論 文

Technical Papers

ユーザ個人の閲覧履歴からのキーワード抽出によるブラウジング支援

Browsing Support by Highlighting Keywords based on User's Browsing History

松尾 豊
Yutaka Matsuo
産業技術総合研究所サイバーアシスト研究センター
Cyber Assist Research Center, National Institute of Advanced Industrial Science and Technology
y.matsuo@carc.aist.go.jp, http://www.carc.aist.go.jp/

福田 隼人
Hayato Fukuta
東京大学工学部*1
School of Engineering, University of Tokyo
hfukuta@miv.t.u-tokyo.ac.jp

石塚 満
Mitsuru Ishizuka
東京大学情報理工学系研究科
School of Information Science and Technology Engineering, University of Tokyo
ishizuka@miv.t.u-tokyo.ac.jp, http://www.miv.t.u-tokyo.ac.jp/~ishizuka/

keywords: browzing support, keyword extraction, χ^2 test, IRM

Summary

We develop a browsing support system which learns user's interests and highlights keywords based on a user's browsing history. Monitoring the user's access to the Web enables us to detect "familiar words" for the user. We extract keywords at the current page, which are relevant to the familiar words, and highlight them. The relevancy is measured by the biases of co-occurrence, called *IRM* (Interest Relevance Measure). Our system consists of three components; a proxy server which monitors access to the Web, a frequency server which stores frequency of words in the accessed Web pages, and a keyword extraction module. We show the effectiveness of our system by experiments.

1. ま え が き

World Wide Web を使って，さまざまな情報を得る機会が増えている．Yahoo!や Asahi.com などのポータルサイトは，多くの人の情報要求に応えることで多くのアクセスを獲得している．このような "one-size fits all" という方向性と対照的に，個人の文脈を意識した情報の提示を行う試みもあり [Suryanarayana 02]，ビジネスでも注目を集めている [Forey 02]．例えば，Yahoo!では，ユーザが興味あるニュースだけを表示する My Yahoo!*1というサービスがあり，Amazon*2ではクリックしたデータをもとにユーザの興味を把握し，本を薦めるマイページというサービスが行われている．Web 上の振る舞いからユーザの興味を把握し，個人化した情報提示を行う研究は数年前から活発である [Mobasher 01]．特に，プロキシサーバを用いて個人の Web の閲覧情報を収集する研究がいくつか行われている．

個人の状況や文脈に基づいて情報を提示するという方向性は，将来，マイボタン [中島 01] などのように，携帯端末による情報へのアクセスが高度にそして容易になるにしたがって必要性が高まってくるだろう．ユーザの情報アクセスが携帯端末というひとつのエージェントを通して行われ，行動の履歴を利用することに対するプライバシーの問題が軽減されるためである．したがって，ユーザが情報にアクセスした履歴を用いて，ユーザの目的や興味に応じた適切な情報支援を行うことは将来的に重要な基盤技術になると考えられる．

一方，現在の検索や要約などの自然言語文書の処理においては，語の重みづけのために TF-IDF[Salton 89] などの手法が用いられており，文や文書の重要度・類似度の計算が行われる場合が多い．基本的に対象となる文書とコーパスを用いて，語の重みを決定しており，どんなユーザが読むかといった個人の文脈による違いは考慮されないのが普通である．しかし，ユーザの文脈により語の重みが異なるのはごく自然なことに思える．例えば，「イチロー球宴 2 年連続トップ確実」という新聞記事があったとしよう．野球に興味のない人であれば，「イチロー」

†1 現在，東日本旅客鉄道株式会社。
*1 http://my.yahoo.co.jp/
*2 http://www.amazon.co.jp/

という語により，この記事が野球に関する内容であることが分かるので「イチロー」という語が重要であると考えるかもしれない．一方，野球に興味がある人にとっては「球宴 2 年連続トップ」，つまりオールスターの得票が 2 年連続でトップであることが重要であると考えるだろう．さらにもっと頻繁にニュースをチェックしている人ならば，トップが「確実」になったことがこの記事の重要な点であると理解するだろう．このように，ユーザの興味によって，同じ文書でも重要な語が異なると考えるのは自然であろう．

そこで，本論文ではプロキシサーバを用いてユーザ個人が閲覧した文書の履歴を取得し，対象とする文書から，そのユーザにとって重要度の高い語を抽出するキーワード抽出法を提案する．さらに，キーワードをハイライトすることでブラウジングを容易にするシステムを構築する．ハイライト表示されている語に注意して読むことで，自分の興味に関連した部分を集中的に読むことができ，興味ある情報を読み飛ばさなくてすむ．また，内容を早く理解することができる．論文や本を読むときに，自分の興味があるところ，重要なところに下線を引きながら読んだ経験があるだろう．本論文では，これを自動的に実現するシステムの構築を試みる．

以下，2 節では関連研究を概観し，3 節で本論文のアイディアとアルゴリズムについて述べる．3 節でアルゴリズムの評価を行った後，4 節でブラウジング支援システムの構成，5 節でブラウジングシステムの評価を行う．6 節で議論を行い，7 節で結論を述べる．

2. 関 連 研 究

ブラウジングの支援に関する研究は，10 年ほど前から行われている．Letizia[Lieberman 95] は，ユーザのブラウジングを補助するためのユーザインタフェースエージェントである．ユーザの振る舞いを監視し，さまざまなヒューリスティックを用いてユーザの興味を予測し，ページ内のリンクの先読みと推薦を行う．WebWatcher[Joachims 97] は，博物館のツアーガイドのメタファーで，ユーザが特定の Web サイトを探索するのを支援するエージェントである．最初にユーザの興味を入力してもらい，プロキシサーバを介すことで，関連したページや推薦するリンクなどを表示する．

ユーザの興味を適切に管理することも重要である．WebMate[Chen 98] はプロキシを用いたシステムであり，ページの閲覧中にユーザが “I like it” というボタンを押すと，ユーザからの正のフィードバックが得られる．ユーザの異なるドメインにおける複数の興味に対応するため，複数の TF-IDF ベクトルによりユーザの興味を管理する．PVA(Personal View Agent)[Chen 01] も同じくプロキシを用いたシステムであり，ユーザの興味を学習し管理する．移りゆくユーザの興味に対応するため人工生命的な考え方を導入し，ユーザが閲覧したページと近い興味カテゴリはエネルギーを得る仕組みとしている．エネルギーが多くなれば，カテゴリはサブカテゴリに分かれ，少なくなればカテゴリは併合する．各カテゴリは，文書のキーワードのベクトル（TF ベクトル）で表現される．Alipes[Widyantoro 99] は，オンラインの新聞や雑誌の記事を対象にした個人化ニュースエージェントであり，ユーザの長期的興味，短期的な正の興味，短期的な負の興味という 3 つのプロファイルを管理し，ニュースのフィルタリングを行う．プロファイルは，それぞれキーワードのリストと重みで表現され，短期的な興味は即座に，長期的な興味はゆっくりと更新される．

この他にも，Syskill&Webert[Pazzani 96] や SiteIF [Magnini 01] など，さまざまなシステムが提案されている．しかし，これらの研究では，語の重みづけに TF-IDF や TF を用いることが一般的であり，本論文のように，語の重みづけをユーザの閲覧履歴を考慮して定義するものではない．

3. キーワードの抽出

Web ページは，画像や動画だけのものを除くと，その多くが何らかの文字情報を含んだテキストである．このテキスト中に含まれる語の重みを計算し，重みの高い語，つまりキーワードを抽出する．

我々は，以前，単一の文書からキーワードを抽出する手法 [松尾 02]（ここでは簡単のため，CF 法（Chi-square measure to Frequent words) と呼ぶことにする）を考案した．本論文における語の重み付けはこの手法の拡張であるので，まず CF 法について簡単に概要を説明する．

3·1 CF 法：頻出語との共起の偏りに基づくキーワード抽出

ひとつの文書が与えられたとき，語の出現頻度を数えることで，頻出語 $g \in G$ を取り出すことができる．ここで語とは，単語もしくは複数の単語からなるフレーズである．次に，語が同一文内に出現すれば共起と考えることで，語の共起の頻度を集計し共起行列を作ることができる．例えば，Alan Turing による有名な論文 “Computing machinery and intelligence”[Turing 50] の頻出語と共起行列 (の一部) は表 1，表 2 のようになる．

仮に，語 w が頻出語 $g \in G$ と全く独立に生起するなら，語 w と語 $g \in G$ が共起する確率は，語 w 単独での生起確率と同様の分布になるはずである．一方，語 w と頻出語 $g \in G$ の間に何らかの意味的なつながりがあれば，この確率は偏ることになる．

図 1，図 2 に，いくつかの語と頻出語との共起確率*3の分布を示す．“kind” や “make” などの語は，どの頻出語

*3 合計が 1 になるように正規化したもの．

表 1 頻出語の出現頻度と確率分布

語	a	b	c	d	e	f	g	h	i	j	Total
頻度	203	63	44	44	39	36	35	33	30	28	555
確率	0.366	0.114	0.079	0.079	0.070	0.065	0.063	0.059	0.054	0.050	1.0

a: *machine*, b: *computer*, c: *question*, d: *digital*, e: *answer*, f: *game*, g: *argument*, h: *make*, i: *state*, j: *number*

表 2 共起行列

	a	b	c	d	e	f	g	h	i	j	Total
a	—	30	26	19	18	12	12	17	22	9	165
b	30	—	5	50	6	11	1	3	2	3	111
c	26	5	—	4	23	7	0	2	0	0	67
d	19	50	4	—	3	7	1	1	0	4	89
e	18	6	23	3	—	7	1	2	1	0	61
f	12	11	7	7	7	—	2	4	0	0	50
g	12	1	0	1	1	2	—	5	1	0	23
h	17	3	2	1	2	4	5	—	0	0	34
i	22	2	0	0	1	0	1	0	—	7	33
j	9	3	0	4	0	0	0	0	7	—	23
…	…	…	…	…	…	…	…	…	…	…	…
u	6	5	5	3	3	18	2	2	1	0	45
v	13	40	4	35	3	6	1	0	0	2	104
w	11	2	2	1	1	0	1	4	0	0	22
x	17	3	2	1	2	4	5	0	0	0	34

u: *imitation*, v: *digital computer*, w:*kind*, x:*make*

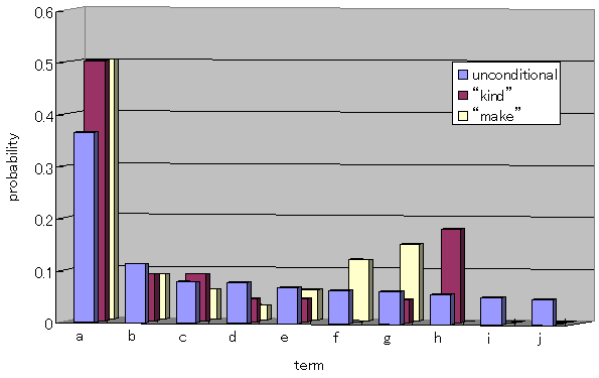

図 1 頻出語と {“kind” もしくは “make”} との共起の分布

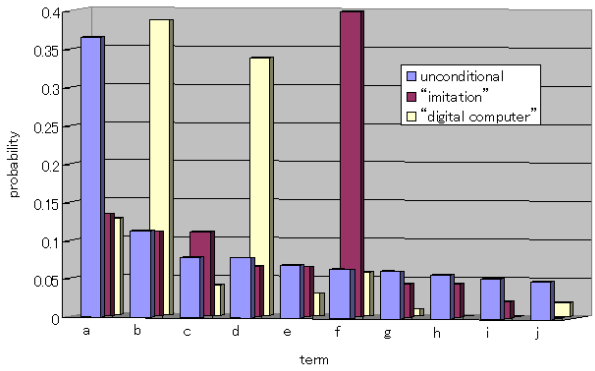

図 2 頻出語と {“imitation” もしくは “digital computer”} との共起の分布

とも共起しているが，“imitation” や “digital computer” は，特定の語とだけ選択的に共起している．このような偏りは，筆者が意味的なつながりを考慮し文書を書き進めていく上で生まれたものであり，分布が偏っている語は文書中において何らかの重要な意味を担っている語であると考えられる．実際，この論文は「機械は思考できるかという問いを imitation game によって置き換える」ことを提案しており，“imitation” や “digital computer” などの語は論文中で重要な語である．

この偏りを測るために，[松尾 02] では，次のように定式化している．頻出語単独での生起確率を理論確率 $p_g(g \in G)$ とし，語 w と頻出語群 G の共起の総数を n_w，語 w と語 $g \in G$ の共起頻度を $freq(w,g)$ とすると，語の共起の偏りを表す統計量 χ^2 は以下の式で与えられる．

$$\chi^2(w) = \sum_{g \in G} \frac{(freq(w,g) - n_w p_g)^2}{n_w p_g} \qquad (1)$$

Turing の論文において頻出語の上位 24 語を G とすると，χ^2 値の上位語は表 3 のようになる*4．

3·2 IRM：ユーザの興味との偏りに基づくキーワード抽出

CF 法は，誰が読むかに関わらず，頻出語との共起が偏った語をキーワードとして抽出するものであった．しかし，読者ごとに文書の捉え方や着目点もさまざまであり，重要と感じる語も読者によって異なるであろう．

*4 なお，[松尾 02] では，さらに頻出語のクラスタリングの処理などいくつかの改善を行っている．

表 3 χ^2 値の上位語

χ^2 値	頻度	語
196.9	16	imitation game
88.9	15	play
62.4	9	human computer
60.1	3	card
57.1	4	future
50.4	10	logic
45.1	7	identification
44.4	6	universality
42.7	30	state

表 4 頻出語に “logic” を加えた場合の χ^2 値の上位語

χ^2 値	頻度	語
196.6	16	imitation game
88.5	15	play
84.4	3	logic system
62.2	9	human computer
60.0	3	card
57.0	4	future
44.9	7	identification
44.2	6	proposition
43.9	5	limitation

表 5 頻出語に “God” を加えた場合の χ^2 値の上位語

χ^2 値	頻度	語
196.2	16	imitation game
113.8	6	animal
88.2	15	play
62.0	9	human computer
59.9	3	card
56.9	4	future
49.8	10	logic
44.7	7	identification
43.9	5	woman
40.8	5	book

CF 法では，共起の偏りを測る対象である G は文書における頻出語であった．ここで，仮に G に “logic” という語を加えてみよう．すると，χ^2 値の計算式には，“logic” との共起の偏りも含まれるようになる．その結果，Turing の論文の例において，上位語は表 4 のように変化する．“logic system” や “proposition” など，“logic” と偏って共起する語が上位に現れている．また，仮に G に “God” を加えてみると，表 5 のように “animal” や “woman” などの語が上位にくる．Turing の論文では,「機械や動物は考えることができない」という宗教的な視点からの批判に対処するために議論を行っている部分があり，そこでは “God” と偏って共起する語として “animal” が抽出される．

このように，G と共起の偏った語は χ^2 値が高くなるので，G を変化させることで χ^2 値の上位語を変えることができる．つまり，G にユーザとなじみの深い語を加えれば，その語に関連した語を抽出することができる．

そこで，式 (1) を，ユーザにとって身近な語との共起を測る形に変更する．ここではプロキシサーバを用いてユーザが閲覧した Web ページからその頻出語を取り出すことにする．ユーザが閲覧した Web ページにおける頻出語は，ユーザがよく知っている身近な語であると考えられる．したがって，ユーザに身近な語の集合を H として，H との共起の偏りが大きい語を取り出せばよい．

ユーザが閲覧した過去のページに何回出現したかにより，各語に対して「身近度」を定義する．ある文書から重要語を抽出する際には，その文書に含まれる語のうちで身近度が高いものを一定数 H として取り出す．そして H との共起の偏りが大きいものを，その文書のユーザにとってのキーワードとして抽出する．さらに，次のようにユーザの興味との関連度の指標 IRM(Interest Relevancy Measure) を定義する．

Definition 1ユーザ k にとってあるページにおける語 w の重要度 IRM を次のように定義する．

$$IRM(w,k) = \sum_{h \in H_k} \frac{(freq(w,h) - n_{w,k}p_h)^2}{n_{w,k}p_h} \quad (2)$$

ここで H_k は，このページに出現する語のうちユーザ k にとって身近度が高い語の集合，$freq(w,h)$ は語 w と語 h の共起の回数，$n_{w,k}$ は語 w と H_k の共起の総数，p_h は語 $h \in H$ 単独でのこのページ中での出現確率とする．

ここでは，H_k は身近度の上位語から 20 語選ぶ[*5]．ユーザの閲覧した文書の履歴がない状態では，各語の身近度がすべて 0 であるが，対象とする文書における語の頻度もカウントした後に H を決定するので，最初のページに関しては頻出語が H になる．ユーザが Web 文書を閲覧するにしたがって，身近度のデータが蓄積され，そのユーザの興味に関連したキーワードが得られるようになる．

なお，Web ページの場合でも，同一文内に出現すれば 1 回共起すると考えるが，Web ページでは文が明示されない場合もあるので，句点，空行を文の区切りとして扱う．また，リストの各行，表の各セルは一文として扱う．

4. IRM の評価

IRM の評価を行うために，次のような評価実験を行った．まず，1998 年，1999 年の新聞記事から，次のような記事群を用意した[*6]．

- 記事群 1: タイトルに「遺伝子」が含まれる記事 20 編．遺伝子治療，遺伝子組み換え作物，ヒトゲノム解読などの記事が含まれる．
- 記事群 2: 記事の内容に「ハイブリッド」を含む記事 20 編．ハイブリッドカーの開発記事，自動車メー

*5 この基準は試行錯誤により定めた．
*6 年月日の早いものから 20 編選んだ．ただし，ニュースダイジェストや，検索語は含むが内容は全く関係ない記事は除いた．

表 6 各手法の評価

記事群	CF 法	TF-IDF	IRM
1	0.280	0.267	0.320
2	0.348	0.284	0.375

カーの提携，環境問題の法制化など，記事群 1 より多様な記事が含まれる．

それぞれの記事群に対して，被験者は決められた順番に 1 記事ずつ読む．各記事に対して，次のようにキーワードを各 10 語抽出する．

- CF 法：閲覧中の記事において，χ^2 値の高い語．
- TF-IDF：それまでに閲覧した記事をコーパスとし，TF-IDF 値が高い語．
- IRM：それまでに閲覧した記事を履歴文書とし，IRM が高い語．

各手法が出した語（重なりがない場合，最大 30 語）をシャッフルし，記事とともにユーザに提示し，「この文書を読む上で興味を持った語にチェックをして下さい．文書の主題ではなく，自分が興味を覚えた語にチェックして下さい.」と指示を行った．被験者は，大学生・大学院生（20 代男性 8 人）であった．

結果を表 8 に示す．いずれの記事群でも IRM 法が最も良い評価を得ている．TF-IDF は，閲覧した文書をコーパスとしているので，コーパスの量が少ないことも評価が悪かった要因のひとつであろう．CF 法はもともと，ある程度長い文書を対象にその文書単独でキーワードを抽出する方法であるので，コーパスの量は関係なく，結果的に TF-IDF 法よりも良い結果となっている．IRM は，CF 法を基準としながら，徐々に身近度の選択が適切になるため，よい結果が得られたと考えられる．また，多様性が高い記事群 2 でも，よい性能を示している*7．

表 7 に例として，3 つの手法を記事群 1 のある記事に適用した場合のキーワードを示す．記事群 1 の文書を 10 文書読んだとした場合の結果である．この記事は，老化に関する遺伝子の研究を紹介した「クローズアップ」という記事で，線虫の寿命遺伝子の研究者である白沢氏の話，たんぱく質を作るクロトー遺伝子，20 代で老化の特徴が現れるウェルナー症，DNA の傷を修復する 3 つの「天使」などの話題について述べられている．この例では，TF-IDF は，この文書中に 42 回出現する「遺伝子」を最も重要度が高く評価してしまう．記事のポイントである詳細な記述に関する語は，あまり抽出できていない．「マウス」は他の文書には出現せずこの文書で 6 回出現するため重要度が高くなっているが，マウスは実験材料に用いられているものであり，この記事のポイントではない．一方，CF 法は，細かい話題についてうまく捉えているが，「現象」や「原因」などの一般的な語も抽出してしまう．IRM では，「遺伝子」「米国」「研究」「老化」などの語は身近語となり，それとの共起が偏っている語が抽出されため，「天使」「寿命 遺伝子」など適切な詳細さの語が抽出されている．

表 7 各手法の上位語の例

順位	CF 法	TF-IDF	IRM
1	老化 症状	遺伝子	天使
2	線虫 寿命	寿命	白沢
3	遺伝子 異常	研究	寿命 遺伝子
4	寿命 遺伝子	発見	たんぱく質
5	現象	マウス	遺伝
6	原因	異常	老化 現象
7	研究所	老化 症状	研究所
8	ウェルナー 症候群	老化 現象	酵素
9	ヒト	ヒト	修復
10	酵素	たんぱく 質	老化 症状

本節における評価は，我々が指定した記事を被験者に読んでもらうことで評価を行ったものであり，実際に被験者がその記事に興味をもっているわけではない．実際には，ユーザは自分が興味を持つ文書を選択的に読むと考えられるので，以下の節では，ユーザが自由にブラウジングを行う状況で支援を行うシステムについて述べる．

5. ブラウジング支援システム

本システムでは，プロキシサーバを用いてユーザが閲覧した Web ページをモニターする．そして，ユーザが閲覧したページに含まれる語がどのくらいの文書で出現したかという頻度を蓄積する．例えば，あるユーザ（実際に 6 節の評価実験で本システムを使用した被験者）が数時間ブラウジングした後，閲覧した文書によく含まれる身近語は表 8，表 9 のようになる．表 8 は，普段ピアノを引いていて楽器に興味がある被験者のデータであり，表 9 はゲームに興味がある被験者である．表から，ユーザの興味を示す語の頻度が高くなっていることが分かる*8．このようにユーザが閲覧する文書をモニターすることで，ユーザにとって身近な語を収集することができる．身近な語と偏って共起する語は，ユーザの興味と関連したキーワードであると考えられる．

そこで，本節では，ユーザが閲覧した Web ページにおける語の頻度を蓄積しておき，現在閲覧しているページのキーワードをハイライト表示するブラウジング支援システムを構築する．ユーザは，自分が今までによく見た語と関連する語がハイライトされるので，内容を容易に理解でき読みやすくなる．

*7 なお，トピックによって重要語の出しやすさやユーザが興味を持っているかどうかが異なるので，記事群 1 と 2 の値を直接比較することはできない．

*8 なお，“document” や “found” などの不要と思われる語も含まれているが，このような語は，ページが見つからないときの “Document not found on this server” などの表示に含まれる語であり，実際に内容のあるページを閲覧するときには影響はない．また，「ページ」や「リンク」「更新」といった語の身近度が高くなっているが，このような語は特定の語と偏った共起をすることが少ないので影響は少ない．

表 8 被験者 1 の身近度

順位	単語	身近度
1	ページ	34
2	Nifty	24
3	Document	20
4	リンク	20
5	Write	19
6	情報	19
…	…	…
23	ピアノ	14
24	研究	13
24	たち	13
24	思う	13
…	…	…
31	音楽	11
31	コーナー	11
…	…	…
59	pianno	9
59	聴く	9
59	趣味	9
59	気軽	9
59	岐阜	9

システムの全体像を図 3 に示す．大きく分けて，プロキシサーバ，キーワード抽出部，そして頻度サーバから構成される．以下，処理の流れを追って説明する．

まず，プロキシサーバは，ブラウザから HTTP のリクエストがくると，要求先の WWW サーバにそのまま送る．応答が返ってきたら順次次のような処理を行う．

(1) テキストファイルかどうかを，ヘッダを見て判断する．テキストファイルでなければ終了．
(2) テキストの行数が 5 行以下であれば終了．
(3) 文字コードを EUC に変換する．
(4) HTML テキストのボディ部分をキーワード抽出モジュールに送る．

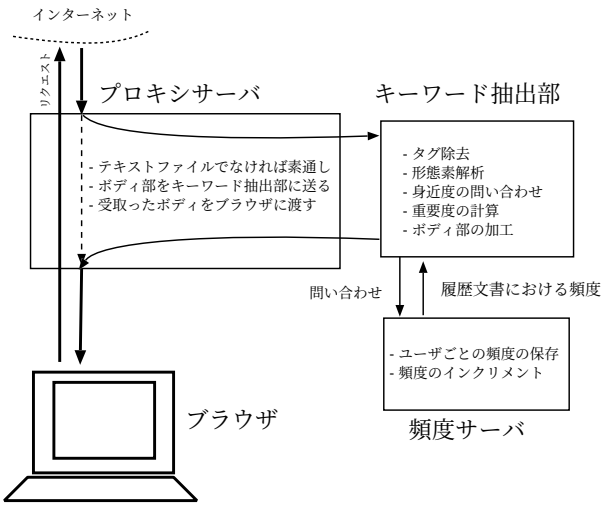

図 3 システム構成

表 9 被験者 2 の身近度

順位	単語	身近度
1	ゲーム	34
2	ページ	24
3	found	20
4	情報	20
5	server	19
6	関連	19
…	…	…
15	更新	17
15	掲示板	17
15	登場	17
15	イメージ	17
…	…	…
57	ネット	9
57	copy	9
57	シリーズ	9
57	ソフト	9
57	game	9

(5) 返ってきたボディ部分とヘッダ部分を合わせて，ブラウザに送る．

なお，終了となった場合には受け取った応答をそのままブラウザに送る．

次に，キーワード抽出モジュールは，送られてきたテキストに対し，次のような処理を行う．

(1) HTML タグを除去し，プレーンテキストにする．
(2) 形態素解析を行い，名詞，動詞，形容詞，未知語だけを取り出す．ストップワードに指定されている語は除去する．
(3) 各語について，身近度を頻度サーバに問い合わせ，身近度の高いものを一定数（上位 30%）選ぶ．
(4) 式 (2) により，各語の重要度を算出する．
(5) もとの HTML のボディ部分に対して，キーワードを赤色，過去の頻度の高い語は青色でハイライトする．具体的には，“” と “” のタグを挿入する．これをプロキシサーバに返す．

頻度サーバでは，各ユーザごとに語の頻度を保持しておく．各語が出現した最終日時も保存しておき，語数の上限を越えた場合には，出現した最終日時が最も早い語をデータベースから削除する．

ユーザは普段のブラウジング操作と同じようにブラウジングを行うことができる．閲覧中のページのところどころ文字が赤や青になって強調される．本システムのスクリーンショットを図 4 に示す．(キーワードは*印が右肩についた語，身近度の高い語は+印が右肩についた語である．)

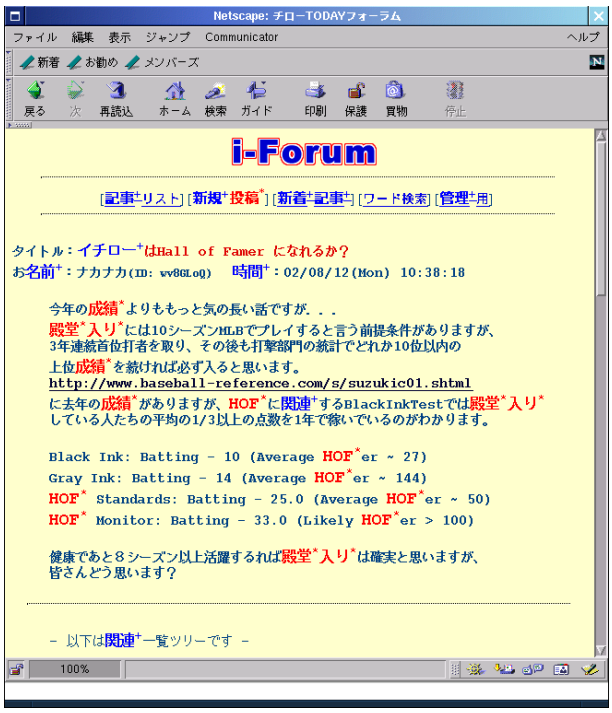

図 4 スクリーンショット

6. 評価実験と考察

Web ページは，ページの内容や記述の仕方，長さなど非常に多様である．多様性が高いために，新聞記事や論文を対象とした語の重み付けと比べ，正しくキーワードを抽出すること，そして適切な評価を行うことが難しい．特に，本システムは，ユーザ個人の興味に近い語を提示することを狙ったシステムであり，客観的な指標で評価することは困難である．

そこで，ブラウジング支援システムをユーザに利用してもらい，利用後にユーザにアンケート調査を行うことで評価を行うことにした．ハイライトする語を決めるアルゴリズムとして以下の 3 つを用い比較を行う．

- システム I：閲覧中のページにおいて CF 法によりキーワードを抽出する．閲覧中のページにおける頻出語を青，χ^2 値の高い語を赤でハイライトするもの．
- システム II：今まで閲覧したページをコーパスとして用い，身近度が高い語を青，TF-IDF 値の高い語を赤でハイライトするもの*9．
- **システム III(本手法)**：今まで閲覧したページ（履歴文書）を用い*10，身近度が高い語を青，IRM 値の高い語を赤でハイライトするもの．

ブラウジング時に，全く異なるトピックのページを閲覧する場合には，システム I と III は同様の処理となる．しかしここでは，ユーザは自分の興味あるトピックに関して選択的に複数ページ読むと仮定している．また，短時間の実験であるため，実験中に閲覧した文書全てを履歴文書として用いた．

評価実験は大学生・大学院生（20 代男性）10 人に対して行った．システム I → II → III の順に使用する被験者群と，システム I → III → II の順に使用する被験者群を同数とし，それぞれのシステムを 30 分程度ずつ，特に課題は定めずに自由にブラウジングをしてもらった．システム I の利用中にも履歴情報が蓄積される．被験者には，赤や青でハイライトされる語がどのように選ばれているのかは知らせず，異なるアルゴリズムで抽出しているとだけ説明した．各システムを試用したあとに，次の質問に 1（全くあてはまらない）〜5（よくあてはまる）の 5 段階で答えてもらった．

- Q1: ブラウジングのしやすさ
- Q2: 赤で表示されている語は興味のある語ですか
- Q3: 興味のある語が赤で表示されていますか
- Q4: 青で表示されている語は興味のある語ですか
- Q5: 興味のある語が青で表示されていますか

また，すべてのシステムを試した後，

- Q6: どのシステムが一番ブラウジングしやすいか
- Q7: どのシステムが一番自分の興味を反映しているか

という質問も行った．また，赤・青で表示される語に対する感想，気づいたことなどを自由記述で書いてもらった．

結果を集計すると表 10，表 11 のようになった．Q1 のブラウジングのしやすさに関しては各システム 3 前後の値であり，それほど大きな差は出ていない．Q2 に関しては，システム I よりも II や III の方がよい評価となっているが，II と III には大きな差はない．Q3 の，興味がある語が赤で表示されているかという質問に対しては，システム III が最もよく，ついでシステム II，システム I という結果となった．Q4，Q5 は青でハイライトされた身近語に関する質問であるが，どのシステムも大きな違いはない．

Q3 に関してはシステム III の優位性を示すことができたものの，Q2 に関してはシステム II と同程度の評価であった*11．しかし，各システムを用いる際に，被験者が閲覧したページ群は同一ではない．そのため，閲覧していたページによっては，キーワードを出しやすいものとそうでないものがあり，システムごとの評価に差が出ることも考えられる．一方，赤でハイライトされるキーワードと青でハイライトされる語の評価は，同一のページ群に対して行われており，被験者は両方の評価を同時に行うので，その相対的な差が重要である．特に，シス

*9 語 w に対する idf の重みづけは $\log(D/df(w))+1$ とした．ただし D は全文書数，$df(w)$ は語 w が出現する文書数である．

*10 なお，システム II では，語の文書頻度（DF）を出すために閲覧文書を用いており，III では閲覧履歴における出現頻度を出すために履歴文書を用いている．

*11 平均の差を検定する t 検定では，5%の有意水準では，Q2 に関してシステム I が他の 2 つより悪いこと，Q3 に関してシステム II がシステム I よりよく，さらにシステム III がシステム II よりよいことだけが示せた．

表 10 平均得点

	Q1	Q2	Q3	Q4	Q5
(I) システム I	2.8	3.2	2.9	2.7	2.7
(II) システム II	3.2	4.0	3.3	2.5	2.5
(III) **システム III**	3.2	4.1	3.8	2.0	2.4

表 11 投票結果

	Q6	Q7
(I) システム I	1	0
(II) システム II	3	2
(III) **システム III**	6	8

表 12 Q2 と Q4, Q3 と Q5 の相対的な比較

	Q2–Q4	Q3–Q5
(I) システム I（参考）	0.5	0.2
(II) システム II	1.5	0.8
(III) **システム III**	2.1	1.4

テム II と III の青でハイライトされる語は，同じ方法 (身近度の高いもの) で抽出されているので，青の評価を基準とした差の比較が可能である．各システムに対して，Q2 と Q4，Q3 と Q5 のポイントの差を計算したものを表 12 に示す．いずれの項目に関しても，システム III が最もよく，次いでシステム II がよいという評価となっている．

Q6，Q7 は，トータルの印象を聞いたものであるが，一番ブラウジングしやすいものという問いには，システム II もしくは III と答える被験者が多かった．ユーザの興味を一番反映しているものは，システム III であった．したがって，本システムがユーザの興味に関連した語をハイライトするという狙い通りの結果が得られていると言えるであろう．

自由記述による被験者の意見をまとめると，「システム III はなかなか興味にあっている」「ニュースサイトのような記事には使いやすい」という意見がある一方で，「提示意味がよく分からない語もある」「個別の語が興味があるかどうかと言われても判断しづらい」という指摘もあった．逆に，「ハイライトされた語によって興味を喚起された」という意見もあった．

7. 議　　論

一般にユーザの興味は，例えば研究，スポーツ，旅行など，複数ある．本手法では，ブラウジング中のページにおける身近度の上位語を身近語として選ぶ．したがって，スポーツのページに頻出する語が研究のページには出現しなければ，身近語には選ばれずキーワード抽出に悪影響は与えない．

しかし，複数の興味にまたがって同じ語が出現する場合もある．これが，複数の分野に共通するユーザの興味を表している場合には問題ないが，一般語（例えば「思う」「日本」など）が身近語として選ばれる場合や，同じ語が異なる意味で用いられている場合には問題がある．前者に対しては，身近度として，各語の履歴文書中の文書頻度ではなく，総出現回数を用いることで対応している．ページのトピックとなる語はその文書内に何度も出現するので，頻度を考慮することで相対的に一般語の身近度を下げることができる*12．後者は難しい課題であるが，シソーラスなどを用いて概念レベルでの処理を行う [Magnini 01] などの方法がある．

Web ページは文書の長さがまちまちである．文書がある程度長ければ，その文書中の頻出語はその文書のトピックをよく表すことになり，頻出語との共起の偏る CF 法は有効である．しかし，長さが短いページでは，例えば，最も頻出する語でも 3 回しか出現せず，しかも同順位のものがたくさんあるという状況も発生する．こういった場合に，ユーザにとって身近な語を共起の対象と選ぶ IRM は有効な手法であると考えている．

前節の評価実験では，課題は定めず，ブラウジングの仕方も統制を行っていない．これは，普段のブラウジング行動の中で，本システムがどのくらい役に立つかを測りたかったためである．(読む文書を統制した評価実験は，4 節で行った.) ユーザの興味には大きく分けて，長く続く長期的な興味とその場限りの短期的な興味があるとされている [Chen 01] が，長期的な興味の変化に対する特性を明らかにすることは今後の課題である．その際，ユーザが閲覧した文書は時系列的なコーパスであることを考慮して，以前に出現した語は重みを下げ，最近出現した語は重みを上げるなどの処理が必要となるだろう．また，ユーザが興味あると評価したページだけ身近語の計算に用いる関連性フィードバックを用いれば，身近語をより正確に取得することができると考えられる．

8. む　す　び

本論文では，文書におけるユーザの興味を考慮したキーワード抽出と，キーワードのハイライトによるブラウジング支援システムについて述べた．

ユーザが今まで多く見た語自体はユーザにとってそれほど興味がなく，それと関連のある語が興味を引く可能性が高い．現在，ユーザの興味を把握し，情報を提示する研究が盛んであるが，その多くが「ユーザの興味と一致する」情報を提示しているように思う．ユーザの興味を踏まえた上で，どのくらい新しい情報，ユーザの興味と離れた情報を提示すればよいかという議論が必要ではないだろうか．

謝　　辞

実験に御協力いただいた方々に感謝いたします．また，査読者の方には非常に有益なコメントを頂きました．ありがとうございました．

*12 さらに改善するには，大規模な文書コーパスから語の生起確率をあらかじめ求めておくなどの処理が必要となるだろう．

◇ 参 考 文 献 ◇

[Chen 98] Chen, L. and Sycara, K.: WebMate: A personal agent for browsing and searching, in *Proc. 2nd International Conference on Autonomous Agents (Agents '98)* (1998).

[Chen 01] Chen, C. C., Chen, M. C., and Sun, Y.: A Web Document Personalization User Model and System, in *Proc. 8th International Conference on User Modelling (UM 2001)* (2001).

[Forey 02] Forey, T. A.: One to One マーケティングを超えた戦略的 Web パーソナライゼーション, 日経 BP 社 (2002).

[Joachims 97] Joachims, T., Freitag, D., and Mitchell, T.: WebWatcher: A tour guide for the world wide web, in *Proc. 15th International Joint Conference on Artificial Intelligence (IJCAI-97)*, pp. 770–775 (1997).

[Lieberman 95] Lieberman, H.: Letizia: An Agent that assists Web browsing, in *Proc. 14th International Joint Conference on Artificial Intelligence (IJCAI-95)*, pp. 924–929 (1995).

[Magnini 01] Magnini, B. and Strapparava, C.: Improving User Modelling with Content-Based Techniques, in *Proc. 8th International Conference on User Modelling (UM 2001)* (2001).

[松尾 02] 松尾, 石塚：語の共起の統計情報に基づく文書からのキーワード抽出アルゴリズム, 人工知能学会誌, Vol. 17, No. 3, pp. 217–223 (2002).

[Mobasher 01] Mobasher, B., Berendt, B., and Spiliopoulou, M.: KDD for Personalization (Tutorial), in *Proc. 5th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD-2001)* (2001).

[中島 01] 中島, 橋田, 森, 伊東, 本村, 車谷, 山本, 和泉, 野田：情報インフラに基づくグラウンディングとその応用 – サイバーアシストプロジェクトの概要 –, コンピュータソフトウェア, Vol. 18, No. 4, pp. 48–56 (2001).

[Pazzani 96] Pazzani, M., Muramatsu, J., and Billsus, D.: Syskill & Webert: Identifying Interesting Web Sites, in *Proc. AAAI-96*, pp. 54–61 (1996).

[Salton 89] Salton, G.: *Automatic Text Processing*, Addison-Wesley, MA. (1989).

[Suryanarayana 02] Suryanarayana, L. and Hjelm, J.: Profiles for the situated Web, in *Proc. of WWW2002* (2002).

[Turing 50] Turing, A. M.: Computing machinery and intelligence, *Mind*, Vol. 59, pp. 433–450 (1950).

[Widyantoro 99] Widyantoro, D. H., Ioerger, T. R., and Yen, J.: An Adaptive Algorithm for Learning Changes in User Interests, in *Proc. of 8th International Conference on Information and Knowledge Management (CIKM'99)* (1999).

〔担当委員：栗原　聡〕

2002 年 8 月 19 日　受理

著 者 紹 介

松尾　豊（正会員）

1997 年東京大学工学部電子情報工学科卒業．2002 年同大学院博士課程修了．博士（工学）．同年より，産業技術総合研究所サイバーアシスト研究センター勤務．仮説推論，数理計画法，キーワード抽出，Web マイニング，ユーザモデリング等に興味がある．受け手にとって価値の高い情報の提示を目指している．情報処理学会，AAAI の各会員．

福田　隼人

2002 年東京大学工学部電子情報工学科卒業．同年 4 月東日本旅客鉄道株式会社入社．7 月より東京電気工事事務所所属．現在に至る．最近は鉄道信号・列車運行管理システムの論理的設計に従事．データマイニング技術を応用した顧客情報収集・分析による鉄道・関連事業との有機的連携，及び新サービスへの展開について興味がある．

石塚　満（正会員）

1971 年東京大学工学部電子卒業．1976 年同大学院博士課程修了．工学博士．同年 NTT 入社，横須賀研究所．1978 年東京大学生産技術研究所助教授，同教授を経て，1992 年工学部電子情報工学科教授．2001 年より情報理工学研究科電子情報学専攻．研究分野は人工知能，知識処理，マルチモーダル擬人化エージェント，ネットワーク化知的情報環境．IEEE，AAAI，情報処理学会，電子情報通信学会，映像情報メディア学会，画像電子学会等の会員．