D-005

学術情報データベースにおける質的データのクラスター形成手法

An approach for clustering qualitative data of the scientific database

菊池　桂† 　　升井　洋志†
Kei Kikuchi 　　Hiroshi Masui

1. はじめに

近年，科学研究をより活発に行うことを目的としたオープンサイエンス化[1]の動きが注目を集めており，多くの研究者たちにとって，無料公開された電子ジャーナルを含めた研究データを利用する機会が増加している．それ伴い，大量にある研究データの関連性を見ることは，その研究分野の発展に有用な情報となる．そこで，論文に対してクラスタリングを適用し，類似している内容をもつ論文をグループ化することは，研究データの関連性の把握と今後の研究動向の推測につながる．

クラスタリングは大規模なデータにおけるデータ解析およびデータ可視化の一手法として利用されている[2]．一般的に対象データ間の相対距離が多次元空間でのユークリッド距離として定義できる場合は，クラスタリングは多次元の分布の偏りとして表現可能である．しかし，学術情報データにおける著者や所属機関といった質的データについては，ユークリッド距離を一意に定義できないためにクラスタリングが困難となる．

本研究では，質的データを多く含む論文において，著者や所属機関といった書誌情報から論文間の相対距離を算出し，クラスタリングを行う手法を提案する．

2. 学術情報データベースにおけるクラスタリングの定義

2.1 クラスタリング

論文に対してクラスタリングを適用するには，文字情報として表された質的データから論文間の類似性を相対距離として数値化する必要がある．論文が互いに類似した内容をもつかどうかの指標には，著者，所属機関，標的核，物理量の 4 項目を用いる．クラスター形成には，一対比較表を用いた独自の手法を採用する．

2.1.1 類似度

論文間の類似度の決定には，著者，所属機関，標的核，物理量の 4 項目を用いる．項目ごとに論文間の重複度合を重複度として数値化し，4 項目の重複度を加味して類似度とした．ここで，各項目の重複度の分布には偏りがあるため，重複度を規格化し，類似度を算出した．

著者の重複度の算出式を以下に示す．著者以外の項目に関しても同様にして重複度を算出する．

著者の重複度

$$m_{Aij} = \frac{\text{文献 i と文献 j の同一著者数}*2}{\text{文献 i の著者数}+\text{文献 j の著者数}} \quad (1)$$

論文間の類似度の算出式を以下に示す．類似度の値が大きくなるほど，論文間の類似性が高くなる．

論文間の類似度

$$d_{ij} = \sqrt{m_{Aij}{}^2 + m_{Iij}{}^2 + m_{Tij}{}^2 + m_{Pij}{}^2} \quad (2)$$

m_{Aij} ：著者の重複度
m_{Iij} ：所属機関の重複度
m_{Tij} ：標的核の重複度
m_{Pij} ：物理量の重複度

2.1.2 クラスター形成

クラスター形成手順は以下のとおりである．

(ⅰ)行と列を文献番号とした類似度の一対比較表を作成する．

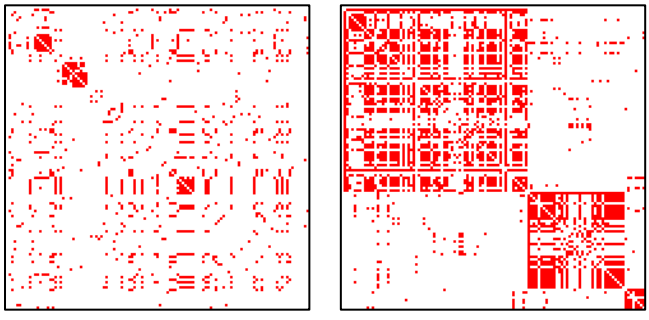
(ⅱ)文献同士が類似しているかどうかを判断するために閾値を決定する．ここで，類似度の値が決定した閾値より大きい場合には，文献 i と文献 j を類似文献であるとし，類似文献同士は同一クラスターに属する可能性がある．

(ⅲ)類似文献の数が最も多い文献を一対比較表の左上に移動させる．

(ⅳ)類似文献同士が連続するように行と列の入れ替えを行い，クラスター形成済みとする．

(ⅴ)クラスター形成済みでない部分に対し，(ⅲ)から(ⅳ)を繰り返す．

図 1 は，閾値を 0.83 としたクラスター形成前とクラスター形成後の一対比較表を表しており，図を見やすくするために拡大図としている．また，類似度の値が 0.83 以上となっている欄に色を付加している．



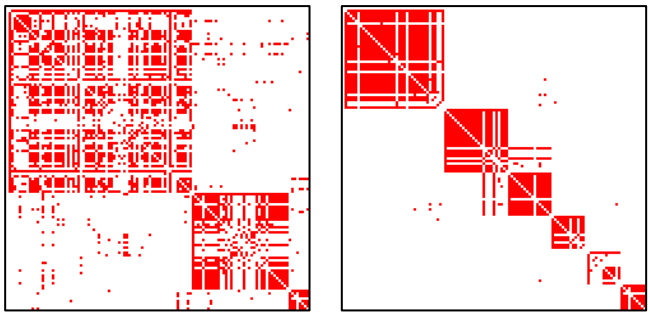
(a)クラスター形成前　　(b)クラスター形成後

図 1:一対比較表

† 北見工業大学，Kitami Institute of Technology

このクラスターの形成手順では，類似文献の数が最も多い文献から，クラスターが形成されることとなり，クラスター内外の類似文献の疎密関係から，クラスター内の論文間の類似性とクラスター間の類似性を視覚的に判断できる．

ここで，閾値によってクラスターの大きさやクラスター数は変化する．クラスター形成の際，閾値を小さな値に設定すると，類似文献となる文献の数が増加し，一つのクラスターが構成する文献数が増加することによって，クラスター数が減少する．逆に閾値を大きな値に設定すると，類似文献となる文献の数が減少し，一つのクラスターが構成する文献数が減少することによって，クラスター数が増加する．そこで，クラスター形成には最適な閾値を決定する必要がある．閾値によるクラスター数の関係を調べたところ，閾値を 1.3 としたときに，別の閾値を設定した時と比べて，クラスター数が増加し，どのクラスターにも属さない文献の数が減少した．そのため，クラスター形成に最適な閾値は 1.3 付近の値といえる．図 2 は図 1 と同様にして，閾値を 0.83 と 1.3 とした場合でのクラスタリング結果を表した一対比較表である．図 2 から，閾値を 1.3 としてクラスター形成をした場合，類似文献の疎密関係から，クラスター内の論文間の類似性は高く，クラスター間の類似性は低いものとなる．



(a)閾値 0.83　　(b)閾値 1.3

図 2:閾値によるクラスタリング結果

2.2 学術情報データベース

本研究で取り扱う学術情報データベースとして，NRDF(Nuclear Reaction Data File)[3]を用いる．

北海道大学大学院理学研究院附属原子核反応データ研究開発センター（JCPRG, Hokkaido University Nuclear Reaction Data Centre）[4]は，日本の加速器で生成された荷電粒子核反応のデータ及び光核反応データの収集し，NRDF として公開している．クラスタリングを適用するデータ対象は，NRDF master files[5]で公開されているデータ番号が D2000 から D2500 の閲覧可能となっているデータファイルとした．

3. クラスタリングの結果

閾値 1.3 とした場合に形成されたクラスターについて，論文数，出版年度，著者数，所属機関数を調べた．

クラスター毎の特徴を図 3 に示す．著者数と研究機関数には従属関係があり，平均著者数と平均所属機関数にはクラスター毎での違いはあまり見られなかった．また，対象とした文献は，2000 年を基準にして 10 年前後に出版された文献であるので，クラスター毎で著しい出版年度の差は見られなかった．

クラスター	論文数	平均出版年度	平均著者人数	平均研究機関数
クラスター1	39	1997	15	5
クラスター2	25	1995	15	5
クラスター3	17	2005	22	8
クラスター4	14	2011	20	7
クラスター5	13	1995	13	5
平均	22	2001	17	6

図 3:クラスター毎の特徴

4. おわりに

本研究では，学術情報データベースに対して，質的データを多く含む論文において，著者や所属機関といった書誌情報から論文間の相対距離を算出し，クラスタリングを行う手法を提案した．

今後の課題として，学術情報データベースにおいて，研究データの関連性の把握と今後の研究動向の推測を視覚的に解釈するために，データ可視化手法について考察していく必要がある．

謝辞

本研究に関して，議論に参加して下さった北見工業大学核科学情報工学研究室の皆様に感謝します．

参考文献

[1]オープンサイエンスの推進について
http://www.mext.go.jp/b_menu/shingi/gijyutu/gijyutu22/siryo/__icsFiles/afieldfile/2016/12/08/1380241_04.pdf
[2]神蔦 敏弘，“データマイニング分野のクラスタリング手法(1) : クラスタリングを使ってみよう!”，人工知能学会誌，Vol.18，No.1 (2003).
[3]NRDF - Search
http://www.jcprg.org/nrdf/
[4]北海道大学大学院理学研究院附属原子核反応データ研究開発センター
http://www.jcprg.org/
[5]NRDF master files
http://www.jcprg.org/master/nrdf.html