Ye, Minjian; Li, Guangzhong

Article

Internet big data and capital markets: a literature review

Financial Innovation

Provided in Cooperation with:
Springer Nature

Suggested Citation: Ye, Minjian; Li, Guangzhong (2017) : Internet big data and capital markets: a literature review, Financial Innovation, ISSN 2199-4730, Springer, Heidelberg, Vol. 3, Iss. 6, pp. 1-18, https://doi.org/10.1186/s40854-017-0056-y

This Version is available at:
https://hdl.handle.net/10419/176448

Ye and Li *Financial Innovation* (2017) 3:6
DOI 10.1186/s40854-017-0056-y

Financial Innovation

REVIEW Open Access

Internet big data and capital markets: a literature review

Minjian Ye and Guangzhong Li*

* Correspondence:
liguangzhong@mail.sysu.edu.cn
Sun Yat-Sen Business School, Sun Yat-Sen University, Guangzhou 510275, China

Abstract

Background: Research in various academic disciplines has undergone tremendous changes in the era of big data. Everyone is talking about big data nowadays, but how exactly is it being applied in research on financial studies?

Results: This study summarizes the sources of Internet big data for research related to capital markets and the analytical methods that have been used in the literature. In addition, it presents a review of the research findings based on Internet big data in the field of capital markets and proposes suggestions for future studies in which big data can be applied to examine issues related to capital markets.

Conclusion: (1) Internet big data sources related to present capital market research can be categorized into forum-type data, microblog-type data and search class data. (2) As for research about investors' sentiments on the basis of Internet big data, the main methods of sentiment analysis include building an inventory of lexical categories, using dictionaries for analysis of lexical categories, and machine learning. (3) Many studies address whether Internet big data can predict capital markets. However, they reach no consistent conclusions, which could be due to limitations of sample and analysis method used. (4) Data collection technique and analysis methods require further improvements.

Keywords: Internet big data, Financial studies, Capital markets

Introduction

This is the era of information explosion and a world overwhelmed by numbers and digits. Research by the International Data Corporation indicates that the global data volume is expected to reach 35 zettabytes (ZB)[1] by 2020. Beyond that, the trend of growth doubling every 2 years will be maintained. This implies that we have entered the era of big data. Simultaneously, the term "big data" has been mentioned repeatedly in both commercial applications and academic research. The World Economic Forum[2] has claimed that big data is a new type of asset class. In *Forbes* magazine, Rotella has labeled big data "the new oil."[3] In addition, research on big data within academia is increasing rapidly; a summary of the number of papers published in 2010–2015 with "big data" as the subject and listed in the CNKI and Web of Science databases is illustrated in Fig. 1.

The growing emphasis that academia has placed on research on big data can be discerned intuitively. In fact, academic research on the topic has surged in the last 3 years. Furthermore, other events show that big data plays an important role in the

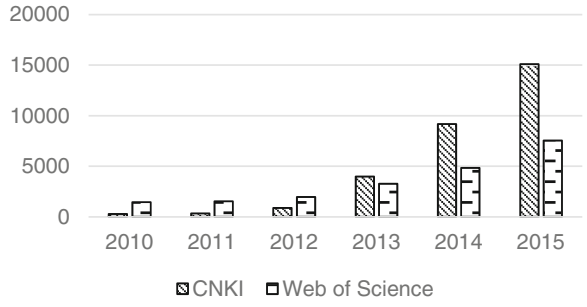

Fig. 1 Number of papers published in 2010–2015 with "big data" as the subject

developmental process of modern societies. These include the launch of the Big Data Research and Development Initiative[4] by the Obama Administration in 2012 and the special report on big data[5] published by the United Nations Global Pulse.

Big data has various unique characteristics, including volume, velocity, variety, and veracity. The application of big data in the financial field has developed rapidly, and such applications are highly valued by various fields in society. The most direct application is the development of various fund products based on big data. The earliest index product based on big data to be launched in China was Dingtoubao by Galaxy Asset Management Co. This product mainly invests in the Teng An 100 Index, a stock market index jointly created by Tencent Holdings Limited and Ji An Financial Information and managed by China Securities Index Co., Ltd. (CSI). This was soon followed by three fund management companies: GF, China Southern, and Bosera. These companies successively developed big data indexes jointly with owners of Internet big data resources, including Baidu, Sina, and Alibaba. This has led to the launch of separate funds, such as the GF's Baifa 100 Index A, China Southern's Big Data 100 Index and Big Data 300 Index A, and Bosera's CSI Taojin Big Data 100 Index A.[6]

As of October 2016, there were 19 funds in the Chinese market named after big data, with a combined scale of funds amounting to 15 billion RMB.[7] One of the funds with outstanding performance is the Dacheng CSI 360 Internet + Big Data 100 Index Fund, which ranked second among stock-oriented funds that were newly launched in 2016. From the second quarter, it ranked top among all the big data funds and was in the top 10 in the Chinese market for stock-oriented funds. Furthermore, it achieved the highest cumulative gain of 25.6%, which not only delighted its fund investors but also surprised those in the big data fund industry.[8]

As big data has been incorporated successfully into business, how should big data be applied in academic research related to capital markets? To answer this question, we must first clearly ascertain where big data is located and who owns it. At present, it seems that big data is controlled mainly by several Internet companies and public service sectors. Some organizations or departments that are relatively rich in data resources are illustrated in Fig. 2. However, considering data availability, most research on big data in the financial field has focused on big data from social networking platforms and search engines. Hence, the review subjects for this study are studies on big data from forums, microblogs, and search engines related to the capital markets. The review was undertaken from the perspective of data acquisition.

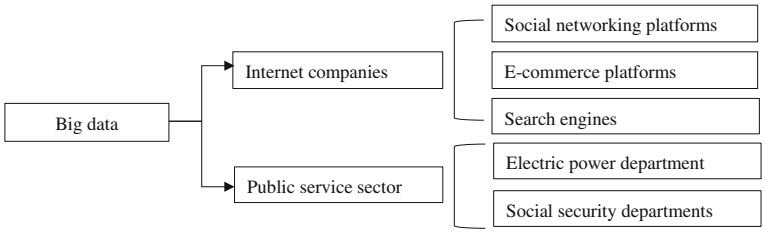

Fig. 2 Distribution of big data

The main review topics are as follows: (i) big data indexes used in research on capital markets and analysis methods and (ii) research findings on big data indexes related to the field of capital markets. These two aspects are discussed in "Internet big data related to capital markets" and "Application of Internet big data to capital markets" sections, respectively. "Conclusions and future research" section provides recommendations on future research directions.

Review

Internet big data related to capital markets

Observing investors' decision-making processes was difficult prior to the advent of the Internet era. Furthermore, it was difficult to directly observe information that investors are concerned about and their views concerning the market. With the ubiquity of information technology and the Internet, an increasing number of investors are gathering information from the Internet for analysis. Through a click of the mouse and inputs via the keyboard, such investment decision behavior has resulted in the accumulation of a huge amount of unstructured information in the virtual network. For example, Google processed 2 million searches per minute in 2012, whereas Twitter users posted approximately 100,000 tweets per minute (James 2012). Over time, these types of data gradually have become the basic sources of big data for research related to capital markets.

Specifically, most research on Internet big data related to capital markets can be categorized into forum-type, microblog-type, and search-type data. Given the significant differences in the structures of the big data, each type is explained and analyzed separately in the following subsections.

Forum-type data

Internet forums provide users with a platform for communication and interaction that transcends time and geographical region. Within the same message board of a forum, investors are able to interact with other investors worldwide. Forums on stocks and shares are generally categorized into separate message boards for different stocks; therefore, investors can have discussions and exchanges on those message boards corresponding to the stocks of their interest. On forum sites, the general approach is to automatically place that the newest posts and posts on a message board that are being discussed most intensely on the home pages. In addition, message board managers evaluate the content quality of posts and bump the good ones to the top or pin these as recommended posts, thereby making them easily found within the sub-board that contains all the pinned or sticky posts. This means that investors can quickly obtain the latest, relatively important, and high-quality information on the stocks that they are

interested in by checking the relevant message boards. However, pinned posts are screened and selected by message board managers, indicating that a certain degree of subjectivity might be involved. Moreover, since individuals are unable to customize the various message boards to focus on information related to a particular topic, it might be inefficient to obtain information from stock message boards given investors' limited attention. In addition, the timeliness of interactions on the boards is poor.

The main Chinese stock message boards include Eastmoney.com, Homeway.com, Iguba.com, and Taoguba.com. For existing studies, the stock message boards of Eastmoney.com are the main data sources selected to represent Chinese forums. Examples of such studies include those by Ackert et al. (2016), Dong and Xiao (2011), Huang et al. (2016), Nan (2015), Shen et al. (2013) and Zheng et al. (2015). Several probable reasons exist for this situation, as follows. (i) Eastmoney.com has always been ranked the most popular[9] among Chinese stock message boards, its users are active, and the volume of its cumulative data is huge. (ii) The time period over which its public data has been saved is longer. For example, its earliest saved data on Gree Electric Appliances Inc. dates back to August 29, 2007, whereas data from Homeway.com are available only for the 3 months. (iii) Information on Eastmoney.com is richer and more detailed. For example, studies by Dong and Xiao (2011) and Huang et al. (2016) required identifying of the geographical region of users who posted messages. Only Eastmoney.com can provide relatively sufficient data to meet such a requirement.

In terms of the basic data indicators, the main types selected for existing studies generally can be classified as quantitative and content class indicators. The former includes the number of posts, clicks, shares, and investors participating in discussions (Ackert et al. 2016; Dong and Xiao 2011; Huang et al. 2016; Jiang et al. 2015; Shen et al. 2013; Zheng et al. 2015). Simultaneously, most literature has evaluated the information contents of forum posts using textual analysis to construct the content class indicators. The main types include indexes based on numbers of opposers and supporters and the degree of divergence in opinions (Antweiler and Frank 2004; Das et al. 2005; Nan 2015; Sabherwal et al. 2011; Zhang and Swanson 2010). In a study on information exchange over the Internet and herding behavior in the stock market, Zheng et al. (2015) selected the natural logarithm for the total number of daily posts as the proxy variable for investors' communication.

Presently, there are two main text classification methods for constructing indicators of market sentiments: manual classification and machine learning. Studies based on Chinese forums mostly have relied on manual classification with supplementation from simple programming classification (Ackert et al. 2016; Nan 2015; Shen et al. 2013). In his study on the impact of online public opinions on private placement of listed firms, Shen et al. (2013) used web crawler technology to collect all posts that contain the word "additional issue." He then randomly selected 3,000 posts from the overall samples for manual classification. In addition, 26 words that appeared in those selected posts were deemed to reflect the tendency to oppose. These words included "opposition," "failure," and "misappropriation of funds." Finally, a program was used to identify whether posts contained any opposing content so that all posts could be categorized as either the opposing or supporting type. The proportions of each type of posts relative to the total number of posts were used to construct the indexes for opposition and support, which were used to gauge the direction of online public opinions.

Based on Shen et al. (2013), Nan (2015) calculated the squared difference between the indexes for opposition and support. This led to the construction of a variable for the degree of divergence in opinions. Similarly, Ackert et al. (2016) manually set up keywords reflecting bullish or bearish market sentiments and then used a program to identify the market sentiments reflected in the posts, thereby building an indicator for trends in market sentiments.

For research on English forums, machine learning or tagging of posts is used to determine textual sentiment. Antweiler and Frank (2004) collected text information from posts made on message boards in Yahoo! Finance and Raging Bull forums that corresponded to 45 stocks. The Naive Bayes algorithm was applied to classify textual sentiment, which was used as the basis for constructing the indicator for the degree of divergence. Separately, Sabherwal et al. (2011) selected data from posts on forums in TheLion.com. Similarly, the Naive Bayes classifier method was applied to build indicators for market sentiments and degree of divergence. Leung and Ton (2015) captured text information from posts made on the HotCopper forum for more than 2,000 stocks and then applied the Naive Bayes algorithm to classify market sentiments and build an indicator for the degree of divergence.

Besides the Naive Bayes method, some scholars have used the maximum entropy method for text classification. Zhang and Swanson (2010) collected data from posts on message boards in Yahoo! Finance that corresponded to 30 U.S. listed companies. The maximum entropy model was used for text classification, leading to the construction of an indicator for market sentiments. In addition, some scholars have realized that the reliability ranking of the person who made the post affects the level of trust of other investors regarding its contents. This causes posts to have different weightings within the overall indicator for market sentiments. Hence, it might be more appropriate to use the reliability-weighted average method instead of a simple weighted average method to build the indicator for overall trends in market sentiments.

Forums on TheLion.com have an exclusive feature that indicates the reliability level of its users. Thus, Sabherwal et al. (2011) used the related data to create a reliability-weighted overall indicator for market sentiments. Gu et al. (2006) calculated the reliability scores for forum users based on the forecast accuracy of their previous posts, creating a reliability-weighted overall indicator for market sentiments. After 2004, some of the data sources (e.g., Yahoo! Finance) began allowing users to tag their personal sentiments about the market when making posts. This enables scholars to directly use the self-disclosed sentiments and tag these for text classification (Kim and Kim 2014).

In addition, some stock message boards have unique data that provide the appropriate context for studies in specific fields. For example, among Chinese stock message boards, only Eastmoney.com provides data on the IP addresses of users who have posted a message. This unique feature facilitates the study on home bias in investors communication. Dong and Xiao (2011) employed this feature; through users' respective IP addresses, the authors identified the geographical regions of participants involved in discussions on stock message boards. Looking at the message board of a particular stock, the authors compared the number of investors from a particular province engaged in discussions against the total number of investors having discussions on that board. This enabled the authors to measure the local level of exchange for that stock in the province. Extending the same approach to other provinces, the authors established

the differences in levels of exchange between two provinces for a particular stock. Similarly, Huang et al. (2016) utilized the IP data of Eastmoney.com and constructed a quantitative indicator of local bias in investor attention. The indicator directly measures the magnitude of home bias in information exchange.

On the HotCopper forum, if investors publish posts that do not comply with stipulated regulations, board managers can make the necessary revisions and state the reason(s) for doing so beneath the affected posts. Tapping into this feature, Delort et al. (2011) examined the reasons for revisions and identified posts that contained keywords relating to the manipulation of stock prices. The event studies method was then used to analyze those posts in order to examine the situation of stock price manipulation known as "pump and dump".

Microblog-type data

This type of data completely differs from forum-type data both in terms of data structure and means of information exchange. In terms of data structure, microblogging on a specific stock does not occur on message boards dedicated to that stock. Hence, the method used for collecting samples definitely impacts the conclusion. On the other hand, the efficiency of information exchange via microblogging is much faster than that via forum-type information. The transmission mode of microblog-type data is that of external chain reactions, which is much more efficient than the internal accumulation mode for forum-type data. Considering that investors have limited attention, microblogs help increase investors' efficiency in accepting information. This is because the follow classification method and push mechanism of microblogging can filter irrelevant information more effectively.

Presently, the main microblogging platforms in China are Sina and Tencent, whereas Twitter is the main microblogging platform outside China. For research related to capital markets, most scholars typically have used the Sina microblog as the source for Chinese microblog-type data. For the English equivalent, Twitter is the main data source.

The basic data used for empirical study are quantity-type indicators, including the number of microblogs published, shares, and comments made (Cheng and Lin 2013; Mao et al. 2012; Ruiz et al. 2012; Sul et al. 2016; Xu and Chen 2016). The main content-type indicators include the frequency with which specific keywords appear or sentiment indexes (Bartov et al. 2015; Bollen et al. 2011a, 2011b; Cheng and Lin 2013; Sprenger et al. 2014a; Xu and Chen 2016). In addition, some scholars have used the large amount of microblogging data on user-follower relationships to construct social network-type indicators (Ruiz et al. 2012; Yang et al. 2015).

For research based on Sina microblog's big data, Xu and Chen (2016) used web search technology to collect all content published by the official microblogs of listed companies. Next, 104 keywords were used to classify the microblogs into disclosure and nondisclosure types. The keywords were then grouped and used to categorize information disclosed by the official microblogs of listed companies into one of the following four categories: those related to the company's business; finance; research and development; and reputation. The number of disclosures made by a listed company via its microblog in a day is treated as the company's level of disclosure, and the frequency with which the disclosure type of keywords appears in the company's microblog

was used to measure the disclosed information contents. Considering that microblogs contain relatively substantial amounts of irrelevant information, Xu and Chen (2016) measured the level of noise in microblogs by subtracting the disclosure type of microblogs from the total number of microblogs published on the same day.

Separately, Cheng and Lin (2013) selected the microblogs of five accredited financial media organizations. Data on the number of microblogs published and commentaries made via these microblogs were used as the sources of sample data. Using sentiment analysis techniques, Cheng and Lin (2013) quantitatively calculated the amount of change in market sentiments at various levels, including the word, sentence, and overall post. The findings were then synthesized to become an indicator for investors' daily market sentiments. In addition, other scholars have used the event studies method to analyze changes in the stock market performance of listed companies prior and subsequent to the launch of their official Sina microblogs (Jin et al. 2016).

With the exception of Mao et al. (2012), who examined the correlation between the number of posts and stock market from the levels of individual stocks, industries, and indexes, most other scholars who have used Twitter's big data for research have focused only on indicators of market sentiments. Sprenger et al. (2014b) collected Twitter data related to listed companies, used the Naive Bayes algorithm for text classification of market sentiments, and constructed a daily overall indicator to reflect those sentiments. Sul et al. (2016) collated information in tweets related to listed companies, examined the text contents, and determined market sentiments by analyzing the lexical categories (parts of speech) using the Harvard-IV dictionary. The authors similarly established a daily overall indicator of market sentiments.

For analytical purposes, Bartov et al. (2015) constructed four tendency indicators of market sentiments using the enhanced Naive Bayes classifier, Harvard-IV dictionary, and inventory of lexical categories by Loughran and McDonald (2011). After gathering information from Twitter on the numbers of posts, fans, and shares, Zhang et al. (2011a) extracted vocabulary related to the words "hope" and "fear." He then used the daily frequency of occurrence of such vocabulary as the indicator of market sentiments.

Some scholars have conducted in-depth mining of relationship class data found in microblogging data to undertake social network research. Ruiz et al. (2012) collected two main categories of data. The first category pertained to the numbers of posts and shares; the second category related to statistical variables of social networks, including the connected component of social networks and centrality of distribution. Yang et al. (2015) collected data on the personal information of Twitter users, in addition to relationship class data on users and their followers, which facilitated the development of a financial social networking structure within Twitter.

Search class data

Currently, Baidu and Google are the mainstream Chinese and English search engines, respectively. In China, Baidu has the most extensive market coverage. The majority of investors obtain information related to the stock market via this search engine (Zhang et al. 2014b). At the global level, Google's market share far exceeds that of other products in the same category. Most studies on the U.S. market have used the search data

in Google Trends (www.google.com/trends/) as their study sample. Search class data directly reflects users' behavioral patterns for accessing information. In addition, keywords can capture the contents of users' information needs. The search behavior of an individual directly expresses the user's information needs, while the aggregate search behavior reflects the level of attention paid to a particular event, or the demand for certain types of information.

The basic search class data are mainly those utilizing the number of searches to measure the popularity of a topic and those that use search results to measure information contents. Although Baidu does not publish raw data on the number of searches, it has built separate indicators on the levels of media and users' attention on a topic by applying big data analytical techniques to the raw data. Most studies have adopted these two indicators as proxies for attention levels. An example is Liu et al. (2014), who considered the absolute value of data on users' and the media's levels of attention as proxy variables for attention levels. However, Zhang et al. (2014b) indicated that the absolute values of attention indicators often ignore the impact of company size on the level of attention received. The authors proposed that instead of the magnitude of the absolute value, the relative level of its changes is more worthy of concern. Based on this principle, Zhang et al. (2014b) used the relative value of the Baidu indicator to measure search intensity. Based on that, the authors further constructed an indicator of positive rate of change in online search intensity, which acted as the proxy for changing trends in the search behavior of investors in the market.

In addition to data on search intensity, Baidu's data on search results are also valuable. Most studies tend to focus on a particular network platform for data collection, and Baidu's search results can reflect practically all of the content information found on the Chinese Internet. From this aspect and based on a mining algorithm for text semantics, Zhang et al. (2011b) used Baidu search keywords to "search for X number of relevant pages." With the search results, the authors used various techniques such as keyword selection, application of the "advanced search" function in the search engine, processing of date formats, and data cleaning to construct an open source information indicator. This indicator allows for continuous day-by-day observation of variables and facilitates the examination of the impact of open source information on asset pricing.

Google publishes two types of data that can be applied for research: (i) raw search data and (ii) search indicators that have been standardized. These allow for more diversified types of indicators to be constructed during research. Joseph et al. (2011) collected search data on the Standard & Poor's (S&P) 500 Index's component stocks. The authors directly defined the standardized search indicators as indicators of investor sentiments. After collecting the search data on the Russell 3000 Index's component stocks, Da et al. (2011a) defined the raw data on search numbers as the search volume index (SVI). After taking the logarithm of that index and performing median adjustment, the authors proposed an adjusted SVI, which they used to measure the level of investor attention. Takeda and Wakao (2014) collated Google search data on the Japanese market and built three indicators of search intensity.

Some scholars have been concerned that the weekend effect and seasonal variations might affect search volumes (Da et al. 2015; Drake et al. 2012). As such, these scholars

often eliminated time trends from the raw data. Drake et al. (2012) argued that online search behavior reflects investor demand for public information. Thus, after collecting daily search data for the S&P 500 Index's component stocks, the authors performed regression of the SVI against a time dummy variable. The residual value was then used to construct an abnormal SVI for measuring changes in demand for information. Da et al. (2015) selected 118 words from the Harvard IV-4 and Lasswell value dictionaries as keywords for searches. Furthermore, the authors used the Google search engine to obtain the raw search data. The latter was subjected to both first-order difference and seasonal trend adjustments to construct the Financial and Economic Attitudes Revealed by Search (FEARS) Index for measuring market sentiments.

Summary of usage of Internet big data

The discussion in "Forum-type data" to "Search class data" subsections indicates that the main research direction for big data is the impact of investor sentiments on the financial sector (Antweiler and Frank 2004; Bartov et al. 2015; Sprenger et al. 2014a; Sabherwal et al. 2011). At present, the main methods for judging sentiments include building an inventory of lexical categories, using dictionaries for analysis of lexical categories, and machine learning.

The first method mostly relies on the collection of text needed for research. A large number of sentiment-type keywords are selected from the sample text to build an inventory of lexical categories, which are used as the judgment criterion for classification of all the text (Nan 2015; Shen et al. 2013; Xu and Chen 2016; Zhang et al. 2011a). The second method is based on the lexical classifications in dictionaries. The relevant keywords are extracted and used for the classification of collated text (Bartov et al. 2015; Da et al. 2015; Sul et al. 2016). In the field of financial research, the main distinction between these two methods lies in the sources of keywords: the source of the first method is text information from the samples, whereas the sources for the second method are actual dictionaries or inventories of lexical categories developed on the basis of dictionaries. Although the first method seems more subjective, this method, which is based on manual screening of keywords from the samples, might be more appropriate in the Chinese context. This is due to vast linguistic variations such that the lexical category of a Chinese word might not be completely consistent across different contexts.

The main dictionary selected for the second method is the Harvard IV-4 (Da et al. 2011a). However, many vocabularies that are not considered negative in the financial context are grouped into the negative lexical category in the Harvard IV-4 (Loughran and McDonald 2011). To address this issue, Loughran and McDonald (2011) developed an inventory of lexical categories that is more appropriate for financial research. However, Sul et al. (2016) highlighted that the inventory by Loughran and McDonald (2011) is more suitable for the analysis of formal financial documents (e.g., 10 K filings). Since microblog class text employs colloquial and popular terminologies, the Harvard-IV dictionary is still deemed more appropriate for their analysis.

The third method-machine learning-uses computer algorithms for the classification of textual sentiment. The main steps of machine learning are as follows: (i) select a specific text as the corpus training set and manually classify the words contained within, (ii) use a computer algorithm, such as the Naive Bayes algorithm, to train the text in

the training set and establish the judgment rules for text classification, and (iii) apply the judgment rules to all text classifications.

The Naive Bayes algorithm is currently one of the most popular training algorithms for classifying textual sentiment (Antweiler and Frank 2004; Leung and Ton 2015; Sabherwal et al. 2011). Through simple comparison, Antweiler and Frank (2004) found that the accuracy of algorithm classification is actually higher than that of manual classification. Other than the Naive Bayes algorithm, some scholars have also used the maximum entropy model for text classification (Zhang and Swanson 2010).

During the process of data collection, posts related to stock information within forum class big data are often concentrated in the same message boards. This makes it relatively simple to determine the corresponding relationship between posts and stocks when collecting data. However, this process is relatively more complex for data of the microblog and search classes. This is because these two classes of data do not have specific boundaries that segregate the various stock data.

When acquiring search data, scholars have to collect corresponding data for a stock based on the keywords relevant to it. The accuracy of the search data is closely related to the keywords used for the search. Keywords available for scholars to choose from include stock code, company name, and names of a company's main products (Da et al. 2011a, 2011b). Keywords used for empirical research must be carefully selected by considering the special features of the research topic. Often, some stock codes or company names are similar to keywords for other objects and events. Examples include the Chinese stock-Zhangjiajie, which also is a city name, and American company Apple, which is also a kind of fruit. Search data for these types of keywords would often contain a lot of other irrelevant information, which must be removed during data collection (Da et al. 2011a).

For the acquisition of microblog class big data related to stocks, it is similarly necessary to select appropriate search rules for text screening to ensure that the search contents contain relevant information. In their study on Twitter, Sprenger et al. (2014a) used hashtags to collect data from text information. Sul et al. (2016) used the $ symbol and a stock's abbreviation as the search criterion.

Review

Application of Internet big data to capital markets

Internet big data and stock market performance

The Internet has revolutionized the manner in which information is transmitted and the pattern by which investors process information (Barber and Odean 2001; Moat et al. 2014). At the present stage, big data from forums, microblogs, and search engines are mainly used to examine their impact on stock market performance. Many scholars believe that the various indicators constructed on the basis of big data (e.g., those of market sentiments, divergence in opinions, and level of attention) impact multiple variables, including stock returns, trading volumes, and volatility. Information extracted from Internet big data can definitely explain stock market performance to a certain extent (Alanyali et al. 2013; Bordino et al. 2012; Gloor et al. 2009; Siganos 2013; Sprenger et al. 2014b; Wysocki 1998).

At the level of individual stocks, Wysocki (1998) initially found that the number of stock-related posts published at night is related to trading volumes. Using posts data

made on Yahoo! Finance and Raging Bull, Antweiler and Frank (2004) discovered that information contained in the posts can help forecast stock volatility and stock returns, although the latter does not have any economic significance. In addition, the greater the divergence in market sentiments, the larger the stock trading volume. Sprenger et al. (2014b) demonstrated that sentiments contained in tweets are significantly correlated with stock returns. The greater the number of daily microblogs, the larger the stock trading volume. Significant correlations were observed between divergence in market sentiments reflected on the microblogs and stock volatility.

Zhang et al. (2011b) conducted searches in Baidu using keywords. Furthermore, the authors defined the number of web pages returned from the searches as measurement indicators for the amount of information contained in social media. After analysis of information on social media and asset pricing, the authors found that the former is a rich source of effective information that affects abnormal stock returns. In addition, Vlastakis and Markellos (2012) found a significant and positive correlation between the volumes of Google searches and both market trading volumes and volatilities. When investors' level of risk aversion increases, Google search volumes also increase. Ruiz et al. (2012) examined the user-follower relationships established in microblogs and concluded that the connection components within social networks and the number of nodes in interaction graphs are significantly correlated with trading volumes and stock prices, with the correlation being stronger for the former than the latter. Nevertheless, trading strategies developed on the basis of correlation between stock prices are superior to basic trading strategies.

Empirical research by Zhang et al. (2014b) revealed that the intensity of online searches by investors impacts short-term stock returns, short-term trading volumes, and cumulative returns. In addition, investors' online searches have stronger explanatory power and better forecasting ability on the stock market compared with the traditional variables of investors' sentiments and level of attention. Shen et al. (2013) discovered that for a company facing negative public opinion about its private placement, the excess returns on stocks subsequent to the private placement notice are significantly negative.

Da et al. (2011a) found that an increase in search volumes often meant a rise in stock prices over the subsequent fortnight, in addition to price reversals within the year. A similar reversal effect exists in the Chinese stock market (Yu and Zhang 2012; Zhang et al. 2014a). Other scholars have analyzed the link between search volumes and initial public offering (IPO) premiums. Da et al. (2011a) found that an increase in searches prior to IPO indicates greater gains on the first day when the shares are listed. The study by Song et al. (2011), based on data from Google Trends, indicated that online search volumes for a pre-IPO stock have better explanatory power and forecasting ability on a company's level of stock sales, excess returns on the first day of trading, and long-term performance. Search volumes can explain 23% of the first-day excess returns and 10% or more of long-term cumulative returns. In addition, results from the empirical research by Nan (2015) indicate a significant and positive correlation between divergence of online opinions in stock message boards and IPO premiums.

For the indexes, Zhang et al. (2011a) found that the proportion of emotion-related vocabulary on microblogs is significantly and negatively correlated with the Dow Jones Index, NASDAQ Index, and S&P 500 Index. However, the proportion is significantly

and positively correlated with the Volatility Index. For the Dow Jones Index, Bollen et al. (2011b) showed that the addition of sentiment-type indicators significantly improves forecasting results. The accuracy of analysis for the daily direction of change in the Dow Jones Index was 86.7%, while the mean absolute percentage error decreased by 6%.

Da et al. (2015) found that the FEARS Index developed using Google's search data can predict trends in short-term returns, volatility changes, and capital flows of mutual funds. Furthermore, the addition of Twitter data improves the model's forecasting accuracy of the S&P 500 Index. Cheng and Lin (2013) showed that sentiment indicators of investors on social media are positively correlated with stock market index returns and trading volumes. The impact of those two factors on sentiment indicators can last for more than 40 trading days.

Mao et al. (2012) undertook a comprehensive analysis of the behavioral characteristics of Internet big data and various aspects of the stock market. They further analyzed the correlation between the number of Twitter posts and the stock market from the levels of individual stocks, industries, and indexes. Furthermore, they showed that Twitter could help predict stock market performance, especially at the indexes level.

However, the results of some empirical research has revealed that Internet big data does not improve forecasting results on the stock market. Kim and Kim (2014) highlighted that market sentiments contained in posts on message boards cannot predict future returns, volatilities, and trading volumes of stocks. In addition, the findings of Tumarkin and Whitelaw (2001) demonstrated that information in posts on message boards cannot predict stock returns or excess trading volumes, thereby supporting the efficient market hypothesis. Although Zhao et al. (2013) proved a positive correlation between search intensities on Baidu and stock returns, the rate of change in the level of attention is not a significant risk factor. The authors concluded that search intensities on Baidu do not systematically affect stock returns.

Although Internet big data was found to impact stock market performance, studies have shown that stock market performance similarly affects the behavioral characteristics of online investors. Kim and Kim (2014) discovered that sentiments in investors' posts are affected by past performance of the stock. Zhang et al. (2014b) indicated that although the stock market can affect online searches, online searching behaviors affect and predict stock market performance to a greater extent. The endogenous problem that exists between online searches and stock returns only has a minimal impact on forecasting ability.

Some scholars believe that the forecasting results of Internet big data on the stock market are affected by other factors. Many of these authors have discussed the relationship between the weight ratio of investor information and forecasting ability. Gu et al. (2006) weighed each post's recommendation by its author's credibility based on the accuracy of his/her past posts. The authors proved that a credibility-weighted recommendation of a stock message board can predict stock returns but a simple-weighted recommendation cannot. Yang et al. (2015) found that for the Twitter social network, weighted sentiment indicators based on critical nodes have better forecasting ability for the financial market than do general sentiment indicators. Zhang et al. (2016) identified financial users who were invited and certified by Sina Weibo as "celebrities." Through event study analysis, the authors determined that posts made by

celebrities could significantly predict stock returns compared to those made by ordinary users. The former contained more future public information and current private information, whereas the latter mostly comprised outdated information, indicating that the role of ordinary users tended toward one of information follower rather than of provider.

However, other scholars have learned that investors who are more influential might not publish information with better forecasting ability. According to Sul et al. (2016), tweets published by users with more followers are unable to predict stock returns, but those by fewer followers significantly impact future stock returns. In this regard, the authors believed that information disseminated by the former is quickly reflected in the stock prices and, hence, is not predictive. This reason was supported when analyzing the number of shares. Sul et al. (2016) further found that the more a piece of information is shared, the poorer is its forecasting results and vice versa. A trading strategy based on the aforementioned findings can achieve annualized returns of 11–15%.

Considering that Internet big data has the advantage of geographical identification, some scholars have introduced home bias into their studies on forecasting ability. Dong and Xiao (2011) established that the phenomenon of home bias exists in communication on stock message boards. There is a greater probability that investors in stock message boards will participate in discussions about local stock information. This home bias significantly impacts stock prices. The larger the proportion of local investors involved in information exchange, the higher are stock prices.

Huang et al. (2016) used IP data from Eastmoney.com to construct a quantitative indicator of investors concerned about home bias. The construction of this indicator is more refined compared with that by Dong and Xiao (2011). Their findings demonstrated that the situation in which investors are concerned about home bias is more severe in less developed regions and that the level of concern is affected by market size, turnover rate, and name of securities. Ackert et al. (2016) found that the advice of opinion leaders has greater investment value. In addition, the authors were more concerned about corporations from "home" and, thus, were more accurate when making related forecasts.

Nevertheless, other scholars have held the view that Internet big data's forecasting ability with regard to stock market performance is affected by other factors, including the difficulty of a stock being arbitraged, level of attention on the company concerned, event type, and disclosure environment. Joseph et al. (2011) compiled search data on S&P 500 component stocks and defined search intensity as the indicator for investors' market sentiments. Search intensity is considered to forecast weekly stock returns and trading volumes steadily. Moreover, the relationship between returns and search volume might be affected by the difficulty of a stock being arbitraged. For companies that received less attention from the market, Blankespoor et al. (2013) established that the posting of information via Twitter can reduce the degree of information asymmetry. Furthermore, the authors established that a positive correlation exists between the level of information dissemination and stock liquidity. In addition, Sprenger et al. (2014a) indicated that advanced stock returns for good news are higher than those for bad news and that the impact of news events on stock markets significantly differs for various event types.

The stocks of a small company are small cap and have weak profitability and poor underwriting capacity. Nan (2015) discovered that the IPO premiums for such stocks are more vulnerable to divergence of opinions among investors on stock message boards. Xu and Chen (2016) empirically revealed that disclosure via microblogs could significantly increase same day excess returns and excess trading volumes for the company's stocks. The level of increase is affected not only by the degree of disclosure intensity and information density of the disclosure but also by noise information in the microblogs. In addition, when microblogs are used to disseminate information that has already been made public, the resultant market response will be stronger than an announcement that is not circulated via microblogs. The impact of disclosures via microblogs is greater for companies that are relatively out of the limelight, and the effect of such disclosures on the trading behavior of individual investors is more significant.

Internet big data and other research related to capital markets

Other studies have been conducted in areas outside of stock market performance. These studies have found that Internet big data can predict the performance of companies. Da et al. (2011b) demonstrated that the search intensity for companies with main products can predict their corporate profitability upon listing. The forecasting effect is especially significant for corporations that have fewer products and for growth companies. In addition, Bartov et al. (2015) discovered that overall sentiment indicators can predict a company's quarterly profitability, in addition to excess earnings after announcement of its quarterly earnings data. This forecasting effect is more pronounced for companies whose information environments are poorer. Shen et al. (2013) found that for companies facing more negative public opinions, the probability of their performance declining after implementation of private placement is higher.

Other scholars' area of interest is the impact of Internet big data on the regulatory mechanism. An example is the empirical research by Shen et al. (2013), who found that companies facing negative public opinions about theirs private placements have a significantly lower probability of their private placement proposals being approved by the relevant departments after evaluation. However, such negative public opinions do not significantly impact the probability of the private placement proposal being passed at shareholders' meetings. Separately, studies about the herding effect have demonstrated that the immediate and next day effects are weakened by online communication. Dynamic interactions exist between this effect in the stock market and online communication; the latter can weaken the herding effect, suppress the continued spread of herding behavior, and improve market efficiency (Zheng et al. 2015).

Conclusions and future research

From the literature review, Internet big data sources related to present capital markets research can be categorized into forum-type data, microblog-type data and search class data. Based on these data, researchers can build some more complicated variables to analyze traditional questions. With regards to research about investors' sentiments based on Internet big data, the main methods of sentiments analysis include building an inventory of lexical categories, using dictionaries for analysis of lexical categories, and machine learning. Many studies analyze whether Internet big data can predict

capital markets. but they reach no consistent conclusions, which might be due to sample and analysis methodlimitations.

Through the summary stated above, we believe that research on Internet big data and capital markets has achieved some results. Nevertheless, there remains room for improvement and enhancement in the methods used for data collection and analysis, as well as the areas covered in the research. In most existing studies about Internet big data and capital markets, data collection was mainly undertaken for forum, microblog, and search classes of big data. From the perspective of data acquisition, the majority of the samples in the literature were treated as representative samples. These included the component stocks of the various indexes and specified stocks of high-tech enterprises. The majority of the samples did not strictly comply with the definition of research based on full samples, which should be an important characteristic of big data research. Moreover, the data source was usually a single and particular platform. Relatively few studies have analyzed data that were comprehensively collected from multiple platforms. Overall research based on big data would benefit if there were bigger breakthroughs in terms of data collection and information aggregation.

The time spans of many studies tended to be relatively short because these were limited by the short traceback time of data from many platforms. In addition, the degree of replicability of research conclusions by other scholars was low, and the conclusions did not facilitate multiangle studies for a specific issue using the same benchmarks as the basis. This issue can be resolved only through the establishment of specialized databases that cater to research. This would require the cooperation of corporations that are sources of big data through the implementation of an appropriate method.

Presently, Twitter has been officially authorized and is building the GNIP database for supporting research (Bartov et al. 2015). However, specialized databases have not yet been developed for other forum- and microblog-class platforms. Separately, most study samples were collected by the study teams themselves programming. During the initial sample collection stage, it was inevitable that microblog- and search-class big data face the problem of noise interference. The direction of future research for sample collection algorithms is to identify methods for the accurate identification and collection of fuzzy but relevant information (Godbole et al. 2007). Big data research samples will undoubtedly be more comparable if specialized databases were to be established through the application of unified technical means across big data platforms, which are suitable for research use.

From the perspective of data analysis, analytical methods used in current studies have remained relatively simple and have room for improvement in terms of accuracy level. Taking the textual analysis technique as an example, Koppel and Shtrimberg (2006) found that the classification accuracy of machine learning algorithms can reach 70.3 and 65.9% for intra-sample and out-of-sample classifications, respectively. In addition, high overall accuracy in classification of textual sentiment has been achieved through the machine learning method. However, even higher accuracy levels would be beneficial for eliminating noise interference in studies, thereby ensuring stability of the conclusions (Nardo et al. 2016).

This is especially the case for textual analysis of Chinese big data. Research on big data and capital markets mainly depends on manual identification, which involves

subjective judgments, causing differences to remain between keywords selected in most studies. This situation introduces a certain degree of interference to the stability of conclusions. Many highly effective algorithms for mining of textual semantics have been introduced progressively into the field of financial research. These include different classifier algorithms coupled with a voting theme (Das and Chen 2007), support vector machines (De Choudhury et al. 2008), and five-stage filtering (Bettman et al. 2010).

At present, studies on Internet big data are mainly focused on their forecasting effects on stock market performances. Relatively few scholars have discussed the impact of Internet big data on corporate behavior. In the era of big data, it is possible to apply Internet big data to the forecasting of not only stock market performance but also companies' performance levels (Da et al. 2011b). Furthermore, Shen et al. (2013) indicated that such data can be used to forecast the probability of decline in companies' future performances. Bartov et al. (2015) found that overall sentiment indicators can predict companies' quarterly profitability.

Simultaneously, the literature has examined the impact of Internet big data on regulatory mechanisms (Shen et al. 2013). In modern societies, companies' management teams are in environments surrounded by massive volumes of information. Are a company's decisions on capital structure, corporate governance, and other corporate behavior somehow affected by these Internet big data? These relationships await further research by scholars.

Endnotes

[1]In computer terminology, ZB refers to one sextillion bytes. The term KB in our daily usage is the acronym for "kilobyte."

[2]World Economic Forum: Unlocking the value of personal data: From collection to usage, 2013.

[3]Rotella, P. (2012, April 2). Is data the new oil? Forbes.

[4]Obama administration unveils "Big data" initiative: Announces $200 million in new R&D investments. http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release_final_2_pdf, 2012.

[5]UN Global Pulse. Big data for development: Challenges & opportunities, 2012.

[6]http://finance.sina.com.cn/money/fund/jjpl/2015-12-31/doc-ifxncyar6085226.shtml.

[7]The size of funds based on big data is derived from the website and the total size is calculated by the author http://fund.eastmoney.com/data/fundsearch.html?spm=search&key=%E5%A4%A7%E6%95%B0%E6%8D%AE#key%E5%A4%A7%E6%95%B0%E6%8D%AE.

[8]http://funds.hexun.com/2016-10-31/186671597.html.

[9]According to the statistics by China Webmaster (http://top.chinaz.com/), the overall popularity ranking of Eastmoney.com's stock message boards far exceeds that of other stock message boards.

Acknowledgements
This paper is funded by National Nature Sciences Foundation of China (No. 71372148).

Funding
National Nature Sciences Foundation of China (No. 71372148).

Authors' contribution
GL gave the main idea of the review paper and gave some improvement suggestions. MY collected references and wrote the main body for the paper. Both authors read and approved the final manuscript.

Authors' information
Minjian Ye is a PhD student at Sun Yat-Sen Business School, Sun Yat-Sen University. His research interests are corporate governance, capital structure, and big data analysis.
Guangzhong Li is a professor of finance at Sun Yat-Sen Business School, Sun Yat-Sen University. He has published several papers in *Review of Finance*、*Journal of Corporate Finance*、*Journal of International Money and Finance*、Journal of Business Finance and Accounting、Journal of Comparative Economics. His research interests are corporate finance, financial institution, and big data analysis.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Received: 6 December 2016 Accepted: 26 March 2017
Published online: 20 April 2017

References
Ackert LF, Jiang L, Lee HS et al. (2016) Influential investors in online stock forums [J]. Int Rev Financ Anal 45:39–46
Alanyali M, Moat HS, Preis T (2013) Quantifying the relationship between financial news and the stock market [J]. Sci Rep 3:3578
Antweiler W, Frank MZ (2004) Is all that talk just noise? The information content of internet stock message boards [J]. J Financ 59(3):1259–1294
Barber BM, Odean T (2001) The internet and the investor [J]. J Econ Perspect 15(1):41–54
Bartov E, Faurel L, Mohanram PS (2015) Can Twitter Help Predict Firm-Level Earnings and Stock Returns? Available at SSRN 2631421
Bettman JL, Hallett AG, Sault S (2010) Exploring the impact of electronic message board takeover rumors on the US equity market. SSRN working paper
Blankespoor E, Miller GS, White HD (2013) The role of dissemination in market liquidity: Evidence from firms' use of Twitter™ [J]. Account Rev 89(1):79–112
Bollen J, Mao H, Pepe A (2011a) Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena [J]. ICWSM 11:450–453
Bollen J, Mao H, Zeng X (2011b) Twitter mood predicts the stock market [J]. J Comput Sci 2(1):1–8
Bordino I, Battiston S, Caldarelli G et al (2012) Web search queries can predict stock market volumes [J]. PLoS One 7(7):e40014
Cheng W, Lin J (2013) Investors' market sentiments on social media and stock market indices [J]. Manage Sci 05:111–119
Da Z, Engelberg J, Gao P (2011a) In search of attention [J]. J Financ 66(5):1461–1499
Da Z, Engelberg J, Gao P (2011b) In search of fundamentals [C]. AFA 2012 Chicago Meetings Paper
Da Z, Engelberg J, Gao P (2015) The sum of all FEARS investor sentiment and asset prices [J]. Rev Financ Stud 28(1):1–32
Das SR, Chen MY (2007) Yahoo! for Amazon: Sentiment extraction from small talk on the web [J]. Manag Sci 53(9): 1375–1388
Das S, Martínez-Jerez A, Tufano P (2005) eInformation: A clinical study of investor discussion and sentiment [J]. Financ Manag 34(3):103–137
De Choudhury M, Sundaram H, John A et al (2008) Can blog communication dynamics be correlated with stock market activity? [C]. Proceedings of the nineteenth ACM conference on Hypertext and hypermedia. ACM 2008:55–60
Delort JY, Arunasalam B, Milosavljevic M et al (2011) The Impact of Manipulation in Internet Stock Message Boards. Int J Bank Account Finance 8(4):1–18
Dong D, Xiao Z (2011) Home bias in the exchange of stock information and its impact on stock prices: evidences from stock message boards [J]. Manage World 01:52–61, +188
Drake MS, Roulstone DT, Thornock JR (2012) Investor information demand: Evidence from Google searches around earnings announcements [J]. J Account Res 50(4):1001–1040
Gloor PA, Krauss J, Nann S et al (2009) Web science 2.0: Identifying trends through semantic social network analysis [C]. Computational Science and Engineering, 2009. CSE'09. International Conference on. IEEE 4:222–215
Godbole N, Srinivasaiah M, Skiena S (2007) Large-Scale Sentiment Analysis for News and Blogs [J]. ICWSM 7(21):219–222
Gu B, Konana P, Liu A et al (2006) Identifying information in stock message boards and its implications for stock market efficiency [C]. Workshop on Information Systems and Economics, Los Angeles
Huang Y, Qiu H, Wu Z (2016) Local bias in investor attention: Evidence from China's Internet stock message boards [J]. J Empir Financ 38:338–354
James J (2012) Data never sleeps: How much data is generated every minute [J]. Domo Blog 2012:8
Jiang C, Liang K, Ding Y, Liu S, Liu Y (2015) Forecasting stock behaviors based on social media [J]. Chin J Manag Sci 01:17–24
Jin X, Shen D, Zhang W (2016) Has microblogging changed stock market behavior? Evidence from China [J]. Physica A: Stat Mech Appl 452:151–156
Joseph K, Wintoki MB, Zhang Z (2011) Forecasting abnormal stock returns and trading volume using investor sentiment: Evidence from online search [J]. Int J Forecast 27(4):1116–1127

Kim SH, Kim D (2014) Investor sentiment from internet message postings and the predictability of stock returns [J]. J Econ Behav Organ 107:708–729
Koppel M, Shtrimberg I (2006) Good news or bad news? Let the market decide [M]. Computing attitude and affect in text: Theory and applications. Springer Netherlands., pp 297–301
Leung H, Ton T (2015) The impact of internet stock message boards on cross-sectional returns of small-capitalization stocks [J]. J Bank Financ 55:37–55
Liu F, Ye Q, Li YJ (2014) Impacts of interactions between news attention and investor attention onstock returns: Empirical investigation on financial shares in China. J Manag Sci Chin 17(1):72–85
Loughran T, McDonald B (2011) When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks [J]. J Financ 66(1):35–65
Mao Y, Wei W, Wang B et al (2012) Correlating S&P 500 stocks with Twitter data [C]. Proceedings of the first ACM international workshop on hot topics on interdisciplinary social networks research. ACM 69:72
Moat HS, Preis T, Olivola CY et al (2014) Using big data to predict collective behavior in the real world [J]. Behav Brain Sci 37(01):92–93
Nan X (2015) Impact of divergence of opinions among online investors on IPO premiums in the new media era: a mining method based on data from stock message boards [J]. Chin Soft Sci 10:155–165
Nardo M, Petracco M, Naltsidis M (2016) Walking down Wall Street with a tablet: a survey of stock market predictions using the web [J]. J Econ Surv 30(2):356–369
Ruiz EJ, Hristidis V, Castillo C et al (2012) Correlating financial time series with micro-blogging activity [C]. Proceedings of the fifth ACM international conference on Web search and data mining. ACM 2012:513–522
Sabherwal S, Sarkar SK, Zhang Y (2011) Do internet stock message boards influence trading? Evidence from heavily discussed stocks with no fundamental news [J]. J Bus Finance Account 38(9–10):1209–1237
Shen Y, Yang J, Li P (2013) Corporate governance role of Internet public opinions: empirical evidences based on private placements [J]. Nankai Bus Rev 03:80–88
Siganos A (2013) Google attention and target price run ups [J]. Int Rev Financ Anal 29:219–226
Song S, Cao H, Yang K (2011) Investor concerns and IPO anomalies: empirical evidence from online search volumes [J]. Econ Res J S1:145–155
Sprenger TO, Sandner PG, Tumasjan A et al (2014a) News or Noise? Using Twitter to Identify and Understand Company-specific News Flow [J]. J Bus Finance Account 41(7–8):791–830
Sprenger TO, Tumasjan A, Sandner PG et al (2014b) Tweets and trades: The information content of stock microblogs [J]. Eur Financ Manag 20(5):926–957
Sul HK, Dennis AR, Yuan LI (2016) Trading on Twitter: Using Social Media Sentiment to Predict Stock Returns [J]. Decision Sciences
Takeda F, Wakao T (2014) Google search intensity and its relationship with returns and trading volume of Japanese stocks [J]. Pac Basin Financ J 27:1–18
Tumarkin R, Whitelaw RF (2001) News or noise? Internet postings and stock prices [J]. Financ Anal J 57(3):41–51
Vlastakis N, Markellos RN (2012) Information demand and stock market volatility [J]. J Bank Financ 36(6):1808–1821
Wysocki PD (1998) Cheap talk on the web: The determinants of postings on stock message boards [J]. University of Michigan Business School Working Paper., p 98025
Xu W, Chen D (2016) Role of information disclosed by the media: empirical evidence from Sina Weibo [J]. J Financ Res 03:157–173
Yang SY, Mo SYK, Liu A (2015) Twitter financial community sentiment and its predictive relationship to stock market movement [J]. Quant Finan 15(10):1637–1656
Yu Q, Zhang B (2012) Investors' limited attention and stock returns: an empirical study using the Baidu Index as indicator of the level of attention [J]. J Financ Res 08:152–165
Zhang Y, Swanson PE (2010) Are day traders bias free? Evidence from internet stock message boards [J]. J Econ Financ 34(1):96–112
Zhang X, Fuehres H, Gloor PA (2011a) Predicting stock market indicators through twitter "I hope it is not as bad as I fear" [J]. Procedia Soc Behav Sci 26:55–62
Zhang Y, Zhang W, Jin X, Xiong X (2011b) Does the Internet know more? Open source information and asset pricing [J]. Syst Eng Theory Pract 04:577–586
Zhang X, Fuehres H, Gloor PA (2012) Predicting asset value through twitter buzz [M]. Advances in Collective Intelligence 2011. Springer Berlin Heidelberg., pp 23–34
Zhang J, Liao W, Zhang R (2014a) Impact of ordinary investors' level of attention on the volume and price of stock market transactions: an empirical study based on the Baidu Index [J]. Account Res 08:52–59, +97
Zhang Y, Li Y, Su Z, Zhang Z (2014b) Can online searches be used to forecast stock market performance? [J]. J Financ Res 02:193–206
Zhang Y, An Y, Feng X et al (2016) Celebrities and ordinaries in social networks: Who knows more information? [J]. Finance Research Letters
Zhao L, Lu Z, Wang Z (2013) Public's search for stocks on Baidu: empirical research on the relationship between stock returns and search volume on Baidu [J]. J Financ Res 04:183–195
Zheng Y, Dong D, Zhu H (2015) Can online exchange of stock information weaken the herding effect? An analysis based on the Chinese stock market [J]. Manage Rev 06:58–67