データマイニングを用いた情報検索手法の研究

井原雄人*1・永田勝也 1
Email: ihara@akane.waseda.jp

*1: 早稲田大学大学院環境・エネルギー研究科

◎Key Words 情報検索，データマイニング

1. はじめに

インターネット上で扱われる情報量は急速に増大しており、日本国内のインターネット上を流通するトラフィック量を国内の主要 ISP6 社の値から推計すると、2008 年には 1373bps であったものが、2011 年には 2365Gbps にまで増加していることがわかる。2009 年に一時減少傾向が見られたが、これは動画などの大容量データの送信に当たり従来は P2P が主流であったのに対して、配信サーバを中核とした形式に変容したためである。また、近年のクラウドコンピューティングによる XaaS の普及より今後も増大し続けるものと予想される。

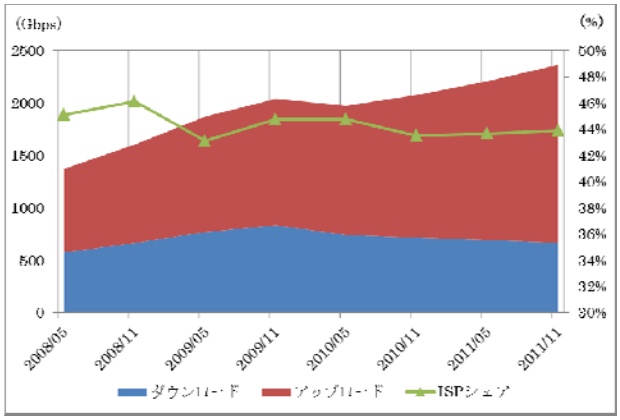

図1 インターネットトラフィックの現状

大学教育において、論文や講義内でのレポートを作成するにあたり、関連情報をインターネットを用いて検索をすることは当たり前として行われているが、情報量の増大により、有意な情報にたどりつくまで時間がかかる現象を発生させている。

そこで本研究では、ベイズ定理を用いた関連語のフィルタリング、検索に用いられているキーワードの絞り込みを行うと、ランダムサーチによる検索アルゴリズムの効率化を組み合わせることで、教育に必要な情報に効率的にたどり着くことのできる情報検索手法を提案する。

2. 情報検索の現状

2.1 検索キーワード数

膨大なインターネット上の情報に対して、ユーザーは自らに必要な情報を探すために、Google や Yahoo といった検索ポータルサイトを利用することが多い。一般的に検索ポータルサイトを利用した検索は、検索ウィンドに対してキーワードを入力することによって行われ、AND 検索や OR 検索などの機能が実装されている。

図 2 及び 3 に示すのは、google.co.jp において、検索において使用された word 数を示したものである。

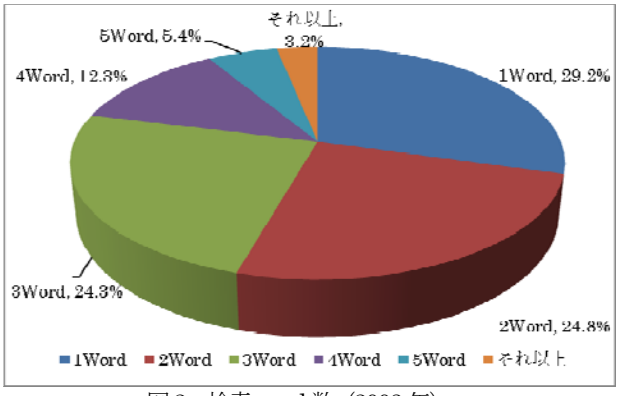

図2 検索 word 数（2003 年）

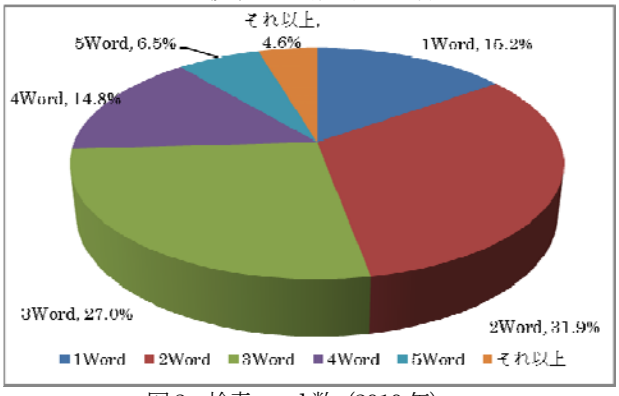

図3 検索 word 数（2010 年）

2003 年と 2010 年を比較した場合、平均 word 数においても 2.47word から 2.79word に増加しており、特に 1word で検索されたものが 2003 年において、3 割を占めていたものが 2010 年では半減している。

この結果からも検索ポータルサイトを利用した検索を行う上で、インターネット上の情報量が増大するにつれ、有意な情報にたどり着くためには、複数のキーワードを用いて検索を行うことが必要不可欠となってきているといえる。

2.2 検索クエリ数

次に、上記の検索 word 数を踏まえた検索クエリ数について示す。実際に検索を行う際のユーザーの行動をして、有意な情報を得るためにまず、少ない word 数で検索を行い、網羅的に得られた情報に対して、検索 word を加えることで情報絞り込みを行っていくこととなる。

つまり 1 回の検索行動において、最終的な優位な情報にたどり着くために複数回の検索が行なわれていることになる。

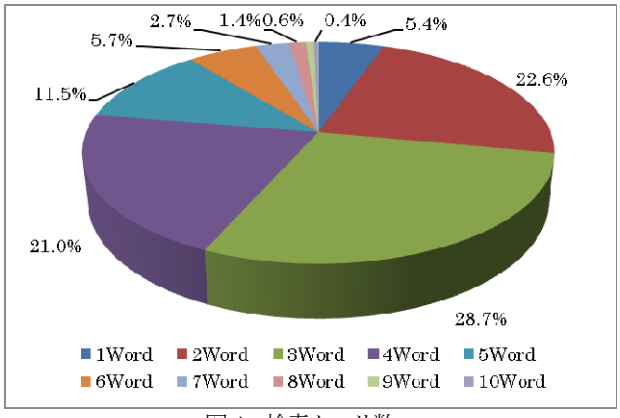

図 4　検索クエリ数

これを踏まえて、2010 年に実際に必要な検索クエリ数の割合を示したものが図 4 である。単純な検索 word 数の増大を示した図 2 及び 3 に比べて、さらに少ない word 数の検索が占める割合が少なくなり、多数の word による検索が重要となってきていることが分かる。

3.　ベイズ定理によるフィルタリング

3.1　キーワードの自動抽出回数

検索にかかる word 数を増加による利便性の低下に対応するための機能として、最初に入力されたキーワードに対して、自動で関連のキーワードを付加するシステムの開発が進められている。しかし、前述の google などでの自動抽出は、ユーザーの検索履歴になど依存して関連度の重み付を行っているため、過去に検索を行ったことのあるカテゴリの検索においては効果が大きいが、新たなカテゴリにおいて検索を行った場合には誤った結果をなることが多い。そこで本研究では、ユーザーの検索履歴に依存せず、データマイニングにより検索対象文章内の単語同士の関連度を用いて判定することとした。

有意な情報にたどり着くまでに必要な関連語の word 数を前述の検索クエリ数の割合より定める。図5は2010 年に google.co.jp 上で「環境」をキーワードとして検索をした 22,940,000 回に対して、検索クエリ数を当てはめたものである。

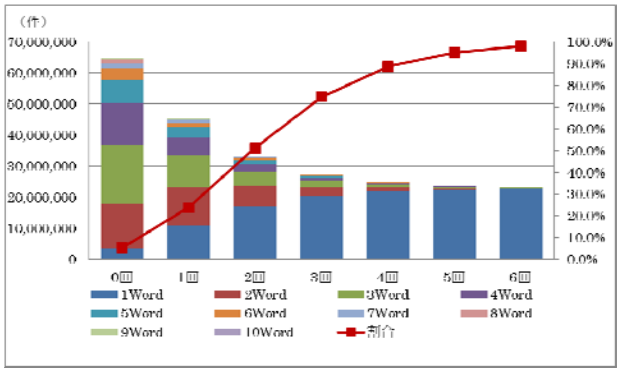

図 5　自動抽出回数と検索クエリ数

自動抽出を行わない場合の検索クエリ数は、64,782,560 回であるのに対し、自動抽出を 6 回まで行うことで 23,215,280 回まで削減され、検索総数の 98%を占めることが出来るようになるために、関連度の自動抽出は 6 回が適当であることが分かった。

3.2　ベイズ定理による判定

抽出された関連語の正誤判定においては、ベイズ定理を利用した重み付けと尤度関数を用いた補正を行う。ベイズ定理とはメールソフトにおける spam メールの判定にも使われている方法であり、情報検索における正誤の判定においては有効性の高い方法である。

また、一度検索を行ったものに対して実際にユーザーが想定した検索結果と異なっていた場合、単純な確率密度による判定を行うだけでなく、前述の事前確率に尤度によるもっともらしさを加えることで判定することができるために、検索のたびに検索精度の向上に繋がると考えられる。

実際に前述の「環境」をキーワードとして検索を行った場合を当てはめると、

$p(w_i|C)$　・・・式①

検索対象となる文書内の C という文節に対して文書内の w 番目に単語 i(今回の場合であれば「環境」)が存在すると式①により仮定する。

$$p(D|C) = \prod_i p(w_i|C) \quad \text{・・・式②}$$

次に、i 対する関連語が C に依存しない、新たな文節 D に存在する可能性は式②で示すことができ、

$$\ln \frac{p(S|D)}{p(\neg S|D)} = \ln \frac{p(S)}{p(\neg S)} + \sum_i \ln \frac{p(w_i|S)}{p(w_i|\neg S)} \quad \text{・・・式③}$$

$$\ln \frac{p(S|D)}{p(\neg S|D)} > 0 \quad \text{・・・式④}$$

式③および④により、0 以下である場合は関連度がないものと判定することとした。

4.　ランダムサーチによる効率化

4.1　ステップ数の削減

従来の検索手法[6]における関連語の抽出は、Suffix Tree Clustering Algorithm（以下、STCA）を用いた全文検索により行われていた。

STCA による検索は、文節や熟語単位でなく単語単位で全文検索であり、精度が高い検索を行なえる反面、検索対象となるデータベースの情報量の増加に比例して検索時間が増大するという欠点があった。

その原因として、検索対象の文章中に関連語が含まれている、いないにかかわらず文章全てを検索してしまうことが考えられる。それに対し本研究では、判定式の一般化を関連語の有無の判定を事前に行う機能を実装することで負荷軽減を行った。

表 1 に示すのは、文章「A」において「環境」に対する関連語の有無を判定する場合のアルゴリズムのステップ数を一般化したものである。

図 5 で示したように関連語の自動抽出を 6 回行った場合、6n+15 回のステップ数が必要であるのに対して、文章中の M 番目に関連語が存在する場合は 4M-2 回で判定することが可能であり、存在しない場合は n=単語としたときに 4n+5 回で存在しないことを証明できることが分かった。この結果より判定式の一般化を行うこ

とで33%程度のステップ数削減が行われた。

表1　判定式の一般化

	存在する場合のステップ数	存在しない場合のステップ数
procedure SEARCH (A,n,環境)	1	1
k←1	1	1
while k<n do	M	n+1
if A[k]=環境 then	M	n
return	1	1
end if	0	0
k←k+1	M-1	n
end-while	M-1	n
return	0	0
end procedure	1	1
合計	4M-2	4n+5

4.2　関連語の有無の事前チェック

上記の判定式の一般化はステップ数の削減には貢献するが n=単語数に依存するために検索対象が大きくなるほど効果が失われることとなる。

そこで関連語の有無を事前に判定することで、単語数に依存しない形のステップ数の削減を行う。文章「A」中に関連語がある場合、k=n+1 番目の単語が関連語であるという形で表すことができる。表 2 に表 1 の判定式に事前判定を踏まえ一般化したものである。

表2　事前判定を含む判定式の一般化

	ステップ数
procedure SEARCH (A,n,環境)	1
A[n+1]←環境	1
k←1	1
while A[k]≠環境 do	M+1-S
k←k+1	M-S
end-while	M-S
if k<n then	1
return	S
end-if	0
return	1-S
end procedure	1
合計	3M-3S+7

関連語を有無について S=1 の場合存在する、S=0 の場合存在しないとした場合、3M-3S+7 回のステップ数で判定することができる。表 1 で示したものに比較した場合、アルゴリズムに関わるプログラムの行数は増加する結果となっている。しかし、表 1 を表 2 と同様に S=1or0 で表した場合 4M-3S+5 回であることから、両者を比較した場合(3M-3S+7)- 4M-3S+5=M-2 と表すことでき、文章中の 2 番目までに関連語が存在する場合は、単語数に依存せず、表 2 で示した事前判定機能を用いた方が 25%程度のステップ数の削減を行うことが可能となった。

5.　有効性の検証

5.1　研究者データベースでの実証

提案手法の有効性を検討するため早稲田大学に所属する2183名の研究者が登録されている研究者DBから「環境」に関連する研究者の検索を行い、従来使われているGoogle APIを用いた検索と提案方式の検索を結果の比較を行った。

Google API は、JavaScript を使用して個別のウェブページに Google 検索を組み込むことができる API で、早稲田大学のウェブページでも採用されている。過去の検索におけるクリック数、被リンク数や meta 情報などを参考に検索順位が決定される。その結果、実際に内包されている情報より、過去の検索ユーザー動向やページ制作者が意図的に組み込んだキーワードなどが優先的に表示される傾向があるといえる。

検索結果として 2183 件中 756 件の抽出データが得られた。また、その中で実際に「環境」というキーワードの使用数を併記したものを表 3 に示す。

表3　googleAPI による検索結果と word 数

順位	名前	″環境″word 数
1	大塚　直	32
2	小松　幸夫	59
3	大和田　秀二	32
4	吉田　徳久	20
5	香村　一夫	43
6	寺島　信義	33
7	川上　泰雄	54
8	川名　はつ子	40
9	斉藤　修	40

この結果からも分かるように、Google API による検索結果は直接「環境」をキーワードに使っているページが優先的に検索されている結果となっているといえる。しかしながら、実際に研究内容を参照してみると、環境を「environment」の意味で使用している以外にも、「ネットワーク環境」や「システム環境」といった「scene」という意味合いで使われている「環境」が word 数としては抽出されているものも多く見受けられ、必ずしも適切な検索結果であるとはいえない。

次に、今回の提案方式を用いた検索を行う過程によって得られた「環境」と共に使用される頻出関連語を表 4 に示す。

表4　頻出関連語

順位	関連語	頻出		順位	関連語	頻出
1	エコ	136		11	河川環境	24
2	自動車	78		12	森林環境	24
3	廃棄物	46		13	影響評価	23
4	環境保全	44		14	化学物質	18
5	自然環境	44		15	都市環境	16
6	温暖化	40		16	住環境	13
7	生態系	34		17	ﾓﾋﾞﾘﾃｨ	6
8	環境教育	32		18	蓄電	6
9	観光	30		19	資源活用	6
10	資源循環	29		20	代替ｴﾈﾙｷﾞ	3

「自然」や「都市」といった熟語は、単独で使用した場合でも名詞となってしまいことから、は改めて「環境」と組み合わせたキーワードとして抽出している。また、抽出数が 2 件以下のものを含めると 436 件のキーワードが該当することとなり、関連付けに用いるには適当でないと判断をして除外することとした。さら

に表 4 の頻出単語によって重み付けされた後の検索結果を表 5 に示す。

表 5　提案方式による検索結果と word 数

順位	名前	″環境″word 数	関連語数
1	北山　雅昭	28	10
2	後藤　春彦	20	10
3	斉藤　修	40	10
4	村山　武彦	12	8
5	森川　靖	22	8
6	天野　正博	23	7
7	高口　洋人	8	7
8	田邉　新一	12	7
9	常田　聡	21	7
10	名古屋俊士	40	7

検索結果の件数は、単純にキーワードが存在するかしないかで選別しているため 756 件と Google API による検索と同一数となった。

しかし、検索順位については使用されている環境 word 数との相関は見られず、必ずしも環境と単語の頻出数ではなく、関連語が多く含まれるものほど上位に来ることが分かり提案方式の有効性が確認された。

さらに両者の検索結果において Google API による検索結果より提案方式の検索に結果において大きく順位が変動した研究者 4 名を抽出したものが表 6 となる。順位 1 が Google AP での検索順位、順位 2 が提案方式による検索順位である。

表 6　順位変動

順位 1	順位 2	名前	専門分野
6	413	寺島　信義	情報
7	403	川上　泰雄	スポーツ工学
8	136	川名　はつ子	福祉環境
10	168	嶋本　薫	情報

各研究者について研究内容を調査した結果、それぞれの研究者の専門分野は環境を「environment」として捉えている内容でなく、「scene」として捉えている情報系、人間工学系を専門としている研究者であることが分かった。このことからも、提案方式による頻出関連語を基とした検索手法が、Google API を使用した検索手法と比較をした場合より、的確な情報の抽出を行っているといえる。

5.2　一般検索への適用

提案手法を google などの一般的な検索ポータルサイトへ適用した場合の有効性について検証を行った。

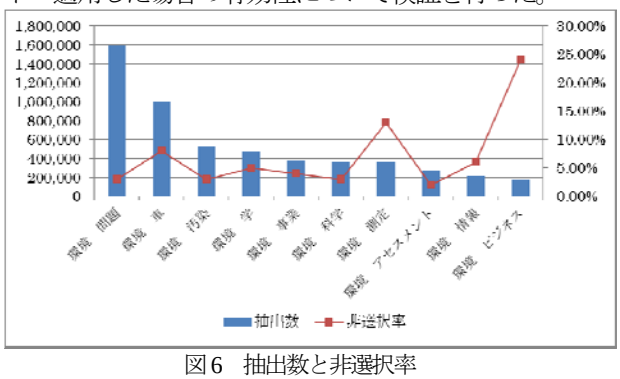

図 6　抽出数と非選択率

図 6 に示すのは google.co.jp で「環境」の関連語を検索した場合の抽出数と非選択率を示したものである。一般検索への適用の場合 5.1 のように検索結果を細かく検証することが困難であることから、検索結果が実際にクリックされたかにより評価することし、非選択率とは表示された検索結果がクリックされず、異なる関連語で検索された割合である。

前述のように google.co.jp 上で「環境」をキーワードした検索は 22,940,000 回/年行われているが、2010 年の 1 年間で関連語として抽出された「問題」においては 1,600,000 件程度の抽出結果が得らえた。

提案手法により抽出された関連語上位 30 件における非選択率の平均は 9.8%であることがわかり、図 3 で示されるように 3word 以上の検索の割合が 52.9%であることから、提案手法による検索は、検索効率の向上において十分な成果を上げていると考えられる。

6.　まとめと今後の展望

本研究で提案された検索手法は、既存の検索手法に比べて効率的に検索が行えることが分かった。しかし、小規模のデータベースにおいては、精度を高く、かつ効率的に検索することができる一方で、大規模なデータベースで使用する場合においては、関連語の自動抽出における効率化は行えるものの、実際の検索時間の削減については実証されていない。

これは、大規模なデータベースに対する検索は、本研究で行われたようなソフトウェア的な改善とサーバやネットワークといったハードウェア的な改善を比較した場合、その寄与度は対象が大規模になればなるほどハードウェアに偏るためである。

今後は、検索対象を徐々に大規模化しつつ、ステップ数だけでなく、実際の検索時間を対象とした、ユーザー効率の定量評価を行っていく必要がある。

参考文献

[1]細野公男、緑川信之、岸田和明, (2007), 情報検索の認知的転回, 丸善株式会社
[2]水田正弘、南弘征、小宮由里子, (2007) ,確立モデルによる Web データ解析法,森北出版株式会社
[3] 総務省 (2010) 情報通信白書,電気通信の利用状況 188-192
[4] Liu Y 他(1998)学術情報検索における未知語の分類とその処理, 情報処理学会全国大会講演論文集, 巻: ５７ｔｈ 号: ３ 頁:3.219-3.220
[5] 藤井敦, 伊藤克亘, 秋葉友良. 事典的 Web 検索サイトの構築. 言語処理学会第 9 回年次大会発表論文集, pp. 129-132, 2003.
[6] 井原雄人,永田勝也.データマイニングを用いた検索手法の有効性に関する研究 日本教育工学会第 26 回全国大会論文集, pp.229–232, 2010
[7] 藤井敦,石川徹夜. World Wide Web を用いた事典知識情報の抽出と組織化電子情報通信学会論文誌, Vol. J85-D-II, No. 2, pp. 300–307, 2002.