学科分类知识库的构建及其在网络资源分类中的作用

向桂林

(中国科学院文献情报中心 北京 100080)

[摘要] 提出利用知识库来对网络资源进行自动分类,讨论知识库中的规则体系,提出统计规则、上下文规则和经验规则,以及这些规则在分类中的作用。

[关键词] 知识库 自动分类 分类规则

[分类号] G254

The Construction of Knowledge Bases and Their Functions in the Classification of Web Resources

Xiang Guilin

(The Library of Chinese Academy of Sciences, Beijing)

[**Abstract**] This paper uses the knowledge base to classify Web resources, elucidates the construction of knowledge base and its application in classifying Web resources, emphasizes the rule system of knowledge base, and puts forward the statistic rule, the context rule and the empirical rule and their application in classification.

[**Keywords**] knowledge base automation classification classification rule

1 问题的提出

网络信息增长迅速,用户在网上寻找信息,犹如大海捞针一样。为了把网络信息有序地组织起来,目前可行的做法是,给定一个学科知识分类体系,把网络上关于该学科的网页、网站和其他电子资源都纳入这个分类体系之中,使用户能有序地浏览网络上关于这个学科的资源。本文以笔者正在从事的项目为例,选择数学学科为实验对象,来说明网络资源自动分类的解决方案。

2 相关的工作

关于网络资源的自动分类问题,国际上已有一些研究团体在进行这方面的工作,比如 WWLib[1] 提出的 ACE(Automatic Classification Engine),就是一个基于向量比较的分类引擎。它选定 DDC(杜威十进制分类法)的 10 个类,认为这 10 个类分别可以由 10 个 n 维的向量来表示,向量的每一维表示一个主题词对该类的贡献度。比如 DDC 的 700 类,可由("arts","decoration","composition","sculpture"……)来表示,且各词对 700 类的贡献度为(0.3,0.2,0.3,0.2,……),同时,认为每个网上资源也可以由一个 m 维的向量来表示,比如资源 R = ("arts","painting","Maecenas","Mona Lisa"……),各词对应的权重为(0.2,0.3,0.2,0.2……),然后借助于余弦相似度算法[2],计算出资源 R 与 10 个类的相似度,最后把 R 归入相似度最大的那一个类目中。国内有人提出了分类体系分类法(已经现存一套分类体系,然后把资

收稿日期:2002—07—12

源归入某一个类)和聚类分类法(假设初始时,没有一个分类体系,在大量的统计、学习基础上,聚集出若干类,然后对新的资源分类)[3]。现在最新的自动分类方法是基于本体论的分类方法,本文后面再作具体介绍。

3 学科知识库的构建

3.1 知识库的含义

知识库的含义有很多种,不同行业有不同的说法,比如在企业界,描述企业工序,管理各种生产资料的系统,被称为知识库;在计算机界,一个大的资料集合,含有源程序代码、帮助文件、FAQ等等,利于程序员自己学习和长进的系统,被称为知识库;在机械行业,把优选系统称为知识库。本文所说的知识库,则是指能完全或部分地代替人的脑力劳动,把网络上的信息资源准确地归到它应属于的类目中的系统。结合本文要讨论的问题,笔者认为知识库由概念体系、分类体系,以及能把概念体系映射到分类体系的规则体系构成。

3.2 知识库建设的原则

知识库的建设要遵循易用性和易维护性原则。易用性是指知识库的构造和数据组织要适应于将要解决的问题。对本文来说,就是要适应于把网络上的信息资源进行分类。

易维护性是指知识库里的知识与人脑中的知识一样,也会过时、老化,因此,就要对知识库进行更新,比如,添加新的术语,增加新的关系,删除过时的术语和关系等等。

3.3 知识库建设平台的选择

目前知识库的建设平台有两种流行的系统,一种是德国人研制的 Ontology Editor[4];另一种是美国斯坦福大学研制的 Protege 2000[5]。前者在欧洲国家用得比较多,但是不支持汉字,也不提供该平台的源代码;后者在欧洲以外的地区用得较广泛,它不仅支持汉字,而且还是用 Java 语言写成的,并提供该平台的全部源代码。同时,斯坦福大学还有专门的 Protege 2000 BBS 站点,上面有许多关于 Protege 2000 的插件,如与关系型数据库的接口插件,与 XML 的接口插件等。总之,选用 Protege 2000,一是因为该系统支持汉字,开发起来方便,能节省不少人力。二是因为它是用 Java 语言写成的,开发出来的知识库可以跨平台使用。

3.4 知识库组成部分及实现

笔者认为,知识库可以由3部分组成:概念体系、分类体系和规则体系(见表1)。

表1 知识库的组成

类	子类
ΘSystem Class	
ΘTerminology	
ΘTaxonomy	
	ΘTaxo-display
ΘRule	
	ΘStatistic
	ΘContext
	ΘEmpiricism

- 概念体系。由数学学科的主题概念及概念之间的关系构成。概念体系来自于数学学科主题词表。主题词表含有的"用"、"代"、"属"、"分"、"参"、"族"关系,集中反映了概念间的3大关系:①等同关系("用"、"代"),这是为了实现一个概念使用一个词来表述。它包含同义词,一个词的新、旧称呼,缩写与全称,近义词(可能是反义词)等关系;②等级关系("属"、"分"),表述词与词之间的上、下位关系,反映词的内涵、外延关系,比如"微分方程"就是"方程"的下位词,因为前者只是后者的一个特例;③相关关系("参"、"族"),表述了词与词之间的关联性,比如,"积分"与"数值方法"两个词之间有相关关系,因为"数值方法"是可以用来求"积分"的。概念体系的组织结构见表2。

表2 概念体系的组织结构

Name	Type	Other Facets
Reference	Instance	{Terminology}
Cname	String	
Count	Instance	
Sub	Instance	{Terminology}
Class	Instance	{Terminology}
Super	Instance	{Terminology}
Gestalt	Instance	{Terminology}
Ename	String	{Terminology}
Use	Instance	
Use-for	Instance	{Terminology}

- 分类体系。《美国数学文摘》和《德国数学文摘》编辑部合作出版了《数学主题分类表》,把数学学科分成63个大类,每个大类中又含有若干二级类目,二级类目中又下分为若干三级类目,……。笔者从事的项目采用的就是这个数学学科分类体系。如

果采用《中国图书馆分类法》(以下简称《中图法》)来做分类体系,是不合适的,因为《中图法》是一个通用的分类体系,并没有针对数学学科作更深入的划分。比如,数学在各行各业的应用,就被《中图法》分在了各个类目之中,而没有冠以专门的分类号,其实,采用杜威分类法,也有这个问题。分类体系的组成结构见表3。

表3 分类体系的组成结构

Name	Type	Other Facets
Class-father	Instance	{Taxonomy}
Edescription	String	
Class	String	
Cdescription	String	

国外有一些自动分类研究是基于本体来分类的,这些研究基本上是从杜威分类法中抽取概念,并形成概念与分类之间的联系[6]。笔者认为这种做法不妥当,因为分类体系与概念体系并不是一回事,前者不具有"用、代、属、分、参、族"的关系,为了弥补这样做的不足,有人提出同义词表、近义词表、相关词表等附属组成部分,搞得比较复杂,整个设计不简洁明了。

中国科学院文献情报中心编制的《数学汉语主题词表》[7],就是一部数学学科主题与分类合一的词表,既可以查到某主题词属于某个类(分类体系采用的是《数学主题分类表》);也可以查到某个类目下面含有多少主题词;还可以查到某个词的"用、代、属、分、参、族"关系。因此,笔者选用了《数学汉语主题词表》作为知识库概念体系的基础(实际上,还有相当一部分属于数学学科的概念,未被收入主题词表,而是被当成了关键词,排除在主题词表之外),选择《数学主题分类表》作为分类体系的基础(因为分类体系在不断地变化,这是由于数学学科在不断发展的缘故,我们要对分类体系作更新)。

● 规则体系。本文所说的规则,其思想来源于标引人员的标引过程。因为计算机目前还不具有思维能力,只能是按程序办事。为了让计算机更加逼近标引人员的工作过程,笔者假设标引人员也是按某些分类规则来对文献分类的。笔者要研究的内容就是把标引人员遵从的分类规则转换为计算机能理解的分类规则,从而让计算机也能自动对文献分类。尽管这种变换是形式上的,而非内容上的,但这是计算机走向智能化的第一步。笔者把规则划分为3个层次:统计规则、上下文规则和经验规则。

统计规则主要是统计"一词多类"的情况。有的主题词、概念词,可以属于一个或多个类目,当从文献中抽取出这种词时,标引人员一眼就能看出这个词在这种情况下属于哪一类,但机器看不出来。我们需要先把概念体系中的所有"一词多类"的词找出来,并在基于大规模的实验数据上进行统计,得出该词属于某个类的几率有多大,比如"语法分析"这个词:它所属的类目及其对类目的贡献度见表4。

表4 "语法分析"所属类目及其对类目的贡献度

所属类目	贡献度
03	0.4
49	0.2
68	0.4

大部分情况是:"一词归一类",那么这个词属于这个类的几率为1.0。可以把几率解释为词对类目的贡献度。统计规则是最基本的规则,目的是解决机器没有人类常识这个问题。统计规则的结构见表5。

表5 统计规则的结构

Name	Type	Other Facets
Term	Instance	{Terminology}
Term-class	Instance	{Taxonomy}
Frequence	Float	

其中,term、term-class的值可以取自于相应的类,而不能从相应的类继承而来,因为并不是每个词都属于多个类。

上下文规则主要用来解决多主题文献应该分到哪一个类目的问题。见表6所举的示例。

表6 多主题文献归类示例

文献中出现的概念词	文献应归到的类目
统计、概率	60
统计、热力学	82
统计、心理学	92
统计	62

多主题的另一种表达形式是:文献中根本未出现某个主题词A,但该文献的主题就是主题词A所表达的内容。如某文献中出现"垂钓"、"轻音乐"、"宾馆"、"房间"、"海滩"等词,要是单独看这几个词,是搞不清楚这篇文献在说什么,但经过人的思考、联想,可以得知:这篇文献的主题是"旅游",尽管"旅游"这个词没有在文中出现。

标引人员是基于自己的背景知识和对各主题搭配的理解进行分类的。通过与标引人员的讨论,笔者发现:这种组配关系是不能穷举的,不仅有两个主题的搭配,更有3个、4个甚至更多主题的搭配。标

引人员是靠思维能力来处理这种多主题搭配情况的，而不是靠"记住"来处理这种情况的。计算机虽然没有思维能力，但是在充分学习的情况下，可以表现出一些智能。常见的就是人工神经网络和推理网络。对人工神经网络的一般描述是：给定一个神经网络，向网络输入一个 n 维的向量，神经网络会输出一个 m 维的向量，输入向量与输出向量之间是一种非数学解析所能表达的函数关系，换句话说，神经网络所表达的函数关系，目前是不能用数学解析的办法来替代的(否则，神经网络也不会成为一门新的学科分支!)。所以，我们可以用神经网络来模拟人的思维(在特定的学科领域中)。神经网络在工作以前，需要大量地学习，以建立神经元之间的联系，学习得越多，这种联系就越牢固、越可靠，对新的输入能产生更准确的输出。具体来说，神经网络的输入可以是若干个主题词，假设用 1 000 个词作为神经网络的输入(这 1 000 个词来自于 467 个族首词，再从 3 000 余个独立词中挑出 500 余个词，原则是保证这 1 000 个词为数学学科的核心词，而且基本覆盖分类体系中的各个类目)，输出就是数学学科分类法的一级类目，共计 63 个类。我们的训练集是《中国数学文摘》，其文摘数据中不仅含有英文、中文摘要和英文、中文主题词，而且人工标引了分类号。上下文规则的结构见表 7。

表 7　上下文规则的结构

Name	Type	Other Facets
Context-Taxo	Instance	{Taxonomy}
Weight	Float	
Context-term	Instance	{Terminology}

上下文规则使用一个神经网络来记忆和联想主题词串与类目之间的联系。神经网络的输入为 1 000 个数学学科的主题词，输出为数学主题分类法的 63 个一级类目。神经网络训练完成后，存储的是一个权重矩阵，矩阵的列为 63 个一级类目，矩阵的行为 1 000 个核心词；

经验规则是用来否定机器给出的某些分类结果。因为"一定是……"这种断言很难给出，但是"一定不是……"这种断言是可以给出的。举个例子："微分方程"和"数值求解"是可以相关的，但"数论方程"和"数值求解"肯定不相关，因为前者的解是整数，后者讨论的是实数解。更通俗一点讲，假如机器给出了"11B、13C、65C"3 个分类号，但是根据经验，"11B、13C"和"13C、65C"这两种组合都不存在，或者说没有意义。因此，只剩下"11B、65C"这种组合。有人可能会问，为什么经验规则一上来就从类目这个层次开始，而不是从概念、主题词这个层次开始？这与知识库的使用很有关系，前面说了，知识库的建设要遵循两大原则，即易用性和易维护性。因为"给出分类号"这个步骤是分类的最后一步，而规则的使用是与分类的步骤相适应的。在数学主题分类表中，编制分类的专家已经指出了某个类与其他类有相关关系，这是笔者要用的第一部分数据；其次，我们有大量的由分类人员标引好的文摘数据，其中，有的记录被给出了 2 个甚至 3 个类号，这是类目与类目可以搭配的实例，这是笔者要用的第二部分数据。对上述两部分数据进行统计，笔者可以得到一个类目间可以产生搭配关系的列表(这个列表不会太大，不会是组合爆炸)。经过标引人员的仔细审查，可以确定一个组合搭配表。当然，如果统计出的搭配组合有漏、有错，都可以通过标引人员更正过来，这就是本规则被称为经验规则的原因，因为有人的因素参予了进来。经验规则的结构见表 8。

表 8　经验规则的结构

Name	Type	Other Facets
Empric-taxo-1	Instance	{Taxonomy}
Empric-taxo-2	Instance	{Taxonomy}
Empric-taxo-3	Instance	{Taxonomy}
Empric-taxo-4	Instance	{Taxonomy}
Empric-taxo-5	Instance	{Taxonomy}

经验规则是用来解决"哪些类能同现"这一问题的。它来自于对分类表和训练数据的统计，并由标引人员确认。本文假定只允许最多 5 个类同现。

4　学科知识库在网络资源分类中的作用

笔者建立的学科知识库，主要用于分类。笔者提出了一套分类办法，并考察了知识库在其中的作用。第一个作用是抽取概念词。笔者在前文讲过，仅仅按词频抽词，有些时候可能会漏掉某些主题。所以，抽取出来的词，要按知识库的概念体系把它们组织起来，让机器真正理解文献中有几个主题概念，这对正确分类至关重要。如果主题分析出错，后面的分类肯定也不会正确。比如：对"指数稳定性研究"这一题名，可以抽出"指数"、"数"、"稳定"、"稳定性"、"指数稳定性"等词，乍一看，有两个主题："数"与"稳定"(按词频)，事实上，这篇文献只有一个主题

概念"指数稳定性",造成这种混乱的原因是抽词的方法不对,我们应该优先抽取最专指的概念,因为越专指的概念,它属于的类目越少,对它分类也越方便。当专指概念也达到若干个时,我们才采取"往上位归类"的办法[8]。比如:我们抽出 B31、B32、B33、A21、A22、A23、C31、C32、C33、C34 10个概念,如果只按词频抽取,就有可能漏掉概念。词频统计见表9(括号中的数字表示词频,"下位词的词频向上位词累加"):

表9 词频与概念抽取的关系

A2(4)	B3(6)	C3(4)
A21(2),	B31(3),	C31(1),
A21(1),	B31(2),	C31(1),
A22(1),	B32(2),	C32(1),
A23(1)	B33(1)	C33(1),
		C34(1)

如果只按词频抽取,会漏掉 C3 系列(C31、C32、C33、C34)的概念。经过"往上位归类",文献的主题就凸现出来,这就是知识库在概念抽取中的作用。

第二个作用是指导分类。分类的过程可分为初级分类、统计规则分类、上下文规则分类、经验规则分类。本文说的类是指数学主题分类表中的63个一级类目。在分类之前,假设抽取出来的词已被归在了几个概念之中,而且知识库中的统计规则有表10所列出的条目。

表10 统计规则示例

贡献度 / 类目 / 概念	80类	01类	90类	62类	13类
A2	1.0				
B3		0.3	0.4		0.4
C3	0.5			0.5	

● 初级分类,就是抽取出来的词属于同一个概念。这可以直接分配一个类号,从而结束分类。

● 统计规则分类,就是抽取出来的词不属于同一个概念,但这些词可能属于同一个类,也可能不属于同一类,这就需要基于统计的规则来分类。以上面的例子(表9、表10)为例,对含有主题概念 A2、B3、C3 的文献,按统计规则分类,分类结果见表11。

这样一来,我们就可以认为,该文献最可能属于80类,但也有可能属于90类、13类。如果基于统计规则分类后,得到的类号非常明显(该类号比其他类号有明显的重要性),就可以结束分类。

● 上下文规则分类。我们可能会遇到这样的情况:经过统计规则分类后,得到的类号具有相等的重要性,这种情况极可能发生在前面举的"旅游"那个例子中。这时候,就得依靠上下文规则来分类。我们把经过概念抽取后得到的概念词(有的已经是1 000个核心词之一,有的不是,如果不是1 000个核心词之一,就参照概念体系,把它引导到核心词上!)输入神经网络,通过神经网络的计算,得到一个或几个分类号。如果只得到一个分类号,就结束分类。如果得到若干个分类号,就再进行基于经验规则的分类。

表11 统计规则在分类中的使用方法

本文献属于某类的可能性	计算方法
80类 = 6	4*1.0 + 4*0.5
01类 = 1.8	6*0.3
90类 = 2.4	6*0.4
62类 = 2	4*0.5
13类 = 2.4	6*0.4

● 经验规则分类。如果神经网络的输出含若干个类目,就得用类目搭配表检查一遍,通常有两种情况:第一种情况是刚好符合类目搭配表中的某项,就作为正式输出;第二种情况是与类目搭配表中的条目不匹配,这又有两种情况,一是设 a、b、c 为神经网络输出的3个类目,a、b 是类目搭配表中的某项,那么,分类系统的输出将是 a、b;二是设 d、e、f 是神经网络输出的3个类目,c、d、e、f 是类目搭配表中的某项,则分类系统的输出将是 d、e、f。

5 结 论

知识库既可以用来对网络资源作分类标引,也可以用来对资源作主题标引。但是本文涉及的分类较粗,适合于做资源导航,不适合于为资源指派一个较专指的分类号,这是以后需要改进的地方。在主题标引方面,本文讨论的知识库已经具有比较完善的解决方案。

知识库还可以辅助检索。有人提出的"概念联想"[9],就是一种提高检准率、检全率的好方法,不过该文提出的同义词典、近义词典等,太复杂,如果构建了知识库,问题就好解决了。

本文提到的上下文规则,实际上指的只是一种概念上下文规则,事实上,还存在一种语义上下文规则[10]。比如,如何判断一个网页是否在讨论计算机专业的课程呢?如果对每一个网页过滤一遍,漫无

目的地找，效率极低，提高效率的办法，就是设计一个语义上下文规则，比如，网页中出现有“computer science”、“teached by ”、“instruction” 等词，我们就可以认为该网页在讨论计算机专业的课程问题。这就比概念上下文规则更有效。同时，我们应看到，语义上下文规则只有在解决特别具体的问题时才有效。不过，这是知识库今后的一个发展方向。

参考文献：

1 The Wolverhampton Web Library. Automatic classification of Web resources using java and deway decimal classification. http://www.scit.wlv.ac.uk/~ex1253/classifier/

2 Fabio Crestani etc.. WebSCSA: Web Search by Constrained Spreading Activation. IEEE Internet Computing, May./Jun. 1999:163-167

3 李 勇. 网络文本数据分类技术与实现算法. 情报学报,2002,21(1):38-41.

4 http://img.cs.man.ac.uk/oil

5 http://protégé.stanford.edu

6 Rudy Prabowo, Mike Jackson, Peter Burdon. Ontology-based automatic classifier for classifying the Web pages. Proceeding of ETCE2002, ASME Engineering Technology Conference on Energy, February 4-5, 2002, Houston, TX, 1-12

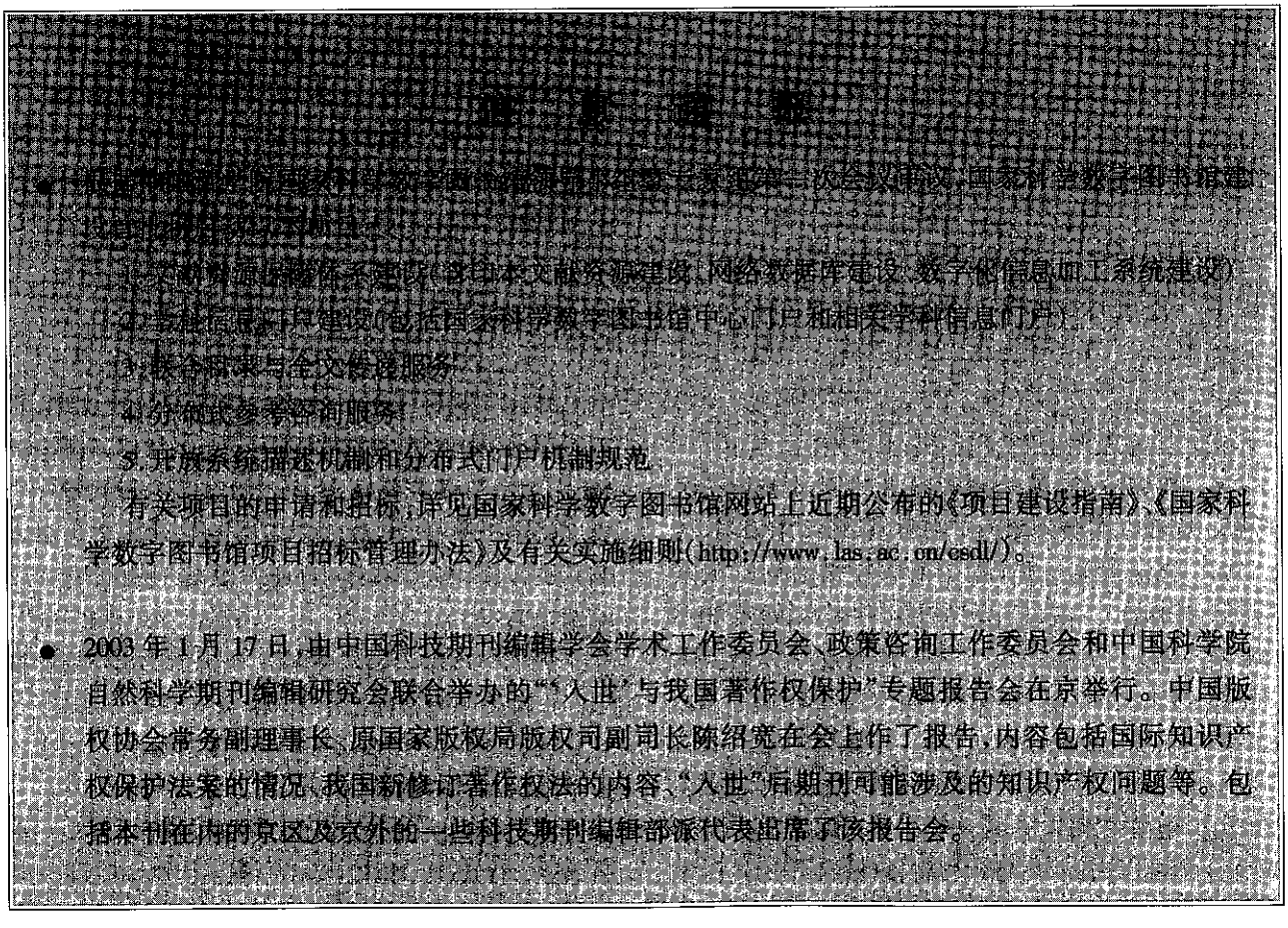
7 王声培等. 数学汉语主题词表. 上海:上海教育出版社,1994

8 韩客松. 中文全文标引的主题词标引和主题概念标引方法. 情报学报,2001,20(2):102-104

9 王兰成. 基于中国档案主题词表的自动标引控制研究. 情报学报,2002,21(2):54-57

10 Mark Craven, Dan Dipasquo. Learning to construct knowledge bases from the World Wide Web. Artificial Intelligence, 2000 (118):69-113

[作者简介] 向桂林，男，1971年生，博士生，发表论文多篇。



3. 开放系统描述机制和分布式门户机制规范

有关项目的申请和招标，详见国家科学数字图书馆网站上近期公布的《项目建设指南》、《国家科学数字图书馆项目招标管理办法》及有关实施细则(http://www.las.ac.cn/csdl/)。

● 2003年1月17日，由中国科技期刊编辑学会学术工作委员会、政策咨询工作委员会和中国科学院自然科学期刊编辑研究会联合举办的“‘入世’与我国著作权保护”专题报告会在京举行。中国版权协会常务副理事长、原国家版权局版权司副司长陈绍宽在会上作了报告，内容包括国际知识产权保护法案的情况、我国新修订著作权法的内容、“入世”后期刊可能涉及的知识产权问题等。包括本刊在内的京区及京外的一些科技期刊编辑部派代表出席了该报告会。