Appendix 8　特徴語の抽出

1.　サイエンスマップにおける研究領域の特徴語の抽出とは

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。しかしながら、全研究領域のコアペーパやサイティングペーパを全て読むことは難しい。そこで、我々は、各研究領域の特徴を表す語（以後、「特徴語」と記す。）を以下の手順を踏むことで自動抽出し、サイエンスマップにおける研究領域の内容を把握することに用いることとした。

なお、本調査で行った特徴語の自動抽出のプログラム開発およびその運用については、VALUENEXコンサルティング株式会社に委託し実施した。

特徴語の自動抽出の流れは以下である。なお、「特徴語」は、論文のタイトルおよびアブストラクトを格納した解析 DB から、研究領域の内容を示す特徴的な言葉を機械的に抽出することで得られる。特徴語の抽出においては、研究領域を構成するコアペーパおよびサイティングペーパをともに分析に用いた。以降では、コアペーパおよびサイティングペーパをあわせて全論文と呼ぶ。

ステップ1：　論文情報の整理

論文情報を研究領域別に分類し、ステップ2で使用できる形に変換する。

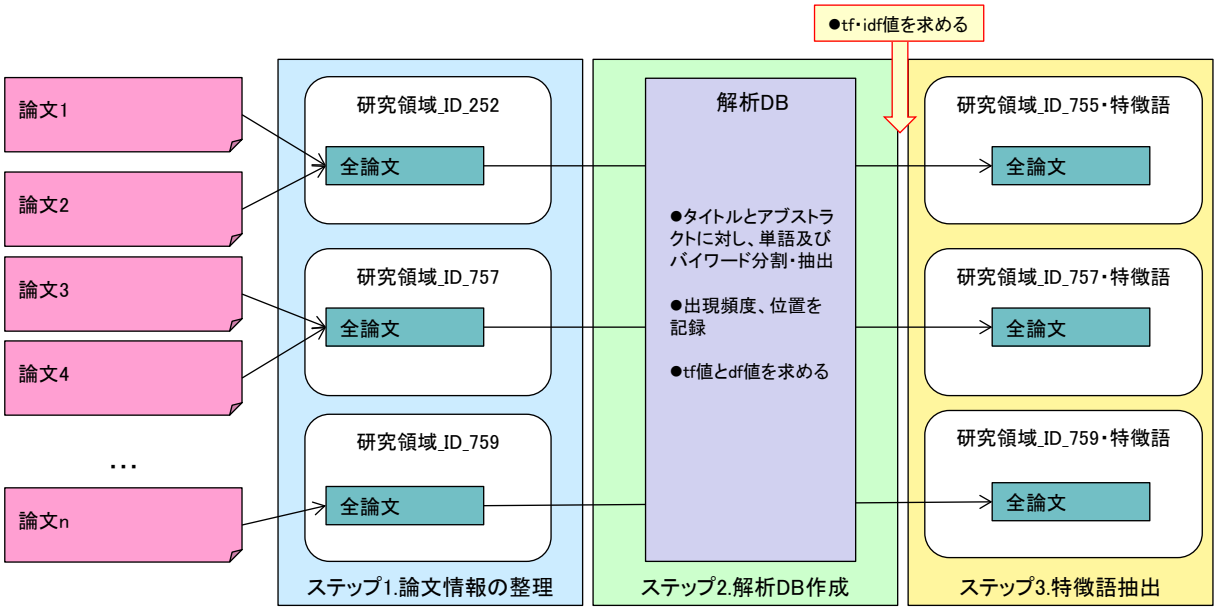
ステップ2：　解析 DB 作成

研究領域別に出現する単語を集計し、解析用データベースを作成する。

ステップ3：　特徴語抽出

ステップ2で作成したデータベースを用いて研究領域別に特徴語を抽出する。

Appendix8_figure 1　特徴語の自動抽出の全体の流れ

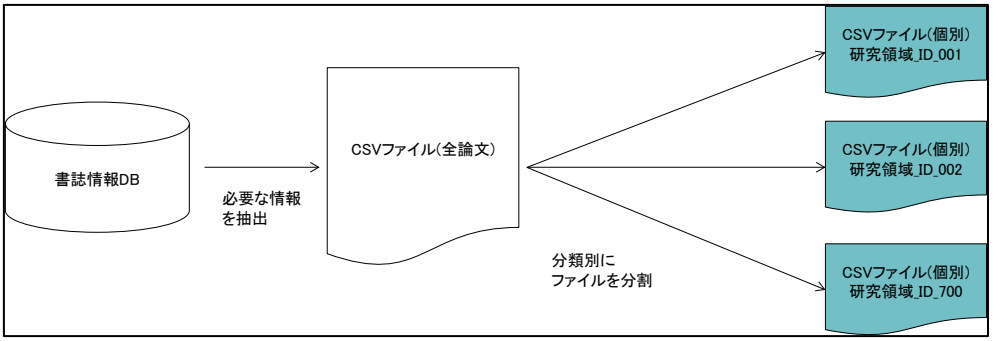


2. ステップ 1: 論文情報の整理

(1) 分析に使用する項目(論文のタイトルやアブストラクト)および個別の論文がどの研究領域に属するかの情報を抽出し CSV 形式のデータとして出力する(以下 CSV ファイル(全論文))。

(2) 抽出した CSV ファイル(全論文)を、研究領域別に、CSV ファイルに分割する(以下、CSV ファイル(個別))。

Appendix8_figure 2 特徴語の自動抽出のステップ 1 の模式図



3. ステップ 2: 特徴語を抽出するための解析 DB 作成

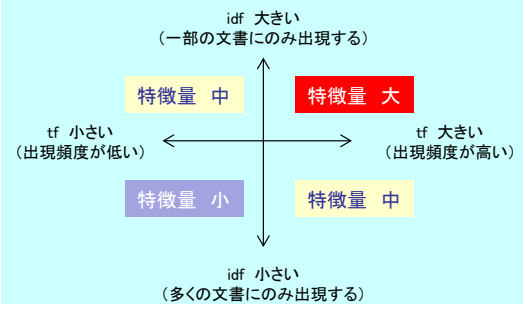
各研究領域の内容を把握するため、各研究領域の特徴語を抽出する。各研究領域の特徴語を抽出するには、各単語が当該研究領域において特徴的に使用されていることを、単語の特徴量として定量的に判定する必要がある。

単語の特徴量を定量的に評価する方法として、広く利用されている tf-idf 法を用いた。ここで tf は単語の出現頻度、idf は逆文書頻度である。

Appendix8_figure 3 tf-idf 法について

TF(Term Frequency)
TFとは、ある文書内で単語がどれだけ多く使用されているのかを示す指標。
ある単語を多く含む文書ほど、その単語について詳しく説明しているものと考える。

IDF(Inverse Document Frequency)
IDFとは、その単語がどれだけの数の文書で使用されているかを示す指標。
多くの文書で使用されている単語より、少ない文書で使用されている単語の方が、その文書の特長をよく表すものと考える。
DFの逆数。

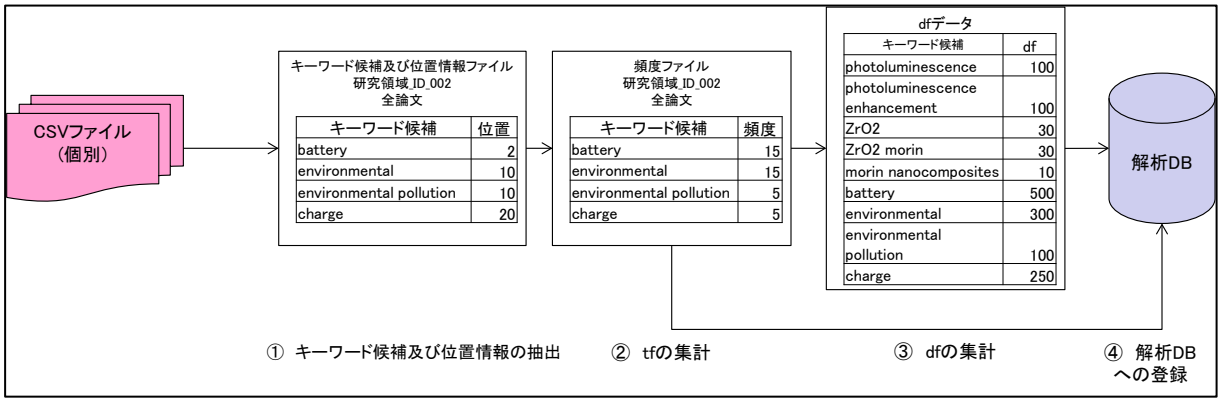


$$tfidf = tf(x, i) \cdot idf(x)$$

特徴語を抽出するための解析 DB 作成は次の手順で行った。

① ステップ 1(論文情報の整理)において抽出した CSV ファイル(個別)毎に含まれるタイトルとアブストラクトを、単語およびバイワード(並んでいる2つの単語)に分割し、キーワード候補を挙げる。この際、キーワード候補がアブストラクト、タイトルのどの位置に出現したかについても記録する。
② CSV ファイル(個別)毎に、各キーワード候補の出現頻度(tf)を集計し、頻度ファイルを作成する。
③ 各キーワード候補について、いくつの CSV ファイル(個別)に出現するか(df)を集計する。
④ 抽出した結果を解析 DB に登録する。なお、この DB は VALUENEX コンサルティング株式会社が文書解析を高速に行うために独自開発したデータベースである。

Appendix8_figure 4 特徴語の自動抽出のステップ 2 の模式図



3－1 語の扱い

語の扱いについては、以下の点を考慮した。

[アブストラクトとタイトルを、単語およびバイワードに分割する]

- 一般的にはひとつの単語をひとつの単位として認識する方法が考えられる。しかしながら、英語の場合、キーワードの並びが意味を持つことがある。例えば、「function」といえば関数や機能等を意味するが、「brain function」では脳機能となる。VALUENEX コンサルティング株式会社で検討を行った所、ひとつの単語をひとつの単位として抽出した上、並んでいる 2 つの単語(バイワード)をひとつの単位として抽出することでより精度の高い特徴語抽出が行えることが判明しており、本調査研究においてもこれを同様の処理を適用した。

[一般的な言葉を不定形にする]

- 活用形や複数形等同じ意味で記述が異なる場合が存在する。これらは原形に戻して処理を行う。

[数値から始まるキーワードは使用しない]

- 数値から始まる文字列は単語としての意味を成さない場合が多いため、使用しない。バイワード

を作成する際には当該文字列はないものとして処理を行う。

[特定のノイズキーワードは使用しない]

- 一般動詞、接続詞、指示代名詞は文章を特徴付けるキーワードとしては不適切かつ頻繁に出現するため、当該文字列は使用しない。バイワードを作成する際は、当該文字列はないものとして処理を行う。例えば「center of the earth」は of、the を排除する。バイワードとしては「center earth」として処理する。

[大文字は原則小文字に変換する]

- 大文字は原則小文字に変換した。ただし、論文中に含まれる大文字のみの文字列や語頭以外の文字が大文字である文字列の場合、省略語の可能性がある(DNA や iPS 等)ため、これらの文字列は大文字のまま取り扱う。なお、論文タイトルがすべて大文字になっている場合もあるため、小文字に変換した文字列も一緒に作成する。
- 仮に論文タイトルで使用されていて、アブストラクトでは小文字で書かれている文字列の場合、大文字で出現する文字列の出現頻度が低いため、キーワード候補とならない。大文字の文字列と小文字の文字列が両方共キーワード候補として抽出された場合は省略語である可能性が高いため、大文字のみをキーワード候補として抽出する。

３－２　TF(Term Frequency)の計算

TF(Term Frequency)とは、ある文書内で単語がどれだけ多く使用されているのかを示す指標である。ある単語を多く含む文書ほど、その単語について詳しく説明していると考える。

以下に TF の計算方法について示す。まず、頻度として何を計測するかについては、以下の2つの候補が考えられる。

① 各論文で各キーワード候補が使用される回数(1 論文に N 回出現したら N と重みづけする)
② 各キーワード候補が含まれる論文数(1 論文に何回出現しても 1 とする)

いずれの方式がふさわしいかを検討するため、それぞれについて特徴量の計算を行い、特徴語の抽出を行った。以下にその結果をまとめる。

- TF をキーワード数とした場合①と論文数とした場合②で、上位キーワードの順位変動はあるものの、キーワードとしては共通して出現していることが多い(一部例外あり)。
- TFをキーワード数とした場合①、研究領域全体ではなく一部の論文で大量に使用されるキーワードが上位に出現する場合がある。
  → TF をキーワード数とした場合、A という論文で 100 回出てきた単語は 100 件の論文で 1 回出てきた論文と同じ特徴量になる。
  → これらのキーワードは領域全体の特徴語ではなく、特定の論文の特徴語となるため研究領域の特徴語としては望ましくない。

上記の検討を踏まえて、本調査研究では、TF の計算方法として②を用いることとした。つまり TF は以下で計算される。

$$tf(x,i)=\frac{n(x,i)}{\sum_{y} n(y,i)}$$

ここで、$n(x,i)$ は研究領域 i を構成する全論文においてキーワード候補 x を含む論文の数であり、分母は、すべてのキーワード候補について、それらを含む論文数の和を取ったものである。

3－3　IDF(inverse document frequency)の計算

IDF(inverse document frequency)とは、その単語がどれだけの数の文書で使用されているかを示す指標である。多くの文書で使用されている単語より、少ない文書で使用されている単語の方が、その文書の特長をよく表すものと考えるものであり df の逆数で表すことが出来る。

全文書数としては通常、研究領域の全論文数であり、本調査研究においてもこれを採用する。

$$idf(x)=\log\left(\frac{N}{df(x)}\right)+1$$

ここで、N は研究領域(サイエンスマップ 2012 では 823 研究領域全て)を構成する全論文数(18,515 件)であり、$df(x)$ はその中でキーワード候補 x を含む論文の数である。

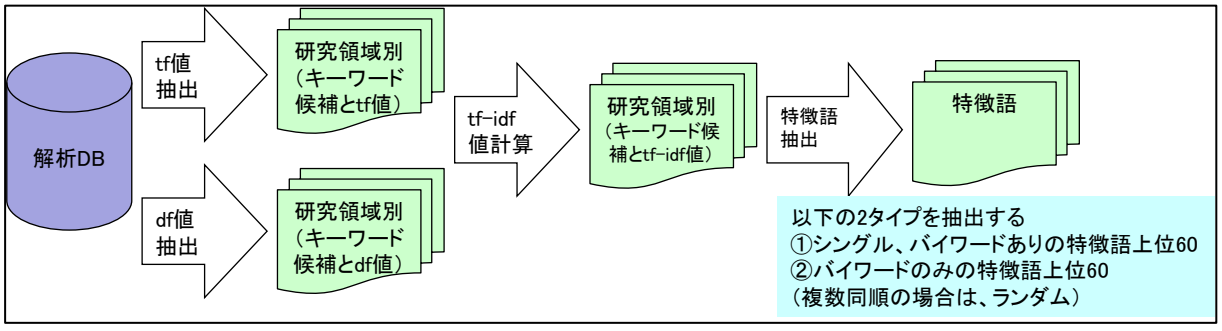
4. ステップ 3:特徴語抽出

ステップ 1 から 2 で構築した解析 DB を用いて特徴語を抽出した。本調査研究では、特徴語から研究領域の内容が分かるように、特徴語をフレーズとして抽出した。以下に、その手順を示す。

4－1　特徴量が上位のシングルワード、バイワードの抽出

ステップ 2 で説明した TF および IDF の値を用いて、tf-idf 値を計算する。tf-idf 値は TF と IDF の積として定義される。tf-idf 値が大きい(特徴量が大きい)ものから、特徴量上位 60 語を抽出した。その際、シングル、バイワードのなかでの上位 60 語と、バイワードのみの上位 60 語の両方を抽出した。

Appendix8_figure 5 特徴語の自動抽出のステップ 3 の模式図



４－２　抽出した特徴語(バイワード)からのフレーズ候補の抽出

各研究領域で特徴量上位 60 のバイワードのみをもちいて、フレーズの抽出を行った。抽出の基本的な考え方は以下のとおりである。

① STEP1（バイワードの伸長によるフレーズ候補の構築）

原文中の特徴語(バイワード)の位置を特定し、特徴語を中心として前後に文字列を1ワードずつ伸ばしたフレーズを作成して出現件数をカウントする。この際、特徴語リストの記述ではなく文章中の記述を使用した。

たとえば、特徴語の「based superconductor」に関して前後に伸ばしたフレーズを抽出する場合、以下のような文字列が抽出される。

- 1 ワード増加させた場合
  - → iron based superconductors　10 件
  - → Fe based superconductors　1 件
  - → based superconductors and　4 件
  - → based superconductors of　4 件
  - → based superconductors is　3 件
- 2 ワード増加させた場合
  - → and iron based superconductors　8 件
  - → or iron based superconductors　2 件
  - → of iron based superconductors　1 件
  - → iron based superconductors and　4 件
  - → iron based superconductors of　4 件
  - → iron based superconductors is　3 件
- 3 ワード増加させた場合
  - → or iron based superconductors and　4 件
  - → and Iron based superconductors or　3 件
  - → of iron based superconductors or　2 件
  - → iron based superconductors and Ni　2 件
  - ・・・・・・・・・
- 4 ワード増加させた場合
  - → or iron based superconductor and Ni　2 件
  - ・・・・・・・・・

フレーズの伸ばし方としては、以下の組み合わせ全てについて行っている（●●：特徴語、○：伸ばしたワード）。フレーズごとに、何ワード伸ばしたか、またどの方向にいくつ伸ばしたか、の情報を取得しておく。

1 ワード増加 ： ○●● OR ●●○

2 ワード増加 ： ○○●● OR ○●●○ OR ●●○○
3 ワード追加 ： ○○○●● OR ○○●●○ OR ○●●○○ OR ●●○○○
4 ワード追加 ： ○○○○●● OR ○○○●●○ OR ○○●●○○
OR ○●●○○○ OR ●●○○○○
・・・・・・・・

② STEP2（フレーズ候補の抽出）

STEP1 で前後に文字列を増加させたフレーズのうち以下の条件にマッチするものを、フレーズ候補として採用した。

[1] 特徴語を一定ワード数増加させた場合に最も多く出現するフレーズが、研究領域に含まれる論文に含まれる当該特徴語件数の 1 割以上ある。

STEP1 で示した「based superconductor」を例とする。研究領域に含まれる論文のタイトル、要約中に based superconductors が 39 件含まれていた場合、3 ワード数増加させた場合には 1 割を超える 4 件の同一フレーズが含まれるが、4 ワード増加させた場合には最大でも 2 件となるので、3 ワード増加が上限となる。

（例）

- 1 ワード増加させた場合 ・・・・ ○
  → iron based superconductors 10 件
  → Fe based superconductors 1 件
  → based superconductors and 4 件
  ・・・・・・・・・
- 2 ワード増加させた場合 ・・・・ ○
  → and iron based superconductors 8 件
  → or iron based superconductors 2 件
  → of iron based superconductors 1 件
  → ・・・・・
- 3 ワード増加させた場合 ・・・・ ○
  → or iron based superconductors and 4 件
  → and Iron based superconductors or 3 件
  → iron based superconductors and Ni 2 件
  → ・・・・・・・・・
- 4 ワード増加させた場合 ・・・・ ×
  → or iron based superconductors and Ni 2 件
  → cupper or iron based superconductors and 1 件
  → ・・・・・・・・・

[2] 上記[1]の条件に合致したワード増加数に該当するフレーズのうち、前後の増加のさせ方が同じフレーズ（例えば○○●●○）をフレーズ候補とする。上記例であれば、以下のようになる。

（例）

- 3 ワード増加させた場合
  - → or iron based superconductors and　4 件　・・・・フレーズ候補（○○●●○）
  - → and Iron based superconductors or　3 件　・・・・フレーズ候補（○○●●○）
  - → iron based superconductors and Ni　2 件　・・・・フレーズ候補としない（○●●○○であり、増加のさせ方が異なる）
  - → of oxide based superconductors and　1 件　・・・・フレーズ候補（○○●●○）
  - ・・・・・・・・・

4－3　抽出したフレーズ候補からのフレーズの抽出

前項の通り抽出したフレーズ候補を確認すると、フレーズ候補の先頭または末尾に接続詞や前置詞が多くみられる。これらは意味のあるフレーズとしては不要なものである。そこで、一旦抽出したフレーズ候補に対し、ストップワードを用いて前後の不要な文字列を削除し、共通性があり、意味のあるフレーズを抽出した。なお、この処理はいったんフレーズ候補を抽出した後に行う。

STEP1：

抽出したフレーズ候補の先頭又は最後にある文字列がストップワードの場合、削除する。ストップワードが続く場合は繰り返し削除する。

STEP2：

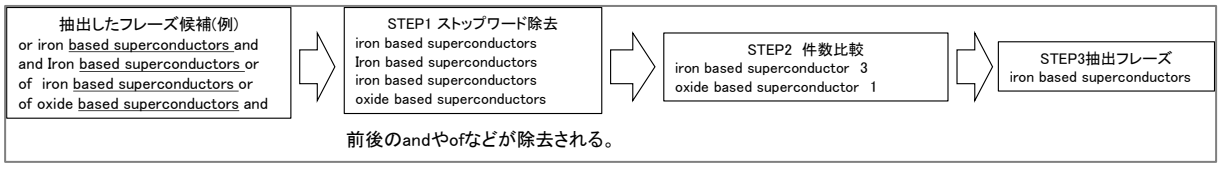
STEP1 が終了した結果を、大文字は小文字に変換し、ハイフンや語末の「s」を除去して集計し、件数比較を行う。

STEP3：

STEP2 で最も多かった文字列のベースとなる STEP1 の文字列をフレーズとする。この際、STEP1 の文字列が複数存在する場合は最も件数が多い文字列をフレーズとする。

以下の例では、STEP2 終了時に「iron based superconductor」を含むものが多い。これを含む STEP1 の文字列としては「iron based superconductors」が最も多いため、これをフレーズとして採用する。

Appendix8_figure 6 フレーズ候補からフレーズへの流れ



5. 特徴語の和訳

これまでのプロセスで得られた特徴語は、タイトルやアブストラクトが英語で書かれた論文をもとに抽出しているので、英語表記となっている。特徴語から研究領域の内容を理解しやすくするために、特徴語(フレーズ)の和訳を行った。報告書に掲載されているのは、和訳を行った特徴語である。

5－1 特徴語の和訳の対象とした研究領域

特徴語の和訳の対象とした研究領域は以下である。

① サイエンスマップ 2012 の 823 研究領域
② サイエンスマップ 2008 およびサイエンスマップ 2010 の研究領域の内、トラジェクトリーマップにおいて、上記の 823 研究領域とつながりを持っている 728 研究領域(サイエンスマップ 2008 で 261 研究領域、サイエンスマップ 201 で 467 研究領域)

それぞれの研究領域から特徴量が上位 20 の特徴語(フレーズ)を抽出し、そこから重複およびアブストラクトの一部に含まれる学会名や雑誌名等のノイズを除いた約 16,000 の特徴語(フレーズ)を和訳の対象とした。

5－2 特徴語の和訳

特徴語の和訳は以下の手順で行った。特徴語の和訳には、約 20 人日を要した。

① 抽出された特徴語(フレーズ)を、Google のウェブサイト翻訳ツールを用いて一括して日本語に翻訳した。
② 上記で得られた和訳の全てについて、目視による確認を行った。抽出された特徴語(フレーズ)の中には断片化されているものも存在する。それらについては、適時フレーズを補完して和訳を行った。なお、目視による確認、フレーズの補完の際には、適時、Web 上の公開情報を参考にした。

5－3 特徴語の和訳精度

特徴語の和訳は、報告書執筆者による仮訳であり、より適切な和訳が存在する可能性がある点については留意願いたい。

なお、特徴語の和訳情報の取得困難さには分野による違いが見られた。一例をあげると、地球温暖化など研究者以外も関心を持つような話題については、Web 上の公開情報から一般的と思われる和訳を見出すことが、比較的容易であった。逆に、素粒子物理学などの最先端の研究でかつ国際共同研究が主に行われている研究領域については、和訳を見出すことが困難な特徴語が多数みられた。これらの特徴語については、英語の特徴語をそのまま掲載している。

ある英語の科学技術用語に対応する日本語が存在するかについては、我が国における研究者コミュニティーの有無、研究者コミュニティーの大きさ、研究の進展の速度、研究の国際化の度合、科学研究と社会とのつながりの度合などが関係していると思われる。しかしながら、この仮説の検証は、本調査研究の範囲を超えることから、今後の研究の進展を待つこととしたい。