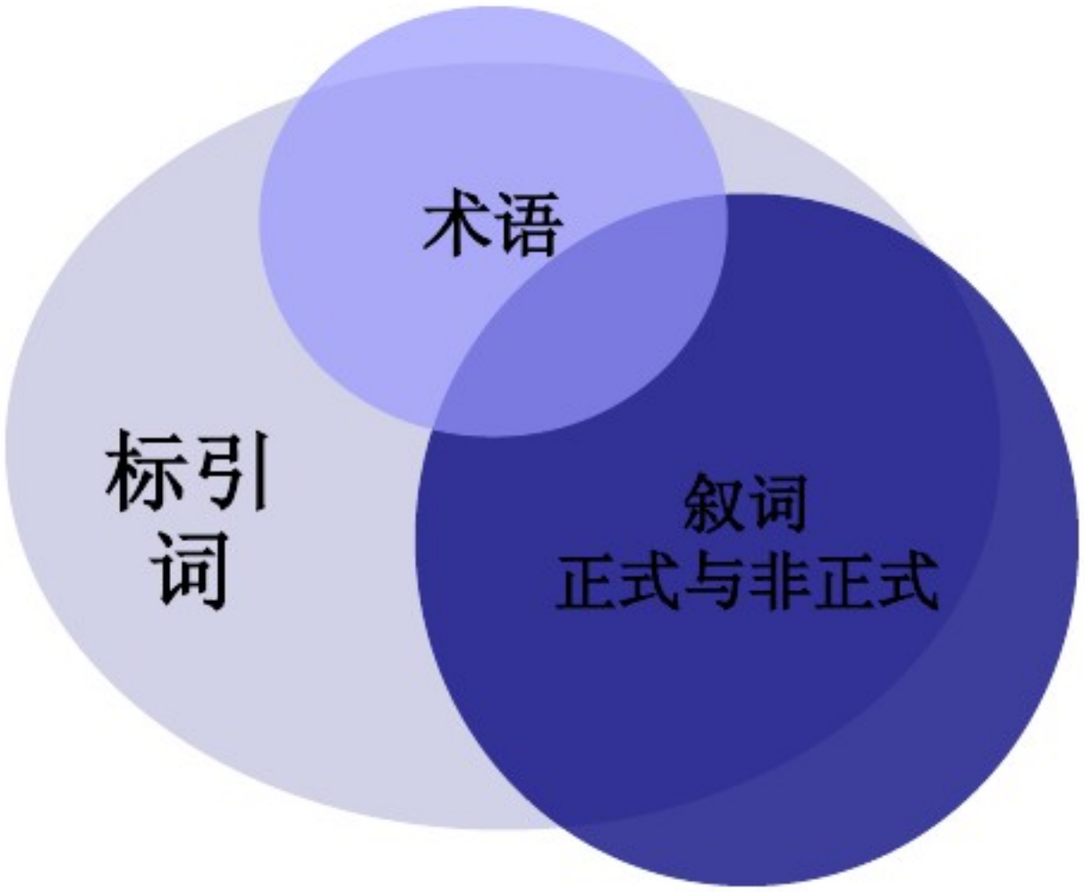
信息自动标引技术

2012年3月27日

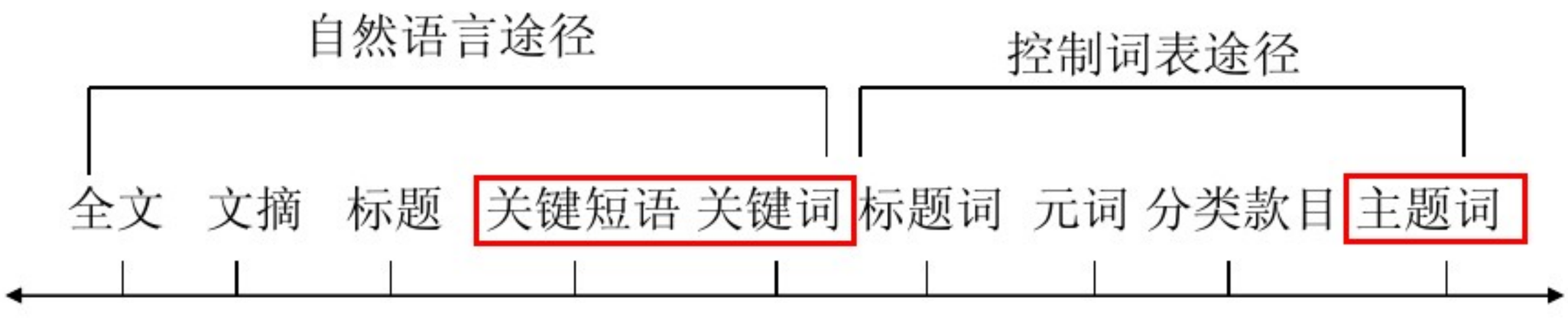
自动标引技术概述

- 自动标引包括关键词自动提取（又称自动抽词标引）与自动赋词标引两种类型。
  - 文本挖掘领域：关键词抽取（Keyword Extraction）
  - 在计算语言学领域：术语自动识别（Automatic Term Recognition）
  - 在信息检索领域：自动标引（Automatic Indexing）
- 自动标引属于文本信息抽取的范畴——文本信息抽取是从文本数据中抽取人们关注的特定的信息。
- 关键词自动提取是一种识别有意义且具有代表性片段或词汇的自动化技术。

- 术语、叙词（主题词）、标引词包含关系图。



- 信息描述颗粒度



自动抽词标引和自动赋词标引

- 自动抽词标引：指直接从原文中抽取词或短语作为标引词来描述文献主题内容的过程。

找到主题词

- 自动赋词标引：指使用预先编制的词表中的词来代替文本中的词汇进行标引的过程 。

转换主题词

自动标引的五十年研究历程

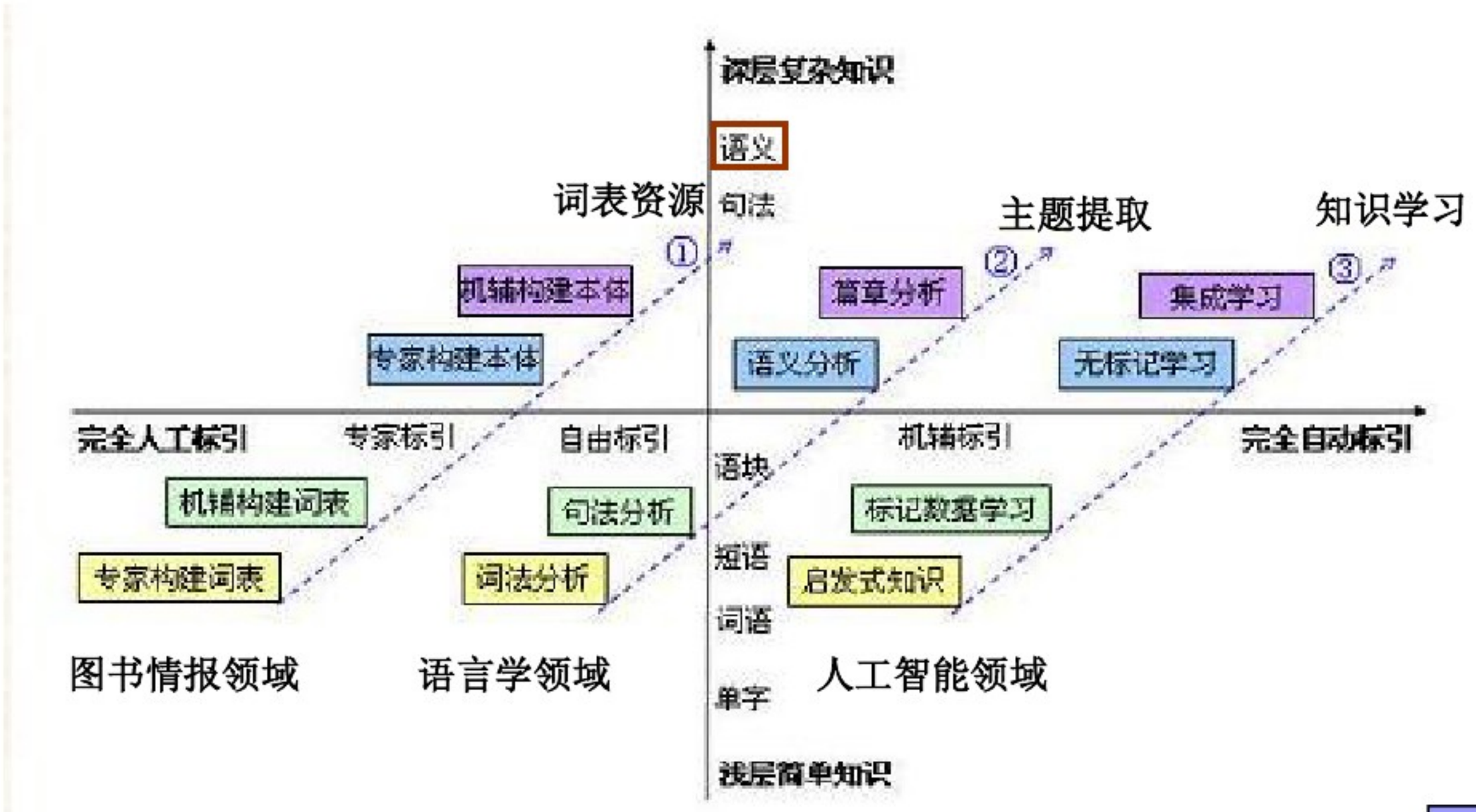
- 1957年开始进行自动标引后开始，到目前为止，自动标引研究经历了50年的发展历程。
- 20世纪90年代初到90年代末，自动标引研究渐渐冷却：
  - **全文索引**逐渐被人采用，并且基本上能满足用户需要；
  - 传统的自动标引方法的效率达到极限；
  - 网络兴起之初的冲击与信息需求环境的改变。
- 随着信息量的增加，全文索引的功能越来越难以满足实际需求，用户需要更加精确的结果。
- 互联网信息服务：例如自动摘要，文档分类与聚类，文本分析，主题检索等都要依赖于关键词自动提取的结果，只有这样才能有希望从根本上提高信息服务质量。

研究历程

- 1957年，Luhn开始自动标引研究，首次将计算机技术引入文献标引领域，开创了以词频为特征的统计标引方法，其理论基础是Zipf定律，该方法具有一定的客观性和合理性，并且简单易行，在自动标引中占有重要地位。
- 基于绝对频率加权法到基于相对频率加权法到贝叶斯、遗传算法、决策树算法等机器学习方法到基于本体的自动标引方法到基于语言模型的关键词提取方法到基于集成学习的关键词抽取。

- 1957年，Luhn开始自动标引研究，首次将计算机技术引入文献标引领域，开创了以词频为特征的统计标引方法，其理论基础是Zipf定律，该方法具有一定的客观性和合理性，并且简单易行，在自动标引中占有重要地位。
- 1958年，Luhn提出基于绝对频率加权法的自动标引方法P.B.Baxendale提出从论题句和介词短语中自动提取关键词
- 1959年，Edmundson与Oswald提出基于相对频率加权法的自动标引方法
- 1960年，Maron & Kuhns提出基于相关概率的赋词标引方法
- 1969年，H.P.Edmundson提出了一些新的加权方法，如提示词(预示词)加权法、题名加权法、位置加权法，并探讨了不同加权法的最优组合问题
- 1970年，Lois L. Earl利用句法分析等语言学方法与词频统计方法相结合的方法来提取关键词
- 1973年，Salton等提出基于词区分值的自动标引方法
- 1975年，Salton等将VSM模型用于自动标引中
- 1983年，Dillon等提出一种基于概念的自动标引方法，研制了FASIT系统；
- 1985年，Devadason提出基于深层结构标引方法；
- 1990年，Deerwester & Dumais等提出潜在语义分析标引法；
- 1993年，Silva & Milidiu提出基于相信函数模型的赋词标引方法；
- 1995年，Cohen提出N-Gram分析法的自动标引方法；
- 1997年，简立峰提出基于PAT树的关键词提取方法；
- 1999年，Frank等人提出基于朴素贝叶斯(Naive Bayes，NB)的关键词提取方法；
- 1999年， Turney 利用遗传算法和C4.5决策树算法等机器学习方法进行关键短语提取的研究；
- 2001年，Anjewierden & Kabel提出基于本体的自动标引方法；
- 2003年，Tomokiyo & Hurst提出了基于语言模型的关键词提取方法；
- 2003年， Hulth利用Bagging算法进行了基于集成学习的关键词抽取；
- 2004年，李素建提出基于最大熵模型的关键词提取方法；
- 2006年，张阔提出基于SVM自动标引模型；
- 2007年，Ercan, G. & Cicekli, I提出基于词汇链的自动标引方法。

研究路线图


深层复杂知识
语义
句法
词表资源
主题提取
知识学习
①
②
③
机辅构建本体
专家构建本体
篇章分析
语义分析
集成学习
无标记学习
完全人工标引
专家标引
自由标引
机辅标引
完全自动标引
机辅构建词表
专家构建词表
句法分析
词法分析
语块
短语
词语
单字
标记数据学习
启发式知识
浅层简单知识
图书情报领域
语言学领域
人工智能领域

三个领域

- **图书情报领域，**主要从资源构建角度进行研究，为主题标引提供了丰富的词表资源；
- **语言学领域，**从语言分析的角度研究了主题提取的机制与方法，利用词法知识、句法知识、语义知识以及篇章知识进行不同层次的主题提取研究；
- **人工智能领域，**主要从机器学习角度对自动标引进行了大量的研究，如利用启发式知识、标记数据的机器学习、无标记的机器学习、集成学习等方法的运用。

两个维度

- 自动化程度维度：先后经历人工标引、机器辅助标引、自动标引等阶段；
- 知识复杂程度维度：先后经历字、词、短语、语块、句法、语义、篇章结构等不同颗粒度的多种知识。

自动标引技术的发展

- 绝大多数标引系统不是完全自动的，标引技术仍然处于实验阶段。
- 自动标引的研究主流方法为*统计学习模型*与*语言知识*（如词类、句法、语义、篇章结构等）的结合。
- 语义检索——本体自动构建。

本体的提出——概念空间

- 概念空间指用于描述领域中概念及其关系的概念模型，通过概念的组织形式来表达领域中的概念以及概念与概念之间的关系。长期以来，领域知识的表达依赖于特定的任务，这样不利于大规模的模型共享、系统集成、知识获取和知识重用，因此需要与任务独立的知识库来表达领域概念空间，从而提出本体的概念。

本体 Ontology 概念

- 哲学概念：被哲学家用于描述事物的本质。
- 本体：人工智能领域将给出构成相关领域词汇的基本术语和关系，以及利用这些术语和关系构成的规定这些词汇外延的规则的集合。
- 本体是构成**语义网**知识结构的基础，通过本体对语义网中的概念和关系以及在此基础上的规则进行定义，从而进行语义上的推理和判断。

本体的特点

- 本体是共享概念化的形式化的明晰的规范。是一整套对某一领域的知识进行表述的词和术语。根据知识结构进行类目的划分组织。
  - 概念化：概念的模型；
  - 明晰的：概念及其应用的明确限定；
  - 形式化的：机器可读，数学表达；
  - 共享：交互知识，组织划分。

本体自动构建研究

- 本体自动构建研究主要集中在自动抽词技术上，机器辅助编制词表的研究可以直接用于赋词标引。随着本体学习的研究的不断深入，本体有望自动或半自动地被构建，并且可用于自动赋词标引当中。基于本体的自动赋词方法是在概念层面上对文本进行标引，并能识别概念之间的关系，标引结果可以用于语义检索当中。

语义分析

- 自动标引主要依据候选对象的若干特征进行分析，将主题表达能力强的候选对象作为标引结果。
- 深层语言知识（词法分析向句法分析过渡）
- 随着深层语义分析和篇章分析研究的不断深入，这些研究成果可用于自动标引任务，提高标引质量。

多标引方法集成

- 标引是一项富有智能性的工作。我们可以借助认知理论对标引任务进行分析和理解。
- 目前还没有一种方法能完全模拟并达到标引员的标引能力。
- 多种标引模型拟合或方法的集成学习要求每个标引模型标引结果存在差别，同时保证标引结果优于随机猜测的结果。
- 多种标引方法的集成学习、寻求更加理想的机器学习方法，并用于自动标引任务中，是今后自动标引研究的趋势之一。

语义理解、自我学习

- 自动标引技术从最初的寻找“关键词”已经发展到被广泛用于文本检索、自动问答、文本知识发现（或称文本挖掘）等领域。
- 随着互联网海量数据规模的进一步扩大，“信息爆炸”问题将变得更加紧迫。对信息资源进行基于主题的自动标引，并进行后续的数据挖掘，不仅能解决高维数据计算问题，并且能从主题或语义层次上对信息资源进行揭示和控制。随着语义网的不断深入研究和应用，自动标引将不断被赋予新的含义和特定任务。