基于概念属性聚类的多视角知识组织系统研究初探

时晨，朱礼军
（中国科学技术信息研究所，北京 100038）

摘要：文章在传统单视角知识组织系统的基础上，提出了多视角知识组织系统的概念与框架，即在传统知识组织系统的构建和服务过程中增加知识的适用语境——视角，并对视角的概念与表达作了简要阐述。此外，针对视角的构建方法，文章提出了基于属性聚类的视角生成方法，并用自建的金融语料与属性列表进行了实证研究，成功抽取出多个视角。

关键词：多视角；知识组织系统；多视角聚类；属性聚类
中图分类号：G254
DOI：10.3772/j.issn.1673—2286.2014.09.05

信息经济与知识经济时代，随着信息技术特别是数据库技术的发展，企业在运营过程产生的大量数据得以存储，但是隐藏在海量数据中的知识并没有得到有效的发掘与利用。

1 现有知识组织系统存在的问题

知识组织系统是为了应对企业机构信息组织需求、提升企业机构知识管理水平而产生的，是对知识进行结构化、系统化表达和阐述的各类工具的总称，比较常见的有叙词表、分类体系、本体等等。有学者曾将知识组织系统分为三大类[1]，如图1所示：

（1）词单（Term Lists），包括同义词环、权威文档、词汇/字典等；

（2）分类体系（Classification），包括图书分类法、知识分类表（taxonomies）、标题表等；

（3）关联组织（Relationship Groups），包括本体、语义网络、概念地图和叙词表等。

对多数知识组织系统而言，构建者并未关注知识视角的问题，但在现实应用中，这个问题却无法避免。对任何一个研究对象而言，从不同的视角上看，具有不一样的属性集合和相关关系集合。例如，能源领域的科学家，对概念词汇“汽油”，更关注汽油的加工、传输、

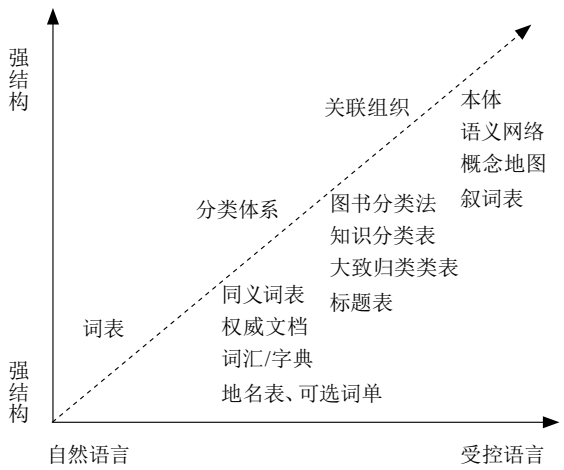

图1 知识组织系统（KOS）一览[1]

保存相关的属性；经济学家更多关注汽油的价格、市场方面的属性；而环境科学家则更多关注汽油在生产和燃烧过程中，产生的二氧化碳气体、造成的环境污染等相关属性。此外，即使在同一个领域，由于概念词汇构建者本身的主观差异，也会造成同一个概念词汇的属性和相关关系集合不一致的现象[2]。

传统的知识组织系统在遇到知识视角这个问题时，一般通过注释等方法加以处理，并未在构建以及服务环节作出系统的应对策略[3,4]。此外，在知识协同构

建过程中，传统做法往往通过专家审定的方式，来试图消除不同构建者的主观差异。而专家审定的结果，只是代表了某一个特定视角和层次来观察的权威知识，舍弃了其他部分实际上具有价值的知识。针对传统的处理方式，可能会造成三个方面的严重后果。一方面，在构建时，大量在某个视角上（语境中）成立的有效知识，被其他视角的审核专家生硬地错判为无效知识，造成知识损失；一方面，在组织时，不同视角的知识交叉混合在一起，没有明确地标识出来其视角信息（适用语境），造成知识体系组织混乱；另外，在服务时，组织的混乱也导致无法根据用户的个性化需求，提供其准确和必要的个性化知识。

为此，本文基于属性聚类的方法，提出了多视角知识组织系统。它力图在系统构建伊始，就通过属性聚类生成知识视角，服务于后续构建多视角知识组织系统或者传统知识组织系统多视角化过程。

关于知识组织系统的视角，知识工程和图书情报领域的专家都在各自的研究中有所提及。在知识工程领域，出于对知识有效性和可靠性要求，很早就开始了对知识库中的规则系统的视角和粒度分析的研究。Mehrotra等[5]认为，一个单一结构化方法或者抽象分类不足以理解一个复杂的知识库系统，由此提出了MVP-CA、多视角聚类分析的思想。MVP-CA提供了一种在规则库中发现层次结构和视角结构的方法，在规则库中进行聚类分析，得出多个相应视角。Acker和Porter[6]开发视角检索器（view retriever），用来从知识库中识别出不同的视角，如“as kind of”视角。它还能发现一些基本的视角，例如结构组成、感知、功能、时态、行为、过程、类别层次、因果联系等。在图情领域，一些传统的知识组织方法在一定程度上已经注意到了知识的视角和粒度问题，例如分面组配法，例如，Kingston[7]通过研究发现，在对一个领域知识进行自然分类的过程中，在顶层类目或者二级类目划分时就常常会发生多视角的现象，以计算机科学分类（ACM分类）为例，如表1所示，在一级类目中，就具备从what、how和why三个视角进行的分类。

1995年左右，科学院计算所曹存根研究员，提出国家知识基础设施（National Knowledge Infrastructure，简称NKI）构想，其中提到的问题，包括研究知识的各种操作，如知识抽象、求精、求同、求异、组合、分解等，为知识服务提供必要的基本设施（国家知识基础设施的意义）。在今天看来，这些问题依然具有相当的挑战性。

表1 ACM一级范畴的多视角现象[6]

	What	How	Why	When	Where	Who
Computer applications What goes inside a computer	Computer Applications Hardware Software	Computing Methodologies Computer Systems Organization Data, Information Systems	Computer Milieux			
Theoretical level		Theory of Computation Mathematics of Computing				

2 多视角知识组织系统

多视角知识组织系统在传统知识组织系统的基础上，引入视角的概念，通过对视角的定义、表达以及自动构建，从而完成对传统单视角知识组织系统的多视角化改造。视角作为知识的适用语境，由基于聚类的机器学习方法自动生成。

2.1 整体框架

在关系型知识组织系统中，我们可以将其核心知识体系抽象为一个四元组，CK={C,A,R,Val}，其中C={C_1,C_2,…,C_n}代表概念集合（亦称作“术语”—term或者“类”—class），A={A_1,A_2,…,A_n}代表属性集合，Val={ Val_1, Val_2,…, Val_n}代表属性值集合，R={R_1,R_2,…,R_n}代表相关关系集合。如图2所示，对于传统的单视角知识组织系统而言，所有概念和属性关系都罗列在系统中，并未对知识的适用语境（视角信息）加以标注，导致这种知识结构固定僵化，适用领域有限，灵活性低，可移植性与复用性较差。

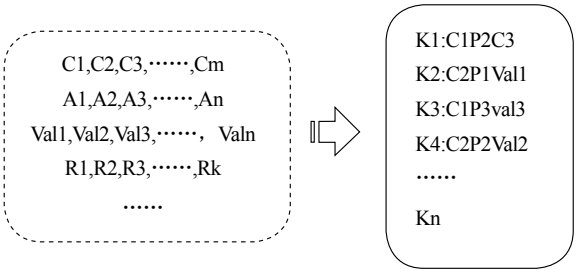

图2 单视角知识组织系统

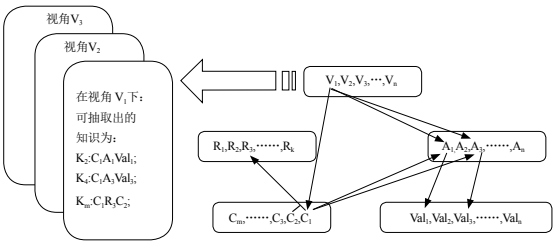

图3 多视角知识组织系统

相比较于传统的知识组织系统，多视角知识组织系统引入了视角的概念，视角代表了对于知识的语境表达，相当于为每条知识增加了应用环境，其下再包含不一样的概念、属性、属性值和关系集合。视角之内，知识具有相似的语境，结合起来能提供面向某个细分领域或是某种职能的全部知识；视角之间，存在着两种关系：完全独立和包容型关系。独立关系是指两个视角下的集合不存在任何的层次关系。包容型视角，是指一个视角下的元素集合是另一个视角的子集。通过视角描述或视角说明，用户能够区分不同视角的用途并快速选择自己所需的视角。如上图3所示，$V=\{V_1,V_2,V_3,\cdots,V_n\}$代表视角集合，假设视角$V_1=\{C_1,C_2;A_1,A_2;R_3\}$，通过知识映射和滤取等操作，视角$V_1$下可以抽取出知识$K_2$、$K_4$和$K_m$，而$V_1$也是对$K_2$、$K_4$和$K_m$的应用语境表达。

2.2 视角构建方法

关于视角的构建方法，由于事先无法得知每个视角的描述，也无法确定视角生成的依据，利用聚类这种无指导的机器学习方法可以较好地解决这些难题。但是，聚类由于缺少人工监督，带有一定盲目性，由聚类直接而得的视角往往与用户视角存在一定差距，所以在语料收集阶段，我们引入存在按类整合后的文档体系作为语料，在一定程度上，能够控制聚类的收敛过程。

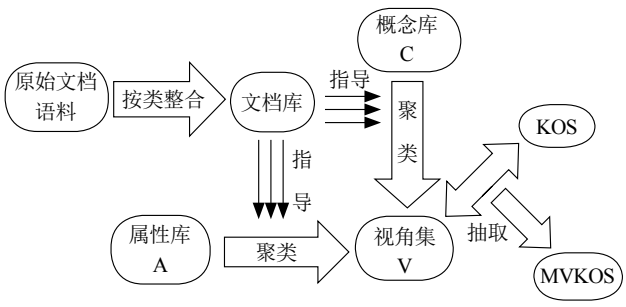

图4 视角构建流程

如图4所示，视角的构建可以分为四个模块：文档语料按类整合、概念聚类、属性聚类以及后续的视角知识关系抽取过程。

（1）文档语料按类整合。知识组织系统中包含着大量的文档资源，它们构成了KOS最原始的语料库，概念或属性都是基于它抽取而来。每一个文档都具有一个分类号（一般为中图分类号），整合相同分类号下的文档，使每个主题形成一篇大文档，当有些分类号过细，其下文档过少时，可适当向上合并。

（2）属性聚类。属性是概念词汇的特征，本身带有明显的应用性，它们的组合可以指征出各种适用语境。利用属性聚类，可以自动学习出KOS的各个视角，而基于分类整合后的文档语料库实施聚类则可以更好地控制聚类过程，减少因语料分散零碎、特征矩阵维度过大、过稀疏造成的聚类误差。

（3）概念聚类。基于按类整合的文档语料库，利用基于统计的方法提取特征矩阵，再根据一定的算法对概念实施聚类，由于是在同一语料库指导下进行的聚类，聚类结果可以与属性视角相关联，从而可以帮助所有概念找到视角归属。

（4）视角知识关系抽取。经过属性聚类与概念聚类过程，视角模型初具规模，后续主要是将原KOS中的知识和关系映射到各个视角之中，从而将KOS多视角化，这也是我们下一步的研究内容。

在四个模块中，属性聚类是视角生成的关键步骤，在后文中，本文将基于一定的数据进行实证研究。

3 基于共词的属性聚类

视角作为视物的角度，本质上就是无穷无尽的，以人工或者以任何一种机器学习的方法进行构建都不能够得到全面、完整的视角，这是一个长期积累的过程。对KOS的多视角化过程来说，要在考虑应用性和成本的基础上，从用户常用视角出发，逐步完善KOS的视角构建。为此，本文基于传统共词分析的思路，提出了一种实用的视角构建方法——基于共词的属性聚类，它能快速、准确地挖掘出最常用的视角。基于共词的属性聚类包含以下步骤：属性共词矩阵的生成与处理、距离测度以及利用系统聚类法实施聚类。

3.1 属性共词矩阵的生成与处理

传统的共词聚类一般通过分析词对出现的频率，以一定的聚类算法，把原本复杂的共词网络转换为若

干类团，从而更加简洁直观地表示出来[8]。共词分析法被普遍用于主题间的关系和领域热点研究，一般认为，词汇对在同一篇文献中共同出现的次数越多，它们之间的关系就越密切。拓展到属性，如果属性对在同一文献中同时出现，也能说明它们之间存在着一定的关系，并且关系的亲密程度也会随着属性对出现的次数而不同。假设属性集合为A={a_1,a_2,a_3,…a_i…,a_n}，对于其中的任意一个元素a_i(i为正整数，0<i<=n)，如果是基于共现的方法，则a_i的特征向量T_i可以表示为：

$$T_i=\{f_i^1,f_i^2,f_i^3,\cdots f_i^j\cdots f_i^n\},\ j\text{为正整数，}0<j<=n$$

$$f_i^j=f_j^i$$

其中，f_i^j代表属性a_i和a_j共同出现的文档数，由于a_i和a_j出现的次序并不影响其共现的次序，故f_j^i与f_i^j相等。

根据上述公式，计算n个属性两两共现的频次，得到n维的共现矩阵。由于属性众多，分布不均，频次相差可能较大，为了消除频次悬殊给结果带来的偏差，引入Ochiia系数[9]，即矩阵中的每一个数都除以相关两个属性总频次乘积的开方，Ochiia系数计算公式如下：

$$\text{Ochiia系数}=f_i^j/\sqrt{f_i*f_j}$$

其中f_i和f_j分别代表属性a_i和a_j总共出现的文档数，经过处理后得到相关矩阵和相异矩阵（1减去相关矩阵）。

3.2 距离侧度与聚类算法选择

层次聚类法也叫系统聚类法、等级聚类法，是常用的词聚类方法。总体而言层次聚类分为自底向上的凝聚层次聚类法和自顶向下的分裂层次聚类法。而聚类过程中的距离测度包括点与点之间的距离以及类与类之间的距离。一般的K-means算法仅仅使用欧式距离作为度量，而层次聚类法提供了多种点与点以及类与类之间的距离测度方法，其中点与点之间的距离测度方法包括欧式距离法、欧式距离平方法以及余弦距离法等等。同时层次聚类法也具有丰富的类间距离度量方法，比如最长距离法（complete linkage）、最短距离法（single linkage）和中间距离法（median method）等，此外，层次聚类法也具有多种结果表示方法，如树状图、饼状图等。本文中，选择凝聚法作为特征矩阵聚类算法，点与点之间的距离选择最为常用的欧式距离方法，类间距离选择能兼顾到类中所有点的中间距离法。聚类过程中，它将n个属性看成n类，然后将相似度最大（距离最小）的两个类合并为一类，所有属性变成n-1类，再从n-1类中找出相似度最大的两个类合并，变为n-2类，依次下去，直至类簇总数变为m时使聚类的准确率最高。其中聚类准确率是指正确归类的属性数（人工判断）占属性总数的比例。每一个类簇代表着一个视角，分析每一个视角的属性组成，判断其是否具有一个明显的主题特征，如果是的话，根据这个视角的主题特征添加视角描述，否则将这个视角定义为模糊视角。

4 实证研究——基于金融培训语料的视角生成过程

在这一章节中，为了验证属性聚类形成视角的可行性，文章将基于一个自行构建典型金融领域语料库与属性列表，通过属性聚类的方法，进行生成视角的尝试。

4.1 数据准备

为了模拟按类整合的语料库，在数据准备阶段，基于主题概念分类体系，分别检索关于各个主题的文档语料。

（1）金融领域典型概念分类体系

本文首先依据《金融公文主题词表》、《中国农业银行公文主题词表》和《信用社（银行）公文主题词表》建立了金融领域典型主题概念体系，其中一级分类8个涵盖票据结算、外汇、存款、贷款、债券、基金、机构法人、电子银行，二级分类由于存在多种分类标准，总共有256个，三级分类110个。

（2）语料库

语料库可以模拟语言应用环境，为了进行属性抽取与聚类分析，必须进行语料库的构建。语料库主要有三个来源：搜索引擎、学术数据库以及金融机构网站资源。来源于搜索引擎资源的特点是用户角度、类型多样，通俗性高。而学术数据库里面的文献是金融领域专家的研究成果，专家角度，专业性高，但多是底层研究，与前端业务尚有一段应用时差。金融机构网站，含各大银行网站，这部分语料包括产品业务介绍、用户指南、招聘信息等，是与前端具体业务关系最为密切的一部分资源，它的特点是银行角度、应用前沿，标准性与应用性高。

（3）属性列表

在依据属性抽取工具从语料库抽取出的属性列表中，本文选择100个在语料库中出现频率较高的典型属性作为聚类实验对象。

4.2 聚类过程

统计100个属性词在语料库中两两共现的次数，生成特征矩阵后，借助Ochiia系数将其转变成100×100的相关矩阵和相异矩阵，将矩阵输入SPSS 19.0，进行系统聚类，如图5所示，得到层次比较明显的树状图（部分），经过整理，最后得到的聚类结果如表2所示。

4.3 结果分析

从聚类结果看，将属性词分15个类时，聚类准确率最高。关于效果的评价，选取两个指标，聚类准确率与视角精确率，其中视角精确率是指精确视角（非模糊视角）数占视角总数的比例。前者反映了聚类方法本身的有效性，后者反映了属性聚类方法生成视角的可行性。经过计算，聚类准确率为91%。如表2所示，从视角表征度看，在15个视角中，13个视角具有比较明确的主题特征，只有视角5（注册资本、发起人）和视角14（金额、名称、有效期）两个视角类目过少且没有比较直观的视角指向性，所以它们属于模糊视角，则视角精确率为86.7%。聚类的准确率与视角精确率均比较高，充分证

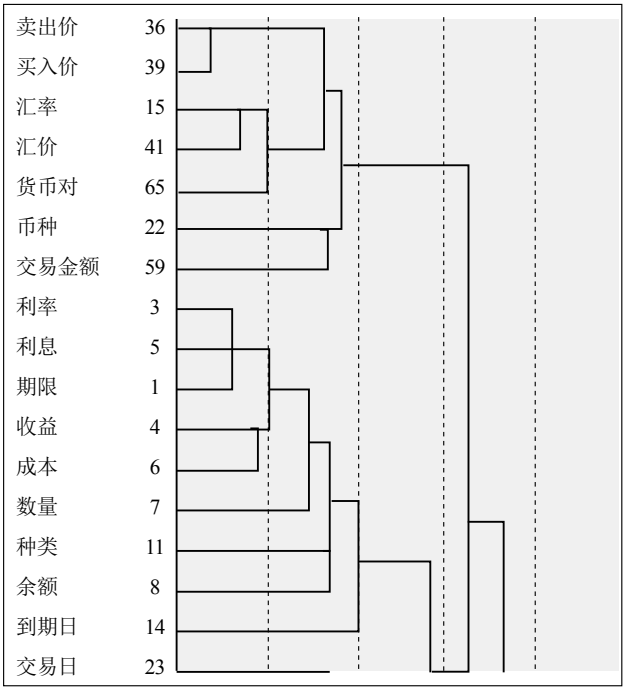

图5 属性聚类树状图部分

表2 属性聚类结果与视角指向

类/视角	项目	视角描述
视角1	卖出价、买入价、汇率、汇价、货币对、币种、交易金额	外汇交易
视角2	利率、利息、期限、收益、成本、数量、种类、余额、到期日	理财产品
视角3	交易日、交易方式、专业、质量、财产、功能、电话、营业网点、限额、城市	理财交易
视角4	面额、偿还期、偿还方式、票面利率、债券价值、股价、报酬、股息、所得税	股票债券买卖
视角5	注册资本、发起人	模糊视角
视角6	发行人、发行方式、发行价格、净资产、净利润、存续期、发行额、信用等级	股票发行
视角7	汇款人、汇出行、结算方式	汇款视角
视角8	申请人、法定代表人、担保范围、担保方式、授信额度	信用卡申请
视角9	手机号、登录密码、密码、户名、积分、增值服务、挂失手续、证件号码、证件类型、卡号	电子银行
视角10	存期、起存金额	存款视角
视角11	涨幅、股票名称、股票代号、基金名称、基金净值、职务、市值、总资产、净额	股票基金买卖
视角12	支付方式、贴息、利息税、贷款额度、期数、职称、年龄、性别、国籍	贷款视角
视角13	代理业务、服务项目、服务对象、利润率	服务视角
视角14	金额、名称、有效期	模糊视角
视角15	收款人、付款人、背书人、被背书人、出票日期、单位名称、最高金额、身份证号码、开户行	票据视角

明属性聚类生成主要视角的方法是可行的。在现实应用中，视角之间可能存在着重叠，经过低频词过滤后，基于属性共词的聚类方法尽管不能挖掘出所有完整的视角，但是它却能够快速地挖掘出常用视角，为后续的视角建设提供了基础，结合其他视角构建方法，借助专家审核与控制，通过视角间的合并与减去等操作，逐步进行视角的补充与完善。

由于我们的语料是按照概念主题检索而得，我们所获得的视角也多与概念主题相关，如“外汇交易视角”对应“外汇主题”，“票据视角”对应“票据结算主题”，“股票债券买卖视角”对应“股票主题”和“债券主题”等等。由此推断，如果语料来源于金融机构内部的各个岗位，那所得的各个视角是否也能对应到相应的岗位？由于这部分视角对于企业内部构建按岗位服务的多视角知识组织系统是极为重要的，所以这个问题也值得进一步研究。

5 结语

本文提出了多视角知识组织系统的概念，在对视角定义作了简要阐述的基础上，阐述了基于按类整合的文档语料库，利用聚类构建视角的方法，重点提出利用属性聚类生成视角的方法。在实证分析中，利用自建的分类语料库以及属性词列表，基于属性共词的系统聚类法，得到15个视角，其中13个为主题明确的视角，验证了属性聚类生成视角的可行性。基于本文的研究，如何利用所得视角对已有知识组织系统实施多视角化过程，即如何通过映射、滤取等操作得到各视角下的知识和关系将是下一步的研究内容。

参考文献

[1] 曾蕾.用于标引、浏览、检索的语义工具Semantic Tools [EB/OL]. (2010-09-11) [2014-02-23]. www.libnet.sh.cn/upload/htmled-itor/File/071213121547.pdf.

[2] ZHU Lijun, SUN Fengjun, OU Jie. Multi-Viewpoint Based Dimension Control to Knowledge Organization System [J]. ICIC Express Letters, Part B: Applications, 2012, 3(6): 1403-1408.

[3] 王琳,赖茂生.信息集成的领域分析研究[J].图书情报知识,2007(3):5.

[4] 董慧,余传明,杨宁,等.基于本体的数字图书馆检索模型研究(III):历史领域资源本体构建[J].情报学报,2006,25(5):11.

[5] MEHROTRA M, WILD C. Multi-viewpoint clustering analysis [C]// Proceedings of the Goddard Conference on Space Applications of Artificial Intelligence, 1993.

[6] ACKER L, PORTER B. Extracting viewpoints from knowledge bases [C]// Proceedings of the Twelfth National Conference on Artificial Intelligence Seattle, Washington, United States, 1994. American Association for Artificial Intelligence.

[7] KINGSTON J. Ontology, knowledge management, knowledge engineering and the ACM classification scheme [C]// Proceedings of the Proceedings of ES’ 02, the 22nd Annual International Conference of the British Computer Society's Specialist Group on Artificial Intelligence, Peterhouse College, Cambridge, December 10-12 2003. Springer-Verlag New York Inc.

[8] 钟伟金,李佳,杨兴菊.共词分析法研究(三):共词聚类分析法的原理与特点[J].情报杂志,2008,27(7):118-120.

[9] 李长玲,雪梅.我国情报学硕士学位论文的共词聚类分析[J].情报科学,2008(1):73-76.

作者简介

时晨，女，1990年生，中国科学技术信息研究所硕士研究生，研究方向：知识组织与数据挖掘，E-mail：shichen2012@istic.ac.cn。
朱礼军，男，中国科学技术信息研究所研究员，通讯作者，E-mail: zhulj@istic.ac.cn。

Research on Multi-view Knowledge Organization System Based on Clustering of Concepts' Attributes

SHI Chen, ZHU LiJun
(Institute of Scientific and Technical Information of China, Beijing 100038, China)

Abstract: Based on the traditional single-view knowledge organization system (KOS), the article proposes the concept and architecture of multi-view KOS. The multi-view KOS introduces the "view" in the process of its construction and service, which represents the knowledge's context. Still, the concept and expression of the "view" are described. In addition, the article proposes the method of concept attributes' multi-view clustering to construct "view" and its feasibility is proved with several effective views extracted using self-built financial corpus.

Keywords: Multi-view; KOS; Multi-view clustering; Attributes' clustering

（收稿日期：2014-08-18）

基于概念属性聚类的多视角知识组织系统研究初探

作者： 时晨， 朱礼军， SHI Chen， ZHU LiJun
作者单位： 中国科学技术信息研究所,北京,100038
刊名： 数字图书馆论坛ISTIC
英文刊名： Digital Library Forum
年，卷(期)： 2014(9)