大数据精准营销中搜狗用户画像挖掘

李恒超，李裕礞，王安然，钱凌飞，任璐，林鸿飞
大连理工大学计算机科学与技术学院，大连，116023
E-mail: hflin@dlut.edu.cn

——大大黑楼战队

目录

1. 数据预处理

1.1. 停用词处理

在各类文本相关的任务中，大多需要先对样本进行分词、去停用词等预置处理。在本任务中，通过对训练数据进行细致的分析，结合人们进行日常检索的先验知识，发现“空格”、“标点”及很多停用词均有助于判别用户的基本属性。如下图所示：

因此在进行分词处理及 TFIDF 特征计算的过程中，均保留了空格、标点以及停用词这些信息，在该任务中是很好的特征。比如我们发现，一个人的受教育程度越高，越喜欢在查询词中添加空格，例如“大连理工大学　图书馆　地址”；而受教育程度越低，使用“之”的概率就越大，因为“之”经常在一些玄幻修真小说名中出现，低年龄和低学历会经常搜类似的小说如《网游之邪龙逆天》。

1.2. 分词处理

我们发现不限制字典长度才能达到最好的预测效果，由于电脑内存的限制，我们的 Bigrams 过滤掉了文档频率低于 5 的词，整个字典长度是 174W。我们分析该语料的特点是低频词特别多，而且这些低频词有很好的预测效力。

该任务中的文本为用户的查询记录，其长度往往较短，因此分词效果便显得较为重要。本文测试了几种常用的分词工具，并使用 Bayes 模型比较了它们在本任务中的效果，下表展示了其中表现较好的几组结果：

分词工具	学历	年龄	性别
JIEBA	58.55%	54.35%	81.83%
NLPIR	57.57%	53.84%	81.12%
THULC	58.54%	54.14%	81.26%
Ngram(1,2)	58.03%	53.87%	81.11%

下表是分词效果：

原单词	JIEBA	THULC	NLPIR
周公解梦大全查询	周公, 解梦, 大全, 查询	周公解, 梦, 大全, 查询	周, 公, 解, 梦, 大全, 查询
中财网首页	中财网，首页	中财网，首，页	中，财网，首，页

2. 特征表示

为了对数据进行更全面的刻画，我们从用户用词习惯、语义相关性及所包含的主题几方面构建了多角度的特征。具体特征如下：

2.1. Bag of Words

我们筛选了至少出现在 5 篇文档中的词语来组成词表，统计 one-gram 及 bi-gram 特征。该特征可以有效体现出不同类别用户的用词习惯。

该特征虽然简单且有效，但其缺陷在于其缺乏词与词之间的语义相关信息，因此我们在文本语义方面采用了更多的特征表示方法。

2.2. Word Embedding

我们使用 Google 公布的 word2vec 工具在搜狗新闻语料上训练得到了常用词的词向量，并将其应用到用户的历史查询词中。该方法得到的词向量可以有效计算出两个词之间语义的相似程度，从而表示出不同用户查询历史的差异。

2.3. Topical Word Embedding

在该任务中，每个用户具有多组查询词，其中有些查询相关性较强，有些则完全不相关。使用主题模型来抽取用户的多个查询主题，更有利于刻画用户的查询习惯、关注方向等。我们将用户的所有查询词拼接在一起，使用 LDA 模型进行主题分析。基于 LDA 的结果，使用 Topical Word Embedding(TWE)[2]模型训练得到每个查询词的词向量。TWE 模型与常用的 Word2Vec 不同在于，其计算出的词向量同时考虑词的上下文及该词所在主题的信息。使用 TFIDF 特征值对用户查询历史中的词向量进行加权平均，可以得到表示整体查询的向量值，可以将其直接作为多个分类模型的输入，完成用户层级的分类任务。

2.4. Doc2Vec 特征表示

为了将文档直接表示成一个固定长度的向量，我们还采用了 Doc2Vec[1]方法，它通过直接构造文档向量，并将该向量加入到该文档中词向量的训练过程，进行共同训练，从而得到能直接体现该文档语义特征的向量。根据训练文档向量的网络结构不同，可分为 Distributed Memory Model(DM)与 Distributed bag of words(DBOW)两种模型，其中 DM 模型不仅考虑了词的上下文语义特征，还考虑到了词序信息。DBOW 模型则忽略了上下文词序信息，而专注于文档中的各个词的语义信息。我们同时采用了 DM 和 DBOW 这两种文档向量表示方法，从而保证构建的特征中信息的完整性。

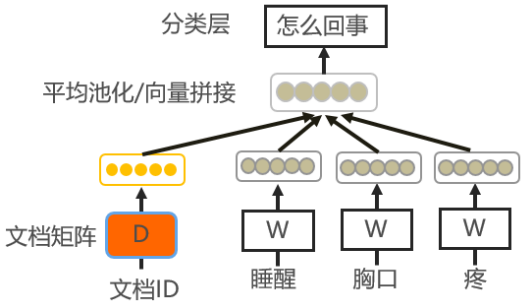

图 1 DM 模型结构图

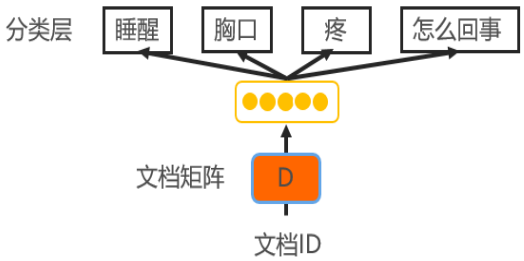

图 2 DBOW 模型结构图

以下在 3 个子任务上使用对 doc2vec 表示向量使用 tsne 降维的效果：

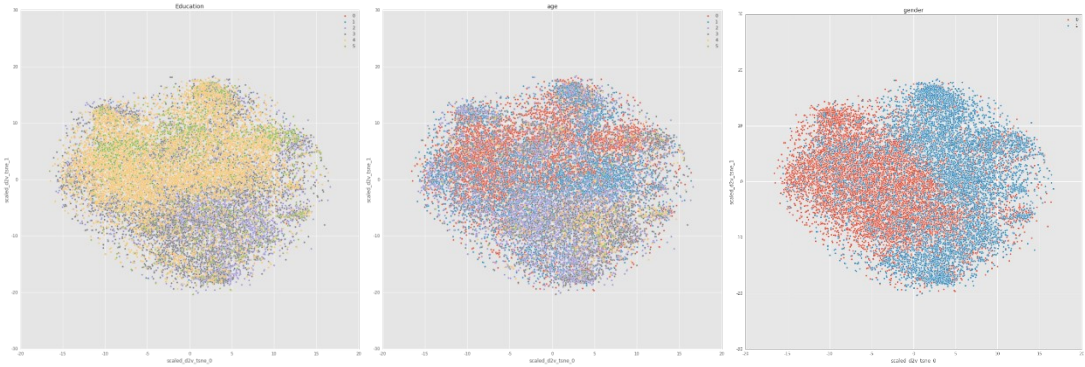

图 3 教育任务上的效果　　图 4 年龄任务表现效果　　图 5 性别任务表现效果

2.5. 人工构建的特征

除以上通过模型学习得到的特征表示之外，我们还人工构建了一些特征，从而完成多视角特征体系的构建。人工构建的特征包括：查询词的个数、查询词的平均长度、查询词的最大长度、有空格的 query 占总查询的比例、英文查询词的比例、所有 query 的语义标准差（该特征用来表示用户查询的多样化程度）。

3. 模型结构

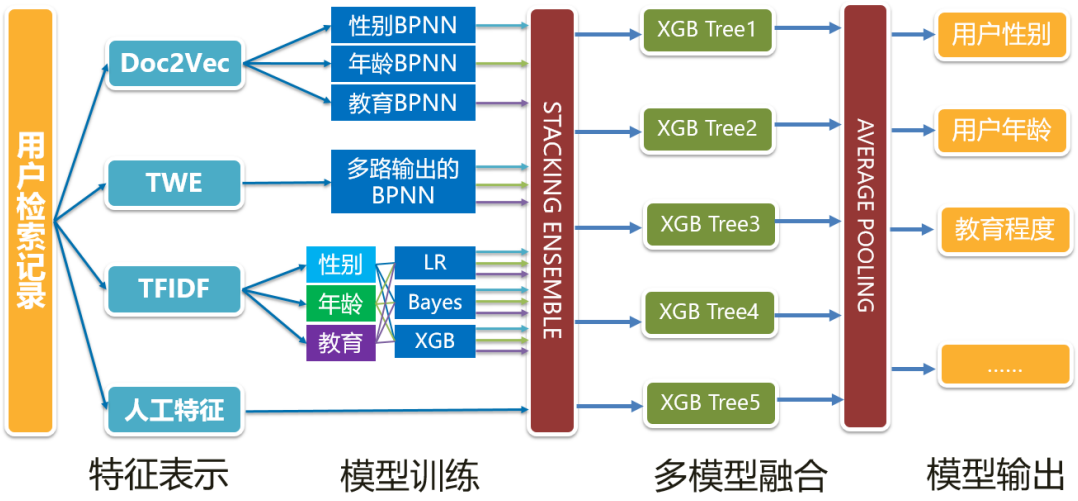

图 6 模型整体结构图

图 6 中给出了第一赛季和第二赛季的综合模型结构图，其中使用了两级的模型结构。第一级中使用了传统机器学习模型与 TFIDF 特征相结合来抽取用户用词习惯的差异，使用神经网络模型与表示学习相结合来抽取查询的语义关联信息。第二级模型中使用了 XGB Tree 模型及 Stacking 多模型融合的方法，来进一步提升模型的准确性与泛化能力。

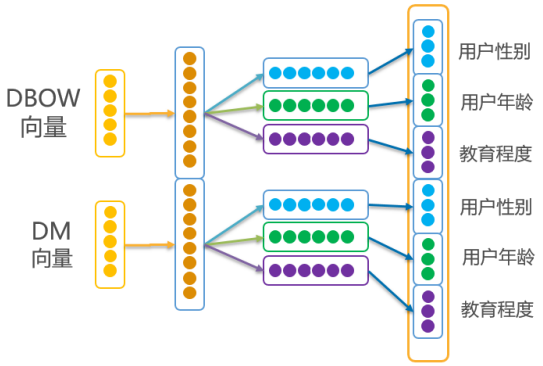

图 7 多输出神经网络结构图

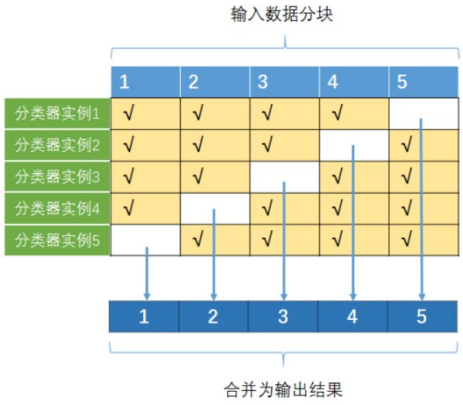

图 8 多模型融合数据分割图

3.1. 基于 TFIDF 的传统机器学习模型

第一层模型中，我们尝试了 sklearn 中的 LogisticRegression, MultinomialNB，BernoulliNB,KNN,SVC, RandomForestClassfier 和 xgboost 中的 gblinear 和 gbtree。其中，由于 tfidf 特征过于稀疏、维度过高，树形模型表现效果很差；由于数据量太大，KNN 和 SVC 算法都不能训练出结果；Gblinear 线下测试要高于逻辑回归，但是线上成绩不如逻辑回归。

分类器	学历	年龄	性别	线下成绩
LogisticRegression	0.6424	0.5978	0.8338	0.6913
Gblinear	0.6484	0.6068	0.8362	0.697
MultinomialNB	0.6121	0.5798	0.8262	0.6727
BernoulliNB	0.6032	0.5712	0.8246	0.6663

表 1　传统机器学习模型的线下成绩

3.2. 基于分布式向量的神经网络模型

我们同时使用了 Doc2Vec 与 TWE 进行用户查询词的表示。我们发现在该语料上，Doc2Vec 的表现对迭代次数和学习率很有关系，所有我们使用 5 折交叉验证分别选择 pv-dbow 和 pv-dm 的迭代次数和学习率。其中 pv-dbow 的迭代 2 次，学习率为 0.025，pv-dm 的迭代次数为 10 次，学习率是 0.05；其中 NN 的参数是 300 维的隐藏层；TWE 中主题数 400，alpha 为 0.5，beta 为 0.1，输出向量维度 400 维。以下是一些实验数据：

分类器	学历	年龄	性别	线下成绩
Dbow-lr	0.6330	0.5947	0.8371	0.6883
Dm-lr	0.6370	0.5918	0.8349	0.6879
Dbow-NN	0.6626	0.6172	0.8429	0.7076
Dm-NN	0.6506	0.6096	0.8383	0.6995
TWE-multiNN	0.6301	0.5990	0.8350	0.6880

表 2　神经网络模型的线下成绩

在该语料上，Doc2Vec 的表现特别好，我们分析原因可能有 2 点：

1) 用户的查询词千差万别，相比于一般常见的电影评论、新闻等数据集，该语料的低频词特别多。Doc2vec 能够对低频词有很好的语义总结，对这些低频词利用更充分；
2) 该语料是很多查询词的拼接，语料中的词序（word order）特征不怎么重要，pv-dbow 的训练方式忽略了词序，天然地适合处理该语料。

3.3. 第二层融合模型

我们的融合技术很大程度上参考了 Combining Predictions for Accurate Recommender Systems [7], Ensemble of Generative and Discriminative Techniques for Sentiment Analysis of Movie Reviews [8]这两篇论文。

第一层模型分别在 3 个子任务上训练，模型输出的概率值作为下一层模型的输入。由于 3 个子任务分别是 6 分类、6 分类和 2 分类，所以我们第一层特征维数是 4*（6+6+2）= 56。下表是一些实验结果：

模型	学历	年龄	性别	线下成绩	线上成绩
Tfidf-ensemble	0.6422	0.6104	0.8328	0.6951	0.7101
Dbow-ensemble	0.6717	0.6332	0.8492	0.7180	0.7116
Tfidf-dm-ensemble	0.6708	0.6332	0.8481	0.7174	0.7213
Tfidf-dbow-dm-ensemble	0.6788	0.6389	0.8516	0.7231	0.7246
Tfidf-dbow-dm-twe-ensemble	0.6790	0.6385	0.8520	0.7232	0.7255

表 3　多任务多模型融合的线上成绩

Tfidf-ens 和 Dbow-ens 是在 3 个子任务分别训练 tfidf-lr 模型，然后使用 xgboost 融合；tfidf-dm-ens 是指 3 个子任务上分别训练 tfidf-lr 模型和 dbow-nn 模型，然后使用 xgboost 融合；Tfidf-dbow-dm-ens 是在 3 个子任务上分别训练 tfidf-lr、dbow-nn、dm-nn 模型，然后使用 xgboost 融合；tfidf-dbow-dm-twe-ens 是融合了 tfidf-lr、dbow-nn、dm-nn、twe 的模型。

在该任务上，融合技术非常关键，我们只用 tfdif-lr 模型然后在其之上使用 xgboost 融合，就能达到 0.710 线上成绩，这样的成绩差不多就能进前 5 名了。融合技术之所有如此关键，我们分析有以下 3 点：

1) 在多分类任务中，一般模型是基于 OneVsRest 或者 OneVsOne，这样分类器只能看到 2 类的分类信息，而使用 stack 技术输出每一类的概率值后，第二层模型可以看到所有的分类结果，然后在其之上做一些阈值判断、相互校验等等。2 分类任务性别的融合效果不如其它 2 个 6 分类任务也验证了这一点。
2) 3 个子任务之间是有一些关联的，比如年龄和学历之间有很大的关联。第二层的模型

能对这样的特征关系有很好的学习。例如，我们试过在预测学历时去掉年龄和性别特征，预测结果有一定程度的降低。

3) 该数据集存在数据不均衡的问题，但由于评价指标是准确率（acc），所以downsample 和 upsample 都没有必要做。借由 xgboost 模型，我们可以很好的学习各个类别中最优的阈值。

4. 数据后置处理——错误分析

错误样本分析可以给模型优化指引方向。在进行错误样本分析的过程中，我们也找到了一些规律。

对于属性值存在空缺的样本，我们首先使用属性值已知的样本作为训练样本，使用 LR 模型训练分类器，再对这部分属性空缺样本进行预测，从而补全空缺值。但我们发现在最终的两级多模型融合得到的结果中，对于教育属性空缺的样例，它们的年龄和性别预测准确率很低；对于年龄属性空缺的样例，教育预测准确率很低；对于性别属性空缺的样例，教育预测准确率很低。具体比较结果如下：

教育属性空缺与未空缺部分的预测准确率比较

	空缺部分	未空缺部分
教育	89.25%	65.25%
年龄	41.54%	65.29%
性别	73.57%	86.15%

年龄属性空缺与未空缺部分的预测准确率比较

	空缺部分	未空缺部分
教育	14.59%	68.38%
年龄	79.17%	62.81%
性别	85.17%	84.97%

性别属性空缺与未空缺部分的预测准确率比较

	空缺部分	未空缺部分
教育	38.10%	68.13%
年龄	64.97%	63.04%
性别	94.48%	84.77%

各属性空缺值占比

字段名	空缺比例
Education	9.28%
Age	1.67%
Gender	2.16%
任意值空缺	11.51%

以年龄属性空缺的样本为例，其共包含样本 1670 个，其中教育属性预测正确的比例为 14.59%；而对于其他年龄属性未空缺的样本，其教育属性预测正确的比例为 68.38%，差距为 53.79%。由此可推测出：存在空缺值的样本，它们的标注质量较差。结合先验知识，我们也可直观地想到，用户的这三项基本属性存在空缺，可能意味着用户信息统计较不充分、信息来源可靠性较差。因该部分样本噪音较大，可能干扰分类器的训练，因此本文最后将存在空缺的样本从训练集中删掉，这样训练出的模型较使用 LR 填充空缺属性值在最终评价指标上有 0.1%~0.2%的提升。

5. 总结与展望

本次竞赛极大地锻炼了我们的团队协作及解决问题的能力，有机会将学习的理论真正应用起来。在竞赛中我们还尝试了很多其他方法，但因一些条件的限制，无法调至最优的水平，其中较有参考意义的方法包括深度学习模型及查询扩展方法。

5.1. 深度学习方法

目前深度学习模型发展迅猛，已经应用到自然语言处理中的多项任务中。本文也尝试采用深度学习模型来解决此问题。使用 Word2Vec 训练得到的词向量，输入到 CNN[6]模型中，经 Pooling 层及 Softmax 分类层，最后输出结果。另外尝试了基于层级式 Attention[5]机制[3]的深度神经网络模型，该模型在微博语料上取得了很好的效果，但在本任务中并无突出表现。

分析原因在于，使用搜狗新闻语料训练得到的词向量只能覆盖本语料中词表的 18%，其中大量检索词无法找到，包括人名、地名及生僻字等；另外，本任务中的文本为查询关键词，并非完整的句子，其语义的不连贯性及表述的不全面性也导致了基于语义词向量的

深度神经网络模型表现不佳。

5.2. 查询扩展与伪相关反馈方法

因查询关键词语义不完整，我们考虑使用搜狗新闻语料及查询扩展方法，补充查询关键词的信息。通过使用 BM25F 相似度计算方法对训练集和测试集的每条查询在新闻语料中进行搜索匹配，筛选出其中匹配得分最高的前 n 条新闻。分别使用这些新闻的标题、内容、以及类别信息，对用户查询关键词进行特征补充，从而增加用户文本的信息量，其新闻类别又可起到特征降维的作用。

在实验过程中发现，由于新闻语料并不充分，内容具有局限性，在该语料上检索得到的结果含有较大噪声，在模型中拼接该部分特征并无明显提升，但该方法也可作为解决该任务的一种思路，未来我们也将继续尝试这部分的工作。

参考文献

[1] Le, Quoc V., and Tomas Mikolov. "Distributed Representations of Sentences and Documents." *ICML*. Vol. 14. 2014.
[2] Liu, Yang, et al. "Topical Word Embeddings." *AAAI*. 2015.
[3] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." *Journal of machine Learning research* 3.Jan (2003): 993-1022.
[4] Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." *arXiv preprint arXiv:1301.3781* (2013).
[5] Yang, Zichao, et al. "Hierarchical attention networks for document classification ." *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2016.
[6] Kim, Yoon. "Convolutional neural networks for sentence classification." *arXiv preprint arXiv:1408.5882* (2014).
[7] Jahrer M, Töscher A, Legenstein R. Combining predictions for accurate recommender systems[C]//Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2010: 693-702.
[8] Mesnil G, Mikolov T, Ranzato M A, et al. Ensemble of generative and discriminative techniques for sentiment analysis of movie reviews[J]. arXiv preprint arXiv:1412.5335, 2014.