Computational Manuscriptology

Nachum Dershowitz*

Abstract

Written text is an ideal source for understanding much about life in the historical past. Manuscripts, including community documents, personal letters, and commercial records all contribute to a fuller understanding of a given place and time. The Digital Age we live in has brought with it the large-scale digitization of such historical records as well as various artifacts. Many large collections of documentary material are currently being digitized around the globe and made available on the Internet. Examples include: 350,000 fragments of discarded medieval codices, scrolls, letters, and documents discovered in the 1890s in a *genizah* repository in the attic of a synagogue in old Cairo and now scattered in over 75 libraries and collections around the world, which have been digitized (in full color, recto and verso, at 600dpi) by the Friedberg Genizah Project; 2,000,000 images and transcriptions of 70,000 pre-1900 Taiwanese deeds and court papers in the Taiwan History Digital Library; more than 15,000 Dead Sea Scroll fragments now undergoing multispectral imaging by the Israel Antiquities Authority; countless Arabic, Sanskrit and Chinese manuscripts; Tibetan Buddhist manuscripts and xylographs; Greek and Coptic papyri; and much more. Thus, the modern scholar of history or of other disciplines in the humanities is often blessed with easily accessible images of hundreds of thousands of readily-available and potentially-relevant full or fragmentary documents, as well as numerous transcribed texts. Unfortunately, until recently digital tools for analyzing such material have been extremely weak. Without appropriate computer aids, the scholar is faced with the challenge of isolating the sought-after needles in a proverbial haystack of online images and texts. That situation is on the verge of significant change. Though state-of-the-art optical character-recognition (OCR) still comes nowhere near providing quality searchable texts for such historical handwritten material, there are now prototypes of a formidable array of other computational tools that can be brought to bear on such documents to aid scholars in their research. In the paragraphs that follow, we briefly describe several such tools on the development of which we have been working, in collaboration with Lior Wolf and other colleagues and students.

2014 DADH

I. Measuring

A relatively simple image-processing task is to infer various physical parameters

* Professor, Computer Science, Tel Aviv University, Israel. Email: nachumd@tau.ac.il.

of a document from its image, measurements such as page size, written area, color, material, line density, and the like, many of which are normally included in catalogs and academic publications. This has been done successfully for the digitized Genizah collection [8]. Using a contrasting background color (like RGB 0, 120, 255) when digitizing and including a (contrasting-colored) ruler and color patch in the image makes such tasks much easier.

II. Paleography

Modern machine-learning methods may be used to classify manuscripts in paleographic terms, according to writing styles or provenance. A classifier may be built from training examples, using features like SIFT descriptors. Or the same features may be used to identify "neighboring" texts from some suitable collection of candidates, and help expedite the paleographic classification of manuscripts in that way. Another useful tool might aggregate graphemes into clusters that can be analyzed for variability. We are applying such methods of digital paleography to Cairo Genizah [9], Dead Sea Scroll [4], and Tibetan [3] corpora.

III. Spotting

Spotting is the task of visually retrieving words or graphemes within a corpus of manuscript images, beginning with a query image that is either cut out of one image or created synthetically in the writing style of the corpus. This topic is the subject of much recent research, including our own prize-winning prototype [6], which allows for extremely rapid querying, while still maintaining high accuracy. The dataset images that are to be queried are preprocessed by a simple binarization operation, followed by the extraction of multiple overlapping candidate targets. Each binary target, as well as the query, is resized to fit a fixed-size rectangle and represented by conventional image descriptors. Then, a cosine similarity operator—followed by maximum pooling over random groups—is used to represent each target or query as a concise 250 dimension vector. Retrieval is performed in a fraction of a second by nearest-neighbor search within that space, followed by an easy suppression of extra overlapping candidates.

2014DADH

IV. Joins

We have developed a scheme for image analysis that enables the computer to compare two images and compute a similarity score based on general handwriting style and also on inferred physical attributes [10]. This has resulted in a coherent and efficient system that helps scholars find candidates for potential joins (that is, matches between leaves written in the same hand that were originally part of the same manuscript). Given a fragment, the tool—now available to scholars on the Genizah website—retrieves and ranks other fragments for similarity [1]. In this way, we have been able to identify thousands of new joins, many of interest to scholars [7], this despite the fact that these documents have already been extensively studied

throughout the twentieth century. We have also investigated interactive re-ranking of results of a given query. The reranking makes use of the similarities between the various results themselves, without employing any additional sources of information. Graphical Bayesian models are used to reinforce retrieved items strongly linked to other retrievals, and repeated clustering measures the stability of the obtained associations, and an active relevance-based reranking process leverages true matches that have low similarity to the query [2]. The same tools are applicable to other collections as well.

V. Intertextuality

Various tools are available for working with texts that have been digitized—typically by manual transcription or a combination of mediocre OCR and manual correction. For example, there are alignment algorithms (such as CollateX) for matching up multiple recensions of the same text, which can help in the preparation of a synoptic critical edition. A related problem, for which tools are not yet available, is that of finding parallel passages (of any desired length) in different texts. Approximate text matching (using dynamic programming and/or indexing) can be used to locate all *similar* passages of one work in another or within a given large corpus. The results can also be used to uncover orthographic peculiarities and systematic lexical variations. Adding morphological, lexical, and semantic awareness (for specific languages of interest) to the approximate-search algorithm would be invaluable. Moreover, using translation-alignment methods should make it possible to search for correspondences within texts that are available in multiple languages, such as the Buddhist corpus (in Sanskrit, Tibetan, Chinese, and additional languages). Similar ideas can be used as a lexical search tool that can match to available texts. Approximate string-matching algorithms can also be used to take tentative and fragmentary readings of degraded documents, such as the Dead Sea Scrolls or partial and noisy OCR results of manuscripts, and use them to locate other occurrences within existing corpora. In some cases, these need to be adapted to take the likelihood of alternate readings into account, as well as expected letter widths and spacing considerations for defective manuscripts.

2014DADH

VI. Alignment

Often original manuscripts are already accompanied by manually-prepared transcripts. An important challenge, for which no practical system exists as of yet, is that of aligning transcript letters to their coordinates in manuscript images. We have built a prototype system that directly matches the digital image of a historical text with a synthetic image created from the transcript for the purpose. The method matches the pixels of the two images by employing a dedicated dense flow mechanism coupled with novel local image descriptors designed to spatially integrate local patch similarities. The various stages of the method make it robust with respect to document degradation, to variations between script styles and to non-linear image

transformations. Preliminary work [5] has shown excellent results for many languages and scripts, but the tool needs to be extended to deal with complex cases, like marginal and interlinear additions. It should also be modified to work in an iterative manner in which, after each application, letter appearances would be learned and employed to resolve ambiguities, thereby improving overall performance.

VII. Reconstruction

One of the most challenging issues in the analysis of large collections of historical manuscripts or handwritten fragments is the virtual pasting together of torn parts of a mutilated folio or of different folios originating from the same original manuscript, that for one reason or another are no longer connected with each other. Until just recently, a Genizah researcher, say, holding half a page in his hand and seeking its other half, didn't have the resources with which to achieve his goal beyond his erudition, memory, a few catalogs, and a lot of luck. A tool is needed to assist scholars attempting to reconstruct the correct placement and sequence of extant fragments (based on contour, handwriting, and other features) within the original manuscript and display it together with proposed reconstructions. A prototype system for placing manuscripts fragments has been built for Genizah research [1]; it needs to be made generic and to be ported for tablet touchscreens. Similar problems have been tackled in the context of fresco reconstruction. Special techniques will be needed for matching the fibers of papyri for some items among the Dead Sea Scroll fragments. A tool that will attempt to reconstruct scrolls based on physical defects of fragments, implementing the methodology of the late Hartmut Stegemann, is also being contemplated. Three-dimensional rendering will be needed to help scholars visualize and evaluate reconstructions.

2014 DADH

References

[1] Adiel Ben-Shalom, Yaacov Choueka, Nachum Dershowitz, Roni Shweka, and Lior Wolf, 2014, “Where is My Other Half?”, *Digital Humanities 2014,* Lausanne, Switzerland.

[2] Itai Ben Shalom, Noga Levy, Lior Wolf, Nachum Dershowitz, Adiel Ben Shalom, Roni Shweka, Yaacov Choueka, Tamir Hazan, and Yaniv Bar, June 2014, “Congruency-Based Reranking”, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* Columbus, OH, USA, pp. 2107–2114.

[3] Nachum Dershowitz and Lior Wolf, July 2013, “Automatic Scribal Analysis of Tibetan Writings”, *Abstracts of the 13th Seminar of the International Association of Tibetan Studies,* Ulaanbaatar, Mongolia.

[4] Nachum Dershowitz and Lior Wolf, July 2014, “From Caves to Cyberspace: AI Aids in the Study of the Dead Sea Scrolls”, *Xth Congress of the European Association of Jewish Studies,* Paris, France.

[5] Tal Hassner, Lior Wolf, and Nachum Dershowitz, Aug. 2013, “OCR-Free

Transcript Alignment", *Proceedings of the 12th International Conference on Document Analysis and Recognition (ICDAR 2013),* Washington DC, USA.

[6] Alon Kovalchuk, Lior Wolf, and Nachum Dershowitz, Sept. 2014, "A Simple and Fast Word Spotting Method", *Proceedings of the Fourteenth International Conference on Frontiers in Handwriting Recognition (ICFHR),* Crete, Greece, pp. 3-8.

[7] Roni Shweka, Yaacov Choueka, Lior Wolf, and Nachum Dershowitz, 2011, "וקרב מחשב באמצעות הגניזה קטעי וצירוף יד כתב זיהוי :אחד אל אחד אותם (Identifying Handwriting and Joining Genizah Fragments by Computer)", *Ginzei Kedem,* vol. 7, pp. 171-207. (In Hebrew.)

[8] Roni Shweka, Yaacov Choueka, Lior Wolf, and Nachum Dershowitz, Feb. 2013, "Automatic Extraction of Catalog Data from Digital Images of Historical Manuscripts", *Literary and Linguistic Computing,* vol. 28, no. 2, pp. 315-330.

[9] Lior Wolf, Nachum Dershowitz, Liza Potikha, Tanya German, Roni Shweka, and Yaacov Choueka, 2011, "Automatic Paleographic Exploration of Genizah Manuscripts", in *Kodikologie und Paläographie im Digitalen Zeitalter 2 — Codicology and Palaeography in the Digital Age 2,* F. Fischer, C. Fritze, and G. Vogeler, eds., Norderstedt: Books on Demand, Germany, pp. 157-179.

[10] Lior Wolf, Rotem Littman, Naama Mayer, Tanya German, Nachum Dershowitz, Roni Shweka, and Yaacov Choueka, Aug. 2011, "Identifying Join Candidates in the Cairo Genizah", *International Journal of Computer Vision,* vol. 94, no. 1, pp. 118-135.

2014 DADH

數位手稿學

Nachum Dershowitz*

摘要

書面文本是了解過去生活的理想來源。手稿，包括社區文件、個人信件和商業記錄都有助於對特定地點和時間的更完善了解。我們現在生活的數位時代，讓歷史紀錄以及不同的文物大規模的數位化變得可能。全球許多大型的文件材料收藏最近正在數位化，並可在網路上取得。其中一個例子就是 Friedberg Genizah 計畫，將 1890 年代在老開羅猶太教會堂頂樓 Genizah 儲藏處找到廢棄的中世紀手抄本、羊皮卷、書信和文件，目前散布在世界各地超過 75 間圖書館與典藏機構，總共有 350,000 個的片段，進行數位化（全彩色、正反面、600dpi 的）。其他例子如臺灣歷史數位圖書館的 2,000,000 個影像和 70,000 份的 19 世紀民間契約及朝廷文件；現在正由以色列古物管理局進行光譜成像，總數超過 15,000 份的死海古卷片段；無數的阿拉伯文、梵文、和中國手稿；藏傳佛教手稿和木板畫；希臘和埃及古語的莎草紙；以及其他更多收藏。所以，現代的歷史學者（或其他人文學科學者）得天獨厚，除了擁有成千上萬、容易取得、隨時可用的影像，潛藏相關的完整或片段文件，還有大量的已經轉錄文字的文本。不幸的是，直到最近以前，分析這些材料的數位工具都極度薄弱。沒有適當的電腦輔助設備，學者面對大海撈針的挑戰，必須在猶如草堆的大量網路圖像和文本之中，將尋覓良久的那些針頭分離出來。這種情況正瀕臨顯著的改變，雖然最先進技術的光學文字識別（Optical Character-Recognition，OCR），仍無法為這些歷史手寫材料提供優質可搜索的文本，但現在已有幾種功能強大的其他電腦化工具原型，可以用來處理這類文件，以幫助學者研究。接下來的段落，將簡要說明幾種我們和 Lior Wolf 團隊正在進行的這類工具發展。

一、測量

對於圖像處理，我們採取一個相對簡單的任務，推算文件圖像以外的其他物理參數，包括頁面大小、書寫範圍、顏色、素材、行距、以及其他類似的量度資料，這些資料通常都會涵蓋在目錄和學術出版物裡。我們已經在 Genizah 數位化計畫有成功完成的經驗[8]。如果在數位化時使用對比的背景顏色（例如 RGB0, 120, 225），且在圖片中附上一個（對比色）標尺和色標，會使這項的工作變得更加容易。

二、古文字學

現代機器學習的方法可以用於古文字學，根據書寫風格或起源，對手稿進行分類。利用尺度不變轉換器 SIFT 的描述功能，由訓練實例中建造一個分類器。

* 以色列臺拉維夫大學資訊科學學院教授，Email：nachumd@tau.ac.il。

或者相同的功能可以用於從適宜候選集合中，辨識「相鄰」文本，並幫助加速那樣的古文手稿分類。另外一個有用的工具，可以將字型聚合成叢集，分析其變異性。我們目前將這類數位化古文字學方法，應用於開羅 Genizah [9]，死海古卷，[4]和西藏[3]語料庫等。

三、查找

查找的任務，從一個查詢圖像開始，不論是從某個影像中切出的部分，或是從語料庫人工創造的書寫形式。要能從一個手稿圖像的語料庫，進行詞或字型的視覺檢索。這個主題是最近許多研究的熱門標題，包括我們自己的獲獎原型系統[6]，這個原型系統可以極快速的查詢下，仍然保持高精準度。要被查詢的數據組影像同時透過一個簡單的二元化運算來事先處理，接著由多個重疊的侯選目標抽取出來。每個二元目標，以及查詢，都被調整大小以適應一個固定大小的矩形，並被傳統的圖像描述再呈現。接著是餘弦相似性的操作—接著最大化匯集了隨機組合—被用於以一個精簡的 250 維向量呈現各個目標或查詢。檢索被快速的呈現一小部份，透過在那個空間最接近的鄰近搜尋，其次是一個簡單的對其他重疊侯選者的抑制。

四、連結點

我們已經發展出一個圖像分析的架構，使電腦能夠比較兩個影像，並根據通用的手寫形式、以及推算的物理屬性，計算一個相似性的得分 [10]。這個統一且有效的系統，其幫助學者找出侯選的潛在連結點（也就是，出於同一人的手寫，來自源於同份手寫稿的片段）。給定一個片段，這個工具（現在能被在 Genizah 網站上的學者使用）以其相似性[1]檢索並排名其他片段。這樣一來，我們能夠辨認出上千個新的連結，讓學者感興趣[7]，儘管事實上這些文件已經在 20 世紀被廣泛的研究。我們同時還研究了對一個既定查詢互動再排名的結果。再排名使在各種結果間使用類似性，而無需使用任何其他的資訊來源。貝葉斯圖形被使用於加強被搜索物件強烈的連結到其他檢索，並重複聚集從連結中獲得穩定性的方法，而一個活性的、以相關性為基礎的再排名過程利用真實的、對查詢[2]有低相關的配對。相同的工具也適用於其他藏品。

五、互文性

各種工具都可以用於已被數位化的文本（通常是由人工轉錄，或結合一個平庸 OCR 與人工校正）。舉例來說，有比對算法（如 CollateX）用以匹配數個相同文本的修訂本，其能幫助摘要式關鍵版本的準備。對於還無工具可使用者，相關問題則是找到不同文本間的平行段落（任何想要的長度）。近似的文本配對（使用動態編程以及／或索引）可以被用於定位兩篇文章或一個已知的大語料庫的所有相似段落。該結果也可用於揭示正投影詞法變化的特殊性和系統性。添加型態、詞彙和語義意識（對於有興趣的特定語言）來近似搜索算法是非常寶貴的。另外，使用翻譯比對的方法應該使搜尋可用於多種語言的文本之對應變得有可能，例如佛教語料庫（梵文、藏文、中文和其他語言）。相似的想法，可以作為詞彙搜索的工具，其能與可取得文本配對。近似字符串匹配算法也可以用於對退化的文件採取試探性的和零散的閱讀，例如死海古卷或部分，以及雜亂的 OCR

手稿結果，以及用它們定位其他在現有語料庫中發生的事件。在某些情況下，這些需要進行調整，以將另類閱讀的可能考慮在內，或是考量有缺陷的手稿，預計調整字母的寬度和間距。

六、對列

原始手稿經常伴隨人工已經準備的轉錄文本。雖然尚無實際完成的系統存在，重要的挑戰就是將轉錄文本的字元與手稿圖像的座標對列。我們建立一個原型系統，可以直接將歷史文件之數位化影像與創造自轉錄文本之合成圖像配對。這個方法將兩個影像的像素配對，採用一個專屬的密集流動機制，與空間整合地方相似性的新設計在地影像描述。透過這個方法不同階段的測試，對文件的原始毀壞狀況、文本形式之間的變異性、以及非線性影像的轉換，都逐漸達成系統穩定。前期工作[5]已經顯示對許多語言和文本優異的結果，但因需面對複雜的工作，需要擴展邊際和並排等補充工具。它同時需要修改以利用迭代方式被運用，在各個應用程序後，字元字體會被學習而減少歧異，從而提高整體表現。

七、重建

在大量歷史文本或手寫片段收藏之分析中，最具挑戰性的工作，就是將源自相同原始手稿不同作品之間，或是同一作品被分開的部分，虛擬地黏貼起來，尤其當這些分開的片段因某些緣故早已沒有互相連結。最近，有一個 Genizah 研究員說過，手裡拿著半頁作品要找到失散的另一半，除了需要博學強記、還需要一小部分目錄和很大部分運氣的幫忙，才足以達成目標。學者想要重建現存片段的排序與正確的位置（基於輪廓外觀、手寫和其他特徵），並將原始手稿與擬定的重建架構呈現在一起，需要開發一些工具來協助。Genizah 研究已經打造一個放置手稿片段的原型系統[1]；未來它會被製作成通用的系統，也可以移植到平板電腦的觸控螢幕，才能解決戶外重建時的類似問題。為了匹配死海古卷片段某些物件之莎草紙的纖維，同樣需要特殊的技術。我們正在考量引用 Hartmut Stegemann 最新方法論來開發一個工具，可以利用某些片段的物理缺陷來重建捲軸。當然也需要援用三維渲染的技術，來幫助學者視覺化並評估重建的成效。

2024 DAH