Web 信息处理与应用 复习笔记

© 2017-1 熊家靖 PB14011026

PART 1： Web Search

一、Introduction

1、web 搜索的挑战：

数据规模大、分布散、不稳定、质量差、无结构、异构、价值低

2、信息检索：

给定查询和信息库，找到相关的文档

3、IR 与 DB 的区别：

DB 数据结构化、有明确语义，查询结构化、匹配要精确、次序不重要

IR 数据半结构化、无明确语义，查询为任意内容、无需精确匹配、次序很重要

4、IR 的任务：

基于用户查询的搜索、信息过滤、分类、问答

5、IR 的基础性问题：

相关性计算、检索模型、评价、信息需求、检索性能

二、Web Crawler

1、网络爬虫的概念：

从一个种子站点集合开始，从 web 中寻找并且下载网页，获取排序需要的相关信息，并且剔除低质量的网页

2、网络爬虫基本过程：

种子装入桶中、每次从桶中取出一个网页、提取出网页所有 url 放入桶中、重复

3、网络爬虫的主要需求：

快、可扩展性、友好性、健壮、持续搜集、时新性

4、网络爬虫的常用策略：

用栈深度优先、用队列广度优先

5、网络爬虫涉及的协议：

HTTP/HTML、DNS/URL、Robots Exclusion（排斥协议）、Sitemp（允许协议）

6、URL 规范化：

协议://主机名[:端口]/路径/[:参数][?查询]#Fragment

7、分布式爬虫的概念：

如何有效地把 N 个网站的搜集任务分配到 M 个机器上去使得分配比较均匀

8、一致性 Hash 的概念：

将网页和机器都映射到环路 Hash 空间，每个机器负责自身位置与后继的网页搜集

三、Text Processing

1、文本处理的概念：

将原始文档转换成词项集以方便索引

2、字符编码的概念：

ASCII：美国信息交换标准代码

Unicode：统一码，满足跨语言、跨平台的需求

UTF-8：针对 Unicode 的可变长度字符编码

3、分词中的概念：

分词：将文档的字符串序列变成词序列

语素：最小的语音语义结合体，是最小的语言单位

词：代表一定的意义，具有固定的语音形式，可以独立运用的最小的语言单位

交叉歧义：网球/场/　网/球场/

组合歧义：我/个人/　三/个/人/

未登录词：未包括在分词词表中但必须切分出来的词，包括各类专名、术语、缩略语等

停用词：在文档中频繁出现或与语料库特性有关的词

4、中文分词的挑战：

汉语是字的集合而不是词的集合

汉字存在着不同的组词方式

汉语虚词众多，大多数汉字在不同的词语中可能为关键字，也可能为停用词

分词歧义

新词的频繁出现

5、常用的分词方法：

机械分词：　正向最大匹配分词 FMM

反向最大匹配分词 BMM / RMM

双向最大匹配分词 BM: FMM + RMM

最少切分分词：图中最短路径

ASM(d, a, m) d 为匹配方向，a 为失败后增/减串长，m 为最大/小匹配

理解分词：　分词时进行句法、语义分析，从而减少歧义

统计分词：　一元文法模型 即最大概率分词

二元文法模型 每个词的概率为前一个词出现后的条件概率

N 元文法模型 每个词的概率为前 N 个词出现后的条件概率

6、词根化和编辑距离的概念：

词根化：使用一系列后缀变换规则对单词进行变换

编辑距离：从 s 转换为 t 使用增加、删除、替换三种操作的最小次数

四、Indexing

1、布尔检索的概念：

利用 AND、OR 或者 NOT 操作符将词项连接起来的查询

2、关联矩阵的概念：

行为词项，列为文档，词项在文档中出现为 1 不出现为 0

3、倒排索引的概念和结构：

以词项为索引，每个词项维护一个链表，表示其出现过的文档集（从小到大）

可以加入扩展项：某词在某文档中的出现词频 TF、某词出现过的文档频数 DF

4、倒排索引的构建：

写出每个文档的 词项 -> 文档 索引

合并所有的索引，词项和文档号均从小到大排列

5、倒排索引的存储：

词项与链表存储在同一个文件中/不同文件中

6、词汇表存储结构：

顺序存储、Hash table、B+-树、Trie 树

7、Zipf' Law：

任意一个词项，其频度和排名的乘积大致是一个常数

五、Queries

1、查询表达的难点：

一个查询可以代表非常不同的信息需求

一个查询可能是其真正需求的一种非常差的表述

2、查询表达的优化：

局部优化：对用户查询进行局部分析，如相关性反馈

全局优化：进行全局分析来产生同/近义词词典，如查询扩展

3、相关性反馈的概念和过程：

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

4、相关性反馈的分类及其各自的概念和特点：

显式反馈：　定义：用户显式参加交互过程，即用户反馈

问题：开销大、查询长、用户不愿意、反馈逻辑难理解

隐式反馈：　定义：系统跟踪用户的行为来推测返回文档的相关性，从而反馈

好处：省却了用户的显式参与过程

问题：对分析的要求高、准确度难保证、可能需要额外设备

伪相关反馈：定义：对于真实相关反馈的人工部分进行自动化

好处：不用考虑用户因素，处理简单，平均效果也不错

问题：准确率难以保证，可能出现查询漂移

5、Ricchio 算法：

新查询向量 = α·原查询向量 + β·平均相关向量 – γ·平均不相关向量

计算过程中出现负值，全部设为 0

基本假设：用户知道使用文档集中的词项来表达初始查询；相关文档出现的词项类似

6、查询扩展的概念：

相关性反馈中，用户针对文档提供附加信息，查询扩展中，用户对词项提供附加信息

7、查询扩展的几种方法：

人工构建同/近义词词典、自动导出同/近义词词典、基于查询日志挖掘查询等价类

六、Ranking

1、Ranking 的难点：

Web 网页的质量参差不齐，大量的网页组织性、结构性比较差

大部分检索用户是没有任何经验的

用户的查询需求存在着巨大差异

2、信息检索模型的概念：

用来描述文档和用户查询的标识形式以及它们之间相关性的框架

形式化表示为：[D, Q, F, R(Di,q)]即[文档表达，查询表达，匹配框架，相关性度量函数]

3、信息检索的实质问题：

对于所有文档，根据其与用户查询的相关程度从大到小排序

4、信息检索模型与搜索引擎排序算法的关系：

好的信息检索模型在相关性上产生和人类决策非常相关的结果

基于好的检索模型的排序算法能够在排序结果顶部返回相关的文档

5、信息检索的分类：

基于集合论的模型：布尔模型

基于代数论的模型：向量空间模型

基于概率论的模型：概率模型、语言模型、推理网络

6、相关系数的概念和计算：

Jaccard：A 与 B 的交中元素的个数 / A 与 B 的并中元素的个数

未考虑词频、文档长度、罕见词信息量

tf(t, d)：词项 t 在文档 d 中出现的次数

相关度不会正比于词项频率

w(t, d): 当 tf > 0 时，1 + lg(tf)； 否则，0

df(t): 出现词项 t 的文档数目

idf(t): lg(N / df) 其中 N 是文档集中文档的数目

tf-idf: (1 + lg tf)·lg(N / df)

随着词项频率的增大而增大

随着词项罕见度的增大而增大

7、向量空间模型 SMART：

D： 每个文档是一个以词项为维度的向量，每个维度的值为词项的 tf-idf 值

Q： 每个查询是一个以词项为维度的向量，每个维度的值为词项的 tf-idf 值

F： 非完全匹配

R： 用文档向量和查询向量的相似度来估计相关性

前提假设：检索到的所有文档相关性不等价、相关性多元、查询关键字互相独立

8、余弦相似度：

两个向量夹角的余弦值，即：两向量的点乘 / 各自模的积

9、向量空间模型的操作过程：

文档和查询表示成 tf-idf 的权重向量

计算两向量余弦相似度

将余弦相似度 Top-K 的文档返回给用户

10、向量空间模型的缺点：

用户无法描述词项之间的关系

tf-idf 高的词项可能会在检索中影响过大

词项之间的独立性假设与实际不符

11、概率模型：

定义随机变量 R、Q、D，相关度 R = 0 或 1

通过计算条件概率 P(R = 1 | Q = q, D = d)来度量文档和查询的相关度

12、PageRank：

PR(a) = (1 – d) + d·sigma(PR(T) / C(T))

每个页面的 pagerank 等于进入它的边的 pagerank 的函数

计算过程：每个网页赋初值，然后迭代计算，直到变化小于一个阈值

优点：给网页提供重要性排序 + 可以离线完成 + 独立于主题

缺点：未区分链接种类 + 对新网页不公平 + 不能单独用于排序

13、HITS：

入步骤：所有权威页面的值等于链向它的中心页面的值之和

出步骤：所有中心页面的值等于其链向的权威页面的值之和

计算过程：所有页面初始为 1，迭代使用入步骤和出步骤

优点：能更好描述互联网的组织特点 + 主题相关 + 查询无关 + 可以单独用于排序

缺点：需要在线计算时间代价大 + 容易受到“链接作弊”的影响

七、Evaluation

1、信息检索评价概述：

评价受主观、情景、认知、时间的影响，重点在于保持公平

2、信息检索评价指标：

效率：时间开销、空间开销、响应速度

效果：准确率、召回率、是否靠前

其他：覆盖率、访问量、数据更新速度

3、效果评价指标：

基于集合：	正确率 P：	返回的相关文档占返回的总文档的比比例
	召回率 R：	返回的相关文档占相关总文档的比例
	F 值：	召回率 R 和正确率 P 的调和平均
	Fβ 值：	召回率 R 和正确率 P 的加权调和平均
		其中 R 的权为β^2，P 的权为 1
基于序：	P@N：	值考虑返回的前 N 个文档时的正确率
	R-Precision：	即 P@相关文档总数
	未插值 AP：	P@相关文档出现位置的平均
	插值 AP：	在召回率 0,0.1,0.2……1.0 上十一点的正确率平均
		不存在某召回率点时，取该点到下一个点之间最大正确率
	简化 AP：	在未插值 AP 中忽略未出现的相关文档
多个查询：	MAP：	所有查询的 AP 的算术平均
	MRR：	第一个相关文档返回的位置的倒数的算术平均
其他：	CGp：	位置 1 到位置 p 的检索结果的相关度之和
	DCGp：	相关度要先除以 log2(i)作为惩罚，其中 i 为出现的位置
	NDCGp：	DCG 的值除以理想化的 IDCG 的值，规范化为[0,1]

PART 2： Web Information Extraction

一、Named Entity Extraction

1、信息抽取的概念：

从语料中抽取指定的事件、事实等信息，形成结构化的数据

能作为一种浅层的文本理解，是信息检索的进一步深化

2、信息抽取与信息检索的关系：

检索是从文档集合中找文档子集，抽取是从文本中获取用户感兴趣的事实信息

检索通常利用统计与关键词等技术，抽取借助于自然语言处理技术

检索通常与领域无关，抽取通常与领域相关

3、MUC-7 定义的信息抽取任务：

命名实体 NE：现实世界中具体或抽象的实体，还包括日期、时间、数量等

模板元素 TE：实体属性，通过槽描述命名实体的基本信息

共指关系 CR：命名实体的等价关系

模板关系 TR：实体之间的各种关系，又称为事实

背景模板 ST：实体发生的事件

4、信息抽取的内容：

实体、属性、关系、事件

关键在于“抽取实体，确定关系”

5、命名实体识别 NER 的概念：

识别文本中的人名、地名等专有名称和有意义的时间、日期等数量短语并加以归类

6、命名实体识别 NER 的难点：

命名实体类型多样、新命名实体不断出现、命名实体歧义、命名实体结构复杂

7、MUC-7 中定义的 NER 内容：

实体类：人名、地名、机构名

时间类：日期、时间

数值类：货币、百分比

注意：人造物、重复指代的普通名词、派生词、以人命名的法律和奖项等不算！

8、命名实体识别 NER 的性能评价指标：

正确率 P：正确数 / 总数　　(正确数 + (1/2)部分正确数) / 总数

召回率 R：正确数 / 总正确数　(正确数 + (1/2)部分正确数) / 总(部分)正确数

F 值：　　P 与 R 的调和平均

9、命名实体识别 NER 的常用方法：

基于词典：词典匹配；难以枚举命名实体、构建词典代价大、难以处理歧义

基于规则：自行构造模板匹配；依赖性强、代价大、建设周期长、可移植性差

基于统计：隐马尔可夫 HMM、最大熵 ME、支持向量机 SVM、条件随机场 CRF

混合方法：混合使用词典、规则和统计

二、Relation Extraction

1、关系抽取的概念：

从文本中识别出两个实体或多个实体之间存在的事实上的关系

2、关系抽取的意义：

提高搜索引擎发现知识的能力

广泛应用于各种知识库的构建

支持知识推理和问答系统研究

3、关系的表示方法：

二元组、三元组、多元组

4、关系抽取的常用方法：

基于规则：针对特定领域的特定关系，设计针对性的抽取规则，代价大，难移植

基于模式：种子关系生成关系模式，基于关系模式抽取新的关系，再迭代生成新的模式和新的关系

基于机器学习：特征向量、核函数

5、DIPRE 系统：

给定种子元组 R

在文档中搜索元组 R 的出现 O

从出现 O 中提取模板 P

使用模板 P 从文档中获取新的元组

6、Snowball 系统：

只使用能匹配很多模板的元组

只使用有多个元组支持的模板

PART 3：Web Data Mining

一、Introduction

1、网络挖掘的概念：

从 web 中挖掘有用的信息和有用的模式

2、网络挖掘的内容与应用：

网络内容挖掘：数据挖掘、数据分类、数据聚类

网络结构挖掘：社区分析、影响力分析

网络用途挖掘：推荐系统

二、Data

1、数据对象、属性、维度、特征：

数据对象是一个数据实例，其属性、维度、特征意思相同，均为描述数据的一个域

2、高维诅咒现象：

数据分类的表现不会随着维数的增加而一直上升，反而到了某个阈值后会下降

因为随着维数上升，每个类的数据变得稀疏，很多测量手段都逐渐失去意义

3、数据预处理的基本方法：

采样：使用有代表性的样本，使得样本与总体在属性上有相似的性质

特征选择：剔除冗余和无关特征

降维：避免高维诅咒、降低数据挖掘的代价、使数据更加清楚、消除噪声

三、Classification

1、监督学习和无监督学习：

监督学习：使用训练样本训练模型，再利用模型解析未知数据进行分类

无监督学习：无训练样本，直接按照未知数据的相似程度建模聚类

2、分类的基本原理：

选定模型后，使用训练数据训练模型参数，之后用模型解析输入数据得到分类

3、数据的向量表示：

用数据的频数或者 tf-idf 表示

4、KNN 算法：

找到与待分类数据距离最近的 K 个数据，然后将其分入频数最高的类中

KNN 无法免疫高维诅咒现象，但是在高维特征独立数较小时，KNN 也适用

5、Logistic regression 算法：

6、如何评价分类效果：

训练误差：训练数据的过程中造成的错误

测试误差：测试的过程中造成的误差 accuracy 为测准率

泛化误差：使用模型在未知记录上造成的分布相同的期望误差

四、Clustering

1、聚类的概念：

聚类是一个把现实或抽象的对象和与它相似的对象组织到一起的过程

2、聚类的基本原理：

聚类内部相似性很高，聚类之间相似性很低

3、层次式聚类算法流程：

计算距离矩阵，默认所有数据点都是一个类

每次找到距离最近的两个类，将其合并，并更新距离矩阵，重复直到只有一个类

4、类的距离定义：

Single-link：使用两个聚类之间最近的点作为聚类的距离

Complete-link：使用两个聚类之间最远的点作为聚类的距离

Average-link：使用所有跨聚类的结点对的平均距离

Centroid：使用聚类重心之间的距离

5、K-means 算法流程：

随机产生 k 个聚类中心点

每个数据点归类到与它最近的那个中心所代表的类

每个类重新计算中心点，返回第二步

算法迭代到所有数据点的类归属不再改变

6、K-means 算法优化目标：

每个数据点到它所属的类中心距离的平方和最小

7、K-means 收敛性分析：

均方差函数单调递减而有界

8、聚类算法的评价标准：

凝聚度：计算各聚类的均方差的和

分离度：不同聚类的重心要尽可能相互远离

专家评判

五、社区分析：

1、图的表示、组成部分以及相关性质：

点、边（有向、无向）

2、社区的概念：

一组结点集，集合内的点之间有很多联系，而集合内的点与集合外的点联系很少

3、社区发现与聚类：

基于结构相似性通过使用层次式聚类或分割式聚类

4、结构相似度计算：

结构差异测度 dij：取两点关联向量的差，向量中两点所在的位置清零，取模

Jaccard 相似度：两点公共邻居数 / 两点无重总邻居数

余弦相似度：两点关联向量的余弦

5、GN 算法：

一对结点之间的最短路径为路上的边贡献一个流

若最短路径有多条，则均分

每次切除一条流量最大的边，然后重新计算流量，迭代进行，直到无边

6、矩阵及性质：

邻接矩阵：相邻为 1，不相邻为 0

度数矩阵：对角线放每个结点的度数，其余地方为 0

拉普拉斯矩阵：度数矩阵减去邻接矩阵，是半正定的

7、 Cut 的性质：

Cut(A, B)表示 A 与 B 之间的边数

Cut(A, B) $=\frac{1}{4}y^T(D-W)y=\frac{1}{4}y^TLy$；当 u∈A，y(u) = 1，当 u∈B，y(u) = -1

$\mathrm{RatioCut(A,B)}=\mathrm{cut(A,B)}(\frac{1}{|A|}+\frac{1}{|B|})$

$\mathrm{NCut(A,B)}=\mathrm{cut(A,B)}(\frac{1}{vol(A)}+\frac{1}{vol(B)})$　　vol(A)表示 A 中结点度数之和

$\mathrm{RatioCut(A,B)}=min_h\frac{h^T(D-W)h}{h^Th}$　st. $h^Te=0$

$\mathrm{NCut(A,B)}=min_{g'}\frac{g'^TD^{-0.5}(D-W)D^{-0.5}g'}{g'^Tg'}$　st. $g'^TD^{0.5}e=0$

8、modularity 的概念：

一种测量网络划分为社区的好坏程度的指标

两结点间的实际边数为A_{ij}，期望边数为$\frac{d_id_j}{2m}$，每个社区内的边数差为$A_{ij}-\frac{d_id_j}{2m}$

每个社区内边数差相加后除以总度数 2m，即为 Q(G, S)属于[-1, 1]

六、影响力分析：

1、度量结点中心性的标准：

Degree centrality：结点的度，可以除以 n-1 标准化

Closeness centrality：结点到其他结点的平均测地距离的倒数

Betweenness centrality：该结点通过的流量，可除以(n-1)(n-2)/2 标准化

Eigenvector centrality：Ax = λx，其中 x 是所有结点的 Eigenvector centrality

2、关系强度：

删除后会造成结点对不连通的边叫桥

删除后造成的结点对的距离增量越大，该关系越不牢固

邻居 overlap 函数：两结点公共邻居数 / (两结点无重总邻居数 – 2)s

3、影响力传播模型：

线性阈值模型 LTM：关联到某结点的激发边的总激发值大于阈值，则该结点被激发

层级传播模型 ICM：激发结点按照边权概率激发周围的结点

区别：LTM 是基于接收者的，ICM 是基于发送者的

LTM 依赖于所有邻居结点，ICM 影响到所有邻居结点

LTM 状态只依赖于阈值，ICM 的状态存在随机性

但是他们都具有子模性质！

4、最大影响结点集：

f(S)是结点集 S 最终能够影响的结点集的大小

最优化问题：max f(S) ，其中 S 大小为 k

贪心算法：每次选取一个对影响集的大小增量最大的结点

近似度：f(S) ≥ (1 – 1/e)*OPT = 0.63*OPT，严格成立，数据无关

5、子模性质：

A 是 B 的子集，对于函数 f( )

如果：f(A+e)-f(A)>=f(B+e)-f(B)成立，则说 f()函数是子模的，即增益递减

七、Recommendation

1、推荐系统基本模型及一般工作流程：

用户兴趣建模 + 推荐算法 + 效果评估 + 大数据库

2、基于内容的推荐算法流程：

分别对用户和项目建立配置文件

通过分析已购买过的内容，建立或更新用户的配置文件

比较用户与项配置文件的相似度，并直接向用户推荐与其配置文件最相似的项目

3、基于内容的推荐算法分析：

优点：简单、冷启动、不受打分稀疏性问题约束、可以解释为什么这么推荐

缺点：多媒体数据难提取、用户潜在偏好难发现、新闻系统等不适用

4、协同过滤（基于用户）推荐算法流程：

利用历史评分信息计算用户之间的相似性

根据相似度得到邻居用户集，利用邻居用户在目标项上的评分信息来预测目标用户

根据计算所得的喜好程度对目标用户进行推荐

5、 协同过滤（基于用户）推荐算法分析：

优点：

缺点：受打分稀疏性问题约束

附：2016 年秋考试记录

【考试时间】

2017 年 1 月 9 日 下午 2:30 — 4:30

【命题老师】

金培权老师、徐林莉老师联合命题

【考试题型】

第一大题：判断题（ 2’ × 10 = 20’ ）

1、金老师和徐老师每人 5 道

2、知识点非常散，但是只要知道都很容易判断

3、ppt 上的每一句话都要看仔细了

第二大题：综合题（ 80’ ）

1、简述 K-means 算法流程

分析 K-means 算法是否一定收敛

分析运行多次 K-means 算法是否会收敛到同样的结果

2、根据题目表格计算 F 值、MAP 值等（作业原题）

3、根据文档信息词根化并去除停用词，然后建立倒排索引（类似作业题）

4、给出 NCut 算法对某 6 结点的图进行社区发现的全过程（要用矩阵计算）

5、证明逻辑回归的分类面为线性超平面