●马张华

论中文信息动态自动聚类的特点和方法体系

摘 要 与传统静态聚类系统相比,动态自动聚类系统有以下特点:聚类是动态进行的,它是在检索结果返回的基础上进行的实时操作;每次聚类的文献对象数量有限;用来作为聚类依据的文献数据只是文献的局部;参与聚类的资源在整个资源集合中的分布是随机的。动态自动聚类方法有:直接将专指性短语作为揭示类目相似性识别的依据;更多使用线性聚类策略;使用等级显示、多维聚类的形式;采用优化算法;扩大预处理的应用。表 1。图 1。参考文献 12。

关键词 动态自动聚类 专用切分词典 中文自动分类 网络资源组织

分类号 G254

ABSTRACT Compared wth traditional static clustering, dynamic automatic clustering has the following characteristics: real-time dynamicness, limited number of documents to be clustered every time, localness of document data to be used for clustering and randomness of the distribution of resources to be clustered in the whole resource set. In this paper, the author summarizes some methods for dynamic automatic clustering, such as using special phrases, using more linear clustering strategies, using hierarchical display and multi-dimensional clustering, using optimization algorithm and expanding the application of pre-processing. 1 tab. 1 fig. 12 refs.

KEY WORDS Dynamic automatic clustering. Special segmentation dictionary. Chinese automatic classification. Network resource organization.

CLASS NUMBER G254

1 研究动态自动聚类的意义

动态自动聚类是一种在文献自动聚类与关键词检索结合的基础上发展起来的自动聚类形式。与传统自动聚类系统不同,这种方法以检索返回的资源为处理对象,选取其中的一部分,对其自动聚类并加以显示,用以改进检索效果。这一方法自 Vivisimo 使用以来,在国外受到极大重视,目前在网上出现的多个系统,如 iboogie、Mooter、Grokker、Webclust 等,都采用了这一形式。这种现象值得我们关注和思考。

动态自动聚类系统的出现,首先是网络关键词检索系统实用需要的结果。关键词搜索引擎作为网上通用性检索工具,至少存在 3 个问题。其一,系统提供的相关性排序效果不理想,不能保证将最符合查询需求的资源排列在前,特别是在返回的资源数量较多,用户所需要的网页排列靠后的情况下,往往很难找到。其二,用户检索时存在着如何确切表达检索需求的问题,普通用户多数情况下只输入表达检索对象的基本词汇,在此情况下,存在着对检索表达的优化问题,通常需要通过对资源的进一步分析,才能明确相关的主题,从而影响检索效率。其三,网络环境中返回结果的总量往往过大,需要有进一步改进应用的手段。以 Google、百度为例,符合检索条件的资源往往动辄数万条,存在着探索新的提供方式,以使检索结果能够得到有效使用的问题。动态自动聚类可以在这些方面加以改进和优化,这也是一些网络搜索引擎研制、开发这类系统的直接原因。

其次,从分类法的角度看,动态自动聚类实际上是对海量环境下新的分类法应用形式的一种探索。传统的文献分类系统,包括自动分类系统或聚类系统基本上是一种全局系统,这类系统根据资源的整体情况建立分类结构,遵循文献保证原则,以适合浏览的方式设置类目。但网络资源数量巨大,无论是类目体系的规模、编制难度以及分类结构的适用性、稳定性等都超出了实际的可能。动态自动分类则正是根据网络环境特点出现的新的分类形态。这种方式不试图建立全局性分类体系,每次都以局部对象为目标建立有限的类目系统;不预先设置固定的分类结构,而是以动态方式实时构建,通过动态操作和局部分类的结合,构成一种适合海量环境的全新的分类形式。这类系统可以与网络分类搜索引擎亦即主题指南形成一种明显的分工:主题指南主要针对特定的使用需要,在精选资源的基础上建立全局性的分类浏览结构;而动态自动聚类系统则解决海量资源的聚类揭示

问题。两种分类方法结合,成为分类法在网络环境中的两种相互补充的应用形式。

此外,这一方法也为自动聚类发展了一种新的使用形式。由于对大容量、整体性资源进行自动聚类的复杂性和有效性方面存在的问题,自从上世纪80年代以来,自动聚类的研究受到了限制。动态自动聚类则使它找到了一种适合这一技术的使用形式,为自动聚类在网络环境下的应用和发展提供了可能性。随着对这一技术形式研究的深入,它必然可以与海量环境下的各种应用,如数据分析、知识挖掘等灵活结合,成为信息资源开发利用的利器,值得我们重视。

2003~2005年期间,我们在建立专用实验库的基础上,对中文动态自动聚类方法进行了研究。本文试图结合我们在实验中的认识,对动态自动聚类系统的特点、适用的方法系统以及中文切分工具等问题进行讨论。

2 动态自动聚类的特点

动态自动聚类系统与传统聚类系统的突出差别在于它的聚类条件不同。早在1985年,P. Willet就试验了以倒排文档检出的有限文献集为对象进行自动聚类,并且将其与静态聚类比较[1]。网络出现后,1993年R. B. Allen, P. Obry和M. Littman提出了以Scatter/Gather的方法,通过在检索过程中与用户交互,结合动态聚类方式,用以改进检索效果[2]。其后,许多学者对网络环境下基于检索返回结果的聚类要求、技术方法等问题进行了一系列探讨。比较典型的,如1998、1999年O. Zamir和O. Etzioni[3-4]以及2004年我国学者曾华军(音译)等[5]对于适合网络使用的新的英文文献聚类方法的探索;2000年Anton Leuski和James Allan对动态聚类和排序处理结合应用的研究等[6]。动态自动聚类的研究和发展实践显示,是否能充分了解动态自动聚类的特点及其要求,是有效进行技术探索的关键。

动态自动聚类系统的聚类操作是在动态有限环境下进行的,与传统静态聚类系统相比,具有以下特点:①聚类是动态进行的,该操作是在检索结果返回的基础上进行的实时操作,而不像传统的全局性聚类系统那样,可以预先进行聚类处理。②每次聚类的文献对象数量有限,通常只处理每次检索返回时排列在前的一定数量的资源,而不像全局性聚类系统那样需要对系统所有的资源聚类,极大增加了充分揭示的可能性。③用来作为聚类依据的文献数据只是文献的局部,而不是全文。例如在元搜索引擎中,据以聚类匹配的对象通常只是返回网页的文摘数据。④参与聚类的资源在整个资源集合中的分布是随机的,往往随着检索查询的不同而变动,因此没有预先设定的类目框架可以套用,应能适用于各种可能出现的情况,具备根据不同主题对象、特点随时进行聚类的能力。

作为一种聚类方法,动态聚类与传统的静态聚类在许多方面,如类目相似性要求、类目显示、类名表达等都存在着相同或相似之处,但上述动态有限环境的特点决定了后者在聚类条件和处理功能要求等方面均存在着较大差异,必须根据其使用需要确定其要求,包括:

(1)要求采用计算开销小,处理速度快的聚类方法。动态聚类是一种实时操作,对降低计算开销、提高计算速度十分敏感,需要选用适合的算法和方式加以处理。这包括使用合理的聚类方式、优化算法、改进语词切分技术和接口技术等。当然也可以结合采用先期处理技术来提高处理速度。

(2)适合采用充分揭示的聚类方式。包括加深内容的揭示程度,对多主题文献或多属关系文献从不同角度重复揭示等,以方便用户从不同角度发现和利用。全局性聚类系统由于涉及的资源数量过多,容易造成类目控制方面的问题,不可能使用充分揭示的方式。针对有限对象聚类,使得这一方式成为可能,特别是在将它作为二次检索手段时尤其有价值。

(3)应选择具有较强聚类能力的方法。搜索引擎中自动聚类的对象通常是检索返回的动态文摘,它们一般由包含检索词的句子组成,与检索主题有较好的相关性,但数据量较少。因此有必要使用聚类能力较强的方法,以便能依据有限数据处理对象,有效识别其特点,并根据情况对内容相同或相关的文献进行聚类操作。

(4)应具有灵活的调整能力。动态系统的每一次聚类都是以有限、局部资源为对象,但作为聚类的总体,这种方法应可以适用于各种情况和所有的资源对象。系统应能够根据不同情况对聚类方式进行调整,以便能够根据条件的变化,以适合的方式将相关内容的文献聚合成类。

此外,在中文自动聚类系统中,还存在使用适合的词汇切分工具或方法,对词汇进行识别的问题。如何根据系统聚类的整体需要,使用有效、适用的方式加以处理,这是中文动态自动聚类过程中必须解决的问题之一。

动态自动聚类系统不能照搬传统自动聚类的技术方法,而应该根据动态自动聚类环境的特点和用户的使用需要,选择或发展适用的方法。

3 动态自动聚类方法系统的发展

尽管目前国内外进行的动态自动聚类试验在采用的技术方法方面存在着多样化的特点,一些传统聚类策略和方法仍为不少动态实验系统所选用。但近年来,根据动态有限环境特点和需要进行的技术探索发展迅速,并已形成了一个新的方法系统。这一方法系统的常用做法包括:

(1)直接将专指性短语作为揭示类目相似性识别的依据。传统文本聚类技术通常将文献作为词集对待,在对文献词汇比较的基础上,通过计算文献词汇的重叠程度,作为相应文献的聚类依据。这一方法可以满足按相似性聚类的要求,但并不一定是以特定主题为中心聚类的。直接以专指语词为中心进行测度的方法,一般不计算文献的整体相似度,只计算文献之间作为聚类依据的短语这一变量方向上的相似度。这种方法可以直接以表达主题概念的词或词组作为切分和聚类的依据,同时引入同义控制等措施,实现在概念层次上聚类。这种方式可以使聚类结果揭示充分,并有较强的适应性。但不适合在文献数量较多的全局性系统中应用。由于动态文摘由包含检索词的句子组成,与检索词联系密切,从试验情况看,依据这一方式的文献聚类结果通常有较高的相关度。

(2)更多使用线性聚类策略。传统聚类方式往往在资源对象整体相似性的基础上按照最优化(亦即贪婪法,greedy)方式聚类,将资源组织成按照相似性程度构建的系统,其中尤以等级聚合聚类法(AHC)使用最多。这类算法通常根据文献的相似性,采用从下往上的方式,从个体文献之间的比较出发,按照文献的相似度,将文献有层次地聚合为等级系统,包括单链法、全链法、类平均法等。这些聚类策略,可以根据文献相似性,按照 greedy 法,建立起一个系统类图,在这一结构中,所有文献之间均按相似性关系组织,并且十分稳定。不足的是:计算开销较大,如单链法的计算复杂性为 $O(n^2)$(O 为聚类文献数,n 为文献中词数),全链法则为 $O(n^3)$;一种资源通常只能归入一个类目,对多主题文献揭示有一定局限性;这种方法一般首先根据相似性确定类,然后再确定类名,因此类名语词在对资源内容的确定和表达上往往有一定的差距;不是以主题内容为中心聚类的,聚类结果并不必然符合主题检索的特点。结合线性方式进行聚类的开销则远比前一种方法少。以语词为中心的聚类方式为例,直接以表达主题概念的词或词组作为切分和聚类的依据,同时引入同义控制等措施,实现在概念层次上聚类,不仅符合以主题为中心聚类的特点,而且可以极大降低计算开销,便于类名显示和对多主题重复反映。由于线性聚类策略计算开销小,因此基于文献词汇整体相似性的聚类系统,也采用这类聚类策略,例如 k-means 法以及一些带有类似特点的如 Buckshot 和 Fractation 等方法,均引入了线性聚类策略[7]。

(3)使用等级显示、多维聚类的形式。等级形式便于有层次地展示资源的内容,方便用户测览,用户可以通过其等级系统,选择相应层次的类目检索。多维聚类包括两方面含义:其一指一个多主题文献可同时在相应门类下重复揭示;其二指在各个层次的类目下,所有的子类都能够得到完整显示,便于用户在相应门类中查阅全部有关的文献资源。前一种情况,如同时涉及"农村经济"和"可持续发展"的某一文献,可同时在两个相应类下得到揭示;后一种情况,如有关"农村经济"的类目,在符合聚类条件时,可以同时作为不同等级的类目在相应位置上显示,例如,在作为一级类聚类显示的同时,部分资源作为"西部大开发"或"可持续发展"等类下的二、三级类目显示。这一类目显示方式,改变了手工文献类表传统的单线揭示方式,通过类目重复设置,使得一类下所有的类目对象和资源都得到完整展示,适合电子类表使用的新的显示形式。如果结合采用以语词为中心的聚类方式,就可以为灵活、有效地聚类和显示提供较好的基础。

(4)采用优化算法。以主题语词中心、线性方式基础上建立的聚类系统为例,这一方式中应予以解决的问题包括同义词分散、不同类之间逻辑关系颠倒、主题类目之间的文献交叉、重合现象,以及以通用词为中心形成的无检索价值类目等,通常采用各种优化的算法进行控制,并在必要时结合相应控制词集加以处理,包括同义控制、等级控制、交叉控制、选择控制等。交叉控制可以根据设置的文献重合度阈值,确定交叉类目是否合并;同义控制、等级控制,以及按照资源情况分散或集中聚类等控制,一般则可以结合一定的控制词表,并建立相应的规则系统予以解决;对类目有效性的处理,往往通过引入必要的词汇识别处理机制解决。

(5)扩大预处理的应用。检索界过去通常将计算的常规检索方式认定为后控制方式。即认为计算机检索系统对复杂主题的检索处理,是在用户提出检索要求后进行的。实际上,在当前的计算机检索系统中,为了减少实时处理时的计算开销,对多数可以预料的操作大都采用了预处理的方式。Google 等关键词搜索引擎中对于网页链接值的计算就是一个典型例子。动态自动聚类系统同样也采用了这一处理方式,特别是在将动态聚类直接应用于特定数据库时,这类预处理的范围可以扩大到所有可能预先估计到的对象,这是这类系统减少实时处理时计算开销的重要措施。

以上是结合动态有限聚类环境发展的新的方法系统的内容。根据这一方法系统的特点,我们建立的实验系统在聚类试验中采用的模式包括:采用以主题为中心的聚类方式,通过短语识别揭示主题对象;使用等级结构、多维聚类,增加揭示深度;广泛采纳优化算法,通过同义控制、等级控制、交叉控制,改进类目有效性;采用优化算法,缩短聚类时间等,取得了较好的聚类效果。图1为采用上述方法系统建立的动态自动聚类系统的显示界面,其中左栏为对于检索返回的前200篇资源的聚类结果。

为了比较以主题词为中心的聚类方法与传统等级聚合聚类方法在动态聚类应用中的效果,我们在进行动态自动聚类试验时,依据传统等级聚合聚类算法中最常用的单链法,建立了一个对照系统。试验显示,在以主题为中心聚类、主题揭示充分程度、多维显示能力、类目均衡度、类名表达、算法复杂性等方面,以主题词为中心的聚类方法均优于单链法。有关内容我们将另撰文介绍。

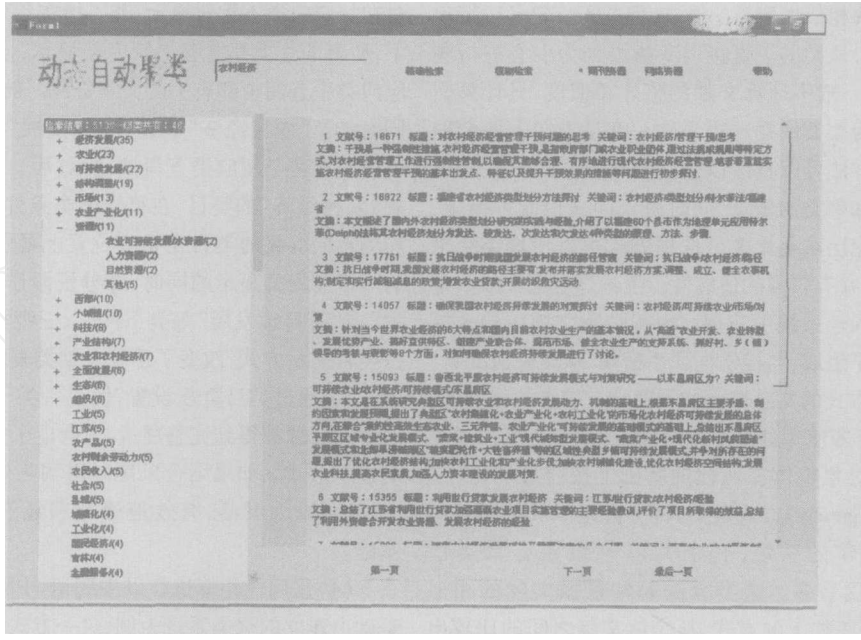

图1 采用主题分类方式的自动聚类系统

4 关于专用中文切分词典

在中文计算机处理的词汇识别方面,国内进行过许多研究,发展了多种中文分词方案[9-10]。其中,实际使用或研究较多的主要有3种:基于基本语词切分词典的切分,基于专用词典的切分,采用基于N-gram法的词汇识别方式。

第一种方式的优点是词汇的识别和切分比较充分,但由于是在普通词汇基础上进行的,在以语词为中心动态聚类时,类目质量容易受到非主题特征语词的影响,同时词汇对主题表达的专指度往往不够充分,在这一基础上引入词汇控制,特别是等级控制有一定困难。基于N-gram法的词汇识别方式有较大发展潜力,但目前离中文语词识别的实际使用需要还有一定差距,仍然无法较好地解决中文短语识别问题,不能满足自动聚类的要求,更谈不上词汇控制的引入。比较而言,建立包括基本概念为基础的词和词组的专用切分词典,有利于识别以主题概念为基础的术

语,使切分更加有效,从而改进聚类质量,优化聚类结果。这一专用控制词集的功能包括:作为切分工具,聚类依据,类名表达依据,对聚类结果加以优化的辅助工具等。通过它,不仅可以用来代替普通切分词典,以此"回避汉语切分的一些技术难点"[11];而且还可以根据引入优化机制的需要,对词汇进行必要处理。要符合这一需要,一般应收入相应专业领域的基本词汇,特别是各种表达专业概念的单词和词组,包括结合抽词需要,收入文本中各种常用的自然语言表达形式,以便可以对文本中的主题表达进行有效的抽取;此外,应结合聚类优化的需要,进行必要的词汇控制。

一般而言,不同的专用控制词集,应允许根据聚类优化的需要,确定各自的收集和控制机制,形成不同的收录和处理特点。以我们在动态聚类系统中研制的专用切分词典情况看,与自动分类系统研究中使用的有指导的词汇系统(或称基于文献分类体系的词汇控制集)相比,其不同主要有:

(1)收入的词汇不从属于预先建立的分类体系。所有的词汇一旦收入,即作为词表的构成单元独立存放,不必预先组织成固定的标题,也不从属于预先建立的分类体系;控制词表中也不预先建立先组方式的等级类表,类目结构根据文献中词汇情况,按照相应规则系统动态生成。

(2)根据动态自动聚类的需要,建立特定的词汇控制机制。词汇控制应根据自动聚类的需要设置。为了使得动态建立的类目系统能够贯彻一定的逻辑原则,并根据动态聚类中各种可能出现的情况,灵活进行聚类处理,可以根据词汇的含义,进行必要的等级处理甚至相关关系处理,建立起词汇关联系统。这类等级系统,不仅涉及的词汇更广,而且处理也更为灵活、粗放,其关系类型,通常需要结合使用的工具和结合计算机处理的可能性确定,不同系统往往可能因其处理方法的不同而采用不同的控制表形式。

(3)根据聚类性能对某些词汇进行必要的标注。例如,对于类似主题标引中通用概念的词汇,无实际聚类价值的高频词等,可根据其聚类能力较弱、容易出现搭配错位等情况,对其作相应标注,以便在聚类阶段结合处理规则合理应用。

表1即为控制词集结构显示的样例。表中Keywords即关键词,该栏收入作为抽词依据的所有自然语言词汇;Guifanci即规范词,收入一关键词对应的规范词;Dengji即为等级词,收入关键词对应的上位词。可以看到,这一专用词集的构成远比基于文献分类法建立的知识库单纯。这一专用词集中建立了同义关系和等级关系,它不仅是动态聚类的依据,也是进行优化控制的辅助工具。只要结合相应的规则系统,即可以在概念的基础上聚类,并在等级关系类目建立的过程中进行概念关系的推理和控制[12]。

表1 专用控制词集样例

Keywords	Guifanci	Dengji
作业管理	作业管理	
ABM	作业管理	
资金流	资金流	
资金流程	资金流	
住宅产业化	住宅产业化	产业化
住宅产业化	住宅产业化	住宅
住宅产业	住宅产业	产业
住宅产业	住宅产业	住宅
住宅产品	住宅产品	产品
住宅产品	住宅产品	住宅
住宅标准	住宅标准	标准
住宅标准	住宅标准	住宅
住院时间	住院时间	时间
住院时间	住院时间	住院
住院	住院	
住房置业担保	住房置业担保	住房置业
住房制度改革	住房制度改革	住房制度
住房制度改革	住房制度改革	制度改革
住房制度	住房制度	住房
住房制度	住房制度	制度
住房证券化	住房证券化	住房
住房证券化	住房证券化	住房
住房需求	住房需求	住房
住房需求	住房需求	需求
住房	住房	
心与迹	心迹	
心迹	心迹	

5 结束语

将知识结构引入自然语言系统有两种方式:一种是通过先控的方式,利用人工标引等方式将知识结构引入系统;另一种是按照自然语言系统的特点,在文本检索的基础上,利用计算机的能力,结合各种词汇控制的方式加以实现。由于网络环境下文本数量的迅速增加,单纯的人工控制系统已远远无法满足使用的需要。根据文本系统的特点,发展各种适用的形式,将知识结构引入系统,是改进自然语言系统使用效果的努力方向之一。动态聚类系统就是其中的一种类型。随着文本检索技术的不断发展,这类努力必将受到更多关注,并逐步发展出一系列有效结合的新形式。

参考文献

1 Willett, P.. Query specific automatic document classification. International Forum on Information and Documentation, 1985. 10(2), 28 ~ 32

2 R. B. Allen, P. Obry and M. Littman. An interface for navigating clustered document sets returned by queries. in: Proceedings of the ACM Conference on Organizational Computing systems, 1993, pp 166 ~ 171

3 O. Zamir, O. Etzioni. Web document clustering: a feasibility demonstration. in: Proceedings of the 19th International ACM SIGIR Conference on Research and Development in Information Retrieval(SIGIR'98), 1998, pp 46 ~ 54

4 O. Zamie, O. Etzioni Grouper. A Dynamic Clustering Interface to Web Search Results. In Proceedings of the Eighth International World Wide Web Conference(WWW8), Toronto, Canada, May 1999

5 Hua-Jun Zeng, Qi-Cai He, Zheng Chen, et al. Learning to Cluster Web Search Results. ACM, 2004

6 A. Leuski, J. Allan. Improving Interactive Retrieval by Combining Ranked List and Clustering. Proceedings of RIAO, College de France, 2000, pp665 ~ 681

7 M. A. Hearst, J. O. Pedersen. Reexamining the cluster hypothesis: Scatter/Gather on retrieval results, in: Proceedings of the 19th International ACM SIGIR Conference on Research and Development in Information Retrieval(SIGIR'96), 1996, PP 76 ~ 84

8,12 马张华,陈文广等.基于控制词集的动态自动聚类研究(技术报告),2005,10

9 张琪玉.情报语言学基础.武汉:武汉大学出版社,1997

10 刘开瑛.中文文本自动分词与标注.北京:商务印书馆,2000

11 侯汉清.文献数据库词表及自动标引技术的研究—新华社电讯稿数据库自动标引系统的研制(研究报告),2000

马张华 北京大学信息管理系教授。通信地址:北京大学。邮编100871。

(来稿时间:2006-03-22)

(上接第48页)健全控制风险责任制和对风险业务控制与使用[11]。风险管理是一门新兴的边缘学科,将风险管理引入图书馆BPR项目实施中是一种有益的尝试,在对图书馆BPR项目的风险进行识别、分析和应对的系统过程中,合理运用BPR项目风险管理的方法,有利于提高图书馆BPR项目的风险管理能力,将图书馆BPR项目引向成功的方向。

参考文献

1,8 [美]保罗·S·罗耶著;北京广联达慧中软件技术有限公司译.项目风险管理:一种主动的策略.北京:机械工业出版社,2005

2 赵林度.供应链与物流管理理论实务.南京:东南大学出版社,2003

3,4 邹凯,何岸,陈能华.面向供应链管理的图书馆业务流程重组.中国图书馆学报,2005(4)

5 尚加宁,图书馆组织工作中的风险管理.情报杂志,2002(2)

6 孙德彬.数字图书馆建设中的风险管理.现代情报,2002(12)

7 Royer Paul S. Risk Management: The Undiscovered Dimension of Project Management. PM Journal, 2000, 31(1)

9 Il Im, Omar A El Sawy, Alexander Hars. Competence and Impact of Tools for BPR. Information & Manegement, 1999, 36

10 张喜年,詹德优.基于供应链管理的图书馆参考服务.情报理论与实践,2006(1)

11 罗志尧,周群芳.现代图书馆的组织风险及其规避.浙江高校图书情报工作,2004(6)

蒋知义 湘潭大学图书馆馆员,在职硕士研究生。通信地址:湘潭大学图书馆。邮编411105。

邹 凯 湘潭大学管理学院博士,教授,副院长,硕士生导师。通信地址:湘潭大学管理学院。邮编411105。

(来稿时间:2006-04-12)