Establishing a Database of Competition between ISPs Based on Python Language

Hanyue He*

International School, Beijing University of Posts and Telecommunications, Beijing 100876, China

*Corresponding author e-mail: 2017213127@bupt.edu.cn

Abstract. Nowadays, we have entered the era of 5G. And the application of the Internet surrounds our daily life. People need to rely on broadband products provided by different Internet service providers (ISP) when they access the Internet. People often compare the speed and the cost of Internet access between different providers. However, monopolistic behavior in the industry is severe in some countries, which may result in users having to withstand higher network access prices and lower quality broadband products. Therefore, the competition between ISPs is a direct influencing factor for users to use the network. This article first highlights the importance of competition among ISPs. And secondly, choose the European authoritative website, crawl the data, and manipulate it at the first step. Finally, this paper summarizes the ISPs competition data sets that can be used for subsequent model analysis. The author will elaborate on the relationship between competition and flow price in subsequent paper.

1. Introduction

Today, people's lives are no longer able to leave the Internet. The speed of network speed and the Internet access price affect people's work and life. Different Internet service providers offer customers different broadband products to choose. However, lack of competition may result in users not having too many choices. Monopoly may result in users having to endure high network access prices and low-quality broadband products. Therefore, the motivation of the project is to study the relationship between competition level and network access. This is not only meaningful to the user, it also helps the ISPs to develop a competitive strategy that suits them.

The main purpose of this paper is to establish a data set that describes the degree of competition between ISPs. The price data comes from a study called European fixed broadband prices. The study provides comprehensive data on fixed broadband retail prices for consumers in the EU, Iceland, Norway, Japan, Korea, Canada and the United States. In the data processing process, select python as a tool to clean data, group data, and so on. In addition, the project also applied web crawler technology in the process of obtaining data. Initially, the project began with a year of data mining. To ensure the accuracy of the target model, finally we used data of many years.

The main work of the article is as follows:
1. Basic introduction to the degree of competition between ISPs.
2. Data crawling of the original web page and storing the data in the database.
3. Further processing and screening of data
2. Background

2.1. Technology Background

**Python for data analytics**

In recent years, data analytics has become increasingly important. How to dig out knowledge from the data has become a concern of people. Python is heavily used for data analysis. The reasons are as follows:

1. Python has a large number of libraries providing a complete set of data analysis tools
2. The Python library is constantly increasing, and the methods of algorithm implementation adopted are also becoming more and more innovative
3. Python can easily port other languages, such as c, Java, etc.

Python is simple to use and has a wide range of data processing libraries, making it an ideal choice for big data processing and scalable applications. Another important aspect of Python is its ease of integration with web applications, all of which support big data processing and quick insight. This kind of fast and dynamic insight (often changing) is valuable to the business. Therefore, they need some powerful languages/platforms/tools to immediately capture this value and remain competitive in the marketplace. Python plays an important role here and supports business requirements [2].

**Pandas**

Pandas provides a number of data structures and functions that enable us to process structured data quickly and easily. Pandas combines Numpy's high-performance array computing capabilities with flexible data processing capabilities for spreadsheets and relational data such as SQL. It provides sophisticated indexing to make it easier to reshape, slice, and diced, aggregate, and select subsets of data. DataFrame is an object of pandas, which is a column-oriented two-dimensional table structure with row and column labels [3].

Because the first dataset we want to process is an Excel spreadsheet, it is very efficient and convenient to use the DataFrame object in the Python Pandas library. The DataFrame is a two-dimensional table structure. In the Design and Implementation section, we can see that the data is sliced, counted, aggregated, etc., using a number of functions in the DataFrame.

2.2. Net-neutrality

The debate about net-neutrality is fierce, and many of these debates are around the Internet access price. Net neutrality requires ISPs to treat all traffic equally on their networks. This means that the wires and towers that are used to access the Internet cannot block or throttle data from certain websites or applications. Internet content providers cannot make special deals with ISPs to make their data faster than others. Some products from Internet content providers like Facebook, and Netflix take up much more bandwidth than others like email, while ISPs can’t charge them extra according to net neutrality. Actually, The United States has announced the abolishment of the principle of net neutrality, which may be detrimental to the development of some small companies. Some large companies can provide higher speed bandwidth to serve customers by paying more. This creates a monopoly.

2.3. Competition among ISPs

The number of access providers that can provide high-speed network access speeds is limited, which means that the competition between these providers in high-speed bandwidth or some certain bandwidth units is insufficient [1].

Because lack of competition, some big ISPs got the worst customer satisfaction ratings. In this case, we can imagine that if those ISPs start blocking, slowing down, or charging more, final customers have to put up with it or go without Internet. That is why we need to establish models with real data, which aim to check whether there are some certain relationships between the degree of competition and Internet access prices/quality.
3. Design and implementation

This project is based on data analytics. We need to ensure which data are related to this project at the beginning. Then we use the model to process the data. But before using the model to process the data, we filter the data at first step (which is more important). This article focuses more on the processing of data.

3.1. Price data

The datasets that were used in this project came from https://ec.europa.eu/digital-single-market/en/news/fixed-broadband-prices-europe. Data was collected from ISP websites between 10th and 24th October 2016. It includes countries, ISPs names, products names, bandwidth units, and access prices information. Through downloading Simulation Tool, we can get the dataset.

For efficient data processing and data visualization, I decided to use Python to process the data. First, I saved the original dataset into the DataFrame, and sliced the DataFrame to obtain the data that were needed. The information needed includes country information, ISP names, products, bandwidth units, and prices. From the original dataset we found a lot of price information, such as LC, EUR, and PPP prices. LC stands for the Internet access price information counted in local currency units. EUR is to unify the Internet access price of different countries into euro. At first, I selected the EUR prices as the data I needed. However, there is a problem. The wage income and living standard of each country are different, so it is not objective to directly compare their Internet access price. Later I discovered that PPP prices represented Purchasing Power Parity. Purchasing power parity, also known as purchasing power level. The price of Internet access, measured at purchasing-power parity, has fallen to take into account the different standards of living in each country.

In addition, since the original data is bandwidth units, we need to calculate the average bandwidth of this bandwidth unit, for example, 25M for the 10-30M bandwidth units.

To compare the price of Internet access in different countries, we also need to calculate the price of Internet access per bytes. Divide the price by the average bandwidth to get the price per bytes. Take the calculated data as a new column in the DataFrame. In this way, we can obtain the processed data.

3.2. The degree of competition

In this section, this paper specifies the processing of crawler code. The number of ISPs in different countries can be used to show the different degrees of competition. For example, Australia has 5 ISPs, so we marked Australia's degree of competition as 5, Belgium has 3 ISPs, and then marked as 3.

When counting the number of ISPs in each country, remove duplicate names from the ISP column and group them by country.

3.3. Processed dataset

Table 1. Static values of processed dataset

| Country | ISP | Products | Price     | Average Bandwidth | Price Per Bytes | P2P       | P2C       |
|---------|-----|----------|-----------|-------------------|----------------|-----------|-----------|
| 38      | 153 | 1331     | 51.443332€| 56.679682Mbps     | 1.971904€/Mbps | 11.673279 | 10.951593 |

Table 1 shows some summary statistics of processed dataset. The column names are the data required for the project. It includes Country, ISP, Products, Bandwidth Units, Price, Average Bandwidth, Price Per Bytes, the number of Peering interconnections of each AS, and the number of Transit interconnections of each AS. Where Average Bandwidth, Price Per Bytes were computed. The number of peering and transit interconnections came from CAIDA. This project analysed Internet access prices and bandwidth of 153 ISPs in 38 countries. On average, each AS has 11.673279 p2p interconnections, 10.95159 p2c interconnections. AS 6939 (USA) has the most p2p interconnections (3156). AS174 (USA) has the most p2c interconnections (4719).
Among the 38 countries, Liechtenstein has the fewest number of ISPs (only 1 ISP). Denmark has the largest number of ISPs (6 ISPs). Most countries have 4-5 ISPs. There is a total of 153 ISPs in 38 countries, with an average of about 4 ISPs per country.

These data have been filtered at the first step after crawling, and the relationship between the degree of competition and the network traffic topology will be discussed in subsequent papers.

4. Conclusion
This paper first introduces the technical background needed for data crawling and analysis, including proficiency in the Python language, the ability to use the pandas package and subsequent processing of data (including data deduplication, filtering, categorization, and storage). Next, it introduces Net-neutrality and the concept of the competition among ISPs, which is very helpful to the overall understanding of this paper. And the importance of this study is also emphasized. After the first step of the data processing, the paper summarizes the data and obtains the degree of competition between Static values of processed dataset and ISPs in different countries.

For subsequent studies, the dataset in this paper will be further researched. We will use the model to explore the impact of the degree of competition between ISPs on Internet access prices and the impact of network topology on the degree of competition.

Reference
[1] Brodkin, J. (2019). FCC report finds almost no broadband competition at 100Mbps speeds. [online] Ars Technica. Available at: https://arstechnica.com/information-technology/2018/02/fcc-report-finds-almost-no-broadband-competition-at-100mbps-speeds/

[2] Chinacpda.com. (2019). Why is Python so important for big data analytics? | Data Analyst CPDA. [online] Available at: http://www.chinacpda.com/jishu/16419.html.

[3] Cnblogs.com. (2019). Python data analysis entry study notes - Little Star ☆ - Blog Park. [online] Available at: https://www.cnblogs.com/zzhzhao/p/5269217.html.