Paper SAS291-2014

Secret Experts Exposed: Using Text Analytics to Identify and Surface Subject Matter Experts in the Enterprise

Richard Crowell, Saratendu Sethi, Xu Yang, Chunqi Zuo, Fruzsina Veress, SAS Institute Inc.

ABSTRACT

All successful organizations seek ways of communicating the identity of subject matter experts to employees. This information exists as common knowledge when an organization is first starting out, but the common knowledge becomes fragmented as the organization grows. SAS® Text Analytics can be used on an organization's internal unstructured data to reunite these knowledge fragments. This paper demonstrates how to extract and surface this valuable information from within an organization. First, the organization's unstructured textual data are analyzed by SAS® Enterprise Content Categorization to develop a topic taxonomy that associates subject matter with subject matter experts in the organization. Then, SAS Text Analytics can be used successfully to build powerful semantic models that enhance an organization's unstructured data. This paper shows how to use those models to process and deliver real-time information to employees, increasing the value of internal company information.

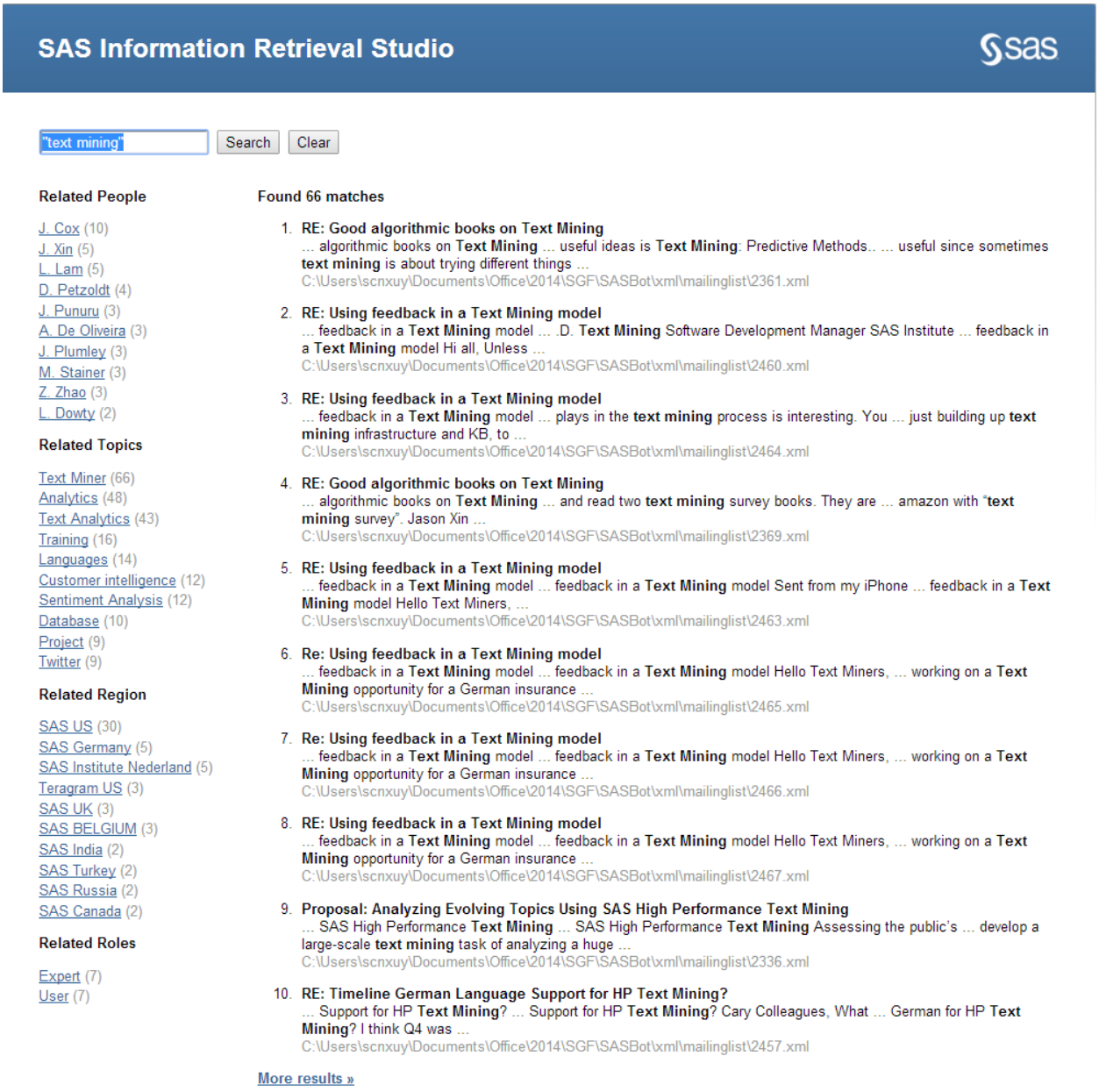
INTRODUCTION

Information retrieval is one of the key enablers to success in today's business environment. SAS Text Analytics offers powerful tools for consolidating, categorizing, and retrieving information across an enterprise, no matter how far-flung or unstructured that information might be. Traditionally, SAS Text Analytics is used for information that is stored in databases. But what happens when the information that you need is stored in someone's head instead of in a database and you're not sure who the right person to ask is? Determining the correct person might not be an issue in a small office, where you can simply ask around. But in a larger enterprise, tracking down subject matter experts is likely to be time-consuming, if not futile. And in a critical situation, that could mean the difference between failure and success.

The good news is that you can use SAS software to solve this problem! As this paper demonstrates, you can configure SAS® Information Retrieval Studio (which is the administrative interface for SAS® Crawler and SAS® Search and Indexing) to point you in the direction of employees who are most likely to be able to answer specific subject-matter questions. After you have performed this configuration process, searching for a topic produces a list of relevant documents along with a special faceted search window that lists the names of the employees who have the most experience dealing with that topic. Optionally, the search can also include a list of regional offices as a filter for where the expertise is located and another filter for differentiating between experts and users.

This paper shows you step-by-step how you can deploy SAS Text Analytics to enable this process, which consists of two principal tasks: using SAS Enterprise Content Categorization to create content categorization projects (as described in the section "Create the Content Categorization Projects") and then using SAS Information Retrieval Studio to configure categorization document processors to enable various search filters (as described in the section "Configure SAS Information Retrieval Studio"). In the first task, you create one project that lists all the employees and their email addresses and another project that lists all the important subject-matter areas. Optionally, you can also create projects to associate employees with their regional offices and to categorize documents as emanating from an expert, a user, or a poster. Then you use SAS Information Retrieval Studio, which provides a rich set of document processors to help you define a document processing workflow that will process the documents as they flow through the system.

Display 1 demonstrates a sample query in the SAS Information Retrieval Studio query web server interface. This list of matches was configured by crawling and indexing all emails sent to a specified mailing list within the last three months. When you search for a term such as "text mining," the query interface returns all relevant emails that mention the term. In addition, on the left, the search engine also returns the list of relevant people in the company who sent the most emails that mention that term, the list of topics to which the search term is related, the list of related regions where these employees are active, and a classification of whether the employee is an expert or simply a user. You can refine your query further by right-clicking terms in the left panel and selecting **Include**, **Exclude**, or **Restrict**. For example, you could include "SAS US" to see list of only those employees who are related to text mining and work in SAS offices in the United States.



Display 1. Search Results

TASK 1: CREATE THE CONTENT CATEGORIZATION PROJECTS

Before you can configure SAS Information Retrieval Studio, you need to create the content categorization projects that will be used to process the documents. In order to set up the expert-finder system presented in this paper, three projects were automatically derived from enterprise databases. These databases were exported into Microsoft Excel and then imported into SAS Enterprise Content Categorization. The fourth project, which is relatively small, was created manually in SAS Enterprise Content Categorization. The following list describes all four projects.

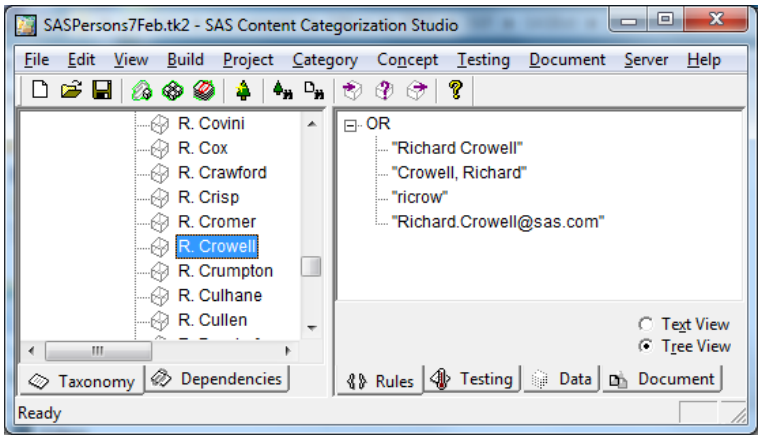
- Employees categorization project (mandatory): The data for this project were derived in an automated fashion from an online database of employee names, which includes the email address and user ID of each employee. The database interface allowed for automatic export to Excel. The data were then cleaned by eliminating any unnecessary columns (such as phone numbers) and deleting any rows in which the email

field was left blank (for example, email addresses of mailing lists). Table 1 provides a sample of the final data that were exported from the company database.

First Name	Last Name	User ID	E-mail Address
Richard	Crowell	ricrow	Richard.Crowell@sas.com
Saratendu	Sethi	ssethi	Saratendu.Sethi@sas.com
Xu	Yang	scnxuy	Xu.Yang@sas.com

Table 1: Employee Data in Microsoft Excel

A custom script was written to export the data from Excel into an XML file that could be imported into SAS Enterprise Content Categorization. The import was accomplished by creating a new project in SAS Enterprise Content Categorization Studio, adding a language, and then right-clicking on the language to select **Import Categorizer from XML**. The structure of the content categorization taxonomy is very simple: categories are employee names, and the rule definitions are a combination of employee names in various forms such as *firstname lastname*, *lastname, firstname*, *userid*, and *email address*. Display 2 demonstrates a sample rule that was derived for one of the authors of this paper, Richard Crowell.



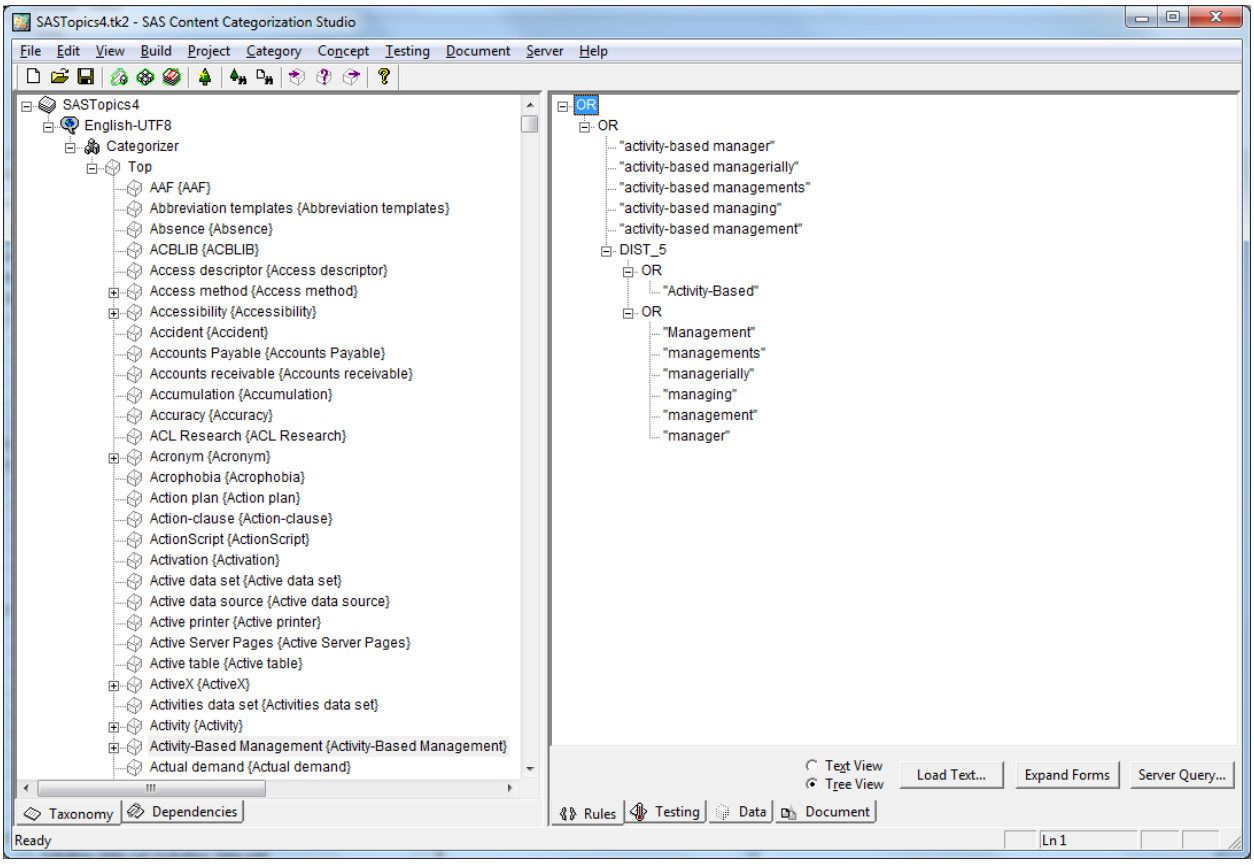
Display 2. Employee Content Categorization Project

At the time of writing this paper, the employees content categorization project consists of 13,769 categories, which represent SAS employees from around the world.

- Topics categorization project (mandatory): The other project that must be created is a list of subject-matter categories. Most enterprises will find that creating this project isn't as straightforward as creating the employees content categorization project. For example, in this paper, the list of categories that represent various subject-matter areas was created by combining the index of the company's internal Semantic MediaWiki with the corporate terminology glossary. The resulting list was cleaned to remove irrelevant topics and general keywords, and then was organized into a hierarchical structure by grouping similar strings. As with the employees content categorization project, these data were organized in an Excel spreadsheet and then imported into SAS Enterprise Content Categorization by means of a custom script.

  The process of collecting terms and identifying relevant topics is likely to lead to multiple discussions within your company. The effort required to create this categorization project can yield benefits that extend beyond the immediate purpose of the expert-finder system, because shared terminology is essential to the efficient exchange of information within an enterprise. There are different ways to achieve the goal of shared terminology: employees might agree to use a standardized set of terms, or tools for normalization might be built into applications. Whatever method you use, it is critical for an enterprise to understand and control its terminology so that its employees can understand each other. The degree of difficulty that you encounter when creating this topics content categorization project is a good indicator of how well your corporate terminology is organized.

The original flat list of topics (without any added hierarchy) was organized into a single column in an Excel worksheet. The custom script used an automated hierarchical structure algorithm to import the Excel worksheet into SAS Enterprise Content Categorization; part of the resulting topics content categorization project is shown in Display 3.



Display 3: Topics Content Categorization Project

- Regions categorization project (optional): In order to reveal the regional offices where expertise is concentrated, a third project was created in SAS Enterprise Content Categorization. This project consists of a list of categories that represent the regional offices; the rule definition for each category consists of the employee name, email address, and user ID of each employee who works in that office. So the category *SAS US* has the following definition (truncated in this example).

```
(OR,(OR,"Richard Crowell","Richard.Crowell@sas.com"),(OR,"Xu
Yang","Xu.Yang@sas.com"),(OR,"Saratendu Sethi","Saratendu.Sethi@sas.com"),(OR,…))
```

Because this project has a relatively small number of unique categories (one for each SAS office location), it was simple and quick to compile. The data were organized in an Excel worksheet before being imported into SAS Enterprise Content Categorization Studio. Table 2 shows two rows of the Excel spreadsheet.

Region	First Name	Last Name	User ID	E-mail
SAS US	Saratendu	Sethi	ssethi	Saratendu.Sethi@sas.com
SAS US	Xu	Yang	xyang	Xu.Yang@sas.com

Table 2: Region Data in Excel

- Role categorization project (optional):

  The final categorization project was designed to allow filtering according to the role of the person associated with a topic according to the following criteria:

  - Expert role: Is the email author answering a question?
  - User role: Is the email author asking a question?
  - Poster role: Is the email author simply passing along third-party information?

  The role content categorization project contains only those three categories: expert, user, and poster. The definitions for those categories contain strings of words that are likely to characterize one of those roles. Following is the (truncated) definition of the Expert category:

```
(OR,"did you try","are you sure it is","assist you","be sure to check","detailed steps
below","does that help","encourage you to","good luck","hope this helps",…)
```

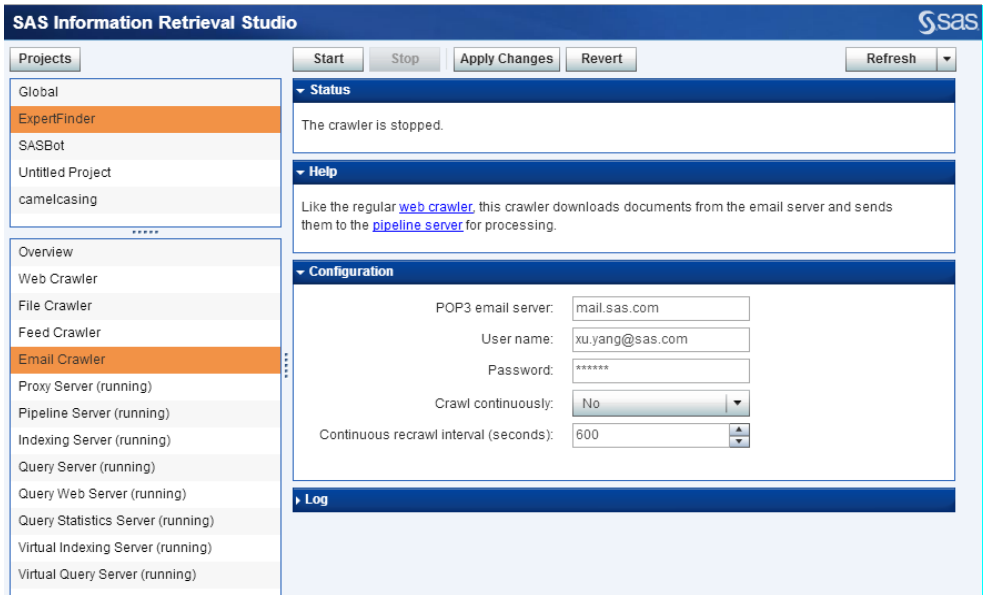
  After you have created the content categorization projects, you can upload them to SAS Enterprise Content Categorization Server from SAS Information Retrieval Studio by selecting **Upload Categorizer** from the **Build** menu.

TASK 2: CONFIGURE SAS INFORMATION RETRIEVAL STUDIO

SAS Information Retrieval Studio is the web-based administrative interface for SAS® Crawler and SAS® Search and Indexing. The following steps show how to set up the document processing workflow to create a functional expert-finder system. The workflow starts by crawling documents from the email exchange server, then uses the uploaded content categorization projects to process the documents, and finally indexes the annotated corpus of relevant documents to make it searchable.

1. Create a new project in SAS Information Retrieval Studio. Give it a meaningful name (such as *ExpertFinder*). Check the Projects window to ensure that only the Proxy Server is running. The other services are unnecessary at this point.
2. Select your documents and set up the email crawler. The success of the expert-finder system depends primarily on your choices of the documents and categorization topics. It is recommended that you use multiple sources and include both structured and unstructured data. The main criterion for selecting any source should be the way in which it is used. (Forums, listservs, and corporate wikis that are specifically designed for answering questions and sharing information are obvious choices.) Locate the channels where such exchanges of information are likely to occur within the enterprise, and then find a way to retrieve those documents and store them in a location that is easier to crawl.

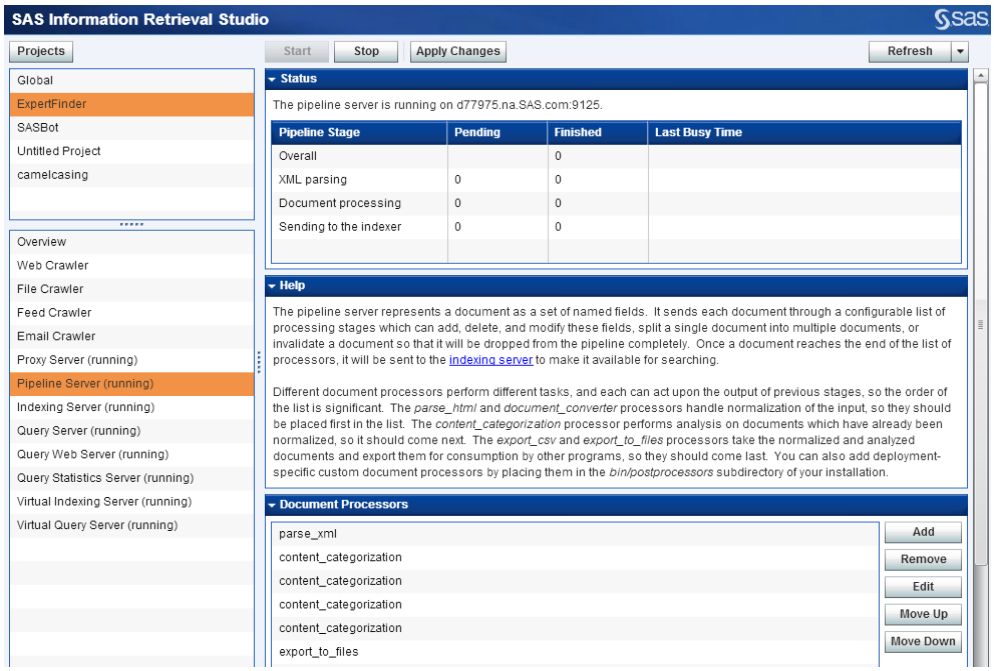
   SAS Information Retrieval Studio provides multiple options for crawling data. You can use the file crawler to crawl file systems and directories, the web crawler to crawl web pages, various social media crawlers to crawl data from social media sites. You can also write your own crawlers and plug-ins. For this paper, a custom email crawler was written, which collected emails from various mailing lists and email accounts from the company's email exchange server. To set up the email crawler in SAS Information Retrieval Studio, click **Email Crawler** in the Projects window and open the **Configuration** pane. Fill in the details of your corporate email exchange server and the username and password credentials for the email account that will be used for crawling. To let the crawler continuously collect emails, select **Yes** in the **Crawl continuously** list and specify the interval in the **Continuous recrawl interval** field. Display 4 shows the example configuration. At this point, verify that the **Status** field indicates that the crawler is stopped. You will turn it on later.



Display 4: Email Crawler Setup

3. Insert document processors in the pipeline server. The following substeps describe the process of configuring various document processors to extract content from emails and categorize them by using the content categorization projects that were created in previous sections:

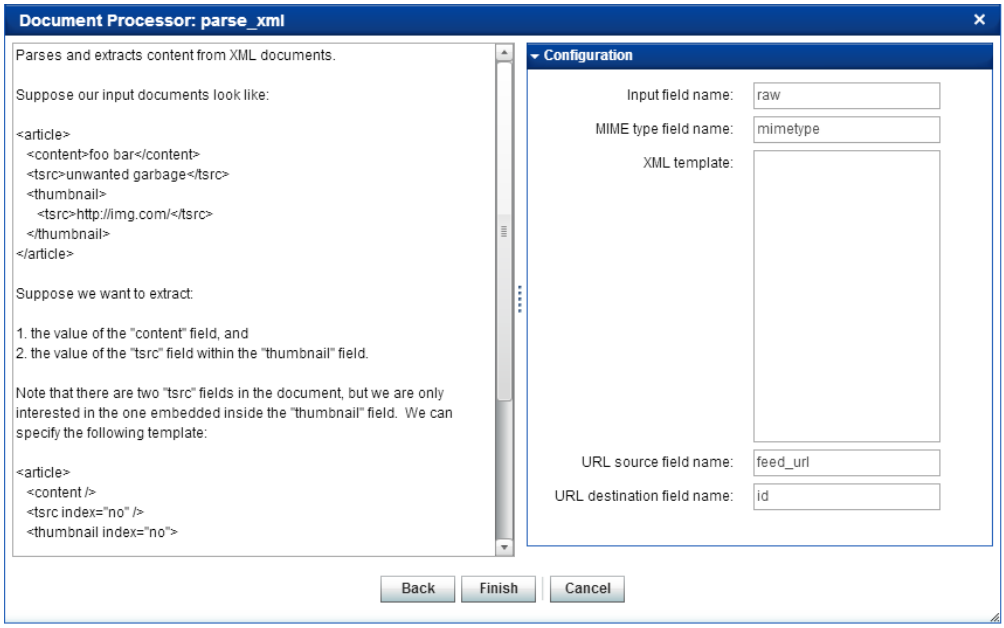
    a. In the left pane, select **Pipeline Server**. You can add or edit various document processors in the **Document Processors** pane (see Display 5). Click **Add** to select the type of processor that you need.



Display 5: Document Processors Setup in the Pipeline Server

b. When the Email Crawler will be started later, it will extract all emails in XML format and feed them into the Pipeline Server which will process them through the document processors. The first document processor extracts the XML fields that have useful content and can be processed.

c. Select **parse_xml** from the list of available document processors, and click **Add**. A Document Processor window appears, as shown in Display 6, which shows the settings that were used for this paper. These settings will cause the email crawler (when you start it in step 6) to extract all emails in XML format and feed them into the pipeline server, which will process them through the document processors.



Display 6: parse_xml Document Processor Settings in the Pipeline Server

d. The next set of document processors adds the content categorization projects that will be used to classify each email according to the employees, topics, regions, and role projects. In the left pane, Select **Pipeline Server**, and perform the following steps, as indicated in Display 7.

1) Click **Add** in the **Document Processors** pane.

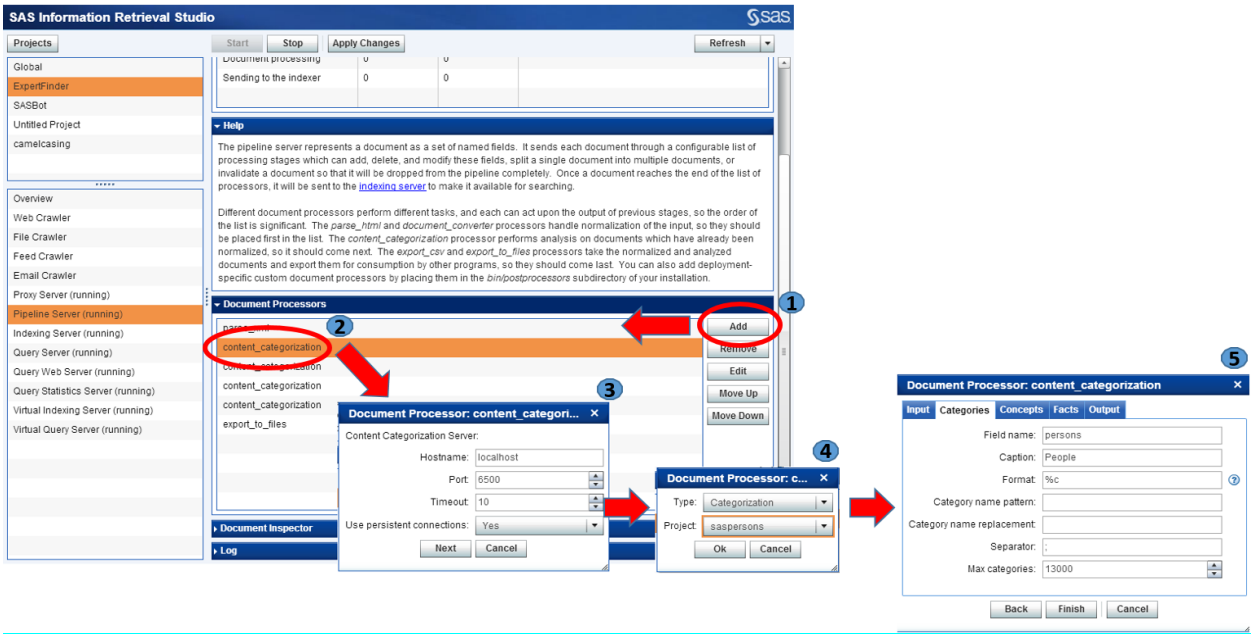
2) Select **content_categorization** from the list of available document processors.

3) In the Document Processor: content categorization window, fill in the details about the machine where SAS Enterprise Content Categorization Server is running. If that server is running on the local machine and using default settings, you can accept all the default parameters. Click **Next**.

4) The **Project** field in next window contains a list of projects that have been uploaded to SAS Enterprise Content Categorization Server. Select the project that you want to process. (This example selects *saspersons*, which is the name of the employees project.) Click **OK**.

5) The next window enables you to add the result of this categorization as document metadata. The example settings in Display 7 demonstrate that if the document is successfully matched to any employee, then the list of matches is added as *persons* metadata into the document. For this example, the default settings of other fields are used.

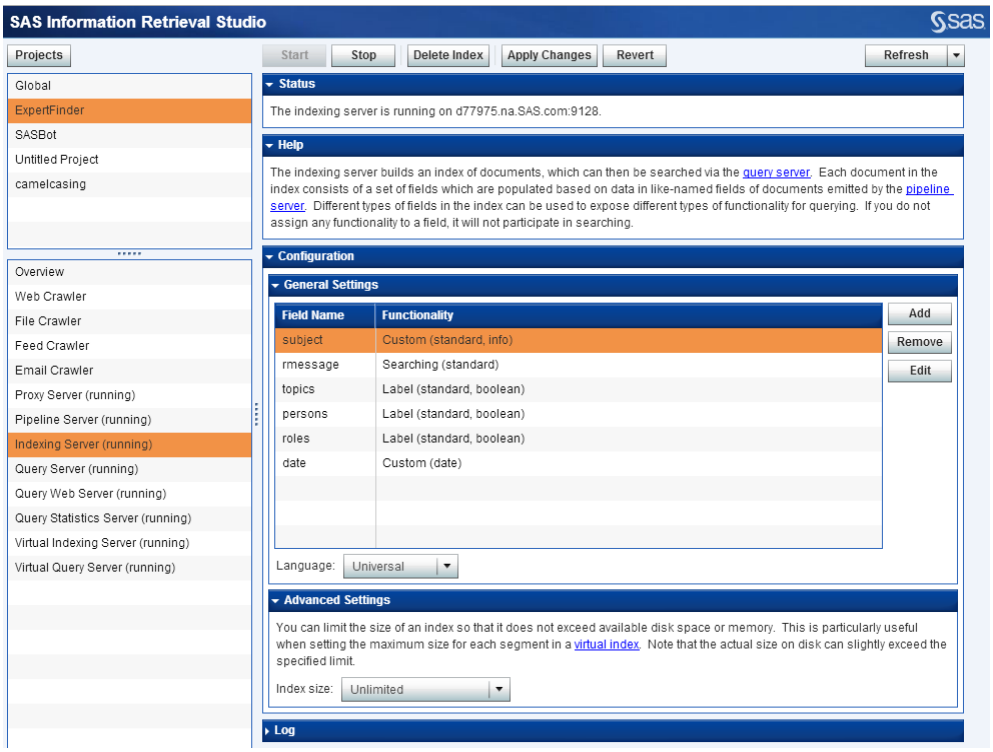
Repeat the steps 1) through 5) to add the remaining content categorization projects (topics, regions, and roles) in a similar fashion.



Display 7: Content Categorization Document Processor Settings in the Pipeline Server

e. During the implementation of this example, an optional document processor “export_to_files” was added. This processor stores all the processed document contents along with the categorization results as XML files into a specified document folder. Exporting the files enables you to rebuild the project later if you want to migrate the project from one physical server to another.

After performing the preceding substeps, click **Apply Changes** at the top of the window to apply all the configurations to the SAS Information Retrieval Studio pipeline server and restart all the necessary modules.

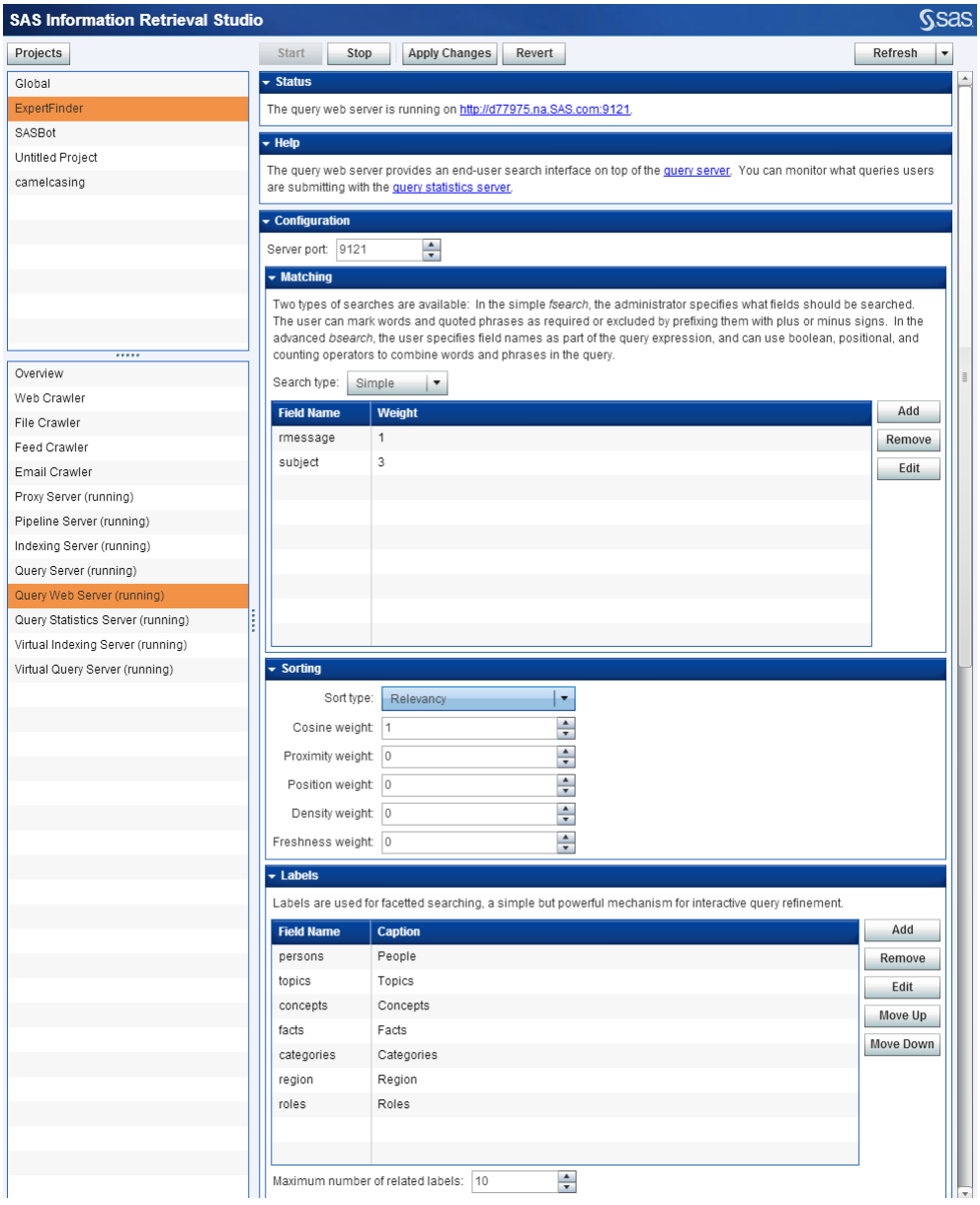


Display 8: Configuration of Indexing Server

4. Configure the search interface. This step describes how to configure the indexing server, which is used to build the search index that is accessed through the query web server in SAS Information Retrieval Studio.

   In the left pane, select **Indexing Server**. The **General Settings** pane specifies the language of the documents, the XML fields where the content is to be searched, and the XML fields that will act as the "filter fields." Display 8 shows that this project has configured the search index to use the **rmessage** field for searching and has configured **persons**, **topics**, **roles,** and **regions** as the filter fields.

   In the left pane, select **Query Web Server** so that you can configure it by adjusting the settings for search. (See Display 9.) In the **Matching** pane, leave the default value (*Simple*) in the Search type field. In the **Sorting** pane, make sure that the **Sort type** is *Relevancy*. The **Labels** pane enables you to use the results of content categorization projects for search filters. The **Field Name** represent the XML fields or metadata names into which results of categorization were stored, and the values in the **Caption** column show how those fields will be exposed in search query results.



Display 9: Configuration of Query Web Server

5. Run the crawler. In order for the Indexing Server to start building the index, you must start each of the following servers in turn by selecting it in the left pane and clicking **Start**: **Proxy Server**, **Pipeline Server**, **Indexing Server**, **Query Server**, and **Query Web Server**.
6. Finally, select **Email Crawler** and click **Start**. The email crawler uses the user credentials to begin crawling emails from the exchange server and starts sending documents to the pipeline server, which parses them, processes them through content categorization projects, and passes them on to the indexing server. The indexing server then builds an index to be used by the query web server to provide a faceted search. Do not shut down any of the servers or close the SAS Information Retrieval Studio interface until all the documents have been completely processed. You can monitor the progress by keeping the pipeline server window open.

When the pipeline server finishes, you can open the query web server interface by clicking the link that is displayed in the **Status** field. Now you can start querying the system for experts by typing in query terms as shown in Display 1.

CONCLUSION

This paper demonstrates the application of SAS Text Analytics (specifically the combination of SAS Information Retrieval Studio and SAS Enterprise Content Categorization) to create a powerful tool for locating and disseminating expert information within your enterprise. The advantage of this method is that it can easily and quickly point out the most helpful and active resources. Future work on this project includes creating a conversation agent that will answer questions that are posted to the company message board or automatically respond to posters by sending emails that contain information about helpful employees in specific subject-matter areas within specific regions. There are many other creative ways to use this expert-finder system within a variety of corporate environments. There are also many possibilities to make this system more sophisticated by adding analysis of response patterns in each conversation, studying network connections in the enterprise community sites, and so on. The method described in this paper is an initial attempt to provide a simple and flexible method of integrating knowledge that is indicated by, but not necessarily contained within, existing documentation or emails.

ACKNOWLEDGMENTS

The authors would like to acknowledge Chris Heller for his initial ideas and many other SAS employees whose continuous mentions of the utility of building an enterprise-wide expert-finder system motivated the process described in this paper. The authors also thank Anne Baxter for her editorial assistance.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact:

Saratendu Sethi
SAS Institute, Inc.
600 SAS Campus Drive
Cary, NC, 27513
+1 (919) 531-0597
Saratendu.Sethi@sas.com

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.