

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と評価

吉田裕平¹ 児玉英一郎² 王家宏² 高田豊雄²

概要：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年々増加傾向にある。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ているが、論文の総数が増えているため、すべての論文を見ることは困難であり、支援が必要と考える。このような視点に基づいた関連研究として、論文の参照構造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研究[1]が知られている。しかし、この研究では、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た場合、過去方向へのリンクのみでリンク構造が生成され、古い論文が中心として推薦されてしまうといった問題があった。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。本提案では、関連研究の手法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を付与する手法を用いており、この点に特徴がある。この手法によって、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を張ることが可能となる。また、本提案システムの有用性確認のため評価を行ったので、その評価結果についても報告する。

キーワード：論文推薦システムのモデル、論文推薦、リンク構造、参照構造

Proposal and Evaluation of Scientific Paper Recommendation System Model Based on Keywords and Reference Structure

YUHEI YOSHIDA^{†1} EIICHIRO KODAMA^{†2}
WANG JIAHONG^{†2} TOYOO TAKATA^{†2}

Abstract: One way to access scientific papers is by the Web. As examples of the Web scientific paper databases, in Japan CiNii Articles and IPSJ Electronic Library are available, and papers can be obtained with using them. Year by year, new research findings appear, and the number of scientific papers published annually is increasing. People such as the fourth-year undergraduate students who are engaged in research work for the first time may encounter difficulties in finding appropriate papers from such a large number of published ones, and thus, could not learn the latest research trend in time. Paper recommendation system become necessary for them. For supporting the people who conduct a paper survey for the first time, a method was proposed in paper [1] which creates a link structure of nodes representing papers from the paper reference structure and ranks the nodes. However, since the link structure is created only by the reference structure, nodes are arranged by the publication years, the link structure is generated in the direction of the publication years, so the older papers are recommended mainly. For the purpose of recommending students to recent scientific papers and papers with similar themes, we propose a recommendation system model of scientific papers, which is based on both the keyword and the reference structure. The links of the keywords of papers have been added to the reference structure created by the method proposed in [1]. Thus, paper databases would have links other than the reference structure, and the link structure is in the future direction, with new paper recommended first. We have evaluated the effectiveness of the proposed system model and report the evaluation results.

Keywords: Scientific Paper Recommendation System Model, Scientific Paper Recommendation, Link Structure, Reference Structure

¹ 岩手県立大学大学院ソフトウェア情報学研究科
Graduate School of Software and Information Science, Iwate Prefectural University

² 岩手県立大学ソフトウェア情報学部
Faculty of Software and Information Science, Iwate Prefectural University

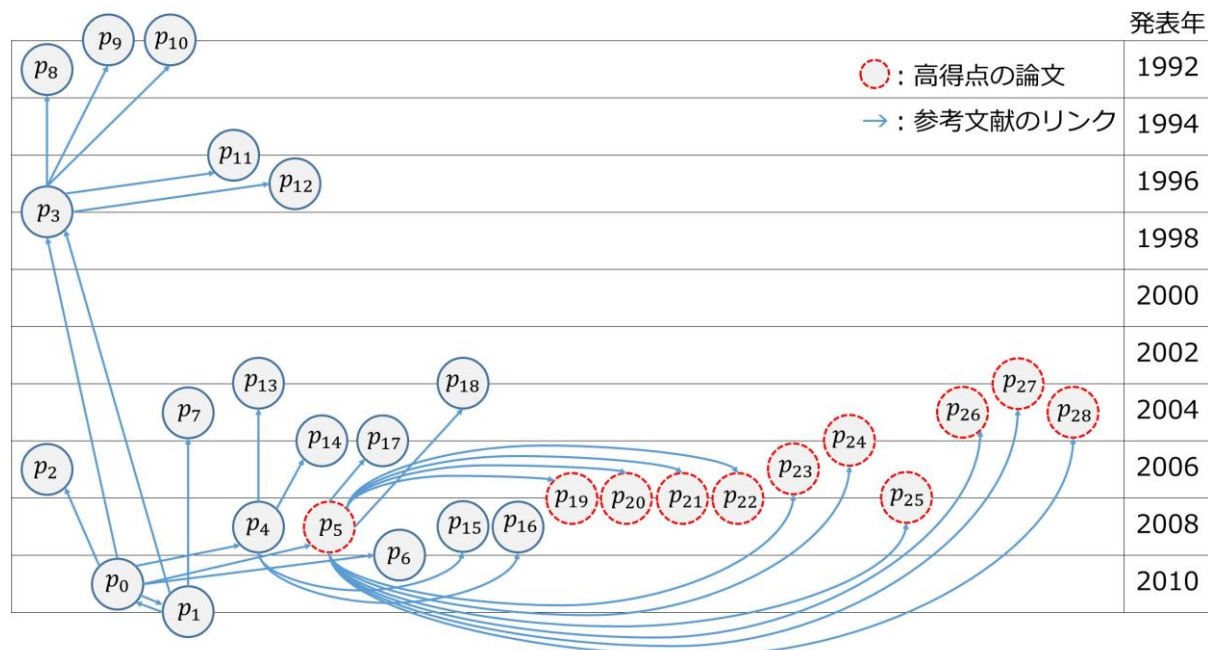


図 1 井坂らの手法で作成されるリンク構造
Figure 1 Link structure created by related research

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」[2]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、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り、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と評価を行う。

2. 関連研究

論文発見に関する研究として、難波らの研究[3]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[1]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 Iask の研究を行った。

Iask の論文推薦手順は以下の通りである。

- ユーザの興味の持った論文の ACM Portal（論文検索サービス）における URL を入力する
- 入力された論文の URL から参照構造（参考文献の集合）を取得・構築する
- HITS（または PageRank）を用いて参照構造内のノードをランク付ける
- ランク付けされた論文のうち、上位 10 件をユーザに推薦する

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、その論文の発行年からそれ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。これに対して、HITS を適用し、hub 度の高いページから参照されている authority 度の高いページを高得点とする。すなわち、ある特定の論文から参

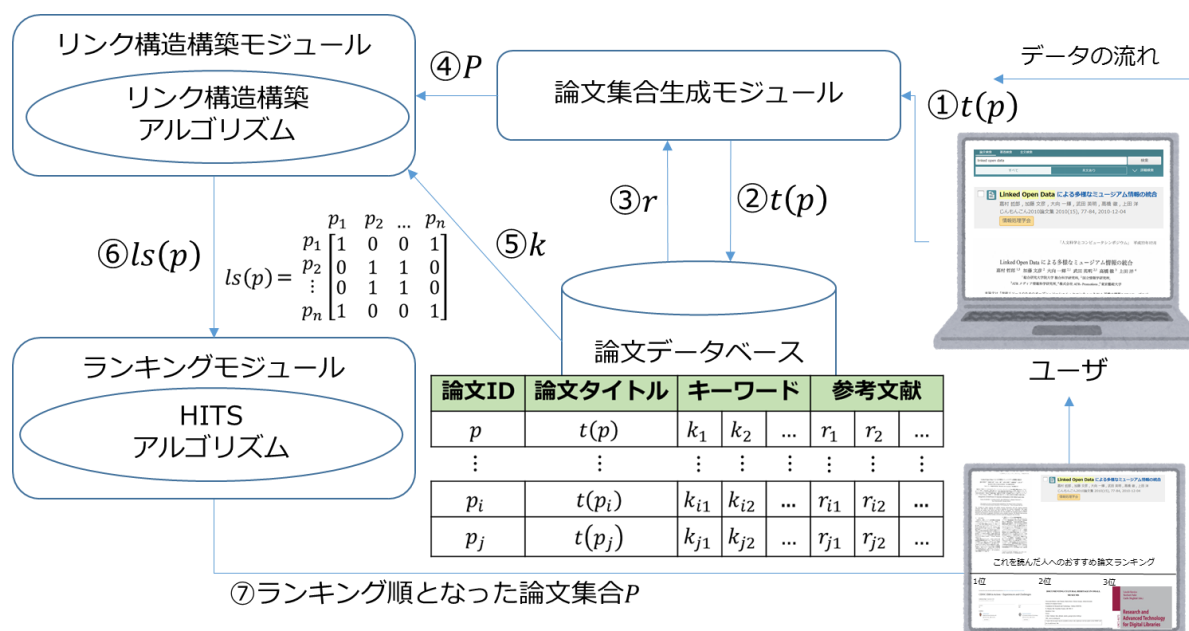


図 2 本提案の論文推薦システムのモデル
Figure 2 Scientific paper recommendation system model

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に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 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論文推薦システムのモデルを提案する。この提案では、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。

本提案の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。

3.1 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、設定することが可能である。

3.2 論文データベース

論文 p の持つキーワード $k \in kw(p)$ 、論文 p の持つ参考文献の論文 $r \in ref(p)$ （論文 ID）などの情報を保存するデータベースである。

3.3 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 の各ノード間のリンク構造に対し、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.

$p_i, p_j \in P$ に対して、 θ を閾値とし、 $Jaccard(p_i, p_j) = |kw(p_i) \cap kw(p_j)| / |kw(p_i) \cup kw(p_j)| > \theta$ のとき、 $ref(p_j) \not\supset p_i$

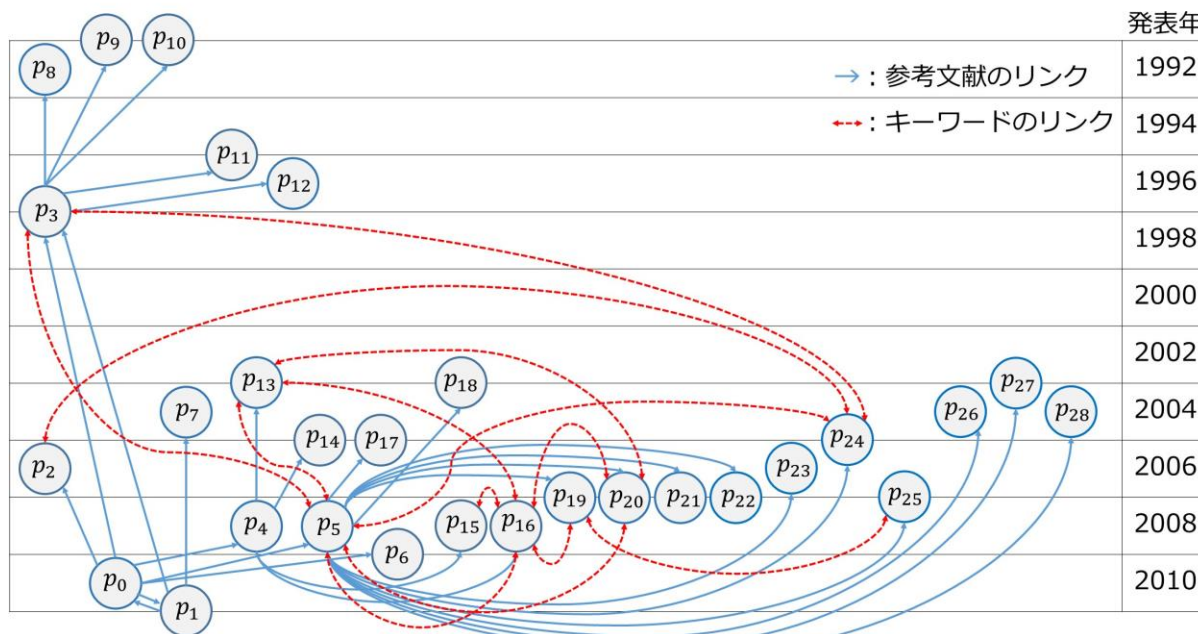


図 3 本提案手法で作成されるリンク構造
Figure 3 Link structure created by the proposed method

p_i ならば、 p_j から p_i へリンクを張り、また、 $ref(p_i) \neq p_j$ ならば、 p_i から p_j へリンクを張る。

図 3 に本提案手法で作成されるリンク構造の例を示す。

3.4 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。

3.5 システムの利用例

本提案システムの利用例を図 4～図 7 に示す。

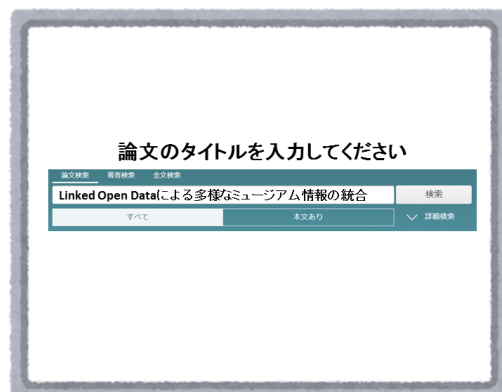


図 5 タイトルの入力画面 (2)

Figure 5 Search screen

図 4 と図 5 に示すように、ユーザはあらかじめ選択した 1 つの論文のタイトルを入力する。



図 4 タイトルの入力画面 (1)

Figure 4 Search screen

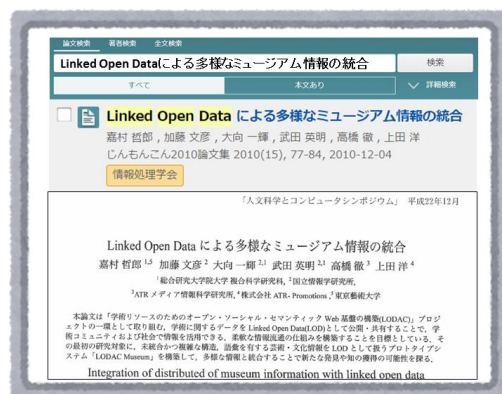


図 6 検索結果の画面 (1)

Figure 6 Search results

表 1 評価に使用したデータ
Table 1 Data used for evaluation

	タイトル	キーワード (抜粋)	発行年	参照構造の 論文数 (3 階層)	参照構造内に 含まれる 正解論文数
①	Linked Open Data による 多様なミュージアム情報の統合		2010	29 個	7 個
②	スポーツジャンルに応ずる ネタバレ特性分析と判定手法の提案	ネタバレ防止, 機械学習, スポーツ, Twitter	2016	25 個	6 個
③	翻訳機能付きのチャットシステム		2012	21 個	7 個
④	CUD に基づいた Web ページの 再ランキング手法に関する研究	情報検索, 文書分類	2017	27 個	8 個
⑤	Web 上に混在する観光情報を活用した 観光地推薦システム	観光地推薦システム, 集合知, 類似度	2012	30 個	8 個
⑥	多言語論文データベースを用いた サーベイ論文検出	サーベイ論文検出, HITS アルゴリズム, 多言語論文データベース,	2002	22 個	6 個
⑦	2D デザインデータのみで 3DCG 化する 自動 3DCG 生成システムの開発		2018	4 個	2 個
⑧	ファイル更新の時系列に着目した フォルダのクラスタリング手法	クラスタリング, アクセス履歴, ファイル整理	2018	28 個	8 個
⑨	一般のオノマトペを対象とした 音象徴的意味による分類手法に関する研究	情報検索, 文書分類	2017	17 個	5 個
⑩	ICT を活用した創造型ものづくり教育の 導入と授業実践	ものづくり, 情報通信技術, 実践教育, プロジェクトベースの学習	2018	12 個	2 個



図 7 検索結果の画面 (2)

Figure 7 Search results

その後, 図 6 のように, 元となる論文が表示される. 元となる論文をクリックすると, 各モジュールによって, 論文集合の生成, リンク構造の構築, ランキング化が実行される.

そして, 図 7 に示したように, ランキング化された論文がユーザに提示される.

4. 評価実験

4.1 評価目的

本評価の目的は, 本提案モデルに基づいて作成したシステムを用いて, その精度を適合率・再現率・F 値によって評価することである.

4.2 評価方法

4.2.1 正解論文の決定

評価対象となる論文 (計 10 個, 表 1 参照) 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う. 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 2 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て設定した. 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 2 正解論文を決定する際の基準

Table 2 Criteria for correct data

	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

表3 井坂らの手法での評価結果

Table 3 Evaluation results of related research

	適合率	再現率	F 値	正解論文の 含有率
論文①	0.30	0.43	0.35	10 位内に 3 個
論文②	0.44	0.67	0.53	9 位内に 4 個
論文③	0.43	0.43	0.43	7 位内に 3 個
論文④	0.22	0.25	0.23	9 位内に 2 個
論文⑤	0.40	0.50	0.45	10 位内に 4 個
論文⑥	0.38	0.50	0.43	8 位内に 3 個
論文⑦	0.50	0.50	0.50	4 位内に 1 個
論文⑧	0.40	0.50	0.45	10 位内に 4 個
論文⑨	0.33	0.40	0.36	6 位内に 2 個
論文⑩	0.25	0.50	0.33	4 位内に 1 個
平均値	0.37	0.47	0.41	

4.2.2 適合率・再現率・F 値の算出

表 1 の計 10 論文に対し、井坂らの手法によってリンク構造を構築する (10 リンク構造)。また、本提案手法によるリンク構造も構築する (10 リンク構造)。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 値を算出した。

4.3 評価に使用したデータ

評価実験に使用したのは、表 1 に示す 10 個の論文である。これらの論文はいずれも情報分野の内容に関するものであり、本実験を行うにあたり無作為に選出したものである。各論文の発行年は、直近 5 年以内に発行されたものが半数以上を占めるが、中には 2002 年に発行されたものなど、比較的古い論文もある。また、各論文の 3 階層までの参照構造内に含まれる論文数の平均は 22 個、参照構造内に含まれる正解論文数の平均は 6 個であった。

4.4 評価結果

井坂らの手法を用いた場合の評価結果は表 3 の通りである。適合率は 0.37、再現率は 0.47、F 値は 0.41 となった。

また、本提案モデル内の手法を用いた場合の評価結果は表 4 の通りである。適合率は 0.61、再現率は 0.78、F 値は 0.68 となった。

4.5 評価結果の考察

本提案手法によって、未来方向のリンクが付与され論文の hub 度が分散し、井坂らの手法では点数が低かったが本提案手法では高得点となる論文が出現した。その点が適合率や再現率の向上に寄与したと考えられる。

また、被験者によって決定された正解論文は、発行年が新しい論文やページ数が多い論文などに限らず、古い論文であっても、その研究分野の内容について基礎から網羅的

表4 本提案モデル内の手法での評価結果

Table 4 Evaluation results of the proposed method

	適合率	再現率	F 値	正解論文の 含有率
論文①	0.50	0.71	0.59	10 位内に 5 個
論文②	0.67	1.00	0.75	9 位内に 6 個
論文③	0.57	0.57	0.57	7 位内に 4 個
論文④	0.67	0.75	0.71	9 位内に 6 個
論文⑤	0.70	0.88	0.78	10 位内に 7 個
論文⑥	0.50	0.67	0.57	8 位内に 3 個
論文⑦	1.00	1.00	1.00	4 位内に 2 個
論文⑧	0.50	0.63	0.56	10 位内に 5 個
論文⑨	0.50	0.60	0.55	6 位内に 3 個
論文⑩	0.50	1.00	0.67	4 位内に 2 個
平均値	0.61	0.78	0.68	

に書かれてある論文などが選ばれていた。

正解論文でありながら、推薦漏れとなった論文としては、正解論文に付与されていたキーワードが独特で、他の論文には付与されていないようなものであったために、キーワードによるリンクが張られず、高得点とならない論文があった。一方、正解論文ではないのに、誤推薦される論文としては、その論文に付与されたキーワードが、他の論文にも多く付与される一般的なものであったために、キーワードによるリンクが張られてしまい、高得点となってしまった論文が挙げられる。

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論文推薦システムのモデルの提案を行った。また、本提案モデルに基づいて作成したシステムを用いて、その精度を確認のため評価を行い、適合率 0.61、再現率 0.78、F 値 0.68 であることを確認した。

今後の課題としては、プロトタイプシステムの実装を完成させ、基本性能評価、作業手順数に関する評価、インタフェースの評価、正解論文の決定方法に関する検証を行うことである。

参考文献

- [1] 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).
- [2] 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).
- [3] 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, Vol.102, No.119, NLC2002, pp.35-41 (2002).
- [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).