CKM: A Shared Visual Analytical Tool for Large-Scale Analysis of Audio-Video Interviews

Lu Xiao*, Yan Luo† and Steven High‡
* † *The University of Western Ontario, Canada,*
**Email: lxiao24@uwo.ca,* †*yluo84@csd.uwo.ca*
‡ *Concordia University, Canada,*
Email: steven.high@concordia.ca

***Abstract*—Our access to human rights violation data has increased with the growing number and size of data collections. We have been combining text-mining and visualization techniques to facilitate big data analysis in human rights research. Taking a user-centered approach, we first surveyed the human rights research literature to understand reported data analysis practices in the field, and then taking a participatory design approach working with oral history researchers to develop a visual analytical tool that facilitates the analysis of collections of audio-video interviews oral history research projects. In this paper we present our current prototype - Clock-based Keyphrase Map (CKM). CKM utilizes Keyphrase technique to identify important topics in the collection and a clock-based visualization to present them in a temporal order. CKM also enables the users to further analyze the collections and share their analysis process with other researchers. We discuss the tool in details including its architecture, the computational and visualization techniques, and the interaction features. Our future plan on evaluation and further development are also discussed in the paper.**

***Keywords*-Keyphrase technique; visual analytics; human rights data; audio-video interviews; collaborative analysis**

I. Introduction

There has been different definitions of Big Data. For example, Boyd and Crawford [1] suggested that, Big Data is fundamentally networked. Its value comes from the patterns that can be derived by making connections between pieces of data (p.2). Fisher et al. [2] identified big data as datasets that are so large that they cannot always fit on a single hard drive. Rotman et al. [3] chose to use the term large scale online environments instead to encompass the large datasets, the tools that are used in the process of creating, archiving, and sharing the datasets, and the interactions around the data. In this paper, we use big data in the sense that Boyd and Crawford [1] deployed the term.

Researchers are increasingly drawn to the opportunities and challenges of accessing and utilizing big data from a variety of perspectives, and it has even been argued that a ”fourth paradigm” of data-driven scientific research is upon us (Hey et al. [4]). In this research program, we are interested in combining text-mining and visualization techniques to facilitate the data analysis processes around big data in human rights research. In our project, we first surveyed the related literature on how researchers analyze big data in general, and then focused on developing a tool that facilitates the analysis of a collection of video interview data in human rights research, the Stories Matter database. These interview data were created, archived, and shared by a research community at Concordia Universitys Centre for Oral History and Digital Storytelling that was engage in recording the life stories of hundreds of people displaced by war, genocide and other atrocity crimes. Once created, the database software has been adopted by oral historians and other qualitative researchers working within projects of varying size, ranging from large interview projects of the size of the Montreal Life Stories Project (www.lifestoriesmontreal.ca) to projects involving a single researcher or graduate student. The scale of the datasets being produced within Stories Matter environments thus varies tremendously.

Our development process is user-centered. At the initial meeting with the database creator and regular user (who is also a co-author of this paper), we asked questions related to the user behavior of the database, e.g. human rights researchers can use the software to watch interview segments that speak to a specific aspect of mass violence, the experience of forced displacement, or how it is later remembered within families or survivor communities. We then constructed several paper prototypes to illustrate our main design ideas and presented to the regular users and maintainers of Stories Matter database. Based on the their feedback, we decided to develop a visualization tool that 1) combines text mining and human tagging approaches to identify and present similarities among interview content, 2) allows the researchers to play the interview section that has is related to an identified topic, and 3) enables the researchers to annotate the interviews and share their annotations within the research community.

In the remaining paper, we first review the related work, and then present our design prototype. We conclude with a list of design features that we will implement in the future and our evaluation plan.

II. Related Work

A. Big Data Analysis

There are a few studies that examined data analysis processes around big data to identify the challenges and opportunities of offering information and communication technologies to aid the process. Heer and Kandel [5] interviewed analysts to understand how they manage big data. They found that most data analysts follow the same general workflow pattern of data discovery and acquisition. For example, the data analysts 'wrangle' data through procedures like reformatting, cleaning and integration, profiling data to explore its contents, and identifying salient features, assessing data quality issues, modeling data to explain or predict phenomena, and reporting results to disseminate findings. In Fisher et al. [2]s interviews of sixteen data analysts at Microsoft, the authors found that analysts would reflect and iterate on the results during the process.

To understand how qualitative methods are used in large scale online environments, Rotman et al. [3] conducted interviews with nine leading qualitative researchers and identified the practical challenges in the process of qualitative analysis of large-scale online environments: identifying entry point for analysis and selecting participants; ephemerality, interface, and cultural change; and applying ethical oversight. From their interview study, Kandel et al. [6] identified five high-level tasks in data analysis: discover, wrangle, profile, model, and report. The authors also maintained that there is an opportunity for visual analytic tools to improve the quality and speed of big data analysis. They suggested that such tools should have features like managing diverse sets of procedures, data sets, and intermediate data products. The collaborative features that help analysts to share the analysis process are also promising.

With the increasing likelihood of accessing and analyzing big data, the interest of facilitating analysis of big data with computer technologies is growing.Heer and Kandel [5] designed a tool to support the process of wrangling data, that is, the process of reformatting data values or layouts, correcting missing values, and integrating multiple data sources. Knudsen et al. [7] explored the possibility of using large, high-resolution multi-touch displays to facilitate data analysis process. They suggested design implications such as supporting sequential visualizations of data, interaction from a distance, and allowing the display of variables and data be fixed on the display or vary relative to the users position. Fisher et al. [2] presented a hypothetical system which addresses the existing challenges of managing big data by including features such as increased interactivity and providing incremental results to the user as soon as they are available rather than making them wait until the full computation is complete. A study by Livnet and Zhang [8] suggested that features that help the analysts promptly interpret information can be more important than helping them discover information patterns in some cases.

Based on our review of these studies about big data analysis and the requirements for tools to support the analysis process, we drew several design implications that were realized in our tool. They are: a) fostering analysts to reflect and iterate the results; b) affording identification of interviewees to analyze; c) revealing similarities among different interviews to help analysts integrate data interpretations; d) allowing interactive visualizations and presenting incremental results; and e) affording analysts to promptly interpret the information, as well as record and share their interpretations. We decided not to focus on supporting data wrangling process as Stories Matter database is already well structured and the data cleaning process is completed when they are added to the database.

B. Text-Mining Techniques

Text mining aims to discover the useful hidden patterns and information from a large scale of structured, semi-structured, and / or unstructured textual data. Current researches in this area mainly tackle the problems of text categorization, clustering, information extraction, topic modelling, summarization, concept linkage, and question answering [9]. The domain-specific problems or issues usually play an very important role in the selection of text mining tools and determine the research direction.

Keyphrases or keywords can help users obtain a feel for the content of a collection, provide sensible entry points, and offer a powerful means of comprehending the document similarity. Most practical key phrases algorithms follow the "learning to extract" framework. That is, documents are first preprocessed to generate a collection of candidate phrases, and various of lexical or semantic word features are designed to characterize each candidate. Then, different kinds of machine learning algorithms/models can be adopted to predict whether a candidate phrase is a keyphrase. Training documents with known keyphrases are used to tune the model parameters and improve the prediction accuracy.

Researchers mainly focus on two directions to improve the performance of keyphrases extraction: 1) better feature engineering and 2) more advanced machine learning models. KEA algorithm is a very early and popular keyphrases algorithm proposed by Witten et.al [10]. They used the *TF* × *IDF* and *first occurrence* as features and build a Naïve Bayes prediction model on each candidate. Zhang et. al [11] significantly enrich the dimension of phrase feature vector and employed the Conditional Random Fields (CRF), which is a state-of-the-art sequence labeling method to calculate the maximum probability of being keyphrase for each candidate. Their results show that CRF model outperform other models such as support vector machine (SVM), multiple logistic regression, etc. More intelligently, Chen et al. [12] proposes an unsupervised two-stage keyword extraction approach. They first utilize the topic coherence and term significance measure to select qualified training documents, and then use them to train an SVM classifier to predict the keywords. Their experiments on course lectures documents showed very promising keyword extraction performance.

C. Text-Mining based Visualization Tools

The use of text-mining techniques allows us to discover patterns or relationships from large amount of textual data. It is another challenging task to effectively and interactively present the users this discovered information so as to facilitate further data analysis tasks. Information visualization techniques have been used to address this challenge in various research and data analysis context. For example, Ong et al. [13] presented an integrated web based application called FOCI (Flexible Organizer for Competitive Intelligence) that uses text mining techniques, user-configurable clustering, trend analysis and visualization techniques to address the problem of managing information gathered from the web.

Jacquemin et al. [14] presented an interactive visualizer, OCEAN, that takes the output of a text miner TEMIS as input to produce a representation of the documents in a 3-dimensional space. To help crime analysts analyze, compare and contrast crime reports in a timely manner, Ku [15] developed a crime report visualization system Textual Analysis of Similar Crimes (TASC) that utilizes a document similarity algorithm that analyzed semantic characteristics between crime reports, and text visualization techniques to present the identified similarities among reports.

Cui et al. [16] presented a visual analytic tool, TextFlow, to illustrate how topics evolve in the textual data sets. TextFlow is an interactive visual analysis tool that helps users analyze how and why the correlated topics change over time. It presents the three-level mining results in a single view. In this three-level mining method, the authors first used a probabilistic model to extract topics from the dataset and model the relationships among topics. They then computed topic merging and splitting relationships using some statistical formula. With another formula, the authors next extracted noun and verb phrases, and named entities to help users better understand the major reasons triggering topic evolution and rank the keywords. TextFlow tool used the river-flow-based visual metaphor: the topic flows of all topics were stacked and aligned based on the time stamp. The tool also supports interactive exploration of the data by allowing hovering and selection in the visualization.

Our tool development is based on Stories Matter database which is well-structured and metadata-rich. The efficiency of traditional text mining tasks like classification or clustering would be limited on such database. On the other hand, the end users of Stories Matter are interested in how the time-series interviews evolve under certain topic(s) or how they are related with each other to support that topic(s). Such unique challenges inspired us to direct our attention to the area of keywords extraction rather than other text-mining techniques to facilitate this analysis process. We also recognize the challenges of presenting and facilitating interaction with the result of text-mining technique, and combine text-mining and visualization techniques in our tool design and development. In the following section, we introduce Stories Matter in more details. We then explain in details the text-mining technique and visualization mechanisms of our tool.

III. Stories Matter - An Oral History Database

Stories Matter database software (http:/storytelling.concordia.ca/storiesmatter/) was developed by Concordia Universitys Centre for Oral History and Digital Storytelling and lets researchers interact directly with the original audio-video interviews. It is an alternative to transcription with respect to analyzing audio-video interviews. Survivor testimony is an integral part of human rights research and it is regularly deployed in global human rights campaigns, truth and reconciliation processes, and court cases. Indeed, eyewitness testimony has proven essential to bringing perpetrators to justice and to building the case for political intervention or societal reconciliation [17]. Personal stories have the power to affect otherwise remote publics, or to build solidarity amongst survivors themselves. Given its political importance, and the growing scale of these interviewing efforts - as evidenced by the Shoah Visual History Foundations 54,000 interviews, it is noteworthy that human rights researchers, truth and reconciliation commissions, and other large testimony projects continue to rely on transcription as their primary interpretative and search tool (Frisch [18]; High et al. [19]). In a transcript, the words spoken are recorded but not the accompanying emotions, body language, pauses, or the sound of their voices (High and Sworn [20]). Much is lost in the act of translation from the spoken word to the written word. Transcripts are also ill-positioned to help us make connections between individual testimonies, to follow threads of meaning across multiple interviews or find wider patterns of significance. As a result, individual life stories often remain detached from one-another or, as Schafer and Smith [21] argue, used as little more than an illustrative device.

Acknowledging these shortcomings of relying on transcriptions in analyzing audio-video interviews, the researchers developed Stories Matter database. In building this database, qualitative researchers work at multiple levels, richly annotating the source interviews at the project level, as well as the level of the interviewee, the interview session (as many oral historians conduct multiple interviews) and the clip level. Users can search via the tag cloud, or via the keywords found in the annotations or transcripts themselves. Additional documentary evidence can be attached to the audio or video recording, such as interviewer reflections or field notes as well as scanned material collected from the interviewees themselves. With the Stories Matter database, the manual clipping and indexing of these interviews transform stand-alone interviews into searchable data-sets, revealing the wider logics that underpin or structure these conversations (Jessee et al. [22]).

Researchers of different oral history groups have been using Stories Matter database to archive their own collections of audio-video interviews. The database makes it possible to share data across research projects and research groups. Moreover, a list of standardized tags are provided in the database, which makes it relatively easy to associate different interviews cross the projects. For example, in the Montreal Life Stories collection the "public remembering“ tag term yields dozens of interview sessions and clips where interviewees speak to the ways in which individuals, families, and communities publicly remember mass violence. In scrolling through the results, and listening to the clips identified, one quickly realizes how many of the interviewees spoke of the "first time“ that they told their story in "public“. From this observation, it would be relatively easy using Stories Matter to zoom in on this aspect listening in as survivors from the Holocaust, the Rwandan and Cambodian genocides, political oppression in Haiti, and atrocity crime committed elsewhere speak to this idea. In making the shift from oral history transcription to oral history datasets, we are therefore encouraged to make connections not only between the stories told by different survivors but by different survivor communities.

IV. Clock-based Keyphrases Map (CKM)

To develop a visual analytical tool for Stories Matter researchers, we first interviewed the director of Stories Matter database project who is also an oral history researcher and a user of the database. Our interview was focused on understanding his data analysis practices and examples of how he used Stories Matter in his research. We then brainstormed several design ideas within the development team and produced low-fidelity prototype to illustrate our design. We presented our design ideas to the director and two other regular users of the database to further understand the design requirements of the tool and engage them in the design brainstorming discussion. Our prototype presented here, the Clock-based keyphrases Map (CKM) was then developed based on the users feedback at that presentation.

Keyphrases are able to summarize and characterize each interviews and help users to grab the main ideas in a much shorter time. More interestingly, a properly designed visual representation and visual analytic tool, as opposed to a textual or numeric representation, allows them to understand a large amount of keyphrases and dig up new knowledge in a parallel manner. To validate this idea, we have designed and built a prototype integrating both keyphrases extraction and information visualization technologies into Stories Matter. The prototype framework is shown in Figure 1, which contains three main components: 1) *Data source management*, 2) *Automatical keyphrases extraction*, and 3) *Keyphrases map visualization*. Data source management module is designed to store and preprocess the raw data and convert them to proper data formats for further text mining and visualization tasks. When the data is cleaned and prepared, various keyphrases extraction algorithms can be adopted and plugged into the Automatical keyphrases extraction module. Generally, the keyphrases extraction contains a training stage and a extracting stage. In the training phrase, users need to manually identify keyphrases for each document. The predication models will be then tuned using training data to characterize the keyphrases from the candidates. In the extracting stage, the probability of being keyphrase for each candidate is calculated and the top ranked candidates are selected as keyphrases. Then, we need to determine what are the objects that need to be visualized. Here, we define an entity as an abstract object that has metadata and set of keyphrases. An entity could be a Stories Matter project, interview session, or video clip, and their metadata includes information like name, ID, time duration, tag, etc. Furthermore, the entities to be visualized are denoted as *visual entities*. Visual entities and the associated keyphrases are combined to form different GraphMLs (GML), which is an XML-based file format for graphs to support the entire range of possible visual graph structure constellations. The keyphrases map visualization module will render the clock-based keyphrase map based on the abstract visual graphs stored in the GML files.

A. Automatical Keyphrases Extraction

Among the keyphrase extraction algorithms surveyed in Section 2, we adopted and refined the most popular Keyphrases Extraction Algorithm (KEA) [10]. KEA uses only two but very representative features, e.g., $TF \times IDF$ and *first occurrence* and Naïve Bayes for keyphrases extraction. As such, it is very computational effective, which is very important for user experience in practice. Despite of its simplicity, KEA still can perform at the current state-of-art. In addition, KEA provides a flexible framework for future extension and refinement. We can conveniently extend the keyword feature vector or adopt more complicated models to improve its prediction accuracy. In the future, we plan to conduct empirical studies to identify the optimal features and discriminative models to balance the computation time and accuracy.

Given a document stream written in a single language, KEA first extracts qualified candidate phrases using lexical methods. More specifically, the document stream will be tokenized first to generate initial tokens set (candidate phrases). The initial token set will be then filtered according to three criterions:1) maximum length is limited to three words 2) can not be proper names and 3) can not begin or end with language specific stop words. The last step is case-fold all candidates and using the classic Lovins stemmer to discard any suffix.

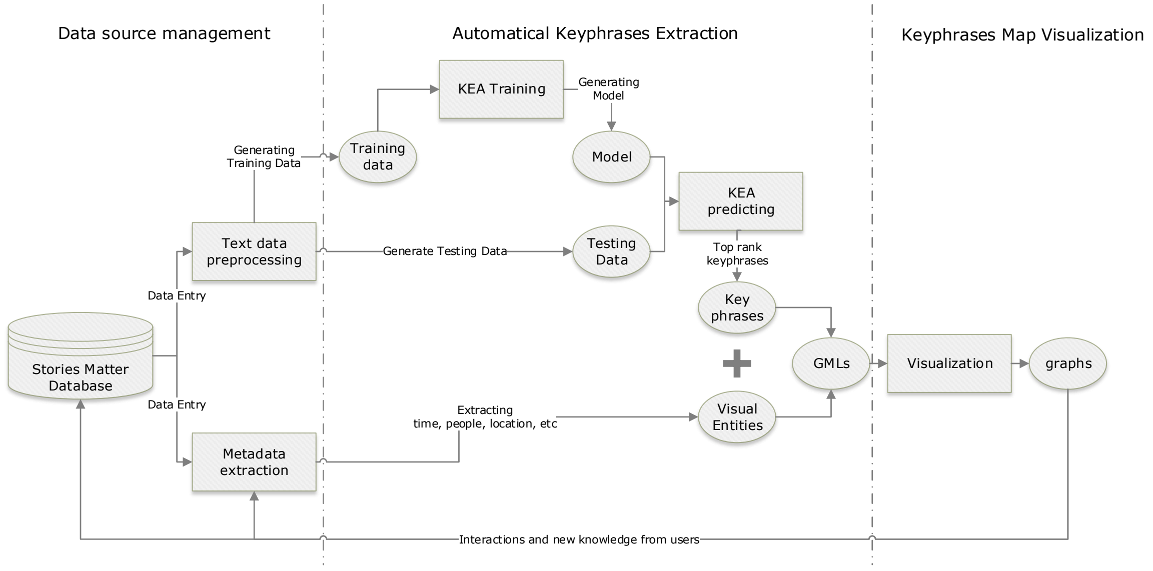


Figure 1. Framework of Clock-based Keyphrases Map

For each token, its TF × IDF and first occurrence values will be calculated as two main features. TF × IDF measures how important or rare a word is to the document with respect to a collection or corpus. The second feature, first occurrence, is the relative distance from first appearance to the start point of the document. To eliminate the effects of the size of document, the first occurrence is typically normalized by the document length.

The training stage utilize a collection of documents with known keyphrass to build a Naïve Bayes model for new keyphrases extraction. To extract the new keyphrases, the Naïve Bayes model calculate the overall probability. Then, the top-K tokens with highest probability with be selected as keyphrases k_j, $j = 1, 2, \ldots, K$ for each document. Note that the parameter K can be specified by users, depending on what detailed level they want the KEA to summarize their interview textual data.

Stories Matter users need to assign various tags T_i, $i = 1, 2, \ldots, m$ on each document D. Such tags indicate the categories/topics of documents and thus they can summarize a document from a more general aspect. Meanwhile, the keyphrases are much more specific to a document since they are the actual words existing in the document. In practical, it is possible that an extracted keyphrase k is similar or even identical to an existing tag T. Thus, we need to filter such abundant keyphrases.

We use the classic Levenshtein distance to measure the difference between a keyphrase and a tag. It is calculated by the minimum number of single-character edits required to change one word into the other. Given the Levensthein distance between a tag T and keyphrase k as $L(T, k)$, the similarity s can be calculated as:

$$s(T, k) = 1 - \frac{L(T, k)}{l_{max}(T, k)} \tag{1}$$

where $0 \leq s \leq 1$ and $l_{max}(T, k)$ is the maximum string length of tag T and k. If the similarity $s(T_i, k_j)$ is greater than a threshold h, k_j will be filtered out since it sufficient similar to the tag T_i. We normally set the h to a very large value to ensure that the extracted keyphrases are as informative as possible.

B. Keyphrases Map Visualization

Suppose an interviewer has interviewed a number of interviewees in order to investigate identified topic(s), he/she could be mainly interested in the following questions:

- *Q1*: How to quickly access and retrieve the desired information?
- *Q2*: Any point of views or content are shared and linked among different interviewees?
- *Q3*: How the interviews evolve?
- *Q4*: How to share my analysis process with other researchers?

Even though the interview data has been summarized and characterized by keyphrases and tags, it would be still difficult for the interviewer to conveniently look into above issues. To solve them more effectively, we borrow the ideas of real clock and map and create a novel visual analytic tool called Clocked-based Keyphrases Map (CKM) to help users reviewing and analyzing the interview data in a visually, intuitive, and parallel manner.

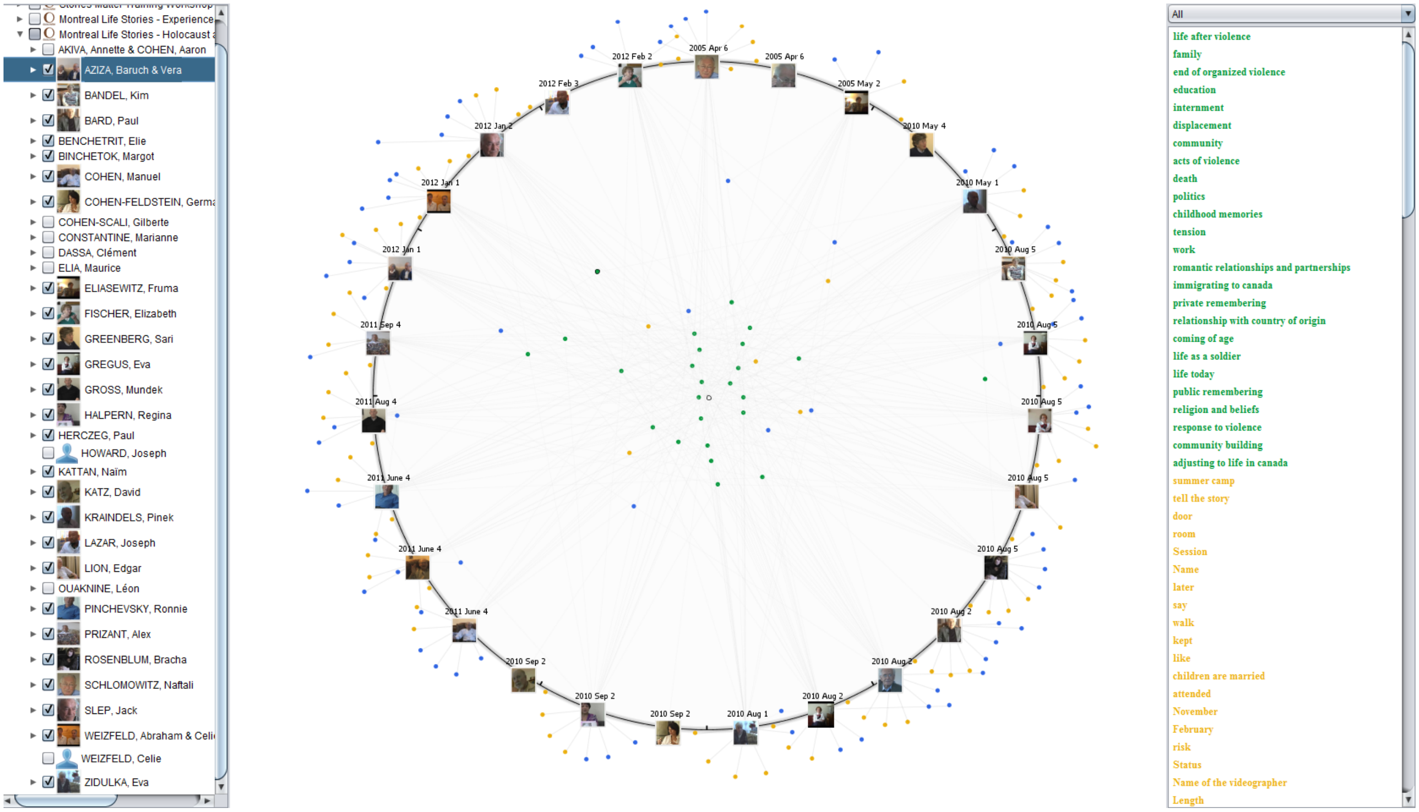


Figure 2. The global view of CKM

The global view of CKM is shown in Figure 2. To better describe and illustrate the validness of our tool, we imported a real interview project with 26 interview sessions (all in English) into the tool. It usually starts with an empty clock and user can add arbitrary number of visual entities on CKM for analyzing. As we can see in Figure 2, all visual entities are placed equally on the clock and sorted in clockwise direction. Each visual entity is visualized as the root of a subtree and links various number of child nodes. A node represents either a tag or a keyphrase and its color means its type. At this point, we do not show the text directly. The main reason is that the perception of textual data is serial, thus visualizing too many text information could overwhelm users. Users can reveal the texts by hovering their mouse on interested nodes or visual entities. More interestingly, the position of a node indicates its commonness or uniqueness to the global visual entities. Nodes closer to the rim mean they are more specific to certain interview sessions and not likely to be shared by other interview sessions. While nodes are at the center of the clock indicate that they could be more likely to be shared by the global visual entities. Figure 2 provides the global view about the CKM, users may find their interesting areas or node patterns and then further interact with the clock using operations such as zoom in/out, pan, highlight to discover the details on demand. However, before explaining the detailed functionalities, we want to first discuss how the node positions are calculated because it is the core of this visual representation.

1) Nodes Localization: Each visual entity on the clock rim acts like an anchor point determining the positions of all nodes. The node positions are calculated based on the relative weights (distance) on each visual entity. Suppose a visual entity V_i, $i = 1, 2, \ldots, M$ has a set of child node n_j^i, $j = 1, 2, \ldots, N^i$, the relative weight of a node $\hat{n}$ ($\hat{n}$ is the child node of a visual entity as well) on visual entity V_i is calculated by:

$$w_i = max(s(\hat{n}, n_j^i)) \qquad (2)$$

E.g., the maximum similarity between $\hat{n}$ and all child nodes of visual entity V_i. The rational is if two visual entities share similar keyphrases or tags, they could be conceptually linked to each other and shared nodes should also be placed on similar physical coordinate. The similarity s between two nodes is calculated using equation 1. After the relative weights on all visual entities are calculated, we will have a weight vector $W = w_1, w_2, \ldots, w_M$ for node $\hat{n}$.

The next important step is use the classic Gaussian Smoothing [23] to remove the noise from the similarity calculation. It is mainly because the similarity calculation occasionally return non-zero values for strings that quite different from each other. For example, the similarity between the word *education* and *coming* is about 0.2 but they are not similar from either lexical or semantic prospect. Such small non-zero value is considered as noise and should be filtered out by Gaussian Smoothing [23]:

$$G(w_i; w_m) = \frac{1}{\sqrt{2\pi}\delta} exp\left(-\frac{(w_i - w_m)^2}{2\delta^2}\right) \tag{3}$$

where w_i is the $i-th$ relative weight on visual entity i and w_m is the maximum value of W. Then, each relative weight will be replaced by its gaussian function to the maximum weight w_m.

$$w'_i = G(w_i; w_m) \tag{4}$$

Note that δ determines the width of the gaussian function and can be adjusted by the user on the runtime. All the adjusted relative weights will be then normalized and used for position calculation, the position of $\hat{n}$ is calculated by the weighted average of

$$\hat{p} = \sum_{i=1}^{M} w'_i P_i \tag{5}$$

where $\hat{p}$ is the physical coordinates of node $\hat{n}$ and the P_i is the physical coordinates of visual entity V_i. If nodes are very close (e.g., very large relative weight) to a certain visual entity, we will claim such nodes as its private nodes and will be placed around that visual entity like its own satellites. In addition, two nodes are collided if they have very similar or identical weight vector. To solve this issue, a global collision detection mechanism has been built to slightly separate any collided nodes.

After all nodes and visual entities are properly placed, users can start utilizing CKM to check the main contents of the selected interview sessions in a very intuitive way.

2) Main Functionalities: The keyphrase map is not merely a visual presentation and summarization of the data. More importantly, it provides users with a novel way to manipulate their data and lead them to their desired information. To achieve this, we design and develop a number of interactions and functionalities for users:

- *Highlighting*: We designed four highlighting strategies: 1) Entity-nodes highlighting 2) Concept linkage highlighting 3) Similar nodes highlighting and 4) Searching highlight. Entity-nodes highlighting enables users to highlight the children nodes of a visual entity when users hover on a entity. If users move the mouse over a node or the keyphrase list view, all visual entities containing this node will be highlighted, which is called concept linkage highlighting. Meanwhile, the nodes with similar relative weight vector will also be highlighted. User can also find their desired information using keywords searching. The retrived nodes or visual entities will also be highlighted. By highlighting the edges linking relevant nodes and visual entities, we believe that users could obtain an intuitive sense about the commonness and uniqueness patterns of focused nodes on the clock. This design, as presented in Figure 3, could effectively solve the question *Q1* and *Q2*.

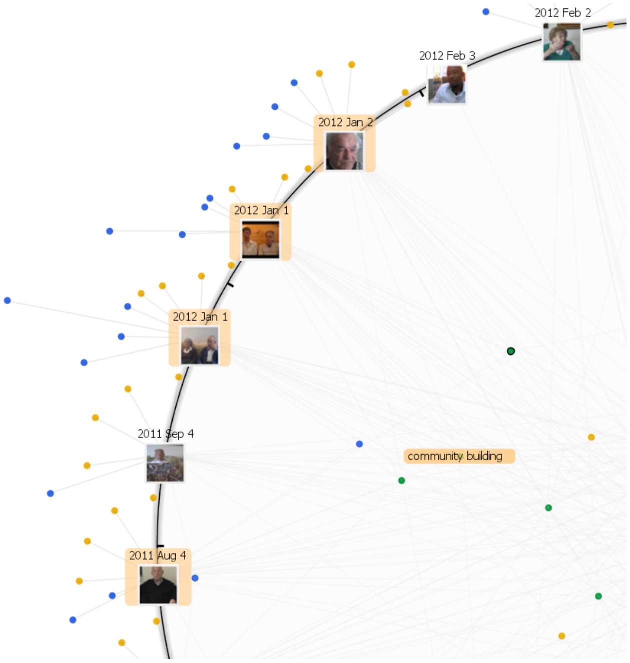


Figure 3. Zoom in to recent interviews and highlight shared nodes

- *Zoom and Pan*: When users noticed some interesting areas on the clock overview, they can zoom in to those areas for more details. As shown in Figure 3, suppose a user want to focus on the most recent interview sessions so he/she zoom in to the end part of the clock. Users are also able to drag and move the clock under the same scale level using pan operation.
- *Filtering and Reconfiguring*: The CKM can be reconfigured based on users' filtering operations. Users can conveniently set up filters and the keyphrase map will be redrawn accordingly. More interestingly, users can set up a special time filter to incrementally render the time-series visual entities. This simulates the slide show of the entire interview process. Users thus can review how their interview projects evolve. We believe such feature can be a satisfied solution to question *Q3*.
- *Context Navigation*: Another important functionality is the context navigation. Similar to the landmarks of a real map, nodes or visual entities on the keyphrase map are visual indexes leading users to their on-demand context and associated video clips. As illustrated in Figure 4, the keyphrase "Budapest“ is highlighted by a user and we can see it is shared by two interview sessions. The user then click on the node to open its

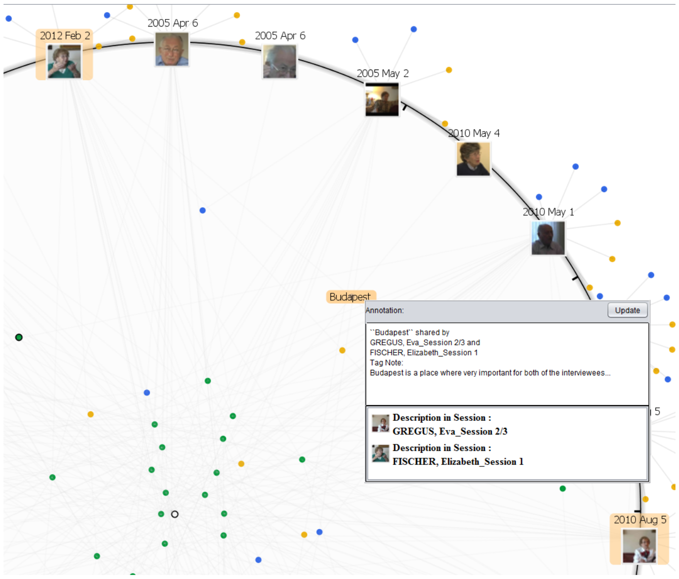
detail panel. In the detail panel, all the context and video clips related to the keyphrase are retrieved, the user can conveniently explore, comparing, and analyze them.

- *Collaborative Annotation and Feedback*: When analyzing the context of visual entity or node, users might come up with new thoughts or ideas. As such, we also build a channel for them to provide their new knowledge and feedback. Firstly, as shown in Figure 4, a node or visual entity can be annotated. In addition to annotating, users can also alert the existing the notes on the map. They can choose to generate new nodes and drag them on the proper position on the keyphrase map. Within a online community, such user-generated content will be shared and edited among different Stories Matter users, which enables a novel human-centric collaborative data processing manner. Using this human-centric collaborative design, people can conveniently and interactively share each other's analysis process, which answers the question *Q4*.

CKM is a systematic integration of text mining and information visualization and designed to significantly improve the functionalities and user experience of original Stories Matter. Such visual analytic tool could be further refined and enhanced in a number of interesting ways and also be scaled up for more general usage. Also, its effectiveness and efficiency need be evaluated in field tests.

V. Future Plan

Our next step is to invite five to ten researchers who use Stories Matter database to use our prototype for a period of time, say, two months. During this period, we will observe several sessions when the researchers use the prototype to help us understand the effectiveness and efficiency of the tool in use context. We will also interview the researchers to collect their feedback on the usability of the prototype and suggestions on the future design.



``Budapest`` shared by
GREGUS, Eva_Session 2/3 and
FISCHER, Elizabeth_Session 1
Tag Note:
Budapest is a place where very important for both of the interviewees...

Description in Session :
GREGUS, Eva_Session 2/3

Description in Session :
FISCHER, Elizabeth_Session 1

Figure 4. Highlighting and navigating to the context and associated video clip of keyphrase "Budapest"

In our previous design, tags and keyphrases are considered on the same semantic level. In fact, a tag typically represents a topic and thus should be more general than keyphrases. Their difference in generality should be better considered and organized. In addition to keyphrase extraction, we plan to use the topic modelling technologies to automatically extract the topics and their words distribution in a document. We will then create a hierarchy that each visual entity will have multiple topics and each topic will be associated with numbers of keyphrases. Once the topic hierarchies have been built, a intelligent scaling strategy, just like Google Map, will be applied. At the top level, users only see the visual entities and their topics. When zoom in, more and more detailed keyphrase will be rendered accordingly. If the number of visual entities on the clock is large, a hyperbolic-view can be utilized to dynamically visualize the complex hierarchical structure.

At last, the clock-based keyphrase map is not restricted for the Stories Matter system. We believe that any type of projects generating time-series textual dataset could benefit from this visual analytic tool. For example, suppose in an online collaborative project, each team member will generate numbers of Email, SMS, documents, meeting records, etc. These textual data could be summarized and visualized by our tool to facilitate the data processing tasks as well as team communication.

VI. Conclusion

In this paper, we present a visual analytical tool that facilitates the qualitative analysis of collections of audio/video interview recordings about human rights. Taking a participatory design approach, weve been working closely with the human rights researchers in the requirement analysis, iterative design, and evaluation process. The current prototype combines combines Keyphrase text-mining and Clock-based visualization techniques, and we plan to conduct an evaluation study of the prototype as well as exploring other text mining techniques such as topic modeling. Developed based on Stories Matter - a database developed and shared by a community of oral history researchers, our tool represents a significant step in the long-term development of the database, as an interpretative and search tool for humanities and social science researchers working with qualitative interview data. The original database was developed to better analyse oral narratives and to follow threads of significance across interviews. As a result, it was an example of little data researchers scaling up. With this new visualization tool, we see big data researchers applying some of the insights and techniques of data-mining and visualization on a qualitative research tool that operates at multiple scales from the narrative analysis of a single interview to large datasets of thousands of hours of audio or video recorded interviews. This visualization tool has the great potential to enhance the capacity of the researcher to make-sense of these wider patterns and connections.

Human rights researchers analyze the relevant data of various sorts format (e.g., textual, audio, and video) and conduct both qualitative and quantitative content analysis. With the increasingly available data sets about human rights violation and their growing sizes, navigating and analyzing such large scale data collections is more and more challenging. Our work presents a concept of proof of combining text mining and information visualization techniques to address this challenge. We anticipate the growing importance of such visual analytic tools that help users conveniently navigate to their on-demanded information and manipulate data, and that facilitate users to analyze the large scale data collection through multiple views and interactive features. There is no similar tool in use within the field of oral history. With our specific CKM design, we expect that it is applicable to different kinds of time-series sequential data stream.

Acknowledgment

We thank Kim Moore and Eve-Lyne Cayouette Ashby who participated in the first design presentation meeting as regular users and maintainers of Stories Matter database. They provided valuable feedback on our design ideas. This project is funded by SSHRC Digging into Data challenge grant.

References

[1] D. boyd and K. Crawford, “Six provocations for big data,” in *A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society*, 2011.

[2] D. Fisher, R. Deline, M. Czerwinski, and S. Drucker, “Interactions with big data analytics,” vol. 19, no. 3, pp. 50–59, 2012.

[3] D. Rotman, Y. He, J. Preece, and A. Druin, “Understanding large scale online environments with qualitative methods,” in *2013 iConference*, 2013.

[4] T. Hey, S. Tansley, and K. Tolle, *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research, 2009.

[5] J. Heer and S. Kandel, “Interactive analysis of big data,” vol. 19, no. 1, pp. 50–54, 2012.

[6] S. Kandel, A. Paepcke, J. Hellerstein, and J. Heer, “Enterprise data analysis and visualization: An interview study,” vol. 18, no. 12, pp. 2917–2926, 2012.

[7] S. Knudsen, M. R. Jakobsen, and K. Hornbæk, “An exploratory study of how abundant display space may support data analysis,” in *the 7th Nordic Conference on Human-Computer Interaction: Making Sense Through Design*, 2012.

[8] J. Livnat and Y. Zhang, “Information interpretation or information discovery: Which role of analysts do investors value more?” vol. 13, no. 3, pp. 612–641, 2012.

[9] A. Srivastava and M. Sahami, *Text Mining: Classification, Clustering, and Applications*. Chapman & Hall/CRC, 2009.

[10] I. H. Witten, G. W. Paynter, E. Frank, C. Gutwin, and C. G. Nevill-Manning, “Kea: Practical automatic keyphrase extraction,” in *The 4th ACM Conference On Digital Libraries*, 1999.

[11] C. Zhang, “Automatic keyword extraction from documents using conditional random fields,” vol. 4, no. 3, pp. 1169–1180, 2008.

[12] Y. nung Chen, Y. Huang, H. yi Lee, and L. shan Lee, “Unsupervised two-stage keyword extraction from spoken documents by topic coherence and support vector machine,” in *The 37th IEEE International Conference on Acoustics, Speech and Signal Processing*, 2012.

[13] H.-L. Ong, A.-H. Tan, J. N. Lee, H. Pan, and Q.-X. Li, “Foci : Flexible organizer for competitive intelligence,” in *The 10th IEEE International Conference on Information and Knowledge Management*, 2001.

[14] C. Jacquemin, H. Folch, and S. Nugier, “Ocean: 2 1/2d interactive visual data mining of text documents,” in *The 10th IEEE International Conference on Information Visualization*, 2006.

[15] C. H. Ku, J. H. Nguyen, and G. Leroy, “Tasc - crime report visualization for investigative analysis: A case study,” in *The 10th IEEE International Conference on Information Reuse and Integration*, 2012.

[16] W. Cui, S. Liu, L. Tan, C. Shi, Y. Song, Z. Gao, X. Tong, and H. Qu, “Textflow: towards better understanding of evolving topics in text,” vol. 17, no. 21, 2001.

[17] H. Greenspan, *On Listening to Holocaust Survivors: Recounting and Life History*. Westport: Praeger, 2010.

[18] M. Frisch, “Three dimensions and more: Oral history beyond the paradoxes of method,” pp. 221–222, 2008.

[19] S. High, M. Jessica, and S. Zembrzycki, “Telling our stories/animating our past: A status report on oral history and new media,” vol. 37, no. 3, pp. 1–22, 2012.

[20] S. High and D. Sworn, “After the interview: The interpretive challenges of oral history video indexing,” vol. 1, no. 2, 2009.

[21] K. Schaffer and S. Smith, *Human Rights and Narrated Lives:The Ethics of Recognition*. New York and Basingstoke: Palgrave Macmillan, 2004.

[22] E. Jessee, S. Zembrzycki, and S. High, “Stories matter: Conceptual challenges in the development of oral history database building software,” vol. 12, no. 1, 2010.

[23] L. G. Shapiro and G. C. Stockman, *Computer Vision*. Prentice Hall, 2001.