Webサイトを利用したニーズの収集と分析に関する研究

鈴木浩之*1　内藤丈資*1　石田　聡*2　伊藤幸希*3

Study on Collection and Analysis of Needs Using Web Site

SUZUKI Hiroyuki*1，NAITO Takeshi*1，ISHIDA Satoshi*2，ITO Yuki*3

抄録

Webサイトに自前のサイト内検索システムを構築し、そのアクセスログを取得することで閲覧者のニーズ情報を収集できるようにした。サイト内検索で使われたキーワードとして、１か月でおよそ150個のサイト内検索情報が収集できた。また、ページ管理システムを開発し、効率的な管理ができるようにするとともに、従来データベースとWebアプリケーションで動的に表示されていた情報をHTML化した。これにより一般の検索サイトによる検索に、より多くヒットするようになった。

キーワード：Webサイト，検索キーワード，アクセス解析

１　はじめに

当センターでは、自主管理しているサーバ上にWebサイトを構築し、情報提供を行っている。一般にWebサイトではアクセスログを活用したマーケティングなどが可能であり[1)]、当Webサイトでも、アクセスログを取得して閲覧の傾向などをとらえコンテンツの改良に役立てている[2)]。

しかし、従来のアクセスログを取得しただけでは、閲覧されたページの情報は取得できるものの、どのような情報を求めているのかというニーズに関する情報を取得することは難しかった。また、改良すべきコンテンツがあっても、対象となるページの量が多く、管理効率の悪さが原因で改良できないページがあるという課題もあった。

そこで本研究では、ニーズ情報を取得できるようにし、収集、分析を行うことを目的としてサイト内検索システムを自前で構築した。また、効率的な改良や、アクセス解析用のHTMLタグまで含めた管理を行えるよう、ページ管理システムを開発した。

*1 電子技術部

*2 戦略プロジェクト推進担当

*3 総務・企画室

２　システム開発

2.1 サイト内検索システム

これまでは、無料で広く使われている外部の検索サイトを、当 Web サイトの URL に限定する設定で組み込み、サイト内検索として利用してきた。しかしこの場合、検索結果を表示するページは当 Web サイト内のページではなく、外部の検索サイトのページとなるため、検索結果に関するアクセスログを取得することはできなかった。

2.1.1 Namazu の利用と自動化

自前のサイト内検索システムを構築するために、Namazu を利用した。Namazu はオープンソースの日本語全文検索エンジンであり、GPL ライインセンスで提供されているため [3)]、一定の条件下で利用や改変が可能である。

Namazu は、コンテンツ情報をあらかじめインデックス化することで高速な検索を可能としてい

るが、コンテンツを更新した場合は再度インデックス化する必要がある。このインデックス化を１日に１回、自動的に実行するようにした。これにより、新しいコンテンツ情報もほとんど遅れることなく検索結果に反映されるようになった。一般の検索サイトを利用した場合は、更新された情報が検索に反映されるまで時間がかかることがある。更新の反映が早いことは、自前のサイト内検索システムの優位な点の一つである。

2.1.2 スタイルシートの適用

当 Web サイトでは、スタイルシートを使って埼玉県の統一フォーマットにしたがったスタイルを適用している。サイト内検索のページにも同じスタイルシートを適用し、統一フォーマットに適合したスタイルになるようにした。（図１）また、サイト内検索の利用方法に関する説明を当 Web サイトで扱う内容に合うように修正することで、利用者に分かりやすくした。

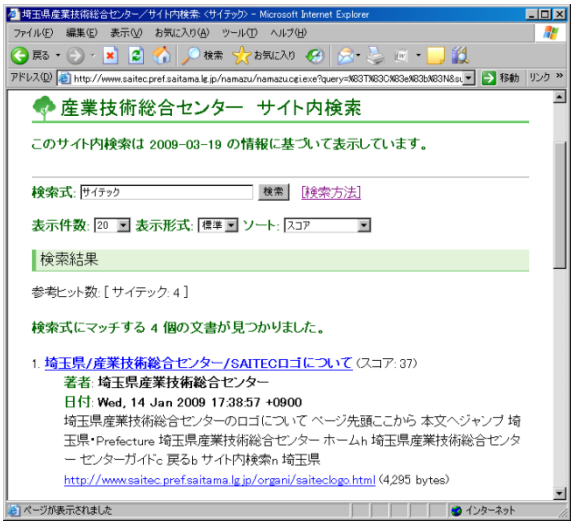

図１　スタイルシートを適用したサイト内検索のページ

2.1.3 PDF ファイルのインデックス化

Namazu は単独で HTML ファイルのインデックス化が可能である。しかし当 Web サイトには PDF ファイルで提供している情報も非常に多いため、PDF ファイルもサイト内検索の対象に含める必要がある。そこで、PDF ファイルからテキストデータを抽出できる Xpdf をサーバにインストールして Namazu と、PDF ファイルもインデックス化されるようにした。

特に、一部の PDF ファイルには内容を安易に複製されることを防ぐためにパスワードを設定してある。そのパスワードに対応できるようにインデックス化プログラムを改良することで、パスワードが設定された PDF ファイルもインデックス化できるようにした。このような修正が可能であることが、自前の構築することで細かな利点となる。

2.1.4 その他の対応

検索結果として該当する情報が見つからなかった場合に、技術相談のページを案内するよう結果表示のページを改良した。

2.2 ページ管理システムの開発

従来行ってきたアクセス解析とそれに基づく Web ページの改良において、いくつかの課題があった。特に大きな課題として、Web アプリケーションとデータベースによる動的なページの問題があった。一般の検索サイトで検索されるためには、HTML 化されたページを用意することが望ましいが、該当するページ数は 200 ページ以上あり、また今後のメンテナンスを考慮すると、個々のページを手作業で管理することは非効率であった。

そこで、ページ管理システムを開発し、以下の機能を持たせた。①機器データベース、及び試験データベースに登録されている情報から、あらかじめ用意したフォーマットファイルに沿って個々の HTML ページを作成する機能。②各 HTML ファイルにアクセス解析のためのタグが正しく設置され、サイト全体として矛盾がないかを確認する機能。③サイト内のリンクに矛盾が無いかを確認する機能。

システムの開発には、Windows 環境において幅広く使われ、使用実績も高い [4),5)]Visual Basic 系列の Visual Basic 2005 を使用した。

2.3 Web ページの改良

サイト内検索システムの動作確認の結果、及びページ管理システムによる HTML ファイルの確認結果から、Web ページの改良として以下のこ

とを行った。

2.3.1 不要なファイルの削除

Web サーバ上に、情報公開の役目を終えた古いページが残っている場合があった。例えばイベントの案内などがこれに該当する。本研究で構築したサイト内検索システムでは、コンテンツの内容と更新日の両方を同時に勘案して結果の順位付けをすることができない。そこでこのようなページを削除した。あわせて、今後も古いページを適宜削除するよう運用体制を見直した。

2.3.2 検索の対象としないファイルの設定

Web サーバ上にページを残しておく必要はあるが、サイト内検索で検索される必要のないページがある。例えば URL 変更の案内などがこれに該当する。これについては Namazu によるインデックス化の対象とならないよう設定した。具体的には、該当するページの HTML 内にインデックス化の対象としないためのタグを設置するとともに、Namazu によるインデックス化の際に、該当タグを含むページを対象に含めないオプションを指定した。

2.3.3 PDF ファイルのタイトル情報の管理

PDF ファイルもサイト内検索の対象となるよう必要な設定を行ったが、検索結果として表示されるタイトルは PDF ファイルのプロパティとして設定されているタイトルである。このタイトル情報は今まで意識的に管理されていなかったので、これを見直して適切なタイトルを付けるようにした。

2.3.4 試験情報、機器情報の HTML 化

開発したページ管理システムを使用して、試験情報及び機器情報を HTML 化したページを新たに追加した。

３　結果及び考察

3.1 ニーズ情報の収集と分析

3.1.1 キーワードの集計と考察

サイト内検索で使われた検索キーワードをアクセスログとして取得し収集した。１か月間に 152 個のキーワードがアクセスログから取得できた。使用されたキーワードを表１のように５つに分類し、分類ごとに検索された件数、そのうち検索結果として当 Web サイト内にヒットがあった件数、及びそのうち実際にページが閲覧された件数を調べた。また、ヒット率、閲覧率を以下のとおり算出して比較した。その結果を図２及び図３に示す。

ヒット率 ＝ ヒット件数 ÷ 検索件数　(1)

閲覧率 ＝ 閲覧件数 ÷ ヒット件数　(2)

表１　キーワードの分類

分類名	内容	具体例
一般技術用語	技術に関連があるが一般的に使われることの多い単語	熱
専門技術用語	試験や機器の名称など、具体性が高く特定の技術に関連がある単語	塩水噴霧試験
研修・講習関連	研修や講習に関連がある単語	セミナー
技術以外の用語	技術とは関係がないと推測される単語	入札
その他・不明	特定の企業名や数字などその単語だけでは目的の推測が難しい単語	（特定の企業名）

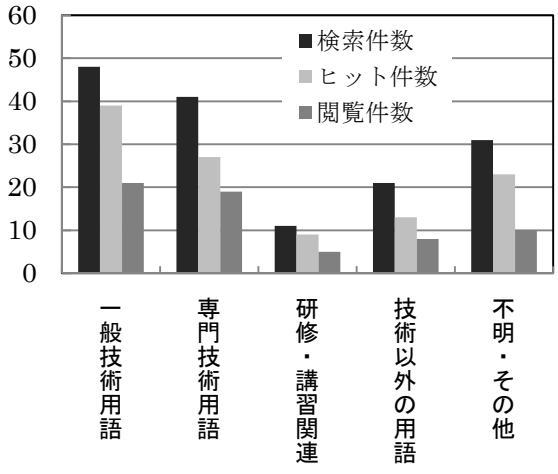

図２　分類ごとの使用件数等

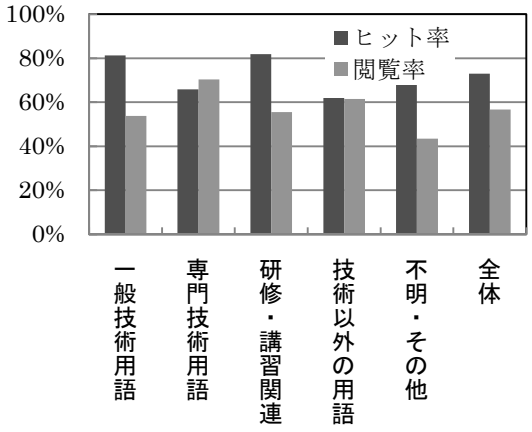

図３　分類ごとのヒット率等

一般技術用語、専門技術用語をあわせて技術用語が多く検索されていることが分かった。ヒット率を見ると、技術用語、研修・講習関連はヒット率が高く、それ以外は低いことが分かる。技術用語のうち、ヒット率は一般技術用語のほうがが高いが、閲覧率は専門技術用語のほうが高かった。一般技術用語は多くのコンテンツで使われているものの、内容が閲覧者の求める情報に一致することが少ないためと推測される。

一方、技術以外の用語やその他・不明の単語の検索もある程度の検索があることが分かったが、ヒット率や閲覧率は高くない。

個々のキーワードについて見たところ、特に突出して使われるキーワードは見られなかった。

これらの結果は、当センターの業務内容や当 Web サイトで提供している情報内容から妥当な結果と考えられる。

3.1.2 ヒットしない理由の調査と考察

次に、サイト内検索でヒットしなかったキーワードについて、ヒットしなかった理由を調査し、表２のように分類し、集計した。その結果を図４に示す。

情報・案内の未掲載に分類される理由が最も多かったが、当センターの業務に対する関連性の高さは様々である。当センターで直接対応できる内容かどうかは、個々のキーワードごとに精査する必要がある。関連性はあるものの当センターで直接扱っていない事項については、適切な他のサイトへのリンクを掲載するなどして閲覧者の利便性を向上させる必要があると考えられる。また、当センターで所有していない機器を検索している事例も見られたが、このような情報を蓄積することで、例えば保有機器の更新に役立てることが期待できる。

表２　検索にヒットしない理由の分類

分類名	内容
表記・用語等の問題	「データー」と「データ」のように表記や用語の違いによりヒットしなかったもの
情報・案内の未掲載	当センターに関連はあるが、キーワードが含まれていなかったものや、関係機関への案内などがなかったもの
業務外	当センターの業務と関連性のないキーワードによる検索
その他・不明	数字や略号など

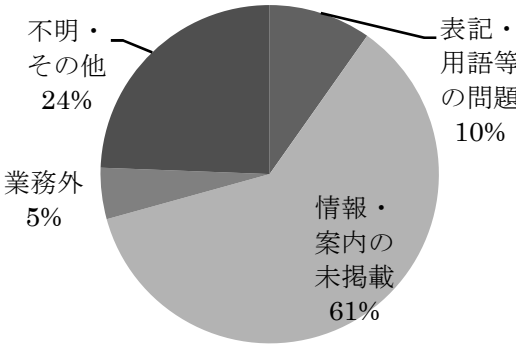

図４　ヒットしない理由と割合

3.1.3 キーワード情報の追加

表記・用語等の問題でヒットしない事例については、HTML にキーワードを追加した。例えば、アルファベット表記しかしていなかった名称に対してカタカナ表記で検索された事例では、HTML のキーワードタグにカタカナ表記を追加した。これにより、サイト内検索でヒットするようにした。

3.2 ページの改良の効果

依頼試験、開放機器のデータを掲載したページ

を HTML 化した結果、それらのページに含まれるキーワードが一般の検索サイトの結果ヒットしやすくなり、より多くの種類の単語にヒットすることが期待できる。そこで、半月間にヒットした検索キーワード数を先の研究時の値 [2)]と比較した。その結果を図５に示す。

図５から、より多くの単語に検索がヒットするようになったことがわかり、コンテンツの改良に効果があったと考えられる。

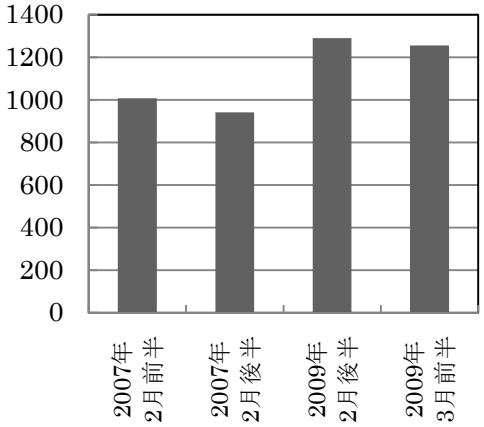

図５　検索キーワード数の変化

４　まとめ

ニーズ情報の取得、及びコンテンツの管理に関して以下のことを行った。

(1) 自前のサイト内検索システムの構築

オープンソースの日本語全文検索エンジン Namazu を利用して自前のサイト内検索システムを構築した。これによりサイト内検索で使われたキーワードをアクセスログから取得できるようになった。

(2) ニーズ情報の収集と分析

サイト内検索で使われたキーワードをニーズ情報として収集し、分類して検討した。現状は、推測される妥当な結果であった。収集と分析を継続することでコンテンツの改良につながると期待できる。

(3) ページ管理システムの開発

ページ管理システムを開発した。これにより、効率的な管理や Web ページの改良が可能となった。

(4) ページの改良

開発したページ管理システムを使って、機器データベース及び試験データベースの情報を HTML 化したページを新設した。その効果として、検索キーワードとしてヒットする単語が増えたことが確認できた。

参考文献

1）石井研二：アクセス解析ログの教科書,翔泳社,(2005)4

2）鈴木浩之，内藤丈資，秋山稔，菊池和尚，大野哲治，匂坂剛：WEB ページのアクセスに関する分析とアクセスの質の向上に関する研究，埼玉県産業技術総合センター研究報告，**5**，(2007)39

3）全文検索システム Namazu, http://www.namazu.org/，2008.6.18

4）安藤昌弘，匂坂剛，筒井大輔：産業技術総合センター情報システムの開発，埼玉県工業技術センター研究報告，**4**，(2002)33

5）石田聡，安藤昌弘，鈴木浩之，本多春樹，小沢一郎：インターネットを用いた開放機器の予約状況公開システムの構築：埼玉県産業技術総合センター研究報告，**6**，(2008)46