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

船木　信宏† 　清木　康†

† 慶應義塾大学 環境情報学部
E-mail: †funaki@mdbl.sfc.keio.ac.jp, ††kiyoki@sfc.keio.ac.jp

あらまし　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．

キーワード　Web 利用技術，問い合わせ処理，情報検索，WWW 検索エンジン

A WWW search system with functions for query-context extraction reflecting user's evaluation

Nobuhiro FUNAKI† and Yasushi KIYOKI†

† Faculty of Environmental Information, Keio University
E-mail: †funaki@mdbl.sfc.keio.ac.jp, ††kiyoki@sfc.keio.ac.jp

Abstract　Current WWW search engines performance and quarity have been improved. However the search results provided by search engines include 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 it is around 7 ± 2 items. Most popular search engines return 10 pages in their search results. This fact detracts from their usability. In this paper, we present a web search system which returns search results reflecting user's query-context. 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 implimentation and exprimental results to clarify the feasibility our system.

Key words　Web Application Technology, Query Processing, Information Retrieval, WWW Search Engine

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（図 1），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

検索エンジン名称	ページ数
Yahoo!Japan	10
Google	10
goo	10
MSN サーチ	10

図 1　主な検索エンジンのデフォルトの検索結果ページ数

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]．ユーザは自身のコンテキストと関連する分類を発見することで情報獲得が容易になる．

[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]．ユーザは検索結果の各ページに対してコンテキストに合致しているかどうか評価を与える必要がある．

[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．

ページ ID	単語	重み
A	りんご	0.8333
A	産直	0.375

図 2　データ構造: 特徴語群例

ページ ID	Positive / Negative
A	Positive
B	Negative

図 3　データ構造: ユーザの評価例

ページ ID	タイトル
A	ニッポンレンタカー
B	ドライブガイド
C	Sony Drive ソニー製品情報

図 4　検索語「ドライブ」に対する検索結果抜粋

[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.（図 2）

[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.（図 3）

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．

[Phase1]

(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 図 4）をブラウザに出力する．

[Phase2]

(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.（下記 (1)，図 5 参照）
(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.（下記 (2)，図 6 参照）
(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.　実　　装

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上に実装した.

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 SOAP [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.（図 7）

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

特徴語	北海道	水戸黄門	コース	ニッポン	ドライブ	レンタカー	カナダ	モンテカルロ	陸前	ラリー	旅行	十勝	自動車
重み	1	1	0.9	0.8	0.6	0.55	0.5	0.5	0.4	0.4	0.35	0.3	0.25

図 5　ページ A の特徴語群

要素	北海道	水戸黄門	コース	ニッポン	ドライブ	レンタカー	カナダ	モンテカルロ	陸前	ラリー
値	1	1	0.9	0.8	0.6	0.55	0.5	0.5	0.4	0.4

図 6　検索語「ドライブ」のコンテキストベクトル Q_1

図 7　ブラウザに検索結果を出力

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

図 8　重み付けの規則

図 9　Positive，Negative ボタン

図 10　強調，非強調表示

る．続いて取得した各語に対して図 8 のような規則で重み付けを行う．

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とする．

$$P_a = \left(\frac{W_1}{MAX(W)}, \frac{W_2}{MAX(W)}, \cdots, \frac{W_n}{MAX(W)} \right)$$

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．（図 9）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．

$$Q_c = \frac{Q_p + EP_i}{a}$$

- 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

検索結果の各ページの P_{b-j} と 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．（図 10）本実装では $X = 0.5$，$Y = -0.5$ とした．

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	Sony Drive 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 11　検索語「ドライブ」に対する検索結果

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

図 12　検索語「ドライブ」のコンテキストベクトル

ページ ID	コンテキストベクトルとの相関量
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 13　検索語「ドライブ」に対する検索結果 2

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．

4.1 実　験　1

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 11）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 12 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 13）

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 14）との相関量に応じた検索結果が出力される．（図 15）

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．

4.2 実　験　2

検索語「ギョウザ」に対して検索語のコンテキストが「ギョウザのレシピが知りたい」場合に，本システムを利用した際の挙動を示す．ここで「ギョウザ」は一般に中国料理の一つであ

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

図 14　検索語「ドライブ」のコンテキストベクトル（学習後）

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 15　検索語「ドライブ」に対する検索結果 3

ページ ID	タイトル
A	ギョウザ・点心 料理レシピ／キッコーマン ...
B	Yahoo!グルメ - レシピ情報 - 焼きギョウザ
C	大宮ぎょうざ (餃子/ギョウザ)OM-GYOUZA
D	餃子スキなら迷わず JOIN
E	ギョウザ・シューマイ類
F	スープギョウザ 2
G	必見！目がテン!?ライブラリー
H	ギョウザ
I	【楽天市場】銀座 直久・生餃子（60 個）：通販“ ...
J	焼きギョウザ

図 16　検索語「ギョウザ」に対する検索結果

要素	kcal	分	中華	風	ギョウザ	レシピ	春巻き	マイ	点心	エスニック
値	1	0.91	0.88	0.88	0.57	0.32	0.25	0.22	0.22	0.17

図 17　検索語「ギョウザ」のコンテキストベクトル

ページ ID	コンテキストベクトルとの相関量
A	4.29
B	0.78
C	0.06
D	0
E	0.38
F	0.71
G	0.62
H	0.57
I	0.02
J	0.90

図 18　検索語「ギョウザ」のコンテキストベクトルとの相関量

る食べ物の意味しか持たない．しかし，ユーザのコンテキストが「ギョウザのレシピが知りたい」場合に,「ギョウザのお店」に関する内容のページはユーザにとって不必要である．意味は同一であるがコンテキストによって表示すべき内容が異なる検索語の場合に，コンテキスト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を容易にすることを本実験では示す.

ユーザは検索語「ギョウザ」を送信，検索結果を出力する（図 16）．ランキングの 1 位に表示された「ギョウザ・点心 料理レシピ／キッコーマン ...」にアクセス，コンテキストに合致するので Positive の評価を与えコンテキストベクトルを得る.（図 17）コンテキストベクトルとの相関量を計算し，反映した検索結果を出力する.（図 18）,

10 の検索結果のうち，コンテキストに合致したページは 6 ある.（検索エンジンの適合率が 60%，A,B,E,F,H,J がコンテキストに合致）本システムが出した評価軸との相関量上位 6 件中 5 件（A,B,F,G,H）が強調表示されている．すなわち，本システムは再現率 83%を得た．この結果より，意味が同一であるがコンテキストによって表示すべき内容が異なる検索語の場合に，コンテキストに合致したページを強調表示することで，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.

4.3 実　験　3

検索語「小泉」に対して検索語のコンテキストが「タレントの小泉今日子」の場合に本システムを利用した際の挙動を示す.

要素	メール	マガジン	内閣	登録	流れ	トップ	毎週	掲載
値	-1	-0.95	-0.50	-0.45	-0.16	-0.16	-0.16	-0.12

図 20　検索語「小泉」のコンテキストベクトル

要素	家具	メール	マガジン	内閣	登録	学習	子供部屋	こども	子供	流れ
値	-1	-1	-0.95	-0.5	-0.45	-0.44	-0.22	-0.22	-0.22	-0.16

図 21　検索語「小泉」のコンテキストベクトル（学習後２）

要素	家具	メール	マガジン	内閣	登録	学習	子供部屋	こども	子供	流れ	koizumi	今日子	kyonkyon
値	-1	-1	-0.95	-0.5	-0.45	-0.44	-0.22	-0.22	-0.22	-0.16	0.85	0.52	0.28

図 22　検索語「小泉」のコンテキストベクトル（学習後３）

ページ ID	タイトル
A	小泉内閣メールマガジン
B	小泉総理プロフィール
C	首相官邸ホームページ
D	小泉産業株式会社／照明・家具
E	***(((KOIZUMIX)))***
F	小泉清人ホームページ
G	小泉文夫記念資料室
H	小泉としあき（衆議院議員）公式ホームページ
I	味噌 天然-味噌 お味噌作りサポート隊!
J	小泉エリ＆トモのホームページ (小泉えり ...

図 19　検索語「小泉」に対する検索結果

ページ ID	評価軸との相関量
A	-2.64
B	-0.16
C	-0.95
D	0
E	0
F	-0.04
G	-0.03
H	0
I	-0.02
J	-0.01

図 23　検索語「小泉」に対する検索結果２

ページ ID	評価軸との相関量
A	-2.64
B	-0.16
C	-0.95
D	-1.19
E	0
F	-0.04
G	-0.03
H	0
I	-0.02
J	-0.01

図 24　検索語「小泉」に対する検索結果３

ここで「小泉」は人名であり複数の人物を指し示している．そのため検索語のみからユーザのコンテキストがどの人物を指しているかを判別することはできない．本実験では，本方式を用

ページ ID	評価軸との相関量
A	-2.62
B	0.01
C	-0.22
D	-1.35
E	1.81
F	0.23
G	-0.12
H	0
I	0.16
J	0.10

図 25　検索語「小泉」に対する検索結果４

いることにより複数の対象を指す検索語からユーザのコンテキストに合致する内容を持つページのみを強調表示し，求めるページへのアクセスが容易になることを示す．

ユーザは検索語「小泉」を送信，検索結果を出力する（図 19）．ランキングの 1 位に表示された「小泉内閣メールマガジン」にアクセス，コンテキストに合致しないため Negative を指定しコンテキストベクトルを得る．(図 20）コンテキストベクトルとの相関量を反映した検索結果を表示する（図 23）．負の相関が最も低くランキング最上位の「小泉産業株式会社／照明・家具」にアクセス，Negative を指定しコンテキストベクトルを学習させる（図 21）．検索結果を表示する（図 24）．負の相関が最も低く（すなわち正の相関が最も高い）ランキング最上位の「***(((KOIZUMIX)))***」にアクセス，Positive を指定しコンテキストベクトルを得る．(図 22）コンテキストベクトルとの相関量を反映した検索結果を表示し，検索を終了する．(図 25）．

10 の検索結果のうち，コンテキストに合致したページは 1 あり（検索エンジンの適合率が 10%）上記手順を踏むことで本システムは再現率 100%を得た．以上により，検索語が複数の対象を指す場合に本方式を用いることで，ユーザのコンテキストに合致した対象のページのみを強調表示し，ユーザが求めるページへのアクセスが容易になることを示した．

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: “Visualizing World-Wide Web 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.