グループウェア　37－11
（2000. 10. 19）

検索ログから抽出した知識の利用

鵜飼孝典
富士通研究所ドキュメント処理研究部
住所: 川崎市中原区上小田中 4-1-1
TEL: 044-754-2671
EMAIL: ugai@flab.fujitsu.co.jp

あらまし

ネットワーク上の情報を利用するために全文検索サービスを利用するがうまく情報に関連した文書を見つけ出すことができないとき，似た事をしている他の人がどう行なっているのかという知識を利用することは有効である．本稿では有用な URL，適切な絞込みキーワード，グループマネージメントための情報など，検索に利用する知識を検索ログから抽出する方法を提案し，得られた知識を検索支援に利用するインターフェイスを持ったシステムの試作について報告する．

キーワード　検索ログ，ログ分析，知識共有，検索支援

Using the Knowledge from Search Engine Log

Takanori Ugai
Fujitsu Laboratories Limited
Address:4-1-1 Kamikodanaka, Nakaharaku, Kawasaki
TEL: 044-754-2671
EMAIL: ugai@flab.fujitsu.co.jp

Abstract

When we search some information with search engines and unable to reach related documents, we ask someone who did same kind of searches how to do it. Using another's knowledge of usage the search engines is quite useful. In this paper, we propose a definition of such knowledge from keyword search logs and a system to share the knowledge and to use it.

key words　search engine log, log analysis, knowledge sharing, information retrieval

1. はじめに

現在，WWW などによって提供されているネットワーク上の情報を利用する方法として，キーワード検索サービスが頻繁に利用されているが，ほしい情報に関連した文書を探し出すことは難しい．このような状況では，過去に似たような情報を求めて，検索を行なった他の人がどう行なったかを，参考にすることが有効である．しかし，他の人に聞いたのでは，聞いた相手の作業を中断することになり，迷惑になる．また，適切なノウハウを持った人を探すことも必ずしも容易ではない．また流行に敏感な人は新しいキーワードについていち早く調べて知識を増やすが，そのような人が積極的にグループに有用な情報を広めることができれば，有効に情報の共有が行なわれる．しかし実際には世の中の動きに敏感ではない人はいつまでも知識を得ることが無いままになりがちである．

これらのような状況において，有用なURL，適切な絞込みキーワード，似たような知識を持つ人，利用するのに適当な情報源，情報を知っておくべきキーワードなどの知識を蓄積し，利用できるようにしてあれば有効である．

ユーザによる検索の履歴である検索ログは (1)問い合わせの履歴なのでコンテンツに比べて，ユーザが探しているものをより直接に反映している，(2)“結果が得られなかった”という情報が得られる，(3)時系列情報を利用できる，(4)ユーザによるプロファイルなどの情報を必ずしも必要としないといった特徴をもつので検索サービスを利用するためのノウハウが蓄積されていると考えられる．

本稿では，組織内向け全文検索サービスや proxy のログから先に述べた知識を抽出する方法を提案し，得られた知識を検索検索支援に利用するインターフェイスシステムの試作について報告する．

以下，第2章ではログから抽出される知識について述べ，第3章では知識を抽出するための処理，第4章ではこの知識を用いて検索支援を行なうために試作したシステムについて説明する．さらに第5章で関連する技術，システムについて述べる．

2. ログから抽出される知識

本章ではログから抽出する知識の概要，知識の形式的な定義，その定義にしたがって得られた知識の直感的解釈について述べる．

2.1. 知識の概要

本章では，本システムで利用するについて述べる．

本稿では検索ログをつぎの４つの要素からなる組の集合とみなす．

t ： 時刻
p ： 人
r ： 検索対象グループ
k ： キーワード

ここで t は検索を実行した時刻を示し，p は検索したユーザを区別する識別子である． r はニュースグループ， フォーラム，検索対象サーバなど検索対象となる文書の集合であり，k は検索で利用されたキーワードである．p に関しては所属組織， 好みなどの属性(プロファイル)が得られれば，異なるバリエーションの知識の抽出を行なうことができる．

そして検索ログを構成する要素間の関連性をユーザが共有する知識とする．

たとえば，キーワードと検索対象グループの関連性が数値として得られたとする．それがあるキーワードがある特定の検索対象グループで良く使われるほどその検索対象グループに対して関連性が大きく，他のグループでも同じような頻度で使われるキーワードは関連性が小さくなるものだとする．そのときキーワードに関連性の大きなニュースグループは過去に多くの人がそのキーワードについて検索を行なったものであり，そのキーワードについて情報が得ら

れるようなコンテンツであることが期待できる．

また検索対象グループ間の関連性は，同じキーワードが検索に利用される検索対象グループは関連性が大きく，違ったキーワードが利用される検索対象グループ同士は関連性が小さくなる．同じようなキーワードが良く使われる検索対象グループ同士は関連性が大きく，例えば検索対象グループがニュースグループであった場合，同じような議論が行われていることが期待される．

2.2. 知識の形式的定義

本稿では，ログを構成する要素である時刻，人，検索対象グループ，キーワードの相互の関連度を知識と定義する．本節では，時刻，人，検索対象グループ，キーワードの相互の関連度を定義する．

記法0：T は時刻の集合，P はユーザの識別子の集合，R は検索対象グループの集合，K はキーワードの集合であるとして，検索を時刻 $t \in T$，人 $p \in P$，検索対象グループ $r \in R$，キーワード $k \in K$ の組 $l = (t,p,r,k)$ とする．

記法1：検索ログを検索の集合 $L = \{(t,p,r,k) \mid t \in T, p \in P, r \in R, k \in K\}$ とする．

記法2：ある期間 $T_1 \subset T$ に利用された検索の集合を $L_1 = \{(t,p,r,k) \in L \mid t \in T_1\}$ とし，検索対象に指定された検索対象グループを $R_1 = \{r \mid (t,p,r,k) \in L_1\}$ とし，検索に指定されたキーワードの集合を $K_1 = \{k \mid (t,p,r,k) \in L_1\}$ とする．

記法3：L_1 の中で，キーワード $k \in K_1$ を指定している検索対象グループの集合を $R_k = \{r \mid (t,p,r,k) \in L_1 \wedge k \in K_1\}$ とする

記法4：L_1 の中で，検索対象グループ $r \in R_1$ を指定している検索で指定されるキーワードの集合を $K_r = \{k \mid (t,p,r,k) \in L_1 \wedge r \in R_1\}$ とする

定義1：L_1 の中で，単語 $k \in K_1$ を検索キーワードに指定した検索の集合を $L_k = \{(t,p,r,k) \in L_1 \mid k \in K_1\}$ とし，L_k のうち検索対象グループ $r \in R_1$ を検索対象に指定した検索の集合を $L_{kr} = \{(t,p,r,k) \in L_k \mid r \in R_1\}$ と定義する．

定義2：L_1 の中で，検索対象グループ $r \in R_1$ を検索対象に指定した検索の集合を $L_r = \{(t,p,r,k) \in L_1 \mid r \in R_1\}$，$L_r$ のうち，キーワード $k \in K_1$ を検索キーワードに指定した検索の集合を $L_{rk} = \{(t,p,r,k) \in L_r \mid k \in K_1\}$ と定義する．

定理1：任意の T_1 について $L_{kr} = L_{rk}$ である．

証明1：

$$L_{kr} = \{(t,p,r,k) \in L_k \mid r \in R_1\}$$
$$= \{(t,p,r,k) \in L_1 \mid k \in K_1 \wedge r \in R_1\}$$
$$= \{(t,p,r,k) \in L_r \mid k \in K_1\} = L_{rk}$$

定義3：(キーワードの検索対象グループにおける重要度) L_1 の中で，検索対象グループ $r \in R_1$ でのキーワード $k \in K_1$ の重要度 I_{rk} を $I_{rk} = |L_{rk}| \times \log(R_1/R_k)$ とする．ただし，集合 A の要素数を |A| とする．

定義4：(検索対象グループ間の関連度) L_1 の中で，検索対象グループ $r_1 \in R_1$ と検索対象グループ $r_2 \in R_1$ の関連度 F_{r1r2} を

$$F_{r1r2} = K_{r1} \times \sum_{ki \in K1} (I_{r1ki} \times I_{r2ki})$$

とする．ここで I_{r1ki} は L_1 の中で，検索対象グループ r_1 でのキーワード $K_i \in K_{r1}$ の重要度であるとする．

定義7で与える重要度は検索対象でセグメント化した検索ログを文書群とみなしたときのキーワードの TF/IDF と同じ定義である．

同様にして，キーワード間の関連度，キーワードと人の関連度，人と人の関連度，時刻，人，キーワードの関連度が定義できる．

2.3. 知識の解釈

本節では前節で定義した知識に対する実際のサービスにおける解釈について述べる．本稿では検索ログを構成する要素の関連度をユーザが共有する知識としている．たとえばキーワードとニュースグループの２項の関連度(K, R)が計算によって得られたとする．キーワードに関連の大きなニュースグループは直感的にはキーワードに深くかかわるニュースグループで，そのキーワードについて情報が得られるような議論が行われていることが期待できる．同様に

- (K, K)：キーワード同士の関連度からは入力キーワードに対して絞込みに使えるキーワードが得られる．
- (P, P)：人同士の関連度からは，例えば似たような興味を持つ人，同じような知識を持つ人が得られることが期待できる．
- (P, K)：人とキーワードの関連度からはある人，あるいは組織で行なわれている業務に必要な情報を示すキーワードが得られる．
- (T, K)：時間とキーワードの関連度からは時間の取り方により例えば季節ごとに決まって現れるキーワード，流行のキーワードなどが得られる．

また３項の関連度を計算することで同様に

- (T, P, K)：時間と人とキーワードの関連度から例えばあるプロジェクトで最近よく使われるキーワードが得られる．
- (P, R, K)：時間と人と検索対象グループの関連度から同じキーワードでもプロジェクトごとに検索する情報が異なるときにふさわしい検索サーバを得ることができる．

3. 知識抽出処理

本章では図 1に示す検索ログから知識を抽出する処理についてそれぞれ述べる．

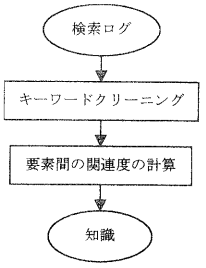

図 1:知識抽出の処理フロー

検索ログ:通常の検索ログ以外に proxy のログでもよい．

キーワードクリーニング: キーワードクリーニングではキーワードの表記揺れの除去，同義語の統一，ミススペルの訂正，および不要語の除去などを行なう．検索を行なうユーザは必ずしも文字列的なキーワードを含む文書を取り出したいわけではなく，多くの場合キーワードが示すコンセプトに関係する情報を取り出したいのであり，そのコンセプトを示す代表としてひとつのキーワードを選んで入力している．例えば“コンピュータ”，“Computer”，“計算機”などの同一コンセプトの代表として“コンピュータ”という文字列を入力していることが多い．そのためユーザが知りたいという観点ではこれらの同一コンセプトを示すキーワードを同一語とみなすことが必要となる．

関連度の計算：定義にしたがって，時刻，人，検索対象グループ，キーワードの相互の関連度を求める．

4. 応用例

本章では，検索ログから抽出した知識を利用する検索インターフェイスを備えたシステムについて述べる．

4.1. システム構成

本システムは組織内からインターネットにアクセスするための proxy サーバのログを用いる．図 2に示すように Proxy サーバのログから検索サービスの利用に関する部分を取り出し，2.2節の定義に従って知識を抽出し，データベースに蓄える．検索支援システムがデータベースに蓄積された知識をユーザに示す．

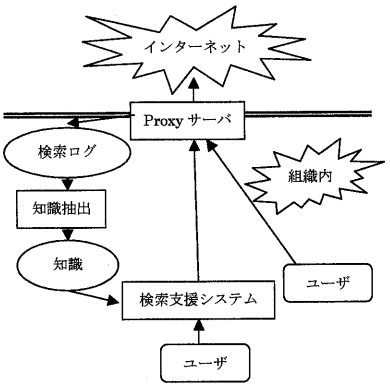

図 2：システム構成

本システムでは全文検索サービスを使うだけで自動的に知識が蓄えられ，自動的に更新される．蓄えられた知識は検索インターフェイス上で利用される．

4.2. 画面例

図 3は試作中のシステムの実行画面である．ここでは検索対象サーバとして@niftyの検索サービスである@search[6]を利用している．初期状態では最近良く使われるようになったキーワードが表示される．特定の組織向けシステムであれば，これらは知っておくことが望ましい（かもしれない，に違いない）キーワードとなる．表示されているキーワードから検索キーワードを選ぶか，または検索キーワードを入力して，検索を行なうと検索結果と共に関連キーワードと関連部署が表示される．関連キーワードを用いて絞込みを行なうことができる．また関連部署には連絡先がリンクされている．

5. 関連技術，関連システム

検索バディ[1] は，Web の著名検索サービスと連動し，同じキーワードで検索した人から情報をもらったり，お勧めのサイトを教えてもらったりできるWeb検索支援ツールであり，電子掲示板のように利用者でWeb 情報を提供し合うシステムである．文献[3]ではネットワーク上の情報資源を効率的に発見，利用するために，流通範囲が限定された，評価情報を伴う口コミを利用する手法を提案している．上記２つで用いられている手法は利用者が意識的に情報を提供するが本稿で提案する手法では，利用者による意識的な情報提供は不要である．

文献[4]ではその時点での Web 文書の参照履歴を基に参照済みのWeb文書群と関連性の高い Web 文書を推定し，次に参照するWeb 文書候補として推薦する方法を提案している．この手法では Web 文書の参照履歴を利用するのに比べ，本稿で提案する検索ログを用いた場合は，履歴に残るキーワードがより的確に探している情報を反映していると考えられる．

文献［2］では本稿とは異なる手法で検索ログから関連語を抽出する技術を提案しており，それは一人の人がある程度の短い時間に続けて入力しているキーワード，ある程度の期間で使用頻度が同じように増減するキーワードを関連語と定義しているものである．

6. まとめ

本稿では，有用な URL，適切な絞込みキーワード，グループマネージメントの支援のために有効な共有すべき情報を検索ログから抽出する方法を提案し，得られた知識を検索インターフェイスに利用するシステムの試作について報告した．

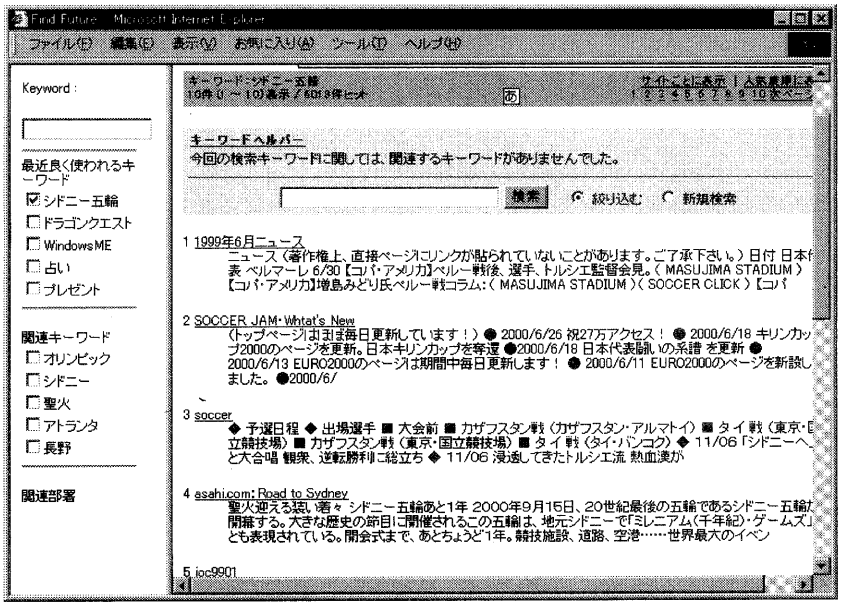

図 3：実行画面例

検索ログから抽出した知識，それを利用したシステムの評価が今後の課題として残されている．評価としては，抽出した知識が知識として妥当であるかという観点とそれを用いたシステムの使いやすさという観点からの評価を次のように行なうことを予定している．ユーザが関連キーワードを利用した頻度の測定と関連キーワードを用いた場合にどれくらいの絞込みが行われたか，その割合を測定することで抽出した知識の妥当性を評価する．またユーザに対するアンケートにより，使いやすさ，関連キーワードの妥当性を評価する．

本稿で提案する手法を用いた知識の抽出は，周りの人が“そこそこ妥当な”使い方をしていて，“そこそこ妥当な”結果を得られていることが前提となっていて，情報共有によって足りない部分を補い合う形になっている．組織内の大部分の人が情報を捜すことができないような状態では，本手法はうまく働かない．しかしながら一般的に組織内には流行に敏感で，上手に検索システムを使いこなす人もいて，そのような人にシステムの使い方を教わる人がいる．このような状況をかんがみた場合，本手法で仮定している状況は妥当であると考えられる．

参考文献

1. 検索バディ http://www.tt.rim.or.jp/~umeda/kbuddy/
2. 大久保他，WWW 検索ログに基づく情報ニーズの抽出，情報処理学会論文誌，Vol.39, No. 7, pp. 2250-2258
3. 大谷，南，口コミによる情報伝達を利用した情報資源管理 情報処理学会 56 回全国大会，Mar. 1998
4. 織田，南，参照履歴を用いた Web 文書推薦方法の提案 情報処理学会 56 回全国大会，Mar. 1998
5. @search, http://www.nifty.com/search/

盛光印刷所