6ZH-03

キーワードと参照構造に基づいた高精度な論文推薦システムのモデルの提案

吉田裕平† 児玉英一郎‡ 王家宏‡ 高田豊雄‡

岩手県立大学大学院ソフトウェア情報学研究科† 岩手県立大学ソフトウェア情報学部‡

1. はじめに

Webを活用して，必要な学術論文を入手する人が増えている．現在，論文データベースをWeb上から検索できるようにしたものとして，CiNii Articlesや情報処理学会電子図書館などが存在している．これらを利用することによって論文の入手が行える．

一方で，毎年新しい研究や論文が現れており，論文の年間発行件数は年々増加傾向にある．文部科学省の「論文分析でみる世界の研究活動の変化と日本の状況」[1]によると，日本の研究機関が発表した論文数は，1981年に約2万5千件であったものが，2015年には約7万6千件へと増加し，約3倍となっている．

学部4年生など研究に初めて取り組む学生は，このように膨大な数の論文の中から興味のある論文を見つけ，最新の研究動向を学ぶことが適切であると考えるが，論文の総数が増えているため，すべての論文を見るのは難しく，支援が必要と考える．

そこで，本研究では，膨大な数の論文の中から最近の論文や関連性のある近いテーマの論文を学生へ推薦することを目的として，キーワードと参照構造に基づいた論文推薦システムのモデルの提案を行う．

2. 関連研究

論文発見に関する研究として，難波らの研究[2]が知られている．難波らは，論文データベースに対し，HITSというWebページのランク付けアルゴリズムを適用し，サーベイ論文の自動検出を試みている．HITSとはauthorityとhubの2つの概念から重要性の高いWebページを検出するアルゴリズムである．このとき，authorityとは検索キーワードに関する重要ページのことであり，hubは優秀なauthorityを数多くリンクしているページのことを指している．この研究内では，authorityは「他の論文から多く参照されている論文」，hubは「それらの論文を多く参照している論文」のことを指している．また，難波らの研究を応用したものとして，井坂らの研究[3]が知られている．井坂らは，論文の参照構造(参考文献の集合を利用)からリンク構造を作成し，ノードをランク付けすることによって，初めて論文サーベイを行うユーザを支援するシステムの研究を行った．井坂らの研究では，参照構造だけでリンク構造を作成しているため，論文全体を年代別に並べると，参考文献の性質によって，一般的には，その論文の発行年よりも前の年方向へのリンク（過去方向へのリンク）によってリンク構造が生成される．

A Proposal of High-Precision Scientific Paper Recommendation System Model Based on Keywords and Reference Structure

Yoshida Yuhei† Kodama Eiichiro‡ Wang Jiahong‡ Takata Toyoo‡

Graduate School of Iwate Prefectural University†

Iwate Prefectural University‡

これに対して，HITSを適用すると，hub度の高いページから参照されているauthority度の高いページが高得点となる．すなわち，ある特定の論文から参照されている論文が高得点となり，その論文の発表年よりも古い論文のみが高得点となってしまうという問題がある．図1に井坂らの手法で作成されるリンク構造の例と高得点の論文の例を示す．

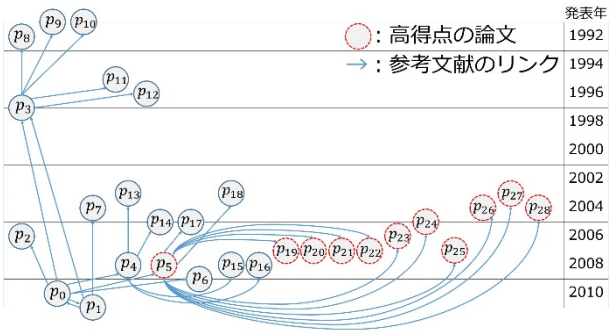

図1 井坂らの手法で作成されるリンク構造の例

図1中のp_0は，「Linked Open Dataによる多様なミュージアム情報の統合」という2010年に発表された論文である．これに対し，HITSを適用すると，p_5のhub度が高くなり，p_5から参照されているp_{19}からp_{28}のauthority度が高くなる．このためp_5の発表年である2008年以前の論文p_{19}からp_{28}が高得点となる．

3. キーワードと参照構造に基づいた高精度な論文推薦システムのモデル

本研究では，本提案のキーワードと参照構造に基づいた高精度な論文推薦システムのモデルを提案する．本提案では，井坂らの研究で作成した参照構造に，論文同士の持つキーワードによるリンクも付与し，参照構造を構築する．これにより，論文データベースにある参照構造に基づいたリンク以外のリンクを張ることができ，論文の発表年順で見ると，古い年から新しい年方向へのリンク（未来方向へのリンク）も張ることができる．これにより，最新の論文であってもauthority度が高くなる可能性が生じる．また，hub度が高い論文が分散し，関連性のある近いテーマの論文が選ばれやすくなると考える．

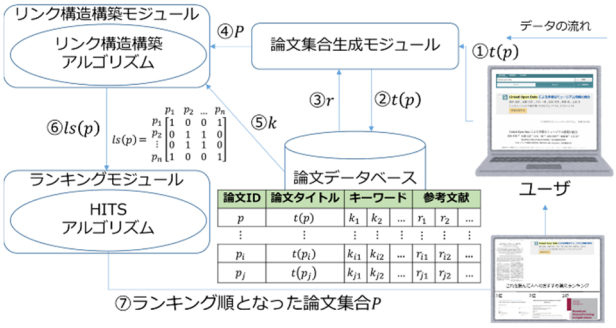

図2 本提案の論文推薦システムのモデル

本提案のキーワードと参照構造に基づいた高精度な論文推薦システムのモデルを図2に示す．ここで，論文（論文ID）をpとし，その論文pのタイトルを$t(p)$，論文pの持つキーワードの集合を$kw(p)$，論文pの持つ参考文献の論文（論文ID）の集合を$ref(p)$，論文pを基点として構築されるリンク構造の隣接行列を$ls(p)$とする．以下，図2の各構成要素の詳細を示す．

・論文集合生成モジュール

ユーザがあらかじめ選択した1つの論文pのタイトル$t(p)$を取得し，$t(p)$を基に論文データベースから，論文pが持つ参考文献の論文r（論文ID）を取得する．参考文献の論文r（論文ID）から，同じように参考文献の論文r'（論文ID）を取得し，同様の処理をn回（指定された回数）繰り返すことで参考文献の論文r（論文ID）から成る論文集合Pを生成する．なお，処理回数のnは，ユーザの目的に合わせて，設定するものとする．

・リンク構造構築モジュール

論文集合生成モジュールによって生成された論文集合Pの各ノード間にリンク構造の隣接行列$ls(p)$を構築する．以下，リンク構造を構築する際のアルゴリズムを示す．ここで，論文pの持つキーワード$k \in kw(p)$は論文データベースから取得するものとする．次のステップにより，Pの要素間のリンク構造を構築する．

Step1．$p_i, p_j \in P$に対して，$p_j \in ref(p_i)$ならば，p_i, p_jをノードとし，p_iからp_jへリンクを張る．

Step2．$m = i, j$とするとき，各$k_{ml} \in kw(p_m)$に対して，キーワードオントロジを利用して，k_{ml}の類義語集合$\overline{k_{ml}}$を取得し，拡張キーワード集合$\overline{kw(p_m)} = kw(p_m) \cup \left(\bigcup_{l=1}^{|kw(p_m)|} \overline{k_{ml}}\right)$を生成する．

Step3．$p_i, p_j \in P$に対して，θを閾値とし，本研究独自のJaccard接続係数$\mathrm{JaccardContinuation}(p_i, p_j)$を式(1)により算出後，$\mathrm{JaccardContinuation}(p_i, p_j) > \theta$のとき，$ref(p_j) \not\ni p_i$ならば，$p_j$から$p_i$へリンクを張り，また，$ref(p_i) \not\ni p_j$ならば，$p_i$から$p_j$へリンクを張る．

$$\mathrm{JaccardContinuation}(p_i, p_j) = \left|\overline{kw(p_i)} \cap \overline{kw(p_j)}\right| / \left|kw(p_i) \cup kw(p_j)\right| \quad (1)$$

図3に本提案手法で作成されるリンク構造の例を示す．

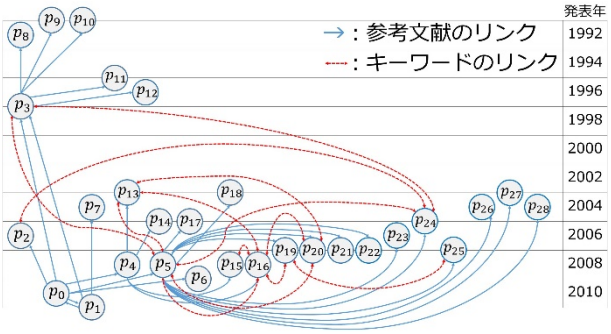

図3　本提案手法で作成されるリンク構造の例

・ランキングモジュール

リンク構造構築モジュールによって完成したリンク構造の隣接行列$ls(p)$を用いてHITSアルゴリズムを適用する．HITSアルゴリズムによって，リンク構造内のノードである論文はランキング化される．そして，最終的にこれらの論文は，ランキング順に並び替えられ，ユーザに提示される．

4.　評価

・評価目的

本評価では，本提案モデルの精度を適合率・再現率・F値によって評価する．

・評価方法

評価の準備として，正解論文の決定を行った．評価対象となる論文（計10個）を収集し，各論文に対し，3階層まで参考文献を辿る作業を行った．その後，参考文献の概要部分を全て被験者に提示し，正解論文を決定した．その際は，表1に示す基準を提示し，5人中3人以上がAとした論文を正解論文とした．なお，この基準は林らの研究[4]で使用されたものである．

表1　正解論文を決定する際の基準

	基準
A	重要なトピックについて言及しており，論文の参考文献として記載するのにふさわしい論文である
B	論文の参考文献として記載するのにふさわしくないが，論文を執筆する際にある程度は参考になる
C	内容が完全に異なり，まったく参考にならない

次に，本提案モデルに従い推薦された論文と正解論文を利用し，適合率・再現率・F値の算出を行った．評価用の計10論文に対し，井坂らの手法によってリンク構造（10リンク構造）を構築し，本提案手法によるリンク構造（10リンク構造）も構築した．このようにして作成した計20リンク構造に対し，HITSアルゴリズムを適用し，authority値とhub値の平均値を求め，その平均値によって各論文をランキング化した．そしてノード数の1/3（つまりランキング上位）の値をnとし，上位n件の適合率・再現率・F値を算出し，比較した．

・評価結果

井坂らの研究での適合率は0.379，再現率は0.482，F値は0.420であった．また，本提案手法を用いた場合の適合率は0.691，再現率は0.877，F値は0.766であった．本提案手法において，論文同士の持つキーワードの高度な活用を行わない場合の適合率は0.615，再現率は0.782，F値は0.683であった．また，ランキング上位をノード数の1/3以内と設定せずに，authority値とhub値の平均値を近い値で区切り，上位グループをランキング上位とした場合の適合率は0.688，再現率は0.903，F値は0.774であった．

5.　おわりに

本研究では，論文の持つキーワードと参照構造に基づいた高精度な論文推薦システムのモデルの提案を行った．本提案手法の有用性確認のため評価を行い，適合率0.691，再現率0.877，F値0.766であることを確認した．

参考文献

[1] 文部科学省 科学技術・学術政策研究所：科学研究のベンチマーキング 2017 ~論文分析でみる世界の研究活動の変化と日本の状 況 http://www.nistep.go.jp/wp/wp-content/uploads/NISTEP-RM262-FullJ.pdf, (参照 2019-5-24).

[2] 難波英嗣，奥村学 ：多言語論文データベースを用いたサーベイ論文検出：サーベイ論文自動作成の実現に向けて，電子情報通信学会技術報告. NLC，言語理解とコミュニケーション, Vol.102, No.119, pp.35--41 (2002).

[3] 井坂徳恭，中山泰一：重要論文検索システムIaskの実装と評価，情報処理学会研究報告，2011-CE-109, pp.1--8 (2011).

[4] 林佑磨，奥野峻弥，山名早人：単語の意味概念行列を用いたキーワード生成による関連論文検索システム，情報処理学会研究報告，Vol.2014-IFAT-115, No.10, pp.1--6 (2014).