2H4-1

専門用語の出現に基づく論文の重要度の分析

Significance Analysis of Papers based on Occurrence of Terms

岡嶋 穣[*1] 松尾 豊[*2] 石塚 満[*3]
Yuzuru Okajima Yutaka Matsuo Mitsuru Ishizuka

[*1*3]東京大学大学院 情報理工学系研究科
Graduate School of Information Science and Techonology, The University of Tokyo

[*2]独立行政法人 産業技術総合研究所
National Institute of Adcanced Industrial Science and Technology

Recently there are many datasets where entities are connected by links and the importance of entities are measured by the link structure. Examples include academic papers, in which the importance of a paper is measured by the number of citations. Because it is sometimes difficult to obtain the link information, we can use a set of features as a surrogate of the link information. In this paper, we propose a new method to measure the link-based importance of entities without using explicit link information. Features which co-occurs with the existence of links are extracted, and added as nodes to a network composed of entities. Our graph-based ranking algorithm can estimate the importance of the entities and the features simultaneously. We conducted an experiment on the scientific papers dataset and compared the results with an existing algorithm.

1. はじめに

今日では、相互にリンクによって関連付けられたエンティティから成るデータの量が、ますます増大する傾向にある。その代表的な例はウェブである。ウェブ上では各ページの重要度はページ間のリンク情報を用いて重要度が計算される。PageRank[1] や HITS[3] などのランキングアルゴリズムである。論文の重要度もリンク情報を用いて計算されるもののひとつである。論文の重要度を測るもっとも基本的な手法は、その被引用数、つまり他論文からリンクされた数を測ることである。

このようにリンク情報が重視され、リンク情報に基づいてエンティティの重要度が決定される一方、必ずしもリンクの情報が全て与えられるとは限らない。例えば発表されたばかりの論文や投稿されたばかりのブログなどは、他の論文や記事から十分なリンクが張られておらず、リンク情報のみを用いて重要度を測定することが難しい。そこで、このような、リンク情報を持っていないエンティティに対しても、リンク情報に基づく重要度がどのような値になるかを推測することができれば、非常に有用な情報となる。例えばエンティティが論文ならば、論文に含まれる語の頻度などが素性として使え、ある語が含まれていればリンクされやすいなどの傾向が分かれば、語がリンク情報を予測する手がかりとなる。しかし、文書に含まれる語は雑多であり、どの語がリンクされた理由なのか判断することは難しく、また、リンクする側の素性との組み合わせによって決定されている場合もあり、多くの難しい問題をはらんでいる。

本研究では、このリンク情報が未知のエンティティの被リンク数を、グラフにおけるランキングアルゴリズムを用いて推定する手法を提案する。また、提案手法を用いて、論文の引用ネットワーク上において、論文の被引用数を推定する実験を行う。

本稿の構成は以下の通りである。第 2. 章で、本研究に関連する研究について述べる。第 3. 章で提案手法について説明する。第 4. 章で論文ネットワークについて実験を行い、SVM(Support Vector Machine [5]) による分類結果と比較する。最後に第 5. 章でまとめと今後の課題を述べる。

連絡先: 岡嶋 穣
東京大学 〒 113-0033 東京都文京区本郷 7-3-1
TEL：03-5841-6774
E-mail：okajima@mi.ci.i.u-tokyo.ac.jp

2. 関連研究

本研究の関連研究は、第一に PageRank に代表されるグラフにおけるランキングアルゴリズム自体であり、またそれらを予測するためのリンクマイニング上の研究である。

2.1 グラフにおけるランキングアルゴリズム

グラフにおけるランキングアルゴリズムとは、以下のような問題を扱うアルゴリズムを指す：ノードの集合とノードを繋ぐリンクの集合が与えられた時に、リンクがどのノードを結びつけているかという情報を手がかりに、どのノードが重要であるかをランク付けする。

ランキングアルゴリズムの最も基本的な原理は、『投票』である。例えば文書 A 中に他の文書 B へのリンクが張られている時は、文書 A の作成者は文書 B が参照するに足る有用な文書であると判断したと考えられる。このように、ノード A からノード B へとリンクが張られている時、A が B を重要なノードと見なして『投票』していると考える。この『投票』の量と質により各ノードの価値が判断される。

PageRank[1] は、グラフ内の各ノード V_i について、以下の式 (1) で定義される。

$$PR(V_i) = (1-d) + d * \sum_{V_j \varepsilon IN(V_i)} \frac{PR(V_i)}{|Out(V_j)|} \qquad (1)$$

$IN(V_i)$ は V_i へとリンクしているノードの集合、$OUT(V_i)$ は V_i からリンクされているノードの集合である。d は damping factor で $0 < d < 1$ である。

HITS(Hypertext Induced Topic Selection)[3] は、ひとつのノードに対して、次式で定義される $HITS_A$ と $HITS_H$ というふたつの数値を割り振る。

$$HITS_A(V_i) = \sum_{V_j \varepsilon IN(V_i)} HTS_H(V_j) \qquad (2)$$

$$HITS_H(V_i) = \sum_{V_j \varepsilon OUT(V_i)} HTS_A(V_j) \quad (3)$$

$HITS_A$ は多くのノードからリンクされているほど大きく、$HITS_H$ は多くのノードにリンクしているほど大きい。$HITS_A$ が最終的なランクである。$HITS_H$ は『投票』者としての価値を表わす。ここでは、重要度 ($HITS_A$) の大きいノードに多くリンクしているノードは『投票』者として価値が高いと仮定されている。

2.2 リンクに基づく重要度の予測

Yang ら [6] は、ウェブ構造について十分な情報が得られないとき、未知部分のリンクの性質を推測し、ランダムにリンクを生成することにより、より正確なランキングを行おうとする研究である。その未知のリンクの性質のひとつとして、ノードの in-degree、すなわち被リンク数の推測を行っている。ノード v_i の真の被リンク数を $d^-(v_i)$、それまでに判明しているリンク数を $fd^-(v_i)$ として、真の被リンク数を以下の式で近似する。

$$d^-(v_i) \approx \frac{n}{(m+m_1)} \cdot fd^-(v_i)(i = 1, 2, ..., n). \quad (4)$$

n はグラフ上のノードの総数、$(m+m_1)$ はクローラーが訪れたノードの総数である。すなわち、既知のリンク数を、全てのノードに対する調査済みのノードの比率で割ることにより、未調査のノードからのリンクを含めた真の被リンク数を推定する。このように推定した上で、他のノードから張られるリンクの数がこの推定値になるように、ランダムにグラフにリンクを追加することになる。クローリング初期の不十分な情報においては、通常の PageRank よりもこの手法のほうが最終的な PageRank をより正確に予見できることが実験で示されている。

この手法は真の被リンク数を既知の被リンク数を等倍して求めるという単純な近似であり、素性情報を用いてより正確な近似を求めようとする本研究とは研究の力点が異なる。本研究との重要な関連は、被リンク数を推定することで、情報が不十分なネットワークをより正確にモデル化し、より的確なランキングを行うことができるという事実である。

Kritikopoukos ら [4] はブログのランキングを対象とする。ブログは通常のウェブページに比べ、投稿 (post) 同士のリンク関係が非常に疎であるために、通常のグラフにおけるランキングアルゴリズムを用いても、重要度の評価をうまく行うことができないと彼らは主張する。そこで、ブログの重要度を正確に反映するために、ネットワーク構造および PageRank のアルゴリズムに変更を加える手法を提案している。ネットワークをより濃密な状態にするために、ブログ同士が持つ類似性に着目する。各投稿は投稿間のリンクだけでなく、トピックへの情報、著者であるブロガーの情報を持っている。そこで、同じトピックに対し言及しているか、同じブロガーが投稿しているかで、ブログ同士の類似度を計算し、その類似度が高いブログ同士をリンクで結ぶようにする。こうしてブログ間のエッジが増加し、より濃密でランキングアルゴリズムを適用しやすいグラフ構造になる。ユーザーのクリック数を元に、この手法は通常の PageRank よりもより適切にブログをランキングしていると評価している。

この研究から、互いにリンクを持たない文書を適切にリンクさせるには、文書が共有するトピックが重要な手がかりになることが理解される。

3. 提案手法

本研究では、次のようなアルゴリズムを用いて、リンク情報を素性情報で代替し、リンク情報未知のエンティティの被リンク数を推定する。

1. リンクと代替させる素性の候補として、リンクと共起する素性を抽出する。
2. エンティティと素性をノードとするネットワークを作成し、ランキングアルゴリズムで素性を評価する。
3. 素性のスコアを元に、未知のエンティティの被リンク数を推定する。

ここで、ある素性があるエンティティAに含まれるとき、Aにリンクしているエンティティにもその素性が含まれる場合、その素性を「リンクと共起する素性」と呼び、そのような素性は、そのリンク情報を代替できる候補であると考える。

以降では、この提案手法の各部分を詳細に説明する。

3.1 リンクと共起する素性の抽出

提案手法においては、リンクと共起する素性は未知のリンク情報を知る手がかりとなると仮定し、まず最初にリンクと共起する素性を抽出する。このことは、リンクしているエンティティとリンクされているエンティティに同時に含まれている素性は、そのリンクが生じた原因である可能性が高い、という仮定に基づいている。これは、特に語を素性とする文書のエンティティに関しては、根拠のある仮定である。ある文書からある文書にリンクが張られるとき、リンクが張られる原因になったトピックは、しばしば両者に共通する語として現れる。これは [4] で、ブログ同士がトピックを共有している場合に、ブログ間のリンクを推定することと同じアプローチである。

それでは、リンクと共起する素性はどのように抽出すればよいだろうか。言い換えれば、素性のリンクとの共起は、どのように数値的に評価できるだろうか。

基本となるアイディアを図 1 に示す。ある素性 f を含むエンティティの集合 (”linked entities”) を考える。さらに、このエンティティの集合にリンクしているエンティティの集合 (”linking entities”) を考える。この時、それぞれの”linking entity”が素性 f を含んでいるかを考える。もし含んでいれば、素性 f はある”linked entity”とある”linking entity”に共有されていることになり、素性 f は一つのリンクと共起していると言える。よって、素性 f が”linking entities”に含まれている比率が高いほど、素性はリンクと良く共起し、リンクが行われる上で重要視されている素性であると考えられる。

この考えを数値で評価するために、本研究では次の式を用いることとする。

$$with\_link(f) = \frac{1}{N_{e \in have(f)}} \sum_{e \in have(f)} \frac{N_{e \in linking(e) have(f)}}{N_{e \in linking(e)} + \alpha} \quad (5)$$

ここで $have(f)$ は素性 f を含むエンティティの集合、$linking(e)$ はエンティティ e にリンクしているエンティティの集合である。また $N_{e \in have(f)}$ は集合 $have(f)$ に含まれるエンティティの総数を表わすものとする。この式は、素性 f にリンクしているエンティティ e にリンクしているエンティティ e’ に、素性 f が含まれる比率を示している。なお α は、e の数が少ないときに値が極端に大きくなることを防ぐための定数（実験では $\alpha = 2$）である。

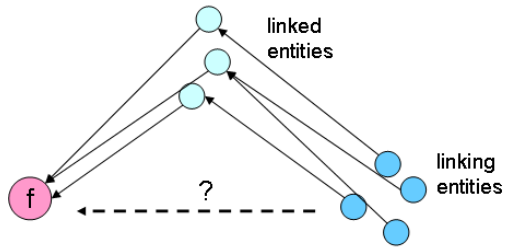

図 1: 語のリンクとの共起の計測方法

提案手法においては、まず素性すべてについて上式を用いてリンクとの共起度を算出し、上位から一定の数をリンクと共起する素性として抽出する。

3.2 グラフ構造を用いた被リンク数の推定手法

リンクと共起する素性を抽出したあとは、それらの素性を用いて、リンク情報を近似する。

ここで、どの素性に、どのようにリンク情報を代替できるようなスコアを与えるかが問題である。まず第一に、リンク関係と多く共起するような素性は、リンク情報を推定する上で重要であるはずである。また、ドメインの中であまりリンクを持たないエンティティは重要ではなく、そのようなエンティティにのみ含まれる素性を重要視するべきではない。さらに、重要なエンティティ同士を結ぶリンクと共起する素性は特に高く評価されるべきである。

このように、どの素性を高く評価するべきかは、実際は、その素性と関連するエンティティやリンクが持つグラフ内での重要度に深く依存する。すなわち、グラフ内での重要度を測るための素性を抽出するためには、逆にエンティティやリンクのグラフ内での重要度を計算し、それに付随する素性を高く評価するという手法が有効であると考えられる。それはすなわち、グラフにおけるランキングアルゴリズムである。

この観点から、本研究では、エンティティとリンクと素性からなるグラフを作成し、その上でランキングアルゴリズムを走らせ、重要なエンティティおよびリンクと共起する素性を高く評価する手法を提案する。

その着想を具体的なグラフ構造で表現したものが図 2 である。このグラフ構造上にランキングアルゴリズムを適用することで、エンティティと素性が同時に評価される。リンク情報は、このネットワークにおいて”linked entities”と”linking entities”のあいだのリンクで表現される。このリンク関係により、”linked entities”は被リンク数に応じたスコアを持つ。この”linked entities”のスコアは、それらのエンティティが含む素性へと推移する。ここで、重要なエンティティに含まれるスコアは重要であるという仮定が満たされる。さらに、これらの素性スコアの一次結合で、”linked entities”のスコアが計算される。すなわち図 2 のグラフ構造上でランキングアルゴリズムを計算することにより、被リンク数に応じたエンティティの重要度と、重要なエンティティに含まれる素性を重要視したエンティティの重要度の一次近似が同時に評価できることになる。

こうして求められた素性のスコアを用いて、未知エンティティのスコアを、含まれる素性のスコアの一次結合として推定することが可能である。これが本提案手法で推定する被リンク数である。

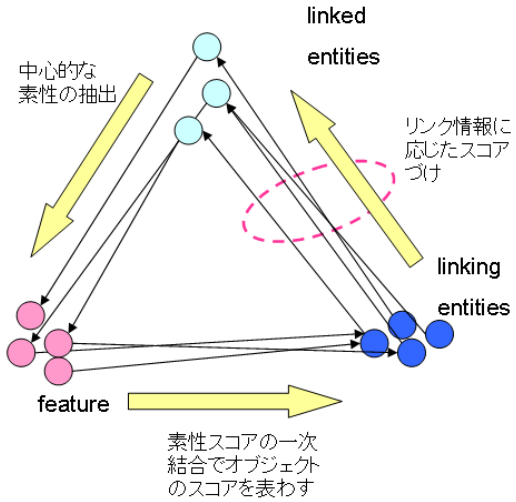

図 2: 提案手法のグラフ構造

4. 実験

本章では、提案手法の実際の適用例として、論文の引用ネットワークを選び、論文の被引用数を予測する実験を行う。

4.1 論文の被引用数を推定する理由

我々が論文を対象とするのは、以下のような理由からである。

まず、論文は引用というリンク情報を持ちながら、個々の論文自身がその内容を的確に表わす多くの素性（＝語、著者、論文誌など）を豊富に含んでいる。つまり、リンクの理由を推測しうる素性が十分量あり、リンク情報を代替させるための素性を抽出する提案手法を試すのに適したネットワークである。

第二の理由は、被引用数が論文のインパクトを測る上で重要であり、推定する意義が大きいことである。論文のインパクトを測るため過去様々な手法が提案されてきたが、その中でも論文の被引用数を用いるのが最もスタンダードな手法である。また、論文自身のインパクトだけでなく、論文誌の重要度を測る際にも論文の被引用数が用いられる ([2])。

4.2 実験データ

本論文では、実験対象として、機械学習をテーマとする論文 785 本を選んだ。これは *Web of Science* で”machine learning”で検索してヒットした、1996 年から 2003 年の論文である。さらに、これらの論文を引用している論文 8224 本を”linking entities”として利用する。論文の素性としては Abstract に含まれる語の tf*idf 値のみを用いた。論文の著者や掲載誌も、被引用数を決定する重要な要素と考えられるが、ここでは第一段階として語だけを素性として用いる。語にステミングをほどこし、一般的すぎる語はストップワードのリストを作り排除した。結果、約 8785 語が素性として抽出された。

4.3 実験１：リンクとの共起語の抽出

提案手法を用い、論文セットにおける各語のリンクとの共起度を推定した。その結果は表 1 のようになった。

機械分類の手法の名前や医学の専門用語が高く評価され、狙い通り、論文同士のリンクの理由になるような特徴的な素性が抽出できていることが分かる。

表 1: リンクとの共起語の抽出結果

順位	語	スコア
1	MCH(平均赤血球血色素量)	0.9167
2	brain-computing	0.8095
3	workflow	0.8077
4	electrophoresis(電気泳動法)	0.7796
5	sarcoma(肉腫)	0.7778
6	mediterranean(地中海の・病名)	0.7692
7	BCI	0.7143
8	transvers	0.7143
9	anti	0.7000
10	gel	0.6707
	:	

表 2: 素性の評価結果

順位	語	スコア
1	data	0.0297
2	learn	0.0239
3	machine	0.0237
4	classify	0.0206
5	tree	0.0143
6	algorithm	0.0136
7	model	0.0139
8	system	0.0133
9	predict	0.0124
10	method	0.0111
	:	

4.4 実験２：被引用数の大きい論文の分類

提案手法による被引用数推定の検証のために、論文集合を被引用数の多い重要な論文とそれ以外の論文に分け、その分類を行うというタスクを試みた。閾値を５として、５回以上引用された論文を正例とし、それ以外の論文を負例とした。

まず最初に、リンクと共起する素性が分類において有効であることを示すために、ＳＶＭを用いて以下の４条件で分類を行い結果を比較した。

1. 語のみを素性として用いる
2. 1 に加え、引用情報を素性として用いる
3. 2 に加え、リンクとの共起度を用いて、語の素性選択 (feature selection) を行う
4. 2 に加え、χ^2 検定値を用いて、語の素性選択を行う

ここでいう引用情報とは、その論文が引用している論文の被引用数の平均など、reference から求まる様々な素性である。

結果は表 3 のようになった。リンクとの共起度を用いて素性選択したときが最もＦ値が高くなっており、被リンク数を推定する上での素性の重要さを測る指標として、リンクとの共起度が優れた指標であることが分かる。

次に、提案手法を用いて同様の分類を行った。その結果は表 3 の下半分に示されている。Ｆ値において大きくＳＶＭを上回っている。

ＳＶＭの分類精度があまり高くならないのは、被引用数の高い論文と被引用数の低い論文が、明確な境界線を持たないことが理由のひとつとして考えられる。被引用数の高い論文と低い論文はベクトル空間上で明確に分離するものではなく、おおまかに入り混じりあい、ゆるやかな傾向の差だけを持つ状態であると考えられる。このような分類に対して、ＳＶＭのような分割する超平面を設定する分類器は向いていないと考えられる。

表 3: 提案手法とＳＶＭの分類精度

	SVM			
	語のみ	引用情報あり	リンク共起 ＦＳ	χ^2 ＦＳ
precision	0.500	0.750	0.607	0.529
recall	0.079	0.143	0.270	0.143
F-value	0.136	0.240	**0.373**	0.225

	提案手法
precision	0.524
recall	0.683
F-value	**0.593**

5. まとめと今後の課題

本研究では、リンクが未知のエンティティに対しリンクによって定まる重要度を推測するために、エンティティと素性をノードにしたネットワークを作成しランキングする手法を提案した。また提案手法を用いて科学論文の被引用数を推定する実験を行い、提案手法の有効性を検証した。

今後の課題としては、素性の抽出、重要度推定結果の正確な検証と評価がある。本当に理想的な素性が抽出されているのか検証を行う必要がある。また、論文ネットワークだけでなく、ウェブ文書や商品ネットワークについても実験を行う予定である。

参考文献

[1] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117, 1998.

[2] E. Garfield. The history and meaning of the journal impact factor. *JAMA*, 295(1):90–93, January 2006.

[3] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[4] Apostolos Kritikopoulos, Martha Sideri, and Iraklis Varlamis. Blogrank: ranking weblogs based on connectivity and similarity features. In *AAA-IDEA '06: Proceedings of the 2nd international workshop on Advanced architectures and algorithms for internet delivery and applications*, page 8, New York, NY, USA, 2006. ACM Press.

[5] Vladimir N. Vapnik. *The Nature of Statistical Learning Theory (Information Science and Statistics)*. Springer, November 1999.

[6] Haixuan Yang, Irwin King, and Michel R. Lyu. Predictive random graph ranking on the web. In *Neural Networks, 2006. IJCNN '06. International Joint Conference on*, pages 1825–1832, 2006.