

北京邮电大学
BEIJING UNIVERSITY OF POSTS AND TELECOMMUNICATIONS

大数据时代的管理

Management in Big Data Era

马宝君 博士 讲师

经济管理学院
电子商务中心
2014年12月20日

1

上次课程小结

- 基本概念
- 分类分析的经典方法
- 预测分析的常用方法
- 分类、预测方法的评估
- 应用案例
- 总结

2

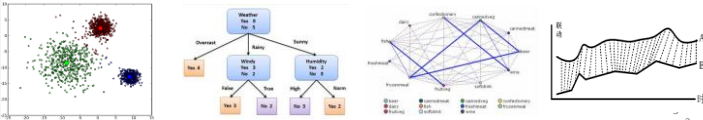
课程回顾：深度业务分析——原方法

- 聚类 (Clustering)
- 分类 (Classification)
- 关联 (Association)
- 模式 (Pattern)
-

类别

联系

轨迹



3

深度业务分析——组合方法及应用

- 信息检索及信息搜索服务 (文本内容、链接)
- 推荐系统及产品推荐
- 舆情分析及商誉构建 (情感)
- 社交网络分析及关系营销
- 用户生成内容 (口碑/评论/社交) 分析
-



4

信息检索基础


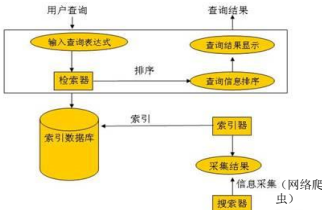
Information Retrieval Foundations



5

信息检索的基本概念

- 信息检索发源于**文档检索**技术，用以处理文档的采集、组织以及检索。近年来由于网络内容的大量增长，而网络中的信息形式很多为文本形式（如文献、新闻、博文、评论和公告等），因此信息检索技术被大量应用到网络检索中
- 信息检索的过程



倒排检索示例

6

文档的表示（基于文档内容）

- 为了便于信息检索，一般文档都需要表示成合适的形式：
 - 一般情况下，一个文档可以表示成为一系列**有权重的关键词**的集合，其中的权重是反映不同关键词相对重要性的数值

$$\begin{matrix} t_1 & t_2 & \dots & t_n \\ d: & w_1 & w_2 & \dots & w_n \end{matrix}$$

- 检索原则
 - 如果一个文档含有**越多的、越重要的查询关键词**，那么这个文档应该越可能被检索到
- 文档预处理
 - 英文：去除标点符号等、去除停用词、词根化处理（Stemming）
 - 中文：去除标点符号等、分词、去除停用词（Stop words）

7

文档向量的权重计算

- w_{ij} ：文档集合 **D** 的第 **j** 个文档 **d_j** 中的第 **i** 个关键词 **t_i** 的重要性(权重)
- 一般可以从两个角度来衡量文档中**关键词的重要性**：
 - 关键词频率（Term Frequency, tf ）：文档 **d_i** 中关键词 **t_j** 出现的次数
$$tf_{ij} = \frac{f_{ij}}{\max\{f_{1j}, f_{2j}, \dots, f_{nj}\}}$$
 - 文档频率（Document Frequency, df ）：整个文档集合 **D** 包含关键词 **t_i** 的文档数量（inverse document frequency: idf_i ）
$$idf_i = \log \frac{N}{df_i}$$
- TF-IDF term weight: $w_{ij} = tf_{ij} \times idf_i$

8

向量空间模型 (Vector Space Model)

- 每一个文档 d 被表示为一个权重的向量：

$$d_i = (w_{i1}, w_{i2}, \dots, w_{in})$$

- 每一个查询 q 也可以被表示为一个权重的向量：

$$q = (w_{1q}, w_{2q}, \dots, w_{nq})$$

- 其中

$$w_{iq} = \begin{cases} \left(0.5 + \frac{0.5 f_{iq}}{\max\{f_{1q}, f_{2q}, \dots, f_{nq}\}} \right) \times \log\left(\frac{N}{df_i}\right) & \text{if } f_{iq} \neq 0 \\ 0 & \text{if } f_{iq} = 0 \end{cases}$$

- 查询 q 与文档 d_j 之间的相关程度可以使用两个向量之间的相似度来衡量，最常用的是cosine similarity

$$\text{cosine}(\mathbf{d}_j, \mathbf{q}) = \frac{\langle \mathbf{d}_j, \mathbf{q} \rangle}{\|\mathbf{d}_j\| \times \|\mathbf{q}\|} = \frac{\sum_{j=1}^{|\mathbf{q}|} w_{ij} \times w_{iq}}{\sqrt{\sum_{j=1}^{|\mathbf{q}|} w_{ij}^2} \times \sqrt{\sum_{i=1}^{|\mathbf{q}|} w_{iq}^2}}$$

检索问题：

根据文档与查询之间的相关程度大小
从高到低排列文档

9

精准的相关性衡量远没有那么容易.....

- 更复杂一些的相关性测度

$$\text{okapi}(d_j, q) = \sum_{t_i \in q, d_j} \ln \frac{N - df_i + 0.5}{df_i + 0.5} \times \frac{(k_1 + 1) f_{ij}}{k_1 (1 - b + b \frac{dl_j}{avdl}) + f_{ij}} \times \frac{(k_2 + 1) f_{iq}}{k_2 + f_{iq}}$$

$$\text{pmw}(d_j, q) = \sum_{t_i \in q, d_j} \frac{1 + \ln(1 + \ln(f_{ij}))}{(1 - s) + s \frac{dl_j}{avdl}} \times f_{iq} \times \ln \frac{N + 1}{df_i}$$

t_i is a term

f_{ij} is the raw frequency count of term t_i in document d_j

f_{iq} is the raw frequency count of term t_i in query q

N is the total number of documents in the collection

df_i is the number of documents that contain the term t_i

dl_j is the document length (in bytes) of d_j

$avdl$ is the average document length of the collection

10

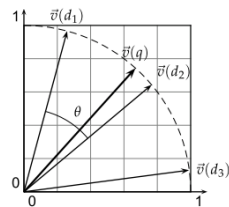
例子 (向量空间模型)

Example: Suppose $\mathbf{D} = (0.2, 0, 0.3, 1)$ and

$\mathbf{Q} = (0.75, 0.75, 0, 1)$.

Using Cosine function, we have

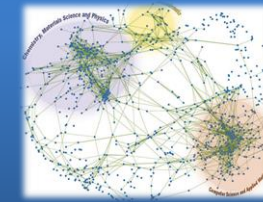
$$\begin{aligned} \text{sim}(\mathbf{D}, \mathbf{Q}) &= (0.15 + 0 + 0 + 1) / (\|\mathbf{D}\| * \|\mathbf{Q}\|) \\ &= 1.15 / (1.063 * 1.458) \\ &= 0.742 \end{aligned}$$



11

链接分析基础

Link Analysis Foundations



12

互联网上的信息获取

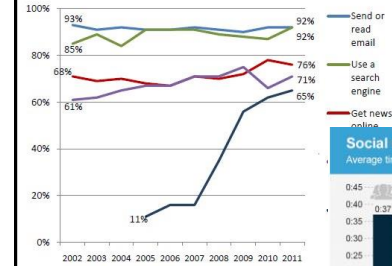
- **搜索 (Searching)/ 检索 (Retrieval)**
 - 向搜索引擎提交一个查询, 以获取想要的网络文档
 - Submit a query to a search engine to find desired documents
- **许多知名的网络搜索引擎**
 - Google, Baidu, Yahoo, Bing, AltaVista...
- **搜索是网络上仅次于Email的第二/四大流行活动**

13

互联网上的信息获取 (续)

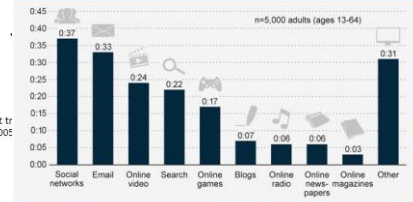
Over time, search and email are most popular online activities

% of Internet users who do each activity



Source: The Pew Research Center's Internet & American Life Project for 2002-2011. Social network site use not tracked prior to February, 2005 activity trends, go to pewinternet.org.

Social Networking is the No. 1 Online Activity in the U.S.
Average time U.S. consumers spent with digital media per day in 2012 (hours/minutes)



statista

Mashable

Source: GFK, IAB

搜索引擎技术

- A search engine is a Web-based search system for finding information on the Web.
- A search engine is essentially a text retrieval system for web pages plus a Web interface.
 - Web pages are widely distributed on many servers.
 - Web pages are extremely dynamic/volatile.
 - Web pages have more structures (extensively tagged).
 - Web pages are extensively linked.
 - Web pages are very voluminous and diversified.



**So what's new
beyond contents???**

15

搜索引擎：如何对搜索结果排序？

- **早期搜索引擎**
 - 根本不评价结果重要性, 而是直接按照某自然顺序 (例如时间顺序或编号顺序) 返回结果
 - 这在结果集比较少的时候还行, 但是一旦结果集变大, 用户叫苦不迭, 试想让你从几万条质量参差不齐的页面中寻找需要的内容, 简直就是一场灾难, 这也注定这种方法不可能用于现代的通用搜索引擎
- **基于检索词的评价**
 - 检索词 (或查询词) 与文档内容相关性
 - 和检索词匹配度越高的页面重要性越高: $tf-idf$ 权重, cosine相似度
 - 非常容易受到一种叫 "Term Spam" 的攻击

Link Analysis

16

链接分析 (Link Analysis)

- 目标
 - 讨论如何将互联网的特性纳入构建好的搜索引擎的过程中
 - Discuss how to take the special characteristics of the Web into consideration for building good search engines
- 网络爬虫 (Web crawler)
- 链接信息的使用 (The use of link information)
-

17

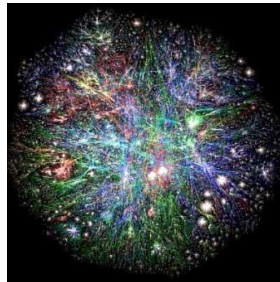
链接信息的使用

- 网页之间的超链接(Hyperlinks)为文档检索提供了新的机会
- 例如 :
 - The ranking score (similarity) of a page with a query can be spread to its **neighboring** pages
 - Links can be used to compute the **importance** of web pages based on citation analysis
 - Links can be combined with a regular query to find **authoritative** pages on a given topic



18

PageRank



19

PageRank背景简介

- Google早已成为全球最成功的互联网搜索引擎，但这个当前的搜索引擎巨无霸却不是最早的互联网搜索引擎，在Google出现之前，曾出现过许多通用或专业领域搜索引擎
 - 例如：雅虎、Inktomi、Overture、LookSmart、MSN、HotBot等。
- Google最终能击败所有竞争对手，很大程度上是因为它解决了困扰前辈们的最大难题：**对搜索结果按重要性排序**；而解决这个问题的算法就是PageRank
- 毫不夸张的说，是PageRank算法成就了Google今天的地位

20

链接信息的使用：PageRank

● PageRank citation ranking

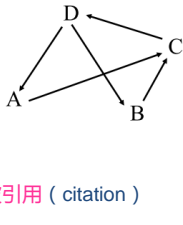
- Page, Brin, et al. The PageRank Citation Ranking: Bring Order to the Web. Technical report, 1998

● 整个网络可以被视为一个巨大的有向图 $G(V, E)$

- V ：网页页面的集合（顶点集合，vertices）
- E ：超链接的集合（有向边集合，directed edges）

● 每一个页面都可能有一些：

- 出向边（outgoing edges, forward links）：出度
- 入向边（incoming links, back links）：入度
- 一个页面的每个后向链接都体现了该页面的一次被引用（citation）



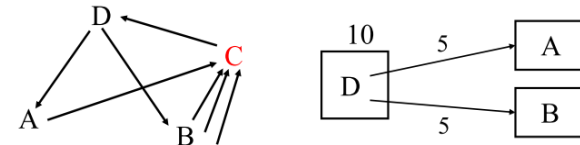
● PageRank是一种基于后向链接的全局网页重要性评估测度

21

PageRank的计算原则

● PageRank基于以下几个基本观点

- 如果一个页面被许多页面链接到（入度较大），那么该页面较重要的可能性较大
- 如果一个页面被一些重要的页面链接到，那么该页面可能也很重要，即使链接到它的页面数量不是很多
- 一个页面的重要性被它所指向的页面平均划分，且传播给这些它所指向的页面（随机访问假设）



22

PageRank的基础模型

- 假设有 N 个网页通过某些超链接互相关联，第 i 个网页的重要性测度（PageRank值， $P(i)$ ）可以由所有链接到它的其他网页的重要性测度值完全决定，即：

$$P(i) = \sum_{j=1}^N A_{ji} \cdot P(j) = \sum_{(j,i) \in E} \frac{P(j)}{O_j}$$

$$A_{ji} = \begin{cases} \frac{1}{O_j} & \text{if } (j,i) \in E \\ 0 & \text{otherwise.} \end{cases}$$

O_j : 第 j 个网页的出度

$A = (A_{ij})_{N \times N}$: 邻接矩阵

$$\begin{cases} P(1) = A_{21}P(2) + A_{31}P(3) + \dots + A_{n1}P(n) \\ P(2) = A_{12}P(1) + A_{32}P(3) + \dots + A_{n2}P(n) \\ \dots \\ P(n) = A_{1n}P(1) + A_{2n}P(2) + \dots + A_{n-1n}P(n-1) \end{cases}$$

$$P = (P(1), P(2), \dots, P(n))^T$$

$$P = A^T P$$

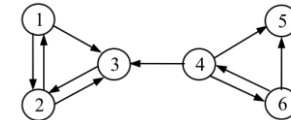
23

PageRank的基础模型

- According to Page & Brin, 1998, the PageRank function is an assessment model of Internet random surfer behavior, which regards that

- The surfer will randomly click an link in the current web page, e.g., j , the probability of clicking any links (outcoming web pages) will be equal, $1/O_j$ (即为 A_{ji})。[也可以用Markov chain model来解释]

$$A = \begin{pmatrix} A_{11} & A_{12} & \dots & A_{1n} \\ A_{21} & A_{22} & \dots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & \dots & A_{nn} \end{pmatrix}$$



- 基础模型的假设：有 N 个网页通过某些超链接互相关联，第 i 个网页的重要性测度（PageRank值， $P(i)$ ）可以由所有链接到它的其他网页的重要性测度值完全决定（强连接图）

24

PageRank的扩展模型

- 假设有 N 个网页通过某些超链接互相关联，第 i 个网页的重要性测度（PageRank值， $P(i)$ ）可以由以下两个部分决定：
 - 所有链接到它的其他网页的重要性测度值（概率： d ）（阻尼因子）
 - 未通过超链接访问直接随机访问页面行为（概率： $1-d$ ）

$$P(i) = (1-d) + d \sum_{j=1}^N A_{ji} P(j) = (1-d) + d \sum_{(j,i) \in E} \frac{P(j)}{O_j}$$

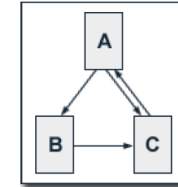
$$\begin{cases} P(1) = (1-d) + d[A_{21}P(2) + A_{31}P(3) + \dots + A_{n1}P(n)] \\ P(2) = (1-d) + d[A_{12}P(1) + A_{32}P(3) + \dots + A_{n2}P(n)] \\ \dots\dots\dots \\ P(n) = (1-d) + d[A_{1n}P(1) + A_{2n}P(2) + \dots + A_{n-1n}P(n-1)] \end{cases} \quad P = (1-d)e + dA^T P$$

25

PageRank的求解

- Suppose there is a network as shown in the figure. For simplicity, set $d = 0.5$, we have:

- $P(A) = 0.5 + 0.5P(C)$
- $P(B) = 0.5 + 0.5(P(A)/2)$
- $P(C) = 0.5 + 0.5(P(A)/2 + P(B))$



- The we can calculate:

- $P(A) = 14/13 = 1.07692308$
- $P(B) = 10/13 = 0.76823077$
- $P(C) = 15/13 = 1.15384615$

According to Page & Brin, usually $d = 0.85$.

26

PageRank求解：power iteration method

- Since the size of Internet is extremely large, e.g., $N \rightarrow \infty$, so the N-ary equations cannot be easily solved.
- Google adopts an approximate approaching method – power iteration. The procedure is as follows:

- Assign an initial PageRank value (=1) for each web page;
- Compute PageRank values for other web pages iteratively;
- After limited iteration, the theoretical PageRank values could be converged on some accuracy level

(即前后两次迭代的结果差异小于给定阈值 ϵ).

```

PageRank-Iterate( $G$ )
 $P_0 \leftarrow e/n$ 
 $k \leftarrow 1$ 
repeat
   $P_k \leftarrow (1-d)e + dA^T P_{k-1}$ ;
   $k \leftarrow k + 1$ ;
until  $\|P_k - P_{k-1}\|_1 < \epsilon$ 
return  $P_k$ 
  
```

27

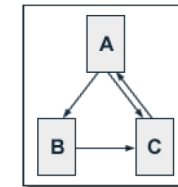
PageRank求解：例子

- The procedure:

- Initially, set $P(A) = P(B) = P(C) = 1$.

Iteration	P(A)	P(B)	P(C)
0	1	1	1
1	1	0.75	1.125
2	1.0625	0.765625	1.1484375
3	1.07421875	0.76855469	1.15283203
4	1.07641602	0.76910400	1.15365601
5	1.07682800	0.76920700	1.15381050
6	1.07690525	0.76922631	1.15383947
7	1.07691973	0.76922993	1.15384490
8	1.07692245	0.76923061	1.15384592
9	1.07692296	0.76923074	1.15384611
10	1.07692305	0.76923076	1.15384615
11	1.07692307	0.76923077	1.15384615
12	1.07692308	0.76923077	1.15384615

$$\begin{aligned} P(A) &= 0.5 + 0.5P(C) \\ P(B) &= 0.5 + 0.5(P(A)/2) \\ P(C) &= 0.5 + 0.5(P(A)/2 + P(B)) \end{aligned}$$



28

PageRank的优势和问题

● Advantages

- Its ability to **fight spam**: Since it is **not easy** for Web page owner to add in-links into his/her page from other important pages, it is thus not easy to influence PageRank.
- It is a **global** measure and is **query independent**: the PageRank values of all the pages on the Web are computed and saved **off-line** rather than at the query time.

● Criticisms

- The **query-independence** nature of PageRank
- PageRank does not consider **time**

29

搜索结果综合排序

● 搜索引擎对搜索结果排序的两个重要因素：

- 搜索结果文档与查询词之间的内容相关性
- 搜索结果文档的链接重要性

- 一个网页文档的排序得分（**ranking score**），可以表示为它与查询的内容相关度以及它自身（**链接**）重要度的加权求和值，即：

$$Ranking_score(q, d) = \begin{cases} w \cdot sim(q, d) + (1 - w) \cdot P(d); & \text{if } sim(q, d) > 0 \\ 0. & \text{otherwise} \end{cases}$$

- 其中, $0 < w < 1$, $sim(q, d)$ 和 $P(d)$ 需要标准化为 $[0, 1]$ 区间的实数.

30

Google: Beyond PageRank



31

PageRank in Web Search

- Google downloads all web pages into local data centers using web crawlers.
 - It is time-consuming, but not too hard technically.
 - Google has been doing this since 1996.
- Based on the downloaded web pages as well as the links, Google has been keeping computing PageRank values for each web page iteratively since 1996, which is clearly a hard work.
- When input a keyword, Google will search local data centers to extract all the web pages containing the keyword.
 - Nothing new, just some classical search algorithms.
- Sort all the web pages based on PageRank values.
 - Nothing new, just some classical sort algorithms.

32

PageRank in Web Search

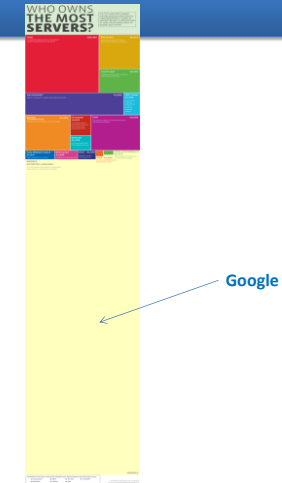
- 无论是方程组解法还是迭代解法，关键在于整个互联网要同时进行计算。这意味着：
 - 理论上说，要下载整个互联网——Google开始此工作于1996年，至今仍未完成。
 - 数据量极端大——Google开发了专用的分布式计算结构来下载和处理如此大量的数据；2001年3月，Google还只有8000台电脑，到2003年已经是10万台。目前，在全球范围，Google拥有36个大小不一的数据中心。以平均每个数据中心有150个服务器集群计，这意味着Google拥有的服务器数量超过20万台。Google不愿透露自己共有多少台服务器，但内行人估计，Google服务器的数量应该远远超过这一数字，而且每天都在增长。更有人认为，Google在全球的数据中心超过45个。
 - 计算量极端大——Google的所有电脑每天都是不停的计算耗电量极大，因此Google开始将其机房搬至新墨西哥州、爱荷华州等地皮便宜，电费低的地区。

<http://www.bundpic.com/link.php?action=print&linkid=6887>

33

Google Servers

- 一份由INTAC公布的图像显示了主流的高科技公司专用服务器的大致数量，例如英特尔有大约10万台服务器运行，而Facebook、AT&T公司和时代华纳有线大约有2-3万台，虽然你没有亲眼看到他们强大的服务器农场，但以下的统计数据可以帮助你了解有趣的事实，例如Google他们的服务器数量是100多万台，占全球的2%。



34

Website Alliance (网站联盟)

- The website alliance means that, in the alliance, the websites will exchange outcoming links among each others.
- For any certain website, e.g., A, the outcoming links will cause $PR(A)$ leak.
- Website alliance will not increase the whole PR value of the alliance, but can further arrange PR values among websites.

35

Additional PageRank Parameters

- Other parameters:
 - Visibility of a link
 - Position of a link within a document
 - Distance between web pages
 - Up-to-dateness of a linking page
 -
- PageRank @ 2008
 - 7 parameters have been involved, but now at least 150 ...

36

如果进行搜索引擎优化？

- **网站必须做到：**
 - 能被搜索到，如采用Big Words作为搜索关键字；
 - 避免被搜索引擎检索到无用的内容，如可以采用防火墙技术；
 - 增加显著度，如将网站内部链接都链入到某个需要强调的网页上；
 - ...
- **网站要避免做到：**
 - 放置无意义的内容；
 - 没有文本，太多flash或图片等富媒体；
 - 没有Big Words；
 - 太多无用的生僻词；
 - 网页更新过快过多；
 - 失效链接；
 -

37

On What Google Make Huge Money?

Online Advertising

- 窄告是一种新型的网络广告模式，不仅适合于各行各业推广宣传品牌、产品等，也适合各种规格的网络广告发布商。



- 窄告就是“窄而告之”、“专而告之”。指客户投放的窄告直接投放到与之内容相关的网络媒体上的文章周围，同时窄告还会根据浏览者的偏好、使用习性、地理位置、访问历史等信息，有针对性地将窄告投放到真正感兴趣的浏览者面前。



窄告的工作流程图

Target Marketing/Advertising

- Targeted marketing is a type of advertising whereby advertisements are placed so as to reach consumers based on various traits such as demographics, purchase history, or observed behaviors.
- Google AdWords
- Google AdSense
- 百度“凤巢”
-



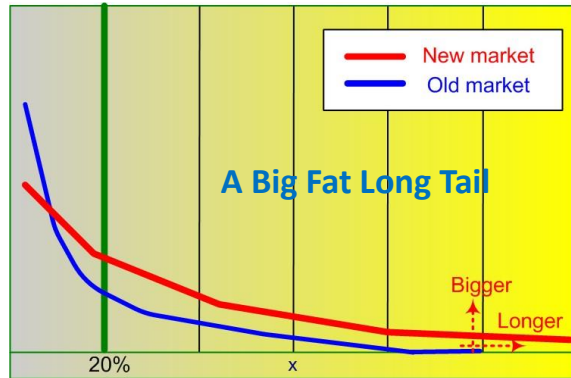
39

Targeting the Long Tail

- **Traditional/1st generation media**
 - TV+Newspaper+Portal+...
 - Only the 20% mainstream consumers could be targeted, the left 80% long tail has to be ignored.
 - Mainstream mass ads → Expensive ads payment
- **Targeted Marketing media**
 - Google+Baidu+...
 - Target each consumer by search engine and further customize the ad.
 - Pay by per click → Cheap payment
 - Actually, Google Adwords engage millions of small advertisers.

40

Targeted Marketing and Long Tail



41

Google's Success Stories



42

Vital Challenges to Google



43

Google关键词欺诈？

- 拥有经营路易威登和其他奢侈品牌的奢侈品集团LVMH对Google将搜索引擎中的关键字“Vuitton”高价卖给他人的行为表示不满。
- 该集团声称，如果有人在Google引擎上输入“Louis Vuitton merchandise”，出现的搜索结果里就会有竞争对手产品的广告，甚至还有销售假冒路易威登产品的店铺信息。
- “Google的广告宣传给了那些出售假冒名牌的商家前所未有的曝光率，这样的曝光率是他们做梦都垂涎的。”路易威登的律师帕特里斯·德坎德在本周二向欧洲法院做出陈述。

44

Google' s Competitors



Baidu 百度



45

Google' s Competitors

- The most potential competitor is Microsoft Bing.



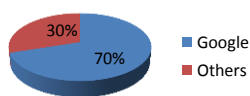
= *But It's Not Google!*

46

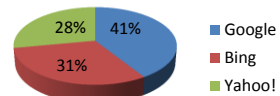
Google' s Competitors

- Michael Kordahi, a programmer, invented an black-box test, in which a user query any keywords and don't know which search engine provide the results. Then, the user should evaluate the quality of the query results.

Market Share



Ratio of Best Evaluation



Preference, not quality, determine user's choice on search engines!

47

New Market, New Challenges

- Mobile Search
- Social Search
- Instant Search
- Amazon?
- 人肉搜索 ???
-



Google's Future



49

小结：信息检索及信息搜索服务

- 深度业务分析——组合方法及应用
- 信息检索基础
 - 基本概念
 - 文档表示、文档权重计算
 - 内容相似度计算
- 链接分析基础
 - 链接信息的利用
 - PageRank算法的思路、基础模型、扩展模型
 - PageRank的求解
 - 搜索结果的综合排序

50

期末课程论文说明

- 主题要求
 - 必须与“大数据管理”相关
 - 建议围绕所学专业背景下的“大数据管理问题”展开
- 内容要求
 - 不少于4000字，版式：word中正文小四字体，1.5倍行距
 - 独立完成，不得大段拷贝或直接引用网上、书上及他人已发布内容，需要适当引用时请在引用位置注明参考文献来源（查重）
 - 论文内容框架（建议）：
 - 1. 学习本课程的心得体会、感受，对本课程教学的建议和意见（必有）
 - 2. 论文背景介绍
 - 3. 论文涉及的大数据问题及管理需求、策略和意义（可举实例说明）
 - 4. 本人对该大数据问题的看法、观点及讨论
 - 5. 总结
 - 6. 参考文献和资料

51

期末课程论文说明（续）

- 论文提交要求
 - 需要以电子版提交，建议提交word版本
 - 作业提交邮箱：bigdata_homework@163.com
 - 作业提交截止时间：第19周周日（2015.01.11）24时
- 其他说明
 - 邮件标题和电子版论文文件请务必按照“学号_班级_姓名.docx”命名，例如“2014211234_2014212103_张三.docx”，也请在邮件中留下姓名、学号及联系方式，以备论文有问题时能够联系到；
 - 请在截止时间之前提交论文（不要在截止时间附近，以避免系统原因过期），过期将不再接收论文提交，成绩为0，请务必注意；
 - 每次提交论文后，作业邮箱都会有“已收到邮件”的自动回复，如未收到自动回复，表示发送不成功，请在截止时间内重新提交；
 - 论文评分的关注重点
 - 有效的课程建议和意见
 - 关注问题的新颖度
 - 个人分析和讨论的深度
 - 论文的整体工作量

52