网络信息检索

第十一讲 网络信息检索应用（2）

董守斌
sbdong@scut.edu.cn

华南理工大学计算机学院
广东省计算机网络重点实验室
Communication & Computer Network Laboratory (CCNL)

主要内容

- 多媒体检索
- 跨语言检索
- 问题回答

多媒体（**Multimedia**）的定义

- **Multimedia** is **media** that uses multiple forms of **information content** and **information processing** (e.g. **text**, **audio**, **graphics**, **animation**, **video**, **interactivity**) to inform or entertain the (user) audience.

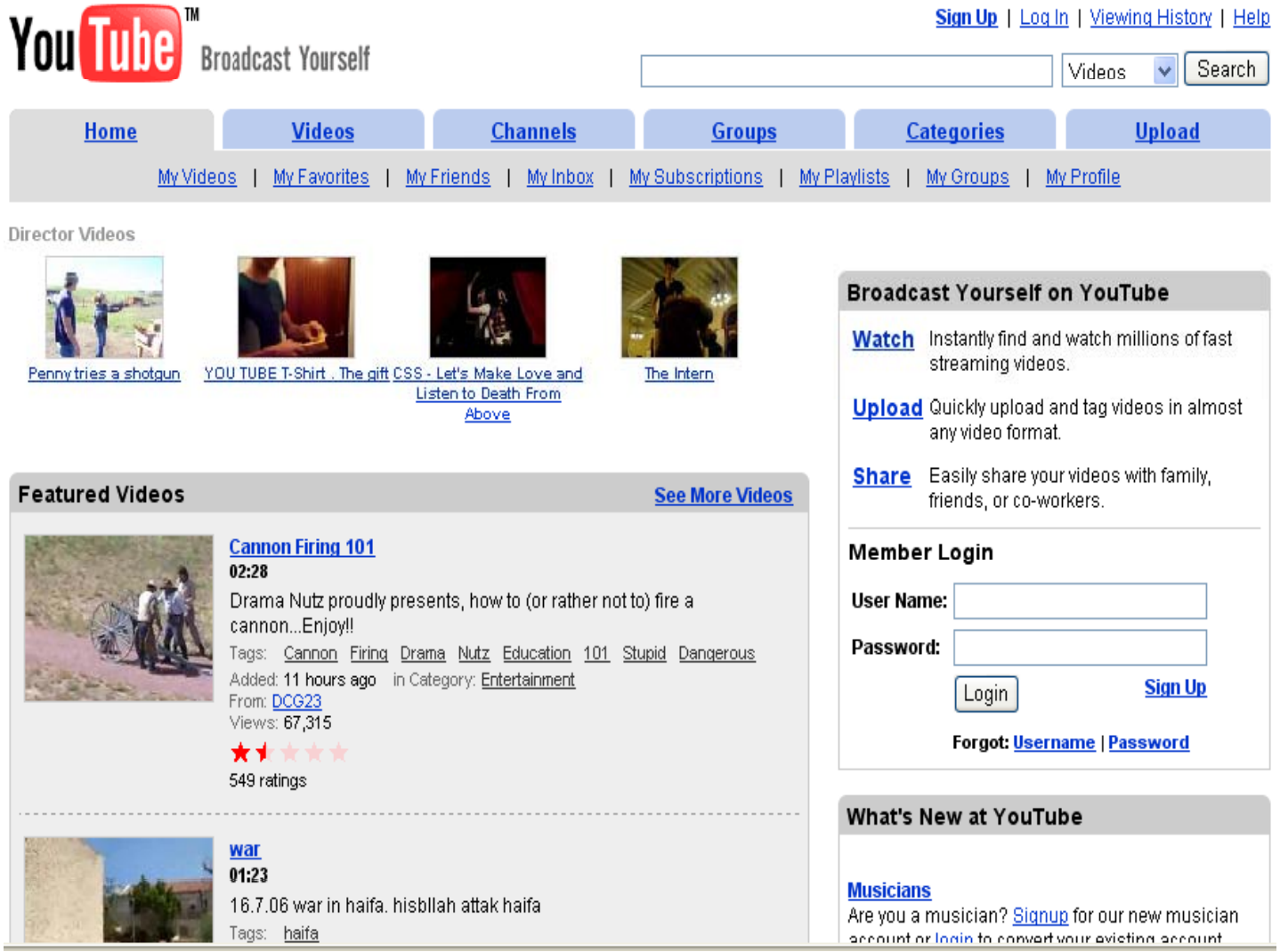
----*from wikipedia*

网络多媒体对象

- 网上存在大量多媒体文档
  - 声音：**mp3/wav/rm…**
  - 图片：**jpg/bmp/gif/tiff/…**
  - 动画：**swf/gif…**
  - 图形：（矢量图形文件）**dwg/dxf/3ds…**
  - 视频：**mov/wmv/mpeg/mpg/rm…**

- **A picture is worth a thousand words!**

YouTube


YouTube
Broadcast Yourself
Sign Up | Log In | Viewing History | Help
Videos
Search
Home
Videos
Channels
Groups
Categories
Upload
My Videos | My Favorites | My Friends | My Inbox | My Subscriptions | My Playlists | My Groups | My Profile
Director Videos
Penny tries a shotgun
YOU TUBE T-Shirt . The gift
CSS - Let's Make Love and Listen to Death From Above
The Intern
Featured Videos
See More Videos
Cannon Firing 101
02:28
Drama Nutz proudly presents, how to (or rather not to) fire a cannon...Enjoy!!
Tags: Cannon Firing Drama Nutz Education 101 Stupid Dangerous
Added: 11 hours ago in Category: Entertainment
From: DCG23
Views: 67,315
549 ratings
war
01:23
16.7.06 war in haifa. hisbllah attak haifa
Tags: haifa
Broadcast Yourself on YouTube
Watch Instantly find and watch millions of fast streaming videos.
Upload Quickly upload and tag videos in almost any video format.
Share Easily share your videos with family, friends, or co-workers.
Member Login
User Name:
Password:
Login
Sign Up
Forgot: Username | Password
What's New at YouTube
Musicians
Are you a musician? Signup for our new musician

YouTube

- 2006年2月的视频片断1500万，目前已达到4000万，平均每个视频片断长2-3分钟，每天有1亿分钟的视频（约合190年的视频）
- 每天的网络流量已经由25TB上升到200TB，即1个月6PB，1年72PB
- 每月的网络流量费$1百万，其内容传送网络提供商Limelight Networks每月的总收入$4百万（1/4的收入来自YouTube）
- 2006年10月9日，Google以价值16.5亿美元的股票收购YouTube

多媒体检索的困难

- 对同一主题，多媒体表达千差万别
- 多媒体对象具有十分复杂的特征，进行特征表示比较困难，对多媒体对象的理解就更困难
- 用户的检索需求也非常复杂，有时是基于低级特征、有些是基于元数据文字描述、有些是基于高级语义特征

语义鸿沟（Semantic Gap）

Raw Media

This is what we have to work with

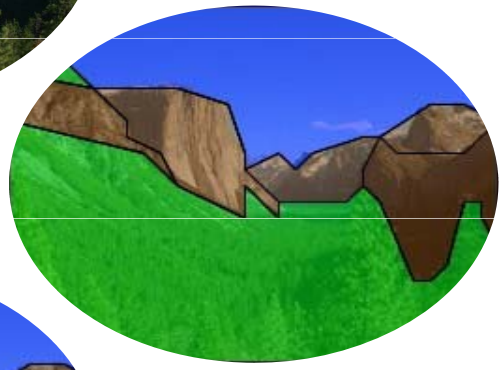
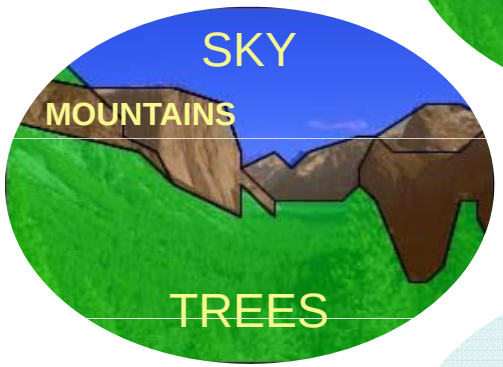


Image-level descriptors



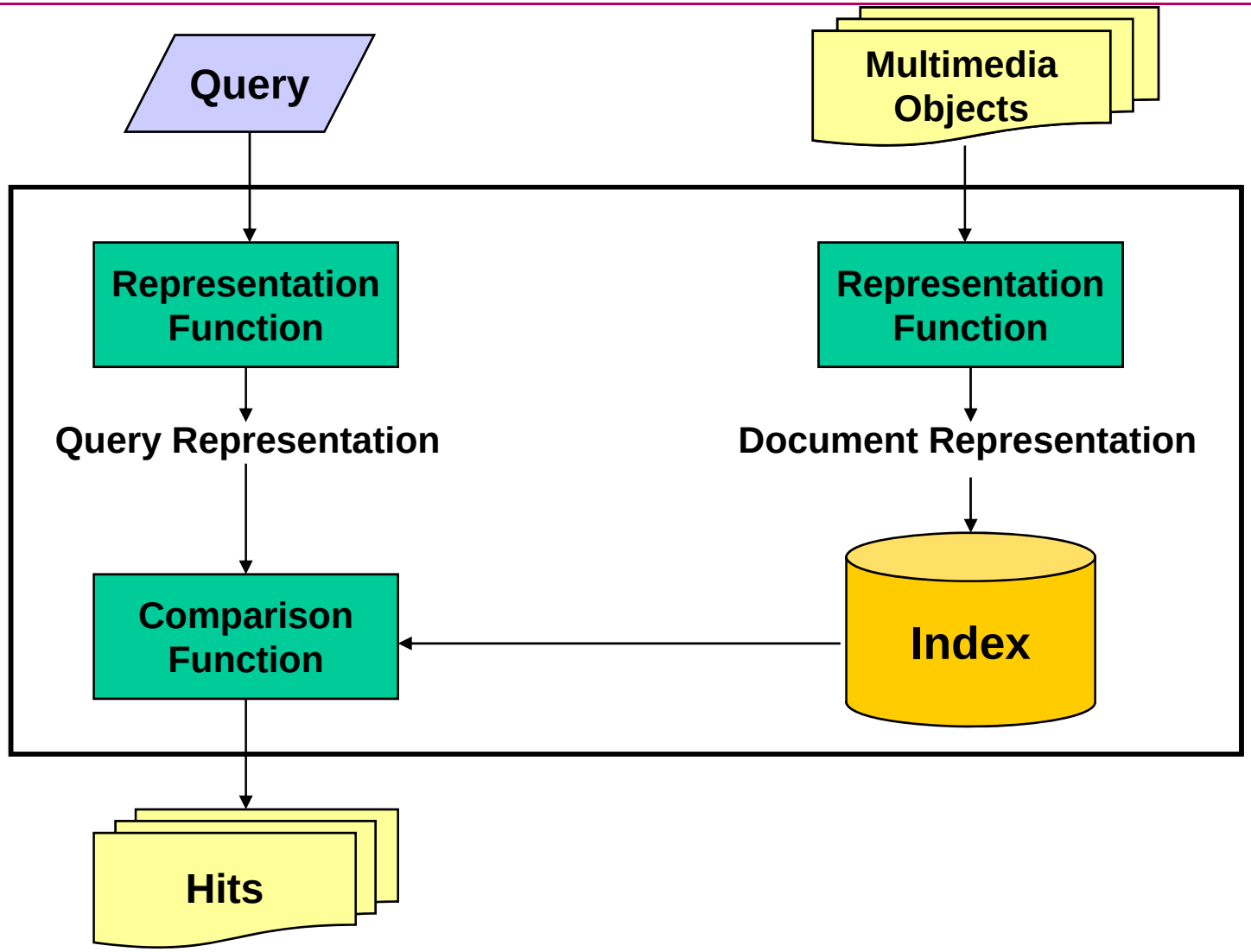
Content descriptors

This is what we want

Photo of Yosemite valley showing El Capitan and Glacier Point with the Half Dome in the distance

Semantic content

信息检索的黑匣子


Query
Multimedia Objects
Representation Function
Representation Function
Query Representation
Document Representation
Comparison Function
Index
Hits

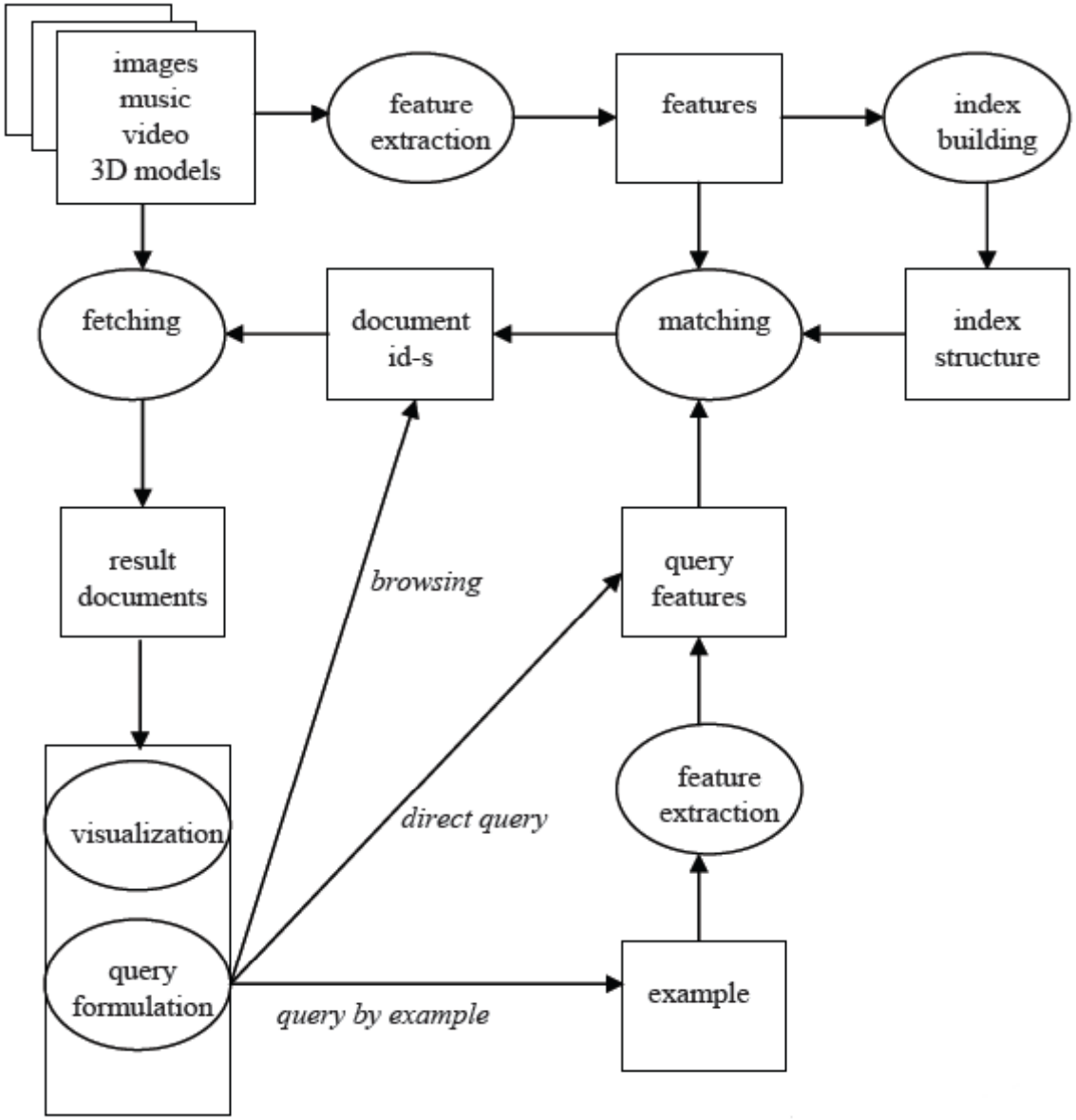
多媒体检索的方法（1）

- 基于关键词检索的方法
  - 人工标注：对多媒体对象进行手工标注，可标注元数据（作者、标题、日期等）或者内容数据（内容关键词）。如**WEB2.0**中提交多媒体对象时的标签（**tag**）数据就是标注文本
  - 自动抽取：
    - 在多媒体对象周围抽取能够表示对象的文本数据用于标注。如在**WEB**中通过图片周围的文字来描述图片
    - 在视频中抽取字幕、对话，从音频中抽取语音，从图片中识别文字等等

多媒体检索的方法（2）

- 基于内容的方法（**Content Based Retrieval，CBR**）
  - 从多媒体对象的内容出发，抽取它们的特征并进行特征表示，在特征层面上进行相似度计算，得到检索结果。
  - 如：基于颜色或形状的图像检索、哼一句歌找整支歌曲、基于概念的检索（如：检索有关“日出”的图片)

多媒体检索的一般框架


images
music
video
3D models
feature extraction
features
index building
fetching
document id-s
matching
index structure
result documents
browsing
query features
visualization
direct query
feature extraction
query formulation
query by example
example

多媒体对象中的特征

- 视觉类媒体的特征：颜色、形状、纹理、空间约束、运动、对象（如太阳）、场景、语义（如日出）等等
- 听觉类媒体的特征：音调、音量、音色、旋律、和谐度、语义（如爆炸声）等

“bag of words” to “bag of features”

	维数	特性	布尔运算	语义
文字	超高(10万级)	稀疏	可	离散
多媒体	高(几千以内)	致密	不可	连续

图像

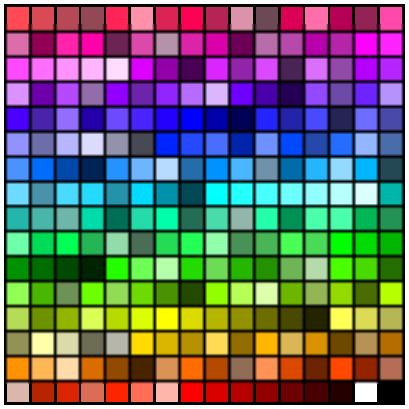
- 二维材料经扫描器扫描、拍照或编辑产生数字化图像。图像的主要规格包括分辨率、颜色表示位数、存储格式、压缩手段等等
- 图像包括：照片**(photo)**、图片**(picture)**、位图**(bitmap)**、电脑绘图**(graphics)**、视频中的帧**(frame)**

A Picture ...

... is comprised of pixels

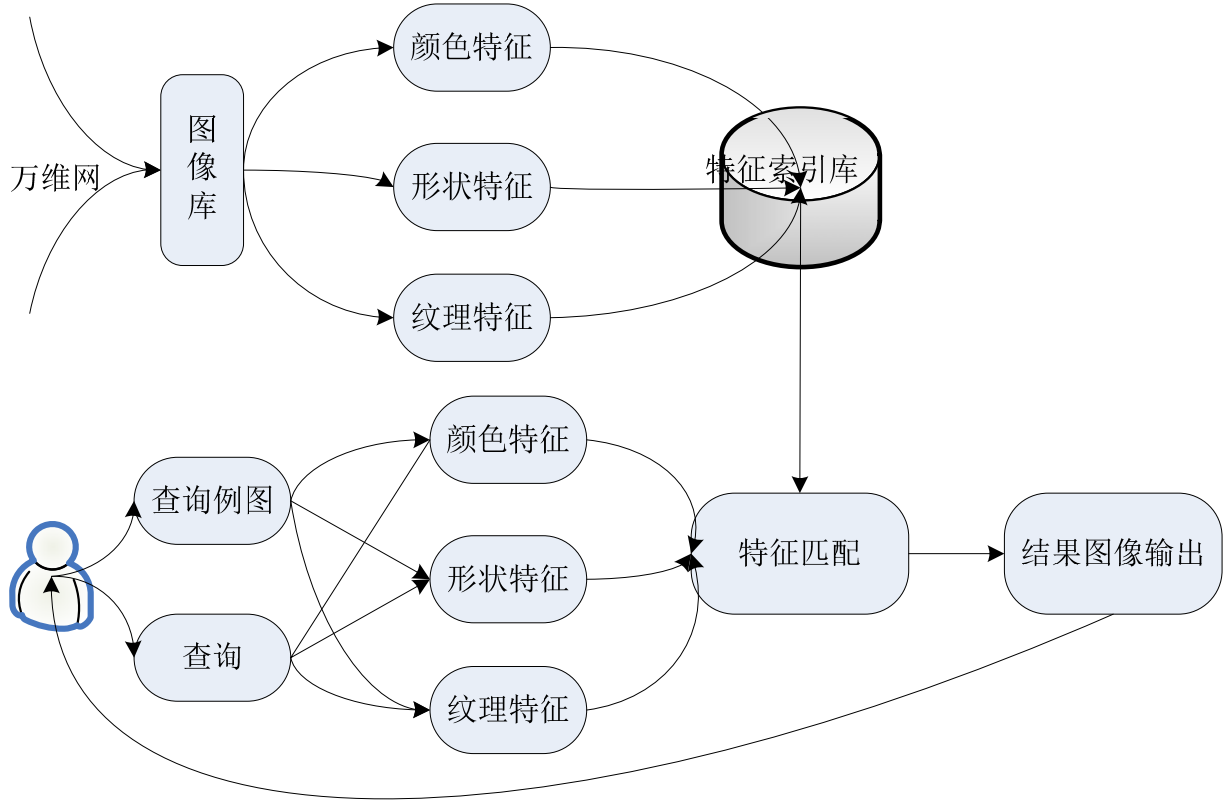
数字图像基本元素

- 像素：组成数字图像的基本元素
  - 黑白图像：每个像素值只能取1或0
  - 灰度图象：每个像素值从0~255取值
  - 彩色图像：每个像素值包含三个分量RGB，每个分量从0~255取值
- 格式：bmp、jpg、gif



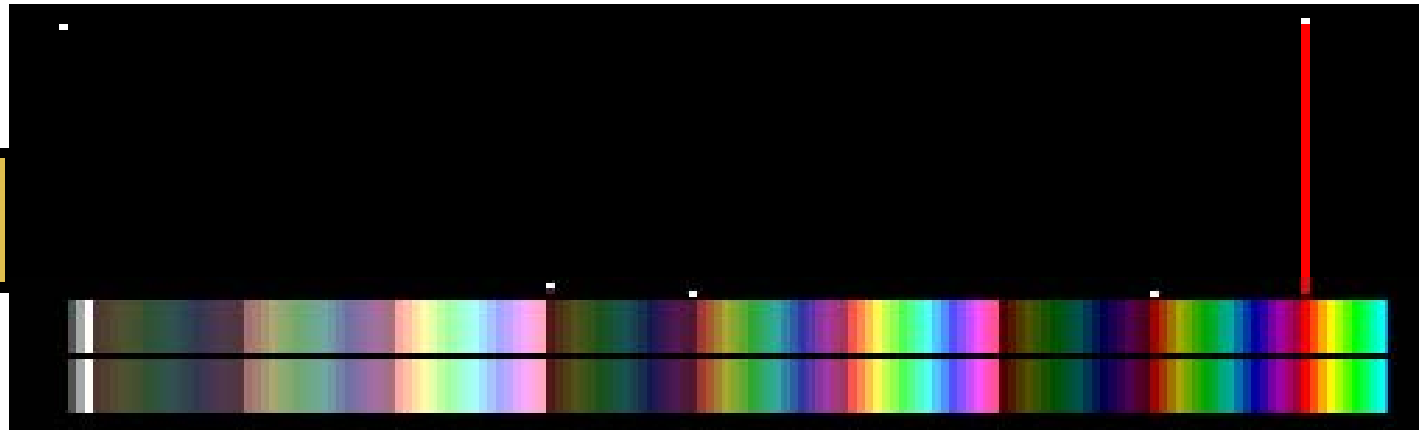
基于内容的图像检索流程

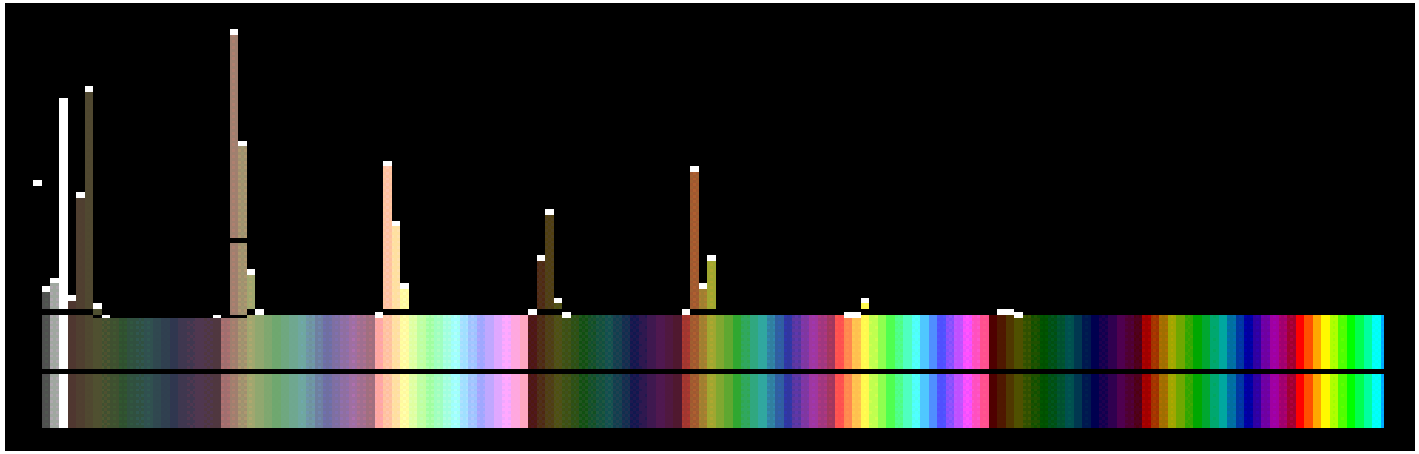
- **Content based image retrieval（CBIR）**



颜色特征

- 颜色模型
  - **HIS**
  - **RGB**
- 颜色直方图

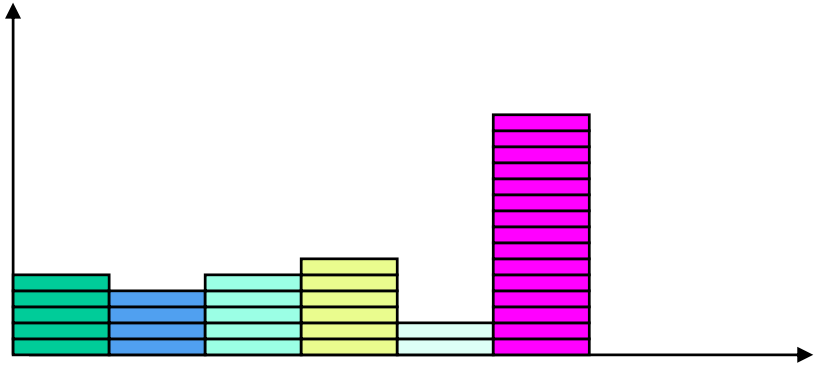




直方图概念

1	2	3	4	5	6
6	4	3	2	2	1
1	6	6	4	6	6
3	4	5	6	6	6
1	4	6	6	2	3
1	3	6	4	6	6

1	2	3	4	5	6
5	4	5	6	2	14

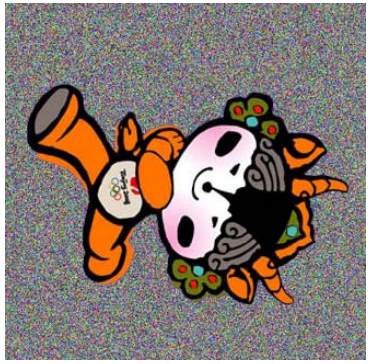


灰度直方图

颜色直方图的例子

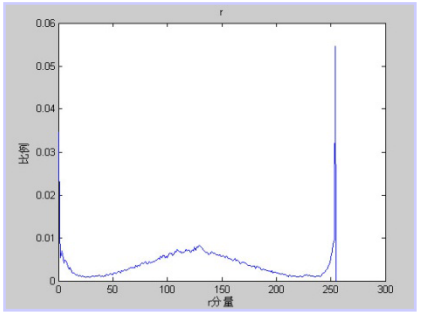
原图

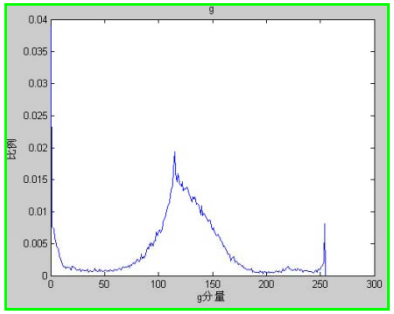
缩小图

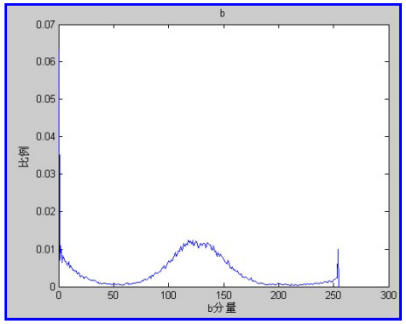


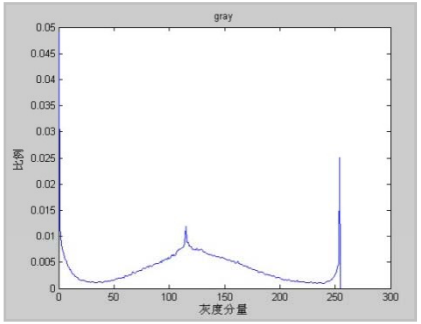
旋转图

位移图







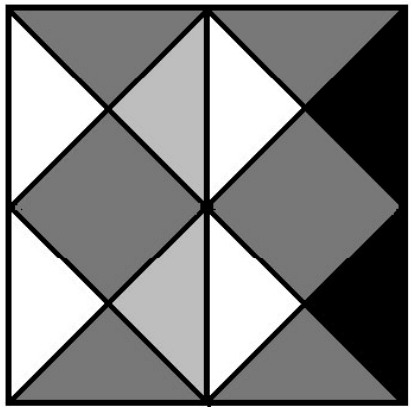


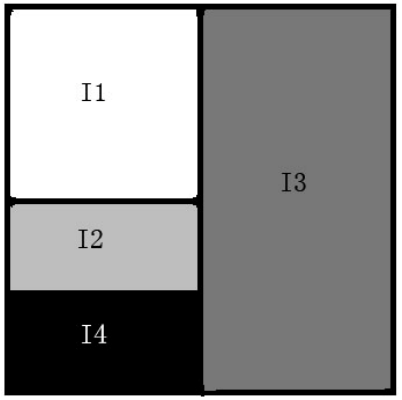
四帧图的R、G、B和灰度的直方图都一样！

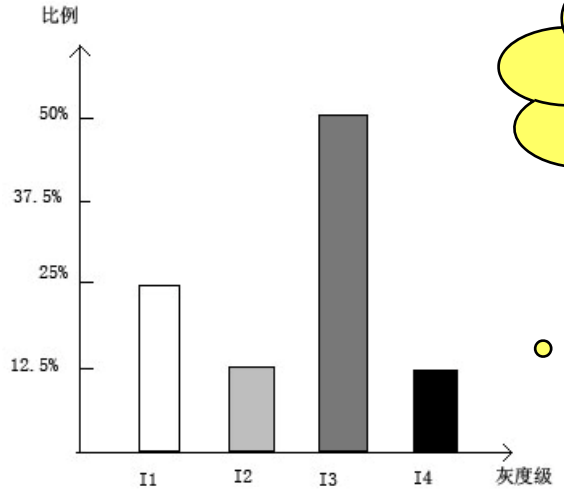
颜色直方图的特点

- 缩放不变性：图像进行了缩放，不引起颜色直方图变化
- 旋转不变性：图像进行了旋转，不引起颜色直方图的变化
- 位移不变性：图像进行了移动，不引起颜色直方图的变化
- 双峰特性：如果图像中的前景和背景分明，直方图出现明显的双峰特性
- 优点
  - 较成熟，广泛使用
- 缺点
  - 丧失了颜色的空间信息
- 改进
  - 累加直方图
  - 分块

丧失了空间信息的典型实例



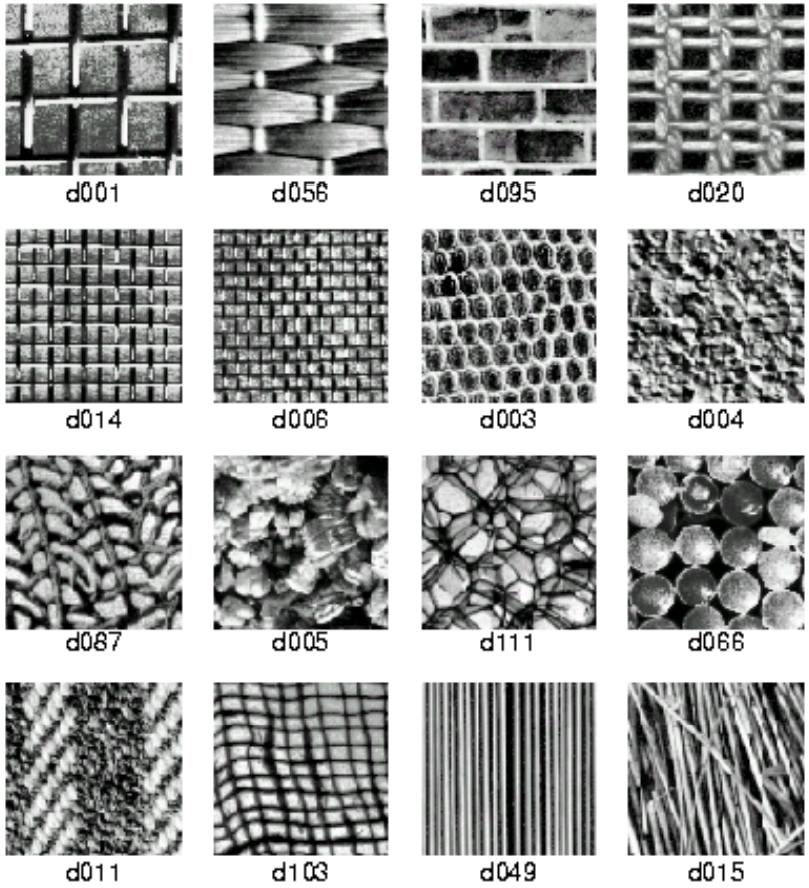




不同的图，相同的直方图

纹理特征

- 某颜色或密度模式的改变



纹理特征提取

- 结构方法是以图像本身结构为基础的，主要针对规则的结构纹理，提取和分析纹理基元
- 统计方法是以人的直观感觉为基础的。它根据像素灰度的统计特征确定纹理特征，如直方图统计特征法、自相关函数法等
  - **Tamura**纹理特征
- 无论从历史发展还是从当前进展来看，纹理的统计分析方法仍然占主导地位

形状特征

- 形状特征提取方法
  - 傅立叶描述子
  - 几何特征：区域的面积、圆形度、形状的纵横比（**AspectRation**）等
  - 不变矩（**Invariant Moment**）

相似度计算

- **Minkowski测度**

$$D(e,d)=(\sum_{k}|He(k)-Hd(k)|^{p})^{1/p}$$

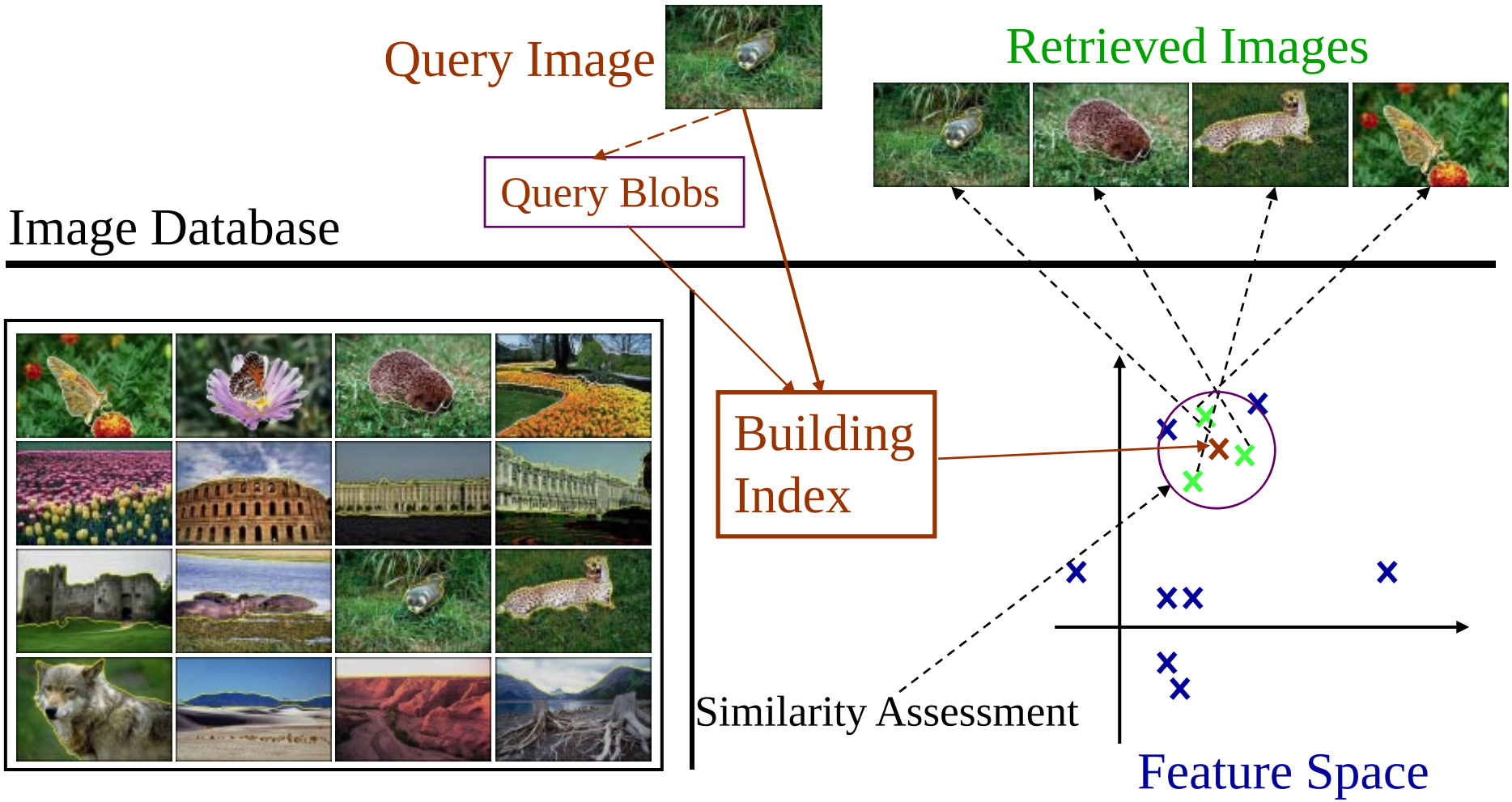
  - $H_e(k)$和$H_d(k)$分别为图像d，e的特征量（向量）
  - p=1,2,3时，对应的分别被称为L1，L2（欧式距离），L3距离

- **Mahalanobis测度（马氏距离）**

$$D(e,d)=\sqrt{(He-Hd)^{T}C^{-1}(He-Hd)}$$

  - C表示特征向量协方差矩阵

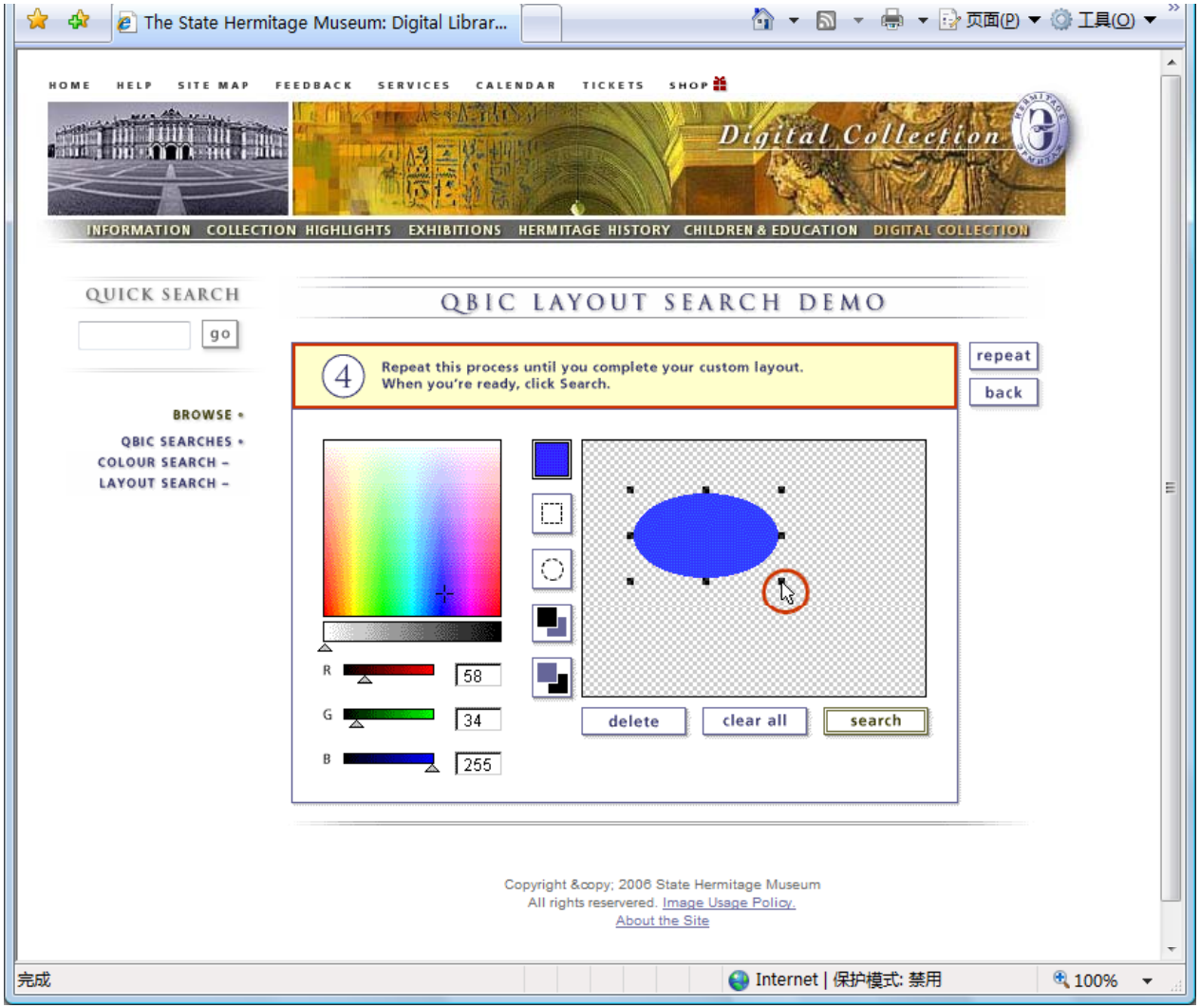
基于内容的图像检索CBIR



查询方式

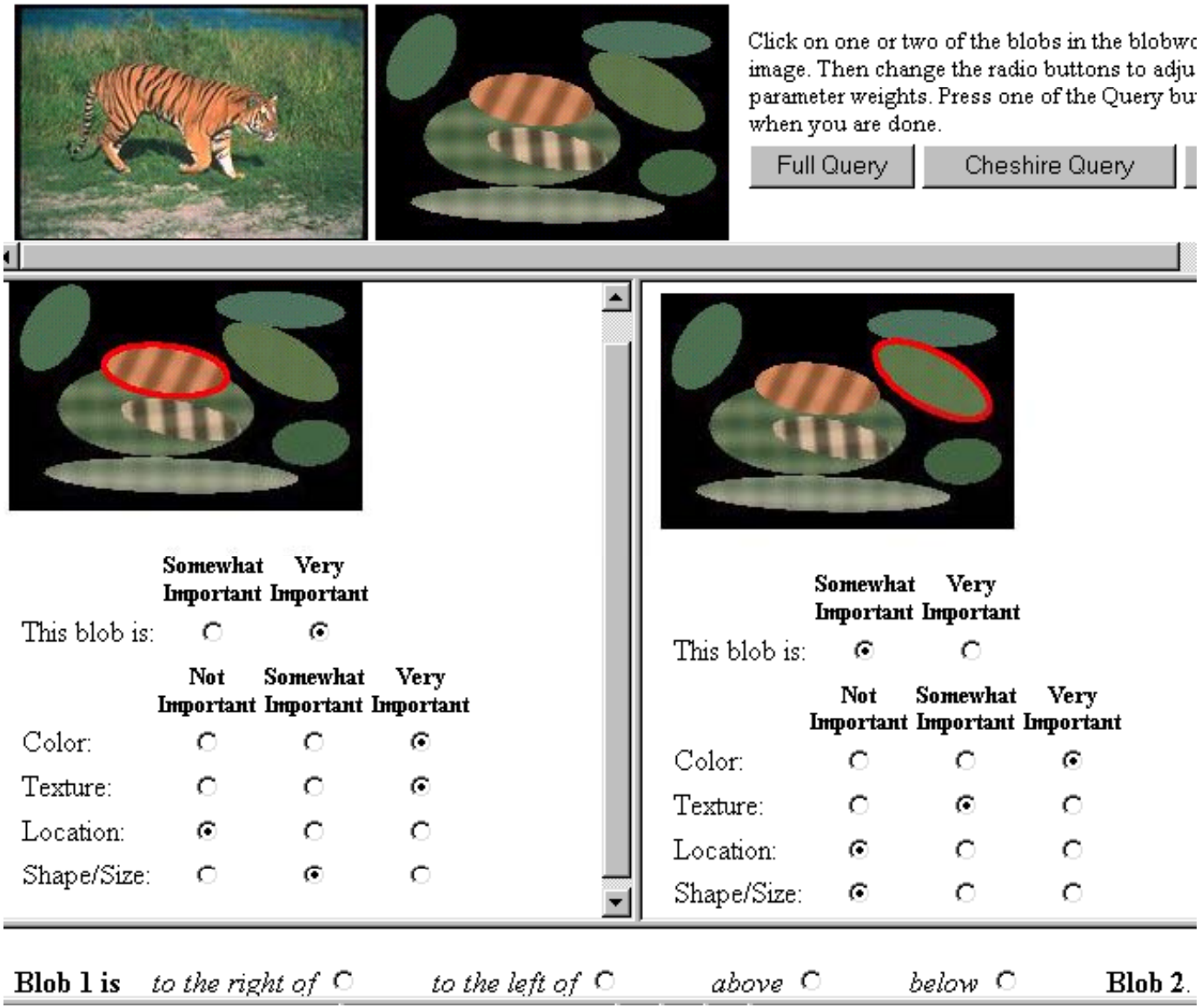
- 样例：根据库中或者库外已有图像或者人工绘制的图像进行检索。比如通过输入一个红色圆形物体来检索相似的图像
- 绘图：手工绘制草图用于检索。如通过勾画衣服形状对服装设计图进行检索
- 属性说明方式：指定特征进行检索。如通过限定人的脸形、五官特征从人脸库中进行检索

IBM的QBIC：基于布局的检索（Layout Search）

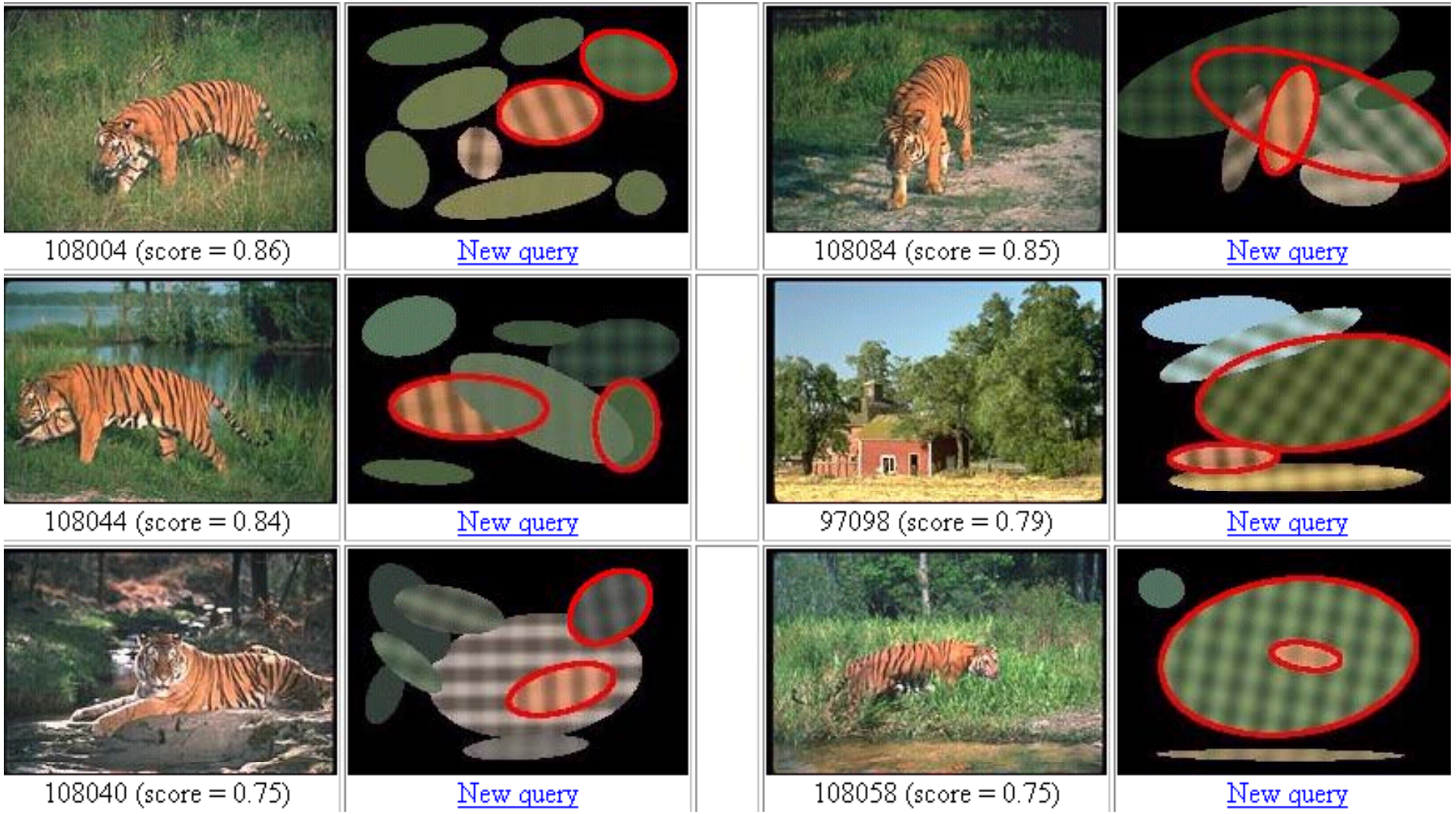


http://www.hermitagemuseum.org/fcgi-bin/db2www/qbicSearch.mac/qbic?selLang=English

Berkeley的Blobworld：基于草图的检索



Berkeley的Blobworld



http://elib.cs.berkeley.edu/blobworld/

Google：基于文字的图像检索

- Google分析图像周围的文字，图像标题以及其他特征来决定图像内容

http://image.google.com

音频（Audio）

- 音频（声音）经过模拟设备记录或再生，成为模拟音频，再经数字化成为数字音频
- 数字音频的主要规格为：采样率（**sampling rate**）及每个样本的位数（**bits per sample**）
- 我们能够听见的音频频率范围是**60Hz～20kHz**，其中语音（**speech**）大约分布在**300Hz～4kHz**之内，而音乐（**music**）和其他自然声响是全范围分布的

音频中的特征层次

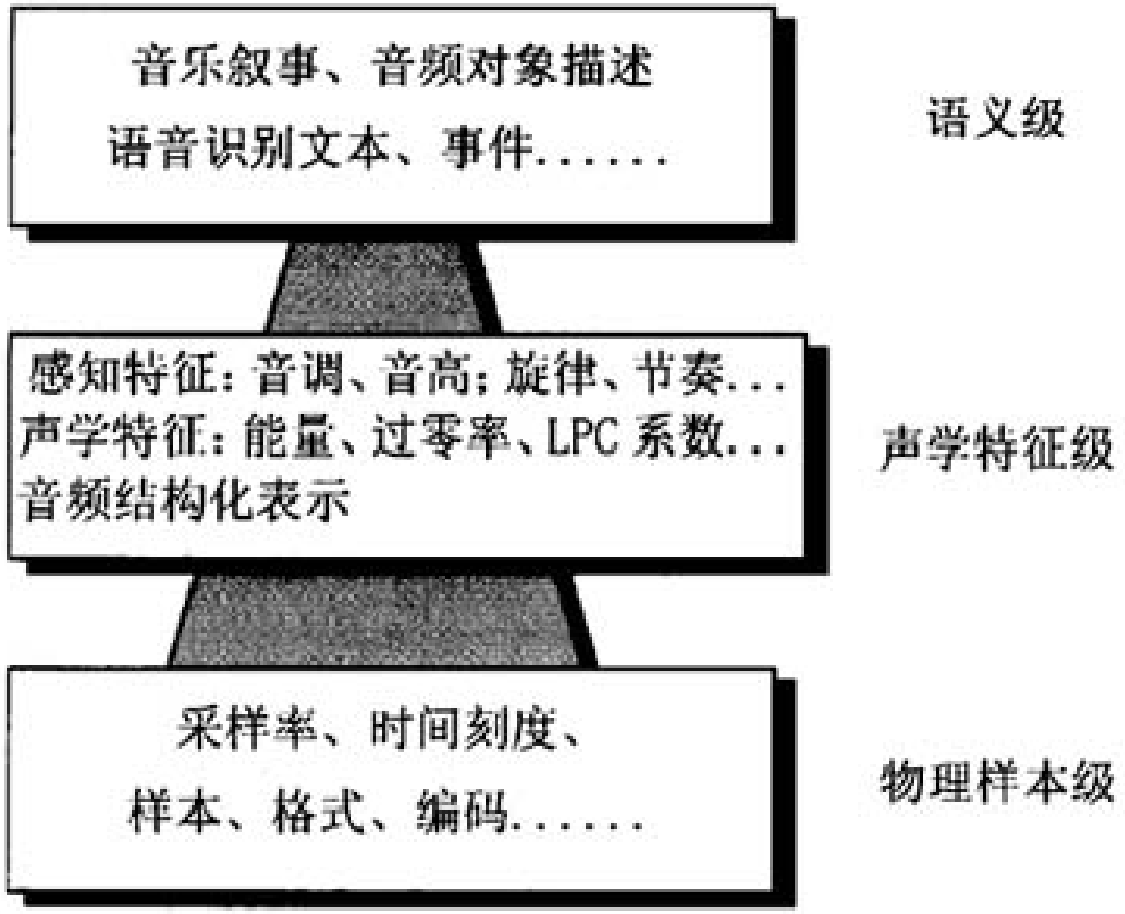


图　音频内容分层描述模型

查询方式

- 样例：用户选择一个声音例子表达其查询要求，查找出与该声音在某些特征方面相似的所有声音。如查询与飞机的轰鸣声相似的所有声音
- 直喻：通过选择一些声学/感知物理特性来描述查询要求，如亮度、音调和音量等
- 拟声：发出与要查找的声音性质相似的声音来表达查询要求。如用户可以发出嗡嗡声来查找蜜蜂或电气嘈杂声
- 主观特征：用个人的描述语言来描述声音。这需要训练系统理解这些描述术语的含义，如用户可能要寻找“欢快”的声音

检索的分类

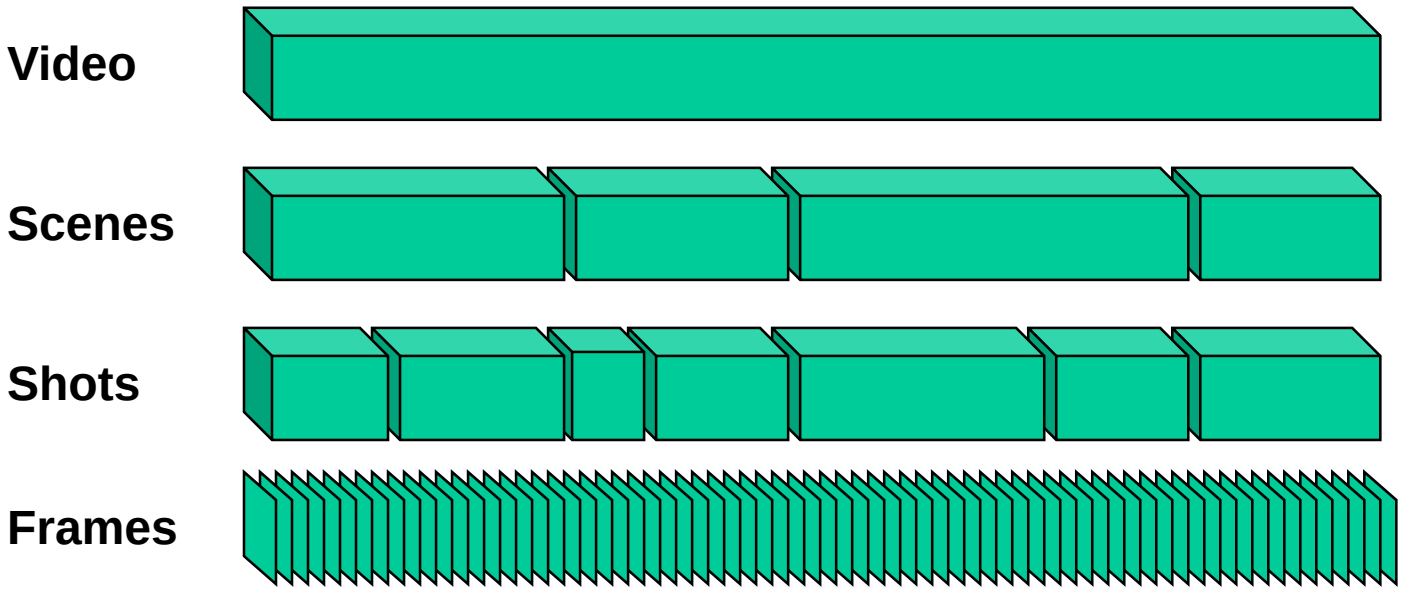
- 语音检索：利用语音识别（**Speech Recognition**）技术，从语音中获取全部文本或者关键文本、或者辨别说话人
- 普通音频检索：以波形声音为对象的检索，这里的音频可以是汽车发动机声、雨声、鸟叫声，也可以是语音和音乐等，这些音频都统一用声学特征来检索
- 音乐检索：以音乐为中心的检索，利用音乐的音符和旋律等音乐特性来检索。如检索乐器、声乐作品等

视频（Video）

- 主要通过视频采集卡从播放画面中采集加工而成。可以看成是在普通图像上增加了时间维度。主要的规格包括：分辨率、每秒播放帧数、压缩方法等
- 常见的视频格式：**.dat、.mov、.rm、wmv、mpg、mpeg等等**
- 每秒播放帧数：电视是**30**帧，电影为**24**帧，对人的感觉而言，至少要每秒**12**帧以上
- 压缩方法：**MPEG （Motion Picture Experts Group）、国内AVS**

视频的特征层次

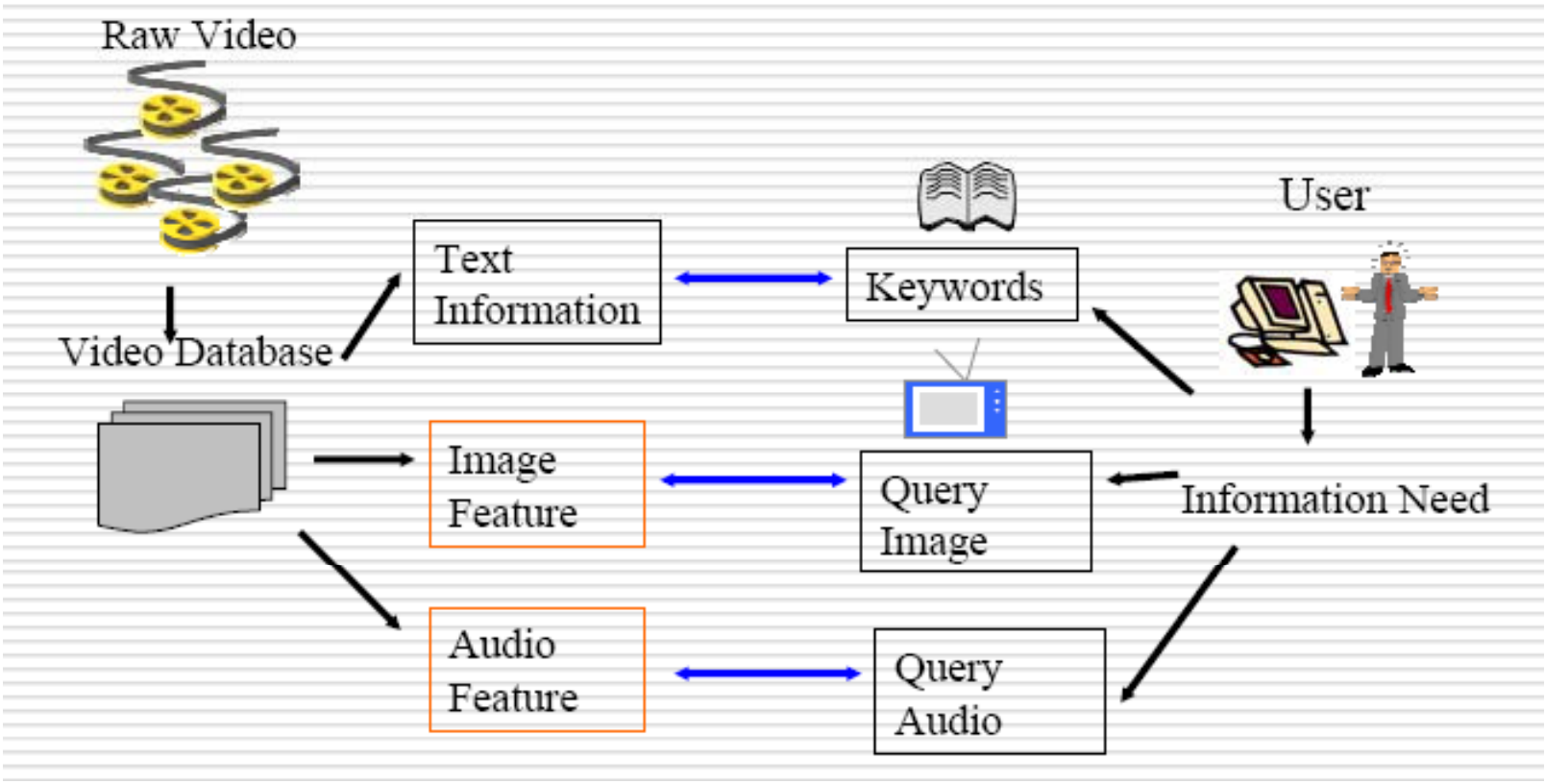
- 帧(Frame)：每个帧可以看成一幅静态图像
- 镜头(Shot)：由连续的帧组成的一个基本拍摄操作单元。镜头可以通过关键帧表示，摄像机操作引起的镜头运动特征也是视频检索中重要的特征内容
- 场景(Scene)：由连续的多个内容相似的镜头组成的一个有意义的单元。场景关键帧可以由镜头关键帧组合而成。关键对象也可以组合



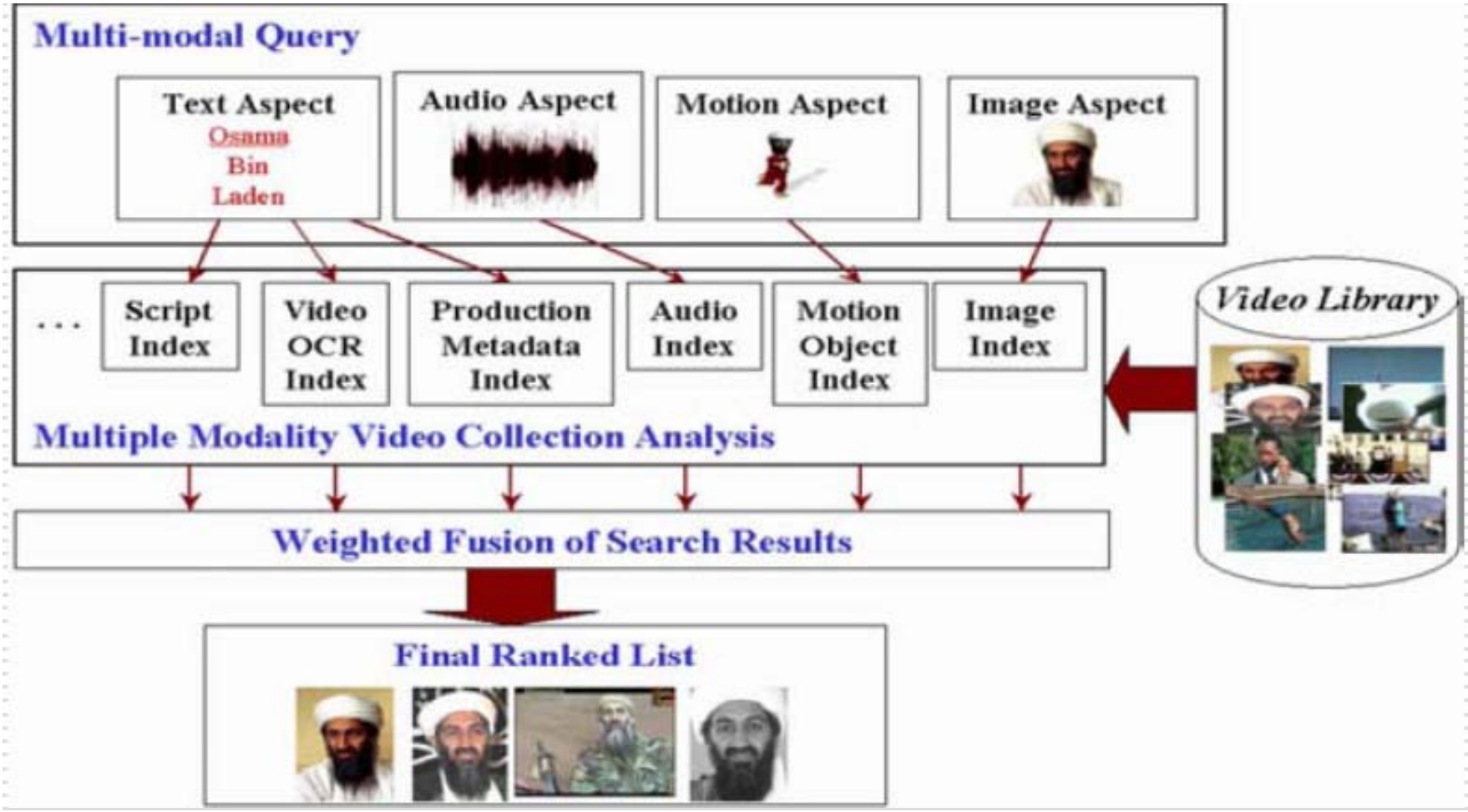
视频的检索

- 基于关键帧的检索：类似于图像检索的方法，利用全部和局部的图像特征进行检索
- 基于运动特征的检索：基于摄像机运动或者像素运动特征的检索
- 基于视频对象的检索：利用视频对象的特性，从库中检索出包含相关视频对象的所有场景或者镜头

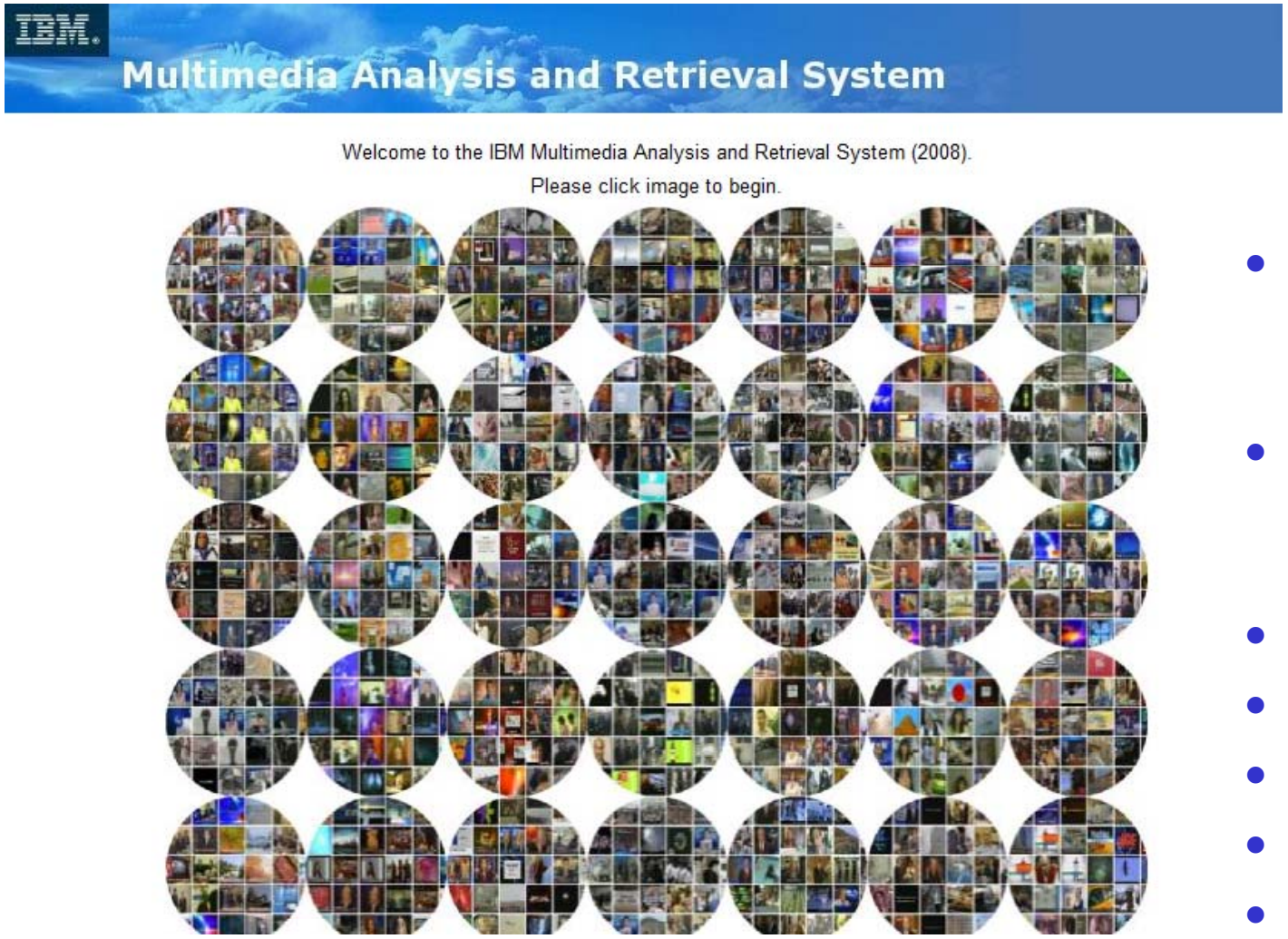
跨媒体检索


Raw Video
Video Database
Text Information
Image Feature
Audio Feature
Keywords
Query Image
Query Audio
User
Information Need

多媒体综合检索服务


Multi-modal Query
Text Aspect
Osama
Bin
Laden
Audio Aspect
Motion Aspect
Image Aspect
...
Script Index
Video OCR Index
Production Metadata Index
Audio Index
Motion Object Index
Image Index
Multiple Modality Video Collection Analysis
Video Library
Weighted Fusion of Search Results
Final Ranked List

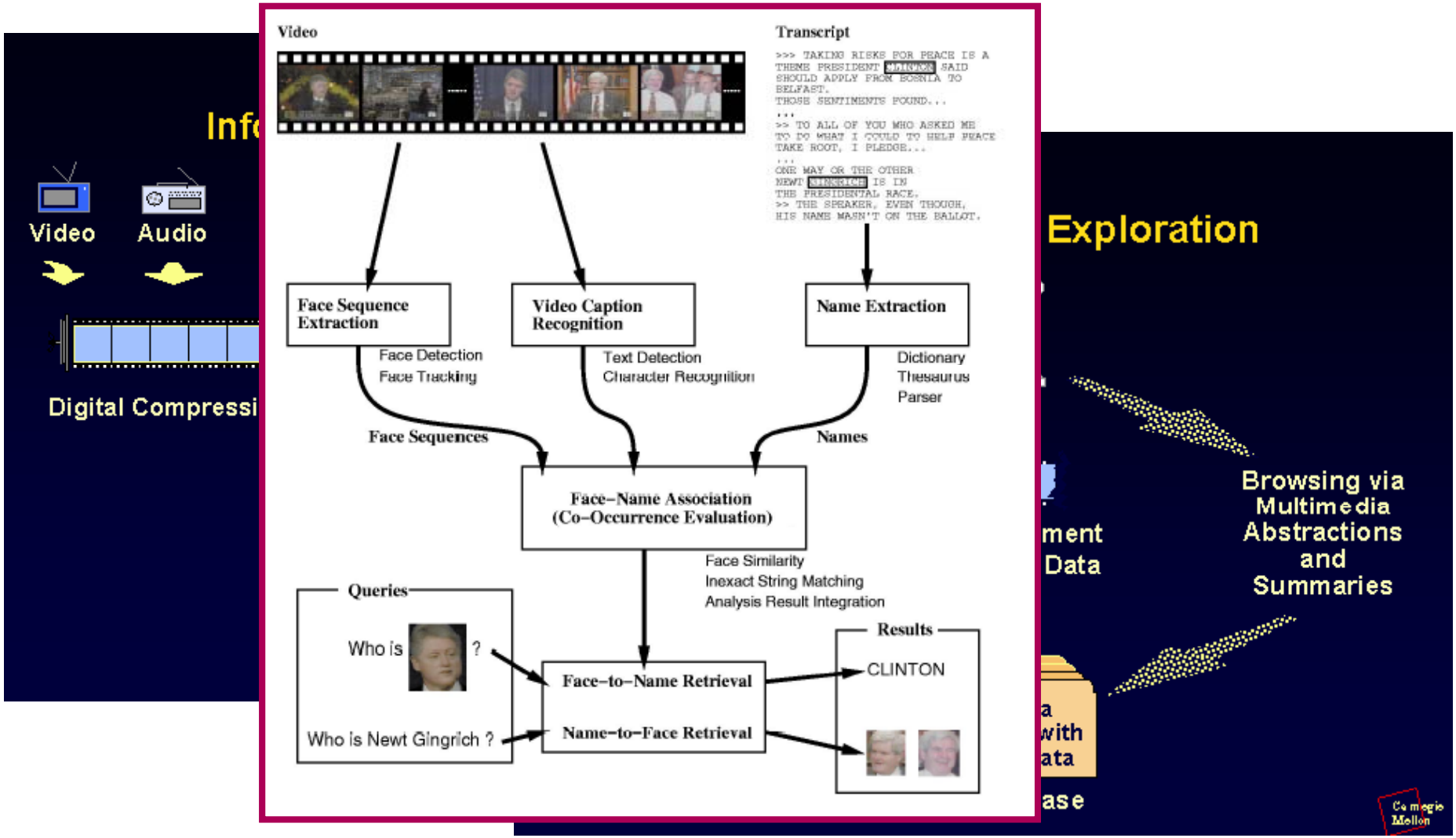
IBM视频搜索引擎



- **MPEG-7** 视频搜索引擎
- 人工标注用于机器学习
- 多模索引
- 图像处理
- 语音识别
- 结构构模
- 特征聚类

http://mp7.watson.ibm.com/

CMU的Informadia



http://www.informedia.cs.cmu.edu/

视频检索评估

- 基于内容的视频检索评测
  - **TREC Video Track （2001，2002）**
  - **TRECVID（至今），http://www-nlpir.nist.gov/projects/trecvid/**

Modality	MAP
Baseline: ASR + Closed Captions (CC)	0.155
ASR + CC + Video OCR	0.177
ASR + CC + VOCR + Image Similarity weighted by query type	0.198
ASR + CC + VOCR + Image Similarity weighted by development set query results	0.207
ASR + CC + VOCR + Image Similarity weighted by development set query results + Person X retrieval	0.218

A. Hauptmann and M. Christel. (2004) Successful Approaches in the TREC Video Retrieval Evaluations. Proceedings of ACM Multimedia 2004.

小结：多媒体检索

- 以搜索引擎为代表的文本检索已经深入人心，得到了用户的认可
- 而多媒体检索却由于技术上的难度目前在应用上并没取得突破，离用户的要求还有较大的距离
- 各大公司投入很大力量进行多媒体检索的研发
- 多媒体检索成为竞争焦点

主要内容

- 多媒体检索
- 跨语言检索
- 问题回答

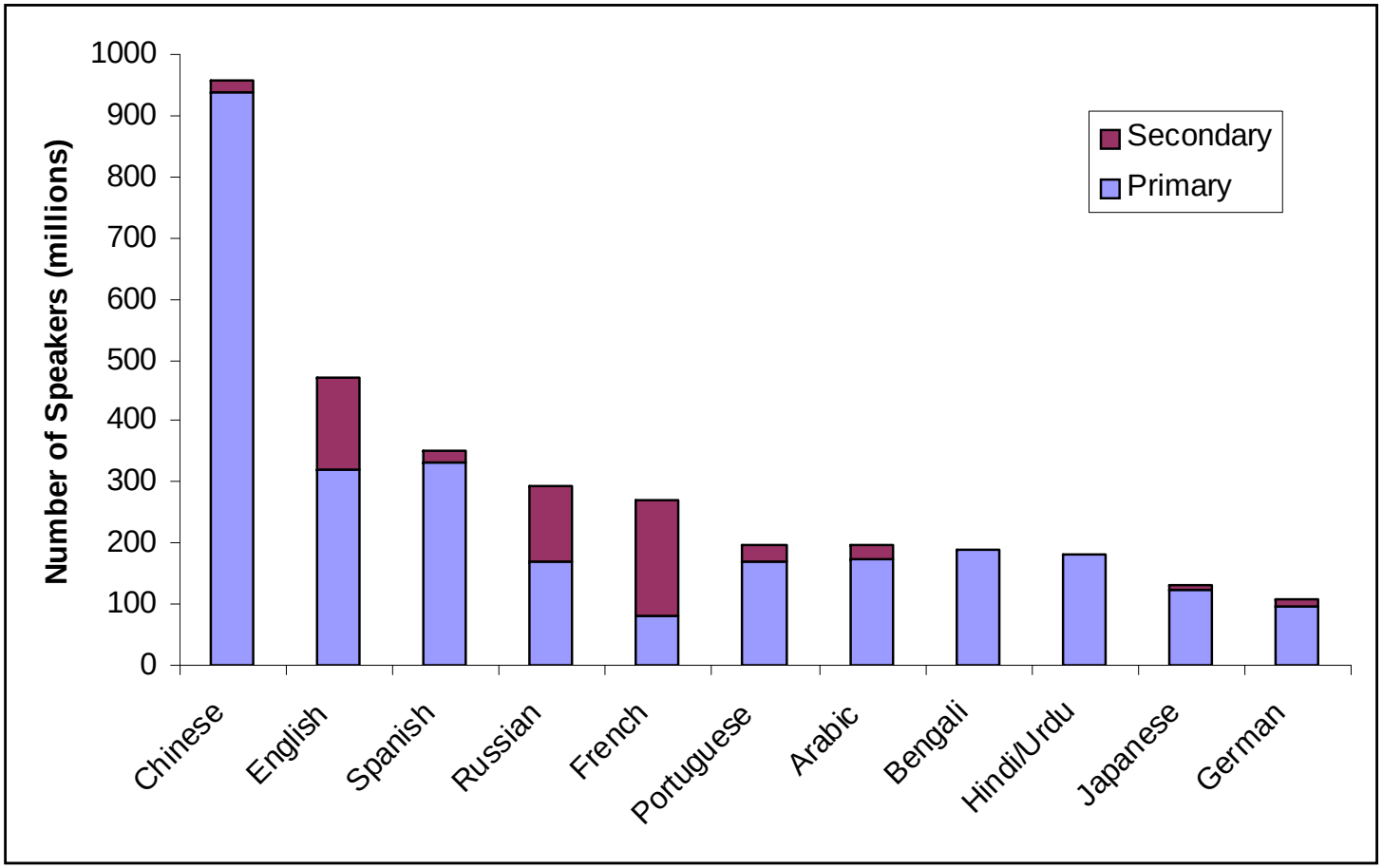
概述

- 跨语言检索(**Cross Language Information Retrieval, CLIR**)：输入某种语言（源语言）的查询，在一个或者多个其他语言（目标语言）的文档库中检索结果
  - 例子：输入“信息检索”，在英文文档库中检索结果
  - 也称为**Cross-lingual IR, Trans-lingual IR**
  - 当只有一种目标语言时，称为双语检索（**Bilingual IR**）
  - 当有多个目标语言时，称为多语言检索（**Multilingual IR**），与此对应，传统的检索称为单语检索（**Monolingual IR** 或者**Single Language IR**）

为什么要进行跨语言检索？

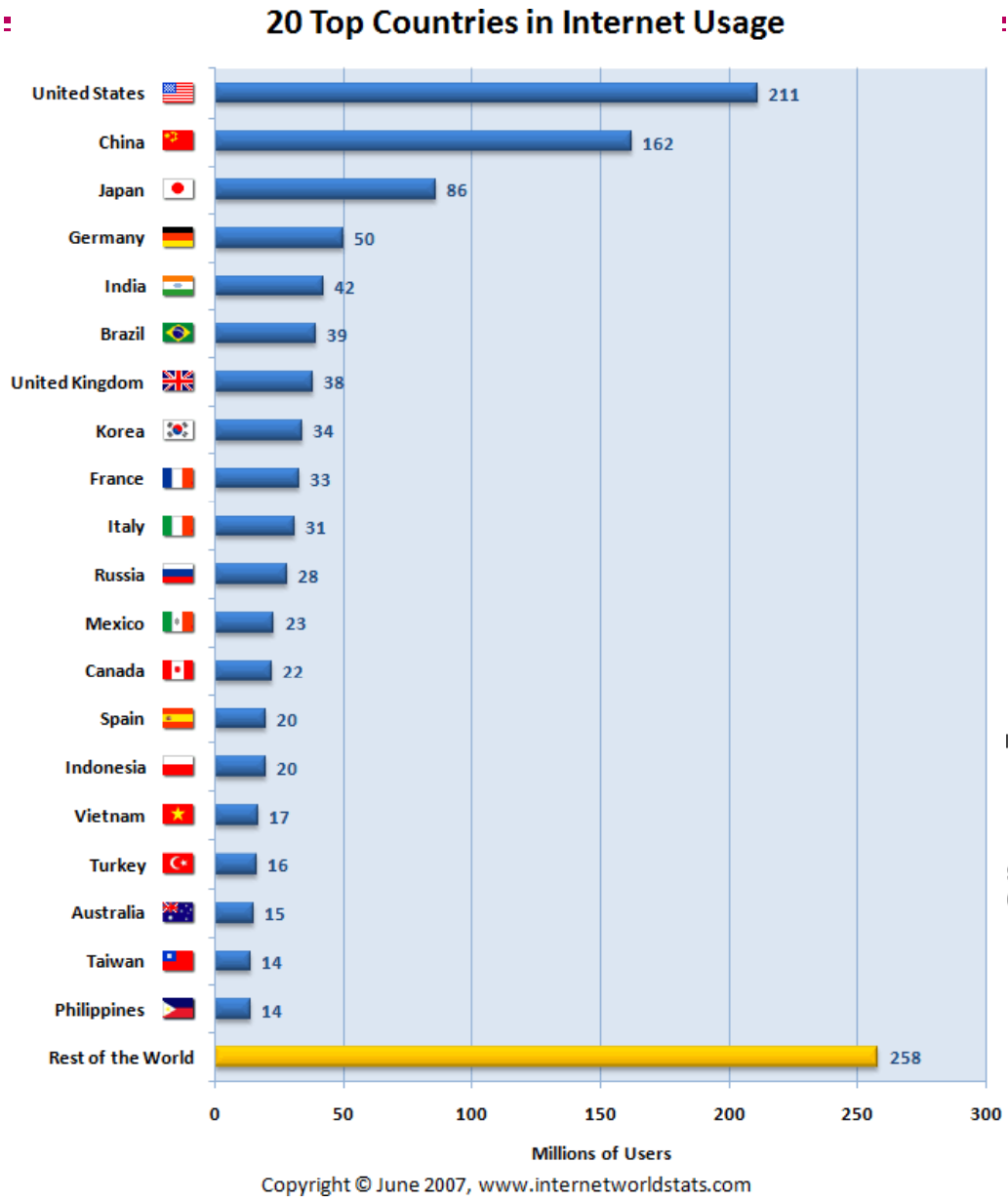
- 人们通常只对自己的母语非常熟悉
- **Web**上拥有大量不同语言表示的文档
- **Web**上拥有采用不同语言标注的多媒体文档
- 人们当然希望只通过输入自己的母语进行查询，而得到所有相关的文档

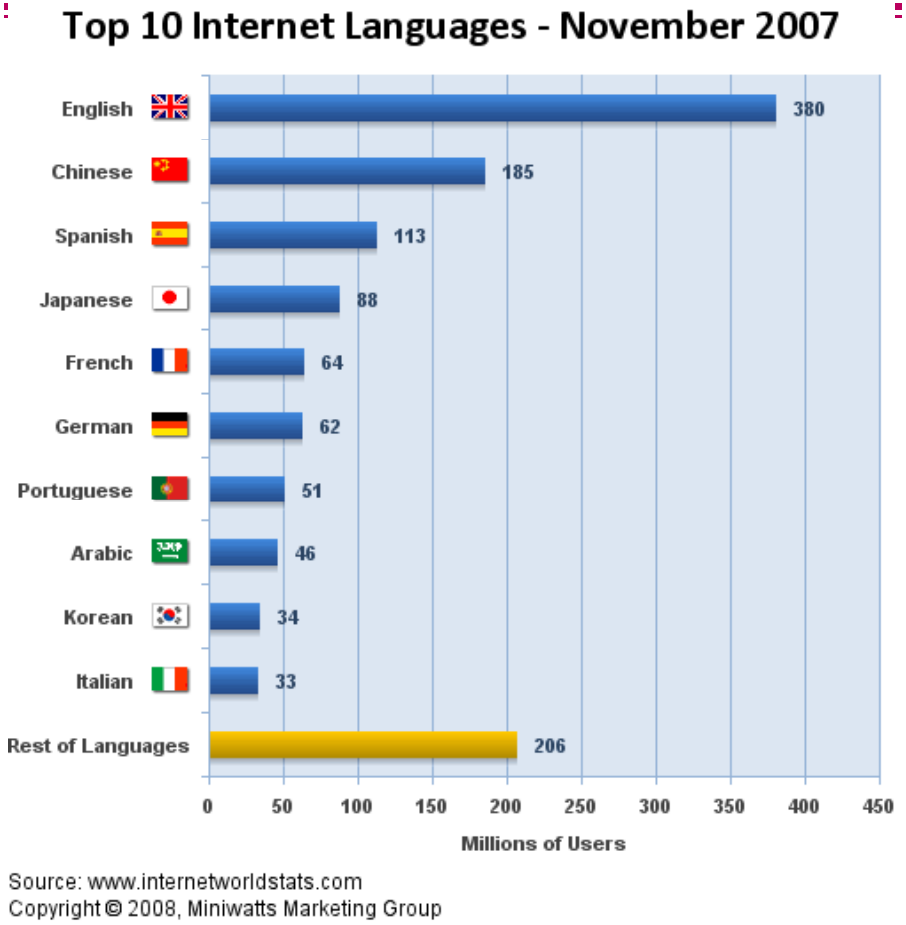
世界上最流行的语言



来源： **Ethnologue (SIL), 1999**

互联网用户及语言类型





Source: www.internetworldstats.com
Copyright © 2008, Miniwatts Marketing Group

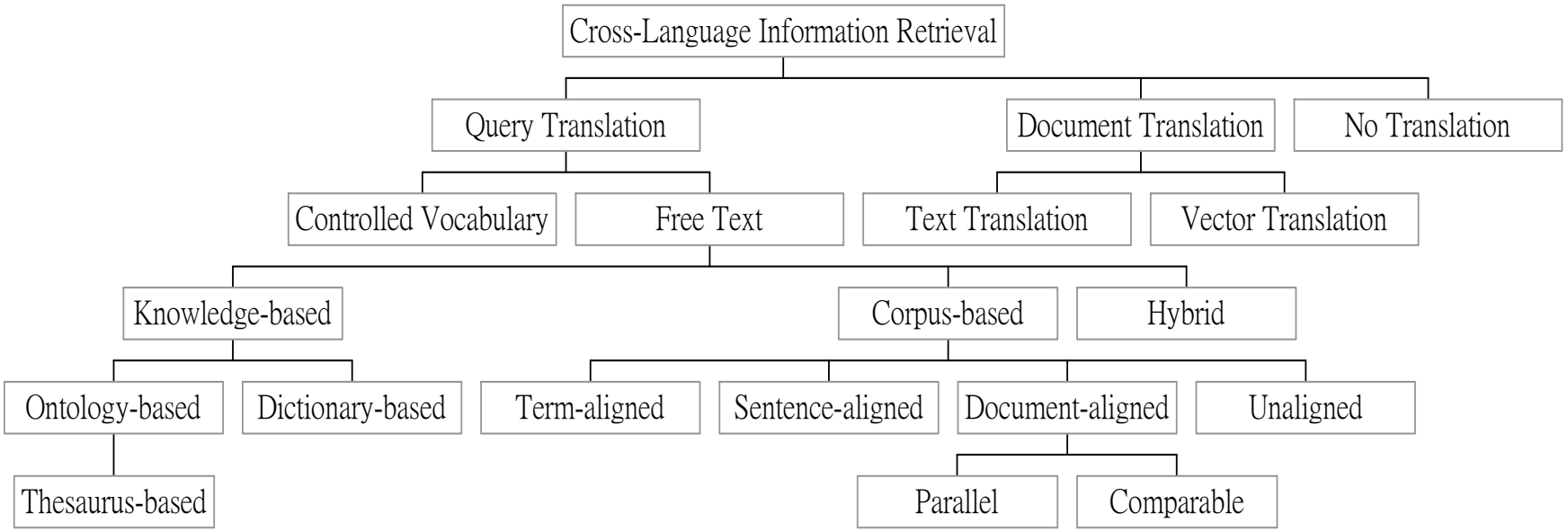
CLIR的应用场合

- 一个研究人员想查其他语言撰写的学术论文；
- 一个旅游者想了解更多的国外相关情况；
- 一个企业想了解国外竞争对手的发展情况；
- ……

CLIR的难点

- **CLIR仍然依赖于单语检索的效果，而单语检索本来就不容易**
  - 目前单语检索仍然主要是基于关键词匹配的方法，不能准确地理解用户的查询需求，检索的准确性仍然不高
- 源语言和目标语言之间可能存在巨大的语言鸿沟
  - 以世界上使用最广泛、使用人口最多的英文**vs.** 中文为例，两种语言不论在词法、句法还是语义处理等方面都有巨大差异
  - 同根语言对之间可能翻译的难度小一些，但是作为不同的语言，仍然具有较大的差异，全自动翻译仍然达不到实用水平

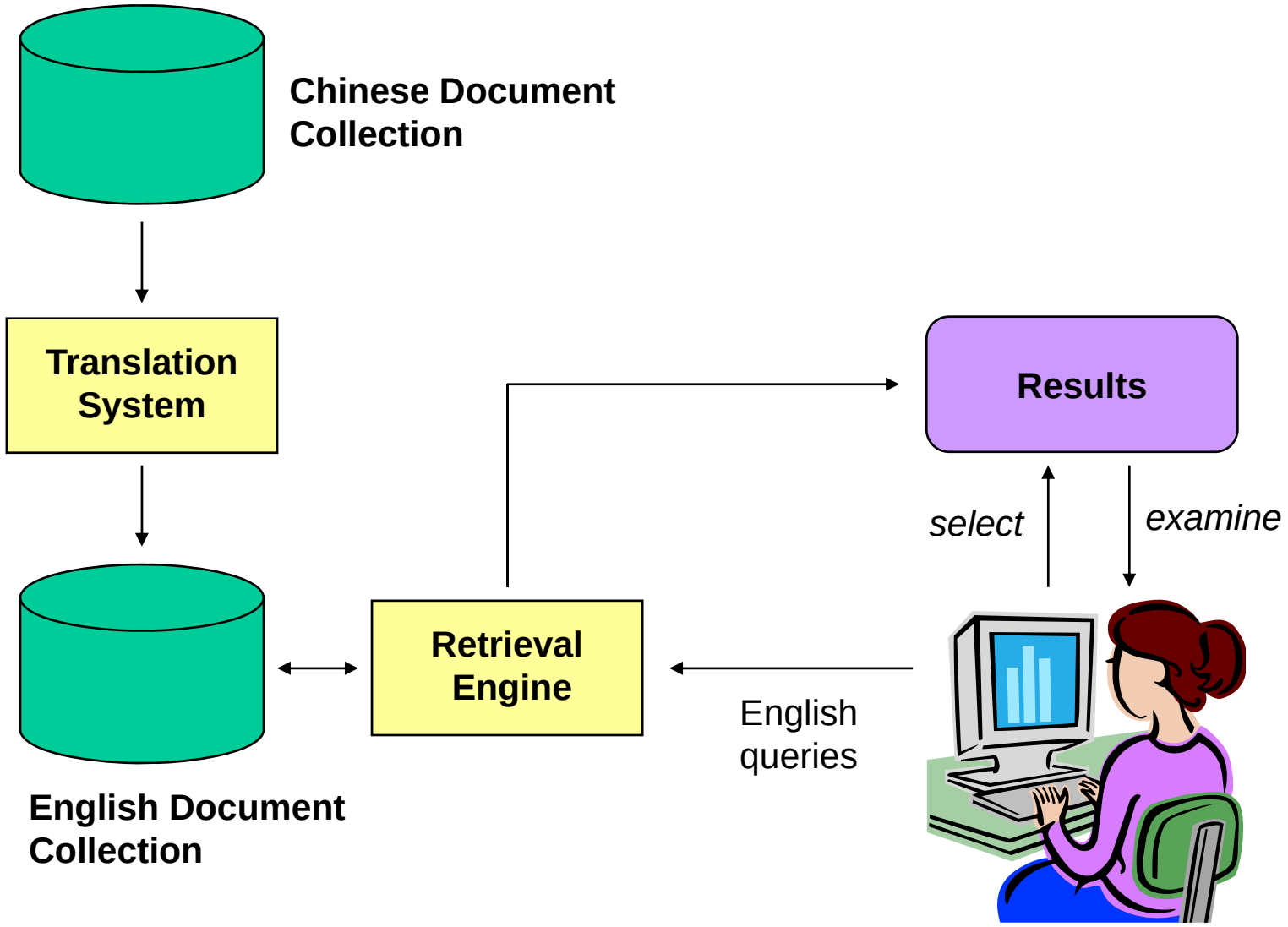
CLIR方法


Cross-Language Information Retrieval
Query Translation
Document Translation
No Translation
Controlled Vocabulary
Free Text
Text Translation
Vector Translation
Knowledge-based
Corpus-based
Hybrid
Ontology-based
Dictionary-based
Term-aligned
Sentence-aligned
Document-aligned
Unaligned
Thesaurus-based
Parallel
Comparable

文档翻译

- 文档翻译：将所有文档翻译成源语言，采用单语检索的方法
- 优点：
  - 检索结果可读
  - 文档的翻译理论上相对准确（可以依靠上下文解决翻译中的歧义）
- 缺点：
  - 翻译量巨大，翻译的时间消耗较大
  - 在多语言检索情况下，需要多个语言对之间的翻译工具

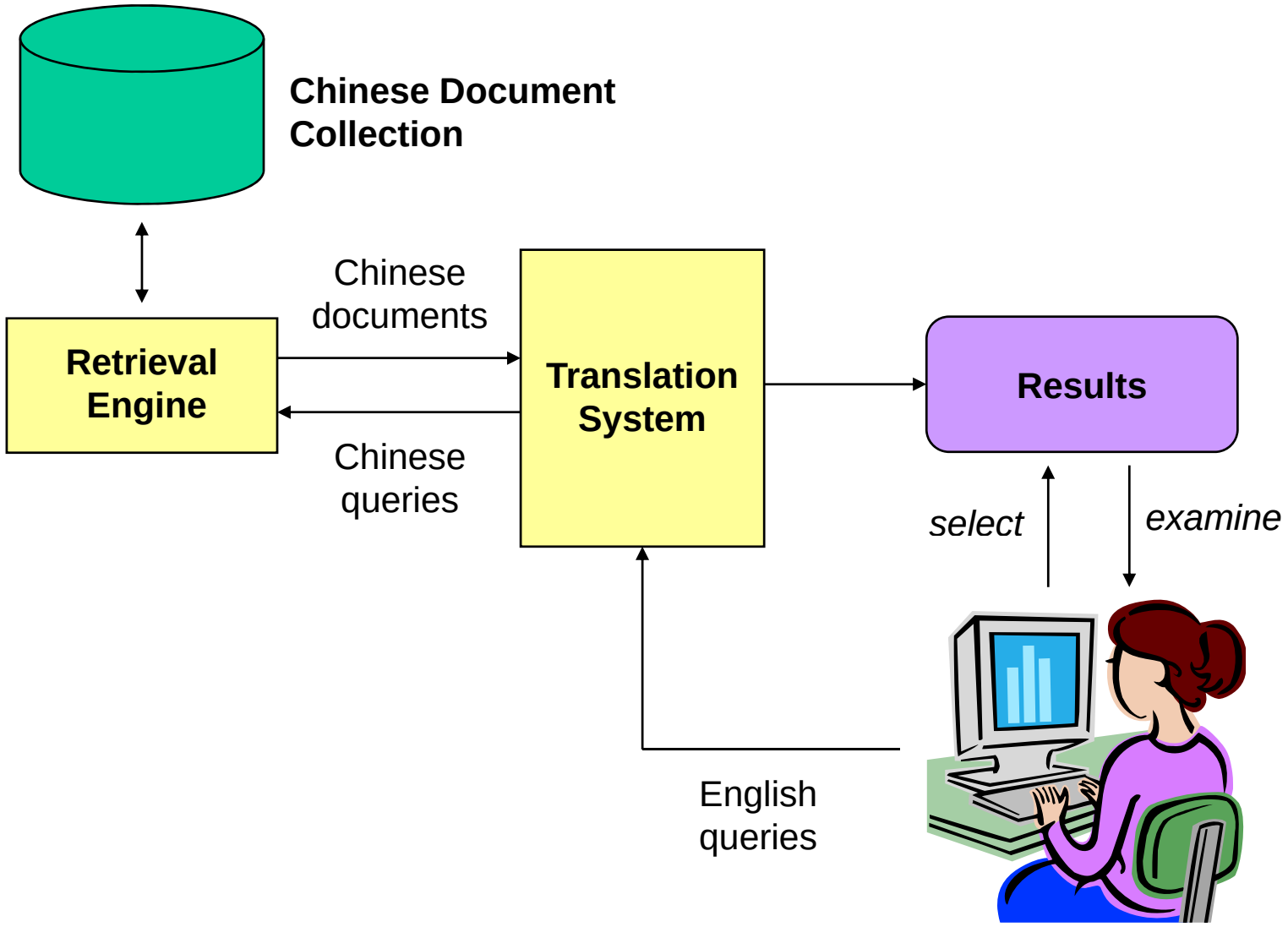
文档翻译图示


Chinese Document Collection
Translation System
English Document Collection
Retrieval Engine
Results
select
examine
English queries

查询翻译

- 查询翻译：将查询翻译成目标语言
- 优点：
  - 翻译量小，相对灵活
- 缺点：
  - 由于查询通常很短，翻译质量难易保证
  - 如果用户不懂目标语言，仍然需要把结果再翻译成源语言。
- 查询翻译方法是目前**CLIR**中的主要方法

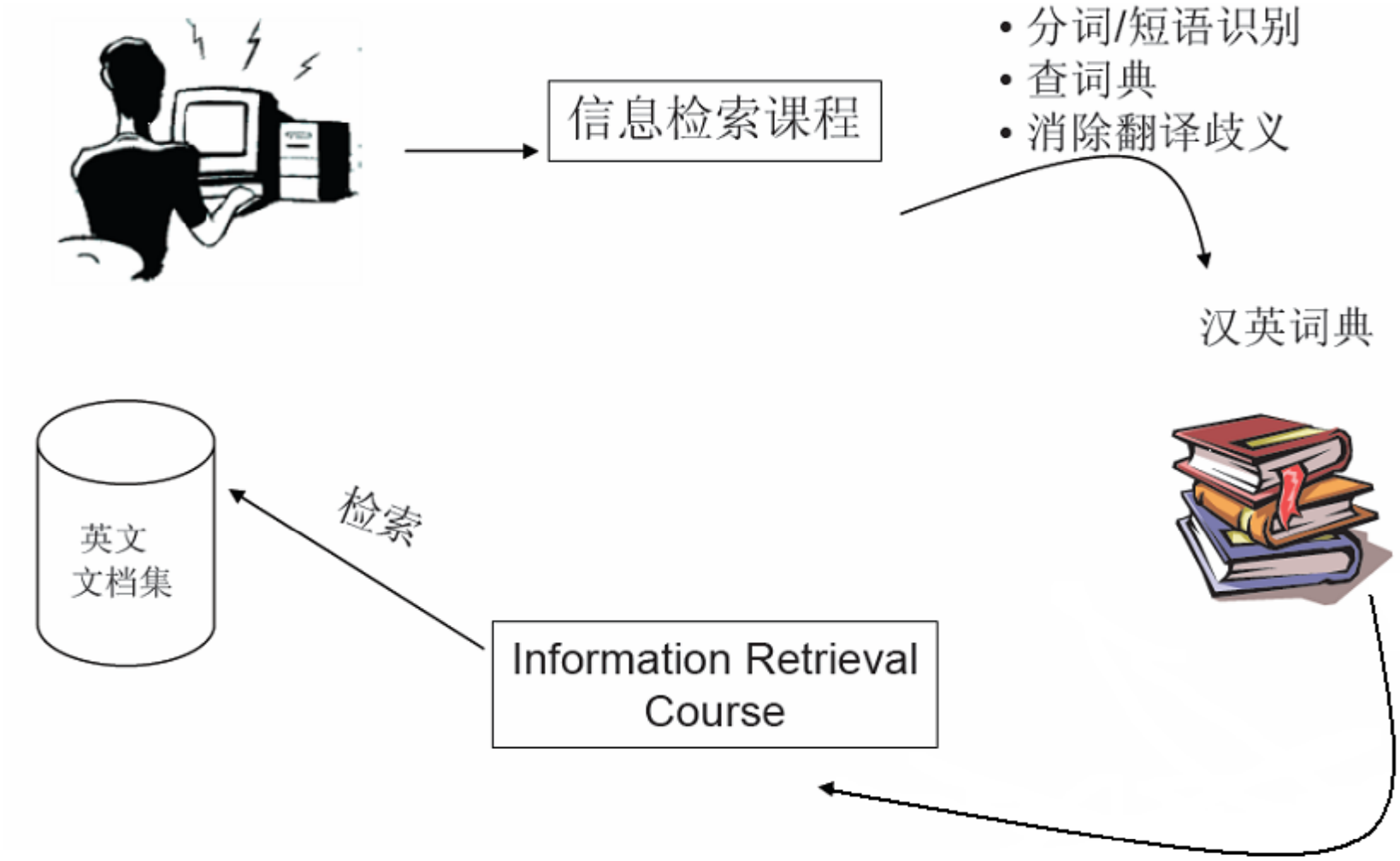
查询翻译图示


Chinese Document Collection
Chinese documents
Retrieval Engine
Translation System
Results
Chinese queries
select
examine
English queries

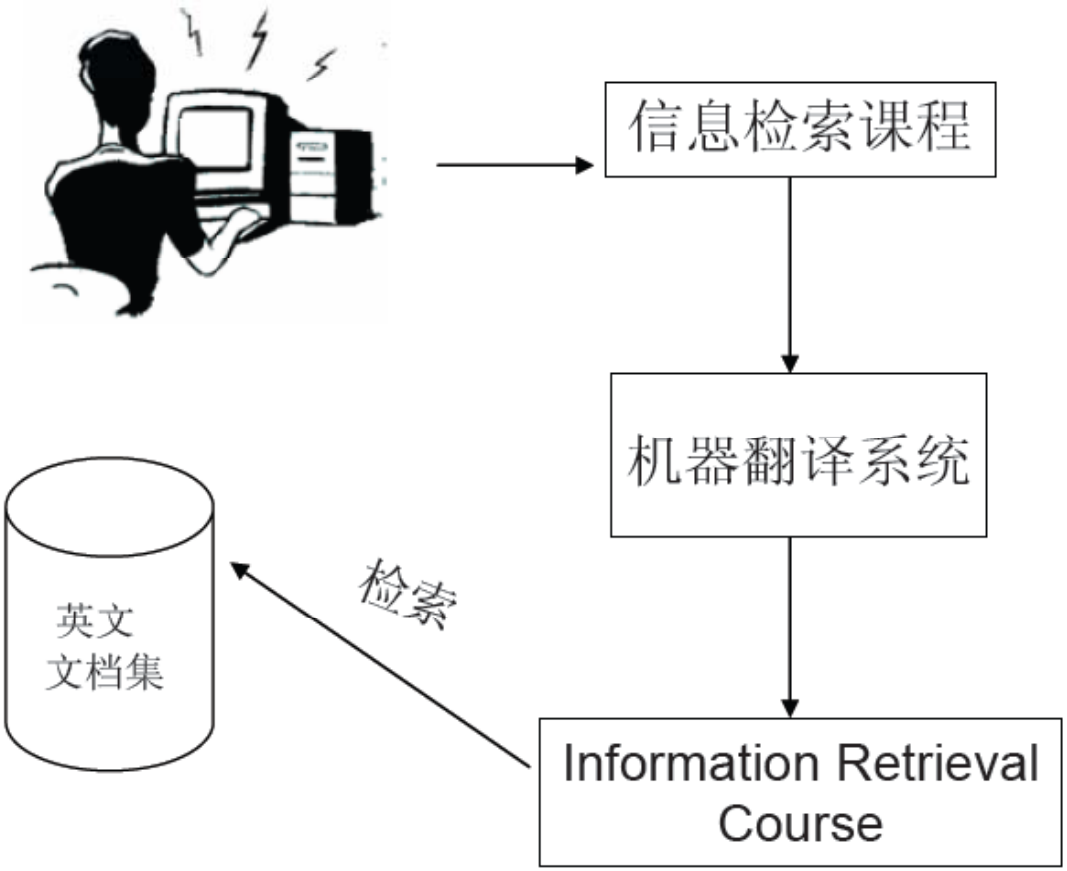
如何翻译？

- 基于词典的方法：通过查词典（双语词典、同义词词典、统计词典等），将源语言的**Term**变成目标语言的**Term**
- 基于机器翻译工具的方法：通过机器翻译工具，将源语言翻译成目标语言
- 基于并行语料库的方法：对于一个查询，先在一个并行语料库中搜索，利用并行语料之间的对齐关系，将源语言搜索结果映射成目标语言

基于词典的方法

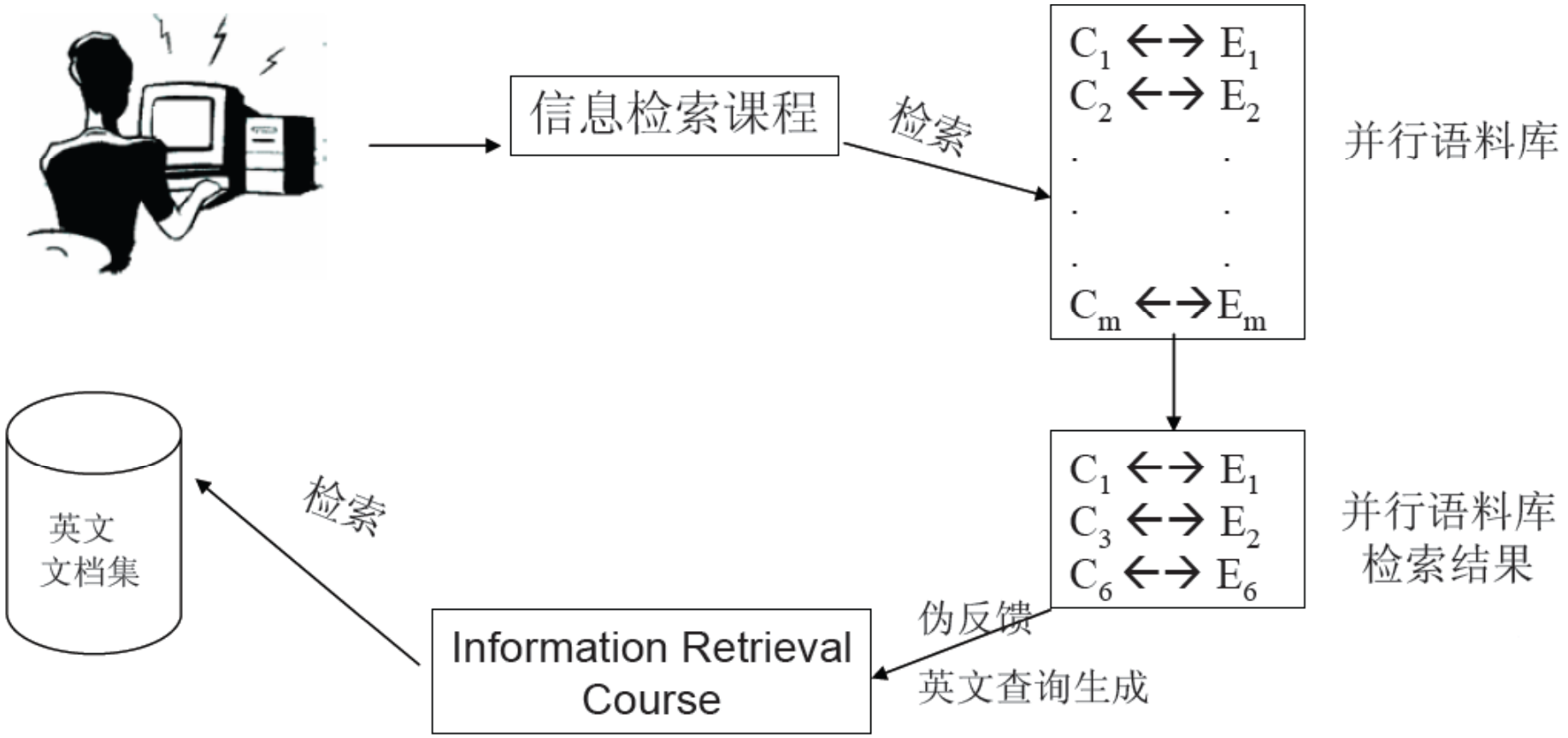

信息检索课程
• 分词/短语识别
• 查词典
• 消除翻译歧义
汉英词典
英文
文档集
检索
Information Retrieval
Course

基于机器翻译（MT）的方法



- 词法分析
- 规则方法/统计方法/EBMT

基于并行语料库的方法


信息检索课程
检索
$C_1 \leftrightarrow E_1$
$C_2 \leftrightarrow E_2$
$C_m \leftrightarrow E_m$
并行语料库
$C_1 \leftrightarrow E_1$
$C_3 \leftrightarrow E_2$
$C_6 \leftrightarrow E_6$
并行语料库
检索结果
伪反馈
英文查询生成
Information Retrieval Course
检索
英文
文档集

Google的翻译

- **Google对白宫主页（2008年2月13日）的翻译截图**

翻译的技术难点

- 一词多译：一个词或者片断有多个可能的译文
  - 一般通过上下文进行排歧
- 未定义词（**Out Of Vocabulary，OOV**）问题：词典里找不到这个词
  - 一般通过并行语料获取及对齐等方法来解决

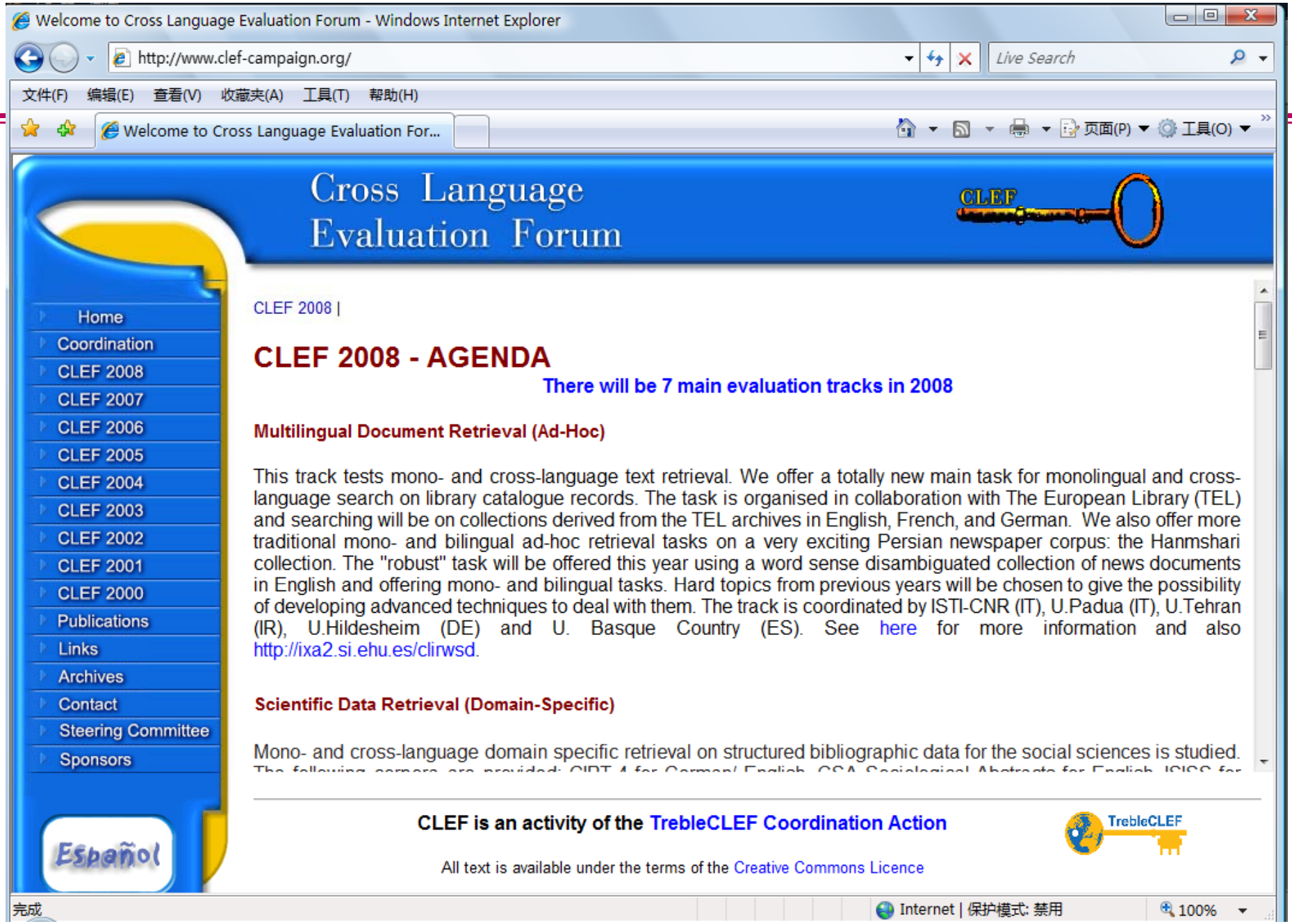
不翻译的方法

- 潜在语义分析**LSI**：将原文和对应的译文建立联系，构建训练集和对应的矩阵，再利用奇异值分解进行分析，得到双语文档的特征信息及映射关系
  - 潜语义索引方法避免机器翻译。但奇异值分解相当耗时，且需要相当数量的训练文档
- **GVSM**：两种语言中具有相同意义的词常常具有类似的共现模式，可以进行相似性比较

Carbonell, J.G., Yang, Y., Frederking, R.E., Brown, R., Geng, Y., and Lee, D. Translingual Information Retrieval: a comparative evaluation,, International Joint Conference on Artificial Intelligence, 1997

国际CLIR评测

- **TREC中CLIR评测：1997年开始设立CLIR评测，近几年取消，转入CLEF和NTCIR**
- **CLEF（Cross-Language Evaluation Forum）：主要针对欧洲语言对之间的检索评测**
  - **http://www.clef-campaign.org/**
- **NTCIR（NII-NACSIS Test Collection for IR Systems）会议：日本国立信息研究所（National Institute of Informatics）主办的信息检索测试集评测会议。主要针对英文及主要亚洲语言的检索评测**
  - **http://research.nii.ac.jp/ntcir/**


Welcome to Cross Language Evaluation Forum - Windows Internet Explorer
http://www.clef-campaign.org/
Live Search
文件(F) 编辑(E) 查看(V) 收藏夹(A) 工具(T) 帮助(H)
Welcome to Cross Language Evaluation For...
页面(P) 工具(O)
Cross Language Evaluation Forum
CLEF
Home
Coordination
CLEF 2008
CLEF 2007
CLEF 2006
CLEF 2005
CLEF 2004
CLEF 2003
CLEF 2002
CLEF 2001
CLEF 2000
Publications
Links
Archives
Contact
Steering Committee
Sponsors
Español
CLEF 2008 |
CLEF 2008 - AGENDA
There will be 7 main evaluation tracks in 2008
Multilingual Document Retrieval (Ad-Hoc)
This track tests mono- and cross-language text retrieval. We offer a totally new main task for monolingual and cross-language search on library catalogue records. The task is organised in collaboration with The European Library (TEL) and searching will be on collections derived from the TEL archives in English, French, and German. We also offer more traditional mono- and bilingual ad-hoc retrieval tasks on a very exciting Persian newspaper corpus: the Hanmshari collection. The "robust" task will be offered this year using a word sense disambiguated collection of news documents in English and offering mono- and bilingual tasks. Hard topics from previous years will be chosen to give the possibility of developing advanced techniques to deal with them. The track is coordinated by ISTI-CNR (IT), U.Padua (IT), U.Tehran (IR), U.Hildesheim (DE) and U. Basque Country (ES). See here for more information and also http://ixa2.si.ehu.es/clirwsd.
Scientific Data Retrieval (Domain-Specific)
Mono- and cross-language domain specific retrieval on structured bibliographic data for the social sciences is studied.
CLEF is an activity of the TrebleCLEF Coordination Action
TrebleCLEF
All text is available under the terms of the Creative Commons Licence
完成
Internet | 保护模式: 禁用
100%

CLEF 2008

- **Multilingual Document Retrieval (Ad-Hoc)**
- **Scientific Data Retrieval (Domain-Specific)**
- **Interactive Cross-Language Retrieval (iCLEF)**
- **Multiple Language Question Answering (QA@CLEF).**
- **Cross-Language Image Retrieval (ImageCLEF)**
- **CLEF Web Track (WebCLEF)**
- **Cross-Language Geographical Information Retrieval (GeoCLEF)**
- **Cross-Language Video Retrieval (VideoCLEF)**
- **Multilingual Information Filtering (INFILE@CLEF)**

小结：跨语言检索

- 跨语言检索的研究在欧洲语言间取得成功, **1998**年**Lisa Ballesteros**的从西班牙语查询条件到英语文档集的实验效果达到了英语单语检索的**90%**
- 亚洲语言间结果还不理想，**2004**年**NTCIR4**评测的中文查询到英文文档集的效率约为**70%** ,英语查询条件到中文文档集的效率约为**20%**
- 跨语言信息检索内容不再局限于文档检索，已经扩展到跨语言图像检索、跨语言语音检索、跨语言交互式检索、跨语言问答系统、跨语言新话题发现和跟踪

主要内容

- 多媒体检索
- 跨语言检索
- 问题回答

什么是问答系统?

- 问答系统（**Question Answering，QA**）：给定一个问题，从大规模文档集合中返回答案的系统
  - 例子：谁获得**2006**年多哈亚运会男子体操全能冠军？杨威
- 比搜索引擎更进一步，不仅仅返回相关的文档，而且直接返回正确答案

QA的分类

- 根据文档集涉及的领域，QA可以分成：
  - 开放域QA（Open domain QA）：文档集涉及的领域非常广泛，体裁风格也不一致，是各种领域、各种风格文档的综合体。如面向整个WEB的QA
  - 受限域QA（Restricted Domain QA）：文档集只涉及某个领域或行业（比如天气预报）、或者较固定书写风格的文档集（产品FAQ、百科全书）

实现方法

- 模板匹配（**Template Matching**）方法
  - 模板：**[NP]** 是谁？孙中山是谁？美国总统是谁？
  - 一个问题提出以后，从已有的模板库中进行匹配，匹配上以后，根据模板对应的处理方法调处理过程
  - 信息抽取（**Information Extraction**）的方法
- 分析问题的类型，然后从可能存在答案的结果文档中抽取答案
  - 通过问题类型分析模块确定问题的类型，然后通过检索返回可能的文档或者段落，最后在这些文档或段落中抽取相应类型的问题答案

例子：就业信息

Subject: US-TN-SOFTWARE PROGRAMMER
Date: 17 Nov 1996 17:37:29 GMT
Organization: Reference.Com Posting Service
Message-ID: <56nigp$mrs@bilbo.reference.com>

SOFTWARE PROGRAMMER

Position available for Software Programmer experienced in generating software for PC-Based Voice Mail systems. Experienced in C Programming. Must be familiar with communicating with and controlling voice cards; preferable Dialogic, however, experience with others such as Rhetorix and Natural Microsystems is okay. Prefer 5 years or more experience with PC Based Voice Mail, but will consider as little as 2 years. Need to find a Senior level person who can come on board and pick up code with very little training. Present Operating System is DOS. May go to OS-2 or UNIX in future.

Please reply to:
Kim Anderson
AdNET
(901) 458-2888 fax
kimander@memphisonline.com

提取的Job模板

computer_science_job
id: 56nigp$mrs@bilbo.reference.com
title: SOFTWARE PROGRAMMER
salary:
company:
recruiter:
state: TN
city:
country: US
language: C
platform: PC \ DOS \ OS-2 \ UNIX
application:
area: Voice Mail
req_years_experience: 2
desired_years_experience: 5
req_degree:
desired_degree:
post_date: 17 Nov 1996

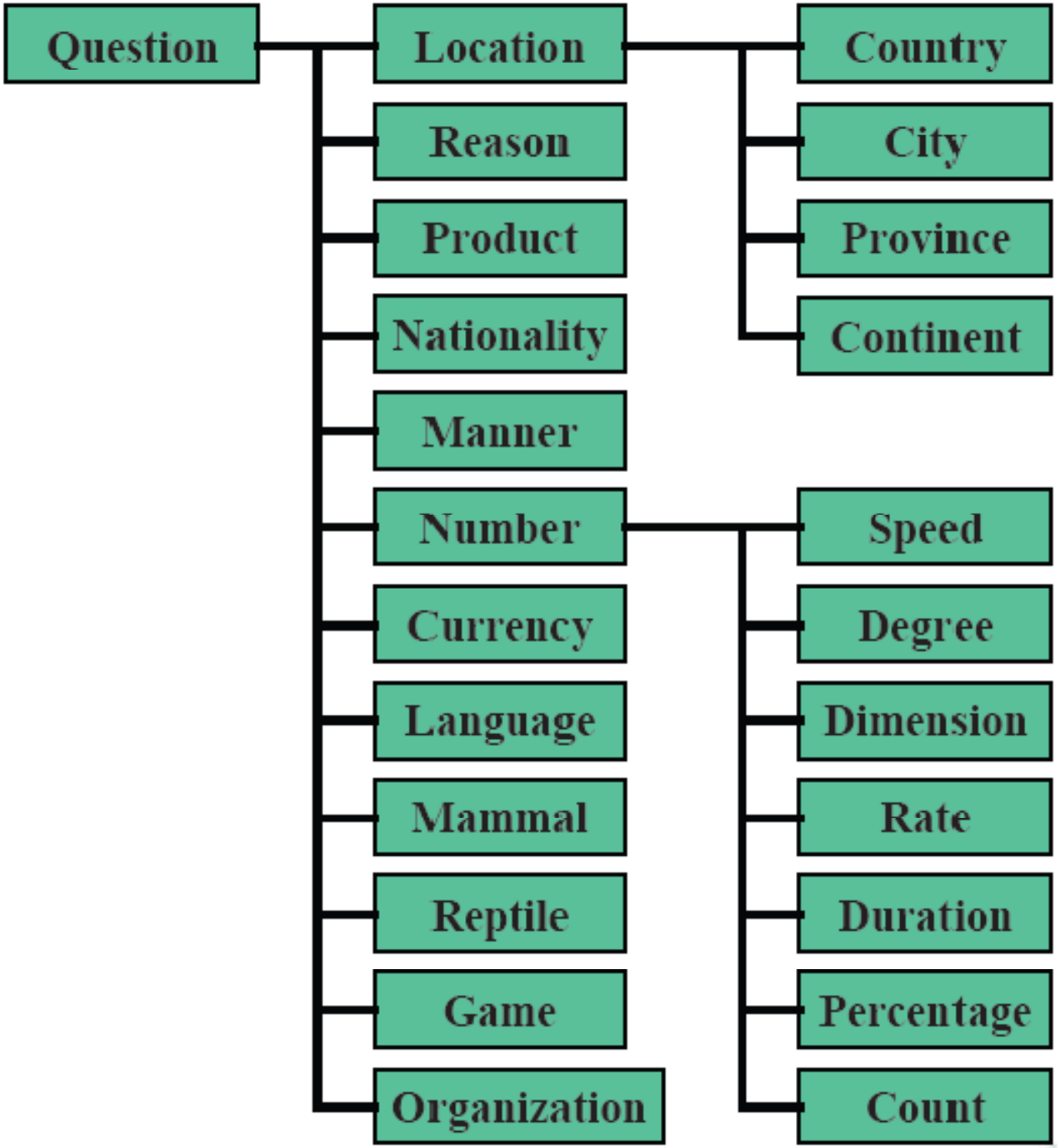
例子：Amazon书籍描述

```
….
</td></tr>
</table>
<b class="sans">The Age of Spiritual Machines : When Computers Exceed Human Intelligence</b><br>
<font face=verdana,arial,helvetica size=-1>
by <a href="/exec/obidos/search-handle-url/index=books&field-author=
            Kurzweil%2C%20Ray/002-6235079-4593641">
Ray Kurzweil</a><br>
</font>
<br>
<a href="http://images.amazon.com/images/P/0140282025.01.LZZZZZZZ.jpg">
<img src="http://images.amazon.com/images/P/0140282025.01.MZZZZZZZ.gif" width=90
   height=140 align=left border=0></a>
<font face=verdana,arial,helvetica size=-1>
<span class="small">
<span class="small">
<b>List Price:</b> <span class=listprice>$14.95</span><br>
<b>Our Price: <font color=#990000>$11.96</font></b><br>
<b>You Save:</b> <font color=#990000><b>$2.99 </b>
(20%)</font><br>
</span>
<p> <br>…
```

提取的Book模板

Title: The Age of Spiritual Machines :
When Computers Exceed Human Intelligence
Author: Ray Kurzweil
List-Price: $14.95
Price: $11.96
:
:

问题类型的例子


Question
Location
Country
City
Province
Continent
Reason
Product
Nationality
Manner
Number
Speed
Degree
Dimension
Rate
Duration
Percentage
Count
Currency
Language
Mammal
Reptile
Game
Organization

典型的TREC问题

1. Who is the author of the book, "The Iron Lady: A Biography of Margaret Thatcher"?
2. What was the monetary value of the Nobel Peace Prize in 1989?
3. What does the Peugeot company manufacture?
4. How much did Mercury spend on advertising in 1993?
5. What is the name of the managing director of Apricot Computer?
6. Why did David Koresh ask the FBI for a word processor?
7. What debts did Qintex group leave?
8. What is the name of the rare neurological disease with symptoms such as: involuntary movements (tics), swearing, and incoherent vocalizations (grunts, shouts, etc.)?

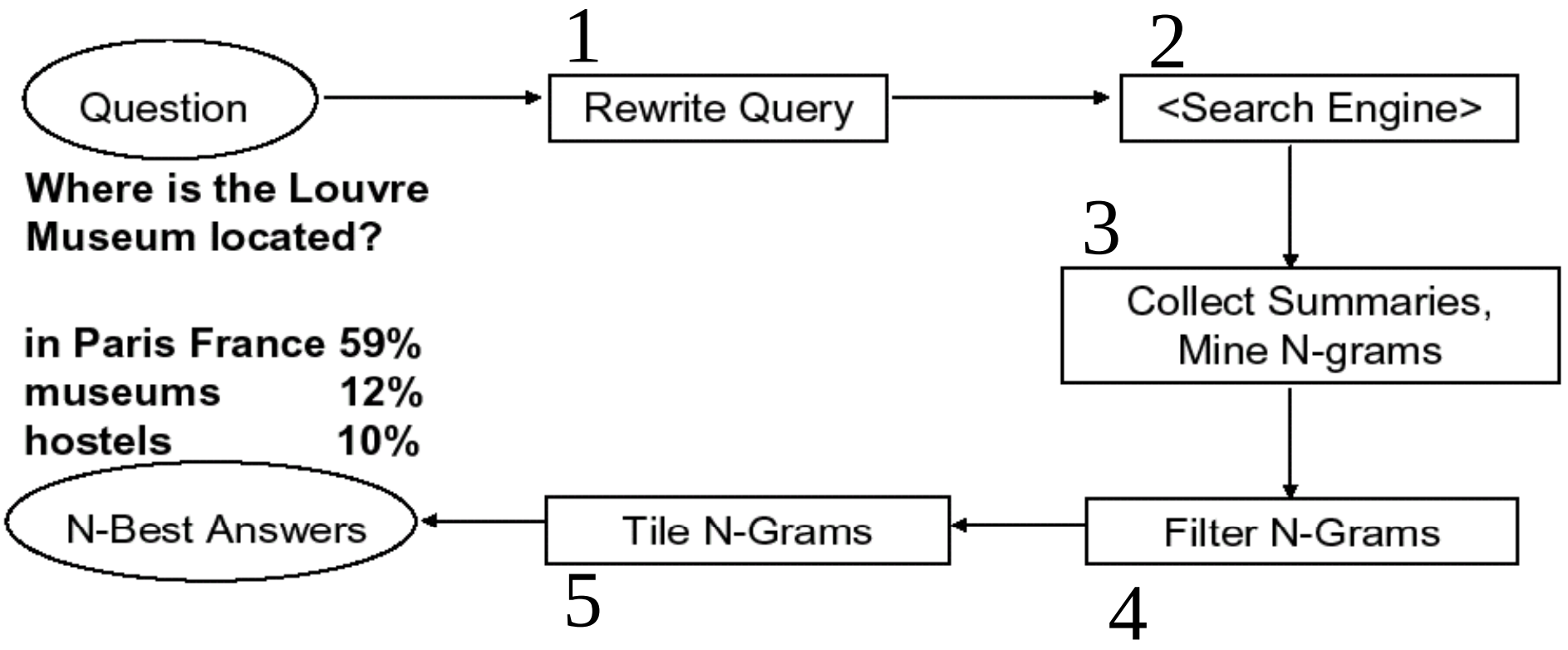
问题类型的判定

- 人工规则：人工总结出一些判定规则
  - 如：**who**　　找人
- 机器学习的方法：建立训练语料，通过统计学习的方法学习到统计规则

答案的抽取

- 以事实型问题（**factoid question**）为例
- 命名实体（**Name Entity**）的识别：人名、地名、机构名等等命名实体的识别
- 命名实体的评分：为命名实体打分，找出最可能的命名实体

例子：AskMSR


Question
1
Rewrite Query
2
<Search Engine>
Where is the Louvre Museum located?
3
Collect Summaries, Mine N-grams
in Paris France 59%
museums 12%
hostels 10%
N-Best Answers
Tile N-Grams
Filter N-Grams
5
4

步骤1：重写查询

- 思路：用户的问题通常与包含答案的句子在句法上很相近（**syntactically quite**）

  - **Where is the Louvre Museum located?**
  - **The Louvre Museum is located in *Paris***

  - **Who created the character of Scrooge?**
  - ***Charles Dickens* created the character of Scrooge.**

查询重写

- 将问题分成七个类别
  - Who **is/was/are/were…?**
  - When **is/did/will/are/were …?**
  - Where **is/are/were …?**

(1) 面向类别的转换规则

例： **“For Where questions, move ‘is’ to all possible locations”**

“Where is the Louvre Museum located”

→ **“is the Louvre Museum located”**

→ **“the is Louvre Museum located”**

→ **“the Louvre is Museum located”**

→ **“the Louvre Museum is located”**

→ **“the Louvre Museum located is”**

Nonsense, but who cares? It’s only a few more queries to Google.

(2) 期望的答案的数据类型 **“Datatype”** (如**Date, Person, Location, …**)

When was the French Revolution? → DATE

步骤2：将查询提交给搜索引擎

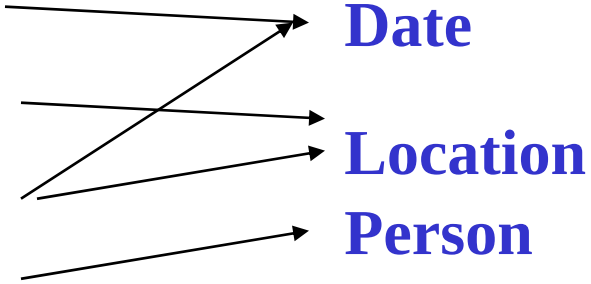
- 将所有的重写发给搜索引擎
- 提取前N个结果（100?）
- 为了加快速度，只获取网页摘要（snippets），而不是实际文档的全文

步骤3：挖掘N-Grams

- 枚举在提取的摘要中出现的所有N字词（N-grams），N=1,2,3
  - 如： “Web Question Answering: Is More Always Better”
    - 双字词（Bigrams）: Web Question, Question Answering, Answering Is, Is More, More Always, Always Better
- 用出现次数计算N字词的权重
- 例子： “Who created the character of Scrooge?”
  - Dickens - 117
  - Christmas Carol - 78
  - Charles Dickens - 75
  - Disney - 72
  - Carl Banks - 54
  - A Christmas - 41
  - Christmas Carol - 45
  - Uncle - 31

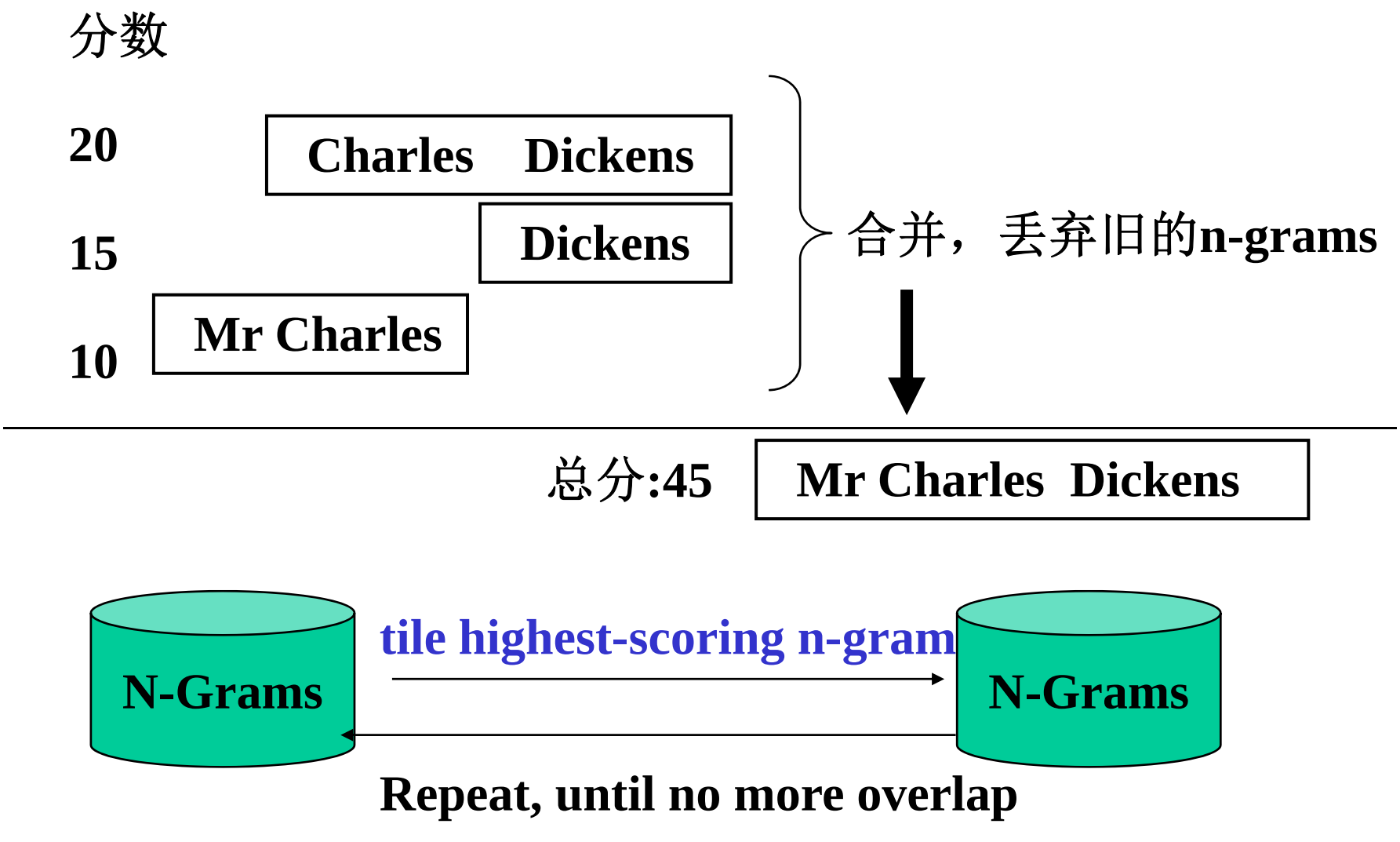
步骤4：过滤N-Grams

- 每个问题类型有一个或多个“数据类型过滤器”（**data-type filters**）=正则表达式
  - **When…**
  - **Where…**
  - **What …**
  - **Who …**



- 增大匹配正则表达式的**n-grams**
- 减低不匹配正则表达式的**n-grams**

步骤5：拼接答案


分数
20
15
10
Charles Dickens
Dickens
Mr Charles
合并，丢弃旧的n-grams
总分:45
Mr Charles Dickens
N-Grams
tile highest-scoring n-gram
N-Grams
Repeat, until no more overlap

QA的评测

- 1999年开始，QA加入到TREC评测中，一直延续到2007年
- 目前基于事实型的问题的正确率最好可以达到70%（2007年TREC QA评测），但是其他类型的问题（如描述性问题）要解决还为时过早

课程小结

- 很活跃的研究领域
- 许多未解决的问题
- 富有挑战性
- 具有广阔的应用前景

推荐阅读和网站

- 《网络信息检索》
  - 第一章：绪论
  - 第九章：中文和跨语言检索
  - 第十章：多媒体信息检索
  - 第十一章：信息分类与聚类
  - 第十二章：信息提取和QA系统

- 北京大学2008季研究生课程“**Web Based Information Architectures**”
  - **http://net.pku.edu.cn/~wbia/**

参考文献

- **Carbonell, J. G., Yang, Y.,. Frederking, R. E., Brown, R., Geng, Y., and Lee, D. Translingual Information Retrieval: a comparative evaluation. Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'97), 1997.**
- **J. Jeon, V. Lavrenko and R. Manmatha, Automatic Image Annotation and Retrieval Using Cross-Media Relevance Models, SIGIR 2003.**
- **A. Hauptmann and M. Christel. (2004) Successful Approaches in the TREC Video Retrieval Evaluations. Proceedings of ACM Multimedia 2004**
- **Hoa Trang Dang,Diane Kelly,and Jimmy Lin， Overview of the TREC 2007 Question Answering Track, In Proceedings of the Sixteenth Text REtrieval Conference, 2008**