SIGIR2018、WWW2018 知识图谱研究综述

肖仰华
复旦大学知识工场实验室
shawyh@fudan.edu.cn
2018.08.16

SIGIR 18、WWW18 知识图谱（KG）研究概览

分类	相关论文
知识获取	上下文事实发现：Weakly-supervised Contextualization of Knowledge Graph Facts，SIGIR18 KG扩充：Enriching Taxonomies With Functional Domain Knowledge，SIGIR18 三元组规范化：CESI: canonicalizing open knowledge bases using embeddings and side information, WWW18
知识挖掘	必有属性挖掘： Are all people married? Determining obligatory attributes in knowledge bases, WWW18 分面标注：Facet annotation using reference knowledge bases, WWW18
知识应用	会话系统： Response Ranking with Deep Matching Networks and External Knowledge in Information-seeking Conversation Systems， SIGIR18 问答系统: Knowledge-aware Attentive Neural Network for Ranking Question Answer Pairs SIGIR18; Never-ending learning for open-domain question answering over knowledge bases, WWW18 推荐系统: Improving Sequential Recommendation with Knowledge-enhanced Memory Networks SIGIR18 DKN: deep knowledge-aware network for news recommendation, WWW18
知识评估	KG补全评价指标：On Link Prediction in Knowledge Bases: Max-K Criterion and Prediction Protocols, SIGIR18 规则评估：Estimating rule quality for knowledge base completion with the relationship between coverage assumption, WWW18

趋势1（知识获取）：基于深度学习的知识获取成为热点

知识获取：上下文知识获取

Weakly-supervised Contextualization of Knowledge Graph Facts, SIGIR18

研究问题：对于给定的KG 事实（实体关系三元组），找出相关的事实

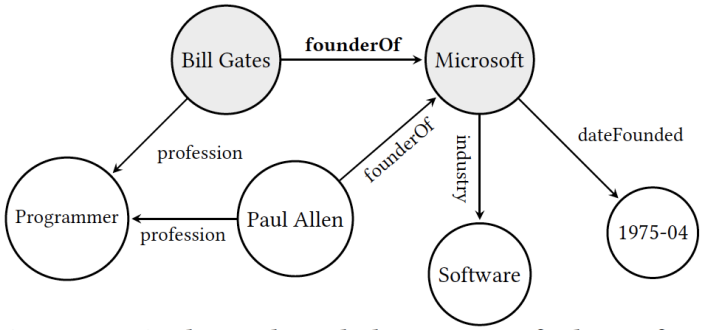
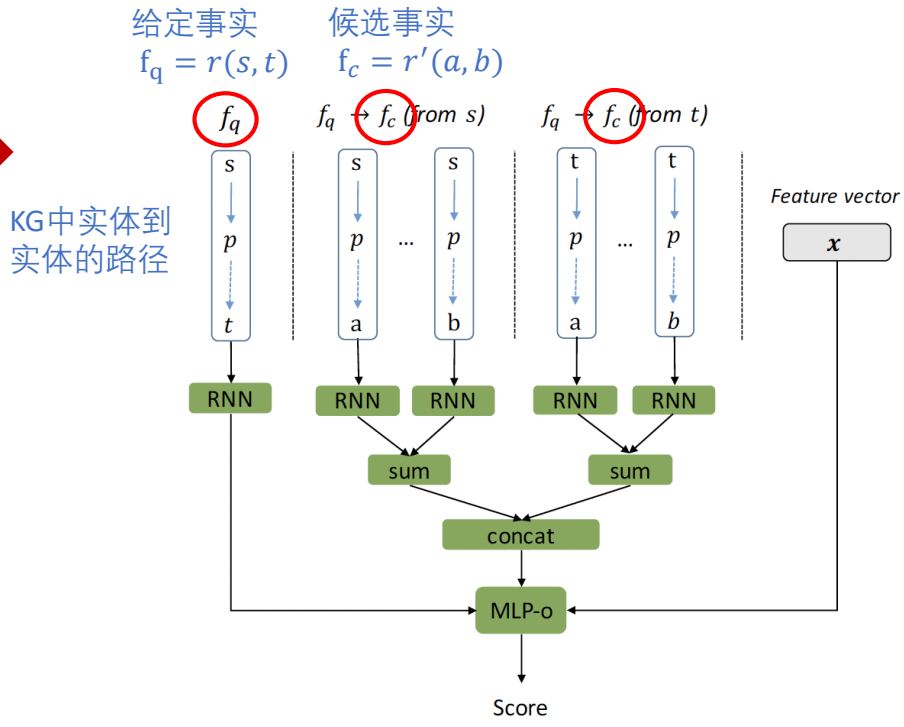


Figure 1: A Freebase subgraph that consists of relevant facts to the query fact *founderOf*(Bill Gates, Microsoft).

存在的挑战：
KG很大，即使在很小邻域找相关的事实，都会产生大量候选事实

候选事实选取算法：找出给定KG事实2-hops以内的事实

候选事实排序模型：



Neural Fact Contextualization Method

知识获取：Taxonomy补全

Enriching Taxonomies With Functional Domain Knowledge, SIGIR18

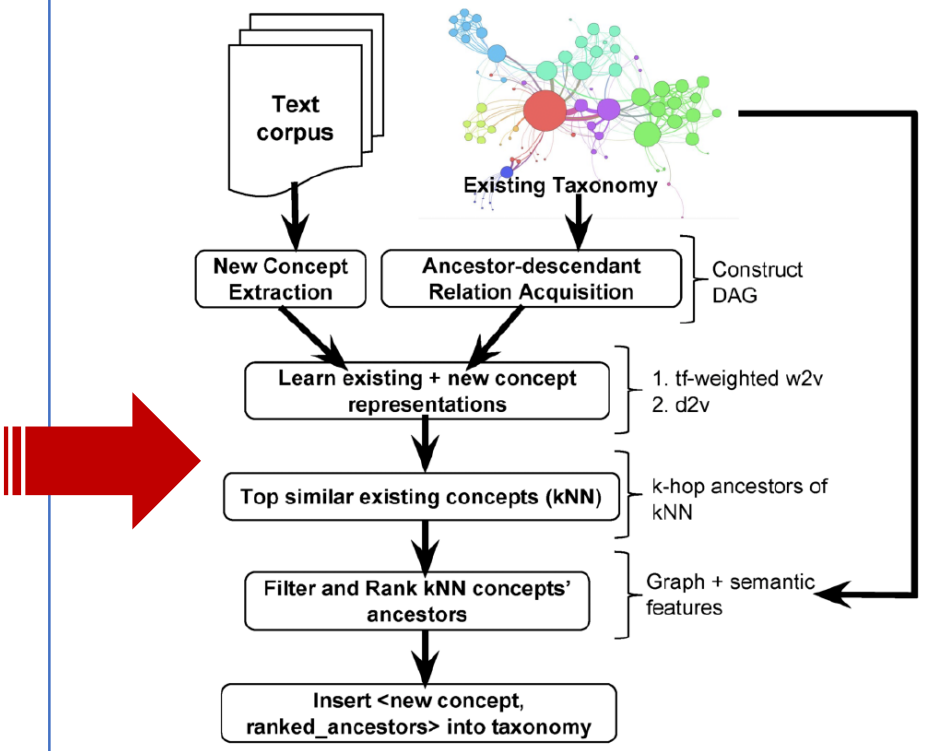
研究问题：知识结构扩充，即对于大量新出现的concepts，如何将其加入已有的知识结构中

存在的挑战：

- 未知concept检测
- 新concept插入已有的知识结构，如何保证新创建的关系的语义完整性

已有工作的不足：

- 受语言限制
- 受领域限制
- 无法用于大规模的知识结构扩充



- 生成concepts的embeddings （w2v + d2v）
- 寻找与新concept相似的K个已有concepts
- 基于图特征和语义特征的learning to rank 排序，确定新concept合适的插入位置（父节点）

知识获取：三元组规范化

CESI: canonicalizing open knowledge bases using embeddings and side information, WWW18

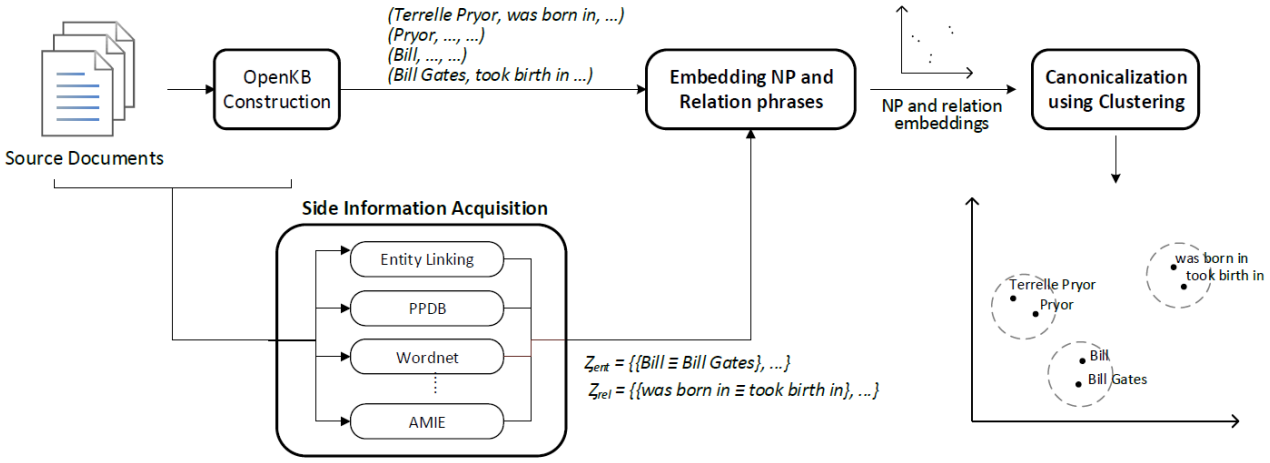
研究问题： 信息抽取，openIE抽取出的三元组规范化

已有工作的不足：

- 现有方法主要通过人工定义的特征将三元组进行聚类，再进行规范化。
- 人工方法进行特征工程代价巨大，次优

解决方案：

- 使用embedding方法学习openIE抽取出的三元组的语义表示
- 在Embedding下的语义空间进行聚类



趋势2（知识挖掘）：具有应用价值的新型知识挖掘问题受到关注

知识挖掘：必有属性挖掘

Are all people married? Determining obligatory attributes in knowledge bases, WWW18

研究问题： 找到知识图谱中概念的必有属性

挑战：

- 知识库中存在数十万的概念，难以利用人工来判断哪些属性是概念必有的
- 开放世界假设，使得无法判断一条没有包含在知识库中的三元组的真假

解决方案：

- **提出了基于概念层次结构来推断概念的必有属性**
- **基本假设**
  - 假设知识库的不完全性在知识库的所有类中都是均匀分布的。如果一个属性在某个概念中分布较稀疏（而在其他概念中分布较密），则可推断它一定不是分布较为稀疏的概念的必有属性

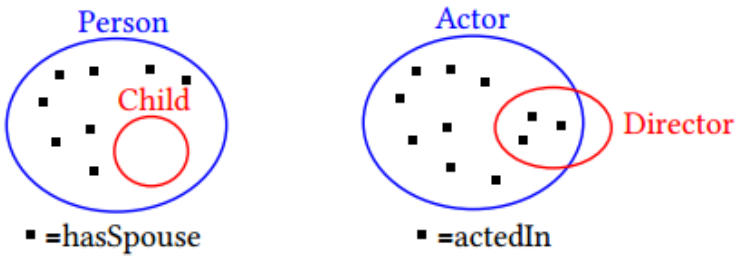


Figure 1: Examples of attributes and classes.

知识挖掘：分面标注

Facet annotation using reference knowledge bases, WWW18

研究问题： 概念的分面（Facet）属性自动标注

已有工作的不足：

- 现有方法只是找到了概念的某个分面属性的所有值，但没有将这个分面属性给标注出来

解决方案：

- 利用知识图谱中三元组的谓词来作为分面的属性名
- 定义了三个度量函数来评估分面属性和谓词的相似度
  - specificity, coverage, and frequency

Figure 1: Facets that characterize books.

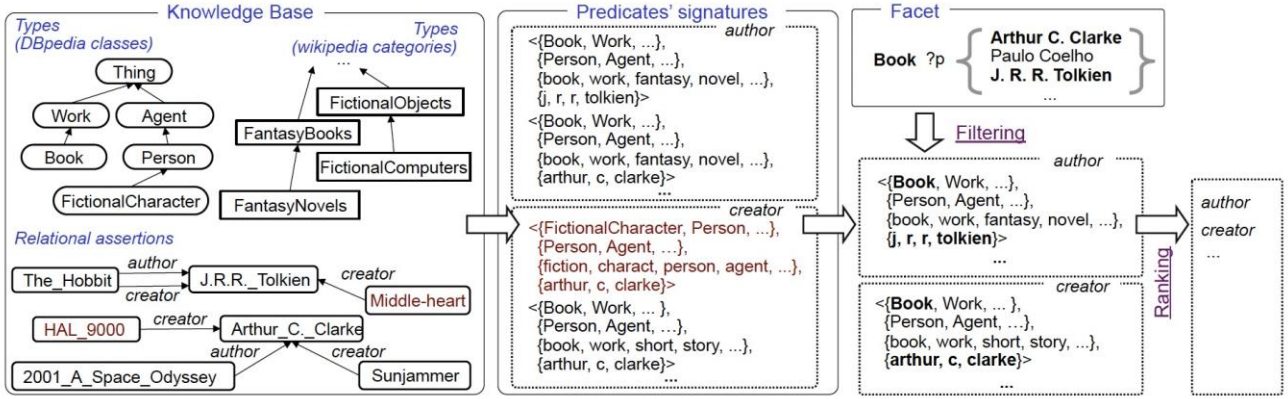


Figure 3: The Facet Annotation process.

趋势3（知识应用）：利用知识图谱等各种背景知识增强数据的描述，显著提升推荐、问答的效果，成为主流趋势

知识应用：会话系统

Response Ranking with Deep Matching Networks and External Knowledge in Information-seeking Conversation Systems, SIGIR18

研究问题：检索式会话系统中的答案排序

已有工作的不足：大多注重用户输入信息和候选答案之间的匹配模式，而忽略了会话所涉及的外部知识

将外部知识融入深度神经网络用于会话排序

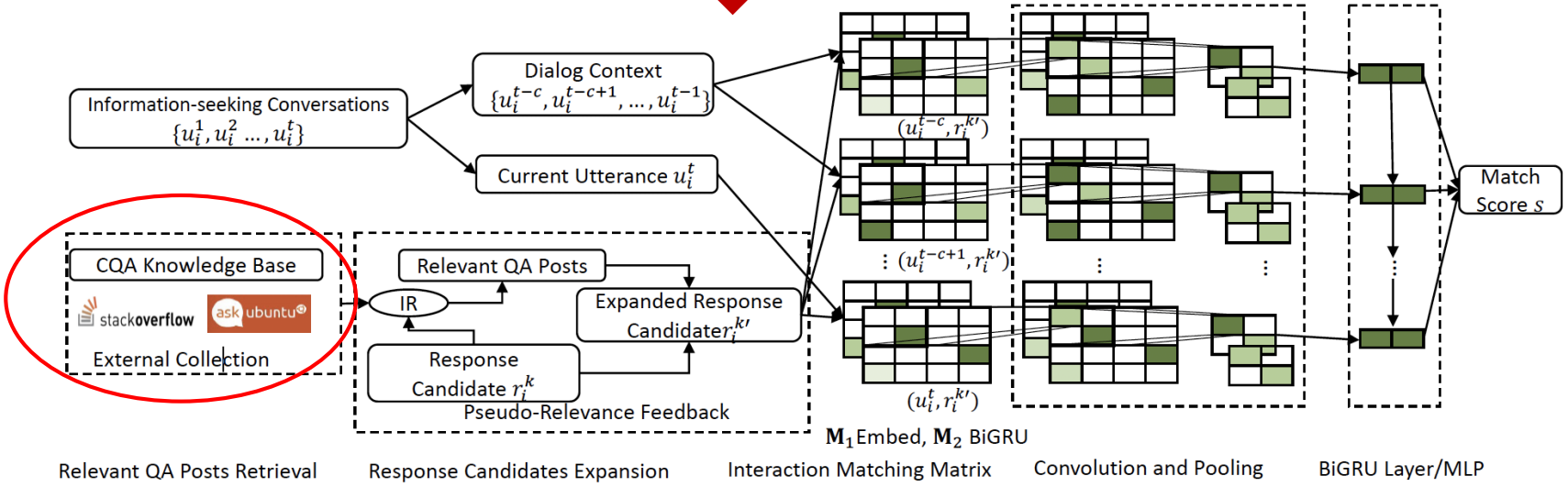


Figure 1: The architecture of DMN-PRF model for conversation response ranking.

知识应用：问答系统

Knowledge-aware Attentive Neural Network for Ranking Question Answer Pairs, SIGIR18

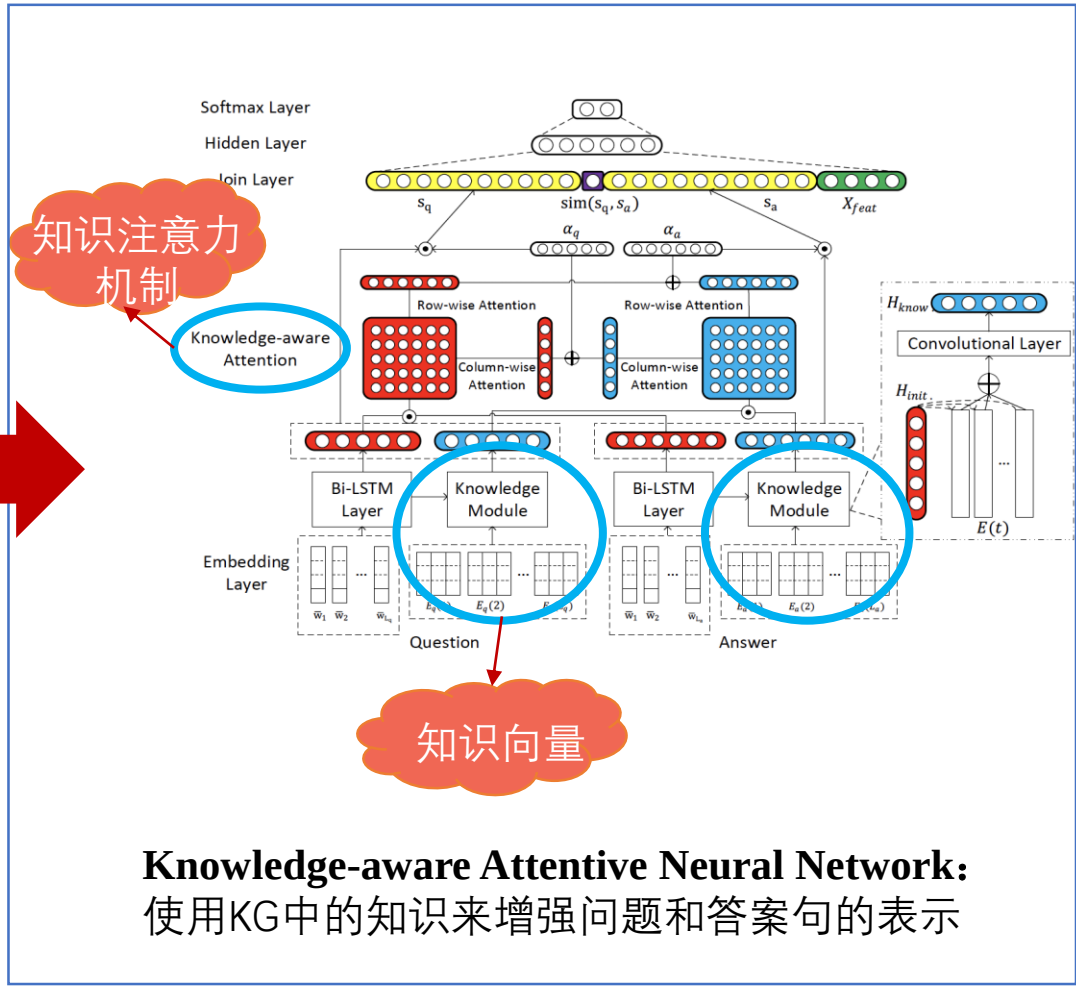
研究问题：问答系统中的答案排序

已有工作的不足：目前多基于神经网络对问题和答案句建模，但忽略了问题的背景信息和隐含的关系

Table 1: Example of QA candidate pairs.

Question	When was ***Pokemon*** first started ?
Positive Answer	Is a media franchise published and owned by Japanese video game company ***Nintendo*** and created by ***Satoshi Tajiri*** in ***1996*** .
Negative Answer	The official logo of ***Pokemon*** for its international release ; "***Pokemon***" is short for the original Japanese title of "***Pocket Monsters***" .

Score (Negative answer) > Score (positive answer)



Knowledge-aware Attentive Neural Network：
使用KG中的知识来增强问题和答案句的表示

知识应用：问答系统

Never-ending learning for open-domain question answering over knowledge bases, WWW18

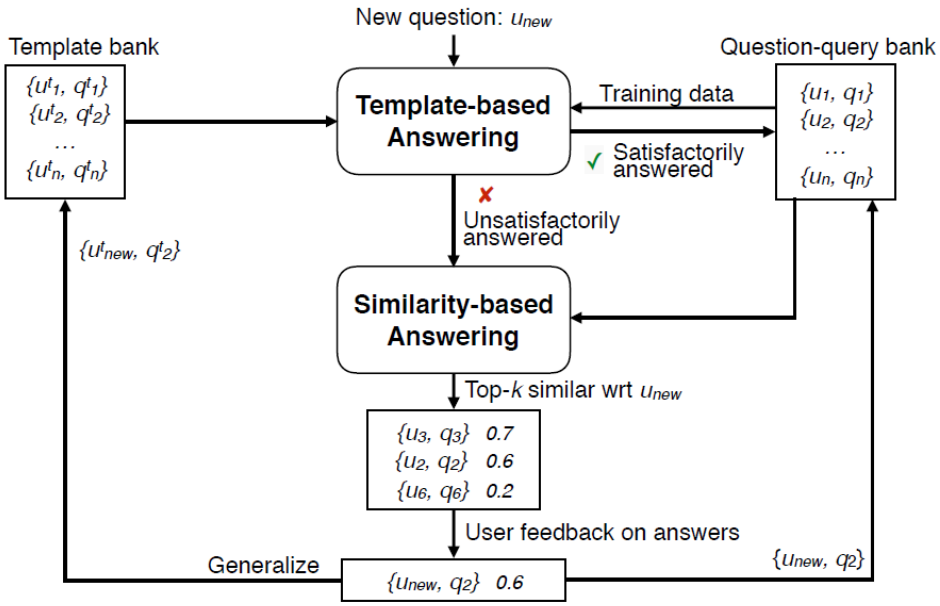
研究问题： 在开放域的KBQA中，将自然语言问题转换为语义表示（如SPARQL）

已有工作的不足：

- 现有工作将训练过程（离线）和问答过程（在线）严格分开，方法存在两类不足：
  - 需要大量的标记数据集，但这些数据集很多时候并没有现成的
  - 无法对训练集没有覆盖的领域进行有效回答

解决方案：

- 提出了一种持续学习的KBQA方法：
  - 在离线训练过程中，从少量训练数据集中自动学习出模板
  - 在在线问答过程中，当遇到模板未覆盖的问题时，出发持续学习模型



知识应用：推荐系统

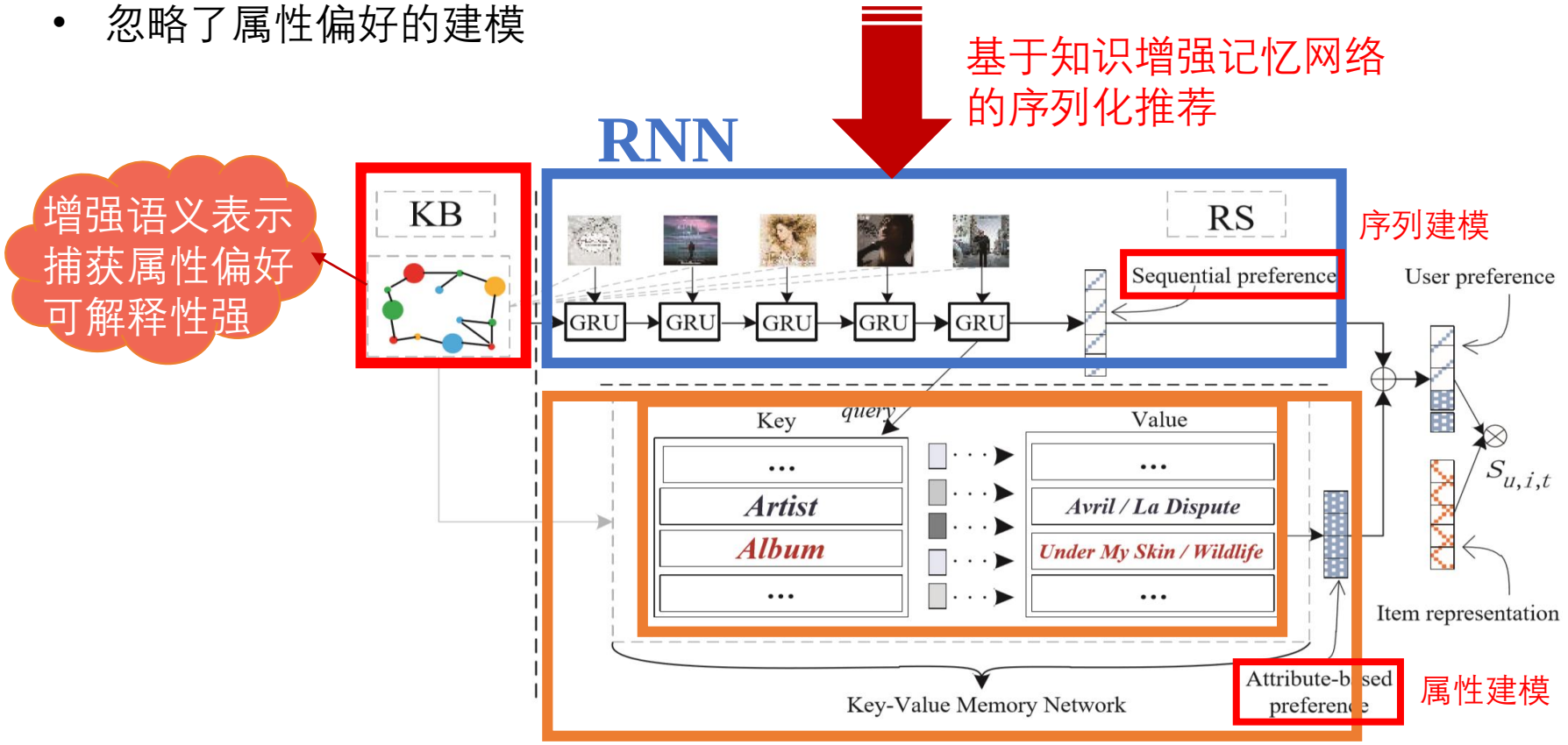
Improving Sequential Recommendation with Knowledge-enhanced Memory Networks, SIGIR18

研究问题：推荐系统中的序列化推荐

已有工作的不足：

- 大多基于RNN，虽然能在一定程度上捕获序列依赖，但难以记忆和维护长期数据
- 忽略了属性偏好的建模

基于知识增强记忆网络的序列化推荐



知识应用：推荐系统

DKN: deep knowledge-aware network for news recommendation, WWW18

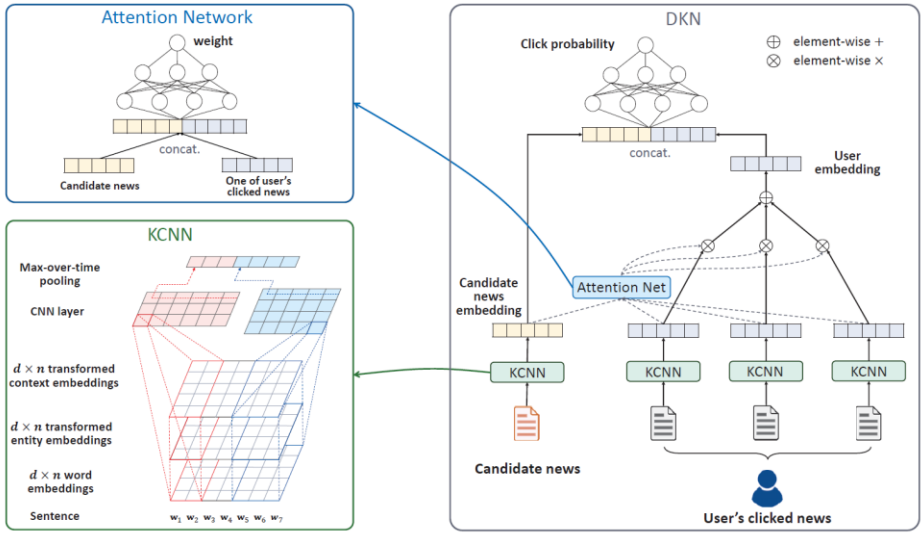
研究问题： 新闻推荐

已有工作的不足：
新闻标题和正文中通常存在大量的实体，实体间的语义关系可以有效地扩展用户兴趣。然而这种语义关系难以被传统方法（话题模型、词向量）发掘

解决方案：

- 使用知识图谱特征学习得到实体向量和关系向量，然后将这些低维向量引入推荐系统，学习得到用户向量和物品向量

利用CNN提取词、实体以及实体上下文中的特征



趋势4（知识评估）：实用化知识评估、针对开放世界假设的知识评估受到关注

知识评估：补全评估

On Link Prediction in Knowledge Bases: Max-K Criterion and Prediction Protocols, SIGIR18

研究问题：KG 补全中的评价指标

（*h, r, ?*）

↓

该返回多少个答案实体

↓

采取Top-K criterion ?

- 不合理，不同任务的答案个数不同
- 由于任务的差异性，全局的K难以设置

→

Max-K criterion

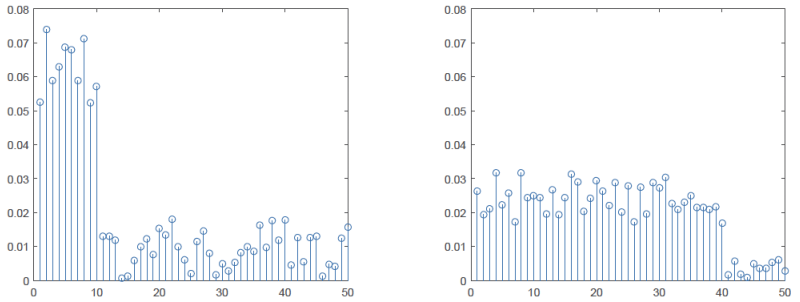


Figure 1: Two example answer predicting distributions.

左图：实体 **1 -10**在预测的概率分布上起统治作用。
右图：没有任何实体在预测的概率分布上起统治作用，但实体**1-40**的概率比其他要大。

- **返回max K个答案来评估模型的性能**
- **对于不同任务更合理，更个性化**

知识评估：规则评估

Estimating rule quality for knowledge base completion with the relationship between coverage assumption, WWW18

研究问题： 知识图谱补全，包括产生新的事实或规则

挑战：

- 无法判断一条没有包含在知识库中的三元组的真假
- 知识图谱中只存在正例，不存在负例

解决方案：

- 提出了一个打分函数来评估从知识库中学习出的一阶规则的质量
- 在估计一个规则的质量时，考虑了不在知识库中的三元组的信息

$$\frac{supp(B \Rightarrow R) + |UP_{B \Rightarrow R}|}{|\mathbb{P}_{B \Rightarrow R}|}$$

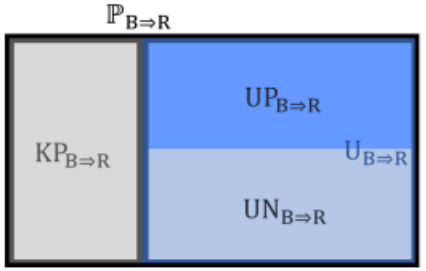


Figure 1: The set of predictions $\mathbb{P}_{B \Rightarrow R}$ can be divided into labeled ($KP_{B \Rightarrow R}$) and unlabeled ($U_{B \Rightarrow R}$) examples. Furthermore, $U_{B \Rightarrow R}$ can be subdivided into the unknown positives ($UP_{B \Rightarrow R}$) and unknown negatives ($UN_{B \Rightarrow R}$).

- 从两个角度对未知信息进行估计：
  - 一个规则覆盖的正例比率在已知KB和未知KB上是相同的
  - 一条关系的所有未标记数据中，有多少比例是正确的

总结

- 深度学习与知识图谱的深度融合已经成为普遍趋势
- 数据驱动已经成为大数据时代知识工程的主要手段
- 知识引导已经成为突破应用瓶颈的重要思路之一

- **以知识图谱为代表的大数据知识工程方兴未艾**