iStory: Intelligent Storytelling with Social Data

Amin Beheshti
Macquarie University
Sydney, Australia
amin.beheshti@mq.edu.au

Alireza Tabebordbar
University of New South Wales
Sydney, Australia
alirezat@cse.unsw.edu.au

Boualem Benatallah
University of New South Wales
Sydney, Australia
boualem@cse.unsw.edu.au

ABSTRACT

The production of knowledge from ever increasing amount of social data is seen by many organizations as an increasingly important capability that can complement the traditional analytics sources. Examples include extracting knowledge and deriving insights from social data to improve government services, predict intelligence activities, personalize the advertisements in elections and improve national security and public health. Understanding social data can be challenging as the analysis goal can be subjective. In this context, storytelling is considered as an appropriate metaphor as it facilitates understanding and surfacing insights which is embedded within the data. In this paper, we focus on the research problem of 'understanding the social data' in general and more particularly the curation, summarization and presentation of large amounts of social data. The goal is to enable intelligent narrative construction based on the important features (extracted and ranked automatically) and enable storytelling at multiple levels and from different views using novel summarization techniques. We implement an interactive storytelling dashboard, namely iStory, and focus on a motivating scenario for analyzing Urban Social Issues from Twitter as it relates to the Australian Government Budget, to highlight how storytelling can significantly facilitate understanding social data.

CCS CONCEPTS

• **Mathematics of computing** → **Exploratory data analysis**; • **Information systems** → **Data analytics**; **Data mining**; **Summarization**; **Web services**; • **Human-centered computing** → **Information visualization**.

KEYWORDS

Storytelling, Data Curation, Knowledge Lake, Data Lake, Summarization

ACM Reference Format:
Amin Beheshti, Alireza Tabebordbar, and Boualem Benatallah. 2020. iStory: Intelligent Storytelling with Social Data. In *Companion Proceedings of the Web Conference 2020 (WWW '20 Companion), April 20–24, 2020, Taipei, Taiwan.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3366424.3383553

1 INTRODUCTION

The large amount of information generated on online social networks, such as Twitter (twitter.com/) and Facebook (facebook.com/), can provide a new slant on business intelligence and can lead to important insights. Examples include extracting knowledge and deriving insights from social data to improve government services [6] and predict intelligence activities [7]. One of the main challenges in this domain is to transform social data into actionable insights. This task is challenging as the analysis goal can be subjective as it depends on analyst's perspective. Accordingly, storytelling is considered as an appropriate metaphor as it facilitates understanding data. In particular, discovery and analysis of narratives (i.e., set of related summaries) will be an important step toward understanding how analysts might reason about the impact of related features. Accordingly, a story will be able to combine data with narratives to reduce the ambiguity of social data, to connect this data with the context and to describe a specific interpretation. Most of the related works [2, 10, 11] in data-driven storytelling presented interactive visualizations to convey data-driven discoveries. However, data storytelling is much more than sophisticated ways to present data visually. The *main challenges in storytelling* include: building the foundation for organizing the raw data, contextualizing the raw data, enhancing the discovery of connected events and entities and finally presenting the data to the end-user in an interactive manner.

In this paper, we focus on the research problem of 'understanding the social data' in general and more particularly the curation, summarization and presentation of large amounts of social data. The goal is to enable intelligent narrative construction based on the important features (extracted and ranked automatically) and enable storytelling at multiple levels and from different views using novel summarization techniques. We present a system, built upon our previous work on data curation [3, 4, 6], to: (i) build the foundation for organizing the raw data: we leverage our previous work, Data Lake [3], to facilitate the organization of big social data; (ii) contextualizing the raw data: we leverage our previous work, Knowledge Lake [4, 5], to facilitate the data curation process and enable automatic extraction and enrichment of hidden features from social data; (iii) summarization: enhancing the discovery of connected entities, concepts and topics; and (iv) presenting data in an interactive manner: we extend our previous work, ConceptMap [15], to enable feature engineering in social data analytics to automatically create features that make machine learning algorithms work.

We design and implement an interactive storytelling dashboard, namely iStory, to facilitate understanding and surfacing insights which is embedded within the social data. We focus on a motivating scenario for analyzing Urban Social Issues from Twitter as it relates to the Australian Government Budget, to highlight how storytelling with data can significantly facilitate understanding social data. The rest of the paper is organized as follows. Section 2 presents an overview of the system. In Section 3 we describe our demonstration scenario.

WWW '20 Companion, April 20–24, 2020, Taipei, Taiwan

ACM ISBN 978-1-4503-7024-0/20/04.
https://doi.org/10.1145/3366424.3383553

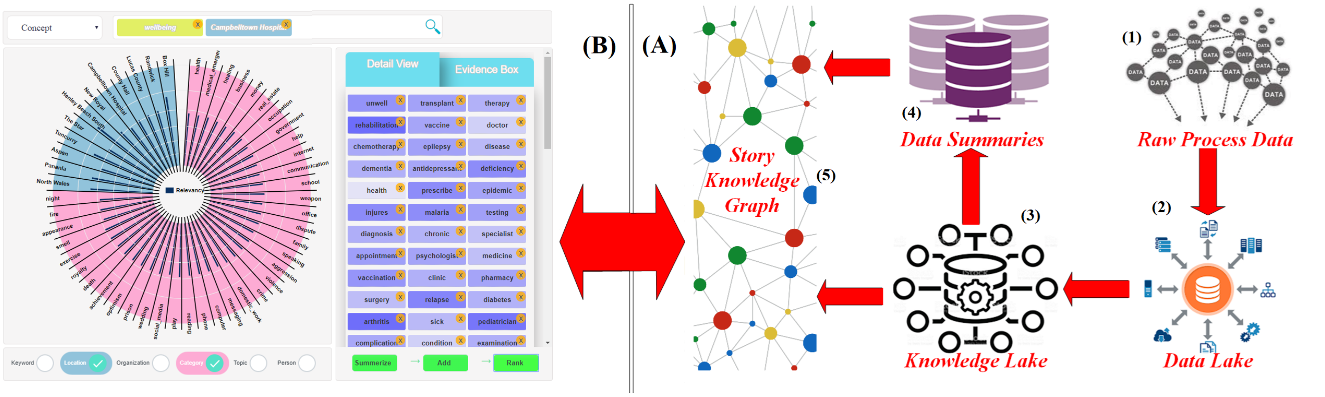


Figure 1: The storytelling engine (A) which is responsible to organize the raw data (A.1) in Data Lake (A.2), transform the raw data into contextualized data and knowledge (A.3), summarize the contextualized data (A.4) and construct the Story Knowledge Graph (A.5); and a screenshot of the digital dashboard (B).

2 SYSTEM OVERVIEW

We present a novel storytelling system to facilitate understanding the social data and to enable interactive exploration and visualization techniques to help analysts quickly identify interesting features and construct narratives. This helps us to put the first step towards enabling storytelling with social data. Figure 1 illustrates the storytelling engine (which is responsible to organize the raw data in Data Lake, transform the raw data into contextualized data and knowledge to build an Intelligent Knowledge Lake [5], summarize the contextualized data and construct the Story Knowledge Graph) and a screenshot of the developed digital dashboard.

2.1 Storytelling Engine

We present a storytelling engine to organize the raw social data in the Data Lake, transform the raw data into the contextualized data and knowledge, summarize the contextualized data and construct the Story Knowledge Graph.

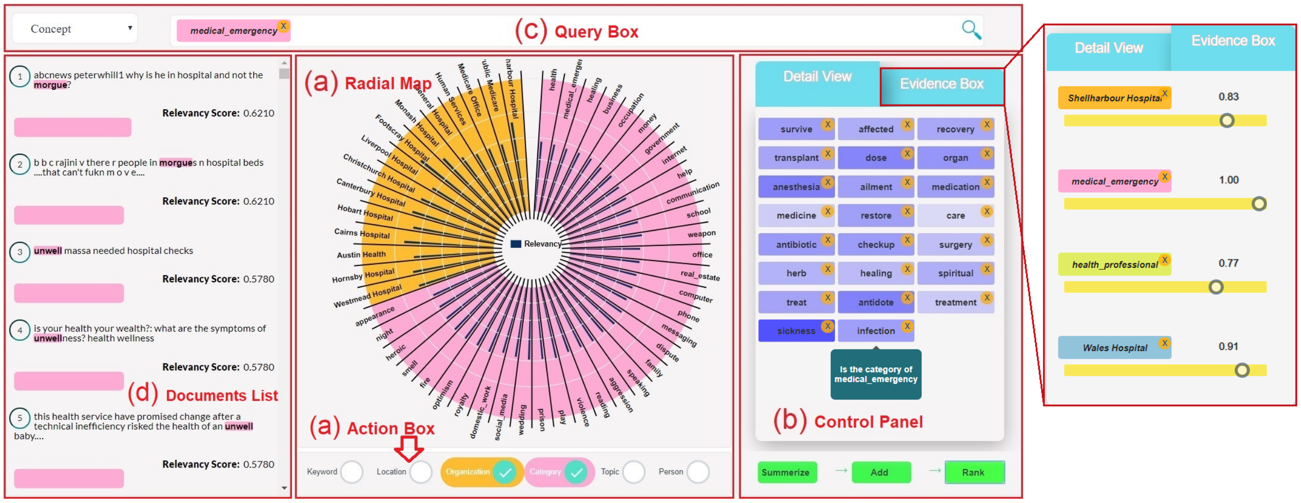
2.1.1 Organizing Social Data. Organizing vast amount of social data gathered from various data islands will facilitate dealing with a collection of independently-managed datasets such as Twitter and Facebook. The notion of a Data Lake has been coined to address this challenge and to convey the concept of a centralized repository containing limitless amounts of raw (or minimally curated) data stored in various data islands. At this level, we use our previous work, Data Lake as a Service [3], to facilitate the organization of social data. The Data Lake service manages multiple database technologies (from Relational to NoSQL databases), exposes the power of Elasticsearch and weave them together at the application layer. Moreover, it offers a built-in design to support security (to provide a database security threats including authentication, access control and data encryption) and Provenance [13] (to collect and aggregate tracing metadata including descriptive, administrative and temporal metadata and build a provenance graph).

2.1.2 Contextualizing Social Data. The rationale behind the Data Lake is to store raw data and let the data analyst decide how to cook/curate them later. At this level, we introduce the notion of Knowledge Lake (i.e., a contextualized Data Lake) and leverage our previous work, Knowledge Lake as a Service [4], to facilitate the data curation process and enable automatic extraction and enrichment of hidden features (e.g., facts, information, and insights extracted from the raw social data) from social data. The Knowledge Lake will automatically link the extracted enriched features to external knowledge bases (such as Google Knowledge Graph[1] and Wikidata[2] as well as other external data and knowledge sources, e.g., government open data[3]. The Knowledge Lake service automatically annotates items in data islands by information about the similarity among extracted information items, classifying and categorizing items into various types, forms or any other distinct class. This will enable enable intelligent narrative construction based on the important features that extracted and ranked automatically.

2.1.3 Story Knowledge Graph. In order to enable storytelling with social data, we introduce a graph-based data model[4]. The data model includes: (A) Entities: such as (i) raw social data, such as a Tweet in Twitter; (ii) content, such as a tweet text, source and time; and (iii) context, such as a named entity, keyword, or topic extracted from the text of a tweet; (B) Relationships among entities: such as Keyword $\xrightarrow{(etracted\text{-}from)}$ Tweet.text or Tweet $\xrightarrow{(contains)}$ Negative-sentiment; and (C) Summaries: which are abstractions that act as higher level entities to store and browse the results for follow-on analysis. For example, a summary could be a set of related tweets which contains same/similar keywords and/or entities. Another example, could be a set of related tweets having positive or negative sentiment [1].

The formalism of the summary enables considering different dimensions and views of a narrative, including the structure (narratives are about something happening), the purpose of a narrative (narratives about actors and artifacts), and the role of the listener

[1] https://developers.google.com/knowledge-graph/
[2] https://www.wikidata.org/
[3] https://data.gov.au/
[4] Graph data modeling is the process in which a user describes an arbitrary domain as a connected graph of nodes and relationships with properties and labels. The Resource Description Framework (RDF) and SPARQL query language are the W3C (w3.org/) standards to organize and query graphs [12].



1 abcnews peterwhill1 why is he in hospital and not the morgue?
Relevancy Score: 0.6210

2 b b c rajini v there r people in morgues n hospital beds ....that can't fukn m o v e....
Relevancy Score: 0.6210

3 unwell massa needed hospital checks
Relevancy Score: 0.5780

4 is your health your wealth?: what are the symptoms of unwellness? health wellness
Relevancy Score: 0.5780

5 this health service have promised change after a technical inefficiency risked the health of an unwell baby....
Relevancy Score: 0.5780

Is the category of medical_emergency

Figure 2: A snapshot of the interactive storytelling dashboard. We extend our previous work, ConceptMap [15], to enable feature engineering in social data analytics by connecting to the Story Knowledge Graph to automatically create features that make machine learning algorithms work.

(narratives are subjective and depend on the perspective of the analyst).

2.1.4 *Summarizing Social Data.*

Data summarization is an effective approach to improve the efficiency and scalability of data analytics tasks while dealing with the big data. Knowledge Lake services enable the execution of various data analytics tasks and facilitate constructing data summaries. At this level, to summarize the large Story Knowledge Graph, we present a novel summarization method which enables analysts to choose a set of features based on their specific goal (e.g., to analyze the social issues trending on social media related to the health category of government budget), and interact with small and informative summaries in an easy way. This will enable analysts to analyze the data from various dimensions. We present three types of summaries: OLAP Style, Group Style and Regular-Expression Style summaries.

OLAP Style Summaries. This type of summary will facilitate the analytics over the Story Knowledge Graph through summarizing the graph based on various dimentions and providing multiple views at different granularities. We leverage our previous work [8, 9], an OLAP (On-Line Analytical Processing) style data summarization technique to isolate the analyst from the process of explicitly analyzing different dimensions such as time, location, activity, actor and more. Instead, the system will be able to use interactive summary generation to select and sequence narratives dynamically. For example, set of dimentions such as topic, location and time can be used to summarize the Graph into set of related (social) users discussing about the same topic (e.g., health) in Australia during 2018.

Group Style Summaries. To summarize the Story Knowledge Graph, we support multiple information needs with one data structure (graph) and one function (machine learning algorithms such as classification). This capability enables analysts summarizing the large graph, by extracting complex data structures such as hierarchies and subgraphs. For example, this type of summary, can partition the Story Knowledge Graph into: (i) tweets having a specific keyword and/or named entity in their text; or (ii) tweets posted on a specific time period and/or from a specific location;

Regular-Expression Summaries. This type of summary can be used to discover patterns (i.e.,transitive relationships among entities) in the Story Knowledge Graph. We support reachability algorithms [14] (such as Transitive Closure, GRIPP, Tree Cover, Chain Cover, Path-Tree Cover, and Shortest-Paths) to enable partitioning the graph into set of related patterns. For example, a partition may include users who posted a tweet with negative sentiment:

$$\text{User} \xrightarrow{(post)} \text{Tweet} \xrightarrow{(contains)} \text{Negative-sentiment}$$

We present a narrative $N = \{S, R\}$, as a set of summaries $S = \{s_1, s_2, ..., s_n\}$ and a set of relationships $R = \{r_1, r_2, ..., r_m\}$ among them, where s_i is a summary name and r_j is a relationship of type 'part-of' between two summaries. This type of relationship enables the zoom-in and zoom-out operations to link different pieces of a story and enable the analyst to interact with narratives. This will enable storytelling at multiple levels and from different views using novel summarization techniques.

2.2 Interactive Storytelling Dashboard

We implement an interactive storytelling dashboard, namely iStory, on top of the Story Knowledge Graph to enable analysts to understand and surface insights which is embedded within the social data. The interactive storytelling dashboard takes the advantage of deep learning and the Story Knowledge Graph to provide a conceptual summary of the information space. In particular, iStory enable users to specify their preferences implicitly as a set of concepts without the need to iteratively investigate the information space. Moreover, the dashboard provides a 2D Radial Map of related features where a user can rank items relevant to her preferences through dragging and dropping. The interactive storytelling dashboard facilitates interaction with analysts to codify their knowledge into *regular* *expressions* that describe paths through the nodes and edges in the graph. The goal is to help users to better formulate their preferences when they need to retrieve varied and comprehensive list of information across a large amount of social data. Figure 2 illustrates a snapshot of interactive storytelling dashboard.

3 DEMONSTRATION SCENARIO

The demonstration scenario focuses on assisting knowledge workers in Australian Federal Budget (budget.gov.au/), in understanding governments' budget in the context of urban social issues. A typical government's budget denote how policy objectives are reconciled and implemented in various categories and programs. In particular, budget categories such as 'Health', 'Social-Services', 'transport' and 'employment' defines a hierarchical set of programs such as 'Aged Care' in Social-Services. These programs refers to a set of activities or services that meet specific policy objectives of the government [6]. Using traditionally adopted budget systems, it would be difficult to accurately evaluate the governments' services requirements and performance. For example, it is paramount to stabilize the economy through timely and dynamic adjustment in expenditure plans by considering related *social issues*. For instance, a problem or conflict raised by society ranging from local to national issues such as Health, Social Security, Public Safety, Welfare Support and Domestic Violence [6]. Therefore the opportunity to link active social issues (e.g., public opinions harvested from Tweets) to budget categories will provide the public with increased transparency, and likewise government agencies with realtime insight about how to make decisions (e.g., reshape policies).

The demonstration scenario consists of three parts: (i) Knowledge Lake: The Treasurer handing down the Budget on 3 May, each year. To properly analyze the proposed budget, we have collected all tweets from one month before and two months after this date. In particular, for these three months, we have selected 15 million tweets, posted on 2016. We demonstrate to the audience how to use the Data Lake service (to persist and index the tweets) and Knowledge Lake service (to transform the raw tweets into contextualized data and knowledge); (ii) Summarization and Narratives: we present to the audience a scenario to focus on the 'health' category of the budget and to summarize the tweets based on several features and to construct the three different types of summaries discussed in Section 2.1.4. (iii) Interactive Digital Dashboard: We demonstrate the interactive digital dashboard and present to the audience a scenario, how the digital dashboard can enable social network analysts to automatically use the features (from both entities and relationships among them) in the Story Knowledge Graph to formulate their preferences, link them to various summaries, construct narratives and tell stories from various dimensions and at different levels.

4 CONCLUSION AND FUTURE WORK

In this paper, we focused on the research problem of the curation, summarization and presentation of large amounts of social data in a succinct and consumable manner to business users. We implemented an interactive storytelling dashboard and focused on a motivating scenario for analyzing Urban Social Issues from Twitter as it relates to the Australian Government Budget. As a result of the move towards combining narratives and analytics with social data, a vast amount of analysis produced in the form of stories, i.e., set of related narratives presenting different perspectives of what happens in the enterprise, will feed and impact the business, creating a more pervasive analytics-driven environment. As the future work, we will extend our model to consider the importance of time and provenance as narratives may have different meanings over time.

ACKNOWLEDGMENTS

- We Acknowledge the AI-enabled Processes (AIP) Research Centre[5] for funding this research.

REFERENCES

[1] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. 2011. Sentiment analysis of twitter data. In *LSM*.

[2] Benjamin Bach, Moritz Stefaner, Jeremy Boy, and et al. 2018. Narrative Design Patterns for Data-Driven Storytelling. In *Data-Driven Storytelling*. AK Peters/CRC Press, 125–152.

[3] Amin Beheshti, Boualem Benatallah, Reza Nouri, Van Munin Chhieng, HuangTao Xiong, and Xu Zhao. 2017. CoreDB: a Data Lake Service. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM*. 2451–2454.

[4] Amin Beheshti, Boualem Benatallah, Reza Nouri, and Alireza Tabebordbar. 2018. CoreKG: a Knowledge Lake Service. *PVLDB* 11, 12 (2018), 1942–1945.

[5] Amin Beheshti, Boualem Benatallah, Quan Z. Sheng, and Francesco Schiliro. 2019. Intelligent Knowledge Lakes: The Age of Artificial Intelligence and Big Data. In *Web Information Systems Engineering - WISE 2019 Workshop, Demo, and Tutorial, Hong Kong and Macau, China, January 19-22, 2020, Revised Selected Papers*. 24–34.

[6] Amin Beheshti, Boualem Benatallah, Alireza Tabebordbar, Hamid Reza Motahari-Nezhad, Moshe Chai Barukh, and Reza Nouri. 2019. DataSynapse: A Social Data Curation Foundry. *Distributed and Parallel Databases* 37, 3 (2019), 351–384. https://doi.org/10.1007/s10619-018-7245-1

[7] Amin Beheshti, Vahid Moraveji-Hashemi, Shahpar Yakhchi, Hamid Reza Motahari-Nezhad, Seyed Mohssen Ghafari, and Jian Yang. 2020. personality2vec: Enabling the Analysis of Behavioral Disorders in Social Networks. In *Proceedings of the 13th ACM International Conference on Web Search and Data Mining, WSDM*.

[8] Amin Beheshti, Francesco Schiliro, Samira Ghodratnama, Farhad Amouzgar, Boualem Benatallah, Jian Yang, Quan Z. Sheng, Fabio Casati, and Hamid Reza Motahari-Nezhad. 2018. iProcess: Enabling IoT Platforms in Data-Driven Knowledge-Intensive Processes. In *Business Process Management Forum - BPM*. 108–126.

[9] Seyed-Mehdi-Reza Beheshti, Boualem Benatallah, and Hamid Reza Motahari-Nezhad. 2016. Scalable graph-based OLAP analytics over process execution data. *Distributed and Parallel Databases* 34, 3 (2016), 379–423. https://doi.org/10.1007/s10619-014-7171-9

[10] Sheena Erete, Emily Ryou, Geoff Smith, Khristina Marie Fassett, and Sarah Duda. 2016. Storytelling with data: Examining the use of data by non-profit organizations. In *Proceedings of the 19th ACM conference on Computer-Supported cooperative work & social computing*. ACM, 1273–1283.

[11] Marat Fayzullin, VS Subrahmanian, Massimiliano Albanese, Carmine Cesarano, and Antonio Picariello. 2007. Story creation from heterogeneous data sources. *Multimedia Tools and Applications* 33, 3 (2007), 351–377.

[12] Mohammad Hammoud, Dania Abed Rabbou, Reza Nouri, Seyed-Mehdi-Reza Beheshti, and Sherif Sakr. 2015. DREAM: Distributed RDF Engine with Adaptive Query Planner and Minimal Communication. *PVLDB* 8, 6 (2015), 654–665. https://doi.org/10.14778/2735703.2735705

[13] Luc Moreau, Juliana Freire, Joe Futrelle, Robert E McGrath, Jim Myers, and Patrick Paulson. 2008. The open provenance model: An overview. In *International Provenance and Annotation Workshop*. Springer, 323–326.

[14] Liam Roditty and Uri Zwick. 2016. A fully dynamic reachability algorithm for directed graphs with an almost linear update time. *SIAM J. Comput.* 45, 3 (2016), 712–733.

[15] Alireza Tabebordbar, Amin Beheshti, and Boualem Benatallah. 2019. ConceptMap: A Conceptual Approach for Formulating User Preferences in Large Information Spaces. In *Web Information Systems Engineering - WISE*. 779–794.

[5] https://aip-research-center.github.io/