Article

In the Shadow of Goitein: Text Mining the Cairo Genizah

Christopher Stokoe, Gabriele Ferrario, and Ben Outhwaite | Cambridge

Abstract

The widespread digitization of manuscripts has brought about an era of unprecedented access to a range of important historical collections. However, the lack of substantive metadata associated with these online digital collections represents a significant barrier to those wishing to navigate them in order to identify manuscripts relevant to a particular research question or theme. We propose a novel solution to cataloguing based around text mining published editions, commentaries and other secondary literature in order to automatically generate a rich searchable electronic catalogue. This research explores a range of techniques from the fields of Information Retrieval (term-weighted vocabularies), Natural Language Processing (named entity recognition) and Text Analysis (topic models). Our initial results demonstrate the potential for these approaches to produce significant volumes of descriptive metadata which, when evaluated in the context of retrieval effectiveness, provide suitable evidence on which to perform analysis and make discoveries. A search engine which recommends manuscripts based on the contents of our automatically derived catalogue achieves a Precision @ 10 of 0.54, which significantly beats a baseline strategy of random selection.

1. Introduction

The Taylor-Schechter Genizah Collection at Cambridge University Library is the single most important collection of medieval Jewish manuscripts in the world.[1] As of June 2013, its 193,000 manuscripts have now been completely digitized and are in the process of being made available online as part of the Cambridge University Digital Library.[2]

Whilst mass digitization has significantly improved access to the collection, it is clear that discovery – the act of directing researchers to a particular manuscript that will answer a given information need – remains a key challenge. This is largely due to the sheer size of the collection coupled with the lack of any substantive metadata describing the content of individual manuscripts. The inability to navigate the collection by content presents a substantial roadblock to the diverse group of scholars looking to exploit this unique source of information about the history of the Mediterranean and Near East. The following catalogue entry describes fragment T-S 24.64 (see fig. 1), which aptly demonstrates the full extent of the problem:

> *T-S 24.64 — letter*
> 55 × 14; 74 lines + marginalia (recto); 7 + 3 lines (verso)
> Paper; 1 Leaf; Torn; Judaeo-Arabic
> A lengthy letter from Ḳalaf b. Isaac to Abraham b. Yiju, middle of 12th c.

Manual efforts to improve the quality of our catalogue by the Genizah Research Unit (GRU) are ongoing, but the scale (in terms of number of manuscript fragments) and complexity (manuscript condition, language constraints and required subject expertise) make the cost of full description, transcription and translation prohibitive. In light of this, we have elected to explore the potential for a technology-based solution.

2. Related work

Recent advances in technology have opened up the possibility of automatically deriving catalogue data from digital images of manuscripts. In particular, the Friedberg Genizah Project[3] has experimented with extracting the shape, size and con-

[1] See Reif and Reif 2002.

[2] Cambridge University Digital Library (2014), URL: http://cudl.lib.cam.ac.uk (accessed on March 14, 2014).

[3] Shweka, Choueka, Wolf, and Dershowitz 2013.

Fig. 1: T-S 24.64, a letter from Ḵalaf b. Isaac to Abraham b. Yiju, middle of 12th c.

dition of a manuscript through the use of image analysis and machine learning. Another recent effort[4] attempted to analyze handwriting features and scribal practice as a means of determining authorship, date and point of origin. Whilst these techniques provide one potential source of descriptive metadata about the physical artifact, they do little to unlock the content of the manuscripts. Most notably, optical character recognition (OCR) accuracy on handwritten manuscripts remains problematic for retrieval purposes.[5] Even if OCR accuracy could be improved, machine translation remains out of reach because of a lack of parallel corpora for the vast range of languages present in the Genizah (in particular, Judeo-Arabic).

3. Methodology

Our approach centers on exploiting the 110 years of scholarship that surrounds the Genizah in order to automatically derive a content-based catalogue from the secondary literature. Extensive written material has been published about the manuscripts in the form of published editions, commentaries and academic papers, many of which include full text translation. Through the use of text mining, we propose taking this rich source of knowledge and using it to produce metadata that will facilitate content-based retrieval of data on the manuscripts.

Our methodology consists of the following steps:

1. Bibliometric analysis in order to identify a suitable corpus of secondary literature.
2. Corpus construction involving OCR and automatic segmentation of the text.
3. Term weighting through association of text from the corpus to a specific fragment.
4. Extraction of named entities and temporal expressions from associated texts.
5. Classification of fragments using topic modeling.

The resulting catalogue has been indexed using a search engine and we have evaluated the output in the context of its potential to provide adequate evidence for resource discovery.

3.1 Bibliometric analysis

For the past 30 years the GRU has been compiling an extensive bibliography[6] by tracking those scholarly works that make use of Genizah fragments as their primary source material. As of December 2013, this bibliography contains over 113,786 citations to 64,265 unique manuscript fragments across 3,643 scholarly works. As part of this research, we have undertaken a detailed analysis of this dataset focusing on the number of fragments cited in each work, co-citation of fragments by author and the co-occurrence of fragments across scholarly works. These measures combine to give us a picture of which authors and works to target in order to maximize our coverage of the collection.

Fig. 2 shows a visualization of the citation information for the fragments in the Taylor-Schechter collection based upon clustering co-occurrence across scholarly works. Note

[4] Levy, Wolf, and Stokes 2013.

[5] Naji and Savoy 2011.

[6] Genizah Research Unit Bibliography (2014), URL: http://cudl.lib.cam.ac.uk/bibliographies/genizah (accessed on March 14, 2014).

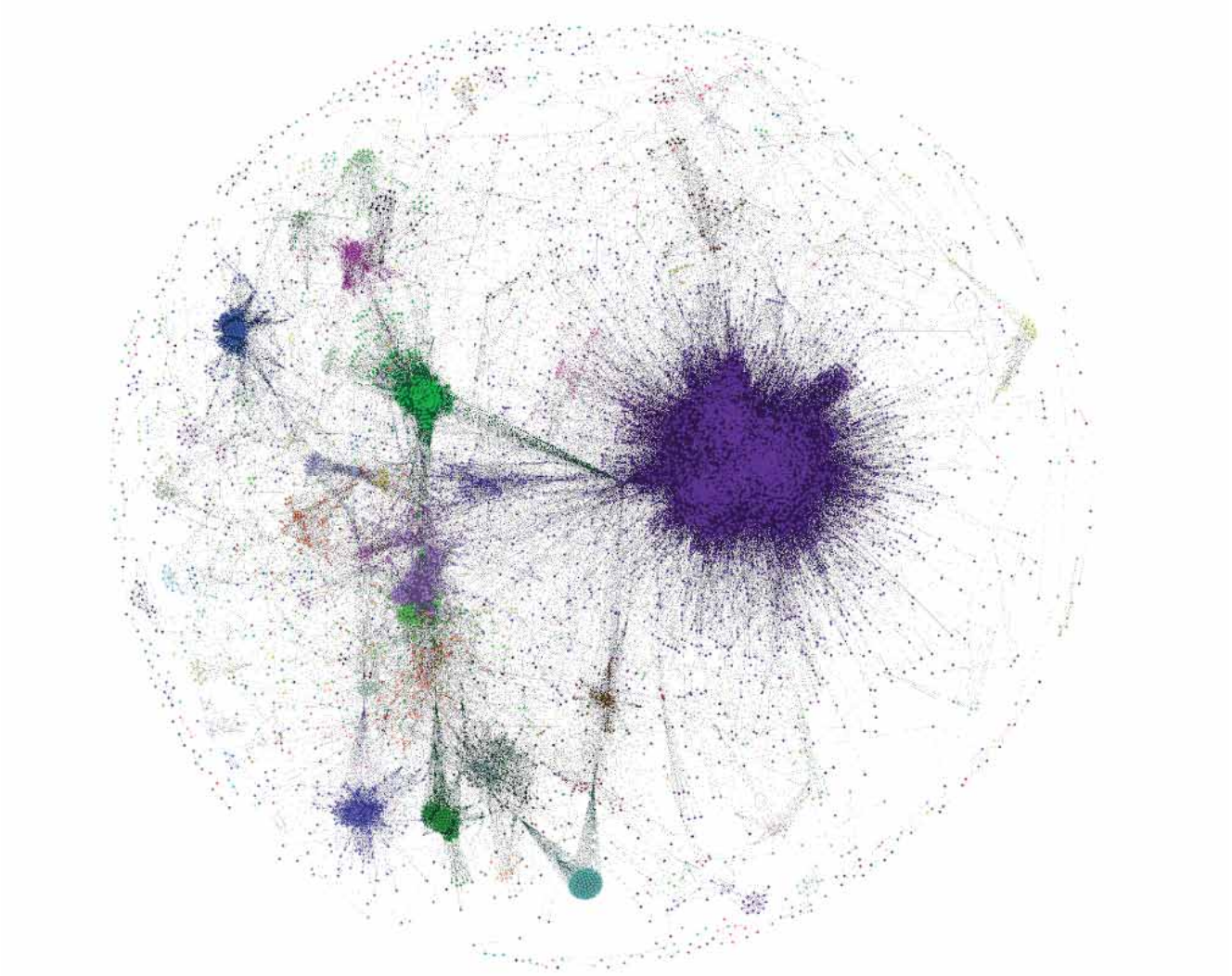

Fig. 2: Visualization showing Genizah fragments clustered based upon co-occurrence in the secondary literature.

that the largest cluster of fragments belongs to a subset of the Genizah that the literature[7] describes as documentary. These fragments consist of the everyday ephemera of life in the classical Genizah period, e.g. letters, accounts, merchants' papers, court depositions and other day-to-day writings. Our analysis clearly shows that the largest single contribution to the literature, measured in terms of fragments discussed, comes from the writings of Shelomo Dov Goitein, principally his six-volume work *A Mediterranean Society: the Jewish communities of the Arab world as portrayed in the documents of the Cairo Genizah*.[8] In addition to the scholarship surrounding the documentary Genizah, there are several other distinct clusters that represent the breadth of Genizah studies, which includes magic, literary works, medicine, liturgy and religious law. In order to maximize our coverage of the collection, we have tried to target a cross section of works that encompasses all of these clusters.

If we consider the example of manuscript T-S 24.64, then our analyses of the bibliography identified 15 scholarly works that are known to have discussed this fragment to varying degrees. The full texts of seven of these were available to us for inclusion within our corpus.

3.2 Corpus construction

The process of building our corpus is ongoing, and scholarly works continue to be added as and when we clear the rights. Accessioning new texts involves format shifting the source material into a machine-readable format (UTF-8) and then automatically segmenting the raw text according to its structure (e.g. page boundary, subsection, chapters). As of December 2013, our corpus contains 38 scholarly works by

[7] Frenkel 2010.

[8] Goitein 1967-1993.

Fig. 3: The top 50 terms associated with T-S 24.64 scaled according to term weight.

25 different authors and represents over 6,500 pages of text about the Genizah. This includes Goitein's *A Mediterranean Society* as well as works by other prominent Genizah scholars including Moshe Gil and Jacob Mann. In total, our corpus contains references to 6,322 fragment classmarks.

Many of the recent works have a native digital edition, but where an electronic source is not available we have had to resort to using a cradle scanner to image a physical copy and then OCR in order to produce machine-readable text. Therefore, approximately half of our corpus consists of text produced using an OCR process that has a reported error rate of approximately 3% when applied to printed material.

3.3 Deriving a term-weighted vocabulary

Using the citation information from the bibliography along with pattern matching for classmark recognition, we automatically associate blocks of text from the secondary literature with a given fragment. With regard to context, our approach tries to identify the boundaries of the paragraph containing the classmark, but if one cannot be detected, then it defaults to including the whole page. Once this mapping is performed, our system generates a vocabulary of the words used to describe the fragment and assigns a set of term weights.

The term-weighting scheme we have used is length-normalized TF-IDF,[9] which in this instance represents a single value measure of the importance of a word to a given fragment relative to the frequency of the word across all fragments. Term-weighting measures are the foundation of the vector space model[10] used in modern information retrieval systems and thus provide a representation which is suitable to enable 'full text' search of the fragments. Fig. 3 shows an example of the resulting vocabulary for manuscript T-S 24.64 expressed as a word cloud with the terms scaled according to their term weight. In total, the vocabulary of terms that we have associated with this manuscript fragment is 3,802, but for the purposes of our experiments we have only used the 50 most commonly occurring terms for inclusion in our catalogue. As we can see from fig. 3, the top 50 terms closely track what little we know about the manuscript, but we have significantly expanded the number of terms that would cause

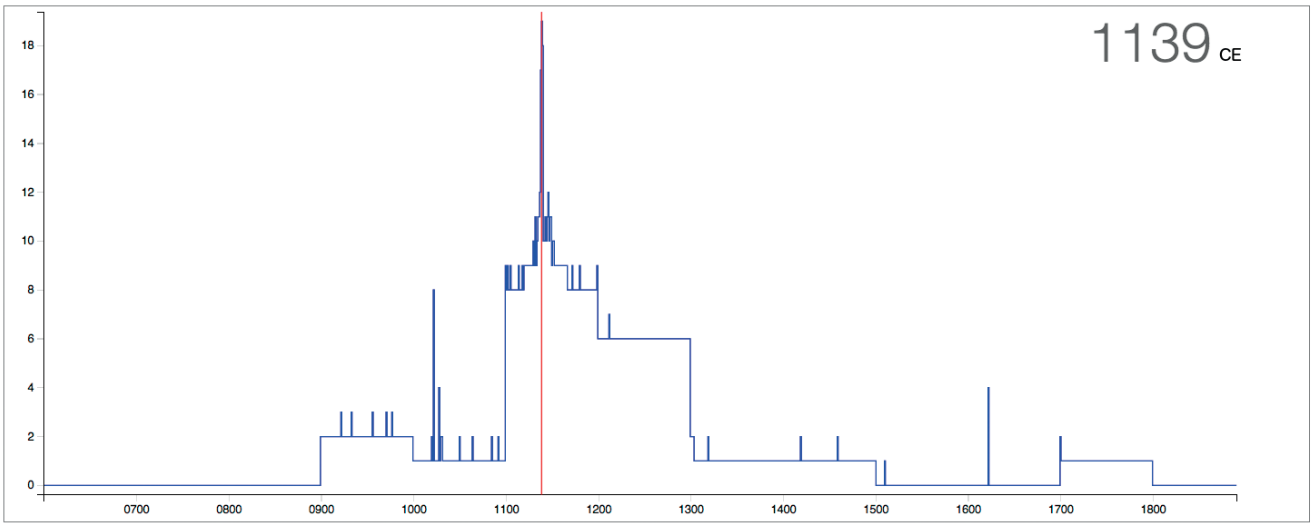

Fig. 4: Temporal graph showing dates associated with T-S 24.64.

[9] Salton and Buckley 1988.

[10] Salton, Wong, and Yang 1975.

this fragment to be surfaced within a search engine or the discovery layer of a library catalogue.

Whilst we recognize that a bag-of-words approach is not necessarily optimal in terms of producing a human-readable catalogue, it is important to highlight the fact that most modern discovery layers simply reduce any free text to this form of representation for the purposes of performing a search. As such, unless a library chooses to surface its raw catalogue information, the user is not necessarily aware that their search results are based upon automatically derived catalogue data.

3.4 Information extraction

Taking our approach a step further, we then attempted information extraction from the blocks of text we had associated with each fragment. In particular, we were interested in extracting any proper names and locations as well as any dating information in the form of temporal expressions.

Our approach to named entity recognition was based on the implementation of a Conditional Random Field (CRF) classifier contained in the Stanford coreNLP toolkit.[11] This is a supervised machine learning approach which required training using a gazetteer of medieval Muslim and Jewish personal names augmented to account for the wide variance in accepted spelling/transliterations commonly found in the literature. We associated names with over 3,500 fragments with varying degrees of success. For fragment T-S 24.64 we identified seven names, which were either derivatives of the known authors or closely related family members.

For dating, we developed a simple rule-based approach (based loosely on the techniques described in Mani and Wilson 2000) in order to extract temporal expressions from the text, focusing on assigning a creation date to each fragment. An extensive list of hand-crafted rules attempts to account for the range of dates present in scholarship which documents a period of over 1200 years using at least three disparate calendar systems. An example of the resulting temporal graph for fragment T-S 24.64 is shown in fig. 4. We can see that we identified a number of potential candidate dates from the literature, but a candidate date of 1139 CE appears over 18 times.

3.5 Topic modeling

By hand-labeling the clusters of fragments that we identified from the bibliography (section 3.1, fig. 2), it was possible to assign an approximate classification to each fragment. This approach only yielded an extremely coarse-grained view of the collection, however.

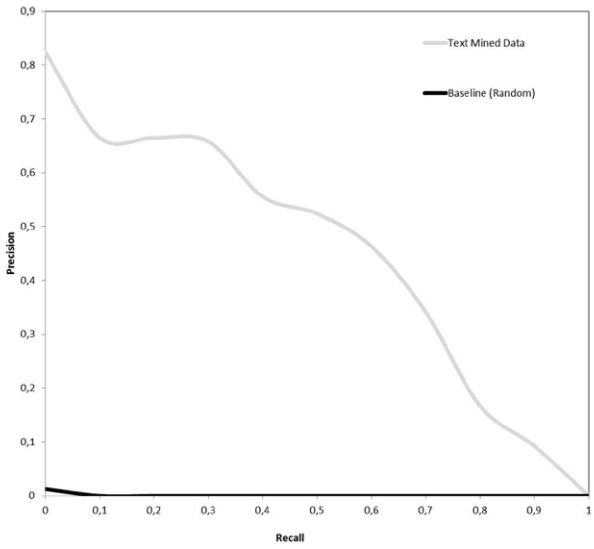

Fig. 5: Interpolated precision / recall graph contrasting retrieval effectiveness of our baseline vs. the text-mined data.

A more fine-grained classification approach was to use Topic Modeling (specifically Latent Dirchlet Allocation[12]) to identify clusters of words that commonly co-occurred within our corpus of secondary literature. Having identified a set of 75 topics, we asked a subject expert to hand-label these based on the top 50 terms most frequently associated with each topic. These labels then became the basis of our classification system. We then applied our topic model to the term-weighted vocabulary that we had generated for each fragment, thus allowing us to fit the fragments to our subject-based classification scheme. The resulting classifications provide a suitable mechanism to browse the collection using content-based themes and highlight the diverse range of topics covered in the Genizah (e.g. travel, religion, society, business, finance, medicine and the occult).

4. Evaluation

The main aim of this work was to generate usable catalogue data to facilitate text retrieval on the fragments. Our approach to evaluation has followed the Cranfield paradigm,[13] which is widely used in information retrieval. Put simply, we have taken a number of real-world queries and made use of a subject expert to identify a set of known relevant fragments. Consider the following query:

[11] Finkel, Grenager, and Manning 2005.

[12] Blei et al. 2003.

[13] Voorhees and Harman 2005.

<title>sufism or mysticism
<narr>Texts referring to Sufi practices, to mystical practices or to interest in 'sodot', qabbala, numerology etc

The title field represents the keywords that a user might enter into a search engine, whilst the narrative contains the guidance provided to the subject expert to help interpret the underlying information need.

Using the metrics of precision and recall, the graph in fig. 5 contrasts the retrieval effectiveness of search results based upon our text-mined data against a baseline strategy of returning 100 random fragments from the 6,322 for which we have catalogue data. As expected, the baseline performs badly with the likelihood of returning a relevant document by random selection being less than 1 in 1000. Contrast this with the results of searching our catalogue and we can see that this approach provides a significant source of evidence as the basis for discovery. In terms of precision @ 10 documents retrieved (which effectively models the first page of a search engine's results), then on average 1 in 2 documents returned were judged relevant by a subject expert, and this holds true all the way to precision @ 30 documents retrieved.

5. Conclusions

In this paper we have outlined a methodology for combining rich citation data with a corpus of secondary literature in order to automatically generate a content-based catalogue for the Taylor-Schechter collection. Our evaluation demonstrates that this approach has produced a weighted vocabulary for 6,322 fragments that, when used as the basis of performing retrieval, significantly outperforms a strategy of randomly selecting fragments. Given the sparseness of existing metadata for this collection, any solution that can recommend relevant fragments is a significant step forward. In this context an average of five relevant results in the first ten retrieved is extremely encouraging. Our exploration of Named Entity Recognition and Topic Modeling has been positive, but our ability to evaluate the work is limited by the lack of a ground truth or gold-standard metadata. Our next step is to engage with several pilot communities in order to produce an appropriate test collection to evaluate these elements more formally.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the support of the Andrew W. Mellon Foundation, which is providing funding for Cambridge University Library's project 'Discovering history in the Cairo Genizah' (2012–14).

REFERENCES

Blei, D., et al. (2003). 'Latent Dirichlet allocation', *The Journal of Machine Learning Research*, 3.4–5: 993–1022.

Cambridge University Digital Library (2014), URL: http://cudl.lib.cam.ac.uk (accessed on March 14, 2014).

Finkel, J., Grenager, T., Manning, C. (2005), 'Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling', in *43rd Annual Meeting of the Association for Computational Linguistics*, 363–370.

Frenkel, M. (2010), 'Genizah Documents as Literary Products' in B. Outhwaite and S. Bhayro (eds.), *From a Sacred Source: Genizah Studies in Honour of Professor Stefan C. Reif* (Cambridge Genizah Studies Series, 1), 139-156.

Genizah Research Unit Bibliography (2014), URL: http://cudl.lib.cam.ac.uk/bibliographies/genizah (accessed on March 14, 2014).

Goitein, S. D. (1967–1993), *A Mediterranean Society: the Jewish communities of the Arab world as portrayed in the documents of the Cairo Genizah*, 5 vols. and index vol. (University of California Press).

Levy, N., Wolf, L., Stokes, P. (2013), 'Document classification based on what is there and what should be there', in *Digital Humanities 2013: Conference Abstracts* (Lincoln, NE: University of Nebraska–Lincoln), 279–82.

Mani, I., and Wilson, G. (2000), 'Processing of News', *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL2000)*, 69–76.

Naji, N., and Savoy, J. (2011), 'Information Retrieval Strategies for Digitized Handwritten Medieval Documents', in *Proceedings of the Asian Information Retrieval Symposium*, Dubai, LNCS #7097 (Berlin: Springer), 103–114.

Reif, S., and Reif, S. (2002), *The Cambridge Genizah Collections: Their Contents and Significance* (Cambridge University Press).

Salton, G., Buckley, C. (1988), 'Term-weighting approaches in automatic text retrieval', *Information Processing and Management,* 24.5: 513–523.

——, Wong, A., and Yang, C. S. (1975), 'A Vector Space Model for Automatic Indexing', *Communications of the ACM*, 18.11: 613–620.

Shweka, R., Choueka, Y., Wolf, L., Dershowitz, N. (2013), 'Automatic extraction of catalog data from digital images of historical manuscripts', *Literary and Linguistic Computing,* 28.2: 315–330.

Voorhees, E. and Harman, D. (2005), *TREC: Experiment and Evaluation in Information Retrieval* (The MIT Press).