RevEx: Visual Investigative Journalism with A Million Healthcare Reviews

Cristian Felix
School of Engineering,
New York University
cristian.felix@nyu.edu

Anshul Vikram Pandey
School of Engineering,
New York University
anshul.pandey@nyu.edu

Enrico Bertini
School of Engineering,
New York University
enrico.bertini@nyu.edu

Charles Ornstein
ProPublica
Charles.Ornstein@propublica.org

Scott Klein
ProPublica
Scott.Klein@propublica.org

ABSTRACT

In this paper, we present *RevEx (Review Explorer)*, a visual analytic platform to help investigative journalists explore large sets of crowd-sourced reviews. *RevEx* uses a combination of faceted search and visualization to single out reviews with a desired combination of parameters and keywords (e.g., reviews with 1 star containing the word "food"). We developed this tool in collaboration with *ProPublica*, an independent non-profit newsroom, to help journalists look into more than one million reviews on healthcare providers from Yelp. The paper describes the main motivation behind the tool, its design rationale, and discoveries made through its use. We also discuss how the same methods can be used to explore other possible data sets and thus support to the development of numerous additional journalistic investigations.

Author Keywords

Visual Analytics; Faceted Search; Crowd-sourced Reviews

INTRODUCTION

In recent years, we have witnessed a tremendous increase in the amount of user-generated content on-line. Social network sites such as *Facebook* and *Twitter*, public forums such as *Stack Overflow*, and crowd-sourced review sites such as *Yelp*, all represent a treasure trove of data that can be mined to explore social and behavioral phenomena. Being able to look into such rich and nuanced sample of human interactions represents a unique opportunity for journalistic inquiry and for the generation of engaging and revealing journalistic pieces.

Large human-generated text data is, however, notoriously complex to analyze. Besides the scalability problem, text data is often very noisy and does not lend itself to statistical analysis and visualization. Furthermore, investigative data analysis is often characterized by a complex iterative and opportunistic process in which new questions are generated and new directions explored as more of the data is analyzed and understood.

In order to support these needs, we developed *RevEx*, a visual analytics tool for journalistic analysis of large sets of crowd-sourced reviews. *RevEx* has been developed through a collaboration with *ProPublica*, an independent newsroom of journalistic investigators, to support their analysis of 1.3 million reviews obtained from the healthcare section of *Yelp*; the large internet aggregator of business reviews.

RevEx accepts large sets of reviews and their structured data as an input and transforms them into an interactive visual interface that allows for their exploration. The interface is based on *keyword* and *faceted search*, to retrieve reviews with specific characteristics (e.g., 1 star reviews of dental practices), and carefully designed visualizations to show statistical distributions of the retrieved reviews and their content (e.g., through statistical charts and tag clouds).

While numerous methods and tools for complex statistical and visual analysis of text data exist, e.g., for topic detection and evolution [12], document clustering [7, 2], and opinion mining [11, 4], our interaction with the investigators revealed a lack of simple highly-interactive querying and sub-setting tools able to: 1) single out reviews of interest and reveal how they distribute across the available structured data and 2) permit to quickly refine generic questions into more specific ones and to easily switch to new questions when needed.

While with current tools journalists can easily do statistical and textual analysis, there is a lack of tools that allow the users to quickly slice the corpora and get new statistics and text summaries about the selected slice. This kind of approach permits to explore the data set more freely and to support the serendipitous nature of exploratory data analysis.

In the rest of the paper, we present *RevEx* in more details. We provide a brief description of related works. Next, we describe the context in which *RevEx* was born, describing the main journalistic intent that motivated its development. Then, we describe the design of the user interface and functions in details, and provide a story of how *RevEx* has been used to publish an article analyzing how people reviews doctors on Yelp. Finally, we conclude with a discussion of how the tool can be used for future journalistic inquiries with similar data sets and future plans for further development.

RELATED WORK

In the following, we present a very brief summary of works related to *RevEx*. *RevEx* draws its inspiration mainly from

two areas of research, namely: *faceted search*, the idea of exploring a collection by applying multiple filters, and visualization tools for investigative journalism.

Faceted Navigation

Faceted navigation has been studied extensively as a method to explore large collections of items (e.g., images, web pages, documents). The basic idea is to use the structured data to let the user narrow down the results through interactive filters. Hearst describes and discuss the technique at length in her book on user interfaces for information retrieval [6]. She also provides guidelines on effective user interfaces design for faceted navigation [5]. Her Flamenco system permits to explore a large set of books and art images through several hierarchical facets. *RevEx* uses the same concept but adds visualization to easily convey the frequency of each value. Ben-Yitzhak[1] adds other structured data statistics aside from the facet values like average price for each document category. *Elastic lists* [10] adds visualization to facets by mapping the frequency to the height of the of the list item. Zhang [13] adds a horizontal bar to represent the frequency of the facet value. *RevEx* also uses bars and other graphical representations to depict frequencies, but it also adds secondary information on sample size. We noticed that frequencies can be misleading when the base rate changes dramatically between the categories of a facet. For instance, an item may have 90% of negative reviews on a total of 5 reviews only, whereas another item may have 70% of negative reviews on 1000 reviews.

Visualization for Investigative Analysis

Some tools have been created to help investigative journalists explore text. Overview [2] uses a clustering and tagging approach to help journalists explore large corpora of text. Kules and Shneiderman[8] developed "Service", a faceted search tool, to conduct a study with journalists while performing exploratory search with a set of US Government websites. The journalists found the faceted version stimulating, satisfying and organized. Jigsaw [9] focuses on highlighting relationships between documents to help investigators find connections between documents and entities within the documents. Vox Civitas [4] helps journalists explore twitter data to understand people's reaction to events like the State of Union. *RevEx* focuses on the analysis of large sets of reviews to understand people's opinion about certain topics of interest and to detect reviews that contain valuable journalistic information. Its main contribution is its seamless exploration of opinion data by linking reviews with structured statistical data and by allowing flexible interactions that permit to quickly specify and edit a query.

JOURNALISTIC INTENT

With the advent of social media and crowd-sourced review sites like Yelp, it is possible to better understand and depict complex social phenomena and their impact on work people, businesses and organizations. Today, businesses and journalists can capture user feedback and emotions more holistically, rather than just looking at stars and ratings. This exploration opportunity has surfaced and gained momentum only in the last few years as businesses and journalists got hold of large text corpora of user reviews.

While businesses look at this data to get a high-level understanding of user feedback, investigative journalists are interested in much deeper mining and analysis. The primary intent of our journalist collaborators is to understand how users talk about certain businesses, and whether or not there are consistencies in user feedback for certain businesses. This also includes the detection and analysis of extreme cases and behaviors, such has doctors giving wrong prescriptions, practices where people wait for a very long time, etc. This is the first time in the data era that we have the opportunity to perform such fine grained analyses on large number of user reviews. At times, these analyses also expose the anomalous behaviors, such as privacy infringements, violation of code of conduct, or law and regulation breaches.

When we asked our collaborators why looking at user reviews is of importance to them, they replied, "*For journalists, user-generated content is important for at least two reasons. The first is to identify hot spots - those business or providers whose ratings are superior to peers and those whose ratings are inferior. Of course, any signal gained through this data needs to be independently validated and matched with other data, such as licensing data, lawsuits, etc. If I am writing about a particular doctor or hospital, the Yelp data can flag me to people who have had experiences there, the good and the bad. The other way in which the data could be useful is the ability to keyword search it, looking for individuals who mention a particular drug or a particular treatment and then contacting them. In this situation, you are not looking into a particular provider necessarily but rather specifically individuals who have used a drug or had a knee replaced. Simply using Google or review sites directly does not allow you to look for either*".

DATA PRE-PROCESSING

Many text documents, including the reviews obtained from Yelp, may have additional numeric and categorical variables associated with them. We often refer to these additional variables (except the text fields) as structured data, these variables bring context to the analysis. Typically, review data includes information about the object that is reviewed (e.g., specialty/type, geographical region, number of reviews), the reviewer and the review itself (e.g., rating).

RevEx does not make any assumptions on the data structure except that it is a collection of text segments associated with categorical data. Therefore, it is possible to use the system with any other review dataset (i.e., RateMyProfessror, Trip Advisor) or even short news. The input data from Yelp contains information about the user, the healthcare provider, the review text, user rating and date of review and the data is delivered and updated in daily batches.

Given the aforementioned relevance of the structured data, we designed *RevEx* to support not only text exploration but also exploration of the associated statistics. All the structured data that comes together with the dataset is used to create visible statistics and visual representations that can be used interactively as filters. For instance, to show how many reviews of each score exist and to allow the end-user to single out reviews with a given score.

To be able to quickly search and browse the collection, the data set is indexed the first time it is imported. In this process, we remove stop-words and use light stemming to reduce words to a common root. The system also indexes bi-grams (pairs of words) for to be shown in the interface. The back-end architecture is designed to consume the daily data, index the content and run the pre-processing modules. The data in the back-end is managed by *elasticsearch* on a single-node cluster. To further enhance its scalability, the data can also be sliced and distributed across multiple machines.

REVEX

The user interface is based on an iterative user interface design process that involved numerous interactions with the group of journalists working with the Yelp data.

The user interface is made of 5 main modules. The *Search* field allows the user to specify a search query to retrieve reviews with a given set of words. The *Timeline* displays the volume of reviews by time and it also allows the user to filter the results by date range. The *Faceted Navigation* contains the list of facets to filter the collection. The *Results View* shows the reviews that satisfy the search and the criteria specified through the facets. The *Significant Terms View* shows terms that are significant for the search. Figure shows the *RevEx* user interface.

Faceted Search and Navigation

To support exploration through structured data, *RevEx* is built based on a *faceted search* paradigm [5]. The general idea behind *faceted search* is to use the variables associated with the data sets to filter the data through interactive user actions that limit their values. For instance, in the Yelp dataset, the user can click on the *category* facet to show only reviews that belong to businesses of a given category. *RevEx* expands the basic concept of *faceted search* by enhancing the facets with carefully designed visual representations of their statistical distribution. For each value in a facet, the system shows a graphical representation conveying information about *absolute* and *relative* frequency, that is, how many reviews exist with a given value and what is their proportion in the given result set. Figure shows an example with *rating* and *categories*.

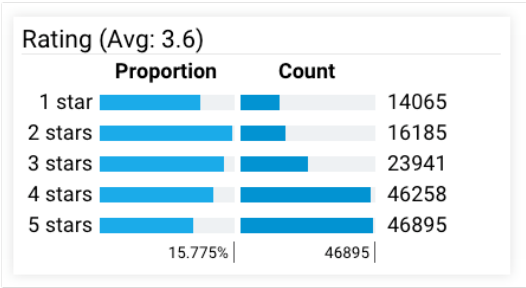

Figure 2. The rating facet for restaurants in the Yelp 2015 Challenge dataset, while the absolute count of reviews shows a predominance of positive reviews. However, the proportion, when normalized by the dataset, shows a predominance of negative reviews.

As the user navigates through the facets, a *breadcrumb* representation helps the user to quickly see which filters have been applied to the search, and also allows the user to remove filters. Given their relevance and pervasiveness in all types of reviews, we use special representations for the *date* and *rating* facets. The *date facet* is represented by a histogram with each bar mapping to either the reviews count or the proportion for each month/year combination. The user can select ranges on the histogram to filter the results by a specific date range. The user can alternate between visualizing count or proportion by using buttons on the left of the histogram. The *rating facet* is represented as a list, with each value having two horizontal bars, one representing the count and other representing the proportion. The user can click on one value (i.e., 1 star) to show only documents with that specific value.

The current implementation also shows the *category*, *state* and *provider name* facets. In these facets we use squares with area mapped to the *count* and *proportion*, respectively. Similarly, to the rating facet, the user can sort by count and proportion and click on a value to filter documents that satisfy the given value.

It is important to notice that a document in the collection may satisfy more than one filter at a time; for instance, a review about a restaurant that sells hamburgers may satisfy the *restaurant* as well as the *fast food* category filter.

Keyword search is another important aspect of the investigative process. Through interaction with our journalist collaborators, we found that search is often associated with confirmatory analysis tasks: where the journalists wants to check some hypotheses she or he may have. This often adds a credibility and trust layer to the process. Sometimes the search is also used in a more exploratory fashion, as when the journalist wants to search for a specific drug name, or a disease, to see what people talk about when the keyword is present in their review. *RevEx* allows the user to initiate an exploration through keyword search as well as a combination of keyword search and facet navigation. These two functions are indeed often used concurrently as when, for instance, one wants to search for reviews containing a specific word and then single out those that contain a negative review.

Result Visualization

After each filter or search interaction, the systems shows two types of results: a list of *review surrogates* that satisfy the search/filter criteria, and a set of terms that are relevant for the search.

The *reviews surrogates* are displayed in the *Results View*, which shows the reviews that match the query. Each surrogate is made of the name of the provider, its category, the date the review was written, the number of stars (representing the rating) and a snippet of the text in which the search terms are highlighted. This helps the journalist to understand the context of the search and easily select interesting reviews to explore further. The results can be sorted by *relevance*, *date* or *rating*.

Another important feature of the system is the presentation of *significant terms* associated with the search result in a *tag cloud*. Every time a new result is generated, that is, for every keyword search and/or facet selection, the system produces

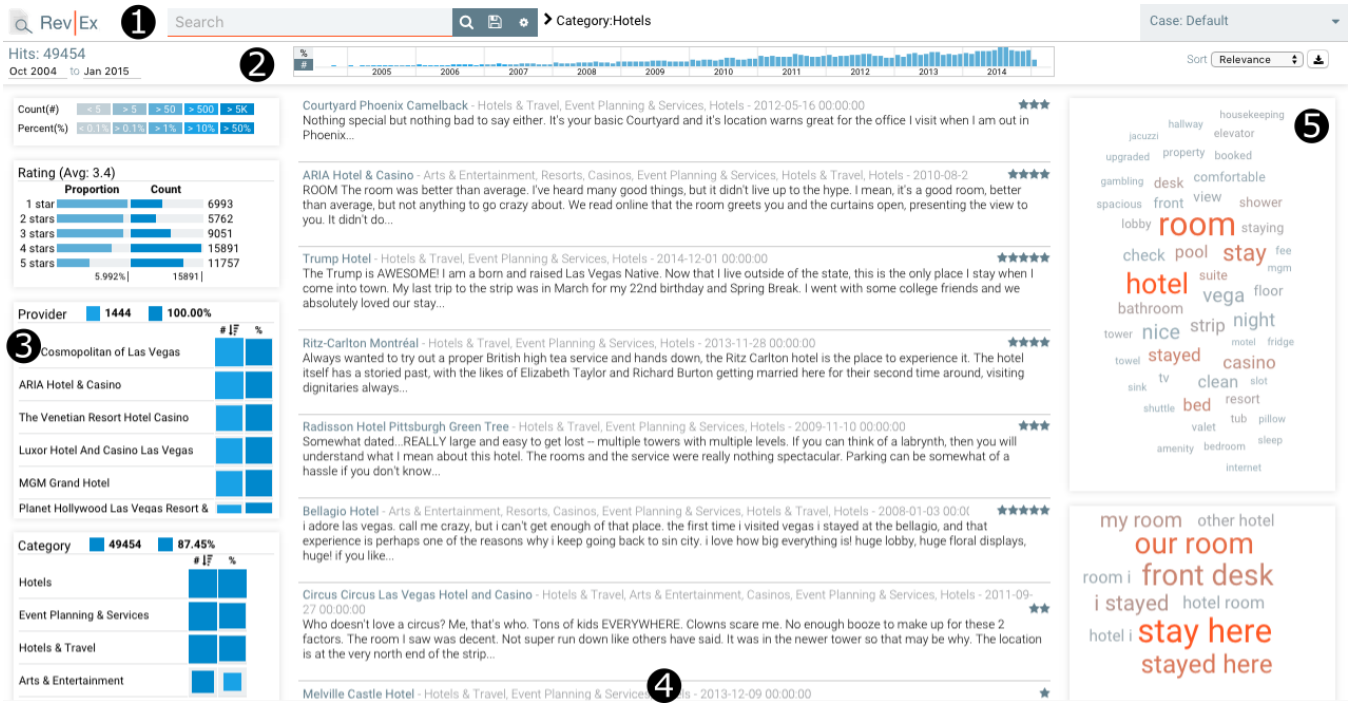

Figure 1. RevEx displaying results for hotels category on the Yelp Challenge 2015 dataset, 1) Search View, 2) Timeline, 3) Faceted Navigation, 4) Results View, 5) Significant Terms View

a list of related terms (words or bi-grams). These terms are computed using the *Google similarity distance* [3] method to single out words that characterize the result set.

The significant terms are displayed using two word clouds: one for the uni-grams and the other for the bi-grams. In each word cloud, *font size* is used to display *word frequency* and *color* is used to convey *relevance* (that is, how specific the word is to the result set). We use a color scale that goes from light blue to a very saturated orange, where orange means high relevance.

This representation has two main purposes. First, it permits to get the gist of people's opinion relative to the current result set. Without such summary, the end-user can get a sense of the result set only by reading a large number of reviews. Second, it provides *hints* on potentially interesting words associated with a given search that the journalist may have an interest to investigate further. In this case, the user can also click on a word in the tag cloud to automatically add that word to the search.

One simple example in which this visual representation is useful is when one is trying to understand what leads people to give negative ratings to a healthcare provider. The user can select 1-star reviews and look at the tag clouds to see that most of the terms are associated to waiting time, which seems to be a major cause of frustration.

Data Provenance

Some investigative explorations can take days or even weeks to be performed, and it is important for the journalist to be able to keep track of interesting documents and searches realized during previous explorations. To support this process, *RevEx* allows the user to save the current state of the search and to provide a label to it. The user can also save interesting documents or facet configurations. To organize all this information, we follow a case-based model with three levels of detail: named *case*, *state* and *document*. Case refers to the broadest umbrella in which the user has a primary purpose defined. States refer to an intermediate facet configuration. And document refers to the actual text the user wrote.

The user can access the content of previous cases or create new ones by using a panel on the top-right of the interface. A list of states and documents is always available in this section to retrieve previous cases. The user can also download results and cases into a spreadsheet so that it can be imported into other applications for further analysis, printed or sent via email.

DISCOVERIES AND IMPACT

In this section, we present a synopsis of the discoveries our collaborators made while exploring the data using *RevEx*. We also highlight how the exploration evolved and what impact the analysis had on the outcome. Further details of these discoveries are provided in one recent National Public Radio (NPR) article that one of our collaborators published as an outcome of his initial experimentation with *RevEx* [1].

In the initial state of the analysis, the journalists were interested to get a sense of the overall corpus. A high-level analysis of the dataset distribution shows that overall there are

[1] http://www.npr.org/sections/health-shots/2015/08/06/429624187/on-yelp-doctors-get-reviewed-like-restaurants-and-it-rankles

many more 5-star (positive) reviews than any other rating. That is, people tend to give more positive than negative reviews. The overview also shows that although Yelp's health reviews date back to 2004, more than half of them were written in the past two years.

A second finding, obtained by filtering the providers according to number of reviews, is that providers with the highest number of reviews generally have poorer ratings. By looking at some of the reviews in the result view, it was also found that mostly elaborate and long reviews typically correspond to lower ratings, and vice-versa.

The next discovery made with *RevEx* was about the distribution of user reviews across health providers. After selecting reviews with only 1 star, the provider called *Western Dental*, a chain of low-cost dental clinics, jumps out as the one with most (negative) reviews. The visual representation based on *absolute* and *relative* frequency also shows that not only there are lots of reviews for this provider, but also that the proportion of negative reviews is the highest among the whole set. When looking at the reviews and the tag cloud, the words "horrible", "wait", "waiting", "worst" stand out. When adding some of these words to the current search one can see that they are extremely frequent: about 1250 of its 3000 reviews used the words "wait" or "waiting" and about 15% of them, the word "worst". When looking at the *location* facet one can also see that these negative reviews are equally distributes across different regions.

A similar kind of analysis was performed by looking at all providers with negative ratings and again what was found is that negative reviews are very often due to problems with customer service and staff behavior. Slicing the data to show only reviews with 1 star revealed that for several low-rated providers, the responses included references to sloppy customer care. For instance, several customers complained about phone calls not being returned and emails seeking comment. In many ways, consumers on Yelp rate health providers in the same way they would do with other businesses like restaurants: on how they feel they have been treated.

During this kind of analysis, the journalist also discovered the word "HIPAA", which refers to the "The Health Insurance Portability and Accountability Act". By looking at reviews that contain this word it was possible to identify cases in which a patient protests against privacy infringement as well as providers mentioning not being able to respond to avoid such infringements.

Through this analysis it was found that in some cases doctors, dentists and other providers threaten or even file lawsuits against people who post negative reviews on Yelp or against Yelp itself. One such example is a lawsuit filed by *Brighter Dental* on a patient who gave a 1-star rating to them on Yelp. The lawsuit was then dropped when the patient took her story to the media.

These kind of analyses are difficult to do without a flexible investigation tool. *RevEx* made these analyses possible by providing the search and faceted navigation mechanism, allowing the journalists to quickly switch from one question to another and by allowing to easily refine each search through multiple parameters.

With some very few exceptions all the finding published in the NPR article were extracted using *Revex*. The system was the starting point of the investigation and it supported the analysis throughout its lifetime. The journalists were able to understand the corpora and find not only interesting statistics but also interesting reviews and easily switch from a high level analysis to very specific details. Reviews and searches were saved using the data provenance features of the system, and the information was later verified and enriched by contacting providers and users or researching in third-part venues.

CONCLUSION AND FUTURE WORK

In this paper, we presented *RevEx*, a visual analytic platform to explore and analyze millions of healthcare reviews obtained from Yelp. The work has been done in collaboration with a team of journalists for *ProPublica*, a non-profit investigative journalism newsroom based in New York City. The tool is designed with journalistic intents and requirements in mind and it supports search and faceted visual navigation. Our collaborators used *RevEx* to understand how customers talk about certain healthcare providers, and whether there are consistent patterns or anomalies in their functioning. The first set of discoveries are presented in an article published at NPR. Some of these discoveries could not have been possible without a specialized tool for text analysis, like *RevEx*.

Exploring crowd-sourced data has proven highly relevant, and it is a way to detect signals that should be investigated further. It is important to mention that this type of data alone is not a proxy for quality. One should never categorically affirm that a doctor is good or bad only based in his or her rating on Yelp. This is the reason why *RevEx* shows specific text to provide context but does not attempt to provide definite answers. The tool should be considered as a way to generate hypotheses and ideas that need to be verified. The best way to use these tools is to identify people and services that could be interviewed to gain more detailed information for a particular story.

As part of our future work, we are collaborating with other media outlets, journalists, data scientists, survey and publishing agencies, who are interested in analyzing similar data sets with similar intents. *RevEx* is highly modular and can be used with almost any text corpus that has a similar structure; which happens to be true for many review data sets. An online demo of the platform with the Yelp Academic Dataset, a public version of the Yelp data, can be found at: http://nyuvis.github.io/revex/.

REFERENCES

1. Ben-Yitzhak, O., Golbandi, N., Har'El, N., Lempel, R., Neumann, A., Ofek-Koifman, S., Sheinwald, D., Shekita, E., Sznajder, B., and Yogev, S. Beyond basic faceted search. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, ACM (2008), 33–44.
2. Brehmer, M., Ingram, S., Stray, J., and Munzner, T. Overview: The design, adoption, and analysis of a visual document mining tool for investigative journalists. *IEEE Trans.*

Visualization and Computer Graphics (TVCG / Proc. InfoVis) 20, 12 (2014), 2271–2280.

3. Cilibrasi, R. L., and Vitanyi, P. The google similarity distance. *Knowledge and Data Engineering, IEEE Transactions on 19*, 3 (2007), 370–383.
4. Diakopoulos, N., Naaman, M., and Kivran-Swaine, F. Diamonds in the rough: Social media visual analytics for journalistic inquiry. In *Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on*, IEEE (2010), 115–122.
5. Hearst, M. Design recommendations for hierarchical faceted search interfaces. In *ACM SIGIR workshop on faceted search*, Seattle, WA (2006), 1–5.
6. Hearst, M. A. *Search User Interfaces*, 1st ed. Cambridge University Press, New York, NY, USA, 2009.
7. Hetzler, E., and Turner, A. Analysis experiences using information visualization. *Computer Graphics and Applications, IEEE 24*, 5 (2004), 22–26.
8. Kules, B., and Shneiderman, B. Users can change their web search tactics: Design guidelines for categorized overviews. *Information Processing & Management 44*, 2 (2008), 463–484.
9. Stasko, J., Görg, C., and Liu, Z. Jigsaw: supporting investigative analysis through interactive visualization. *Information visualization 7*, 2 (2008), 118–132.
10. Stefaner, M., Urban, T., and Seefelder, M. Elastic lists for facet browsing and resource analysis in the enterprise. In *Database and Expert Systems Application, 2008. DEXA'08. 19th International Workshop on*, IEEE (2008), 397–401.
11. Wanner, F., Rohrdantz, C., Mansmann, F., Oelke, D., and Keim, D. A. Visual sentiment analysis of rss news feeds featuring the us presidential election in 2008.
12. Wei, F., Liu, S., Song, Y., Pan, S., Zhou, M. X., Qian, W., Shi, L., Tan, L., and Zhang, Q. Tiara: a visual exploratory text analytic system. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM (2010), 153–162.
13. Zhang, J., and Marchionini, G. Coupling browse and search in highly interactive user interfaces: a study of the relation browser++. In *JCDL*, vol. 4 (2004), 384–384.