Technical Disclosure Commons

Defensive Publications Series

April 2021

Interactive Tool for Researching Large Unstructured Document Collections

Niv Efron

Daniel Chirpich

Jim Albrecht

Roni Rabin

Guy Ronen

See next page for additional authors

Inventor(s)
Niv Efron, Daniel Chirpich, Jim Albrecht, Roni Rabin, Guy Ronen, Amos Yoffe, Shlomo Urbach, and Tali Rosen Shoham

Interactive Tool for Researching Large Unstructured Document Collections

ABSTRACT

Reviewing large document collections is an activity that arises commonly in certain professional contexts such as investigative journalism. Such document collections can arise in many use contexts such as investigative journalism; academic research; litigation, arbitration or other legal context; audit; research using document archives; etc. The collections may include a large number of documents, including scanned images of documents or handwritten documents, and are often devoid of structure or organization. This makes it difficult to sift through such collections and identify important pieces of information. This disclosure describes a tool that enables easier access to such collections and features that support review and research based on such document collections. Automated techniques such as optical character recognition, entity recognition, indexing, etc. are utilized to process the document collection to index the documents and to generate timelines, connection graphs, or other views on the collection. A user interface is provided that enables users to search the collection, view event timelines, make annotations, take notes, and collaborate with others. The described techniques facilitate sensemaking and can help surface latent insight.

KEYWORDS

- Investigative journalism
- Document review
- Document collection
- Document archive
- Unstructured data
- Cloud storage
- Entity recognition
- Latent insight
- Entity graph
- Deep semantics
- Sensemaking

BACKGROUND

Investigative journalists deal with an ever-increasing amount of data of various types such as publicly available information, data obtained through open records requests, information obtained from confidential sources, material leaked to the press, etc. Such data can run into several terabytes and can include millions of individual documents. Typically, most such data is unstructured text that is contained in PDF files, majority of which are scans of paper documents.

Journalists work with such data by collecting it from various sources and storing it in folders and files on their computers and/or on cloud storage services. They review and analyze the documents individually or as a group. However, there is currently a dearth of tools geared toward helping journalists navigate and make sense of large document collections. Specialized applications that help journalists analyze documents typically offer only a subset of the necessary functionality. For instance, some applications are geared toward manual annotations and searching documents while others provide automation capabilities such as optical character recognition (OCR), entity detection and annotation, translation, etc.

Journalists encounter several limitations when using general-purpose applications for their work. For example, web search is restricted to data available on the open web, rather than the specific collection of documents of interest to the journalists. The navigation and search capabilities of device and cloud filesystems are geared toward files created by the user or others in the user's organization rather than the kinds of document collections typically utilized in investigative journalism. For instance, the indexing capabilities of typical filesystems may be insufficient for massive document corpora often involved in investigative journalism.

Owing to the lack of effective specialized tools, journalists often find it slow and laborious to harvest and research large document collections and connect the dots involving the

people, organizations, locations, and time periods contained within them. As a result, the societal benefits of investigative journalism work come at a rather high cost per story. Other professionals such as academic researchers, lawyers, auditors, etc. also face similar challenges when reviewing document collections.

DESCRIPTION

This disclosure describes a tool for automatically processing a collection of documents and enabling interactive user capabilities for navigating, searching, and exploring the collection to facilitate sensemaking and to surface latent insight. Such document collections can arise in many use contexts such as investigative journalism; academic research; litigation, arbitration or other legal context; audit; research using document archives; etc. The collections may include a large number of documents, including scanned images of documents or handwritten documents, and are often devoid of structure or organization. Users can use the tool with one or more document collections of interest by creating a filesystem folder corresponding to each collection of documents. For instance, an investigative journalist can employ the tool for working with various documents obtained in connection to their investigation for a news story.

When a user creates a folder with a document collection of interest, the tool is used for backend processing of the files in the collection. Such backend processing involves various automated document analysis operations, such as text recognition via OCR (if needed), entity extraction and annotation, indexing, etc. The user can then use the user interface (UI) of the tool for researching the collection in an interactive manner by leveraging various features such as navigation, search, highlighted annotations, knowledge panels, timelines, entity lists and connections, comment insertion, etc.

For instance, the user can use these interactive capabilities to gain high-level familiarity with the contents of the collection, quickly find documents connected to a specific keyword or entity, skim document contents via annotations, embed notes and comments, generate a timeline of events mentioned in the documents, generate a graph of connections among people or entities mentioned in the documents, apply one or more filters, etc. Using backend processing coupled with the interactive UI, the tool enables users to perform such sensemaking activities relatively quickly. Visual presentation of the processed information within the UI can further help surface latent insight by enabling the user to leverage visual cognitive abilities not easily feasible with the manual approach.

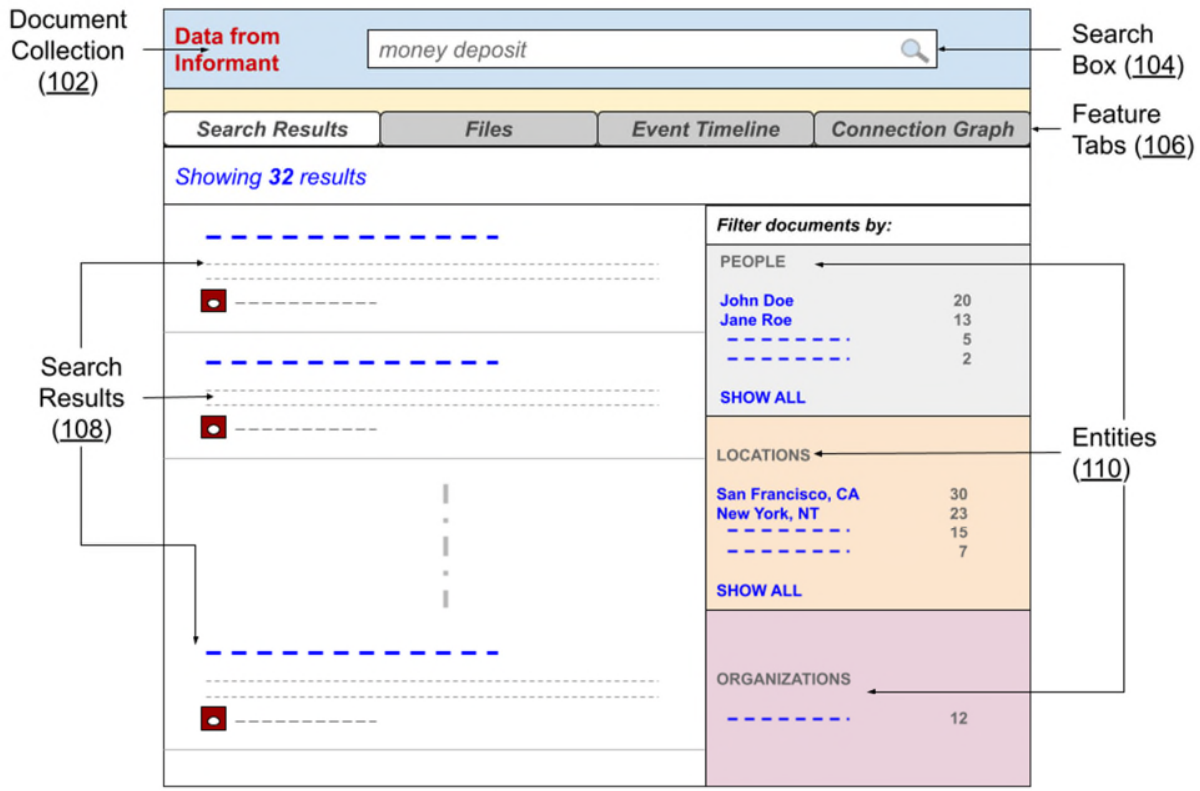

Fig. 1: User interface of the tool to research a large collection of unstructured documents

Fig. 1 shows an example of a user interface of a tool for review of large document collections per techniques of this disclosure. A user enters a query within the search box (104) to research a large collection of unstructured documents (102) loaded within the tool and automatically processed by the backend as described above.

In response to the query, the user interface is updated to include a list results (108) of the search with each result showing a title, snippet, and a link to the corresponding document within the collection. The knowledge panels on the right contain the various extracted entities (110) that occur within the documents along with the corresponding number of documents that include the entity. The user can choose to filter the obtained search results using one or more of the entities. Various other features of the tool are available via the corresponding UI tabs (106). For instance, the user can use the tabs to access individual documents, view a timeline of the events (that can include automatically identified events as well as manually added events), or explore a graph that shows connections among the various parties (or entities) in the documents.

The backend processing of the documents in the collection can be performed using suitable automated techniques for the specific purposes. For instance, entity extraction can be performed using a suitably trained machine learning model. The individual UI features can be provided using standard user interface design techniques. For instance, the search function can be invoked using a search bar and provide familiar search features, such as semantic query expansion, query corrections and suggestions, etc. The tool integrates the various backend and UI features into a single application geared toward researching document collections. The tool permits users to store a document collection in a folder in a filesystem on their device and/or in the cloud. Users can specify whether and how specific documents in the collection may be

processed automatically, set permissions for other users to access documents within the collection, place restrictions on documents, etc.

With user permission, the tool can also include collaborative capabilities to permit a group of users to work together on a document collection. Such capabilities can include common mechanisms such as shared access to the collection, shared notes and responses, chat, etc. For instance, a team of journalists can use the collaborative capabilities to work on a collection by dividing work, commenting on each other's notes, generating shared artifacts (e.g., timelines, entity annotations, etc.), etc.

The tool described in this disclosure can be implemented as a standalone application or integrated within other relevant platforms such as cloud storage, device filesystems, collaborative workspaces, etc. When integrated within other applications or platforms, the tool can appropriately leverage the capabilities of the applications or platforms for providing or augmenting the various features described herein. For example, integration within a cloud storage platform can be designed such that users of the tool can use the familiar and powerful collaboration and commenting features of the platform when performing the corresponding tasks within the tool.

The tool described in this disclosure enhances the efficiency and user experience (UX) of researching and making sense of large collections of unstructured documents, such as those typical in investigative journalism. Moreover, the UI features of the tool can help journalists surface insight that might otherwise go unnoticed. The tool can help reduce the tedium and time involved in the manual approach of analyzing documents, thus potentially reducing the overall cost per news story and contributing to more trustworthy journalism that facilitates a less corrupt society. While the foregoing discussion describes document collections in the context of

investigative journalism, other contexts such as document review during litigation, audit, research using document archives, etc. can also be supported.

Further to the descriptions above, a user may be provided with controls allowing the user to make an election as to both if and when systems, programs or features described herein may enable collection of user information (e.g., information about a user's document collections, user-generated content such as document annotations or notes, social actions or activities, profession, or a user's preferences), and if the user is sent content or communications from a server. In addition, certain data may be treated in one or more ways before it is stored or used, so that personally identifiable information is removed. For example, a user's identity may be treated so that no personally identifiable information can be determined for the user, or a user's geographic location may be generalized where location information is obtained (such as to a city, ZIP code, or state level), so that a particular location of a user cannot be determined. Thus, the user may have control over what information is collected about the user, how that information is used, and what information is provided to the user.

CONCLUSION

This disclosure describes a tool that enables easier access to document collections and features that support review and research based on such document collections. Such document collections can arise in many use contexts such as investigative journalism; academic research; litigation, arbitration or other legal context; audit; research using document archives; etc. Automated techniques such as optical character recognition, entity recognition, indexing, etc. are utilized to process the document collection to index the documents and to generate timelines, connection graphs, or other views on the collection. A user interface is provided that enables users to search the collection, view event timelines, make annotations, take notes, and

collaborate with others. The described techniques facilitate sensemaking and can help surface latent insight.

REFERENCES

1. Google News Initiative – Google News Initiative https://newsinitiative.withgoogle.com/
2. https://journaliststudio.google.com/pinpoint/