网络空间智慧搜索相关理论和技术

基于相似性的高性能精准实体搜索

周晓方

Soochow Advanced Data Analytics Lab

苏州大学先进数据分析研究中心

报告提纲

- 关于搜索
- 实体搜索的理解
- 实体搜索理论与技术
- 小结

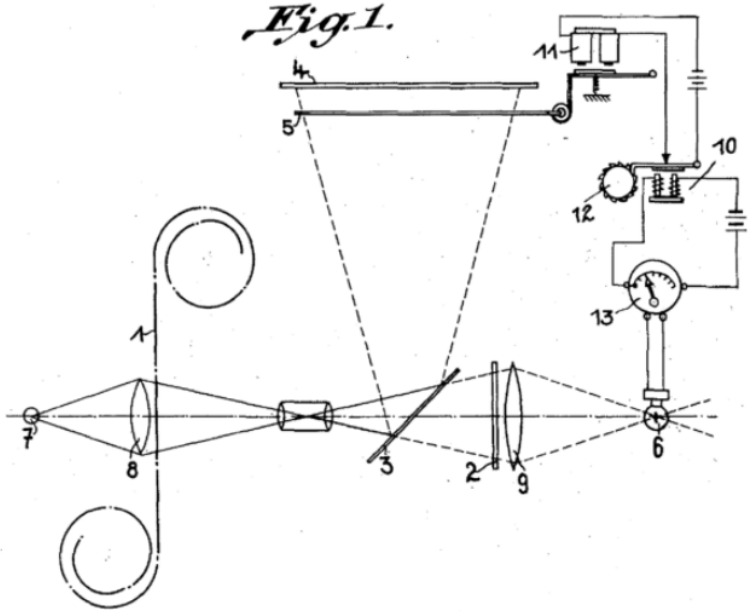
搜索的历史

Dec. 29, 1931. E. GOLDBERG 1,838,389

STATISTICAL MACHINE

Filed April 5, 1928

Fig. 1.



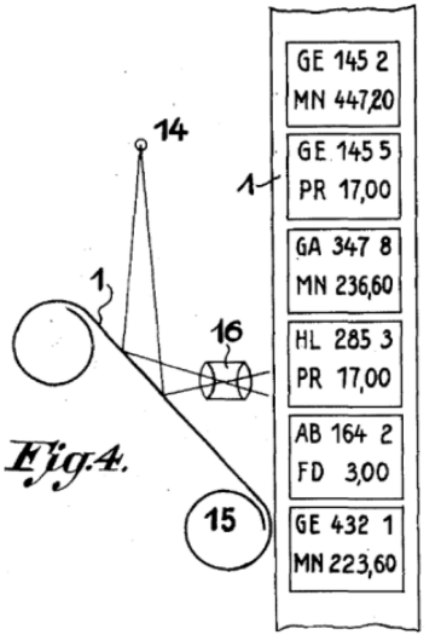


Fig. 4.

Fig. 2.

Fig. 3.

Inventor:

Emanuel Goldberg

现代搜索

	1970－1990	1990－2010	2010-?
需求	记录搜索	文本搜索	实体搜索
应用	企业	Web	网络空间
技术	数据库	信息检索	智慧搜索

搜索的开放性

	数据库搜索	Web搜索	智慧搜索
数据类型	封闭	封闭	开放
数据来源	封闭	开放	开放
相似性度量	封闭	半封闭	开放

需求变化

- 信息源扩展
  - 已有：结构化数据，海量文本数据
  - 增加：时序数据，时空数据，多媒体数据，大图数据
  - 来源：移动互联网，物联网，传感器网络，社交网络，各种监控网络
- 信息需求扩展
  - 已有：单值匹配，关键字匹配
  - 增加：相似性匹配，语境（context）理解，多源动态整合，时空关联，结果加工与理解
- 结论
  - 下一代的搜索技术的核心是网络空间智慧搜索
  - 网络空间智慧搜索信息的载体和搜索的对象是实体

实体（Entity）

- 实体是现实世界或网络空间的一个独立对象
- 数据来源
  - 已有ID的数据
    - Key，URL / URI，DOI / ISBN，RFID Tag，身份证，网络用户，车牌，地点等
  - Named entity
    - 各种实体抽取技术
  - 可以识别的数据
    - 人脸，指纹，车牌，场景
  - 虚拟实体
    - 概念，关系，事件，交互，行为等

实体特性

- 内容属性
  - 内容属性是数据库技术和信息检索技术的管理对象
  - 原生属性：实体固有的原始表述
  - 抽象属性：通过特征提取获得的具有应用针对性的属性
- 时空属性
  - 位置，时间及衍化
  - 空间：地理空间，各种网络空间，维度空间；时间包括版本
- 关系属性
  - 同一实体在不同空间及不同表达之间的关系
  - 与环境之间的关系
  - 实体群体之间的相互关系

现有技术的不足

- 现有的成熟搜索技术主要针对原生属性
  - 在多媒体检索等方面也开始了对抽象属性的研究
  - 这一类研究面临着高维空间检索带来的一系列问题，具体表现在检索的低效和不可扩展性（“维度诅咒”）和抽象属性和用户搜索意图的脱节（“语义鸿沟”）
  - 各种探索尚不具备系统性和可扩展性
- 时空属性和关系属性需要新思路，新理论和新技术
  - 大规模实体抽取和管理， 精准、灵活、有语境的实体表达
  - 统一且有针对性的相似性度量， 用户意图理解
  - 结果表达与可解释性，数据空间理解，交互支持
  - 处理效率

理论与技术

搜索技术	• 搜索模型，语言模型，用户建模，到排表，排序算法，交互方式
实体抽取	• 实体抽取 / 实体识别，实体表达，信息补全，实体链接，实体建模
相似查询	• DB+IR检索，时空检索，相似视频检索，图检索，自定义相似函数，近似查询，用户反馈
高效处理	• 高维空间索引，多空间倒排表及使用，数据采集与存储，大数据处理平台
其它支持	• 非结构化数据管理，计算语言学，数据质量管理，知识库的构建与使用，溯源支持，数据空间探索

信息链与智慧搜索


数据
信息
知识
智慧
数据获取
关键词抽取
组织
搜索
数据获取
实体抽取
集成
分析
搜索
解释
异构多源，
分布实时
大规模，
有语境，
可回溯
Just-in-time,
not just-in-case
灵活实时
隐私安全
高效
交互

信息搜索与信息使用

警告

- 智慧搜索进一步模糊了搜索与使用的边界
- 过度考虑信息使用会影响搜索的通用性和高效性
- 搜索：快速、精确地找到相关信息
- 智慧搜索创新点：扩展相关性，保持快速、精准

小结

- 下一代搜索是网络空间智慧搜索，有需求，有基础，有挑战
- 网络空间智慧搜索可以带动信息技术，计算机科学和信息系统的突破，为人类社会发展作出巨大贡献
- 基于实体的相似性搜索可以成为网络空间智慧搜索的切实载体
- 具体设想：建立一个基于Web，视频和移动对象的大型实体搜索系统，进行大规模实体抽取和互联研究，为扩展现有技术，测试新技术和可用性及性能，获得并积累用户反馈提供一个平台

谢谢！

Soochow Advanced Data Analytics Lab
苏州大学先进数据分析研究中心