1G2-2

論文の引用情報を用いた論文被引用数予測
Predicting Citation Count of Papers using Reference of the Paper

関 喜史
YoshiLawrencemi Seki

松尾 豊
Yutaka Matsuo

東京大学大学院工学系研究科
School of Engineering, the University of Tokyo

Citation count of paper is important indicator of paper for readers. Many studies have observed a correlation between bibliometric measures and citation counts. However, there are few studies focusing on predicting future citation count of paper. In this paper, we propose to use reference information of paper as new feature to predict citation counts of future. We use support vector machine to train our model and predict future count. that preference information is good resource for future prediction of citation counts.

1. 背景と目的

論文の被引用数は研究の評価指標として知られている[Gross 27]が,その重要性は電子ジャーナルの普及に伴い近年変化しつつある.

電子ジャーナルは大きく普及しており,学術図書館研究委員会の調査によれば,研究者の内 70%が最近読んだ論文は電子的に入手したと回答している[佐藤 10].それに伴い研究者の研究スタイルは大きく変化している.電子ジャーナル登場以前に論文を探す際は,コアジャーナルの目次から論文を探索する方法が一般的であった.現在では,電子ジャーナルでキーワードを元に論文を検索すると,キーワードを含む論文のリストが表示される.多くの場合そのランキングは論文の被引用数に基づいており,研究者は被引用数を元に論文を探すことができる.このように論文被引用数は,研究の評価指標としてだけでなく論文の検索指標として用いられるようになっている.電子ジャーナルの登場以後,全体では論文の引用文献数が増えているが,その引用は一部の論文に集中しているという報告[横井 10]がある.これは被引用数の高い論文ほど,研究者の目に止まりやすくなっていることが理由の一つと考えられる.

しかし被引用数は論文が発表されてから数年経たないとわからない指標であり,過去の論文と最新の論文を同列に評価できないという点で検索指標として用いるには問題がある.そこで本研究では,書誌情報を用いて最新の論文の将来の被引用数を予測することを目的とする.

2. 既存研究とその課題

被引用数と書誌情報の相関を示すことで,被引用数に影響を与える書誌情報を探す研究は過去数多く行われている.しかし,実際に予測モデルを生成するという問題にまでは踏み込んではいない.予測を実現した例として Lokker らがイギリスの医学誌 BMJ のデータベースにおいて,発表から 2 年後の被引用数を発表後 3 週間後に得られる指標から回帰分析を行い,その結果生成した回帰モデルで精度の検証を行っている[Lokker 08].その結果データセットの中で被引用数が上位1/2 に入る論文を感度(sensivity)83.3%,特異度(specicity)71.5%,上位 1/3 に入る論文を感度 66.1%, 特異度 82.2% で予測できることを示している.また Lawrence らは MEDLINE のデータベースに対して,Support Vector Machine(SVM)を用いて 10 年後の被引用数が閾値以上になるかどうかで正例負例を分け,予測を行い 80%以上の精度が得られることを示した[Lawrence 10].このように高精度の予測ができている研究もあるが検討すべき点は未だ多くある.

まず予測に用いる特徴量が汎用的でないことが挙げられる.Lokker らの研究では電子ジャーナル上のアクセス数や,論文に対するアラートの設定数,BMJ 側の評価値などが特徴量として採用されている.また Lawrence らの研究では MEDLINE データベース上における MeSH term というタグ情報を用いている.

また用いるデータセットと実際のデータとの間にギャップがある.世の中に存在する論文のほとんどは一度も引用されず,引用を一定数獲得する論文はごく一部である.しかし Lawrence らの研究では,データセットの半分以上が 10 年後に 50 以上の引用を獲得した論文で占められており,実際のデータセット上での予測の再現には疑問がある.

そしていずれの研究においても,実際に予測モデルとして用いることが出来るかの検証がなされていない.Lokker らの検証方法は,予測モデルを 2005 年の一部の論文の 2 年後の被引用数を用いて生成し,2005 年の残りの論文に対して 2 年後の被引用数を予測することで検証するものである.また Lawrence らの検証方法は 1991 年,1992 年の 10 年後の被引用数を学習し,1993 年,1994 年の論文の被引用数の予測精度を検証するものである.いずれの検証方法でも,モデルを生成するために用いた学習データに未来のデータが使われている.Lokker らの検証方法において,モデルの生成に用いたデータは 2005 年に発行された論文の 2007 年時点での被引用数である.検証に用いたデータは 2005 年の物であるため未来のデータで生成されたモデルで予測を行なっていることになる.Lawrence らの検証方法も同様に,モデル生成に用いたデータは 1991 年,92 年の 10 年後の被引用数なので 2001 年,2 年のデータである.検証に用いたデータは 1993,94 年の物であるため未来のデータから生成されたモデルを用いて予測を行なっていることになる.以上のことから過去の研究における検証では,予測が行えると結論づけることができない.

以上を踏まえ,本研究では汎用的な特徴量を使うこと,実際のデータセットにおける引用数の分布を踏まえること,そして最新の論文への適用を考慮した詳細な検討を行うことの 3 点をコンセプトとして論文被引用数予測を行う.

連絡先:関 喜史, 東京大学大学院工学系研究科, 東京都文京区弥生 2-11-16 工学部 9 号館 03-5841-1161,seki@biz-model.t.u-tokyo.ac.jp

3. モデル生成

3.1 学習データ

本研究では 1996~2007 年の間の IEEE 関係論文誌 25 誌の論文,計 35,786 件とそれらの論文に対する被引用数の情報を学習,評価のデータとして取り扱う.

クラス分類の方法としては Lawrence らの手法[Lawrence 10]と同様に一定期間語の被引用数が閾値以上となるか,それ未満かで分類する.閾値は被引用数 20,50 とする.閾値以上のデータを正例,そうでないものを負例として取り扱う.期間は 2 年,5 年と設定する.そのため 2 年後の予測に用いる論文は 2005 年まで,5 年後の予測には 2002 年までの論文を用いる.

素性には過去有効とされている Minger らが用いた書誌情報の内[Minger 10]今回の予測には参考文献の数,参考文献のうち最近 5 年間に発表された論文の数,著者人数,ページ数,Impact Factor[Garfield 72]について検証する.

そして本研究では新たに参考文献の被引用数を用い素性を提案する.過去の研究で参考文献の数と被引用数については相関が認められている[Lokker 08][Minger 10][芳鐘 09].また引用ネットワークを解析することで特徴的な論文を抽出する研究[梶川 09]や,参考文献情報を用いた論文推薦システムの提案もなされている[鷲崎 01].しかし,参考文献の質とその論文の被引用数の関係については過去検討されていない.本研究では,参考文献の被引用数を予測モデルの素性として用いることで,参考文献の質が論文の被引用数と関連していることを示す.今回素性として提案する特徴量は最大被引用数,平均被引用数,被引用数 20 以上の参考文献数,被引用数 50 以上の参考文献数,被引用数 100 以上の参考文献数である.

3.2 予測モデル生成

SVM のライブラリとして LIBSVM を用いる[Chang 01].カーネル関数は RBF カーネルを用いることとし,パラメータは 10-fold クロスバリデーションテストで F-value を用いて決定する.

SVM を用いて学習・評価を行う場合には正例と負例のデータ数が 1:1 であることが望ましいが,実際の論文データにおいて被引用数が 20 以上になる論文は少ない.例えば 5 年後に 20 を超える論文は全体の 10%程度である.そのため,正例に対して負例が同じ数になるようにランダムに選択して学習データ,検証データを生成している.

4. モデルの評価

4.1 素性の選択による精度の比較

まず 3.1 節で提示した 10 個の特徴量の予測結果に与える影響を検討し,予測モデルの生成に用いる特徴量を決定する.まず,
・すべての特徴量を素性として用いたモデル
・書誌情報のみを素性として用いたモデル
・参考文献の被引用数を素性として用いたモデル
の 3 つのモデルを生成しそれらの予測結果を検討する.図1にそれぞれの素性を用いて,1996 年の論文から学習した 5 年後の被引用数が 20 以上かを予測するモデルの Accuracy を示す.

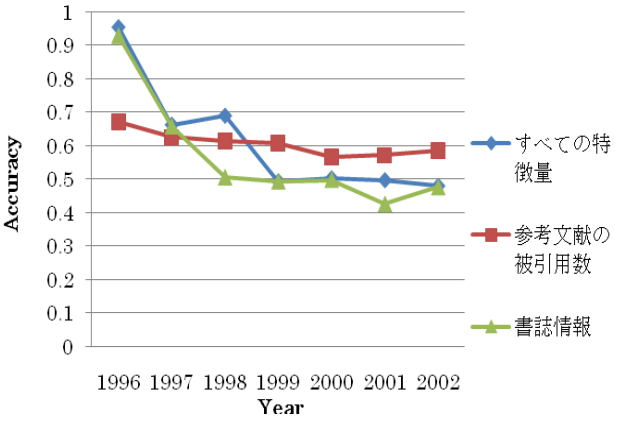

図.1 用いた素性の違いによるモデルの予測精度の変化

学習データとして用いた 1996 年の論文では,すべての特徴量を素性として用いたモデルと書誌情報を用いたモデルは 9 割以上の再現率を示しているのに対して,参考文献の被引用数を用いたモデルでは 6 割前後の再現率となっている.しかし 1997 年の論文ではほぼ同等となり,1998 年の論文では書誌情報を用いたモデルの精度ランダムに選択した場合と同等の 5 割になっている.そして 1999 年にはすべての特徴量を素性として用いたモデルも 5 割となっている.それに対して参考文献の被引用数を用いたモデルの精度は 6 割前後で安定している.以上の結果から書誌情報で素性として用いた特徴量の中に再現率を大きく高めるものの,予測に対して用いることのできないものが含まれていることがわかる.

さらに詳しく検討するため,参考文献の被引用数を素性として用いたモデルに素性として参考文献の数,参考文献のうち最近 5 年間に発表された論文の数,著者人数,ページ数,Impact Factor を1つずつ加えたモデルを生成し,それぞれの特徴量が予測結果に与える影響を検討する.図 2 に 1996 年の論文で学習した 5 年後の被引用数を,閾値 20 で予測するモデルの予測結果を示す.

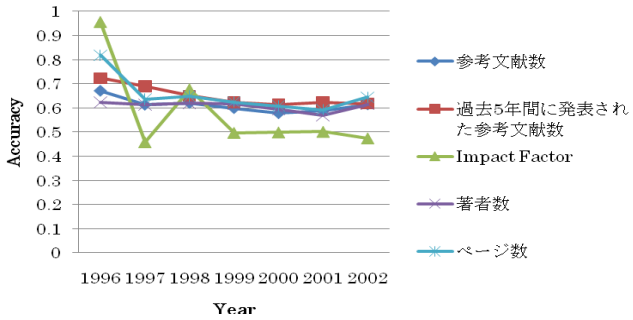

図.2 書誌情報がモデル予測結果に与える影響

Impact Factor を素性に加えたモデルでは再現率は 9 割を超えるものの,1997 年以降の予測結果は 5 割とランダムと同等の結果となっている.それ以外の指標では 6 割前後の予測結果を示していることから図1でみられた結果の要因は Impact Factor であったことが分かる.

4.2 モデルの最適化

以上を踏まえて,3.1 節で示した特徴量から Impact Factor を除いた 9 つの特徴量を素性として生成した予測モデルの評価を行う.図 3 に 1996 年の論文から学習した,5 年後の被引用数が 20 以上かを予測するモデルの予測結果を示す.

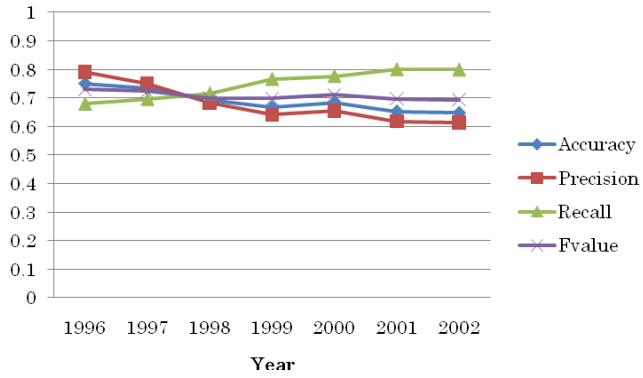

図.3 構築した予測モデルの予測精度

年数が経過しても Precision は 6 割程度,F-Value も 0.7 程度を示しており一定の予測が行えることを示している.

1996 年の 5 年後の被引用数は 2001 年のデータである.そのため 2002 年の論文に対して 1996 年の論文で生成した 5 年後の被引用数を予測するモデルでの予測結果を検証することで,学習に未来のデータを用いていない検証を行うことができる.2002 年の論文に対する予測結果は Accuracy 0.65,Precision 0.61,Recall 0.8 F-value 0.69 であり,十分予測を行うことが出来ているといえる.このことから,本研究で生成したモデルを用いて,未来のデータを用いることなく論文被引用数予測が出来ていることが示された.表 1 に 2 年後の被引用数を閾値 20 で予測するモデルにおいて,未来のデータを用いずに予測を行った場合の精度を示す.いずれのモデルでも十分な予測が行えており,本研究で提案した手法で予測が行える可能性をしめしている.

表 1 2 年後の閾値 20 での予測結果

学習データ	1996	1997	1998	1999	2000	2001	2002
検証データ	1999	2000	2001	2002	2003	2041	2005
Accuracy	0.7	0.79	0.88	0.82	0.72	0.73	0.65
Precision	0.65	0.73	0.82	0.8	0.7	0.68	0.6
Recall	0.89	0.94	0.97	0.85	0.78	0.84	0.89
F-value	0.75	0.82	0.89	0.82	0.74	0.75	0.72

5. 実際のデータセットへの適応

5.1 実際のデータセットにおける問題点

ここまで正例:負例が 1:1 の場合には予測が可能であることを示した.しかし,2 節で述べた通り,実際の論文データでは正例と負例の割合は負例が遥かに多い.図 4 に生成したモデルを実際の論文データに適用した際の結果を示す.モデルの生成に用いたのは 1996 年の論文で 5 年後の被引用数を閾値 20 で分類する.なお 1996 年の論文 4447 件のうち,5 年後の被引用数が 20 以上となる論文は 380 件と 1 割以下である.

このように精度は非常に落ちる.もともと正例の割合が 1 割以下であることからランダムに選択した場合の Precision は 0.1 以下であり,それと比較して精度は多少高いが,大きな改善はみられない.Accuracy と低さと比較して Recall が高いことから,正例であるデータは正例と判断されているが,本来負例である多くのデータが正例と分類されていることがわかる.

本研究ではモデル生成の際にデータの正例と負例の割合 1:1 とした.このとき多くの負例のデータを学習データから除外しており,除外した負例のデータから得られる特徴は生成したモデルにおいて学習できていない.これが多くの負例が正例と判断されてしまった要因であると推測する.

この仮説を検証するため,実際に負例の選び方を変えて予測モデルを生成し,その精度を検証した.1996 年の 5 年後の被引用数を予測する予測モデルを,負例となる論文がすべて含まれるように負例を分割することで複数個生成し,それらすべての予測モデルで 1996 年の全論文データを予測した際の Accuracy の分布を図 5 に示す.生成されたモデルの数は 11 個であった.

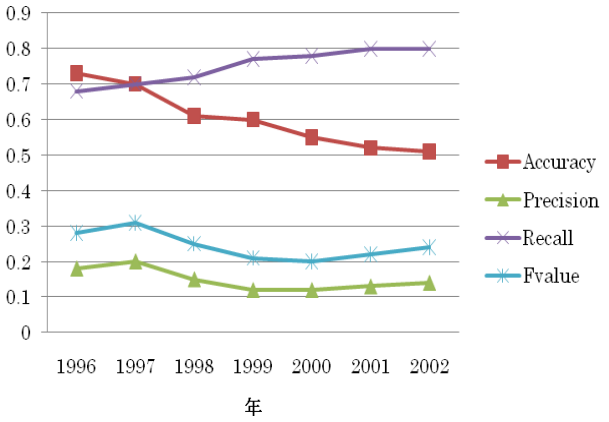

図.4 論文全体に対する予測結果

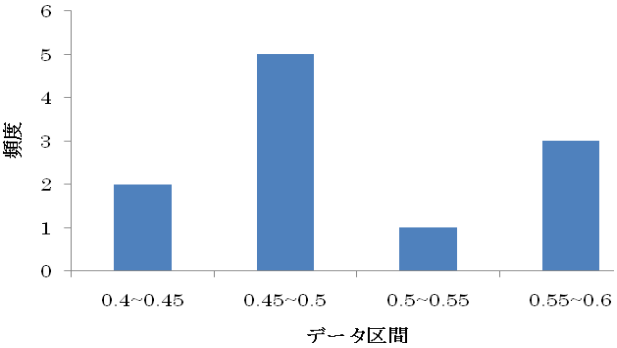

図.5 モデルの Accuracy の分布

5.2 被引用数予測手法の提案

図 5 より負例の選び方によって生成されたモデルによる予測精度が変化してしまうことが示された.これより正しい予測を行うためには負例の特徴全体を反映する必要が.しかし SVM を用いて適切なモデル生成を行うためには正例と負例が1:1である

必要がある.そこで負例を分割することで生成された予測モデルをすべて利用することを考える.

生成されたモデルをすべて論文データに対して適用し,その結果いくつのモデルで正例と判断されたかによって,その論文を正例と負例に分類することを考える.この方法により正例と負例が1:1でモデルを生成しながら,すべての負例の特徴を反映させることが可能となる.

図 6 は 1996 年の 5 年後の被引用数予測システムの精度の変化を示したものであり,モデル数の閾値によって精度がある程度変化することを示している.予測モデルの数は 11 個である.なおもともと正例の数は全体の 10%ほどのため,Precision がそれ以上であれば一定の予測が出来ているといえる.F-value をみるとモデル数の閾値を 10 に設定したときに最大の値となり,F-value 0.29.Precision 0.17, Recall 0.81 となる。適切な閾値を設定すれば正しい予測が行えるが,その値をどう設定するかという点を考慮していく必要がある.また,さらに正例の数が負例に比べて極端に少なくなると予測モデルの数が多くなり,Recall がとても小さくなってしまう.そのため,どの予測モデルでシステムを構成するかということも重要である.

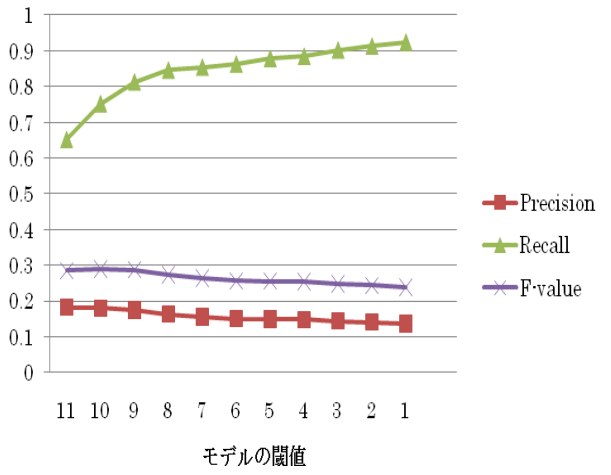

図.6 モデル数の閾値とその分類精度

6. まとめ

本研究では最新の論文の評価指標とするために,将来的な被引用数を予測することを目的として,SVM を用いて被引用数予測モデルの生成を行い,被引用数予測の可能性を検討した.

汎用的な特徴量を用いた予測モデルで最新の論文に対する予測が行えることを示した.加えて参考文献の被引用数を予測モデルの素性として用いることで,論文の被引用数に参考文献の被引用数が一定の影響を与えていることを示した.

そしてどの年の論文データをモデルとし,どの年の論文データを用いて精度の検証を行うかによって,精度に差がでることを明らかとした.特に Impact Factor のように学習段階では高い再現率を実現するものの,モデルとしたデータとの時間が離れるほどに予測精度が落ちていくような特徴量も存在する.そのため被引用数予測の検証においては,未来のデータを使用しないでよそくが行えるかということを検討する必要がある.

実際の論文データの分布においては,正例と負例の割合が大きく違うことで単純な学習が難しいという問題が明らかになった.そこで予測モデルを組み合わせることで,正例が極端にすくない状況でも予測を行うシステムを提案し,その可能性をしめした.

今後は予測モデルのさらなる精度向上や,他の論文データに対しても適用可能かという検証を行っていく.そして予測システムにおいてモデルの選択や閾値設定といった課題の解決が求められる.

参考文献

[Chang 01] Chih-Chung Chang and Chih-Jen Lin:LIBSVM: a Library for Support Vector Machines available at : http://www.csie.ntu.edu.tw/cjlin/libsvm (2001)

[Garfield 72] Garfield E.: Citation analysis as a tool in journal evaluation, Science, Vol.178,(1972).

[Gross 27] P.L.K Gross, E.: College Library and chemical education, Science, Vol.66,(1927).

[Lawrence 10] Lawrence D. F. and Aliferis,C.F..:Using content-based and bibleo-metric features for machine learning models to predict citation counts in the biomedical literature. ,Scientometrics, Vol. 85, (2010).

[Lokker 08] Lokker, C., McKibbon,K.A.,McKinlay, R. J., Wilczynski, N. L., and Haynes, R.B.: Prediction of citation counts for clinical articles at two years using data available within three weeks publication : retrospective chort study ,BMJ , Vol. 336 (2008)

[Mingers 10] John Mingers, Fang Xu:The drives of citations in management science journals.,European Journal of Operational Reasearch,205,(2010).

[梶川 09] 梶川 裕矢,森 純一郎:ネットワーク指標を用いた学際的な論文の抽出,情報知識学会誌,Vol.19,(2009)

[佐藤 10] 佐藤 翔:電子リソースの普及と研究活動への影響,カレントアウェアネス,(2010)

[横井 10] 横井 慶子:電子ジャーナルが研究者の文献利用へ及ぼす影響:国立大学所属研究者発表論文の引用文献分析,2010 年度三田図書館・情報学会研究大会研究発表,(2010).

[芳鐘 09] 芳鐘 冬樹,辻 慶太,小野寺 夏生:論文引用に影響を与える要因:負の二重回帰による検討,日本図書館情報学会春季研究集会発表要項,(2009)

[鷲崎 01] 鷲崎 弘宣,深澤 良影:パターン:引用からの品質,第7回パターン研究会,(2001).