プロキシログ解析に基づくトップページの抽出と検索

成 凱[†]　平野 真太郎[‡]　上林 弥彦[†]

†京都大学情報学研究科社会情報学専攻 〒606-8501 京都市左京区吉田本町
‡京都大学工学部情報学科 〒606-8501 京都市左京区吉田本町
E-mail: †{chengk,yahiko}@db.soc.i.kyoto-u.ac.jp, ‡shin@db.soc.i.kyoto-u.ac.jp

あらまし Web ブラウザを立ち上げた時初めに現れるページは「トップページ」もしくは「ホームページ」と呼ばれ，利用者の興味を確実に反映したコンテンツであり重要である．また，このページは重要なページへのリンクを含んでいることが期待される．しかし，トップページは各利用者によって異なり共有することはできない．本研究では大規模 ISP のプロキシログの利用状況からトップページの抽出を行い，さらにトップページを対象とした検索手法を開発した．プロキシログ解析より得られる利用状況をコンテンツの重要度の計算に反映させることで今までにない検索が可能となった．これによりプロキシログという実データが示す利用者集合の特性を反映した検索の実現と情報の共有が可能となると考えられる． 利用者の集合は地域性などの特性を持つものと考えられ，本研究では特定の地域にいる利用者を例に示す．今までの検索エンジンでは扱うのが困難であった興味の時間的推移といった問題も扱うことができると考えられる．実験には冬季のデータを用いたため，冬季のコンテンツが検索結果の上位に現れることが確認でき，この方法が従来用いられてきた静的な順位付けより優れているとう結果が示せた．

キーワード 情報検索，Web とインターネット，データウェアハウス，パーソナライゼーション

Extraction and Retrieval of Top Pages By Analyzing of Proxy Log

Kai CHENG[†]　Shintaro HIRANO[‡]　and　Yahiko KAMBAYASHI[†]

†Department of Social Informatics, Kyoto University
‡Department of Information Science, Kyoto University
Yoshidahonmachi, Sakyo-ku, Kyoto-shi, 606-8501 Japan

E-mail: †{chengk,yahiko}@db.soc.i.kyoto-u.ac.jp, ‡shin@db.soc.i.kyoto-u.ac.jp

Abstract The page that appears first when a web browser is started is called the “Top Page” or “Homepage”. Top Page reflects user’s interests certainly and is expected to include links to important pages. Top Page of one user can be useful for other users. But no methods for sharing them have been proposed. In this study, we extract Top Pages from a proxy access log and develop a new search engine which treats them. By applying user behavior captured by analysis of the proxy access log to a page-scoring algorithm of a search engine, the search engine can satisfies users’ needs which will change according to user location and season, for example.

Keyword Information Retrieval , Web and Internet , Data Warehouse , Personalization

1. はじめに

インターネットに接続してブラウザを立ち上げた時に，利用者が初めに見るページは「トップページ」もしくは「ホームページ」,「開始ページ」と呼ばれる．トップページは利用者が Web ブラウザに設定しているページや，友人からの推薦によってアクセスされたページなどがあると考えられる．トップページは利用者が苦労して集めたよいリンクの集合や，利用者の興味や仕事と最も合った質の良い情報が含まれるサイトである可能性が高く，利用者の興味を確実に反映したコンテンツであることが予想される．しかし，利用者のトップページを調べることは困難であり，利用者同士が共有することができない．

このようにトップページは重要であるにもかかわらず，これまでトップページの検出，共有を実現する手法は提案されてこなかった．本研究では利用者のアクセス履歴を用いたトップページの抽出方法を提案し，それらを共有するための検索サービスについて検討した．さらに利用者がトップページの次にアクセスするページ(セカンドページ)がトップページのリンクを辿ったものであるかを調べることによって，質の良いトップページを集めることも可能である．トップページからのリンクを用いていない場合はアクセスされたページは利用者がブ

ックマークの中から選択したページと考えることができる．これにより質の良いトップページが分かると共に，ブックマークという利用者にとって重要なページを知ることが可能である．

コンテンツのリンク構造用いて，ハブとオーソリティーの関係から重要なページの検出を行う方法[4]があるが，このような静的な構造を用いるより，実際に利用者がどのように使ったかという情報から得られるトップページのほうが重要なページである可能性が高い．

検索システムにおいて各利用者の興味に関する情報を用いてその個人向きの検索精度をあげる手法も使われている．例えばある利用者のグループがどのような内容に興味を持っているかを知りたい場合は各個人の興味情報を集める必要がある．これにはプライバシーの問題があり困難である．あるグループの利用者に対するトップページ集合はその利用者集合が持つ興味を反映しているといってよい．これは個人を特定しなくても良いためプライバシーの問題はない．本稿ではこのような目的のために，トップページの検出方法とそれを利用した検索について述べている．実際のデータは，京都市のインターネットサービスプロバイダーから提供されたもので，その履歴データから求めたトップページ情報を用いてページの順位付けを行いその有用性を確認した．個人情報を用いて検索をさらに個人向けにすることも可能であるが，その部分には新規性がないのでここでは検討していない．

Web 上での利用者の活動を記録した物としてプロキシログがある．そのログ上では利用者は IP アドレスによって表現されている．このプロキシログを用いてトップページリストの抽出，リストを基にした実データの取得そして共有を試みようとした時，個別の利用者のアクセス履歴を入手することが前提である．しかし，IP アドレスからでは正確な入手は困難である．本研究では，どの利用者がどの IP アドレスをいつ利用しているかを記すラディウスログを利用することでこの問題を解決した．利用価値のあるこれらのログ，特にラディウスログの利用はプライバシーの問題に抵触する可能性があるため実現が難しく，これらの情報を利用した研究は貴重である．利用者特定したアクセス履歴から抽出したトップページの実データの取得時にはノイズやフレーム処理，URL 転送，ページ更新の問題に配慮した．

現在一般に普及している検索システムでは利用者は検索するクエリーを提出するだけで地域や年齢などの利用者に関する情報の提示はしない．多くの検索システムはコンテンツのクエリーに対する重要度の計算をする際に，コンテンツ中に含まれるキーワードやリンク構造のみを用いて計算しているので利用者に返される結果は，検索空間やその中のコンテンツの変化がない限り，常に同じものになる．利用者の特性や利用履歴を各利用者の個人サイトが記憶している場合，それを用いて質問修正と結果のフィルタリングをすることによって検索を各個人向きにすることは可能である．利用者の特性の例には地域性や検索される時期(季節)がある．携帯端末における利用者の位置を考慮する手法として[7]がある．また，EC(電子商取引)サイト[8]では利用者の履歴を利用している．しかし，先に述べた理由でグループ対応の情報の反映は不可能であった．

本研究では ISP の保持するプロキシログを利用することで，個別の利用者ではなく ISP の会員という利用者集合のトップページを抽出すると共に，それらを利用した地域性や検索する時期といった特性を考慮した検索システム[UTPS(Usage-aware Top Page Search System)]を開発した．これまでの検索エンジンでは扱うのが困難であった興味の時間的推移の問題に対処できると考えられ，位置情報や検索履歴が不十分な利用者にも対処することができると考えられる．

利用者の特性として，地域性のほかに家族の状況や財政状況などもインターネットの利用状況に影響があると考えられる．UTPS は ISP の新サービスとしての利用が期待できる．

以下 2 章ではトップページの抽出について述べ，3 章では利用状況を反映したトップページの検索，4 章では開発した UTPS の実装及び評価，5 章では関連研究，最後に 6 章で本研究の結論と将来研究について触れる．

2. プロキシログによるトップページの抽出

インターネットに接続してブラウザを立ち上げた時に，利用者が初めに見るページは「トップページ」もしくは「ホームページ」,「開始ページ」と呼ばれる．ここで扱うトップページとして次のものが考えられる．

1. **利用者が Web ブラウザにトップページとして設定したもの**
   普段最初に利用されることから，トップページには利用者の Web 活動の拠点であるページが設定されていると考えられる．Web ブラウザにおいて利用者が興味のあるページの URL を保存しておくことで，アクセスを簡単にするツールとして bookmark などがあるが，トップページはその bookmark の中で利用者にとって最も重要なページであると考えられる．
2. **他者の推薦や自分の bookmark によりアクセスしたページ**
   Web ブラウザを立ち上げていない状態で，友人からのメールやメーリングリストに含まれるページの紹介を参考にしてアクセスしたページや bookmark の中から利用者が選んでアクセスしたページ．

これらのトップページは利用者の Web 活動の起点であり，利用者の興味に基づいてアクセスされたページであ

ると考えられ，利用者の興味を確実に反映したコンテンツであると言える．有名サイトが多いと考えられる 1 型のトップページに対して，2 型のトップページは特に利用者の興味や個性を現していることが期待できる．これらのトップページを利用者同士で共有することができれば，利用者の新しい知識(Web ページ)の入手が可能になることが期待できる．

トップページの抽出方法として，プロキシログを利用してトップページを抽出する方法を提案する．まずプロキシログについて説明する．プロキシログは利用者の Web 上での活動を時間順に保持したものであり，主要な情報のみを記したテーブルは以下のとおりである．

- **Proxy(Time , IP , Size , URL)**

Time はリクエストされた時刻，IP は URL リクエストした IP アドレス，URL，Size はそれぞれリクエストされた URL とそのデータのサイズである．しかしこれらの情報だけでは，IP アドレスがどの利用者なのか特定することはできない．そこで，本手法では利用者特定を可能にするためにプロキシログに加えてラディウスログを利用する．

2.1. ラディウスログを用いた利用者の特定

ラディウスログは, 誰がどの期間にどの IP アドレスを利用しているかを記録するアクセス履歴である．ラディウスログのテーブルは以下の通りである．

- **Radius(U_ID , IP , Login , Logout , TEL_NO)**

U_ID は利用者 ID , IP は利用 IP アドレス, Login , Logout , TEL_NO は順に接続開始時刻，接続終了時刻，接続元電話番号を表している．これらの情報を用いたプロキシログ上における利用者特定は図 1 に示すとおりに行われる．プロキシログとラディウスログの IP アドレスが一致し，リクエストした時刻が接続開始時刻と接続終了時刻の間にあることが結合条件である．

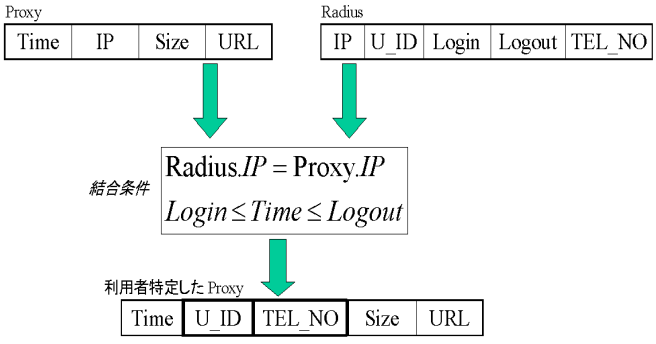

図 1.ラディウスログを用いた利用者特定

プロキシログとラディウスログをこの方法によって結合し利用者を特定することによって，個々の利用者の行動を追うことが可能となる．

2.2. 利用者別のアクセス履歴の作成

利用者を特定した新しいプロキシログを利用者 ID ごとに記録することによって利用者別のアクセス履歴を生成することができる．生成した利用者別のアクセス履歴からトップページの抽出を行う．利用者を特定したアクセス履歴は時刻に関して昇順に並んでいる．利用者 ID ごとにアクセス履歴を割り振れば利用者別のアクセス履歴が生成できる．本研究ではこれらのデータを京都市の運営する大規模 ISP，Kyoto I-net から提供されたものを利用したが利用者のプライバシーを尊重し，利用者名が特定できないようにデータを変換して用いている．

2.3. トップページの抽出

利用者別のアクセス履歴を用いてトップページの URL を抽出し，実データを取得する方法について述べる．先ほど利用したラディウスログの Login(接続開始時刻)の情報を利用しトップページを抽出する．Login よりも後で，なおかつもっとも近い時刻にリクエストされた URL が本研究での定義におけるトップページである．トップページ決定イメージは次の図 3 の通りである．2 番目にアクセスされたページ(セカンドページ)が，トップページのリンクを辿ったものであるかを調べることで質の良いトップページが分かると共に，ブックマークという利用者にとって重要なページを知ることも可能である．

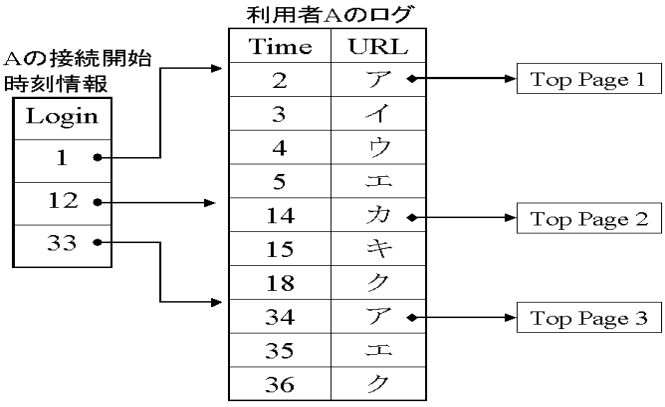

図 3.ある利用者Ａのトップページの抽出イメージ

このようにして決定したトップページのリストを作り，それを基にトップページの実データを取得する．実データとは Web コンテンツの HTML ソースを意味する．実データの取得時における問題には，現在は削除されてしまっている可能性，残っていてもエラー表示の内容しか含まない可能性，フレーム情報のみでコンテンツの実部分が含まれない可能性，内容はなく転送コードが記載されている可能性がある．これらの問題をデータのサイズや，取得した HTML ソースの内容解析によって解決し実データを収集した．実データの収集に関しては 4.1.2 節にて詳しく述べる．

3. 利用状況を反映したトップページの検索

多くの人がトップページとして利用しているページは当然として，少数の人がトップページとしているページの中には，他の利用者にとって重要なものが含まれている可能性がある．本研究では UTPS を開発することで利用者の興味，情報の共有を実現した．本研究ではトップページを共有する方法として検索を用いた．トップページのコンテンツを考慮せずに利用頻度順に並べることで，利用者に推薦方法がある．しかし，この方法では利用者の興味に合ったトップページを探すことは難しい．検索を用いることで利用者が自分の興味をキーワードで表すことができるので，自分の興味に関連したトップページを知ることが可能となる．

さらにプロキシログから分かる利用者集合の特性をコンテンツの重要度の計算に反映することで，地域性や検索する時期を反映した検索が可能となる．ここでは提案するトップページ検索の手法について述べる．

3.1. コンテンツの関連性による検索

既存の検索手法としてコンテンツの関連性を用いるものがある．関連性としてコンテンツに含まれるキーワードの出現状況を計算に用いる手法, TF/IDF 法[1]がある．キーワードの出現頻度をベクトル化し，提出されたクエリーのベクトルとより類似しているベクトルを含むコンテンツが利用者の望むものであると判断するベクトル空間モデルがある．単語の重み付けの方法である TF/IDF 法は，コンテンツを特徴付けるキーワードの性質として次の二つの要素を用いる．

1. Term Frequency

   コンテンツ中に良く含まれるキーワードはそのコンテンツを良く特徴付ける．式は以下のように定義する．

   $tf(t,d)$ = コンテンツ d における
   キーワード t の頻度

2. Inverse Document Frequency

   キーワードは出現するコンテンツの数が少ないほどその出現元のコンテンツを特徴付ける．キーワード t の重要度を次の式で定義する．

$$idf(t) = \log\left(\frac{DBsize}{freq(t, DB)}\right) + 1$$

   DB は対象としたコンテンツ集を保存したデータベース，$freq(t,DB)$はキーワード t のデータベース DB に現れる頻度，$DBsize$ は DB のサイズである．

これらの二つの値を利用することでコンテンツ d におけるキーワード t の重要度 $Weight(t, d)$を次のように計算することができる．

$$Weight(t,d) = tf(t,d) * idf(t) \quad \text{<式 1>}$$

検索の対象となるコンテンツ集を TF/IDF 法を用いてベクトル空間モデルで表現する．検索されたキーワードに対して重みの大きいコンテンツを検索結果として返すことで検索を実現している．

しかし利用者の住んでいる地域や検索する時期といった利用者の特性によって検索に求める情報は異なることがあると考えられる．ゆえに検索におけるコンテンツの重要度の計算にはコンテンツの関連性だけでなく地域性や検索する時期を考慮する必要がある．

3.2. プロキシログに反映した地域, 時期の特性

コンテンツの重要度の計算において地域性や検索する時期といった利用者の特性を考慮し，検索に利用することができれば，より利用者の嗜好を反映した検索が可能となる．

プロキシログは利用者の Web 上での活動を記録している．故に，そのログには利用者の興味が反映していると考えられる．これらは動的に変化するため，プロキシログから利用者の興味や特性を抽出することができればその利用価値は大きい．プロキシログからコンテンツの重要度の計算に利用できる利用状況として *frequency* と *recency* の二つの概念を提案する．

- ***frequency***
  長期的によく利用されるコンテンツは重要である
- ***recency***
  最後に利用された日時が新しいほどそのコンテンツは重要である

これらの要素は利用者の提出するキーワードには依存せず，プロキシログという利用履歴によって値が決まる．そのため，利用するプロキシログを変えること，つまり ISP を変えることで，同じコンテンツであってもこれらは違う値になる．これは，ISP を一つのコミュニティーと見た場合，コミュニティーごとのパーソナライゼーションであるといえる．要素にコミュニティーの地域性などの特性が反映されることが期待できる．この *frequency* と *recency* を考慮することで利用者集合の特性を反映した検索が可能になると考えられる．利用者集合を特定の性質を持つ利用者に制限してその集合の特性を求めることもできる．次にこれらの要素をコンテンツの重要度の計算に用いる方法を次に述べる．

3.3. 利用履歴情報を反映した検索

前節で述べた利用状況を TF/IDF 法に加味することで新しいコンテンツの重要度の計算方法を提案する．コンテンツ d の利用頻度を $c(d)$と表記する．$c(d)$の集計の方法として，どの時期の利用頻度を重視するかを決めるこ

とができる．長期的な利用頻度を重視する方法として，利用するアクセス履歴に出現する回数を集計する方法(ALL_COUNT)や，最近の利用頻度を重視して集計する方法(λ-aging[9]，Sliding Window)がある．コンテンツ *d* の *frequency* を *f(d)*とし，次式により定義した．

$$f(d)=\begin{cases}0.3 & c(d)\le \dfrac{m}{1.7}\\ 2-\dfrac{m}{c(d)} & c(d)>\dfrac{m}{1.7}\end{cases}$$

m は全てのコンテンツのアクセス頻度の中央値もしくは平均値である．この式は重みである *f(d)*の範囲を指定していて，最大で 2 とし，ある一定の頻度(中央値)に満たないコンテンツは全て同じ重み(0.3)として扱うことを意味している．

またコンテンツ *d* の *recency* を *g(d)*とし次式により定義する．

$$g(d)=\begin{cases}1 & time\ (d)\le 24\ (1day)\\ 1.5 & 24<time\ (d)\le 168\ (1week)\\ 2 & 168<time\ (d)\le 336\ (2weeks)\\ 3 & 336<time\ (d)\le 504\ (3weeks)\\ 4 & 504<time\ (d)\le 672\ (4weeks)\\ \vdots & \end{cases}$$

コンテンツ *d* が最後にアクセスされた時刻と検索を行う日の午前 0 時までの時間の差(単位 *h*)を *time(d)*とする．*d* が最後にアクセスされた時間が 1 日前，1 週間前，2 週間前と増えていくにつれて *g(d)*の値は大きくなる．最近利用されたものほど値が小さくなる．

これらを<式 1>に考慮したコンテンツ *d* におけるキーワード *t* の重要度 *NewWeight(t,d)*の計算式は次式により定義される．

$$NewWeight\ (t,d)=tf\,(t,d)*idf\,(d)*e^{\frac{f(d)}{g(d)}}$$

先に述べた *frequency*，*recency* の範囲は関連度によるコンテンツの重要度 *tf(t,d)*idf(t)*との兼ね合いによるものである．この計算式においてキーワードとコンテンツに依存する変数は *tf(t,d)*と *idf(t)*であり，*f(d)*と *g(d)*はコンテンツのみに依存する変数である．*frequency* が大きくなればなるほど重要度は大きくなり，*recency* が小さければ小さいほど重要度は大きくなる．コンテンツの重要度の計算をこの式で行うことによってプロキシログの示す利用者集合の地域性や検索する時期などの特性を反映した順位付けが可能となると考えられる．

4. システムの実装及び評価

4.1. 大規模 ISP Kyoto I-net の利用データ

4.1.1. 利用データのプロフィール

本研究では京都市の ASTEM(京都高度技術研究所)の運営する ISP，Kyoto I-net のプロキシログとラディウスログを利用して実験を行った．会員数は 2 万人以上で，解析に利用したデータは 02/01/15 から 02/02/14 の 1 ヶ月分である．27 台のプロキシサーバが稼動しており，各々が独立にプロキシログを記録している．実験に利用したレコード総数は 116,839,901 にも及ぶ．

プロキシログに現れるファイルタイプが Image であるものを除いた URL アドレスを出現回数によって並べると，Yahoo!や Google のような有名なサイトの他に My Yahoo!や MSN の hotmail などのサービスで普及している"マイページ"と呼ばれる個人の Web ページや，オークションなど，コンテンツの内容の変化が激しいと思われる Web ページが上位を占めていることが分かった．

プロキシログとラディウスログの情報を結合することによって利用者を特定したアクセス履歴では，今までできなかった解析が可能となる．Radius のメンバ TEL によって利用者がダイアルアップ接続を利用しているか，常時接続を利用しているかを調べることができる．これにより利用者の接続速度を知ることができる．図 4 は左から順に**ナローバンドユーザー，ブロードバンドユーザー，LAN ユーザー別の利用者数**(Sum of Users)，利用者が 1 日にどれだけ HTML タイプのファイルのリクエストをしているかを示す**１人当たりの１日の平均 HTML リクエスト数**(Ave Daily REQUESTS)，リクエストした HTML の中でユニークな HTML の数をしめす**１人当たりの１日の平均利用ページ数**(Ave Daily Pages)を表している．

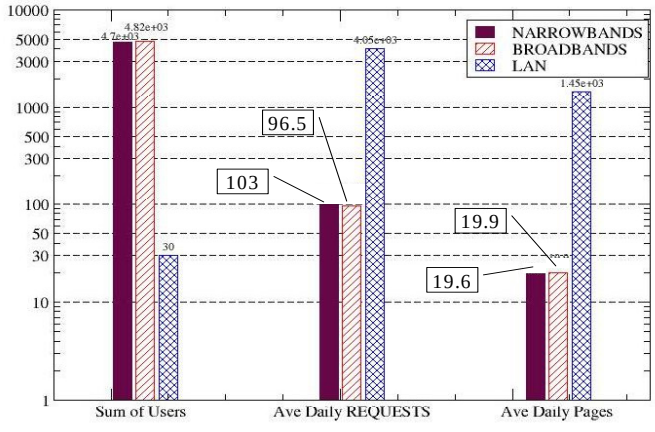

図 4.利用者のタイプ別 Web 利用状況

ナローバンドユーザー4703 人，ブロードバンドユーザー4820 人，LAN ユーザー30 人で合計 9553 人であるが，図 4 よりナローバンドユーザーとブロードバンドユーザーとの Web 利用状況の差異は殆どないことが分かる．１利用者当たりの１日の平均リクエスト数と 1 日の平均利用ページ数を見てみるとナローバンドユーザー，ブロードバンドユーザーと共にリクエストの内の約 8 割がある一

定の Web ページに対して繰り返し行われていることが分かる．

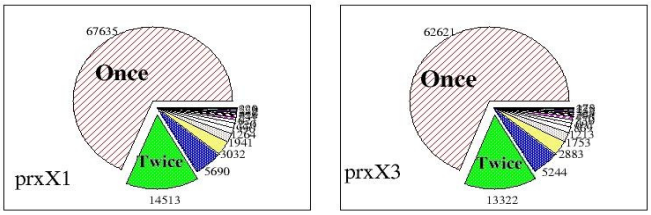

図 5.Web ページ再利用状況

図 5 では利用したプロキシサーバのうち，2 つのサーバのアクセス履歴 1 ヵ月における Web ページの利用状況を表している．アクセス履歴に表れた，ファイルタイプが HTML である URL を利用回数によって分類したものである．約 7 割の Web ページが 1 回(Once)しか利用されていないことが分かる．

図 4，図 5 から利用者が一部の Web ページをよく繰り返し見ていること，Web ページ全体の中で数回しか見られることのないページが多いことから，一部の Web ページが多くの利用者に，繰り返しよく利用されていることが分かる．

これらのことらから，プロキシログを利用したサービスとして Web に広く分散しているコンテンツを収集し利用者に提供しようとした場合，提供する中身となるページのサンプリングが,非常に大事であることが分かる．

純粋に利用頻度の大きいコンテンツを利用者に提供すれば，オークションなどの内容の変化が激しく，すぐ無くなってしまう可能性が高いコンテンツが多く含まれる可能性が高い．また既に利用者が知っているページを提供する可能性があるため，良いとは考えられない．この問題に対処するために，再利用率の低いコンテンツの中から重要なコンテンツを発掘することが必要となる．本研究ではトップページという利用者の強い興味に基づき利用されたページを対象とすることで，データの質の信頼性を保ち，アクセス数の少ないトップページも検索を用いることで利用者に提供することでこの問題に対処した．

4.1.2. トップページの抽出

2 章で述べた手法によって集めたトップページの中で，ファイルタイプが HTML であるものは 32,938 個であった．これらの URL を基に 2002/12/19 に，実際に HTML ソースを取得したものが共有(検索)の対象である．取得した HTML ソースは以下の条件を満たす．

1. 取得時のステータスコードが 200 番台であること
2. ファイルサイズが 1KB を超えること
3. HTML ソースがフレームの時，本体の中身の取得
4. 別の URL への自動転送が行われる時，転送先の中身の取得

ページが実在(残存)することが前提である．ファイルサイズが小さいものは文書量が少なく，内容のないページであることが多いため除く必要がある．これによりエラー表示のみを内容とする HTML ソースも大部分除くことができる．3 と 4 の条件によって HTML ソースを取得する時も，1 と 2 の条件を満たすことが必要である．

32,938 のトップページの中から，ランダムに選び出したページの中で上記の条件を満たしたトップページは 4,192 ページである．これらのページから得られたキーワードの数は 503,843 個であった．アクセス履歴の記録時期と HTML ソースを実際に取得した時期の間に，約 10 ヶ月の時間のずれがあるため，HTML ソースを取得した際，なくなっている可能性があるため，1 の条件を強調する必要があった．予備実験として 32,938 のトップページの中からランダムに 200 ページを取得した時，ステータスコードが 200 番台である割合を調べた．5 回の実験の結果は，平均 76%であった．これは本実験で集めたトップページの信頼性を示すものである．本実験では 10 ヶ月という時間のずれのため，正確なセカンドページの調査は出来ないと考え行っていない．

これらのトップページは，9,553 人分の利用者特定したアクセス履歴より抽出したものである．9,533 人のアクセス履歴に現れた，ファイルタイプが HTML でユニークなページは，5,234,099 個ある．抽出したトップページの数はその 0.63%に当たる．HTML ソース取得時の条件を満たしたトップページは全体の 0.08%である．これは膨大な Web ページの中から，重要なページである可能性が高いページに絞り込むことができたと考えることができる．

	URL	出現回数
1	http://www.yahoo.co.jp	35859
2	http://www.asahi.com	2473
3	http://www.google.co.jp	2038
4	http://messenger.jp.netscape.com/ja/bookmark/4_5/messengerstart.html	1851
5	http://livepage.apple.co.jp	1228
6	http://www.goo.ne.jp	1131
7	http://home.jp.netscape.com/ja/	944
8	http://jp.msn.com/	675
9	http://auction.yahoo.co.jp/	572
10	http://lycos.co.jp	526

表 1.人気トップページ

抽出したトップページを，トップページとしての出現回数の多い順に記したものが表 1 である．表 1 よりポータルサイトや検索エンジン，ニュースサイト，オークションが人気であることが分かる．出現回数の大きいトップページを調べると，多くの利用者が様々な情報にアクセスできる可能性の高いページを，トップページとして利用していることが分かる．

4.2. Namazu に基づく検索システム

検索システムのベースとして Namazu-2.0.10[2]を利用した．Namazu は日本語全文検索システムであり，ページの重要度計算に TF/IDF 法を利用している．さらに Web ページを対象とする時，キーワードのフォントの大きさやタグによりキーワードの重みを変えている．本研究では Namazu のコンテンツの重要度の計算方法を，前章で提案したコンテンツの重要度の計算方法に変更することでトップページ検索システムを開発した．*frequency* において利用頻度の集計を ALL_COUNT の方法で行ったが，利用したデータが 1 ヶ月分と少ないため，結果的に短期的な傾向を重視したものになった．利用頻度の中央値 *m* は 3 である．

開発した検索システムのインターフェース＜図 6 参照＞において，利用者はキーワードを入力する．そして利用状況をどのように考慮するかを以下の 4 通りから選択できる．

図 6.インターフェース

1. 考慮しない
   - ➢ 利用状況を考慮しない．$f(d)$=0 で計算．
2. 最近利用
   - ➢ 最近利用されているものを重視する．$f(d)$=1 で計算．
3. よく利用
   - ➢ 長期的に良く利用されているものを重視する．$g(d)$=1 で計算．
4. 最近よく利用
   - ➢ 最近良く利用されているものを重視する．

検索結果をソートする方法も選択が可能である．コンテンツの重要度，コンテンツのサイズ等で検索結果をソートできるようになっている．さらに表示形式として，コンテンツの要約を付けるか付けないかを選択することができる．次に UTPS の評価について述べる．

4.3. UTPS の実験評価

検索システムの評価として *frequency* の有用性の検証を行う．キーワードを"スポーツ"とし，利用状況を”考慮しない"と"よく利用"で検索を行いその検索結果の比較を行う．検索結果をコンテンツの重要度順に並べるようにした．2 つの検索結果をそれぞれ図 8A, 図 8B に示す．以下この検索結果の比較を行う．図中の表には結果に現れたページの順位と，コンテンツの説明を載せている．表の右端には，利用状況を考慮しないで検索した時の順位からの変動を矢印(↑，↓)で表している．"*"は 11 位以降から 10 位以内に順位を上げたものについている．

利用状況を考慮しないで検索した場合，上位 10 位にはスポーツ用品を対象とした電子商取引(EC)や，サッカー，占い，競馬，メジャーリーグ(MLB)に関するコンテンツが現れている．とりわけ電子商取引関連のコンテンツが多いことが分かる．*frequency* を考慮した検索結果は，EC が順位を下げ，サッカーやスキー，競馬と言った季節に相応しいスポーツに関するコンテンツが順位を上げている．サッカーはワールドカップや選手の移籍の話題，競馬はこの時期に地元の京都競馬場で開催されるレース多いことから影響が出たと推測される．これはアクセス履歴より計算した利用者集合の"地域性"や"時期"の特性が正しく反映した結果と考えることができる．

図 8A．利用状況を考慮しない検索結果例

図 8B．frequency を考慮した検索結果例

recency を考慮した場合は，長期的な利用頻度を重視する *frequency* に比べ，短期的な利用頻度を反映した検索が期待できる．”プレゼント”のようなキーワードに対する利用者の興味の変化が激しいと考えられるキーワードで検索を行った場合に日によって検索結果に違いが現れることを確認した．例えば，バレンタインデー付近では，香水や宝石の EC のサイトが検索結果の上位に現れるよ

うになるといった具合である．

幾つかの実験で利用状況を考慮した場合，Yahoo!やgoo に代表される，様々な情報にアクセスできるポータルサイトが上位に現れる傾向があることが分かった．これはポータルサイトの利用頻度が大きいことを意味している．さらにポータルサイトはリンクを多く含むことから，検索式に入力したキーワードがリンクとして含まれていることが多いと考えられる．この場合，入力したキーワードのリンク URL にアクセスすることで,利用者が望む情報に辿りつくことが期待できる．

5. 関連研究

プロキシログを解析する研究は多くあるが，本研究のようにトップページに注目した研究はない．プロキシログの持つ情報を利用して新しいサービスを提案，開発する研究として[3]などがある. [3]はアクセス頻度が大きいコンテンツが人気コンテンツであるという仮定の元に，アクセス頻度の大きいコンテンツを推薦するシステムを開発し，情報の共有を目指した研究である．

プロキシログを検索エンジンの改良に利用した研究が行われている[6]. 既存のディレクトリ型検索エンジンを改良する研究[5]がある．プロキシログに残されている主要検索エンジンの検索結果を利用して，提出されたクエリーと，それに基づき利用者に返された検索結果の中から，実際に利用者が選択したと思われるページを集め検索履歴とし，それを用いてクエリーをクラスタリングすることで，Web ページのグループわけを行い，既存のディレクトリ型検索エンジンの Web ページのグループ分けを改善している．この研究では実際に利用者が実際に利用したページを重要視することで検索エンジンの改善に成功している. 実際に利用されている Web ページが重要であるという考えが正しいことをしめしている．

6. おわりに

本研究の新規性は，利用者のアクセス履歴であるプロキシログとラディウスログを利用して個人の Web 活動の細かい追跡を可能にしたことである．その結果の一つとしてトップページを抽出した. これは固定 IP サービスを提供している ISP においてもラディウスログにあたる利用者の接続情報を用いることができればアクセス履歴の利用者特定は可能である．

本研究では，利用者を特定したアクセス履歴からトップページを抽出し，集めたトップページを対象とした検索システム[UTPS]を開発した．これにより，時間の移行に対応した利用者同士の質の高い情報共有ができると考えられる． ISP の新サービスとしての利用が期待できる．

本研究では，コンテンツの重要度の計算に，利用状況として *frequency*，*recency* の 2 つの要素を考慮することにより今までにない検索手法を提案した．開発した検索システムにおいてはプロキシログに表れる利用者集合の地域性や時期などの特性が反映されることが確認できた．

利用者の興味や地域性の特性を検索に反映する手法である，利用者の位置情報を用いた携帯端末での検索手法や検索履歴を利用した EC サイトでの検索手法に比べ，本研究の提案する手法は，位置情報がない利用者や検索履歴がない時でも，つまりどんな利用者に対してもサービスが可能なことである．これらのことは提案した検索手法が既存の静的な検索システムよりも優れていることを示していると言える．

本研究ではコンテンツの重要度を決める利用状況として *frequency*，*recency* を提案したが，これらに加えて更新頻度などの Web ページの内容性質によってコンテンツの重要度を決める方法が考えられる．今後はこのようなアクセス履歴に表れる利用状況のコンテンツの重要度計算へのより良い加え方の考案と共に，今回は扱えなかったトップページの次にアクセスされたページを考慮することによって，システムの充実を図ると共に，複数の ISP のプロキシログを利用し，ISP 間におけるトップページ集合の差異を調べることで，ISP の地域性などの特性を調べることを考えている．

文献

[1] G. Salton and C.S. Yang, “On the Specification of Term Values in Automatic Indexing,” Journal of Documentation 29(4), pp351-372.December 1973

[2] Namazu, http://www.namazu.org/

[3] 杉井俊彦, 北英彦, 林照峰, “アクセス回数を利用した WWW の人気ホームページ道案内システム.” 情報処理学会研究報告, vol97, No13, pp235-240, 1997

[4] Jon M and Kleinberg, “Authoritative sources in a hyperlinked environment,” Proc. of the 9th ACM-SIAM Symposium on Discrete Algorithms, pp 668-677, 1998

[5] M. Hansen and E. Shriver, “Using Navigation Data to improve IR functions in the context of Web search,” CIKM2001

[6] T. Mukai, K. Cheng and Y. Kambayashi, “A Usage-Aware Information Retrieval Method in World Wide Web. Master Thesis,” Department of Social Informatics, Kyoto University, 2002

[7] N. Yamada, R. Lee, H. Takakura and Y. Kambayashi, “Classification of Web Pages with Geographic Scope and Level of Details for Mobile Cache Management,” Second International Workshop on Web and Wireless Geographical Information Systems, in conjunction with WISE 2002 3rd International Conference on Web Information Systems Engineering, 2002

[8] Amazon, http://www.amazon.com/

[9] C. Cortes and D. Pregibon, “Giga-Mining,” Knowledge Discovery and Data Mining, pp174-178, 1998

[10] Y. Kambayashi and K. Cheng, “Capacity Bound-free Web Warehouse,” First Biennial Conference on Innovative Data Systems Research, 2003