



数据科学与计算机学院
School of Data and Computer Science

自然语言处理

Natural Language Processing

权小军 教授

中山大学数据科学与计算机学院

quanxj3@mail.sysu.edu.cn

问答与对话（上）

本节内容

- ◆ 问答系统概述
- ◆ 问答式检索系统
- ◆ 社区问答系统

本节内容

◆ 问答系统概述

- 研究背景

- 发展历史

◆ 问答式检索系统

◆ 社区问答系统

搜索的局限

- 据统计，全球数据量以每年58%的速度增长，2020年将达到40ZB，是2011年数据总量的22倍



IDC统计2014年数据量是多少

网页 新闻 贴吧 知道 音乐 图片 视频 地图 文库 更多»

百度为您找到相关结果约359,000个

域名商2014年度报告:中国数据域名总量跌至22万 - IDC评述网

2015年1月26日 - IDC评述网(idcps.com)01月26日报道:据域名统计机构WebHosting.info数据,截至2014年12月29日,中国数据域名总量已降至221,397个,年净减80,640个,与...
www.idcps.com/news/201... - 百度快照 - 70%好评

IDC | 199IT互联网数据中心 | 中文互联网数据研究资讯中心-199IT

2014年Q4全球PC出货量之差 Gartner与IDC数据相差290万台 2014年第四季度PC出货量呈现稳定趋势,由于统计的方式不同,两家大型统计公司在第四季度PC出货量数字上略有分...
www.199it.com/archives... - 百度快照 - 91%好评

2014年Q4全球PC出货量之差 Gartner与IDC数据相差290万台

2015年1月14日 - 两家大型统计公司在第四季度PC出货量数字上略有分歧:根据Gartner统计2014年第四季度PC出货量达8370万台,同比增长1%;根据IDC统计,2014年第四季度出货量...
www.199it.com/archives... - 百度快照 - 91%好评

域名商年度报告:2014年DNSPod域名总量达174万 - IDC评述网

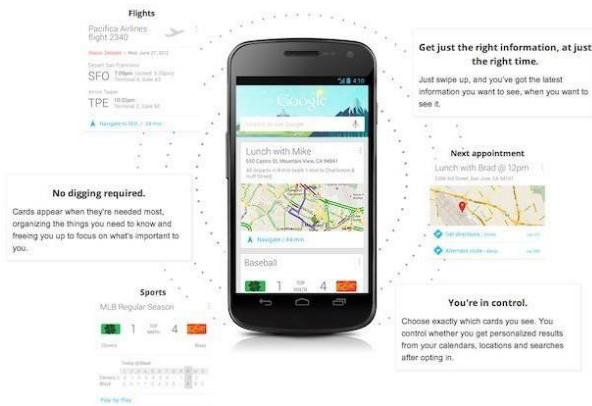
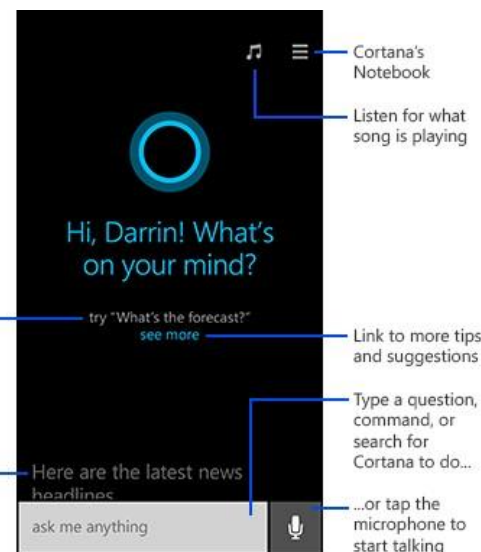
2015年1月15日 - IDC评述网(idcps.com)01月15日报道:据域名统计机构WebHosting.info数据,截至2014年12月29日,DNSPod域名总量已增至174万,年净增长220,393个域名,而...
www.idcps.com/news/201... - 百度快照 - 70%好评

2014年全球软件开发者数量统计_中国行业研究网

2014年1月6日 - 数据分析公司IDC发布最新报告称:2014年全球的软件开发者数量达到1850万,其中1100万是专业的软件开发人员,另外750万是开发爱好者,占全体程序猿的4成之...
www.chinairm.com/news/... - 百度快照 - 94%好评

交互方式的转变

■ 移动互联网以及可穿戴设备的飞速发展需要高效、准确的自然语言形式的信息服务方式



问答系统

问题：谁获得1987年的诺贝尔文学奖？

答案：约瑟夫·布罗茨基

谁获得1987年的诺贝尔文学奖？

网页 新闻 贴吧 知道 音乐 图片 视频 地图 文库 更多»

百度为您找到相关结果约35,300个

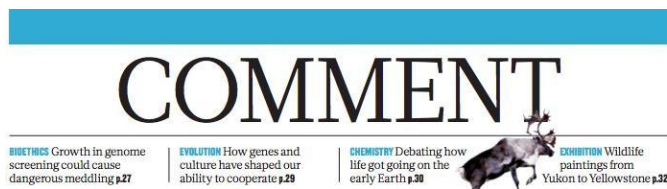
[1987和1988年都获得诺贝尔文学奖的中国作家是谁 百度知道](#)
3个回答 - 提问时间: 2013年01月10日
1988纳吉布·马哈富兹《街魂》他通过大量刻画入微的作品—洞察一切的现实主义,唤起人们树立雄心—形成了全人类所欣赏的阿拉伯语言艺术。1987约瑟夫·布罗茨基《从彼得...
[zhidao.baidu.com/link?...](#) - 80%好评

[1987年诺贝尔文学奖得主列里的喜 新浪博客](#)
2009年...然而七年之后,又是不到十年,布罗茨基取得了和米沃什一样的成就,获得了1987年诺贝尔文学奖,成为诺贝尔文学奖...
[blog.sina.com.cn/s/blo...](#) - 百度快照 - 88%好评

[谁获得1987年的诺贝尔文学奖? 学位论文\(共66篇\) 百度学术](#)
1987年诺贝尔文学奖得主布罗茨基谈话...译林》
J·勃罗斯基获1987年诺贝尔文学奖 《外国文学研究》
一九八六年诺贝尔文学奖获得者沃·索英卡诗选 《诗刊》
[查看更多相关论文>>](#)
[xueshu.baidu.com](#)

[谁获得了1982年的诺贝尔文学奖? 百度知道](#)
4个回答 - 提问时间: 2013年09月18日
1982年诺贝尔文学奖:加夫列尔·加西亚·马尔克斯(1928~)哥伦比亚记者、作家。主要作品有长篇小说《百年孤独》、《家长的没落》、《霍乱时期的爱情》、《迷宫中的将军》...
[zhidao.baidu.com/link?...](#) - 百度快照 - 80%好评

问答系统是下一代搜索引擎的基本形态



Search needs a shake-up

On the twentieth anniversary of the World Wide Web's public release, **Oren Etzioni** calls on researchers to think outside the keyword box and improve Internet trawling.

Two decades after Internet pioneer Tim Berners-Lee introduced his World Wide Web project to the world using the alt.hypertext newsgroup, web search is on the cusp of a profound change — from simple document retrieval to question answering. Instead of poring over long lists of documents that contain requested keywords, users need direct answers to their questions. With sufficient scientific and financial investment, we could soon view today's keyword searching with the same nostalgia and amusement reserved

for bygone technologies such as electric typewriters and vinyl records.

But this transformation could be unreasonably delayed. As a community, computer scientists have underinvested in tools that can synthesize sophisticated answers to questions, and have instead focused on incremental progress in lowest-common-denominator search. The classic keyword search box exerts a powerful gravitational pull. Academics and industry researchers need to achieve the intellectual 'escape velocity' necessary to revolutionize

search. They must invest much more in bold strategies that can achieve natural-language searching and answering, rather than providing the electronic equivalent of the index at the back of a reference book.

Today, that 'book' is distributed over billions of web pages of uneven quality, and much effort has been directed at ranking the most useful results. Such engines readily index billions of documents, but overwhelm their users with millions of results in response to simple queries. This quandary only worsens as the number of web pages



Prof. Oren Etzioni

Turing Center
University of Washington

以直接而准确的方式回答
用户自然语言提问的自动
问答系统将构成下一代搜
索引擎的基本形态

— 《Nature》 2011.8

© 2011 Macmillan Publishers Limited. All rights reserved.

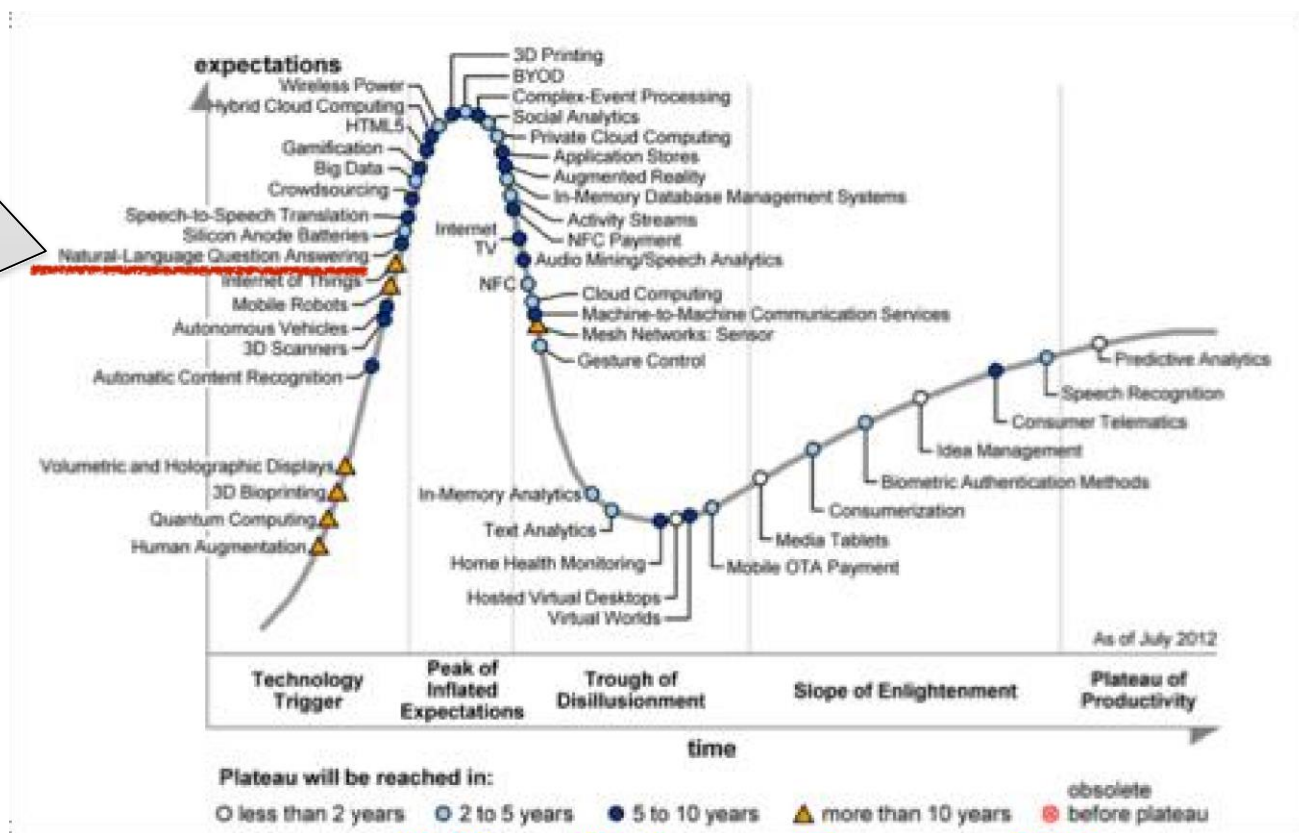
4 AUGUST 2011 | VOL 476 | NATURE | 25

问答系统是未来信息技术的重要突破口

■ 世界权威IT市场调查咨询公司高德纳（Gartner）2012年8月发布《2012新兴技术成熟度曲线》

自然语言问答

在未来5-10年将成为主要新兴技术关注点，极有可能迎来技术突破



问答系统概述: 定义

◆ 定义

输入：自然语言的问句，而非关键词的组合

例如：谁获得1987年的诺贝尔文学奖？

输出：直接答案，而非文档集合

例如：约瑟夫·布罗茨基

问答系统任务

大规模知识库



自然语言问句

我国有两个省接壤的省份最多，与它们接壤的省分别有哪些？

涉及到自然语言理解、知识表示、知识推理等人工智能的难点问题

QA研究对于自然语言理解的意义

- ❑ 美国认知心理学家G.M.Olson认为，判别计算机理解自然语言的4个标准是：QA、Summarization、Paraphrase和MT。计算机只要达到以上标准之一，就认为它理解了自然语言；
- ❑ 自然语言理解：词语层面、句子层面、篇章层面、篇章之间、语言之间的基本问题，在QA中都会出现；
- ❑ QA和信息检索密切相关，信息检索中的基本问题在QA中也同样存在；
- ❑ QA是自然语言理解研究人员追求的目标之一，它的研究会带动自然语言理解的发展；

技术应用：WATSON



辅助
医疗



金融辅
助决策

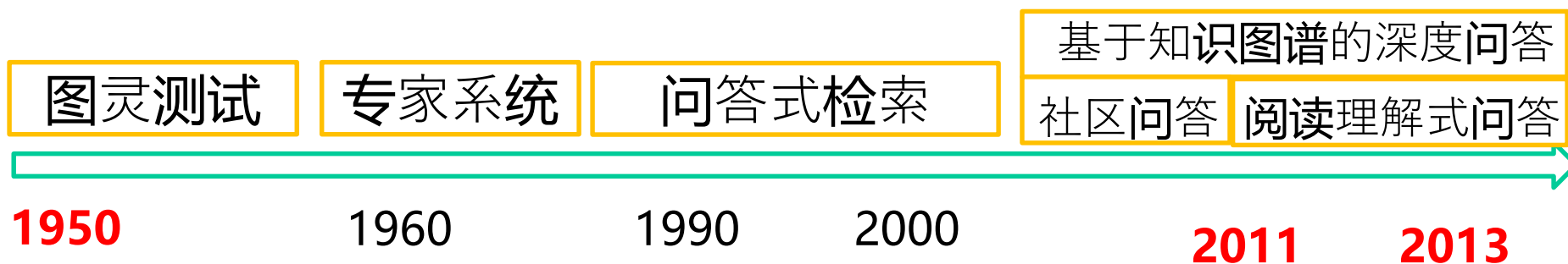


企业
服务

WATSON应用于辅助医疗

- 全美每年20%的诊断（1.5 million）是错误的或者不完全的
- 从海量医疗档案以及医学期刊进行Deep Mining，挖掘有用的医疗信息，用户的病历作为Clue，挖掘最合适的诊疗方案
- 医生不再需要阅读海量的医学期刊，以获取知识的更新
- 用户也可以及时获得医疗咨询服务

问答发展历程



问答发展历程

□ 1950年，A. M. Turing 提出“图灵测试”

- 一个人在不接触对方的情况下，通过一种特殊的方式，和对方进行一系列问答。如果在相当长时间内，他无法根据这些问题判断对方是人还是计算机，那就可以认为这个计算机具有同人相当的智力，即这台计算机是能思维的。

理想模型

问答发展历程

□ 基于知识推理的问答系统

□ 特点：答案或者从知识库中检索得到，或者在知识库上经过推理得到；

□ 代表性系统：

- 自然语言界面的专家系统：如：MIT开发的数学符号运算系统 MACSYMA (1960年左右)
- 基于本体的问答系统：如陆汝钤院士主持开发的Pangu (2000年)

问答发展历程

□ 基于知识推理的问答系统

□ 特点：答案或者从知识库中检索得到，或者在知识库上经过

对话者：“动物园有一头黑熊死了。”

计算机：“黑熊是怎么死的？”

对话者：“据说黑熊是吃塑料袋死的。”

计算机：“准是哪个不文明的游客投进去的。”

统

年)

问答发展历程

□ 基于知识推理的问答系统

□ 优点：性能良好，对于用户提出的许多问题，回答准确，具有一定的推理能力；

□ 缺点：

- 人工建设知识库非常困难，知识库规模有限，领域有限；
- 如果用户的问题超出了知识库的范围，系统性能很差

问答发展历程

□ 问答式检索系统

- 搜索引擎为人们的信息获取提供了可能，但搜索引擎无法清楚表达人们的信息需求意图，返回的信息太多；
- 为了克服搜索引擎的不足，问答式检索系统应运而生；
- 主要特点：利用信息检索以及浅层自然语言处理技术从大规模文本库或者网页库中抽取出答案；
- 代表性系统：
 - MIT开发的Start (<http://start.csail.mit.edu/index.php/>)

问答发展历程

□ 问答式检索系统

□ 优点：

□ 相对于基于知识推理的问答系统而言：不受知识库规模限制，不受领域限制，更加接近真实应用需求；

□ 相对于搜索引擎而言：问答式检索系统接受的是自然语言形式的提问，由于自然语言处理技术的应用，对用户意图的把握更加准确，呈现给用户的答案更加准确；

□ 缺点：目前问答式检索系统仅能处理有限的简单问题，如 Factoid 问题等；

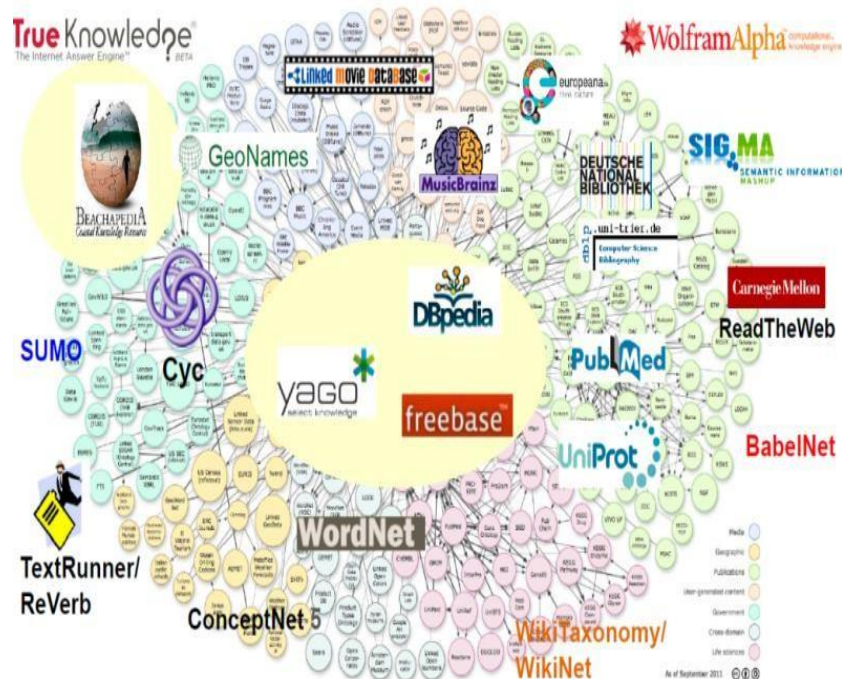
问答发展历程

□ 社区问答系统：

- 随着Web 2.0的兴起，基于用户生成内容(User-Generated Content,UGC)的互联网服务越来越流行，社区问答系统应运而生，为问答系统的发展注入了新的活力和生机；
- 用户可以提出任何类型的问题，也可以回答其它用户的问题，通过问答方式来满足人们的信息查找和知识分享需求；
- 代表性系统：
 - 英文：Yahoo! Answers等；
 - 中文：百度知道、新浪爱问、**知乎**等；

问答发展历程

基于知识图谱的问答系统



自然语言问句

我国有两个省接壤的省份最多，与它们接壤的省分别有哪些？

大规模知识库

内蒙古：黑龙江、吉林、辽宁、河北、山西、陕西、宁夏、甘肃
陕西：内蒙古、宁夏、甘肃、四川、重庆、湖北、河南、山西

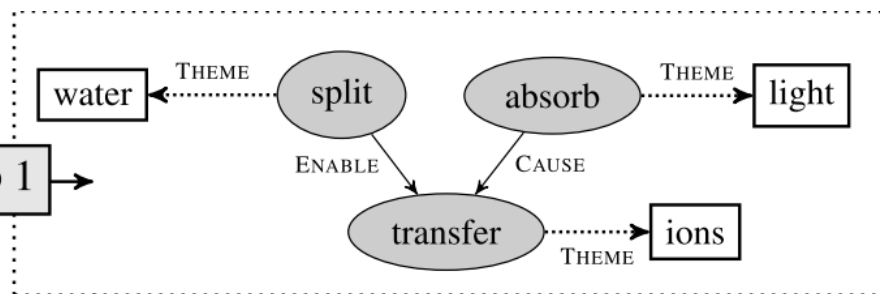
问答发展历程

□ 阅读理解系统

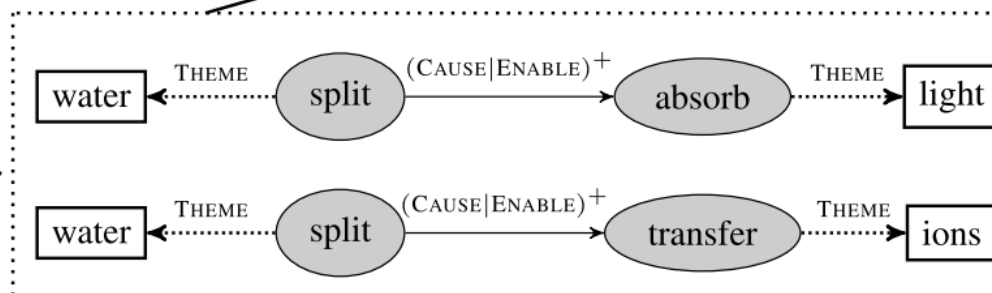
“... **Water is split**, providing a source of electrons and protons (hydrogen ions, H^+) and giving off O_2 as a by-product. **Light absorbed** by chlorophyll drives a **transfer of the electrons and hydrogen ions** from water to an acceptor called $NADP^+$...”

- Q What can the splitting of water lead to?
a Light absorption
b Transfer of ions

Step 1



Step 3: Answer = b



Step 2

小结

| | 知识推理系统 | 问答式检索 | 社区问答 | 知识图谱问答 | 阅读理解 |
|-------|--------|-------|-------|--------|----------------|
| 数据源规模 | 小 | 大 | 大 | 大 | 大 |
| 数据源质量 | 没噪 | 噪声较小 | 有噪 | 有噪 | 有噪 |
| 查询类型 | 受限制 | 受限制 | 不受限制 | 不受限制 | 不受限制 |
| 典型应用 | 专家系统 | 事实性问答 | 一般性问答 | 事实性问答 | 事实性问答 一般性问答 |

本节内容

◆ 问答系统概述

◆ 问答式检索系统

□ 问答式检索方法

□ 问答式检索应用：Watson

◆ 社区问答系统

问答式检索方法

- 信息检索 + 信息抽取
- 信息检索 + 模式匹配
- 信息检索 + 自然语言处理技术
- 基于统计翻译模型的问答技术

信息检索 + 信息抽取

❑ **方法描述**：从问句中提取关键词语，用信息检索的方法找出包含候选答案的段落或句子，然后基于**问答类型**用信息抽取的方法在这些段落和句子中找出答案

❑ **检索**：段落或者句子级排序，利用不同类型关键词的加权组合

❑ **答案抽取**：根据问答类型从排序后的段落或句子中抽取答案

❑ **特点**：

❑ **优点**：技术相对成熟，易于开发

❑ **缺点**：准确率一般，不能推理

信息检索 + 模式匹配

□ 方法描述：

□ 基本思想：对于某些提问类型，问句和包含答案的句子之间存在一定的答案模式，该方法在信息检索的基础上根据这种模式找出答案。

□ 例如，询问“某人生日年月日”类提问的部分答案模式：

- 1.0 <NAME> (<ANSWER> -)
- 0.85 <NAME> was born on <ANSWER>
- 0.6 <NAME> was born in <ANSWER>
- 0.59 <NAME> was born <ANSWER>
- 0.53 <ANSWER> <NAME> was born
- 0.50 – <NAME> (<ANSWER>

信息检索 + 模式匹配

□ 包括两阶段的任务:

- 离线阶段: 获取答案模式
- 在线阶段: 首先判断当前提问属于哪一类, 然后使用这类提问的所有模式来抽取候选答案

□ 模式获取方法:

- 表层字符串匹配(Ravichandran ACL 2002)
- 深层句法分析(Lin NLE 2001)

□ 特点:

- 优点: 对于某些类型的问题(如生日问题等) 效果良好
- 缺点: 无法表达长距离、复杂关系, 没有推理能力

信息检索+自然语言处理技术

□方法描述:

- 对问句和答案句进行浅层句法分析, 获得句子的浅层句法、语法表示, 作为对前两种方法的补充和改进

□涉及到的自然语言处理技术主要包括:

- 命名实体识别技术(Ravichandran ACL 2002)
- 句法分析技术(Lin NLE 2001)
- 逻辑表示(Harabagiu TREC 2000; Moldovan ACL 2001)
- ...

信息检索+自然语言处理技术

□特点:

- 优点: 能够从句法、语义的角度解析答案
- 缺点: 技术还不成熟

□代表系统:

- Sanda Harabagiu等人研发的系统(Harabagiu TREC 2000)
，该系统在TREC QA Track 评测中获得好成绩，且具有较大的领先优势

基于统计翻译模型的问答技术

□方法描述：

- 把提问句看作答案句在同一语言内的一种翻译

□特点：

- 过分依赖于训练集

四类问答技术的比较分析

- **基于信息检索和信息抽取的问答技术**：相对简单，容易实现。但它以基于关键词的检索技术(或称为词袋检索技术)为重点，只考虑离散的词，不考虑词之间的关系。因此无法从句法关系和语义关系的角度解释系统给出的答案，也无法回答需要推理的提问
- **基于模式匹配的问答技术**：虽然对于某些类型提问(如定义，出生日期提问等)有良好的性能，但无法找到所有提问的答案模式，长距离模式和表达复杂关系的模式的获取也很困难，同样无法实现推理

四类问答技术的比较分析

- ❑ **基于自然语言处理的问答技术：**可以对提问和答案文本进行一定程度的句法和语义分析，从而实现推理。但自然语言处理技术还不成熟，除一些浅层的技术(汉语分词、命名实体识别、词性标注等)外，其他技术还没有达到实用程度。所以这种技术的作用还有限，只能作为对前两种方法有效补充
- ❑ **基于统计翻译模型的问答技术：**在很大程度上依赖训练语料的规模和质量，而对于开放域问答系统，这种大规模训练语料的获取是非常困难的

新的方向

基于深度学习的方法：把回答问题的过程看作一个黑盒子，通过复杂的神经网络和超大规模的数据集训练出一个拟合能力强大的模型；

本节内容

◆ 问答系统概述

◆ 问答式检索系统

□ 问答式检索方法

□ 问答式检索应用：Watson

◆ 社区问答系统

问答式检索应用：Watson

- ❑ 沃森 (Watson)：2011 年，IBM 研发的计算机“沃森”正在美国智力竞赛节目《危险边缘 Jeopardy!》中上演“人机问答大战”，战胜人类选手；
- ❑ DeepQA 问答系统是 Watson 实现的核心。



问答式检索应用：Watson

□ 强大的硬件平台

- 90台IBM服务器、360个CPU 组成，每个CPU主频可达 4.1GHz
- 15TB内存、每秒可进行80万 亿次运算
- Linux操作系统
- 分布式计算



问答式检索应用：Watson

□强大的知识资源

- 存储了大量图书、新闻和电影剧本资料、辞海、文选和《世界图书百科全书》等海量的资料
- 每当读完问题的提示后，沃森就在不到三秒钟的时间里对资源库"挖地三尺"，在长达2亿页的资料里展开搜索

问答式检索应用：Watson

□ 自然语言处理、信息抽取和知识工程技术

□ 统计机器学习理论、句法分析

- 基于句法和语义结构的文本表示

□ 主题分析

- 确定问题主题，然后确定给定主题的最佳文本资源

□ 信息抽取

- 对象-属性-值数据模型，该模型支持对半结构化数据源中的信息进行有效的检索，以回答自然语言问题

问答式检索应用：Watson

□ 自然语言处理、信息抽取和知识工程技术

□ 大规模知识库集成和知识推理技术

- 将大量不同来源的资料转化为该系统的知识资源，并利用这些知识进行推理，以发现答案的矛盾和差异之处

□ 信息检索技术

- 寻找和检索最可能包含准确答案的文字

□ 知识工程

- 将多源异构知识统一集成和表示，方便后续的推理

本节内容

- ◆ 问答系统概述
- ◆ 问答式检索系统
- ◆ 社区问答系统
 - 研究背景
 - 主要任务和方法

传统问答式检索系统的问题

□ 目前问答式检索系统仅能处理有限的简单的问题，如 Factoid 问题等

而用户的提问是复杂多样的

社区问答的出现

- 随着Web2.0的兴起，基于用户生成内容的互联网服务越来越流行，社区问答服务的应运而生，为人们快速获取信息提供了更方便的选择
- 用户可以提出任何类型的问题，也可以回答其它用户的问题，区别于传统问答式检索系统仅能回答有限类型的问题

社区问答的出现

| 网站名称 | 发布时间(年) | 网址(http) |
|---------|---------|------------------------|
| 新浪爱问知识人 | 2004 | iask.sina.com.cn |
| 百度知道 | 2005 | zhidao.baidu.com |
| 雅虎知识堂 | 2006 | knowledge.yahoo.com.cn |

最新：知乎、Reddit、Quora

| | | |
|----------------|------|-------------------|
| Naver | 2002 | kin.naver.com |
| WikiAnswers | 2002 | wiki.answers.com |
| Answerbag | 2003 | www.answerbag.com |
| Yahoo! Answers | 2005 | answers.yahoo.com |

本节内容

◆ 问答系统概述

◆ 问答式检索系统

◆ 社区问答系统

□ 研究背景

□ 主要任务和方法

主要任务

□ 预处理

- 问题分类

- 作弊检测

□ 与回答新提交问题相关的研究

- 相似问题检索

- 答案质量评估

- 专家用户推荐

□ 与用户体验相关的研究

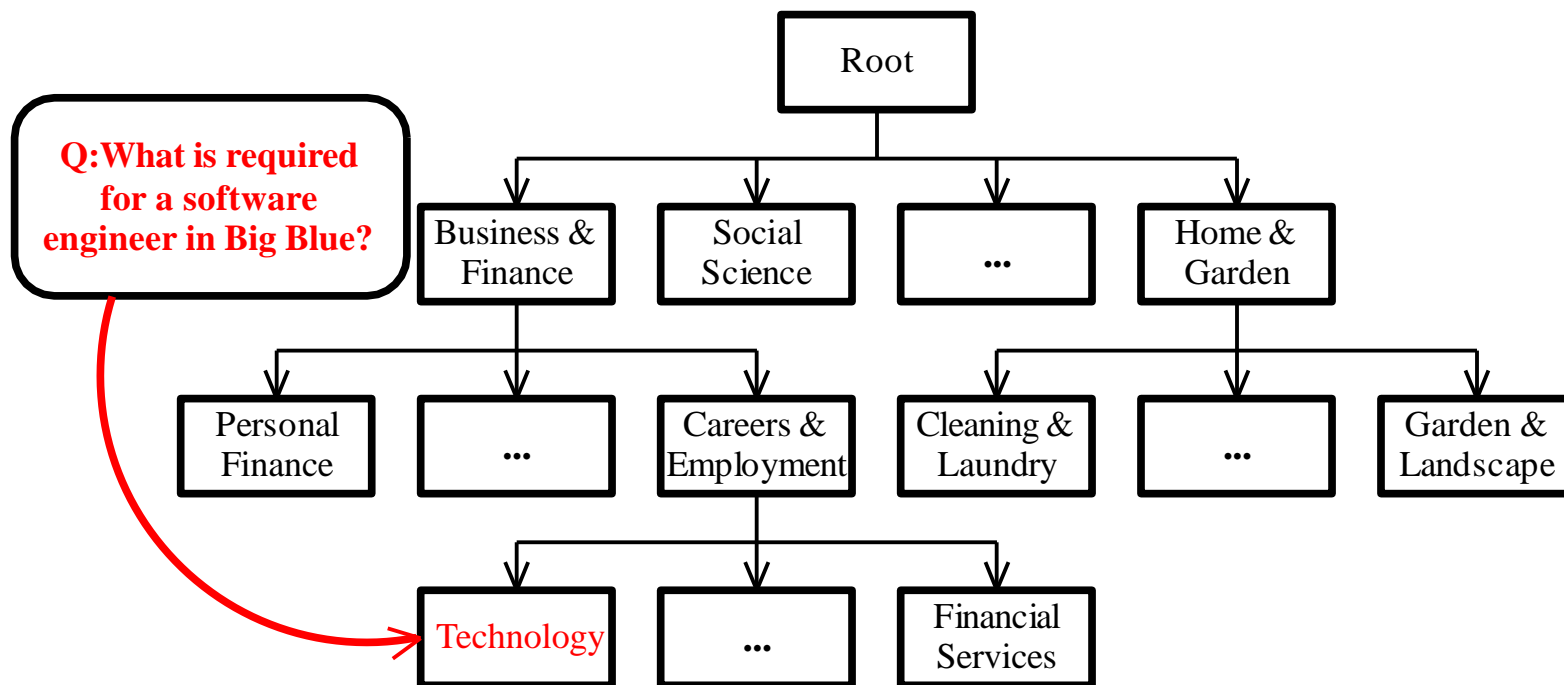
- 用户满意度预测

- 用户社区结构挖掘

- 热点话题检测

问题分类

□ **任务：** 将用户提问自动分到社区问答系统对应的类别体系中



Yahoo! Answers 的类别体系结构

问题分类

□ 难点:

- 类别多且分布不均衡(Yahoo! Answers中类别: 1,535)
- 用户的提问往往很短, 包含的信息少

□ 问题:

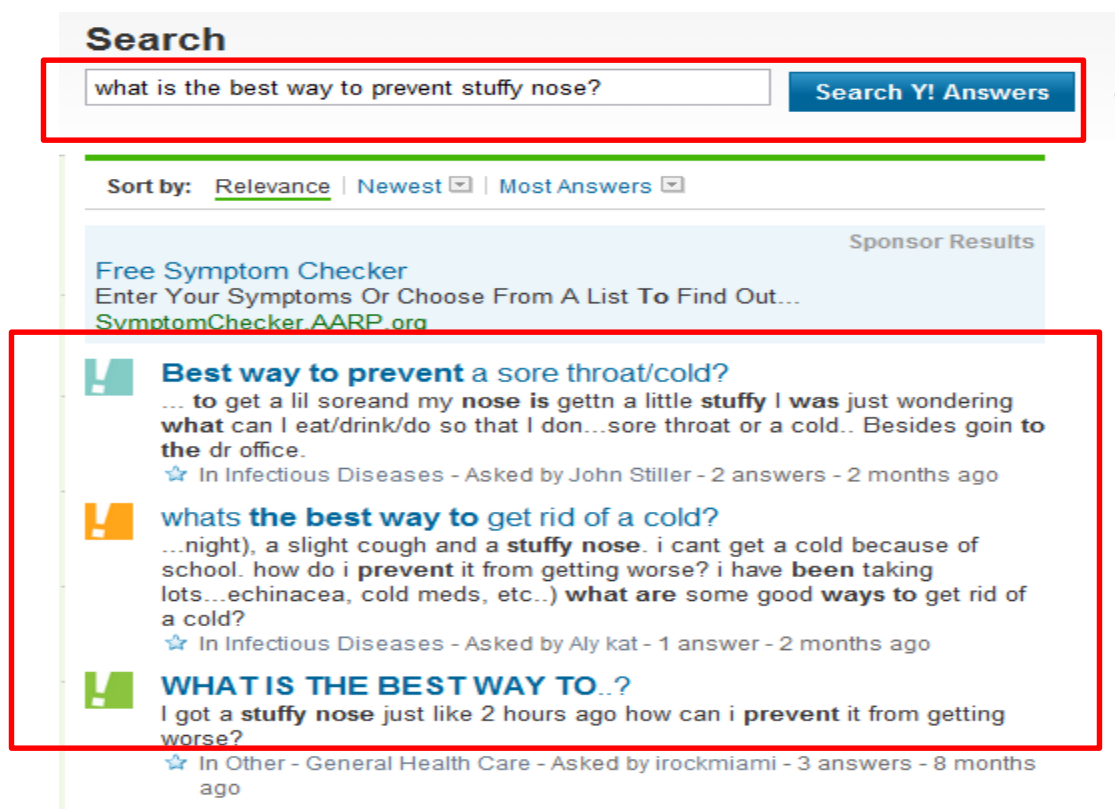
- 准确率很低, 导致马太效应
- 数据稀疏

借用传统文本分类和问答式检索中问题分类的方法, 无法取得满意的结果

用户的提问一般很短, 比如在Yahoo! Answers中用户提问的平均长度为11.2个词

相似问题检索

□ **任务：** 问题检索指基于大规模用户产生的问答数据集提供的信息检索服务，即从数据集中找出与用户提问相似的问题，这些相似问题的答案可以作为用户提问问题的答案候选



用户的查询

返回的相关问题

相似问题检索

核心、挑战、方法

- 核心：计算两个问题的相似度
- 挑战：问题一般较短，包含的信息很少，词汇鸿沟问题很严重
- 主要方法有：
 - 传统的信息检索模型，比如VSM, BM25等
 - 语言模型(Zhai SIGIR 2001; Jeon CIKM 2005)
 - 基于词的翻译模型(Jeon CIKM 2005; Xue SIGIR 2008)
 - 基于短语的翻译模型(Zhou ACL 2011)

答案质量评估

□ 社区问答作为一种社会媒体，数据是由用户自动生成的，造成质量差异的主要因素有：

- 用户水平的高低

- 自然语言表述上的多样性

□ 通过对答案质量的自动分析，将全部候选答案按质量高低排序后展现给用户，可以节省用户的浏览时间，增强用户满意度

答案质量评估

问题：谁比我帅？

答案：哈哈，我啊

问题：What can I do tonight?

答案：Sleep!!!



低质量！

答案质量评估

问题：网上买手机应该注意什么？

答案：你买之前一定问好，有没有保修，保修凭证是什么。快递是什么，几天到。还有就是看店铺了，最好找皇冠以上店铺，问清楚，你要买的翻新机有什么问题，可能有什么问题。

高质量

问题：How to get from north london to thorpe park?

答案：Take the Underground to Waterloo Station. From Waterloo take an SWT Reading or Windsor train to Staines Station. They run four times per hour and the journey takes 35 minutes.

答案质量评估

□ 主要方法:

- 采用统计机器学习方法，比如分类或回归等

□ 核心：如何选取特征

□ 文本特征

□ 非文本特征

- 回答者的问题采纳率
- 答案长度(也可以称为是文本化的特征)
- 用户推荐的次数
- 页面点击次数

答案质量评估

| 特征 | 因素 | 维度 | 代表工作 |
|-----|------|--------|---------------------------|
| 文本 | 答案 | 准确性 | Harper 2008; Blooma 2008 |
| | | 完整性 | Blooma 2008 |
| | | 可读性 | Zhu 2009 |
| | | 合理性 | Blooma 2008 |
| | | 相关性 | Blooma 2008; Zhu 2009 |
| | | 新颖性 | Zhu 2009 |
| | | 内容长度 | Jeon 2006; Lee 2007 |
| 非文本 | 提问者 | 权威性 | Blooma 2008 |
| | 回答者 | 权威性 | Jeon 2006; Bian 2009 |
| | 问题 | 主题类别 | Jeon 2006 |
| | 答案 | 时间、位置等 | Jeon 2006; Bian 2009 |
| | 问答页面 | 点击信息等 | Jeon 2006; Agichtein 2008 |

答案质量评估

- ❑ 质量评估问题是一个非常主观的工作
- ❑ 目前的主流方法是采用统计机器学习方法，将其看成是一个分类或回归的问题
- ❑ 难点：主观性较强，难以评测或对比

专家用户发现

□ 专家用户发现（回答的准确）

- 在社区问答系统中，专家用户给出的答案更有可能被提问者或者其他投票者选为最佳答案
- 如果一个用户回答的质量较高的话，往往称为是专家用户（只考虑一个领域，不涉及具体的提问）

专家用户发现

- ❑ 任务：在社区问答系统中找出哪些用户相对更加权威
 - ❑ 用户回答问题的情况
 - ❑ 用户提问的情况
- ❑ 这里“专家用户”是个相对概念

专家用户发现

□主要方法:

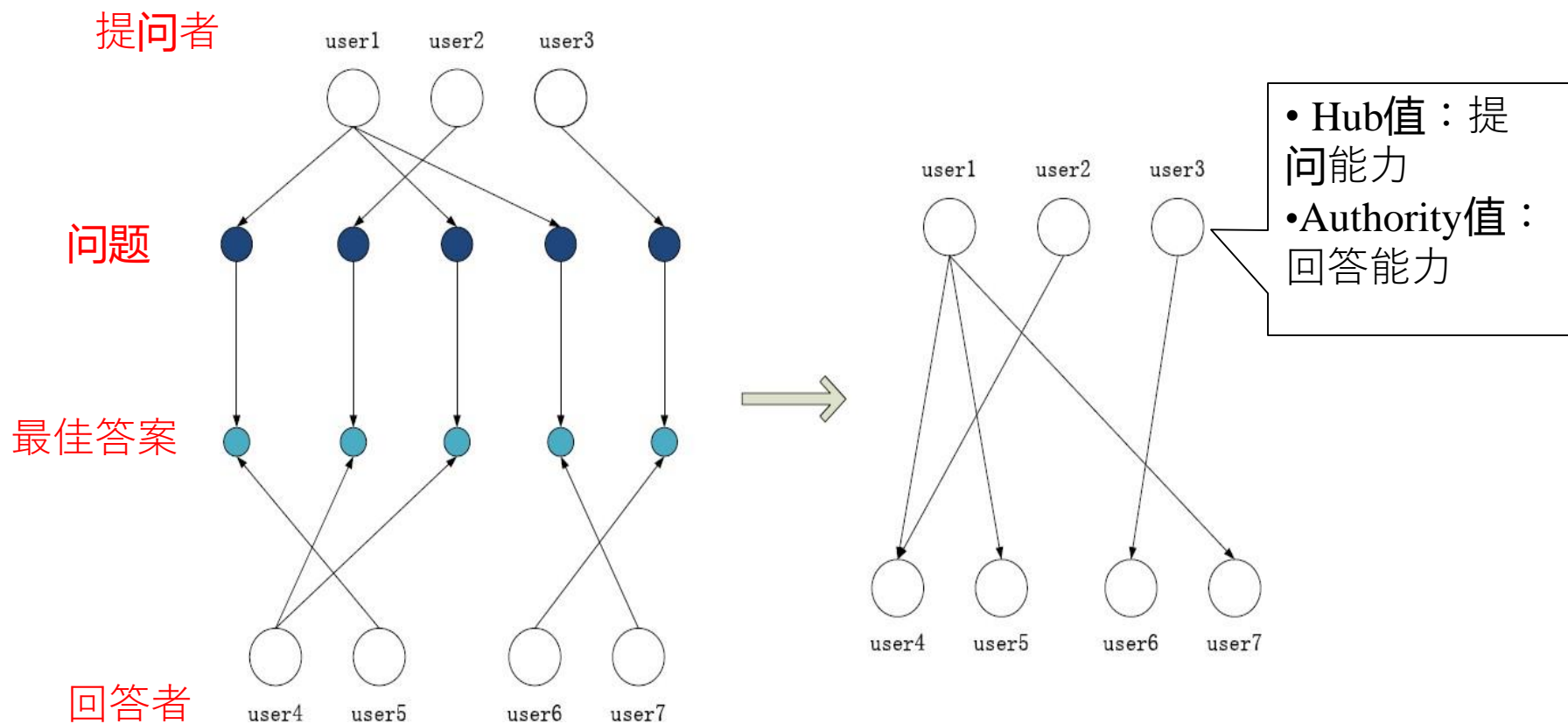
- 基于图结构的专家用户发现(Jurczyk CIKM 2007)
- 基于启发式的专家用户发现(Bouguessa KDD 2008)

基于图结构的专家用户发现

□方法：采用类似网页排序任务中的PageRank算法和HITS算法

- 通过问答建立用户之间关系，如果一个用户对另一个用户提出的问题做了最佳回答，那么这两个用户之间就存在一条有向边
- 一个用户既可以提出问题又回答许多问题，整个图呈网状结构
- 对每个顶点计算Hub值（出度）和Authority值（入度），分别对应用户的提问能力和回答能力

基于图结构的专家用户发现



基于图结构的专家用户发现

- ▣ 顶点的Hub值为其所有指向它的顶点的Authority值的加权和

$$Hub(i) = \sum_{j=1, \dots, N} w(i, j) Authority(j)$$

- ▣ 顶点的Authority值为其所有指向它的顶点的Hub值的加权和

$$Authority(j) = \sum_{i=1, \dots, M} w(i, j) Hub(i)$$

Thank you!

权小军 中山大学数据科学与计算机学院