计算机信息检索

第5章信息过滤(Information Filtering)

课前思考题

- 信息过滤的概念是什么？它和一般的信息检索、信息分类、信息抽取有什么区别？
- 信息过滤的类型有哪些？基于内容的过滤和基于协作的过滤有什么不同？
- 信息过滤的构成和各部分功能如何？
- 信息过滤系统如何评估？

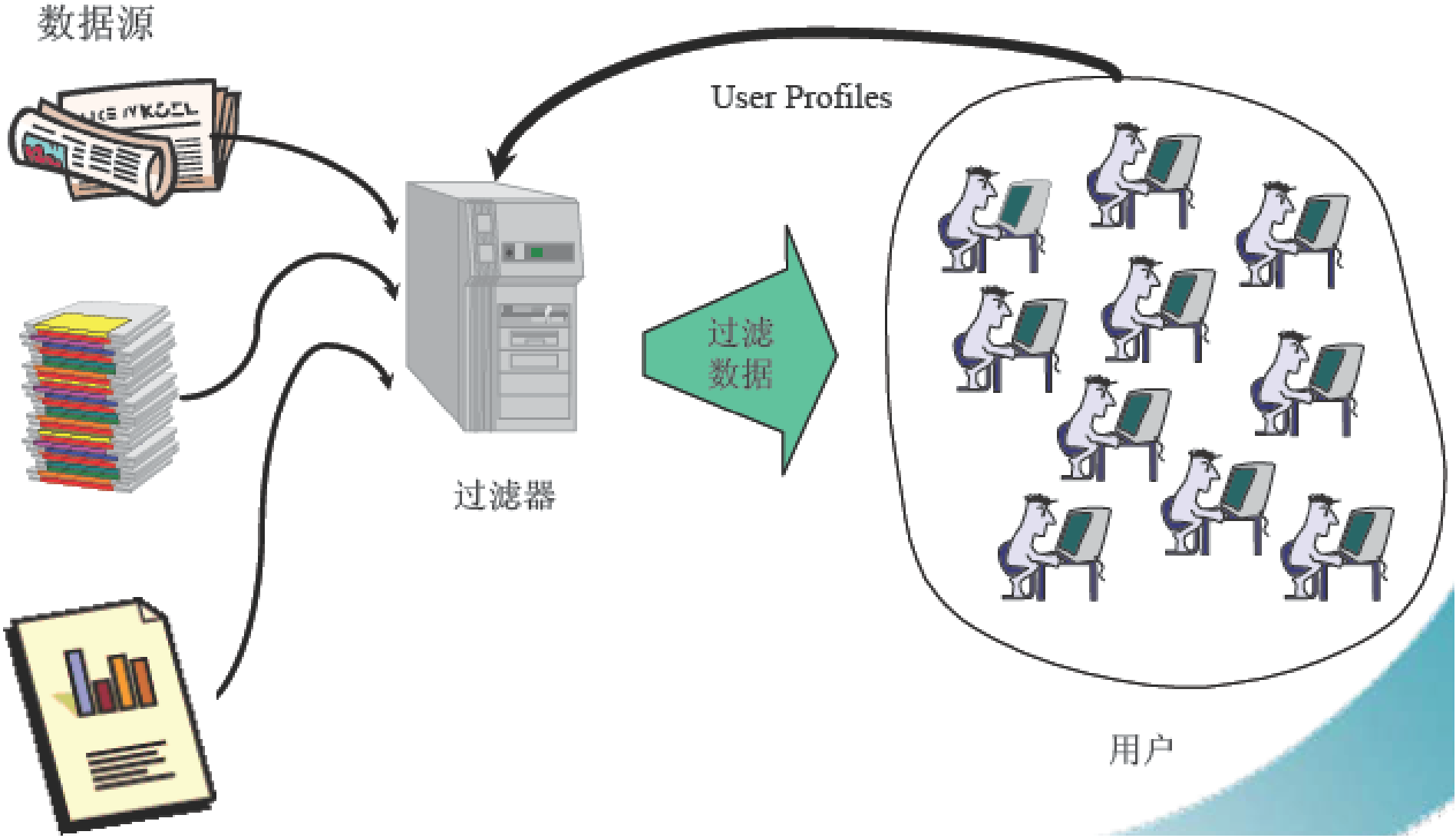
提纲

1. 信息过滤的基本概念
2. 信息过滤系统的分类
3. 信息过滤系统的组成
4. 信息过滤系统的评估
5. 信息过滤的现状及发展趋势

信息过滤的定义

- 从*动态*的信息流中将满足用户兴趣的信息挑选出来，用户的兴趣一般在较长一段时间内比较稳定不会改变(静态)。
- 其他名称：
  - Selective Dissemination of Information(SDI)，定题服务，来自图书馆领域。
  - Current Awareness, 来自Data Mining。

信息过滤系统示意图



信息过滤系统的特点

- 新信息的产生速度很快，人的兴趣变化速度赶不上信息的变化速度。可以说，人的兴趣变化比较缓慢，可以看成相对静态的和稳定的。
- 信息过滤主要借用信息检索和用户建模(User modeling)两个领域的技术。
- 用户的需求或者兴趣通常采用User Profile建模来表示。
- 新信息到来的时候，根据用户的User Profile，有选择地挑出信息给用户。

信息过滤系统数据流图

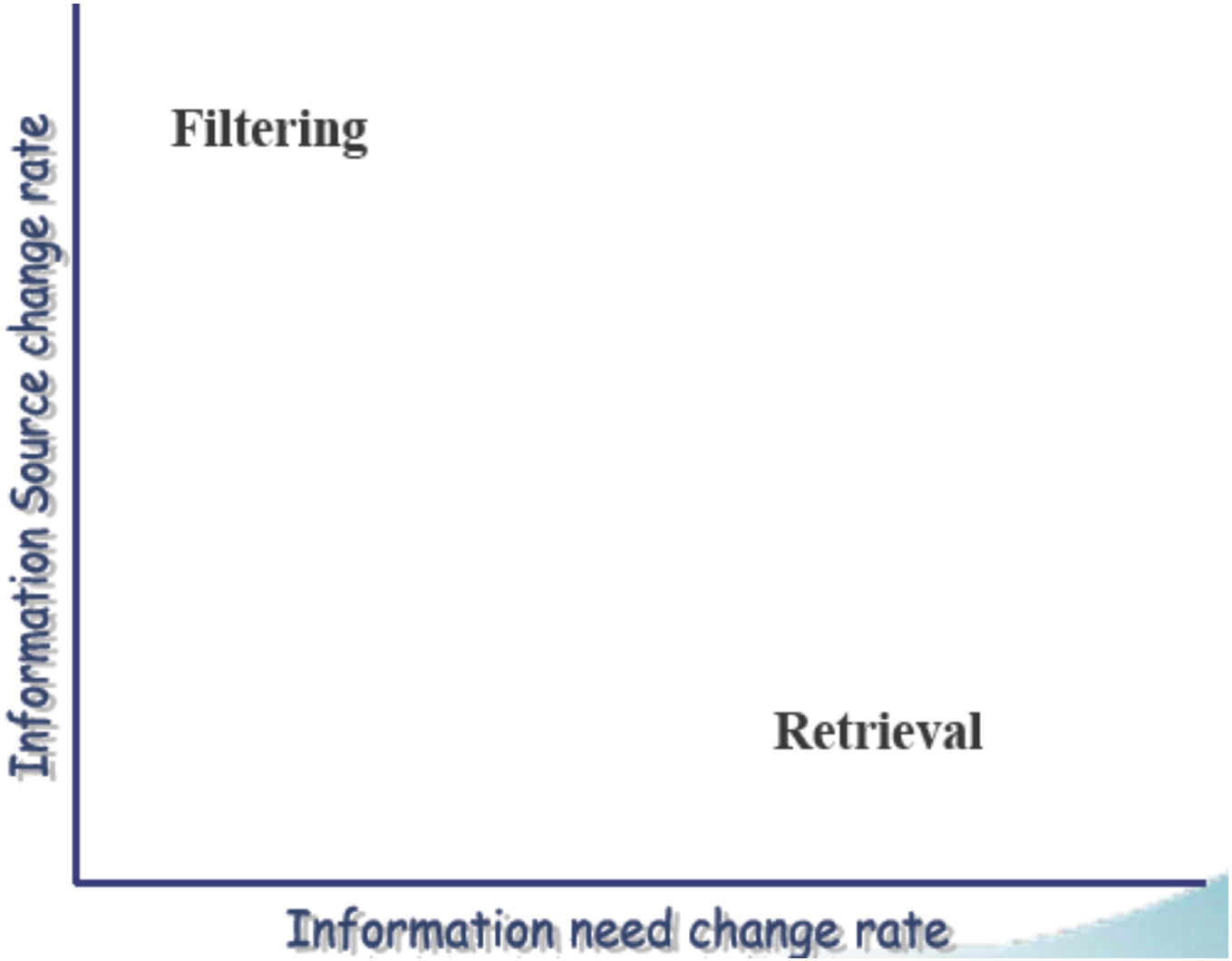
- **Collection**
- **Selection**
- **Display**

看上去很象IR！

IF vs. IR (1)

- **IF是可以看成广义IR的一部分，即和Adhoc Retrieval相对的一种任务模式。IR通常采用Pull模式，而IF通常采用Push模式。**
- **IF一般都借用狭义IR中的表示和计算方法。**
- **和Adhoc Retrieval相比：**
  - IR可以认为面向一次性的查询而使用，而IF是面向用户的长期需求的重复使用
  - IF信息源动态，用户需求(采用User Profile来表示)相对静态；IR检索信息源相对静态，用户需求(采用Query来表示)动态变化
  - IF用户要对系统有所了解，IR不需要。
  - IF一般要关注用户建模，涉及用户隐私问题。而IR一般不需要。

IF vs. IR (2)



IF vs. IC (Info. Classification)

- **IF可以采用IC中的分类算法。**
- **某些场合下人们所称的“信息过滤”实际就是一个IC问题。如不经过用户Profile调整的垃圾邮件过滤。**
- **IC中的Category通常不会变化，相对而言，IF的User Profile会动态调整。**

IF vs. IE

- **Information Extraction是从无格式数据源中抽取相关字段的过程。比如抽取恐怖事件的时间、地点、人物等字段。**
- **IE中不太关注相关性，而只关注相关的字段。IF中要关注相关性。**

IF 的一些应用

- 搜索引擎检索结果的过滤：Google
- 个人的邮件过滤
- 新闻订阅和过滤
- 浏览器过滤
- 面向儿童的过滤系统
- 面向客户的过滤系统和推荐系统

提纲

1. 信息过滤的基本概念
2. 信息过滤系统的分类
3. 信息过滤系统的组成
4. 信息过滤系统的评估
5. 信息过滤的现状及发展趋势

按信息过滤方式分

- **主动(Active)的IF系统**
  - 主动搜集信息，并将相关信息发送给用户
  - 通常采用Push操作
  - 会造成信息过载问题，所以该系统要尽力建立精确的User Profile。
- **被动(Passive)的IF系统**
  - 不负责为用户搜集信息
  - 通常用于邮件和新闻组信息过滤

按信息过滤地点分

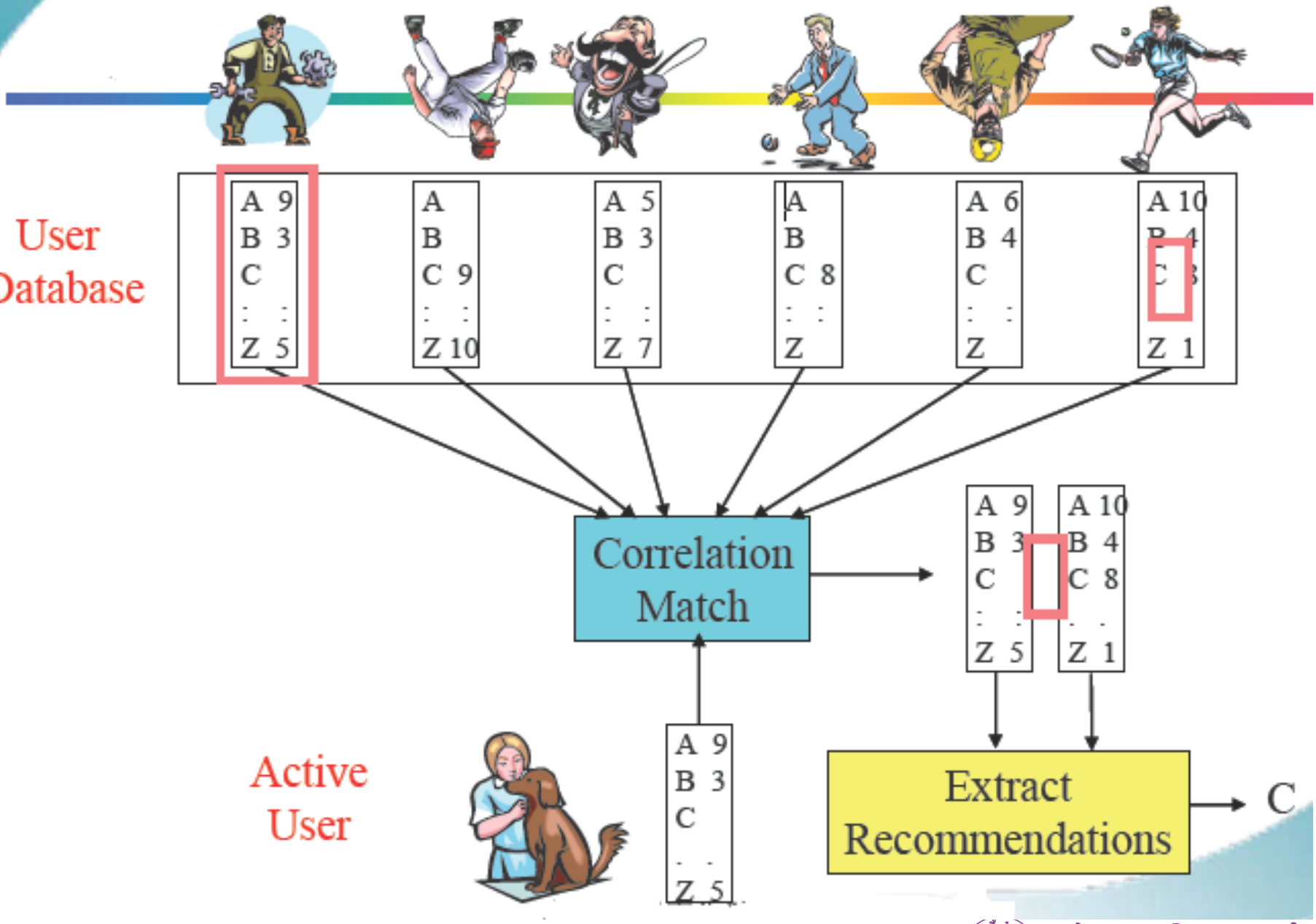
- **在信息源端过滤**
  - 将用户的Profile发送给信息提供者，后者将和用户Profile匹配的信息回送给用户
  - 这种服务通常也称为Clipping service
  - 用户通常需要付费，代表系统：Dialog的Alert服务
- **在过滤服务器端过滤**
  - 信息提供者将信息发送给过滤服务器
  - 过滤服务器根据用户的Profile将匹配信息发给用户
  - 代表系统SIFT
- **在用户端过滤**
  - 是一个局部过滤系统
  - 如Foxmail或outlook的过滤功能。

从过滤方法分

- **基于感知的过滤(Cognitive filtering)**
  - 也称为基于内容的过滤(Content-based filtering)
  - 将文档内容和用户的Profile进行相似度计算
  - 代表系统CiteSeer
- **基于社会的过滤(Sociological filtering)**
  - 也称为协同过滤(Collaborative filtering）
  - 对某个用户的Profile进行匹配时，通过用户之间的相似度来计算Profile和文档的匹配程度
  - 基于社会过滤的系统常常称为推荐系统(Recommendation systems)
  - 社会过滤常常使用用户建模(User modeling)及用户聚类(User clustering)等技术。
  - 社会过滤一般不单独使用，常常和基于内容的过滤配合使用。
  - 代表系统：RINGO、GroupLens

社会过滤的一个实际例子

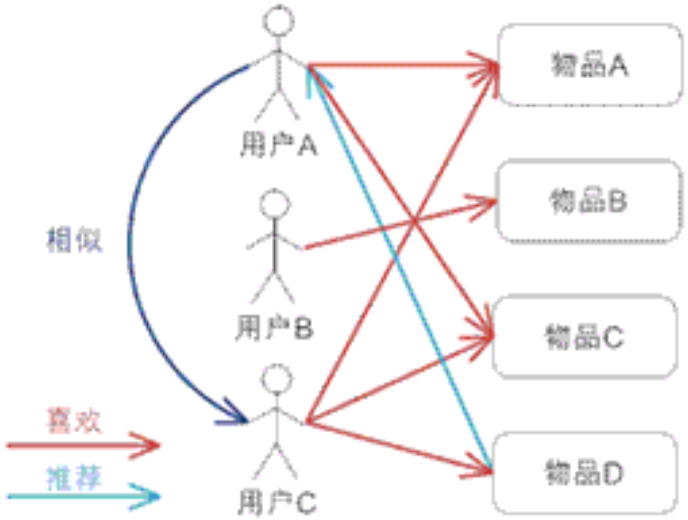
	书1	书2	书3	书4	书5	书6
用户1	√	√		√		√ ?
用户2	√ ?	√		√		√
用户3	√		√		√	√ ?
用户4	√		√		√ ?	√



基于用户的 CF 的基本原理

- 基于用户对物品的偏好找到相似用户，然后将相似用户喜欢的物品推荐给当前用户；
- 用户对所有物品的偏好作为一个向量，来计算用户之间的相似度；
- 根据相似用户的对物品的偏好，预测当前用户没有偏好的未涉及物品

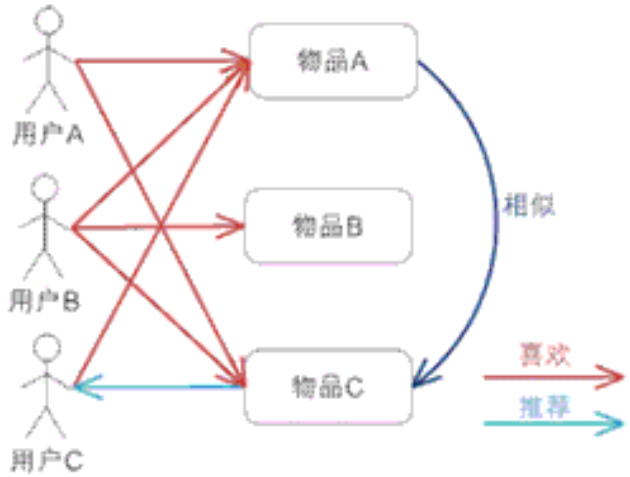
用户/物品	物品A	物品B	物品C	物品D
用户A	√		√	推荐
用户B		√		
用户C	√		√	√



基于物品的 CF 的基本原理

- 基于用户对物品的偏好找到相似的物品，然后根据用户的历史偏好，推荐相似的物品给用户；
- 所有用户对某个物品的偏好作为一个向量来计算物品之间的相似度；
- 根据用户历史偏好预测当前用户还没有表示偏好的物品。

用户/物品	物品A	物品B	物品C
用户A	√		√
用户B	√	√	√
用户C	√		推荐



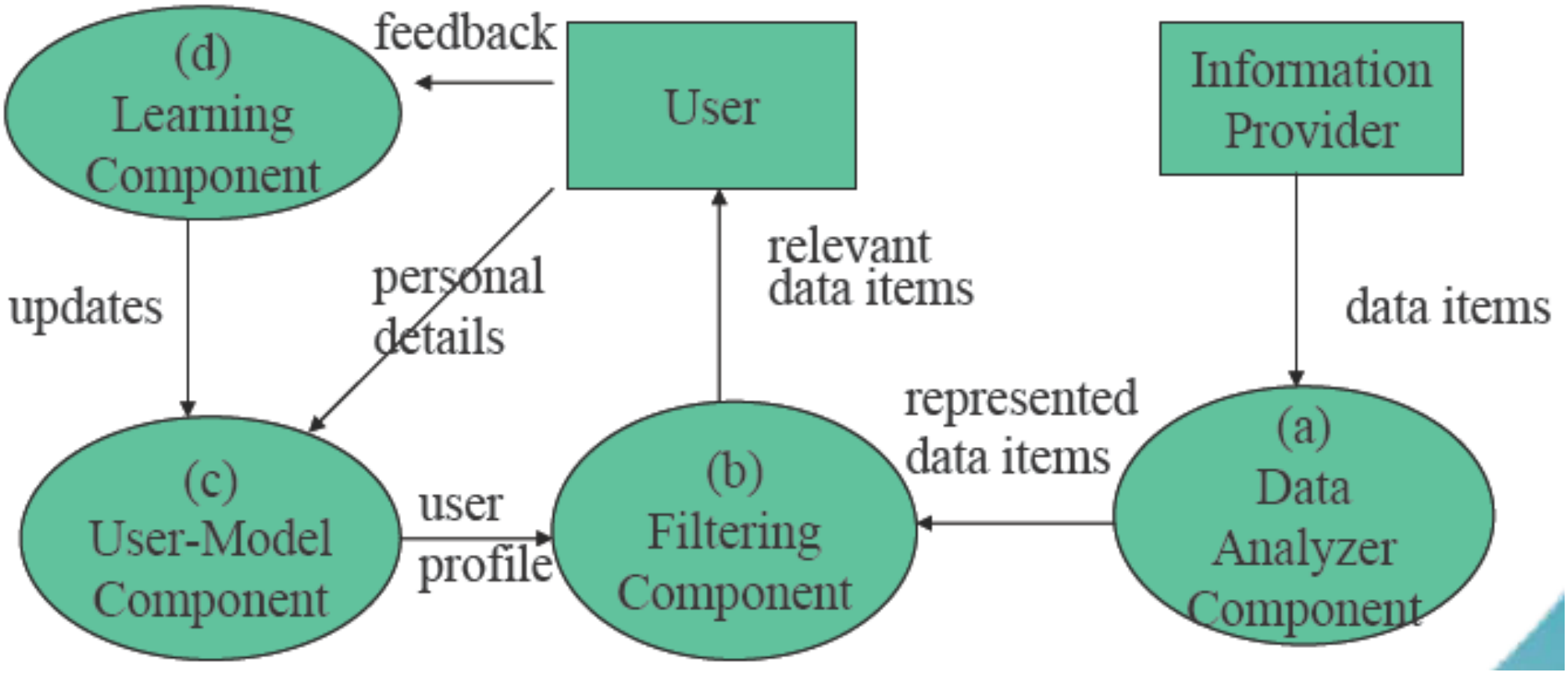
从获得用户兴趣的方法分

- **显式方法**
  - 用户填写表格或用户提交关键词
  - 代表系统：SIFT、BackWeb
- **隐式方法**
  - 记录用户的行为，包括：时间、次数、上下文、行为(保存、废弃、打印、浏览、点击)等。
  - 代表系统：GroupLens
- **介于显式和隐式之间的方法**
  - 文档空间方法：将用户标注过的文档作为正例，新来的文档和它们比较，选择相似度大的文档。
  - 代表系统：SIFTER
- **显式和隐式相结合的方法**
  - 原型参考（Stereotypic inference）：开始定义一些默认的Profile，根据用户的过滤过程进行修改。
  - 代表系统：UM

提纲

1. 信息过滤的基本概念
2. 信息过滤系统的分类
3. 信息过滤系统的组成
4. 信息过滤系统的评估
5. 信息过滤的现状及发展趋势

一般组成



Data-analyzer component

- 靠近信息提供方
- 从信息提供方获得或搜集数据
- 分析文档并将文档转化成相应表示(如布尔模型表示、向量空间模型表示等等)
- 将上述表示传给过滤模块

User-model component

- 显式或隐式地获得用户的一些相关信息
- 构建用户Profile模型(规则表示模型、向量模型、文档中心模型等等)
- 将用户模型传给过滤模块
- 用户模型必须要和文档表示模型具有可比性

用户建模不仅仅用于过滤(Beyond Filtering)

- 搜人：基于用户行为和特征
- 发现潜在的合作者
- 数据挖掘
  - 用户分类与广告投放
- 发现用户的兴趣转移

Filtering component

- IF系统的核心模块
- 将User Profile和文档的表示进行相似度计算
- 做出二值判定或者根据概率大小将文档进行排序
- 用户可以对过滤结果进行判定
- 判定信息传给学习模块以便对用户的Profile进行调整。

Learning component

- 根据用户的反馈信息对用户的**Profile**进行调整，以便提高以后的过滤效果
- 检测用户的兴趣转移

IF系统中的两个概念

- 基于统计的系统(System based on the statistical concept)
- 基于知识的系统(System based on the knowledge-based concept)

基于统计的IF系统

- **用户建模模块:**
  - Profile采用Term的权重向量来表示(如VSM, LSI)
- **过滤模块:**
  - 相关系数计算，Cosine距离
  - 概率检索模型(PRM)
  - 采用Bayes分类器进行计算
- **学习模块**
  - 进行相关反馈和查询重构(如采用Rocchio公式)

基于知识的IF系统

- **采用规则(Rule-based)或者语义网(Semantic-nets)的过滤系统**
  - 规则：如果…那么…
- **User profile采用语义网(如利用wordnet)**
  - 基于神经网络的过滤系统
  - 基于遗传算法的过滤系统

IF系统中的用户建模

- **建模数据的获取办法：**
  - 显式方法: 填写表格，直接交互
  - 隐式方法：对用户行为的观察
- **模型中的数据：**
  - 浅层语义：如关键词
  - 增强的用户模型中包含更多关于用户的高级知识(如背景经历)
- **采用构架(Underlying Architecture)**
  - Agent/neural networks for auto inferred model
  - VSM/LSI for explicit inference
  - Concept model for intelligent systems
  - Keyword system for statistically-based systems

IF系统中的学习

- **学习方法**
  - 基于观察进行学习
  - 基于反馈进行学习
  - 基于用户的训练进行学习(user-train learning)
- **学习频率(Frequency of learning)**
  - 出现紧急情况下的学习(Critical learning)
  - 定期学习

提纲

1. 信息过滤的基本概念
2. 信息过滤系统的分类
3. 信息过滤系统的组成
4. 信息过滤系统的评估
5. 信息过滤的现状及发展趋势

IF系统的评估方法

- **Evaluation by Experiments**
- **Evaluation by Simulation: such as TREC**
- **Analytical Evaluation**

评估指标(1)

- **正确率和召回率(Precision & Recall)**
- **基于统计的评价指标**
  - 相关系数(Correlation)：用户评估的结果排序和系统评估的结果排序的序相关系数
- **其他基于集合的评价指标**
  - Utility=(A*R+)+(B*N+)+(C*R-)+(D*N-), R+N+R-N-分别表示选出来的结果中真正相关文档的个数、不相关文档的个数、未选出来结果中相关文档的个数及不相关文档的个数，A、B、C、D是加权系数。
  - ASP(average set precision)=P*R, 当P or R=0, ASP 不可用

评估指标(2)

- **面向用户(User-oriented)的指标**
  - Coverage Ratio=|Rk|/|U|=|A∩U|/|U|, A是用户找出的文档集合，U是用户已知的相关文档集合，Rk是系统找出的用户已知的相关文档集合
  - Novelty=|Ru|/(|Ru|+|Rk|)，Ru是系统找出的用户未知的相关文档集合

提纲

1. 信息过滤的基本概念
2. 信息过滤系统的分类
3. 信息过滤系统的组成
4. 信息过滤系统的评估
5. 信息过滤的现状及发展趋势

现状

- IF 系统不可缺少
- 但是目前的IF系统并不十分可靠(unreliable)
- 商用的IF系统的相关度在50%左右
- TREC实验的结果也不尽如人意
- 用户宁愿读一些不相关信息，也不愿意丢掉重要相关信息
- 还有很长的路要走。

关于用户建模

- 集成各种方法来表示用户的兴趣(不仅仅是关键词、还应该包括用户的一些特性或者参数)
- **Profile**更新及更新时间
- 必须包含一个学习模块
- 必须跟踪用户兴趣随时间的变化

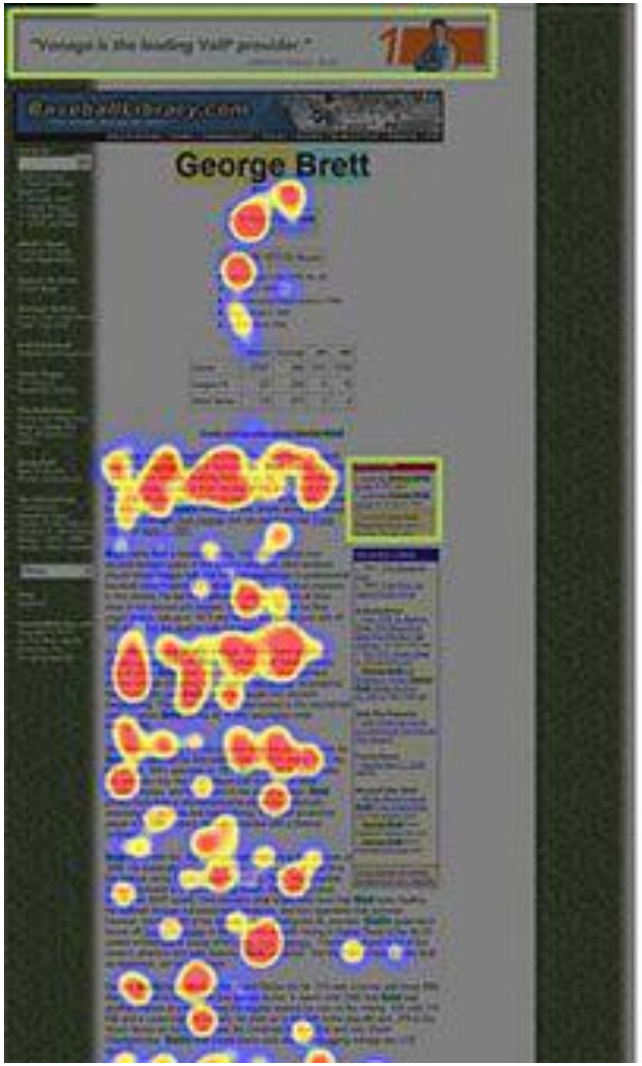
关于过滤技术

- 目标：宁愿返回一些不相关文档，也要返回更多的相关文档
- 应该走多种方法相结合的道路。
- 研究方向：
  - 智能过滤Agent：非集中式，基于信用，Agent之间互相竞争和合作，也不断进化
  - 可视化技术
  - 多媒体过滤：如视频点播VOD, not text-based
  - 多语言过滤(multilingual filtering)

IF中其他需要考虑的问题

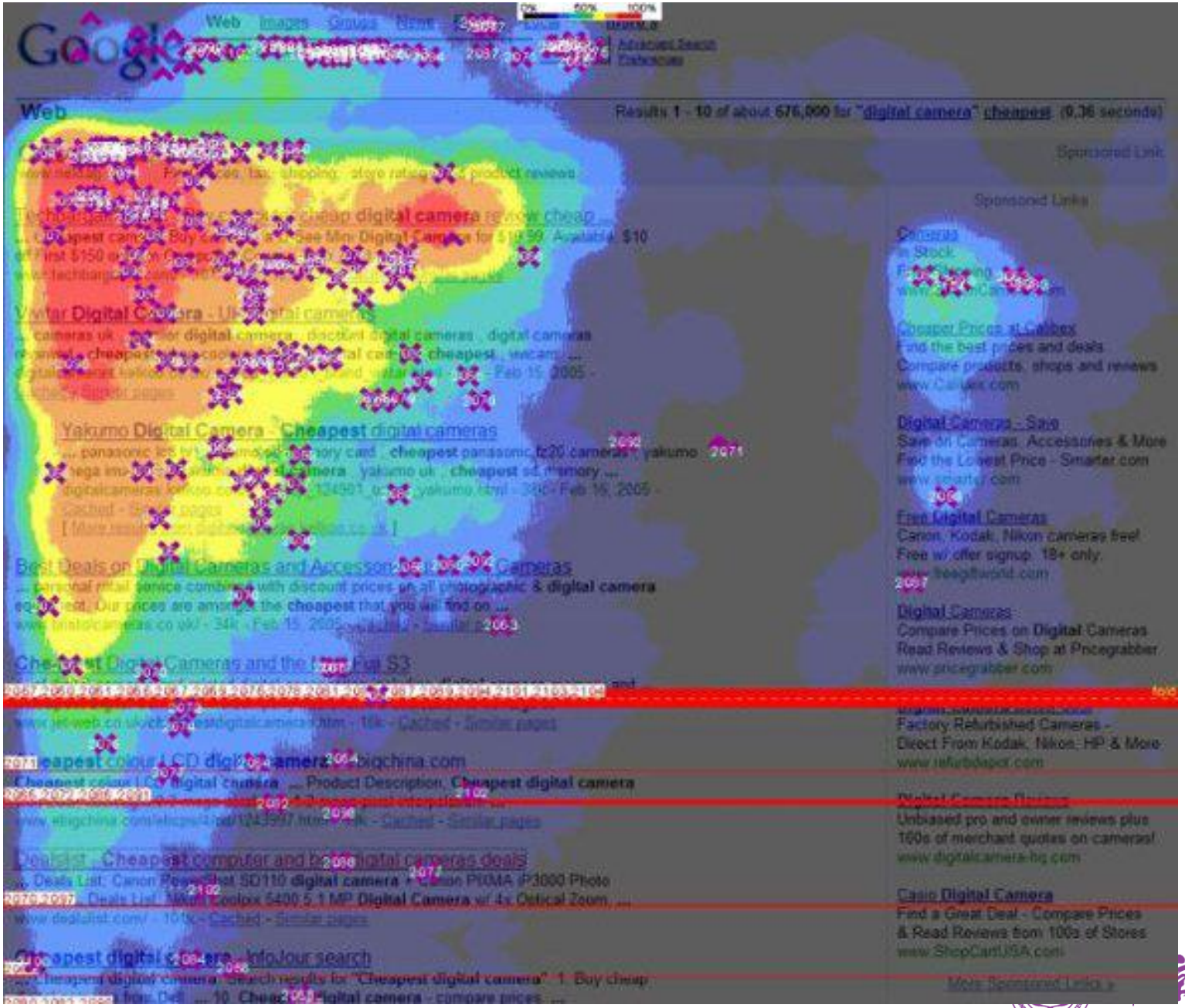
- **Protecting privacy**
  - What absolute assurances can we provide?
  - How can we make remaining risks understood?
- **Non-cooperative users**
  - How can the effect of spamming be limited?

Banner盲点



Facebook上人们眼神停留最多的地方

Google前五条结果最受关注



即使放一张再大的脸，人们也会看左边的文字

简历最初6秒关注的地方

1. Professional experience

TECHNOLOGY INNOVATION EXECUTIVE

PROFESSIONAL EXPERIENCE

VICE PRESIDENT

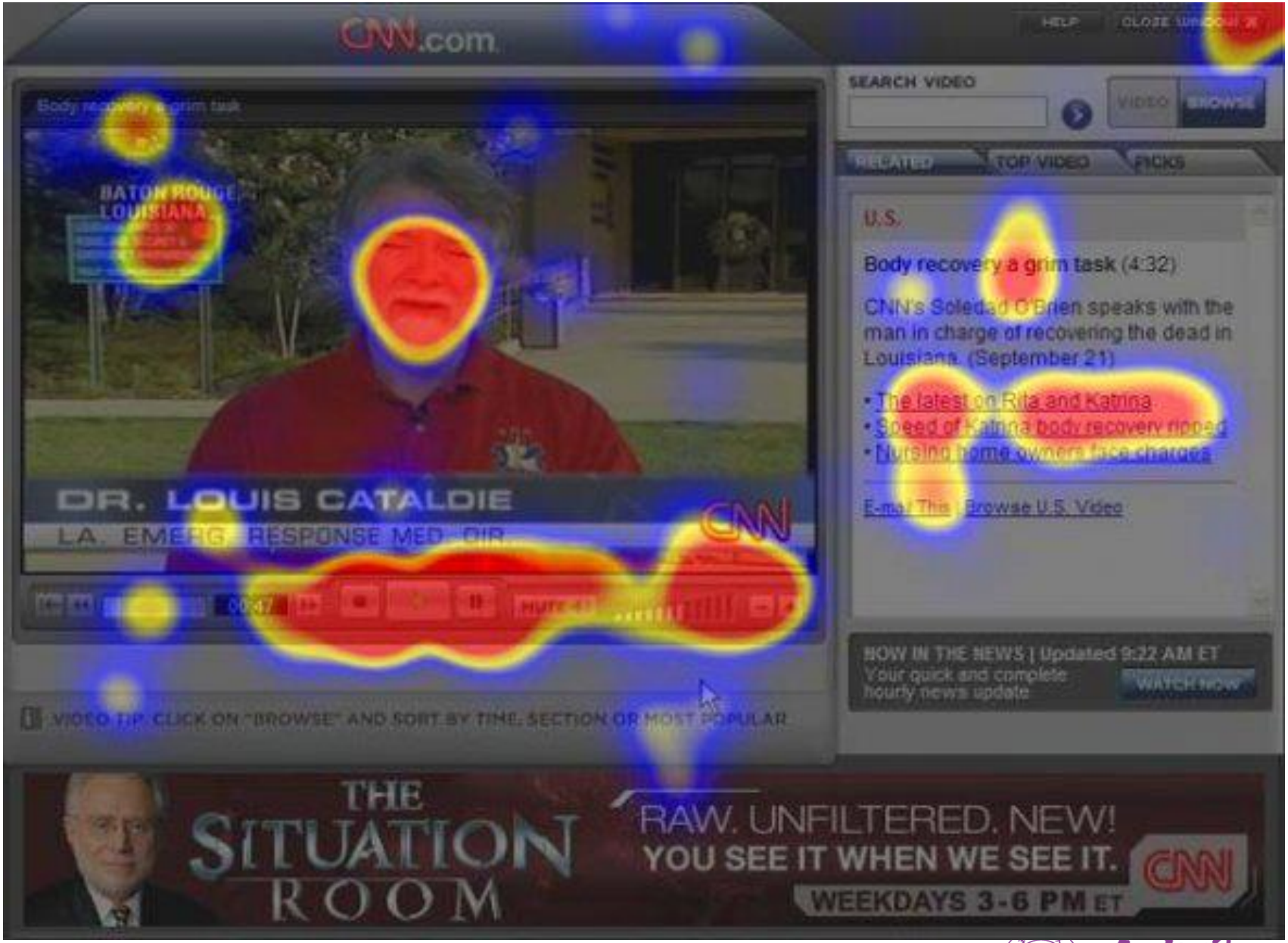
DIRECTOR – NETWORK ARCHITECTURE

PERFORMANCE ANALYSIS CONSULTANT

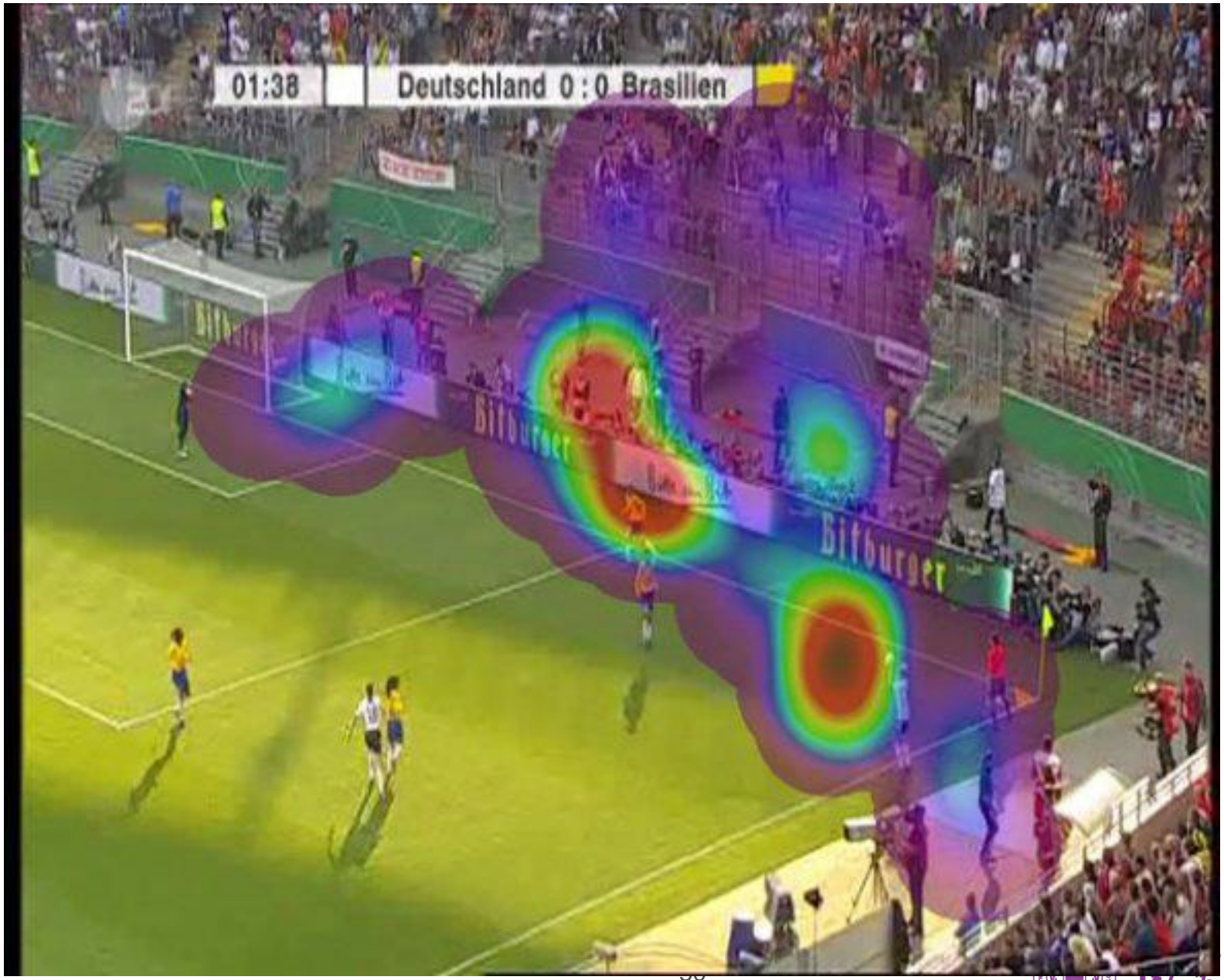
EDUCATION

聚焦在广告页的图片和大标题

没人看下一行的广告



足球比赛，人们还会关注对面的观众席



The End