2G4-1

ユーザインタラクションに基づく情報検索
Information Retrieval based on the User Interaction

下山 洋一[*1]
Yoichi Shimoyama

高木 友博[*1]
Tomohiro Takagi

[*1] 明治大学 理工学研究科 基礎理工学専攻 情報科学系
Computer Science Course, Graduate School of Science and Technology, Meiji University

Abstract: We propose an interactive information retrieval system using RSV. In general, ordinary information retrieval systems use tf-idf. However, the systems usually cannot understand users' true requests. We propose a system using RSV to overcome the problem, and compare it with the ordinary system.

1. はじめに

我々は情報検索において，ユーザとシステムとの間でインタラクションを行うことにより，ユーザの情報検索の意図をくみ取り，検索精度を向上させることを目的としている．

実際に検索をするにあたって，いくつか問題が存在する．たとえば，ユーザが必要としている情報について，ユーザ自身の知識が乏しい場合がある．また，十分に知識があっても，検索システムに対してどのように質問を投げかければよいか分からない場合がある．このような時は大抵，ユーザの潜在的な要求が検索システムに伝わらず，結果としてユーザは不満足な検索結果を得る．

このようにユーザの意図が検索システムに十分に伝わらない問題に対処する手法のひとつとして，ユーザフィードバックが挙げられる．この手法は，検索システムがユーザの検索質問を受けたときに，その意図を明確にするためにユーザとのやりとりを行い，システムがその意図の絞り込みや修正を行うものである．本研究では，このユーザフィードバックを行うにあたって，一般的に用いられる TF-IDF の替わりに Robertson Selection Value (RSV) を用いることで，的確な情報検索を行うことを試みる．

2. ユーザの意図に基づいた情報検索

多くの場合，検索要求と検索質問とは異なる．検索要求はユーザが欲しいと思った要求を表し，検索質問は実際にユーザがシステムに入力として渡したキーワードのことを意味する．したがって，ユーザがどれだけ的確に自分の要求をシステムに伝えられているかが，これらの違いに影響することとなる．検索の熟練者ほど，検索対象での単語の使われ方を想定し，自分の要求を漏れなく，かつゴミが混ざらないよう検索質問を構成させる傾向がある．これに対して検索初心者は，検索質問を構成させるだけの語彙を想起することができず，検索者の満足とは程遠い検索結果を手にすることが多い．

この問題を解決しようとするのがユーザフィードバックとなる．ユーザフィードバックとは，ユーザとシステム間でやり取りを何度が行うことで，検索要求をシステム側に理解させる手法である．従来の 1 回のみキーワードを入力する検索に比べ，検索質問の背後にある不確定要素を減らすことが可能になるため，ユーザの意図をより検索システムに伝えることが可能になる．

ユーザフィードバックの具体例として，インドネシアに旅行を予定しているユーザが観光目的でジャワ島について調べる場合について考える．ユーザはジャワ島を意味する”java”と入力し検索を行うとする．このとき，システム側は”java”という単語の意味を表しているドキュメントを返すのが最も理想的となる．ところが，このシステムの結果にはプログラミング言語としての”java”や，コーヒー豆を表す”java”などユーザが意図していない意味の”java”が含まれ，ユーザはこの中から自分が必要とするドキュメントを探し出さなければならない．

次に，今度はユーザが”java island”の 2 単語を入力し検索するものとする．先ほどとは違い，”island”という言葉を加えたことでプログラミング言語としての”java”や，コーヒー豆を表す”java”は検索されなくなるだろう．しかしながら，この場合もあらゆるジャワ島のドキュメントが検索され，観光という目的にあったジャワ島の情報は自分で探し出さなければならない．

このように実際の検索ではユーザが自分の検索要求を的確に表現することは非常に困難である．この問題を解決する 1 つの手法が，ユーザフィードバックである．ユーザフィードバックとは，ユーザが検索質問としてキーワードを入力した後，そのキーワードの背後に潜む不確定要素をシステム側はユーザに提示し，それをそのユーザに評価してもらうことにより，より検索要求を満たした検索結果を得ることを目的としている．ジャワ島の例では，ユーザが欲しがっている情報がジャワ島なのか，プログラミング言語の Java なのか，コーヒー豆の Java なのか，どういう目的で情報を探しているのかなどをシステムがユーザに問いかけることにより，ユーザ自身が検索結果から必要な情報を探し出す手間を省くことができる．

3. システム概要

構築したシステムの概要を，以下の図1に示す．

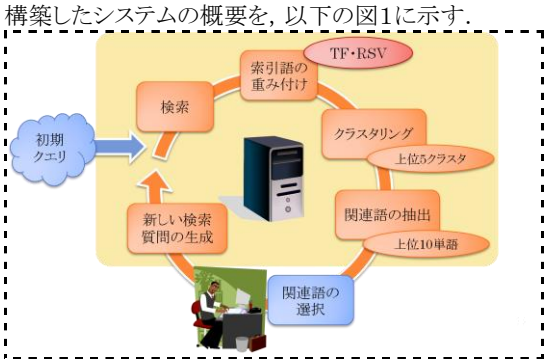

図 1　システム概要図

ユーザは，システムが提示した関連語のうち，本当に欲しい情報と関連する語を選択し，システムに返答する．このようにイ

連絡先：下山 洋一，明治大学大学院 理工学研究科 基礎理工学専攻 情報科学系，川崎市多摩区東三田 1-1-1，044-934-7483，shimoyama@cs.meiji.ac.jp

ンタラクションを繰り返すことにより，ユーザの要求を満たすことができると考えられる．

3.1 索引語の重み付け

この処理では，検索結果として返された上位 100 文書の索引語に対して重み付けを行う．従来手法と提案手法とで，異なる索引語の重みを与える．従来手法では，索引語の重みとして TF-IDF を用いる．RSV を用いる提案手法では，TF-IDF のかわりに TF×RSV を用いる．

RSV(Robertson Selection Value) は，Stephen Robertson らにより研究されている，正解文書を検索するのに有効な言葉を選択する指標のうちのひとつである．

n　：単語 t を含むコーパス内での文書数

r　：単語 t を含む正解文書内での文書数

$p = \frac{r}{R}$　：正解文書が単語 t を含む確率

$\bar{p} = \frac{n-r}{N-R}$　：不正解文書が単語 t を含む確率

$$RSV(t) = (p-\bar{p})\log\left(\frac{p(1-\bar{p})}{(1-p)\bar{p}}\right)$$

3.2 クラスタリング

重み付けされた上位 100 文書を K-means method Clustering (クラスタ数:20)でクラスタリングする．クラスタリングを行うことにより，文書集合を分類することができ，抽出する関連語に幅を与えることが出来る．そのことによりユーザに選択の幅を与えられるので，正確なフィードバックを得られると考えられる．

3.3 関連語の抽出

前節で生成されたクラスタのうち，クラスタに含まれる文書のクエリに対するコサイン値の平均値の上位 5 クラスタを抜き出す．そして抜き出されたクラスタの重心ベクトルから，値の高い順に各クラスタから 10 単語ずつ抽出する．ただし，クエリと同じ単語，および既に抽出された単語は除き，別の単語を提示する．もし既に以前に提示されているなどしてクラスタから単語が抽出できない場合は提示しない．

3.4 索引語選択

抽出された索引語をユーザに提示する．ユーザは提示された索引語の中から検索に有用だと思うものを選択し，システムは選択された索引語をユーザフィードバックとして得て，それによって検索質問拡張を行う．

3.5 新しい検索質問の生成

選択された索引語を元の検索質問に加えることにより，新しい検索質問の生成を行う．新しい検索質問は次のように生成する．ここでは，予備実験より α =0.8 とした．

$$q_{New} = \alpha \cdot q_{Original} + (1-\alpha) q_{Expandec}$$

4. 実験

実験では，提案手法である RSV を用いたシステム，および従来手法である TF-IDF を用いたシステムをユーザに使ってもらい，繰り返し検索をそれぞれの手法について 5 回ずつ行ってもらった．毎回の検索質問およびその検索結果をログとして保存して，後でその分析を行った．また，システムに対するユーザの主観的評価を得るためアンケートを行った．アンケートでは，検索の難易度を 5 段階で評価してもらった．

なお，実験において，コーパスは AQUAINT を利用した．文書数は約 103 万文書である．また，トピックは TREC HARD Track 2005 で用いられた 50 トピックのなかから，無作為に 10 トピックを選出して用いた．

4.1 実験結果

(1) アンケート

ユーザのアンケートより，従来手法と提案手法とで，どちらがより容易に感じたかを以下にまとめた．

topic	RSV	TF-IDF
307	1	7
336	4	2
341	3	0
344	1	2
353	5	0
354	4	1
389	4	1
399	4	1
408	5	2
638	4	2

表 1　より容易に感じた人数

10 トピック中，Topic 344 を除く 9 トピックにおいて，有意な差が認められ，8 トピックにおいて提案手法である RSV を用いたシステムが簡単だと感じたユーザが多かった．

(2) 検索質問の長さ

従来手法と提案手法とのインタラクションを 5 回終えた時のクエリ長の差の平均を以下に示す．

topic	差
307	-2.8
336	5.2
341	7.4
344	7.3
353	7.3

topic	差
354	11.7
389	6.6
399	6.3
408	2.3
638	9.3

表 2　クエリ長の差の平均（提案手法－従来手法）

Topic 408 を除く 9 トピックにおいて有意な差が認められ，8 トピックにおいて提案手法の方がクエリ長が長いことが分かった．また，前節でのアンケート結果と比較してみると，ほとんどのトピックにおいて，ユーザが検索が容易に感じた手法の方が，クエリ長が長いと言うことが分かった．これらにより RSV を用いることで，クエリに関連のある単語をユーザに提示できると言える．

(3) 検索精度

ここでは Average Precision，R-Precision，P@10 の3つの尺度で検索精度を分析する．

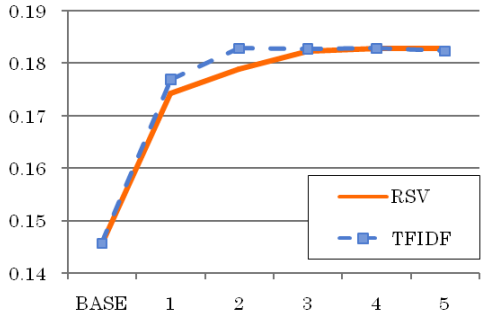

図 2 Average Precision

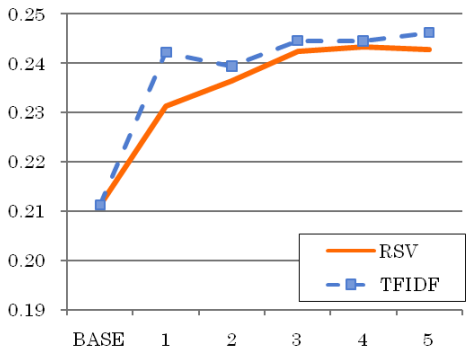

図 3 R-Precision

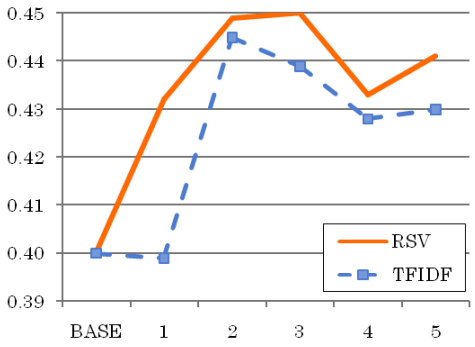

図 4 P@10

Average Precision はわずかに TF-IDF を用いた手法が上だが，あまり有意な差は認められなかった．R-Precision もわずかに TF-IDF を用いた手法が上で，フィードバック 1 回目については有意な差があるが，それ以降についてはあまり有意な差は認められなかった．P@10 は RSV を用いた手法の精度が良く，特にフィードバック 1 回目については有意な差が認められた．しかし，これら 10 トピック平均のデータからだけでは，特にどちらの検索精度が良いといった有意な差があるとは言えない．

4.2 考察

アンケート結果，およびクエリ長については有意な差は認められたが，客観的な検索精度については有意な差が認められなかった．この節では，実際に検索質問に使われた索引語を分析し，その原因を探る．

(1) Topic 344: Abuses of E-mail の場合

まず，ユーザによる関連語の選択の差の影響を減らすため，初回のフィードバックにおいて過半数（6 人以上）が選択した単語のみを抜き出し，初期クエリと，過半数が選んだクエリとの検索精度を比較する．

	Query	AP	R-Precision
BASE	**abuse email**	0.0624	0.1545
RSV	**abuse email** spam internet provider *mail address*	0.1143	0.1951
TF-IDF	**abuse email** spam internet provider	0.1261	0.2114

表 3

検索語に”mail”と”address”の多い RSV を用いた手法の方が，TF-IDF を用いた手法に比べ，Average Precision，R-Precision ともに精度が低くなっている．さらに詳しく，”mail”と”address”を単独で加えた場合について検証してみる．

Query	AP	R-Precision
abuse email spam internet provider *address*	0.1167	0.1870
abuse email spam internet provider *mail*	0.1275	0.2114

表 4　クエリの差による検索精度の違い

RSV を用いた手法では，主に検索語に関連する単語が多く，提示される索引語の意味的な幅が比較的狭い場合がある．そのため，初期クエリに現れない意味を引き出すためには不利な場合があると考えられる．

(2) Topic 408: Tropical Storms の場合

(1) と同様の比較をしてみると以下のようになる．

	Query	AP	R-Precision
BASE	**tropical storm**	0.1012	0.1639
TF-IDF	**tropical storm** hurricane typhoon meteorological *flooding damage death*	0.1310	0.1803
RSV	**tropical storm** hurricane typhoon meteorological *flood weather*	0.1123	0.1639

表 5　初期クエリと初回フィードバック後の精度

表より，二つの手法での単語の差は「”flooding”，“damage”，“death”」と「”flood”，“weather”」である．ここで Topic 408 の内容に注目する．

Num	408
Query	tropical storms (熱帯低気圧)
Request	どのハリケーンやタイフーンなどの熱帯低気圧が重大な財産被害や犠牲者を出したか．その日付，被害のあった範囲，被害および犠牲者の程度について．但し，小規模の熱帯低気圧に関しては除く．

表 6　Topic 408 の内容

Topic 408 の内容から重要なのは「被害」や「犠牲者」についての情報であることが分かるが，RSV を用いた提案手法では検索語が「気象」に関するものばかりであり，比べて TF-IDF を用いた手法では索引語に”damage”や”death”といった被害に関する単語が抽出できていることが分かる．このことにより検索精度に差が生まれて来ているということが分かる．

RSV を用いた手法では，主に検索語に関連する単語が多く，提示される索引語の意味的な幅が比較的狭い傾向がある．そのため，初期クエリに現れない意味を引き出すためには不利な場合があると考えられる．

(3) Topic 354: Journalist Risks の場合

ここでは，検索語に関連する単語が多いことの弊害を，別の角度から観察する．

Query	AP	R-Precision
journalist risk	0.0855	0.1915
journalist risk reporter hostage journalism	0.1504	0.2606
journalist risk reporter hostage journalism *correspondent reporting dangerous journalists*	0.1496	0.2660

表 7　クエリ追加時との精度比較

表の 3 つ目のクエリは，2 つ目のクエリより検索語が大幅に拡張されているにもかかわらず，ほとんど精度に変化が見られない．RSV を用いて抽出され，なおかつ既にクエリに存在する検索語と似た意味を持つ索引語は，検索結果に大幅な変化を与

えないことが言える. なぜならば RSV が高い索引語は, すでに上位文書に含まれている確率が高く, そしてそれ以外の文書に含まれていないことが多いためである.

以上の(1), (2), (3)をまとめると次のようなことが言える. RSV はクエリに近い索引語を集めてしまう. その結果として, 元々クエリに存在する検索語と似た意味をもつ索引語が多く表示され, 検索結果がほとんど変化しないことになる. また, 逆に TF-IDF を用いた手法においては, 提示される単語の意味的な幅が広すぎてしまい, ユーザはその選択を難しいと感じることが多いが, 選択された索引語は IDF が高いので, 短いクエリ長でも巧く文書集合を絞り込むことが出来る.

5. まとめ

本実験によって, 繰り返し検索における RSV を用いたユーザフィードバックの効果について検証した. その結果, 客観的な検索結果においては, 従来手法の TF-IDF と比較して明確な有用性を示すには至らなかったが, ユーザに対しても比較的選びやすい索引語を抽出できることが判明した. また, RSV と IDF を比較することにより, それぞれの利点・欠点を具体的に示すことができた.

今後の課題としては, 実際に客観的な検索精度の向上が求められる. RSV のみを用いてクエリに関連する索引語ばかりを抽出するのではなく, 実際に検索に貢献できるよう, IDF などの文書の絞り込みの力が強い要素を用いることが必要であると考えられる.

参考文献

[北 02] 北研二, 津田和彦, 獅々堀正幹: 情報検索アルゴリズム, 共立出版, 2002.

[徳永 99] 徳永健伸: 言語と計算 5 情報検索と言語処理, 東京大学出版会, 1999.