2ZF-1

科学技術論文の二層構造化

加藤　俊弥†　　管村　昇‡
工学院大学大学院 工学研究科 情報学専攻†‡

1 はじめに

研究活動を進める上で，関連研究の動向を知るために文献の調査は重要である．しかし，ある文献を理解するためには，他の複数の文献を読まなくてはならない場合が多い．これは初めて研究活動を行う学生などにとって大きな負担となる．このような負担軽減のために，文献検索支援システムが考案されてきた[1][2][3]．しかし，これらシステムでは，キーワードベースの検索が多い，ある特定の専門用語を理解するために，どの文献を読めば良いか直感的に理解しにくい，文献の重要度の指標に，発表年に依存しやすい被引用件数を利用することが多いなどの課題がある．

上記の問題の解決策として，本研究では文献を科学技術論文に絞り，キーワードではなく論文の引用情報を入力とし，同一空間上に論文と専門用語，またそれらの接続関係を提示，論文の重要度に専門用語への接続数を利用することを提案する．本稿では，提案手法のための構成技術の抽出方法とそれを実装した検索システムの概要と検証実験について論じ，手法の有効性を検証する．

2 科学技術論文の二層構造化

まず本研究では，科学技術論文は暗黙的に図 1 のような多層構造をしていると捉える．最上位に論文があり，その論文を複数の構成技術が支え，その構成技術を複数の学問分野が支えているという構造である．この構造を明確に提示することで，論文と構成技術間の意味的な繋がりを明示できる他，その背後にある学問分野への繋がりも明らかにできると考える．

本研究では，多層構造化の前段階として，図 2

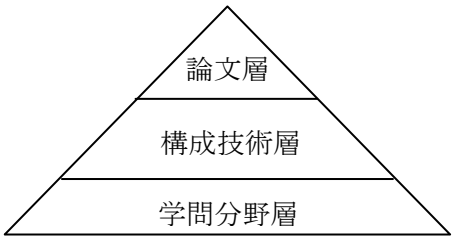


図 1 科学技術論文の多層構造

のように論文層と構成技術層の二層に絞り，引用ネットワークによって繋がれた論文群を論文層，その論文群より抽出された専門用語群を構成技術層とし，同一空間上に配置する．

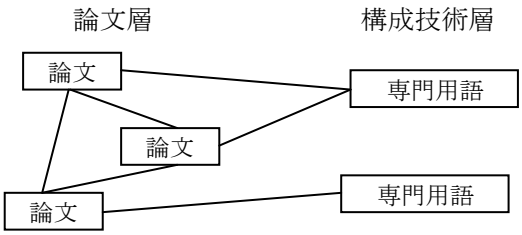


図 2 引用で接続された論文群とその構成技術

3 構成技術の抽出法

構成技術層の作成するため，アブストラクトから構成技術を抽出する．専門用語抽出を目的とした先行研究より「TermExtract」[4]を利用することとした．「TermExtract」は単語の隣接情報を用いて専門用語を抽出し，独自の重要度を算出する．「TermExtract」で抽出した専門用語では，重要度の低い用語ほど一般性が高く，また最も高い重要度の用語は研究カテゴリ自体に近い用語である．一般性が高い用語を削除するため，順位に基づき 21 位以下を削除する．また，研究カテゴリ自体に近い用語は，論文層のほぼすべての論文に対し接続することが予想でき，接続関係を明示する必要がないため，分別をする．研究カテゴリ自体に近い用語の分別には，引用で接続された他の論文の専門用語の重要度をフィルタとして用いた．本研究では，こ

Two-Layer Structure of the Technical Papers
†Toshiya KATO ‡Noboru SUGAMURA
†Department of Informatics,Graduate School Of Engineering Kogakuin University

のフィルタで除去された研究カテゴリ自体に近い用語をマクロ用語，残った用語をミクロ用語と呼ぶこととする．抽出された複数のミクロ用語の表記ゆれなどを吸収するためにクラスタリングを行う．クラスタリングには，レーベンシュタイン距離を文字数で正規化したものを用いて，距離の近いものをクラスタとしてまとめた．この各クラスタを構成技術として配置する．

4 構成技術の妥当性の検証実験

抽出された構成技術(マクロ用語とミクロ用語)の妥当性を評価するため，専門家 2 名によるアブストラクトからの手動での構成技術抽出との比較と，自動で抽出された構成技術内に含まれる無関係な用語を抽出する実験を行った．

対象となる論文群は，論文 5 つからなる論文群と論文 7 つからなる論文群の 2 つである．専門家には各論文のタイトル・著者・発表年月・著者キーワード・アブストラクトを提示し，アブストラクトから手動で専門用語に当たる名詞を抽出させた後，その論文群より自動抽出された構成技術群より，無関係な用語を抽出させた．

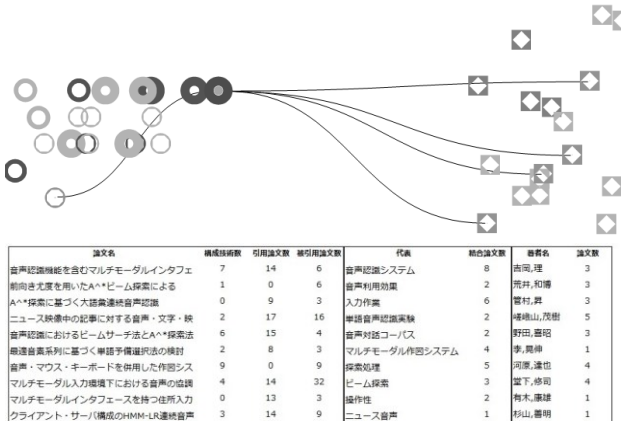
実験の結果，手動で抽出した 58 用語中 27 用語(46.5%)が自動で抽出されていた．また，自動抽出された構成技術群のうち，無関係な用語と判定されたものは，75 用語中 16 用語で，全体の 21.3%であった．最低限重要な用語は抽出出来ていると考えられる．

自動で抽出できなかった構成技術については，フィルタリングにより除去されているものが多かった．また，複数の論文で出現するが，手動では抽出されなかった無関係ともいえない用語が約 30%あり，これらの用語を抽出できたことは利点であると考えられる．

5 検索システムへの実装と評価

提案手法を実装した検索システムの構築を行い，評価を行った．検索システムは，図 3 のように左側に論文層と右側に構成技術層にそれぞれノードを配置し，論文層のノードは引用情報で結ばれ，論文層と構成技術層のノードは，論文に構成技術が出現する場合に結ぶものとした．論文層のノードは縦軸に時間を取り，発表年が古いものほど上に配置され，被引用件数が多いものは大きく表示し，構成技術層のノードへの接続数が多いものは色を濃く表示した．また，論文と構成技術と著者について別途，表による提示も行った．

システムの評価として，研究経験の浅い学生 6 名を対象に，検索システムを一通り操作した後，改善点や今後利用したいかを尋ねるアンケートを実施した．ユーザビリティに関して否定的な意見はあったが，6 名ともこのような検索システムを利用したいと答えた．



論文名	構成技術数	引用論文数	被引用論文数
音声認識機能を含むマルチモーダルインタフェ	7	14	6
前向き尤度を用いたA*ビーム探索による	1	0	6
A*探索に基づく大語彙連続音声認識	0	9	3
ニュース映像中の記事に対する音声・文字・映	2	17	16
音声認識におけるビームサーチ法とA*探索法	6	15	4
最適音素系列に基づく単語予備選択法の検討	2	8	3
音声・マウス・キーボードを併用した作図シス	9	0	9
マルチモーダル入力環境下における音声の位置	4	14	32
マルチモーダルインタフェースを持つ住所入力	0	13	3
クライアント・サーバ構成のHMM-LR連続音声	3	14	9

代表	結合論文数
音声認識システム	8
音声利用効果	2
入力作業	6
単語音声認識実験	2
音声対話コーパス	2
マルチモーダル作図システム	4
探索処理	5
ビーム探索	3
操作性	2
ニュース音声	1

著者名	論文数
吉岡,理	3
荒井,和博	3
曽村,昇	3
嵯峨山,茂樹	5
野田,喜昭	3
李,晃伸	1
河原,達也	4
堂下,修司	4
有木,康雄	1
杉山,雅明	1

図 3 実装したシステムの概形

6 まとめと今後の課題

本稿では，論文検索支援のための多層構造化の前段階として，二層構造化を行い，二層構造化を実装したシステムの評価を行った．構成技術層を作成するための，構成技術抽出では専門家による手動抽出と比較して，約半数の用語がされた．実装したシステムについて，肯定的な評価を得られた．今後の課題として，本研究の本来の目的である，多層構造化が最も重要であり，また構成技術抽出の精度の向上が必要である．

参考文献

[1]鈴木 雅人．リッチインターフェースを備えたグラフィカル論文検索支援システム．ヒューマンコンピュータインタラクション研究会報告 2008

[2]杉本 雅則，小山 照夫，掘 浩一，大須賀 節雄，絹川 博之，間瀬 久雄．文書間の関連性を可視化することによる文献検索システム．自然言語処理研究会報告，1996

[3] 謝 英双，三末 和男，田中 二郎．キーワードの頻度推移と文献の被引用数を視覚化した文献検索ツール．情報処理学会全国大会，2010

[4]TermExtract http://gensen.dl.itc.u-tokyo.ac.jp/(2014/1/13 アクセス)