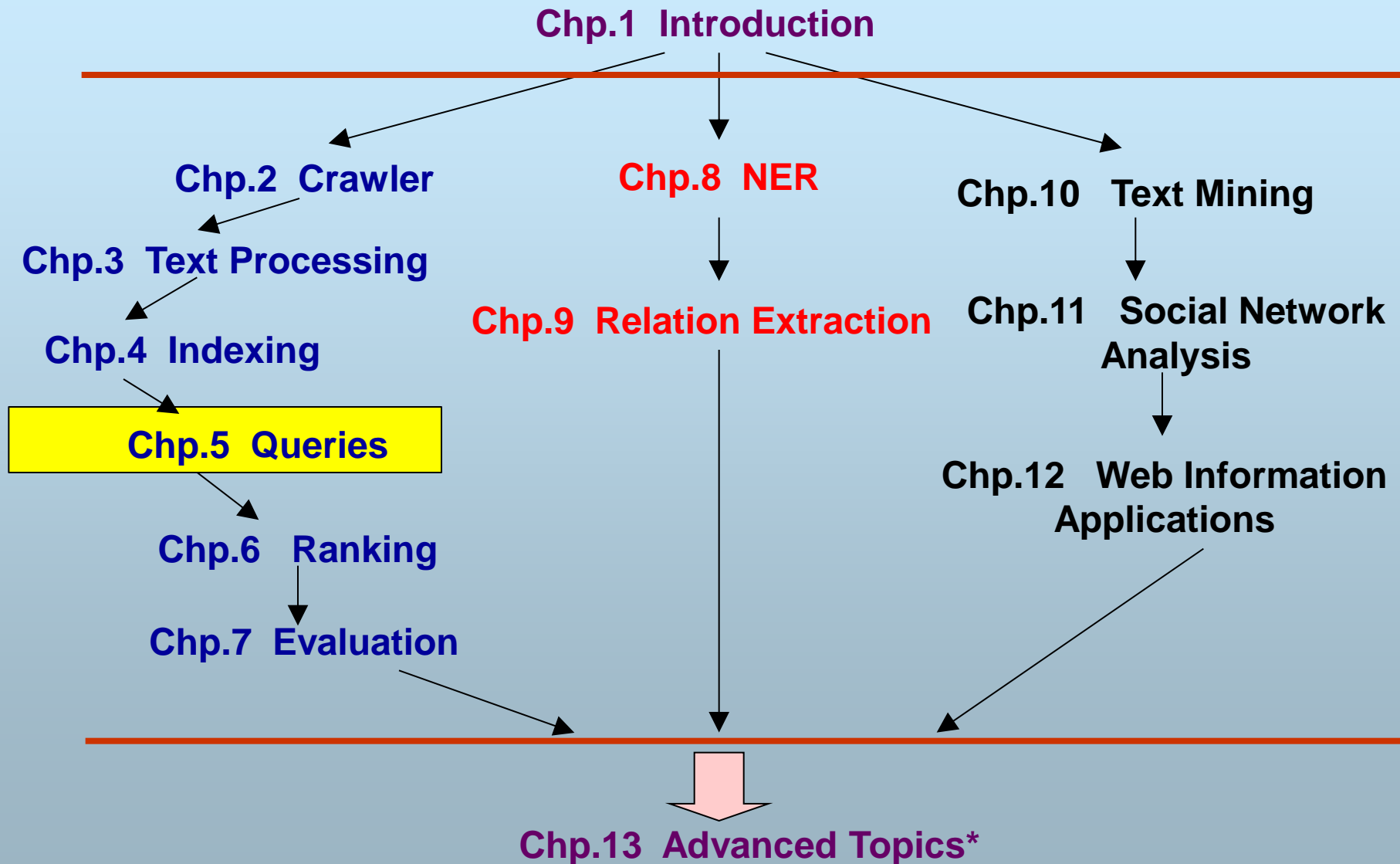


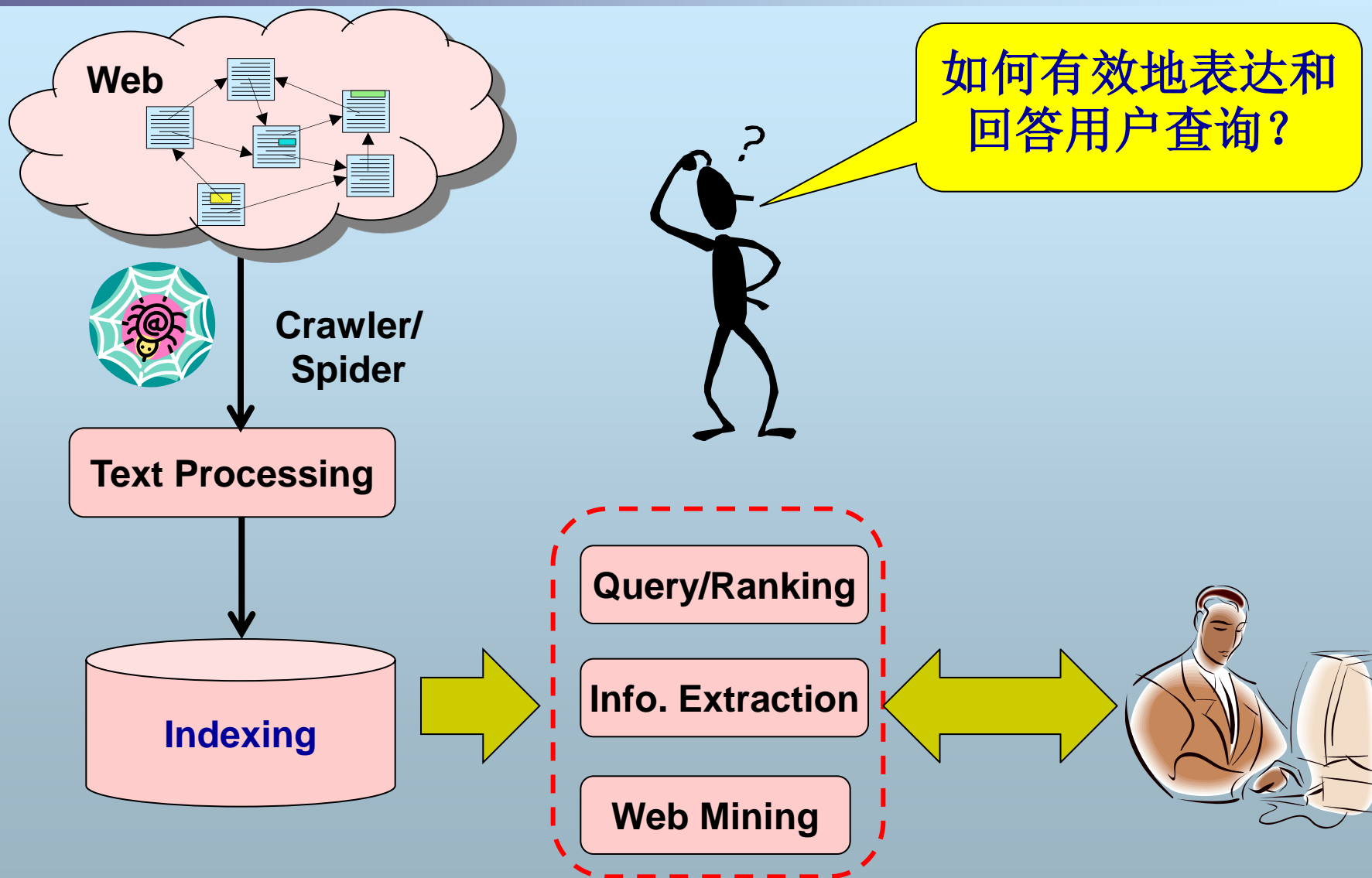
Queries

A thick, wavy orange line that spans the width of the slide, positioned below the title 'Queries'.

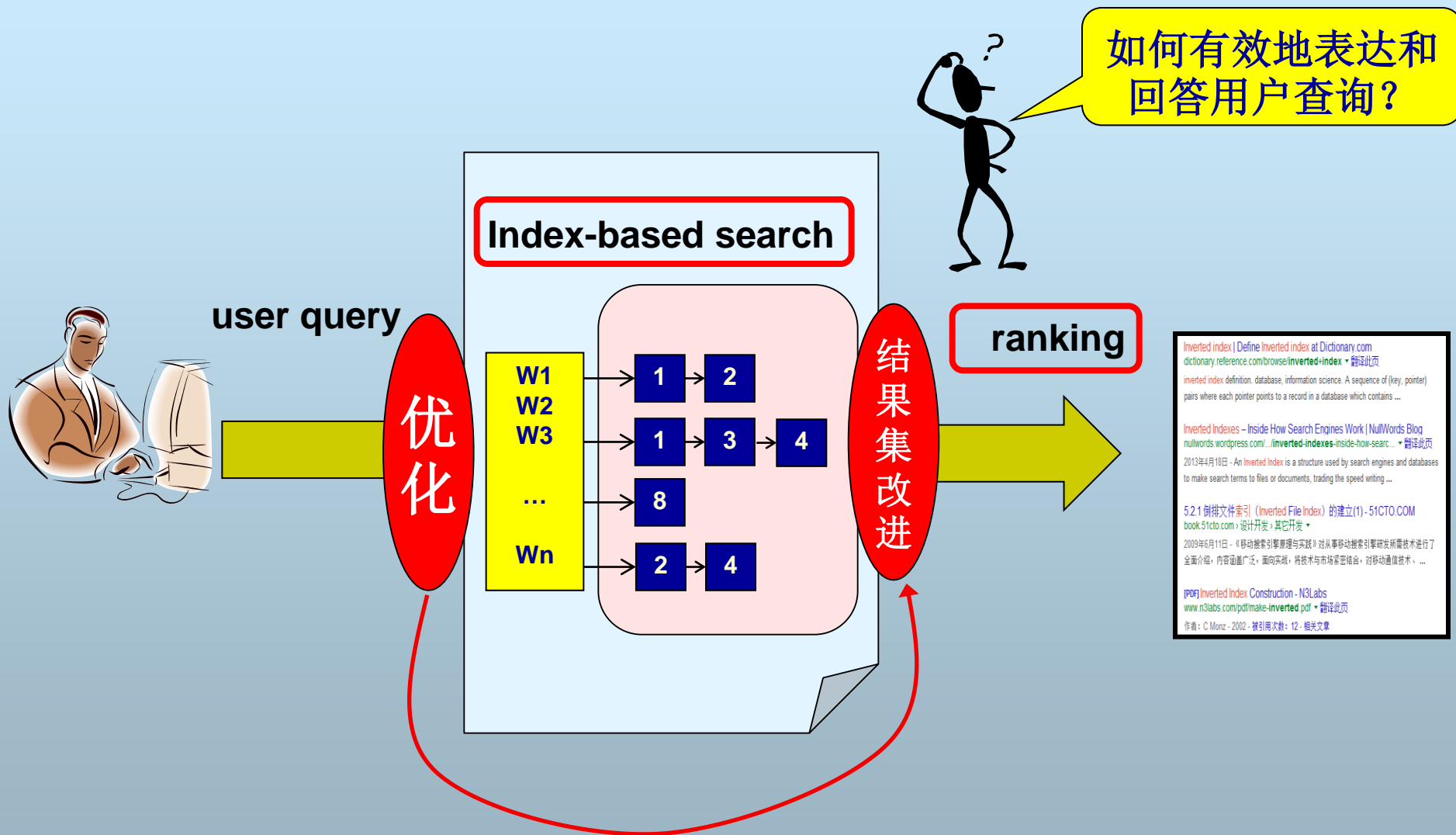
课程知识结构



本章讨论的问题



Web查询处理的过程



Inverted index | Define Inverted index at Dictionary.com
dictionary.reference.com/browse/inverted-index • 翻译此页

inverted index definition, database, information science. A sequence of (key, pointer) pairs where each pointer points to a record in a database which contains ...

Inverted Indexes - Inside How Search Engines Work | NullWords Blog
nullwords.wordpress.com/.../inverted-indexes-inside-how-search-... • 翻译此页

2013年4月18日 - An Inverted Index is a structure used by search engines and databases to make search terms to files or documents, trading the speed writing ...

5.2.1 倒排文件索引 (Inverted File Index) 的建立(1) - 51CTO.COM
book.51cto.com • 设计开发 • 其它开发 •

2009年6月11日 - 《移动搜索引擎原理与实践》对从事移动搜索引擎研发所需技术进行了全面介绍, 内容涵盖广泛, 面向实践, 将技术与市场紧密结合, 对移动通信技术、...

[PDF] Inverted Index Construction - N3Labs
www.n3labs.com/pdf/make-inverted-pdf • 翻译此页

作者: C. Monz - 2002 - 被引用次数: 12 - 相关文章

本章主要内容

- 查询表达
- 相关性反馈
- 查询扩展

一、查询表达

■ Information Retrieval

Given a **query** and a **corpus**,
find **relevant documents**.

- ◆ **query**: user's expression of the information need
- ◆ **corpus**: the repository of retrievable items
- ◆ **relevance**: satisfaction of the information need



查询表达

排序 (next chp.)

一、查询表达

- 问题：如何准确、正确地表达用户查询？
 - A query can represent very different information needs
 - ◆ table: furniture, data structure, ...
 - ◆ office: a work place, software
 - A query can be a poor representation of the information need
 - ◆ Query terms will not always appear in the index, e.g., **plane** vs. **aircraft**
 - ◆ Some (new) queries are difficult to express.

② 钢铁锅，含眼泪喊修瓢锅！ 请问大神这是什么歌曲？

一、查询表达

- 局部(Local)优化方法: 对用户查询进行局部的分析
 - 相关性反馈 **relevance feedback**
- 全局(Global)优化方法: 进行一次性的全局分析(比如分析整个文档集)来产生同/近义词词典 (thesaurus)
 - 查询扩展 **query expansion**



二、相关性反馈

- 用户在查询后标记相关/不相关文档，然后（迭代）更新查询以获得更好的结果

- **Motivation**

- You may not know what you're looking for, but you'll know when you see it
 - ◆ “find me more documents like this...”
- Query formulation may be difficult; simplify the problem through iteration

Initial Query

Baidu 图片

新闻 网页 贴吧 知道 音乐 图片 视频 地图 百科 文库

本田 像轿车又有点像SUV

百度一下

宅男最爱的游戏美女内涵图 ^{NEW}

相关搜索 本田suv 广汽本田suv 东风本田suv 本田suv汽车大全 2014款本田suv 新款本田suv 本田suv图片



讴歌 MDX 大幅优惠
狂降三万·仅售64万元



本田飞度SUV提前曝光
北美车展正式首发



Revised Query

Baidu 图片

新闻 网页 贴吧 知道 音乐 图片 视频 地图 百科 文库

本田 像轿车又有点像SUV 跨界车



百度一下

宅男最爱的游戏美女内涵图 ^{NEW}

相关搜索

启辰r50跨界车图片

本田suv车图片

跨界图片

跨界服饰

跨界摩托车

艺术家的跨界作品

跨界鞋



车型
轴距 (mm)
悬挂系统
容量 (L)
变速箱
功率 (kW)
扭矩 (Nm)
整车质量 (kg)
整备质量

二、相关性反馈

- User issues a (short, simple) query
- The **user** marks returned documents as relevant or non-relevant.
- The **system** computes a better representation of the information need based on feedback.
- Relevance feedback can go through one or more **iterations**.

Idea: it may be difficult to formulate a good query when you don't know the collection well, so iterate

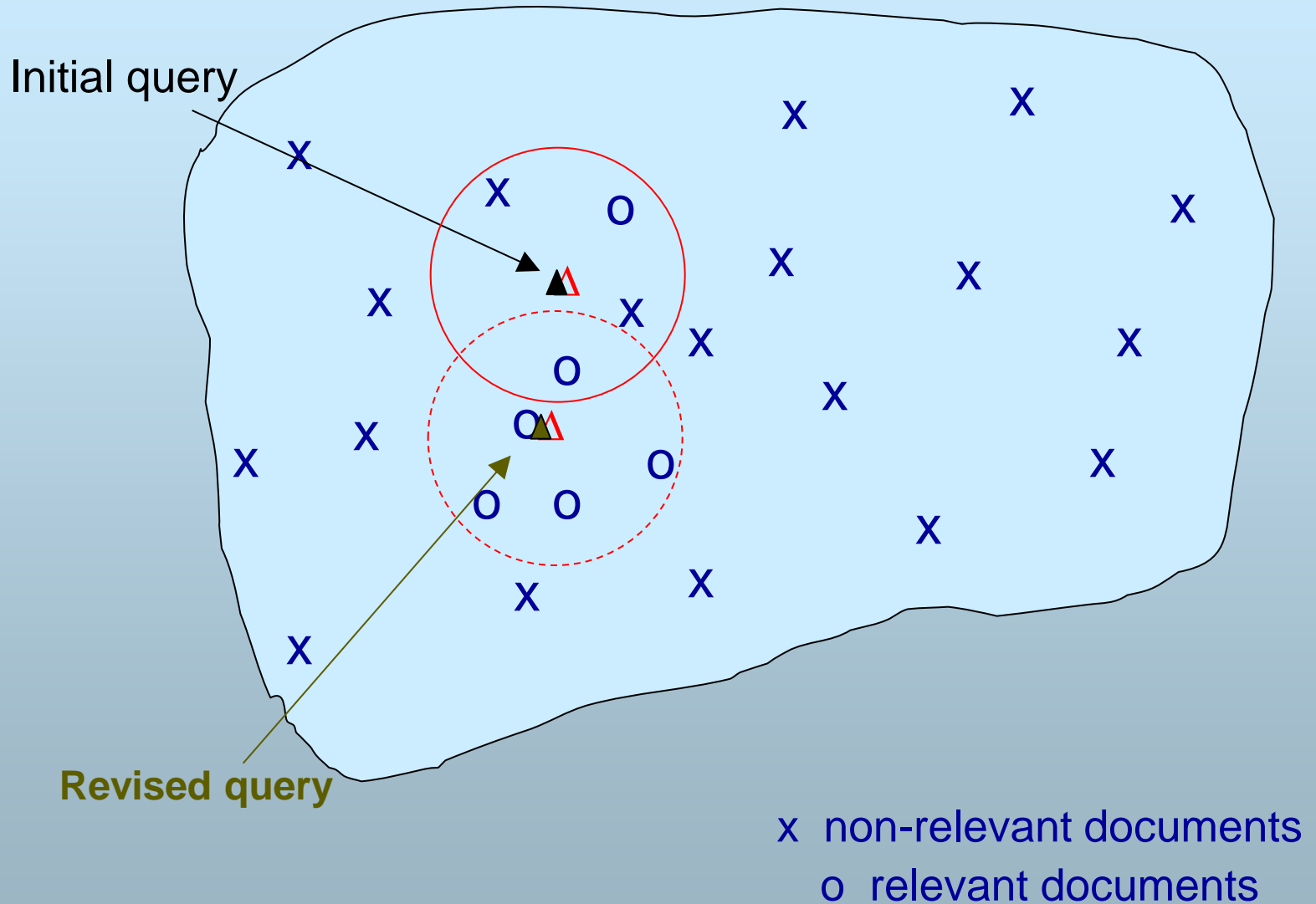
通常用术语“**ad hoc retrieval**”来表示那种无相关反馈的常规检索

二、相关性反馈

■ 相关性反馈如何工作？

- **Let's assume that there is an optimal query**
 - ◆ The goal of relevance feedback is to bring the user query closer to the optimal query
- **How does relevance feedback actually work?**
 - ◆ Use relevance information to update query
 - ◆ Use query to retrieve new set of documents
- **What exactly do we “feed back”?**
 - ◆ Boost weights of terms from relevant documents
 - ◆ Add terms from relevant documents to the query
- **Note that this is hidden from the user**

二、相关性反馈



2、相关性反馈示例1

初始查询:

[new space satellite applications] 初始查询的检索结果: ($r = \text{rank}$)

r	
+ 1	0.539 NASA Hasn't Scrapped Imaging Spectrometer
+ 2	0.533 NASA Scratches Environment Gear From Satellite Plan
3	0.528 Science Panel Backs NASA Satellite Plan, But Urges Launches of Smaller Probes
4	0.526 A NASA Satellite Project Accomplishes Incredible Feat: Staying Within Budget
5	0.525 Scientist Who Exposed Global Warming Proposes Satellites for Climate Research
6	0.524 Report Provides Support for the Critics Of Using Big Satellites to Study Climate
7	0.516 Arianespace Receives Satellite Launch Pact From Telesat Canada
+ 8	0.509 Telecommunications Tale of Two Companies

用户将一些文档标记为相关 “+”.

2、相关性反馈示例1

2.074	new	15.106	space
30.816	satellite	5.660	application
5.991	nasa	5.196	eos
4.196	launch	3.972	aster
3.516	instrument	3.446	arianespace
3.004	bundespost	2.806	ss
2.790	rocket	2.053	scientist
2.003	broadcast	1.172	earth
0.836	oil	0.646	measure

相关性反馈后
扩展的查询

原始查询: [new space satellite applications]

2、相关性反馈示例1

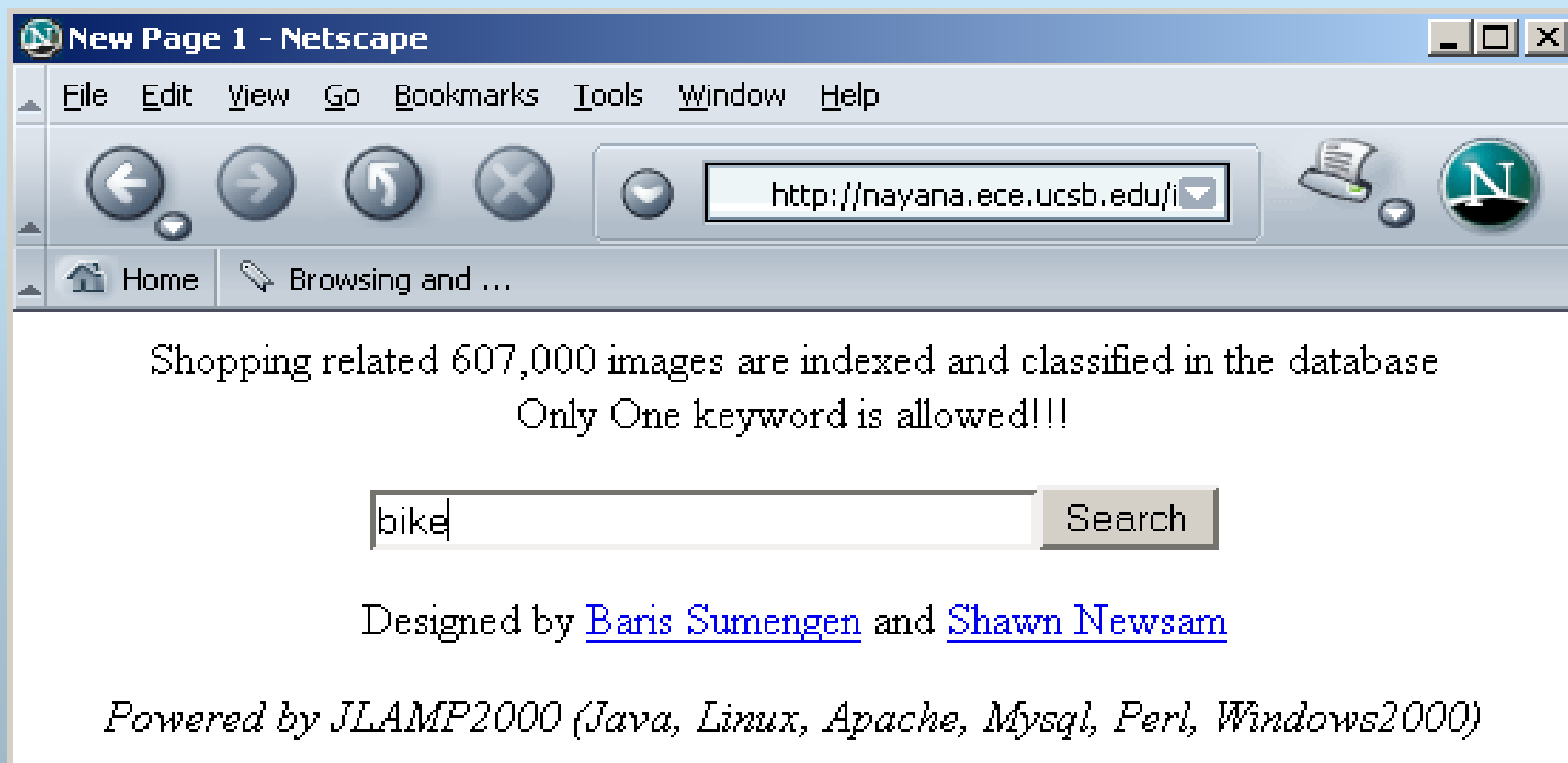
■ 基于扩展查询的检索结果

rank

- *2 1 0.513 NASA Scratches Environment Gear From Satellite Plan
- *1 2 0.500 NASA Hasn't Scrapped Imaging Spectrometer
- 3 0.493 When the Pentagon Launches a Secret Satellite, Space Sleuths Do Some Spy Work of Their Own
- 4 0.493 NASA Uses 'Warm' Superconductors For Fast Circuit
- *8 5 0.492 Telecommunications Tale of Two Companies
- 6 0.491 Soviets May Adapt Parts of SS-20 Missile For Commercial Use
- 7 0.490 Gaping Gap: Pentagon Lags in Race To Match the Soviets In Rocket Launchers
- 8 0.490 Rescue of Satellite By Space Agency To Cost \$90 Million

2、相关性反馈示例2

■ Image search engine






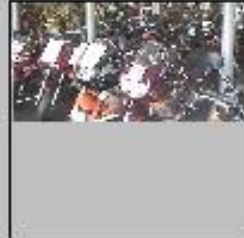








2、相关性反馈示例2

■ 初始查询结果

Initial search results interface showing a grid of images and their associated IDs and similarity scores.

Navigation buttons: Browse, Search, Prev, Next, Random

					
(144473, 16458)	(144457, 252140)	(144456, 262857)	(144456, 262863)	(144457, 252134)	(144483, 265154)
0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0
					
(144483, 264644)	(144483, 265153)	(144518, 257752)	(144538, 525937)	(144456, 249611)	(144456, 250064)
0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0

2、相关性反馈示例2

■ 相关性反馈

Interface showing a grid of 12 images related to bicycles and motorcycles, with navigation buttons (Browse, Search, Prev, Next, Random) at the top.

The images are arranged in two rows of six. The first row contains:

- Image 1: A white scooter. (144473, 16458)
- Image 2: A person riding a bicycle. (144457, 252140)
- Image 3: A white folding bicycle. (144456, 262857)
- Image 4: A black motorcycle. (144456, 262863)
- Image 5: A diamond-shaped logo for "BIKING 2000 BIKE OF THE YEAR". (144457, 252134)
- Image 6: A group of people riding bicycles. (144483, 265154)

The second row contains:













- Image 7: A red motorcycle. (144483, 264644)
- Image 8: A blue bicycle. (144483, 265153)
- Image 9: A stack of bicycle magazines. (144518, 257752)
- Image 10: A blue motorcycle. (144538, 525937)
- Image 11: A black bicycle seat. (144456, 249611)
- Image 12: A green bicycle. (144456, 250064)

Each image is accompanied by a set of three numerical values (e.g., 0.0, 0.0, 0.0) below it.

2、相关性反馈示例2

■ 相关性反馈后的查询结果

[Browse](#) [Search](#) [Prev](#) [Next](#) [Random](#)

					
(144538, 523493) 0.54182 0.231944 0.309876	(144538, 523835) 0.56319296 0.267304 0.295889	(144538, 523529) 0.584279 0.280881 0.303398	(144456, 253569) 0.64501 0.351395 0.293615	(144456, 253568) 0.650275 0.411745 0.23853	(144538, 523799) 0.66709197 0.358033 0.309059
					
(144473, 16249) 0.6721 0.393922 0.278178	(144456, 249634) 0.675018 0.4639 0.211118	(144456, 253693) 0.676901 0.47645 0.200451	(144473, 16328) 0.700339 0.309002 0.391337	(144483, 265264) 0.70170796 0.36176 0.339948	(144478, 512410) 0.70297 0.469111 0.233859

3、相关性反馈分类

■ Explicit Feedback

- 用户显式参加交互过程
- Also known as User Feedback

■ Implicit Feedback

- 系统跟踪用户的行为来推测返回文档的相关性，从而进行反馈。

■ Pseudo Feedback

- 没有用户参与，系统直接假设返回文档的前 k 篇是相关的，然后进行反馈。
- Also known as Blind Feedback

4、向量空间模型与质心

- 相关性反馈技术最早由Rocchio于1965年提出并最终应用于Salton领导研制的SMART系统中
- Salton在SMART系统中提出了著名的向量空间模型
- 向量空间模型 (Vector Space Model, VSM)
 - 每个文档表示为一个词项权重构成的向量
 - ◆ 因此文档也可看成是多维空间中的一个点
 - 每个查询也表示为一个词项权重构成的向量
 - 查询处理时通过比较查询和文档向量之间的相似度进行匹配

$$d_j = (w_{1,j}, w_{2,j}, \dots, w_{t,j})$$
$$q = (w_{1,q}, w_{2,q}, \dots, w_{t,q})$$

Rocchio J.J. (1971). **Relevance Feedback in Information Retrieval**. In Salton G. (Ed.), The SMART Retrieval System (pp. 313-323). Englewood Cliffs, N.J.: Prentice-Hall, Inc.

Gerard Salton, A. Wong, C. S. Yang: **A Vector Space Model for Automatic Indexing**. Commun. ACM 18(11): 613-620 (1975)

4、向量空间模型与质心

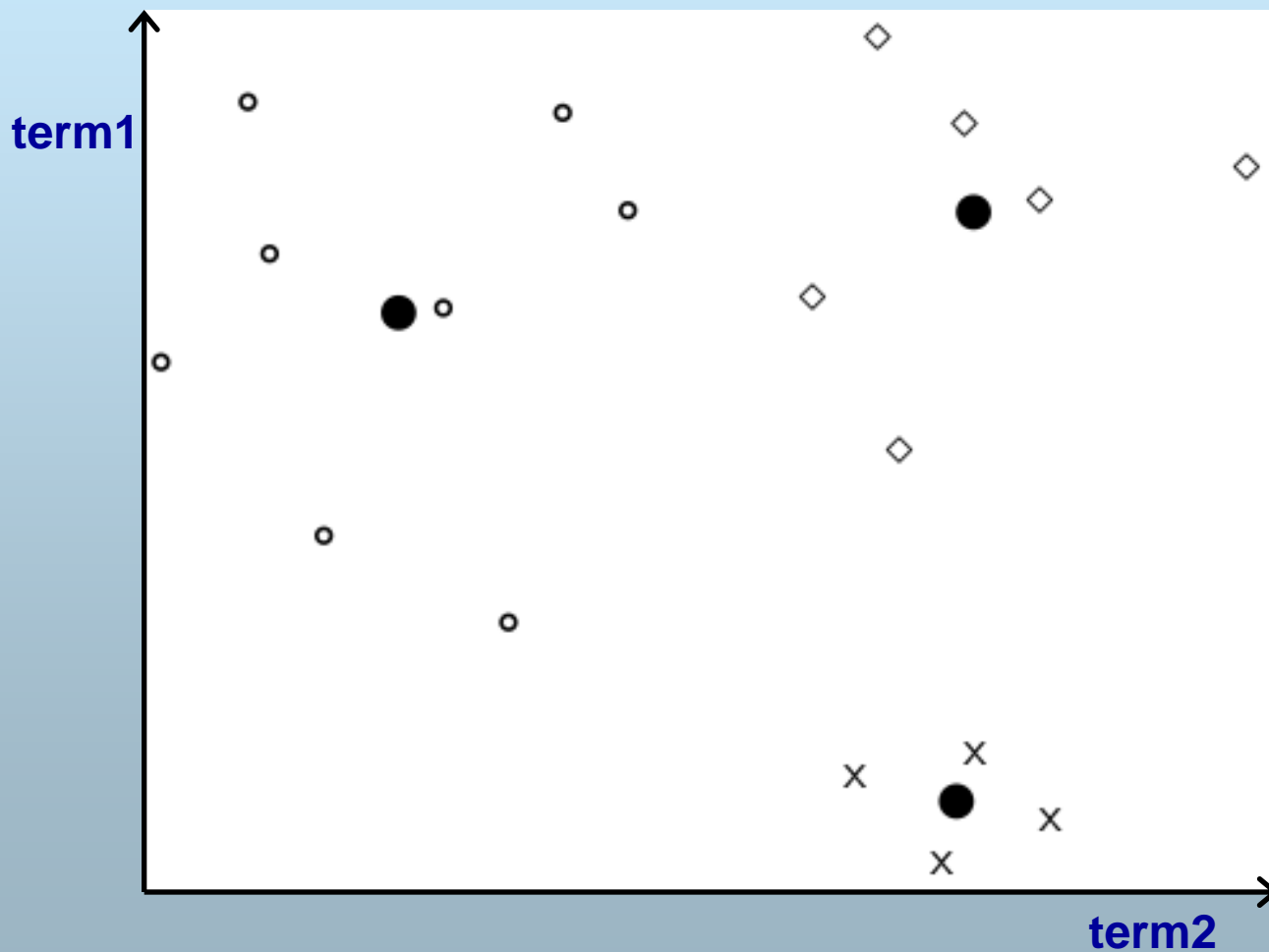
- 质心是一系列点的中心
- 由于在**VSM**中文档表示成高维空间中的点，因此，我们可以采用如下方式计算文档的质心

$$\vec{\mu}(C) = \frac{1}{|C|} \sum_{d \in C} \vec{d}$$

- 其中**C**是文档集合

4、向量空间模型与质心

■ 质心的例子



5、Rocchio算法

- Rocchio算法提供了一种将相关反馈信息融入到向量空间模型的方法。
- 出发点是最大化 $\text{sim}(Q, C_r) - \text{sim}(Q, C_{nr})$
- 最优查询向量应能够完美地区分相关文档与不相关文档

$$\vec{Q}_{opt} = \frac{1}{|C_r|} \sum_{\vec{d}_j \in C_r} \vec{d}_j - \frac{1}{N - |C_r|} \sum_{\vec{d}_j \notin C_r} \vec{d}_j$$

Q_{opt} = optimal query

C_r = set of rel. doc vectors

N = collection size

但是实际中我们无法获知全部相关文档集合 C_r

5、Rocchio算法

■ SMART(Salton, 1971)中的实现

$$\vec{q}_m = \alpha \vec{q}_0 + \beta \frac{1}{|D_r|} \sum_{\vec{d}_j \in D_r} \vec{d}_j - \gamma \frac{1}{|D_{nr}|} \sum_{\vec{d}_j \in D_{nr}} \vec{d}_j$$

q_m	modified query vector
q_0	original query vector
α, β, γ	weights (hand-chosen or set empirically)
D_r	set of known relevant doc vectors
D_{nr}	set of known irrelevant doc vectors



- 新查询向相关文档靠拢而远离非相关文档
- α vs. β/γ 设置中的折中: 如果判定的文档数目很多, 那么 β/γ 可以考虑设置得大一些
- 计算过程中若出现负的权重一律设为0

5、Rocchio算法

Rocchio 算法示例

query vector = $\alpha \cdot$ original query vector
+ $\beta \cdot$ positive feedback vector
- $\gamma \cdot$ negative feedback vector

Typically, $\gamma < \beta$

Original query

0	4	0	8	0	0
---	---	---	---	---	---

 $\alpha = 1.0$

0	4	0	8	0	0
---	---	---	---	---	---

Positive Feedback

2	4	8	0	0	2
---	---	---	---	---	---

 $\beta = 0.5$

1	2	4	0	0	1
---	---	---	---	---	---

 (+)

Negative feedback

8	0	4	4	0	16
---	---	---	---	---	----


 $\gamma = 0.25$

2	0	1	1	0	4
---	---	---	---	---	---

 (-)

New query

-1	6	3	7	0	-3
----	---	---	---	---	----



0	6	3	7	0	0
---	---	---	---	---	---

5、Ricchio算法

- 正反馈 (positive) vs. 负反馈 (negative)
 - 正反馈价值往往大于负反馈
 - 比如，可以通过设置 $\beta = 0.75$, $\gamma = 0.25$ 来给正反馈更大的权重
 - 很多系统甚至只允许正反馈，即 $\gamma = 0$
 - ◆ It's harder for user to give negative feedback
 - ◆ Use only the docs that were positively marked
 - ◆ Users can be expected to review results and to take time to iterate

6、相关性反馈中的假设

- **A1:** 对于某初始查询，用户知道在文档集中使用哪些词项来表达
- **A2:** 相关文档中出现的词项类似 (因此，可以基于相关反馈，从一篇相关文档跳到另一篇相关文档)

6、相关性反馈中的假设

■ 假设A1不成立的情形

- 假设 A1: 对于某初始查询，用户知道在文档集中使用哪些词项来表达
- 不成立的情况：用户的词汇表和文档集的词汇表不匹配
 - ◆ 例如： plane / aircraft; laptop / notebook

6、相关性反馈中的假设

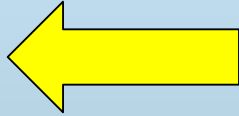
■ 假设A2不成立的情形

- 假设A2: 相关文档中出现的词项类似
- 假设不成立的查询例子
 - ◆ 文档子集之间使用不同的词汇表
 - **Burma / Myanmar(缅甸)**
 - ◆ 查询的答案集合本身需要不同类的文档组成
 - **Pop stars that worked at Burger King**
 - ◆ 通用概念，通常以多个具体概念出现
 - **Felines 猫科**

7、相关性反馈存在的问题

- 相关反馈开销很大
- 相关反馈生成的新查询往往很长
- 长查询的处理开销很大
- 用户不愿意提供显式的相关反馈
- 有时很难理解，为什么会返回(应用相关反馈之后)某篇特定文档

Where are we?

- 查询表达
- 相关性反馈
 - **Explicit Feedback**
 - **Implicit Feedback** 
 - **Pseudo Feedback**
- 查询扩展

8、隐式相关性反馈

- 通过观察用户对当前检索结果采取的行为来给出对检索结果的相关性判定。
- 判定不一定很准确，但是省却了用户的显式参与过程。
- 对用户行为的分析也可以用于个性化信息检索 (Personalized IR)和数据挖掘。



04 大赛时间	06 奖项设置
2013年9月1日——2013年12月6日	一等奖：1个，奖励人民币10万元
报名时间：2013年9月1日——2013年10月30日	二等奖：3-5个，奖励人民币3万元
提交作品：2013年9月15日——2013年11月15日	三等奖：10-15个，奖励人民币1万元
作品评选：2013年11月16日——2013年11月25日	
作品终审：2013年11月26日——2013年12月1日	
结果公布：2013年12月6日	

8、隐式相关性反馈

■ 用户行为的种类

● 鼠标键盘动作：

- ◆ 点击链接、加入收藏夹、拷贝粘贴、停留、翻页等等

● 用户眼球动作

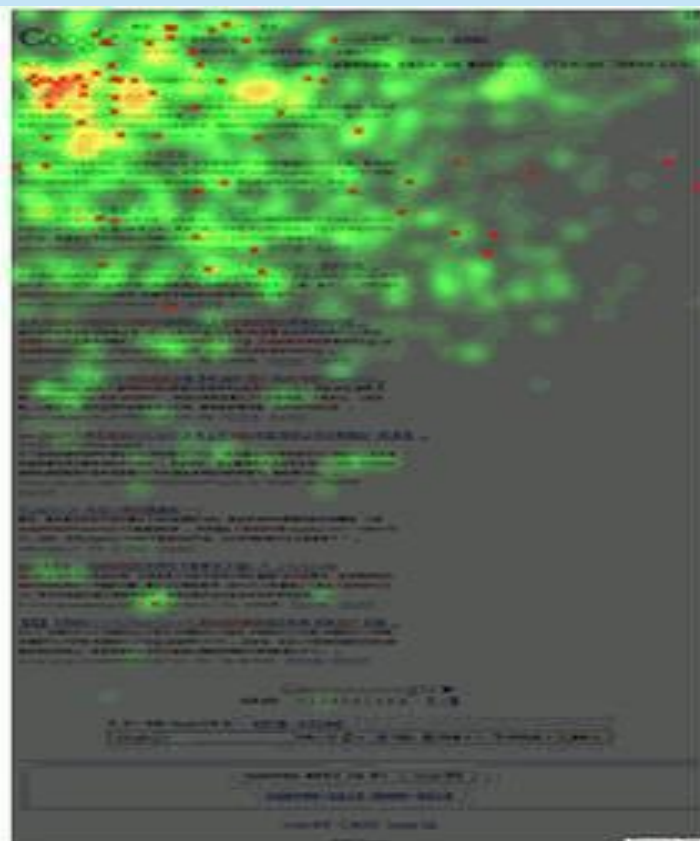
- ◆ **Eye tracking**可以跟踪用户的眼球动作
- ◆ 拉近、拉远、瞟、凝视、往某个方向转

眼球动作(通过鼠标轨迹模拟)

■ 视觉注意机制

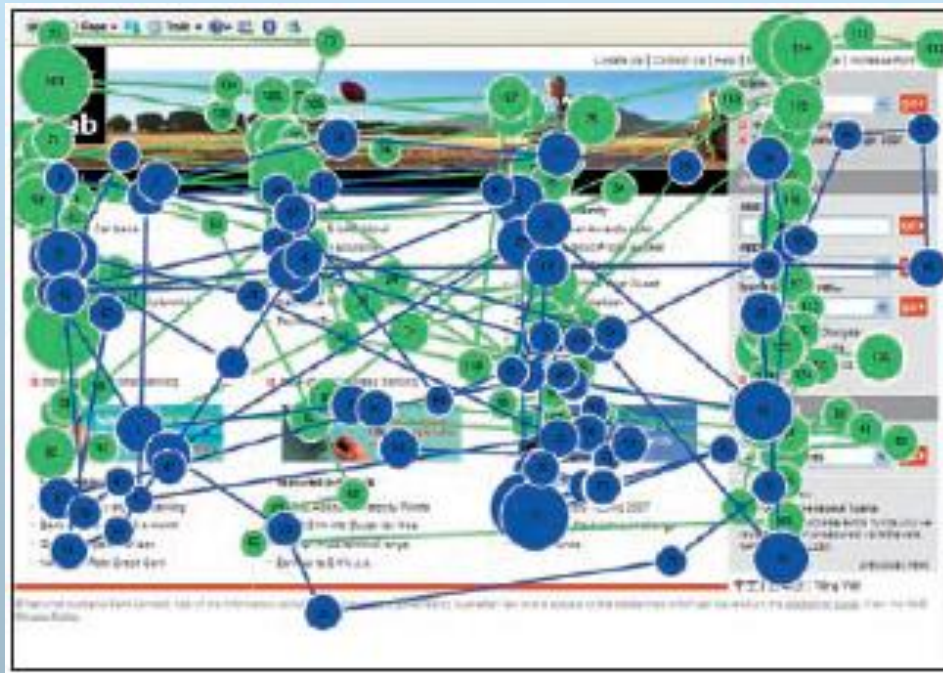


Baidu



Google

眼球动作(通过Eye tracking设备)



8、隐式相关性反馈

■ 优点

- 不需要用户显式参与，减轻用户负担
- 用户行为某种程度上反映用户的兴趣，具有可行性

■ 缺点

- 对行为分析有较高要求
- 准确度不一定能保证
- 某些情况下需要增加额外设备

9、伪相关性反馈

- **Pseudo-relevance feedback**
- 伪相关反馈对于真实相关反馈的人工部分进行自动化
 - 对于用户查询返回有序的检索结果，假定前 k 篇文档是相关的
 - 进行相关反馈 (如 **Rocchio**)
- 平均上效果不错
- 但是对于某些查询而言可能结果很差
- 几次循环之后可能会导致查询漂移(**query drift**)
 - 例如，如果需要查询的是铜矿，而且位于前面的一些文档都是关于智利的铜矿，那么在查询方向上会逐渐偏向于那些与智利有关的文档。

9、伪相关性反馈

■ 优点

- 不用考虑用户的因素，处理简单
- 很多实验也取得了较好效果

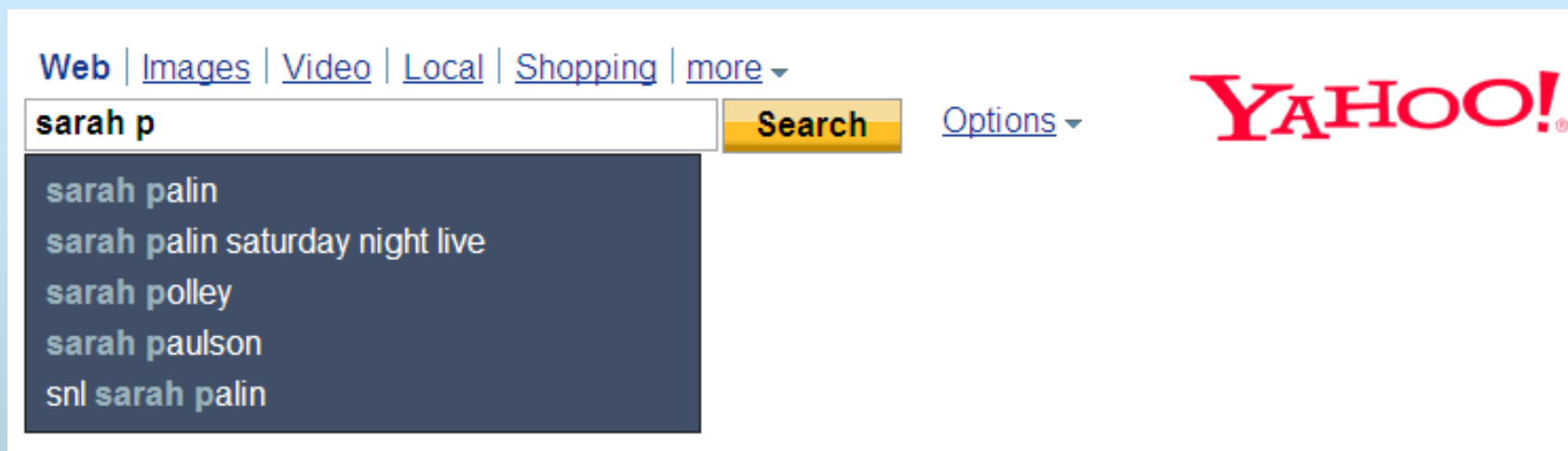
■ 缺点

- 没有通过用户判断，所以准确率难以保证
- 不是所有的查询都会提高效果

三、查询扩展

- In **relevance feedback**, users give additional input (relevant/non-relevant) on **documents**, which is used to reweight terms in the documents
- In **query expansion**, users give additional input (good/bad search term) on **words or phrases**

1、查询扩展示例



2、查询扩展类型

■ Manual thesaurus

- 人工构建同(近)义词词典
- 如 PubMed

■ Automatically derived thesaurus

- 自动导出同(近)义词词典
- 比如，基于词语的共现统计信息

■ Refinements based on query log mining

- 基于查询日志挖掘出的查询等价类
- Web上很普遍

3、人工同义词典例子

PubMed提供生物医学方面的论文文献搜索服务

The screenshot displays the PubMed website interface. At the top, there are logos for NCBI, PubMed, and the National Library of Medicine (NLM). Below the logos, a navigation bar includes links for PubMed, Nucleotide, Protein, Genome, Structure, PopSet, and Taxonomy. The main search area features a search bar with the text 'cancer' and buttons for 'Go' and 'Clear'. Below the search bar, there are tabs for 'Limits', 'Preview/Index', 'History', 'Clipboard', and 'Details'. On the left side, there is a sidebar with links for 'About Entrez', 'Text Version', 'Entrez PubMed', 'Overview', 'Help | FAQ', 'Tutorial', 'New/Noteworthy', 'E-Utilities', 'PubMed Services', 'Journals Database', 'MeSH Browser', 'Single Citation', and 'MetaBox'. The main content area shows the 'PubMed Query:' section with the query text: `("neoplasms"[MeSH Terms] OR cancer[Text Word])`. At the bottom of the main content area, there are buttons for 'Search' and 'URL'.

4、Thesaurus自动构建

- 通过分析文档集中的词项分布来自动生成thesaurus, 基本的想法是计算词语之间的相似度
- **Definition 1: Two words are similar if they co-occur with similar words.**
 - 例如: “car” \approx “motorcycle”, 因为它们都与 “road”、“gas” 及 “license”之类的词共现, 因此它们相似
- **Definition 2: Two words are similar if they occur in a given grammatical relation with the same words.**
 - 例如: Entities that are grown, cooked, eaten, and digested are more likely to be food items, and thus are similar.

4、Thesaurus自动构建

■ 基于共现关系的thesaurus示例

词语	同(近)义词
absolutely	absurd whatsoever totally exactly nothing
bottomed	dip copper drops topped slide trimmed
captivating	shimmer stunningly superbly plucky witty
doghouse	dog porch crawling beside downstairs
makeup	repellent lotion glossy sunscreen skin gel
mediating	reconciliation negotiate case conciliation
keeping	hoping bring wiping could some would
lithographs	drawings Picasso Dali sculptures Gauguin
pathogens	toxins bacteria organisms bacterial parasite
senses	grasp psyche truly clumsy naive innate

5、基于搜索日志的查询扩展

- query log是搜索引擎查询扩展的主要方式
 - 例 1: 提交查询 [herbs] (草药)后, 用户常常搜索[herbal remedies] (草本疗法)
 - ◆ “herbal remedies” 是 “herb”的潜在扩展查询
 - 例 2: 用户搜索 [flower pix] 时常常点击URL photobucket.com/flower, 而用户搜索 [flower clipart] 常常点击同样的URL
 - ◆ “flower clipart”和 “flower pix” 可能互为扩展查询

本章小结

■ 查询表达的难点

■ 相关性反馈(Relevance Feedback)

- 在初始检索结果的基础上, 通过用户指定哪些文档相关或不相关, 然后改进检索的结果
- 最著名的相关性反馈方法: **Rocchio**

■ 查询扩展(Query Expansion)

- 通过在查询中加入同义或相关的词项来提高检索结果
- 相关词项的来源: 人工编辑的同义词词典、自动构造的同义词词典、查询日志等等。