資料編３　キーワード検索システム 運用の手引き

Ver.2.00　H16.2.29

はじめに

検索の仕方は、「入力しかたは？」と「やり直し検索のヒント」の画面で表示されます。

実際に「or検索」や「前方一致検索」などの操作をした上で、本「手引き」をご活用ください。

①検索するキーワードを入力した際に使用するKAKASI辞書と、既に蓄積されているインデックスデータを作成したKAKASI辞書が同一である必要があります。本システムを使用して新たに登録語を追加した場合は、登録語の違いによって切り分け方が異なる場合が生じ、過去のデータが検索できないこともあります。

②対象となるデータのフォルダ名やファイル名が、日本語の表記でも利用できます。しかし、フォルダ名やファイル名に使用されている漢字や記号の中に、日本語EUCコードにない漢字（外字を含む）があると、「ページが見つかりません　HTTP404ファイル未検出」などと表示されて利用できません。（各ファイル内のテキストデータに、日本語 EUC コードにない漢字（外字を含む）がある場合は支障ありません。但し、検索結果の見出し文に「　」のように表示されてしまいます。また、その日本語EUCコードにない漢字（外字を含む）を入力して検索することはできません。）

③通常は、最長一致でマッチしたキーワードのみを検索します。そのため、検索対象のテキスト文が「群馬県太田市」で、インデックス登録されていると、「群馬」や「太田」と入力しても、検索結果に表示されません。これは、分かち書き辞書の問題ですので、検索者の使用する言語力、検索経験などに応じた分かち書き辞書を使用したり、作成したりする必要があります。（本システムは、対応済みです。）

④検索しようとするデータに文字や文が存しないデータは、ファイル名のみを検索の対象をすることで、検索することができるような仕様になっています。

Ⅰ　サーバ機のメンテナンスについて

Win9x系のOSは、サーバ機用に開発されたものではありません。特に、システム管理のためにメモリを消費するために、時間の経過に伴ってフリーズ状態になることがあります。これらを回避するためには、下記の表に基づいて、一旦コンピュータを電源OFFにし、新たに電源を投入するようにしてください。

Win95,Win98,Win98SE	WinNT	Win2000,XP
ほぼ毎日の電源OFF	１ヶ月に１回以上の電源OFF	必要に応じての電源OFF

Ⅱ　設定ツールの【sachi.exe】をインストールすると、インターネット上で配布されている【KAKASI】にはない機能－分かち書き辞書への登録を事後的に追加・修正できる機能が利用できます。

まずは、その機能が使えるようにしてみましょう。

①デスクトップ上に「辞書追加修正ツール」のアイコンがあるかをを確認してください。（もし、ない場合はC:\kakasi\share\kakasiを開いて、辞書マークのアイコンをデスクトップ上へコピーしてください。）

②「辞書追加修正ツール」のアイコンをダブルクリックすると、２つのアイコンが表示されます。

③「追加修正シート」のアイコンをダブルクリックすると、メモ帳が開きます。そこへ登録したい語を「ひらがな(半角スペース)漢字」のように、登録の組ごとに改行しながら記します（下記の記述例を参照）。

```
ぐんまけんそうごうきょういくせんたーきょういくじょうほう 群馬県総合教育センター教育情報
えぐちきち 江口きち　　←先頭文字が漢字であれば、登録できます。
あめりかんどりーむ アメリカンドリーム　←カタカナは登録しなくても、検索処理できます。
かえるのがっしょう かえるの合唱　　←先頭の文字が漢字でない場合は、登録できません。
```

但し、日本語EUCコードにない漢字（外字を含む）は、記述できても登録できません。右側の漢字を含ん

だ言葉を分かち書き辞書に追加しますが、先頭の一文字は漢字でなければなりません。
（これは、Namazuシステムの内部では、EUCコードで処理がなされるためです。また、登録してしまっても処理がなされないだけで、システムの不具合が起きないことを確認しています。）
④追加したい語の記述を終えたら、＜ファイル(F)＞をクリックして、表示される＜上書き保存(S)＞をクリックし、上書きした後にダイアログの右上の×をクリックして、メモ帳を終了します。
⑤今度は「辞書追加修正ツール」フォルダ中の「新辞書作成ボタン」をダブルクリックします。黒いDOS画面が表れて、10秒ぐらいで自動的に終了します。これで、先に記した言葉が分かち書き辞書に追加されました。
⑥もし、後で追加した言葉を修正、削除したくなったら、②～④の操作をして、該当する言葉を修正、または行ごと削除してください。（万一、登録語に多くなった場合は、メモ帳の＜検索(S)＞ボタンを押して、該当の登録語を探すことができます。）
しかし、最初からKAKASI辞書にある言葉は、修正、削除できません。あくまで、追加した登録語のみに有効です。そのため本システムでは、最小構成語で作成した辞書を標準として採用しています。
さらに、作成した登録語で検索できるようにするには、最初からインデックス化をやり直す必要があります。但し、対象のデータが判明していれば、最新の日付に上書き更新し、タスクを開いて実行するだけで、その登録語を反映した検索をすることが可能です。
なお、画像データファイルなども、内部照合するインデックス方法をとっていますので、10GBのデータの場合、およそ10時間～15時間の時間を要します。(USB1.0等で外部ハードディスクを設置している場合は、さらに数倍の時間を要します。)

Ⅲ　ファイル名を付ける際の留意事項（特に、テキストデータがない画像ファイル等の命名方法について）
「半角(または全角)の空白」「記号」「英数字」「ひらがな」「カタカナ」を、文字間にうまく挿入することが、記された言葉を正しくインデックス化できるような分かち書きに切り分けるコツです。
①まず、フォルダ名、ファイル名に外字などの日本語EUCコードにない特殊な文字を使用すると、ネットワークからブラウザで開く際に、データファイルのある場所を特定できなくなります。ファイル名のテキスト文も認識されずに、文字化けします。（但し、＝「げたマーク」とよばれる記号に置き換えられた場合は、使用可能です。）結果として、URIがリンクしているところをWebサーバから認識できず、検索結果が表示されてもファイルを開くことができません。インデックス化をしている時に文字化けしているフォルダやファイルが原因です。前後に渡って化けてしまうので、前後の正しく表示されたファイルの情報を参考にして修正してください。どんな文字が化けるかについては、例えば「　橋」の「　」（いわゆる「はしご高」と呼ばれる和製漢字）は文字化けしますが、「關口」の「關」は問題なく使用できるなど、日本語文字コードの問題は複雑です。詳しくは日本語EUC文字コード表で確認してください。
検索結果で見出し文に　　や＝やが見られるのは、Namazu内部ではEUCコードで処理されているためで、その文字以外の検索は可能です。また、そのファイルを実行するのにも支障はありません。もし、目障りであるならば、検索結果表示画面で「見出し文の長さ」のウィンドウ右側の▼をクリックして「短く」を選択してください。一時的に表示されなくなります。
また、根本的に表示しないようにするには、データドライブのindexA\template\NMZ.result.normal.jaファイルをEUC対応エディタで開いて、[<dd>$ {summary}]と記された行を削除します。
さらに、ファイルの中味が空（データが無い）の場合、ファイル名が付いていても、インデックス化の段階で、除外されるような仕様です。したがって、検索することはできません。
②ファイル名を「001～」などと、ナンバー付けするアプリケーションがありますが、インデックス化は英数字が「ISO9660」のように一体化してなされるために、「05GUNMA」などは、「GUNMA」とキーワード入力しても検索できません。特別の意味がないときは、数字と英字とを続けて記さないようにしてください。この場合は「05 GUNMA」と半角のスペースを空けることにより、切り分けることができます。

③ひらがなも空白や記号で分離させれば、検索することが可能です。「うどんづくり」のように、ひらがなが続く場合は、「うどん･づくり」「うどん(作り方)」のように、空白(半角、全角どちらでも可)を入れたり、中点や括弧などで区切ったりするようにしてください。また、「なすの花」では、ひらがなの「なすの」でインデックスされてしまいますので、「ナスの花」と記述すると、よいでしょう。続けて「なす・茄子・なすび・茄・ナスビ」などとシソーラス化をすれば、「ナス」の同義語から全て検索できるようになります。

但し、ファイル名には、「\/:;*?"<>|」の記号は使えません。同様に「ここではきものをぬぐ」のような短い文章は「ここで、はきもの をぬぐ」なのか、「ここでは、きもの をぬぐ」などのように読点や空白を入れるなどの工夫が必要です。(奈良先端科学技術大学院大学の自然言語処理学講座からフリーソフトとして配布されている【Chasen】辞書は、ひらがなやカタカナのみで記述された文章の形態素解析に対応しています。Windows上で使用するためには、Cygwinなどを使用してLinux版のNamazuシステムをインストールし、UNIX版の【Chasen】を分かち書き辞書として導入する必要があります。)

④「１年数学」と「年」「数学」の二つの意味の漢字が続く時は、「数学１年」「１年の数学」のように記すことにより、間に記号がなくても「英数字、ひらがな及びカタカナは区別」という法則にしたがって切り分けられます。

「岡田家庭訓往来」のように漢字のみが続く場合は、「岡田家・庭訓往来」(または、「岡田家の庭訓往来」)のようにしないと、「岡田」「家庭」「訓」「往来」のように切り分けられてしまうので、記号や空白などを上手に使って、言葉本来の意味が適切に反映されるように切り分けてください。

⑤カタカナは、そのまま切り分けられてインデックス化されますが、小文字で表される促音・拗音を正しく記したり、「エントランススペース」のように長く記されるものは、「エントランス スペース」のように分けて記したりします。また、「ヴィジュアル」「ビジュアル」など、両方の表記が使われるものについては「ヴィジュアル(ビジュアル)」のように、どちらでも検索できるようにしておくとよいでしょう。

⑥ファイルの種類については、「mpg」「群馬県　JPG」のように拡張子名を入れて、単一語検索、and検索などをすることができるので、特別の意味がない限り、ファイル名に記述する必要はありません。

⑦データベースとして構成する際は、ファイル名に表記する際の約束ごとを決めておくと、後の管理が楽になります。例えば、「視聴覚 液晶プロジェクター使用状況」「城西町 区民センター文化祭」のように、そのデータの所属を明示した後、ファイル内容を記すなど、記述の仕方を決めておくと、特定の観点（この例では「校務分掌」や「地区名」）から、検索することが可能となります。ぜひ校内のルールを決めて、誰でも簡単に検索できるようなデータベース化を推進してください。

Ⅳ　各校のコンテンツ集の作成について

G-TaKコンテンツ集と同様に、各校で作成したコンテンツを階層構造のフォルダに整理（特に、キーワード検索するだけの対象ならドンブリ勘定でよいのですが…）し、サーバ機のデータドライブにコピーするだけで、キーワードによる検索ができるようになります。

①データドライブの[\dataA]以下に、「○○小」フォルダを作成し、その中に各校のコンテンツデータをコピーします。(Ⅳの「ファイル名の付け方」に基づいて、写真なども１枚ずつファイル名を付けてください。)

②タスク(「導入編」Ⅳ(9)⑨参照)を実行させるだけで、コンテンツデータの検索ができるようになります。

Ⅴ　Appendix（追録）

[既知の問題]　なお、②と③の問題は故意に発生させたものです。

①サーバ機がインデックス化をしている際に、キーワードを入力して検索したり、クライアント機から検索のリクエストがあると、一瞬ロックした状態になりますが、特に問題はありません。

②もし、インデックス化の最中に障害が起きて、タスクからインデックス化ができなくなった場合は、データドライブの[\indexA]のフォルダの中から、[NMZ.lock1]、[NMZ.lock～]（～は数字）などと記されたファイルを、すべて削除してください。その後、再度タスクを実行させてインデックスを作成します。

③②の処理でも解決しない場合は、過去のインデックス化したデータに破損が及んでいる可能性があります。その場合は、[\indexA]のフォルダの[template]フォルダ以外のすべてのフォルダを削除してください。([template]フォルダは、検索表示画面の設定ファイル等があるので、絶対に削除しないでください。)

その後、再度タスクを実行させてインデックス化しますが、すべてのインデックス化をやり直すために、1GBあたり１～1.5時間の処理時間が必要になります。(この時間は、外付けハードディスクの接続速度、サーバ機にインストールされている、他のアプリケーションの設定などにより、大きな差異が見られることが判明しています。)

④インデックスの最中に「システムリソースが不足しています。…」と警告表示が出た場合には、<OK>ボタンを押さずに、右上の×印をクリックしてウィンドウを閉じてください。そのまま、インデックスを続行することができます。(<OK>ボタンを押すと、処理が中止されてしまいます。なお、後ほどWindowsの設定で仮想メモリの最大値を増やしてやる必要があります。)

⑤クライアント機からダウンロードしたファイルを書き換える場合には、サーバ機の被検索ファイルのあるフォルダを「共有する」に設定しておく必要があります。

例えば、Windows98の場合は該当フォルダをポイントし、右クリックした後に表れるメニューから、<共有(H)>をクリックします。表示された「○○のプロパティ」から、<共有>タブをクリックします。表示された「共有する(S)」のラジオボタンをクリックし、反転して有効になる「フルアクセス(F)」のラジオボタンをクリックすると、完了します。

クライアント機からアプリケーションのファイルを書き換える際は、ブラウザのウィンドウの中に既にアプリケーションが起動していますので、ファイル名と保存場所を指定して書き換えてください。

特に「一太郎」は、注意が必要です。Ver.10 以前の場合は、ブラウザのウィンドウ外に起動しますので、メニューバーの「ファイル(F)」等を使用して書き換えることができます。Ver.11 ～ 2004 については、ブラウザのウィンドウ内に起動しますが、メニューバーの「ファイル(F)」を使用して書き換えることができません。ウィンドウ内をダブルクリックすると、ブラウザのウィンドウ内とは別の「一太郎」（こちらが本来の「一太郎」です。）が起動しますので、ファイル名と保存場所を指定して書き換えます。

なお、書き換えられては困るデータのあるフォルダは、絶対に共有設定しては、いけません。

⑥サーバ機に圧縮ファイル解凍ソフトがインストールされ、実行ファイルとして関連づけられていないと、インデックスできなかったり、クライアント機からダウンロードできないことがあります。使用する解凍ソフトによって結果が異なりますが、フリーソフトの【range141.exe】(解凍レンジ)のみを使用している場合は、正常に動作することを確認しています。

⑦クライアント機のブラウザ（インターネット エクスプローラ5.5など）に、使用するアプリケーションのファイルが関連づけられていないと、検索したファイルを開けないことがあります。（ブラウザが、標準設定のままならば開くことができます。また、ファイル自体のダウンロードは可能です。）ブラウザの詳細設定で＜既定の設定に戻す＞ボタンをクリックして、標準設定にしてください。

⑧OSのみをインストールしただけでシステムを構成すると、データ内にテキスト文のあるファイルは、インデックスに対応していない旨のメッセージが表示され、インデックスできません。インデックスする際に、Wordのファイルとして読み出すことへの関連付けがなされていないためです。最も簡単な対処方法は、Word2000以上のバージョンをインストールすれば、すべて解消します。正常にインデックスできるようになったら、Wordをアンインストールしても支障なく、動作します。

⑨設定ツール【sachi.exe】は、G-TaK検索用として標準設定するために、ファイル名が［index～.html］となるデータを検索できません。タスクの「実行するファイル名」の記述のうち、「|index*.html」を削除して実行すると、検索できるようになります。但し、G-TaKの利用では「目次」部分が数多く表示されます。

⑩サーバを設定せず、スタンドアローンで使用する場合は、Namazuアドオンの【srchs092.exe】を利用してください。入手先→ http://www.syam.net/library/search-s/

⑪「すーほ」などの語句は、「ー」の部分がカタカナとして取り扱われ、検索できないことがあります。