Generation of Big Data from Behaviors of Web-based Foreign Language Learners in China

Chen Xiangyu
School of Foreign Languages
Southeast University
Nanjing, China
xiangyu0509@163.com

Chen Meihua
School of Foreign Languages
Southeast University
Nanjing, China
meihuachen123@163.com

***Abstract*—The enormous amount of published multimedia resources for foreign language learning on the internet in China belong to content data, which do not reflect any information about learners. Current content browsing applications support recording basic learning behaviors, such as login time, online period and post content, but do not record details like mouse movements, mouse wheel scrolls, or GPS locations of learners. These detailed behaviors of learners provide essential information to analyze the learning process. Therefore, the big data technology applied in foreign language learning should focus on gathering and data-mining the detailed behavioral data from learners rather than improving the amount or the high-definition of content data provided for learning. Tin Can API provides a solution to synchronize the behavioral data of learner individuals across different learning platforms.**

Keywords—big data; foreign language learning; Tin Can API

I. INTRODUCTION

For enterprises, data analytics have been being considered asan advantage to the point that those who are not utilizing data analysis are regarded as at grave disadvantage [1]. As the world's second largest economy, China has fully embraced the concept and application of big data in the research fields such as commerce, transportation, medicine and education. Foreign language education in China has also begun to explore the use of big data to study learners' behaviors. Foreign language learning activities are engaging learners nearly every day, which generates substantial amounts of data. These data need to be consciously recorded by teachers that are activity organizers. However, foreign language teachers are not all IT experts. They have encountered difficulties in both differentiating content data from behavioral data and designing activities to gather effective data. When foreign language teachers communicate with web learning application developers, these difficulties would cause misunderstandings as why, what kind of, andhow data should be gathered from learners' behaviors as big data.

II. DATA MINING IN FOREIGN LANGUAGE LEARNING BEHAVIORS

A well-known case of data mining is the story of beer and diapers in Wal-Mart, although its authenticity is not proved [2]. In the story, after an analysis of fathers' habits of purchasing beer and diapers together, the supermarket put the two commodities next to each other, resulting in sale increase for both.

From this story, it can be seen that the target of data mining is behaviors rather than the receivers of behaviors. In this case, it is the purchase rather than the commodities. If the analysis only focused on the sales data of the two separate commodities, the relationship between them would never be discovered.

Fabricated or not, the case of 'beer and diapers' illustrates the meaning of data mining and correlation. So it is safe to coin a case in the area of foreign language learning for illustration. This paper makes up a case called 'English intonation and IP address'. An English teacher in a Chinese

university found at the end of one semester that most boys' English intonation had been significantly improved while girls' did not show such a change. Meanwhile, a campus network administrator found that a group of IP addresses frequently visited a server at evenings through the whole semester. This server was later confirmed as one serving an online team battling game. And these IP addresses were all from the boys in the English class. The English teacher finally discovered that when the boys were playing this game, they imitated some English phrases or short sentences from the roles in the game. They even began to shout these utterances to express feelings such as joy or anger when they were not playing the game. The teacher concluded that this game with authentic English sounds helped improve the boys' intonations. This case showed that it is the game behavior that correlated the two seemingly irrelevant concepts, intonation and IP. And through this behavioral data analysis, a positive factor for improving intonation is pinpointed by the teacher.

Theoretically, correlation analysis could be carried out between any unrelated behavioral data and the data we are interested in to explore the most unlikely factors that might have significant effect on learners' outcomes. Another case could be fabricated to illustrate the correlation between 'GPS locations and spoken English level'. The spoken English scores of the students in one class in a Chinese university and their frequently visited places were input into a correlation analysis. The places were recorded by GPS system in their smartphones and represented as dots on a map. Results showed that there was a densely dotted place by students of higher scores, a pub. The explanation was that a large number of regulars in the pub were oversea students in China and the English learners had more practice of spoken English here with foreigners. In real practice, many hidden factors of foreign language learning still exist. Correlation analysis is an effective way to find or notice these factors.

III. Content Data and Behavioral Data in Foreign Language Learning In China

When the notion of big data was newly introduced into the foreign language learning in China, some may have the incorrect impression that big data should be files that are vast in size and large in number. For example, there are two language teaching videos of the same content, but due to different video resolutions, one is 2 Gigabytes and the other is 500 Megabytes. The former seems to be more like big data. However, the English contents in both videos are the same. If the students do not watch it and learn it, no behavioral data could be generated. Another example: there are two databases of examination papers, one containing 10,000 papers and the other 1,000. Again, the former seems more suitable to be big data. But if the students do not take any tests, no learning behaviors would be recorded. The above two examples are content data of learning resources, which carry no information of learners. Therefore, the size and number of content data should not be used to judge the effectiveness of big data in foreign language learning.

The behavioral data of foreign language learners are generated when the learners are interacting with the content data. Taking English videos for instance, the traditional behavioral data include the number of hits, dates of hits, watching progress, etc. With the advanced technology in big data, more detailed behavioral data of the video viewer could be recorded. For example, during the playback of an English video, a learner might watch a certain part of the video more than one times. This part may only last a few seconds. Then the images, audio and subtitles of this part could be extracted and recorded for analysis to see which elements have attracted the interest of or caused some difficulty to the learner. For another example, while watching a long video, a learner may drag the progress bar to skip some part of the video or close the video before finishing watching it. Then the timestamps for these behaviors could be recorded to establish an attention-time curve of the learner. Also, the corresponding images, subtitles and audios at these timestamps could be analyzed to see which elements have decreased the learner's motivation to continue watching.

Big data technology also affords the recording of keyboard and mouse movements to keep a thorough log of learners' actions. Taking the computer-based English test for instance, it could record more than the traditional information of answers to the questions. It could record the frequency of Backspace to calculate the spelling ability of the learner or to provide a vocabulary list for the learner to make more practice. From a

thorough log of keyboards, the writing ability of the learner would be assessed by both writing result and writing process [3]. In addition to keyboard log, the record of mouse movement could help analyze the learner's reading behavior. Some learners move the mouse over the words they read. Therefore, through the speed and position of mouse movement the difficult and easy words or sentence structures for the learner could be guessed.

IV. Generation of Detailed Behavioral Data in Foreign Language Learning

Some traditionally recorded data of learners are not effective behavioral data. One crucial problem of recording learners' behaviors is that many learning activities happen but the details are not recorded. The traditional data of learners' behaviors include the time stamps, time period, times and scores of learning, but the details within the learning process are missing. For example, a piece of English news is published on a webpage for the learners to read. It is easy to count how many times this webpage is visited and to log the time stamp when the learners visit it. However, it is hard to know how learners understand the news content and how they perceive the language details. Actually, the hit times and timestamps are information about the content data, the news, rather than the behavioral data of readers. Even some comments on certain language points made by English teachers are only a prediction of learners' behaviors, which could not reflect learners' true responses to the news. To record detailed learners' behaviors, the following techniques are needed.

Plug-in programs that record learners' detailed behaviors could be added into multimedia browsing software. Text and video are the two main media types that are used in foreign language learning resources in China. When learners read texts, mouse events like cursor movement, click, selection and mouse wheel scrolls are triggered. Learners could also hover the cursor on a certain word to show its translation by some screen word capturing dictionaries. The above mentioned mouse events or user behaviors all have the corresponding pixel coordinates on the screen, at which words or sentences could be recorded. When learners watch videos, they often drag the progress bar, pause, resume, turn to full screen, adjust the volume, etc. The time stamps of these learner actions reflect the time points when the learners' cognitive status of the images, audio or subtitles may have changed. The recording of the detailed reactions of texts and video could help analyze the thorough learning process. Modern technologies even make it possible to record the eye movement to track the text or other visual elements that the learner perceive on the screen.

Dividing the learning resources into smaller learning modules can help record detailed learner behaviors. Learning contents with larger size easily reduce learners' motivation. For example, when reading an English passage of 1,000 words on a web page, a learner could show a high interest and keep a strong motivation for the first 400 words. During the first 400 words, the actions of cursor movements and screen word capturing might be frequent. However, for the rest 600 words, the learner's interest may drop dramatically, with the decrease of numbers of screen word capturing and the increase of skimming indicated by more mouse wheel scrolls. Therefore, the valid gathering of behavioral data is concentrated in the first 400 words. The analysis of the behavioral data from the rest 600 words would be a waste of computing resource and a meaningless raise of analysis complexity. If the content data of learning resources could be divided into smaller learning modules, like 300-word text in the above example, the distribution of valid behaviors would be narrowed down. So the effectiveness of data analysis can be improved. What is more, when the learning module size is reduced, the learners could use more time fragments to practice the language [4].

V. Communication of Behavioral Data and Tin Can API

Communication and synchronization of behavioral data are essential because the data for one learner may be generated from various learning platforms and distributed on different storage systems. In the concept of big data there are two kinds of data, object data and personal data, which are equally significant. Object data refers to the information gathered from a large number of its users for one object, like all readers' attitudes toward a particular book. Personal data refers to the information gathered from any sources related to this individual, like all the living and working information of one person. Personal data can be used to provide customized services for individuals, which is the real value of this big data

era. And the key technology of processing personal data lies in the communication and synchronization of behavioral data.

An individual language learner's learning behaviors, which fall into the category of personal data, need data communication and synchronization to integrate the recorded information by various platforms. For example, learner *S* began practicing English listening on Website *A* in the year 2012. Website *A* was later closed in 2014. Then *S* turned to Website *B* to continue English listening practice. And from 2013, *S* began to practice spoken English on a smartphone APP named *C*. The service providers *A*, *B* and *C* all stored the learning behavioral data of *S*. If the listening history of *S* needs to be analyzed, the problem is that the data from year 2012 to 2014 was lost on Website *A*. When *S* changed from *A* to *B*, the data could not be migrated from *A* to *B*, which brought the problem of data synchronization. If *S*'s comprehensive abilities of English listening and speaking need to be assessed, the data from *A*, *B* and *S* should all be integrated, which brought the problem of data communication. These situations call for an industrial standard of synchronization and communication of behavioral data.

Tin Can API uses unified data format to gather learners' behavioral data, on line or off-line, formal or informal [5]. All the learning activity providers that have implemented Tin Can API would use a simple syntax to communicate, thus synchronizing individuals' learning data. Here again takes the English learner *S* as an example to illustrate how Tin Can API works. In this example, *S* uses his email address as the universally unique ID. Website *B* and smartphone APP *C* both support Tin Can API. After *S* listened an English dialogue on *B*, *B* would generate a statement like this: *{S(email address) listened dialog(dialog URL)}* and send this statement to *B*'s Learning Record Store(LRS). Then *S* watched an English video on APP *C*, when *C* would generate a statement like this: *{S(email address) watched video(video URL)}* and send it to *C*'s LRS. When *S* wanted to check his own learning record, LRS-B and LRS-C would communicate and provide this cross-platform information to the user. Tin Can API provides an overall and effective solution for the communication and synchronization of behavioral data in the big data time.

VI. CONCLUSION

The large amount of learning resources belong to content data, from which no valuable information of foreign language learners could be found. The behavioral data of learners come from the details recorded during the learning process, which is the major target for data mining. The current multimedia browsing technology can be revolutionized to record more detailed learner behaviors like mouse movements. The breaking down of larger learning content into smaller modules could also help the recording of detailed behaviors. Tin Can API provides the way of data communication for individual foreign language learners. For the foreign language education practitioners in China, after truly understanding why, what kind of, and how data should be collected as big data, can the real value of big data be achieved

ACKNOWLEDGMENT

This paper was supported by 'Jiangsu Social Science Fund' in China (14YYA001), 'the Fundamental Research Funds for the Central Universities' in China (2242015S20018), and 'the Fundamental Research Funds for the Central Universities' in China (2242014S20073).

REFERENCES

[1] David Smith. 5 real-world uses of big data. Jul. 17, 2011. https://gigaom.com/2011/07/17/5-real-world-uses-of-big-data/[OL]. 2015-04-11.

[2] Donald Fisk. Beer and Nappies–A Data Mining Urban Legend. http://web.onetel.net.uk/~hibou/Beer and Nappies.html [OL].2015-04-05.

[3] Wang Haixiao.Reform in the Teaching of College English Writing in the Big Data Era[J]. Modern Distance Education Research,2014,03:66-72+86.

[4] Liu Runqing. The Scientific Research of Foreign Language Education in Big Data Era[J]. Contemporary Foreign Languages Studied,2014,07:1-6.

[5]Poltrack, J., Hruska, N., Johnson, A., & Haag, J. (2012). The next generation of scorm: Innovation for the global force. In The Interservice/Industry Training, Simulation & Education Conference (I/ITSEC) (Vol. 2012, No. 1).