部分文書クラスタリングによる未解読文書の解読可能性の判定

安形輝（亜細亜大学） agata@asia-u.ac.jp　安形麻理（慶應義塾大学） mari@slis.keio.ac.jp

1. 未解読文書の解読可能性の判定

線文字Bの解読[1]に代表されるように、未解読文字や文書の研究は、内容の解読に焦点を当てたものが多い。しかし、長年の解読の試みにも関わらず未解読のままの文書は、その内容が無作為に（デタラメに）作成されたために、そもそも解読できないという可能性もある。

著者らは、既知の多くの言語に有効なテキスト処理手法は未知の言語に対しても応用できると仮定し、文書内容の解読ではなく、本文が何らかの構造を持つかどうかを検討することにより、該当文書の解読可能性そのものを判定する手法を提案する。そして、実際にこの手法を未解読文書に応用した結果を報告する。

具体的な判定手順は次の通りである。①未解読文書のテキストデータを未解読のまま対象データとし、テキスト分類や情報検索の手法を応用することにより、部分文書同士の類似度測定を行う、②そこから得られた文書構造を、図やページ順などの他の手がかりと比較する、③テキスト分類手法などから得られた構造と他の手がかりによる構造の一致の度合いから、該当文書がデタラメな捏造文書であるのかどうかを判断する。

2. 実験対象：ヴォイニッチ写本

2.1 ヴォイニッチ写本の概要

本研究ではWilfred M. Voynichが1912年にイタリアのイエズス会修道院モンドラゴーネで発見した、通称「ヴォイニッチ写本」[2]を判定対象とした。この写本は未知の言語または暗号で書かれており、多数の解読の試みにも関わらず未解読のままで、真贋論争も絶えない[3]。

写本の制作地や年代は不明であるが、15世紀末から16世紀の中欧とされることが多い（装丁は18～19世紀）。現在は、102葉の羊皮紙から成る。アラビア数字で付与されたフォリオ番号は116まであり、番号からは14葉が欠落していると推測される[4]。前半の7丁は8葉ずつから構成されるが、後半の折丁は不規則で、一部の葉は2～6回折り畳まれている。

本文はヴォイニッチ文字と呼ばれるアラビア数字やアルファベットに似た未知の文字で書かれている。この写本と類似の言語や文字の資料は見つかっておらず、唯一の資料と言える。大部分のページには緑、茶、黄、青、赤のインクを用いた挿図があり、ページ大のものも多い。挿図の内容は、植物、天文学、小さな裸の女性、十二宮図、薬草の調合用壺などである。植物などの同定は成功していない。

ヴォイニッチ写本の真正性の根拠としてたびたび指摘されてきたのは、200ページを超える大部な写本であり、作成の労力を考慮すると捏造の可能性は極めて低いということである。ただし、強い動機を持つ捏造者を想定すると、この説が成立するとは考えにくい。また、外的な証拠としては、当該写本を思わせる特徴を持つ写本に言及した17世紀のAtanasius Kircher宛の複数の書簡が現存している[2]。

2.2 解読の初期の試み

解読の最初の試みはペンシルバニア大学のWilliam Romaine Newboldによって1928年に発表された[5]。彼は暗号の一部の解読に成功し、この写本は13世紀英国のRoger Baconの著作であると主張し、大きな反響を巻き起こした。しかし、その後のJ. M. Manlyによる研究で、Newboldの解読手法は著しく主観的かつ不完全であることが指摘された[6]。

その後も、趣味的なものから暗号専門家による研究まで、様々な解読の試みがあり、一部を解読したとする主張も散発的に見られる[7]。中には第二次大戦中の日本軍のパープル暗号を解読したWilliam F. Freedmanもいるが、彼はこの写本は「ア・プリオリなタイプの人工的もしくは普遍的言語を作成しようとする初期の試みである」という見解をアナグラムで残している[8]。しかし、多くの試みにも関わらず、首尾一貫した解読内容を提示できた研究はない。

2.3 統計・テキスト処理手法による既往研究

真正な写本、つまり、暗号または未知の言語で書かれた写本とする科学的な根拠として、写本の本文が言語学的な特徴に従うという、統計的手法やテキスト処理手法を応用した研究成果が挙げられる。最近では、1998年にG. LandiniとR. Zandbergenが、テキストデータが

ジップ（Zipf）の法則に従うことを指摘した[9]。

また、冗長暗号で書かれているのならば情報エントロピーは低くなると予想される。しかし、Zandbergen の研究によれば[10]ヴォイニッチ写本には少なくともラテン語と同程度の多様性があるため、冗長暗号ではないと考えられる。

本研究のように、文書クラスタリング技術を応用した研究は少ない。学術雑誌に掲載された論文はないが、ウェブ上では内部報告や個人サイトで5件の実験結果が公開されている。1件目はM.E. D'Imperioによる研究[11]があるが、これは複数の人間が書写したとするCurrier の仮説[12]を検証する目的で行われた。他の 3 件も Currier の仮説検証が目的であり、仮説を支持する結果が得られている。

Jorge Stolfi のみは、Currier の検証ではなく、解読の手がかりを得るために、クラスタリング技術を用いた実験を行った。これは、ページをセクション同士のペアに分け、シュミットの直交化を用いて図示したものである[13]。しかし、トランスクリプション（翻字）の正確さ等が検証されていないにも関わらず、頻出単語上位 50 語しか用いていない点や、独自のセクション分けを行っている点に問題がある。

2.4 ヴォイニッチ写本を捏造とする研究

捏造説の根拠としては、W. Voynich の発見以前の来歴が不確かである、挿図の植物等が同定できない、暗号専門家による解読の試みが悉く失敗している、出現頻度の高い文字列の連続が多い、などの点が指摘されてきた。初期の捏造説には Michael Barlow によるものがある[14]。彼は、D'Imperio によるヴォイニッチ写本についての網羅的な研究報告書[15]に基づき、どの分野の専門家もこの写本から意味のある内容を見出せないことから、W. Voynich 自身による捏造だと解釈するのが自然であると結論付けている。

また、最近の研究に、2004 年のキール大学の Gordon Rugg による本文の復元実験[16]がある。「カルダーノ・グリル（Cardano grille）」という簡単な道具を使うとヴォイニッチ写本と似た特徴を持つ文書を比較的短期間に作成できることから、この写本が精巧な捏造であり、作成者は 16 世紀の錬金術師 Edward Kelly（Dee の召使）であろうと結論付けたものである。Rugg の研究は、*Nature Science Update*[17]他、一般誌を含む数多くの雑誌に取り上げられ[18]、日本語にも訳されるなど[19]、近年のヴォイニッチ研究の中では最も注目を集めている。

2.5 既往研究のまとめ

ヴォイニッチ写本の既往研究をまとめると、解読の手がかりを得ようとする研究は一定の成果を挙げているものの、数多くの解読の試みに成功したものはなく、最近では捏造説を支持する有力な研究成果が発表されるなど、真贋論争も再燃している。また、文書クラスタリング技術を応用した研究はあるが、写本が真正であるという前提で Currier の仮説検証を目的に行われている。つまり、解読可能性の判定そのものに関する研究は行われていない。

本研究では、文書クラスタリング技術を応用し、ヴォイニッチ写本が構造を持たないデタラメな文書であるのか、つまり、そもそも解読が可能な文書であるかどうかの判定を行った。

3. 実験の手順

3.1 部分文書の単位

ヴォイニッチ写本の解読可能性を検証する上で、その部分文書単位をどう設定するかという問題がある。後述のトランスクリプションは段落単位で行われているため、段落、ページ、複数ページから成る葉、挿図から推測したセクション、などの単位を考えることが可能である。本研究では、恣意的な判断が最小限となるように、物理的なページを単位として用いた。

3.2 元にする図表との対応

従来の研究と同様に、挿図から推測されるセクションごとに分析する。セクションの分け方は研究によって様々であり、非常に細かく分けている例も見られる。ここでは所蔵館であるイエール大学バイネッケ図書館の目録の記述に基づき、表1の通り、6セクションに分けた。

なお、「十二宮図」セクションの半数のページにはほとんど文字がないため、本研究のようにテキスト分類手法を応用する場合には、正しく扱うことが困難なセクションだといえる。

表1 本研究のセクション分け

フォリオ番号	セクション	ページ数
f. 1r～65v	植物（Plant）	116
f. 66r～73v	天文(Astoro.)	26
f. 75r～84v	生物(Bio.)	20
f. 85r1～85v6	十二宮図（Zodiac）	8
f. 87r～102v2	薬草(Herb.)	32
f. 103r～116r	レシピ（Recipe）	23

3.3 トランスクリプション

未解読のヴォイニッチ写本のテキストデータを分析するには、トランスクリプションが必要となる。ヴォイニッチ写本に関しては複数のトランスクリプションが存在する。ここでは、写本の画像と照合したときに比較的忠実で、かつ、全ページ分のデータの存在が確認できた高橋健によるトランスクリプションを用いた[20]。

3.4 単語の切り分け

ヴォイニッチ写本では一定の文字列の間に空白が挿入されており、空白で単語が区切られていると推測されるため、空白と改行により、単語の切り分けを行った。トランスクリプションでは、判別不能文字は“＊”で、図や直前の文字の形態等に関する注釈は“{}”内に示されている（例えば、“{plant}”は植物の図があることを示す）。ここでは、“＊”はそのままとしたが、注釈には意味が含まれると判断し、除去した。

以上のルールを適用し、トランスクリプションを単語ごとに切り分けた場合のヴォイニッチ写本のテキストの基本的統計を表2に示した。

表2　単語等の基本的な統計

異なり単語数	7907 語
平均単語数／ページ	166.0 語
平均単語長	5.0 文字
総ページ数	225 ページ

3.5 単語重みの算出

単語の重み付けは情報検索で一般的な TF-IDF 法を用いた。ただし、文献単位ではなく、ページ単位で単語の重み付けを行った。

$$w_{ij} = tf_{ij} \cdot idf_j = \frac{f_{ij}}{\max_{j=1,\ldots M} f_{ij}} \cdot \log\left(\frac{N}{n_j}\right)$$

なお、ページ内の単語数には大きなばらつきがあるため、単語の出現頻度はページ内単語の最大出現頻度で正規化をした。

3.6 類似度行列の作成

ページ同士の類似度行列作成には、キャンベラ距離を用いた。この距離は非類似度を示すため、値が小さいほどページ同士の類似度が高いことを意味する。

$$d_{ij} = \sum_{k=1}^{n} \frac{\left|x_{ik} - x_{jk}\right|}{\left|x_{ik}\right| + \left|x_{jk}\right|}$$

3.7 クラスター分析手法

代表的な手法として、階層的クラスター分析では「最短距離法」「最長距離法」「群平均法」「ウォード法」を、非階層的クラスター分析では「k 平均法」を用いてそれぞれ実験を行った。

4. 実験結果

4.1 セクション同士の類似度

類似度行列を各ページ間で算出し、セクションごとの平均類似度を算出したものが表3である。この表では、各行ごとに類似度の最も高いものの枠を太線に、値を太字にして示した。「十二宮図」以外は類似度が最も高いのは同じセクション同士の交差する箇所となっている。つまり、あるセクションに属するページは、他セクションのページよりも、同じセクションのページとの類似度の方が高いという結果となった。

4.2 クラスター分析の結果

紙幅の制限のため各手法により作成された全ての図を示すことはできないが、階層的クラスター分析の「ウォード法」で作成されたデンドログラムにおいて「生物」セクションが含まれるクラスターを中心とした一部を示す（図 1）。

図1のように「生物（Bio.）」、「天文(Astro.)」、「レシピ(Recipe)」の一部は、ほぼ同じセクションのページのみから構成されるクラスターにまとまった。一方、「植物(Plant)」と「薬草(Herb.)」のページ群に関しては、同じセクションのページ群でまとまったのは半分程度であり、残りは「植物」と「薬草」が混在するクラスターが構成された。しかし、「植物」と「薬草」は、元々、図自体が明確な区別ができないもの同士であり、ページ内の本文の特徴が似たとしても不思議ではな

表3 セクション同士の平均類似度

	植物	天文	生物	十二宮図	薬草	レシピ
植物	**7612.1**	7715.7	7653.8	7666.0	7659.6	7686.3
天文	7715.7	**7619.9**	7638.9	7645.6	7670.2	7646.7
生物	7653.8	7638.9	**6942.7**	**7302.9**	7575.8	7242.6
十二宮図	7666.0	7645.6	7302.9	7350.7	7596.7	7346.2
薬草	7659.6	7670.2	7575.8	7596.7	**7568.2**	7593.4
レシピ	7686.3	7646.7	7242.6	7346.2	7593.4	**7148.8**

い。また、「十二宮図（Zodiac）」は一部のページ群はまとまっているが、文字がほとんどないページは、全体に分散する結果となった。

これらの傾向は、他の階層的クラスター分析においても同様であることが確認された。

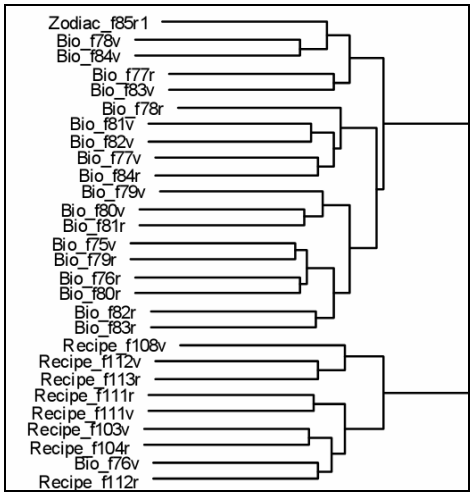

図1 作成されたデンドログラムの一部

4.3 結果の考察

今回の実験結果からはヴォイニッチ写本は挿図によるセクションごとに部分文書同士に一定のまとまりがあることが確認された。既往研究[16]で示唆されたようにデタラメに捏造された場合は、このような挿図と部分文書との対応はないはずである。従って、今回の結果からはヴォイニッチ写本は少なくとも構造を持つ文書であり、デタラメではないという結論が導かれる。

また、この結果は、ヴォイニッチ写本が暗号で書かれているとしても、複式換字式のように文書の進行と共に変換方式が変化していく暗号ではないことを示唆している。

5. 結論

本研究では未解読文書に対する解読可能性の判定手法を提案し、実際に未解読文書の一つであるヴォイニッチ写本に対して提案手法を適用した結果を報告した。このような解読可能性の判定は、どの程度、未解読文書を解読する努力を傾注すべきかを判断するうえで、有効な判断材料になると考えられる。

【注・引用文献】

1) Chadwick, John. 線文字Bの解読. 東京, みすず書房, 1997, 239, 5p.(原著の第二版の翻訳)

2) New Haven, Conn., Yale University Library, MS Beinecke 408. Kircher 宛の書簡 1 通も収蔵(MS 408A)。

3) Kennedy, Gerry; Churchill, Rob. ヴォイニッチ写本の謎. 松田和也訳. 東京, 青土社, 2006, 380, viiip.

4) A.G. Watson は、フォリオ番号は John Dee (1527-1608)の筆跡だとしている。Watson, A.G.; Roberts, R.J. ed. John Dee's Library Catalogue. London, Bibliographical Society, 1990, p. 172-173 (DM93).

5) Newbold, William Romaine; Kent, Roland Grubb. Cipher of Roger Bacon. Whitefish, MT, Kessinger, 2003.（原著 1928 年の再版）

6) Manly, John Matthews. Roger Bacon and the Voynich MS. Speculum. vol. 6, no. 3, 1931, p. 345-391.

7) 例えば Brumbaugh, Robert S. The Voynich 'Roger Bacon' cipher manuscript: Deciphered maps of stars. Journal of the Warburg and Courtauld Institutes. vol. 39, 1976, p. 139-150.がある。また、1976 年にはヴォイニッチ写本解読の会議も開かれた。

8) Reeds, Jim. William F. Freedman's transcription of the Voynich manuscript. Cryptologia, vol.19, 1995, p.1-25.(アナグラムの訳は文献 3 より引用)

9) Landini, Gabriel; Zandbergen, René. A well-kept secret of mediaeval science: the Voynich manuscript. Aesculapius, vol.18, 1998, p.77-82.

10) Zandbergen, René. From digraph entropy to word entropy in the Voynich manuscript(2000). http://www.voynich.nu/wordent.html. [最終確認日:2006-09-30]

11) D'Imperio, M.E. An Application of Cluster Analysis and Multiple Scaling to the Question of "Hands" and "Languages" in the Voynich Manuscript. NSA Technical Journal, vol.23, no.3 (Summer 1978) (http://www.voynich.nu/extra/dimperio92c.html から改訂版[1992]が入手可能)[最終確認日:2006-09-30]

12) Currier, Prescott H. Some Important New Statistical Findings. the Proceedings of a Seminar held on 30th November 1976 in Washington D.C. edit by M. D'Imperio(http://www.voynich.nu/extra/curr_main.html から本文が入手可能) [最終確認日:2006-09-30]

13) Jorge Stolfi. Scatterplots of VMs pages. http://www.dcc.unicamp.br/~stolfi/voynich/98-06-19-page-plots/plots.html [最終確認日:2006-09-30]

14) Barlow, Michael. The Voynich manuscript -by Voynich? Cryptologia, vol.10, 1986, p.210-216.

15) D'Imperio, M.E. The Voynich Manuscript: an elegant enigma. Aegean Park Press, 1981, 148p.(Cryptographic Series , No 27)

16) Rugg, Gordon. An elegant hoax?: A possible solution to the Voynich manuscript. Cryptologia, vol.28, no.1, 2004, p. 31-46.

17) Whitfield, John. World's most mysterious book may be a hoax: The Voynich manuscript may be elegant gibberish. Nature Science Update. 17 December 2003(http://www.nature.com/nsu/031215/031215-5.html より本文が参照可能)

18) Rugg, Gordon. The Mystery of the Voynich Manuscript: New analysis of a famously cryptic medieval document suggests that it contains nothing but gibberish. Scientific American. July 2004, p. 104-109.

19) Rugg, Gordon. ヴォイニッチ手稿の謎. 日経サイエンス, 2004 年 10 月号, 2004, p.86-92.（文献 18 の翻訳）

20) Jorge Stolfi. Reeds/Landini's interlinear file in EVA, version 1.6e6. http://www.dcc.unicamp.br/~stolfi/voynich/98-12-28-interln16e6/[最終確認日:2006-09-30]