博士学位請求論文審査報告

報告番号　甲　乙　第　　号

氏名：岸田　和明

論文題目：Large-Scale Multilingual Document Clustering
（大規模な多言語文書クラスタリング）

論文審査担当者

主査	慶應義塾大学文学部教授 文学研究科委員	田村　俊作
副査	慶應義塾大学文学部教授 文学研究科委員	谷口　祥一
副査	国立情報学研究所教授	神門　典子
学識確認	慶應義塾大学文学部教授 文学研究科委員	田村　俊作

論文の要旨

異なる主題の文書が入り混じった集合を自動的に均質な部分集合に分割する文書クラスタリングは，ウェブ情報資源や電子文書の組織化に重要な役割を果たす。このため，その技術の研究および実用化が進んでいるが，分割対象となる集合が大規模で，なおかつ異なる言語で書かれた文書を含んでいる場合には，効果的かつ効率的な手法はこれまで見いだされてこなかった。本論文は，この大規模な多言語文書クラスタリングを実現するため，情報検索の分野で培われてきた言語横断検索の技術を活用し，さらに，大規模なクラスタリングが効率的に実行されるよう新たな改良を試みている。併せて，英語・ドイツ語・フランス語・イタリア語の新聞記事から成る一定規模の文書集合にそれらの手法を適用し，十分に効果的かつ効率的であることを，実験を通して実証している。

目　次

リング）
3.1 Introduction to Document Clustering Technique（文書クラスタリング法の概要）
3.2 Hierarchical Document Clustering（階層的文書クラスタリング）
3.3 Flat Partitioning（平坦な分割）
4. Probabilistic and Matrix-based Document Clustering（確率および行列に基づく文書クラスタリング）
4.1 Probabilistic Document Clustering（確率的文書クラスタリング）
4.2 Matrix-based Document Clustering（行列に基づく文書クラスタリング）
4.3 Graph-based Document Clustering（グラフに基づく文書クラスタリング）
5. Experiments of Large-Scale Multilingual Document Clustering（大規模な多言語文書クラスタリングの実験）
5.1 Scalable Multilingual Document Clustering Technique（スケーラブルな多言語文書クラスタリング技術）
5.2 Experiment of Term Disambiguation Techniques（語義曖昧性解消技術の実験）
5.3 Experiment of Monolingual Document Clustering（単一言語文書クラスタリングの実験）
5.4 Experiment of Multilingual Document Clustering（多言語文書クラスタリングの実験）
6. Conclusion（結論）

論文の内容

図書館・情報学分野では，長年に渡って，図書や雑誌論文などの文書に対する自動分類の研究が進められてきた。その中でも，分類体系の存在を事前に仮定しない状況において，主題的に多種多様な文書の集合を均質な部分集合に分割する場合，この自動分類は特に文書クラスタリング（document clustering）と呼ばれ，そのための技術は，ウェブ情報資源や電子文書が急増する中で，重要性を増しつつある。文書クラスタリングは，例えば，ある特定の技術に関する特許の最新動向を析出するための特許マップの自動作成や，あるテーマに関連したニュース記事群の自動要約を行う際にも，中心的な役割を果たすため，図書館・情報学以外でも，いわゆるテキストマイニングの中心的要素として，多くの関心が集まっている。

文書クラスタリングの問題の中で，未解決のまま残されてきたのが，本論文のテーマである多言語文書クラスタリング（multilingual document clustering）である。この場合には，英語やドイツ語，フランス語などの異なる言語で書かれた文書を含んだ集合が想定され，それを言語に関係なく，主題的に凝集した複数の部分集合に分割することになる。この文書集合自体が十分に小さければ，この問題の解決は容易であり，例えば，何らかの機械翻訳ソフトウェアと統計パッケージとを併用すれば，多言語文書クラスタリングを実行できる。つまり，すべての文書を単一の言語に翻訳できれば，この問題は，単一言語での文書クラスタリングに縮退し，また文書集合が小さければ，汎用的な統計パッケージによるクラスタリング処理が可能となる。しかしながら，機械翻訳とクラスタリングのアルゴリズムは，情報検索アルゴリズムとは異なり，スケーラブルではない。つまり，対象となるデータが大きくなれば指数関数的に計算量が

増え，コンピュータの処理能力をはるかに超えることになる。このため，「大規模な」多言語文書クラスタリングの実行は非常に難しい。

この問題を解決するために，本論文では，大規模な文書集合に対するクラスタリング手法に独自の工夫を加えるとともに，言語横断検索の技術の適用を試みている。言語横断検索は，例えば英語で書かれた文書の検索を日本語の質問文で行うような，異なる言語間での検索を意味し，このような検索を実現するために，汎用的な機械翻訳ソフトウェアに依存しない，情報検索の特性に合わせた独特の翻訳手法がこれまで種々開発されている。本論文は，その中でも，計算量が少なく，かつ品質の高い手法を選択し，結果的に，大規模な多言語文書クラスタリングの実行に成功している。

本論文は 6 つの章で構成されている。第 1 章で研究の背景や目的等を説明した後，第 2 章では，まず，標準的な情報検索技術を概観している。これは，文書クラスタリングが本来的に情報検索の技術を基盤としているためであり，具体的には，ブール検索，自動索引法，ベクトル空間モデル，確率的検索モデル，質問拡張，テキスト処理，適合性の概念，検索実験の手法が順に論じられている。この中でも，文書中に含まれる語の重みを算出するための自動索引法や，文書間の類似度を測定するためのベクトル空間モデル，文書のテキストから語を自動抽出するためのテキスト処理は，これ以降，本論文の中で繰り返し取り上げられている。

第 2 章の後半部分では，言語横断検索の技術を網羅的に展望している。具体的には，異なる言語の語の間を関連付けるための照合技法，語義曖昧性解消技術，言語横断検索のための確率モデルなどが順に論じられている。この中で，「対訳辞書に基づく文書翻訳」と呼ばれる手法を多言語文書クラスタリングに採用すべきことが結論される。この手法は，文書集合が大規模になってもそれほど急激に計算量が増えない「準スケーラブル」な手法であり，この点で，他の手法よりも，本論文の目的に適している。ただし，対訳辞書に基づく手法では，語の意味の曖昧性（多義性）の解消が欠かせない。なぜなら通常，辞書には，各語のもつ様々な意味が網羅的に掲載されており，その語を使用しているテキストの文脈に応じて，その中から適切なものを選択しなければならないからである。この手法についても様々なものが開発されているが，実用性と，大規模なテキストへの適用可能性の観点からは，語の共起頻度に基づく手法を採用すべきことが，ここでの議論から導かれている。

第 3 章と第 4 章では，文書クラスタリングの技術を包括的に論じている。文書クラスタリングの技術は，テキストマイニングに関心が集まるようになった 2000 年代以降，コンピュータサイエンスや統計学の分野で盛んに研究され，現在では，その理論や手法は非常に多岐に渡っている。それに対して，これらの技術の全体を詳細に論じた図書や論文は未だ出版されておらず，その細部に至るまで，慎重に吟味する必要がある。そこで本論文では，それらを大きく 2 つに分け，第 3 章において，主として図書館・情報学分野および情報検索分野で開発された諸技術を論じ，第 4 章では，2000 年代以降急速に発展した確率論や線形代数論に基づく手法を俯瞰している。この 2 つの章では，理論やアルゴリズムの単なる整理・体系化に留まらず，小規模の事例を用いて各アルゴリズムを実際にコンピュータで動作させ，実証的な観点からも，その特徴を比較し議論している。

第 3 章ではまず，文書クラスタリングの基本的な事項として，文書ベクトルの構成法，類似

度の計算法，基本的なクラスタリング法，素性選択，クラスタリング結果の評価法等を整理した上で，階層的クラスタリング法，k-means 法（Hartigan-Wong 法，球面法など），leader-follower 法，自己組織化マップ，ファジイクラスタリング等を順に詳しく論じている。文書中には一般に数多くの語が含まれ，その主題表現としての文書ベクトルは高次元となるため，通常のクラスタリング法はうまく動作しないことが多い。それに対して，情報検索分野で伝統的に用いられてきた leader-follower 法は，情報検索のベクトル空間モデルで標準的に使用される余弦係数に基づいて文書集合を分割するため，先の問題は回避でき，効率と効果の両方の点で優れている。本章では，各クラスタリング法のアルゴリズムを精査し，leader-follower 法の利点を明確にするとともに，他の手法の中では，特に文書集合向けに工夫された球面 k-means 法が，反復計算を必要とするものの，大規模な文書クラスタリングに適用可能なことを示唆している。

一方，第 4 章では，確率論に基づく手法として，確率的混合モデル，確率的潜在意味分析（PLSA)，潜在的ディリクレ配分法（LDA)，階層的ディリクレ過程（HDP）混合モデルなどが取り上げられている。また，行列の分解に基づく手法としては，潜在的意味索引法（LSI)，非負行列分解（NMF）等を議論し，加えて，グラフ理論に基づく手法として，スペクトラルクラスタリングについても言及している。確率的混合モデルの場合には，1 件の文書が，複数の確率分布の混合から生成されると仮定する。当該文書の生成に最も寄与した分布をその文書が帰属するクラスタと見なせば，文書クラスタリングが可能になるが，そのためには，帰属の程度をデータから推定する必要があり，通常，EM アルゴリズムあるいはギブスサンプリングが使用される。この章では，これらの統計的推定法を詳細に論じ，文書クラスタリングへの適用可能性を検討している。その結果，これらの混合モデルは，文書ベクトルが高次元であるため，実際には，大規模な文書クラスタリングに応用する場合に困難となることを確認している。

PLSA や LDA では，混合モデルとはやや異なり，主題に関する潜在的なクラスを想定し，それらの潜在クラスから，語や文書が生成されると仮定する。この潜在クラスを 1 つのクラスタとして捉えれば，ある特定の文書に対する潜在クラスの確率を使って文書クラスタリングを実行できる。さらに，HDP 混合モデルは，無限個の確率過程の混合を仮定することで LDA を拡張したものであり，「無限個」の混合が，実際のデータに応じて有限個に収束した場合，これはクラスタ個数をも同時に推定したことを意味する。この点，HDP 混合モデルは非常に有用である。

第 4 章の後半部分で議論される行列に基づく手法では，「語×文書」の重み行列を分解することにより，クラスタリングを実行する。図書館・情報学分野では伝統的に，特異値分解に基づく LSI を用いることが主流であったが，NMF のための簡便なアルゴリズムが開発されて以来，一般的には，NMF の利用が増加している。ただし，この分解は一種の近似であり，この近似解を得るために，行列の積の演算を一定回数繰り返す必要がある。確率に基づく手法と行列に基づく手法とに共通している点は，各パラメータに無作為に初期値を割り当てた上で，反復計算に基づいてクラスタの集合を求めることであり，第 4 章では，実際にコンピュータ上で小規模な実験を繰り返し，初期値の割り当ての影響や反復計算の回数の評価を試みつつ，大規模文書クラスタリングへの適用可能性を探っている。

第 2 章から第 4 章までの議論により，大規模な多言語文書クラスタリングを実行するには，計算量の点からは，翻訳手法として「対訳辞書による文書翻訳」および「語の共起頻度に基づく曖昧性解消法」，クラスタリング法としては「leader-follower 法」を採用することが示唆されたことになる。球面 k-means 法や HDP 混合モデルも，クラスタリングアルゴリズムとして魅力的ではあるが，一定回数の反復計算が必要な点で，やや適性を欠く。第 5 章では，これらの手法の有効性を実証的に確認するための 3 つの実験結果を報告している。第 1 の実験では，言語横断検索における自動的な翻訳手法の比較を試みている。語の共起関係に基づく曖昧性解消法としては， Best pair 法，Best sequence 法，Best cohesion 法の 3 つのバリエーションが存在するが，それらの効果と効率を，検索実験用テストコレクション CLEF を使った言語横断検索実験（英語・ドイツ語・フランス語の 3 言語による検索質問からイタリア語の記事集合約 15.7 万件への検索）により確認している。理論的には 3 つの手法のうち，Best sequence 法が曖昧性解消の正確性の点で望ましいが，計算量が非常に多く，本論文の目的からは不適切である。そのため，Best pair 法か Best cohesion 法を使わざるを得ないが，実験の結果，Best cohesion 法が，比較的計算量が少ないにもかかわらず，曖昧性解消の効果がさほど落ちないことが明らかになった。この結果から，本論文の多言語文書クラスタリングでは Best cohesion 法を活用することが示唆された。

第 2 の実験の目的は，単一言語での大規模文書集合に対するクラスタリング法の効果と効率とを実証的に分析することにある。まず，leader-follower 法の性能を明らかにするために，単一言語（英語）の文書集合（1996 年 8 月の Reuters の英語記事 6,374 件）に対して，leader-follower 法に加えて，反復計算に基づく球面 k-means 法，HDP 混合モデルの 3 つの手法を適用し，その効率と効果とを比較評価した。その結果，予想通り，実験上でも leader-follower 法の計算量が格段に少ないことが示されたのに加えて，その効果（妥当なクラスタ集合を出力する程度）も，他の 2 つの手法に比べて遜色ないことが明らかになった。これは，計算量が少ないにもかかわらず，妥当なクラスタ集合を出力するという費用対効果の点での leader-follower 法の優位性が実証されたことを意味する。次に，より大規模な文書集合への leader-follower 法の適用可能性を調べるために，Reuters の英語記事を順次追加して，単一言語の文書クラスタリングを再度試みた。その結果，実行中に生成されるクラスタの主題表現ベクトルの肥大化を抑制するという独自の工夫を加えることにより，記事数が 7 万件を超えても，leader-follower 法は安定的に妥当なクラスタ集合を生成することが明らかとなった。文書クラスタリングにとっては「7 万件」の集合はかなり大きく，この実験結果は，本論文で提案する多言語文書クラスタリングの手法が，大規模な文書集合に対しても実行可能であることを示唆している。

最後の第 3 の実験では，最終的に，第 1 の実験の結果から採用された翻訳手法と，第 2 の実験の結果から選択されたクラスタリング法とを組み合わせて，1.3 万件以上の Reuters の新聞記事から構成される多言語（英語・ドイツ語・フランス語・イタリア語の 4 言語）文書集合のクラスタリングを試みている。より具体的には，この実験では，(a)多言語文書クラスタリングの性能の確認，(b)語義曖昧性解消（Best cohesion 法）の効果の分析，(c)素性選択の効果の分析（高い重みをもつ一定数の語を抽出し，それを用いてのクラスタリングの実行），(d)文書翻

訳とクラスタ翻訳との性能比較がなされ，それぞれの結果が分析されている。ここで，「クラスタ翻訳」とは，第 1 段階で言語別にてクラスタリングを実施し，その結果として得られたクラスタを翻訳する方法であり，翻訳されたクラスタに対して，再度，クラスタリングを実行し（第 2 段階），最終的なクラスタ集合を生成する。クラスタリングを 2 段階に分ける点で計算量が多くなるが，仮に第 1 段階で均質かつ少数のクラスタが生成されれば，クラスタを翻訳する計算量は，文書を個別に翻訳する場合に比べて減少することが期待できる。また，上記の(b)と(c)については，パラメータをいくつか変えて実験を繰り返し，その結果得られた評価指標の値を「データ」と見なして分散分析を試みることにより，各要因がクラスタリングの効果に影響するかどうかを調べている。これらの実験の結果，(1)文書を直接，対訳辞書を用いて翻訳する「文書翻訳」の場合には，多言語文書クラスタリングが有意に成功すること，それに対して，(2)クラスタ翻訳の場合には，有効な結果をもたらさないこと，また(3)文書翻訳の場合には，語義曖昧性解消（Best cohesion 法）はクラスタリングの性能をある程度向上させること，(4)同じく文書翻訳の場合には，素性選択で選ばれる語数が多いほど，クラスタリングの性能がほぼ確実に向上することが明らかになった。

この実験結果を受けて第 6 章で全体的な結論が述べられている。以上，本論文は，言語横断検索技術を活用した多言語文書クラスタリング法を考案し，いくつかの実験を通して，この手法が効果および効率の両方の点で優れており，きわめて実用性の高いものであることを実証的に示している。

審査結果

図書や新聞，雑誌記事，学術論文などの公的な出版物に対する書誌コントロールのために，メタデータを作成することは，図書館員などの情報専門家が果たすべき重要な使命である。それに対して，出版過程を経由せずに公開・利用されるウェブページや，機関・組織内で頻繁に作成し更新される電子文書を，その種の情報専門家がメタデータの作成を通じて管理する状況は考えにくい。

例えば，ある時事的な話題についてインターネット上の書き込みを収集・類別し，その傾向を知ろうとするような場合，自動化された何らかの手段を用いることが不可欠である。さらに，時事的な話題に対しては，既存の分類は不適切なため，話題に合わせた適切な分類を自動的に作成・適用すること，すなわち文書クラスタリングを行うことが求められる。また，多言語に渡る書き込みを一覧するには，まず言語に関わりなく検索・類別した上で，関連するものが閲覧できるようになっていれば便利である。本論文が試みている多言語文書クラスタリングは，このように有用性の極めて高い技術である。

しかしながら，その有用性にもかかわらず，難度の高さから，小規模なものを除き，多言語文書クラスタリングはほとんど研究されてこなかった。本論文は，大規模な文書集合に対して一定の性能を示す手法を考案した世界で初めての試みである。

多言語文書クラスタリングが困難な原因は，本論文でも述べられているとおり，機械翻訳のアルゴリズムと文書クラスタリングのアルゴリズムの両者がともに，スケーラブルではなく，対象文書およびその集合が大規模になるにつれて計算量が爆発的に増えてしまう点にある。例

えば情報検索アルゴリズムの場合には，インターネットの検索エンジンで実践されているように，並列処理を効果的に適用できるため，大規模文書集合に対する処置が可能であるが，機械翻訳や文書クラスタリングにおいては，その種の処置は難しい。

本論文では，クラスタリング法が大規模文書集合に対して安定した結果をもたらすように改良を加えるとともに，言語横断検索における翻訳手法を援用することでこの問題の解決を図っている。これらの文書クラスタリング技術と言語横断検索技術とは，それぞれ別個に発展したものであり，なおかつそれぞれ特有の難しさを持っている。これらの技術を扱っている研究領域も基本的には異なっており，この 2 つの組み合わせに着目したこと自体に，本論文の高い独創性が認められる。

異なる領域の手法を組み合わせるという課題に対し，本論文ではまず，両領域における研究の徹底的なレビューにより，適切な手法の同定を行っている。特に，第 3 章と第 4 章における文書クラスタリングのアルゴリズムのレビューは，単なる研究動向の把握に留まらず，小規模な実験により各手法の性能の比較評価にまで踏み込んでおり，それ自体に高い独創性が認められる。テキストマイニングに多くの関心が集まった結果，さまざまな分野で結果的に文書クラスタリングに関係する研究が試みられ，数多くの手法が提案されている。その多くは，統計学的に複雑で，これまでに出されたいくつかの展望論文はその詳細に踏み込むことができず，それらの理論・技術の詳細を分野横断的に理解することは難しい状況にあった。本論文はその手法の多くを網羅しつつ，一貫した視点からそれぞれの仕組みを体系的に整理している。さらには, そのうちの主なものに対して独自にプログラムを作成し, 小規模な人工的データを用いて，実際の「動き方」を確認している。これにより，局所解への収束などの，各手法の原文献では報告されていない問題の所在が詳らかになるとともに，各手法の性能を比較評価することが可能となった。この種の横断的な比較評価は，文書クラスタリングの領域ではこれまで十分に行われてこなかったものである。もちろん，対象が小規模な人工的データであることなどから，本論文の比較評価だけでは確定的な結論は導き得ないものの，逆に，対象データの単純さは，数学的に複雑な各手法の仕組みの理解にはむしろ貢献している。

第 2 章の検索技術に関するレビューも同様に網羅的なものであり，以上のようなしっかりした基盤の上に，本論文の中核となる第 5 章の諸実験は構築されている。それらの実験もまた，丁寧に計画され，信頼性が高い。5.2 節における，言語横断検索技術として「語の共起関係に基づく曖昧性解消」を用いる手続きも，5.3 節における単一言語での文書クラスタリングも，実装の難しい技術や大規模な文書集合を用いたことなど，いずれも容易ではない課題に答えるものであり，それぞれが単独でも十分に評価すべき成果である。実際,「語の共起関係に基づく曖昧性解消」の実験と「単一言語文書クラスタリング」の実験とは，それぞれ別個に国際的な査読付き雑誌に掲載された原著論文に基づいたものである。

これらを総合的に組み合わせた多言語文書クラスタリング法が本論文の主たるテーマであるが，これに関する実験結果（5.4 節）も同様に，国際的な査読付き雑誌にすでに原著論文として掲載され，一定の評価を受けている。多言語文書クラスタリングの困難さから，4 言語に渡る一定規模の多言語文書集合に対してクラスタリングが成功した例はこれまでになく，この原著論文が世界で最初の成果である。この点で，本論文は，多言語文書クラスタリング研究の今

後の発展にとってメルクマールとなる成果と捉えることが可能であろう。

本論文の不十分な点を敢えて指摘するならば，実験が Reuters の記事集合のみでなされている問題がある。例えば，leader-follower 法は，反復計算に基づく手法に比べて，そのクラスタリング結果が，処理される文書の順序により大きな影響を受ける可能性がある。本論文の実験に用いた新聞記事は，順序が時間順に確定している文書群であるため大きな問題とはならないが，そうした順序をもたない文書集合の場合に，leader-follower 法がどの程度の性能を示すのかには不確定さが残る。例えば情報検索の分野では，複数の異なる文書集合に対して実験を繰り返すことによって，より信頼性のある結果を導くことが常識となっている。多言語文書クラスタリングの実験に使用可能でかつ良質な文書集合として，Reuters を用いることは妥当ではあるものの，今後はさらに研究を進め，複数の異なる文書集合に対して実験を繰り返すことにより，さらに手法を改良することが望まれる。

本論文は，すでに述べた 3 編の国際的な査読付き雑誌に掲載された論文に加えて，査読付き雑誌に掲載されたその他の多数の関連論文の諸成果をまとめた 1 つの集大成である。なお課題は残るものの，その独創性や，結果の妥当性・信頼性の点で高く評価されるべき論文であり，審査員一同，博士（図書館・情報学）の学位にふさわしいものと判断する。