Asian Review of World Histories 4:1 (January 2016), 125-132

doi: http://dx.doi.org/10.12773/arwh.2016.4.1.125

Conclusion and Future Work

Provenance and Validation from the Humanities to Automatic Acquisition of Semantic Knowledge and Machine Reading for News and Historical Sources Indexing/Summary

Andrea NANETTI, Chin-Yew LIN and Siew Ann CHEONG*

Abstract

This paper, as a conlcusion to this special issue, presents the future work that is being carried out at NTU Singapore in collaboration with Microsoft Research and Microsoft Azure for Research. For our research team the real frontier research in world histories starts when we want to use computers to structure historical information, model historical narratives, simulate theoretical large scale hypotheses, and incent world historians to use virtual assistants and/or engage them in teamwork using social media and/or seduce them with immersive spaces to provide new learning and sharing environments, in which new things can emerge and happen: "You do not know which will be the next idea. Just repeating the same things is not enough" (Carlo Rubbia, 1984 Nobel Price in Physics, at Nanyang Technological University on January 19, 2016).

Key words

Provenance, Validation, World Histories, Big Data, Data Science

For the first time ever, our society has acquired the technological capacity to organize and retrieve myriad records of human experiences at any time and from any digitally connected place, and apply these elements to any kind of decision-making process. For what concerns world histories, automatic acquisition of historical knowledge and machine reading applied to historical sources indexing/summary and automatic narratives solutions can move from the historian and the reporter experiences in finding out more and more background information surrounding any event under investigation.

The optimist in us may think that because of this ability to automatically link as much as needed knowledge to relevant courses of action will help us solve most of our problems eventually. We see this optimism as early as 1945, in Vannevar Bush's book *As We May Think*. In this book, Bush discussed the problem of storing and retrieving 'cultural records,' and proposed the prototype networked personal computer "Memex" and the hyperlink as the solution. 70 years on, we still do not have the complete solution that Bush envisaged, but believe that every step towards a better organisation of the treasure of human experiences and to make it universally accessible and useful is a step forward in the advancement of learning and the evolution of human society.

Nevertheless, the pessimist in us realizes a problem. With advances in Information and Communication Technology (ICT), we are today generating digital information with unprecedented volume, velocity, and variety. In the face of this Big Data problem, it is entirely possible that we understand less and less of the sum total of human experiences. Lest we think that Big Data is a new and contemporary problem, let us also point out that Big Data has always been recognised as a problem by scientists (e.g., Renaissance Italy, Enlightenment, Idealism, Positivism. . .). In the past, the only solution has always been to select something at the detriment of something else kept in latency or deleted. And it is what we would do today after the digital revolution. In a sense, the problem and the theoretical

solution remain the same one already experienced by human societies before us.

However, we can make a difference this time round, because we are in command of ICT technology that did not exist until now. In the past, an expert well versed in the analysis of certain data and information can let his or her ideas be known by recording them down on traditional media such as books, paintings, music, and plays. Other experts can 'consume' these records, refine their own ideas, and create records of their own. This resulted in two decoupled processes: (1) human-to-media, and (2) media-to-human. After the invention of printing, process (2) became much faster than process (1), because humans process printed knowledge in parallel, but can only create knowledge serially, using very slow technology. Today, with new media replacing much of traditional media, we have the reverse situation. Because content creation is so easy with present-day ICT, even non-experts hop onto the bandwagon, to produce knowledge and pseudo-knowledge in parallel. In contrast, humans continue to rely largely on their organic computing power to process the deluge of content. Process (1) becomes so much faster than process (2) that as a society, we are no longer effectively aggregating knowledge. Unless we eliminate this bottleneck soon when we are going up against Big Data of today, the horizon of our collective knowledge will shrink when we have to go up against Bigger Data of tomorrow.

Based on the discussions above, we identify two key problems that are related to each other. The first problem is the weak coupling between processes (1) and (2), such that when one is slower than the other, a bottleneck is formed. The second problem is the validation and provenance of information or the varacity problem in Big Data jargon. When process (1) was slower than process (2) in the past, there was plenty of time to verify if a piece of information was correct, and where the information really came from. When information comes so quickly in the modern era, few of our old, manual ways for validating and establishing provenance still work. Today, we find

many initiatives around the world to build smart cities, smart nations, and even super smart societies. None of these initiatives will amount to anything, unless these two problems are solved.

To kill two birds with one stone, we propose to bring the humanities to the fore, to study societies and ask what is it that they desire, and how modern technology can help to realize these. Fundamentally, people want to interact with other people (family, friends, colleagues, and even total strangers). Therefore, instead of thinking that people create and consume media, we should think of them as creating and consuming 'interactions'. The purpose of creation is to consume. It makes no sense to write a book if the author does not want anyone to read it. When we think of processes (1) and (2) in this way, we realize that it is possible, without having to invent brand new technologies, to couple the two processes in a way that restores the symmetry between creator and consumer.

For example, let us imagine someone posting a series of old photographs on a town that had fallen into ruin for social and economic reasons. This person writes in his comment of one particular photograph how he remembers the bustling town square as a kid. Someone else, who chanced upon these photographs, realized that she was also from this town. She affirms the buzz in the town square, and also shares information on another photograph, information that the poster of the old photographs was not aware of. All these activities attract a third, fourth, ... persons, some former residents of the town, others children of former residents, still others who have visited the town as tourists. A few visitors to these photographs are professional historians who know different aspects of the distant past of the town. Together, they draw in new sources of information on the town and piece together a micro-history of the town going back several hundred years, and also identified possible reasons for its decline. Encouraged by this case study, many others started contributing old photographs and maps of old towns that are still around or have gone extinct, to build up their micro-histories. Eventually, people noticed common threads that link the destinies of these towns, and started building up a history of the broader region. This knowledge

aggregation process, spurred by the desire of people to discover their roots, can then continue to go deeper (further back in time or more details for a given historical period), or broader (more towns and cities), or higher-level (from towns to cities to nations).

However, we have outlined a best-case scenario, where interactions between the creator and consumer encouraged the consumer to also assume the role of a creator, attracting other consumers who would in turn become creator contributors. These interactions can happen by chance, but there is a much higher chance that they do not occur, in which case the series of old photographs languish in cyberspace, never to be discovered. A study[1] by John R. Frank et al. indicates the median lag time for news articles pubished after 2001 and got cited by a Wikipedia article in the Living People category about 1.5 year. We would like to create an online environment, populated by data mining and machine learning tools, where these interactions do not happen by chance, but are instead engineered.

To achieve this, we need to learn what people do online. Our objective is not simply to profile them, but to use their online profiles to define correlations between people. If people have gone to the same university, or universities in the same city around the same time, there is some chance that they know each other, or come from the same town. Even if they do not know each other, by virtue of coming from the same town implies that they are more likely to have shared interests, compared to with people from different towns. In a nutshell, people who are part of a shared history (of the town, for example) will not only have many digital footprints in common while they shared in the history, but will also be conditioned by this shared history to leave highly similar digital footprints after they part from each

* Andrea Nanetti, School of Art, Design and Media, Nanyang Technological University; Chin-Yew Lin, Knowledge Mining Group, Microsoft Research; Siew Ann Cheong, School of Physical and Mathematical Sciences, Nanyang Technological University

[1] John R. Frank, Ian Soboroff, Max Kleiman-Weiner, and Dan A. Roberts, "Entity-Oriented Filtering of Large Streams," http://www.nist.gov/tac/publications/2012/presentations/KBA-2012-overview.ppt.

other. This is the basis for us identifying groups of people we can act on. By acting on them, we mean to use individualized weights on searches by them and newsfeeds to them, so as to enhance the probability of them discovering each other and interact. By using modern ICT to provide positive feedback to human-media-human interactions leading to more human-media-human interactions, we will have closed the loop between processes (1) and (2).

As Donald A. Norman theorized in his 1993 book *Things That Make Us Smart*, "the complex interaction between human thought and the technology it creates, arguing for the development of machines that fit our minds, rather than minds that must conform to the machine. Humans have always worked with objects to extend our cognitive powers, from counting on our fingers to designing massive supercomputers. But advanced technology does more than merely assist with thought and memory—the machines we create begin to shape how we think and, at times, even what we value. Norman, in exploring this complex relationship between humans and machines, gives us the first steps towards demanding a person-centered redesign of the machines that surround our lives."

The next step in the research will be a focus on validation and provenance, which are common and recurrent (and not yet solved) issues in the papers presented at international scholarly conferences between 2013 and 2015 by Andrea Nanetti and his interdisciplinary research team, from various syntropic and complementary perspectives: with Siew Ann Cheong and Mikhail Filippov (Complexity Science) in 2013 in Kyoto at the Culture and computing conference, with Francesco Perono Cacciafoco (Linguistics) and Mario Giberti (Architecture and Mapping) in 2014 in Glasgow at the International Congress of Onomastic Sciences, with Siew Ann Cheong in 2015 in Jinan at the International Congress of Historical Sciences, with Angelo Cattaneo (History of Cartography), Siew Ann Cheong, Keng We Koh (Maritime Trade in Asia) and Chin-Yew Lin (Computer science) in 2015 in Singapore at the Congress of the Asian Association of World Historians and in Rio de Janeiro at the International Cartographic Conference, and with Anna Simpson

(Social Media Studies) in 2015 in Barcelona at the International Conference on Social Media Technologies, Communication, and Informatics.

In the history discipline, validation and provenance have relied heavily on authorities. For example, if an ancient piece of writing is used by many historians as a source for their works, then this ancient piece of writing becomes authoritative, and events recounted within are more trusted than suspected. As another example, a historian who has done good work on a particular historical era, and is cited by many historians working on the same historical era, also becomes authoritative. His claims are again more trusted, and accepted as valid. This reliance on authority stems from few historical events having direct material evidence, and sources are writings by participants or observers. Such a validation method is not always reliable, but remains useful even in the Big Data era. In the present day, most events are recorded electronically. These records are our primary sources. However, such records, whose provenance are known, and whose sources are authoritative, are generally too confusing and arrive too quickly in too large a volume. Therefore, many of our decisions are based on secondary interpretations of the primary data. For example, when a famous market commentator goes on TV to give a poor prognosis of a stock, bringing in multiple sources of information, investors trust the commentator, and proceed to dump the stock. Some of these investors may themselves be trusted on their own social networks, and should they post their actions on social media, may encourage more investors to sell the stock. This process can continue on the network of trust and turn into a cascade, even when the information leading to the poor initial prognosis is wrong. A good example is a false report of United Airlines filing for bankruptcy on September 8, 2008.[2] On that day, United Airlines saw its stock plunge from $12 to $3 in less than an hour. We have seen numerous other examples of bad decision making

[2] http://www.pbs.org/newshour/bb/media-july-dec08-unitedstock_09-09/

like this in social media, but also on traditional media like TV and newspapers.

Computer scientists have been working hard on the problem of validation and provenance, but have unnecessarily restricted themselves to the technological realm, i.e. they compare information in electronic form to see if they can find a body of information that corroborates with each other. Separately, they have been working on recommendation engines to learn the trustworthiness of sources from user ratings.

Ultimately, people are the sources and users of information that makes its way into cyberspace, and we must be aware that even a single human word encodes multiple ideas that come from multiple agents and activate different reactions according to the cultural experiences of the receiver. By closing the loop, we not only have data mining and machine learning tools to evaluate the validity and establish the provenance of information and knowledge, but can also rely on the passive and active interactions of humans with such information and knowledge as means to validate them and check their source. After all, not all facts are true, and not all truths are factual. From our study of human history, we know that societies routinely manufactured truth, as part of their projects to build common futures. To understand and engineer knowledge aggregation in the Big Data era, we must understand that there are many such projects for the future playing out. Our effort to build validation and provenance into the closed loop of human-media-human interactions must therefore take into account the fact that a few of these projects succeed, but many of them fail, and the many that reside in cyberspace today are in constant competition with each other. In short, we must be able to tell which facts that we want are true, and which truths we want supported by facts, when we go about our validation and provenance business.