学校代码：10285

学　　号：20155227004

蘇州大學

SOOCHOW UNIVERSITY

硕士学位论文

（专业学位）

基于时空上下文共现的用户关系强度预测

Users Relationship Strength Prediction Based on Spatio-temporal Context Co-occurrence

研究生姓名	徐 彩 旭
指导教师姓名	严 建 峰
专 业 名 称	计算机技术
研 究 方 向	大数据研究与应用
所 在 院 部	计算机科学与技术学院
论文提交日期	2018 年 5 月

苏州大学学位论文独创性声明

本人郑重声明：所提交的学位论文是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含其他个人或集体已经发表或撰写过的研究成果，也不含为获得苏州大学或其它教育机构的学位证书而使用过的材料。对本文的研究作出重要贡献的个人和集体，均已在文中以明确方式标明。本人承担本声明的法律责任。

论文作者签名：徐彩旭　　日　　期：2018.6.23

苏州大学学位论文使用授权声明

本人完全了解苏州大学关于收集、保存和使用学位论文的规定，即：学位论文著作权归属苏州大学。本学位论文电子文档的内容和纸质论文的内容相一致。苏州大学有权向国家图书馆、中国社科院文献信息情报中心、中国科学技术信息研究所（含万方数据电子出版社）、中国学术期刊（光盘版）电子杂志社送交本学位论文的复印件和电子文档，允许论文被查阅和借阅，可以采用影印、缩印或其他复制手段保存和汇编学位论文，可以将学位论文的全部或部分内容编入有关数据库进行检索。

涉密论文□

本学位论文属 在_____年___月解密后使用本规定。

非涉密论文☑

论文作者签名：徐彩旭 日 期：2018.6.23

导 师 签 名：尹建峰 日 期：2018.6.27

基于时空上下文共现的用户关系强度预测

摘　　要

目前，基于时空数据的用户关系强度预测成为了众多学者研究的热点。前人研究工作主要集中在上下文感知预测或者时空共现预测，并且没有考虑时间上下文信息。本文提出转化的思想从多视角巧妙地将时空上下文和时空共现进行融合，并进一步提升预测精度。本文所做的工作分为三个方面：第一，基于不同的时空视角，提出多视角时空上下文共现的预测方法；第二，提出了基于视角融合的预测方法；第三，给出用户关系强度预测的应用解决方案。本文使用SNAP开源数据集Brightkite和Gowalla作为实验，在这个数据集上再进一步切分成训练集、验证集和测试集。本文主要贡献包括以下三点：

(1) 本文的提出的多视角上下文共现方法巧妙地将时空数据中用户间的关系转化为自然语言处理领域（**N**atural **L**anguage **P**rocessing）中同义单词的关系，从而巧妙实现时空上下文和时空共现的融合，并且本文方法还考虑时间上下文信息。该方法首先从多视角生成时空上下文序列。然后利用NLP领域中的工具分别基于多视角提取用户上下文共现特征，该特征表征用户的签到时间的共现、空间的共现、时间上下文和空间的上下文信息。最后利用机器学习技术基于多视角分别进行关系强度预测。实验表明，本文多视角方法中最好的Day-Location视角比EBM算法在Brightkite数据集在相同的Precision下Recall要最高提高10%，在Gowalla数据集上最高提高8%。

(2) 根据特征级的特点，本文提出时空上下文共现特征融合的方法（**F**eature **F**usion）。FF方法基于其中两个视角特征所表征信息的互补性，将这两组特征进行融合，再结合机器学习技术进行关系强度的训练与预测。同时，本文也给出了基于多视角决策融合的方法(**D**ecision **F**usion)。实验表明，FF方法相比本文提出的最好的Day-Location视角在Brightkite数据集上AUC指标提升3.6%，在Gowalla数据集上提升4.3%。FF方法相比DF方法在Brightkite数据集上AUC提升1.4%，在Gowalla数据集上提升1.6%。并且FF方法比目前最好的方法SCI在Brightkite数据集上AUC提升6.1%，在Gowalla数据集上提升2.4%。

(3) 本文提出一种社交网络关系强度预测的应用框架。该框架包括以下几个模块：数据的存储与管理模块包含数据结构化、数据的存储和数据可视化这三个子模

块；数据建模模块主要将两种融合方法FF和DF方法模块化；模型评估模块进行预测方法性能的全方位评估，同时给出LIFT曲线、ROC曲线和PR曲线，并输出关系强度最强的用户对。

关键词： 用户关系强度，时空数据，上下文序列，上下文共现特征

作　　者： 徐彩旭

指导教师： 严建峰

Users Relationship Strength Prediction Based on Spatio-temporal Context Co-occurrence

Abstract

Recently, users relationship strength prediction based on spatio-temporal data has become a hot topic for many researchers. Previous researches had mainly focused on context-aware or spatio-temporal co-occurrence, and the time context is seldom considered. This paper proposes a novel transformation idea that artfully merge spatiotemporal context and spatiotemporal co-occurrence together from multiple views, and further improves the prediction accuracy. Our work is mainly from three aspects. First, we propose multi-view relationship strength prediction based on spatiotemporal context co-occurrence. Second, we propose relationship strength prediction method based on view fusion. Third, an application solution for predicting user relationship strength is given. We use the SNAP open source dataset Brightkite and Gowalla as our experiments, and we split the dataset into training set, validation set and testing set. The main contributions of our work are listed as follows:

(1) The multi-view context co-occurrence method is presented in this paper, which artfully transforms the relationship between users in spatiotemporal data into the synonymous word relationship in **N**atural **L**anguage **P**rocessing domain, which artfully realize the fusion between context and co-occurrence. And this method also considers the time context information. Our method firstly generates spatiotemporal context sequences from multiple views, and then use tool in the NLP domain to extract user context co-occurrence feature based on multi-views. The feature represents users' check-in time co-occurrence, space co-occurrence, time context and space context. Finally, we use machine learning techniques to predict the relationship strength based on multiple views. Experiments show that the recall of the best Day-Location view in multi-views is 10% higher than for the Brightkite dataset under the same precision, 8% higher for the Gowalla dataset than the EBM method.

(2) In this paper, we propose feature fusion (**FF**) based on context co-occurrence feature according to the characteristics of feature level. The FF method is based on the com-

plementarity of two view feature, merge the two sets of feature, and then use machine learning techniques to train and predict the strength of the relationship. At the same time, this paper also gives a multi-view **D**ecision **F**usion (**DF**) method. Experiments show that the FF method improves 3.6% in AUC on the Brightkite dataset and 4.3% in AUC on the Gowalla dataset compared to the best day-Location view proposed in this paper. The FF method increases the AUC by 1.4% on the Brightkite dataset, and 1.6% on the Gowalla dataset compared to the DF method. Moreover, the FF method is 6.1% higher on the Brightkite dataset and 2.4% higher on the Gowalla dataset than the current best method SCI.

(3) This paper proposes an application framework for predicting the strength of social network relationships. The framework includes the following modules: The data storage and management module contains three sub-modules (data structure, data storage and data visualization); The data modeling module mainly modularizes the two fusion methods (FF and DF); The model evaluation module performs a comprehensive evaluation of performance of the forecasting method, and gives the LIFT, ROC and PR curve. Finally, the output the strongest user relationship strength pairs in descending sort.

Keywords: Users Relationship Strength, Spatio-temporal Data, Context Sequence, Context Co-occurrence Feature

Written by Caixu Xu

Supervised by JianFeng Yan

目　录

第一章 绪论

1.1 课题研究背景

随着互联网技术的迅速发展、地理定位技术的产生、传感器设备和全球定位系统（GPS）的普及[1]，用户更容易与朋友之间分享他们的签到数据，人们也越来越愿意分享他们的地理签到位置信息，所以导致大量的签到时空数据产生。当一个用户使用移动设备进行通信或签到时，即产生用户签到数据。该签到数据类型隐藏着各类实体、以及实体与实体间的属性关系，有其独特的研究价值和应用价值。这些数据很多时候同时包含时间维度和空间维度的信息，故我们称之为时空数据。时空数据挖掘作为一个新兴的研究领域，正致力于新兴的计算技术来分析海量、高维的时空数据，揭示时空数据中的有价值的知识。与此同时，基于时空数据的社交网络关系预测也正在逐渐兴起。典型的社交网络有Brightkite、Gowalla、Foursquare、街旁网和Yelp等。在社交网络中，用户不仅可以通过签到（Check-in）来跟踪和分享他们的位置信息，而且人们可以从时空数据的社交网络中获取很多信息，比如，人们可以在社交网络中，查询各个地方的美食、餐馆、酒店和旅游攻略等等。时空数据真实地反映了用户的位置信息，使得用户的网络世界和真实世界紧密地结合。由于这些优势，基于时空数据的社交网络受到人们的广泛喜爱并且得到快速发展。

移动传感设备的便捷吸引了数以百万的用户，用户在各个地方产生大量的时空数据。用户间的众多关系隐藏在这些时空信息中，互联网公司可以考虑利用这些时空数据来进一步优化他们的服务。不过目前互联网公司亟待需要合适的方法来挖掘这些隐藏着的潜在用户关系。通过对时空数据行为特征的分析和理解，把握人们的行为和思想倾向，从而为市场营销、信息传递、舆情的管理和导向等一系列关系到民生的现实问题提供了有效的技术支持[2]。企业也需要更好地理解这些时空签到数据，并提供相对准确的推荐信息，从而尽可能更好地识别用户的意图，进而更好地服务于用户。用户间的社交关系强度为更好地理解用户和挖掘用户需求提供了进一

步的可能。

时空数据的获取方式有显式获取和隐式获取两种。显式获取主要通过用户所产生带有地理标签的内容（如文本、图像、语音和视频）或通过手机APP（如切客、嘀咕、售票系统等）直接获取。隐式获取主要通过移动设备（如基站、GPS或WiFi热点）及信用卡交易等来源间接获取。这些数据的规模十分庞大，并且能真实地反映出现实生活中人们的活动规律。因此吸引了来自各领域的学者和专家对其进行研究和探索，他们希望对这些时空数据的分析能够帮助解决或者优化一些社会难题（比如城市规划、公共交通系统设计和流行病学等），或者通过研究可以发现新的社会规律。用户间的关系强度是社交网络中重要的组成部分，通过分析用户的时空数据可以预测他在现实生活中的社交关系强度。社交关系强度在各个领域得到广泛地应用。比如在社交网络中，社交关系分析可以用来进行好友推荐，避免用户在海量数据中进行筛选，使得用户很快找到新朋友，提高用户体验[3]；比如在社会安全领域，用户关系强度预测可以用来协助破案，通过已知的部分犯罪分子及其相关信息，挖取隐藏在关系网络中的信息，对潜在的犯罪团伙进行相关的排除检查[4,5]；在电子商务领域，用户关系强度预测可以帮助构建推荐系统，通过用户间的社交关系针对性地进行广告投放、优惠活动推出和客户推荐等，有效地节约用户的浏览时间并提高用户的购买欲等[6,7]。随着时空数据的增加，利用人工方式处理和分析时空数据以获得用户关系信息变得越来越困难。因此，如何利用非人工方式获取时空数据中的用户关系变得更加迫切。相对于传统数据类型，时空数据隐藏着社交网络结构更加复杂，但其内容更加丰富，从而使得社会网络数据的组织与管理以及对数据的处理成为了工业界关注的焦点。

基于时空数据的用户关系强度预测存在很多挑战。用户的移动和签到模式虽然存在很大随机性，但基于共现的方法得到了有效的证实[8]。通常可以直观得出这样的结论：因为共同的职业或者有共同的兴趣爱好，有较强关系强度的用户相比陌生人而言，有更高的可能性在同一场合一起出现，比如说同事正常在工作日一起出现，

好朋友经常一起出现在咖啡厅喝咖啡，晚上两亲密好友一起在私人场所讨论问题等等。基于这样的直觉，众多学者通过共现特征（即用户同时出现）来预测用户间关系强度[8–11]，同时也有学者提出基于上下文感知的方法来预测用户关系强度[12]。

基于时空数据的用户关系强度预测具有重要的意义。然而，时空数据的爆炸式增长，使得通过非人工方式快速、准确地预测用户社交关系已经成为一种现实的迫切需求。

1.2 课题研究现状

基于时空数据的用户关系强度预测方法根据它们的侧重点大致可以分成三种：(1) 基于轨迹相似的预测方法；(2) 基于上下文感知的预测方法；(3) 基于共现特征的预测方法。本节将分别介绍这三类预测方法的特点和研究现状。

1.2.1 基于轨迹相似的预测方法

Ying等人[13]首先为每个轨迹点指定一个预定义的标记，然后通过衡量用户位置序列信息中位置的标记信息的相似度来推荐用户，并提出了一种新颖的轨迹度量方法。Xiao等人[14]首先用历史位置信息来建立用户的GPS轨迹，例如餐馆、电影院等，然后通过使用最大行程匹配算法测量不同用户间的相似度。Chen等人[15]首先为用户经常访问的位置建立模型，然后应用频繁序列挖掘技术来提取用户频繁访问的地点序列，然后利用这些地点序列建立携带时空语义信息的用户移动画像。与此同时，Chen等人还聚焦于利用轨迹模式的相似性来衡量用户的相似性[16,17]，他们提供一个新颖的工具MinUS基于他们的签到数据来计算用户间相似性，该工具集成了识别用户轨迹模式的最好的模式挖掘技术，该工具可以管理移动数据、并且构建和比较用户的轨迹模式。

这类方法认为两两用户之间如果存在较多相似的轨迹特征，则两用户之间存在社交关系的可能性较大。基于轨迹相似的预测方法的优点是通过该方法可以获得用户的轨迹信息，且根据用户的轨迹信息的相似度度量方法可以得到用户的社交强度。

但该方法存在两个缺点：一是基于时空数据获得用户轨迹信息相对比较复杂；二是此类方法弱化了共现的概念，轨迹相似并非意味着共现，此类方法容易遭受检测停留点的错误[10]。

1.2.2 基于上下文感知的方法

在基于上下文的方法中，上下文通常包含社交上下文、用户偏好上下文、位置上下文和时间上下文。Bagci等人[12]提出基于随机行走的上下文感知朋友推荐算法(RWCFR)。该方法考虑了当前用户的上下文（当前用户社交关系上下文、当前用户喜好上下文和当前位置上下文）来提供关系推荐，他们建立当前用户无向无权图来表征社交网络的用户间关系、位置间关系及社交关系，同时他们提出随机行走方法来排序关系强度得分。

这类方法证实了上下文能一定程度刻画用户社交关系。但时空数据中常常含有时空共现信息，充分利用时空共现信息能进一步提升用户社交关系强度的预测[8]。

1.2.3 基于共现特征的方法

因为基于共现的方法挖掘了时空共现特征，该方法已经被证实相比于基于轨迹的方法而言能提高社交关系评估的精度[10]。基于共现的方法[9–11]通过实验已经得到较好的结果。

Crandall等人[8]通过来自社交网络Flickr中的带有地理标签的三千八百多万图片证实了共现特征对推理社交关系强度有重要的影响，社交关系强度将会随着共现次数的增多而急剧上升，同时也会随着共现时差的缩短而增多。同时，Cranshaw等人[18]还发现在公共地点的频繁共现可能来源于巧合且共现特征面临数据稀疏的问题。为了解决以上的问题，研究者提出了地点熵的概念，给每一个地点分配一个权重。我们给用户访问数多的地点赋予一个相对小的权值，表示该地点很可能是一个公共场所（比如：咖啡馆、图书馆）。相反地，给用户访问数少的地点赋予一个相对大的权重，表示该地点很可能是一个私人场所。

Pham等人[9]提出了基于熵的模型（EBM）的方法，该方法通过多样性和带权重的频次来分析人们在时间和空间上的共现，此外，该方法还将每个地点的特点考虑在内从而来弥补受限的地点位置信息，最终来评估社交关系强度。

Zhou等人[10]提出主题感知社交强度推理（TAI）的方法，该方法引入了从共现信息中挖掘主题（即共现单元），通过每个共现单元对社交强度的贡献来训练每个主题，然后利用这些共现单元来衡量两个用户之间的社交强度。

Njoo等人[11]提出了一种社交链接推理的框架（SCI），该框架从共现信息中量化了三种类型的关键特征（共现多样性特征、共现稳定性特征和共现持续性特征）。这三类特征用来区分朋友之间的共现和陌生人之间的偶遇，然后利用这些共现特征结合机器学习技术来预测用户社交关系，SCI模型的预测效果是目前最好的方法。

因为基于共现的预测方法全面挖掘了地点特征和共现特征，避免了用户之间的共现偶然性，所以基于共现特征的预测方法已经被证实比基于轨迹的预测方法精度更高[10]。

1.3 目前存在的主要问题

目前，针对时空数据预测用户关系强度的方法主要有三类：基于轨迹相似的预测方法、基于上下文感知的预测方法和基于共现特征的预测方法，各种方法的侧重点各不相同。针对上述研究现状的分析，主要存在的问题总结如下：

基于轨迹相似的预测方法获得用户时空轨迹信息相对比较复杂；同时此类方法弱化了共现的概念，轨迹相似并非意味着共现，此类方法容易遭受检测停留点的错误。

基于上下文感知的预测方法利用了各个角度的上下文信息（用户偏好上下文、社交关系上下文和位置上下文）来进行用户关系强度的预测。这类方法没有考虑用户的时间上下文，而用户签到的时间先后顺序能反映用户间的关系强度。同时这类方法没有考虑到时空的共现信息，而时间和空间共现的信息能很好地刻画用户间的

关系强度。

基于共现特征的预测方法从不同的角度来分析用户间的共现情况。但这类方法未考虑到时间和空间上下文信息，而用户间签到位置和时间的先后顺序能一定程度描述用户间的关系强度。

1.4　课题研究内容

本文的主要目的是从用户签到的时空数据中挖掘出隐藏在用户间的关系强度信息，即通过用户的签到时空数据来找出两两用户之间是否存在社交关系强度。签到时空数据是同时具有时间维度和空间维度的数据，时间维度的信息主要以签到时间表示，空间维度的信息主要以签到位置表示。本文分别从时间和空间维度着手提取时空数据中的上下文和共现信息，从而充分挖掘时空数据中预测关系强度的信息，并进一步提升关系强度预测的精度。表1-1显示了本文方法与基于上下文感知的方法和基于共现特征方法的不同，由表可知本文的方法综合考虑到时间共现、时间上下文、位置共现、位置上下文和用户签到周期性这些特征。

表 1-1　本文的方法和其它方法的比较

特点	RWCFR 2016 [12]	EBM 2013 [9]	TAI 2014 [10]	SCI 2017 [11]	本文方法
位置上下文	√				√
时间上下文					√
位置共现		√	√	√	√
时间共现		√	√	√	√
签到周期性					√

纵观文献中的方法，基于共现特征的方法[8–11]通过实验证实了共现特征的重要性，基于共现特征的方法无一例外的考虑了位置共现和时间共现，本文的方法也同时考虑了位置共现和时间共现。位置上下文对关系强度的预测也起着重要的作用[12]，本文的方法也考虑了位置上下文，同时因为用户关系强度会受签到时间上下文信息的影响，所以本文引入了时间上下文。用户签到的周期性对关系强度预测具

有一定作用[19]，本文通过一个特殊的视角将用户签到的周期性考虑在内。本文方法对于时空上下文和共现的融合并非简单传统的融合，上下文与共现的融合具有一定的新颖性。如图1-1，本文巧妙地将时空签到数据转化成用户签到的上下文序列，进而将时空数据中用户对的关系转化为自然语言处理领域（**N**atural **L**anguage **P**rocess）中单词对的关系。

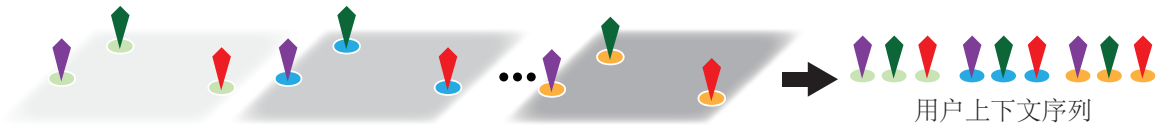


图 1-1 时空数据转化为用户上下文序列

本文首先提出了一种基于多视角的时空上下文共现特征预测方法，该方法同时考虑了位置上下文、时间上下文、位置共现、时间共现以及签到周期性；此外，针对多视角的特点，本文提出了基于视角融合的预测方法；最后，本文提出了针对工业界应用的解决方案。本文的主要研究工作和创新点如下：

(1) 目前大多数研究只针对于上下文或者共现。本文提出基于多视角时空上下文共现的方法进行不同视角上下文和共现的融合，本文基于时空上下文共现的方法并非简单地进行上下文和共现的融合。此外之前的方法没有考虑到时间的上下文，本文很好地考虑了时间的上下文。本文的方法分别从两个主视角（时间-空间视角、空间-时间视角）构建两种类型的上下文序列，时间-空间视角又将时间分为不同的时间粒度视角，基于不同的时间粒度分别构建三种视角上下文序列，其中一种时间视角描述了用户签到周期性这一特点。不同类型的上下文序列携带不同程度的时空上下文和时空共现信息。本文创新地将该上下文序列当作NLP领域中的语料，并利用NLP的工具分别基于多个视角提取上下文共现特征。用该上下文共现特征表征用户，分别基于多个视角利用机器学习技术训练并预测用户间的关系强度。实验验证了该方法的可行性。

(2) 根据每个视角的特点，分别提出了基于时空上下文共现特征融合（**F**eature **F**usion）的方法和基于多视角决策得分融合（**D**ecision **F**usion）的方法。FF方

法分析用户签到的周期性特点，将两个表征信息互补视角的上下文共现特征进行融合，用两个视角的特征组合去表征每个用户，并结合机器学习技术进行关系强度预测。DF方法是根据各个视角的特点提出一种可以很好拟合数据分布的融合策略，将多个视角的预测得分通过加权融合的方式得到最终的用户关系强度。实验结果证实了DF方法的性能超过各个单一视角的性能，此外本文提出的FF方法的性能超过了DF方法。

(3) 本文基于融合策略提出了用户关系强度预测的整体架构。该架构将海量的签到数据处理成结构化数据；并将结构化的签到数据进行数据的存储与管理；该架构同时提供了用户签到数据可视化功能；该架构分别基于FF方法和DF方法进行建模预测。本架构使用全方位的模型评估指标Lift曲线、ROC曲线和PR曲线。同时该架构还推荐出最有可能存在关系强度用户对，企业可以利用这些最有可能强关系的用户对信息提升对用户的理解，从而更好地服务于用户。

1.5　文章组织结构

全文共五章，组织结构如下：

第一章介绍了课题研究背景、课题的研究现状、目前存在的主要问题、课题研究内容和文章组织结构。

第二章主要介绍了基于多视角时空上下文共现用户关系强度预测方法。首先介绍了本章节所使用到的词向量，词向量是文档序列中单词的数学表征，本文巧妙地把它迁移过来表征用户上下文序列中的用户；然后介绍了本文使用到的机器学习模型XGBoost的基本原理；其次详细说明了多视角时空上下文共现用户关系强度预测的方法；最后给出实验结果与分析。

第三章介绍基于视角融合的用户关系强度预测方法。首先根据特征级的特点，提出了基于时空上下文共现特征融合（FF）的方法；然后根据决策级的特点，提出了基于多视角决策融合（DF）的方法；最后对融合方法的相关实验结果进行比较和

分析。

第四章介绍用户关系强度预测的应用。首先介绍用户关系强度预测的整体框架；然后介绍了用户签到数据的可视化；其次给出Lift曲线、PR曲线和ROC曲线对模型进行全面的评估；最后给出工业界的应用的解决方案。

第五章是对本文工作的总结和对未来工作的展望。

第二章　多视角时空上下文共现的预测方法

本章首先介绍了词向量表征与生成；然后介绍了所使用的机器学习模型XGBoost的原理；接着提出多视角时空上下文共现预测关系强度的方法，该方法巧妙地将时空数据中的用户间关系转化为NLP中同义词的关系，并从多视角提取用户上下文共现特征，同时结合机器学习技术进行用户关系强度的预测；最后给出实验结果与分析。

2.1　词向量

词向量是用来将自然语言中的词进行数学化的一种方式，本节主要从词向量的表征和词向量的生成两个方面进行介绍。

2.1.1　词向量的表征

词向量（Word Embedding）在NLP领域中起着非常重要的作用，它是自然语言中将词数学化的产物。通常词向量一共有两种表示方法，一种是One-Hot Representation表示方法，另一种是Distributed Representation表示方法。

One-Hot Representation是最直观词表示方法，这种方法把每个词表示为一个很长的向量[20]。这个向量的维度是词表大小，向量中绝大多数维度的值为0，只有一个维度的值为1，这个值为1的维度就代表了当前的词。这里举个例子说明：“话筒”表示为[0 0 0 0 1 0 0 ...]，“麦克”表示为[0 1 0 0 0 0 0 ...]。这种方法通常采用稀疏编码方式存储，也就是给每个词分配一个数字ID，比如以上的例子话筒记为4，麦克记为1 （假设从0开始编码），如果要编程实现的话，用Hash表给每个词分配一个编号就可以了。然而，这种表示方式的不当之处也非常的明显：(1)向量的维度会随着句子中词的数量增大而增大；(2) 任意两个词之间都是孤立的，根本无法表示出在语义层面上词与词之间的相关信息。

传统的One-Hot Representation仅仅将词符号化，不包含任何语义信息。Distributed

Representation表示可以避免One-Hot Representation的一些缺点，本文提取特征的方法中所表示的用户上下文共现特征即是采用Distributed Representation的方式。基于分布的词表示方法根据建模的不同，主要可以分为三类：基于矩阵的分布表示、基于聚类的分布表示和基于神经网络的分布表示[21]。我们在这里聚焦于基于神经网络的分布表示方法，神经网络词向量表示技术通过神经网络技术对上下文、以及上下文与目标词之间的关系进行建模[22]，由于神经网络较为灵活，这类方法的最大优势在于可以表示复杂的上下文，从而在词向量中包含更丰富的语义信息。Begio等人[22]在2001年提出神经网络语言模型（Neural Network Language Model），该神经网络方法在学习语言模型的同时，也首次得到了非常重要的神经网络分布表示的副带产物词向量（Word Embedding）。

2.1.2 词向量的生成

词向量生成的一个重要工具是word2vec，word2vec是一种更快训练神经网络语言模型的高效实现。word2vec提供了两种重要模型[23,24]分别为**C**ontinuous **B**ag-**o**f-**W**ords模型和Skip-gram模型。与此同时对于CBOW和Skip-gram两个模型分别给出两套框架，它们分别是基于Hierarchical Softmax的框架和Negative Sampling的框架[25]。在这里我们聚焦于Skip-gram模型，此模型的目标函数[25]正是本文方法所构建序列的目标函数，同时本文方法选用Skip-gram模型的Hierarchical Softmax的框架。

Skip-Gram模型主要包含三层：输入层、投影层和输出层，在已知当前词w_t的前提下，预测其上下文$Context(w_t)$即w_{t-2},w_{t-1},w_{t+1},w_{t+2}（见图2-1）。对于基于Skip-gram 的模型，其优化的目标函数则如公式2-1。

$$Obj = \sum_{w \in C} logP(Context(w)|w) \tag{2-1}$$

图2-2给出Skip-gram模型的Hierarchical Softmax框架实现，它包括三层：输入层、投影层和输出层。下面以样本$(w, Context(w))$为例，对这三层分别做简要的说明。

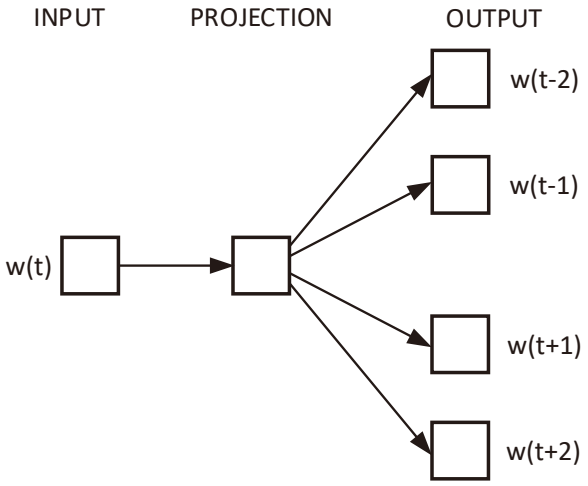


图 2-1　Skip-gram模型

(1) 输入层（Input Layer）：只含当前样本的中心词w的词向量$v(w) \in \mathbb{R}^m$；

(2) 投影层（Projection Layer）：在这里是恒等投影，把$v(w)$投影到$v(w)$。

(3) 输出层（Output Layer）：输出层对应一棵二叉树，它是以语料中出现过的词当叶子结点，以各词在语料中出现的频次当作权值构造Huffman树[26]，从而为利用Hierarcical Softmax技术奠定了基础。在这棵Huffman树中，叶子结点的个数共N个，即是词典中词的数目（对应图2-2中浅色的结点），非叶子结点数目为$N-1$个（对应图2-2深色的结点）。

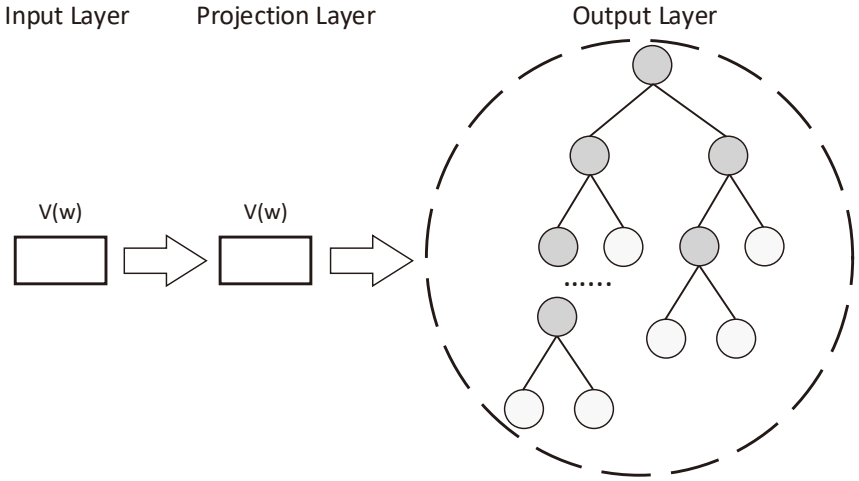


图 2-2　Skip-gram模型Hierarchical Softmax实现

Hierarchical Softmax是word2vec中用于提高训练速度的一项关键技术，在具体介

绍算法之前，首先介绍该算法中涉及的各个变量符号及其意义，我们在推导过程中使用的向量默认为列向量。考虑到Huffman树中的某个叶子结点，假设它对应词典D中的词w，表2-1是word2vec的skip-gram模型的Hierarchical Softmax技术实现的算法符号及其表示意义。

表 2-1　word2vec算法符号定义

符号	表示意义
p^w	从根结点出发到达w对应叶子结点的路径
l^w	路径p^w中包含结点的个数
p_1^w,p_2^w,...,$p_{l^w}^w$	路径p^w中的l^w个结点，其中p_1^w表示根结点，$p_{l^w}^w$ 表示词w 对应的结点
d_2^w,d_3^w,...,$d_{l^w}^w \in \{0,1\}$	词w的Huffman编码，它由l^w-1位编码构成，d_j^w 表示路径p^w 中第j个结点对应的编码（根结点不对应编码）。这里约定Huffman树的左结点为1，右结点为0。
θ_1^w,θ_2^w,...,$\theta_{l^w-1}^w$	路径p^w中非叶子结点对应的向量，θ_j^w表示路径p^w 中第j 个非叶子结点对应的向量

对于Hierarchical Softmax框架，已知的是当前词w，需要对当前词的上下文$Context(w)$中的词进行预测，目标函数为2-1，公式2-1的关键是条件概率函数$p(Context(w)|w)$ 的构造，现在对公式2-1的$P(Context(w)|w)$ 进行改写：

$$P(Context(w)|w) = \prod_{u \in Context(w))} p(u|w) \tag{2-2}$$

上式中的$p(u|w)$可按照Hierarchical Softmax思想[27]改写为：

$$p(u|w) = \prod_{j=2}^{l^u} p(d_j^u|v(w), \theta_{j-1}^u) \tag{2-3}$$

其中，$\sigma(x)$函数为神经网络中常用的激活函数之一sigmoid函数，其具体的数学定义为：$\sigma(x) = \frac{1}{1+e^{-x}}$。在这里，我们将$\sigma(v(w)^\top \theta_{j-1}^u)$定义为分到Huffman树右子树结点的概率，则$1-\sigma(v(w)^\top \theta_{j-1}^u)$定义为分到Huffman树左子树结点的概率，则最终$p(d_j^u|v(w), \theta_{j-1}^u)$可以改写成公式2-4：

$$p(d_j^u|v(w),\theta_{j-1}^u) = [\sigma(v(w)^\top\theta_{j-1}^u)]^{1-d_j^u}\cdot[1-\sigma(v(w)^\top\theta_{j-1}^u)]^{d_j^u} \tag{2-4}$$

将公式2-4代入公式2-3，其次将合并后的公式2-3代入公式2-2，再将合并后的公式2-2代回到2-1。依次回溯代入化简可得到最终对数似然目标函数2-1的具体表达式（即公式2-5）：

$$\begin{aligned} Obj &= \sum_{w\in C} log \prod_{u\in Context(w)} \prod_{j=2}^{l^u} [\sigma(v(w)^\top\theta_{j-1}^u)]^{1-d_j^u}\cdot[1-\sigma(v(w)^\top\theta_{j-1}^u)]^{d_j^u} \\ &= \sum_{w\in C} \sum_{u\in Context(w)} log \prod_{j=2}^{l^u} [\sigma(v(w)^\top\theta_{j-1}^u)]^{1-d_j^u}\cdot[1-\sigma(v(w)^\top\theta_{j-1}^u)]^{d_j^u} \\ &= \sum_{w\in C} \sum_{u\in Context(w)} \sum_{j=2}^{l^u} \{(1-d_j^u)\cdot log[\sigma(v(w)^\top\theta_{j-1}^u)] + d_j^u\cdot log[1-\sigma(v(w)^\top\theta_{j-1}^u)]\} \end{aligned} \tag{2-5}$$

至此，公式2-5就是word2vec工具的Skip-gram模型Hierarchical Softmax框架的目标函数，该模型训练优化方式是通过随机梯度上升法对其进行优化，梯度的更新是通过反向传播的原则进行更新[28]。这里我们为了下面梯度推导的方便，将公式2-5中大括中的内容$\{(1-d_j^u)\cdot log[\sigma(v(w)^\top\theta_{j-1}^u)] + d_j^u\cdot log[1-\sigma(v(w)^\top\theta_{j-1}^u)]\}$ 记为$Obj(w,u,j)$，即为公式2-6：

$$Obj(w,u,j) = (1-d_j^u)\cdot log[\sigma(v(w)^\top\theta_{j-1}^u)] + d_j^u\cdot log[1-\sigma(v(w)^\top\theta_{j-1}^u)] \tag{2-6}$$

至此，我们需求分别计算出$Obj(w,u,j)$关于$v(w)$和θ_{j-1}^u的梯度。在这里我们首先计算$Obj(w,u,j)$关于θ_{j-1}^u的梯度：

$$\begin{aligned}\frac{\partial Obj(w,u,j)}{\partial \theta_{j-1}^u} &= \frac{\partial}{\partial \theta_{j-1}^u}\{(1-d_j^u)\cdot log[\sigma(v(w)^\top \theta_{j-1}^u)] + d_j^u \cdot log[1-\sigma(v(w))^\top \theta_{j-1}^u]\} \\ &= (1-d_j^u)\cdot[1-\sigma(v(w)^\top \theta_{j-1}^u)]\cdot v(w) - d_j^u \cdot \sigma(v(w)^\top \theta_{j-1}^u)\cdot v(w) \\ &= \{(1-d_j^u)\cdot[1-\sigma(v(w)^\top \theta_{j-1}^u)] - d_j^u \cdot \sigma(v(w)^\top \theta_{j-1}^u)\}\cdot v(w) \\ &= [1-d_j^u - \sigma(v(w)^\top \theta_{j-1}^u)]\cdot v(w)\end{aligned} \tag{2-7}$$

所以，θ_{j-1}^u的更新公式可写为公式2-8：

$$\theta_{j-1}^u = \theta_{j-1}^u + \eta\cdot[1-d_j^u - \sigma(v(w)^\top \theta_{j-1}^u)]\cdot v(w) \tag{2-8}$$

由于$Obj(w,u,j)$中的参数$v(w)$和θ_{j-1}^w是对称关系，所以我们也可以很容易地计算出$Obj(w,u,j)$ 关于$v(w)$的梯度：

$$\frac{\partial Obj(w,u,j)}{\partial v(w)} = [1-d_j^u - \sigma(v(w)^\top \theta_{j-1}^u)]\cdot \theta_{j-1}^u \tag{2-9}$$

所以，$v(w)$的更新公式可写为公式2-10：

$$v(w) = v(w) + \eta\sum_{j=2}^{l^u}[1-d_j^u - \sigma(v(w)^\top \theta_{j-1}^u)]\cdot \theta_{j-1}^u \tag{2-10}$$

算法1以样本$(w, Context(w))$给出Skip-gram模型随机梯度上升更新的伪代码：

算法 1: Word2vec的Skip-gram模型的Hierarchical Softmax框架

for each $u \in Context(w)$ **do**
  $e = 0$
  for $j = 2$ *(to)* l^u **do**
    1: $q = \sigma(v(w)^\top \theta_{j-1}^u)$
    2: $g = \eta(1-d_j^u - q)$
    3: $e = e + g\theta_{j-1}^u$
    4: $\theta_{j-1}^u = \theta_{j-1}^u + gv(w)$
  $v(w) = v(w) + e$

2.2　XGBoost分类器

XGBoost（**E**xtreme **G**radient **Boost**ing）[29]是大规模并行的决策树集成学习（Ensemble Learning）[30]的工具，数据科学家广泛地使用该工具。决策树算法的关键是对节点的分裂，XGBoost中决策树主要是以回归树的形式。所以本节首先介绍回归树的分裂，然后介绍XGBoost原理及其实现。

2.2.1　回归树分裂

XGBoost中的回归树是一种特殊的分类回归树，分类回归树(**C**lassification **A**nd **R**egression **T**ree) [31]是机器学习领域中应用最广的归纳推理算法之一，并且是广泛使用的树模型学习方法。CART既可以用于分类问题也可以用于回归问题，它是一种逼近离散值函数的方法，对噪声数据有很好的健壮性且能够学习较为复杂的空间分界面。XGBoost模型中使用到的CART树是二叉树，该算法根据训练数据集，从根结点开始，不断地将样本空间一分为二，最后在叶子结点得到该样本的得分。通常，CART树中分成两种类型的结点：非叶子结点和叶子结点。CART树的生成是一种启发式构树的过程，其关键是对节点的分裂，经典的决策树算法ID3是通过信息增益（Information Gain）作为衡量训练样例集合纯度的标准[32]，从而对非叶子结点进行分裂，XGBoost模型中CART树的分裂采用了一种新的高效合理的方式来定义增益值并进行非叶子结点的分裂[29]。

这里的回归树是采用了二阶泰勒展开的方法[33]并在正则化函数部分作了改进。XGBoost是Boosting的集成方式[29]，即模型的训练是以增量的方式（后一棵树依赖于前一棵树的状态）。这里假设上一棵树的对第i个样例的预测结果为$\hat{y}_i^{t-1}$，第i个样例的真实值为y_i，则定义损失函数为$Loss(y_i, \hat{y}_i^{t-1})$。$g_i$为第$i$样本的损失函数$Loss(y_i, \hat{y}_i^{t-1})$在$\hat{y}_i^{t-1}$的一阶偏导，$G_L$为当前左子树结点的一阶偏导之和（即左子树$g_i$之和），$G_R$为当前右子树结点的一阶偏导之和（即右子树$g_i$之和）；$h_i$为第$i$样本的损失函数$Loss(y_i, \hat{y}_i^{t-1})$在$\hat{y}_i^{t-1}$的二阶偏导，$H_L$为当前左子树结点的二阶偏导之

和（即左子树h_i之和），H_R为当前右子对结点的二阶偏导之和（即右子树h_i之和）；λ和γ为自定义的正则项常数，这两个参数是为了防止模型的过拟合。

算法 2: 回归树贪心策略分裂

Input : 前t-1棵树的预测值$\hat{y}^{t-1}$;当前结点的样本结合D
Output: 决策树T

创建一个子树的根节点T;
$Gain = 0$;
G=Loss在$\hat{y}^{t-1}$处当前结点样本一阶偏导之和;
H=Loss在$\hat{y}^{t-1}$处当前结点样本二阶偏导之和;
for $k = 1$ **to** m **do**
    $G_L = 0, H_L = 0$;
    for $j \in$ *sorted(当前第k个特征的所有样本)* **do**
        $g_j = Loss$在$\hat{y}^{t-1}$处x_j样本的一阶偏导;
        $h_j = Loss$在$\hat{y}^{t-1}$处x_j样本的二阶偏导;
        G_L=G_L+g_j, H_L=H_L+h_j;
        G_R=G-G_L, H_R=H-H_L;
        $score = max(score, \frac{1}{2}[\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{G^2}{H+\lambda}] - \gamma)$;
for each $d \in D$ **do**
    if d满足叶子结点条件 **then**
        $T.isLeaf = True$;
    else
        $T.isLeaf = False$;
        将$T.samples$按照$score$所对应特征样本的值将D分成两份D_1, D_2;
        $P.left.samples = d_1$;
        $P.right.samples = d_2$;
        $T.add(\{P.left, P.right\})$;
return T;

XGBoost实现了一种贪心算法将非叶子结点拆分成两个结点进行启发式构树，由此可以定义一棵树的增益（公式2-11）。各项的含义为：（1）新增左叶子结点的得分；（2）新增右叶子结点的得分；（3）原根结点的得分；（4）新增叶子复杂性的代价（该参数默认为0；该值越大模型就越学得的模型保守）。对于决策树分裂的关键是通过公式2-11查找到最优的分裂，算法2呈现了贪心搜索策略查找最优的分裂。

$$Gain = \frac{1}{2}[\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{(G_L+G_R)^2}{H_L+H_R+\lambda}] - \gamma \tag{2-11}$$

2.2.2 XGBoost原理

XGBoost是大规模并行的树集成机器学习算法，自从被提出以来就获得广泛的使用，它是使用增量技术进行监督学习方法，同时它有训练速度快、子树之间可以并行和模型可解释性强等优点。

在众多的竞赛项目中，例如Kaggle数据挖掘比赛，众多选手选用XGBoost作为他们的学习器并取得很好的成绩。在KDDCup 2015 比赛中，前10名的选手全都选用XGBoost作为他们的分类器。它对回归树的分裂进行了优化，并对Boosting树的分裂结点进行暂存从而实现多棵树生成的并行。算法3给出XGBoost基本训练过程，对于给定n个样本m维特征的数据集：$D = \{(x_i, y_i)\}(|D| = n, x_i \in \mathbb{R}^m, y_i \in \mathbb{R})$，运用$K$个增量函数来预测输出结果（公式2-12），其中$F$是分类回归树（即CART树）。

$$\hat{y} = \sum_{k=1}^{K} f_k(x_i), f_k \in F \tag{2-12}$$

算法 3: XGBoost基本训练过程

Input ：训练集D
Output: 集成树F

for $k = 1$ **to** K **do**
　使用数据集D和前$k-1$棵树构建第k棵回归树T_k，构建时重复下面的步骤直到满足结束条件：
　　(1) 根据回归树的分裂标准从这m个特征从选出一个最好的进行分裂;
　　(2) 将节点分成两个子节点。
Return 集成树F

2.3 多视角时空上下文共现特征预测方法

本节首先给出的重要符号定义；然后分别从二个主视角Location-Time（即空间-时间视角）和Time-Location（时间-空间视角）提出上下文共现特征预测的方法，其中Time-Location视角根据时间粒度的不同，本文分别提出Day-Location、

Hour-Location和Minute-Location三个子视角[34]。每个视角我们分别详细介绍了上下文序列的生成、上下文共现特征的提取和分类器的学习与预测。图2-3显示了多视角时空上下文共现特征预测方法的概要图。

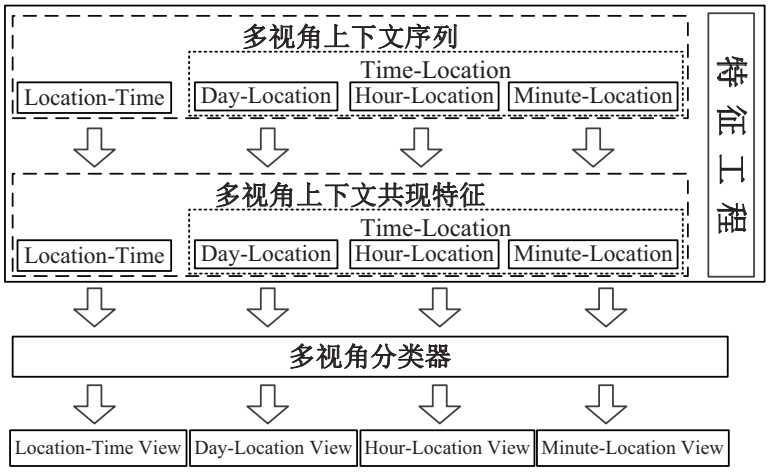


图 2-3 多视角时空上下文共现的预测方法

表 2-2 算法符号定义

变量	英文全称	中文全称
u	user id	用户标识
U	$\{u_1, u_2, u_3,...\}$ all different user id set	不同用户标识的集合
t	check-in timestamp	用户签到时间戳
l	(longitude, latitude), location identification	用户签到地点
c	$\{u, t, l\}$, user check-in information	用户在特定时间地点签到信息
C	$\{c_1, c_2, c_3, ...\}$ all check-in data set	所有用户签到数据的集合
C_t	$\{c_1, c_2, ..., c_n\}$, all c at the same time period	在相同时间域签到的数据集合
$C_{t\&l}$	elements in C_t rank according nearest distance	元素在C_t 中按最短距离排序
$C_t^{Sequence}$	$\{C_t^1, C_t^2, ..., C_t^N\}$,elements rank according time order	按签到时间顺序排序的序列
C_l	$\{c_1, c_2, ..., c_m\}$, all c at the same location	在相同地点域签到的数据集合
$C_{l\&t}$	elements in C_l rank according time order	元素在C_l 中按时间顺序排序
$C_l^{Sequence}$	$\{C_l^1, C_l^2, ..., C_l^M\}$,elements rank according nearest distance	按签到距离最短排序的序列
Φ	$\{u_1, u_2, u_3, ...\}$, a sequence with spatiotemporal context co-occurrence information	时空上下文共现信息序列

我们首先对本文方法所需的符号给出定义。在表2-2中，本文列出算法中使用到的参数符号。本文用$u \in U$ 标识每个用户；用$c \in C$ 标识每个用户的签到数

据，每个签到数据c用$\{u,t,l\}$反映一个用户u在特定位置l特定时间t的签到信息，每个用户u通常对应多条签到数据c；$C_t=\{c_1,c_2,c_3,...,c_n\}$表示所有签到数据在相同时间域的集合；$C_{t\&l}$表示所有在$C_t$中的数据按签到位置的距离最短原则进行排序后的序列；$C_t^{Sequence}=\{C_t^1,C_t^2,C_t^3,...,C_t^N\}$表示元素块$C_t$按时间顺序排序后的序列；$C_l=\{c_1,c_2,c_3,...,c_m\}$ 表示所有在相同地点域签到的数据；$C_{l\&t}$表示所有在C_l中的数据按照签到时间先后顺序排序后的序列；$C_l^{Sequence}=\{C_l^1,C_l^2,C_l^3,...,C_l^M\}$表示元素块$C_l$按照签到位置的距离最短原则进行排序后的序列；$\Phi$表示只包含用户标识的上下文序列，$\Phi$同时携带时空上下文和时空共现信息。以上符号之间存在内在关联$\sum_{i=1}^{M}|C_l^i|=\sum_{j=1}^{N}|C_t^j|=|C|=|\Phi|$，$\Phi$通常包含重复的用户标识。

2.3.1　空间-时间视角上下文共现预测

这节描述了如何从原始的签到数据中构建空间-时间视角的上下文序列，时间信息通过签到时间来描述，空间信息通过签到位置的经纬度来描述。并从该上下文序列中提取上下文共现特征，最后利用该特征组合进行关系强度的预测。

算法 4: 空间-时间上下文序列

Input : C
Output: Φ

```
Define Φ = ∅;
C_l^Sequence = SortLocationByDistance(C);
for all C_l in C_l^Sequence do
    C_l&t = SortTime(C_l);
    for all c in C_l&t do
        Φ.append(c.userid);
return Φ;
```

空间-时间(Location-Time)上下文序列： 该空间-时间上下文序列生成过程在算法4中给出。其中$SortLocationByDistance$函数是根据签到位置距离最短原则进行排序并产生序列$C_l^{Sequence}$。$C_l^{Sequence}$中元素的距离最短原则描述如下：第一个位置标识是随机选择一个用户标识作为第一个位置标识；离第一个用户标识位置最近的一个

用户标识作为第二个位置标识；之后离前一个用户标识位置最近的一个用户标识位置作为当前位置；依此类推，离第$(M-1)$个用户标识位置最近的用户标识位置指定为第M个用户标识位置。由于时间信息是一维信息，所以$SortTime$函数是采用快速排序算法依据时间顺序进行元素的排序。算法4返回值就是由用户标识组成的空间-时间下文序列Φ。如图2-4所示，注意不同椭圆的颜色的距离和顺序，以及菱形的颜色的顺序。因为我们所使用的数据集是基于位置社交网络的数据集，位置的稀疏性使得我们没有考虑基于位置域来划分。此空间-时间上下文序列捕捉了强位置共现、同时携带最短位置上下文和最短时间间隔上下文信息。

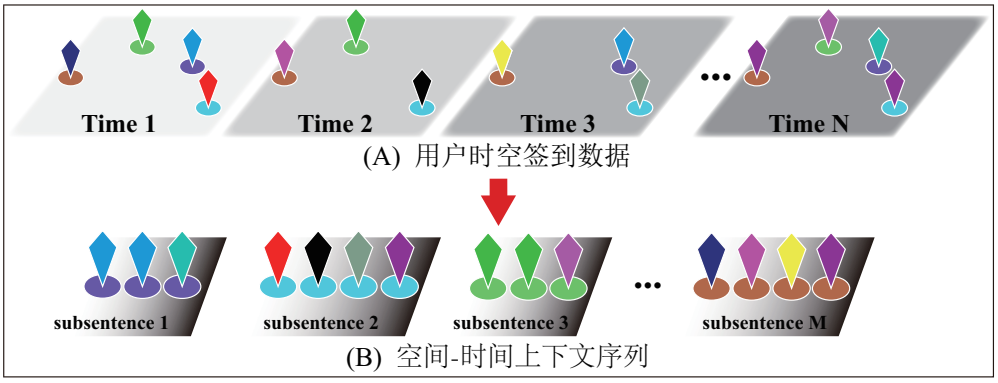


图 2-4　菱形代表用户，椭圆型代表位置，时间变化通过背景色的深浅来表示。子图(A)表示在原始数据中，特定的时间域内，不同的用户在不同的位置签到；子图(B)表示空间-时间上下文序列，即所有用户在相同的位置签到的上下文序列。

上下文共现特征： 本文的上下文共现特征不同于传统的方法，并非是简单的上下文和共现特征的融合。我们巧妙地利用word2vec工具从上述的上下文序列中提取上下文共现特征。工具word2vec通常用在语料中（这里的语料通常可以看作单词的序列）查找同义词，我们把它迁移过来查找用户对之间的关系强度。该工具将大量语料作为输入，从而形成高维的空间，每个词都可以被指定为空间中的一个向量[24]。在工具从语料学习完之后，向量空间中具有相似上下文的词向量间的余弦非常的接近[23,24]，所学得的词向量也是本文所需要的上下文共现特征。我们从空间-时间视角的上下文序列中提取用户上下文共现特征（即Context Co-occurrence Feature）。

具体地，给定一个上下文序列$\Phi = \{u_1, u_2, u_3, u_4, ..., u_T\}$，该序列携带时空上下文语义表征信息，我们的目标是最大化平均对数概率2-13：

$$\begin{aligned} Obj &= \sum_{u \in \Phi} logP(Context(u_t)|u_t) \\ &= \sum_{t=1}^{T}[\sum_{j=-k}^{k} logP(u_{t+j}|u_t)] \end{aligned} \tag{2-13}$$

其中Φ是序列中的所有元素，$Context(u_t)$为用户u_t的上下文，k为上下文窗口的大小。内循环从$-k$到k计算在给定u_t的状态下预测u_{t+j}的对数概率，外循环遍历上下文序列中所有的用户，窗口两端的值用边界值进行填充。每个用户关联两个可学习的参数向量w_u和v_u，它们分别是用户u可学习的“输入”和“输出”向量[23]，在给定用户标记u_j的情况下正确预测u_i的概率可定义为：

$$p(u_i|u_j) = \frac{exp(w_{u_i}^{\top} v_{u_j})}{\sum_{l=1}^{U} exp(w_l^{\top} v_{u_j})} \tag{2-14}$$

其中U是上下文序列Φ中不同的用户标记，模型的优化方法是采用随机梯度下降，梯度的计算方式是利用反向传播的原则[23]。每个用户的上下文语义特征v_u（在自然语言处理领域也称为词向量）能够学习到，词向量v_u所捕捉了词的共现和词的上下文信息对我们非常有用。

在空间-时间视角中，该上下文共现特征捕捉用户强位置共现、最短时间上下文和最短位置上下文。

分类器的学习与预测： 本文选择XGBoost分类器作为学习器。每个用户标识通过用户上下文共现特征来表征，两用户的特征的组合作为分类器的输入。考虑到用户间的关系是双向的，即用户A与用户B有关系也就等价于用户B与用户A有关系，所以特征组合中相对较大的对应位置的向量值放在前面，相对较小的对应位置的向量值放在后面。本文用用户上下文共现特征组合在训练集上训练学习器，并用训练

好的分类器在测试集上进行预测。分类器的输出$\hat{y}$（即公式2-12）是用户关系强度的相对得分，本文为了最终性能评估的方便将输出的用户关系强度值按降序排序。

2.3.2 时间-空间视角上下文共现预测

时间-空间（Time-Location）上下文序列： 该时间-空间上下文序列的整个生成过程在算法5中给出。由于时间信息是一维信息，所以$SortTimeByGranularity$函数采用$QuickSort$快速排序根据时间域的先后顺序进行排序并产生序列$C_t^{Sequence}$，时间参数τ可以精确到天、小时、和分钟（即$24h, 1h, 1/60h$）。序列$C_t^{Sequence}$中的元素块C_t中的元素按照距离最短原则进行排序。C_t中的元素排序的距离最短原则描述如下：如果之前无位置标识，随机选择一个用户标识位置作为第一个位置；否则，离前一个用户标识位置最近的一个用户标识位置作为当前位置；依此类推，离第$(n-1)$个用户标识位置最近的用户标识位置被指定为第n个用户标识位置。如图2-5(B)所示，该算法的返回值就是时间-空间上下文序列Φ。此上下文序列从不同的粒度捕捉了时间共现、同时包含最短时间上下文和最短位置上下文的信息，我们根据时间粒度的取值，继续将时间划分为Day-Location、Hour-Location和Minute-Location序列。

算法 5: 时间-空间上下文序列

Input : C
Output: Φ

Define $\Phi = \varnothing$;
$C_t^{Sequence} = SortTimeByGranularity(C, \tau)$;
for all C_t **in** $C_t^{Sequence}$ **do**
    $C_{t\&l} = SortLocationByDistance(C_t)$;
    for all c **in** $C_{t\&l}$ **do**
        Φ.append($c.userid$);
return Φ;

*Day-Location*上下文序列： 时间粒度参数τ设置为天（即$24h$），如图2-5，每个时间域设置为一天：第一个时间域设置为第一天；第二个时间域设置为第二天；依

此类推，第N个时间域为第N天。相同时间域C_t中的元素按距离最短原则进行排序。因为人们签到周期性是基于天的[19]，比如说人们在工作日时间通勤，再比如人们每天晚上家里签到等等。所以Day-Location上下文视角是一种特殊类型的视角，由于人们签到的周期性是以天为单位的，所以Day-Location 上下文序列捕捉了人们签到周期性。并且同时携带天共现、最短时间上下文和最短位置上下文的信息。

*Hour-Location*上下文序列：整个生成过程与Day-Location上下文序列的过程类似，但是时间粒度参数τ设置为小时（即$1h$），相比于Day-Location 上下文序列而言，这将产生越来越多的时间块，但是在每个时间块中产生更短的子序列。Hour-Location 上下文序列捕捉适度的时间共现和时空上下文的信息。

*Minute-Location*上下文序列：整个生成过程类似于Hour-Location上下文序列，但是时间粒度参数τ 设置为分钟，相比于Hour-Location上下文序列而言，这将产生越来越多的时间块，但是在每个时间块中产生更短的子序列。此上下文序列捕捉强时间共现、并且携带时空上下文的信息。

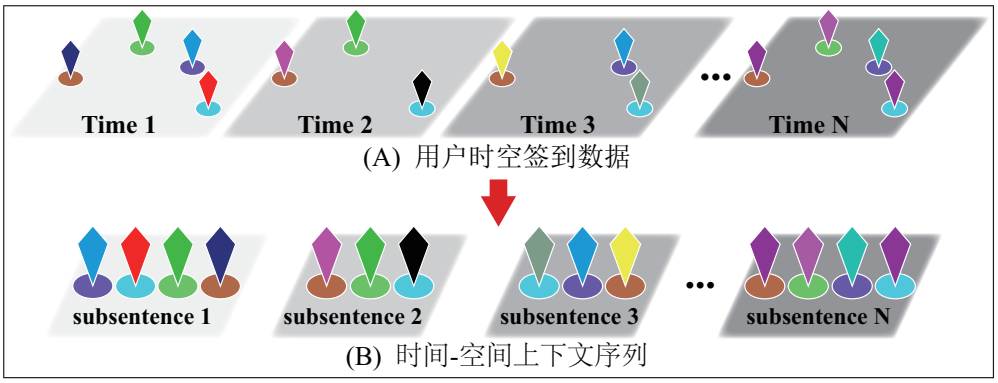


图 2-5　菱形代表用户，椭圆型代表位置，时间变化通过背景色的深浅来表示。子图(A)表示在原则数据中，特定的时间域内，不同的用户在不同的位置签到；子图(B)表示时间-空间上下文序列，即所有用户在相同的时间域签到的上下文序列。

上下文共现特征：不同的时间-空间上下文序列描述不同程度时间共现和不同程度时空上下文。时间-空间上下文共现特征的生成与空间-时间上下文共现特征的生成方法相同，都是巧妙地利用word2vec工具从上述的上下文序列中提取上下文共现持征（即Context Co-occurrence Feature）。序列中的具有相似上下文用户标识的向量

余弦通常非常地小。时间-空间视角与空间-时间视角生成过程虽然相同，但两者表征的信息确完全相反。空间-时间上下文共现特征表征强位置共现、最短时间上下文和最短空间上下文，而时间-空间上下文共现特征表征时间共现、最短空间上下文和最短时间上下文。

分类器的学习与预测： 时间-空间视角的分类器设置与空间-时间视角的分类器参数设置相同。用用户所对应的上下文共现特征组合表征用户对，然后利用分类器在训练集上用户对进行学习。学习后的分类器在测试集上的输出值即为用户关系强度，最终将用户关系对按强度值降序排序。

2.4 实验

2.4.1 实验数据集和数据预处理

本文实验采用的数据集是基于两个真实的数据集（Brightkite数据和Gowalla数据）[35]。Brightkite曾经是一个基于位置的社交网络服务应用服务，用户可以用它进行签到，并分享他们的位置信息。Brightkite数据集包括58,228名用户在2008年4月到2010年10月之间发生的4,491,143条签到数据和用户间的214,078 对社交链接关系数据。其中用户签到的数据大小为373,164KB，用户社交链接关系的数据大小为4,472KB。Gowalla也是一个基于位置的社交网络服务应用，其于2009 年在美国德克萨斯州的奥斯汀成立，用户也能通过签到来分享他们的地理位置。2011年Facebook将Gowalla收购，Gowalla在次年3月关闭应用。Gowalla数据集包括196,591名用户在2009年2月到2010年10月之间发生的6,442,890条签到数据和用户间的950,327对社交链接关系数据。其中用户签到数据集大小为385,397KB，用户社交链接关系的数据大小为21,601KB。

Brightkite数据集和Gowalla数据集的数据结构是相同的，它们的签到数据集中记载了用户的签到时间、地点信息。在签到数据集中每一个用户都用唯一的用户标识u标志，其中地点信息用经度和纬度表示，签到数据被处理成三元组$< u, t, l >$，

其中l通过经纬度来表示。表2-3记录着数据中用户的签到数据样例。其中用户的时间维度的信息通过用户的签到时间t来体现，用户的空间维度的信息通过用户签到的位置（即经纬度）来体现。

表 2-3　用户签到数据样例

用户标识(u)	签到时间(c)	经度	纬度
58186	2008-12-03T21:09:14Z	-105.317215	39.633321
58186	2008-11-30T22:30:12Z	39.633321	-105.317215
58187	2008-08-14T21:23:55Z	41.257924	-95.938081
58187	2008-08-14T07:09:38Z	41.257924	-95.938081
58190	2009-04-08T07:01:28Z	46.421389	15.869722

表 2-4　用户签到数据样例

用户标识(u_1)	用户标识(u_2)	标签
0	1	True
0	2	True
0	123	False
0	3	True
0	121	True
4	123	False
4	0	False
3	4	False

在用户对关系数据集中，主要记载了存在社交关系的两个用户。原始数据集中的标签是非常不平衡的[11,35]。通常有四种广泛使用的方法来处理数据不平衡问题[36]：(1)不平衡训练；(2)过采样；(3)欠采样和(4)指定样本权重。我们这里使用欠采样来解决数据的不平衡问题。因为原始数据中未提供负样例，原始数据集只提供正样例（所有标签都为真），除正样例外任意两个用户都不存在关系，所以可以合成负样例使得负样例的数目等于正样例的数目。本文将用户对数据也被处理成一个三元组$< u_1, u_2, label >$，$label$标签标记着两个用户是否存在关系。表2-4是处理后数据集中用户关系对的样例。本文按70%,20%和10%的比例来分别划分用户关系对数据为训练集、验证集和测试集。

2.4.2 实验环境

本节中的实验运行在服务器上，CPU型号为Intel(R) Xeon(R) CPU X5690@ 3.47GHz，24核，128G内存。

服务器的操作系统是CentOS，上面配置了Python 3.6。本文使用的数据均以结构化形式存储在MySql上，对数据的预处理、特征工程均基于Python。由于XBoost对Boosting Tree做了优化并可以并行，所以多个内核可达到并行执行以充分利用CPU资源。

2.4.3 实验设定

本文通过算法4获得携带时空上下文信息的空间-时间上下文序列。本文通过指定参数τ的粒度分别为天、小时和分钟（即$24h, 1h, 1/60h$），在算法5中，获得三种类型的时间-空间上下文序列，不同的粒度捕捉不同程度时空上下文与共现。通过算法4和算法5，本文可以最终生成四种类型携带不同程度上下文共现信息的序列。

Word2Vec工具提供了skip-gram框架，该框架的目标函数2-1与本文方法的目标函数2-13一致，所以我们选择skip-gram框架。因为计算Skip-gram框架的$\bigtriangledown logP(u_i|u_j)$时间复杂度是与$U$成正比的，所以公式2-14的计算是非常耗时的。框架skip-gram对于Full Softmax一个高效的替代是Hierarchical Softmax，后者大大地降低了计算$logP(u_i|u_j)$的时间复杂度（大约是U的对数）[27]。因此我们采用Skip-Gram框架的Hierarchical Softmax方法[23,27]来提取上下文共现特征。对于word2vec有一些重要参数需要设置[37]，词向量$v(w)$的大小$size$设置为200，滑动窗口设置为10（即公式2-13中k的值设置为10），学习率η为0.01，其它参数值设置为默认，本文选用Skip-gram框架的Hierarchical Softmax方法。在工具分别从四种类型的上下文序列中学习完之后，便可以获得用户上下文共现特征。基于四个视角，可以获得四种类型的上下文共现特征，本文用200维的上下文共现特征来表征每个用户。

本文从四个表征不同程度上下文共现信息的视角分别训练四种分类器。在本

文的实验中，我们选择XGBoost分类器做预测，每个用户通过上下文共现特征表征。考虑到用户关系的双向性，两用户特征对应位置特征的较大值放置到前面，然后用户特征组合作为模型的输入。模型的输出值是用户关系强度的相对概率。对于XGBoost模型有一些重要的参数需要设置：提升树参数设置为gbtree；为了避免过拟合树的最深度设置为3；学习率设置为0.1；目标函数设置为binary:logistic；迭代次数设置为1000；其它参数设置为默认。根据四种类型的特征分别训练四种类型的分类器，如图2-3所示，随后用学习完的多样化分类器在测试集上预测关系强度，即Location-Time、Day-Location、Hour-Location 和Minute-Location 四种类型的得分。

2.4.4　评价标准

本节采用二分类领域通用的性能评价指标：精确率（Precision）、召回率（Recall）来评价模型的性能，精确率和召回率也在近期的二分类任务评估中应用[38,39]。根据预测结果，可以得到表2-5的4种预测情况。

表 2-5　二分类问题的结果混淆矩阵

真实情况	预测情况	
	正例	负例
正例	True Positive (真正例)	False Negative (假负例)
负例	False Positive (假正例)	True Negative (真负例)

根据表2-5，可以具体定义精确率（Precision）和召回率（Recall）。根据以上的公式，Precision反映了预测出来的正样例中真正的正样例所占的比例（即公式2-15），Recall反映了真正的正样例中被预测出来的正样例的比例（即公式2-16）。

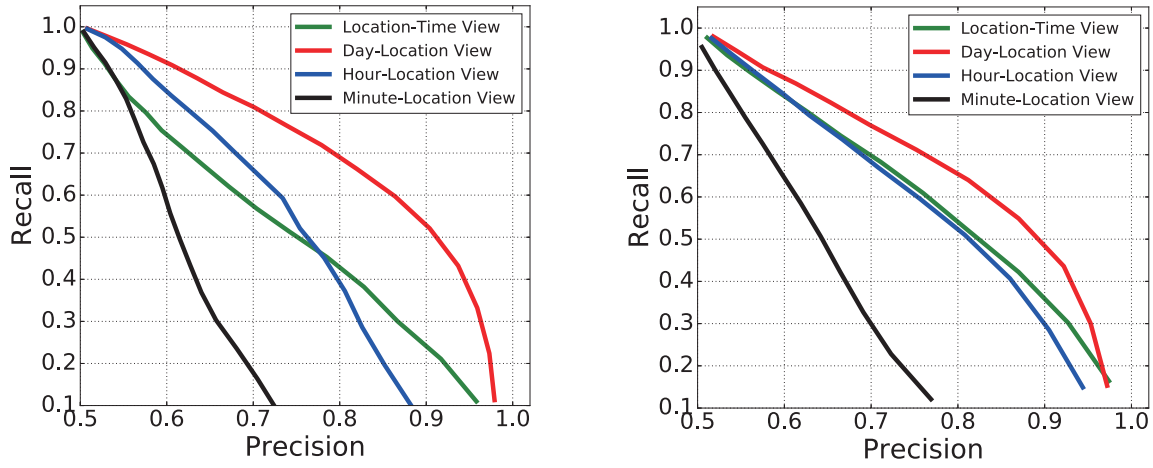
$$Precision = \frac{TP}{TP + FP} \tag{2-15}$$

$$Recall = \frac{TP}{TP + FN} \tag{2-16}$$

PR曲线是一个评估分类器性能的常用指标之一，本文是以Precision为横坐标、Recall为纵坐标。PR曲线显示了针对不同阈值的Precision和Recall之间的权衡，曲线下方面积越大代表更高的Precision和Recall，更高的Precision代表更低的假正例，更高的Recall代表更低的假负例。由此可见，Precision和Recall这两个指标的值越大，我们提出的上下文共现预测的方法效果就越好。

2.4.5　实验结果与分析

结果1： 图2-6呈现了四个视角间的比较。在两个数据集上，Day-Location视角的性能均超过其它三个视角的性能；Minute-Location视角的性能最差，远低于Hour-Location和Day-Location两个视角的性能；Location-Time视角在Brightkite数据集上效果和Hour-Location视角差不多，而在Gowalla数据集上Location-Time 视角超过了Hour-Location视角。Location-Time视角在Brightkite数据集上稳定性较差，而在Gowalla数据集上相对稳定。



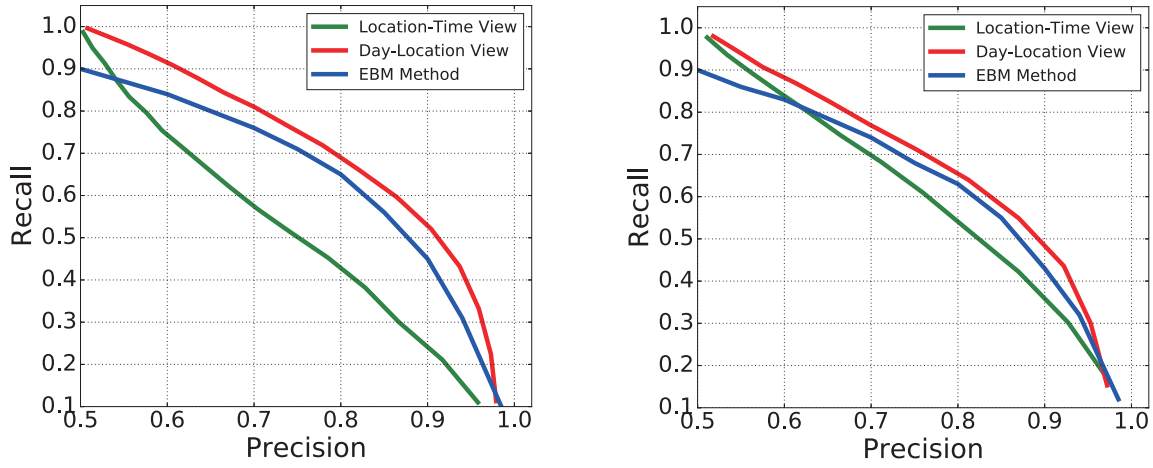
(a) Brightkite的PR曲线　　(b) Gowalla的PR曲线

图 2-6　多视角上下文共现方法的PR曲线

分析1： Day-Location视角是一种特殊的Time-Location视角，此视角能更好地描绘了周期性的时间共现、最短空间上下文和最短时间上下文，人们的周期性签到的特性和数据集的稀疏性使得以粗粒度的Day-Location视角在两个数据集比

其它两个Time-Location视角更好。Location-Time视角捕捉了强位置共现、最短位置上下文和最短时间上下文，因为Brightkite数据集位置的灵动性与稀疏不均性，所以在该视角的性能在Brightkite数据集上不是非常稳定，而在Gowalla数据集位置稀疏度相对稳定，所以在该视角的性能相对较稳定。Hour-Location视角捕捉适度的时间共现和时空上下文，共现粒度是以小时为单位，在很多场合是非常适用的，在Brightkite 和Gowalla数据集效果一般。因为基于位置社交网络的数据稀疏性，所以Minute-Location在当前数据集上比其它两个Time-Location 视角效果要差，尽管此视角在当前数据集上效果较差，不过该视角捕捉了强时间共现和最短时空上下文，此视角对于短时间共现频繁发生的稠密数据非常适用。

结果2： 图2-7呈现了本文方法中两个视角与算法EBM比较。在Brightkite数据集上，Day-Location视角在相同的Precision下Recall比EBM算法[9] 最高提高约10%；在Gowalla数据集上，Day-Location视角在相同的Precision下Recall比EBM算法最高提高约8%。Location-Time视角在两个数据集上性能均低于EBM算法。



(a) Brightkite的PR曲线　　(b) Gowalla的PR曲线

图 2-7　本文的方法和EBM方法的对比

分析2： Location-Time视角捕捉了强位置共现，因为数据集位置稀疏性，所以在该视角的性能差于EBM算法，但该视角很好地捕捉了强位置共现，在稠密的位置数据集效果会很好。Day-Location视角很好地描述了用户周期性签到行为，并且该方

法不仅考虑了时空共现，而且还考虑到位置上下文与时间的上下文，该视角方法在两个数据集上均超过了EBM算法。

2.5 本章小结

本章首先介绍了词向量表征形式，同时介绍了词向量的生成方式；然后介绍了XGBoost中回归树分裂的基本原理及分裂过程，同时引入了XGBoost模型；接着引出了本文提出的多视角上下文共现特征的预测方法；通过多个视角性能的比较，分析了各个视角所携带不同的表征信息并分析了结果；通过与EBM算法比较，本文的Day-Location视角在Brightkite和Gowalla数据集上均超过EBM算法。

第三章　基于视角融合的预测方法

在第二章中给出多时视角时空上下文共现的方法，每个视角表征不同程度的时空上下文和时空共现信息。本章基于每个视角的不同特性，分别提出基于时空上下文共现特征融合的方法（**F**eature **F**usion）和基于多视角决策融合的方法（**D**ecision **F**usion）。两者融合的精度比各个单一视角的方法得到进一步提升。由于特征之间的互补性，基于时空上下文共现特征融合的方法（FF）比基于多视角决策融合的方法（DF）效果好。

3.1　基于时空上下文共现特征融合预测

本节首先介绍特征级融合的基础知识；然后详细分析用户签到的周期性特性；最后基于特征级的特点提出基于时空上下文共现特征融合的方法。

3.1.1　特征级融合

根据融合对象的不同，我们可以将融合分为数据级融合、特征级融合和决策级融合[40]。当特征类型相同，可以对特征进行融合。特征级融合就是特征层面上的融合，是将提取的特征进行融合与分析[41,42]。特征级融合是中间层次的融合，整个过程如图3-1所示。首先，从原始数据中提取表征不同程度信息的特征；然后用特征的组合来表征实体；最后对融合的特征进行机器学习识别。特征与决策分析息息有关，因此，特征级融合最大限度给出了决策分析需要的特征信息。

空间-时间视角即在2.3.1节中，空间-时间上下文共现特征携带强位置共现、相同位置最短时间上下文共现和位置之间最短距离上下文共现信息。而时间-空间视角即在2.3.2节中，时间-空间上下文共现特征携带时间共现、相同时间域最短位置上下文共现和时间最短上下文信息。因为这两个主视角特征呈现互补的形式，所以我们将这两组互补的特征进行融合。

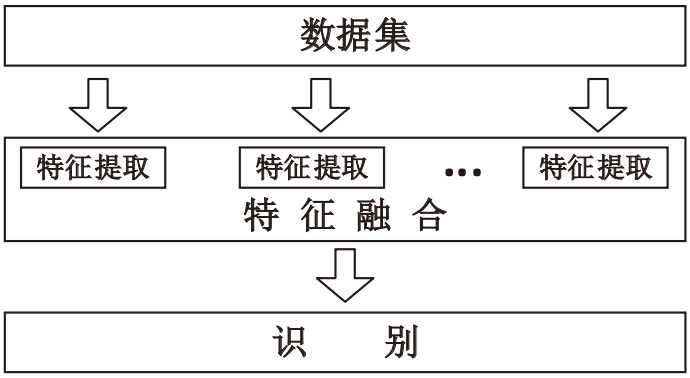
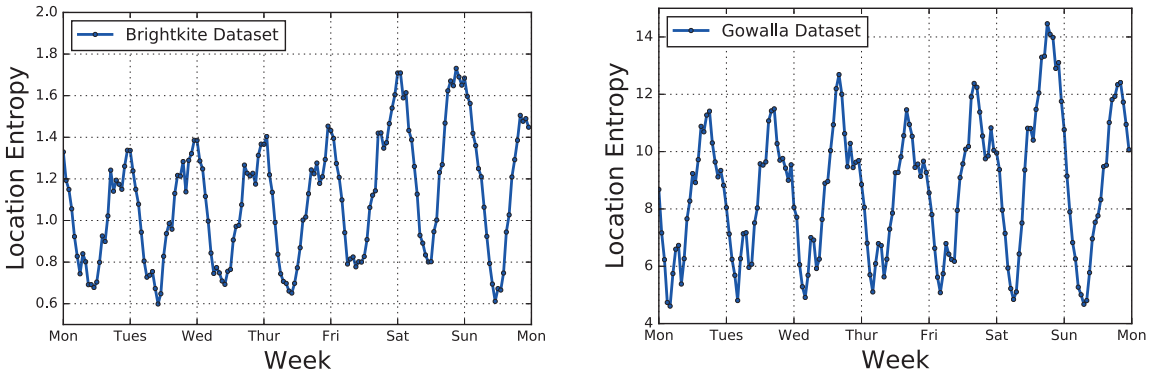


图 3-1　特征级融合

3.1.2　用户签到的周期性

时间-空间视角中分为Day-Location、Hour-Location和Minute-Location视角。Day-Location视角包含Hour-Location和Minute-Location视角所表征的信息，数据集的稀疏性使得以粗粒度天的视角更能拟合当前数据分布。Day-Location视角描述了以天为粒度的共现、同一天中位置最短距离的上下文共现和最短时间上下文共现信息。与此同时，它表征了其它两个视角所没有的信息，Day-Location 视角还捕捉了人们签到的周期性，比如人们在工作日时间正常通勤，再比如说人们每天晚上在家里签到等等。正因为这个原因，所以Day-Location视角的性能高于Hour-Location和Minute-Location视角的性能。



(a) Brightkite数据集的位置熵　　(b) Gowalla数据集的位置熵

图 3-2　位置熵的周期性

如图3-2所示，在Brightktie数据集和Gowalla数据集上，位置熵在工作日呈现很

强的以天为单位的周期性。因为每个人都在家，所以每天早上6点左右平均位置熵最低。随后当人们在高峰时段的时候，位置熵会逐渐提升。晚上人们社交行为，导致位置熵逐渐提升。凌晨人们的社交行为越来越少，所以导致位置熵逐渐降低。依据人们签到的周期性，本文特征融合的方法聚焦于时间-空间视角中的Day-Location视角。

3.1.3　基于上下文共现特征融合的方法

基于特征融合的概要图如3-3所示。本文首先从两个主视角（Location-Time视角和Time-Location视角）分别构建用户上下文序列，其中Time-Location视角本文选取Day-Location视角，该视角包含其它两个视角表征信息，同时该视角表征用户强签到周期性。然后基于两个主视角分别提取用户上下文共现特征，因为Location-Time视角和Day-Location视角的表征信息的互补性，所以我们将这两组互补的特征进行融合，同时融合的特征包含用户签到的周期性信息。最后我们将两个视角用户上下文共现特征的组合作为新的特征输入到XGBoost模型进行学习和预测。

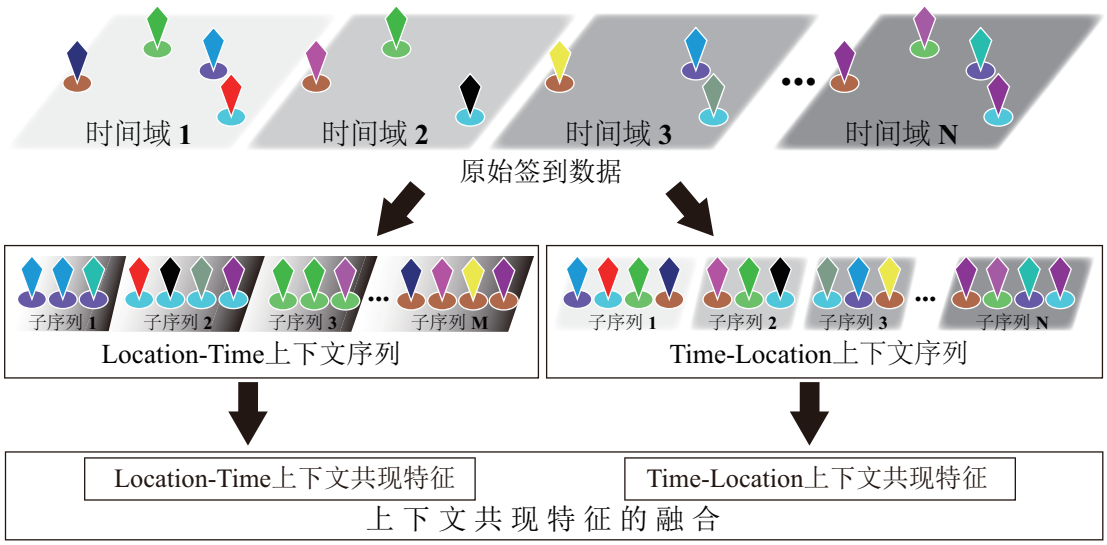


图 3-3　基于上下文共现特征融合

3.2　基于多视角决策融合预测

本节首先介绍决策级融合的基础知识；然后基于决策级的特点提出基于多视角决策融合的方法。

3.2.1　决策级融合

决策级融合是将每个特征的识别结果进行融合，然后实现联合识别的过程[43,44]。决策级融合是一种高层次融合，并且是属于联合决策结果，这种联合决策结果的精度通常比单个结果更高。决策级融合能有效地反映各个侧面不同类型的信息，且能反映不同类型的不同特征信息。当一个或几个识别信息有误时，通过适当地融合，进行再识别，从而获得正确的结果。

决策级融合对原始数据进行数据分析、特征提取和特征识别。图3-4是决策级融合的整个过程，首先是根据数据集提取特征，然后根据特征来分别识别决策目标。接着，将数据集所提取特征的识别预测结果进行融合，再根据识别结果进行融合来获取最终的结果。

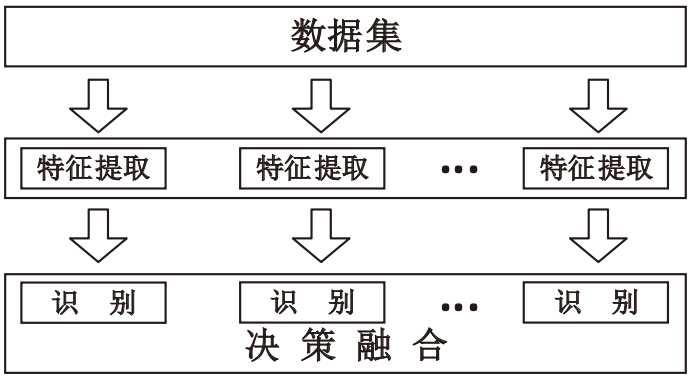


图 3-4　决策级融合

3.2.2　基于多视角决策融合的方法

基于决策融合方法的概要图如图3-5所示。本文首先从多视角构建携带时空语义信息的上下文序列；然后从携带不同时空上下文共现信息的上下文序列中提取多视

角上下文共现特征；接着基于多视角分别用上下文共现特征训练模型并进行关系强度的预测；最后对各视角的得分进行融合得到最终的用户关系强度。

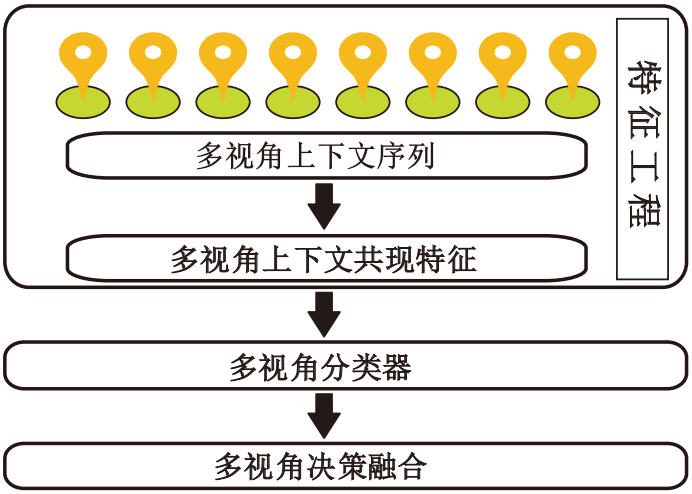


图 3-5　基于多视角决策融合

第二章节的多视角时空上下文共现的方法可以获得四种类型的关系强度得分（即Location-Time得分、Day-Location 得分、Hour-Location 得分和Minute-Location 得分），我们考虑到每个视角的特点并提出了基于多视角决策融合方法。我们的决策融合方法是一种高层推理融合，能够反映不同类型视角的信息，同时也能避免单一视角表征信息的不足。决策融合具体可以通过以下的公式3-1计算:

$$RelationshipStrength = \sum_{d \in D} Score_{d_i} \cdot W(d_i) \tag{3-1}$$

其中D是多视角，d_i是多视角中的一个视角，$W(d_i)$是单个视角的权重，$Score_{d_i}$是单一视角的用户关系强度得分，$W(d_i)$ 的具体计算通过以下的方式：首先通过以下方式计算$W'(d_i)$：

$$W'(d_i) = \frac{Val\_AUC_i - Val\_AUC_{min}}{Val\_AUC_{max} - Val\_AUC_{min}} + \sigma \tag{3-2}$$

其中Val_AUC_{max}是多个视角验证集上AUC最大值，Val_AUC_{min}是多个视角验证集上AUC 最小值，σ是足够小的值（比如0.01），然后通过将$W'(d_i)$归一化

得到$W(d_i)$。最后根据值$Score_{d_i}$和值$W(d_i)$通过公式3-1计算出最终的用户关系强度$Relationship$，并将关系强度结果以降序的顺序排序。

尽管该决策融合策略非常的简单，但是它是非常有效的，公式3-2给予更多的权重给更能拟合当前数据集分布的视角，正是由于真实数据时间和空间稠密的不确定性使我们的融合策略更加有效。

3.3 实验

本实验利用第二章实验部分2.4.1的处理好后正负样例平衡的数据集，分别基于两种融合策略分别进行评估。

3.3.1 实验设定

基于时空上下文共现特征融合的方法： 特征提取阶段有一些重要的参数需要设置。word2vec参数设置为：滑动窗口的大小设置为10，词向量的大小设置为200（即用户上下文共现特征向量的大小），学习率设置为0.01，其它参数值设置为默认。分别基于Location-Time和Day-Location这两个视角提取上下文共现特征，然后将这两组特征进行融合。由此每个用户被映射400维的向量，将用户对的向量组合作为XGBoost模型的输入。XGBoost模型存在一些重要的参数：提升树参数设置为gbtree；树的最大深度设置为3；提升树的学习率设置为0.1；目标函数设置为binary:logistic；迭代次数设置为2000次，其它参数设置为默认。

基于多视角决策融合的方法： 分别从四个视角提取上下文共现特征，特征提取阶段的word2vec参数设置与基于特征融合的方法是相同的。我们可以获取四种类型的用户上下文共现特征，也就是Location-Time特征、Day-Location特征、Hour-Location特征和Minute-Location特征。分别基于这四种类型的特征进行XGBoost分类器的训练。XGBoost模型的参数将迭代次数设置为1000次，其余参数和基于时空上下文共现特征融合的方法参数设置保持一致。可以在训练集上训练出四种不同视角的模型，训练好的模型在验证集上计算出Val_AUC，四个视角的Val_AUC分

别是：Location-Time视角为0.7406；Day-Location视角为0.8493；Hour-Location视角为0.7479；Minute-Location视角为0.6581。然后基于公式3-2计算出每个视角的权重，其中σ设置为0.01，四个视角的权重分别为；0.2274，0.5203，0.2471和0.0052。最后根据公式3-1计算出最终的用户关系强度。

3.3.2　评价标准

在本章方法中，除了用第二章实验部分的性能指标精确率（Precision）和召回率（Recall）来评价模型以外。还引入了AUC(Area Under ROC Curve)来直观评价预测方法的优劣。在介结AUC之前，我们要先介绍下假正类率（False Positive Rate, FPR）和真正类率（True Positive Rate，TPR），其中FPR和TPR的定义如下：

$$TPR = \frac{TP}{TP + FN} \tag{3-3}$$

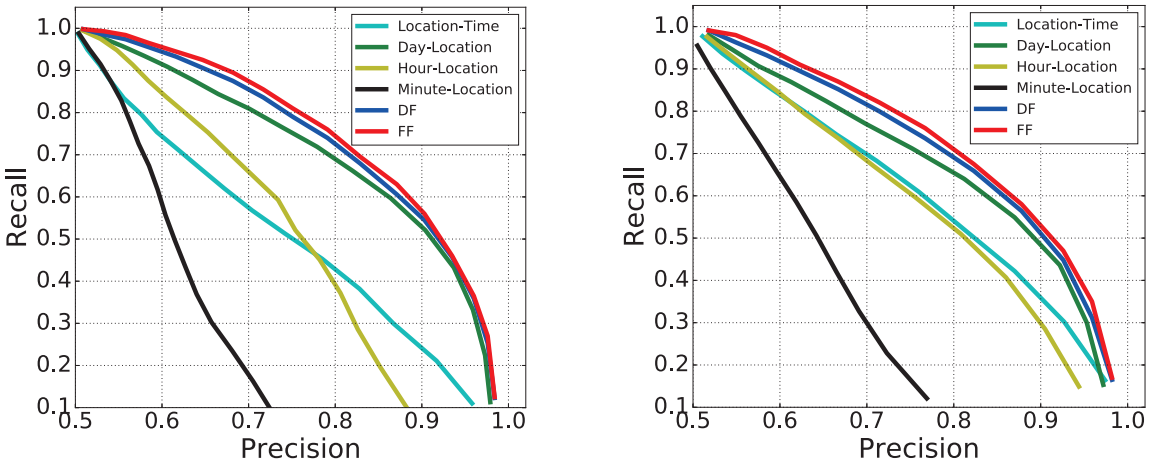
$$FPR = \frac{FP}{FP + TN} \tag{3-4}$$

ROC曲线是不同阈值下真正类率（TPR）和假正类率（FPR）的变化关系。将FPR作为横坐标、TPR作为纵坐标，可以通过二维平面曲线刻画ROC。对于ROC曲线，如果预测方法越好，则其ROC曲线就越接近左上角。但是，ROC曲线是一个曲线图，无法给出模型好坏的度量值。AUC就是ROC曲线下的面积的大小，AUC的值越大，则表示模型的性能越好。AUC可以更好地比较模型的好坏。所以本文采用AUC（**A**rea **U**nder **C**urve）来评估模型的好坏。

3.3.3　实验结果与分析

结果1： 图3-6呈现了两种融合策略和其它四个视角分别在Brightkite和Gowalla数据集上的比较。基于多视角的**DF**方法在两个数据集均超过了其它四个视角。通过融合四个视角的得分，**DF**方法效果超过了任何一个单一视角，在相

同Precision下Recall均高于Location-Time视角、Day-Location视角、Hour-Location视角和Minute-Location视角。具体地，**DF**在Brightkite数据集上当Precision为0.7时Recall超过最好的Day-Location视角5%；在Gowalla数据集上当Precision为0.7时Recall超过Day-Location视角4.8%。



(a) Brightkite的PR曲线　　(b) Gowalla的PR曲线

图 3-6　融合策略与单一视角方法比较的PR曲线

分析1： 在本文的**DF**方法中，四个视角不同的侧重点满足了多样性原则，并且每个视角捕捉不同程度的时空共现与时空上下文信息，公式3-2对于能更好刻画当前数据集的视角给予更多的权重，实验的结果证实了**DF**方法能更好地平衡了单一视角的不足。

结果2： 图3-6同时呈现了**FF**方法在两个数据集均超过了其它四个视角，并且超过了**DF**方法。具体地，**FF**方法在Brightkite数据集上当Precision为0.7时Recall超过最好的Day-Location视角6.9%；**FF**在Gowalla数据集上当Precision为0.7时Recall超过最好的Day-Location视角6.5%。**FF**方法在Brightkite数据集上当Precision为0.7时Recall超过**DF**方法1.9%；**FF**方法在Gowalla数据集上当Precision为0.7时Recall超过**DF**方法1.7%。

分析2： 基于特征融合的方法（FF）性能超过基于决策融合的方法（DF）。FF方法中没有采用Hour-Location视角和Minute-Location视角。因为Day-Location携

带了其余两个Time-Location视角的表征信息，数据的稀疏性使得Day-Location视角更能拟合数据的分布，与此同时Day-Location 携带签到的周期性信息。由于Day-Location与Location-Time视角特征的互补性，这里去除短时间共现Hour-Location和Minute-Location特征的影响，导致FF方法的性能分别在Brightkite和Gowalla数据集上均超过了DF方法的性能，同时超过了其它四个视角。

结果3： 表3-1给出了FF方法、DF方法和其它单一视角的AUC比较情况。DF方法比其它单一视角的AUC值都要好，DF方法比最好的Day-Location视角的AUC值在Brightkite数据集上要提高2.2%，在Gowalla数据集上要提高2.7%。在Brightkite数据集上，FF方法比Day-Location方法AUC值提高3.6%，FF方法比DF方法的AUC值要提高1.4%；在Gowalla数据集上，FF方法比Day-Location方法AUC值提高4.3%，FF方法比DF方法的AUC值要提高1.6%。

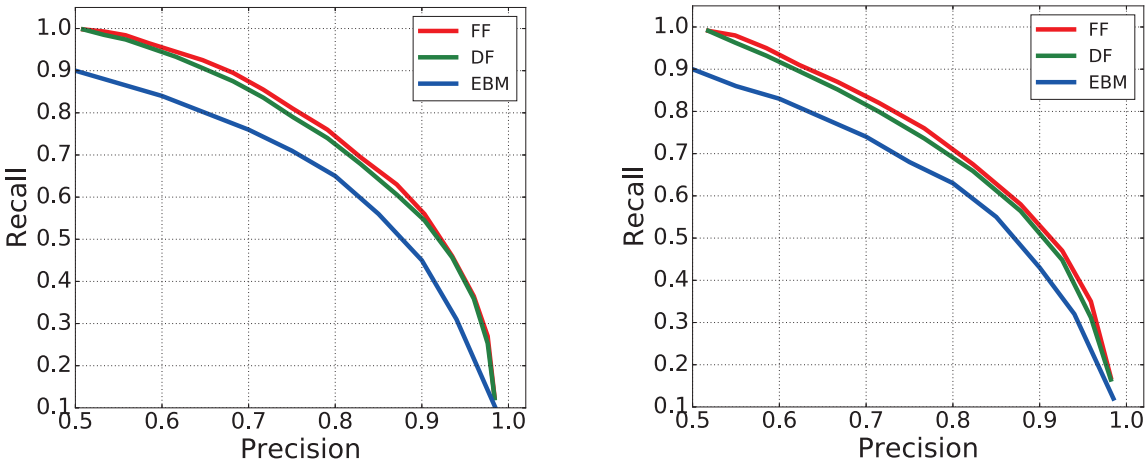
表 3-1　融合与其它视角的AUC比较

测试集	Brightkite数据集	Gowalla数据集
Location-Time视角	0.730691	0.778668
Day-Location视角	0.829309	0.806326
Hour-Location视角	0.747866	0.761170
Minute-Location视角	0.635823	0.644667
FF	**0.865329**	**0.849248**
DF	**0.851635**	**0.833099**

分析3： DF方法将四个视角的得分融合，四个视角满足多样性并且每个视角都有效果，DF方法平衡单一视角的不足。FF方法根据其中最好的视角Day-Location和Location-Time视角的特征特点，进行融合，因为特征的互补性，同时减少了短时间共现视角的噪音，使FF方法超过DF方法。

结果4： 图3-7显示了FF方法和DF方法与EBM算法的比较。DF方法与FF方法均超过了EBM算法。具体地，在Brightkite数据集上当Precision为0.7时，DF超过EBM算法9.3%，FF超过EBM算法11.2%；在Gowalla数据集上当Precision为0.7时，DF超过EBM算法8.4%，FF超过EBM算法10.1%。

分析4： FF方法与DF方法不仅考虑了时空的共现，而且还考虑了时空的上下文，



(a) Brightkite的PR曲线　　(b) Gowalla的PR曲线

图 3-7　融合方法与EBM方法比较的PR曲线

与此同时携带用户签到周期性这一特性，而EBM考虑时间的共现和位置的共现。所以导致本文提出的两种融合方法超过了EBM方法。

结果5： 图3-8比较了FF方法与SCI方法[11]和PGT方法[45]。其中SCI方法是目前最好的方法，PGT方法同时考虑了个人因素、全局因素并且结合时间因素来评做用户关系强度。FF方法在两个数据集上均超过了PGT方法。下面主要量化FF方法与目前最好的SCI方法，在Brightkite数据集上，FF方法在AUC指标上超过SCI方法6.1%；在Gowalla数据集上，FF方法在AUC指标上超过SCI方法2.4%。

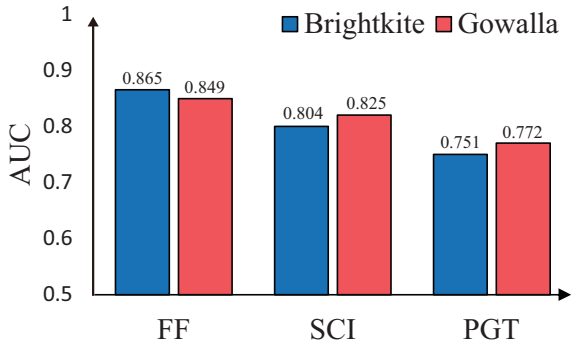


图 3-8　FF方法与SCI、PGT的比较

分析5： SCI方法是目前为止最好的方法，该SCI框架量化了时空的共现特征，然后利用这些共现特征结合机器学习技术来预测用户的社交关系。SCI框架中量化的

共现特征中考虑到用户签到的周期性，在FF方法中Day-Location视角也考虑到其周期性。本文的FF方法挖掘了时空共现特征的同时，还考虑到时空的上下文。

3.4　本章小结

本章首先依据视角特征的互补性提出基于时空上下文共现特征融合（FF）的方法；然后依据决策级的特性，提出基于多视角决策融合（DF）的方法；最终对这两种融合的策略进行了量化的比较。实验表明，基于融合的方法均超过了第二章各视角的性能，由于特征的强互补性，基于特征融合的方法比基于多视角决策融合的方法更好。同时，本文提出的FF方法在两个数据集上均超过了目前最好的SCI方法。

第四章　用户关系强度预测的应用

本章首先介绍用户关系强度的整体框架。然后介绍了整体框架中的数据存储与管理模块，存储与管理模块包括数据结构化、数据的存储和数据可视化功能。接着介绍了数据建模模块。最后提供一种用户关系强度的应用解决方案。

4.1　用户关系强度预测的整体框架

如图4-1所示，该框架主要包括数据的存储与管理、数据建模和模型性能评估三大模块。

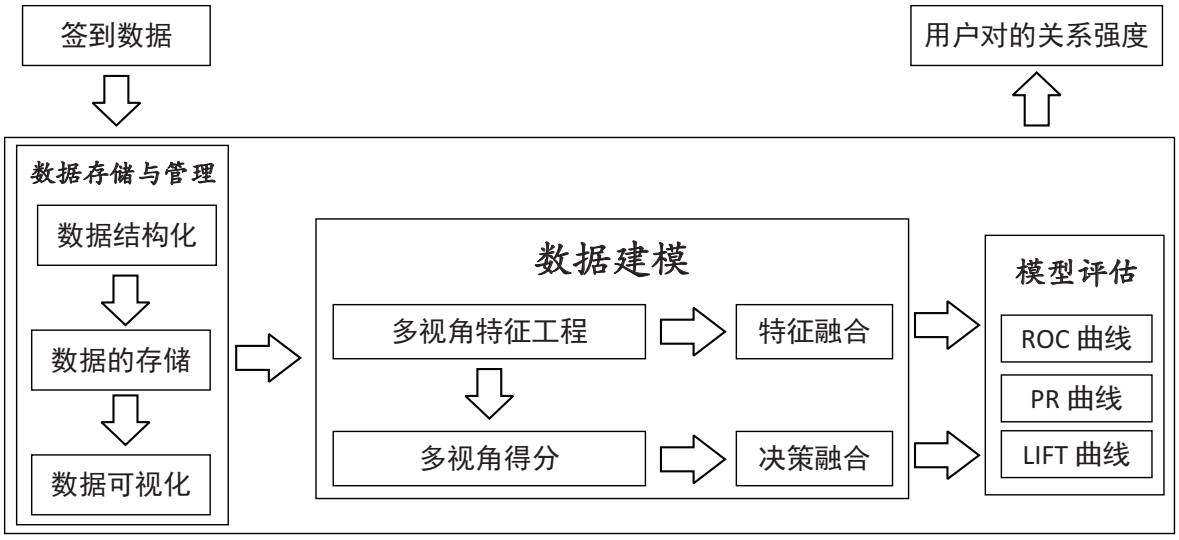


图 4-1　关系强度预测的整体框架图

该框架从互联网网站或者从移动设备端获取用户签到的时空数据，通过数据结构化模块将非结构化数据转化为结构化数据；由于经过数据结构化模块后，数据呈现结构化的形式，所以数据的存储采用传统的关系型数据库；数据可视化模块将用户的签到数据以热力图的形式显示在地图上，为数据建模进一步作出决策提供参考；数据建模模块首先将第二章节的多视角时空上下文共现的方法进行模块化，然后将第三章节的提出的特征融合和决策融合的方法模块化；框架输出用户对的关系强度，用户的关系强度值以降序的顺序排序，关系强度的值越高，两个用户之间存在关系

的可能性就越高。框架还给出了ROC曲线、PR曲线、LIFT曲线三种曲线对模型的性能进行全面的评估。同时，模型还推荐出最有可能存在关系的前TOP对用户。

4.2 数据存储与管理

4.2.1 数据结构化

数据结构化模块主要是将非结构化数据进行结构化处理。该模块主要包括数据清理、数据变换和数据简化[46]。数据清理主要是去除签到数据中的噪声数据和无关数据，处理遗漏数据和清洗脏数据，去除空白数据域和知识背景上的白噪声。其次数据变换主要是找到数据的特征表示。最后再经过数据简化，有些数据属性对发现任务是没有影响的，这些属性的加入会大大影响挖掘效率，甚至还可能导致挖掘结果的偏差。

在本文的框架中，主要包括对用户签到数据进行结构化处理，将用户签到数据处理成$< user, timestamp, location >$的形式，其中$location$主要以经纬度的形式表示；同时该模块生成平衡的正负样例，将用户间的关系数据处理成$< user01, user02, label >$，其中$label$是布尔型变量，它标记两用户间是否存在关系。

4.2.2 数据的存储

由于数据经过数据预处理模块后已经转化为结构化数据。对结构化数据的存储本文采用传统的数据库MySQL[47,48]。用户的签到数据是以一条签到数据为主键，即通过$< user, timestamp, location >$来标识一条记录。用户间的关系对数据是以两个用户标识为主键，即通过$< user01, user02 >$来标识一条记录。用户签到数据和用户间关系对数据通过用户标识进行关联，一个用户标识对应多条签到数据，或者与多个用户存在关系。我们为用户的签到数据可提供增加、删除、查询权限，在这里，签到数据具有时效性，比如用户几年前的签到数据不足以反映用户现在的社交关系，

所以删除功能是系统数据库根据时效性统一删除。本文数据的存储模块未提供修改的功能，因为用户的签到数据通过移动设备采集或者通过Web签到传递到后台，反映的是用户的真实数据，所以不提供随意篡改的功能。

4.2.3　数据可视化

数据可视化一般指将结构或非结构数据转化成适当的可视化图表，可视化能将数据以更加直观的方式展现出来，使数据更加客观、更具说服力[49]。签到数据具有稠密不均的特性，本文采用热力图的形式呈现用户签到信息。通过热力图，我们可以清晰地发现签到点的聚集范围、聚集程度、颜色越深，表示签到点越密集并且随着地图缩放，热力图形状、颜色都会相应发生变化。

图4-2呈现Brightkite数据集的签到分布。Brightkite社交网络数据集允许用户通过移动终端登录后利用GPS定位当前所在地进行签到（Check-in），用户也可以使用电脑登录，自行输入地点后搜寻确定。Brightkite数据集相比Gowalla数据集具备更强的灵活性。

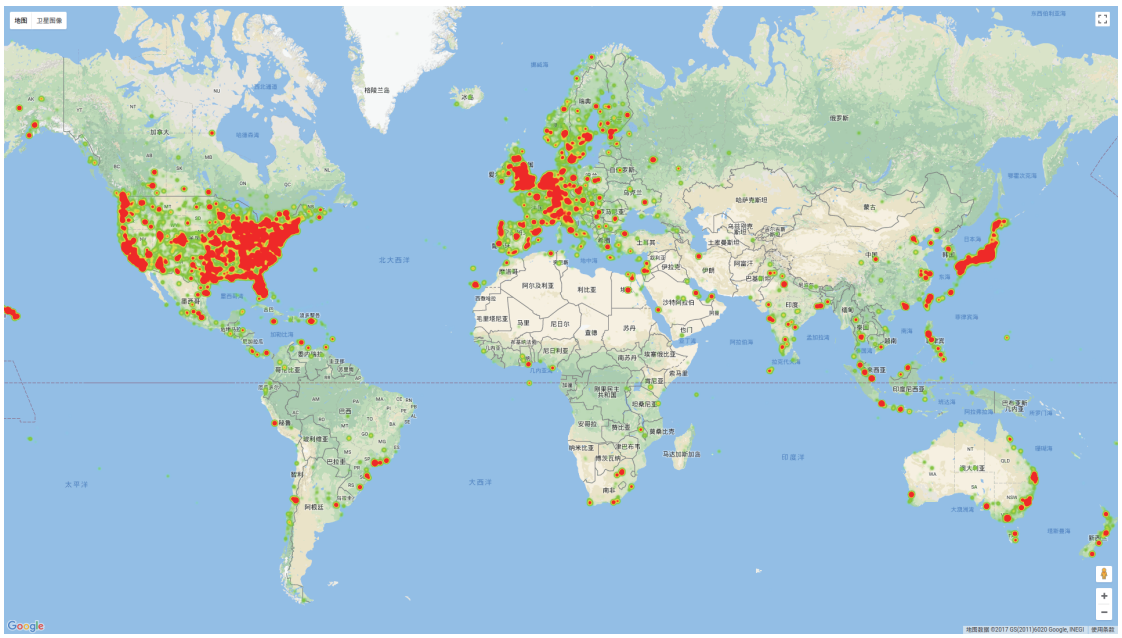


图 4-2　Brightkite数据集的可视化

图4-3呈现Gowalla数据集的签到分布。Gowalla数据集在欧洲的分布相比Brightkite数

据集而言更加密集。

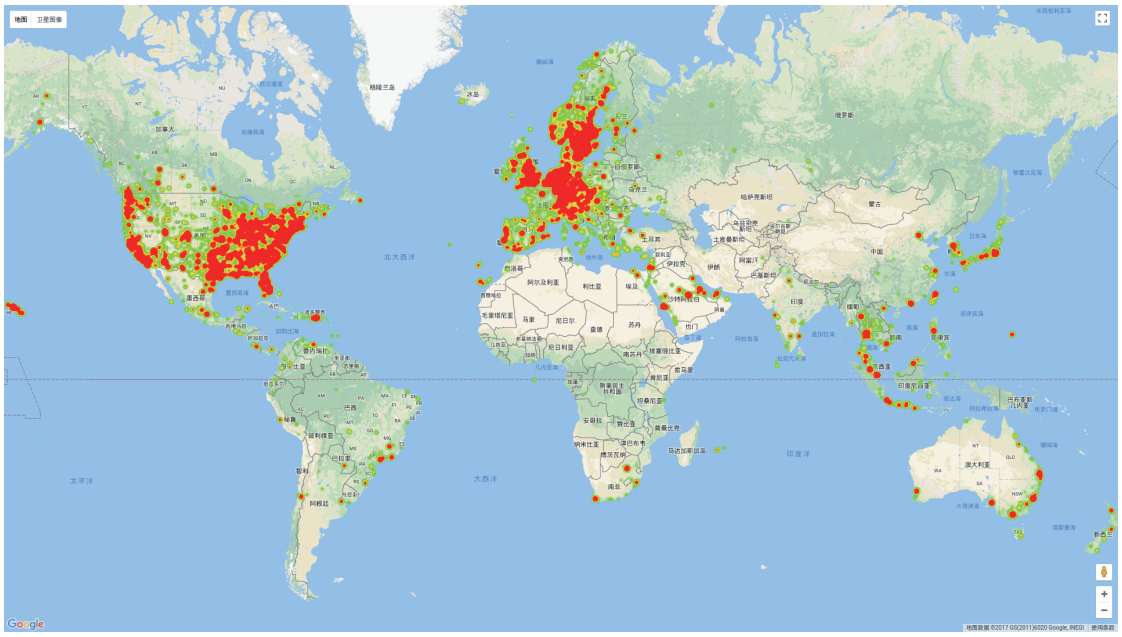


图 4-3　Gowalla数据集的可视化

4.3　数据建模

数据建模主要是对第二章节和第三章节的内容进行模块化。首先对签到数据进行多视角特征工程，多视角主要分为时间-空间、空间-时间两个主视角，然后特征工程主要包括上下文序列的构建和上下文共现特征的生成，利用上下文共现特征结合机器学习技术基于多视角进行用户关系强度得分预测。

根据上下文共现特征的互补性进行特征融合，提出的基于时空上下文共现特征融合的方法（FF）进行关系强度的预测，同时输出用户关系强度可能性最大的前Top对用户。同时根据多视角的预测得分提出基于多视角决策融合的方法（DF），同时输出用户关系强度可能性最大的前TOP对用户。

4.4　用户关系强度预测的应用

本节首先给出关系强度预测问题的模型性能评估曲线，然后给出结果的展示与分析，与此同时还推荐出最有可能存在关系强度的前TOP对用户。

4.4.1 应用性能评估指标

本文的框架的模型的评估主要采用三种曲线来进行评估，分别是ROC曲线、PR曲线和LIFT曲线，表4-1给出了这三种类型曲线的概要说明。

表 4-1 评估指标名词说明

名词	解释
TP	True Positive，模型把正样本预测为正样本的数目
TN	True Negative，模型把负样本预测为负样本的数目
FP	False Positive，模型把负样本预测为正样本的数目
FN	False Negative，模型把正样本预测为负样本的数目
TPR	True Positive Rate，TP/(TP+FN) 即模型把正样本预测为正样本占所有正样本的比例
FPR	False Positive Rate，FP/(FP+TN) 即模型把负样本预测为正样本占所有负样本的比例
ROC曲线	把FPR作为横坐标，TPR作为纵坐标，从0-1改变分类阈值得到的曲线
Precision	TP/(TP+FP)，模型把正样本预测为正样本占所有预测为正的样本的比例
Recall	TP/(TP+FN)，模型把正样本预测为正样本占所有正样本的比例
PR曲线	Precision作为横坐标，Recall作为纵坐标
Depth	Positive Rate，即预测为正样本的比例，等于(TP+FP)/(TP+FN+FP+TN)
Lift	衡量使用模型与不使用模型相比，模型预测正样本能力提升多少倍
Lift曲线	Lift作为纵坐标，Depth作为横坐标； 从0-1改变阈值得到曲线；ROC曲线关注的是覆盖率，Lift曲线关注的是命中率

ROC（**R**eceiver **O**perating **C**haracteristic）曲线是首先是在二战中的电子工程师和雷达工程师发明的，用来侦测战场上的敌军载具（飞机、船舰），也就是信息检测理论。后来被引入机器学习领域[50]，近几年被应用于模型性能评估[51]以及其他领域，最近越来越多地应用于数据挖掘（Data Mining）领域。对于ROC曲线，我们根据学习器的预测结果对样例进行排序，按此顺序逐个把样本作为正例进行预测，每次计算出两个重要的值即FPR、TPR的值。然后分别以它们为横、纵坐标作图，ROC曲线以FPR（**F**alse **P**ositive **R**ate）为横坐标、TPR（**T**rue **P**ositive **R**ate）为纵坐标，最终便得到了ROC曲线。

PR曲线是另一个常用的指标[52]。Precision和Recall是一对矛盾的度量，一般来说，Precision高时，Recall往往偏低；而Recall高时，Precision往往偏低。本文对学习器的预测结果进行排序，划分不同的阈值，根据不同的阈值分别计算出Precision和Recall，然后以Precision为横坐标、Recall为纵坐标作图，便可得到PR曲

线。PR曲线是不同阈值之间Precision和Recall之间的权衡，与ROC曲线相同，曲线下的面积越大模型的性能就越好。当数据类型不平衡时，PR曲线相比ROC曲线而言有更好的稳定性。

Lift是一种目标模型的预测和分类能力性能的度量方法。该度量方法衡量的是使用模型与不使用模型（即随机分类）相比，模型的预测正样本准确率的能力提升了多少倍。使用模型预测正样本的准确率通过Precision来衡量，其中Precision=TP/(TP+FP)；不使用模型预测正样本的准确率通过用正例的比例来估算，即准确率为(TP+FN)/(TP+FP+TN+FP)。Lift的值为使用模型与不使用模型的比值。显而易见，二者的比值越大，说明模型的预测能力越好。其中Depth为模型预测为正样本的比例，等于(TP+FP)/(TP+FP+TN+FP)，故Depth的取值为[0,1]。Lift曲线是以Depth为横坐标，Lift为纵坐标，Depth从0至1之间改变分类阈值得到曲线。Lift曲线也可以被认为是ROC曲线的一个变种，不过ROC曲线关注的是覆盖率，Lift曲线关注的是命中率[53]。

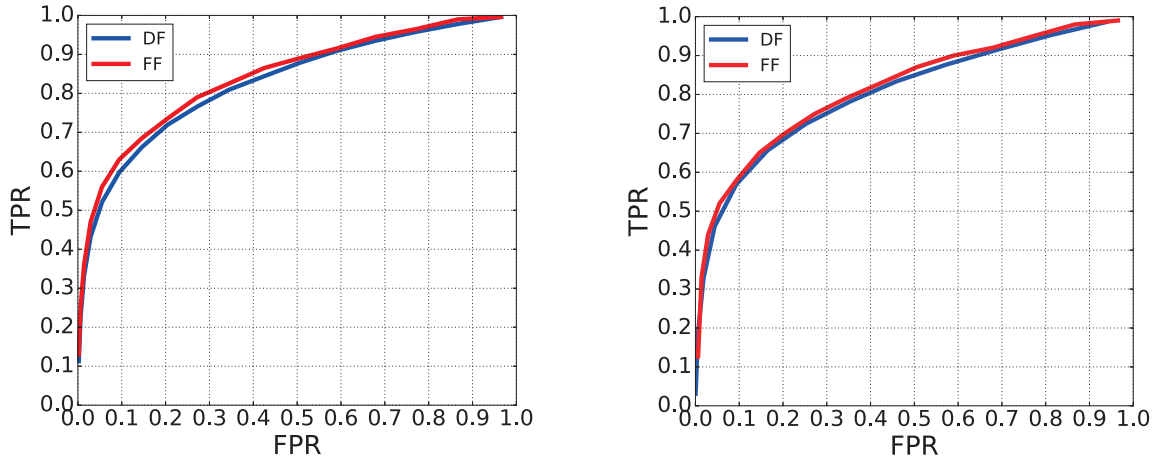
本文框架给出平台输出的三种类型曲线，这三种类型的曲线为模型的性能提供不同方面的评估，方便用户判断模型的优劣。与此同时，本文针对应用还推荐出关系强度最强的前TOP对用户。该框架根据学习器的预测结果对样例进行排序，排在前面的是平台认为“最可能”是正例的样本，排在最后的则是平台认为“最不可能”是正例的样本，按此顺序逐个把样本作为正例进行预测，则每次可以计算出当前的查准率（Precision）和查全率（Recall）。该框架主要输出相应的TOP对用户以及其所对应的Precision、Recall。前TOP对用户很有可能是强关系，企业可以利用这些关系强度值进一步理解用户的需求，法院部门可以通过已知的部分犯罪份子找出其它犯罪团伙等等。

4.4.2　结果展示与分析

结果1： 图4-4和图4-5给出了特征融合方法（FF）和决策融合方法（DF）的ROC曲线和PR曲线比较。在Brightkite和Gowalla数据集上，FF方法均超过了DF方

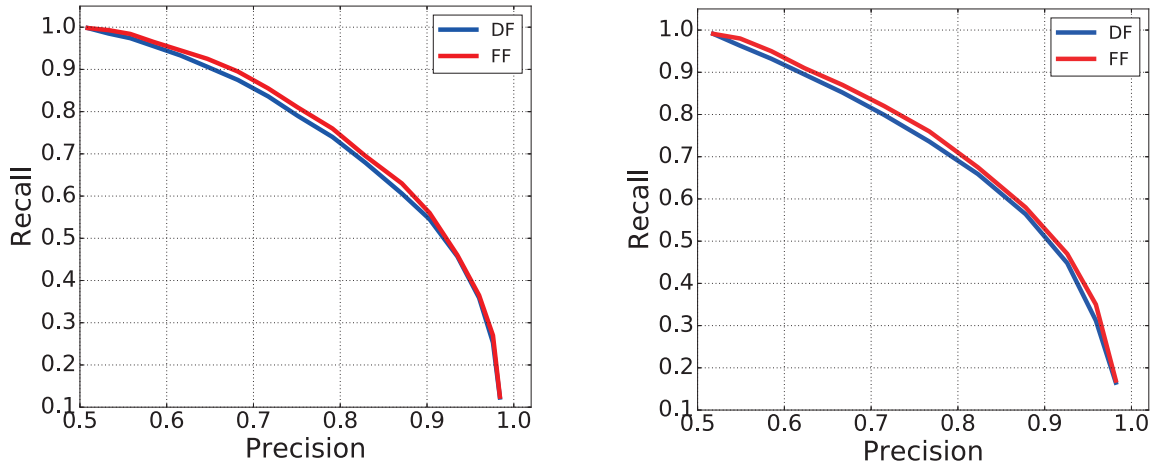
法的效果。

分析1： 在当前两个数据集上，FF方法均超过了DF方法，主要是因为数据的稀疏性导致以小时和分钟为粒度的视角性较差，但DF方法可以对稠密的签到数据有很好鲁棒性。对当前的两个数据集，FF方法很好地利用了两个视角特征的互补性。所以本文在当前测试集上选用FF方法作为用户关系强度的预测。



(a) Brightkite ROC曲线　　(b) Gowalla ROC曲线

图 4-4　DF与FF融合方法的ROC曲线



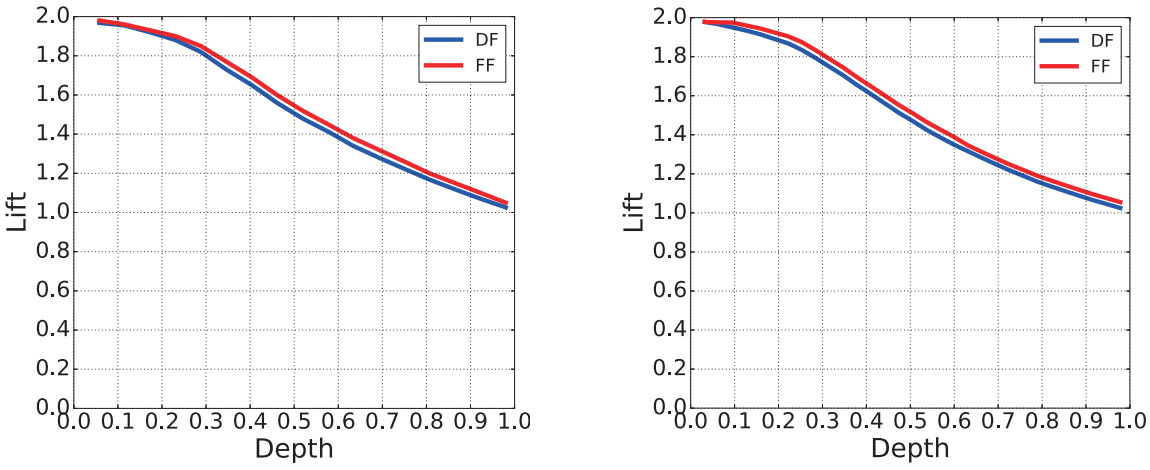
(a) Brightkite PR曲线　　(b) Gowalla PR曲线

图 4-5　DF与FF融合方法的PR曲线

结果2： 图4-6给出了DF方法与FF方法LIFT曲线的比较，同样，FF方法的比DF方法要好。由图可知，当Depth<0.3时，LIFT曲线下降速度缓慢；而当Depth>0.3，

LIFT曲线下降速度变快。

分析2： 本文通过分析LIFT曲线的斜率，建议采用Depth<0.3的预测结果，因为Depth<0.3之前的模型的预测精度相比随机猜测的精度要高。



(a) Brightkite Lift曲线　　(b) Gowalla Lift曲线

图 4-6　DF与FF融合方法的LIFT曲线

结果3： 表4-2和表4-3分别显示了选用FF 方法在Brightkite数据集和Gowalla数据集推荐出的前TOP对用户。根据LIFT曲线当Depth<0.3时，FF方法在Brightkite数据集上给出了前7000个用户对，在Gowalla数据集上给出了前24000条用户对，这些用户对的关系强度按降序排序。

表 4-2　Brightkite数据集前TOP对用户推荐

TOP N	Precision	Recall	Depth
1000	0.979	0.123	0.058
2000	0.973	0.234	0.115
3000	0.959	0.342	0.173
4000	0.937	0.442	0.231
5000	0.904	0.531	0.288
6000	0.864	0.607	0.346

表 4-3　Gowalla数据集前TOP对用户推荐

TOP N	Precision	Recall	Depth
4000	0.978	0.123	0.063
8000	0.962	0.243	0.126
12000	0.941	0.356	0.189
16000	0.913	0.461	0.252
20000	0.870	0.549	0.315
24000	0.823	0.623	0.378

分析3： 表中同时给出取前TOP对用户时的Precision和Recall的变化，Precision会随着N的增加而减少；而Recall会随着N的增加而增大。企业可以根据正负样例的营收比来自定义选择N的取值，例如需要选出的用户对中尽可能高关系强度则尽

可能的减小N的值则可以提高Precision；再例如犯人的排查则需要增大N的值来提高Recall。

4.5 本章小结

本章首先给出了用户关系强度预测的整体框架。框架中包含数据的存储与管理、数据建模、模型评估这几个重要的模块。然后对这几个重要的模块进行详细的说明。最终对用户关系强度预测给出一种应用解决方案。

第五章　总结与展望

5.1　论文工作总结

本文基于时空数据进行用户关系强度预测，该社交网络中的用户签到信息包含时间和空间信息。前人研究发现上下文感知预测和时空共现特征有利于关系强度的预测，不过前人只考虑了上下文或者共现一种因素，并且从未考虑到时间上下文信息。本文通过妙巧利用上下文序列的方法将时空上下文和时空共现进行融合，并且同时考虑了时间上下文信息，从而精度得到提升。本文的方法挖掘上下文共现特征来更好的预测关系强度。本文的研究内容和成果主要包括以下几个方面：

(1) 本文首先提出多视角时空上下文共现的方法进行关系强度的预测。前人发现两个用户经常在相同的地点短时间内发生签到行为，则可以认为他们具有相似的日常活动，进而彼此存在关系的概率越大。基于以上的特性，我们从两个主视角分别建立上下文共现序列，通过上下文序列将时空数据中用户间的共现关系转化为NLP领域中同义单词共现的关系，本文利用word2vec工具分别基于多视角提取用户上下文共现特征，该特征表征用户的时间上下文、时间共现、空间上下文和空间共现，同时还表征用户的签到周期性行为。用该特征表征用户并结合机器学习技术来进行用户关系强度的预测。实验表明，本文方法中最好的Day-Location视角比EBM算法在Brightkite数据集上在相同的Precision下Recall最高提高10%，在Gowalla数据集上最高提高8%。

(2) 本文考虑到两个视角的上下文共现特征的互补性，空间-时间视角的上下文共现特征表征强地点共现、最短时间上下文和弱最短空间上下文，时间-空间视角的上下文共现特征表征强时间共现、最短空间上下文和弱最短时间上下文，提出了基于时空上下文共现特征融合的方法（FF）。同时，本文也给出基于决策加权融合（DF）。实验显示，FF方法相比最好的Day-Location视角在Brightkite数据集上AUC指标提升3.6%，在Gowalla数据集上提

升4.3%，FF方法相比DF方法在Brightkite数据集上AUC提升1.4%，在Gowalla数据集上提升1.6%。同时，FF方法比目前最好的方法SCI在Brightkite数据集上AUC提升6.1%，在Gowalla数据集上提升2.4%。

(3) 同时本文提出一种社交网络关系强度预测的应用框架。该方案包含以下几个模块：(1)数据的存储与管理模块：数据的结构化处理、数据的存储和数据的可视化功能；(2)数据建模：将DF方法和FF方法模块化；(3)同时我们给出模型的评估曲线，分别为ROC曲线、PR曲线和LIFT曲线，并推荐出关系强度最强的用户对。

5.2 未来工作展望

本文主要是用户关系强度预测问题提出了基于时空上下文共现的预测方法。虽然通过实验说明了本文的方法在精度上超过了EBM和SCI算法，但仍有一些问题需要进一步研究：

(1) 我们采用的Brightkite和Gowalla数据集具有稀疏性，位置信息的上下文特征不能很好的体现，创建特定稠密地点的子集可以更能体现本文所提出的上下文共现特征的重要性。

(2) 本文没有充分挖掘该地点是公共场所还是私人场所。我们认为如果用户在私人场所出现，即使出现一次，该场所对用户之间的关系强度的影响也是非常大的。

(3) 目前本文研究的方法主要都是针对离线预测问题，模型的预测的时效性还存在一定的不足，并且模型不能进行增量更新。因此接下来需要从时效性和可增量更新这两个角度去考虑做进一步的研究。

参考文献

[1] 刘大有, 陈慧灵, 齐红and 杨博. 时空数据挖掘研究进展[J]. 计算机研究与发展, 2013, 50(2): 225–239.

[2] 史岭峰. 基于社交网络好友关系的图查询算法研究与应用[D]. 南京: 南京理工大学, 2012.

[3] Daisuke Takagi, Yoshikazu Homma, Hiroki Hibino, Satoru Suzuki and Yoshihiro Kobayashi. Single-walled carbon nanotube growth from highly activated metal nanoparticles [J]. Nano letters, 2006, 6(12): 2642–2645.

[4] E Rutger Leukfeldt. Cybercrime and social ties [J]. Trends in organized crime, 2014, 17(4): 231–249.

[5] Nasrullah Memon and Henrik Legind Larsen. Practical algorithms for destabilizing terrorist networks [A]. International Conference on Intelligence and Security Informatics [C]. Berlin, Heidelberg: Springer, 2006: 389–400.

[6] Ashwin Machanavajjhala, Aleksandra Korolova and Atish Das Sarma. Personalized social recommendations: accurate or private [J]. Proceedings of the VLDB Endowment, 2011, 4(7): 440–450.

[7] Zan Huang, Xin Li and Hsinchun Chen. Link prediction approach to collaborative filtering [A]. Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries [C]. Denver, CO, USA: ACM, 2005: 141–142.

[8] David J Crandall, Lars Backstrom, Dan Cosley, Siddharth Suri, Daniel Huttenlocher and Jon Kleinberg. Inferring social ties from geographic coincidences [J]. Proceedings of the National Academy of Sciences, 2010, 107(52): 22436–22441.

[9] Huy Pham, Cyrus Shahabi and Yan Liu. Ebm: an entropy-based model to infer social strength from spatiotemporal data [A]. Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data [C]. New York, USA: ACM, 2013: 265–276.

[10] Ningnan Zhou, Xiao Zhang and Shan Wang. Theme-aware social strength inference from spatiotemporal data [A]. International Conference on Web-Age Information Management [C]. Cham: Springer, 2014: 498–509.

[11] Gunarto Sindoro Njoo, Min-Chia Kao, Kuo-Wei Hsu and Wen-Chih Peng. Exploring check-in data to infer social ties in location based social networks [A]. Pacific-Asia Conference on Knowledge Discovery and Data Mining [C]. Cham: Springer, 2017: 460–471.

[12] Hakan Bagci and Pinar Karagoz. Context-aware friend recommendation for location based social networks using random walk [A]. Proceedings of the 25th international conference companion on world wide web [C]. Montr é al, Qu é bec, Canada: International World Wide Web Conferences Steering Committee, 2016: 531–536.

[13] Josh Jia-Ching Ying, Eric Hsueh-Chan Lu, Wang-Chien Lee, Tz-Chiao Weng and Vincent S Tseng. Mining user similarity from semantic trajectories [A]. Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks [C]. San Jose, California: ACM, 2010: 19–26.

[14] Xiangye Xiao, Yu Zheng, Qiong Luo and Xing Xie. Finding similar users using category-based location history [A]. Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems [C]. San Jose, California: ACM, 2010: 442–445.

[15] Xihui Chen, Jun Pang and Ran Xue. Constructing and comparing user mobility profiles for location-based services [A]. Proceedings of the 28th Annual ACM Symposium on Applied Computing [C]. Coimbra, Portugal: ACM, 2013: 261–266.

[16] Xihui Chen, Piotr Kordy, Ruipeng Lu and Jun Pang. MinUS: Mining user similarity with trajectory patterns [A]. Joint European Conference on Machine Learning and Knowledge Discovery in Databases [C]. Berlin, Heidelberg: Springer, 2014: 436–439.

[17] Xihui Chen, Ruipeng Lu, Xiaoxing Ma and Jun Pang. Measuring user similarity with trajectory patterns: Principles and new metrics [A]. Asia-Pacific Web Conference [C]. Cham: Springer, 2014: 437–448.

[18] Justin Cranshaw, Eran Toch, Jason Hong, Aniket Kittur and Norman Sadeh. Bridging the gap between physical location and online social networks [A]. Proceedings of the 12th ACM international conference on Ubiquitous computing [C]. Copenhagen, Denmark: ACM, 2010: 119–128.

[19] Eunjoon Cho, Seth A Myers and Jure Leskovec. Friendship and mobility: user movement in location-based social networks [A]. Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining [C]. San Diego, California, USA: ACM, 2011: 1082–1090.

[20] Joseph Turian, Lev Ratinov and Yoshua Bengio. Word representations: a simple and general method for semi-supervised learning [A]. Proceedings of the 48th annual meeting of the association for computational linguistics [C]. Uppsala, Sweden: Association for Computational Linguistics, 2010: 384–394.

[21] 来斯惟et al. 基于神经网络的词和文档语义向量表示方法研究[J]. 2016.

[22] Yoshua Bengio, Réjean Ducharme, Pascal Vincent and Christian Jauvin. A neural probabilistic language model [J]. Journal of machine learning research, 2003, 3(Feb): 1137–1155.

[23] Tomas Mikolov, Quoc V Le and Ilya Sutskever. Exploiting similarities among languages for machine translation [J]. arXiv preprint arXiv:1309.4168, 2013.

[24] Tomas Mikolov, Kai Chen, Greg Corrado and Jeffrey Dean. Efficient estimation of word representations in vector space [J]. arXiv preprint arXiv:1301.3781, 2013.

[25] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado and Jeff Dean. Distributed representations of words and phrases and their compositionality [A]. Advances in neural information processing systems [C]. . 2013: 3111–3119.

[26] Jan Van Leeuwen. On the Construction of Huffman Trees. [A]. ICALP [C]. . 1976: 382–410.

[27] Frederic Morin and Yoshua Bengio. Hierarchical Probabilistic Neural Network Language Model [A]. Aistats [C]. Citeseer, 2005: 246–252.

[28] David E Rumelhart, Geoffrey E Hinton and Ronald J Williams. Learning representations by back-propagating errors [J]. nature, 1986, 323(6088): 533.

[29] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system [A]. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining [C]. San Francisco, California, USA: ACM, 2016: 785–794.

[30] Zhi-Hua Zhou. Ensemble methods: foundations and algorithms[M]. CRC press, 2012.

[31] Leo Breiman, Jerome Friedman, Charles J Stone and Richard A Olshen. Classification and regression trees[M]. CRC press, 1984.

[32] Tom M Mitchell. 机器学习[M]. 机械工业出版社, 2003.

[33] Jerome Friedman, Trevor Hastie, Robert Tibshirani et al. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors) [J]. The annals of statistics, 2000, 28(2): 337–407.

[34] Xu Caixu, Yan Jianfeng, Yang Lu, Xu Guanggen and Shi Hongbin. Context Co-occurrence Based Relationship Strength Prediction in Spatiotemporal Data [A]. International Conference on Computer and Mechatronics [C]. . 2018.

[35] Jure Leskovec and Andrej Krevl. {SNAP Datasets}:{Stanford} Large Network Dataset Collection [J]. 2015.

[36] Haibo He and Edwardo A Garcia. Learning from imbalanced data [J]. IEEE Transactions on knowledge and data engineering, 2009, 21(9): 1263–1284.

[37] Xin Rong. word2vec parameter learning explained [J]. arXiv preprint arXiv:1411.2738, 2014.

[38] 石鸿斌, 严建峰, 白瑞瑞, 徐彩旭and 徐广根. 多粒度时序特征在离网预测中的应用[J]. 计算机应用研究, 2018, 36(04): 1–7.

[39] 徐广根, 杨璐, 严建峰, 徐彩旭and 石鸿斌. 基于LDA主题模型的用户电信轨迹恢复算法[J]. 计算机应用研究, 2018, 36(08): 1–8.

[40] Federico Castanedo. A review of data fusion techniques [J]. The Scientific World Journal, 2013, 2013.

[41] Ren C Luo, Chih-Chen Yih and Kuo Lan Su. Multisensor fusion and integration: approaches, applications, and future research directions [J]. IEEE Sensors journal, 2002, 2(2): 107–119.

[42] Thomas E Fortmann, Yaakov Bar-Shalom and Molly Scheffe. Multi-target tracking using joint probabilistic data association [A]. Decision and Control including the Symposium on Adaptive Processes, 1980 19th IEEE Conference on [C]. Albuquerque, NM, USA: IEEE, 1980: 807–812.

[43] Éloi Bossé, Pierre Valin, Anne-Claire Boury-Brisset and Dominic Grenier. Exploitation of a priori knowledge for information fusion [J]. Information Fusion, 2006, 7(2): 161–175.

[44] Marleen Morbee, Linda Tessens, Hamid Aghajan and Wilfried Philips. Dempster-Shafer based multi-view occupancy maps [J]. Electronics letters, 2010, 46(5): 341–343.

[45] Hongjian Wang, Zhenhui Li and Wang-Chien Lee. PGT: Measuring mobility relationship using personal, global and temporal factors [A]. Data Mining (ICDM), 2014 IEEE International Conference on [C]. Shenzhen, China: IEEE, 2014: 570–579.

[46] 刘明吉and 王秀峰. 数据挖掘中的数据预处理[J]. 计算机科学, 2000, 27(4): 54–57.

[47] AB MySQL. MySQL database server [J]. Internet WWW page, at URL: http://www.mysql. com (last accessed/1/00), 2004.

[48] AB MySQL. MySQL: the world's most popular open source database [J]. http://www.mysql. com/, 2005.

[49] 李田丁and 王莉. 浅谈大数据时代的数据挖掘和数据可视化[J]. 图书情报导刊, 2016, (1): 100–101.

[50] Kent A Spackman. Signal detection theory: Valuable tools for evaluating inductive learning [A]. Proceedings of the sixth international workshop on Machine learning [C]. Elsevier, 1989: 160–163.

[51] DJ Peres, C Iuppa, L Cavallaro, A Cancelliere and E Foti. Significant wave height record extension by neural networks and reanalysis wind data [J]. Ocean Modelling, 2015, 94: 128–140.

[52] Jesse Davis and Mark Goadrich. The relationship between Precision-Recall and ROC curves [A]. Proceedings of the 23rd international conference on Machine learning [C]. Pittsburgh, Pennsylvania, USA: ACM, 2006: 233–240.

[53] Miha Vuk and Tomaz Curk. ROC curve, lift chart and calibration plot [J]. Metodoloski zvezki, 2006, 3(1): 89.

发表文章目录及科研项目

1.发表论文情况：

[1] **Caixu Xu**, JianFeng Yan, Lu Yang, Guanggen Xu and Hongbin Shi. Context Co-occurrence Based Relationship Strength Prediction in Spatiotemporal Data [C]. Proceedings of the 2018 International Conference on Computer Modeling, Simulation and Algorithm. CMSA 2018. （EI Indexed，已发表）。

[2] **Caixu Xu** and Ruirui Bai. Inferring Social Ties from Multi-View Spatiotemporal Co-occurrence [C]. APWeb-WAIM 2018 Data Science Workshop. (CCF C类，已发表)。

[3] 石鸿斌，严建峰，白瑞瑞，**徐彩旭**，徐广根. 多粒度时序特征在离网预测中的应用[J]. 计算机应用研究.（已发表）。

[4] 徐广根，杨璐，严建峰，**徐彩旭**，石鸿斌. 基于LDA主题模型的用户电信轨迹恢复算法[J/OL]. 计算机应用研究.（已发表）。

2.专利成果：

[1] **徐彩旭**，徐广根，石鸿斌。基于位置社交网络的用户关系强度预测方法、装置及设备。申请号：201711422233.0

3.软件著作权：

[1] **徐彩旭**，严建峰。基于时空数据的用户社交关系预测软件。登记号：2016SR307760

[2] 蒋贤进，**徐彩旭**，刘纯平。轨迹相似度分析与预测系统。登记号：2018SR102432

4.学科竞赛：

[1] **徐彩旭**，蒋贤进，卢奇。第五届“中国软件杯”大学生软件设计大赛《针对以经纬度或经纬度带时间定义的不同轨迹》决赛三等奖。(国家级决赛三等奖)

5.实习：

2017/6-2017/10 百度-北京中关村-大数据应用（已获留用offer）。参与百度黑名单系统的构建：编写分布式爬虫增量抓取Web失信数据；利用Hadoop对数据进行统

计、处理和管理；利用MySql进行结构化数据的存储管理；黑名单模型的调测，并结合外部数据提高黑名单模型Precision、Recall、Lift、AUC和PR-AUC的评估。

6.参与科研项目：

[1] 参与“融合文本网数据的深度学习技术研究”，国家自然科学基金（61572339）

[2] 参与“基于超图的主题建模算法研究”，国家自然科学基金（61373092）

[3] 参与“基于时空相关性的无线传感器网络节能问题研究”，国家自然科学基金（61272449）

[4] 参与“基于二型模糊逻辑的多核程序数据竞争与死锁检测方法研究”，国家自然科学基金（61202029）

[5] 参与“基于大规模数据中心的面向大数据应用的云管理平台-大规模机器学习技术”，江苏省科技支撑计划（BE2014005-4）

致　谢

三年的研究生生涯弹指一挥间就要过去了。刚进实验室时，师兄师姐们准备毕业的情景还历历在目，如今自己也要毕业，踏上社会了。在这三年中，我的收获有很多，不仅仅是在学术方面，更是在待人处世方面。在这里我想要对所有帮助过我的老师，同学和亲人表示衷心的感谢!

首先，我要感谢我的导师严建峰老师。严老师对我们的要求非常的高，做学术的过程中容不得半点沙子。严老师每周让我们给他汇报自己的研究进展，他会有针对的提出指导建议。严老师一直坚持理论结合实际的观点，在很早就创造机会，让我进苏州一家大数据创业公司做数据挖掘相关的项目。有了实际的项目的支撑，让我学习起来更有目的性，这比只是看书本上的知识学更加的具体。而且，项目的压力也可以督促我严格要求自己，锻炼了自己的动手能力和团队协作能力。

然后，我想感激我的师兄师姐们。虽然你们不在实验室，但还一直牵挂着我。每当遇到什么不懂的问题，你们总会很热情地帮我解答。和你们讨论，具有创新性的想法像泉水一样涌出来，收获颇丰，讨论过程在轻松高效的状态下结束。谢谢你们，有了你们，我的三年更加多彩。

其次，我想感谢师弟师妹们。我们经常就某一个具体的问题展开讨论。我们不仅是学术上的伙伴，还是生活上的朋友，时常会一起出去聚餐，聊聊各自在的进展。硕士求学生涯即将结束，感激你们的陪伴。愿你们一切安好，珍重再见。

接着，我要感谢我的父母，感谢你们对我的关心和支持。在生活上处处为我着想，一次次给我送来生活必需品，让我无后顾之忧，专心进行科研。在我离校实习时，也经常打电话来嘘寒问暖。

最后，感谢各位老师在百忙中对我的论文的进行评审。

徐彩旭

二〇一八年二月二十三日