessary to develop a method for automatic acquiring，processing and classifying the literatures. This paper designs and implements an automatic classification system for massive online academic literatures based on the experimental researches done before. This system is a modular design which consists of four models of automatic fetching，term－document matrix processing，ontology integrating and semantics-driven classifying. This is proven that it can automatically accomplish the acquiring，processing and classifying online academic literatures.

[Keywords] academic literature　automatic classification　parallel processing