Text Mining: The Gold Rush of Big Data

Assistant Professor Jiang Jing, a data analyst expert from the SMU School of Information Systems, performs data mining on social media platforms to gain insights.

By Alan Aw

SMU Office of Research (15 May 2014) – Every second, troves of information are generated on social media channels such as Twitter and Facebook. But the majority of this data cannot be analysed meaningfully, because it is complex and difficult to understand.

According to IBM, 95 percent of digital data generated daily is unstructured. Referring to content from social media or the Internet, unstructured data presents an opportunity for businesses and governments to predict human behaviour and map influencer networks. Such data is growing at double the rate of traditional data, which is stored in databases and commonly accessed by a structured query language.

“The goal of mining is to discover knowledge from data. Thus, text mining is about distilling knowledge from sources of textual data, which include news articles, emails, and even Facebook messages or Short Message Service (SMS),” says Assistant Professor Jiang Jing from the Singapore Management University (SMU) School of Information Systems (SIS).

Text mining, she explains, is different from searching the Internet. When using a search engine such as Google and Baidu, we are looking for information that has already been organised, such as a Wikipedia article on a topic.

Making sense of unstructured data, on the other hand, is a huge challenge for data analysts. A computer that analyses swathes of data must “learn” the rules governing the language in order to capture meaning from it. Creating information, and in turn knowledge from textual data, is known as natural language processing, an area which represents the basis of Professor Jiang’s research.

“Text mining works on a computational approach to language. We may not be excellent at using the language, but it is important for us to identify its structure, grammar, and vocabulary, so that a computer can ‘learn’ them, and apply it to the processing of textual documents,” advises Professor Jiang.

Tweeting your heart out

Professor Jiang’s research contributions are linked to the popular micro-blogging platform Twitter.

“There are many things we can find out by studying the Twitter platform. For example, we try to understand what people talk about on Twitter and their personal interests, so as to make relevant recommendations to each individual,” says Professor Jiang.

According to her, text mining can potentially be used to shed insights into difficult questions that cannot be easily answered through surveys. “Traditionally, social scientists need to conduct surveys to ask people about their feelings or reactions. But a lot of these opinions are readily available on the Web nowadays, so we need to find a way to extract, categorise and organise the information that can best address the questions,” she says.
In collaboration with fellow SIS colleague Professor Lim Ee-Peng and a visiting student from Peking University, Professor Jiang applied text mining to compare topics published on traditional media platforms and Twitter. The researchers found that even though both media formats covered similar topics, Twitter users covered celebrities and brands that were less emphasised by traditional media. Thus, companies can direct their marketing efforts to the micro-blogging platform to enhance brand popularity, she advises.

Apart from enhancing marketing strategies, Professor Jiang also worked with Professor Lim, Dr Palakorn Achananuparp, and other researchers from the SMU-Carnegie Mellon Living Analytics Research Centre to study local community sentiment. To do so, they analysed the tweets and followed links of the Twitter community in Singapore.

"In a piece of work that I am proud of, we developed a model for measuring individual influence on Twitter. Instead of using the traditional approach of using number of followers to measure influence, we tried to understand the underlying motives. We realised that when a user follows another user, he/she is likely to be followed by others as well. This is because the users share similar interests, which are captured through their use of common phrases," she says.

**Staying trendy is part of the job**

Professor Jiang says data analysts must always stay ahead of social trends, such as being familiar with the latest mobile messaging platforms that are in use. “People use Whatsapp and Wechat to talk to their friends these days. Text mining can only be useful if we obtain commonly available data, so the challenge is to extract knowledge from these new text platforms.”

Text mining of unstructured information – the wild west of big data – is not without its challenges, notes Professor Jiang. These include security issues that are related to private information, and integration of information in different languages.

"There are security issues when companies have access to this rich data, which holds huge potential for knowledge," she notes. "Another equally significant phenomenon is the convergence of people, cultures and language. Already in Singapore, you see people who use different languages interchangeably. One potentially significant breakthrough would be to understand the structure of multilingual exchanges to perform text mining on multilingual platforms, so that we can gain a deeper understanding of ideas, culture, people, and our society."

Back to Research@SMU Issue 14 (May 2014)