

从文本表示看NLP与IR

孙乐

中国科学院软件研究所

CIPS2008

大纲

- **NLP与IR**
- 文本表示
- **IR模型**
- 文本表示下的融合方案
- 小结

NLP与IR

- **Query**（查询条件） & **Collection**（文档集） **NLP**
- 传统的基于关键词的检索系统无法处理由于同义词，多义词带来的性能损失，也不能判断用户对所输入的几个关键词的关注程度，用户也很难通过关键词来设定准确的需求
- 将各个层面的**NLP**技术引入**IR**在九十年代初期曾引起人们的普遍关注，特别是对词义排歧**WSD**技术在**IR**中的应用进行了集中研究。然而，大部分的研究结果令人沮丧，不少实验结果甚至表明采用**WSD**技术会带来检索性能的下降。
- 目前人们普遍的看法是**NLP**不会对**IR**的性能带来明显的益处

NLP与IR

但是给出这样的结论缺乏足够的依据：

- ✓ 大部分自诩为采用**NLP**技术的**IR**系统实际上只是采用了较低层次的**NLP**技术，比如词形变化
- ✓ 现有的大部分系统只是对查询条件采用了**NLP**技术，而不是对查询条件和整个文档集
- ✓ 由于较高层**NLP**的技术实现的复杂性，大多数**IR**领域的研究人员并不了解如何正确引入较高层次的**NLP**技术
- ✓ 缺少真正的能够给出每层**NLP**技术对**IR**性能贡献的实验结果
- ✓ 传统文本检索系统本质上只是将文本看作一组无序的词串，利用简单的词频统计来模糊计算相关性，不利于引入相对复杂和精确的**NLP**技术

NLP与IR

但是给出这样的结论缺乏足够的依据：

- ✓ 大部分自诩为采用**NLP**技术的**IR**系统实际上只是采用了较低层次的**NLP**技术，比如词形变化
- ✓ 现有的大部分系统只是对查询条件采用了**NLP**技术，而不是对查询条件和整个文档集
- ✓ 由于较高层**NLP**的技术实现的复杂性，大多数**IR**领域的研究人员并不了解如何正确引入较高层次的**NLP**技术
- ✓ 缺少真正的能够给出每层**NLP**技术对**IR**性能贡献的实验结果
- ✓ 传统文本检索系统本质上只是将文本看作一组无序的词串，利用简单的词频统计来模糊计算相关性，不利于引入相对复杂和精确的**NLP**技术

文本表示

文本表示模型

A类 表层信息 (局部)

- *词*
- *N-gram*
- *POS, 短语 (句法)*

B类 深层信息 (全局)

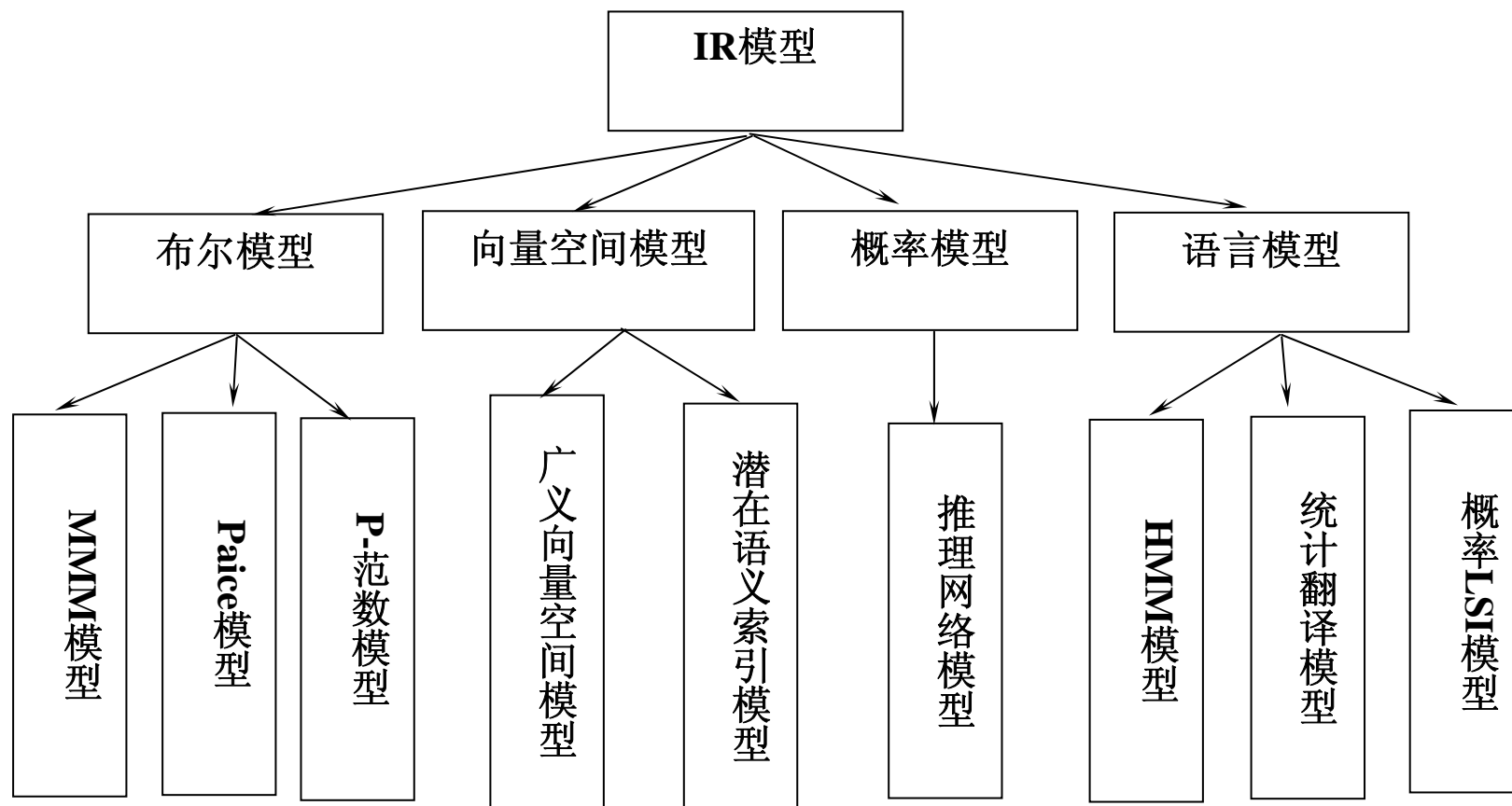
- *DWC (Distributional Words Clustering)*
- *LSI (Latent Semantic Indexing)*
- *LDA (Latent Dirichlet Allocation)*

文本表示

几种文本表示模型比较

- *BOW* (Bag of Words, 词袋)
优点: 简单方便, 缺点: 信息量低
- *VSM* (向量空间模型)
优点: 支持各种权重策略, 缺点: 不支持概念建模
- *LSI* (隐含语义索引)
优点: 支持概念建模
缺点: 不支持大规模语料建模(模型参数和文档数成正比)
- *LDA* (Latent Dirichlet Allocation)
优点: 支持概念建模, 支持大规模语料建模
缺点: 模型较复杂

IR模型

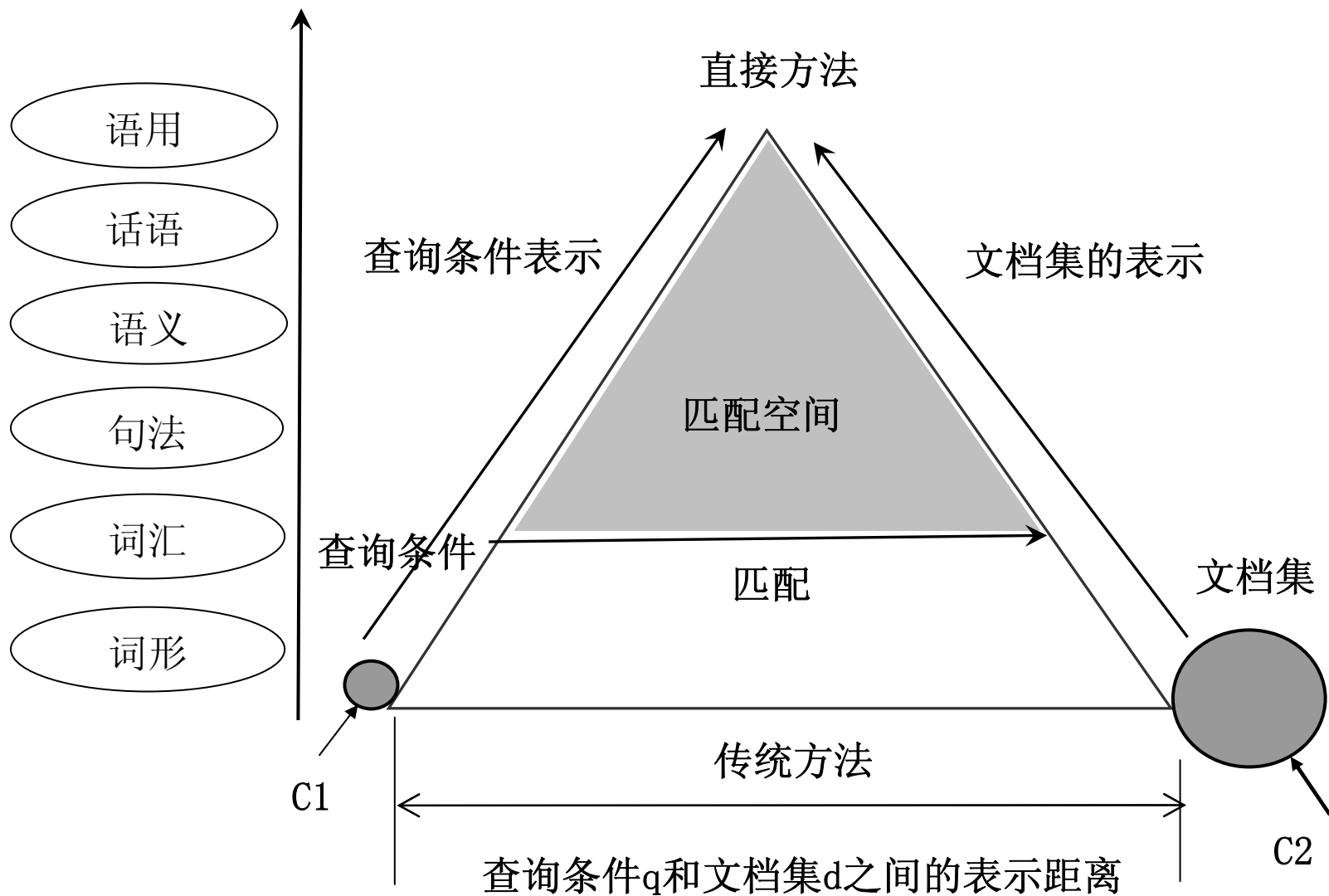


IR模型

四种模型之间的比较

模型	布尔模型	向量空间模型	概率模型	语言模型
提出时间	20世纪50年代	20世纪60年代	20世纪80年代	20世纪90年代末
理论基础	集合论	代数理论	概率论	概率论/随机过程
相关文档判断	二元无序	非二元有序	非二元有序	非二元有序
系统实现难度	简单	简单	较难	简单
部分匹配支持	不支持	支持	支持	支持
文本表示方法	词	词向量	词	N-gram
学术代表系统	无	SMART	INQUERY	LEMUR
商业运用情况	采用	常采用	采用	未采用

NLP技术与IR相融合的体系结构



小结

- 文本表示是**NLP**与**IR**相融合的关键技术之一，通过在文本表示中引入**NLP**技术可以减小查询与文档集之间的匹配空间
- 目前文本表示方法中使用的要么是局部信息，要么是全局信息，这两者的结合有可能使一些深层次的**NLP**技术引入**IR**
- 随着计算机硬件性能的不断提高，新的**IR**模型将会出现，从文本表示来看，现有的**SVM IR**，**LM IR**模型将有可能被以**LDA Model**为代表的图模型所取代

请各位老师批评指正！
