情 報 学 基 礎　66－14
デジタル・ドキュメント 32－14
（２００２．３．１５）

時間間隔を用いた検索履歴のモデル化

鈴木　俊輔　　山名　早人
早稲田大学　理工学部　情報学科

Google など既存の検索エンジンでは、あらかじめ生成しておいたインデックスに基づいた検索を行う。したがって、同一の検索語で検索した場合、ユーザの求める情報が異なっても同一の検索結果が返される。このような問題を解決するために、本稿では、検索履歴からユーザの意図を推測することを目的として、検索履歴の解析を行い、ユーザの検索パターンのモデル化を行った。その結果、再検索までの時間間隔を 2 パターンに分類できた。さらに、再検索しているユーザの検索語入力パターンの 91%を 9 パターンに分類することができた。このパターンを利用することにより、検索エンジン側でユーザの意図に合致した検索結果を返すための指針を得ることができた。

Search Pattern Modeling based on its Search Interval

Shunsuke Suzuki　　Hayato Yamana
School of Science and Engineering, Waseda University

The conventional search engines searches based on the pre-generated index. Thus, when some users search with the same query, the search engine returns same result, even if they want to obtain different results. In order to solve such a problem, in this paper, we propose the user modeling scheme based on the user's search pattern to speculate the user's intention. Consequently, we have classified the search interval to re-search into two patterns. Furthermore, we have classified 91% of a user's queries into nine patterns. Using these patterns, the search engines will be able to return the results that suite the user's intention.

1. はじめに

インターネットの普及により、WWW 上に蓄積される情報が急増している。このように急増する多種多様な情報を検索するために Yahoo![1]や goo[2]など数々の検索エンジンが存在する。

検索エンジンとは、検索語をユーザに入力してもらい、あらかじめ生成しておいたインデックスに基づいて検索を実行し、結果をユーザに返すシステムである。しかし、その検索品質には問題があることが指摘されており、品質向上のために様々な工夫が試みられている。その１つに、Google[3]では HTML のリンク構造をランキングに利用する PageRank[4]を利用している。PageRank は、「重要なサイトからの被リンク数が多いほど重要な情報である」という経験則に基づき検索結果をランキングする手法である。Google は PageRank が高く評価され、現在 Yahoo!や Excite[5]などの検索エンジンのバックエンドとし

て利用されている。

Google など既存の検索エンジンでは、前述の通りあらかじめ生成しておいたインデックスに基づいた検索を行う。このように、既存の検索エンジンでは、求める情報が異なっても同一の検索語で検索した場合、同一の検索結果が返されてしまう。しかし、同一の検索語で検索した場合でもユーザによって求める情報は異なる。

このような問題を解決する手法として、検索履歴からユーザの意図を推測する手法が必要となる。そのためにはユーザのモデル化が必要と考え、本論文では検索履歴の解析とユーザの検索のモデル化を行った。

2. 検索エンジンの現状

本節では、既存の検索エンジンの現状と問題点を挙げ、関連研究を紹介する。

2.1 検索エンジンの現状と問題点

現在 Yahoo!や goo など様々な検索エンジンが存在するが、その中で Google がもっとも良いと評価され、Yahoo!や Excite など既存の検索エンジンの多くがバックエンドに Google の検索エンジンを使用している。

しかし、Google など既存の検索エンジンではあらかじめ生成しておいたインデックスに基づいた検索を行う。このため、既存の検索エンジンでは、求める情報が異なっても同一の検索語で検索した場合、同一の検索結果が返される。しかし、同一の検索語で検索した場合でもユーザによって求める情報は異なる場合があり、ユーザがどういう意図で検索語を入力したのかを判断し、ユーザが求める情報を検索結果として的確に返すことが求められている。

2.2 関連研究

2.2.1 検索エンジンのパーソナル化

2.1 の問題の解決策としてユーザに合わせた検索結果の表示が必要である。一例として、ユーザに合わせて検索結果を表示する検索エンジンのパーソナル化が考えられる。ユーザに合わせ検索エンジンをカスタマイズすることによって、検索に関してユーザの傾向や趣向を反映させることが可能になる。しかし、パーソナル化に関してはユーザごとに履歴を取る必要がある。[6]では、ユーザの検索履歴をブラウザで取得することにより、アクセス履歴だけでなくブラウザ上でのアクションなどを得ることが出来た。そして、ＵＲＬごとの統計情報やアクセスパターンからユーザにとって重要度の高いページへの直接リンクを提供している。その結果ユーザごとに異なるランキング表示が実現された。

2.2.2 絞り込み検索の支援

絞り込み検索とは検索結果が多すぎた場合や、ランキングの上位に求める情報がなかった場合に検索語を増やし、求める情報だけを抽出しようという方法である。絞り込み検索が利用できればユーザは求める情報を的確に得ることが出来る。しかし、検索の専門知識を持たないユーザにとって自分の求める情報に関連する的確なキーワードもしくはキーワードの組み合わせを入力することは困難である。そこで、検索エンジン側で絞り込み検索の候補語を提示し、絞り込み検索を支援する手法が考えられている。

[7]では、不特定多数のユーザの検索履歴に基づき、検索結果の関連語を抽出し、検索エンジン側が検索の際に検索結果とともに検索語の関連語を絞り込み候補語として提示することにより、ユーザの絞り込み検索の支援を行っている。

2.2.3 ユーザの差別化

検索エンジンのパーソナル化を行うためには、ユーザの詳細な行動をモニタリングするために[6]の提案のようにブラウザとの連携が重要である。しかし、ユーザの検索語パターンと再検索までの

時間間隔等の検索エンジン側で得られる情報のみを用いてユーザの意図が判断できれば、ブラウザとの連携は不要となる。そこで、本論文では検索エンジン側で得られる情報のみを用いて、ユーザの検索のモデル化を行う。

3 検索履歴の解析手法

本節では、同一ユーザによる検索の傾向を調べ、ユーザの検索のモデル化を行う。まず、時系列で並んでいる検索語のデータから同一ユーザによる検索語を抽出する。その後、同一ユーザによる検索語入力をパターン化する。

3.1 解析に用いた検索履歴

解析にあたっては Microsoft Accessibility[8]および Microsoft Product Support Service のサポート技術情報検索[9]に入力された 2001 年 4 月 20 日から 2001 年 11 月 5 日までの検索履歴を使用した。使用する検索履歴は個人情報を含んでおらず、検索語と検索のタイムスタンプのみである。

3.2 解析手順

データの解析の手順は次の通りである。

(1) 同一ユーザによる検索の同定

入力された各々の検索語は独立の検索としてログ上に記録されているため、まず、同一ユーザによる一連の検索を同定する必要がある。本稿では同一ユーザを同定するために 10 分以内に同じ検索語で検索している場合を同一人物とみなした[1]。ただし、複数の検索語で検索している場合は１番目に入力されている検索語だけを比較し、同一であった場合、同一ユーザによる検索とみなした。

(2) 検索時間間隔抽出

同一ユーザの検索語入力の時間間隔を集計する。

(3) 検索のパターン分類

検索語の増減、変化によってパターンに分類する。

(4) 検索パターン毎の時間間隔抽出

パターンごとの割合およびパターンごとの時間間隔を調べる。

4 解析結果

本節ではデータを解析した結果を示す。

4.1 検索の時間間隔

第３節で示した手法により抽出した同一ユーザの検索の時間間隔について調べた。検索エンジンを利用して自分の欲しい情報を探す際のユーザの行動をモデル化すると図１のように表すことができる。図中の各ＳＴＥＰの意味は下記の通りである。

ＳＴＥＰ１ 欲しい情報を得るための検索式を検討し入力する。

ＳＴＥＰ２ 検索エンジンから得た検索結果を見て、自分の要求に合致するページがあるかどうか評価し、該当するＵＲＬを選択する。

ＳＴＥＰ３ ＳＴＥＰ２で選択したページを閲覧する。

通常、求める情報が１回の検索で得られることは少ない。図１の(1),(2)のようにＳＴＥＰ２で検索結果が不適であったり多すぎたりするときには、ＳＴＥＰ１に戻って、異なる検索語を入力したり、検索語を変えたりなどの試行錯誤による、連続した検索を行う。また、図１の(2)のようにＳＴＥＰ３における閲覧の結果、目的の情報でなかったと判断したときには、ＳＴＥＰ２に戻って別の検索結果を選択したり、ＳＴＥＰ１に戻って再検索を行う。このようにユーザは上記のステップを繰り

[1] １日分の検索履歴を調査した結果、１０分以内に同一語を用いた検索が集中していたため、一つの指標として 10 分を閾値とした。

返しながら、求める情報を得るための検索を行う。

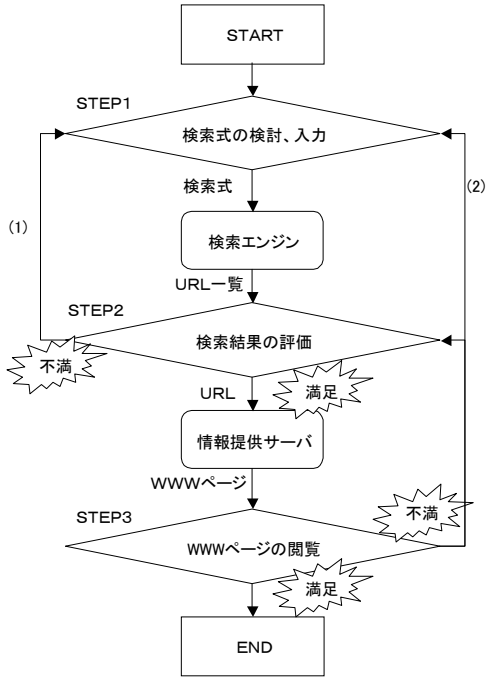

図 1 検索エンジンのユーザの行動

次に、同一ユーザの連続する検索の時間間隔を表 1 および図 2 で示す。

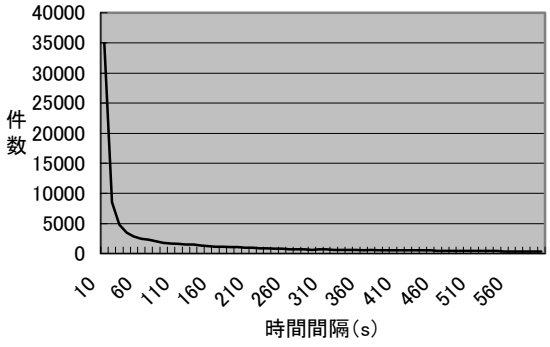

図 2 同一ユーザによる検索の時間間隔

表 1 および図 2 から、多くの検索間隔が非常に短く（1 分未満で 57%,2 分未満で 69%）なっている。これは図 1 の(2)のようにＳＴＥＰ３からの戻りが少なく、図 1 の(1)のようにＳＴＥＰ２からの戻りが多いことを示している。多くの検索エンジンではＳＴＥＰ２の検索結果として、ＵＲＬ以外にタイトルやアブストラクトを表示するため、検索結果で表示されたページの内容をある程度推測できる。そのため、ＳＴＥＰ３からの戻りは比較的少なく、ＳＴＥＰ１とＳＴＥＰ２の間で比較的短い時間間隔での頻繁な繰り返しが多くなっていると推測される。

検索間隔をパラメータにして、下記のような 2 つの検索モデルを定義する。

(1) ページを閲覧せずに次の検索を行う場合（時間間隔：短）

再検索の理由

検索結果が多すぎたため

検索結果が少なすぎたため(1 件も表示されなかった場合も含む)

スペルミスで検索し、すぐ気づいて直した場合

表示された 10 件のなかに欲しいページがなさそうだった場合

(2) 実際にページを閲覧した場合（時間間隔：長）

再検索の理由

表示されたタイトルおよびアブストラクトを見て、欲しいページと判断し実際にページを閲覧した場合

表 1　同一ユーザによる検索の時間間隔

時間間隔(s)	件数	割合	時間間隔(s)	件数	割合	時間間隔(s)	件数	割合
0-10	34940	35.0%	200-210	992	1.0%	400-410	497	0.5%
10-20	8489	8.5%	210-220	851	0.9%	410-420	558	0.6%
20-30	4804	4.8%	220-230	878	0.9%	420-430	526	0.5%
30-40	3481	3.5%	230-240	787	0.8%	430-440	490	0.5%
40-50	2823	2.8%	240-250	778	0.8%	440-450	508	0.5%
50-60	2470	2.5%	250-260	737	0.7%	450-460	434	0.4%
60-70	2274	2.3%	260-270	744	0.7%	460-470	472	0.5%
70-80	2000	2.0%	270-280	669	0.7%	470-480	434	0.4%
80-90	1793	1.8%	280-290	664	0.7%	480-490	424	0.4%
90-100	1729	1.7%	290-300	668	0.7%	490-500	441	0.4%
100-110	1599	1.6%	300-310	679	0.7%	500-510	409	0.4%
110-120	1529	1.5%	310-320	618	0.6%	510-520	440	0.4%
120-130	1472	1.5%	320-330	645	0.6%	520-530	434	0.4%
130-140	1365	1.4%	330-340	582	0.6%	530-540	420	0.4%
140-150	1250	1.3%	340-350	577	0.6%	540-550	357	0.4%
150-160	1144	1.1%	350-360	566	0.6%	550-560	397	0.4%
160-170	1117	1.1%	360-370	578	0.6%	560-570	348	0.3%
170-180	1091	1.1%	370-380	561	0.6%	570-580	379	0.4%
180-190	1049	1.1%	380-390	538	0.5%	580-590	368	0.4%
190-200	948	1.0%	390-400	514	0.5%	590-600	363	0.4%
左側の数値は以上、右側の数値は未満を表す						合計	99692	100.0%

実際に、検索に関して素人の 20 代の男女 4 人のユーザに(2)のように実際にページを閲覧してもらい、検索エンジンに戻って再検索を行ってもらった。その結果時間間隔は 56,68,74,96 秒であった。この結果に基づき、実際にページを閲覧した場合の最短の時間間隔は 60 秒程度と推測され、今回、(1)と(2)の閾値を 60 秒と設定した。

4.2 検索語のパターン

複数回（2 回以上）検索しているユーザの検索の回数を調べると表 2 のようになる。

表 2 同一ユーザによる連続検索回数

連続検索回数	人数	割合
2	8698	78.52%
3	1770	15.98%
4	399	3.60%
5	129	1.16%
6	49	0.44%
7	21	0.19%
8	3	0.03%
9	4	0.04%
10回以上	4	0.04%
合計	11077	100.00%

表 2 より 4 回以上の検索を行っているユーザは全体の５%ほどであることがわかった。そのため以下では連続する 3 回までの検索について検索パターンを分類する。

各パターンは表３の通りである。また、それぞれの検索パターンにおける再検索の推測理由は表４の通りである。

表 3　検索パターン

No.	パターン		例
1	A−AB	最初の検索語から検索語の数を増やす場合。	1回目：“excel”
			2回目：“excel”，“ ダウンロード”
2	AB−AC	先頭の検索語は変えずに後ろの検索語を他の検索語に変える場合。	1回目：“IPアドレス”，“ プリンタ”
			2回目：“IPアドレス”，“ プリントサーバー”
3	AB−A	1回目に複数個の検索語で検索を行ったが、2回目で検索語を減らす場合	1回目：“windows2000”，“ファイルキャッシュ”
			2回目：“windows2000”
4	A−AB−AC	2回目の検索で検索語を増やし、3回目の検索語で後ろの検索語を変えている場合。	1回目：“フォント”
			2回目：“フォント”，“追加”
			3回目：“フォント”，“ダウンロード”
5	A−AB−A	2回目の検索で検索語の数を増やし、3回目の検索で再び検索語を減らし元に戻す場合。	1回目：“キーボード”
			2回目：“キーボード”，“OSK”
			3回目：“キーボード”
6	A−ABC−AB	2回目の検索で2個増やしたが、3個目で再び1個減らす場合。	1回目：“ユーザインターフェイス”
			2回目：“ユーザインターフェイス”，“ガイドライン”，“Windows”
			3回目：“ユーザインターフェイス”，“Windows”
7	AB−A−AC	1回目の検索で複数の検索語を入れて検索し、2回目で検索語を増やし、3回目で再び検索語を増やす場合。	1回目：“ProxyServer”，“フィルタ”
			2回目：“ProxyServer”
			3回目：“ProxyServer”，“設定”
8	AB−A−AB	検索語を減らしたあと、再び同じ検索語を増やす場合	1回目：“ユーザー”，“切り替え”
			2回目：“ユーザー”
			3回目：“ユーザー”，“切り替え”
9	AB−AC−AB	検索語を変えた後、再び同じ検索語に戻す場合	1回目：“ユーザー”，“切り替え”
			2回目：“ユーザー”，“切替”
			3回目：“ユーザー”，“切り替え”

A,B,Cはそれぞれ異なる検索語を示す

表 4 各パターンと再検索の推測理由

パターン		検索結果が多すぎた場合	検索結果が少なすぎた場合（0の場合も含む）	検索結果を見たが、欲しい情報がなかった場合	実際にページを閲覧した結果、欲しい情報がなかった場合	1回目の検索結果の方がふさわしいと判断した場合	表記のゆれを訂正
1	A−AB	○		○	○		
2	AB−AC	○	○	○	○		○
3	AB−A		○	○	○		
4−a	A−AB	○		○	○		
4−b	AB−AC	○	○	○	○		○
5−a	A−AB	○		○	○		
5−b	AB−A					○	
6−a	A−ABC	○		○	○		
6−b	ABC−AB		○				
7−a	AB−A		○	○	○		
7−b	A−AC	○		○	○		
8−a	AB−A		○	○	○		
8−b	A−AB					○	
9−a	AB−AC	○	○	○	○		○
9−b	AC−AB					○	

4−aとはパターン4の1回目、4−bとはパターン4の2回目をあらわす

5 考察

第 4 節の解析結果についての考察を行う。

5.1 各パターンの割合

第 4 節で分類したパターンごとの全体に占める割合を表 5 で示す。

表 5 パターンごとの割合

No.	パターン	件数	割合
1	A−AB	3939	35.6%
2	AB−AC	2030	18.3%
3	AB−A	2729	24.6%
4	A−AB−AC	335	3.0%
5	A−AB−A	95	0.9%
6	A−ABC−AB	521	4.7%
7	AB−A−AC	151	1.4%
8	AB−A−AB	15	0.1%
9	AB−AC−AB	266	2.4%
その他		996	9.0%
計		11077	100.0%

表 5 より、2 回の検索（パターン 1～3）が約 80%を占め、3 回までの検索でパターン分けした場合、全体の 91%を第 4 節で挙げた 9 つのパターンに分類することができた。第 4 節で検索回数が 3 回以内の割合が 95%であったのに対し、表 5 でパターンに分類できた割合が 91%になったのは、その他に分類されている中に、4 回以上の検索とともに、パターン 1～9 に当てはまらなかった 3 回以内の検索も含まれているからである。

5.2 各パターンごとの時間間隔の比較

第 4 節で、実際のページを閲覧した場合の閾値を 60 秒と設定した。この閾値を用いて、各パターンごとに実際のページを閲覧した場合と閲覧しなかった場合を分類し、表 6 に示す。

表 6 パターンごとの時間間隔

パターン		60秒未満	60秒以上	平均(s)	標準偏差
1	A−AB	27.3%	72.7%	198.82	185.99
2	AB−AC	38.4%	61.6%	183.65	182.14
3	AB−A	46.0%	54.0%	158.02	175.94
4−a	A−AB	23.9%	76.1%	205.16	183.02
4−b	AB−AC	52.2%	47.8%	124.15	151.69
5−a	A−AB	21.1%	78.9%	196.08	177.92
5−b	AB−A	48.4%	51.6%	149.36	162.05
6−a	A−ABC	27.4%	72.6%	212.90	197.71
6−b	ABC−AB	58.9%	41.1%	119.45	154.16
7−a	AB−A	35.1%	64.9%	191.51	193.48
7−b	A−AC	31.1%	68.9%	184.01	175.35
8−a	AB−A	26.7%	73.3%	260.47	260.02
8−b	A−AB	13.3%	86.7%	296.20	182.91
9−a	AB−AC	41.7%	58.3%	157.69	167.94
9−b	AC−AB	48.1%	51.9%	151.55	169.89

4−aとはパターン4の1回目、4−bとはパターン4の2回目をあらわす

表 6 より実際にページを閲覧した割合はパターンごとに違いがあることがわかる。2 回の検索（パターン 1～3）を比較してみると、パターン 1 に比べ、パターン 2,3 は時間間隔が短い場合が多い。これは、検索語を増やす場合は新しい検索語を考える必要があり、再検索までの時間間隔が長くなっていると推測される。

6 問題点

本節では、今回のモデル化に関しての問題点について述べる。

本稿の解析の結果、検索語のパターンを 9 パターン、再検索までの時間間隔を 2 パターンに分類し、検索の意図を表 4 の通りに推測することが出来た。この検索語のパターンごとに検索エンジン側で検索結果のランキングや検索結果の表示件数を変えることでユーザの意図にあう情報を検索結果として出力することが出来ると考えられる。

例えば、再検索の際に検索語を減らす場合、2 回目の検索結果の中には 1 回目の検索結果が必ず含まれる。しかし 1 回目の検索結果にユーザの求めるページがなかった場合、2 回目の検索結果に 1 回目の検索結果を表示しない。このようにユーザの意図にあわないページを排除できる。

しかし、時間間隔について 2 種類に分類し、そ

の閾値を全パターン一律に設定したが、5.2 で述べたように、新たな検索語を考える場合と考えない場合など、実際のページを閲覧しない場合であっても、パターンごとに時間間隔は異なる。そのため、パターンごとに異なる閾値を設定する必要がある。

7　おわりに

Google など既存の検索エンジンでは、求める情報が異なっても同一の検索語で検索した場合、同一の検索結果が返されてしまう。このような問題を解決する手法として、本論文では検索履歴からユーザの意図を推測する手法を考案し、そのために必要なユーザの検索のモデル化を行なった。その結果、再検索の時間間隔を 2 つのパターンに、複数回検索しているユーザの検索パターンの 91%を 9 つのパターンに分類することに成功した。この検索語のパターンごとに検索エンジン側で検索結果のランキングや検索結果の表示件数を変えることで、ユーザの意図に合致する情報を検索結果として返すことが出来るようになると考えられる。

謝辞

本研究を行うにあたり、検索履歴のデータを提供して下さったマイクロソフトアジアリミテッドの木戸冬子さんに深く感謝致します。そして、私を助けてくれた山名研究室の同輩の方々に厚く御礼申し上げます。

参考文献

[1] Yahoo! : http://www.yahoo.co.jp/

[2] goo : http://www.goo.ne.jp/

[3] Google : http://www.google.co.jp/

[4] Lawrence Page, Sergey Brin, Rajeev Motwani and Terry Winograd : ”The PageRank Citation Ranking: Bringing Order to the Web”, 1998 http://www.ab.stanford.edu/~backrub/pageranksub.ps

[5] Excite : http://www.excite.co.jp/

[6] 石川雅弘, 後藤文太郎 : “WWW アクセス活動と Web コンテンツの情報統合いおける履歴抽出精度の向上とその応用”, 第 62 回情報処理学会全国大会 予稿集, No.3, pp.173-174, 2001

[7] 原田昌紀, 清水奨 : ”WWW 検索システムにおける不特定多数の操作履歴の活用”, 情報処理学会研究報告 97-DPS No.81,pp61-66,1997

[8] マイクロソフト　アクセシビリティサイト : http://www.microsoft.com/japan/enable/

[9] Microsoft Product Support Service : http://support.microsoft.com/