



知识图谱与数据应用——智能推荐

周晶, 孙喜民, 于晓昆, 边新宁

(国网电子商务有限公司, 北京 100069)

摘要: 知识图谱技术是新兴的互联网技术, 具有较好的前瞻性和较强的技术先进性, 有着广泛的应用前景。知识图谱技术提供从“关系”的角度分析问题的能力, 可以对数据进行深度挖掘, 将自然语言转化为计算机语言, 最大限度地展示数据的价值, 可以服务于智能搜索、智能推荐、风险预警、智能运营、智能客服、舆情监测、设备预警等业务, 大幅度地提高企业生产效率。研究了综合聚类算法、SVD 分解算法、基于商品的协同顾虑推荐算法、基于用户的协同顾虑推荐算法和商品相似度算法等, 结合推荐策略提出了基于知识图谱数据应用的智能推荐系统。证明了以知识图谱作为基础工具的数据应用的可实施性和可部署性, 能够全面满足用户在互联网平台的获取信息的需求。

关键词: 知识图谱; 数据挖掘; 智能推荐

中图分类号: TP311

文献标识码: A

doi: 10.11959/j.issn.1000-0801.2019064

Knowledge graph and data application: intelligent recommendation

ZHOU Jing, SUN Ximin, YU Xiaokun, BIAN Xinning

State Grid Electronic Commerce Co., Ltd., Beijing 100069, China

Abstract: Knowledge graph technology is an emerging internet technology with good forward-looking and strong technological advancement, and has broad application prospects. Knowledge graph technology provides the ability to analyze problems from the “relationship” perspective. It can deeply mine data, transform natural language into computer language, and display the value of data to the greatest extent. It can serve intelligent search, intelligent recommendation, risk early warning, intelligent operation, intelligent customer service, public opinion monitoring, equipment early warning and other businesses, greatly improving the production efficiency of enterprises. The comprehensive clustering algorithm, SVD decomposition algorithm, commodity-based collaborative concerns recommendation algorithm, user-based collaborative concerns recommendation algorithm and commodity similarity algorithm were studied. Combined with recommendation strategy, an intelligent recommendation system based on knowledge graph data application was proposed. It proves the feasibility and deployability of the data application with knowledge graph as the basic tool, which can fully meet the user's demand for information on the internet platform.

Key words: knowledge graph, data mining, intelligent recommendation



1 引言

互联网高速发展的今天,人工智能(artificial intelligence, AI)已成为新的生产力,而知识图谱(knowledge graph/vault)是 AI 进步的阶梯。早在 20 世纪,知识图谱的概念雏形已出现,最具代表性的是美国哈佛大学心理学教授 Stanley Milgram (1933—1984 年)提出的“六度分隔理论(six degrees of separation)”^[1],理论表明:“任何两个陌生人之间所间隔的人不会超过 6 个。”六度分隔理论的应用场景是描述联接人与社区的人际关系网。关系推理如图 1 所示。

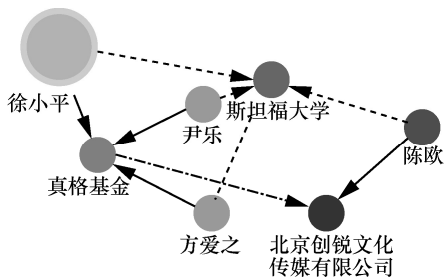


图 1 关系推理

2012 年 5 月,Google 正式提出知识图谱的概念,主要用于描述真实世界中存在的各种实体和概念及实体概念之间的关联、关系,把所有不同种类的信息连接在一起而得到的一个关系网络,提供了从“关系”的角度分析问题的能力,用于增强其搜索引擎功能的知识库。从本质上讲,知识图谱是一种揭示实体之间关系的语义网络,可以对现实世界的事物及其相互关系进行形式化的描述,近年来,语义链网络已经发展为支持信息物理社会智能^[2]。现在的知识图谱已被用来泛指各种大规模的知识库。

2 基于知识图谱的数据应用

2.1 知识图谱技术及图谱构建

知识图谱是描述真实世界存在的各种实体或概念及其关系所构成的语义网络图,语义网络图的节点表示实体或概念,边由属性或关系构成,

知识图谱技术是一种工具或方法论。通过对海量数据进行采集、知识抽取、知识融合、知识存储、知识编辑、知识标注等数据处理流程,获取“实体—关系—实体”“实体—属性—属性值”等三元结构的关系数据并存储,把数据包含的知识进行抽取和应用,形成计算机可读的知识库,完成知识图谱构建,通过数据访问、图谱计算、图谱可视化等功能进行知识图谱数据交互。数据抽取示例如图 2 所示。



图 2 抽取数据示例

2.2 知识图谱的数据应用场景

知识图谱的数据应用概括为两个范畴:通用知识图谱应用和行业知识图谱应用,前者强调广度,后者精于深度。通用知识图谱的实体应用场景主要面向互联网的搜索、推荐、问答等业务,通过无限的迭代,扩充知识的广度,强调更多的是关联关系中的实体;而行业知识图谱面向的是特定领域,一个领域对应一个规则,需要考虑不同场景、不同用户目标对象存在的差异化,具备丰富而严格的数据结构,并需要一定的深度和完备性。基于知识图谱的数据应用如图 3 所示。

目前,知识图谱典型的应用场景有智能搜索、智能推荐、风险预警、智能运营、智能客服、舆情监测和设备预警等。其中,智能搜索应用场景引入知识图谱技术,构建基于语义网的智能搜索功能,搜索结果不再是简单的文字模糊匹配,搜索模块可读懂搜索意图,进行意图识别,返回智能搜索结果;在智能客服系统的应用中可以构建业务知识库,构建机器人智能客服,节省人工客服资源;在风险预警方面,可以自动推理出风险

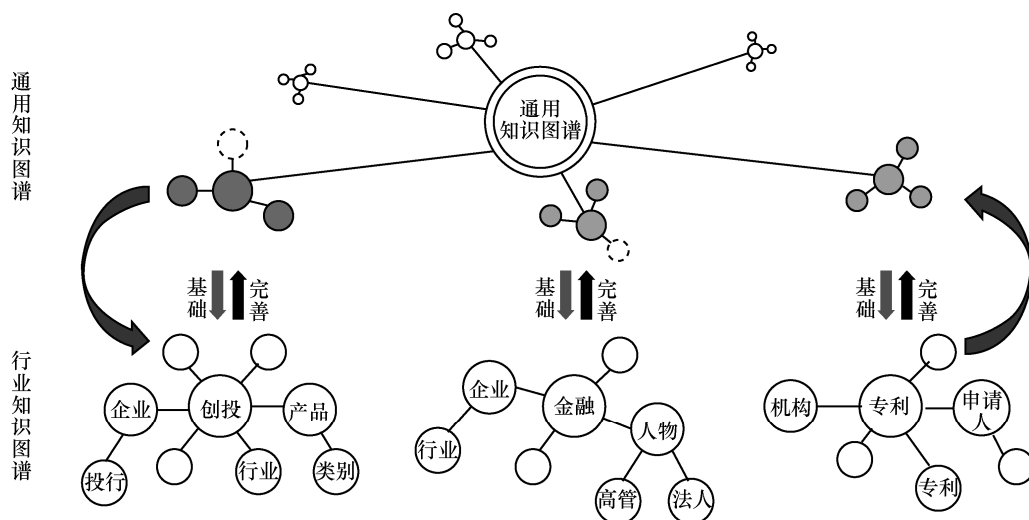


图3 基于知识图谱的数据应用

点，提前预防，比如电商运营平台的商品价格异常、用户购买行为异常等。

2.3 基于知识图谱的互联网电商数据应用

传统的互联网电子商务，在搜索、推荐、数据挖掘及计算方面还不能给人提供富有智能感的交互式场景，对用户体验会造成一定影响，从而影响用户的购买意愿。而随着语义技术的发展，利用知识图谱技术及围绕知识图谱的一系列技术，商城消费者可以进行智能的搜索交互操作，并提供智能的商品推荐，从而使消费者能够便捷、轻松地从海量商品中找到目标，大幅度提升用户体验，促进购买。图4为国网电商的知识图谱构建示例，图谱数据通过可视化技术进行关系展示，同时，关系数据还用于国网电商平台的智能搜索、智能推荐等业务功能模块。

3 基于知识图谱的数据应用—智能推荐

智能推荐是知识图谱在互联网电商平台数据的应用场景之一。智能推荐从业务逻辑描述，就是对海量的商品信息进行筛选、过滤，将用户最关注、最感兴趣的商品信息展示给用户，提升用户的购物体验。从业务实现来说，就是引入知识图谱技术，抽取商品信息、用户信息，对信息进

行过滤，构建智能推荐需要的商品知识库、用户信息知识库，并形成知识图谱，通过协同过滤^[3]的推荐算法，实现商品面向用户的智能推荐。智能推荐应用业务逻辑如图5所示。

3.1 商品知识图谱构建

商品知识图谱构建过程就是商品信息数据知识抽取、知识融合的过程，知识图谱构建平台在获得电商平台商品数据后，构建完成商品知识图谱，以此作为基础数据，并对融合后的知识进行存储和应用。在获取数据的过程中，采用ETL技术同步元数据，经过模式转换、模式映射、属性抽取等技术手段，保证数据的一致性，电商平台的商品数据发生变化时，知识图谱库数据库中的数据要同步更新。知识图谱数据库在创建时，数据库表、表类字段、表之间的关联关系严格按照索引时的逻辑进行建库，对字段进行划分，用户、商品、商品分类、商户、订单数据建模的关系如图6所示。

3.2 用户画像知识图谱构建

用户画像知识图谱构建需要融合用户基础信息知识和用户行为信息知识。用户基础信息抽取面向的是用户在电商平台提交的基础信息，包括用户的年龄、地域、职业、学历、购物偏好等；

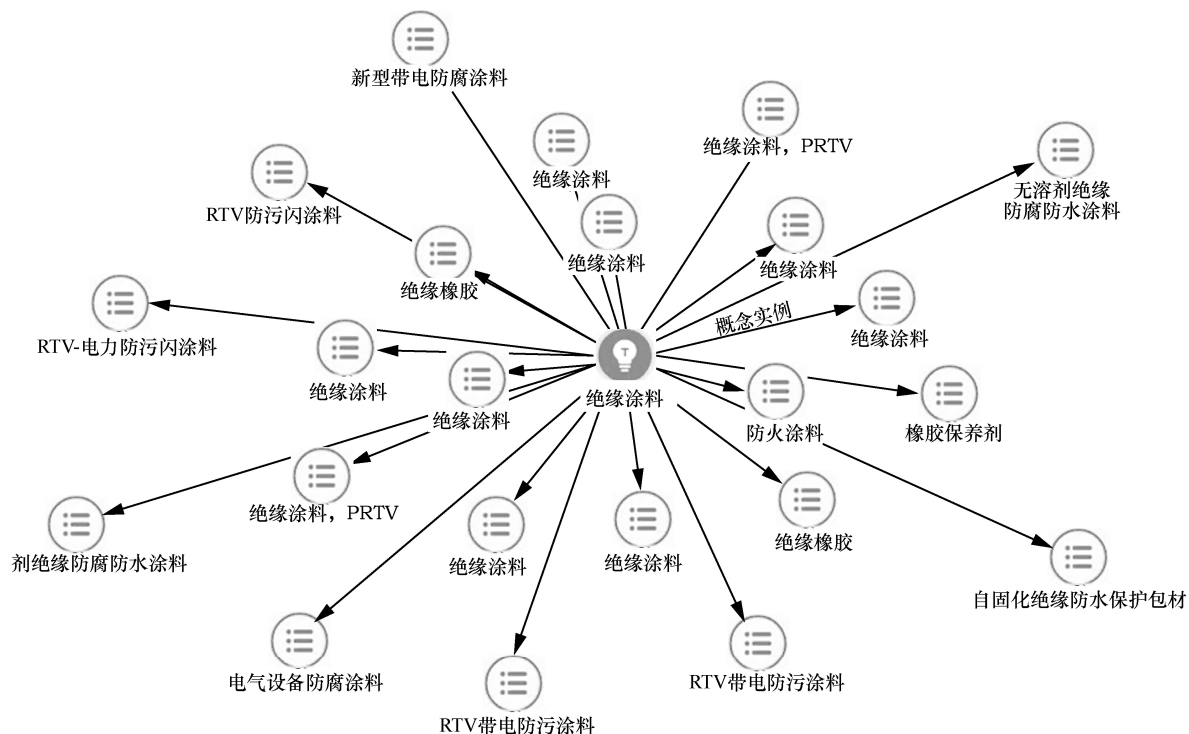


图4 电商知识图谱构建示例

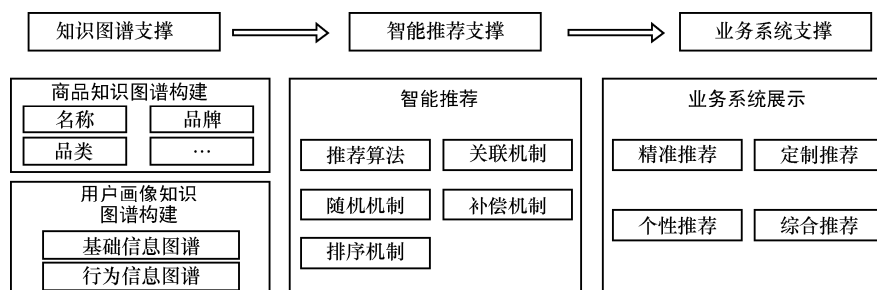


图5 智能推荐应用业务逻辑

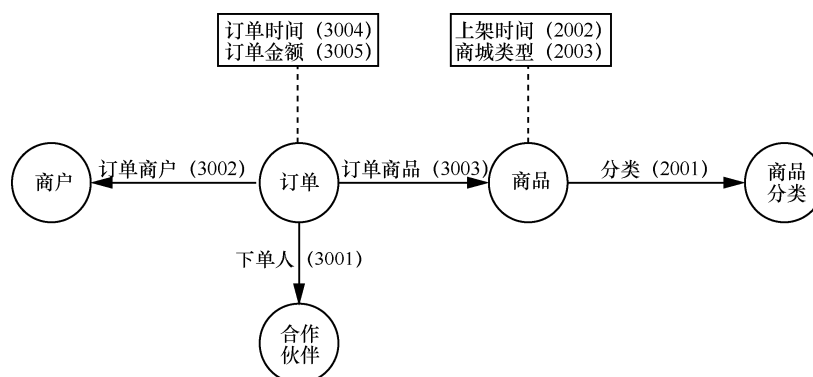


图6 知识图谱库数据建模关系

用户行为信息抽取面向的是用户在电商平台产出的行为数据，包括用户浏览、收藏、评论、加入购物车、下单、付款的商品信息。分别抽取这两类

信息的实体数据，进行数据融合，构建用户画像知识图谱。基于用户画像的推荐引擎原理如图7所示。

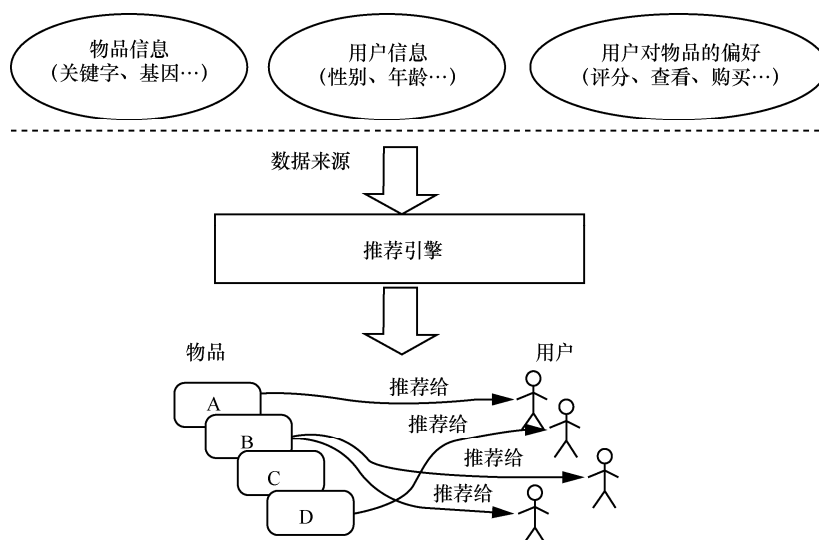


图7 基于用户画像的推荐引擎原理

3.3 推荐算法

知识图谱存储的数据为结构化数据，可以作为基础数据为大多数的数据挖掘、数据分析、机器学习特征提取与建模使用，也可以直接在知识图谱的模型上，利用聚类及 SVD 进行推荐计算。

3.3.1 电商平台智能推荐应用数据模型

在互联网电商平台，向用户推荐商品的场景较多，有平台主页、商品详情页、订单页等场景推荐，推荐的数据模型也可以多样化展示，目前使用较多的模型如下：

- 数值区间推荐；
- 推荐同商品类目下，在同一数值区间的物品，数值区间可以根据业务来制定；
- 基于数值统计来推荐；
- 热销商品推荐；
- 基于重点业务来推荐；
- 根据业务，按条件加权推荐。
- 基于推理规则推荐；
- 同品类推荐；
- 语义关联推荐；
- 根据业务规则来进行推荐。

3.3.2 基于知识图谱的聚类

互联网电商平台智能推荐聚类的目的是把相

似度匹配的商品进行汇聚，使用 K 最近邻 (K -nearest neighbor, KNN) 分类算法进行聚类计算，其原理为如果一个样本在特征空间中的 k 个最相似（即特征空间中最邻近）的样本中的大多数属于某一个类别，则该样本也属于这个类别。在推荐系统中，基于向量空间模型计算推荐对象和待推荐对象之间的相似度，取相似度最高的 N 个作为最终的推荐结构。

3.3.3 知识图谱模型的 SVD 矩阵分解

SVD 是一种主成分分析以及降维的方法，思路就是一个对称矩阵 A 可以分解成对角形：

$$A = U^T \Lambda V \quad (1)$$

其中， U^T 代表用户的 r 行矩阵 U ， V 代表商品的 r 列矩阵 V ，使用 SVD 矩阵的目的是找到用户—商品矩阵中最大的 r 个奇异值，进行分解构成 r 行的矩阵 U 和 r 列的矩阵 V ，以此对用户的打分进行预测^[4]。

3.3.4 基于商品的协同顾虑推荐算法

基于商品的协同顾虑推荐 (item-based collaborative filtering) 的原理是在计算邻居时采用商品本身，而不是从用户的角度，即基于用户对商品的偏好找到相似的商品，然后根据用户的历史偏好，推荐相似的商品。从计算的角



度看,就是将所有用户对某个商品的偏好作为一个向量来计算商品之间的相似度,得到商品的相似商品后,根据用户历史的偏好预测当前用户还没有表示偏好的商品,计算得到一个排序的商品列表作为推荐。

例如,对于商品 A,根据所有用户的历史偏好,喜欢商品 A 的用户都喜欢商品 C,得出商品 A 和商品 C 比较相似,而用户 C 喜欢商品 A,那么可以推断出用户 C 可能也喜欢商品 C,如图 8 所示。

用户/物品	物品A	物品B	物品C
用户A	✓		✓
用户B	✓	✓	✓
用户C	✓		推荐

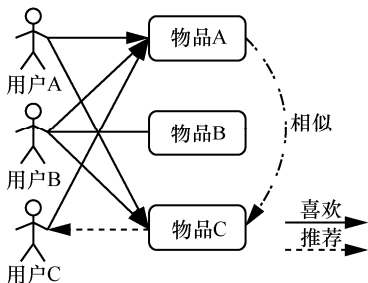


图 8 基于商品协同推荐示例

3.3.5 基于用户的协同过滤推荐

基于用户的协同过滤推荐 (user-based collaborative filtering) 算法是基于用户对商品的偏好找到邻居用户,然后将邻居用户喜欢的推荐给当前用户。将一个用户对所有商品的偏好作为一个向量来计算用户之间的相似度,找到 K 邻居后,根据邻居的相似度权重以及他们对商品的偏好,预测当前用户没有偏好的未涉及商品,计算得到一个排序的商品列表作为推荐。

例如,对于用户 A,根据用户的历史偏好,这里只计算得到一个邻居——用户 C,然后将用户 C 喜欢的商品 D 推荐给用户 A,如图 9 所示。

用户/物品	物品A	物品B	物品C	物品D
用户A	✓		✓	推荐
用户B		✓	✓	
用户C	✓		✓	✓

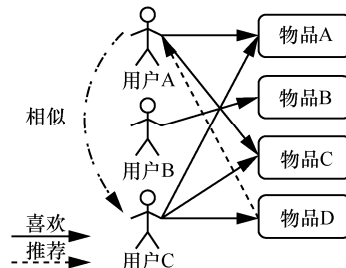


图 9 基于用户协同推荐示例

3.3.6 相似度算法

两个用户之间的相似度算法可以使用欧几里德距离计算公式^[5],可以简单而精准地评价两个用户的关联关系,原理是通过计算两个用户在散点图中的距离来判断不同的用户是否有相同的偏好:

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (2)$$

其中, (x, y) 代表用户在欧几里德空间的点集, $d(x, y)$ 为用户在欧几里德空间的距离。

皮尔逊相关度评价是另一种计算用户间关系的方法^[6],评分数据不规范时皮尔逊相关度评价能够给出更好的结果,计算式如下:

$$\rho_{X,Y} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E((X - \mu_X)(Y - \mu_Y))}{\sigma_X \sigma_Y} = \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - E^2(X)} \sqrt{E(Y^2) - E^2(Y)}} \quad (3)$$

其中, X, Y 为代表用户变量, ρ 为用户间的线性相关关系, E 为数学期望值, $\mu_X = E(X)$, $\sigma_X^2 = E[(X - E(X))^2] = E(X^2) - E^2(X)$ 。

以上 4 个算法为智能推荐中最常用的基础算法,在一些特殊的业务场景,还有基于商品自身相似度推荐的算法等。综合以上算法,计算用户

和商品的相似度, 再根据权重配置重新计算, 用于搜索结果的排序。

4 推荐策略

推荐策略根据不同的业务场景而不同, 主要是电商平台前端展示推荐商品的规则, 通过推荐算法, 可以获取 3 类商品数据。

- 强关联商品: 按照用户和商品知识图谱的相似度关系, 相似度值最高的商品定义为强关联商品。
- 弱关联商品: 按照用户和商品知识图谱的聚类结果, 与推荐目标用户相似度较高用户聚类的商品, 且这些商品与目标用户关联度较低的商品定义为弱关联商品。
- 无关联商品: 按照用户和商品知识图谱的聚类结果, 被聚类次数较高的商品但与目标用户没有关联关系的商品为无关联商品。

推荐策略可以围绕这 3 类商品的投放比例对用户推荐数据, 实现不同场景推荐商品不同, 不同用户推荐商品不同, 相同用户在不同场景推荐商品不同, 达到个性化推荐、千人千面的目的。7:2:1 的智能推荐策略如图 10 所示。

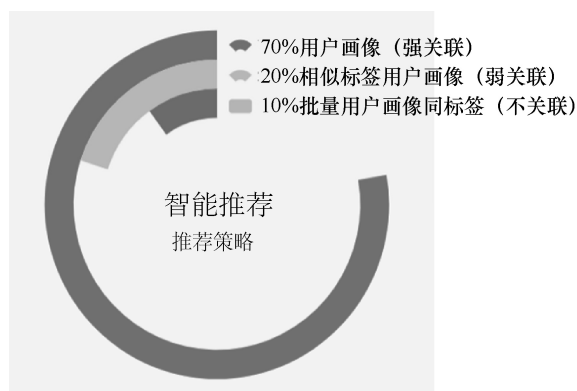


图 10 7:2:1 的智能推荐策略

为避免推荐的商品反复出现, 干扰用户判断, 推荐展示的商品数据可以通过提取关联度最

高的前 10% 占比的商品, 从这批数据里进行随机抽取展示。

为避免出现部分用户推荐数据不足, 可以通过补偿机制来填充缺失数据, 当获取的强关联商品不足排序条数时, 用弱关联的商品进行补全, 弱关联的商品条数不足补偿强关联条数时, 用不关联的商品进行补偿展示; 当获取的弱关联商品不足排序条数时, 用不关联的商品数据补全。

5 结束语

本文提出的基于知识图谱数据应用的智能推荐系统, 主要面向的是互联网用户, 通过抽取物联网产出的数据, 以知识图谱技术作为支撑的工具和方法, 实现互联网平台智能推荐系统的数据应用^[7]。通过知识图谱技术对电商平台用户、商品数据进行知识抽取、知识融合处理, 获取有效的用户、商品数据, 构建形成用户知识图谱、商品知识图谱, 综合聚类算法、SVD 分解算法、基于商品的协同顾虑推荐算法、基于用户的协同顾虑推荐算法、商品相似度算法等算法, 结合推荐策略, 为多维度、多业务、定制化业务场景的用户提供人性化、个性化和千人千面的内容推荐系统。

基于知识图谱技术的数据应用系统模型的原型(智能推荐系统)证明了以知识图谱作为基础工具的数据应用的可实施性和可部署性, 能够全面满足用户在互联网平台的获取信息的需求。另外, 本系统同样可应用到电商平台外的其他物联网平台系统^[8]。

参考文献:

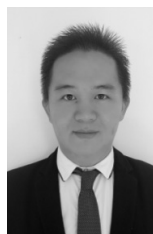
- [1] 邓肯·J·瓦茨. 六度分隔[M]. 陈禹, 等译. 北京: 中国人民大学出版社, 2011.
WATTS D J. Six degrees[M]. Translated by CHEN Y, et al. Beijing: People's University Publication House, 2011.
- [2] ZHUGE H. Communities and emerging semantics in semantic



- link network: discovery and learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 21(6): 785-799.
- [3] MEHDI E, FRANCESCO R, NEIL R. A survey of active learning in collaborative filtering recommender systems[J]. Computer Science Review, 2016.
- [4] HANSEN P C. The truncated SVD as a method for regularization[J]. BIT, 1987(27): 534-553.
- [5] DEZA E, DEZA M M. Encyclopedia of distances[M]. Heidelberg: Springer, 2009.
- [6] RODGERS J L, NICEWANDER W A. Thirteen ways to look at the correlation coefficient[J]. The American Statistician, 1988, 42(1): 59-66.
- [7] 郭佳睿, 魏进武, 张云勇. 大数据助力运营商提升规模化运营核心力策略[J]. 电信科学, 2018, 34(1): 120-125.
- GUO J R, WEI J W, ZHANG Y Y. Strategies for enhancing core capability of large-scale operation for national telecom operators assisted by big data[J]. Telecommunications Science, 2018, 34(1): 120-125.
- [8] 张建强, 张高毓. 区块链技术在物联网中的应用分析[J]. 电信科学, 2018, 34(Z1): 104-110.

ZHANG J Q, ZHANG G Y. Analysis application of blockchain technology in internet of things[J]. Telecommunications Science, 2018, 34(Z1): 104-110.

[作者简介]



周晶 (1984-), 男, 国网电子商务有限公司工程师, 主要研究方向为电子商务、人工智能以及大数据应用。

孙喜民 (1979-), 男, 国网电子商务有限公司技术工程师, 主要研究方向为图像识别、人工智能以及大数据应用。

于晓昆 (1979-), 男, 国网电子商务有限公司数据产品架构师, 主要研究方向为电子商务、人工智能以及大数据应用。

边新宁 (1977-), 男, 国网电子商务有限公司产品设计师, 主要研究方向为电子商务、人工智能以及大数据应用。