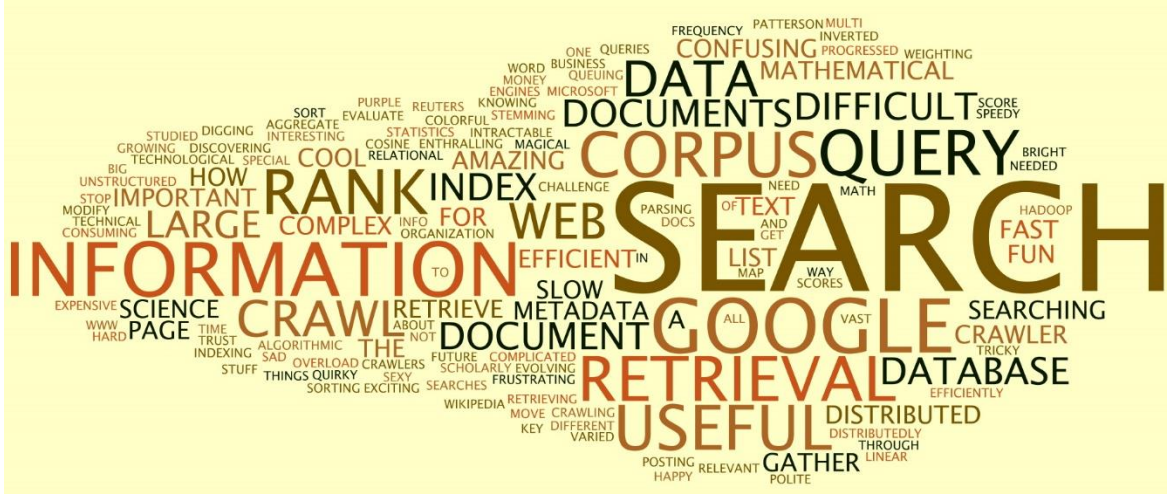

SEARCH
INFORMATION
CORPUS
QUERY
RANK
WEB
INDEX
CRAWL
GOOGLE
RETRIEVAL
USEFUL
DATA
DOCUMENTS
DOCUMENT
DIFFICULT
MATHEMATICAL
CONFUSING
IMPORTANT
LARGE
COMPLEX
COOL
AMAZING
HOW
FOR
THE
TO
IN
A
OF
TEXT
LIST
FAST
FUN
SLOW
METADATA
RETRIEVE
SCIENCE
PAGE
DATABASE
DISTRIBUTED
SEARCHING
CRAWLER
GATHER
FREQUENCY
PATTERSON
MULTI
INVERTED
PROGRESSED
WEIGHTING
ONE
QUERIES
WORD
BUSINESS
QUEUING
MONEY
ENGINES
MICROSOFT
KNOWING
PURPLE
REUTERS
SORT
EVALUATE
STEMMING
COLORFUL
SCORE
SPEEDY
STUDIED
DIGGING
AGGREGATE
INTERESTING
STATISTICS
INTRACTABLE
GROWING
DISCOVERING
COSINE
ENTHRALLING
MAGICAL
RELATIONAL
TECHNOLOGICAL
SPECIAL
BRIGHT
NEEDED
BIG
UNSTRUCTURED
STOP
MODIFY
TECHNICAL
CONSUMING
CHALLENGE
NEED
MATH
PARSING
DOCS
INFO
ORGANIZATION
AND
GET
HADOOP
MAP
WAY
SCORES
EFFICIENT
EXPENSIVE
WWW
HARD
TIME
TRUST
INDEXING
ALGORITHMIC
SAD
ABOUT
NOT
ALL
VAST
TRICKY
OVERLOAD
CRAWLERS
FUTURE
COMPLICATED
SCHOLARLY
EVOLVING
STUFF
THINGS QUIRKY
SEXY
SORTING
EXCITING
SEARCHES
FRUSTRATING
EFFICIENTLY
WIKIPEDIA
RETRIEVING
MOVE
CRAWLING
KEY
DIFFERENT
VARIED
DISTRIBUTEDLY
THROUGH
LINEAR
POSTING
HAPPY
RELEVANT
POLITE

信息检索

讲课内容

- 目前的信息检索(搜索)技术实质上是融合了文本及多媒体检索、数据挖掘、机器学习和自然语言处理的综合学科。
- 因此本课程的内容包括：
  - 信息检索的基本知识
    - 布尔检索
    - 倒排索引
    - 向量空间模型
    - 文档评分
    - 检索评价
    - 检索模型
  - 简单的自然语言处理
    - 语言模型
  - 信息检索/知识发现/机器学习/数据挖掘中的经典算法
    - 分类
    - 聚类

参考书籍

- Christopher D. Manning, Prabhakar Raghavan & Hinrich Schütze, **Introduction to Information Retrieval**, Cambridge University Press 2008 Electronic version (draft) can be downloaded from

http://www-csli.stanford.edu/~hinrich/information-retrieval-book.html

- 王斌译，**信息检索导论 修订版**，人民邮电出版社出版，2019年7月。

考核方式

- 平时作业 + 期末考试
  - 期末考试 60%
  - 平时40%
    - 不定期考勤 5%
    - 实验：开发一个“轻型搜索引擎”(35%)
    - 只要能想到就能做到，发挥你的奇思妙想，下一个就“微软、Google”诞生在你们中间！！！

Why 搜索技术？

信息过载—数据爆炸性的增长，而人的处理能力有限

- Internet上网站总数

1995	2004	2006	2010	2011	2012	2014
1.8万	500万	1亿	2.55亿	5.55亿	6.34	10亿

- 网页

  Web页面数量近千亿，数据总量约10万亿GB (均为估计数据)

- 视频

48	YouTube上每分钟上传视频的小时数
250万	被上传到YouTube新闻相关视频的小时数
883亿	Google sites（incl. YouTube）每月观看的视频数
2000万	Facebook 上每月上传的视频数
15648303	谷歌视频网站的独立访客访问数量，视频领域排名第一

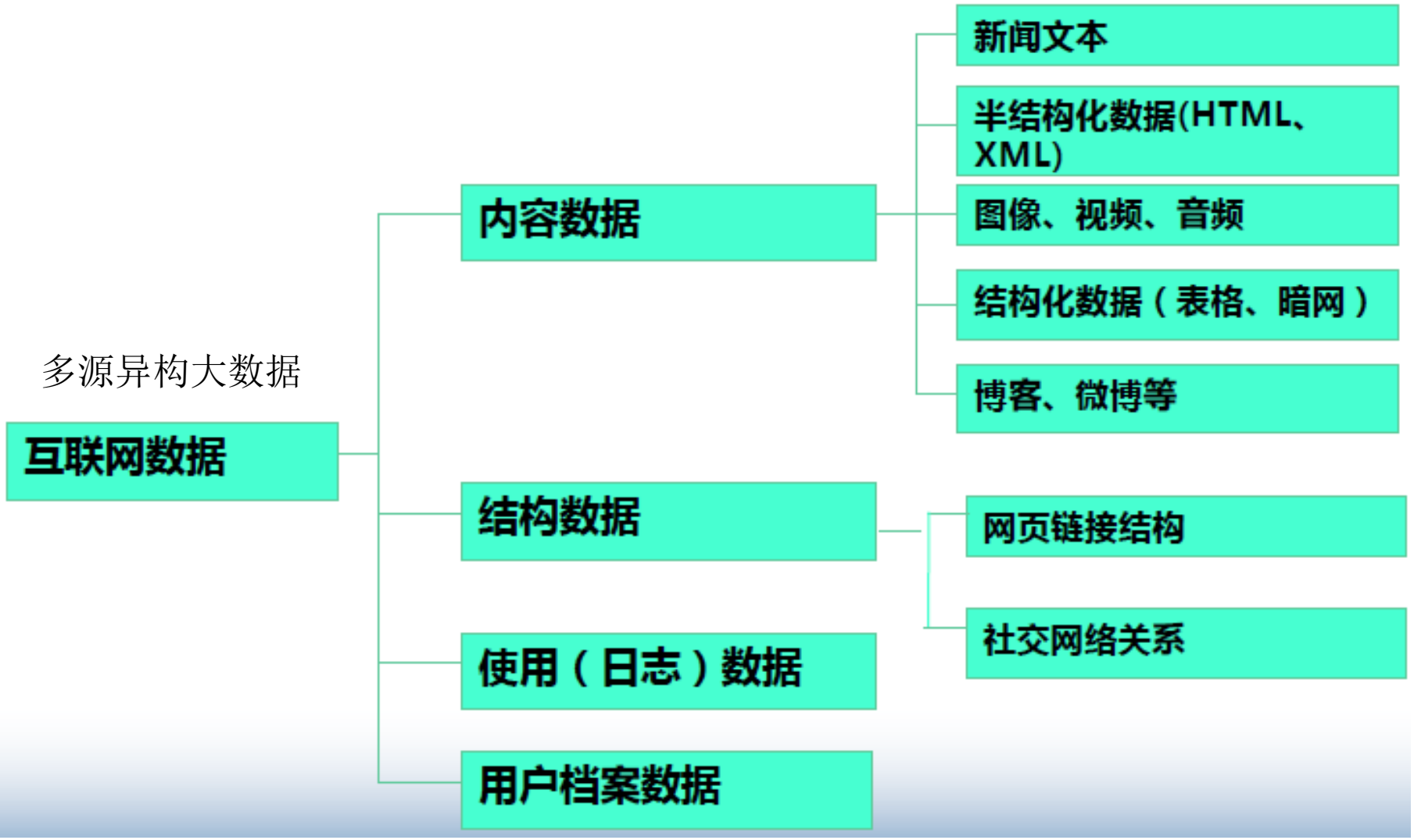
- 搜索

2012	**谷歌的总搜索次数1.2万亿**
2016	**百度每天收到搜索请求60亿次，谷歌大约有55亿次**

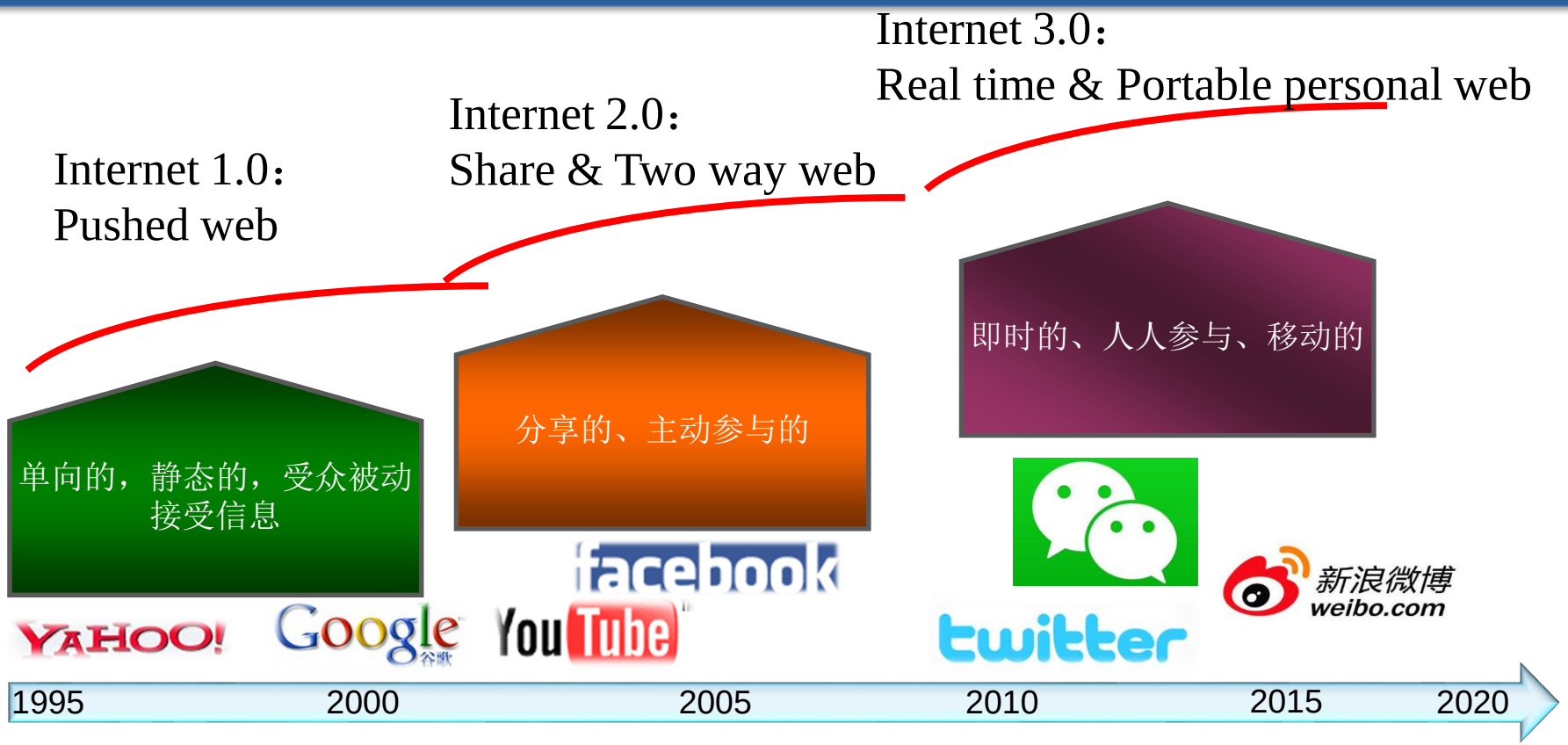
- 全球智能移动终端互联网用户数量已达到30亿(2016),13亿(2012年底)

67亿	手机订阅的数量
31%	美国互联网用户中，有31%都在使用平板电脑或电子阅读器
1.3EB	2012年每月的全球移动数据流量约为1.3EB(1EB=2^10PB)
59%	全球移动数据流量59%的份额来自于视频
500MB	智能手机平均每月消耗的数据流量为500MB
504kbps	全球所有手机平均移动网络的连接速度为504kbps
1820kbps	全球智能手机平均移动网络的连接速度为1820kbps

Web数据类型众多


多源异构大数据
互联网数据
内容数据
新闻文本
半结构化数据(HTML、XML)
图像、视频、音频
结构化数据（表格、暗网）
博客、微博等
结构数据
网页链接结构
社交网络关系
使用（日志）数据
用户档案数据

国际互联网发展历程



●Facebook的到来意味着社交网络以及Internet2.0发展的成熟，而Twitter以及新浪微博的到来，预示着全球互联网进入Internet3.0时代。

●Internet3.0，意味着：受众消失，微众来临；成为生产消费者之前，个体首先在自组织网络中实时连接起来；微众不再是调查统计意义上抽象的某类群体或者大众分众窄众，微分到个体；而微众联动又足以脱离媒体影响，自主完成交流分享、消费决策过程。这意味着即时的、人人参与的以及移动的智能网络到来。

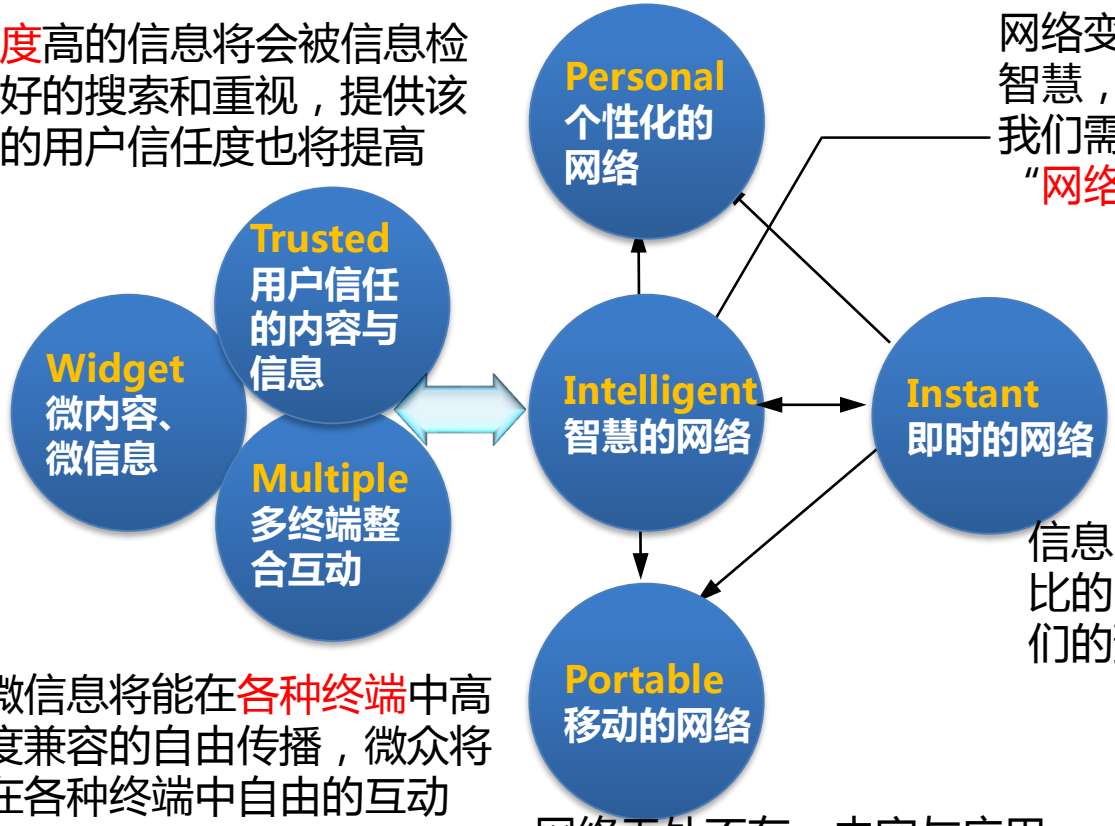
Internet3.0特征图谱

网络应用与服务、网络内容与信息将提供高度的个性化服务，如个性化的web os

可信度高的信息将会被信息检索更好的搜索和重视，提供该信息的用户信任度也将提高

网络变得更加聪明，更加智慧，甚至比我们更知道我们需要什么。
“网络比你更懂得你自己”

微众首先在自组织网络中实时连接起来，在此过程中微内容与信息能够得到自由整合与聚合

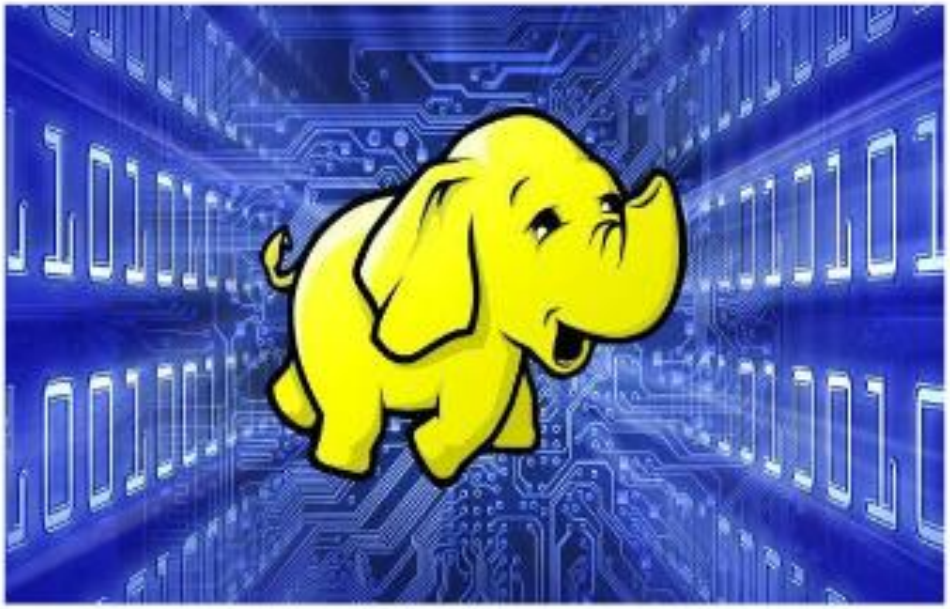


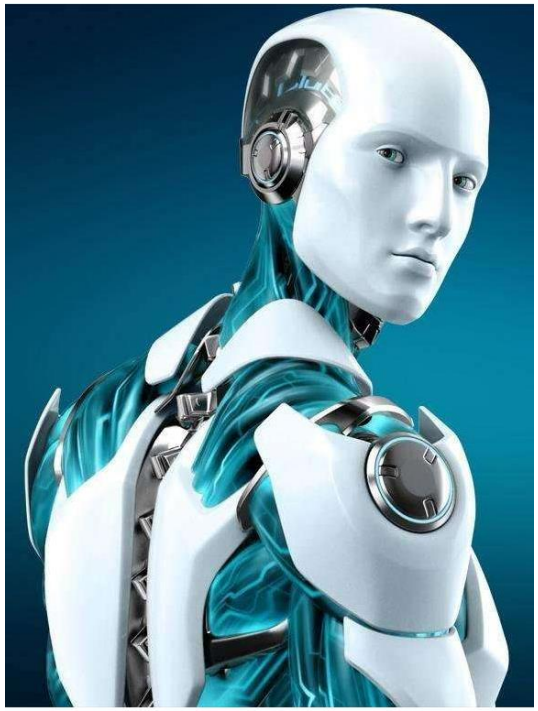
信息的传播速度变的无比的即时，远远超越我们的预期

微信息将能在各种终端中高度兼容的自由传播，微众将在各种终端中自由的互动

网络无处不在，内容与应用变得更适于移动时代的使用

- Web 3.0
  - 大众从消费者变成生产者
  - 社会网络兴起
- **人类跑步进入大数据(Big Data)时代！！！**
- **一转身又跑步进入人工智能(AI)时代！！！**



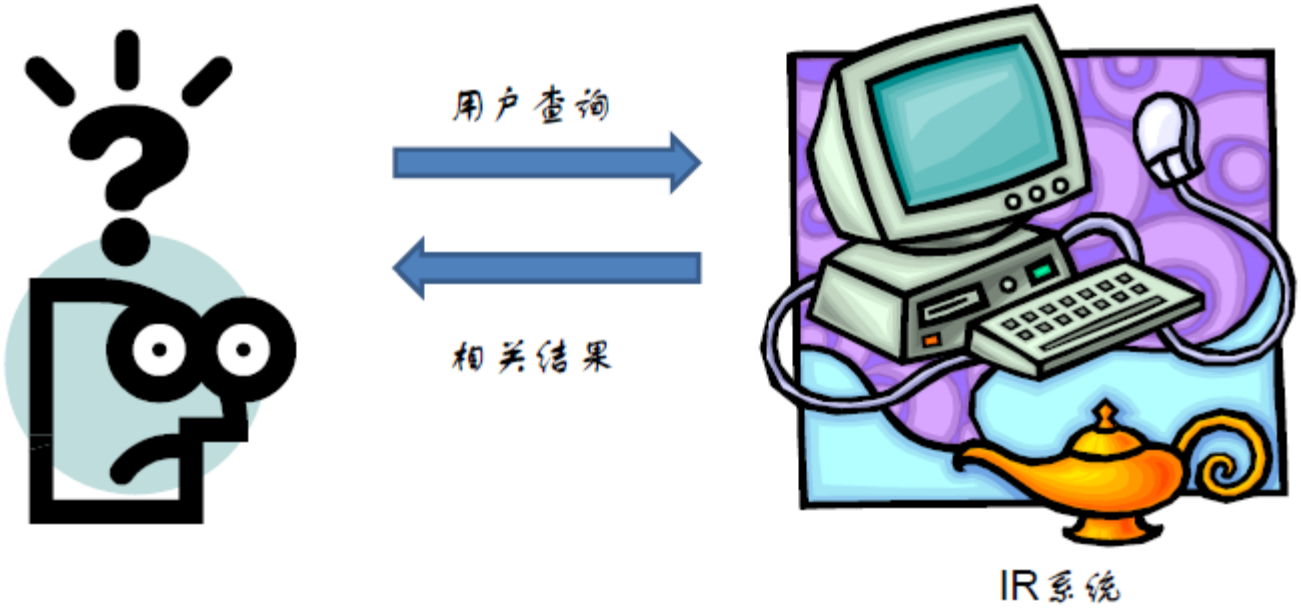


大数据的4V特性

体量Volume	*非结构化数据*的超大规模和增长 总数据量的80~90% 比结构化数据增长快10倍到50倍 是传统数据仓库的10倍到50倍
多样性Variety	大数据的异构和多样性 很多不同形式（文本、图像、视频、机器数据） 无模式或者模式不明显 不连贯的语法或句义
价值密度Value	大量的不相关信息 对未来趋势与模式的可预测分析 深度复杂分析（机器学习、人工智能Vs传统商务智能(咨询、报告等）
速度Velocity	*实时分析*而非批量式分析 数据输入、处理与丢弃 立竿见影而非事后见效

信息过载，如何解决？

- **搜索技术应运而生！**
- 从大规模非结构化数据(通常是文本)的集合(通常保存在计算机上)中找出满足用户信息需求的资料(通常是文档)的过程
- 作为一门学科，是研究信息的获取(acquisition)、表示(representation)、存储(storage)、组织(organization)和访问(access)的一门学问。



搜索系统 甚至可以是试衣间！

本课程的意义

市场发展的需求

- 用户需要信息检索技术：互联网的信息量太大，寻找信息耗时耗力
- 公司需要信息检索技术：信息检索技术可以挣大钱，搜索引擎改变了人们获取信息的方式，Google、Microsoft、Baidu，还有一些公司如~~Tecent~~、Sina、Sohu、360、神马都加入到这个搜索技术的竞争
- 人才的竞争：搜索人才人数出现缺口，非常抢手，待遇如日中天
- 是不是泡沫?：2000年左右出现的网络泡沫和现在的互联网有什么不同，搜索引擎在其中占什么位置(2010)?

课程特点

- 不是教如何使用信息检索工具(学校有专门的课程)，而是了解信息检索工具背后的基本原理和技术，并且能够进行深层的研究或开发相关的应用。

- **基本原理 + 广泛实践**

搜索技术的历史

历史分段

- 计算机出现以前
- 计算机出现以后
- Internet出现以后

SIGIR2020 - The **43rd** International ACM SIGIR Conference on Research and Development in Information Retrieval

- **计算机出现以前**
  - 约4000年前，人类就开始有目的地组织信息，一个典型的例子就是图书中的目录。
  - 随后，逐渐出现**索引**的概念，即从一些词和概念指向相关信息或者文档的指针。
  - 计算机问世以前，人们主要通过手工方式来建立索引。

  - 例子：词典(拼音检字、部首笔画检字等)

- **1948**
  - C. N. Mooers在其MIT的硕士论文中第一次创造了“**Information Retrieval**”这个术语。
- **1960－70年代**
  - 人们开始使用计算机为一些小规模科技和商业文献的摘要建立**文本检索系统**。
  - 产生了**布尔模型**(Boolean Model)、**向量空间模型**(Vector Space Model)和**概率检索模型**(Probabilistic Model)。康奈尔大学的Salton领导的研究小组是该领域研究的佼佼者。
  - 伦敦城市大学的Robertson及剑桥大学的Sparck Jones是概率模型的倡导者。

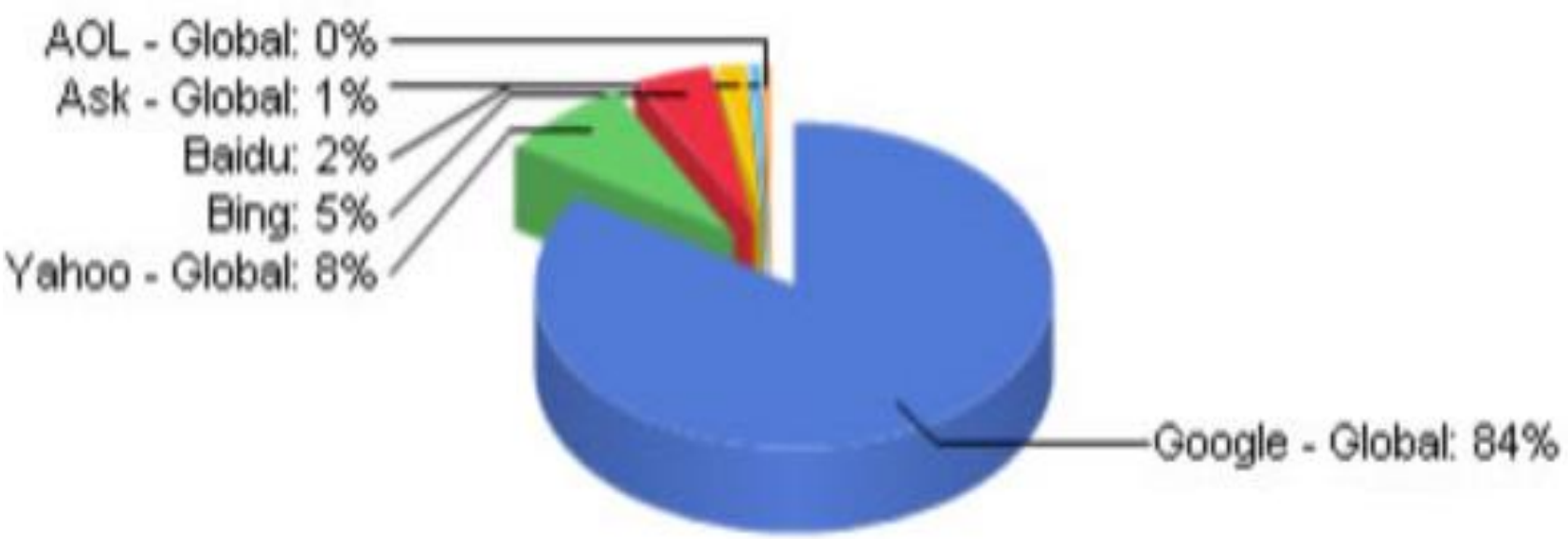
- **1980年代**
  - 出现了一些商用的较大规模数据库检索系统
  - Lexis-Nexis
  - Dialog
  - MEDLINE

- **1986**：Internet正式形成
- **1990’s**
  - 第一个网络搜索工具：1990年加拿大蒙特利尔麦吉尔(McGill)大学开发的FTP搜索工具Archie
  - 第一个Web搜索引擎：1994年美国CMU开发的Lycos
  - 1995：斯坦福大学博士生开发的**Yahoo**
  - 1998：斯坦福大学博士生开发的**Google**，提出PageRank计算公式。
  - 1998：基于语言模型的IR模型提出。

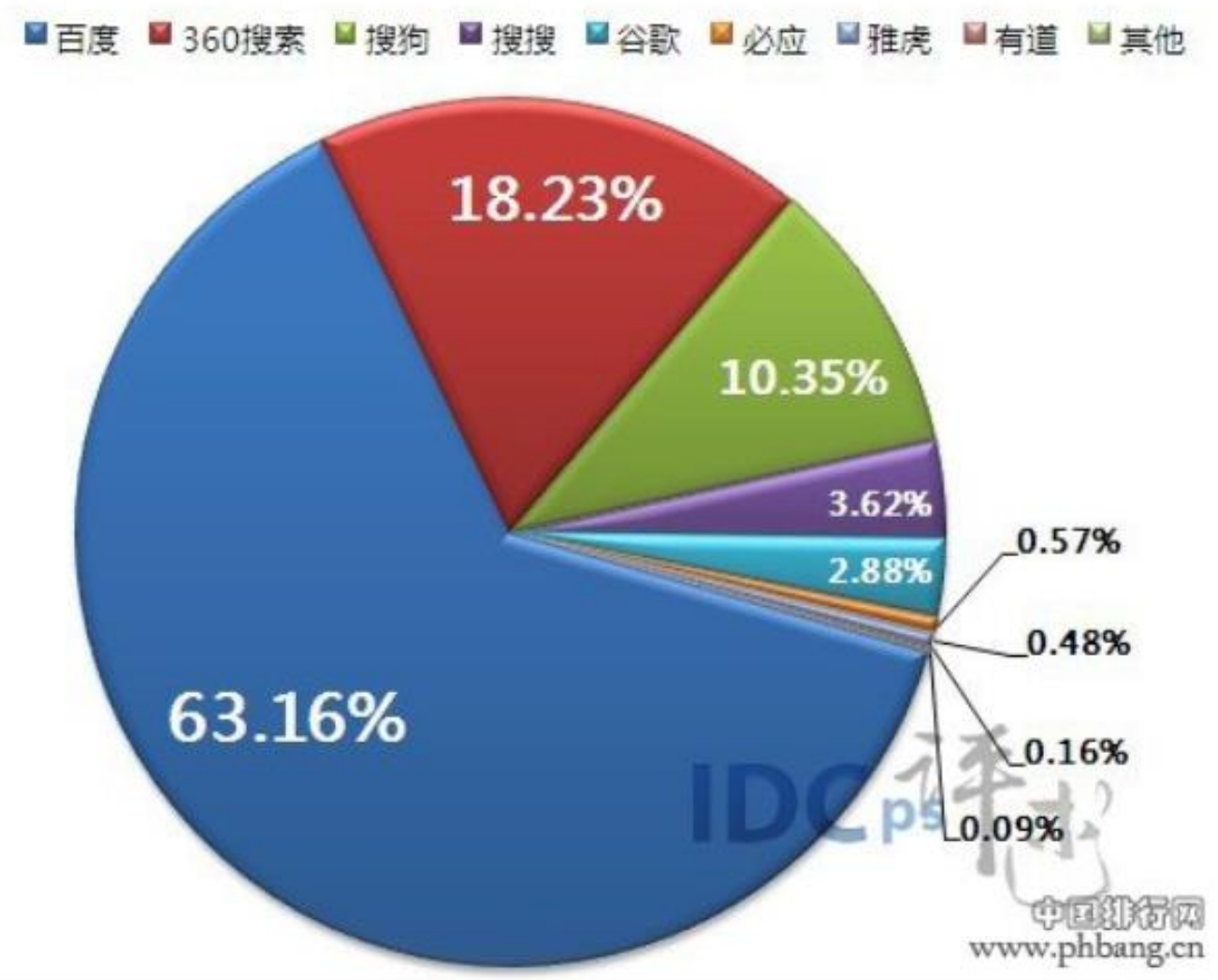
- **2000's**
  - 信息抽取
  - Whizbang
  - Fetch
  - Burning Glass
  - 问答系统
  - TREC Q/A track
  - 2000年，百度成立

- **2000以来的其它重要事件**
  - 多媒体 IR
    - Image
    - Video
    - Audio and music
  - 跨语言 IR
    - DARPA Tides
  - 文本摘要
    - DUC评测
  - 自动问答
  - 自动对话

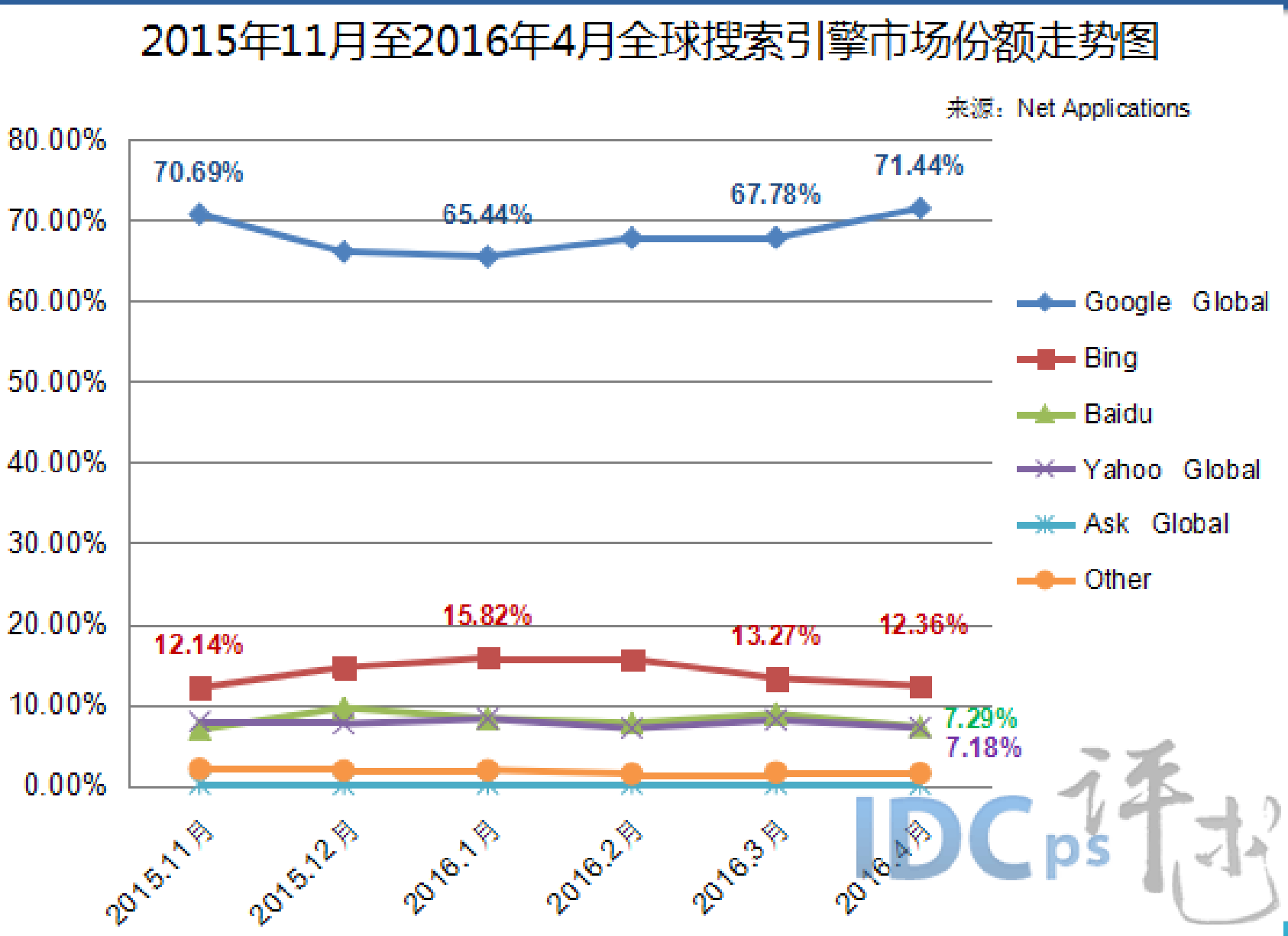
2012年12月全球搜索市场份额



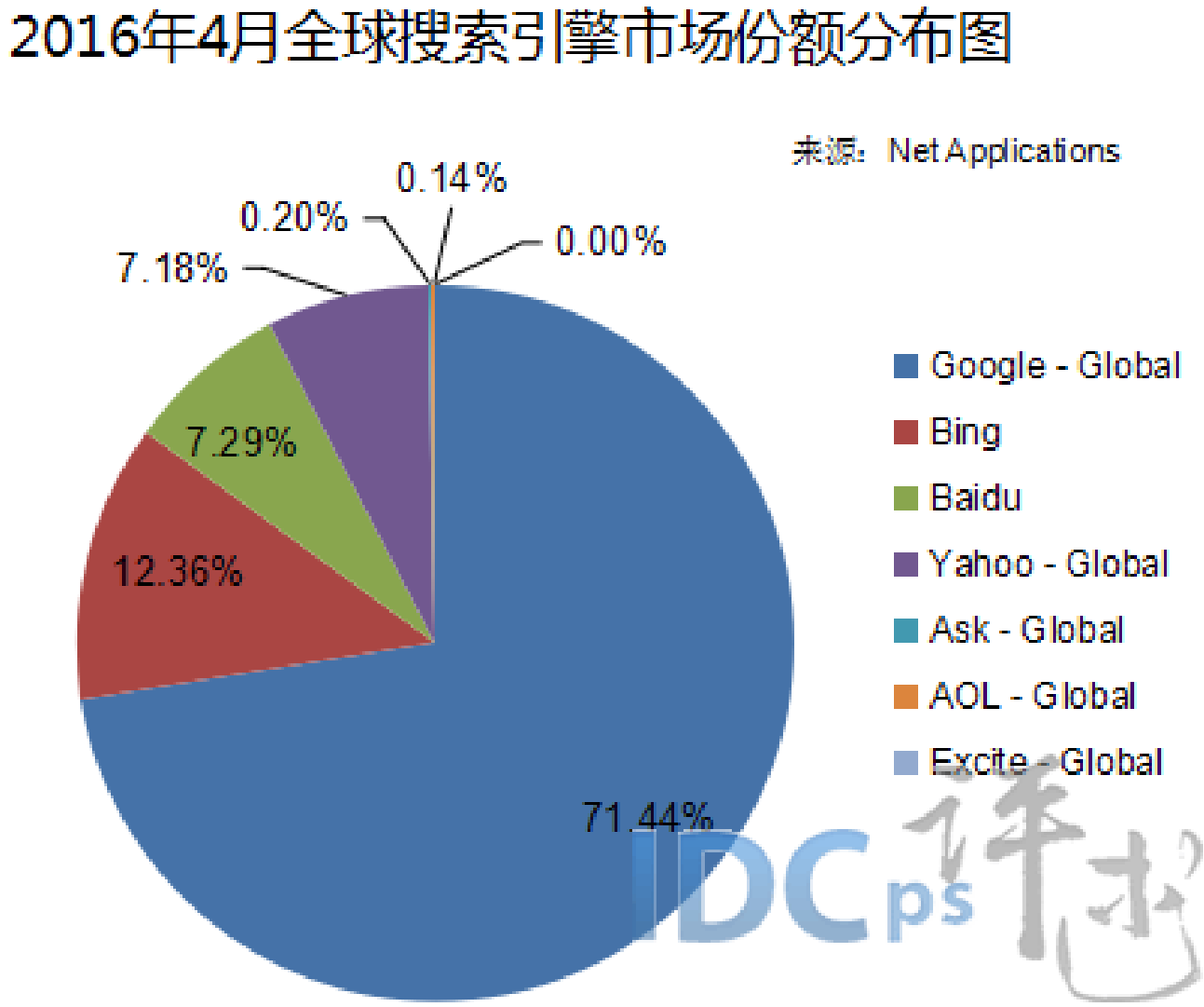
2013年8月国内搜索市场份额



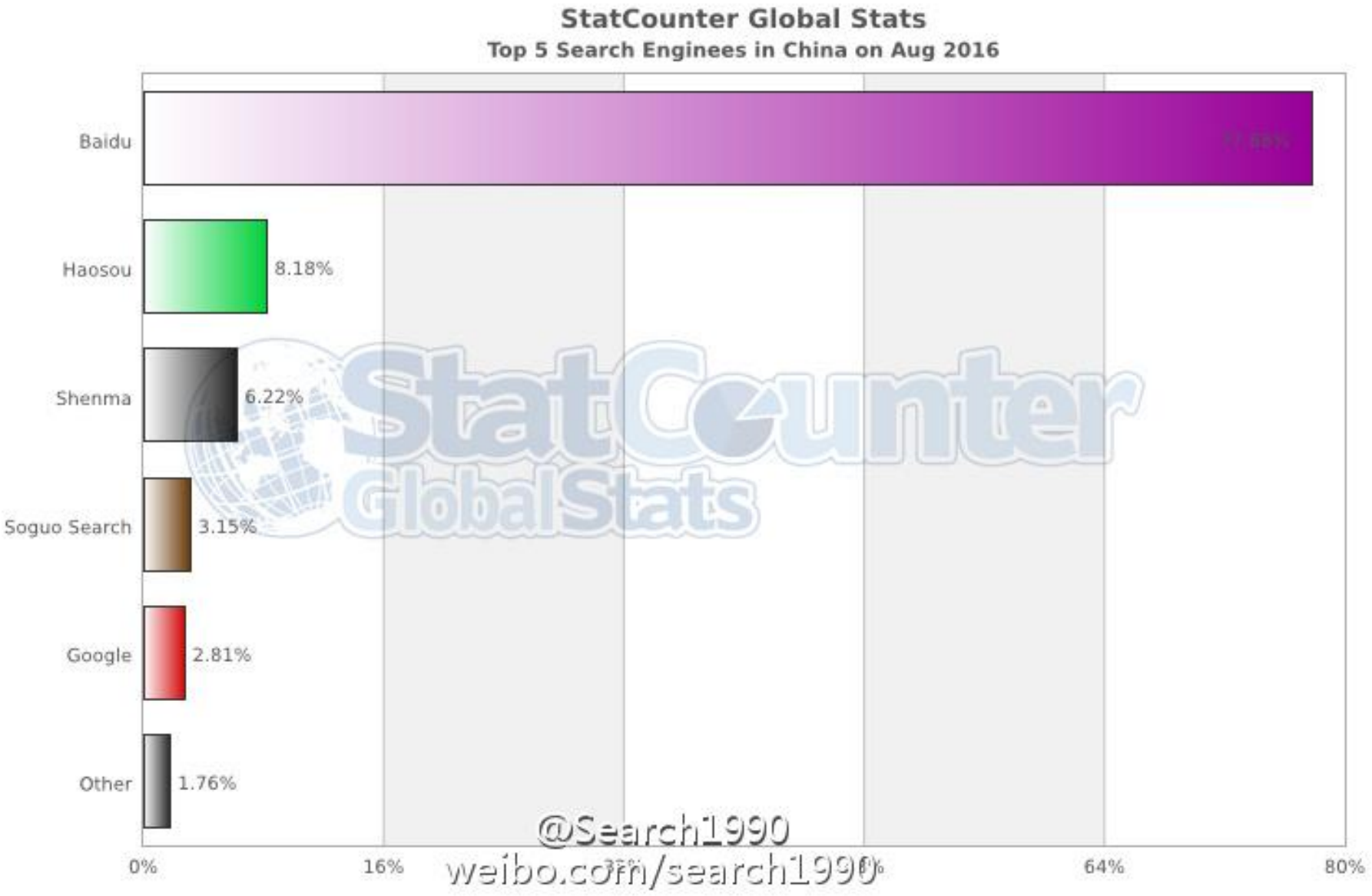
2015年11月至2016年4月全球搜索市场份额走势



2016年4月全球搜索市场份额

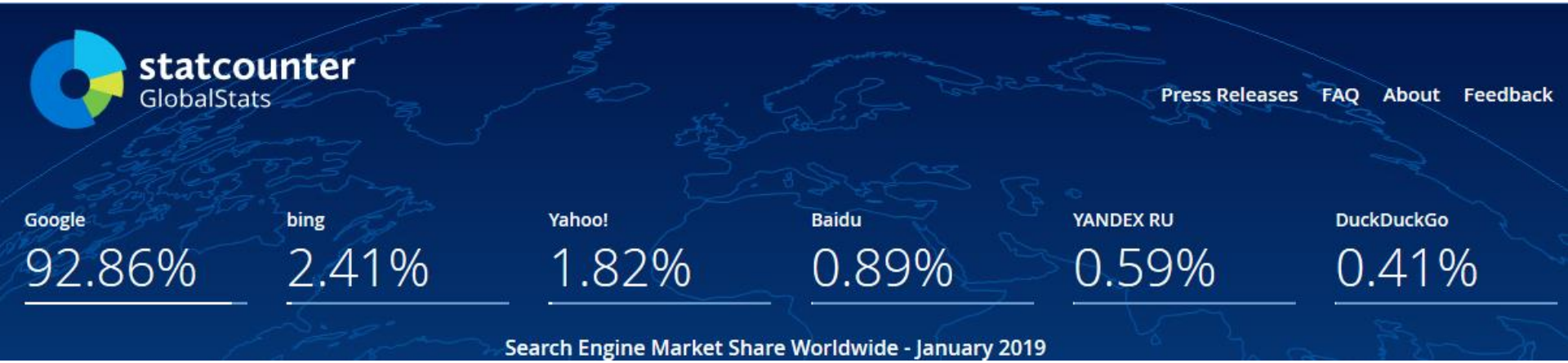


2016年8月国内搜索市场份额

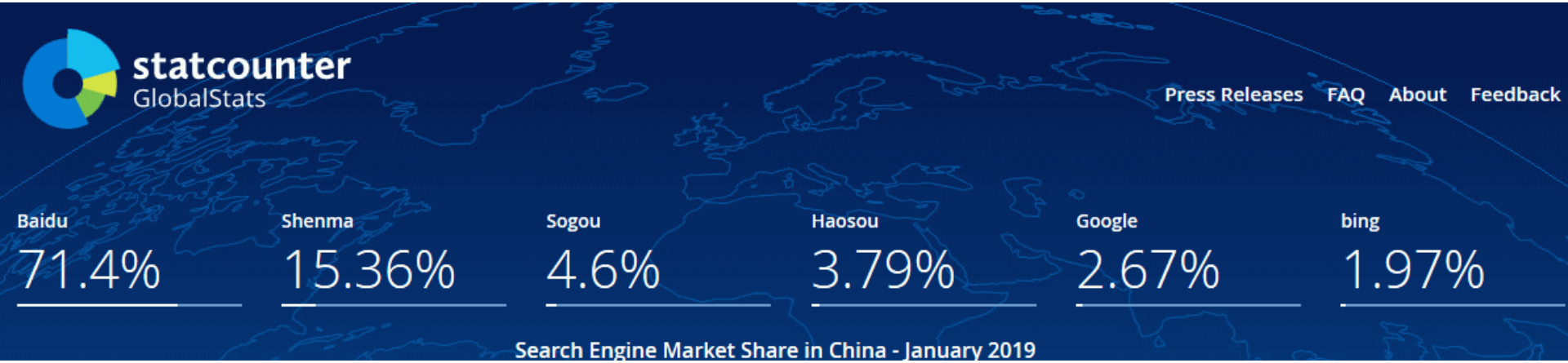


2017年1月全球搜索市场份额

2019年1月全球搜索市场份额



国内搜索市场份额



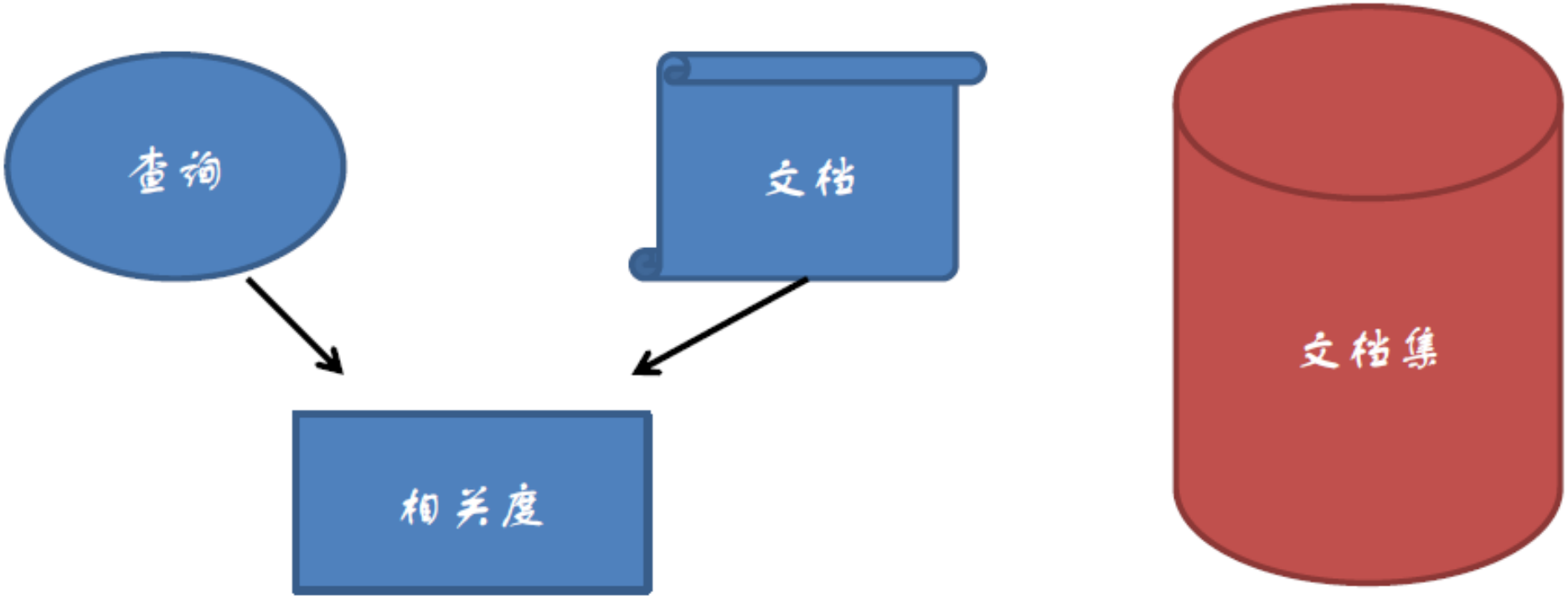
搜索的基本概念

- **用户需求(User Need，UN)**：用户需要获得的信息
  - 严格地说，UN只存在于用户的内心，但是通常用文本来描述，如查找与2016奥运会相关的新闻，有时也称为主题(Topic)
  - UN提交给检索系统时称为查询(Query)，如 2016奥运会，对同一个UN，不同人不同时候可以构造出不同的Query，比如上述需求也可表示成2016奥运会新闻，Query在IR系统中往往还有内部表示

- **文档(Document)**：检索的对象
  - 可以是文本，也可以是图像、视频、语音等多媒体文档，text retrieval / image retrieval / video retrieval / speech retrieval / multimedia retrieval
  - 可以是无格式、半格式、有格式的
- **文档集合(Collection)**：所有待检索的文档构成的集合
  - 也称为Repository，Corpus，Dataset

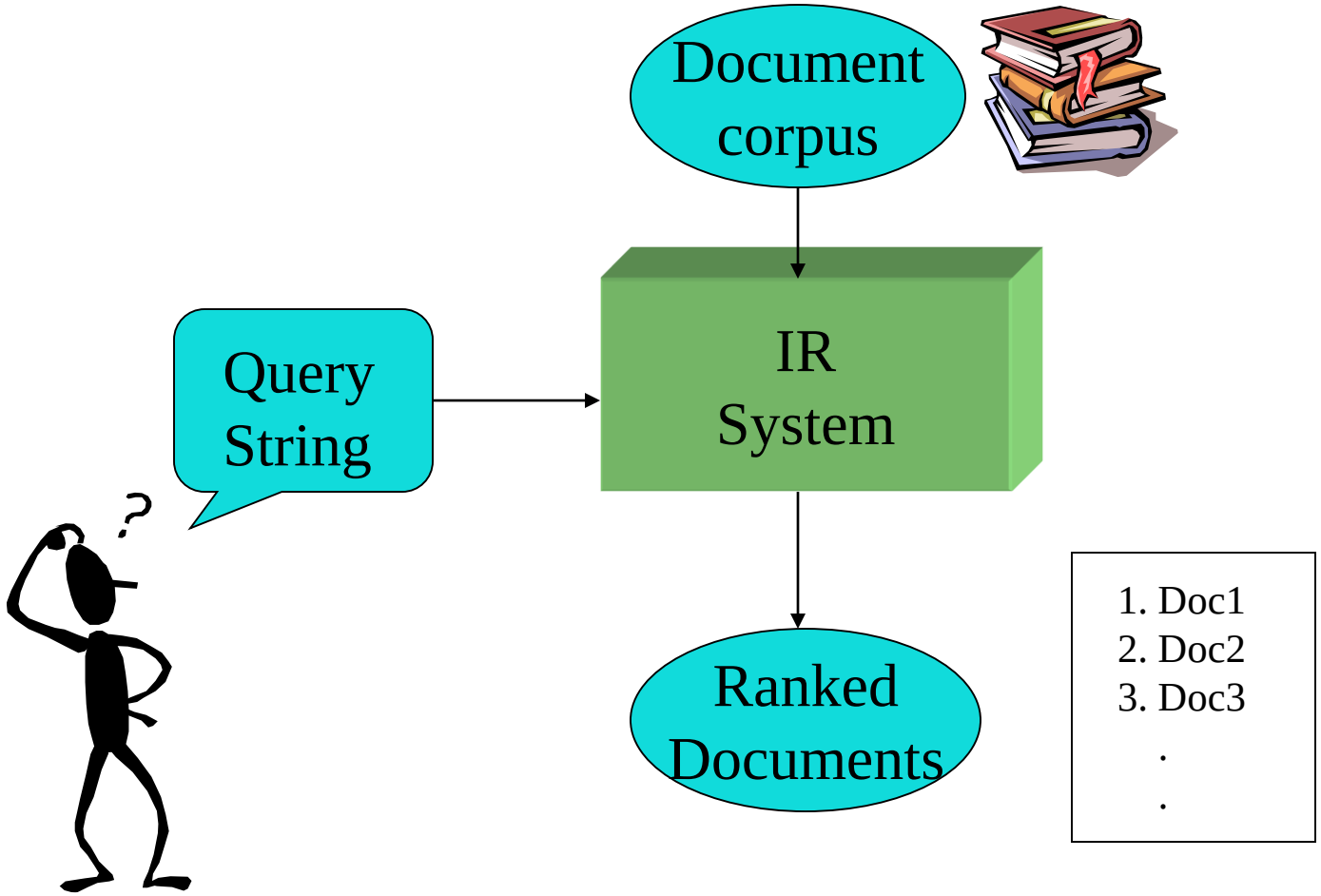
- **相关(Relevant)、相关度(Relevance)、相似度(Similarity)**
  - 相关取决于用户的判断，是一个主观概念
  - 不同用户做出的判断很难保证一致
  - 即使是同一用户在不同时期、不同环境下做出的判断也不尽相同

- **定义“相关性”的两个角度：**
  - **系统角度**：系统输出结果，用户是信息的接受者。
    - 这种理解置用户于被动的地位，研究的重心落在系统本身。
    - **主题相关性**：检索系统检出的文档的主题即核心内容与用户的信息需求相匹配。
    - 系统角度相关并不和用户脱节。系统角度定义的相关简单可以计算。
  - **用户角度**：观察用户对检索结果的反应，是系统输出向用户需求的投射。
    - 相关性被认为是用户方面的属性。
    - 用户角度定义的相关目前仍然难以计算。
- 现代信息检索研究中仍然主要采用**系统角度**定义的主题相关性概念，当然也强调考虑用户的认知因素。


查询
文档
相关度
文档集

- 形式上说，信息检索中的**相关度**是一个函数f，输入是查询Q、文档D和文档集合C，返回的是一个实数值 R, $R = f(Q,D,C)$
- **信息搜索:** 给定一个查询Q，从文档集合C中计算每篇文档D与Q的相关度并排序(Ranking)。
- 相关度通常只有相对意义，对一个Q，不同文档的相关度可以比较，而对于不同的Q的相关度不便比较。
- 相关度的输入信息可以更多，比如用户的背景信息、用户的查询历史等等。
- 现代信息检索中相关度不是唯一度量，如还有：重要度、权威度、新颖度等度量。或者说这些因子都影响“相关度”。

信息检索系统的基本组成



- 用户接口(User Interface)：用户和IR系统的人机接口
  - 输入查询(Query)
  - 返回排序后的结果文档(Ranked Docs)并对其进行可视化(Visualization)
  - 支持用户进行相关反馈(Feedback)
- 用户的两种任务：retrieval 或者 browsing
- IR的两种模式： pull (ad hoc) 或者 push (filtering)
  - **Pull**: 用户主动发起请求，在一个相对稳定的数据集合上进行查询，比如百度搜索
  - **Push**: 用户事先定义自己的兴趣，系统在不断到来的流动数据上进行操作，将满足用户兴趣的数据推送给用户。典型就是推荐系统，比如今日头条。

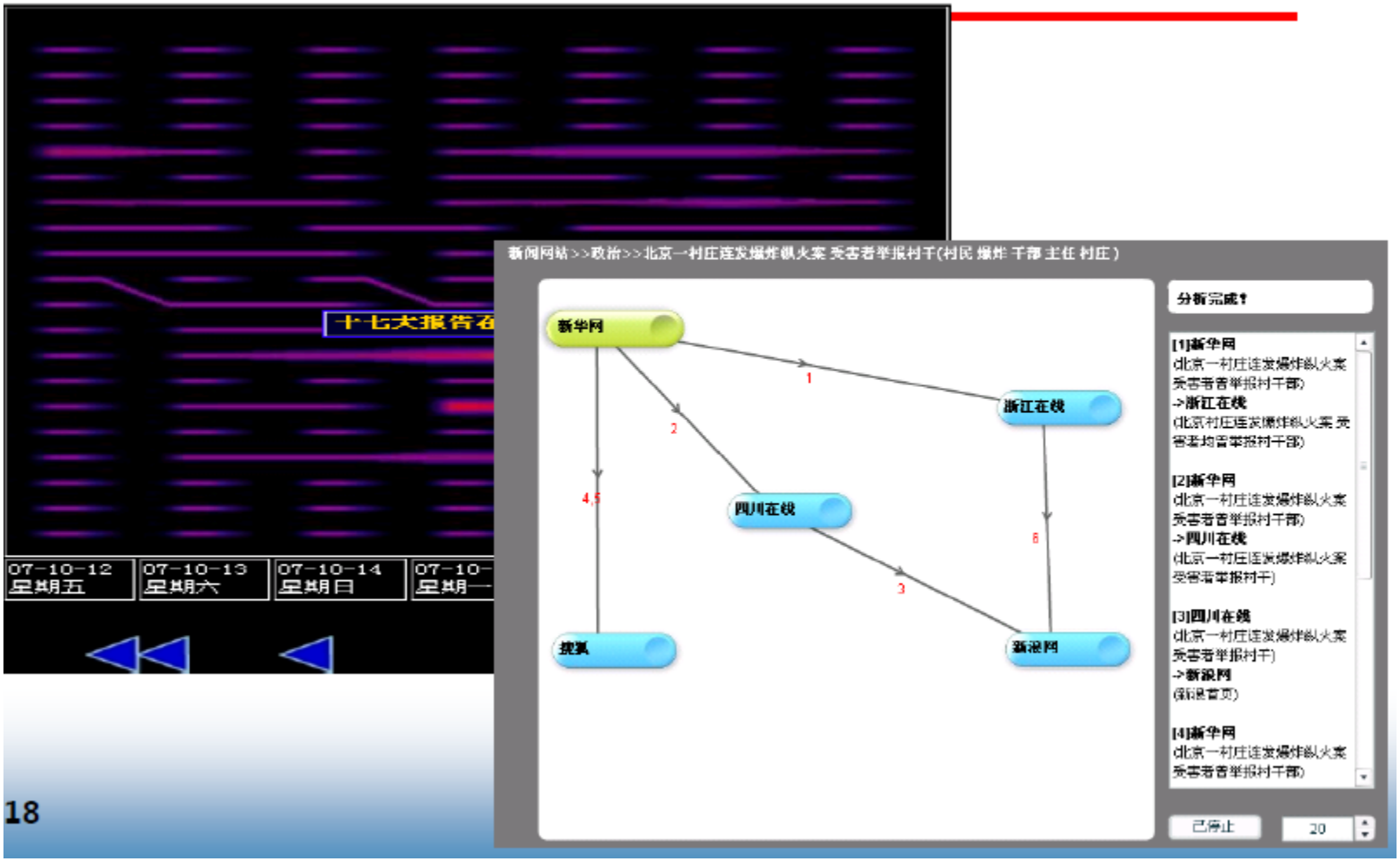
- **文本处理(Text Operations)**：对查询和文本进行的预处理操作
  - 中文分词(Chinese Word Segmentation)
  - 词干还原(Stemming)
  - 停用词消除(Stop Word Removal)
- **查询处理(Query Operations)**：对经过文本处理后的查询进行进一步处理，得到查询的内部表示(Query Representation)
  - **查询扩展**(Query Expansion)：利用同义词或者近义词对查询进行扩展
  - **查询重构**(Query Reconstruction)：利用用户的相关反馈信息对查询进行修改

- **文本索引(Indexing)**：对经过文本处理后的文本进行进一步处理，得到文本的内部表示(Text Representation)，通常基于索引项(Term)来表示
  - 向量化、概率计算
  - 组成倒排表进行存储
- **搜索(Searching)**：从文本中查找包含查询中索引项的文档
- **排序(Ranking)**：对搜索出的文档按照某种方式来计算其相关度
- **Logical View**：指的是查询或者文档的表示，通常采用一些关键词或者索引项(index term)来表示一个查询或者文档。

搜索技术的应用

- 垂直搜索

- 舆情监控

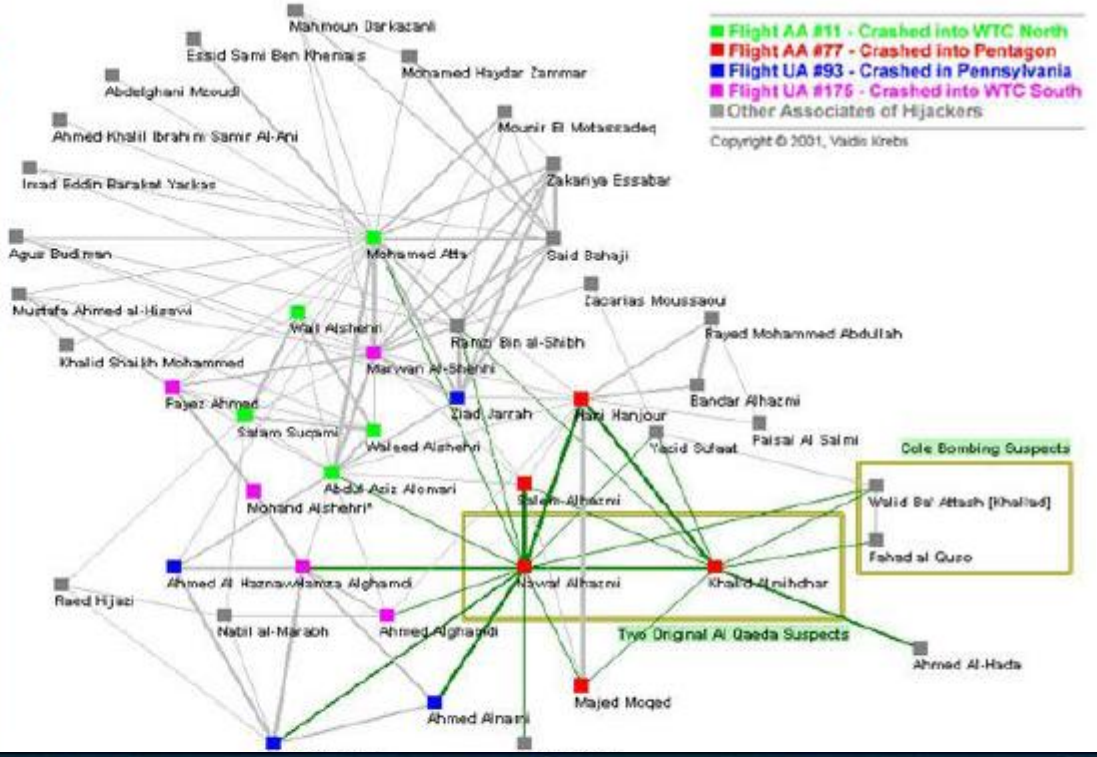

十七大报告
07-10-12 星期五
07-10-13 星期六
07-10-14 星期日
新闻网站>>政治>>北京一村庄连发爆炸纵火案 受害者举报村干(村民 爆炸 干部 主任 村庄)
新华网
浙江在线
四川在线
搜狐
新浪网
1
2
4.5
6
3
分析完成!
[1]新华网
(北京一村庄连发爆炸纵火案 受害者举报村干部)
->浙江在线
(北京村庄连发爆炸纵火案 受害者均举报村干部)
[2]新华网
(北京一村庄连发爆炸纵火案 受害者举报村干部)
->四川在线
(北京一村庄连发爆炸纵火案 受害者举报村干)
[3]四川在线
(北京一村庄连发爆炸纵火案 受害者举报村干)
->新浪网
(返回首页)
[4]新华网
(北京一村庄连发爆炸纵火案 受害者举报村干部)
已停止
20
18

- 情报与反恐

BASICS OF SOCIAL
FOR COUNTERTER

By Rebecca Garcia

It is well-known that extremist organizations have embraced the Internet for its vast and immediat worldwide reach for recruitment. In 2006, the University of Haifa's Gabriel Weimann published groundbreaking research in his book *Terror on the Internet: The New Arena, the New Challenges*, and discovered approximately 4,300 terrorist-related websites



- 预测

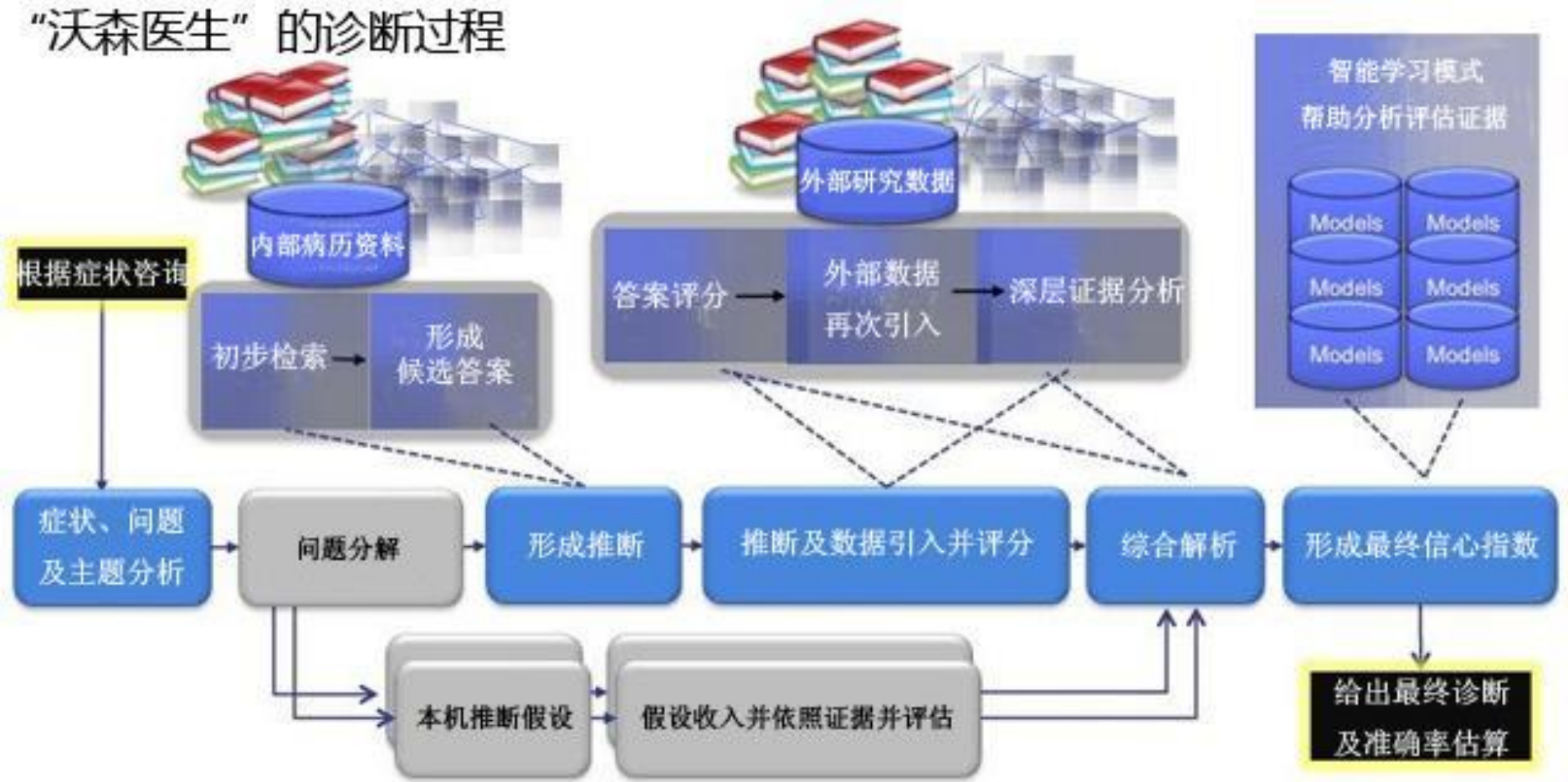
实例：2013年奥斯卡大奖花落谁家？

微软纽约研究院专家
David Rothschild

Oscars Ballot Predictor（奥斯卡投票预测器），为所有24个类别的奥斯卡得奖奖项提供实时预测

- 谷歌AlphaGo

- 精准医疗


"沃森医生"的诊断过程
内部病历资料
外部研究数据
智能学习模式
帮助分析评估证据
Models
Models
Models
Models
Models
Models
根据症状咨询
初步检索
形成
候选答案
答案评分
外部数据
再次引入
深层证据分析
症状、问题
及主题分析
问题分解
形成推断
推断及数据引入并评分
综合解析
形成最终信心指数
本机推断假设
假设收入并依照证据并评估
给出最终诊断
及准确率估算

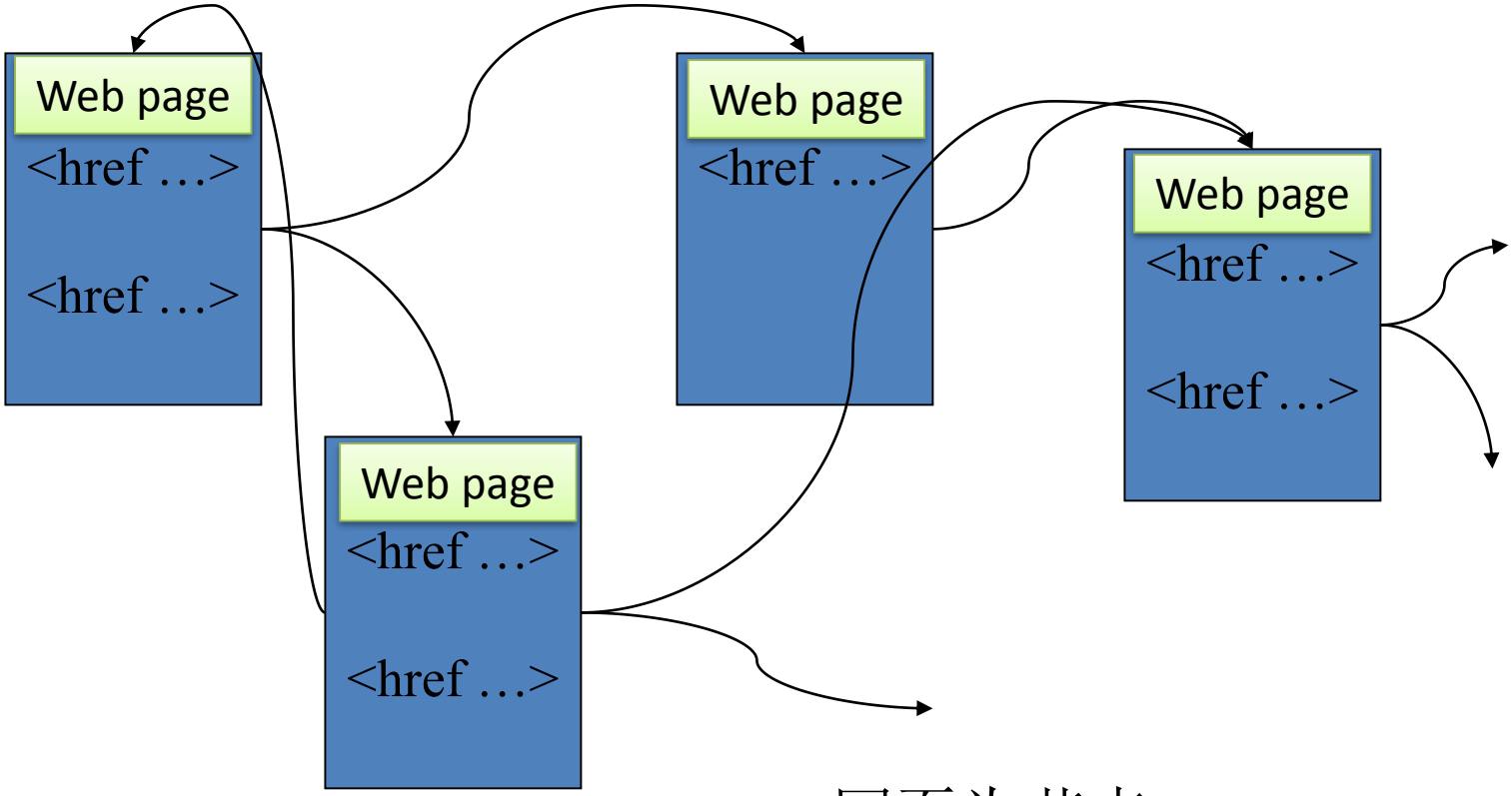
Web 搜索技术

- WWW发展迅速，Web文本数据以P=1000T计。而多媒体数据，如图片、视频、音频信息以1000P计，而且正快速增加。
- Web可以看做巨型的、**非结构化**的、无所不在的数据库，除了传统的书面文档，目前口语形式的文档正在迅速增加(微博、微信、短信等)
- Web的发展要求有效的工具来管理、检索、过滤信息：Data mining
- 结论：信息获取、组织和搜索将成为现代社会的不可缺少的基础服务。

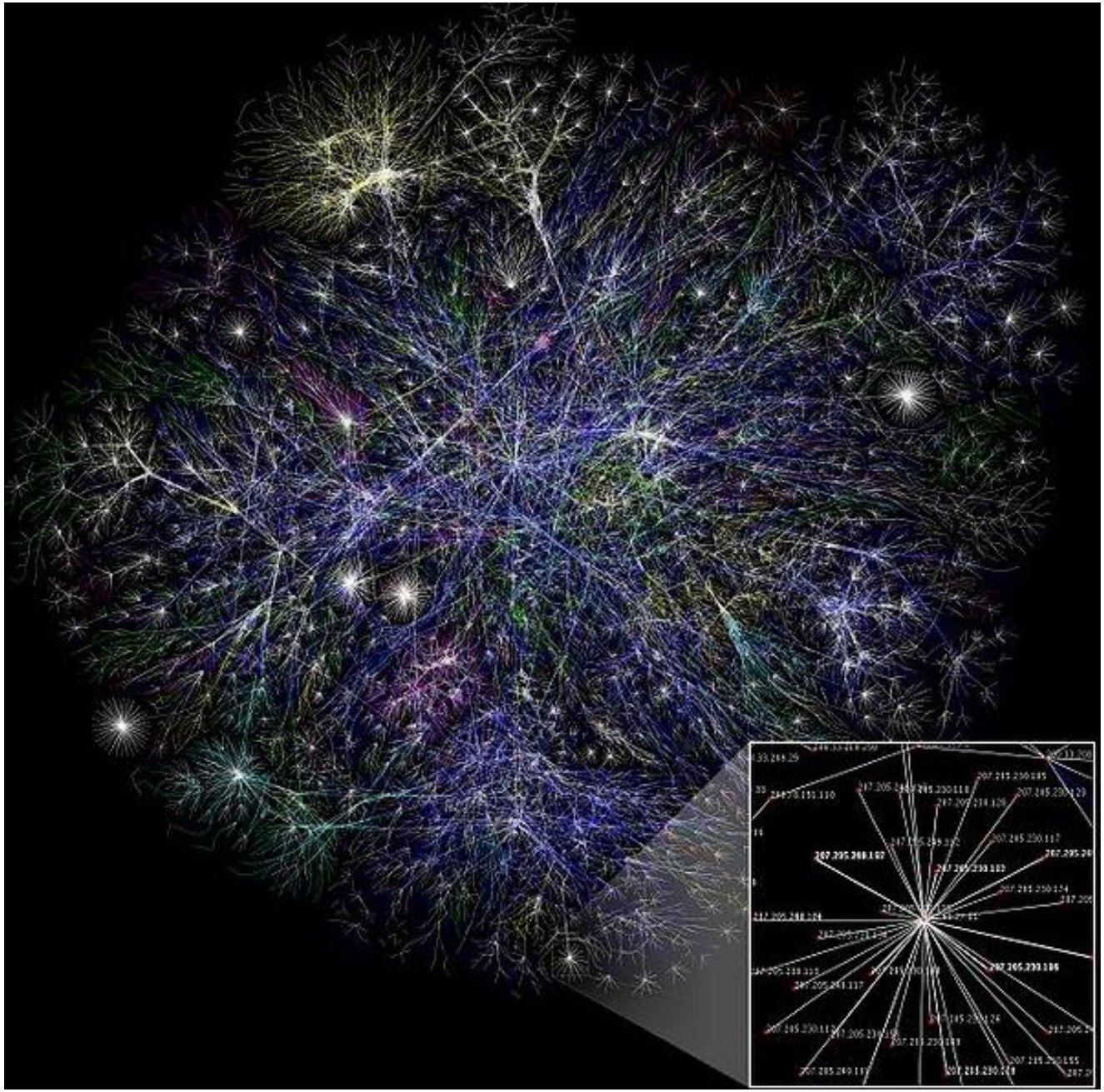
World Wide Web是什么?

- The world wide web (web) is a network of information resources. The web relies on three mechanisms to make these resources readily available to the widest possible audience:
- 1. A uniform naming scheme for locating resources on the Web (e.g., URLs).
- 2. Protocols, for access to named resources over the Web (e.g., HTTP).
- 3. Hypertext, for easy navigation among resources (e.g., HTML).

Web资源的拓扑结构(有向图)



网页为节点
网页中的HyperLink为有向边



- 搜索引擎市场的激烈竞争
  - Google，Bing，Baidu，阿里巴巴，京东，360，搜狗，神马对搜索引擎市场激烈竞争，公司上市造就数以百计的千万富翁，对与搜索相关的人才需求强烈。
  - MSRA建立搜索研究中心，努力追赶，Bing依然差距明显。
  - 主要门户网站，电子商务网站纷纷推出搜索引擎产品，如阿里巴巴、京东等
    - 面向领域的搜索引擎

目前最受北美IT毕业生青睐的公司

- Google
- Apple
- Facebook
- Microsoft
- Yahoo!
- 其中有3个公司的主要业务之一是信息检索。
- 微软已经认为过去没有重视搜索是犯了巨大错误，目前正在飞速追赶。
- 目前Google、微软在美国正在建立巨型的数据处理中心，保存通过Internet、卫星等收集的数据。
- 上述公司的搜索部门在面试时，一般会问关于现代检索的核心算法、概念等。
- 从全球范围看，急需能开发搜索、数据挖掘和大数据分析的人才，招人数量巨大。

国内IT

- 百度
- 搜狗
- 360
- 人民搜索
- 一搜
- 爱问
- 阿里巴巴，淘宝
- 腾讯
- 京东
- ……

- **中国大学生评出的2008最佳雇主中有半数和搜索有关。**

市场的需求

- 所有的网站都需要搜索引擎，但目前很多借用Google等，搜索效果不好，因为不专业，Google的排序算法可能并不适合该网站的页面链接分析。
  - 微博、twitter 、微信等实时信息检索
  - Facebook等涉及社会网络的实时信息检索以及事件发现
  - Flickr、Youtube 等共享图片和视频信息的检索
  - 医学和生物领域的检索
  - 无论是政府还是商家，都更重视网络上的舆情(商业情报)，包括讨论的热点及反应出的问题。而这些必须基于对网站、论坛、博客、社会网络等媒体内容的搜集和分析。

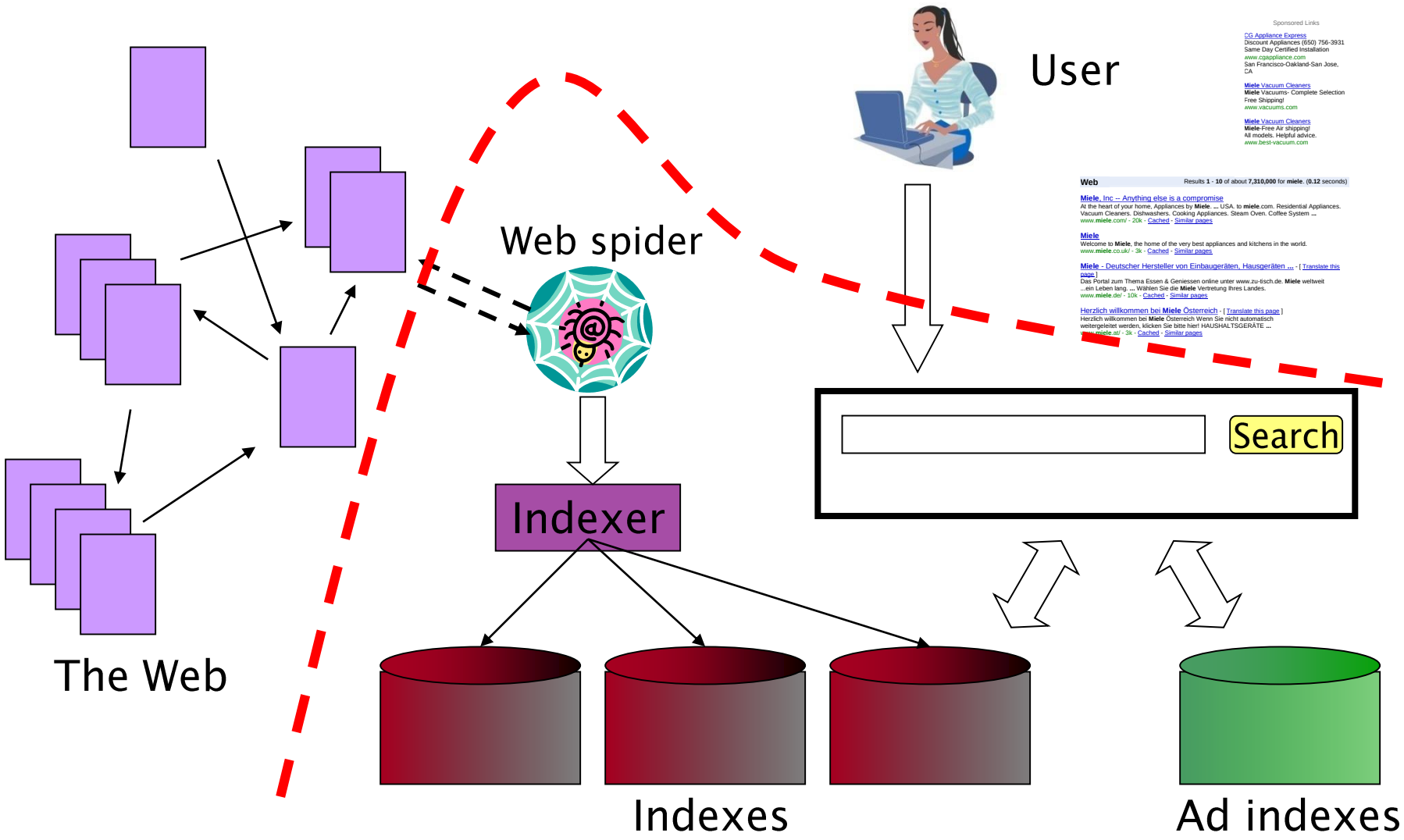
搜索技术现状

- 目前基本认为是处于关键词(bag of words)匹配的阶段
- 未来的远景图
  - 个性化
  - 专业化
  - 多媒体检索
  - 智能化
  - 自动问答
  - 自动对话
- 检索形式的创新，不是输入关键词，可能是任务(Task)，得到的是解决方案(Solution)。

和以往IR研究的区别

- 原来是图书馆学的专业方向，但目前的研究和以前的研究相差太大了。
- 目前的研究者来自
  - 数据库、人工智能、计算机算法、人机交互、多媒体、移动通讯、模式识别、地理信息和卫星图像处理等，都加入到这个领域。

Web search basics



进入2000后对 IR 的研究

- **2000's**
  - 基于网页链接分析技术
    - Google
  - 网页的排序技术
    - PageRank
  - 网页的分析技术
    - 基于块的分解，内容提取和理解

IR 新课题1：自然语言理解

- 自动文本摘要或主题的提取
  - 抽取式 Extractive Summarization
  - 生成式Abstractive Summarization
- 会话体文本的处理
  - 微博，博客，微信
- 自然语言的复杂性

➢ 话说普京与奥巴马赛跑，普京第一，奥巴马第二

➢ 俄罗斯媒体：在国际领导人赛跑中，普京勇夺冠军，奥巴马倒数第一！

➢ 美国媒体：在国际领导人赛跑中，奥巴马勇夺亚军，普京倒数第二！

- 自然语言的复杂性
  - 南京市长江大桥
    - 南京 市长 江大桥
    - 南京市 长江大桥
  - 兵乓球拍卖完了
    - 兵乓 球拍 卖完 了
    - 兵乓球 拍卖 完了

- 人工智能的终极目标！

IR 新课题2：多媒体检索

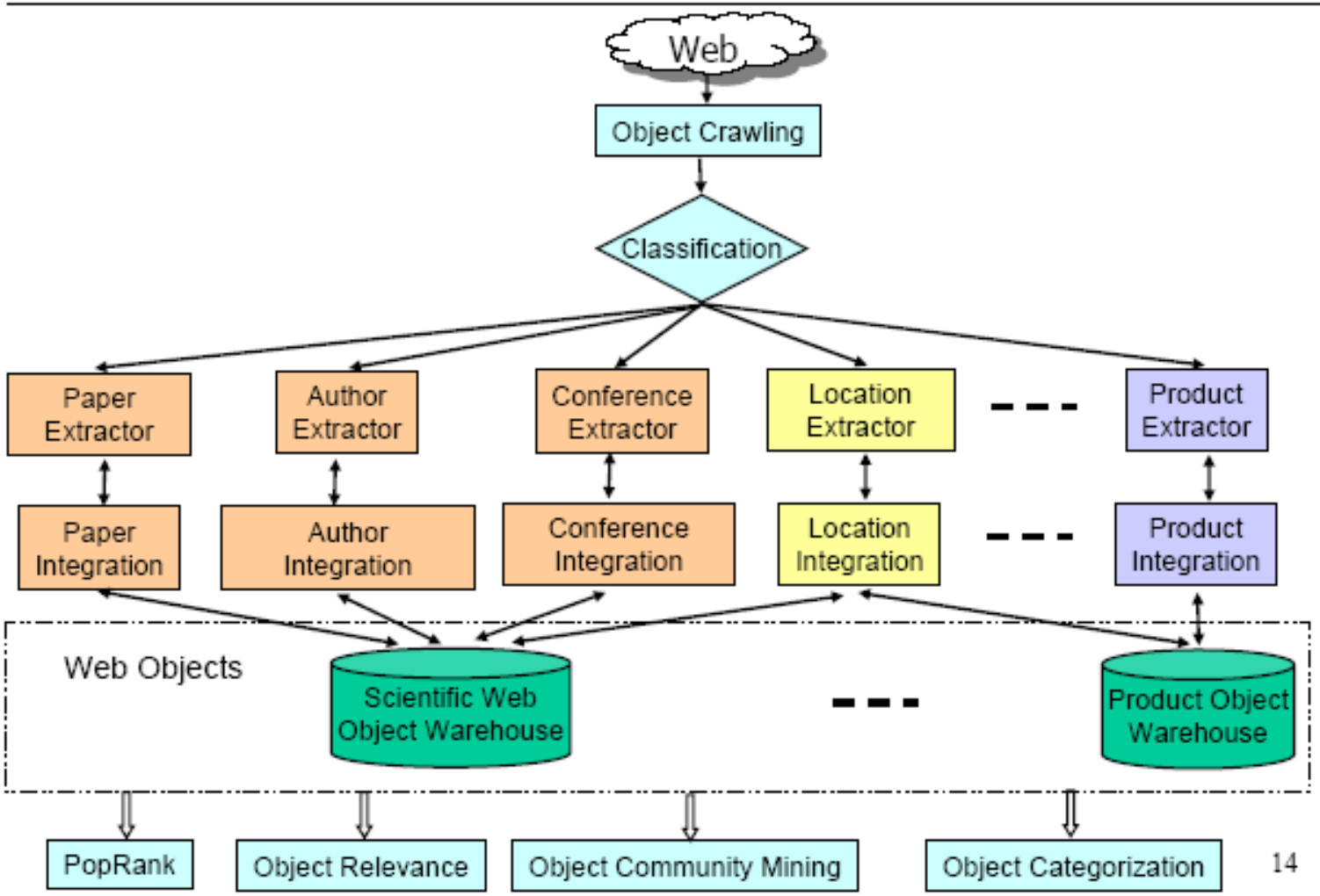
- Multimedia (多媒体的检索技术)
  - Image
  - Video
  - Audio and music
- 图像的自动标注,使得对图像的检索可以借助关键词进行检索

Image				
Original Annotation	people pool swimmers water	cars formula tracks wall	clouds mountain sky water	field foals horses mare
Automatic Annotation	water people swimmers pool	cars tracks wall formula	sky mountain clouds park	field horses foals mare

IR新课题3：垂直检索技术

- 目前通用的搜索引擎对Internet网页的覆盖率小于50%。未来计算机的存储和运算能力都不可能100%的覆盖，需要面向具体领域的专用搜索引擎
  - 如就业、股票、宾馆饭店、地图、天气、商品查找、任务、旅游，.....
- 垂直检索也是未来利用移动通讯设备检索的支撑技术

微软的科技论文垂直检索系统

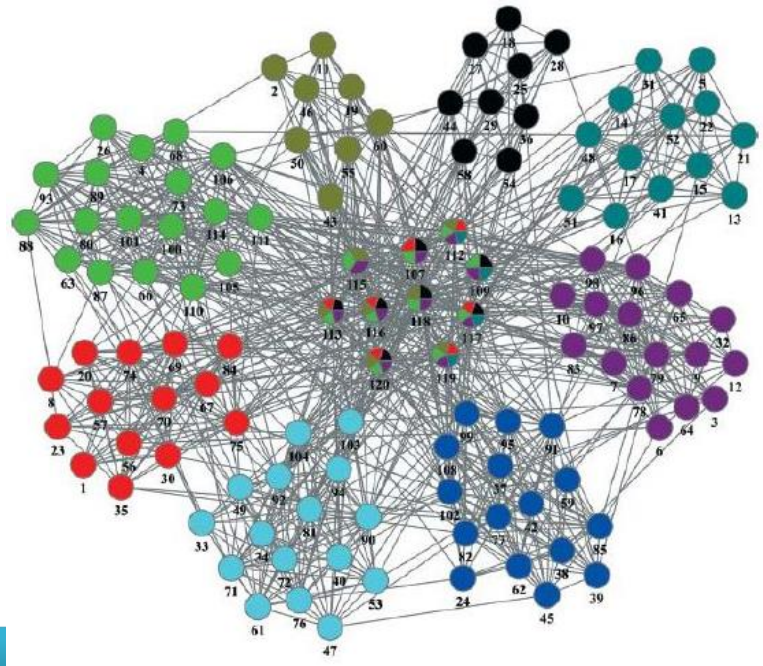


IR 新课题4: 移动搜索-Mobile Search

- Google、百度、NEC、微软等早就已经开始了对手机检索系统的研究。
- 解决的科学与技术问题
  - 如何在小屏幕上显示用户的检索结果，如何在小键盘上输入不方面的情况下进行文本检索？
  - 能否通过语音检索？
  - 能否通过图片进行检索？因为手机更方便语音和图像。

IR 新课题5：对社会媒体信息检索

- 微博、Twitter等实时信息检索
- Wordpress (博客)、Yahoo!answers、百度知道(论坛)
- Flickr、YouTube(共享视频)
- Facebook、人人(社交网络)
- Wikipedia(在线百科词典)



现实世界和虚拟世界

- 目前我们生活在2个世界
  - 现实世界和网络上的虚拟世界
- 两者的关系是什么，以及如何利用虚拟世界了解现实世界，是目前的研究热点
  - 虚拟世界中用户之间的关系网络及社区划分
  - 虚拟世界中的领袖，利用虚拟世界的信息，发现现实世界的突发事件
    - 地震、疾病的传播、突发事件的发现、追踪和分析

如何利用虚拟世界研究现实世界

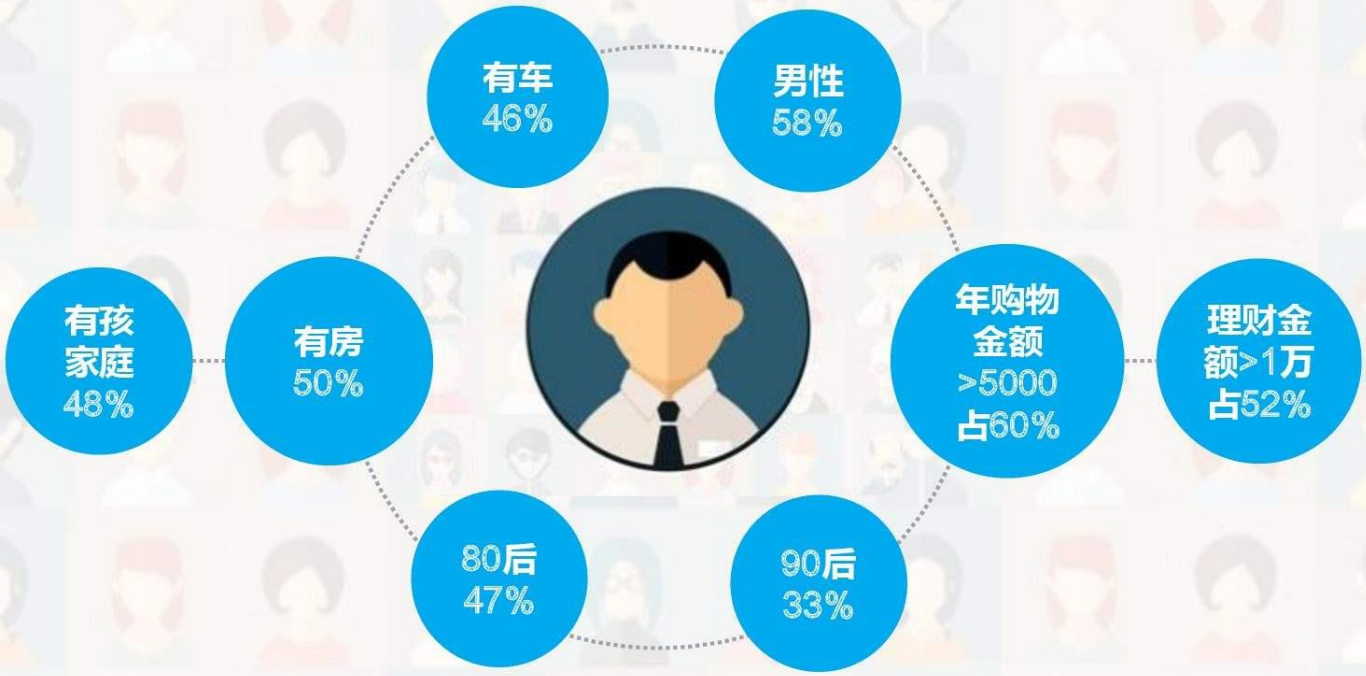
- 网上人气榜的排名
- 网上对产品、任务、事件的评论
- 网上所反映的最关心的热点和焦点是什么？
- 网上对事物、人物、事件的正面和反面评价的比例是什么？
- 作弊(Spam)网页、重复(Duplicate)网页及水军(Water Army)网页的发现

行为分析-用户画像

活跃互联网保民典型画像

注：对最为活跃的30%互联网保民进行的画像分析

CBNData 第一财经商业数据中心　蚂蚁金服 ANT FINANCIAL



IR新课题7：自动问答(Question-Answer)

- 定义：允许用户以自然语言方式询问，系统从单语或多语文档集中查找并返回确切答案或者蕴含答案文本片段，是一种新型信息检索方式。
  - 姚明的身高是多少？
  - 从青岛校区院到火车站怎么走？
  - 树上有三只鸟，又飞来了两只，现在有多少只？
- 问答系统是下一代搜索引擎的基本形态

IR-based QA

基于关键词匹配 + 信息抽取，仍然是基于浅层语义分析

Community QA

依赖于网民贡献，问答过程仍然依赖于关键词检索技术

YAHOO! ANSWERS BETA

KB-based QA

Knowledge Base

WolframAlpha computational knowledge engine

沃森智能问答IBM Watson

- 沃森(Watson)：2011年，IBM研发的超级计算机“沃森”在美国知识竞赛节目《危险边缘Jeopardy!》中上演“人机问答大战”，战胜人类选手Ken和Brad

辅助医疗

金融辅助决策

企业服务

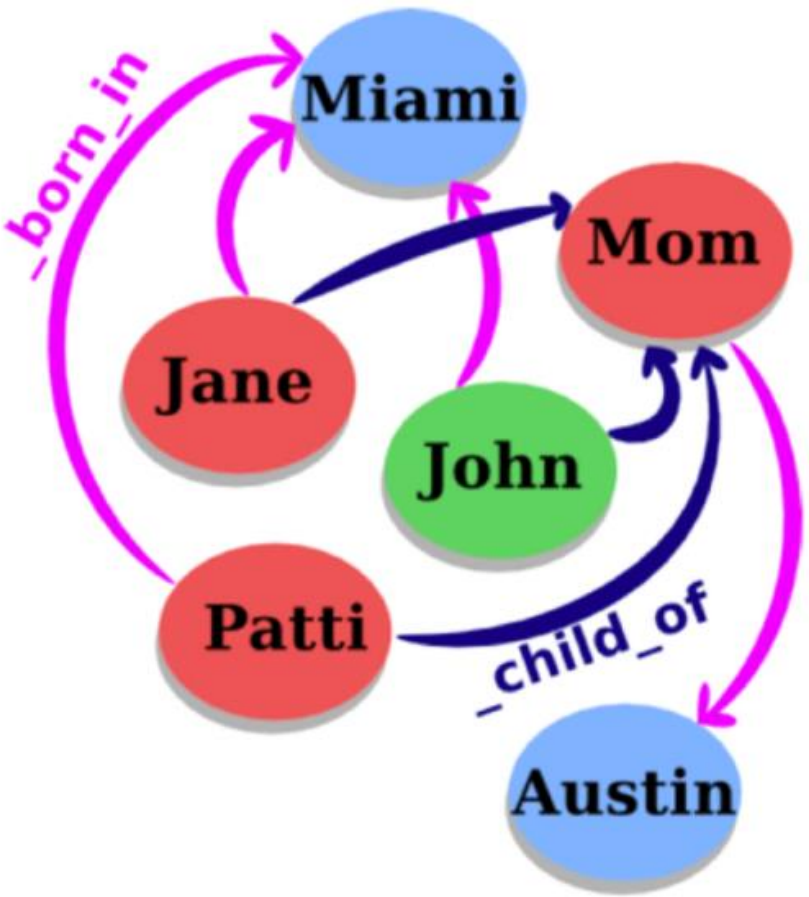
IR新课题8：知识发现

知识图谱的构建是目前的最大热门领域之一

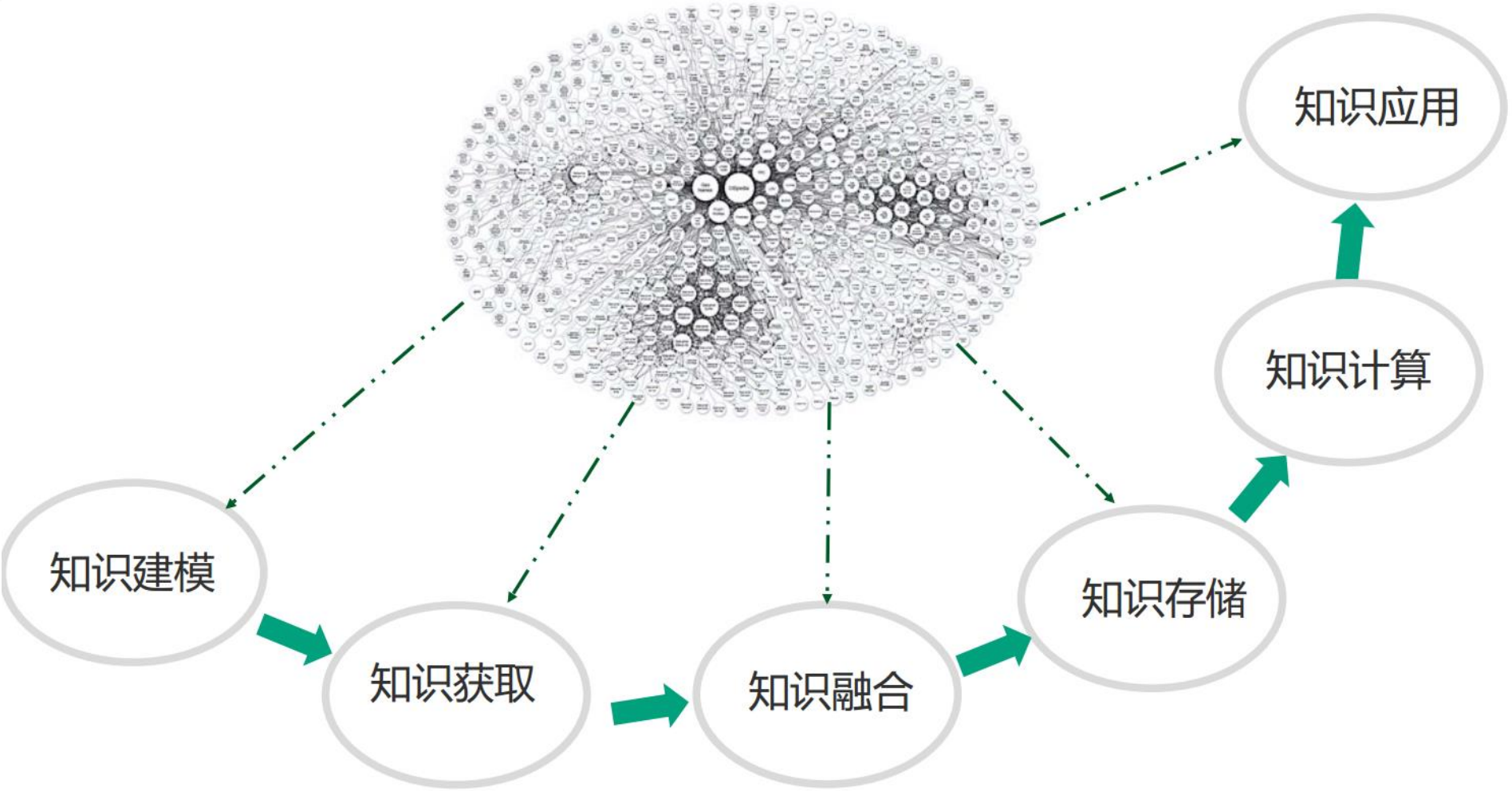
- Wiki 百科
- Freebase
- DBpedia
- Yago
- 百度百科
- 互动百科

知识图谱→实体与关系

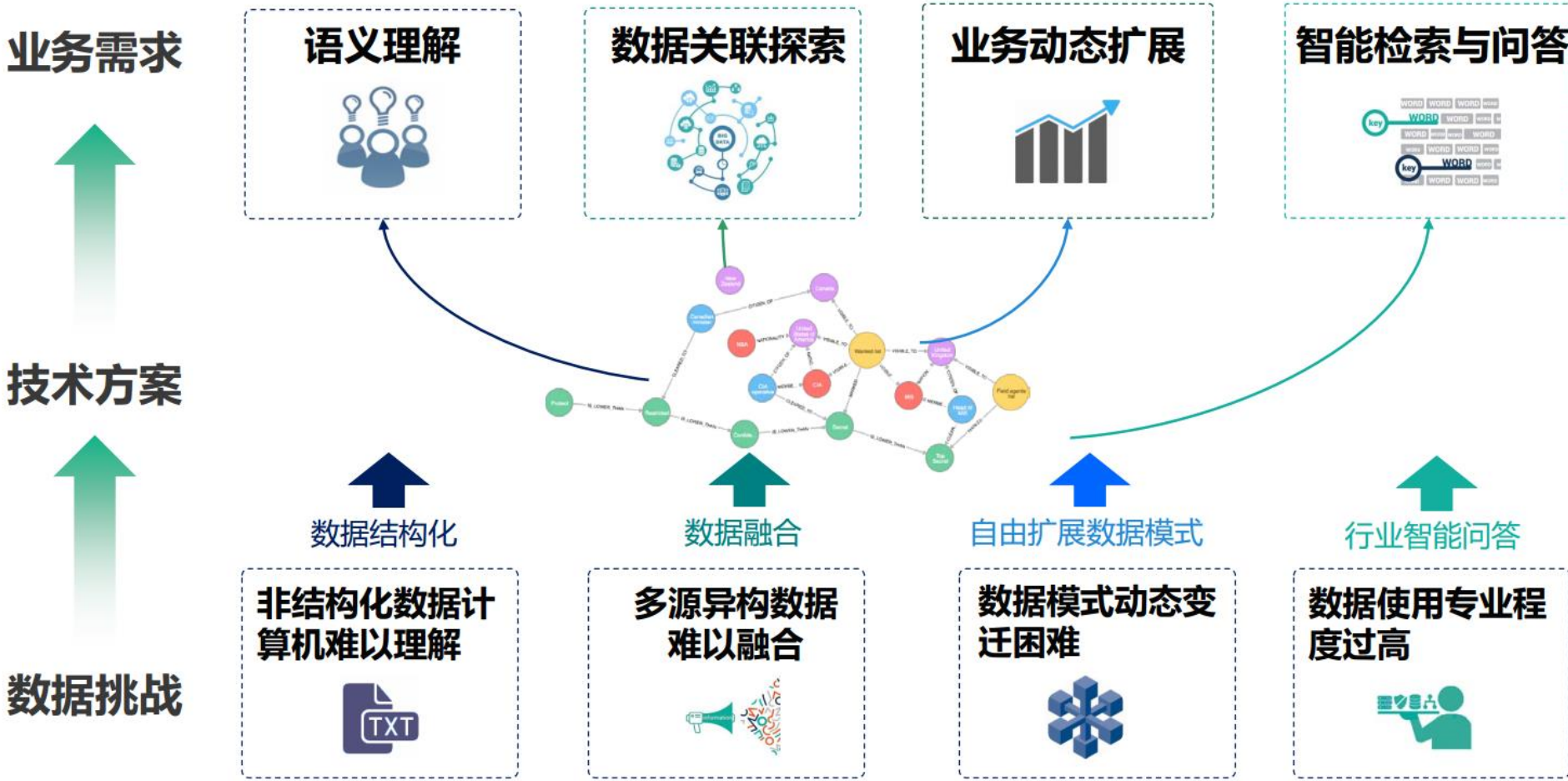
- 知识图谱包括实体与关系
  - 节点代表实体
  - 连边代表关系
- 事实可以用三元组表示
  - (head, relation, tail):
- 代表知识图谱
  - WordNet: 语言知识
  - Freebase: 世界知识



知识图谱的生命周期



知识图谱应用

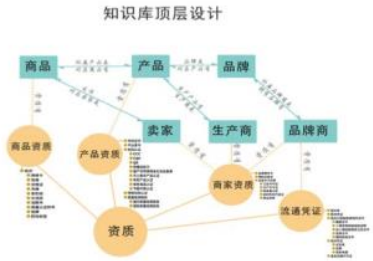

业务需求
技术方案
数据挑战
语义理解
数据关联探索
业务动态扩展
智能检索与问答
数据结构化
数据融合
自由扩展数据模式
行业智能问答
非结构化数据计算机难以理解
TXT
多源异构数据难以融合
数据模式动态变迁困难
数据使用专业程度过高

行业知识图谱应用一览

金融证券

生物医疗

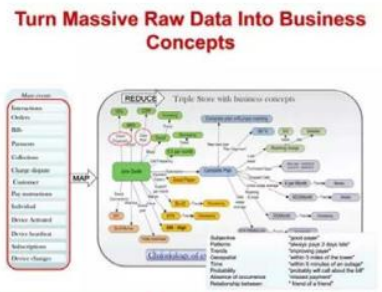
图书情报



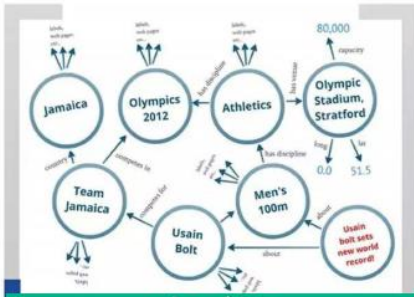
电商

农业

政府



电信



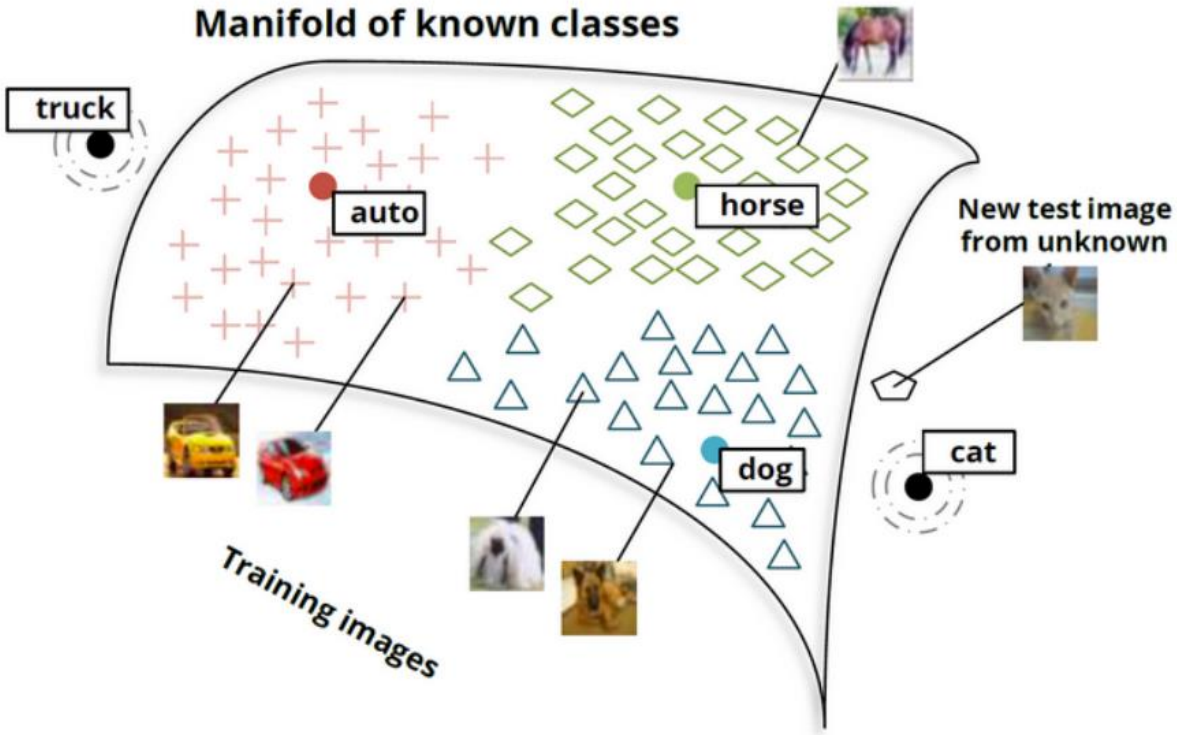
出版

知识图谱应用→自动推理

- 梁启超的儿子的老婆的情人的父亲

知识的表示学习-Represent Learning

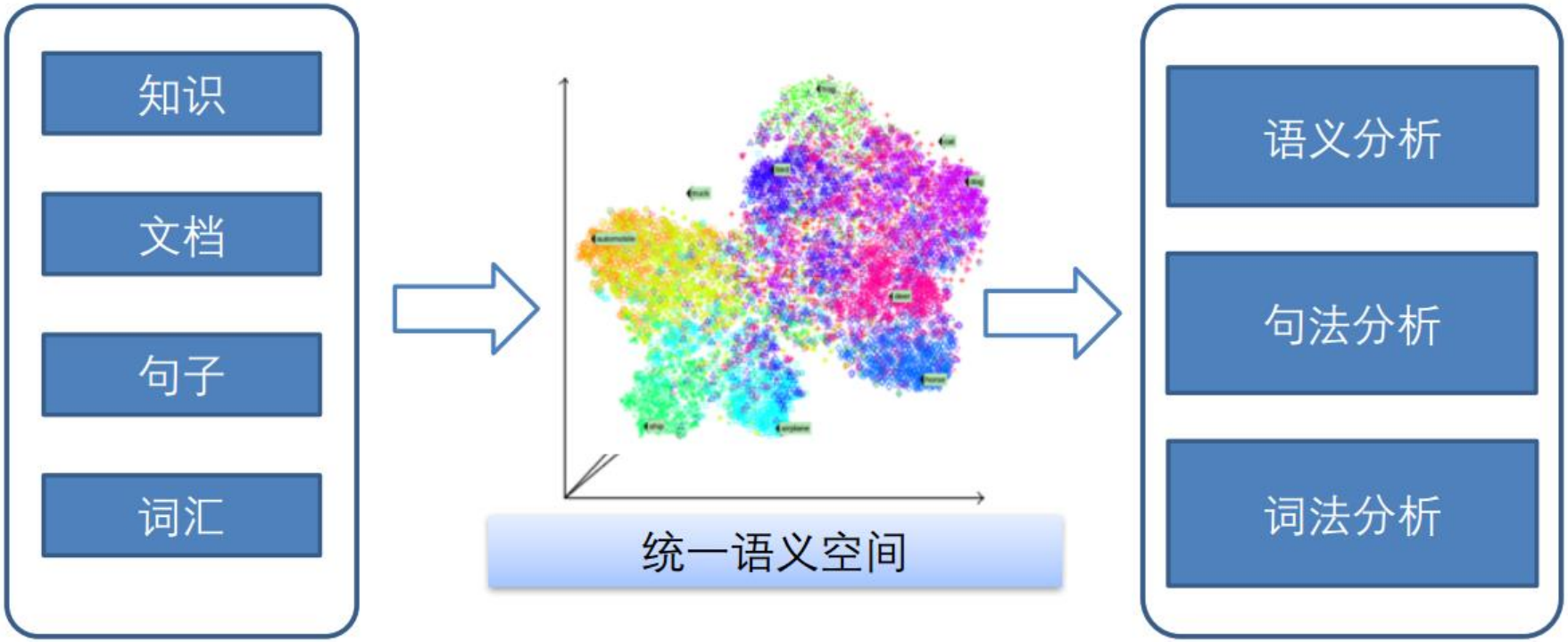
- Distributed Representation (Word Embeddings)
- 每个词被表示成稠密、实值、低维向量



- 男人-女人=国王-皇后

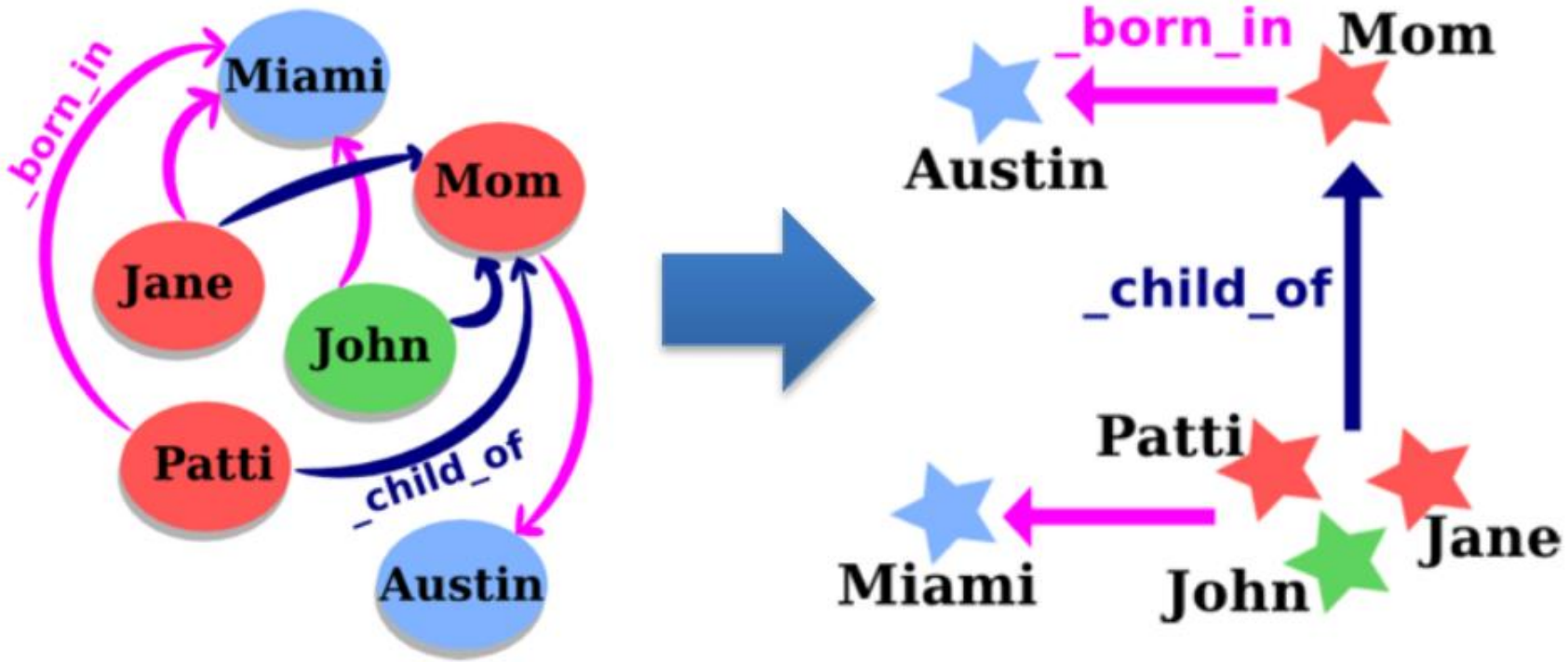
分布式表示的意义

- 解决大数据NLP的数据稀疏问题
- 实现跨领域、跨对象的知识迁移
- 提供多任务学习的统一底层表示



表示学习对知识图谱的表示

- 对每个事实 (head, relation, tail)，将其中的relation作为从head到tail的翻译操作



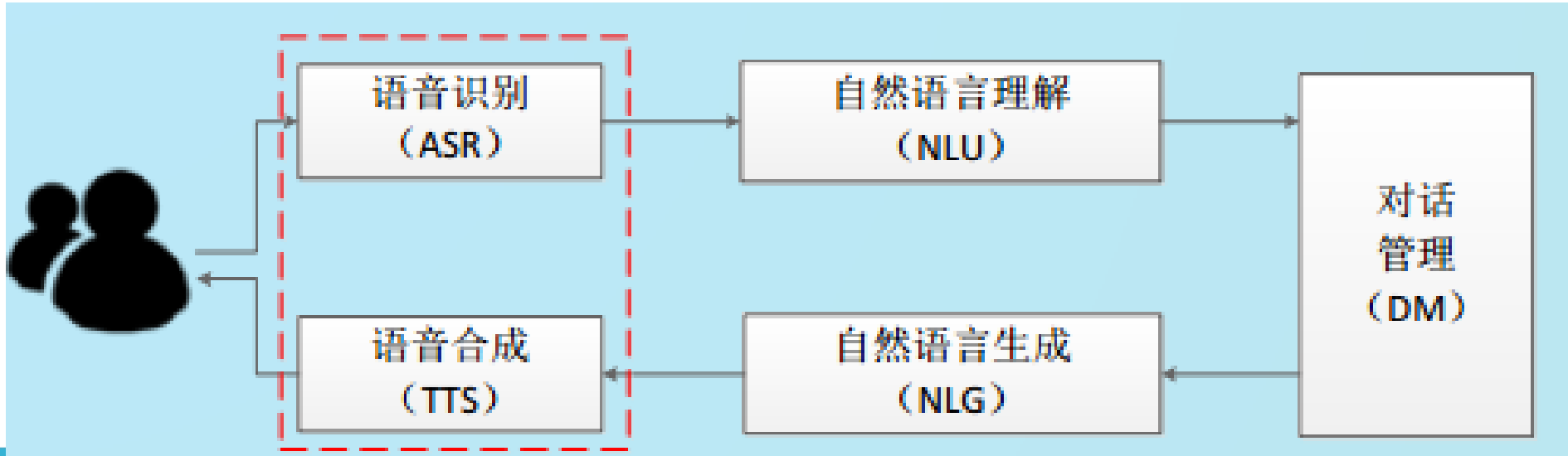
IR新课题9：行为分析、舆情监控

- 通过对网络上的信息收集，了解公众对某种社会现象或社会问题的具有一定影响力和倾向性的共同意见。
- 网络环境下舆情信息的主要来源有
  - 新闻评论、BBS、聊天室、博客、聚合新闻(RSS)、微博、微信等
- 利用Facebook、Twitter等发现人类自身活动等
- 利用Twitter、Facebook等发现野生动物的分布等

IR新课题10：自动对话

- 聊天机器人，是一种通过自然语言模拟人类进行对话的程序。
- 研究源于图灵(Alan M. Turing)在1950年在《Mind》上发表的文章《Computing Machinery and Intelligence》，提出了“机器能思考吗？”(“Can machines think?”)的设问，并且通过让机器参与一个模仿游戏(Imitation Game)来验证“机器”能否“思考”，进而提出了经典的图灵测试(Turing Test)。
  - 图灵测试是指测试者在与被测试者(一个人和一台机器)隔开的情况下，通过一些装置(如键盘)向被测试者随意提问。进行多次测试后，如果有超过30%的测试者不能确定出被测试者是人还是机器，那么这台机器就通过了测试，并被认为具有人类智能。
- 图灵测试被认为是人工智能的终极目标，图灵本人因此也被称作“人工智能之父”。

- 从应用场景的角度来看，可以分为在线客服、娱乐、教育、个人助理和智能问答五个种类。
  - 微软：小冰
  - 微软：Cortana
  - 百度：小度
  - Facebook：M
  - 苹果：SIRI

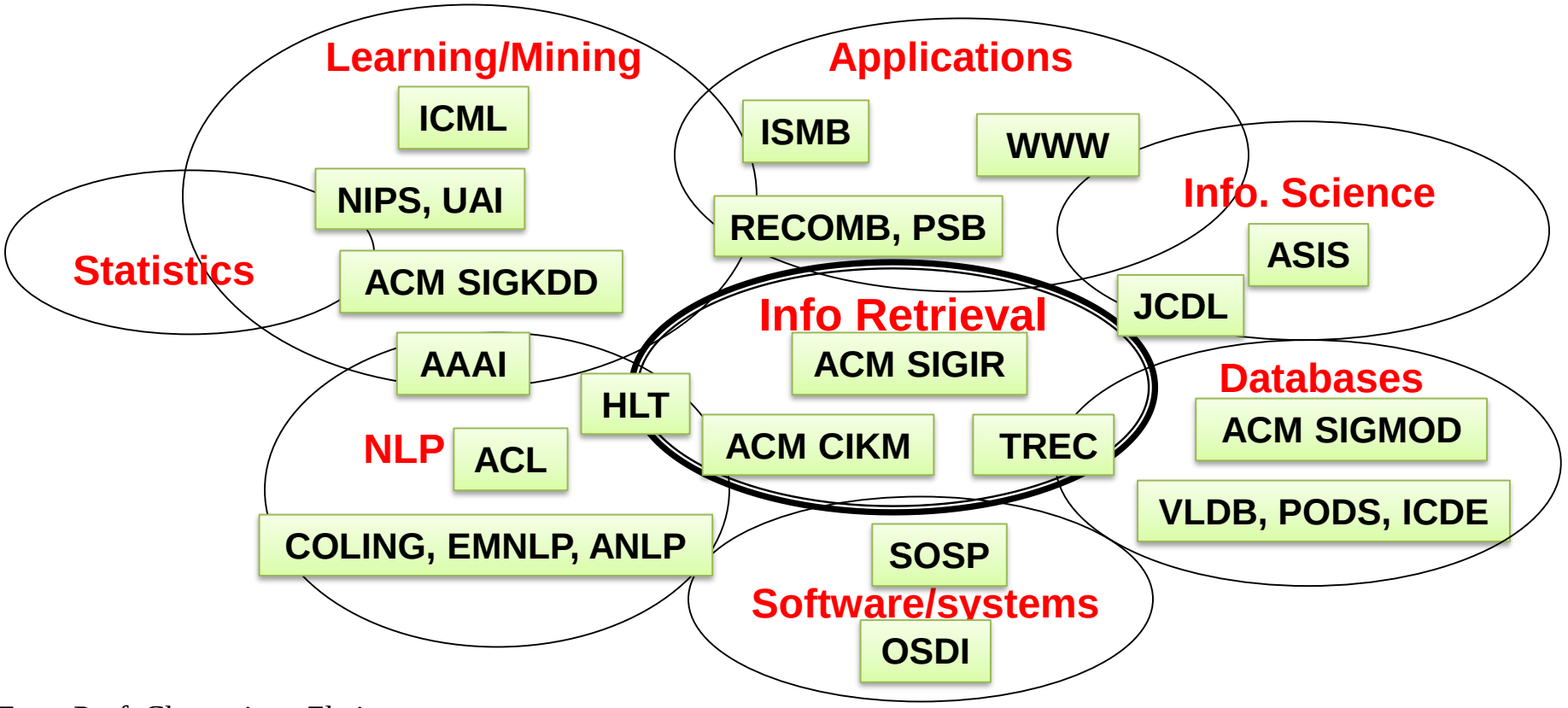


结论

信息检索技术

- 是一项飞速发展的科学技术。
- 是一项和人们生活密切相关的技术。
- 是计算机科学研究领域中为数不多的**理论和应用**密切相关的研究领域，即理论研究可直接导致应用系统的产生。
- 由于研究内容比较新，容易出成果。
- 和搜索技术相关的研究是易于自主创业的研究领域之一！

IR及相关研究领域重要会议



*From Prof. Chengxiang Zhai

重要工具

- Lemur、Indri：包含各种IR模型的实验平台，C++
- SMART：向量空间模型工具，C编写
- Weka：数据挖掘工具，Java编写
- Lucene：开源检索工具，Java版本受维护，存在各种语言编写的其他版本
- Nutch：开源爬虫，Java版本
- ElasticSearch：基于Lucene的企业级搜索引擎
- Sphinx：开源检索工具，C++
- Larbin：采集工具，C++
- Firtex：检索平台，C++，计算所开发
- 更多：http://www.searchtools.com/tools/tools-opensource.html