Building of Visual Analysis System for Design of Youth Sports Shoe Products Based on Comment Mining

Shengbo Qian¹ᵃ, Kaixin Zhang²ᵇ
¹Hefei Thomas School, Hefei, China
²University College London, London, The United Kingdom
ᵃsimplewy@126.com,ᵇZhangkaixin1021@163.com

Abstract: This paper proposes a visual analysis system for data collection, pre-preprocessing and sentiment analysis to support whole-process innovative design of youth sports shoes. This system aims to analyze the sales data of sports shoe e-commerce sellers before and after design, provide reference for key design elements in innovative design of sports shoes based on 3D printing and realize dynamic closed design of pre-design data investigation and analysis and post-design feedback data analysis. With dynamic data collection technology based on a crawler, product feature extraction model, sentiment analysis model and big data visualization technology, this system can better satisfy data analysis requirement for innovative design of sports shoes.

1. INTRODUCTION

With development of the society, everyone pays much attention to health. Thus, the demand for different shoes increases much on the market. With quick development of the market, it also expedites development of domestic sports shoe brands. Especially the demand of the youth sports shoes is very huge and the requirements of the styles are very strict. It represents the popularity trend of sports shoe. Based on predictions of the American investigation agency Grand View Research, the market scale will exceed 95 billion USD in 2025. There are plentiful sport shoe enterprises in China and China is the largest sports shoe producer and exporter in the world. Now the demand of Chinese users for sports shoes is transforming from single batch to function and personalization. Such diversified and variable user requirements bring huge challenges to development of new sports shoe products [1][2].

During recent years, with quick development of e-commerce, e-commerce websites provide more and more products. More and more youth consumers are accustomed to purchasing sport shoe products via the e-commerce websites. To improve product satisfaction of consumers and improve shopping experiences of consumers, most of e-commerce websites provide special areas for consumers to publish comments on purchased products. Thus, comments on network products are quickly growing. More and more evidences prove that comments will affect purchase decision of consumers [3]. In addition, global marketing science is transforming from commodity-oriented logic to the user service-oriented logic. Innovation modes of enterprises also changes in a disruptive manner and extends from pure dependence on internal users to users of enterprise users. Thus, user innovations become a hot topic in research on marketing and innovation management theory [3]. In practice, with maturity of Internet technology and extensive applications of social software, some enterprises start to find high-value innovation ideas from the network[4] and present personalized products[5]. Thus, it is very important to perform data...
mining and analysis on the related comments of the network sports shoe purchasers prior to design and comments on designed samples and mine product features focused by consumers and their sentiment opinions on these product features. These information can be used to guide design of sport shoe products [6].

Based on the school-class innovation project “Sport Design and Making”, the online dynamics and history e-commerce sales data on sports shoe are collected during recent years, the statistical analysis method and data mining method are used to analyze data collected by using the network questionnaire forms and online crawlers. Finally, we use SolidWorks software and 3D printing technology to design sports shoe based on the consumer preferences on product features. Based on our design, we make a personalized sports shoe product based on combination of the best features, and continuously optimize design based on analysis on the network comment data of designed samples.

2. CURRENT CONDITIONS OF RESEARCH ON E-COMMERCE PRODUCT COMMENT ANALYSIS

Before the sports product is designed, it is required to fully know preferences of target groups. This project aims to design a sports shoe for youth. Before it is designed, we select historical sale data of several focused e-commerce brands to analyze preferences of key design elements of sports shoe. With the network, the network investigation questionnaire is designed to collect the investigation data at target schools. Based on collected historical sales data and investigation data, we design a sports shoe by using SolidWorks and 3D printing technology, make a sample at the factory, and pre-sell this product sample at the e-commerce website. We collect comment on sold product samples by using the crawler. Thus, the product comment mining is the key technology in system building. By collecting user’s comments on products published at the Web, the natural language processing technology is used to find user’s comments on product performance from plentiful text data. The comment mining involves research on two aspects, including product feature extraction and comment opinion extraction and attitude positivity determination [7][8]

Product feature extraction: It indicates to extract product features from product’s use comments. The product features include product properties or function, product component, properties or function of product component and related product concepts. The product features can be automatically extracted by using natural language technologies such as part of speech tagging, syntax analysis and text pattern to analyze sentences in product comments and automatically discover product features. Quan et al [9] extracted product features without supervision by measuring the semantic similarity distance of the domain vector. Wang et al [10] proposed one hybrid association rule mining algorithm to identify hidden features. Li Shi et al [11] proposed the product feature extraction algorithm based on Apriori association analysis algorithm from the features of Chinese languages. Ma Baizhang et al [12] trained the text model by using the LDA (Latent Dirichlet Allocation) and got the product features by combining the Synonymy Thesaurus and filtering rules.

Comment opinion extraction and attitude positivity determination: It indicates to extract opinions in user’s comments on products and determine attitude. MISSEN et al [13] computed the similarity between the sentences to identify and subjective sentences by using the machine learning method to determine if sentences to identify are subjective sentences. Kim and Hovy established the feature word form of the subjectivity determination by using the manually defined method [14]. Ye Qiang et al [15] proposed the continuous 2-part-of-speech combination mode to automatically determine subjectivity degree of sentences in the Chinese language field. Kim [16] used the sentence positions and sentiment words to classify sentence sentiment as the classification feature based on the traditional N-gram model.

3. FRAMEWORK OF VISUAL ANALYSIS SYSTEM OF SPORTS PRODUCT COMMENTS

The visual analysis system of sports shoe product based on comments is a visual analysis platform for design of sports shoe product. The whole system is designed as B/S architecture and includes comment information collection module, data pre-processing module and visual display module. The system’s functional structure and implement framework is shown as the figure 1. The business process of this system is described as follows: First, parse the specified product comment data from product comment
pages by using the comment information collection module and store them to the database. Later, the data pre-processing module cleans data to get the standardized data. Finally, import cleaned data into the visual display module for visual analysis of comment information of sports shoe products.

3.1. **Comment Information Collection Module**
The comment information collection module mainly collects information at the focused sports product websites, including name, color and styles of sports shoe products, comment content, product score, praise number, comment time, comment properties, crawling time, and stores these data into the database.

3.2. **Data Pre-Processing Module**
The data pre-processing module mainly cleans collected product comment data, including HTML labels, meaningless English characters, digital characters and expression text characters and segment the comment corpus according to punctuations (full stop, comma, etc.) and space character. The word segment system ICTCLAS2011 of the Chinese Academy of Sciences is used to segment sentences in the comments and tag part of speech to get the sentence library. The nouns or noun phrases (including gerunds and adjective nouns) are selected as the product features and the transaction files of the association rules can be obtained via the part of speech filtering.

3.3. **Visual Analysis Module**
The visual analysis module counts and analyzes pre-processed standardized product comment data and visualizes regular data, including statistical analysis on product category information and statistical analysis on specific products. In addition, this module can analyze subject in the comment texts and display analysis results on the user interface via the visual interface.

4. **IMPLEMENTATION OF VISUAL ANALYSIS SYSTEM OF SPORTS PRODUCT DESIGN BASED ON COMMENT**

4.1. **Data Collection**
By analyzing the webpage source code structure of sports shoe products, the product comment information is asynchronously loaded by using Ajax and is saved as texts. Thus, specific crawler programs are written to collect related data. By writing python network crawler scripts, we deal with regular anti-crawler strategies by using the Selenium automatic sequence test tool [17][18]. This tool
can simulate a person to browse webpages, capture required product comment data, and store them to local database. For the data collection process, refer to the figure 2.

FIG. 2 DATA COLLECTION PROCESS

The breadth-first strategy is used as the data collection strategy in the data collection process in the figure 2, namely first position the category url of sports products, crawl source codes of this page, and get the url of specific products. The Web Driver is called by using Selenium to start the local browser and simulate manual operation in crawling. Data is dynamically loaded on this product platform. Thus, the manual clicking operations can be simulated to get returned real-time dynamic data and get positioning node data via Xpath. The front-end webpage source codes of comments are shown as follows:

```
<div class="comment_list">
  <h3 style="text-indent: 2em;">Comment list</h3>
  <hr>
  <div class="comment">
    <div class="imgdiv">
      <img class="imgcss" src="/images/user.jpg"/>
    </div>
    <div class="conmment_details">
      <div style="float:left;">
        <span class="comment_name">Dabai </span> <span>Ahead of 22 mins</span>
      </div>
      <div class="del">
        <span class="show_reply_list">View replies</span> <i class="icon layui-icon">
          Praise (164)</i>
        <a class="del_comment" data-id="1"> Delete </a>
      </div>
      <div class="comment_content">Except some problems in express delivery, this shoe is very beautiful, is very ok, is light in wearing, and is comfortable in running, OK</div>
    </div>
  </div>
  <div class="reply_list">
    <div class="reply">
      <span class="reply_name">Xiaobai </span>Reply <span class="reply_name">Xiaohei </span>: <span class="reply_content">Hello </span>
    </div>
  </div>
</div>
```
After the crawler program parses the webpage source codes, the pyquery is called in python to locate and match corresponding strings, further parse data on product comments and store these data into the local database. The key codes of this process are shown as the figure 4.

```python
import requests
import time
import random
import re
import json

url = 'https://rate.tmall.com/list_detail_rate.htm'

f = open('votes.csv','w',encoding='gbk')
f.write('comment content, reply of shop owner, nickname \n')

for i in range(99):
    t = str(time.time()).split('.
    pagram = {
        'currentPage': i+1,
        '_ksTS': '%s_%s' % (t[0], t[1]),
        'callback': 'jsonp%s' %(int(t[1])+1)
    }
    time.sleep(random.random())
    response = requests.get(url, params=pagram)
    data = response.text
    data = re.findall(r'{.*}', data)[0]
    data = json.loads(data)
    data = data['rateDetail']['rateList']
    print(data)
    for item in data:
        f.write('%s,%s,%s' %
            (item['rateContent'].replace(',', ', '),
             item['reply'].replace(',', ', '),
             item['displayUserNick']))
```

**FIG. 4 KEY CODES FOR PARSING COMMENT DATA**

Above codes can extract product names, consumer nickname, product comment, comment contents, comment praise and comment publishing time in comments from web pages and later the crawled data are stored to local My SQL database. The crawled and stored data are shown as the figure 5.

**TABLE. 1 INSTANCE OF CRAWLING DATA**

| Product-name       | Nickname | Comments                                                                 | Time         |
|--------------------|----------|--------------------------------------------------------------------------|--------------|
| White customized   | Xiaobai  | The matched color is very beautiful, the sole hardness is proper, and I like it very much | YY-MM-DD     |
| sports shoe        |          |                                                                          |              |

![FIG. 3 EXAMPLE OF COMMENT WEBPAGE SOURCE CODE](image-url)
White customized sports shoe  
Xiaoai  
The pattern style is simplified, the vamp is ventilated well, and is comfortable in wearing  
YY-MM-DD

White customized sports shoe  
Xiaohui  
This shoe has a heel with proper height, is comfortable in walking, and its style is personalized.  
YY-MM-DD

4.2. Pre-processing of Sports Shoe Product Comment

(1) Clean data: The crawled contents may include plentiful HTML tags which shall be removed. A small number of non-text contents can be deleted by the regular expression (Re) of Python and complicated non-text contents can be removed by using Beautiful Soup library. Except HTML tags, the corpus includes replies using partial expression symbols and meaningless digit and English characters and punctuations. These data are not favorable to model training and shall be removed. The original comment data is shown as the table 2 and cleaned data is shown as the table 3.

| TABLE 2 ORIGINAL COMMENT DATA EXAMPLE OF ANTA SPORTS SHOE |
| SN | Subject text | Product configuration |
|----|--------------|-----------------------|
| 1  | Except some problems in express delivery, this shoe is very beautiful, is very ok, is light in wearing, and is comfortable in running, OK | Tone: tint  
Matched color: white and blue and white and black  
Pattern style: brand mark  
Heel height: middle-height heel  
Sole hardness: soft sole |

| TABLE 3 EXAMPLE AFTER DATA CLEANING |
| SN | Original text | Processed texts |
|----|---------------|----------------|
| 1  | Except some problems in express delivery, this shoe is very beautiful, is very ok (Non-Chinese characters), is light in wearing, and is comfortable in running, OK (Non-Chinese characters) | Except some problems in express delivery, this shoe is very beautiful, is light in wearing, and is comfortable in running |

(3) Feature word extraction and processing module: This module mainly extracts sentiment words in the comment texts by using association rule algorithm, e.g. feature words such as speed, freight fee and prices and adjectives such as good, bad and worse. After processing, the comment text data is shown as the table 4.

| TABLE 4 EXAMPLE AFTER EXTRACTION AND PROCESSING OF FEATURE WORDS |
| SN | Original text | Processed texts |
|----|---------------|----------------|
| 1  | Except some problems in express delivery, this shoe is very beautiful, is light in wearing, and is comfortable in running | Express delivery has some problems, this shoe is very beautiful, is light in wearing, and is comfortable in running |

(4) Part of speech tag processing module of comment text: This module segments comment texts into words by using the words segment system of the ICTCLAS system, tag the corresponding part of speech, and delete meaningless words according to “Stop Word Library”. The processed comment text data are shown as the table 5.

| TABLE 5 EXAMPLE AFTER PART OF SPEECH TAGGING |
| SN | Original text | Processed texts |
|----|---------------|----------------|
| 1  | Express delivery /n some /k | Express delivery /n problem |
problem/n, shoe/n very/k beautiful/a, wear /v well /a light /v, run /n comfortably /a
/n, shoe /n very/k beautiful/a, wear /v well /a light /v, run /n comfortably /a

(5) Text dependency structure processing module: This module mainly analyzes dependency relation between components in the text language units by using the text part of speech and generates the set of comment text dependency structure, shown as the table 6.

| SN | Original text                                      | Processed texts                                      |
|----|----------------------------------------------------|------------------------------------------------------|
| 1  | Express delivery /n problem /n, shoe /n very/a beautiful/v, wear /v well/a light /v, run/n comfortably/a | (Express delivery, problem), (<EOS>, beautiful), (wear, well), ( (<EOS>, wear), (run, comfortably)… |

(2) Design of stop word table: The frequency distribution of stop words such as statement label designator and pronoun is similar in the obtained product comment data. On the one hand, it increases the similarity between texts and is unfavorable for clustering. On the other hand, the use frequency of the stop words is very high, they will occupy the density of the feature words meaningful for clustering and make feature words insignificant. Thus, they shall be removed. The stop word table used in this paper is formed by consolidating and screening the stop words tables from Harbin Institute of Technology, Machine Learning Intelligence Lab of Sichuan University and Baidu stop words table. Considering significance of the sports shoe product in the context, some adverbs are reserved to indicate degree, e.g. “very”, “very much”, “significantly”, “equivalently”, “highly” and “emphasize”, which compose the stop word table of the sports shoe products as the supplementary.

4.3. Visual Analysis of Sports Shoe Product Comments

4.3.1. Regular data visualization: The regular data visualization indicates to count, analyze and visually display design data of sport products by field properties based on comments. For this project, visual analysis includes automatic statistical analysis of comment consumers, automatic statistics of product score, visual distribution of product feature attention and time distribution of product activity. The target users can be pictured by basic user properties such as gender, age, area and user behavior, purchase list, collection list and shopping car list. We can get the buyer’s feedback to commodities from the buyer score and comment data. The figure 5 and 6 show the product type trend and consumer age distribution based on crawling data.

FIG. 5 SALE TREND OF SPORTS SHOE TYPE (SALE DATA IN 2018-2019)
4.3.2. **Subject discovery visualization:** The Visualization of the subject model based on comments aims to mine key words and subject information from the design information cleaned sports shoe products based on comments in the visual analysis module of this system, form effective clustering subjects by using certain model computing method, and visually display them. As a unsupervised machine learning method, the Latent Dirichlet Allocation (LDA) can find hidden pattern in the texts with the Bayes probability at the document, subject and word layer as the kernel structure. It does not only count word frequencies and regards the text contents as the multinomial distribution of the subjects. Each subject can be regarded as the multinomial distribution of the words. It can simulate document generation and compute the hidden subject composition in the document and subject tags of the part of speech [14]. Thus, this system discovers the subjects of the comment data of the sports shoe product based on LDA subject model, writes the python codes to integrate LDAvis [15] to visualize subject model information in the comments. For specific visual results of the comment subject and product features, shown as the figure 8 and 9.

**FIG. 6 AGE DISTRIBUTION OF CONSUMERS OF DIFFERENT SPORTS SHOE PRODUCTS (2018-2019)**

**FIG. 7 SPORTS SHOE SUBJECT EXTRACTION BASED ON LDA**
The left diagram shows the subject modeling results of the LDA model. Each circle indicates a subject, the circle size indicates the scale of text content of this subject and the circle distance indicates the similarity between subjects. When one subject is selected, the right diagram shows the top N words in this subject.

5. CONCLUSIONS
With rising of the maker culture, more and more innovative products emerge in different industries. The originality of any commodity shall be from the market demand and innovative improvement. By analyzing comment data of related products, questionnaires and third-party sales data prior to design and analyzing online sale feedback comment of designed samples, the product design elements can be selected more scientifically. A visual analysis system based on sports product comments is designed and implemented in this paper. This system crawls and pre-processes the information on sports shoe products to visualize the regular data of product comments and subject discovery. This system provides references and supports for design originality and product improvement of maker’s sports shoe via automatic collection, statistics and comment data visualization. The functions of the sports shoe visual analysis system of sports shoes based on comments designed and developed in this paper can be further extended and the adaptation and efficiency of algorithms and multi-dimensional analysis functions shall be further optimized for higher practicability.

REFERENCES
[1] Song Shuguang. Thorough Analysis on Mass Sports Value Orientation from Sports Shoe Culture [J]: Sports and Sciences, 2015 (5) : 79- 86
[2] Zhang Dengfeng. Investigation on Sports Shoe Consumption Conditions of Middle-School Students and Reason Analysis, 2009 (1) : 598- 604.
[3] CUI A S, WU F. The impact of customer involvement on new product development: contingent and substitutive effects[J]. Journal of Product Innovation Management, 2016, 34(1): 60-80.
[4] BILGRAM V, BREM A, VOIGT K. User-centric innovations in new product development—systematic identification of lead users harnessing interactive and collaborative online tools[J]. International Journal of Innovation Management, 2008, 12 (3): 419-458.
[5] Xiao Jinghua, Wu Yao and Liu Yi et al. R&D Innovation of Consumer Data Participation: Dual-Case Research on Collaborative Evolution View of Enterprise and Consumer [J]. Management World, 2018, 34 (8): 154-173.

[6] Song Xiaobing, Xu Kexin and Wu Yuzhen. Can User Design Try to Do Everything? Research on Influences of Self-Building on Preferences of User-Designed Products [J]. Management World, 2017(5) : 119-130.

[7] Yang Li. Sentiment Tendency Identification and Topic Mining Technology of Online Product Comments[J]. Electronics Technology and Software Engineering, 2018 (1) : 185.

[8] Wang Yong, Zhang Qin and Yang Xiaojie. Research on Product Feature Extraction Method in Chinese Network Comments [J]. Modern Book Intelligence Technology, 2013, (12):70－73.

[9] QUAN C Q, REN F J,Unsupervised product feature extraction for feature-oriented opinion determination[J].Information Sciences, 2014, 272(10): 16-28.

[10] WANG W, XU X, WAN W. Implicit feature identification via hybrid association rule mining[10].Expert Systems with Applications, 2013, 40(9): 3518-3531.

[11] Li Shi, Ye Qiang and Li Yijun. Research on Product Feature Mining Method of Chinese Network Customer Comment [11]. Management Science Journal, 2009, 12(2): 142-152.

[12] Yan Zhijun and Ma Baizhang. Feature Extraction Method of Network Comment Products Based on Latent Dirichlet Allocation Distribution Model [12]. Computer Integrated Manufacturing System, 2014, 32 (1): 96-103.

[13] MISSEN M M S, BOUGHANEM M, CABANAC G. Opinion mining: reviewed from word to document level[J].Social Network Analysis and Mining, 2013, 3(1): 107-125.

[14] Kim S M, Hovy E. Identifying and analyzing judgment opinions[14] Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, New York, USA, 2006.

[15] Ye Qiang, Zhang Ziqiong and Luo Zhenxiong. Research on Automatic Determination Method of Chinese Subjectivity for Internet Comment Sentiment Analysis [J]. Information System Journal, 2007, 1(1): 79-91.

[16] KM S, HOVY E. Determining the Sentiment of Opinions[C] Proceedings of the 20th international conference on Computational Linguistics.Stroudsburg, PA, USA: Association for Computational Linguistics, 2004: 1367-1373.

[17] Zeng Weihui. Design and Implementation of Network Crawler System Supporting AJAX [D]: [Master’s Degree Thesis]. Anhui: University of Science and Technology of China, 2009

[18] Sui Junxiong. Research on Website Information Collection Technology based on Network Crawler [D]: [Master’s Degree Thesis]. Liaoning: Maritime Affairs University Of Dalian, 2014