Web 上の学術リソースの検索

難波英嗣
広島市立大学大学院 情報科学研究科
nanba@hiroshima-cu.ac.jp

1. はじめに

ある研究分野の成果として作られたツールやデータなどを学術リソースと呼ぶとき、Web 上に存在するこのようなリソースを、非専門家でも容易に検索・入手できるシステムがあれば、最新の研究成果を利用した商用サービスの実現が容易になり、産学連携の活性化も期待できる。筆者はこのようなシステムの開発を行っており、本稿では、その仕組みについて述べ、システムの動作例を紹介する。

近年、学術リソースを作成者自ら Web 上で公開することも少なくないが、現状では、非専門家がこれらを有効に活用できる環境が整っているとは必ずしも言えない。これには以下に述べる 2 つのケースが考えられる。第一のケースは、非専門家が必要とするツールがひとつのパッケージとしてまとまっていない場合である。例えば、画像を入力とし、その画像に人物が写っているかどうか、あるいは海や山が写っているかどうかを自動的に判定できるようなツールを探している場合を考える。この時、Web 検索エンジンを用いて「画像分類＋ダウンロード」といったクエリで検索しても、該当するツールは容易には見つからない。第二のケースは、関連ツールが非常に数多く存在する場合である。例えば、全文検索システムを探している場合に、Web 検索エンジンを用いて「全文検索＋ダウンロード」といった検索すると、膨大な数の検索システムが見つかる。しかし、これらの中には、技術的に見てかなり古く、最新の研究成果と比べると検索精度の点で十分でないものも含まれており、結果的に検索者が求めるツールにたどりつくことが難しくなる。

そこで、次の 2 点を考慮した学術リソースの検索について考える。一点目は、検索者が探しているツールそのものがなくても、複数のリソースを組み合わせて実現できる場合、それらを列挙する。例えば、上述の第一のケースにおいて、もしこの検索者が機械学習に関する知識を持っていれば、例えば、mirflickr[1]のような画像データと OpenCV[2]などのコンピュータビジョン向けライブラリを組み合わせることで、画像分類器を作成することができる。二点目は、最新の研究論文で提案されている手法、あるいは最新の手法と比べ、それほど見劣りしない、言い換えれば最新手法と比較できる程度に新しい技術を用いたリソースを検索する。

上記 2 点を考慮した学術リソースの検索を実現するため、本研究では、研究論文中で提案手法の有効性を調べるために使われるベースライン手法に着目する。一般に、ベースライン手法には、最新の技術と比較できる程度の性能をもつ従来技術が用いられる。このような技術は、同じ研究テーマの他の研究論文でも同様にベースライン手法として採用されることが多く、ツールあるいはライブラリやデータとして一般公開されていることも少なくない。この場合、論文中では、それらのリソースの入手方法として URL が記載されることが多い。そこである分野の論文集合で高い頻度で記載されている URL を抽出できれば、それがその分野の学術リソースになっている可能性が高い。本稿では、このような考え方に基づいた学術リソース検索システムについて述べる。さらに、この考えを特許にも適用し、特許中の URL の抽出を行ったシステムも紹介する。

本稿の構成は以下のとおりである。2 章では、関連研究について述べる。3 章では、学術リソース検索システムの概要および構築手順について述べる。4 章ではシステム動作例を紹介し、5 章で本稿をまとめ、今後の課題について述べる。

2. 関連研究

Web ページと論文間の引用(リンク)関係を扱った研究は、大きく、以下の 2 つのグループに分けることができる。

1. オンライン・ジャーナルなどの Web 上でアクセス可能な論文がどのような Web ページから引用されているのかを分析する研究 [Kousha 2007][Vaughan 2005]
2. 論文中でどのような Web ページを引用しているのかを分析する研究 [Lawrence 1999][Yang 2012]

以下、それぞれの関連研究について述べる。

Kousha ら[Kousha 2007]は、Web 上に存在する論文に引用(リンク)している Web ページについて、様々な側面から分析を行っている。この分析

[1] http://press.liacs.nl/mirflickr/
[2] http://opencv.org/

の観点として、例えば、引用元の Web ページを国別に分類したり、サイトのドメイン(org/com/edu など)別に分類したりしている。また、ある論文(誌)を引用する平均 Web ページ数とインパクトファクタの間には一定の相関があることを報告している。インパクトファクタとの相関に関して、Vaughan ら[Vaughan 2005]も同様の結果を報告している。

Lawrence ら[Lawrence 1999]は、論文中で引用されている Web ページの持続性について調査している。Web ページは時間が経過するにつれ、ページそのものが消滅してしまうことがある。そこで、論文中で引用されている Web ページがどのくらい消滅するのかを、論文の著作年ごとに分けて集計している。その結果、論文が発表されて 5 年以上経過すると、論文中で引用されている Web ページの過半数は消滅すると報告している。一方で、こうして消滅してしまった Web ページの大半は、URL が変わっているだけで、検索エンジンなどを利用して同一内容の Web ページをすぐに見つけることができたり、同一でなくても関連性の高いページを見つけることができたりするとも報告している。Yang ら[Yang 2012]は、 3 つのデータセット(the Chinese Social Science Citation Index / Communication of the ACM, IEEE Computer / MEDLINE)について、論文中の URL を分析している。分析の観点として、以下のものが挙げられる。

- Web サイトのドメイン: com/net/org/edu/gov/ac/int など
- Web ページのタイプ: html/pdf/doc/ppt/動的なもの(php/jsp/asp)
- URL の頻度
- URL の深さ(URL に含まれる/の数)
- URL の長さ: URL の文字数

以上の関連研究を概観すると、Web ページと論文間の引用の性質を統計的に分析することに焦点を当てている。これに対し、本稿では、論文中での Web ページの引用を、第三者が再利用するためのシステムを構築することに主眼を置いている点が従来研究と異なる。

3. 学術リソース検索システムの構築

本章では、学術リソース検索システムの構築方法について述べる。

3.1. 学術リソース検索システム構築の手順

学術リソース検索システムの構築は、以下の 3 つのステップから構成される。

1. 文書データから正規表現を用いて URL を抽出する。
2. 抽出された URL にアクセスし、Web ページを収集する。Web ページにアクセスできない場合[3]、その URL は、このステップで除外する。また、今回は学術リソースとしてツールやデータの検索を目的とし、論文などは対象外とするため、拡張子が pdf および txt となっているものは事前に除外する。
3. ステップ 2 で収集できた Web ページを対象に、学術リソース検索システムを実装する。

ステップ 3 については、2 種類の方法で実装している。3.2 節では、これらの方法について述べる。

1.2 学術リソース検索システムの概要

学術リソース検索システムを構築する第一の方法は、3.1 節のステップ 2 で収集された Web ページを対象にした全文検索システムを構築することである。図 1 に、第一の方法によるシステムの概要を示す。

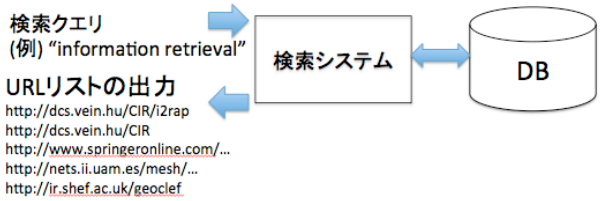

図 1　学術リソース検索システムの概要(その 1)

図 1 において、ユーザが"information retrieval"(情報検索)を検索クエリとして入力すると、全文検索システムは、クエリと類似度が高い文書を順に出力する。検索結果は、URL のリストとして出力される。第一の手法では、結果として出力する文書には必ず検索クエリが含まれているため、無関係な文書が出力されにくいという利点がある。一方で、1 章で述べたように、検索者本人が、検索された複数のリソースを組み合わせて目的のシステムを構築する必要がある時、第一の手法では、検索できない場合がある。例えば、1 章で例に上げた画像分類の場合、画像データ(mirflickr)、プログラムのライブラリ(OpenCV)の他に機械学習ツールが必要となるが、機械学習ツールを公開している Web ページ内に「画像分類」という用語が含まれているとは限らない。

そこで、この問題を解決する第二の手法を図 2 で説明する。

[3] Web ページにアクセスできないケースは、(1)Web ページそのものが消滅している場合と、(2)ステップ 1 で抽出された URL が正しくない場合の 2 通りが存在する。

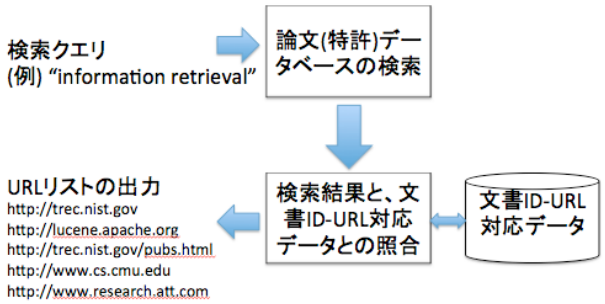

図 2　学術リソース検索システムの概要(その 2)

図 2 において、ユーザが“information retrieval”(情報検索)を検索クエリとして入力すると、システムは、まず、検索クエリと類似度の高い論文を検索する。次に検索された論文中に含まれる URL を、3.1 節手順 2 で構築された「文書 ID-URL 対応データ」と照合し、結果として得られた URL 集合を頻度順に並べ出力する。第二の手法では、検索クエリを含まない Web ページも出力するため、第一の手法よりも高い再現率が期待できる。なお、図 1 と 2 では、論文を対象にした場合について述べたが、論文を特許に置き換えることで、全く同様の手順で特許中の URL を検索するシステムを実現できる。

4. システムの動作例

3 章で述べた手法に基づき、学術リソース検索システムを構築した。4.1 節では、システム構築に用いたデータについて述べ、4.2 節で、実際の出力例を示す。

4.1. 学術リソース検索システム構築に用いたデータ

システムの構築には、(1) CiteSeer[4]全文データ、(2) 言語処理学会年次大会論文および論文誌データ(ANLP-20 コーパス)、ACL Anthology[5]、(3) 米国特許、(4) 日本国公開特許公報を用いた。データの詳細および抽出された URL 数を表 1 に示す。表 1 からわかるとおり、全体的に論文から抽出できた URL 数に比べ、特許から抽出できた URL 数が非常に少ない。ただし、これにはいくつか理由がある。まず、CiteSeer 全文データは、大半が情報科学分野の論文であることから、農業、機械、電気などあらゆる分野の文献を含む特許と比べ、Web ページを引用する確率が潜在的に高いと言える。また、日本国公開特許公報中で引用される URL 数が極端に少ない理由のひとつは、URL が全角で記述されるケースが少なくないことに加え、例えば、ハイフンを漢字の一や罫線記号で記述される場合などもあったため、全角を半角に変換するという前処理は行ったものの、抽出漏れがかなりあると思われる。

表 1　学術リソース検索システム構築に用いたデータ

文書データ	文書数	URL 数 (異なり数、頻度 2 以上)
CiteSeer 全文データ	2,232,117	378,364
ANLP-20 & ACL Anthology	27,200	956
米国特許 (2006-2011)	1,111,717	18,812
日本国公開特許 公報(2006-2011)	2,118,122	2,913

同じ学術リソースでも ANLP-20 & ACL Anthology から抽出できた URL の文書数に対する比率は CiteSeer と比べるとやや少ない。これは、ANLP-20 の年次大会のデータにおいて、おそらく日本語用の OCR ソフトが利用されたため、URL 周辺の文字列が正しく認識されていないケースが少なからずあったためであると考えられる[6]。とはいえ、自然言語処理関連の URL が一定数は抽出できている。図 3 は、ANLP-20 コーパスから実際に抽出された URL の一例であり、抽出件数もあわせて記載している。

60 http://mecab.sourceforge.net
57 http://chasen.org
38 http://chasen.naist.jp/hiki/ChaSen
23 http://chasen.aist-nara.ac.jp
23 http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html
22 http://chasen-legacy.sourceforge.jp
16 http://svmlight.joachims.org
15 http://www.goo.ne.jp
15 http://www.aozora.gr.jp
15 http://ja.wikipedia.org/wiki

図 3　ANLP-20 コーパスから抽出された URL

4.2. システムの出力例

特許から抽出された URL が少ないため、ここでは、論文を対象にした結果のみを掲載する。図 4 は、 “statistical machine translation” (統計的機械翻訳)を入力とし、第二の手法を用いて学術リソース(CiteSeer)を検索した結果である。

[4] http://citeseerx.ist.psu.edu/index
[5] http://aclweb.org/anthology/

[6] 論文誌に関しては TeX から変換された XML データを利用したため文字認識誤りに関する問題は発生していない。

14 http://www.fjoch.com/GIZA++.html
12 http://www.nist.gov/speech/tests/mt
11 http://www.clsp.jhu.edu/ws99/projects/mt/final
8 http://cs.jhu.edu
8 http://mi.eng.cam.ac.uk
7 http://www.iccs.inf.ed.ac.uk
5 http://www.csie.ntu.edu.tw

図 4 “statistical machine translation”(統計的機械翻訳)を検索クエリとし、第二の手法を用いて学術リソース(CiteSeer)を検索した結果

図 4 において、トップに表示されている URL は、統計的機械翻訳の代表的なツールである GIZA++の Web ページである。2 位に表示されている URL は、機械翻訳の評価ツールに関するページであり、概ね妥当な結果が得られていると考えることができる。

図 5 は、“翻訳”を入力とし、第二の手法を用いて学術リソース(ANLP-20 & ACL Anthology)を検索した結果である。日本語をクエリとした場合でも、トップに表示されている URL は図 4 と同じく GIZA++の Web ページであるが、2 件目や 5 件目は日本語用のリソースが検索できていることがわかる。

2 http://www.fjoch.com/GIZA++.html
2 http://unicorn.ike.tottori-u.ac.jp/toribank
1 http://www.speech.sri.com/projects/srilm
1 http://www.ldc.upenn.edu
1 http://www.kc.t.u-tokyo.ac.jp/nl-resource/juman.html

図 5 “翻訳”を検索クエリとし、第二の手法を用いて学術リソース(ANLP-20 & ACL Anthology)を検索した結果

次に、第一の手法を用いて学術リソース(CiteSeer)を検索した結果を図 6 に示す。

http://dcs.vein.hu/CIR/i2rap
http://dcs.vein.hu/CIR
http://www.springeronline.com/3-540-26166-4
http://ir.shef.ac.uk/geoclef

図 6 “information retrieval”(情報検索)を検索クエリとし、第一の手法を用いて学術リソース(CiteSeer)を検索した結果

図 6 において、トップおよび 2 位の URL は、情報検索ツールを公開しているグループの Web ページである。3 位の URL は情報検索に関する書籍の Web ページである。

第一の手法で情報検索に関する Web ページが検索できているものの、代表的な情報検索ツールが結果の上位にランクされていないなど、全体的に見ると、第二の手法の方が良好な結果が得られているように思われる。

5. おわりに

本稿では、著者が現在構築中の学術リソース検索システムについて述べた。定量的な評価を行っていないため、断定するまでには至っていないものの、システムの出力結果を見る限り、第二の手法では概ね良好な結果が得られたと言える。

謝辞

本システムを構築するにあたり、CiteSeer 全文データを提供していただいたペンシルベニア州立大学の Lee Giles 博士、ANLP-20 コーパスを提供していただいた言語処理学会に感謝の意を表する。

参考文献

[Kousha 2007] Kayvan Kousha and Mike Thelwall (2007) “How Is Science Cited on the Web? A Classification of Google Unique Web Citations” Journal of the American Society for Information Science and Technology, 58(11), pp.1631-1644.

[Lawrence 2001] Steve Lawrence, David M. Pennock, Gary William Flake, Robert Krovets, Frans M. Coetzee, Eric Glover, Finn Arup Nielsen, Andries Kruger, and C. Lee Giles (2001) “Persistence of Web References in Scientific Research” Computer, 34(2), pp.26-31.

[Vaughan 2005] Liwen Vaughan and Debora Shaw (2005) “Web Citation Data for Impact Assessment: A Comparison of Four Science Disciplines” Journal of the American Society for Information Science and Technology, 56(10), pp.1075-1087.

[Yang 2012] Siluo Yang, Ruizhen Han, Jingda Ding, and Yanfei Song (2012) “The distribution of Web citations” Information Processing & Management, 48, pp.779-790.