Receptor: A Platform for Exploring Latent Relations in Sensitive Documents

Hitarth Narvala
University of Glasgow, UK.
h.narvala.1@research.gla.ac.uk

Graham McDonald
University of Glasgow, UK.
graham.mcdonald@glasgow.ac.uk

Iadh Ounis
University of Glasgow, UK.
iadh.ounis@glasgow.ac.uk

ABSTRACT

Many government and public organisations have a requirement to release their official documents to the public and therefore need to review such documents to identify and protect any sensitive information that they contain. When reviewing a document for sensitivity, reviewers often use information from other documents within the collection to assist in their decisions. It can be difficult for the reviewers to find related documents in large digital collections when they are performing sensitivity review. Receptor is a new solution that aims to provide sensitivity reviewers with the ability to explore a collection of documents to discover latent relations, between for example entities and events, that can be a reliable indicator of sensitive information. The system provides novel scalable graph search and exploration functionalities as well as interactive visualisations of the latent relations between related entities, events, and documents to enable users to identify hidden patterns of sensitivity.

ACM Reference Format:
Hitarth Narvala, Graham McDonald, and Iadh Ounis. 2020. Receptor: A Platform for Exploring Latent Relations in Sensitive Documents. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20), July 25–30, 2020, Virtual Event, China.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3397271.3401407

1 INTRODUCTION

Sensitivity review is the process of manually reviewing documents to identify and protect sensitive information, such as personal or confidential information. Many government and public organisations around the world have to release their documents to the public to comply with Freedom of Information (FOI) laws. These documents may contain sensitive information and therefore need to be sensitivity reviewed prior to their release. In sensitivity review, reviewers often use information from multiple related documents and sources when they are making a judgement about sensitivity, by using known sensitive information about related entities or events from other documents. For example, information about a person having business relationships with a member of a royal family could make discussions about that person more likely to be sensitive. However, the person's relationship with the royal family may not be apparent in a single document or in the document that contains the sensitive information. Moreover, in large collections of digital documents, the documents that discuss related entities are usually distributed throughout the collection. Therefore, when performing sensitivity review, it is often not easy for a reviewer to identify the relationships that constitute a reliable indicator of potential sensitivity. Hence, there is a need for tools to help the reviewers to explore the relationships between key entities and related documents.

A range of systems have been deployed in the past to support document analysis and review (e.g. [1, 2]). However, these systems either do not principally focus on sensitivity review, lack augmentation of document search with the exploration of relationships in a collection or essentially focus on users having sufficient understanding of information retrieval (IR) concepts for analysing the documents. Instead, in this paper, we propose a new system, Receptor,[1] which aims to enable sensitivity reviewers to find and explore the latent relations in a collection for identifying potential sensitive or non-sensitive information. The name "Receptor" is derived from its scientific reference to a structure capable of receiving stimuli and generating informative impulse. The system provides predictions about whether the documents in the collection are sensitive or not sensitive and automatically extracts relations between documents, entities and events. Reviewers can explore the collection with interactive visualisations and advanced search functionalities such as entity exploration, faceted search and building complex queries to find documents that are likely to contain sensitive information. The system incorporates information extraction techniques, constructs a graph to represent extracted information and implements efficient graph search and exploration functionalities to provide reviewers with insights from the relationships.

2 RELATED WORK

A number of systems have been proposed for document review and analysis. In the following, we briefly describe the most relevant systems. Abualsaud *et al.* [1] proposed a reviewing system for high-recall IR tasks, such as electronic discovery (eDiscovery), systematic review and the construction of test collections. The system integrated an active learning classifier and search functionality with a reviewing interface that enables reviewers to navigate a collection, review documents (assisted by keyword highlighting) and judge the relevance of a document. The system is focused on processing documents efficiently and providing a user-friendly document assessment interface. However, differently from the system of Abualsaud *et al.* [1], Receptor focuses on indicating sensitivities by automatically finding latent relations between entities and events that appear in sensitive or non-sensitive documents. Recently, some efforts have been initiated to build tools to support finding sensitive information in a collection most notably, ePADD [4] and BulkReviewer.[2]

SIGIR '20, July 25–30, 2020, Virtual Event, China
© 2020 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-8016-4/20/07…$15.00
https://doi.org/10.1145/3397271.3401407

[1] Video Capture of Receptor is available at: https://youtu.be/-e6m7lRIcsc

[2] https://github.com/bulk-reviewer/bulk-reviewer

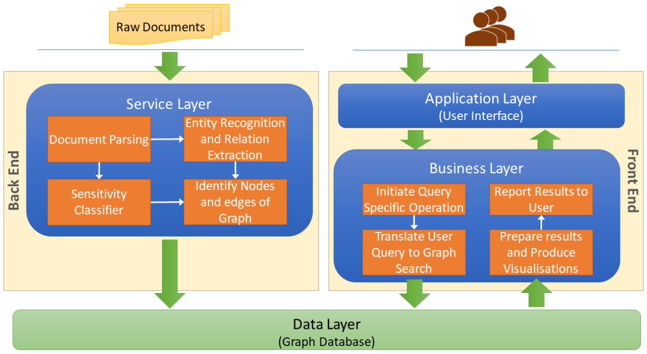
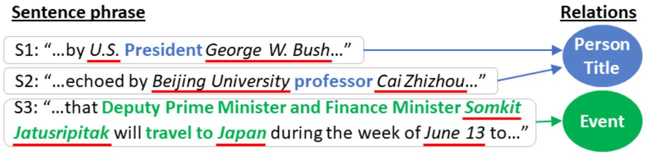

Raw Documents
Back End
Service Layer
Document Parsing
Entity Recognition and Relation Extraction
Sensitivity Classifier
Identify Nodes and edges of Graph
Application Layer
(User Interface)
Business Layer
Initiate Query Specific Operation
Report Results to User
Translate User Query to Graph Search
Prepare results and Produce Visualisations
Front End
Data Layer
(Graph Database)

Figure 1: Layered system architecture



S1: "…by *U.S.* President *George W. Bush*…"

S2: "…echoed by *Beijing University* professor *Cai Zhizhou*…"

S3: "…that Deputy Prime Minister and Finance Minister *Somkit Jatusripitak* will travel to *Japan* during the week of *June 13* to…"

Figure 2: Extracted relations for entities (underlined in red)

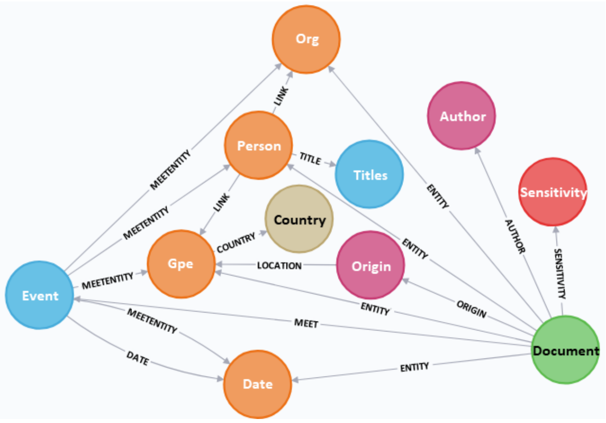


Figure 3: Graph Structure

ePADD [4] is an open-source software that supports the discovery and appraisal of email archives. ePADD incorporates natural language processing (NLP) techniques, including named entity recognition (NER), to assist users in searching within email collections. ePADD also supports search using a sensitive-lexicon to identify restricted or legally protected information across emails. However, ePADD is mainly focused on features that are typically specific to email archives, such as sender, recipient and attachments. Differently from ePADD, Receptor's key aim is to effectively capture and visualise latent relations independent of the collection type. BulkReviewer also assists a reviewer to find sensitive personal information, such as social security numbers, phone numbers and credit card numbers, within a file system. On the other hand, the system lacks generalisability to other types of sensitive information such as confidential information about health and safety or national security, whereas Receptor focuses on indicating sensitivities by extracting relations regardless of the type of associated sensitive information.

Further, Warcbase [2] and Open Semantic Search[3] were proposed to analyse collections through search and data visualisations. Warcbase [2] is a system for the exploration of Web archives, which provides temporal browsing functionalities and visualisations of Web graphs to analyse frequent mentions of named entities (person names, locations, organizations, etc.). Although, similar to ePADD, Warcbase is designed specifically for Web documents. Additionally, in comparison to Receptor, Warcbase is confined to only NER, whereas Receptor goes beyond NER and performs relation extraction and entity resolution to capture the semantic relationships between documents. Open Semantic Search is an open-source platform of research tools powered by Apache Solr, which provides utilities for text-mining and data visualisations. The system primarily supports research-related tasks that can only be performed by users having sufficient understanding of IR concepts, in contrast to Receptor, which comprises a familiar search interface with easy to use features that can be used by sensitivity reviewers without any explicit training.

3 RECEPTOR'S ARCHITECTURE

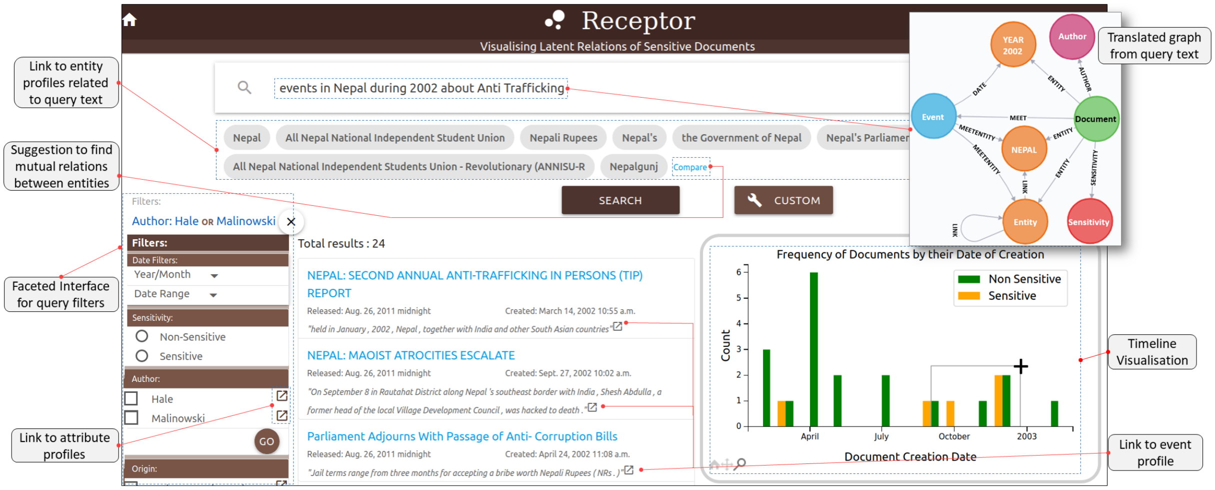
As discussed in Section 1, the goal of Receptor is to assist a sensitivity reviewer to quickly discover latent relations between documents, entities and events, which can be a reliable indicator of sensitive or non-sensitive information. To achieve this, the system must support: (1) a sensitive classifier to provide predictions of sensitive information within documents for reviewers to check, (2) the extraction of a document's attributes, entities and events, and the identification of their associated relationships, (3) a scalable and dynamic storage structure to represent the extracted relations, (4) an integrated interface for full-text and complex search to find related sensitive and non-sensitive documents and (5) interactive visualisations to explore the latent relations.

The system architecture, presented in Figure 1, is separated into four distinct layers, namely the data layer, the service layer, the business layer and the application layer:

Data layer: The data layer is built on a graph database that provides flexibility towards storing and accessing the discovered relations. The graph database is also scalable as the number of nodes and associations grows, compared to RDBMS or NoSQL databases where the scalability of associations is impacted by the schema size.

Service layer: In this layer, the source data is passed through the information extraction pipeline to perform the following tasks: (1) extracting document attributes (2) Named Entity Recognition (3) Syntactic Dependency Parsing (4) Entity Resolution and (5) Information Enrichment using external sources. The extracted elements from tasks (1) and (2) are categorized as, first, "Document Attributes" such as Creation Date, Author and Origin, and second, "Entities" such as Person, Organization and Location. In task (3), the document sentences, which contain the extracted entities are further scanned for identifying two types of relationships as illustrated in Figure 2. First, "Person-Title" to represent the title a person holds within an organisation or a location. e.g. sentence S1 in Figure 2 is used to extract the title *"President"* of *"U.S."* for *"George W. Bush"*. Second, "Event-Information" to represent a dated event involving a person, organisation and/or location e.g. sentence S3 in Figure 2 is used to extract the event context: *"Deputy Prime Minister Finance Minister and Somkit Jatusripitak travel to Japan"* between *"Somkit Jatusripitak"* and *"Japan"*. In tasks (4) & (5), the system uses external sources (e.g. Dbpedia) to enrich the extracted elements with

[3] https://www.opensemanticsearch.org/


Receptor
Visualising Latent Relations of Sensitive Documents
Link to entity profiles related to query text
Suggestion to find mutual relations between entities
Faceted Interface for query filters
Link to attribute profiles
Translated graph from query text
Timeline Visualisation
Link to event profile
events in Nepal during 2002 about Anti Trafficking
Nepal
All Nepal National Independent Student Union
Nepali Rupees
Nepal's
the Government of Nepal
Nepal's Parliamer
All Nepal National Independent Students Union - Revolutionary (ANNISU-R
Nepalgunj
Compare
SEARCH
CUSTOM
YEAR 2002
Author
Event
Document
NEPAL
Entity
Sensitivity
DATE
ENTITY
AUTHOR
MEET
MEETENTITY
MEETENTITY
ENTITY
ENTITY
LINK
SENSITIVITY
LINK
Filters:
Author: Hale OR Malinowski
Filters:
Date Filters:
Year/Month
Date Range
Sensitivity:
○ Non-Sensitive
○ Sensitive
Author:
□ Hale
□ Malinowski
GO
Origin:
Frequency of Documents by their Date of Creation
Count
0
1
2
3
4
5
6
April
July
October
2003
Document Creation Date
🟩 Non Sensitive
🟧 Sensitive

Total results : 24

NEPAL: SECOND ANNUAL ANTI-TRAFFICKING IN PERSONS (TIP) REPORT
Released: Aug. 26, 2011 midnight Created: March 14, 2002 10:55 a.m.
"held in January , 2002 , Nepal , together with India and other South Asian countries"

NEPAL: MAOIST ATROCITIES ESCALATE
Released: Aug. 26, 2011 midnight Created: Sept. 27, 2002 10:02 a.m.
"On September 8 in Rautahat District along Nepal 's southeast border with India , Shesh Abdulla , a former head of the local Village Development Council , was hacked to death ."

Parliament Adjourns With Passage of Anti- Corruption Bills
Released: Aug. 26, 2011 midnight Created: April 24, 2002 11:08 a.m.
"Jail terms range from three months for accepting a bribe worth Nepali Rupees (NRs .)"

Figure 4: Search Results

additional information. The documents are also predicted as sensitive or non-sensitive by the deployed sensitivity classifier, which is incorporated as an interchangeable component. Receptor currently integrates the classifier from McDonald *et al.* [3]. Finally, the system generates a graph representation of the collection. Figure 3 illustrates the nodes for raw documents, sensitivity predictions, document attributes (in pink), entities (in orange) and entity relations (in blue) along with the edges representing the links between the extracted element, document and relation nodes.

Business Layer: The business layer incorporates interaction with the application layer to receive user search queries and report the results. User queries are automatically translated into graph search queries according to the corresponding query operation such as full-text search, event search or searching for mutual attributes between multiple entities. The query results are reported back to the application layer along with meaningful visualisations.

Application Layer: The application layer comprises a web-based user interface, which enables users to effectively perform exploratory search. The interface presents the resulting documents with enriched information through network visualisations showing relations between documents and key entities as well as through a timeline visualisation of related documents based on their creation date.

Receptor is implemented in python. spaCy[4] is used for all of the NLP operations and the web-interface is implemented using the Django[5] framework. The graph database is supported by Neo4j.[6]

4 KEY FUNCTIONALITIES

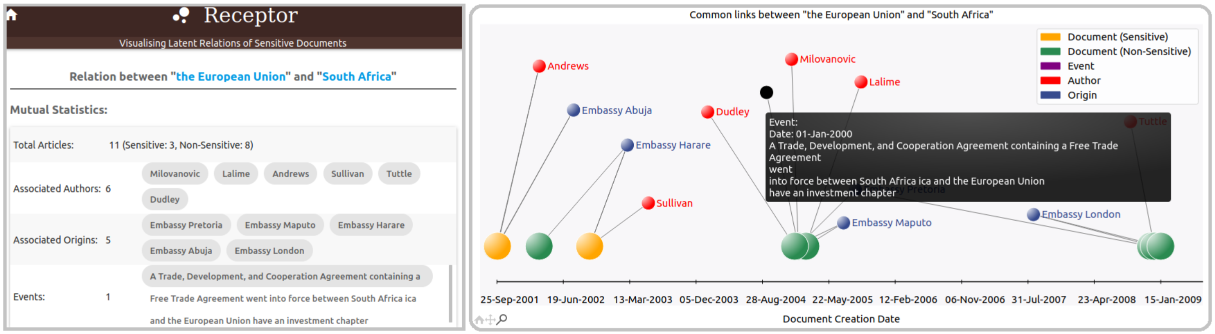
Receptor provides numerous functionalities as per the above architecture discussion, which we summarise below.

Exploratory Search: As previously discussed, reviewers need to find related documents to analyse which documents, entities or events are indicators of sensitive information. The system integrates three distinct types of search functionalities. First, traditional search with simple textual queries implemented using full-text indexing. Second, faceted search with various search filters such as document creation date, authors and origins. Third, custom query builder to express *complex queries* with boolean conditions for searching documents related to particular attributes, entities or events. Figure 6 shows the query builder interface with various options as search conditions and the visualisation of the query syntax to explain the query to the user. The complex queries from the faceted search and query builder are implemented as graph search functionalities by automatically translating the text query into nodes and edges of the graph and efficiently traversing the graph structure to provide relevant results to the reviewer. For example, Figure 4 illustrates how the text query *"events in Nepal about Anti Trafficking during 2002"* is automatically mapped into a graph comprising document nodes relevant to the keyword *"Anti Trafficking"*, which are linked to event nodes having links to entity node *"Nepal"* and all the date nodes from year *"2002"*. Author nodes are also included to map the specified filter condition.

Profile Generation: Elements such as entities, authors or origins can be associated with many documents in a collection, and some of these elements can have more links to sensitive documents. Therefore, reviewers often need to know how a particular attribute or an entity appears in the collection to analyse potential sensitivities. Receptor captures this information as *profiles* of individual document attributes, entities and events. For example, a person profile will comprise details about which documents mention the particular person, which authors/origins produced these documents, what is the timeline of creation of these documents as well as which of these documents are sensitive and non-sensitive. As shown in Figure 4, reviewers can access the profiles from the highlighted entities and events in the search results and can also explore the mutual relations between two profiles, for example common documents and events between a country and an organisation as illustrated in Figure 5(a).

Interactive Visualisation of Latent Relations: Illustration of the latent relations between documents as well as entities is essential to quickly gain insights over the extracted elements and their associated documents. This is supported by providing two types of interactive visualisations. First, a timeline visualisation of documents in search results and profiles to show the frequency of documents created in a particular time frame. Timeline visualisation as shown in Figure 4 can be interacted with by clicking through a time frame to filter the search results by creation date and/or by the predicted sensitivities. Second, a network visualisation of the profiles to illustrate links between entities, events, document attributes and their related documents. Network visualisations as shown in

[4] https://github.com/explosion/spaCy
[5] https://djangoproject.com
[6] https://neo4j.com/neo4j-graph-database



Receptor

Visualising Latent Relations of Sensitive Documents

Relation between "the European Union" and "South Africa"

Mutual Statistics:

Total Articles:	11 (Sensitive: 3, Non-Sensitive: 8)	
Associated Authors:	6	Milovanovic Lalime Andrews Sullivan Tuttle Dudley
Associated Origins:	5	Embassy Pretoria Embassy Maputo Embassy Harare Embassy Abuja Embassy London
Events:	1	A Trade, Development, and Cooperation Agreement containing a Free Trade Agreement went into force between South Africa ica and the European Union have an investment chapter

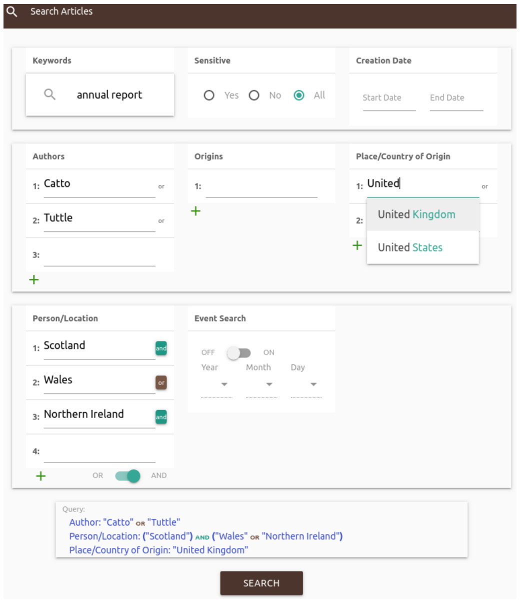
a.

Common links between "the European Union" and "South Africa"

Event:
Date: 01-Jan-2000
A Trade, Development, and Cooperation Agreement containing a Free Trade
Agreement
went
into force between South Africa ica and the European Union
have an investment chapter

b.

Figure 5: (a) Mutual relations between two entities; (b) Network Visualisation of latent relations



Search Articles

Keywords: annual report

Sensitive: ○ Yes ○ No ● All

Creation Date: Start Date End Date

Authors: 1: Catto or; 2: Tuttle or; 3:

Origins: 1:

Place/Country of Origin: 1: United or; 2:

United Kingdom

United States

Person/Location: 1: Scotland and; 2: Wales or; 3: Northern Ireland and; 4:

OR AND

Event Search: OFF ON; Year Month Day

Query:
Author: "Catto" OR "Tuttle"
Person/Location: ("Scotland") AND ("Wales" OR "Northern Ireland")
Place/Country of Origin: "United Kingdom"

Figure 6: Custom Query Builder

Figure 5(b) capture the search results of the underlying graph and can be interacted with by clicking to access the profile or document content depending on the node type or by dragging the nodes to promptly discover any trends such as elements having more links to sensitive or non sensitive documents. Both visualisations represent sensitive and non-sensitive documents in different colours to distinguish them easily. Moreover, the documents are arranged in the order of their creation date to maintain the chronology.

Query suggestions and automatic query generation: Sensitivity reviewers are often unaware of information in the collection that could provide valuable additional context for their review. For example, if a document discusses the illegal activity of a named individual, it could be important to know if the individual has had close ties with senior political figures. Hence, the reviewers require assistive inputs to enhance the formulation of queries from their information needs. The system provides query suggestions to assist reviewers by checking if the initial text query contains any entity or event information and suggests them to search for any events related to the entity or to compare multiple entity profiles to illustrate their mutual relations. The system further automatically re-writes the query corresponding to the user's requirement. The system also intercepts other information like event dates from the textual query to narrow down the search results.

5 EVALUATION

We performed a user study with 14 participants to evaluate Receptor where the users were provided with various search tasks comprising both full-text search and graph search. User behaviour and actions were logged to measure the CPU time for end-to-end query execution. We specifically evaluated the latency for queries producing large results i.e. 500 to 1000 documents. The mean latency for queries with full-text search was found to be 0.5040s with an average result count of 777.125, while the graph search queries executed with a mean latency of 0.4725s with an average result count of 819.5. This clearly depicts the efficiency of the implemented graph search functionality in comparison to the full-text search even for larger sets of results. Qualitative analysis was also performed through a user questionnaire to rate the relevance of results on a 5-point scale. Our participant users rated the provided results as 90.47% relevant.

6 CONCLUSIONS

In this paper, we have presented Receptor, a system to assist sensitivity reviewers with searching large collections to find latent relations. The system supports a modular architecture with a seamless integration of information extraction and scalable graph search techniques. The web-based interface of Receptor provides advanced search functionalities and interactive visualisations of latent relations between the documents, entities and events to enable reviewers to explore the data and determine any trends in sensitive information. The system can be deployed, for example, in governments or public organisations to assist with sensitivity review. Moreover, its modular architecture enables Receptor to be tailored as a general collection enrichment and knowledge discovery solution. As future work, we plan to investigate additional approaches for extracting latent relations, for example, by developing reinforcement learning approaches.

REFERENCES

[1] M. Abualsaud, N. Ghelani, H. Zhang, M. Smucker, G. Cormack, and M. Grossman. 2018. A System for Efficient High-Recall Retrieval. In *Proc. SIGIR 2018*.

[2] J. Lin, I. Milligan, J. Wiebe, and A. Zhou. 2017. Warcbase: Scalable Analytics Infrastructure for Exploring Web Archives. *J. Comput. Cult. Herit.* 10, 4 (2017).

[3] G. McDonald, C. Macdonald, and I. Ounis. 2017. Enhancing Sensitivity Classification with Semantic Features Using Word Embeddings. In *Proc. ECIR 2017*.

[4] J. Schneider, C. Adams, S. DeBauche, R. Echols, C. McKean, J. Moran, and D. Waugh. 2019. Appraising, processing, and providing access to email in contemporary literary archives. *Archives and Manuscripts* 47, 3 (2019), 305–326.