6U-09 アクセス履歴を利用したWeb情報の共有手法

木村祥久† 尾内理紀夫† 後藤滋樹† ††

† NTTソフトウェア株式会社 インターネット技術センター

†† 早稲田大学 理工学部 情報学科

1．はじめに

インターネットのWWWによる情報流通は拡大の一途を辿り，ディレクトリサービスや検索サービスを利用しても目的とする情報を入手することが困難になってきている．これを解決するため，PA-search[1]ではソーシャルフィルタリング[2]や協調フィルタリング[3]，推薦システム[4]の手法を用い，検索母集合の質を高めることで適合率の高い情報検索システムを実現している．このシステムはユーザグループが利用するHTTPプロキシサーバのアクセス履歴から検索母集合を構築することで，以下の特徴を備える．

・グループ内でWeb情報を共有する

・自動的に情報を蓄積するためサービスの維持運営に必要なコストが低い

本論文ではアクセス履歴から検索母集合を構築する情報検索システムにおいて，Web情報を分類して蓄積する新しい手法と，その手法に基づいて実際に構築したシステムについて述べる．

2．URLの分類と情報の蓄積

無作為に情報収集するロボット系情報検索システムと比較して，アクセス履歴から検索母集団を構築する情報検索システムは検索母集合の質を高めることができる．しかし，アクセス履歴内にも質の低い情報が少なからず存在しているため，アクセス履歴内のWeb情報を分類し，質の低い情報を除去することで検索母集合の質を向上させる．

本論文でのアクセス履歴とはユーザグループが利用するHTTPプロキシサーバのアクセスログを表す．以下の前提条件によりアクセス履歴から利用者を特定できるものとする．

・クライアントはHTTPプロキシサーバへ直接アクセスしている（他のプロキシを経由していない）

・クライアントと利用者は一対一に対応している（1クライアントの利用者は一人）

利用者毎のアクセス履歴に着目し，以下の３種類にURLを分類する．

(1)定番

利用者の行動と情報の質との関連を調査したところ，ポータルサイトのトップページへのアクセスは前回のアクセスから時間的に間隔をおいていた．利用者はポータルサイト等をブックマークに登録するため，離散的なアクセスになると考えられる．逆に連続的なアクセスはフレームによる自動ダウンロードなど利用者の意図しないアクセスが多かった．このことから離散的なアクセス先のWeb情報を検索母集合に加えることで利用者間のブックマークを共有することが可能になる．この離散的なアクセス先のURLを「定番」と分類する．

(2)推薦

アクセス履歴は利用者がアクセスした結果として記録される受動的なものであるが，能動的な利用により情報を埋め込むことができる．利用者が推薦したいWeb情報を閲覧した状態で，ブラウザのリロードを短時間に２回実行する．この動作により同一Web情報に対する連続的なアクセスがアクセス履歴に記録される．アクセス履歴を解析することで推薦されたWeb情報へのアクセスを容易に抽出することが可能である．このようにして推薦されたURLを「推薦」と分類する．

(3)その他

「定番」や「推薦」に属さないURLを「その他」と分類する．

情報を蓄積する初期の段階では適合率よりも再現率を高める必要があるためURLの分類に関わらず検索母集合に加える．その後，情報の蓄積度を見ながら適合率を高めるため，URLの分類を利用して質の低い情報を検索母集団から除去していく．このようにして検索母集団のサイズを適切に保つことで，適合率を高めるだけでなく情報検索システムにかかる負荷を抑えるこ

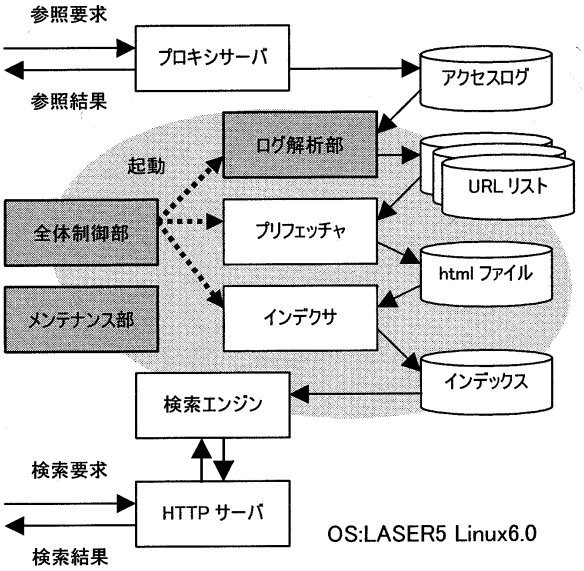
A technique for sharing Web information by using access logs
Yoshihisa Kimura
Internet Technology Laboratory
NTT Software Corporation

とができる.

情報の除去は「その他」に分類される URL リストからアクセス日時の古い URL を削除する．この作業は自動化することが可能であり，サービスの維持運営に必要なコストを抑えることができる.

3．システム構成

Web 情報を共有するシステムとして，プロキシサーバのアクセスログを利用し，過去に閲覧された Web 情報に対して全文検索するシステムを PC サーバ上に構築し運用した．システム構成を下図に示す.



①全体制御部は cron により定期的（１時間ごと）に起動される

②全体制御部はログ解析部を起動する．ログ解析部はプロキシサーバ squid[5]のアクセスログから URL を分類し，結果を URL リストに蓄積する

③全体制御部はプリフェッチャ wget[6]を起動する．プリフェッチャは URL リスト上の html ファイルを収集する

④全体制御部はインデクサ namazu[7]を起動する．インデクサは収集した html ファイルからインデックスを作成する

⑤HTTP サーバは検索要求に対して検索エンジン namazu からの検索結果を返す

⑥メンテナンス部は cron により定期的（１ヶ月毎）に起動され URL リストから１ヶ月以前にアクセスされた「その他」に分類される URL を削除し，html ファイルおよびインデックスファイルを削除する

4．運用結果と今後の課題

実際にシステムを運用した結果，情報の共有化という点で利用者に高い効果を認められた．ブックマークの登録と整理の必要性が減った点で好感を得られた.

利用者数が３人と少なく，ログ収集期間が２ヶ月と短いため「定番」を抽出するのに最適なアクセス間隔時間を把握することができなかった．長期に渡って調査を実施し精度を高めたい.

情報検索システムにおいて，検索結果に URL の分類を反映しなかった．検索エンジンのスコアと連動することで検索効率の向上が期待できる．今後は検索エンジンとの連携を高める方法を検討したい.

情報検索システムにおいて，URL を分類する情報管理手法について述べた．評価したシステムは単純な構造であるが，有益な情報を少ないコストで共有化することができた.

5．参考文献/URL

[1] 清水，神林，佐藤，風間，グループ指向 WWW 検索アシスタント PA-search の実現，
<http://www.ingrid.org/w3conf-japan/97/shimizu/pas-info.html>
[2] U.Shardanand, P.Maes, Social Filtering: Algorithms for Automating “Word of Mouth”,
<http://www.acm.org/sigchi/chi95/Electronic/documnts/papers/us_bdy.htm>
[3] P.Resnick, N.Iacovou, M.Suchak, P.Bergstrom, J.Riedl, GroupLens: An Open Architecture for Collaborative Filtering of Netnews,
<http://ccs.mit.edu/papers/CCSWP165.html>
[4] W.Hill, L.Stead, M.Rosenstein, G.Furnas, Recommending And Evaluating Choices In A Virtual Community Of Use,
<http://www.acm.org/sigchi/chi95/Electronic/documnts/papers/wch_bdy.htm>
[5] <http://squid.nlanr.net/>
[6] <http://www.gnu.org/software/wget/wget.html>
[7] <http://openlab.ring.gr.jp/namazu/>