インターフェイスの街角– WebGlimpse

増井 俊之

大規模な Web サイトでは、検索機能を備えるのが当然になりつつあるようです。ユーザビリティ工学および Web デザインの大家である Jakob Nielsen 氏のコラム「Alertbox」[1]によれば、Web ユーザーの約半分は〝検索中心主義 (search-dominant)〟であり、ページ上に検索用のダイアログをみつけるとすぐに検索を実行するそうです[1]。Nielsen 氏は、200 ページを超えるようなサイトはかならず検索機能を備えるべきであり、とくに重要な点として、

- すべてのページに検索ボタンを配置する
- サイト全体の検索ができるようにする
- AND/OR 検索などは補助機能として用意しておけばよい

といった項目を挙げています。

私が公開しているページ[2]もかなり大きくなってきたので、最近になってようやく 1999 年 11 月号で紹介した〝お手軽検索システム〟を使った検索機能を導入しました。

検索用のダイアログを付けると、それだけで一人前のページになったような気がして嬉しいものです。自分が作ったページですから、内容はよく知っているはずなのですが、それにもかかわらずけっこう便利に使っています。〝お手軽検索システム〟はソースが短くて簡単に導入できるのが取り柄であり、私の Web ページ程度の規模であれば十分実用的です。しかし、大規模なサイトで利用するとなると、

- 複数のキーワードが指定できない
- キーワードのスペルが間違っていると検索できない
- 実行速度が遅い

といった点が問題になるので、もっと複雑で高速な検索システムが必要でしょう。

最近の検索システム

World-Wide Web とサーチエンジンが隆盛であるためか、このところ Information Retrieval（情報検索）に関する研究も活況を呈していて、新しい書籍が数多く出版されています（図 1）。

たとえば、テキスト検索アルゴリズムで有名なチリ大学の Ricardo Baeza-Yates は、近年の情報検索に関する研究をひろくまとめた『Modern Information Retrieval』[2]を出版し、そのなかの検索インターフェイスや視覚化技法についての章を Web 上で公開しています。情報圧縮や機械学習の分野では、ワイカト大学の Ian H. Witten が、大量のテキスト情報の圧縮と検索をテーマとする『Managing Gigabytes』[6]の第 2 版を出版しました[3]。一方、各種のビジュアルな情報検索システムを開発しているピッツバーグ大学の Robert R. Korfhage も、『Information Storage and Retrieval』[3]という本を出版しています。

最近は、Web 上でも多種多様な検索システムやサービスが提供されています。SearchTools.com は、いろいろ

1 これに対し、つねにリンクをたどって情報をみつけようとする〝リンク中心主義者 (link-dominant)〟は全体の 20%で、残りはその中間だそうです。検索機能がそれほど頻繁に使われているのだろうか、という気もしますが、1997 年当時の米国の話なので現在の日本とは状況が違うかもしれません。Alertbox は、Web ページで公開されている、ユーザビリティ工学に関する隔週刊のコラムです。

2 http://www.csl.sony.co.jp/person/masui/

3 この本で紹介されているシステム〝MG〟は、http://www.mds.rmit.edu.au/mg/で公開されています。

図 1　最近出版された情報検索分野の書籍

な検索サービスやツールの紹介と評価をおこなっています。日本語が使える検索システムに関しては、京大の馬場 肇さんが Web ページ上で詳しいサーベイを公開しています。

SearchTools.com では約 100、馬場さんのページでは約 70 のシステムが紹介されています[4]。これほど多くなると、どの検索システムが目的に適しているかをみきわめるのも容易ではありません。

とはいうものの、できるかぎり検索対象の性質に合った検索システムを選びたいものです。たとえば、個人のファイルやメールなどが対象の場合には、検索に多少時間がかかってもインデックスは小さいほうがよいでしょう。この種のファイルには綴りに間違いがある可能性が高いので、曖昧検索機能があったほうが便利です。しかし、同義語などを利用した意味検索はとくに必要ないと思います。

一方、不特定多数のユーザーが訪れる企業の Web サイトなどでは、インデックスが大きくなっても高速な検索が求められます。漠然とした概念からでも検索可能なシステムのほうがよいので、キーワード検索だけでなく、意味検索もおこなえるほうがよいでしょう。

これらの点を考えあわせると、次のような項目が選択のポイントになるのではないでしょうか。

- 自然言語処理
  - 形態素解析をおこなうか
  - 意味処理をおこなうか
  - 単語の重要度を利用するか
- インデックス
  - 完全なインデックスを使うか
  - シグネチャのような簡易インデックスを使うか
  - まったく使わないか
- パターンマッチの手法
  - 正規表現を許すか
  - 曖昧表現を許すか
  - AND/OR 検索を許すか

今回は、日本ではあまり利用されていませんが、曖昧検索などの興味深い特徴をもつシステム〝WebGlimpse〟を紹介します。

4 日本の Web サーバーでは、フリーの検索システムである Namazu や Freya などがひろく使われているようです。

agrep と Glimpse

WebGlimpse [4]は、ファイル検索システム Glimpse [5]を Web 検索用に拡張したものです。Glimpse は、Udi Manber と Sun Wu がアリゾナ大学在籍中に開発した曖昧検索プログラム agrep[7]をファイルシステム全体の検索に拡張したものです。

WebGlimpse/Glimpse のソースコードは公開されていますが、残念ながらフリー・ソフトウェアではなく、教育機関以外で使用する場合はライセンスが必要です[5]。

agrep

1998 年 1 月号で紹介したように、agrep は grep に曖昧検索などの機能を加えて高速化したものです。agrep は UNIX 標準の grep コマンドと同じように使えるだけでなく、引数で曖昧度（何文字のミスマッチを許すか）を指定することができます。

たとえば、英単語辞書に対して曖昧度を指定せずに起動した場合は、標準の grep と同じように完全なパターンマッチをおこなうため、

```
% agrep masui words
%
```

のように何も出力されません。ここで、引数 –1 を指定して 1 文字誤りを許すようにすると、

```
% agrep -1 masui words
massif
massive
mastic
mastiff
swimsuit
%
```

5 教育機関以外のユーザー向けに、30 日間の試用版も用意されています（内容は教育機関向けのものと同じです）。ライセンス料などの詳細は、http://webglimpse.net/licensing.html を参照してください。

といったように文字列 `masui´に近い単語のリストが得られます。

agrep では、`+´以外のほぼすべての正規表現に対応しています。また、`.*´の代わりに `#´を使うことができます。

```
% agrep '(foo|bar)#(foo|bar)' words
barbarian
barbaric
barbarism
barbarous
barefoot
%
```

さらに、AND/OR 検索もサポートしています。複数のパターンを `;´で区切って指定すると、それらのパターンをすべて含む行だけが出力されます（AND 検索）。また、`,´で区切って指定するといずれかのパターンを含む行が出力されます（OR 検索）。

```
% agrep 'foo;bar' words
barefoot
% agrep 'foo,bar' words
afoot
Aldebaran
archfool
bar
barb
.....
%
```

このように、agrep は grep よりも検索システムに適した機能をもっています。

Glimpse

Glimpse は agrep のアルゴリズムを利用してファイルシステムの曖昧検索を可能にするシステムです。agrep では高速化のためのさまざまな工夫が凝らされており、多くの場合は grep よりも高速な検索が可能です。それでも、検索対象のすべてのファイルに対して agrep を適用すると膨大な時間がかかります。そこで、Glimpse ではあらかじめ **glimpseindex** コマンドで検索対象のファイルに含まれる単語のインデックスを作成します。検索を実行する **glimpse** コマンドは、最初にこのインデックスを検索し、その後にファイル本体を検索するという 2 段構成になっています。この手法は、11 月号で紹介したお手軽検索システムで、単語がファイルに含まれるかどうかを高速に判断するために、シグネチャ・ファイル（簡易インデックス）を使ったのと似ています。

ファイルのどの位置にどの単語が含まれるかを示す正確なインデックスを作成しておけば検索は高速になりますが、インデックスは巨大になってしまいます。Glimpse では、単語がファイルのどのあたりに含まれるかというおおよその情報だけを簡易インデックスに含めることによって、インデックスのサイズを全ファイルの 2〜3%に収めるようにしています。検索の第 1 段階の、簡易インデックスを対象とした検索にも agrep の高速化アルゴリズムが使われています。

Glimpse の入手

Glimpse は下記の WebGlimpse のページから入手できます。

- http://webglimpse.org/

執筆時点の最新版は 4.12.6 で、Solaris や HP-UX、AIX、IRIX、OpenBSD、Linux などの OS に対応しています。

オリジナルの Glimpse は日本語に対応していませんが、インテック・システム研究所の石田 茂さんによる日本語パッチ[6]を適用すれば日本語が扱えるようになります。

■ パッチは新しいバージョンにもあたる？

インデックス・ファイルの生成

さきほども述べたように、Glimpse を利用するにはまず glimpseindex コマンドでインデックス・ファイルを作成する必要があります。引数には、検索対象となるディレクトリを指定します（図 2）。

この図からも分かるように、インデックス・ファイル.glimpse_index、ファイル名のリストである.glimpse_filenames など、`.glimpse´で始まるいくつかのファイルがホーム・ディレクトリに作成されます。この例では、約 50MB のファイルに対して合計 1.5MB 程度のインデックスが作成されています

glimpseindex では、デフォルトで `.Z´などの拡張子をもつファイルはインデックスの対象から除外されます。さらに、ホーム・ディレクトリに、

6 ftp://www.isl.intec.co.jp/pub/person/ishida/freeware/glimpse/

図 2　glimpseindex コマンドの実行

```
% glimpseindex /user/masui/DOC

This is glimpseindex version 4.1, 1997.

Indexing "/user/masui/DOC" ...

Size of files being indexed = 50931557 B, Total #of files = 10093

Index-directory: "/mnta/masui"
Glimpse-files created here:
-rw-r--r-- 1 masui SonyCSL      48 2月 25日 18時39分  .glimpse_exclude
-rw-r--r-- 1 masui SonyCSL       3 2月 25日 18時31分  .glimpse_exclude~
-rw------- 1 masui SonyCSL  461602 2月 26日 15時07分  .glimpse_filenames
-rw------- 1 masui SonyCSL   40372 2月 26日 15時07分  .glimpse_filenames_index
-rw------- 1 masui SonyCSL       0 2月 26日 14時46分  .glimpse_filetimes
-rw------- 1 masui SonyCSL 1164776 2月 26日 15時07分  .glimpse_index
-rw------- 1 masui SonyCSL    6100 2月 26日 15時07分  .glimpse_messages
-rw------- 1 masui SonyCSL     868 2月 26日 15時07分  .glimpse_partitions
-rw------- 1 masui SonyCSL   80296 2月 26日 15時07分
.glimpse_statistics
%
```

図 3　glimpse の実行

```
% glimpse Manber
/user/masui/DOC/bib/m/a/Manber:Glimpse: Author: Udi Manber
/user/masui/DOC/bib/m/a/Manber:WebGlimpse: Author: Udi Manber
/user/masui/DOC/bib/w/u/Wu:agrep:  Udi Manber
/user/masui/DOC/paper/Pen/CHI98/CHI98CameraReady.bib: Author = {Sun Wu and Udi Manber},
/user/masui/DOC/UnixMagazine/0004/if0004.tex: \citation{Manber:WebGlimpse}
/user/masui/DOC/UnixMagazine/0004/if0004.tex: \citation{Manber:Glimpse}
.....
%
```

```
.glimpse_exclude
```

というファイルを用意し、不要なファイル名のパターンを記述すれば、それらのファイルは検索の対象になりません。

```
% cat $HOME/.glimpse_exclude
~$
\,v$
/#
\.log$
\.eps$
\.ps$
\.dvi$
%
```

検索の実行

glimpse コマンドを実行すると、glimpseindex によって作成されたインデックスを使って検索がおこなわれます。

図 3 に示した例では、〝Manber〟という文字列を含む文献ファイルや文書が検索されています。

Glimpse では、agrep と同様な曖昧検索が可能で、パターンについても同じものが使えます。

```
% glimpse -1 Manber
.....
% glimpse 'foo;bar'
.....
```

glimpse コマンドに –F オプションを付け、検索対象のファイル名を指定することもできます。この機能を利用すれば、特定のディレクトリ以下のファイルのみを検索対象にするといった指定が簡単におこなえます。

さらに、検索対象のディレクトリに移動せずに、grep と同様のファイル検索をおこなうこともできます（図 4）。

図 4　検索対象のディレクトリに移動せずに検索

```
% glimpse -F paper Manber
/user/masui/DOC/paper/Pen/CHI98/CHI98CameraReady.bib: Author = {Sun Wu and Udi Manber},
%
```

ファイル名によるフィルタリングは、agrep を利用してファイル名のリスト (~/.glimpse_filenames) と指定したパターンとを照合しておこなわれます。したがって、ファイル名に ``–v パターン″と指定すれば、指定したパターンを含まないファイルのみを検索対象とすることもできます。

glimpseindex によるインデックス作成には時間がかかるので、cron などで起動するとよいでしょう。

WebGlimpse

WebGlimpse は、Glimpse を検索エンジンとして Web サイトの検索をおこなうシステムです。執筆時点の最新版は 1.7.7 で、Glimpse のところで紹介した Web ページから入手できます。

前節で説明したように、Glimpse では特定のディレクトリ以下のインデックスを作成して検索をおこないます。WebGlimpse には、あるページから一定のリンク距離内にあるページや最近更新されたファイルに限定して検索できるという特徴もあります。一般的な検索エンジンでは、Web サイト全体あるいはサイト単位での検索は可能ですが、特定のページの ``近く″のページだけを対象にすることはできません。WebGlimpse のインデックス作成方式を使えば、関連が深いと思われるページだけを対象にできるので効果的な検索が可能になります。

現在、キーワードによるフィルタリングとブラウジングを融合するさまざまな検索手法の研究が進められています。2 月号では、このような試みの 1 つとして「LensBar」を紹介しましたが、WebGlimpse なども同じ目的をもつ方式と捉えてよいのではないでしょうか。

WebGlimpse のインストール

WebGlimpse のパッケージの大半は、Glimpse を呼び出す Perl スクリプトから構成されています。

リンク先のページを取得するプログラム httpget を作成したり、ディレクトリなどの設定をおこなうために、最初に **wginstall** というシェル・スクリプトを起動します (このスクリプトは、perl コマンドのパスを調べて wginstall.pl という Perl スクリプトを起動します)。

図 5　WebGlimpse による検索

表示される質問に答えるかたちで検索対象のディレクトリや URL などを指定していくと、glimpseindex によってインデックスが作成され、glimpse を CGI として起動するための wgindex.html ファイルが作成されます。

WebGlimpse の実行例

図 5 は、wgindex.html による WebGlimpse の検索画面です。表示オプションなどのほかに、キーワードの曖昧度が指定できるといった特徴があります。

この例では、検索対象である LEGO MindStorms メーリングリストの記事アーカイブから、``cibermastar″をキーワードとして検索しようとしています。CyberMaster というのはヨーロッパで販売されている LEGO のロボットキットの名前ですが、このように綴りを間違えたままキーワードを指定しても、曖昧度を指定しておけば正しい検索結果が得られます (図 6)。

この例では、検索ダイアログは図 5 のように独立したページになっていますが、検索される各ページに自動的に検索ダイアログを埋め込むオプションも用意されています。

おわりに

今回は、曖昧検索システム Glimpse と、それを Web 検索に適用した WebGlimpse を紹介しました。インストールにやや手間がかかることと、そのままでは日本語が

関連 URL

WebGlimpse	http://webglimpse.org/
SearchTools.com	http://www.searchtools.com/
日本語全文検索エンジンソフトウェアのリスト	http://www.kusastro.kyoto-u.ac.jp/~baba/wais/other-system.html
Ricardo Baeza-Yates	http://www.dcc.uchile.cl/~rbaeza/
Modern Information Retrieval	http://www.sims.berkeley.edu/~hearst/irbook/
Ian H. Witten	http://lucy.cs.waikato.ac.nz/~ihw/
MG	http://www.mds.rmit.edu.au/mg/
Namazu	http://www.namazu.org/
Freya	http://www.ingrid.org/ja/project/freya/
Alertbox	http://www.useit.com/alertbox/

図 6　検索結果の表示

使えないという問題があり、Web 検索システムとして現状で最善の選択とはいえないかもしれません。ただし、曖昧検索などの効果を体験するには有効でしょう。

Glimpse を開発した Manber 氏は、現在は Yahoo!のチーフ・サイエンティストとして活躍しているそうです。しかし、Yahoo!の検索エンジンでは曖昧検索が使えないところをみると、Glimpse の技術は Yahoo!では利用されていないようです。

Glimpse を開発したときのエピソードとして、どうしてもみつからなかった知人の住所が Glimpse を使ったらすぐに分かったという話が文献[5]に書いてありました。その住所は論文募集のメールに記してあったため、住所録ファイルを収めたディレクトリをいくら探してもみあたらなかったというわけです。個人情報の管理では、この話のように、「情報を見たのは確実だが、どこで見たかが思い出せない」ことが多いような気がします。数多くの検索システムのなかから、用途に適したものを組み合わせて上手に活用することが重要だと思います。

（ますい・としゆき　ソニー CSL）

[参考文献]

[1] Jakob Nielsen, *Search and You May Find*, July 7, 1997 (http://www.useit.com/alertbox/9707b.html)

[2] Ricardo Baeza-Yates and Berthier Ribeiro-Neto, *Modern Information Retrieval*, ACM Press Series, Addison-Wesley, May 1999

[3] Robert R. Korfhage, *Information Storage and Retrieval*, John Wiley & Sons, June 1997

[4] Udi Manber, Mike Smith and Burra Gopal, "Webglimpse — combining browsing and searching", In *USENIX Technical Conference*, pp.195–206, January 1997

[5] Udi Manber and Sun Wu, *GLIMPSE: A tool to search through entire file systems*, Technical Report TR 93-34, Department of Computer Science, The University of Arizona, Tucson, Arizona, 1993

[6] Ian H. Witten, Alistair Moffat and Timothy C. Bell, *Managing Gigabytes: Compressing and Indexing Documents and Images*, Van Nostrand Reinhold, New York, 1999

[7] Sun Wu and Udi Manber, "Agrep — a fast approximate pattern-matching tool", In *Proceedings of USENIX Technical Conference*, pp.153–162, San Francisco, January 1992