信息检索与数据挖掘

第7章 相关反馈和查询扩展

课程内容

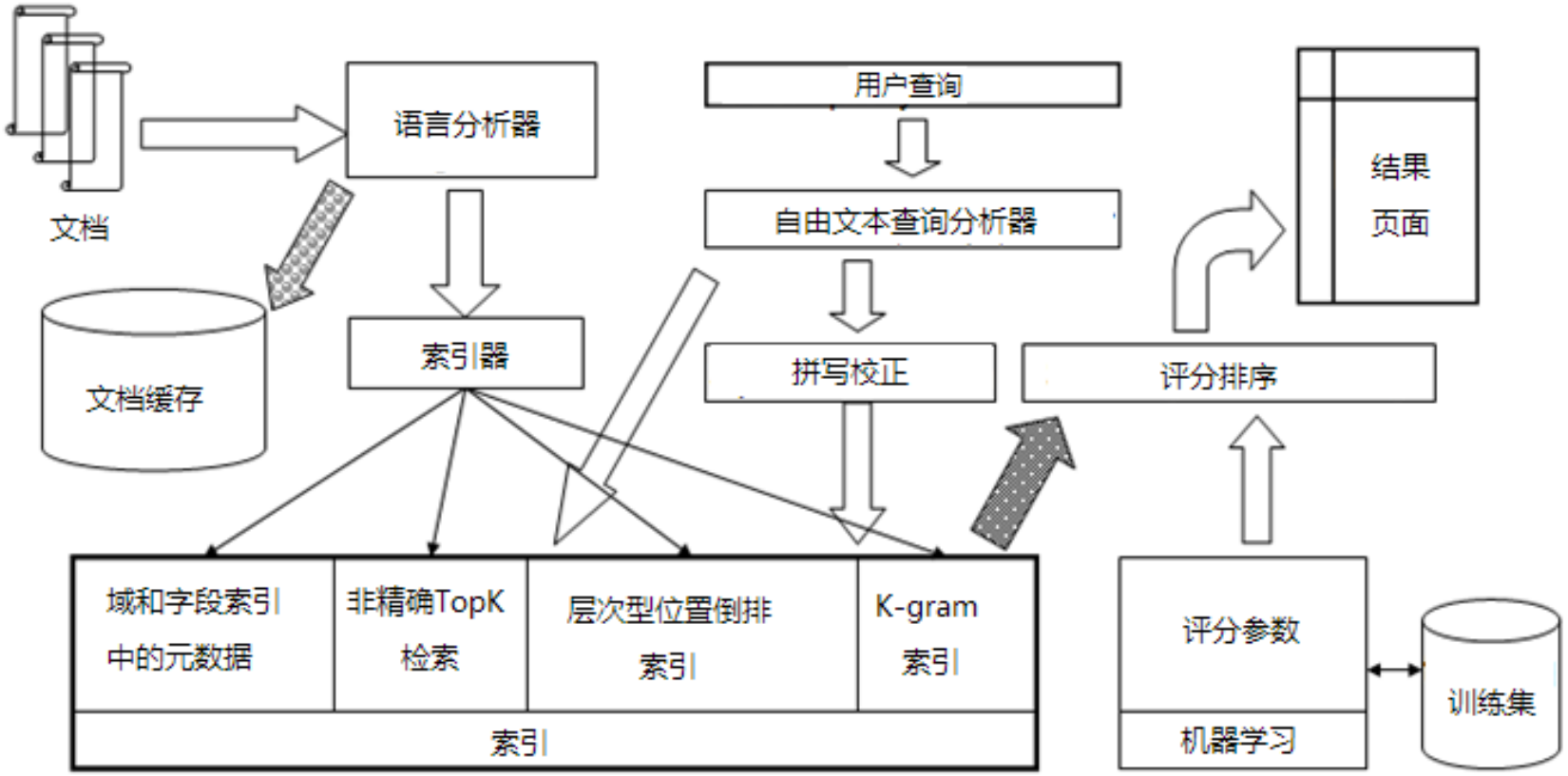
- 第1章 绪论
- 第2章 布尔检索及倒排索引
- 第3章 词典查找及扩展的倒排索引
- 第4章 索引构建和索引压缩
- 第5章 向量模型及检索系统
- 第6章 检索的评价
- 第7章 相关反馈和查询扩展
- 第8章 概率模型
- 第9章 基于语言建模的检索模型
- 第10章 文本分类
- 第11章 文本聚类
- 第12章 Web搜索
- 第13章 多媒体信息检索
- 第14章 其他应用简介

本讲内容

- 查询优化概述
- 相关反馈(relevance feedback)
  - 相关反馈概述
  - Rocchio 相关反馈算法
  - 隐式相关反馈
  - 伪相关反馈
  - 相关反馈的假设条件及评价方法
- 查询扩展(Query expansion)

回顾：检索系统

能否让查询结果更相关？



查询优化

信息需求≠查询

喜马拉雅 高度 喜马拉雅 高度 百度一下

网页 新闻 贴吧 知道 音乐 图片 视频 地图 文库 更多»

百度为您找到相关结果约4,100,000个 搜索工具

喜马拉雅山脉海拔：

平均7000-8000米以上

喜马拉雅山脉 （梵语：hima alaya，意为雪域），藏语意为"雪的故乡"。位于青藏高原南缘边缘，是世界海拔最高的山脉，其中有110多座山峰高达或超过海拔7350米。是东亚大... 详情>>

来自百度百科 | 报错

喜马拉雅山脉_百度百科

喜马拉雅山脉作为一个影响空气和水的大循环系统的气候大分界线,对于南面的印度次大陆和北面的中亚高地的气象状况具有决定性的影响。由于位置和令人惊叹的高度,大...

地理情况 气候特征 主要资源 水系情况 人文历史 更多>>

baike.baidu.com/

喜马拉雅山海拔高度?_百度知道

2个回答 - 提问时间: 2013年05月27日

最佳答案: 喜马拉雅山脉位于西藏自治区与巴基斯坦、印度、尼泊尔、锡金、不丹等国边境上,东西绵延2400多公里,南北宽约200—300公里,由几列大致平行的山脉组成,呈向南...

https://zhidao.baidu.com/quest...

喜马拉雅山的高度是怎么量出来的呢	3个回答	2009-12-18
喜马拉雅山的准确高度是多?	4个回答	2016-02-22
喜马拉雅山有多高?	2个回答	2008-04-04

更多知道相关问题>>

高度!高度!高度!MP3_电台节目免费下载-喜马拉雅FM

电台频道最近更新了高度!高度!高度!MP3,您可以免费下载高度!高度!高度!等电台节目,精彩纷呈,不容错过!-喜马拉雅听书

喜马拉雅_百度搜索

https://www.baidu.com/s?wd=%E5%96%9C%E9%A9%AC%E6%8B%89%E9%9B

喜马拉雅 百度一下

网页 新闻 贴吧 知道 音乐 图片 视频 地图 文库 更多»

百度为您找到相关结果约20,900,000个 搜索工具

为您推荐: 喜马拉雅电台 喜马拉雅app 喜马拉雅睡前故事 喜马拉雅在线收听

喜马拉雅FM-国内专业音频分享平台,随时随地,听我想听!

国内专业音频分享平台,随时随地,听我想听!4亿用户选择的在线音频平台。马东、郭德纲、吴晓波等20多万大咖入驻,1亿多条原创有声内容覆盖有声书、儿童、相声评书、...

www.ximalaya.com/explore/ V3 - 百度快照

喜马拉雅FM_电台节目在线收听-喜马拉雅FM

6天前 - 喜马拉雅FM播客节目精选。 喜马拉雅FM播客节目精选。 收起03-嫁人当嫁王小川? 13 856 00:00/20:30 00:00 3天前 下载到手机 赞(14) 评论(45) 转采(3) ...

www.ximalaya.com/89141... V3 - 百度快照

喜马拉雅好声音_网络电台_主播-喜马拉雅FM

欢迎收听喜马拉雅好声音网络电台,在这里您可以了解更多喜马拉雅好声音主播,个人电台动态信息。喜马拉雅FM,听我想听!

www.ximalaya.com/zhubo... V3 - 百度快照

喜马拉雅开放平台

喜马拉雅开放平台 open.ximalaya.com 将喜马拉雅海量音频内容开放给第三方合作方。通过移动应用SDK和完善的接入文档,让接入喜马拉雅音频内容更便捷。

open.ximalaya.com/ - 百度快照 - 204条评价

喜马拉雅FM-国内专业音频分享平台,随时随地,听我想听!

国内专业音频分享平台,随时随地,听我想听!4亿用户选择的在线音频平台。马东、郭德纲、吴晓波等20多万大咖入驻,1亿多条原创有声内容覆盖有声书、儿童、相声评书、...

www.ximalaya.com/downl... V3 - 百度快照

信息需求≠查询

关键词：
知否知否 应是绿肥红瘦

信息需求：
(1)李清照的词
(2)网络小说
(3)电视剧

为何有时用户感觉召回率低？

- 在大多数文档集中，同一概念可以用不同的词来表达，这个现象称为一义多词（synonymy），它会对大部分信息检索系统的召回率产生影响。
  - 比如，输入查询aircraft时我们希望能找到包含plane的文档，当然，这里的plane指的是飞机（airplane），而不是木工刨（woodworking plane）。
- 另外，我们也希望在查找thermodynamics时能够与特定环境下的heat匹配上。

一义多词
LSI可以帮助我们发现相关文档

搜索中提高召回率的方法

- 本讲的主题：两种提高召回率的方法—**相关反馈**及**查询扩展**

- 考虑查询q: [aircraft] . . .
- 某篇文档 d 包含 “plane”, 但是不包含 “aircraft”
- 显然对于查询q，一个简单的IR系统不会返回文档d，即使d是和q最相关的文档
- 我们试图改变这种做法：
- 也就是说，我们会返回不包含查询词项的相关文档。

关于召回率Recall

- 本讲当中会放松召回率的定义，即(在前几页)给用户返回更多的相关文档。

- 这可能实际上会降低召回率，比如，将jaguar扩展为jaguar(美洲虎；一种汽车品牌)+panthera(豹属)
- 可能会去掉一些相关的文档，但是可能增加前几页返回给用户的相关文档数

提高召回率的方法

- 本章主要讨论系统中进行**查询优化（query refinement）**的各种方法，包括全自动的方法和用户参与的方法。
- 局部(local)方法
  - 对用户查询进行局部的即时的分析
  - 主要的局部方法：相关反馈(relevance feedback)
- 全局(Global)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)
  - 利用该词典进行查询扩展

小结：查询优化的动机

查询优化（query refinement）

- 查询不能准确表示信息需求→召回率低？
- 查询优化的目标？
  - 提高召回率
- 查询优化的可能途径？
  - 局部(local)方法：相关反馈(relevance feedback)
  - 全局(Global)方法：查询扩展

本讲内容

- 查询优化概述
- 相关反馈(relevance feedback)
  - 相关反馈概述
  - Rocchio 相关反馈算法
  - 隐式相关反馈
  - 伪相关反馈
  - 相关反馈的假设条件及评价方法
- 查询扩展(Query expansion)

相关反馈的基本思想

- **相关反馈：**用户对初始返回结果的相关性进行反馈
  - 用户提交一个查询
  - 用户将部分结果标记为相关或者不相关
  - 系统根据用户的反馈，对信息需求进行优化，将其表示成更好的形式
  - 相关反馈可以进行多次
  - Idea:如果不能很好地了解文档集合，就很难把自己的信息需求转化成查询，进行多次相关反馈可以有所帮助。

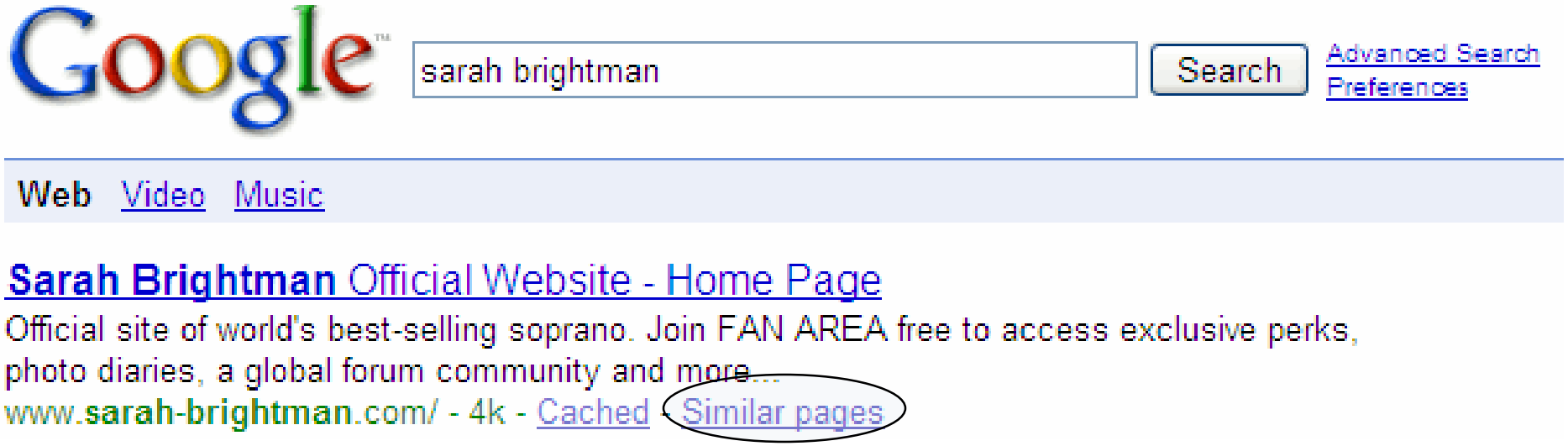
相关反馈分类

- 用户相关反馈或显式相关反馈(User Feedback or Explicit Feedback): 用户显式参加交互过程

- 隐式相关反馈(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。

- 伪相关反馈或盲相关反馈(Pseudo Feedback or Blind 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。

相关反馈的例子1：类似页面



类似页面→相关推荐

oogle Ed Sheeran

全部 新闻 视频 图片 地图 更多 设置 工具

找到约 16,900,000 条结果 （用时 0.48 秒）

紅髮艾德- 维基百科，自由的百科全书
https://zh.wikipedia.org/zh-hans/紅髮艾德 ▾ 转为简体网页
爱德华·克里斯托弗·希兰，MBE（英语：Edward Christopher Sheeran，1991年2月17日 - ），以其藝名紅髮艾德知名，是英國的一名歌手、吉他手與唱片製作人。希兰出生於西約克郡哈利法克斯，並在薩福克郡弗瑞林姆長大。18歲時，他曾於薩里郡吉爾福德當代音樂學院（英语：Academy of Contemporary Music）就讀大學。2011年年初， ...

出道地点: 英國萨福克郡弗瑞林姆 **唱片公司**: Asylum（英语：Asylum Records）; 大...
音乐类型: 流行; 民谣流行; **国籍**: 英國

早年生活 · 巡迴演唱會 · 爭議事件

Ed Sheeran - Wikipedia
https://en.wikipedia.org/wiki/Ed_Sheeran ▾ 翻译此页
Edward Christopher Sheeran, MBE is an English singer, songwriter, guitarist, record producer, and actor. Sheeran was born in Halifax, West Yorkshire, and raised in Framlingham, Suffolk. He attended the Academy of Contemporary Music in Guildford as an undergraduate from the age of 18 in 2009. In early 2011, Sheeran ...

Labels: Asylum; Atlantic; Elektra **Origin**: Framlingham, Suffolk, England
Instruments: Vocals; guitar **Genres**: Pop; folk pop;

X (Ed Sheeran album) · Ed Sheeran · Ed Sheeran discography · Ed Sheeran album

Ed Sheeran
www.edsheeran.com/ ▾ 翻译此页
Ed Sheeran is pleased to announce eight additional dates to his first... 06 Feb 2018. Read More · Ed Sheeran Announces Brits'... Ed Sheeran is happy to announce a special show as part of this year's... 22 Jan 2018. Read More · Snowglobe yourself into the Perfect... Create your own #PerfectSnowGlobe video and share it ...

Ed Sheeran - YouTube
https://www.youtube.com/user/EdSheeran ▾ 翻译此页

艾德·希兰

歌手

爱德华·克里斯托弗·希兰，MBE，以其艺名艾德·希兰知名，是英国的一名歌手、吉他手与唱片制作人。希兰出生于西约克郡哈利法克斯，并在萨福克郡弗瑞林姆长大。18岁时，他曾于萨里郡吉尔福德当代音乐学院就读大学。 维基百科

生于： 1991 年 2 月 17 日（27 岁），英国哈利法克斯
身高： 1.73 米
配偶： Cherry Seaborn（2018年結婚）
家长： 伊莫金·希伦， 约翰·希伦

歌曲


你的样子
÷ · 2017 年


完美无瑕
÷ · 2017 年

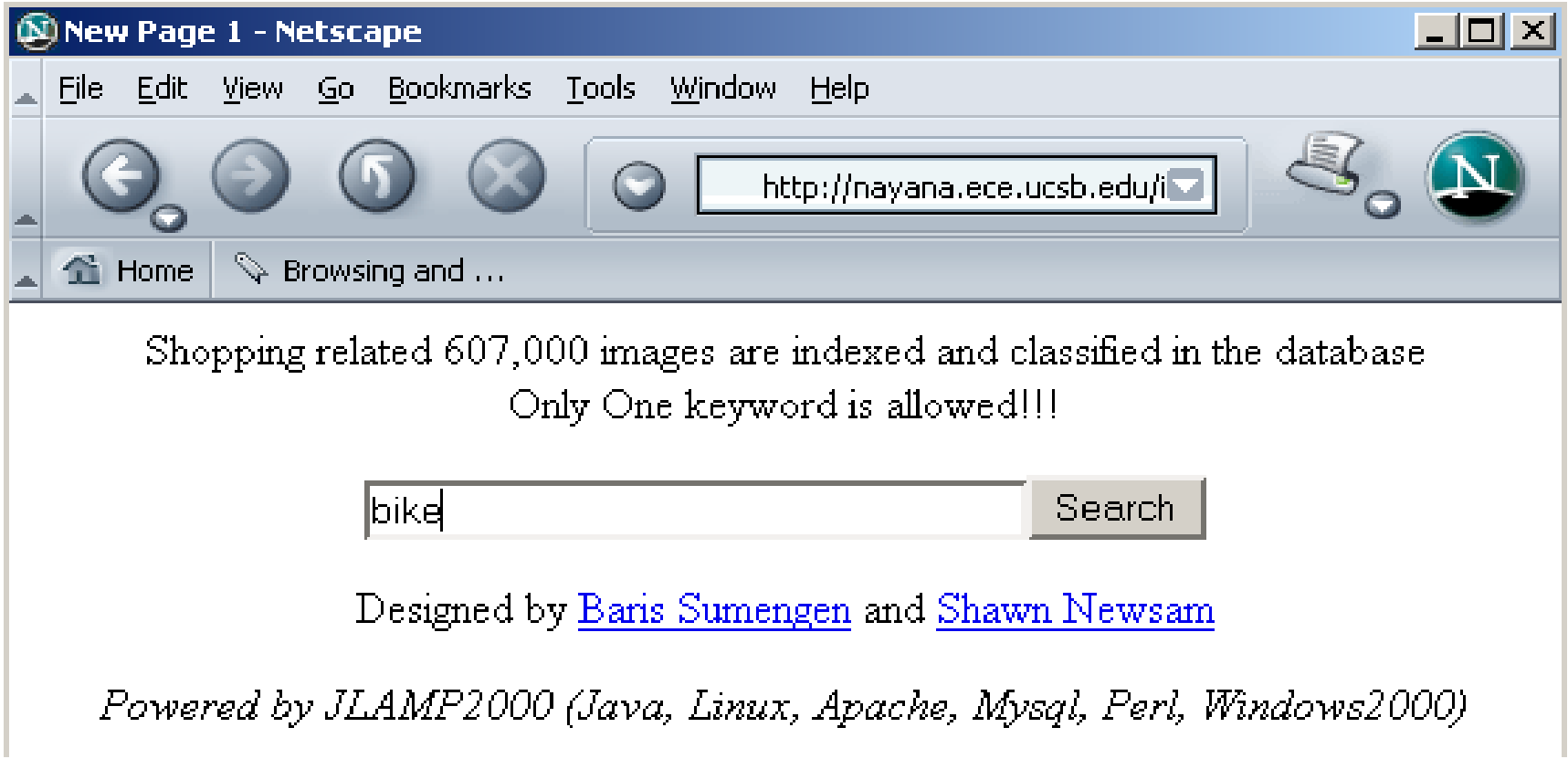

自言自语
X · 2014 年

还有25+项

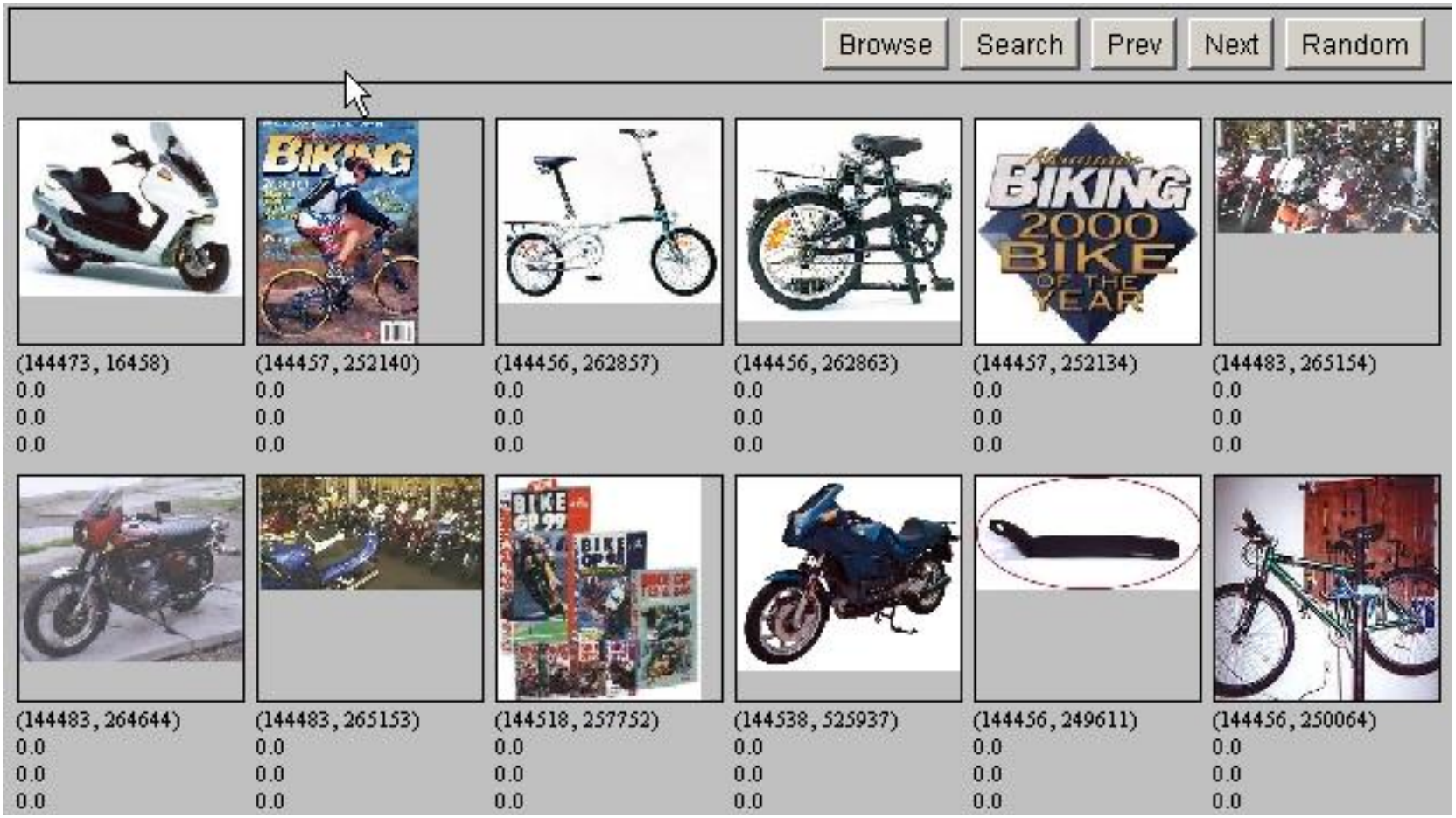
相关反馈的例子2：

- Image search engine

  http://nayana.ece.ucsb.edu/imsearch/imsearch.html

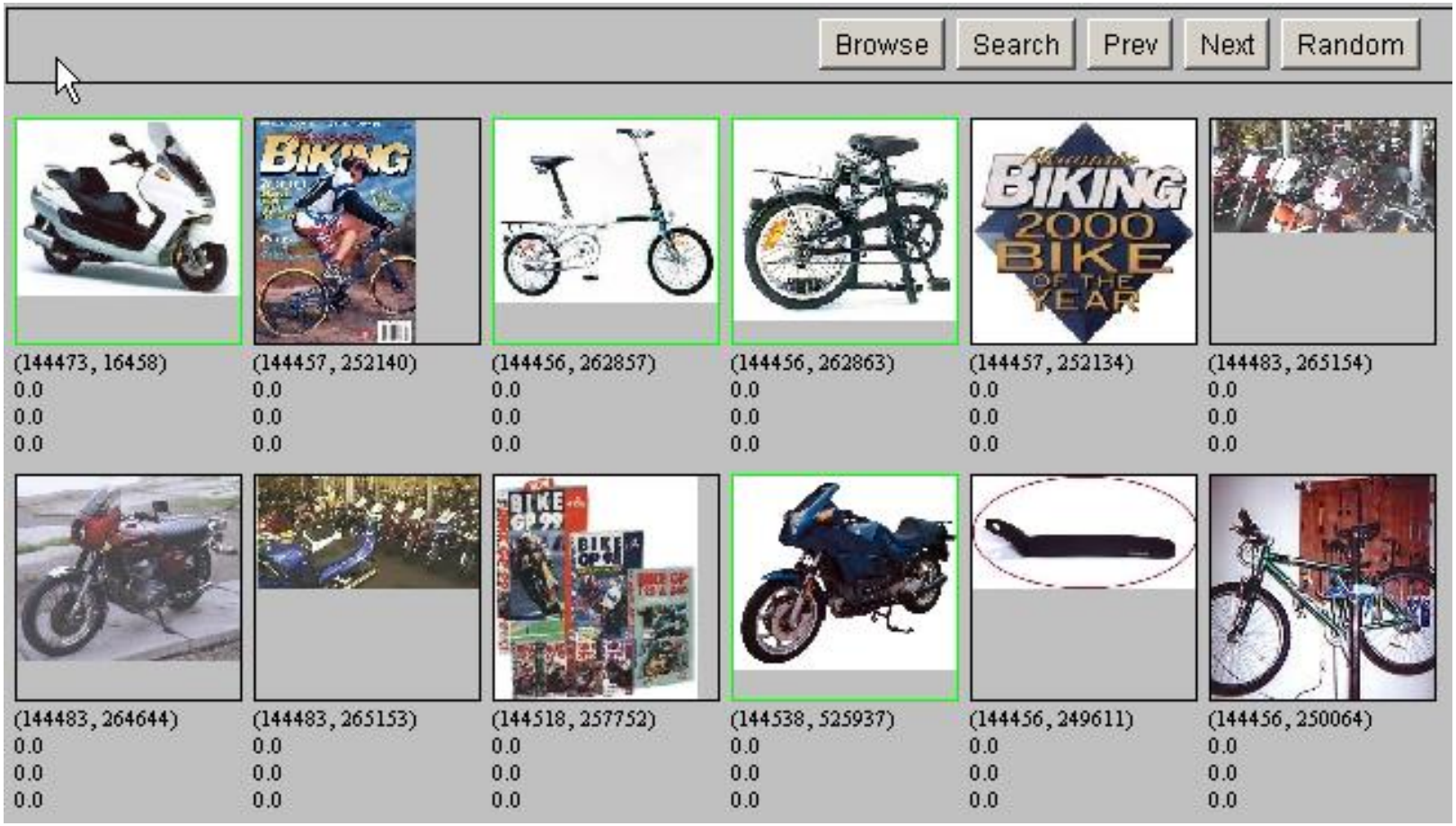


首次查询的返回结果

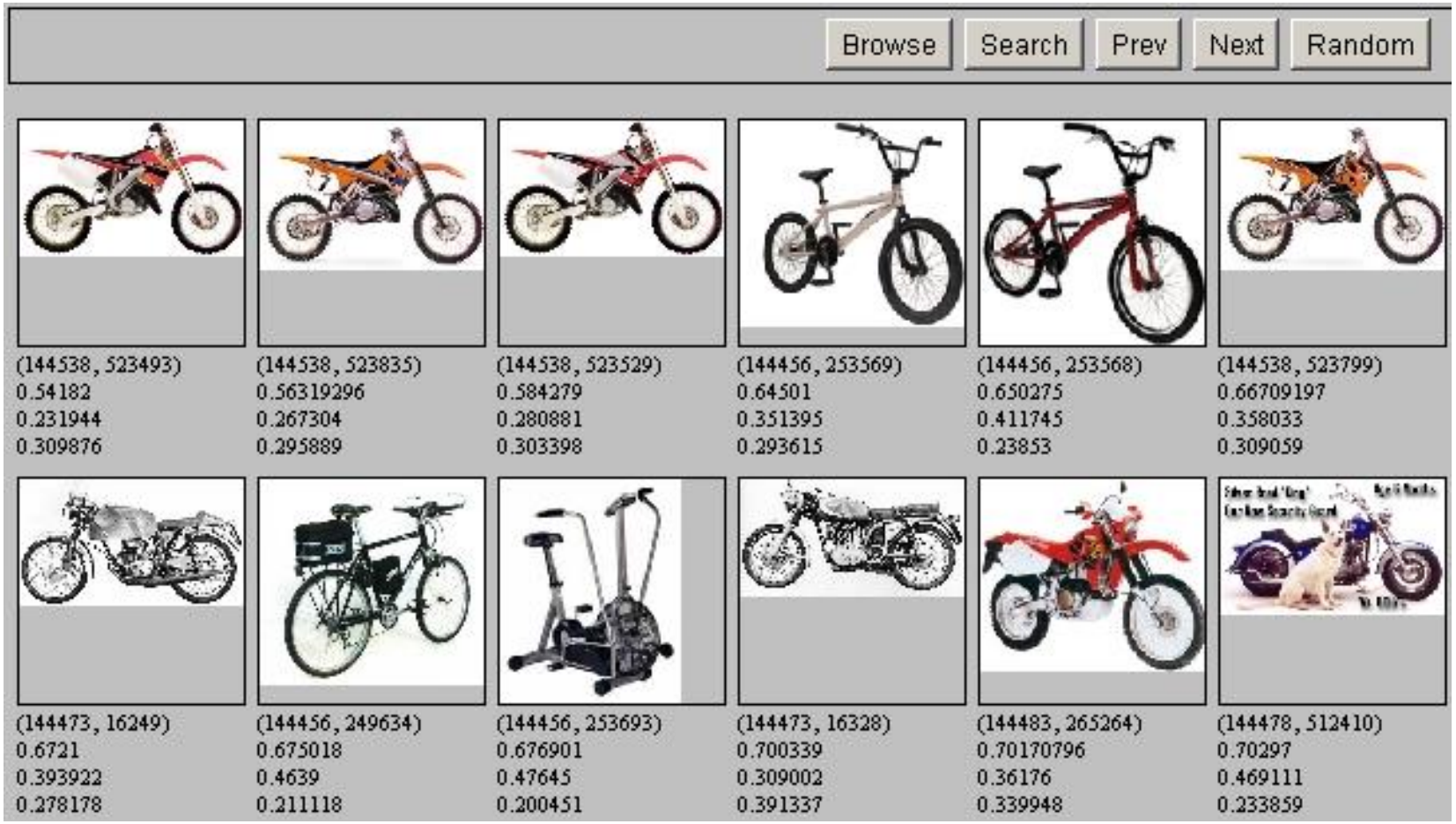

Browse
Search
Prev
Next
Random
BIKING
BIKING 2000 BIKE OF THE YEAR
(144473, 16458)
0.0
0.0
0.0
(144457, 252140)
0.0
0.0
0.0
(144456, 262857)
0.0
0.0
0.0
(144456, 262863)
0.0
0.0
0.0
(144457, 252134)
0.0
0.0
0.0
(144483, 265154)
0.0
0.0
0.0
(144483, 264644)
0.0
0.0
0.0
(144483, 265153)
0.0
0.0
0.0
(144518, 257752)
0.0
0.0
0.0
(144538, 525937)
0.0
0.0
0.0
(144456, 249611)
0.0
0.0
0.0
(144456, 250064)
0.0
0.0
0.0

相关反馈

绿框表示用户认为相关的结果



相关反馈后再次检索的结果

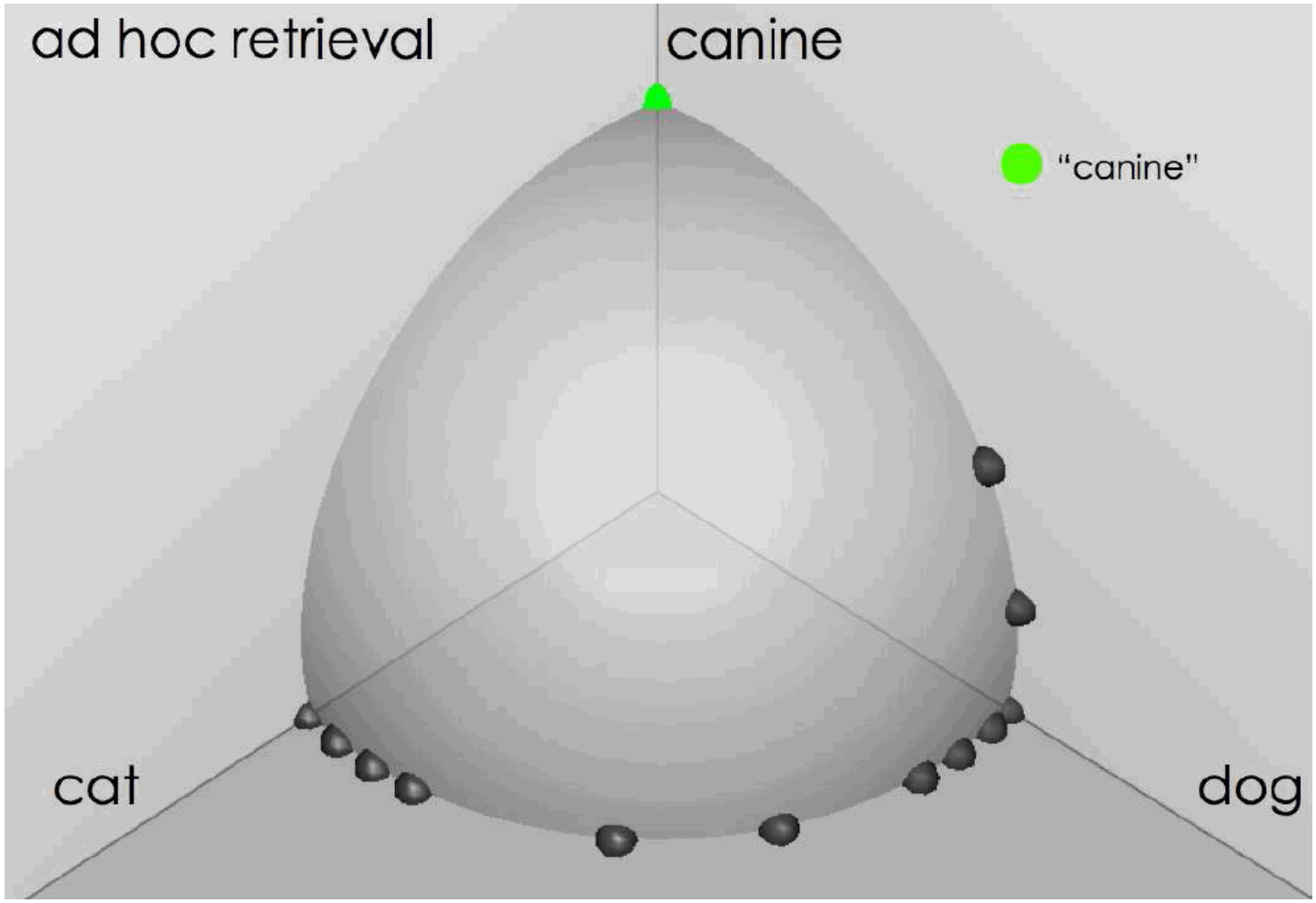


相关反馈的例子3：向量空间的例子: 查询“canine”

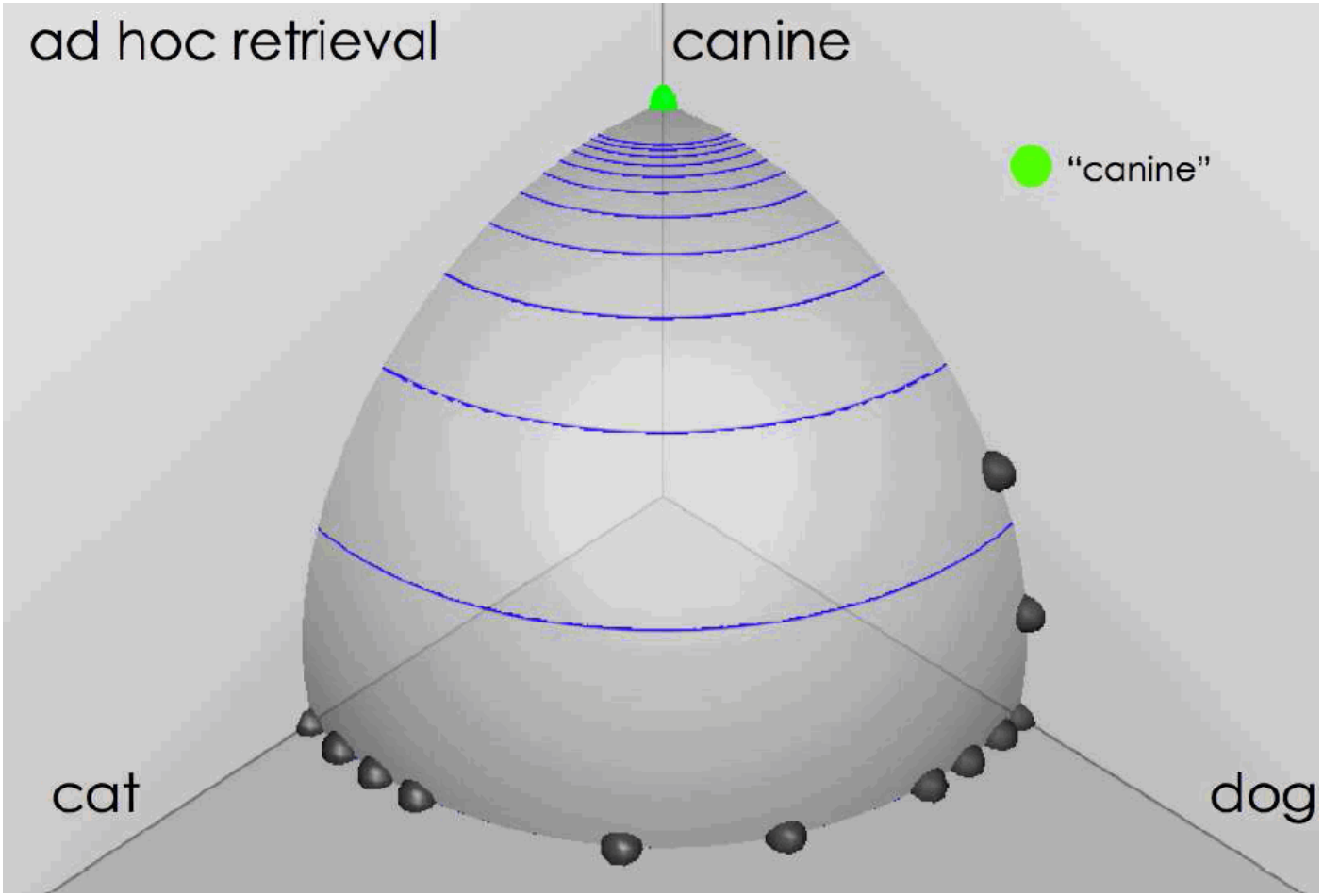
Canine ['kenaɪn]
adj. 犬的；犬齿的；犬科的；似犬的
n. 犬；[解剖] 犬齿

Source:

Fernando Díaz



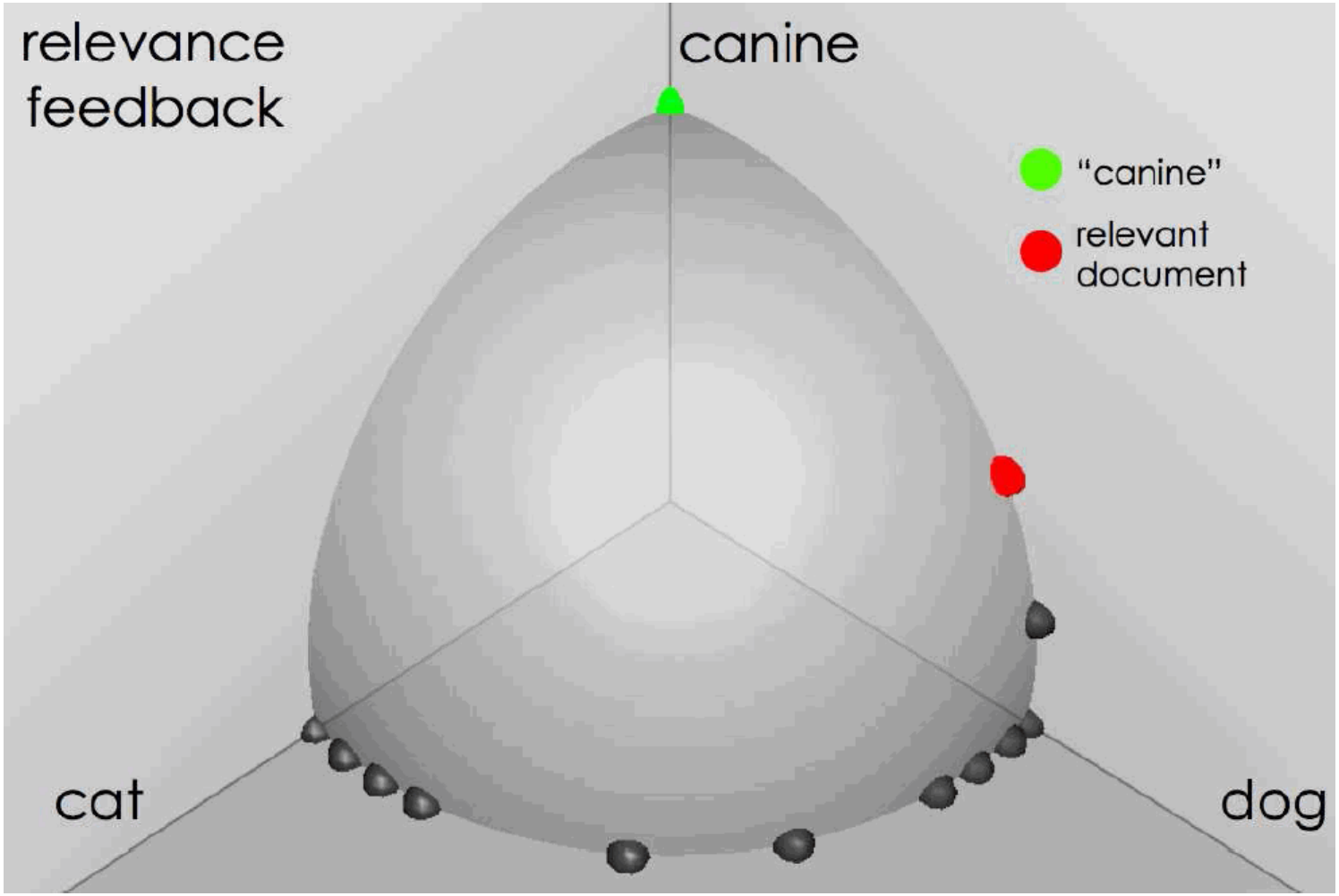
文档和查询 “canine” 的相似度



Source:

Fernando Díaz

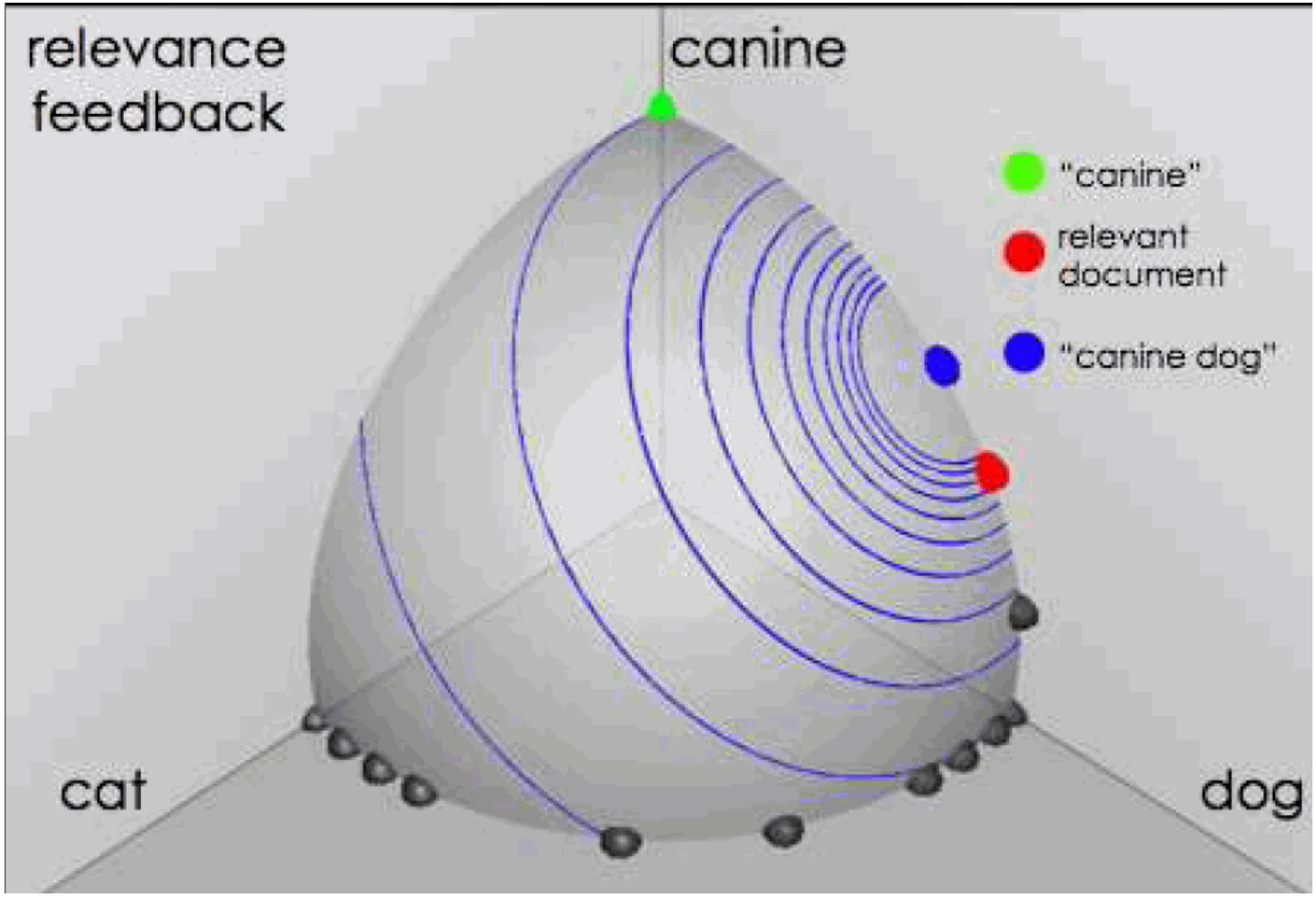
用户的反馈：选择一个认为相关的文档



Source:

Fernando Díaz

查询扩展后的结果



Source:

Fernando Díaz

例4：一个实际的例子

- 初始查询： *New space satellite applications*

+ 1. 0.539, 08/13/91, NASA Hasn't Scrapped Imaging Spectrometer
+ 2. 0.533, 07/09/91, NASA Scratches Environment Gear From Satellite Plan
3. 0.528, 04/04/90, Science Panel Backs NASA Satellite Plan, But Urges Launches of Smaller Probes
4. 0.526, 09/09/91, A NASA Satellite Project Accomplishes Incredible Feat: Staying Within Budget
5. 0.525, 07/24/90, Scientist Who Exposed Global Warming Proposes Satellites for Climate Research
6. 0.524, 08/22/90, Report Provides Support for the Critics Of Using Big Satellites to Study Climate
7. 0.516, 04/13/87, Arianespace Receives Satellite Launch Pact From Telesat Canada
+ 8. 0.509, 12/02/87, Telecommunications Tale of Two Companies

- 用户使用“+”标记相关的文档

相关反馈之后，扩展了的查询

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

查询扩展成18 个带权重的词项

扩展查询的结果

2 **1. 0.513, 07/09/91, NASA Scratches Environment Gear From Satellite Plan**

1 **2. 0.500, 08/13/91, NASA Hasn't Scrapped Imaging Spectrometer**

3. 0.493, 08/07/89, When the Pentagon Launches a Secret Satellite, Space Sleuths Do Some Spy Work of Their Own

4. 0.493, 07/31/89, NASA Uses 'Warm' Superconductors For Fast Circuit

8 **5. 0.492, 12/02/87, Telecommunications Tale of Two Companies**

6. 0.491, 07/09/91, Soviets May Adapt Parts of SS-20 Missile For Commercial Use

7. 0.490, 07/12/88, Gaping Gap: Pentagon Lags in Race To Match the Soviets In Rocket Launchers

8. 0.490, 06/14/90, Rescue of Satellite By Space Agency To Cost $90 Million

小结：相关反馈

- **相关反馈：**用户对初始返回结果的相关性进行反馈
  - 用户提交一个查询
  - 用户将部分结果标记为相关或者不相关
  - 系统根据用户的反馈，对信息需求进行优化
- **相关反馈的方式**
  - **用户相关反馈或显式相关反馈**
    - User Feedback or Explicit Feedback
  - **隐式相关反馈**
    - Implicit Feedback
  - **伪相关反馈或盲相关反馈**
    - Pseudo Feedback or Blind Feedback

本讲内容

- 查询优化概述
- 相关反馈(relevance feedback)
  - 相关反馈概述
  - Rocchio 相关反馈算法
  - 隐式相关反馈
  - 伪相关反馈
  - 相关反馈的假设条件及评价方法
- 查询扩展(Query expansion)

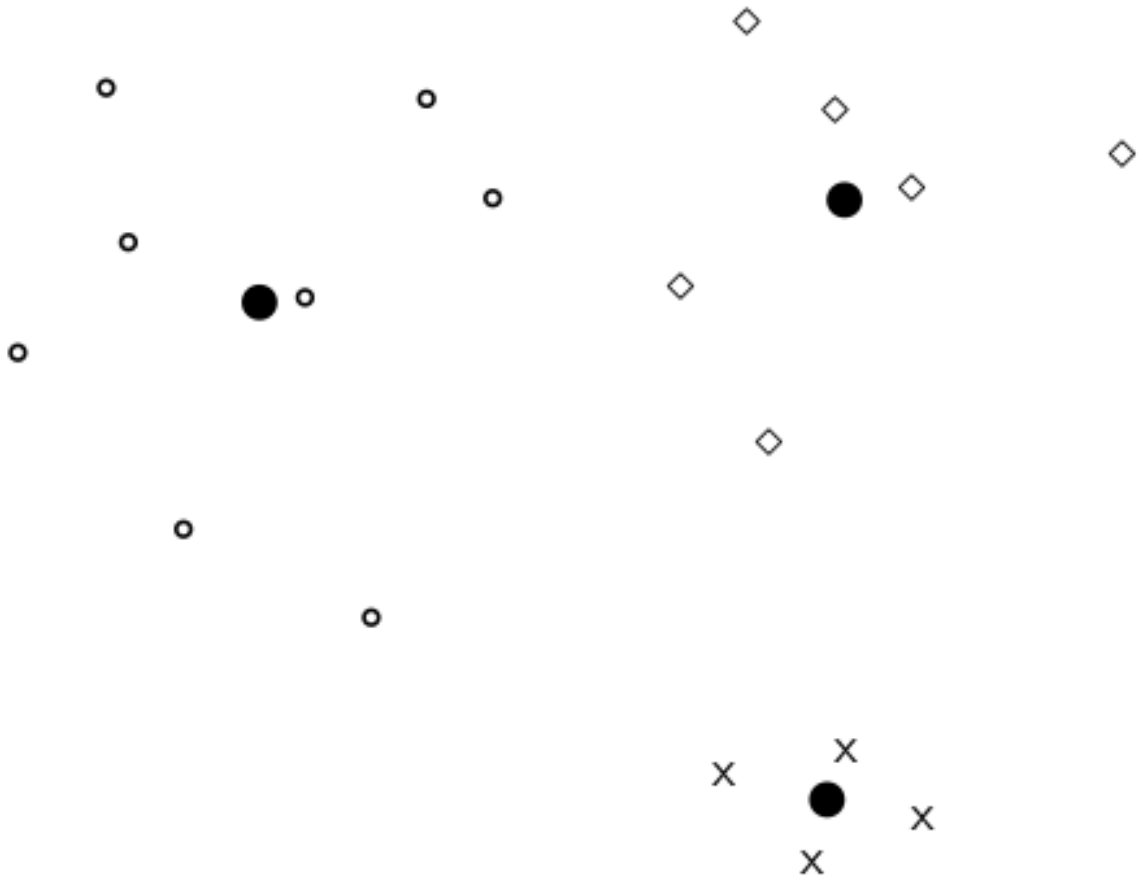
相关反馈中的核心概念：质心

- 质心是一系列点的质量的中心
- **回顾一下：我们将文档看作高维空间中的点**
- 我们可以采用如下方式计算文档的质心

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

其中D是一个文档集合，$\vec{v}(d) = \vec{d}$ 是文档d的向量表示

质心的例子



罗基奥（Rocchio） 算法

- Rocchio 算法使用向量空间模型来收集相关反馈
- Rocchio 算法试图寻找一个查询 $\vec{q}_{opt}$，使得：

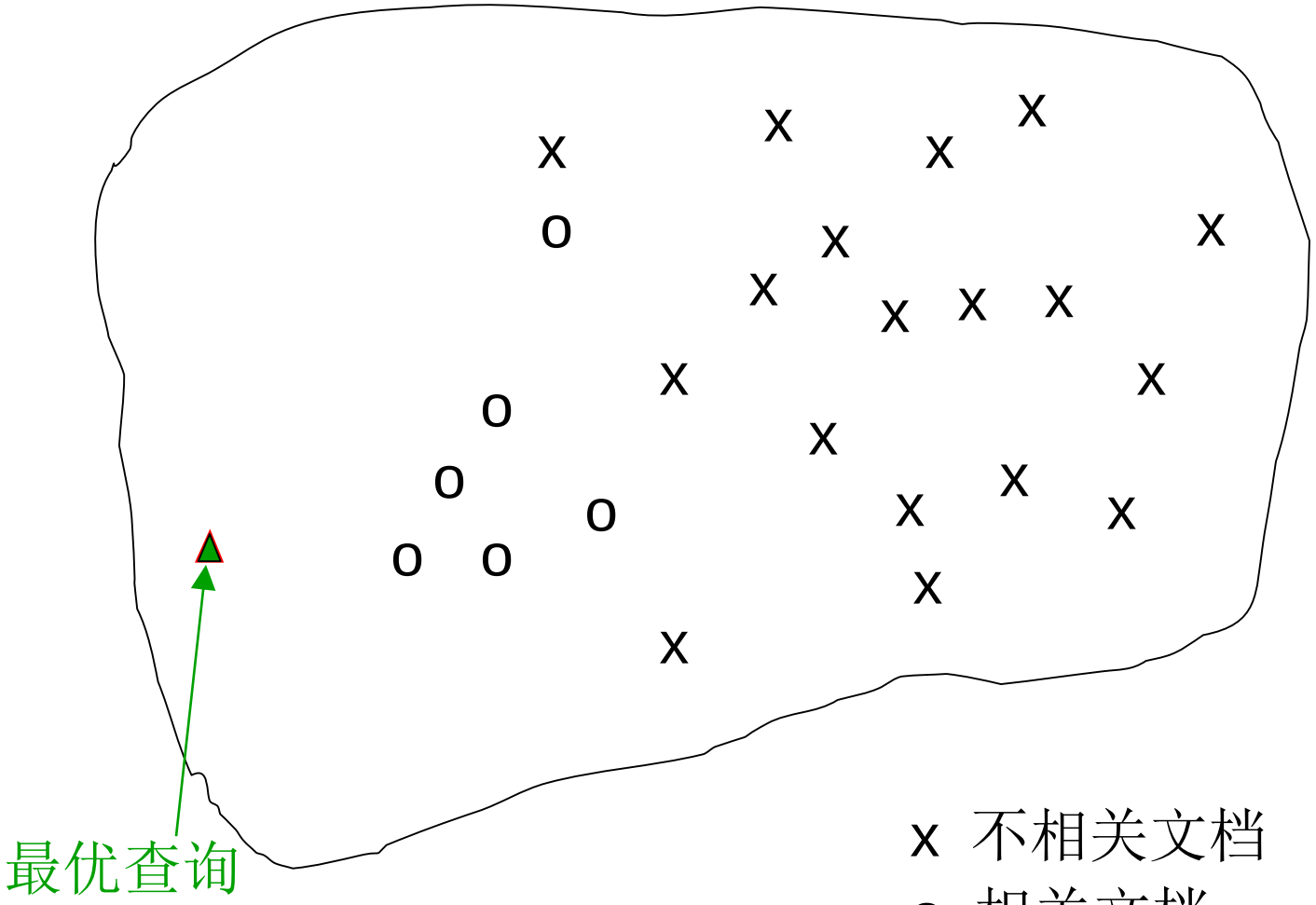
$$\vec{q}_{opt} = \arg\max_{\vec{q}} [\cos(\vec{q}, \vec{\mu}(C_r)) - \cos(\vec{q}, \vec{\mu}(C_{nr}))]$$

- 试图将相关文档和不相关文档分开

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

- 问题是，我们并不知道哪些文档是真正相关的

理论上的最好的查询



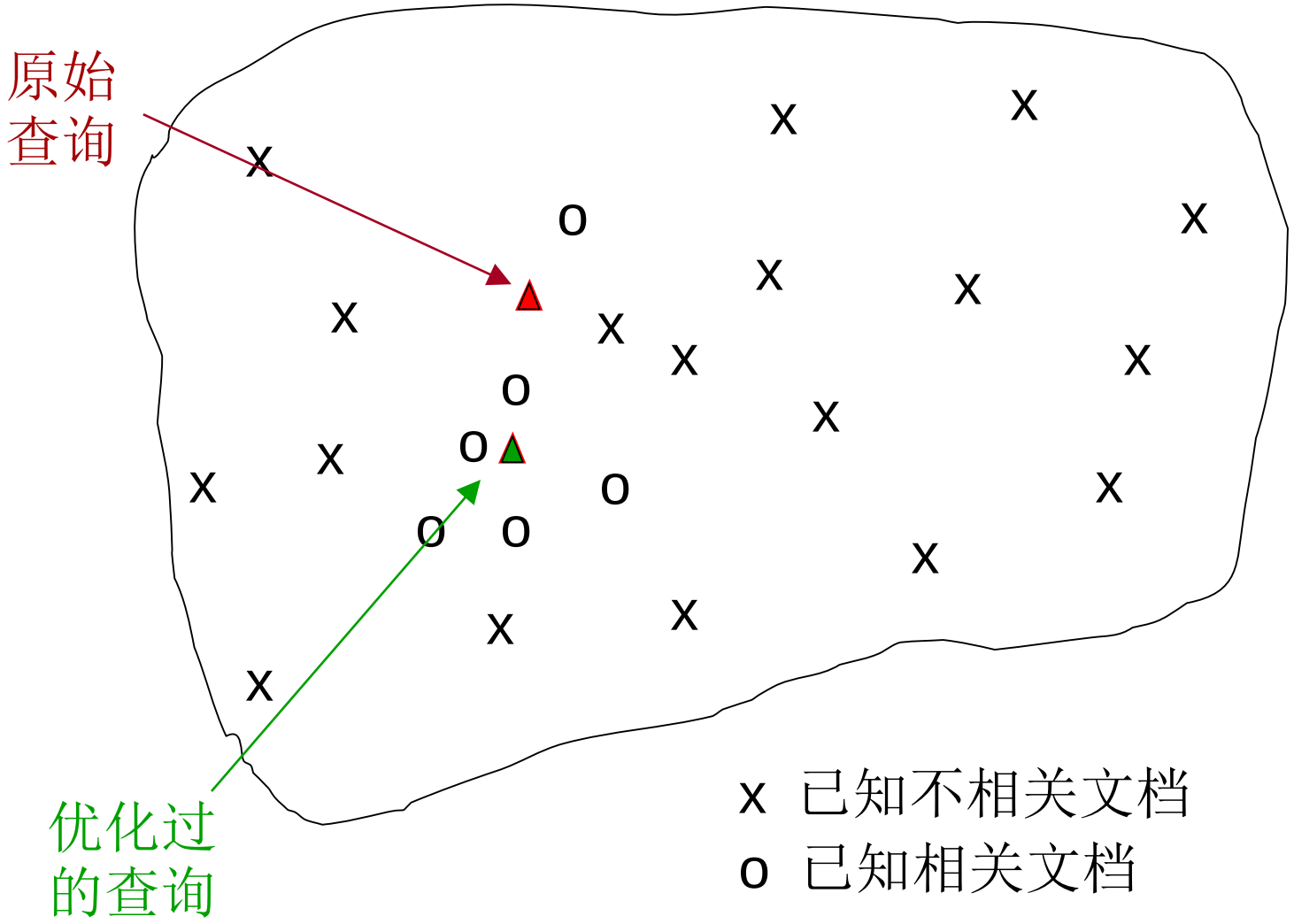
Rocchio 1971 算法 (SMART)

- 实际使用:

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

- D_r = 已知相关文档的向量集合
- D_{nr} = 已知的不相关文档的向量集合
  - 注意和文档集合D_r 和 D_{nr}*不同*
- q_m = 优化过的查询向量; q_0 = 原始的查询向量; α,β,γ: 权重 (手工或者根据经验设定)
- 新的查询向量向相关文档向量移动，远离不相关文档向量

对初始查询的相关反馈



需要注意的细节

- α、β、γ的权衡: 如果很多文档已经评价了相关度，那么β、γ应该大一些.
- 查询向量的某些权值可能为负数
  - 忽略负的权值

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

- 正反馈比负反馈更有价值(即设置$\gamma < \beta$; e.g. $\gamma = 0.25$, $\beta = 0.75$).
- 许多系统只使用正反馈($\gamma=0$). ←Why?

小结：Rocchio 相关反馈算法

- 文档的质心

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

- **Rocchio 算法的目标**

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

- **Rocchio 1971 算法 (SMART)**

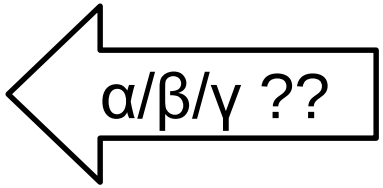
$$\vec{q}_m = \alpha\vec{q}_0 + \beta\frac{1}{|D_r|}\sum_{\vec{d}_j \in D_r}\vec{d}_j - \gamma\frac{1}{|D_{nr}|}\sum_{\vec{d}_j \in D_{nr}}\vec{d}_j$$

本讲内容

- 查询优化概述
- 相关反馈(relevance feedback)
  - 相关反馈概述
  - Rocchio 相关反馈算法
  - 隐式相关反馈
  - 伪相关反馈
  - 相关反馈的假设条件及评价方法
- 查询扩展(Query expansion)

Web上的相关反馈

- 一些搜索引擎提供“相似或相关网页”的功能(这是一种简单形式的相关反馈)
  - Google (link-based)
  - Altavista
  - Stanford WebBase



- 有些搜索引擎没有，因为很难向普通用户解释清楚什么是相关反馈:
  - Alltheweb
  - bing
  - Yahoo

Excite搜索引擎的相关反馈

- 一般而言，在Web 搜索中很少使用相关反馈技术。Excite搜索引擎是个例外，它一开始就提供了完整的相关反馈，不过，由于很少有人用，该功能后来及时被取消。
  - 在Web 上，很少有人会用到高级搜索界面，而且大部分人都希望在一次交互中完成搜索任务。此外，相关反馈技术在Web 上很少利用也可能反映出其他两个因素：一是相关反馈很难向普通用户解释清楚；二是相关反馈主要用来提高召回率，而Web 搜索用户很少关注是否获得足够的召回率。
- Spink 等人（2000）给出了在Excite 搜索引擎中相关反馈技术使用的结果。在所有用户的查询会话（query session）中，只有4%的会话使用了相关反馈功能，这大都是通过点击每个结果后面的“ More like this” 链接来实现的。大约70%的用户仅仅浏览了第一页的结果，对于使用相关反馈技术的用户来说，当时的检索效果大约提高了2/3。

相关反馈的问题

- **长的查询对典型的IR系统是低效的.**
  - 将结果返回给用户的耗时较长.
  - 检索系统的消耗大.
  - 能部分解决这个问题的方法:
    - 相关反馈时只对重要的查询词项重新计算权值
      - 比如按照词频，取前20
- 用户一般不太情愿提供明确的反馈
- 在使用了**相关反馈**之后，可能很难解释某个文档为什么会被返回

间接相关反馈

- 现实：用户不愿意反馈→无法使用显示相关反馈

- 在反馈过程中，我们也可以利用间接的资源而不是显式的反馈结果作为反馈的基础。这种方法也常常称为隐式相关反馈（**implicit relevance feedback**）。隐式反馈不如显式反馈可靠，但是会比没有任何用户判定信息的伪相关反馈更有用。
- 此外，尽管用户往往不愿意提供显式相关反馈，但是在一个如Web 搜索引擎一样的具有高访问量的系统中，**收集用户的大量隐式反馈信息是十分容易**的。

间接相关反馈

- 在 Web 上，DirectHit 引入**一种文档排序的思路**，即对于某文档，如果用户浏览的次数越多，那么它的排名也越高。换句话说，这里假设了用户对链接的**点击能够反映出该页面的相关性**。这种方法基于很多假设，比如结果列表中的文档摘要片段能够为用户判定文档的相关与否提供提示信息。
- **DirectHit将用户点击频率高的文档排在前面.**
  - 点击的多的页面被认为是相关的.
  - 从用户的点击记录中挖掘信息，进行相关反馈
- **这种方法是全局的，并不依赖特定用户或查询.**
  - 这是点击流挖掘（clickstream mining）的典型应用场景
- 现在这是通过机器学习产生排序的一部分

隐式相关反馈

- 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**非当前检索行为或非检索相关行为的分析**也可以用于提高检索的效果，这些是**个性化信息检索**(Personalized IR)的主要研究内容，并非本节的主要内容。

用户行为种类

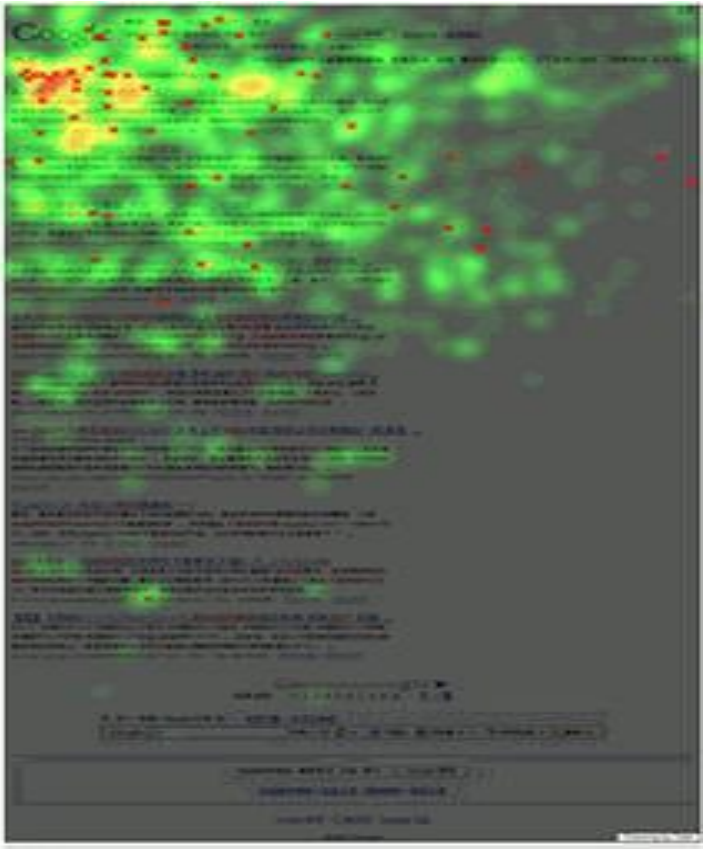
- **鼠标键盘动作：**
  - 点击链接、加入收藏夹、拷贝粘贴、停留、翻页等等
- **用户眼球动作**
  - Eye tracking可以跟踪用户的眼球动作
  - 拉近、拉远、瞟、凝视、往某个方向转

点击行为(Click through behavior)

FIELD	VALUE
User ID	1162742023015
Time stamp	06/Nov/2006:00:01:35
Query terms	嫁给警察的理由
URL	http://bbs.cixi.cn/dispbbs.asp?Star=4&boardid=46&id=346721&page=1
Page number	1
Rank	7
Anchor text	姑娘们，你们愿意<FONT color=#cc0033>嫁给警察</FONT>吗？[慈溪社区]

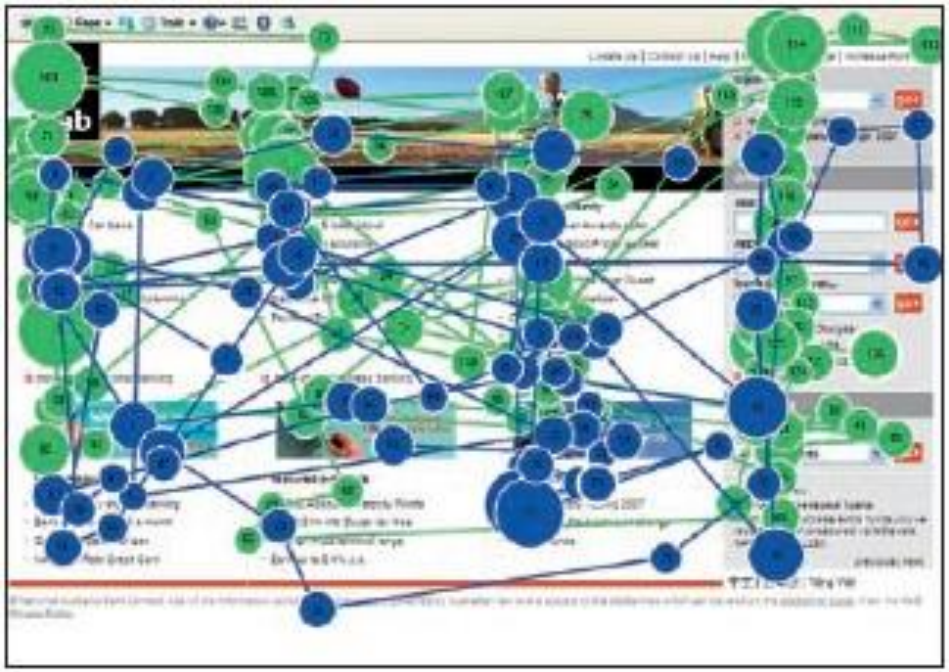
眼球动作(通过鼠标轨迹模拟)

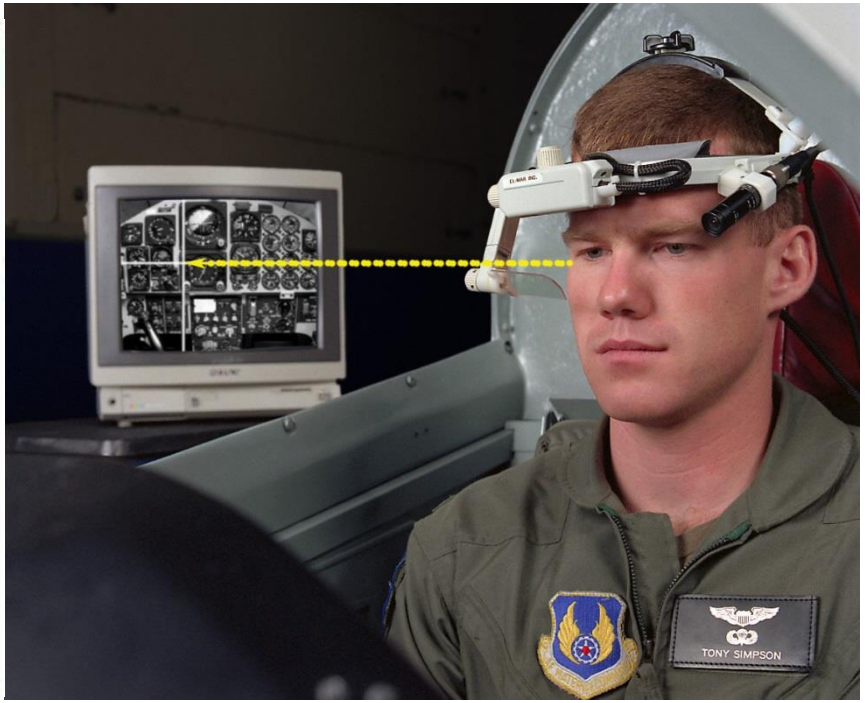
Baidu



Google

关于Eye tracking





小结：隐式相关反馈

- implicit relevance feedback
  - 隐式相关反馈，或称间接相关反馈
  - 利用间接的资源而不是显式的反馈结果作为反馈的基础
- 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

本讲内容

- 查询优化概述
- 相关反馈(relevance feedback)
  - 相关反馈概述
  - Rocchio 相关反馈算法
  - 隐式相关反馈
  - 伪相关反馈
  - 相关反馈的假设条件及评价方法
- 查询扩展(Query expansion)

伪相关反馈

- 伪相关反馈（pseudo relevance），也称为盲相关反馈（blind relevance feedback），**提供了一种自动局部分析的方法。**它**将相关反馈的人工操作部分自动化**，因此用户不需要**进行额外的交互就可以获得检索性能**的提升。
- **伪相关反馈的算法:**
  - 根据用户的查询，检索出结果列表
  - 假设列表中前k个结果是相关的.
  - 进行相关反馈(e.g., Rocchio)
- **平均效果很好，但对于某些查询可能错的很严重**。几步迭代后就可能出现严重的偏移。
- 为什么?

TREC4上的伪相关反馈实验

- 康奈尔大学利用SMART 系统在TREC 4 上的实验（Buckley 等人 1995），其中同时对比了两种长度归一化方式的结果（L 方式和l 方式）。**在进行伪相关反馈时每个查询增加20 个词项。**

	Precision at $k = 50$	
词项权重计算	无反馈	伪反馈
lnc.ltc	64.2%	72.7%
Lnu.ltu	74.2%	87.0%

ddd.qqq，前3 位代表文档向量的权重计算方法，而后3 位字母代表查询。第1 位字母表示权重计算中的tf 因子，第2 位表示df 因子，第3 位表示归一化形式。

词项频率tf	文档频率df	归一化方法
n(natural) $tf_{t,d}$	n(no) 1	n(none) 1
l(logarithm) $1+\log(tf_{t,d})$	t(idf) $\log\frac{N}{df_t}$	c(cosine) $\frac{1}{\sqrt{w_1^2+w_2^2+\cdots+w_M^2}}$
a(auamented) $0.5+\frac{0.5\times tf_{t,d}}{\max_t(tf_{t,d})}$	p(prob idf) $\max\left\{0,\log\frac{N-df_t}{df_t}\right\}$	u(pivoted unique) $1/u$(Section 17.4.4)
b(boolean) $\begin{cases}1 \text{ if } tf_{t,d}>0\\0 \text{ otherwise}\end{cases}$		b(byte size) $1/CharLength^a, a<1$
L(log ave) $\frac{1+\log(tf_{t,d})}{1+\log(ave_{t\in d}(tf_{t,d}))}$		

小结：伪相关反馈

- **算法流程**
  - 根据用户的查询，检索出结果列表
  - 假设列表中前k个结果是相关的.
  - 进行相关反馈(e.g., Rocchio)
- **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

本讲内容

- 查询优化概述
- 相关反馈(relevance feedback)
  - 相关反馈概述
  - Rocchio 相关反馈算法
  - 隐式相关反馈
  - 伪相关反馈
  - 相关反馈的假设条件及评价方法
- 查询扩展(Query expansion)

相关反馈的假设

- 相关反馈的成功依赖于某些假设。
- **A1：用户对于初始查询有充分的认识。**
  - 可能出现的问题：拼写错误、跨语言 IR、用户的词汇表和文档集的词汇表不匹配
- **A2：相关文档之间非常相似。**
  - 相关文档的词项分布相似
  - 不相关的文档的词项分布和相关文档的词项分布不相似
    - 所有相关文档都聚集在某个原型（prototype）周围，形成一个簇。
    - 或者: 有不同的原型，但是它们的词汇有很大重合。
    - 相关文档和不相关文档的相似度很小

不满足A1假设的情况

- 用户没有足够的知识来建立一个初始的查询。
- 比如:
  - 拼写错误(小田田布兰妮)
  - 跨语言的搜索(hígado)
  - 用户的词汇和文档集合里的词汇不吻合
    - 硬盘/磁碟
    - Laptop/Notebook

不满足A2假设的情况

- 相关文档聚成几个不同的簇
- 这种情况可能发生的情形:
  - 文档子集使用不同的词汇，如Burma/Myanmar（缅甸）
  - 某个查询的答案本身就需要不同类的文档来组成，如Pop stars that worked at Burger King
- 通用概念需要由多个具体概念体现
- 文档当中精心编辑的内容往往可以解决上述的问题

相关反馈策略的评价

- 使用初始查询 q_0，然后计算“查准率-查全率”曲线
- 使用相关反馈后修改的查询 q_m，然后计算“查准率-查全率”曲线
- **方法一、在整个文档集合上评价**
  - 有显著的改善,但是有作弊的嫌疑
  - 部分原因是会把已知的相关文档排的很前
  - 需要用用户没有看到的文档集合来评价
- **方法二、使用剩余的文档集合来评价(总的文档集合减去评价过相关性的文档)**
  - 评价结果往往比初始查询的结果差
  - 但是这种方法更现实
  - 可以用来有效地比较不同相关反馈方法之间的相对效果

相关反馈策略的评价

- **方法三、使用两个文档集合**
  - 在第一个文档集合上使用初始查询q_0 ，并进行相关反馈
  - 在第二个文档集合上使用初始查询q_0和修改过的查询q_m进行评价

- 从**经验上说，一轮相关相关反馈很有用。两轮相关反馈的效果就不那么明显。**
- 对于相关反馈的作用，最好的评价方法或许是进行用户调查，特别是是采用一种基于时间的比较方法：和采用其他方法（如查询重构技术）相比，用户采用相关反馈技术找到相关文档的时间是否更短？或者说，在一个固定的时间内用户能否找到更多的相关文档？这些代表用户效用性的指标最公平，也更贴近真实的应用。

评价的误区

- 评价不同相关反馈方法的效用的时候，必须考虑消耗时间的要素.
- 代替相关反馈的方法：用户修改并重新提交查询.
- 相对于判断文档的相关性，用户可能更愿意修改并重新提交查询.
- 没有证据能表明相关反馈占用了用户的时间就能给用户带来最大的效用.

小结：相关反馈的假设条件及评价方法

- 相关反馈的成功依赖于某些假设
  - A1：用户对于初始查询有充分的认识
  - A2：相关文档之间非常相似
- 评价：比较查询修改前后的 P-R曲线
  - 方法一、在整个文档集合上评价
  - 方法二、使用剩余的文档集合来评价
  - 方法三、使用两个文档集合

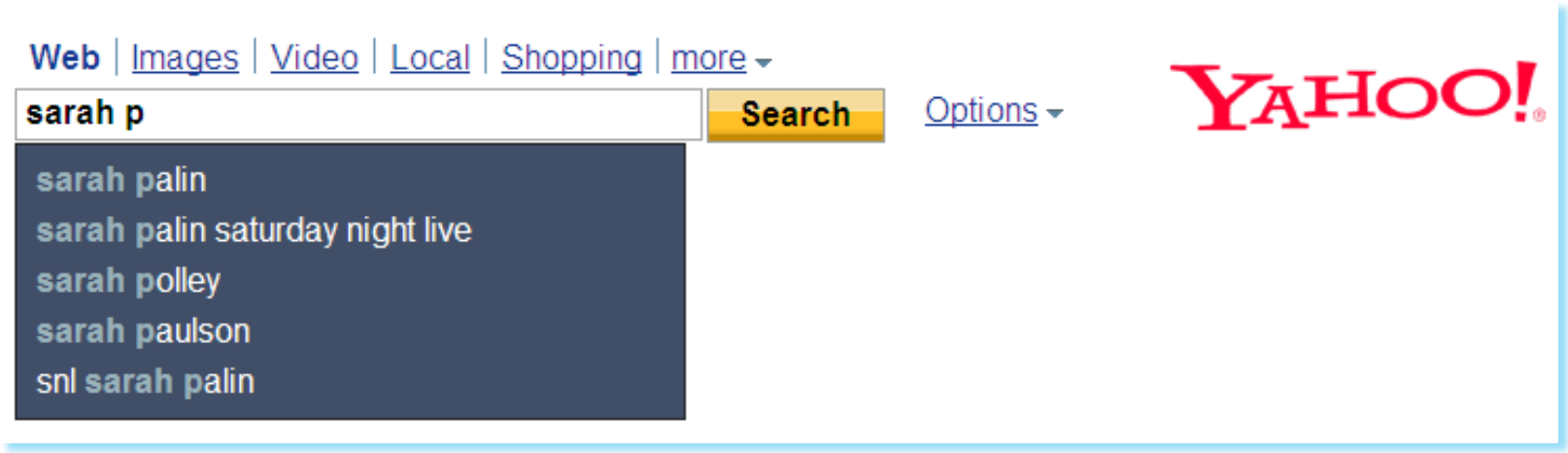
本讲内容

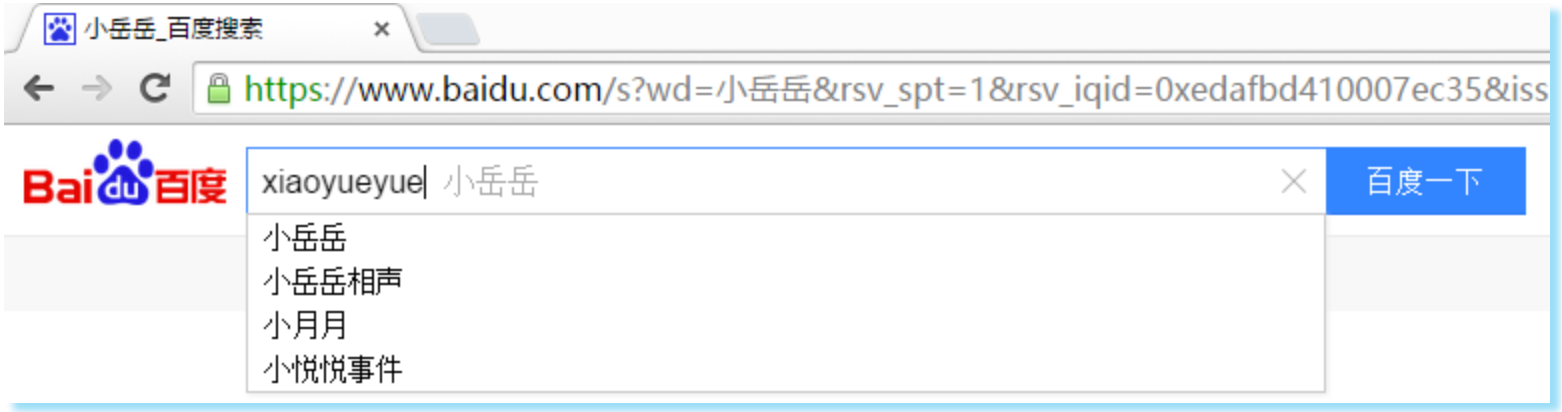
- 查询优化概述
- 相关反馈(relevance feedback)
  - 相关反馈概述
  - Rocchio 相关反馈算法
  - 隐式相关反馈
  - 伪相关反馈
  - 相关反馈的假设条件及评价方法
- 查询扩展(Query expansion)

查询重构

- **什么是查询重构？**
  - 在相关反馈中，通过判定文档相关和不相关，用户会给文档以额外的输入信息，该信息可以用于对查询词项进行权重的重新调整。另一方面，在查询扩展（query expansion）中，用户会对查询词或短语给出额外输入信息，比如可能推荐额外的查询词项。
- **查询重构可能的实现方法**
  - 简单辅助用户进行查询扩展
  - 采用人工词典的方法
  - 自动构建词典的方法

查询提示


Web | Images | Video | Local | Shopping | more
sarah p
Search
Options
YAHOO!
sarah palin
sarah palin saturday night live
sarah polley
sarah paulson
snl sarah palin


小岳岳_百度搜索
https://www.baidu.com/s?wd=小岳岳&rsv_spt=1&rsv_iqid=0xedafbd410007ec35&iss
Bai du 百度
xiaoyueyue 小岳岳
百度一下
小岳岳
小岳岳相声
小月月
小悦悦事件

查询推荐


小岳岳_百度搜索
https://www.baidu.com/s?wd=小岳岳&rsv_spt=1&rsv_iqid=0xedafbd410007ec35&issp=1&f=3&rsv_bp=1&rsv_idx=2&ie=utf-8&tn=baiduho
Bai du 百度
小岳岳
百度一下
网页 新闻 贴吧 知道 音乐 图片 视频 地图 文库 更多»
百度为您找到相关结果约8,900,000个
搜索工具
为您推荐: 小岳岳五环之歌 小岳岳春晚 小岳岳相声 小岳岳表情包
这个是我看过最逗的了~笑的肚子疼~小岳岳自己都憋不住 -频道:...
立即播放 时长: 39分钟
这个是我看过最逗的了~笑的肚子疼~小岳岳自己都憋不住 正...
v.pps.tv/play_3733...html - 百度快照
相关搜索
小岳岳相声 岳云鹏 小月月
欢乐喜剧人小岳岳 小岳岳五环之歌 小岳岳表情包
小岳岳 了不起的挑战 小悦悦 男孩撞脸小岳岳
1 2 3 4 5 6 7 8 9 10 下一页>
相关人物 展开
岳云鹏 曹云金 白慧明 郭麒麟
中国德云社相声演员 多次登上央视春晚舞台 于谦的现任妻子 郭德纲的大儿子
贾玲 苗阜 王自健 郭冬临
发起并创立了酷口相声 火车头艺术家 相声第二班创始人 小品演员央视春晚常客
相关艺人 展开
郭德纲 赵四 于谦 小黄飞

拼写校正

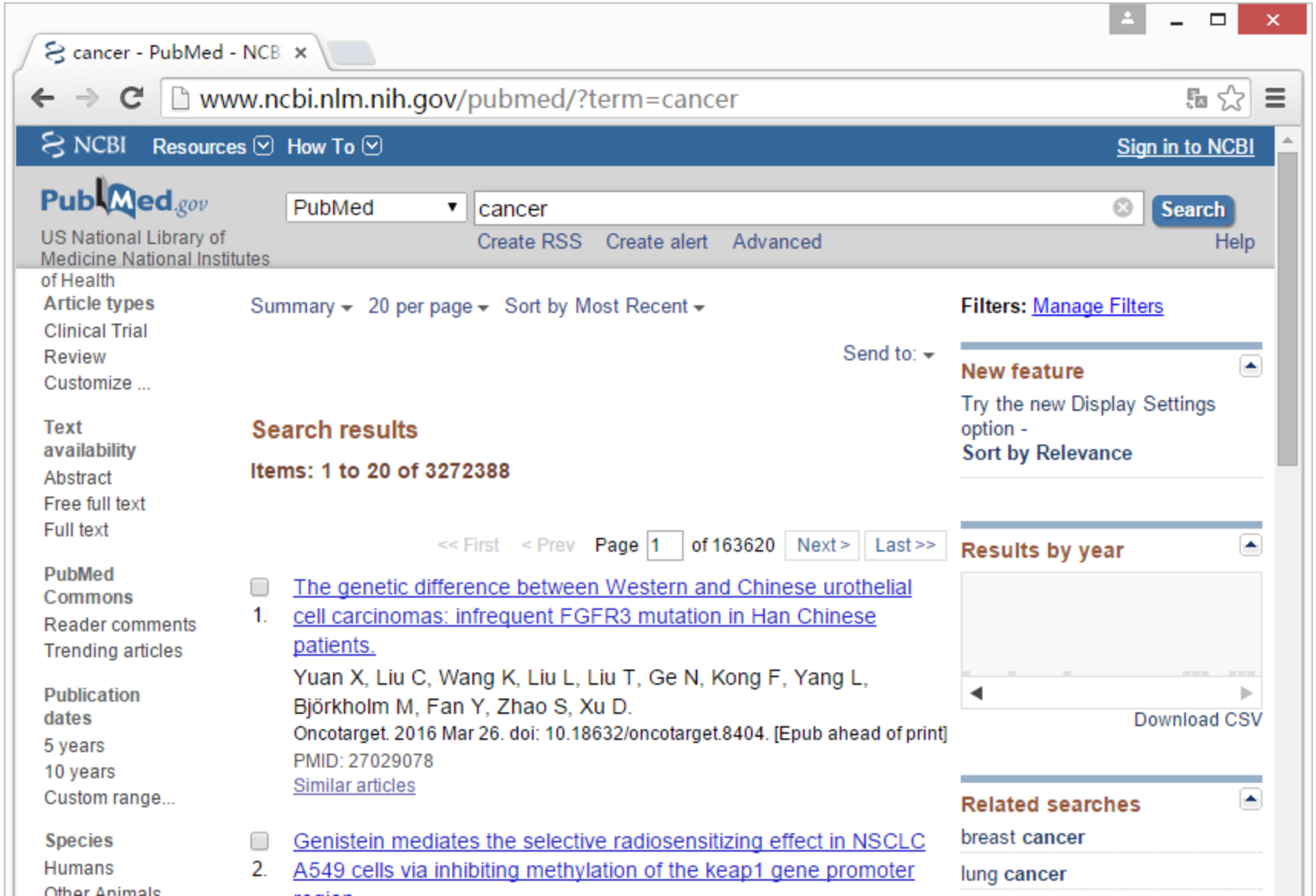

zootopiann_百度搜索
zootopai - 必应
cn.bing.com/search?q=zootopai&go=提交&qs=n&form=QBRE&pq=zc
必应
zootopai
网页 图片 视频 学术 词典 网典 地图 更多
提示:当前显示为 全部结果 | En 英文搜索 | 仅中文
包含 zootopia 的结果。
是否只需要 zootopai 的结果?
疯狂动物城 高清预告片 - 豆瓣电影
上映：2016-03-04
类型：喜剧 · 动作 · 动画 · 冒险
导演：拜伦-霍华德 · 瑞奇-摩尔
主演：金妮弗-古德温 · 杰森-贝特曼 · 伊德里斯-艾尔巴
剧情：故事发生在一个所有哺乳类动物和谐共存的美好世界中，兔子朱迪（金妮弗·古德温 Ginnifer Goodwin 配音）从小就梦想着能够成为一名惩恶扬善的刑警，凭借着智慧 ... 豆瓣电影
9.3
豆瓣评分

如何扩展用户的查询?

- 利用人工编纂的同义词辞典
  - E.g. MedLine: physician, 同义词: doc, doctor, MD, medico
  - 这些同义词可以作为查询
- 全局的分析: (static; of all documents in collection)
  - 同义词辞典的自动生成
    - 统计词汇的共现（co-occurrence）
  - 利用对查询日志的挖掘进行优化
    - Web中最常用的
- 局部的分析: (动态的)
  - 分析查询的结果文档集合

人工编纂的范例

cancer → cancer OR neoplasms



利用同义词辞典进行查询扩展

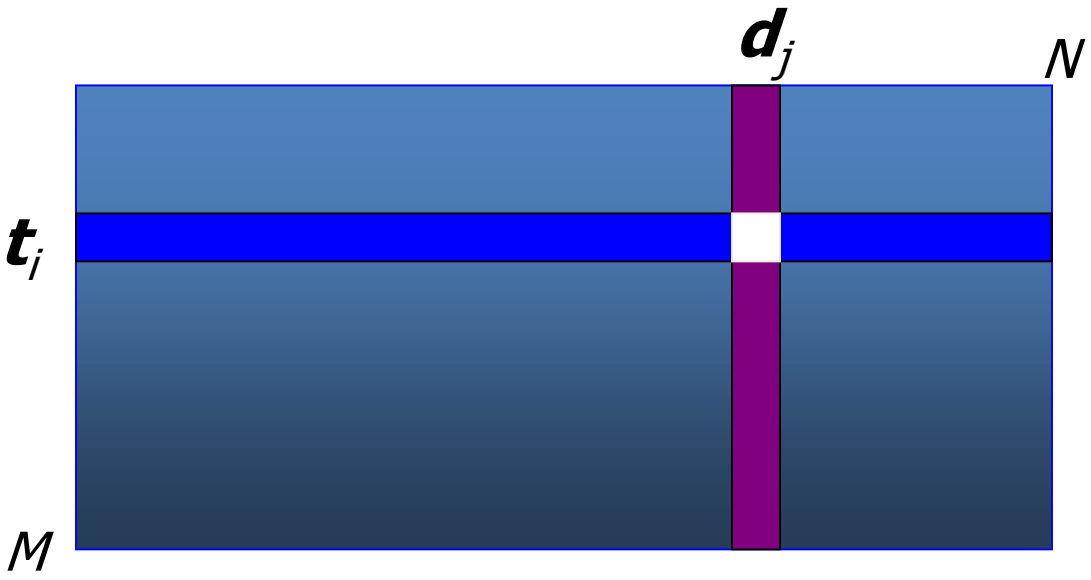
- **对查询词汇 *t*, 使用辞典中的同义词或者词汇进行扩展**
  - feline → feline cat
- 相对于原始的查询词汇，可以给扩展的词汇分配更小的权重.
- 通常可以增加查全率
- 在科研和工程领域广泛应用
- 可能会明显地降低查准率,特别是对于含混不清的查询词汇.
  - “interest rate” → “interest rate fascinate evaluate”
- 人工编纂同义词辞典需要很大代价

同义词辞典的自动生成

- 通过分析文档集合，可以得到两个词的相似性
- 定义1: 如果两个词经常和同样的词同时出现，那么这两个词相似.
- 定义2: 如果两个词经常和同样的词在某种语法关系里出现，那么这两个词相似.
- 你可以“削，吃，收获”“苹果，梨”，那么“苹果”和“梨”就相似.
- 简单采用词的共现更鲁棒，但采用语法关系更准确。

共现同义词辞典

- 计算词和词之间相似性的简单方法： $C = AA^T$ ， A 是词项-文档矩阵。
- $w_{i,j} = (t_i, \boldsymbol{d}_j)$的归一化的权值，使得A中的行向量长度为1
- 对每个词项t_i选择C 中相似度最高的词项作为同义词



如果A是一个词项-文档出现矩阵（0/1矩阵），C中的元素为?

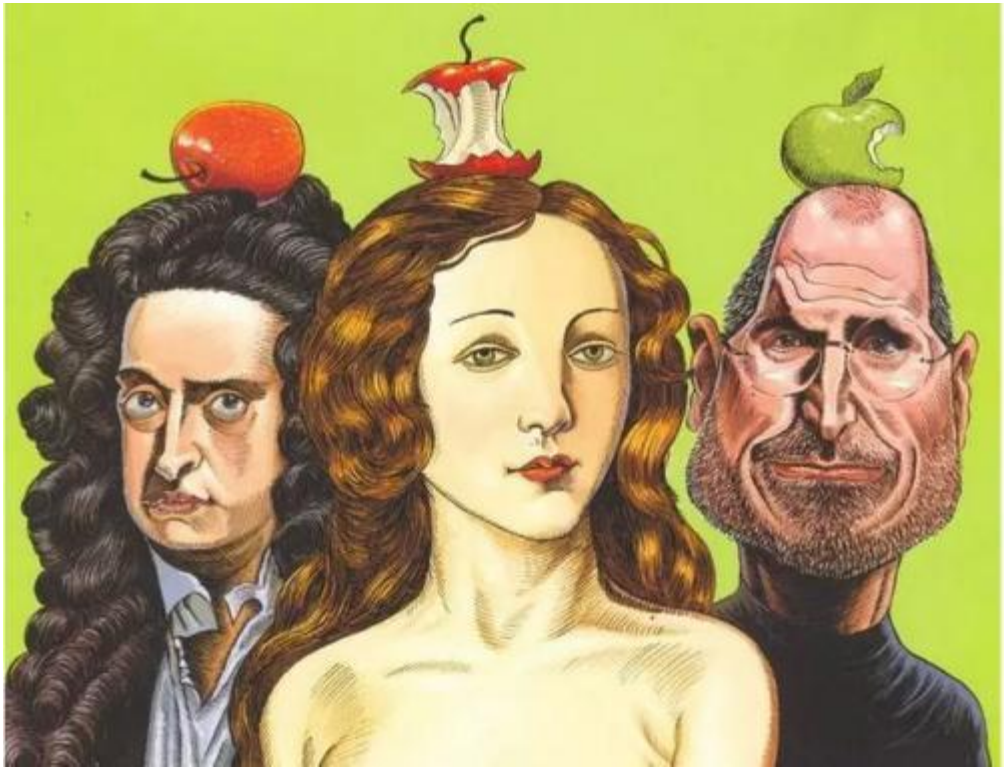
自动生成同义词辞典的例子

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

自动生成同义词辞典的讨论

- 词项关联的质量是一个问题.
- 有歧义的查询词可能会引入统计上相关，但意思不相关的词.
  - “Apple computer” → “Apple red fruit computer”
- 问题:
  - 假正率（False positives）: 不相似的词被认为相似
  - 假负率（False negatives）: 相似的词被认为不相似
- 由于扩展的查询词和原查询词很相关，扩展的查询也未必能得到更多的相关文档.

与“苹果”的统计相关



第一个是圣经中亚当和夏娃吃掉的那个苹果（改变了地球）
第二个是砸在牛顿头上的苹果（改变了物理界）
第三个是乔布斯手中被咬了一口的苹果（改变了数码界）
第四个？

小结：查询扩展(Query expansion)

- 什么是查询重构？
  - 针对查询词给出额外输入信息
- 查询重构可能的实现方法
  - 简单辅助用户进行查询扩展
    - 查询提示、查询推荐、拼写校正
  - 采用人工同义词词典的方法←代价很大
  - 自动构建同义词词典的方法
    - 一种方法是简单地使用词共现信息。
    - 另一种方法是采用浅层语法分析器来分析文本得到词汇之间的语法关系或语法依存性。

本讲内容要点

- 查询优化：对查询进行修改
- 相关反馈(relevance feedback)
  - 在初始检索结果的基础上，根据用户交互指定相关或不相关，或采用其他方法指定
  - Rocchio 相关反馈算法←最著名的方法
  - 隐式相关反馈
  - 伪相关反馈
  - 相关反馈的假设条件及评价方法
- 查询扩展(Query expansion)
  - 通过在查询中加入同义或者相关的词项来提高检索结果

课后练习

- 习题9-3
- 习题9-4
- 习题9-7