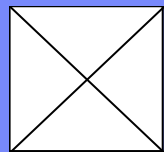
Advanced web technology

Web高级开发与应用技术

Web语义与搜索



Web 3.0

- 什么是Web3.0

“My prediction would be that Web 3.0 will ultimately be seen as applications which are pieced together.”

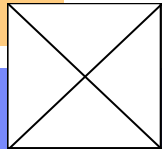
— Eric Schmidt

Google’s CEO

“…….distinction between professional, semi-professional and consumers will get blurred, creating a network effect of business and applications.” — *Jerry Yang* Yahoo founder,

think maybe when you've got an overlay of scalable vector graphics - everything rippling and folding and looking misty – on Web 2.0 and access to a semantic Web integrated across a huge space of data, you'll have access to an unbelievable data resource.“

——World Wide Web 创始人蒂姆-伯纳斯-李



Web 3.0

- 什么是Web3.0

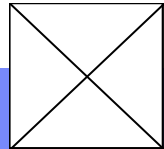
其实很简单Web3.0将是语义网的天下

----Qiantu.org

“Referred to as Web 3.0, the effort is in its infancy, and the very idea has given rise to skeptics who have called it an unobtainable vision…”

----纽约时报

“Web 3.0 will be 10 megabits of bandwidth all the time, which will be the full video Web, and that will feel like Web 3.0” — *Reed Hastings* Netflix founder



Web 3.0

- 什么是Web3.0

从web2.0到web3.0则会通过更加个性化的技术革新使得互联网的表现形式更为丰富。例如3D(三维)、 3G等新技术在互联网的运用

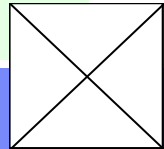
——阿里巴巴软件经理王涛

“web3.0，就是让个人和机构之间建立一种互为中心而转化的机制，也就是说个人在一定程度上可以转化为机构，机构在一定的环境下也可以像个人一样，拟人化的进行他们的商业行为，而进一步拉近和网民的距离……”

------Mezi.Bulunbulei博士

博客会是搜狐web3.0中相当重要的一个元素，也是网民的一个主要入口。而这个全新的“声•色”版博客增加了视频功能，将全面支持视频内容的上传和分享，让用户把视频、音乐、图片、文字随意支配于掌上

——搜狐CEO张朝阳



Web 3.0

- **Web3.0相关技术**

"Web 3.0: A Vision for Bridging the Gap between Real and Virtual"

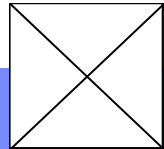
  - **Semantic Web**
    - machines can read it and understand it as much as humans can, without ambiguousness.
    - The first challenge is the effort to link existing content to semantic meaning by using some sort of metadata.
    - The second challenge is to develop a set of applications that make use of this newly generated metadata-based knowledge.

Radar Networks

Garlik for personal data management on the Web

Yahoo Food Site

Joost Internet TV Platform



Web 3.0

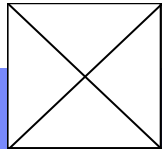
- **Web3.0相关技术**
  - **The 3D Web or Web 3D**
    - the interactions occurring between avatars are kept in the virtual world
    - A new interaction dimension that can be incorporated with online social virtual worlds is the sense of touch, or haptics

Second Life

IMVU

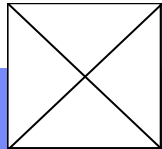
Active Worlds

Red Light Center



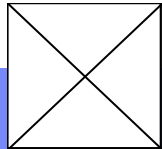
Web 3.0

- **Web3.0相关技术**
  - **The Media Centric Web**
    - Future search engines should take media as input and be able to search for similar media objects based on its features and not only based on textual metadata.(Ojos Riya photo sharing tool or Like.com)
    - Systems should be able to recognize hand gestures, voice, and even people's faces and moods, and respond in a multimodal fashion as well



Web 3.0

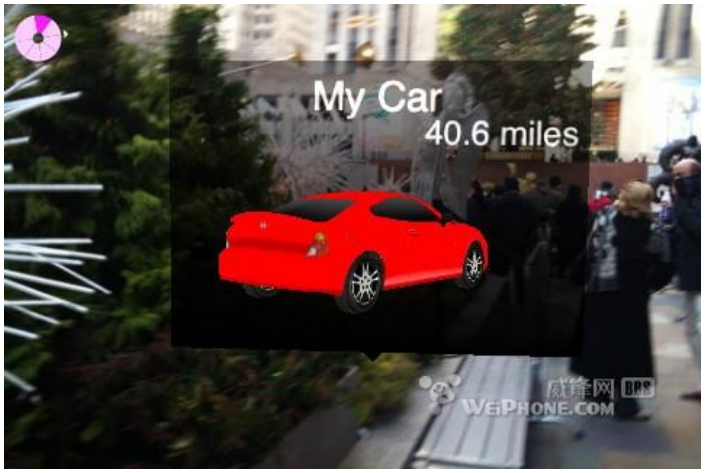
- **Web3.0相关技术**
  - **The Social Web**
    - The social web components have been put in place since web 2.0, this is a field that will evolve along with the Web itself
    - Examples: IMVU
  - **The Pervasive and Ubiquitous Web**
    - we envision the Web3.0 to go beyond the use of the traditional web by including natural ways of interacting with real-life objects that typically have not been considered as computing entities



Web 3.0

- **The Pervasive and Ubiquitous computing**

google glass



Web 3.0

- **Web3.0相关技术**
  - **使用基于XML的精确的描述**
    - RDF(Resource Description Framework)
    - OWL(Web Ontology Languages)
  - **Microformats(微缩格式）:**将基本的语义学埋置入HTML
    - XFN:反映互联网上人与人之间的关系
    - hCard 可以注释HTML : 解决了一个个人信息的问题
    - hCalendar，它允许页面作者自己去描述事件

Wikipedia 定义语义网 :"a project that intends to create a universal medium for information exchange by putting documents with computer-processable meaning (semantics) on the World Wide Web"

Web 3.0

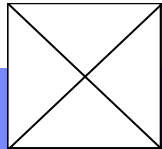
- **Microformats(微缩格式）**
  - 给没有实际意义的tag添加富有实际意义的属性（attribute）内容
  - 将信息标准化的目的是为了聚合

```
<a href="http://creativecommons.org/licenses/by-nc-sa/2.5/cn/">遵守我的版权</a>
```

```
<a href="http://creativecommons.org/licenses/by-nc-sa/2.5/cn/" rel="license">遵守我的版权</a>
```

  - XFN示例

```
<a href=“http://www.machenlei.com/” rel=“met colleague”>老冯</a>
```



Web 3.0

阿里妈妈(alimama)

智慧园（witpark.com）

Web 3.0

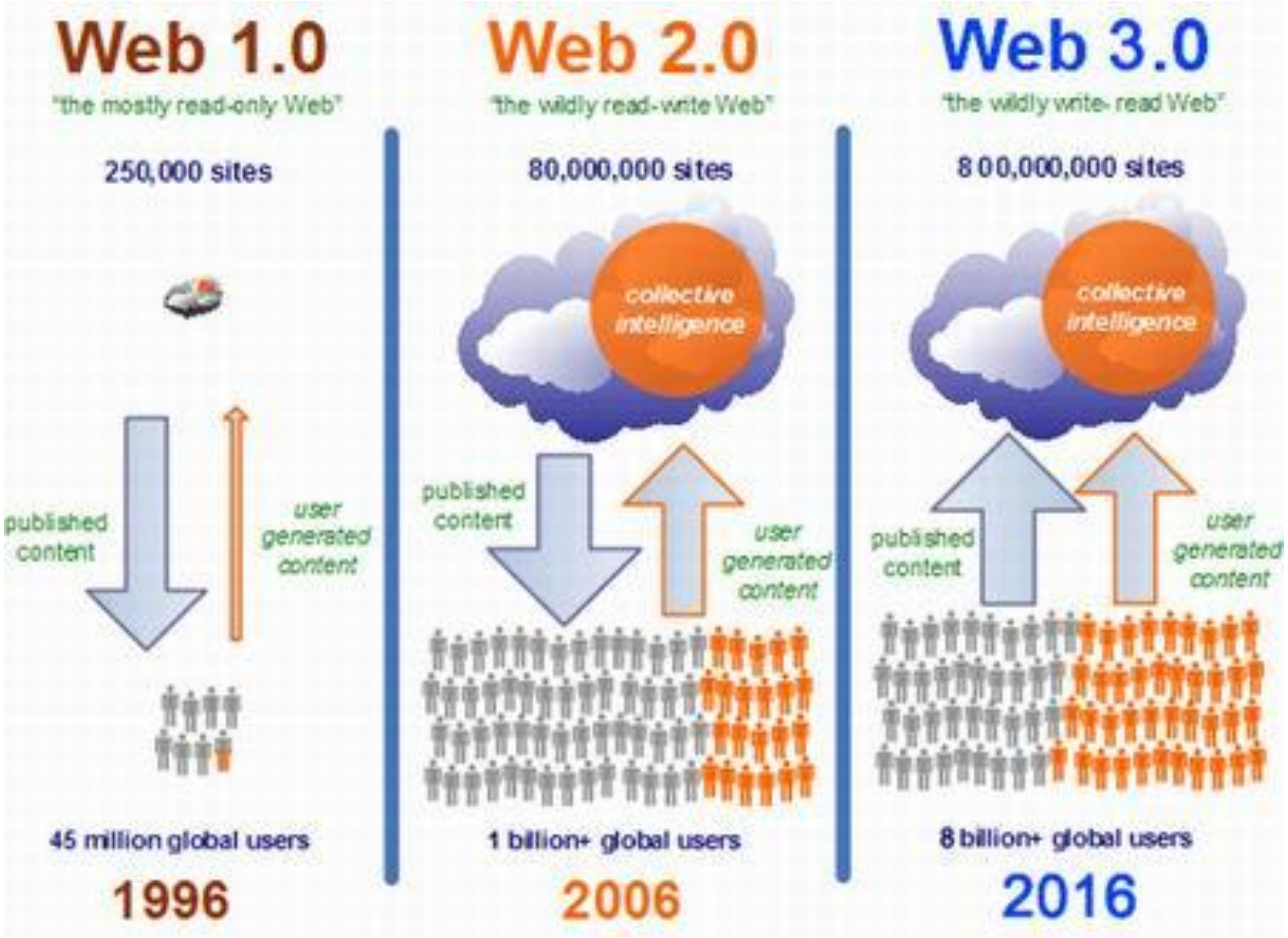
Web 1.0: Centralized Them.

Web 2.0: Distributed Us.

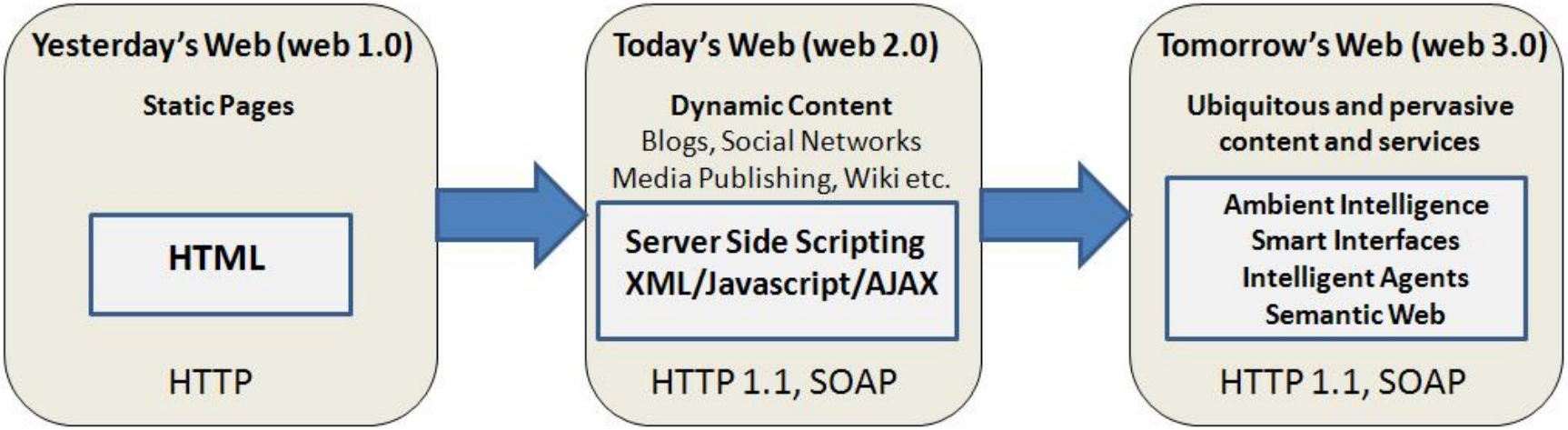
Web 3.0: Decentralized Me

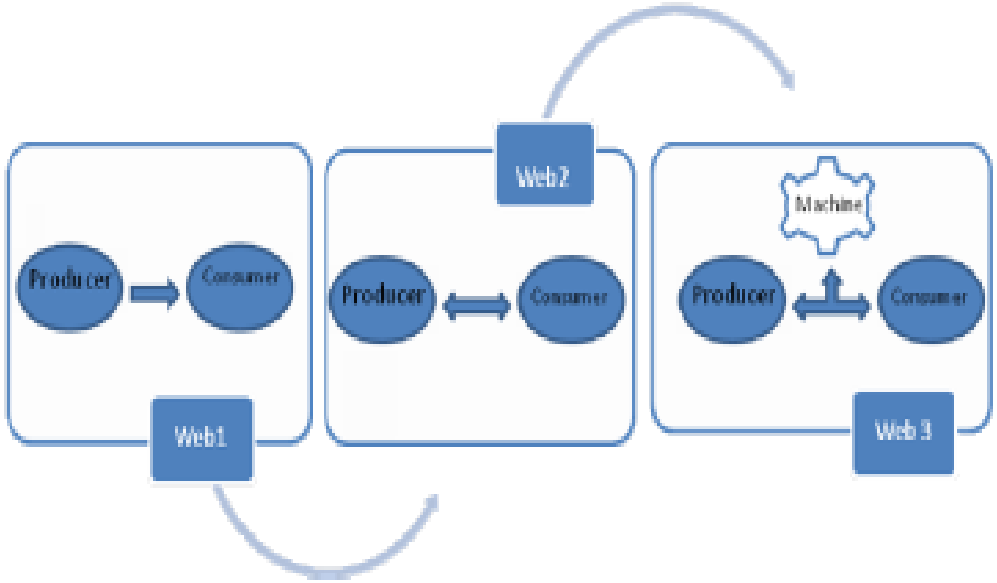
http://www.readwriteweb.com/archives/web_30_is_it_about_personalization.php

Web 3.0


Web 1.0
"the mostly read-only Web"
250,000 sites
published content
user generated content
45 million global users
1996
Web 2.0
"the wildly read-write Web"
80,000,000 sites
collective intelligence
published content
user generated content
1 billion+ global users
2006
Web 3.0
"the wildly write- read Web"
800,000,000 sites
collective intelligence
published content
user generated content
8 billion+ global users
2016

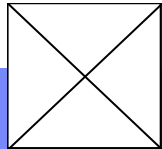
Web 3.0


Yesterday's Web (web 1.0)
Static Pages
HTML
HTTP
Today's Web (web 2.0)
Dynamic Content
Blogs, Social Networks
Media Publishing, Wiki etc.
Server Side Scripting
XML/Javascript/AJAX
HTTP 1.1, SOAP
Tomorrow's Web (web 3.0)
Ubiquitous and pervasive content and services
Ambient Intelligence
Smart Interfaces
Intelligent Agents
Semantic Web
HTTP 1.1, SOAP

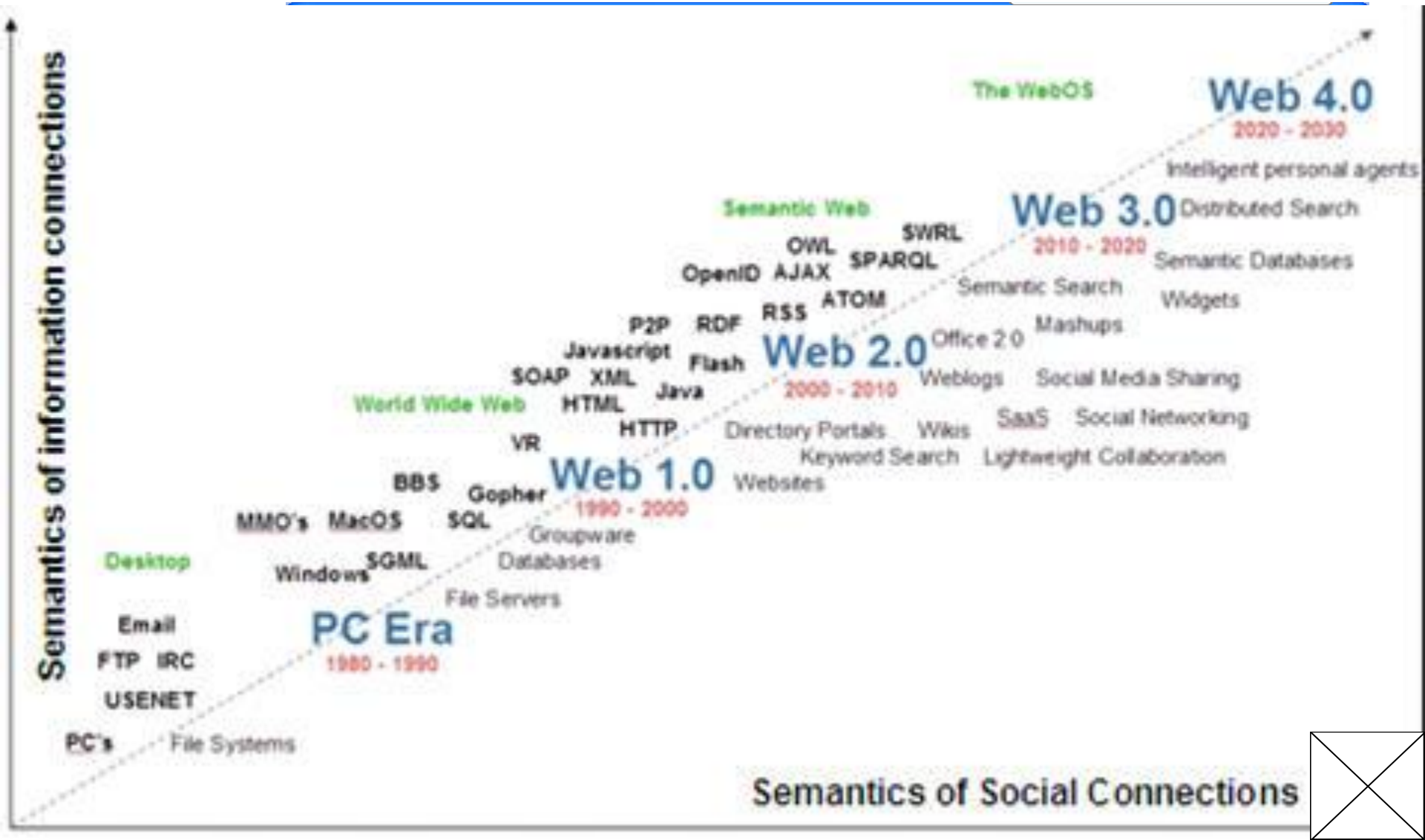

Producer
Consumer
Web1
Web2
Producer
Consumer
Machine
Producer
Consumer
Web 3

Web 3.0

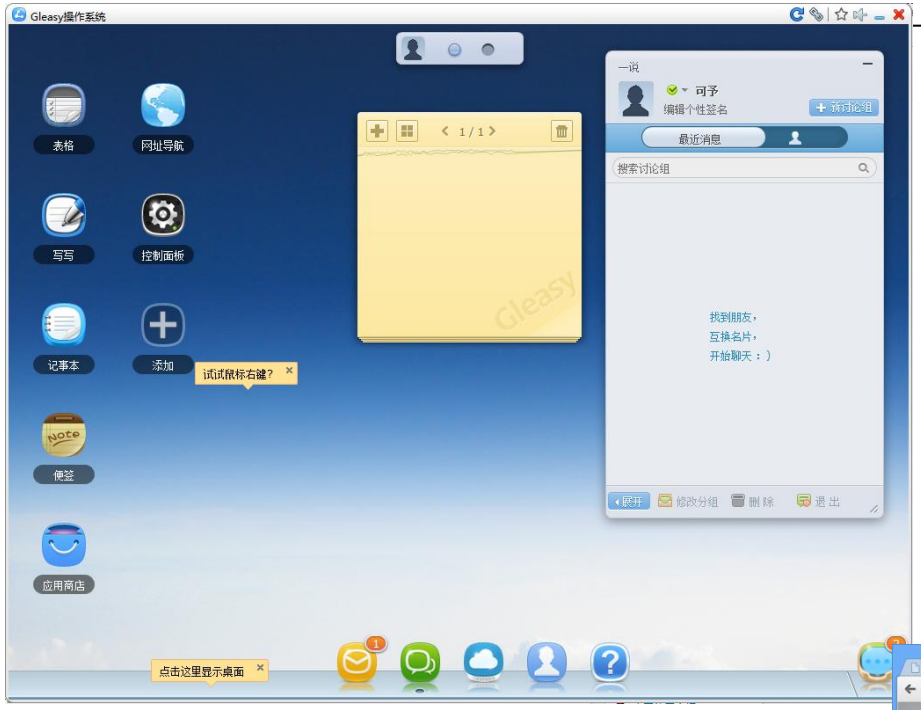
Web 1.0	Web 2.0	We b3.0
FrontPage	MySpace	SIOC-project.org
Encarta	Wikipedia	Dbpedia
Streetmap/MapQuest	Google earth	3-D Street View
PC games	Online games	Online 3D-games
Home video	YouTube	Yet to come
Mp3.com	iTunes	Yet to come
Microsoft Office	Google Docs	Yet to come

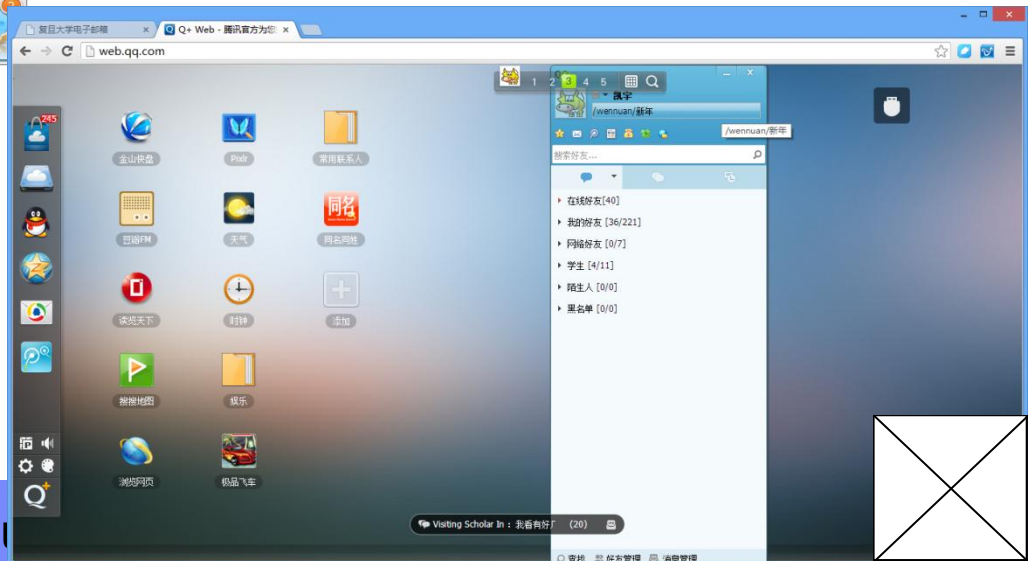


From semantic Web (3.0) to the WebOS (4.0)



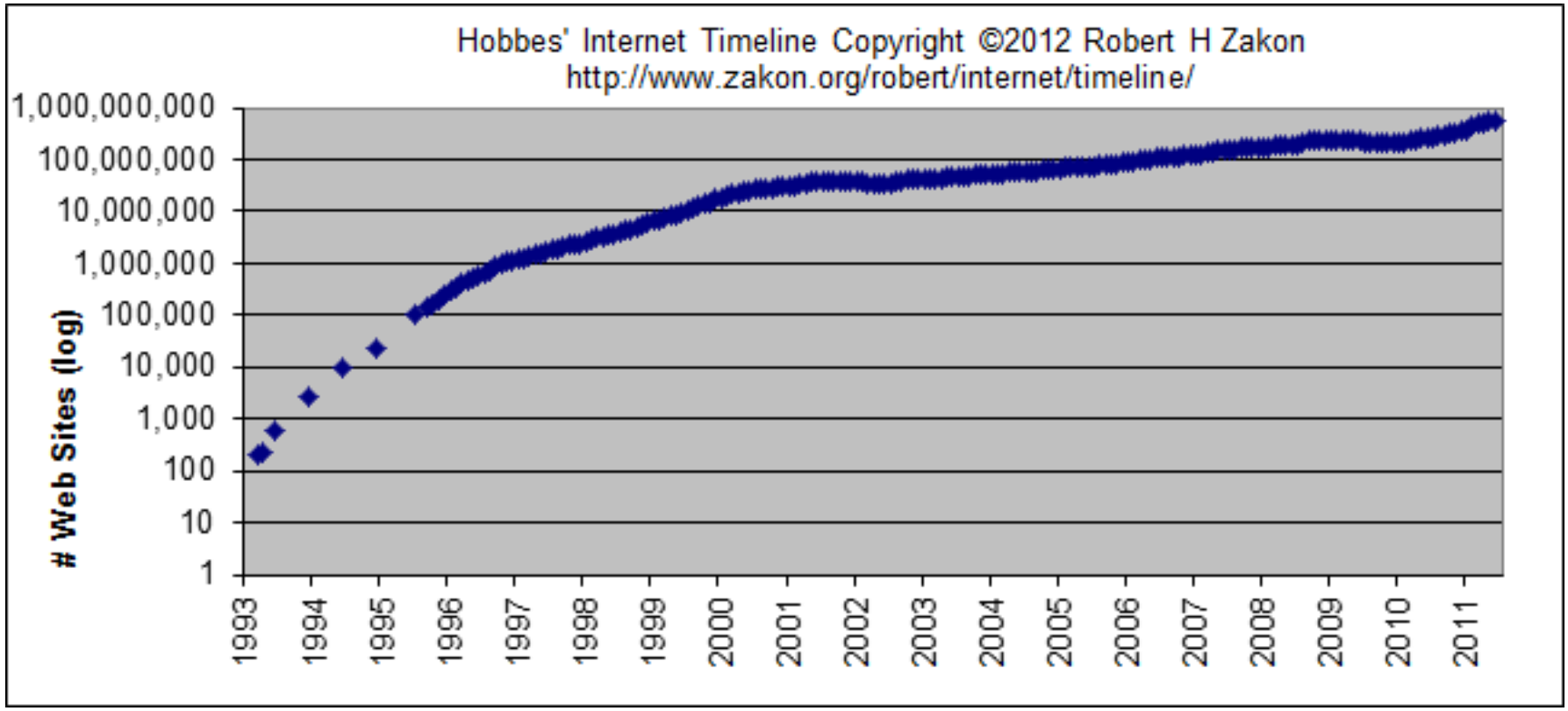
WebOS 示例


Gleasy操作系统
表格
网址导航
写写
控制面板
记事本
添加
试试鼠标右键？
便签
应用商店
点击这里显示桌面
一说
可予
编辑个性签名
最近消息
搜索讨论组
找到朋友，
互换名片，
开始聊天：）
展开
修改分组
删除
退出

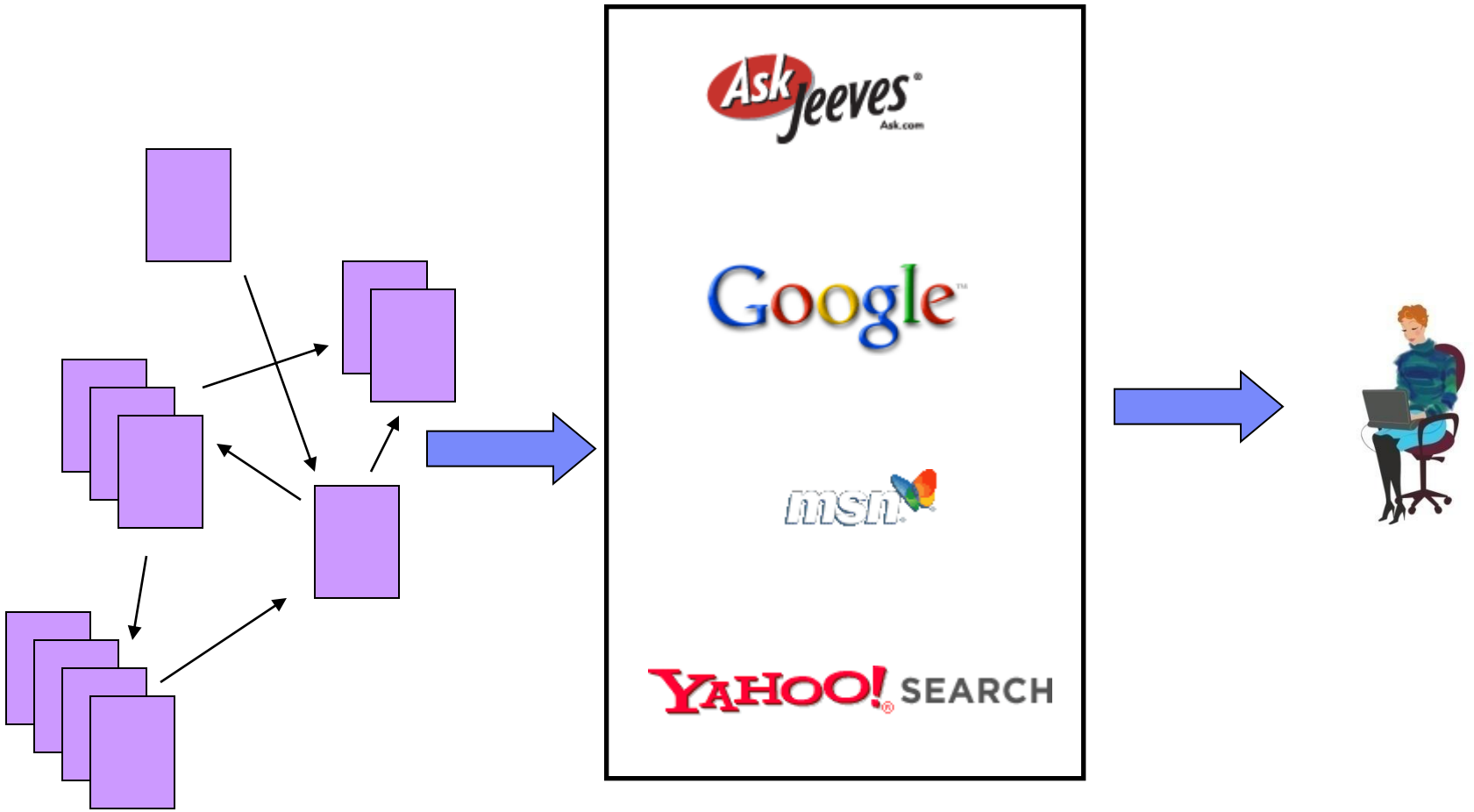

web.qq.com
在线好友[40]
我的好友 [36/221]
网络好友 [0/7]
学生 [4/11]
陌生人 [0/0]
黑名单 [0/0]

搜索引擎简介

Web站点的增长



搜索引擎简介

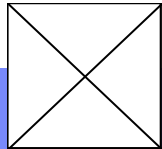

Ask Jeeves
Ask.com
Google
msn
YAHOO! SEARCH
The Web
Content aggregators
Content consumers

Web 搜索引擎历史

- 1993, 早期的 web robots (spiders) 用于收集 URL:
  - Wanderer: Perl-based web crawler
  - ALIWEB (Archie-Like Index of the WEB)
  - WWW Worm (indexed URL's and titles for regex search)
- 1994, Stanford 博士生 David Filo and Jerry Yang 开发手工划分主题层次的雅虎网站.

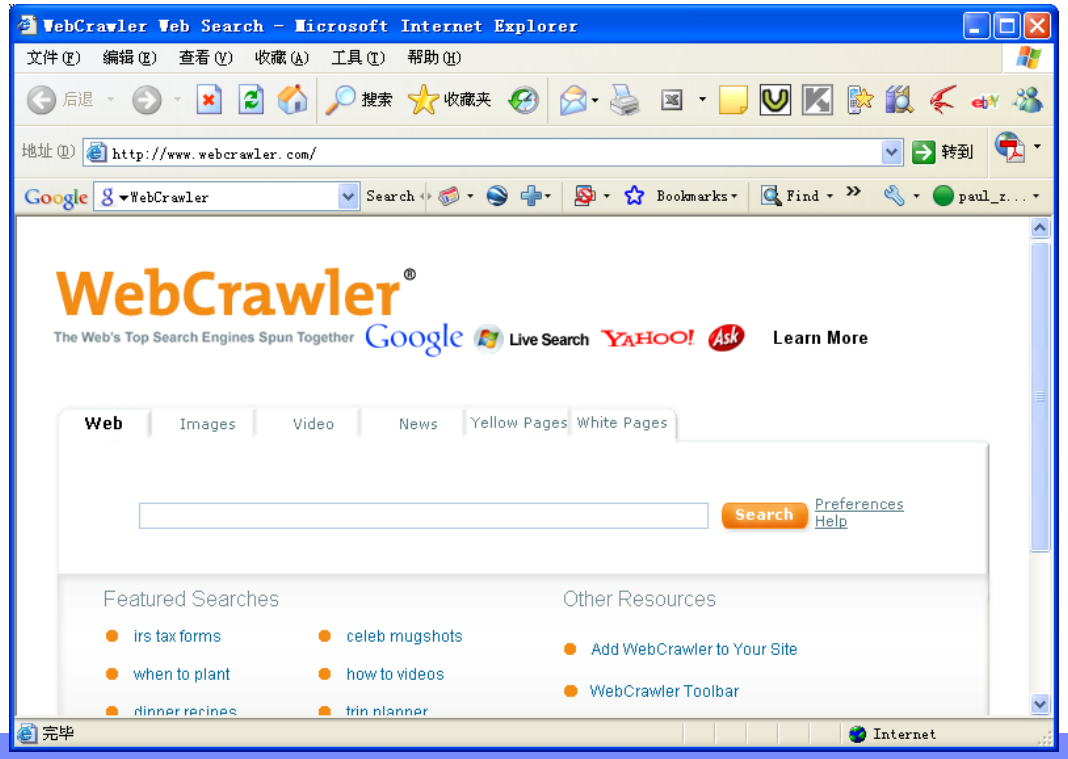
Web 搜索引擎历史

- 1994年初，WebCrawler是互联网上第一个支持搜索文件全部文字的全文搜索引擎
- Lycos（Carnegie Mellon University Center for Machine Translation Announces Lycos ）具有相关性排序，还提供了前缀匹配和字符相近限制，第一个在搜索结果中使用了网页自动摘要，远胜过其它搜索引擎的数据量
- 1995年12月,DEC的AltaVista第一个支持自然语言搜索的搜索引擎，实现高级搜索语法（如AND, OR, NOT等)
- 1997年9月15日博士生Larry Page注册了google.com的域名，Google在Pagerank、动态摘要、网页快照、多文档格式支持、地图股票词典寻人等集成搜索、多语言支持、用户界面等功能上具有革新;尤其是应用链接分析根据权威性对部分结果排序 。



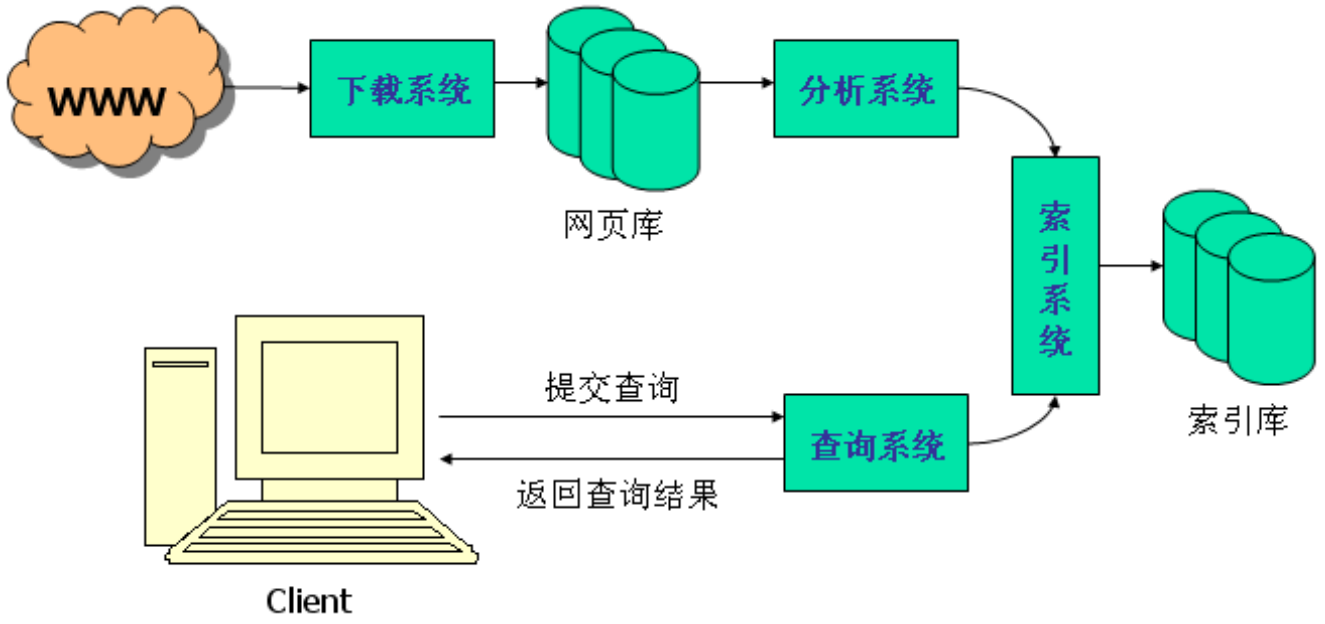
搜索引擎简介

- 分类
  - 目录式搜索引擎：早期的Yahoo
  - 全文搜索引擎：Google及百度等第二代商用搜索引擎
  - 元搜索引擎：WebCrawler



搜索引擎简介

- **搜索引擎的体系结构**
  - **下载系统：网络蜘蛛(Spider)，广度，深度优先**
  - **分析系统：分词，PageRank**
  - **索引系统：正排索引；倒排索引**
  - **查询系统：检索模型**

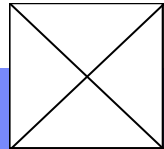


搜索引擎简介

- **分词**
  - 基于字符串匹配
  - 基于理解
  - 基于统计
- **PageRank**

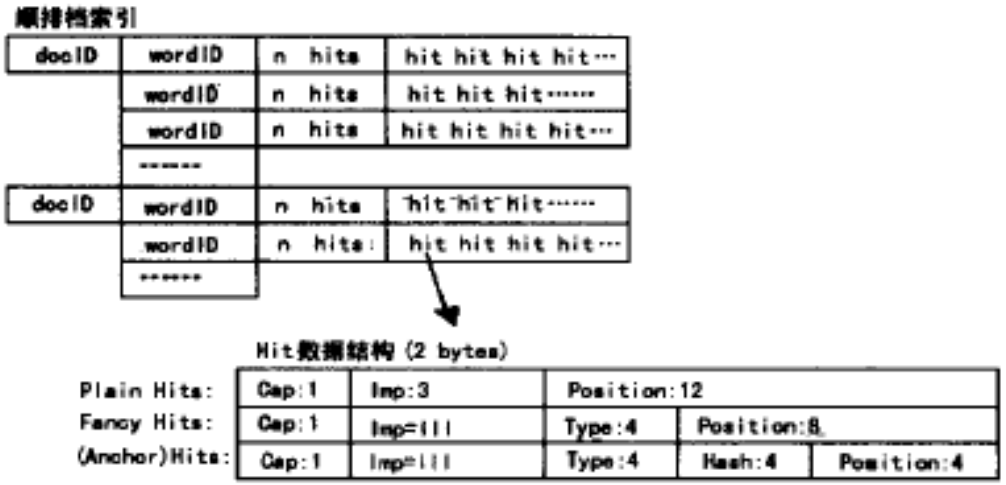
$$PR(A) = (1-d) + d(PR(T1)/C(T1) + \cdots\cdots + PR(Tn)/C(Tn))$$

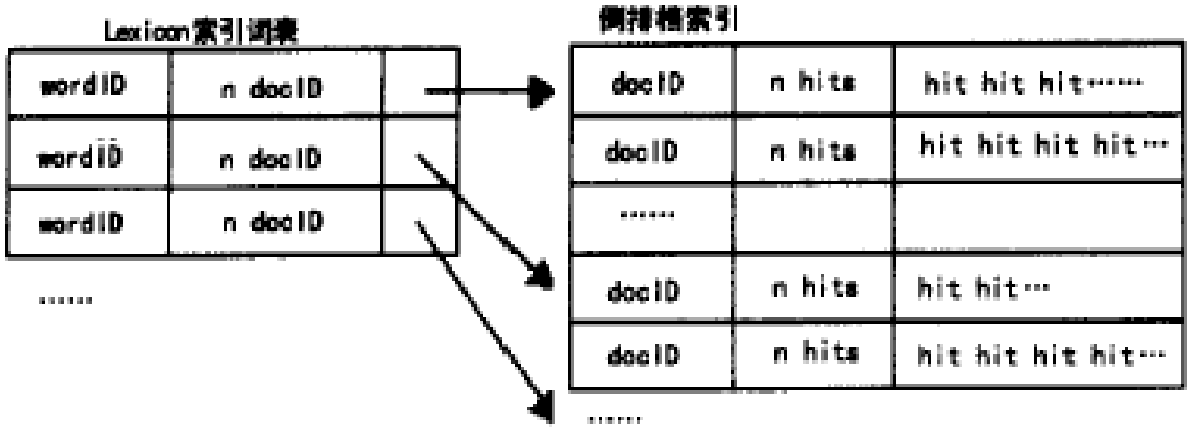
- **PR (A)**：网页A 的PageRank 值；
- **T1, T2, …, Tn**：网页A 的链入网页；
- **PR (T i)**：网页T i 的PageRank 值(i= 1, 2, …n) ；
- **C (T i)**：网页T i 的链出网页的数量(i= 1, 2, …n) ；
- **d**：一个衰减因子, 0< d< 1, 通常取值为0. 85。



搜索引擎简介

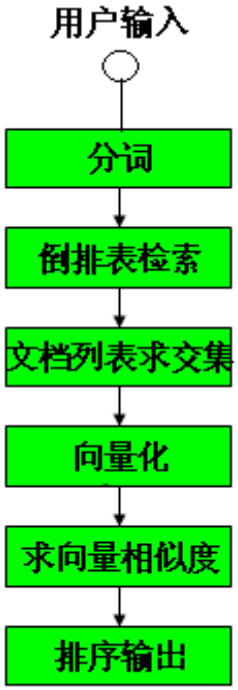
- 索引系统
  - 正排索引
  - 倒排索引





搜索引擎简介

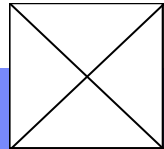
- 查询系统



- 搜索引擎的评价标准
  - 查全率（Recall）
  - 查准率（Precision）

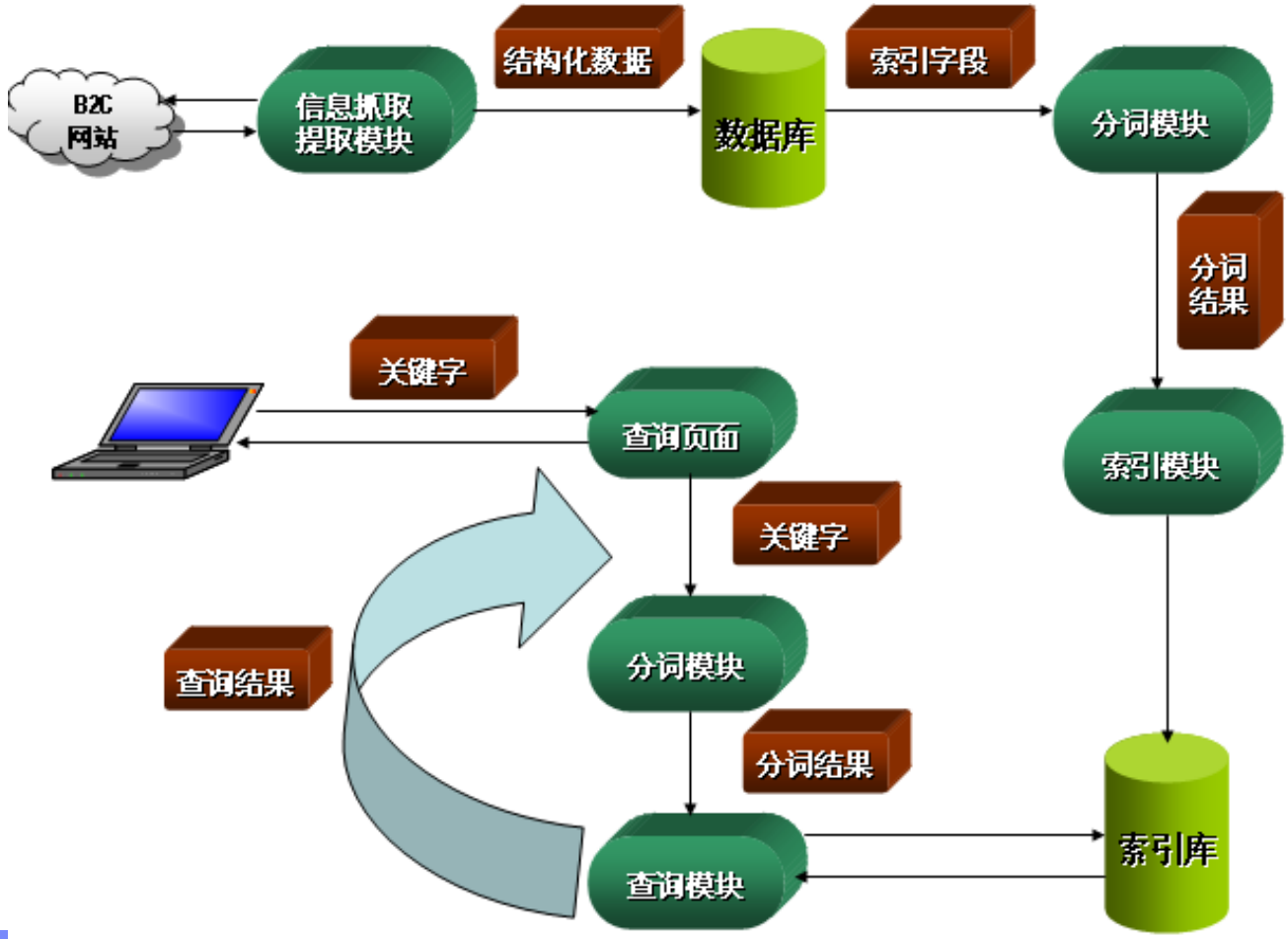
$$recall = \frac{\text{number of retrieved relevant documents}}{\text{total number of relevant documents}}$$

$$precision = \frac{\text{number of retrieved relevant documents}}{\text{total number of retrieved documents}}$$



搜索引擎简介

- 垂直搜索引擎



搜索引擎简介

- 智能搜索引擎?

“这是Web 将以一种智能的方式为我们服务、为我们完成乏味任务的时代的开始。Web 和信息量的增长速度非常快，开发智能化的搜索系统是势在必行的。”

— Medstory的创始人、首席执行官阿莱因

20岁获得物理学博士学位，麦克·阿瑟“天才人物”奖获得者，mathmatic 软件的开发者，美国的传奇式计算机科学家Stephen Wolfram近日在其博客中称研发了一个名为“Wolfram|Alpha”的新型搜索引擎，该引擎可能以其更加智能的搜索功能而成为谷歌搜索引擎的巨 大竞争对手。

Web挖掘简介

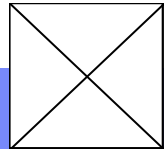
数据挖掘(Data Mining)从*大量的、不完全的、有噪声的、模糊的、随机的*实际应用数据中，提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程。

- 传统的数据挖掘在结构化数据上进行
  - 关系型表格
  - 电子表格
  - 以表格形式存储的纯文本

- 随着WWW和文本文件规模的不断增大，Web挖掘和文本挖掘变得越来越重要

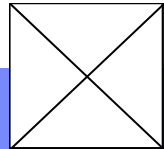
Web挖掘简介

- Web数据挖掘的目标:从Web上寻找有价值的信息
  - Hypertext
  - 网页内容
  - 使用日志
- Web挖掘的分类
  - Web结构挖掘
  - Web内容挖掘
  - Web使用挖掘



Web结构挖掘

- 从WWW 上的组织结构和链接关系中推导知识
  - 超文本文档间的关联关系使得WWW 不仅仅可以揭示文档中所包含的信息, 同时也可以揭示文档间的关联关系所代表的信息
  - 利用这些信息可以对页面进行排序, 发现重要的页面
  - 挖掘Web 结构的目的是发现页面的结构和Web结构, 在此基础上对页面进行分类和聚类, 从而找到权威页面



Web内容挖掘

- **直接挖掘文档内容(Web content extraction)**

Given a web page *S* .

Determine a mapping *W* that populates a data object *R* from the objects in *S*.

W must also suitable for similar web page.

W4F(w3 wrapper factory)

XWRAP

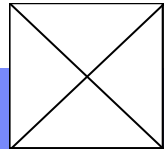
Road Runner

NLP based tools

RAPIER

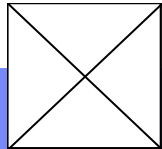
WHISK

Web-harvest



Web内容挖掘

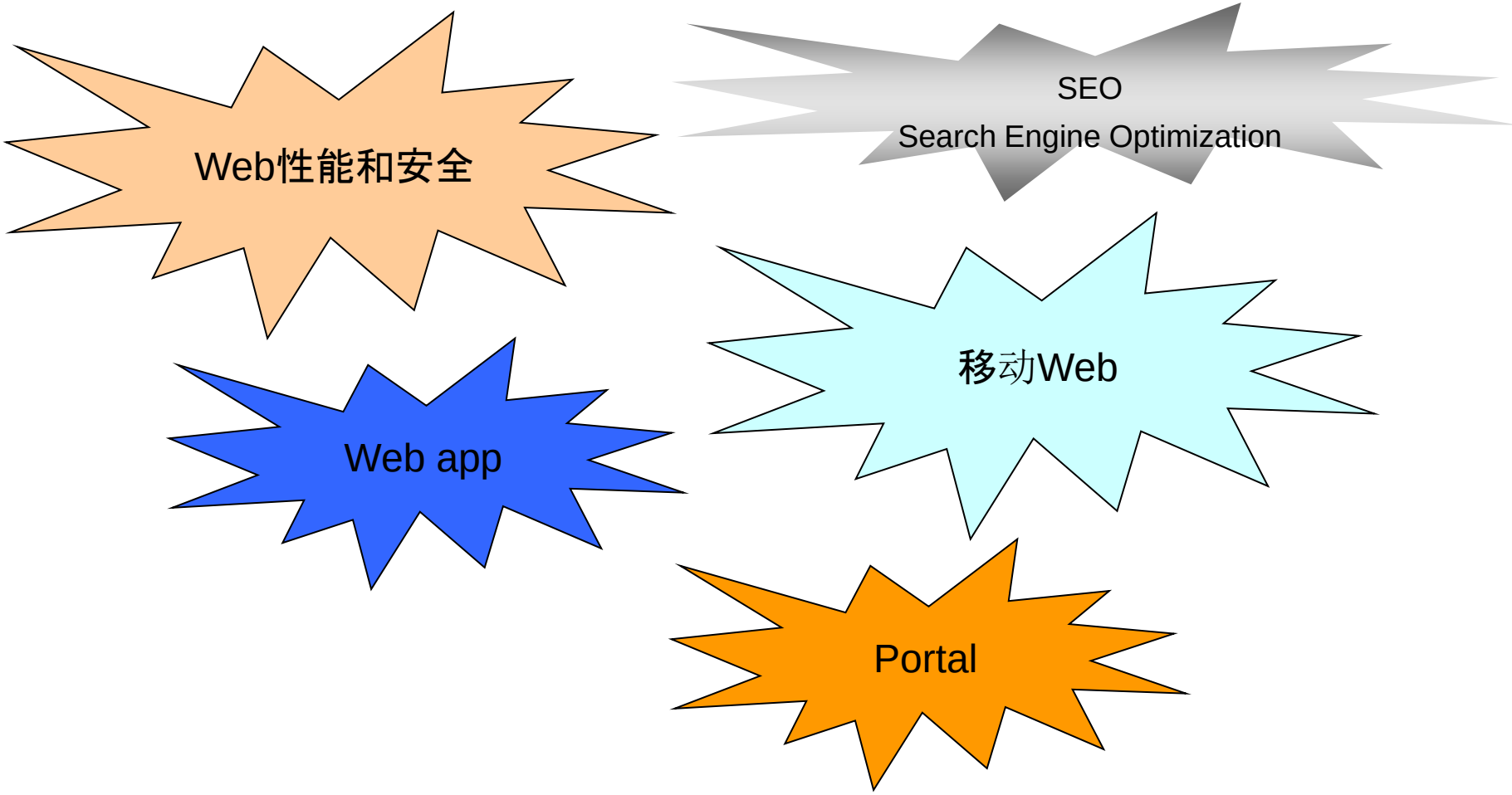
- 对搜索引擎查询结果的进一步处理
  - Web 查询语言
    - WebOQL 是一个用于Web 页重构的查询语言, 利用Web 文档的图树表示形式, 可从在线的文档站点或导游指南中获取信息
    - Ahoy 利用像搜索引擎一类的互联网服务来获取与个人有关的服务, 利用试探法识别文档中显示该文档作为个人主页的句法特征



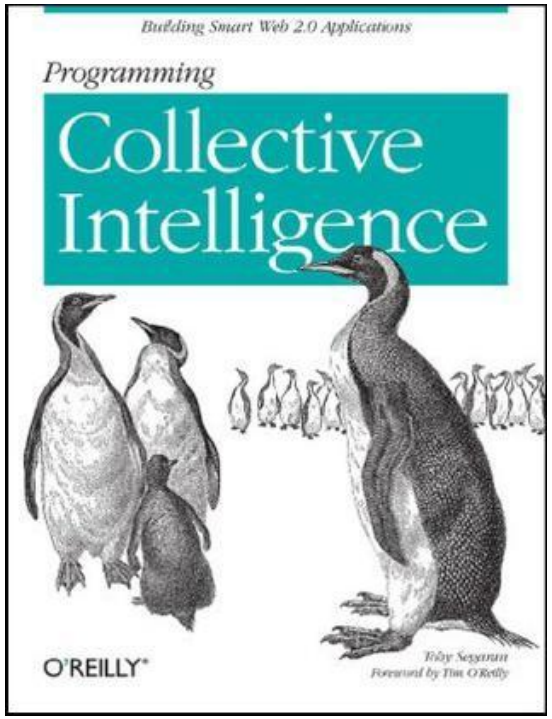
Web使用挖掘

- 主要目标是从Web的访问记录中抽取感兴趣的模式
- WWW 中的每个服务器都保留了访问日志, 记录了关于用户访问和交互
- 帮助理解用户的行为, 从而改进站点的结构,或为用户提供个性化的服务
  - 一般的访问模式追踪
    - 通过分析使用记录来了解用户的访问模式和倾向,以改进站点的组织结构。
  - 个性化的使用记录追踪
    - 倾向于分析单个用户的偏好, 其目的是根据不同用户的访问模式, 为每个用户提供定制的站点。

Still more


Web性能和安全
SEO
Search Engine Optimization
移动Web
Web app
Portal

Reference books and website



http://www.programmableweb.com