

## 北京大学研究生课程期末考题，2004 年秋

课程名称：网络信息体系结构；时间：2005 年 1 月 6 日

### 一、名词或概念解释：(21 分)

(1). 信息检索模型 (Information Retrieval Models)

$\langle D, Q, F, R(q_i, d_j) \rangle, q_i \in Q, d_j \in D$

- $D$ : 来源于  $D_0$ , 是  $D_0$  的一种逻辑表示
- $Q$ : 来源于  $Q_0$ , 是  $Q_0$  的一种逻辑表示
- $F$ : 一个“框架”, 给出  $D_0$  和  $Q_0$  的表示方法, 以及它们之间的关系
- $R$ : 基于表示 ( $D$  和  $Q$ ) 确定一个查询和一篇文档相关性的函数

(2). 信息抽取 (Information Extraction)

信息提取是通过分析非结构化文本, 提取预先定义好的实体、关系或事件, 把非结构化的文本转化为结构化的信息库

(3). 特征选取 (Feature Selection)

文本分类中, Remove terms in the training documents which are statistically uncorrelated with the class labels.

(4). 文本聚类 (Text Clustering)

□ Cluster: a collection of data objects

- Similar to one another within the same cluster
- Dissimilar to the objects in other clusters

(5). Topical locality

- web graph was not constructed randomly. Web network is topically clustered.
  - Radius-1 hypothesis:  $u$  positive and  $u$  link to  $v$ , then  $v$  is positive with higher probability than random chosen Extract outlinks of relevant pages  $\diamond$  classifier
  - Radius-2 hypothesis:  $u$  points to many pages  $v$  with large  $R(v)$ , then  $u$  is good HUB, having high relevant children.

(6). Small world network

Diameter of graph is small ( $\log N$ ) as compared to overall size

- Empirical study of Web-graph reveals small-world property
  - Average distance ( $d$ ) in simulated web:

$$d = 0.35 + 2.06 \log (n)$$

#### (7). Text test collection

測試集 (Test Collection) 配合測量準則 (Measures) 來評估系統的模式。所謂測試集，是在規範化環境中測試系統效能的機制，包括測試文件集 (Document Set)、測試問題 (Queries)、及相關判斷 (Relevance Assessment) 等三個部分。其研究設計的概念是假設在給定的查詢問句與文件集中，某些文件是與查詢問句相關的。系統的目的是檢索出相關的文件，並拒絕不相關的文件，因此採用召回率 (Recall) 及精確率 (Precision) 作為測量準則。

**注：** 上述是答案要点，来自课程 ppt。表述方式没有固定限制。

## 二、问答和计算题 (60 分)

1. 若你被要求设计一个通用搜索引擎，目标是覆盖中国所有的静态网页（预计规模在 3 亿），其搜集系统将面临哪些挑战，相应的有哪些技术手段可以应用？

答：应覆盖下列方面

- Scalable
  - Parallel , distributed
- Fast
  - Bottleneck? Network utilization
- Polite
  - DoS, robot.txt
- Robust
  - Traps, errors, crash recovery
- Continuous
  - Batch or incremental

2. 给定一个有向图  $G = (V,E)$ ,  $V=\{1,2,3,4\}$ ,  $E=\{<1,2>, <1,4>, <2,3>, <2,4>, <3,1>, <3,4>, <4,2>\}$ , 试利用 Power Iteration 算法, 给出节点声望值计算的前三次迭代结果 (设初值均为 0.25)。

迭代算法:

- 给定邻接矩阵  $E$ ,
- 初始化向量  $p_0$ , 使得  $\text{Sigma}(p_0)=1$
- 对于  $k = 1, 2, \dots$ , 执行如下步骤
  - $x = E^T p_{k-1}$ ,      基本迭代
  - $p_k = x/\|x\|$ ,      规格化步骤

$$E^T = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \end{bmatrix}$$

$$p(0) = [0.25 \quad 0.25 \quad 0.25 \quad 0.25]$$

$$p(1) = [0.25 \quad 0.50 \quad 0.25 \quad 0.75]$$

$$p(2) = [0.25 \quad 1.00 \quad 0.50 \quad 1.00]$$

$$p(3) = [0.50 \quad 1.25 \quad 1.00 \quad 1.75]$$

注: Normalization doesn't change the result distribution of the iteration algorithm; it's just a guardian against the overflow in calculation. So it's omitted here.

注: 结果因不同的规格化方法的选择而不同, 但数值比例应该合上述答案一致。

3. 给定一个查询输出文档相关性序列表示如下: 0 1 1 1 0 0 1 0 1 0 1 1, 其中“1”表示对应的文档相关,“0”表示不相关。假设系统文档集合中共有 10 个相关文档。试给出针对该查询的“11 点标准召回率的精度值”(可以任意选用一个插值方法, 但要指明)。

插值方法:

a)  $P(r_j) = \max P(r), r_j \leq r \leq r_{j+1}$

- i. 取在下一个标准召回率之间的已知召回率对应的最大精度值

b)  $P(r_j) = \max P(r), r_j \leq r$

- i. 取往后的已知召回率对应的最大的精度值 (这得到的是阶梯函数, 单调性)。

011100101011

1.原始 P-R 数据

|           |     |      |      |      |      |      |      |     |     |     |
|-----------|-----|------|------|------|------|------|------|-----|-----|-----|
| Recall    | 0.1 | 0.2  | 0.3  | 0.4  | 0.5  | 0.6  | 0.7  | 0.8 | 0.9 | 1.0 |
| Precision | 1/2 | 2/3  | 3/4  | 4/7  | 5/9  | 6/11 | 7/12 |     |     |     |
|           | 0.5 | 0.67 | 0.75 | 0.57 | 0.56 | 0.55 | 0.58 |     |     |     |

2.未召回的 recall 点 precision 置 0

3.插值方法 a)

$P(r_j) = \max P(r), r_j \leq r \leq r_{j+1}$  取在下一个标准召回率之间的已知召回率对应的最大精度值

|           |      |      |      |      |      |      |      |      |     |     |     |
|-----------|------|------|------|------|------|------|------|------|-----|-----|-----|
| Recall    | 0.0  | 0.1  | 0.2  | 0.3  | 0.4  | 0.5  | 0.6  | 0.7  | 0.8 | 0.9 | 1.0 |
| Precision | 2/3  | 2/3  | 3/4  | 3/4  | 4/7  | 5/9  | 7/12 | 7/12 | 0   | 0   | 0   |
|           | 0.67 | 0.67 | 0.75 | 0.75 | 0.57 | 0.56 | 0.58 | 0.58 | 0   | 0   | 0   |

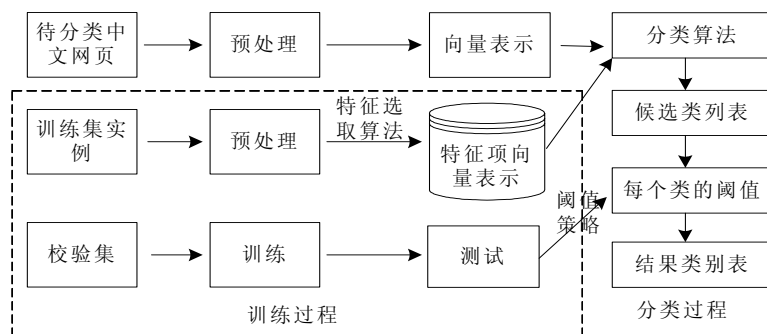
插值方法 b)

$P(r_j) = \max P(r), r_j \leq r$  取往后的已知召回率对应的最大的精度值

|           |      |      |      |      |      |      |      |      |     |     |     |
|-----------|------|------|------|------|------|------|------|------|-----|-----|-----|
| Recall    | 0.0  | 0.1  | 0.2  | 0.3  | 0.4  | 0.5  | 0.6  | 0.7  | 0.8 | 0.9 | 1.0 |
| Precision | 3/4  | 3/4  | 3/4  | 3/4  | 7/12 | 7/12 | 7/12 | 7/12 | 0   | 0   | 0   |
|           | 0.75 | 0.75 | 0.75 | 0.75 | 0.58 | 0.58 | 0.58 | 0.58 | 0   | 0   | 0   |

4. 阐述文本自动分类的一般过程。在文本分类中有哪些阈值选取策略？

分两步：(1)训练过程 (2).分类过程；可参考下图描述的过程加以叙述。其中，常见的分类器的构造方法有：DTree, Rocchio, NB, kNN, SVM, LLSF, Nnet。



常见的阈值选取策略有：位置截尾法(RCut)：以文档为中心的方法。

比例截尾法(PCut)：以类别为中心的方法。最优截尾法 (SCut)：上述两种方法的综合。

5. 如何用支持向量机 (SVM) 构造文本分类器？

(1)。 处理两类问题，多类问题可以归结为两类问题

(2)。 对两类问题，构造支持向量机 (SVM)。

寻找能分开这两类的最优超平面， $g(x) = w^T x + w_0$ ，

A two-category classifier with a discriminant function of the form (1) uses the following rule:

Decide  $w_1$  if  $g(x) > 0$  and  $w_2$  if  $g(x) < 0$

Ū Decide  $w_1$  if  $w^T x > -w_0$  and  $w_2$  otherwise

If  $g(x) = 0$   $x$  is assigned to either class

(3) 用课堂上讲过的画图描述并解释更好。

6. 如何利用潜在语义标引 (LSI/SVD) 将文档和词项表示在一个二维或三维的空间中？对一个新的查询，如何在该空间中找到与之相关的文档？

构建词频矩阵  $A$ ，(  $m$  个词，  $n$  个文档)

对  $m \times n$  矩阵  $A$ ，做 SVD 分解：

$$A = U \Sigma V^T$$

对上面的公式解释一下，中间的矩阵  $\Sigma$  为对角矩阵，奇异值从大到小排列。

$k=2$  或  $3$  时，利用下式 (1)，可将  $m$  维的文档投影到  $2$  或  $3$  维空间中，类似地用下式 (2) 对  $n$  维的词项也做投影。

$$\hat{d} = d^T U_k \Sigma_k^{-1}, \quad \hat{t} = t V_k \Sigma_k^{-1}.$$

先将查询串用单个词表示，然后用上式 (1) 投影到该空间中。最后在该空间内，用某种距离度量如向量夹角的余弦公式，计算文档间的相似度。

三、综合题 (19 分)：按照下述描述和要求完成相关工作

给定文档语料：

D1: 北京安立文高新技术公司

D2: 新一代的网络访问技术

D3: 北京卫星网络有限公司

D4: 是最先进的总线技术。。。

D5: 北京升平卫星技术有限公司的新技术有。。。

利用中文切分词软件，分别得到用 “/” 分开的一些字词：

D1: 北京/ 安/ 立/ 文/ 高新/ 技术/ 公司/

D2: 新/ 一/ 代/ 的/ 网络/ 访问/ 技术/

D3: 北京/ 卫星/ 网络/ 有限/ 公司/

D4: 是/ 最/ 先进/ 的/ 总线/ 技术/ 。。。

D5: 北京/ 升/ 平/ 卫星/ 技术/ 有限/ 公司/ 的/ 新/ 技术/ 有。。。

你的任务是设计一个针对这些文档的信息检索系统。具体要求是：

(1). 给出系统的有效词汇集合（说明取舍原因）。

去除停用词。选取停用词或取文档集中  $df$  极高的，作为与文档集相关的停用词，或使用通用停用词表，比如包括汉语的常用虚词等。

例如取“的”作为停用词，得到有效词汇表：

北京 高新 技术 公司 网络 访问 有限 总线 卫星 .....

(2). 写出  $D1$  和  $D2$  在  $VSM$  中的表示（使用  $tf*idf$ ，写出各项的数字表达式，具体数值不必实际计算出来）。

向量表示（空间维数依赖（1））；选定一种  $idf$  的取值方法，如  $\log(N/n_i)$ 。

$D1$  表示为： $(\log(5/3), \log(5), \log(5/4), \log(5/3), 0, 0, 0, 0, 0\dots)$

$D2$  表示为： $(0, 0, \log(5/4), 0, \log(5/2), \log(5), 0, 0, 0\dots)$

(3). 画出系统的倒排文件示意图。

标注词的位置较好，但不做要求，如下只列文档亦可。

|     |             |
|-----|-------------|
| 北京  | D1 D3 D5    |
| 高新  | D1          |
| 技术  | D1 D2 D4 D5 |
| 公司  | D1 D3 D5    |
| 网络  | D2 D3       |
| 访问  | D2          |
| 有限  | D3 D5       |
| 总线  | D4          |
| 卫星  | D5          |
| ... | ...         |

(4). 按照向量夹角的余弦计算公式，给出针对查询“技术的公司”的前 3 个反馈结果。

首先可以作定性分析：

这几个文档长度相当，文档的  $|Wd|$  相对来说差别不大。采用向量夹角的余弦计算公式计算查询与文档间的相似度时，查询的  $tf*idf$  值是固定的，故此主要考察文档的  $tf*idf$  值。包含“技术”又包含“公司”的有两个文档： $D1$ ， $D5$ ，由于  $D5$  中“技术”的词频为 2 高于  $D1$  中的词频，故  $D5$  的相似度大于  $D1$ 。在剩余的三个文档中，“技术”的  $idf$  值低于“公司”的  $idf$  值，故此选用  $D3$ 。可以定性的推算出排序结果为： $D5$ ， $D1$ ， $D3$ 。

实际计算的方法:

为简化计算,不妨取 idf 为  $N/df$ , 词汇表取上述 9 个词 (单字略去)。计算过程如下:

$$\text{Cos}(Q, D_d) \propto \frac{1}{|W_d|} \sum_{t \in Q} f_{d,t} \times f_{q,t} \times \left(\frac{N}{df_t}\right)^2$$

$$|W_d| = \sqrt{\sum_{t \in D} \left(f_{d,t} \times \frac{N}{df_t}\right)^2}$$

$N=5$  为常数, 上面计算公式中不影响结果序, 可略去。

根据 (2) 的倒排文件, 可以列出下表:

| 分子 | 技术                               | 公司                           | Sum (*144) |
|----|----------------------------------|------------------------------|------------|
| D1 | $\left(\frac{1}{4}\right)^2$     | $\left(\frac{1}{3}\right)^2$ | 9+16=25    |
| D2 | $\left(\frac{1}{4}\right)^2$     | 0                            | 9          |
| D3 | 0                                | $\left(\frac{1}{3}\right)^2$ | 16         |
| D4 | $\left(\frac{1}{4}\right)^2$     | 0                            | 9          |
| D5 | $2 * \left(\frac{1}{4}\right)^2$ | $\left(\frac{1}{3}\right)^2$ | 9*2+16=34  |

| Wd | Inside radical sign   | Result (*√144) |
|----|---|----------------|
| D1 | $\left(\frac{1}{3}\right)^2 + 1 + \left(\frac{1}{4}\right)^2 + \left(\frac{1}{3}\right)^2$                          | $\sqrt{185}$   |
| D2 | $\left(\frac{1}{4}\right)^2 + \left(\frac{1}{2}\right)^2 + 1$   | $\sqrt{189}$   |
| D3 | $\left(\frac{1}{3}\right)^2 + \left(\frac{1}{3}\right)^2 + \left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2$ | $\sqrt{104}$   |

|    |   |              |
|----|---|--------------|
| D4 | $(\frac{1}{4})^2 + 1$   | $\sqrt{153}$ |
| D5 | $(\frac{1}{3})^2 + 2 * (\frac{1}{4})^2 + (\frac{1}{3})^2 + (\frac{1}{2})^2 + 1$ | $\sqrt{230}$ |

易知， $D5 > D1 > D3 > \dots$