

一种基于知识图谱与内容的推荐算法

张雅歌, 胡雪若白, 黄 洁, 张 克

战略支援部队信息工程大学, 河南 郑州

收稿日期: 2022年2月20日; 录用日期: 2022年3月15日; 发布日期: 2022年3月22日

摘 要

推荐系统作为信息筛选工具, 通过对用户个人信息记录分析以实现用户兴趣预测, 推荐满足用户需求的物品列表, 其核心是推荐算法。现有的推荐算法大多需要收集用户大规模的数据训练来实现, 存在数据稀疏性的问题, 且当用户需要购买新物品时, 往往无法进行有效的推荐。针对该问题, 本文提出了一种基于知识图谱和内容的推荐算法。首先, 解析用户记录并构建知识图谱, 设计基于知识图谱的相似度算法, 对知识图谱中的实体与关系进行相似度计算; 其次, 基于文本词嵌入对内容进行向量化表示, 并构建基于内容的相似度算法; 最后, 本文融合了基于知识图谱的相似度算法和基于内容的相似度算法, 通过权重计算选取得分高的值作为推荐结果。与其它模型在MovieLens数据集上进行了实验对比, 实验结果表明, 本文所提出的算法在召回率和准确率方面有明显的提升。

关键词

知识图谱, 内容, 推荐算法

A Recommendation Algorithm Based on Knowledge Graph and Content

Yage Zhang, Ruobai Huxue, Jie Huang, Ke Zhang

PLA Strategic Support Force Information Engineering University, Zhengzhou Henan

Received: Feb. 20th, 2022; accepted: Mar. 15th, 2022; published: Mar. 22nd, 2022

Abstract

As an information screening tool, the recommendation system can predict users' interests and recommend the list of items that meet users' needs through the analysis of users' personal information records. Its core is the recommendation algorithm. Most of the existing recommendation algorithms need to collect users' large-scale data for training, which has the problem of data

sparseness. When a user wants to buy new items, it is often unable to make effective recommendation. In this paper, we propose a novel recommendation algorithm based on knowledge graph and content. Firstly, a content graph was constructed to obtain the similarity between entities and relationships based on graph embedding. Secondly, the content was vectorized based on text word embedding, and a content-based similarity algorithm was constructed. Finally, a fusion method combined graph-based similarity method and content-based similarity method was proposed, and then the entities with high scores would be selected as the recommendation result through weight estimation. The result of comparative experiments on the MovieLens dataset shows that the algorithm proposed has improved the precision and recall on performance.

Keywords

Knowledge Graph, Content, Recommendation Algorithm

Copyright © 2022 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

随着互联网的发展,人们在享受网络便利的同时,也越来越受到海量信息的影响,无法快速搜寻到满足自己真正需求的内容。推荐系统应运而生,其是智能学习的一个研究领域,目标是根据用户的喜好给用户有效的推荐结果。推荐系统可以通过物品的内容信息,根据用户的历史偏好数据、用户个人兴趣、网络社交信息、所在环境等信息,判断用户目前最需要或者最感兴趣的内容(物品或服务) [1],已经广泛应用于商务网站,资讯和旅游网站等[2]。

推荐系统的核心是推荐算法的研究,目前主流的推荐算法主要有三类,包括协同过滤推荐算法、基于内容的推荐算法以及基于深度学习的推荐算法。协同过滤推荐算法的核心思想就是充分利用与目标用户兴趣偏好相同的用户群体喜好来进行高精度推荐[3]。文献[4]中介绍了一个成熟的推荐系统 LibFM (Lithium Bis Fluoromalonato),该系统依赖于协同过滤算法,利用邻域模型分析用户与物品特征之间的相似度,为用户提供推荐服务。但是由于协同过滤推荐算法主要是根据用户的历史行为数据发掘偏好,因此推荐质量取决于用户的历史数据集,存在数据稀疏性问题。基于内容的推荐算法是将用户偏好与项目内容特征进行匹配,进而为用户提供推荐结果[5],其原理就是通过对目标用户的中意项目元数据,根据内容相似程度为目标用户产生新的推荐。因此较为直观,应用较为广泛,但其对于复杂属性的项目较难处理。随着神经网络的发展,基于深度学习的推荐算法也越来越受到青睐。文献[6]中提出了图协同神经过滤算法,从卷积神经网络出发,提取物品特征,聚合多邻数特征以构建物品向量,通过提升物品相似度的计算结果来提升推荐性能。Sharma 等人[7]专注于提高基于神经网络技术的推荐系统的有效性,通过多个推荐方法的排列组合以进行推荐。文献[8]提出了深度知识注意力网络(Deep Knowledge-Aware Networks, DKN),融合了实体嵌入和词嵌入,其中实体嵌入采用 TransH (Translating on Hyperplanes)学习,但是 DKN 对实体的表征不明确,无法为个性化推荐提供有用的信息。与传统推荐方法相比,基于神经网络的推荐算法无需人工提取特征,具有处理复杂属性项目的的能力。但其需要大量的数据作为支撑,无法解决数据稀疏性问题。

随着知识图谱的出现,将知识图谱作为辅助信息引入到推荐系统中,通过知识图谱中实体之间丰富的关系能丰富项目和项目的联系、增强推荐算法的挖掘能力,从而有效弥补数据稀疏性的问题[9]。文献

[10]中利用通过异构信息网络(Heterogeneous Information Network, HIN)学习实体之间的路径信息,能够利用任意用户指定的实体连接路径,并返回相似度最高的实体。但是由于用户定义的元路径很难达到最佳,所以当近邻数达到一定程度时,其推荐性能作用有限。

综上所述,单一的推荐算法存在或多或少的缺陷,譬如数据稀疏性问题、处理复杂属性困难等问题,尤其是当用户需要购买新物品时,其推荐效果不佳。为了解决这个问题,本文提出了一种基于知识图谱与内容的推荐算法。首先,解析用户记录并构建知识图谱,基于图谱嵌入计算得到实体与关系的相似度;其次,基于文本词嵌入对内容进行向量化表示,并构建算法对内容进行相似度计算;最后,本文融合了基于知识图谱的相似度算法和基于内容的相似度算法,通过权重计算选取得分高的值作为推荐结果。通过7折交叉验证的方法表明,在一个完备的推荐系统中,近邻数的选取并不是越大越好,而是应该进行适当的选择。通过对比验证的方法表明,在相同近邻数的情况下,本文提出的融合算法具有较高的召回率和准确率,这是由于本文的算法在基于内容的推荐算法基础上,利用知识图谱实体之间的多关系和属性进行了融合相似度计算,提升了算法的性能。

2. 架构流程

2.1. 推荐系统架构

一个完整的推荐系统如图1所示。首先,通过Lambda架构[11]进行数据采集,底层数据搜集起来,包括用户的行为日志、一些后台的业务数据以及通过爬虫得到一些数据信息等;然后,将搜集到的数据信息传递到推荐算法模块,基于这些信息来选择合适的推荐算法;之后,进入用户服务模块,通过调参等步骤确定模型的具体参数,构建新的推荐模型,当有新的用户数据到来之际,推荐模型将推荐的结果反馈给用户;最后,用户基于这些推荐结果产生一些行为数据,即用户行为反馈模块,推荐系统将这些用户的行为反馈通过Lambda架构收集起来,进行分析处理,继而为用户提供更加完善的服务。

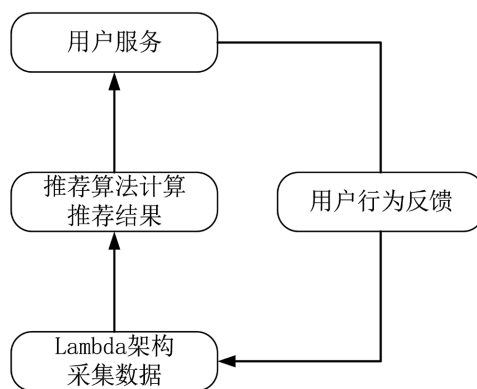


Figure 1. The framework of recommended system
图1. 推荐系统整体架构图

其中,大数据Lambda架构提供了一个结合实时数据和Hadoop预先计算的数据环境和混合平台,能够提供实时的数据视图。推荐算法模块基于用户的历史行为数据预估用户对于某商品的点击率,进而将推荐结果反馈给用户,实现商品的推荐。由推荐系统整体架构可知,推荐算法的优劣直接决定了推荐系统的好坏,所以推荐算法的研究就尤为重要。

2.2. 基于内容的推荐算法流程

一般来说,基于内容的推荐算法包括以下三步[12]:

- 1) 抽取该物品的一些特征用来表征这个物品;
- 2) 根据用户的历史行为数据信息学习出用户感兴趣的物品的特征, 即构造用户偏好文档;
- 3) 比较步骤(2)得出的用户偏好文档和待推荐物品的特征(推荐项目文档), 确定出关联性最大的一组物品作为推荐列表并将其推荐给用户。

本文中, 基于内容的推荐算法流程如图 2 所示。

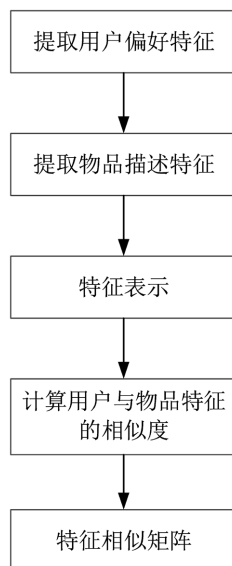


Figure 2. The procedure of content-based recommendation algorithm
图 2. 基于内容的推荐算法流程

2.3. 基于知识图谱的推荐算法流程

基于知识图谱的推荐算法流程如图 3 所示。首先, 根据物品知识及用户知识构建物品知识图谱, 并将原有的物品信息及用户信息转化为三元组的形式; 然后, 对上述三元组进行知识表示, 将构建好的物品知识图谱数据作为表示模型的输入; 最后, 计算求得物品的相似度, 生成物品相似度矩阵, 并依据相似度矩阵给出推荐列表。

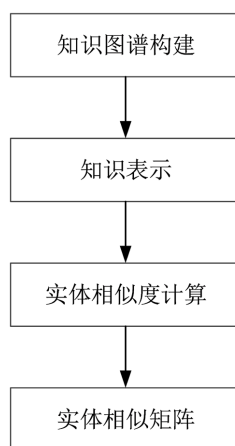


Figure 3. The procedure of recommendation algorithm based on knowledge graph
图 3. 基于知识图谱的推荐算法流程

基于内容的推荐算法较为直观，但其对于复杂属性的项目较难处理。基于知识图谱的推荐算法通过知识图谱中实体之间丰富的关系能丰富项目和项目的联系、增强推荐算法的挖掘能力，适合处理属性复杂的项目，两者进行结合，可以有效解决数据稀疏性的问题和新物品的“冷启动”问题。

3. 基于知识图谱与内容的推荐算法

基于知识图谱与内容的推荐算法如图 4 所示。包括三个部分：基于知识图谱的相似度计算，基于内容的相似度计算和基于权值的融合算法，其中输入为用户的喜好信息。在基于知识图谱的相似度计算部分，构建基于内容的知识图谱，并基于知识图谱嵌入计算得到实体与实体关系的相似度。在基于内容的相似度计算部分，基于文本词嵌入对内容进行向量化表示，并构建算法对内容进行相似度计算。在基于权值的融合算法部分，融合了基于知识图谱的相似度计算和基于内容的相似度计算结果，通过权重计算，选取准确率得分高的值作为推荐结果。

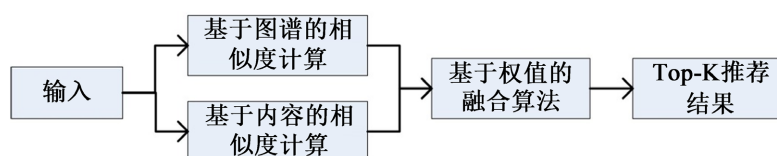


Figure 4. The framework of the algorithm based on knowledge graph and content
图 4. 基于知识图谱与内容的推荐算法框架

3.1. 基于知识图谱的相似度计算

虽然不同用户对于同一种类型的物品的描述会有所不同，但是其表达语意大都类似。例如，对于电影《东邪西毒》，用户 A 认为它是武侠片，用户 B 认为它是动作片，这是因为武侠片和动作片时有部分交叠的。利用不同用户的偏好特征，挖掘其语义特征，能够更完善地对物品进行描述，将更有利于对用户感兴趣物品的判断。因此，本文利用概念知识图谱，通过对概念的文本及知识的语义描述，在知识图谱中发现相似的概念。

基于知识图谱的相似度计算方法如下：首先，根据物品知识及用户知识构建物品知识图谱，并将原有的物品信息及用户信息转化为三元组的形式，将这些三元组存储在知识图谱中，作为进一步的查询基础。然后通过 TransE (Translating Embeddings) 模型对物品三元组进行知识表示，其中 TransE 模型以 $h+r \approx t$ 为基本思想对实体和关系做低维空间映射[13]，知识的表示嵌入如图 5 所示。如“金陵十三钗”向量为实体 h ，“电影类型”向量为关系 r ，“战争片”向量为实体 t ，因此可获得物品实体的低维稠密向量表示为 e^h 。

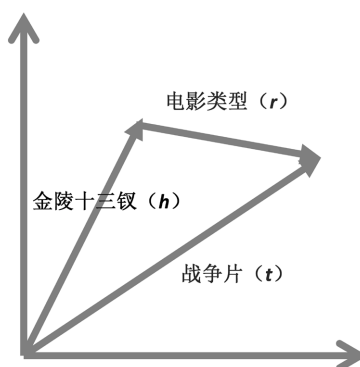


Figure 5. An example of algorithm TransE
图 5. TransE 算法案例

假设物品 I_i 用向量表示为 $I_i = (e_{i1}, e_{i2}, \dots, e_{ki}, \dots, e_{im})$ 。其中, e_{ki} 为物品的第 k 个属性, m 是属性的最大数量, 物品通过知识图谱嵌入成为 m 维向量。构建物品的知识图谱时, 采用基于 TransE 模型的训练损失函数, 由于损失函数采用欧式距离, 所以在计算物品相似度时也采用欧式距离:

$$d(I_i, I_j) = \sqrt{\sum_{k=1}^m (e_{ki} - e_{kj})^2} \quad (1)$$

因此, 物品间的相似度可以定义为:

$$sim_{graph}(I_i, I_j) = \frac{1}{1 + d(I_i, I_j)} = \frac{1}{1 + \sqrt{\sum_{k=1}^m (e_{ki} - e_{kj})^2}} \quad (2)$$

根据以上算法, 构建了知识图谱对物品的知识表示, 计算了两个物品的相似度, 由此可以生成一个物品相似度的矩阵。

3.2. 基于内容的相似度计算

基于内容的相似度计算算法提取用户的偏好特征以及物品的描述特征, 利用特征表示方法, 在特征表示的基础上计算用户的偏好特征与物品的描述特征之间的相似度, 从而实现高效准确的推荐[14], 具体步骤如下: 首先对文本中的评价信息进行中文分词; 其次利用词频-逆文档频率(Terms Frequency-Inverse Document Frequency, TF-IDF)计算词的权重; 再次利用 word2vec 计算用户特征和物品特征的词向量, 用向量表示词的语义信息; 然后利用词向量构建物品向量, 当一个物品包含多个关键词, 且每一个关键词对应一个词向量时, 可以利用加权算法构建物品向量; 最后基于用户特征和物品特征的代表向量的相似度做推荐[15]。

本文基于余弦相似度对物品向量的相似度做出评价, 将用户对某个物品的在 n 个方面的评分当成一个 n 维向量值 S_{ni} , 则某个物品 I_i 的评分向量为 $I_i = \{S_{i1}, S_{i2}, \dots, S_{ki}, \dots, S_{ni}\}$, S_{ki} 代表物品第 k 个方面的评分向量值。同样地, 物品 I_j 的评分向量为 $I_j = \{S_{j1}, S_{j2}, \dots, S_{kj}, \dots, S_{nj}\}$ 。计算 I_i 与 I_j 的余弦相似度:

$$sim_{cos}(I_i, I_j) = \frac{\sum_{k=1}^n S_{ki} \cdot \sum_{k=1}^n S_{kj}}{\left\| \sum_{k=1}^n S_{ki} \right\|_2 \cdot \left\| \sum_{k=1}^n S_{kj} \right\|_2} \quad (3)$$

如果余弦相似度得到的值大, 则说明两个物体间的相似度高, 可推荐性就高; 反之, 余弦相似度的到的值越小, 说明两个物体间的相似程度低, 可推荐性小。

3.3. 基于权值的融合算法

基于权值的融合算法结合基于知识图谱计算出的相似度与基于内容计算出的相似度, 选择合适的权值进行算法融合, 得到最终的综合相似度; 最后, 将综合相似度的结果作为是否推荐的依据。由于知识图谱中含有丰富的实体信息, 而用户的文本信息中则蕴含了用户的偏好特征, 故分别设置其加权系数为 α 和 β , α 和 β 的取值范围均在 $[0,1]$ 之间, 且 $\alpha + \beta = 1$, 关于 α 和 β 合适的取值需要通过大量的实验来获得。融合后的物品间相似度结果, 如下式所示:

$$sim(I_i, I_j) = \alpha \cdot sim_{cos}(I_i, I_j) + \beta \cdot sim_{graph}(I_i, I_j) \quad (4)$$

通过融合后的相似度计算, 可以对用户未点击过的物品进行评分预测, 选出前 k 个预测评分最高的结果, 将其排序后推荐给用户。用 P_{ui} 表示用户 U 对物品 I_i 评分, P_{ui} 的数学表达式如下:

1) 深度知识注意力网络(Deep Knowledge-Aware Networks, DKN): 融合了实体嵌入和词嵌入, 为每个输入新闻生成一个知识感知嵌入向量, 然后对用户进行动态建模, 其中物品和书籍的标题用于生成词嵌入, 实体嵌入由 TransH(Translating on Hyperplanes)学习[8];

2) LibFM (Lithium Bis Fluoromalonato): 将用户 ID、商品 ID 以及从 TransR (Translating Relations)获得的实体嵌入结合起来, 作为 LibFM 的输入[4];

3) PER (Personalized Entity Recognition): 将知识图谱视为一个异构信息网络(Heterogeneous Information Network, HIN), 并提取基于元路径的特征来表示用户和项目之间的连通性[10]。

4.3. 评估准则

在实验验证中, 采用准确率和召回率进行推荐算法的性能评价。准确率(Precision, P)表示是预测用户感兴趣的物品为正确的结果的数量。计算公式为:

$$P = \frac{TP}{TP + FP} \quad (6)$$

其中, TP 为预测结果为正确的数量, FP 为预测结果为错误的数量。

召回率(Recall, R)表示推荐给用户的物品占用户所感兴趣的物品的比例。其计算公式为:

$$R = \frac{TP}{TP + FN} \quad (7)$$

其中, FN 是把用户不感兴趣的物品预测为感兴趣的物品的数量。

4.4. 近邻数对推荐性能的影响分析

本文通过 7 折交叉验证的方法, 在实验中随机选取 70% 的数据作为训练集, 30% 的数据作为测试集。由于基于知识图谱的相似度计算的精度是由节点的近邻数所决定的, 近邻数越高, 精确越低, 但推荐结果越能覆盖用户感兴趣的物品。因此, 为了求得最佳结果, 分别取近邻数为 5、10、25、50、70、100, 并且重复进行 10 次取平均值作为最终结果。本文选取 $\alpha = 0.7, \beta = 0.3$ 。

图 7 表示了不同近邻数下召回率的变化关系, 随着近邻数的增加, 召回率越来越高, 意味着覆盖率越来越高。图 8 表示了不同近邻数下准确率的关系, 随着近邻数的增加, 准确率越来越低, 意味着推荐

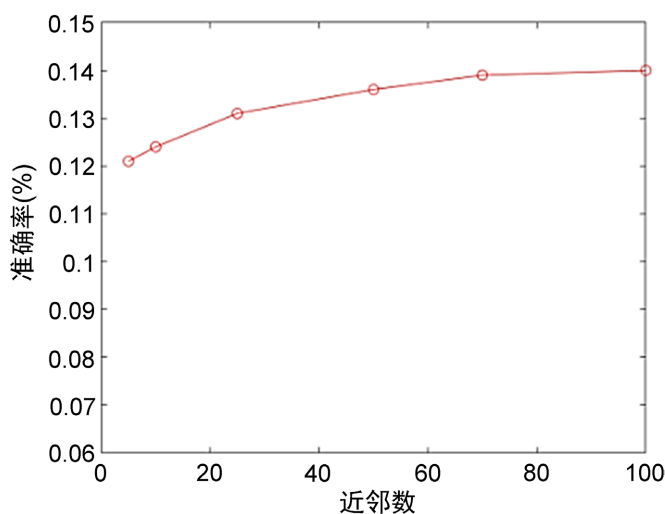


Figure 7. The movie graph used in experiment

图 7. 构建完成的电影图谱

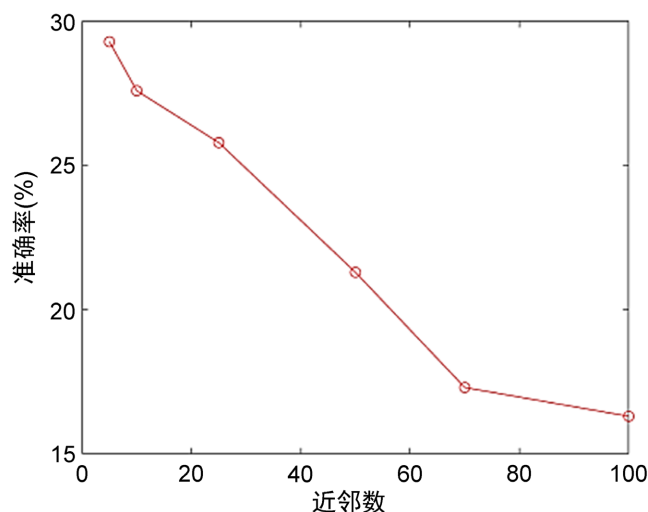


Figure 8. The precision of different nearest neighbor
图 8. 不同近邻数下的准确率

结果的错误概率越来越高。从图 7 和图 8 中可以看出，当选取不同的近邻数时，推荐给用户的物品的召回率与准确率有所差异，这是由于当选取的近邻数越多时，对用户的偏好特征掌握越详细，因此推荐给用户的物品占用户所感兴趣的物品的比例会有所提升，但是当近邻数足够大时，对用户的偏好掌握已经足够，因此召回率会趋于平稳。但是当选取的近邻数越多时，相似的用户及物品也就越多，剔除掉用户曾经购买或点击过的物品，剩余的物品依旧数量庞大，用户点击或购买的物品占推荐物品的总数会有所下降，因此准确率会有所下降。可见，在一个完备的推荐系统中，近邻数的选取并不是越大越好，而是应该进行适当的选择。

4.5. 对比实验

通过计算不同近邻数下的召回率和准确率来验证本文算法的性能。近邻数选择 1、5、10、25、50、100，取 10 次结果并求平均值。实验结果如图 9、图 10 所示。

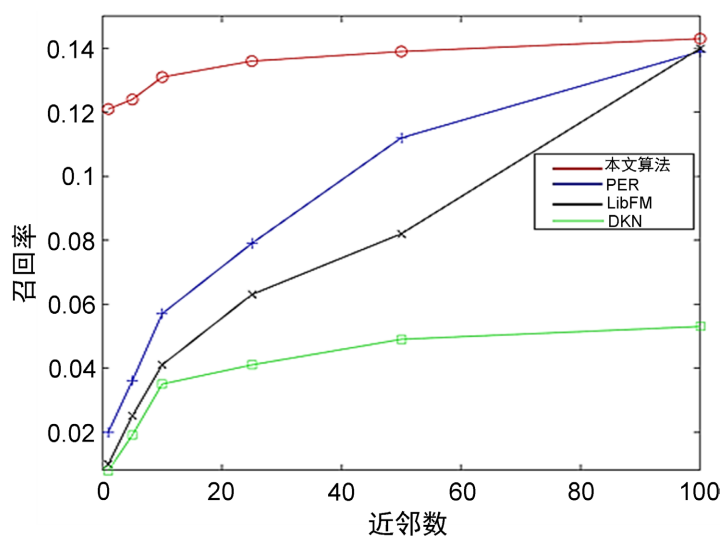


Figure 9. The recall of different nearest neighbor
图 9. 不同近邻数的召回率

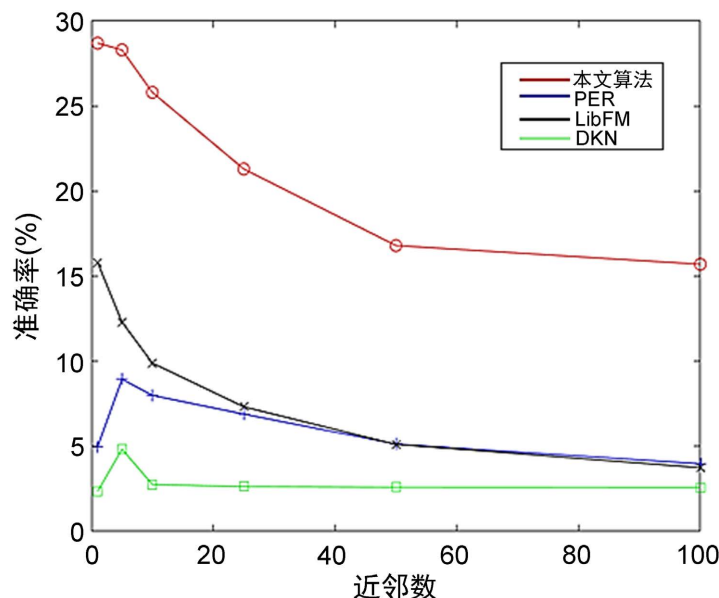


Figure 10. The precision of different nearest neighbor
图 10. 不同近邻数的准确率

图 9 展示了不同近邻数下的召回率。近邻数越多，召回率越高，意味着覆盖率越高。在相同近邻数下，本文所提算法的召回率均高于其它算法。图 10 展示了不同近邻数下的准确率。近邻数越多，准确率越低，这表明推荐的结果中出现错误推荐的数量越高。在相同近邻数下，本文算法的准确率均高于其他算法。这表明，利用实体之间的多关系和属性进行相似度计算能够提升算法的性能。由上图所示对比结果可以进一步看出：与其他算法相比，DKN 的表现最差，这是由于 DKN 对实体的表征不明确，无法为个性化推荐提供有用的信息；PER 算法的推荐结果良好，但是由于用户定义的元路径很难达到最佳，所以当近邻数达到一定程度时，PER 算法的推荐性能作用有限；LibFM 算法作为一种通用的推荐工具，在进行推荐时，准确率一直保持在较高水准，这表明 LibFM 算法可以很好地利用知识图谱的知识进行个性化推荐。而本文所提算法与上述推荐算法相比，当近邻数在 100 时召回率提升了约 1%，准确率提升了约 10%，说明本文算法是有效的。

5. 结束语

本文提出了一种基于知识图谱与内容的推荐算法。利用用户数据解析内容并构建知识图谱，在知识图谱相似度计算和内容的相似度计算基础上，设计融合算法，完成对内容的启发式推荐。与通用算法的对比实验证明了本文算法具有较高的召回率和准确率。

后续研究开展可以考虑联合实体的多个关联关系及多跳关系，利用其他的 Trans 系列模型或者知识表示模型进行知识图谱实体嵌入，以更充分地学习实体的表示向量，描述实体的语义信息，实现更进一步的推荐。

基金项目

本文得到《*****特性分析与智能判证识别技术》课题(10601010202)资助。

参考文献

- [1] 吴月萍, 郑建国. 协同过滤推荐算法[J]. 计算机工程与设计, 2011, 32(9): 3019-3021.

-
- [2] 欧艳鹏. 知识图谱技术研究综述[J]. 电子世界, 2018(13): 54+56.
- [3] 纪文璐, 王海龙, 苏贵斌, 柳林. 基于关联规则算法的推荐方法研究综述[J]. 计算机工程与应用, 2020, 56(22): 33-41.
- [4] Liao, C., Han, K.S., Baggetto, L., *et al.* (2014) Synthesis and Characterization of Lithium Bis(fluoromalonato)borate (LiBFMB) for Lithiumion Battery Applications. *Advanced Energy Materials*, **4**, 2101-2105.
- [5] Raza, S. and Ding, C. (2019) Progress in Context-Aware Recommender Systems—An Overview. *Computer Science Review*, **31**, 84-97. <https://doi.org/10.1016/j.cosrev.2019.01.001>
- [6] Wang, X., He, X.N., Wang, M., Feng, F.L. and Chua, T.S. (2019) Neural Graph Collaborative Filtering. *Special Interest Group on Information Retrieval*, Paris, 21-25 July 2019, 165-174. <https://doi.org/10.1145/3331184.3331267>
- [7] Sharma, S. and Rana, V. (2017) Web Personalization through Semantic Annotation System. *Advances in Computational Sciences and Technology*, **10**, 1683-1690.
- [8] Lian, J., Zhou, X., Zhang, F., *et al.* (2018) xDeepFM: Combining Explicit and Implicit Feature Interactions for Recommender Systems. *The 24th ACM SIGKDD International Conference*, London, 19-23 August 2018, 1754-1763.
- [9] 徐兵, 弋沛玉, 王金策, 彭舰. 知识图谱嵌入的高阶协同过滤推荐系统[J]. 计算机科学, 2021, 48(11): 244-250.
- [10] Xiong, Y., Zhu, Y. and Yu, P.S. (2015) Top-k Similarity Join in Heterogeneous Information Networks. *IEEE Transactions on Knowledge & Data Engineering*, **27**, 1710-1723. <https://doi.org/10.1109/TKDE.2014.2373385>
- [11] 韩嫒. Lambda 架构在处理海量高并发数据中的应用[J]. 现代电视技术, 2016(12): 70-72.
- [12] 秦冲, 赵铁柱, 柳毅. 个性化推荐算法的研究及发展综述[J]. 东莞理工学院学报, 2021, 28(3): 51-59.
- [13] Dechter, R. (1996) Bucket Elimination: A Unifying Framework for Probabilistic Inference. *Proceedings of the 12th Conference on Uncertainty in Artificial Intelligence*, Portland, 1-4 August 1996, 211-219.
- [14] 肖文强, 姚世军, 吴善明. 一种改进的 top-N 协同过滤推荐算法[J]. 计算机应用研究, 2018, 35(1): 105-108+112.
- [15] 张国锋, 徐雷, 余方超, 李大双. 一种改进的基于用户协同过滤的推荐算法研究[J]. 电视技术, 2019, 43(14): 5-10.