検索ログ分析結果を利用して 知識を持つ人を探すことを支援するシステム

鵜飼孝典，三末和男
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@jp.fujitsu.com, misue.kazuo@jp.fujitsu.com

概要

だれでも容易に情報を発信でき,多くの情報を得ることができるようになった．しかしながら経験から得られる知識やノウハウは文書になっていないことが多い．我々はプロジェクトの作業履歴を蓄積して作業履歴を照合することにより人を探し出す方法を提案し，作業履歴として検索ログを用いたキーワードハイパーインデックス(KHI)について述べる．KHI は部署，キーワード，URL の3つの項目に対応したページがキーワード検索ログを分析して計算した関連度にしたがってハイパーテキストでそれぞれリンクしたものとして実現した．さらに必要な付属情報を一覧でき，多くの関連情報を閲覧できるように多数の異種データベースを結合することでより適切な人を探すことができるようにした．

キーワード： 検索ログ，ログ分析，検索支援,コミュニティ発見

A system supporting discovery “someone who knows” with keyword search log analysis

Takanori Ugai, Kazuo Misue

Fujitsu Laboratories Limited

Address:4-1-1 Kamikodanaka, Nakaharaku, Kawasaki
TEL: 044-754-2671
EMAIL: ugai@jp.fujitsu.com, misue.kazuo@jp.fujitsu.com

Abstract

IT technology makes easy information broadening and retrieving. However there are only few written documents on knowledge and know-how from experience. We propose the way to explorer “someone who knows” with matching the activities’ log. We describe a service named Keyword Hyper Index that shows persons who know the knowledge. The system consists of group name page, keyword page and URL page. The pages are linked each other based on relevance extracted from a search engine. The system let users find better persons with additional information from a variety of databases.

Search engine log, log analysis, knowledge sharing, information retrieval

1. はじめに

Web の普及により，だれもが容易に情報を発信でき，以前に比べてはるかに多くの情報を得ることができるようになった．しかしながら経験から得られる知識やノウハウは，Web 上に限らず文書になっていないことが多い．

そのような知識やノウハウを求めるには，結局は知っていそうな「人」に尋ねることが多いが，そもそも「知っていそうな人」をどうやって知るかが問題である．一般的に利用される「口コミ」では，自分が知る狭い範囲，あるいは自分が所属するコミュニティでしか人を探せないという難点がある．共通する知識や興味を持つ人の暗黙的なコミュニティは，数多く潜在すると推測されるが，そのほとんどが認知も利用もされていないと思われる．

我々は，これまで検索サービスや proxy のログから様々な知識を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システムを試作してきた[1]．それらの検索ログを知識ソースとして利用するメリットとしては次が挙げられる．

(1) ユーザの要求を検索結果(コンテンツ)よりも直接的に反映している．
(2) ユーザのプロファイル情報(の近似情報)を自然に蓄積できる．
(3) 検索サービスのユーザ全体を対象にした知識を獲得できる．

我々は，検索ログを活用することで，ユーザに余分な手間を掛けさせることなく，同様な情報を求める人，あるいは共通する情報に興味を持つ人からなる，暗黙的なコミュニティを探せそうだと考えた．

本稿では，そのような暗黙的なコミュニティの断片を提示することで，求める知識を持っていそうな人の発見を助けるシステム「キーワードハイパーインデックス(KHI)」について述べる．本システムは，キーワード検索と同様のインターフェイスで知識を持っている人を示し，(1)絞り込みを支援するための関連キーワード，(2)類似文書を見つけるための URL をキーとした検索結果，(3)求める知識を持っている人の所属する関連部署，の 3 つの項目をハイパーテキストで結合したインターフェイスを提供する．

以下，第 2 章では人の探し方についていくつかのアプローチを述べ，第 3 章では検索ログから人，キーワード，URL の関連度を抽出する方法について説明する．第 4 章では抽出した関連度を用いて実現したキーワードハイパーインデックスの概要を説明し，第 5 章で提案した手法の妥当性について考察する．

2. 人の検索

2.1. 背景

だれもが容易に情報を発信でき，多くの情報を得ることができるようになった現在でも経験から得られる知識やノウハウは文書になっていないことが多い．このような経験に基づく知識を得たい場合においては，過去に同じような経験をした人に聞くことがもっとも有効である． たとえばシステムエンジニア（SE）が開発中のシステム障害について調べる場合，解決の方法を文書としては見つけることができない場合でも，同様のシステムを開発した経験のある SE に聞くことで解決できることがしばしばある．そこでは同じような経験を持つ人を探すことが課題となる．

2.2. アプローチ

同じような経験を持つ人を探すという課題に対し，つぎの 3 種類のものを蓄積するアプローチによる解決が考えられる．

(1) **経験**：それぞれの SE が自分の経験を蓄える．蓄えられた経験そのものの照合により人を探し出す．
(2) **産物**：プロジェクトの最終成果物，中間生成物，メモなどを蓄える．経験そのものの照合に代えて，蓄えられた産物を照合することにより人を探し出す．実際には産物に出現するキーワードなどで照合することになる．
(3) **作業履歴**：プロジェクトの作業履歴，プロジェクトで利用した補助システムのシステムログを保存する．同じような経験をする人は，類似の作業履歴を残すであろうという仮定に基づき，作業履歴を照合することにより人を探し出す．

たとえば，Solaris を利用した検索サービス構築時に Oracle 周りで障害が起こったとしよう．

(1)のアプローチは，その障害そのものについて障害内容を蓄積するものである. 知識を求める者にとっては非常に効率が良いが, 知識の蓄積者にとっては, 本来の業務に対して余分な作業が発生するという問題がある．(2)のアプローチは，当該プロジェクトによって生成されるドキュメントを蓄積しておき, まず類似プロジェクトを探し，そこから該当知識を得るというものである．これらのアプローチはこれまでにも知識管理システムで行なわれている[2,3].

さて，このような障害に直面した人の多くが，「Solaris」や「Oracle」に関する情報を収集したと考えられる．そこで，過去に「Solaris」や「Oracle」に関する情報を収集した人を探すのが，アプローチ(3)のアイデアであり，本稿で提案するものである.

3. ログからの関連度抽出

我々が構築するシステムでは, キーワード検索サービスと同様に，キーワードから聞くべき人を探す方法と，人の ID から同じような検索を行なった人を探す方法を提供する． そのためにシステムはキーワードと人の関連度と, 人と人の関連度を検索ログから抽出し, 関連度の大きな人を表示する. 本章ではログから抽出する関連度の概要と形式的な定義について述べる.

3.1. 抽出したい関連度

過去にあるキーワードを用いて特に多くの回数検索を行なっている人をキーワードに関連度の大きな人とする．そのキーワードに関連度の大きな人は, そのキーワードを用いて過去に何回も調べたことがあり, そのキーワードに関して多くの知識をもっている人であることが期待できる.

同じキーワード同じような頻度パターンで検索に利用した人同士は関連度が大きく， 違ったキーワードを利用する人同士は関連度が小さくなるように関連度を定義する．このとき,関連度の大きな人同士は, 同じような知識を共有していることが期待される.

3.2. 関連度の形式的定義

検索ログをつぎの 3 つの要素からなる組の集合とみなす.

p ： 人
k ： キーワード
r ： URL

ここで p は検索したユーザを区別する識別子である． k は検索で利用されたキーワードであり，r は k をキーとして検索した結果からアクセスした URL である.

本節では，人，キーワード，URL の相互の関連度を定義する.

記法０： 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，検索を人 $p \in P$，URL $r \in R$，キーワード $k \in K$ の組 $l = (p,r,k)$ とする.

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

記法 2：検索の集合 L に対して， 検索の結果から利用ユーザの集合を $P = \{p \mid (p,r,k) \in L\}$ とし， 検索に指定されたキーワードの集合を $K = \{k \mid (p,r,k) \in L\}$ とする.

記法 3：L の中で， キーワード $k \in K_1$ を指定したユーザの集合を $P_k = \{p \mid (p,r,k) \in L \wedge k \in K\}$ とする

記法 4：L の中でユーザ $p \in P$ にアクセスした検索において指定されたキーワードの集合を $K_p = \{k \mid (p,r,k) \in L \wedge p \in P\}$ とする

定義 1：L の中で，キーワード $k \in K$ を検索キーワードに指定した検索の集合を $L_k = \{(p,r,k) \in L \mid k \in K\}$ とし， L_k のうちユーザ $p \in P$ が検索した検索の集合を $L_{kp} = \{(p,r,k) \in L_k \mid p \in P\}$ と定義する.

定義 2：L の中で，ユーザ $p \in P$ が検索した検索の集合を $L_p = \{(p,r,k) \in L \mid p \in P\}$， L_p 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 $L_{pk} = \{(p,r,k) \in L_p \mid k \in K\}$ と定義する.

定義 3：(キーワードのユーザに対する関連度) L の中で，ユーザ $p \in P$ に対するキーワード $k \in K$ の重要度 I_{pk} を $I_{pk} = |L_{pk}| \times \log(R/R_k)$ とする． ただし， 集合 A の要素数を |A| とする.

定義 4：(ユーザ間の関連度) L の中で，ユーザ $p_1 \in P$ とユーザ $p_2 \in P$ の関連度 F_{p1p2} を

$F_{p1p2} = K_{p1} \times \sum_{ki \in K1} (I_{p1ki} \times I_{p2ki})$ とする．ここで I_{p1ki} は L の中で，ユーザ p_1 に対するキーワード $K_i \in K_{p1}$ の関連度であるとする．

定義３で与える関連度はユーザ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 URL の関連度，キーワード間の関連度などが定義できる．

3.3. ユーザのグループ化

上の定義に従って，実際の検索ログ(第 4 章で述べる)を利用して予備実験を行なったところ，利用頻度の多いユーザが何に対しても聞くべき人として関連付けられるという結果になった．

具体的には，各キーワードに対して最も関連度の大きいユーザ(つまり聞くべき人)が，全ユーザの 5%に集中した．関連度の上位 5 位までに含まれるユーザに聞くべき人を拡大しても，全ユーザの 7%にしかならなかった．つまりどのようなキーワードを入力しても大体同じ人が聞くべき人として表示されるということである．

これは検索の利用頻度の偏りによる弊害であると考え，本システムでは p をユーザの所属組織(部署)でグループ化することで，その偏りを分散させることにした．グループ化した場合，全キーワードに対する関連度の最も大きなグループは全体の 36%になり，5 番目までに含まれるグループはさらに 52%にまで拡大された．

グループあたりのキーワードの種類が，個人あたりのキーワードの種類より多くなるため，回数の多さよりも特徴的なキーワードの出現が関連度に大きく作用するようになるためである．50%程度が聞かれ役になる状況になれば，グループ間のコミュニケーションがとられ，広い範囲で知識を探すことができるようになると考えている．

4. キーワードハイパーインデックス（KHI）

本章では，検索ログから抽出した関連度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. システム概要

本システムは検索サービスのログから人，キーワード，URL の間の関連度を抽出し，蓄える．上記３つの項目に対応したページが関連度にしたがってハイパーテキストでそれぞれリンクされたものである．部署，キーワード，URL のそれぞれのページには関連部署，関連キーワード，関連 URL が３章で定義される関連度の大きな順に表示される．KHI は次の３つの特徴をもつ．

ハイパーテキスト：部署，キーワード，URL の３種類のページが相互に前章で定義した関連度に従ってリンクで結合されている．リンクをたどるだけで関連する部署やキーワードに関する情報を閲覧できる．

一覧性：部署，URL，キーワードのそれぞれのページでは Web の分析などで得られる付随する情報を一覧できる．

多様性：KHI のハイパーテキスト構造だけでなく，電子電話帳，キーワード検索などの多くの情報システムと相互に結合することにより，多くの情報が得られるようになっている．

4.2. 画面構成

図 1はシステムの実行画面例である．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．これらは知っておくことが望まし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すると，(a)のキーワードページが表示される．関連部署とともに関連キーワード，関連 URL が表示される．関連部署には連絡先がリンクされている．関連キーワードをチェックして検索ボタンを押すとキーワード検索サービス用いて絞込み検索を行なうことができる．関連部署の名前をクリックすると(b)の部署のページが表示される．部署のページは電子電話帳と連動し，電話番号，メールアドレスが検索でき，電話やメールによる問い合わせを容易にする．関連 URL をクリックすると(c)の URL ページが表示される．

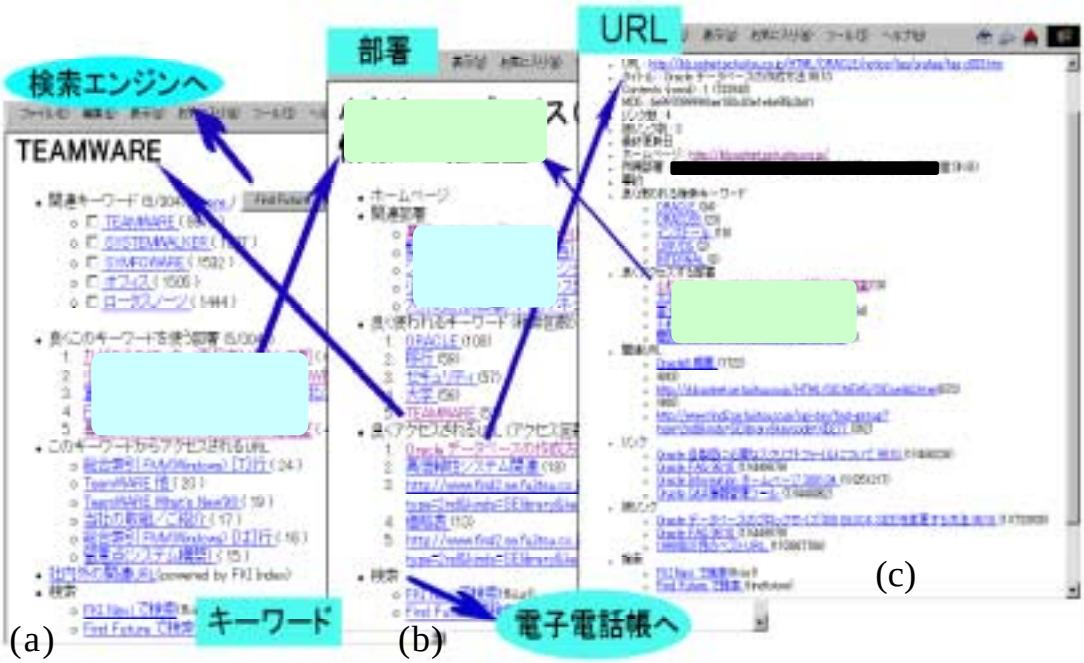

図 1:実行画面例

更新日時，要約，リンクページ，被リンクページを表示し,関連情報を使って，より新しく，より良質な情報が記載された文書を探すことができる．

キーワード，URL，部署のそれぞれのページにおける関連部署，関連キーワード，関連 URL の項目には次のような機能を実装し, 各種のデータベースを簡便に検索できるインターフェイスを実現した

- 電子電話帳検索によって電話番号やメールアドレスを検索できるようにした．
- Internet の各種キーワード検索，社内キーワード検索へのインターフェイスを各ページにリンク付けを行なった．
- URL に関しては更新日時，タイトル，要約，その URL を提供している組織や部署の特定を行ない，その URL に実際にアクセスしなくてもある程度の中身がわかるようになっている．

4.3. 実装

本システムは関連度抽出にイントラネット内の検索エンジンのログを用いた． 今回実装のために用いた検索ログは以下のような大きさである．

- ログの収集期間:1999 年 4 月から 2001 年 2 月 (1 年 10 ヶ月)
- 総検索数：1,533,597 回
- ユーザ数：134,254 人
- グループ数：13,415 グループ
- キーワード数(キーワードクリーニング後：後述)：48,235
- アクセス URL 数：69,531

図 2 に示すように,検索エンジンのログから 3.2節の定義に従って関連度を抽出し，データベースに蓄える. 検索ログから関連度抽出にかかった時間は以下のとおりである．

全キーワードと全グループの関連度計算：38 時間

全グループ間の関連度計算： 15 時間

利用計算機：

PentiumIII 850MHz 1GBMem

Solaris8 for x86

KHI のインターフェイス合成部がデータベースに蓄積された 3 章で定義した関連度にしたがって，関連部署，関連 URL，関連キーワード表示する．ハイパーテキストは WWW を用いて実現した．

検索ログからキーワードを抽出する際に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を行なう. 本システムの利用者は必ずしも文字列的なキーワードに関連する人を探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持つ人を探すということを想定していて, 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させるようにしている. 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.

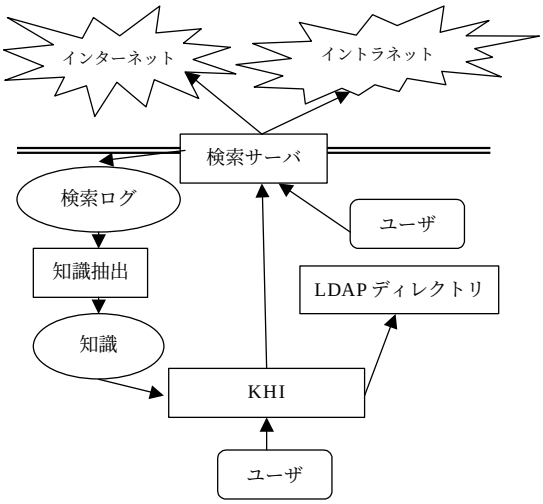

図 2：システム構成

5. 考察

KHI を使って得たい知識を持つ人にどれくらい到達できるかについては本システムが試験的に利用され始めたばかりで評価が行なえていない. ユーザに対するアンケートの実施,分析を行ない, またあらかじめ決めた問題に対する実験を行ない本システムの有効性を確認する予定である.

以下では, われわれが採用した検索ログの分析結果から人を探すための関連度を抽出するアプローチの妥当性について考察する. われわれが提案しているログの分析手法では過去に多くの回数調べた人を, そのキーワードについてよく知っているので聞くべき人であるとしている. これは, 検索を多く行なった人はそのキーワードに関する知識をもっているという仮定に基づいている. この仮定は一般的には妥当であると考えられる. 一方でそのキーワードに関する知識を持っている人の中にそのキーワードを頻繁には使っていない人がいるとも考えられる. キーワード検索以外の方法によって多くの知識を得ている人はわれわれのアプローチでは探し出すことが出来ない. 例えばあるシステムの開発者が知っているパラメータのチューニングのノウハウを持った人は本システムでは見つけることが出来ない. これを補うためには 2 章で挙げた 2 つのアプローチを併用することが有効であると考えている.

6. まとめ

本稿では, 経験から得られる知識やノウハウなど文書になっていないものを探す場合に, 求める知識を持っている人を見つけるサービスを提供するキーワードハイパーインデックス(KHI)について述べた. また同じような経験をする人は, 類似の作業履歴を残すであろうという仮定に基づき, プロジェクトの作業履歴を蓄積して, 作業履歴を照合することにより人を探し出す方法を提案した. KHI では作業履歴としてキーワード検索ログを用いた. システムは検索ログを構成する部署, キーワード, URL の 3 つの項目に対応したページを関連度にしたがってハイパーテキストでそれぞれリンクしたものとして実現し, 必要な付属情報を一覧でき, 多くの関連情報を閲覧できるように多数の異種データベースを結合した. KHI を使って得たい知識を持つ人にどれくらい到達できるかについての評価が今後の課題である.

参考文献

1. 鵜飼, 検索ログから抽出した知識の利用, 情報処理学会グループウェア研究会, Oct 2000,pp61-66

2. 黒瀬, 事例 2：ナレッジマネジメントとその支援技術,人工知能学会誌 Vol.16, No.1 2001/1, pp54-63

3. 黒瀬, ソフト・サービス部門のワークスタイルの変革, 情報処理学会誌 Vol.40 No.3,pp308-311