大学機関リポジトリのアクセスログに現れる検索語と論文題名との関係について

木下，仁

馬場，謙介

伊東，栄典

廣川，佐千男

出版情報：情報処理学会火の国情報シンポジウム2011, pp.B-3-2-, 2011-03. 情報処理学会
バージョン：
権利関係：

大学機関リポジトリのアクセスログ解析

木下仁[†]　馬場謙介[††]　伊東栄典[†]　廣川佐千男[†]

近年、研究成果公開とアーカイブのために機関レポジトリが充実してきている。自分の論文のアクセス状況は、研究者にとっても意味のある情報といえる。本発表では、アクセスログに現れる検索語と論文題名の関係について分析結果を紹介する。

Analysis of access log of university organization repository

Jin Kinoshita[†]　and Kensuke Baba[†]　and Eisuke Ito[†]
and Sachio Hirokawa[†]

Recently, organization repository has been enhanced for the research result opening to the public and the archive.
It can be said that the access status of own articles is significant information for the researcher. In this paper, we report the analysis result about the relation between the retrieval word and the article title.

1. はじめに

大学や研究機関において、「オープンアクセス」の概念に基づく活動の一つとして研究成果の保存と発信を行う機関リポジトリの設置と整備が進んでいる。

機関リポジトリ（Institutional Repository，以下 IR）とは、研究機関がその知的生産物を電子的形態で集積し保存・公開するために設置する電子アーカイブシステムであり、北海道大学の HUSCAP[4]や筑波大学の Turips-R[5]が挙げられる。IR の設置には 2 つの目的がある。まず、機関の研究成果を自主的に保存・公開することにより、オープン・アクセス化に寄与すること、また、出版されないものや失われやすいもの（灰色文献。例えば学位論文や研究報告書類）を確保し保存していくことが挙げられる。

IR のアクセスログ解析は、ダウンロード数やアクセス元などの基礎分析は行われている[1],[2]が、ログ中の検索語を用いた研究は行われていない。また、アクセスログを用いた研究者のためのシステムの実装は行われている[3]が、本研究では利用者に対するシステム案を考えた。

九州大学では学術情報リポジトリ QIR [6]の運用を 2006 年 4 月から開始し, 学内研究者の著作文献を蓄積・公開している。しかしながら，QIR の論文登録数は十分ではない。九州大学の研究者情報データベース[9]には，学内の研究者が書いた論文のタイトルが多数記録されているが、そのような論文の多くは QIR に蓄積されていない。

原因の一つとして、現在 QIR に論文を登録することによる有効性があまり明らかにされていないため、登録を行う研究者が少ないのではないかと考えた。そこで、本研究では QIR のアクセスログの解析を行い、論文登録数の増加に繋がる手掛かりを探った。

具体的にはまずアクセスログ中の検索語と論文タイトルとの関連を分析した。論文著者が考えている自身の研究内容を表わすキーワードと、検索者が考えるキーワードを比較することで、世間からどのように期待されているのかが分かると考えた。また、アクセスログの検索語を用いた関連研究者の推薦を行うシステムを提案する。論文を検索した際に、他の利用者が過去に検索で用いたキーワードを使い、関連研究者を推薦する。論文中には出現しないようなキーワードを発見することができれば、結びつけることが困難であった人物を新しい関連研究者として推薦が行えるのではないかと仮説を立て、研究を行った。

九州大学.
Kyushu University

2. 九州大学機関リポジトリ QIR

2.1 QIR

九州大学学術情報リポジトリ（Kyushu University Institutional Repository，QIR） は，九州大学内で生産された知的生産物を保存・公開することを目的とした学術情報資源管理システムであり，平成 18 年 4 月より，主に九州大学附属図書館によって管理・運用されている。QIR では，学術論文の他，会議発表委資料等の著作物も蓄積し，原則として無償で配布され、ユーザー登録をせずに誰でも閲覧できる。

QIR は MIT 開発の Dspace[8]をカスタマイズしたシステムである。QIR では，学術論文の他，会議発表委資料等の著作物も蓄積しており，多くの著作物は公開されている。

2.2 文献数

2011 年 2 月現在、QIR には 15948 件の論文が登録されている。表 1 に 2010 年 9 月末の QIR コンテンツ数を示す。

表 1　QIR 登録文献数の種類と数

種類	文献数	割合
紀要論文	11172	74%
学術雑誌論文	1226	8%
会議発表論文	744	5%
その他	525	3%
テクニカルレポート	429	3%
一般雑誌記事	264	2%
プレプリント	163	1%
会議発表用資料	148	1%
研究報告書	134	1%
図書	116	1%
学位論文	106	1%
教材	34	0%
総数	15061	100%

表 1 に示すとおり，QIR のコンテンツの 75%が学内の紀要論文である。一方，国際会議や学協会論文誌など,外部で発表された論文は 10%程度しか蓄積されていない。

2.3 データ構造

OAI-PMH プロトコルを使うと、XML 形式で、各文献のメタデータを入手できる。OAI-PMH とは、データの自動収集によってメタデータを交換するためのプロトコルの名称である。XML の形式を用いて、HTTP プロトコル上でクライアントとサーバ間のデータ転送を行っている。

QIR の保持する各論文には，それぞれ一意な ID が割り振られており、ID は，「2324/15937」のように，2 つの数字をスラッシュ「/」でつないだものである。2324 は九大の IR であることを示し、15937 は論文の ID である。この ID は「https://qir.kyushu-u.ac.jp/dspace/handle/2324/15937」の形式で使われている。

3. アクセスログ

ここでは、本研究が解析対象とする QIR のアクセスログについて述べる。

3.1 ログの形式

QIR のアクセスログは Apache Web サーバの形式になっている。図 1 にログの一例を示す。この例では IP アドレス 133.11.90.71 から 2010 年 10 月 15 日 16 時 22 分 42 秒にアクセスされたことが分かる。この論文は文献 ID が 15937 であることを示している。

```
133.11.90.71 - - [15/Oct/2010:16:22:42 +0900] "GET /dspace/bitstream/2324/15937/1/itou-dlw30.pdf HTTP/1.1" 200 617717 "http://www.google.co.jp/search?hl=ja&client=firefox-a&hs=VKf&rls=org.mozilla%3Aja-JP-mac%3Aofficial&q=%E4%B9%9D%E5%B7%9E%E5%A4%A7%E5%AD%A6+%E3%82%B7%E3%83%A9%E3%83%90%E3%82%B9%E5%8F%8E%E9%9B%86&aq=f&aqi=&aql=&oq=&gs_rfai=" "Mozilla/5.0 (Macintosh; U; Intel Mac OS X 10.6; ja-JP-mac; rv:1.9.2.10) Gecko/20100914 Firefox/3.6.10"
```

図 1　ログの例

3.2 ログのサイズ

本論文では 2008 年 6 月 24 日から 2010 年 10 月 31 日までのアクセスログを分析した。全体のアクセス数（行数）は 40,975,348 件である。アクセスログには利用者による検索だけでなく、検索エンジン企業がデータ収集のためにアクセスしているクローラー（Bot）のログが多数含まれる。本論文では Bot によるアクセスを除去した結果を分析対象とした。表 2、図 2 に月毎のアクセス数を示す。

表 2 QIR アクセス数

年月	アクセス数	Bot 除去	検索語
2008 年 6 月	228,139	148,424	7,286
2008 年 7 月	888,127	603,654	30,102
2008 年 8 月	958,115	625,863	25,575
2008 年 9 月	1,019,150	781,366	28,084
2008 年 10 月	1,028,906	776,563	32,733

2008 年 11 月	1,098,943	881,358	31,185
2008 年 12 月	1,910,913	1,643,814	26,852
2009 年 1 月	1,802,189	1,497,137	32,136
2009 年 2 月	1,588,639	1,261,470	28,179
2009 年 3 月	1,727,886	1,229,198	28,066
2009 年 4 月	1,467,650	1,237,298	31,158
2009 年 5 月	1,544,752	1,297,575	36,070
2009 年 6 月	1,015,578	832,820	27,517
2009 年 7 月	1,305,121	1,089,496	36,703
2009 年 8 月	1,048,173	824,380	27,293
2009 年 9 月	1,191,433	862,767	27,868
2009 年 10 月	1,350,651	1,001,781	35,268
2009 年 11 月	1,423,128	1,004,269	33,318
2009 年 12 月	993,631	732,227	27,098
2010 年 1 月	1,264,880	882,823	35,963
2010 年 2 月	1,239,528	787,767	29,963
2010 年 3 月	1,253,421	832,615	33,640
2010 年 4 月	1,496,417	1,027,272	43,462
2010 年 5 月	1,474,454	1,023,120	52,579
2010 年 6 月	1,446,687	1,067,283	46,880
2010 年 7 月	1,727,042	1,170,169	36,575
2010 年 8 月	1,633,859	1,239,403	28,475
2010 年 9 月	1,955,813	1,557,449	34,681
2010 年 10 月	3,892,118	3,549,786	48,167

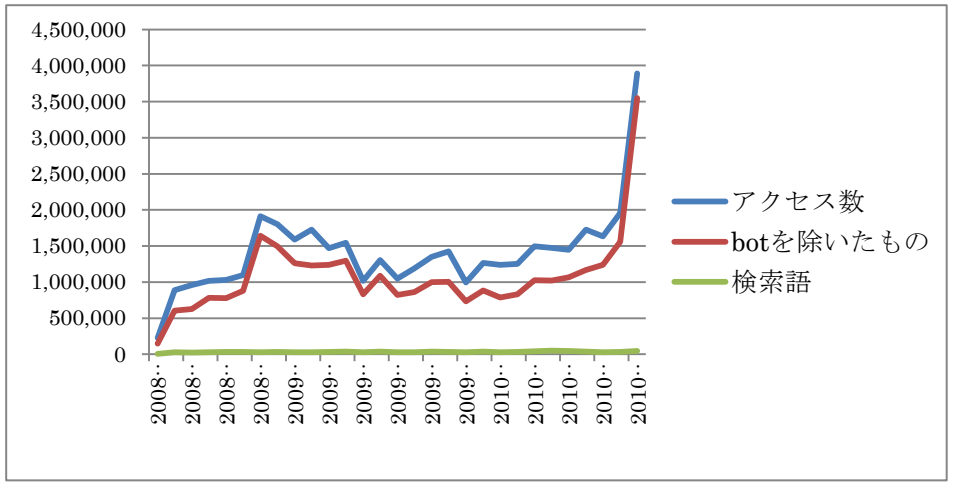

図 2　QIR アクセス数

3.3　検索語の形式

本論文では検索エンジンから QIR に辿り着いた利用者のアクセスログを分析するが、特に検索語について調べる。しかし、URL のパラメータの引数として与えられた検索語は URL エンコードされているため、デコードを行い可読な単語を復元する必要がある。アクセスログのリファラを調べてみると、56 種類の検索エンジンから検索されていることが分かり、ほとんどが google からのものである。表 3 に 2010 年 9 月のアクセス元上位 10 件を示す。

表 3　アクセス元

アクセス元	件数
google.co.jp	30258
search.yahoo.co.jp	12502
hyoka.ofc.kyushu-u.ac.jp	10824
google.com	4858
scholar.google.co.jp	2654
ci.nii.ac.jp	1452
ezsch.ezweb.ne.jp	1344
cgi.search.biglobe.nejp	1044
google.co.in	848

qir.kyushu-u.ac.jp	836

4. 検索語と論文タイトルの関連

4.1 著者と利用者のキーワードのずれ

論文著者が自身の研究内容を表すキーワードは論文の題名に現れると考えられる。論文を求めて検索してくる人の検索語は必ずしも論文題名と一致しているとは限らない。著者の考えるキーワードと、論文を求めて外から検索してくる人の考えるキーワードのずれがどの程度あるのかをQIRのアクセスログにより調べてみた。これにより、著者は自分の論文を求めている人が自身をどのように認識､評価しているかが分かる。また、自分が予想していなかったような思いがけないキーワードを提示することにより新たな研究のきっかけを与えることができる。利用者に対しては、自分が知らなかった有効な検索語を与えることができれば、検索の手助けになる。

4.2 著者と利用者のキーワード合致度の定量化

検索語と論文題名との合致度を表わす 3 つの指標h_1, h_2, h_3を次のように定義した。

論文d_iのタイトルをt_i,著者をa_iとする。d_iをアクセスしたログに現れる検索語の集合を$Q_i = \{q_{i1}q_{i2} \dots q_{ik}\}$とする。また、$T_i$を$t_i$に含まれる単語の集合とする。このとき、

$$h_1(d_i) = \begin{cases} 1, \dots \exists x \in t_i \wedge x \in Q_{i,} \\ 0, \dots \dots \dots \dots \dots \dots \text{else.} \end{cases}$$

$$h_2(d_i) = \frac{|T_i \wedge Q_i|}{|Q_i|}$$

$$h_3(d_i) = \frac{\sum_{w \in T_i \wedge Q_i} tf(w)}{\sum_{w \in Q_i} tf(w)}$$

ただし、tf(w)はアクセスログにおける単語 w の出現回数である。

h_1は1 か 0 の値を持ち、検索語のうち， 1 つだけでもタイトルに含まれる場合 1、含まれない場合は 0 とする。h_2 は 検索語のうち, タイトルに含まれる語の割合である。h_3 は検索語毎の出現頻度を考慮したタイトルに含まれる語の割合である。

4.3 形態素解析

検索エンジンのパラメータとして与えられる検索語には、長い文も多く含まれている。そこでまず、検索語に対して、形態素解析を行う。形態素解析には Mecab[10]を用いた。MeCab は 京都大学情報学研究科と NTT コミュニケーション科学基礎研究所が共同研究で開発したオープンソースの形態素解析エンジンである。Mecab は言語, 辞書,コーパスに依存しないよう汎用的に設計されている。

4.4 キーワード合致度の計算例

表 4 に論文 ID と hit 率の例を示す。例えば、ID が 15937 の論文のタイトルは「Web シラバス統合による教育情報ライブラリ構築」であり、「syllabus」、「情報」、「web マイニング」など 19 個のキーワードで検索されている。形態素解析を行うと 22 個のキーワードとなり、その中の 2 個がタイトルに含まれている。また、アクセスログ中のこれらの単語はそれぞれ 1 回ずつ現れている。したがって、h_1, h_2, h_3は以下のような値となる

表 4 論文 ID と hit 率

ID	h_1	h_2	h_3
15937	1	0.09091	0.09091
15938	0	0	0
15939	1	0.07407	0.06504
1594	0	0	0
15940	1	0.1186	0.1129

4.5 解析結果

図 3,4 に年度毎の hit 率の分布を示す。

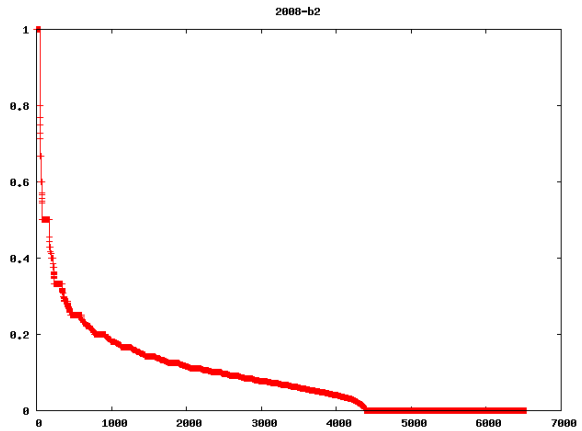

図 3 2008 年度h_2の分布

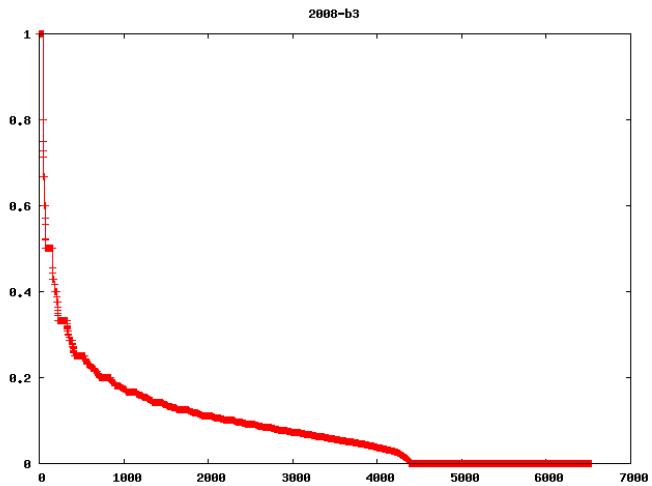

図 4　2008 年度h_3の分布

4.6 考察

論文タイトルとアクセスログ中の検索語とがマッチする割合を hit 率と定義し､2008 年 6 月から 2010 年 10 月までの期間を調べた。3 年とも同じ傾向にあり、年や月毎の変化はあまり見られなかった。結論としては著者のキーワードと利用者のキーワードのずれの大きさは、論文により様々であり、一般的な傾向は見られなかった。

h_2,h_3が 0 のものと 0.5 より上位のものをランダムに 30 件取り出し､値が高いものと低いものに見られる特徴を調べてみた。

h_2,h_3が 0 のものはタイトル以外で検索されている。著者名や本文中に現れる単語や表現の他に、論文によって目次や論文のキーワード、抄録が記載されており検索対象になっている。また、漢字、スペルのミスや大文字、小文字の違いでマッチしていないものがある。

h_2,h_3が 0.5 以上のものはタイトルに論文内容を表わすキーワードが表れている。または､タイトルを直接検索している。また､タイトルが長いか､検索語の数が少ない。特に値が 1 のものは検索語数が 1 つのものが多い。

5. 関連研究者の推薦

ある論文をアクセスした利用者に関連研究者の推薦を行う方法として、共著者の推薦や参考文献の引用元に出てくる人物を推薦する方法が考えられる。共著者推薦や参考文献の人物だけでは推薦される人物が共同で研究を行った人だけに限られる可能性が高い。そのため、論文とは直接関わりはないが、研究分野で類似している論文を推薦することができない。良い論文があったとしても発見されることなく埋もれてしまう可能性がある。

そこで、アクセスログによる検索語を用いた関連研究者の推薦を行うことを考えた。論文を求めている人が検索を行った際に、アクセスログの検索語を用いて再度検索を行うことにより、別の研究者の推薦を行う。これにより、見つけることが困難であった新たな関連研究者の発見に役立てることができる。

推薦の手法を比較するため共著者推薦、論文タイトルのキーワードによる推薦、アクセスログの検索語による推薦を行った。検索語を用いた推薦により、共著者以外の研究者が関連著者として推薦されることが望ましい。

本研究では汎用連想検索エンジン(GETA)[11]を用いた。論文 ID と著者の頻度ファイル、論文 ID とタイトルのキーワードの頻度ファイル、論文 ID と検索語の頻度ファイルを用意しておき、検索を行う。

5.1 汎用連想検索エンジン GETA

汎用連想計算エンジン（GETA）は、文書検索における頻度付き索引データ（どの文書にどの単語が何回出現するというような)を典型とする大規模かつ疎な行列 を対象として、行と行あるいは列と列（具体的には文書間および単語間） の類似度を内積型メジャーで高速計算するツールである。連想検索をはじめ、文書分類、単語間類似度計算など､大規模文書 の分析に必要な要素技術をサポートすることを目的としている。「汎用連想計算エンジン（GETA）」は、情報処理振興事業協会（IPA）が実施した「独創的情報技術育成事業」の研究成果であり、ソースコードを含むプログラム及びドキュメンテーション を無償で提供している。

5.2 共著者による推薦

著者から論文 ID を割り出し、その論文 ID から再度検索を行い、関連研究者として共著者を推薦する。表 5 に「廣川佐千男」を入力とした際の、共著者推薦の上位 10 件を示す。

表 5　QIR における「廣川佐千男」の共著者上位 10 件

共著者	weight
中藤哲也	22.46
山田泰寛	22.31
YamadaYasuhiro	22.31
NakatohTetsuya	21.18
松永吉広	13.69

MatsunagaYoshihiro	13.69
関隆宏	13.45
SekiTakahiro	13.45
野口正人	12.72
NoguchiMasato	12.72

ここで weight は GETA によって定められる類似度の値であり、検索語を含むアクセスログ d における単語 q について、以下のような weight(d|q)として求められる。

$$\mathrm{weight}(d|q) = \frac{\sum_i^n wq(t_i|q)wd(t_i|d)}{\mathrm{norm}(d)}$$

$wq(t_i|q)$は単語t_iの q 中での頻度、$wd(t_i|d)$は単語t_iの d 中での頻度である。それぞれの単語t_iについて$wq(t_i|q)wd(t_i|d)$を掛けた総和が q と d の類似度になる。d の長さが様々であるため、d の長さ norm(d)で類似度の正規化を行っている。

5.3 論文タイトルのキーワードによる推薦

まず、論文著者から論文 ID のリストを返し、ID からそれぞれの論文のタイトルのキーワードを返す。そのキーワードから再度検索を行い、キーワードを含む論文 ID、その論文の著者と辿り、得られる著者を関連研究者として推薦する。表 6、表 7 に「廣川佐千男」を入力とした際の、タイトルのキーワードによる共著者推薦の例上位 10 件を示す。

表 6 「廣川佐千男」のもつ論文タイトルのキーワード

タイトルのキーワード	weight
Web	14.57
Calculus	6.452
検索	6.399
文字	5.786
列	5.387
頻度	5.089
シラバス	4.966
マッシュ	4.966
Databases	4.896
単語	4.801

表 7 論文タイトルから推薦された「廣川佐千男」の関連研究者

関連研究者	Weight
TadauchiOsamu	43.23
タケシタヒトシ	43.07
TakeshitaHitoshi	43.07
多田内修	42.92
竹下齊	41.39
HirashimaYoshihiro	40.15
ヒラシマヨシヒロ	39.92
平嶋義宏	39.92
タダウチオサム	39.80
モリアキラ	39.30

推薦される関連研究者の評価については後の節で詳しく考察する。

5.4 アクセスログの検索語による推薦

論文著者を入力とし、論文 ID のリストを返し、ID からアクセスログ中の検索語をリストとして返す。その検索語を用いて再度検索を行い、検索語を含む論文の ID、その論文の著者と辿り、得られる著者を関連研究者として推薦する。表 8、表 9 に「廣川佐千男」を入力とした際の、アクセスログの検索語による共著者推薦の例上位 10 件を示す。

表 8 「廣川佐千男」のもつアクセスログ中の検索語

アクセスログの検索語	weight
Web	14.91
文字列	14.63
mashup	13.81
Wrapper	12.34
XML	12.10
WEB	12.03
html	11.56
特徴語	11.29
Hirokawa	11.27
crawler	10.83

表 9　アクセスログ中の検索語から推薦された「廣川佐千男」の関連研究者

関連研究者	weight
江口弘美	42.46
エグチヒロミ	42.46
EguhiHiromi	42.46
北野雅治	38.69
KitanoMasaharu	38.63
キタノマサハル	38.61
江頭和彦	35.84
EgashiraKazuhiko	35.84
KaiSatoshi	35.21
HeskethJ.D.	35.15

著者からランダムに 100 人を選び、共著者以外の人物が推薦されるかを調べた。84%の著者から共著者でない関連研究者を推薦することに成功した。しかし、研究内容が関連していない人物を推薦してしまっている。そこで、著者と推薦された関連研究者とのキーワードの関係について調べてみた。

5.5 著者と関連研究者の共起語

「廣川佐千男」の論文 67 件と「江口弘美」の論文 54 件において、検索語はそれぞれ 1319 個、2903 個ある。両者に現れるキーワードは 176 個存在する。この 176 個のキーワードをみると、「system」や「pdf」、「kyushu」といった研究内容を表わさず、他の多くの論文に現れうる単語が多く目についた。そこで各単語の出現頻度を調べた。表 10 にその上位 10 件を示す。

表 10　「廣川佐千男」と「江口弘美」に現れる単語と出現頻度上位 10 件

単語	出現頻度
of	92
in	55
and	54
and	42
on	42
is	32
by	24
pdf	19
&	17
japan	9

5.6 考察

共著者推薦、論文タイトルのキーワードによる推薦、アクセスログの検索語を用いた推薦という関連研究者の推薦を行う際の 3 つの手法を考えた。特にアクセスログの検索語による推薦は論文の著者ではなく、読み手の方が考えたキーワードを用いている。そのため、関連研究者として推薦される人物は著者も思いがけない人が推薦されるのではないかと仮説を立て研究を行った。

アクセスログ中の検索語を用いて共著者以外の関連研究者を推薦することは、高い割合で成功している。しかし、推薦された研究者は研究内容が一致しておらず、共著者でない新しい研究者を推薦するという期待していた結果が出なかった。

著者と推薦された研究者間の検索に使われたキーワードを見てみると、「of」、「in」、「and」といった研究内容を直接表わさず、他の論文にも多く現れるような単語で検索が行われていることが分かった。そのような単語は、研究内容を表わす単語に比べると出現頻度が圧倒的に多いため、研究内容が異なる研究者を関連著者として推薦してしまっていると考えられる。

このシステムを実装するには、直接意味を持たない単語をあらかじめ除いておき、著者や論文の内容を表わす単語を用いて行う必要がある。また、アクセスログのデータは年々増えていくのでこの研究を継続することで研究者間の結びつけが新しく見つかり、発展していくことが期待できる

6. おわりに

九州大学では学術情報リポジトリ QIR が 2006 年 4 月から運用されているが、論文の登録数は不十分であり、ほとんどが学内の紀要論文である。原因の一つとして、QIR に論文を登録することへの有用性が明らかにされていないからではないかと仮説を立て、アクセスログの分析を行うこととした。

アクセスログの解析として、論文タイトルと検索語の関連についての研究とアクセスログ中の検索語を用いた関連研究者の推薦を行った。論文タイトルと検索語の関連は、著者のキーワードと検索者のキーワードとの間にずれがどれほど存在しているかを明らかにすることで、著者への評価の提示や検索支援に繋がるのではないかと考えた。ずれの大きさは論文により様々であるため、一般的な傾向はあまりみられなかった。アクセスログ中の検索語を用いた関連研究者の推薦では、共著者でない人物を関連研究者として推薦することは成功していた。しかし、意味を持たず、出現頻度の多い単語が多数含まれていたため、著者と研究内容が関連しない人物を推薦してしまっていた。

今後このシステムを実装するためには、論文内容に関わりがなく、意味をなさない単語を排除する必要がある。

参考文献

[1]馬場謙介,伊東栄典,吉松直美,星子奈美, 機関リポジトリの有効性分析, DEIM Forum 2010 F7-3,2010
[2]佐藤翔,逸村裕 機関リポジトリ収録コンテンツにおける利用数とアクセス元、アクセス方法、コンテンツ属性の関係, 三田図書館・情報学会研究大会発表論文集, pp.9-12,2009
[3]井上創造,藤井達朗,小林健,池田大輔, 学術情報リポジトリ活性化のための足跡機能, 九州大学附属図書館研究開発室年報 2007/2008, pp.17-22, 2008
[4]HASCAP 北海道大学学術成果コレクション,http://eprints.lib.hokudai.ac.jp/dspace/index.jsp
[5] Turips-R つくばリポジトリ,http://www.tulips.tsukuba.ac.jp/dspace/
[6] QIR 九州大学学術情報リポジトリ, https://qir.kyushuu.ac.jp/dspace/
[7]Ranking Web of World Repositories,http://repositories.webometrics.info/ (Aug. 12,2010).
[8]Dspace,http://www.dspace.com/ja/jpn/home.cfm
[9]九州大学研究者情報,http://hyoka.ofc.kyushu-u.ac.jp/search/index.html
[10] MECab, http://mecab.sourceforge.net/
[11] 汎用連想計算エンジン GETA,http://geta.ex.nii.ac.jp/geta.html