

	4	进行实验	2020.08 - 2020.10
	5	向导师、学科组作阶段性研究成果汇报	2020.11
	6	撰写学位论文	2020.12 - 2021.02
	7	研究生在学科组汇报论文撰写情况；导师介绍研究生学习的全面情况和对毕业论文的意见	2021.04
	8	论文完成，申请论文答辩	2021.04
	9	论文答辩	2021.05
备注			

- [42] P. Li, Z. Wang, Z. Ren, L. Bing, and W. Lam, “Neural rating regression with abstractive tips generation for recommendation,” in Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval, 2017, pp. 345–354.
- [43] T. Zhou, Z. Kuscsik, J.-G. Liu, M. Medo, J. R. Wakeling, Y.-C. Zhang, Solving the apparent diversity-accuracy dilemma of recommender systems, Proceedings of the National Academy of Sciences 107 (10) (2010) 4511-4515.
- [44] T. Zhou, J. Ren, M. Medo, Y.-C. Zhang, Bipartite network projection and personal recommendation, Physical Review E 76 (4) (2007) 046115.
- [45] Y.-C. Zhang, M. Blattner, Y.-K. Yu, Heat conduction process on community networks as a recommendation model, Physical Review Letters 99 (15) (2007) 154301.
- [46] M. Ye, P. Yin, W.-C. Lee, D.-L. Lee, Exploiting geographical influence for collaborative point-of-interest recommendation, in: Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, ACM, 2011, pp. 325-334.

指导教师签字:

2020 年 6 月 29 日

硕士学位论文开题论证报告

报告人姓名	张晓洁	年级	2018	专业	计算机科学与技术
论文题目	基于用户点评信息的推荐和餐厅倒闭可解释预测算法的研究				

指导教师意见:

该生对推荐和可解释预测算法相关的知识与理论研究比较透彻，参考了许多的文献资料，具有一定的研究价值。本课题结构合理，内容完整，主要观点突出，是学生学习方向的延续，对于提高学生的能力有利。同意该课题开题。

指导教师（签名）:

2020 年 6 月 29 日

- [2] F. Yu, A. Zeng, S. Gillard, M. Medo, Network-based recommendation algorithms: A review, *Physica A: Statistical Mechanics and its Applications* 452 (2016) 192-208.
- [3] S. P. Borgatti, A. Mehra, D. J. Brass, G. Labianca, Network Analysis in the Social Sciences, *Science* 323, 892(2009).
- [4] V. Colizza, A. Barrat, M. Barthelemy, A. Vespignani, The Modeling of Global Epidemics: Stochastic Dynamics and Predictability, *Bull. Math. Biol.* 68, 1893(2006).
- [5] C.H. Yeung, D. Saad and K. Y. M. Wong, From the physics of interacting polymers to optimizing routes on the London Underground, *Proc. Natl. Acad. Sci. USA* 110, 13717 (2013).
- [6] L. Lu, M. Medo, C. H. Yeung, Y.-C. Zhang, Z.-K. Zhang, T. Zhou, Recommender Systems, *Phys. Rep.* 519, 1 (2012).
- [7] D. J. Watts, S. H. Strogatz, Collective dynamics of “small-world” networks, *Nature* 393, 440(1998).
- [8] A.-L. Barabasi, R. Albert, Emergence of Scaling in Random Networks, *Science* 286, 509(1999).
- [9] S. Fortunato, Community detection in graph, *Phys. Rep.* 486, 75 (2010).
- [10] A. Zeng, S. Gualdi, M. Medo and Y.-C. Zhang, Trend prediction in temporal bipartite networks: the case of MovieLens, Netflix, and Digg, *Advs. Complex Syst.* 16, 1350024 (2013).
- [11] A. Zeng, S. Gualdi, M. Medo and Y.-C. Zhang, Trend prediction in temporal bipartite networks: the case of MovieLens, Netflix, and Digg, *Advs. Complex Syst.* 16, 1350024 (2013).
- [12] Q.-M. Zhang, A. Zeng and M.-S. Shang, Extracting the information backbone in online systems, *PLoS One* 8(5), e62624 (2013).
- [13] L. Lu and T. Zhou, Link Prediction in Complex Networks: A Survey, *Physica A* 390, 1150 (2011).
- [14] K.-I. Goh, A.-L. Barabási, Burstiness and memory in complex systems, *Europhys. Lett.* 81 (4) (2008) 48002.
- [15] Y. Borghol, S. Mitra, S. Ardon, N. Carlsson, D. Eager, A. Mahanti, Characterizing and modelling popularity of user-generated videos, *Perform. Eval.* 68(11) (2011) 1037–1055.
- [16] Y.-H. Eom, S. Fortunato, Characterizing and modeling citation dynamics, *PLoS One* 6 (9) (2011) e24926
- [17] S. Fortunato, A. Flammini, F. Menczer, Scale-free network growth by ranking, *Phys. Rev. Lett.* 96 (21) (2006) 218701.
- [18] D. Goldberg, D. Nichols, B. M. Oki and D. Terry, *Commun. ACM* 35 (1992) 61.
- [19] J. B. Schafer, D. Frankowski, J. Herlocker and S. Sen, *Lect. Notes Comput. Sc.* 4321 (2007) 291.
- [20] S. Maslov and Y.-C. Zhang, *Phys. Rev. Lett.* 87 (2001) 248701.
- [21] Y.-C. Zhang, M. Blattner and Y.-K. Yu, *Phys. Rev. Lett.* 99 (2007) 154301.
- [22] A. Said, A. Bellogn, J. Lin, A. D. Vries, Do recommendations matter?: news recommendation in real life, in: Companion Publication of the ACM Conference on Computer Supported Cooperative Work & Social Computing, ACM, 2014.
- [23] H. Liao, M. S. Mariani, M. Medo, Y.-C. Zhang, M.-Y. Zhou, Ranking in evolving complex networks, *Physics Reports* 689 (2017) 1-54.
- [24] Z. Lu, Z. Dou, J. Lian, X. Xie, Q. Yang, Content-based collaborative filtering for news topic recommendation, in: Twenty-ninth AAAI conference on artificial intelligence, 2015.

$$L = p_1 L_y + g_1 (\lambda_o L_o + \lambda_c L_c) + \lambda_l \|\Theta\|_2^2$$

其中, p_1 和 g_1 分别是预测任务和解释任务的权重, λ_o 和 λ_c 是整体相似性和概念相似性的权重。最后 $\lambda_l \|\Theta\|_2^2$ 是调节参数。

2.2.3 实验

在这部分研究内容中, 我们使用的数据集是大众点评网站 2010 年间用户对商店的评论文本。因为我们的模型需要大量的数据进行训练, 为了保证实验效果的有效性, 我们选择了数据量排名前三名的城市的数据, 分别是上海、北京、广州。这三个城市的基本数据统计如表 3。

	reviews	users	shops	restaurants	closed
上海	581713	126158	11508	10251	3312
北京	361635	75283	8476	5067	1308
广州	87166	20532	2247	1932	509

表 3: 三个城市的统计数据。

预测任务方面, 我们选择了三个常用的机器学习算法和三个深度学习算法作为基线方法, 分别是 CNN, RNN, NRT, GBDT, SVM, LR。图 3 表明, 我们的方法在三城市的 AUC 中始终优于基线方法。DCA 在上海、北京和广州分别比基准方法提高了 $18.12\% \sim 50.74\%$ 、 $19.41\% \sim 59.41\%$ 和 $19.20\% \sim 49.52\%$ 。这证明了我们模型的有效性。另外, 我们还剔除了 DCA 模型中的概念层选择器, 构造了 DCA-C 变体, 实验结果标明 DCA-C 的 AUC 比 DCA 会低一些, 这也说明了概念在预测任务的重要性。

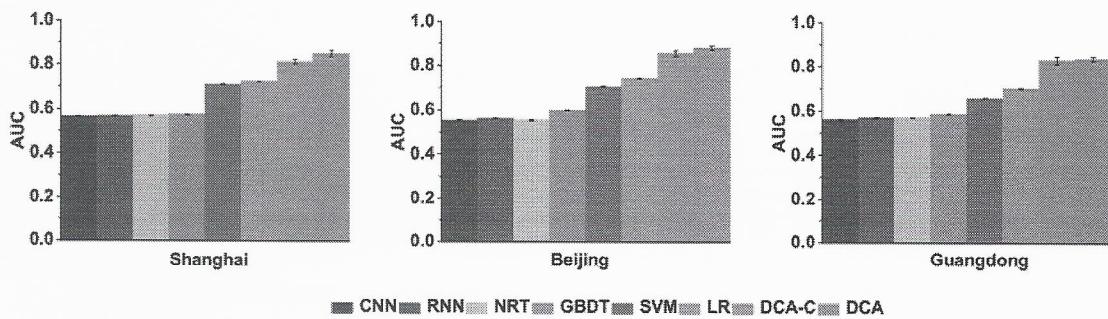


图 4: 三个数据集中, 我们的方法和基线方法在准确性上的比较结果。

解释任务方面, 我们选择了 Lexrank 和 NRT 算法作为基线算法。首先分析整理的文本相似性, 表 4 显示, 我们的方法在 BLEU 和 ROUGE 指标上始终优于基线。以 BLEU 数据为例, 我们的方法在广州数据和北京数据上分别比基线提高 $2.99\% - 36.07\%$ 和 $7.23\% - 49.5\%$, 在上海数据上分别提高 $0.67\% - 11.64\%$ 。我们还给两个基线算法添加了概念信息, 分别构造了变形 Lexrank+C 和 NRT+C, 实验结果标明, 添加了概念信息的基线方法也比原始的基线方法有提升, 这也充分说明了概念在解释任务起到的积极作用。

	U	I	L	$\langle k_i \rangle$	$\langle k_\alpha \rangle$
Yelp	123368	41958	804789	6.5	19.6
Brightkite	37303	15651	201308	92.3	3.4
Foursquare	2293	61852	211670	5.4	12.9

表 1：三个数据集的基本信息：用户数量 U，项目数量 I，链接数量 L，用户平均度 $\langle k_i \rangle$ ，项目平均度 $\langle k_\alpha \rangle$ 。

除了 THybridS 作为基线方法之外，还选择了另外 3 个对比方法：第一个是 Fusion 算法[46]，该算法使用两种不同的评分，一种基于用户相似度，另一种基于地理位置。分数分别计算，然后合并在一起。第二个是 closest-item，我们在 THybridS 中添加了空间距离，但是我们也想将空间距离本身的性能作为基准性能来考虑。我们设计了一个 Closest-item 方法，它只使用地理信息进行推荐。它根据产品与用户之间的距离对产品进行排序，然后生成推荐列表。第三个是 SVDPP，是传统的机器学习算法。

在图 2 展示了在三个数据集上，改变 lambda 参数和距离参数，recall，precision 和新颖度的热度图。可以看出，这三个指标的趋势基本一致。

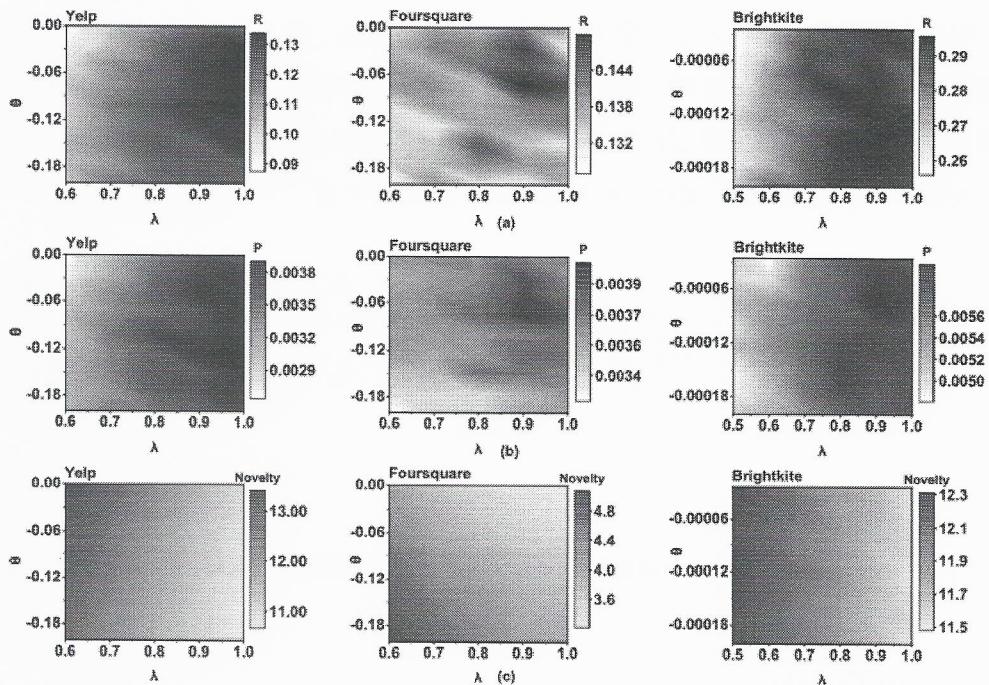


图 2：Yelp, Foursquare, Brightkite 数据集上的 precision、recall 和新颖度热度图

表 2 展示了在三个数据集，我们的方法和 THybridS 的比较结果。在这三个数据集中，recall 值有明显的提高：Yelp 的 recall 优化了为 5.5%，Foursquare 的优化了 8.7%，Brightkite 的优化为 2.5%。另外，我们的方法在 precision 上始终优于 THybrids。具体来说，在 Yelp、Foursquare 和 Brightkite 上，Geo-THybridS 分别比 THybrids 提高了 8.3%、2.6% 和 1.8%。AUC 并没有因为在这三个地方都添加了地理修改而得到改善。这是因为数据集中项目的数量很多，项目的排名不会显著影响 AUC 的值，例如，项目的排名从 20 名变化到 50 名，AUC 评分不会发生很大变化。但对于用户来说，这是非常重要的，所以我们使用 recall 作为主要的评估指标。Mean distance 指标是用来评估推荐列表上的项目距离用户的平均距离，可以明显的看出，我们的方法推荐出来的产品与用户的距离更近，这也符合我们的常识。

评论文本中包含的概念的最大数目。概念权重矩阵 AC 的计算如下：

$$AC_{ij} = F(c_{u,i})^T M_c F(c_{v,j})$$

在使用平均池化操作从 AC 中计算出表示用户和餐厅的表示向量 \vec{c}_u 和 \vec{c}_v

$$\begin{aligned}\vec{c}_u &= (\text{Gumbel}(\text{avg}_{\text{col}}(AC)))^T c_u \\ \vec{c}_v &= (\text{Gumbel}(\text{avg}_{\text{row}}(AC)))^T c_v\end{aligned}$$

(3) 多指针学习

由于在评估商店时可以同时考虑评论和概念，所以我们支持聚合多个指针。以用户 u 为例，用户的协同注意力嵌入向量表示 $\bar{u} = [r_u : \vec{c}_u]$ ，由评论级用户嵌入和概念级用户嵌入组成。在多指针设置中，我们使用不同的 Gumbel 噪声运行选择器多次，得到多个样本： $\{\bar{u}_1, \dots, \bar{u}_p\}$ 。同样，我们获得了一组共同关注项嵌入的样本： $\{\bar{v}_1, \dots, \bar{v}_p\}$ 。然后我们使用一个非线性层 $\text{ReLU}(\cdot)$ 来聚合它们。

$$\begin{aligned}x_u &= \text{ReLU}(W[\bar{u}_1, \dots, \bar{u}_p] + b) \\ x_v &= \text{ReLU}(W[\bar{v}_1, \dots, \bar{v}_p] + b)\end{aligned}$$

1.2.2 Factorization machine(FM)模型

FM 用于实现不同特征之间的两两交互 $[x_u, x_v]$ ，预测公式如下。

$$\hat{y} = f(x) = w_0 + \sum_{i=1}^n w_i x_i + \sum_{i=1}^n \sum_{j=i+1}^n \langle p_i, p_j \rangle x_i x_j$$

由于任务是预测餐厅关闭情况，这是一个二元预测问题，所以通常使用 sigmod 交叉熵损失函数。

$$L_y = \frac{1}{2|\Omega|} \sum_{(u,i) \in \Theta} (-[y \log \hat{y} + (1-y) \log (1-\hat{y})])$$

其中， Ω 代表了训练集， y 是餐厅真实倒闭情况。

1.2.3 Gated Recurrent Unit(GRU)模型

近期研究表明，LSTM 和 GRU 在生成文方面表现良好。考虑到 GRU 在相似性能下具有较少的参数和较高的计算效率，我们的实验使用 GRU 来生成解释文本。

$$p(l_t | l_1, l_2, \dots, l_{t-1}, \hat{y}) = \zeta(h_t)$$

其中， l_t 是生成的在 t 时刻的单词。 $\zeta(\cdot)$ 是一个 softmax 函数。

$$\zeta(z^{(i)}) = \frac{e^{z^{(i)}}}{\sum_{n=1}^N e^{z^{(i)}}}$$

h_t 是 t 时刻的隐藏状态，依赖于 $t-1$ 时刻的隐藏状态。

$$h_t = \text{GRU}(h_{t-1}, l_t)$$

出示隐藏状态 h_0 表达如下：

$$h_0 = \tanh(W_u x_u + W_v x_v + W_u \hat{y} + b)$$

然后将隐藏状态输入到最终的输出层，生成时间为 t 的单词分布。

$$d_t = \zeta(W_d h_{t-1} + b_d)$$

解释任务的损失值由两部分组成，第一部分是生成文本整体的相似性：

$$L_o = \frac{1}{|\Omega|} \sum_{(u,v) \in \Theta} \sum_{t=1}^T (-\log d_{t,\bar{l}_t})$$

另外一个损失值是概念相似性的损失值：

- (1) 在推荐算法中，我们在不获取用户当前位置的情况下，使用 DBSCAN 算法预测用户的位置。既保护了用户的隐私，又提高了推荐的准确性。
- (2) 在可解释预测模型中，相比传统的破产预测使用的金融数据，我们使用的数据来自网络服务，数据规模大且更容易获取。
- (3) 在可解释预测模型中，从评论文本数据中构造中文概念集，提高了预测的准确性和解释性。
- (4) 在可解释预测模型中，将预测任务和解释任务进行联合学习，提高了模型的泛化能力。

二、论文研究的主要内容，方案和拟采用的研究方法、手段；已进行的科研工作基础和已具备的科学实验条件（包括文献资料及主要实验仪器设备准备情况等），对其他单位的协作要求；论文总工作量（估计），论文初稿的进度以及预期结果：

1、论文研究的主要内容，方案和研究方法

1.1 基于扩散的位置感知推荐系统

1.1.1 二分网络

大多数数据可以用网络(或称为图)表示。网络由节点组成，在[43]中，当两个节点之间有关系时，通过一条链路连接起来。在这项工作中，我们使用具有两种类型节点的数据：用户和项目。链接只能发生在用户和项目之间(即不能发生在两个用户或两个项目之间)。这就是我们定义的二分图网络。为了简单起见，我们对非用户节点都统称位项目，它们可以是与用户交互的任何东西。例如，用户享受的服务、购买的产品、发布的评论信息等。

我们用拉丁字母表示用户，用希腊字母表示项目。与数据相关的时间用 t 表示。用户集被定义为 U ，项目集用 I 表示。网络的邻接矩阵 A 的元素表示用户和项目之间的关系。如果用户 i 连接到项目 α ，则有 $a_{i\alpha}=1$ ，否则则有 $a_{i\alpha}=0$ 。用户 i 的用户的度 $k_i(t)$ 表示当时间为 t 时，连接到该用户的项目数。相似地，项目的度 $k_\alpha(t)$ 代表 t 时刻连接到项目上的用户数量。

1.1.2 扩散推荐算法

在第一部分研究内容中，我们是以基于概率扩散[44]的推荐系统为基础进行算法改进的。选择了两种阔算推荐算法，第一个是基于随机游走过程的物质扩散推荐算法，首先每一个商品把自己的能量平均分给所有购买过它的用户。用户的能量值则是从所有商品所得到的能量值得总和；接下来，每一个用户再把自己的能量平局分给所有购买过的商品，商品的能量则是从所有用户收到的能量值得总和。它倾向于推荐那些度较大（较流行）的物品，相当于一个凸透镜，将用户的视野汇聚在那些较流行的节点上，具有较高的推荐精度。而基于热扩散过程 [45]的方法首先每一个用户的温度等于所有他购买过的商品的温度的平均值，接下来每一个商品的温度等于所有购买过他的用户的温度的平均值。热传导倾向于推荐那些度较小（较不流行）的节点，相当于一个凹透镜，把用户的视野发散到了那些较不流行的物品，但更倾向于推荐对象的多样性。由于两者都是基于扩散的物理过程，所以它们可以优雅地合并在一起。这两个过程的合并产生了一种同时比两个过程分开更精确和更多样化的方法。

THybridS 的推荐分数首先通过传播和过程获得，然后根据最近的网络活动调整分数。在数学上，传播矩阵为：

影响。而在预测分析方面，基于机器学习或深度学习的各类预测算法也在不断推动商业模式的变革。比如电商行业，基于大数据，根据客户点击与购买记录，利用算法推测客户喜好，展开精准营销。然而很多机器学习模型（深度学习首当其冲）的可解释性不强，这也导致在真正的商业应用中无法被广泛地采纳。这是因为企业决策者在做经营决策时无法接受一个不可解释的结论，更无法接受如果预测出来的结果并不准确，用户却不知道如何优化当前的模型。由于业务场景千变万化，没有一套通用的预测算法可以解决所有问题。既然场景、模型都有那么多的选择，企业管理者对模型的信任都会比较谨慎。因此无论我们提供给客户的解决方案的最终目标是什么，客户都需要一个可解释、可关联、可理解的解决方案，这是建立信任的必要因素，因为它代表安全、责任与可靠！此外，借助模型的可解释性，用户可以通过调整可控因素，获得最优预测结果，为企业管理者提供更多可操作的决策方法。而作为解决的提供商，我们也可以在模型的可解释性中受益，从而验证并持续改进我们的工作。

由此可以看出，将异构信息应用到网络科学服务中，无论是对于消费者还是商家，都能够提供更加准确和贴心的服务异构信息为研究人员提供了探究问题的崭新思路方法，越来越多来自不同学科的研究人员参与到信息挖掘的工作中来，将产生巨大的实际应用效果，并有助于信息科学和其他学科的深入交叉。

2、国内外研究现状

(1) 推荐算法

协同过滤算法是推荐系统应用最广泛的算法，主要分为基于近邻的协同过滤算法 (memory-based CF)和基于模型的协同过滤算法(model-based CF)。基于内存的协同过滤算法又分为基于用户的协同过滤算法(UserCF) 和基于物品的协同过滤算法(ItemCF)[27]。基于用户的协同过滤算法是最早出现的推荐算法，算法首先计算和目标用户兴趣相似的用户集合，然后为目标用户推荐该相似用户集合 中用户喜欢且未接触过的物品。基于物品的协同过滤算法是目前业界应用最多的算法，该算法给用户推荐那些和他们之前喜欢的物品相似的物品。基于模型的协同过滤算法主要通过机器学习和数据挖掘模型，利用分类、回归、矩阵分解等算法提取用户和物品的隐含模式进行推荐。其中代表性的有基于贝叶斯信念网络的算法[28]、基于聚类模型的算法[29]、基于回归模型的算法[30]、基于矩阵分解模型的算法[31]等。

推荐系统是人工智能和机器学习研究的一个分支。由于深度学习在许多领域取得了巨大成功，许多研究人员开始将深度学习与推荐算法结合，解决推荐系统的各种问题(如冷启动、稀疏性等)，提高推荐性能。当前基于深度学习的推荐算法研究分为 4 类：利用辅助信息的深度学习推荐算法；基于模型的深度学习推荐算法；动态深度学习推荐算法；基于标签的深度学习推荐算法。其中，利用辅助信息的深度学习推荐算法又可分为 3 种类型：利用辅助信息仅提取用户(或物品) 的特征表示；利用辅助信息分别提取用户和物品的特征表示；利用辅助信息提取用户和物品的共同特征表示。

20世纪 90 年代开始出现基于位置的服务和社交网络，以及基于位置的社交网络[32,33]。自这些网络出现以来，推荐系统开始使用地理信息[34,35]。最近的工作[36]采用了一组基于马尔科夫的预测器和一系列的位置推荐算法来挖掘基于位置的社交网络。另一项工作是基于位置的推荐，将协同过滤与位置[37]相结合。所得到的方法是推荐精度和计算效率之间的权衡。在[38]中，选择了一种不同的方法。将地理区域建模为两个层次系统，分别考虑了局部尺度和区域尺度。[39]研究了新增的地理信息和用户之

撰写要求

硕士研究生中期考核通过者，进行开题报告并进入学位论文工作阶段。硕士学位论文开题报告撰写要求：

一、开题报告内容

1. 选题的目的和来源，课题研究的意义、学术和应用价值以及国内外研究动态；
2. 选题的基本内容、构思、创新点及初步见解；
3. 课题拟采用的研究方法和手段；
4. 课题研究程序、实验方案和预期达到的目标；
5. 论文写作进度安排及所需提供的条件、设备和经费来源。

二、开题报告会由导师指导小组（3-5人）进行集体评议，写出评语，评定考核成绩。

三、开题报告后，研究生应根据评审小组的意见，对选题方案进行修正、补充和提高，并填写《硕士研究生学位论文开题报告书》，按规定的程序报批备案后，方可进入论文写作阶段。

四、开题报告后，若学位论文课题有重大变动，应重新作开题报告，并按程序重新报批。

五、《硕士研究生学位论文开题报告书》填写内容必须属实，字迹端正清楚，由学院（部）存档备查。