Analysis of Stock Browsing Patterns on Yahoo Finance site

Chenglin Chen

chenglin@cs.umd.edu

Introduction

Yahoo finance [1] is the largest business news Web site and one of the best free Stock Chart Websites in the United States. It provides a charting service which is clear, easy to use, and very basic. According to comScore [2], there are more than 37.5 million monthly unique visitors to this website so it would be interesting to gain some insights for the stock browsing patterns of Yahoo finance website users.

Dataset

When users come to Yahoo Finance and search for a stock quote, they input the ticker or the name of the stock and click “Get Quotes” to get the quote along with the stock price chart and other data. Yahoo finance also suggests other stocks people view while viewing this particular stock. For example, if users get a quote for AAPL, they will see this message: People viewing AAPL also viewed: APPL PCLN GOOG AMZN MA CMG.

I think using this suggesting view feature is a good way to build a stock browsing network and classify some Web Usage Patterns. If I start with some stocks and get the suggested stocks for them, I will get the first level network. Then with the new list of stocks I can get to the next level, so on and so forth. It will be like the 1.0, 2.0 or 3.0 network of the original stocks.

This network will be a directed graph.

Analysis

To get the size of the graph under control, I start with the 30 stocks of the Companies in the Dow Jones Industrial Average[3] and get its 1.0 network as the first try.

The resulted Figure 1 shows the Dow30_1.0 network. It has 73 vertices and 182 edges. The colors of the vertices represent their groups, the sizes of the vertices are their in-degrees. The graph algorithm is Harel-Koren Fast Multiscale.

Looking at this small network, first thing that we notice is that there are two separated graphs. The green one on the top left consists of Bank of America (BAC), CitiBank (C), JPMorgan Chase (JPM) and Goldman Sachs (GS), etc. These are the stocks in the financial sector. The other part
of the graph, though connected, is clearly divided into groups like Technology, Basic Materials and consumer Goods.

In the orange group on the bottom right, two vertices (VZ and T) have very similar connections with other vertices in the same group. These two are Verizon and AT&T. No wonder!

So there are some interesting viewing patterns in the Dow30_1.0 network. How about bigger networks?

Figure 1: The Dow30_1.0 network. The colors of the vertices represent their groups. The sizes are their in-degrees. The graph algorithm is Harel-Koren Fast Multiscale.

To get a bigger network, a 3.0 network of the original Dow30 stocks is created. This directed graph has 194 vertices and 905 edges. Carefully study of this network using NodeXL gives the following three insights.
**Insight 1: Yahoo Finance users tend to browse stocks by sector/industry with some exceptions (caused by typo, maybe)**

Figure 2 is the Dow30_3.0 network. It shows that all vertices are grouped in a way that’s similar to the stock sector groups, which means that Yahoo Finance users tend to browse stocks by sectors.

For example, the biggest group is the dark blue group in the top left. This group consists of Yahoo (YHOO), Microsoft (MSFT), DELL (DELL), EBAY (EBAY), Oracle (OCLR), etc… It’s the Technology sector. And the orange group on the bottom right is the Financial sector.

![Dow30_3.0 network](http://nodeXL.codeplex.com)

Figure 2: The Dow30_3.0 network with each of its groups in their boxes. The colors of the vertices represent their groups. The sizes are their in-degrees. The graph algorithm is Harel-Koren Fast Multiscale.

But there are some exceptions:
(1) IBM (IBM) belongs to the technology sector and is grouped with McDonald (MCD), NIKE (NKE) and other Consumer Goods companies.

(2) Apple (AAPL) is in IT, but it points to APPELL PETE CORP (APPL) which is totally out of place. The only logical explanation is that this is caused by typo. What people really interested in is Apple (AAPL), they should either put in the complete name “Apple” or the ticker “AAPL” to get the quote. But I guess the combination of a little carelessness and the auto complete feature on the website causes the mistake. This mistake happens quite frequently, really, since this edge shows in our most frequently viewed network.

Insight 2: If Yahoo Finance users want to browse stocks in the financial sector, they tend to start from there and stay there.

Figure 3: The Dow30_3.0 network with a NodeXL radial layout. The colors represent different groups. From top, in clockwise, the groups are green(REIT), red(Energy), orange(Financial), yellow(communication), dark blue(Technology), blue(Consumer Goods), dark green(Basic Materials) respectively.

Figure 2 shows how the vertices are divided into groups. To see how well the different groups link together, a NodeXL radial layout of this the same network is showed in figure3. As in figure2,
the colors represent groups. The labels of the vertices are not showed to allow a better view of the graph’s structure.

We can see that the connections within groups are strong and the connections among groups are not that strong. For example, there are very few edges coming in to or out from the orange financial sector group. This means that if Yahoo Finance users want to browse stocks in the financial sector, they tend to start from there and stay there.

**Insight 3: Freeport-McMoRan Copper & Gold Inc (FCX) is likely the most viewed stock.**

To find outliers, we can also calculate and visualize vertex metrics to find important individuals. Figure 4 shows the Dow30_3.0 network mapping In-Degree to the X axis and Betweenness Centrality to the Y axis. Edges are hidden.

![Figure 4: The Dow30_3.0 network mapping In-Degree to the X axis and Betweenness Centrality to the Y axis. Edges are hidden.](image)
Looking at figure4, we can easily indentify that Freeport-McMoRan Copper & Gold Inc (FCX) is likely the most viewed stock since it has the highest in-degree.

There are other outliers:

(1) Johnson & Johnson (JNJ) has the second highest in-degree but not a very high Betweenness Centrality, so it has lots of views but it’s not likely the only connector to its connected vertices.

(2) Intel (INTC) and Pfizer (PFE) are also special. Neither of these two has very high in-degree, but they both have high Betweenness Centrality. This means that they probably are some important bridges to other vertices groups.

**NodeXL Critique**

NodeXL is an excellent tool especially designed for social network data analysis with visualization as a key component.

Good features of NodeXL:

(1) It is free and open source.
(2) It provides a wide range of basic network analysis and visualization features such as Dynamic Filtering, Powerful Vertex Grouping and Graph Metric Calculations.
(3) It has direct connections to Social Networks (Twitter and Facebook), and it can import and export graphs in GraphML, Pajek, UCINet, and matrix formats.

Things to be improved:

(1) NodeXL gets really slow and crashes when it deals with large dataset (30,000 vertices).
(2) It would be nice if NodeXL also runs on Mac.
(3) Sometimes the auto snapping of the graph window gets in the way when users try to better utilize the limited screen display.
(4) Lack of easy reversal of action. An “undo” button would be very useful when the user is trying out different settings of the analysis and visualizations.

**References**

