2005－MPS－53（9）
2005／3／10

P2P を用いた検索語共有による Web 検索支援システム

丸山　崇[†]　斉藤栄一[††]
堀田知宏[†††]　北英輔[†]

膨大な情報源である Web 上において，検索対象分野に関する知識が乏しい場合，検索者の知識から導く検索語では検索要求を具体的に表現できず，目的の情報に到達しにくい．本研究ではキーワードの情報源に検索語を用い，URL と結びつけることで，検索された Web ページに関連したキーワードを提供するシステムを提案する．検索語はスペースで区切られた形式で検索エンジンに入力され，キーワードを得るために文字列を解析する必要がないため，検索語の劣化が起きない．また，検索語を URL で結びつけるため，URL が存在する検索語のみ共有化されるので Web ページと関連のあるキーワードを提供できる．更に，キーワードのソーティングを行うことで，多数のキーワードからユーザの検索要求に適合したキーワードを提示できるシステムとした．

Searching Assistant System by Common Search Words Using P2P

Takashi Maruyama,[†] Eiichi Saito,[††] Tomohiro Hotta[†††]
and Eisuke Kita[†]

If the Web user has no knowledge, he cannot express search words form his knowledge and it's difficult to get information of purpose. In this paper, we propose the Searching Assistant System by common search words using P2P. This system connects keywords for using search words and URL, and it shows keywords related to Web. Using this system, Web users are able to share search words connected with URL, so they can obtain keywords related to Web.

1. はじめに

近年通信回線の大容量，定額制の普及によってインターネットを利用した情報発信が盛んになり，Web 上で利用できる情報量が増大している．インターネット利用者が求める情報にたどり着くために最も使用するサービスが検索エンジンである．しかし，検索対象分野に関する知識が乏しい場合，検索者の知識から導く検索語では検索要求を具体的に表現できず，目的の情報に到達しにくい [1]．検索支援システムはそのような検索語不足を解決するシステムである．以下,「検索語」とはユーザが検索エンジンに入力する語であり,「キーワード」とは検索支援システムによって提示される語である．これまでの検索支援システムではキーワードを得るための情報源が 2 種類ある．1 つは検索エンジンの検索結果から得た文章群で，その文章から単語を抽出しキーワードを得る方法である．もう 1 つは検索対象とは別途にキーワードのデータベースを用意しておく方法である．前者の方法は，文章群の文字列からキーワードを抽出するのに形態素解析を用いる．形態素解析を用いた場合，文字列を形態素まで分解するため，複数の形態素からなる語が分解されてしまい有効な絞込みができるキーワードを提供するのが難しい．後者の方法では，検索対象となっている Web ページの文章集合とは別に用意された類義語や関連語は，ユーザの知識不足を補えるが，そのキーワードを含む Web ページが存在するとは限らない．本研究ではキーワードの情報源に検索語を用いる．検索語を P2P を用いて共有化し，URL と結びつけることで，検索された Web ページに関連したキーワードを提供するシステムを提案する．検索語はスペースで区切られた形式で検索エンジンに入力され，キーワードを得るために文字列を解析する必要がないため，検索語の劣化が起きない．また，検索語を URL と結びつけることで URL が存在する検索語のみ共有化され，Web ページ

† 名古屋大学大学院 情報科学研究科 複雑系科学専攻
Graduate School of Information Science, Nagoya University
†† 名古屋大学大学院 人間情報学研究科
Graduate School of Human Informatics, Nagoya University
††† 名古屋大学 情報文化学部
School of Informatics and Sciences, Nagoya University

と関連のある語を提供できる．更に，各キーワードごとの適合度を求め，それに基づいたソーティングを行うことで，多数のキーワードからユーザの検索要求に適合したキーワードを提示できるようにした．

第 2 章では本研究で提案する検索支援システムの説明を述べる．第 3 章では本システムの実行例を述べる．第 4 章では本研究のまとめを述べる．

2. 提案するシステム

本章では提案するシステムの概要，アルゴリズムを説明する．本システムの特徴はキーワードの情報源に検索語を用いて，データベース型の情報提供を行うことである．検索語をキーワードとして用いる理由は，それ自体が絞込みを可能にする語を含み，かつ，検索対象の Web ページに関連した語を提供できるからである．

第 2.1，2.2 節において，本システムの特徴である検索語の絞込みと，キーワードの提示について述べる．第 2.3 節では提示するキーワードのソーティングについて，第 2.4 節では本システムが採用する P2P について述べる．第 2.5 節では本システムのアルゴリズムについて述べる．

2.1 他のユーザの検索語による検索語の絞込み

検索者が入力する検索式は，検索対象分野の専門知識を有する人の場合，一般の人より多くの検索語を含み，精度の良い検索結果を得ている [1] ．本研究では，検索対象分野の専門家が用いる検索語は，その分野の知識が不足している人にとって絞込みを可能にし，検索を支援するキーワードを用いていると仮定する．検索語の共有によって，知識が不足しているユーザは同じ検索対象の専門家が用いる検索語を得ることができ，自身の検索語の知識不足を補えると考える．

2.2 検索対象分野のキーワードの提示

知識が乏しい人の検索結果は，検索要求に適合した文章群の一部しか検索されていない [1] ．本研究では，検索要求に適合した文章群は検索対象の専門家が検索エンジンを用いて得た文章群とほぼ同じとし，検索対象の知識が不足している人の検索結果の中に専門家の検索結果が含まれていると仮定する．検索エンジンの性質上，同じ検索対象に対して検索対象を絞り込むキーワードを含まない検索は，その語を含む検索結果に加えて，多くの Web ページがヒットしてしまうためである．本システムではこの性質を利用し，検索対象の知識が不足している人の検索結果の中には専門家がたどり着いた URL が含まれるとし，他のユーザが同じ URL を得るのに用いた検索語を提示する．

2.3 キーワードのソーティング

検索支援システムにより多数のキーワードが表示された場合，検索対象の知識が少ない人はどのキーワードが検索要求に適合しているのか判断するのが難しく，キーワードを１つ１つ調べるのは非常に時間と労力がかかる作業である．本研究ではキーワードごとに入力した検索語に対する適合度を求め，その値に基づいたソーティングを提案する．適合度とは提示されたキーワードがどれくらいユーザが入力した検索語にマッチしているかを表した値である．このソーティングによりユーザは検索要求に適合したキーワードを選択できると考える．

2.4 P2P 方式の採用

本システムの実装は P2P を用いる．P2P は厳密な運用は求めず，システム全体としての安定性を求めるシステムに向いている [2] ．本システムでは P2P の実現に JXTA を用いる．JXTA とは Sun Microsystems 社が開発した P2P 型アプリケーションを容易に開発できる環境を提供するオープンソースプロジェクトである．プラットフォームに依存しない技術を目指しており，携帯電話や PDA を含んだ，あらゆるデバイスで使用できることを目指している．また，プログラミング言語に依存しない技術を目指しており，C 言語，Perl，Pysthon，Ruby にバインディングするプロジェクトが進行中である．

2.5 提案システムのアルゴリズム

本節では提案する検索支援システムのアルゴリズムについて述べる．本システムは「キーワードの発見と提示」,「キーワードのソーティング」から構成されている．第 2.5.1 節では「キーワードの発見と提示」について，第 2.5.2 節では「キーワードのソーティング」について述べる．

2.5.1 キーワードの発見と提示

図 1 は「キーワードの発見と提示」のフローチャートである．ユーザが本システムに検索対象の検索語を入力すると，その語を用いて検索エンジンで検索を実行し，検索結果の URL を得る．検索語と URL をローカルピアのデータベースに保存し，その後 URL を P2P ネットワークにクエリーとして送信する．リモートピアでそのクエリーを受信するとリモートピア内のデータベース内を検索し，同じ URL が存在する場合，その URL を得るのに用いられた検索語をキーワードとして送信メッセージに加え，同じ URL が存在しないなら，メッセージ”`Nothing_in_URLSERVICE`”を送信メッセージに加えてクエリーを送信してきたピアにメッセージを返信する．メッセージを受信したピアでは，検

索語と同じキーワードと”Nothing_in_URLSERVICE”を除き表示する．

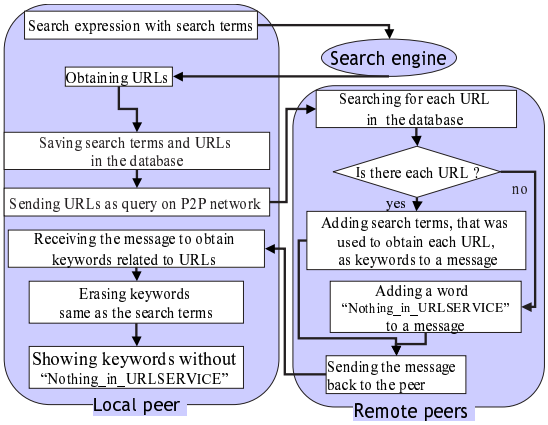

図 1 Procedure for obtaining keywords

2.5.2 キーワードのソーティング

図 2 は「キーワードのソーティング」のフローチャートである．本システムからローカルピアに提示されたキーワードを検索エンジンに入力して関連した URL を得る．得た URL をリモートピアに送信し，結びつきがあるキーワードを得る．このリモートピアから得たキーワードの中に含まれる検索語の数を評価することで，ローカルピアから提示されたキーワードの適合度得る．適合度の計算は以下の式を用いる．

$$F = \sum_{i=1}^{p} f_i \qquad f_i = 100^q$$

F: キーワードの適合度
f_i: 各 URL ごとに得られる適合度の部分値
p: 検索エンジンから得た URL の数
q: 各 URL に含まれる検索語数

3. 実験結果

本節では PC4 台を用いた実験結果を示す．実験をする前に，各ピアのデータベースに Java 関連の語，Linux 関連の語をデータベースに登録している．

本検索支援システムを用いて，Java の開発を支援するエディタを検索した結果（**図 3**）が**表 1** である．検索語として”java”，”エディタ”を使用した．本システムにより提示されたキーワードの中に検索目的に合うキーワードとして”秀丸エディタ”，”magcup”，”eclipse”，”emacs”が存在した．秀丸エディタは幅広くしようされているエディタであり，

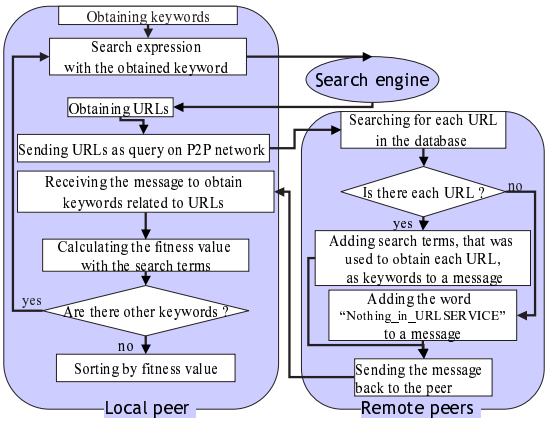

図 2 Procedure for sorting

プログラムソースの編集によく使用される．magcup は Java により作成された Java プログラム開発用のエディタである．eclipse は Java の統合ソフトウェア開発環境である．emacs は UNIX 環境で最も広く普及しているテキストエディタであり，Java の編集機能がある．ソーティングによりこれらの語が上位に並んでいるため，提示されたキーワードをうまくソーティングできたといえる（**図 4**）．

表 1 Result of search:1

本システムにより提示されたキーワード	秀丸エディタ horb ide ブラウザ 軽い linux 並列 ライブラリ c 言語 magcup javac 開発環境 eclipse 分散オブジェクト …
ソーティング	秀丸エディタ javac magcup eclipse swt emacs ：

次に，”linux”，”gui”，”プログラミング”を検索語として使用した結果（**図 5**）が**表 2** である．上位に，”qt”，”gtk+”，”xlib”，”motif”，”tcl/tk”などが提示された．これらのキーワードは Linux もしくはマルチプラットフォームに対応した GUI ツールキットやライブラリ，X Window System のプログラミング言語であり，いずれも Linux の GUI プログラミングに深く関連しているキーワードである．また，”gtk+”，”gui ツールキット”，”tcl/tk”は，”gtk”と”+”，”gui”と”ツールキット”，”tcl”と”/”と”tk”などのようにキーワードが分解されず，知識のあるユーザが検索エンジンで用いた検索語がキーワードとして提示できているといえる．

更に，本システムはキーワードと URL の結びつき

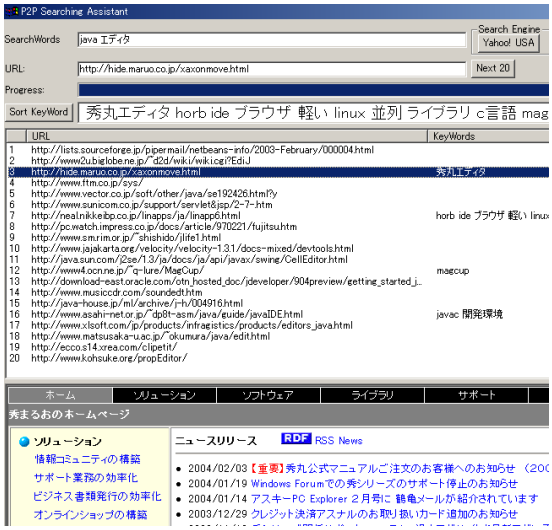

図 3 Execution screen:1, Before sort

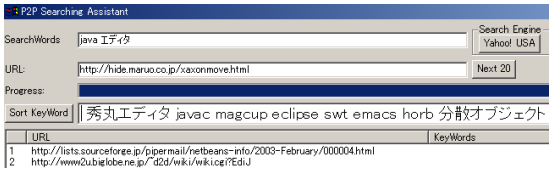

図 4 Execution screen:1, After sort

を用いることで，検索エンジンでヒットした URL にどのようなキーワードが含まれているか提示することができる（図 3，図 5）．これにより，検索目的に合う Web ページの発見を支援し，検索者の負担を軽減できるといえる．

表 2 Result of search:2

検索語	提示されたキーワード
linux	qt
gui	gtk+
プログラミング	ライブラリ
	tk
	fox
	gui ツールキット
	xlib
	motif
	tcl/tk
	：

4. ま と め

本研究では検索対象の知識が少ない人が検索対象の専門家が用いる検索語を用いて検索を絞り込むためのシステムを提案した．本システムは，複数のユーザが検索エンジンで用いた検索語を P2P を用いて共有化

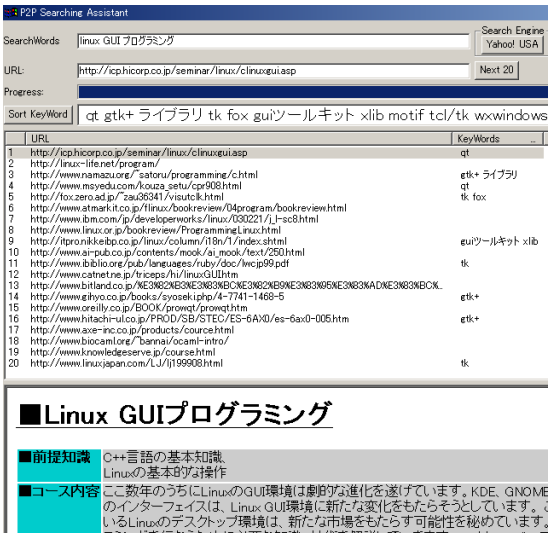

図 5 Execution screen:2

し，URL で検索語を結び付けることで，検索された Web ページに関連するキーワードを提示する検索支援システムを提案した．更に，検索語に対するキーワードの適合度を求めその値によりソーティングする機能を設け，提示された多数のキーワードからユーザの検索要求に適合したキーワードを選択し易いようにした．

本システムではピアの数が膨大になるにつれて，検索に要する時間が増加し，更に，ネットワーク全体に負荷をかけてしまう．これを解決する方法として，ユーザの嗜好情報を利用したクラスタリングが考えられる．同嗜好同士のピアのデータベースを参照することで，ネットワークに負荷をかけず，効率のよいキーワードの探索と提示が可能になる．

謝辞

本研究を遂行するにあたり，21 世紀 COE プログラム「計算科学フロンティア」から援助を頂いた．ここに記して謝意を表する．

参 考 文 献

1) 木谷強，高木徹，木原誠，関根道隆，”フルテキストと抽出キーワードを利用した情報検索”，情報処理学会報告，96-NL-115，pp.129-134，1996.
2) 横澤誠，”サーバーに依存しない P to P 型システムの設計”，知的資産創造，pp.54-67，2002，野村総合研究所．