|  |
| --- |
| 基于知识图谱的限定领域问答系统 |



1. 项目背景

机构部门的干部任免需要考察干部的综合信息，而这些信息的来源、存储和检索需要一个集成部署的平台。以往的信息管理系统采用结构化的数据存储和基于规则的精确查询。为了优化存储、精简系统以及使搜索结果更丰富、可推理，本文基于知识图谱构建了一个基于知识图谱的组织干部领域问答系统。

1. 要解决的问题
2. 知识库的构造
3. 自然语言处理
4. 用到的软件和技术

软件：kettle、kafka、neo4j、elasticsearch

StandFord CoreNLP、Tinkerpop

技术：

1. 从关系数据库到图数据库

数据源为关系型数据库，将结构化、半结构化、文本数据抽取

处理后导入消息队列，再自动化的创建图到图数据库中。

1. 模型定义

实体定义、属性决策、关系建立

1. 语义网络搜索

对搜索关键词进行语义分析，提取相关概念，查询知识库，对返 回结果进行推理、统计、排序。

1. 系统整体架构

自然语言文本

|

分词、命名实体识别、词性标注 —— 分词模型

|

实体对齐（同义词替换。。） —— 词典

|

主题词（实体、关系）提取 —— 词典

|

实体链接 —— 知识库

关系抽取 —— 模型

|

1. 引言

互联网的发展，给世界带来了越来越多的数据，所以说我们正在步入数据时代。数据描绘着我们的生活，改变着我们的生活。如今，随着Linking Open Data 等项目的全面展开，语义Web数据源的数量激增，大量RDF数据被发布。互联网正从仅包含网页和网页之间超链接的文档万维网(Document Web)转变成包含大量描述各种实体和实体之间丰富关系的数据万维网(Data Web)。在这个背景下，Google、百度和搜狗等搜索引擎公司纷纷以此为基础构建知识图谱，分别为Knowledge Graph、知心和知立方，来改进搜索质量，从而拉开了语义搜索的序幕。

知识图谱的定义，知识图谱主要的目标是用来描述真实世界中间存在的各种实体和概念，以及它们之间的关联关系。

知识图谱旨在描述真实世界中存在的各种实体或概念。其中，每个实体或概念用一个全局唯一确定的ID来标识，称为它们的标识符(identifier)。每个属性-值对(attribute-value pair，又称AVP)用来刻画实体的内在特性，而关系(relation)用来连接两个实体，刻画它们之间的关联。知识图谱亦可被看作是一张巨大的图，图中的节点表示实体或概念，而图中的边则由属性或关系构成。上述图模型可用W3C提出的资源描述框架RDF[2] 或属性图(property graph)[3] 来表示。知识图谱率先由Google提出，以提高其搜索的质量。

为了更好地理解知识图谱，我们先来看一下其在搜索中的展现形式，即知识卡片(又称Knowledge Card)。知识卡片旨在为用户提供更多与搜索内容相关的信息。更具体地说，知识卡片为用户查询中所包含的实体或返回的答案提供详细的结构化摘要。从某种意义来说，它是特定于查询(query specific)的知识图谱。例如，当在搜索引擎中输入“姚明”作为关键词时，我们发现搜索结果页面的右侧原先用于置放广告的地方被知识卡片所取代。广告被移至左上角，而广告下面则显示的是传统的搜索结果，即匹配关键词的文档列表。

知识图谱的作用：

第一个就是用来搜索，这也是Google提出知识图谱的初衷；

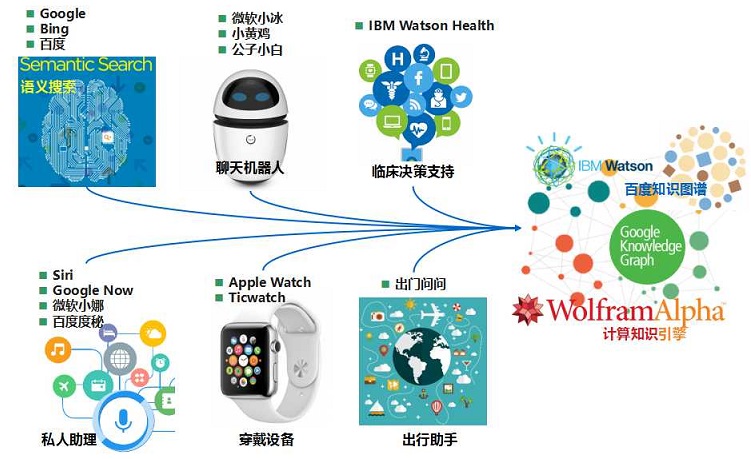
第二个就是聊天机器人，我相信在坐的各位或多或少都有用过，比如说微软的小冰，王昊奋老师公司的公子小白；

第三个就是用来做问答，有一个贡献非常大的，就是IBM Watson，这在业内非常有名；

然后现在还有很多私人的助理，比如说我们苹果手机里的Siri，还有微软的小娜，百度的度秘；

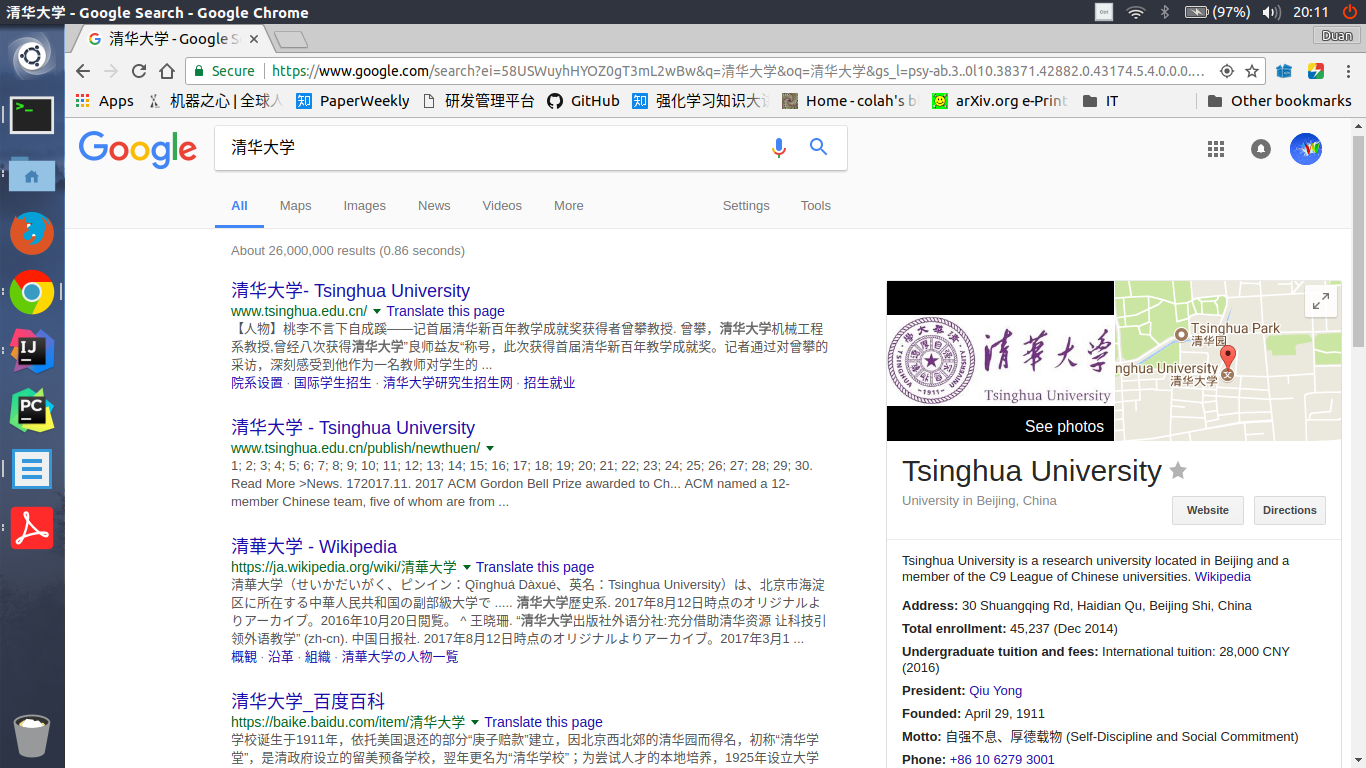
同时我们还有很多的穿戴设备，它里面也用到了知识图谱相关的技术，比如iWatch里面就用了；

最后一个就是我们出行的助手，国内做的非常好的“出门问问”，它也是用了相关的技术。



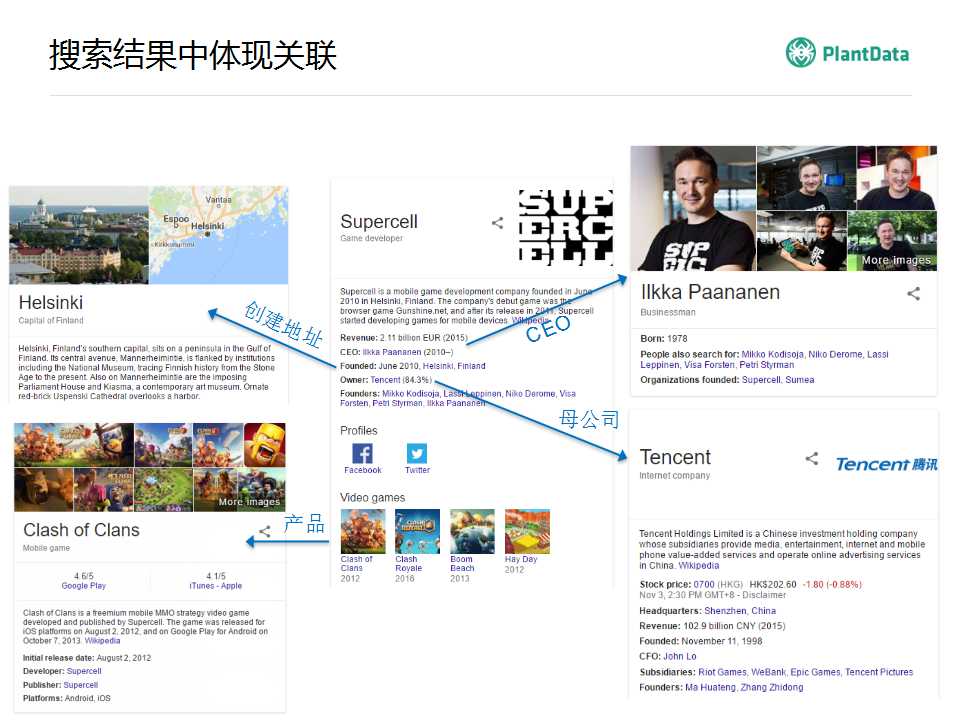
知识图谱的第一个应用，也就是知识图谱提出的初衷——用来改善搜索。

1、为搜索提供结构化结果



如果我们现在去 Google 搜清华大学的时候，我们可以看到右边有一个知识卡片，它里面包含了北理工的很多基本介绍，同时还有一些基本的属性，这是知识图谱用在搜索里面给我们带来的第一个变化；

搜索结果中体现关联：



在社交网络中也用到了知识图谱，如果用过Facebook，它就用了知识图谱的技术去链接里面的人物、地点和事件，正如我们举的两个例子：如果去搜喜欢哈佛大学的人，它可以直接给你答案；第二个例子，搜去过哈佛大学的人，它同样可以给你答案；

自动问答：

自动问答目前也是一个非常热门的方向，这可能是面向应用最直接的方式，目前不管是学术界还是工业界都在做相关的研究，这里有两个例子，百度的度秘，苹果的Siri，可以看到自然语言问答的结果。

通过知识图谱，我们在搜索引擎里面可以看到丰富的结果，在右边我们可以看到和清华相关的高校，以及相关的其他一些事物，这就是知识图谱给我们带来的搜索的改变。

对于开放领域的知识搜索，返回页面链接就已经足够。但很多领域我们需要返回更精确的结果，这时问答系统就顺应潮流而发展起来。其中，基于知识图谱的问答系统由于利用了语义网络的结构特性，具有精确查询，可推理，可扩展等优点，

1. 论文研究背景和意义

本文构建的是基于知识库的领域问答系统，即KB-QA。首先是建造一个领域的知识库。在现实世界中，知识多是非结构化展现的，如一句话、一段自然语言文本、一个网页。这些知识易于人的阅读和理解，