2F4-6

研究レポートシステムのクラスタリング解析

Clustering analysis of Research Report System

大西　祥代[*1]　廣安　知之[*2]　三木　光範[*2]
Sachiyo Onishi　Tomoyuki Hiroyasu　Mitsunori Miki

[*1]同志社大学　大学院工学研究科
Graduate School of Engineering, Doshisha University

[*2]同志社大学　工学部
Department of Engineering, Doshisha University

In our laboratory, students and researches open their research results on the web as ISDL reports. These reports are more than 1300 and many people are accessing them. However, The link between reports hardly exist nor there are no cluster information. If the effective mechanisms exist, these reports can be treated more effectively. In this paper, we tried to cluster these reports and illustrated their relations as the network. The research keywords in the laboratory were defined and the relations between the keywords were analyzed from the result of the related level and clustering between the reports.

1.　はじめに

現在，我々の研究室では，学生が研究報告，文献調査などのレポートを HTML 化し Web 上で公開している．研究レポートの公開目的は，我々の研究活動を広く外部に知ってもらい，研究の活性化を図ることにある．これまでに約 1300 本ものレポートが公開され，これらのアクセスログを解析した結果，外部からの一定数のアクセスが確認でき，研究室の内部外部共に有用なデータとして活用されていることがわかった．そこで，公開しているレポートをより効率的に閲覧できる仕組みや，レポートの作成を支援する仕組みを構築することで，レポートシステムをより有効的に活用することができると考えられる．例えば，現在の研究レポートはレポート間でのリンクがほとんどなく，Web 上に独立した状態で存在している．そこで，研究レポート間の関連度からレポートのネットワークを形成し，閲覧者に関連レポート情報を提示する仕組みを作る．また，レポートについてクラスタリング解析を行うことで，我々の研究を表すキーワード群を可視化し，レポートの閲覧や作成を支援するシステムを構築する．

2.　研究レポートシステム

2.1　研究レポートシステムの概要

我々の研究室では学生が研究報告や文献調査などを HTML 形式で記述し，Web 上で公開する仕組みを持っている．研究レポートの公開は，研究活動を一般に公開し，研究を活性化することを目的としている．これまでに公開している研究レポートに対して他の研究者から問い合わせがあったケースも複数存在する．2002 年より公開を始め，現在 (2007 年 3 月) までに 1368 本のレポートが公開されている．

2.2　アクセス状況

研究レポートのアクセスログの解析を行い，レポートへのアクセス数を調査した．図 1 は，2006 年 5 月から 11 月のアクセス数を示す．レポート全体に対して，外部から平均 44112 件/月のアクセスが確認できた．また，表 1 は，2006 年 11 月のアクセス数上位 5 つのレポートであり，これらのレポートのように，一定数のアクセスが確認でき，外部から有効に活用されている可能性があると考えられる．また，研究レポートの参照元を調査したところ，検索エンジンからのアクセスが大多数を占め，他の研究レポートからのアクセスは全体の 6%ということがわかった．さらに，類似した内容のレポートにも関わらず，アクセス数に 10 倍以上の差があり有効に活用されていないレポートも存在することがわかった．

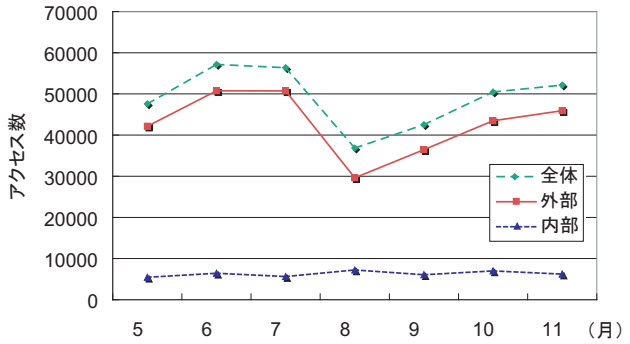

図 1: アクセス数 (2006 年 5 月〜11 月)

表 1: 月間アクセスランキング

順位	研究レポートタイトル	アクセス数
1	最短経路問題におけるアルゴリズム [ダイクストラ法]	795
2	【IT 用語】動画圧縮技術	718
3	PHP を用いた XML-RPC の基礎	713
4	Web アクセスログ解析ソフト Analog の利用方法	699
5	【IT 用語】PCI バス	676

2.3　研究レポートシステムに求められる改善点

2.2 節より，研究レポートは外部からの一定のアクセスがあり，外部ユーザによる閲覧や有効活用の可能性があると考えられる．そこで現状のシステムについて，以下の点を改善する必要がある．

- 研究レポート間のリンクがほとんどない状態である
  アクセスログの解析から検索エンジンからの閲覧が多いことがわかった．レポートに関連レポート情報が付加されていれば，レポート単体のアクセスからより多くの閲覧を促すことができると考えられる．

連絡先: 大西 祥代，同志社大学大学院工学研究科，
〒 610-0394 京都府京田辺市多々羅都谷 1-3 香知館 KC101，
0774(65)6921，onishi@mikilab.doshisha.ac.jp

- 研究レポートの作成を支援する仕組みがない
  レポートを作成，公開，閲覧するという仕組みに加え，作成支援を行うことで，より研究を活性化させることができると考えられる．

2.4 提案する研究レポートシステム

2.3 節のことから，以下のような研究レポートシステムを構築する．

- 関連研究レポートの提示
  現在，研究レポート間でのリンクはほとんど存在しないため，レポート閲覧の効率が悪い．そこで，自動的にレポート間でリンクを作成し，関連レポートの情報を閲覧者に提示する．まず，レポート同士の関連度計算を行い，関連の高いレポートを自動的に抽出する．本システムでは，GETA(Generic Engine for Transposable Association)*1を利用してシステムを構築する．

- 研究キーワードの可視化
  レポート間の関連度計算を行うことで，研究レポートのネットワークを作成することができる．そこで研究レポートネットワークに対しクラスタリング解析を行い，グループ化されたレポートを特徴付けるキーワードを抽出する．そして，図 2 のようにキーワード間の関連性や重要性を可視化することで，閲覧支援，レポート作成支援を行えるシステムを構築する．キーワードの重要性にはグループのレポート本数や，アクセス数を考慮する．キーワード間の関連性には，同じクラスタのキーワードなら関連性が高くなるよう可視化を行う．研究キーワードを可視化することにより，以下のような効果を得られる可能性が考えられる．
  - 研究テーマの一覧や，注力されているテーマを把握することが容易になり，研究レポートの閲覧支援が可能になる
  - キーワードの重要性や関連性を見ることで，書くべき研究レポートの傾向が把握でき，研究レポートの作成支援が可能になる

以上のシステムを目指し，まず研究レポート間の関連度を求め，レポートネットワークを生成し，クラスタリング解析を行う．

3. 研究レポートのクラスタリング

3.1 研究レポートネットワークの作成

研究レポート間の関連度を計算し，レポートのネットワークを作成する．関連度計算には GETA を利用する．GETA は検索する文書の単語の出現頻度を行列で表現したものに対して，文書間の類似度を高速計算するツールである．単語の出現頻度の際の形態素解析には，MeCab*2を利用する．また，関連度計算には tf*idf 法を用いる．

図 3 に得られたネットワークを示す．ノードがレポートを表す．なお，図 3 はネットワークの一部分である．この情報を付加することにより関連レポートの提示が可能となる．

*1 GETA http://geta.ex.nii.ac.jp/
*2 MeCab http://mecab.sourceforge.net/

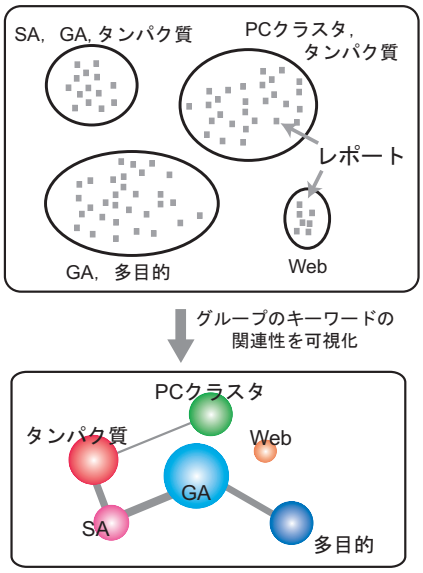

図 2: キーワードの可視化

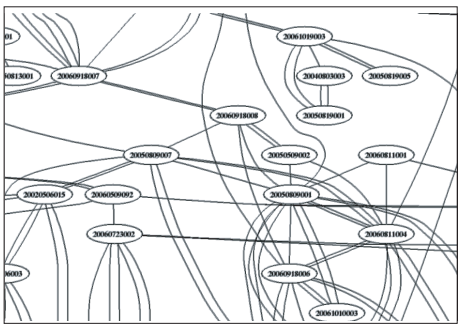

図 3: 研究レポートネットワーク

3.2 クラスタリング方法

3.1 節で得られた関連度を元に，クラスタリング解析を行う．クラスタリング解析には GETA が提供するクラスタリングライブラリを利用する．対象となるデータは 1368 本のレポートで，クラスタ数は 100 に設定した．クラスタリングは下記の手順で行う．

1. 各レポートだけから成るクラスタを作る
2. 全てのクラスタ間の距離を計算する
3. 最も距離の近いクラスタのペアを合併する
4. 合併してできたクラスタと，他のクラスタとの距離を更新する
5. 任意のクラスタ数になるまで 3，4 を繰り返す

本研究ではクラスタ間の距離計算に最短距離法を用いる．また，初期のクラスタ間距離には tf*idf 法で重み付けしたベクトル間の *cos* 値を利用する．

3.3 クラスタリング結果

クラスタリングを行った結果の一部を表 2 に示す．表 2 のように，ディーゼルエンジン燃料噴射の最適化に関するレポートが一つのクラスタとみなされている．このクラスタを表すキーワードとして，噴射，NOx，燃料，ディーゼルエンジン，燃焼

表 2: クラスタリング結果

研究レポートタイトル	クラスタを表すキーワード
近傍培養型遺伝的アルゴリズムを用いたディーゼルエンジン燃料噴射率の多目的最適化 Ver.1	噴射・NOx・燃料・段階・SFC・Soot EGR・ディーゼルエンジン・燃料
設計変数空間における端切り法を用いたディーゼルエンジンの燃焼噴射スケジューリング問題の最適化	
多目的最適化実問題 ーディーゼルエンジン燃料噴射スケジューリング問題ー	
n段階噴射におけるディーゼルエンジンの噴射スケジュールの最適化を実現するためのプログラムの実装と性能調査(1)	
噴射別ディーゼルエンジン燃料噴射スケジュールの最適化の検討（１） など他13件	

などが抽出された．図 3 のネットワークと表 2 のディーゼルエンジンなどをキーワードに持つクラスタを照らし合わせたものを図 4 に示す．実線で囲んだノードが一つのクラスタに分類されており，ネットワーク図でも一つのグループとみなすことができた．図 5 もネットワークとクラスタリング結果を照らし合わせたもので，ネットワーク図では一つのグループだが，クラスタリング結果では，4 つのクラスタで形成されていた．

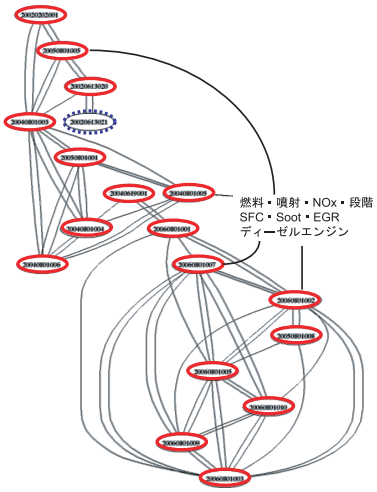

図 4: ネットワークとクラスタリング結果の比較 (1)

3.4 考察

クラスタリングを行いネットワークと比較した結果，ネットワーク図でグループとして確認できる部分が，クラスタとなっている部分もあった．しかし，図 5 のように複数のクラスタに属するレポートが集まって，ネットワーク上で一つのグループになっている部分もあった．これは，図 5 のグループには英語の文献調査のレポートが含まれており，クラスタリングではこれらのレポートが一つのクラスタとみなされていたことが原因だと考えられる．また，クラスタリングして得られたキーワードの中には単語が分割されており，キーワードとして理解が難しいものが多かった．これは，形態素解析に問題があると考えられる．研究のキーワードとなる語は複合語からなる場合も多く，複合語を分割しない形態素解析が必要である．さらに，クラスタリング数やクラスタ間の距離計算のアルゴリズムの検討も必要である．

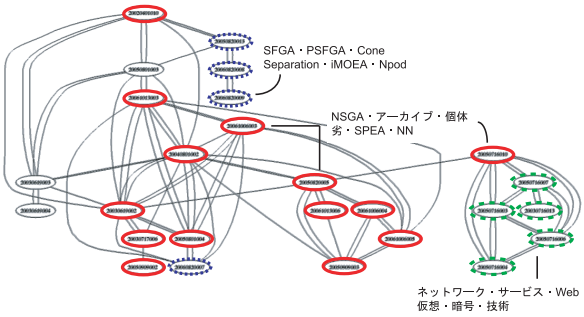

図 5: ネットワークとクラスタリング結果の比較 (2)

4. 今後の展望

研究レポートのネットワークにより，レポート間の関係を把握することが可能となったので，この情報をレポートに組み込み，レポートの閲覧者に提示する必要がある．また，クラスタリング解析により研究レポートのグループ化やキーワード抽出が可能となったので，それらの情報の可視化を行う．また，作成したレポートネットワークは，複雑に関連する部分もあるが，関連が少なく，ネットワークから孤立した部分も多数存在することがわかった．そこで，システムがネットワークを繋ぐハブの役割を果たすようなレポートを推測し，レポートのテーマを提示するレポート作成支援システムを目指す．

5. まとめ

本研究では研究レポート間の関連度よりレポートネットワークを形成し，これらのクラスタリングを行い，クラスタを特徴付けるキーワードを抽出した．クラスタリングの結果，ネットワーク上のグループとクラスタリング結果が合致する部分もあったが，複数のクラスタから成るグループもあった．これらは形態素解析を考慮する必要があり，クラスタ数やクラスタ間の距離に用いるアルゴリズムも検討する必要がある．今後は，クラスタリングから得られたキーワードを可視化することで，研究の全体像を視覚的に把握し，研究にフィードバックできるシステム構築を目指す．また，レポートネットワークを最適化するようなレポートを，システムが推薦することにより研究の活性化を図るシステムも求められる．

参考文献

[1] 高野 明彦，西岡 真吾，今一 修，岩山 真，丹羽 芳樹，久光 徹，藤尾 正和，徳永 健伸，奥村 学，望月 源，野本 忠司
汎用連想計算エンジンの開発と大規模文書分析への応用

[Cutting 92] D.R.Cutting, D.R.Karger, J.O.Pedersen, and J.W.Tukey: Scatter/Gather: a cluster-based approach to browsing large document collections. In Proceedings of the 15th Annual ACM SIGIR Conference on Research and Development in Information Retrieval, pp.318-329 (1992)

[Anderberg 1973] M.R.Anderberg. Cluster Analysis for Applications. Academic Press (1973) 「クラスター分析とその応用」西田英郎監訳，内田老鶴圃 (1988)