# 位相的データ解析を応用した医薬品テキスト情報からの特徴抽出のアプローチ

塩野義製薬(株)　北西　由武　　岡山大学　石岡　文生
岡山大学　　飯塚　誠也　　岡山大学　栗原　考次

## 1.　はじめに

近年，データ・情報が蓄積され巨大化してきているが，データベースの量的データのみから知見を得るような方法は限界が生じ始めており，テキストや画像などの質的データも解析対象に含める動きが活発化している．特に「質的データを特徴抽出，特徴量計算のために如何に定量化するか」への課題解決アプローチは目的も様々で算出方法も多様である．また，テキスト情報は膨大であるが故に解析処理にも効率性も求められる．本発表では医薬品の質的データ（主に医薬品情報を含んだテキストデータ）を対象に，前処理の影響分析も含めて位相的データ解析を適用した結果を報告し，その解析アプローチの特性と効率性，更なる応用について報告する．

## 2.　テキストデータの前処理

医薬品テキスト情報として，本発表では Wikipedia 日本語版を利用する．具体的には約 100 万件の記事が格納されている xml ファイルとそれらのカテゴリ情報が格納されているテーブル(https://dumps.wikimedia.org/jawiki/)から医薬品に関係する記事約 2000 件を抽出した．抽出したテキストデータに対し，特に医学，薬学の単語識別性能を向上させた辞書を適用し，形態素分析を行った．定量化するにあたって，局所的重みである TF(Term Frequency)値と大域的重みである IDF(Inverse Document Frequency)値を掛け合わせ，文章の長さで正規化した TF*IDF 値を用いた．つまり，記事毎に TF*IDF 値を算出し，記事×単語の情報行列を作成し，これを解析（分類）に用いる基本データ：X とした．

## 3.　位相的データ解析の利用

解析（分類）に利用した位相的データ解析は主に①Filtering, ②Binning, ③Partial Clustering, ④Connecting の 4 つのプロセスから成る．テキストデータから作成した基本データ：X を①Filtering として，(f,g):X→Y と低次元化された空間で近さを定量化（例えば距離）する．次にデータ駆動型解析のポイントとして，データに応じ，②Binning で範囲を規定し，③Partial Clustering を行う．さらに Clustering ノードを エッジによって④Connecting することでデータを形状化する． この形状化したデータを用いることで，さらなる解析への応用も可能となる．当日の発表では，テキストデータの定量化処理と位相的データ解析の詳細とそれらの性能評価，適用結果についても報告する．

参考論文

1)　Edelsbrunner, H. et al. Topological Persistence and Simplification. Discrete Comput Geom. 28, 511–533 (2002)
2)　Gunnar Carlsson, Topology and data. Bull. Amer. Math. Soc. 46, 255-308 (2009)
3)　Manish Saggar, et al. Towards a new approach to reveal dynamical organization of the brain using topological data analysis. Nature Communications. 9, Article number: 1399 (2018)