Exploration and Discovery of User-generated Content in Large Information Spaces

Luca Chiarandini*
Web Research Group
Universitat Pompeu Fabra
Tànger, 122-140
Barcelona, Spain
chiarluc@yahoo-inc.com

ABSTRACT

The accumulation of large collections of social media data poses new challenges for the design of exploratory experiences, such as when a user browses through a collection to discover content (*e.g.* exploring photo collections, network of friends, etc). Cardinality and characteristics of the set, together with volatility of the information, resulting from fast and continuous creation, deletion and updating of entries, trigger novel research questions. In this context, we plan to investigate and contribute to the data analysis, and user interface design of exploratory experiences. The proposed approach is an iterative process where analysis and design phases are performed in cycles. The long-term vision is to understand the underlying reasoning in order to be able to automatically replicate it.

Categories and Subject Descriptors

I.2 [**Computing Methodologies**]: Artificial Intelligence; H.5.2 [**Information Systems**]: Information Interfaces and Presentation—*User Interfaces*

General Terms

Design, Human Factors

Keywords

Human Computer Interaction, Data mining, Social media

1. INTRODUCTION

Exploration and discovery of content is a research topic that deals with the traveling through possibly unfamiliar data in order to making sense of it. Nowadays the amount of content is very large and increasing, rendering the problem more challenging, especially in the field of *social media*. Traditional ways of representing information (such as tables, lists, etc.), which have been suitable for a small amount of data, fail to scale. Information representation needs to account for the users' need, inherent capabilities and particular skills.

*Also Yahoo! Research Barcelona

WSDM'12, February 8–12, 2012, Seattle, Washington, USA.
ACM 978-1-4503-0747-5/12/02.

Good methods for exploring and presenting the content of a collection may help in many ways. Firstly, they help the user understanding how the collection is composed or how it evolves over time. Secondly, in search, it is sometimes hard when not even impossible for the user to formulate the query in the required form (*e.g.* text, keywords); therefore, the user needs to browse through content. In addition, it helps the user not only in finding what she needs but also retrieving content she did not expect but is interested in (*serendipity*). Finally, browsing through a collection can give the *sense of space*, *i.e.* make the user aware of how this is structured and being able to orientate herself in it [3].

During the Ph.D. we plan to investigate how the characteristics of large collections of *user-generated content* could be used in order to design the exploratory experience. The methodology uses techniques from *data mining* and *machine learning* to analyze the data and *human-computer interaction* to design the interaction. We will exploit an iterative approach in which the result of the analysis phase drives the design process. At the same time, collecting data about how users interact with the machine, allows us to trigger again the analytical step.

2. CONTEXT AND RELATED WORK

Morse [7] defined browsing as *search, hopefully serendipitous*). Bawden [1] identifies three kinds of browsing: *a*) *purposive*, *i.e.* the deliberate seeking for new information in a defined (albeit broad) subject area; *b*) *capricious*, *i.e.* random examination of material without a definite goal; and *c*) *exploratory* or *semi-purposive*, *i.e.* in search, quite literally, of inspiration. We will mainly consider the third kind.

Kaplan and Haenlein [6] define social media as *a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, which allows the creation and exchange of user-generated content.* They use two dimensions for the categorization (Table 1). On the one hand, the categorization considers *media richness* [2], *i.e.* the amount of information they allow to be transmitted in a given time interval. On the other hand, we consider *self-disclosure*, *i.e.* the conscious or unconscious revelation of personal information (*e.g.*, thoughts, feelings, likes, dislikes).

In the context of User Interfaces (UI), many authors devised methods in order to automatically build interfaces and adapt them to different devices. Nichols et al. [8] present a run-time, fully-autonomous UI generator based on a declarative language. More recently, Gajos and Weld [4] describe

Self-disclosure	Media-richness			
		Low	*Medium*	*High*
	High	Blogs	Facebook	Virtual social networks (*e.g.* Second life)
	Low	Wikipedia	Youtube	Virtual game worlds

Table 1: Social media categorization with examples for each category.

SUPPLE, which interprets the process of generating the UI as an optimization problem. Several design criteria may be expressed as cost functions. Our work differs from existing solutions under multiple aspects: *a*) the design of the interface is not static but *dynamically adapts* to users' behavior *b*) the goal of the interface is not just accessing or editing the data but should allow the user to browse through the dataset; therefore, the correspondence between the data point and interface is not trivial *c*) in our work, the context is the Web and in particular Social Media applications; this changes many aspects of the interaction (*e.g.* the user may want to read more than write).

3. RESEARCH QUESTIONS

The research task requires the understanding of the two sides of the interaction: *a*) *Large information spaces*: Which are the characteristics of large content-rich collections? How do they change over time? Understanding this may also suggest new ways of indexing, summarizing or querying the collection as well as give hints about possible UIs and exploration mechanisms. *b*) *User*: What does the user look for when browsing? Could we model or predict serendipity? What changes in the behavior of the user between browsing and accessing to information (*e.g.* web search, catalog querying, etc.)? User modeling has been extensively analyzed in the context of websites.

In parallel, many questions are related to the *interface* between these two entities: How could we increase the frequency of serendipitous encounters? Could we take into account intuitive knowledge of the user in the design of interfaces? How could dataset and user behavior models enhance the design process? Can UIs adapt to user's behavior to improve usability? Can browsing be a *social* experience?

4. METHODOLOGY

We address two aspects in this thesis: *a*) *Analysis* (Machine learning and data mining), in order to understand the characteristics of the dataset and model user's behavior *b*) *Design* (Human-computer interaction), to improve the user experience on the the basis of the results of the previous phase. The results of the analysis are used to extract design guidelines and improve both user experience and interaction mechanism. The integration of an automated data collection mechanisms not only allows us to evaluate the goodness of the method but also to extract information to iterate the process. Whereas in a first moment the iterations are done manually through studies and data mining, we aim at the automation of the process.

This methodology will be applied in different fields to test its generality. We have access to data about heterogeneous websites in the form of click- and pageview-logs (*i.e.* the information about where the user clicks and which pages she views). We plan to apply the analysis/design methodology to: *a*) *News sites*: portals where users are able to find the latest news, be informed about the happenings in the world or simply browse interesting content *b*) *Media and document browsing sites*: systems in which users are able to browse or optionally search for informative or multimedia content.

5. CONCLUSION

The Ph.D. research will contribute in settings in which the user browses through and explores a content-rich dataset. Social media is a particularly suitable context for this research since it involves very heterogeneous content and datasets (multimedia data, social network graphs, etc.).

Acknowledgement

This Ph.D. is under the supervision of Dr. Ricardo Baeza-Yates and Dr. Alejandro Jaimes.

References

[1] D. Bawden. Information systems and the stimulation of creativity. *Journal of Information Science*, 12(5):203, 1986.

[2] R. Daft and R. Lengel. Organizational information requirements, media richness and structural design. *Management science*, 32(5):554–571, 1986.

[3] M. Dörk, S. Carpendale, and C. Williamson. The information flaneur: A fresh look at information seeking. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, pages 1215–1224. ACM, 2011.

[4] K. Gajos and D. S. Weld. Supple: automatically generating user interfaces. In *Proceedings of the 9th international conference on Intelligent user interfaces*, IUI '04, pages 93–100, New York, NY, USA, 2004. ACM.

[5] E. Goffman. The presentation of self in everyday life. 1959. *Garden City, NY*, 2002.

[6] A. Kaplan and M. Haenlein. Users of the world, unite! the challenges and opportunities of social media. *Business Horizons*, 53(1):59–68, 2010.

[7] P. Morse. On browsing: the use of search theory in the search for information. *Bulletin of the Operations Research Society of America, Vol. 19 supplement, p.1*, 1971.

[8] J. Nichols, Myers, B. A., M. Higgins, J. Hughes, Harris, T. K., R. Rosenfeld, and M. Pignol. Generating remote control interfaces for complex appliances. In *Proceedings of the ACM Symposium on User Interface Software and Technology*, Papers: infrastructure for ubicomp, pages 161–170, 2002.

[9] J. Short, E. Williams, and B. Christie. *The social psychology of telecommunications*. John Wiley and Sons Ltd, 1976.

[10] R. Spence. A framework for navigation. *International Journal of Human-Computer Studies*, 51(5):919–945, 1999.