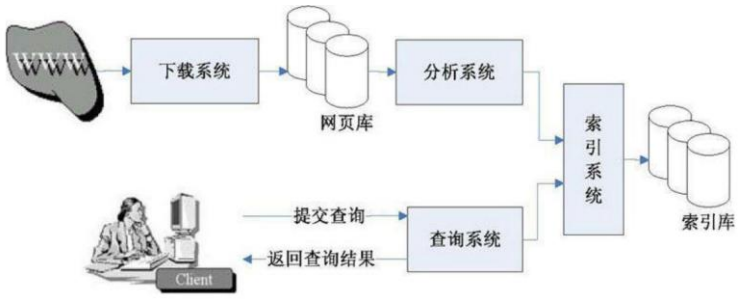
第二章 万维网网页信息的表达及解析

提交日期：4 月 21 日
提交形式：作业以 word 或 pdf 形式，提交到助教邮箱，命名方式为“姓名”+“学号”
提交邮箱：邮箱地址 cwt@bupt.edu.cn ，邮件主题需包含 "信息表达"这四个字作为关键（以方便自动归档），提交成功后会有自动回复。

2.3 搜索引擎与分析系统

1、【简答题】
请结合下图简述全文搜索引擎中下载系统、分析系统、索引系统和查询系统的功能。



2、【综合题】
（1）N-gram 模型是自然语言处理中一种基于_____的语言模型，其基本思想是：将文本里面的内容按照字进行大小为 N 的滑动窗口操作，形成长度为___的片段序列，每一个片段称为_____，在所给语句中对所有的 gram 出现的_____进行统计。再根据_____中每个 gram 出现的频数进行比对可以得到所给语句中每个 gram 出现的_____。整句的概率就是各个词出现概率的_____。
（2）请列举两种 N-gram 的应用。

（3）假设我们采用 2-gram 语言模型，且有下图的统计结果，请计算 s1=” <s> i want to eat food </s>”和 s2=” <s> i eat chinese food </s>”哪个句子更合理，并给出计算公式。(提示：可以假设 p(i|<s>)=x, p(</s>|food)=y，x, y~(0, 1))

各词出现的次数

i	want	to	eat	chinese	food	lunch	spend
2533	927	2417	746	158	1093	341	278

n=2词序列频度

I-I频次：5　I-want频次：827　I-food频次：0

	i	want	to	eat	chinese	food	lunch	spend
i	5	827	0	9	0	0	0	2
want	2	0	608	1	6	6	5	1
to	2	0	4	686	2	0	6	211
eat	0	0	2	0	16	2	42	0
chinese	1	0	0	0	0	82	1	0
food	15	0	15	0	1	4	0	0
lunch	2	0	0	0	0	1	0	0
spend	1	0	1	0	0			

问题：数据稀疏

2.4 搜索引擎索引系统

1、【简答题】
请解释索引系统中的索引［S. Brin 1998］包含的 3 个概念?

命中（Hit）:

正向索引（Forward Index）:

倒排索引（Inverted Index）

2、【简答题】
请简要比较正向索引和倒排索引的优缺点?

3、【应用题】
请根据给出的 3 个文档补全下面的带有单词频率、文档频率和出现位置信息的文档倒排索引。

文档编号	文档内容
1	The cat sit on the mat
2	The cat catches mice
3	Here is a little cat

单词 ID	单词	文档频率	倒排列表（DocID;TF;<POS>）
1	The		
2	cat		
3	sit		
4	on		
5	mat		
6	catches		
7	mice		
8	Here		
9	is		
10	a		
11	little		

2.5 搜索引擎查询系统

1、【填空题】

_____进行的一次查询，是相对于搜索引擎查询系统而言的，结果是搜索结果网页；

_____对索引库进行的一次检索，是相对于搜索引擎索引系统而言的，结果是与查询词相关的文档列表；普通用户提交给查询系的关键词称为“_____”；经过查询系统分词，提交检索代理的称为”_____”。

2、【简答题】

分析“布尔模型”的优缺点?

3、【综合题】

（1）名词解释

词频:

停用词:

逆文档频率:

TF-IDF 值:

（2）TF-IDF 计算，根据以下 5 段内容，请根据词表计算词频、逆文档频率（逆文档频率可以通过**“文档数”/“包含该词的文档数“**简单计算）和 TF-IDF 值。

内容 1： 水果有水果，水果，水果，水果，水果

内容 2： 水果有苹果，桃子，西瓜，菠萝，梨子

内容 3： 蔬菜都很好吃，我最爱吃茄子了

内容 4： 苹果，梨子都是很好吃的水果

内容 5：好吃的水果有西瓜，苹果，葡萄，其他水果还有菠萝，猕猴桃

词表	水果	有	苹果	桃子	西瓜	菠萝	梨子	蔬菜
词频-内容 1								
词频-内容 2								
词频-内容 3								
词频-内容 4								
词频-内容 5								
文档频率								
逆文档频率								
TF-IDF 简易计算								
内容 1								
内容 2								
内容 3								
内容 4								
内容 5								