



T O V E K

---

# Tovek Tools: Harvester

---

version 4.26 and above

---

## User Guide

---







# Copyright

---

Information contained in this document may change without notice. The software described herein is subject to and protected by copyright law and international agreements related to intellectual property. It can be used solely in strict accordance with the License Agreement. No part of this publication can be reproduced, transmitted, transcribed by any electronic or mechanical means including photocopying and storing in retrieval systems, for any purpose, without the prior explicit permission of Tovek.

© Copyright Tovek. All rights reserved.  
<http://www.tovek.com>

## Registered trademarks

TOVEK and InfoRating are registered trademarks of Tovek.

Analyst's Notebook is a registered trademark of i2 Limited.

## Other trademarks recognition

Microsoft, Excel, Windows and Windows NT are registered trademarks of Microsoft Corporation.

All other companies and product names are trademarks or registered trademarks of their respective owners.

**TOVEK**

<http://www.tovek.com>

# Contents

---

<b>Introduction .....</b>	<b>7</b>
Manual Version .....	7
Publications Related to Tovek Tools .....	7
Help .....	7
Start Working with Harvester .....	8
<b>Data in Harvester.....</b>	<b>11</b>
<b>Harvester Window.....</b>	<b>12</b>
<b>Main Pane .....</b>	<b>14</b>
Connection Chart .....	14
Displaying Data in Chart .....	15
Working with Chart .....	16
Groups of Words.....	19
Retrieving Documents .....	21
Copying from the Chart.....	22
Chart Settings.....	23
Document Text .....	26
Working with Text View .....	27
<b>Lists.....</b>	<b>28</b>
Word List .....	28
Related Words.....	29
Word History.....	31
Descriptors .....	31
Word Neighbourhood .....	32
Result List.....	34
Document Fields.....	36
<b>Searching for Documents using Tovek Agent.....</b>	<b>38</b>
Customizing TT Query.....	39
<b>InfoRating Context Queries.....</b>	<b>41</b>
Word List and Related Words List .....	41
Descriptors Panel .....	41
Word Neighbourhood .....	41
Chart View.....	42
<b>Exports.....</b>	<b>43</b>
Html/XML Export .....	43
Export of Documents to Tovek InfoRating.....	43
<b>Settings.....</b>	<b>44</b>
Default Settings .....	44
Settings of Current Repository .....	46
<b>Printing from Harvester .....</b>	<b>48</b>
Print .....	48
Print Preview .....	48
Page Breaks Preview .....	48

**Index ..... 49**

## Introduction

---

Harvester analyses the content of input documents and extracts relevant terms and their relations.

## Manual Version

This manual corresponds to the Tovek Tools version stated on the first side of the manual. For further program versions Release Notes will be published, which will be available on [www.tovek.com](http://www.tovek.com).

## Publications Related to Tovek Tools

Tovek Tools Search Pack

Tovek Tools Analyst Pack

Tovek Tools: Query Editor

Tovek Tools: InfoRating

Tovek Tools: Harvester

Tovek Tools: Fulltext Plug-in for Analyst's Notebook

Verity Query Language

## Help

Tovek Tools contains help accessible from **Help** menu. After selecting the **Help Content F1** command you can see the overview help for all Tovek Tools programs and after selecting **About** command you will see information about the licence owner, expiration and Tovek Tools version. The version number is required for a support demand.

## Start Working with Harvester

Harvester uses a set of statistical measures to determine if a term or relation between two terms is relevant or not. The relevancy depends on the distribution of the occurrence of the term over all analyzed documents, so the result also depends on the count of input documents.

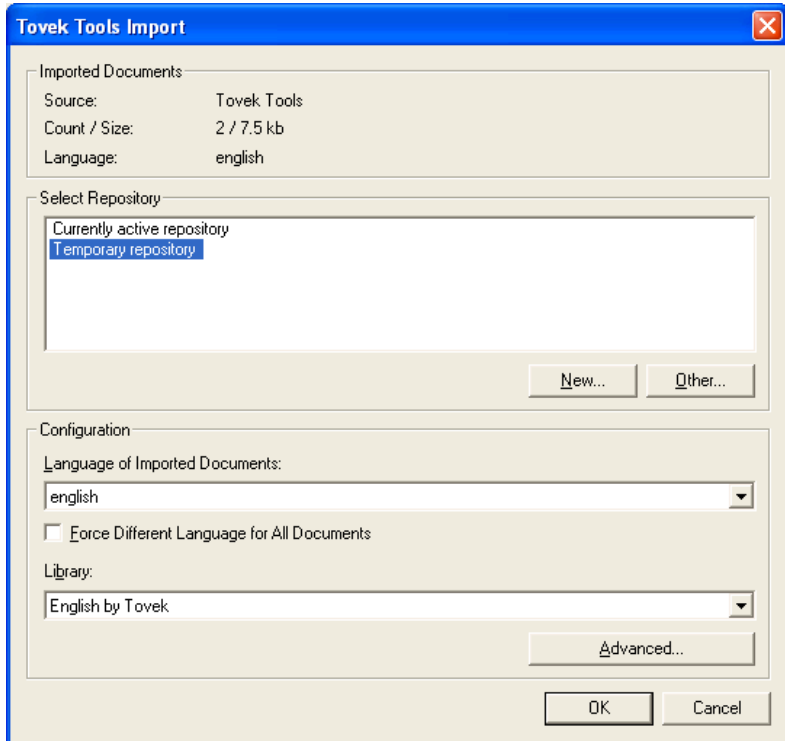
Harvester allows you to examine the result of the analysis in several ways:

- list relevant words and their relations,
- browse graphically through the relations between relevant words,
- view clusters of relevant words,
- read documents highlighted by selected relevant words,
- examine the neighbourhood of a relevant word,
- detect trends.

Input documents can be transferred from Tovek Agent to Harvester in the two following steps:

1. selecting documents from result list of Tovek Agent query - by standard Windows methods or by marking documents for export (see Tovek Tools Search Pack Guide, Marking documents for export),
2. transferring them into Harvester - selected documents can be exported by **Harvester Export** command from the **Tools** menu or from the context menu of selected documents.

After successful export, the Harvester window opens together with the following dialogue window:



In the **Imported Documents** part of the dialogue there is information about the document source (usually Tovek Tools), number of transferred documents, their complete size and their languages.

In the **Select Repository** part you can select whether the documents should be imported to:

- new temporary repository (area for analysis) – **Temporary repository** option. The temporary repository is not stored on the disk until you store it by **Save** or **Save As** command.
- current repository – **Currently active repository** option. This option is available only if you perform a repeated import to an already existing repository. Newly imported documents will be added to the repository and then analyzed.

- new repository – the **New** button, it allows you to create and name a new analysis where the documents will be imported.
- already saved repository – the **Other** button, it allows you to select and open an existing repository and extend it with the imported documents.

The trend detection feature of Harvester works with time information attached to the documents. This information can be supplied in the form of a date field that is imported together with the documents. If there isn't any date field, then today's date is used as the date of the document's origin (for more details on how to export documents together with a date field, please, refer to the Tovek Tools Search Pack Guide).

All documents in one repository must be in the same language. The last part of the dialogue called **Configuration** allows you to specify documents of which language should be analyzed if the import contains documents of different languages. You are also able to enforce a language setting for all imported documents by selecting the checkbox **Force Different Language for All Documents**. In this case the language combo box contains all languages supported by Harvester.

In the **Library** combo box you can select a language component that should be used while computing document abstracts. These libraries are responsible for determining the type of each word in the analyzed text and building its base form. Only nouns and unknown words are involved in further steps of content analysis.

The **Advanced** button lets you change the default settings for new repositories (see Default Settings Chapter).

After pressing the **OK** button, all documents with the selected language will be added to the selected repository and analyzed.

All examples and pictures that can be found in this documentation have been prepared based on documents from the Reuters demo fulltext database<sup>1</sup>. These documents have been analyzed using default settings. The date field "Date" has been used as the basis for trend detection.

---

<sup>1</sup> Available for download at <http://www.tovek.com>

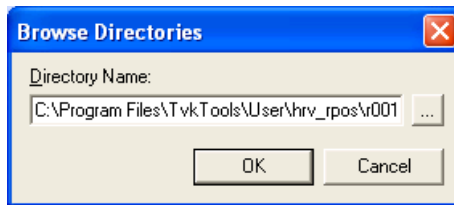
---

## Data in Harvester

---

Analysis of input documents is always performed within one *repository*. This repository holds all statistical values, document indexes, document abstracts and configuration of the concrete analysis. This allows incremental updating of the analysis with new documents or removing old ones.

Repositories can be stored in file system as a set of files organized in a directory structure. To store an active repository, run the **File/Save Repository** command (Ctrl+S) or the **File/Save Repository As** command. If you save the repository for the first time, or if you use the second command, the following dialogue appears where you can select the destination directory:



To open a saved repository, use the **File/Open Repository** command (Ctrl+O) and select the directory that contains repository files. Only if you select a proper directory will the **Ok** button be enabled.

The **File/Close Repository** command closes the currently opened repository. The **File/New Repository** command (Ctrl+N) lets you save the currently opened repository (if there is any) and creates a new empty repository.



Charts created based on an analysis are handled independently of the repositories. You can store the current chart in a file using the **File/Save Chart** or **File/Save Chart As** commands. Charts are stored in files with *.ths* extension.



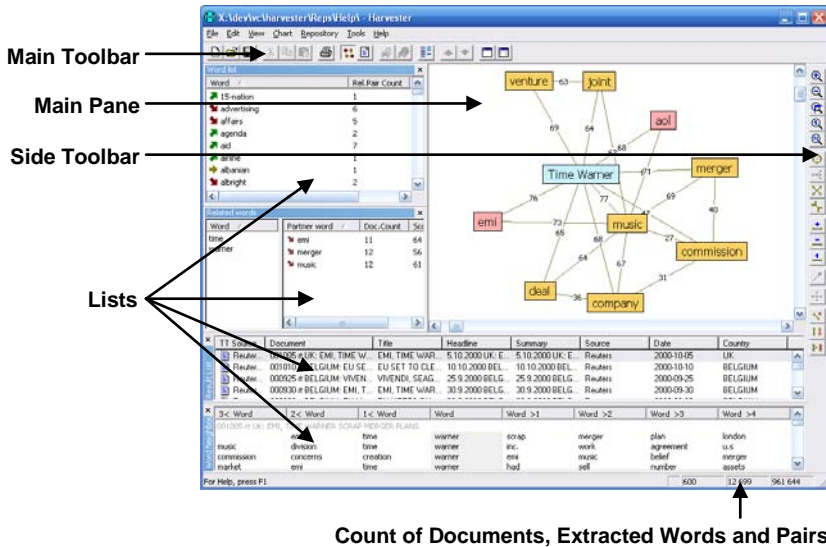
Saved charts can be loaded into Harvester using the **File/Open Chart** command. When you load a chart, it will be restored in the chart view. Words of the chart that are not relevant in the current repository will be greyed and not active. All other words can be used to extend the chart as described in further sections.



The **File/New Chart** command lets you save the current chart and clears the main pane.

# Harvester Window

The default view of the Harvester application window is shown in the following picture:



It consists of several components. Results of an analysis are accessible through the main pane of the application and through a set of lists.

The main pane can display either extracted relevant words and their relations in graphical form or highlighted text of a selected document. Each of these views has an associated side toolbar which brings the most important commands for working with the view. Side toolbars switch together with their views.



Many aspects of the result of the analysis are accessible through six lists:

- **Word List** displays list of all extracted relevant words.
- **Related Words** list shows words that are related to all words specified in the left part of the list.
- **Word History** shows a graph of the occurrence of selected words in time.

- **Descriptors** pane lists either all extracted descriptors or descriptors for specified words only.
- **Word Neighbourhood** shows neighbouring words of a given word.
- **Result List** displays information about a document matching specified words.

## Main Pane

The main pane of Harvester is used to display connection charts or document texts. To switch between the two views, you can either use buttons on the main toolbar or commands in the **View** menu.

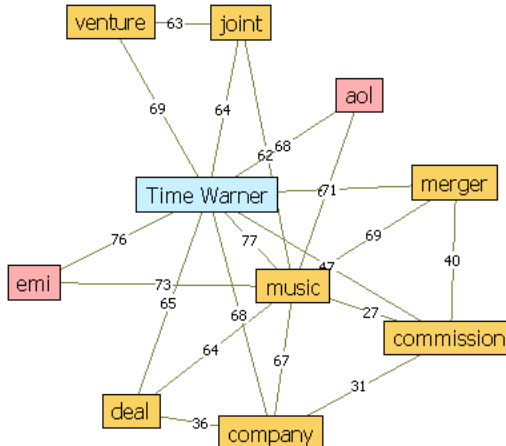
Toolbar Button	Command name	Keyboard shortcut	Function
	<b>Connection Chart</b>	<b>Alt+1</b>	Displays the connection chart of words
	<b>Document Text</b>	<b>Alt+2</b>	Displays the text of the last opened document.

## Connection Chart



The chart view allows you to inspect relations between extracted relevant words. To switch to this view, use the **View/Connection Chart** command or press **Alt+1** or use the **Connection Chart** toolbar button.

The following picture shows a sample chart with word clusters that have been extracted during the analysis.



The chart contains three types of nodes and links between them:

music

Known relevant words – nodes coloured yellow by default, represent all words that have been recognized by the language component.

aol

Unknown relevant words – nodes coloured pink by default, represent words unknown for the language component. These words are often new names of people, companies or products. In the above picture the company names “eml” and “aol”, for example.

Time Warner

Groups of words – nodes coloured light blue by default, represent user defined groups of words. The above picture shows one group node called “Time Warner”, which has been created as a connection of two words “time” and “warner”.

–63–

Connections between nodes are labelled with a value between 0 and 100 which describes how related the connected words are (0 – not related, 100 – both mean the same). There are two types of connections – a *descriptor* connection and a *simple* connection. A descriptor connection links two words that are part of a descriptor.

## Displaying Data in Chart

There are several ways to add data to the chart:

1. Dragging words from lists and dropping them onto the chart allows you to add single words. If you drag and drop a word that is contained in a word group, then the appropriate group will be added to the chart.

2. To add selected words to the chart, you can also use the **Add to Chart** command from the context menu of a word in a list.
3. Using the **Chart/Insert Related** command, you can add all related words of all selected words to the chart.
4. Double-click a selected word in the chart to add all related words.
5. To get an overview of all words and their relations, you can use the **Chart/Show Clusters** command. In this way all relevant words compounding descriptors will be added to the chart and reorganized so that you can easily recognize the clusters that they build. With **Ctrl** this command will add all words compounding pairs.

## Working with Chart

The most common commands for working with a chart are accessible through the chart's side toolbar and are described below.

### *Changing the Chart Size*



**Zoom In** button or **Chart/Zoom/Zoom In** command or **PgUp** key enlarges an object's representation in the chart.



**Zoom Out** button or **Chart/Zoom/Zoom Out** command or **PgDn** key lessens an object's representation in the chart.



After you press **Zoom to Area** button or start **Chart/Zoom/Zoom to Area** command and mark by mouse an area, all the items in it will change their size to fit the diagram pane.



**Actual Size** button or **Chart/Zoom/Actual Size** command or **Home** key changes the display scale to original icon size.



**Fit in Window** button or **Chart/Zoom/Fit in Window** command or **End** key changes the size of items to fit them all into the diagram pane.

### *Modifying Layout of the Chart*



After you press **Circular Layout** button or run **Chart/ Layout/Circular Layout** command, the objects will rearrange into a circle.



If you select one object and press **Tree Layout** button or run **Chart/Layout/Tree Layout** command, the objects will rearrange into hierarchical structure under the selected object.



**Group Layout** button or **Chart/Layout/Group Layout** command rearranges the most tied objects into a circle and places the rest of the objects around them.



**Connections Layout** button or **Chart/Layout/Connections Layout** command rearranges objects with the least crossings of their links.



**Increase Edge Length** button or **Shift+PgUp** shortcut enlarges the distance between objects (increases the length of links between them).



**Decrease Edge Length** button or **Shift+PgDn** shortcut lessens the distance between objects (decreases the length of links between them).



**Original Edge Length** button or **Shift+Home** shortcut changes the distance between objects to the default one.



**Fixed Position** button fixes selected objects in a diagram. These objects will have a different icon (red pin added) and will stay in the same position even after the layout change. If you do fixing for several selected objects, some of them fixed and some not, all of them will be released (pin removed).

### *Modifying the Chart*



**Insert Related** button, **Chart/Insert Related** command, double-click on object or **Ctrl+Enter** shortcut adds related words to the chart.



**Hide Not Connected** button or **Chart/Hide/Hide Not Connected** command removes all nodes having no connection from the chart.



**Hide Simple Edges** button or **Chart/Hide/Hide Simple Edges** command removes all objects having only one connection from the chart.



**Leave Only Descriptors** button or **Chart/Hide/Leave Only Descriptors** command removes all objects that are not part of any descriptors from the chart.

To hide selected nodes of the chart, use the **Chart/Hide/Hide Selection** command or press **Delete** key.

### *Selecting Chart Nodes*

Selecting nodes of a chart can be done by mouse or using the commands described below. To select a single node, just click it with the right mouse button. To add a node to the selection, press the **Ctrl**

key and click it. Neighbouring nodes of a node can be selected by pressing the **Alt** key and clicking the node.

To select all displayed nodes run the **Edit/Select All (Ctrl+A)**.

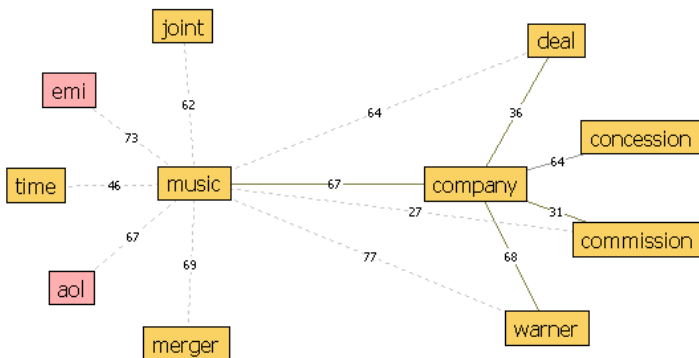
Using the **Edit/Select Single Nodes** command you are able to select all not connected nodes of the chart.

To mark nodes that are related to all of the currently selected nodes, use the **Edit/Select Shared Nodes** command.

### **Exploration Mode**

The chart can be viewed in two different modes.

- The default mode can be selected by pressing **Alt+8** or by running the **Chart/Graph Mode** command. In this mode you can work with the chart as described in the above section.
- The alternative mode is called Exploration mode and can be turned on by pressing **Alt+9** shortcut or by running **Chart/Exploration Mode** command. This chart mode is intended for exploring the neighbourhood of words in the chart and for browsing the relations between words.



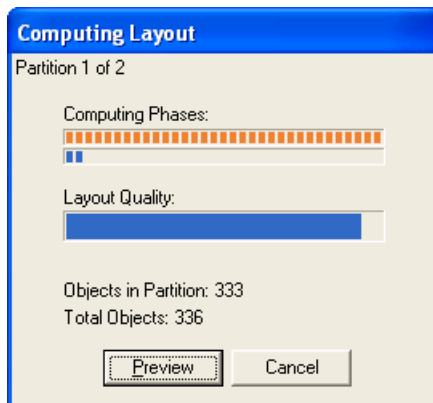
In the Exploration mode the selected word is placed in the centre of the chart and words that are directly related to the selected one are placed around it. Relations between these words and the selected one are drawn as solid lines. In the example above the selected word is

"*company*" and the directly related words are "*deal*", "*concession*", "*commission*", "*warner*" and "*music*".

While browsing relations between words, the chart also displays the previously selected word (if it is related to the currently selected one) and all words related to it. These words are then connected with dotted lines. In the example above the previously selected word is the word "*music*" and words related to it are "*aol*", "*commission*", "*deal*", "*eml*", "*joint*", "*merger*", "*time*" and "*warner*".

### Updating the Chart

A progress dialogue appears if the updating of the chart lasts for a long time. It allows you to view temporary snapshots of the chart before the computation has been finished (**Preview** button) or to interrupt the computation process (**Cancel** button).



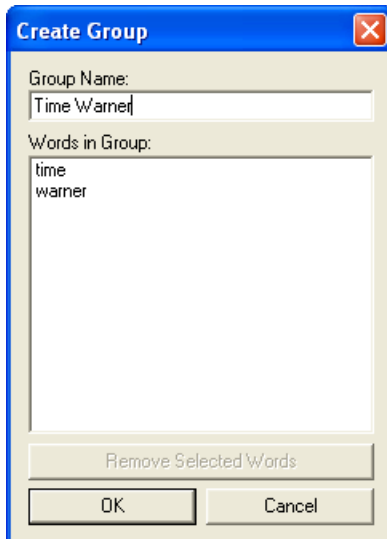
## Groups of Words

Harvester algorithm works on the basis of words extracted from analyzed documents. These are the words that you can browse in the chart view or use in the lists. But sometimes the word level is too fine grained with respect to content presentation. And for such cases Harvester allows you to define named groups of words. These groups are then displayed in the chart in place of the words. The definition of such groups can be exported in a file and reused within any following analysis.

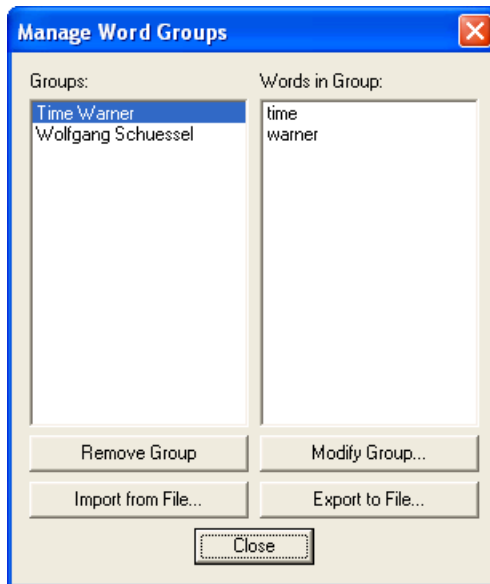
To define a new group of words, select all the words you want to include within the chart first. To create the group, run the **Tools/Create Word Group** command (the same command can also be found in the context

menu of selected word in chart). This command opens a dialogue where you can enter a name of the group. Press the **Ok** button to complete the group definition.

Note: A word can be included in one group only.



Word groups that have been defined can be further managed using the **Tools/Manage Word Groups** command, which opens the following dialogue:



A list of all defined word groups is shown on the left side of the dialogue. On the right side there is a list of words contained in the currently selected group. This dialogue allows you to remove the group definition by pressing the **Remove Group** button or to modify them using the button **Modify Group**. To store all known group definitions in a file, use the button **Export to File**, select the destination file and press **Ok** button. Using the button **Import from File** you can load previously stored group definitions. After selecting a file, all group definitions will be added to the already existing ones except for those that contain a word already assigned to a group. If there is such a group definition in the file, then a message box will pop up saying that not all group definitions could get imported.

To finish managing groups, close the dialogue by pressing **Close** button.

## Retrieving Documents

When you identify interesting words or word relations in the chart, you can either retrieve documents that contain the words directly from the repository or you can use these words for refining search in Tovek

Agent (see Searching for Documents using Tovek Agent) or as context queries in InfoRating analysis (see InfoRating Context Queries).

To search for documents containing selected words directly in Harvester press the **F5** key or use the **Tools/Fill Lists** command (this command is also part of the chart context menu). The result of the search will be displayed within all involved lists.

It is also possible to append the selected words to the last query and so to refine the previous search. This can be done by running the **Add to Query** word context menu command.

## Copying from the Chart

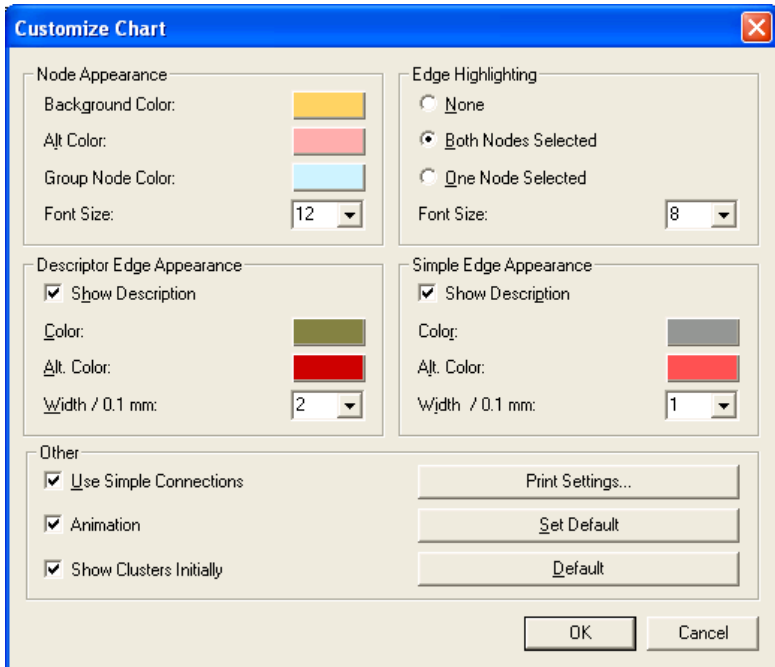


Objects selected in the chart view can be copied onto the clipboard using the **Edit/Copy** command, by pressing **Ctrl+C** or the **Copy** toolbar button. All selected words will be copied as strings each on a separate line; groups will be represented by their names followed by all contained words.

If the **Shift** key is held down while copying the chart, a bitmap representation of its selected part will be placed on the clipboard.

## Chart Settings

The **Chart/Customize Chart** command opens a dialogue for customizing the appearance of the chart.



The **Node appearance** section of the dialogue allows you to change the look of chart nodes. You can change the background color depending on the node type:

- **Background color** for a default word node,
- **Alt. color** for not recognized words and
- **Group node color** for defined groups of words.

Changing the **Font size** influences all nodes regardless of their type.

In the **Edge Highlighting** section you can specify when an edge of the chart should be highlighted, i.e. drawn with the specified alternative color. This section also allows changing the size of edge description.

The next two dialogue sections influence the appearance of chart edges. You can specify colors used to draw edge lines, their thickness and visibility of their description for both types of edges.

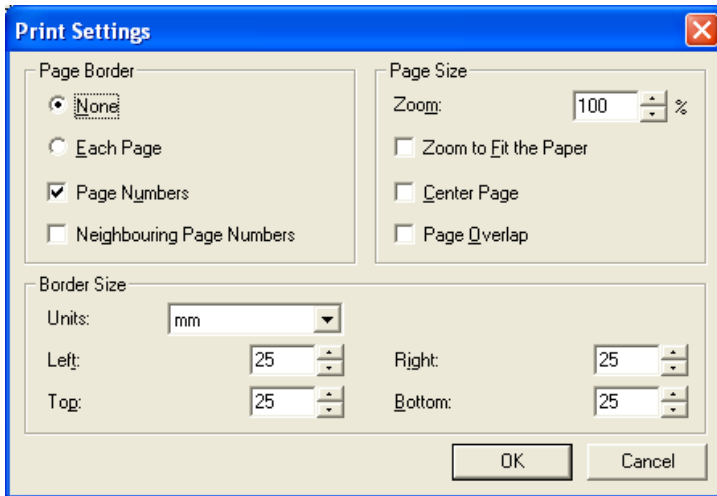
The last section **Other** contains three check boxes:

- If the **Use Simple Connections** check box is selected, then a simple connection will also be displayed in the chart. Otherwise these connections will be skipped.
- Using the **Animation** check box one can disable all animations shown while working with the chart, i.e. zoom animation, layout animation. It is recommended to switch off animations if the chart is very complex and it takes too long to redraw it.
- If **Show Clusters Initially** is selected after documents have been imported into a repository, resulting clusters are shown.

By default the **Customize Chart** dialogue only changes settings for charts of the current repository. But it also allows changing of the default chart appearance which is used for all new repositories as the initial settings. To store current values as default settings, press the **Set Default** button. To restore default settings for the current repository, use the button **Default**.

## Print Settings

Using the button **Print Settings** you can modify how the chart will be printed:



In the **Page Border** section you can select how the border of each printed page should look. It is possible to disable the border at all (radio button **None**) or to add a border to each drawn page (radio button **Each page**). You can also select whether page numbers and/or number of neighbouring pages should be printed.

The section **Page Size** lets you specify the size of the printed diagram. It can be defined either in percentage of the default size or you can select the check box **Zoom to Fit the Paper** to print the chart on one page. The chart can be centered on the page by selecting the **Centre Page** check box. And if more than one page is needed for printing the chart, you specify whether the pages should overlap or not (check box **Page Overlap**).

In the section **Border Size** you can specify the size of the border of the printed page in millimetres or inches.

The way the chart will be split on pages can be checked using the **File/Print Preview** command or by displaying page borders directly in the chart.

## Document Text



The document text view allows you to read the text of a document with highlighted relevant words that have been selected. To switch to this view, use the **View/Document Text** command or press **Alt+2** or use the depicted toolbar button.

The following picture shows a sample document text with highlighted selected terms.

25.9.2000 BELGIUM: EU TO CONSULT WEDNESDAY ON AOL, WARNER, EMI DEALS.  
BRUSSELS, Belgium (Reuters) - European Union regulators will consult experts from the 15 member states on Wednesday before taking their final decision on whether to clear America Online Inc.'s proposed merger with Time Warner Inc.  
At the same time, the EU's merger Advisory Committee will be asked to give its view on the separate, but linked, joint venture between EMI Group Plc and Time Warner's Warner Music unit, EU officials said on Monday.  
The Commission, which has the power of life or death over major mergers in the EU, will present the committee with a draft decision in each case plus a list of concessions offered by the companies to meet its concerns.  
"There'll be an exchange of views on Wednesday, but the Commission has to get the member states to buy into what it's doing," said an official at a rival company being consulted by the Commission on the two deals.  
"It's not clear where things stand."  
The Commission has until October 24 to rule on the AOL-Time Warner merger and until October 18 to reach a decision on the Warner-EMI Music joint venture, but is more likely to rule at its meeting on October 4.

Document text is retrieved using a query sent to Tovek Tools. That is why the text is available only if it is available in Tovek Tools. Once the text has been loaded into Harvester, it is stored in an internal application cache so that all following requests for the document text can be resolved very quickly. The highlighting is computed on the Harvester side.

## Working with Text View

The most common commands for working with text are accessible through the side toolbar:



**Highlight Abstract** button switches between two different highlighting modes for reading the document text. In the default mode only words of the current query are highlighted. An example of such highlighting is displayed in the picture above. In the alternative mode (**Highlight Abstract** button is held down) all words of the document's abstract are highlighted.

25.9.2000 BELGIUM: EU TO CONSULT WEDNESDAY ON AOL WARNER EMI DEALS  
BRUSSELS, Belgium (Reuters) - European Union regulators will consult experts from the 15 member states on Wednesday before taking their final decision on whether to clear America Online Inc.'s proposed merger with Time Warner Inc.  
At the same time, the EU's merger Advisory Committee will be asked to give its view on the separate, but linked, joint venture between EMI Group Plc and Time Warner's Warner Music



**First** button or **View/Highlighted Words/First** command moves the selection to the first highlighted word.



**Previous** button or **View/Highlighted Words/Previous** command (Shift+F8) moves the selection to the previous highlighted word.



**Next** button or **View/Highlighted Words/Next** command (Shift+F9) moves the selection to the next highlighted word.



**Last** button or **View/Highlighted Words/Last** command moves the selection to the last highlighted words.



**Mark for Export** button or **Tools/Mark for Export** command (Ctrl+E) toggles the mark for export state of the displayed document.



**Text Styles** button opens a dialogue window for customizing the look of the document text.

For more details about viewing text of a document see the Tovek Tools user guide, section Tovek Viewer.

## Lists

Lists are dockable windows that can be arranged according to the user needs and that display some additional information or results. There are six lists available in Harvester:

- Word List
- Related Words
- Word History
- Descriptors
- Word Neighbourhood
- Result List

All of these lists can be switched on or off separately (by commands in **View/Lists** menu) or all together using the **View/Hide Lists** command (**F12**). All lists can be switched off by **Chart Only** toolbar button.



**Restore Lists** toolbar button restores lists layout according to standard layout. This standard layout can be replaced by the current layout by pressing the **Restore Lists** toolbar button while holding the **Ctrl** key.



### Word List




Word list displays the list of all words that have been analyzed to be relevant with a set of additional data.

Word	Rel.P...	Doc...	Score	Trend
✖ affairs	5	130	23	-0,025
✖ agenda	2	78	25	-0,033
✖ aid	5	91	26	-0,015
✔ airline	1	12	40	0,012
✔ albanian	1	20	35	0,035
✖ albright	2	36	31	-0,010
✖ ally	1	34	28	-0,011

The list can be sorted according to any displayed column either by clicking the column header or using the **Sort Descending** or **Sort**

---

**Ascending** header context menu commands. The following fields can be made part of the list:

- **Word** contains the string representation of the base form of the word
- **Doc. Count** shows the count of documents containing the word
- **Rel. Pair Count** contains the count of related relevant words.
- **Score** is a value between 0 and 100 showing the importance of the word for describing a theme.
- **Trend** defines the trend of a word's occurrence in time. Negative value means that the word occurrence is falling; positive value indicates that the occurrence is growing. The trend is also indicated by the arrow icon displayed at the beginning of each line:
  -  occurrence of the word is falling
  -  occurrence stagnant
  -  occurrence is growing

To select fields that should be displayed in the list, use the **Fields** header context menu command.

Word list can be used as a starting point for exploring the results of the analysis. The default action on the list is the **Fill Lists** action. It can be performed either by double-clicking a word or by pressing **F5** or using the **Fill Lists** context menu command. It is also possible to run this action for several selected words at once. This command changes the harvester query that is displayed in the Related words list and the content of all other lists will be automatically changed to match the new query.

To add all selected words to the chart run the **Add to Chart** context menu command or drag and drop the selection to the desired position within the chart.

The selected words from the word list can be used to search with Tovek Agent (see Searching for Documents using Tovek Agent) or to build context queries for InfoRating (see InfoRating Context Queries).

## Related Words

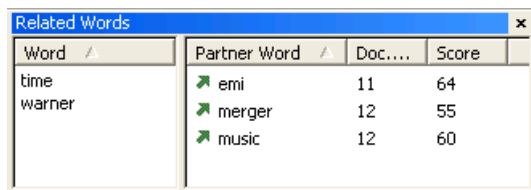
Related words list consists of two parts.

On the left side there is a list of words that build the current *harvester query*. The information showed in another list depends on this query (except the Word list):

- **Related Words** shows all words related to the query.
- **Word History** displays the occurrence history of the query words.
- **Descriptors** list displays all descriptors of the last word that has been added to the query.
- **Word Neighbourhood** shows the neighbourhood of the last word that has been added to the query.
- **Result List** displays all analyzed documents that contain all query words.

You can remove words from the query simply by selecting them in the left list and pressing the **Delete** key (or by running the **Remove** context menu command). If you want to remove all not selected words from the query, use the **Fill Lists** word context menu command.

The harvester query can also be extended by dragging words from other lists or from the chart and dropping them on the query list.



Word	Partner Word	Doc....	Score
time	emi	11	64
warner	merger	12	55
	music	12	60

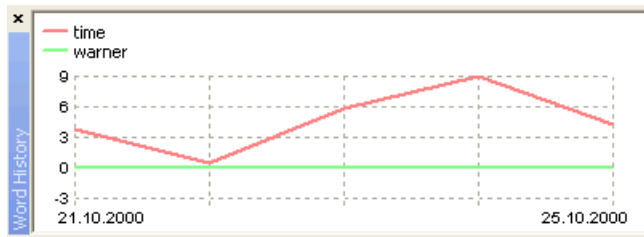
The list on the right side of the Related words list shows all words that are related to all query words. This list also brings the following additional information for these words:

- **Doc. Count** is the count of documents where the word appears together with all query words.
- **Score** is a value between 0 and 100 showing how much the word is related to the query words.
- **Trend** defines the trend of group containing the word and all query words.

Fields that should be displayed in the list can be selected by running the **Fields** header context menu command.

## Word History

Word History graphically represents occurrence of query words in time.



The horizontal axis shows the date when the word appeared. This value is derived from the date of source documents – either it is the value of the document's first date field or it is the date when the document has been analyzed when there is no date field available.

On the vertical axis, the occurrence count is denoted either as an absolute or normalized value. To switch between absolute or normalized values, use the **Norm Slope Graph** context menu command.

## Descriptors

In context of Harvester, a descriptor is the product of the statistical analysis. A descriptor consists of three words where each two words build a relevant pair.

Word 1	Word 2	Word 3	Score	Trend
emi	music	time	58	0,028
emi	time	warner	64	0,028
merger	music	time	50	0,034
merger	time	warner	55	0,034
music	time	warner	60	0,028

This Descriptors list contains three columns **Word 1**, **Word 2** and **Word 3** that hold the words of descriptors alphabetically sorted. There are two additional columns available:

- **Trend** showing the average trend of words building the descriptor.
- **Score** representing the importance of the descriptor.

To select fields that should be visible, use the **Fields** header context menu command.

Descriptors can be added to the chart directly from the Descriptors list by running the **Add to Chart** context menu command or just by dragging and dropping them on the chart view. They can also be used to build context queries for InfoRating (see InfoRating Context Queries) or to refine search queries for Tovek Agent (see Searching for Documents using Tovek Agent).

The Descriptors list can display either a list of all descriptors or descriptors for the current query only. To switch between these two modes, use the **All Descriptors** context menu command.

## Word Neighbourhood

The list Word Neighbourhood shows neighbouring words of the last word that has been added to the current query.

	3< Word	2< Word	1< Word	Word	Word >1	Word >2	Word >3
000925 rt BELGIUM: VIVENDI, SEAGRAM OFFER CONCESSIONS TO EU REGULATORS.	[1] america	[1] online	[1] time	[1] warner	[1] will	[1] have	[1] major
000930 rt BELGIUM: EMI, TIME WARNER IN LAST-DITCH EFFORT TO SAVE JOINT VENTURE.	[2] belgium	[1] ruling	[6] time	[10] warner	[2] merger	[1] october	[1] venture
	[1] vivendi	[1] reuters	[2] aol-time		[1] spokesw...	[1] non-aol	[1] torres
	[1] october	[1] reach	[1] emi		[1] music	[1] music	[1] sell
	[1] final	[1] online	[1] decision		[1] last-ditch	[1] have	[1] save

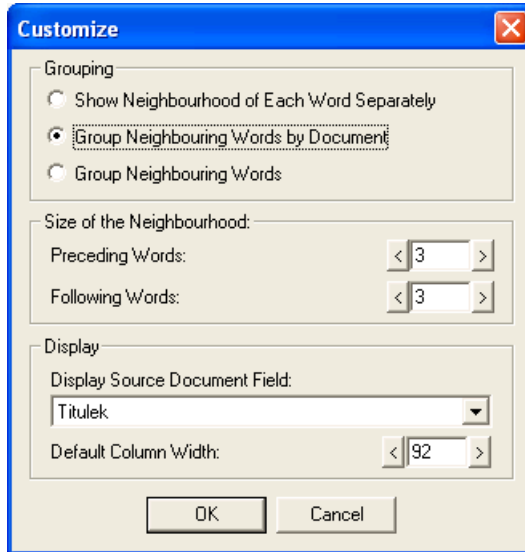
Neighbouring words are computed based on the analyzed documents and show relevant words that precede and follow the last query word.

By default the list is divided into parts by document. Each part is labelled with information about the document; all occurrences of the last query word follow together with their neighbouring words. This gives you an overview about the context in which the last query word appears in the analyzed documents, but it is of course possible to open the document text and inspect its occurrence there. To view the text of the labelled document double click its label or run the **View Text** context menu command.

You can use the listed words similarly as in all other lists to enrich the chart (drag and drop, **Add to Chart** context menu command), to create

context queries for InfoRating (see InfoRating Context Queries) or to refine the search in Tovek Agent (see Searching for Documents using Tovek Agent).

It is possible to define how the neighbourhood information should look using the **Customize Word Neighbourhood** context menu command. This command opens the following dialogue window:



In the **Grouping** section of the dialogue, you choose how the neighbouring words should be grouped:

- **Show Neighbourhoods of Each Word Separately** – for each occurrence of the last query word in a document text, one line containing the neighbouring words is generated.

EMI, TIME WARNER IN LAST-DITCH EFFORT TO SAVE JOINT VENTURE  
 belgium emi time warner last-ditch effort save  
 belgium reuters time warner inc emi group

- **Group Neighbouring Words by Document** – neighbouring words are grouped so that each word appears only once for each document in each column. In the following example you can see that the last query word '*warnar*' appears ten times in the given document where six times the preceding word is '*time*'.

EMI, TIME WARNER IN LAST-DITCH EFFORT TO SAVE JOINT VENTURE  
 [2] belgium [1] ruling [6] time [10] warnar [2] merger [1] october [1] venture  
 [1] vivendi [1] reuters [2] aol-time [1] spokes... [1] non-aol [1] torres

- **Group Neighbouring Words** – selecting the last option causes that the grouping of neighbouring words is not limited to documents but is applied across all words. In this case there will be no information about the source document in the list, just the neighbouring words.

[ 9] billion [ 7] giant [47] time [103] warner [26] emi [24] music [ 5] october  
 [ 8] october [ 6] clear [22] aol-time [11] music [ 6] october [ 5] dominate

In the **Size of the Neighbourhood** section of the **Customize Word Neighbourhood** dialogue you can customize how many preceding and following words should be in the list. The numbers can be set to a value between 0 and 20.


In the **Display** section some aspects of the appearance of the list can be customized. You can select which document field should represent a concrete document in the list. If you select the option **None**, the list will not display any special lines for documents, otherwise the list will be divided into sections according to source documents, and the selected field will be used as the section label (unless the **Group Neighbouring Words** option is selected). The value of the **Default Column Width** setting is used when new columns in the list view are created. The value is given in pixels.


## Result List

The Result List displays analyzed documents that match the current harvester query. Each matching document is represented by a single line that contains its field values.

Title	Country	Date
EU TO CLEAR AOL-TIME WARNER, BLOCK EMI/WARNE...	BELGIUM	2000-09-28
EU TO CLEAR AOL-TIME WARNER WITH CONDITIONS - ...	BELGIUM	2000-10-10
EU STILL MULLING VERDICT ON AOL, WARNER, EMI DE...	BELGIUM	2000-09-27
EU SET TO CLEAR AOL-TIME WARNER, U.S. PROBE RE...	BELGIUM	2000-10-10
EU SAYS NO DECISION YET ON AOL, TIME WARNER, EMI	BELGIUM	2000-10-04
EU RECONSIDERING WARNER EMI MUSIC DEAL SOURCE	BELGIUM	2000-10-02

The mark for export state of the document is reflected in its icon:

 document will not be exported,

 document is marked for export and will be exported as part of exports from the **Tools** menu.

You can mark documents for export by running the **Mark for Export** context menu command (**Tools/Mark for Export** main menu command)

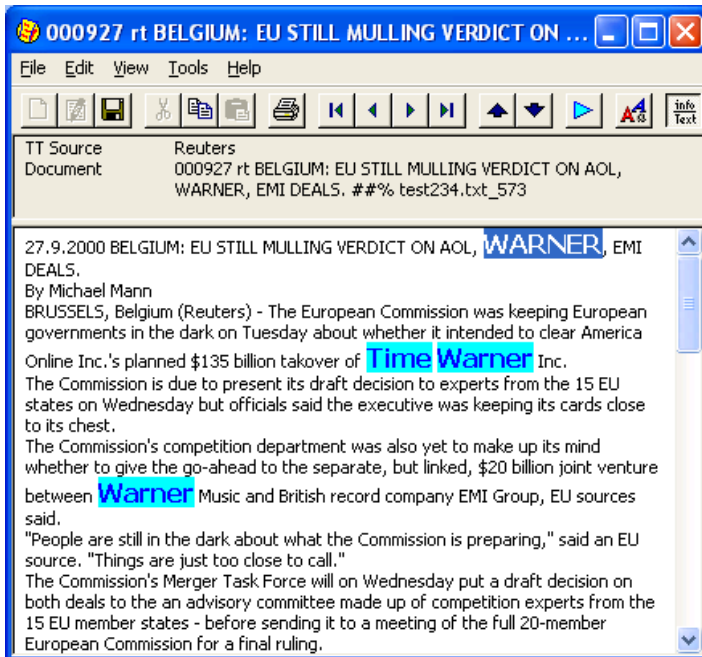
or by pressing the **Ctrl+E** key shortcut. Applying this command for the second time deletes the marking from all selected documents.



When the **Documents for Export** button is down, the result list shows all documents that are marked for export regardless of the current harvester query. To return to the normal state of the Result List panel, release the button. The same functionality is also available through the **Documents for Export** context menu command or the **Tools/Documents for Export** command.

To open the text of the selected document, you can use the **View Text** context menu command or you can simply double click the document within the result list. If it is the first time you are viewing the text of this document, then it will be retrieved using Tovek Agent and stored in the internal Harvester cache. All other accesses to this text are realized through this cache until you restart the Harvester application.

In the same way as in Tovek Agent, you also can use the external text viewer Tovek Viewer. To open a document in the external viewer, press **Shift** while opening it:



For a detailed description of Tovek Viewer, refer to the Tovek Tools documentation.



To display the text of the following document, use the **Next Document** toolbar button or the **View/Next Document** command or press **F9**.



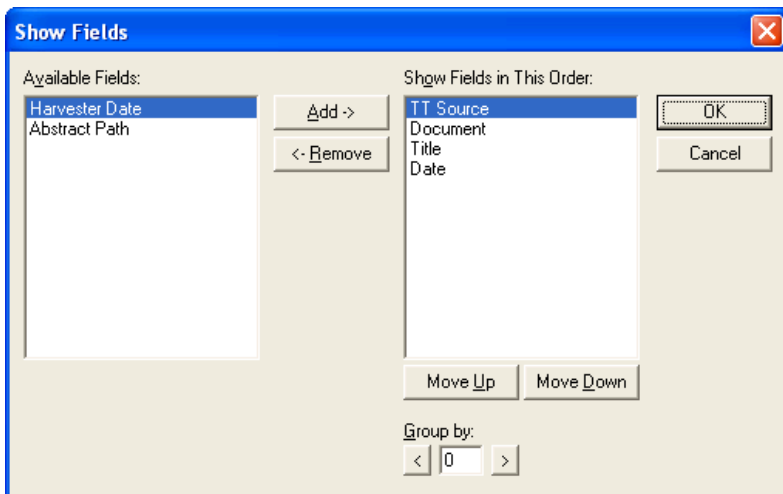
To display the text of the previous document, use the **Previous Document** toolbar button or the **View/Previous Document** command or press **F8**.

## Document Fields

The result list can display all available document fields. There are a few default document fields that are available for all documents; all other fields that should be available must be imported together with the documents. The default fields are the following:

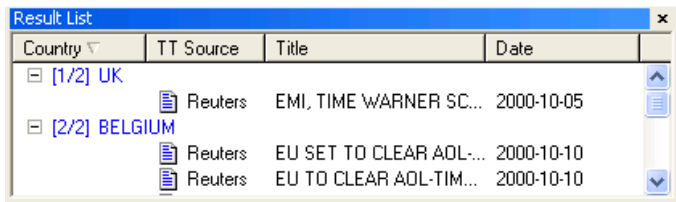
- **Document** – document identifier, either a document key or a path to the document file.
- **TT Source** – source of the document
- **Harvester Date** – date used to compute trends; derived from a documents date field or the time when the document was analyzed.
- **Abstract Path** – path where the document's abstract is stored.

To select visible document fields, run the **Fields** header context menu. This command opens the following dialog, which lets you customize the result list:



Fields listed in the right list box will be shown in the result list in exactly the same order. Use the **Add ->** and **Remove <-** buttons to move fields between the list boxes holding available or visible fields. To change the order of visible fields, use the buttons **Move Up** and **Move Down**.

The result list also enables you to group listed documents according to values of their fields. To group documents according to a given field, move this field to the first position and set the **Group by** value to 1. The following picture shows a result list where documents are grouped according to the *Country* field.



The screenshot shows a window titled "Result List" with a table of search results. The table has four columns: "Country", "TT Source", "Title", and "Date". The results are grouped by country. The first group is "UK" with one document. The second group is "BELGIUM" with two documents. Each document row includes a small icon of a document, the source "Reuters", a truncated title, and the date "2000-10-05" or "2000-10-10".

Country	TT Source	Title	Date
[1/2] UK	Reuters	EMI, TIME WARNER SC...	2000-10-05
[2/2] BELGIUM	Reuters	EU SET TO CLEAR ADL...	2000-10-10
	Reuters	EU TO CLEAR ADL-TIM...	2000-10-10

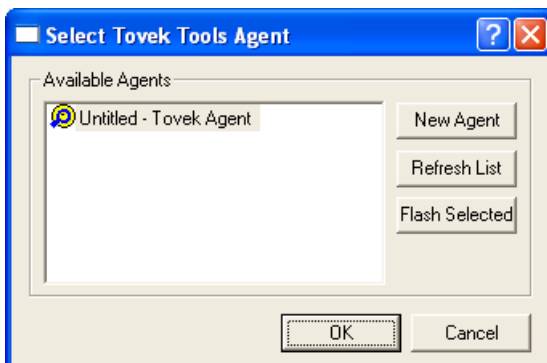
You can also use more than one field to group documents by increasing the **Group by** value. To switch off the grouping, set the **Group by** value to 0.

## Searching for Documents using Tovek Agent

If you are also interested in documents that are not part of the analysis, you can still use the extracted keywords for searching using Tovek Agent.



The search can be started from the lists or from the chart view by pressing the **Search using Tovek Tools** button or by running the **Tools/Search using Tovek Tools** command (this command can also be found in the context menu of the lists or the chart view).



When you start a search for the first time, the above depicted dialogue window appears. In this dialogue you must select the Tovek Agent instance that should be used to run the search. Using the button **New Agent**, it is possible to start a new instance of Tovek Agent.

Queries that are sent to Tovek Agent are generated according to the current selection and according to the query settings.

By default the created query contains all keywords connected by NEAR operator that are further connected with the original query by the <AND> operator. The original query is the query that has been used to retrieve the analyzed document.

## Customizing TT Query

The default way of generating the query is not necessarily always suitable. Harvester allows you to customize the resultant query using the following dialogue:

**Customize TT Query**

Tovek Tools Query

Use Original Query

Connect With New Using

New Query Settings

Query Prefix\*

Query Postfix\*

Word Prefix

Word Postfix

Word Delimiter

\* Used only for queries which consist of more than one word

Sample Queries

Query Containing the Word "hello":

Query Containing the Words "How", "are", "you":

OK Cancel

This dialogue can be opened through the **Tools/Customize TT Query** command.

In the first window section you can decide whether the original query should be part of the generated one. In addition, you can select the operator that should be used to connect the original query with the generated one. By default the original and the generated queries are connected using the <AND> operator.

The second section, **New Query Settings**, allows you to modify how the generated part of a query will look. In this section you can specify

the **Query Prefix** and **Query Postfix** that surround the whole generated query part. Similarly you can define the **Word Prefix** and **Word Postfix** that surround each word of the new query. The last setting is the **Word Delimiter**. The entered string will be placed among the query words.

The last part of the dialogue shows two sample queries that illustrate how the generated queries will look. The first sample is a query with only one word; the second one will be generated for more than one word.

## InfoRating Context Queries

---

Context queries can be exported from following views and lists by running the **Tools/InfoRating Context Export** command or by running the **InfoRating Context Export** context menu command. These commands generate a set of context queries according to current selection and send this set to a running InfoRating application.

### Word List and Related Words List

You can create context queries from selected words shown in these two lists. Harvester will generate one context query for each selected word. Generated queries are very simple, query text and query id equal to the selected word.

### Descriptors Panel

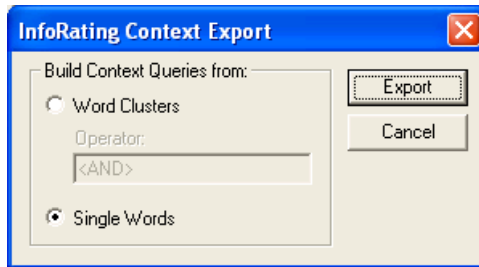
Context queries created according to selected descriptors are somewhat more complex than the ones based on single words. Query text of such queries is build from descriptor words connected by <AND> operator. Query id consists of the same words connected with underscore. So for the descriptor (*company, music, warner*), Harvester generates a query *company\_music\_warner* with query text *company <AND> music <AND> warner*.

### Word Neighbourhood

Queries exported from this panel contain all neighbouring words listed in the selected line connected with <ACCRUE> operator. Their id is built from the central word which is extended by a running number. This number will be reset when the application has been restarted.

## Chart View

It is also possible to create context queries from the selected part of the chart. After running the previously mentioned commands the following dialogue window appears:



It offers you two possibilities. The first and the default one is to build queries based on the chart clusters. If you select this option, the selected part of the chart will be analyzed and the words will be grouped so that each two selected words that are connected will be put into the same group. One query is generated for each group. It contains all words from the group connected with the given operator (by default <AND> operator).

The second possibility is to build simple context queries using single words only. The result is the same as if you were to build the queries from the same word within the Word List.

# Exports

---

## Html/XML Export

Selected documents (see Document Text and Result List) can be exported as one HTML or XML file by **Html/XML Export** command from **Tools** menu or by **Ctrl+T** shortcut. A detailed description of all possibilities to export documents in html or xml can be found in the Tovek Tools and Tovek Agent manual.

## Export of Documents to Tovek InfoRating

Documents marked for export (see Document Text and Result List) can be transferred by **Tools/InfoRating Export** command to the Tovek InfoRating document pane (see Tovek InfoRating Guide).

## Settings

---

**Default settings** will be used when you create a new repository (analysis), **Settings of Current Repository** will be used for an active repository only.

If you make changes concerning calculation in an active repository, you have to recalculate statistics by the **Repository/Recompute Statistics** command.

### Default Settings

The **Tools/Default Repository Settings** command opens the window Default repository settings, where you can change setting on the respective tabs:

- **General**

In the **Query** part you can set whether to work with a word together with its stems (**Use Stemmer** option) and whether to redraw a diagram immediately after the query changes (**Show Stemmed Query** option).

In the **Other** part you select whether words without related words should be displayed too (**Display Words with no Partners too** option)

The **Use Active Repository without Asking** and **Use Temporary Repository without Asking** options make it possible to affect how Harvester imports documents (see Selecting data for Harvester). If you select one of these options, the respective repository will be used automatically without prompting you. If you select both options, you can moreover predetermine which repository will be preferred (**Prefer Active to Temporary Repository** option).

- **Abstract**

The **Language Configuration** part is not active for the standard setting and is used only to change the language or language component to be used for active analysis abstracts calculation.

In the **Content** part you can set whether to work with nouns only (**Only Nouns** option), whether to include unrecognized words in the analysis (**Include Unrecognized Words** option), whether to include

numbers also (**Include Numbers** option) and what minimal length of word should be included in the analysis (**Minimal Word Length** option).

In the **Document Age** part you can select whether to work with all documents or only with documents not older than the given number of days.

- **Word Statistics**

Based on word statistics values you can impact which words are suitable for further processing in the current analysis. Word suitability is derived based on three statistical parameters:

1. minimal number of documents where the word appears. The standard value of this parameter is 3.
2. sum of relative occurrences characterizes how often the word appears in documents. The standard value of this parameter is 0.3.
3. variational occurrence coefficient

In the **Count** part you can increase the number of relevant words by moving the scroll bar to the right.

In the **Time Statistics** part you can change settings for Word history (see Harvester Window above). You can set how many intervals will be on the horizontal axis (**Step Count** option), how long one interval will be (**Step Span in Days** option) and what change of word occurrence (or pair or trio occurrence) will show graphically (**Slope Bound** option). By moving the scroll bar to the right, you can increase the number of occurrences needed for a change in the graph.

After pressing the **Advanced** button, the scroll bars change to number values and you can more precisely define relevant words and slope bound.

- **Pair Statistics**

In the **Build** part you can set the maximal distance between words to still be considered as a pair.

In the **Resolution of the Statistics** part you can set **Connection Count** and what the **Word Distance** and **Word Dependency** for two words to be bound into a pair will be.

After pressing the **Advanced** button, the scroll bars change to number values and you can more precisely define the parameters:

1. minimal number of documents where the pair of words appears
2. minimal near factor per document determines how far apart the words in a pair can be (the lower the near factor, the larger the distance can be)
3. minimal correlation coefficient of occurrence of both words in a pair ensures that both words appear in similar documents
4. relative near factor is a value similar to near factor per document, but is based on different input values
5. dependency characterizes the dependency of word occurrence in a document on the second word occurrence
6. dependency with occurrences characterizes dependency of value of word occurrence on the value of the second word occurrence

- **Company Values**

In the **Company Values** part you can reset made changes to default values. By **General Settings**, **Abstract Setting**, **Word Statistics Settings**, **Pair Statistics Setting** and **Time Statistics Settings** buttons you can reset settings on respective tabs to defaults. The change shows after pressing the **OK** or **Apply** buttons.

## Settings of Current Repository

**Repository/Settings of Current Repository** opens the window for setting property of open repository; you can change the settings on respective tabs:

- **Information**

In the **Repository Path** part there is information as to where the file with active repository is located and as to its size.

In the **Statistics** part there is information about the number of documents, words, relevant words, pairs, relevant pairs and descriptors in an active repository.

The values on this tab are for information only, you cannot change them.

- **General**

The settings on this tab are the same as in **Default Settings** (see **Default Settings** above), but the settings performed here are for an active repository only.

- **Abstract**

The settings on this tab are the same as in **Default Settings** (see **Default Settings** above), but the settings performed here are for an active repository only.

- **Word Statistics**

The settings on this tab are the same as in **Default Settings** (see **Default Settings** above), but the settings performed here are for an active repository only.

- **Pair Statistics**

The settings on this tab are the same as in **Default Settings** (see **Default Settings** above), but the settings performed here are for an active repository only.

- **Company Values**

The settings on this tab are the same as in **Default Settings** (see **Default Settings** above), but the usage performed here is for an active repository only.

## Printing from Harvester

---

### Print

You can print a diagram by **File/Print** command. You can select printer, print range and number of copies.

### Print Preview

Before printing the diagram, you can display a print preview by **File/Print Preview** command.

A larger diagram will be divided into several pages according to Print Settings (see Connection Chart, Chart Settings). You can see how the diagram will be divided into several pages directly in the diagram by Paper Breaks Preview (see Paper Breaks Preview).

### Page Breaks Preview

If the diagram is divided into several pages (see Connection Chart, Chart Settings), you can display the dividing lines between pages by activating the **Chart/Page Breaks Preview** command.

## Index

---

### A

Abstract, 27, 44, 46, 47  
Abstract Path, 36  
Actual Size, 16  
Add to Chart, 16, 29, 32  
Add to Query, 22  
Advanced, 10  
All Descriptors, 32  
Alt. Color, 23  
Animation, 24

### B

Background Color, 23  
Border Size, 25  
Build, 45

### C

Centre Page, 25  
Circular Layout, 16  
Close Repository, 11  
Cluster, 24  
Company Values, 46, 47  
Configuration, 10  
Connection Count, 45  
Connection Chart, 14  
Connections Layout, 17  
Content, 44  
Context Export, 41  
Count, 45  
Create Word Group, 19  
Current Repository, 44, 46  
Currently Active Repository, 9  
Customize Chart, 23  
Customize TT Query, 39  
Customize Word  
  Neighbourhood, 34

### D

Decrease Edge Length, 17

Default Column Width, 34  
Default Repository Settings, 44  
Default Settings, 44, 47  
Descriptors, 13, 30, 31, 41  
Display, 34  
Display Words with no Partners  
  too, 44  
Doc. Count, 29, 30  
Document, 36  
Document Age, 45  
Document Text, 26  
Documents for Export, 35

### E

Edge Highlighting, 24  
Exploration Mode, 18  
Export, 41, 43  
Export to File, 21

### F

Fields, 29, 31, 32  
Fill Lists, 22, 29, 30  
First, 27  
Fit to Window, 16  
Font Size, 23  
Force Different Language for All  
  Documents, 10

### G

General, 44, 46, 47  
Graph Mode, 18  
Group Neighbouring Words, 34  
Group Neighbouring Words by  
  Document, 33  
Group Node Color, 23  
Grouped Layout, 17  
Grouping, 33

### H

Harvester Date, 36

Harvester Export, 8  
Harvester Query, 30  
Help, 7  
Hide Lists, 28  
Hide Not Connected, 17  
Hide Selection, 17  
Hierarchy layout, 16  
Highlight Abstract, 27  
Html/Xml Export, 43

## Ch

Chart Only, 28  
Chart Settings, 23  
Chart View, 42

## I

Import From File, 21  
Imported Documents, 9  
Include Unrecognized Words, 44  
Increase Edge Length, 17  
InfoRating Context Export, 41  
InfoRating Export, 43  
Information, 46  
Insert Related, 16, 17

## L

Language Configuration, 44  
Last, 27  
Leave Only Descriptors, 17  
Library, 10

## M

Main Pane, 14  
Manage Word Groups, 21  
Manual Version, 7  
Mark for Export, 27, 34  
Minimal Word Length, 45  
Modify Group, 21

## N

New Agent, 38  
New Chart, 11

New Query Settings, 39  
New Repository, 11  
Next, 27  
Next Document, 36  
Node Appearance, 23  
Norm Slope Graph, 31  
Nouns, 44

## O

Only Nouns, 44  
Open Chart, 11  
Open Repository, 11  
Original Edge Length, 17  
Other Publications, 7

## P

Page Border, 25  
Page Overlap, 25  
Page Size, 25  
Pair Statistics, 45, 46, 47  
Paper Layout, 48  
Prefer Active to Temporary  
Repository, 44  
Preview, 19  
Previous, 27  
Previous Document, 36  
Print, 48  
Print Preview, 25, 48  
Print Settings, 25  
Publications, 7

## Q

Query, 44  
Query Postfix, 40  
Query Prefix, 40

## R

Recompute Statistics, 44  
Rel. Pair Count, 29  
Related Words, 12, 29, 30, 41  
Remove Group, 21  
Remove Simple Edges, 17  
Repository, 9, 11, 44, 46

---

Repository Path, 46  
Resolution of the Statistics, 45  
Restore Lists, 28  
Result List, 13, 30, 34

## S

Save Chart, 11  
Save Repository, 11  
Score, 29, 30, 32  
Search using Tovek Tools, 38  
Select All, 18  
Select Repository, 9  
Select Shared Nodes, 18  
Select Single Nodes, 18  
Set Default, 24  
Settings, 23, 44  
Settings of Current Repository,  
44, 46  
Show Clusters, 16  
Show Clusters Initially, 24  
Show Neighbourhoods of Each  
Word Separately, 33  
Show Stemmed Query, 44  
Simple Edges, 17  
Size of the Neighbourhood, 34  
Slope Bound, 45  
Statistics, 46  
Step Count, 45  
Step Span in Days, 45

## T

Temporary Repository, 9  
Text, 26  
Text Styles, 27  
Time Statistics, 45, 46  
Tovek Agent, 8, 38  
Tovek Tools, 38  
Tovek Viewer, 35  
Trend, 29, 30, 32

TT Query, 39  
TT Source, 36

## U

Unrecognized Words, 44  
Use Active Repository without  
Asking, 44  
Use Simple Connections, 24  
Use Stemmer, 44  
Use Temporary Repository  
without Asking, 44

## V

Version, 7  
View Text, 32, 35

## W

Word, 29  
Word Delimiter, 40  
Word Dependency, 45  
Word Distance, 45  
Word Group, 19  
Word History, 12, 30, 31  
Word List, 12, 28, 41  
Word Neighbourhood, 13, 30,  
32, 41  
Word Postfix, 40  
Word Prefix, 40  
Word Statistics, 45, 46, 47

## Z

Zoom In, 16  
Zoom Out, 16  
Zoom to Area, 16  
Zoom to Fit the Paper, 25