Computer Science and Application 计算机科学与应用, 2020, 10(2), 265-275
Published Online February 2020 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2020.102028

Hans汉斯

Design and Implementation of Intelligent Quotation System for Printing Products

Zhengxin Li[1], Wangyu Liu[2], Zhe Zhang[3], Xiuli Shao[1]

[1]College of Computer Science, Nankai University, Tianjin
[2]Tianjin Huayue Color Printing Co. Ltd., Tianjin
[3]Nankai University Binhai College, Tianjin
Email: lzx@mail.nankai.edu.cn

Received: Jan. 30th, 2020; accepted: Feb. 11th, 2020; published: Feb. 18th, 2020

Abstract

In order to achieve a more scientific prediction and quotation of printed cartridge products, this paper designs and implements an intelligent quotation system for printed products. Its main functions are intelligent quotation and intelligent robot. Among them, the intelligent quotation function is based on the printing product parameters and historical order data, as well as other quotation information obtained from the Internet through crawler technology. The regression analysis equation is calculated by the least square method, and the estimated price of the printed products is given. In addition, an intelligent quotation robot based on TF-IDF algorithm is designed. After practical use, the intelligent quotation system realized in this paper can better assist enterprises in price evaluation and analysis of future orders.

Keywords

Intelligent Quotation, TF-IDF Algorithm, Regression Analysis, Reptile, Intelligent Question Answering

印刷产品智能报价系统的设计与实现

黎正鑫[1]，刘网钰[2]，张　喆[3]，邵秀丽[1]

[1]南开大学计算机学院，天津
[2]天津市华跃彩色印刷有限公司，天津
[3]南开大学滨海学院，天津
Email: lzx@mail.nankai.edu.cn

收稿日期：2020年1月30日；录用日期：2020年2月11日；发布日期：2020年2月18日

文章引用: 黎正鑫, 刘网钰, 张喆, 邵秀丽. 印刷产品智能报价系统的设计与实现[J]. 计算机科学与应用, 2020, 10(2): 265-275. DOI: 10.12677/csa.2020.102028

摘　要

为实现对印刷药盒产品较为科学的预估报价，本文设计实现了印刷产品智能报价系统。其主要功能是智能报价和智能问答。其中，智能报价功能依据印刷产品参数与历史订单数据，以及通过爬虫技术从网上获取的其他报价信息，利用最小二乘法计算回归分析方程，给出印后产品的预估价格。此外，系统还设计了基于TF-IDF算法的智能报价问答系统。经实际使用，本文实现的智能报价系统能较好地辅助企业对未来订单的价格评估和分析。

关键词

智能报价，TF-IDF算法，回归分析，爬虫，智能问答

1. 引言

目前，数字化深入各行各业，报价服务系统已成为近年来一个重要研究热点和应用热点。对于企业来说，精准的报价可以保证企业在竞争激烈的市场环境中脱颖而出，更加有的放矢。将信息技术与商业流程进行结合，高效的报价系统的研发需要提上日程[1]。

产品报价是报价系统中的一项重要工作。在报价过程中，首先要注意报价的合理性，为企业保留合理的利润空间，其次要注意报价的及时性。通常来讲，报价的格式可以按照难易程度来区分为标准报价、包括内部配置的报价和完全基于定制的报价服务三种。前两种存在不灵活和不系统的问题，后一种虽然可以使客户满足自己的个性需求，但依然存在沟通少、参考少的问题[2]。

本文设计并实现了印刷产品智能报价系统。系统采用机器学习中非常经典的多元回归算法来拟合出用来智能报价印刷产品的多元函数，数据来源是历年的订单和从 web 获得的印刷产品信息，首先通过用户的相关输入，在历史订单数据库中筛选出与用户相关的行，将这些数据用来作为训练数据，通过最小二乘法拟合出回归方程[3] [4] [5]，并通过该方程对用户的输入进行计算得出一个价格的评估。然后利用 java 爬虫技术爬取其他印刷产品的相关详情，通过规格等参数给出估价，并综合从历史订单中的报价给出最终的报价，最后将所有信息通过 JSP 页面显示给用户[6] [7] [8] [9]。系统的设计增强了智能报价系统的灵活性。并且借助智能问答，让客户拥有良好的沟通体验印刷产品智能报价系统的设计与实现。

2. 系统概述

系统功能设计

本系统主要有 2 大模块，分为智能报价及智能问答模块。其中智能报价模块中包括了用户填写相关印刷产品参数、爬取网页上相关印刷产品信息、结合历史订单和从网页获得的印刷产品的相关数据根据用户输入进行智能报价的功能。智能问答模块包括了问题的相似匹配以及对相似问题的处理并输出，从而实现和用户的问答交流，用户可以向该系统询问有关印刷产品的信息，如：查询价格、尺寸，材质，包装、类型等。图 1 为系统功能图。

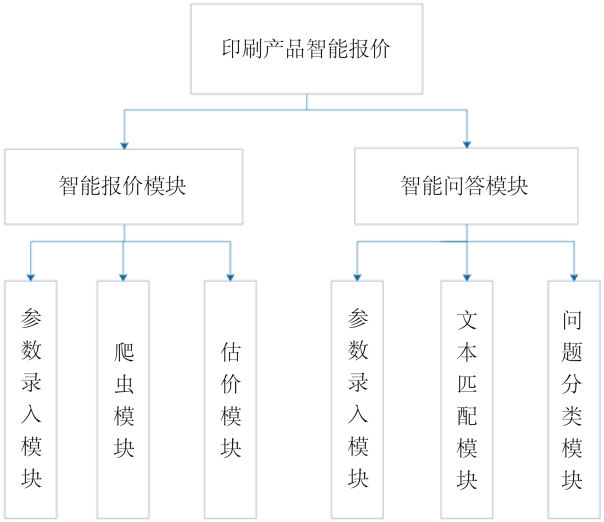

Figure 1. System function diagram
图 1. 系统功能图

在智能报价模块中，印刷产品的价格评估一般考虑产品的尺寸，即长宽高，产品规格，数量，加工程序，然后给出预估价。图 2 为基础报价的计算流程。

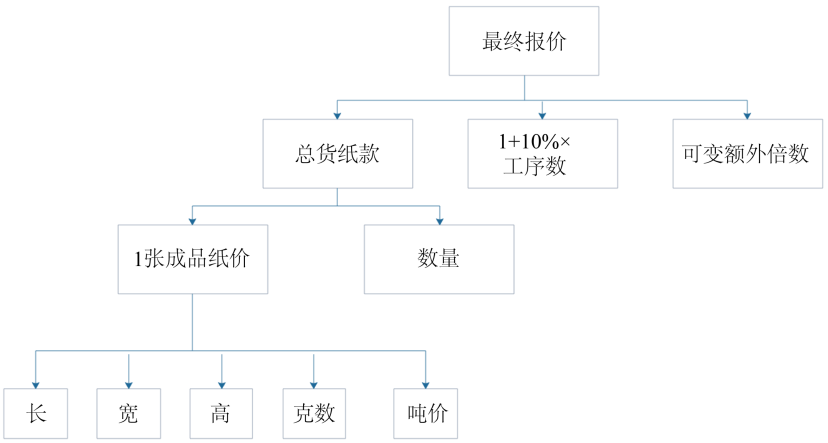

Figure 2. Basic quotation flow chart
图 2. 基础报价流程图

但单纯的计算公式不具备永久的实用性，而且历年价格不一，此时，历史订单中的数据就可以起到很好的参考作用。历史订单中记录着产品的编号，产品的名称以及交易价格，然后根据产品编号从产品列表中取出产品的信息，如尺寸，规格，加工程序等等，就可以获得所有关于尺寸，规格，加工程序，价格数据，从这些数据中筛选出与用户的输入相关的数据，通过回归分析给出合理报价，并将数据显示给用户[10] [11] [12]，对于用户而言不仅能得要一个全面的预估价，还能了解历年的价格详情，具有较强的现实意义。除此之外，利用爬虫技术，从官网获取其他印刷产品，如印刷产品的规格，价格等信息，建立产品规格和价格的回归方程[13] [14]，根据用户的输入规格计算出爬虫给出的报价，结合历史订单中的报价给出最终报价。

在智能问答模块中，论文基于文档相似匹配算法 TF-IDF 算法找出相似度最高的问题，然后处理问题

的答案，有三种处理方法，一种直接输出，一种是间接输出，如查询操作，需要在遍历数据库再给出答案，另一种是引导式按步骤回答，如查询商品，需要用户按照系统给定步骤依次输入筛选条件，最终给用户显示推荐产品。

3. 智能报价

智能报价的报价依据主要来自两个方面，一个是通过爬虫技术获得当前纸张的价格参数，另外一个是从历史订单中获得历史的纸张价格以及参数，通过这两个因素综合给出对应用户需求的预估价格。

3.1. 基于 web 爬虫爬取网上相关售价

爬虫技术主要是下载 web 网页中的文件，并解析 web 文件，获取自己想要的节点元素处的值。

步骤 1：首先选取一部分种子 URL (或初始 URL)，将其放入待爬取的队列中。如放入 LinkedList 或 List 中。

步骤 2：判断 URL 队列是否为空，如果为空则结束程序的执行，否则执行第 3 步骤。

步骤 3：从待爬取的 URL 队列中取出待爬的一个 URL，获取 URL 对应的网页内容。在此步骤需要使用响应的状态码(如 200，403 等)判断是否获取数据，如响应成功则执行解析操作；如响应不成功，则将其重新放入待爬取队列(注意这里需要移除无效 URL)。

步骤 4：针对已经响应成功后获取到的数据，执行页面解析操作。在这里我们需要获取的数据就是印刷产品的相关参数。

步骤 5：针对 3 步骤已解析的数据，将其进行存储[15] [16] [17]。

这里我们采用 JSoup 技术来进行网络爬虫。图 3 为爬虫的基本流程。

如图 3 所示，首先我们在队列中添加我们所要爬取的网页的 URL，然后从队列中依次取出一条 URL，若队列为空则结束，否则通过 JSoup 的 connect 函数来获取对应 URL 的整个文档并解析，然后选择所要获取信息的节点标签，可通过对应的 class 来获取标签节点，这里我们只需要获得商品价格，长宽高以及规格所对应的标签，然后取出标签里面的值作为我们后续预估价格的一个参考依据。

3.2. 本系统报价信息采集功能实现

图 4 为参数录入界面，用户需要填写印刷产品的相关参数，填完之后点击预估价按钮，数据送往服务端，服务端对数据进行合法检测。

用户提交后，用户输入的长，宽，高，印刷产品规格，印刷产品数量，加工程序将送至服务器，服务器端接收数据，根据历史订单和爬虫获得的信息结合用户输入给出合理报价，其中主要用到回归分析算法，以下是对算法的分析和如何应用该算法来给出报价。

3.3. 回归分析算法

为了对用户输入信息给出预股价，本文采用回归分析算法，具体方法为：历史订单中存在产品编号，价格等信息，根据产品编号在产品列表中找到该产品的尺寸，价格，规格，加工程序等等，分别对这些数据和从 web 爬取获得的数据计算多元回归方程，然后根据前端用户输入的参数分别计算价格，结合这两个价格来生成最终的报价，以下为回归分析算法介绍：

多元线性回归模型的一般形式是：

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \cdots + \beta_n X_n + \varepsilon \tag{1}$$

其中

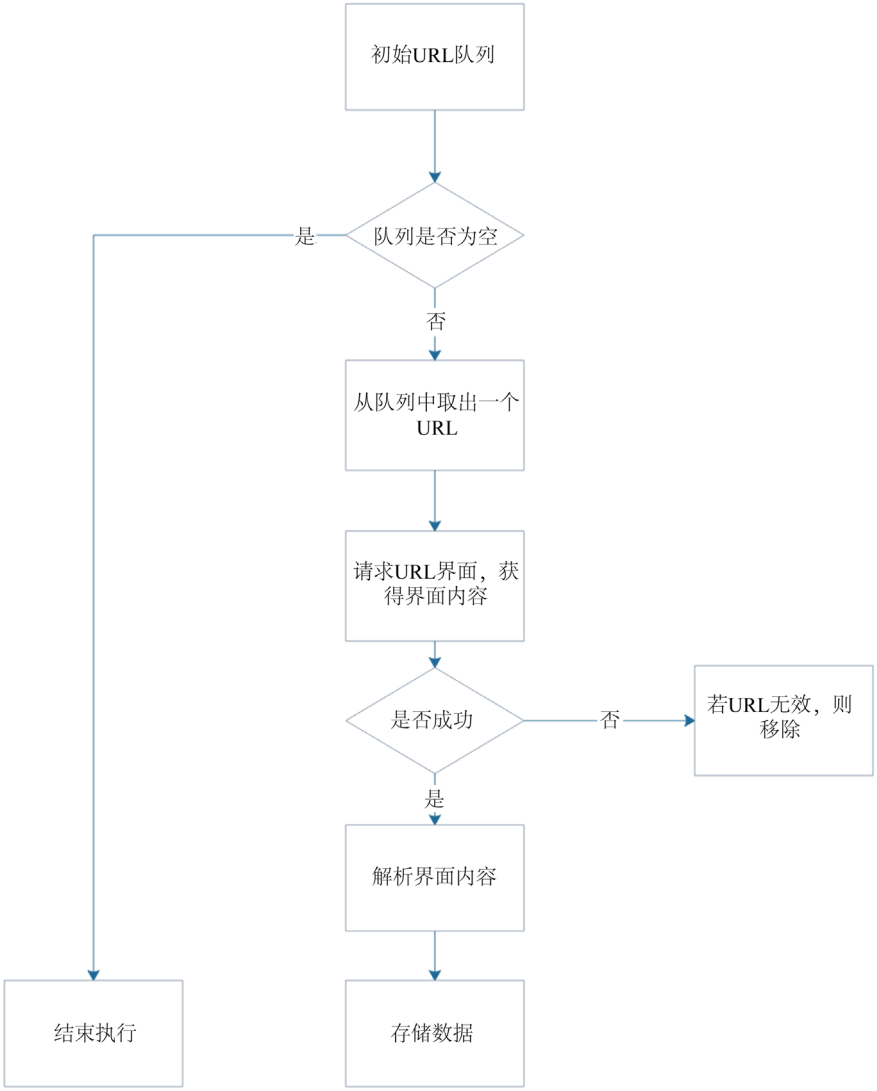

Figure 3. Basic flow chart of web crawler
图 3. Web 爬虫基本流程图

Figure 4. Parameter input interface
图 4. 参数输入界面

$$\beta = XX^{-1}X\acute{y} \quad [18]\ [19] \tag{2}$$

$$X = \begin{bmatrix} 1 & x_1^1 & \cdots & x_n^1 \\ 1 & x_1^2 & \cdots & x_n^2 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_1^m & \cdots & x_n^m \end{bmatrix} \tag{3}$$

X 代表数据库中纸张的长宽高，y 为纸张的历史价格。

最后，我们可以通过 $\beta = X'X^{-1}X'y$ 来计算相关系数[18] [19]，并对用户的参数输入进行价格评估。

在 2.2 中，用户输入了印刷产品的长，宽，高，印刷产品规格，数量，加工程序，根据历史订单中产品的编号从产品列表中找到该产品的尺寸，价格，规格，加工程序，根据用户选择的加工程序和印刷产品规格作为条件来筛选这些数据，选取具有相同的加工程序和印刷产品规格的数据作为分析数据，选取这些数据的尺寸(长，宽，高)构成上面的系数矩阵X，记长为 X_1，宽为 X_2，高为 X_3。价格构成向量 y，根据上式计算 $\beta = X'X^{-1}X'y$，即回归系数，再将用户输入的长宽高带入回归方程中计算出第一个估价。对于从 web 获取的数据，选取与用户提交相关的数据，如尺寸，规格构成系数矩阵 X，记规格为 X_1，尺寸为 X_2。价格构成向量y，同样计算回归方程，最后根据用户输入的规格，尺寸计算出第二个价格。通过 $p = X\beta$ 计算两个训练结果向量(训练价格)，由于两种估价参考意义不同，很显然，历史订单能够很好的反应价格，而从 web 获取的价格仅作参考，所以需要给两个价格赋予一定的权重，求解

$$\min_{\alpha,\beta,\alpha+\beta=1}\left\|\alpha p_1 + \beta p_2 - y\right\|_2 \tag{4}$$

即：

$$\min_{\alpha,\beta,\alpha+\beta=1}\left\{\sum_{i=1}^{n}\left(\left(\alpha p_1^i + \beta p_2^i\right) - y_i\right)^2\right\} \quad \alpha \text{ 和 } \beta > 0 \tag{5}$$

对其求偏导，令偏导数等于 0，计算得出

$$\alpha = \frac{\sum_{i=1}^{n}\left(p_2^i - y^i\right)\left(p_2^i - p_1^i\right)}{\left\|p_1 - p_2\right\|_2}, \quad \beta = 1 - \alpha \tag{6}$$

使得两个价格的期望值与真实值具有最小平方损失。最后根据用户的输入计算出两个估价，根据 α,β 求的两个价格的期望，即给用户的最终报价。

下图(图 5)是该计算过程的流程图。

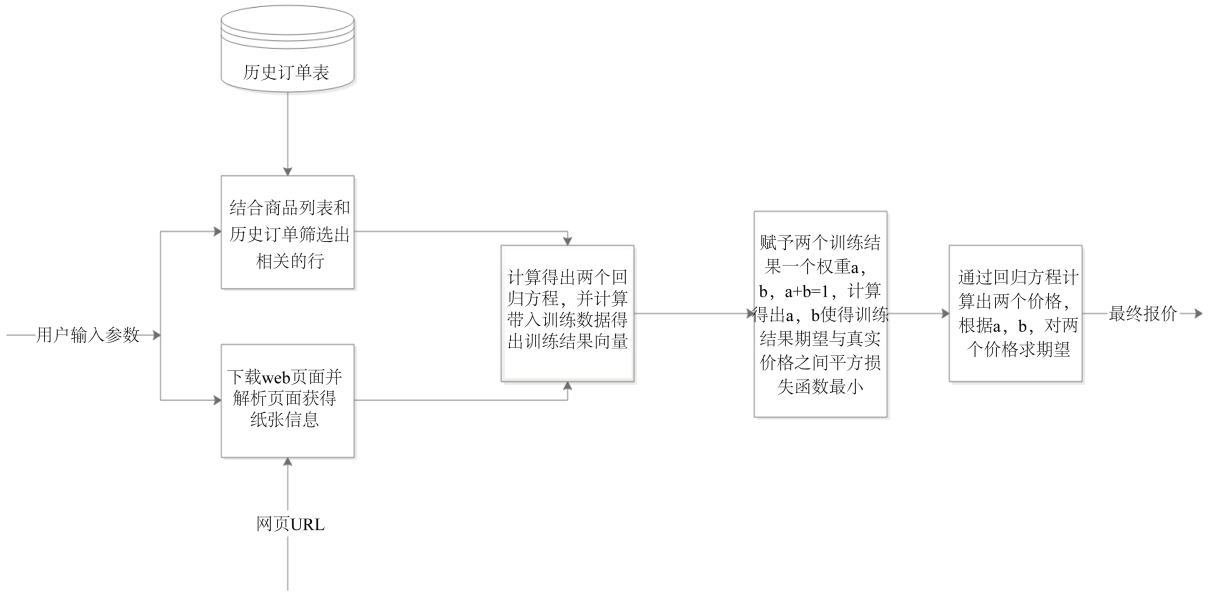

Figure 5. Intelligent quotation flow chart
图 5. 智能报价流程图

3.4. 模型评估

利用最小二乘法对历史数据进行拟合生成长宽高系数，该系数对用户输入长宽高进行加权作为反馈用户的预测价格，为了验证预测的准确性，将历史订单中数据 X 送入模型 $f, f \in R^3 \to R$，得到所有预测价格 $\hat{y}$，通过 $\|y-\hat{y}\|$ 来衡量模型预测准确率，其中 y 为历史订单中的真实价格。

Table 1. Model evaluation form
表 1. 模型评估表

规格	工序	误差
80 g	打孔，压横	0.103
80 g	单面凸击，单面烫金	0.97
80 g	打孔，单面凸击	0.473
80 g	单面覆亮膜，模切‖异型	0.424
80 g	双面覆亚膜，单面烫银	0.877

从表 1 可以看出，随机固定规格，工序的纸张而言，价格的平均误差均未超过 1，说明该线性拟合效果完全符合预期，即通过线性回归方程来计算预估价具有合理性。

3.5. 基于 Echarts 的图表绘制

ECharts 是一个使用 JavaScript 脚本语言实现的开源可视化库，可以非常流畅的运行在各个终端上，兼容当前绝大部分主流浏览器(IE，Firefox，Chrome，Safari 等等)，底层依赖轻量级的矢量图形库 ZRender 提供直观，交互丰富，可以高度个性化定制数据可视化图表。

ECharts 提供了常规的柱状图，折线图，K 线图，散点图，饼图和用于统计的盒形图，用于地理数据可视化的地图热力图、线图，用于关系数据可视化关系图、treemap 旭日图，多维数据可视的平行坐标，还有用于 BI 的漏斗图，仪表盘，并且支持图与图之间的混搭。

除了已经内置的许多具有丰富功能的图表，ECharts 还提供了自定义系列，开发者只需要传入一个 renderItem 函数，就可以从数据映射到任何你想要的图形，更棒的是这些都还能和已有的交互组件结合使用而不需要操心其它事情。这里我们对相关数据进行绘图就采用百度提供的 javascript API。

4. 智能问答

本系统设置该模块主要是为了解答用户的问题，如查询印刷产品名，尺寸，材质，价格等等。主要通过 TF-IDF 算法计算用户问句与数据库中问句的相似度，将相似度最高的数据库项目返回给用户。当用户问句与数据库中的问题相似度为 0 或者低相似性，我们则返回用户默认答案，也就是问答系统所具有的默认功能。

4.1. 基于 TF-IDF 算法完成文档的相似匹配

设 q 为查询文档，及对应为用户的输入，d_i 为数据库中第 i 个待匹配文档。

首先将获取 q 查询文档中的关键字序列，构成关键字向量，将查询文档，匹配文档表示成关键字向量，如下：

$$q = (x_1, x_2, x_3, x_4, \cdots, x_n) \tag{7}$$

$$d = (y_1, y_2, y_3, y_4, \cdots, y_n) \tag{8}$$

$x_i = c(\omega_i, q)$，表示第 i 个关键词在查询文档中出现的次数，

$$y_i = c(\omega_i, d) * \mathrm{IDF}(\omega_i) \tag{9}$$

$$\mathrm{IDF}(\omega_i) = \log \frac{M+1}{k} \tag{10}$$

M 是文档集中文档数量，k 是词在多少个文档中出现[20]。

我们计算 q 与 d 两个向量夹角的余弦值，有关证明表明，余弦值越大，文档越相似。最后我们引入 Pivoted Length Normalization 算法，该算法用于解决文档过长所产生的影响，计算公式为：

$$\mathrm{normalizer} = 1 - b + b * \frac{d}{\mathrm{avdl}} \tag{11}$$

其中，b 为惩罚因子，当 b 为 0 时，表示没有惩罚，当文本余弦值一样活着接近时，文本越短越相似，在问答系统中，文本长度较短，所以我们选取惩罚因子较小的 $b = 0.2$，避免相似度过分依赖文档长度而造成分析错误，最终通过公示

$$\mathrm{final} = \frac{q \cdot d}{|q| * |d| * \mathrm{normalizer}} \tag{12}$$

来计算相似度，最终使得相似度最高的 d_i 作为与用户最匹配的问题，并将该问题对应的答案返回给用户。

4.2. 问题分类

将数据库中的问题分为三类问题，一种是直接回答的问题，第二种为间接回答，如该问题是类似于查询操作，答案为需要查询的 sql 语句，第三种为步骤回答，即可能需要问答多次才返回最终结果，如价格查询，商品查询，需要先输入类型，包装，尺寸等等，最后产生结果。

首先将用户的问题去掉停用词，并拆分成关键词序列 q，分别将数据库中的每一问题转换为词向量 d，通过 final 公式计算相似度，选择最大相似度的问题，如果数据库中的每个问题都不能匹配问题，即数据库中并无此问题记录，这时推荐默认的回答，这个默认回答主要介绍了系统的所能回答的范围。

当匹配到问题时，该问题存在一种类型。数据库中的问题分为三类，直接回答，间接回答，按步骤回答。首先通过用户的输入从数据库中找到与最相似的问题，数据库中标记有该数据为直接回答，间接回答，还是按照系统预定的步骤回答，如何该问题的标记为直接回答，则直接返回内容字段，例如问题：你好。该问题在数据库中标记为直接回答，则系统直接将内容：你好返回给用户，问题结束；如果该问题的标记为间接回答，在该系统中，间接回答问题均属于查查询问题，数据库中对应的内容字段为一个 sql 查询语句，系统则会执行该查询语句，返回相应的查询结果，如：查询数据库中印刷产品的类型，在内容字段则存储一个 sql 语句：select * from table where…，系统执行该语句返回查询结果；如果该问题为需要按照步骤回答的问题，如查询价格，对于这一问题的标记为按照系统给定的步骤回答，这是系统会给出一列问题，用户回答相应问题，最终系统给出查询结果，如先让用户输入尺寸，然后输入规格等等，最终系统返回一个对应相应输入的价格，问题结束，如果用户输入的问题不是当前给出的问题，则系统停止该问题的处理，重新在数据库中寻找用户输入所对应的问题。具体流程成如图 6 所示。

5. 系统展示

如图 7 所示，为智能报价系统的展示，从第一副图可以看出，系统中提供多种参数选择，如规格，工序，最后系统将挑选出对相应参数的历史订单记录，同时查询从网页上爬取的相关纸张规格信息，若存在相一致的规格则保留，否则丢弃，并对这些商品的长宽高与价格拟合两个回归方程，根据公式(6)给出两个预测结果的权值，最终根据用户输入的长宽高给出最终的预测价格。

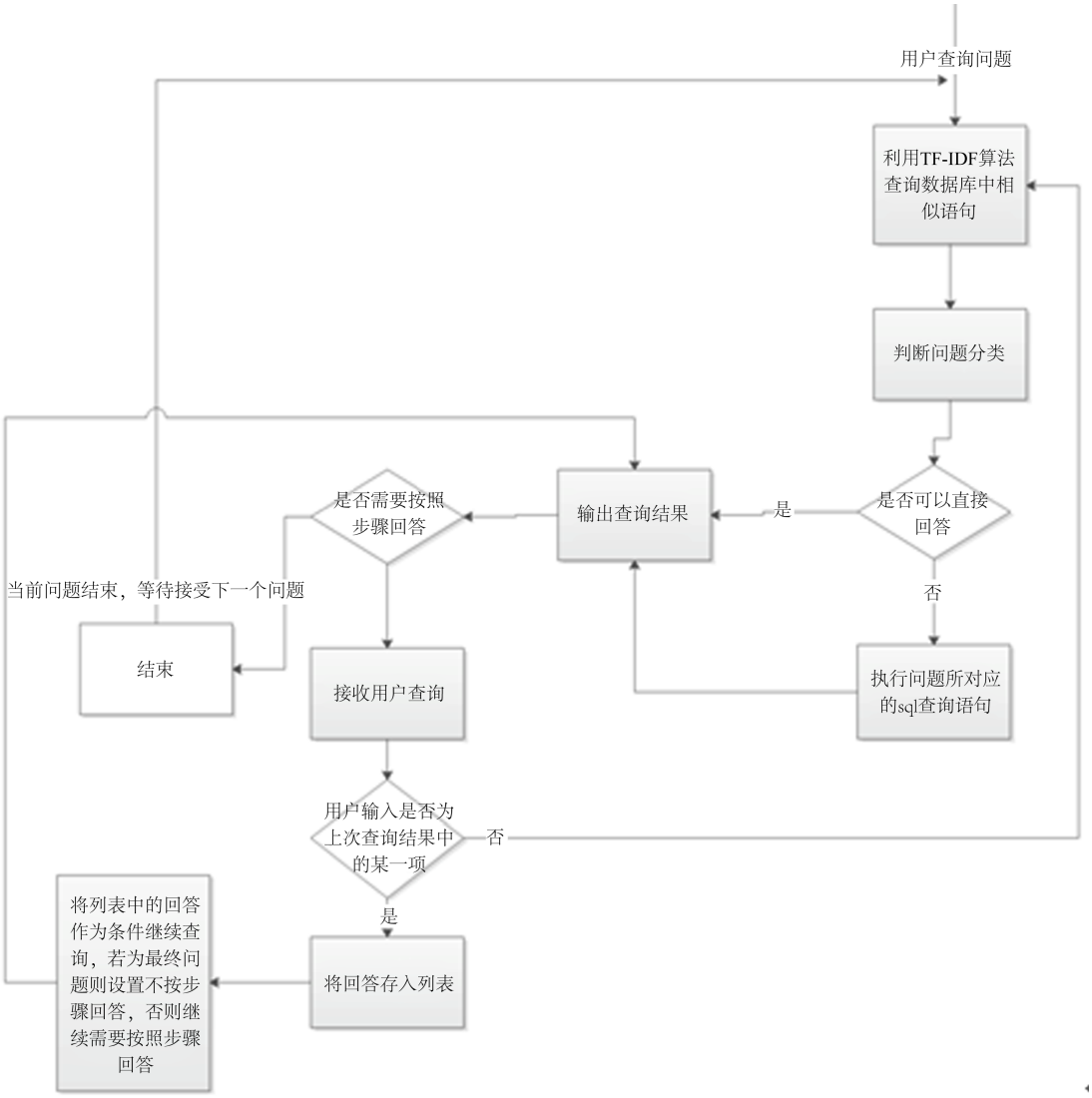

Figure 6. Intelligent Q & A flow chart
图 6. 智能问答流程图

以下是根据您的选择为您绘制的图表

长	宽	高	规格	单面覆亮膜	单面覆哑膜	打孔	单面击凸	双面覆亮膜	双面覆哑膜	压痕	横切II异型	单面烫金	单面烫银	价格
13.0	15.0	4.0	80 g	否	否	是	否	否	否	是	否	否	否	47.59
12.0	12.0	14.0	80 g	否	否	是	否	否	否	是	否	否	否	50.63
6.0	7.0	14.0	80 g	否	否	是	否	否	否	是	否	否	否	34.38
6.0	5.0	13.0	80 g	否	否	是	否	否	否	是	否	否	否	29.85

前一页 下一页

上表格是与您的选择相关的历史订单详情，根据上述参数，我们为您的估价为：4.750418313099942，
分析得出回归方程：价格=1.0166788226994439×长+1.9765001399532647×宽+0.9921761431721903×高+0.7655382491063534

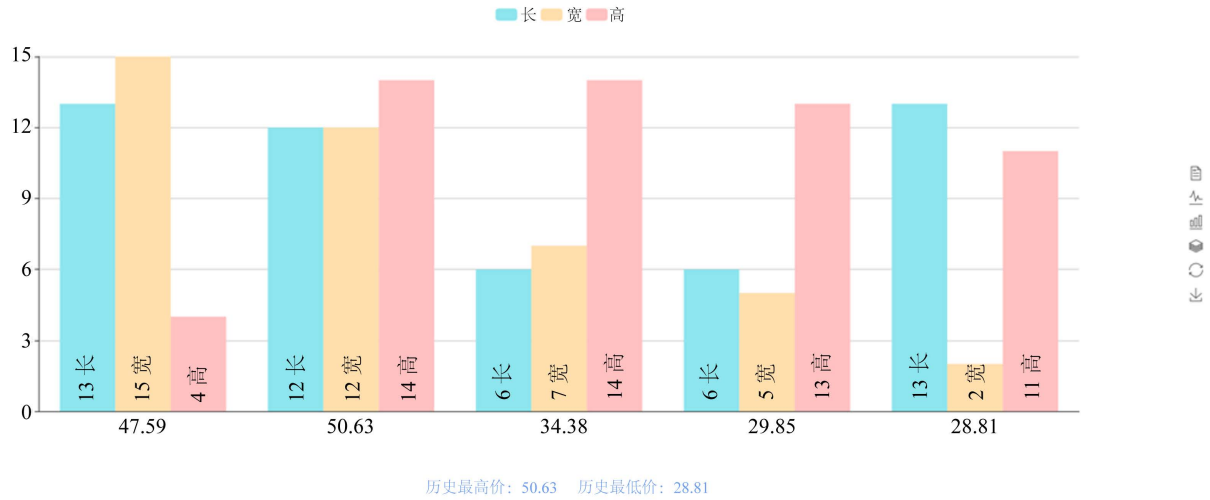

Figure 7. Display of intelligent quotation system
图 7. 智能报价系统展示

6. 结束语

本系统的设计主要在于为用户提供一个关于印刷产品智能报价的功能，并且根据用户的合法输入找出历史订单中相关性较强的数据，对这些数据运用最小二乘法计算拟合得出多元回归方程，通过该方程并结合 WEB 页面其他信息综合估计出报价，并将从 WEB 中爬取的这些信息显示在 WEB 页面供用户做参考，最后设计了一个智能问答系统，用来回答用户输入的问题。在整篇论文中，对系统的模块，以及各个模块的功能作出了详细的介绍，并介绍了其中的主要算法，最小二乘法，TF-IDF 算法，以及如何运用最小二乘法给出价格评估，如何利用 TF-IDF 算法匹配相似问题并应用到智能问答模块。并简要介绍了网络爬虫的基本流程，以及如何运用 java 爬虫技术。最后通过第三方 javascript API，提供强大的图像制作方法。

基金项目

天津市智能制造专项资金项目 201810602，201907206，201907210；天津市互联网先进制造专项资金项目 18ZXRHGX00110。

参考文献

[1] 王晓东. 算法设计与分析[M]. 北京: 清华大学出版社, 2014.
[2] 耿祥义, 张跃平. JSP 大学实用教程[M]. 北京: 电子工业出版社, 2007.
[3] 百度有限公司. Echarts.JS. https://echarts.baidu.com
[4] 周秀媛, 陈娜, 李晓斌. 基于“HTML” “5”的“Web”交互界面设计[J]. 科技展望, 2016, 26(18): 6.
[5] 余节约, 田培娟. 印刷工艺原理[M]. 杭州: 浙江大学出版社, 2010.
[6] 胡海波. 电子变压器加工报价软件设计[J]. 科技创新导报, 2017, 14(32): 126-127.
[7] 陆珂. 制造企业的产品快速核报价系统设计——以无锡 VLK 有限公司为例[J]. 中国高新区, 2017(17): 19-20.
[8] 米登斌. 基于加工特征的产品报价系统开发及其关键问题研究[D]: [硕士学位论文]. 合肥: 合肥工业大学, 2017.
[9] 许力. 智能控制与智能系统[M]. 北京: 机械工业出版社, 2007.
[10] 戴汝为, 王珏. 巨型智能系统的探讨[J]. 自动化学报, 1993(6): 645-655.
[11] 罗文, 瞿少成, 程建平, 陈梦婷. 基于 Android 的纸盒智能排版及快速报价系统[J]. 计算机应用与软件, 2019,

36(5): 25-28+50.

[12] 徐春雷, 周竞, 余璟, 吴海伟, 王勇. 基于强化学习模型的需求侧用户智能报价策略研究[J]. 智慧电力, 2018, 46(10): 32-37.

[13] 黄群钿, 章绵生. PCB 智能化报价系统的实现[J]. 印制电路信息, 2011(5): 49-52.

[14] 李义华, 李夏苗. 基于多智能体的供应链报价协商系统及其实现[J]. 中南林业科技大学学报, 2010, 30(2): 107-111.

[15] 曹素娥. 基于爬虫技术的就业信息管理平台设计[J]. 电子技术与软件工程, 2019(18): 47-48. http://kns.cnki.net/kcms/detail/10.1108.TP.20190927.1149.056.html

[16] Terasaka, S., Kikuchi, U. and Torii, K. (2019) Radiation Imaging Using a Compact Compton Camera Mounted on a Crawler Robot Inside Reactor Buildings of Fukushima Daiichi Nuclear Power Station. *Journal of Nuclear Science and Technology*, **56**, 801-808. https://doi.org/10.1080/00223131.2019.1581111

[17] 熊慧芳. 网络爬虫关键技术的应用探讨[J]. 计算机产品与流通, 2019(9): 171.

[18] 肖黎, 张彩霞. 基于主成分回归分析的我国农产品贸易逆差影响因素研究[J]. 中国经贸导刊(中), 2019(9): 11-14+76. http://kns.cnki.net/kcms/detail/11.3876.f.20190927.1341.006.html

[19] 代磊, 李雪婷. 基于多元线性回归模型的二手房价格影响因素分析——以成都市某区为例[J]. 河南建材, 2019(5): 80-82.

[20] 张会昌. 基于领域词典的中文文本相似度匹配[D]: [硕士学位论文]. 济南: 山东大学, 2014.