Library and Information Science No. 61 2009, p. 1–23.

原著論文

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：ヴォイニッチ写本の事例

Determining the Possibility of Deciphering an Unintelligible Text by Text Clustering: The Case of the Voynich Manuscript

安　形　　輝
Teru AGATA

安　形　麻　理
Mari AGATA

Résumé

Purpose: One of the most common approaches to understanding an undeciphered text is to identify and then decipher the underlying code. If a document remains unintelligible or undeciphered for a long period of time even after many attempts at decoding it, the possibility of it being "gibberish" must be considered. This study proposes a method to detect the existence, or non-existence, of a coherent structure within a previously non-translated text in order to determine the possibility of deciphering it.

Methods: The present method begins with the assumption that natural language-processing methods that are commonly employed in analyzing known languages can be applied to an undeciphered text. To detect a coherent structure in a text, the similarity of every pair of partial document is measured, and then the similarity matrix is analyzed by clustering methods. The next step is to compare the detected structure with the sections suggested by other clues such as illustrations and the page order. Thus, it is determined whether an undeciphered text contains an identifiable structure which corresponds to the latter, or whether it is "gibberish" containing no order or structure.

Results: We applied the proposed method to the Voynich Manuscript, which is a renowned undeciphered text. The results clearly demonstrate that the text of the Voynich Manuscript possesses an identifiable structure, and that the structure corresponds to the existing sections

安形　輝：亜細亜大学，東京都武蔵野市境 5–24–10
Teru AGATA: Asia University
e-mail: agata@asia-u.ac.jp
安形麻理：慶應義塾大学，東京都港区三田 2–15–45
Mari AGATA: Keio University
e-mail: mari@slis.keio.ac.jp
受付日：2008 年 3 月 12 日　改訂稿受付日：2009 年 3 月 9 日　受理日：2009 年 4 月 19 日

of the manuscript suggested by the accompanying illustrations. Thus, the results strongly suggest that the Voynich Manuscript is not “gibberish”; additional attempts to decipher its contents would be justified. The present experiment proves the usefulness of applying this method to a previously non-deciphered text.

I. 未解読文書の解読可能性の判定

A. 未解読文書と「捏造文書」

線文字 B[1] やロゼッタストーンの解読に代表されるように，未解読文字や未解読文書の研究は内容の解読に焦点を当てたものが多い。つまり，未知の言語で書かれている文書の内容を判読しようという研究方法であり，これは当然のことだといえる。未解読文書には，失われた古代の言語で書かれたもののほかに，暗号書も含まれる。暗号の歴史は古く，作成や解読には様々な道具が使われてきた[2]。暗号解読理論の発展やコンピュータなどの新たなツールの登場により，過去の暗号書の解読が進んでいる。例えば，1500 年頃に書かれてから未解読のままだったヨハネス・トリテミウス (Johannes Trithemius) の有名な暗号書 *Steganographia* の第 3 巻を，Jim Reeds がコンピュータを使った分析によって解読したという事

例は 1998 年のことである[3]。こうした解読の試みは，考古学から軍事にいたるまで，様々な分野で社会に恩恵をもたらしてきている。

しかし，長年の解読の試みにもかかわらず未解読のままであるという文書が存在する場合には，読み解くための正しい手がかりが明らかになっていないという理由のほかに，その内容が無作為に（言い換えれば全くデタラメに）作成されたものであるため，そもそも解読できないという可能性を考える必要があるだろう。本論文中では，そのようにデタラメに作成することを「捏造」，デタラメに作成された意味をなさない文書を「捏造文書」と呼ぶこととする（英語の gibberish と同等）。それに対し，意味をなす文書を真正文書とする。なお，ここでいう「捏造文書」とは，文書が意味内容を持たない無意味な文書であるということを指すだけであり，一般的に捏造 (hoax) という言葉で指すような，人を欺く意図をもって作られた文書であるかどうかは問わない。また，偽書，偽作，贋作のように，著者や時代などを偽っているものの，書かれた内容は一応の意味をなしているというものは含まない。以下では，前者を「捏造文書」，後者を捏造文書として区別する。

「捏造文書」であれば解読はそもそも不可能であるし，真正な文書，つまり，文字通りの未解読文書であれば，解読の鍵が見つかっていないだけで解読は理論的には可能であるということになる。

そこで，長いこと未解読であり続けている文書があった場合に，あるいは「捏造文書」である可能性が高いと思われる文書があった場合に，該当文書の解読の可能性を判定することができれば，どの程度の努力を解読に傾注すべきかを判断したり，「捏造文書」に対する実りのない試みを続けることを回避したりできるため，解読の前段階として有用であると考えられる。

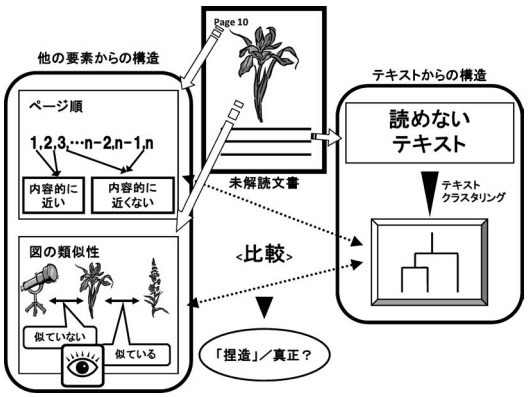
B.　本研究の目的

本研究では，本文が何らかの構造を持つかどうかをテキスト処理手法の応用結果から検討することにより，該当文書が「捏造文書」であるかどうかを判断し，その文書の解読可能性そのものを判定するための手法を提案する。これは，文書内容の解読を目指すものではなく，解読の前段階において用いるための手法である。さらに，実際にこの手法を有名な未解読文書である，通称「ヴォイニッチ写本」に適用した結果を報告する。

C.　手順

本研究で提案する具体的な判定手順を簡潔に述べると，以下の通りとなる（第 1 図）。

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，文書の部分同士の類似度測定を行い，クラスタリング手法により，文書構造の有無を検証する。未知の文字または記号であっても，それを一定の規則に従って既知の文字体系に翻字 (transliteration) すれば，テキストデータとして扱うことが可能である。
2) テキストデータから得られた文書構造を，図表やページ順などの他の手がかりから得られた文書構造と比較する。
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。



第 1 図　未解読文書の解読可能性の判定

この判定手法では，既知の多くの言語に有効なテキスト処理は，未知の言語に対しても応用できるという前提に基づいている。

文書の部分とは，処理・分析の単位とする文書の一部分のことであり，例えば章，ページ，段落，文，などが考えられる。以下では，これを部分文書と呼ぶ。ただし，未解読文書においては，単語や文，節，章といった，一般的に文書が持っている構造があるかどうかさえも明らかではない場合が想定される。その場合は，恣意的な解釈が入る可能性が最も低くなると想定される，物理的なページを単位とすることがよいと考えられる。

提案手法は，真正な文章は何らかの構造を持っているということ，また，真正な文書においては，物理的に近い位置にある部分文書同士（例えば，近くのページ同士）の文章内容ほど，物理的に遠くにある部分文書同士（例えば，離れたページ同士）よりも類似しているということに基づいている。また，その文書が図や写真を含んでいるのであれば，似たような図が共通して見られる部分文書同士は，全く異なる種類の図が見られる部分文書同士よりも近い意味内容を持つということが経験的に予想されるため，同様に手がかりとなる可能性が高い。つまり，花の図を含む部分文書同士は，花の図と望遠鏡の図をそれぞれ含む部分文書同士よりも，意味的な類似度が高いと予想される。一方，無作為に作成された「捏造文書」であれば，部分文書同士の類似度にそうした一定の傾向は見られないはずである。

この前提が成り立つならば，未解読文書があったときに，部分文書のテキストデータの構造と他の手がかりによる構造とが一致していれば，全体として何らかの一貫性のある構造を持つ文書であり，「捏造文書」ではない可能性が高いことになる。逆に，構造が一致しない場合には「捏造文書」であり，そもそも意味をなしていないために解読ができない可能性が高いと判断できる。

ただし，この前提は既往研究で検証された事実ではない。そこで，本研究では対象となる未解読文書をテキスト処理した結果を，無作為に作成された「捏造文書」，および，主題的にまとまった構造を持つ真正文書と比較することで検証を行った。

以下では，提案手法を用いた実験結果を報告するとともに，技術的な詳細について述べる。

II. 実験対象：ヴォイニッチ写本

A. ヴォイニッチ写本の概要

1. 物理的な特徴

本研究では，Wilfred M. Voynich が 1912 年にイタリアのイエズス会の僧院ヴィラ・モンドラゴーネで発見したとされる，通称「ヴォイニッチ写本」(New Haven, Conn., Beinecke Rare Book and Manuscript Library, Yale University Library, MS 408) 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かかわらず未解読のままで，真贋論争も絶えない[4]。作者についても様々に推測されており，発見者 Voynich による捏造ではないかという意見も根強い。近年では，インターネット上に活発なメーリングリストも存在している[5]。

この写本を実験対象として選んだのは，有名な未解読文書であるというだけでなく，後述するように画像データ，画像データから翻字したテキストデータともに入手できるためである。

標題紙やコロフォン等がないためヴォイニッチ写本の制作地や年代は不明であるが，同写本を所蔵するイエール大学バイネッケ図書館の目録では，15 世紀末から 16 世紀の中欧において作られたのではないかと推測されている[6]。ただし，装丁は 18〜19 世紀の新しいものである。全ページのカラー画像は所蔵館のウェブサイト上で公開されている[7]。

物理的な大きさは 225×160 mm で，102 葉の羊皮紙から構成されている。前半の 7 丁は 4 枚の羊皮紙を二つ折りにした 8 葉ずつから構成されているが，後半の折丁は非常に不規則で，一部の葉は，ちょうど現在の折り込み地図などのように，2〜6 枚の羊皮紙が折り畳まれてできている（2 枚折り畳まれているのが 5 葉，3 枚が 3 葉，4 枚が 1 葉，6 枚が 1 葉ある）。ページの順序が写本成立時と同じであるかどうかを判断できる手がか

りは少ない。中世の写本には，ページ末尾や丁末尾に次ページの一語または数語が示されていることが多いが（捕語），この写本には捕語はなく，ページの順序を確認する手がかりとはならない。アラビア数字で付与されたフォリオ番号は，本文とは違う手になると考えられるものではあるが，116 まであるため，現状では少なくとも 14 葉が欠落していると推測されている（以下の 14 葉：12, 59–64, 74, 91–92, 97–98, 109–110）。ただし，フォリオ番号が振られていない箇所もある。また，各丁の最終葉裏の下部には通常のアラビア数字で折丁記号が書かれている。本文と同じ手になるかどうかについての議論は特になく，いつの時点で付与されたものであるかは不明であるが，現在の折丁構成と一致している[8]。

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によって，一段組みで書かれている。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。特定の言語との同定や関連づけに成功した既往研究はない。文字数についても諸説あるが，句読点がなく，非常に繰り返しが多いという見解は共通している。第 2〜7 図からわかるように，ページの右や下に余白が多く見られるので，通常のヨーロッパ言語のように単語から構成され，左から右に書かれていると考えられる。ほとんどのページには緑，茶，黄，青，赤のインクを使って彩色された素朴で奇妙な挿図があり，ページ大の場合も多い。こうした挿図は本文と一体となってレイアウトされており，後世の挿入である可能性はない。挿図の内容は，植物，天文学，水浴びをしている小さな裸の女性たち（ニンフ），十二宮図，薬草の調合用壺などのように見える。ただし，数多くの植物の挿図にもかかわらず，いずれも実際の植物などとの同定は成功していない。なお，ヒマワリや唐辛子などアメリカ大陸原産の植物が描かれているとして写本制作年代を 1493 年以降だと推測する説もあるが[9]，一般的な合意は必ずしも得られていない。

2. 来歴と真正性の根拠

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「捏造文書」である可能性は極めて低いということである。暗号解読の専門家達が何の規則性も見いだすことができないような無意味な文字列をこれほど大量に書くことは，人間にとってはきわめて難しい。また，高価な素材である羊皮紙に書かれており，15, 16 世紀のものだとすると誰もが容易に作成できるものではない。ただし，それにより莫大な経済的な利益を得ることができるなど強い動機を持つ「捏造」者を想定すると，この議論の説得力は高いとはいえない。

来歴は真正性の外的な根拠となりうるが，Voynich によって発見される以前の来歴を示す直接的な証拠は見つかっていない。Voynich によれば，第 1 葉表 (f.1r) の余白に紫外線を当てると “Jacobi de Tepenecz” という署名が読み取れたという[4), 10)]。Voynich は，これは Jacobus Horcicky de Tepenecz（1622 年没）のことであり，彼がその称号を許された 1608 年から没年である 1622 年までのどこかでこの写本を所有していた証拠であるとした。

この写本が存在していたという外的な証拠としては，17 世紀のイエズス会士で古代言語や暗号の専門家としても知られていた碩学 Atanasius Kircher に宛てられ，当該写本を思わせる特徴を持つ写本に言及した複数の書簡が現存していることが挙げられる。バイネッケ図書館では写本とともに発見された書簡 1 通も収蔵している (MS 408 A)。これは，プラハ大学の Marcus Marci（1667 年没）が 1665 年に Kircher にある写本の解読を依頼したもので，そこには，当該写本はボヘミア王兼神聖ローマ皇帝ルドルフ 2 世（在位 1576–1612 年）が 600 ダカットで購入したもので，皇帝は Roger Bacon の著作だと考えていた旨が述べられている[11)]。ただし，その後の約 250 年間，この写本についての言及と思われるものは見つかっていない。

また，写本学者 A. G. Watson と R. J. Roberts

は，エリザベス女王に寵愛された数学者・哲学者・魔術師 John Dee (1527–1609)[12] の蔵書を再構築するなかで，ヴォイニッチ写本は Dee 自筆の蔵書目録には掲載されていないものの，彼の蔵書であった書物のうちの一冊であると結論づけ，フォリオ番号は Dee によって付与されたものだと考えている[13]。なお，Voynich はルドルフ 2 世に写本を売ったのは，1584 年に謁見した Dee だと主張した。しかし，いずれも，この写本が Voynich が発見するよりも前から存在していたということの決定的な証拠とはなりえていない。

B. 解読の初期の試み

ヴォイニッチ写本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[14]。彼は暗号の一部の解読に成功し，この写本は 13 世紀英国のフランチェスコ会士であり自然科学の先駆者とも称される Roger Bacon の暗号による著作であると主張し，大きな反響を巻き起こした。しかし，Newbold の没後，J. M. Manly の研究により，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，解読は白紙に戻った[15]。

その後も，アマチュア研究者の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[16]。1976 年にはこの写本についての会議も開催された。1978 年までの研究状況については，アメリカの国家安全保障局から出版された M. E. D'Imperio による網羅的な報告書に簡潔にまとめられている[17]。しかし，これまでの多くの試みにもかかわらず，首尾一貫した解読内容を提示できた研究はない。

例えば，解読に取り組んだ者のなかには，第二次世界大戦中に日本陸軍が用いた，いわゆるパープル暗号（九七式印字機）の解読で知られる William F. Freedman もいたが，彼も解読には至らず，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遺している[18]。イギリス陸軍所属の暗号学者 John H. Tiltman も，語頭や語尾に置かれる傾向がある文字の存在を指摘し，文字を 3 種類に分類したうえで，1951 年に Freedman と同様の結論に達している[17] [p. 42–44]。

なお，1970 年代に行われた少し方向性の異なる研究としては，軍の暗号専門家であった Prescott Currier のものが挙げられる。Currier は，接頭辞の分析などから，植物の挿図があるセクションは 2 種類の筆跡かつ 2 種類の言語で書かれており，写本全体では 12 の異なる筆跡が確認されるとした[17] [p. 43][19]。

C. 統計・テキスト処理手法による既往研究

ヴォイニッチ写本が真正な写本，つまり，暗号または未知の言語で書かれた意味をなす写本だ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。

1998 年には，Gabriel Landini と René Zandbergen が，ヴォイニッチ写本の単語の出現頻度が Zipf の法則に従うことを指摘した[20]。Zipf の法則はもともと George Kingsley Zipf が英語における単語の出現頻度について発見したべき乗法則の一種であるが，経験的に英語以外の他の自然言語に関しても成立することが知られている。つまり，彼らの結果は，ヴォイニッチ写本のテキストが自然言語と同様の特徴を持つことを示している。

さらに，Zandbergen は，ヴォイニッチ写本の本文の情報エントロピーを調査した。この写本には同じ文字が繰り返し出現することが多く，一見すると冗長性が高そうである。しかし，Zandbergen の結果は，ヴォイニッチ写本の本文には少なくともラテン語や英語と同程度の多様性があることを示した。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ため，この写本の本文は冗長暗号で書かれてはいないと考えられる[21]。

一方，ヴォイニッチ写本に対して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研究成果はないものの，Zandbergen によるウェブサイトでは内部報告や個人サイトに

おける実験5件が紹介されている[22]。例えば，D'Imperioによる研究[23]は，写本が真正であるという前提で，複数の写字生がいるというCurrierの仮説[19]を検証している。他の3件も同様にCurrierの仮説を検証することを目的として行われた実験であり，仮説を支持する結果が得られたとしている。しかし，いずれの実験もその手順の詳細が明らかでないため，再現することや妥当性の検証を行うことが困難な分析にとどまっている。

5件のうち，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ている。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[24]。Stolfiは5つのクラスターに分かれたと結論づけている。しかし，トランスクリプションの正確さやヴォイニッチ文字の種類等が検証されていないにもかかわらず，分析にZandbergenによる頻出上位50語しか用いていない点や，似ている語をノイズとしてまとめてしまっている点，独自の細かなセクション分けを行ったため他の研究との比較ができない点などに問題がある。

D. ヴォイニッチ写本を「捏造」とする既往研究

「捏造」説の根拠としては，Voynichの発見以前の来歴が不確かである，数多くの挿図の植物等が一つも同定できない，暗号専門家による解読の試みがことごとく失敗している，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」説としてはMichael Barlowによるものがある[25]。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[17]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いだせないことから，Voynich自身による「捏造」だと解釈するのが自然であると結論づけている。

また，最近の研究に，2004年に発表されたキール大学のGordon Ruggによる本文の復元実験[26), 27)]がある。Ruggは，1550年にGirolamo Cardanoによって考案された「カルダーノ・グリル(Cardano grille)」[28]という簡単な道具を使うと，ヴォイニッチ写本とよく似た特徴を持つ文書を比較的短期間（3, 4カ月）で作成できると指摘した。彼は，この写本が意味内容を持たない精巧な「捏造」であると考え，作成者は16世紀の錬金術師Edward Kelly (1555–1597/8)であろうと結論付けている。Kellyは1582年から1589年までDeeに雇われており，彼とともにルドルフ2世の宮廷に滞在していた[29]。また，現在British Libraryに所蔵されている，『エノクの書（*Book of Enoch*)』(London, British Library, Sloane MS. 3189)と呼ばれる未解読文書を書いた人物でもあり，それまでにもヴォイニッチ写本の捏造者の候補として名前が挙がってきた人物である。Ruggの研究は*Nature Science Update*[30]ほか，一般誌を含む数多くの雑誌に取り上げられ[31]，日本語にも訳されるなど[32]，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。

さらに，2007年にはRuggの論文と同じ暗号学専門誌にAndreas Schinnerの論文が発表された[33]。これはランダムウォークによるマッピングとトークンの反復の統計に基づく研究である。ヴォイニッチ写本の本文は暗号化ではなく確率過程によって生成された可能性が高いとし，捏造説およびRuggの仮説を補強する結果であると結論づけている。

E. 既往研究のまとめ

ヴォイニッチ写本の既往研究は，次のようにまとめることができる。この写本の性質についての仮説は，(1) 暗号書，(2) 未知の言語あるいは人工言語による写本，(3)「捏造文書」，の三つに大別できるが，議論に終止符を打つだけの証拠はない。解読の手がかりを得ようとする研究は一定の成果を上げているものの，長年にわたる数多くの試みにもかかわらず，解読に成功した研究はない。LandiniとZandbergenによる単語の出現頻度に関する1998年の研究はヴォイニッチ写本の真正性についての科学的な根拠を示したが，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。つまり，

これまでの研究を概観すると，ヴォイニッチ写本が真正な文書であるかどうかを検討する研究，言い換えれば解読可能性の判定そのものの検討は，あまり行われてこなかったことがわかる。

そこで，本研究ではヴォイニッチ写本の文書構造の有無に着目し，真正性の判定を行うこととした。具体的には，文書クラスタリング技術を応用し，部分文書同士の関係を分析することにより，ヴォイニッチ写本が構造を持たない「捏造文書」なのかどうか，つまり，そもそも解読が不可能な文書であるのかどうかの判定を行った。

III. 実 験 環 境

A. 実験の概要

提案した判定手法では，部分文書中のテキストから得られた文書構造と，他の手がかりとを比較することにより，「捏造文書」であるかどうかの判定を行う。今回研究対象としたヴォイニッチ写本の場合，文書構造を得るための手がかりとして，ほとんどのページに描かれた挿図やフォリオ番号を用いることができる。そこで，テキストから得られた文書構造と挿図から推測される構造，および，ページ順から推測される構造との比較を行った。

テキストからの文書構造を抽出するために文書クラスタリングを行った。具体的な手順としては，以下のようになる。

1) 部分文書内のトークンすべての重みを算出する（TF-IDF 法）
2) トークンの重みから類似度関数（距離関数）によって部分文書同士の類似度を算出し，類似度行列を作成する
3) クラスター分析手法（Ward 法）を用いて類似度行列から部分文書のクラスタリングを行う

本章の前半 B〜D 節では文書構造を抽出する前提となる実験の元データの処理単位，トークン，翻字について，後半 E〜H 節ではセクション同士の平均距離やクラスタリング手法を用いてテキストからの構造の抽出を行う手法について，最後の I 節ではクラスタリング結果の評価手法について記述した。

B. 部分文書の単位

提案手法によってヴォイニッチ写本の解読可能性を検証するうえでは，その部分文書単位をどう設定するかという課題がある。この写本には章節などの意味的な区切りを示す空白ページ，冒頭の頭文字，見出し語などの一般的な文章構造の視覚的な提示は行われていないため，そのような分割単位を識別することはできない。それ以外の部分文書単位の候補としては，段落，ページ，葉，挿図から推測したセクションなど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，物理的なページを部分文書の単位として用いた。折り畳まれて複数ページからなる葉には複数ページにまたがる挿図が存在する場合もあるが，それらも物理的なページごとに分割して分析単位としている。

C. テキスト以外から推測される構造

実験対象写本のページ順が写本作成時と同じではないという可能性は考えられるが，フォリオ番号が遅くとも 16 世紀には付与されており，いつの時点のものかは不明であるものの現在の折丁構造と一致する折丁記号もあるため，ここではページ順という文書構造を一つの手がかりとした。

また，実験対象写本には多数の挿図が含まれているため，挿図を二つ目の手がかりとして用い，挿図から推測されるセクションごとに分析を行うこととした。セクションの分け方は研究者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が，内容が未解明である以上，細かく分ける根拠は薄弱である。ここでは，所蔵館であるバイネッケ図書館の目録の記述に基づき，第 1 表に示したように，「植物 (Plant)」，「天文（または占星術）(Astro.)」，「生物 (Bio.)」，「十二宮図 (Zodiac)」，「薬草 (Herb.)」，「レシピ (Recipe)」の 6 セクションに分けた。描かれているものに基づいておおまかにカテゴリ分けしたもので，ヴォイニッチ写本研究においては最も一般的な分け方である。各セ

第 1 表　セクション分け

フォリオ番号		セクション	ページ数
開始	終了		
f. 1r	f. 65v	植物 (Plant)	116
f. 66r	f. 73v	天文 (Astro.)	26
f. 75r	f. 84v	生物 (Bio.)	20
f. 85r1	f. 86v6	十二宮図 (Zodiac)	8
f. 87r	f. 102v2	薬草 (Herb.)	32
f. 103r	f. 116r	レシピ (Recipe)	23

第 2 図　ヴォイニッチ写本：植物セクション (f. 16v) (Beinecke Rare Book and Manuscript Library, Yale University Library, MS 408) 画像出典：http://beinecke.library.yale.edu/dl_crosscollex

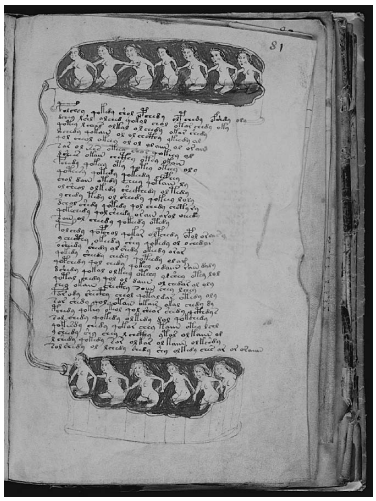
クションは以下のような特徴を持つ。

「植物」セクションはヴォイニッチ写本中で最も量が多く，半数以上のページ数を占める（第 2 図)。各ページには一つの植物の図が大きく描かれ，テキストは図を囲む説明文のような形で書かれている。

「天文」セクションは天文学に関係していると思しき図が大きく描かれており，テキストは比較的少ない（第 3 図)。複数のページが折り畳まれ構成されている葉もあり，複雑な構造をしている。

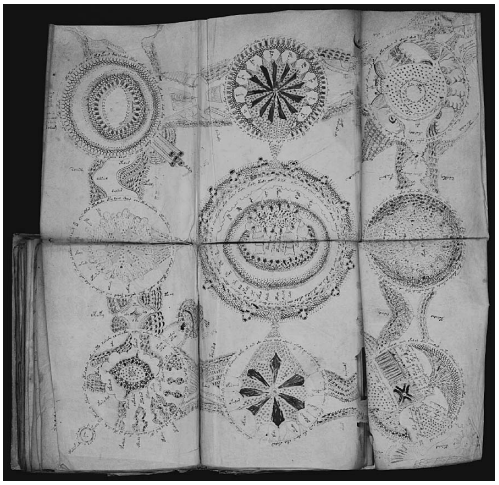
「生物」セクションの挿図には，複数の小さな裸

第 3 図　ヴォイニッチ写本：天文セクション (f. 71r) (Beinecke Rare Book and Manuscript Library, Yale University Library, MS 408) 画像出典：http://beinecke.library.yale.edu/dl_crosscollex



第 4 図　ヴォイニッチ写本：生物セクション (f. 81r) (Beinecke Rare Book and Manuscript Library, Yale University Library, MS 408) 画像出典：http://beinecke.library.yale.edu/dl_crosscollex

の女性が何かをしている姿が描かれている（第 4 図)。このセクションは「植物」セクションとは異なり，テキストが中心で挿図はページの空いている部分に描かれているように見える。



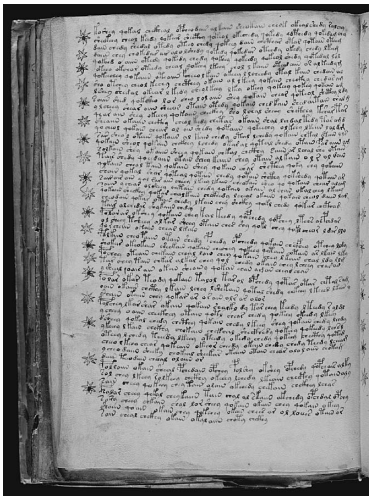
第５図　ヴォイニッチ写本：十二宮図セクション (f. 86v) (Beinecke Rare Book and Manuscript Library, Yale University Library, MS 408) 画像出典：http://beinecke.library.yale.edu/dl_crosscollex

「十二宮図」セクションは最もページ数が少なく，「天文」セクションと同様に複数のページが折り畳まれ構成されている葉がある（第５図）。半数のページには図ばかりで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予想される。

「薬草」セクションには２種類のページが含まれている。１種類は，「植物」セクションと区別ができないような１ページに一つの植物の絵が大きく描かれているページである。もう１種類は，それと似ているものの，根など植物の一部分のみが描かれている，複数の植物が同じページにある，薬草壺らしき物体が描かれている，などの点が異なるページである（第６図）。

「レシピ」セクションは他のセクションと異なり，星印のような行頭文字が描かれている以外には図はなく，テキストのみのセクションである（第７図）。

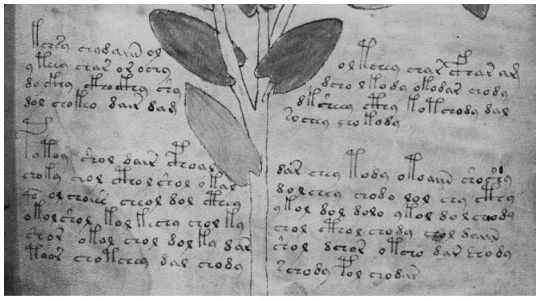
第６図　ヴォイニッチ写本：薬草セクション (f. 99r) (Beinecke Rare Book and Manuscript Library, Yale University Library, MS 408) 画像出典：http://beinecke.library.yale.edu/dl_crosscollex



第７図　ヴォイニッチ写本：レシピセクション (f. 111v) (Beinecke Rare Book and Manuscript Library, Yale University Library, MS 408) 画像出典：http://beinecke.library.yale.edu/dl_crosscollex

D.　翻字データ

未解読のヴォイニッチ写本に対してテキスト処理を行うためには，ヴォイニッチ文字を機械可読型の文字セットに置き換えた翻字データが必要となる。本研究では，近年のヴォイニッチ研究で一般的に使われている文字セットに従った。これは，Zandbergen らが考案した European Voynich



kchsy.chadaiin.ol-{plant}oltchey.char.cfhar.am-
yteeay.char.or.ochy-{plant}dcho.lkody.okodar.chody-
d!o.ckhy.ckho.ckhy.shy-{plant}dksheey.cthy.kotchody.dal-
dol.chokeo.dair.dam-{plant}sochey.chokody=
potoy.shol.dair.cphoal-{plant}dar.chey.tody.otoaiin.shoshy-
choky.chol.cthol.shol.okal-{plant}dolchey.chodo.lol.chy.cthy-
qo.ol.choeee.cheol.dol.cthey-{plant}ykol.dol.dolo.ykol.do.!!ch!!ody-
okol.shol.kol.kechy.chol.ky-{plant}chol.cthol.chody.chol.daiin-
shor.okol.chol.dol.ky.dar-{plant}shol.dchor.otcho.dar.shody-
taor.chotchey.dal.chody-{plant}schody.pol.chodar=

第8図　第1葉裏の画像データ（部分）と翻字データ　画像出典：http://beinecke.library.yale.edu/dl_crosscollex

Alphabet（以下，EVA）[34] と呼ばれる文字セットである。EVA は写本中に 10 回以上出現する文字を置き換えた 27 個の基本 EVA 文字，希に出現する拡張 EVA 文字，さらに句読点，メタ記号（レイアウト情報や翻字者によるコメントなどを示すための記号）から構成されている。なお，EVA の登場以前の研究では独自の文字セットが用いられることが多く，結果として少なからず異なった文字セットが複数存在する。EVA はそれらを比較し問題点を洗い出したうえで構築されており，正確性や一貫性の点で問題が少ないと考えたため，採用することとした。

ヴォイニッチ写本の EVA による翻字データには翻字者の異なる複数のデータが存在する。ここでは，EVA で翻字されており，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翻字データを用いた[35]。ヴォイニッチ文字は手書き文字かつ未解読であるため，翻字データの正確性について検討することは困難であるが，高橋によるデータは，筆者らがヴォイニッチ写本の画像と照合したうえで十分に信頼できると判断した。また，本研究の手法では，たとえ少々不正確な翻字が行われていたとしても，それらが文書中で一貫していれば問題にはならないと考えた。

例として，第 8 図に第 1 葉裏の画像データと高橋による翻字データを示す。

E. トークンの切り分け

ヴォイニッチ写本の言語構造は不明であるが，第 8 図からわかるように，一定の文字列の間に空白が挿入されており，多くのヨーロッパの言語と同様に空白で単語が区切られる分ち書きがされていると推測できる。しかし，空白で区切られた文字列がヴォイニッチ写本の言語における単語かどうかは不明である。ここでは空白で区切られた文字列をトークンとして扱い，本研究におけるテキスト処理の最小単位とした。また，ヨーロッパ言語のハイフネーションによる単語分割のように，一つのトークンが行をまたぐかについても不明であるが，テキスト処理においては行をまたがないものと仮定した。したがって，写本のテキストのトークン切り分けは空白と改行により行った。

EVA による翻字データには，単にある文字を置き換えた記号以外にも，特定の意味を付与された記号も含まれている。“*” や “?” は判別不能文字を示し，“{ }” 内に囲まれた部分は図や直前の文字の形態等に関する注釈の記述である（例えば，“{plant}” は植物の図があることを示す）。ここでは，“*” や “?” はそのままとした。しかし，注釈部分のテキストには，図の説明など，翻字者の解釈が追加されて本来のテキストとは異なる意味が含まれていると判断し，“{ }” で囲まれた部分はデータから除去した。また，高橋の翻字データには段落単位が示されているが，見た目の段落が意味を持っているかどうかは不明であるため，今回の分析ではその情報は用いないこととした。

以上のルールを適用し，翻字データをトークン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な統計を第 2 表に示す。

F. トークンに対する重みづけ

切り分けられたトークンの重みづけには，情報検索で一般的な TF-IDF 法を用いた。TF-IDF 法

第 2 表　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

は語の出現頻度 (TF: term frequency) と文書頻度の逆数 (IDF: inverse document frequency) から構成されるもので，TF と IDF の値の算出法には多くのバリエーションがある。本研究では部分文書の単位をページと設定したため，IDF は逆文書頻度ではなく逆ページ頻度となっている。

具体的には式 (1) のような形でトークンの重みづけを行っている。

$$w_{ij}=tf_{ij}\cdot idf_j=\frac{f_{ij}}{\max f_{ij}}\cdot\log\left(\frac{N}{n_j}\right) \qquad (1)$$

この式で w_{ij} はページ i におけるトークン j の重みを示している。式において前半部分 tf_{ij} の f_{ij} はページ i におけるトークン j の出現頻度を，max f_{ij} はそのページ内における最頻出トークンの出現頻度を示し，トークンの出現頻度を最頻出トークンの頻度で正規化したものを意味する。正規化した理由は，ページ内の単語数には大きなばらつきがあるためである。また，後半部分 idf_j は，n_j が写本全体においてトークン j が出現するページ数を，N が全ページ数を示している。

G.　部分文書同士の類似度算出

上記のトークンに対する重みづけデータを用いて各部分文書（ページ）同士のテキストデータの類似度を算出した。つまり，1 ページ目と 2 ページ目，1 ページ目と 3 ページ目，1 ページ目と 4 ページ目…，2 ページ目と 3 ページ目…，とすべてのページ同士について類似度を算出した。

類似度行列作成には，式 (2) で示されるキャンベラ距離を用いた。ベクトル空間モデルなどで文書間の類似度算出のために一般的に使われるコサイン相関係数ではなく，距離関数を用いた理由は，代表的な階層的クラスター分析法である重心法やウォード法では，類似度ではなく非類似度（距離）をパラメータとして必要とするためである。また，クラスタリングではユークリッド距離といった距離関数を用いることが一般的であるが，キャンベラ距離は値が小さく，差が少ないデータ同士に対しても非常に感度が高いとされているため[36] [p. 20]，本実験ではキャンベラ距離を採用した。

$$d_{ij}=\sum_{k=1}^{n}\frac{|w_{ik}-w_{ij}|}{|w_{ik}|+|w_{jk}|} \qquad (2)$$

この式 (2) で d_{ij} は部分文書 i と部分文書 j の距離を示し，それらの部分文書におけるトークンすべての重みから算出されている。この値は非類似度を示す距離であるため，値が小さいほど部分文書同士の類似度が高いことを意味する。このように，類似度ではなく距離関数を用いているため，以下では類似度行列ではなく距離行列と記述する。

H.　クラスター分析手法

次に，ページ同士の距離行列からクラスター分析手法を用いて文書クラスタリングを行った。クラスター分析手法は，結果がデンドログラムで表現される階層型手法と k-means 法などの非階層型手法に大別できる。文書クラスタリングでは規模の大きいデータに対しては計算量が少ない非階層型手法が用いられることが多いが，今回のヴォイニッチ写本の部分文書集合は比較的小規模であり，階層型手法を適用することが十分に可能である。そこで，結果についてデンドログラムからカットして，階層ごとの詳細な分析を行うことが可能な階層型手法を用いた。

また，階層型クラスター分析手法には，凝集 (agglomeration) 型と分割 (division) 型というクラスター階層の構築方法による類別がある。凝集型手法は個々のクラスターを併合していき最終的に一つのクラスターにまとめるもの，分割型手法は一つのクラスターを分割していき最終的に個々のクラスターに分割するものである。文書クラスタリングの領域では分割型ではなく凝集型クラス

ター分析手法が用いられることが多いため，本研究でも凝集型手法を用いた。

凝集型の階層型クラスタリング手法を部分文書に適用するための具体的な手順を以下に挙げる。

1) 部分文書それぞれをクラスターとする
2) クラスター同士の距離が最も小さいクラスター同士を併合する
3) 併合された結果に基づいて距離行列を再計算する
4) すべてのクラスターが一つのクラスターに併合されるまで 2) と 3) を繰り返す

凝集型で階層型の代表的なクラスタリング手法としては「単連結法 (single linkage method)」,「完全連結法 (complete linkage method)」,「群平均法 (median method)」,「重心法 (centroid method)」,「ウォード法 (Ward method)」がある。どの手法においてもクラスタリングの手順は上述の形で行い，クラスター間の距離算出方法以外は共通である。本研究では，これらの代表的な 5 手法すべてについて部分文書のクラスタリングをそれぞれ行った。

I. 評価手法

文書クラスタリングにおいて正しいクラスタリング結果（以下，正解集合）が存在する場合にはそれらを基準とした評価を行うことが可能であるが，未解読文書は未解読という性質上，正解集合の情報を入手することは困難なことが多い。ただし，今回対象とするヴォイニッチ写本の場合，挿図から推定されたセクション分けがあるため，それを正解集合と見なすことで評価を行った。

階層型クラスター分析手法を用いた場合にクラスタリング結果の評価を行うには，出力された階層構造をある基準でカットする必要がある。今回実験対象とした文書の場合，その文書に含まれるセクション数でカットするものとした。つまり，ヴォイニッチ写本の挿図から推定されたセクション数は 6 であるため，6 クラスターとなる高さでカットしている。

文書クラスタリング結果の評価には様々な尺度が用いられてきた。Zhao らの研究[37]をはじめとする文書クラスタリング評価に関する研究や文書クラスタリングに関するレビュー[38]を検討した結果，本研究では，エントロピー (Entropy)，純度 (Purity)，全体的な F 尺度 (F-measure) である F スコア (F-score) を用いることとした。

以下にそれぞれの評価尺度の説明を行う。その際に，クラスター数 k のクラスタリングの結果 C を $C=\{C_1, C_2, \cdots C_k\}$，正解集合 A を $A=\{A_1, A_2, \cdots A_k\}$，集合に含まれるデータ数を $|X|$ と表記する。例えば，正解集合を A_h とし，クラスターを C_k としたときに共通するデータ数は $|A_h \cap C_k|$ となる。

1. エントロピー

エントロピーはクラスタリング結果の評価において最も標準的な尺度の一つである。0 から 1 の値をとり，値が低いほどクラスタリングの結果が良好であることを示す。クラスタリング結果全体に対するエントロピーは，以下の式 (3) のように算出する。これはクラスターごとにエントロピー E_i を算出し，クラスター C_i に含まれるデータ数に応じた重みつき平均を取ったものである。

$$\mathrm{Entropy}=\sum_{i=1}^{K}\frac{|C_i|}{N}E_i \tag{3}$$

ここで N はクラスタリング対象の全データ数を示している。クラスターごとの E_i は以下の式 (4) で算出する。この式において，$P(A_h|C_i)$ は式 (5) で示されるようにクラスター C_i の文書が正解集合 A_h に属する確率である。

$$E_i=\sum_{h=1}^{k}P(A_h|C_i)\log P(A_h|C_i) \tag{4}$$

$$P(A_h|C_i)=\frac{|A_h\cap C_i|}{|C_i|} \tag{5}$$

2. 純度

純度はエントロピーと同様にクラスタリング結果の評価において標準的な尺度の一つである。この尺度は以下の式 (6) のように算出し，値が高いほどクラスタリングの結果が良いことを意味す

る。クラスタリング結果全体に対する純度は，クラスターごとに純度 P_i を算出し，エントロピーと同様に重みづけ平均をとったものである。

$$\text{Purity} = \sum_{i=1}^{k} \frac{|C_i|}{N} P_i \tag{6}$$

各 P_i は式 (7) のとおり算出される。

$$P_i = \frac{1}{|C_i|} \max_h |C_i \cap A_h| \tag{7}$$

3. *F* スコア

F スコアは情報検索における標準的な尺度である *F* 尺度 (*F*-measure) に基づく評価尺度である。*F* 尺度 F_{hk} は正解集合 A_h とクラスター C_k のペアに対する再現率 R_{hk} と精度 P_{hk} の調和平均で式 (10) によって算出される。再現率は網羅性を示す尺度であり，式 (8) のように正解集合 A_h 中の文書がクラスター C_k に属する確率である。精度は正確性を示す尺度で，式 (9) で示すとおりクラスター C_k の文書が正解集合 A_h に属する確率となる。

$$\text{Recall}_{hk} = \frac{|A_h \cap C_k|}{|A_h|} \tag{8}$$

$$\text{Precision}_{hk} = \frac{|A_h \cap C_k|}{|C_k|} \tag{9}$$

$$F\text{-measure}_{hk} = \frac{2R_{hk}P_{hk}}{R_{hk}+P_{hk}} \tag{10}$$

ただし，*F* 尺度は正解集合とクラスターのペアごとの値であり，クラスタリング結果全体の評価尺度ではない。全体的な評価尺度である *F* スコアは，式 (11) のように正解集合 A_h に対して *F* 尺度 F_{hk} が最大となるようなクラスター C_k を取り出して，それらの重みつき平均を算出したものである。*F* スコアは 0 から 1 の間の値を取り，値が高いほどクラスタリングが良好であることを意味する。

$$F\text{-score} = \sum_{h=1}^{k} \frac{|A_h|}{N} \max_k F_{hk} \tag{11}$$

IV. 実験結果

A. ページ・セクションの類似度

1. ページ同士の類似度

各ページ間の距離行列は量的に掲載することが難しいため，その一部分（10 ページ分）を第 3 表に示す。この表では，前述のようにページ同士の類似度を距離で算出しているため，値が小さいほどページ同士が類似していることを示している。

第 3 表において 1 行目にある第 1 葉表 (f.1r) は最初のページであるが，他のページとの距離の値がすべて 7907.0 である。この値は距離行列全体を通して一定であり，他のページとこのページに出現するトークンに全く重複がないことを意味する。

2. セクション同士の類似度

部分文書同士の距離行列ではセクション同士の関係を把握することが困難である。そこで，セクションごとの大まかな傾向を見るため，セクションごとにページ同士の平均距離を算出し，第 4 表に示した。これは，6 セクションのそれぞれについて，そこに含まれるページの距離を平均した値である（セクションとページの関係については第 1 表参照）。例えば，「植物」セクションに属する各ページ同士の距離を平均すると（1 葉表と 1 葉裏，1 葉表と 2 葉表，…1 葉表と 65 葉裏，1 葉裏と 2 葉表，1 葉裏と 2 葉裏…の平均）7643.7 となり，これは「植物」セクションと「天文」セクションそれぞれに属するページ同士の距離の平均（1 葉表と 66 葉表，1 葉表と 66 葉裏…，1 葉表と 73 葉裏，1 葉裏と 66 葉表，1 葉裏と 66 葉裏…の平均）である 7744.7 よりも小さい数字であるので，類似度がより高いということがわかる。

第 4 表では，各セクションの行ごとに平均距離（非類似度）の最も小さいものの枠を太線に，値を斜体にし下線を入れて示した。「十二宮図」セクションを除く 5 セクションでは，平均距離が最も

第 3 表　距離行列の一部（10 ページ分）

	Plant_f1r	Plant_f1v	Plant_f2r	Plant_f2v	Plant_f3r	Plant_f3v	Plant_f4r	Plant_f4v	Plant_f5r	Plant_f5v
Plant_f1r	0.0	7907.0	7907.0	7907.0	7907.0	7907.0	7907.0	7907.0	7907.0	7907.0
Plant_f1v	7907.0	0.0	7615.7	7619.0	7428.8	7531.0	7621.8	7663.9	7420.6	7661.4
Plant_f2r	7907.0	7615.7	0.0	7417.2	7491.7	7530.6	7672.6	7564.7	7455.2	7747.9
Plant_f2v	7907.0	7619.0	7417.2	0.0	7505.6	7678.2	7750.8	7499.2	7432.8	7772.9
Plant_f3r	7907.0	7428.8	7491.7	7505.6	0.0	7345.5	7597.0	7643.0	7250.0	7393.8
Plant_f3v	7907.0	7531.0	7530.6	7678.2	7345.5	0.0	7639.6	7346.1	7424.8	7702.3
Plant_f4r	7907.0	7621.8	7672.6	7750.8	7597.0	7639.6	0.0	7674.1	7686.8	7771.5
Plant_f4v	7907.0	7663.9	7564.7	7499.2	7643.0	7346.1	7674.1	0.0	7325.0	7699.8
Plant_f5r	7907.0	7420.6	7455.2	7432.8	7250.0	7424.8	7686.8	7325.0	0.0	7537.0
Plant_f5v	7907.0	7661.4	7747.9	7772.9	7393.8	7702.3	7771.5	7699.8	7537.0	0.0

第 4 表　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	***7643.7***	7744.7	7711.9	7699.3	7698.4	7716.8
天文	7744.7	***7659.0***	7731.7	7691.9	7717.4	7697.6
生物	7711.9	7731.7	***7013.0***	***7401.0***	7660.6	7307.3
十二宮図	7699.3	7691.9	7401.0	7445.2	7656.5	7419.9
薬草	7698.4	7717.4	7660.6	7656.5	***7635.9***	7639.9
レシピ	7716.8	7697.6	7307.3	7419.9	7639.9	***7203.7***

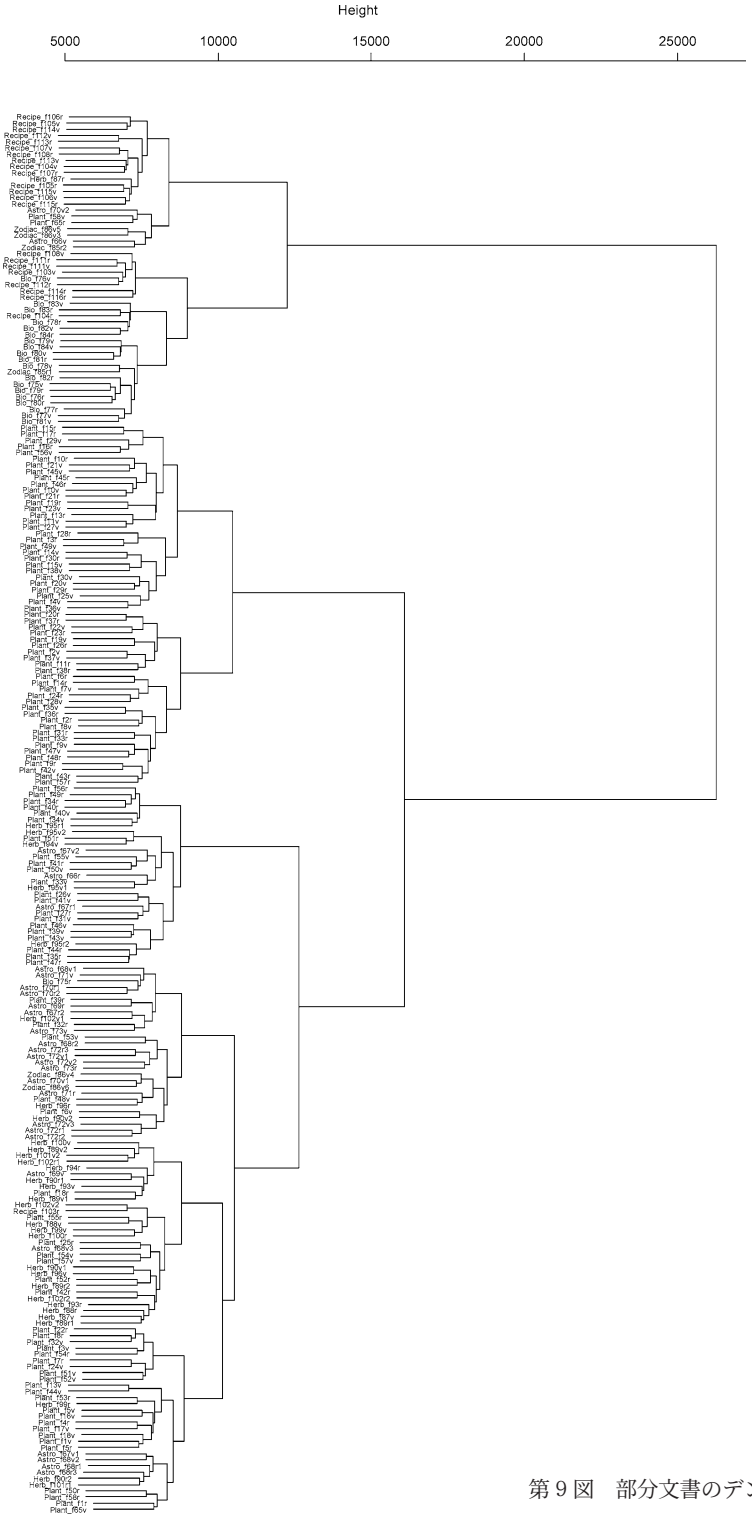
小さいのは同じセクション同士の交差する箇所となっている。つまり，あるセクションに属するページは，他セクションに属するページよりも，同じセクションの別のページに対してのほうが高い類似度を示しているのである。この結果は同一セクションに含まれるページ同士が近い関係であることを意味している。これは，挿図から見たセクション分け（大まかな構造）とテキストからの構造が一致した結果であるといえる。

唯一の例外である「十二宮図」は，「生物」セクションとの平均距離が最も小さくなっている。ここはテキストがほとんど含まれないページから構成されており，他のセクションと比較してページ数も少なく，テキスト処理を適切に行うこと自体が困難なセクションであるため，こうした結果になったと考えられる。

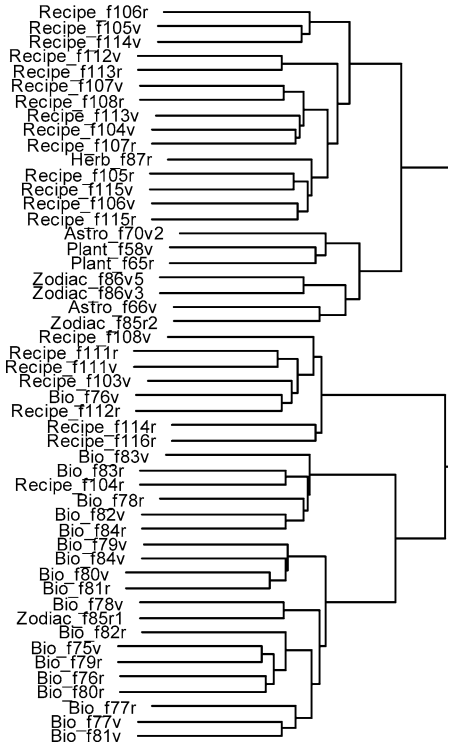
B.　部分文書クラスタリングの結果

本研究では代表的な五つの階層型クラスター分析手法を用いてヴォイニッチ写本の部分文書クラスタリングを行った。ここでは例として，分類精度が最も高いといわれる「ウォード法」を用いた場合の結果のデンドログラムを第 9 図に示す。図で各ページデータは「セクション名の英語略称_フォリオ番号（さらに，折り畳まれている場合はその順番）」で示している。例えば，「Bio_f82r」は，「生物セクションに含まれる第 82 葉表」のテキストデータであることを示す。

全体を概観するために高さが 15,000 の個所でデンドログラムを分割すると，三つのクラスターに分けられる。上のクラスター（第 10 図）には「生物」「レシピ」セクションのページがほぼまとまっており，「植物」「天文」「十二宮図」のページ



第９図　部分文書のデンドログラム



第 10 図　作成されたデンドログラムの一部

が少数含まれている。真ん中のクラスターは「植物」セクションのページのみから構成されている。また，右のクラスターには「植物」「薬草」「天文」セクションのページと「十二宮図」セクションの 1 ページのみが含まれる。

また，個々のセクションという点からこのデンドログラムを見ると，「生物」，「レシピ」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。また，「十二宮図」については，一部のページ群はまとまっているものの，文字がほとんど含まれないページは全体に分散する結果となった。

一方，「植物」と「薬草」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る結果となった。挿図に合わせて本文を記述した場合，「植物」と「薬草」はそもそも図自体が類似しているため，ページ内の本文の特徴が似たとしても不思議ではないと考えられる。一方，Rugg が主張するように中世の暗号器によって無作為に文書が「捏造」されたと想定した場合には，セクションごとの類似度をここまで自在に操作することが可能かという疑問が出てくる。つまり，テキスト処理技術に関する知識がない時代（言い換えればどのような処理がされるか予想できない時代）に，自然言語的な特徴を失わないまま，ページの特徴がセクションごとに類似するようにしつつ，似た挿図を持つ離れた場所のセクション同士に似たような特徴を持たせたうえで，そのような文書を無作為に作成できるかという問題である。この点で「植物」「薬草」の混在は Rugg の説に対する反証の一つとなる。

さらに，ページ番号との対応関係に着目すると，比較的近いページ同士が同じクラスターにまとまる傾向が見られた。隣り合うページ同士が最初に併合されたクラスター内に出現した例を「レシピ」セクションにおいて挙げると，第 107 葉裏と第 108 葉表，第 111 葉表と第 111 葉裏，第 112 葉裏と第 113 葉表である。

C. クラスタリング結果の評価と他文書との比較

本結果に対して文書クラスタリングで標準的に用いられているエントロピー，純度，*F* スコアといった評価尺度の値を算出した。しかし，これらの値を他の文書に対する結果なしに単独で解釈することはできない。そこで，比較対象となる文書を選定し，ヴォイニッチ写本と同様の形で部分文書クラスタリングを行った結果と比較し，分析を行った。

1．比較対象文書

性格が異なる 2 種類の文書を対象として比較実験を行った。一つは構造を持たない無作為な「捏造文書」である Rugg によるソフトウェア版「偽ヴォイニッチ写本」データ，もう一つがセクション構造を持つ真正文書として中世の写本『全ての事物の第五精髄の考察についての書』である。

a. Rugg のソフトウェア版「偽ヴォイニッチ写本」

ヴォイニッチ写本と比較する文書として，最初

に候補に挙がるのは，Rugg による前述の既往研究において「偽ヴォイニッチ写本」として無作為に作成されたデータである。Rugg によるデータで公開されているものとしては，カルダーノ・グリルを用いて手作業で作成された画像データと，ソフトウェア的に作成されたテキストデータの二つのバージョンが存在する。

手作業で作成された画像データは数ページと分量が少なく，挿図がごく一部にしか含まれていない。画像データであるためトランスクリプションが必要な点，挿図がないページに関してはセクション分けができない点，（2009 年 3 月時点で）Rugg のウェブサイトにアクセスできない点から，比較対象文書として扱うことが難しい[39]。

ソフトウェア版「偽ヴォイニッチ写本」データ (Fake Voynich Software Version) は，データ内のコメントによれば，Rugg が同僚である Jonathan Knight に依頼して作成したものである。元データの配布元である Rugg のウェブサイトは（2009 年 3 月時点で）アクセスできないため入手できないが，ヴォイニッチ研究者の一人である Stolfi のウェブサイトに転載され公開されている[40]。このデータは 3 章，3 ページとページ数は少ない。しかし，総語数は 1,950 語と長さがあり，ページあたりの平均語数は 650 語である。本物のヴォイニッチ写本のページあたりの平均語数が 166 語であることから考えると，このデータの 1 ページは本物の 3 ページないし 4 ページに相当する分量となっている。また，Rugg による再現手法の説明からは，章ごとにソフトウェア的にカルダーノ・グリルないしテーブル自体を変更し，ヴォイニッチ写本におけるセクション分けに準ずる操作がされていると考えられる。

以上のことから，ソフトウェア版「偽ヴォイニッチ写本」データを比較対象文書として用いることとした。このデータのページあたりの語数は本物に比べ非常に多いため，本物に合わせてページあたりの平均語数が 166 語前後となるように，20 行ごとに分割し 1 ページとした。また，元の 1 ページを 1 セクションと見なした。結果として「偽ヴォイニッチ写本」の実験データは 3 セクション 12 ページ（各セクションは 4 ページ）となった。

b. 『全ての事物の第五精髄の考察についての書』

検証実験のための比較対象文書として，主題的なセクション分けがなされている真正文書が必要であった。そこで，ヴォイニッチ写本と同時代に作成された写本を選定対象としたが，翻字されテキストデータが入手可能な中世の文書はそれほど多くは存在しない。ここでは (1) 主題的なセクションに分かれている，(2) 一定以上の長さを持つ（フォリオ数が 2 桁以上），(3) 一人の著者による文書である，(4) 翻字されたテキストデータが入手可能である，(5) 科学に関する中世の文書である，という条件から文書データを選定した。

結果として Irma Taaivitsainen らによる *Middle English Medical Texts* という CD-ROM[41] に収録された『全ての事物の第五精髄の考察についての書 (Liber de consideratione de quintae essentiae omnium rerum)』（以下，『第五精髄』とする）を対象とした。

CD-ROM に収録された『第五精髄』に関する説明によれば，この文書はフランスのフランシスコ会修道士 John of Rupescissa (1310–1360s) によって書かれた錬金術的要素の強い医学書を中世英語に翻訳したものである。

Middle English Medical Texts に収録されているテキストデータは，グラスゴー大学図書館の John Ferguson コレクションの一冊 (MS Ferguson 205) を翻字したものである。第五精髄（アリストテレスの言うところの完全元素である第五元素）の製法について書かれた部分（第一の書）は第 1 葉表から第 33 葉裏までである。このデータは 11 のカノンに分けられ，カノンごとに異なる主題が扱われているため，カノンごとに分けたものをセクション分けと見なすことが可能である。本実験では部分文書クラスタリングの難度がヴォイニッチ写本と同じ程度となるようセクション数を揃えて，前書きを除いた最初から 6 セクション分，具体的には第 1 葉表から第 16 葉表までの 31 ページ分のテキストデータを用いた。なお，ページ途中で次セクションが始まる場合は

第 5 表　比較対象文書の基本統計

	ヴォイニッチ写本	偽ヴォイニッチ写本	第五精髄
異なり語数	7,907	635	1,960
単語数	37,359	1,950	10,402
ページ数	225	12	36
平均単語数（ページ）	166.0	162.5	288.9
平均単語長	5.0	5.5	4.3
セクション数	6	3	6

第 6 表　部分文書クラスタリング結果に対する評価

	ヴォイニッチ写本	偽ヴォイニッチ写本	第五精髄
エントロピー	0.40	0.97	0.39
純度	0.71	0.42	0.64
F スコア	0.57	0.41	0.60

二つに分割したため，クラスタリングの対象となった総ページ数は 36 ページである。

2.　評価と比較結果

ヴォイニッチ写本の部分文書クラスタリングと同様の手順で，部分文書としてのページ内のトークンの重みづけをした。テキスト処理を行った結果，ヴォイニッチ写本と比べ，比較対象文書のテキストの特徴は第 5 表のようになった。

各文書についてトークンへの重みづけから部分文書同士の類似度を算出し，階層型クラスター分析手法を適用した。クラスタリング結果を文書に含まれるセクション数に応じてカットし，セクションを正解集合として，エントロピー，純度，*F* スコアという評価尺度を算出した結果を第 6 表に示す。

第 6 表では，「偽ヴォイニッチ写本」はセクション数が一番少なく，文書クラスタリングの課題としては難易度が低いにもかかわらず，値が高いほど悪い結果であるエントロピーでは 0.97 と他の二文書よりも非常に高い値を，純度，*F* スコアでは他の二文書よりも明らかに低い値を示している。つまり，偽ヴォイニッチ写本については，すべての評価尺度において部分文書クラスタリングが成功していないことが明らかである。これは，無作為に作成された文書には何らの構造も期待できないという本手法が前提とした経験的な予想に合致する。一方で，『第五精髄』のエントロピーは 0.39 と低く，純度は 0.64, *F* スコアは 0.60 と良い値を示している。この結果は「捏造文書」と真正文書は部分文書クラスタリングによって明確に区別できることを示している。前者には構造がなく，後者にはセクションによるまとまりが見られるからである。

本物のヴォイニッチ写本については，『第五精髄』とほぼ同様の値となっている（例えば，エントロピーはヴォイニッチ写本が 0.39,『第五精髄』は 0.40）。『第五精髄』は未解読文書ではなく，セクションごとに主題がまとまっている文書である。このことは，ヴォイニッチ写本が挿図によるセクション分けに対応する形で一定のまとまりがある文書，つまりセクションに応じた構造を持つ文書であることを示していると考えられる。

D.　考察

今回の実験結果からは，提案手法を用いることで，Rugg が主張するような手法で無作為に作成された「捏造文書」と真正文書を区別できること

が明らかになった。

ヴォイニッチ写本については，部分文書（ページ）同士には，挿図によるセクション分けに対応する形で一定のまとまりがあること，近いページの本文同士のほうが離れたページの本文同士よりも類似度が高いこと，が確認された。つまり，本文の構造と挿図によるセクションやページ順という別の手がかりの構造とが一致しているということになる。Rugg による既往研究[26]で示唆されたようにデタラメに「捏造」された場合は，このような挿図と本文の部分文書との対応は生じないはずである。比較実験で検証したように，ソフトウェア版「偽ヴォイニッチ写本」に関してはセクションと部分文書が明らかに対応していない結果となった。一方，本物のヴォイニッチ写本については，主題に対応したまとまりのある真正文書と同様の結果が出ている。

したがって，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，デタラメな「捏造文書」ではないという結論が導かれる。つまり，正しい手がかりが得られれば，ヴォイニッチ写本の解読は可能であると考えられる。

同時に，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ともに変換方式が変化していくような暗号ではないことを示している。そのような場合には，セクションごとのまとまりは見られないはずだからである。一方で，単純な暗号であれば，これまでの取り組みの中ですでに解読されているはずである。このことからは，ヴォイニッチ写本は，暗号ではなく，既存の言語体系によらない人工言語または未知の言語で書かれた可能性が高いと考えられる。この結論は，Freedman などの説を支持するものである。

なお，興味深いことに，第 1 葉表には他のページと共通するトークンが全く出現しなかった。第 1 葉表については，挿図がないことから特殊なページであるという可能性が既往研究でも指摘されていたものの，この事実は筆者らが調査した範囲のヴォイニッチ写本関係文献では言及されておらず，新しい発見だと考えられる。トークン単位での分析で第 1 葉表に他のページと重複が全くないことは，この最初のページは特殊なページであることを強く示唆している。例えば，ヴォイニッチ写本の本文が暗号で書かれていた場合の暗号鍵である，あるいは，本来の写本作者とは異なる誰かが他のページに似せて作成し追加した，といった可能性を考えることができる。

V. 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。実験の結果，ヴォイニッチ写本は一貫性のある構造を持つ文書であり，デタラメな「捏造文書」ではないという結論を導くことができた。本研究の結果は，未知の言語で書かれていたり，内容が未解読な文書であったりしても，テキスト処理手法を応用して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，該当文書が「捏造文書」であるかどうかを判断し，解読可能性の判定を行うことが可能であることを示している。このような解読可能性の判定は，どの程度，未解読文書に対して解読の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。

また，この手法の副次的な効果としては，図書館等が「捏造」であるかどうか疑わしい資料について判断する際の参考となりうることが挙げられる。例えば，今回の実験対象としたヴォイニッチ写本は，Voynich による売り込みの努力にもかかわらず，いずれの図書館も購入を控え，最終的には古書籍商 H. P. Kraus が購入し，後にバイネッケ図書館に寄贈したものである。そのような場合，「捏造文書」ではないことが早期に明らかになっていれば，研究用途として購入する可能性もあるだろう。

最後に，本手法が有効である文書とはどのようなものかを述べる。本手法は，文書の構造を判断するものであるため，前述のトリテミウスの *Steganographia* のように，暗号文のままでも意味のある文書の体裁をとっているものや，複式換

字式で書かれた暗号文書に対しては不向きである。むしろ，本実験対象と同様に，暗号で書かれているのかどうかの判断がつかない，解読不可能であり続けている文書に対して適用する際に有用であろう。例えば，前述の British Library 所蔵の『エノクの書』や，ハンガリー科学アカデミー図書館が所蔵する 19 世紀の捏造写本だと言われるローホンク写本 (Rohonc Codex)(Budapest, Magyar Tudomanyos Akademia, K 114)[42]，20 世紀のイタリアの建築家 Luigi Serafini が著した架空平行世界の百科事典『コデックス・セラフィニアヌス (Codex Seraphinianus)』[43] などの未解読文書の解読可能性の判定において有効だと期待できる。

注・引用文献

1) Chadwick, John. 線文字 B の解読. 大城功訳. みすず書房，1997, 239p.（原著の第二版の翻訳）.
2) Singh, Simon. 暗号解読：ロゼッタストーンから量子暗号まで. 青木薫訳. 新潮社，2001, 493p.
3) Reeds, Jim. Solved: The ciphers in Book III of Trithemius's Steganographia. Cryptologia. 1998, vol. 22, no. 4, p. 291–319.
4) この写本についての概説書（2002 年の BBC ドキュメンタリーに基づくもの）は邦訳で読むことができる。 Kennedy, Gerry; Churchill, Rob. ヴォイニッチ写本の謎. 松田和也訳. 青土社，2006, 380p.（原著 2004 年の翻訳）.
5) Gillogly, Jim; Reeds, Jim. "Voynich Manuscript Mailing List HQ." Voynich Manuscript Mailing List HQ, http://www.voynich.net/, (accessed 2009–03–09).
6) Beinecke Rare Book & Manuscript Library. MS 408. Yale University Library Finding Aid Database. http://webtext.library.yale.edu/beinflat/pre1600.ms408.htm, (accessed 2009–03–09).
7) Beinecke Rare Book and Manuscript Library. "Beinecke Rare Book and Manuscript Library: Digital Images & Collections Online." Digital Images Online, http://beinecke.library.yale.edu/dl_crosscollex, (accessed 2009–03–09).
8) 例えば 3 丁末には "3 q" というように，通常のアラビア数字に加え，丁 (quire) を示す "q" が書かれているように見えるが，"q" はヴォイニッチ文字であるという可能性もある。
9) O'Neill, Hugh. Botanical observations on the Voynich MS. Speculumn. 1944, vol. 19, no. 1, p. 126.
10) 後出の文献 17), p. 1 では，"Jacobj à Tepenece" と多少異なる綴りが紹介されている。
11) 書簡の複写画像については後出の文献 17), p. 80，その全文の翻刻は文献 6)，邦訳は文献 4) の p. 32–33 を参照。D'imperio によれば，Marci は 1665 年もしくは 1666 年にルドルフ 2 世の主治医をしていた。ただし，Marci は主治医を勤めるには若すぎたはずだという反論もある。ルドルフ 2 世に売るためであれば，前述の「強い動機を持った捏造者」を仮定することも可能である。
12) John Dee については，以下の文献を参照. French, Peter J. ジョン・ディー：エリザベス朝の魔術師. 高橋誠訳. 平凡社，1989, 345p.
13) Watson, A. G. ; Roberts, R. J., ed. John Dee's Library Catalogue. Bibliographical Society, 1990, p. 172–173 (DM93).
14) Newbold, William Romaine; Kent, Roland Grubb. Cipher of Roger Bacon. Kessinger, 2003, 314p.（原著 1928 年の再版）.
15) Manly, John Matthews. Roger Bacon and the Voynich MS. Speculum. 1931, vol. 6, no. 3, p. 345–391.
16) 例えば Brumbaugh, Robert S. The Voynich 'Roger Bacon' cipher manuscript: Deciphered maps of stars. Journal of the Warburg and Courtauld Institutes. 1976, vol. 39, p. 139–150. がある。
17) D'Imperio, M. E. The Voynich Manuscript: An elegant enigma. Aegean Park Press, 1981, 148 p. (Cryptographic Series, No. 27).（原著 1978 年の再版）.
18) Reeds, Jim. William F. Freedman's transcription of the Voynich Manuscript. Cryptologia. 1995, vol. 19, p. 1–25.（アナグラムの日本語訳は文献 4)，p. 199 より引用）.
19) Currier, Prescott H. "Some important new statistical findings". Proceedings of a Seminar held on 30th November 1976 in Washington D. C. ed. by M. D'Imperio.（http://www.voynich.nu/extra/curr_main.html に転載されている）(accessed 2009–03–09).
20) Landini, Gabriel; Zandbergen, René. A well-kept secret of mediaeval science: The Voynich Manuscript. Aesculapius. 1998, vol. 18, p. 77–82.
21) Zandbergen, René. From digraph entropy to word entropy in the Voynich Manuscript. The Voynich Manuscript. 2002–08–13. http://www.voynich.nu/extra/wordent.html, (accessed 2009–03–09).
22) Zandbergen, René. Voynich MS: Analysis Section (5/5). The Voynich Manuscript. 2002–10–05. http://www.voynich.nu/a_synt.html, (accessed 2009–03–09).

23) 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. 1978, vol. 23, no. 3, p. 59–75. http://www.voynich.nu/extra/dimperio92c.html から改訂版[1992]が入手可能, (accessed 2009–03–09).
24) Stolfi, Jorge. Scatterplots of VMs pages. Bem-vindo ao Instituto de Computação da UNICAMP. 1998–07–06. http://www.dcc.unicamp.br/~stolfi/voynich/98–06–19-pageplots/, (accessed 2009–03–09).
25) Barlow, Michael. The Voynich Manuscript: By Voynich? Cryptologia. 1986, vol. 10, p. 210–216.
26) Rugg, Gordon. An elegant hoax?: A possible solution to the Voynich Manuscript. Cryptologia. 2004, vol. 28, no. 1, p. 31–46.
27) 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. 2004, vol. 291, no. 1, p. 104–109.
28) 特定の場所に穴のあいた板を用いる暗号作成方法で，通常は穴の部分に伝えたいメッセージを一文字ずつ書き，残りの場所は他の文字で埋めて作文することにより，メッセージを隠す．復号には同じ板を用いる。ただし，Rugg は，文字を並べた紙の上に板を置き，穴の部分の文字を順番に並べていくことにより，無意味な文書を作成するという再現実験を行った。
29) Oxford University Press. Oxford Dictionary of National Biography Online. Oxford Dictionary of National Biography. http://www.oxforddnb.com, (accessed 2009–03–09).
30) 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/news/2003/031217/full/news031215–5.html, (accessed 2009–03–09).
31) 例えば，Another twist in the tale. Economist. 2004, vol. 370, no. 8375, p. 71.
32) 例えば，文献 27) の翻訳である Rugg, Gordon. ヴォイニッチ手稿の謎．日経サイエンス．2004, vol. 34, no. 10, p. 86–92.
33) Schinner, Andreas. The Voynich Manuscript: Evidence of the hoax. Cryptologia. 2007, vol. 31, no. 2, p. 95–107.
34) http://www.voynich.nu/extra/eva.html に EVA alphabet の概要とコード表などがある (accessed 2009–03–09)。
35) EVA による高橋健の翻字データは以下のサイトで公開されている。Stolfi, Jorge. Reeds / Landini's interlinear file in EVA, version 1. 6e 6. Bem-vindo ao Instituto de Computação da UNICAMP, http://www.dcc.unicamp.br/~stolfi/voynich/98-12-28-interln 16 e 6/, (accessed 2008–03–09).
36) Gordon, A. D. "2 Measures of similarity and dissimilarity". Classification. 2nd ed. Chapman & Hall, 1999, p. 15–34. (Monographs on statistics and applied probability, 82).
37) Zhao, Ying; Karypis, George. Criterion Functions for Document Clustering. 2001, UMN CS 01–040, 30p.（http://glaros.dtc.umn.edu/gkhome/fetch/papers/vsclusterTR01.pdf より本文が参照可能）(accessed 2009–03–09).
38) 岸田和明．文書クラスタリングの技法：文献レビュー．Library and Information Science, 2003, no. 40, p. 33–75.
39) ソフトウェア版「偽ヴォイニッチ写本」の配布元サイト http://www.keele.ac.uk/depts/cs/staff/g.rugg/voynich/software_version.html,（2009年3月現在アクセス不能であるが，Internet Archive 経由ではアクセスができる。http://web.archive.org/web/20050316224351/http://www.keele.ac.uk/depts/cs/staff/g.rugg/voynich/software_version.html, [accessed 2009–03–09]）.
40) ソフトウェア版「偽ヴォイニッチ写本」を転載しているサイト http://www.dcc.unicamp.br/~stolfi/voynich/Notes/tr-stats/dat/voyp/grs/tot.1/raw.evt, (accessed 2009–03–09).
41) Taaivitsainen, Irma; Pahta, Paivi; Mckinen, Martti. Middle English Medical Texts. John Benjamins Pub Co., 2005. (CD-ROM).
42) Csaba, Csapodi. A "Magyar Codexek" elnevezésü gyüjtemény (K31-K114). (Catalogi collectionis manuscriptorum bibliothecae academiae scientiarum hungaricae vol. 5). Bibliotheca academie scientiarum hungaricae, 1973, p. 109 (K114).
43) Luigi, Serafini. Codex Seraphinianus. Rizzoli, 2006, 305p.（原著 1981 年の再版）.

要　　旨

【目的】未解読文書に関する研究は，文書内容の解読に焦点を当てたものが多い。しかし，長年にわたって解読不能である文書は，何らかの意図で作成された意味をなさない「捏造文書」であり，そもそも解読自体ができない可能性もありうる。本研究の目的は，文書構造の有無から解読可能性そのものを判定する手法を提案することである。

【方法】既存の多くの言語に応用可能なテキスト処理技術は未解読文書に対しても有効であるという前提に基づき，未解読文書の部分文書同士の類似度をクラスタリング手法によって分析することにより，首尾一貫した文書構造の有無を検証する。次に，本書構造と，図表やページ順など他の手がかりから導かれる構造との対応関係を比較・分析することによって，「捏造文書」を判定する。

【結果】提案手法を用いて有名な未解読文書であるヴォイニッチ写本を分析した結果，本文の構造と挿図・ページから推測される構造が一致することが明らかになった。つまり，ヴォイニッチ写本は一貫性のある構造を持つ文書であり，「捏造文書」ではない可能性が高いと判定できる。実験により，提案手法の適用可能性を示すことができた。