DKN: Deep Knowledge-Aware Network for News Recommendation

结合知识图谱中的表示学习，加上神经网络，进行新闻推荐。

Accepted by WWW'18，论文链接：https://arxiv.org/abs/1801.08284

摘要

在线新闻推荐系统主要解决新闻信息过载的问题，同时尽可能提升用户的使用体验，注重个性化。一般来说，新闻的语言是高度凝练的，有许多命名实体，并且包含了很多的常识。目前现有的一些推荐算法不关注、或者说无法理解新闻中的一些"知识"以及我们所了解的常识，仅仅通过一些简单的模式匹配、主题相关进行机械的推荐，不具有理论上的扩展性。此外，新闻具有时效性的特征，用户的口味也会随着时间变化，因此本文提出了 DKN 模型，将知识图谱表示融入新闻推荐中。DKN 是一个预测点击率的基于内容的深度推荐模型。

DKN 的核心是一个多通道、命名实体对齐的、融合了知识的卷积神经网络（KCNN），从语义和知识层面上来表示新闻。KCNN 把新闻中的词和实体作为通道（channel），并且在卷积过程中显式地监督两者的对齐关系。为了处理用户喜好的多样性，加入了 Attention 动态地对用户阅读历史进行加权，选择候选新闻。在大量的相关实验中，取得了最好的结果。

简介

现在人们阅读新闻的习惯已经从传统的纸媒到电视，更多的则是互联网。各大新闻网站中的新闻数量庞大，如何对用户进行推荐则是一个难题。和传统的通用推荐系统对比，新闻推荐更为突出的主要有以下三个挑战：

1. 新闻具有很强的时间敏感性，容易过期，过时的新闻很快就会被更新的新闻所取代。从而传统基于 ID 的协同过滤方法就很有限了；
2. 用户在阅读新闻的时候是话题敏感的，通常对多个话题都具有倾向性。如何从多种多样的阅读记录中，对用户的多种主题偏好信息进行表示，是新闻推荐系统做好的关键；
3. 新闻文本语言高度凝练，还包含了大量的知识实体和常识，而且用户更倾向于去读那些具有相关命名实体（如人名）的其他新闻。

经典的语义模型或者主题模型，都是从词的共现信息或词聚类结构上挖掘新闻之间的关系，仅仅抓住了语义信息，从而给用户的新闻推荐就会变窄，局限在一个话题中。而本文提出的 DKN 能够从中挖掘新闻之间的潜在知识层面上的联系，引入知识图谱中的信息，是一种十分合理的扩展。

传统的协同过滤方法不同的是，DKN 是一种基于内容的 CTR（点击率）预测模型：给定一个候选新闻和用户之前的浏览历史，预测用户点击候选新闻的概率。在 DKN 中的主要步骤：

- 对新闻中的每个词都在知识图谱中找到对应的实体，使用他的邻居实体来增强新闻的知识层面的信息。
- 通过 KCNN（Knowledge-aware CNN）把新闻词语和知识层面上信息表示为一个 embedding，KCNN 与其他相关工作的不同之处有：
    - KCNN 是多个通道的，新闻输入与图像的 RGB 通道类似，这里的通道包含了新闻中词的

embedding，实体的 embedding 及相关实体的 embedding；
    - 词语-实体对齐信息，把一个词和对应的实体在多个通道内进行对齐，通过某种转换函数（映射）来消除词向量和实体向量空间的异构性。或者可以这样理解，KCNN 保证了多个通道内词语的表示的一致性，并且显式地减少不同 embedding 空间的隔阂。
- 通过 KCNN，得到的新闻 embedding，与用户点击过的新闻通过 attention，加权平均得到一个用户的表示，最后通过 DNN 来计算候选新闻被用户点击的概率

最终，本文提出的 DKN 模型在 Bing News 推荐上得到了显著的性能提升。

背景知识

知识图谱表示（KGE）

典型的知识图谱是由许多的三元组 (h, r, t) 构成的，h 表示头实体，t 表示尾实体，r 表示实体之间的关系。知识图谱表示完成的任务就是把实体和关系用一个低维的、稠密的向量表示出来，能够最大程度的保留原本的结构化和语义信息。目前有多种 translation-based 方法，如 TransE、 TransH、TransR 和 TransD，这些模型对 h、t、r 都有各自的处理，大多数是映射到相同的空间，对某一种关系进行正负样本采样，基于正样本的要优于（随机或其他更为巧妙方式采样得到的）负样本的要求，来优化求解实体和关系的 embedding。

CNN 文本表示

传统方法是通过 BOW 来表示的，但是不能包含句子的序列化信息，并且得到的特征向量十分稀疏，所以现在通常的做法是得到一个低维的句子分布式表示向量。CNN 在图像处理上的成功，也可以用在文本中。本文采用的 CNN 是 Kim CNN，在一个长度为 n 的句子 $[w_1, w_2, \cdots, w_n]$ 上，具体操作如下：

- $w_{1:n}$ 是一个 $d \times n$ 的矩阵，d 是词向量的维数
- 卷积操作通过不同的窗口大小（l）的若干个卷积核 $h \in \mathbb{R}^{d \times l}$，得到每一个 filter 下的表示 c_i
    - 在一个确定的卷积核下，$c_i = f(h * w_{i:i+l-1} + b)$，这里 f 是一个非线性函数，$*$ 代表卷积，b 是偏置；
    - 最后可以得到 $n - l + 1$ 个特征，组成一个 Feature Map：$c = [c_1, c_2, \cdots, c_{n-l+1}]$；
    - 在上面加一层 max-over-time pooling，得到该卷积核下的特征
- 通过多个不同大小的卷积核可以得到多个视角下的特征表示，全部拼接在一起，就可以得到最终的文本表示

问题定义

给定一个用户 user i，他的点击历史是 $\{t_1^i, t_2^i, \cdots, t_{N_i}^i\}$，这里的 t_j^i 表示的是新闻的标题，每一个标题都是一个词语的序列，$t = [w_1, w_2, \cdots]$，其中每个词都可能与知识图谱中的某个实体对应。具体来说，对于新闻标题：*Trump praises Las Vegas medical team*，Trump 和 "Donald Trump" 这个人物实体对应，Las 和 Vegas 与 "Las Vegas" 这个地理实体对应。本文的任务即为根据这些输入信息（点击历史，新闻标题和知识图谱中对应的实体），预测该用户对特定的候选新闻是否会点击。

Deep Knowledge-aware Network

DKN 架构

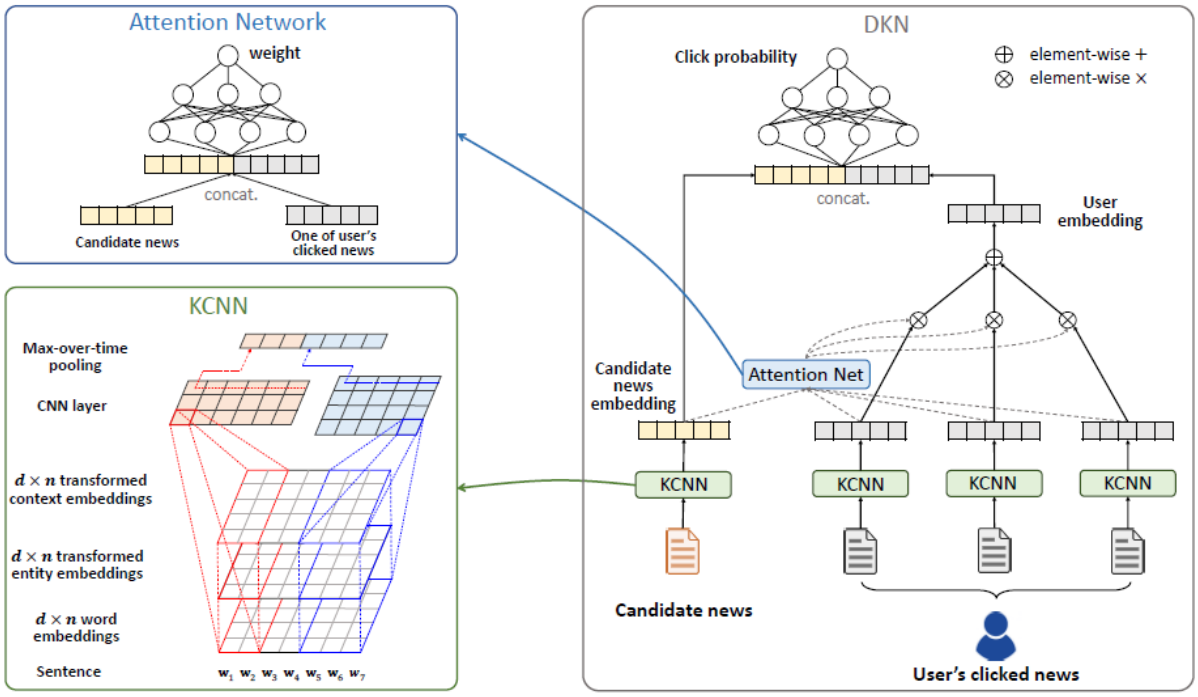

Figure 3: Illustration of the DKN framework.

DKN 的输入包括候选新闻和用户点击过的新闻（标题），通过 KCNN 提取特征得到 embedding 表示，相较传统的 CNN，可以融合知识图谱中的信息，来得到一个更好的句子表示。对用户点击的新闻（标题）如法炮制，得到历史新闻的 embedding，结合 Attention 网络，计算权重后加权，得到用户的 embedding。将候选新闻的 embedding 和用户 embedding 拼在一起，通过一个深度网络计算点击的概率。

知识提取（升华）

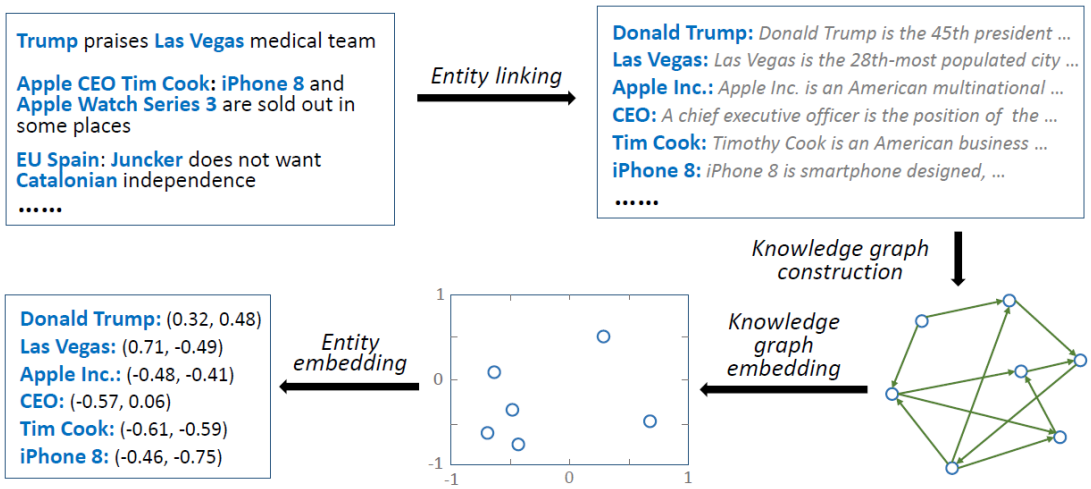

Figure 4: Illustration of knowledge distillation process.

四个步骤：

1. 实体链接：将新闻标题中的文本与知识图谱中的实体对应，并且进行消歧；
2. 利用识别出来的实体和关系，从原本知识图谱中得到一个子图。考虑到仅仅使用这些识别出的实体和他们之间的关系，知识子图会很稀疏，所以还加入与这些实体有 one-hop 的连接的实体和对应的关系；

3. 在构建的子知识图谱上利用一些学习方法得到实体的表示；
4. 从实体的表示上，得到对应单词的词向量。

注意到，现有的知识图谱表示学习方法，保留了绝大多数的是结构信息，很多实体相关的语义信息很有限，对于推荐系统来说是很大的损失，所以还利用了额外的上下文信息，即实体可以通过它的 one-hop 邻居实体节点的平均来表示。用 e 表示实体的 embedding，$\bar{e}$ 表示上下文实体的平均。

KCNN

在 KCNN 的输入中，新闻标题的表示包括原有词的 embedding 和识别出来的实体的 embedding，作者认为简单的拼接，即句子表示为 $[w_1, w_2, \cdots, w_n, e_{t_1}, e_{t_2}, \cdots]$ 太过于粗鲁，有以下三点问题：

- 拼接丢失了词语与对应实体之间的对应信息
- 词向量和实体向量不是通过一个统一的模型得到的，拼接后在一个语义空间上进行卷积是不合适的
- 词向量和实体向量必须有相同的维数，在实际情况中，两者有各自的最优的维数，不一定相同

本文提出的解决方案是在 KCNN zhong 利用多通道（multi-channel）和词语-实体对齐（word-entity-alignment）。

对于每一篇新闻的标题 $t = [w_1, w_2, \cdots, w_n]$，有一个与之对应的实体的表示 $g(e_{1:n}) = [g(e_1), g(e_2), \cdots, g(e_n)]$ 和相关的上下文实体 $g(\bar{e}_{1:n}) = [g(\bar{e}_1), g(\bar{e}_2), \cdots, g(\bar{e}_n)]$，其中的 $g(e) = Me$，M 为一个映射矩阵，或者加上一个非线性层，即 $g(e) = \tanh(Me + b)$。如果某个词没有对应的实体，就设置为 0 向量。

最后得到 3 个长度为 n 的词（实体）向量表示矩阵，在此基础上模拟图像中的 RGB 3 通道处理方式，整个的输入组合成一个三维张量，第三维（通道）第一个是词语层面，第二个是实体层面，第三个是相邻实体层面的表示。最后用若干个卷积核加上一层 max-pooling，就可以得到新闻标题的表示 $e(t) = [\tilde{c}^{h_1}, \tilde{c}^{h_2}, \cdots, \tilde{c}^{h_m}]$，将 m 个卷积核得到的向量再拼接得到。

Attention-based 用户兴趣提取

在有了用户点击过的历史新闻的表示后，简单的做法是把这些新闻的表示作平均作为用户的表示。但是用户的喜好是不断变化的，对于某一篇新闻的点击行为对于不同的候选新闻的影响是不同的，这里就引入了注意力机制来处理。对于用户点击过的新闻 t_k^i 和候选新闻 t_j 通过一个 DNN $\mathcal{H}$ 和 softmax 层计算权重后，再加权，是一种典型的 MLP Attention 方式，得到用户基于注意力机制的表示 $e(i)$。

据此得到的用户表示后，又用了一个 DNN $\mathcal{G}$ 来给出用户 i 点击新闻 t_j 的预测概率：
$p_{i,t_j} = \mathcal{G}(e(i), e(t_j))$

实验

数据集来自 Bing News，时间跨度为 2016 年 10 月 16 日到 2017 年 6 月 11 日作为训练集，6 月 12 日到 8 月 11 日的为测试集，用的知识图谱是微软的 Satori。值得一提的是，在做过数据分析后，有几点发现：

- 大多数新闻的寿命基本上不超过 2 天
- 新闻标题中，平均包含 3.7 个知识图谱实体

基线实验对比

评价指标选择的是 F1 和 AUC，除了 LibFM 其他的都是深度模型，除了 DMF，其他都是基于内容的过滤方法。

Table 2: Comparison of different models.

Models*	F1	AUC	p-value**
DKN	**68.9 ± 1.5**	**65.9 ± 1.2**	–
LibFM	61.8 ± 2.1 (-10.3%)	59.7 ± 1.8 (-9.4%)	$< 10^{-3}$
LibFM(-)	61.1 ± 1.9 (-11.3%)	58.9 ± 1.7 (-10.6%)	$< 10^{-3}$
KPCNN	67.0 ± 1.6 (-2.8%)	64.2 ± 1.4 (-2.6%)	0.098
KPCNN(-)	65.8 ± 1.4 (-4.5%)	63.1 ± 1.5 (-4.2%)	0.036
DSSM	66.7 ± 1.8 (-3.2%)	63.6 ± 2.0 (-3.5%)	0.063
DSSM(-)	66.1 ± 1.6 (-4.1%)	63.2 ± 1.8 (-4.1%)	0.045
DeepWide	66.0 ±1.2 (-4.2%)	63.3 ± 1.5 (-3.9%)	0.039
DeepWide(-)	63.7 ± 0.9 (-7.5%)	61.5 ± 1.1 (-6.7%)	0.004
DeepFM	63.8 ± 1.5 (-7.4%)	61.2 ± 2.3 (-7.1%)	0.014
DeepFM(-)	64.0 ± 1.9 (-7.1%)	61.1 ± 1.8 (-7.3%)	0.007
YouTubeNet	65.5 ± 1.2 (-4.9%)	63.0 ± 1.4 (-4.4%)	0.025
YouTubeNet(-)	65.1 ± 0.7 (-5.5%)	62.1 ± 1.3 (-5.8%)	0.011
DMF	57.2 ± 1.2 (-17.0%)	55.3 ± 1.0 (-16.1%)	$< 10^{-3}$

* "(-)" denotes "without input of entity embeddings".

** p-value is the probability of no significant difference with DKN on AUC by t-test.

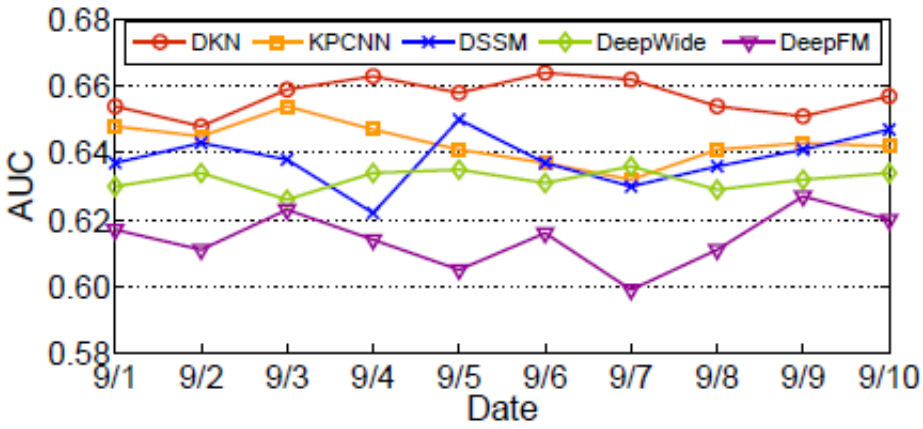

Figure 7: AUC score of DKN and baselines over ten days (Sep. 01-10, 2017).

实验结果分析

为了体现新闻中实体的重要性，在模型方法中，以 "(-)" 表示不使用知识图谱信息的情况。可以看到：

- 大部分基线实验在融合了实体信息后，都有性能的提升
- DMF 的性能较差，因为它是基于用户协同过滤的方法，没有考虑到新闻的时效性特征
- 深度学习模型都要比 LibFM 好一些，推断是因为它更能抓住数据之间一些非线性的相关性
- DSSM 比 DeepWide 和 YouTubeNet 要好一些，可能是因为对词做了 hash
- KPCNN 是基线实验中最好的一组，是因为它用 CNN 能从新闻标题中提取出一些文本模式和语义信息

- DKN 结果最好，因为它首先使用了词-实体对齐的机制，还通过注意力机制对用户的点击历史动态建模
- DKN 的结果比较稳定，方差较小

DKN 上的各种改进：

- 使用实体及其相关实体的 embedding，能提升模型性能
- 使用 DKN + TransD 学到的实体表示，是最好的结果，因为 TransD 是其中最复杂、最精致的知识图谱表示学习模型
- 实体 embedding 的学习过程中，加入一层映射和 tanh 层，能使模型自适应学习这种变换，增强模型的对新闻之间关系的建模能力

Case Study

Table 4: Illustration of training and test logs for a randomly sampled user (training logs with label 0 are omitted).

	No.	Date	News title	Entities	Label	Category
training	1	12/25/2016	Elon Musk teases huge upgrades for Tesla's supercharger network	Elon Musk; Tesla Inc.	1	Cars
	2	03/25/2017	Elon Musk offers Tesla Model 3 sneak peek	Elon Musk; Tesla Model 3	1	Cars
	3	12/14/2016	Google fumbles while Tesla sprints toward a driverless future	Google Inc.; Tesla Inc.	1	Cars
	4	12/15/2016	Trump pledges aid to Silicon Valley during tech meeting	Donald Trump; Silicon Valley	1	Politics
	5	03/26/2017	Donald Trump is a big reason why the GOP kept the Montana House seat	Donald Trump; GOP; Montana	1	Politics
	6	05/03/2017	North Korea threat: Kim could use nuclear weapons as "blackmail"	North Korea; Kim Jong-un	1	Politics
	7	12/22/2016	Microsoft sells out of unlocked Lumia 950 and Lumia 950 XL in the US	Microsoft; Lumia; United States	1	Other
	8	12/08/2017	6.5 magnitude earthquake recorded off the coast of California	earthquake; California	1	Other
			.....			
test	1	07/08/2017	Tesla makes its first Model 3	Tesla Inc; Tesla Model 3	1	Cars
	2	08/13/2017	General Motors is ramping up its self-driving car: Ford should be nervous	General Motors; Ford Inc.	1	Cars
	3	06/21/2017	Jeh Johnson testifies on Russian interference in 2016 election	Jeh Johnson; Russian	1	Politics
	4	07/16/2017	"Game of Thrones" season 7 premiere: how you can watch	Game of Thrones	0	Other

通过一个用户的浏览数据发现，用户点击新闻很大程度上选择一些相关的关键词、实体的其他新闻，也会有些相关的主题信息，这些在 DKN 模型都是考虑到的因素。用户曾经读过关于 Tesla 汽车的新闻，但能够预测出会阅读 Ford 相关的新闻，是依赖于知识图谱中的相邻实体的信息的。通过知识图谱还找到了 Trump 相关的 Jeh Johnson，Russian 相关的新闻。通过注意力机制，可以很好地区分用户在汽车领域、政治领域的不同的喜好变化和对当前推荐新闻的关系。（训练集中的第 2 条新闻和测试集中的第 1 条新闻，训练集中的第 4、5 条新闻和测试集中的第 3 条之间都有比较强的 attention 权重）

参数敏感性分析

大多数参数都比较常规，设置过大的 embedding 维数容易引入噪声导致过拟合，过多的卷积核也会带来过拟合的问题，所以合适就好。

结论

本文是深度学习技术，在推荐系统领域的一个发展，重点放在了新闻推荐上，还融合了知识图谱表示的方法。对于新闻推荐中存在的问题和特点：新闻具有时效性和较多的实体，有针对性地提出了 DKN 模型，解决了三个挑战：

- DKN 是一个基于内容过滤的深度推荐系统模型；
- 为了利用知识图谱中的信息，通过 KCNN 来融合文本的语义层面、实体层面上的异构表示；
- 使用了注意力机制对用户的兴趣进行动态提取。

既然如此，能不能在一些 Baseline 实验上进行一些改进，也融合知识图谱、用户社交网络的信息呢？新闻的时效性是不是也可以通过 RNN 来处理，尤其是用户兴趣提取这块，用 RNN + 注意力机制可行吗？