Research Article

International Journal of Distributed Sensor Networks

International Journal of Distributed Sensor Networks
2017, Vol. 13(1)

DOI: 10.1177/1550147716684842
journals.sagepub.com/home/ijdsn
SAGE

An inside look into the complexity of box-office revenue prediction in China

Jia Xiao[1], Xin Li[1], Shanzhi Chen[2], Xuhui Zhao[1] and Meng Xu[1]

Abstract

In this article, we discuss various elements contributing to exerting influence on box office in China, which are divided into internal and external factors. Since these factors could merely be quantified by online data sources partially or inaccurately, we propose that relativity analysis is more reasonable than precise revenue prediction. Trailer is selected as the combination of movie content and online behavior prior to releasing. Indexes from seven mainstream video websites are retrieved by the designed big data system which is integrated with the Internet of things technology. Correlation coefficients of different time periods are calculated. We apply multiple linear regression with stepwise method in modeling and prove that watching counts of 1 week before releasing on Youku is the barometer of market performance, especially the first week revenues. We also manifest the power of influential users through constructing Sina Weibo acquisition and analysis system.

Keywords

Revenue, big data, linear regression, Internet of things, video website, correlation coefficient

Date received: 1 September 2016; accepted: 27 November 2016

Academic Editor: Xuyun Zhang

Introduction

The Internet of things (IoT) and big data[1] have gotten more and more attention with the development of social and economy in China. Their combination[2] is opening splendid opportunities for a large number of interesting fields, especially everyday objects. Taking the sphere of entertainment as an example, big data integrated with IoT technology[3] and actually the people behind are broadly applicated in industry upgrading and optimization. Movie industry is growing by leaps and bounds here, and the national box-office revenue has increased from 13 billion in 2011 to 44 billion in 2015, with an average growth rate of nearly 40% yearly. More and more investors have been putting their capital into this field. Alibaba and Jingdong (e-commerce giants in China) even created online financial products promoted by their platforms to attract civilian flows. It is known to all that a movie could be the box-office smash or loss in one month. Not only is revenue the most important indicator of the success of a film, but also the profit of the investors. It is depicted in Figure 1 that the revenues of domestic films maintain growing trend all these years.

According to film industry records, China has begun to keep explicit box-office receipts since 2012. We retrieve all the data from m1905.com with Java-based web crawler named jsoup to analyze the overall distribution of revenues from 2012 to April 2015.

As shown in Figure 2, three-fourths of the revenues are extremely low. In contrast, less than one-fourth occupy a large cut of the market. That means if we

[1]State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing, China
[2]State Key Laboratory of Wireless Mobile Communications, China Academy of Telecommunications Technology, Beijing, China

Corresponding author:
Jia Xiao, State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, No 10, Xitucheng Road, Haidian District, Beijing 100876, China.
Email: tonyalex2010@163.com

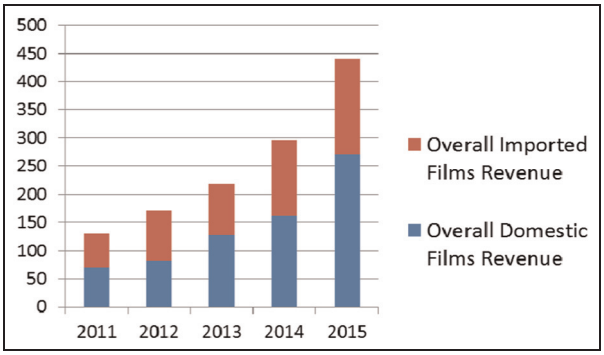

Figure 1. Comparison of release and revenue rates.

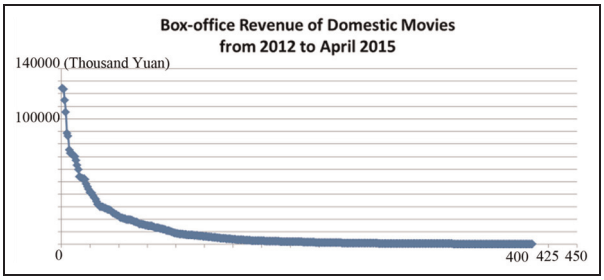

Figure 2. Revenue distribution of movies in China.

predict the profit of any movie is poor, the accuracy rate is higher than 75%. Obviously, the probability of a box-office hit is less than 25% vice versa. This phenomenon follows Zipf's law. To mitigate business risk effectively and increase average returns significantly, revenue prediction in China is deeply worthy of studying.

However, most of previous works discussed foreign movies rather than domestic films in China. Moreover, their methods were merely applied to part of works, either a specific period of time or certain types of movies. Third, most researchers claim that they could achieve the target of accurate prediction to some extent. Last but not least, nearly none of them combined the three characteristics of video content, user-generated feedback, and multiple data sources.

This article discusses the complexity of box-office prediction in detail and focuses more on the relativity analysis rather than the accuracy of forecast. First, box-office revenue is a result of multi-factors, some of which are extremely difficult to quantify. Second, there are a lot of noises contained in these online data and it is never easy to purify and process them with diverse algorithms. Third, film is an art form over time, and historical mathematical models could not predict the so-called black swan event. Accordingly, we select trailers in seven domestic video websites as a combination of the three features mentioned in the last paragraph and collect relevant indexes with the designed big data analysis system which is integrated with the IoT technology. Then, we utilize the multiple regression method to analyze the relationship between revenues and various indexes from different websites and reveal one potential rule. Due to social media's character of high popularity, it can be precious resources for promotion and hype. We also construct Weibo acquisition and analysis system to visualize the information diffusion from promoters to recipients.

Related work

In this area, scientists have been maintaining a high degree of concern. Generally speaking, these articles could be categorized into two types based on data sources. One type focuses on quantifying the elements extracted from the film, which is more conventional. And, they could only get small amount of feedback. Meanwhile, more and more researchers devote to solving this problem with another approach. There are online signals contributing to reflecting receipts, such as web search counts and reviews or ratings on social networks and video websites.

Film is a work of art comprised of several innate elements, such as producer, director, performer, script, genre, and award. Researchers have been trying to reveal the hidden relationship between them and market performance with small amount of feedback. However, different people hold different viewpoints:

T King[4] found that there was no correlation between critical ratings and revenue.
Y Zhang and X Zhang[5] insisted that revenues had positive correlation with capital, negative correlation with piracy and nothing to do with valence.
MB Houston and G Walsh[6] showed that consumers' quality perception and awards were the key elements leading to motion picture success.
P He[7] revealed that creation, time length, advertising, and investment were the key elements of a successful film with high income.
Hu et al.[8] indicated that the combination of actor and director was the most influential factor among all the elements and director weighed more.
M Wasserman et al.[9] performed distribution and correlation analysis to prove the connection between user voting data and the economic statistics.
J Eliashberg et al.[10] developed a methodology extracting three textual features from scripts to predict revenue of a movie at the point of green-lighting, when only its script and estimated production budget were available.

Recently, film-makers are committed to extensive media hype. Market would react to these propagation

incentives. Therefore, researchers could collect online data to solve this problem:

J Krauss et al.[11] introduced a new web mining approach that combined social network and automatic sentiment analysis, which showed that the discussion patterns on IMDb could predict Academy Awards nominations and box-office success.
S Goel et al.[12] found that search counts were highly predictive of future outcomes. The results suggested that search query volume could be used to forecast the opening weekend revenue.
A Oghina et al.[13] extracted surface and textual features with data from Twitter and YouTube to predict IMDb movie ratings.
Google Whitepaper[14] highlighted the magic power of search counts in revealing their correlation with box-office receipts, although these data were not open.
R Yao and J Chen[15] introduced sentiment analysis and machine learning methods to study the relationship between the online reviews for a movie and its box-office revenue performance.
M Mestyan et al.[16] proved that the popularity of a movie could be predicted much before its release based on data extracted from the entry to the movie in Wikipedia, the online encyclopedia.
S Moon et al.[17] applied machine learning techniques and linear modeling to develop a model for predicting the near-weekend ticket sales and the ideal number of screens using web-based external factors, such as online reviews, star ratings, and search volume.
S Thigale et al.[18] used sentiment analysis of Twitter data for the hype creating among the mob and showed that social media expressed a collective knowledge which can yield a powerful and accurate indicator of future revenues.
J Du et al.[19] utilized Tencent microblog in extracting two sets of features, which were count-based features and content-based features, to predict box offices of certain movies in China. But this source is getting more and more invalid due to losing users.
T Kim et al.[20] proposed a novel approach to the box-office forecasting of motion pictures using social networking service (SNS) data and several machine learning–based algorithms. A genetic algorithm was adopted for building models.

As A Bhave et al.[21] and MT Lash and K Zhao[22] had proposed, both internal and external factors of films were crucial in predicting revenues based on papers listed above, which are exactly what we suggest and analyze below. There are also many other papers[23–47] that proposed valuable methods to solve different aspects in data collection and analysis process, such as security, parallelism, performance, and algorithm, which gave us a lot of enlightenment.

Table 1. Elements of internal factors.

Factor	Element
Investment	RMB
Title	*Word*
Script	Word
Time Length	Minute
Schedule	New Year, Summer, Festival Movie Season
Rival	High/Medium/Low
Genre	Comedy/Action/ Literary/Horror/ Suspense/Crime/ Adventure/Child/ War/Eroticization/ History/Romance/ Animation/Family/ Biography/Musical/ Fantasy/Sci-fi
Cast	*Director*
	Producer
	Leading Actor
	Leading Actress
	Supporting Actor
	Supporting Actress
Award	International Award
	Domestic Award
Advertising	*Attention Rate*

Difficulty of quantifying revenue with internal factors

We work with some professionals in this field to analyze the internal factors, which are categorized into 10 types, as shown in Table 1. Based on results which have been investigated, we divide them into three categories, which are invalid type (gray mark), indirect quantifiable type (italic mark), and segmental type (bold mark).

Invalid factors

Not all factors in Table 1 are directly correlated with revenue. The most obvious one is time length, which could be crawled from Youku website. We take all movies in 2014 as a sample set to calculate correlation coefficient of the two with equation (1). The result is 0.42555, which means relativity is fairly weak

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (1)$$

According to film industry law, large capital input generally equals to big revenue. However, there is no

Table 2. Domestic films with grand foreign awards.

Title	Year	Domestic revenue (million yuan)
A Simple Life	2011	68.10
White Deer Plain	2012	129.75
A Touch of Sin	2013	No record
Unbeatable	2013	118.2
Black Coal, Thin Ice	2014	103.6

public source to obtain investment value. Even if not, it is usually exaggerated by film-makers or producers. How could researchers fulfill precise prediction with false input? Not to mention some works with small investment yield high output. As a one-time box-office champion of domestic movies, the capital investment of "Lost In Thailand" (short for "L") is a little bit lower than the medium level. Therefore, this element is invalid in prediction.

Script is a potential source for text mining, which could be probably used as a method of performance assessment. However, to the best of our knowledge, there is still no public way to obtain scripts of domestic movies that have not been released yet. We think this approach is beyond our capacity.

According to the rules of domestic movie awards, the film can only participate in contest after its release. Hence, the element of domestic award would be no avail in prediction. When it comes to international award, the situation is different. Some literary movies would probably suffer a defeat in revenue without these awards. But these samples, which are shown in Table 2, are extremely limited. Based on these facts, this element could not be put into widely used in prediction and it can also be partly reflected through other means, such as online behavior.

Indirect quantifiable factor

1. *Descriptive elements*. Advertising plays exceeding role in engaging audiences. Nowadays, there are multiple channels, both online and offline, for advertising delivery. Typical examples are Weibo (microblog), WeChat, variety show, and outdoor billboard. However, it is extremely difficult to quantify the influence of advertising. Online and offline promotions would have effects on each other. Both of them would exert impact on people, while human behavior could also generate feedback on advertising vice versa. Furthermore, influences of the three are intertwined and difficult to be distinguished, except online behavior might be partially quantified by public data. Eventually, only part of these factors can stimulate purchasing, which turns into box-office revenue (Figure 3).

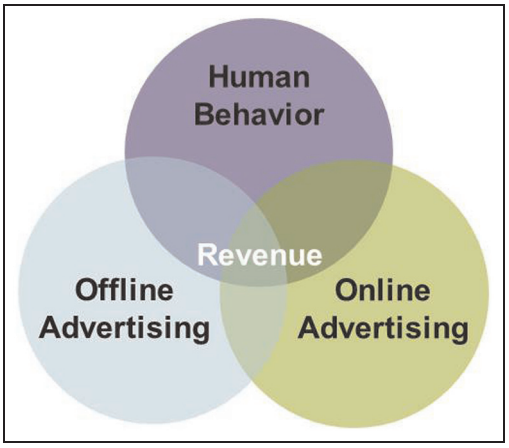

Figure 3. Triangle relationship.

As far as cast is concerned, we think it is the kernel of each movie. Six key elements of this factor are listed in Table 1. However, there is a common character of these elements, which is the opaque market price. That means it is hard to measure them. Since input is fuzzy, we need resort to online methods to quantify them.

Title is the name of a movie. Comparing with script, its fatal flaw lies in the number of words. That means the information available for data mining is very limited. Therefore, the number of occurrences and film sequels are some important evaluating indicators which are worthy of considering.

2. *Online data sources*. The four factors which are analyzed above could only be partially measured by online behavior. Prediction might be possibly accomplished only could we get accurate anticipation and feedback of potential candidates. There are three main public channels to collect online data indicating audience attention, which are search engines, video websites, and social networks (Table 3).

3. *Prediction deviation analysis*. We do admit that online world is the reflection of real world. But the premises of prediction are the volume, variety, and velocity of online data. Only the three features achieved might guarantee the characterization of present behavior.

It is known to all that many group-buying websites in China offer online discount ticket business, like Guevara, Maoyan, Meituan, Nuomi, and Dianping. However, their data are not publicly available. Various jealousy traditional corporations even Internet giants enter this business urgently, such as China Merchants Bank and NetEase. Neither their data are reachable

Table 3. Mainstream online data sources.

Channel	Mainstream	Target
Search engines	Baidu	Baidu Search Index
	360	360 Search Index
	Sougou	Sougou Search Index
Video websites	Youku	Various Indexes
	iQiyi	
	Sougou Video	
	Tencent Video	
	LeTV	
	Thunder Kankan	
	56.com	
Social networks	Sina Weibo	Volume and Valence
	Wechat	Public Account

publicly. Mobile Internet is a new drive for growth in ticket business. More and more apps are soaring each day. Their data are also helpful in customer figure portray but still difficult to access. Consequently, the problem of data deficiency is the first reason contributing to prediction deviation.

Second, it is questionable to identify the precise audience range. The data from different channels are generated by different people or the same person. Taking an example, a man might use Baidu to obtain movie information after he received the signal from outdoor advertisement. An audience probably wandered around Youku after using the 360 search engine. More likely, a film fan updated reviews on his Weibo after employing web search and video website. Moreover, superstars have strong offline fan groups even without Internet, just like Jackie Chan. Unfortunately, only part of these users would be filmgoers finally, while some of them are not. It is possible to achieve accurate prediction only when precise audience range is identified (Figure 4).

Setting two typical Weibo marketing as an example, Han Han and Jingming Guo are two famous young writers in China. They have more than 30 million Sina Weibo followers. Both of them released their film works in July 2014. Hence, they utilized Weibo as a fantastic platform for film propagation. Industry insiders believed that main audience would be their Weibo fans until Word-of-Mouth (WoM) went into effect. Commonly, movies have limited life cycle for several weeks in theaters. Therefore, we depict the trend of reviews, retweets, and revenues of the first week to analyze their relationship. From Figures 5 and 6, we can conclude that the correlation between the numbers of daily retweets, reviews, and revenues is weak, which means that active Weibo fans are not equal to filmgoers. It would probably be invalid if we try to identify accurate audience range with Weibo behavior.

To study Weibo data more visually and vividly, we construct the Weibo acquisition and analysis system with Java based on application programming interfaces (APIs) supplied by Sina Weibo. We also take this sample released by Jingming Guo and Han Han to do retweet and review analyses. Based on the calculation results of betweenness, degree, and average path length, Figures 7–10 manifest clearly that the appeal of big V and verified users overwhelms that of grassroots.

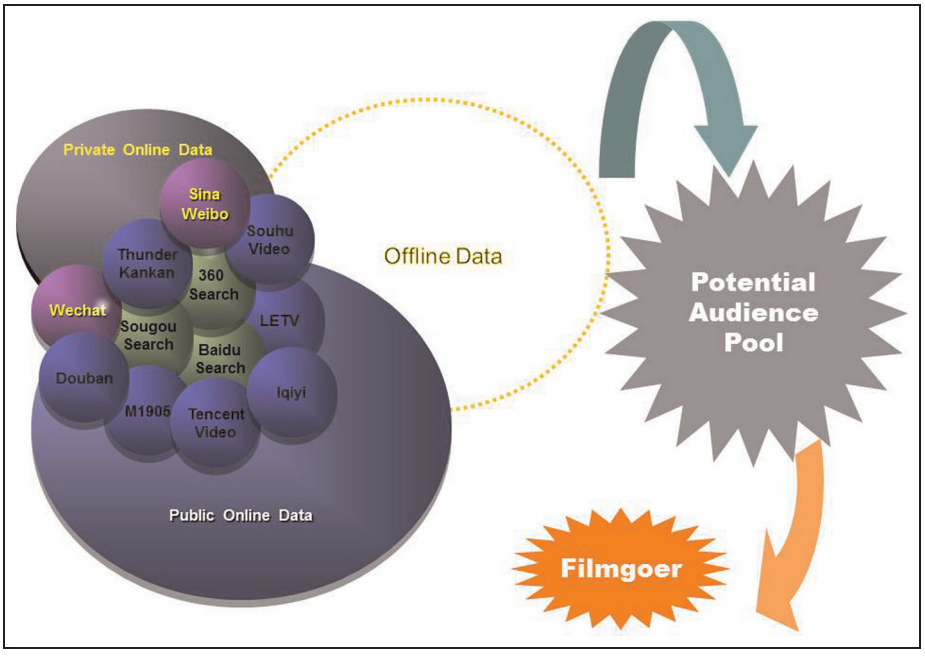

Figure 4. Relationship between data and audience.

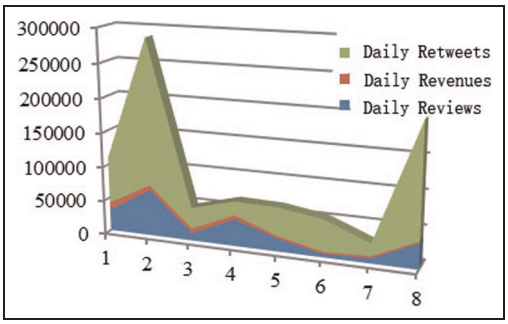

Figure 5. "Tiny Times 3.0" by Jingming Guo.

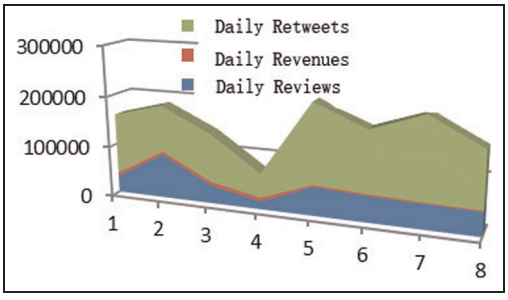

Figure 6. "The Continent" by Han Han.

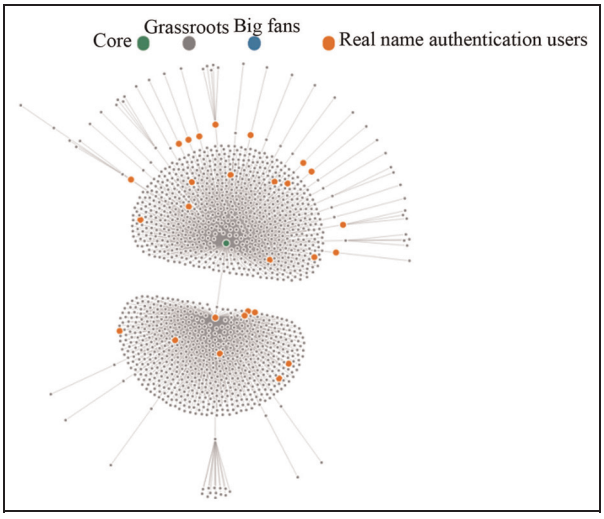

Figure 7. Retweet analysis released by Jingming Guo.
Left: weibo.com/1188552450/Bdwfp23jH?from=page_1035051188552450_profile&wvr=6&mod=weibotime

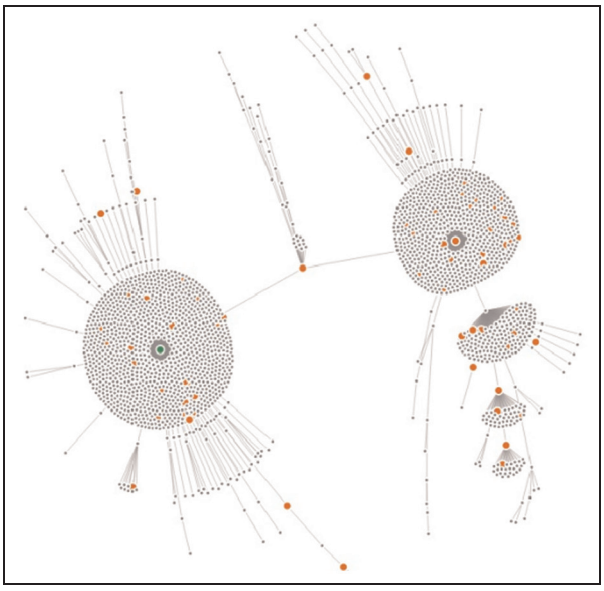

Figure 8. Retweet analysis released by Han Han.
Right: weibo.com/1191258123/B5Lw6zuPP?from=page_1035051191258123_profile&wvr=6&mod=weibotime

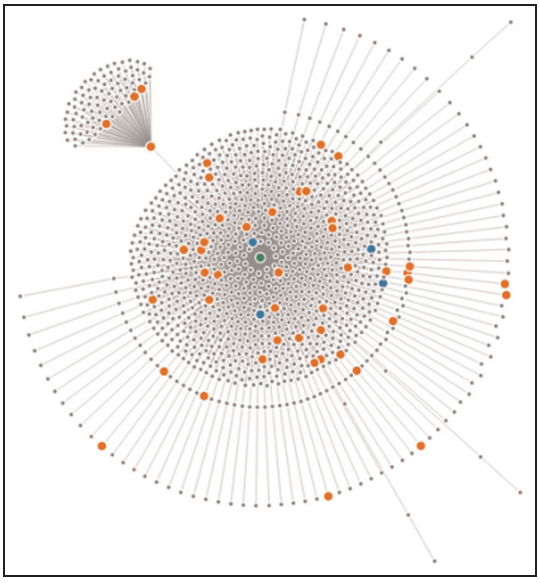

Figure 9. Review analysis released by Jingming Guo.
Left: weibo.com/1188552450/Bdwfp23jH?from=page_1035051188552450_profile&wvr=6&mod=weibotime

The interaction among influential users is much more pervasive than ordinary people. Film marketers find that social media can play an important role in raising potential audience's attention and interest in an early stage of film marketing.[48] Therefore, if any film marketing teams aim to improve the spread of Weibo advertisement, we suggest they should have some big V followers rather than grassroots in giving a helping hand to retweet and review propaganda Weibo (Table 4).

Velocity is the third element of online data, which means that time is important to prediction. Generally speaking, the closer to the screening date, the better is the prediction result. However, the effect of WoM begins to surge after releasing, while other factors start

Table 4. Layer distribution of each type of users.

Layer	①	②	③	④
1	17/0/683	17/1/945	38/3/1012	52/8/1623
2	8/0/621	1/0/83	8/0/167	38/2/375
3	0/0/18	10/0/529	0/0/2	1/0/8
4	None	6/0/196	None	None
5	None	3/0/62	None	None
6	None	1/0/43	None	None
7	None	0/0/4	None	None

Real authenticated/big V/grassroot

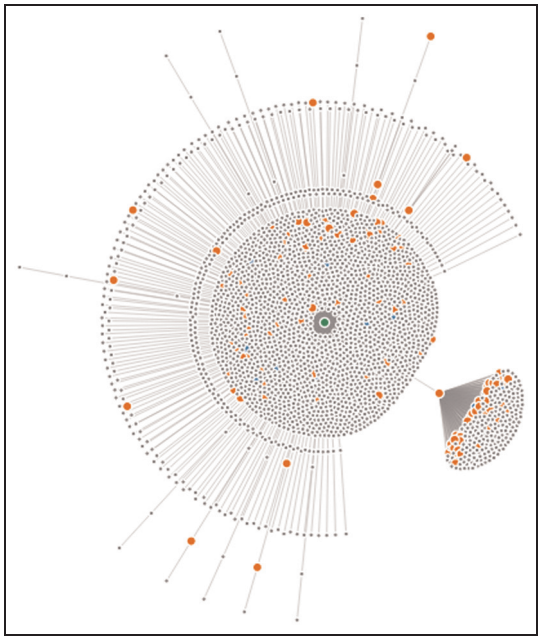

Figure 10. Review analysis released by Han Han.
Right: weibo.com/1191258123/B5Lw6zuPP?from=page_1035051191258123_profile&wvr=6&mod=weibotime

Table 5. Mainstream online data sources.

Search index	Name	
	G	T
Baidu Index of Title	460,640	97,846
360 Index of Title	110,298	27,456
Sougou Index of Title	18,838	2793
Baidu Index of Director	31,938	5736
360 Index of Director	10,045	2078
Sougou Index of Director	1088	176
Baidu Index of Leading Actor	31,938	4343
360 Index of Leading Actor	10,045	1125
Sougou Index of Leading Actor	1088	243
Baidu Index of Leading Actress	32,415	13,706
360 Index of Leading Actress	7197	6286
Sougou Index of Leading Actress	836	2369
Baidu Index of Supporting Actor	4821	4559
360 Index of Supporting Actor	2406	1985
Sougou Index of Supporting Actor	361	429
Baidu Index of Supporting Actress	12,712	3134
360 Index of Supporting Actress	4693	1526
Sougou Index of Supporting Actress	2036	366
Revenue (million yuan)	514.7	886.6
WoM of good from Douban (%)	37.7	70.1

WoM: Word-of-Mouth.

to decay. If the gap between anticipation and WoM is too large, it would be a disaster for prediction job. In 2014, Baidu published revenue prediction of the movie named "The Golden Era" publicly 10 days ahead of screening date based on analysis of its big data technology for the first time, which was declared as over 200 million yuan while the actual value is just 50 million. For another example, there were two competitive blockbusters of New Year movie season before 2015, which were "Gone with The Bullets" (short for G) and "The Taking of Tiger Mountain" (short for T). And their WoM went two extremes. The winner is the one with better reputation but much lower indexes on releasing date. Their contrast chart is shown below. We could summarize that it is difficult to predict the performance of a movie purely depending on the online indexes before screening date (Table 5).

Segmental type

1. *Genre*. Genre summarizes the theme of a movie. The input and output of a film vary drastically according to the difference of genre, which means it could be applied to identify the potential target viewers. Plus, the appetite of market for genre is changing with circumstances. Therefore, it is necessary to classify movies in accordance with genre. But the classification of a movie is not the same in different video websites. For instance, "The Continent," which is mentioned above, is labeled as drama by Youku, while is also marked as literary love film by m1905. We might try to make prediction, but it is questionable that which benchmark we should adopt.
2. *Schedule*. Superficially, schedule, especially holiday, is a potential promotion for the overall box

Table 6. Nine cases of two rivals.

G's WoM (known)	T's WoM (unknown)	G's revenue	T's revenue
Good	Good	Decreased	Decreased
Good	Medium	Little decreased	Decreased
Good	Bad	Not decreased	Seriously decreased
Medium	Good	Decreased	Increased
Medium	Medium	Decreased	Decreased
Medium	Bad	Not decreased	Decreased
Bad	Good	Seriously decreased	Largely increased
Bad	Medium	Decreased	Increased
Bad	Bad	Decreased	Decreased

WoM: Word-of-Mouth.

office during specific period. However, this rule is uncertain actually because the same schedule would lead to fierce competition.

Taking the example of the National Day Holiday in 2014 as an example, the overall revenue of this week is 1.08 billion, while the weekly average value of this year is about 0.6 billion. The film titled "Breakup Buddies" occupied 58% of the proportion. The rest opponents would be inevitably depressed owing to this box-office monster. Therefore, we deem that the factor of schedule should not be analyzed individually. It is more advisable to analyze this element with a combination of another one, which is discussed following as the competing product.

3. *Rival*. There would be several movies on screen at the same time. As far as one film is concerned, some opponents have been shown for days before its release, while other competitors would participate in competition after its release. Hence, all these rivals should be taken into consideration when predicting. According to experts in our team, blockbusters' rivals are not low-cost films, but still blockbusters. We take the movie "The Taking of Tiger Mountain" (short for "T") mentioned above as an example, which is categorized into high level according to the investment. From the perspective of film row piece rate, there are four powerful rivals when it is released on the date of 24 December 2014, which belongs to the New Year movie season. They are "Fleet of Time" ("F" for short), "Women Who Know How to Flirt Are the Luckiest" ("W" for short), "Meet Miss Anxiety" ("M" for short), and "G." The row piece rates are shown in Figure 11. The horizontal axis represents the number of days, and the vertical axis represents the percentage of row piece.

The graph above contains limited amount of information, except that these rivals need to make room for

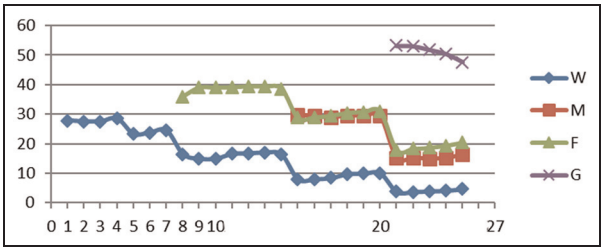

Figure 11. The row piece rates of four competitors.

the new movie "T." However, if we take a look at the box-office revenue trend chart, it will come to a different conclusion. It is manifested that movies whose names are "W," "M," and "F," respectively, had already waned nearly to the close before "T" was shown, and "G" was the only rival to T. We can also deduce from the chart that "G" had started to decline rapidly. No wonder that Lorenza Munoz had said most of box-office revenue was determined at the very early stage after release.[49] Having only one opponent ("G") seems to make the situation simpler to "T." Actually, as far as only WoM is concerned, there are at least nine cases for prediction, as shown in Table 6. Notice that the release time of "G" is earlier than "T" (Figure 12).

Even though the WoM of "G" has been known when "T" is predicted, there are still three cases to decide and neither could we obtain the WoM of "T" in advance. To achieve the object of accurate prediction, it is necessary to quantify the extent of impact of "G", which is very difficult clearly. Moreover, this circumstance is a simple example. It is probable that there are no less than four competitors during Spring Festival. Therefore, the number of cases will grow exponentially. The worst situation is that some future competitors would join in this contest, which would cause serious interference to the predicted target. Not to mention the existence of black swan event, just as the former box-office champion "Lost in Thailand" of domestic films, which is merely a medium investment movie. Nearly,

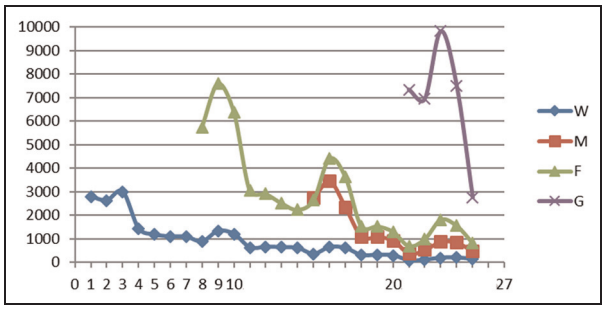

Figure 12. The revenue trend of four competitors.

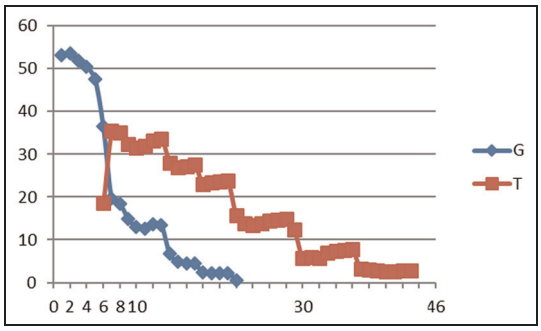

Figure 13. The row piece rate of G and T.

nobody could forecast that it overwhelmed so many blockbusters.

Difficulty of quantifying revenues with external factors

Box-office revenue is the confluence of both internal and external factors. Besides internal factors we analyzed above, there are some vigorous external factors. We aim to analyze them from macro, micro, and artificial perspectives, respectively.

From the macro level, market trend plays an important role like an invisible hand. And this factor is dynamic. Taking domestic action movie as an example, the number of action films is 35 in 2013, comparing with 24 in 2014. There are 15 action films whose box-office revenue is over 100 million yuan in 2013, the same with that in 2014. We can conclude that the number of domestic action films is decreasing, while the proportion of high-revenue films is increasing. It is more advisable to do prediction based on data collected last year rather than the year before last.

Ditto, accidental incident is the force that cannot be neglected during the prediction process based on the micro angle. For example, sudden disaster, star scandal, even state propaganda had successfully exerted influence on revenues in history. The list of this case is long. It is worthy of studying that how to make use of these incidents to hype rather than to sap in marketing. Obviously, this factor is also difficult to quantify in advance of screening.

When it comes to artificial factor, we must analyze a special position in China, the cinema manager, who is powerful enough to affect box-office revenue through adjusting row piece volume. It would be decreased to stop loss with a small attendance and increased to raise income with a large attendance by this position. We still take "G" and "T" as an example. Their change curve chart of row piece rate is shown in Figure 13. The row piece rate of G experienced a rapid decline from the sixth day after releasing. It was only on show for 22 days, while "T" had been on screen for at least about 40 days. This could be all attributed to cinema manager. They kept on changing their strategy according to market reaction. Furthermore, the number of screening days is uncertain to each film in different theater chains because of the same reason. Besides, State Administration of Radio, Film and Television (SARFT) has reigned supreme in China. These artificial factors are exactly what made revenue prediction extremely difficult to be precise, to say nothing of box-office cheating.

Correlation analysis with trailer

Revenue is a tussled product of multiple factors, some of which could be partially quantified, while others are out of control. The most realistic approach is making correlation analysis in a fixed period of time based on reasonable and representative online data. We mentioned three signal sources above, which are search engines, social networks, and video websites. It is known to all that Sina Weibo is the sole survivor, and Tencent Weibo has already been withdrawn from the market in China. Most data from WeChat are not open but private, so the value of these social networks is limited. Data extracted from search engines are correlated with attention rate but have little correlation with plots of movies actually. We have discussed above the huge influence of WoM after releasing, which is closely linked with film content rather than promotion in advance. We have been speculating which element could combine the advantage of both attention rate and movie content. Enlightened by the explosion of video websites, we finally find the exact representative, trailer. Although movie trailers are pervasively consumed via video websites, limited research has been done on the relationship between them and their direct impact on box-office revenue in China. Nevertheless, its commercial effects had been proved by market scientists.[50] We can collect the data of play, support, trample, collection, and so on, to study.

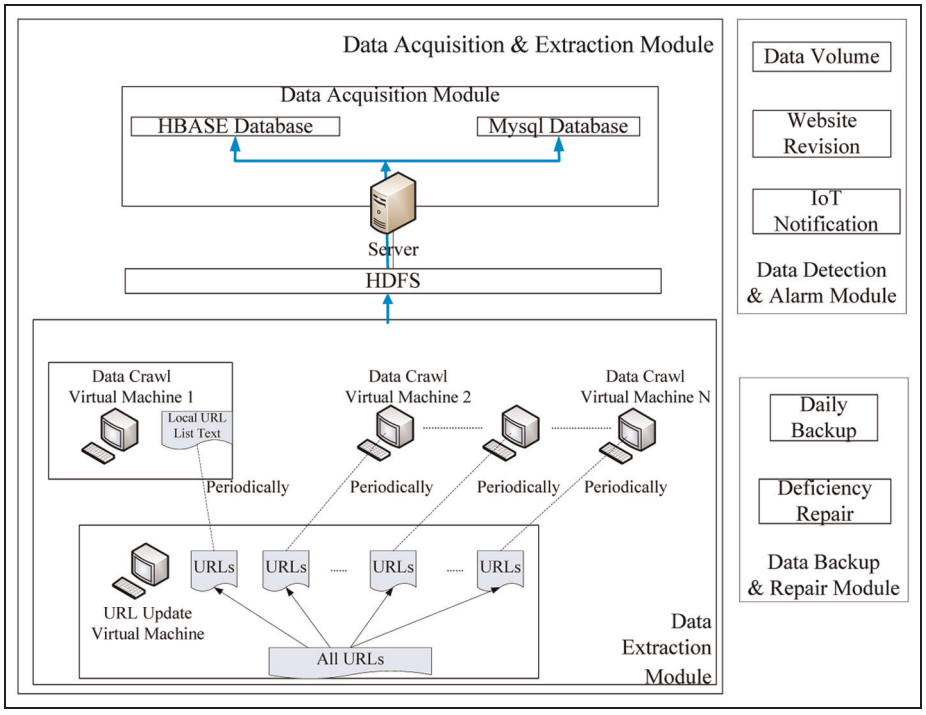

Figure 14. Big data analysis system architecture.

System design and data acquisition

We have built an entire integrated big data analysis system to collect, analyze, and display data gathered from several top video websites based on traditional and distributed databases, which are MySQL and Hadoop, respectively. Seven mainstream domestic video websites were planned to be included as our targets, which are Youku, iQiyi, Sohu Video, Tencent Video, LeTV, Thunder Kankan, and 56.com. Unexpectedly, there existed amount of uncontrollable abnormal values retrieved from Sohu Video and 56.com and the two sites were omitted consequently. The system is comprised of three main modules, which are data acquisition and extraction module, data detection and alarm module, and data backup and repair module (Figure 14).

Data acquisition and extraction module is responsible for collecting primary data from all video websites and extracting useful data to store in HBase and MySQL database. Data detection and alarm module is responsible for detecting whether daily data volume is beyond capacity standard and when the target websites are revised. We introduce the IoT technology solution here to inform the administrator for notification since errors would cause serious consequences to the prediction result. We develop an application that is connected to the wristband. If an error occurs, the server would email to the administrator and call the app on the mobile phone simultaneously. Then, the app will trigger the wristband and the mobile phone to vibrate at the same time. We also adopt the method of vibration frequency to distinguish the error type. A vibration indicates the capacity error and two equals to revision error. The administrator would take different approaches to restore different faults (Figure 15).

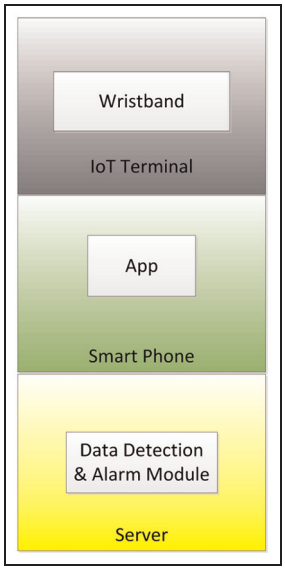

Figure 15. Infrastructure chart of data detection and alarm system.

Data backup and repair module is responsible for backing up daily data as .bak tables and comparing the previous day's data to locate deficient entries, which will be repaired by applicable established rules.

Table 7. Correlation coefficients of data retrieved from five mainstream video websites and m1905.

Element	fr	or	fr	or	fr	or	fr	or	fr	or
aW	0.621	0.603	0.487	0.687	0.291	0.178	0.584	0.665	*0.797*	*0.728*
dW	0.592	0.691	0.374	0.544	0.311	0.276	0.429	0.433	0.696	0.663
wW	***0.861***	***0.842***	0.529	*0.707*	0.321	0.255	0.654	*0.747*	*0.753*	*0.749*
aC	0.147	0.095	0.230	0.331	0.456	*0.726*	0.422	0.482	*0.737*	0.613
dC	0.180	0.182	0.311	0.448	0.436	*0.724*	0.101	0.086	0.502	0.411
wC	0.188	0.195	0.290	0.463	0.481	*0.732*	0.042	0.023	0.608	0.523
aT	0.615	0.578	0.298	0.432	0.183	0.075	0.302	0.237	0.667	0.556
dT	0.622	0.590	0.417	0.645	0.537	0.355	0.337	0.263	0.436	0.316
wT	*0.710*	0.657	0.421	0.672	0.236	0.112	0.419	0.314	0.456	0.352
aS	0.360	0.323	−0.008	−0.010	0.409	0.243	0.328	0.277	0.089	0.088
dS	0.023	0.021	0.283	0.480	0.432	0.261	0.343	0.292	0.262	0.213
wS	0.611	0.523	0.127	0.233	0.473	0.296	0.293	0.217	0.227	0.210
Instruction	Youku, $N = 56$		Tencent, $N = 64$		LeTV, $N = 43$		iQiyi, $N = 55$		Thunder, $N = 81$	

Table 8. Elements of model.

Element	aW	dW	wW	aC	dC	wC	aT	dT	wT	aS	dS	wS	or	fr
Variable	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	y_1	y_2

Descriptive data

Table 7 provides correlation coefficients calculated by formula (1) between first week revenues (short for fr), overall revenues (short for or) gathered from m1905, and five groups of counts retrieved from Youku, Tencent Video, LeTV, iQiyi, and Thunder Kankan in three time periods, which are average watching counts (short for aW), 1 day before releasing watching counts (short for dW), 1 week before releasing watching counts (short for wW); average commenting counts (short for aC), 1 day before releasing commenting counts (short for dC), 1 week before releasing commenting counts (short for wC); average topping counts (short for aT), 1 day before topping counts (short for dT), 1 week before releasing topping counts (short for wT); average stepping counts (short for aS), 1 day before releasing stepping counts (short for dS) and 1 week before releasing stepping counts (short for wS), respectively.

The correlation coefficients between 0.7 and 0.8 are in italic, while values larger than 0.8 are in bold italic. We could conclude the following. First, watching and commenting counts are more correlated with revenue than the other two elements, which are topping and stepping counts. Second, the values of 1 week before releasing are more correlated with box-office revenue than all the values of the other two periods, which are average (from beginning to 1 day before releasing) and 1 day before releasing. Third, watching counts of 1 week before releasing on Youku are more correlated with box-office revenue, especially the first week revenue, than all the other elements.

Analysis model

1. *Overall revenue model*. To realize prediction, we set up equation system with elements discussed above, which are listed in Table 8. We make a linear regression model with stepwise method based on R language. We also use mean absolute percentage error (MAPE) as the evaluation criteria

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|\Pr ed_i - Actu_i|}{Actu_i} \times 100\% \tag{2}$$

First, we build the model with data retrieved from Thunder Kankan. Through comparing the values of MAPE, the second equation is chosen

$$y_1 = \begin{cases} 3275.821 + 0.245x_3 \\ -48.440 + 0.327x_3 - 190,211.963x_{10} \\ -1995.193 + 0.315x_3 - 181,009.621x_{10} + 219.457x_8 \\ -2507.037 + 0.386x_3 - 171,402.186x_{10} + 210.966x_8 \\ -2200.752 + 0.314x_3 - 146,585.638x_{10} + 177.754x_8 \end{cases} \tag{3}$$

$$MAPE = \begin{cases} 21.86256 \\ 6.96629 \\ 8.77421 \quad (\%) \\ 10.03678 \\ 10.22466 \end{cases} \tag{4}$$

The other four data sources are processed with the same method. The result is as follows

$$y_1 = \begin{cases} -48.440 + 0.327x_3 - 190,211.963x_{10}(Thunder) \\ 4496.257 + 0.052x_3(Tencent\ Video) \\ 8407.542 + 379.089x_6 + 5709.872x_{12}(LeTV) \\ -2874.233 + 0.353x_3(Youku) \\ 3029.240 + 0.002x_3 + 7.821x_9(iQiyi) \end{cases} \quad (5)$$

The values of MAPE are listed as follows

$$MAPE = \begin{cases} 6.96629 \\ 18.78758 \\ 29.36338\,(\%) \\ 12.84939 \\ 28.67843 \end{cases} \quad (6)$$

Based on values of MAPE, we would apply equations of the first (Thunder Kankan) and the fourth (Youku) to predict the overall revenues of the subsequent films in the following paragraphs.

2. *First week revenue model*. The first week revenue model is also built based on six data sources with the same method

$$y_2 = \begin{cases} -409.72 + 1.045x_1 + 83.996x_9 - 80,440.694x_{10} + 0.326x_3(Thunder) \\ 4534.881 + 0.017x_3(Tencent\ Video) \\ -1458.215 + 4304.121x_3 + 55.886x_5 - 1340.265x_7(LeTV) \\ -138.550 + 0.154x_3(Youku) \\ 1641.068 + 0.001x_3 + 13.655x_9 - 18.036x_7(iQiyi) \end{cases} \quad (7)$$

The values of MAPE are listed as follows

$$MAPE = \begin{cases} 5.31326 \\ 27.59791 \\ 16.48063\,(\%) \\ 7.08527 \\ 17.93491 \end{cases} \quad (8)$$

According to the values of MAPE, the calculation of equations of the first (Thunder Kankan) and the fourth (Youku) would be taken to predict the first week revenue of the following films.

3. *Analytical results*. Trailers of successive films are kept propagating through channels listed above. And the designed big data system has never stopped collecting data from these video websites. To verify availability and accuracy of the prediction model, we adopt equations (9) and (10) extracted from equations (5) and (7) to predict the overall and the first week revenue of the films released during the following month to the latest samples in the modeling data sets, whose trailers are broadcasted by Thunder Kankan and Youku

$$y_1 = \begin{cases} -48.440 + 0.327x_3 - 190,211.963x_{10}(Thunder) \\ -2874.233 + 0.353x_3(Youku) \end{cases} \quad (9)$$

$$y_2 = \begin{cases} -409.72 + 1.045x_1 + 83.996x_9 - 80,440.694x_{10} + 0.326x_3(Thunder) \\ -138.550 + 0.154x_3(Youku) \end{cases} \quad (10)$$

Table 9. MAPE of overall and first week revenues.

Values of website	or	fr
MAPE of Thunder (%)	10.3043	4.0506
MAPE of Youku (%) with negative results	35.7641	2.7744
MAPE of Youku (%) without negative results	1.2042	1.4259

Instruction: number of samples from Youku is 14 and number of samples from Thunder is 9.

MAPE: mean absolute percentage error.

Based on results listed in Table 9, we could conclude that compared with other video websites, audience watching behavior of trailers on Youku could be the barometer of film's market performance in the period of 1 week before releasing, which is the same with the third point in last section.

Conclusion

In this article, we discuss the difficulty of predicting accurate box-office receipts. All the elements which exert influence on revenues could be split into two types: internal factors and external factors. We also divide internal factors into three classes which are invalid factors, indirect quantifiable factors, and segmental factors. It is impossible for us to retrieve offline data, and open online data are the only channels which could be counted on. However, the inborn nature of data deficiency, distortion even pollution will have a great impact on the prediction job, much less those external factors which are totally beyond control. These elements get tangled with each other to take effect on revenues. Therefore, prediction algorithm is difficult to select or design. Since input is inaccurate and processing method is supposed to change with circumstances, the output is absolutely never easy to be precise. We design Weibo acquisition and analysis system to depict

the retweet and review structure diagrams clearly. The power of influential users is also displayed vividly.

In this complicated situation, relativity analysis is more practicable. Since the efficiency of WoM plays a more important role after releasing, it is more advisable for us to select an appropriate reference. Hence, trailer is the best choice which combines the virtue of timing and content. Through constructing big data collection and analysis system integrated with the IoT technology, it is proved that watching counts of 1 week before releasing on Youku could be the indicator of market performance of a film, especially the first week revenue, based on both correlation analysis and linear regression with stepwise method.

Future work can be continued in several directions. First, we plan to study those factors contributing to receipts which cannot be solved by linear regression. Second, comment content on video websites is an interesting predictive source and natural language processing would be applied as an analysis tool.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work is supported by "the Fundamental Research Funds for the Central Universities" (grant no. 2014PTB-00-02) and the National Science Fund for Distinguished Young Scholars (grant no. 61425012) in China.

References

1. Qi L, Xu X, Zhang X, et al. Structural balance theory-based e-commerce recommendation over big rating data. *IEEE T Big Data*. Epub ahead of print 16 September 2016. DOI: 10.1109/TBDATA.2016.2602849.
2. Qi L, Dou W, Hu C, et al. A context-aware service evaluation approach over big data for cloud applications. *IEEE T Cloud Comput*. Epub ahead of print 23 December 2015. DOI: 10.1109/TCC.2015.2511764.
3. Qi L, Dou W and Chen J. Weighted PCA-based service selection method for multimedia services in cloud environment. *Computing* 2016; 98(1): 195–214.
4. King T. Does film criticism affect box office earnings? Evidence from movies released in the U.S. in 2003. *J Cult Econ* 2007; 31: 171–186.
5. Zhang Y and Zhang X. Analysis of factors that influence income of movies. *Economic Forum* 2009; 4: 130–132.
6. Houston MB and Walsh G. Determinants of motion picture box office and profitability: an interrelationship approach. *Rev Manag Sci* 2007; 1: 65–92.
7. He P. Analysis of factors that affect the box office income of a film. *Chin Film Market* 2011; 11: 8–10.
8. Hu X, Li B and Wu Z. The analysis of the factors which influence film box office. *J Commun Univ China Sci Technol* 2013; 20(1): 62–67.
9. Wasserman M, Mukherjee S, Scott K, et al. Correlations between user voting data, budget, and box office for films in the Internet Movie Database. *J Assoc Inform Sci Technol* 2015; 66: 858–868.
10. Eliashberg J, Hui SK and John Zhang Z. Assessing box office performance using movie scripts: a kernel-based approach. *IEEE T Knowl Data En* 2014; 26: 2639–2648.
11. Krauss J, Nann S, Simon D, et al. Predicting movie success and academy awards through sentiment and social network analysis. In: *Proceedings of the European conference on information systems (ECIS 2008)*, 1 January 2008, p.116.
12. Goel S, Hofman JM, Lahaie S, et al. Predicting consumer behavior with web search. *P Natl Acad Sci USA* 2010; 104: 17486–17490.
13. Oghina A, Mathias B, Manos T, et al. Predicting IMDB movie ratings using social media. In: *Proceedings of the 34th European conference on advances in information retrieval (ECIR'12)*, Barcelona, 1–5 April 2012, pp.503–507. New York: ACM.
14. Google Whitepaper. Quantifying movie magic with Google search. *Industry Perspective + User Insights*, June 2013, http://ssl.gstatic.com/think/docs/quantifying-movie-magic_research-studies.pdf
15. Yao R and Chen J. Predicting movie sales revenue using online reviews. In: *Proceedings of the 2013 IEEE international conference on granular computing (GrC)*, Beijing, China, 13–15 December 2013. New York: IEEE.
16. Mestyan M, Yasseri T and Kertesz J. Early prediction of movie box office success based on Wikipedia activity big data. *PLoS ONE* 2013; 8(8): e71226.
17. Moon S, Bae S and Kim S. Predicting the near-weekend ticket sales using web-based external factors and box-office data. In: *Proceedings of the IEEE/WIC/ACM international joint conference on web intelligence and intelligent agent technologies*, Warsaw, 11–14 August 2014, pp.312–318. New York: IEEE.
18. Thigale S, Prasad T, Makhijia UK, et al. Prediction of box office success of movies using hype analysis of Twitter data. *Int J Invent Eng Sci* 2014; 3(1): 1–6.
19. Du J, Xu H and Huang X. Box office prediction based on microblog. *Expert Syst Appl* 2014; 41: 1680–1689.
20. Kim T, Hong J and Kang P. Box office forecasting using machine learning algorithms based on SNS data. *Int J Forecasting* 2015; 31: 364–390.
21. Bhave A, Kulkarni H, Biramane V, et al. Role of different factors in predicting movie success. In: *Proceedings of the 2015 international conference on pervasive computing*, Pune, India, 8–10 January 2015. New York: IEEE.
22. Lash MT and Zhao K. Early predictions of movie success: the who, what, and when of profitability. *Lect Notes Comput Sc* 2016; arXiv:1506.05382v2 [cs.AI]: 345–349.
23. Xia Z, Wang X, Zhang L, et al. A privacy-preserving and copy-deterrence content-based image retrieval scheme in cloud computing. *IEEE T Inf Foren Sec* 2016; 11: 2594–2608.

24. Fu Z, Wu X, Guan C, et al. Toward efficient multi-keyword fuzzy search over encrypted outsourced data with accuracy improvement. *IEEE T Inf Foren Sec* 2016; 11: 2706–2716.
25. Zhou Z, Wang Y, Jonathan Wu QM, et al. Effective and efficient global context verification for image copy detection. *IEEE T Inf Foren Sec* 2016; 12: 48–63.
26. Li J, Li X, Yang B, et al. Segmentation-based image copy-move forgery detection scheme. *IEEE T Inf Foren Sec* 2015; 10(3): 507–518.
27. Xia Z, Wang X, Sun X, et al. A secure and dynamic multi-keyword ranked search scheme over encrypted cloud data. *IEEE T Parall Distr* 2016; 27(2): 340–352.
28. Fu Z, Ren K, Shu J, et al. Enabling personalized search over encrypted outsourced data with efficiency improvement. *IEEE T Parall Distr* 2016; 27: 2546–2559.
29. Pan Z, Lei J, Zhang Y, et al. Fast motion estimation based on content property for low-complexity H.265/HEVC encoder. *IEEE T Broadcast* 2016; 62: 675–684.
30. Pan Z, Zhang Y and Kwong S. Efficient motion and disparity estimation optimization for low complexity multiview video coding. *IEEE T Broadcast* 2015; 61(2): 166–176.
31. Gu B and Sheng VS. A robust regularization path algorithm for ν-support vector classification. *IEEE Trans Neural Netw Learn Syst*. Epub ahead of print 24 February 2016. DOI: 10.1109/TNNLS.2016.2527796.
32. Gu B, Sun X and Sheng VS. Structural minimax probability machine. *IEEE Trans Neural Netw Learn Syst*. Epub ahead of print 14 April 2016. DOI: 10.1109/TNNLS.2016.2544779.
33. Gu B, Sheng VS, Yeow Tay K, et al. Incremental support vector learning for ordinal regression. *IEEE Trans Neural Netw Learn Syst* 2015; 26(7): 1403–1416.
34. Yuan C, Sun X and Lv R. Fingerprint liveness detection based on multi-scale LPQ and PCA. *China Commun* 2016; 13(7): 60–65.
35. Zhang Y, Sun X and Wang B. Efficient algorithm for k-barrier coverage based on integer linear programming. *China Commun* 2016; 13(7): 16–23.
36. Xia Z, Wang X, Sun X, et al. Steganalysis of LSB matching using differences between nonadjacent pixels. *Multimed Tools Appl* 2016; 75(4): 1947–1962.
37. Shen J, Tan H, Wang J, et al. A novel routing protocol providing good transmission reliability in underwater sensor networks. *J Internet Technol* 2015; 16(1): 171–178.
38. Fu Z, Sun X, Liu Q, et al. Achieving efficient cloud search services: multi-keyword ranked search over encrypted cloud data supporting parallel computing. *IEICE T Commun* 2015; 98(1): 190–200.
39. Ren Y, Shen J, Wang J, et al. Mutual verifiable provable data auditing in public cloud storage. *J Internet Technol* 2015; 16(2): 317–323.
40. Gu B, Sheng VS, Wang Z, et al. Incremental learning for ν-Support Vector Regression. *Neural Networks* 2015; 67: 140–150.
41. Zheng Y, Jeon B, Xu D, et al. Image segmentation by generalized hierarchical fuzzy C-means algorithm. *J Intell Fuzzy Syst* 2015; 28(2): 961–973.
42. Chen B, Shu H, Coatrieux G, et al. Color image analysis by quaternion-type moments. *J Math Imaging Vis* 2015; 51(1): 124–144.
43. Ma T, Zhou J, Tang M, et al. Social network and tag sources based augmenting collaborative recommender system. *IEICE T Inf Syst* 2015; 98(4): 902–910.
44. Wen X, Shao L, Xue Y, et al. A rapid learning algorithm for vehicle classification. *Inform Sciences* 2015; 295(1): 395–406.
45. Xia Z, Wang X, Sun X, et al. Steganalysis of least significant bit matching using multi-order differences. *Secur Comm Network* 2014; 7(8): 1283–1291.
46. Guo P, Wang J, Li B, et al. A variable threshold-value authentication architecture for wireless mesh networks. *J Internet Technol* 2014; 15(6): 929–936.
47. Xie S and Wang Y. Construction of tree network with limited delivery latency in homogeneous wireless sensor networks. *Wireless Pers Commun* 2014; 78(1): 231–246.
48. Tang W-H, Yeh M-Y and Lee AJT. Information diffusion among users on Facebook fan pages over time: its impact on movie box office. In: *Proceedings of the 2014 international conference on data science and advanced analytics*, Shanghai, China, 30 October–1 November 2014, pp.340–346. New York: IEEE.
49. Oh S, Ahn J and Baek H. Viewer engagement in movie trailers and box office revenue. In: *Proceedings of the 2015 48th Hawaii international conference on system sciences*, Kauai, HI, 5–8 January 2015. New York: IEEE.
50. Boksem MAS and Smidts A. Brain responses to movie trailers predict individual preferences for movies and their population-wide commercial success. *J Marketing Res* 2015; 52: 482–492.