# ISDL レポートの時系列クラスタリング

水野 珠季

## 1　はじめに

情報通信技術の発展に伴い，インターネット上では様々な情報が公開され，入手可能となっている．また，医療や流通など様々な分野で，これまで活用されてこなかった大量の情報が蓄積されている．近年では，これら多種多様かつ大量の蓄積された情報を解析し，活用しようという動きが盛んになっている [1)]．蓄積された情報の活用法としては，ユーザの嗜好や興味を抽出し，それらにマッチした新たな情報をユーザに提示するという情報推薦や，ニュースなどの大量の時系列文書から話題やキーワードを抽出するといったトピック抽出などが挙げられるが，本研究では後者に着目した．個人がインターネット上などに蓄積された大量の情報を個々にチェックし，最新の話題や傾向を把握することは困難であり，これを支援するシステムが求められている．また，トピックの推移を把握することは，今後のトレンド予測にも役立つと考えられる．

そこで本研究では，時間軸にそって文書を内容によって分類し，時間の経過に伴うトピックの推移を把握できるシステムの開発を目指す．このシステムを実現し，論文などの文書の集合に適用することで，関連文書の検索や研究テーマなどのトレンドの把握が容易になる．また，トピックが変化していく様子を追うことで，今後取組むべき課題の発見にもつながると考えられる．

## 2　提案システム

本研究では，ISDL レポートを時系列クラスタリングの対象とする．ISDL レポートとは，本研究室において研究成果や文献調査の結果を外部に公開する目的で書かれた HTML 形式の文書である．

### 2.1　概要

提案するシステムでは，ISDL レポートを年度単位で内容によって分類する．その結果得られたレポートのグループをキーワードとともに提示することで，本研究室における研究テーマの推移を捉えることができる．

### 2.2　ISDL レポートの分類

ISDL レポートを分類する手法としては，クラスタリングを利用する．クラスタリングとはデータ解析手法の一つで，事前に定義された基準に従って分類するのではなく，一つの多様な集団をより同質的なサブグループに分類する手法である．レポートをクラスタリングするには，レポート同士の関連性を定義する必要がある．そのため，レポートの内容を解析し，関連するレポートにリンクを張ることでレポートの集合をネットワークとして表現する．

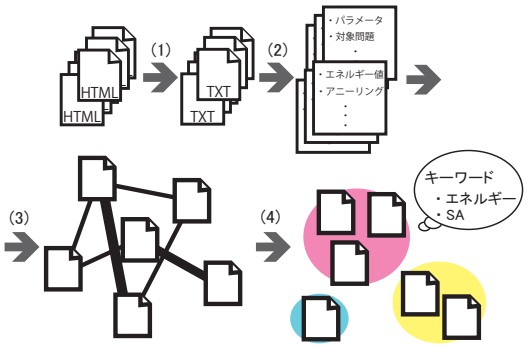


Fig.1　レポートの分類手順（出典：自作）

レポートを分類する手順を Fig. 1 に示す．以下の説明は Fig. 1 中の番号と対応する．

(1) レポートの HTML タグを除去する
(2) レポートを形態素解析し，名詞を抽出する
(3) レポートの関連度を求め，ネットワークを生成する
(4) レポートのネットワークをクラスタリングする

レポートのクラスタリングは年度単位に適用する．ここでの年度単位とは，2002 年度，2003 年度と年度ごとに個別に適用するのではなく，開始年度から該当年度までの全レポートに対して適用するものとする．さらに，単純に年度単位に独立してクラスタリングを行うだけでは，過去のクラスタリング結果と大きく異なった結果になる可能性がある．これではトピックの推移の把握に適さないため，本システムでは時系列的な変化を考慮したクラスタリング手法として，榊ら [2)] の提案する制約付きクラスタリングを用いる．

## 3　提案システムの実装

今回，提案するシステムのうち，ISDL レポートの関連度を求め，それを重みとしたレポートのネットワークを生成する部分の実装を行った．

### 3.1　HTML タグの除去と形態素解析

レポートの HTML ソースから HTML タグを除去し，レポート本文を取得する．取得したレポート本文に対して，MeCab[*1] を用いて形態素解析を行い，名詞を抽出する．この際，「情報通信技術」のような複合語については，連続する名詞を連結することで一つの単語として抽出している．

### 3.2　ベクトル表現

レポートを単語の重みベクトルとして表現する．単語の重みは tf-idf によって求める．tf-idf は，文章中の特

[*1] http://mecab.sourceforge.net/

徴的な単語（キーワード）を抽出するためのアルゴリズムであり，$tf$（Term Frequency:単語の出現頻度）と $idf$（Inverse Document Frequency:逆出現頻度）の二つの指標で計算される [3]．文書 $j$ における単語 $i$ の $tf$ は以下の式で表す．

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{1}$$

$n_{i,j}$ は文書 $j$ における単語 $i$ の出現回数を示す．単語 $i$ の $idf$ は以下の式で表す．

$$idf_i = log\frac{|D|}{|\{d_j : d_j \ni t_i\}|} \tag{2}$$

$|D|$ は総ドキュメント数，$|\{d_j : d_j \ni t_i\}|$ は単語 $i$ を含むドキュメント数である．$tf$ と $idf$ を掛け合わせて以下の式を得る．

$$tfidf_{i,j} = tf_{i,j} \cdot idf_i \tag{3}$$

式 (3) で得た値を用いて，レポートを各単語を次元としたベクトルで表現する．本システムでは計算量を抑えるためベクトルの次元数を 50 とし，tfidf 値の上位 50 個の単語を用いてレポートをベクトル化する．

### 3.3 関連度計算とレポートのネットワーク化

ベクトルによって表現されたレポートの関連度を求める．2 つのレポートの関連度は式 (4) のようにベクトルが成す角度の余弦で表す．

$$sim(rep_1, rep_2) = \frac{\overrightarrow{\mathrm{V}}(rep_1) \cdot \overrightarrow{\mathrm{V}}(rep_2)}{|\overrightarrow{\mathrm{V}}(rep_1)||\overrightarrow{\mathrm{V}}(rep_2)|} \tag{4}$$

式 (4) で求めた関連度を重みとし，レポートのネットワークを生成する．レポートのネットワークは，隣接行列 $A$ によって表現する．$A$ は，レポート数を $n$ とすると $n$ 次正方行列となり，$A$ の要素 $a_{ij}$ はレポート $i$ とレポート $j$ の関連度 $sim(rep_i, rep_j)$ となる．

## 4 クラスタリング結果の検討

3 章で生成したレポートのネットワークをクラスタリングすることで，その妥当性を検証する．クラスタリングには，Newman[4] の提案するネットワークを対象とした階層的クラスタリング手法を用いる．今回，2008 年度に公開されたレポート 79 本を対象にクラスタリングを行った．その結果，Fig. 2 に示すような 20 個のクラスタに分割された．Fig. 2 において，一つの円が一つのクラスタを表し，円の大きさがクラスタ内のレポート数を表す．また，単語はクラスタを代表するキーワードである．このキーワードは，クラスタを構成する全レポートの各単語の tfidf 値を合計し，値の大きい順に 5 つ抽出している．

クラスタリング結果の詳細の一部を Table 1 に示す．Table 1 の 1 はクラスタリングに関するレポート 2 本と文献リスト 1 本から成るクラスタで，キーワードにも「クラスタ」という単語や「凝集法」「Newman」といったクラスタリングの手法名が含まれており，妥当な結果だと

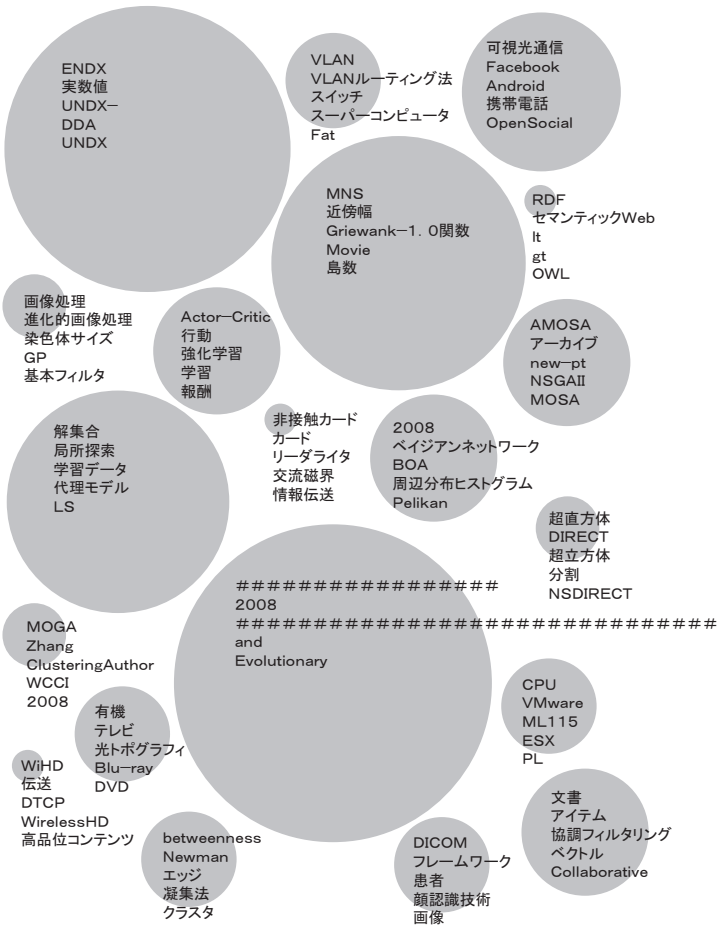


Fig.2 クラスタリング結果（出典：自作）

Table1 クラスタリング結果の一例（出典：自作）

| | レポートタイトル | キーワード |
|---|---|---|
| 1 | 凝集法と k-means 法 | betweenness |
| | ネットワークを対象とした | Newmwn |
| | クラスタリング | エッジ |
| | ISDL Report：文献リスト： | 凝集法 |
| | ネットワークからのコミュニティ抽出 | クラスタ |
| 2 | 【IT 用語】 光トポグラフィ | 有機 |
| | 【IT 用語】液晶, プラズマ, | テレビ |
| | 有機 EL テレビ | 光トポグラフィ |
| | 【IT 用語】HD DVD と | Blu-ray |
| | Blu-ray Disc のまとめ | DVD |
| 3 | ISDL Report：文献リスト： | ########## |
| | 協調探索を用いた多目的最適化 | 2008 |
| | 【ISDL Report：文献リスト： | ############## |
| | DIRECT を用いた多目的最適化 | and |
| | ISDL Report：文献リスト： | Evolutionary |
| | 主要な多目的 GA | |
| | ISDL Report：文献リスト：対話型 | |
| | 遺伝的アルゴリズムにおける個体生成 | |

言える．しかし 2 では，テレビや DVD といった AV 機器関連のレポートと医療機器である光トポグラフィに関するレポートが同一のクラスタに分類されてしまった．このように，本来関連が低いと思われるレポートが含まれるクラスタがいくつか見られた．これは，レポート全体として見ると共通する語は少ないが，特徴的なある一

語が共通していたためである．このように特徴的な一語に大きく影響される要因はベクトルの次元数を 50 に制限したことにあると考えられ，今後，適切な次元数を検討する必要がある．また 3 はレポート数が最大となったクラスタで，文献リストが書かれたレポートのみによって構成されていた．これは，文献リストがすべて決まった形式で書かれているためだと考えられる．文献リストには文献のタイトルや著者名しか書かれておらず，tfidf 法による重み付けには適さない．そのため，他のレポートとは区別して処理する必要があるだろう．

## 5　まとめと今後の課題

本研究では，時系列文書からトピックを抽出し，その推移を把握できるシステムの開発を目指し，その一例として ISDL レポートを時系列クラスタリングするシステムを提案した．本報告では，提案システムの一部として，レポートの形態素解析，関連度計算などについて実装を行い，得られたレポートのネットワークにクラスタリングを適用することで，提案手法の妥当性について検討を行った．今後は，4 章で述べた検討事項に基づいて形態素解析，関連度計算の精度向上を図るとともに，制約付きクラスタリングやインタフェースの実装を行い，ISDL レポートのテーマの推移を把握できるシステムを実現する．

## 参考文献


1) 情報大航海プロジェクト
http://www.meti.go.jp/policy/it_policy/daikoukai/index.html

2) 榊剛史, 松尾豊, 石塚満: 制約付きクラスタリングを用いた論文分類. 人工知能学会第 20 回全国大会論文集,1A1-1(2006)

3) 澁谷 翔吾, 廣安 知之, 三木 光範,
ベクトル空間法を利用した類似度計算
http://mikilab.doshisha.ac.jp/dia/research/report/2008/1110/002/report20081110002.html

4) M.E.J. Newman, Fast algorithm for detecting community structure in networks, Phys. Rev. E 69, 066133(2004)