検索結果評価による問い合わせのコンテキスト抽出機能を有するWWW検索システム

A WWW Search System with Functions for Query-context Extraction Reflecting User's Evaluation

船木 信宏♥　　清木 康♦

Nobuhiro FUNAKI　Yasushi KIYOKI

WWWにおける検索エンジンの精度は上がっているが，検索結果として表示されるページ数は依然として膨大である．人間の短期記憶で判断できる項目数は限定的（ある研究では7±2と言われている）であり，検索エンジンが与えられた検索語に対して表示する検索結果のページ数が平均10以上であることは，ユーザビリティを損ねていると言える．

本稿では，ユーザビリティを向上させるために，検索エンジンを利用する際の“ユーザのふるまい”（直前のアクセス内容）から，与えられた検索語のコンテキストをダイナミックに学習し，抽出されたコンテキストをもとにして各検索結果ページとの相関量を計算，相関量に応じて検索結果を絞り込んで表示するシステムを実装し，検証する．

Performance and quality of current WWW search engines have been improved. However, those search engines provide many search results including a large amount of pages. The number of items that can be discerned from person's short-term memory is limited. Some research results show that only around 7 items can be memorized in short-term. This fact detracts from the usability of search results. In this paper, we present a web search system which returns search results reflecting user's query-contexts. The system realizes a new WWW search environment for giving an appropriate ranking for web pages, according to a context obtained by user's behavior. We show the system implementation and experimental results to clarify the feasibility of our system.

1. はじめに

WWWにおいて検索エンジンの利用は拡大している．検索エンジンの精度についてはGoogleのPageRank[1]に代表されるランキング方式等により，検索語に対する適合率が向上しているが[2]，既存の検索エンジンが検索語に対する結果として表示する際のページ数は総じて10程度であり，ユーザはその中から自身のコンテキストに合ったウェブページ（以下ページと表記）を探して選択しなければならないという問題がある．The Magical Number Seven[3] で知られるように，人間の短期記憶で判断できる項目の数は7±2と言われている．ウェブにおいては，ページ内のメニューの項目数を7前後にするとユーザは直感的に選択できて良いとされている．これに則って判断すれば，検索エンジンが結果として返すページ数は人間がどのページにアクセスするか判断するには，多いと言える．一方で，検索結果の2ページ目以降にアクセスして見るユーザの割合は5%という調査結果があり[4]，最初に表示される10件の中からコンテキストに合ったページを発見できるかどうかが重要である．また，既存の検索エンジンは，入力された検索語に対してパターンマッチングとランキングで結果を返す．そのためユーザのコンテキストは反映されていない．しかし，一般に検索語に利用する単語の数は2～3程度と言われており[5]コンテキストを数語から判別するのは困難である．

既存のランキング方式の検索エンジンに代わる手法として以下の先行研究が挙げられる．

- クラスタリング
  検索語に対して返された検索結果を関連するページごとに分類する手法を提案している（[6][7][8]）．
- 適合フィードバック
  検索結果に対して，ユーザのフィードバックを得ることで検索の精度を高める手法を提案している（[9][10][11]）．
- メタ検索
  検索語によるパターンマッチングではなく，検索語と検索対象それぞれからメタデータを抽出し，意味的に近いページを検索する手法を提案している（[12]）．
- 可視化
  検索語や検索対象を可視化する．視覚的な情報を検索結果で用いることでユーザビリティを高める手法を提案している（[13][14]）．

これら手法を用いることにより検索エンジンの精度は向上しているが，多くの場合，ユーザに検索語を入力する以上の行動を必要としたり，インターフェイスが既存の検索エンジンと著しく異なり，ユーザビリティが高いとは言えない．本稿では，検索エンジンを利用する際の“ユーザのふるまい”（直前のアクセス内容）をもとに，与えられた検索語のコンテキストをダイナミックに学習し，検索結果を絞り込んで表示する方式を提案する．ここで検索語のコンテキストとは，検索語に多様な意味がある場合に，ユーザが意図する意味を確定する情報（文脈）のことを指す．ユーザの自然なふるまいをもとにコンテキストを確定することにより，ユーザビリティを損ねることなく検索エンジンの精度を向上させる．

2. 提案方式

2.1 概要

本方式は，膨大な検索結果からの情報獲得を容易にするために，ユーザの検索語におけるコンテキストを反映し，検索結果を表示することを実現する．本方式の特徴は，ユーザの発行した検索語から検索エンジンを介して出力される検索結果からページにアクセスし，ユーザの本来意図している内容を含むページをユーザが指定することにより，そのページを検索語のコンテキストとして確定し，そのコンテキストを表すベクトルをコンテキストベクトルとして生成することによって，相関量計算によりユーザの意図に合致する検索結果を強調表示し，ユーザの要求に近いページ群として表示することを可能にするものである．

ユーザにとって検索語のコンテキストに当てはまらない

♥ 学生会員　慶應義塾大学環境情報学部
funaki@mdbl.sfc.keio.ac.jp
♦ 正会員　慶應義塾大学環境情報学部
kiyoki@sfc.keio.ac.jp

ページは表示される必要がない．例えば，「ドライブ」という検索語に対してユーザのコンテキストが「(車を運転する)ドライブでいい行き先を知りたい」だった場合に「CD-ROMドライブ」や「ハードディスクドライブ」に関するページは不要である．これは検索語が単に多義語であるだけでなく，コンテキストによってユーザが求めるページの内容が異なりそれによって表示すべきページが変わることを示している．こういったユーザ固有の動的なコンテキストを与えられた検索語だけで判別することは不可能である．ユーザが検索語を複数入力することでコンテキストを判別しやすくなるが，既存のパターンマッチングの検索エンジンでは「ドライブ+行き先」といった検索語が入力された場合，「行き先」という検索語がページ内に含まれていなければ検索結果に出力されない．検索結果の適合率を下げる可能性もある．ユーザにとって自身のコンテキストを示す検索エンジンにとって最適な検索語を入力することは非常に困難である．

そこで本方式では，検索語に加えて“ユーザのふるまい”を利用してコンテキストを判別する．ここでユーザのふるまいとは，ユーザが検索結果から選択したページにアクセスすることを指す．ユーザがページにアクセスし，そのページの内容に満足した場合は検索が終了する．満足しなかった場合は，検索結果に戻る．戻る際に本方式ではユーザに，アクセスしたページの内容がコンテキストに合致していたか否かの評価を与えさせる．システムは，ユーザのふるまいから，アクセスしたページの内容と内容に対するユーザの評価を取得する．このふるまいは，検索エンジンを利用する際にユーザにとって自然な動作であり負担をかけることがない．また，検索語を増やす必要が生じたり従来の検索エンジンとインターフェイスが大幅に変わることもないので，ユーザビリティを損ねない．ユーザのふるまいごとにダイナミックにコンテキストを判別し，ふるまいが複数回行われるごとに学習し検索結果に反映する．検索結果には，コンテキストに合致したページを強調して表示，ないしは合致しないページを非強調表示することで，ユーザの求める情報へのアクセスを容易にする．

2.2 データ構造

本方式で扱うデータ構造は以下のとおりである．

- 検索語
  ユーザから入力される文字列.本方式では日本語による一単語のみを扱う．
- 検索対象
  ウェブ全体．検索語を日本語とするため，検索対象も日本語のページを扱う．
- 検索対象の特徴語群
  検索対象となるページの特徴語とそれに対する重みを数値で表す．
- ユーザの評価
  検索エンジンが検索対象から検索結果を出力し,ユーザがその中から選んでアクセスしたページに対して評価を与える．評価は，ユーザのコンテキストに合致したときにPositiveを,合致しなかったときにNegativeの評価を与える．

2.3 アルゴリズム

本方式のアルゴリズムを以下に示す．Phase1が検索エンジンの動作，Phase2が本方式の中心である．

- Phase.1
  - Step.1 ユーザから入力された検索語を受け取る．
  - Step.2 検索語を検索エンジンに送り，検索結果（ページのランキング）をブラウザに出力
- Phase.2
  - Step.3 検索結果の中からユーザが選んでアクセスしたページAの内容を出力．ページAの特徴語を抽出する．特徴語の抽出方法は，1) ページから名詞を抽出，重み付けをする 2) HTMLの構造に従って重みを加算する，によって行う．
  - Step.4 ページAに対するユーザの「Positive」ないしは「Negative」の評価を受け取り，Step.3で抽出した特徴語とあわせて評価軸となるコンテキストベクトルを生成する．
  - Step.5 Step.4で生成したベクトルをもとにStep.2で取得した検索結果の各ページとの相関量を計算する．各ページのベクトルはStep.3と同様に特徴語を抽出し，生成する．
  - Step.6 相関量の高いページを強調表示,相関量の低いページを非強調表示してブラウザに出力する．

以下，Step.3～6をユーザの検索が終了するまで繰り返す．

1. 重みの定義
   重みとは，ページの内容をより特徴付ける語を，高い数値で表す．
2. コンテキストベクトルの生成方法
   Positiveの評価は，アクセスしたページAの内容がユーザの与えた検索語のコンテキストに合致した場合に指定される．このとき，Step3で抽出したページAの特徴語群が検索語のコンテキストを表すものとみなす.検索語のコンテキストをコンテキストベクトルQ_1と表したとき，Q_1はページAの特徴語群を要素とし各特徴語に対してStep3で求めた重みを値とする．Q_1との相関量が高い検索結果の各ページは適切なコンテキストを持っていると言える.それら適切なコンテキストを持っているページは検索結果において強調表示する.その結果として,検索結果でユーザは強調表示された適切なコンテキストを持ったページへのアクセスが容易となる．
   Negativeの評価は，アクセスしたページAの内容がユーザの与えた検索語のコンテキストに合致しなかった場合に指定される．この場合，検索語のコンテキストをコンテキストベクトル$(-1)Q_1$で表す．これはQ_1の各要素の正負を反転させたベクトルである．$(-1)Q_1$との間で負の相関の高いページは不適切なコンテキストを持っているとし，検索結果において非強調表示する．検索結果でユーザはコンテキストの合致しないページに誤ってアクセスすることがなくなる．
   ページAの次にアクセスしたページBにおいてPositiveないしはNegativeの評価が与えられたとき,コンテキストベクトルベクトルQ_1とページBの特徴語群ベクトルを加算したコンテキストベクトルQ_2を生成する．合成された（学習した）コンテキストベクトルQ_2との相関量(内積)によって検索結果を強調/非強調表示する．検索が終了するまで学習は繰り返される．

3. システム実現方式

前節で述べたアルゴリズムをブラウザから操作可能なウェブアプリケーションとしてサーバシステム上に実現した．

3.1 検索エンジン

検索エンジンには，GoogleのAPI[15]を利用した．Google

のAPIは，XMLとHTTPを利用したプロトコルである[16]を用い，プログラムからGoogleの検索エンジンにアクセスすることを可能にする．ブラウザでユーザがフォームに入力した検索語をAPIに送信する．SOAPを介して検索結果のランキング上位10件を取得し，ページのタイトル，ページのURLをブラウザに出力する．

3.2 特徴語抽出によるコンテキストベクトル生成

検索結果の各ページの特徴語を抽出する．特徴語抽出にはまず，MeCab[17]による形態素解析を行う．形態素解析された結果から名詞と未知語（品詞が判別できなかった単語）を取得する．続いて取得した各語に対して表１のような規則で重み付けを行う．

表１ 抽出語に対応する重み
Table1: Weight for each extracted keyword.

規則	加算される重み
名詞	+1
未知語	+1
固有名詞	+2
H1に含まれる	+2
Titleに含まれる	+3
検索語の前後	+3

固有名詞は特に特徴づけるとみなし，重みを強くする．HTMLによるページは半構造データであり，見出し語を表すh1およびページのタイトルを表すtitle要素に含まれる文字列はページの特徴を表している単語が含まれている可能性が高い．したがってh1，title要素に含まれる名詞および未知語は重みを強くする．また，検索語の直前直後にある名詞および未知語も検索語のコンテキストを表しているとみなし，重みを強くする．ページの特徴語をすべて抽出した後，最も重みの強い語の値で正規化する．各語とその重みを要素としたベクトルをページA～Jのベクトルとし，検索語のコンテキストを表すコンテキストベクトル(P_A:Aはページ ID)とする．

$$P_A = \left(\frac{W_1}{\underset{j=1,n}{MAX}(W_j)}, \frac{W_2}{\underset{j=1,n}{MAX}(W_j)}, \cdots \frac{W_n}{\underset{j=1,n}{MAX}(W_j)} \right)$$

3.3 コンテキストの学習

Positive, Negativeの評価について，検索結果のページに戻る際にユーザがページの内容に満足した（検索語のコンテキストに合致した）場合はPositive，ページの内容に満足しない（検索語のコンテキストに合致しない）場合はNegativeを指定する．アクセスしたページAの特徴語（P_A）の重みの強い上位10語を取り，Positive, Negativeの評価と前回までの評価軸（Q_p）から新たに学習された，検索語のコンテキストベクトルQ_cを以下の式で示す．

$$Q_c = \frac{Q_p E P_A}{a}$$

- Q_p: 学習の対象となっているコンテキストベクトルQ_c
- E: *Positive=1, Negative=-1*
- a: 定数（Qの要素の合計値を平均化するもので，本実装では*a=2*と設定している）

3.4 表示方法

検索結果のページB～Jの$P_{b,j}$とP_A から求めたQ_cとの相関量を計算する．相関量は内積値から求める．相関量に対して閾値X，Yを設けてX以上のページは強調表示，Y以下のページは非強調表示する．本実装では*X=0.5, Y=-0.5*とした．

4. 実験

ここでは，本方式を実現したシステムを対象とした実験について述べる．本実験では，検索対象となるウェブを日本語のページに限定した．

検索語「ドライブ」に対して検索語のコンテキストが「CD-ROMなど記憶媒体を読み書きする装置」だった場合に，本システムを利用した際の挙動を示す．ここで「ドライブ」とは一般に「CD-ROMなど記憶媒体を読み書きする装置」「（車を運転する）ドライブ」の2つの意味を持つ多義語である．意味は異なるが同じ語源であり検索語として「ドライブ」とだけ与えられた場合には意味の違いを判別することができない．本実験により，本方式を用いることで検索結果の各ページに含まれる「ドライブ」がどちらの意味で使われているかを反映した検索結果を表示し，ユーザが求める内容のページへのアクセスが容易になることを示す．

ブラウザのフォームに検索語「ドライブ」を入力し，送信，検索結果が出力される（図１）．ユーザは，まず検索結果のランキングの1位に表示されているページIDがAのページにアクセスする．このページは「車を運転する」意味の「ドライブ」に関するページであるため，ユーザの検索語のコンテキストに合致しない．よって，Negativeを指定し検索結果に戻る．ここでコンテキストベクトルは図２に示すとおりである．各ページと得られたコンテキストベクトルの相関量に応じた検索結果が出力される（図３）．

ページID	タイトル
A	ドライブガイド
B	ニッポンレンタカー
C	SonyDrive ソニー製品情報
D	MediaDrive-メディアドライブ株式会社
E	関東近郊ドライブマップ（メイン）
F	ドライブ A GO GO!
G	[Smap]～エスマップ～
H	比叡山ドライブウェイ
I	株式会社ハーモニック・ドライブ・システムズ
J	ハイウェイドライブカレンダー

図１ 検索語「ドライブ」に対する検索結果
Figure1: Search results (titles) for the term of “drive”

要素	値
ドライブ	-1
ガイド	-0.38
交通	-0.33
プレゼント	-0.33
道路	-0.33
ゲーム	-0.33
リンク	-0.33
旅行	-0.33
観光	-0.33
トラベル	-0.33
クルマ	-0.33

図２ 検索語「ドライブ」のコンテキストベクトル
Figure2: Context vector of “drive”

ページID	コンテキストベクトルとの相関
A	-2.15
B	-0.98
C	-0.04
D	-0.07
E	-1.23
F	-1
G	-0.5
H	-0.38
I	-0.57
J	-0.14

図３ 検索語「ドライブ」に対する検索結果２
Figure3: The other example of search results for the term of “drive”

IDがA, B, E, F, G, Iのページが非強調表示される．非強調表示されていない検索結果の中でランキングが最も高いCのページにアクセスする．このページの内容はユーザの検索語のコンテキストと合致している．よってPositiveの評価を与えて検索結果に戻る．前回の評価軸から学習したコンテキストベクトル（図４）との相関量に応じた検索結果が出力される（図５）．

要素	値
ドライブ	-1
ガイド	-0.38
交通	-0.33
プレゼント	-0.33
道路	-0.33
ゲーム	-0.33
リンク	-0.33
旅行	-0.33
観光	-0.33
トラベル	-0.33
クルマ	-0.33
情報	0.17
無償	0.21
デジタル	0.32
案内	0.42
ソニー	0.46

図４ 検索語「ドライブ」のコンテキストベクトル（学習後）
Figure4: Context vector of “drive” after learning

ページID	コンテキストベクトルとの相関量
A	-2.15
B	-0.55
C	0.95
D	0.49
E	-1.11
F	-1
G	-0.48
H	-0.38
I	-0.46
J	-0.08

図５ 検索語「ドライブ」に対する検索結果３
Figure5: Search results of “drive”

10の検索結果のうち，コンテキストに合致したページは2あり（検索エンジンの適合率が20%）それらは相関量の上位2件に入っている（本システムの再現率が100%）．この結果より，検索語が多義語だった場合に，本方式を用いることによりユーザのふるまいからコンテキストを判別し，そのコンテキストを反映した検索結果を表示することによってユーザが求める内容のページへのアクセスが容易になることが示された．

5. 結論

本稿では，検索エンジンを利用する際のユーザのふるまいをもとに，与えられた検索語のコンテキストをダイナミックに学習し検索結果を絞り込んで表示するシステムを実現した．本システムの特徴は，検索エンジンを利用する際のユーザに特別な行動を取らせユーザビリティを損ねることなく検索を容易にする点である．本実験を通じて，検索エンジンが返す検索結果にユーザのコンテキストに合致しないページが多く含まれている際に効果的であることを示した．検索エンジンの検索結果の適合率が100%に近い場合には，ユーザの検索がすぐに終了するため本システムは必要なく，検索エンジンからの検索結果10件からコンテキストに合致したページを探すため，本システムが効果を発揮するには検索エンジン自体にある程度の適合率の高さが必要である．よって，今後の課題として検索エンジン自身の精度に左右されない方式をつくることが挙げられる．本システムではGoogleを利用したが，それに加えて他の検索エンジンの結果も併用するメタ検索エンジンを利用する，あるいは相関量を計算する対象をGoogleの検索結果の上位10件以上とするなどの方法が考えられる．ユーザのコンテキストを判断する上で，本システムでは，ユーザの直前のアクセス内容のみを利用したが，長期間に渡るユーザの興味分野，静的な属性も反映した方式を今後実現する予定である．

[文献]

[1] L. Page, S. Brin, R. Motwani and T.Winograd: “The PageRank Citation Ranking: Bringing Order to the Web”, Stanford Digital Library working paper SIDL-WP-1999-0120 (version of 11/11/1999).
[2] 福島俊一: “検索エンジンの仕組みと技術の発展”，情報の科学と技術，54 巻2 号，2004，pp. 66-71
[3] George A. Miller: “The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information”, *The Psychological Review*, 1956, vol. 63, pp. 81-97
[4] 米IDC: “検索結果の2 ページ目以降を探すユーザーの割合は、たった5 %にすぎない”，2002
[5] JANSEN,B.J.,SPINK,A.,BATEMAN,J.,AND SARACE- VIC, T. :“Real life information retrieval: a study of user queries on the web.”, *ACM SIGIR Forum 32*, 1 (1998), 5–17.
[6] Oren Zamir, Oren Etzioni:“A Feasibility Demonstration”
[7] Oren Zamir, Oren Etzioni, Oren Madanim:“Grouper: A Dynamic Clustering Interface to Web Search Results”, *Computer Networks*, 1999
[8] Vivisimo http://vivisimo.com/
[9] Rocchio, J.J.: “Relevance feedback in information retrieval”, *The SMART Retrieval System Experiments in Automatic Document Processing*, Prentice Hall, Inc., pp.313-323 (1971)
[10] 辻祐樹，藤本典幸，萩原兼一: “検索質問に含まれる単語と適合文書内の単語の距離に着目した適合フィードバックの改善”，*DEWS2004*，I-1-04
[11] 田中貴志，中島伸介，田中克己: “適合フィードバックにおける複数ユーザの対話からの動的質問修正”，*DEWS2003*，6-B-04
[12] 大橋英博，清木康: “情報通信分野を対象とした意味的連想検索機構によるWWW 検索エンジンの実現”，情報処理学会研究報告 2001-DBS-125(I), pp.233-240, 2001.
[13] S. Mukherjea and Y. Hara: “VisualizingWorld-WideWeb search engine results”, *Proceedings of 1999 IEEE International Conference on Information Visualization*, 1999, pp.400-405
[14] 松田耕史: “統計的手法によるWeb 検索補助システムSeezleの開発”，未踏ソフトウェア創造事業採択プロジェクト2003-2004
[15] Google Web APIs http://www.google.com/apis/
[16] SOAP Specifications http://www.w3.org/TR/soap/
[17] MeCab: Yet Another Part-of-Speech and Morphological Analyzer http://chasen.org/ taku/software/mecab/
[18] G. Salton: “Developments in Automatic Text Retrieval”, *Science*, Vol. 253, pages 974-979, 1991.

船木 信宏 Nobuhiro FUNAKI
慶應義塾大学環境情報学部在学中．日本データベース学会学生会員．

清木 康 Yasushi KIYOKI
1983 年慶應義塾大学大学院工学研究科博士課程修了．工学博士．1984 年筑波大学電子・情報工学系講師，助教授を経て，1996 年慶應義塾大学環境情報学部助教授，1998 年同大学教授，現在に至る．データベースシステム，知識ベースシステム，意味的連想検索，マルチメディアデータベース，感性データベースの研究に従事．ACM，IEEEーCS，情報処理学会，日本データベース学会会員．