CassBeth

Document Analysis Manual

2006

Prepared By

CassBeth Inc
Cherry Hill, NJ, 08003

Contact Information

sat@cassbeth.com

www.cassbeth.com

856-751-9795

Forward

Thank you for considering our tools. We believe that we have significant new products and technologies that will change the way we create and review documents. It is no secret that the computer and the Internet revolution have provided us with so much data that we are unable to effectively consume this data. Our products and technology address this serious issue and we are working so that they will become as common as word processing.

Each tool is packaged with online help. This document is at a higher level than the online help. We strongly recommend that you read this document and the online help content.

Table of Contents

1.0 Introduction

Thank you for considering our document analysis tools. We believe that we have significant new products and technologies that will change the way you create and review documents. It is no secret that the computer and the Internet revolution have provided us with so much data that we are unable to effectively consume this data. Our products and technology address this serious issue and we are working so that they will become your tools of choice.

Computers and the Internet have revolutionized our world. With that revolution two broad trends have surfaced, Information Overload and Loss of Human Assistants and Mentors.

1.1 Information Overload

There is too much to read and not enough time to read it. The Internet has become the expert systems that were attempted in previous decades but rather than disparate solutions, there is one big brain with a vast amount of information. It has become the ultimate expert system providing insight and solutions to millions every hour of every day. This is possible because people are willing to offer and access this information.

1.2 Loss of Human Assistants and Mentors

Automation has removed many layers of human assistants and mentors that previously existed. We now have a significant void with significant impact on productivity. So we have documents to read but no staff to do the highlighting, digesting, and summarizing (VU-Graph form or white papers). We also have separated our mentors from apprentices so that it becomes impossible to write a simple technical specification. Check lists or style guides are not written, and if they do exist are lost or ignored by the organization as being an old technique. But progress is only possible if someone or something performs the role of the assistants and mentors.

2.0 Background

Our products perform detailed analysis of documents using search criteria defined, saved, and exchanged by the user. They are like a threshing machine separating the wheat from the chaff so that any document can be clearly written, read, and understood.

Our document analysis products are based on an Analysis Tool Engine that works like an Internet search engine except instead of returning web pages they return text blocks from a document. The search criteria have more attributes and settings than an Internet search engine. The search settings are grouped into services and rules that are saved as templates. The products come packaged with default templates that contain default services and rules. The user loads a template, submits a document for analysis, if needed modifies the services and rules in the original template, and saves the new settings as a new template. The product runs on a personal computer and uses the default web browser for the interface. It feels like a web page Internet experience where an agent is working for the user to answer questions about a document as needed.

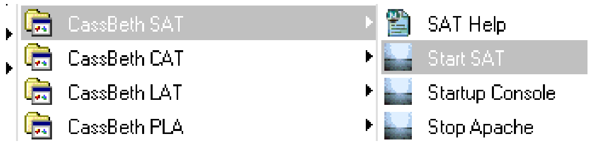
Our products are sold with either an annual or a perpetual license. They use custom activation keys bound to the client name and an expiration date. The latest versions of our document analysis products are downloadable off the CassBeth.com web site.

This surfaced while trying to address a need identified by Carnegie Mellon University to help engineers write better specifications. We quickly realized that a tool framework was needed where the rules for specification analysis must easily change because each organization has a unique view of what is a good specification. We also realized that the rules are not proprietary or hidden but open for all to view. We also realized that this framework applies to everyone who writes reads and tries to understand documents.

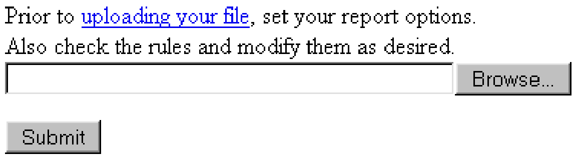
3.0 Installation

Download the software from our web site into a known directory. Double-click the executable and follow the install wizard instructions. Do not modify the install location. If you have a CD, insert the CD and wait for the install wizard. If the install wizard does not automatically start, access the CD and locate the install<product>.exe or the index.html files and double click. Follow the install instructions.

After you have installed the software, locate the icon on the desktop or start menu and start the application. This will start an internal Apache server, automatically open your default wen browser, and take you to the default internal web page that is the control panel. If your Apache server fails to start make sure there is no other web server running or that your firewall has not blocked this internal application. Some applications now use an internal web server, shut them down.

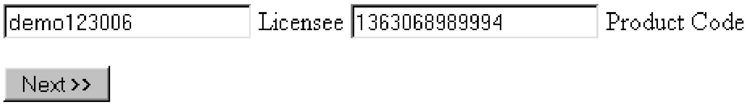


Once the application is started, press the <submit> button.



Prior to uploading your file, set your report options.
Also check the rules and modify them as desired.

The first time the <submit> button is pressed you will be prompted for activation keys. Enter the keys you have been provided. Respond to the license agreement a second time. Select the link to start using the software. You have now completed installation.

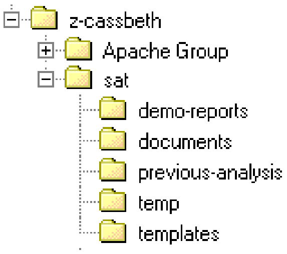


demo123006 Licensee 1363068989994 Product Code

4.0 Getting Started

4.1 Document Location

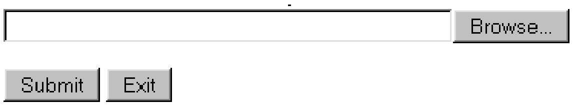
Locate the document you would like to analyze. You can keep the document it its current location or you can move it to the <documents> folder of your tool using windows explorer. Moving it to the tool folder will allow you to use the <library> hyperlinks making it more convenient for you to quickly access your work. The <documents> folder is located in the <z-cassbeth> folder immediately off your drive.



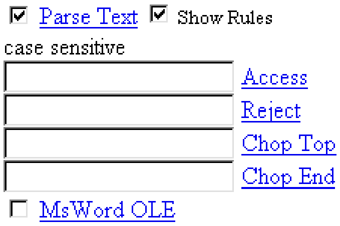
You have a choice, you can keep the document in its current form or you can save it as a text document. If you save it as a text document, open it up a second time in text format and remove the table of contents and any back end information that may have been in appendices.

4.2 Loading a Document

To load a document select the <browse> button and navigate to your document.



If you did not convert the document to text format, you may need to parse the document. If the document is <html> or <htm> the tool will automatically check the <strip html tags> option. If the document is another format you must select the <Parse Text> option.



☑ Parse Text ☑ Show Rules
case sensitive
Access
Reject
Chop Top
Chop End
☐ MsWord OLE

Expand the parse text rules by selecting <Show Rules>. When enabled the tool will look for complete thoughts based on punctuation. These complete thoughts will be converted to stand alone objects that will be returned by the tool rules when you begin analysis. If you do not like the parse demarcation points for an object, then use the text format of the document and manually create the parse demarcation points by placing each standalone object on its own line.

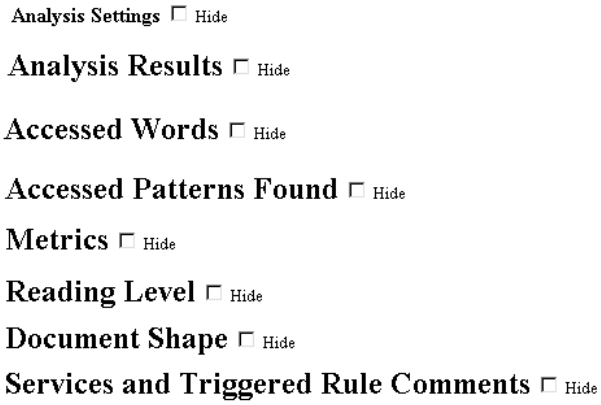
You can preview your download by selecting the <Show Processed Upload> check box. However if the document is very large you may just want to start rule processing and see how your objects are parsed as various rules are triggered.

☑ Show Processed Upload

Remember to upload the document each time you change you parse rules or the analysis results may not match the hyper links for each object when you select a PUI link.

4.3 Processing a Document

Once you have uploaded and parsed you document, select a service from one of your templates and press the submit button. The analysis will return different display areas. Use the check boxes to enable and disable the display areas as needed.


Analysis Settings □ Hide

Analysis Results □ Hide

Accessed Words □ Hide

Accessed Patterns Found □ Hide

Metrics □ Hide

Reading Level □ Hide

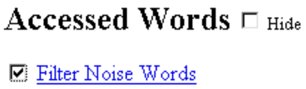
Document Shape □ Hide

Services and Triggered Rule Comments □ Hide

5.0 Creating Your Own Rules

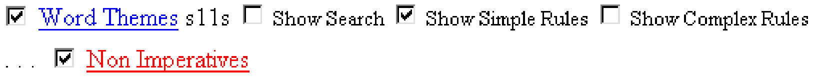
Become familiar with the various services and rules in the different templates. This can happen only after repeated use where you subject a document to these services. You will then be able to determine if you need to modify an existing template and its services, create a new service within an existing template, or create new services in a new template.

One of the best ways to start creating a new service is to "profile a document". This is setting a rule to dump all the keywords in a document. The results are provided in two display areas. The first is sorted by word count and the second is sorted alpha numerically by the words. When a document is profiled all the words are dumped from the document. We recommend that prior to running this service you select the option to filter the noise words.

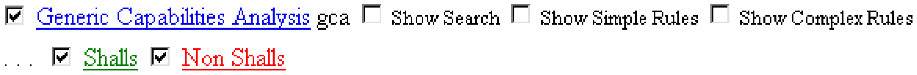

Accessed Words □ Hide

☑ Filter Noise Words

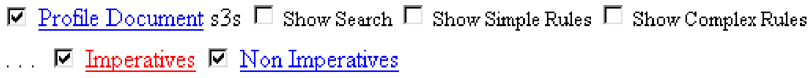
All of the templates have a service that profiles a document. You can start there or you can create your own service and rule to profile a document. These are all examples of rules in different templates that profile a document. Even though the names are different they do the same thing.



☑ Word Themes s11s □ Show Search ☑ Show Simple Rules □ Show Complex Rules
. . . ☑ Non Imperatives

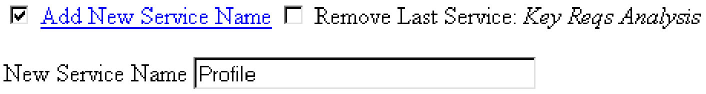


☑ Generic Capabilities Analysis gca □ Show Search □ Show Simple Rules □ Show Complex Rules
. . . ☑ Shalls ☑ Non Shalls



☑ Profile Document s3s □ Show Search □ Show Simple Rules □ Show Complex Rules
. . . ☑ Imperatives ☑ Non Imperatives

To create you own rule to profile a document, select the <Add New Service Check Box>, press the submit button, name the new service, and press the submit button.



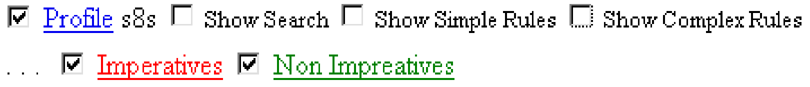
☑ Add New Service Name □ Remove Last Service: *Key Reqs Analysis*

New Service Name Profile

Next follow these steps:

1. Select your new <Profile> service
2. Select <Show Simple Rules> and <Show Complex Rules>
3. Press <submit>
4. Fill in the rule <Name> with "Imperatives"
5. Enter "RED" for the <Color>
6. Fill in the <Access Object > text area with "shall"
7. Select <Hide Accessed Objects> and <Count Accessed Words>
8. Press <submit>
9. Repeat sequence for second rule except set name to "Non Imperatives", Color to "GREEN", fill in <Access Object > with ".", and <Reject Object> with "shall"

You now have a new service to profile your document. Select the rules, uncheck <Show Simple Rules> and <Show Complex Rules>, Select the <Filter Noise Words>, and hide all the display areas except for <Accessed Words> and <Analysis Results>.



☑ Profile s8s □ Show Search □ Show Simple Rules □ Show Complex Rules
. . . ☑ Imperatives ☑ Non Impreatives

To start the process of gaining insight into new rules and services, look at the keywords. Examine the words with large counts. Check one or more words, press the submit button and look at the objects. Start to take note of new keywords, services, and rules. Do not forget to look at words with very low counts.

Start to create a new service and a rule. For initial runs check the <Count Accessed Objects> and <Count Accessed Patterns> options. Run a what-if sequence with your new rule even if it is just with one or two Access Object words. Turn off the original "Profile" service that you used to create this service and rule. As objects are returned in context with the new profile and the new service with the simple rule, new rules will start to surface. Enable the "Profile" and new service and rules in alternate sequences.

If needed take hand written notes on possible new services, rules, and patterns as you work rather than try to author the rules on the fly. Once you start this process you will quickly start to overflow with ideas and you may want to organize them before you invest too much time in creating a final new template, services, and rules. In fact, treat this current template as a temporary product that is used to start the idea process.

Eventually you will create and or modify one or more rules in a service and one or more services. You need to save this work as a new template. If you would like to save a new service and its rules:

1. Select the Template Comments option and press the Submit button
2. Enter the Title and Description for this new template and press the Submit button
3. Once you have completed authoring this new service and its rules use the browser File save option
4. Select the option to only save the HTML page, NOT the whole page
5. Save this new template into z-cassbeth\sat\templates

You can save this new template anywhere, but we suggest you use the SAT Library Directories.

6.0 Heuristics of Document Analysis

To understand this section you should read the Getting Started and Creating Your Own Rules sections.

6.1 Current Template or Slightly Modified Template

1. Locate the best template to start the initial analysis of the document. If needed try various templates and services until you decide on the starting template.
2. If the analysis can be accomplished with a current template, run the analysis by selecting each service and filtering the display as desired.
3. If needed enable the <Show Search> option and tweak the access criteria.
4. Save and export the results as needed.

6.2 New Template or Massively Modified Template

1. If no template is applicable, pick a template you will modify and grow.
2. Start with a Profile Document Service.
3. Try to locate some keywords.
4. Check these keywords, re-run the analysis, and look at the objects.
5. Create a simple Search service and enable the <Show Search> option.
6. Do simple searches based on your Profile Document mining efforts.
7. Add a service and one or more rules based on your keyword and search findings.
8. Keep expanding the services and rules until you feel you are done.
9. Do not be afraid to write notes on paper.
10. At some point group and organize your notes into what will be your final template.
11. Save your initial working template(s).
12. Create your final template based on your notes and initial template efforts.

7.0 Reading Level

To determine the reading level the Count options need to be enabled in the rules. Some templates offer an existing service like <The Generic Capabilities Analysis> service that mines all the words in a report and has a Count option enabled. To determine the reading level:

1. Select the file to be analyzed by pressing the Browse button
2. Select a service with <Count Accessed Words> enabled
3. Check the <Hide> options in the unwanted display areas
4. Press the Submit button

You can scroll through the report or select the links under Report Areas to access the findings. In this case there are findings in the report areas that can be accessed by selecting the Accessed Words, Reading Level, and Metrics links. Select the Reading Level link.

8.0 Document Shape

To determine the document shape you need to create a service that has a rule setting for each paragraph level in the document and select the <Count Child Objects> option. There are templates, which include this service, like the <Generic Structure Analysis> service. When you submit the document for this analysis remove the table of contents to prevent data skew.

1. Select the file to be analyzed by pressing the Browse button
2. Select the <Generic Structure Analysis> service or what you created

5. Check the <Hide> options in the unwanted display areas

3. Press the Submit button

You can scroll through the report or select the links under Report Areas to access the findings. In this case there are findings in the report areas that can be accessed by selecting the Analysis Results, Metrics, and Document Shape links. Select the Document Shape link.

9.0 Control Panel

The control panel is used to set document parsing parameters, filter your display results, reset your analysis, and exit the application.

Control Panel	**Description**
Default Rules	This resets all parameters and starts the engine with the default rules.
No Rules	This resets all parameters and starts the engine with no rules. It is a clean slate for the advanced user.
File to upload	Upload the file using the browse button. Once a file is uploaded it stays in the temp directory. Each time the engine is ReStarted all temp directory contents is deleted.
PUI Mask	If your are exporting from another tool and wish to preserve the PUI, use a mask that looks like the PUI in the export.
Imperatives	Imperatives are words and phrases that command something must be provided. Imperatives in descending order of strength: Shall, Must, Must Not, Is Required To, Are Applicable, Responsible For, Will, Should
Process Only Imperatives	Descriptive text is filtered from the analysis. Only objects matching the imperative pattern are accessed.
Parse Text	Use this option to parse your file into meaningful objects. Obviously the best approach for analysis is to use a document that is properly parsed. However if that is not possible, use this option to attempt an automated parse.
Access - Parse Text	As part of trying to parse a document you may need to remove garbage text. Use this option to access only desired text.
Reject - Parse Text	As part of trying to parse a document you may need to remove garbage text. Use this option to remove undesired text such as tables of contents, headers, footers, and other irrelevant text from a document.
Chop Top - Parse Text	This will remove all the text starting from the top of a document up to the first instance that this pattern is detected. This pattern is part of your valid results.
Chop Bottom - Parse Text	This will remove all the text starting from the bottom of a document up to the first instance that this pattern is detected. This pattern is part of your valid results.
MsWord OLE - Parse Text	If you have Microsoft Word installed, this option will load the file using Microsoft Word services. This will remove garbage text that is found when a Microsoft Word binary file is loaded and viewed as text. The first time this option is executed, Microsoft Word creates a VBE folder within the engine folder. You can leave or delete this folder.
Strip HTML Tags	Filters tags in uploaded HTML files. This is done by removing all carriage return line feeds, since many HTML editors will split text across lines. Objects are established by looking for <BR>, <P>, <LI> and <H.> tags.
Strip Blank	Strips blank lines. Use this filter when working with norm values entered for each rule.

Control Panel	**Description**
Lines	The norm is a percentage of the total lines and there is no reason to skew the data with blank lines.
Access Object	This is a global filter that is applied to the analysis results. Placing a pattern in this text box will only return objects with the pattern. Use this to refine your analysis. The report state is maintained when it is saved.
Reject Object	This is a global filter that is applied to the analysis results. Placing a pattern in this text box will remove objects with the pattern from the results. Use this to refine your analysis. The report state is maintained when it is saved.
Access Risk	This is a global filter that is applied to the analysis results. Placing a pattern in this text box will only return objects associated with a risk that matches the pattern. Use this to refine your analysis. The report state is maintained when it is saved.
Show Processed Upload	If you don't trust your upload, you can view the uploaded file to be processed by the engine. This is also a useful feature when uploading binary documents such as word format files. Use this if you also want a full context view of the findings.
Show Comment Details	Checking this box will attach the rule comments to each reported object text item. For the daily user this becomes noise. Leaving this item unchecked provides links for each reported object text item to a common area that summarizes all the service descriptions and rule comments.
Hide All Comments	Checking this box will hide all object comments in the Analysis Results report area. This allows a user to copy and paste all mined objects without dealing with non document data.
Hide Checked Items	Next to each object in the Analysis Results is a check box. The user can select any of these check boxes. Use this to track your decisions for each object, such as this is not an issue at this time. Selecting the Hide Checked Items is a display filter. When the Hide Checked Items display filter is checked, all the checked objects will be hidden. The report state is maintained in all cases when it is saved. The check boxes are tracked by the PUI. If the PUI changes, then the check box settings may no longer be valid, and must be re-examined.
Save Results	When this option is checked a tab delimited file is saved as "srbd.xls" in the "previous-analysis" folder. This can be used to update your System Requirements Database (SRDB) using an Excel import. This can also be used as a copy and paste from Excel to MsWord. This will paste as a MsWord table. This file is over written each time the analysis is executed. There are 4 ways to communicate your results to the team: 1. Save the HTML report, after you are done, 2. Copy and pasted the HTML report areas into MsWord, 3. Use the srdb.xls file as an import into your SRDB, 4. Use the srdb.xls file to copy and paste into MsWord
Filter Noise Words	Checking this box will filter all noise words. Once checked upon submit the user is presented with a text field to modify the noise words. Changing the pattern from . | to .....| in the noise list will filter all words with less than 5 letters.
Save Metrics	When this option is checked a tab delimited file is saved as "metrics.xls" in the "previous-analysis" folder. This can be used to update your Metrics using an Excel spreadsheet. This can also be used as a copy and paste from Excel to MsWord. This will paste as a MsWord table. This file is appended each time the analysis is executed. Each append event includes the file name and a time stamp. This file needs to be maintained by the user and deleted when it gets too big.
Browse	Used to upload a file.The file is placed in the temp directory. Once a file is uploaded, it will stay for all analysis until the engine is restarted.

Control Panel	Description
Submit	This submits your settings to the web server so that the engine can process your request.

10.0 Rule Attributes

Rules are defined by setting various text areas and processing options. The rules are grouped into services. So to view and modify the rules, you must enable a service. Once the service is visible you must enable the show simple and or complex rule options.

Attributes	Type	Level	Comment
Name	TextField	Simple	The name of the rule. You can use numbers and letters to try to order the results in the metrics table.
Color	TextField	Simple	This applies color to the text mined by this rule. The colors can be hex for RGB or names such as: red, green, blue, yellow, orange, purple, navy, etc.
Norm Metric	TextField	Simple	This is an external value once entered is reported in the metrics table next to the current run in the Metrics Report area.
Case Sensitive	CheckBox	Simple	The mining patterns are case insensitive unless this box is checked.
Access Object	TextField	Simple	This is the pattern used to access an object. It can be any regular expression recognized by PERL. It is shown in Analysis Results.
Previous Object	TextField	Complex	This is the pattern used to access a previous object. It can be any regular expression recognized by PERL. It works in conjunction with the Access Object parameter. Placing a parameter in this field increases the processing time. It is shown in Analysis Results.
Next Object	TextField	Complex	This is the pattern used to access a next object. It can be any regular expression recognized by PERL. It works in conjunction with the Access Object parameter. Placing a parameter in this field increases the processing time. It is shown in Analysis Results.
Reject Object	TextField	Simple	This is the pattern used to reject an object. It can be any regular expression recognized by PERL.
Comment	TextArea	Simple	This is a user comment reflecting what it means when this rule is triggered. It is shown either in Analysis Results and the Comments report areas.
Hide Accessed Objects	CheckBox	Complex	This hides objects in the Analysis Results so that other services can be supported that offer keyword mining results.
Show Child Objects	CheckBox	Complex	This will show a child object in the Analysis Results. The parent is defined by the access pattern. A new parent is identified for each access pattern within an entire service. When a parent is identified the child count is reset.
Count Child Objects	CheckBox	Complex	This will report the number of child objects each time a new parent is detected either within this rule or another rule in the

Attributes	Type	Level	Comment
			same service. This count is used to determine the Document Shape report.
Count Accessed Patterns	CheckBox	Complex	The access patterns are concatenated for a service and counted as the document is processed by each rule. The results are provided in the Accessed Patterns report.
Count Accessed Words	CheckBox	Complex	Every time an object is accessed, all its words are mined and counted. The results are provided in the Accessed Words report. This report is used to calculate the Reading Level report.
Count Rejected Words	CheckBox	Complex	Every time an object is rejected, all its words are mined and counted. The results are provided in the Accessed Words report. This report is used to calculate the reading level.

11.0 Report Areas

There are several report areas and they become populated based on the rule definitions. The report areas are Analysis Results, Accessed Words, Accessed Patterns, Metrics, Doc Shape, Reading Level, and Comments.

11.1 Analysis Results

This report is created when a rule requests a pattern to be accessed from an object. It is the main area that outputs the object text. Placing patterns in Access Object, Previous Object, Next Object mines the document and presents it to the user. The Comment is provided based the Show Object Comments and Hide Object Comments check boxes.

When Hide Accessed Objects is checked, the object text is not provided, but the mining still happens so that other rule processing can be applied, such as the Counting operations.

This area is also populated when looking for Duplicates or when Show Child Objects is checked. Looking for Duplicates

11.2 Accessed Words

This report is created when Count Accessed Words or Count Rejected Words is checked and there are mined objects based on the patterns entered. The objects are parsed and all words in all the accessed objects are identified. There is a Noise Filter than can be applied by checking the Filter Noise Words check box. The Noise filtered can be tuned by modifying the Noise Patterns. The default is unchecked so you should probably check this box and immediately re-run the report.

11.3 Accessed Patterns

This report is created when Count Accessed Patterns is checked. This report is like the Accessed Words report except the words that are identified are only those that were actually found in the mined objects, not all the words in the object. Also, there is a list of words that were not found in any of the objects.

11.4 Metrics

This report is created when a service and a rule is enabled to support some analysis view. The item is placed in the metrics table even if the rule is not triggered. The metrics are analyzed and a judgment is made to determine if the document is a specification or non specification document.

11.5 Doc Shape

This report is created when Count Child Objects is checked. This is actually a report within the Metrics report area.

11.6 Reading Level

This report is created when Count Accessed Words or Count Rejected Words is checked and there are mined objects based on the patterns entered. This is actually a report within the Metrics report area.

11.7 Comments

This report is created when a Service is selected. It provides the service description, rule comments, and summarizes the rule settings. The links in the Analysis Results area take the user to this report area so that a mined object can be fully understood and analyzed.