12 智能信息检索与挖掘

新加坡国立大学蔡达成教授：展望未来——多模态会话搜索的机遇和挑战

整理：智源社区 陈佳

在 2020 年 6 月 23 日上午的“2020 北京智源大会 | 智能信息检索与挖掘专题论坛”中，来自新加坡国立大学的蔡达成教授做了关于多模态会话搜索相关研究的介绍。

蔡达成，新加坡国立大学计算机学院创院院长、KITHCT 讲席教授、清华大学—新加坡国立大学下一代搜索技术研究中心主任。他是国际知名的计算机科学与技术专家，在多媒体与信息检索领域享有盛誉，是国际计算机学会多媒体专委会（ACM SIGMM）杰出技术贡献奖获得者，也先后担任包括 MM 和 SIGIR 在内的多个国际顶级学术会议的大会主席，先后发表多篇国际顶级会议与期刊论文，获得 MM，SIGIR，ICDM，MMM 等高水平国际会议的最佳论文奖。

以下是智源社区编辑整理的蔡达成演讲要点:

关于信息检索的研究开始于 20 世纪 50 年代，在大约 60 年代末 70 年代初的时期，最受欢迎的模型是向量空间模型以及 TF–IDF 模型等等，这其中有很多的模型直到今天仍然被广泛使用。另一个比较有代表性的工作是在 1998 年提出的 PageRank 算法，不仅仅是在信息检索领域，该算法在其他的领域也大放异彩。到了 2013 年，大家开始把注意力集中到词向量表示的评估上，诞生了像 Wordvec 这样非常有影响力的工作。而现在，我们开始展望未来，必须要探讨一下经典的 IR 算法中的一些局限性。首先，是单方向查询模式（即只有用户可以向系统提交查询），它假设用户提交的查询是精确的，且系统可以理解的用户意图等等，但是实际上并不一定是如此。如上所述，我们需要考虑多方的对话因素来帮助用户提升搜索体验并帮助系统能够更好的理解用户； 其

次，是目前对查询以及各种信息进行建模的局限性，绝大部分系统主要是利用文本信息（Text-based）去挖掘用户的意图。然而，近年来随着智能手机的普及，很多人会开始在移动端输入其他形式的查询，例如图片。因此，利用多模态查询来作为检索系统的输入会在不久的将来成为一种常态； 第三点是关于用户查询意图的不确定性和流动性。根据大量系统的反馈数据，用户的意图往往会随着搜索过程产生一定的转移。在这样的情况下，一个搜索系统也必须期待着用户随时变化的意图。

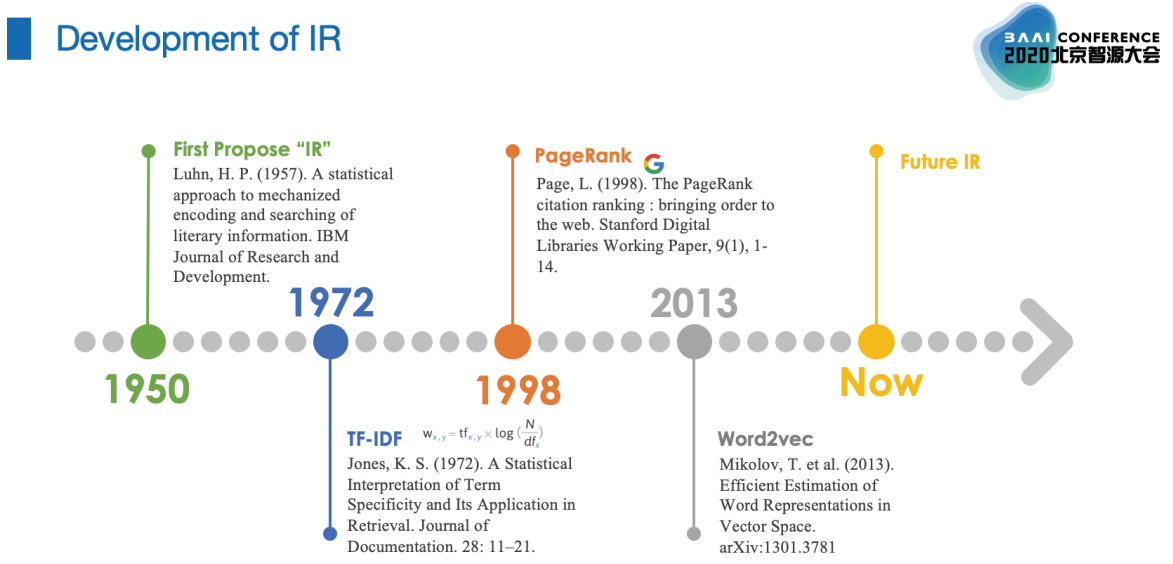


图 1：信息检索的发展

一、多模态会话搜索介绍

信息检索的发展趋势，应该是和新兴的技术息息相关的。首先是多模态处理技术（Multimodel Processing）。新的多模态模型可以处理更为丰富的信息，例如图片、视频和音频数据等等。除此之外，用户画像、历史信息也渐渐被融入到模型中，从而能够使得模型更准确地理解用户意图。另一个领域是对话系统（Dialogue System），对话系统主要通过和用户进行交互来达到它的目的。因此，现在的一个主要趋势应该是如何从文本过渡到多模态信息，例如，构建多模态对话系统（Multimodel Dialogue System）、多模态推荐系统（Multimodel Recommendation System）等等。另一个趋势则是，如何从单向（Unidirectional）的查询转变为交互式（Interactive）的查询，例如会话推荐（Conversational Recommendation）以及会话结构式知识库搜索（Conversational Structured Knowledge Base Search）等等。

这里要强调一下会话搜索（Conversational Search）和对话系统（Dialogue System）的区别。二者之间的差别并不大，但是有一些关键的区别。例如对话系统有以下特点：1）目的是与用户在宽泛的主题下谈话，2）可能包含搜索式或者非搜索式的对话。而对于会话搜索来说，往往包含比较明确的目标，即用户在会话中通过修改查询来明确自己的搜索意图。但是二者有一些需要共同关注的点，包括： 怎样去实时地理解用户的意图，如何去追踪用户的对话状态并对历史信息进行进模，如何学习好的策略去干预用户并引导做用户喜欢的事情以及如何进行人机协调，等等。

举一个关于多轮会话推荐系统的例子。会话开始，用户对智能体说“我想要一个新的手机”，接着智能体问用户

“你想要什么样的操作系统?”用户回答“iOS”。这里智能体会问用户一些关于意图的属性，使得它们可以更好地预测用户想要什么，尽可能让用户能留在当前对话系统，而不是觉得无聊就走了，因此推荐系统也会尽可能快地进行推荐。有的时候，系统的推荐可能会失败，比如例子中的用户认为 iPhone 11 的价格过于昂贵，因此拒绝了系统的推荐。但是这是没关系的，因为大部分用户都会继续使用系统，那么接下来系统的回复需要权衡用户是否知道足够的信息并在适时的时候（比如用户接受了某个属性）进行推荐。

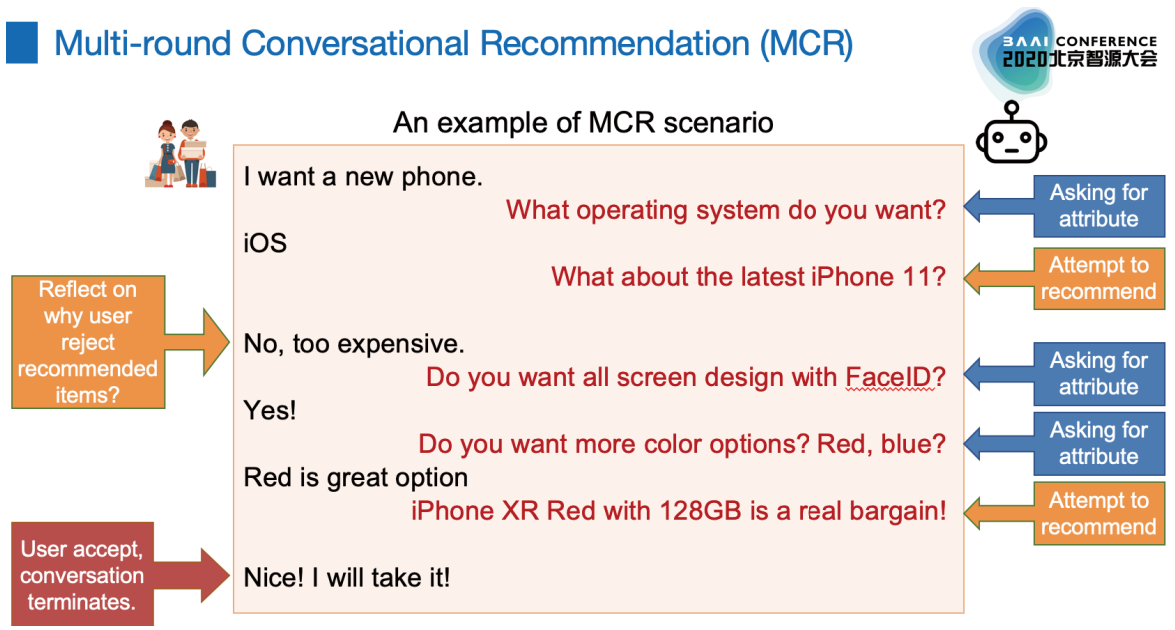


图 2：多轮会话推荐系统举例

二、多轮会话推荐系统实例

所以系统实际上是由两个关键组件构成的，首先用户对系统提出包含意图要求的查询，随之系统需要采取一系列的措施，去决策询问一些属性或者推荐一些商品。决策的过程是多因素影响的，我们将会在下文进行阐述。给定系统推荐的属性，用户将需要作出回应，比如拒绝属性或者表达对该属性的喜爱度。如果用户接受这个推荐，那么很大概率他将会结束本次对话，但实际上结束对话一般需要比较多轮的互动。总的来说，会话一般有两种结果，一个是用户比较积极，接受了系统的推荐并结束本次会话，另一个则是用户比较消极，可能就直接中止了会话。因此我们想要优化的是，如何询问用户正确的问题，去吸引用户停留在我们的推荐系统上。主要的研究问题包括：①推荐什么样的商品，②询问用户什么属性，以及③决定去询问还是推荐的策略。那么我们的主要目标就是：在尽可能短的互动轮次中给用户进行成功的推荐。

Workflow of MCR

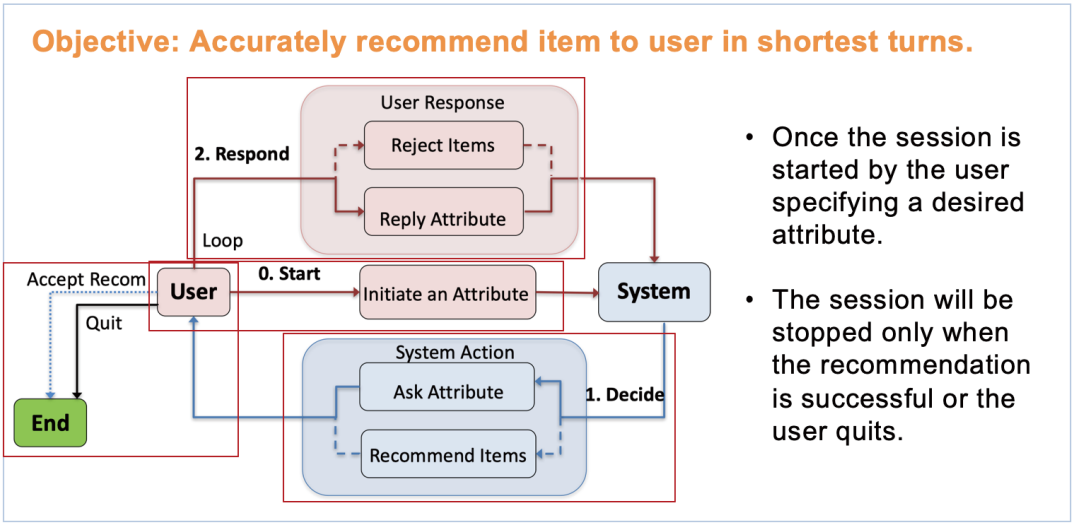


图 3：多轮会话推荐工作流

这里来介绍一个我们组发表在 WSDM 2020 上的一篇工作，其主要由一个推荐模块和一个会话模块构成。整个模型分为三个阶段，分别是：估计阶段 (Estimation Stage)、动作阶段 (Action Stage) 以及反馈阶段 (Reflection Stage)，接下来将逐一介绍各个阶段。

System Overview – EAR

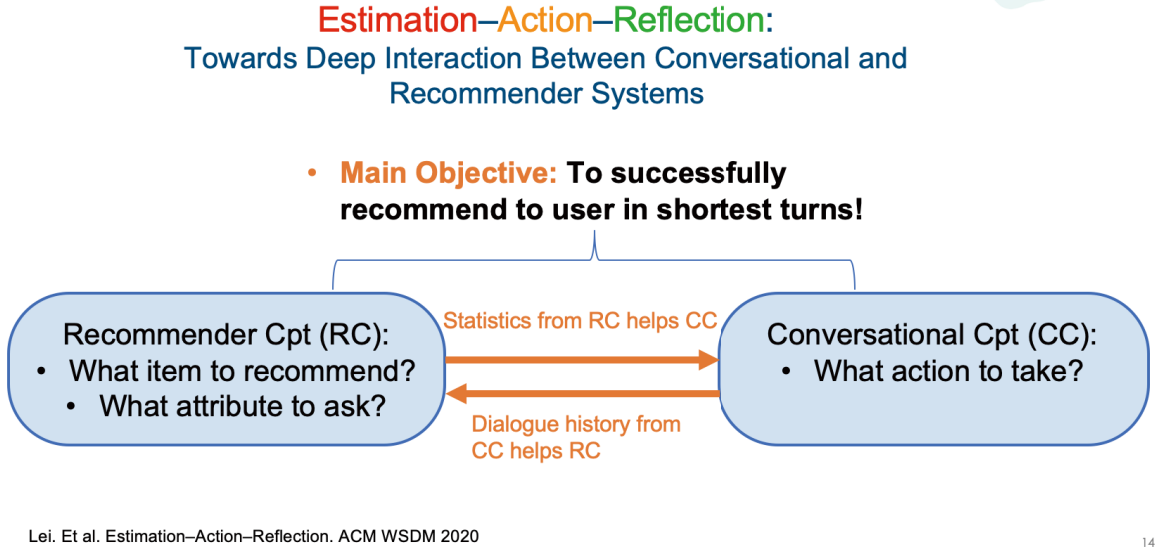


图 4：EAC 模型

首先是一个学习的过程，命名为估计阶段（Estimation Stage），我们需要对推荐的商品或者询问的属性进行预测，即①满足给定的属性我们应该推荐哪些商品；②以及给定用户已经确认的属性我们下一个应该询问用户什么属性。这里我们设计了一个属性相关的因式分解机，去同时对商品和属性的排序进行优化。

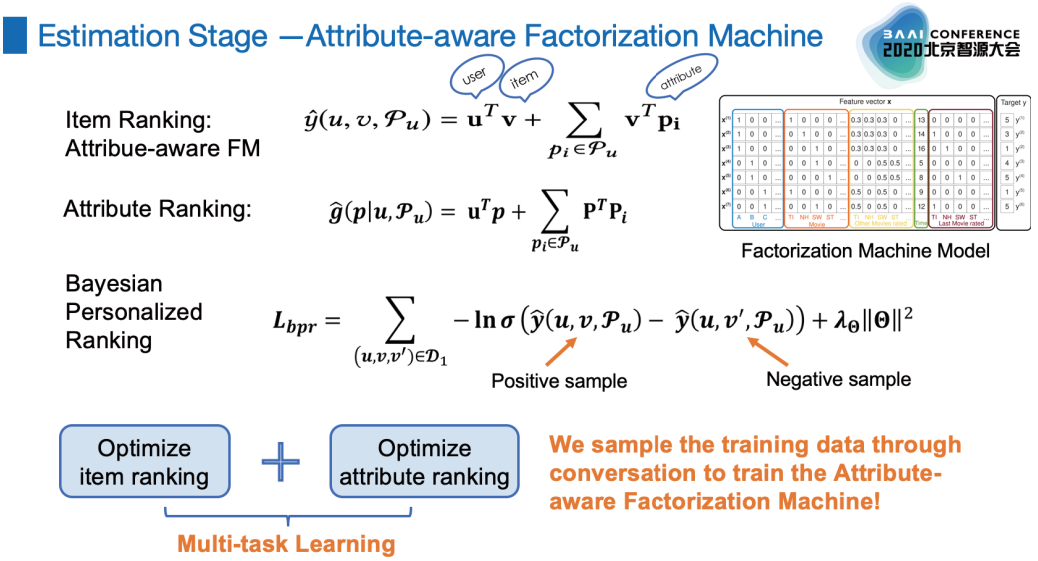


图 5：估计阶段——属性相关的因式分解机

接下来是动作阶段（Action Stage），系统需要维护一个策略去决定当前应该是询问用户某个属性还是给用户进行推荐。这里我们利用强化学习 Policy Gradient 算法去进行建模，维护了一个 2 层的前馈神经网络。其中回报函数由四部分组成，分别是：①当推荐成功时获得一个较大的正回报；②当系统成功询用户一个属性时获得一个较小的正回报；③当推荐失败时获得一个较大的负回报；④当会话持续过长时获得一个较小的负回报。

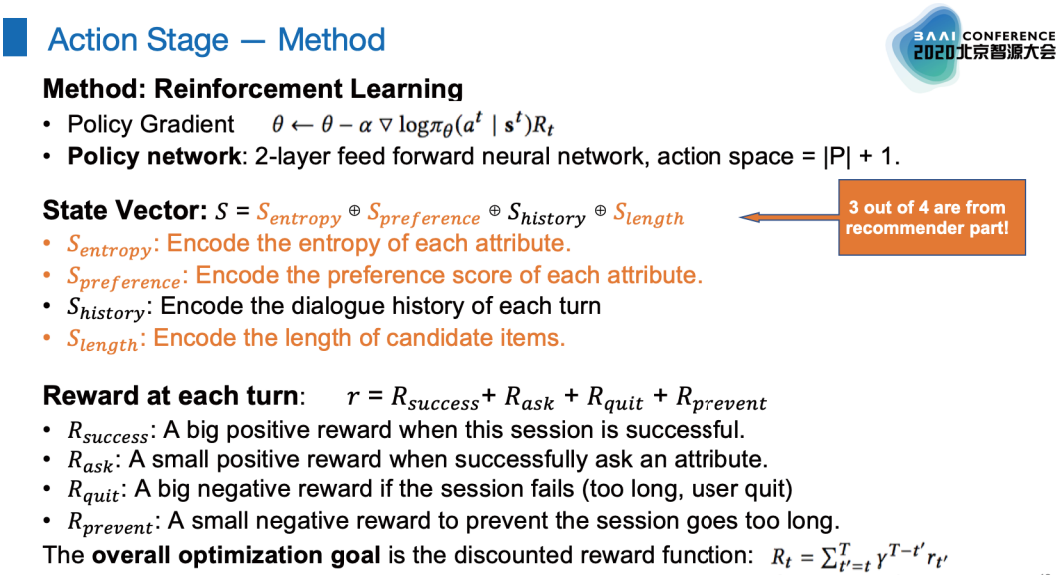


图 6：动作阶段——强化学习

最后一个阶段就是反馈阶段（Reflection Stage），主要是利用用户的在线反馈去更新对用户的建模。一个简单而有效的方法就是将最近被用户拒绝的商品作为负例加入到我们的推荐系统中，对用户的估计进行更新。

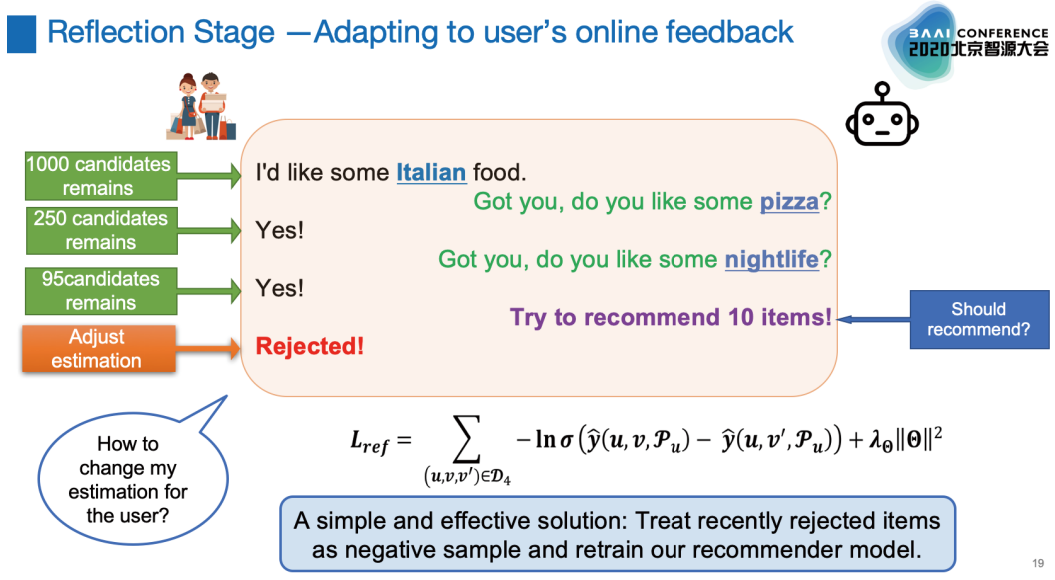


图 7：反馈阶段——强化学习

在实验设置方面，我们使用了真实场景下用户和商品的互动作为正例，构建了用户模拟器。在会话开始的时候，模拟用户会把目标商品“记住”，然后在接下来的对话中根据这个商品给智能体的询问做出反馈。当智能体给用户推荐一个商品时，用户需要根据之前“记住”的商品目标去检验推荐的商品是否满足他的需求； 而当智能体向用户询问一个属性是，用户需要根据“记住”的商品是否包含这个属性进行回应。

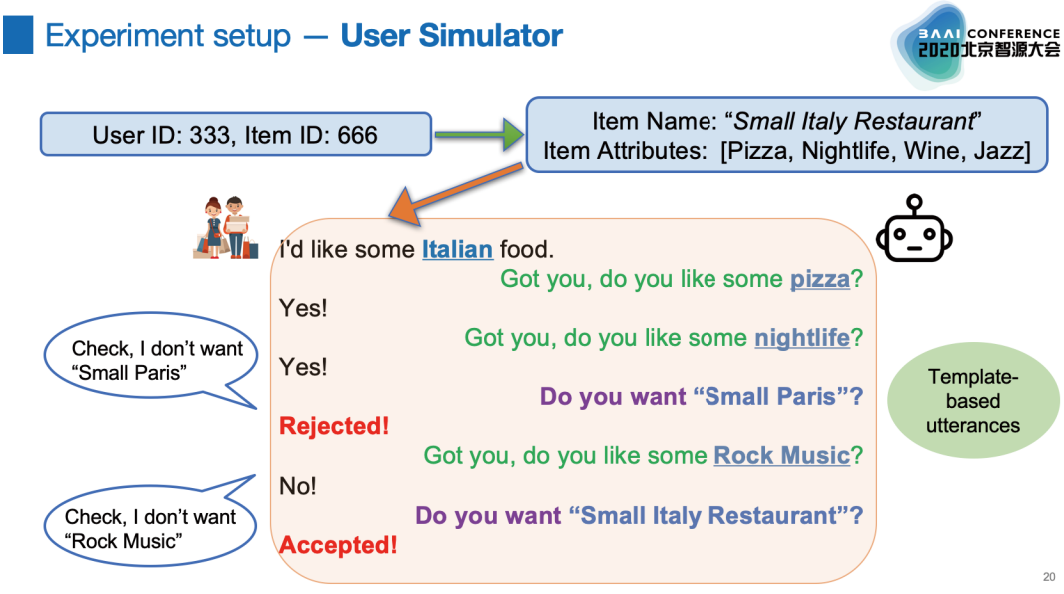


图 8：模拟用户的交互过程

我们采取的评价指标包括 SR@k（直到第 k 轮任务的成功率）以及 AT（平均轮次）。根据实验结果，我们发现 EAR 模型在 Yelp 和 LastFM 两个数据集上都取得了比强基线模型 CRM 更高的 SR@k 值。

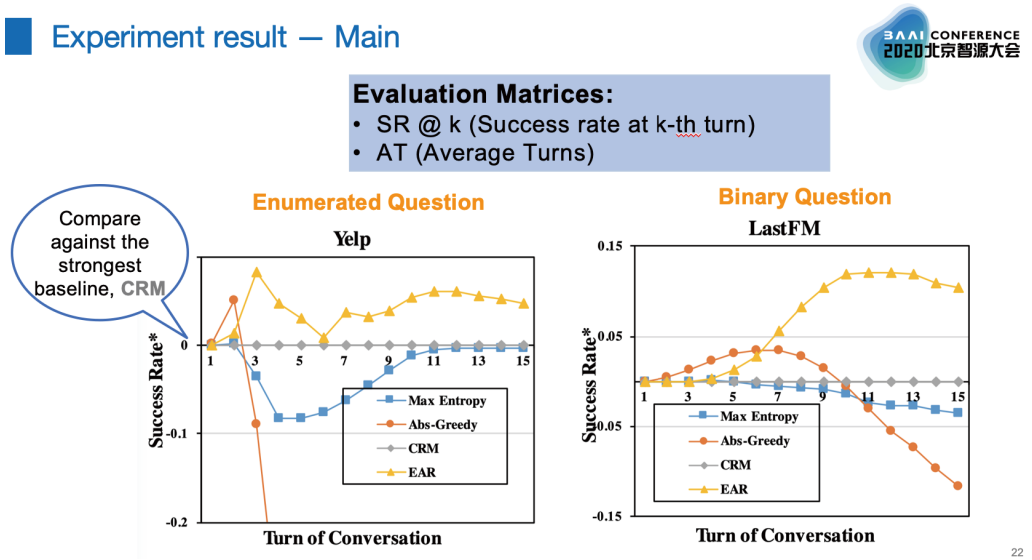


图 9：EAR 模型效果

EAR 模型在会话推荐系统上比已有的模型具有更好的性能，但是它也有一些局限性，比如动作空间太大、忽略了用户 – 商品属性的一些结构化信息。为此，我们提出了一个基于图上路径推理的模型——CPR，它可以利用图上路径的一些限制来更好地进行推理和学习。首先，当用户提交一个查询时，他会有一些特定的查询需求 (Requirements)，这些需求很大程度上是基于一些属性的。因此，我们可以将这些属性中需求传播到其他的商品（图 10 中的黄色线段），对于这个过程我们使用了和之前介绍的 EAR 模型中相同的因式分解机进行建模，经过该步骤我们可以更新和当前的用户状态最符合的商品信息。

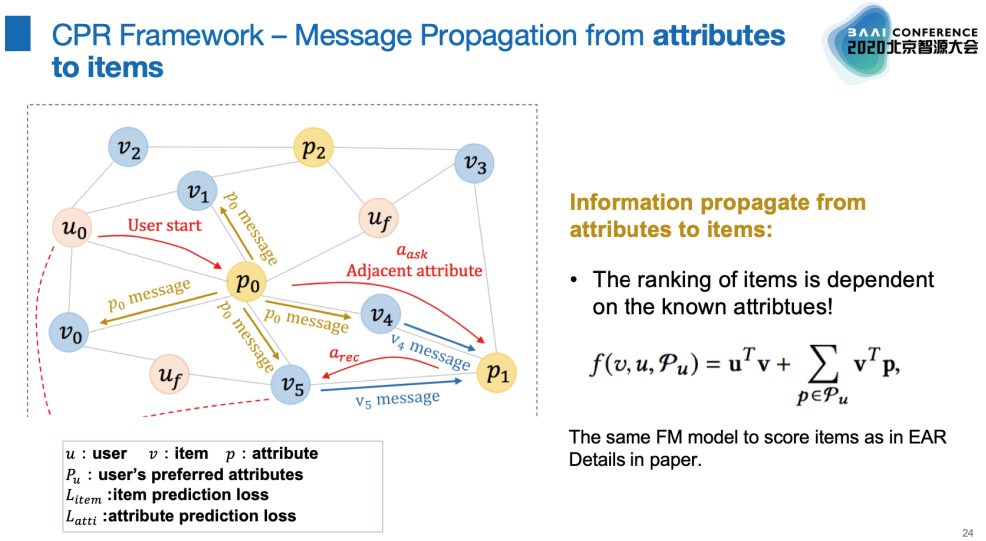


图 10：CPR 模型中更新最相关商品

由于智能体接下来决定向用户询问的属性是和剩余候选商品相关的，因此我们还需要将信息从商品向属性进行传递。这里我们利用了剩余候选商品集合的属性熵来表示和当前用户状态下各个属性分数。本过程详见图 11，其中信息传播为蓝色线段。

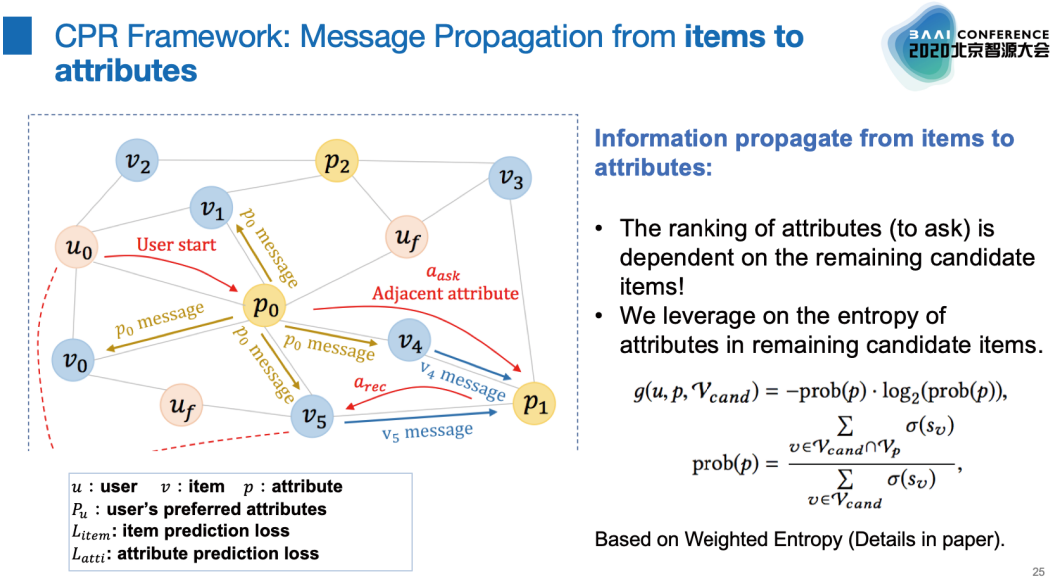


图 11：CPR 模型中更新最相关属性

在已知最相关属性和商品之后，系统需要决定何时进行推荐，何时去询问用户属性。这里我们使用了策略网络 (Policy Network，见图 12) 进行建模，网络结构和 EAR 中的较为相似，但是和之前不同的是，这里的的动作空间降低为 2。从实验结果来看，CPR 模型在会话推荐任务上比其他模型取得了更好的效果，包括 EAR 模型，说明了该模型的有效性。

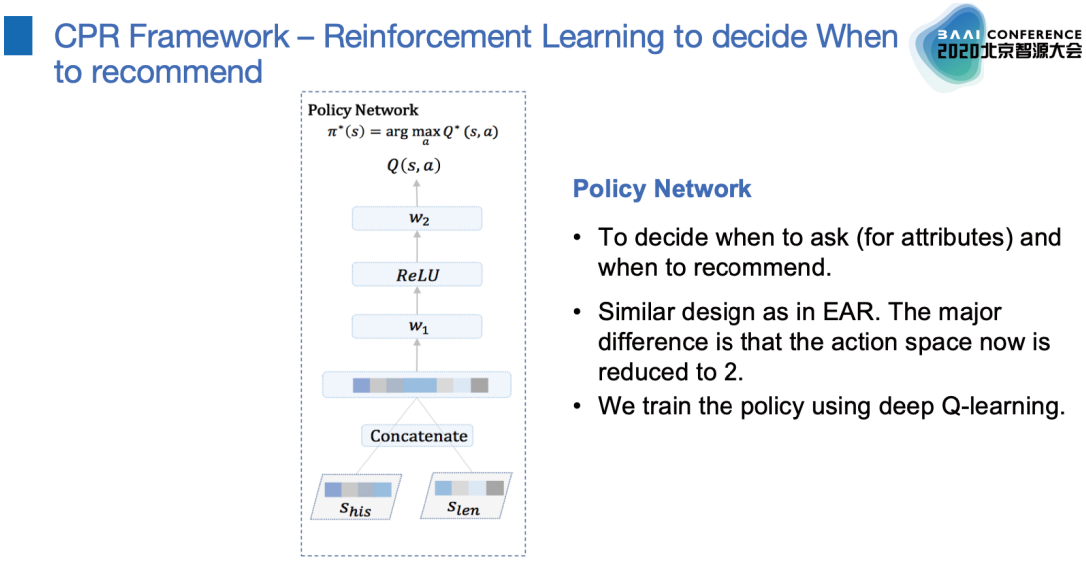


图 12：CPR 模型的策略网络

四、会话搜索当下面临的挑战

关于会话搜索和对话系统，我们还面临一些挑战。包括如何去对多模态上下文和历史进行建模，如何融入领域知识以及用户模型，制定系统的交互策略等等。关于多模态的上下文信息，我们可以更多地利用一些用户信息例如地理信息、地域偏好以及画像信息等等，也可以去使用一些搜索上下文包括搜索历史、搜索结果质量等数据。推荐系统会对这些信息进行隐式的收集，但是有一个问题就是怎样去避免这些反馈信息的偏向性（Bias）。在多模式会话搜索中，除了对上下文进行建模外，更重要的问题是对会话历史进行恰当的建模。我们在这里举一个例子。为了找到正确的位置，我们需要一个从会话历史到结构化需求的良好映射。我们可以借助对话状态跟踪器来处理会话历史记录，对话状态跟踪的研究可以从口语对话系统开始。我们在这里举一个例子。给定一个对话历史，对话状态追踪的目的是将它解析成结构化的槽 – 值对（Slot–value pair）。现有的对话状态跟踪工作大多依赖于一个领域本体（Ontology），它定义了一组时隙和候选值。在这种情况下，对话状态跟踪被视为一个分类任务。不同的特征（例如手工提取的特征、语义特征、神经特征等等）都被利用起来。我们可以看到基于规则的模型、生成模型和判别模型，它们通常表现得更好。近年来，随着大型数据集的面世，研究人员开始在缺乏全面的领域本体的情况下执行对话状态跟踪（DST），并通过从对话历史或知识源生成单词来处理未知的时隙值。通常情况下，对话历史被作为编码器的输入，然后系统为每个特定的插槽生成一个值。在这里，Seq2seq 模型被广泛应用，复制机制也被证实是有效的。还有的工作将其视为一个机器阅读任务，给定一个值槽，系统从对话框历史中提取相应的值。

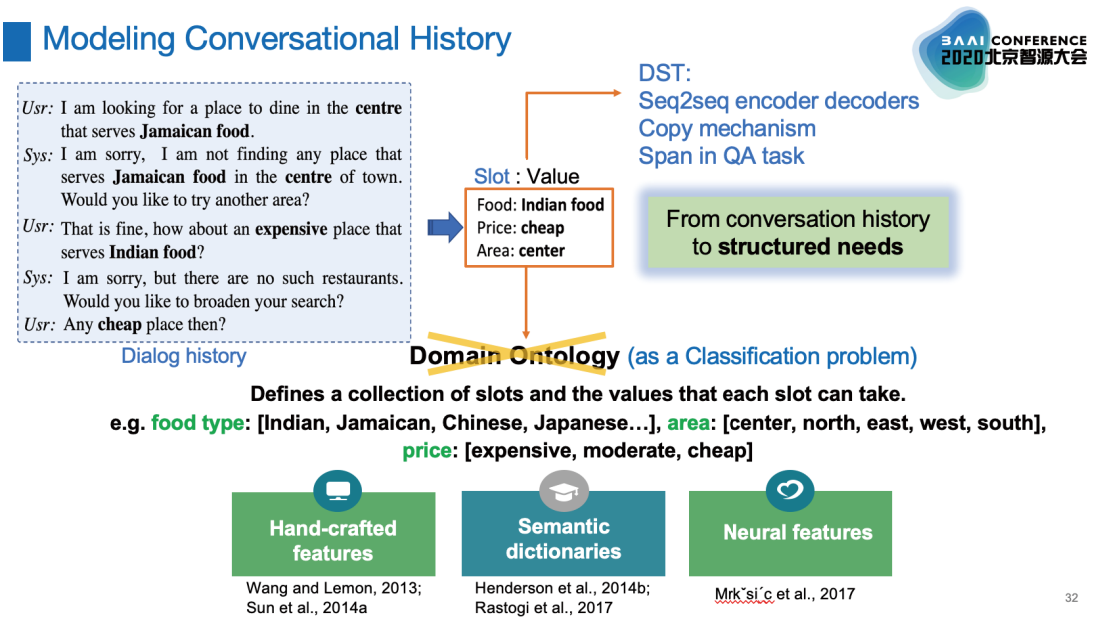


图 13：会话历史建模

就多模态而言，需要处理的问题更多。跟踪多模式对话状态可以与示例类似（见图 15），即给定一个用户的话语，系统将其解析为结构化的表示并给出相应的响应（生成四幅图像），接着用户给出反馈。在这种情况下，为了生成准确的状态，系统需要正确理解图像的语义、这些图像的引用并识别插槽，然后推断出正确的槽值。系统除了要对用户话语进行表示之外，还要考虑视觉和文本信息之间的异质性，以及细粒度的实体识别等等复杂问题。

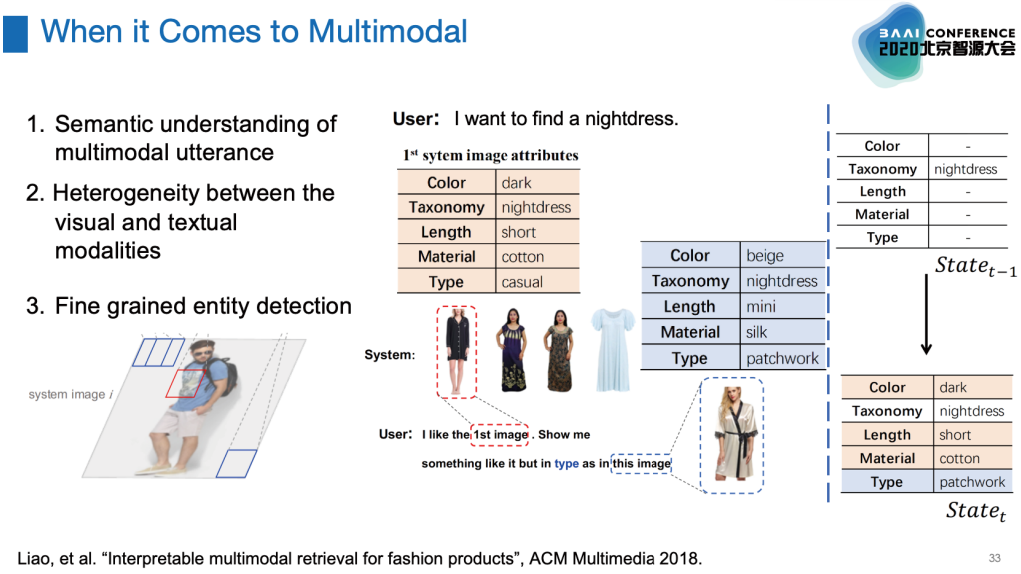


图 14：多模态交互示例

下一个挑战是融入领域知识。多模态会话搜索系统智能化的另一个大问题在于知识。例如，在时尚会话搜索场景下，为了正确地回答用户的问题（比如，用户对系统的推荐的红色裙子作出反馈“有没有类似的蓝色裙子”），我们需要了解人类对属性和关系的感知以及关于匹配风格的知识。一般我们可以采用多模态知识记忆网络来整合知识，这里可以举一个例子（见图 15）。在本例中，当用户查询关于蓝色溜冰服的匹配提示时，匹配的候选对象（如银色细高跟鞋）可能不会在会话上下文甚至整个训练语料库中与之同时出现（Co-occur）。因此，我们使用知识三元组来丰富系统，构建了一个领域知识库，其中包括由领域专家制定的 300K 以上个三元组，然后使用 EI 树模型提取特征并存储在内存网络中。当接收查询时，它根据输入查询检索合适的知识并给其添加权重。

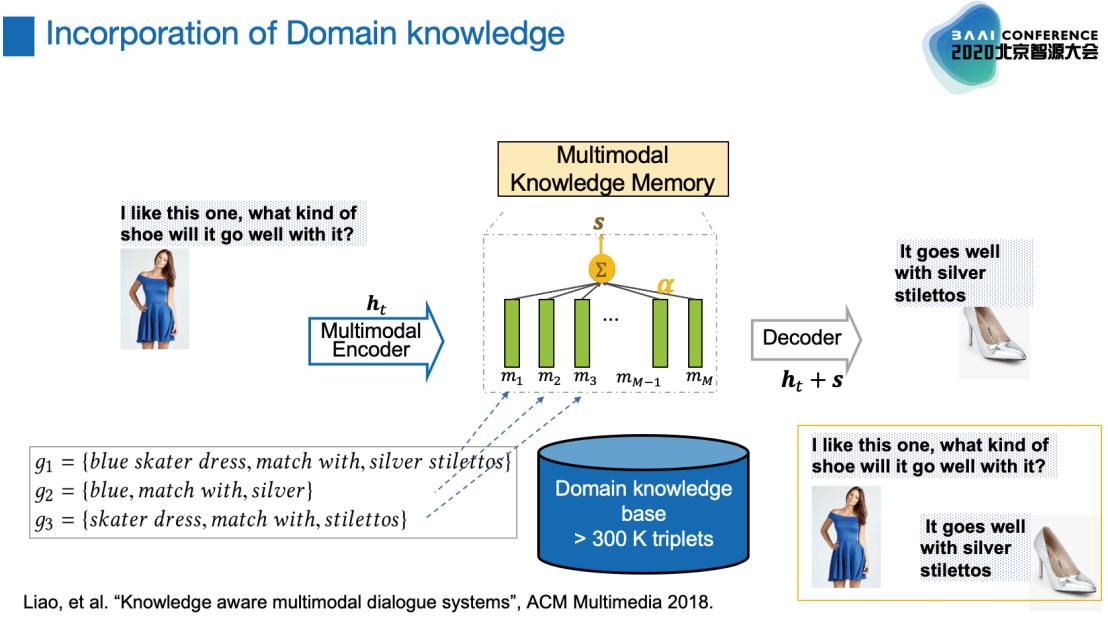


图 15：融入领域知识

第三个挑战是学习更好的交互策略，系统需要学会何时进行推荐、何时询问用户，又被称为问题生成 QG (Question Generation)。可以从三个方面进行概念化，分别是输入 (Input)、焦点 (Focus) 以及感知层次 (Cognitive Level)。①首先是输入的形式，不仅包括文本，还包括图像、知识库以及在会话搜索设置下对用户目标的更新理解。在此基础上，核心问题将是决定“何时问”、“问什么”和“如何问”。传统的 QG 主要侧重于文本输入，可以由问答系统进行建模。而到了最近，关于 QG 的研究还扩大了来源范围，包括知识库和图像。②第二个部分是焦点，是系统去询问用户的策略，包括系统何时去询问以及询问的内容。在对话的初始阶段，用户可能不清楚自己的意图，更偏向于浏览搜索结果，那么这个时候系统可能会更多地进行询问。为了学习好的系统策略，我们需要对于用户的状态进行追踪，并做出一些决策或者干预。关于询问内容，系统需要指出一些对于用户来说重要的属性和方面，并使用自然语言去进行表达。这里可以采取一些方法，例如： 基于规则的方法 (Rule-based Methods，包括 Transformation-based, Template-based)、基于神经网络的方法 (Neural-based Methods) 等等。③最后是感知层次，目前 QG 正逐渐从浅层次往深层过渡。其中浅层次的 QG 一般考虑单一句子、不需要推理以及先验知识，使用语义转换这样的方法就可以取得比较好的效果。但是，在更加复杂的场景下，我们需要考虑深层次的 QG 问题，包括多跳推理 (Multi-hop Reasoning) 以及人类提出的问题 (Human-raised Question)。对于多跳推理，需要根据上下文的多个句子进行建模，并且使用多条信息进行推理，但是不需要先验知识。最后是处理人类问题，需要使用所有的用户输入并进行多跳推理和常识推理，还需要已知先验知识。未来可以考虑在一些深层次 QG 问题上进行更多的探究。

Learning to Ask or Learning to Intervene

Question generation can be conceptualized in three aspects: inputs, focus, and cognitive level.

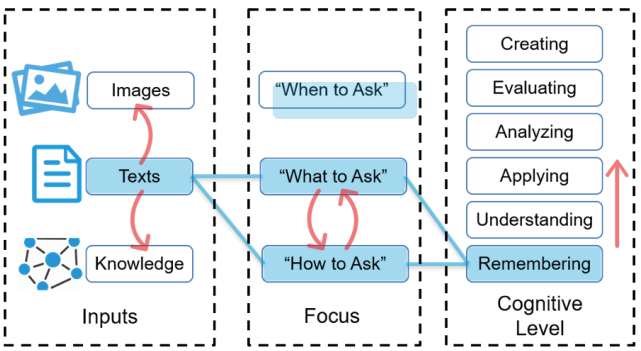


图 16：对话策略学习涉及的三个方面

第四个挑战是关于评估与数据集。目前来说，对于多跳对话系统研究来说，最大的瓶颈其实还是在数据集以及评估方式。现阶段最主要的评估方式有两个，一个是构造用户模拟器 (User Simulator)，另一个是使用多回合会话数据集。其中，用户模拟器需要满足一定的标准，包括①鲁棒性： 需要在各个场景下工作良好；②多样性： 需要覆盖各种用户类型和需求；③覆盖度： 需要尽可能包括各种可能的用户例子；④拟人化： 需要生成自然语言。构造用户模拟器，可以使用基于规则的方法和基于历史的方法。关于多回合会话数据集，图 17 列举了一部分。我们可以观察到三种趋势： 从纯文本模态向交叉模态的转变、从单个域扩展到同时处理多个域、着重于对任务型对话系统中搜索与推荐场景的研究。然而，现有的涉及多模态模型以及会话搜索的数据集缺乏真实的交

互场景，也无法处理不同的任务。目前仍然迫切需要一个会话式的搜索数据集，它可以利用多模式信息，处理跨域任务、建模用户配置文件或偏好，并提供知识库或后端数据库。综上所述，我们需要一个相对全面的多模态会话研究环境，支持不同的对话任务。

Training Resources: Existing Conversational Datasets

Datasets	# Dialogs	# Utters	Types	Domains	User Profile	Modality
CMU DoG (Zhou et al., 2018a)	4K	130K	Chitchat	Movie	No	text
IIT DoG (Moghe et al., 2018)	9K	90K	Chitchat	Movie	No	text
PERSONA-CHAT (Zhang et al., 2018)	10K	162K	Chitchat	Persona chat	Yes	text
Wizard-of-wiki (Dinan et al., 2019)	22K	202K	Chitchat	1365 topics from Wikipedia	No	text
OpenDialKG (Moon et al., 2019)	3K	38K	Chitchat	Sports, music	No	text
KdConv (Zhou et al., 2020)	4.5K	86K	Chitchat	Movie, music, travel	No	text
Facebook Rec (Dodge et al., 2016)	1M	6M	Rec.	Movie	No	text
REDIAL (Li et al., 2018)	10K	163K	Rec.	Movie	No	text
GoRecDial (Kang et al., 2019)	9K	170K	Rec.	Movie	Yes	text
OpenDialKG (Moon et al., 2019)	12K	143K	Rec.	Movie, Book	No	text
DuRecDial (Liu et al., 2020)	10.2K	156K	Rec., chitchat, QA	Movie, music, food, new, etc.	No	text
DSTC2 (Henderson et al. 2014)	1.6K	23K	Restaurant search	Restaurant	No	text
FRAMES (Asri et al., 2017)	1K	20K	Constrained search	Flight, hotel, budget	No	text
KVRET (Eric et al., 2017)	3K	14K	Info search, navigation	In-car assistant	No	text
MultiWOZ (Budzianowski et al., 2018)	8K	115K	Venue search & tasks	Hotel, restaurant, attraction etc.	No	text
VisDial (Das et al., 2017)	123K	2.4M	Image-grounded QAs	Topic constrained by image	No	multimodal
GuessWhat (Vries et al., 2017)	155K	1.6M	Image-grounded QAs	Topic constrained by image	No	multimodal
IGC (Mostafazadeh et al., 2017)	4K	25K	Image-grounded QAs	Topic constrained by image	No	multimodal
MMD (Saha et al., 2017)	150K	6M	Fashion search	Fashion	No	multimodal

Search & recommendation

cross domains

cross modality

图 17：目前可用的会话数据集

第五个挑战是将会话搜索扩展到其他的内容搜索，例如结构化知识库搜索。现有结构化知识库搜索的最大缺点是数据库搜索和自然语言查询之间的不对称性，这种不对称性导致了不完全性（Incompletion）和模糊性（Ambiguity）两个问题。不完全问题意味着用户的初始查询可能不完整。模糊性问题是指用户的话语中可能存在一些不准确、模糊的描述。图 18 说明了这两个问题。给定目标 SQL 和数据库方案，我们使用两个用户查询来解释这两个问题。首先是不完全问题。在第一个查询中，用户忽略了“保险”一词，使得很难识别保险和相关内容。二是模糊性问题。在第二个查询中，“name”可以识别为“full name”或“short name”。然而，将会话与结构化知识库搜索相结合还存在一些挑战，主要是会话策略的问题。首先，我们需要有效地找到最不确定的部分，要求用户进行确认，不确定性估计的性能决定了我们交互的效率。其次，我们必须设计一个友好的对话协议来与用户交互。这个问题必须是人类用户可以理解和回答的，而不是只能由模拟智能体来回答。

Asymmetric Problem in Database Search via Natural Language

- Incompletion: User's query cannot always include everything
- Ambiguity: User may use inaccurate description

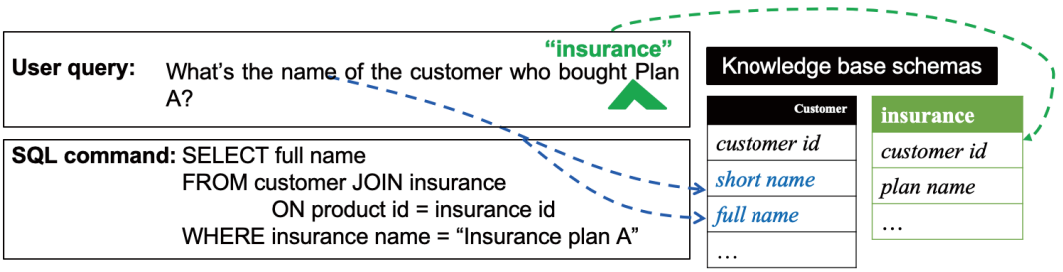


Figure 8. Illustration of the incompletion and ambiguity issues

图 18：结构化知识库搜索举例

第六个挑战是对模型中的偏差进行建模。搜索中常见的三种偏差，分别是位置偏差（Position Bias）、流行性偏差（Popularity Bias）和点击诱饵偏差（Clickbait Bias）。一般来说，搜索或排序模型是从用户的隐式反馈中学习的，比如用户点击数据，因为它们很容易被大量收集，而要收集用户的明确反馈是非常昂贵和耗时的。虽然点击数据很廉价，但它们通常受到许多因素的影响。一个主要原因是，只有部分商品被展现给用户，并且商品之间并不是被公平地呈现。一个解决办法是使用逆偏好加权（Inverse Propensity Weight，IPM），即对于那些相关但是不受欢迎的商品赋予较高的权重。

五、总结

信息检索已经从单向查询和基于文本的方式发展到交互式和多模态的形式。很多系统假设用户已知如何去简洁有效地进行搜索，但大多数时候，用户是不确定的并且他们的需求在不断变化，因此我们需要通过会话搜索来弥补这种不对称的差距。所以实现会话搜索的挑战包括：让用户在没有压力的情况下初始地自由浏览、在用户需要的时候提供帮助以缩短搜索过程。会话搜索可以认为是用户浏览、搜索和对话的无缝连接综合体，在这里我们提到了关于会话搜索的六大挑战。目前搜索、对话和推荐之间的界限正在打破，因此为了解决复杂的搜索问题，我们需要综合考虑更多的因素。

- In this talk, I presented 6 key challenges:
  - Modelling multimodal context and history
  - Integrating domain knowledge and user models
  - Interaction strategies
  - Extension to other Content Search
  - Evaluation and Resources
  - Modeling Biasness in Search
  - Plus many others
- The boundaries between search, conversation and recommendation are breaking down. We need to look at the combination of all to tackle the complex search problems

图 19：会话搜索的六大关键挑战

问答环节

文继荣：谢谢蔡老师，我知道时间给您留的有点短。尤其是最后一页未来我们如果要做很好的方向有很多工作要做，这个领域还是有很大的空间去发挥的。

我的听的过程中间也看见了很多非常有意思的点，第一个问题我先问一下蔡老师，您刚才提到你们做了一个多模态会话数据集？这个是公开吗？

蔡达成：快要结束了，还没有真正做完，主要的重点就是用真正的用户模拟去生成问的还有回答的，就找不一样背景的用户，而且给他不同的任务，它会用不同的方式来问问题，回答的人根据这个直接回答，我们希望通过这个能够找出同样的问题有很多不同的设置和问法。

文继荣：所以工作量很大。

蔡达成：如果你们需要的话电子邮件给我，我们将来肯定会很乐意的分享给大家，因为这个领域需要一个好的引领，不然很难再走下去。

文继荣：我这个问题是帮听众问的，不光是做搜索，做推荐，做对话，有很多的数据，但是现在也没办法就是这种情况，一个领域要往前快速稳定的发展还真的要有这样的工作，今天最大的收获就是您做的团队开始做这个事情了，如果能分享给大家还是有极大的帮助。

蔡达成：对，我觉得评估是很重要的问题，现在大型的评估一定要完成自动的。模拟用户也很重要，很多人都在做，现在的做法是模拟器要对准 task，所以变成比较单元化。如果真正用的话其实用户会改变他的注意力，还没有人真正讨论这个问题，第一个演讲的裴老师讲了，用户可能喜欢意大利餐，看了之后改变主意喜欢中餐，怎么去处理这样的问题？这个也是很重要的。

文继荣：对，因为这个问题其实是做搜索也好推荐也好，如果真正做这个领域就知道这可能是最重要的问题。

今天上午的老师都讲到这个问题，如果没有一个很好的评估方法的话，很多方法是比较难以评价的。刘兵老师的问题也是一样。

蔡达成：刘兵老师是更 open 的。

文继荣：像这种怎么去做。

蔡达成：都很难做的。

文继荣：非常难做，您这个也很难。但是我觉得这些东西反正始终得有人做，做出来不是完美的，但是还是会往前推进一大步，跟整个 IR 的发展史上标注了很多数据，但是它至少使得大家有一个公认的东西往前进步。

蔡达成：这个我觉得很重要，不然这些数据要发展要很难。很多时候很多好的概念都是在工业界提了很久了才开始接受，你们大团队应该想一下这种方式，让全世界能够更开始的解决一些工业界想要的问题，主要还是数据集的问题。

文继荣：因为我以前也在微软，工业界确实有很多优势，比如说它直接上线以后做 A/B Test，有间接的评价你这个方法的好坏。这个也是我们将来要跟它们更多合作的一个地方。

蔡老师我最后问您一个问题，因为您今天讲的两个都非常难的问题，如果说它们俩碰撞在一起以后带来的最大的挑战是什么？因为它们各自单独的发展都在往前走，但是这两个在一起了。刚才说的数据集就是一个大的问题，还有在其它的方面也是一样。

蔡达成：如果我们分开做难题更大，因为每个领域多媒体内容理解，每个领域都要做，如果两个配合起来其实有更好的可能，有一些历史信息。把两个方向综合在一起，我觉得从领域的发展是很重要的想法，是一个好的方向。

裴健：搜索皆智能，智能皆搜索

转载自：AI 科技评论

作者：陈大鑫

6 月 23 日，加拿大西门菲莎大学教授裴健在第二届北京智源大会智能信息检索与挖掘专题论坛上做了《 智能搜索：从工具到思维方式和心智》的报告。

裴健，是加拿大皇家科学院和加拿大工程院的两院院士。裴老师是国际著名的数据科学、数据挖掘和数据管理专家，专长于通过数据战略制定、数据资产管理、数据资源整合和数据产品设计研发把数据和技术转化为业务能力和效益。他同时是多家企业的顾问，提供高端战略咨询和技术咨询服务。其论著被引用九万七千多次。

裴健在这次的演讲中提出了三个核心观点：

第一，搜索皆智能，搜索以人为核心，以满足人的信息需求为目的，所以它天然就包含了智能成分。

第二，智能皆搜索，我们要做到智能必须要用到搜索的方法，目前人工智能的很多应用都是搜索任务，智能和搜索同行。

第三，智能搜索不仅是一个单纯的技术问题，更是一个与人相关的问题，我们必须要一起努力，使得每个人都不会被落下，让智能搜索服务全人类。

在演讲最后，中国人民大学教授、智源首席科学家文继荣与裴健老师进行了精彩的问答互动：

智能搜索和智能推荐可能比我们想象中更深刻地影响到我们每天的生活，比如有一个问题，你的第一反应是不是去搜一下？或者说你想获取什么信息，你会第一时间打开如头条、微博、知乎这样的一些 APP，然后去看它给你推荐了一些什么？

做搜索、推荐、数据分析的责任是非常重大的，如果这方面做得不好，在极端情况下就有可能改变我们下一代甚至改变人类的思维方式，改变我们对世界的看法，因为一个人对整个世界的看法更多地是由他接收到的信息、他的经历所塑造的。如果我们的信息推送和用户检索到的信息是有问题的，比如提到的信息是有偏见的，比如**我看什么就给我推荐什么，那我就进入了信息减法的世界，我可能会失去了解这个世界的更多可能性**。

通过这次精彩的演讲和问答互动，我们可以从智能推荐或者个性化推荐等技术中看到一些人文关怀和哲学反思。

人文关怀： 老人会不会因为不会用智能手机、不会用电脑而享受不了智能搜索带来的红利？比如说残疾人和在偏远地区、经济不发达地区的人会不会因为达不到智能搜索的入门门槛而被慢慢抛弃？我们应该如何解决这些问题？

哲学反思： 随着我们越来越依靠智能搜索、个性化推荐，我们是否会失去了解这个世界的更多可能性？我们是否会失去一部分原有的“自由意志”？究竟是我们驯化了这个信息流世界还是被其驯化？

以下是**裴健演讲正文**:

今天我报告的题目是智能搜索：从技术工具到思维心智。首先，让我们来简单回顾一下搜索的基本概念。在搜索当中，我们假定用户有信息需求。用户的信息需求往往不能直接被搜索系统直接理解，于是用户把信息需求转化为搜索系统的查询。搜索系统得到用户的查询，找到相应的结果，可能是一些文档、图片、图像或者是生

成的内容，返回给用户。用户可以根据这些是否是所需要的，产生相应的反馈，搜索系统根据用户的反馈来决定是否需要去对搜索进行增强。这样一个过程不断循环，直到用户信息需求得到了满足，整个搜索过程就结束了。这个过程听起来非常得完美，很简洁。但在实际当中，搜索并不是那么简单，要比这个复杂得多。

图 1：检索的需求分类

一、搜索皆智能

在实际生活当中，“用户信息需求是固定的”这个假设命题其实是个伪命题。在很多情况下，用户的信息需求不断变化。更麻烦的是，用户本身可能并不清楚自己的信息需求到底是什么。举个例子来说，比如我听说某个小区有新冠肺炎的新感染案例，发出一个“新冠肺炎感染病例”的查询，那么这个查询到底是想问什么呢？用户自己可能并不清楚，在很多时候用户可能是发出一个查询先问一下，看搜索引擎给返回什么样的信息。用户和搜索引擎的交互过程就是一个探索的过程，用户的信息需求在不断变化。在“新冠肺炎感染病例”的例子里，用户可能想问的是这个感染病例是不是得到了治疗？感染病例的具体情况是怎样的？看到搜索引擎的回答后，用户可能马上想到这个感染案例对小区的生活，如出行、购物等，有什么影响？大家可以看到信息需求是不断变化的，我们在搜索过程中不能假定用户的信息需求是不变的。信息系统必须想办法去理解用户的真实信息需求，为用户提供探索的工具。因此，**搜索本身从一开始就是智能的，因为它把人摆在了整个过程的中心**。

图 2：搜索皆智能

下面举个例子来讲一下搜索过程为什么是一个探索的过程。在 VLDB–2019 的会议上，我的研究小组发表了一篇社团搜索的文章。和很多已有的社团搜索工作不太一样，我们假定在每一个网络节点上都有一个数据库。如果这个网络结点是一个人，那么这个数据库就可以是这个人以往购买东西的整个历史。如果这个网络结点是一个论文作者，那这个数据库就是他以前发表的所有论文的集合。我们关心在这样一个网络里面怎样找到社团?

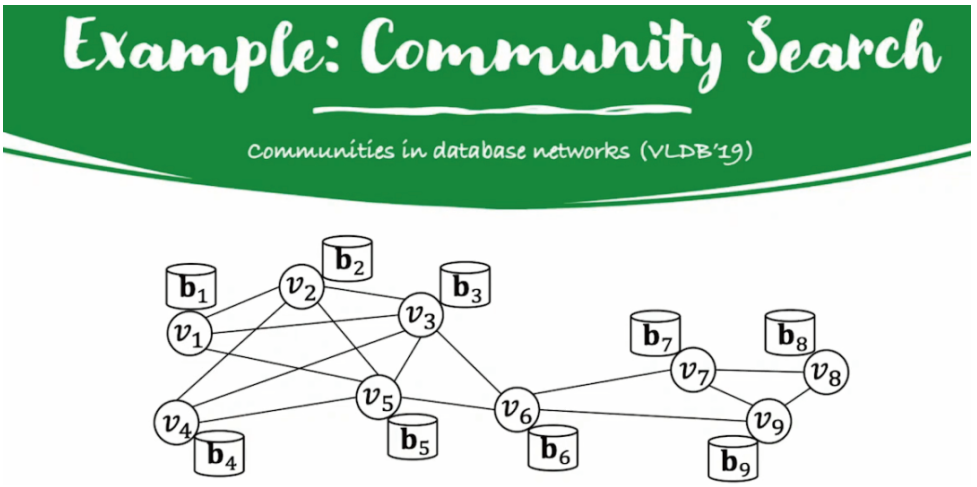


图 3：社团搜索

我们希望社团内成员之间不但有很密切的网络连接关系，还能在数据库上有非常相似的模式。 这样社团有什么用呢？举例来说，在论文作者的数据网络上，用户可能关心的是能不能找到那些用数据挖掘方法来研究人脸识别和图象检索的社团。我们的搜索首先形成了一个查询模式 a1。

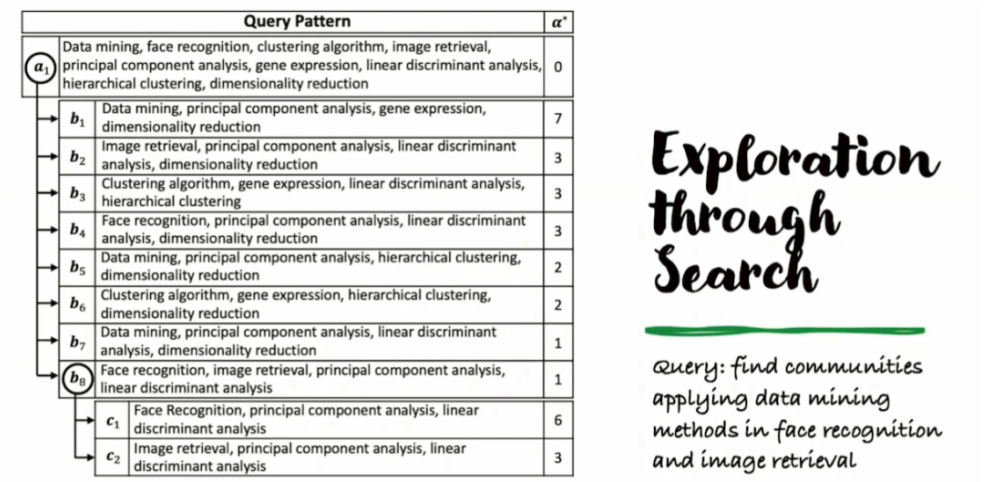


图 4：查询模式 a1 示意图

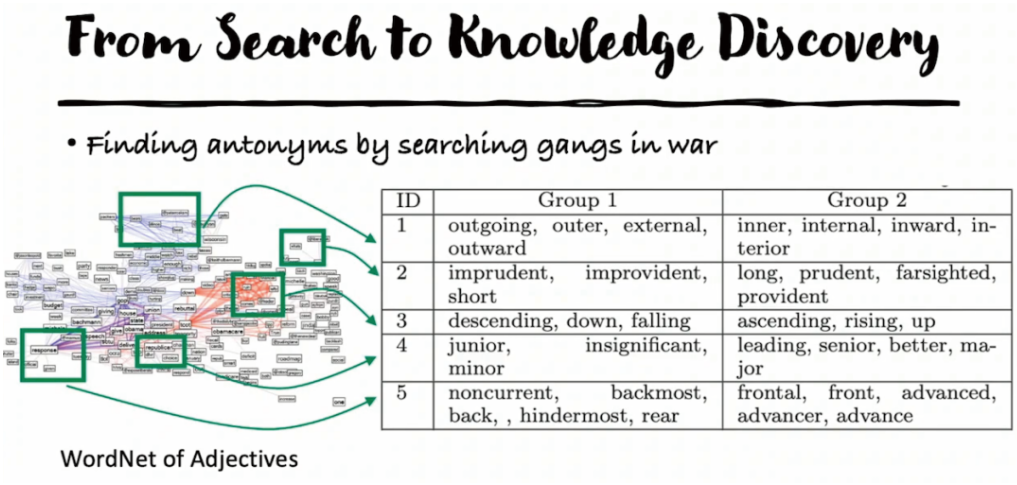
同时，我们的搜索算法还能够提供针对 a1 的各种细化，比如 b1、b2、直到 b8。在这些细化当中我们会专门看各个具体的分支，包括算法具体分支和问题具体分支。这些分支给用户带来探索方向和探索方便。这种探索可以进一步往下走。比如说 b8 可以进一步探索到 c1、c2 两种具体的情况。整个过程是一个不断深入、不断尝试、不断修正的探索过程。

二、智能皆搜索

搜索皆智能，搜索要用到大量的人工智能技术，所以我们要通过人工智能技术去理解用户的信息需求。同时，智能很复杂，智能的每一个任务都需要多多少少用到搜索技术。什么是智能？智能是关于连接的，我们需要把不同的数据、不同的知识点连接起来； 智能是关于推理的，我们需要对数据、对知识进行相应的推理； 智能是关于泛化的，我们有具体的观察，我们希望通过若干具体的观察、具体的例子来泛化来概括成通用的规律； 智能还需要去做具体化，我们有一些通用的原则，要把它用到具体的事例里面，提高具体事例处理的效率和效果。所有这些都需要搜索相应的数据，搜索相应的知识，搜索相应的连接。所以智能皆搜索，智能离不开搜索，智能必须通过搜索来实现。

图 5：智能皆搜索

下面举个一个我们 KDD 2016 论文中的例子来介绍我们怎么通过搜索来达到知识发现。我们可以在 WordNet 的网络上找到很有意思的一些社团，每一个社团内部非常相似，社团成员之间有很强的关联，同时，社团之间非常对立，有非常大的差异。



ID	Group 1	Group 2
1	outgoing, outer, external, outward	inner, internal, inward, interior
2	imprudent, improvident, short	long, prudent, farsighted, provident
3	descending, down, falling	ascending, rising, up
4	junior, insignificant, minor	leading, senior, better, major
5	noncurrent, backmost, back, , hindermost, rear	frontal, front, advanced, advancer, advance

图 6：从检索到知识发现

当我们把这个问题用在形容词网络里面，我们就找到了大家在胶片上看到的 Group1 和 Group2 这样对立的社团，所以我们把它叫做 gangs in war。大家仔细看，每一个社团内部是一组同义组，Group1 和 Group2 之间是反义词关系。我们用智能搜索带给我们新知识，我们可以在词的网络上自动发现同义词和反义词。

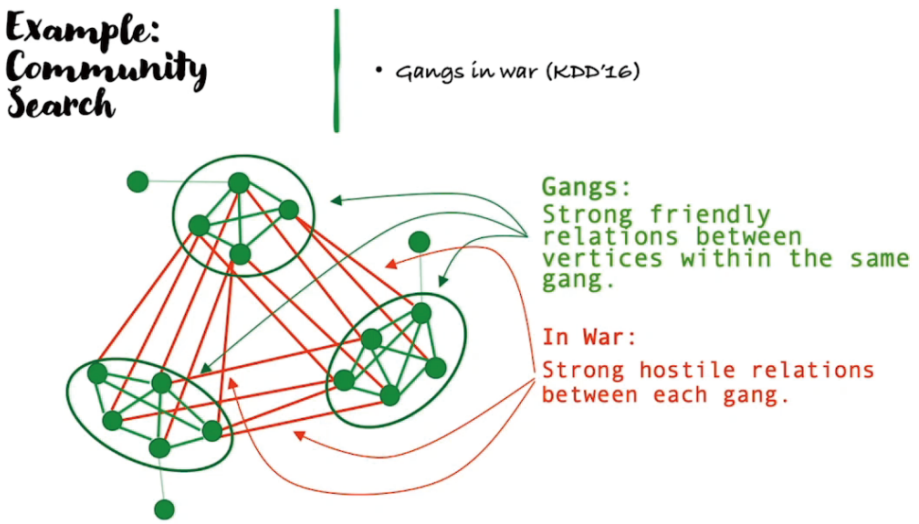


图 7：社团探索示意

刚才我们讲了搜索皆智能，智能皆搜索，所以智能和搜索是密不可分的，二者紧密结合，搜索和智能同行。这里包括两个意思，第一，我们需要用搜索的技术来达到更好的人工智能。像我刚才举的例子，我们可以通过很好的搜索来自动的发现知识，同时我们需要用很多的智能技术和计算来使得搜索更加有效。这里的智能不单单只是人工智能，还包括了很多真正的人的智能，因为我们最终的搜索是为人服务的。

图 8：搜索与智能同行

这里举一个例子，这是我们最近刚刚完成的一个论文，我们研究的是基于 Web-scale 的多语言问答系统。问答系统有很多，在很多商用的搜索引擎里面都有相应的问答功能。当一个用户给出一个问题，例如说想知道感

冒症状，搜索引擎可以总结出像下图左边的信息卡，这个信息卡上会列出相应的感冒的症状甚至是治疗的方法。这给用户带来了很多的便利，在一定程度上这也是对知识的抽取和总结。

Type	Type	behavior
Explicit	*Click*	*Up-vote/Down-vote*
Implicit	*Re-query*	*Reformulation*
	Click	*Answer Click* *Answer Expansion Click* *Outside Answer Click* *Related Click*
	Browsing	*Browse*

Name	Type	Description
RFRate	*Re-query*	rate of re-query
AnswerCTR	*Click*	CTR of answer
AnswerOnlyCTR	*Click*	CTR with only click on answer
AnswerSatCTR	*Click*	satisfied CTR of answer
AnswerExpRate	*Click*	CTR of answer expansion
OTAnswerCTR	*Click*	CTR outside of answer
OTAnswerOnlyCTR	*Click*	CTR with only click outside of answer
OTAnswerSatCTR	*Click*	satisfied CTR outside of answer
BothClickCTR	*Click*	CTR of both click on/outside of answer
RelatedClickRate	*Click*	CTR of related queries
NoClickRate	*Browsing*	no click rate
AbandonRate	*Browsing*	abandonment rate
AvgSourcePage-DwellTime	*Browsing*	average source page dwell time
AvgSERPDwellTime	*Browsing*	average SERP dwell time

Web-scale Multi-lingual QA System

KDD 2020

图 9：Web-scale Multi-lingual AQ System

当搜索引擎给出这样一个答案时，这个答案是否满足了用户的信息需求？这个答案的知识是否正确？是否有用？我们希望能够得到用户的反馈，我们希望用户用人的智能来帮助机器进行学习。这里有一个挑战。很多情况下，用户看过答案但并不一定给出一个显式的反馈，理解人的反馈是一个非常复杂的过程。在这篇论文里面，我们系统地研究了如何观察、推理用户对搜索引擎所给出的问答信息的反馈，如何对用户的行为进行挖掘，抽取相应的反馈信号，用这些信号来改进我们的 QA 系统。

Intelligent Search Serving People

Table 5: Performance comparison between our methods and baselines on the DeepQA dataset. All ACC and AUC metrics in the table are in percentage , where the sign % are omitted.

Model	Method	Pre-training Data Size	Performance on Different Fine-tuning Data Size (AUC/ACC) 5k	10k	20k	30k
BiLSTM	Original	-	60.45/58.21	61.30/59.92	61.55/61.99	62.40/61.74
	$FBQA_{ctr}$	0.5m	59.90/57.60 (-0.55/-0.61)	61.25/58.25 (-0.05/-1.67)	61.40/60.50 (-0.15/-1.49)	60.65/59.29 (-1.75/-2.45)
		1.0m	60.25/58.45 (-0.20/-0.24)	61.35/58.12 (+0.05/-1.80)	62.65/57.43 (+1.10/-4.56)	61.35/60.69 (-1.05/-1.05)
		4.0m	60.50/56.99 (+0.05/-1.22)	59.75/58.39 (-1.55/-1.53)	60.90/59.15 (-0.65/-2.84)	62.25/61.73 (-0.15/-0.01)
	$FBQA_{FA}$	0.5m	61.95/59.66 (+1.50/+1.45)	62.50/60.96 (+1.20/+1.04)	62.85/62.74 (+1.30/+0.75)	64.23/62.50 (+1.83/+0.76)
		1.0m	62.80/60.44 (+2.35/+2.23)	63.20/61.20 (+1.90/+1.28)	63.45/63.00 (+1.90/+1.01)	65.57/63.05 (+3.17/+1.31)
		4.0m	**64.13/62.15 (+3.68/+3.94)**	**65.45/63.33 (+4.15/+3.41)**	**65.46/64.17 (+3.91/+2.18)**	**67.35/64.35 (+4.95/+2.61)**
BERT	Original	-	69.31/64.86	71.81/67.76	72.47/67.07	75.28/68.26
	$FBQA_{ctr}$	0.5m	67.35/62.76 (-1.96/-2.10)	72.96/66.66 (+1.15/-1.10)	75.11/68.26 (+2.64/+1.19)	77.76/71.07 (+2.48/+2.81)
		1.0m	72.33/67.06 (+3.02/+2.20)	73.76/67.36 (+1.95/-0.40)	76.16/69.16 (+3.69/+2.09)	77.42/68.26 (+2.14/+0.00)
		4.0m	72.19/65.66 (+2.88/+2.90)	73.92/67.96 (+2.11/+0.20)	76.81/67.96 (+4.34/+0.89)	77.94/69.36 (+2.66/+1.10)
	$FBQA_{FA}$	0.5m	72.26/65.27 (+2.95/+0.41)	76.03/68.87 (+4.22/+1.11)	77.79/69.47 (+5.32/+2.40)	77.92/69.47 (+2.34/+1.21)
		1.0m	73.53/66.37 (+4.22/+1.51)	76.29/68.97 (+4.48/+1.15)	78.63/68.77 (+6.16/+1.70)	79.82/70.17 (+4.54/+1.91)
		4.0m	**76.53/68.57 (+7.22/+3.71)**	**78.17/68.57 (+6.36/+0.81)**	**79.79/71.17 (+7.32/+4.10)**	**81.03/71.57 (+5.78/+3.31)**

图 10：Intelligent Search Serving People

上图是在一个全球化商业搜索引擎数据集上面所做的实验结果。当我们的系统考虑了用户真正的已知反馈之后，整个搜索效果比不用这个反馈的系统好得多。同时，我们可以看到一个非常有意思的现象： 这种智能搜索所发现的知识可以在不同的领域进行迁移。

Model	AUC of fr-FR & de-DE			
	5k	10k	30k	50k
Original	73.05/71.46	73.99/73.15	76.23/75.84	76.82/77.11
$FBQA_{FA}$	76.43/76.64	77.26/76.22	79.28/78.83	80.31/79.76

图 11： 智能搜索所发现的知识可以在不同的领域进行迁移

举个例子来说，在整个模型建立的过程中，我们用的是英语数据，在英语数据里面，我们抽取了相应的问答和相应的用户反馈。英语里发现的知识完全是可以往别的语言迁移，如德语和法语。迁移的效果很好，在法语的数据集上面我们用了很少的大概 5K 的数据就能够达到如果没有跨语言的迁移、没有反馈的时候需要用 50K 的数据才能达到的效果。也就是说，通过应用用户的反馈，我们能够大大减少相应的数据需求，我们的确可以通过智能化的方法理解用户，并让用户把人类智能来帮助我们的机器。我们的技术已经在一个大型商用搜索引擎的多语言服务中上线应用。

三、智能搜索，与人相关

因为搜索的主体是人，所以搜索并不简单是一个技术问题。最近在《纽约时报》有一篇很好的文章，题目就是 Tech is global. right？（技术是全球化的，对吗？）对，技术是全球化的。

图 12： 搜索关乎人类的方方面面

这篇文章谈到了很多先进的美国企业把相应的技术和平台用到别的国家和地区效果不好。这里面涉及到很多因素，特别是很多与人有关的因素，例如说本地化、可理解性、文化、公平性、隐私保护、模型的复杂性、安全性等等。如果我们要把智能搜索做好，就必须密切考虑人的因素。其中，深度学习模型的复杂性是一个重要的因素。我们最近刚刚完成了一篇 KDD2020 的论文在这方面做了一些探索。模型复杂性本身是一个很复杂的问题。在很多场合下，人们可能只是简单地比较两个模型之间的准确度或者别的一些性能指标，但哪怕两个模型的性能在测试集上是完全一样的，并不意味着这两个模型的本质是一样的，也不意味着它们捕捉了同样的客观现实。

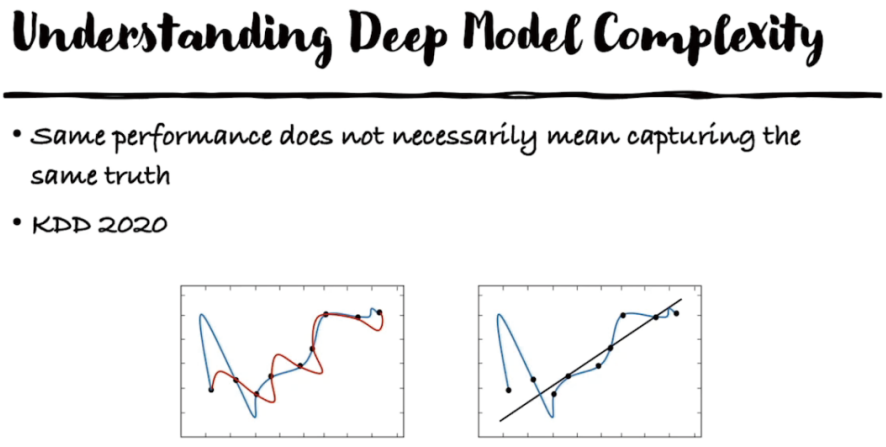


图 13：Understanding Deep Model Complexity

举个具体例子，上图这两个模型在相应的数据点上是完全一致的，但是这两个模型其实差得非常远。因此我们需要有一个系统的方法来衡量模型的复杂度、来衡量模型到底有没有对数据过拟合。我们的 KDD2020 上的论文就在这方面给出了一些新的方法。 与模型和搜索方法很相关的另外一个问题是可解释性，一个模型要获得大家的信任，它必须有良好的可解释性。我们认为，模型的可解释性一定要满足两个原则。第一是准确性： 如果我用一个模型来解释另外一个模型，那这两个模型必须在数学上等价。如果不等价，解释就可能会有问题。第二，模型的解释必须是一致的。一致是什么意思呢？如果我有两个非常相似的样例，它们相应的解释也应该非常相似，这才能够符合人的直觉。可解释性问题的核心是把一个黑盒子转化为一个白盒子。

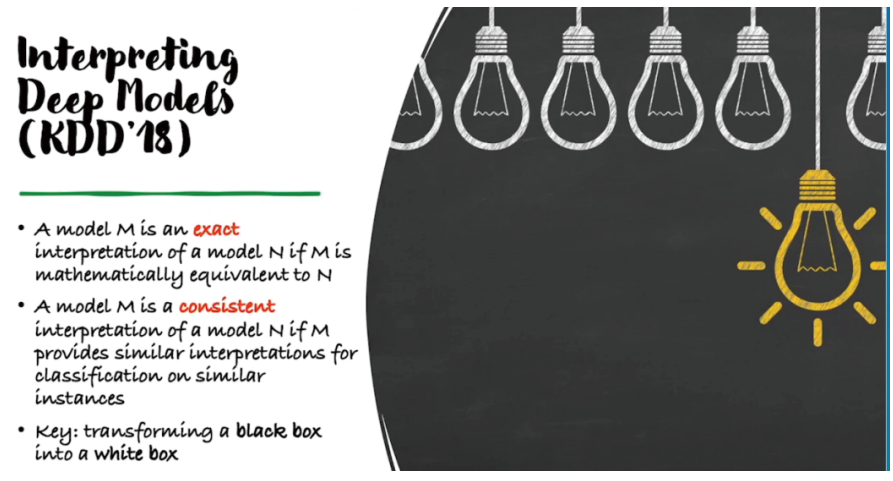


图 14：Interpreting Deep Models

我们 KDD2018 的论文通过把一个深度网络转化为一个基于内部神经元状态的向量，给出基于多胞体 (Polytope) 的解释。这样所得到的解释是精确的：从数学上解释的模型和原来的深度网络等价。同时，解释也是是一致的：如果两个点很相近，它们落在同一个多胞体里面，它们就会遵从相应的相同的线性分类器，所以它们的相应解释也会是一样的。

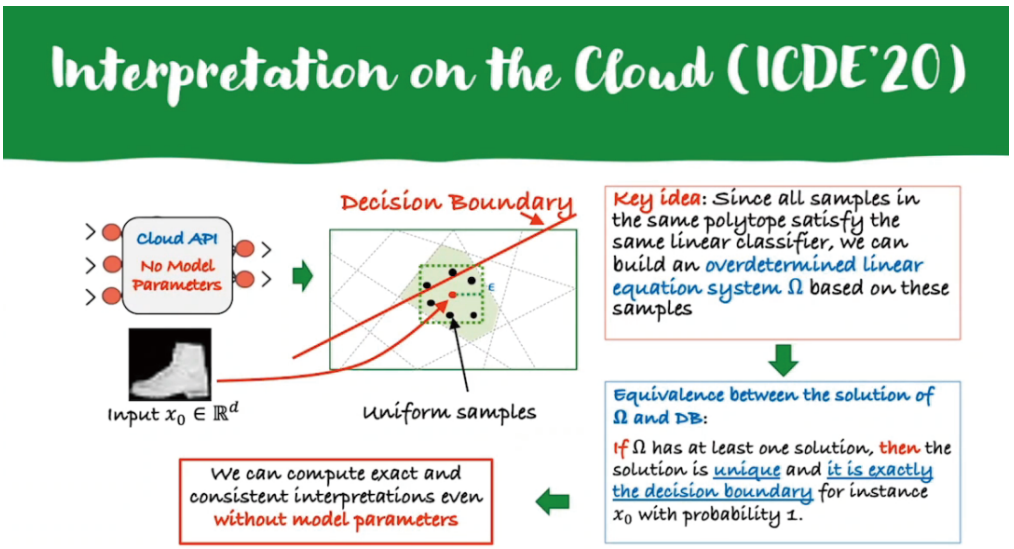


图 15：Interpretation on the Cloud

在今年的 ICDE 论文中，我们把解释模型推到了云端。以往的解释工具往往需要知道整个模型的参数，甚至要知道相应的很多训练数据。在今年的 ICDE 论文里，我们提出可以把整个模型作为一个黑盒，然后给出准确和一致的解释，不需要知道模型的参数和训练数据。这里核心的想法是：如果我们有若干的样例，这些样例落在同一个多胞体里面，它们就应该遵循同一个线性分类器，于是我们可以建造一个线性方程式系统，用此来为我们提供相应的解释。关于怎么样把智能搜索做好，我们讲了很多。但是我们应该要充分认识到一点，智能搜索也在不断地改变我们的心智。

图 16：Intelligent Search is Changing Our Minds

在国外大家经常说这句话：如果你遇到一个不了解的事情怎么办？用搜索引擎查一下 (google it)。在国内也有俗语叫：内事问度娘，外事问谷歌。搜索的过程和结果很深刻地改变了人类的思维和学习方式。在某些方面搜索拓宽了我们获取信息的渠道和速度，使得我们更聪明。但在另一些方面，我们可能会过度依赖智能搜索，在很多地方会变得笨了。这里，信息的准确性和公平性变得非常重要。在这次的疫情当中，我们都知道虚假信息是非常严重的一个问题。很多小道消息、虚假消息通过社交媒体传播产生了很坏的作用。最近推特干了一件很有意思的事情，他们用了一个简单聪明的办法来对付虚假消息。他们观测到有很多人在社交媒体里面看到一个有意思的标题就转发了，但并没有看过那个文章。于是推特在你转发一个没有看过内容的推特时，提示用户其实没有看过这篇文章。这个提示对于降低虚假消息的传播会有很大的帮助。但是这种帮助是有代价的，它需要我们牺牲一定的隐私。推特需要知道你看过什么才知道你有没有看过自己转发的东西。这里有一个挑战性的均衡：我们到底需要保留什么样的隐私，怎么样制止虚假消息的传播。

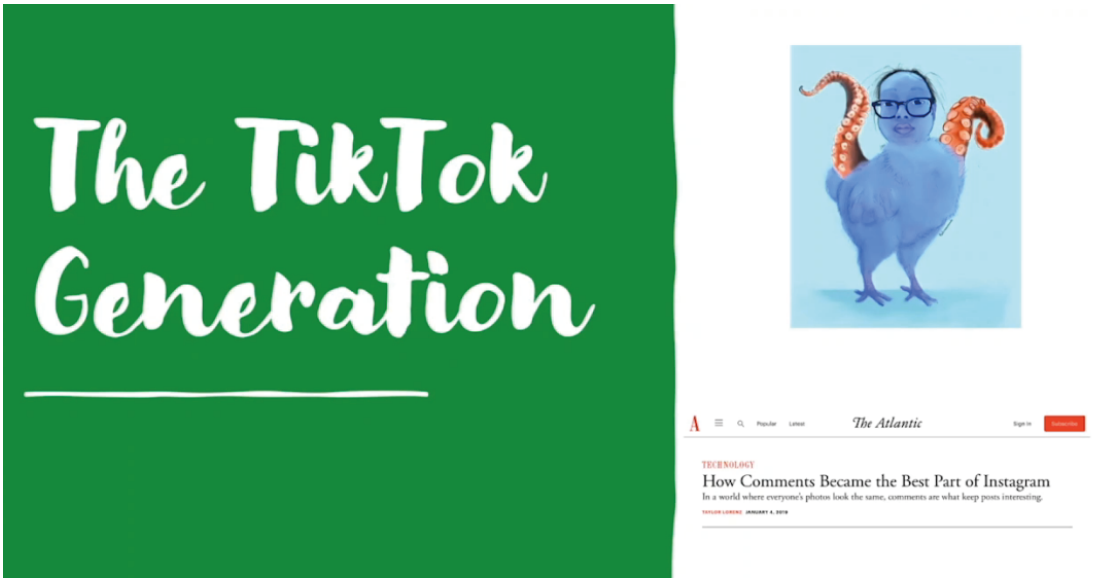


图 17：The TikTok Generation

我们知道抖音在国内外都非常成功，已经出现了一代新的人类叫作 Tik ToK Generation。它们通过智能的搜索和推荐技术把人与人连接起来，把内容和内容连接起来。在内容创造上 Tik ToK Generation 以及这类新媒体有一些重要的特点，其中之一就是媒体内容本身不是那么重要，反而对媒体的评论和媒体的跟进会更重要。大家经常跑到很多新媒体上并不是看它真正的内容，而是看后面跟着的评论。由于智能搜索和智能推荐技术的发展使得人与人之间的连接、内容与内容之间的连接、人与内容之间的连接更加容易、更加广泛。很多人原来并不需要互相认识，但是通过这个智能搜索和智能推荐他们会联系在一起，形成长期的交互，这就导致了我们现在面临着新一代所谓的热情经济。

	The Gig Economy	The Passion Economy
Monetization Model	One-time revenue: pay per trip, per session, etc.	Ongoing revenue based on building an audience
Services Offered	Narrow, commoditized services	Wide variety of creative products and services
Software Stack	On-demand platforms that commoditize providers	Marketplaces that emphasize the individuality of providers SaaS tools that enable providers to run their own businesses
Relationship Between Consumer and Provider	Limited ability for consumer engagement	Platforms encourage direct interaction and loyalty between the service provider and consumer
Levers for Growing the Business	Doing more: more time spent, miles driven, jobs completed, etc.	Expanding audience and offering a differentiated service or product

图 18：智能搜索是热情经济的关键

跟传统的零工经济经济相比，热情经济有一系列新特点。举例来说，热情经济从业者不断地产生新内容，不断地吸引更多的观众获得相应的营收，这是以往很多经济模式不具备的。同时由于智能搜索、智能推荐和平台的连接作用使受众面会大大提高，更多有创意的产品和服务可以以更低的成本推向服务市场，这些也给我们带来很多新机会和新挑战。热情经济完全是基于新的技术、新的软件、新的媒体。智能搜索是热情经济的核心技术，通过技术的进步使得平台更加有效、内容开发更加方便、创业更加快捷、创业者和受众的联系更加紧密、交互更加方便。 智能搜索彻底改变了我们的生活。可以说智能搜索已经变成了我们无时无刻、无处不在的需求和工具。智能搜索同时也会产生很多新的挑战。其中一个核心的挑战是我们怎么确保智能搜索服务于社会的每一个人，没有人因为各种限制而被智能搜索遗弃。

图 19：Intelligent Search for All and Social Good

举个例子来说，老人们会不会因为不会用智能手机而享受不了智能搜索带来的红利？又比如说，残疾人、偏远地区和经济不发达地区的人们会不会因为达不到智能搜索的基础设施入门门槛而被抛弃？这些都是我们需要考虑的问题。我们都知道现在医院挂号经常需要用智能手机来预约，但是很多老人，特别是那些七八十岁、

八九十岁的老人，并不会使用智能手机，用起来也很不方便。他们怎么才能获得信息渠道并消费这些信息？这些都是我们做智能搜索的人需要认真考虑和抓紧行动的方向。我个人认为智能搜索远远不仅仅是一个技术问题，也远远不仅仅是一个人工智能的问题，它是一个非常复杂的全社会的系统工程。

三、问答环节

文继荣：对智能搜索和智能推荐来说，所谓的智能就是越来越了解你，以人为中心来了解你，它给你的信息越来越趋近于你过去的兴趣和经历，但是这样会不会使你失去了解这个世界多样性的可能？在整个大的框架方面或者在整个研究方向上面，有没有更多的深刻思考？

裴健：智能搜索化、智能推荐已经成为下一代人类重要的信息入口，也是非常重要的信息出口。只能搜索把握了这一进一出，对未来的人类有很大的塑造能力。这也许是大家做技术的时候并没有特别深思的一个问题。我们一点一滴的技术贡献会怎样改变未来人类学习的方式、思考的方式和所知所行。这里面涉及很多问题。例如说我们可以通过可适应性使得我们的教育效率提高，使得一个人更容易学习。但是可适应性在一定程度下又可能有缺陷。我们如果过分迁就人类的惰性，就可能会使一部分最聪明的人失去了挑战更高高度的机会。再例如，到底让智能搜索受众学什么？怎么保证整个环境公平性？大家开始去思考，但是远远没有答案。我在演讲的最后也强调了这不是简单的技术问题，这是全社会的很复杂的问题。

观众提问：感觉像谷歌、百度这些巨头已经形成了垄断，其它的搜索引擎或者其它的新的搜索工具经历了多年都没有成长起来。请问这些研究智能搜索方向的人除了到这些巨头公司工作以外，还有没有更好的出路？

裴健：搜索仍在不断地创新，现在所有商用搜索引擎最头疼的事情是越来越多的高质量信息不在公开的互联网上，而在相对封闭的社交媒体上。例如说在朋友圈有很多质量高的信息源，但这种信息源是通用搜索引擎查不到的。怎样把这些信息源整合起来形成一种更强大的搜索能力？这是一个有意思的研究方向。现在很多的搜索跟广告、商业模式结合起来，是商业驱动、利润引导。最近原谷歌的两位高管创办了一个新公司，这个公司做的搜索引擎 Neeva 号称不会有广告，而且要打通一些社交媒体，使得搜索的面更广。这些新业务模式不管成功与否都是非常有意义的尝试。智能搜索从就业、创业的角度来说有很广阔的前景。同时智能搜索会涉及到我们生活中的方方面面，例如说在 IOT 环境下怎么做智能搜索？这些都是现有的面向通用 web 搜索所不能涵盖的，也会是很有意思的方面。

文继荣：现在搜索引擎不管从主要的核心功能还是到形态上已经几十年没有变化了，实际上现在很多东西都在变，比如说裴老师讲的热情经济，还有国内的一个网红经济，现在都是影响非常大的。我昨天看了一个新闻说的非常好玩，浙江余姚区网红可以评为国家级创新人才，不知道真的假的。实际上这个世界在飞速的变化，可能很多时候你认为没有变化空间的时候就是会开始很大变化的时期。就搜索来说，我觉得就直观感受而言还远远达不到我们真正想要了解的世界。这次新冠病毒期间，我觉得甚至可以开一个研讨会来讨论一下这中间的很多问题，人们在获取信息时出现了很多问题，有虚假信息问题也有信息多样性问题。这些信息对大家的影响是巨大的，你可以经常感觉到整个朋友圈都在转发和讨论一个信息。尤其大家在家里没办法面对面交谈，你可以通过控制信息来控制大家的观点和情绪，这个事情我觉得是非常重要的。我们人类将来会走向更加数字化的阶段，从搜索和推荐的角度对信息进行获取和处理，我觉得我们到了一个全新的时期，我们需要去探索。

北京大学教授刘兵：开放世界的人工智能和持续学习

整理：智源社区 张文涛

6 月 23 日，第二届北京智源大会上，刘兵教授做了《开放世界的人工智能和持续学习》的报告。

刘兵，北京大学讲席教授，伊利诺伊州芝加哥分校教授，是数据挖掘，尤其是 Web 信息挖掘领域世界级的领军人物，他的很多工作在领域内具有重要影响力。

图 1 报告主题

开放世界里的人工智能学习和持续学习是很困难也很关键的工作，是通向通用人工智能的必经之路。在本次演讲中，刘兵教授主要讲述了怎么让机器在开放域里面去学习，而不去专门地干涉机器，让它和其他的 agent 交互地学习、持续地学习。跟随让机器在工作中学习这个主线思路，刘兵教授也分享了近几年自己相关的一些工作。最后，也对观众和文继荣教授的一些问题给出了很多独到的见解，相信会给大家带来很多启迪。

以下是刘兵教授的演讲正文:

一、传统机器学习 vs 持续机器学习

传统的机器学习是非常孤立的，有一个任务和一些数据，我们就可以用模型去解决。如图 2 所示，传统的机器学习有几个问题，首先是它要求我们处在一个封闭的世界，我们现在学的东西就是我们所有将来会看见的东西，而机器在使用的时候不会看见任何新事物； 另外，知识在不同的任务中没有任何积累。但现实世界相当复杂，不可能所有事物都在随时学习，同时世界也在时刻发生变化。想要解决这两个问题，就需要在开放世界中去学习。

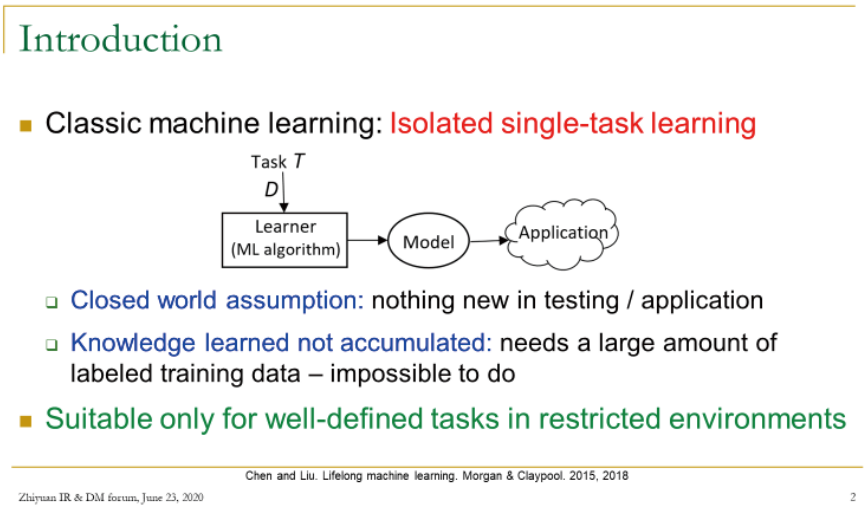


图 2：传统机器学习的局限

与之相对的持续学习有两个典型例子。一是自动驾驶，真实路况十分复杂，此外还有各种各样的突发状况，没办法把所有的情况都考虑进去。但我们人类很善于处理这些情况，即便一个从来没有见过的场景，人也大概知道该去怎么处理，但机器在这种场景下就会比较麻烦，这就是开放场景的特征。另一个是对话机器人，因为语言高度的灵活性，它在设计的时候无法预测用户的问题。即使是在一个很小的领域，如订车票，用户的表述也可以让工程师无法想象。所以，有没有一种办法能让机器自己去学习？这就是一个持续学习（Continual Learning）的问题。

人类在持续学习方面并不存在问题，我们知道如何将不知道的东西慢慢学下来。知识是积累的，人在对话的时候，能够一边对话一边学习。持续学习，最早也有人叫终身学习，具体任务就是当我们已经在 1–N 这 N 项任务上都完成了学习，当我们碰到第 N+1 项任务怎么用之前 N 项任务中学到的知识来帮助这项任务的学习。

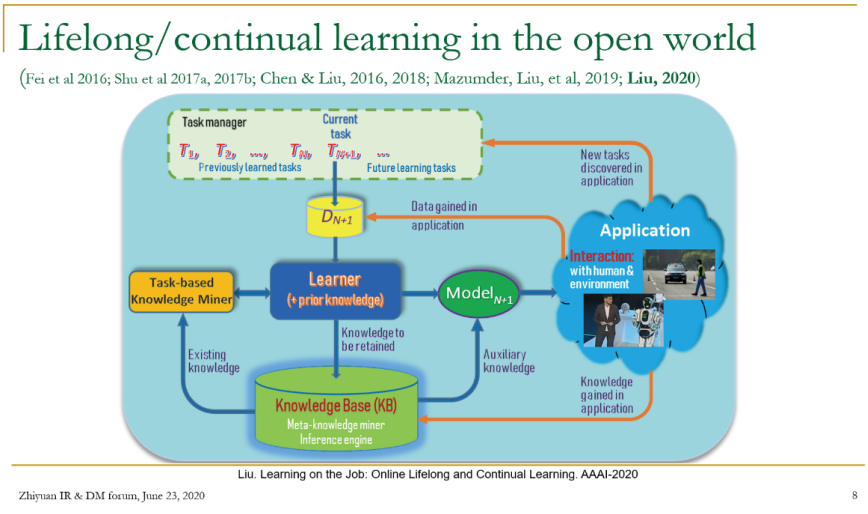


图 3：持续机器学习架构

我们要用过去的知识帮助学习下一个任务，一个任务学习之后会存到下面的知识库（Knowledge base）里，知识库同时也可以进行反馈，实现在工作中的继续学习。持续学习是一个不断学习的过程，我们不能在学习的过程中遗忘之前学到的知识，研究持续学习方向的很多人都在致力于解决这一问题。这涉及到知识积累与适应（Adaptation）的问题，适应即是针对新的情况做出处理。最后，是图 3 中上方橙色的线，表示我们需要在工作中、在实际应用中去学习。刘兵在列举了一个关于自动驾驶的简单例子，他们当时在上海实验一台自动驾驶车辆，当到达一个地方，车在车前检测到一个小石子后停止了前行，只能让驾驶员负责驾驶。这种情况其实很容易处理，如与驾驶员做一次交互，当被告知没问题后就可以继续行驶。

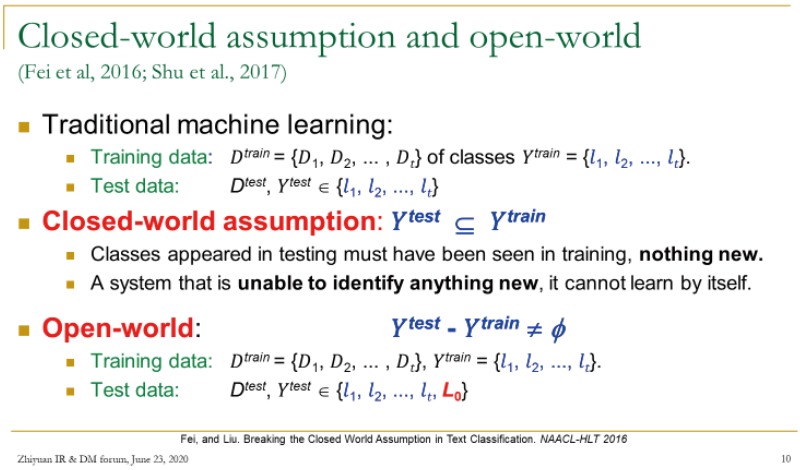


图 4：封闭世界的假设与开放世界

还有一种情况，如遇到一个没有见过的新的问题形成了一个新的任务，只要学习这个任务后，下次遇到同样的任务就会有经验，即不要有封闭世界假设，因为真正的世界很难假设。封闭世界假设的定义也很简单，如图 4，我们的测试数据的类是训练数据的一个子集，也就是说在测试集里不可以出现新的东西，但如果不知道新的东西就不可能自己去学。而开放世界假设这两者不是互相依赖的，也会有新的事物出现。在实际情况下，需要自己去学习处理这种情况。

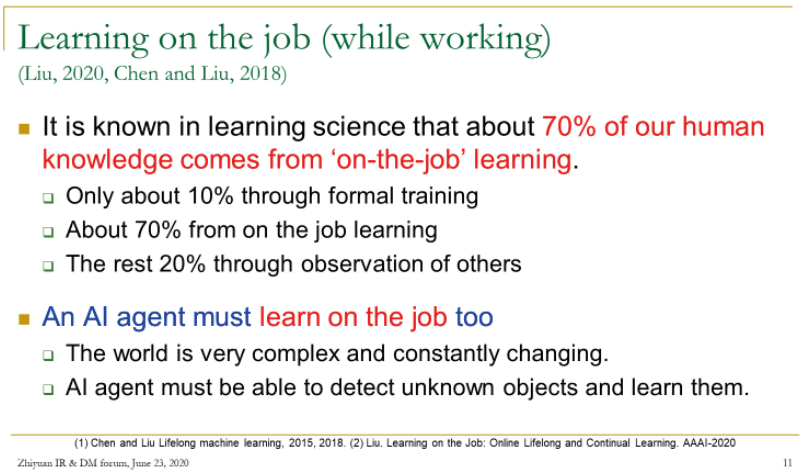


图 5：在工作中学习

而对于在工作中学习，社会科学研究显示： 大概 70% 人的知识是通过工作获得的，这也是很重要的一部分。如在自动驾驶的场景中遇到一个陌生的物体，机器不知道能不能通行，但如果前面有一辆车正常行驶，自己就也可以正常通过。对于 AI agent 也是如此，真实世界非常复杂并且在持续发生变化，我们很难把所有的现象人为地设置进去，因此在工作中学习非常重要。

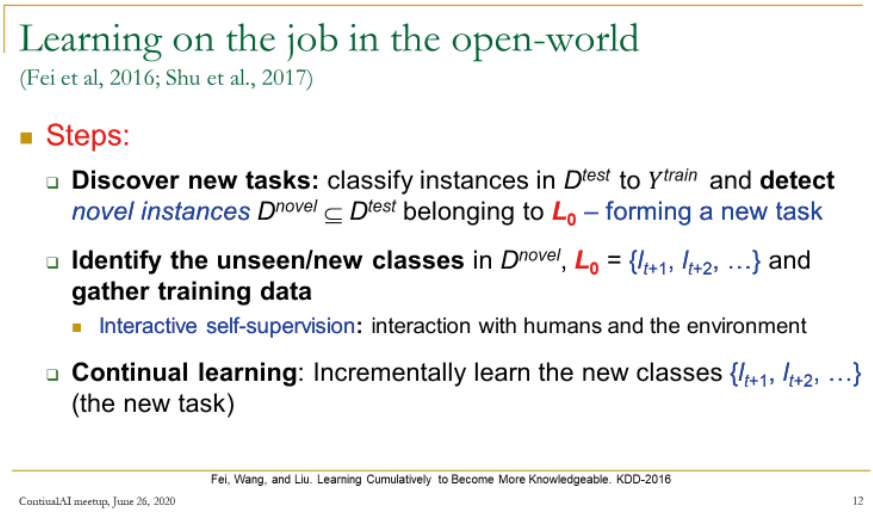


图 6：在开放世界中学习的步骤

总的来说，我们在开放世界中学习的时候有以下几个步骤： 第一步就是需要能够在一个开放的环境下发现新的任务，然后发现未见过的新类型，最后累积已经训练过任务的知识来服务于之后的学习。同时我们的系统需要建立交互的自监督，当机器对一件事不确定的时候，可以通过与人和环境的交互，在工作中获取一些信息。

二、持续学习的挑战

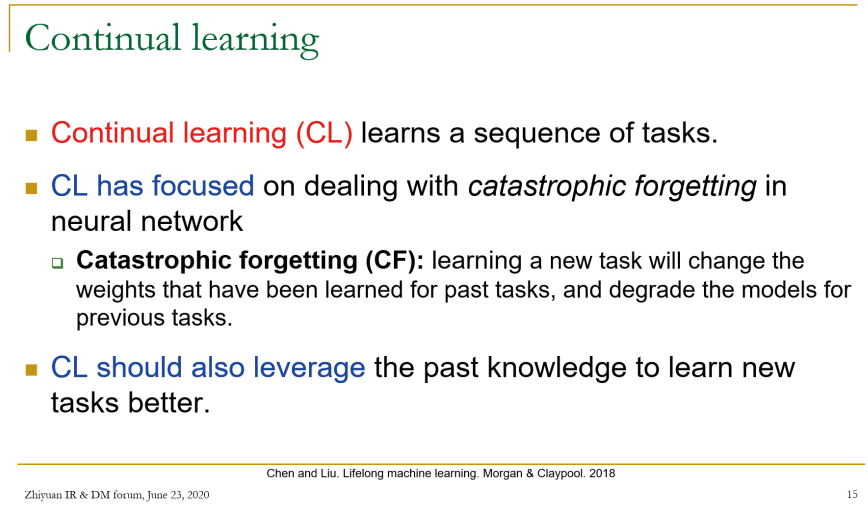


图 7：持续学习的两个挑战

持续学习有两个主要的挑战，一是如何能够持续学习新的知识而不会将以往的经验遗忘，这被称为灾难性遗忘。比如在现在广泛使用的神经网络模型中，知识被存储在网络的权重里，学新的东西就会把过去的权重改变，等于说把过去的东西遗忘了，这样会产生很多问题。第二个挑战是，机器之前可能学过很多东西，有些东西有用，有些东西没用，我们如何选择有用的东西也是一个问题。

三、现有的方案

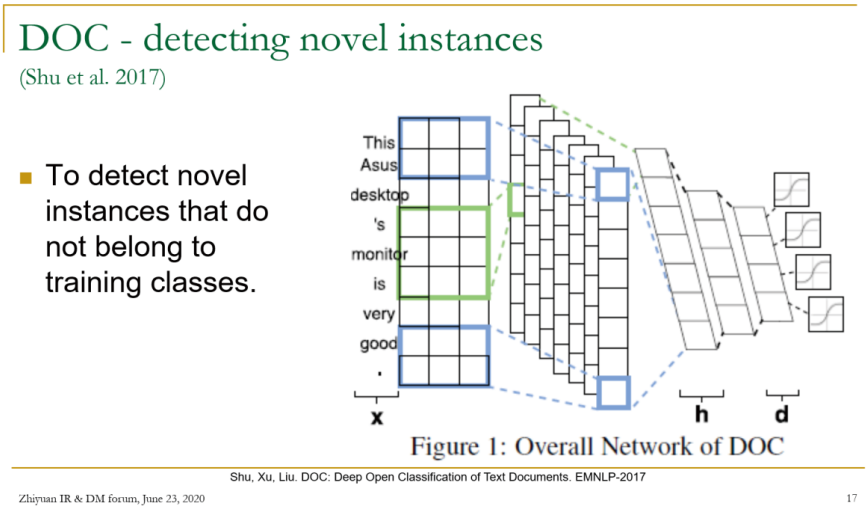


图 8：DOC 结构图

我们来看一下现有的一些工作，如图 8 所示，DOC 改进自一个传统的 CNN，网络的最后一层改成 Sigmoid 层，变成一个 One–Against–The–Rest 的分类器，接着设置一个阈值，来剔除掉不确定分类的样本，从而检测出不属于训练样本类别中的样例。

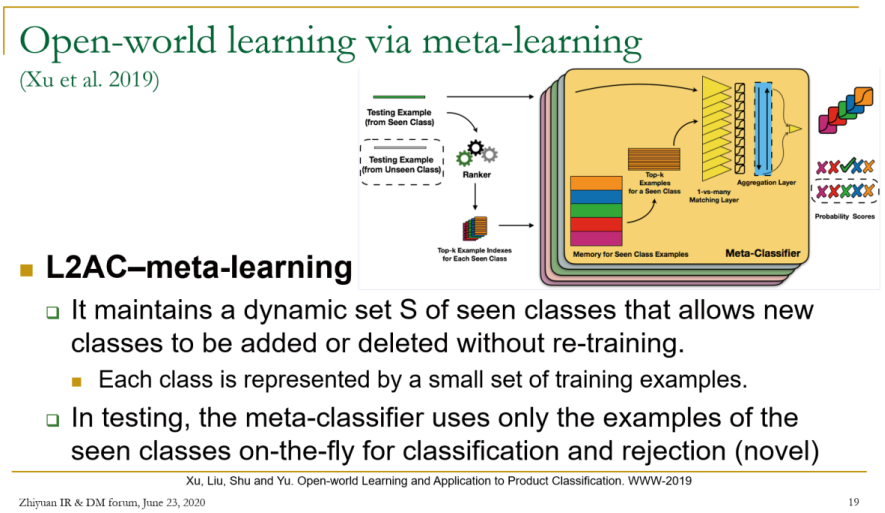


图 9：L2AC–meta–learning 结构图

另一个工作 L2AC–Meta–Learning 是通过元学习来进行的，它的思想是去比对见到的东西和以前见过的哪些东西比较相似，对于一个样本，我们通过元学习中训练得到的距离来判断是否属于已经见过的类别。我们通过以上的技术来使系统更好地发现新的类别，更好地服务于持续学习。

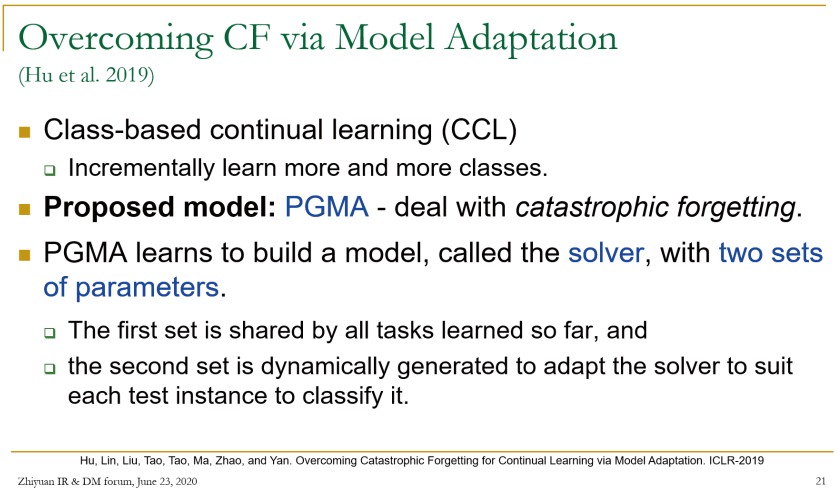


图 10：灾难性遗忘的解决办法

对于灾难性遗忘问题，也有相应的模型。可以将模型的参数分成两部分，一部分对于不同的任务来说都是相同的，可以学到一些通用知识。另一部分则是对于新的任务和样例动态生成的，因此新的任务不会影响旧的权重。

上述方法能让我们的模型不会遗忘已经学到的知识，但仍然帮助不了新的任务。而对于怎么使用过往的知识，这个可以用 KAN (Knowledge Accessibility Network) 系统来解决。它和迁移学习比较相似但又不完全相同，迁移学习一般假设目标数据不够，而我们这个场景目标对象也有数据，而且迁移可以来来回回发生，并且能自动地从过去的任务中挑选有用的知识。

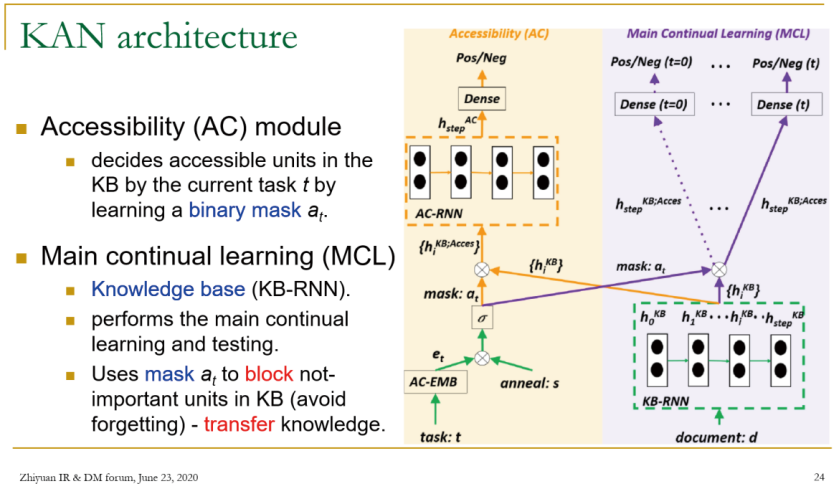


图 11：KAN 结构图

如图 11 所示，这个模型大致的思路是，训练两个模块，第一个是训练一个 Binary Mask，来屏蔽掉过往知识中对现在任务没有帮助的部分，从而避免这些无用知识的影响。第二个部分是主要的持续学习模块，它基于这个训练好的 Mask 能更好地将屏蔽过后剩下的有用的知识迁移到新的任务上。

四、在对话中的持续知识学习

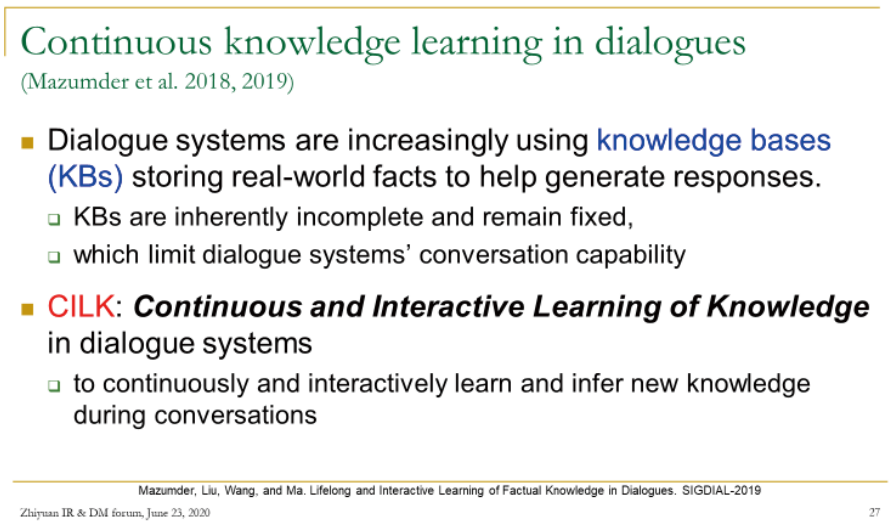


图 12：对话中的持续学习

在对话中做持续学习，要求我们不把形式定死，而是能够意识到自己不知道的信息，并在对话中进行学习。这种情况在人的对话场景下是十分常见的。

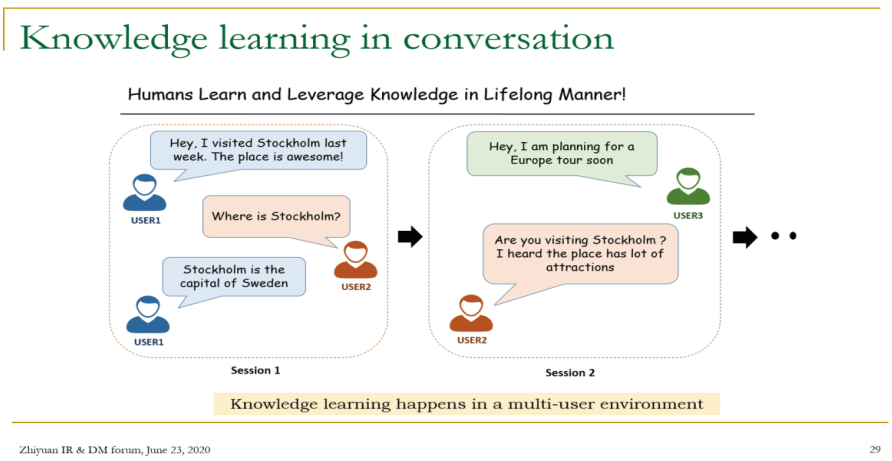


图 13：对话中持续学习的举例

举个简单的例子，可以看到图 13 中的 USER2 在和 USER1 对话时不知道斯德哥尔摩的信息，在对话的过程中学习到了这个信息，并把这个信息运用在和 USER3 的对话中。

在对话里有很多可以学习的方式，第一个就是直接抓对话里的知识，然后就是通过提问的方式去获得正确的信息，最后如果不能回答用户的提问我们也可以问一些和这个问题有关的其他问题，基于这些问题来做推理。所以，第三种形式基本把前两种包含了，这也是我们主要关注的一种设定。

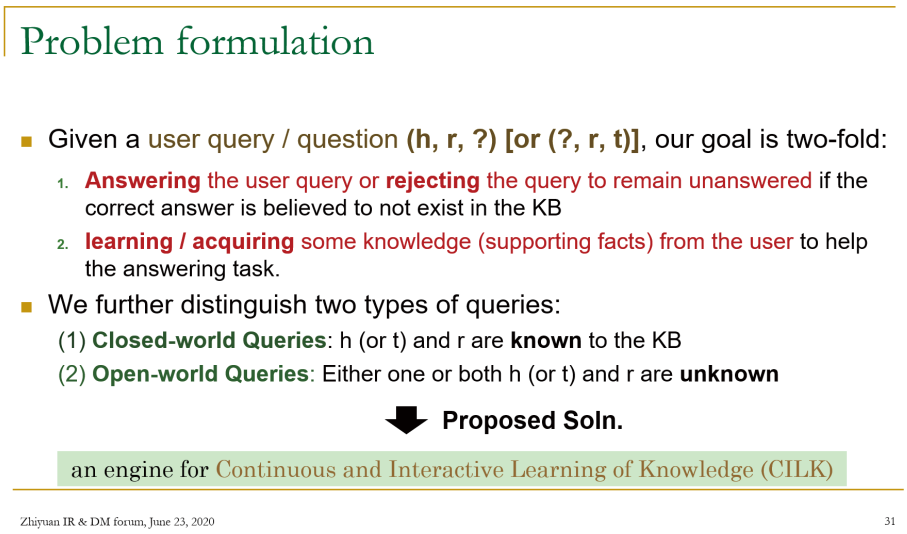


图 14：对话中持续学习的问题定义

具体来说，我们可以类似知识图谱将知识看作一个三元组，当我们面临一个封闭世界的询问，即三元组中的概念。我们在之前积累的知识库中进行判断，但如果面临一个开放世界的问题，即概念在知识库中不存在，那么它就是一个开放世界查询。我们需要通过和用户的交互获得一些可以帮助推理的事实，来解决这样的提问。

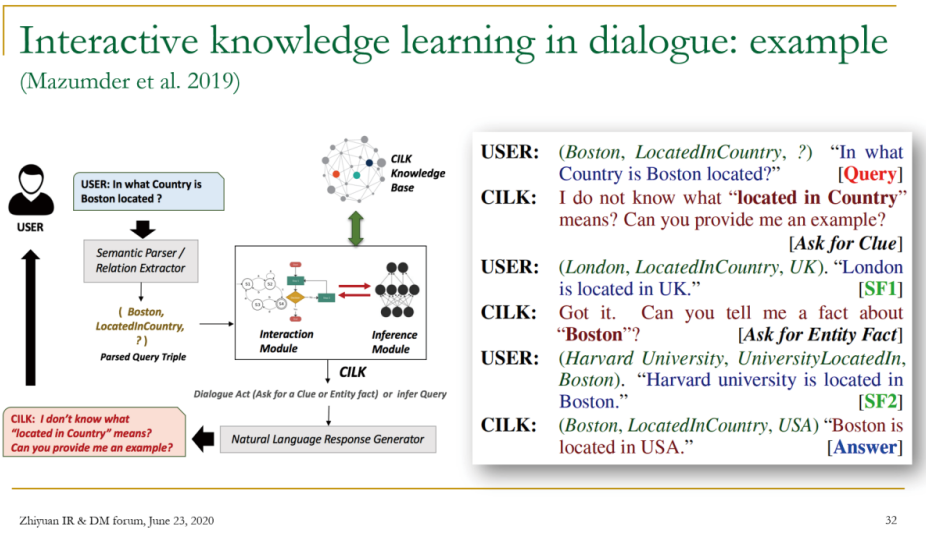


图 15：CILK 原理图

我们可以看到一些解决这样问题的模型，比如 Continuous and Interactive Learning of Knowledge (CILK)。在这个 CILK 处理询问的例子中，我们将自然语言的提问通过语义解析类似的过程处理成一个三元组的提问，如图 15 所示，当系统不知道“处在哪个国家”含义的时候，系统向用户提了一些提示性的问题获得“处在哪个国家”的关系例子，那么我们通过这样交互中得到的事实信息，通过模型的推理模块进行分类，就可以得到问题三元组的答案。

五、总结

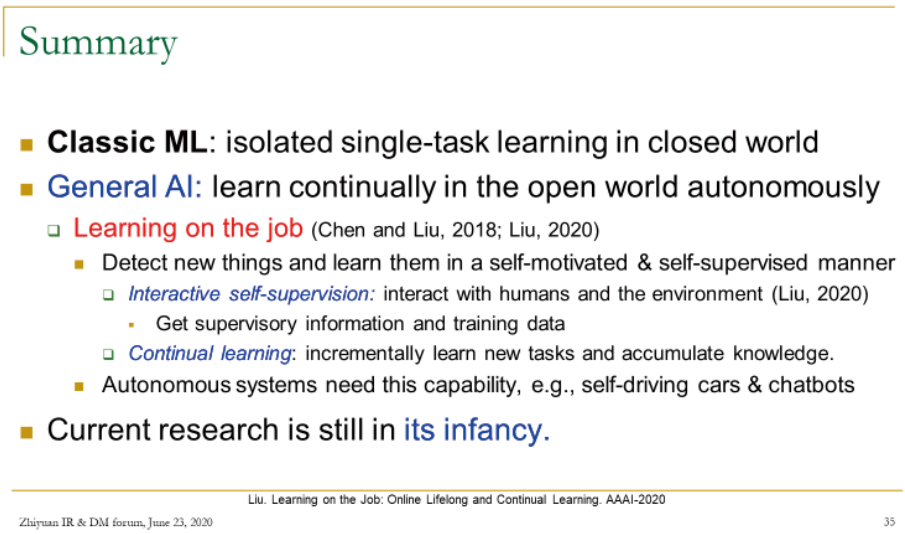


图 16：报告总结

传统机器学习是通过人把数据喂给机器去学，在封闭的世界里学习。将来的问题是怎么让机器在开放域里面学习，不需要人专门地干涉机器，让它跟人和环境自主地交互和学习，持续地去学习，这是非常难的问题。当前的许多研究仍旧比较简单，很多事情我们可以现在开始着手研究。

美国亚利桑州立大学教授刘欢：挖掘社交媒体虚假信息的挑战

整理：智源社区 王建勇

6 月 23 日，美国亚利桑州立大学教授刘欢在第二届北京智源大会上做了《挖掘社交媒体虚假信息的挑战》的报告。

图 1：刘欢报告现场

刘欢，是社交媒体数据挖掘领域的领先学者。刘欢的研究兴趣集中在数据挖掘、机器学习、社会计算等方面，并在社交媒体挖掘领域做出了卓越的成就，因此在 2014 年获得了美国总统创新奖。同时，刘欢是社会计算、行为文化建模和预测国际系列会议的创始组织者，也是《大数据前沿中的数据挖掘和管理》一书的主编以及《社会媒体挖掘 : 导论》一书的合著者，目前是 AAAI、IEEE、ACM、AAAS 的会士。根据 Google Scholar 统计，其论文引用高达 50000 多次。

每年的 315 晚会都会揭穿一批虚假的商品，引发社会的广泛关注。然而在社交媒体上，也存在着大量的虚假信息，这些虚假信息给社会经济带来了极大的危害。相对于实体商品而言，社交媒体上的虚假信息没有实体形态，表达的是一种观点或者感受，因而更加难以处理。虚假信息挖掘是社交媒体挖掘领域中一个非常重要的问题。在本次报告中，来自亚利桑那州立大学的刘欢主要介绍了虚假信息的特点危害性以及相应的挑战，其主要的观点有:

1. 社交媒体中的虚假信息是时刻存在的，社交媒体中的虚假信息危害十分巨大，能够造成无法估量的经济损失;
2. 社交媒体虚假信息的挖掘是一项挑战性很强的工作，其挑战性来自于虚假信息的数据的收集、检测，解释以及对于虚假信息的缓解和防范等方面;
3. 在社交媒体的虚假信息挖掘上，数据是十分重要的，但是大规模的标记数据是不可行的。对于虚假信息的防范可能需要多学科的联合。

刘欢在报告中用新冠疫情信息在社交媒体上的传播作为例子阐述了虚假信息在社交媒体上传播的危害性。特别是在健康与生命领域，虚假信息的传播不仅仅会导致经济损失，更严重地还会危及人民生命。在这次新冠疫情中，虚假信息在社交媒体上传播，也给疫情的防治带来了巨大的困难。对于病毒的理解不够深入以及感染人数的急速骤增，共同导致了网络上虚假信息泛滥，加剧了社会恐慌。涉及药物的虚假宣传，导致公众对于滥用药物，进而威胁生命安全。因此，社交媒体虚假信息挖掘是社交媒体挖掘的重要任务，具有重大的经济和社会价值。

下面，是智源社区编辑整理的刘欢演讲要点。

一、社交媒体信息需要“打假”

虚假信息检测并不是一个新的问题，它一直贯穿于社交媒体的发展进程中，并对社交媒体的发展不利影响。首先，我们要先对相关概念进行了解，错误信息和虚假信息是其中最为基础的两个概念，错误信息包含虚假信息、虚假新闻、谣言、都市传说、垃圾邮件和钓鱼网站。在对虚假信息有所了解之后，我们应该如何处理虚假信息呢?

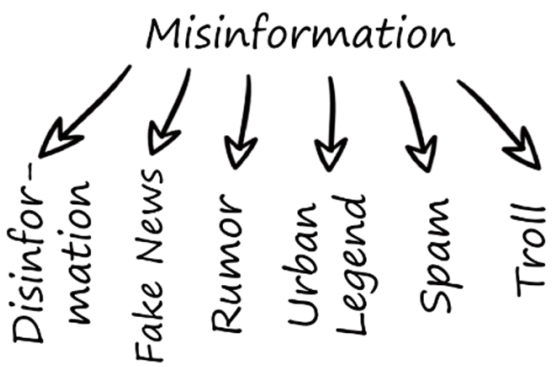


图 2：虚假信息的概念

在 SIGKDD–2019 的会议上，刘欢的研究小组发表了关于如何定义、操作以及检测社交媒体中错误信息的文章。关于虚假新闻和虚假信息的研究，在社会科学领域已经开始很久了，这篇文章建立在早期的研究结果的基础上，。此外，刘欢团队已经发表一本关于虚假新闻检测的书，并且还有一本将会在近期发表。这些书对已有工作进行了介绍，并且对现有的算法进行了改进，提升了算法的检测能力。

在社交媒体中，存在很多误导性信息:

1. 用健康、保健产品来替代药物，例如社交媒体中发布的未经证实的“预防措施”和“治疗手段”；
2. 各种阴谋论，例如在实验室中设计流行性生物武器的论调；
3. 欺诈和诈骗信息，例如虚假的疫苗信息和虚假的捐款网站等。

有时，甚至官方的新闻媒体也会出现发布错误健康信息的情况，人们往往对相关健康问题缺乏足够的认知，容易受到错误信息的影响，在缺少治愈方法的情况下，人们通常会绝望地相信自己在社交媒体上检索到的任何“治愈方法”，在疫情期间，这个问题曾在世界各地出现。

由于阴谋论可以通过多种方式进行传播，导致其难以被检测和阻止，此外，人们心中的恐惧与恐慌也加速了阴谋论的传播。近年来，社交媒体已经成为阴谋的主要传播途径之一，在疫情期间，阴谋论也大量出现在社交媒体中，例如新冠病毒是人为投放的。

社交媒体因其易于访问和广泛传播的特点逐渐成为信息共享的流行方式之一。 多年来，使用社交媒体平台的人数正在快速增长，越来越多的人选择在社交媒体上获得新闻。但是，社交媒体是一把双刃剑，它在传播信息的同时，也会传播虚假信息。从心理学的角度来看，我们作为人类非常容易受到假新闻的攻击。社会科学中的“确认偏差理论”表明，人们倾向于相信符合其现有知识的信息，无论它是假的还是真的。此外，虚假新闻可能会对社会产生不利影响：它可能会使读者困惑，误导人们获取虚假信息。

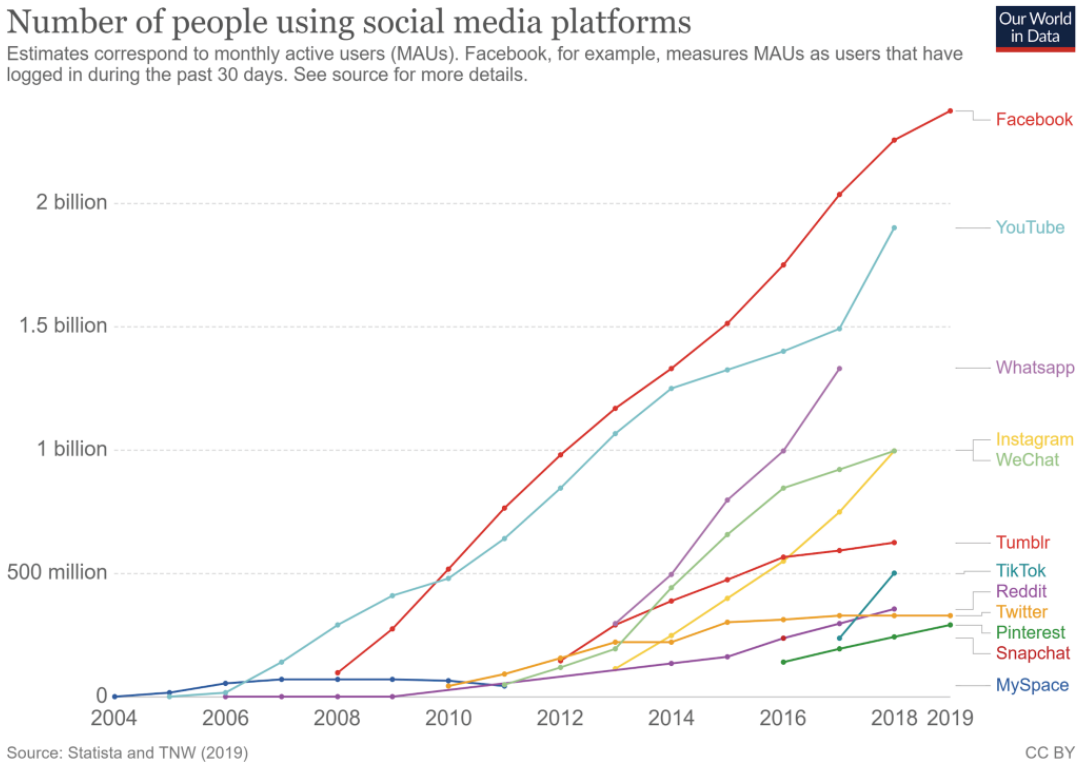


图 3：社交媒体使用情况

在现实生活中，充斥在社交媒体中的错误信息导致了非常严重的损失，例如由美国联邦贸易委员会公布的数据显示，在疫情期间，错误信息的传播已经为美国带来数千万美元的直接损失，间接损失更是达到数亿美元。如何处理这些错误信息，已经成为一个亟待解决的问题。

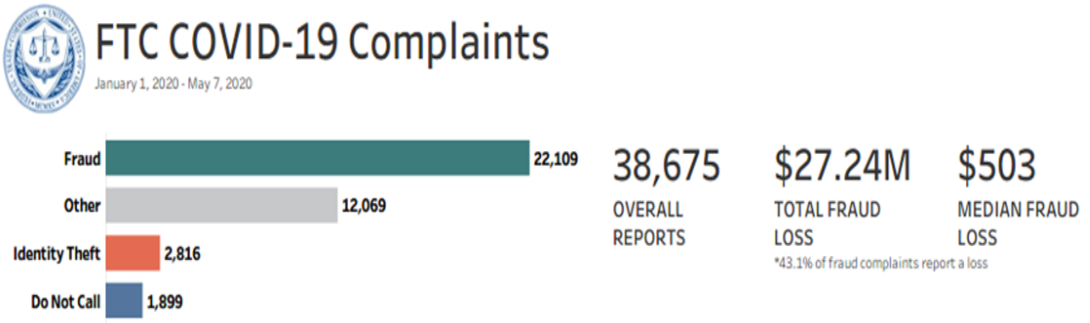


图 4：新冠病毒虚假信息造成的危害（数据来自于 FTC）

二、信息打假，困难重重

虚假信息充斥着社交媒体，但打击虚假信息却困难重重，原因到底出在哪里呢？刘欢老师以假新闻为例进行了阐述。

首先，实际场景下的假新闻检测并不像机器学习比赛那样，能够获得一个已经有标签的数据集，并对各类方法的效果进行准确的评估。其次，假新闻检测的复杂性往往体现在多个维度上，只从一方面着手并不能完全解决问题。目前，我们所面临的紧迫挑战主要在于数据，检测，可解释性和虚假信息的缓解与遏制等方面。之后也将着重阐述虚假信息的缓解与遏制相关内容。

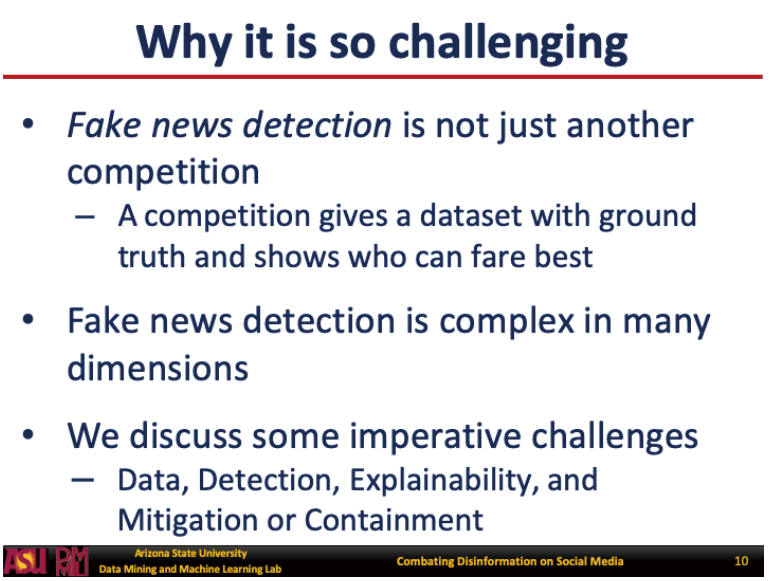


图 5：虚假新闻检测的难点

对于虚假信息检测，如果我们需要信息的真实标签 (Ground Truth)，那么就必须进行事实核查 (Fact-checking)。但事实核查不仅需要领域专家的参加 (如下图中对有关沃尔玛的虚假信息辟谣的例子)，还往往伴随着密集的劳动和大量的时间消耗。那么面对这些问题，我们应该如何快速获得信息的真实标签呢?

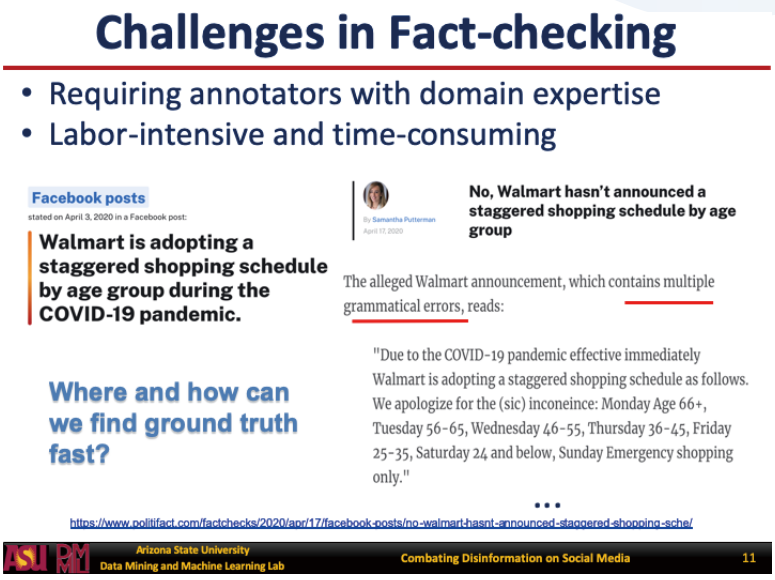


图 6：事实核查面临的挑战与举例

答案就是必须依靠数据来快速获得事实。以 COVID–19 为例，与 COVID–19 相关的数据集是横跨多学科多领域的，包括时空数据、社交媒体数据和学术文章等。虽然有很多相关的数据，但人们可能没有途径去获得这些数据。现在已经着手开始构建一个元数据仓库（Meta–data Repository），旨在将这些分散的、异构的公开数据集中起来，并希望以此推进相关领域的合作。

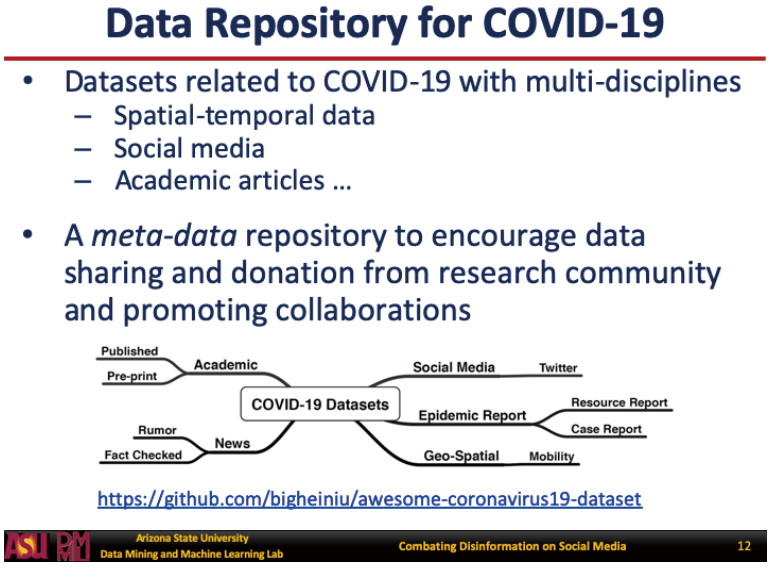


图 7：COVID–19 数据仓库

对于假新闻检测中的挑战，主要体现在新闻内容和社交情境（Social Context）上。自媒体还没有兴起的时候，官方媒体在新闻发布前，都会对新闻的内容进行确认，这很大程度上遏制了假新闻的产生。但现在的很多新闻，会故意在内容上误导读者，以此博人眼球，新闻发布前也没有进行内容核查。对于这些新闻，由于其主题、风格和媒体平台的多样性，检测难度急剧增大，以前一些有效的检测方法也可能失效。此外，对于社交媒体，情况又有所不同。社交媒体中的互动，如“点赞”、“踩”、“评论”等，虽能用于帮助假新闻的检测，但其数量巨

大，不完整，无组织，有噪声。因此如何找到高效的方法来利用这些丰富的社交信号，也是我们亟需解决的一个问题。

图 8：虚假新闻检测的挑战

可解释的假新闻检测，并不是对假新闻背后的因果关系进行阐述，而是在新闻内容或评论中找到支持判定结果的部分。可解释性是重要的，因为我们不能完全依靠数据，我们还需要依靠专家的领域知识。如果我们能够提供这些可解释的特征，那么这将可以帮助并鼓励专家与数据之间的协作。

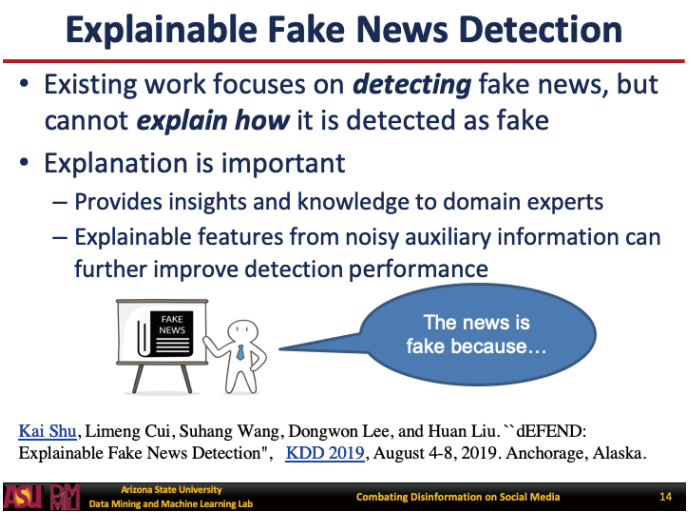


图 9：可解释的虚假新闻检测

虚假信息风险缓解（Mitigation）面临着诸多挑战。首先，人们可能会产生疑问，虚假信息风险缓解是否要比虚假信息检测更容易？但实际上，风险缓解和检测是两个不同的问题，不论检测的准确率有多高，我们都不能保证成功地进行了风险缓解。作为计算机科学家，我们经常对数据驱动的方法非常自信，但事实上风险缓解却要更加复杂。原因在于，风险缓解涉及用户，而每个用户都是信息（包括虚假信息）的传播点，所以风险缓解的复杂度体现在新的维度上：1. 用户有自己的判断和观点。对于相同的信息，不同的用户会有不同的反应；2. 用户

在社交媒体上不是孤立的。

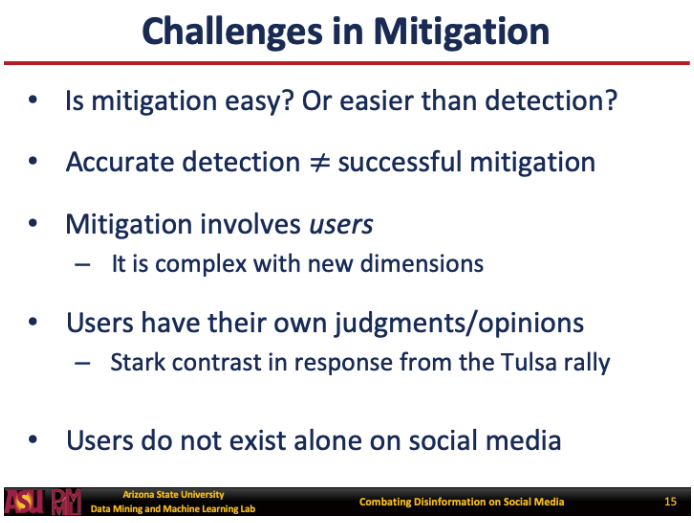


图 10：虚假信息风险缓解面临的主要挑战

缓解虚假信息的负面影响面临的困难恰好解释了“愚蠢”这个词的含义，即“知道了真相，看到了真相，但依旧相信谎言”。然而这种“愚蠢”比任何疾病都更有传染性。用户或多或少存在不理性的情况，虽然我们可能自我感觉良好。我们可以问任何一个人这样一个问题：“我们易受到虚假信息的影响吗?”人们通常会回答：“不，我对假消息免疫。别人可能会被假消息欺骗，但我不会。”然而回声室效应（Echo Chambers）让我们更加固执己见。我们可以检查一下我们的社交网络，我们能从中找到不同的意见吗？往往当强烈的反对意见在一个群组里出现时，很快就会有人退出群组。另外互联网的“过滤气泡”（Filter Bubbles）让我们的信息来源无形中受到了限制，我们能看到的都是我们想看到的，而其他的重要信息都被这无形的“过滤气泡”过滤掉了。我们可以看到，自己每天的新闻来源基本来源都非常有限，而我们仅从这些有限的来源中获取新闻显然是不够的。这告诉我们，我们不仅要做到开放包容，还要承认自己也会犯错误。

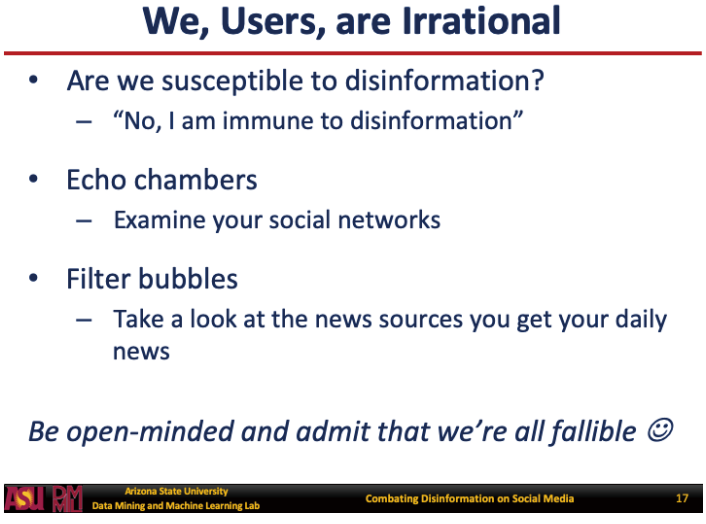


图 11：用户本身存在缺陷

除上面所提到的外，还有一些挑战不容忽视。通常我们都会急切地想要传达自身的想法，但却忽视了会话接受性 (Conversational Receptiveness)，即忽视了如何让自己的观点更容易被人接受。另一个问题在于，人们在日常生活中往往会犯比较低级的逻辑错误。比如误认为“如果 $A \rightarrow B$, 那么 $\neg A \rightarrow \neg B$”。

图 12：信息打假的经验与教训

总结整个报告，我们了解到了以下几点：1. 假新闻检测是困难的，因为目标和主题都是动态的。2. 数据是关键，但想要快速地标注大量数据是不切实际的。3. 在新闻产生的前期进行检测是至关重要的，因为后期检测中即使使用数据驱动方法效果也有限。4. 风险缓解并不容易，因为每个人都有自己的偏好。

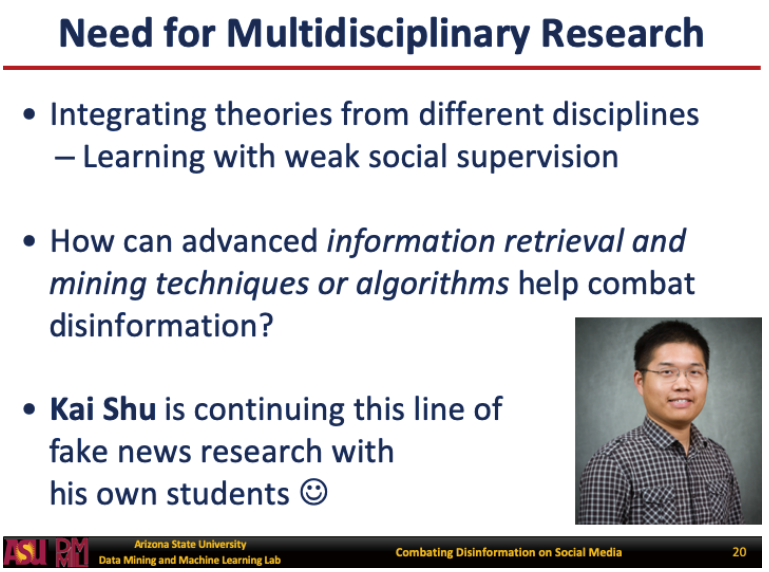


图 13：未来展望

为了应对这些挑战，我们需要多学科的共同研究。整合不同学科的理论，将数据驱动的方法与其他学科的方法结合起来，例如在社会弱监督下进行学习。那么如何将信息检索和挖掘的技术或算法用于帮助对抗虚假信息呢？刘欢认为这之中还有很多工作要做，需要更多的人投入进来。现在中国在这方面做得很好，相关研究的资金充足，也有大量人才的投入。刘欢的学生 Kai Shu（上图）也将继续进行这方面的研究。

问答环节

文继荣： 我先问刘老师一个问题，虚假信息的检测重要性是无可置疑的，在将实验室技术应用到产品里去这方面，你有没有一些经验跟大家分享？假如今日头条要开始在我的内容推送里面加入虚假检测，它应该怎么考虑这个问题？

刘欢： 虚假检测很难，像今日头条、Facebook、Twitter、Instagram，它们都有大量的用户人群。如果只是简单地将某些人或网站加入黑名单，由于检测结果存在“假阳”和“假阴”，这样可能会损害到公司的业务，但是不做又不行，引发谣言会招致处罚。更重要的是，这个工作光靠计算机科学家是不够的，必须要和社会学家、记者协同进行才可以。一个方面是人性，谈到人性方面的时候，我有一次在另外一个地方作报告，下面有一个非常著名的计算机学家，他就说：“这很容易，直接告诉那个人这是假新闻就行了”。但你告诉他这是假的，他可能还会来跟你吵架，还有可能比以前更坚决地相信这件事情，这就是在美国大选的时候出现的事情，好多人利用这些东西来分化人群。那该怎么办？裴健老师刚提到一个办法，其实就是延缓一下，如一两秒钟，我们必须提醒用户，无论是运用显式还是隐式的延缓策略。

文继荣： 在实际情况中，虚假信息的定义很难。比如在这次疫情发展期间，在疫情初期的很多观点在当时是很难判断的，当时很多东西还处于未知的状态，大家可能提出不同的看法，如说到底传不传人？是否应该采种隔离措施？在今天这些问题可能是有答案的，但是那个时候是没有的，是一个发展的过程。如果我们给它们贴上了虚假信息的标签，可能会阻碍这些观点的传播和碰撞，这些情况的存在就使得问题变得更加复杂。

刘欢： 我们需要意识到，科学发展的进程中，其实是一个不断纠错的过程。如果我获得更多的数据，我就能得到更多的信息，就更加接近真实的答案；相反，如果我没有太多的数据，我就无法做出准确的判断。科学是一个进程，是在不断发展的。虚假信息的检测，同样需要数据的支撑。

文继荣： 我觉得从这个思路来讲的话，这种虚假信息的鉴别可能就像刚才刘老师说的，在很多时候并不是非黑即白的一个情况，甚至说它这里面可能有部分是真的，有部分是存疑的，有部分是已经被证实的。但是在鉴别虚假信息过程中，我们怎么才能对明显有害的部分进行控制？又同时不能阻碍真理越辨越明的过程。

刘欢： 这确实是这样，特别是在国内，为什么？国内经常会出现一哄而起的情况，其实真理是越辨越明的，需要一定的时间。我小的时候经常看到国内有人去打鸡血的情况，后来有一段时间流行吃生的茄子，在当时即使你去制止这种情况，也很少有人会相信。不过挑战就是机遇，这也证明我们的方向还有很长时间可以去做。

智源首席科学家文继荣：下一代智能信息检索技术的发展方向

整理：智源社区 秦绪博

在 2020 年 6 月 23 日上午的“智能信息检索与挖掘专题论坛”中，本场论坛的主持人，智源首席科学家，中国人民大学高翎人工智能学院执行院长文继荣教授做了本场论坛的开幕致辞。致辞中，文继荣教授对本场论坛所覆盖的主题——智能信息检索与挖掘的发展历史和未来的研究方向做了介绍。

自上世纪末以来，搜索引擎技术已经成为了人类从大规模数据中获取信息的最为主要的，也是最为成功的手段之一，先进的商业搜索引擎使得人们获取信息的手段相比二十年、五十年前有了飞跃的进步。但是在经过此前的高速发展阶段之后，在最近的十年，人们搜集和获取信息的方法和技术出现了一个相对停滞的阶段。文继荣指出，现有的信息检索技术需要一轮新的变革，而未来的下一代信息检索技术的一种可能的解决方案，应当是基于智能交互的个人智能信息助手，它可以支持自然语言交互，并具备知识增强和个性化满足用户信息需求的能力。最后，文继荣指出，下一代智能信息检索技术的发展，需要多个研究方向的学者们通力合作，并对智源研究院智能信息检索与挖掘平台的未来发展进行了展望。

以下是智源社区编辑整理的文继荣演讲要点。

一、信息检索技术的历史和现状

智能信息检索与挖掘是智源人工智能研究院成立的第三个主要研究方向，它聚焦于智能信息检索和数据挖掘，主要目标是如何利用现代的人工智能、数据挖掘等相关技术，来帮助人们更好的获取信息。众所周知，搜索引擎是目前人们获取信息的主要手段，也是商业上比较成功的工具，在谷歌、百度等伟大的公司的努力下，目前我们的信息获取能力，相对于二十到五十年前已经有了飞跃式的进步。

图 1：搜索引擎的背景

但是现在回过头来看，我们可以发现，经过上世纪末、本世纪初的飞速发展以后，我们的信息获取手段在最近十年进入到了一个相对停滞的阶段，好像大家都已经对现有的信息检索技术比较满意了。然而另一方面，互联网上的数据总量在过去十几年内爆炸式地增长，但我们使用的信息获取工具却并没有相应的随着发展。我们开

设这个方向，是希望能从学术上，甚至从产业化上，为人类的下一代信息获取工具的发展做出贡献。在 2000 年左右，谷歌曾经提出，要把世界上所有的信息都组织出来为人所用，从目前来看，这样的目标还远未实现。

二、下一代的智能信息检索工具

我们认为从整个信息检索技术的各个不同方面来看，现有的技术都需要一次新的革命和飞跃。例如从用户的角度来看，用户现在需要更丰富的手段来获取信息，除了传统搜索需求（例如搜索网页和图片）之外，用户还希望搜索引擎能完成更加复杂的信息分析，甚至直接辅助进行复杂决策；从信息获取的场景来看，用户希望能随时随地获取信息，比如使用手机或者自动驾驶的时候，而不再像以前一样只能通过桌面 PC 获取信息；从数据的形式上也是如此，现在手机环境有很多不同的 App，对应很多不同形式的数据，使得传统面向 HTML 型网页的搜索引擎自然显示出了很多局限性。另外，在搜索结果的评价方面也有很多挑战——目前新数据的获取方式越来越丰富，那么我们要如何评价系统性能的好坏，如何判断一个搜索系统是否能满足用户的信息需求?

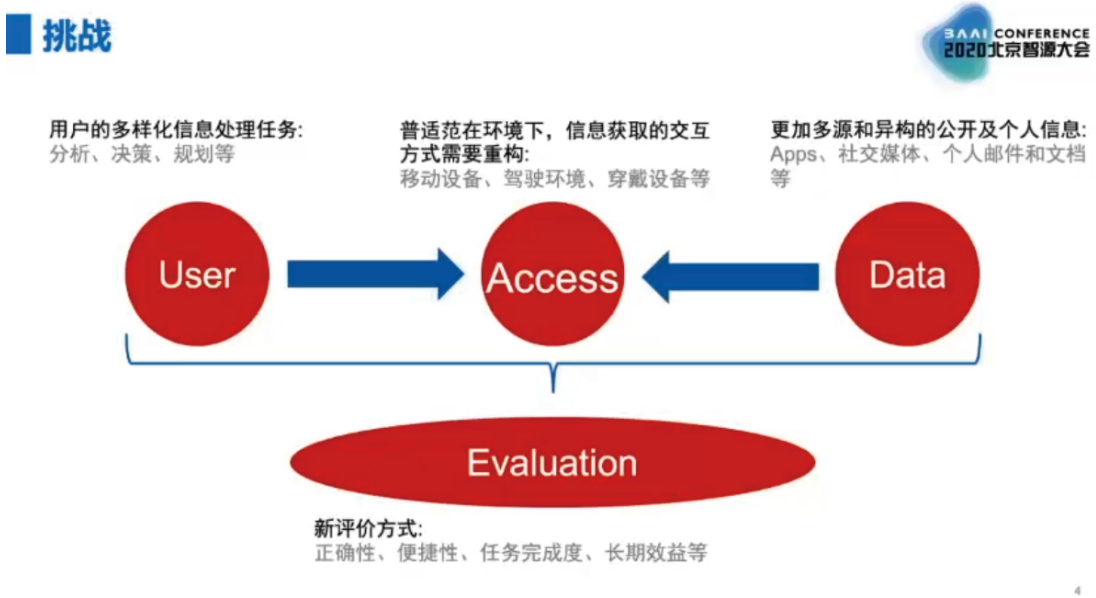


图 2：当前搜索引擎面临的挑战

从去年开始，我们这个方向的智源学者们进行了一系列的讨论。我们认为未来智能信息检索工具的形态，应当是基于智能交互的个人智能信息助手。我们的各位学者们都来自于各个细分的领域，因此需要有一个共同的目标，可以把大家的工作呈现出来，让大家都能围绕这个目标去开展自己的研究——进一步讲，我们希望未来的智能信息检索系统，能够充分地利用目前人工智能领域各个相关方向的研究成果，从不同的方面尽可能地提升用户实际体验。因此，我们提出了个人智能信息助手——我们希望它能支持自然语言交互，支持对话式的检索；我们希望它能更加个性化，可以满足你的信息需求；我们还希望它是知识增强的，能够给用户带来知识和答案，解决用户的问题；最后，我们希望它可以整合不同类型的数据，在各种各样的数据上进行智能的搜索，充分利用各种形态的数据来满足用户的需要。

图 3：打造基于自然交互的个人智能信息助手

以上就是我们提出的研究路线，我们将从理论、算法和系统等多个方面来解决这个问题。我们也跟其它方向的学者们进行合作，例如数理基础、认知基础理论等。我们的目标是，从算法层面上实现基于自然语言的交互，基于深度语言的模型以及知识生成的信息展示，最终构建下一代的个人信息智能助手。

三、智源研究院平台未来的发展

目前，我们的各位成员都是来自于北京地区各个高校和中科院的优秀学者，其中还有很多青年学者。我们希望大家在智源研究院这样的新平台下，能够安心地做自己想做的研究，通力合作，为了一个共同的目标，把我们国家在信息检索、数据挖掘这个方向的发展水平推进到一个新的高度。

我们当时提了一个口号，希望经过我们的努力，最后能够打造一个智能信息检索与挖掘的北京学派，这个目标实际上也并非那么遥不可及。我们认为，经过我们的合作，一方面在今后吸收更多的优秀学者加入，另一方面跟国内和国际的学术界通力合作，最后应当能够实现这样的目标。