根据学校课堂纪律的要求

请同学们坐在前五排

数字媒体技术基础

Meng Yang

www.smartllv.com

SUN YAT-SEN University

机器智能与先进计算教育部重点实验室

智能视觉语言学习研究组

Course Outline

- 9 数字媒体检索技术
- 9.1 基于标签的媒体检索
- 9.2 基于内容的媒体检索
- 9.3 多媒体内容的索引
  - 9.3.1 静态相关性TF-IDF
  - 9.3.2 动态相关性PageRank
- 9.4 媒体检索反馈技术

目 录

数字媒体检索技术

9.3 多媒体内容的索引

9.3.1 静态相关性 TF-IDF

- TF-IDF同时考虑了词频和逆文档频率，计算方法：

$$TF - IDF = TF * IDF$$

如果某个单词在一篇文章中出现的频率高，并且在其他文章中很少出现，则认为此词或者短语具有很好的类别区分能力，适合用来分类。

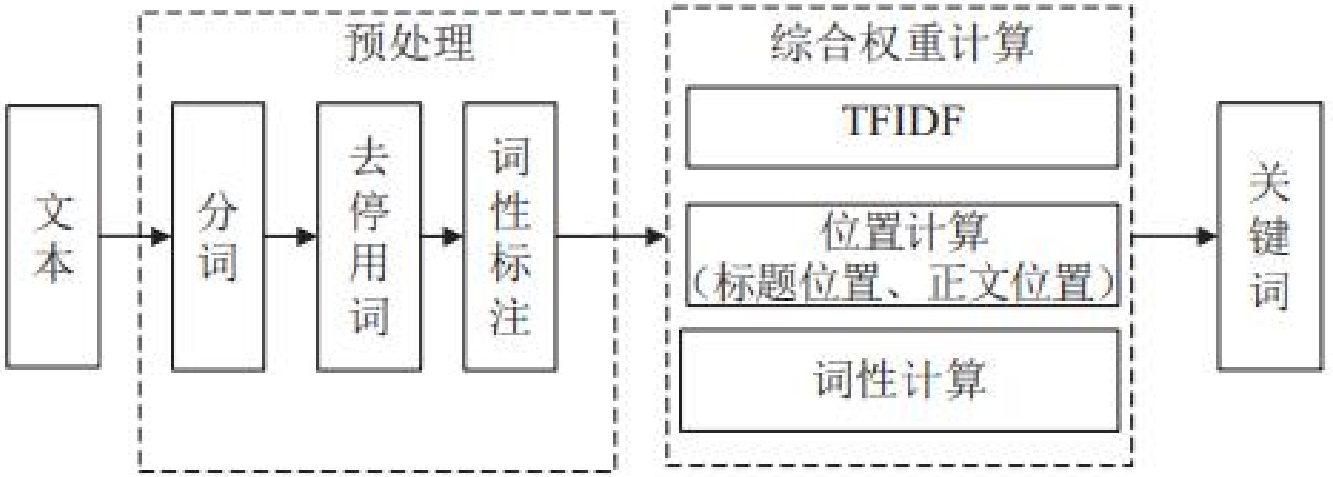
TD-IDF值可作为文章里各词的权重，然后便能对各词做加权求和，得到文章的表示。

TF-IDF在检索中的应用

- TF-IDF算法，主要依赖词频进行权重计算，但存在过度依赖词频的噪声.
- 因此，可以先利用TF-IDF算法完成权重计算后，再对每个词进行位置判断。
- 在考虑标题位置的同时也考虑该词在正文中的位置. 将同时出现在标题和正文中的词，其因子设置为最高；只出现在标题中的次之；而只出现在正文中的因子设为逐次递减的不为零的低值。
- 完成位置特征引入后，再进一步考虑引入词性特征. 对提取出来的关键词进行词性标注. 按名词、动词和其他词汇的顺序从高到低设置为不为零的值。

TF-IDF在检索中的应用

- 关键词提取流程



TF-IDF在检索中的应用

- 权重计算函数：

$$Weight(i) = W_{tf}(i) * W_p(i) * W_c(i)$$

- 其中，$Weight(i)$为候选词 i 的综合权重；$W_{tf}(i)$为 TF-IDF提取词 i 得到的权重；
- $W_p(i)$为位置因子权重，计算方法是：根据提取词在文章中出现位置进行赋值. W_p =3，出现在标题和正文中； W_p =2，仅出现在标题中；W_p =1，仅出现在正文中.
- $W_c(i)$为词性因子权重，计算方法是：根据提取词的词性来进行赋值. 提取词是名词性词汇 W_c =3，动词性词汇 W_c =2，其他词汇 W_c =1.

9.3.2 动态相关性 PageRank

- 早期的搜索引擎大都采用分类检索的方式，即靠人工辨别来对网站进行分类，类似hao123这样的网站。

9.3.2 动态相关性 PageRank

- 随着时代的发展，网页变得越来越多，用人工识别的方式对网页进行识别变得越来越不现实。

- Google的两名创始人Larry Page与Sergey Brin开始对网页排序问题进行研究，依据学术界评判论文的重要程度的方法，查看论文引用次数，将这种方法用到了网页排序中，PageRank算法就产生了。

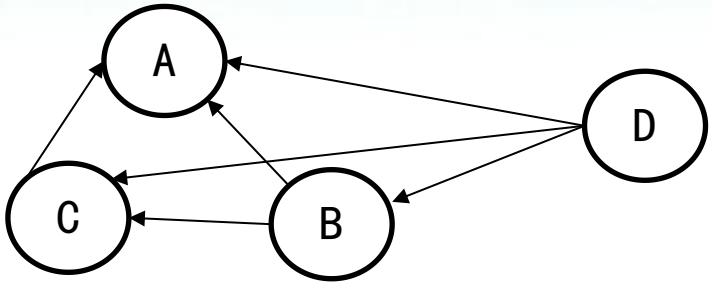
PageRank

- 假设一个由4个网页组成的群体：A，B，C和D。如果所有页面都只链接至A，那么A的PR（PageRank）值将是B，C及D的Pagerank总和。

$$PR(A) = PR(B) + PR(C) + PR(D)$$

- 假设B链接到A和C，C只链接到A，并且D链接到全部其他的3个页面。一个页面总共只有一票。所以B给A和C每个页面半票。以同样的逻辑，D投出的票只有三分之一算到了A的PageRank上。

$$PR(A) = \frac{PR(B)}{2} + \frac{PR(C)}{1} + \frac{PR(D)}{3}$$



PageRank

- 对于一个页面A，那么它的PR值为：

$$PR(A) = (1-d) + d\sum_{i=1}^{n}\frac{PR(T_i)}{C(T_i)}$$

- d 为阻尼系数，其意义是，在任意时刻，用户到达某页面后并继续向后浏览的概率。

PageRank

- 经过迭代计算至收敛，最终页面得到的PR值为该页面的分数。
- 网站被按照他们的PageRank算法分数从大到小排序，那些位于序列顶端的网站被认为是值得信赖的。在这些网站汇总标签为可信赖的网站被选作种子。
- PageRank算法高排名的页面都有很高的入度，说明指向一个网页的重要网页越多这个页面就越重要。

PageRank


Google
多媒体应用
全部
图片
新闻
视频
地图
更多
设置
工具
找到约 51,700,000 条结果 （用时 0.46 秒）
https://baike.baidu.com › item › 多媒体技术应用
多媒体技术应用（信息技术领域用语）_百度百科
多媒体技术应用是当今信息技术领域发展最快、最活跃的技术，是新一代电子技术发展和竞争的焦点。多媒体技术融计算机、声音、文本、图像、动画、视频和通信 ...
主要内容1： 数据压缩，图像处理
主要内容2： 多媒体数据库和基于内容检索 ...
主要内容3： 多媒体著作工具
主要内容 · 应用现状
https://baike.baidu.com › item › 多媒体应用软件
多媒体应用软件_百度百科
多媒体应用软件主要是一些创作工具或多媒体编辑工具，包括字处理软件、绘图软件、图像处理软件、动画制作软件、声音编辑软件以及视频软件。这些软件，概括 ...
简介 · 多媒体制作软件 · 教育培训
https://zh.wikipedia.org › zh-hans › 多媒体
多媒体- 维基百科，自由的百科全书
多媒体（Multimedia），在電腦應用系统中，组合两种或两种以上媒体的一种人机交互式資訊交流

数字媒体检索技术

9.4 媒体检索反馈技术

相关反馈

- 在实际检索系统中，数据库和信息检索模型是相对稳定的，而用户的需求却是相对多变的。

- 用户提交的检索请求往往简短而模糊，即使用户具有明确的查询目的，也不能清晰地构造一个查询。有时查询本身也会有歧义，往往不能准确地描述用户的信息需求。

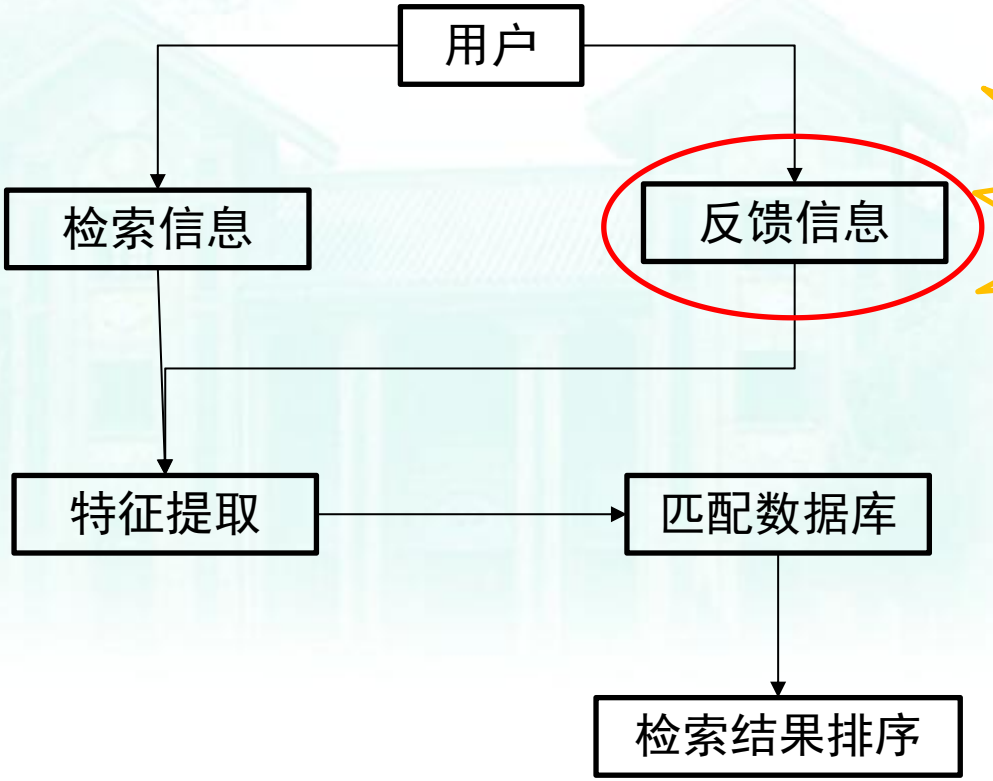
相关反馈

- 相关关反馈技术是提高信息检索性能的有效技术之一。

- 采用相关反馈技术可以修改查询，减少在获取用户查询需求过程中不稳定因素对检索系统造成的负面影响，提高获取用户检索需求的准确性。系统不需要用户构造新的查询条件，而是通过用户的反馈信息来主动实现查询修改，最终返回令用户满意的检索结果。

相关反馈

- 相关反馈基本流程



相关反馈

- 从如何获取反馈信息方式的角度，可以将相关反馈技术分为：
  - 显式相关反馈
  - 隐式相关反馈
  - 伪相关反馈

显式相关反馈

- 显式相关反馈技术主要用于文本信息检索领域，要求用户有一个明确的检索目的并提供查询词项；要求检索系统可以根据查询词项，给出检索结果，同时系统提供一个明确的接口用以接收用户反馈信息。
- 用户按照自己的检索目的，对系统给出的检索结果做出相关与否的标记，这些反馈信息即为来自用户的显式反馈信息，系统可以利用这些信息更新查询结果。

显式相关反馈

- 如百度百科中的点赞功能。

隐式相关反馈

- 在实际检索系统中，往往很难直接获得用户给出的反馈信息。尽管用户做出评价后，可以进一步获得更好的检索结果，但是大多数用户希望简化操作，享受更短的响应时间，因此，带有用户显式反馈功能的检索系统实用性较差，很难得到推广。
- 但是人们在进行信息检索浏览时，往往会给出一些非常隐晦的反馈。比如用户对某个页面感兴趣，则该页面被点击的频率很高，用户停留时间长等。因此，通过收集其他资源可以得到间接的用户反馈信息，即隐式相关反馈。

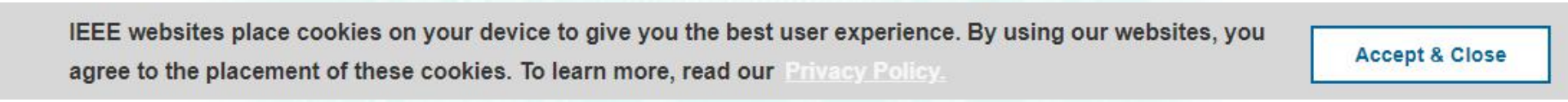
隐式相关反馈

- 隐式相关反馈减轻了用户的负担，虽然用户不直接参与反馈，但是仍然可以改进检索质量。

- 通过分析用户的行为来发现用户的兴趣和爱好，比如通过收集用户浏览记录等信息间接分析用户的偏好，通过文档的全局点击率分析文档的重要性等。结合用户的检索需求进行检索优化。

隐式相关反馈

- 如浏览网页时，弹出如下信息。



伪相关反馈

- 收集分析用户的访问行为有时涉及获取用户隐私信息，并不被用户所接受。另外，在缺少用户信息的时候，无法获得直接或者间接的反馈信息。

- 此时可以采用伪相关反馈，也叫做盲相关反馈技术，实现结果重排序，提高信息检索性能。

伪相关反馈

- 伪相关反馈是相关反馈技术中常用的一种方式。它既不需要用户去对检索的结果进行评价，即不需要用户的交互操作，也不必捕捉用户的点击与浏览行为，而是直接从系统检索结果本身获得反馈信息。

伪相关反馈

- 常用的伪相关反馈通常是将首次检索结果排序靠前的前N项作为相关文档，对前N项结果进行分析扩展用户的初始查询。

- 例如使用TF-IDF的方法从这些文档中选择前20-30个词语；将这些词语加入到查询中，然后再去匹配查询所返回的文档，最终返回最相关的文档。