ノイズを含む語のクラスタリング
～歴史的科学文献の情報検索の予備実験として～

岡本 里夏　　神門 典子

総合研究大学院大学　　国立情報学研究所
複合科学研究科　　総合研究大学院大学
情報学専攻博士課程

{lica, kando}@nii.ac.jp

1　はじめに

歴史的な書籍や文献を機械可読形式にする取組みは国内外で多く行われており，科学史研究においても実際の書籍を図書館で閲覧するのではなく，過去に発行された科学論文をインターネット上で機械可読形式になったものを閲覧する機会が多くなってきた．そのような環境で歴史的な文献を数多く調査する場合，一般的には OCR によってその内容がテキスト形式で PDF ファイル等に埋め込まれた情報を用いて全文検索を行う．現状では，機械可読形式になった歴史的な科学文献を全文検索する上で，次の主な 3 つの問題点が存在する．

(1) OCR エラーに起因するテキスト情報の不正確さ
- 綴りの間違い
- 文や語の区切りの間違い
- レイアウト・エラー等による内容の消失

(2) 正書法や専門用語の変遷
(3) 数式の表示形式の変遷

本論文では (1) の『OCR エラーに起因するテキスト情報の不正確さ』に着目する．自然言語処理的手法を用いて歴史的科学文献から科学史研究のために有用な情報を抽出するために，OCR エラーや旧正書法や歴史的綴りを含む語（以下「ノイズ語」とする）を含んだ文献を対象に，特別な辞書やルールを使う事なく，ノイズ語を含んだままで全文検索する手法を検討する．

1.1　研究の動機

以前科学史の分野にて 19 世紀のドイツの物理学者 R.Clausius について行った際 [6] に問題になったのが，インターネット上にある文献の OCR ノイズの多さである．ノイズが多いテキストでは，語や文の境目自体が信用できない場合が多くある．そのため，それらの文書の解析をするためには，まずノイズを処理する必要があったが，そのほとんどは手作業で行わざるを得ず，非常に時間がかかった．

一方，19 世紀の文献はその保存状態に比較的問題が少ないとはいえ，数式表現の違いや，専門用語の変遷など，文献の調査を進める上で手間がかかる事が多かった．数式表現は現在では使われていない形式の記号があったり，現在使われている記号であっても意味が違うものもある (ex. ライプニッツの微分記号が使われていても，実際の意味するところは偏微分の場合がある)．これらを解決して省力化を図り，より本質的な研究に時間を当てようと考えたのが，この研究の動機である．

2　Annalen der Physik コーパスの作成

筆者は現在，フランス国立図書館所蔵の Annaen der Physik という著名な物理学の科学論文誌の中から 1799 年から 1943 年の PDF ファイルをもとにテキストコーパスを作成中である．（フランス国立図書館のデジタルアーカイブ "Gallica" より非営利の利用として許諾をうけた．）Gallica にある Annalen der Physik 誌（1779 年～1943 年）は PDF ファイル形式で 362 件あり，そのうち今回使用するデータは OCR でテキストデータ化済みの PDF ファイル 138 件，PDF ファイル（画像とテキストデータ）の合計は約 4GB である．

3 ノイズを含む語のクラスタリング

3.1 クラスタリングに着目した理由

ノイズを含む文書の情報検索を行う場合，多くの手法が提案されている．一方，昨今では機械学習の分野の研究が盛んであり，教師無しの機械学習の一手法であるクラスタリングを用いて分類をすることも多く提案されるようになった．ルールや辞書を作ってノイズの処理を行ったり語の正規化を行うことは現状では人の手が入ることも多く，制作コストが高い．一方，仮に教師無し機械学習の手法をノイズを含む文書の検索処理に適用できれば，比較的低コストで行えるのではないかと考えたのが，クラスタリングに着目した理由である．

3.2 ノイズ語の定義

本稿では，OCR エラーや歴史的綴り，あるいは言語に特有の格変化などにより，辞書の見出しより変化した綴りを持つ語を「ノイズ語」と定義する．

3.3 編集距離アルゴリズム

今回クラスタリングには編集距離を利用する．Perl モジュールのアーカイブである CPAN(http://cpan.perl.org/index.html) には，編集距離は Levenshtein 距離, Damerau-Levenshtein 距離, Wagner-Fischer 距離，Brew 距離を計算するモジュールが登録されている．これらのモジュールを利用して編集距離の計算を行った．

3.3.1 類似度

今回設定されたキーワードからクラスタリングを行うにあたっては，それぞれのアルゴリズムより計算された編集距離 D_{edit} と比較したキーワードの文字数 N を用いて計算する類似度を利用した．類似度 s は次のとおりである．

$$s = 1 - \frac{D_{edit}}{N}$$

4 ノイズ語クラスタリングのための予備実験

4.1 目的

ノイズを含む歴史的科学文献 (主に物理学の論文) を検索するために、キーワードと検索エンジン Indri[4] でインデックスされた語の編集距離を利用してクラスタリングする方法を検証する．

4.2 前提

この予備実験は次の事項を前提とする．

- ノイズ語の元となる OCR エラーや歴史的綴りは，共に元のキーワードに対して文字編集操作を行った結果とみなす．
- ノイズ語と元のキーワードの違いの原因（OCR エラー，歴史的綴り，あるいは格変化や時制の変化などによる語形の変化なのかどうか）は一切区別しない．
- ノイズ語について，語幹抽出 (stemming) や辞書見出し化 (lemmatizing) などの文字列の正規化操作は行わない．
- 編集距離を計算する元となるキーワードは，大文字は用いず，小文字のみを使用した辞書見出しの形式にする．

4.3 仮説

この予備実験に際して筆者が立てた仮説は次の通りである．

(1) 検索の為に語の正規化をしなくても，ある程度までは編集距離によって似たものをクラスタリングできる．
(2) ノイズ語とキーワードの編集距離を測り，類似度によってクラスタリングしたものを同義語群として扱い，それらの語群をクエリとして投入することで，ノイズ語ではない元のキーワードだけのクエリよりも多くの関連性のある検索結果を得る事ができる．

本稿では, (1) についての検討を行う．

4.4 準備

4.4.1 実験環境

今回の実験に使用した機材は MacBook Air 1.8GHz Intel Core i7 で，メモリは 4GB 1333Mhz DDR3 である．

4.4.2 Annalen der Physik の PDF ファイル

予備実験に使用する Annale der Physik のデータは，フランス国立図書館の電子図書館である "Gallica" からダウンロードした．

4.4.3 検索用インデックスの作成

今回の実験では Indri 4.5 の IndriBuildIndex の機能を使用し，基本的な機能のみでインデックスの作成を行った．

4.4.4 キーワードのリストアップ

キーワードのリストアップに際しては，来春出版予定である R.Clausius の力学的熱理論論文集の翻訳書の

ためにリストアップしたキーワードの中から一単語のみで構成されるドイツ語の用語 102 個を選んだ．これらのキーワードは R.Clausius の論文に出てきたままの綴りを用いた．そのため現代のドイツ語の綴りとは違うものが含まれている．

4.4.5 編集距離計算のための Perl スクリプト

この予備実験では Perl のスクリプトを使って，Indri で作成したインデックスよりノイズ語を抽出し，編集距離を計算した．

4.4.6 ノイズ語の区別

ノイズ語の区別は人手により確認した．

4.5 実験結果

4.5.1 実行時間

R.Clausius の熱力学等の論文から選択した 102 語のキーワードと Indri でインデックスされた語 756,345 語の編集距離の計算にかかった時間は合計で 31 時間 15 分 54 秒であった．

4.5.2 編集距離

ここでは Levenshtein 距離の計算において類似度でソートした上位 5 番目までをいくつかのキーワードについて示す [表 1 参照]．ここに示したノイズ語のソートは Levenshtein 距離の類似度で行った．'hub' の 3 文字が一番少ない文字数で，'electrolyse' を含む 11 文字が今回一番多かった文字数であり，最長の文字数は 'elementarstrahlenbüschel' であった．編集距離のアルゴリズムは 4 種類と文字操作コストをかえて 1 種類試した．文字操作コストをかえて試した 1 種類以外は類似度の上位にクラスタリングされたものはほぼすべて同じ値となった．そのため，この表からは一部のみ表示した．

4.5.3 考察

今回行った 5 種類の編集距離の算出方法では，編集距離の値は，Levenshtein, Damerau-Levenshtein, Wagner-Fischer, Brew について大きな差はみられなかった．これら 4 種類と比較的差がでたのは，置換操作のコストを 2（1 字削除 +1 字挿入）とした Brew 距離の場合であった．これは Brew 距離のアルゴリズムというよりも，編集距離のコストの考え方によるものと予想される．つまり，Perl モジュールで他のアルゴリズムで置換のコストを変更できていれば，同様の結果が出たと思われる．

一方，より少ない文字数では類似度に大きな差がある．これは文字数で正規化する形で類似度を計算しており，少ない文字数では編集距離による影響をより受けやすいためと考えられる．また，より多い文字数の場合，類似度は上位 10 位以内程度では，より少ない文字数の場合よりも大きな差はみられない．これもより少ない文字数と同様の理由で，類似度の計算の性質上，文字数が多いと編集距離のコスト差が平均されてしまうためと考えられる．

クラスタリングされた語の妥当性については，10 件前後を元の PDF 文書を手動で検索し，目視で確認を行った．詳細な分析は今後の課題であるが，語の文字数が少ない場合ほど OCR エラーにより分割された目的の語ではない別の語の一部を拾うケースが多かった．長めの文字数を持つものは，元のキーワードそのものではないが，関連する語（ex. electrplyse: 電気分解, electrolyt: electrolyte[電解質] の複数形等）がクラスタリングされているケースも見られた．また，今回一番長い語である elementarstrahlenbüschel について，この語そのものあるいはそのノイズ語がクラスタリングされる事はなかった．この語は「輻射線束の成分」を意味しており，ある種の専門用語である．ドイツ語の専門用語は複数の単語を合成してできる複合語が多く見られる．この事から，類似度の上位の語は複合語を構成する

文字数の多少で類似度が影響を受けることは，クラスタリングの指標として好ましくないと考える．この影響を避けるには，より影響を受けにくい類似度の計算方法を工夫するか，あるいは，ドイツ語の複合語のような長大語（語の組み合わせにより文字数が多くなった語）を分割する必要がある．しかし，問題点で述べたように，ノイズが多いテキストでは，語や文の境目自体が信用できない場合が多くあるため，誤った語や文の区切りのまま複合語を分割することは，さらにノイズを増やす結果につながりかねない．クラスタリングの指標として類似度を使いやすくするためには，計算方法の工夫や対象とする語の文字数を一定の範囲にとどめる事で文字数の影響を排除する必要がある．

5 関連研究

ドイツ語の歴史的綴りに着目した研究としては [1], [2] がある．またノイズを含んだテキストデータの扱いについては [5] に多くの手法が紹介されている．音声認識によるテキスト生成時のノイズ処理については [3]

表 1　3 種類の編集距離アルゴリズムの比較 (上位 5 番目まで. 小数第 3 位以下切捨て)

クラスタ	出現	ノイズ語	Levenshtein		Brew		Brew (置換コスト 2)	
hub	頻度	y/n/—	編集距離	類似度	編集距離	類似度	編集距離	類似度
hub	14	—	0	1.00	0	1.00	0	1.00
zub	11	n	1	0.66	1	0.66	2	0.33
ub	238	n	1	0.66	1	0.66	1	0.66
tub	3	n	1	0.66	1	0.66	2	0.33
sub	121	n	1	0.66	1	0.66	2	0.33
electrolyse	頻度	y/n/—	編集距離	類似度	編集距離	類似度	編集距離	類似度
electrolyte	40	n	1	0.90	1	0.90	2	0.81
elektrolysen	1	y	2	0.81	2	0.81	3	0.72
eleklrolyse	1	y	2	0.81	2	0.81	4	0.63
elekfrolyse	1	y	2	0.81	2	0.81	4	0.63
electroly	1	y	2	0.81	2	0.81	2	0.81
elementarstrahlenbüschel	頻度	y/n/—	編集距離	類似度	編集距離	類似度	編集距離	類似度
elementarstrahlen	4	n	8	0.68	8	0.68	8	0.68
elementarparalleledurch	1	n	10	0.60	10	0.60	18	0.28
strahlenbuschel	1	n	11	0.56	11	0.56	12	0.52
mentarstrahlungen	1	n	11	0.56	11	0.56	14	0.44
etementartheitchen	1	n	11	0.56	11	0.56	15	0.40

が，語幹処理や辞書見出し化がノイズを低減する効果があると述べている．

6　おわりに

本論文では，ノイズ語を含む歴史的科学文献を全文検索するために行った予備実験について述べた．今回行った予備実験の結果をふまえ，今後はノイズ語を含んだ文献の情報検索に有効な手法についてさらなる検討をしていきたい．

参考文献

[1] Andrea Ernst-Gerlach and Norbert Fuhr. Generating search term variants for text collections with historic spellings. In *Lecture Notes in Computer Science*, Vol. 3936, pp. 49–60. Springer Berlin / Heidelberg, 2006.

[2] Andrea Ernst-Gerlach and Norbert Fuhr. Retrieval in text collections with historic spelling using linguistic and spelling variants. In *Proceedings of the 2007 conference on Digital libraries - JCDL '07*, pp. 333–333, New York, New York, USA, June 2007. ACM Press.

[3] David Grangier, Alessandro Vinciarelli, and Herv00E9 Bourlard. Information retrieval on noisy text. Technical Report Idiap-Com-08-2003, 2003.

[4] Trevor Strohman, Donald Metzler, Howard Turtle, and W. Bruce Croft. Indri: A language model-based search engine for complex queries, 2005. poster presentation.

[5] L. Venkata Subramaniam and IBM Research India. Noisy text analytics, 2010.

[6] Eri YAGI and Rika OKAMOTO. A history of entropy through various methods : Specially focused on technical term analysis. *Historia scientiarum. Second series : international journal of the History of Science Society of Japan*, Vol. 20, No. 1, pp. 47–56, 2010.