Interactively Producing Purposive Samples for Qualitative Research using Exploratory Search

Orland Hoeber
University of Regina
Department of Computer Science
Regina, SK, Canada
orland.hoeber@uregina.ca

Larena Hoeber
University of Regina
Faculty of Kinesiology and Health Studies
Regina, SK, Canada
larena.hoeber@uregina.ca

Ryan Snelgrove
University of Waterloo
Department of Recreation and Leisure Studies
Waterloo, ON, Canada
ryan.snelgrove@uwaterloo.ca

Laura Wood
University of Waterloo
Department of Recreation and Leisure Studies
Waterloo, ON, Canada
laura.wood@uwaterloo.ca

ABSTRACT

An important step in conducting qualitative research on large collections of text is reducing the size of the collection to one that is manageable. While it is common to use a variety of simple sampling methods, the limitation of these approaches is that their mechanisms do not consider the relevance of the data. We have developed exploratory search methods that leverage visual analytics to produce purposive samples of large qualitative datasets. In this paper, we outline how exploratory search strategies lead to purposive sampling, and put this in the context of the interactive information retrieval literature.

CCS CONCEPTS

•**Information systems** →*Users and interactive retrieval;* •**Human-centered computing** →*Visual analytics;*

KEYWORDS

exploratory search, purposive sampling, qualitative research methods

1 INTRODUCTION

A common task of qualitative research on textual data is to assign codes to significant pieces of data (e.g., phrases, sentences, paragraphs) and then seeking patterns and relationships among the codes. Doing so allows a researcher to translate a large set of textual data into a constrained vocabulary that is more easily kept in mind, and therefore more easily understood. The challenge is that coding data is tedious and time consuming due to the need to carefully read the text before assigning codes. This challenge is greater when dealing with large textual datasets such as those from social media services; it is often necessary to take measures to reduce the amount of data to be coded and analyzed further.

The general approach for data reduction is to sample the data. Stratified sampling seeks to reduce the amount of data to consider by selecting a sub-population of the whole [8]. For the task of analyzing social media data, this may be done by categorizing the stakeholders (e.g., in the context of sporting events: fans, media, athletes, coaches, organizers), and choosing all the posts from those within a specific stakeholder category. Randomness may also be added to the stratified sampling method in two ways: random selection of the stratification (e.g., randomly choosing which stakeholders to follow), or random selection of the data within a specific stratum (e.g., random selection of posts from the selected stakeholders) [8]. Systematic sampling may be used as an alternative to random sampling, where a rule-based mechanism is employed (e.g., select every n^{th} post). Fundamentally, these sampling methods may overlook or ignore important qualities of the larger dataset, such as missing important aspects of the data, interactions among key stakeholders, and the temporal relationships between the data and other events occurring that are inspiring people to post their thoughts and opinions on social media [5].

A fundamentally different approach to data reduction is to perform purposive sampling of the data by carefully choosing a subset based on relevance to the topic of interest. However, doing so in the context of qualitative research requires that the entire dataset be considered, limiting the feasibility when the dataset is large. Our solution to this problem has been to leverage technology to support the human effort of qualitative research [9]. The interactive creation of purposive samples of the data allow the dataset to be reduced to include all relevant posts for a given topic of interest, ensuring that important features are not missed. Furthermore, by maintaining the temporal aspect of the data, the qualitative features can be studied in order, and in consideration of the real-world context in which the posts are situated. In the specific context of studying public opinion posted on Twitter, we have developed a software system called Vista (Visual Twitter Analytics) [4], which enables the visual exploration, creation, and export of purposive samples. In the remainder of this paper, we will explain how purposive sampling is enabled through exploratory search, and how visual analytics approaches can enhance the process. We will also discuss search strategies that allow a qualitative researcher to focus their search activities as they develop inductive research questions.

CHIIR 2017 Workshop on Supporting Complex Search Tasks, Oslo, Norway.

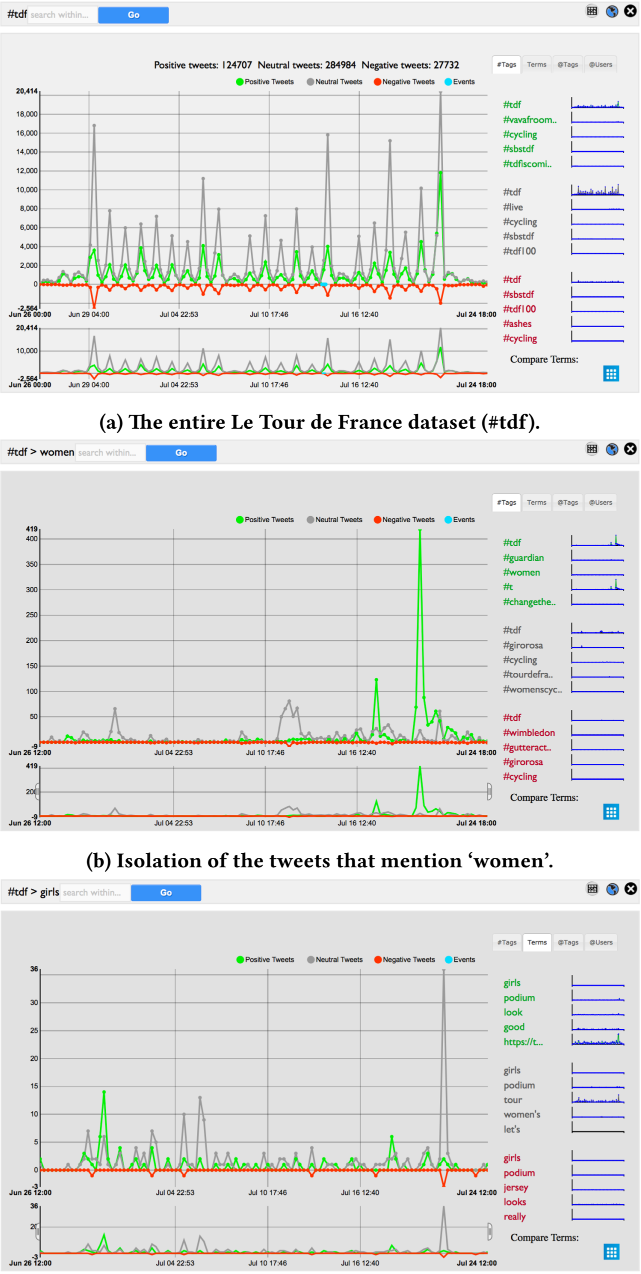
2 EXPLORATORY SEARCH LEADING TO PURPOSIVE SAMPLING

A necessary first step is to collect a large set of data that captures as much of the information relevant to a high-level interest as possible. For social media services such as Twitter, this may be done by choosing many hashtags and query terms in order to capture the full breadth of public interest in a topic, and collecting the data over an extended period of time. Doing so means that the researcher does not need to identify the specific research questions to be pursued *a priori*, but may instead collect a large dataset that has a high probability of capturing salient aspects that emerge as the topic or event of interest unfolds. This is important in situations such as critical event analysis in sport, where it is difficult to predict what issues or micro-events might occur in advance.

With such a large dataset, exploratory search [11] is a valuable approach for enabling a researcher to inductively develop specific research questions to pursue in detail. In particular, the interactive nature of the searcher engaging within the search process [3] allows for potential avenues of interest to be pursued, considered, saved, synthesized, and evaluated in the context of developing research questions. For example, one might collect all of the tweets posted during a mega-sporting event such as Le Tour de France using their official hashtag #tdf, and interactively explore the data to find what issues people are commenting upon. The discovery of gender issues within a predominantly male event may lead to the development of research questions to be pursued within the data, supported by searching for various different embodiments of this issue within the tweets.

Our particular approach has been to use visual analytics [6] to enable the (re)searcher to take an active and informed role in the search process. Vista [4] provides a visual overview of the temporally changing sentiment of the collected Twitter data, presents visual overviews of the top terms, hashtags, user mentions, and authors, provides a geovisualization of the tweet source locations, and enables sub-querying and exclusions to create visually comparable sentiment timelines (see Figure 1). For the purposes of exploring the data to discover emergent events and issues, the sentiment timeline provides a pre-analysis of the data to draw attention to times during the event when there are strong positive, negative, or divisive sentiment. The visual overviews of the terms/hashtags/mentions/authors allow for the recognition of relevant and irrelevant topics, making it easy for the (re)searcher to isolate the data relevant to the topic, or exclude it from the whole. Textual querying is supported, allowing the (re)searcher to construct queries based on their knowledge of the possible issues and micro-events that may have occurred. If spatial and temporal aspects of the data are also relevant, queries can be generated that limit the temporal range and the spatial extent of the data.

An unique feature of Vista is the mechanism by which it supports the comparison and analysis of multiple sets of search results. Each query of the data adds a new section within the visual overview of the data, drawn as a sentiment timeline. As a result, the (re)searcher can generate multiple queries of the data and visually compare the temporal pattern of the stakeholders engagement with the associated issues. Keeping the timelines synchronized enables the visual identification of patterns and relationships across the sets of search results.



(a) The entire Le Tour de France dataset (#tdf).

(b) Isolation of the tweets that mention ‘women’.

(c) Isolation of the tweets that mention ‘girls’.

Figure 1: Using Vista, samples of the data can be isolated focusing on the use of the terms ‘women’ and ‘girls’ and stacked upon one another, allowing visual comparison between them and back to the entire dataset under consideration.

Fruitful avenues of exploration of the data can be maintained and refined, and new hypotheses can be investigated and discarded if they do not reveal useful information. Through the careful selection and experimentation of what terms and hashtags to include and exclude from the data, a purposive sample of the data can be selected and evaluated in an interactive manner, and ultimately exported for further analysis. In addition, during the traditional qualitative analysis of the exported data, the researcher can readily return to Vista to explore newly emergent topics, as well as inspect the details of embedded links and the authors to support their coding and analysis tasks.

In the context of supporting the qualitative study of emergent issues within large datasets, it is useful to consider the importance of serendipity within the (re)search process [1]. The ability to easily generate queries of the data and visually evaluate the results allows for new avenues of inquiry to be readily pursued. Should the searcher stumble upon some topic that was not considered, it can be isolated from the data, and new research questions may be inductively developed to study this aspect of the data.

Considering the use of Vista to produce a purposive sample in light of the theory on interactive information retrieval, we can consider the process to be strongly influenced by exploratory search [11] and sensemaking [7]. In particular, the researcher may start with a vague and under-defined goal for searching within the data, and may seek to develop knowledge and understanding by interactively querying the data to isolate potentially important subsets. The process of starting with a large set of data, identifying a potential research question, and iteratively sub-querying the data to both inductively refine the research question and isolate the tweets that are relevant to the issue is an evolutionary search process. Researchers using the system may employ berry-picking strategies [2] to learn and develop an understanding of what is being sought. This ultimately leads to the co-development of research questions to ask of the data and complex queries (textual, spatial, temporal) to isolate the data to answer the questions.

Because of the need to ensure that all relevant information is discovered, a structured information seeking process is beneficial, such as the task-based information seeking model proposed by Vakkari [10]. Initial assessments of the data, exploratory sub-querying, inspection of the tweets, and preliminary development of a set of possible research questions to pursue within the data can be considered *pre-focus* tasks. Once the researcher settles on a research question to delve into, they may issue a series of sub-queries to isolate the relevant data, and use the visualization of the sentiment timelines to verify the patterns in the data, representing the *focus formulation* tasks. With sufficient data selected via the purposive sample, the corresponding tweets may then be exported and analyzed within traditional qualitative research software and methods, which constitutes the *post-focus* tasks. Such a structured task-centric model of the information seeking process meshes well with the structured research methods that are commonplace in qualitative research.

3 CONCLUSION

The primary contribution of this paper is the presentation of a qualitative research mechanism that leverages interactive exploratory search and visual analytics to enable the dynamic development of purposive samples that address emergent research questions. Our ongoing work is to refine and enhance Vista to further support the interactive exploration, discovery, and isolation of subsets of textual data, producing topically-complete samples that make the application of traditional qualitative methods tractable.

REFERENCES

[1] Paul André, Jamie Teevan, and Susan T Dumais. 2009. From x-rays to silly putty via Uranus: serendipity and its role in web search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 2033–2036. DOI: http://dx.doi.org/10.1145/1518701.1519009

[2] Marcia J Bates. 1989. The design of browsing and berrypicking techniques for the on-line search interface. *Online Review* 13, 5 (1989), 407–431. DOI: http://dx.doi.org/10.1108/eb024320

[3] Nicholas J. Belkin. 2015. People, Interacting with Information. *ACM SIGIR Forum* 49, 2 (2015), 13–27. DOI: http://dx.doi.org/10.1145/2766462.2767854

[4] Orland Hoeber, Larena Hoeber, Maha El Meseery, Kenneth Odoh, and Radhika Gopi. 2016. Visual Twitter analytics (Vista): Temporally changing sentiment and the discovery of emergent themes within sport event tweets. *Online Information Review* 40, 1 (2016), 25–41. DOI: http://dx.doi.org/10.1108/OIR-02-2015-0067

[5] Brett Hutchins. 2014. Twitter: Follow the money and look beyond sports. *Communication & Sport* 2, 2 (2014), 122–126. DOI: http://dx.doi.org/10.1177/2167479514527430

[6] Daniel A Keim, Gennady Andrienko, Jean-Daniel Fekete, Carsten Görg, Jörn Kohlhammer, and Guy Melançon. 2008. Visual analytics: Definition, process, and challenges. In *Information visualization: Human-centered issues and perspectives*, Andreas Kerren, John T Stasko, Jean-Daniel Fekete, and Chris North (Eds.). Springer-Verlag, Berlin Heidelberg, 154–175. DOI: http://dx.doi.org/10.1007/978-3-540-70956-5_7

[7] Peter Pirolli and Daniel M. Russell. 2011. Introduction to this Special Issue on Sensemaking. *Human-Computer Interaction* 26, 1-2 (2011), 1–8. DOI: http://dx.doi.org/10.1080/07370024.2011.556557

[8] Jeremy C. Short, David J. Ketchen, and Timothy B. Palmer. 2002. The role of sampling in strategic management research on performance: A two study analysis. *Journal of Management* 28, 3 (2002), 363–385. DOI: http://dx.doi.org/10.1177/014920630202800306

[9] Ramine Tinati, Susan Halford, Leslie Carr, and Catherine Pope. 2014. Big data: Methodological challenges and approaches for sociological analysis. *Sociology* 48, 4 (2014), 663–681. DOI: http://dx.doi.org/10.1177/0038038513511561

[10] Pertti Vakkari. 2003. Task-based information searching. *Annual Review of Information Science and Technology* 37, 2 (2003), 413–464. DOI: http://dx.doi.org/10.1002/aris.1440370110

[11] Ryen W White and Resa A Roth. 2009. *Exploratory Search: Beyond the Query-Response Paradigm.* Morgan & Claypool Publisher, San Rafael, CA. DOI: http://dx.doi.org/10.2200/S00174ED1V01Y200901ICR003