计算机信息存储与检索

第1章 信息检索的相关概念(IR concepts)

课前思考题

① 信息检索的定义？

② 信息检索中的用户需求、查询、相关度都是什么含义？

③ 信息检索和其他相关学科是什么关系？

④ 信息检索系统由哪些部分组成？各部分的功能是什么？

提纲

- 信息检索的基本概念
- 信息检索的历史
- 信息检索和其他学科的关系
- 信息检索的基本流程

信息过载(Information overload)

- **“…全世界每年产生1到2 EB(1 EB = 1024 PB)信息，相当于地球上每个人大概产生250MB信息。其中纸质信息仅占所有信息的0.03%...”(Lyman & Hal 03)**
- **静态网页有上百亿，动态及隐藏网页至少是静态网页的500倍。**
- **Tom Landauer认为人的大脑只能存储200M信息量，一辈子只能接触6G的信息量。**

Internet增长

Figure: Internet Domains (1989-1997) [see below for 2000-]

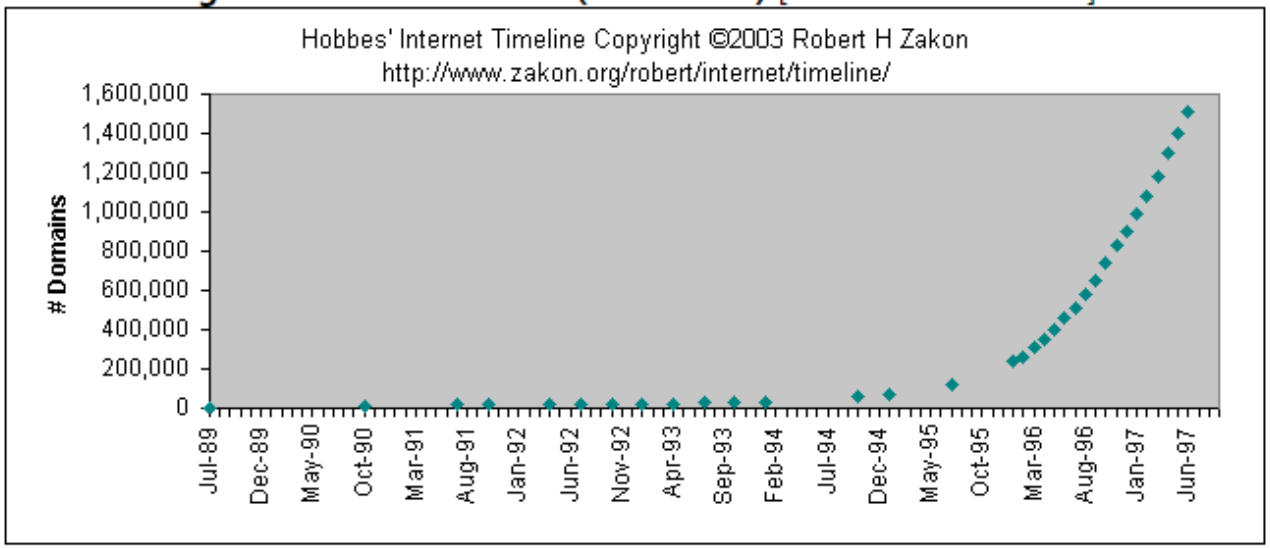
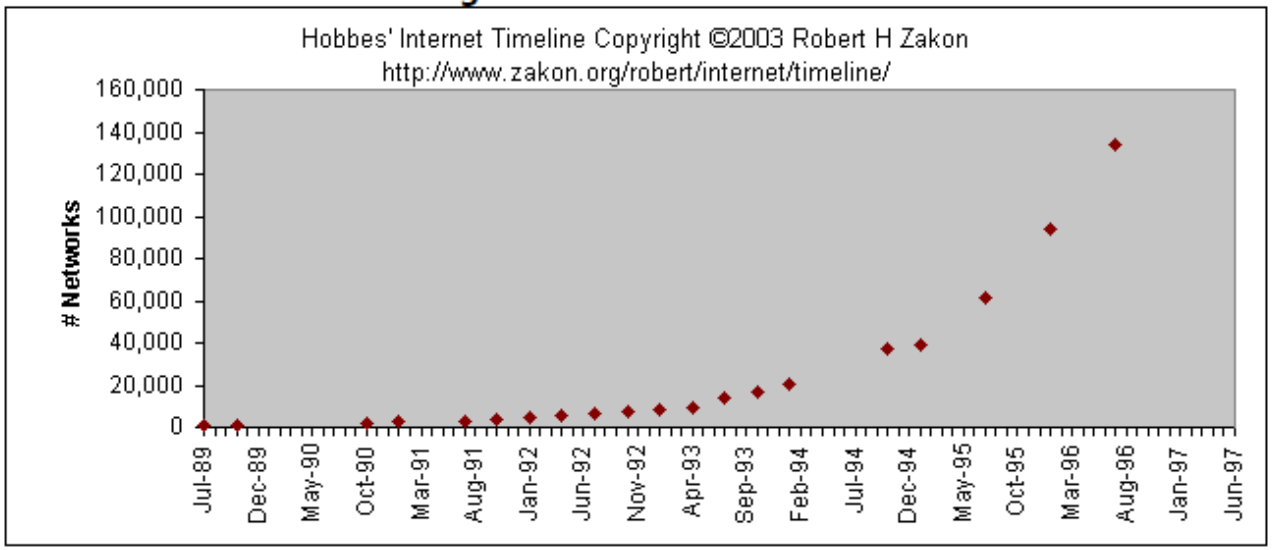


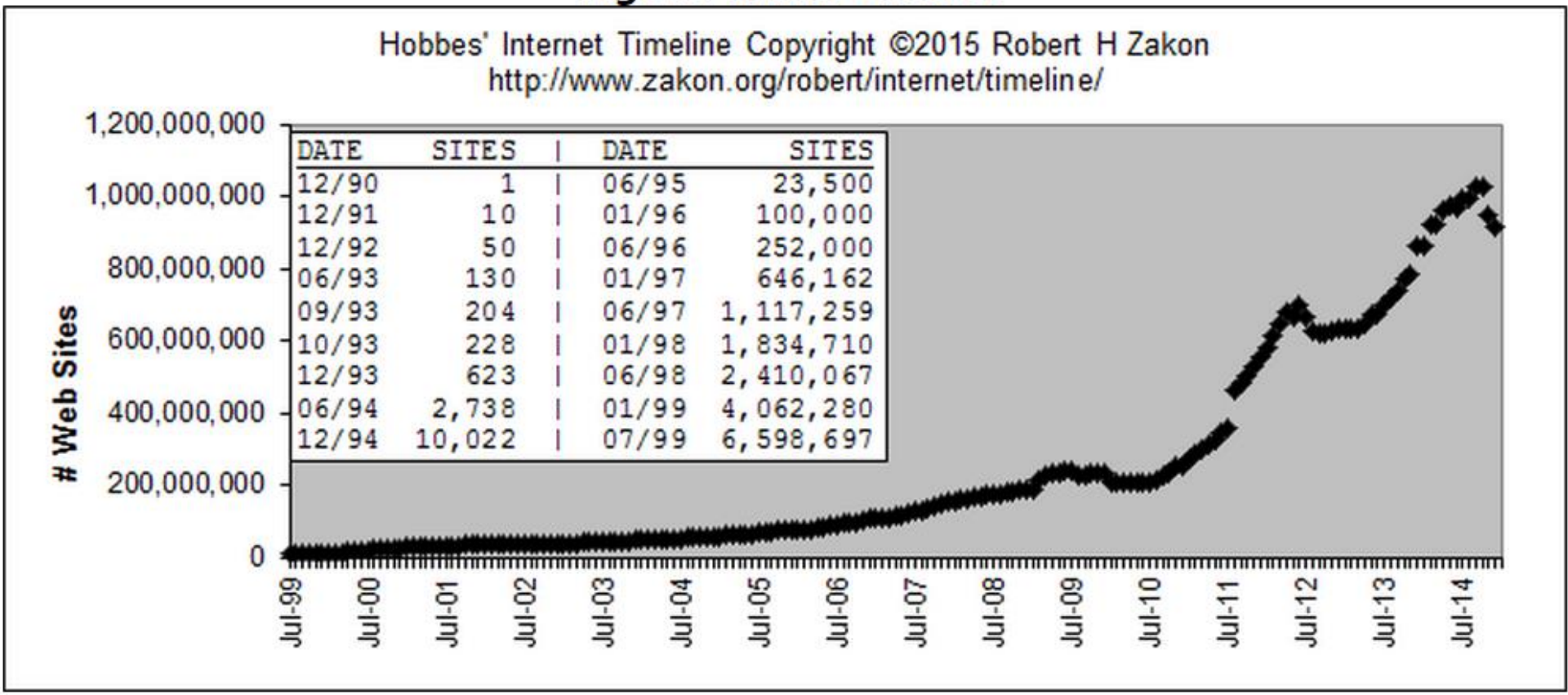
Figure: Internet Networks



南京理工大学
NANJING UNIVERSITY OF SCIENCE & TECHNOLOGY

Internet增长

Figure: WWW Growth

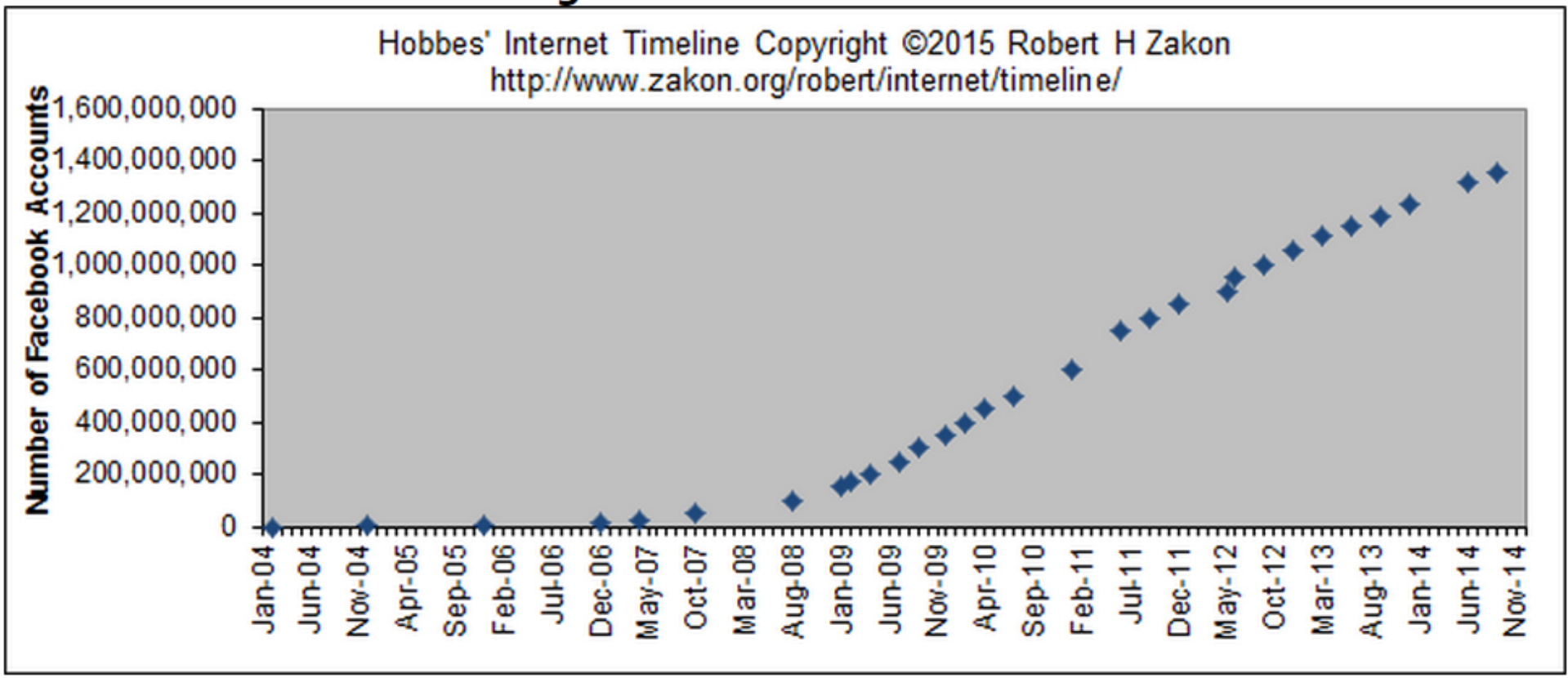


DATE	SITES	DATE	SITES
12/90	1	06/95	23,500
12/91	10	01/96	100,000
12/92	50	06/96	252,000
06/93	130	01/97	646,162
09/93	204	06/97	1,117,259
10/93	228	01/98	1,834,710
12/93	623	06/98	2,410,067
06/94	2,738	01/99	4,062,280
12/94	10,022	07/99	6,598,697

南京理工大学
NANJING UNIVERSITY OF SCIENCE & TECHNOLOGY

Internet增长

Figure: Facebook Accounts



南京理工大学
NANJING UNIVERSITY OF SCIENCE & TECHNOLOGY

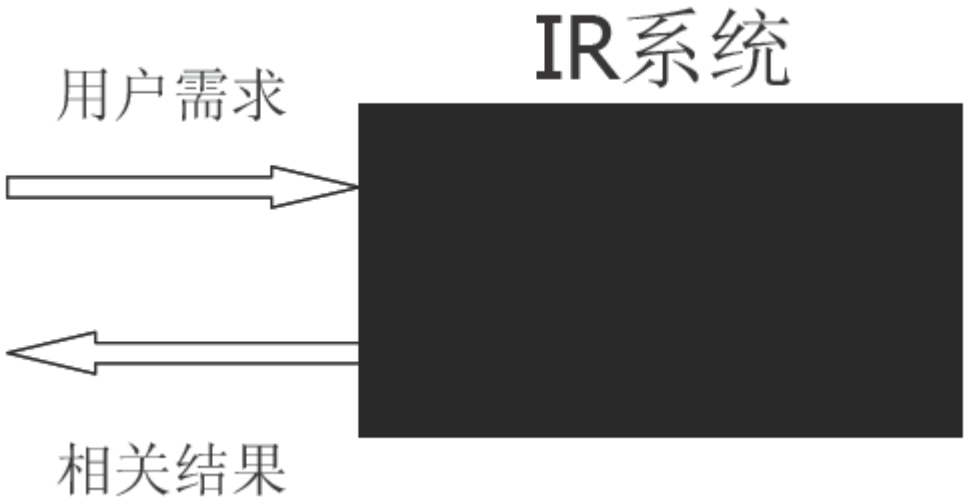
全球数字化进程加快

- 1998年，提出数字化地球的概念。
- 1998年，数字中国战略构想。
- 1999年，提出数字北京概念。
- 全世界启动了数字图书馆、数字博物馆在内的一系列工程，另外包括虚拟博物馆、数字电影、交互电视、会议电视、远程教育、遥感、GPS等在内的服务或应用也产生大量文本和多媒体数据。

问题

- 一方面，人们可以获得的信息的来源非常广泛。
- 另一方面，人们如何快速、准确、全面地获得自己所需要的信息？
- 非常困难！
  - 信息量太大，而且信息冗余度大、质量良莠不齐、格式不一、位置分散、关联复杂、语言繁多
  - 用户需求的表达和理解非常困难
  - 信息的理解非常困难—自然语言文本、图片、视频

信息检索是研究如何解决上述问题的一门学科！



信息检索(Information Retrieval) (1)

- **Information Retrieval这个术语产生于Calvin Mooers 1948年在MIT的硕士论文。**
- **Information Retrieval(IR)：从文档集合中返回满足用户需求的相关信息的过程。**
- **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。**

信息检索(2)

- **信息检索可以看成计算机科学(Computer Science)和图书情报学(Library & Info. Science)的交叉学科。**
  - 以计算机为手段，信息为处理对象
  - 和其他学科也融合：语言学、认知科学等
- **检索来自英文单词Retrieval，有些人把它翻译成获取。其本义是“获得与输入要求相匹配的输出”。**

信息检索(3)

- **和我们平时所理解的搜索意义上的检索不一样：IR不仅仅是搜索，IR系统也不仅仅是搜索引擎。**
  - 例1：返回与“信息检索”相关的网页 -> 搜索引擎(Search Engine, SE)
  - 例2：毛主席的生日是哪天？ -> 问答系统(Question Answering, QA)
  - 例3：返回联想PC的型号、配置、价格等信息 -> 信息抽取(Information Extraction, IE)
  - 例4：订阅有关NBA的新闻 -> 信息过滤(Information Filtering)、信息推荐(Information Recommending)
- **也可以这样说，狭义的IR通常是指Information Search，而广义的IR包含非常多的内容(SE, QA, IE, …)。**
- **本课程介绍的是广义的IR。**

信息检索的基本概念(1)

- **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“NBA总决赛”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“NBA总决赛”，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“2014NBA总决赛”。
  - Query在IR系统中往往还有内部表示。

信息检索的基本概念(2)

- **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
  - 可以是无结构、半结构化、结构化的
- **文档集合(Collection)：所有待检索的文档构成的集合**
  - 也称为Repository，Corpus

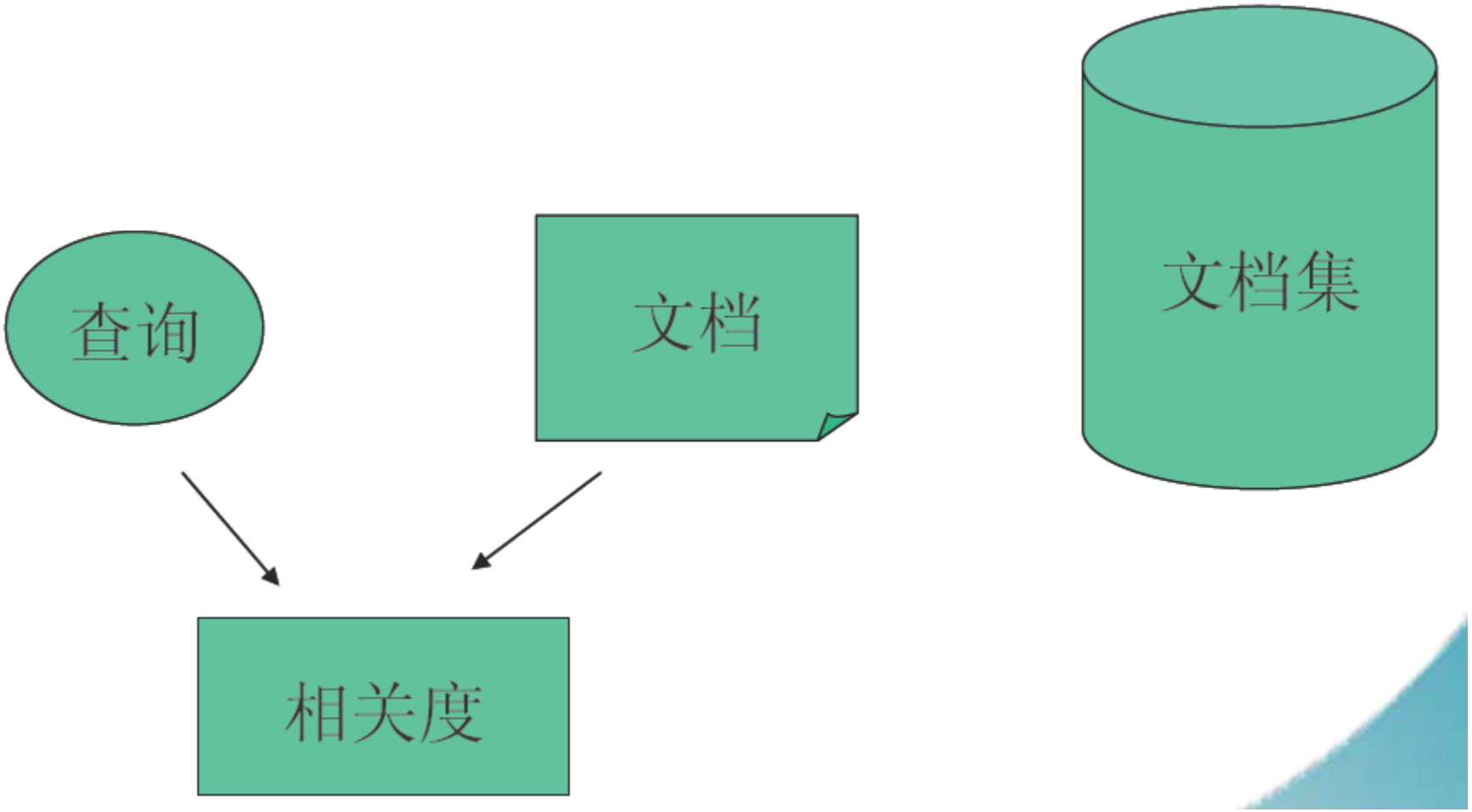
信息检索的基本概念(3)

- **相关(relevant、相关度relevance)**
  - 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

信息检索的基本概念(3)

- 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，基于这种理解，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

信息检索的基本概念(4)



信息检索的基本概念(5)

- 形式上说，信息检索中的相关度是一个函数**R**，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=f(Q, D, C)
- 信息检索就是给定一个查询**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”。
  - Google中据说用了上百种排名因子

信息检索和数据库检索

	信息检索	数据库检索
检索对象	无结构、半结构数据 如网页、图片……	结构化数据 如：员工数据库
检索方式	通常是近似检索 如：每个结果有相关度得分	通常是精确检索 如：姓名==“李明”
检索语言	主要是自然语言 如：查与超女相关的新闻	SQL结构化语言

提纲

- 信息检索的基本概念
- 信息检索的历史
- 信息检索和其他学科的关系
- 信息检索的基本流程

历史分段

- 计算机出现以前
- 计算机出现以后
- Internet出现以后

IR历史(1)

- **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。

IR历史(2)

- **计算机出现之后**
- **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SparckJones是概率模型的倡导者。

IR历史(3)

- **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

IR历史(4)

- **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。

IR历史(5)

- **1990年代的其他重要事件:**
  - 评测会议
    - NIST TREC
  - 推荐系统的出现
    - Ringo
    - Amazon
    - NetPerceptions
  - 文本分类和聚类的使用

IR历史(6)

- **2000's**
  - 信息抽取
    - Whizbang
    - Fetch
    - Burning Glass
  - 问答系统
    - TREC Q/A track
  - 2001年，百度成立

IR历史(7)

- **2000以来的其他重要事件：**
  - 多媒体IR
    - Image
    - Video
    - Audio and music
  - 跨语言IR
    - DARPA Tides
  - 文本摘要
    - DUC评测

提纲

- 信息检索的基本概念
- 信息检索的历史
- 信息检索和其他学科的关系
- 信息检索的基本流程

相关研究领域

- 图书情报学(Library & Info. Science)
- 数据库管理(Database Management)
- 人工智能(Artificial Intelligence)
- 自然语言处理(Natural Language Processing)
- 机器学习(Machine Learning)

图书情报学(Library and Information Science, LIS)

- **IR最初起源于LIS**
- **LIS主要关注IR中的用户方(人机交互、用户界面、可视化)**
- **LIS关注人类知识的高效分类**
- **LIS关注文献的引用分析(citation analysis)和文献计量(bibliometrics)**
- **近年来数字图书馆方面的工作使得LIS和IR日益融合。**

数据库管理系统(Database Management, DM)

- **DM主要面向关系表中的结构化数据而非自由文本。**
- **DM主要集中于高效解决形式化语言(如SQL)定义的查询。**
- **DM中不论是查询还是数据都具有明确的语义。**
- **近年来半结构化的XML数据的出现使DM和IR逐渐融合。**

人工智能(Artificial Intelligence, AI)

- AI关注知识的表示、推理和智能行为。
- AI中知识的形式化表示
  - 一阶谓词逻辑(First Order Predicate Logic)
  - 贝叶斯网络(Bayesian Networks)
- 近年来Web本体及智能信息Agent方面研究使得IR和AI相互融合。

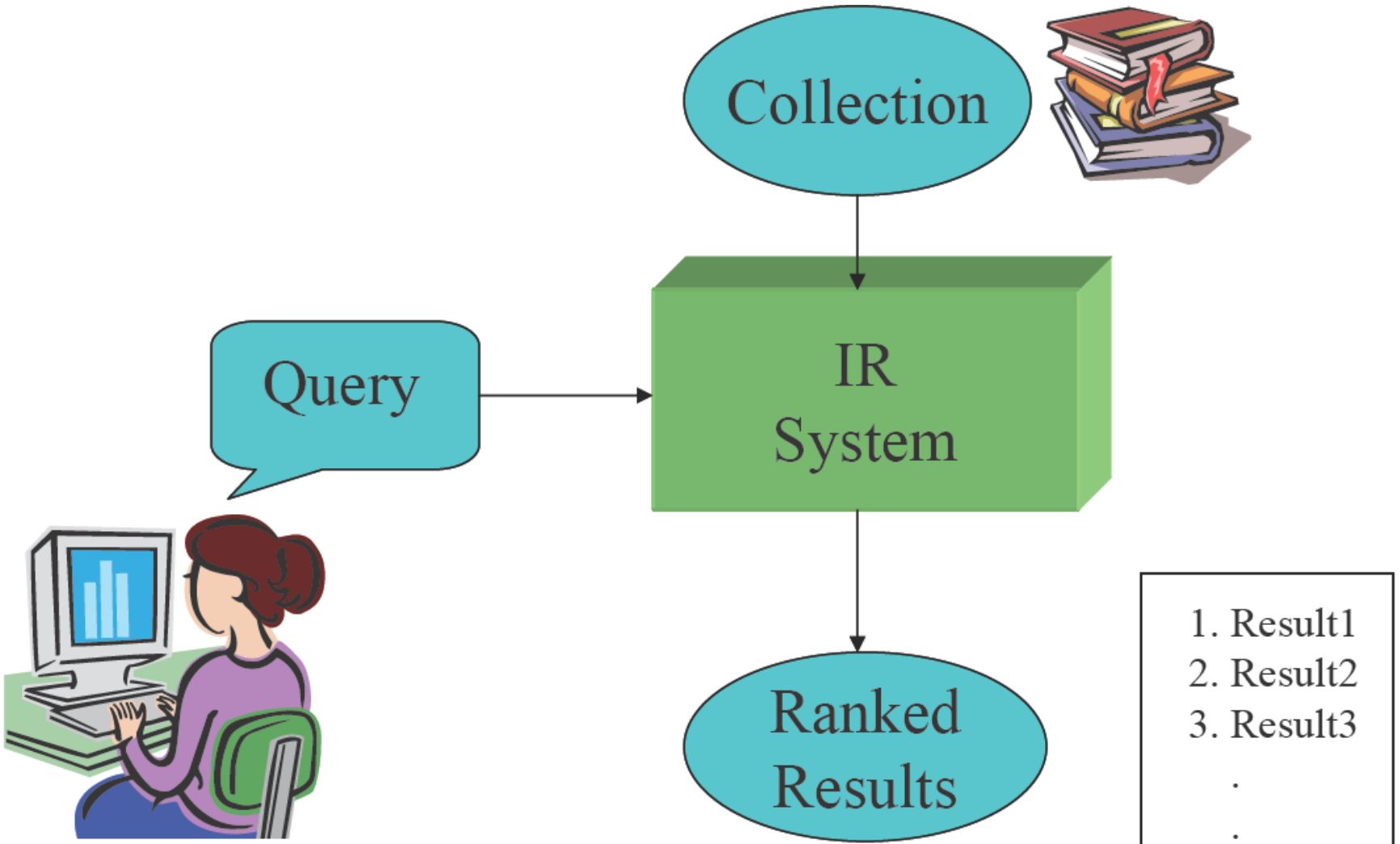
自然语言理解(Natural Language Processing,NLP)

- **NLP关注自然语言文本的语法(syntactic)、语义(semantic)及语用(pragmatic)分析。**
- **NLP可以分析短语结构和语义，使得IR可以在短语上、或者从语义上进行处理，而不是仅仅基于单个关键词。**
- **NLP和IR天生就是融合的。**

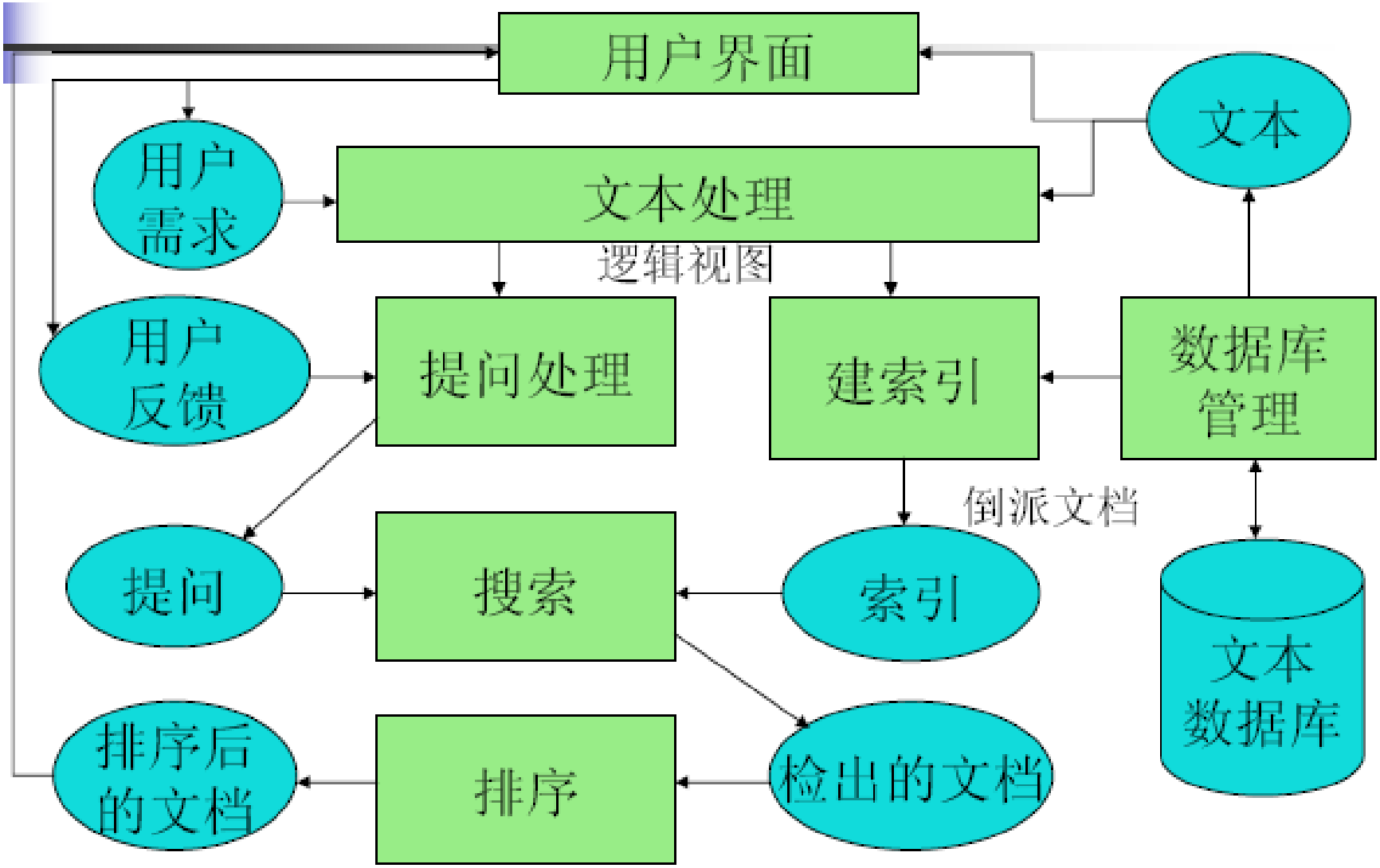
提纲

- 信息检索的基本概念
- 信息检索的历史
- 信息检索和其他学科的关系
- 信息检索的基本流程

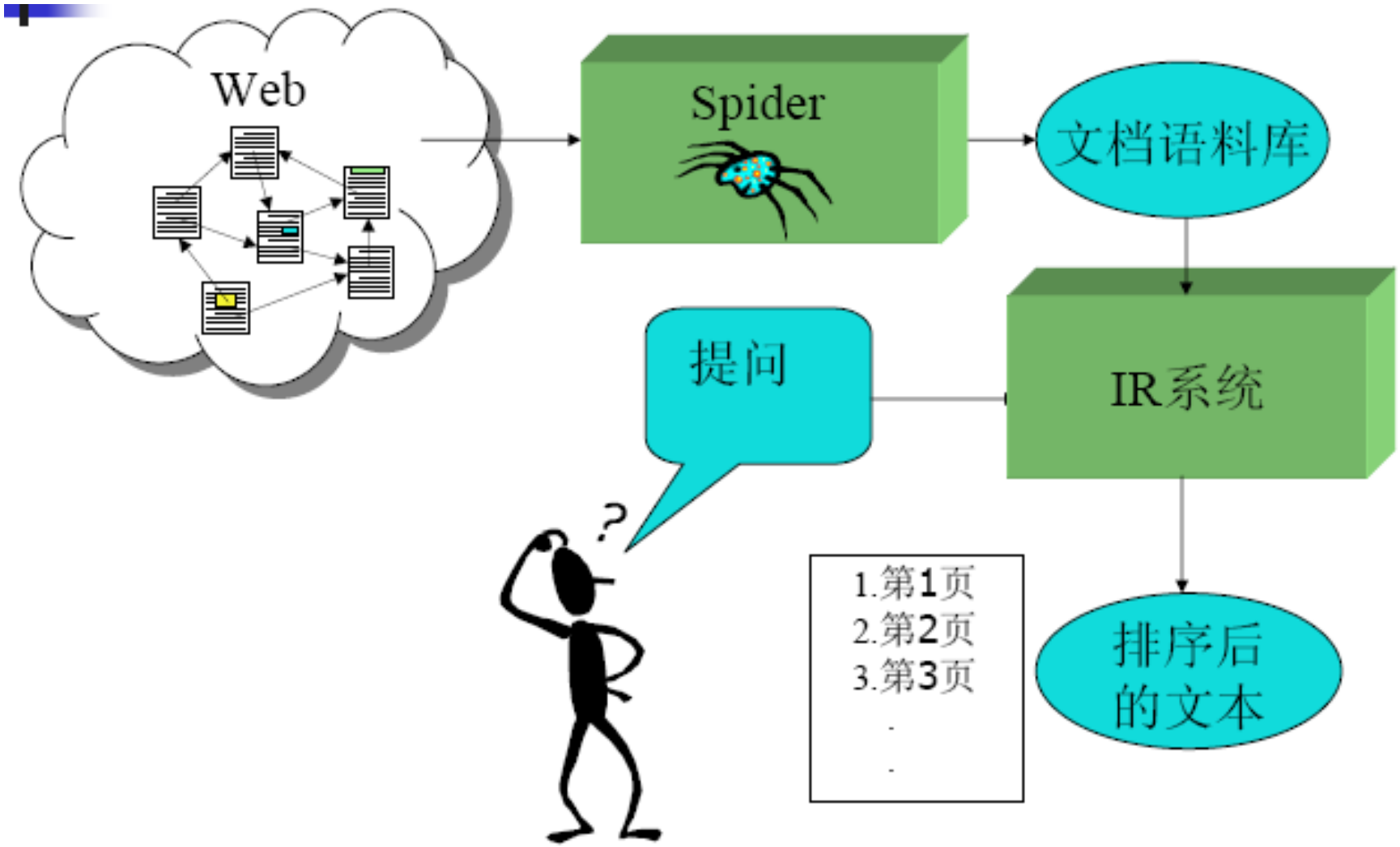
IR系统示意图


Collection
Query
IR
System
Ranked
Results
1. Result1
2. Result2
3. Result3
.
.

IR系统的组成框架



Web搜索引擎



IR系统的组成部分(1)

- 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: 用户是主动的发起请求，在一个相对稳定的数据集合上进行查询
  - 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

IR系统的组成部分(2)

- **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(Stopwordremoval)
- **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
- **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)，通常基于标引项(Term)来表示**
  - 向量化、概率计算
  - 组成倒排表进行存储

IR系统的组成部分(3)

- **搜索(Searching)：从文本中查找包含查询的文本**
- **排序(Ranking)：对搜索出的文本按照某种方式来计算其相关度**
- **Logical View：指的是查询或者文本的表示，通常采用一些关键词或者标引项(index term)来表示一段查询或者文本。**

本章小结

- 信息检索是一门交叉学科，不仅仅是搜索
- 信息检索中的用户需求、查询、文档、文档集、相关度概念
- 信息检索和其他学科领域的关系
- 信息检索的组成和流程

课后思考题

- 信息检索的定义是什么？请列举几种信息检索的应用。
- 信息检索的基本流程如何？各组成部分的功能是什么？

The End