

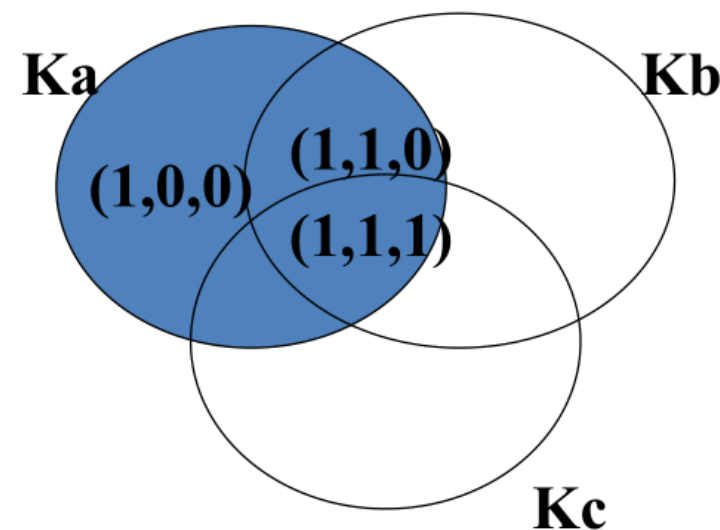
Web信息处理与应用

第五节 查 询

徐童 2021.10.11

- **布尔检索的概念**

- 在布尔检索中，文档被表示为**关键词的集合**。
- 所有的查询式都被表示为关键词的**布尔组合**。
 - 采用“与、或、非”关系加以连接
- 相关度计算
 - 一个文档当且仅当它能够满足布尔查询式时，才会将其检索出来。
 - 检索策略是**二值匹配**。



- 布尔检索的优缺点

优点

- 查询简单，易于理解
- 使用布尔表达式，可以方便地控制查询结果
- 可通过扩展来包含更多功能

缺点

- 功能较弱，不支持部分匹配
- 所有匹配文档均返回，不考虑权重和排序
- 很难进行自动的相关性反馈

- 布尔检索的重要局限性

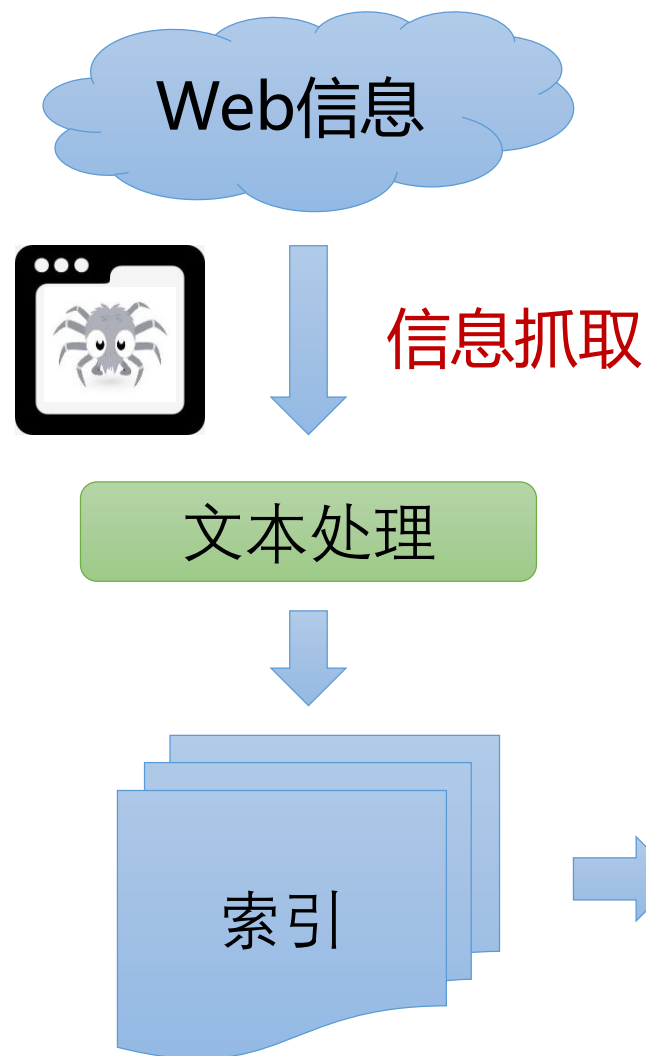
盛宴 or 饥荒

- 需要花费大量的精力设计查询条件（Query），才能获得较为合适的结果
 - 搜索“中国科学技术大学”，可以得到将近3000万条结果。
 - 搜索“中国科学技术大学的XX老师”，结果无限趋近于0
 - 如何获得数量适中、内容符合需求的查询结果？

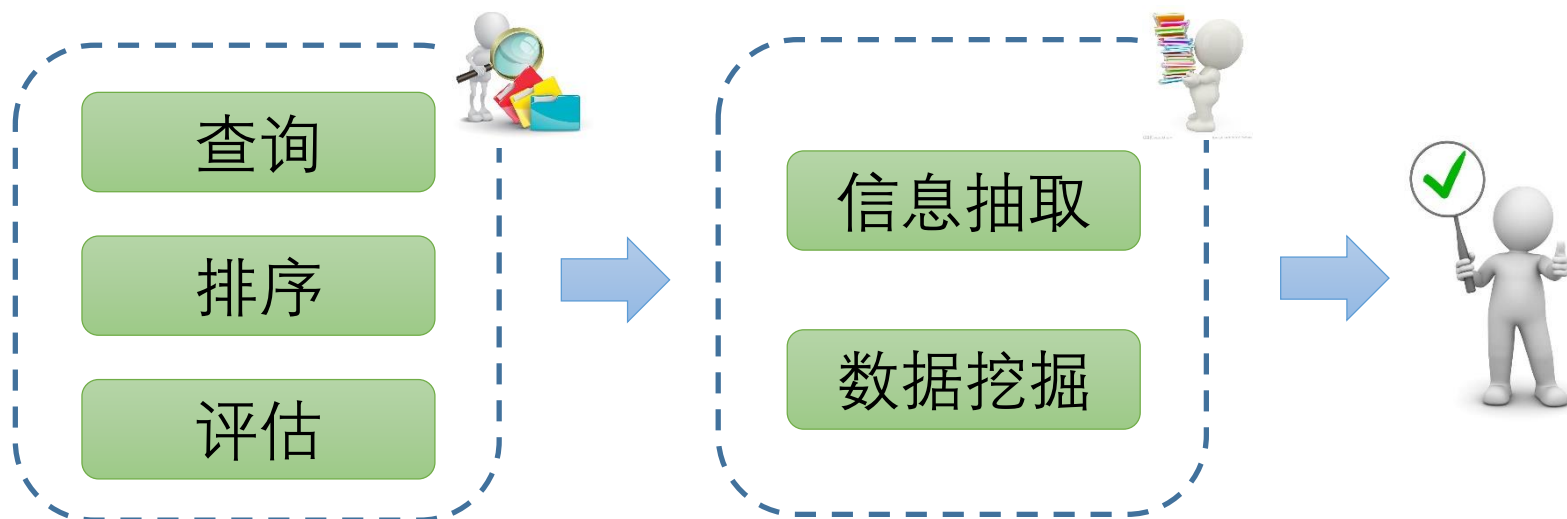
- **更一般的检索方式**
- 采用排序方式代替严格匹配模式
 - 在排序检索中，系统根据文档与查询的**相关性排序**返回文档集合中的合适文档，而不是简单匹配查询条件
 - 自由文本查询：用户查询条件是**自然语言描述**，而不是由查询词项构造的表达式。
- 当系统给出的是**有序**的查询结果时，结果数目将不再是个问题
 - 着眼于给出Top N结果，而不是完整结果



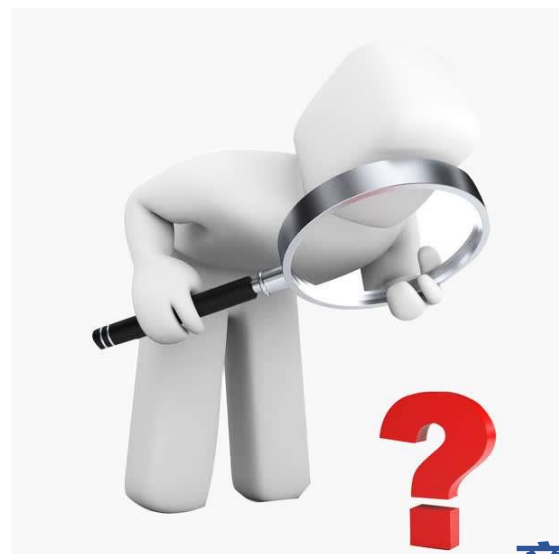
- 本课程所要解决的问题



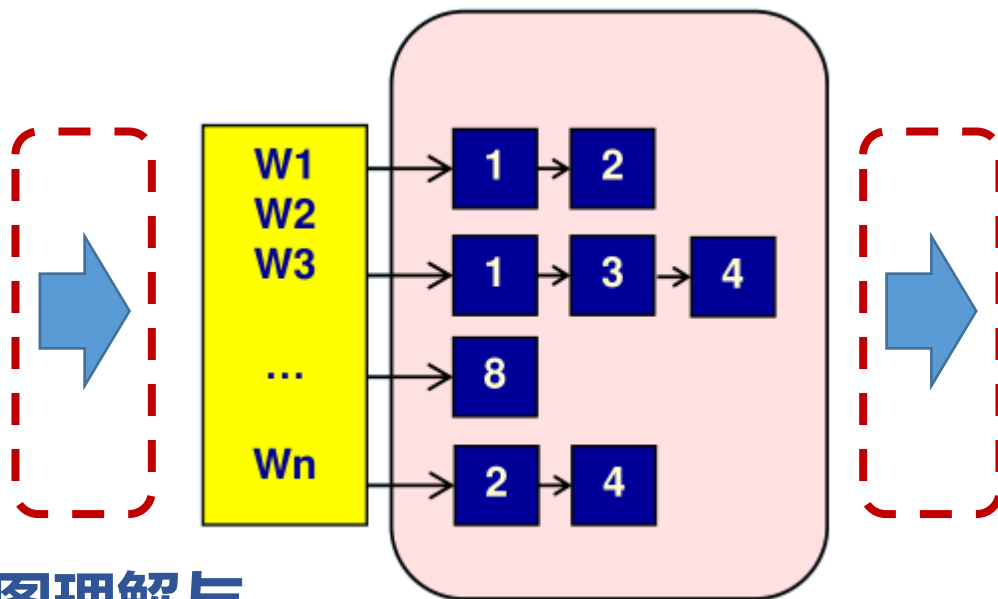
第四个问题：
如何有效理解与响应
用户的查询条件？



• 查询理解所涉及的环节



意图理解与
查询优化



查询结果改进



满足用户需求的第一步，在于准确理解用户的查询意图

- 查询表达理解
- 相关性反馈
 - 常用技术
 - 反馈分类
- 查询扩展
- 情境感知的查询理解

- **查询表达的重要性**
- 从信息检索定义的视角看查询条件的重要性
 - 信息检索：给定查询条件，从文档集合中找出与查询条件相关的文档
 - 查询条件：用户对信息需求的表达
 - 文档集合：待检索的文档库
 - 相关度：返回文档对于信息需求的满足程度



- 为什么查询条件难以理解?
- 用户表达的精简性和歧义性
 - 常用简单词汇表达查询需求, 缺乏精准描述



[动物世界 央视网\(cctv.com\)](#)

2019年9月28日 - CCTV-3综艺频道《动物世界》《动物世界》栏目已经走过20多年,通过专家的讲述、优美的画面、感人的故事去告诉观众、打动观众,使观众认识到我们不能没有...

[tv.cctv.com/lm/dw...](#) - 百度快照

[动物世界|动物世界全集视频 - CCTV1直播网](#)



栏目标题:动物世界 播放频道:CCTV-1综合 播出时间:每天00:20(除周二) 持续时间:30分钟 栏目介绍:《动物世界》栏目于1981年12月31日开播,主旨在于向电视观众介绍...

[www.cctv1zhibo.com/don...](#) - 百度快照

or



- 为什么查询条件难以理解?

- 用户表述方式的差异性

- 可能是用户所用词汇与索引词项的差异，如同义词、方言等
- 也可能是表述方式的糟糕，或者信息错漏导致的误导

钢铁锅,含眼泪喊修瓢锅 这是什么歌? 🍵 50

我来答

分享

举报

浏览 168546 次

39个回答

#热议# 等的就是你! 有奖内测即将开始!



热心网友

2018-10-16

《海阔天空》

演唱: Beyond

- 为什么查询条件难以理解？
- 用户表达中可能存在侧重点
 - 然而，侧重点无法直接从字面意义上看出



- 理解用户查询的几种方式

- 最基本的途径：基于查询的自然语言处理
- 引入相关性反馈
 - 用户直接对查询结果进行评价
- 引导用户表达真实查询意图（查询建议与查询扩展）
- 借助其他信息，完善对于用户的理解
 - 用户间接性反馈
 - 情境信息与情境建模



- 查询表达理解
- **相关性反馈**
 - 常用技术
 - 反馈分类
- 查询扩展
- 情境感知的查询理解

- 何为相关性反馈 (Relevant feedback)

- 用户在查询后标记相关/不相关，然后迭代更新查询，以获得更好的结果
- 相关性反馈的动机
 - 你也许无法表达想要找的内容，但是你至少能够判断所看到的内容
 - “为我提供更多 *相似的文档*.....”
 - 精准的查询条件或许无法一蹴而就，但可以通过迭代逐渐趋于精准

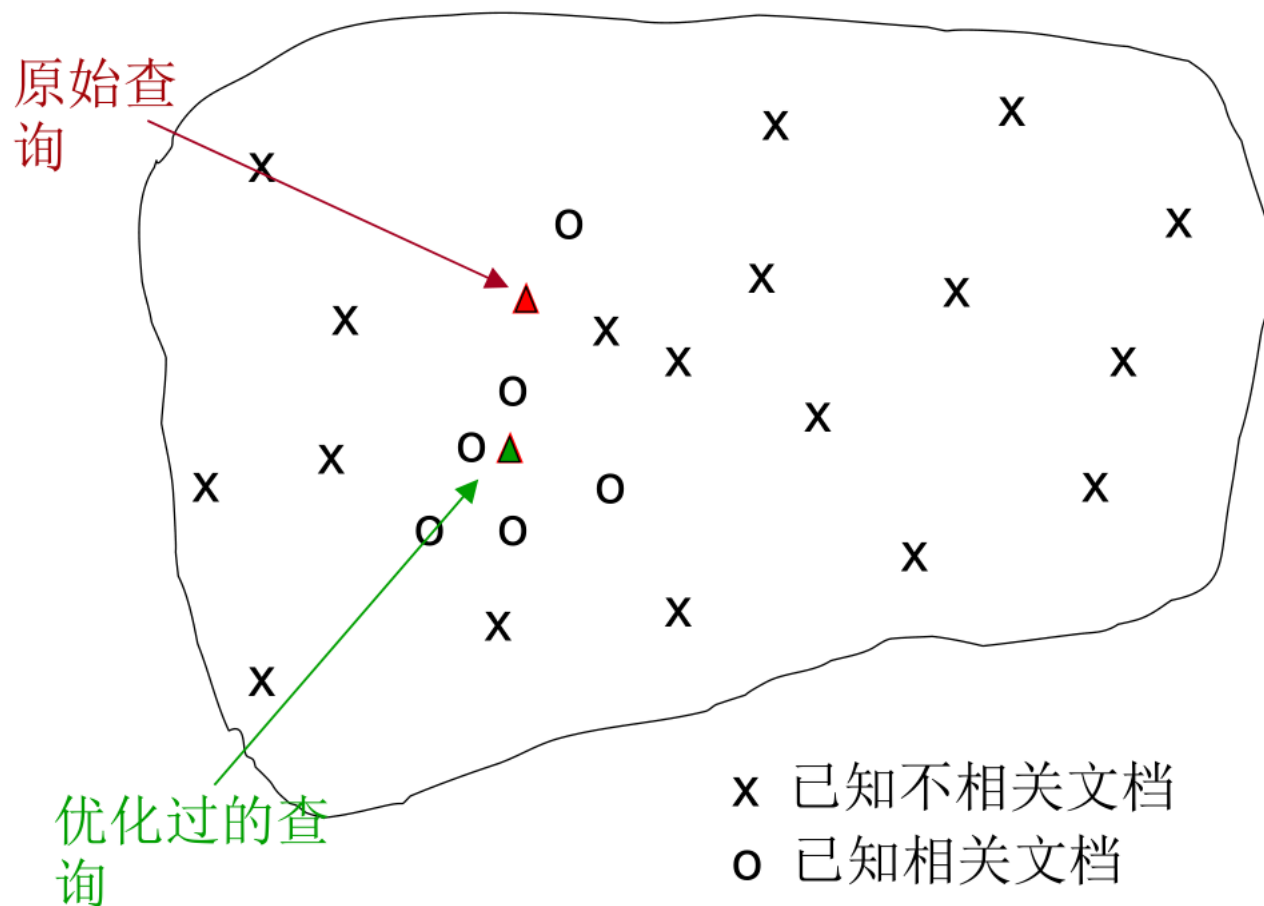


• 相关性反馈的基本流程

1. 首先，用户提出一个查询条件（Query）
 2. 对于返回的文档，用户标出相关与不相关的部分
 3. 系统根据用户反馈，获得用户信息需求更为准确的描述
 - a) 基于相关性信息，更新查询条件，如为不同词项添加不同权重，或在查询条件中添加新的词项
 - b) 基于新查询条件，获取新的结果文档并再次提交用户进行评估
- 上述过程将根据情况进行一次或多次的迭代，从而不断接近最优查询条件

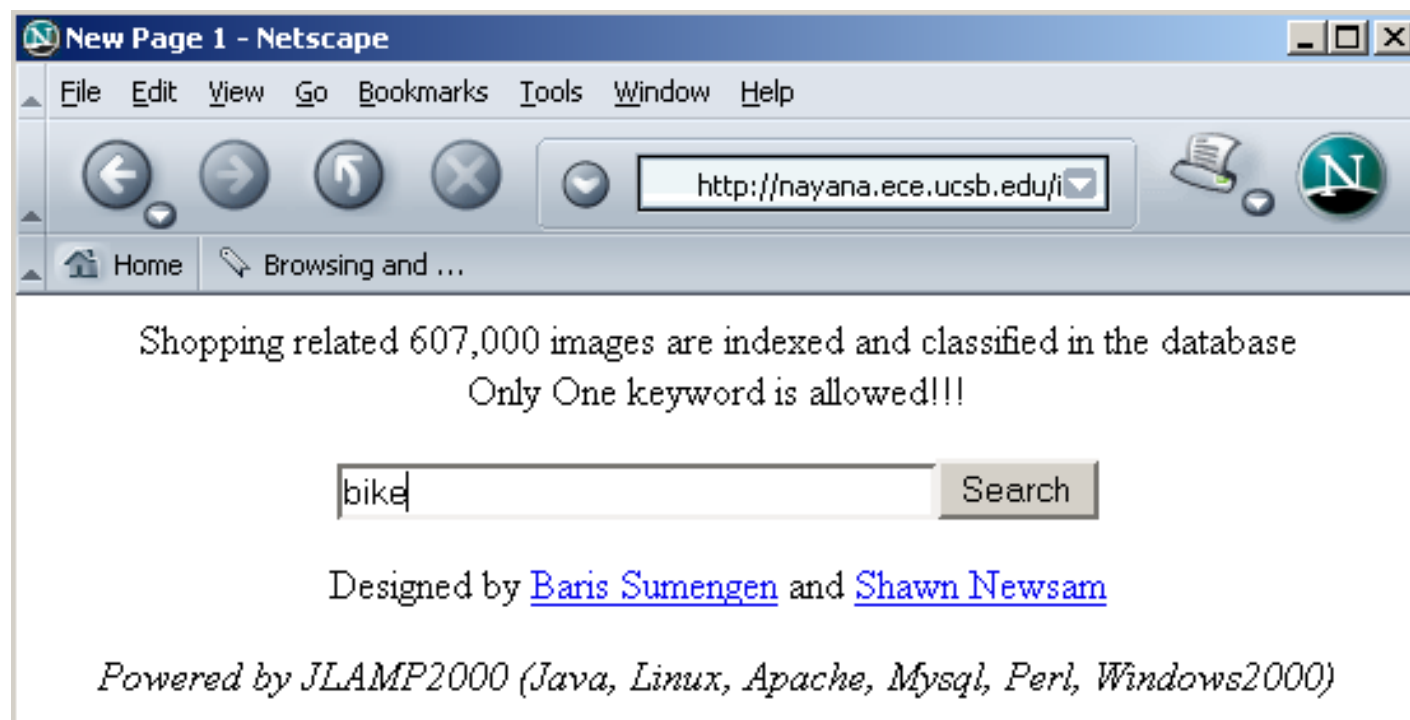
• 相关性反馈的最终目的

- 通过相关性反馈，获得用于表达用户查询意图的最优查询条件
 - 常见方式：为已有词项添加不同权重，或增加新的词项
 - 这一过程应对用户隐藏



• 相关性反馈实例

- 用户的初始查询需求：搜索 “Bike” 相关的图片



- 相关性反馈实例

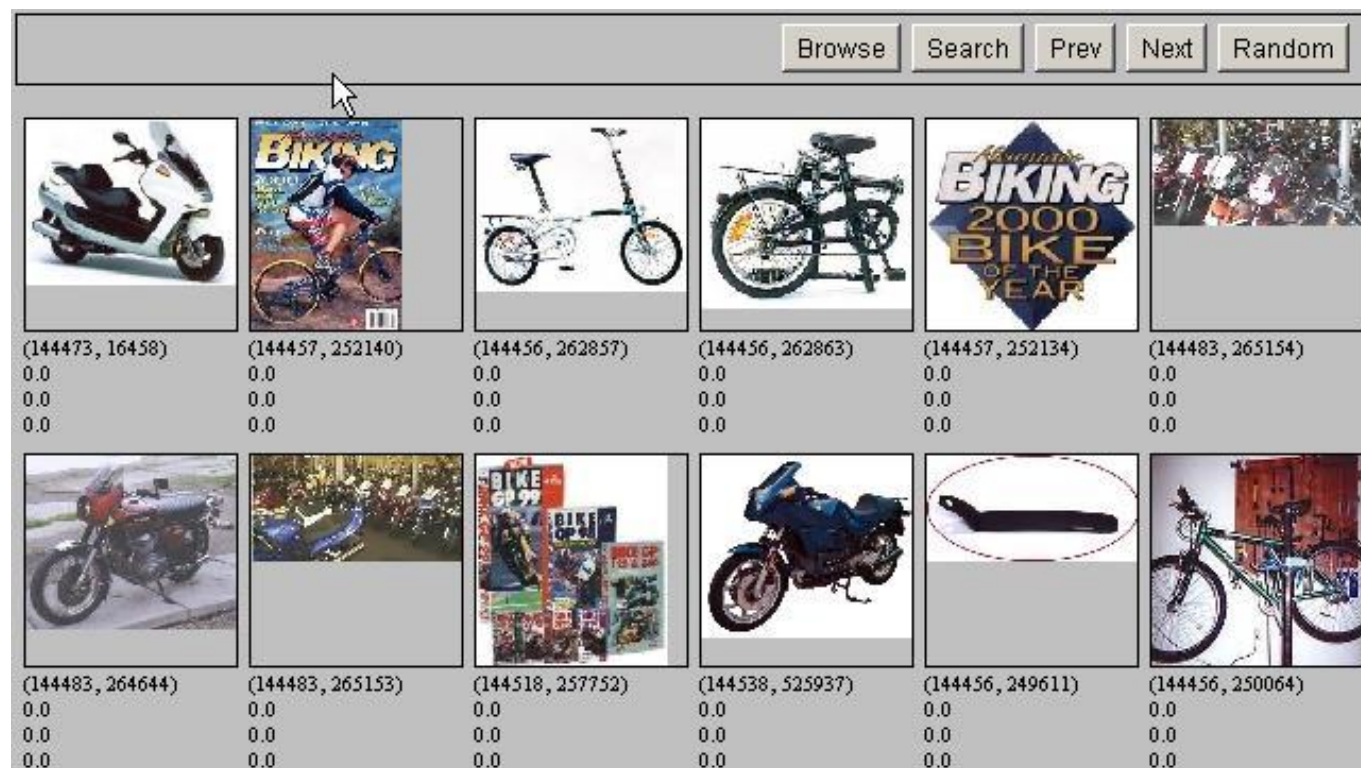
- 用户的初始查询需求：搜索 “Bike” 相关的图片

bike →



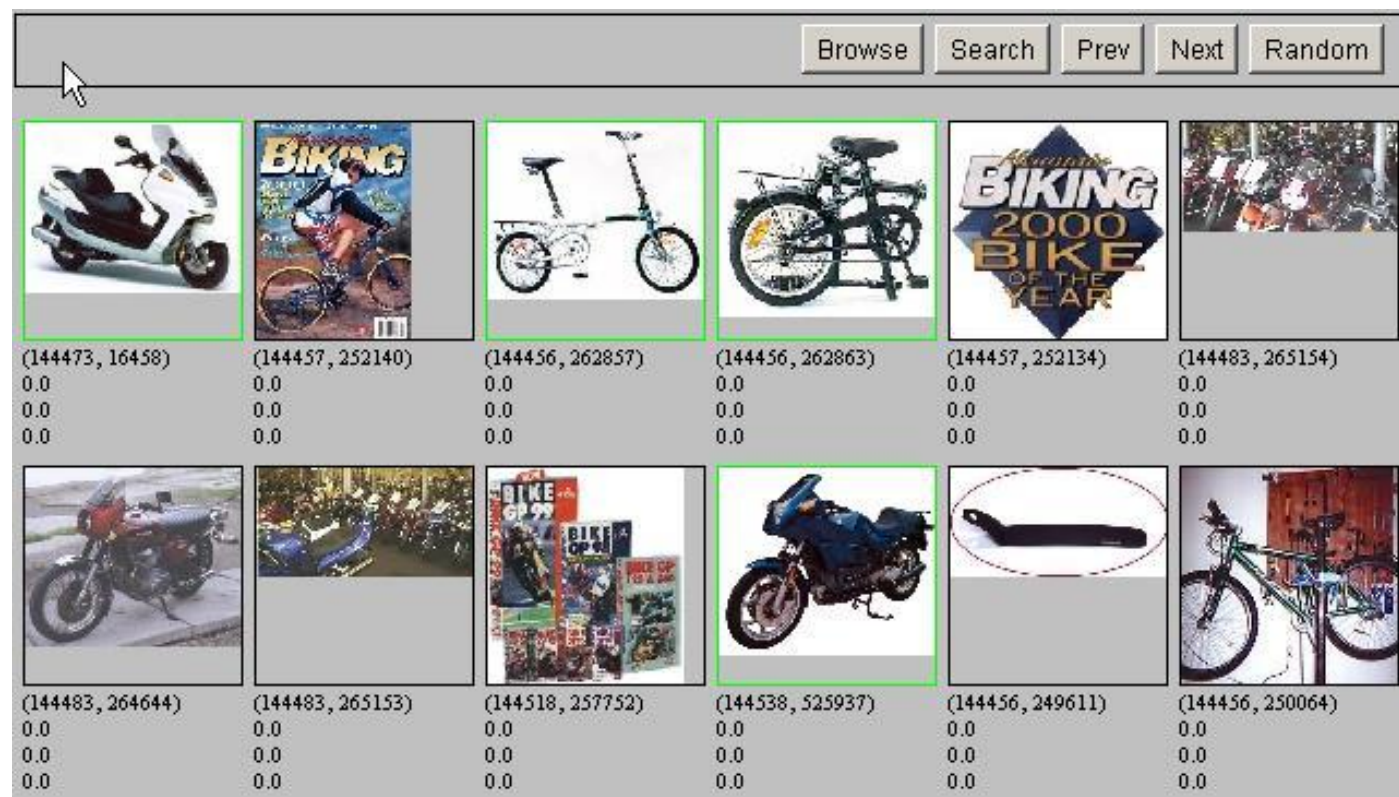
相关性反馈实例

- 基于查询条件的初始检索结果



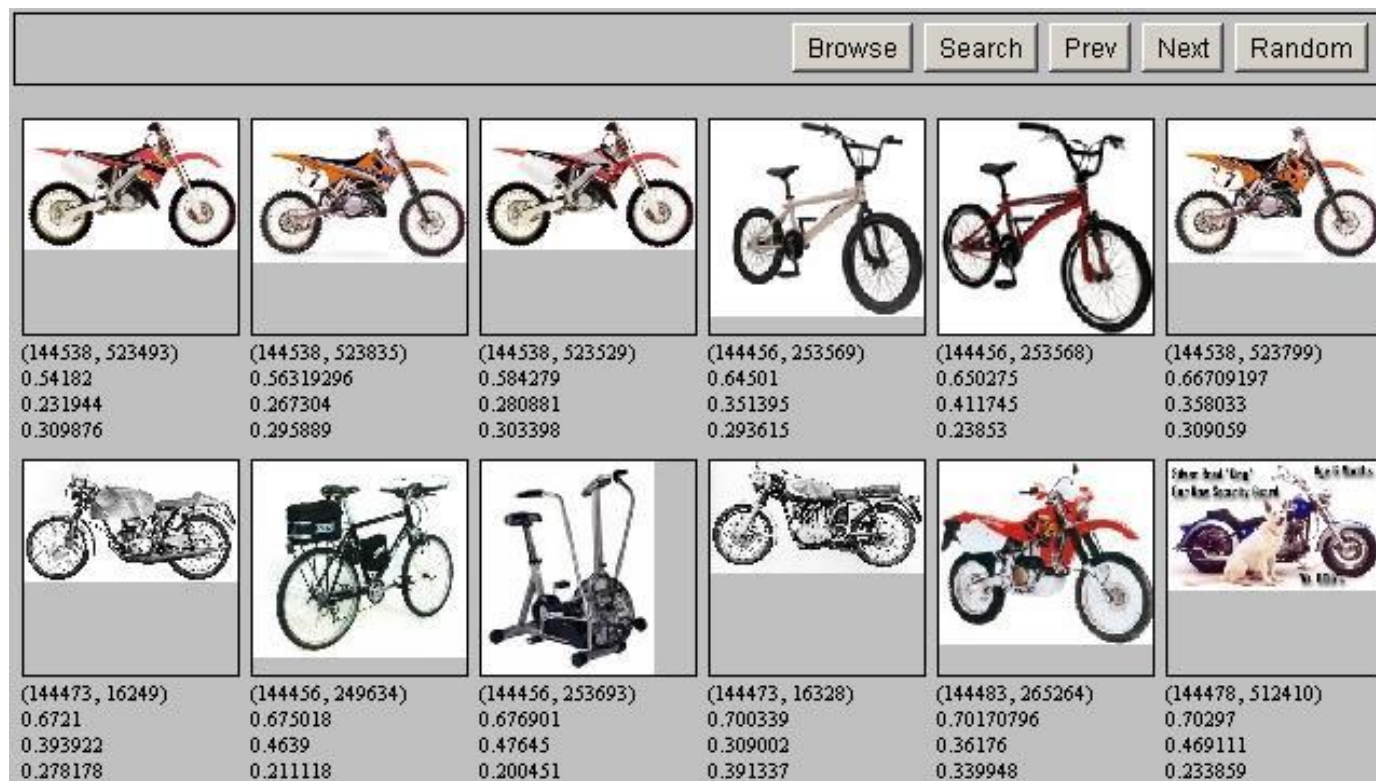
相关性反馈实例

- 用户对部分相关图片进行了标记（或点击等行为）



相关性反馈实例

- 基于用户反馈，得到了更新的搜索结果，以车型图片为主



- 查询表达理解
- 相关性反馈
 - 常用技术
 - 反馈分类
- 查询扩展
- 情境感知的查询理解

- **前提：如何衡量查询与文档之间的相关性？**
- 一个基本的想法：将查询和文档表征为向量，通过相似度评估相关性
- 由此，Salton在SMART系统中提出了著名的向量空间模型
 - 回顾：1960年代，康奈尔大学的Gerard Salton研发了SMART系统，被视作信息检索的鼻祖
- 向量空间模型（Vector Space Model, VSM）
 - 每个文档和查询视作一个词项权重构成的向量
 - 查询时通过比较向量之间相似性来进行匹配



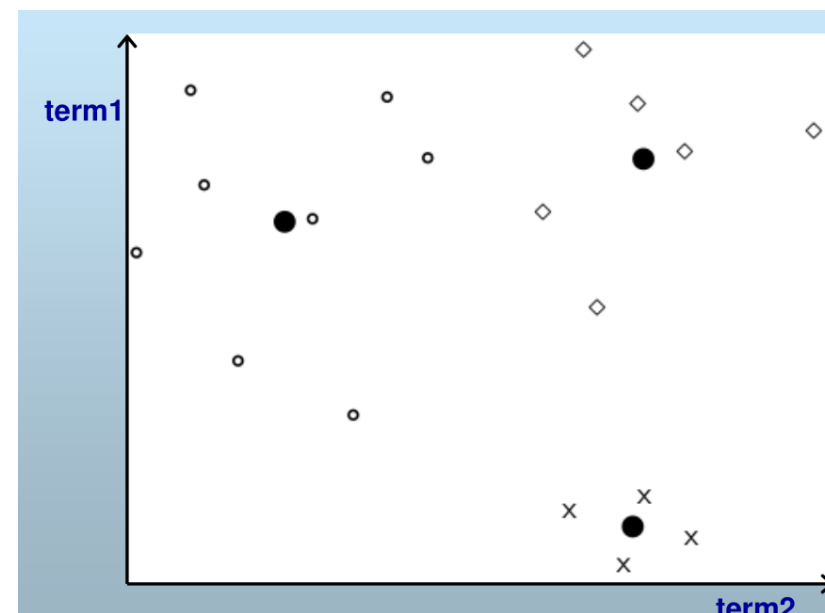
Gerard Salton

- **前提：向量空间模型中的质心概念**

- 某种意义上说，由向量表示的文档，可以视作高维空间中的一个点
- 由此，“质心”就是一系列点（文档）的重心
 - 我们可以用如下公式来计算一类文档的质心

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

- 其中，C是文档的集合



- 基于向量空间模型的Rocchio算法

- 1970年前后，罗基奥（Rocchio）提出了相关性反馈技术，并应用于Salton领导研制的SMART系统中。
- 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办法
 - 其思想在于试图找到一个完美的最优查询向量，以满足以下目标

$$\vec{q}_{opt} = \arg \max_{\vec{q}} [\cos(\vec{q}, \vec{\mu}(C_r)) - \cos(\vec{q}, \vec{\mu}(C_{nr}))]$$

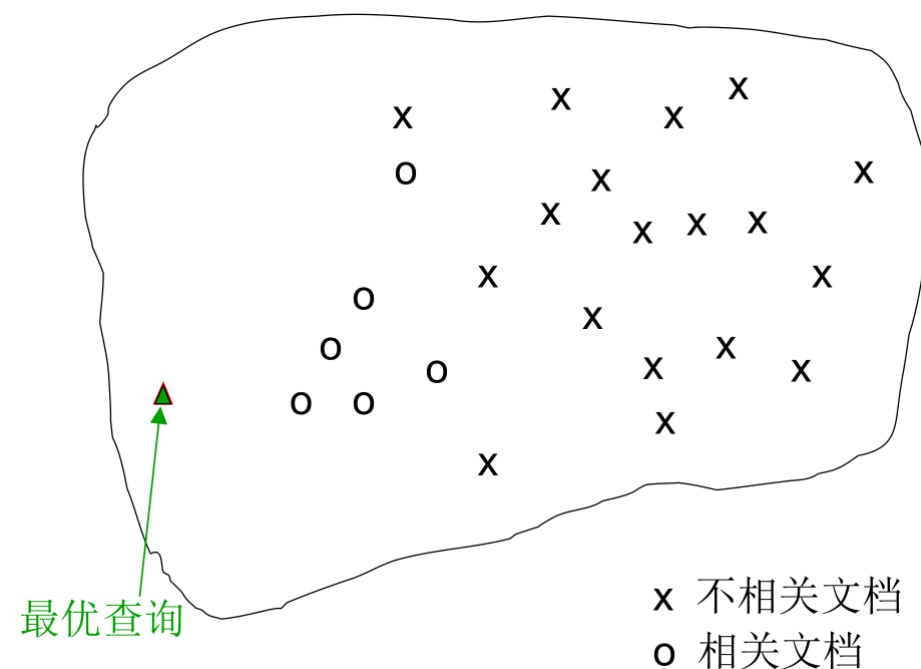
- 即：使得查询尽可能离与之相关的文档更近，离与之不相关的文档更远。

- 基于向量空间模型的Rocchio算法

- 其中，理想情况是，在可知完整的相关/不相关文档集合的情况下，可以使用以下公式来获得一个完美的查询：

$$\vec{q}_{opt} = \frac{1}{|C_r|} \sum_{\vec{d}_j \in C_r} \vec{d}_j - \frac{1}{|C_{nr}|} \sum_{\vec{d}_j \notin C_r} \vec{d}_j$$

- 问题在于，我们事实上并不可能获得完整的相关/不相关文档集合。



- 基于向量空间模型的Rocchio算法

- Rocchio算法（1971）实际使用的近似方法如下：

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

- 其中， D_r 为已知相关文档的向量集合， D_{nr} 为已知不相关文档的向量集合
- q_0 为初始查询向量。 α 、 β 、 γ 为权重，根据手工调节或经验设定
- 由此，新的查询向量将逐渐向相关文档向量移动，远离不相关文档向量

• Rocchio算法示例

Rocchio 算法示例

$$\begin{aligned} \text{query vector} = & \alpha \cdot \text{original query vector} \\ & + \beta \cdot \text{positive feedback vector} \\ & - \gamma \cdot \text{negative feedback vector} \end{aligned}$$

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
---	---	---	---	---	---

- 罗基奥算法与正负反馈

- 正反馈 vs 负反馈
 - 正反馈的价值往往大于负反馈
 - 用户更关心符合需求的标准答案，而不是错误答案
 - 相应的，可以通过设置 $\beta > \gamma$ 来给予正反馈更大的权重
 - 很多系统甚至只允许正反馈，即 $\gamma = 0$
 - 收集真正的负反馈往往比较困难

- **相关性反馈中的基本假设**

- 两个基本假设，可能会影响相关性反馈的性能
- 假设A1：对于某个查询，用户知道在文档集中应使用哪些词项来表达
 - 用户必须用足够的知识来建立一个不错的初始查询
 - 不成立的情况：用户不知道如何表达，或词汇表与文档集词汇表不一致
- 假设A2：相关文档中出现的词项类似，因此可以通过相似文档来相互搜寻
 - 在高维空间中，相似文档应该能够形成一个紧密的簇
 - 不成立的情况：文档词汇表不一致，或是通用概念的不同实体

- **相关性反馈存在的问题**

- 相关性反馈可能影响用户体验
 - 用户不愿意提供显式的相关反馈
 - 用户不希望因为相关性反馈（迭代）而显著延长搜索时间
- 相关反馈生成的新查询往往很长，降低系统效率，增加计算开支
 - 一种做法是只改变重要词项权重而不增加新词项，但效果有限
- 有时很难理解，为什么经过相关性反馈后，会返回不相关的文档
 - 被相关性反馈捕捉的词项，未必是用户需要的内容

- 查询表达理解
- 相关性反馈
 - 常用技术
 - 反馈分类
- 查询扩展
- 情境感知的查询理解

- **常见的相关性反馈类型**

- 显式反馈 (Explicit Feedback)
 - 用户显式地参与交互过程
- 隐式反馈 (Implicit Feedback)
 - 系统追踪用户行为来推测返回文档的相关性
- 伪反馈 (Pseudo Feedback)
 - 在没有用户参与的前提下，直接假设返回结果是相关的，并进行反馈

- 显式反馈

- 最基础的显式反馈：用户点击记录
 - 显而易见的缺陷：只有正样本
 - 用户不点击，不代表完全不相关！
- 拓展的显式反馈：收集负面评价的渠道日益丰富



- 显式反馈

- 更为复杂的显式反馈：用户评论

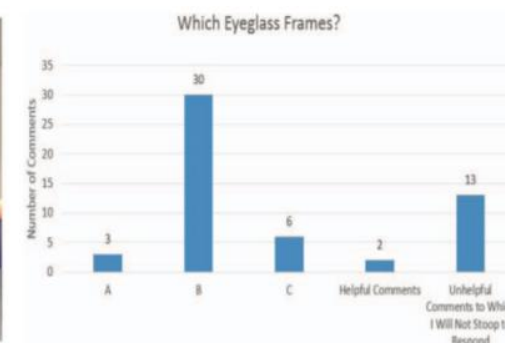


- 基于用户评论，可以收集更为完整的相关性反馈
- 同时，对于网页质量有更为可靠的判断



- 隐式反馈

- 通过观察用户对当前检索结果采取的行为，来判断检索结果的相关性
 - 判定不一定很准确，但省却了用户的显式参与行为
- 常见的用户行为种类
 - 鼠标键盘动作，如点击、停留、翻页、拷贝等
 - 用户眼球动作，如凝视、移动、拉近、拉远等



- 隐式反馈

- 鼠标键盘动作可能揭示用户身份特征
 - 他山之石：《暗算》， “手迹” 识别报务员
 - 不同用户在击键频率、时延、习惯、错误率等方面存在一定差异

[发明专利] 通过键盘鼠标输入习惯识别实现操作用户身份判别的方法 有效

申请号: [CN201110110807.7](#)

文献下载

申请日: 2011-04-29

公开/公告日: [2011-09-14](#)

公开/公告号: CN102184359A

主分类号: [G06F21/00](#)

• 隐式反馈

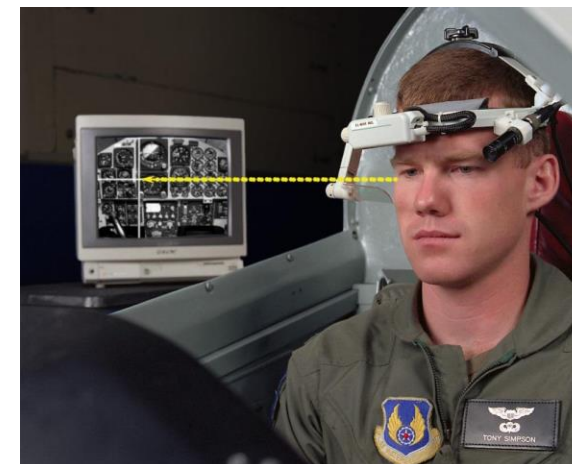
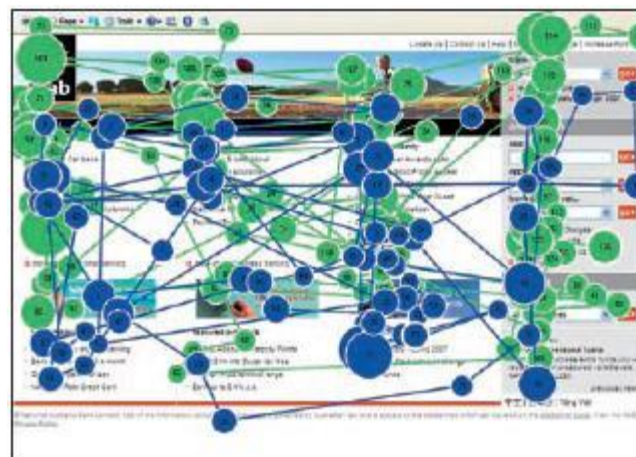
- 用户眼球动作，可以揭示用户关注的内容及关联（视觉注意特征）



Baidu

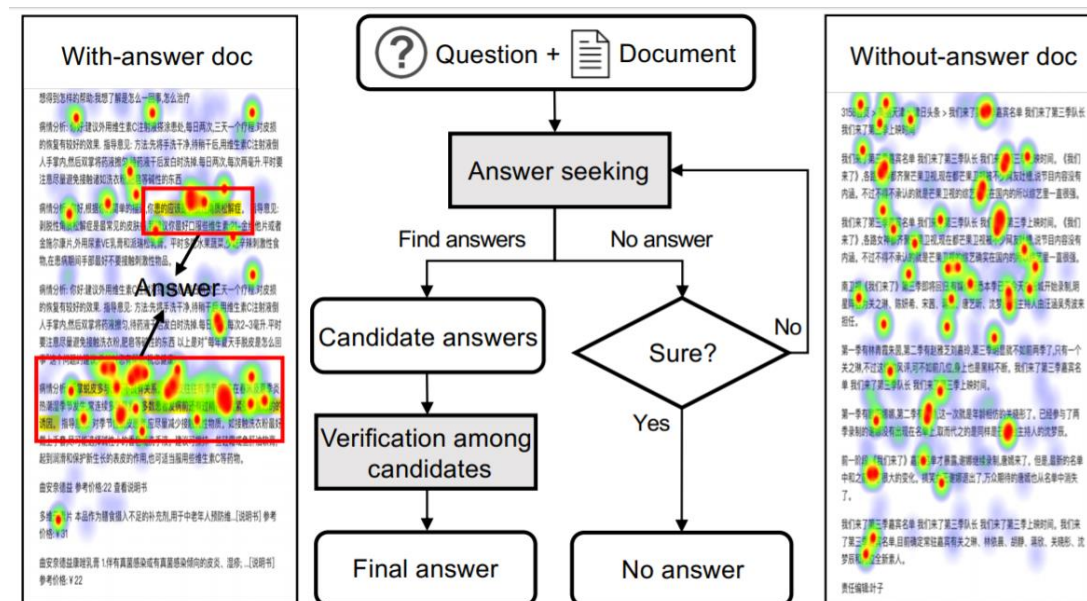


Google



• 隐式反馈

- 借助眼球动作捕捉，还可以支撑其他相关的应用
- 例如，判断文本之间的相关性，甚至揭示问题的答案



• 隐式反馈的优缺点

- 优点：
 - 不需要用户显式参与，减轻用户负担，提升用户体验
 - 用户行为某种程度上可以反应其兴趣，因此具有可行性
- 缺点：
 - 对行为分析有着较高的要求
 - 准确度难以保证
 - 某些情况下需要增加额外设备（且很贵！）

请问tobii眼动仪多少钱可以买到?

我来答

3个回答

#热议# 《平凡的荣耀》搞笑名场面盘点，这部剧讲了什么？



chadbai

Lv6

TA获得超过178个认可 2012-07-05

关注

Tobii X1 10w 左右

Tobii T60 T120 X60 TX300 在几十万 30-50w

Tobii glasses 不详

本回答被提问者和网友采纳

10



评论

分享

举报



匿名用户

2012-11-03

眼动仪国内价格比较隐晦，这些年竞争激烈了，应该价格很便宜了，你可以去找他么国内的总代理打电话问价格去，可以拿别的牌子去压价，感觉应该在几十万。

1



评论

分享

举报

- **伪相关性反馈**

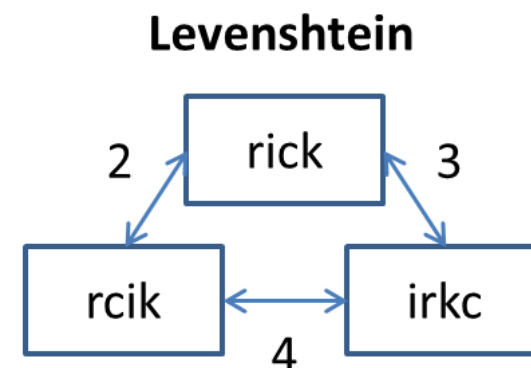
- 无需用户参与反馈过程，而直接根据检索结果自动反馈，较为简单
 - 对于用户查询返回的有序结果，假定前K篇文档是相关的
 - 在此基础上，进行相关性反馈（如借助Rocchio算法）
- 实验证实，利用伪相关性反馈技术可以提升检索的效果
- 但是，伪相关性反馈存在显著的隐患：
 - 结果未经用户判断，难以保证其准确性
 - 某些查询可能结果很差，甚至出现查询漂移（被Top K文档带了节奏）

- 查询表达理解
- 相关性反馈
 - 常用技术
 - 反馈分类
- **查询扩展**
- 情境感知的查询理解

- **拼写错误处理：可视作一种特殊的查询扩展**
- 用户在输入查询条件时，往往容易出现拼写错误（>10%）。
 - 通常采用基于词典或编辑距离的方式进行检查和校对。
 - 某种意义上说，拼写错误检查也是一种查询的建议和完善。



- **复习内容：拼写错误处理的判别方式**
- 通常采用基于词典或编辑距离的方式进行检查和校对。
- 编辑距离 (Levenshtein Distance)
 - 指两个字符串之间转换所最少需要的编辑操作步数。
 - 允许的一步编辑操作包括替换、插入或删除一个字符。
 - 例如：Distance(“Kitten”, “Sitting”) = 3
 - kitten → sitten (substitution of "s" for "k")
 - sitten → sittin (substitution of "i" for "e")
 - sittin → sitting (insertion of "g" at the end).



• 什么是查询扩展

- 在相关性反馈中，用户针对文档是否相关给出反馈，这些反馈将被用来重新计算查询词项的权值。
- 而在查询扩展中，用户针对词项的合适程度给出反馈，这些反馈将被用来构建更为完整的查询条件。
- 暗含的功能：将具有歧义的查询词展示出来，供用户选择和确认。



• 什么是查询扩展

- 事实上，类似查询扩展的服务，在其他领域中也有着广泛的应用
- 例如，人机交互/智能客服中的服务列表




- **查询扩展的实现**

- 利用同义词词典，可以实现查询条件的扩展
 - 对于某个查询词汇，使用词典中的同义词或相关词进行扩展
 - Feline (猫科) → Feline cat
 - 相对于原始的查询词汇，可以给扩展的词汇设定更小的权重
 - 一般而言，查询扩展有助于提升查询的召回率（找得更全）
 - 但是，可能会影响准确率，尤其在扩展词存在歧义的情况下
- **编纂和维护同义词词典需要很大的代价**

查询扩展的实现

- 先前课上讲过的例子：基于同义词的查询扩展


DBLP FILTER



Sort by
 ☐ relevance
 ☒ importance
 ☐ date

Scholar

About 28 results (5.57sec)


 (1998~2013)

Since Time

[Since 2013](#)
[Since 2012](#)
[Since 2009](#)
[Custom range...](#)

Sort By

[Sort By Relevance](#)
[Sort By Importance](#)
[Sort By Date](#)

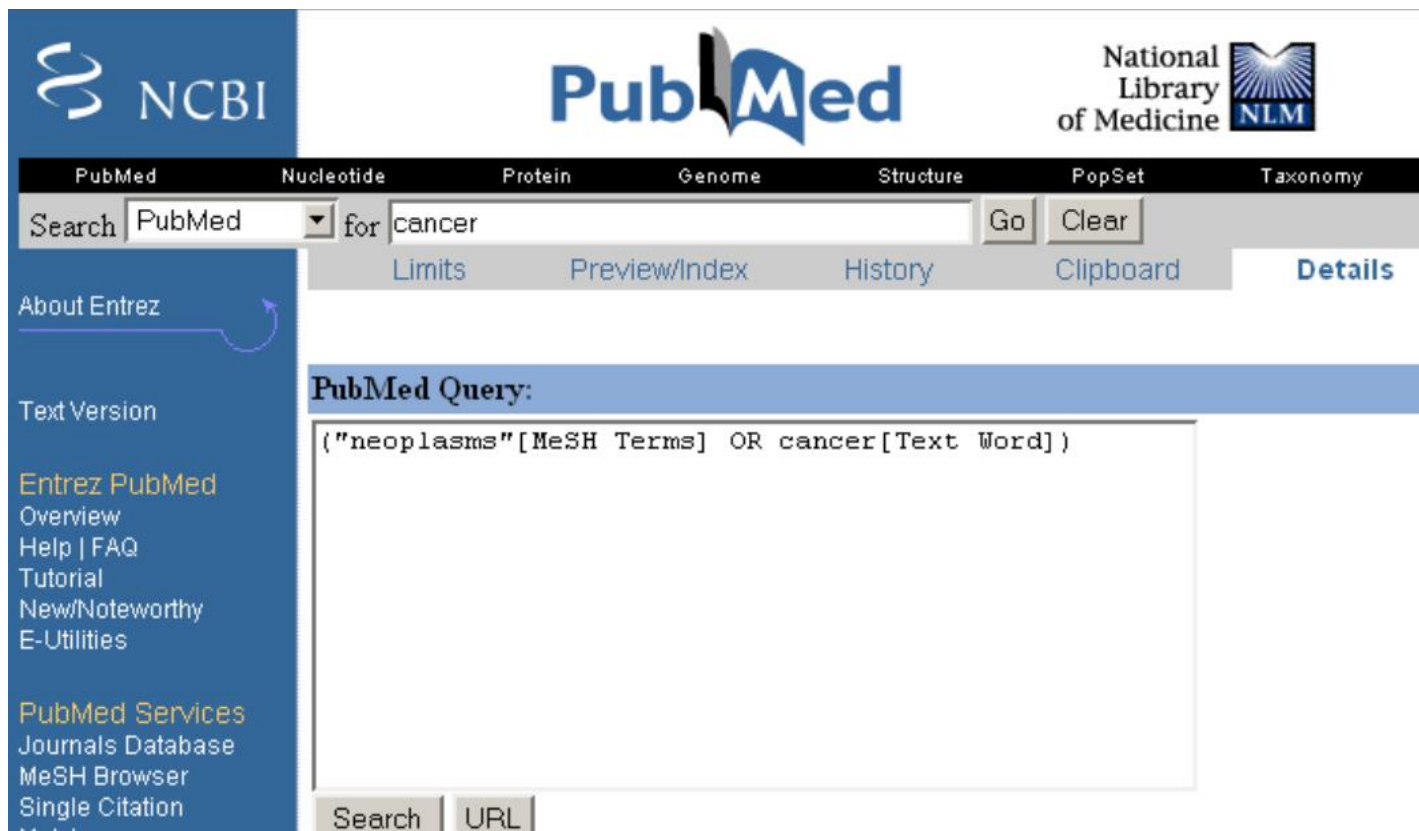
A	EE	Scholar	A Data Model and Data Structures for Moving Objects Databases. (Luca Forlizzi and Ralf Hartmut G and ü) <i>ACM Conference on Management of Data (sigmod) [2000]</i> Cited by 353
A	EE	Scholar	Scientific Data Repositories: Designing for a Moving Target. (Etzard Stolte and Christoph von Praun and Gustavo Alonso) <i>ACM Conference on Management of Data (sigmod) [2003]</i> Cited by 43
A	EE	Scholar	A Data Model for Moving Objects Supporting Aggregation. (Bart Kuijpers and Alejandro A. Vaisman) <i>IEEE International Conference on Data Engineering (ICDE) [2007]</i> Cited by 20
B	EE	Scholar	Spatio-Temporal Data Types: An Approach to Modeling and Querying Moving Objects in Databases. (Martin Erwig and Ralf Hartmut G and ü) <i>GeoInformatica (GeoInformatica) [1999]</i> Cited by 367
B	EE	Scholar	A generic data model for moving objects. (Jianqiu Xu and Ralf Hartmut G and ü) <i>GeoInformatica (GeoInformatica) [2013]</i>
B	EE	Scholar	An Object-Field Perspective Data Model for Moving Geographic Phenomena. (Kyoung-Sook Kim and Yasushi Kiyoki) <i>Database Systems for Advanced Applications (DASFAA) [2010]</i>
C	EE	Scholar	Place: A Distributed Spatio-Temporal Data Stream Management System for Moving Objects. (Xiaopeng Xiong and Hicham G. Elmongui and Xiaoyong Chai) <i>International Conference on Mobile Data Management (MDM) [2007]</i> Cited by 18
C	EE	Scholar	An analytic solution to the alibi query in the space-time prisms model for moving object data. (Bart Kuijpers and Rafael Grimson and Walled Othman) <i>International Journal of Geographical Information Science (IJGIS) [2011]</i> Cited by 3
C	EE	Scholar	A Scaleless Data Model for Direct and Progressive Spatial Query Processing. (Sai Sun and Sham Prasher and Xiaofang Zhou) <i>International Conference on Conceptual Modeling (ER) [2004]</i> Cited by 2
C	EE	Scholar	Efficient Strip-Mode SAR Raw-Data Simulation of Fixed and Moving Targets. (Ozan Dogan and Mesut Kartal) <i>IEEE Geoscience and Remote Sensing Letters (LGRS) [2011]</i>
			Computational data modeling for network-constrained moving objects. (I aurvnas Sneicvs and Christian S. Jensen and Augustas Kliavs) <i>GIS</i>

- **查询扩展的类型**

- 利用人工编纂的同义词词典
 - 例如，第三节“网页文字处理”中提到的How-Net、大词林等
- 全局分析与同义词词典的自动生成
 - 基于统计词汇之间的共现关系（Co-occurrence），自动构建词典
- 基于搜索日志进行优化
 - 通过查询日志，挖掘查询的等价类

- 人工编纂词典的例子

- PubMed, 提供生物医学领域的论文文献搜索服务

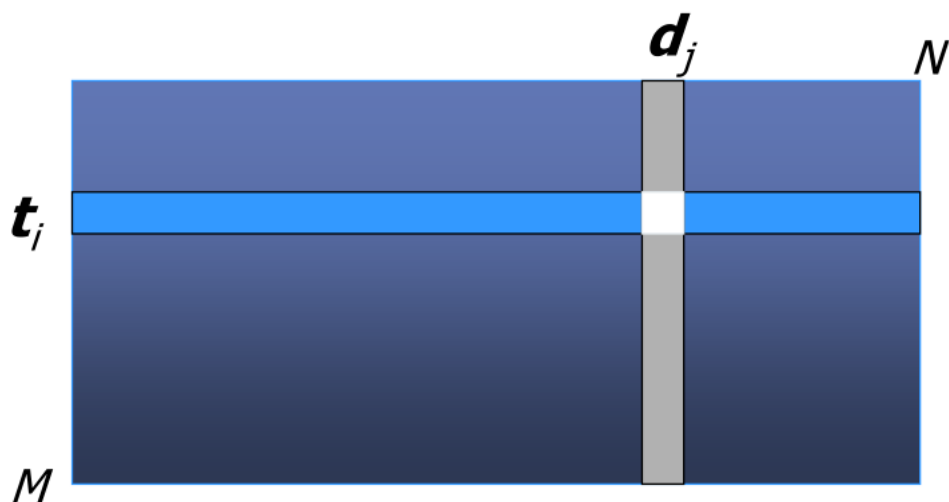


- **自动构建同义词词典的两种思路**

- 通过分析文档集中的词项分布，来自动生成同义词/近义词词典。
- 基本的想法是计算词语之间的相似度，常见的两种思路如下：
 - 思路1：如果两个词经常和相似的词共同出现，则它们很可能是相似的
 - E.g., Car和Motocycle很可能相似，因为它们和road、gas等经常共现
 - 思路2：如果两个词经常与相同的词以一种特定的语义关系共同出现，那么他们很可能是相似的
 - E.g., 可烹调并食用的实体往往都属于食物

- 基于共现的同义词词典构建

- 最简单的方法：给定词项-文档矩阵 A ，计算 $C=AA^T$
 - 对 A 需进行归一化，使行向量的大小为1
 - 对于每个词项，选择 C 中相似度最高的若干词项作为同义词



- **同义词词典质量的讨论**
- 词项关联的质量是一个问题
 - 有歧义的查询词可能导致统计上相关，而意思上不相关的词
 - 如 “Apple Computer” 可能导致 “Apple Red Fruit Computer”
 - 同时，由于扩展的查询词与原查询词高度相关，扩展后的查询也未必能够获得更多的相关文档。

- **基于搜索日志的查询扩展**

- 搜索日志目前是搜索引擎查询扩展的重要方式
 - 实例1：提交查询Herbs（草药）后，用户往往搜索Herbal remedies（草本疗法）。
 - 此时，Herbal Remedies是Herbs的潜在扩展查询
 - 实例2：用户搜索Flower Pix与Flower Clipart时，往往都会点击URL：photobucket.com/flower
 - 此时，Flower Pix与Flower Clipart可能互为潜在扩展查询

- **基于搜索日志的查询扩展**

- 发散思考：基于搜索日志的关联规则扩展
 - 还记得第一章讲过的“啤酒与尿布”的故事吗？
 - 如果两个词项经常共同出现，那么两个词项很可能是相关的
 - 注意，这里的相关不代表任何因果关系
 - 相应的，相关词项可用于搜索的查询扩展
 - 相关技术将在数据挖掘部分介绍

交易号	产品
T01	啤酒
T01	尿布
T02	啤酒
T02	尿布
T03	尿布



- 查询表达理解
- 相关性反馈
 - 常用技术
 - 反馈分类
- 查询扩展
- 情境感知的查询理解

- **回顾：用户查询条件的常见问题**

- 在许多时候，用户输入的查询条件缺乏足够的精准性
 - 用户查询条件存在歧义，难以判断真实意图
 - 用户查询条件过于精简、语义信息不够完整
- 在缺少直接来自用户的反馈时，往往需要借助其他信息来协助判断



- 用户查询条件的常见问题
- 用户查询条件存在歧义，难以判断真实意图



or



- 用户查询条件的常见问题

- 此时，查询上下文能够帮助我们判断用户的真实意图



+ 搜索历史



=



- 用户查询条件的常见问题

- 用户查询条件过于精简、语义信息不够完整



or



- 用户查询条件的常见问题

- 基于查询时的环境信息，可以填补查询条件中的缺失要素



北京 + 非秋季 →



杭州 + 非冬季 →



- 情境的概念与意义

- 在上述两个例子中，我们都是用了其他信息帮助我们理解用户意图。这一类信息被我们称作 “情境信息” (Context Information)
 - 从计算机学科的视角出发，“情境”一词可定义为“所有与人机交互相关，用于区分标定当前特殊场景的信息”。
 - 基于这一定义，服务提供者可借助情境信息，为用户提供更精确的信息检索和过滤服务。

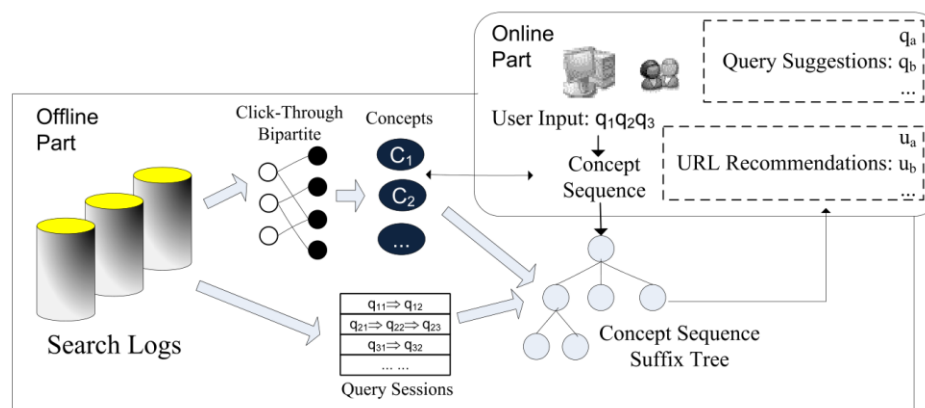
- **搜索中的基础情境：上下文**

- 直观上，查询上下文可以帮助更好的理解用户查询词。
 - 用户的搜索行为往往具有一定的连贯性
 - 相应的，同一查询会话中的查询词和点击的 URL 往往是相关的
 - 一个小问题：如何拆分查询会话？

查询会话 ID	查询会话
S_1	Beautiful mind \Rightarrow Gladiator \Rightarrow Russel Crowe \Rightarrow Russel Crowe movies \downarrow www.imdb.com/title/tt0172495
S_2	Roma \Rightarrow Roma history \Rightarrow Gladiator \downarrow www.exovedate.com/the_real_gladiator_one.html

- **基于上下文感知的搜索：基本流程**

- 线下训练阶段（模型准备阶段）
 - 首先，将查询词归纳为查询概念，从而避免查询词稀疏性的影响
 - 其次，建立模型，描述查询概念之间的关联关系，支撑上下文感知
- 线上服务阶段（感知查询阶段）
 - 首先，切分会话，判断与当前查询相关的上下文查询与点击记录
 - 其次，根据已有查询记录理解用户当前意图，并进行精准查询



- 查询概念归纳

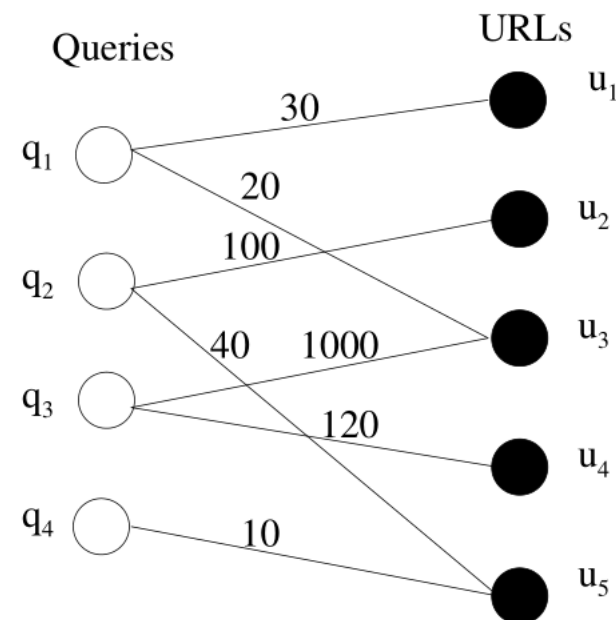
- 用户可能用不同的查询词描述同样的信息需求

- 例如, NRC Beijing和NRC GEL指同一个研究机构 (诺基亚北京研究院)

- 查询概念: 一组有着相同语义的查询词

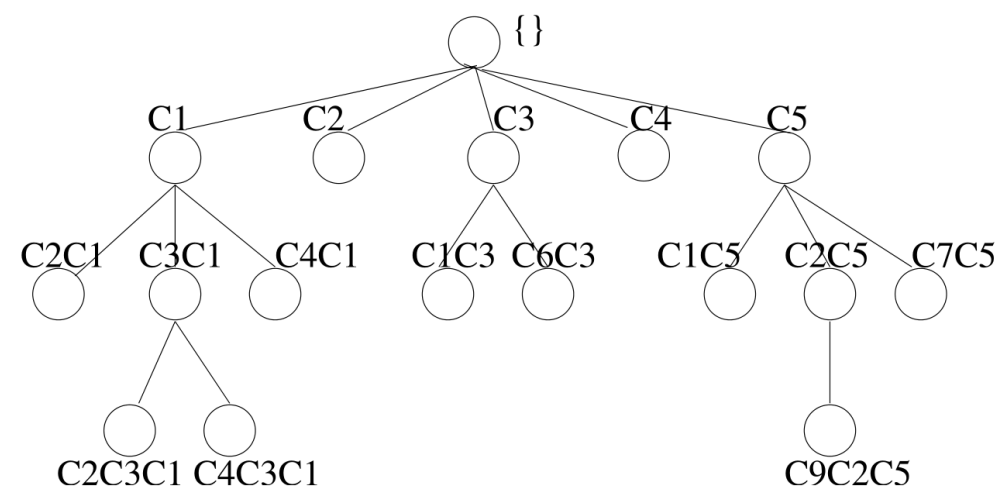
- 可以解决查询词的稀疏性问题
 - 同时, 更规范地解释查询上下文

- 一种启发式方法: Query-URL的二部图聚类



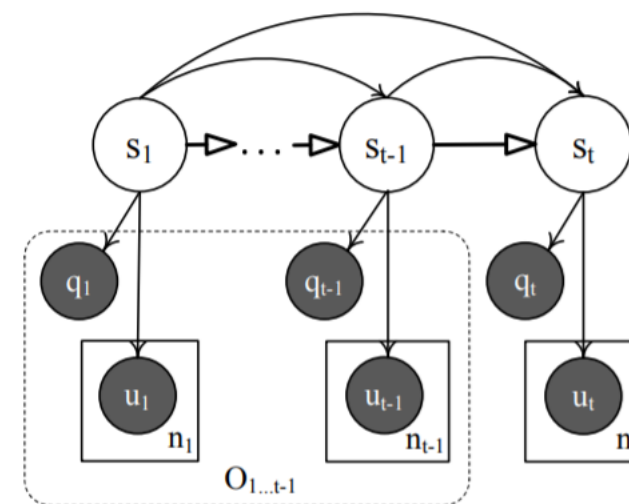
- 基础的上下文感知方法

- 借助查询概念归纳，我们将会话内的查询序列转化为了查询概念序列
- 接下来，核心任务在于如何从序列中抽取查询概念的序列模式
 - 考虑特定长度以内的所有序列模式
 - 保留频率高于阈值的模式，并存储为后缀树 (Suffix Tree)
 - 当面临上下文感知任务时，根据已有序列找到相应节点，从而获得候选查询建议



- 进阶的上下文感知方法

- 前述方法虽然有效利用了上下文信息，但将查询词限制在一个查询意图中，同时仅能推荐查询扩展，而不能帮助判断文档相关性
- 引入隐马尔科夫模型(HMM)，将用户意图 s 视作隐变量，查询上下文 q 与点击记录 u 视作观察值。
- 采用可变长度的HMM模型，避免了一阶HMM模型当前状态仅与前一时刻状态相关的局限性，更灵活地适应不同长度的会话并确保查询之间的相关性。



- 上下文感知的不同类型

- 我们已经证实了上下文有助于更好理解用户的查询意图。然而，上下文信息是否具有不同类型？各种类型所起到的作用有何不同？
- 几种常见的查询上下文类型：
 - 查询重组
 - 查询特化
 - 查询泛化
 - 一般关联



• 上下文感知的不同类型

- 类型一：查询重组
 - 用户的后续查询仅仅只是先前查询的重新表述，目的不变或类似
 - 在此情况下，先前点击的内容往往不再被点击（即使内容相关）

Query 1: "homes for rent in atlanta"		Query 2: "houses for rent in atlanta"	
×	Atlanta homes for rent - home rentals - houses for ren... Rentlist is directory of Atlanta home rentals featuring links to... http://www.rentlist.net		Atlanta homes for rent - home rentals - houses for ren... Rentlist is directory of Atlanta home rentals featuring links to... http://www.rentlist.net
	Homes For Rent, lease in Atlanta suburbs. Can't sell ... Atlanta homes for rent, homes for lease in Gwinnett and north... http://atlantahomesforrent.com		Homes for Rent in Atlanta, GA Houses, Apartments and Homes for Rent in Atlanta, GA Find ... http://www.usrentallistings.com/ga/atlanta
	Rentals.com - Homes for Rent, Apartments, Houses ... Atlanta Home Rentals; Austin Home Rentals; Charlotte Home... http://www.rentals.com		Atlanta Home Rentals, Homes for Rent in Atlanta ... Atlanta Rentals - Homes for Rent in Atlanta, Apartments, Re... http://www.rentals.com/Georgia/Atlanta
×	Atlanta Home Rentals, Homes for Rent in Atlanta ... Atlanta Rentals - Homes for Rent in Atlanta, Apartments, Re... http://www.rentals.com/Georgia/Atlanta		Homes For Rent, lease in Atlanta suburbs. Can't sell ... Atlanta homes for rent, homes for lease in Gwinnett and north... http://atlantahomesforrent.com
	Homes for Rent in Atlanta, GA Houses, Apartments and Homes for Rent in Atlanta, GA Find ... http://www.usrentallistings.com/ga/atlanta	×	Atlanta Homes for Rent, Rental Properties, Houses for ... Search for Homes for Rent in Atlanta, Georgia for free. View li... www.rentalhouses.com/find/GA/AtlantaArea/ATLANTA

• 上下文感知的不同类型

- 类型二：查询特化
 - 在用户的后续查询中，对先前查询中部分内容进行了更为具体、深入的查询
 - 在此情况下，先前查询中较为泛指的内容将被略过

Query 1: "time life music"		Query 2: "time life Christian CDs"	
×	Welcome to TimeLife.com Homepage TimeLife.com: The best in music & video from a name you can... http://www.timelife.com		Welcome to TimeLife.com Homepage Enjoy 138 romantic classics on 9 CDs from top artists like John... http://www.timelife.com
	Time-Life - Wikipedia, the free encyclopedia Time-Life is a creator and direct marketer of books, music, vid... http://en.wikipedia.org/wiki/Time-Life_Music		Time Life Music & Video As Seen On TV Christian ... Time Life Music & Video CD & DVD Collections ... http://www.asseenontvmusic.com/timelife.html
	Welcome to TimeLife.com Music Shop online for exclusive music CDs, music collections, & musi... http://www.timelife.com/webapp/wcs/stores/servlet/Categor...		Welcome to TimeLife.com Music Shop online for exclusive music CDs, music collections, & musi... http://www.timelife.com/webapp/wcs/stores/servlet/Categor...
	Contemporary Country (Time-Life Music) - Wiki... Contemporary Country was a 22-volume series issued by Time-... http://en.wikipedia.org/wiki/Contemporary_Country_(Time-...	×	Songs ... Time Life 10 CD Collection... Christian Music CD/Album review of Songs 4 Ever Time Life 10 CD Collection... http://www.titletrakk.com/album-cd-reviews/songs-4...
	Time Life Canada Homepage The most comprehensive country music collection dedicated to... http://www.timelife.ca	×	Christian Band - Newsong - More Life - CD Review of ... Christian Band - Newsong - More Life CD Review ... Three yea... http://christianmusic.about.com/cs/cdreviews/fr/aafpr09080...

• 上下文感知的不同类型

- 类型三：查询泛化
 - 在用户的后续查询中，对先前查询中部分内容进行了更泛化的查询
 - 体现了用户对于该查询更广泛的兴趣，而不是局限在某一特定话题
 - 在本实例中，用户从单纯的游戏网站转为查询游戏介绍和历史（如维基百科）

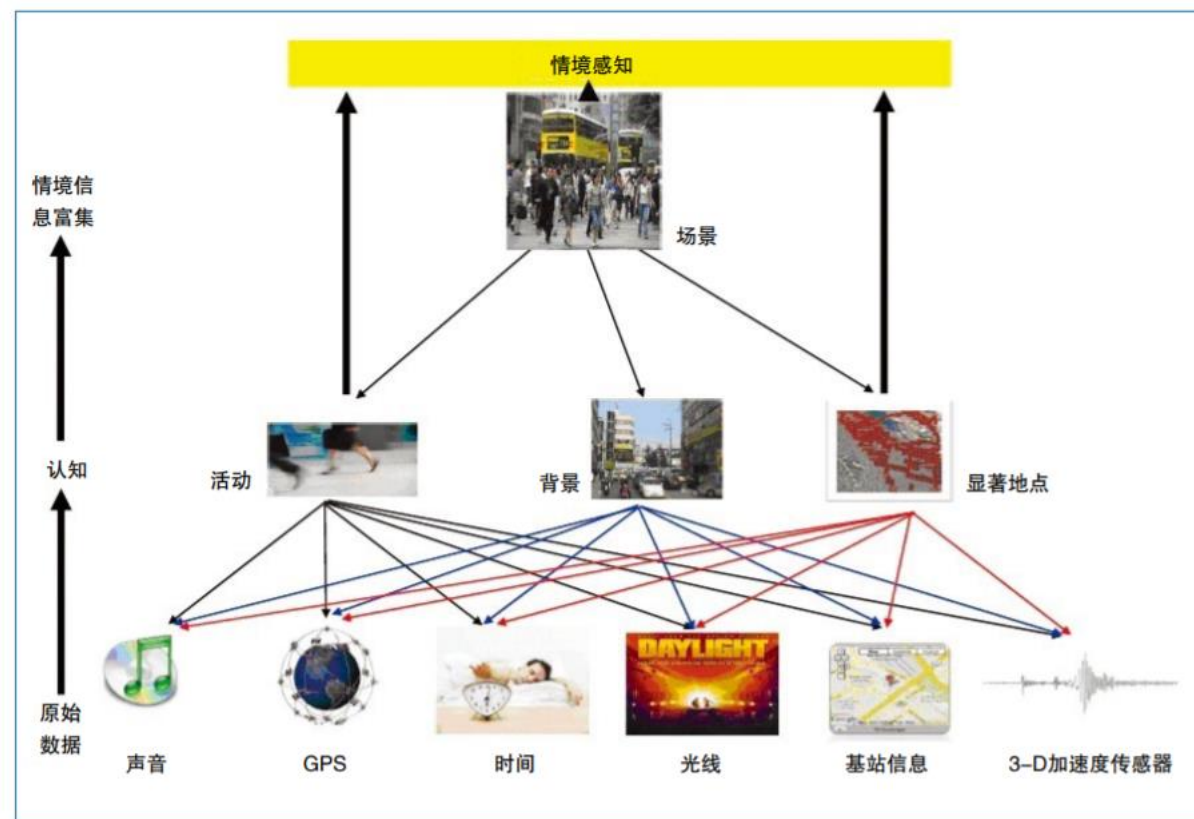
Query 1: "Free online Tetris"		Query 2: "Tetris game"	
×	Tetris Friends Online Games - Play Free Games Featuri... Play free online games featuring Tetris. Play single-player and ... http://tetrisfriends.com		Tetris Friends Online Games - Play Free Games Featuri... Play free online games featuring Tetris. Play single-player and ... http://tetrisfriends.com
×	Play Free Tetris Game Online Play this classic, original, Flash Tetris Game online for free. http://www.gametetris.com		Tetris game Free online game: Make lines with falling blocks! Russia's finest... http://www.play.vg/games/6-Tetris.html
	Free Tetris Game Free tetris game - Play free tetris games online, learn about tet... http://www.tetrislive.com	×	Tetris (Game Boy) - Wikipedia, the free encyclopedia Tetris was a pack-in title included with the Game Boy at the ha... http://en.wikipedia.org/wiki/Tetris_(handheld_game)
	4FreeOnlineGame.com - Free Online Tetris Game 4FreeOnlineGame - Free Online Tetris Game ... This is the all ... http://www.4freeonlinegame.com/Tetris	×	Tetris - non-stop puzzle action Tetris logo, Tetris theme song and Tetrminos are trademarks of... http://www.tetris.com
	Tetris - Play Tetris. Free online games © Adoption Media, LLC 1995 - 2010 This site should not subst... http://games.adoption.com/free-online-games/Tetris		Free Tetris Game Free tetris game - Play free tetris games online, learn about tetr... http://www.tetrislive.com

• 上下文感知的不同类型

- 类型四：一般关联
 - 借助先前查询，可以补全用户在当前搜索中的特定意图
 - 更接近之前所笼统叙述的“上下文”情境感知的概念

Query 1: "Xbox 360"		Query 2: "FIFA 2010"	
×	Xbox.com Home Xbox.com is your ultimate source for all things Xbox and Xb... http://www.xbox.com		FIFA.com - The Official Website of the FIFA World Cup The Official Website of the 2010 FIFA World Cup South Africa™ http://www.fifa.com/worldcup/index.html
	Xbox 360 - Wikipedia, the free encyclopedia The Xbox 360 is the second video game console produced by ... http://en.wikipedia.org/wiki/Xbox_360		2010 FIFA World Cup - Wikipedia, the free encyclopedia The template below has been deprecated (see discussion), and ... http://en.wikipedia.org/wiki/2010_FIFA_World_Cup
×	Xbox.com Xbox 360 Find out more about Xbox 360, the awesome lineup of games ... http://www.xbox.com/en-US/hardware		FIFA.com - Fédération Internationale de Football Associa... The official site of the international governing body of the sport ... http://fifa.com
	Microsoft Xbox Xbox 360 delivers the most powerful console, the next genera... http://www.microsoft.com/xbox	×	FIFA 10 Soccer : FIFA 2010 - EA Sports Games Improvement in Management Mode, Flick Passes, Ball Physics, ... http://www.ea.com/games/fifa-soccer
	Xbox 360 - Gizmodo This No-Name HTPC Remote Has a Keyboard, Can Work W... http://gizmodo.com/tag/xbox-360		FIFA 2010 World Cup in South Africa A surprise in the 2007 Asian Cup! The Iraqis win it! In spite of ... http://southafrica2010.wordpress.com

- **更复杂的情境信息**
- 移动智能终端的发展与移动搜索服务的普及，推动了更丰富情境信息的获取。
- 对于用户所处状态描述更为完整、准确
- 用户的信息需求类型也更为丰富、多样



- 情境感知的用户需求

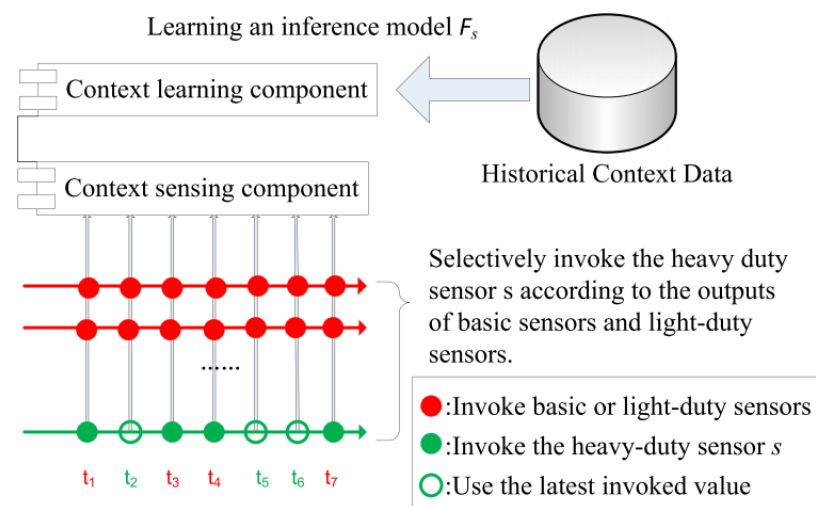
- 描述用户所处状态及意图的，不再是单一的上下文（查询或查询概念），而是一组复杂且相互关联的情境信息。
- 单一解读某一种情境要素，可能无法得到用户当前状态的准确描述

Timestamp	Context record
t_1	{{(Holiday?:No),(Time range: AM8:00-9:00),(Speed: High),(Audio level: Low)}}
t_2	{{(Holiday?:No),(Time range: AM8:00-9:00),(Speed: High),(Audio level: Middle)}}
t_3	{{(Holiday?:No),(Time range: AM8:00-9:00),(Speed: High),(Audio level: Middle) ,(Interaction: Music)}}

t_{58}	{{(Holiday?:Yes),(Time range: AM10:00-11:00),(Movement: Move)(Location: Shop),(Audio level: Middle)}}
t_{59}	{{(Holiday?:Yes),(Time range: AM10:00-11:00),(Movement: Move)(Location: Shop),(Audio level: Middle)}}
t_{60}	{{(Holiday?:Yes),(Time range: AM10:00-11:00),(Movement: Move)(Location: Shop),(Audio level: Middle)}}

- 情境感知的用户需求

- 一个有趣的案例：如何借助情境感知的模式化行为减少开支。
 - 通过情境识别模式化路径，减少传感器（如GPS）收集次数并降低能耗。
 - 发散思维：这一思想是否能够用在缓存机制的使用上？



本章小结

查询

- 查询条件理解
- 相关性反馈
 - 概念与技术: Rocchio算法
 - 不同类型的相关性反馈
- 查询扩展
- 情境感知查询