■デ変研 TF ライブラリと連携した テキスト抽出処理

検索したいドキュメントが、MS-Office / PDF / 一太郎などのファイルフォーマットであった場合、デ変研TFライブラリと連携し、対象フォーマットを解析して、テキスト情報を取出し、そのテキスト情報をインデックスに登録することができ、高速検索処理を実現しています。
以下のファイルフォーマットに対応しています。
※詳しくは「デ変研TFライブラリ」をご参照ください。

■対応文書

Microsoft Word
97 / 98 / 2000 / 2002(XP) / 2003 / 2007 / 2010 / 2013 / 2016 / 2019
Microsoft Excel
97 / 2000 / 2002(XP) / 2003 / 2007 / 2010 / 2013 / 2016 / 2019
Microsoft PowerPoint
97 / 2000 / 2002(XP) / 2003 / 2007 / 2010 / 2013 / 2016 / 2019
Microsoft Visio
2002(XP) / 2003 / 2007 / 2010 / 2013 / 2016 / 2019
Microsoft Word for Mac
98 / 2001 / 2004 / 2008 / 2011 / 2016 / 2019
Microsoft Excel for Mac
98 / 2001 / 2004 / 2008 / 2011 / 2016 / 2019
Microsoft PowerPoint for Mac
98 / 2001 / 2004 / 2008 / 2011 / 2016 / 2019
Microsoft XPS 1.0
JustSystems 一太郎
Ver.5-Ver13 / 2004 - 2019
Adobe Systems Acrobat
4.0 / 5.0 / 6.0 / 7.0 / 8.0 / 9.0 / X / XI / DC
(一部未対応の形式があります)
PDF
1.2 / 1.3 / 1.4 / 1.5 / 1.6 / 1.7
RTF 1.0 – 1.9
テキスト文書
JIS (ISO-2022-JP) / EUC-JP / Shift_JIS / UTF-8 / UTF-16
マークアップ言語
HTML / XML / SGML
ODF (Writer / Calc / Impress)1.1 / 1.2
OpenOffice 3.0 / 3.1 / 3.2 / 3.3
LibreOffice 3.4

■デ変研 MFX ライブラリと連携した メールや圧縮ファイル展開

検索したいドキュメントが、メールに添付された文書ファイルであったり、圧縮ファイルに内包されたフォーマットファイルであった場合には、デ変研MFXライブラリと連携し、対象ファイルを解析して、さらに、添付ファイルや圧縮ファイルの展開を行って、フォーマットファイルを取出し、さらに、デ変研TFライブラリと連携することで、テキスト情報を取出し、そのテキスト情報をインデックスに登録することができ、高速検索処理を実現しています。
以下のメールや圧縮ファイルに対応しています。
※詳しくは「デ変研MFXライブラリ」をご参照ください。

■対応メール形式

1 つのメール展開 EML
（EML は、E-Mail 形式のことで、
RFC822 に準拠したものをいいます）

圧縮ファイル zip (winzip / pkzip : 圧縮形式 / 自己解凍形式)
lha (lh1 / lh5 / lh6 / lh7 : 圧縮形式 / 自己解凍形式)
tar＋gzip / tgz / gzip
rar (圧縮形式 / 自己解凍形式)
bzip2
7z (圧縮形式 / 自己解凍形式)
※それぞれの圧縮形式において、パスワード付きのものを除きます。

アーカイブ形式 tar / gnutar

■対応 OS

Red Hat Linux
AS3 / ES3 / WS3 /AS4 / ES4 / WS4 / EL5 / EL6 / EL7 / EL8

■構成

メモリ 1GB以上
HDD利用量 500MB以上（コマンド、環境ファイルの保存領域）
（検索インデックスは、検索ファイル数に応じて別途のディスク容量が必要です）

※Linuxは、32bit版と64bit版の両方を対応に含めます。
※他のOS・コンパイラ・開発環境下でご不明な点は、お問い合わせください。
※ハードウェアの搭載メモリは推奨2GB以上で、メモリ量が多い方が
大きな文書に対応できます。

Dehenken

株式会社 データ変換研究所 Dehenken Limited

本　　社 〒604-8155 京都市中京区錦小路通室町東入占出山町 308 ヤマチュウビル 1F
TEL 075-254-8780 FAX 075-254-8790
横浜営業所 〒231-0048 神奈川県横浜市中区蓬莱町 2-4-7 澤田聖徳ビル 204
URL : http://www.dehenken.co.jp/ E-Mail : info_ml@dehenken.co.jp

品質マネジメントシステム ISO 9001:2008 の認証取得

株式会社データ変換研究所は、2011年9月27日付けで全社統一の品質マネジメントシステムとしてDNV GL より ISO-9001:2008 の認証を取得しました。(現在は 2015 年版に移行)

認証の対象は「ソフトウェアプロダクトのデザイン・開発・製造」です。

Certificate No. : 02523-2011-AQ-KOB-RvA
Initial certification date : 27 September, 2011
Valid : 27 September, 2017 - 27 September, 2020

OEMビジネス向け全文検索エンジン

Cyclopeエンジン

CYCLOPE

Dehenken

Dehenken

●漏れのない検索エンジン（N-gram 方式を採用）

「Cyclopeエンジン」は、N-gram方式による漏れのない全文検索を行います。「Cyclopeエンジン」で採用している方式は、日本語は 2-gram（UTF-16 により他国の文字も可）、ASCIIは 4-gram という形式で行っています。例として『日本語 ABcd』は、「日本/本語/語AB/ABcd/Bcd/cd/d」のように区切り、これらをインデックスとします。

●すべての対象文書を検索します

「Cyclopeエンジン」は、ヒットした対象文書リストの全数を漏れなく列挙します。一般に検索エンジンは、高速化のためにいわゆる「足きり」をし、結果を速く出力します。本エンジンは、一致する全数を最後まで絞り出し、ヒットした対象文書を総て取り出します。また、検索時に限界個数（例えば1000件）を設定した場合には、指定された数までで打ち切ることもできます。

●管理できる対象文書は無制限

1つのインデックスで管理できる文書数は約10億です（30bitでの範囲。またはOSのシステムによる制限まで可）。1つのインデックスファイルの内部は、複数のデータブロックに分かれています。データブロックは、インデックス作成時に1000ファイル、あるいは、指定されたメモリ量のいずれかに達したときに書き込みが行われます。1つのデータブロックの最大は4GBで、複数の連結ができます。

●複数のインデックスを横断して検索できます

検索時には複数のインデックスファイルを横断して（串刺しして）行うことができます。システムのデザインとしては、日ごとのデータに対して1つのインデックスを作成するという日ごとインデックス型のものにすると、期間指定による横断検索が容易に実装できます。また、対象データのバックアップ時に、対象データとともに検索インデックスを含めることにより、リカバリ時に検索インデックスも回復させる仕組みにすることもできます。

組み込み利用の全文検索エンジン

Cyclopeエンジン

「Cyclopeエンジン」は、ソフトウェアにオリジナルの全文検索の機能を備えたいソフトウェア・メーカー様、クラウドサービスベンダ様のための、OEM 向けに組み込んで利用する全文検索エンジンです。全文検索の機能を搭載した強力なアプリケーションの実現をご支援します。

●ゆらぎによる漏れの吸収

アルファベットの大小文字の同一視や、全半角文字の同一視の設定が可能です。これにより、「ABC」という検索キーワードで、「ABC」「abc」のいずれでも漏らさず検索できます。ゆらぎの設定情報は、将来のマルチリンガル対応に向けて、UTF-8の文字コードにて定義できます（EUC-JPも可）。

●検索条件の指定と部分的条件付検索

検索時には and、or、not、( ) の条件を指定することができます。また、カスタマイズによりますが、部分的条件付検索をデザインすることができます。部分的条件付検索とは、一部の情報だけをインデックスに別フィールドとして登録し、その部分に特化した検索を実装することをいいます。部分的条件付検索は、日付、件数、場所（位置）指定です。メールの全文検索「MailCyclope」を例に挙げると、Subject / From / To等のヘッダの一部や、本文、添付情報、添付ファイル名等が、部分的条件付検索によって実現しています。

※部分的条件付検索の実装は、カスタマイズ案件となります。別途、お問い合わせください。

●ログによる処理の監視情報の提供

本検索エンジンの利用者が処理の状況を把握するために、ログ情報を出力させて、動作状況を追跡しやすくしています。インデックス生成時に、異常データによりプログラムが予期せぬ中断をするときや、大量データに遭遇して処理がタイムアウトするというような想定外の事態発生時の対応を考え、プログラムを監視型にしています。その監視プロセスのログ情報により、エラーの発生したデータを特定し、調査を迅速に行えるようにしています（以上の説明は「MailCyclope」を対象にしています）。

●クラスタリング検索コマンド

複数のインデックスを同時に検索できるクラスタリング検索コマンドを用意しています。クラスタリング機能により、大量の文書によるインデックス検索時間を複数のマシンに分散させて、検索が完了するまでの時間を大幅に短縮させることができます。また、クラスタリング検索コマンドには、監視機能も用意していますので、分散環境におけるプロセスの動作状況を把握できます。

●提供コマンド例

「MailCyclope」を例に、「Cyclopeエンジン」を用いたコマンド例についてご紹介します。多様なニーズやサポート側の立場に立ったコマンドをご提供しています。コマンド例とそのコマンドに含まれるライブラリを示します。

コマンド名	処理内容	TF含む	MFX含む
mc_index	インデックス作成コマンド	○	○
mc_search	検索コマンド	○	○
cmc_search	クラスタリング検索コマンド	○	–
cmc_daemon	cmc_search用の常駐デーモン	○	○
merge_index	インデックスのマージコマンド	–	–
dump_index	インデックスのダンプコマンド	–	–

●多くのソフトウェアベンダが採用

大手電機メーカ系のソフトウェア企業様や、新進気鋭のソフトウェアベンチャー企業様に採用されています。使用対象はアプリケーション組込型エンジン、電子メールフィルタリング、本文添付・ファイルの全文検索機能、通信パケット記録装置のフォレンジックシステム等です。

Cyclope

●MailCyclope

メールの全文検索のためのアプリケーションです。2005年個人情報保護法が運用されて以降、日々送受信されるメールを蓄積し、証拠保全するシステムが利用されています。MailCyclopeは、その大量蓄積メールの中から特定キーワードを含んだメールを、高速に検索するために使用されています。デ変研TFライブラリとデ変研MFXライブラリを組み合わせ、添付ファイルの中の圧縮ファイルの中のPDFファイルの中のキーワードまで検索対象にできます。

●HTTP Cyclope

HTTPプロトコルの蓄積されたストリームデータを、GETしたもの、PUTしたもの、POSTしたものに分類し、全文検索することができます。

●FTP Cyclope

FTPプロトコルの蓄積されたストリームデータを、全文検索することができます。

●Book Cyclope (XML Cyclope)

XML形式で記述されたテキスト情報を、全文検索することができます。書籍データを対象としているのでBook Cyclopeとしています。

●Log Cyclope

Logを全文検索するコマンド群です。Logデータは、文書データと比較するとインデックスが拡大しやすい傾向のデータです。そこでLogデータを複数行にわたってチャンキング（Chunking：処理をまとめる）してインデックスデータ量を圧縮し、検索コマンドではさらにChunked Scopeの実データをパターン一致確認（UNIXのGREPコマンドの動作イメージ）して、検索ノイズをなくしています。

●OfficeCyclope

社内文書をサーバに格納して全文検索するシステムです。GUIはPHPを使用して実現しています。オフィス利用を前提としてDehenkenより販売しています。

●Brain Cyclope

クラウド環境での社内文書検索の頭脳ソフトウェアの実現を目的として商品開発をしています。