Big Text Visual Analytics in Sensemaking

Lauren Bradel, Nathan Wycoff, Leanna House, Chris North
Department of Computer Science and Department of Statistics, Virginia Tech
Blacksburg, Virginia, USA
{lbradel1, nathw95, lhouse, north}@vt.edu

***Abstract*— Learning from text data often involves a loop of tasks that iterate between foraging for information and synthesizing it in incremental hypotheses. Past research has shown the advantages of using spatial workspaces as a means for synthesizing information through externalizing hypotheses and creating spatial schemas. However, spatializing the entirety of datasets becomes prohibitive as the number of documents available to the analysts grows, particularly when only a small subset are relevant to the tasks at hand. To address this issue, we applied the multi-model semantic interaction (MSI) technique, which leverages user interactions to aid in the display layout (as was seen in previous semantic interaction work), forage for new, relevant documents as implied by the interactions, and place them *in context* of the user's existing spatial layout. Thus, this approach cleanly embeds visual analytics of big text collections directly into the human sensemaking process.**

Keywords—sensemaking, interaction design, visual analytics

I. Introduction

While professional analysts are undoubtedly inundated with "too much data," it is crucial to remember that this problem plagues everyday users as well. In this paper, we consider the challenge of exploratory data analysis using large document collections such as scientific literature, news, or the world-wide web. Unfortunately, users are notoriously bad at formulating explicit queries in such tasks [31]. We applied multi-model semantic interaction [13] to this problem to allow the system to passively construct queries through interpreting user interactions, filter returned web results to those which are most relevant to the user's interests, spatially arrange them according to similarity, and visually indicate document and text saliency. All of this is done in a single spatial workspace, thus placing foraged information *in context* of the existing spatial layout, allowing the user to continue synthesizing documents without the need to context switch.

Take, for example, a researcher conducting a literature review on a new research topic. This researcher likely does not know the taxonomy of this topic, which could easily span multiple sub-fields, all of which may use slightly different language in describing similar topics. In such a case, it is difficult to easily gain a comprehensive understanding of the topic through explicit querying. How can the researcher be confident that she has not overlooked important papers that fall in the intersection between known components of the topic? We propose that multi-model semantic interaction takes steps to alleviate this concern. Instead of repeatedly typing queries, reading abstracts, and curating results for later synthesis, the researcher can conduct all foraging actions directly from her synthesis space. For example, if she finds two highly relevant papers that are from different aspects of her topic, she can overlap these two documents in order to retrieve any relevant papers that combine aspects of both papers. This can help to "fill in the gaps" in the researcher's literature review and hopefully allow her to avoid missing crucial information.

Another common exploratory data analysis task is investigating current events in the news. In order to gain a comprehensive understanding of a particular topic, users must frequently consider multiple sources for information, not only to fill in knowledge gaps, but also due to the differing opinions of individual journalists. For example, a user may be interested in tracking political candidates for an upcoming election. They may begin their analysis by searching for a specific candidate's name. This query could return articles from major news networks, opinion pieces, local news outlets, personal blogs, or satirical news websites. The user is then tasked with injecting feedback to steer the underlying user interest model to create a subset of documents that cover various aspects of the candidate, their campaign, as well as comparisons to additional political candidates. While a literature review may be focused enough to pull from specific digital libraries with standardized formatting (e.g. IEEE, ACM), exploring a news topic requires pulling articles from a wider number of sources with varied formatting and advertisements. This presents additional challenges in terms of article parsing.

These scenarios demonstrate an opportunity for big data text analytics. Previous work has shown that users leverage implicit query formation to retrieve relevant information [13], but this technique has not been applied to such a large scale of data. Dealing with vastly different levels of data scale (e.g. a small curated working set of documents vs. the internet) presents a set of research challenges in terms of performance, model coordination, interaction design, and visual encodings. We discuss these challenges and present an extension to our existing visual analytics tool prototype, StarSPIRE [13], that enables these aforementioned scenarios to be performed through integration with external web search services such as Bing and IEEE Xplore.

We present a method to integrate information retrieval with information synthesis by presenting foraged results from external search services *in context* of the user's current analytical state via a spatial "near = similar" metaphor. Other systems tend to treat information retrieval as a separate task, but we intend to remove the intermediary steps to create a *cohesive and integrated* sensemaking environment that does not force users out of their "cognitive zone" [29] of

978-1-4673-7343-2/15/$31.00 ©2015 IEEE

information synthesis. Instead of exiting their synthesis space to execute a query from external data sources, judge results, and import them to a workspace, these actions can be done directly and automatically from the spatial workspace with results being placed within the existing schema [Figure 1].

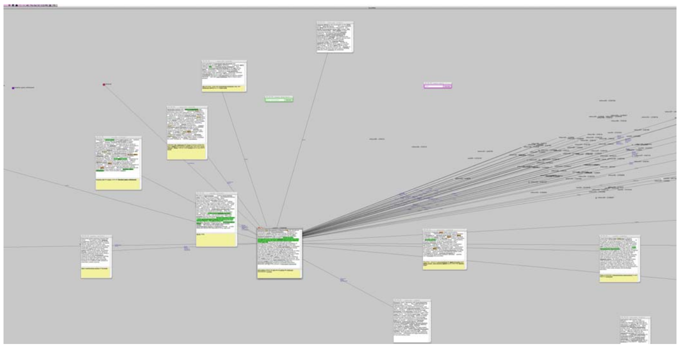


Figure 1. StarSPIRE: analyst's workspace during a literature review task.

Furthermore, working within a spatial metaphor allows users to directly manipulate data at vastly different levels of scale [21]. The user is able to focus on a small working set of documents while having the entirety of large external sources at their fingertips. We present a means to steer the underlying models at these varying levels of scale while addressing the challenges associated with this undertaking.

Finally, we present an updated visualization pipeline and the associated implementation of StarSPIRE that addresses the multi-scale nature of these exploratory data analysis tasks. We discuss the system architecture, model coordination, interaction mapping, and visual encodings. A critical component of this work is the connection to existing external search services, enabling visual analytics methods such as semantic interaction to exploit the benefits of successful information retrieval systems. Through this work, we have achieved near-real-time big data analytics for text-based sensemaking in exploratory data analysis tasks.

II. Related Work

A. *Spatializations for Sensemaking*

Prior research has highlighted the utility of spatializations for text analysis [3, 5, 12, 23, 26, 35, 42, 49, 54, 55]. Large spatial workspaces have been found beneficial in affording a flexible workspace that allows users to externalize knowledge and create semantic schemas [4]. However, this knowledge externalization is typically achieved through parametric interactions (e.g. [36]), many of which require users to go outside the spatial metaphor by manipulating control panels [21]. Furthermore, parametric interaction does not easily scale to big data problems. In unstructured text data, dimensions map to the terms or entities contained in the documents. Thus, the dimensionality of the data grows extremely large as the number of documents increases. Aside from navigating through the flood of dimensions, altering multiple models becomes extremely tedious. If multiple models are used for layout and/or retrieval, the user must update the dimensional weights or parameters for each model and potentially at multiple levels of scale. To remove this redundancy, we contain the interaction within the spatial metaphor and translate interactions into parametric feedback.

For tools that allow users to stay within the spatial metaphor, parametric interaction is still common. For example, Dust & Magnet allows users to manipulate spatial landmarks to adjust the spatialization of multi-variate data. However, these landmarks are attributes of the data, not points themselves. The users only have control over the parameters in the space. Similarly, VIBE allows users to designate keywords as spatial landmarks [42]. In MSSI, users can designate specific data points as spatial landmarks. These landmarks attract other data points (e.g. documents) based on the high-dimensional data instead of a single attribute or dimension. For text data, this enables users to focus more on the high-level semantics of the document contents rather than merely on specific individual keywords.

Systems exist which allow users to directly manipulate data points and interpret this feedback via a dimensionality reduction algorithm to generate a new view that better reflects the user's understanding of the data [15, 24, 34]. These methods inherently suffer from scalability issues. Users expect a quick interaction-feedback loop in order to remain in their "cognitive zone" [29], but calculations on thousands, let alone millions, of data points take from minutes to hours to complete. It is more practical to break the problem into multiple levels of scale and perform dimensionality reduction on a subset of a much larger data set, using information retrieval techniques to add additional information to the workspace.

B. *Semantic Interaction*

Semantic interaction serves as means for analysts to work with data within a spatialization instead of altering algorithms or the raw data [20, 21]. The concept of direct manipulation for visual analytics is an evolution of direct manipulation for information visualization [46]. This is particularly important when the analyst is a non-expert in the layout algorithm(s).

Semantic interaction can be viewed as a form of visual-to-parametric interaction (V2PI) [34]. This type of interaction involves mapping user interactions to algorithmic parameters. For example, in [34], users are presented with a MDS layout and are able to impart feedback on the layout (through highlighting or moving data points), which then iterates to generate a new spatialization that better matches the user's understanding of the relationships within the data. Similarly, DisFunction [15] converts user interaction on a two-dimensional spatialization into feedback on the high-dimensional data, generating a spatialization that better matches the meaning imparted by the user.

Typograph [22] uses varying levels of data abstraction to visualize large text corpora. Users can drill down to see the details of documents at different levels of detail. The MSI technique implemented in StarSPIRE, in comparison, addresses the scalability challenge by constantly updating a small working set of documents. Documents in StarSPIRE are either open or closed, whereas Typograph extracts topics, keywords, and document snippets.

While current forms of semantic interactions have shown to be successful, they are limited in the number of data items they can handle simultaneously (typically less than 1000). Thus, semantic interaction alone is not adequate for tackling the challenge of big data.

It is not practical to display thousands or millions of documents using any of the above models. In addition to a poor interaction-feedback loop time, the user would not be able to distinguish the points. Instead, many researchers have turned to topic modeling to give the user an overview of the topics and their distribution in the dataset. This can be a good method for establishing a starting point for analysis in addition to gaining an overview of the data. However, through an informal requirements analysis done with intelligence analysts, we found that they frequently have a specific topic to research or even a handful of "starting point" documents.

Therefore, we found it to be more practical to store the initial data in a database and use information retrieval algorithms to fetch additional documents for the user. Similarly to how semantic interaction helps to steer the layout model, it can be used to steer information retrieval models by changing either the model itself, input parameters, or both. A multitude of information retrieval algorithms and models exist that could be used in a semantic interaction context. Latent Dirichlet Allocation (LDA) uses probabilistic topic modeling to group similar documents [11]. Latent Semantic Indexing (LSI) uses a method similar to principle component analysis to reduce the high dimensional data (in this case, the term-document matrix), and then constructs a query into the lower-dimensional space using a set of terms [33]. Additional potential models include probabilistic relevance model, Bayesian logistic regression, boolean models, and vector space models [25].

C. *Information Retrieval*

The information retrieval aspect of this work is closely related to content-based recommendation systems [6, 43]. These systems track user interests to build a profile of a user and their interests in order to query for additional relevant items. The data involved is often high-dimensional, typically from facets of an item or associated metadata (e.g. item type, category, production information, genre). However, these systems typically rely on pre-defined characteristics, whereas we are operating on unstructured text data models that are capable of having entities added or removed dynamically.

Additionally, this work is closely related to query-by-example systems, which differ from context-based recommendation systems (e.g. [47, 48]) in that query-by-example systems use a set of user-defined query objects whereas recommendation systems aggregate recommendations over all (or a recent selection of) user selections. Query by example systems have enjoyed a wide implementation across data types [38], from unstructured text documents [8], to multimedia [16, 30], to musical selections [28].

Systems such as Adaptive Information Retrieval [10] use relevance feedback to augment future retrieval requests to return results that are better tuned to the user(s). Other systems use visualizations to construct queries (e.g. geographical and temporal bounding [2], expressive constructors [19, 37], dynamic control panels [50], dynamic query interfaces [1, 27, 47]). However, these mechanisms still often fail to place results in context of existing retrieved results, which is important for maintaining situational awareness [7, 52].

Attempts have been made to visualize information retrieval results (e.g. term distribution charts [32], self-organizing semantic maps [39], hierarchies [17], collages [18], word trees [53]), but these techniques have not been widely adopted. Information retrieval results are typically visualized as a ranked list of results [40, 41]. Presenting results in this format is suitable for targeted queries where the user may view a handful of results at most (e.g. a web search for a specific culinary recipe). However, when the user is presented with hundreds of viable documents worth reading (e.g. an intelligence analysis task) that relate in complicated, intricate, and fuzzy ways, a linear list becomes less than ideal [14].

Work by Ruotsalo et al. has demonstrated the use of direct manipulation to influence information retrieval algorithms [44]. User interactions within a radial topic spatializations were used to infer possible user intent to tune search results, working on the principle that searches evolve incrementally [51], similarly to the incremental formalism seen in sensemaking and spatial organization [45]. They found that these interactions did not replace the need for conducting explicit searches, but that the users in the condition that allowed for the use of the spatial interface performed better than those who did not have this technique available.

Other systems provide mechanisms for visualizing search results beyond the typical ranked list (e.g. term distribution charts [32], self-organizing semantic maps [39]), but these methods do not provide the nuanced spatial interactions that the Ruotsalo system does. While ranked lists are well-suited to narrow and specific searches, they may not be as well suited for exploratory data analysis. For example, conducting a literature review requires exploring multiple facets of a topic. A simple ranked list of results does not yield insight into documents that are mixtures of different topics.

III. Research Challenges

Creating an analytical tool that facilitates exploratory data analysis across multiple models operating at vastly different scales of data comes with a substantial set of research challenges. These include system architecture considerations, interaction and visualization design, and data scale concerns.

A. *Performance*

Dealing with information retrieval requests on big data inevitably requires researchers to address performance issues, particularly in terms of performance and result accuracy.

1) *Speed*

A quick interaction-feedback loop is critical for keeping users engaged in their analysis. As such, the coordinated models (information retrieval, document relevance, and visualization) ought to be optimized to maintain real-time interaction. Several avenues could be chosen and/or combined to facilitate this. As this research is concerned primarily with interaction and visualization design, multiple external existing

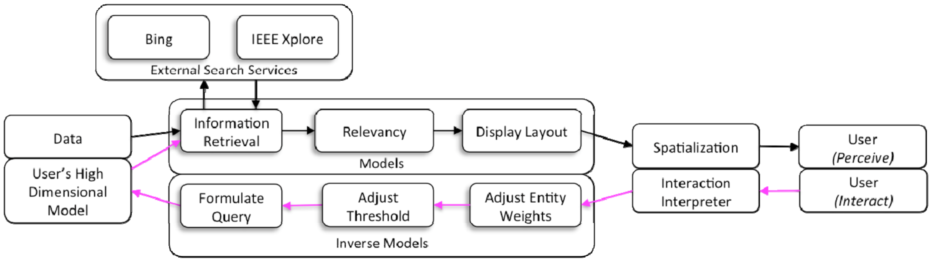


Figure 2. Visualization pipeline indicating web integration for big data analytics, controlled through a single spatialization.

retrieval APIs were leveraged in this extension to StarSPIRE. Additionally, web scraping and entity extraction can be done progressively to give the user an approximate set of document results while processing the remainder in the background.

2) *Accuracy*

Precision and recall are important aspects of any information retrieval task. The quality of retrieved documents can be evaluated objectively as well as subjectively. The information retrieval model is primarily responsible for the objective quality of retrieved results. However, these results can be tuned to the user's interests to provide subjectively better results. Using the user's interest model, the retrieval and data relevance models can be adjusted to execute nuanced queries and filter in documents that the user is more likely to find pertinent to their analysis. It is important to consider how these two notions of data relevance (user's opinion and web search opinion) compliment or contrast each other. Systems can gain insight into the user's perception of the quality of results through relevance feedback. The downside of exploiting existing external retrieval engines is the limited API for specification of user interest. Thus, we use a multi-level approach that follows the retrieval with a more detailed relevance analysis based on the user model.

B. *Data Storage*

When dealing with data on a very large scale, it is important to consider how the data and retrieved results should be stored. Possible storage options include web hosting, cloud architectures, databases, or local memory. These obviously have varying limitations in terms of the amount of data that can be stored and associated retrieval speeds. For this extension to StarSPIRE, the source dataset is left on the web (existing websites able to be accessed by a search engine) while a small working set of documents is stored in local memory. This enables us to exploit the cached storage methods used by major search engines. Previous iterations of StarSPIRE have connected to existing databases containing tens of thousands of documents while also maintaining a working subset in memory. Ultimately, the design decision for data storage should appropriately match the intended dataset in order to ensure optimal retrieval speeds.

C. *Interactions*

Interactions must be carefully designed to best match the user's current analytical reasoning process. This task grows complicated when interpreting interactions across multiple models. For example, how can the system differentiate between searching on the existing set of documents displayed in the workspace and searching over the entire external dataset? How would such an intention be detected? This is a difficult question to answer, particularly because there are individual differences between users and what strategies they employ in a spatial document analysis tool. We chose to search all data repositories simultaneously unless explicitly specified by the user. The user is given the option to toggle external databases on and off if they wish to restrict their interactions to what is currently on the display.

Previous work with StarSPIRE has tested when to launch information retrieval requests. We first required users to explicitly request additional information via a query button, then altered the interface to execute such requests after every interaction. Continually interpreting and acting on interactions removes the need for users to step out of their synthesis process to explicitly forage for information.

D. *Visual Encodings*

Extensive work has been done to tune the visual encodings in StarSPIRE, but adding large-scale information retrieval introduces new aspects of feedback that may be of interest to the user. Such facets include, but are not limited to, the novelty, recentness, and relevance of the retrieved results. Currently, all previously mentioned encodings are carried over into this iteration of StarSPIRE. It is important to consider how this visual feedback can be integrated with existing visual encodings in a manner that clearly conveys system feedback to users in an easy to interpret manner.

E. *Foraging and Synthesis Integration*

In order to keep users focused on their sensemaking task and avoid context switching, it is important to enable foraging and synthesis actions in the same workspace. If this is successfully accomplished, synthesis actions (e.g. highlighting) can drive information foraging and foraging actions can drive information synthesis (e.g. clustering documents). We have chosen a spatial layout where document proximity indicates similarity and documents are visually encoded to indicate data relevance. As new documents are added to the workspace, they are mapped to the current visual encodings and arranged according to their similarity to existing documents in the workspace. Thus, new documents are placed *in context* of the documents already in the workspace.

Thought must be given to how new documents are presented to the user, particularly to how users are alerted to their presence. StarSPIRE picks random initial positions for documents, which then move according to the weighted force-directed layout to a stable state. We have found that this strikes a balance between blatantly interrupting the user and slipping in unseen. Other mechanisms for keeping foraging and synthesis in context of one another should be investigated.

Additionally, foraging actions (e.g. information retrieval, entity extraction, relevancy evaluation) should not interfere with the user conducting synthesis actions. Ideally, foraging actions should be done in the background without the user having to wait for the system to process query requests.

IV. System Description

StarSPIRE [13] is a visual analytics tool prototype that provides users with a spatial workspace to view and arrange documents, facilitated by a modified force-directed layout. All interactions are interpreted and processed sequentially through a series of models: layout, document relevancy, and information retrieval [Figure 2]. Within documents, extracted terms are underlined and highlighted according to the feedback users have given the system through natural actions such as highlighting text, writing notes, or moving documents [Figure 3]. The layout algorithm places similar documents closer to each other and emphasizes terms with large weights and entities that co-occur between documents [Figure 4]. Documents are arranged using a node-link diagram and documents can be shown as closed nodes or as open text windows. To avoid a cluttered workspace, edges linking documents (based on entity co-occurrence) are only shown radiating from the currently selected node or document. We constructed the set of interactions available through working with and observing real-world analysts who offered usability feedback in informal and formal test settings.

StarSPIRE implements the Bing Search API and the IEEE Xplore digital library in order to expand its corpus to include innumerably many documents from the internet. The Bing API allows the client to send a query, and receive either JSON or XML representation of the Search Engine Results Page (SERP). IEEE Xplore is programmatically accessed, and its HTML parsed in accordance with its terms of use. The goal of access to these sources is to gather relevant text from web pages rich with images, colors and video. Future work includes the integration of multimedia content, but this work is limited to the text content of each page.

Queries slated for execution by the StarSPIRE Webscraping Module (WM) can be created as a result of either implicit or explicit interactions by the user. Explicit interaction can include use of the search feature to type a query, for example “Computer Science”. Alternatively, a query can result from implicit action, such as a user “combining” two documents by dragging them together. For example, if two documents on the subject of the murder of Nemtsov, a Russian political leader, were combined, the query sent to the WM might look like “Russia Putin Nemtsov Murder Opposition”. These two kinds of interaction are treated the same by the WM. In the case of Bing, part of the query indicates which source type should be searched for (i.e. full web, news, shopping).

If Bing results are requested by the user, the WM will then send the query to Bing servers and receive the SERP JSON. Currently, the WM performance with News sources considerably outperforms the full web. Next, the SERP JSON is parsed to store information about each news article, including its title and URL. The HTML is extracted from the URL and parsed for content. If the URL returned from Bing leads to anything other than a standard web page (i.e. PDF, PowerPoint), it is ignored. Parsing of such documents remains as future work. The process of parsing the HTML into plain text is considerably more reliable for News articles, as they tend to be more similarly and simply structured.

If IEEE Xplore results are requested by the user, the WM will subsequently navigate to the results page of the relevant query on Xplore. This approach was taken as a proof of concept for connecting to specific digital library search services through its generic search UI. It will extract and parse HTML from the URLs of each of the links on the results page. As with Bing, weights on tokens are not implemented for this source. Next, relevant information is extracted from the HTML. This process is more accurate with the IEEE source than with the Bing source, as each page on Xplore has the same HTML architecture, while Bing has the potential to return documents from vastly different web sites with each query. As such, a tailor-made parsing system is used for Xplore HTML.

When the text is successfully extracted from the HTML, the text is then parsed for entities. New entities are attached to the document in which they were found, and then the rest of known documents are searched for the new entities in order to relate them to one another. Whether the user requested Bing or IEEE, the information stored about the online documents is transferred into the main data structure for StarSPIRE, which the system subsequently processes before it is displayed to the user. These documents are analyzed for relevance to the user by StarSPIRE’s relevancy model to create the working set of documents shown in the spatial workspace. This means that the documents are first gathered using Microsoft’s/IEEE’s relevancy scheme, and then filtered to a subset according to what StarSPIRE believes the user is interested in through the relevancy model. Finally, documents are positioned on the screen where they fit in with the existing spatial layout model, placing the search results in context. The document nodes take on visual encodings based on the current user interest model.

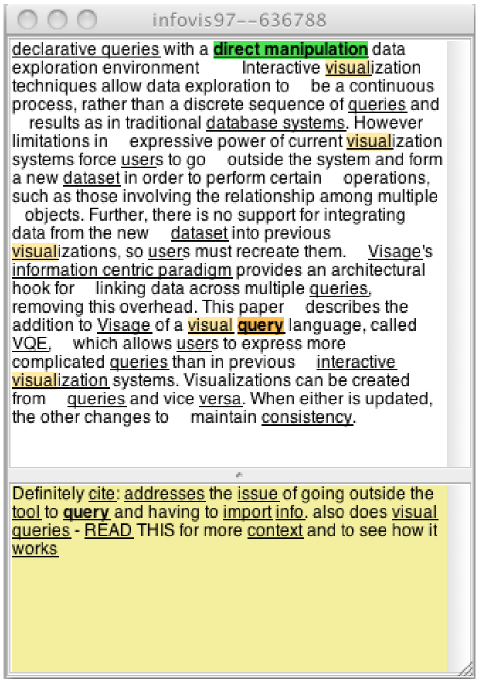
The retrieval model differs considerably from the relevancy model. The information retrieval model is implemented by the external search service (such as Bing or IEEE Xplore) and operates as a black box, using a set of unknown heuristics to rank the retrieved results. Conversely, the relevancy model is directly controlled by StarSPIRE and is constantly tuned to the user’s current interests. By sorting the results retrieved from the black box by StarSPIRE’s internal relevancy model, we can attempt to balance what the external system deems to be relevant and what the user is interested in. It is likely that these two models will have different rankings of the top relevant results. This combination of models may also serve to relieve the cognitive tunneling issue previously observed with StarSPIRE by retrieving results that an outside source (e.g. web-based search engine) believes to be relevant instead of only honing in on the user’s narrow focus. In practice, we have observed that the relevancy model naturally limits the number of documents displayed to a few hundred by pruning off documents that fall below the current relevance threshold.

V. USE CASES

Using StarSPIRE, we successfully completed the two scenarios mentioned in the introduction: literature review and investigative journalism.

A. *Literature Review*

Given that this paper is situated at the intersection of visual analytics, information visualization, and information retrieval, we used StarSPIRE to find additional related work regarding information retrieval from the information visualization and visual analytics communities. This is quite a broad task, making it a good example of exploratory data analysis. In order to ensure that the analysis would not be biased by recently published papers at conferences the analyst had attended, we restricted the dataset to paper abstracts from 1995 to 2009 from the IEEE Information Visualization (InfoVis) and Visual Analytics Science and Technology (VAST) conferences. This resulted in 454 unique paper abstracts, which were processed with LingPipe [9] to extract entities.



declarative queries with a direct manipulation data exploration environment Interactive visualization techniques allow data exploration to be a continuous process, rather than a discrete sequence of queries and results as in traditional database systems. However limitations in expressive power of current visualization systems force users to go outside the system and form a new dataset in order to perform certain operations, such as those involving the relationship among multiple objects. Further, there is no support for integrating data from the new dataset into previous visualizations, so users must recreate them. Visage's information centric paradigm provides an architectural hook for linking data across multiple queries, removing this overhead. This paper describes the addition to Visage of a visual query language, called VQE, which allows users to express more complicated queries than in previous interactive visualization systems. Visualizations can be created from queries and vice versa. When either is updated, the other changes to maintain consistency.

Definitely cite: addresses the issue of going outside the tool to query and having to import info, also does visual queries - READ THIS for more context and to see how it works

Figure 3. Zoomed in document from the literature review scenario.

The analysis began with a simple query for "retrieval." This search resulted in finding multiple relevant paper abstracts that served as starting points for further investigation. As the spatialization evolved, notes were added to documents to make connections based on common themes. This helped to link paper abstracts that used slightly different terminology to describe similar concepts, such as "iterative query refinement" and "visual query language" [Figure 3]. After 75 minutes of analysis, a final set of thirteen previously unidentified documents were selected as being highly relevant and worth citing in this paper. These documents were then rearranged by overarching topic to make re-finding by topic easier [Figure 4].

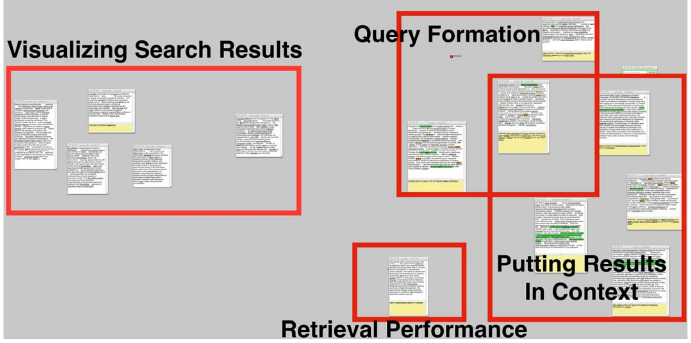
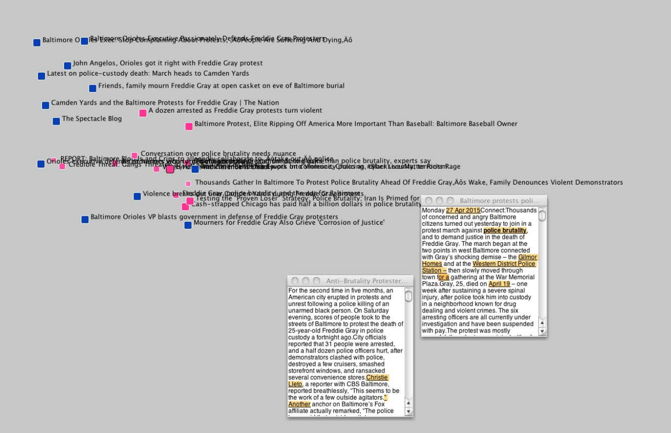


Figure 4. Final spatial layout with labeled clusters

B. *Investigative Journalism*

The investigative reporting scenario linked StarSPIRE to the Bing web search engine. While StarSPIRE is capable of searching the entire web for potential documents with Bing, the scope of this task was limited to news articles. The web documents are labeled with the webpage title. As such, it is typically possible to eliminate clearly irrelevant documents.

For this task, we chose to select a current and controversial topic: police brutality and the ensuing protests. This topic was chosen because it has passionate and opinionated reporters on both sides and has received an overwhelming amount of national coverage. This analysis began with a search for "police brutality," which returned articles that appear to be about recent incidents in the United States. The most relevant article is opened automatically by the system.


Baltimore Orioles Executive Passionately Defends Freddie Gray Protesters
John Angelos, Orioles got it right with Freddie Gray protest
Latest on police-custody death: March heads to Camden Yards
Friends, family mourn Freddie Gray at open casket on eve of Baltimore burial
Camden Yards and the Baltimore Protests for Freddie Gray | The Nation
A dozen arrested as Freddie Gray protests turn violent
The Spectacle Blog
Baltimore Protest, Elite Ripping Off America More Important Than Baseball: Baltimore Baseball Owner
Conversation over police brutality needs nuance
Thousands Gather In Baltimore To Protest Police Brutality Ahead Of Freddie Gray,Äôs Wake, Family Denounces Violent Demonstrators
Testing the "Proven Loser" Strategy; Police Brutality; Iran Is Primed for
Cash-strapped Chicago has paid half a billion dollars in police brutality
Baltimore Orioles VP blasts government in defense of Freddie Gray protesters
Mourners for Freddie Gray Also Grieve 'Corrosion of Justice'
Baltimore protests poli...
Anti-Brutality Protester...

Monday 27 Apr 2015Connect:Thousands of concerned and angry Baltimore citizens turned out yesterday to join in a protest march against police brutality, and to demand justice in the death of Freddie Gray. The march began at the two points in west Baltimore connected with Gray's shocking demise – the Gilmor Homes and at the Western District Police Station – then slowly moved through town for a gathering at the War Memorial Plaza.Gray, 25, died on April 19 – one week after sustaining a severe spinal injury, after police took him into custody in a neighborhood known for drug dealing and violent crimes. The six arresting officers are all currently under investigation and have been suspended with pay.The protest was mostly

For the second time in five months, an American city erupted in protests and unrest following a police killing of an unarmed black person. On Saturday evening, scores of people took to the streets of Baltimore to protest the death of 25-year-old Freddie Gray in police custody a fortnight ago.City officials reported that 31 people were arrested, and a half dozen police officers hurt, after demonstrators clashed with police, destroyed a few cruisers, smashed storefront windows, and ransacked several convenience stores.Christie Lleto, a reporter with CBS Baltimore, reported breathlessly, "This seems to be the work of a few outside agitators." Another anchor on Baltimore's Fox affiliate actually remarked, "The police

Figure 5. Retrieval results from overlapping two documents.

The protests in Baltimore seem interesting, particularly because they are the most recent focus in the news. A second document with a seemingly relevant title is opened, which results in other documents changing in size to reflect the changes in document relevance. These documents are then overlapped to search for information that matches the terms that co-occur between the documents. The result is a number of documents about Baltimore protests, many of which contain the name Freddie Gray. We highlight the name in order to indicate our interest to the system, which retrieves additional documents. This results in many highly relevant documents to analyze which were previously unknown [Figure 5].

VI. Discussion

The use cases presented here demonstrate how a semantic interaction tool can be used to complete real-world tasks in real-time, allowing the user and computer to jointly curate, arrange, and analyze large document collections while preserving the context of previously completed actions.

A. *Use Case Performance*

Approximately half of the relevant documents retrieved from both use cases came from implicitly constructed queries. The remaining half came from explicit searching or query by example. In both use case scenarios, interesting documents were obtained that did not match any of the initial search terms, showing that users are able to explore previously unknown regions of the information space. These results show extensive promise for applying semantic interaction to exploratory data analysis tasks. However, more extensive studies are needed to quantitatively and qualitatively evaluate the design decisions made in the development of this extension.

B. *Limitations*

There are several limitations to this system that should be noted. Currently, all documents are parsed for entities at once, which can delay the time needed to import documents. The retrieval, relevancy ranking, and parsing process should be improved to stream in results or prioritize the very top retrieved documents. Additionally, using outside algorithms for information retrieval limits the amount of model steering that can be done at the information retrieval level. Furthermore, web-based information retrieval does not directly translate to protected databases containing millions of documents. Additional algorithms will be needed to access such datasets.

C. *Applications*

In this instance, we exploited the search engines for Bing and IEEE Digital Library. A previous version of StarSPIRE connected with an existing document modeling and matching system connected to a large database of approximately 13,000 documents. In all of these instances, StarSPIRE must be adapted to work with the existing APIs. For example, the database search could place emphasis on different search terms by repeating the terms in the search query. StarSPIRE converted the existing entity weighting scheme to this format by linearly increasing query term frequency as the term weight increased quadratically. With Bing and IEEE, StarSPIRE creates a query string by ordering the search terms by their associated weight. These modifications demonstrate how we can take advantage of existing recommendation systems while still using the user's interest model as input for these systems. By using such existing tools, we are able to leverage their strengths, such as filtering out duplicate articles so that the user can focus on broader aspects of the topic of interest. However, it may be advisable to cast an even wider net for potential results when using external services, since the retrieval algorithms themselves are unknown. By collecting a larger sample of potentially relevant results, we can lower the risk of missing important documents. This work could be extended to cache the additional potential results in StarSPIRE's local database. They could then be added to the workspace if deemed relevant during future interactions without requiring additional external queries.

Future work could explore leveraging additional recommender systems, both on the small and large scale. For example, a team of analysts could be linked such that they co-create a model of their interests in the data.

VII. Conclusion

The endeavor to create a system in which foraged results are placed in context of an information synthesis space has raised many research challenges. We have addressed several of them in terms of design decisions explored and made, although many remain as open research questions. Through this work, we have enabled users to perform common exploratory data analysis tasks while having access to a nearly unlimited amount of data, while keeping interactions and feedback in context of the user's current analytical state. We have been able to successfully complete these tasks, allowing StarSPIRE to be used in real-world applications while maintaining a quick interaction-feedback loop.

Acknowledgments

This research was supported in part by NSF grants IIS-1218346 and IIS-1447416.

References

1. Ahlberg, C. and Wistrand, E., IVEE: An information visualization and exploration environment. in *Information Visualization, 1995. Proceedings.*, (1995), IEEE, 66-73.
2. Alper, N. and Stein, C., Geospatial metadata querying and visualization on the WWW using Java TM applets. in *Information Visualization'96, Proceedings IEEE Symposium on*, (1996), IEEE, 77-84, 128.
3. Alsakran, J., Chen, Y., Zhao, Y., Yang, J. and Luo, D. STREAMIT: Dynamic visualization and interactive exploration of text streams *IEEE Pacific Visualization Symposium*, 2011.
4. Andrews, C., Endert, A. and North, C. Space to think: large high-resolution displays for sensemaking *Proceedings of the SIGCHI conference on human factors in computing systems*, ACM, 2010, 55-64.
5. Andrews, C. and North, C. Analyst's Workspace: An Embodied Sensemaking Environment For Large, High-Resolution Displays *IEEE visual analytics science and technology*, IEEE, 2012, 123-131.
6. Ansari, A., Essegaier, S. and Kohli, R. Internet recommendation systems. *Journal of Marketing research*, *37* (3). 363-375.
7. Bahrami, A., Yuan, J., Smart, P.R. and Shadbolt, N.R., Context aware information retrieval for enhanced situation awareness. in *Military Communications Conference, 2007. MILCOM 2007. IEEE*, (2007), IEEE, 1-6.
8. Baldonado, M.Q.W. and Winograd, T., SenseMaker: an information-exploration interface supporting the contextual evolution of a user's interests. in *Proceedings of the ACM SIGCHI Conference on Human factors in computing systems*, (1997), ACM, 11-18.
9. Baldwin, B. and Carpenter, B. LingPipe. *Available from World Wide Web:* http://alias-i. *com/lingpipe*.
10. Belew, R.K., Adaptive information retrieval: Using a connectionist representation to retrieve and learn about documents. in *ACM SIGIR Forum*, (1989), ACM, 11-20.
11. Blei, D.M., Ng, A.Y. and Jordan, M.I. Latent dirichlet allocation. *J. Mach. Learn. Res.*, *3*. 993-1022.

12. Bradel, L., Endert, A., Koch, K., Andrews, C. and North, C. Large high resolution displays for co-located collaborative sensemaking: Display usage and territoriality. *International Journal of Human-Computer Studies*, *71* (11). 1078-1088.
13. Bradel, L., North, C., House, L. and Leman, S., Multi-model semantic interaction for text analytics. in *IEEE Conference on Visual Analytics Science and Technology (VAST), 2014*, (2014), IEEE, 163-172.
14. Bradel, L., Self, J.Z., Endert, A., Hossain, M.S., North, C. and Ramakrishnan, N., How analysts cognitively "connect the dots". in *Intelligence and Security Informatics (ISI), 2013 IEEE International Conference on*, (2013), 24-26.
15. Brown, E.T., Liu, J., Brodley, C.E. and Chang, R., Dis-function: Learning distance functions interactively. in *Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on*, (2012), IEEE, 83-92.
16. Burtner, R., Bohn, S. and Payne, D., Interactive visual comparison of multimedia data through type-specific views. in *IS&T/SPIE Electronic Imaging*, (2013), International Society for Optics and Photonics, 86540M-86540M-86515.
17. Clarkson, E., Desai, K. and Foley, J.D. Resultmaps: Visualization for search interfaces. *Visualization and Computer Graphics, IEEE Transactions on*, *15* (6). 1057-1064.
18. Derthick, M., Christel, M.G., Hauptmann, A.G. and Wactlar, H.D., Constant density displays using diversity sampling. in *Information Visualization, 2003. INFOVIS 2003. IEEE Symposium on*, (2003), IEEE, 137-144.
19. Derthick, M., Roth, S.F. and Kolojejchick, J., Coordinating declarative queries with a direct manipulation data exploration environment. in *Information Visualization, 1997. Proceedings., IEEE Symposium on*, (1997), IEEE, 65-72.
20. Endert, A. Semantic Interaction for Visual Analytics: Inferring Analytical Reasoning for Model Steering, Virginia Polytechnic Institute and State University, 2012.
21. Endert, A., Bradel, L. and North, C. Beyond control panels: Direct manipulation for visual analytics. *Computer Graphics and Applications, IEEE*, *33* (4). 6-13.
22. Endert, A., Burtner, R., Cramer, N., Perko, R., Hampton, S. and Cook, K., Typograph: Multiscale spatial exploration of text documents. in *Big Data, 2013 IEEE International Conference on*, (2013), IEEE, 17-24.
23. Endert, A., Fox, S., Maiti, D., Leman, S. and North, C. The semantics of clustering: analysis of user-generated spatializations of text documents *Proceedings of the International Working Conference on Advanced Visual Interfaces*, ACM, Capri Island, Italy, 2012, 555-562.
24. Endert, A., Han, C., Maiti, D., House, L., Leman, S. and North, C., Observation-level interaction with statistical models for visual analytics. in *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on*, (2011), IEEE, 121-130.
25. Faloutsos, C. and Oard, D.W. A survey of information retrieval and filtering methods.
26. Fiaux, P., Sun, M., Bradel, L., North, C., Ramakrishnan, N. and Endert, A. Bixplorer: Visual analytics with biclusters. *Computer* (8). 90-94.
27. Fishkin, K. and Stone, M.C., Enhanced dynamic queries via movable filters. in *Proceedings of the SIGCHI conference on Human factors in computing systems*, (1995), ACM Press/Addison-Wesley Publishing Co., 415-420.
28. Ghias, A., Logan, J., Chamberlin, D. and Smith, B.C., Query by humming: musical information retrieval in an audio database. in *Proceedings of the third ACM international conference on Multimedia*, (1995), ACM, 231-236.
29. Green, T.M., Ribarsky, W. and Fisher, B. Building and applying a human cognition model for visual analytics. *Information visualization*, *8* (1). 1-13.
30. Gupta, A. and Jain, R. Visual information retrieval. *Communications of the ACM*, *40* (5). 70-79.
31. Hearst, M. *Search user interfaces*. Cambridge University Press, 2009.
32. Hearst, M.A., TileBars: visualization of term distribution information in full text information access. in *Proceedings of the SIGCHI conference on Human factors in computing systems*, (1995), ACM Press/Addison-Wesley Publishing Co., 59-66.
33. Hofmann, T., Probabilistic latent semantic indexing. in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, (1999), ACM, 50-57.
34. Hu, X., Bradel, L., Maiti, D., House, L. and North, C. Semantics of Directly Manipulating Spatializations. *Visualization and Computer Graphics, IEEE Transactions on*, *19* (12). 2052-2059.
35. i2. Analyst Notebook, 2007.
36. Jeong, D.H., Ziemkiewicz, C., Fisher, B., Ribarsky, W. and Chang, R., iPCA: An Interactive System for PCA‚Äêbased Visual Analytics. in *Computer Graphics Forum*, (2009), Wiley Online Library, 767-774.
37. Koch, S., Bosch, H., Giereth, M. and Ertl, T., Iterative integration of visual insights during patent search and analysis. in *Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on*, (2009), IEEE, 203-210.
38. Lew, M.S., Sebe, N., Djeraba, C. and Jain, R. Content-based multimedia information retrieval: State of the art and challenges. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP)*, *2* (1). 1-19.
39. Lin, X., Soergel, D. and Marchionini, G., A self-organizing semantic map for information retrieval. in *Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval*, (1991), ACM, 262-269.
40. Maron, M.E. and Kuhns, J.L. On relevance, probabilistic indexing and information retrieval. *Journal of the ACM (JACM)*, *7* (3). 216-244.
41. Masui, T., LensBar-visualization for browsing and filtering large lists of data. in *Information Visualization, 1998. Proceedings. IEEE Symposium on*, (1998), IEEE, 113-120, 159.
42. Olsen, K.A., Korfhage, R.R., Sochats, K.M., Spring, M.B. and Williams, J.G. Visualization of a document collection: The VIBE system. *Information Processing & Management*, *29* (1). 69-81.
43. Pazzani, M.J. and Billsus, D. Content-based recommendation systems. in *The adaptive web*, Springer, 2007, 325-341.
44. Ruotsalo, T., Peltonen, J., Eugster, M., Głowacka, D., Konyushkova, K., Athukorala, K., Kosunen, I., Reijonen, A., Myllymäki, P. and Jacucci, G., Directing exploratory search with interactive intent modeling. in *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, (2013), ACM, 1759-1764.
45. Shipman, F.M. and Marshall, C.C. Formality Considered Harmful: Experiences, Emerging Themes, and Directions on the Use of Formal Representations in Interactive Systems. *Computer Supported Cooperative Work (CSCW)*, *8* (4). 333-352.
46. Shneiderman, B. Direct manipulation. *B. Shneiderman*.
47. Shneiderman, B. Dynamic queries for visual information seeking. *Software, IEEE*, *11* (6). 70-77.
48. Shrinivasan, Y.B., Gotz, D. and Lu, J., Connecting the dots in visual analysis. in *Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on*, (2009), IEEE, 123-130.
49. Stasko, J., Görg, C. and Liu, Z. Jigsaw: supporting investigative analysis through interactive visualization. *Information visualization*, *7* (2). 118-132.
50. Tanin, E., Beigel, R. and Shneiderman, B., Design and evaluation of incremental data structures and algorithms for dynamic query interfaces. in *Information Visualization, 1997. Proceedings., IEEE Symposium on*, (1997), IEEE, 81-86.
51. Teevan, J., Dumais, S.T. and Horvitz, E., Personalizing search via automated analysis of interests and activities. in *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, (2005), ACM, 449-456.
52. Tesone, D.R. and Goodall, J.R., Balancing interactive data management of massive data with situational awareness through smart aggregation. in *Visual Analytics Science and Technology, 2007. VAST 2007. IEEE Symposium on*, (2007), IEEE, 67-74.
53. Wattenberg, M. and Viégas, F.B. The word tree, an interactive visual concordance. *Visualization and Computer Graphics, IEEE Transactions on*, *14* (6). 1221-1228.
54. Wise, J.A., Thomas, J.J., Pennock, K., Lantrip, D., Pottier, M., Schur, A. and Crow, V., Visualizing the non-visual: spatial analysis and interaction with information from text documents. in *Information Visualization, 1995. Proceedings.*, (1995), IEEE, 51-58.
55. Wright, W., Schroh, D., Proulx, P., Skaburskis, A. and Cort, B., The Sandbox for analysis: concepts and methods. in *Proceedings of the SIGCHI conference on Human Factors in computing systems*, (2006), ACM, 801-810.