**融合移动互联网/社交网数据的城市轨道交通网挖掘~~实时路线推荐~~**

**融合社交网数据的轨道交通路线推荐**

**融合商家评论和城市轨道交通客流相关度研究**

**基于多源异质信息和时空模型的轨道交通客流引导与预测**

**城市轨道交通用户兴趣点挖掘和预测**

**报告正文**

**（一）立项依据与研究内容**（4000-8000字）：

1．**项目的立项依据**（研究意义、国内外研究现状及发展动态分析，需结合科学研究发展趋势来论述科学意义；或结合国民经济和社会发展中迫切需要解决的关键科技问题来论述其应用前景。附主要参考文献目录）；

# 研究意义

移动互联网的快速发展已经对人们衣食住行方方面面产生着影响。城市轨道交通系统在日常出行中也占有绝对的比例优势，也同时是未来出行模式的重点。在现有情况下，发掘移动互联网中与出行目的，如吃喝玩乐相关的数据与城市轨道交通客流之间的相关性对商业分析、轨道交通规划、城市规划、客流引导等方面具有十分重要的价值。但是目前该领域的研究工作还存在很多难点：1）可借鉴的研究成果十分有限：城市轨道交通和移动互联网都是近几年快速发展的新兴事物，基础研究工作很少；2）传统角度理解上的行业跨度较大，起步晚，具有相应研究兴趣和基础的工作较少；3）行业跨度也造成数据关联挖掘难度大：轨道交通的客流数据是比封闭的内容，只有城市地铁运营商具有，一般研究单位难以获得。4）目前快速发展的机器学习，甚至是深度学习理论和技术在相关领域的探索比较少。智能化挖掘的基础薄弱。

课题组与国内多城市的轨道交通运营商合作开展客流预测和分析的项目，同时在数据挖掘，移动互联网应用领域有丰富的科研和工程经验。二者的结合促成本课题组希望利用机器学习和深度学习等技术手段开拓该问题的研究工作。

本课题拟将移动互联网的文本数据融合到路线推荐中，构建结合商家和客流信息的路线推荐模型。本课题首先提取站点周边的商家标签，并与气象、节日等多维度信息融合，构建基于多维信息的客流预测模型。最后，将商家和客流信息融入到路线推荐模型中，研究商家、客流、历史行为数据对于路线推荐的影响，完成移动互联网数据和轨道交通规划的结合。研究工作将从下述三方面展开：

1. **基于评论和微博的商家标签优化；**

当前的轨道交通规划通常都局限于交通领域本身的数据，而往往会忽视基于移动互联网的文本数据，而这部分数据通常表达了人们的出行意愿，隐含了大量有价值的潜在出行信息。移动互联网上的文本数据主要包括微博和商家评论数据，针对文本类型的数据从中挖掘商家、商家提供的服务内容(如菜品、环境、服务等)、地点、站点、季节、热点事件等信息，从而为进一步挖掘这类信息与客流模型的相关性以及后续的路线推荐奠定基础。

商家标签对于用户出行意愿具有指导意义,尤其是针对以吃喝玩乐为出行目的的用户来说，站点周边的商家信息对出行目的地的选择有重大参考价值，而以往的研究在客流预测和路线推荐中，都没有将商家信息考虑在内。同时，商家标签的提取是构建商家画像的基础，而只有从商家的属性、菜系、客单价、服务质量、热销菜品等数据分析中，才能最真实最直接的了解到商家的经营情况，才能向用户展示最全面的商家信息。目前主流的评论网站中，已有的商家标签是通过标签提取算法通过对每一条数据提取标签汇总得到的，但是由于算法的不完善使得一条评论通常没有标签或者只有一个标签，最终导致商家标签的准确度和完善度都不高。除了评论数据外，来自微博的文本数据同样蕴含着大量对商家的评价信息且未被利用，因此研究如何利用评论数据和微博数据提高商家标签的准确度，对客流预测和路线推荐都有着重要价值。

在本课题中，商家标签可分为两大类，分别为服务类（如环境、服务、口味、交易方式）和菜品类(推荐的菜品)，由于商家标签的复杂性和多样性，本课题针对这两类标签采取不同的提取方法。其中服务类的标签采用基于语义的多标签分类算法，而菜品类标签的提取则可以转化为对实体关系的识别问题。对多标签分类和实体关系识别的研究，是文本挖掘领域的难点问题，也是本课题拟解决的重点问题。

1. **基于多维数据的时空客流预测模型构建；**

在轨道交通领域，随着在一、二线城市轨道交通的不断扩建和发展，地铁已经成为最重要的轨道交通出行方式之一。日益增长的客流量对地铁的运营提出了更高的要求。例如若出现客流的爆发式增长，交通部门需要对其采取及时有效的管控措施，疏通客流，开放闸机，减少由客流高峰引起的安全事故。同时随着客流的不断增加，交通部门也需要不断规划新的地铁路线，减少各线路的运营压力。对下一时段客流的有效分析和预测，能够让交通部门掌握客流发展动向，拿到一定的主动权，从而更好的应对客流的变化。

客流量数据属于典型的时间序列数据，现有经典的时间序列数据预测方法大致有ARIMA、回归模型、神经网络模型等。这些模型对于时间序列数据处理能力较好。但是这些方法只是利用了历史数据进行挖掘分析，针对的是单一的时间序列数据。实际生活中，影响客流的因素多种多样，除了客流历史数据，站点位置、气象、时令节日、站点周边商家的情况等因素也会对一个城市内的客流也会造成不同程度的影响。典型的预测模型很难将这样多种因素考虑在内，从而丢失了很多有用的信息。挖掘这些因素对客流的影响程度或者与找出与客流的关系对于商家或者活动、客流预测、市政建设都有指导意义。因此综合多因素对多站点短期客流预测有重要意义。同时客流预测模型作为基础研究，对于路线推荐、交通规划等都有重大的参考意义。

1. **基于商家标签和客流的路线推荐模型构建；**

在乘客的出行计划中，可以将目的地划分为两种类型。一种类型是固定目的地，即乘客已经明确要去的地方，是不可更改的，例如对上班人群来说，目的地就是公司所在地。第二种类型的目的地则比较灵活，乘客的出行主要是为了休闲娱乐。这种情况下，乘客出行的目的地在一定程度上具有可规划性，因此构建一个良好的路线推荐模型，既能满足用户的需求，又能减缓交通压力。

用户在以休闲娱乐为目的出行的情况下，通常并未想好自己要去的具体目的地，此时目的地是商家粒度的，由于商家数量巨大，用户查找这些商家需要消耗大量的时间。同时，用户在查找商家的时候，并不清楚商家所处地铁站点的客流量情况。路线推荐模型通过分析用户曾经去过的商家，结合站点周围的商家、商家标签和站点客流量信息，向用户推荐可能感兴趣的站点，从而帮助用户制定更合理的出行计划，节省用户的时间，同时帮助用户避开客流量大的站点，提高出行效率和用户感受。

# 国内外研究现状

补充 基于商户标签和客流的路线推荐：研究现状

传统的路线推荐算法，如Dijkstra、A\*等，都是在已知起点和终点的情况下，推荐从起点到终点的一条路线，推荐的结果带有某种约束条件，如最短时间、最少换乘等。这类算法并不能给以休闲娱乐为目的的用户提供可供参考的终点站，并且只考虑到了单一的客流情况。目前的研究均未考虑到移动互联网文本数据中的潜在出行信息。本课题将通过对评论社交网站上的评论和微博的文本挖掘，提取商家标签，并和天气、节日等多维信息相融合，构建客流预测模型，最终结合商家标签信息和客流预测模型来构建路线推荐模型，从而将移动互联网的文本数据融合到客流预测和路线推荐中。下面将从商家标签优化，客流预测以及路线推荐三方面描述国内外研究现状。

1. 商家标签优化

本课题中的商家标签将分为两类，服务类和菜品类。其中对服务类标签的提取本质上属于多标签分类问题，即通过对评论和微博的分析，对商家提取多个服务类标签，标签维度包括环境、口味、交易方式等不同的服务。菜品类标签则不同于服务类标签，是从评论和微博文本中提取出推荐菜品和不推荐菜品，因此本质上可以视为实体关系识别问题，识别菜品和商家的关系(推荐或不推荐)。下面将分别对多标签分类和实体关系实体的研究现状进行说明。

1. 多标签分类

在文本分类领域，对于二分类和多分类问题的研究已经非常成熟，然而在实际应用中文本常常不只属于某个单一的类，文本数据具有多样性和模糊性，例如在社交网络中，用户对自己的文本设置标签时，会提取出文本中吸引人的多个标签。而在商家推荐系统中，每个商家依赖于用户的评价得到自己的标签如服务好，环境好，性价比高等。将这里的标签映射到分类类别，就演化为了多标签分类问题。对这类问题的研究已经逐渐成为文本分类领域的热点问题之一。

目前, 文本分类技术取得了大量研究成果, 也得到一定的应用。然而, 其仍然面临着数据偏斜、非线性、多标签、标注瓶颈等问题。其中多标签分类, 指的是一个文本与不止一个类别相关联。在实际任务中, 常常会出现多标签分类的情况。一般而言, 对多标签分类问题的研究主要从以下三个角度出发[7]。

(1) 假设类别相互独立, 在此前提下, 最简单、最常用的方法为将多标签分类问题转换为多个二元分类问题, 综合各二元分类的结果作为最终分类结果。如Joachims利用支持向量机算法实现了这种分类方法[2]。此外, 还有基于排序的方法, 在训练时学习得到一个排序函数, 据此对文本和类别的匹配情况进行打分, 将文本划分到分值高的类别。如Crammer等通过计算出每个类别的权重向量, 进而计算文本特征向量与类别权重向量的内积, 排序决定所属类别[3]。该类方法大多简单易行, 且有大量高效算法可以直接利用, 但对于类别间有关联的情况难以获得很好的性能。

(2) 考虑类别间的关联。一般通过构建主题模型解决多标签分类问题。如Ueda等提出一种产生式体系, 包含任意两个类别间的关系[4]; Zhang等提出双层主题分类模型, 基于实例间的差异构建模型[5]。

(3) 利用半监督学习算法对未标记文本进行学习。这类方法综合考虑了同类别文本间的关系和不同类别文本间的关系。如Liu等通过求解带约束的非负矩阵获得最优的样本标签[6]。这类方法能有效地利用类别间的关系, 但是学习过程较为复杂。总体而言, 多标签分类问题比单标签分类问题更为复杂。

已有的多标签分类方法大多从第一个角度解决问题，即假设分类标签直接是独立的，这样做的优点是模型构造简单，但是这样的方法用于服务类标签的分类将会制约算法效果的提升，因为服务类标签间是有关联的。从角度二解决问题虽然在关联性方面有一定提升，但是也存在不足，具体体现在标签关联分析的全面性方面，且算法的复杂度较高。因此如何在模型中考虑标签的关联性同时保证算法的复杂度成为本课题的重点研究内容之一。

1. 实体关系识别

实体关系抽取的研究目前主要采用的是有监督分类算法，有监督算法面临的主要问题在于人工标注语料效率低下，因此Mintz[1]提出了远程监督方法，通过对齐已有的知识库来完成语料的自动标注。远程监督有一个强假设条件，如果两个实体对存在某种关系，那么任何包含这两个实体的句子都存在这种关系。实际上这个假设并不成立，因此会引入较多的噪声数据。为了缓解噪声数据带来的负面影响，Riedel[2]引入了多示例学习，并结合SampleRank来判别实体关系。Riedel证明了多示例学习的有效性。除了多示例之外，随着深度学习的发展，神经网络也被应用于关系抽取中。Zeng [6]提出了PCNN，和CNN不同的是，PCNN根据实体所在位置，将句子切分成三段进行池化，从而得到更多和实体相关的上下文信息，并且基于多示例学习中的at-least-one假设，在包含实体对的所有句子中，选择概率最大的句子作为实体对的表示。虽然这种方法减少了噪声带来的影响，但是也丢失了大量的有用信息。Lin[7] 在PCNN上进一步改进，通过引入注意力机制完成动态加权，解决了丢失大量信息的问题。Wu[11] 则在此基础上引入对抗性训练来提高关系抽取的鲁棒性和准确性，在词向量表示阶段加入了对抗性噪声并给出了新的损失函数。从目前的国内外研究现状和发展趋势来看，如何优化远程监督方法，减少噪声干扰，提高实体关系识别的质量是实体关系抽取领域急需解决的难点和热点问题。

本课题首次将实体关系识别应用于菜品推荐场景，传统的菜品推荐方法都是基于用户历史行为和偏好，实行个性化推荐[]，而该方法不适用于本课题的场景，因为我们无法获取评论网站或者微博用户的历史点餐信息，只能基于评论文本内容本身提取商家的菜品推荐信息，因此如何将实体关系识别应用于菜品推荐场景中，也是本课题的重点研究内容之一。

1. 客流预测

目前，国内外学者对交通领域的客流预测方法进行了大量的研究，很多研究人员都会将时间因素作为客流的重要特性融入到模型的构建过程中。最常见的就是使用时间序列模型对客流建模并进行分析和预测，比如ARIMA。郝勇等人[1]利用ARIMA方法对上海城市轨道交通的日客流量数据进行模拟和预测。这篇文章比较详细的描述了ARIMA阶数计算和参数选择的流程，包括平稳性检验，差分平稳性检验，自相关函数计算，AIC检验选择参数模型等。在火车客流预测中，Wei Xu等[2]人通过整合多种模型对春节期间的客流进行预测，建模过程中使用到了ARIMA模型的预测结果作为最终结果的一部分，建模过程还使用到了神经网络、线性回归、SVR等模型。Shu等人[3]利用SVR对高铁乘客的行为特征进行了深入分析，从他们的试验中可以知道，价格，速度，乘车时间，环境，安全性和整体满意程度是决定乘客选择高铁出行最主要的几个原因，同时证明了在高铁客流出行预测问题里，SVR比传统的Ligical模型准确率更高。赵宇刚等人[4]通过调研北京地铁乘客集散情况，特别是换乘站选址期间该地区乘客的集散特点，以及出现大量客流的原因，发现城市轨道交通的换乘站具有不同线路接驳和大量客流快速集散两方面特性。并且[4]研究得出换乘站选址应优先考虑是城市内部既有的或规划中的商业、居住、商务聚集区等重要的客流集散点和铁路客运站等大型城市对外交通中心。因此站点客流特点和站点周边的环境有较为密切的关系。陈蔚珊等人[5]利用聚类手段分析广州轨道交通枢纽站周边的商业演变情况，通过研究发现换乘站会带动站点周边商业的发展，从轨道交通的特点来看，换乘站能够产生较大的客流量，因此在地铁站周边的商家能够更加有效的吸引用户消费。而商业和以购物为目的的客流之间存在互相促进的作用，商业的发展必然会导致客流特点的变化，因此站点周边的商业繁华程度和站点的客流的特点存在着关联关系。在预测模型上，由于客流数据属于时间序列数据，对于时间序列数据的预测模型近年来深度学习模型有较好的表现。LSTM模型作为优化的循环神经网络模型对于前后有依赖的时间序列数据有较好的学习能力。文献[6]研究了室内时空客流获取与预测过程中的相关方法及技术，引入机器学习和深度学习方法构建模型，从而使室内时空客流的统计和预测更具实用性和准确性。常见的深度学习模型大多是针对单一时间序列的分析，对于多源数据的分析预测目前的研究较少。已有的成果有Qin Y等人[7]提出了一种双阶段的基于注意力的RNN网络，用于对室内温度进行预测。Li X等人[8]提出了一种包含上下文信息的LSTM-CL模型，用于预测血压，输入数据除了血压的历史数据外还有用户的其他身体状况数据，模型可以综合考虑多源数据进行预测。

目前已有的客流预测模型大多是只考虑了时间因素的针对单一客流数据进行的建模分析，常用的如回归模型、神经网络模型、ARIMA方法已经取得了较好的成果。但是一些学者研究发现了很多对客流有影响的因素，这些因素的加入对预测模型有着很大的帮助。如何将这些因素有效的融合到模型的学习中一直是研究的重点，目前的一些方法大多只是利用回归模型进行建模，将这些因素视作特征。很少有学者利用神经网络模型将对客流有影响的多因素加入到模型中，尤其是商家信息。因此针对多因素的多源数据的客流预测有着重要的研究意义

1. 路线推荐

路线推荐的核心在于推荐算法。推荐系统算法主要包括基于内容的推荐算法和协同过滤推荐算法，基于内容的推荐算法通过计算物品之间的相似度向用户进行推荐，协同过滤推荐算法通过用户的历史行为向用户进行推荐。文献[1,2]讨论了基于用户的协同过滤算法，根据用户的历史行为计算用户之间的相似度，然后把相似用户的兴趣点推荐给当前用户。基于用户的协同过滤算法随着用户数量的增多，计算时间会变长，而物品的增长速度通常慢于用户的增长。文献[3,4]提出了基于物品的协同过滤算法，向用户推荐与他之前感兴趣的物品相似的物品。上述两种方法在数据稀疏的情况下，预测性能会急剧下降，因此发展出了多种基于模型的协同过滤算法。文献[5,6]提出了基于聚类的推荐算法，首先对用户进行聚类，然后根据同一类簇内的其他用户的兴趣向用户进行推荐，这类算法在用户数量增加的情况下，聚类时间会相应增长。文献[8]使用奇异值分解(SVD)，根据用户对物品的历史反馈，将用户和物品影射成隐式向量，根据隐式向量计算他们之间的相关性。文献[9]扩展了SVD方法，同时考虑了用户和物品之间的隐式反馈，来计算用户和物品之间的相关性。文献[10,11]利用矩阵分解技术，从用户-物品评分矩阵中得到每一个用户和物品的潜在语义向量，利用潜在语义向量计算用户对物品的喜好程度。矩阵分解技术只能够得到用户和物品的浅层语义信息，深度学习模型可以得到用户的深层次的语义信息，因此研究者提出了大量基于深度学习模型的协同过滤算法。文献[13]将深度神经网络应用在YouTube推荐系统当中，综合考虑了用户的历史行为和属性信息。文献[14]扩展可矩阵分解技术，利用深度神经网络根据用户-物品的评分矩阵，计算用户和物品的语义向量，语义向量之间的余弦距离当作用户对物品的偏好。文献[15]除了考虑用户的历史行为信息，还考虑了用户的属性信息，增强了算法对新用户的处理能力。文献[16]扩展了协同过滤方法，利用神经网络根据用户和物品的语义向量计算用户对物品的偏好程度。

推荐系统算法能够很好的处理用户的历史行为，同时结合用户和物品本身的属性信息，向用户进行个性化推荐。目前的推荐系统算法对用户历史行为的理解还处在比较浅层的方面，并且只使用了少量用户和物品本身的属性信息。如何更深入的理解用户的历史行为，如何加入更多的辅助信息帮助推荐算法提升准确性，是推荐系统算法领域当前的研究方向和难点。同时，在轨道交通的路线推荐领域，目前的研究尚未将移动互联网的文本数据考虑在内，如何挖掘潜在的出行信息，并融合到路线推荐中，是本课题研究的重点内容。

总结相关领域的研究工作的优缺点和不足，。。。。。。

基于上述研究需求和本课题组的前期研究基础，以及对本课题需要解决的问题和方法的清楚认识和深刻理解，我们拟申请本课题。

# 参考文献

1. 金燕. 国内外UGC质量研究现状与展望[J]. 情报理论与实践,2016,39(03):15-19.

郝海涛,马元元. 基于加权关联规则挖掘算法的电子商务商品推荐系统研究[J]. 现代电子技术,2016,39(15):133-136.

禹蒲阳. CBA分类算法的一种改进[J]. 计算机应用与软件,2010,27(08):241-243+254.

S. S. Bucak, R. Jin and A. K. Jain, "Multi-label learning with incomplete class assignments," CVPR 2011, Providence, RI, 2011, pp. 2801-2808.

陈自洁. 多标签分类问题的图结构描述及若干学习算法的研究[D].华南理工大学,2015

PLuaces O, Díez J, Barranquero J, et al. Binary relevance efficacy for multilabel classification[J]. Progress in Artificial Intelligence, 2012, 1(4):303-313.

Multilabel classification via calibrated label ranking

Schapire RE, Singer Y. Boostexter: A boosting-based system for text categorization.[Machine Learning, 2000,39(2):135-168](http://dx.doi.org/10.1023/A:1007649029923).

Zhang ML, Zhou ZH. ML-KNN: A lazy learning approach to multi-label learning.[Pattern Recognition, 2007,40(7):2038-2048](http://dx.doi.org/10.1016/j.patcog.2006.12.019)

Zhang M L, Zhou Z H. A Review on Multi-Label Learning Algorithms[J]. IEEE Transactions on Knowledge & Data Engineering, 2014, 26(8):1819-1837.

朱郁筱, 吕琳媛. 推荐系统评价指标综述[J]. 电子科技大学学报, 2012, 41(2):163-175.

路永和,李焰锋. 改进TF-IDF算法的文本特征项权值计算方法[J]. 图书情报工作,2013,57(03):90-95.

Mikolov T，Sutskever I，Chen Kai，et al． Distributed representations of words and phrases and their compositionality［EB /OL］． ( 2013-10- 16) ． http: / /arxiv． org /pdf /1310．

Bengio, Y., Ducharme, R., Vincent, P., & Janvin, C. (2003). A neural probabilistic language model. The Journal of Machine Learning Research, 3, 1137–1155.

靳延安,李玉华,刘行军. 不同粒度标签推荐算法的比较研究[J]. 计算机应用研究,2012,29(02):504-509.

张朝恒,何小卫,陈勇兵. 基于社交网络信息的协同过滤推荐算法[J/OL]. 计算机技术与发展,2017,(11):1-11(2017-08-01).

张新猛,蒋盛益,李霞,张倩生. 基于网络和标签的混合推荐算法[J]. 计算机工程与应用,2015,51(01):119-124.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky.2009. Distant supervision for relation extraction without labeled data. In Proceedings of ACL-IJCNLP,pages 1003–1011.

Sebastian Riedel, Limin Yao, and Andrew McCallum.2010. Modeling relations and their mentions without labeled text. In Proceedings of ECML-PKDD, pages 148–163.

Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel SWeld. 2011. Knowledgebased weak supervision for information extraction of overlapping relations. In Proceedings of ACLHLT,pages 541–550.

Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher D Manning. 2012. Multi-instance multi-label learning for relation extraction. In Proceedings of EMNLP, pages 455–465.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014. Relation classification via convolutional deep neural network. In Proceedings of COLING, pages 2335–2344.

Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In Proceedings of EMNLP.

Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2016. Neural relation extraction with selective attention over instances. In Proceedings of ACL. volume 1, pages 2124–2133.

Thien Huu Nguyen and Ralph Grishman. 2015. Relation extraction: Perspective from convolutional neural networks. In Proceedings of NAACL-HLT. Pages 39–48.

Xiaotian Jiang, Quan Wang, Peng Li, and Bin Wang. 2016. Relation extraction with multi-instance multilabel convolutional neural networks. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. pages 1471–1480.

Yankai Lin, Zhiyuan Liu and Maosong Sun. 2017. Neural Relation Extraction with Multi-lingual Attention. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Volume 1, pages 34-43.

Yi Wu, David Bamman and Stuart Russell. 2017. Adversarial Training for Relation Extraction. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 1779–1784.

Meng Zhang, Yang Liu, Huanbo Luan and Maosong Sun. 2017. Adversarial Training for Unsupervised Bilingual Lexicon Induction. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pages 1959–1970.

Pengfei Liu, Xipeng Qiu, Xuanjing Huang. 2017. Adversarial Multi-task Learning for Text Classification. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pages 1-10.

程健一. 中文电子病历实体关系抽取研究[D].哈尔滨工业大学,2016.

陈涛. 多示例学习算法及其应用研究[D].华南理工大学,2013.

杨锦锋,于秋滨,关毅,蒋志鹏. 电子病历命名实体识别和实体关系抽取研究综述[J]. 自动化学报,2014,40(08):1537-1562.

黄蓓静,贺樑,杨静. 远程监督人物关系抽取中的去噪研究[J]. 计算机应用与软件,2017,34(07):11-18+31.

石锋. 面向中文新闻文本的实体关系抽取研究[D].哈尔滨工业大学,2017.

刘凯,符海东,邹玉薇,顾进广. 基于卷积神经网络的中文医疗弱监督关系抽取[J]. 计算机科学,2017,44(10):249-253.

郝勇, 朱海燕. 基于客流n日均量的地铁客流量的时间序列分析[J]. 铁道运输与经济, 2009, 31(10):42-46.

Xu W, Qin Y, Huang H. A new method of railway passenger flow forecasting based on spatio-temporal data mining[C]// The, International IEEE Conference on Intelligent Transportation Systems, 2004. Proceedings. IEEE, 2004:402-405.

Shu K, Jing L, Mei L, et al. Prediction based on support vector machine for travel choice of high-speed railway passenger in China[C]// International Conference on Management Science and Engineering. IEEE, 2011:28-33.

赵宇刚, 毛保华, 刘明君,等. 我国城市轨道交通换乘车站选址规划相关问题研究[J]. 物流技术, 2011, 30(5):1-3.

陈蔚珊, 柳林, 梁育填. 广州轨道交通枢纽零售业的特征聚类及时空演变[J]. 地理学报, 2015, 70(6):879-892.

李邦鹏. 基于深度学习的室内时空客流预测[D]. 浙江大学, 2017.

Qin Y, Song D, Chen H, et al. A Dual-Stage Attention-Based Recurrent Neural Network for Time Series Prediction[J]. 2017:2627-2633.

Li X, Wu S, Wang L. Blood Pressure Prediction via Recurrent Models with Contextual Layer[C]// International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, 2017:685-693.

Breese J S, Heckerman D, Kadie C. Empirical analysis of predictive algorithms for collaborative filtering[C]//Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence. Morgan Kaufmann Publishers Inc., 1998: 43-52.

Resnick P, Iacovou N, Suchak M, et al. GroupLens: an open architecture for collaborative filtering of netnews[C]//Proceedings of the 1994 ACM conference on Computer supported cooperative work. ACM, 1994: 175-186.

Linden G, Smith B, York J. Amazon. com recommendations: Item-to-item collaborative filtering[J]. IEEE Internet computing, 2003, 7(1): 76-80.

Sarwar B, Karypis G, Konstan J, et al. Item-based collaborative filtering recommendation algorithms[C]//Proceedings of the 10th international conference on World Wide Web. ACM, 2001: 285-295.

Ungar L H, Foster D P. Clustering methods for collaborative filtering[C]//AAAI workshop on recommendation systems. 1998, 1: 114-129.

Yu Y, Wang C, Gao Y, et al. A coupled clustering approach for items recommendation[C]//Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer, Berlin, Heidelberg, 2013: 365-376.

Miyahara K, Pazzani M J. Collaborative filtering with the simple Bayesian classifier[C]//Pacific Rim International conference on artificial intelligence. Springer, Berlin, Heidelberg, 2000: 679-689.

Billsus D, Pazzani M J. Learning Collaborative Information Filters[C]//Icml. 1998, 98: 46-54.

Koren Y. Factorization meets the neighborhood: a multifaceted collaborative filtering model[C]//Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2008: 426-434.

Koren Y, Bell R. Advances in collaborative filtering[M]//Recommender systems handbook. Springer US, 2015: 77-118.

Koren Y, Bell R, Volinsky C. Matrix factorization techniques for recommender systems[J]. Computer, 2009, 42(8).

Mnih A, Salakhutdinov R R. Probabilistic matrix factorization[C]//Advances in neural information processing systems. 2008: 1257-1264.

Covington P, Adams J, Sargin E. Deep neural networks for youtube recommendations[C]//Proceedings of the 10th ACM Conference on Recommender Systems. ACM, 2016: 191-198.

Xue H J, Dai X Y, Zhang J, et al. Deep Matrix Factorization Models for Recommender Systems[J]. static. ijcai. org, 2017.

Dong X, Yu L, Wu Z, et al. A Hybrid Collaborative Filtering Model with Deep Structure for Recommender Systems[C]//AAAI. 2017: 1309-1315.

He X, Liao L, Zhang H, et al. Neural collaborative filtering[C]//Proceedings of the 26th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, 2017: 173-182.

2．**项目的研究内容、研究目标，以及拟解决的关键科学问题**（此部分为重点阐述内容）**；**

# 研究内容

本课题针对商家评论和客流模型相结合的路线推荐中的理论问题展开研究工作，具体工作如下:

1. 以轨道交通站点为坐标的商家评论优化，挖掘商家评论和微博评论的文本数据，提取与出行相关的商家特征，并进行优化；
2. 基于天气、季节、商家和客流等多因素的实时客流建模，研究多因素的向量表示和利用深度学习模型的客流建模；
3. 结合商家和客流的路线推荐，研究利用深度学习模型、协同过滤方法的路线推荐，研究商家、客流、历史行为数据对推荐的影响情况。

课题的研究工作涉及文本挖掘、客流预测、多因素融合的推荐方法，同时研究上述理论问题在轨道交通出行中的应用方案。具体研究内容如下：

### 基于评论和微博数据的商家标签优化

本课题文本数据的来源主要由两部分组成:商家评论文本和微博文本。当前主流的点评网站上，存在着标签不完善和数据不完善的问题，而本课题将通过多标签分类和实体关系识别的方法来优化商家的标签，并解决商家标签不够准确，不够完善且维度不统一的问题，最终从文本类型的数据中挖掘商家、商家提供的服务内容—如菜品、环境、服务、口味等标签。课题将从下述两方面研究标签的提取和优化:

* 服务内容类型标签提取（如：环境、服务、口味、交易方式等）

商家标签代表了商家特征，一般是来自于用户的点评信息。现有商业系统中标签提取方法保证了提取效率，准确度较低，标签缺失问题较严重，不适用于后续算法直接处理。因此完善商家标签，提升标签的质量，是后续计算的基础。本课题提出了一种嵌入语义的多标签的方法，通过将点评文本映射为多标签向量实现对文本关联性的量化，为后续量化和计算建立基础。

1. 文本表示的信息包含度和准确度很大程度决定了分类模型的最终效果。因此如何取得较好的点评文本表示是获取商家标签中重要的一环。本课题基于文本表示的用途提出对每个样本使用两套向量化方案。一个是使用one-hot向量用于语义分析，为模型额外提供了标签关联信息从而优化了模型效果；另一种是使用word2vec和TF-IDF用于生成基础特征向量。
2. 在提高算法准确率的同时也要注重算法的复杂度，LDA语义模型的主要参数是主题数目k是一个超参数，为了找到最优k值，往往需要设置的搜索区间往往非常广，随着样本数量n和标签类别数量m的增长，算法的时间代价可能达到难以承受的地步。本课题研究上述多标签分类模型的优化策略，通过提出一种最优k值区间预搜索算法以将算法复杂度降低到一个可接受的范围。
3. 如何将大量的微博数据利用起来以辅助商家标签的构建。本课题拟采用的方法是先设计一个针对微博文本的二分类器，以提取出商家评论微博；然后将是商家评论的微博作为上述多标签分类模型的输入。这样增加了评论数据使得商家标签的获取更加准确。

* 菜品标签的提取

菜品标签的提取可转化为实体关系识别问题，识别菜品和商家之间的关系(是否推荐)，目前在实体关系识别领域，面临的主要问题是远程监督方法会引入较多噪声，因此本课题拟提出基于多示例和注意力交换的方法对训练集进行降噪处理，研究内容如下:

1. 结合多示例算法来研究实体关系的抽取。多示例算法将每个实体对视为一个包，将所有包含该实体对的句子视为包中的示例。对实体关系的分类是基于包而不是基于示例。将多示例算法应用于关系抽取的好处在于，可以通过为包构建合理的特征表示，来减弱包内噪声示例带来的负面影响。
2. 基于注意力机制进行包间示例的转移。对于噪声示例而言，其真实表达的关系和标注的关系是不同的，因此通过示例的转移，可以将噪声转移到正确的关系下，从而剔除噪声。
3. 研究实体关系抽取方法在商家菜品标签提取场景上的应用。本课题需要基于文本评论和微博数据，将评论中的商家和菜品作为实体对，提取商家和菜品的关系。

### 基于多维数据的客流预测模型构建

本研究点综合考虑站点，气象，节日，周边商家等对地铁客流的影响，研究设计网络模型对客流进行有效预测。具体研究内容如下：

1. 找出对客流数据有影响的相关因素并形成结构化的特征。对客流数据、周边商家数据、气象、节日等相关数据进行建模统计，并利用特征选取等特征工程进行分析。找出会影响客流的主要因素，并用关联规则挖掘等技术挖掘出有价值的规则属性。同时，形成数值结构化的特征用于后续模型的输入。
2. 对客流数据进行向量化编码表示，将站点属性、商家等信息加入到客流的编码表示中。设计编码器对数值属性的客流数据进行向量化表示，可以融合类别和数值特征。考虑多因素的客流数据向量化表示对预测效果有积极作用。
3. 利用基于注意力机制的神经网络模型对客流预测进行建模。设计多阶段的神经网络模型。第一个阶段，在输入部分引入注意力机制，自适应地学习多源输入之间的相互关系，对地铁全网线路之间的关系进行学习。第二个阶段，利用带注意力机制的LSTM层提取客流量数据在时间维度上的前后关系。最终利用模型对客流数据进行有效预测。

### 基于商家标签和客流的路线推荐模型构建

本研究点综合考虑用户的历史行为、站点周围的商家、商家标签、和站点的客流量信息，向用户进行目的地推荐。为了使目的地推荐获得更加准确的效果，需要对站点构建画像，并且对用户的历史行为进行建模。本研究点的数据按照分类方式不同可分为：数值型和文本型；离散型和连续性；静态、动态和实时类型。数据种类多并且数据量大，综合考虑这些数据是一大难点和挑战。针对以上任务，本研究点的研究内容分为以下几个方面：

1）多种类型的数据建模方法。每种类型的数据都有各自的特点，本研究点根据不同数据的特点选用不同的数据建模方法。

2）地铁站点画像构建。地铁站点画像是进行目的地推荐的重要依据，站点画像构建主要包括以下几个方面：站点地理位置及与其他站点之间的拓扑关系等；站点周围商家情况，包括商家类型、数量和商家标签等等；站点客流量信息，包括站点各个时段客流量大小和客流高峰时间分布等。

3）设计实现基于站点周围商家和客流信息的路线推荐模型。模型根据用户的历史行为、商家信息、站点画像和站点客流量等信息，向用户推荐可能感兴趣的站点。

# 研究目标

本课题拟通过对移动互联网文本信息的挖掘提取出商家标签，并将商家信息融合到客流预测模型中，构建基于多维数据的客流模型，最终结合商家信息和客流模型完成路线推荐，辅助轨道交通的规划，从而将两者有机地结合起来。本课题的研究目标是：

1）优化商家标签，通过多标签分类提取服务类标签，通过实体关系识别提取菜品类标签，解决当前商家标签不准确，不完善的问题并将标签统一到相同维度中；

2）基于多维数据构建客流预测模型，结合商家标签、气象、时间等维度，提高客流预测的准确度；

3）基于商家标签和客流构建路线推荐模型，既考虑用户的出行意愿，又结合实际的客流情况，从而提高出行效率。

# 拟解决的关键科学问题

要完成上述研究内容和目标，需要解决以下关键科学问题：

1. 商家标签优化

本课题拟采用多标签分类来完成服务类标签的提取，需要解决以下问题:

1. 多标签分类的输出空间降维问题

输出空间降维的主要目的是降低需要处理的标签规模，减少分类器复杂度。输出空间降维需要解决的问题是向量间的编解码。在这一过程中需要尽可能对标签向量的信息进行保真，高维数据向低维数据转换的过程中伴随着信息的损失，而过度的信息损失会对模型的效果产生极大影响。

1. 多标签分类的标签关联性挖掘

多标签分类的难点是如何将标签的关联建模与模型的优化求解结合起来。以Label Ranking为例，算法是利用关联标签和非关联标签的关系构造排序结果进而转化为二阶关联模型，这种启发式的建模方式缺乏样本信息的监督，因此在准确度指标上很难取得突破性的成果。

1. 基于语义识别的标签关联分析方法的参数优化问题

语义模型的参数优化直接影响了整个多标签分类模型的效率，因为不仅要关注算法的准确性，同时也要将算法的复杂度降低到了一个可接受的范围。

针对菜品类标签的提取，本课题将其视为实体关系识别问题，识别菜品和商家的关系。需要解决的问题如下:

1. 基于注意力进行包间示例转移

通过注意力机制，可以得到示例在不同关系注意力下的权重，如何基于该权重，将示例转移到正确的关系下，是识别实体关系需要解决的关键问题之一。

1. 将实体关系识别应用于菜品标签提取场景

实体关系识别作为通用性的基础研究算法，要应用到菜品标签提取场景中，需要进行适应性的改进，如何将商家和菜品作为实体对，抽取商家和菜品之间的推荐关系，也是待解决的关键问题之一。

1. 客流预测
2. 分析对地铁客流量造成影响的主要因素

实际场景中，客流量不仅仅与其历史数据有关，往往还与许多其他方面的其他因素息息相关。比如气象、节日和周边商家的情况等等。这些因素都会对路网造成或多或少的影响。如何找出具体影响客流量的特征对于后续模型的挖掘有着重要的意义，也对预测模型的预测准确率有着重要的意义。如何准确找出影响客流的特征需要利用一系列特征工程进行挖掘，如相关性，方差，特征降维，特征筛选等等手段。

1. 对客流数据进行合理有效地向量化编码

向量化表示常见的场景如在自然语言处理中的词向量表示，训练一个简单的多层神经网络模型，利用中心词预测其上下文的词语。最后将训练好的网络参数取出，即为词语的向量化表示，该向量化表示考虑的是词语之间的共现关系。在客流数据中，数据本身就是数值数据，研究难点在于设计编码器模型，将客流数据和相关的特征一起训练，并利用词向量的思路，取模型的中间参数作为最终的客流向量化表示。

1. 针对多源数据设计有效客流预测模型

常见的循环神经网络模型对于单一的时间序列数据有较好的学习能力，能学习长期的时间依赖关系。但是不能直接处理多源的时间序列数据，针对此，研究设计可以接收多源输入的预测模型，模型可以自动学习多输入之间的相互关系，并有效预测目标时间序列数据。

1. 路线推荐

（1）地铁站点画像构建

地铁站点画像要尽可能全面的包含站点的特征信息。除了考虑站点所处的地理位置之外，还要考虑站点的客流量和客流高峰的时间分布。移动互联网上站点周围商家的评论文本同样包含大量与站点有关的隐含信息，挖掘出这些评论里面隐含的信息，会在很大程度上提升推荐的效果。

（2）基于商家标签和站点客流量的路线推荐模型

常见的推荐模型只考虑用户的历史行为，推荐结果准确性不高。将用户的历史行为、商家信息、站点画像和站点客流量等信息，尤其是商家的评论文本加入到模型中，可以提高推荐的准确性。输入数据的维度多，数据量大，需要模型有很好的性能

3．**拟采取的研究方案及可行性分析**（包括研究方法、技术路线、实验手段、关键技术等说明）；

# 研究方案

本课题拟通过移动互联网的文本数据来挖掘和出行相关的商家特征，并和客流模型相结合，最终构建出基于商家和客流的路线推荐模型。整体方案分为三部分内容，第一部分利用文本挖掘技术，得到站点周边的商家标签，该部分包含了潜在的用户出行特征，将作为后续客流模型和路线推荐模型的基础；第二部分则结合了商家标签，天气，时间等多因素构建了客流模型，该部分对交通出行中的客流情况进行了有效合理的建模，为后续的路线推荐提供了重要参考；第三部分则融合了上述的商家标签信息和客流信息，构建了路线推荐模型，既考虑到了用户的出行意愿，同时可以减缓交通压力，从而改善了出行效率。整体路线图如下所示:

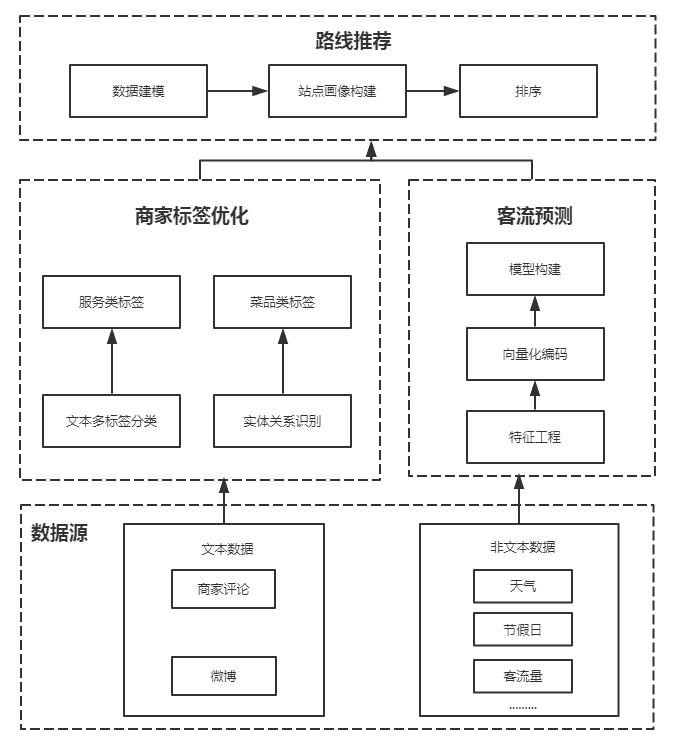


图 整体技术路线图

### 基于评论和微博数据的商家标签优化

本课题基于评论和微博文本数据优化商家标签，并将标签分为服务类和菜品类。其中服务类标签采用多标签分类算法进行提取，可以提高标签的准确性，并且将标签统一到相同的维度下。菜品类标签则采用实体关系识别的方法进行提取，因为菜品类标签数量过于庞大，无法采用多标签分类，因此将其转化为实体关系识别问题。整体流程如下:

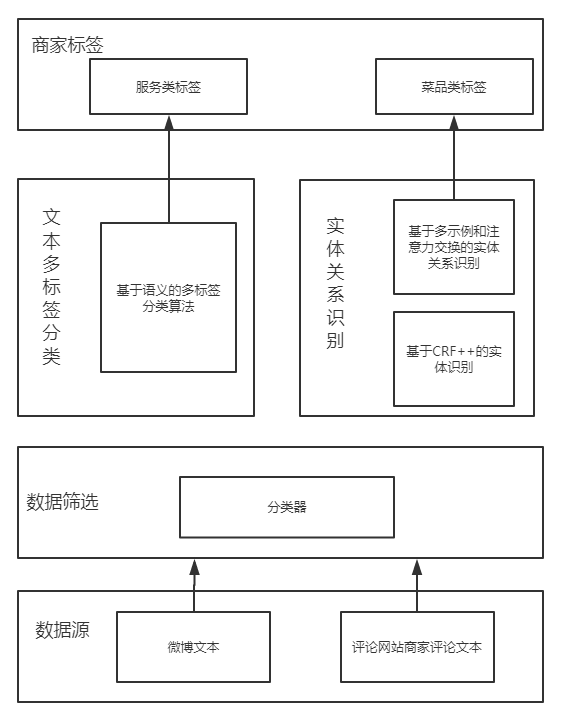


图 商家标签优化流程图

* + - 1. 微博数据筛选

本课题研究内容的数据源有两种，一种是美团、大众点评网站下的各种评论数据；另一种是微博中针对城市轨道交通站点周边商家的文本数据。其中网站下的评论数据针对性较好，无需对数据进行筛选清除操作即可用本课题提出的基于语义嵌入的文本多标签分类模型进行标签的补充；而微博中针对商家的数据包含许多与评论无关的数据，因此若想利用微博中的数据作为网站评论数据的扩展以给商家打标签，首先要过滤与评论无关的数据。

微博数据筛选的目的是清除不是评论的数据，而保留属于商家评论的数据。本课题通过设计一个二分类器来达到此目的。其中训练集由网站商家评论作为正样本，由微博中新闻、报告等作为负样本所共同组成。二分类模型图如下：

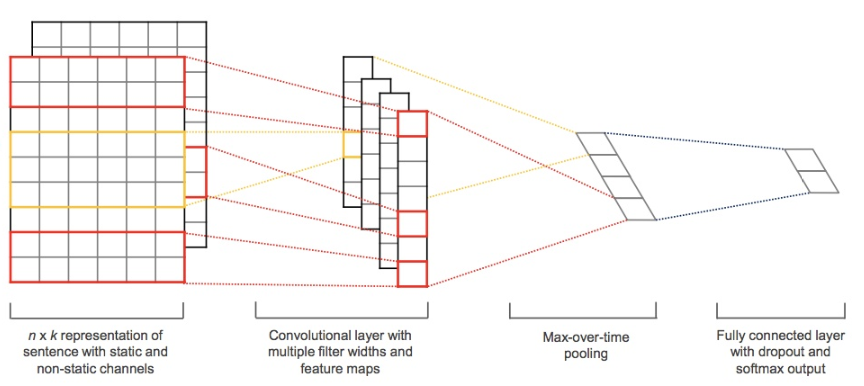


图 微博数据分类模型

因为微博文本大多都是短文本，因此本课题采用针对短文本效果较好的textCNN模型进行微博文本分类。详细过程是第一层是上图最左边的n\*k的句子矩阵，每行是词向量，这里使用word2vec训练得到的词向量维度为k, 这个可以类比为图像中的原始像素点了。然后经过有filter\_size=(2,3,4) 的一维卷积层，之所以选择不同size的卷积窗口是为了获取不同词数级别的语义特征。第三层是一个1-max pooling层，这样不同长度句子经过pooling层之后都能变成定长的表示了，最后接一层全连接的 softmax 层，输出每个类别的概率。

经过这个分类模型后，属于商家评论的微博文本和网站评论文本即可作为本课题后续提出的基于语义嵌入的文本多标签分类模型的输入，输出为商家的多个标签。

* + - 1. 基于语义的文本多标签分类模型的服务类标签提取

传统的多标签分类算法通常将标签看作独立的，然而现实中标签通常具有相关性，如服务类标签中“干净”标签和“环境好”标签通常有较强的相关性。因此本课题提出基于语义嵌入的文本多标签分类模型，以解决原始Binary Relevance方法在标签关联挖掘上的不足。

本课题提出的模型嵌入了语义分析模块，通过对文本标签的语义分解完成了关联建模，语义分析模块作为所有单标签分类器的公共模块能够连接所有标签，同时在模型整体参数优化的过程中还包含了对语义模块的参数调整，因此模型可以利用所有样本信息以及样本—标签交互信息来优化语义分析模块，突破了传统算法在标签关联挖掘上的不足。

* + - * 1. 模型结构

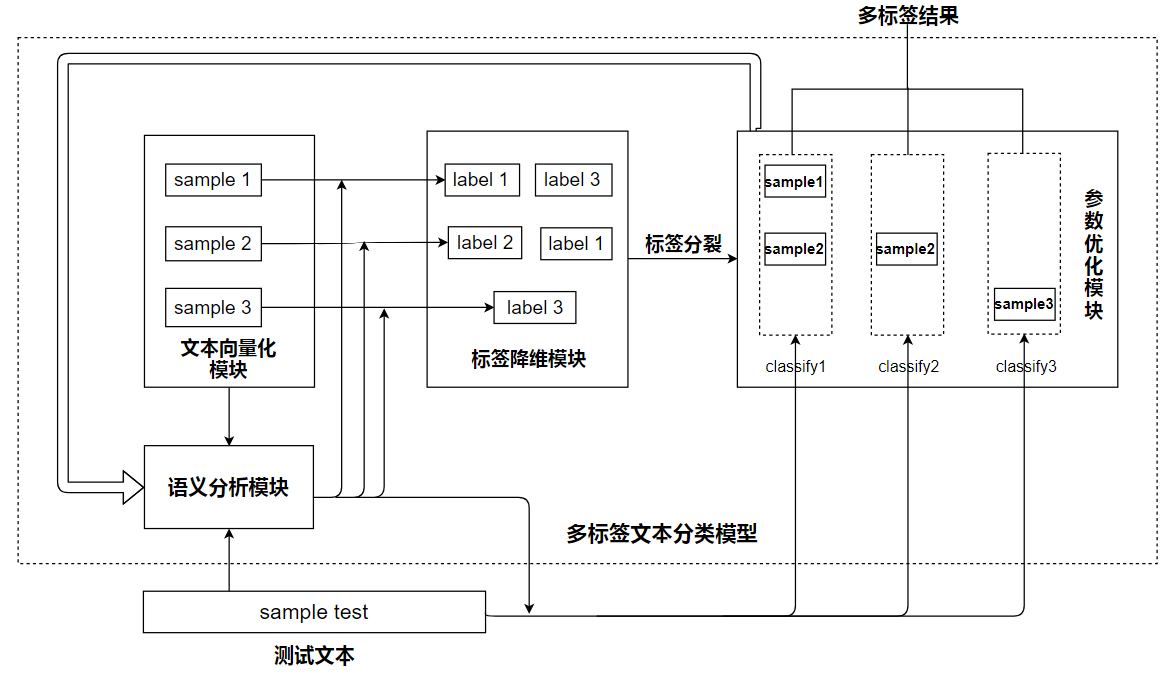


图 基于语义嵌入的多标签分类模型结构

上图详细的展示了基于语义嵌入的文本多标签分类模型的结构。图中的虚线部分代表了整个多标签分类模型。模型包括文本向量化模块、标签降维模块、语义分析模块、模型优化模块。其中文本向量化模块负责对输入的文本信息进行向量空间的映射，标签降维模块负责对多标签样本进行简化拆解，语义分析模块负责提取每个文本的语义信息融入模型的输入特征向量中。模型优化模块需要完成三部分内容，一是目标函数的设计，二是使用合适的基分类器组合多标签分类器，三是设计模型整体优化方案。

完整的多标签模型应该由四部分构成：特征工程、模型结构设计、确定目标函数以及设计优化策略。特征工程的工作由文本向量化模块、标签降维模块、语义分析模块协同完成，而模型结构、目标函数以及优化策略的设计与实现则由模型优化模块完成。

* + - * 1. 文本向量化模块

文本向量化的目标是为了将文本信息表示成计算机可以理解的形式，从而为后续文本分析模块提供操作基础，顺利完成自然语言处理任务。在分类模型中常用的文本向量化方法分为哑编码表示（One-hot representation）和词嵌套（Word Embedding）。

本模块所生成的文本向量主要用于两方面工作：一是文本语义分析；二是用于构建文本基础特征。

对于文本基础特征向量的构造，需要关注以下三点内容：向量大小、数据稀疏性、以及特征表达能力。word2vec作为一种词向量表达方式在控制向量大小和解决数据稀疏性上有着独特的优势，不过遗憾的是，word2vec作为神经网络模型的产物在传统分类模型上应用效果不佳，在特征表达力上劣于one-hot representation。但是如果将word2vec和神经网络分类模型结合却可能取得比传统分类器组合更好的效果，其原因在于神经网络模型在训练过程中还会根据现有数据动态的调节word2vec的词向量，使之更加拟合数据环境。因此本模块同时保留两类向量化方案并根据后续基分类器的选择动态调整所使用的方法。同时介于TF-IDF在各类文献中相比于BOW的效果提升，模块选择TF-IDF作为one-hot representation的代表方法。

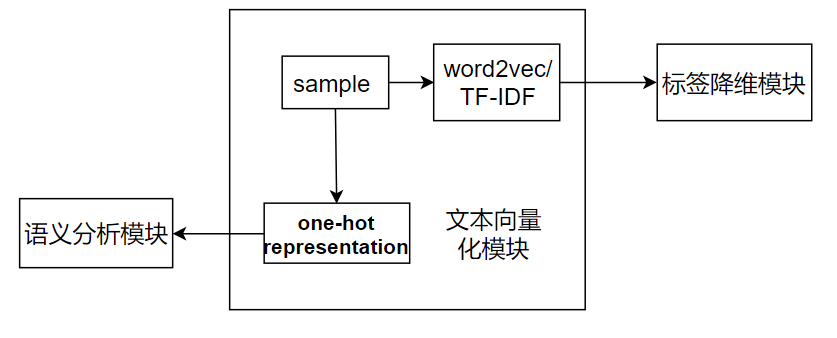


图 文本向量化模块示意图

上图展示了本模块的具体实现细节，对于每个样本使用两套向量化方案，one-hot向量用于语义分析，word2vec和TF-IDF用于生成基础特征向量。word2vec向量可以使用谷歌开源工具包生成，one-hot和TF-IDF则通过对文档集的词语统计生成。具体流程如下：

统计文档集D中的总词数记为N,并用N维0向量初始化每个文档。

对所有，统计词数并更新文档向量对应位置上的值，这就是one-hot方法下的文本向量。

TF-IDF在此基础上引入了IDF加权，根据文档集D计算每个词的IDF值，再使用这些IDF值更新one-hot向量每一维的权重即可得到TF-IDF向量。

表格 1 不同向量化方法的困惑度指标对比

|  |  |  |  |
| --- | --- | --- | --- |
| 模型 | word2vec | TF-IDF | one-hot |
| LDA | 1652 | 1276 | 1195 |
| pLSI | 1723 | 1312 | 1303 |

上表给出了不同向量化方法在语义分析上的指标对比，这里随机选取30W知乎数据作为测试集，测试向量化方法在困惑度指标上的效果。可以看出one-hot方法更加适合进行语义分析。

* + - * 1. 标签降维模块

标签降维模块包含两部分内容：一是整合样本特征；二是对文本多标签进行降维。

1. 样本特征整合：

标签降维模块的输入信息分为两部分，一是从文本向量化模块获取的基础特征向量，二是从语义分析模块获取的语义特征向量。

基础特征向量的两种类型：word2vec和TF-IDF。word2vec对每个词进行向量化，然后将样本的所有词向量按词序首尾相连形成文本向量。这种方式带来的问题是样本维度的不一致，因为样本之间的长度通常是不同的。为了对word2vec特征进行规范，可以采用填补和切分两种方法。但是填补极大的增大了分类器的负担，且在样本长度偏差较大的时候会出现向量稀疏问题。切分方式以一定的样本信息为代价解决了填补方式可能存在的问题。一般情况下，切分方式和填补方式在模型准确度上的差别不大，而在算法时间效率上切分的优势却十分显著，因此本课题使用切分作为word2vec向量的规范方法。

TF-IDF属于One-hot representation，在向量维度上是统一的，因此无需进行规范化。语义分析的输出向量同样是One-hot representation，对于这两类特征的整合可以直接使用向量连接操作完成。word2vec类型的基础向量则需要规范化后与语义向量连接。另外在这里还需要记录整合后的连接点位置，因为部分模型对基础特征有特殊的变换操作，后续将详细说明。

1. 标签降维

本课题选择Binary Relevance对标签集进行降维，单独考虑每个标签在样本中的分布，构造专属于这个标签的正负样本，从而获得了完整的二分类数据。依次对所有标签重复操作可以实现从多标签分类到单标签分类的降维，下面用形式化的语言详细描述该过程。在此之前，按下表进行预定义。

表格 2 标签降维符号定义表

|  |  |
| --- | --- |
| 符号 | 含义 |
| L | 标签集 |
| S | 样本集合 |
|  | 样本特征 |
|  | 样本标签集合 |
| B | 输出的二分类样本集合 |

1. 设Bl={}，从标签集中选择任意标签 ，若L为空集则结束。
2. 对所有，若 ，C=T ，Cl=1；如果 ，C=T ，Cl=0。将新样本C添加到Bl中。
3. 输出Bl到标签l对应的二分类器中。从L中删除标签l，跳转到步骤1。
   * + - 1. 语义分析模块

语义分析的目的是对整个文本集进行语义建模，从而挖掘出文本的潜在语义特征。更进一步，能够通过文本的语义信息对标签进行向量化建模，从而深度挖掘多标签之间的关联关系，提高多标签分类模型的效果。

1. LDA语义分析方法

本模块选用LDA作为语义分析模型，并对原模型做了改进，就是对LDA收敛边界的自动识别。

Gibbs抽样是主流的LDA参数求解方法。Gibbs抽样以马尔科夫链的收敛性为基本原理，每次对联合分布的其中一个分量进行抽样，剔除此分量重新计算联合分布并根据新分布求解此分量的值，对这个过程循环可得到最终的联合分布。Gibbs在复杂环境下的求解十分简单，但存在的问题是收敛条件不好判断。针对此问题，本文提出一种基于文本困惑度指标的收敛边界识别方法。困惑度(perplexity，PP)作为语言模型中常用的评价指标可以用来衡量模型对现有文档集的拟合程度，具体公式如下：

C:\Users\Gsy\AppData\Roaming\Tencent\Users\835373492\QQ\WinTemp\RichOle\9P}J%DW2B]{M8QEMSN(JHP9.png

其中，代表文档集中出现的每一个词在语言模型下出现的概率，分母N是文档集中的总词数。困惑度是语言模型中较为权威的衡量指标， LDA训练的目标也是为了使模型拥有一个最小的困惑度，因此可以通过困惑度值的变化来判断LDA的收敛情况。具体步骤如下：

1. 对文档集中所有词w随机初始化主题，同时计算对应的文档-主题分布U和主题-词分布V。
2. 对于所有(D代表训练文档集),将w抽离文档集，更新U和V两个分布。同时根据U，V按如下公式给w分配主题,其中代表所有主题的集合。和可以通过U和V获取。

C:\Users\Gsy\AppData\Roaming\Tencent\Users\835373492\QQ\WinTemp\RichOle\V@ZTY`UV(4]N(EGINMS(U3R.png

1. 计算困惑度，同时计算困惑度提升速度：。如果提升速度小于阈值λ，则LDA模型已经收敛，输出结果。若仍大于阈值，则更新并跳转至步骤1，继续对模型迭代。

上述步骤中无需每次迭代都进行提升速度的计算，实验表明在模型可能在1至2次迭代过程中没有提升甚至效果下降，但在后续的多次迭代后能取得不错的提升。因此可以以多次迭代为间隔计算提升速度。实验数据表明20次左右的区间得到的提升函数较为平滑。

1. 对文档集中所有词w随机初始化主题，同时计算对应的文档-主题分布U和主题-词分布V。
2. 对于所有(D代表训练文档集),将w抽离文档集，更新U和V两个分布。同时根据U，V按如下公式给w分配主题,其中代表所有主题的集合。和可以通过U和V获取。

C:\Users\Gsy\AppData\Roaming\Tencent\Users\835373492\QQ\WinTemp\RichOle\$L8FH$1)W_`299QGJV96WA2.png

1. 计算困惑度，同时计算困惑度提升速度：。如果提升速度小于阈值λ，则LDA模型已经收敛，输出结果。若仍大于阈值，则更新并跳转至步骤1，继续对模型迭代。

上述步骤中无需每次迭代都进行提升速度的计算，实验表明在模型可能在1至2次迭代过程中没有提升甚至效果下降，但在后续的多次迭代后能取得不错的提升。因此可以以多次迭代为间隔计算提升速度。实验数据表明20次左右的区间得到的提升函数较为平滑。

1. 标签关联挖掘分析

本模块进行文本的语义分析，其目的在于通过语义挖掘标签集之间的潜在联系。标签是文本在特定视角下的一种分类，因此标签之间通常在语义上拥有重合关系，进而演变为了文本的多标签表示。

如果同一类标签-语义的组合在多个训练样本中出现则可以认为标签很可能是包含这些语义的，因此针对某个标签，可以通过对所有关联样本的语义统计获得标签的语义向量。大量文本的训练能够促使向量向真实分布靠拢。

文本标签集建模的目的在于找到合适的标签表达方法描述标签间可能存在的关联关系，本课题使用语义向量进行标签建模有以下两点优势，其一文本标签作为一种文本类别本质上是多维语义的糅合，因此建模方式具有理论上的支持；其二向量化的数据形式相比于图结构在大规模量化运算上有很大优势。下面简述本文的标签建模流程：

1. 使用LDA对训练集进行语义分析，确定语义数目k和每个样本的语义分布。
2. 初始化所有标签向量，向量长度为语义数量k，初始值为0。
3. 对于所有(D代表训练文档集)，根据语义分布得到d的语义向量更新样本d对应标签的向量，对于所有的(代表文档d对应的标签集)，。
4. 对所有标签向量进行归一化处理即可得到最终的标签向量建模。
   * + - 1. 模型优化模块

前面几个模块已经完美的完成了特征工程方面的任务，本模块需要完成模型构造相关的工作，本模块将根据主流基分类器设计出契合的多标签模型结构，同时模块还会就模型优化效率的问题进行深入研究。

1. 模型结构设计

多标签分类与传统分类最大的区别在于样本的类别不是确切的单类别，因此很难将传统分类算法应用于多标签分类。本文采用Binary Relevance方法处理上述问题。针对不同的基分类器，需要以不同的模型结构去适应。

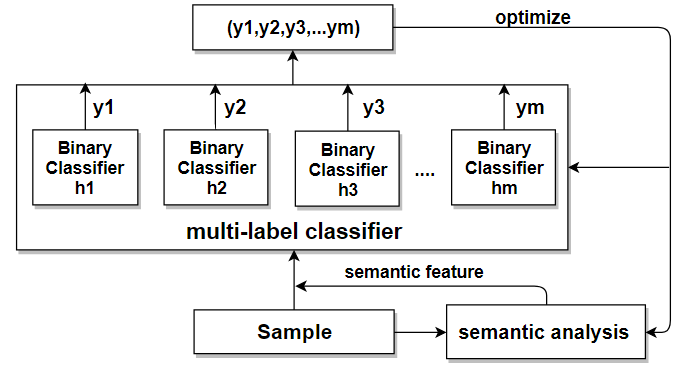


图 4 传统分类模型下的多标签结构图

传统机器学习方法在结构上相对简单，模型间的融合性也较差，因此可以用完全独立化的方法来组合分类器，具体结构如图所示。模型在一阶方法的基础上引入了语义分析模块，通过语义特征影响所有子分类器的结构，模型在优化过程中不仅有对子分类器参数的优化同时还包含了对语义分析参数的优化，文本的语义分布是标签在语义空间上的分解，每个标签代表特定的语义组合，因此我们可以用文本的语义分析来解释标签之间的关联关系，从而提升多标签文本分类的效果。用X代表样本输入空间，代表样本输出空间，多标签模型可以定义为一个X→Y的映射。因此模型的输出可以定义为式4-19的形式:



图中代表m个二分类器，为整体模型对sample的预测结果,y的每个分量取值区间为[0,1]，代表样本属于对应标签的概率，按照图4-9结构易知 。需要注意的是式4-19中的x代表样本的特征向量，依前文所述，这里的特征向量由样本基础向量和语义特征向量组合而成，语义特征为模型额外提供了标签关联信息从而优化了模型效果，本文将这种方法称为语义特征增强。在传统分类模型下的语义嵌入中，保持语义特征不变，分类器参数的优化等价于各个子分类器参数的独立优化。

1. 优化策略设计

本课题提出的多标签分类模型的参数主要是LDA语义分析模型参数。

LDA语义模型的主要参数是主题数目k，与一般的参数不同，主题数k是LDA模型的超参数，k的取值能够直接决定LDA模型的结构，因此难以使用梯度下降这种方式对参数k进行优化。针对上述问题，本课题提出了一种基于线性搜索的参数拟合方法。注意到k只能是整形数值，且实际应用中LDA模型的最优k值也通常在一个较为稳定的范围内波动，本课题提出了一种最优k值区间预搜索算法。

为了找到最优k值，需要设置的搜索区间往往非常广，随着样本数量n和标签类别数量m的增长，算法的时间代价可能达到难以承受的地步。

为了解决上述问题，本文提出了一种预搜索算法用于缩小参数k的搜索区间。算法的核心思想是通过LDA模型自身的最优k值(记为),来确定多标签模型的最优k值(记为)的波动区间。算法采用前文提到的困惑度(preplexity，PP)作为LDA的评价标准。此时，为最小化情况下主题数k的取值。算法使用基于密度的自适应LDA方法[28]确定的取值。

以上小节详细阐述了基于语义嵌入的文本多标签分类模型的各个组成模块的功能和实现方式，最终可以得到一个考虑到标签关联性且复杂度适中的多标签分类模型。将微博中的评论数据与商家评论数据作为该分类模型的输入，则输出为每条评论中的各个服务标签如“服务好”、“菜品佳”、“口味好”等。对所有评论的各个服务标签统计最终可以得到可作为部分商家画像的商家服务类标签。

* + - 1. 基于实体关系识别的菜品类标签提取

和服务类标签不同的是，菜品类标签是为了提取出该商家值得推荐的菜品，从而为用户的用餐选择提供更直观的指导意义。在本课题中，可以将该问题转化为实体关系识别问题，通过判别出现在评论文本中的菜品和商家之间是否是推荐关系，来完成最终的标签提取。和基于规则的方法相比，实体关系识别不需要认为制定大量的规则，可拓展性和鲁棒性好，可以节省大量的人力，从而提高效率。

基于实体关系识别的菜品类标签提取主要分为两步，菜品实体识别和关系识别，由于现有的命名实体识别研究成果较为丰富，并且有现成工具可以使用，因此不是我们的研究重点，本课题将直接使用CRF++工具进行菜品实体识别，并重点阐述实体关系识别方案。

* + - * 1. 基于CRF++的实体识别

CRF++利用了条件随机场模型(CRF)进行命名实体识别，它是当前最流行的实体识别工具之一。CRF++是可用于连续数据标注的简单、可定制并且开源的CRF工具，并且在windows系统下有可执行的工具包。

CRF模型需要从大量的标注文本中学习命名实体的潜在上下文关系和语义环境，考虑到人工标注的工作量较大，本课题首先通过网络搜集现有的词库(例如搜狗词库)，采用字典匹配的方法来自动进行标注。完成CRF模型训练后，再对训练语料识别新的菜品或商家实体，并补充到实体词典中，不断重复上述过程，直到无法发现新的实体时停止。流程图如下:

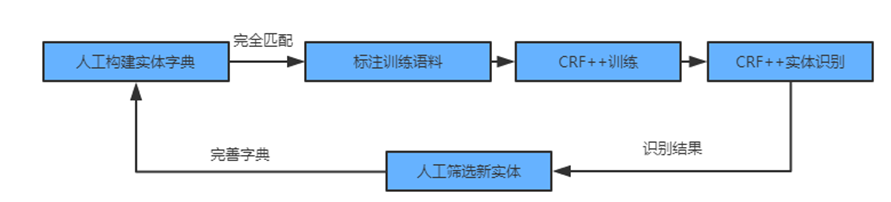


图 菜品实体识别流程图

* + - * 1. 基于多示例和注意力交换的实体关系识别

经过上述的菜品命名实体识别后，可以得到菜品实体词典，通过词典的匹配，能够识别出微博或评论文本中的菜品名称，接下来我们需要判断菜品和商家是否为推荐关系。该问题转化为菜品和商家实体关系识别问题，本课题拟提出基于多示例和注意力交换的深度学习模型来完成实体关系识别。

在关系识别领域，面临的主要问题是人工标注语料的工作量较大，而当前的解决方案是采用远程监督方法，利用现有的开放知识库进行语料的自动标注。在本课题中，由于目前没有针对菜品推荐的知识库，因此人工标注工作难以避免，但我们仍可以采用远程监督方法，可以极大地减少人工标注量，提高工作效率。

由于远程监督有一个强假设条件，如果两个实体对存在某种关系，那么任何包含这两个实体的句子都存在这种关系。实际上这个假设并不成立，因此会引入较多的噪声数据。为了缓解噪声数据带来的负面影响，目前最好的方法采用了多示例和注意力结合的深度学习模型。

多示例的核心思想是将每个实体对视作一个包，所有包含该实体对的句子视为包内的示例，多示例方法是基于包级别而不是示例级别进行分类，可以有效减弱引入噪声句子而带来的负面影响。注意力机制的核心思想则是通过训练学习到包内不同示例的权重，权重是基于示例向量和关系向量之间的相似度计算出来的，如果示例对于关系的贡献较大，那么权重就较高。根据示例的权重，对包内示例进行加权求和得到包的特征表示向量，最终基于包向量进行关系分类。基于注意力的包特征向量构建模型如下所示:

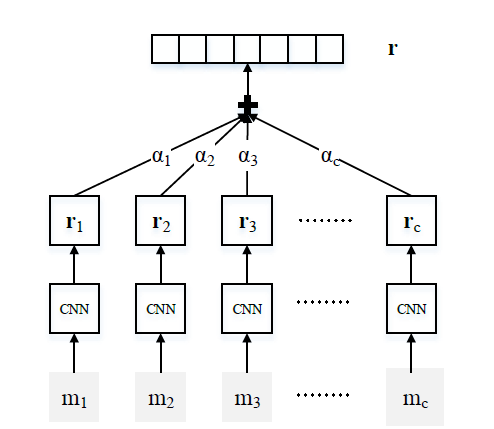


图 基于注意力的包向量构建模型

其中mi表示包含某实体对的文本示例，ri表示经过卷积神经网络CNN后的文本向量，ai表示文本示例的权重，通过加权求和后即可得到r向量，即该实体对的包向量。其中，ai的计算公式如下：





q又称为查询向量,可以看做对关系的向量表示，该向量由模型自动学习得到。最终将包向量输入到softmax层进行分类，目标函数为交叉熵:



上述模型虽然取得了目前最好的效果，但是仍然有改进空间。本课题基于注意力来交换包间的示例，从而完成降噪并提高分类效果。具体方法如下:

1. 每次训练完模型之后，计算每个包内的每个示例在不同关系下的权重；
2. 执行示例的转移操作。将示例转移到权重最大的关系下形成一个新的包，即对示例的关系标签进行修改。示例的转移相当于对包进行了切割操作，将原来的包切割成若干个小包，每个小包的都属于不同的关系，最终得到了新的训练集；
3. 将新的训练集加入到模型中训练，重复上述过程，直到满足指定迭代次数为止；

经过上述方法的改进，可以识别出包中的噪声示例，通过修改噪声示例的标签，最终达到降低噪声的目标。

* + - * 1. 实体关系识别在菜品标签提取场景上的应用

上一节阐述了本课题针对当前实体关系识别研究的不足之处进行改进的方法，本小节将探究将实体关系识别应用到菜品标签提取场景上的应用。

考虑到评论文本和微博数据的不同特点，我们对两种数据源采用不同的处理方式。对于评论文本，通常文本内容中不会提到商家名称，因为评论都是针对具体的商家进行的，此时的商家实体是隐式的，因此在提取(菜品，商家)实体对时，默认将商家名称添加到文本的末尾，评论所属的商家名称在评论网站上可以获得。对于微博文本，则通常会在内容中显示指出评论的商家名称，此时直接使用命名实体识别方法即可获取(菜品，商家)实体对。

在菜品标签提取场景中，每一对(菜品，商家)实体对都构成了一个包，此时包中的示例就是所有包含该实体对的评论文本或微博文本，而菜品和商家的关系分为三种，推荐，不推荐或者没有关系。此时，通过本课题提出的基于多示例和注意力交换的实体识别方法即可识别出菜品和商家之间的关系，得到每个商家的推荐菜品。具体流程如下:

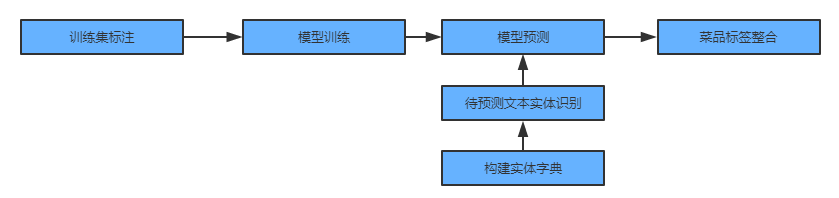


图 菜品标签提取流程图

1. 采用基于CRF++的实体识别方法，构建菜品实体字典和商家名称字典；
2. 通过远程监督的方法标注训练集，首先人工标注部分的(菜品，商家)实体对的推荐关系，然后根据远程监督的思想，所有包含这些实体对的文本全部自动进行标注；
3. 采用基于多示例和注意力交换的方法完成模型的训练。远程监督会引入较多噪声，通过本章提出的方法，可以降低噪声的影响；
4. 通过实体字典匹配的方式，识别新到来的文本中的菜品实体和商家实体，并构建实体对；
5. 在预测阶段，每条文本都视为一个独立的包，并预测(菜品，商家)实体对的关系；
6. 对商家的推荐菜品进行整合，提取出最终的菜品类标签。

通过上述流程，最后完成了实体关系识别方法在菜品标签提取场景上的应用，减少了人工标注的工作量，并提高了标签的准确率。

### 基于多维数据的客流预测模型构建

* + - 1. 针对多源数据情况下的预测问题

目前对时间序列数据的预测方法有很多，传统方法包括：卡尔曼滤波分析法，ARIMA方法，回归方法，灰色模型法，时间序列分析法等等。这些方法大多都只是针对单一时间序列数据的分析，预测某单一时间序列。除了这些传统方法，神经网络模型在时间序列预测上也有较好的效果，比如循环神经网络，由于其独特的结构设计，模型对序列数据的处理表现较好，对复杂的时间关系有较强的学习能力。但是这些模型也只是适用于单一的时间序列数据，对多源的数据支持度不高。这些模型在预测时，只考虑在时间维度上的历史数据对预测值的影响。但是实际生产生活中的时间序列数据往往不只与历史数据有关，一些其他的序列数据或者其他的特征数据都会对研究目标有影响。

在客流数据中，影响客流的不仅仅有历史客流数据，还包括如气象、节日、位置、周边商家等等因素，这些因素对于城市客流有着很大的影响。因此对客流进行预测时，仅仅考虑其历史数据往往会忽视很多有用信息。因此，如何找出这些有用的相关特征，并有效的利用这些特征，使之能融合到客流模型的预测中是研究的重点。方案的基本流程如图所示：



图 客流预测方案流程

* + - 1. 数据源获取

首先，搜集各个来源的数据源，包括原始客流量数据，气象数据，地铁站点数据，全年节假日数据，以及站点周围的商家数据。其中，客流量数据信息如下表：

|  |  |
| --- | --- |
| 字段名称 | 字段描述 |
| tradeId | 订单id，区分不同订单的标志 |
| machineId | 乘客取票机器id，区分不同的出票机器 |
| ticketPrice | 乘客购买的单程票的总价格 |
| ticketCount | 本次订单的购票数量 |
| ticketTime | 乘客取票的时间，若为空则乘客未取票 |
| stationStart | 单程票记录的起点站 |
| stationEnd | 单程票记录的终点站 |

由于是原始数据，数据中包含了一些用于测试的样本，以及一些不能和客流对应的订单数据。本文根据字段的定义，对数据进行一定的筛选，剔除掉不合理的订单或者是无效的订单，合理的订单数据需要满足如下要求：

(1)起点站和终点站不能相同；

(2)记录中必须包含取票时间。因为需要将订单数据转化为进站的客流数据，而未取票的订单不能确定乘客的取票时间，其中包括未付款订单，取消的订单，未取票的订单；

(3)去掉取票时间为凌晨的记录。因为取票机器大多在地铁站内部，取票时必须保证地铁站是在运营状态，这部分数据为测试数据；

(4)订单的购票金额要合理。通常地铁采用路程计价方式，数据中最低票价为2元。

预处理之后是将原始的订单数据转化成客流数据。客流指的是一段时间内进入站点的乘客数量，使用两种时间段划分方式：以天为单位的划分和以1小时为单位的划分。前者是进行短期客流预测的训练数据和测试数据，后者是为了发现不同站点之间的相似性，缩小时间粒度以进行更精确的计算。

考虑到当天气情况变差时，道路变得拥堵，这时大部分乘客会选择地铁出行，这对于地铁的客流量会造成不小的影响，也说明气象因素会对客流有影响。对于气象信息，搜集了包括天气类型，气温，风速，湿度等气象数据。

地铁站点作为乘客上下班通勤，出行娱乐的主要交通工具，具有很强的空间特点，有学者研究发现地铁乘客的出行目的大多以上下班、上学、公务办事、私人购物等为主。因此某一个站点的客流会和其所在的位置，站点与站点之间的位置关系，站点周边的商业的特点等因素有一定关系。比如部分地铁站点，由于其地处热门商区或者住宅区，本身的客流量可能比其他站点的多一些。这些由于站点本身的属性决定的特征属于站点特征。

同时，某一时间窗内的客流特点与该时间点的含义有一定关系。例如在火车客流分析中，节假日的出行客流较平时都会有大幅度增加，这时的时间点本身便带有了节假日的属性。例如有学者对2005年南昌客运站春运期间的客流进行了调研，发现客流在正月初二到正月初十期间是客流高峰期。因为正月初二到初十不仅是春节放假期间，同时也是探亲，旅游的高峰期，所以这段时间被赋予了传统节假日的含义。因此客流和时间点本身的含义有一定的关系。统计全年内所有节假日的信息作为一项数据源。

除此以外，如果一个地铁站点处于商区附近且周边的商家好评较多，那么这个站点的客流量也会有一定的增加。因此，站点周边商家的评价信息也对客流量有一定的影响。根据地理位置搜集站点周边的商家在大众点评网站上的评价信息作为站点商家信息。

* + - 1. 特征工程

以上数据源作为客流预测的基本数据源。第二步便是对搜集的数据源进行预处理和特征选择等特征工程的处理。对于特征选择来说通常从两个方面考虑来选择特征：

(1)特征是否发散：如果一个特征不发散，例如方差接近于0，也就是说样本在这个特征上基本上没有差异，这个特征对于样本的区分并没有什么用。

(2)特征与目标的相关性：与目标相关性越高的特征，说明其重要性越大，对结果的影响越大，在特征选择时对其优先选择。

使用上述方法对部分特征进行筛选，并将最终的特征分为时间维度特征和非时间维度特征。时间维度特征包括客流量数据，气象数据，节假日信息等，非时间维度特征包括站点信息，周边商家评价等。对特征在时间维度上进行分类有助于下面模型的学习。

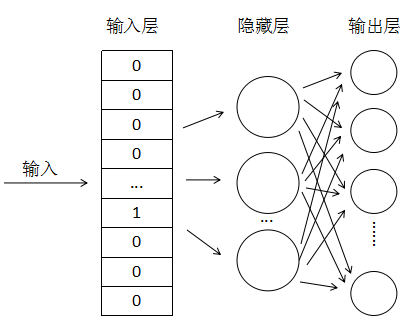
* + - 1. 建模分析

第三步，便是对客流进行建模预测。拟采用两个步骤进行建模，一是对客流数据进行向量化表示，二是针对多源数据，同时利用向量化的客流数据设计模型进行预测。在第二步中提取的非时间维度特征不会随着时间的变化而变化，对于时间序列数据看似没有影响。但是对于一个城市内的所有地铁线路，每个站点的空间特征是不同的，因此对于不同的站点，客流数据的变化也是不一样的。这些非时间维度特征属于每个站点独特的属性，标志着每个站点的特有的特征。将这些空间特征编码进客流数据中可增加“客流量”这一数据的信息量。将原本属于数值属性的数据融入了一些相关的类别特征。

在将空间特征编码进目标数据中时，利用词向量表示的思路。在自然语言处理中，一个基本问题是如何计算一段文本序列出现的概率，语言模型给出了基本解决框架。对于一段文本序列，它的概率表示为



即将序列的联合概率转化为一系列条件概率的乘积。考虑到巨大的参数空间，一般采用其简化版Ngram模型。常见的词向量word2vec是google在2013年推出的计算词向量的一种方式。它可以将one-hot转换为低维度的连续值，也就是稠密向量。word2vec模型其实就是简单化的神经网络。

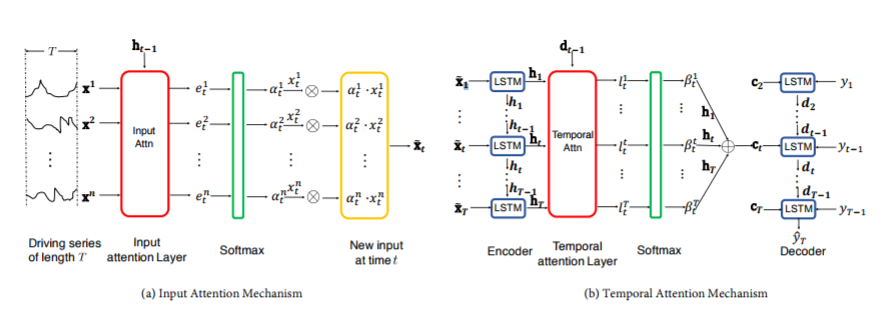


简化的模型包含三层结构，输入层，隐藏层，输出层。输入是one-hot数据，隐藏层没有激活单元，也就是线性单元，输出层维度和输入层一样。获取的词向量其实就是隐藏层的输出单元，或者说是输入层和隐藏层之间的权重。词向量的训练中，根据中心词预测上下文内容，利用隐藏层的单元表示为词向量。这样的思路同样可以用在客流数据的向量化表示中。

将与之相关的空间特征拼接在客流量这一数值后，训练任务同样为预测任务，即根据历史客流量和相关空间特征对下一时刻的客流进行预测。同样使用简单的神经网络模型进行训练。取隐藏层的单元表示作为客流量的向量化表示。这是相当于利用神经网络学习到了客流量和空间特征的“高级”表示。这样的表示对于后续模型的预测起到了积极地影响。

接下来，便是预测模型的设计。利用上述的方式将城市内所有的站点进行向量化表示，得到了各个站点新的时间序列数据，同时加上第二步中提取的时间维度特征，这些数据组成了模型使用的多源数据。普通的循环神经网络模型对于单一序列的学习能力较强，对于多源数据支持度不高。针对此问题，引入了注意力机制。注意力机制最早用在图像邻域，受到人类注意力机制的启发。人类观察图像的时候会将注意力集中在重要的部分。因此借鉴这样的思路，对于图像的学习不会将每个像素点同等对待，而是对于重要的部分提升更高的注意力。同样的思路可以引入到自然语言处理的领域，比如机器翻译，对于源句子，不是把每个单词都同等对待，而是翻译不同部分时对源句子的不同部分有不同的权重。

利用注意力机制的思路，设计如图所示的模型结构。



在输入阶段第一次引入注意力机制，自适应地学习输入之间的相互关系，学习站点之间客流数据之间的关系，可以对整个城市地铁全网的客流学习整体的相互关系。例如，如果两个站点之间存在依存关系，那么一个站点的客流变多，那么一段时间后另一个站点的客流也会变多。这里在模型的第一阶段是可以学习到这样的变化关系的。注意力机制可以关注到对目标站点影响较大的其余站点。在第二阶段，对于时间序列数据的预测，也同样引入注意力机制，将注意力放在整个序列中对目标时间点影响最大的数据上。例如可能上一时刻的数据对于下一时刻的影响较大，或者是前一天的同一时间点的客流对目标时刻客流的影响较大。第二阶段的注意力机制将对这方面的影响进行学习。这样可以利用模型自适应地学习多源数据之间的相互影响，自动找出站点与站点之间的关系，站点与气象，节日之间的关系，并以此预测出目标客流。

本课题的客流预测模块按照图6的流程，对数据源进行收集处理，分析对客流有影响的因素并形成结构化的特征；然后将客流量数据结合特征进行向量化编码；最后设计针对多源数据的神经网络模型对客流进行有效预测。

### 基于商家标签和客流的推荐模型构建

* + - 1. 数据建模方法

本课题中的数据源包括用户历史行为、商家信息、站点画像、商家评论信息和站点客流量信息。

用户历史行为使用一个矩阵表示，由两种实体集合组成：M个用户集合和N个商家集合，中的每一项是用户对商家的评分。原则上可以是任意的实数，但本课题中评分为正数，并且，其中0值代表用户未对此商家进行评分，评分值越高代表用户对商家越满意。根据用户-商家评分矩阵，用户的历史行为信息可以表示成。

用户属性信息即除了用户历史行为之外的其他特征，包括职业、爱好和年龄等基本特征，和从历史行为中统计出来的统计特征，如经常去的商家类型、历史消费平均值等。这些特征中既有离散值，又有连续值。对于离散值，使用one-hot进行编码，将特征值转化成01向量的形式；对于连续值，除了使用原始值，还加入等子特征，加强特征的非线性表示能力；一部分动态的连续值可以进行离散化，加强模型的稳定性，如对年龄进行离散化，将年龄分成20~30等区间，用户年龄增长1岁并不会对模型造成影响。

商家信息包括商家的类型、商家标签、人均消费价格、商家所处的站点等基本特征，除此之外，还包括大量的用户对于商家的评论文本。这些评论之中包含了大量的有用信息，如果能有效地加以利用，可以大量提升模型的效果。本研究点利用循环神经网络对评论文本进行建模，挖掘出评论中对推荐有用的隐含的信息。

站点客流量对用户的出行体验有重要的影响，假设有S个站点，站点集合可以表示成，代表站点k客流量的具体值。推荐模型计算出用户去往站点的可能性为，模型根据站点客流量对进行调整,调整之后的得分为：

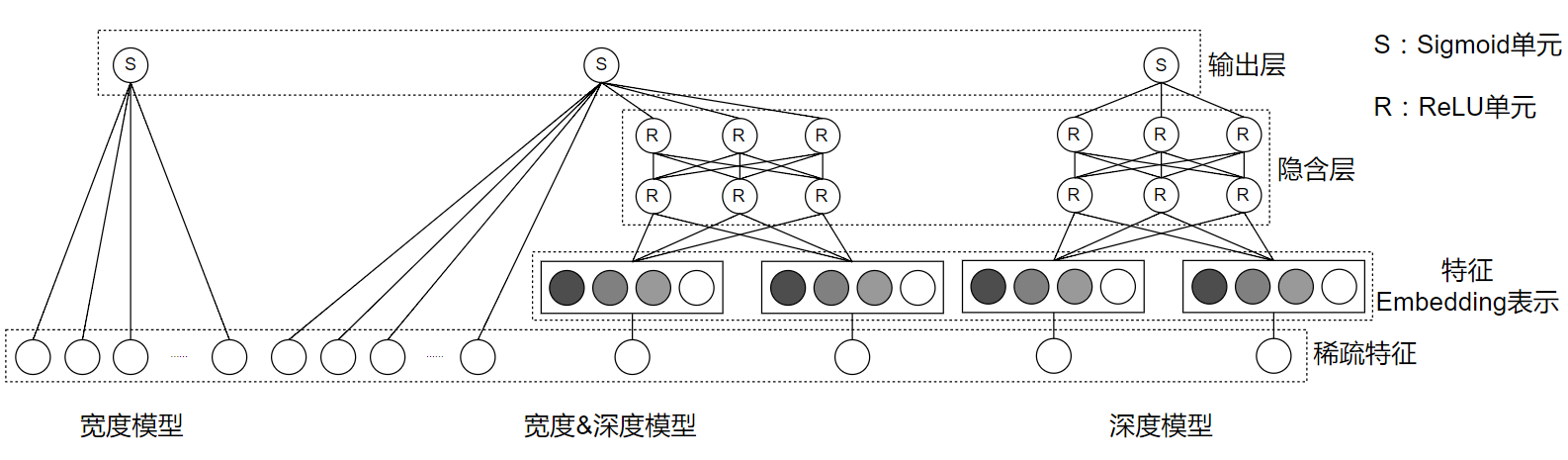
将最为用户去往站点的可能性的最终得分。

* + - 1. 站点画像构建

本研究点从以下几个维度构建站点画像。

1. 站点地理位置和拓扑关系。首先统计各个站点的地理位置信息，包括站点是否位于市中心、站点与市中心的距离等。站点是否位于市中心采用0-1编码，站点与市中心的距离使用欧式距离计算。站点之间的拓扑关系包括站点是否是起终点，是否是换乘车站等，站点拓扑关系使用0-1编码。
2. 站点客流量信息。利用统计分析的方法统计出全天各个时段站点的平均客流量、站点的客流高峰时段、工作日与周末、节假日客流关系等。
3. 站点周围商家信息。利用商家的标签信息，统计站点周围各类型商家的数量。
   * + 1. 路线推荐模型设计
          1. 模型训练

模型训练的主要功能为基于训练数据集训练得到用于预测用户去往站点概率的算法模型。本研究点输入数据种类多、数据量大，并且每种数据有不同的表现方式，综合考虑多种数据的特征，使用深度&宽度模型（Wide & Deep Model），其结构如下图所示：



模型的深度部分对输入变量进行深层次的归纳，提取出里面包含的深层次语义信息；而宽度部分相当于一个线性模型，对输入变量的各分量进行加权求和。

用户的历史行为信息，具有维度高和稀疏性的特点，因此使用深度神经网络进行深层次的编码，提取出深层语义信息。而对于用户和商家的属性信息，使用一层神经网络进行处理，以保留其原始特征所蕴含的语义信息。使用循环神经网络对商家的评论文本信息进行挖掘，提取出里面的有用信息。本课题设计的模型结构如下图所示：



常用的损失函数包括交叉熵损失函数、绝对值损失函数和平方查损失函数等。模型根据用户去往各个站点的概率，对站点进行排序，因此使用交叉熵损失函数：

其中，n表示训练集样本数，表示用户真实概率，表示模型预测概率。

排序模型训练方法按照原理的不同可以分成三种：PointWise方法、PairWise方法和ListWise方法。在PointWise方法中，排序问题首先被转换成等价的回归问题，再通过训练算法模型拟合打分函数。Pairwise方法不是单独计算某一候选项的排序得分，而是将分析的重点转向判断待排序项之间的顺序关系是否合理。ListWise方法对每一个候选项打分后，利用Luce模型计算候选项集的每一种排列情况的概率值，之后再利用余弦相似度或KL距离等作为损失函数来衡量模型输出序列的概率分布与真实序列的排序概率分布之间的差距。由于我们无法得知站点对用户优先度的相对顺序，因此模型采用PointWise的方法进行训练。

* + - * 1. 选择负样本

对于用户每次出行，只能得到用户去了哪个站点，训练数据中只有正样本。对此，我们需要在剩余的站点中选择负样本，从而使得模型能够更好的训练。选择负样本的方式有多种：1）余下的所有站点当作负样本，这种方法简单，但是缺点也很明显，就是会导致正负样本严重不平衡。2）从余下的站点中中随机选取一部分作为负样本，这种方法虽然不会导致正负样本不平衡，但是会有一定的随机性。3）计算站点之间的相似度，选择与用户目的站点相似的其他站点作为负样本，这种方法避免了选择的随机性，并且不会导致正负样本失衡，因此，模型选择这种方法选择站点负样本。

模型使用站点的画像数据计算站点之间的相似度，相似度计算方式使用余弦距离：

根据距离计算公式，选择与目标站点相似的前N个站点当作模型的负样本。

* + - * 1. 排序

排序主要功能是使用训练得到的模型对站点进行打分，预测用户去往各个站点的概率，并按照概率的高低返回排序结果。由于每个城市地铁站点数量固定，排序时会对每个站点都进行预测，这样确保不会错过每个用户可能去往的站点。

# 可行性分析

**研究目标明确：**本课题的研究目标来源于Mashup服务中用户隐私信息保护问题的实际需要，定位于应用基础研究，确定了明确的研究目标。

**技术方案可行：**实现本课题的研究目标需要解决一些关键的科学问题，课题组针对Mashup服务中后端服务与Mashup主站的信息交互技术以及Mashup服务中的授权访问方法进行了前期研究，调研并总结了需要解决的关键科学问题，对这些关键科学问题进行了深入的分析并已经有了初步的解决思路，包括初步确定提出基于挖掘方法对后端服务中的信息源进行划分；初步提出了面向Mashup服务中授权访问的隐私信息感知方法；初步验证了本课题所提出的技术方案是可行的。同时课题组对后续研究内容有清楚认识和深刻理解，所提出的研究思路和技术途径明确、具体，针对性强。

**研究基础具备：**申请人及所在的课题组长期从事服务安全与访问控制领域的关键技术研究，发表了多篇SCI论文。在移动互联网、服务计算等新型技术和需求下对服务安全体系结构、访问控制、授权访问和组合服务隐私保护方面已有较为深入的研究，为开展Mashup服务隐私信息感知的授权访问方法的研究创造了条件。同时课题组内人员具有良好的知识交叉结构，为本课题的顺利开展奠定了重要的研究基础。

**实验环境支撑：**课题组所属并依托于北京邮电大学网络与交换技术国家重点实验室，具有良好的科研环境和一流的实验条件。课题组现已拥有多种高性能容错计算机、多个不同厂家的交换机平台、较多的高性能工作站和丰富的软件，自行开发并构造了一整套面向下一代电信网和互联网的安全评测平台（NTISAP）和相关测试仪表，自行研发和建设了下一代互联网业务支撑平台（NSSP），自行研发了融合网络业务支撑环境（ISSE），并拥有一批自行开发的网络安全模拟测试工具。

综上所述，本课题无论在关键理论和技术研究方面，还是在系统实验仿真方面，研究隐私感知的Mashup服务授权访问机制都具有可行性。

4．**本项目的特色与创新之处；**

**结合移动互联网的文本数据进行客流预测和路线推荐是本课题的特色之处。**传统的轨道交通规划数据来源单一，尤其是移动互联网数据和轨道交通结合的研究成果欠缺，而本课题则基于多维数据进行挖掘，特别是利用商家评论和微博文本，来挖掘出行相关的商家特征，并融合到客流预测和路线推荐中。课题的研究工作涉及文本挖掘、客流预测、多因素融合的推荐方法，同时研究上述理论问题在轨道交通出行中的应用方案。同时本课题在如下三个方面具有创新性：

1. **商家标签优化**

本课题将商家标签分类服务类标签和菜品类标签，针对服务类标签提取，将其转化为多标签分类问题，而对菜品类标签提取，则将其视为实体关系识别问题。

优化文本多标签分类的关键在于标签关联挖掘，本课题在传统分类模型中融入了LDA语义分析模块，并将其命名为语义嵌入模型。模型通过对标签的语义解析来挖掘标签之间的关联性，本文设计模型的整体结构并测试了模型在不同文本表示方法、基分类器以及优化策略下的运行效果，结合理论研究和实验结果实现了语义嵌入的多标签分类模型。

本课题针对多标签分类的创新性体现在以下四点：

1. 文本关联挖掘的难点在于文本建模，传统的方法通常使用文本词向量进行建模，容易出现维度爆炸和数据稀疏等问题，而本文提出的标签化文本建模方法，既解决了词向量存在的问题，又能够最大限度的表达文本特征。
2. 传统多标签分类算法在考虑标签关联性的时候大多与模型的学习过程相分离，只考虑标签与标签之间的相关性，忽略了样本的先验知识对于标签关联性挖掘的重要作用。而本文提出了基于语义识别的标签关联分析方法，通过语义元素解释标签之间的关联性并创造性的将标签间的关联分析（语义分析）融合于模型的最优化过程中从而提升算法的效果。
3. 语义模型的参数优化直接影响了整个多标签分类模型的效率。本文利用语义模型最优参数与分类模型最优参数之间的统计关系巧妙避开了大范围线性搜索带来的时间消耗，成功的将算法复杂度降低到了一个可接受的范围。
4. 利用了大量的微博数据作为输入之一，扩充的数据源，提高了模型的容错性。

针对实体关系识别的创新点体现在以下方面:

1. 结合多示例方法，基于注意力权重进行包间示例的交换，从而降低噪声，提高远程监督的效果；
2. 首次将实体关系识别方法应用于标签提取场景中，通过识别菜品和商家之间的关系来抽取菜品标签。
3. **基于多维数据的客流预测模型**

一般时间序列预测问题大多是针对单一维度的时间序列数据。数据在时间维度上的数值表示的就是当前时间点上的实际数值。结合自然语言处理中的词向量的思路，将模型参数引入到词向量的表示中。设计合理的编码器模型，可以将影响客流量的非时间维度特征融入到客流量的向量化表示中，能增加“客流量”这一数值数据的信息量，训练模型，并将模型的参数作为客流数据的向量化表示。这样的向量不仅包含了客流量的信息，还融合了其他一些和站点，商家等有关的信息。这些信息的加入对后续的预测提供有力支持。

同时，传统时间序列分析模型大多是针对单一序列的分析。而且只是利用该序列的历史数据预测未来值。但在轨道交通场景中，影响客流量的不仅仅只有其历史数据，有很多比如气象、节假日、站点位置、周边商家信息等。在客流预测中，将这些因素考虑在内可以提高预测的准确率。本项目针对多源数据设计的网络模型，在输入时接收不同的序列数据，自适应的选取合适的序列，在预测目标序列时，综合考虑各个序列对其的影响。在模型设计时，在输入部分引入注意力机制，针对多源的序列数据进行加权选取，学习站点之间客流数据之间的关系，可以对整个城市地铁全网的客流学习整体的相互关系。

1. **基于商家标签和客流的推荐模型**

。

5．**年度研究计划及预期研究结果**（包括拟组织的重要学术交流活动、国际合作与交流计划等）。

## 年度研究计划

|  |  |
| --- | --- |
| **2017.1-2017.12** | 收集资料和调研；  研究包含隐私信息的Mashup服务授权访问的建模，对该控制访问模型中的角色进行分析，明确角色之间的关联关系，角色之间的交互动作，交互中的隐私信息流及隐私信息内容定义；研究并建立完整模型和隐私信息流图。基于模型中的隐私信息流及定义实现隐私信息划分方法的确定。 |
| **2018.1-2018.12** | 基于确定的划分方法，提出隐私划分方法的实验方案并进行实验，在范例数据集上进行方法测试，基于测试和实验结果继续优化方法。同步进行隐私信息挖掘、封装方案设计并对范例数据进行封装。组织学术交流活动，文章和专利撰写及提交。 |
| **2019.1-2019.12** | 设计和实现隐私感知方法，包括方法的流程、步骤、核心感知算法，设计实验进行验证。设计和实现基于迁移图的控制机制，确定原型系统的搭建方案。参加国际会议，文章和专利撰写及提交。 |
| **2020.1-2020.12** | 搭建原型系统，设计范例，仿真验证完整控制机制。修改完善机制并撰写论文和专利，课题结题准备，演示系统搭建和结题资料撰写。 |

## 预期研究结果

对隐私感知的Mashup服务授权访问进行建模，建构后端服务中非结构与半结构化信息的划分和封装方法，实现后端服务隐私数据的细粒度的划分和封装；提出Mashup服务授权访问中后端服务隐私信息感知方法；设计和实现基于迁移图控制全生命周期的授权单步控制，完成研究报告。

1．反映研究成果的高水平论文15篇以上，其中被SCI收录5篇以上；

2．申请发明专利5个；

3．培养3名以上博士、5名以上硕士。

**（二）研究基础与工作条件**

1．**研究基础**（与本项目相关的研究工作积累和已取得的研究工作成绩）；

## 已取得的研究工作成绩

近几年来，课题组长一直从事服务计算、服务安全的研究工作，作为项目负责人主持1项242安全计划项目、多项企业合作项目，作为主研人参加了国家973、863、国家自然科学基金、重大专项、242等多个国家项目的科研工作。所承担的863计划项目《大数据分析技术在输变电设备状态评估中的研究及应用》、《下一代网络中关于业务能力开放的安全机制研究》、《面向下一代电信网的安全测试评估技术及工具》、国家自然科学基金项目《开放式综合业务体系结构的研究》对融合网络中的**服务计算框架、服务安全、服务隐私保护、行业大数据分析及应用**进行了深入研究，包括组合服务的安全框架、服务的授权访问模型、服务授权策略检测方法、组合服务漏洞检查方法、服务组合过程中的信息泄露。所承担的242信息安全计划项目《电信网安全防御研究》、《电信网应用业务系统安全检测平台》、IBM产学研计划项目《Mobile Internet中服务安全问题研究》提出了互联网服务的隐私保护流程、服务授权中的隐私策略并研发了服务应用认证原型系统和工具。

本课题申请人在Mashup服务安全研究工作方面的进展顺利，已经发表与本课题研究工作密切相关的第一作者SCI论文3篇，探讨了在移动互联网、云计算环境下的组合服务安全体系结构和访问控制模型，细化了该模型并针对该模型进行了原型系统的设计和实现。

在研究成果中申请人提出了基于补偿机制的组合服务访问控制模型，该模型在组合服务访问控制中属性缺失情况下能够提高认证效率、提升认证服务质量改善用户交互友好性和提高系统处理效率。具体工作包括提出了基于属性可信度补偿的访问控制模型（TCBAC模型）；提出了基于个体属性可信度和相似群体属性可信度进行聚合的可信度补偿聚合方法；提出了基于绝对时间衰减因子和同类型操作相对时间衰减因子的属性可信度计算方法。通过实验对比了采用TCBAC模型后，组合服务可到达率的提升效果，系统遭受恶意用户攻击的概率以及对于系统资源消耗的变化情况。本文还展示了引入TCBAC模型后的SOA交互流程。这部分工作对本课题中隐私感知的Mashup服务授权访问机制的研究工作具有借鉴作用。

基于上述研究工作，课题组继续深入研究，针对流行的Mashup服务授权访问机制进行研究，图7描述了该授权访问机制的示例。

Mashup授权访问主要流程如下：

1）将原始服务所提供的信息源进行封装成xml文件类型(SEF.xml)提供给Mashup站点进行访问授权, 解决其信息异构问题。

2）将原始服务所提供的SEF.xml与已有隐私策略进行形式化融合，输出形式化授权文件，作为Mashup与多后端服务间信息交互时的向导授权文件。



图7 面向Mashup应用的隐私保护的授权访问方法示例

3）为每一份授权文件建立生命周期状态机，提高Mashup应用服务中信息授权效率，不同授权服务将会在整个Mashup使用过程中以不同的状态存在，减少用户隐私在Mashup服务应用中的暴露时间，细粒度的保护用户隐私不受侵犯。

**课题组近期发表的与课题相关的论著：**

I．期刊论文：所有作者（通讯作者以“\*”标出），论文标题，期刊名称，卷(期), pp起始页码，发表年份

[1] \***Yan Dan-Feng**，Tian Yuan，Yang Fang-Chun. Privacy-preserving authorization method for mashups. Security and Communication Networks, Volume 8 Issue 18, 2015, 12, p 4421-4435. SCI 检索源期刊

[2] \***Yan Danfeng**, Sun Jing, Zhang Liying, Yang Fangchun. Trust-Compensation-Based Access Control Model for Web Services. CHINA COMMUNICATIONS,2012,9(12):8-21.SCI IDS号:064PW,SCI收录号: :000313088900002

[3] \***Yan Danfeng**, Yang Fangchun, Yeap Tet. Service Security Architecture and Access Control Model for Cloud Computing. CHINA COMMUNICATIONS,2011,8(6):44-50.SCI IDS号:839WD,SCI收录号: 000296399300006

[4] \***Yan Dan-Feng**，\*Tian Yuan，Huang Jun-Lin，Yang Fang-Chun. Privacy-aware RBAC model for web services composition. Journal of China Universities of Posts and Telecommunications, v 20, n SUPPL. 1, p 30-34, August 2013. EI检索号: 20134216863876

[5] \***Yan Dan-feng**, Yang Fang-chun. Design and performance evaluation of management service logic execution environment in NGN. China University of Posts and Telecommunications,2007.10, p 105-110.EI检索号:080111007387

[6] \***Yan DanFeng**，Yang FangChun. Research of universal service management model in NGN. Tien Tzu Hsueh Pao/Acta Electronica Sinica, v 35, n 10, p 1828-1832, October 2007. EI检索号:074810951677

[7] \***Yan Danfeng**, Yang Fangchun. Service Management Server in NGN. Computer Engineering, 33(15): 118-120, 2007.

[8] \***Yan Dan-Feng**, Yang Fang-Chun. NGN management services modeling with multi-view. Beijing YoudianDaxueXuebao/Journal of Beijing University of Posts and Telecommunications, v 29, n SUPPL., p 25-28, May 2006. EI检索号：06269966280

II．会议论文：所有作者（通讯作者以“\*”标出），论文标题，会议名称，会议时间，pp起始页码，会议地址，发表年份，说明

[1] \***Yan Danfeng**, Huang Junlin, Tian Yuan, Zhao Yao, Yang Fangchun. Policy Conflict Detection in Composite Web services with RBAC. Proceedings of 2014 21th IEEE International Conference on Web Services, IEEE ICWS 2014.

[2] \***Yan Danfeng**, Tian Yuan. Privacy policy composition of privacy-aware RBAC model for composite WEB services. Proceedings of 2013 5th IEEE International Conference on Broadband Network and Multimedia Technology, IEEE IC-BNMT 2013, p 312-316, 2013. EI检索号: 20142717891513

[3] \***Yan Danfeng**, Feng Rui, Huang Junlin, Yang Fangchun. Host security event track for complex network environments based on the analysis of log. Proceedings - 2012 IEEE 2nd International Conference on Cloud Computing and Intelligence Systems, IEEE CCIS 2012, v 2, p 807-811, November 13, 2013. EI检索号: 20135117115190

[4] \***Danfeng Yan**, Haili Wu, Junlin Huang, Fangchun Yang. Research on Testing and Detection for Clouding Applications. Proceedings 2012 IEEE 2nd International Conference on Cloud Computing and Intelligences Systems. Hangzhou, China. Oct.2012 (EI检索源)

[5] \***Danfeng Yan**, Wen Hu, Xiang Yang, Fangchun Yang. Private Data Analysis In the Cloud Computing Environment. Proceedings 2012 IEEE 2nd International Conference on Cloud Computing and Intelligences Systems. Hangzhou, China. October,2012.(EI检索源)

[6] \***Yan Danfeng**, Song Dawei, Luo Lin, Yang Fangchun. A fault detecting algorithm based on spatio-temporal correlation in wireless sensor network. Proceedings - 2011 4th IEEE International Conference on Broadband Network and Multimedia Technology, IC-BNMT 2011, p 162-167, 2011. EI检索号: 20121214869255 (该论文被评为大会优秀论文)

[7] \***Yan Danfeng**, Wang Juan, Yang Fangchun. The research on security framework for Mobile Internet Services. Proceedings - 2011 4th IEEE International Conference on Broadband Network and Multimedia Technology, IC-BNMT 2011, p 75-79, 2011. EI检索号: 20121214869237

[8] \*Huang Rui, \***Yan Danfeng**, Yang Fangchun. Research of security metric architecture for next generation network. Proceedings of 2009 IEEE International Conference on Network Infrastructure and Digital Content, IEEE IC-NIDC2009, p 207-212, 2009. EI检索号: 20101012756318

[9] \*Rui Liu, \***Danfeng Yan**, Fan Lin, Fangchun Yang. Optimization of hierarchical vulnerability assessment method. Proceedings of 2009 2nd IEEE International Conference on Broadband Network and Multimedia Technology, IEEE IC-BNMT2009, p 458-462, 2009. EI检索号: 20100412656519

[10] Qibo Sun, \*Shangguang Wang, **Danfeng Yan**, Fangchun Yang. ARM-CPD: Detecting SYN flooding attack by traffic prediction. Proceedings of 2009 2nd IEEE International Conference on Broadband Network and Multimedia Technology, IEEE IC-BNMT2009, p 443-447, 2009. EI检索号: 20100412656516

[11] \***Danfeng Yan**, Fangchun Yang, Yu Lu. Service-based quantitative calculation of risk for NGN. Proceedings of 2009 2nd IEEE International Conference on Broadband Network and Multimedia Technology, IEEE IC-BNMT2009, p 306-310, 2009. EI检索号: 20100412656485

[12] \*Chao Deng, \***Danfeng Yan**, Yun Yang, Fangchun Yang. A domain-oriented distributed vulnerability scanning mechanism. Proceedings of 2009 2nd IEEE International Conference on Broadband Network and Multimedia Technology, IEEE IC-BNMT2009, p 831-836, 2009. EI检索号: 20100412656376

[13] \***Yan Danfeng**, Yang Fangchun. Vulnerability analysis of service architecture in NGN. 2009 International Conference on E-Business and Information System Security, EBISS 2009. EI检索号: 20094212377225

[14] \*Qian Shuang, \***Yan Danfeng**, Yang Fangchun. The Research on Interface Sensibility of Vulnerabilities in IP Multimedia Subsystem. IET Conference Publications, n 544 CP, p 283-287, 2008, China-Ireland International Conference on Information and Communications Technologies, CIICT 2008. EI检索号: 20092912195924

[15] \***Yan Danfeng**, Yang Fangchun; Wang Wenbin.Research and implementation of security evaluation system for NGN. IET Conference Publications, n 544 CP, p 329-333, 2008, China-Ireland International Conference on Information and Communications Technologies, CIICT 2008, Telecommunications Science, 2006. EI检索号:20092912188595

[16] \*Wang Wenbin, Sun Qibo, **Yan Danfeng**. Research on privacy-preserving problem for data-sharing in IDS. IET Conference Publications, n 544 CP, p 320-323, 2008, China-Ireland International Conference on Information and Communications Technologies, CIICT 2008. EI检索号: 20092912188593

[17] \*Li Qin, \***Yan Danfeng**, Yang Fangchun. Research of a hybrid distributed network intrusion detection system. IET Conference Publications, n 544 CP, p 301-305, 2008, China-Ireland International Conference on Information and Communications Technologies, CIICT 2008. EI检索号: 20092912188589

III．专利：发明人，专利名称，授权时间，授权国别，专利号

[1] \***闫丹凤**/孙其博/杨放春/李静林/王尚广，一种SIP消息洪泛攻击的检测方法，2012.2，中国，ZL200910076311.5

[2] 孙其博/\***闫丹凤**/杨放春/龙湘明/王尚广, 基于CUSUM算法检测SIP消息洪泛攻击的装置和方法,2011年5月，中国，ZL200910076310.0

[3] \*赵耀/邹华/杨放春/李晓亮/孙其博/刘志晗/**闫丹凤**/林荣恒，一种支持融合网络业务的云平台及其工作方法，2012.8，中国，ZL201010527644.8

[4] \*邹华/杨放春/赵耀/刘志晗/孙其博/**闫丹凤**/于晓燕/林雪华/起鹏飞; 李静林/龙湘明/刘京，一种会话控制服务方法、装置和系统，2014.04，中国，ZL201010194717.6

[5] \*张冬梅/\***闫丹凤**/王鲁华/舒敏/周春燕/钟金鑫/邓明，基于沙箱技术的客户端蜜罐网页恶意代码检测方法与装置，2013.03，中国，ZL201110025474.8

[6] \***闫丹凤**/宋大伟/胡平/黄俊霖，基于多目标测试用例生成的通行控制算法测试装置与方法，2014.1中国，ZL201210071191.1

[7] \***闫丹凤**/孙敬/王娟/管宁，一种电信应用业务安全测试评估通用平台系统及其方法，（2012.7），中国，201210065960.7

[8] \***闫丹凤**/孙其博/杨放春/王文彬/李沁，一种网络安全态势评估方法及其系统，（2009.05），中国，200810240733.7

[9] \***闫丹凤**/冯瑞/周广，一种基于多主机日志关联的入侵检测方法，（2014.05），中国，201410101730.0

[10] \***闫丹凤**/吴海莉/徐佳，一种社交网用户异常行为的分析方法，（2014.06），中国，201410101728.3

IV．课题组获得的奖励情况

本人前期从事的组合服务授权访问相关研究工作，在课题组完成的“开放融合业务支撑环境”工作中作为代表性成果和重要支撑材料，于2012年和2010年分别获得奖励；本人早期从事的服务控制执行环境的相关研究工作，在课题组获得的若干奖项中作为重要支撑材料。本人获得的奖项如下：

[1]杨放春/苏森/王红熳/张海滨/邹华/双锴/赵耀/孙其博/施智辉/刘志晗/**闫丹凤**/于晓燕/詹舒波/李静林/徐鹏，融合业务支撑环境关键技术与应用，科技部，国家科学技术进步奖，二等奖2012。本人作为主要研究人员承担该获奖中融合业务支撑环境安全框架关键技术研究。完成了服务安全框架的设计和实现，隐私保护的访问控制方法的设计和实现工作。

[2]杨放春/詹舒波/苏森/邹华/赵耀/双锴/王红熳/孙其博/张文涛/刘志晗/**闫丹凤**/李静林/于晓燕，开放式综合业务支撑环境关键技术，教育部，科学技术进步奖，二等奖，2010。本人作为主要研究人员完成了开放服务安全防护要求制定，提出了开放服务安全架构和安全模块的功能规范。

[3] 融合业务支撑环境技术，**闫丹凤**（第11完成人），中国通信学会，科学技术进步奖，一等奖，2010。本人作为主要研究人员完成了服务安全执行环境的设计和实现。

[4] 独立智能外设系统软件，**闫丹凤**（第6完成人），中国通信学会，科学技术进步奖，二等奖，2006。本人作为主要研究人员完成了智能外设业务监控和管理模块的设计和实现。

[5] CIN系列智能网系统，**闫丹凤**（第9完成人），教育部，科学技术进步奖，一等奖，2004。本人作为主要研究人员完成了CIN智能网业务执行环境的设计和实现。

本人所完成的工作在与企业合作项目中得到应用，在中国电信运营商、互联网企业、企事业单位的信息化部门中发挥着重要作用。

2．**工作条件**（包括已具备的实验条件，尚缺少的实验条件和拟解决的途径，包括利用国家实验室、国家重点实验室和部门重点实验室等研究基地的计划与落实情况）；

## 2.1课题组现有的支撑条件

本课题组长期从事服务安全、服务计算、网络安全测试评估等相关项目，课题组目前已经具有较强的科研支撑条件。课题组所属北京邮电大学网络与交换技术国家重点实验室，是国家在网络交换与智能控制技术领域唯一的国家级重点实验室，具有良好的科研环境和一流的实验条件，现已拥有多种高性能容错计算机、高性能服务器、多个不同厂家的交换机平台、较多的高性能工作站和丰富的软件，自行开发并构造了一整套面向下一代电信网和互联网的安全评测平台和相关测试仪表，并拥有一批自行开发的网络安全模拟测试工具。这些支撑条件为本课题的开展提供了较好环境。

## 2.2 依托单位承诺提供的支撑条件

课题组依托单位保证向课题组提供所需要的实验环境（包括实验场地、网络接入、必要的硬件和软件系统），以及支持研究工作的服务支撑平台、组合服务仿真环境、安全测试评估工具的科研平台和环境。保证为课题组提供一切与研发工作相关的便利条件，保证必需的人员配置和工作时间。同时，还将根据课题研究的需要增加人力和物力的投入。本课题依托的北京邮电大学网络与交换技术国家重点实验室十分重视对青年教师的培养工作，在对青年教师对外交流、课题实施要求的必要条件和人员配备提供最高有限级别的支持工作。

3．**正在承担的与本项目相关的科研项目情况**（申请人和项目组主要参与者正在承担的与本项目相关的科研项目情况，包括国家自然科学基金的项目和国家其他科技计划项目，要注明项目的名称和编号、经费来源、起止年月、与本项目的关系及负责的内容等）；

无

4．**完成国家自然科学基金项目情况**（对申请人负责的前一个已结题科学基金项目（项目名称及批准号）完成情况、后续研究进展及与本申请项目的关系加以详细说明。另附该已结题项目研究工作总结摘要（限500字）和相关成果的详细目录）。

无

**（三）其他需要说明的问题**

1. 申请人同年申请不同类型的国家自然科学基金项目情况（列明同年申请的其他项目的项目类型、项目名称信息，并说明与本项目之间的区别与联系）。

无

2. 具有高级专业技术职务（职称）的申请人或者主要参与者是否存在同年申请或者参与申请国家自然科学基金项目的单位不一致的情况；如存在上述情况，列明所涉及人员的姓名，申请或参与申请的其他项目的项目类型、项目名称、单位名称、上述人员在该项目中是申请人还是参与者，并说明单位不一致原因。

无

3. 具有高级专业技术职务（职称）的申请人或者主要参与者是否存在与正在承担的国家自然科学基金项目的单位不一致的情况；如存在上述情况，列明所涉及人员的姓名，正在承担项目的批准号、项目类型、项目名称、单位名称、起止年月，并说明单位不一致原因。

无

4. 其他。

无