Why are some websites researched more than others? A review of research into the global top twenty

Mike Thelwall

How to cite this article:


Abstract

The web is central to the work and social lives of a substantial fraction of the world's population, but the role of popular websites may not always be subject to academic scrutiny. This is a concern if social scientists are unable to understand an aspect of users' daily lives because one or more major websites have been ignored. To test whether popular websites may be ignored in academia, this article assesses the volume and citation impact of research mentioning any of twenty major websites. The results are consistent with the user geographic base affecting research interest and citation impact. In addition, site affordances that are useful for research also influence academic interest. Because of the latter factor, however, it is not possible to estimate the extent of academic knowledge about a site from the number of publications that mention it. Nevertheless, the virtual absence of international research about some globally important Chinese and Russian websites is a serious limitation for those seeking to understand reasons for their web success, the markets they serve or the users that spend time on them. The sites investigated were Google, YouTube, Facebook, Baidu, Wikipedia, QQ, Tmall, Taobao, Yahoo, Amazon, Twitter, Sohu, Live, VK, JD, Instagram, Sina, Weibo, Yandex, and 360.

Keywords

Webs; Major websites; Popular websites; Scholarly interest; Academic research; Success reasons; Comparative research; Cybermetrics; Google; YouTube; Facebook; Baidu; Wikipedia; Yahoo; Amazon; Twitter; Review article.

1. Introduction

The web is an important source of information and communication for work, social and personal reasons. Social scientists therefore need to understand how it affects people's lives. For example, an inability to access websites can cause problems in richer nations (Dutton; Reisdorf, 2019), knowledge of usage patterns can help marketing initiatives (Mariani; Di-Felice; Mura, 2016), and investigations of social network sites are important to understand modern political movements (Bolsover; Howard, 2019). Websites offer many different affordances and it is important to investigate the most popular sites and genres so that their individual values can be known. Despite this, there are apparently no academic studies of the extent to which popular websites have been researched, a gap that the current paper addresses.

There have been previous reviews of internet-related research, especially in the early days of the web, finding that no academic discipline dominates it (Hunsinger, 2005; Silver, 2004), and the consequent lack of a systematic approach allows important sites to be ignored. An investigation of (online) social media research 2004-11 found it to be exponentially growing, with most articles originating from the USA (61%) or UK (11%) (Coursaris; Van-
A later study confirmed the exponential growth but, with different methods, found China to be second most productive country (Gupta; Dhawan; Gupta, 2015). There have also been reviews of social science research about individual websites, such as Facebook (Caers; De-Feyter; De-Couck; Stough; Vigna; Du-Bois, 2013), or an aspect of use of a website, such as health information on YouTube (Gabarrón; Fernández-Luque; Armayones; Lau, 2013) but there are not enough of these to be helpful for a systematic comparison of website research.

This article assesses whether globally popular websites have attracted substantial research interest, in terms of published journal articles related to them or mentioning them. The extent of academic interest in a website seems likely to be affected by a range of factors other than its overall popularity, however, and the following are hypothesised to be relevant: Country of use (popularity in countries that do not publish in international journals may lead to less research); Age (newer sites have had less time to be researched); provision of raw data for other purposes (such data increases the chance that a site is exploited for tangential research); novelty (more novel sites are more important to investigate), whether it contains user-generated data to investigate (Chi, 2008) and whether it supports activities within extensively researched areas. The citation impact of research about a website may also be an indicator of the extent to which the site is considered important within academia. Factors that may affect citation rates include country of use (since there are national differences in average citation impact) and website age (since newer websites have had less time for follow-up research about the site to cite earlier research about it).

2. The content of Scopus-indexed research focusing on a top 20 website

This section briefly reviews research into each of the top 20 websites in April 2019 (according to Alexa: Table 1; see the Methods for more details) published in Scopus-indexed journal articles. It focuses on research from a social science perspective that deals with affordances of the site for users, how the site is used, or characteristics of the site that are relevant to uses of it. Each section is mainly based on a search for the name of the site in Scopus-indexed journal article titles to restrict attention to research that has the site as a major focus or component. This section is not a traditional literature review in the sense of critically analysing the research covered. Instead, it attempts to characterise the scope of the papers. This gives insights into what has been researched but not how it was researched or what was discovered.

Table 1. Background information about Alexa.com’s top 20 websites in April 2019.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Site</th>
<th>Birth</th>
<th>Country</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Google</td>
<td>1998</td>
<td>USA</td>
<td>Search, email, scholar, maps, others</td>
</tr>
<tr>
<td>2</td>
<td>YouTube</td>
<td>2005</td>
<td>USA</td>
<td>Video sharing</td>
</tr>
<tr>
<td>3</td>
<td>Facebook</td>
<td>2004</td>
<td>USA</td>
<td>Social network</td>
</tr>
<tr>
<td>4</td>
<td>Baidu</td>
<td>2000</td>
<td>China</td>
<td>Search engine, email, maps, others</td>
</tr>
<tr>
<td>5</td>
<td>Wikipedia</td>
<td>2001</td>
<td>USA</td>
<td>Encyclopaedia</td>
</tr>
<tr>
<td>6</td>
<td>QQ</td>
<td>1999</td>
<td>China</td>
<td>Instant messaging</td>
</tr>
<tr>
<td>7</td>
<td>Tmall</td>
<td>2008</td>
<td>China</td>
<td>Retail (B2C)</td>
</tr>
<tr>
<td>8</td>
<td>Taobao</td>
<td>2003</td>
<td>China</td>
<td>Retail (C2C)</td>
</tr>
<tr>
<td>9</td>
<td>Yahoo</td>
<td>1994</td>
<td>USA</td>
<td>Search, email, others</td>
</tr>
<tr>
<td>10</td>
<td>Amazon</td>
<td>1994</td>
<td>USA</td>
<td>Bookstore, retail (B2C)</td>
</tr>
<tr>
<td>11</td>
<td>Twitter</td>
<td>2006</td>
<td>USA</td>
<td>Microblog messaging</td>
</tr>
<tr>
<td>12</td>
<td>Sohu</td>
<td>1996</td>
<td>China</td>
<td>Search, online gaming, others</td>
</tr>
<tr>
<td>13</td>
<td>Live</td>
<td>2005</td>
<td>USA</td>
<td>Email, calendar, office software, others</td>
</tr>
<tr>
<td>14</td>
<td>VK</td>
<td>2006</td>
<td>Russia</td>
<td>Social network</td>
</tr>
<tr>
<td>15</td>
<td>JD</td>
<td>1998</td>
<td>China</td>
<td>Retail (B2C)</td>
</tr>
<tr>
<td>16</td>
<td>Instagram</td>
<td>2010</td>
<td>USA</td>
<td>Photo sharing, video sharing</td>
</tr>
<tr>
<td>17</td>
<td>Sina.com.cn</td>
<td>1998</td>
<td>China</td>
<td>Search, email, maps, others</td>
</tr>
<tr>
<td>18</td>
<td>Weibo</td>
<td>2009</td>
<td>China</td>
<td>Microblog messaging</td>
</tr>
<tr>
<td>19</td>
<td>Yandex</td>
<td>1997</td>
<td>Russia</td>
<td>Search, email, maps, others</td>
</tr>
<tr>
<td>20</td>
<td>360.cn</td>
<td>2005</td>
<td>China</td>
<td>Internet security, games, others</td>
</tr>
</tbody>
</table>
Google

Google is primarily known as a web search engine. There have been many studies of search engine user queries based on their search engine log files (Jansen; Spink; Saracevic, 2000) but none of these involved Google since it did not share its logs with researchers. The search behaviour of Google users has also been investigated, seeking to understand the search process (Lorigo; Pan; Hembrooke; Joachims; Granka; Gay, 2006). One article also analysed Google search output, seeking evidence of coverage or bias (Vaughan; Thelwall, 2004).

Many early papers also analysed the impact of Google on specific use contexts, such as library services (Norris, 2006), on the findability of online services (Ashmore; Gross, 2006), on specific types of user, such as deaf web users (Smith, 2006), or on specific information needs, including health-related (Pérez-López; Pérez-Roncero, 2006). Some research not focusing on Google searching nevertheless found it to be an important component of information behaviour, as in the example of an investigation into problem solving in a chemistry laboratory (Shultz; Zemke, 2019). Overall, however, since the web has become the default source for many types of information, there are many purposes and contexts for web searching and so it probably can never be fully understood. Nevertheless, most researchers are probably experienced users of Google and so detailed studies may be less important than they were in the early years of the search engine.

Google provides many services other than web search, including email, word processing, spreadsheets, discussion groups, academic search, image search and mapping. Each of these may generate dedicated bodies of social science research but this has not always occurred. For example, in April 2019 Scopus reported 341 journal articles containing the phrase “Google Earth” in their titles, compared to 248 for Google Scholar, 91 for Google Maps, 58 for Google Books, 29 for Google Docs, 13 for Google Drive, 11 for Gmail, and 1 for Google Photos. It seems unlikely that the volume of research reflects the volume of use; Gmail is probably far more widely used than Google Scholar. The case of Google Scholar is interesting because it fits well into the active research area of scientometrics by providing citation data (Bar-Ilan, 2008). Moreover, it provides a convenient source of data for analysis and so can be researched relatively quickly and cheaply, at least compared to standard Google searches. Thus, the volume of research about a service may be influenced by the pre-existence of academics with relevant expertise and the ease with which the service can be researched. In contrast, Gmail may be little researched because email is well understood and Google’s offering is not fundamentally novel.

YouTube

There is a substantial body of research into aspects of YouTube culture, such as an analysis of Black women natural hair vloggers (Neil; Mbilishaka, 2019), factors that helped the Ice Bucket Challenge to spread virally (Kwon, 2019), YouTube commenting (Murthy; Sharma, 2019; Thelwall; Sud; Vis, 2012) and Mexican protest songs (corridos) as an alternative collective memory about traumatic events (Castillo-González; Leetoey, 2019). The specificity of these examples nevertheless suggests that their results would not generalise easily and that YouTube cultures may be too varied to be investigated systematically.

YouTube is also used by social scientists as a source of public evidence about a topic, even though the site itself is not well understood. For example, teams have investigated parents’ video reactions to their children’s autism diagnosis (Lloyd; Osborne; Reed, 2019), family reactions to being affected by the opioid crisis in the US (Johnson; Worth; Brookover, 2019), and squirrel videos to help understand their behaviour (Jagiello; Dyderski; Dylewski, 2019). Many articles are also hybrid, discussing the YouTube aspect of a wider issue, such as the representation of medical aspects of gender transformation (Miller, 2019).

There are many investigations into the quality of health information on YouTube, concerned that patients may find misleading information. Topics covered include disc herniation (Gokcen; Gumussuyu, 2019), prostate cancer (Loeb et al., 2019), premature ejaculation (Kaynak; Kaya; Aykaç, in press) and facelifts (Derakhshan; Lee; Bham; Barbarite; Shaye, 2019). YouTube may be important for health information for users that prefer an explanation from a person to reading (for example) an encyclopaedia article.

The value of YouTube in support of formal education has been recognised in many disciplines. These include dance (Iannone, 2019), science (Thelwall; Mas-Bleda, 2018) and clinical medicine (Rangarajan; Begg; Somani, 2019). Its support for research has also been assessed (Kousha; Thelwall; Abdoli, 2012).

The business, economics, politics or law of YouTube has been studied, including its international cultural adaptation strategies (Mohan; Punathambekar, 2019), the

It seems unlikely that the volume of research reflects the volume of use; Gmail is probably far more widely used than Google Scholar

The specificity of these examples nevertheless suggests that their results would not generalise easily and that YouTube cultures may be too varied to be investigated systematically
effectiveness of vlogger brand endorsements (Munnukka; Maity; Reinikainen; Luoma-aho, 2019; Xiao; Wang; Chan- Olmsted, 2018), social media advertising strategies (Feng; Xie, 2019), political advertisements (Sohal; Kaur, 2018), reactions to Russian political events (Ushkin, 2014) and Russia Today’s YouTube programming strategy (Orttung; Nelson, 2019).

**Facebook**

Facebook has been extensively researched since its academic beginnings. It was preceded by Friendster and MySpace, which were arguably more novel social network sites, but it has become more successful and eclipsed both in popularity. Social science Facebook research seems to typically investigate why and how (Chen; Kuo; Hsieh, 2019) people or organisations (Lam; Au; Chiu, 2019) use the site in many different educational, or social contexts. Other types of use are also investigated, such as politics (Yuan; Feng; Liu, 2019). Thus, Facebook seems to be particularly well researched in academia, at least from a social sciences perspective. This may be because it is frequently used by academics and its communications are relatively open (in comparison to email) and rich (in comparison to Twitter).

**Baidu**

The Baidu website, which is mainly known for its web search engine, is mentioned in only 11 Scopus-indexed social science journal article titles. Five used the Baidu Index tool that reports the volume of Baidu searches for a given keyword to track interest in topics through user searches, such as AIDS in China (Li et al., 2019). Another analysed the construction of pages in Baidu Baike, which is a Chinese version of Wikipedia (Cheng; Dong, 2018). Three papers compared Baidu to Google in terms of results or political background (e.g., Jiang, 2014) and two were about legal issues (e.g., Zhang, 2011). There seems to be no research in Scopus focusing on how people use Baidu searches, and only two articles investigating the processes behind the construction or use of Baidu Baike (Cheng; Dong, 2018; Wang; Sun; Shen; Zhang, 2018).

Presumably there are also many relevant Chinese-language articles that are not in Scopus.

**Wikipedia**

Wikipedia has been discussed for the new way in which it creates and disseminates knowledge (Hartellus, 2010) as well as for issues relating to the creation and maintenance of high quality content (Hara; Shachaf; Hew, 2010), including sources of references (Rodríguez-Mateos; Hernández-Pérez, 2018), and the motivations of its volunteers (Yang; Lai, 2010). The lack of female contributors has been a topic of concern (Ford; Wajcman, 2017). Wikipedia has generated issues for education, as a potential teaching resource (Moy; Locke; Coppola; McNeill, 2010; Vetter; McDowell; Stewart, 2019) as well as for the accuracy of its content as well as for referencing issues in assignments: the importance of primary sources and the credibility of editable information sources (Rand, 2010).

The content of Wikipedia is also sometimes analysed from an accuracy perspective, similarly to YouTube. Medical research, for example, has examined the accuracy of relevant Wikipedia pages, on the assumption that they are likely to be consulted by patients (Leithner et al., 2010; Simpson; Le; Malicka, 2018). WikiProject Medicine within Wikipedia is a related project to ensure the accuracy of medical pages (Heilman et al., 2011). Some articles have also assessed evidence for political or gender bias (Gauthier; Sawchuk, 2017; Stahel, 2018), or discussed potential coverage gaps (Luyt, 2018; Thelwall; Sud, 2018).

In addition to the potential educational and health information values of Wikipedia, it has also been discussed to understand its role as an open access intermediary, conveying scientific information to the public (Minguillón; Lerga; Albar; Lladós-Maslorenes; Meseguer-Artola, 2017; Teplišský; Lu; Duede, 2017).

Wikipedia is also extensively used as a knowledge resource for natural language processing applications (Azad; Deepak, 2019; Ray; Singh; Joshi, 2010), which is peripheral to user issues, and is sometimes assessed as a source of citation impact evidence (Kousha; Thelwall, 2017; Pooladian; Borrego, 2017; Thelwall, 2016).

A few studies have analysed aspects of Wikipedia use. One analysed temporal trends in searching, showing that external events trigger increases in relevant search volumes (Geiß; Leidecker; Roessing, 2016; Segev; Sharon, 2017). Another compared male and female student use of the site in Spain, finding that both genders used it equally, but males were less critical of its content (Obregón-Sierra; González-Fernández, 2019). Student perceptions of the value of Wikipedia for education have also been surveyed on a small scale (Blikstad-Balas, 2016; Cummings; DiLauro, 2017). The perspective of users outside of education concerning Wikipedia has largely been ignored, however, except in terms of news-related searches.
QQ
Two articles have investigated use contexts for Tencent QQ instant messaging: factors associating with knowledge sharing in groups (Yuan; Liu, 2017) and reasons why students may use it to support their learning (Ma; Au, 2014). The marketing effectiveness of the site and its business model viability have also been investigated (Huang; Kim; Kim, 2013), as have users’ privacy concerns (Meng; Zuo, 2008). Whilst this is a small set of articles, they are insightful for the user perspective.

Tmall
Only three Scopus-indexed journal articles mention Tmall in their titles, none of which focus on the site itself.

Taobao
The e-commerce site Taobao is the subject of a small number (38 Scopus title search matches) of insightful Scopus-indexed journal articles that investigate aspects of its use. For example, one paper analyses the Taobao villages phenomenon, which involves villages reorganising their production and distribution to make commodities for sale through the site (Qi; Zheng; Guo, 2019). Another investigates factors affecting consumer reluctance to trust the site (Han; Kim, 2017). Other articles also exploit data from the site to investigate economic models or theories, such as returns policies that optimise profits (Zhou; Hinz, 2016).

Yahoo
The role of the news services offered by Yahoo has been analysed from a business perspective for insights into its market niche (Li, 2017) or business model (Rindova; Yeow; Martins; Faraj, 2012), and aspects of its internal business processes have also been investigated (Matsuo, 2015; Pathak; Bathini; Kandathil, 2015). Early research investigated the topics that users searched for (Segev; Ahituv, 2010) and the directory system used by Yahoo to organise web pages (Callery; Proulx, 1997). The accuracy of the answers given on Yahoo Answers has been assessed for medical topics (Ohigashi et al., 2017), as discussed above for YouTube and Wikipedia. In contrast, music-related answers have been investigated from a music behaviour perspective, focusing on the types of questions asked and the types of answers given (Hertzum; Borlund, 2017). Also in contrast, the knowledge present in Yahoo Answers has been analysed from a Wikipedia-like perspective (Shen; Li; Liu; Grant, 2015). Political issues related to Yahoo in China have also been discussed (Stevens; Xie; Peng, 2016).

Yahoo is also used as a data source for natural language processing and related research (Figuerola; Gómez-Pantoja; Neumann, 2019; Olivares; Vivanco; Figuerola, 2018), and has been used for scientometric impact analysis (Zahedi; Shirazi; Dehghani, 2010).

Amazon
The core functions and economic model of Amazon have been topics of investigation. The usefulness of Amazon reviews for consumer purchases has been analysed (Chi, 2008; Lee; Trimi; Yang, 2018), as have aspects of its business model (Hadgkiss; Morris; Paget, 2019; Ritala; Golnam; Wegmann, 2014; Zhu; Liu, 2018) and legal issues (Rühl, 2018). Its recommender system technology is important to the site and has wider commercial value (Smith; Linden, 2017). Other than indirectly through purchase decisions, however, the user perspective has been mostly ignored for Amazon.com. One topic-limited and indirect exception is a study of reactions in Goodreads.com to its takeover by Amazon.com (Albrechtslund, 2017).

The Amazon.com website is also sometimes used as a convenient source of data to test marketing theories, such as the efficacy of convenience pricing online (Chenavaz; Drouard; Escobar; Karoubi, 2018). The Mechanical Turk environment for recruiting cheap labour has been exploited by many academic papers as a tool for research, and its value has also been analysed in a meta-study (Berinsky; Huber; Lenz, 2012).

Twitter
The role of Twitter as a news source has been examined from multiple perspectives, including perceived credibility of sources (Edgerly; Vraga, 2019), cross-cultural information flows (Mao; Menchen-Trevino, 2019) interaction with news channels (Ackland; O’Neill; Park, 2019) and agenda setting for the mainstream media (Quinn; Prendergast; Galvin, 2019). Sports news has also been investigated, for example analysing gender bias in official sports team accounts (Grace; Mueller, 2019). Perhaps also related to news coverage, politics is widely discussed on Twitter and investigated (O’Boyle, 2019), including for evidence of the partisanship and bias of journalists (Lacatus, 2019; Usher; Holcomb; Littman, 2018). Social campaigns, such as the #MeToo movement, have also been investigated (Xiong; Cho; Boatwright, 2019). Various ways of exploiting Twitter in education have been proposed and evaluated (Luo; Xie, 2019; Ober, 2019). Twitter-based marketing campaigns have also been evaluated, including an international comparison of alcoholic drinks (Gupta; Lam; Pettigrew; Tait, 2019).
Twitter has also been used as a convenient source of public opinion information by researchers that are not interested in the site itself. For example, cycling-related tweets reveal information about cyclists’ motivations (Das; Dutta; Medina; Minjares-Kyle; Elgart, 2019), tweets about a walking trail might reveal problems with parts of it (Wilson; Lovelace; Evans, 2019), and tweets about a tobacco product may reveal how users react to it (Malik; Li; Karbasian; Hamari; Johri, 2019).

The usage patterns of groups have been investigated in Twitter, including teachers (Carpenter; Kimmons; Short; Clements; Staples, 2019), politicians (Spierings; Jacobs; Linders, 2019), journalists (Tandoc; Cabañas; Cayabyab, 2019), and non-profit organisations (Dong; Rim, 2019).

Overall, social science Twitter research seems to investigate many different types of user and use. Reflecting the widespread use of Twitter in social science research, there are also methodological articles on effective sampling of tweets (Hino; Fahay, 2019) and biases in Twitter samples (Jiang; Li; Ye, 2019). Twitter research seems therefore to be highly developed. This may be due to the ease of access of tweets through a free public applications programming interface, its widespread use, and, presumably, the familiarity of many Western researchers with it.

**Sohu**

Three Scopus-indexed journal articles mention search and online gaming site Sohu in their titles but none focus on the site.

**Live.com**

No Scopus-indexed journal articles mention search and services portal Live.com in their titles.

**VK**

Research on the Russian VKontakte social network seems to broadly follow Facebook research in topic, but is much rarer and mostly available in Russian. For example, one paper (with only the abstract in English) analysed a Bhuddist VK discussion group (Badmatsyrenov; Skvortsov; Khandarov, 2018) and another analysed discussion of a terrorist attack in Romania on the site (Barabash; Bobryshova; Lepilkina; Karabulatova, 2018), and another investigated self-presentation in the site (Shchekoturov, 2017). VK data has also been used to map HIV networks (Rykov; Koltsova; Meylakhs, 2016) and teenage interests (Polivanova; Smirnov, 2017).

**JD.com**

One journal article focuses on Amazon-like retail site JD.com, analysing the quality of its reviews, and proposing a system to automatically classify this (Liu; Fu; Liu; Sun, in press).

**Instagram**

As relevant to an image-based site, much Instagram research addresses visual aspects through either qualitative or quantitative methods. Self-presentation strategies are a common theme (Yau; Reich, 2019), including the relationship between posts and body image (Baker; Ferszt; Breines, 2019) and special groups, such as female body builders (Marshall; Chamberlain; Hodgetts, 2019) and American skateboarders (Dupont, in press). The role of Instagram in aspects of everyday lives has been addressed through user attitudes towards current and potential future romantic partners (Lee; Choi; Lee; Sung, 2019), coping with cancer (Stage, 2019) and promoting healthy behaviours (Santarossa; Coyne; Lisinski; Woodruff, 2019). User reactions to viewing Instagram posts have also been analysed from the perspective of the emotions generated (De-Vries; Möller; Wieringa; Eigenraam; Hamelink, 2018).

Influential sets of pictures have also been examined, such as one from North Korea (Holiday; Anderson; Lewis; Nielsen, 2019), and dieticians’ food photograms (Inan-Eroglu; Buyuktuncer, 2018).

Marketing is important for Instagram, as for YouTube, because of its novel affordances. Issues covered include the effectiveness of influencers in promoting brands (De-Verim; Hudders, 2019; Konstantopoulou; Rizomyliotis; Konstantoulaki; Badahdah, 2018) and visual marketing strategies used in the site (Laestadius; Wahl; Pokhrel; Cho, 2019). Individual market segments have also been addressed, such as tourism (Kuhzady; Ghasemi, 2019) and television (Martin-Quevedo; Fernández-Gómez; Segado-Boj, 2019).

Overall, Instagram seems to have been extensively researched from various user perspectives as well as for marketing.

**Sina.com.cn**

No Scopus-indexed journal articles mention sina.com.cn and none seem to mention the Sina web company (except in the context of Sina Weibo) in their titles.

**Weibo**

The potential for Weibo to promote healthy behaviours has been investigated, such as through posting fitness information from electronic devices (Dong; Chen; Wang, 2019) and by identifying the factors that help health messages to be shared virally (Liu; Lu; Wang, 2017). It has also been investigated for its ability to facilitate social support for people with long-term illnesses (Han; Li; Qu; Zhu, 2018).
Politics has also been analysed, such as by identifying ideological positions expressed through weibos (Huang; Gui; Sun, 2019), international differences in the framing of news discussions (Bolsover, 2017), public discussion of “Brother Wristwatch” 腕表哥, a corrupt safety official caught out on Weibo from crowdsourced images of him wearing expensive watches (Feng; Wu, 2018), how discontent with some government actions is expressed (Wu, 2018), government censorship of the site (Cairns; Carlson, 2016; Vuori; Paltemaa, 2015), and how it can facilitate political involvement for college students (Wang; Shi, 2018).

Weibo is used for brand marketing and therefore it is important to analyse commercial strategies, including from linguistic (Li; Wu, 2018) comparative (Wen; Clark; Kang; Fine, 2016) perspectives. Other use cases examined include librarianship through a comparison with Twitter (Huang; Chu; Liu; Zheng, 2017), journalist posting strategies (Fu; Lee, 2016), information dissemination by non-government organisations (Zhou; Pan, 2016), rural migrant workers’ posts (Zhang, 2013) and the effectiveness of Weibo for crisis communication (Ngai; Jin, 2016).

Like Twitter, Weibo has been used as a convenient source of public sentiments about topics unrelated to the site, such as green buildings (Liu; Hu, 2019) and tourist opinions about attractions (McCann; Pek, 2018). It has also been used for information about the external environment, such as through reports about air quality and pollution (Graminius; Haider, 2018), and as a natural language processing resource (Ling; Marujo; Dyer; Black; Trancoso, 2016).

As with Wikipedia and Twitter, Weibo has been used as a source of evidence about the impact of academic research and the extent to which science is discussed (Yu; Xu; Xiao; Hemminger; Yang, 2017).

Weibo has also been investigated from a more general perspective to understand how discussion topics evolve in the site, and how this contrasts with Twitter (An; Yu; Lin; Du; Zhou; Li, 2018).

Many of the above studies have analysed aspects of the user experience on Weibo, users have also been surveyed to find out which benefits they believe they get from it (Gan, 2018), how users may be anxious about their posts (Li; Lin, 2016) and privacy and anonymity conditions have been investigated (Chen; Li; Hu; Li, 2016). Overall, therefore, there is a rich and diverse academic literature about Sina Weibo.

Yandex
No Scopus-indexed journal articles mention Russian search engine Yandex in their titles. Like Google, Yandex offers a range of services from email to maps.

360.cn
No Scopus-indexed journal articles mention Chinese internet security and games site 360.cn in their titles. None of the matches of the search for 360 seem to be about the site either. This seems to be a background service that many not be directly mentioned in research that uses it.

Summary
The amount of Scopus-indexed research mentioning the top 20 websites in article titles is highly variable, from extensive coverage of Facebook, Twitter, YouTube and Weibo to no research focusing on Yandex and 360.cn. Part of the difference is likely to be due to language (because Scopus is dominated by English) and first mover (the first popular website of a type should attract the most attention due to novelty) issues. In addition, some websites have opportunities for substantial user content generation, e.g., Facebook, Twitter, YouTube, Weibo but not Google, Yandex, 360.cn. Also, more flexible sites seem to attract more attention than those with a single activity focus, such as Tmall and Taobao.

Whilst sites that are flexible in use and allow user content generation have attracted the most research, typically from the social sciences investigating social issues, usage patterns or commerce, it seems likely that this research is far from comprehensive. For example, whilst there is a large volume of research covering multiple aspects of YouTube, it seems likely issues relevant to YouTube are too complex to be analysed with any degree of completeness. Similarly, it seems likely that academia could investigate only a small fraction of Facebook uses because it is a flexible community site. This is exacerbated for Facebook by its default privacy policy so that user behaviours are typically hidden. Thus, paradoxically, the sites that are most researched may also have the most scope for further research into other aspects of their use.

3. Methods
The above section briefly reviewed research with a major focus on a top 20 website, using the heuristic of mentioning the site in the title. The reminder of the article is a bibliometric analysis of the prevalence and citation impact of articles that mention these websites in their abstract, keywords or title. This wider scope allows an analysis of the influence of the sites on research, a different focus.
The top 20 of the Alexa.com top 500 global websites list in April 2019 was used as the source of popular websites. Alexa.com uses a range of sources of evidence, including browser extensions and website scripts, to estimate the number of visitors to major websites over the previous three months. Alexa.com uses a range of sources of evidence, including browser extensions and website scripts, to estimate the number of visitors to major websites over the previous three months, which it then uses to rank websites based on traffic volume. The exact sources of data and algorithm combining them are commercial secrets but the site has credibility through its owner, Amazon, its long history (since 1996) and prior use in academic research (e.g., Han; Kim, 2017; Lui, 2015; Vaughan, 2012). Nevertheless, its traffic data should be treated as estimates rather than precise values. In addition, it is not clear how it deals with web traffic through smartphone apps rather than web browsers and it does not adequately rank popular internet services that do not primarily operate through their websites, such as WhatsApp. Another limitation for the current study is that the figures do not consider whether a website is relatively new and cannot therefore have had much research published with it.

The Scopus citation index was used to estimate the prevalence and citation impact of international research about the chosen websites. Scopus was used in preference to the Web of Science because it has wider coverage, and particularly for non-English and non-Western sources (Mongeon; Paul-Hus, 2016). Scopus queries were used to identify articles mentioning one of the top 20 websites. Since the websites could be mentioned peripherally in the full text of an article, the queries were restricted to the parts of an article that normally signal its core content: the title, abstract and keywords. The queries were tested and modified when they produced a high proportion of false matches, as in the case of QQ, Live, VK, JD, and Sina (Table 2). There were a few ornithology matches for Twitter (about 2 per year) but this proportion is too low to greatly influence the results. The Russian name of VK (ВКонтакте) produced only a single match and so its English translation was used. The Scopus queries were submitted on 18 April 2019. Documents not within academic journals were subsequently excluded, as were documents from before a website existed, or from after 2018 (for field normalisation purposes, as described below). The results are nevertheless limited by the comprehensiveness of the English-language queries used to identify relevant articles.

Table 2. The Scopus queries submitted on 18 April 2019 to identify articles mentioning one of Alexa's top 20 websites.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Site</th>
<th>Query</th>
<th>Full query</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Google</td>
<td>Google</td>
<td>TITLE-ABS-KEY(Google) AND DOCTYPE(ar)</td>
</tr>
<tr>
<td>2</td>
<td>YouTube</td>
<td>YouTube</td>
<td>TITLE-ABS-KEY(YouTube) AND DOCTYPE(ar)</td>
</tr>
<tr>
<td>3</td>
<td>Facebook</td>
<td>Facebook</td>
<td>TITLE-ABS-KEY(Facebook) AND DOCTYPE(ar)</td>
</tr>
<tr>
<td>4</td>
<td>Baidu</td>
<td>Baidu</td>
<td>TITLE-ABS-KEY(Baidu) AND DOCTYPE(ar)</td>
</tr>
<tr>
<td>5</td>
<td>Wikipedia.org</td>
<td>Wikipedia</td>
<td>TITLE-ABS-KEY(Wikipedia) AND DOCTYPE(ar)</td>
</tr>
<tr>
<td>6</td>
<td>QQ</td>
<td>Tencent QQ</td>
<td>TITLE-ABS-KEY(Tencent QQ) AND DOCTYPE(ar)</td>
</tr>
<tr>
<td>7</td>
<td>Tmall</td>
<td>Tmall</td>
<td>TITLE-ABS-KEY(Tmall) AND DOCTYPE(ar)</td>
</tr>
<tr>
<td>8</td>
<td>Taobao</td>
<td>Taobao</td>
<td>TITLE-ABS-KEY(Taobao) AND DOCTYPE(ar)</td>
</tr>
<tr>
<td>9</td>
<td>Yahoo</td>
<td>Yahoo!</td>
<td>TITLE-ABS-KEY(Yahoo!) AND DOCTYPE(ar)</td>
</tr>
<tr>
<td>10</td>
<td>Amazon</td>
<td>Amazon.com</td>
<td>TITLE-ABS-KEY(Amazon.com) AND DOCTYPE(ar)</td>
</tr>
<tr>
<td>11</td>
<td>Twitter</td>
<td>Twitter</td>
<td>TITLE-ABS-KEY(Twitter) AND DOCTYPE(ar)</td>
</tr>
<tr>
<td>12</td>
<td>Sohu</td>
<td>Sohu</td>
<td>TITLE-ABS-KEY(Sohu) AND DOCTYPE(ar)</td>
</tr>
<tr>
<td>13</td>
<td>Live</td>
<td>live.com</td>
<td>TITLE-ABS-KEY(live.com) AND DOCTYPE(ar)</td>
</tr>
<tr>
<td>14</td>
<td>VK</td>
<td>Vkontakte</td>
<td>TITLE-ABS-KEY(Vkontakte) AND DOCTYPE(ar)</td>
</tr>
<tr>
<td>15</td>
<td>JD</td>
<td>jd.com</td>
<td>TITLE-ABS-KEY(jd.com) AND DOCTYPE(ar)</td>
</tr>
<tr>
<td>16</td>
<td>Instagram</td>
<td>Instagram</td>
<td>TITLE-ABS-KEY(Instagram) AND DOCTYPE(ar)</td>
</tr>
<tr>
<td>18</td>
<td>Weibo</td>
<td>Weibo</td>
<td>TITLE-ABS-KEY(Weibo) AND DOCTYPE(ar)</td>
</tr>
<tr>
<td>19</td>
<td>Yandex.ru</td>
<td>Yandex</td>
<td>TITLE-ABS-KEY(Yandex) AND DOCTYPE(ar)</td>
</tr>
<tr>
<td>20</td>
<td>360.cn</td>
<td>360.cn</td>
<td>TITLE-ABS-KEY(360.cn) AND DOCTYPE(ar)</td>
</tr>
</tbody>
</table>

The Scopus citation count was extracted for each journal article and used to calculate the mean normalised log citation score (MNLCS) (Thelwall, 2017). This is a variant of the original mean normalised citation score (MNCS) (Waltman; Van-Eck; Van-Leeuwen; Visser; Van-Raan, 2011), using logged citation counts to stop small numbers of very highly cited articles unduly influencing the results. The MNLCS has a value of 1 if the articles tended to have the same number of citations as other articles from the same field and year. MNLCS values above 1 indicate more citations than the world average for the field and year of publication, and values below 1 correspond to below world average citation counts for
the publishing field and year. When articles were recorded within multiple Scopus categories, the average (arithmetic mean) of all Scopus categories was used as the denominator.

4. Results

There are vastly different amounts of research relating to the top 20 websites (Figure 1), ranging from 16,362 for Google to 1 each for Sina and 360.cn (both with complex queries). There is not a strong trend for sites with more research to have higher impact research about them, however, with a weak Spearman correlation of only 0.247 between these two.

For websites mentioned by relatively many articles, there is a clear trend for the volume of research to increase over time (Figures 2 and 3). There is evidence of peaking or levelling off for Wikipedia and Weibo, whereas Amazon has three years of decline before increasing again. Whilst research mentioning Google and Facebook is increasing rapidly, despite their age, it is no longer growing approximately exponentially.

Research about major websites usually has above world average citation impact (for its field and year) when the difference is no longer growing approximately exponentially. The two main exceptions are that Yahoo research since 2002 and Wikipedia research since 2012 have had close to world average citation impact.

5. Discussion

Some of the global 20 most popular websites (according to Alexa.com) seem to have almost no journal articles closely related enough to them for a mention in the article title, abstract or keywords. Most of the websites with almost no Scopus-indexed research are from China (Tencent QQ, Taobao, Sohu, JD, Sina, 360.cn) or Russia (VK, Yandex). There is a small amount of research about the Chinese site Baidu, but this is relatively low for its popularity. In contrast, Chinese site Weibo has a substantial amount of research and is therefore an anomaly for the language issue mentioned above. This anomaly is nevertheless in line with the high volume of Twitter-related research since both sites have similar affordances (an applications programming interface allowing automated access to content). Thus, overall, site nationality (Table 1) is enough to explain the low position of most of the sites. A similar issue has been observed within another repository of knowledge, Wikipedia, which has greater coverage of topics relevant to an English-speaking audience despite its international variants (Thelwall; Sud, 2018).

Whilst China is a major contributor to international academic output, its social science research is much less international (Liu; Hu; Tang; Wang, 2015), explaining the lack of Scopus-indexed research about Chinese websites. Scopus has weaker coverage of non-English research, particularly in the Social Sciences and Humanities (Mongeon; Paul-Hus, 2016), exacerbating the problem. China accounted for only 3.5% of Social Sciences articles in Scopus and Russia accounted for 0.9% by April 2019. Websites that are in Chinese or Russian therefore have two obstacles for researchers: relatively few researchers publishing international academic social science journal articles can read them and relatively few may know them well enough to consider them interesting or relevant. The problem may be further exacerbated by the non-English sites being referred to in different ways and being mainly known by their original Chinese or Russian names.

The website live.com is has little research that directly mentions it probably because it hosts services that are better known by other names, such as Microsoft OneDrive, Outlook and Office. This may also affect some of the Chinese websites.

There are vastly different amounts of research relating to the top 20 websites, ranging from 16,362 for Google to 1 each for Sina and 360.cn.
Figure 2. The number of Scopus-indexed journal articles 1996-2018 mentioning the specified website in their abstract, title or keywords (black dots) and the MNLCS of these articles (blue dots with 95% confidence interval error bars). The websites are the most popular 10 according to Alexa.com in April 2019.
Why are some websites researched more than others? A review of research into the global top twenty

Figure 3. The number of Scopus-indexed journal articles 1996-2018 mentioning the specified website in their abstract, title or keywords (black dots) and the MNCS of these articles (blue dots with 95% confidence interval error bars). The websites are the most popular 11-20 according to Alexa.com in April 2019.
Twitter and Weibo have a relatively large amount of research mentioning them. This may be because they can be easily used as a source of evidence (tweets, weibos) in text processing computer science or linguistics research. For example, 52% of Weibo and 46% of Twitter research in Scopus is classified as Computer Science or Mathematics. Each site is more extensively researched in its own country, with US first authors outnumbering mainland Chinese first authors by 6.2 to 1 for Twitter and mainland Chinese first authors outnumbering US first authors by 7.3 to 1 for Weibo.

Another anomaly in the graphs is that YouTube has about a third as much research as Facebook, despite Facebook being less popular. This could be explained by Facebook having been more popular than YouTube in the past. Another possible explanation is Facebook’s origins within academia and, presumably, its academic-friendly affordances. In contrast, YouTube may be thought to be primarily an entertainment site, even though it is used in education (e.g., Barry et al., 2016). The greater flexibility of Facebook to be a site for organising groups and providing different types of information (including videos) is probably also a factor because it provides a wider variety of use types to investigate.

Of the factors considered in the introduction, website age does not seem to be relevant to explain the level of research interest in a top 20 website because all the sites are over ten years old and there is not a trend for newer sites to be discussed less, at least within this list. The novelty of a website is a possible explanatory factor given that the Chinese and Russian websites are slightly younger than similar sites from the USA, but it seems likely that the language or geography issue is more important.

The usefulness of Google’s affordances for academic research methods help to explain the amount of research about it. Almost a third (31%) of papers mentioning Google discuss Google Scholar, typically as a source of data for citation analysis or literature searches. Another 14% discuss Google Maps, typically using it as a tool in computing or geographical information systems research (e.g., “Timely reporting and interactive visualization of animal health and slaughterhouse surveillance data in Switzerland”). These studies give no social science insights into the use of Google. The same is true for some of the other sites. For example, 10% of Amazon research uses its Mechanical Turk service and a few (12) YouTube articles only used a video dataset derived from it, “YouTube Action” or UCF11. Similarly, Facebook is used to recruit survey or interview participants around a focused topic, such as, “The psychosocial impact of ptoasis as a symptom of myasthenia gravis: A qualitative study”, or to track people, “Patients with outdated details were tracked with the national health insurance database and social media (Facebook)” (Kowalewski; Olszewski; Kwiatkowski; Gałązka-Świderek; Cichoń; Paśnik, 2017), although most Facebook research seems to investigate an aspect of the use of the site itself (e.g., “Connected motherhood: Social support for moms and moms-to-be on Facebook”).

Academic knowledge requirements

Although the results show gaps in knowledge about important social web sites, the empirical data cannot prove that the gaps are important. It is impossible for academic research to generate all possible knowledge and so a useful consideration is whether the omissions will be a substantial disadvantage in practice. Site users, including businesses, can get information from other sources, including professional magazines, blogs, white papers, networks, and their own experience or analysis. Thus, a lack of academic knowledge does not imply that a site is being underexploited or poorly used. Apparently relatively straightforward sites, such as those for e-commerce, may not need much research to understand their fundamentals.

Research often serves educational uses so academic studies of websites may be less important when those websites are not relevant to degree courses. This may apply to the standard office services offered by Live.com and within other websites (e.g., email from Google, Yandex, Sina, Baidu). Thus, part of the issue is whether academic research is the most relevant avenue for publishing research about popular websites.

Despite the above consideration, the lack of knowledge is not a guarantee that a topic is unimportant. The case of neglected tropical diseases (Hotz et al., 2007; Mathers; Ezzati; López, 2007) is a well-known example of a knowledge area that is recognised as being under-researched for its societal importance. Thus, the current paper can identify research gaps but qualitative judgements are needed to assess whether the gaps are important.

Citation impact

Research related to top websites seems to usually have above average citation impact, except that there is a (not statistically significant) trend for the Chinese and Russian sites to have below world average citation impact or, in the case of Weibo, lower citation impact than the similar site Twitter. This is probably due to research in these countries having lower overall average citation impact (Elsevier, 2013; Fairclough; Thelwall, 2015) rather than being a characteristic of the studies. Nevertheless, the lower citation impact, and the dominance of Western sites in international research, may serve as a disincentive for researchers seeking international recognition to study Chinese and Russian sites.
6. Conclusions

This study found vastly different levels of academic publishing related to the globally most popular twenty websites (according to Alexa.com). The main reason for differences in the amount of research about a site seems to be the following.

- Less research for non-English websites. Generalising this, there is likely to be even less research about non-English websites serving an area with few social science researchers publishing in English. Such research may be available in local language journals (where it may be most useful and most read) but seems unlikely to be visible to an international audience.
- Less research for websites that host services rather than being primary portals.
- More research about flexible sites that can be adapted for different social purposes, such as organising groups, disseminating information, or communicating with friends.
- More research about websites that provide an easily accessible source of data for computing-related research. Because of this, even if a website is important enough to be mentioned in an article title, abstract or keywords, the article’s findings may reveal nothing about the site.
- More research about websites that provide affordances for research methods. As above, this type of research may not be informative about the site itself.

Thus, the level of academic interest in a website does not accurately reflect its global importance from a social science perspective or the depth of academic knowledge about it.

Despite the limitations of the methods, the results give evidence that several globally important non-US websites are mostly receiving little international research attention. This is a problem for developers seeking to understand the factors behind the success of individual websites, for marketers needing information about the role in the information, entertainment and retail infrastructures of the audiences that they serve, and for social scientists wanting deeper international insights into how users interact with websites for aspects of their daily lives (e.g., health, socialising, media consumption). Thus, there is a need for more international comparative research into popular websites as well as studies that focus on international websites. Even though such research may have lower citation impact, editors and reviewers should be aware of this important gap when evaluating the contribution of such papers.

7. References


Han, Xi; Li, Bei; Qu, Jiabin; Zhu, Qinghua (2018). “Weibo friends with benefits for people live with HIV/AIDS? The implications of Weibo use for enacted social support, perceived social support and health outcomes”. *Social science & medicine*, v. 211, n. 1, pp. 157-163. https://doi.org/10.1016/j.socscimed.2018.06.016


Lee, Kang; Liu, Meiliang; Feng, Yi; Ning, Chuanyi; Ou, Weidong; Sun, Jia; Wei, Wudi; Liang, Hao; Shao, Yiming (2019). “Using Baidu search engine to monitor AIDS epidemics inform for targeted intervention of HIV/AIDS in China”. *Scientific reports*, n. 9, article 320. https://doi.org/10.1038/s41598-018-35685-w


Why are some websites researched more than others? A review of research into the global top twenty


Liu, Jie; Fu, Xiaodong; Liu, Jin; Sun, Yunchuan (in press). “Analyzing and assessing reviews on JD.com”. *Intelligent automation & soft computing*. https://doi.org/10.1080/10798587.2016.1267244


Loeb, Stacy; Sengupta, Shomik; Butaney, Mohit; Macaluso Jr, Joseph N.; Czarniecki, Stefan W.; Robbins, Rebecca; Braithwaite, R. Scott; Gao, Lingshan; Byrne, Nataliya; Walter, Dawn; Langford, Aisha (2019). “Dissemination of misinformative and biased information about prostate cancer on YouTube”. *European urology*, v. 75, n. 4, pp. 564-567. https://doi.org/10.1016/j.euro.2018.10.056


Santarossa, Sara; Coyne, Paige; Lisinski, Carly; Woodruff, Sarah J. (2019). “#fitspo on Instagram: A mixed-methods approach using Netlytic and photo analysis, uncovering the online discussion and author/image characteristics”. Journal of health psychology, v. 24, n. 3, pp. 376-385. https://doi.org/10.1080/01972240903423477


Shen, Haiying; Li, Ze; Liu, Jinwei; Grant, Joseph-Edward (2015). “Knowledge sharing in the online social network of Yahoo! answers and its implications”. IEEE transactions on computers, v. 64, n. 6, pp. 1715-1728. https://doi.org/10.1109/TC.2014.2322598


Thelwall, Mike; Mas-Bleda, Amalia (2018). “YouTube science channel video presenters and comments: Female friendly or vestiges of sexism?”. *Aslib journal of information management*, v. 70, n. 1, pp. 28-46. https://doi.org/10.1017/AJIM-09-2017-0204


