# Research on Network Media News Visualization Based on FDCD-TFIDF Weighting Algorithm

**Luqi He, Yuan Long, Ning Wang*, Zhanxiang Ma, Zeping Ma**
*School of Data Science, Guangzhou Huashang School, Guangdong, 511300, China*

**Abstract: In the era of big data, online media has become an important source of news. However, the sheer volume of data makes it difficult for users to extract useful information from it. Therefore, news visualization has become an important means to help users understand the world more easily and quickly. Taking sports news data as an example, this paper uses FDCD-TFIDF algorithm to carry out important news weights on network media news in the era of big data, so that news push users can be more accurate and grasp current hot news more accurately. Meanwhile, this paper also explains the principle of FDCD-TFIDF and why in this research, the algorithm is more suitable for this kind of research. The accuracy of visual interface results also verifies the accuracy of the algorithm.**

**Keywords: Python; Data Visualization; Big Data; FDCD-TFIDF; Visualization**

## 1. Introduction

In the era of digitization, informatization and networking, big data has penetrated into every place of our lives, and more and more people use the internet to get sports news[1]. with the entry of the era of big data, data has become an important production factor in the field of news, but at the same time, the difficulty for people to obtain key information from news has greatly increased, and how to efficiently and accurately process and present these massive data has become an important issue for people in the internet era. the emergence of data visualization technology provides a new way to deal with this problem. therefore, exploring how to visualize sports news data through network media, so that users can obtain and analyze news data more intuitively and conveniently, has become a topic worth studying.

at present, a lot of research in the world is exploring how data visualization technology can be applied to the field of journalism. in some countries, the relevant research is relatively early, and the application of data visualization technology is more mature. for example, espn[2] in the united states uses a large number of data visualizations on its website to present sports news data. with the advent of the era of big data, more and more researchers begin to pay attention to the application of data visualization in the field of sports news. However, at present, there is still a lack of relatively research and perfect research results, this topic will take sports news as an example to explore and study on this basis.

the research object of this topic is the data visualization technology of online media sports news. by selecting the data sets of basketball and football, this paper studies the use of data visualization technology to present online media sports news, so as to provide readers with more intuitive sports news reading experience. specific operations are as follows:

1. select appropriate data sets and conduct data processing and analysis;
2. design the corresponding data visualization research according to the analysis results;
3. implement and test the function and performance of data visualization research.

## 2. Data Feature Extraction

TF-IDF (Term Frequencyinverse Document Frequency, Word frequency-Inverse document frequency)[3] is a text statistical method for assessing the importance of a word in a file set or corpus in a document. The basic idea is that the importance of a word is directly proportional to the number of times it appears in the current document, and is inversely proportional to how often it appears in the entire file set. The more common a word is, the lower the IDF value is[4]. Multiplying the TF and IDF gives the TF-IDF value, and the higher the value, the higher the word is in this file, and the more likely it will be the key word of the article.

TFIDF is currently the most commonly used feature-weighting processing method nowadays, but it has certain limitations, mainly including the following two points:

### 2.1 The Imbalance in the Distribution between the Categories of the Datasets Was Not Considered

Most of the category distribution in real data sets is unbalanced, and different categories often have certain differences. However, the TFIDF algorithm does not take into account this difference, and the calculated feature vector weights are only based on the number of documents.

When the category distribution of the data set is quite different, especially for the weak category distribution, the calculated weight value will be too small, which will affect the classification accuracy, and cannot correctly reflect the distribution difference of the text vector between the various categories in the data set.

### 2.2 There Is No Correct Response to the Difference in the Distribution of the Text Vector in the Classification System

The inter-class distribution of text vector needs to consider that when the feature item ti has a large word frequency tfij value in class Cj, and the word frequency tfij value in other classes is small, the feature item should well reflect the degree of difference in text category, and should be given high weights.

Low weights are given when present in most categories and with little proportion in the categories. The intra-class distribution of text vectors needs to consider that when the feature term ti is within the Cj class, the feature term with a more consistent within-class distribution should be given high weights. When the feature item appears in only a few documents of the same kind, and rarely appears in such other documents, the feature item may be a special term, and can not reflect the category information of the text well, and should be given low weights[5]. Therefore, this section cites a TFIDF improvement algorithm, FDCD-TFIDF, based on the word frequency distribution factor and the category distribution factor.

Interclass distribution factor (Inter-class distribution factor):

Reflects the distribution of feature items between document classes and classes. It can be obtained by calculating the quotient between the number of documents containing feature item ti, aij, in the Cj class and the non-Cj class containing feature item ti, with the following formula:

$$\alpha = \log\left(2 + \frac{a_{ij}}{c_i + 1}\right) \tag{1}$$

Inclass distribution factor (Intra-class distribution factor):

Reflects the distribution of feature items between document classes and classes. It can be obtained by calculating the quotient between the number of documents containing feature item ti, aij, in the Cj class and the non-Cj class containing feature item ti, with the following formula:

$$\beta = \log\left(2 + \frac{a_{i,j}}{b_j + 1}\right) \tag{2}$$

Category Distribution Factor (Category distribution factor):

Reflects the distribution information of the various categories of the document. It can be obtained by dividing the total number of documents in the data set by N divided by the quotient of the total number of documents nj contained in the category Cj, defined as follows:

$$\gamma = \log\left(\frac{N}{n_j}\right) \tag{3}$$

So the improved weight calculation formula is as follows:

$$\mathrm{FDCD} - \mathrm{TFIDF} = \mathrm{tf}_{i,j} \times \mathrm{idf}_i \times \alpha \times \beta \times \gamma \tag{4}$$

## 3. Visual Analysis

This module aims to clarify the requirements of the function, data set and performance of the visualization program in the research.

The functional requirements of this study include the following:

1. Data visualization: It is necessary to realize the visualization of the collected sports data, including charts, maps and other different forms of visualization, in order to intuitively understand the data.
2. Data analysis: It is necessary to support the analysis of data, such as trend analysis, correlation analysis, cluster analysis, etc., in order to further explore the regularity and characteristics of data.

3. Data screening and sorting: Users need to be supported to screen and sort data according to their own needs, such as screening and sorting according to time, players, and the second of the competition field, so as to find the required data[6].

The data set required for this study mainly includes the following aspects:

1. Data set of basketball players in basketball events.

2. Dataset of football players in football competitions.

The above data sets need to include basic athlete or news information such as name, nationality, club, competition level, etc.

The performance requirements of this study mainly include the following aspects:

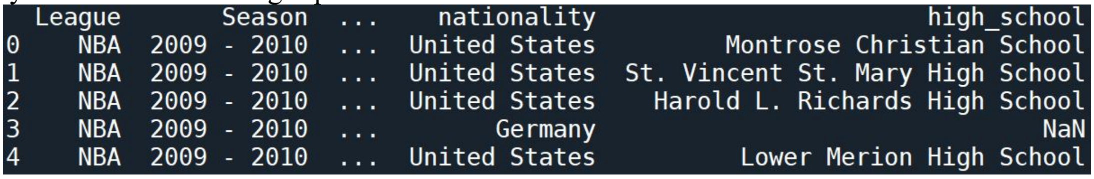
1. Data import speed: The research needs to be able to import a large number of data sets quickly.

2. Speed of data analysis: Research needs to be able to analyze data sets and generate corresponding visual charts in a short time.

## 4. Chapter One Research Design

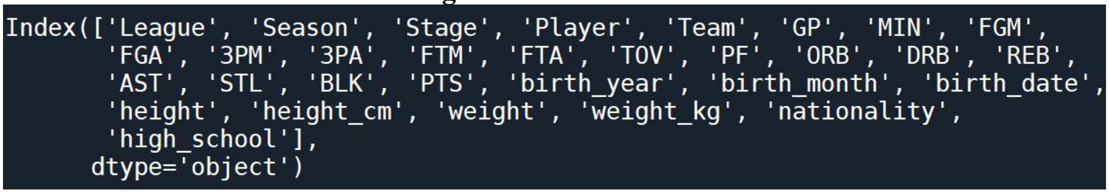
### 4.1 Data Acquisition

In this section, the crawled data is partially displayed. Figure 1 is the crawled data header, Figure 2 is the crawled data column name, and Figure 3 is the data classification name of each season.



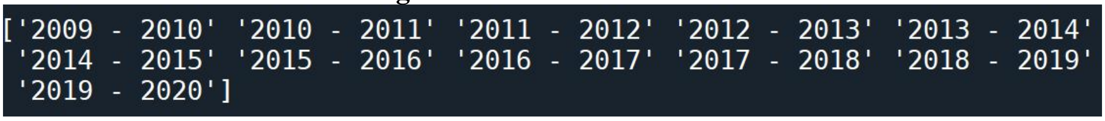
```
  League        Season  ...    nationality                    high_school
0    NBA  2009 - 2010  ...  United States       Montrose Christian School
1    NBA  2009 - 2010  ...  United States  St. Vincent St. Mary High School
2    NBA  2009 - 2010  ...  United States    Harold L. Richards High School
3    NBA  2009 - 2010  ...        Germany                             NaN
4    NBA  2009 - 2010  ...  United States         Lower Merion High School
```

**Figure 1. Data Header**



```
Index(['League', 'Season', 'Stage', 'Player', 'Team', 'GP', 'MIN', 'FGM',
       'FGA', '3PM', '3PA', 'FTM', 'FTA', 'TOV', 'PF', 'ORB', 'DRB', 'REB',
       'AST', 'STL', 'BLK', 'PTS', 'birth_year', 'birth_month', 'birth_date',
       'height', 'height_cm', 'weight', 'weight_kg', 'nationality',
       'high_school'],
      dtype='object')
```

**Figure 2. Data column names**



```
['2009 - 2010' '2010 - 2011' '2011 - 2012' '2012 - 2013' '2013 - 2014'
 '2014 - 2015' '2015 - 2016' '2016 - 2017' '2017 - 2018' '2018 - 2019'
 '2019 - 2020']
```

**Figure 3. Season information**

### 4.2 The Way of Sports News Visualization in Network Media -- Bar Chart

Bar charts are a common form of news visualization, often used to show quantities between individual items or to compare differences in data between different categories. The characteristics of the bar chart can be summarized as follows:

1. Easy to compare and highlight: The bar chart makes it very easy to compare the quantity or data differences between different groups, because they are displayed on the same chart with the same metric, which helps the audience better understand the differences and connections between the data; In addition, you can highlight certain groups or data to help readers focus more on specific information or results, for example, you can set columnar attributes with different thicknesses and stripes, or use darker colors to emphasize changes in data. As shown in Figure 3, this bar chart mainly uses seaborn and matplotlib libraries to draw a bar chart about the number of players in each league. The gradient colored columns used for decoration in the chart well highlight the differences in data, which is precisely an advantage of visualizing data in sports news.

2. Good scalability: When new data or new groups need to be added, the bar chart is usually a very flexible and easy to identify the new data, it can be easily extended to multiple arrays or multiple dimensions, only in the same coordinate system can be changed to quickly compare the differences between multiple groups of arrays.

3. Good presentation effect: bar charts usually have striking visual effects and significant information presentation, making it easier for people to understand and remember the data.

At the same time, they have good aesthetics and readability in the design of colors, label fonts, etc., to help users better understand the data. In Figure 4, the same operation uses the matplotlib library and seaborn library to visualize the data as a bar chart. By changing the color of each column, the presentation of each data item is easier to distinguish than the ordinary bar chart.

To sum up, bar chart, with its clear and concise graphical effect and easy comparison characteristics, has become one of the most popular tools in many network media sports news reports and story data visualization[7]. It can not only convey the basic information of data, but also present the relationship and changes between data, which is a powerful and effective way of data display.

Seaborn library and Matplotlib library were mainly used to count the number of players in each league in nba dataset. Matplolib library was used to build a basic framework for drawing bar charts, and Seanborn library countplot function was used to draw bar charts. The order element is used to specify that the bar chart is arranged in descending order by the number of players.

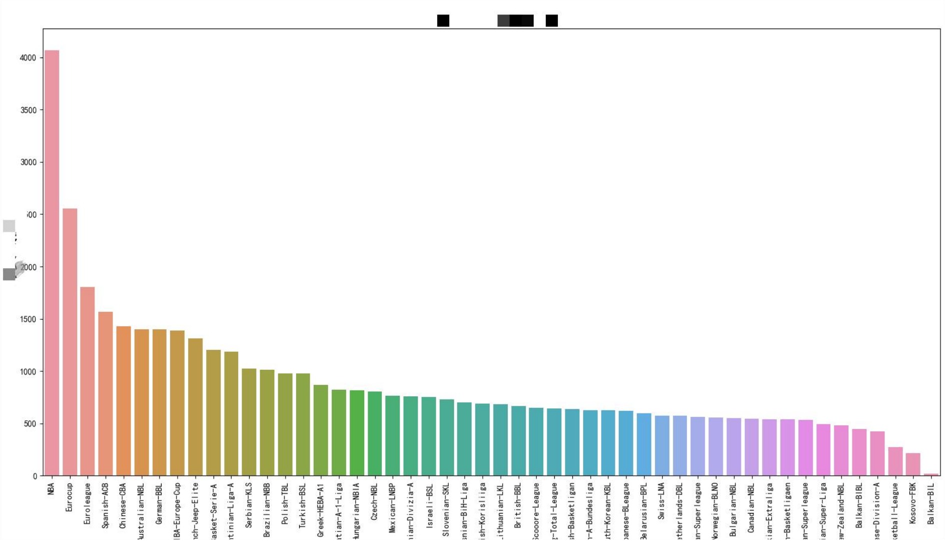
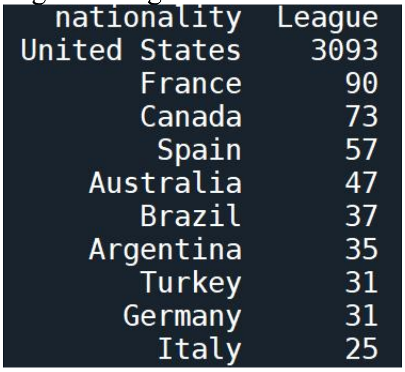


**Figure 4. Number of Players in the League Bar Chart**

Visualization of player nationality distribution in NBA League.Extract NBA game data.The nationality information of players in the nba data set was counted. Groupby function was used to group players of the same nationality and calculate the number of players. Then iloc function was used to screen out the first and second columns, and the number of players from the top ten countries was output and sorted in descending order.Figure 5 shows the results.



```
     nationality  League
United States     3093
       France       90
       Canada       73
        Spain       57
    Australia       47
       Brazil       37
    Argentina       35
       Turkey       31
      Germany       31
        Italy       25
```

**Figure 5. Distribution of Players by Nationality**

### 4.3 The way of sports news visualization in network media -- line chart

A line chart is a common type of data visualization chart that describes trends and changes in data by connecting data points into lines. Line charts are often used to represent time series data, where the horizontal coordinate represents time and the vertical coordinate represents changes in values. The difference between bar chart and line chart is mainly that bar chart is used to show and highlight the differences between different categories of data, while line chart is more suitable for showing the trend of data change to its degree.

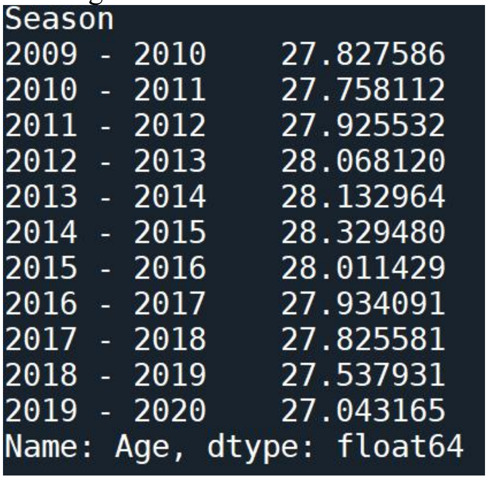
The characteristics of line charts can be summarized as follows:

1. Intuitive: The line chart intuitively presents the trend and change of data, and its presentation of data details can well highlight the change point of data and help readers quickly understand the meaning of data.

2. Easy to compare multiple groups: The line chart can present multiple groups of data, and distinguish different groups of data through different colors or line lines, so that readers can compare and analyze more easily.

Due to its intuitions and readability, line charts are widely used in news visualization, such as displaying data such as stock prices and temperature changes[8]. In the visual application of sports news, as shown in Figure 5 from the competition statistics table of Hupu app, line charts can be used to display the changes of athletes' scores in competitions. It can also show the team's points change in different stages, the trend change of a certain indicator during the season, and so on. When conducting comparative analysis, line charts of different categories can also be superimposed and presented at the same time to visually compare the differences and connections between different categorical variables.

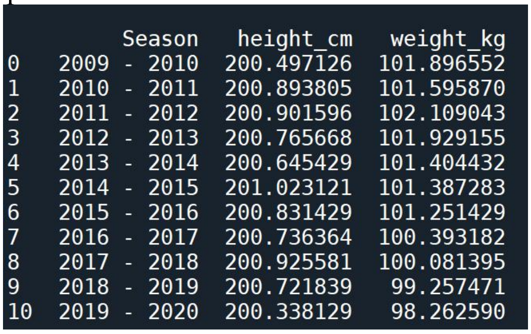
As shown in Figure 6 the average age of NBA players by season. The method of age calculation is to first split the season string of each row according to "-", select the second part as the end year of the season, convert it into floating point data with astype(float), subtract the year of birth from it to get the age of the player in the season, and finally use groupby function to group the column data according to the season. The mean function is used to calculate the average age of the players in each group and output the average age for each season[9].Figure 6 shows the results.



```
Season
2009 - 2010    27.827586
2010 - 2011    27.758112
2011 - 2012    27.925532
2012 - 2013    28.068120
2013 - 2014    28.132964
2014 - 2015    28.329480
2015 - 2016    28.011429
2016 - 2017    27.934091
2017 - 2018    27.825581
2018 - 2019    27.537931
2019 - 2020    27.043165
Name: Age, dtype: float64
```

**Figure 6. Average Age of NBA Players by Season**

In Figure 7, we can check the average weight and height of NBA players in each season. After extracting the season, weight, and height information, groupby() was used to group according to the season, mean() was used to calculate the average value, and reset_index() was used to reset the index, so as to reduce the complexity of multilevel indexes and the duplication of indexes[10].



```
        Season   height_cm   weight_kg
0   2009 - 2010  200.497126  101.896552
1   2010 - 2011  200.893805  101.595870
2   2011 - 2012  200.901596  102.109043
3   2012 - 2013  200.765668  101.929155
4   2013 - 2014  200.645429  101.404432
5   2014 - 2015  201.023121  101.387283
6   2015 - 2016  200.831429  101.251429
7   2016 - 2017  200.736364  100.393182
8   2017 - 2018  200.925581  100.081395
9   2018 - 2019  200.721839   99.257471
10  2019 - 2020  200.338129   98.262590
```

**Figure 7. Average Weight and Height of NBA Players by Season**

Figure 8 shown the change trend chart of the average weight and height of NBA players in each season. Seaborn library and Matplotlib library were mainly used to draw line charts according to the change trend of average weight and height of players in nba data sets in each season. Matplolib library was used to build a basic framework for line chart drawing, and Seaborn library lineplot function was used to draw line charts[11]. It includes the data for the horizontal and vertical coordinates and the data set used, then creates a subplot that shares the horizontal coordinates with ax via the twinx() method, and again draws the subplot using the lineplot() function, where the subplot is set red for differentiation using the color parameter.

Figure 9 shown NBA players with the best points, most assists, most rebounds, most steals, most blocks, most 3-point shots and highest 3-point percentage in each season. First, after calculating the average score based on the main information, all the rows in the Regular_Season are selected[12]. Secondly, groupby is used to group them according to the season, and a transform method is used to identify the scoring champions based on the highest value of 'PTS/G'. Then idx is used to select the rows containing the information of the scoring champions. Finally, select the main information and filter out the information of each scorer according to the above results.

Figure 10 shows the best scoring player line chart for each NBA season. The drawing method is the same as that shown in Figure 10 After importing the basic data, set(ylim=(27,37)) limits the Y-axis range to 27-37 to facilitate the observation of data. Secondly, the iterrows method is used to traverse best_scorers line by line. And use the annotate method to add each scorer's name to the line chart.

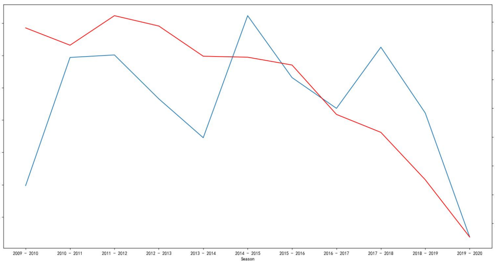
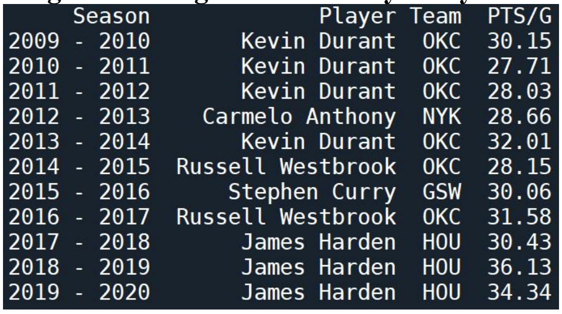


**Figure 8. The Variation Trend of Average Weight and Height of NBA Players by Season**



```
     Season                Player Team  PTS/G
2009 - 2010          Kevin Durant  OKC  30.15
2010 - 2011          Kevin Durant  OKC  27.71
2011 - 2012          Kevin Durant  OKC  28.03
2012 - 2013       Carmelo Anthony  NYK  28.66
2013 - 2014          Kevin Durant  OKC  32.01
2014 - 2015     Russell Westbrook  OKC  28.15
2015 - 2016         Stephen Curry  GSW  30.06
2016 - 2017     Russell Westbrook  OKC  31.58
2017 - 2018          James Harden  HOU  30.43
2018 - 2019          James Harden  HOU  36.13
2019 - 2020          James Harden  HOU  34.34
```

**Figure 9. NBA Top Scorers by Season**

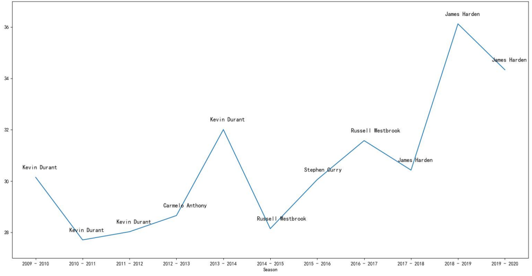
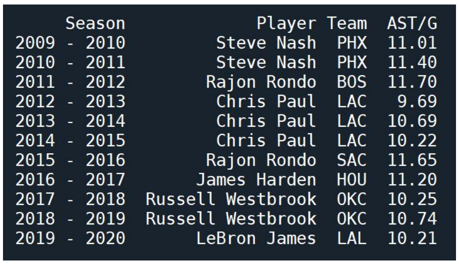


**Figure 10. NBA Top Scoring Players by Season**

Look at the most assisted players in the NBA by season, using the same calculation principle as that shown in Figure 11.



```
     Season             Player Team  AST/G
2009 - 2010         Steve Nash  PHX  11.01
2010 - 2011         Steve Nash  PHX  11.40
2011 - 2012        Rajon Rondo  BOS  11.70
2012 - 2013         Chris Paul  LAC   9.69
2013 - 2014         Chris Paul  LAC  10.69
2014 - 2015         Chris Paul  LAC  10.22
2015 - 2016        Rajon Rondo  SAC  11.65
2016 - 2017       James Harden  HOU  11.20
2017 - 2018  Russell Westbrook  OKC  10.25
2018 - 2019  Russell Westbrook  OKC  10.74
2019 - 2020       LeBron James  LAL  10.21
```

**Figure 11. NBA Players with the Most Assists by Season**

The drawing method is the same as that shown in Figure 11-14. After importing basic data, set(ylim=(9.5,12)) limits the Y-axis range to 9.5-12 for easy observation of data, and then use the iterrows method to traverse best_pas line by line[13]. And use annotate to add the name of each king to the line chart.

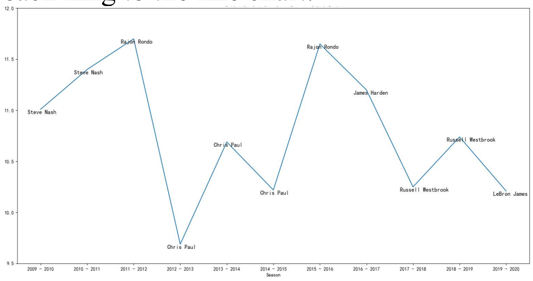


**Figure 12. Most Assists by NBA Season**

## 5. Visual Research Presentation

The homepage is shown in Figure 15.

The part of the Home page includes the name of the website, the introduction of the website and the list bar, and the part selected in the red box is the list bar containing the name of each part, including the home page, League league, Player Information, Player Comparison, Team, etc. Click on the name to jump to the corresponding section to view the news content[14].

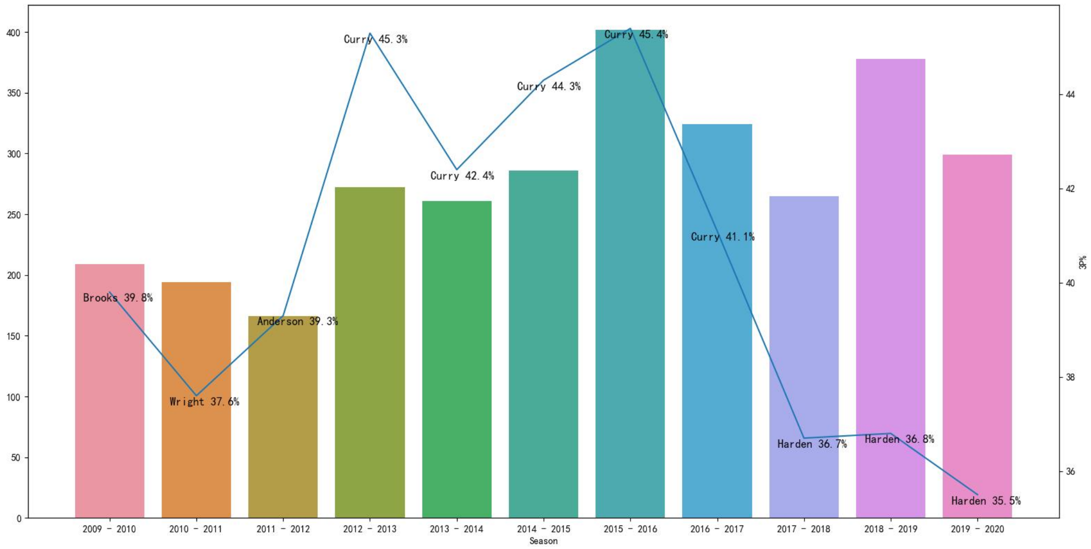


**Figure 13. NBA Players with the Highest 3-Point Shooting Percentage by Season**

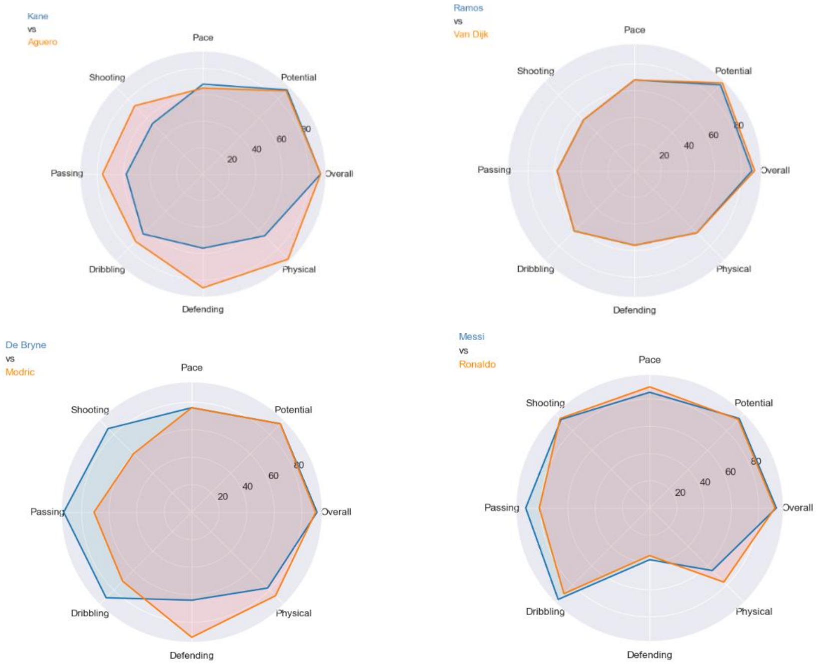
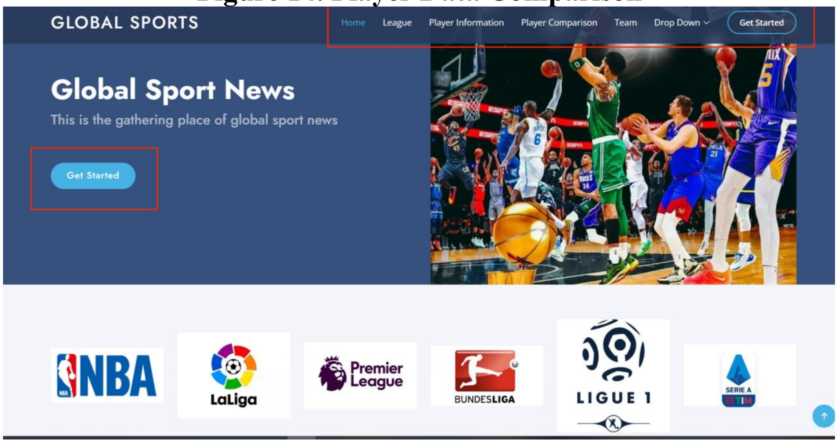


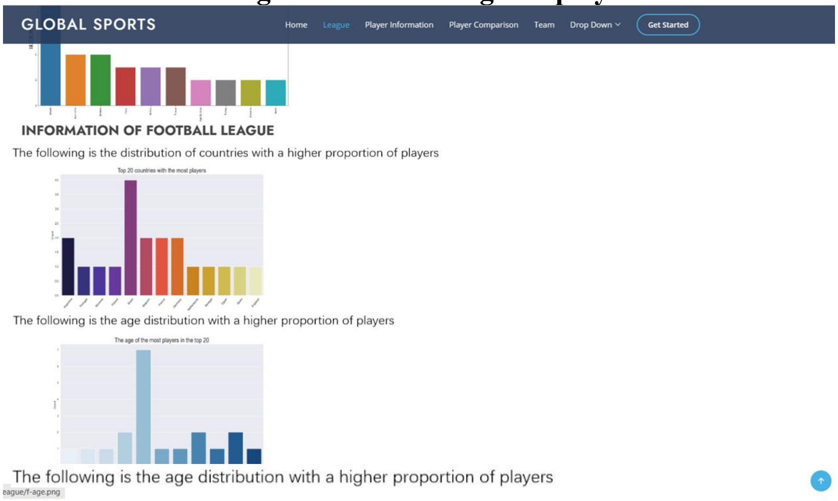
Figure 14. Player Data Comparison



Global Sport News

This is the gathering place of global sport news

Figure 15. Home Page Display



INFORMATION OF FOOTBALL LEAGUE

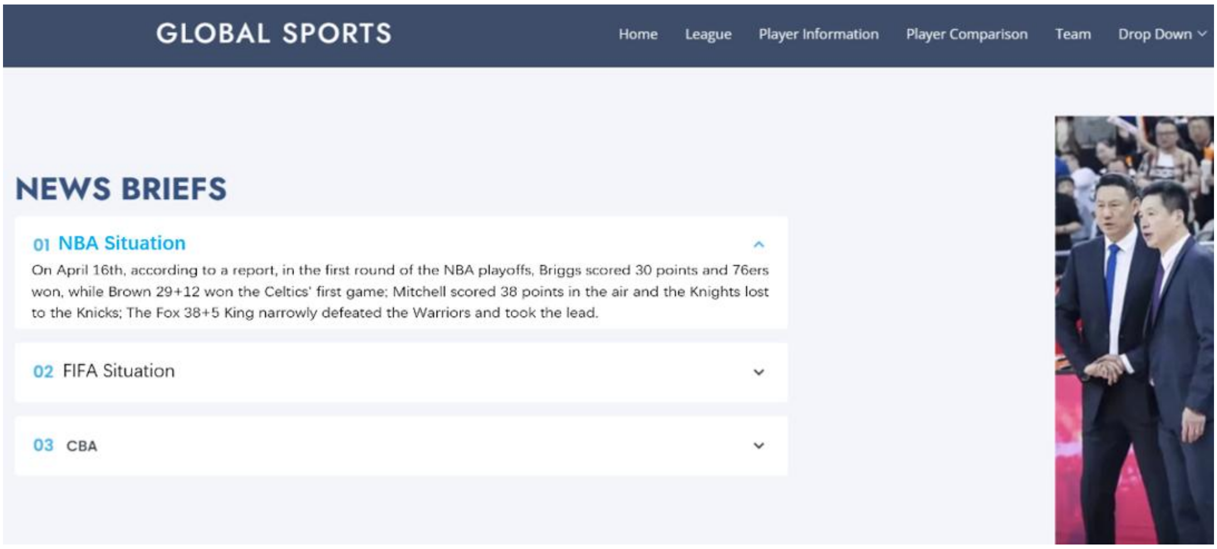
The following is the distribution of countries with a higher proportion of players

The following is the age distribution with a higher proportion of players

The following is the age distribution with a higher proportion of players

Figure 16. League Information Page Display

As shown in Figure 16, the league information page is divided into NBA and Football league. The NBA section shows the nationality distribution of players in the NBA league and the nationality distribution of players in the 2019-2020 season, and the football league section shows the countries with the highest proportion of players and the age distribution of players. The above visualizations can be enlarged by clicking.



NEWS BRIEFS

01 NBA Situation

On April 16th, according to a report, in the first round of the NBA playoffs, Briggs scored 30 points and 76ers won, while Brown 29+12 won the Celtics' first game; Mitchell scored 38 points in the air and the Knights lost to the Knicks; The Fox 38+5 King narrowly defeated the Warriors and took the lead.

02 FIFA Situation

03 CBA

**Figure 17. League Information Page Display**

As shown in Figure 17, the news page is the sports news briefing within 24 hours, which is divided into NBA, football, CBA three parts, click the arrow will pop up the news briefing box, the right side is the day's top content to display pictures.

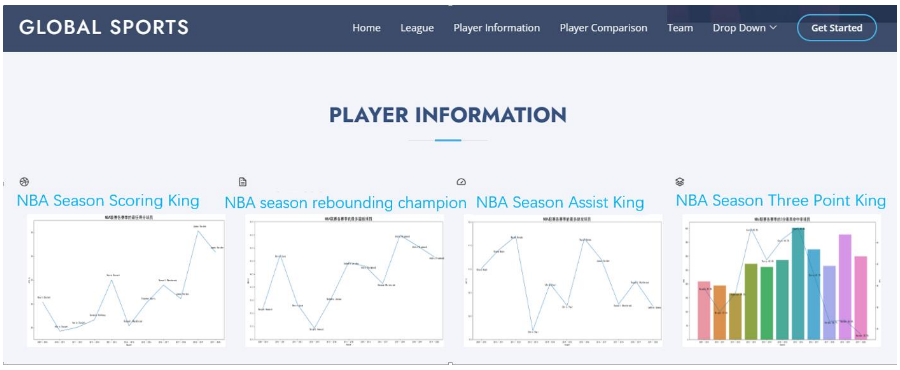


**Figure 18. Player Data Information Page Display**

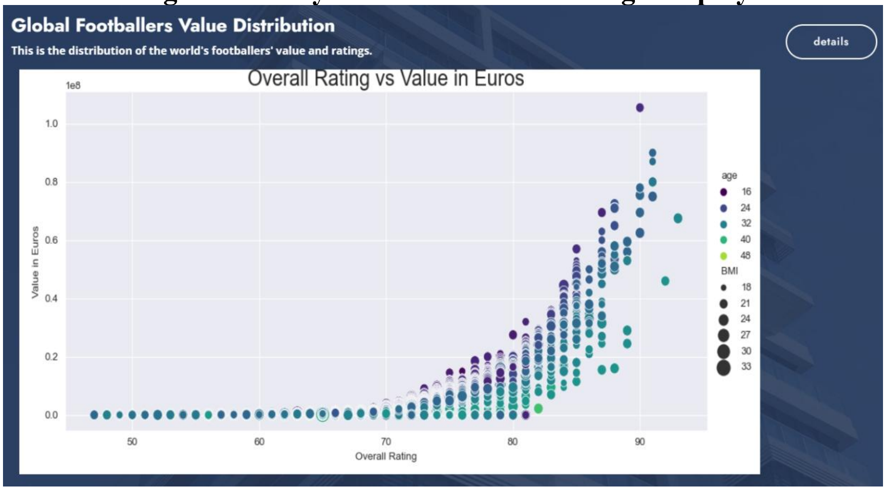
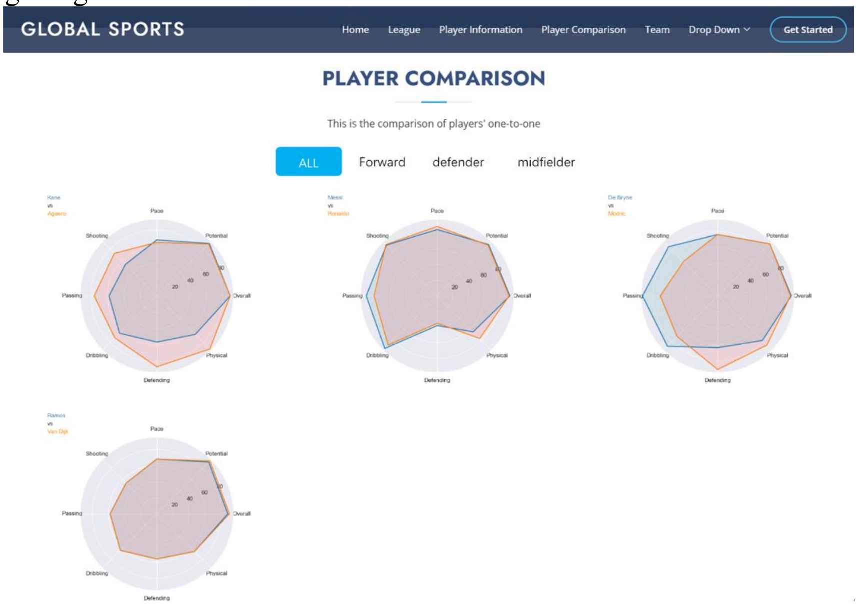


**Figure 19. Player Data Information Page Display**

As shown in Figure 18, the player data information page is used to display the detailed data of NBA players and football league seasons, in which the NBA part shows the data information of the season scoring champion, the season rebounding king, the season assisting king and the season three-point king. Click on the four titles to jump to the detailed data page. The football section shows in Figure 19, the rating distribution and price distribution of global football players. Click the white box in the upper right corner to jump to the detailed information page.

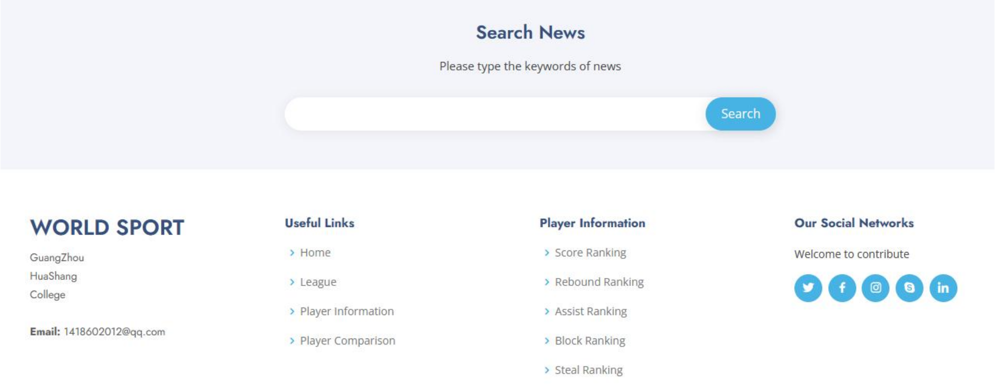


PLAYER COMPARISON

This is the comparison of players' one-to-one

Figure 20. Player Comparison Page Shown

The Player comparison page is used to display the comparison of players in the same position, as shown in Figure 20. The list box is divided into four parts, including all players, forwards, defenders, and midfielders. In addition, each picture is classified according to its corresponding position. Clicking the position option in the list box will pop up the comparison of the corresponding position. Moving the mouse to the picture will pop up the detailed information of the comparison of players and the function of enlarging and jumping the detailed information of the comparison picture, including the position classification and players[15].



Search News

Please type the keywords of news

WORLD SPORT

GuangZhou
HuaShang
College

Email: 1418602012@qq.com

Useful Links

Player Information

Our Social Networks

Welcome to contribute

Figure 21. The Final Page

As Figure 21 shows, the end page includes a news search box, creator information, shortcuts to jump to content, and ways to submit. In the news Search box, enter keywords and click Search to search for related news.

## 6. Conclusion

Taking NBA basketball and FIFA football as cases, this study uses FDCD-TFIDF algorithm to explore the realization method and significance of sports news data visualization in network media. Through the research, the following conclusions are drawn:

Network media sports news plays an indispensable role in modern sports communication, which is of great significance and value.

In this study, NBA basketball and FIFA football are selected as cases, Python programming language and data visualization tools are used for data processing and display, and the presented results are analyzed.

The research results and achievements of this study are the further exploration of data visualization in sports information, which can provide certain reference and reference for the data visualization of sports news in network media.

**Acknowledgments**

This project is sponsored by the “Quality Engineering” Project of Guangzhou Huashang University and the Youth Academic Program of Guangzhou Huashang University under grant HS2023ZLGC08, 2023HSQX040. Project Name: Design and Analysis of Algorithms (mixed online and offline) and Research on Cloud Computing AI Empowering the Development of Audiovisual Industry.

## References

[1] Yearning. Research on the communication advantages and application strategies of news visualization in sports news reports. Journal of Science and Technology, 2018 (5):150-152.

[2] Zhang Wei. The Inspiration of ESPN Brand Culture to the development of sports media in China. Sichuan sports science, 2022, 9 (02): 33 + 30-97 DOI: 10. 13932 / j.carol carroll nki sctykx. 2022.02.07.

[3] Wu Keke. The impact of Big Data Era on Sports News Communication. Cradle of Journalists, 2022 (08):162-164.

[4] Cong Hongyan, Li Hongmeng, Song Xinyi. Combining the era of media sports news data visualization research. Journal of xi 'an sports institute, 2020 ((4): 449-456. The DOI: 10.16063 / j.carol carroll nki issn1001-747 - x. 2020.04.010.

[5] LU Yi. The Dissemination and optimization of Network sports news in the all-media era. News Front, 2017 (24):141-142.

[6] Wu Dan, Wan Xiaohong, Peng Yufeng. The Development Approach and future Prospect of Sports Communication Research in China since the 21st century (2000-2020). And sports science, 2022 (01): 45-58, DOI: 10.16469 / j.carol carroll ss. 202201004.

[7] Hong Lihua, Zhou Weihong, Huang Qionghui. Research on data visualization based on Python. Science and Technology Innovation and Application, 2022, 12 (33):36-40.DOI: 10.19981/j.CN23-1581/G3.2022.33.009.

[8] XIAO Huiming. Research review of Python technology in data visualization. Electronic Testing, 2021 (13):87-89. (in Chinese) DOI: 10. 16520/ j.cnki.1000-8519.2021.13.029.

[9] Shui Qingyan, Guo Ning. Sports news data mining and analysis in the context of Big Data. China Newspaper Industry, 2021 (06):110-111.DOI:10.13854/ j.cnki.cni.2021.06.056.

[10] Li Junfeng. A comparative analysis of Big Data technology and traditional statistical analysis methods. Modern Marketing (Management Edition), 2020 (02):98. ( in Chinese) DOI: 10.19921/j.cnki.1009-2994.2020.02.077.

[11] Wu Yupeng. Application research of Machine Learning in Data preprocessing. Information and Computer (Theoretical Edition), 2022, 34 (13):16-18. ( in Chinese)

[12] Fang Ji, Xie Huimin. Application research of Python in Big Data Mining and analysis. Digital Technology and Applications, 2019, 38 (09):75-76+81. DOI:10.19695/j.cnki.cn12-1369.2020.09.29.

[13] Chen Hantao. Research on the visualization of Network Media Sports News in the era of Big Data. Shanghai University of Sport, 2017.

[14] Zhan Di. Investigation on the narrative types of news Visual production -- Based on the analysis of the visual reports of Sina.com and Xinhuanet.com. Journalism University, 2018(1):9.

[15] Yan Zhilong. Research on the production and dissemination mode of network media news in the era of Big Data. News Research Guide, 2021. 12 (23):98-100