

Softwarehaus Graf & Partner

TopicSource

Ver. 1.8

SOFTWAREHAUS

Graf & Partner

Säulengasse 17, A-1090 Wien
Tel. und Fax: +43 - 1 / 310 24 39
Email: office@grafsoft.co.at
Homepage: www.grafsoft.at

Contents:

1	Überblick	3
2	Details.....	4
2.1	Volltextdatenbank.....	4
2.2	Textquellen.....	5
2.3	Internet Crawler.....	5
2.4	Kategorisierung.....	6
2.5	Frequenzanalyse.....	7
2.6	Begriffswolke.....	10
3	Technische Angaben.....	11

1 Überblick

TopicSource ist ein Werkzeug zum Wissensmanagement auf der Basis einer Volltextdatenbank, hauptsächlich für Internet- und Intranet-Lösungen. Man könnte aber auch eigene Programme damit erstellen, oder Erweiterungen für MS Office.

Mögliche Anwendungen und einsetzbare Features

- Internet-Suchmaschinen, ein Beispiel ist <http://www.newswatch.at>
- Verwaltung eigener Textdokumente und Tabellen
- Verwaltung von und Textanalyse in Emails
- Auto-Beschlagwortung von Texten auf der Basis von vordefinierten Suchbegriffen
- Auto-Beschlagwortung von Texten auf der Basis von vordefinierten Hierarchien
- Definition von topic-Maps
- Textanalyse mit Wortfrequenzen und Ko-Frequenzen – man kann Schwerpunkte in Texten erkennen, ohne die Texte selbst lesen zu müssen
- Textanalyse mit Begriffswolken, die entweder aus der Gesamtheit der Texte oder den bei einer Suche gefundenen Texte gewonnen werden

2 Details

2.1 Volltextdatenbank

- Konkurrenzlos schnell, vergessen Sie die Indizierung von Office oder ähnliche Hilfsmittel
- Liest auch PDF, HTML und Office-Dokumente
- Kennt auch die Positionen der gefundenen Begriffe innerhalb des Textes, zum Hervorheben während der Anzeige
- Unterstützt Umlaute und verschiedene Zeichensätze
- Kennt Stopwörter
- Vielfältige Einstellmöglichkeiten
- Einfache Suche, aber auch sehr komplexe Sucheingaben möglich

Die nachfolgenden Programmierbeispiele sind eher für Techniker gedacht. Sie sollen die Einfachheit des Systems zeigen. Zum Feintuning gibt es noch hunderte andere Prozeduren und Methoden, die im Handbuch nachgelesen werden können.

Programmbeispiel (VB):

Herstellen einer Datenbank

```
` Datenbankobjekt herstellen
Set obj = CreateObject("Topicdb.ttopicdb")
` Am Server anmelden
obj.openserver ("localhost","root","MyPassword")
Textdatenbasename="MyText"
Topicdatenbasename = "Mytopic"
` Datenbanken herstellen wenn sie nicht schon da sind
if not obj.databaseexists (textdatenbasename) then obj.createtextdatabase
(textdatenbasename)
if not obj.databaseexists (topicdatenbasename) then obj.createtopicdatabase
(topicdatenbasename)
obj.textdatabase = textdatenbasename
obj.topicdatabase = topicdatenbasename
obj.createtexttables
obj.createtopictables
` Felder definieren
obj.createfield ("Filename")
obj.createfield ("Filecontent")
```

Text einfügen und Index automatisch nachziehen

```
` Die folgenden Zeilen brauchen wir immer, in Zukunft werden sie nicht
angeführt
Set obj = CreateObject("Topicdb.ttopicdb")
obj.openserver ("localhost","root","MyPassword")
obj.textdatabase = "MyText"
obj.topicdatabase = "Mytopic"

` Einen Text einfügen
Obj.createdocument(false)
Obj.Addtext "filecontent","This is my text",true,true
```

Suchen und Ergebnisse ansehen

```
Hitcount = obj.search („Water“)
somestring = obj.hitsasstring
```

```

array_var = split (somesstring, ",")
` Zum ersten Treffer gehen
Obj.gotodocument (array_var(0))
` Text abholen
t = obj.fieldtext ("Filecontent")

```

2.2 Textquellen

Fast alle Daten auf der Festplatte oder aus dem Internet können verarbeitet werden.

Das System kann auch automatisch Daten aus dem Intranet spiegeln, und die gespiegelten Inhalte suchbar machen.

Alle Office-Formate können gelesen werden, ebenso PDF und HTML. Wir haben auch die Technologien weitere Formate einzulesen.

Programmbeispiel (VB):

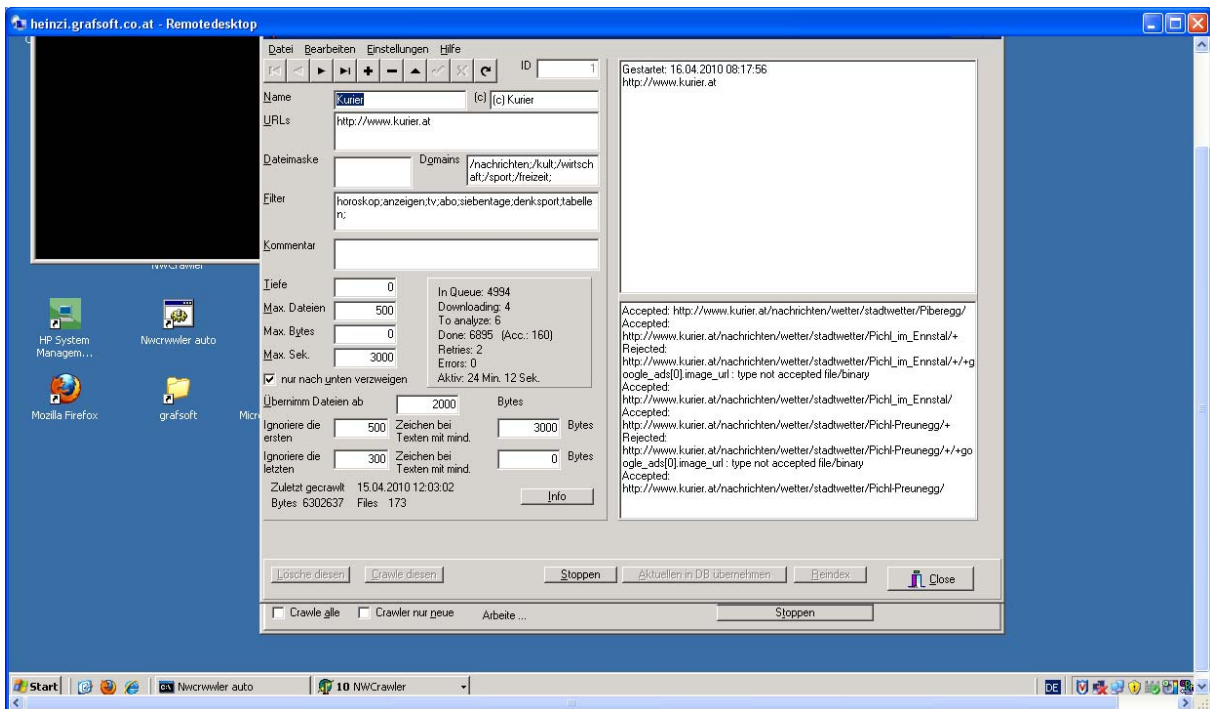
```

` Wir sagen dem System welche Dateien es lesen soll
obj.setinfo "Watchfiles", "Directories", "f:\texte" & chr(13) & chr(10) &
"\server\d\Office"
obj.setinfo "Watchfiles", "Textfiledir", "d:\dbtext"
obj.readfiles (true)
` das war's: gelöschte Dokumente werden vergessen, neue und geänderte neu
indiziert.

```

2.3 Internet Crawler

Wir verfügen über Programme zum gezielten Einlesen von Daten aus dem Internet- So können bestimmte Seiten überwacht oder tägliche Nachrichten eingelesen werden. Das Aktualisieren der Datenbank erfolgt automatisch.



Es gibt folgende Einstellungsmöglichkeiten:

- Zu akzeptierende Dateitypen
- Zu ignorierende Dateitypen
- Nur nach unten verzweigen J/N
- URLs mit bestimmten Substrings einschließen
- URLs mit bestimmten Substrings ausschließen
- Nur Dateien akzeptieren, die ein bestimmtes Datum enthalten (z.B., von heute oder gestern)
- Maximale Tiefe
- Maximale Anzahl an Dateien
- Maximale Anzahl an Bytes
- Maximale Crawlzeit
- Maximale Crawlzeit pro Seite
- An jedem Wochentag oder nur an bestimmten crawlen?
- Mindeste Dateigröße
- Die ersten x Zeichen ignorieren
- Die letzten x Zeichen ignorieren
- Zahl der Seiten, die synchron gecrawlt werden
- Zahl der Threads pro Crawler

2.4 Kategorisierung

Wenn wir Obst suchen, meinen wir Äpfel, Birnen, Weintrauben ...

Städte sind Berlin, Wien, London ...

In England liegen London, York, Chester ...

Sie können beliebig viele Kategorien verwenden und miteinander vernetzen:

Europa -> Österreich -> Tirol -> Innsbruck -> ...

Die Art der Kategorisierung kann auch mehrdimensional sein. Stellen Sie sich Zeitwörter vor, z.B. „Berlin liegt in Deutschland“ oder „Ein Auto hat Räder“ oder „Lachse gehören zu den Fischen“.

Auch Synonyme können definiert werden: Ferment = Enzym, Aubergine = Melanzani.

Dazu kommt noch die automatische Suche nach Schlagwörtern. Sie definieren: Wenn in einem Dokument (*Brand und Löschen*) oder *Feuerlöscher* vorkommt, handelt es sich um einen Artikel über Sicherheit. Oder:

Damit ist sogenanntes automatisiertes Tagging möglich.

The screenshot shows a Mozilla Firefox browser window with the address bar containing the URL: `http://www.grafsoft.at/rezepte/showresults.asp?Page=6&MaxDocs=10&Searchstring=&NumberOfHits=90`. The page content is a search results page for 'Geflügel' and 'Huhn'. It shows two search results:

Keywords	Geflügel, Huhn
Treffer Nummer:	52 Ändern Löschen
Nr / Datum	347 / 12.01.2010 18:48:49
Quelle	Wienerin Dez. 1988
Titel	Perlhuhnbrust auf Lauchrahm mit Gemüsebandnudeln
Abstract	Perlhuhn oder Huhn, Lauch, Karotten, Sellerie, Zwiebel, Zucchini, Weißwein, Obers, Champignons, Butter, Nudeln. Brüste und Keulen abtrennen. Aus dem Rest mit dem Grünen von Lauch, 2 Karotten, Sellerie und 1/2 Zwiebel Geflügelfond herstellen. Lauchrahm: Geflügelfond mit 1/4 Wein, Zwiebel und 2 blättrig geschnittenen Champignons 110 min reduzieren lassen. Obers dazu, zur Chreme einkochen, abseihen, mit 5 dkg Butter mixen und den in Streifen geschnittenen blanchierten Lauch dazu. Bandnudeln in Salzwasser kurz kochen, kalt abspülen. Karotte und ungeschälte Zucchini in Streifen schneiden. Karotten nebeneinander ohne Wenden in Wasser mit Salz und Butter knackig garen, Zucchini erwärmen. Nudeln dazu. Brüste in Butter braten, etwas ruhen lassen und auf dem Lauchrahm anrichten.
Keywords	Geflügel, Gemüse, Huhn, Karotten
Treffer Nummer:	53 Ändern Löschen
Nr / Datum	346 / 12.01.2010 18:48:49

At the bottom of the browser window, a status bar shows: 60 / 100 Flow: 100% Links: 0/37 Visits: n/a Fertig

Das obige Beispiel zeigt auch wie man es halb-richtig machen kann. Hier wurde nach „*huhn* or *huehn*“ gesucht, um Hühner zu finden und sie auch gleich dem Geflügel zuzuweisen. Dass damit auch Perlhühner gefunden werden, war ein Versehen – oder nicht? Das müsste ein Zoologe entscheiden.

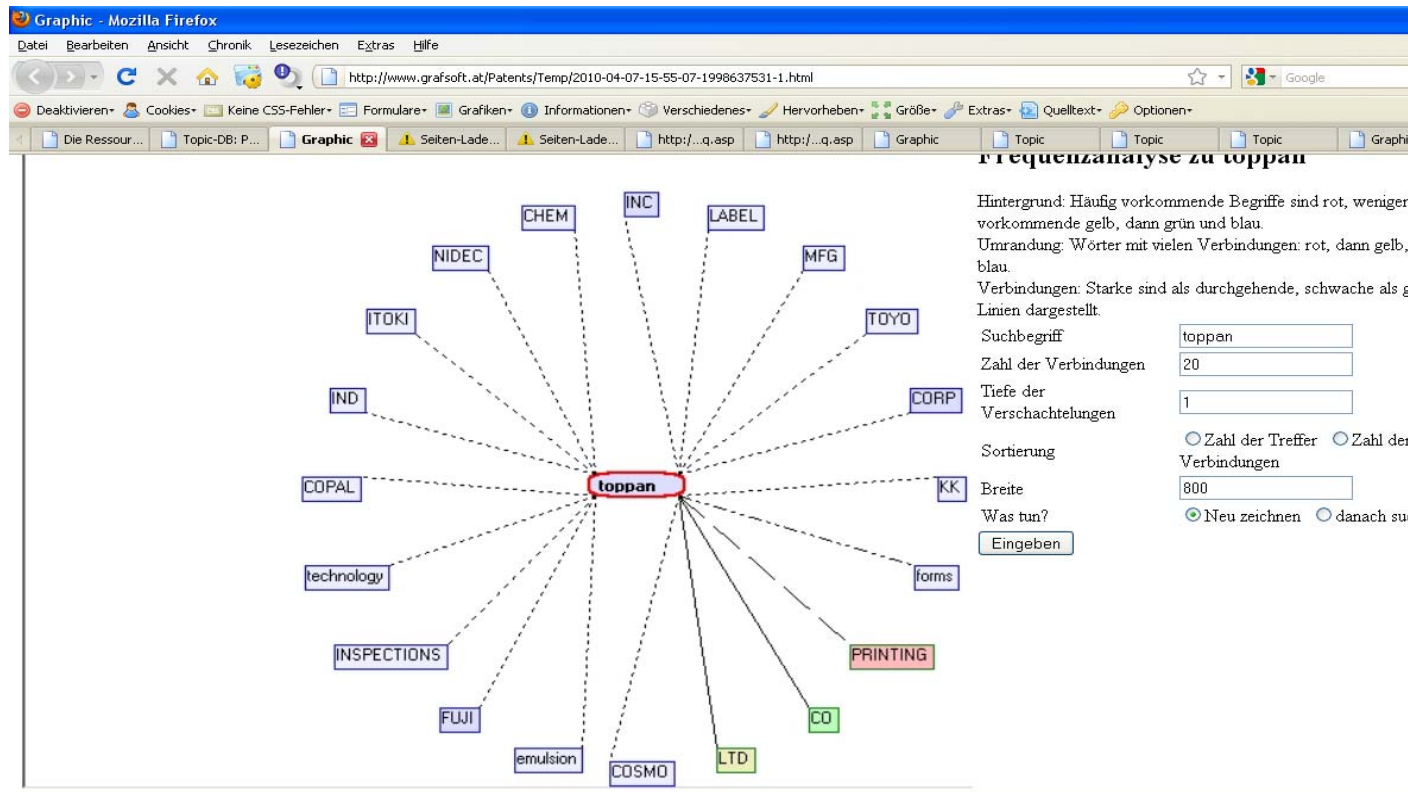
Programmbeispiel (VB):

```
Obj.Addreplacetopic (Afrika, Ägypten, Is_Upper, Is_lower, false)
obj.autokwsearch ("Autokeywords")
obj.autokeywords (1)
```

2.5 Frequenzanalyse

Manchmal ist es interessant zu wissen, welche Wörter in einer Sammlung von Texten häufig vorkommen. Noch interessanter kann es sein, zu sehen, welche Wörter in der Nachbarschaft welcher Wörter vorkommen.

Das folgende Beispiel zeigt eine Analyse aus einer Patentdatenbank. In der Mitte sieht „Toppan“ für die Patentanmelder Toppan Printing Ltd und Toppan Ink Mfg. Rundherum sieht man Firmennamen wie Copal, Fuji, Cosmo und Toyo.



Warum? Weil Toppan offensichtlich mit diesen Firmen gemeinsam Patente angemeldet hat.

Nr / Datum	7951 / 07.02.2010 13:24:42
Number	2004-050416 JAPIO
Title	DRYER
Inventor	FUJITA YUICHI; MATSUNAGA KAZUO; HOTTA KOICHI; KAWASHIMA HIROYUKI; YASUDA HIDEKI; TOMINAGA YASUMASA; MIZUNO TAKAHIRO; WAKAMATSU TOSHIO; NAGUMO MIKIO
Applicant	TOPPAN PRINTING CO LTD TOYO INK MFG CO LTD FUJI KIKAI KOGYO KK TERUMO KOGYO:KK AJINOMOTO ENGINEERING CORP
Patno	JP 2004050416 A 20040219 Heisei
Application	JP 2002-206763 (JP2002206763 Heisei) 20020716
Priority	JP 2002-206763 20020716
Source	PATENT ABSTRACTS OF JAPAN (CD-ROM), Unexamined Applications, Vol. 2004
Class	ICM B41F023-04 ICS F26B013-10
Abstract	PROBLEM TO BE SOLVED: To provide a dryer, with which an aqueous coating liquid is efficiently dried without developing a bad effect such as the swelling of a sheet due to too much heating and, at the same time, which is used both for drying an organic coating liquid and for drying the aqueous coating liquid in combination. SOLUTION: In the dryer having a boxlike casing 4 with the sending-in slit 2 and the taking-out slit of a sheet 1 coated with the coating liquid, a plurality of guide roller 5 arranged in the


```
"showgraph.asp?thestring=", "&count=", "&depth=", "&order=" ) )
```

2.6 Begriffswolke

Begriffswolken sind häufig im Internet zu finden. Mit Topicsource sind sie auch leicht herzustellen.

The screenshot shows the NewsWatch website in a Mozilla Firefox browser. The page title is "NewsWatch-Startseite - Mozilla Firefox". The address bar shows "http://www.newswatch.at/default.aspx". The page content includes the NewsWatch logo, a search bar, and a word cloud titled "Begriffswolke". The word cloud contains various terms such as "Karriere", "Suche", "Sudoku", "Themen", "Newsletter", "Services", "Inhalt", "Vergessen", "Events", "heute", "Außenpolitik", "margin", "Innenpolitik", "Kanäle", "hier", "Wintersport", "Wien", "Europa", "display", "Dienst", "rsaaqu", "Kino", "Reise", "aktuellen", "kleine", "Mobil", "none", "Profil", "Medien", "Family", "Angemeldet", "Familie", "Redaktion", "com", "Anzeigen", "Politik", "Trinken", "Hilfe", "Film", "Essen", "Shop", "Tabellen", "Wirtschaftsblatt", "Login", "Kunst", "Motorsport", "Thema", "Presse", "Info", "Mittwoch", "Gewinnspiele", "Volkszeitung", "Spiele", "Startseite", "Immobilien", "freizeit", "Menschen", "Preisvergleich", "Börsenkurse", "Sport".

Sie haben unter anderen folgende Möglichkeiten:

- Begriffswolken aus dem gesamten Text oder den Ergebnissen einer Suche herstellen
- Begriffe aus Wolken ausschließen
- Schriftgrößen einstellen
- Einen Link zu jedem Wort definieren
- Zufallsverteilt umsortieren

Programmbeispiel (VB):

```
obj.cloud_setup 1000,7,24,0,0,from,4,0
for i=lbound(hitarr) to ubound(hitarr)
    obj.gotodocument hitarr(i)
    t=obj.fieldtext ("Text")
    obj.cloud_adddtext t
next
obj.cloud_keepfirst 5000,true
obj.cloud_sortrandom
response.write (obj.cloud_create (80))
```

3 Technische Angaben

Technisch gesehen handelt es sich um eine DLL, die mit allen Internet-Programmiersprachen zusammenarbeiten kann, aber auch mit Sprachen wie .C, Visual Basic (auch Microsoft Office) usw. arbeitet.

Voraussetzungen:

Windows NT bis 7 oder Windows Server
MySQL oder MS-SQL, andere Datenbanksysteme auf Anfrage.

Installation:

Sehr einfach. Kopieren und registrieren (mit regsvr32)

Für das Internet-Crawling und die Zusammenführung von Texten gibt es EXE-Dateien.

Wir können auch in kurzer Zeit kundenspezifische Anwendungen erstellen.