武汉邮电科学研究院硕士学位论文

开题报告

专 业： 信号与信息处理

研究方向： 数据挖掘

论文题目： 基于信息熵与信任机制的

地点个性化推荐研究

研 究 生： 张成 学号： 20160053

导 师： 王峥

报告日期： 2017-11-28

|  |
| --- |
| 1. 课题的来源、目的及意义 |
| 近几年，互联网购物兴起和一些B2C网站的发展，人们已经习惯于从互联网中获得自己需要的物品。但同时也提出了一个非常困难和急切需要解决的问题，即如何从爆炸性的海量数据中提取出我们需要的数据。在海量数据的情况下，仅仅使用搜索引擎会出现数据量过大还需要人为的筛选的情况，而且这种技术也显得有些单调和过时。  推荐系统正是在这样的信息过载情况下出现的，它能够根据用户的喜好，推荐给用户所关心和需要的信息，给用户提供个性化的服务，从而让用户有更好的体验。基于位置的社交网络（Location Based Social Networks，LBSN ）能够依据信息需求者当前位置，而实时地进行个性化信息推荐，在LBSN中，一个重要的功能就是地点推荐--帮助用户发现其可能感兴趣的地点。并且基于位置的社交网络己经在社交功能上附加了游戏、旅游、O2O（Online to Office）、支付等高附加值的业务，LBSN庞大的用户基数为LBSN服务的开展提供了巨大的潜在商业价值，而个性化地点推荐在丰富用户体验、提高用户忠诚度、付费用户转化等方面具有不可替代的地位。  但相比传统的推荐技术，地点个性化推荐面临着一些新的挑战，其中最为关键的就是用户-地点签到矩阵的稀疏性和用户相似性的计算问题。研究学者围绕这两个问题进行了大量的研究，但很少有学者从用户行为时间段相似性来来缓解稀疏性和用“熵理论”来计算用户相似性，且在进行地点推荐时大多只利用了好友信息，对非好友信息未能充分利用。  因此，本课题在用户的签到信息和好友信息基础上，采用用户行为时间段相似性来缓解矩阵稀疏性的问题，并基于信息熵理论来计算用户相似性和结合友信任机制来进行地点个性化推荐。 |
| 1. 本课题的国内外研究现状及发展趋势分析   （在文献调研的基础上完成此部分内容，并列出至少10篇相关文献。） |
| 位置服务（Location-Based Service，LBS）与社交网络逐渐融合形成了基于地理位置的社交网络（Location-Based Social Networks，LBSN）。相比传统在线社交网络，基于位置的社交网络通常具有“地点签到”（Check-in）功能，同时用户可以对他们所在的地点发表评论、图片和视屏等。因此，基于位置的社交网络上包含了众多的用户信息、时空信息、好友关系、评论内容、标签信息等，但大量数据也加剧了“信息过载”[1]问题，为了解决用户面临的信息选择问题，合理的地点推荐成为LBSN研究的重要方向之一。  最初，Ye等在文献[2]中将地点推荐引入到LBSN中，他们认为朋友关系会对用户的地点选择有很大的影响，因此，提出了一种基于朋友的协同过滤方法；随后，Ye等在文献[3]中又分析了地理位置对用户地点访问行为的影响，他们假设用户更喜欢访问距离近的地点，从而建立幂分布模型来模拟地理距离与访问可能性大小的关系；除了幂分布外，Cho等[4]和Cheng等[5]通过多中心髙斯分布来模拟地理位置的影响。  也有学者将用户的评论信息融入地点推荐中，如用情感分析的方法[6]，对用户评论进行分析，从中可以得到用户对地点的情感分值，然后结合用户在地点的签到数据，构建了新的用户－地点评分矩阵，并在新矩阵的基础上通过协同过滤方法得到地点评分预测值；此外，时间信息也吸引了学者们的关注，孙光福等[7]提出一种基于时序消费行为的最近邻建模方法，通过用户（产品）的相互影响关系，产生更为精准的推荐结果。  同时也有学者们开始探索利用其他的用户信息在LBSN系统中进行地点推荐。谢怡等[8]以位置服务推送作为出发点，结合用户兴趣建模，解决移动推荐系统中用户情景需求的问题，提高移动个性化推荐系统的推荐精度、效率和质量；周而重、黄佳进等[9]通过社交网站上用户与其好友之间、用户与签到地点以及地点与地点之间的关联，从用户的网络签到行为中总结出用户的出行特点，融合用户对地点的个人偏好程度、地点自身属性对用户的影响程度以及用户好友对地点的推荐程度，来筛选出候选地点中满足用户个性化需求的地点。  另外，还有一些学者们在传统的推荐方法基础上进行改进并应用于地点推荐中，如Berjani等[10]提出一种正则化的矩阵分解（Regularized Matrix Factorization,RMF）方法，他们认为LBSN上地点推荐最主要的问题是缺少明确的用户对地点的评分，而只是单纯将用户签到数据作为地点评分，因此他们将用户－地点签到矩阵通过一定的映射规则转化到一个隐式因素空间中，并采用正则化奇异值分解（Singular Value Decomposition,SVD）模型对地点进行评分预测；Leung等[11]认为，由于用户－地点的签到矩阵十分巨大，想要从众多的地点中寻找到相似的地点是一项很大的挑战，他们提出将用户、地点进行划分聚类，得到相似用户类在相似地点类上的新的签到矩阵，然后通过传统的协同过滤方法进行地点推荐。  总体来说，目前的研究主要是基于用户－地点签到矩阵，并结合其他一些影响因素，如社交关系、地理位置、时间因素以及签到相关的内容信息，如评论、地点标签等等。虽然现有的研究取得了一定的成果，但是还存在着不足之处；  （１）在对时间因素的研究中，虽然有学者将时间因素加入到推荐的模型[12]中，也有学者研究了用户行为的时间模式，如时序性PS，不一致性PG等，但目前对时间因素的研究中缺少探究用户行为在不同时间段之间的隐含关系，如用户在中午１２点左右与下午１８点左右都会访问餐厅类地点，也就是说，用户在不同时间段之间的地点访问行为具有一定的相似关系，而现有的研究往往都忽略了这一点；  （2）目前很多地点推荐使用的是协同过滤的方法，基于用户－地点的签到矩阵，来寻找相似用户或相似地点，从而产生推荐。但是由于地点数据量巨大，而且许多地点受访问的次数很少，所以，用户－地点的签到矩阵非常稀疏，而基于这种稀疏性数据计算相似用户或相似地点较为困难，这也是造成协同过滤方法准确率有待提高的一个重要原因，因此需要解决数据稀疏性问题，或通过其他途径计算相似度。  （3）对好友关系信息的利用方面，大多只针对好友的地点-签到信息对目标用户进行地点推荐，而忽略了非好友的地点-签到信息对目标用户的推荐作用。  【1】任磊.推荐系统关键技术研究[D].华东师范大学.2012  【2】Ye M,YinP,LeeWC.Location recommendation for location-based social network[C]//Proceeding of the 18th SIGSPATLAL Internationl Conference on Advances in Geographic Information Systems.ACM,2010:458-461  【3】Ye M,YinP,LeeWC,etal.Exploiting geographical influence for collaborative point-of-interest recommendation[C]//Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval.ACM,2011:325-334  【4】Cho E，Myers S A，LeskovesJ.Friendship and mobility:user movement in location-based social networks[C]//Proceedings of the 17th ACM SIGKDD international conference on Konwledge discovery and data mining.ACM,2011:1082-1090  【5】Cheng C,YangH,KingI,etal.Fused matrix factorization with geographical and social influence in location-based social networks[C]//Twenty-Sixth AAAI Conference on Artifical Intelliengence.2012  【6】卢思雨. 基于情感的用户建模和推荐方法研究[D]. 中国科学院大学, 2016  【7】孙光福，吴乐等.基于时序行为的协同过滤推荐算法[J]，软件学报，2013（11）：56-67  【8】谢怡.基于用户兴趣的位置服务推荐系统研究[D].南京邮电大学，2015  【9】周而重, 黄佳进, 徐欣欣. 一种基于用户网络签到行为的地点推荐方法[J]. 计算机科学, 2015, 42(10):232-234.  【10】BerjariB,Strufe T.A recommendation system for spots in location-based online social networks[C]//Proceedings of the 4th Workshop on Social Network Systems.ACM,2011:4  【11】Leung K W T，Lee D L，Lee W C.CLR:a collaborative location recommendation framework based on co-clustering[C]//Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval.ACM,2011:305-314  【12】范家兵，王鹏，周渭博等.在推荐系统中利用时间因素的方法[J].计算机应用，2015,35[5]:1324-1327 |
| 1. 课题中待解决的关键问题 |
| 1.矩阵稀疏性问题:  用户-地点评分矩阵的高稀疏性会在多个方面直接或间接影响协作过滤算法的推荐质量，形成相应的稀疏性问题。在相似性计算方面，用户或地点的相似性度量依赖于不同对象的公共评分数据，而稀疏评分矩阵将造成对象间公共评分数据数量严重不足，进而导致相似度计算过于片面且带有较大偏差。而在评分预测方面，基于用户的协作过滤算法需要使用目标用户的邻居评分对目标项目进行评分的预测，而在评分矩阵过分稀疏的情况下，很难或无法确定目标用户的邻居，导致推荐算法的推荐覆盖率降低，甚至无法实现推荐；而基于项目的协作过滤算法则需要使用目标用户的已有评分实现评分预测，由于稀疏性问题的存在，用户的历史评分数据过少，无法确定目标项目的己访问邻居项目，也会导致推荐的准确性下降甚至失败。稀疏性问题本质上是一种信息缺失的表现，在无法获得足够多用户评分数据的情况下，协作过滤算法的推荐质量难以保证，因此稀疏性问题是妨碍协作过滤推荐算法发展的主要问题，是推荐系统研究的热点核心问题。  2.用户之间相似性的度量方法：  在度量用户相似性方面，常用的计算方法主要有余弦相似性、修正的余弦相似性和相关相似性，这些计算方法都会用到用户对项目的评分矩阵，但一般LBSN中只有用户-地点签到信息，并没有评分信息，在对用户-地点签到矩阵进行赋值时难以找到合适的方法。目前常用的评分方法主要为：一是简单地将用户的地点签到行为二值化，即用户在某地点签到过，则将该用户-地点矩阵对应的元素置为1，否则置为0（该方法太过武断，没有考虑用户重复签到地点的偏爱性）；二是考虑用户的签到次数，将用户在某地点的签到次数转换成用户对该地点的评分，从而构建用户-地点评分矩阵，虽然这种转化的方式符合用户签到次数与地点的之间潜在的兴趣关系，但方法比较粗糙，难以取得较好的推荐效果。因此，挖掘出一种合适的度量用户相似性的计算方法，对提高推荐结果质量是十分有帮助的。  3.好友关系信息的应用方法：  如何设计一种算法将目标用户的好友签到信息与非好友的签到信息结合起来进行地点推荐，充分的利用好友信息与非好友信息是一个难点。 |
| 四、课题的研究内容 |
| 在推荐系统发展所面临的主要问题中，稀疏性问题是影响推荐系统预测准确性的关键问题，其对推荐系统及算法的影响是长期的，而且稀疏性问题与信息系统的用户和信息项目的规模是成正比的，规模越大的信息系统受到稀疏性问题的影响也越为严重。本文将研究的重点聚焦在推荐算法的研究上，针对稀疏性问题和用户间相似性计算问题对协作推荐算法各执行过程的影响，通过对推荐算法核心执行过程的改进和完善，以实现提高推荐算法整体预测准确性的目标。本文的主要研究内容包括：  （1）考虑时间因素的影响，分析LBSN上用户行为所具有的时间特征，挖掘不同时间段之间用户行为的相似关系，并对用户行为在不同时间段之间的相似性进行建模，进而利用时间相似性模型对用户－地点签到矩阵进行填补，缓解稀疏性。  （2）在社交网络中，用户在网络中的任何行为都会产生信息，如果信息的大小能够度量，就能有效地帮助用户筛选出对自己信息量大的信息。本文引入“信息熵”的概念来计算信息量，为用户构建签到行为的向量，将度量用户相似性的问题转化为度量该向量的余弦相似性。  （3）融合基于信息熵理论和好友信任机制进行地点推荐算法设计，使得融合后的算法能够取得更优的综合推荐效果。 |
| 五、拟采取的研究方法、实验方案、技术路线 |
| 1.时间属性分析  为了更好的验证用户行为的时间段相似性，我们首先对LBSN上的签到行为进行观察。从Foursquare网站随机抽取三个地点，并获取三个地点所有用户的签到记录，用曲线图描述用户签到行为与时间的变化关系如下：    上图中，横轴表示时间，单位为小时，1-24时分为24个时间段，即t1,t2,…t24,纵轴表示地点的签到次数。可以发现，地点L1被访问的时间段主要集中在t12和t18时间段，这类地点为餐馆、小吃街等；地点L2 被访问的时间段主要集中在t10、t11和t15时间段，这类地点主要为办公场所、图书馆和咖啡馆等；地点L3的被访问时间主要集中在t24和t1等午夜时间段，这类地点主要为夜店、KTV和酒吧等。可以看出，用户访问地点的行为具有明显的时间导向，且在相邻的时间段或特定的不同时间段内具有相似性。因此，本文提出一个时间相似性，来填补用户-签到矩阵，缓解稀疏性。  时间相似性的计算：    如上图所示，矩阵C是一个用户的地点签到矩阵，矩阵值X是用户对该地点的评分（用户在该地点的签到次数）；空白代表用户用户未曾访问过该地点。将用户-地点签到矩阵划分成24个签到子矩阵C1-C24，每个子矩阵即为当前时间段内用户的地点签到行为。计算任意两个时间段内用户的行为相似性（同一用户），采用余弦相似性作为度量时间相似性的方法，具体计算公式如下：    其中，sim(ts，tk)代表ts时间段与tk时间段之间用户行为的相似度大小，是用户u在ts时间段内访问地点L的次数。M为用户的数量，N为地点的数量。  经过余弦相似度计算之后，可以得到一个24\*24的相似度矩阵，主对角线上的元素为1，非主对角线上的元素取值在（0,1）区间，代表两个时间段之间用户行为的相似度大小。在利用时间相似性进行填补时，重点考虑与当前时间段签到行为最相似的时间段，基于相似性最高的时间段矩阵进行填补，矩阵填补公式如下：    经填补后的矩阵即为上图中的。可以看到，矩阵的稀疏性得到一定的缓解。  2.基于信息熵理论的用户相似性地点推荐：  “熵”原本是热力学中的名词，热力学中的“热熵”是表示分子状态混沌程度的物理量。香农借用“熵”的概念，用“信息熵”来描述信源的不确定度。熵值越大说明系统状态越混乱，携带的信息就越少；熵值越小说明系统状态越有序，那么携带的信息越多。有关信息熵的两个关键计算公式如下：      H(x)为信源的信息熵，该信源有n中不同的状态，h(xi)表示状态xi的不确定性，即状态xi所含的信息量。可以理解为，当某一信息描述的事件百分之百会发生，并没有减少人们对事件的任何不确定性，则该信息的信息量为零;而某一信息描述的事件不可能发生，给人带来几乎不可信的感受，则该信息带来的信息量为无穷大。  用户在地点签到的行为本身是携带信息的，如果把用户看作信息论中的信源，则用户在不同地点签到可以理解为信源的不同种状态，因此，可以从信息论的角度度量用户签到行为所产生的信息量。假设用户u一共在r个不同的地点签到，其签到地点集Lu={l1,l2,…,lr}中，用户在每一地点上出现的概率记为{Pr(l1),Pr(l2),…,Pr(lr)},则用户在每一点的熵值可以表示为：    其中h(li)表示用户在地点li(i=1,2,…,r)签到的这个状态产生的熵值；Pr（li）表示用户在地点li签到的概率。将用户看作是信息论中的信源，则不确定性即为用户本身的信息熵，表示为：    其中Hu表示用户u的信息熵，代表用户u签到的随机性。对于用户签到集Lu中的每一地点，定义地点的信息熵如下：    其中表示地点li对用户u的地点信息熵。特别的，对于用户未签到的地点j，Pr（Ij）=0,故=0，意味着该地点对用户的信息熵也为0。用户的签到行为可以表示为一个“签到信息向量”：,为后续计算表达式方便，将其简记为：，Hui表示地点li对用户u的地点信息熵。  每个地点对于不同用户来说重要程度是不一样的，地点信息熵表明该地点为用户签到偏好带来的信息量，可看作是一种兴趣权重。若两个用户对于很多地点持有相近的兴趣权重，则可认为他们对地点的兴趣是相似的。因此，通过计算用地点信息熵所表示成的签到信息向量间的相似性，可以反映用户兴趣的相似性，计算时选用经典的余弦相似性：    其中，S（u，v）为用户u和用户v的相似性；和分别表示用户u和用户v的签到信息向量；n为LBSN中地点总数；和分别表示地点li对用户u和用户v的信息熵。  选取与目标用户u相似性最大的前n个用户作为目标用户的最近邻集合N，将最近邻集合中所有签到地点中目标用户u未签到的地点作为地点集LN,对LN中每一个地点计算推荐值IR：    其中，Suv为最近邻用户v和目标用户u的相似性，Cvi为用户v在地点i的签到次数。选取地点推荐值中最大的前p个地点作为候选地点集L。考虑到地点集LN中地点可能都只被一个邻居用户签到过，且签到次数均为1，那么计算得到的地点推荐值可能一样，候选地点集L中待推荐结果不是真正的有序。因此引入好友关系对推荐值IR进行改进。  定义地点i的好友推荐权重ωi为：    其中，为目标用户u的好友数，是在地点i签到过的好友数。则地点集LN中的地点i的推荐值改进为：IR´ui=еωi×IRui。选取推荐值最大的前p个地点构成候选推荐地点集L1。  3.基于好友信任的地点推荐  分析LBSN中的好友，如果该好友是用户有目的添加的，那么该好友可能是认识的人（无论是现实生活中己经认识，还是通过LBSN中的地点认识）；或者访问地点偏好相似的人（仅仅是在LBSN中的好友，未在线下相互认识）。因此，好友的可信度可以从好友的熟悉度和好友的兴趣相似两方面来考虑。熟悉度可以间接地看作用户标记的显性信任，因为送种信任关系来源于线下的真实世界社交圈，用户之间即使没有偏好的相似性也可能产生相同的签到行为；兴趣相似性即为从用户己有的签到记录中推理出的隐性信任。  熟悉度方面，采用Konstas I 等人提出的社交影响因子，公式为：    其中，Fu和Fv分别表示用户u和用户v的集。该公式用来表明用户u和用户v的熟悉程度，即：两个用户的共同好友越多，越有可能在现实生活中认识。  兴趣相似性方面，可以从用户签到地点相似性出发，公式为：    其中，其中，为用户v在地点i的签到次数，k(v)为用户v的度，即用户的总签到次数；k(i)为地点i的度，，Suv即可表示两个用户u和v之间的相似度。  因此，好友的可信度由好友的熟悉程度和好友的相似性构成。好友可信度的计算公式为;    其中，u为可调参数，u∈[0,1]。当u=0时，表示好友信任度直接由社交影响因子得到；当u=1时表示好友可信度直接由好友间的相似性得到。  将目标用户u好友中可信度最高的前f个好友作为用户u的可信好友集，可信好友集的所有签到地点中，目标用户为签到的地点构成集合LF，对于LF中的每一个地点i，计算地点的推荐值TR：    其中，Tuv为好友v的可信度，Cvi为好友v在地点i的签到次数。TRui值最大的前p个地点作为候选推荐地点集L2.  4.融合用户相似性和好友信任机制的地点推荐算法流程：  对于目标用户u，将所有其他用户划分成好友集F和非好友集NF。对非好友集NF采用基于信息熵理论的地点推荐算法产生一个候选推荐地点集L1；对好友集采用基于好友信任机制的地点推荐算法产生一个候选推荐地点集L2；最后对L1和L2采用线性加权的方法得到最终的地点推荐列表LF，计算公式为：    其中，η为可调参数，η∈[0,1], η=0表示仅通过好友信任推荐算法得到推结果，η=1表示仅通过非好友用户相似性得到推荐结果。流程图如下： |
| 六、论文的创新点或实用性 |
| 1.在地点推荐中，用户的行为与时间密切相关，经过分析可以得知用户行为在相邻的时间段内和特定的不相邻时间段内具有相似关系。本文尝试构建一个时间段相似模型，缓解用户-地点签到矩阵的稀疏性；  2.在基于位置服务的社交网络中，用户在地点的签到行为也伴随着信息的产生，本文引入信息熵理论计算这些信息量，并为用户构建签到行为的向量，将度量用户间相似性问题转化为度量该向量的余弦相似性问题是本章计算用户兴趣相似性的基本思路；  3.在单纯的用信息熵理论计算用户相似性的地点推荐算法上融合好友信任机制，充分利用好友信息与非好友信息，使得融合后的算法能够取得更优的推荐效果。 |
| 七、研究计划进度和预期成果 |
| 计划进度：  1、2017年10月-2014年11月,确定论文题目，了解国内外的研究现状及最新的发展趋势；  2、2017年12月-2018年3月，熟悉各种推荐系统常用算法，用Python实现常见的推荐算法，为后面的推荐系统打下基础；  4、2018年4月-2018年6月，算法的调优与对比，python推荐系统的编写和调试，小论文的撰写与发表。  6、2018年7月-2018年9月，完成论文写作、修改定稿。  预期成果：  在Foursquare数据集上利用时间段相似性缓解矩阵稀疏性的问题，采用熵理论改善用户间相似性的计算方法，结合好友信任机制提高推荐结果的准确度（相较于常用的推荐算法）。 |

|  |
| --- |
| 八、导师意见  （请对项目的意义、具体内容、创新点和取得预期成果的可能性等进行评价） |
| 本论文针对基于位置服务的社交网络的个性化信息推荐，提出了一个缓解用户-地点签到矩阵的稀疏性的模型，利用信息熵概念优化用户兴趣相似性的计算，并加入好友信任机制提高推荐效果。  论文目标清晰，内容有一定创新性，工作量适中，同意开题。  导师签字： |
| 九、研究生部审核意见 |
| 审核老师签字： |