中图分类号：

论文编号：

北京航空航天大學

中法工程师学院

工程师毕业实习报告

报告标题

作者姓名	陈千奔
实习领域	大数据算法应用
实习岗位	应用研究算法工程师
企业导师	王国华
培养院系	中法工程师学院

Research Application of Corporate Profile and Public Opinion System Based on Big Data

A Dissertation Submitted for the Degree of Engineer

Candidate:

Qianben Chen

Supervisor:

Guohua Wang

Sino-French Engineer School

Beihang University, Beijing, China

中图分类号：TP242

论文编号：10006ZY1724130

工程师毕位实习报告

基于大数据的企业画像与舆情系统的应用研究

作者姓名	陈千奔	申请学位级别	工程师/硕士
企业导师姓名	王国华	职　　位	应用研究算法岗
专业方向	ITR	实习方向	大数据算法应用
实习时间自	2021 年 05 月 07 日	起至	2021 年 11 月 05 日止
报告提交日期	年　月　日	答辩日期	年　月　日

关于实习报告的独创性声明

本人郑重声明：所呈交的报告是本人在指导教师指导下独立进行实习工作所取得的成果，论文中有关资料和数据是实事求是的。尽我所知，除文中已经加以标注和致谢外，本论文不包含其他人已经发表或撰写的研究成果，也不包含本人或他人为获得北京航空航天大学中法工程师学院或其它教育机构的学位或学历证书而使用过的材料。与我一同工作的同志对研究所做的任何贡献均已在论文中作出了明确的说明。

若有不实之处，本人愿意承担相关法律责任。

实习报告作者签名：陈千奔　　　　日期：2021 年 10 月 29 日

实习报告使用授权书

本人完全同意北京航空航天大学中法工程师学院有权使用本实习报告（包括但不限于其印刷版和电子版），使用方式包括但不限于：保留实习报告，按规定向国家有关部门（机构）送交实习报告，以学术交流为目的赠送和交换实习报告，允许实习报告被查阅、借阅和复印，将实习报告的全部或部分内容编入有关数据库进行检索，采用影印、缩印或其他复制手段保存实习报告。

保密实习报告在解密后的使用授权同上。

实习报告作者签名：陈千奔　　　　日期：2021 年 10 月 29 日

企业导师签名：王国华　　　　日期：2021 年 10 月 30 日

中文摘要

为了参与腾讯云大数据计数影响力的构建与传播，发现数据价值并支撑 B 端应用，本实习的主要目的是负责数据类算法的研究与落地，对于 B 端客户的需求，利用腾讯云大数据平台进行海量数据的分析与挖掘。其中，我主要参与了两大项目的建设：

1. 企业画像指数构建。我们对 2.5 亿企业的风险、发展、活跃度等各个指数层面进行评估，并且建立一套指数自动评估机制。数据的预处理部分是指数构建的重点。考虑到数据量极大、分布离散、特征稀疏等特点，需要利用离散编码、特征选择、稀疏矩阵特征处理等技术进行深度处理；考虑缺少标签的原因，我们将会使用半监督学习的算法，并使用 XGBoost 强学习器拟合特征，得到应用于不同指数层面的最优函数回归结果。

2. 舆情系统热点发现。我们将相似新闻进行聚合，并分析其主要内容，主要包含两步骤：新闻聚合和脉络分析。新闻聚合是文本聚类的重要应用，我们定义话题即新闻的集合，发现热点话题就是将新闻数据合理聚类的过程。在技术角度上，我们使用一种独特的两步聚类方式，来解决可能存在的簇数目不定和效率低下问题；脉络分析模型用于获取指定话题的脉络，即由这个话题按时间轴展开的相关关键新闻，这一新闻链能客观地展示事件的发展和推理过程。在技术角度上，我们使用 DBSCAN 去噪、BERT 相似度计算和维特比算法等获得时间轴上的新闻链。

上述内容将详细展示在实习内容板块中，除此之外，主要内容还包括公司及岗位介绍、实习描述、自我评估和感谢。在公司及岗位介绍部分，对腾讯云及大数据算法应用实习岗位进行介绍；在实习描述部分，回顾实习职责及体会；在自我评估部分，总结实习中的亮点与不足，总结实习中的收获；在致谢部分，对实习中给予我帮助的人表达感谢。

关键词：数据预处理、半监督学习、XGBoost、聚类算法、关键词挖掘、命名实体识别、文本分类、脉络分析

ABSTRACT

In the construction and dissemination of Tencent Cloud, I mainly participated in two major projects:

1. Corporate Profile Index. We evaluate the risk, development, and activity of 250 million companies at various index levels, and then establish an automatic index evaluation mechanism. The preprocessing of data is the main challenge of index construction. Regarding the characteristics of extremely large data volume, discrete distribution, sparse features, etc., it is necessary to use discrete coding, feature selection, sparse matrix feature processing and other technologies for in-depth processing; considering the lack of labels, we will use semi-supervised learning algorithms, and the XGBoost learner is used to fit the features in order that the optimal function regression results applied to different index levels are obtained.

2. Opinion System Hot Topic Discovery. We aggregate similar news and analyze their main content. This program is mainly composed of two major steps: news aggregation and timeline summarization. Firstly, news aggregation is an important application of text clustering. We define topics as a collection of news, and we should discover hot topics. From the perspective of technique, we use a two-step clustering method to solve the problems of uncertain number of clusters and low efficiency in the calculation; Secondly, the timeline summarization model is used to obtain news chain on the timeline for a specific topic. This news chain can objectively show the development of topic. From the perspective of technique, we use DBSCAN for denoising, BERT similarity for calculation and Viterbi algorithm for news chain.

The above content will be displayed in detail in the section of internship content. In addition, the report also includes company and post introduction, narrative of the internship, self-evaluation and letter of thanks. In the section of company and post introduction, I will introduce Tencent Cloud and the post of big data algorithm application internship; in the section of narrative of the internship, I will review the internship responsibilities and experience; in the section of self-evaluation, I will summarize the gains and deficiencies during the internship; In the section of letter of thanks, I would like to express my gratitude to those who helped me during the internship.

Keywords: data preprocessing, semi-supervised learning, XGBoost, clustering algorithm, keyword mining, named entity recognition, text classification, timeline summarization

目 录

Chapter 1 Introduction of the facts

.1 Introduction of the company

Tencent is a world-leading internet and technology company that develops innovative products and services to improve the quality of life of people around the world. Founded in 1998 with its headquarters in Shenzhen, China, Tencent's guiding principle is to use technology for good. Our communication and social services connect more than one billion people around the world, helping them to keep in touch with friends and family, access transportation, pay for daily necessities, and even be entertained. Tencent also publishes some of the world's most popular video games and other high-quality digital content, enriching interactive entertainment experiences for people around the globe. Tencent also offers a range of services such as cloud computing, advertising, FinTech, and other enterprise services to support our clients' digital transformation and business growth. Tencent has been listed on the Stock Exchange of Hong Kong since 2004. A list of products of Tencent is shown in Figure 1:

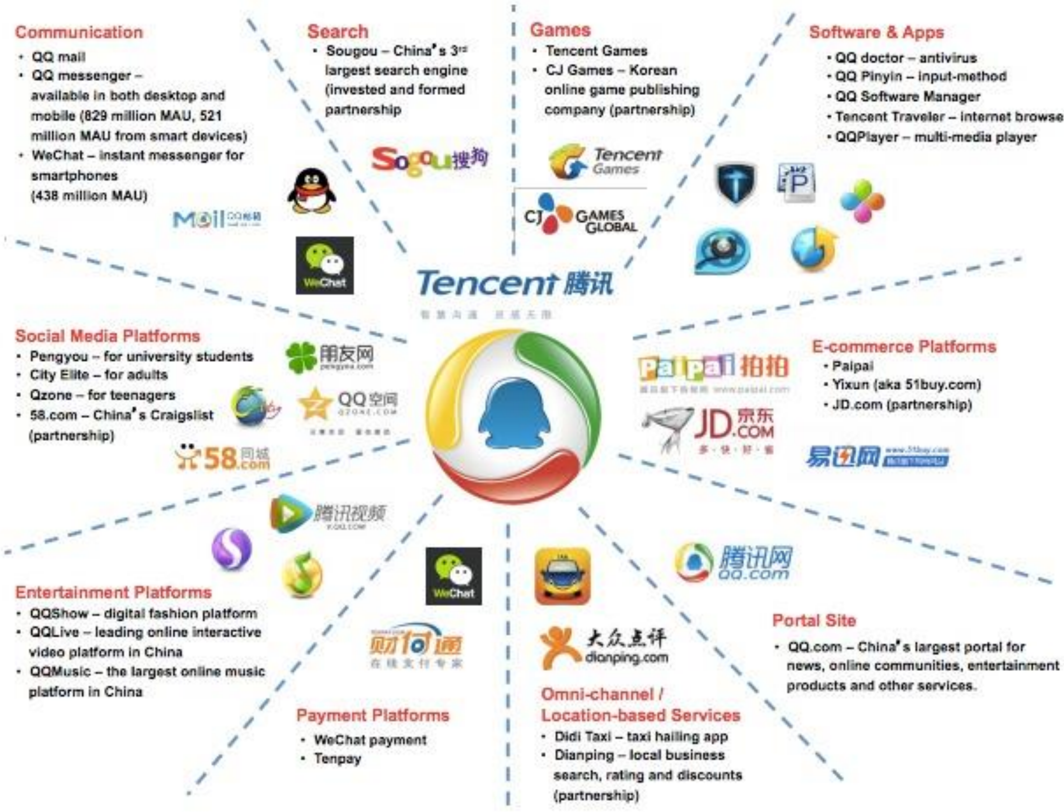

Communication
• QQ mail
• QQ messenger – available in both desktop and mobile (829 million MAU, 521 million MAU from smart devices)
• WeChat – instant messenger for smartphones (438 million MAU)
Search
• Sougou – China's 3rd largest search engine (invested and formed partnership
Games
• Tencent Games
• CJ Games – Korean online game publishing company (partnership)
Software & Apps
• QQ doctor – antivirus
• QQ Pinyin – input-method
• QQ Software Manager
• Tencent Traveler – internet browser
• QQPlayer – multi-media player
Tencent 腾讯
Social Media Platforms
• Pengyou – for university students
• City Elite – for adults
• Qzone – for teenagers
• 58.com – China's Craigslist (partnership)
E-commerce Platforms
• Paipai
• Yixun (aka 51buy.com)
• JD.com (partnership)
Entertainment Platforms
• QQShow – digital fashion platform
• QQLive – leading online interactive video platform in China
• QQMusic – the largest online music platform in China
Payment Platforms
• WeChat payment
• Tenpay
Omni-channel / Location-based Services
• Didi Taxi – taxi hailing app
• Dianping – local business search, rating and discounts (partnership)
Portal Site
• QQ.com – China's largest portal for news, online communities, entertainment products and other services.

Figure 1 List of products of Tencent

Tencent Cloud, a cloud computing brand built by Tencent Group, provides world-leading cloud computing, big data, artificial intelligence and other technology products and services to government agencies, corporate organizations, and individual developers in various countries and regions around the world. Excellent technological capabilities create a wealth of industry solutions, build an open and win-win cloud ecosystem, promote the construction of the industrial Internet, and help all industries to achieve digital upgrades.

Over the years, Tencent Cloud has based on the technical training of QQ, Qzone, WeChat, and Tencent's game real business. From infrastructure to refined operation, from platform strength to ecological capacity building, Tencent Cloud has integrated all mentioned above into the market to provide enterprises and entrepreneurs with a cloud service experience including cloud computing, cloud data, and cloud operation.

The changes that cloud computing has brought to IT and the entire commercial market are no longer empty talks. Traditional enterprises have been transformed in a fundamental sense in the cloud era. Large enterprises gain a steady stream of vitality in the cloud, and small and medium-sized enterprises can quickly face the market to obtain opportunities and development through the cloud. In the future, more companies will join the cloud world, and Tencent Cloud will strive to build a public cloud service platform with the highest quality and the best ecology. Let companies focus more on business and trust Tencent Cloud with the construction of the infrastructure.

Figure 2 Company Logo of Tencent Cloud

.2 Presentation of the post

The post of this internship is named "Tencent Cloud Big Data Application Algorithm". The main task is to be responsible for the analysis and mining of massive data, discover the value of data and support the application of B-end products. While given a detailed program, it is necessary to develop and implement algorithms to optimize the problem, which including but not limited to risk control, recommendation, graph calculation, NLP algorithm, etc. Programs, such as Construction of Corporate Profile Index and Public Opinion System Hot Topic Discovery, are conducted so as to disseminate the influence of Tencent Cloud's big data technology.

My post belongs to Application Group 2, Tencent CSIG and locates in Sigma Building, Zhichun Road, Haidian District, Beijing, which is nearby Beihang University. Our group consists of 15 staff in total, among which most work in Shenzhen and only three work in Beijing, including my supervisor, another colleague and me. The group leader mainly works in Shenzhen and sometimes comes to Beijing on business. We have colleagues who are responsible for product, back-end development or algorithm research separately while my post concentrates on algorithm research.

For work communication, I write a work report to my supervisor and group leader everyday and participate in the weekly meeting for information exchange and mission progress report. I also have a close connection with other colleagues for work issues.

.3 Presentation of the internship

The presentation of the internship will be unfolded in two aspects: Construction of Corporate Profile Index and Public Opinion System Hot Topic Discovery.

Firstly, in consideration of the Construction of Corporate Profile Index, we will analyze the status of 250 million companies in various aspects: risk, development, activity etc. by giving them evaluation scores with our big database. This problem can be considered as a regression problem. Despite numerous successful examples in the literature, the application of such optimization approach still faces some major challenges in big data field. There may be several questions:

1) As there exist massive discrete variables, how can we take advantage of these features in a fast way with billions of data? An efficient quantification method should be taken into consideration. With all the raw data, we are supposed to conduct smart data filtering and feature selection to reduce data size.

2) While the obtained data is sparse, will the performance of our model be interrupted? A useful approach is use matrix factorization or low-rank matrix completion algorithm for data preprocessing.

3) How can we build a model with few labeled datasets? Semi-supervised methods lead a better way to train our model.

Secondly, Public Opinion System aims at discovering hot topic within numerous news and then generating a timeline development context by reasoning their logic with other historic news. We will resolve this problem by building a pipeline. In research and engineering, this kind of work is rare and challenging. There may be several questions:

1) when is the result updated? Although news library is updated every moment, there is no need to calculate results instantly. Instead, we pull news data every two hours and conduct calculation.

2) How should we define a hot topic? To the needs of our product, we consider a hot topic as a clustering of news. Such a topic must have at least a certain amount of news at a period and should have some topic information such as topic name, summary, keywords to give clients a total impression of all the news included. Thus, clustering approaches should be developed.

3) How should we define a topic name, summary and keywords? For generating topic name or summary, there should be some abstractive or extractive algorithm. As for the keywords, perhaps Named Entity Recognition (NER) or other supervised approaches can be helpful.

4) How can we generate a timeline development context? We can consider this as a Timeline Summarization[1] (TLS) task. We need to model the time series information of the input news and summarize important news in chronological order inside a specific topic.

.4 Plan of the report

(1) Introduction of the facts

This section presents the company I worked for, Tencent Cloud. In addition, this part introduces the main job of an algorithm engineer and gives general information about the Tencent Cloud.

(2) Narrative of the internship

This part presents all kinds of tasks performed as part of my internship, including the engineering part, the algorithmic part and the daily communication part.

(3) Internship content

This section gives an account of information about the work I did during my internship. The details of two programs, Corporate Profile Index and Public Opinion System Hot Topic Discovery will be unfolded. In each part, the context of problem will be presented. Following the theoretical base, the result of benchmarks will be presented and analyzed along with the modification and improvement that has been made during the internship.

(4) Self-evaluation

This section analyzes the good and bad work I did and gives a summary of the things I learned during my internship.

(5) Letter of thanks

This part is a letter of thanks to the company and university.

Chapter 2 Narrative of the internship

During my internship, I was involved in the development of two grand programs: Corporate Profile Index and Public Opinion System Hot Topic Discovery. I investigated all the works including the theoretical bases, the used tools and communicates with colleagues for deep cooperation before writing code and conducting experiments.

.1 Corporate Profile Index

First of all, I will introduce Corporate Profile Index program, which lasts two months in the beginning of my internship. I will show the enrollment of my task in three aspects: the algorithm, the engineering, and the daily communication.

From the perspective of algorithm, I was responsible for investigating and implementing the most advanced regression algorithm related to corporate profile. Regarding the data preprocessing, a pipeline is built, including discrete type encoding, data filtering, feature selection and sparse feature processing. I implemented and compared the discrete type encoding methods such as one-hot encoding, WoE encoding and target encoding for discrete variables. In order to reduce data size and resolve overfitting problem, I successively employed a rule-based data filtering approach and a clustering-based feature selection approach. Also, given that our data is sparse, I investigated some sparse feature processing approaches and finally decided to use a deep learning approach to solve this problem. Regarding the model, I separately used the score card model and XGBoost model and applied them for the regression problem. Finally, for the lack of label, I deployed the PU-learning which has labeled dataset and unlabeled dataset as input.

From the perspective of engineering implementation, I participated in the deployment of various modules of the automatic updating system. I successfully built a Daily update mechanism by using the Airflow tool to renew the Corporate Profile Indexes. This is realized by daily pulling newest datasets, training models, saving results to HDFS, Hive and ElasticSearch successively.

From the daily communication perspective, I mainly communicate with my supervisor, Guohua Wang and one of my colleagues, Zefeng Weng on this project and I get along well with them. When I encounter a difficult problem relative to the algorithms, I cooperate with Guohua;

if the problem is relative to Big Data tools, I ask Zefeng for help. I appreciate their selfless help and I am glad to help them whenever they need. We all feel lucky to help each other and eager to process together.

.2 Opinion System Hot Topic Discovery

Secondly, I will present Public Opinion System Hot Topic Discovery program. In condition that this program is still under development and has not yet been launched, I will only show my task in two steps: the algorithm and the daily communication.

From the perspective of algorithm, I mainly focus on proposing a new version based on the existing system to improve effectiveness. The existing system is designed by some of my colleagues but still encounters some problems, including clustering error, topic information incorrectness etc. Until then, how to generate a timeline development context for every topic was not decided yet. Thus, my job is to continue on the previous version by improving clustering effects and accuracy of topic information as well as developing a newest approach to generate timeline development contexts. Firstly, for improving clustering effects, I designed a two-step clustering algorithm that not only is not restricted to the number of clusters but also has a high efficiency. Secondly, for reducing noise, I designed a classifier as well as building a dataset to train this classifier to distinct noise data with other data. Thirdly, for improving accuracy of topic information such as topic name and topic keywords, I used Longest Common Subsequence (LCS) and Named Entity Recognition (NER) models to separately solve these two problems. Finally, in order to reason the logic of every topic, I designed an algorithm that generates a timeline development context as a solution.

From the perspective of daily communication, this project is also mainly supervised by my supervisor, Guohua Wang. In addition, this project contains knowledge related to Natural Language Processing so that I ask Xiusen Gu for help who is also specialized in this domain.
In detail, Guohua offers me suggestions about the construction of TLS system and ask me to write survey on the relative state-of-the-art papers. Xiusen offers me existing codes about online NER models and keywords extraction models and we discuss their pros and cons. Thorough mutual communication, we gain knowledge and improve efficiency.

Chapter 3 Internship Content

This section gives detailed information about the work I did during my internship. I will present Corporate Profile Index in three aspects: algorithm, engineering and daily communication followed by Opinion System Hot Topic Discovery which I will present in two aspects: algorithm and daily communication.

.1 Corporate Profile Index

The project Corporate Profile Index is essential to our team because it is capable to analyze all companies in various aspects to help us understand our data. What I should do is to give continuous technique support and update our product in every life cycle. I succeeded in implementing two versions iteratively and deployed one of the versions online. Figure 3 shows the homepage of our product.

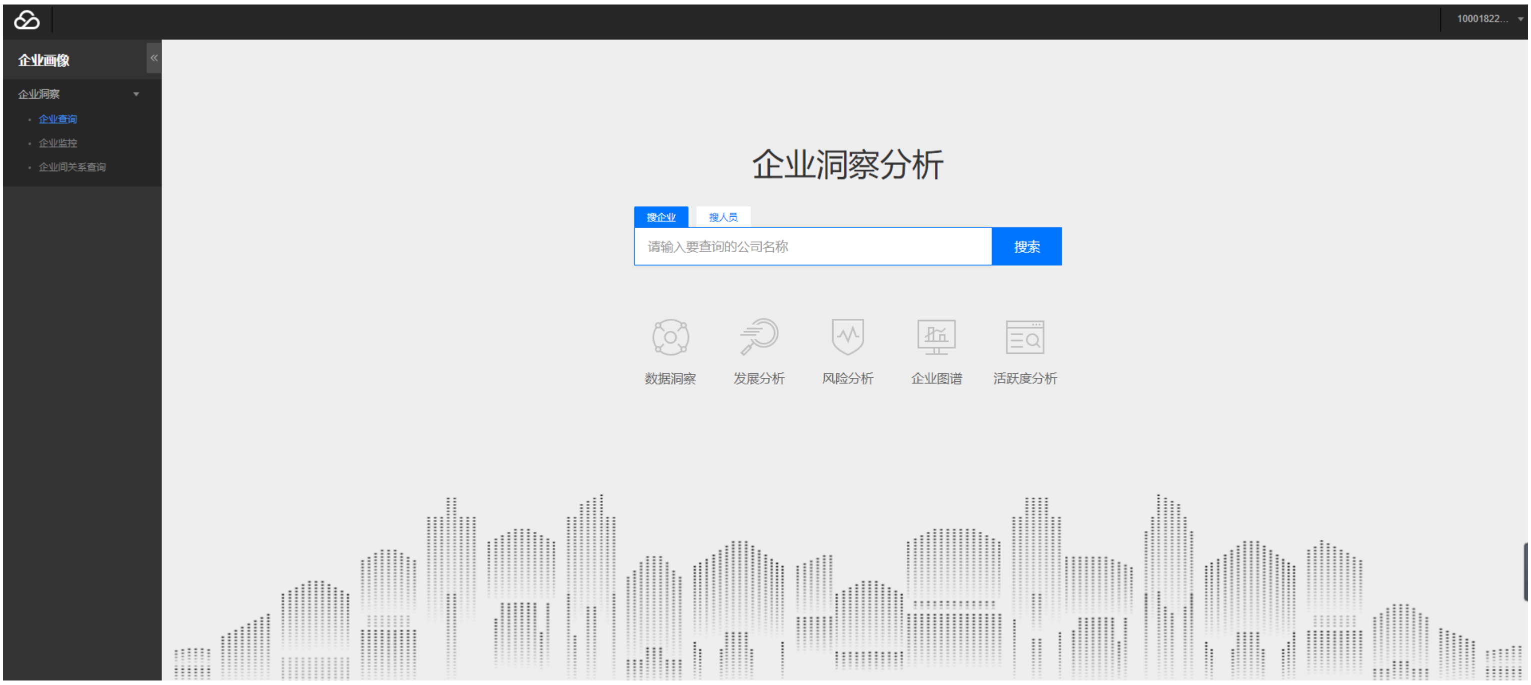


Figure 3 Homepage of our product Corporate Profile Index

.1.1 Algorithm

In the beginning, my algorithm for evaluating different aspects of a company by predicting index scores is the score card model. This model is already deployed online. The essential of this model is to assign weights to every attribute, which is done by experts. Besides, I assume that the score distribution is Gaussian so that we should adjust the score distribution. I use Box-

Cox Transformation to constrain the scope of data. The mathematical expression of Box-Cox Transformation is shown in the following:

$$y_i^{(\lambda)} = \begin{cases} \dfrac{y_i^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0, \\ \ln(y_i) & \text{if } \lambda = 0, \end{cases}$$

where λ is the coefficient of transformation. I use Grid Search to obtain the optimal λ, which should maximize the log-likelihood function.

Secondly, for iterative update, I designed a semi-supervised algorithm by using XGBoost model to build more robust scoring system. The algorithm is mainly composed of four steps: feature preprocess, label construction, model construction and evaluation.

Feature Preprocess

Regarding that our data are complex and difficult to analyze, I design a pipeline to preprocess data.

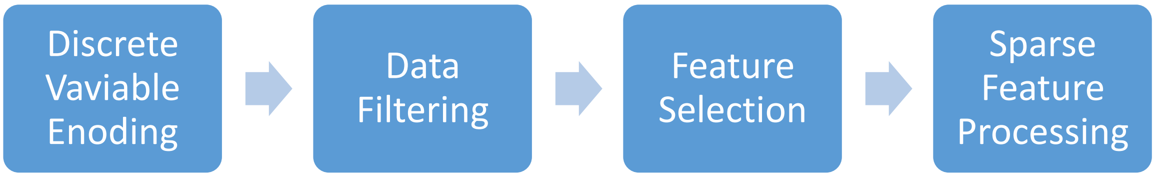
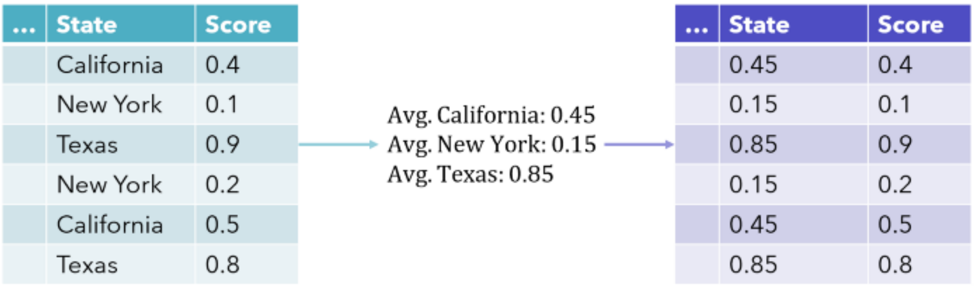


Figure 4 Main steps of feature preprocess

The first step is Discrete Variable Encoding. Among all possible methods that I investigate, I mainly compare one-hot encoding, WoE encoding and target encoding. Considering that one-hot encoding may vastly increase the size of feature dimension and is hard to optimize, or that WoE encoding demands label which is not accessible in our case, we finally choose target encoding. It is worth mentioning that time complexity and space complexity of target encoding are both low. An illustration is drawn in Figure 5:



...	State	Score
	California	0.4
	New York	0.1
	Texas	0.9
	New York	0.2
	California	0.5
	Texas	0.8

Avg. California: 0.45
Avg. New York: 0.15
Avg. Texas: 0.85

...	State	Score
	0.45	0.4
	0.15	0.1
	0.85	0.9
	0.15	0.2
	0.45	0.5
	0.85	0.8

Figure 5 Target encoding example

Secondly, we will do Data Filtering. We have plenty of data to describe big companies while the data of small companies, such as restaurants and markets, are hard to access. Therefore, data are always dense for the former and sparse for the latter. In our algorithm, sparse data usually leads to unstable solution space and can cause overfitting problem. To this end, I filter data that are sparse in the feature space. In other words, small companies are usually ignored.

Thirdly, we will operate Feature Selection. In case that we have a quantity of features, feature selection is crucial to avoid overfitting. In our case, I choose Silhouette Coefficient to select feature. Silhouette Coefficient is a supervised algorithm that can analyze the result of clustering. For labels, I use the rules to explicitly define good companies and bad companies. All that have high valuation, high registered capital and at least one branch are good companies. On the other hand, all that are in either suspension, liquidation, cancellation, or revocation are bad companies. A specific feature that can significantly separate the two clusters has high Silhouette Coefficient. The mathematical expression is shown in Figure 6:

$$s(i)=\frac{b(i)-a(i)}{\max\{a(i),b(i)\}} \qquad s(i)=\begin{cases}1-\dfrac{a(i)}{b(i)}, & a(i)<b(i)\\ 0, & a(i)=b(i)\\ \dfrac{b(i)}{a(i)}-1, & a(i)>b(i)\end{cases}$$

Figure 6 Silhouette Coefficient

where $a(i)$ represents the average value of dissimilar degree from sample i to other points in the same cluster and $b(i)$ represents the minimum value of the average dissimilar degree from sample i to other clusters.

Finally, we will do Sparse Feature Processing. I investigate two kinds of approaches: Matrix Factorization and Low-Rank Matrix Completion. Considering time complexity, I decide to use matrix factorization. A representative algorithm is named Latent Factor Model[2] (LFM). Given a company-feature matrix R, our objective is to fill the unknown values. LFM splits matrix R into two sub-matrices: company-latent matrix P and latent-feature matrix Q so that R = P * Q. Note that the latent dimension is smaller than the company/feature dimension. We construct a L2 loss function that calculates the L2 difference of every visible value:

$$E\left(p_u, q_i\right) = \sum_{(u,i)\in\kappa} \left(r_{ui} - q_i^T p_u\right)^2$$

We can also add a regular term to two sub-matrices:

$$E\left(p_u, q_i\right) = \sum_{(u,i)\in\kappa} \left(r_{ui} - q_i^T p_u\right)^2 + \lambda\left(\left\|q_i\right\|^2 + \left\|p_u\right\|^2\right)$$

Finally, we update parameters using gradient descent:

$$\frac{\partial E}{\partial p_u} = -2q_i \cdot e_{ui} + 2\lambda p_u$$

$$\frac{\partial E}{\partial q_i} = -2p_u \cdot e_{ui} + 2\lambda q_i$$

After we learn the two sub-matrices, we reconstruct R using P * Q to predict the unknown values.

Label Construction

We use rules to construct labels. There are two main principles to choose:

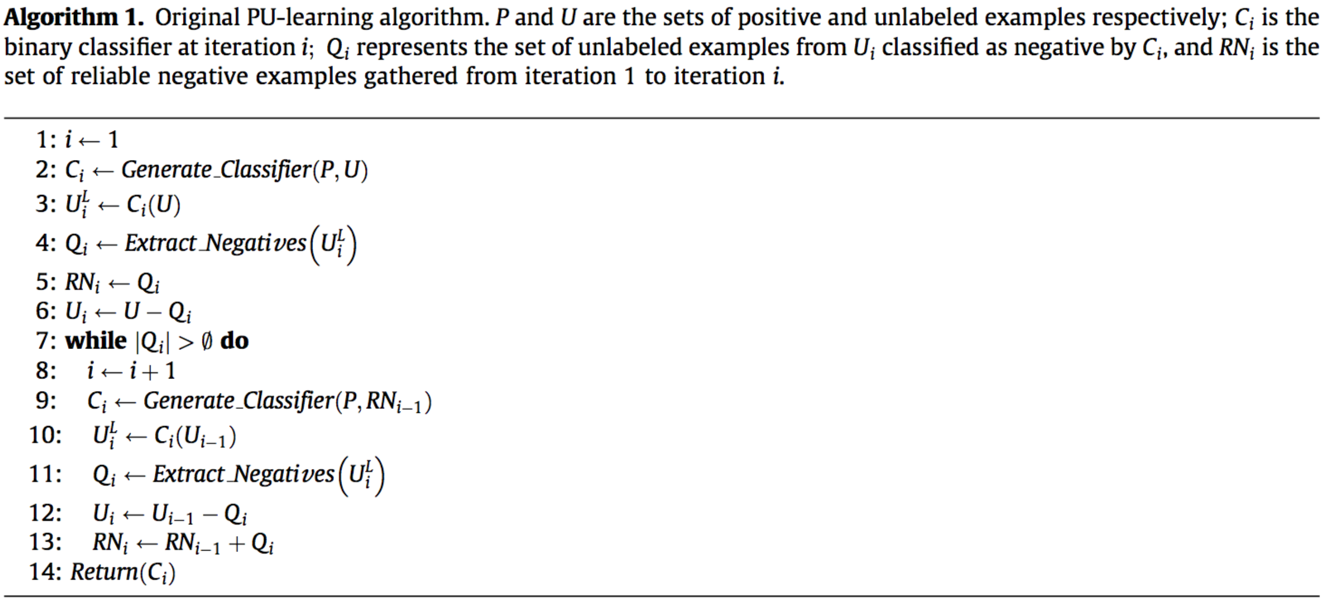
1) Only data with high credibility are labelled. In this case, I label 1 thousand data out of 250 million data.

2) Only positive labels are considered. Owing that negative data are hard to detect and positive data are, on the other hand, performant in every aspect and then easier to detect, we only label positive data. Afterwards, we will use semi-supervised algorithm.

Model Construction

For iterative update, we consider an innovative semi-supervised algorithm: PU Learning[3][4] along with XGBoost[5].

PU Learning aims at using only positive-labeled data and unlabeled data to train classifiers. I use two-step strategy of PU Learning: The first step of PU learning is to identify a set of reliable negative documents (set RN) from the unlabeled set U; and second step of PU leaning is then to build a classifier using positive set P, reliable negative set RN and remaining unlabeled set U' (U'=U-RN). Iteratively, we can build numerous classifiers and average their predict scores on unlabeled set as outputs. The mathematical expression of the algorithm is shown in the Algorithm 1:

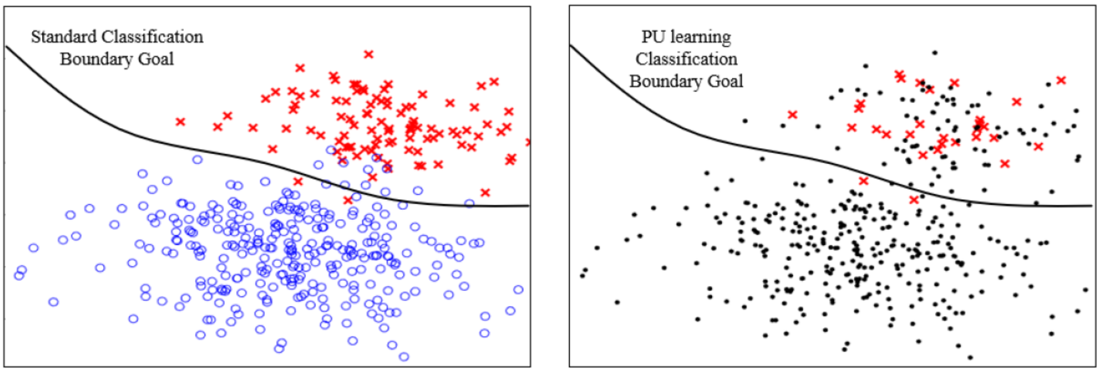


Algorithm 1. Original PU-learning algorithm. P and U are the sets of positive and unlabeled examples respectively; C_i is the binary classifier at iteration i; Q_i represents the set of unlabeled examples from U_i classified as negative by C_i, and RN_i is the set of reliable negative examples gathered from iteration 1 to iteration i.

```
$i \leftarrow 1$
$C_i \leftarrow Generate\_Classifier(P, U)$
$U_i^L \leftarrow C_i(U)$
$Q_i \leftarrow Extract\_Negatives\left(U_i^L\right)$
$RN_i \leftarrow Q_i$
$U_i \leftarrow U - Q_i$
while $|Q_i| > \emptyset$ do
    $i \leftarrow i + 1$
    $C_i \leftarrow Generate\_Classifier(P, RN_{i-1})$
    $U_i^L \leftarrow C_i(U_{i-1})$
    $Q_i \leftarrow Extract\_Negatives\left(U_i^L\right)$
    $U_i \leftarrow U_{i-1} - Q_i$
    $RN_i \leftarrow RN_{i-1} + Q_i$
$Return(C_i)$
```

Algorithm 1 PU Learning

An illustration of PU Learning can be depicted in the Figure 7:



(a) Supervised classification learning problem with all labels known.

(b) Positive Unlabeled learning problem with only a percentage of known positive labels.

Figure 7 Illustration of PU Learning. This figure is captured in [4].

Evaluation

Since we lack labels for our 250 million data, it’s hard to evaluate results. Thus, we ask experts to establish a set of authoritative score card weights, which plays an important role in score card weights and can be regarded as labels to test our PU Learning approach. Here we have a weight table example in Table 1:

tag_table_name	tag_name	lower_bound	upper_bound	score
graph_prop	eid_investment_moneys	\	0	0.05
graph_prop	eid_investment_moneys	0	1000	0.5

Table 1 An example of score card weights

I also visualize the classification results of good and bad companies to evaluate models. Figure 8 shows that by PU Learning, we obtain more good-labeled data surrounded by the original good samples.

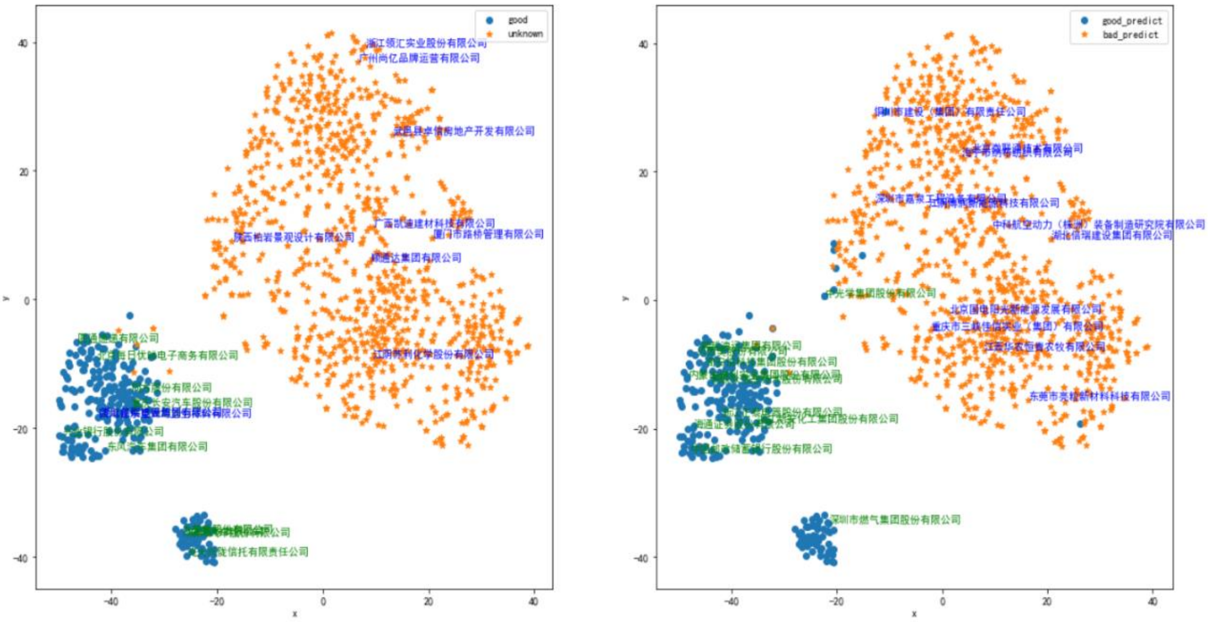
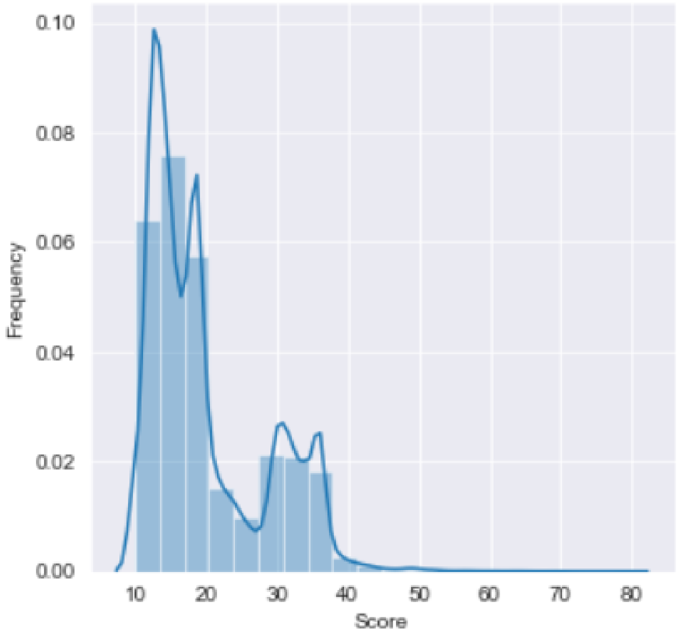


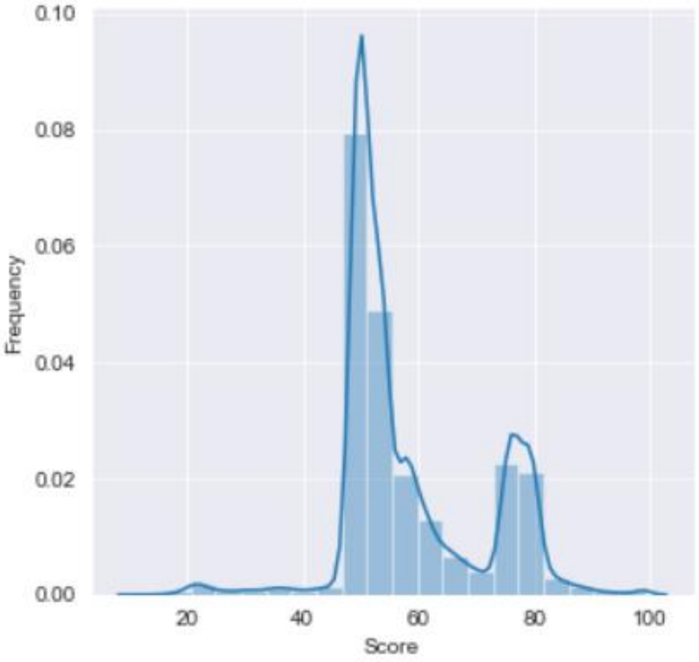
Figure 8 Visualization of PU Learning results. In the figures, every point represents a company data. The left figure depicts the distribution of labels (good and unknown data) while the right figure depicts the distribution of predict results (good predictions and bad predictions)

It is worth mentioning that different customers have different opinions on companies, a unified standard may not be satisfactory in all cases. Thus, our model needs to be interpretable.

To this end, we either compare weights given by experts with our result or examinate the coefficients of features.

Finally, I succeeded in calculating Corporate Profile Index in various aspects. Figure 9 separately shows the distribution of development index, the risk index and the activity index of all companies.





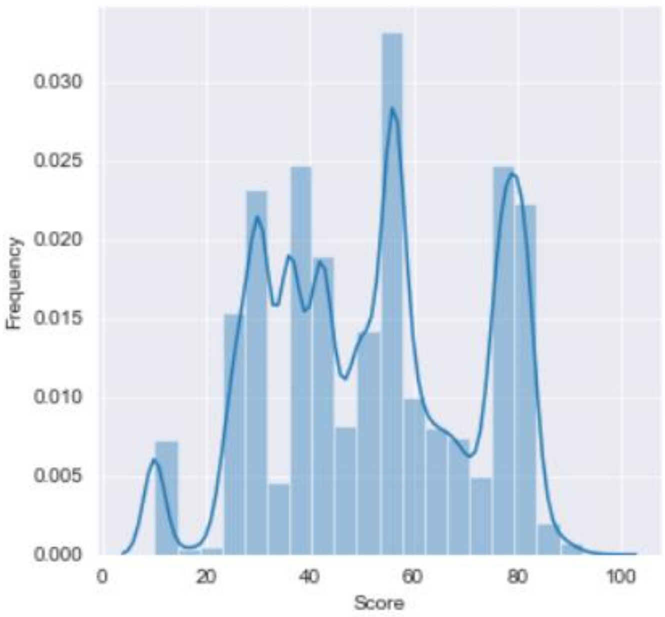


Figure 9 Index distributions of development (up), risk (middle) and activity (down).

We can find that the index distribution of activity is not smooth because the used features are relatively sparser.

.1.2 Engineering

The engineering task I was involved in during my internship was the automatic update mechanism that daily integrates all the python and shell scripts including pulling and preprocessing data, training model, calculating profile indexes, saving to HDFS, Hive, ElasticSearch successively by Airflow.

Figure 10 Logo of Apache Airflow

Airflow is a tool for automating and scheduling tasks and workflows that can be repeated on a regular basis. I deployed a daily process pipeline described as the following:

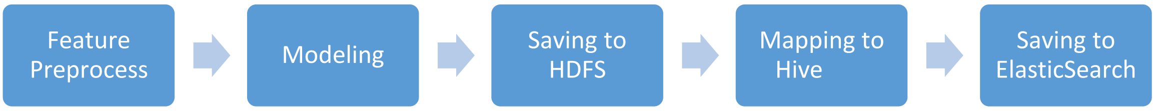


Figure 11 Daily process pipeline of the program Corporate Profile Index

There are five main steps in my pipeline. I will present every step in detail.

Feature Preprocess

In this step, I write a python script to pull and preprocess data. As described before, the data preprocessing is crucial and should be considered carefully before model training. While the data size is tremendous, big data framework should be considered. I preprocess data using Apache Spark, which is a lightning-fast cluster computing technology based on Hadoop MapReduce to efficiently compute data.

Figure 12 Logo of Spark

Modeling

In this step, I write a python script which builds the model (either the score card model or the XGBoost model), trains the model with processed data and outputs predict index scores. The main principles are presented in the above.

Saving to HDFS

HDFS is a distributed filesystem that allows us to store data across multiple machines or nodes in a cluster and allows multiple users to access data. Thus, the huge output results can be saved to HDFS in an efficient way.

Figure 13 Logo of HDFS

Mapping to Hive

Hive is a data warehouse infrastructure tool to process structured data in Hadoop. I build a Hive external table to map the HDFS structured data file, so that there is no need to repeat storage. What's more, the Hive external table points to HDFS, which will be updated as HDFS is updated, and can provide a complete SQL query language (HQL) to process data.

Figure 14 Logo of Hive

Saving to ElasticSearch

The eventual objective aims at putting index scores on ElasticSearch, which is the distributed search and analytics engine at the heart of the Elastic Stack and provides near real-time search and analytics for all types of data, for that our development colleagues can consult index scores easily and efficiently and then develop products.

Figure 15 Logo of ElasticSearch

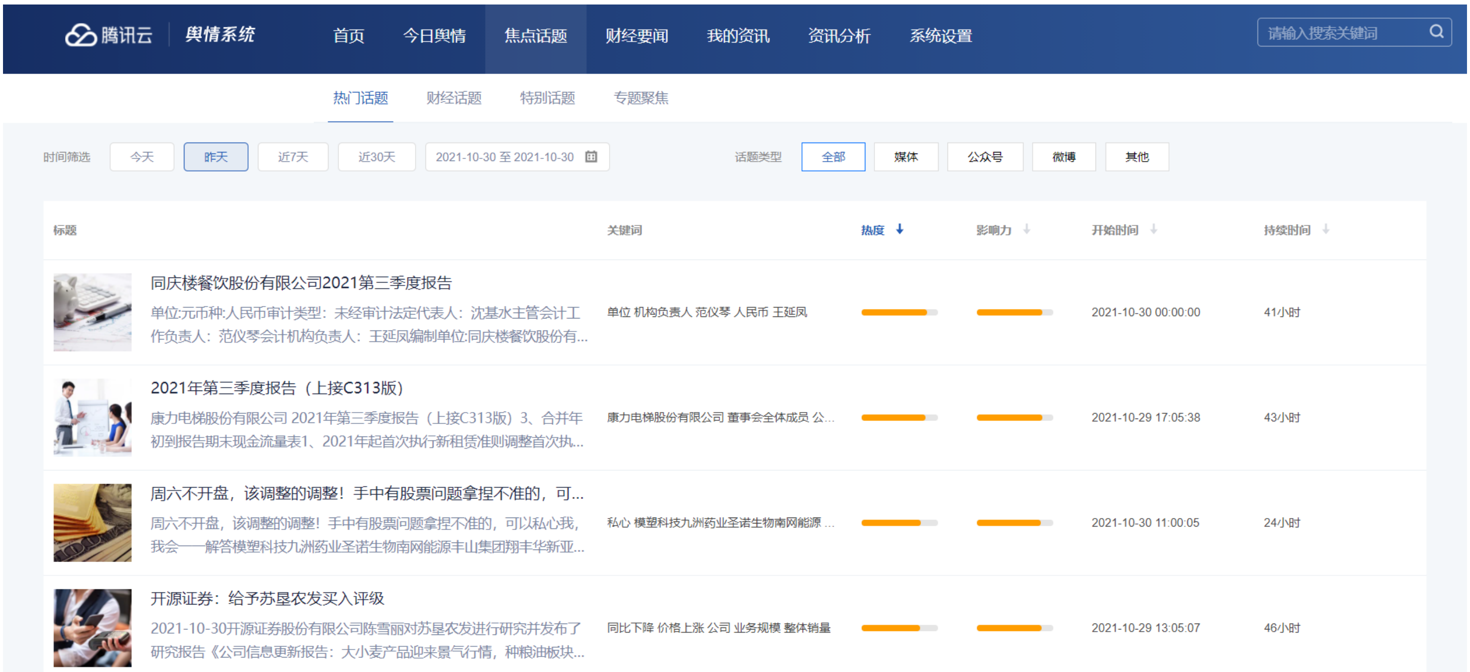
.1.3 Daily Communication

From the perspective of daily communication, I get along well with my supervisor, Guohua Wang, and actively ask him when I encounter a technique problem. He offers me some recommendations according to his knowledge base and experience, which is proven to be kind of enlightenment.

Furthermore, before I started my internship, I merely had the opportunity to be involved with Spark, HDFS, Hive, ElasticSearch and Airflow. Thus, I had trouble using these big data tools and thanks to one of my colleagues, Mango Weng, who gave me advice to learn them, I achieved a satisfactory result and put it online.

.2 Opinion System Hot Topic Discovery

Hot Topic Discovery is one of the main components of our Opinion System. Out of thousands of news, we should detect the hot topic and analyze their context to help our clients understand what’s happening in the world. We mainly focus on financial domain. Following one of my colleagues, I continue to improve the performance of the Hot Topic Discovery. Figure 16 shows the homepage of our product Hot Topic Discovery.



标题	关键词	热度	影响力	开始时间	持续时间
同庆楼餐饮股份有限公司2021第三季度报告 单位:元币种:人民币审计类型：未经审计法定代表人：沈基水主管会计工作负责人：范仪琴会计机构负责人：王延凤编制单位:同庆楼餐饮股份有...	单位 机构负责人 范仪琴 人民币 王延凤			2021-10-30 00:00:00	41小时
2021年第三季度报告（上接C313版） 康力电梯股份有限公司 2021年第三季度报告（上接C313版）3、合并年初到报告期末现金流量表1、2021年起首次执行新租赁准则调整首次执...	康力电梯股份有限公司 董事会全体成员 公...			2021-10-29 17:05:38	43小时
周六不开盘，该调整的调整！手中有股票问题拿捏不准的，可... 周六不开盘，该调整的调整！手中有股票问题拿捏不准的，可以私心我，我会一一解答模塑科技九洲药业圣诺生物南网能源丰山集团翔丰华新亚...	私心 模塑科技九洲药业圣诺生物南网能源 ...			2021-10-30 11:00:05	24小时
开源证券：给予苏垦农发买入评级 2021-10-30开源证券股份有限公司陈雪丽对苏垦农发进行研究并发布了研究报告《公司信息更新报告：大小麦产品迎来景气行情，种粮油板块...	同比下降 价格上涨 公司 业务规模 整体销量			2021-10-29 13:05:07	46小时

Figure 16 Homepage of our product Opinion System Hot Topic Discovery

.2.1 Algorithm

I invent a pipeline to solve this problem. What I design is based on the previous version. I will present every step in the pipeline in detail and make it clear about my contributions. The main steps are depicted in Figure 17:

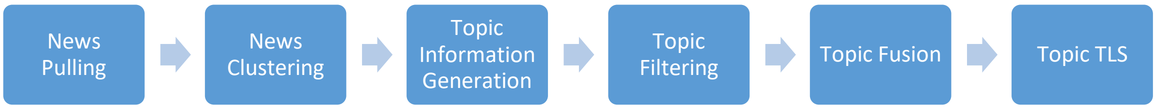


Figure 17 Pipeline of the Opinion System Hot Topic Discovery project

News Pulling

The original news data are stored in ElasticSearch. Although data are updated every moment, there is no need to calculate results instantly. Instead, we pull news data from ElasticSearch every two hours and conduct calculation.

News Clustering

The previous news clustering approach is Principal Component Analysis (PCA) for decomposition and K-Means for clustering. Nevertheless, the existing solution needs to set a hyperparameter K, which is unreasonable in our case where data stream is sometimes big and sometimes small so that the number of clusters should be not fixable. Furthermore, PCA and K-Means may pass loss from one to the other, which may increase loss. To this end, I invented a two-step clustering approach:

1) Non-Negative Matrix Factorization (NMF). NMF is much better than PCA in our case, for that NMF can operate decomposition and clustering at the same time, which reduces loss transmission. Also, NMF is a soft-clustering approach that obtains the probability of every input allocated to a specific cluster. This is logical since an input news data may belong to different topics with high probabilities. [6] demonstrates shows that NMF performs better than K-Means in public dataset.

2) Single-Pass Algorithm (SP). SP is a greedy algorithm that in influenced by the input order. SP takes outputs of NMF as inputs and do re-clustering inside every cluster result of NMF. The biggest convenience of SP is that it is not restricted to the number of clusters. Furthermore, it can be run in parallel in each cluster result of NMF to make it faster.

After two-step clustering, we obtain clustered news data, each cluster of which can be considered as a topic.

Topic Information Generation

After aggregating news into topics, I generate topic information including topic name, representative keywords and entities that can describe every clustering result. To be specific:

1) Topic name: I use the Longest Common Subsequence (LCS) algorithm. Given a set of news data with their titles, we can consider the LCS among all the titles as topic name. For example, if we have three news data in a cluster with their separate titles “武汉经开区一人未履行防控责任被拘”, “武汉经开区一企业责任人未依法履行防控责任被拘留” and “流调中未如实报告工人去向 武汉一企业责任人被拘留”, we obtain “武汉一企业责任人被拘” as the topic name by LCS (shown in Figure 18).

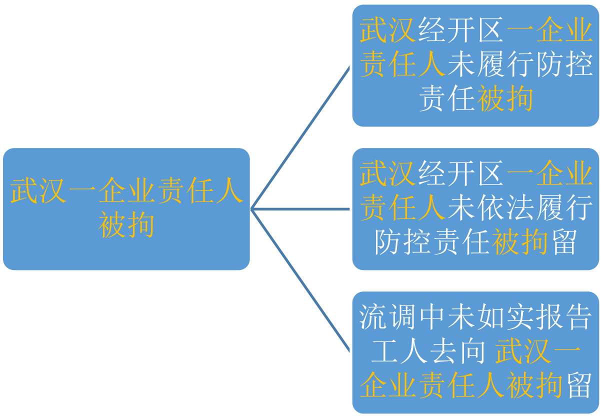

武汉一企业责任人被拘
武汉经开区一企业责任人未履行防控责任被拘
武汉经开区一企业责任人未依法履行防控责任被拘留
流调中未如实报告工人去向 武汉一企业责任人被拘留

Figure 18 An example of the extraction of topic name out of three news titles

2) Keyword extraction. Keywords based on news and public opinion include entities and their related attributes, financial events, etc. These keywords represent the core content of the entire topic and can be extracted from the news titles and contents inside every cluster. In the process, we use an algorithm named TopicRank[7]. The process is shown in Figure 19. I will present the process in detail.

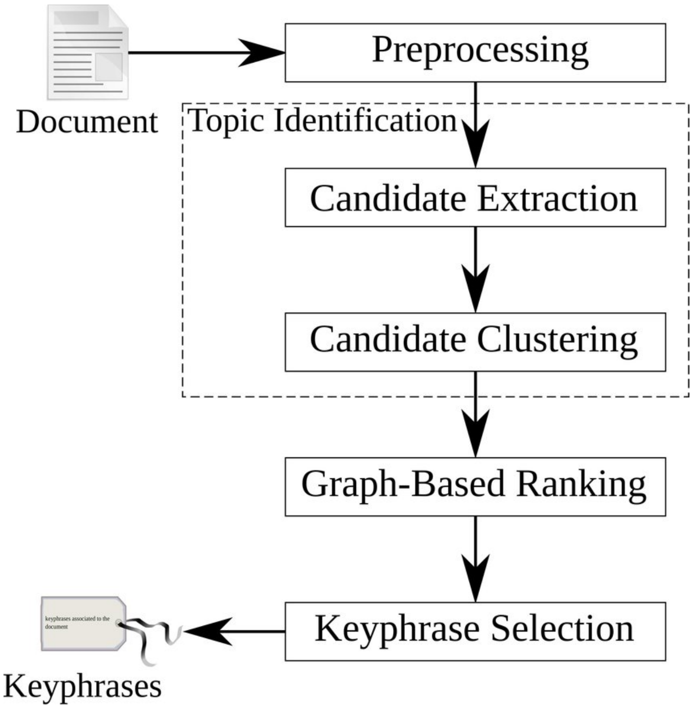


Figure 19 TopicRank process

Firstly, we should do topic identification according to the preprocessed data. We extract noun phrases to characterize the topic of the document. Words in the phrase that overlap more than 25% are considered as similar phrases, and the Hierarchical Agglomerative Clustering (HAC) algorithm is used to cluster similar phrases.

Secondly, we construct graph. let G = (V, E) be a complete and undirected graph where V is a set of vertices and the edges E a subset of $V \times V$. Vertices are topics and the edge between two topics t_i and t_j is weighted according to the strength of their semantic relation. t_i and t_j have a strong semantic relation if their keyphrase candidates often appear close to each other in the document. Therefore, the weight $w_{i,j}$ of their edge is defined as follows:

$$w_{i,j} = \sum_{c_i \in t_i} \sum_{c_j \in t_j} \text{dist}(c_i, c_j)$$

$$\text{dist}(c_i, c_j) = \sum_{p_i \in \text{pos}(c_i)} \sum_{p_j \in \text{pos}(c_j)} \frac{1}{|p_i - p_j|}$$

Thirdly, we run PageRank[8] algorithm in the graph until convergence to find the core topics which contain keyphrases.

Finally, once the topics are sorted, the top K topics are selected, and each topic selects one of the most important key phrases as output, and all topics generate top K keyphrases in total.

An illustration of TopicRank is depicted in Figure 20.

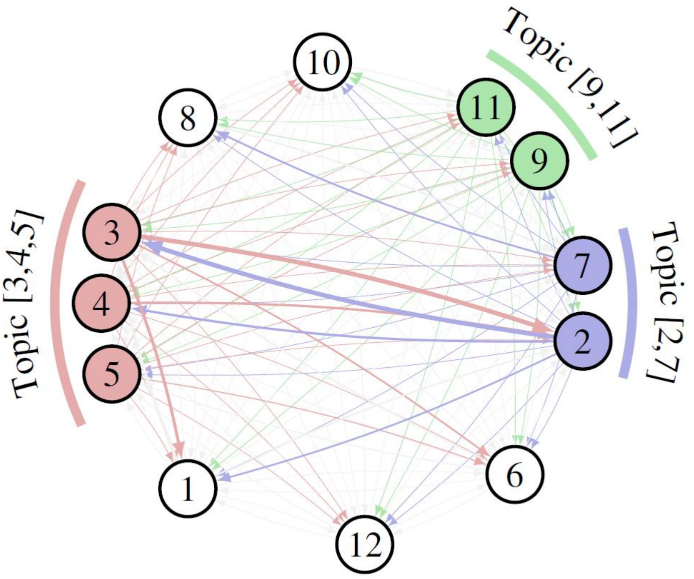


Figure 20 Illustration of keyword extraction algorithm: TopicRank

Note that this algorithm is already implemented by one of my colleagues and I can directly deploy it. Extracted keywords are crucial for Topic TLS afterwards.

3) Named Entity Recognition. NER refers to extracting company entities and related information from various public opinion texts and unstructured texts, which may tell readers who (persons, governments and organizations), what (companies and products), when (time),

where (places) the texts are related to. Faced with this problem, we consulted the latest papers on entity recognition. After comparing BiLSTM-CRF, ID-CNN-CRF[11], Lattice[10], BERT[9] and other models, we decided to use the Lattice model for deployment.

The Lattice model simultaneously labels characters based on word information and character information, and selects entities such as companies, names, and institutions. Compared with CRF, the Lattice model is based on word vectors and word vectors, and extracts text information from the semantic level. It analyzes the text in the word granularity and word granularity at the same time so that it controls the entity boundary in a better way; compared with the BERT model, the Lattice model has a smaller volume and is more convenient to deploy. In the experimental environment, the training accuracy of the Lattice model reached about 85%, which was much higher than the experimental performance of CRF 65%, so we finally decided to use the Lattice model.

However, we found that Lattice processed much slower in Kafka and caused a serious data backlog due to lack of online GPU resources. Therefore, upon investigating an improved model based on Lattice[12] that merges word-level information into character-level information and performs a unified encoding, which operates 10 times faster than before. Finally, we decided to use this model. The illustration of the model is depicted in Figure 21:

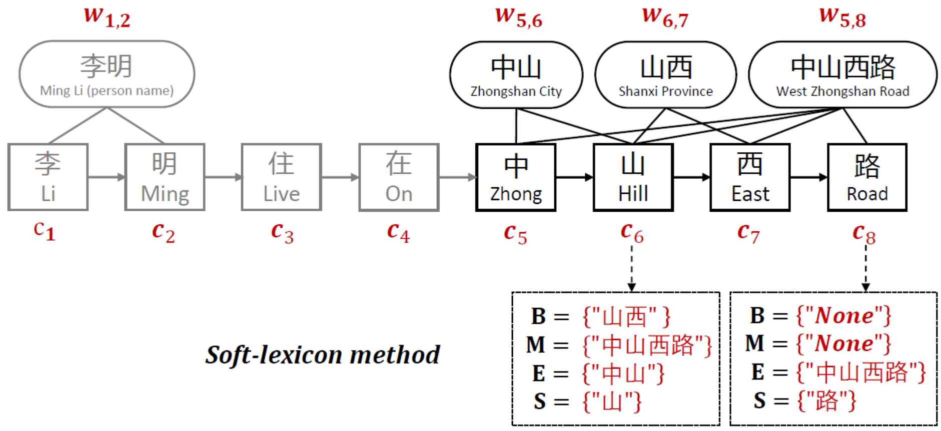


Figure 21 Illustration of improved Lattice[12]

Note that the codes are available at https://github.com/v-mipeng/LexiconAugmentedNER. What I mainly have done is preprocessing our own NER data, augmenting data embedding and deploying the improved Lattice model. Regarding the data preprocessing, we have 10,000 long news data with average length equal to 1800. The named entities and the relationship between entity pairs are labeled. There are 6 kinds of labels. Table 2 shows the statistic results of labels.

Label Classification	Government	Organization	Company	Product	People	Region
Total Number	20476	22637	215277	32697	68959	136049
Average Length	6.8	10.2	8.6	7.2	2.7	2.2

Table 2 Statistics of labels

I have compared results within two different labeling methods and within two different embedding methods. BIO and BMES are two typical ways for NER labeling. According to [12], the char and word embeddings are trained by Gigaword Chinese Tree Bank dataset (we note embedding method A). However, we can also improve pre-trained knowledge by using ERNI[13] as char embedding and Tencent AI Lab embedding[14] as word embedding (we note embedding method B). The results are shown in Table 3.

Labeling Method	BIO	BMES	BIO	BMES
Embedding Method	A	A	B	B
P	0.861	0.894	0.868	0.905
R	0.863	0.880	0.856	0.899
F1	0.862	0.887	0.862	0.900

Table 3 Evaluation results of improved Lattice on our own data

We find that using BMES plus embedding method B achieves the best score. This is probably because that BMES has a stronger ability to divide the boundary and improved embedding contains richer semantic and syntactic knowledge with larger pre-trained datasets.

Eventually, we extracted entities, including persons, governments, organizations, companies, places, products and time from every topic, which may be extremely helpful for Topic TLS afterwards.

Topic Filtering

In this section, we mainly aim to filter out noise data by classifier. I chose Naïve Bayes Classifier to split noise data with other data. In fact, Naïve Bayes Classifier is based on a strong hypothesis and is easy enough, which may result in fast computation but moderate performance.

However, in this step, we only want to filter out noise data such as advertisement and personal talks, which may be obviously different from other data in grammar, semantics or wording and thus are easy to separate. We have 2000 labeled datasets to train our model and the test accuracy reached about 90%.

Topic Fusion

Due to the fact that we pull data and operate calculation every two hours, the timeline is segmented. Thus, similar topics may appear both in historic and current operations. This section targets at merging historic and current topic if they are similar than a threshold to ensure that the newest topics are different from that in the database.

Topic TLS

Unlike traditional document summarization, timeline summarization[1] (TLS) needs to model the time series information of the input events and summarize important events in chronological order. However, in our case, instead of making abstractive summarization out of news data, we aim to extract news data with their structured information out of related data in chronological order so that clients can click on the extracted data to view details. To tackle this challenge, we propose a framework which can be described in the following[1]:

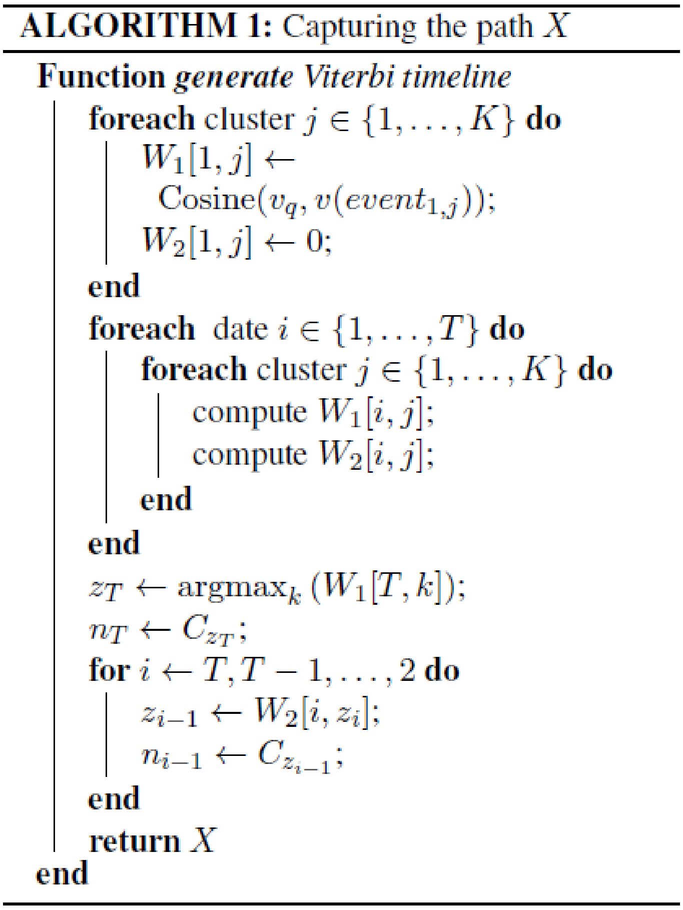
1) News Recall*. In this step, we use topic keywords and named entities to recall similar news in our database.

2) News denoising#. In this step, we use TFIDF to calculate similarity such that recalled news of a low similarity with the topic should be removed. Furthermore, we use DBSCAN to operate drift clustering. News that are not concerned by drift are regarded as noise data.

3) News alignment*#. In this step, we firstly use BERT similarity to do clustering according to their titles and publish times by calculating vectorized features of titles and giving a punishment to time difference. Then, we assume that the probability distribution of occurrence of each news is uniform. We define the launch probability from the event to each news according to the chapter-level similarity, and the transition probability from each news of the previous day to each news of the day to form a Markov chain. Finally, the Viterbi algorithm is used to find the optimal path, and the news on this path is regarded as the

[1] Note that the steps marked with * were developed by my colleagues and that marked with # were developed by me. For the step of news alignment, my supervisor proposed Viterbi algorithm for alignment while I improved clustering results by considering multiple aspects such as title BERT similarity and time penalty.

timeline context of the topic. The algorithm can be expressed in the following:



```
ALGORITHM 1: Capturing the path $X$
Function generate Viterbi timeline
    foreach cluster $j \in \{1, \ldots, K\}$ do
        $W_1[1, j] \leftarrow$
          Cosine$(v_q, v(event_{1,j}))$;
        $W_2[1, j] \leftarrow 0$;
    end
    foreach date $i \in \{1, \ldots, T\}$ do
        foreach cluster $j \in \{1, \ldots, K\}$ do
            compute $W_1[i, j]$;
            compute $W_2[i, j]$;
        end
    end
    $z_T \leftarrow \text{argmax}_k (W_1[T, k])$;
    $n_T \leftarrow C_{z_T}$;
    for $i \leftarrow T, T-1, \ldots, 2$ do
        $z_{i-1} \leftarrow W_2[i, z_i]$;
        $n_{i-1} \leftarrow C_{z_{i-1}}$;
    end
    return $X$
end
```

Algorithm 2 Our Topic TLS algorithm

An illustration of this algorithm is depicted in Figure 22:

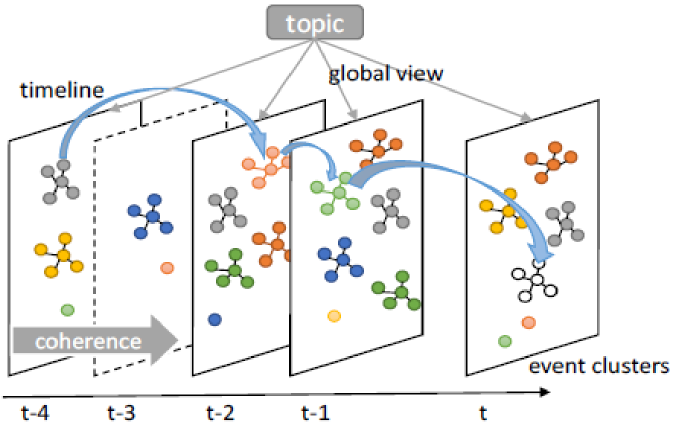


Figure 22 Illustration of our own model. Note that this figure is captured in [15] where my supervisor is one of the authors.

Eventually, we managed to operate TLS for every topic to reason their logic and find useful historic news. An example of the result is shown in Table 4:

Type	Topic/News Name	Publish Time	Keywords	Named Entities
Topic	普通医疗口罩可防德尔塔变异株	2021-09-10 18:53:00	变异、狡猾、口罩、疫情、全国	德尔塔、拉姆达
News 1	#普通医疗口罩可防德尔塔变异株#网友：我现在出门不戴口罩都感觉自己在裸奔	2021-09-10 18:53:00	口罩、变异、网友、裸奔	德尔塔
News 2	拉姆达变异株比德尔塔更强？	2021-09-09 12:26:10	变异、疫情	德尔塔、拉姆达
News 3	Delta 毒株不断突破疫苗防线 #美国人不愿意再戴上口罩#	2021-09-07 15:10:00	疫苗、口罩、疫情	德尔塔、Delta

Table 4 An example of topic TLS result

.2.2 Daily Communication

From the perspective of daily communication, I get along well with my supervisor, Guohua Wang, and cooperate with him not only in idea construction but also in code writing. Following what he has done in the Opinion System Hot Topic Discovery program, I communicate with him actively and propose some inspirational ideas based on the newest research in 2021.

Furthermore, I have a deep connection with another colleague, Xiusen Gu, who is also specialized in Natural Language Processing and have implemented the keyword extraction and Named Entity Recognition models, with which I have managed to operate Topic TLS process.

Chapter 4 Self-Evaluation

Looking back on my internship experience, I have both good and some weaknesses in my work.

For the good performance, I showed strong learning and communication skills. Frankly speaking, a lot of the work content is new knowledge for me. Especially the knowledge related to system development. The problems encountered in real engineering are different from what I learned in school. On one hand, in school, the only metric for evaluating a model is accuracy. However, in engineering, it is necessary to pursue faster computation and the lowest possible resource consumption while ensuring a high accuracy rate. For example, in the Opinion System Hot Topic Discovery program, the entire coding process needs to be run every two hours for update so that time complexity for our algorithms should be considered while accuracy is ensured. On the other hand, in school, datasets are given or are available online in most cases while in engineering, it is necessary to build and preprocess a specific dataset according to the needs of our task. In conclusion, the learning ability and adaptivity are kind of crucial in work.

In addition, I have good communication skills. I get along well with my colleagues, and I am open-minded to ask them for advice if I don't understand a problem. At the end of my internship, I can communicate with my colleagues on an equal footing and discuss solutions together. Nor only my supervisor but also my other colleagues praise me for my work and agree to give me an appointment after probation.

As for the shortcomings in my internship, my ability to plan work still needs further improvement. Due to the lack of work experience, especially in the early stage of the internship, I have deviations in judging the time needed to complete the task, which leads to the unreasonable time arrangement. I tried to remedy this problem through active communication with my colleagues. In the later part of the internship, with the increase of experience, I was able to allocate time more reasonably and arrange tasks more reasonably.

I learned a lot from this internship experience. In terms of professional skills, my programming skills have improved significantly. In business scenarios, where systems are going to be used and the robustness and correctness of the programs need to be of a high standard, Tencent's high requirements for programming code also allowed me to learn a more standardized way of programming. In terms of system development skills, I learned about the engineering issues that need to be addressed. In my future work, I will analyze the performance, algorithm prediction ability, computation time, and resource consumption from multiple perspectives to find the best solution. In terms of research capabilities, I learned more efficient ways to conduct algorithmic experiments. As the company has more computer computing resources, it can conduct multiple sets of experiments simultaneously in a parallel way. In terms of day-to-day communication, my internship at the company has helped me to develop my communication skills. I was able to communicate with my colleagues on an equal footing and express my opinions when appropriate.

At the end of my internship, I was honored by my colleagues, supervisor and group leader and seized the opportunity of becoming a regular member.

Chapter 5 Letter of thanks

At the end of my internship report, I would like to express my sincere gratitude to the people who have offered me practical, cordial and selfless support during this internship.

Firstly, I am grateful to my supervisor, Guohua Wang, with whom I spend most of time and learn most of things during my internship. He has strong professional ability and strict working attitude. In terms of workplace, he guides me to be familiar with everything in work. In terms of professional skills, he always respects my ideas and encourages me to give it a try. When I encountered difficulties, he helped me to find problems and provided me with ideas to solve them. I am lucky to have such a supervisor. Even though the internship is over, we are still good friends.

In addition, I would like to thank every colleague at Tencent. They are very kind and eager to help me in any way they can. I would like to thank them for accepting me as an intern and for making it such a wonderful experience.

Last but not least, I would like to express my gratitude to Beihang University and Centrale Pékin. The six and a half years of study at Centrale Pékin has been very rewarding. I not only learned about mathematics, physics and computer science, but also broadened my horizons and developed the ability to conduct research from a multidisciplinary perspective, which has helped me greatly during my internship. I would like to express my sincere gratitude to Centrale Pékin.

参考文献

[1] Steen J, Markert K. Abstractive Timeline Summarization[C]//Proceedings of the 2nd Workshop on New Frontiers in Summarization. 2019: 21-31.

[2] Jenatton R, Le Roux N, Bordes A, et al. A latent factor model for highly multi-relational data[C]//Advances in Neural Information Processing Systems 25 (NIPS 2012). 2012: 3176-3184.

[3] Bekker J, Davis J. Learning from positive and unlabeled data: A survey[J]. Machine Learning, 2020, 109(4): 719-760.

[4] Jaskie K, Spanias A. Positive and unlabeled learning algorithms and applications: A survey[C]//2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA). IEEE, 2019: 1-8.

[5] Chen T, Guestrin C. Xgboost: A scalable tree boosting system[C]//Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016: 785-794.

[6] Liu B, Han F X, Niu D, et al. Story Forest: Extracting Events and Telling Stories from Breaking News[J]. ACM Transactions on Knowledge Discovery from Data (TKDD), 2020, 14(3): 1-28.

[7] Bougouin A, Boudin F, Daille B. Topicrank: Graph-based topic ranking for keyphrase extraction[C]//International joint conference on natural language processing (IJCNLP). 2013: 543-551.

[8] Page L, Brin S, Motwani R, et al. The PageRank citation ranking: Bringing order to the web[R]. Stanford InfoLab, 1999.

[9] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[10] Zhang Y, Yang J. Chinese NER using lattice LSTM[J]. arXiv preprint arXiv:1805.02023, 2018.

[11] Strubell E, Verga P, Belanger D, et al. Fast and accurate entity recognition with iterated dilated convolutions[J]. arXiv preprint arXiv:1702.02098, 2017.

[12] Ma R, Peng M, Zhang Q, et al. Simplify the usage of lexicon in Chinese NER[J]. arXiv preprint arXiv:1908.05969, 2019.

[13] Sun Y, Wang S, Li Y, et al. Ernie: Enhanced representation through knowledge integration[J]. arXiv preprint arXiv:1904.09223, 2019.

[14] Song Y, Shi S, Li J, et al. Directional skip-gram: Explicitly distinguishing left and right context for word embeddings[C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). 2018: 175-180.

[15] Liang D, Wang G, Nie J. A Dynamic Evolutionary Framework for Timeline Generation based on Distributed Representations[J]. arXiv preprint arXiv:1905.05550, 2019.

致谢

在实习报告的最后，我要向在实习期间给予我实际、亲切和无私支持的人们表示衷心的感谢。

首先，感谢我的导师王国华，在实习期间，我和他一起度过了最多的时间，学到了最多的东西。他具有很强的专业能力和严谨的工作态度。在职场方面，他引导我熟悉工作中的一切；在专业技能方面，他总是尊重我的想法，鼓励我去尝试。当我遇到困难时，他帮助我发现问题，并为我提供解决问题的思路。我很幸运有这样的导师。

另外，我要感谢腾讯的每一位同事。他们非常善良，愿意以任何方式帮助我。我要感谢他们接受我作为实习生并让我拥有如此美妙的实习经历。

最后，我要向北京航空航天大学和中法工程师学院表示感谢。在 中法工程师学院的六年半学习非常有收获。我不仅学习了数学、物理和计算机科学，而且开阔了我的视野，培养了从多学科角度进行研究的能力，这对我的实习帮助很大。