JSiSE Research Report
vol.31,no.1(2016-5)

論文検索データから歴史を見る

土屋　敏明[*1]，鈴木　一史[*2]
*1 放送大学大学院, *2 放送大学

How to Determine Historical Context in Web Search Data

Toshiaki TSUCHIYA[*1], Motofumi SUZUKI[*2]
*1 Graduate School of the Open University of Japan, *2 The Open University of Japan

The Internet has been growing explosively through the use of information technology based on highly improved communication speed and computer ability. If someone investigates the historical background of an event in the field of technology on the Internet, the information is considered as "recent data" in most cases. The information that exists on the Internet has been optimized for "current" users' interest for only a few years. It is considered difficult to organize historical context using primary information. Therefore it is necessary to conceive organization of context in time series using data on the Internet. In this study, methods that receive a large amount of data from the Internet and some results of the trial are discussed.

キーワード: 情報検索，ビッグデータ，歴史, データベース，インターネット

1. はじめに

今後インターネットに広がる情報を有効活用することは教育システムを考える上でも重要である．インターネットは，通信速度やコンピュータ能力の劇的向上による情報技術を利用して爆発的に成長している(1)．インターネット上に存在する情報は，現在のユーザによる数年間の興味や行動を対象に最適化されている(2)．そのため，ある技術分野の事象の歴史的文脈をインターネット上で調査する場合，ほとんどの情報源は何かの恣意的作用による順序付けられた上位の情報であり，「最近のデータ」となってしまう．これは，Wikipedia で整理された 2 次的情報を活用するような行為に象徴される(3)．このような状況では，一次情報を使用して歴史的文脈を客観的に整理することは困難であると考えられる．インターネット内の一次的情報源のデータを使用して，時系列な文脈を整理する方法を考案する必要がある．本研究では，具体的キーワードを設定してインターネットから大量のデータを取得する複数の方法を実行して，結果を比較検討する．そこからさらに，歴史的推移をする時系列データを抽出して，その特性について議論する．第一報として，インターネットからの情報取得方法の検討を行い，その特性について報告する．

2. 情報取得方法の検討

2.1 キーワードの設定

インターネットからの取得情報の特性を考えるために，ある技術分野のキーワード具体的に設定し，取得された情報を整理する．そのために，1980 年代にスーパーコンピュータの出現と発達ともに自然科学分野の研究開発において注目され(4)，さらに近年のインターネット上の「ビッグデータ」を処理し認識するための方法としての「visualization（可視化)」というキーワードに注目した(5)(6)．このキーワードを元に，インターネットでキーワードに関する情報を取得し，歴史的経緯を整理する方法について検討する．

2.2 仮説と取得対象データ

キーワードに関するデータはインターネット上に無数に存在する．この研究では情報検索において最初に表示され，見出しとして機能している記事名に注目する．記事名を時系列で大量に取得し，分析すること

で，キーワードに関する歴史的文脈を客観的に捉えることができるという仮説を立てている．

2.3 インターネット情報の取得方法

インターネットの WWW 上には無数のウェブサイトがあり，そこで提供されるウェブサービスがある．ウェブサイトに対する情報検索について以下の方法を試行し，得られる情報の傾向について整理した．

ウェブサイトに存在する情報は，デジタルデータであるが，その形態は入手経路や取得可能性，公開性等により分類される．個別ウェブサイト url の入手経路は，受動的な場合であり，ポータルサイト，SNS により url を取得する方法である．能動的な url 入手経路は，キーワードを使い検索エンジンを使う方法や専用データベース･サービスを利用する方法である．また，取得可能性による分類，例えば「無料か，有料か」という分類も考えられる．さらに言えば，インターネットに接続している情報であっても，公開性のある一般に公開している情報なのか，企業等の私的組織内での管理下の情報なのかという分類も考えられる．

インターネット上のサービスは，個人が無料で使用できる代わりに，サービスの仕様が明らかになっていない部分が多いことがある．そこで，キーワードを「visualization」として，ウェブ情報取得方法について以下の複数の方法を試行し，得られる情報の傾向について整理した．

表 1 ウェブ情報取得方法

サービス概要	サービス名	データソース
通常のウェブ検索	Google	WWW
専門的なウェブ検索	Google Scholar	WWW
専門データベース	放送大学ディスカバリーサービス	EBSCO 社

2.4 ウェブ情報の評価方法

分析に足る大量にデータを取得する場合に，以下のような評価項目が考えられる．

① 記事等の結果の件数
② 取得可能ページ数
③ 取得可能件数
④ ページの定型性（プログラムでの取得容易性）
⑤ １件の記事等の定型性（プログラムでの取得容易性）
⑥ １件あたりの，属性情報，リンク（リンクを飛ばずに取得できる属性）
⑦ 取得性，手動，自動，リスク

2.5 ウェブ情報取得方法の検討

2.5.1 方法１：Google ウェブ検索サービス(7)

一般的なウェブ検索サービスとして Google を利用した．Google は，インターネットの WWW 上のウェブサイトから，ある条件に基づき複数のページを「スパイダー」,「ロボット」呼ばれるプログラムを使って自動的に取得していく「ウェブクローリング」によって世界中のウェブサイトを巡回し，自社の大規模サーバーにページをコピーし，独自のデータ整理とデータ抽出ルールに基づく検索サービスを提供している。キーワード「visualization」を使った検索結果の表示画面の冒頭が図１であり，結果の概要を表２に示す。

2.5.2 方法２：Google Scholar(8)

“Google Scholar”は，Google 社が提供する学術文献用ウェブ検索サービスである。WWW にウェブサイトとしてオープンされた複数の文献データサービス等を大量に組み合わせたデータベースにより学術論文に関する検索サービスとなっている。キーワード「visualization」を使った検索結果の表示画面の冒頭が図２であり，結果の概要を表３に示す。

2.5.3 放送大学ディスカバリーサービス(9)

情報を取得するために，最も簡単な方法はその情報に関連するデータベース・サービスに対して，キーワード等によりデータを検索し，結果を取得するという方法である。ウェブブラウザの画面から必要項目を打ち込み検索することで，検索結果がウェブページ，つまり html 形式で表示される。

これについては，放送大学附属図書館が提供している文献検索サービス「放送大学ディスカバリーサービス」を使うこととした。キーワード「visualization」を使った検索結果の表示画面の冒頭が図３であり，結果の概要を表４に示す。

3. 考察

複数のウェブ情報取得方法の検討から，以下のことがわかった．

3.1 方法１（Google）による情報取得

Google の検索結果及び検索機能では，今回の研究目的に不向きである．

3.2 方法２（Google Scholar）による情報取得

Google Scholar では論文発表年での検索条件で，検索結果を分割しながらウェブクローリングでページを取得していくこと試みた。しかし，検索結果は約 10 万件であったとしても，1 ページ 100 件で最初の 1000 件までしか html ページを表示されないことがわかった。Google Scholar では，検索可能ページ数，件数に限定があり，今回の研究目的に不向きである．

3.3 方法３（放送大学ディスカバリーサービス）による情報取得

放送大学ディスカバリーサービスでは，ページ数が多いため，手動でのデータ取得が困難である．

「タイトルに『visualization』を含む」という条件で放送大学附属図書館の「ディスカバリーサービス」で検索すると，約 10 万件の検索結果がある。1 ページで最大 50 件のため，2000 ページ程度を保存する必要がある。これは 1 ページの保存作業に 2 分かかるとして，4000 分（＝60 時間）であり，余り実用的な方法とは言いがたい．そこで，プログラミングによる自動取得やデータ整形を表 5 の方法で検討した．

4. 今後の課題

インターネット上の公開情報は比較的短い期間での個人の欲求や社会の傾向に関する情報が広く分布しているため，情報検索の結果は「刹那的な」に最適化されている傾向があることがわかった．その意味では，個人がアクセスできる長期間の時系列的な情報の蓄積所としての公共機関やデータベース・サービスは重要と考えられた。

インターネット上のデータベースサービスとして検索結果を一括で保存する機能を活用してデータを入手した．これは方法３に対して慎重にウェブクローリングしたデータと同等のデータであると考えられる．今後，このデータ（XML 形式）を整理して分析を実施することを検討していきたい．

図 1 方法１の試行結果の冒頭（Google）

表 2 方法１の結果概要

検索結果特性	結果等
検索ヒット数	約 54, 600, 000 件
上位 1 位	辞書サービス
上位 2 位	Wikipedia
上位 3 位	学会
上位 4 位	最近の技術紹介
上位 5 位	最近の技術紹介
期間設定	最近 1 年程度
検索可能ページ数	30
検索可能件数	約 300 件
年代順並び替え	不可
件名限定検索結果	約 398, 000 件

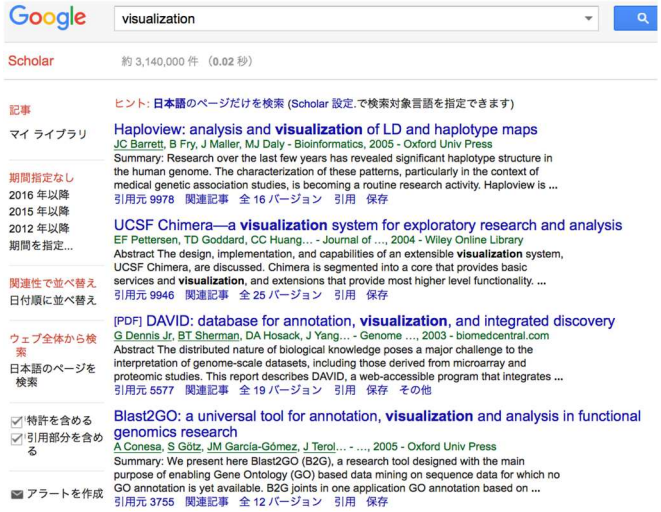

図 2 方法２の試行結果（Google Scholar）

表 3 方法２の結果概要

検索結果特性	結果等
検索ヒット数	約 3, 140, 000 件
上位 1 位	Oxford Univ Press
上位 2 位	Wiley Online Library
上位 3 位	オープン文献サービス
上位 4 位	Oxford Univ Press
上位 5 位	Oxford Univ Press
期間設定	任意
検索可能ページ数	10
検索可能件数	約 1000 件
年代順並び替え	可能
件名限定検索結果	約 135, 000 件

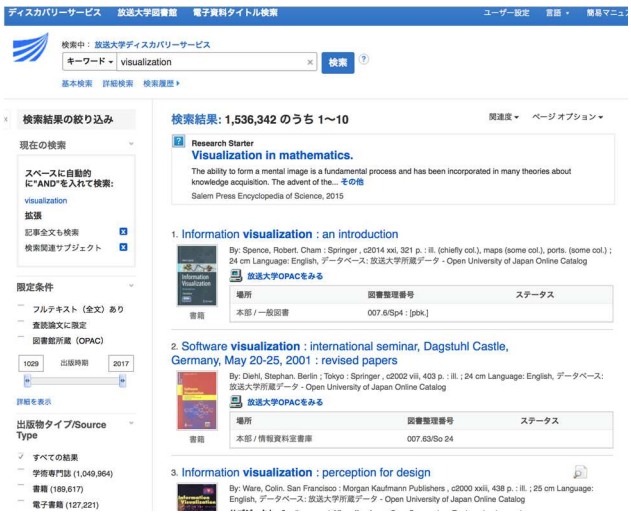

図 3 方法３の試行結果
（放送大学ディスカバリーサービス）

表 4 方法３の結果概要

検索結果特性	結果等
検索ヒット数	1, 536, 342 件
上位 1 位	関連書籍
上位 2 位	関連書籍
上位 3 位	関連書籍
上位 4 位	関連書籍
上位 5 位	関連書籍
期間設定	任意
検索可能ページ数	全て
検索可能件数	全て
年代順並び替え	可能
件名限定条件	100, 738 件

表 5 プログラミングツール

目 的	ツール名称	概 要
クローリング	CasperJS	ヘッドレスブラウザを操作するためのスクリプト
スクレイピング	Python	汎用プログラミング言語
	BeautifulSoup4	スクレイピング用ライブラリ

参 考 文 献

(1) 総務省: “ＩＣＴコトづくり検討会議 報告書”,(2013)
(2) 岡崎, 松尾, 石塚 : ” 関連する複数新聞記事からの重要文抽出法”, 第 3 回 AI 若手の集い, 人工知能学会誌, 第 17 巻, 第 5 号, pp.646-648 (2002)
(3) Wikipedia: https://ja.wikipedia.org/
(4) 山田: “超大型科学用コンピュータ（スーパーコンピュータ）”, 日本航空宇宙学会誌, vol. 28, no. 318, pp.10−16 (1980)
(5) 白山, 桑原: “計算流体力学における可視化技術とその問題点,” 日本物理學會誌, vol. 45, no. 7, pp.483–490 (1990)
(6) M.Cox, D.Ellsworth: “**Managing Big Data for Scientific Visualization**”, ACM Siggraph (1997)
(7) Google: https://www.google.co.jp
(8) Google Scholar: https://scholar.google.co.jp/
(9) 放送大学ディスカバリーサービス: http://eds.a.ebscohost.com/eds/Search/