Appendix. 8　特徴語を用いた研究領域群の抽出

8-1 サイエンスマップにおける特徴語を用いた研究領域群の抽出とは

サイエンスマップにおいて、研究領域の内容を把握することは重要なステップである。そこで、Appendix 8 に記したように、論文のタイトルやアブストラクトを用いて各研究領域の特徴を示す語「特徴語」を抽出した。しかし、サイエンスマップ 2014 では研究領域数が 844 あり、それぞれの研究領域の特徴語に目を通すことは容易ではない。そこで、この特徴語を基に、ある程度同様の研究内容とみなせる研究領域群（複数の研究領域を包含したまとまり）を自動的に抽出することにより、サイエンスマップ全体の内容について把握できるように試みた。

なお、本調査で行った「特徴語を用いた研究領域群の抽出」のプログラム開発及びその運用については、VALUENEX 株式会社に委託し実施した。

8-2　サイエンスマップにおける特徴語を用いた研究領域群の自動選択アルゴリズム

8-2-1　研究領域群候補の抽出

研究領域群候補の作成は、次に示す 6 つの手順によって行った。

(1)　手順 1：マップをメッシュ（400）に分割する。

(2)　手順 2：メッシュに含まれる論文数（密度）を計算する。

(3)　手順 3：もっとも密度の高いメッシュについて、以下を行う。

①　特徴語を集計する。

（ア）「Appendix. 7」で抽出した「シングル、バイワードありの特徴語上位 60（不定形）」を用いる。

（イ）特定の特徴語が含まれる研究領域数を集計する。メッシュに 5 つの研究領域が含まれ、A という特徴語が 3 領域で出現する場合は 3、2 つの研究領域で出現する場合は 2 となる。比較に使用する特徴語の数は 60 を最大値とする。複数の領域が含まれ、特徴語の数が 60 以上あった場合、含まれる研究領域数が多い単語から上位 60 語が対象となる。

②　その他のメッシュに含まれる特徴語と比較し、同じ特徴語（共通特徴語と呼ぶ）の件数をそれぞれ集計する。

③　共通特徴語が 2 以上であり、かつ、一定の範囲内（距離 11，注：距離はメッシュ数）に含まれるメッシュを一つの領域群候補とする。

(4)　手順 4：手順 3 で領域群候補に設定されなかったメッシュの内、最も密度の高いメッシュについて、手順 3 と同様の処理を行う。

①　手順 3 ですでに他の研究領域群候補に設定されたメッシュが選択されても、一つのメッシュが複数の領域群候補に属することを許すため、新たな領域群候補のメンバーに含める。

(5)　手順 5：手順 4 を実施すると、一定の距離範囲にあり、共通特徴語が 2 以上あるものは特定の研究領域群候補に属することとなる。いずれの研究領域群候補にも属さず、研究領域を含むメッシュが存在する場合、手順 4 を再実行する。研究領域を含み、いずれの研究領域群候補にも含まれないメッシュがなくなるまで手順 4 を繰り返す。したがって、研究領域群候補を作成する段階では、各研究領域はいずれかの研究領域群候補に所属する。

(6) 手順 6：研究領域群候補に含まれる研究領域が 10 以下の研究領域群を削除する。

(7) 手順 7：領域群候補について以下の処理を行う。

① 領域群に含まれる研究領域の中で、X、Y 軸の最大値、最小値及び中心の XY 座標を求める。

② ①で求めた値について、中心の XY 座標を中心とし、(X 最大値−X 最小値)を X 方向の長さ、(Y 最大値−Y 最小値)を Y 方向の長さとした楕円を領域群の候補とする。

③ 各領域群候補の候補に対し、仮定した楕円同士を比較し、以下の楕円を最終的な研究領域群として残す。

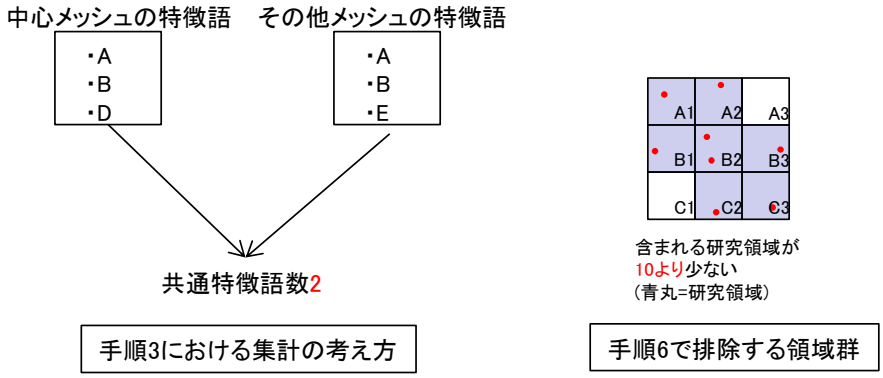
(A) 他の楕円に内包されない楕円である。

(B) 他の楕円と交差している楕円の内、中心点が一定以上離れている。

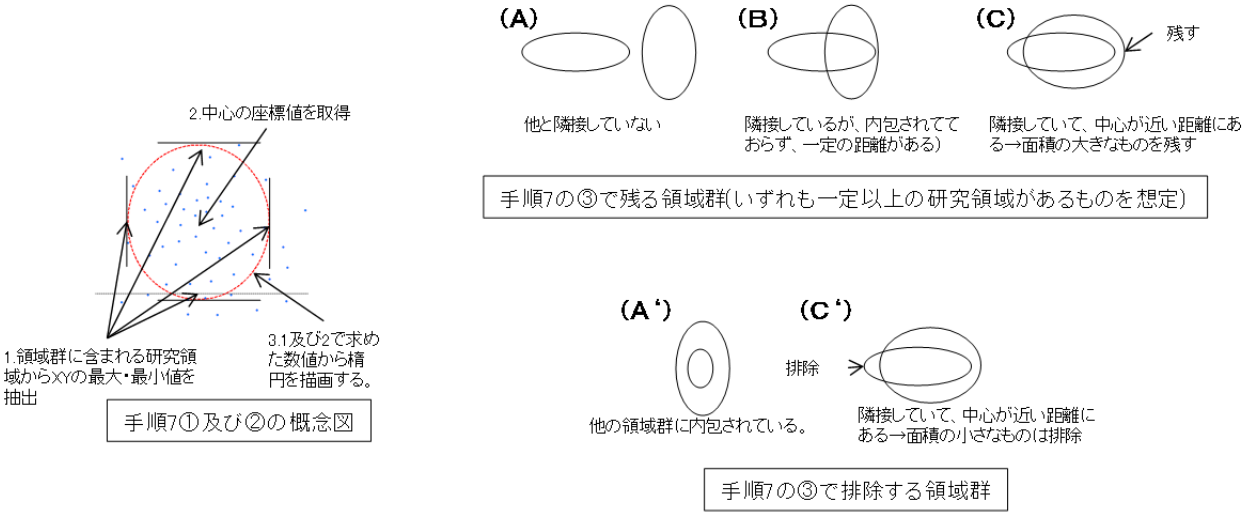
ここでは、楕円の式が $X^2/A^2 + Y^2/B^2 = 1$ とした場合に、中心点 x1、y1 が $x1^2/A^2 + y1^2/B^2 > 0.5$ を対象とした。

(C) 他の楕円と交差し、楕円の中心点が一定距離以内にある場合、面積の大きな楕円を残す。

付録図表 8-1 手順 3 と手順 6 における考え方



付録図表 8-2 手順 7 の考え方



8-2-2　研究領域群候補の削除並びに統合による研究領域群の決定

ここまでのステップで得られた研究領域群を、削除並びに統合することで、最終的な研究領域群を決定した。

研究領域群の中には、他の領域群に囲まれており、その占めるエリアのほとんどがいずれかの領域群と重なっているものが存在する。この領域群に含まれる研究領域の特徴語を分析した場合、領域群に含まれる上位特徴語の多くが周辺の領域群と共通する。

つまり、この領域群は、特に領域群として設定する必要がない。そこで、中心点が一定以上離れているけれども、他の研究領域群と重なりが大きい研究領域群を削除するプロセスを入れた。具体的な手順を以下に示す。

(1)　手順 1：マップをメッシュに分割する(600)。

楕円の面積が複数の他の楕円に内包されている領域を解析的に導くことは難しいため、マップ全体をメッシュ分割し、各メッシュがどの領域群に含まれているかを計算、ある領域群に含まれているメッシュの多くが他のメッシュに含まれている場合は削除する、というアルゴリズムで削除を行なっている。

このアルゴリズムを採用する場合、メッシュが荒くなると計算誤差が大きくなるため、メッシュサイズを、領域群を設定する場合よりも細かく設定している。

(2)　手順 2：各メッシュがどの領域群に含まれるか確認する。

A)　一つのメッシュが複数の領域群に含まれる場合、複数の領域群に含まれることを容認する。

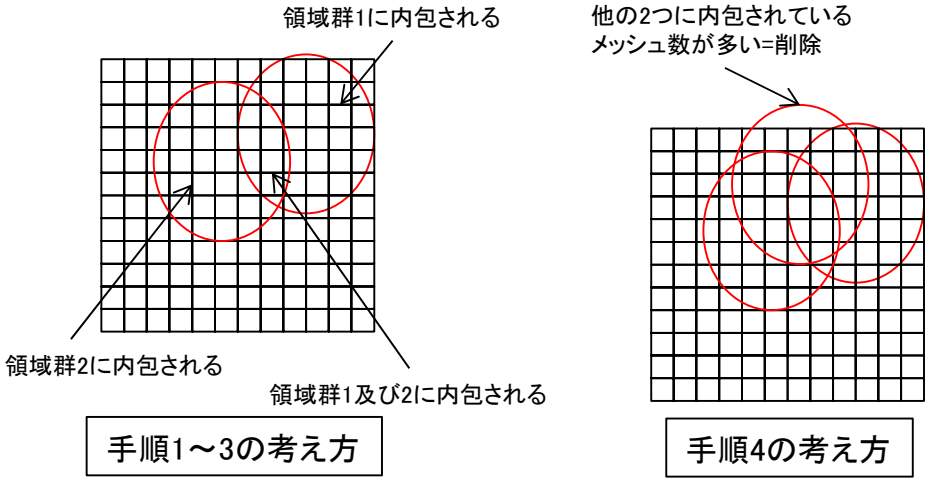
(3)　手順 3：各領域群に含まれるメッシュについて以下を計算する。

A)　各領域群に含まれるメッシュの数。

B)　A)の内、他の領域群に含まれるメッシュの数。

(4)　手順 4：手順 3 で B の数が A の数に対して一定割合以上ある場合(80%以上)、他の領域群に内包されるものとして除外する。ただし、該当する領域群が複数存在する場合、他の領域に含まれる割合が最も高いものを 1 つ除外し、手順 3 に戻る。

付録図表 8-3 手順 1～3 と手順 4 の考え方



(5) 手順 5: 手順 4 の結果残った領域群に対して、各領域群に含まれる研究領域に共通する特徴語上位 60 語(不定形)を抽出する(研究領域群別上位特徴語とする)。

上位かの判定は、領域群として設定したエリアに含まれる研究領域の特徴語について、特徴語別に当該特徴語を含む研究領域数を求め、その件数が多いものから順に研究領域群別上位特徴語とする。

(6) 手順 6: 手順 5 で抽出した研究領域群別上位特徴語について、領域群間の共通件数を計算し、一定割合以上共通する場合(50%以上)、領域群を統合した。

付録図表 8-4 手順 6 の考え方

	領域群1	領域群2	領域群3
特徴語1	A	B	A
特徴語2	B	A	I
特徴語3	C	C	M
特徴語4	D	E	N
特徴語5	E	I	O
特徴語6	F	J	L
特徴語7	G	K	P
・・・	・・・	・・	・・・
特徴語n	H	L	Q

領域群1と領域群2は共通が50%以上。
->統合する。

領域群3は共通性が低い。
->統合せず残す。

手順6の考え方