Title	Toward theory and method of hybrid data collection
Author(s)	Maddah, Mahed; Lukyanenko, Roman; Chiarini Tremblay, Monica
Editor(s)	Parsons, Jeffrey Tuunanen, Tuure Venable, John R. Helfert, Markus Donnellan, Brian Kenneally, Jim
Publication date	2016-05
Original citation	Maddah, M., Lukyanenko, R. & Chiarini Tremblay, M. 2016. Toward theory and method of hybrid data collection. In: Parsons, J., Tuunanen, T., Venable, J. R., Helfert, M., Donnellan, B., & Kenneally, J. (eds.) Breakthroughs and Emerging Insights from Ongoing Design Science Projects: Research-in-progress papers and poster presentations from the 11th International Conference on Design Science Research in Information Systems and Technology (DESRIST) 2016. St. John, Canada, 23-25 May. pp. 93-95
Type of publication	Conference item
Link to publisher's version	https://desrist2016.wordpress.com/ Access to the full text of the published version may require a subscription.
Rights	**©2016, The Author(s).**
Item downloaded from	http://hdl.handle.net/10468/2572

Toward Theory and Method of Hybrid Data Collection

Mahed Maddah[1], Roman Lukyanenko[1], Monica Chiarini Tremblay[1]

[1] College of Business, Florida International University, Miami, FL USA
{ mmadd011, rlukyane, tremblay}@fiu.edu

Traditionally, information systems were designed to collect data in a structured format [1]. Structured format provides consistent data for data consumers. The explosive growth of social media (e.g., Facebook, Twitter), however, consistently demonstrates the advantages of a different approach to data collection and storage. On social media, people have more freedom to generate content as they leverage the unstructured collection process (e.g., via open textboxes). Applying traditional structured approaches to these settings limits user expressions and prevents users from conveying all the information they want [1], [2]. Unstructured data from social media are challenging to organize, integrate, and aggregate for analysis because they are variable, heterogeneous and sparse. This motivates us to develop novel approaches to collection and storage that combine the advantages of both formats.

We propose a hybrid solution, in which data begins as unstructured and gradually gains structure based on the popular social media practice (e.g., on Twitter and Facebook) of hashtagging. Hashtags are typically inserted by social media users to tag parts of unstructured content [3]. Hashtags do not restrict user input as traditional pre-defined structured fields do, but allow for content to be better categorized, searched and integrated with other content that has the same or similar hashtags.

Despite the benefits, hashtags appear to be quite sporadic and random [3]. Thus, most tweets are posted without any hashtags, and similar tweets may and may not use hashtags [4]. Using hashtags is not mandatory and requires conscious effort and understanding of how and why to apply them. To make hashtags more effective, we propose a novel solution based on machine learning (ML) to make the hashtags more predictable and automatically generate and suggest hashtags. We consider research in cognitive psychology to gain a deeper theoretical understanding of the underlying psychological mechanisms behind human decision to utilize hashtags. We further contribute by suggesting an important connection between structure of human memory and data collection and storage on social media.

When people create information on social media, they frequently reflect on their personal experiences or share notable experiences of others (e.g., friends, public figures). Psychology divides human declarative memory into semantic and episodic [5]. Semantic memory comes from the general world knowledge (shared among humans). Studies show that structured data formats is suitable for capturing semantic memory [6]. On the other hand, episodic memory is the memory of autobiographical events and comes from person’s own experiences, which are necessarily unique.

As much of social media content is based on episodic memory, storing it in unstructured format (i.e., free-form text that allows for unabated representation of unique events) appears consistent with human cognition. In fact, imposing structure on this information may inhibit or even distort it [7]. Hashtag thus create a simple way for people to relate their own episodes to similar those of others. Yet, if a hashtag is not utilized, it is more challenging to integrate similar content, due the information's unique and personal nature.

Following the cognitive psychology foundations we thus develop a novel method to enhance the usage and quality of hashtags - "Autotag". Autotag has multiple stages. First, we automatically identify episodic content (which could be distinct from other uses of social media, such as link sharing, or product promotion). Second, we use ML to predict whether a particular social media post or tweet should have a hashtag (this is done by training ML on a variety of historic social media content). This is challenging as content similarity does not appear to be enough to predict hashtag usage [4]. Thus, informed by cognitive psychology, we include context and user profile variables. Finally, we extend existing auto hashtag algorithms [4], [8] by introducing elements of episodic memory theory to predict specific hashtags. We evaluate the ML method by comparing its predictions to real historic social media data and conducting a laboratory experiment with potential social media users.

The Autotag approach carries important implications for social media and information management theory and practice. We believe that, using the theory of human memory, we can design more effective mechanisms that relate and connect unique experiences of people communicated via social media. Implementing this approach could reduce sparseness and heterogeneity of unstructured data without limiting its freedom and could be used in a wide range of applications.

References

[1] R. Lukyanenko, J. Parsons, and Y. Wiersma, "The IQ of the Crowd: Understanding and Improving Information Quality in Structured User-generated Content," *Inf. Syst. Res.*, vol. 25, no. 4, pp. 669–689, 2014.

[2] X. Zhang, Y. Yu, H. Li, and Z. Lin, "Sentimental Interplay between Structured and Unstructured User-Generated Contents-An Empirical Study on Online Hotel Reviews," *Online Inf. Rev.*, vol. 40, no. 1, 2016.

[3] M. A. H. Khan, M. Iwai, and K. Sezaki, "Towards urban phenomenon sensing by automatic tagging of tweets," Networked Sensing Systems (INSS), 2012, pp. 1–7.

[4] J. She and L. Chen, "Tomoha: Topic model-based hashtag recommendation on twitter," presented at the International World Wide Web Conferences Steering Committee, 2014.

[5] E. Tulving, "Episodic and semantic memory 1," *Organ. Mem.*, vol. 381, no. 4, 1972.

[6] E. Goldstein, *Cognitive psychology: Connecting mind, research and everyday experience*. Cengage Learning, 2010.

[7] D. Brown, A. W. Scheflin, and C. D. Hammond, *Memory, trauma treatment, and the law.* WW Norton & Company, 1998.

[8] F. Godin, V. Slavkovikj, W. De Neve, B. Schrauwen, and R. Van de Walle, “Using topic models for twitter hashtag recommendation,” International World Wide Web Conferences Steering Committee, 2013, vol. 593–596.