主　論　文　要　旨

報告番号	甲　乙　第　　号	氏　名	岸田　和明

主 論 文 題 名：

Large-Scale Multilingual Document Clustering

（大規模な多言語文書クラスタリング）

（内容の要旨）

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（または文献）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同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，インターネット上の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。

例えば，インターネットの検索エンジンに，複数の意味をもつ多義的な語が投入された場合，その検索結果には，異なる主題を論じた様々なウェブページが含まれることになる。この際に，このウェブページ集合に対して文書クラスタリングを実行すれば，それらを同一主題に関するいくつかのグループに分割することが可能で，利用者はその中から自分の意図した主題に対応するグループのみを選択することによって，不適切なウェブページの閲覧を避けることができる（実際に，この種のシステムは「クラスタリング検索エンジン」として実用化されている）。さらには，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，文書クラスタリングは中心的な役割を果たす。

従来的な図書や新聞・雑誌記事などの公的な出版物（publication）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（information specialists）が果たすべき重要な使命である。それに対して，公的な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される電子文書に対して，その種の情報専門家がメタデータを作成して管理する状況は考えにくい。このような情報資源を効果的かつ効率的に組織化するには，文書クラスタリングをはじめとする自動的な手法の援用が必須であり，このため，上で述べたように，文書クラスタリングに関する研究が精力的に進められている。また，図書館・情報学以外にも，いわゆるテキストマイニングの中核技術として，文書クラスタリングに多くの関心が集まっている。

文書クラスタリングの技術的な問題の中で，未解決のまま残されているのが，本論文のテーマ

「多言語文書クラスタリング」（multilingual document clustering）である。この場合には，英語や独語，仏語などの複数の言語で書かれた文書が雑多に集まった集合が想定され，それを言語に関係なく，主題的に均質な部分集合に分割することになる。もちろん，この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る。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（例：英語）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するため，文書集合が小さければ，汎用的な統計パッケージで処理することが可能である。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く，本論文が対象とする文書集合の規模では，一つのシステムとして成功した例は，これまでない。

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技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る。この 2 つを組み合わせることは，おそらく世界的に初めての試みである。後者の「言語横断検索」とは，例えば英語文献の検索を日本語で書かれた質問文で行うことを意味し，この検索の実行のために，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技法が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た。

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の説明をした後，第 2 章では，情報検索の標準的な技術と言語横断検索の技術を幅広く展望する。さらに，第 3 章と第 4 章では，文書クラスタリングの技法を，「伝統的な方法」（第 3 章）と「確率・行列に基づく方法」（第 4 章）とに分け，それぞれ詳しく論じる。第 5 章では，第 2～4 章の議論の結果として，新たな多言語文書クラスタリング法が提案され，その効果と効率を実証的に明らかにするための 3 つの実験が報告される。第 6 章は結論である。以下，第 2～5 章の要旨を順に述べる。

第 2 章では，まず，標準的な情報検索技術を概観する（第 1 節）。これは，文書クラスタリングが本来，情報検索の技術を基盤としているためであり，この技術の要点を整理することは，文書クラスタリングの問題を論じる際には欠かすことはできない。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，検索実験の方法を順に論じていく。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，次節以降，本論文の中で繰り返し取り上げられることになる。

続いて第 2 章の第 2 節では，言語横断検索の技法を網羅的に展望する。すでに述べたように，言語横断検索の技術により，汎用的な機械翻訳ソフトウェアに頼らない，情報検索の目的に適した「テキストの翻訳」が提供される。本論文の目的に沿って，「大規模」な文書集合に適用可能な翻訳技法を特定することがこの節の目的である。具体的には，言語横断検索における照合技法（異なる言語の語の間を関連付ける技法），翻訳技法，曖昧性解消法，言語横断検索のための確率モデルなどを順に論じる。この中で，一般に「対訳辞書に基づく文書翻訳」と呼ばれる方法を本論文は採用することになる。この方法は，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，他の方法よりも，本論文の目的に適している。

ただし，対訳辞書に基づく方法では，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，言語横断検索において，多数のノイズ（不適合文書）が出力されることになるからである。この方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点からは，語の共起頻度に基づく方法を採用すべきことが，この節の議論から導かれる。例えば，「Mercury」という語が文書中で「planet」と共起する傾向がある場合，それは神話の「マーキュリー」などの意味ではなく，「水星」を指示するために用いられていることが推察される。この種の共起関係を利用した曖昧性解消法を試してみることには十分な価値が認められ，多言語文書クラスタリングに対してそれが有効かどうかを実験的に確かめる必要がある（第 5 章の第 1 節で，実際にこのための実験を試みる）。

次に，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るようになったため，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ま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技術を俯瞰する。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(Java 言語によりソースコードを作成し，実装），実証的な観点から，その特徴を比較・議論する。特に，2000 年代以降の文書クラスタリング技術には，理論的・数学的に複雑なものが多く，本論文における実証比較は，これらの技術に対する理解を深めるものである。

具体的には，第 3 章では，第 1 節にて，文書クラスタリングの基本的な事項（文書ベクトルの構成方法，類似度の計算の方法，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方法）を整理した上で，第 2 節以降，階層的クラスタ分析（単連結法，完全連結法，群平均法，Ward 法），k-means 法（基本的な方法，Hartigan-Wong 法，インクリメンタルな方法，球面法など），SKWIC 法，leader-follower 法，自己組織化マップ（SOM），ファジイクラスタリングなどを順に詳しく論じていく。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。その点，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，この点で，効率と効果の両方の点で優れている。通常のクラスタリング法では，標準的なユークリッド距離が用いられるが，データが高次元のベクトルの場合には，余弦係数のほうが計算の効率がよく，また，ベクトル間の類似度を容易に測定することができる。しかも，leader-follower 法の場合，文書集合が収められたファイルを 1 回ないし 2 回走査するだけで結果が得られるため，対象となる文書数が膨大な状況でも適用しやすく，目的関数を最適化するための反復計算が必要な他の方法に比べて有利である。すなわち，leader-follower 法では，文書ベクトルをファイルから順次読み込み，その時点で存在しているクラスタのうち，最も類似したものに割り当てていく（十分に類似したクラスタが存在しない場合には，当該文書が新たなクラスタとして登録される）。一方，反復計算が必要な方法の中では，特に大規模な文書集合向けに工夫された球面（spherical）k-means 法が，本論文の目的には適している。この方法は，通常，生成されたクラスタ内の「密度」の最大化を目的として，その最適解を反復計算により近似的に求めるために，leader-follower 法よりも文書ファイルを数多く走査する必要があるものの，基本的には余弦係数に基づくクラスタリング法であり，この点，文書クラスタリングに適している。

一方，第 4 章では，確率的な方法として，確率的混合モデル（ポアソン混合，多項混合，von Mises-Fisher 分布の混合），ベイズ型多項混合モデル，確率的潜在意味分析（PLSA)，latent Dirichlet allocation（LDA），階層的ディリクレ過程（HDP）混合モデルを取り上げる，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），

主成分分析，IRR アルゴリズム，COV-rescale アルゴリズム，非負行列分解（NMF），非負テンソル分解（NTF）を議論し，加えて，グラフに基づく手法として，スペクトルクラスタリングについても論じる(この方法は最終的には,行列の計算に帰着するため)。各手法はそれぞれ比較的複雑な確率あるいは行列の計算に依拠しており，その詳細を明確にすることを目的として，この章では，それらを統一的な数学的な記号法の下で体系的に整理する（もちろん，それ以前の章との整合性も考慮する）。

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。この確率分布としては，情報検索では伝統的にポアソン分布が使われてきたが（ポアソン混合），むしろ，機械学習の分野では，多項分布が用いられることが多い（多項混合）。いずれの場合にも，パラメータの異なるそれぞれの分布に混合係数を掛け，それらを足し合わせたモデルが想定されるが，それらのパラメータをデータから推計することにより，各文書が主としてどの分布に帰属するのかを判別できる。これによって，文書クラスタリングが可能になるが，そのためには当然，パラメータの推定が必要であり，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。第 4 章第 1 節ではこれらの推計方法を詳細に論じ，文書クラスタリングへの適用可能性を明らかにする。なお，混合モデルの構成要素となる確率分布としては，文書クラスタリングを応用目的とした場合，von Mises-Fisher 分布が適している。さらには，ベイズ推定の考え方を使って拡張されたベイズ型多項混合モデルも，局所的な最適解への到達を避けるという点で，優れた方法である。しかしながら，これらの混合モデルは，文書ベクトルが高次元であることから，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が生じる。

PLSA では，混合モデルとはやや異なり，主題（トピック）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（その推計には EM アルゴリズムが使用される）。一方，LDA は PLSA をベイズ推定の考え方により拡張した手法であり，さらに，HDP 混合モデルでは，無限個の確率過程の混合を仮定することで LDA を拡張している。実際に，「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しており，この点，HDP 混合モデルは非常に有用な手法である。

第 4 章の第 2 節で議論される行列に基づく方法では，「語×文書」または「文書×語」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。伝統的には，特異値分解に基づく LSI 法を用いることが情報検索分野では主流であったが，NMF のための便

利なアルゴリズムが開発されて以来，一般には，NMF が使われることが増えている。NMF では，例えば「語×文書」行列を，クラスタ個数を指定して分解することにより得られる「文書×クラスタ」の行列中の値を使って文書クラスタリングを行うことができる。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。確率に基づく方法と行列に基づく方法に共通している点は，通常，パラメータに無作為に初期値を割り当てた上での反復計算に基づいてクラスタの集合を求めることである。このため，反復的な計算が必要であり，初期値の割り当てによっては，局所的な最適解に収束してしまう。第 4 章では，すでに述べたように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ながら，大規模文書クラスタリングへの適用可能性を探る。

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，翻訳技法として「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」，クラスタリング技法としては「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，反復計算が必要な点で，やや適性を欠く。第 5 章では，実際にこれらの方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する。第 1 の実験では，言語横断検索における自動的な翻訳技法の実証的な比較を試みている。語の共起関係に基づく曖昧性解消法としては，具体的には，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・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・仏・独」の 3 言語から伊語の文書集合約 15 万 7 千件への検索）により実証的に確認することがこの実験の目的である。理論的には 3 つの方法のうち，Best sequence 法が曖昧性解消の正確性の点では望ましいが，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，実験の結果として，Best cohesion 法が，比較的計算量が少ないにもかかわらず，曖昧性解消の正確さがそれほど落ちないことが実証的に示された。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る。

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，球面 k-means 法，HDP 混合モデルの 3 つの方法を適用し，その効率と効果とを比較評価した（ここでの Reuters の記事集合は，この種の評価が可能なように編纂され，多くの研究者に使用されているテストコレクションである）。

比較対象である球面 k-means 法と HDP 混合モデルは，第 3 章と第 4 章で議論された反復計算に基づく「より洗練された方法」のうち，大規模な文書集合に適用可能なアルゴリズムとしてこの実験において採用されたものである。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，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（ただし，オリジナルの方法に若干の独自の改良を加えている）。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方法が，ここでの実験で用いられた文書集合よりも大規模なデータ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技法と，第 2 の実験の結果から選択されたクラスタリング技法とを組み合わせて，1 万件以上の Reuters の新聞記事から構成される多言語文書集合（英・独・仏・伊の 4 言語）のクラスタリングを試みる。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)語の曖昧性解消（Best cohesion 法）の効果の分析，(c)素性選択の効果の分析（語の重みが上位 x 語を各文書から抽出してのクラスタリングの実行，ここで，x=10,20,30,50,100），(d)文書翻訳とクラスタ翻訳との比較がなされ，それぞれの結果が分析された。ここで，「クラスタ翻訳」とは，第 1 段階で言語別で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，もし第 1 段階で均一かつ少数のクラスタが生成されれば，クラスタの翻訳の計算量は，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される。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た。実験の結果，(1)文書を直接翻訳する場合には，多言語クラスタリングが成功する（翻訳を省略し，言語別でのクラスタリング結果を単純に併合したものよりも，評価指標の値が高くなる）こと，(2)それに対して，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，(3)文書を直接翻訳する場合には，語の曖昧性解消（Best cohesion

法）はクラスタリングの結果の質をある程度向上させること，(4)文書を直接翻訳する場合には，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となった。

以上，本論文は，言語横断検索技法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じて，この方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た。