

基于 MPEG-4 的视频检索技术研究^{*}

许平 谈宁韡 郭胜 陆桑璐 谢立

(南京大学计算机软件新技术国家重点实验室 南京 210093)

(南京大学计算机科学与工程系 南京 210093)

A Survey of Video Retrieval Based on MPEG-4 Standard

XU Ping TAN Ning-Wei GUO Sheng LU Sang-Lu XIE Li

(State Key Laboratory for Novel Software Technology, Department of Computer

Science and Engineering, Nanjing University, Nanjing 210093)

Abstract There are so many multimedia resources now. But it is difficult to look for any video thing among them. MPEG-4 is a major technique used for video coding. In the paper, some methods based on MPEG-4 coding standard for video retrieval are introduced. The focus is on converting video data to textual data, then video retrieval is converted to textual retrieval. XML technique is involved naturally.

Keywords MPEG-4, XML, DIVX, A/VO, OCI, XMT

1 引言

今天,人类的科学文明发展到了很高的水平。人们足不出户,就可在家里欣赏到声形并茂的视听内容。我们可以根据需求去选择视听材料。拿电影来举例,如果我们知道所要查找的影片名字时,这很好办;但如果我们仅凭童年的记忆依稀记得某个场景的一个细节,这仍然是一个未解决的问题。当人们可以哼着旋律找到心目中的歌曲时,视频检索就成了一个亟待解决的问题。视频检索涉及的多媒体文件格式,目前最有发展前景的莫过于基于 MPEG-4 标准的了,所以本文就基于 MPEG-4 的视频检索技术进行研究。

众所周知,活动图像专家组 MPEG(Moving Picture Experts Group)从 1992 年起逐步定义了 MPEG-1, MPEG-2, MPEG-4 视频音频编码标准。MPEG-4 的高压缩率和高图像还原质量使得其应用得到了迅速的发展。

MPEG-4 希望建立一种统一的多媒体数据格式,能被多媒体传输、存储、检索等应用领域普遍采纳。MPEG-4 引入的关键概念 AV(Audio/Visual)对象,是 MPEG-4 为支持基于内容编码而提出的。在 MPEG-4 中所见的视频/音频已不再是过去 MPEG-1, MPEG-2 中图像帧或者帧频的概念,而是一个个视听场景(AV 场景),这些不同的 AV 场景由不同的 AV 对象组成。AV 对象是听觉、视觉,或者视听内容的表示单元。其中最基本的单元是原始 AV 对象,它可以是自然的或合成的声音、图像。原始 AV 对象又可以进一步组成复合 AV 对象。整个 MPEG-4 就是围绕如何高效编码 AV 对象,如何有效组织、传输 AV 对象而编制的。因此,AV 对象的编码是 MPEG-4 的核心编码技术。MPEG-4 中基于对象的思想,给基于内容的检索开拓了新的思路。

通常意义上说的 MPEG-4 文件格式实际上指的是 Divx。Divx4 文件格式使用 MPEG-4 的视频压缩和 mp3 的音频压

缩标准。新近推出的 Divx5 使用一些新的技术,支持 MPEG-4 视频标准的 SVP(Simple Visual Profile)和 ASP(Advanced Simple Profile)。鉴于 MPEG-4 的应用发展和视频检索技术的需求,基于 MPEG-4 标准的视频检索技术是有应用价值的。我们的目标在于将视频信息以文本的方式表现。实际上本文不涉及具体的文本检索技术及图像识别技术,我们的重点在于根据 MPEG-4 的特点将其转化为文本信息。

2 基于 MPEG-4 的视频检索方法

这里,依照视频检索依赖技术可分为 3 类:基于 MPEG-4 本身;基于 XML 技术;两者结合(XMT)。

2.1 基于 MPEG-4 本身:使用 OCI(Object Context Information)

(1)背景 MPEG-4 把多媒体内容看作是一组 AV(音频视频)对象。一般说来,MPEG-4 中所有的信息都以流的形式传播。MPEG-4 用基本流作为传输的基本单位,用对象描述符来刻画,可以包括场景描述、AV 对象数据,甚至对象描述符自己,完全或部分包含单个音频或视频对象的编码表示、场景描述或控制信息的数据。在流上附加有用的元信息是非常方便的。流描述本身也是以基本流的形式传输。这种分离促进了 MPEG-4 内容的编辑和总体操作。

(2)可用于检索的 OCI 部分 对象内容信息 OCI 是 MPEG-4 中定义的附加的辅助信息。可以把附加的辅助信息关联到一个对象描述符上,以文本的形式描述这些流传输的内容。分层的结构化的描述符集组成了对象描述结构的主要部分。对象描述符(OD)是最高层次的描述符,我们首先介绍描述符自己的语义,然后进一步讨论它们以流方式传输的隐含内容。

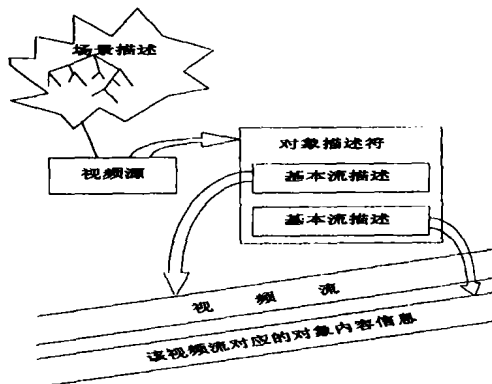
对象描述符 一个对象描述符可以聚集一组连接 MPEG-4 场景的描述内容。我们所关注的 OCI 属于附加的描

^{*} 本文得到国家八六三技术研究发展计划(编号 2001AA113050)资助。许平 硕士研究生,主要研究方向为分布式计算、网络多媒体。谈宁韡 硕士研究生。郭胜 硕士研究生。陆桑璐 副教授,主要研究方向为分布式计算。谢立 教授,博士生导师,主要研究领域为分布与并行计算机系统。

述符。对象描述符除了包括基本流描述符之外,还可以有一些与基本流相关的辅助流。主要包括三种辅助信息:首先,可以通过 OCI(Object Content Information)传输 AV 对象内容的语义信息,然后可以附加知识产权管理和保护信息,最后可以用时钟参考流来传输时间基信息。其他的信息还可以包括 QoS 描述。由标准化组织对描述符的自定义进行规范,我们可以在 AV 对象加入任意的辅助描述符来传送用户所需的信息。这里的 OCI 是我们现在所关心的焦点,可帮助我们进行基于内容的视频检索。

(3)对象内容信息(Object content information,OCI) 以上我们已知 OCI 属于附加的描述符,主要包括一组 OCI 描述符,表达由基本流构成的 AV 对象的一组特征,同时和某一个特定的对象描述符相关。在这里可以使用带关键字的描述符,描述内容的文本信息、语言和内容码率信息,以及内容项的日期和名字及对象内容信息的作者。

可以在对象描述符中直接包括这些描述符来指出静态的 OCI 属性。如果 OCI 在与对象描述符相关的媒体流的生命周期之外改变的话,那么可以将 OCI 流关联到对象描述符上。在 OCI 上标明开始和持续时间,成为 OCI 事件,同样用流的形式传输,称为 OCI 流。这意味着一个 MPEG-4 的表示可以进一步从语义上分解为表示的实际内容。OCI 流和 OCI 描述符一样,可以被关联到音频、视频和场景描述流上。所关联的对象将影响到它们的表示范围。一旦被关联到对象描述符上,比如说音频、视频流,OCI 显然只能描述对应的流。当关联到场景描述流上,OCI 描述的任何事物,反过来又被场景描述所描述。在这种情况下,OCI 流可以非常有用,通过一个场景描述传输语义上有意义的 AV 事件^[3]。



(4)在 MPEG-4 标准中提出的 OCI,对相应的 AV 给予内容描述,这为我们对视频进行基于内容的描述提供了很好的切入点。利用 OCI 对场景描述中的 AV 对象进行描述,利用 OCI 流关联到场景描述流中,这就建立了 AV 对象的文本描述。这里还有一个关键技术,就是 OCI 流的编解码。我们可以对现有的编/解码器进行改造。在解码时分离出 OCI,并且根据其对应流找到相应的视频内容。这样,利用 OCI 流,基于内容的视频检索转化为 OCI 流中携带的文本的检索。这种策略为我们提供了一种可行的方法,极大地降低了检索的数据量,简化了检索,提高了检索的效率。

2.2 基于 XML 的表示方法

(1)XML(可扩展标记语言)简述 XML 可扩展标记语言是 Web 数据使用的通用语言,具有结构化、规范性、可扩展性及简洁的特点。XML 能让开发人员将来自各种应用程序的结构化数据传送给桌面,以在本地计算和表示。并且,XML 允许为特定应用程序创建独特的数据格式。由于 XML 的这

些特性,它很适合用来精确地描述 MPEG-4 位流的语法和语义,并且很容易被转换成程序。

(2)将 MPEG-4 用 XML 语言表示的可行性分析 基于 XML 的描述可以支持被描述内容和结构的随机访问。XML 标签的属性可以传输值。因此,传统的二进制格式中的时空依赖性在基于 XML 的表示中被打破,可以自定义 MPEG-4 的 XML 格式的表示。

XML 作为一种标记语言,标记集的定义决定了最终的 MPEG-4 的表示,并且可以将所有的 XML 的标签组织成一个树结构。MPEG-4 的可视内容有其内容组织的基本树型结构。一般说来,一个可视的对象由几个视频对象层组成,每个视频层包括一些宏块,且每个宏块又分为几小块(比如说在 4:2:0 格式中 6 块)。这个树结构自然被映射成 XML 表示的 XML 标签。每个 MPEG-4 视频格式对应于 XML 树的一个节点。

我们可以把 MPEG-4 视频格式的所有部分映射成 XML 标签。这些 XML 标签指出了目标视频的结构化信息,其属性值则记录了关于视频内容的重要的描述信息。在这里要将 MPEG-4 的视频内容都用 XML 的结构化形式记录下来,面临的关键问题是标签集的定义。所幸的是,MPEG-4 中基于对象的特点以及基本流的管理方式,都给我们定义标签集带来了极大的便利。后面我们将会看到一个相关的实验。

2.3 可扩展的 MPEG-4 文本格式(XMT)

上面我们分别介绍了着眼于 MPEG-4 本身的方法和基于 XML 的方法,这里我们还可以利用可扩展的 MPEG-4 文本格式(XMT)。XMT 是使用文本语法表示 MPEG-4 场景描述的结构,允许内容的提供者交换其内容,并且增强了 W3D 组织开发的 X3D 和同步多媒体综合语言(SMIL)的互操作能力。

MPEG-4 文本格式(XMT)的结构包括两层文本格式: XMT-A 和 XMT-Ω。XMT-A 是 MPEG-4 二进制表示的 XML 镜像,提供了一对一的确定映射,并可以与 X3D 互操作。XMT-Ω 是基于 W3C SMIL 的 MPEG-4 的特征的高层抽象,是一种基于 XML 的语言,允许用户描述时间行为、多媒体表示的层次以及授权者的动态交互式多媒体表示。XMT-A 提供了在 VRML 和 MPEG-4 之间的互操作,XMT-Ω 提供了 SMIL 和 MPEG-4 之间的互操作^[2]。

XMT 主要适用范围是场景描述。我们知道,在 MPEG-4 中,场景是个重要的内容单元。当对视频进行基于内容的检索时,如果仅仅停留在场景的层次上进行,那么我们可以利用 XMT 提供的文本信息进行。但是使用 XMT 不能像 XML 那样自定义需要的标签集,只能使用已有的标准。并且,XMT 定义的初衷在于为授权者提供维持意图的文本格式,所以我们推荐使用前面的方法。

3 实验分析

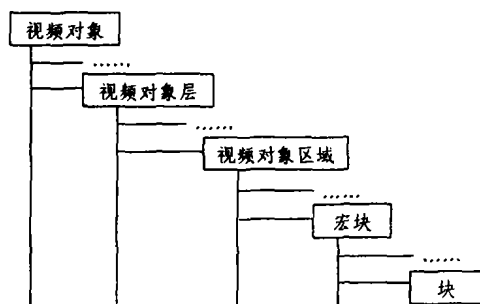
3.1 基于 XML 的检索方法

如上所提,我们首先要解决的问题就是标签集的定义。我们现在的的问题是,究竟要多少 XML 标签才能完全描述一段 MPEG-4 的视频剪辑。

我们先考察一下 Xiaoming Sun 等的实验,他们的工作重点是基于 XML 的 MPEG-4 的视频表示、传输和恢复。

(1)标签集 在他们的实验中使用这样一个完全的集合来表示 MPEG-4 内容信息。在他们定义的完全集中,大部分

XML 标签中有 3 个属性都是需要的 (“pos”, “bits” 和 “value”)。pos 指出了视频位流的位置, bits 表明在视频位流中元素所占的位数; value 是这个元素的实际值, 可以是原始的二进制值, 也可以是经过转换之后的值。MPEG-4 视频结构如下:



XML 文档反映这样的层次结构比较简单。从中我们可以看出, MPEG-4 的 VO 对象为基于内容的检索提供了很好的机制。

(2) 索引 在 Xiaoming Sun 等的实验中提供了索引机制。XML 描述提供了树型结构。XML 文档被解析后, 这个 XML 树就在内存中建立起来了。XML 树中的任何一个元素可以用命名匹配或索引来标志。也就是说, 为了访问 XML 树中的一个指定的元素, 名字和 XML 节点的索引号都是必不可少的, 它们指明了从根到目标节点的路径。

(3) 剪裁 一个完整的 MPEG-4 视频的 XML 完整描述能够提供视频内容和结构的精确描述。但是, 这也会产生大量 XML 文档, 造成巨大开销, 所以对整个视频序列产生 XML 表示是不切实际的。在 Xiaoming Sun 等的实验中, 提供了剪裁机制来减少开销。这个系统提供了 3 种过滤机制: ①指定产生一个特别的 XML 标签; ②指定某一个特别宏块的 XML 标签是否产生; ③指定某一个视频包的 XML 标签是否产生。被屏蔽的内容节点在 XML 树中将不会存在, 并且这个节点的任何引用都将返回空值。

(4) XML 文档的产生和解析 Xiaomig Sun 等人使用 XML 解析器和 MPEG-4 解码器同时工作, 以产生和解析 XML 文档。对于基于 XML 的 MPEG-4 的解码过程, XML 文档首先在初始阶段被解析, 而在内存中建立一个 XML 结构树。可以随机访问这个结构数, 并且 MPEG-4 的解码过程可以与这个数据结构交互, 避免任何可能出现的错误。传统的 MPEG-4 解码过程的顺序性可以变得不必要。

一般说来, MPEG-4 的 XML 文档在服务器端产生, 有两种情况: 第一种情况, XML 解析器和 MPEG-4 编码器同时工作, 在编码时产生 XML 文档; 第二种情况, 对编码的视频流进行独立的编码扫描, 这个扫描过程可以作为任何进一步视频流需求的预处理。在客户端, 可以通过 XML 解析器析取出 value 的值, 传送给 MPEG-4 的解码器。大部分 XML 标签有属性 “value”, 这个属性中记录了视频中内容的真实数值^[1]。

设计 XML 文档, 得到这样关于 MPEG-4 内容的文本的结构化表示, 将基于 MPEG-4 视频的内容检索转化为对 XML 文档的检索。在 Xiaoming Sun 的方案中, 是将该 XML 表示用于 MPEG-4 视频内容的保护, 这与我们的目的不同。所以定义的 XML 标记集显然是有差别的, 而 XML 标记集的定义是基于 XML 的视频检索的关键。因此, 实现这一方案的焦点, 应该在于考虑定义标记集, 目标是减少 XML 文档的开销, 并提供快速检索功能。

3.2 多媒体内容描述模型 InfoPyramid

由 BM T. J. 研究中心的 Chung-Sheng Li 等提出。这个方案虽然不涉及 MPEG-4 格式的特性, 但是其思想却是很通用, 并且在实际中得到了应用。

(1) 使用多媒体不同格式的分离 多媒体内容通常不是一个单独的媒体格式。比如说, 一个视频剪辑可以包括视频、两种或更多语言的音频以及标题文本部分。对于一个病人, 可以用核磁共振、CT 扫描、超声波等来获得同样内容的三维描述。

对于一定的查询和检索任务, 可能不总是能获得最合适的内容形式。所需的形式可以通过转换形式得到。比如说, 可以将一个视频剪辑转换成关键帧, 而文本也可以合成为语音。每部分的内容都可以有多种解决方法。在构建图像和视频金字塔中, 存在许多简化技术。比如说, Flashpix 提供了存储和检索静止图像的机制。在不同的方案中, 由原始数据或转换后的数据获得的特征和语义可以构成一个特征和语义金字塔。用多个抽象层次来组织, 多个抽象层次以一个结构化的方法促进前向的搜索和浏览。一个层次可以是特征、语义、对象描述、注解或元数据。

还有其他方法和规则的概念。利用方法, 从数据特征中产生内容描述, 或者分析、操作, 转换不同形式, 或者以不同的方式处理数据。Infopyramid 提供了保证灵活应用的规则。方法和规则提供了不同的形式、解决方案和抽象之间的连接。InfoPyramid 提供的是一个数据、数据的抽象以及内容描述数据的连续体, 而不是数据和内容描述之间的强制分离。

在 InfoPyramid 模型中, 为了使得内容描述可以在搜索引擎之外使用, 定义了标准的一组内容描述符原语和处理、表示的过程, 并且使用 XML 来表达内容描述^[4]。

(2) 电视新闻的获取和索引 这个模型被应用到一个视频数据库系统中, 自动进行电视新闻的获取和索引, 并且可在因特网上搜索。系统获取电视新闻播放和文字说明流, 然后把电视新闻分段成独立的新闻故事, 每个故事的文本信息传送给一个文本索引。用户在因特网上可以通过文本查询方式来查询这个新闻故事的数据库。

在数据库中, 新闻以不同的方法表示, 比如 AVI 视频用关键帧, 文本用标题, 音频也独立表示。这个 InfoPyramid 可以得到其 XML 表示(见附录)。

使用何种形式的新闻内容, 依赖于查询, 由搜索机制决定正确的查询。比如说, 基于文本的查询就访问文本描述, 基于视频的查询可以使用关键帧。这里的应用范围不仅仅是搜索, 当需要返回视频摘要时, InfoPyramid 就返回一个关键帧序列。当需要一个特定的新闻时, 就返回这个新闻的视频信息。

(3) 信息金字塔 这是一个多媒体内容描述方案, 这种描述符合 MPEG-7 的致力目标, 可以被 MPEG-7 应用。

IBM 研究中心提出的基于 XML 的多媒体内容描述方案 InfoPyramid, 这个方案已经在一个实际的多媒体数据库中实现。这里提出的 XML 结构值得借鉴, 但是关于关键帧的选取也是一个很重要的问题, 需要采取好的算法才能达到理想的效果。

总结 基于内容的视频检索是一个非常有意义的工作。今天是信息时代, 在大量的多媒体资源面前, 人们要查找自己需要的内容, 无异于大海捞针。所以, 用视频内容和音频内容进行自动检索, 显然是一个理想的层次。利用 MPEG-4 自己定义的 OCI, 生成 OCI 流, 可以与视频信息同时传递, 并且可

与基本流中的对象联系起来。OCI 中提供的内容信息,可以将基于内容的视频检索转化成基于文本的检索。缺点在于需要改造现有的编解码器,并且表示方式不够灵活多样。另外,MPEG-4 协议中定义的 XMT 也是可利用的对象,但是 XMT 的定义初衷是保留、维持原作者的创作意图,所以同样在使用上不能完全适应检索的需要。MPEG-4 格式的对象结构,使我们用结构化的 XML 可扩展标记语言来表示更为容易。利用 MPEG-4 标准和 XML 技术的结合,将基于内容的视频检索转化为对 XML 文档的检索,是一条捷径。已有的工作与此有相关性,但是还没有完全实现检索的例子。这里需要解决的问题是,怎样定义 XML 标签集,才能完全表示 MPEG-4 视频内容,且 XML 的表示文档要有合理的开销。合理的索引结构也是必要的。我们进一步的工作就是要解决这些问题。

参考文献

- 1 Sun Xiaoming, Shi Zhi, Kuo C C J. XML-based MPEG-4 Video Representation, and Error Resilience. Circuits and Systems, 2002 IEEE International Symposium on, 2002, 2: 676~679
- 2 Kim M, Wood S, Cheok L-T. Extensible MPEG-4 Textual Format. In: Proc. ACM Multimedia 2000
- 3 Herpel C, Eleftheriadis A. MPEG-4 Systems: Elementary Stream Management. Image Communication Journal TUTORIAL ISSUE ON THE MPEG-4 STANDARD, 2000
- 4 Li C-S, Mohan R, Smith J R. Multimedia Content Description In The InfoPyramid. In: IEEE Proc. Int. Conf. Acoust., Speech, Signal Processing (ICASSP), Seattle, WA, May 1998
- 5 Law C, Furht B. The Mpeg-4 Standard for Internet-based Multimedia Application. Handbook of Internet computing, 1999. 41~65
- 6 Hunter J, Newmarch J. An Indexing, Browsing, Search and Retrieval System for Audio visual Libraries. <http://archive.dstc.edu.au/RDU/staff/jane-hunter/ECDL3/paper.html>
- 7 Salembier P. An overview of MPEG-7 Multimedia Description Schemes and of future visual information analysis challenges for

content-based indexing. Int. Workshop on Content-based Multimedia Indexing, CBMI'01, Key note speech, Brescia, Italy, Sep. 2001. 351~361

- 8 王相海, 张福炎. 多媒体视频编码研究. 计算机科学, 2001, 28(11): 17~21
- 9 Chang S-F, et al. A fully automated content based video search engine supporting spatio-temporal queries. ACM Multimedia, 1997
- 10 Koenen R. Overview of the MPEG-4 Standard, ISO/IEC JTC1/SC29/WG11 N4030, March 2001
- 11 Zhong D, Chang S-F. Spatio-temporal video search using the object based video representation. In: Proc. of IEEE Int'l Conference on Image Processing, Santa Barbara, CA, Oct. 1997. 21~24
- 12 Hampapur A. Semantic Video Indexing: Approach and Issues. ACM Sigmod Record, 1999, 28(1): 32~39
- 13 Jose M. Martinez (UPM-GTI, ES). Overview of the MPEG-7 Standard (version 6.0), ISO/IEC JTC1/SC29/WG11 N4509, Pataya, Dec. 2001
- 14 Fallside D C. XML Scheme part0; Primer. March 2001. <http://www.w3.org/TR/xmlschema-0/>

附录

InfoPyramid 的 XML 表示:

```
<NEWS-STORY>
  <Program>ABC Evening News</Program>
  <date>11/2/97</date>
  .....
  <Video>
    <Component>
      <Content-type>video/avi</Content-type>
      <URL>http://foo.com/news1.avi</URL>
      <bit-rate>1Mbs</bit-rate>.....
    </Component>
    <Component>
      <Content-type>video/Bamba</Content-type>
      .....
    </Component>
    .....
  </Video>
  <transcript>.....</transcript>
</NEWS-STORY>
```

(上接第59页)

术语集的搜索引擎内容描述方法,并将之结合到 gGLOSS 方法和文[5]的计算方法中,使之对于支持布尔查询和支持向量查询的搜索引擎选择都有效。从实验结果可知,它有效提高了引擎选择的精确度。

参考文献

- 1 Gravano L, Chang K, Garcia-Molina H, Lagoze C, Paepcke A. STARTS Stanford Protocol Proposal for Internet Retrieval and Search. In: Proc. of the 1997 ACM SIGMOD International Conference On Management of Data, 1997. 207~218
- 2 Gravano L, Garcia-Molina H, Tomasic A. Gloss: Text-Source Discovery over the Internet. ACM Transactions on Database Systems, 1999, 24(2)
- 3 孟卫一, 吴宗寰. 集成搜索引擎的文本数据库选择. 计算机研究与发展, 2001, 38(4): 396~404
- 4 Meng W, Liu K L, Yu C, et al. Determine text databases to search in the internet. In: Proc. of Very Large Database Conference, New York City, 1998. 14~25

- 5 Meng W, Liu K L, Yu C, et al. Estimation the usefulness of search engines. In: Proc. of IEEE Int'l Conf on Data Engineering, Sydney, Australia, 1999. 146~153
- 6 Gravano L, Garcia-Molina H. Generalizing GLOSS to vector-space databases and broker hierarchies. In: Proc. of Very Large Database Conference, Zurich, Switzerland, 1995. 78~89
- 7 Callan J, Lu Z, Croft W. Searching distributed collections with inference networks. In: Proc of ACM SIGIR Conf. Seattle, 1995. 21~28
- 8 Gravano L, Garcia-Molina H. Generalizing GLOSS to vector-space databases and broker hierarchies. In: Proc of Very Large Database Conf. Zurich, Switzerland, 1995. 78~89
- 9 Dreilinger D, Howe A. Experiences with selecting search engines using metasearch. ACM Trans on Information System, 1997, 15(3): 195~222
- 10 Fan Y, Gauch S. Adaptive agents for information gathering from multiple, distributed information sources. In: Proc of 1999 AAAI Symposium on Intelligent Agents in Cyberspace, Stanford University, 1999. 40~46