

知识图谱在大数据反欺诈领域的应用与实践



点融黑帮 (/u/dd56464a4e4d) (+关注)

2016.09.23 11:04* 字数 2471 阅读 2255 评论 0 喜欢 21 赞赏 1

(/u/dd56464a4e4d)

1、为什么要用大数据来反欺诈？

近些年来互联网金融蓬勃发展，特别是P2P的兴起，颠覆了传统的银行贷款模式，给大众带来快速便捷的金融服务；在P2P行业中，借款端的风险是P2P公司面临的主要风险，而借款端的风控水平可以说决定了一家P2P公司的核心竞争力。

借款端风险的一个主要来源是欺诈风险，传统的反欺诈手段主要依赖于信息的人工审核，而身份证、手机号码、银行流水等材料的伪造成本非常低，各类信贷服务机构均不得不投入大量的人力用于核实信息主体的身份及其提供材料的真实性；在这种形式下大数据反欺诈成为了P2P平台提高风险控制水平的新思路。

大数据反欺诈，即是通过数据的采集和分析，找出欺诈者的蛛丝马迹，挖掘其数据的矛盾点和可疑点，识别和预防欺诈事件的发生。大数据收集了大量异构、多样化的信息，包括可交叉验证信息主体所提供的信息以及第三方信息来源的真实性，尤其是对于第三方信息来源，信息主体想要进行长时间、全方位的伪造，非常困难，成本较高，并且事实上经常不可行，因此大数据具有较强的反欺诈能力。

2、面对的挑战

大数据反欺诈技术又可以分解为两个子问题，第一个问题是在用户的授权下如何收集用户的相关数据，包括去哪里收集和收集哪些数据，为此我们对接了大量的第三方数据提供商的系统，还在用户的授权下，利用网络爬虫抓取公开的互联网数据，从而不断完善和丰富数据集，增加覆盖维度；第二个问题是如何整合和利用已收集的数据解决反欺诈问题，由于数据来源多，数据异构碎片化，结构(structure)、半结构(semi-structure)和无结构(adhoc)数据共存，并且规模庞大增长迅速，因此这一过程的挑战在于如何整合异构的数据源，如何有效的利用已有的数据进行交叉验证。

为了应对这个挑战，我们利用图(Graph)的数据结构，将不同渠道的碎片化、异构数据整合成为机器可以理解的知识，构建了知识图谱(Knowledge Graph)，借助规则引擎(Rule Engine)，实现了欺诈的识别与防御。

3、知识图谱的概念

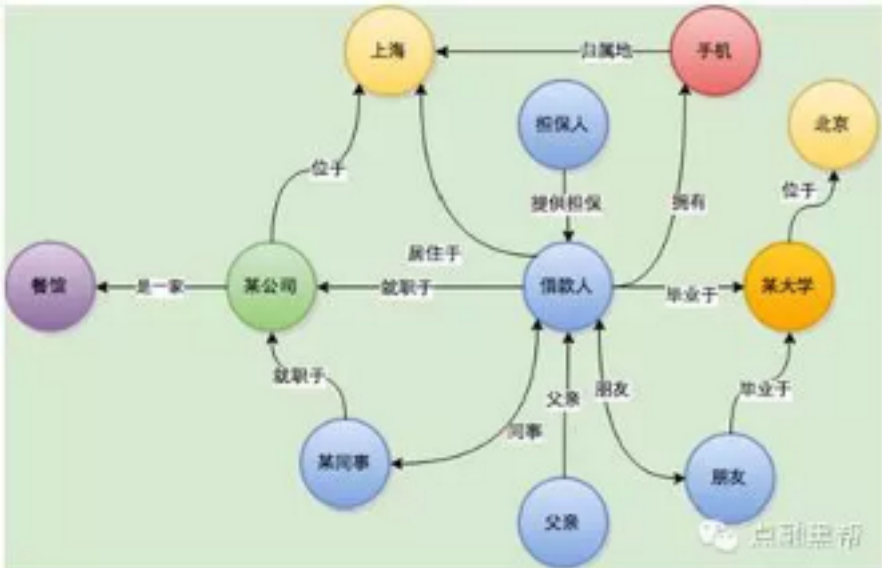
知识图谱是一种基于图的数据结构，其目的是将真实世界所存在的实体，知识以及概念等描述成机器可以理解的数据结构，将数据转化为知识；图的节点(Point)是真实世界所存在实体，由一个全剧唯一的ID来标识和索引，每个实体可以带有若干不同的属性(Property)，用来刻画实体的特性，而图的边(Edge)则用来描述两个实体的关系，例如is-a关系，表示一个实体是另一个实体的一种，或是has-a关系，表示一个实体具有另一个实体，这样的关系都是用来刻画实体之间的关联关系。知识图谱可以看作一个巨大的网络，是由数据绘制出来的一张知识图。

知识图谱最先由Google提出(<http://googleblog.blogspot.sg/2012/05/introducing-knowledge-graph-things-not.html>)，用于提升搜索引擎质量。举一个简单的例子，当我们用Google搜索“刘德华的老婆”时，Google返回了朱丽倩的信息，说明Google是理解了搜索框中的内容才进行的搜索，而不是简单的字符串检索，这就是一个知识图谱的应用场景。



4、知识图谱在反欺诈场景的应用

在反欺诈场景中，知识图谱聚合各类数据源，逐步绘制出借款人的profile，从而针对性质的识别欺诈风险。以一个借款人举例，借款人可以有身份证号，手机号，学历等个人信息，属于个人的属性信息；而借款人可以有担保人或是亲属好友，借款人与担保人之间的关系（也就是边Edge）是被担保与担保的关系，借款人与其亲属好友之间的关系是父亲、母亲、同事、同学等关系；借款人也具有住址，银行流水，工作单位等信息。这些信息可以来自于多个渠道，例如可以由借款人自己填写，或是积累的历史数据，或是数据提供商提供，或是在互联网上获得，甚至通过推理得到，往往具有冗余性；信息通过图的形式连结，展示出了借款人的profile。



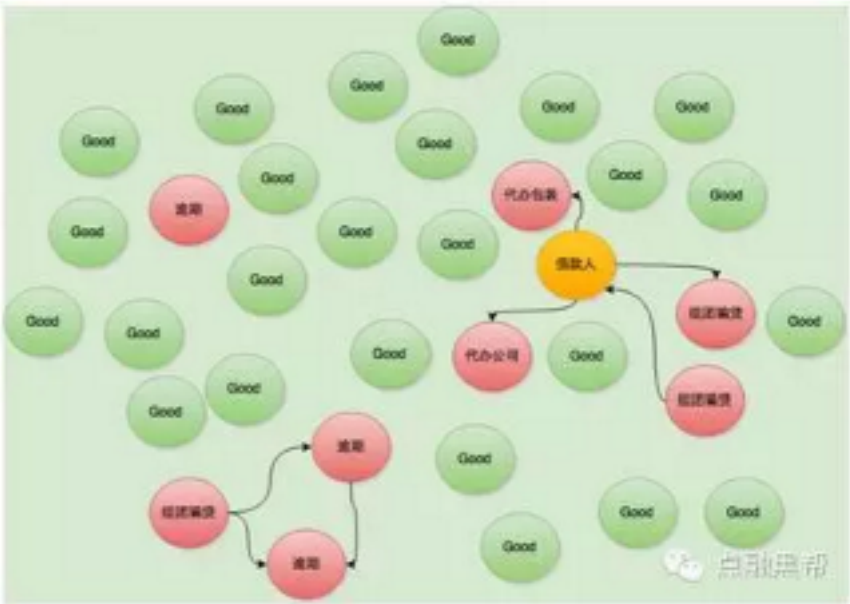
4.1、识别数据造假

当融合来自不同数据源的信息构成知识图谱时，有一些实体会同时属于两个互斥的类别（例如同时在两个不同的城市工作），或某个实体所对应的一个Property（同一个人的住址）对应多个值，这样就会出现不一致性，这个不一致性即可判定为潜在的可疑点。

通过这种不一致性检测，我们利用绘制出的知识图谱可以识别潜在的欺诈风险。在P2P行业，欺诈风险主要的骗术包括个人信息造假、工作单位虚假、代办包装、虚假联系人、组团骗贷等。以识别数据造假为例，利用知识图谱我们可以通过借款人的身份信息PII(Personal Identify Information)，例如手机号或是身份证号，直接索引到个人的全部信息，并以此与借款人的填写信息进行不一致性检测；也可以通过借款人的其他信息进行推理出其相关信息进行验证，举一个例子，我们可以通过借款人的身份证号和姓名可以获得他的学历信息和年龄，通过学历信息和年龄可以推算出其工作年限，再根据其所在城市，行业，职位，结合互联网上的招聘网站数据推理出其薪水范围，进而验证他的收入水平；甚至可以通过不同借款人之间的同事关系，验证其工作单位的真假。

4.2、组团欺诈和代办包装

除了对数据造假进行验证外，由于图结构带来的天然关联检索的特点，知识图谱可以识别潜在的代办包装或是组团骗贷。我们利用征信公司提供的欺诈数据，拥有的代办包装公司数据，互联网公开欺诈黑名单，行业黑名单联盟等数据开发了大量的标签数据，对实体（包括公司和人）贴上标签，例如逾期，虚假手机号，代办包装或是组团骗贷等标签，当借款人进行申请贷款时，如果我们发现他和bad people/company/info具有较多的关联关系，那么这个人有很大的可能是欺诈，从而识别出风险。




与搜索引擎的场景不同，知识图谱在反欺诈场景中具有较低的应用门槛，数据量较少时也可以进行低程度的交叉验证，而随着数据量的积累和增多，知识图谱也会越来越完善，其反欺诈能力也会越来越强。我们建立了名为“Matrix”的大数据反欺诈系统，在借款人提交借款申请开始即介入整个风控流程，对接多个数据源以获取借款人的数据信息，在各个环节建立checkpoint，通过可配置的规则引擎在各个checkpoint执行预定的逻辑，识别和防御欺诈风险。



结语

这篇文章介绍了点融网在大数据反欺诈领域的尝试与实践，比较系统的介绍了知识图谱技术在反欺诈领域的应用。知识图谱的构建离不开数据的积累，也需要知识库、自然语言理解、机器学习和数据挖掘等多方面知识的融合；知识图谱使得机器能够理解现实世界的实体和关系，正如Google所说，a “graph”—that understands real-world entities and their relationships to one another: things, not strings.

本文作者：程书欣（点融黑帮），现任点融网研发工程师，关注大数据风控技术，主导研发点融网反欺诈系统Matrix。



点融黑帮 (/u/dd56464a4e4d)

写了 737715 字，被 4901 人关注，获得了 3529 个喜欢
(/u/dd56464a4e4d)

+ 关注

点融黑帮——一个充满激情和梦想的技术团队，吸引了来自金融及信息科技领域的顶尖人才。我们正在用技...

小礼物走一走，来简书关注我

赞赏支持



(/u/11e5d581f623)

♡ 喜欢 (/sign_in?utm_source=desktop&utm_medium=not-signed-in-like-button)

| 21

更多分享

(http://cwb.assets.jianshu.io/notes/images/5948222)



下载简书 App ▶

随时随地发现和创作内容



(/apps/download?utm_source=nbc)



登录 (/sign_in?utm_source=desktop&utm_medium=not-signed-in-comment-form) 后发表评论


评论


^

智慧如你，不想发表一点想法 (/sign_in?utm_source=desktop&utm_medium=not-signed-in-nocomments-text)咩~


被以下专题收入，发现更多相似内容

- 

工具癖 (/c/2mvgxp?utm_source=desktop&utm_medium=notes-included-collection)
- 


@IT·互联网 (/c/V2CqjW?utm_source=desktop&utm_medium=notes-included-collection)
- 


今日看点 (/c/3sT4qY?utm_source=desktop&utm_medium=notes-included-collection)
- 

程序员 (/c/NEt52a?utm_source=desktop&utm_medium=notes-included-collection)
- 

亮书房 (/c/fd71d2f6495f?utm_source=desktop&utm_medium=notes-

included-collection)


 知识图谱学习 (/c/7ee9079aaeed?utm_source=desktop&utm_medium=notes-included-collection)

 数据科学 (/c/102149797c26?utm_source=desktop&utm_medium=notes-included-collection)

展开更多 ▾

【刘知远】知识图谱——机器大脑中的知识库 (/p/8bf571284747?utm_cam...

作者：刘知远（清华大学）；整理：林颖（RPI）本文来自Big Data Intelligence知识就是力量。——[英]弗
兰西斯·培根1 什么是知识图谱在互联网时代，搜索引擎是人们在在线获取信息和知识的重要工具。当用户输入..


 墨白找 (/u/9ee413023cf1?utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

(/p/cdbed82a34bc?




utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendation)
[3/4]我所经历的大数据平台发展史（三）：互联网时代·上篇 (/p/cdbed82a...

//我所经历的大数据平台发展史（三）：互联网时代·上篇http://www.infoq.com/cn/articles/the-development-history-of-big-data-platform-paet02 编者按：本文是松子（李博源）的大数据平台发展史...

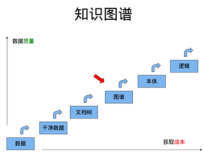
 葡萄喃喃呓语 (/u/2c67926c48ce?utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

浙江大学译美国白宫”大数据“白皮书 (/p/0457b6046c1c?utm_campaign=m...

大数据：抓住机遇、保存价值 美国总统行政办公室浙江大学历史数据2014 年 5 月 大数据：抓住机遇、保存
价值“即使大数据技术重塑了我们周围的世界，今天的发言也将帮助我们持续贯彻自身的价值观念。”这份评...


 Albert陈凯 (/u/185a3c553fc6?utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

(/p/bd15e0f50eb9?



utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendation)
知识图谱技术解剖 (/p/bd15e0f50eb9?utm_campaign=maleskine&utm_c...

本体、知识库、知识图谱、知识图谱识别之间的关系？ 本体：领域术语集合。 知识库：知识集合。 知识图
谱：图状具有关联性的知识集合。 知识图谱本质上是语义网络，是一种基于图的数据结构，由节点(Point)...


 方弟 (/u/1ffaf4faed6?utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

(/p/6cba1710d6f3?



utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendation)
大数据金融反欺诈将一步步破碎羊毛党的黄粱美梦 (/p/6cba1710d6f3?utm...


今年两会上，总理在《政府工作报告》中指出，当前系统性风险总体可控，但对不良资产、证券违约、影子
银行、互联网金融等积累风险要高度警惕。互联网金融风险连续四年被写进政府工作报告中，可见风控对于..

 刘旷 (/u/599bcbf0439f?utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

(/p/d5527bb50f5a?



utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendation)
水中文理 (/p/d5527bb50f5a?utm_campaign=maleskine&utm_content=n...

 陌_e63b (/u/62144fc9d088?


utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

(/p/ab78c78fa131?



utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendation)
根植意念，托举生命 (/p/ab78c78fa131?utm_campaign=maleskine&utm_...


大讲堂第一天早上课程中，居老师让30多个孩子上台，给每个孩子都注入能量，托举生命。其中有年仅六岁的孩子徐亮。当居老师问孩子，你是谁时？徐亮因为有了辅导基础，响亮地回答：“我是标准男...

 陌上花开_c14e (/u/6de27a9cd9ca?

utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

精度更高的随机数生成函数 rand_s (/p/5706d5ccf605?utm_campaign=ma...

(简书处女写) 此方法只适用于WINDOWS下 rand() 结合 srand() 函数可以有效地获取随机数序列 大多数情况下已经够用，然而srand()产生的seed每秒更新一次。考虑到频发调用随机数生成函数，存在1秒内需要不同...

 RainING1947 (/u/acafed8003b3?


utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

(/p/423721ee47ac?



utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendation)
技术书看不懂怎么办 (/p/423721ee47ac?utm_campaign=maleskine&utm_...

作为平时主要写业务逻辑的程序开发，工作中很少直接与线上服务器接触，上生产环境的服务器一般也就是用vim调试程序，因此对于运维方面的知识懂得很少。很多运维知识也只是听说过概念罢了。近来，由于工...

 钟森龙 (/u/f9338eda7dda?


utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

(/p/36a19ae65a91?



utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendation)
沈阳远大装备科技成功研发“负刚度隔振技术” (/p/36a19ae65a91?utm_cam...

近日，由沈阳远大装备科技有限公司研发的“负刚度隔振技术”即将亮相2017第三届“中国国际舰船技术与装备展览会暨海军建设论坛”。该项技术为高能密度准零刚度隔振器提供了一个极其紧凑且承载能力大，高效可...

 乖乖s (/u/c3266f83a1d4?

utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

