様式 C-19

科学研究費補助金研究成果報告書

平成 21 年 4 月 1 日現在

研究種目：基盤研究(B)
研究期間：2005～2008
課題番号：17300071
研究課題名（和文）　生物分類データベースを対象としたラフ集合理論に基づく概念体系比較に関する研究
研究課題名（英文）　Comparison of taxon concept hierarchies in biodiversity databases using Rough set theory
研究代表者
伊藤　希（YTOW NOZOMI）
筑波大学・大学院生命環境科学研究科・講師
研究者番号：90251016

研究成果の概要：ラフ集合理論を生物分類データの解析に応用し、データの不完全さ・不均質さを許容しつつ学名データ全体から浮かび上がる分類体系を抽出する手法を開発した。また、ラフ集合理論と形式概念分析との関係性に着目し、それぞれの生物分類群が有する性質に基づいて分類を行なうことと、それぞれの分類群に含まれない生物を列挙することが理論的に同等であることを示し、ラフ集合理論による分析の生物分類学的な位置づけを明らかにした。

交付額

（金額単位：円）

	直接経費	間接経費	合　計
2005 年度	3,300,000	0	3,300,000
2006 年度	1,800,000	0	1,800,000
2007 年度	1,800,000	540,000	2,340,000
2008 年度	1,800,000	540,000	2,340,000
年度			
総　計	8,700,000	1,080,000	9,780,000

研究分野：生物多様性情報学
科研費の分科・細目：情報学　感性情報学・ソフトコンピューティング
キーワード：データベース、ラフ集合、形式概念分析、オントロジー、GBIF

１．研究開始当初の背景

地球上のどこに、どの様な生物が、どれだけ存在しているか、という情報のデータベース化が国際連携により行なわれている。そのために必要となる基礎情報のひとつに生物分類すなわち生物名とその性質との対応づけがあるが、リンネ以来 250 年を超える知見の蓄積がある一方、最新の分類研究による新規知見の追加とそれに伴う既存分類体系の見直しも行われている。こうした情報は博物館などの研究組織や研究者個人によってそれぞれの目的に応じて集積され、徐々に公開されつつあるが、分散して行われているが故に相互の整合性に関しては検証されていない。それゆえ、すべてを包括的に含みながら同時に整合した分類体系を提供するデータは存在していない。データベースに蓄積提供されつつあるこうした断片的生物分類情報を有効利用するためには、分散した情報源を統合的に解析し、内在する矛盾点の指摘も含めて提示する手法が必要とされる。この際には、これまでに提案されたさまざまな分類体系について相互比較が必要となるが、新種発見や遺伝子配列の様な新手法の導入による新知見の追加はそもそも分類体系のフレームワークを変更してしまうため直接的な単純比較した結果の解釈は困難をともなう。フレームワークを変更せざるを得ない様な状

況下にあっても比較検討を可能とする手法として、不完全情報の扱いに優れたラフ集合理論の応用について検討を行なった。研究開始時点では生物分類概念データ交換方式の標準化が TDWG (Tanoxmic Database Working Group, http://www.tdwg.org) で行なわれており、制定される交換方式により分散データベースアクセスを行うことを想定して研究計画を立案した。

２．研究の目的

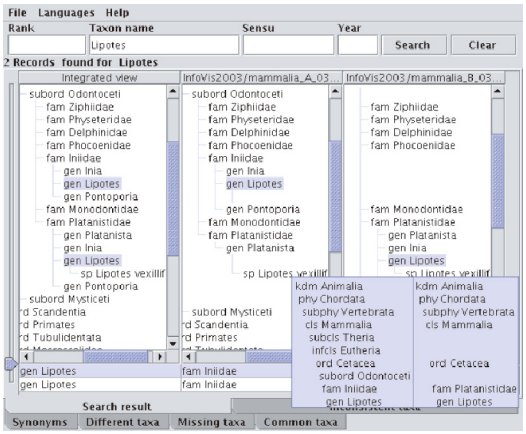
分散配置された生物分類データベースに蓄積された情報から整合性を持った分類体系群を抽出するには、分類体系の異同ならびに整合性を判定するための理論的枠組み、実際にそうしたデータベースにアクセスする手法、ならびに、得られたデータについて整合性判定を行った結果を評価する手法が必要となる。この三つの技術を開発し、実際の生物分類データベースへアクセスして理論の検証を行なうことを目的とした。

３．研究の方法

生物分類データベースにラフ集合理論を適用するため、生物分類学のモデル化を行った。このモデルに基づき、ラフ集合の適用による分類体系の比較検討について理論的側面から検討した。また、その意味づけを行うために、形式概念分析との比較を行った。ラフ集合論の基本となるのは対象とそれが有する性質のマトリクスであり、性質により同値類を導入して粗視化を行なうが、生物分類データベースで得られる「性質」に相当する情報すなわち形質情報は分類群によってまちまちである一方、分類群名と分類体系ほど広く得られるわけではない。また、データの整合性という観点では分類群名と分類体系の方が形質情報よりも深刻であるので、分類群名をその分類体系における同値類に与えられたラベルと看做して粗視化はすでに行われているものとして扱った。

TDWG 標準である分類概念スキーマ (Taxon Concept Schema, TCS) はデータベース間でのデータ交換を想定した XML スキーマであり、データの表現方法は規定しているがデータ転送プロトコルを指定してはおらず、さまざまなデータ転送プロトコルと組み合わせて利用できる。候補となるデータ転送プロトコルのうち、生物多様性情報分野での一般的利用を想定して開発されたプロトコルである TAIPR (TDWG Access Protocol for Information Retrieval) は、TDWG での標準化を経て広く使われる様になると予想された。そこで、生物分類データベースへのアクセス手法として TAPIR クライアントライブラリをプログラミング言語 Java により開発し、別途検討したラフ集合理論を用いたプロトタイプツールで利用可能とした。また、分散データベースへのアクセスを実現すべく、UDDI を利用して GBIF (Global Biodiversity Information Facility, 地球規模生物多様性情報機構) レジストリよりデータ提供者情報を取得するメカニズムを Java で実装し、Java アプリケーションソフトウェアでの利用を可能とした。

TCS は研究期間の早い時期に TDWG 標準となったが、複数のデータベース間での実際のデータ交換に使われるには研究期間中には至らなかった。そこで、学名の用例を分類体系を問わず蓄積している公開データベース uBio (http://www.ubio.org) をデーターソースとして利用することとし、uBio へのアクセスライブラリを Java により実装した。このライブラリを用いて uBio にアクセスし、ラフ集合理論を実際のデータに適用する際の問題点を検討した。



ユーザインタフェース画面例

４．研究成果

比較的単純な状況として、ある属に新種が追加された場合を想定する。生物種をラフ集合における同値類としてモデル化すると、新種の記載前後で当該属概念の負集合すなわち当該属に含まれない分類群名の集合は一致し、正集合すなわち当該属に含まれなければならない分類群名の集合は新種名だけ異なることになる。新種記載後の観点から新種記載前の体系をラフ集合論的に記述すれば、新種は記載前の体系において境界領域に存在していたことになる。一方、ある属のある種が別の属に再分類された場合を考えると、再分類前後で正集合も負集合も異なっている。新種追加というのは当該種が当該属の性質を満足するという判断に基づいておこなわれるものであるから、記載前後で当該属概念が異なると想定するのは合理性を欠く。再分類によりある種が一方の分類体系ではある

属に含まれ、他方の分類体系では別の属に含まれるのであるから、この二つの分類体系における当該分類群は異なる概念であると考えるべきである。従って、分類概念の整合不整合は、新規分類群が追加された場合であっても、比較すべき分類概念対の一方の正集合と他方の負集合を比較し、共通部分があれば不整合であると判断できる。この際の共通部分の判断は、その分類体系において当該分類概念に含まれる全ての分類群名と排除される全ての分類群名との比較によるが、全体としては比較対象となる分類体系に共通する分類群名についてのみ比較すればよいことになる。このことは、一方の分類体系でたとえば属と種の間に亜属が設けられた場合であっても、その亜属名は一方に分類体系にしか登場しないため、当該亜属の有無が属の整合性判定には影響しないことを意味する。このように、ラフ集合理論の正集合と負集合を用いることで、分類体系の直接的影響を受けることなく分類概念の整合不整合を判定できることがわかった。比較対象の分類体系互いに不整合となる分類概念が含まれている場合、その分類体系もまた互いに不整合であるので、そうした場合についてはラフ集合理論により検出可能である。概念不整合が全く存在しない分類体系の対には、トリビアルな場合として共通部分が全くない、独立な分類体系が含まれる。それ以外の、何らかの共通部分を有する、整合な概念からのみ構成される分類体系について、更なる検討を行なった。なお、不整合な概念対を含む分類体系対から不整合部分を除いた残りの部分も同様に扱うことが可能である。

ラフ集合理論による概念比較では、たとえば亜属の有無といった差異は整合性判定に影響しない。このことは、分類体系を構成する概念をラフ集合理論により比較しただけでは、体系を構成する概念に不整合がない体系間では概念体系の比較ができない場合があることを意味する。一方、ある分類概念の、分類体系における上位分類群の連鎖、たとえば、ヒトから哺乳類、脊椎動物を経て生物全体に至る包含関係は、対象とする分類概念の正集合、負集合、境界領域のいずれにも属しておらず、この包含関係に関する情報は概念の整合性判定では利用されていない。すなわち、比較対象の分類体系について、包含関係の連鎖を比較することにより概念体系の比較が可能となり、概念比較と概念体系比較を分離できたことになる。

上位分類群の連鎖は、分類群の包含関係であると同時に、それと双対な分類対象が有する性質の包含関係でもある。このことと、概念比較には概念を構成する要素をラフ集合理論により扱うことが有効であるのに対し、概念体系比較については概念の包含関係を利用することが有効であることとの関係を調べるため、形式概念分析（Formal Concept Analysis）とラフ集合近似された分類体系との比較検討を行なった。形式概念分析における概念を構成する対象集合が分類群すなわち生物個体の集合に対応するのは自明である。形式概念体系における個体集合と属性集合の双対性から、属性集合に対応する集合は、分類群の有する包含関係とは逆向きの順序関係をもつ必要がある。ある分類群に属する個体の集合の補集合はこの条件を満たし、属性集合と順序同形であることが示された。このことは、形式概念分析における属性集合は、その概念に属していない個体の集合と等価であることを意味する。同様の事はラフ集合近似においても成立し、ある概念の負集合が形式概念分析の意味での属性集合に対応することが示された。正集合と負集合、あるいは、ある概念に分類される個体の集合とその概念から排除される個体の集合の組は、相応する形式概念と(ラフ集合の場合は粗視化の程度の近似で) 一対一で対応しているのであるから、属性集合を扱うことと負集合ないし対象概念から排除される個体の集合を扱うことは操作的に対応していることになる。これにより、正集合と負集合により分類群を扱うことと、形式概念分析により概念を扱うことが操作的には同一視できることが示された。

実際のデータベースへの適用可能性を検討するため、実装したライブラリをすでに開発しているユーザインタフェースと組み合わせ、uBio を対象とした検証用ツールとして用いた。利用シナリオとしては、特定の分類群について分類体系を抽出して比較する場合と、特に分類群を指定することなく対象データベースに含まれる分類体系について総体的解析をする場合が考えれるが、データベースのアクセスへの様態としては前者がよりインタラクティヴであり、後者はむしろバッチ処理に近い利用方法である。後者は対象データベースに含まれる全ての分類群名について前者のアクセスを行うことと同等であるので、ここでは前者を想定しつつ述べる。

複数の分類体系間での分類概念の整合性を検証するためには、個々の概念に含まれる分類概念の名前と、含まれない概念の名前の全てが必要となる。素朴にはデータベース全体のデータが必要であるが、uBio の様に分類群名を指定することでそれが含まれるデータソース名を得るインタフェースがあれば、データベースとのデータのやりとりを格段に減らすことができる。とはいうものの、目的とする分類群がより上位で多くの下位分類群を有する場合には分類群の整合性から調べるのは効率が悪く、目的とする分類群

を包含する上位分類群の連鎖を比較することで分類体系の異同判定を先行して行ない、一種のプレフィルタリングとすることが有効である。上位分類群鎖の比較によって異なる分類体系と判断された場合については、確認のため双方の分類体系に含まれる下位分類群名に共通部分があるかどうかによる独立性試験を行ない、独立でなかった場合に異なる分類体系であるとすればよい。独立であった場合には、そもそも対象とした分類群が同名異物であったのであり、分類体系の異同比較そのものに意味はない。上位分類群鎖の比較で異なる分類体系とは判定されなかった組については、目的とする分類群名の下位分類群により整合性判定と独立性判定を行ない、比較対象とする分類体系の異同を判定できる。分類体系全体の異同判定の場合には、原理的にはこの手順を全ての分類名について繰り返せばよいが、必要となる操作の多くが文字列集合に関する集合演算であり、データベースクライアントで処理を行うよりはデータベース上で直接処理する方が効率は良いと期待される。現実的には、対象データベースについて解析しインデクシングする様なサービスポータルを構築する事が有効であろう。現在、uBio の後継ともいうべき Global Name Architecture の検討・開発が GBIF を中心として行われているが、それに基づいた学名情報ポータルなどはそうしたサービスポータルの実装例となり得る。

分類群に含まれる標本集合についての集合演算による分類概念比較はすでに提案されているが、本研究の様な欠落のあるデータには適用できず、また概念体系比較に踏み込んだ手法でもなかった。その意味で、本研究は従来にない手法を提供するものである。生物多様性データの提供という点で欧米に遅れをとらざるを得ない日本からの生物多様性情報学への貢献としては、学名データベースや分類概念データベースの国際協力による整備が軌道に乗りつつある現状もあり、潜在的に少なからざるインパクトを有する。今後は理論を応用したツールを一般向けに提供することで、より広い範囲での貢献と実際のデータの詳細な解析によるより深い検証が進むものと期待される。
また、TAPIR ライブラリはプロトコルの標準化とほぼ同時の公開であり、タイムリーかつ今後の活用が見込まれるものである。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 1 件）
Nozomi Ytow, David R. Morse and Dave Roberts (2006) Rough Set Approximation as Formal Concept *Journal of Advanced Computational Intelligence and Intelligent Informatics* **10**(5) 606-611 査読有

〔学会発表〕（計 3 件）
①伊藤希（2008）「言語としての生物学的階層」科学基礎論学会秋の研究例会 2008 年 11 月 22 日於慶応大学
②伊藤希（2008）「進化・変化・形式化」生物基礎論研究会第 2 回研究会 2008 年 9 月 14 日於京都大学
③伊藤希 「文献データベースは何を提供すべきか」日本動物学会第 77 回大会 2006 年 9 月 21 日於島根大学

〔図書〕（計 1 件）
伊藤希 「生物多様性情報学から見た分類」、片倉晴雄・馬渡峻輔共編『動物の多様性』培風館 2007 年 145 - 172 頁

〔その他〕
http://www.nomencurator.org/
全体的な背景説明のほか、開発ソフトウェア、サンプルデータ等がダウンロード可能

TAPIR クライアントライブラリ
https://digier.svn.sourceforge.net/svnroot/digir/TapirJChirp
http://wiki/tdwg.org/twiki/bin/view/TAPIR/TapirJChirp

６．研究組織
(1)研究代表者
伊藤 希 (YTOW NOZOMI)
筑波大学・大学院生命環境科学研究科・講師
研究者番号：90251016

(2)研究分担者
佐藤 聡 (SATO AKIRA)
筑波大学・大学システム情報工学研究科・講師
研究者番号：90285429

(3)海外研究協力者
Dave Roberts
Natural History Museum, Zoological Department
Head of Microbiology Research Group

David R. Morse
The Open University, Faculty of Mathematics, Computing and Technology
Senior Lecturer