2ZJ-04

キーワードと参照構造に基づいた論文発見手法の提案

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

必要な学術論文を入手する方法の一つとしてWeb の活用が考えられる．現在，論文データベースを Web 上で検索可能とした CiNii Articles や情報処理学会電子図書館などが存在しており，それらを利用することによって論文の入手が行えている．

しかし，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増加傾向にある．文部科学省の「解説　論文成果に見る我が国の状況」[1]によると，日本の研究機関が発表した論文数は，1988 年に約 4 万件であったものが，2008 年には約 7 万件へと増加し，約 1.75 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困難であると考える．

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発見手法の提案を行う．

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．

An Approach to Searching for Research Papers Using Keywords and Citation Structure
Yoshida Yuhei† Kodama Eiichiro‡ Wang Jiahong‡ Takata Toyoo‡
Graduate School of Iwate Prefectural University†
Iwate Prefectural University‡

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，過去方向へのリンクのみでリンク構造が生成される．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．

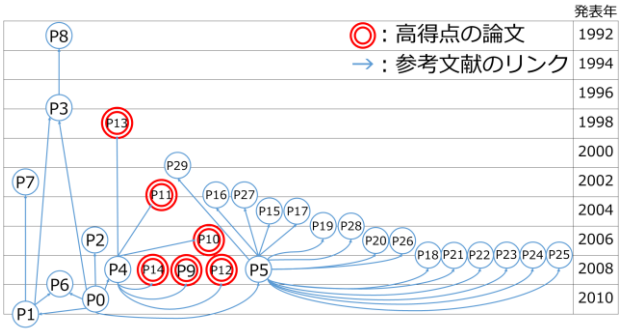

図 1　井坂らの手法で作成されるリンク構造の例

図 1 中の P0 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，P4 の hub 度が高くなり，P4 から参照されている P9 から P14 の authority 度が高くなる．このため P4 の発表年である 2008 年以前の論文 P9 から P14 が高得点となる．

3. 予備調査

予備調査では，学生が推薦されて嬉しい論文を明確にするため，ソフトウェア情報学部の 4 年生(5 名)を被験者として，P0 から P29 までの論文を提示し，推薦されて嬉しい論文の調査を行った．5 名中 3 名以上が推薦されて嬉しいと答えた論文を黒いノードとして図 2 に示す．図 2 から分かるように，最近の論文 P1 や関連性のある近いテーマの論文 P5，P23 などが選ばれている．

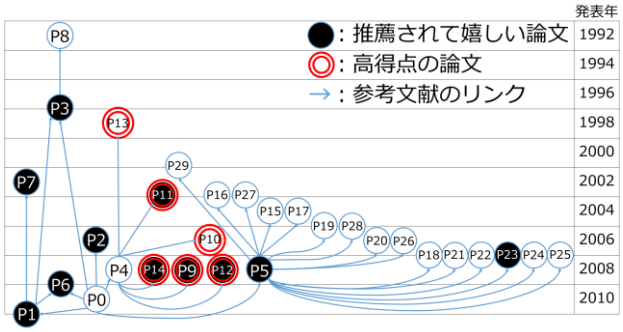

図 2　学生が推薦されて嬉しい論文

4.　キーワードと参照構造に基づいた論文発見手法の提案

本研究では，キーワードと参照構造に基づいた論文発見手法を提案する．この手法では，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する．これにより，論文データベースにある参照構造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．

本提案の，キーワードと参照構造に基づいた論文発見手法を以下に示す．

(1) 下記のキーワードと参照構造に基づいたリンク構造構築アルゴリズムに従い，リンク構造を構築する．

(2) 構築したリンク構造に対し，HITS を適用し，論文のランク付けを行い，このランクに従い，論文を発見する．

・キーワードと参照構造に基づいたリンク構造構築アルゴリズム

考察対象の論文集合をPとし，$p_i \in P$に対して$ref(p_i)$をp_iの参考文献の集合，$kw(p_i)$をp_iの持つキーワードの集合とする．このとき，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．

(1) 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．
(2) p_i，$p_j \in P$に対して，θを閾値とし，
$\mathrm{Jaccard}(p_i, p_j)$
$= |kw(p_i) \cap kw(p_j)| \,/\, |kw(p_i) \cup kw(p_j)| > \theta$のとき，
$ref(p_j) \ni p_i$ならば，p_jからp_iへリンクを張り，また，
$ref(p_i) \ni p_j$ならば，p_iからp_jへリンクを張る．

図 1 の論文集合に対し，本提案アルゴリズムを適用した場合のリンク構造を図 3 に示す．

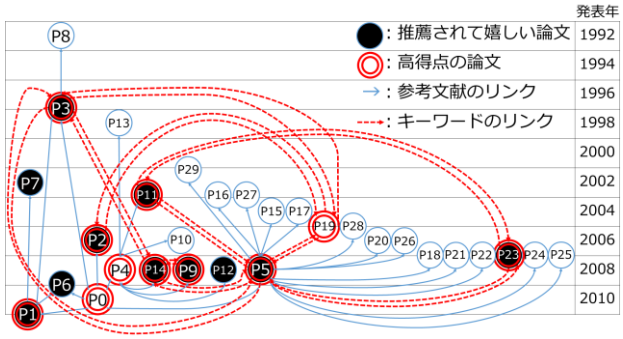

図 3　本提案手法で作成されるリンク構造の例

5.　評価

・評価目的と評価方法

本提案手法の有用性確認のための評価実験を行った．評価目的は，本提案の有用性を適合率，再現率，F 値によって確認することである．評価に用いる正解論文は，被験者 5 人に対して，評価対象の論文(10 個)とその参考文献を辿って求めた論文を提示し，3 人以上が推薦されて嬉しいと判断した論文とした．評価対象の論文に対して井坂らの研究と本提案手法のリンク構造をそれぞれ作成し，それに HITS を適用した．論文のランク付け後，ノード数の 1/3 の値をnとするときの上位n件の適合率，再現率，F 値を算出し，比較を行った．

・評価結果

井坂らの研究での適合率は 0.36，再現率は 0.47，F 値は 0.40 となった．また，本提案手法を用いた場合の適合率は 0.59，再現率は 0.78，F 値は 0.67 となった．

6.　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論文発見手法の提案を行った．本提案手法の有用性確認のため評価を行い，適合率 0.59，再現率 0.78，F 値 0.67 であることを確認した．

参考文献

[1] 文部科学省「解説　論文成果に見る我が国の状況」 http://www.mext.go.jp/b_menu/hakusho/html/hpaa201001/detail/1296363.htm (20171205)

[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)．