文章编号：2095-2163(2021)06-0082-06 中图分类号：TP319 文献标志码：A

服装消费数据的分析与可视化平台研究

赵 娟，刘国华，史 倩，赵士源
（东华大学 计算机科学与技术学院，上海 201600）

摘 要：当前服装数据的分析与可视化存在数据单一等缺陷，以至于企业及消费者不能有效处理和利用网络购物平台中产生的海量消费数据。为此，本文以行业统计数据、服装评价、服装类新闻等 3 类关键数据为对象，设计并实现了服装消费数据的分析与可视化平台。平台针对文本类型的服装评价和新闻，运用 NLP 技术进行预处理；根据评价数据的主题差异性特点，分别采用基于词向量和欧氏距离的 k-means、基于 tf-idf 的 k-means 思想进行分类，并将 2 种结果结合进行分类与分析；根据服装潮流新闻关键词的特征，采用 tf-idf 对其进行关键词的提取与分析。最后采用 ECharts、词云、交互式等方法将各类数据进行可视化并嵌入到平台中。

关键词：服装；消费数据；聚类与分类；分析与可视化

Research on analysis and visualization platform of clothing consumption data

ZHAO Juan, LIU Guohua, SHI Qian, ZHAO Shiyuan
(School of Computer Science and Technology, Donghua University, Shanghai 201600, China)

【Abstract】At present, the analysis and visualization of clothing data has the defects of single data, so that enterprises and consumers can not effectively process and use the massive consumption data generated in the online shopping platform. By taking the three key data types such as industry statistics, clothing evaluation, and clothing trend news as the research objects, the analysis and visualization platform for clothing consumption data is designed. In view of the text-type clothing reviews and news, the natural language processing technology is used by the platform for preprocessing. In line with the subject difference features of the evaluation data, it adopts the k-means idea based on word vector and Euclidean distance and the k-means idea based on tf-idf respectively to carry out the classification, and makes combination of the two results for classification and analysis. Based on the features of clothing trend news keywords, the tf-idf is adopted to carry out the extraction and analysis on the keywords. Finally, the methods of ECharts, word cloud, interaction and so on to visualize and embed all kinds of data into the platform.

【Key words】clothing; consumption data; clustering and classification; analysis and visualization

0 引 言

随着各大电商平台不断兴起，为人们购物需求提供了极大便利，越来越多的人依赖于网络购物。因此，在网络各大购物平台中积累了大量的服装消费数据。这些数据不仅包括服装消费的数值型统计数据，而且包含服装评价、服装潮流新闻的文本数据等。将这些海量的数据进行有效的再利用，对服装的消费数据进行分析和可视化，能够帮助企业从海量消费数据中挖掘出有意义的知识，帮助服装生产企业的决策者做出正确决策，从而指导企业进行目标明确的生产、销售等一系列服务，这对于促进服装企业甚至整个服装行业的数字化转型升级有着极其重要的意义。

目前关于服装消费数据的分析和可视化研究主要分为以下几类：

(1)运用典型的深度学习方法进行文本分类，包括 RNN、CNN 等方法[1-2]。在此基础上运用 LSTM+CNN 的思想将文本数据进行分类并为其添加主题标签[3-5]。

(2)运用深度学习及 LDA 主题模型，实现了基于服装评价信息的情感分析和可视化[6-7]。

(3)运用 python 和 ECharts 工具，对服装的型号、款式、性能等进行分析，并以词云、旭日图等简单易懂的方式进行可视化[8-11]。

这些研究在一定程度上可以为消费者、服装生产厂家及服装销售商提供数据分析和可视化的参考

基金项目：上海市工业互联网创新发展专项项目(2019-GYHLW-004)。

作者简介：赵 娟(1997-)，女，硕士研究生，主要研究方向：工业互联网、自然语言处理；刘国华(1966-)，男，博士，教授，主要研究方向：大数据、数据库、隐私保护；史 倩(1998-)，女，硕士研究生，主要研究方向：工业互联网、自然语言处理；赵士源(2000-)，男，本科生，主要研究方向：自然语言处理。

通讯作者：刘国华 Email：ghliu@dhu.edu.cn

收稿日期：2021-04-12

方式，为其提供有针对性的指导。但是，在现有研究及可视化平台中，还没有综合多种服装消费数据进行统一的分析和可视化，而且对服装评价文本也仅是基于情感主题进行分析，不能满足服装消费者、服装销售商、服装生产商了解服装不同主题的需求。

针对上述问题，本文以服装行业统计数据、服装评价、服装潮流新闻为研究对象，搭建了服装消费数据的分析与可视化平台。平台中采用分词、去停用词、词性标注等方法对文本数据进行预处理后，针对服装评价数据采用基于欧氏距离、词向量的 k-means 思想和基于 tf-idf 的 k-means 思想相结合的方式，对其进行基于不同主题的分类；针对服装潮流新闻采用 tf-idf 方式进行关键词的提取；采用 ECharts 图表、词云、交互式可视化等方法将 3 类数据进行可视化，并将各种可视化结果集成到平台中。本文所实现的服装消费数据的分析与可视化平台充分考虑到了不同角色人群的需求和不同数据的具体特点，让数据分析和可视化更加人性化，弥补了基于情感主题分析以及可视化数据类型单一的不足。

1　平台体系结构

服装消费数据分析与可视化平台的体系结构如图 1 所示。结构中的关键模块包括：文本预处理、文本数据分析、数据可视化等。

(1)文本预处理模块：对文本类型数据运用自然语言处理技术进行处理。处理步骤依次为分词、去停用词、词性标注及过滤。

(2)文本分析模块：在该模块中对 2 种文本数据分别进行了分析。对于评价数据采用将基于欧氏距离和词向量的 k-means 思想与基于 tf-idf 的 k-means 思想相结合的方式进行分类与分析；对于服装潮流新闻数据采用 tf-idf 进行关键词提取及分析。

(3)数据可视化模块：针对服装消费的 3 类数据，分别采用不同的方式进行可视化。针对服装评价，通过词云显示单件衣服评价关键词，采用交互式可视化方式展示单件衣服不同部位的评价信息，另外使用 ECharts 图表对评价中的搭配推荐和除衣服部位的其它服装主题进行可视化；针对服装潮流新闻数据，采用词云对行业热点词进行可视化，并且采用列举的方式来显示服装搭配关键词；对于行业统计数据，采用 ECharts 中的图表对其进行可视化。

最后按照平台布局，将各数据模块统一集成到服装消费数据的分析与可视化平台中。

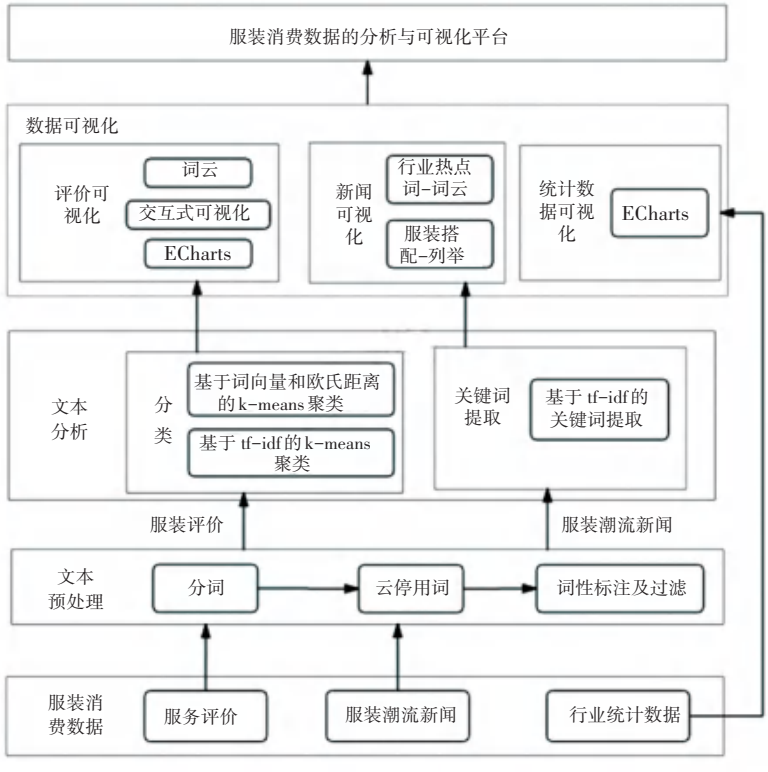

图 1　服装消费数据分析与可视化平台体系结构

Fig. 1　Architecture of analysis and visualization platform for clothing consumption data

2 文本预处理及分析

本文搭建的服装消费数据的分析与可视化平台,采用的数据具有数据量大、数据形式多样、数据可靠且有价值的特点。

2.1 数据预处理

为了保证最终结果的准确性和可信度,需要对文本数据进行预处理操作,预处理阶段主要包括分词、去停用词、词性标注与过滤。

(1)分词。本文运用 python 的第三方中文分词库-jieba,实现将文本划分为词语的集合。如将“时髦与保暖兼顾搭配法:内厚外薄,舒适又百搭!”进行分词。结果为:“时髦”、“与”、“保暖”、“兼顾”、“搭配”、“法”、“:”、“内厚”、“外薄”、“,”、“舒适”、“又”、“百搭”、“!”。

(2)去停用词。根据自定义的停用词词典,将以上分词结果中的“与”、“又”等没有实际意义的词去除,以保证分析及可视化结果的有效性。

(3)词性标注与过滤。由于文本信息的分析是基于关键词而进行的,而且关于消费的重要信息多体现在名词、形容词类的关键词中。因此,本文对去停用词之后的结果进行了词性标注,并且筛选过滤掉其它词性的词语,如“兼顾”、“保暖”等。

2.2 基于主题的服装评价分类与分析

基于预处理后的结果,对评价数据进行分类。在此环节,主要采用基于欧氏距离和词向量的 k-means 思想与基于 tf-idf 的 k-means 思想相结合的方式,对服装评价进行基于主题的分类与分析后,将 2 种分类的结果进行合并确定最终的分类结果

k 均值聚类算法是一种无监督的聚类算法。其基本思想为:对于给定的数据集,预定将其分为 k 个类别,并从数据中选取 k 个对象作为聚类的中心,然后计算其它数据对象与各个聚类中心点之间的距离,之后计算每个中心点中距离的均值,将均值做为新的中心点,通过这种方式进行多次迭代,将各个对象划分到效果较好的聚类中心点范围,以此来达到分类的效果。

2.2.1 基于欧氏距离和词向量的 k-means 聚类步骤

(1)运用 word2vec 工具,根据给定的语料库训练模型,以此来计算每条文本数据的词向量,将其作为各文本,用于计算距离的值。

(2)随机选取 k 个词向量值作为 k-means 聚类的中心点。

(3)计算其它数据对象与中心点的距离,该距离通过两点之间的欧氏距离来确定。按照距离中心点的距离最小化原则,将所有数据对象都分配到各个中心点中。欧式距离是指两点之间的实际距离,由公式(1)确定。

$$\rho = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}. \tag{1}$$

其中,ρ 为点 (x_2, y_2) 与点 (x_1, y_1) 之间的欧氏距离。

(4)计算每个中心点中的数据对象与中心点间欧氏距离的均值,将均值作为下一次迭代的中心点。

(5)迭代执行(3)、(4)步,直到中心点不再改变或者达到设置的迭代最大次数。将最后一次迭代的结果作为最终的分类结果。

2.2.2 基于 tf-idf 的 k-means 聚类步骤

(1)计算每条文本中所有关键词的 $tf - idf$ 值,生成一个 $\boldsymbol{m} \times \boldsymbol{n}$ 的矩阵。m 表示文本的数量,n 表示所有文本包含的不重复关键词的总和,如公式(2)所示。在矩阵中,一行代表一个文本,矩阵中的每个值表示每个关键词的 $tf - idf$ 值。如 $\boldsymbol{A}_{11}$ 表示第一条文本中是否出现了第一个关键词,如果没有出现,则该处值为 0,反之该处的值为 $tf - idf_{11}$。其中,$tf - idf$ 代表词频 - 逆文档频度,用于衡量一个词对于一个文件集或一个语料库中的一份文件的重要程度;tf 表示词频,用以衡量一个词语在其所在文件中的重要程度;idf 表示逆文本频率指数,用以衡量一个词语的普遍重要性。如果一个词语在一篇文章或者一个文件中出现的频率较大,而在其它文件中很少出现,则可以认为这个词语具有一定的类别区分能力。因此可以用于进行基于主题的服装评价分类,具体计算过程如公式(3) - (5) 所示。(式中各字符的定义见表 1)

$$\begin{bmatrix} A_{11} & \cdots & A_{1n} \\ \vdots & \ddots & \vdots \\ A_{m1} & \cdots & A_{mn} \end{bmatrix}, \tag{2}$$

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}, \tag{3}$$

$$idf_i = \lg \frac{|D|}{|\{j : t_i \in d_j\}|}, \tag{4}$$

$$tfidf_{i,j} = tf_{i,j} \times idf_i. \tag{5}$$

(2)从中随机选取非连续的 k 行作为 $k-means$ 的中心点。

(3)计算其它文本对象与中心点的距离。该距离由两个点的矩阵作差之后的二范数确定,之后按

照与中心点的距离最小化原则，将所有数据对象都分配到各个中心点中。

(4)计算每个中心点中的数据对象与中心点间距离的均值，将均值作为下一次迭代的中心点。

(5)迭代执行(3)、(4)步，直到中心点不再改变或者达到设置的迭代最大次数。将最后一次迭代的结果作为最终的分类结果。

最后，基于两种思想分别得到的分类结果，按照求并集的思想求得最终的服装评价分类结果。

表 1　公式中字符的定义

Tab. 1　Definition of characters in formula

字母	定义
$tf_{i,j}$	词语 i 在文件 d_j 中出现的频率，评价词语在文件 d_j 中的重要程度
n_{ij}	词语 i 在文件 d_j 中出现的次数
k	文件 d_j 中不同词语的个数
$\sum_k n_{k,j}$	文件 d_j 中所有词语出现的次数总和
idf_i	词语 i 的逆向文件频率
$\mid D \mid$	数据集中文件的总数
$\mid \{j:t_i \in d_j\} \mid$	包含词语 t_i 的文件数量，公式中要 + 1
$tfidf_{i,j}$	词频-逆向文件频率，该值越大，词语 i 的区分度越高

2.3　服装潮流新闻的关键词提取

本文将爬取到的服装潮流新闻进行预处理，消除不相关词的干扰，并将服装潮流新闻中包含的所有关键词整合到一起，作为运用 tf-idf 算法提取关键词的语料库；采用基于 tf-idf 思想，对每条新闻进行关键词的提取，并对提取的结果采用关键词匹配方式，分别提取行业热点词和服装搭配关键词。

2.4　数据可视化

在平台各数据的可视化中，对 3 类数据分别采用多种不同方法进行可视化。

2.4.1　服装消费统计数据

该平台对服装消费的统计类数据直接采用 ECharts 中的柱状图、折线图、饼状图、散点图等方法进行可视化。其中关于服装的部分产销率指标的可视化结果如图 2 所示，以柱状图的形式显示更易了解其变化趋势。

2.4.2　服装评价数据可视化

按照基于词向量和欧氏距离的 k-means 思想进行评价分类时，k 取值为 6；采用基于 tf-idf 的 k-means 思想进行分类时，k 值取 8。最后将分类结果使用 translate 工具进行翻译得出最终显示的内容。

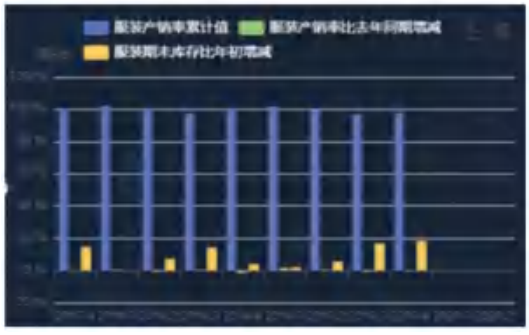

图 2　产销率指标

Fig. 2　Production and marketing rate index

评价数据主题分为 3 类，分别为衣服部位主题、搭配推荐主题、其它主题。平台中对于服装评价各类主题的展示效果如下：

(1)关于服装部位的主题采用交互式方法进行可视化。在平台中用户可以通过滑动鼠标选择不同部位来查看其对应的评价信息。如图 3 所示，当鼠标滑动到腰部位置时，会显示对于腰部的评论信息。

图 3　服装评价主题-衣服部位

Fig. 3　Clothing evaluation theme-clothing parts

(2)对于单件衣服评价中的搭配推荐。在平台中以 ECharts 的条形图进行了可视化，如图 4 中所示，显示了对于各种搭配推荐的支持用户数量。

图 4　服装评价主题-搭配推荐

Fig. 4　Clothing evaluation theme-collocation recommendation

(3)对于除服装部位的其它主题，采用 ECharts 的饼状图进行可视化。如图 5 中，当用户点击“大

小”这一主题时,即可在平台中看到相应的评价,以及各方面不同评价占该主题所有评价的比例。

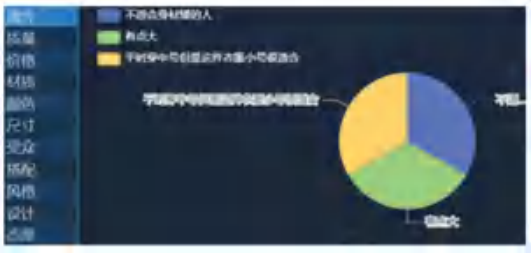

图 5 服装评价主题-服装属性

Fig. 5 Clothing evaluation theme-clothing attributes

另外,对于单件衣服的所有评价,在平台中对其预处理之后,按照各关键词出现的频率,以词云方式进行了可视化,其结果如图 6 所示。在图中可以明显看出“颜色好看”比“长度适中”更显眼,表明对于该件衣服,关于“颜色好看”的评论信息较多。

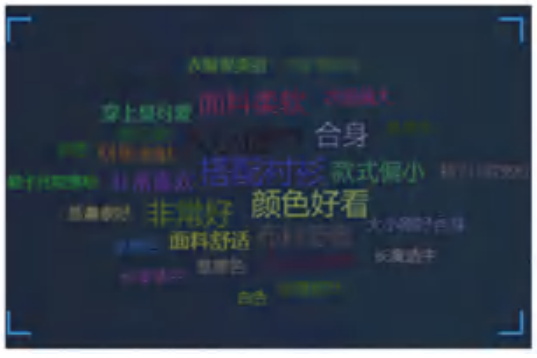

图 6 服装评价词云

Fig. 6 The wordcloud of clothing evaluation

2.4.3 服装潮流新闻

在平台中对服装潮流新闻按照平台提供的预处理和关键词提取模块进行关键词提取之后,对近三年的行业热点词采用词云的方式进行了可视化,如图 7 所示。

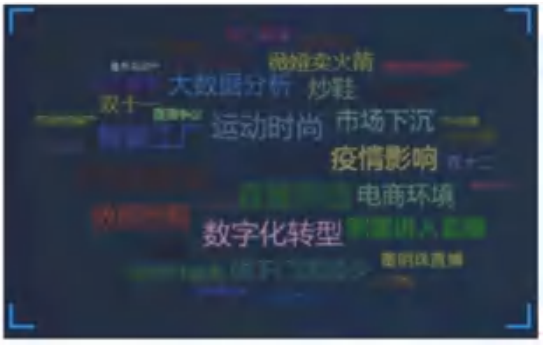

图 7 行业热点词词云

Fig. 7 The wordcloud of industry hot words

另外,平台也对新闻中的潮流穿搭关键词采用归类列举的方式进行了可视化,如图 8 所示。如流行的“男装女穿”穿搭,包括工装连体裤搭配短靴、带帽卫衣等。

图 8 潮流穿搭

Fig. 8 Fashion wear

3 平台搭建与分析

3.1 数据集

该平台中所涉及的数据集包括服装消费统计数据、服装评价数据、服装潮流新闻。

(1)服装消费统计数据。服装消费统计数据从国家统计局网站获取,用于分析服装行业整体发展趋势,包括近几年批发及零售的相关指数、居民服装消费指数等。

(2)服装评价数据。服装评价数据是关于女装的评论数据集,用于对服装评价进行分析和可视化。

(3)新闻数据。中国服装协会网提供了国内及国际上关于服装行业的新闻数据,本文采用爬虫方式获取了部分服装潮流新闻数据。

3.2 平台环境及工具

实现该平台所使用的环境及各模块所用工具见表 2。

表 2 实验环境及工具

Tab. 2 Experimental environment and tools

用途	工具
编程环境	Python3.7.6
编辑器	Visual Studio Code 1.54
分词	jieba
可视化	ECharts、wordcloud
词向量	Word2vec
自然语言处理库	nltk
爬取数据	Selenium、BeautifulSoup

3.3 平台效果展示

以与服装消费相关的服装行业统计数据、服装评价、服装类新闻 3 类数据为对象,分别对其进行分析及可视化,然后将各可视化结果通过 Vue 组件进行整合,统一集成到平台主页中。服装消费数据的

分析与可视化平台整体效果如图 9 所示。

图 9　服装消费数据分析与可视化平台界面

Fig. 9　The interface of the analysis and visualization platform for clothing consumption data

4　结束语

本文以服装行业统计数据、服装评价、服装类新闻 3 类与服装消费相关的数据为对象，分别对其进行了分析和可视化，并根据服装消费数据的特点设计了服装消费数据的分析与可视化平台体系结构。不仅采用基于词向量和欧氏距离的 k-means 与基于 tf-idf 的 k-means 聚类相结合的方式实现了评价数据基于主题的分类分析、采用 tf-idf 算法对服装类新闻进行了关键词提取、采用多种不同的可视化方式对服装消费数据进行了可视化，而且设计并实现了服装消费数据的分析与可视化平台。

本文中的消费数据种类和及消费数据的分析方法有待进一步研究，以完善服装消费数据的分析与可视化，并实现更高效更深入的信息挖掘。

参考文献

[1] LIU Pengfei, QIU Xipeng, HUANG Xuanjing. Recurrent Neural Network for Text Classification with Multi-Task Learning[C]// IJCAI 2016: 2873-2879.

[2] RAFIEI A, REZAEE A, HAJATI F, et al. SSP: Early prediction of sepsis using fully connected LSTM-CNN model[J]. Computers in Biology and Medicine, 2021, 128:104110.

[3] YUE Z, QI L, SONG L. Sentence-State LSTM for Text Representation[C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018.

[4] LOGANATHAN K, RS KUMAR, NAGARAJ V, et al. CNN & LSTM using python for automatic image captioning[J]. Materials Today: Proceedings, 2020(6245).

[5] HUANG Z, WEI X, KAI Y. Bidirectional LSTM-CRF Models for Sequence Tagging[J]. Computer Science, 2015.

[6] 李文江. 基于深度学习的商品评价数据分析系统[D]. 大连:大连海事大学,2018.

[7] 廖桦涛. 多电商平台服装产品评论信息综合分析与可视化[D]. 天津:天津师范大学,2017.

[8] 李晓久,刘皓. 基于数据分析的羊毛衫领口性能评价方法的确定[J]. 东华大学学报(自然科学版),2005,31(5):59-63.

[9] 刘睿智,赵守香. 基于 SPSS 数据分析的服装号型设计[J]. 服装报,2019,4(6):504-509.

[10] 刘旭婷, 李春青, 荆妙蕾, 等. 基于 Python 的消费者服装购买数据分析研究[J]. 计算机科学与应用,2021,11(1):1-7.

[11] 易小群,李天瑞,陈超. 面向评论文本数据的旭日图可视化[J]. 计算机科学,2019,46(10):14-18.

(上接第 81 页)

[4] ZHONG H, GUO Z, TU C, et al. Legal judgment prediction via topological learning[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018: 3540-3549.

[5] YANG W, JIA W, ZHOU X I, et al. Legal judgment prediction via multi-perspective bi-feedback network[J]. arXiv preprint arXiv:1905.03969, 2019.

[6] CHEN H, CAI D, DAI W, et al. Charge-Based Prison Term Prediction with Deep Gating Network[J]. arXiv preprint arXiv: 1908.11521, 2019.

[7] VLEK C, PRAKKEN H, RENOOIJ S, et al. Constructing and understanding Bayesian networks for legal evidence with scenario schemes[C]// International Conference on Artificial Intelligence and Law. ACM, 2015:128-137.

[8] WALKER V R. Visualizing the dynamics around the rule-evidence interface in legal reasoning[J]. Law, Probability and Risk, 2018, 6(1-4):5-22.

[9] LUO B, FENG Y, XU J, et al. Learning to Predict Charges for Criminal Cases with Legal Basis[J]. 2017.

[10] WANG P, FAN Y, NIU S, et al. Hierarchical matching network for crime classification[C]//Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2019: 325-334.

[11] PETERS M E, NEUMANN M, IYYER M, et al. Deep contextualized word representations[J]. arXiv preprint arXiv: 1802.05365, 2018.

[12] 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.

[13] YANG Z, YANG D, DYER C, et al. Hierarchical attention networks for document classification[C]//Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies. 2016: 1480-1489.

[14] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Advances in neural information processing systems. 2017: 5998-6008.