Visual Search Analytics: Combining Machine Learning and Interactive Visualization to Support Human-Centred Search

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

Abstract

Searching within large online document collections has become a common activity in our modern information-centric society. While simple fact verification tasks are well supported by current search technologies, when the search tasks become more complex, a substantial cognitive burden is placed on the searcher to craft and refine their queries, evaluate and explore among the search results, and ultimately making sense of what is found. Visual search analytics provides a means for relieving this burden through a combination of automatic machine learning and interactive visualization. The goal is to automatically extract and infer relevant information during the search process and present this to the searcher in a visual format that allows for quick interpretation and easy manipulation of the search process, providing support for the full range of human-centered search activities.

1 Introduction

A substantial portion of human knowledge exists in textual formats within online domain-specific document collections. Examples include Wikipedia, the ACM Digital Library, Engadget, Twitter, Associated Press, the vast document collections within corporate environments, and the ever-growing public collections indexed by the top web search engines. For small collections or for simple fact verification tasks, browsing or using basic search interfaces are often satisfactory; however, to support complex information seeking activities within large collections, it is necessary to provide powerful search facilities that support exploratory search and analytical reasoning about the information that has been found.

In recent years, there have been significant advancements in search technologies. Web search companies like Google and Microsoft can index billions of documents, and return matches to user-supplied queries within fractions of a second. Open source information retrieval frameworks such as Apache Lucene can efficiently index large unstructured document collections and provide the backbone for custom search systems. Amid all of these development efforts on the back-end of the search process, the interfaces to the vast majority of search systems have remained largely unchanged. The searcher is provided with a query box in which to type a description of what is being sought, and the search results are provided in a list-based format that requires document-by-document inspection.

While such interfaces work well for highly targeted search tasks such as fact verification, their ability to support complex search activities such as disambiguation and exploration are limited. Because of the fundamental differences in why and how people search within various large online document collections, the one-size-fits-all approach to search interfaces may not be appropriate in all settings. For example, one might initiate a search within an online encyclopedia such as Wikipedia in order to find specific facts and explanations about a topic of interest. However, because of the lack of specific knowledge about the topic, the initial query may be ambiguous or may not even match

In: U. Kruschwitz, F. Hopfgartner and C. Gurrin (eds.): Proceedings of the MindTheGap'14 Workshop, Berlin, Germany, 4-March-2014, published at http://ceur-ws.org

the terminology used within the encyclopedia. In the process of exploring among the search results, new information will be acquired that will allow for the refinement of the query in order to re-focus it more precisely on the topic of interest, as well as to develop new search interests based on serendipitously discovered information. By contrast, one might conduct a search within a corporate document collection to re-find a specific policy document. Because of the common language used within such documents, even for a very specific and accurate query, there may be many search results to evaluate. By re-constructing approximately when the document was last viewed, the collection of documents can be narrowed down to a more manageable size. The differences in these search processes exemplify the need for further research and study of how search interfaces can support complex information seeking tasks within a broad range of online search contexts.

In this paper, *visual search analytics* is introduced as a special class of visual analytics, with a focus on supporting human-centred search activities [Hoe12]. The more general, multidisciplinary research domain of visual analytics combines data processing and machine learning with information visualization and human-computer interaction, with the goal of supporting data exploration, analytical reasoning, information synthesis, and decision-making [TC06, KAF+08]. Visual search analytics applies this philosophy to search contexts, using intelligent visual methods for guiding query refinement, explaining the composition of the search results, supporting document evaluation and comparison, allowing interactive filtering and re-ranking, and enabling analytical reasoning, sense-making, and exploration among the search results. By enhancing the abilities of people to search within large document collections, the ever-present big data and information overload problems can be managed.

The remainder of this paper is organized as follows. The core concepts of information visualization, visual analytics, and visual text analytics are presented in Section 2. Section 3 outlines the fundamental features of visual search analytics, and explains its importance for supporting human-centred search activities. Section 4 presents a high-level research agenda for the advancement and study of visual search analytics, along with a critical discussions of its limitations. A summary of the primary contributions of this work are provided in Section 5.

2 Related Work

Information visualization is the field of study that explores the use of computer-generated graphical representations as a method for conveying abstract information to a user. It provides mechanisms for linking the data being processed within a computer system and the mind of the user, via the human vision system [War04]. The goal of information visualization is to take advantage of the parallel processing capabilities of human visual perception [WGK10], allowing people to see information, visually interpret patterns and relationships, and minimize the need to read or examine specific details. By making such visual representations interactive, users are able to manipulate and control the visualization as they seek to understand the data being shown [YaKSJ07].

Although information visualization has been explored in many different application domains, much of the focus has been on either the human perception of graphical entities, or the application of visualization approaches to various types of data (e.g., multidimensional data, graph data). The lack of focus on human-centric problem solving and data analysis tasks has lead to the promotion of a new field of research: visual analytics. Visual analytics combines data processing and machine learning with information visualization and human-computer interaction, with a specific focus on supporting data analysis activities such as exploration, reasoning, information synthesis, and decision-making [TC06, KAF+08]. The ultimate aim is to take advantage of the powerful analytic capabilities of the computer whenever possible, using the resulting information to support the cognitive abilities of the user through interactive visualization.

Although text and document visualization have been active research domains for many years [HHWN02, Hea95, WTP+95], much of the recent work in this area has followed the visual analytics approach of combining automated processes with interactive visualizations, resulting in visual text analytics [AdOP12, CCP09, DWS+12, GLK+13, KKRS13, WLS+10]. These approaches generally focus on providing visual tools for exploration among document collections. While they may include basic keyword search as a means for filtering the data, very little attention has been given to supporting the core tasks associated with searching among the textual data.

Others have studied how visualization can enhance the search interface [Hea09] or the information seeking process [MW09], and have identified the importance of providing additional support to searchers within the context of exploratory search [WKDs06, WR09, WsS10]. For example, HotMap [HY09] provides lightweight visual encodings of the correspondence between query terms and search results, and WordBars [HY08] visually represents the relative frequency of the top terms within the search results set. Both of these approaches provide interactive methods

for re-ranking the search results based on the addition input provided by the searcher, and have been shown to be helpful when the search tasks are complex or difficult [Hoe13].

3 Visual Search Analytics

Visual search analytics extends the normal keyword search paradigm by automatically extracting salient information from the query, search results, and/or the entire document collection, using visualization to convey this information to the searcher, and allowing the searcher to interactively engage in the fundamental tasks of query specification and refinement, search results evaluation and exploration, and knowledge discovery and management. This domain is related to a number of other important research areas, including big data [Rus11], text mining [AZ12], and visual text analytics [AdOP12]. Many large document collections exhibit the big data traits of volume, variety, and velocity, which must be effectively managed. The goal of text mining is to automatically infer structure from unstructured text, and the goal of visual text analytics is to use visualization to enable the human element of text analysis with the support of automated methods. Visual search analytics draws from advances in these domains, applying text mining within the context of the big data problems of large document collections, and focusing on a very important and far-reaching subclass of visual text analytics problem domains: search.

Figure 1 illustrates a process-oriented framework for visual search analytics, extending the traditional search framework with automatic methods for information extraction and interactive visualization to facilitate communication with the searcher. Rather than the document-centric focus that is common within the traditional search framework, a human-centred approach is taken, with the ultimate goal of supporting the searcher's knowledge discovery and decision-making activities [Hoe12]. In particular, extracting, modelling, and learning from the query and corresponding search results set provides the information upon which to base the visualization and interaction features, supporting the fundamental search tasks of crafting and refining the query, evaluating and exploring among the search results, and ultimately making sense of what was found.

While the specific approaches for information extraction depend on the details of the available data within a given document collection setting, they are generally divided into two categories: statistical modelling and machine learning. Statistical modelling focuses on inferring structure from the unstructured textual data, and range in complexity from simple term frequency calculations to more complex approaches from the domain of natural language processing [MS99] such as sentiment analysis and named entity extraction. Machine learning attempts to learn generalized models of the data, and include approaches such as document clustering or sentiment classification, or more complex methods such as topic modelling using latent dirichlet allocation [BNJ03] or various graph-based inference approaches [Mur12]. In addition, one must consider whether these approaches should be applied to the queries, the search results, or even the entire document collection. A fundamentally important step in any visual search analytics research is to choose appropriate methods for the extraction of useful information upon which to base the visual representations and provide interactive tools to aid the searcher.

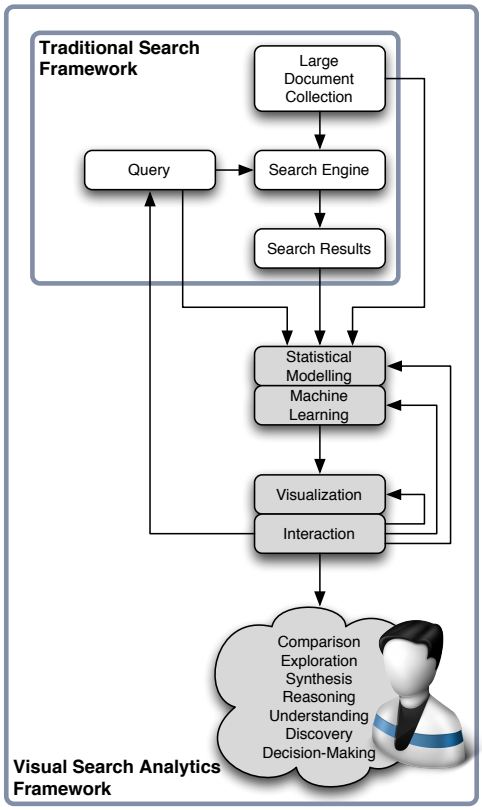

Figure 1: A framework for visual search analytics.

The visualization and interaction methods selected within any visual search analytics research may be guided by Shneiderman's information seeking mantra: "overview first, zoom and filter, then details-on-demand" [Shn96], focusing on supporting human-centric search processes [Hoe12, Hoe08]. More specifically, the information extracted may be used to provide visual overviews of the search results, as well as perhaps the query and the entire document collection; zooming and filtering may be implemented via query

refinement, faceted navigation, and/or search results re-ranking; and accessing details-on-demand for specific search results is needed in order for the searcher to examine individual documents in detail.

In the design of the visualization features, it is important to consider the fundamental principles and theories that describe how and why users perceive and interpret visual information, including the Gestalt Laws [Kof35], colour theory [Her64], and the work of Bertin [Ber83] and Tufte [Tuf01]. In addition, Pirolli & Card's information foraging theory [PC99] provides a useful basis for understanding how visualization may be used to convey helpful information to searchers as they seek to fulfill their information seeking goals. An important consideration in any work on visual search analytics will be how to effectively abstract and visually convey the complexity of the textual information to the searcher, drawing upon and contributing to the more general field of visual text analytics.

4 Research Agenda

In order to realize the potential value of visual search analytics, the principles must be explored in the development of interactive search interfaces for a number of different large online document collection settings. Potential avenues for research include searching within encyclopedias and online digital libraries, blogs and microblogs, news websites, private corporate document collections, and the web in general. These document collections each feature important differences not only in the textual data available, but also in the types of common information seeking behaviours of the searchers. While the specifics for a given document collection must be carefully studied, a general summary of such data and behaviours is provided in Figure 2.

More specifically, there is a need for the design, development, and study of visual search analytics prototypes that explore the broad range of statistical modelling and machine learning approaches for extracting meaningful information from within textual data, and the visual and interactive techniques for presenting this information to the searchers to support their information seeking tasks. By focusing on the human-centred aspects of search, query refinement can be supported, the composition of the search results can be illustrated, visual document evaluation and comparison can be enabled, search results can be filtered, re-ranked, and explored, and the cognitive activities of analytical reasoning, sense-making, and decision-making can be enhanced.

An important aspect of such research will be to conduct a well-planned series of user evaluations for each prototype at various levels of scale and complexity [Hoe09]. Whenever possible, comparisons should be made to carefully selected baseline systems representing the state-of-the-art and/or industry standard search approaches, and measurements should be taken to capture not only absolute retrieval effectiveness, but also the searchers' perceptions of usefulness and ease of use. Analyzing the time taken to complete a search task should be done with careful consideration of the specific search activity being supported, noting that the extended engagement with an exploratory task, for example, may be considered a beneficial result. The outcomes of such evaluations can be used to identify aspects that need improvement, allowing for the incremental refinement of the prototype. Successful evaluations will build confidence in the value and benefits of the combination of specific machine learning and interactive visualization approaches employed in the creation of the visual search analytics prototype.

The ultimate goal of this research agenda will be to formalize the common elements of search across multiple online document collection settings, identify the reasons for the differences, and study how visual search analytics approaches support both the common and unique elements in each search setting. This will lead to further refinement of the framework and generalization of the evaluation results across multiple search settings and task types, allowing it to be used as the starting point when developing visual search analytics interfaces for new and emerging application domains.

While this paper has proposed visual search analytics as an approach for supporting human-centred search activities in a broad range of large online document search settings, it should not be considered a silver-bullet solution to all search problems in all situations. In addition to the computational cost of modelling and learning from the search data, there is also a cognitive cost imposed on the searcher to learn how to interpret the visual representations and make effective use the interactive features. For search tasks that already have a high cognitive overhead and are frequently being performed (e.g., exploratory search in online digital libraries), searchers may be willing to accept a temporary increase in cognitive load, with the expectation that once they learn the features, the visual search analytics system will relieve the cognitive burden associated with the complex search task and provide a more effective way of finding relevant information. However, for search tasks that are already simple in nature (e.g., targeted search on the web) or infrequent (e.g., the occasional search for a document within a corporate intranet), the overhead of a visual search analytics approach may not make sense and a traditional search interface may be more readily accepted. As a result, one can expect some resistance to change if the value of the visual search analytics ap-

Data Features	Encyclopedias & Digital Libraries	Blogs & Microblogs	News Websites	Corporate Doc. Collections	The Web
unstructured text	✓	✓	✓	✓	✓
existing meta-data	✓	✓	✓	✓	✓
temporal features	✓	✓	✓	✓	✓
geospatial features	✓	✓	✓		
document relationships	✓	✓	✓		✓
author relationships	✓	✓		✓	
general topics	✓		✓		✓
focused topics		✓	✓	✓	
Search Behaviours					
targeted search	✓	✓	✓	✓	✓
exploratory search	✓		✓		✓
re-finding	✓			✓	✓

Figure 2: A summary of the common data features and search behaviours within five large online document collection search settings.

proach is not carefully measured against the current difficulty and frequency of searching within the target setting.

5 Conclusions

This paper proposes visual search analytics as subclass of visual analytics, and as an avenue for new research focused on providing greater support for the human-centred elements of search within online document collection settings. In conducting such research, one must consider that searchers in different settings have different motivations for conducting their searches, which lead to different search behaviours that must be supported. The one-size-fits-all approach of providing a simple query box and search results list provides limited support for the complexity of search tasks beyond simple fact verification. As the sizes of the document collections in these settings continue to grow, searchers increasingly face information overload problems making it more and more difficult to find the information they are seeking. The goal of visual search analytics is to leverage the power of automatic and intelligent information processing approaches, using these to provide the basis for visual and interactive support to the searcher, allowing them to conduct their search tasks in a more effective manner by supporting exploration, analysis, and sense-making among the information provided.

By exploring the application of visual search analytics within various different online document collection settings, the common themes among the different search activities will lead to the refinement of the visual search analytics framework proposed in this paper. This framework may then be applied to a wide range of search settings beyond the specific document collections discussed. These include searching within email, desktop files, image collections, and textual data within other visual analytics problem domains. The value of such a framework is that it will provide guidance from both the collection/document perspective, as well as the searcher behaviour perspective.

Further research on visual search analytics will be significant and important because of the ubiquity of textual data and the difficulty in analyzing it. In any domain where such text is important, taking a visual search analytics approach will move search beyond a simple filtering mechanism, making it a fundamental tool for analyzing and understanding the textual information. While text is everywhere, it is seldom used to its fullest potential; visual search analytics is the key to enhancing the human-centred aspects of search and unlocking the value of textual data.

References

[AdOP12] Aretha B. Alencar, Maria Cristina F. de Oliveira, and Fernando V. Paulovich. Seeying beyond reading: A survey on visual text analytics. *WIREs Data Mining and Knowledge Discovery*, 2(6):476–492, 2012.

[AZ12] Charu C. Aggarwal and Cheng Xiang Zhai, editors. *Mining Text Data*. Springer Science+Business Media LLC, Philidelphia, PA, 2012.

[Ber83] Jaques Bertin. *Semiology of Graphics*. Translated by W. J. Berg. University of Wisconsin Press, Madison, WI, 1983.

[BNJ03] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet al-

location. *Journal of Machine Learning Research*, 3(1):993–1022, 2003.

[CCP09] Christopher Collins, Sheelagh Carpendale, and Gerald Penn. DocuBurst: Visualizing document content using language structure. *Computer Graphics Forum*, 28(3):1039–1046, 2009.

[DWS+12] Wenwen Dou, Xiaoyu Wang, Drew Skau, William Ribarsky, and Michelle X. Zhou. LeadLine: Interactive visual analysis of text data through event identification and exploration. In *Proceedings of the IEEE Conference on Visual Analytics Science and Technology*, pages 93–102, 2012.

[GLK+13] Carsten Görg, Zhicheng Liu, Jaeyeon Kihm, Jaegul Choo, Haesun Park, and John Stasko. Combining conceptual analyses and interactive visualization for document exploration and sensemaking in Jigsaw. *IEEE Transactions on Visualization and Computer Graphics*, 19(10):1646–1663, 2013.

[Hea95] Marti Hearst. TileBars: Visualization of term distribution information in full text information access. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pages 59–66, New York, NY, USA, 1995. ACM.

[Hea09] Marti Hearst. *Search User Interfaces*. Cambridge University Press, Cambridge, UK, 2009.

[Her64] Ewald Hering. *Outlines of a Theory of Light Sense (Grundzge der Lehr von Lichtsinn, 1920)*. Harvard University Press, 1964.

[HHWN02] Susan Havre, Elizabeth Hetzler, Paul Witney, and Lucy Nowell. ThemeRiver: Visualization thematic changes in large document collections. *IEEE Transactions on Visualization and Computer Graphics*, 8(1):9–20, 2002.

[Hoe08] Orland Hoeber. Web information retrieval support systems: The future of web search. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence - Workshops (International Workshop on Web Information Retrieval Support Systems)*, pages 29–32, 2008.

[Hoe09] Orland Hoeber. User evaluation methods for visual web search interfaces. In *Proceedings of the International Conference on Information Visualization*, pages 139–145, 2009.

[Hoe12] Orland Hoeber. Human-centred web search. In C. Jouis, I. Biskri, J-G Ganascia, and M. Roux, editors, *Next Generation Search Engines: Advanced Models for Information Retrieval*, pages 217–238. IGI Global, 2012.

[Hoe13] Orland Hoeber. A longitudinal study of HotMap web search. *Online Information Review*, 37(2):252–267, 2013.

[HY08] Orland Hoeber and Xue Dong Yang. Evaluating WordBars in exploratory web search scenarios. *Information Processing and Management*, 44(2):485–510, 2008.

[HY09] Orland Hoeber and Xue Dong Yang. HotMap: Supporting visual explorations of web search results. *Journal of the American Society for Information Science and Technology*, 60(1):90–110, 2009.

[KAF+08] Daniel A. Keim, Gennady Andrienko, Jean-Daniel Fekete, Carsten Görg, Jörn Kohlhammer, and Guy Melançon. Visual analytics: Definition, process, and challenges. In Andreas Kerren, John T. Stasko, Jean-Daniel Fekete, and Chris North, editors, *Information Visualization: Human-Centered Issues and Perspectives*, LNCS 4950, pages 154–175. Springer, Berlin, 2008.

[KKRS13] Daniel A. Keim, Miloš Krstajić, Christian Rohrdantz, and Tobias Schreck. Real-time visual analytics for text streams. *IEEE Computer*, 46(7):47–55, 2013.

[Kof35] Kurt Koffka. *Principles of Gestalt Psychology*. Harcourt-Brace, New York, 1935.

[MS99] Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, MA, 1999.

[Mur12] Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, Cambridge, MA, 2012.

[MW09] Gary Marchionini and Ryen W. White. Information seeking support systems. *IEEE Computer*, 42(3):30–32, March 2009.

[PC99] Peter Pirolli and Stuart Card. Information foraging. *Psychological Review*, 106(4):643–675, 1999.

[Rus11] Philip Russom. *Big Data Analytics*. TDWI Research, Renton, WA, 2011.

[Shn96] Ben Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings of IEEE Symposium on Visual Languages*, pages 336–343, 1996.

[TC06] James J. Thomas and Kristin A. Cook. A visual analytics agenda. *IEEE Computer Graphics and Applications*, 26(1):10–13, 2006.

[Tuf01] Edward Tufte. *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, CT, 2001.

[War04] Colin Ware. *Information Visualization: Perception for Design*. Morgan Kaufmann, San Francisco, second edition, 2004.

[WGK10] Matthew Ward, Georges Grinstein, and Daniel Keim. *Interactive Data Visualization: Foundations, Techniques, and Applications*. A K Peters, Natick, MA, 2010.

[WKDS06] Ryen W. White, Bill Kules, Steven M. Drucker, and m. c. schraefel. Supporting exploratory search. *Communications of the ACM*, 49(4):37–39, 2006.

[WLS+10] Furu Wei, Shixia Liu, Yangqiu Song, Shimei Pan, Michelle X. Zhou, Weihong Qian, Lei Shi, Li Tan, and Qiang Zhang. TIARA: A visual exporatory text analytic system. In *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining*, pages 153–162, 2010.

[WSS10] Max L. Wilson, m. c. schraefel, and Ben Shneiderman. From keyword search to exploration: Designing future search interfaces for the web. *Foundations and Trends in Web Science*, 2(1):1–97, 2010.

[WR09] Ryen W. White and Resa A. Roth. *Exploratory Search: Beyond the Query-Response Paradigm*. Morgan & Claypool Publisher, San Rafael, CA, 2009.

[WTP+95] James A. Wise, James J. Thomas, Kelly Pennock, David Lantrip, Marc Pottier, Anne Schur, and Vern Crow. Visualizing the non-visual: Spatial analysis and interaction with information from text documents. In *Proceedings of IEEE Information Visualization*, 1995.

[YaKSJ07] Ji Soo Yi, Youn ah Kang, John T. Stasko, and Julie A. Jacko. Toward a deeper understanding of the role of interaction in information visualization. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1224–1231, 2007.