信息检索导论--中文文本分类

授课人：李永可

邮箱：lyk@xjau.edu.cn

文本分类问题的提出

- 假想图书馆的图书资料不加以分类, 结果如何?
- 随着互联网技术的飞速发展，各种电子文本数据的数量急剧增加
- 信息数据量的爆炸性增长使得传统的手工处理方法变得不切合实际

目 录

文本分类的基本概念

文本分类(Text Categorization或Text Classification，TC)

- 是根据给定文本的内容，将其判别为事先确定的若干个文本类别中的某一类或某几类的过程。
- 这里所指的文本可以是媒体新闻、科技、报告、电子邮件、技术专利、网页、书籍或其中的一部分。
- 由于类别是事先定义好的，因此分类是有指导的（或者说是有监督的）

文本分类的基本概念

- 分类体系一般人工构造
  - 政治、体育、军事。。。
  - 中美关系、恐怖事件。。。
- 分类系统可以是层次结构，如yahoo!
- 分类模式
  - 2类问题，属于或不属于(binary)
  - 多类问题，多个类别(multi-class)，可拆分成2类问题
  - 一个文本可以属于多类(multi-label)
- 这里讲的分类主要基于内容
- 很多分类体系：Reuters分类体系、中图分类

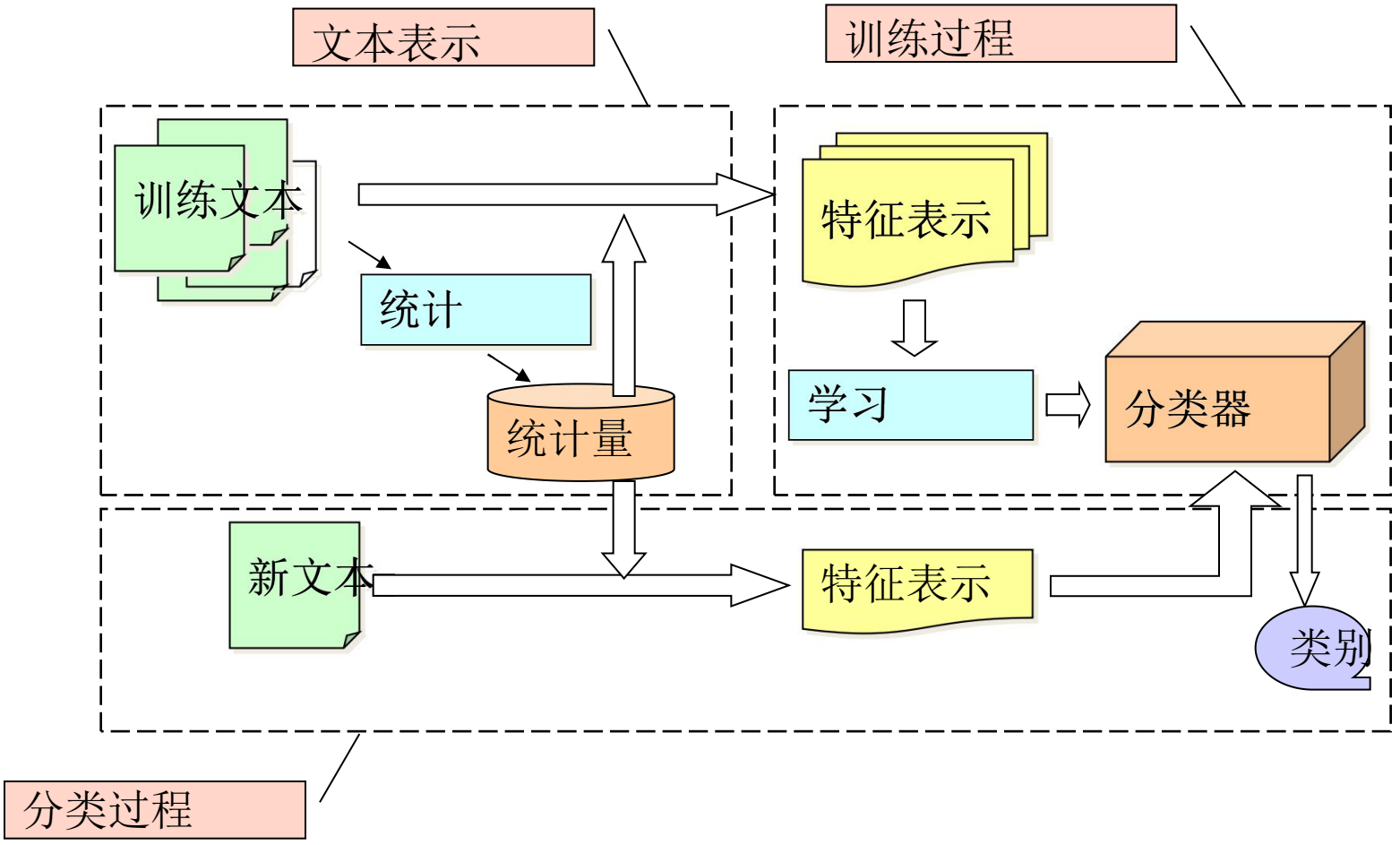
文本分类的基本概念

应用领域

- 冗余过滤　在数字图书馆和搜索引擎的建设中
- 组织管理　图书馆利用图书分类法来管理所收藏的图书资料
- 智能检索　搜索引擎的构建过程中
- 信息过滤　“人找信息” 成为“信息找人”
- 其它应用　元数据提取、构建索引、文本过滤

文本分类的基本概念

文本分类的一般过程



目 录

文本表示-中文分词

中文的预处理要比英文的预处理要复杂的多，因为汉语的基元是字而不是词，句子中的词语间没有固定的分隔符（如空格），因此必需对中文文本进行词条切分处理。

- 基于词典和规则的方法，应用词典匹配、汉语词法、约束矩阵等知识进行分词
- 基于统计的方法:将汉语基于字与词的统计信息，如相邻字间互信息、词频及相应贡献信息等应用于分词
- 混和方法

文本表示-向量空间模型

- 向量空间模型(Vector Space Model, 简称VSM)
- 文档（Document）：

  — 泛指一般的文献或文献中的片断（段落、句子组或句子），一般指一篇文章。

- 项（Term）：

  当文档的内容被简单地看成是它含有的基本语言单位（字、词、词组或短语等）所组成的集合时，这些基本的语言单位统称为项，即文档可以用项集（Term List）表示为 $D(T_1, T_2, ..., T_n)$

  其中 T_k 是项, $1 \le k \le n$

文本表示-向量空间模型

- 项的权重（Term Weight）：

— 对于含有n个项的文档 $D(T_1,T_2,...,T_n)$，项常常被赋予一定的权重，表示它们在文档中的重要程度，即

D = $D(T_1,W_1;T_2,W_2;...;T_n,W_n)$

为了简化分析，可以暂不考虑 T_k 在文档中的先后顺序并要求 T_k 无异（即没有重复）

- 这时可以把 $T_1,T_2,...,T_n$ 看成一个n维的坐标系，而 $W_1,W_2,...,W_n$ 为相应的坐标值，因而 $D(W_1,W_2,...,W_n)$ 被看成是n维空间中的一个向量

文本表示-向量空间模型

- 相似度（Similarity）：

当文档被表示为VSM，常用向量之间的内积来计算：

$$Sim(D_1, D_2) = \sum_{k=1}^{n} W_{1k} * W_{2k},$$

或用夹角余弦值来表示：

$$Sim(D_1, D_2) = \cos\theta = \frac{\sum_{k=1}^{n} W_{1k} * W_{2k}}{\sqrt{(\sum_{k=1}^{n} W_{1k}^2)(\sum_{k=1}^{n} W_{2k}^2)}},$$

目 录

特征选择

- 目的:
  - 为了提高程序的效率，提高运行速度
  - 为了提高分类精度
    - 一些通用的、各个类别都普遍存在的词汇对分类的贡献小
    - 在某特定类中出现比重大而在其他类中出现比重小的词汇对文本分类的贡献大
    - 对于每一类，我们应去除那些表现力不强的词汇，筛选出针对该类的特征项集合

特征选择

常用方法

- 文档频率DF
- 信息增益IG
- 互信息MI
- χ^2统计量（CHI-2）

特征选择

常用方法-文档频率DF

- Document frequency，文档频率，简称DF
- 指在训练语料中出现某词条的文档数
- Term的DF小于某个阈值去掉(太少，没有代表性)
- Term的DF大于某个阈值也去掉(太多，没有区分度)

特征选择

常用方法-信息增益IG

- 对于特征词条t和文档类别c，IG考察c中出现和不出现t的文档频数来衡量t对于c的信息增益，定义如下：

$$IG(t) = -\sum_{i=1}^{m} P(c_i)\lg P(c_i) + P(t)\sum_{i=1}^{m} P(c_i \mid t)\lg P(c_i \mid t) + P(\overline{t})\sum_{i=1}^{m} P(c_i \mid \overline{t})\lg P(c_i \mid \overline{t})$$

特征选择

常用方法-信息增益IG

- 信息增益的优点在于，它考虑了词条未发生的情况，即虽然某个单词不出现也可能对判断文本类别有贡献。

- 但在类分布和特征值分布是高度不平衡的情况下其效果就会大大降低了。

特征选择

常用方法-互信息MI

- 互信息(Mutual Information)在统计语言模型中被广泛使用。
- 它是通过计算特征词条t和类别c之间的相关性来完成提取的。其定义如下:

$$MI(t,c) = \lg \frac{P(t \wedge c)}{P(t) \times P(c)}$$

特征选择

常用方法-互信息MI

- 如果用A表示包含特征词条t且属于类别c的文档频数，B为包含t但是不属于c的文档频数，C表示属于c但不包含t的文档频数，N表示语料中文档的总数，t和c的互信息可由下式计算：

$$MI(t,c) \approx \lg \frac{A \times N}{(A+C) \times (A+B)}$$

特征选择

常用方法- χ^2 统计量（CHI-2）

- 它度量特征词条t和文档类别c之间的相关程度，并假设t和c之间符合具有一阶自由度的分布。
- 特征词条对于某类的统计值越高，它与该类之间的相关性越大，携带的类别信息也越多。
- 反之，统计量也是反映属性t和类别c之间的独立程度。当值为0时，属性t与类别c完全独立。

特征选择

常用方法-χ^2 统计量（CHI-2）

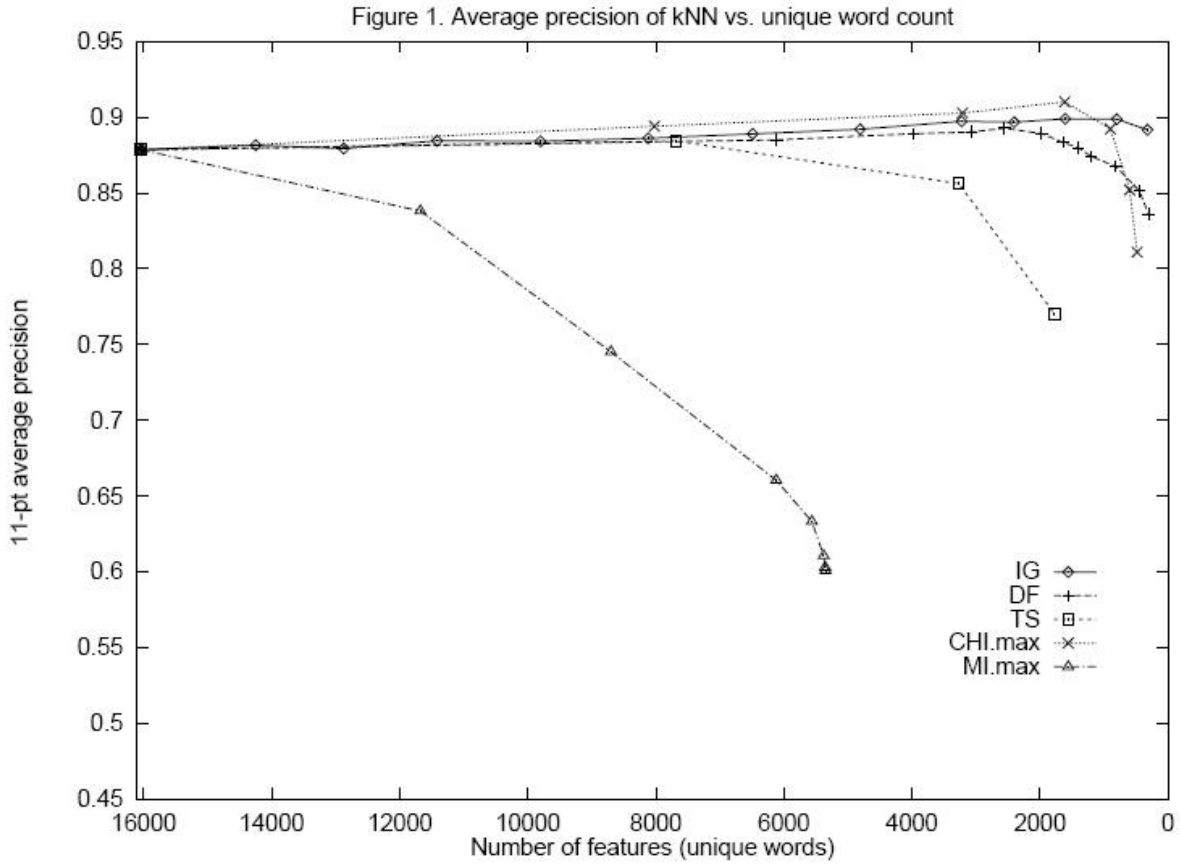
- 令N表示训练语料中的文档总数，c为某一特定类别，t表示特定的词条
- A表示属于c类且包含t的文档频数，B表示不属于c但是包含t的文档频数
- C表示属于c类但是不包含t的文档频数，D是既不属于c也不包含t的文档频数.其定义为：

$$\chi^2(t,c)=\frac{N(AD-CB)^2}{(A+C)(B+D)(A+B)(C+D)}$$

	c	~c
t	A	B
~t	C	D

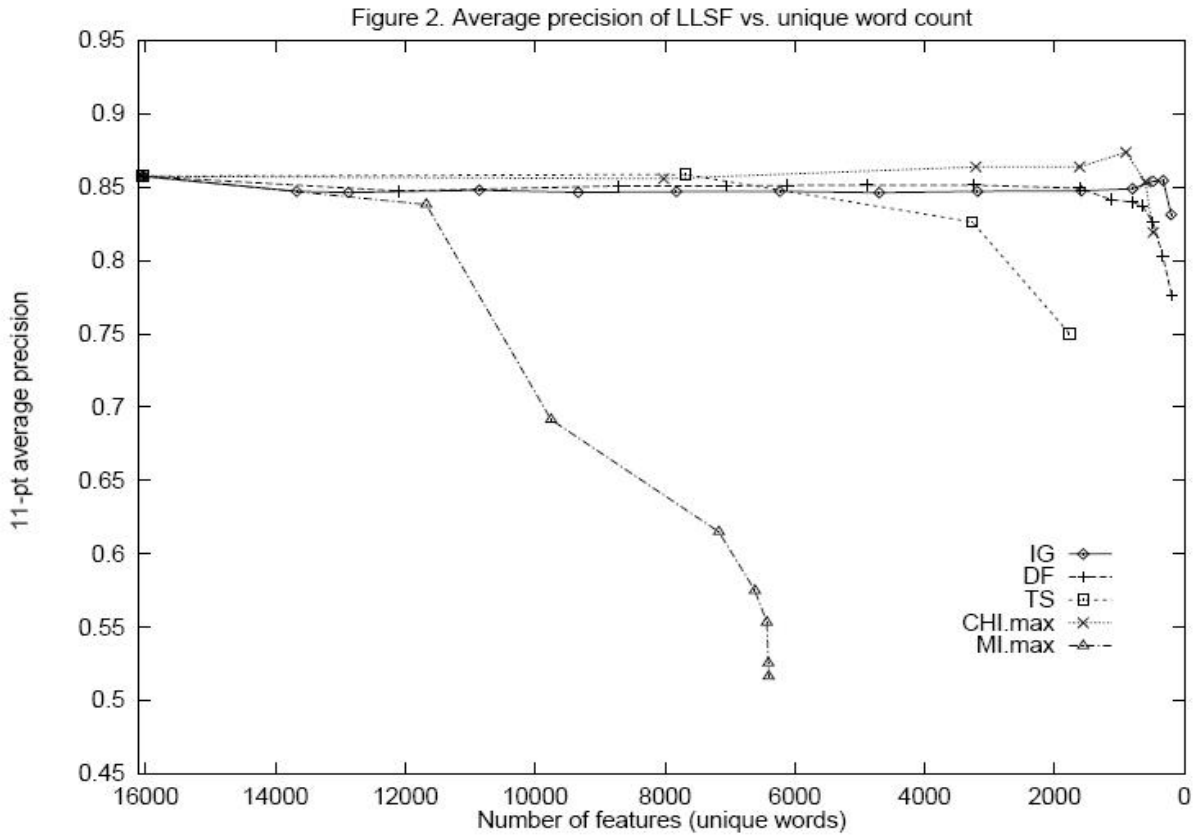
特征选择

特征选择方法性能比较



特征选择

特征选择方法性能比较



注：以上实验结果来自于Yang, Y., Pedersen J.P. A Comparative Study on Feature Selection in Text Categorization Proceedings of the Fourteenth International Conference on Machine Learning (ICML'97), 1997, pp412-420.