Analyzing the Web from Start to Finish

Knowledge Extraction from a Web Forum using KNIME

Bernd Wiswedel Bernd.Wiswedel@knime.com
Tobias Kötter Tobias.Koetter@uni-konstanz.de
Rosaria Silipo Rosaria.Silipo@knime.com

Table of Contents

Summary

In recent years, the web has become the source of all possible information. Web contents are used more and more frequently to find out about customers' and real-world people's preferences. This whitepaper implements all parts of the process of extracting information from the web using KNIME.

In order to produce a practical example and at the same time to learn more about the KNIME community users, this analysis is focused on data from the KNIME Forum. The analysis is divided in 4 parts.

The first workflow is a web crawler. It is dedicated to web content extraction and data reorganization which renders it suitable for the following analysis. It has often been said that one of the best features of KNIME is the community behind it. Indeed, the download of the forum content is outsourced to one of the community nodes. The following steps for data extraction and reorganization, on the opposite, mainly rely on the KNIME XML node collection.

A few basic statistical measures are calculated to get insights about the forum performance as an indirect measure of the KNIME community performance. Here users can be posters and commenters at the same time. While the total number of users and of posts over time gives a measure of the community growth, the average number of comments for each post can be considered a measure of forum answer efficiency.

The topics discussed in the KNIME Forum represent another big source of information: they clearly describe the evolution of the users' interests and wishes over time. A full workflow has been implemented to classify topics and detect topic shifts in time, using text mining techniques to describe the post contents and predictive analytics to classify them. The results show that popular topics have always been "data manipulation" and "data mining", i.e. tasks where KNIME quality is well known. They also show that the interest for the "flow control" category has been growing strongly and steadily over time, re-enforcing the choice to keep improving these category nodes even further.

Finally, a fourth workflow examines how the forum users interact with each other in different discussion groups. Depending on the discussed topics, experts emerge quickly here from the user network graph.

Those are four very popular areas in the analysis of web contents which are easily exportable to other business contexts: web crawling, web analytics, topic detection, and description of user interaction. All workflows are available on the KNIME public server and the KNIME software can be downloaded from www.knime.com.

Web Analytics and the Desire for Extra-Knowledge

The analysis of web sites and especially of forums and social networks has become necessary for many companies, in order to know what their customers wish, need, and use. However, the extraction of knowledge from web sources is a complex task and involves many different steps in a wider variety of disciplines. Often, employees as well as tools are specialized in only one of those disciplines and therefore can carry out only one of the tasks required for the knowledge extraction from the web, be this web crawling, database storage, th text analytics for topic detection, sentiment analysis, user network representation, statistics, web analytics, or other such tasks.

The ambitious goal of this whitepaper is to implement and describe a full approach to the analysis of web forums and social media from the beginning to the end, through:

- data collection via a web crawling algorithm and XML parsing functions
- simple statistics to see the evolution of the forum in time
- a topic detection application using text mining techniques
- a representation of the user network via network analytics
- the productization of all these steps on a KNIME Server

KNIME (www.knime.com) has been selected as the tool for the implementation of this project. The choice fell on KNIME, not only because KNIME has a graphical interface, is easy to use, and is reliable, but also because it contains all the necessary tools for all of the steps required by this project.

Indeed, many nodes are available in the KNIME Core for basic statistics. A full suite can be installed from the KNIME extensions for text mining as well as for network analytics, containing all of the nodes you need to perform sentiment analysis, topic detection, and network graph representations. Finally, a web crawler node has been made available via a community extension provided by the Palladian project (http://www.palladian.ws/).

All these features are powerful already by themselves, but combined with all the other data processing, data mining, and reporting functionalities, they turn KNIME into a very powerful tool for any analytics task, and in particular for extracting data and knowledge from the web communities.

For this whitepaper we focused on the KNIME Forum (http://tech.knime.org/forum) where community users exchange questions and answers about technical KNIME topics. The goal was to gain insights about the evolution in time of the KNIME software and about the user wishes and needs for the future.

The KNIME open source platform can be downloaded from the KNIME site at www.knime.com. All workflows reported in this whitepaper were developed using KNIME 2.7. However, an updated version for KNIME 2.8, including the forum data and the metanode template, is available from the KNIME public server at knime://publicserver/050_Applications/050007_ForumAnalysis.

The KNIME Forum

The goal of this whitepaper is to show how to extract data and therefore knowledge from web sites. As a data mining company, we decided it would be best if we actually learned something from our own data. So, we centered the project on the KNIME Forum data.

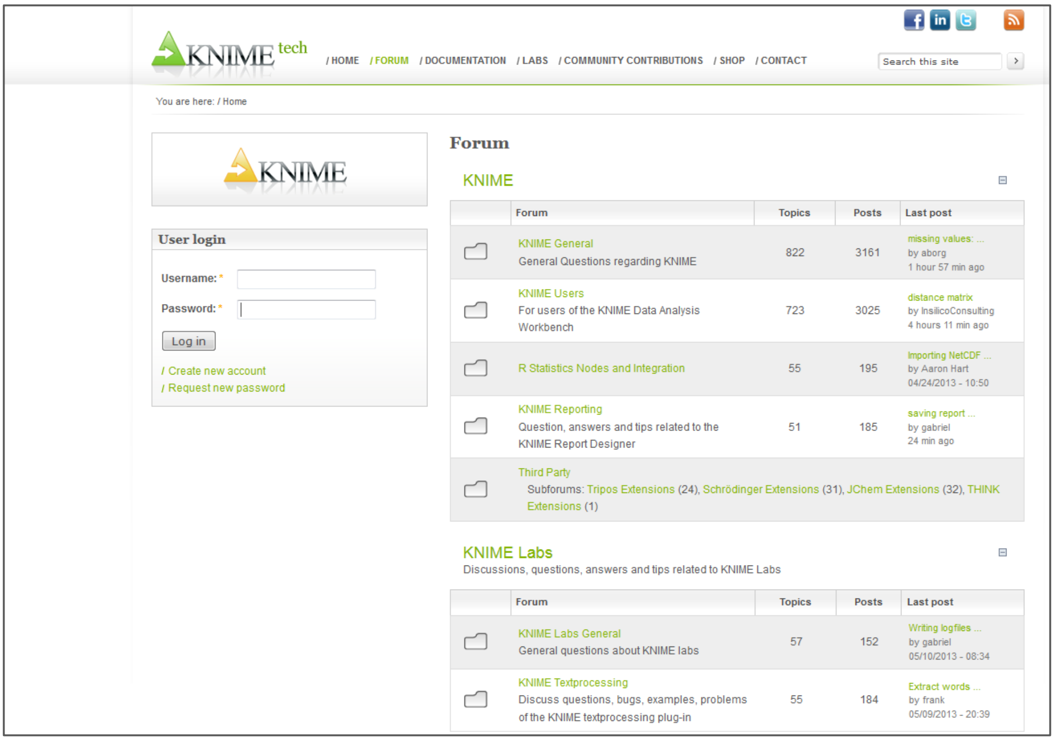
The KNIME Forum (http://tech.knime.org/forum) is the main community place for KNIME users. Here KNIME users post questions to get help from other KNIME users to solve some particular problem. Usually and quickly, at least one KNIME user knows the answer and is willing to help with the problem at hand.

You need to officially register to take part into the KNIME Forum. After logging in, you can then place your post or comment on other users' posts. The Forum is divided in 5 topic-dependent parts, with posts about:

- KNIME Core and Extensions
- KNIME Labs
- Development of new KNIME nodes
- KNIME Server
- Community Contributions

Each one of these sub-forums consists of many categories, like KNIME General, Text Processing, Third Party nodes, and so on (see Fig. 1).

Figure 1. The KNIME Forum



Forum

KNIME

Forum	Topics	Posts	Last post
KNIME General General Questions regarding KNIME	822	3161	missing values: ... by aborg 1 hour 57 min ago
KNIME Users For users of the KNIME Data Analysis Workbench	723	3025	distance matrix by InsilicoConsulting 4 hours 11 min ago
R Statistics Nodes and Integration	55	195	Importing NetCDF ... by Aaron Hart 04/24/2013 - 10:50
KNIME Reporting Question, answers and tips related to the KNIME Report Designer	51	185	saving report ... by gabriel 24 min ago
Third Party Subforums: Tripos Extensions (24), Schrödinger Extensions (31), JChem Extensions (32), THINK Extensions (1)			

KNIME Labs
Discussions, questions, answers and tips related to KNIME Labs

Forum	Topics	Posts	Last post
KNIME Labs General General questions about KNIME labs	57	152	Writing logfiles ... by gabriel 05/10/2013 - 08:34
KNIME Textprocessing Discuss questions, bugs, examples, problems of the KNIME textprocessing plug-in	55	184	Extract words ... by frank 05/09/2013 - 20:39

The Data

In order to retrieve all posts and comments from the KNIME Forum we could have accessed the underlying database storing the forum content. However, this would have been a very specialized approach, which is not universally applicable to other systems and hence not of interest in the scope of this article. To be able to extract data from any web site, with no access to the underlying database, we need to crawl and read all (linked) pages.

There are a number of open source as well as commercial web crawling tools available. One option could have been to start and run one of those web crawling tools via the "External Tool" node from inside a KNIME workflow. However, as it often happens with KNIME, surfing a bit around the KNIME community, we discovered that a web reader node had already been developed and made available to KNIME users for free by Palladian (http://www.palladian.ws/). The node reads a given online HTML web site and extracts all information in further processable XML format.

The KNIME Forum actually hosts five discussion groups: KNIME; KNIME Labs, Community Contributions, Enterprise Tools, and KNIME Development.

The most visited place is of course KNIME, since this is where all general posts about KNIME usage are.

The KNIME Labs forum page contains all questions about nodes in the KNIME Labs category. Just as a reminder, KNIME Labs contains a number of new node extensions, which are not yet part of the standard KNIME distribution but are still made available as a sneak preview for the KNIME users.

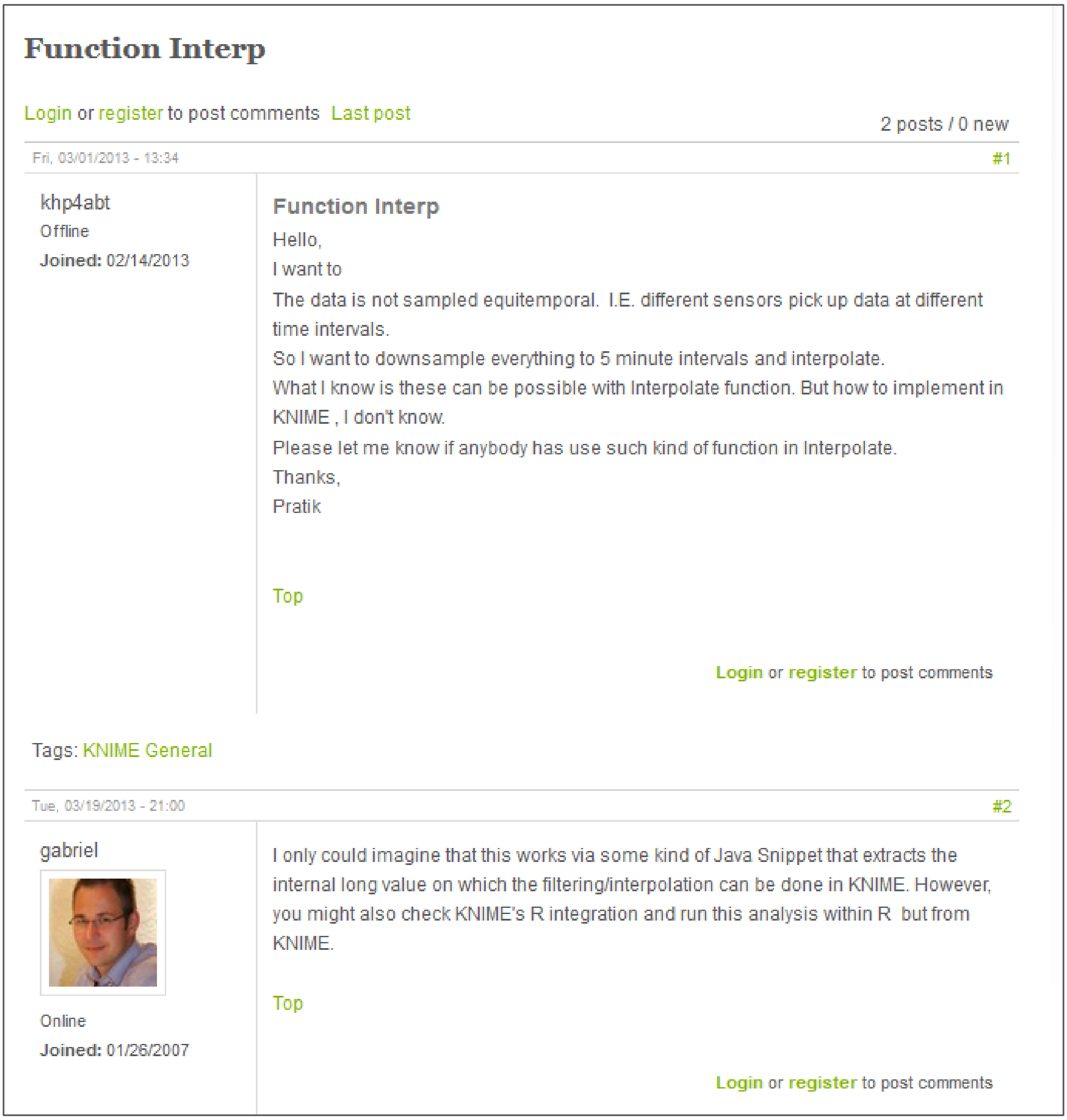
Community Contributions hosts all discussions about community developed nodes.

Recently a full discussion group has been opened for the KNIME Enterprise Tools. Indeed, while the KNIME Enterprise Tools become more widely adopted, a category for discussions on this topic has become necessary.

The last discussion group, KNIME Development, hosts all posts about standards and guidelines in node development and the KNIME tools available to make the development of custom nodes easier. Each discussion group separates its posts in a number of categories. The KNIME discussion group, for example, contains 5 categories: KNIME Users, KNIME General, R Statistics nodes and Integration, KNIME Reporting, and Third Party. In total, the KNIME Forum hosts around 20 categories.

An individual forum discussion thread consists of an initial post and a number of comments (Fig. 2); a thread can span over several pages. The biggest challenge, after pulling down all the web pages, is to resynchronize each post with all its comments.

Figure 2. An example of a forum thread



Function Interp

Login or register to post comments Last post

2 posts / 0 new

Fri, 03/01/2013 - 13:34 #1

khp4abt
Offline
Joined: 02/14/2013

Function Interp

Hello,

I want to

The data is not sampled equitemporal. I.E. different sensors pick up data at different time intervals.

So I want to downsample everything to 5 minute intervals and interpolate.

What I know is these can be possible with Interpolate function. But how to implement in KNIME , I don't know.

Please let me know if anybody has use such kind of function in Interpolate.

Thanks,

Pratik

Top

Login or register to post comments

Tags: KNIME General

Tue, 03/19/2013 - 21:00 #2

gabriel
Online
Joined: 01/26/2007

I only could imagine that this works via some kind of Java Snippet that extracts the internal long value on which the filtering/interpolation can be done in KNIME. However, you might also check KNIME's R integration and run this analysis within R but from KNIME.

Top

Login or register to post comments

The Analysis

The analysis of the content of the KNIME Forum has been split in 4 parts. The corresponding workflows can be downloaded from the KNIME public server named EXAMPLES and available in the KNIME Explorer panel in the KNIME Workbench.

1. The web crawler workflow to collect all posts and comments from the forum.
2. A workflow to calculate some simple statistics about the forum contents and the user behavior over the past 6 years.
3. An attempt to classify the posts in terms of discussed topics and to observe the topic shift over time.
4. The definition of a user network for a given category of the forum and for a given topic. That is basically a description of who is talking to whom about a certain topic.

In this whitepaper, we explore all of these workflows. We start from the workflow to crawl the web, we continue with the workflow for some statistics on the forum data, then we implement a topic classification application, and we end with a representation of the social structure in each forum category.

The "WebCrawler" Workflow

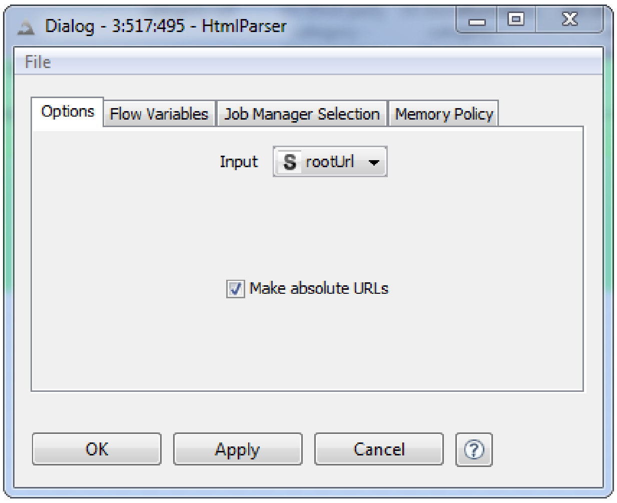
We need now a web crawler application to download all posts and comments from the KNIME Forum. Palladian (http://www.palladian.ws/) has developed an HTML parser node that connects and downloads web pages. The "WebCrawler" workflow uses this node to download contents from a website followed by a number of XML Parser nodes to extract meaningful pieces of text from the downloaded contents.

The HtmlParser Node from Palladian

The HtmlParser node is based on the Java HTML5 Parser, named Validator.nu (http://about.validator.nu/). The parser is designed to work as a drop-in replacement for the XML parser in applications that already support XHTML 1.x content and use SAX, DOM or XOM to interface with the parser. The node execution is resource intensive, especially on the targeted web site, and quite time consuming. So be careful when running it, not to stall the target web site!

The configuration window of the HtmlParser node is quite simple. The node just needs to know the list of URLs of the target websites, in the form of a data column (Fig. 3). When enabled, all relative URLs are converted to absolute ones. This simplifies/permits further processing steps with the URLs obtained from the input data column.

Figure 3. Configuration Window of the HtmlParser node



Dialog - 3:517:495 - HtmlParser

Input: S rootUrl

☑ Make absolute URLs

The web address for the KNIME Forum (http://tech.knime.org/forum) is provided as the only content of a Table Creator node. The HtmlParser node then, during execution, connects to the web site, downloads the raw HTML content, and presents it at its output port in form of an XML column.

Notice that the HtmlParser node belongs to the community contributions and therefore it is not part of the core KNIME Desktop features. The node has to be installed separately as a KNIME Extension from the Palladian package (http://tech.knime.org/community/palladian) in the Community Contributions.

The XML Parser Node

After execution, the HtmlParser node presents a huge XML cell at the output port, containing all threads from the forum. In order to isolate the details of each forum thread, some XML parsing has to be applied.

KNIME offers a full category, named XML, dedicated to XML operations: reading, writing, converting, combining, and of course parsing. The node that is needed here is the XPath node.

This node performs an XPath query on a column of XML cells and can produce another XML cell (a Node), a collection of XML cells (a Node-Set), or just a single value (number, string, Boolean, etc…) depending on the result of XPath query. The node supports XPath 1.0.

For example, let's consider the following XML document:

```
<inventory>
     <book year="1999">
         <title>Harry Potter and the Sorcerer's Stone</title>
         <author>J. K. Rowling</author>
     </book>
     <book year="2000">
         <title>Harry Potter and the Chamber of Secrets</title>
         <author>J. K. Rowling</author>
     </book>
</inventory>
```

Using "dns" as the prefix of root namespace:

- `//book` returns a collection of XML cells (Node-Set) , which, after ungrouping, shows as:

```
<book year="1999">
   <title>Harry Potter and the Sorcerer's Stone</title>
   <author>J. K. Rowling</author>
</book>
```

```
<book year="2000">
   <title>Harry Potter and the Chamber of Secrets</title>
   <author>J. K. Rowling</author>
</book>
```

- `//book[@year=1999]` returns again a collection of XML cells (Node-Set) similar to the previous one, but containing only the items with `year = 1999`.
- `//book/title` returns a collection of XML cells (Node-Set) , which, after ungrouping, shows as:

```
   <title>
       Harry Potter and the Sorcerer's Stone
   </title>
```

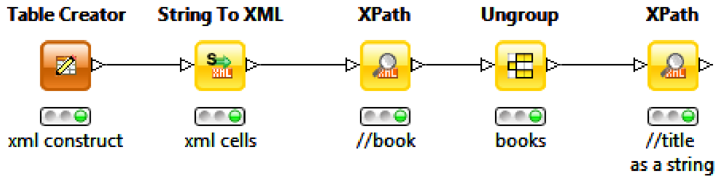
```
   <title>
       Harry Potter and the Chamber of Secrets
   </title>
```

- //book/@year Notice that collection outputs are only possible for the XML cells. The XPath node does not produce string, integer, double, or any other type of collections. So, using an XPath node, using //book/@year as XPath query, produces a collection of XML cells, which after ungrouping is visible as:

```
    <fragment year=1999>
    </fragment>
    <fragment year=2000>
    </fragment>
```

- To isolate the titles as strings, you need to extract the <book> items first as a collection of cells with //book XPath query, then ungroup the collection, and finally extract the <title> items as strings with XPath query //title in an XPath node again (Fig. 4). The return type of the last XPath node is “String (String cell type)”.

Figure 4. Extract titles from xml cell as //book/title



- To isolate the years of publication, you need to extract the <book> items first as a collection of cells with //book XPath query, then ungroup the collection, and finally extract the @year attributes as integer with XPath query //@year in an XPath node again. The return type of the last XPath node is “Number (Integer cell type)”.
- You can find more complex examples of XML parsing at w3cschools.com.

Before continuing with parsing the html document collection produced by the HtmlParser node, let’s have a look in more detail at the XPath node’s configuration window. For example, let’s use the configuration window of the XPath node that extracts all book titles as strings (Fig. 5).

First of all, the node needs the name of the XML column to work on, “books” in this case, which we can remove after the XML parsing has been executed. Then, it needs the name for the newly produced column: we used “titles”.

The most important part is, of course, the XPath query, in this case //title. A relative spacious textbox is reserved to host the XPath query.

Finally, the definition of the type of results coming out of the XML parsing is needed. A few options are possible:

- Node-Set (collection of XML cells)
- Node (XMl cell)
- Number (Double or Integer)
- Boolean
- String

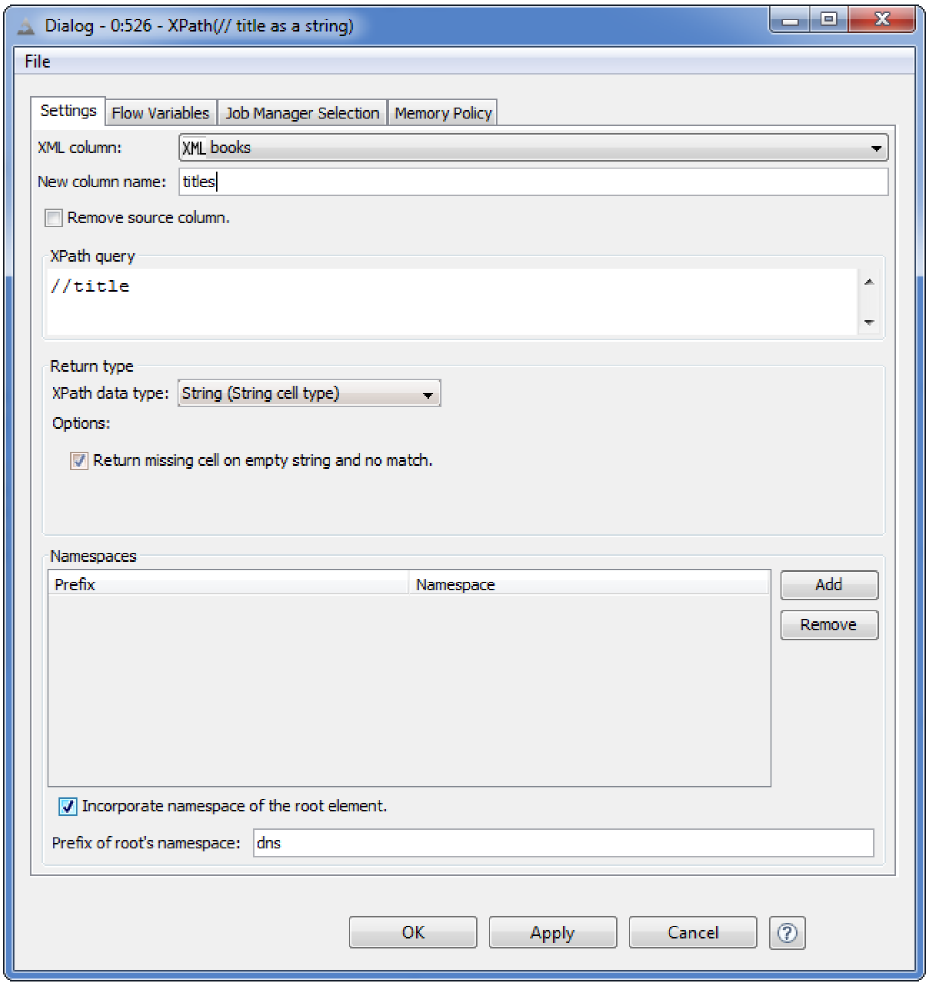
An error in selecting the wrong output type can be recovered using missing cells, by enabling the option “return missing cell on empty string or no match”. If the return type is an XML type (collection or cell), enabling the option "return missing cells on no match" allows error recovery via the <fragment> item.

If your XML document contains namespaces, you can define them in the XPath node's configuration window.

The root namespace can be incorporated by default in every XPath query, by enabling the corresponding option at the bottom of the configuration window. The default prefix for the root namespace can be choosen arbitrarily and needs to be used within the query. In the example we used "dns".

An example workflow for XML parsing can be found on the KNIME public server under 013_XMLProcessing/013004_XMLProcessing_HarryPotter.

Figure 5. XPath node configuration window



Dialog - 0:526 - XPath(// title as a string)

XML column: XML books

New column name: titles

□ Remove source column.

XPath query

```
//title
```

Return type

XPath data type: String (String cell type)

Options:

☑ Return missing cell on empty string and no match.

Namespaces

Prefix	Namespace

☑ Incorporate namespace of the root element.

Prefix of root's namespace: dns

To extract information from the html document generated by crawling the KNIME Forum, we start from the <forum-details> item, as root, by splitting it in two sub-items: <forum-name> and <forum-subforums>. After ungrouping these two item collections, the <href> items are extracted to get the full list of threads on each page as string documents. This sub-workflow has been encapsulated in the meta-node named "TOC of Forum".

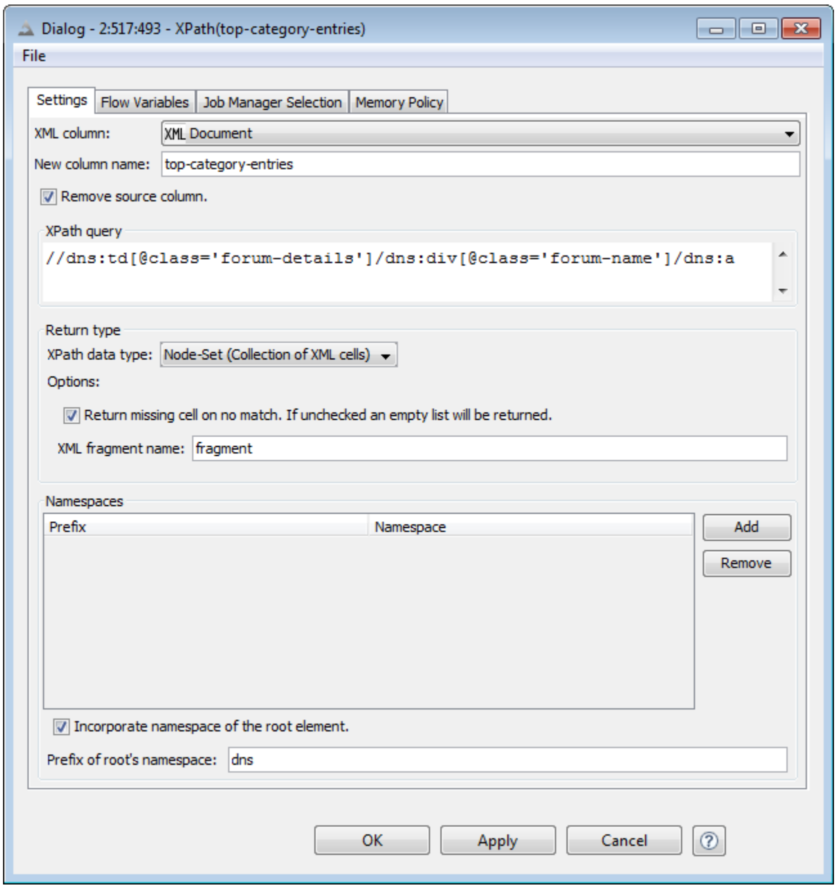
The upcoming sequence of XPath nodes, contained in the "Category Page Count" meta-node rescues all pieces of the same threads spread across different forum pages.

The final XPath query used in our workflow to extract the top category entries from the forum raw data is the following:

```
//dns:td[@class='forum-details']/dns:div[@class='forum-name']/dns:a
```

as shown in the configuration window shown in figure 6.

Figure 6. The WebCrawler Workflow



Dialog - 2:517:493 - XPath(top-category-entries)

File

Settings | Flow Variables | Job Manager Selection | Memory Policy

XML column: XML Document

New column name: top-category-entries

☑ Remove source column.

XPath query

```
//dns:td[@class='forum-details']/dns:div[@class='forum-name']/dns:a
```

Return type

XPath data type: Node-Set (Collection of XML cells)

Options:

☑ Return missing cell on no match. If unchecked an empty list will be returned.

XML fragment name: fragment

Namespaces

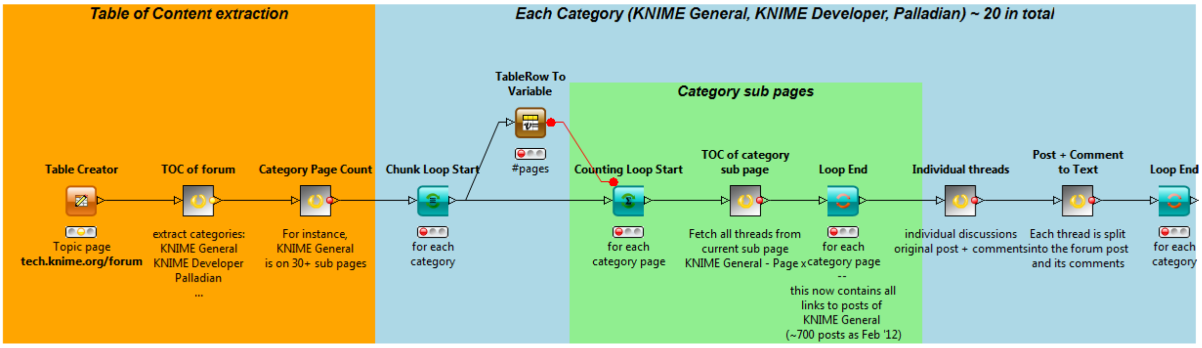
Prefix	Namespace

☑ Incorporate namespace of the root element.

Prefix of root's namespace: dns

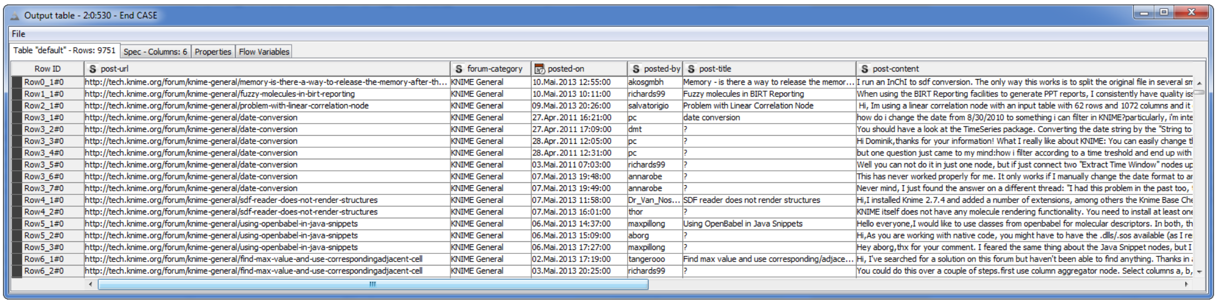
For each category, then, all links to subpages are explored and the html content is downloaded using a combination of the HtmlParser node and of XPath nodes as before. The complete "WebCrawler" workflow is shown in figure 7. This workflow has been later incorporated in a metanode and used as the first node in the upcoming workflows. To see its structure, please refer to the the first metanode in the next workflows.

Figure 7. The WebCrawler Workflow



The final result is a data table with a row for each post (either question or comment) including its URL in the forum, the category where it was posted, the date when it was posted, its author, its title, and its content (Fig. 8).

Figure 8. Post data after crawling the KNIME Forum.


Output table - 2:0:530 - End CASE
File
Table "default" - Rows: 9751
Spec - Columns: 6
Properties
Flow Variables

Row ID	S post-url	S forum-category	posted-on	S posted-by	S post-title	S post-content
Row0_1#0	http://tech.knime.org/forum/knime-general/memory-is-there-a-way-to-release-the-memory-after-th…	KNIME General	10.Mai.2013 12:55:00	akosgmbh	Memory - is there a way to release the memor…	I run an InChI to sdf conversion. The only way this works is to split the original file in several sm
Row1_1#0	http://tech.knime.org/forum/knime-general/fuzzy-molecules-in-birt-reporting	KNIME General	10.Mai.2013 10:11:00	richards99	Fuzzy molecules in BIRT Reporting	When using the BIRT Reporting facilities to generate PPT reports, I consistently have quality iss
Row2_1#0	http://tech.knime.org/forum/knime-general/problem-with-linear-correlation-node	KNIME General	09.Mai.2013 20:26:00	salvatorigio	Problem with Linear Correlation Node	Hi, Im using a linear correlation node with an input table with 62 rows and 1072 columns and it
Row3_1#0	http://tech.knime.org/forum/knime-general/date-conversion	KNIME General	27.Apr.2011 16:21:00	pc	date conversion	how do i change the date from 8/30/2010 to something i can filter in KNIME?particularly, i'm inte
Row3_2#0	http://tech.knime.org/forum/knime-general/date-conversion	KNIME General	27.Apr.2011 17:09:00	dmt	?	You should have a look at the TimeSeries package. Converting the date string by the "String to
Row3_3#0	http://tech.knime.org/forum/knime-general/date-conversion	KNIME General	28.Apr.2011 12:05:00	pc	?	Hi Dominik,thanks for your information! What I really like about KNIME: You can easily change t
Row3_4#0	http://tech.knime.org/forum/knime-general/date-conversion	KNIME General	28.Apr.2011 12:31:00	pc	?	but one question just came to my mind:how i filter according to a time treshold and end up with
Row3_5#0	http://tech.knime.org/forum/knime-general/date-conversion	KNIME General	03.Mai.2011 07:03:00	richards99	?	Well you can not do it in just one node, but if just connect two "Extract Time Window" nodes up
Row3_6#0	http://tech.knime.org/forum/knime-general/date-conversion	KNIME General	07.Mai.2013 19:48:00	annarobe	?	This has never worked properly for me. It only works if I manually change the date format to ar
Row3_7#0	http://tech.knime.org/forum/knime-general/date-conversion	KNIME General	07.Mai.2013 19:49:00	annarobe	?	Never mind, I just found the answer on a different thread: "I had this problem in the past too,
Row4_1#0	http://tech.knime.org/forum/knime-general/sdf-reader-does-not-render-structures	KNIME General	07.Mai.2013 11:58:00	Dr_Van_Nos…	SDF reader does not render structures	Hi,I installed Knime 2.7.4 and added a number of extensions, among others the Knime Base Che
Row4_2#0	http://tech.knime.org/forum/knime-general/sdf-reader-does-not-render-structures	KNIME General	07.Mai.2013 16:01:00	thor	?	KNIME itself does not have any molecule rendering functionality. You need to install at least one
Row5_1#0	http://tech.knime.org/forum/knime-general/using-openbabel-in-java-snippets	KNIME General	06.Mai.2013 14:37:00	maxpillong	Using OpenBabel in Java Snippets	Hello everyone,I would like to use classes from openbabel for molecular descriptors. In both, th
Row5_2#0	http://tech.knime.org/forum/knime-general/using-openbabel-in-java-snippets	KNIME General	06.Mai.2013 15:09:00	aborg	?	Hi,As you are working with native code, you might have to have the .dlls/.sos available (as I re
Row5_3#0	http://tech.knime.org/forum/knime-general/using-openbabel-in-java-snippets	KNIME General	06.Mai.2013 17:27:00	maxpillong	?	Hey aborg,thx for your comment. I feared the same thing about the Java Snippet nodes, but I
Row6_1#0	http://tech.knime.org/forum/knime-general/find-max-value-and-use-correspondingadjacent-cell	KNIME General	02.Mai.2013 17:19:00	tangerooo	Find max value and use corresponding/adjace…	Hi, I've searched for a solution on this forum but haven't been able to find anything. Thanks in a
Row6_2#0	http://tech.knime.org/forum/knime-general/find-max-value-and-use-correspondingadjacent-cell	KNIME General	03.Mai.2013 20:25:00	richards99	?	You could do this over a couple of steps.first use column aggregator node. Select columns a, b,

The "WebCrawler" workflow has been eventually wrapped into a meta-node named "Knime Forum Data" to be used in all other workflows to fetch the forum data.

Web Analytics and Simple Statistics

Many questions can be answered digging in the forum data: about the users, about the KNIME platform, about the efficiency and the scalability of the forum itself, and finally about the discussed topics. Especially discussed topics can help with the development of new features and issue identification.

Here we list a few common questions that have risen when we started analyzing the forum data. How many users does the KNIME Forum have? How active are they? Has the forum grown over the past years? How long is the response time to be expected for a new post? Is then the forum a useful channel to get information? What are the discussed topics? Have they changed in time? Who's talking to whom? Who seems to play the role of the accredited expert on a particular topic?

In this section, we concentrate on the forum users and on the role that the forum has been playing in answering their questions. In particular, we would like to measure the growth of the KNIME Forum in terms of number of users and posts. We would also like to measure the efficiency of the KNIME Forum in answering questions about the KNIME platform and to check whether such efficiency decreases with a higher volume of posts and users. Finally, we would like to quantify how much of the forum is KNIME driven and how much is community driven. That is: Is the forum too strictly controlled by KNIMers or is it the natural expression of the KNIME community? A simple statistics should suffice for such tasks (workflow "KNIME Forum simple statistics").

Forum Growth over the Years

Figure 9. Number of active users per year from 2007 to 2012

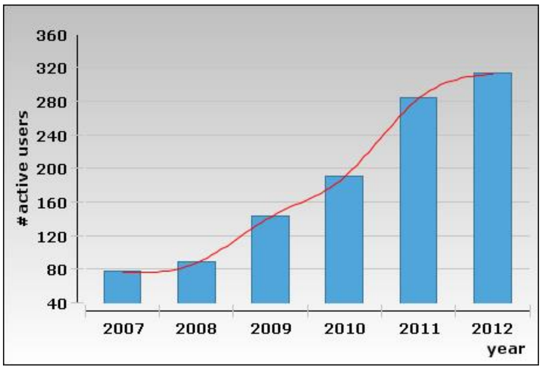
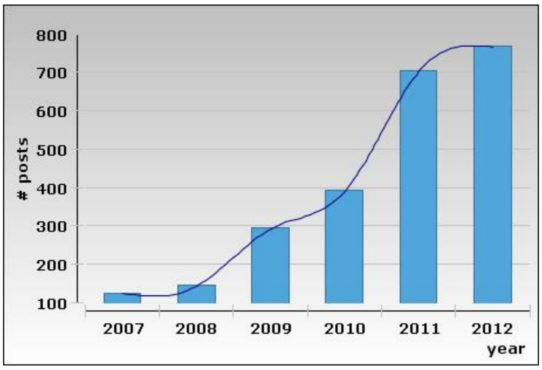


Figure 10. Number of posts per year from 2007 to 2012



After extracting the year from the posting date in the parsed data, the number of active users was counted for each year. This was obtained by means of a GroupBy node aggregating by year and applying the unique count to the post author field. After connecting the workflow to a simple report, the bar chart in figure 9 was derived, showing how the absolute number of active users has grown in the time interval 2007-2012. The same plot could be drawn at a higher time resolution considering the time month+year, instead of just year. However, the plot would not change much, besides some more noise in the peak variations.

Similarly, the number of posts for each year could be displayed. According to how the data have been parsed, a post-url with title represents the source post, that is the post with the question; but a post-url without title represents an answer to the original post. To count the missing values in the title field by post-url means to count how many answers each post has got. Aggregating again by year and counting the number of unique post-url defines the number of posts for each year in the KNIME Forum (Fig. 10).

Both the number of active users and the number of posts have been growing steadily over the last 6 years. A big increase in number of posts is clearly visible between 2010 and 2011, when KNIME became more widely used and many more users joined the forum.

Forum Efficiency and Scalability

So, the forum base has been growing. But has the KNIME Forum managed to grow with its base? Has the forum efficiency suffered from the bigger load of posts, questions, and answers? In short, has the KNIME Forum proven to be capable of scaling? The second part of the "KNIME Forum simple statistics" workflow measures the forum efficiency and scalability in terms of response time and number of comments.

Figure 11. Scakability of the Answer Process per year from 2007 to 2012

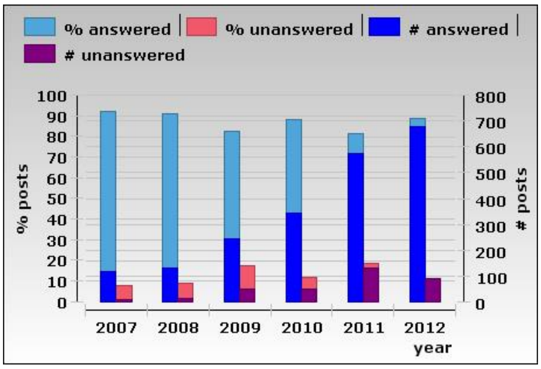
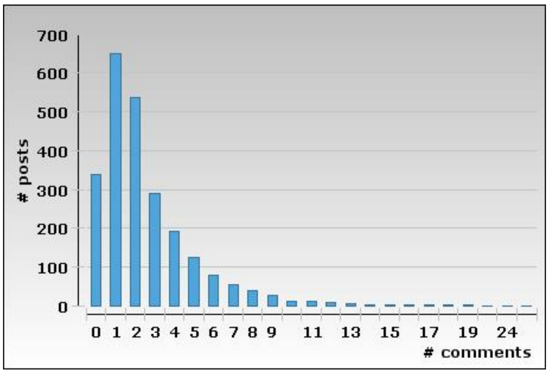


Figure 12. Response Time in Number of Comments for all posts posted between 2007 and 2012



As seen in the previous section, the number of rows with missing title was counted for each post-url. Such rows represent comments to the original posted question. The total number of answered and unanswered post, as well as their percentage, was calculated by year and reported in the corresponding report in the bar chart in figure 11. There you can clearly see the total number of posts growing, both answered and unanswered (blue and purple bars), while the percentages of answered vs. unanswered posts remain stable, between 80%-90% of answered posts vs. less than 20% of unanswered posts each year. This alone shows that the KNIME forum has been quite reliable over the years in its capability to deliver answers to posted questions, even if the number of posts has been growing fast.

So, it has been able to deliver answers, but how fast? Posting a question about KNIME today and getting an answer in two weeks is of not much help. In the next part of the workflow the response time was measured, as the number of days or as the number of comments needed to get the answer.

Let's talk about number of comments. With a GroupBy node, we count the number of comments (rows with missing titles) for each post-url. Then we count the number of post-url with a specified number of comments across all available years. The result is the bar chart in figure 12, where the number of comments is reported on the x-axis and the number of posts receiving that number of comments is reported on the y-axis.

Of course, we need to keep in mind that not all comments are useful. KNIME Forum users are very polite and many of the comments are just thank you notes. But still the number of comments for each post gives us an idea of how fast it was to answer the posted question.

We also need to keep in mind, that a few posts are actually announcements or, worse, spam, especially in the early time of the KNIME Forum life. Most of these posts lie in the first bar of posts with 0 comments of the bar chart in figure 12. In the same chart, we see that most posts are answered by only one or at maximum two comments, where probably the second comment is a thank you note. Only a few posts receive more than 2 comments in their lifetime. This is good hope for new KNIME forum subscribers: somebody out there already knows the answer to their question and will quickly share it on the forum.

Figure 13. Average Number of Days till first Comment per year from 2007 to 2012

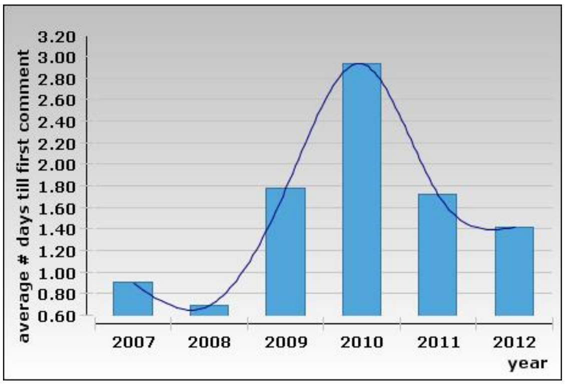
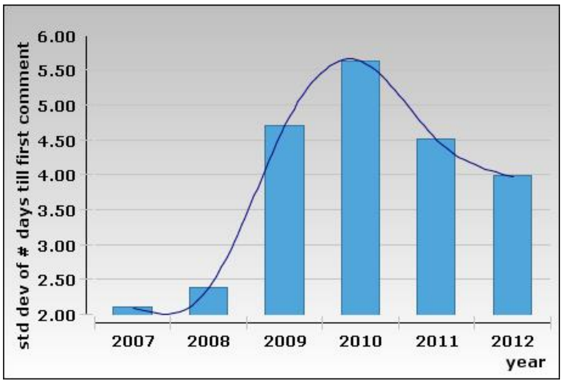


Figure 14. Standard Deviation of the Number of Days till first Comment per year 2007 and 2012



Let’s move now to the number of days needed to get an answer on the forum. Sorting the comments by date and calculating the difference between the post date and the first comment date in number of days offers a first measure of the response time (in number of days) and therefore of the forum efficiency.

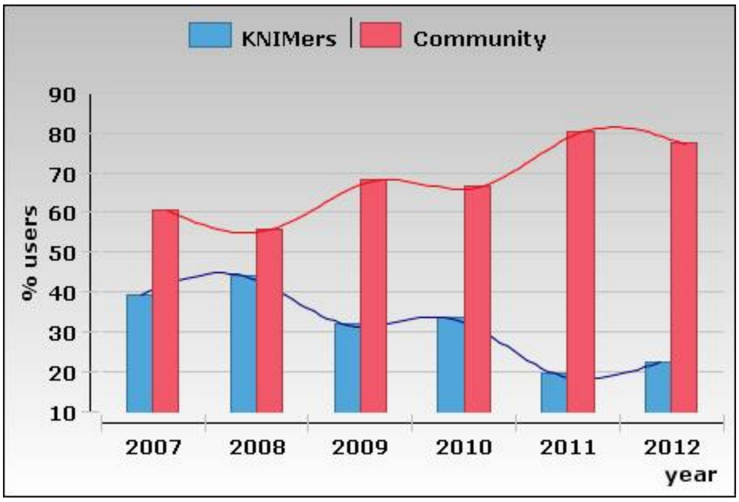
The average and the standard deviation of the response time (difference between first comment date and original post date) has been calculated for all posts in each year and reported in figures 13 and 14. In most years, you needed to wait one or maximum 1.5 days till you got a first (and probably final, based on figure 12) answer. In 2010 things looked a bit more complex. But this is due to spam messages that never got answered and are still part of the KNIME forum data. Indeed, this phenomenon can also be observed in figure 14, where in 2010 the standard deviation of the number of days to pass till first answer is much higher than in the other years, taking into account those spam posts with no answer that drive the average measure up.

Forum Ownership

And finally, who is answering all these questions? Is the KNIME Forum a spontaneous place or is it fully controlled by the KNIME employees?

Here, we checked the post-author field of the data. Since the KNIME employees, who are active in the KNIME Forum, are not that many and easily identifiable, we could isolate the KNIME-originating answers. All other answers must have come from the KNIME community. We then counted the number of comments per each year from KNIME employees and from the KNIME community and transformed the results from absolute numbers to percentages. In figure 15, the percentage of KNIME answers (blue bars) decreases over the years with the increasing number of posts, while the community contribution takes over, covering 80% of the post answers in 2011 and 2012.

Figure 15. Users Answers: KNIME vs. Community



Topic Classification

So, the number of users in the KNIME Forum has been growing and they have been discussing more and more over the years. The question now is: What have they been discussing about? Are there recurrent themes? Are there important themes, like show stopper bugs? In this section, we would like to discover the discussion topics and track the topic changes over the years. This section's workflow is "KNIME Forum Classify Posts".

In the forum data we can find discussion texts, but a set of pre-defined topic categories to classify such texts is missing. Usually when the target classes are missing we resolve to some clustering technique. The idea then would be to extract the keywords from each post thread and cluster all post threads on the basis of the used keywords. However, this procedure might be sub-optimal in this case: on one side, because each topic category can be described by many similar keywords; on the other side, because at the end we still need to associate a topic to each cluster.

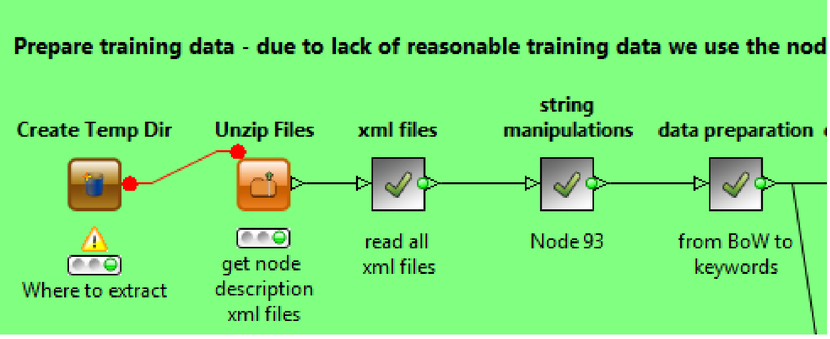
Alternatively, we can try to borrow a new topic ontology from some documentation site available within KNIME. Indeed, a node description can be found for each selected node in the right panel in the KNIME workbench. The node description text comes from an XML file available within KNIME. The node name could be used as the topic class and the description text could be used as the source text describing that topic class. All those description XML files have been collected, zipped, and used to build the topic ontology for the forum topic detection.

Unzipping Files and the "drop" folder

As usual, the first part of the workflow is dedicated to clean up and organize the data for the upcoming classification: on one branch, the zipped file containing the XML files with the node descriptions; on the other branch the results of the web crawling workflow. The first text-processing part in the branch in the workflow, processing the XML description texts, is shown in figure 16.

The "Unzip Files" node unzips the file into a temporary location, defined by the previous "Create Temp Dir" node. Both paths - the path of the zip file and the path of temporary location - are passed to the "Unzip Files" node as flow variables. The temporary location path is provided by the "Create Temp Dir" node. The input zip file path is provided via the "drop" folder utility.

Figure 16. Extracting text documents from a zipped file



The "drop" folder utility is available to all reader nodes. In the system folder of the reader node you can insert a subfolder, named "drop", and some files in the "drop" folder. With such a folder configuration, the reader node creates as many flow variables as many files in the "drop" folder. Each flow variable is named "knime.node (drop) <filename>" and contains the current path to the referred file.

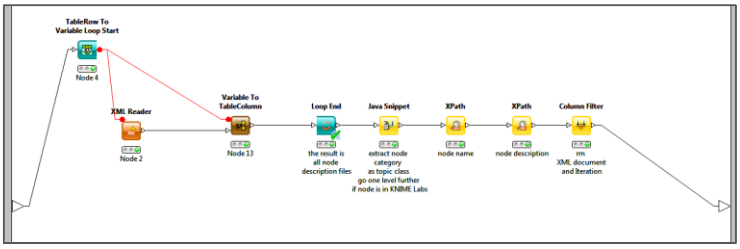
If the resetted workflow is exported including data, the final zip file contains the workflow and the folder structure of the input files for the reader node. Also, referencing the input file by means of its "drop" path allows moving the workflow to another machine without having to reconfigure the reader node. The "drop" folder utility is thus very useful to move workflows around, keeping the reader nodes validly configured.

Text Pre-processing

The node description files are in XML format. The "xml files" meta-node reads all the description files with an XML Reader node inside a loop, then uses a Java Snippet node to extract their parent category from their URI, and finally with a few XPath nodes extracts the node name and description from the XML file content.

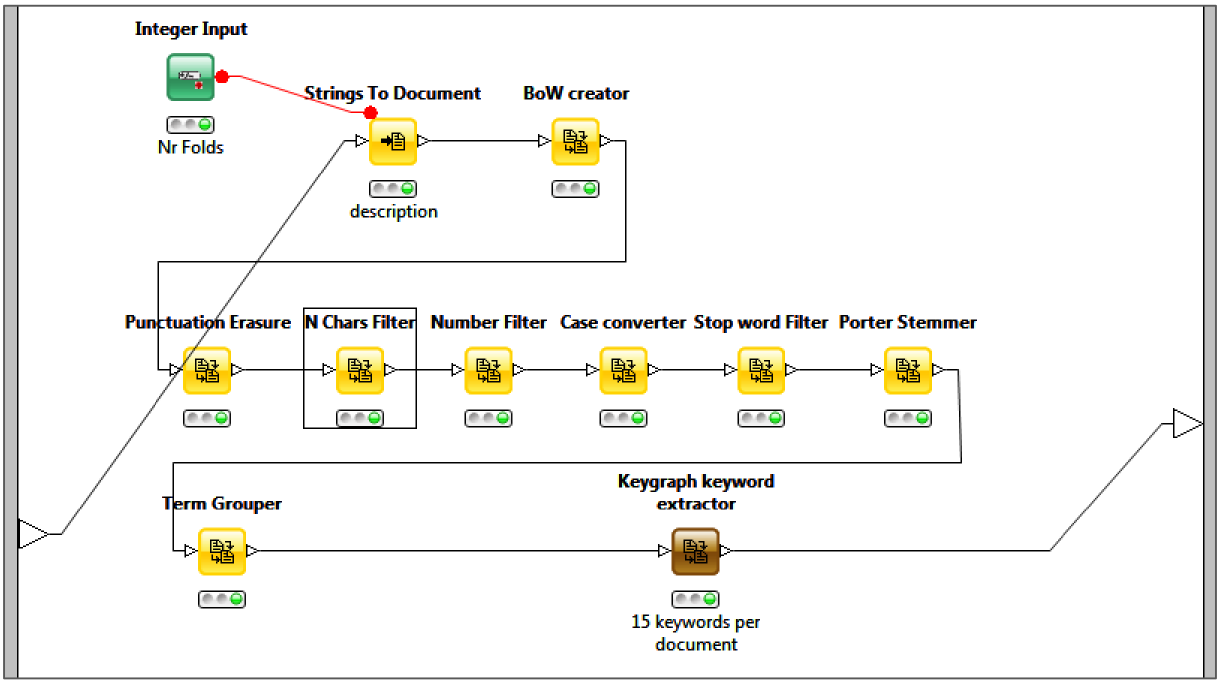
The „string manipulations" meta-node only contains a number of String Manipulation nodes to remove copyright and other content unrelated sentences from the text descriptions.

Figure 17. The sub-workflow in the"xml files" meta-node


TableRow To Variable Loop Start
Node 4
XML Reader
Node 2
Variable To TableColumn
Node 13
Loop End
the result is all node description files
Java Snippet
extract node category as topic class go one level further if node is in KNIME Labs
XPath
node name
XPath
node description
Column Filter
rm XML document and Iteration

At last, the "data preparation" meta-node prepares the text data for classification, by converting all description text strings into Document type, extracting the bag of words, getting rid of all non-content related words, reducing words to their stems, deleting conflicting tags, and finally extracting the keywords for each document (Fig. 18). The Integer Input Quickform node defines the number of folds for the cross-validation node hosting the classification algorithm down the workflow.

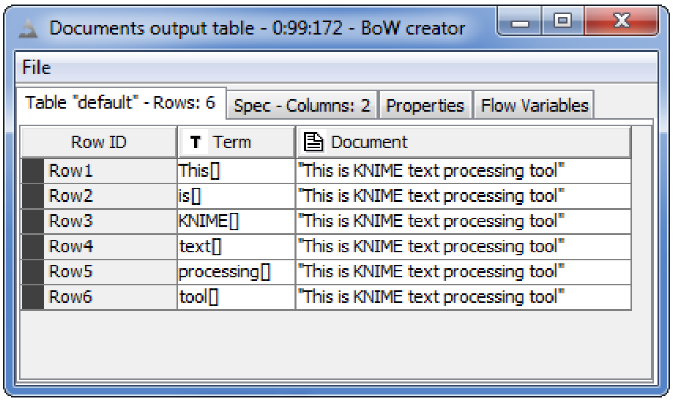
Figure 18. Sub-workflow in the "data preparation" meta-node



Note that the document type input is required by most text processing nodes to operate. A document data includes some additional information in the shape of tags besides the text, like the author name, the title, the document source, the document category, the publication date, and the document type.

The BoW Creator node creates a bag of words (BoW); that is it extracts each word from each document. For example, the sentence "this is KNIME text processing tool", after the conversion to document type and the BoW creation, produces the data table displayed in figure 19, where the words are isolated and still associated to the original document.

Figure 19. BoW for "This is KNIME text processing tool".


Documents output table - 0:99:172 - BoW creator
File
Table "default" - Rows: 6
Spec - Columns: 2
Properties
Flow Variables

Row ID	T Term	Document
Row1	This[]	"This is KNIME text processing tool"
Row2	is[]	"This is KNIME text processing tool"
Row3	KNIME[]	"This is KNIME text processing tool"
Row4	text[]	"This is KNIME text processing tool"
Row5	processing[]	"This is KNIME text processing tool"
Row6	tool[]	"This is KNIME text processing tool"

Now, words need to be reduced to their stem. Indeed, if we classify "process" and "processing" as separate words, the result might be suboptimal. We need to use both words as the same input value; that is, we need to use their original stem. The Porter Stemmer node produces the stem for each term in the input data table, according to the Porter stemmer algorithm [1]. In our previous example, an input data table with the list of terms "process", "processing" would produce the following list of stemmed terms: "process[]", "process[]" with the associated document on the same row.

The Term Grouper node is used here just to delete conflicting tags in instances of the same word. For example, if we have words tagged for "good" and "bad" and a word happens to have tag sometimes "good" and sometimes "bad", all tags for that word will be deleted. We use this node to get rid of tag inconsistencies.

The Keygraph keyword extractor is the main node of this text processing meta-node. This node analyses documents and extracts relevant keywords using a graph-based approach [2]. In a very rough summary, the Keygraph keyword extractor node sorts terms by frequency measures and by edges in a graph. The most frequent and the most connected terms obtain the highest scores and therefore are the most appropriate candidates as keywords for the document. The node takes a list of terms, with the document they belong to, and produces a significance score for each term. Only the top 15 most significant keywords are kept for each document.

This part of the workflow has been applied to the XML node description documents, to prepare the training set. On the other branch, the post texts coming from the KNIME Forum have been processed similarly, including the removal of duplicates, the cleaning of html tags, the bag of word creation, the stemming, and the keyword extraction. We only used the posts from the KNIME Users and from the KNIME General discussions for the test set, since they are the most populated and most general discussions anyway.

Since the two sets, the training and the test set, come from two related but different sources, it might happen that the two word sets do not completely overlap. The meta-node "fix domain" fixes a possible domain discrepancy among word sets, by including the words present in one of the two sets and missing in the other.

Model Training and Evaluation: the Decision Tree Ensemble

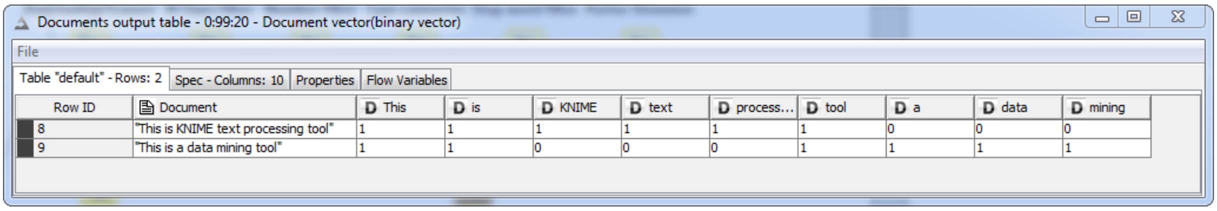
Each node description document has now a set of terms with significance measure and it has been tagged with the node category. The goal is to use such category and such list of terms to train a classifier and to apply the classifier later to the forum posts, in order to associate a discussion topic to each post.

As a classification model a decision tree ensemble was adopted. KNIME implements the decision tree ensemble via the Tree Ensemble Learner node and the Tree Ensemble Predictor node. The Tree Ensemble Learner node builds an ensemble of decision trees, like a random forest variant. Each of the decision tree models is trained on a different set of rows and/or on a different set of columns, randomly selected at each iteration. The output model is an ensemble of decision tree models. The Tree Ensemble Predictor node applies all decision trees to each data row and uses the simple majority vote for prediction.

The Tree Ensemble Learner node is quite complex, because of the many settings to be provided for the decision tree ensemble training. Indeed, it has three setting tabs in the configuration window: one to select the attributes (that is the data columns), one to select the decision trees training parameters (like the entropy measure), and one to select the ensemble settings (like the number of trees to be trained).

As we said at the beginning of this section, for now we have only one column with a list of terms and one column with a significance measure. This does not give enough details even for the most powerful classification algorithm. Thus, usually, text mining techniques change the data perspective from a long list of terms to the matrix of keywords found in each document. The matrix representation gives more information about which keywords have been found in the document and which ones have not. Using the example in figure 19, we want to move from that list-based to the vector-based representation shown in figure 20. Here the document "This is KNIME text processing tool" has been represented as having the word "KNIME" but not the word "data".

Figure 20. The terms matrix representation of a document

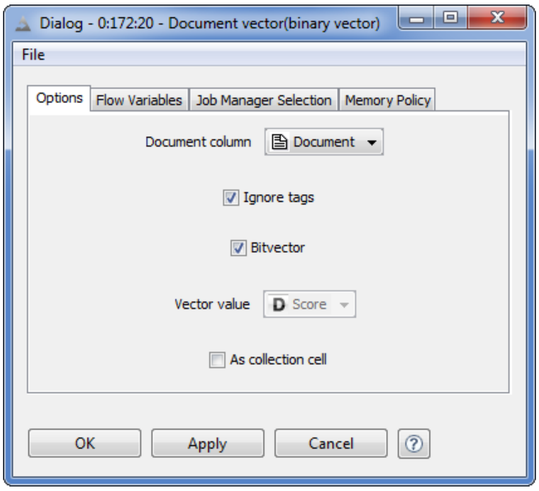

Documents output table - 0:99:20 - Document vector(binary vector)

Row ID	Document	This	is	KNIME	text	process...	tool	a	data	mining
8	"This is KNIME text processing tool"	1	1	1	1	1	1	0	0	0
9	"This is a data mining tool"	1	1	0	0	0	1	1	1	1

The Document Vector node creates a document vector for each document in the terms space. The row matrix dimension is the number of distinct terms in the input column. The feature vectors can take Boolean values (1= the term is in the document; 0= the term is not in the document) or the values of another column, for example a significance measure column. Options “Bitvector” and “Vector value” are mutually exclusive and define the type of cell content. Since the training set we used is not a native data set of the forum, but just an approximation, we decided not to rely on the significance measures produced by the “Keygraph keywords extractor” node and we enabled the “Bitvector” option (Fig. 21).

We have now an appropriate set of input features to feed the Tree Ensemble Learner. We still need a column to work as target for the decision tree. That is, we need to define the class for each one of the node description documents. Previously, we assumed that the node category would have been the class of each document and therefore we embedded the node category into the document category. Now it is the moment to convert the document category into a class for each document, i.e. each matrix row. The Category To Class node extracts the category from a document cell and appends it into a new column, named “Document class”.

Figure 21. The "Document vector" configuration window


Dialog - 0:172:20 - Document vector(binary vector)

Document column: Document

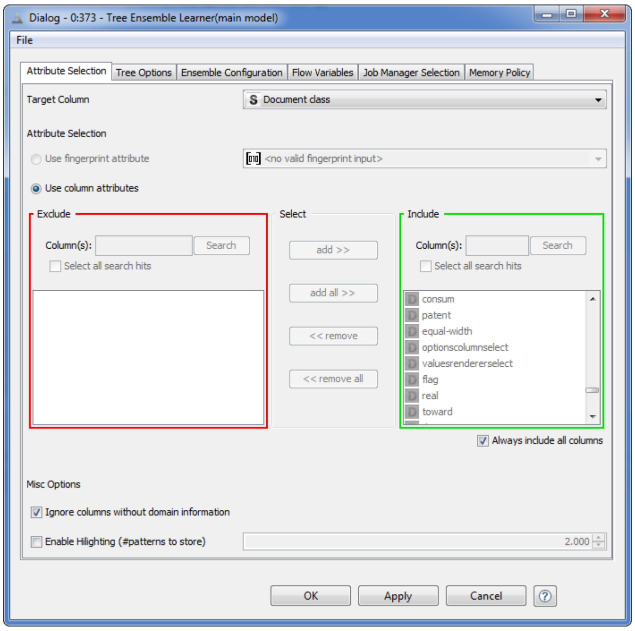
☑ Ignore tags

☑ Bitvector

Vector value: Score

□ As collection cell

Figure 22. "Attribute Selection" tab in the configuration window of the "Tree ensemble learner".


Dialog - 0:373 - Tree Ensemble Learner(main model)

Target Column: Document class

Attribute Selection

○ Use fingerprint attribute <no valid fingerprint input>

● Use column attributes

Exclude: □ Select all search hits

Include: □ Select all search hits

consum
patent
equal-width
optionscolumnselect
valuesrendererselect
flag
real
toward

☑ Always include all columns

Misc Options

☑ Ignore columns without domain information

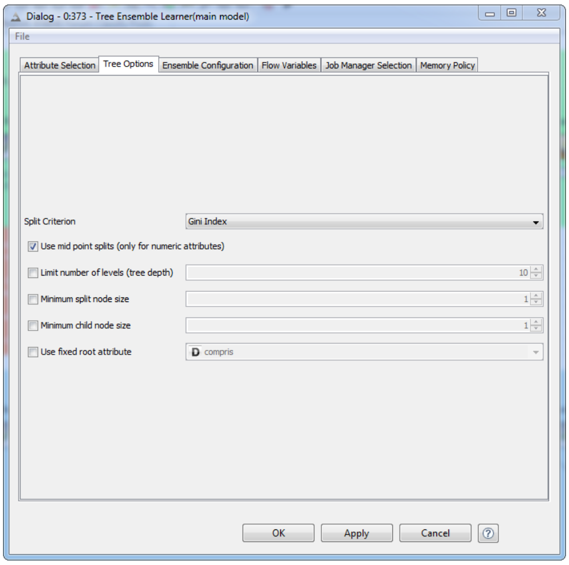
□ Enable Hilighting (#patterns to store) 2.000

The Tree Ensemble Learner is then fed with the term binary features and with the document class as the target for the decision trees (Fig. 22). The Gini index is adopted as entropy measure and the size of nodes and leaves is not limited (Fig. 23). And finally, for the ensemble configuration (Fig. 24) all rows are sampled (fraction=1) using the random sampling with replacement strategy (bootstrapping) and only some columns are randomly selected, i.e. the square root of the number of columns, to train each tree.

In theory, we could use the whole data set coming from the node description files as training set, since the real test set is coming from a different source. However, just for evaluation purposes, we create a test set from the same node description data set to evaluate the performance of the ensemble tree. It is clear that if the ensemble tree fails the classification task on a test set with the same characteristics as the training set, it will have no hope of performing better on a loselier related data set.

On the last part of the upper branch of the “KNIME Forum Classify Posts” workflow (Fig. 25), the node description based data set is split in two parts (80%-20%). 80% of the data is used to train the Tree Ensemble Learner node, with the same configuration settings as described above, and the remaining 20% is used to apply the ensemble model using the Tree Ensemble Predictor node. Evaluation is carried out by a Scorer node and shows a 0.754 accuracy on the test set.

Figure 23. "Tree Options" tab in the configuration window of the "Tree ensemble learner".



Dialog - 0:373 - Tree Ensemble Learner(main model)

File

Attribute Selection | Tree Options | Ensemble Configuration | Flow Variables | Job Manager Selection | Memory Policy

Split Criterion: Gini Index

☑ Use mid point splits (only for numeric attributes)

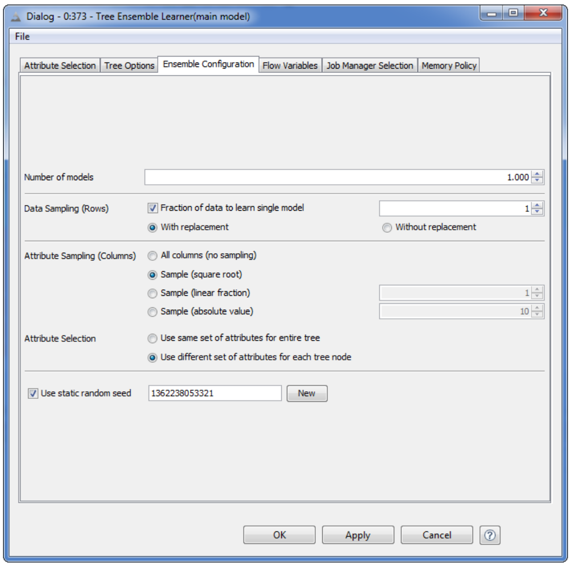
☐ Limit number of levels (tree depth): 10

☐ Minimum split node size: 1

☐ Minimum child node size: 1

☐ Use fixed root attribute: D compris

Figure 24. "Ensemble Configuration" tab in the configuration window of the "Tree ensemble learner".



Dialog - 0:373 - Tree Ensemble Learner(main model)

File

Attribute Selection | Tree Options | Ensemble Configuration | Flow Variables | Job Manager Selection | Memory Policy

Number of models: 1.000

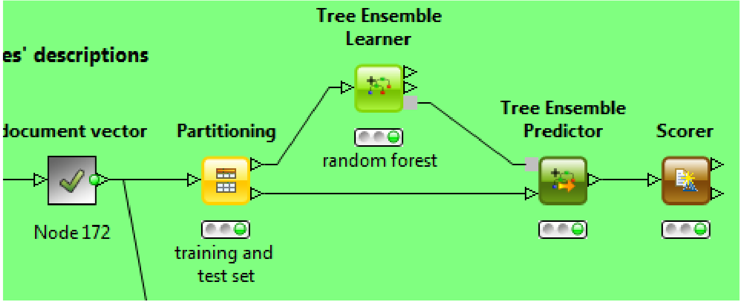
Data Sampling (Rows): ☑ Fraction of data to learn single model: 1
● With replacement ○ Without replacement

Attribute Sampling (Columns):
○ All columns (no sampling)
● Sample (square root)
○ Sample (linear fraction): 1
○ Sample (absolute value): 10

Attribute Selection:
○ Use same set of attributes for entire tree
● Use different set of attributes for each tree node

☑ Use static random seed: 1362238053321 New

Figure 25. Tree Ensemble evaluation part of the workflow



The real ensemble model to be used is trained on all node description data, to give more chances to the model to perform well on foreign data.

The same term vectorization described at the beginning of this section is applied again to all documents coming from the forum posts. A Tree Ensemble Predictor node uses the ensemble model trained on the node description data to produce a topic class for the forum posts. For example, a post containing, among others, the terms plot[], domain[], scatter[], size[], color[], and dimension[] is classified as belonging in a first instance to class-category "data views" and in a second instance to class-category "flow control".

Topic Shift Detection

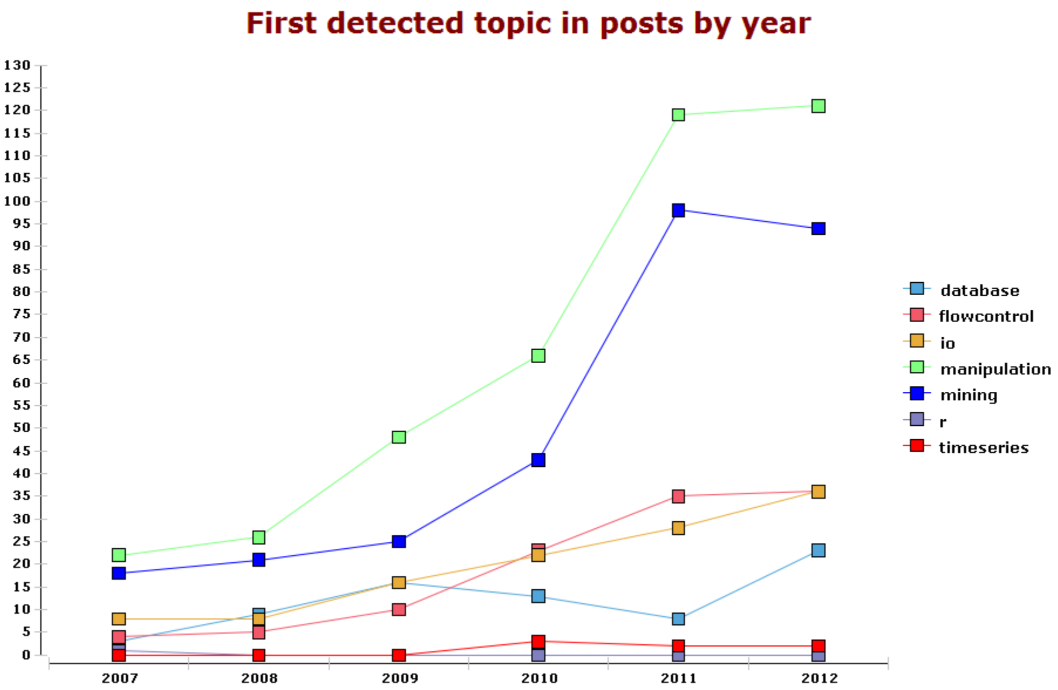
Indeed, most posts are not just simple questions on how to run a specific node in a specific category. More often they ask how to combine nodes from different categories. It is hard then to associate one post to one category only. The most frequent situation has one post associated to two or even more categories.

For example, the post "How can I loop on a list of files and concatenate the contents?" includes the "io" category, the loops in the "flowcontrol" category, and the concatenation in the "data manipulation" category. So, if the option "Append individual class confidences" is enabled in the Tree Ensemble Predictor node, the results show all possible output categories with their corresponding probabilities for each post. For the post mentioned before, the ensemble tree would produce "io" with highest probability, followed by "flowcontrol", and then "data manipulation" categories.

Looping on all class categories and isolating the three class categories with highest probability assigns the top three output classes, i.e. with highest likelihood, to each forum post: topic 1, topic 2, and topic 3. Finally, the number of posts is counted for each year and for each category in topic 1, topic 2, and topic 3, and a final report shows the number of posts by class category per year and in each one of the three topics.

Figure 26 shows the dominance of posts having data mining or data manipulation as main topic. The chart for the secondary and tertiary topic also shows a dominance in the number of posts about data mining nodes and data manipulation nodes in most years, but especially in 2011 and 2012.

Figure 26. Number of posts in the forum about a certain topic over the years



We know from the previous sections that the number of posts has grown dramatically in 2011 and 2012. From this topic detection analysis, we know that the biggest increase has been in data mining and data manipulation related posts, which is also the feeling a user gets when browsing the KNIME Forum.

Notice the little numbers for “r” and “timeseries” related questions. Indeed, while the part of data mining and data manipulation is quite mature and abundantly used, time series and R nodes are still in their prime infancy as of KNIME 2.7. With KNIME 2.8 both categories have been majorly improved by introducing the “Lag Column” node for time series lagging and the “R (Interactive)” category for interactive R coding. It would be interesting to know if the number of posts on these two topics would increase in 2013 and beyond.

Another interesting topic is “database”. The number of posts concerning databases reached its height in 2009 when most database nodes have been introduced, to slow down till 2011 and pick up again in 2012. In 2009 and in 2012 there have been requests via the KNIME Forum for additional database nodes and features.

“flowcontrol” related posts also increased after 2009, being the “flowcontrol” category virtually empty before 2009.

In conclusion, the strength of KNIME lied almost exclusively in the data mining and data manipulation nodes till 2012. However, flow control and io posts are also rapidly growing, while time series and r are probably waiting for the big improvements of KNIME 2.8 in 2013. Database issues and requests have been resolved or answered in 2009, till a new wave of requests for non-structured databases arrived in 2012.

Identify Contents and Users

At this point we know that many people have come to the KNIME Forum and we know what they have been talking about over the years. We would like now to know who the main actors in the different discussion groups are and how they interact with each other.

The goal here is to generate a report with one page for each discussion group to contain:

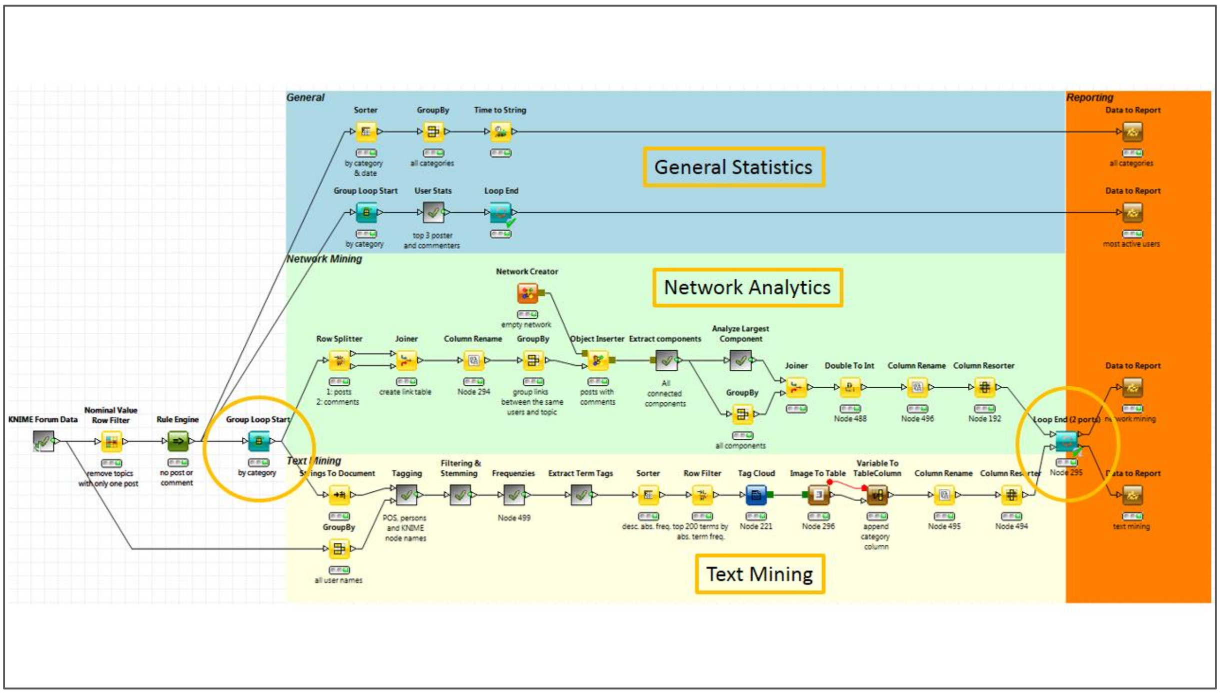
- A word cloud of the discussed topics
- A user interaction graph
- Some general statistic about the most active posters and the most active commenters

It would also be useful to have a first page in the report with the list of the analyzed discussion groups, each linked to its report page.

Starting again from the forum data in figure 8, in the workflow named “KNIME Forum Text and Network Analysis” (Fig. 27), discussion groups (forum-category field in the forum data table) with only one topic have been removed, leaving circa 20 forum categories. A GroupBy node then extracts the list of all those categories and a Group loop finds out the three most active posters and the three most active commenters for each forum category. Both data tables are exported to the report.

At this point, the big loop starts with the Group Loop Start node and loops over all forum categories, producing the word cloud and the user graph for each one of them. In the loop we have two branches. The lower branch calculates and visualizes the word cloud of the forum category as an image; the upper branch calculates and visualizes the graph representing the user interactions as an image. Both images are included as image cells in the output data tables.

Figure 27. Workflow "KNIME Forum Text and Network Analysis"



The Tag Cloud

In order to produce the word cloud for each forum category, we need to use the KNIME text processing libraries again.

All words in a document are tagged with the part of speech, being that a verb, a noun, an adverb, etc., using the POS Tagger node. User names in the document are also recognized and tagged as well as KNIME node and category names with two separate Dictionary Tagger nodes. Finally, the Bag of Words is created with the BoW Creator node. All those nodes are stored in the “Tagging” metanode.

The “Filtering & Stemming” metanode filters out pause words, numbers, adverbs, conjunctions, and short words in general as non relevant. The Kuhlen Stemmer [3] is applied to reduce words to their stem, the Term Grouper node groups all similar terms together, and the RegEx Filter node filters all terms matching the “0x.+” regular expression, that is all terms starting with “0x”, which are system log reports.

Metanode “Frequencies” calculates the frequencies to use as term size in the word cloud.

- IDF (Inverse Document Frequency) of a term is the logarithmic measure of the total number of documents divided by the number of documents containing that term. This is a measure of how important this word is for this particular group of documents.
- TF absolute is the number of times that a word appears in a document.
- TF relative is the number of times that a word appears in a document divided by the number of words in the same document.
- (TF relative *IDF) gives the measure of how representative a word is for that document and for its category.

Metanode “Extract Term Tags” extracts all previously encapsulated tags (Part Of Speech, Poster names, and KNIME node and category names) for filtering and coloring of the word cloud. The average frequency measures, the most frequent POS, poster and node name are calculated for each term across all documents. Finally, a color is assigned to nouns (NN), verbs (VB), poster names (PERSON), node names (DATE), and category names (TIME) in a Color Manager node.

Figure 28. Tag cloud using (IDF*TF relative) for KNIME Textprocessing forum category

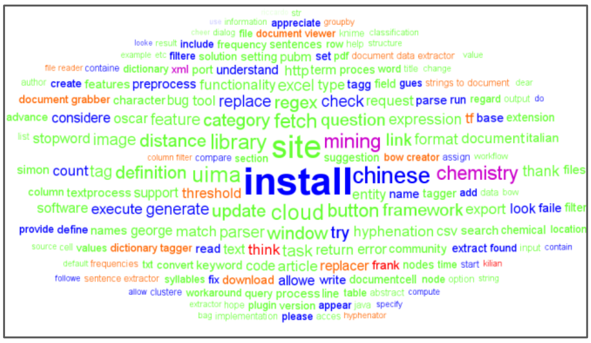
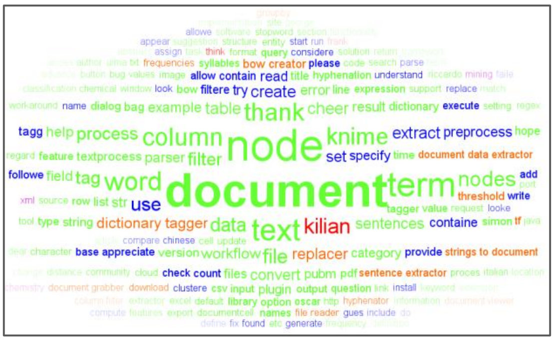

appreciate
file document viewer knime classification
include frequency sentences row help structure
filtere solution setting pubm set pdf document data extractor value
file reader containe dictionary xml port understand httpterm proces word title change
author create features preprocess functionality excel type tagg field gues strings to document dear
document grabber character bug tool replace regex check request parse run regard output do
advance considere oscar feature category fetch question expression tf base extension
stopword image distance library site mining link format document italian
column filter compare section suggestion bow creator assign workflow
simon count tag definition uima install chinese chemistry thank files
column textprocess support threshold entity name tagger add data bow
software execute generate update cloud button framework export look faile filter
provide define names george match parser window try hyphenation csv search chemical location
source cell values dictionary tagger read text think task return error community extract found input contain
default frequencies txt convert keyword code article replacer frank nodes time start kilian
followe sentence extractor syllables fix download allowe write documentcell node option string
allow clustere workaround query process line table abstract compute
extractor hope plugin version appear java specify
bag implementation please acces hyphenator

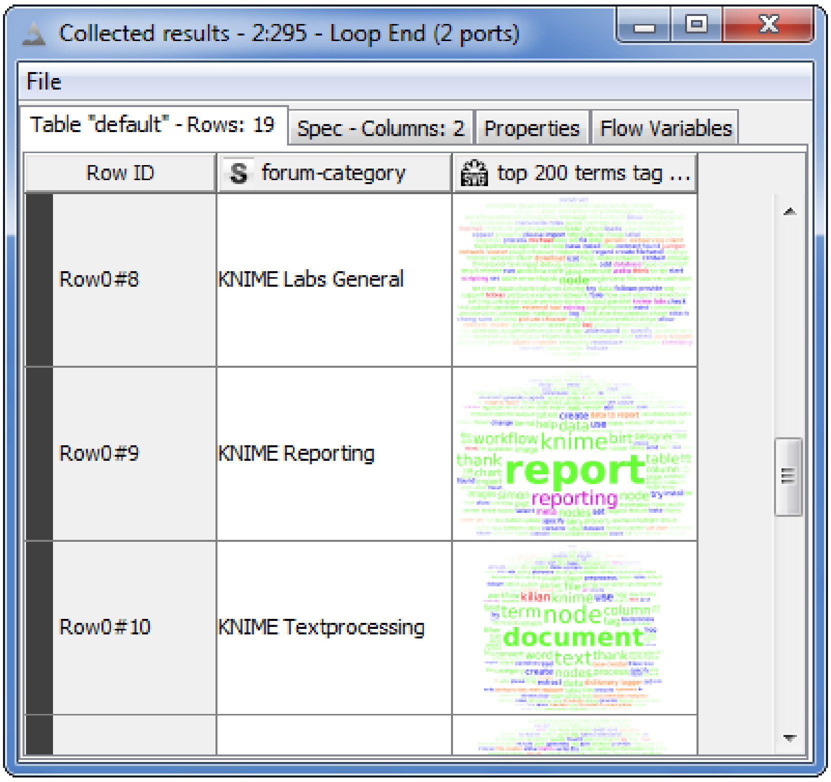
Figure 29. Tag cloud using (TF absolute) for KNIME Textprocessing forum category


allowe stopword
appear structure entity start run
assign think format query considere solution
frequencies syllables bow creator please code search parse
button bug values image allow contain read title hyphenation understand
classification chemical window look bow filtere try create error line expression support replace
workaround name dialog bag example table thank cheer result dictionary execute setting regex
tagg help process column node knime extract preprocess hope
regard feature textprocess parser filter set specify time document data extractor
followe field tag word document term nodes add threshold write
xml source row list str use tagger value request looke
tool type string dictionary tagger data text kilian sentences containe simon java
compare chinese cell update
character base appreciate version workflow file replacer category provide strings to document
community cloud check count files convert pubm pdf sentence extractor proces
document grabber download clustere csv input plugin output question link install
extractor excel default library option oscar http hyphenator
compute features export documentcell names file reader gues include
define fix found etc generate

The most frequent 200 terms, based on absolute frequency, for each forum category are isolated and fed into the Tag Cloud node. This produces a colored image of the word cloud of the 200 most frequent words across all documents in the selected discussion group. The term size in the word cloud can be proportional to (IDF*TF relative), IDF, TF relative, or TF absolute. The tag clouds for the KNIME Textprocessing discussion group using (IDF*TF relative) and (TF absolute) respectively are shown in figures 28 and 29.

The Image To Table node converts the image output of the Tag Cloud node into a data cell of type Image in a KNIME data table. This is then collected by the second output port of the Loop End (2 Ports) node and exported into the associated report as the "text mining" data set (Fig. 30).

Figure 30. Output Data Table of the text mining loop branch.


Collected results - 2:295 - Loop End (2 ports)

File

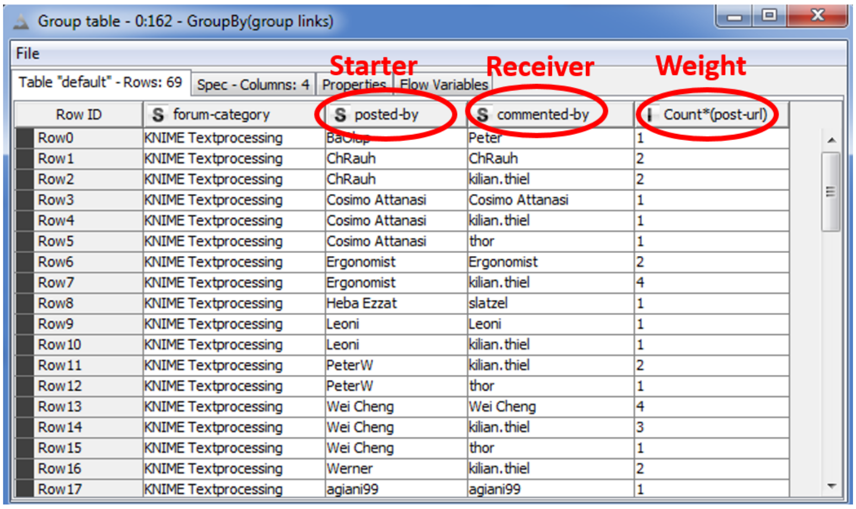
Table "default" - Rows: 19 | Spec - Columns: 2 | Properties | Flow Variables

Row ID	forum-category	top 200 terms tag …
Row0#8	KNIME Labs General	
Row0#9	KNIME Reporting	
Row0#10	KNIME Textprocessing	

Build the User Network

The upper branch of the Group loop is dedicated to Network analysis, to answer the question of who is talking to whom. First of all, by means of a Row Splitter, a Joiner, and a GroupBy node we build a data table as depicted in figure 31, with a source (the poster name), a target (the commenter name), and a weight (the number of comments) for every edge.

Figure 31. Data table to build the network graph with a source, a target, and a weight for every edge.


Group table - 0:162 - GroupBy(group links)
File
Table "default" - Rows: 69
Spec - Columns: 4
Properties
Flow Variables
Starter
Receiver
Weight

Row ID	S forum-category	S posted-by	S commented-by	I Count*(post-url)
Row0	KNIME Textprocessing	BaOlap	Peter	1
Row1	KNIME Textprocessing	ChRauh	ChRauh	2
Row2	KNIME Textprocessing	ChRauh	kilian.thiel	2
Row3	KNIME Textprocessing	Cosimo Attanasi	Cosimo Attanasi	1
Row4	KNIME Textprocessing	Cosimo Attanasi	kilian.thiel	1
Row5	KNIME Textprocessing	Cosimo Attanasi	thor	1
Row6	KNIME Textprocessing	Ergonomist	Ergonomist	2
Row7	KNIME Textprocessing	Ergonomist	kilian.thiel	4
Row8	KNIME Textprocessing	Heba Ezzat	slatzel	1
Row9	KNIME Textprocessing	Leoni	Leoni	1
Row10	KNIME Textprocessing	Leoni	kilian.thiel	1
Row11	KNIME Textprocessing	PeterW	kilian.thiel	2
Row12	KNIME Textprocessing	PeterW	thor	1
Row13	KNIME Textprocessing	Wei Cheng	Wei Cheng	4
Row14	KNIME Textprocessing	Wei Cheng	kilian.thiel	3
Row15	KNIME Textprocessing	Wei Cheng	thor	1
Row16	KNIME Textprocessing	Werner	kilian.thiel	2
Row17	KNIME Textprocessing	agiani99	agiani99	1

An empty Network object is allocated with the Network Creator node and shaped with the data table of figure 31 by means of an Object Inserter node. The result is a Network object representing the user graph for this forum category.

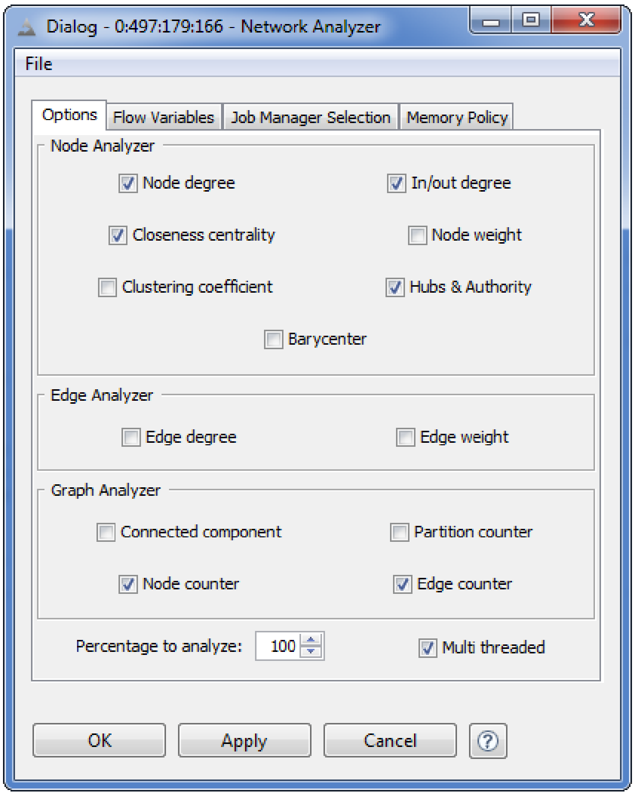
The metanode “Extract Components” calculates the number of edges, the number of nodes, and the number of objects (= number of edges + number of nodes) for each connected component of this graph.

The “Analyze Largest Component” metanode includes a sub-metanode, named “Create Network View”. This metanode uses the Network Analyzer node to calculate the node degree, the in/out degree, the closeness centrality, the Hubs and Authority score, the number of nodes, and the number of edges of the largest connected components.

- The node degree counts the number of incident edges. The percent degree is the percentage of edges in comparison to the total number of edges.
- The in/out degree counts the number of incoming/outgoing edges. The percent incoming/outgoing degree is the percentage of edges in comparison to the total amount of edges.
- The closeness centrality index divides the number of nodes by the sum of the distances from the analyzed node to all other nodes. The node with the highest value is the most central node of the graph [4].
- The Hubs & Authority scores measure each node’s authority and hub feature [5][6]. The essential idea is that a node is a hub to the extent that it links to authoritative nodes, and is an authority to the extent that it is linked from 'hub' nodes. To calculate these scores the implementation of JUNG the Java Universal Network/Graph Framework [7] is used.

Now the network graph visualization. The maximum score between the hub and authority score is kept, renamed as “Max Score”, and used in a Size Manager node to resize the nodes later in the network view. After that, nodes are classified as Hub if their Hub score is above 0.1, as authority if their authority score is above 0.1 and as neutral otherwise via a “Rule Engine” node. A “Color Manager” node colors in gray all neutral nodes, in orange all authoritative nodes, and in blue all hub nodes of the network graph.

Figure 32. Configuration Window of the "Network Analyzer" node inside the "Create Network View" metanode.



Dialog - 0:497:179:166 - Network Analyzer

File

Options | Flow Variables | Job Manager Selection | Memory Policy

Node Analyzer

☑ Node degree　☑ In/out degree

☑ Closeness centrality　□ Node weight

□ Clustering coefficient　☑ Hubs & Authority

□ Barycenter

Edge Analyzer

□ Edge degree　□ Edge weight

Graph Analyzer

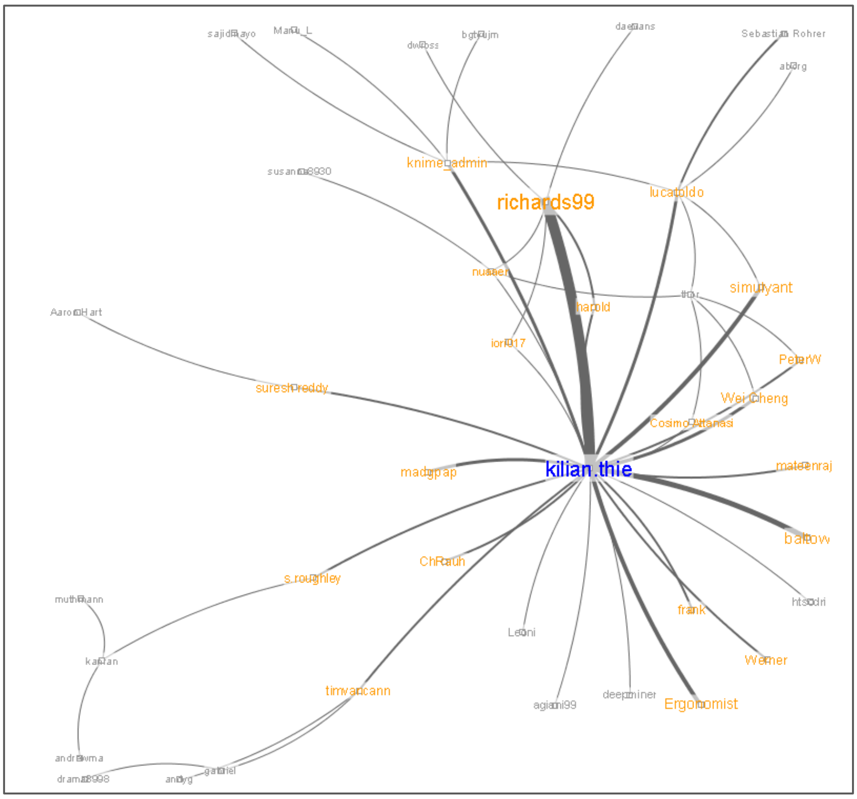
□ Connected component　□ Partition counter

☑ Node counter　☑ Edge counter

Percentage to analyze: 100　☑ Multi threaded

OK | Apply | Cancel | ?

Figure 33. User Network for the KNIME Textprocessing forum category.



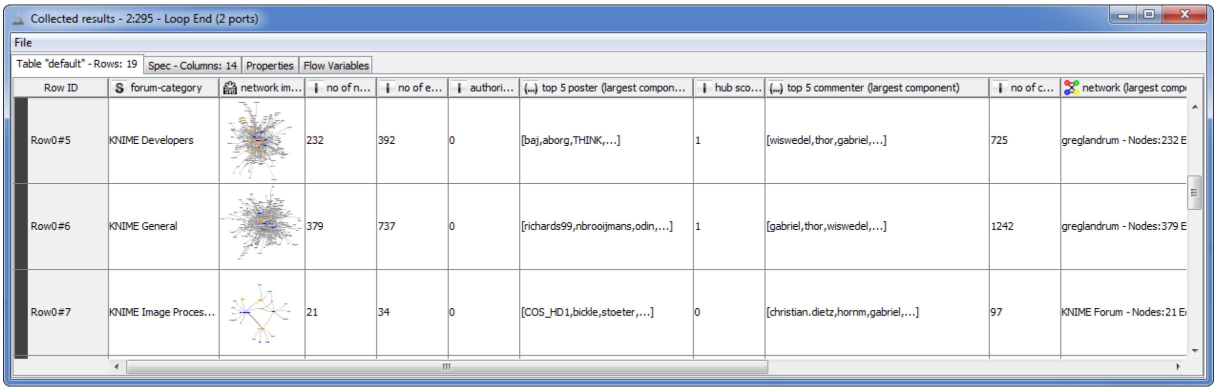
The node size (~Max Score) and color (Hub/Authority/neutral) is attached to each node by means of a Multi Feature Inserter node. The edge width, derived from the number of comments, becomes an edge property by means of a second Multi Feature Inserter node. And finally, the Network Viewer node produces an image of the network graph, which for the KNIME Textprocessing forum category has the shape reported in figure 33.

In this graph image (Fig. 33), you can see that Kilian Thiel is at the center of the graph and colored in blue. This means he talks to more or less everybody in this discussion group and mainly comments to other users' questions. Being Kilian the key developer of the KNIME text processing nodes, it is no surprise that he answers all questions to all users on this topic.

While the network image is exported into a data table with an Image To Table node, the average authority score and the average hub score of respectively the top 5 authority users and the top 5 hub users are also calculated.

The output data table contains the network image, the forum category, the average authority score and the top 5 authority users on which it has been calculated, the average hub score and the top 5 hub users on which it has been calculated, the sum of the edge weights, the total number of nodes and edges, the total number of objects, and the largest component of the network. All those data tables are then collected at the first output port of the Loop End (2 Ports) node and exported into the associated report as the "network mining" data set (Fig. 34).

Figure 34. Output Data Table at the upper port of the "Loop End (2 Ports)" node.



Collected results - 2:295 - Loop End (2 ports)

Row ID	forum-category	network im...	no of n...	no of e...	authori...	top 5 poster (largest compon...	hub sco...	top 5 commenter (largest component)	no of c...	network (largest comp
Row0#5	KNIME Developers		232	392	0	[baj,aborg,THINK,...]	1	[wiswedel,thor,gabriel,...]	725	greglandrum - Nodes:232 E
Row0#6	KNIME General		379	737	0	[richards99,nbrooijmans,odin,...]	1	[gabriel,thor,wiswedel,...]	1242	greglandrum - Nodes:379 E
Row0#7	KNIME Image Proces...		21	34	0	[COS_HD1,bickle,stoeter,...]	0	[christian.dietz,hornm,gabriel,...]	97	KNIME Forum - Nodes:21 E

Build the Summary Report

The report of the "KNIME Forum Text and Network Analysis" workflow shows a first page (Fig. 35) with the list of forum categories as links to their detail page and a series of details pages, one for each forum category, including some general statistics, the tag cloud, and the network graph of this discussion group. The details page for the KNIME Textprocessing forum category is shown in figure 35.

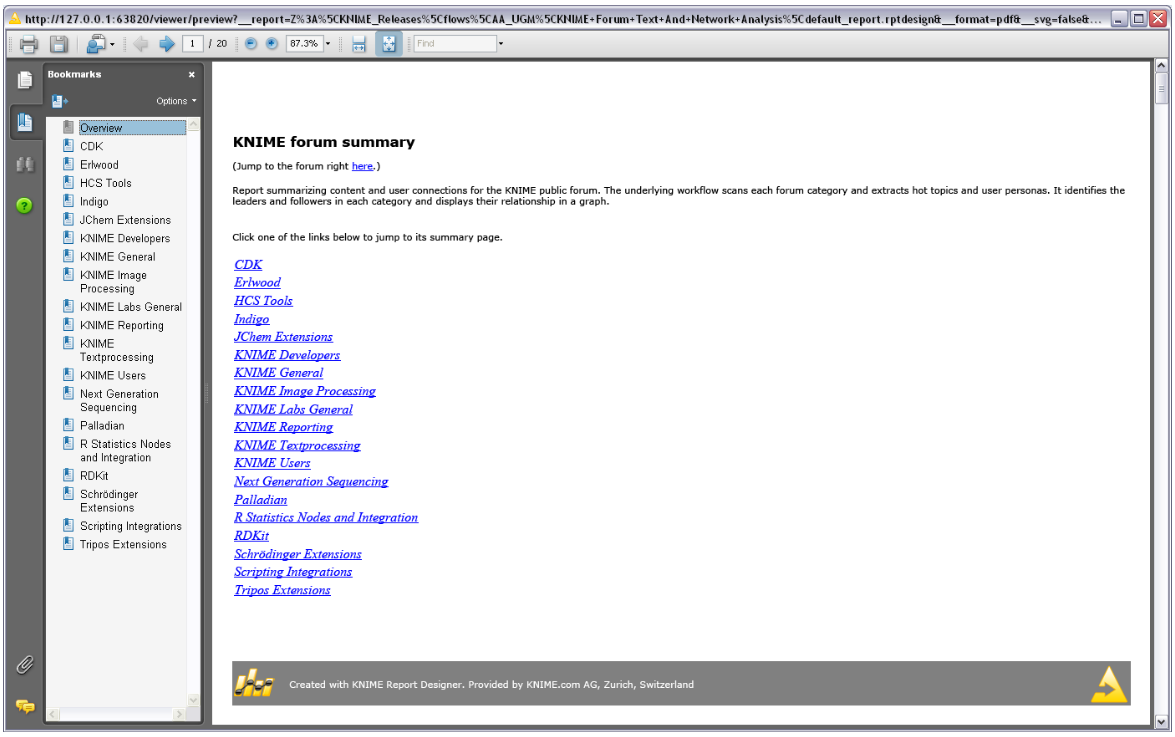
Of this report we would like to describe three interesting aspects:

- How to generate the links
- How to create dynamic text, displaying the current value of some parameters
- How to produce one report page for each one of the forum categories

Multi-page report: one page for each forum category

In this report we want to create one page for each forum category, including pictures and other category related facts. How can we build one page of the report for each category's details?

Figure 35. First page of the report with the list of forum categories. Each category is linked to the category's detail page in the same report.



KNIME forum summary

(Jump to the forum right here.)

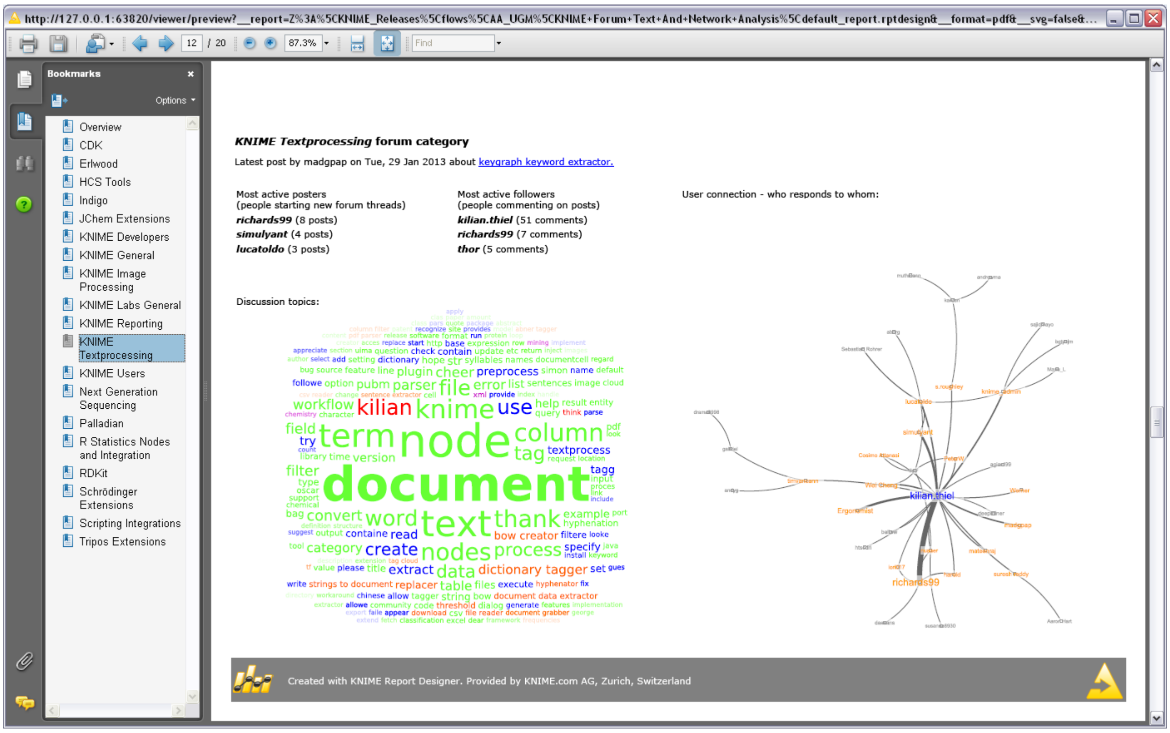
Report summarizing content and user connections for the KNIME public forum. The underlying workflow scans each forum category and extracts hot topics and user personas. It identifies the leaders and followers in each category and displays their relationship in a graph.

Click one of the links below to jump to its summary page.

CDK
Erlwood
HCS Tools
Indigo
JChem Extensions
KNIME Developers
KNIME General
KNIME Image Processing
KNIME Labs General
KNIME Reporting
KNIME Textprocessing
KNIME Users
Next Generation Sequencing
Palladian
R Statistics Nodes and Integration
RDKit
Schrödinger Extensions
Scripting Integrations
Tripos Extensions

Created with KNIME Report Designer. Provided by KNIME.com AG, Zurich, Switzerland

Figure 36. The details page of the KNIME Textprocessing category in the report of the “KNIME Forum Text and Network Analysis” workflow.



***KNIME Textprocessing* forum category**

Latest post by madgpap on Tue, 29 Jan 2013 about keygraph keyword extractor.

Most active posters
(people starting new forum threads)
richards99 (8 posts)
simulyant (4 posts)
lucatoldo (3 posts)

Most active followers
(people commenting on posts)
kilian.thiel (51 comments)
richards99 (7 comments)
thor (5 comments)

User connection - who responds to whom:

Discussion topics:

Created with KNIME Report Designer. Provided by KNIME.com AG, Zurich, Switzerland

First of all, we create a table from the category list from the “all categories” data set, with only one column binded to the “forum-category” field. The data cell of this table produces as many rows as many forum categories and the data cell of each row displays one forum category only. Thus, the data cell of this table is the place where to shape the layout of the category’s detail page.

In the data cell of the table, a grid is placed to host a few text items to display the facts, an image with the word cloud from the “text mining” data set, and an image with the network graph from the “network mining” data set, all for the table row’s forum category.

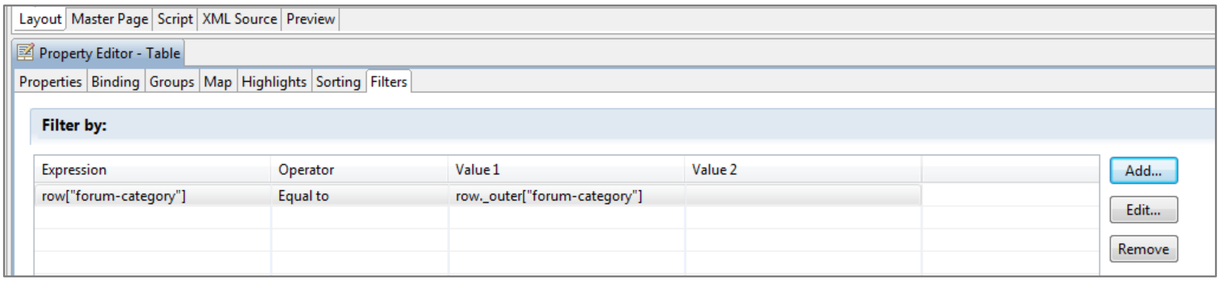
While the text items display contents from the “all categories” data set, the network graph and the word cloud display images from the “text mining” and the “network mining” data set respectively. Here we need a new table binded to the “text mining” or the “network mining” data set, with only the image for the same forum category as the one in the data row of the external table. Both internal tables consist of only one data cell hosting the image of the word cloud and the network graph respectively.

The last part to implement is to link the word cloud or the network graph with the forum category of the data cell in the external table. In order to do that, in the "Properties" panel of the internal table, in the "Filters" tab, we added a filtering condition via the Expression Builder editor to match the value of the “forum-category” field in the internal table with the value of the “forum-category” field in the external table. That is:

```
row["forum-category"] Equal to row._outer["forum-category"]
```

row._outer["forum-category"] represents the field "forum-category" of the external table. If you use the "Expression Builder" option, in the bottom panel for the category "Available Column Bindings" you should see two tables: one is the internal table and one is the container table.

Figure 37. The "Filters" tab in the "Property Editor" panel of the internal table



Filter by:

Expression	Operator	Value 1	Value 2
row["forum-category"]	Equal to	row._outer["forum-category"]	

To conclude, a Page Break is added at the end of the data cell of the external table, exactly to reserve the page to the forum category’s details.

The result of the filtering rule defined above is a page for each forum category and a variable number of pages.

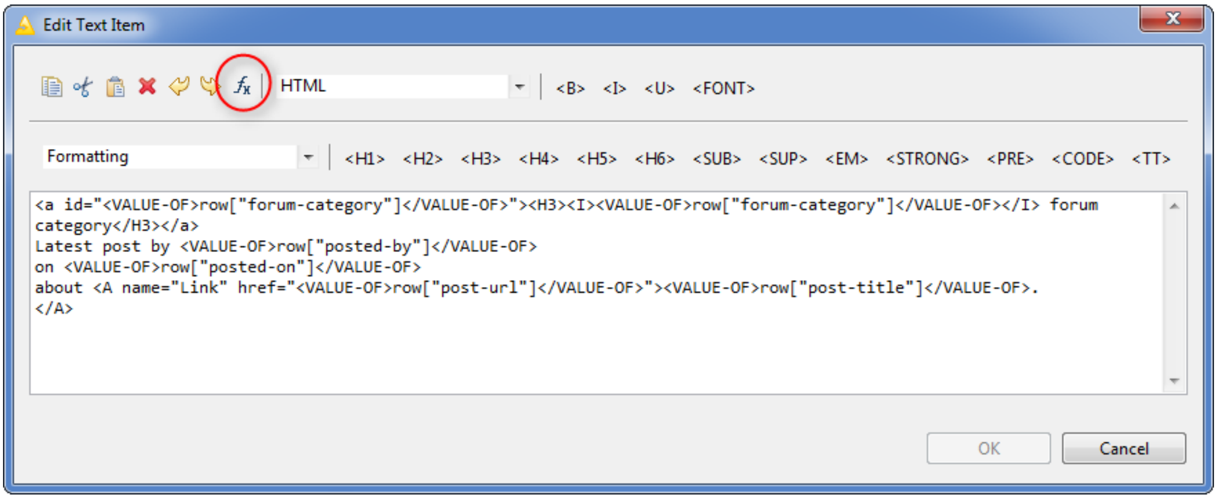
Dynamic Text

The first part of the details page is a text containing a title and something about the latest post in the forum category. All information coming from the “all categories” data set. In order to visualize these facts, we created a “Text” report item.

The “Text” report item is a very versatile tool, since it can contain static text as well as dynamic text, with many formatting options like pure text or HTML. We opted for HTML and for the “Formatting” style.

The text is written in the text editor reported in figure 38 using HTML formatting commands. In addition, in order to write a dynamic text, we relied on the “Expression Builder” option (see red circle in Fig. 38), which, mainly via the <VALUE-OF> … </VALUE-OF> item, allows to display values from the current data set.

Figure 38. The text editor for the "Text" report item.


Edit Text Item
HTML
<B> <I> <U> <FONT>
Formatting
<H1> <H2> <H3> <H4> <H5> <H6> <SUB> <SUP> <EM> <STRONG> <PRE> <CODE> <TT>
OK
Cancel

```
<a id="<VALUE-OF>row["forum-category"]</VALUE-OF>"><H3><I><VALUE-OF>row["forum-category"]</VALUE-OF></I> forum
category</H3></a>
Latest post by <VALUE-OF>row["posted-by"]</VALUE-OF>
on <VALUE-OF>row["posted-on"]</VALUE-OF>
about <A name="Link" href="<VALUE-OF>row["post-url"]</VALUE-OF>"><VALUE-OF>row["post-title"]</VALUE-OF>.
</A>
```

Links in a KNIME Report

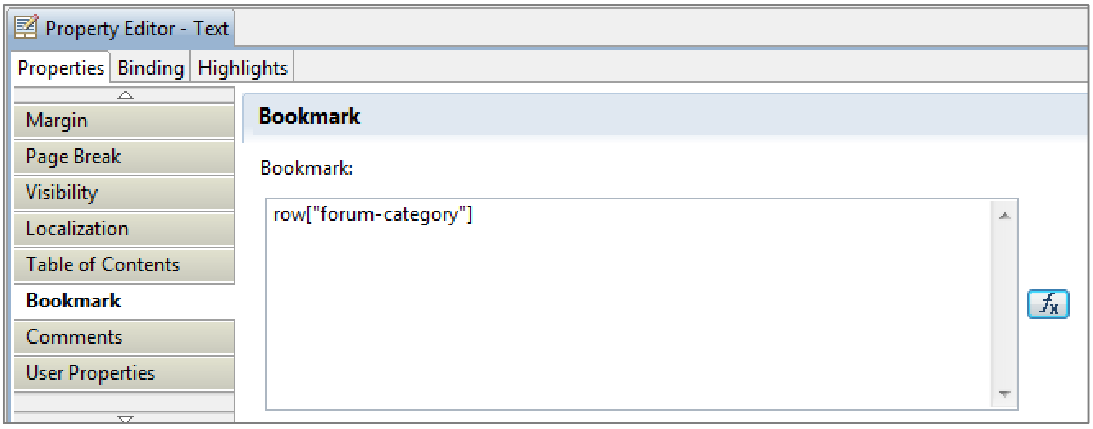
For the first page of the report we created a table based on the “all categories” data set. Of all fields, only the “forum-category” field has been kept as the only column of the table. This data cell must be linked to the start of the details page of the same forum category.

At the top of each row of the external table, where the Text report item shows the statistical facts of the forum category, a “Bookmark” property was set. As a Bookmark, *row[“forum-category”]* was selected through the “Expression Builder” editor, meaning that the bookmark is placed at the start of each row, as defined by the field “forum-category”, in the external table.

Now, going back to the table in the first page, the property “Hyperlink” for the only data cell of this table has been set to “Internal Bookmark” and then to *row[“forum-category”]* as the only available bookmarks in this report.

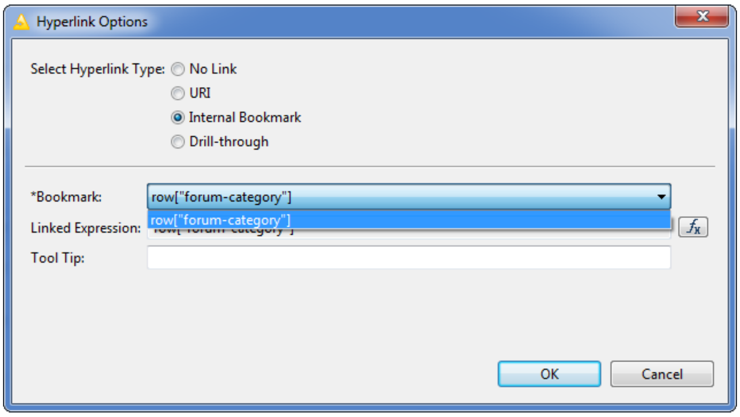
This combination of Bookmark and Hyperlink properties allows for the linking of one row in one table to the corresponding row in another table binded to the same data set.

Figure 39. The Bookmark property for the first report item in the data cell of the external table



Bookmark: row["forum-category"]

Figure 40. The Hyperlink property for the data cell of the forum category table in the first page of the report.



Hyperlink Options

Select Hyperlink Type: ○ No Link
○ URI
● Internal Bookmark
○ Drill-through

*Bookmark: row["forum-category"]

Linked Expression: row["forum-category"]

Tool Tip:

Moving into Production

We are almost at the end of this project. We have developed three workflows, one for the basic statistics, one for the topic detection, and one for the network analysis. We need now to move these three workflows into production. This means the following.

- Data needs recollecting from the web site at regular times. At other times the workflows run just to display the results. A selective procedure is needed to enable the web crawling at times and the data reading at other times.
- The switch in the data collection/reading part has to be controlled via a parameter
- The data collection/reading part is common to all workflows and might be shared from a central repository.
- Results have to be displayed on a web browser

All these requirements, in addition of course to remote execution and scheduling capabilities, require a KNIME Server for the production phase. In this section, we illustrate how each step has been implemented and the final results displayed on the Web Portal.

Selective Data Reading

Do you remember the „Web Crawler" workflow in figure 7? That small workflow collected and reorganized all data coming for a web site. That workflow was then transformed into the first metanode, named "KNIME Forum Data", in all the other three following workflows (Fig. 27). The execution of this metanode takes around 30 minutes, which is quite a long time to wait. In general we would like to avoid refreshing the data, if it is not necessary.

For this purpose, a CASE Switch block has been introduced. The "CASE Switch" node starting the block is controlled by a flow variable, with possible values 0 and 1, and it is connected to two branches. The upper branch consists of the nodes of the "Web Crawler" workflow and the bottom branch of just a Table Reader node to read the previously stored data.

The switch logic has been hidden in a metanode, named "Variable Switch" (Fig. 44).

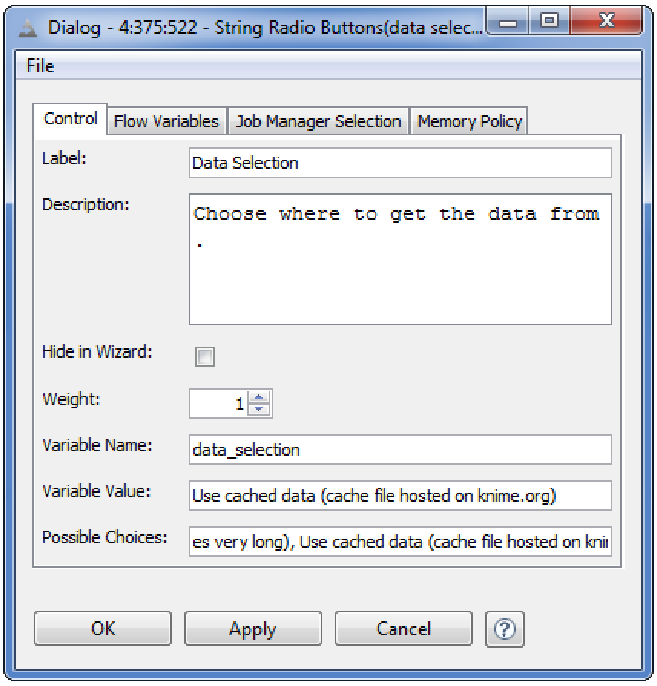
Controlling the KNIME Workflows via Quickform Nodes

The CASE switch block is controlled via a flow variable with possible values 0 and 1 only, i.e. the port index of the "CASE Switch" node. This value has to be controlled by the user. That is, when it is the time to redownload the web site content, the user must enable the upper branch of the switch block, otherwise he must enable the lower branch and just read the currently available data.

How to control flow variable values is always an issue. If other users run your workflow you do not want them to open the workflow editor and change the variable value every time, especially if the workflow has to run on the KNIME WebPortal. In this last case, i.e. when the workflow has to be run and controlled from a web browser, we need a user interface to insert the flow variable values.

Setting flow variable values and user interfaces always come together in the Quickform nodes. Quickform nodes create flow variables and set their values. In addition, when running on the KNIME WebPortal they produce a user interface to guide the user in selecting the right values. The same user interface is also visible in the configuration window of the metanode that might contain them.

Figure 42. Configuration Window of the "String Radio Buttons" Quickform node



Dialog - 4:375:522 - String Radio Buttons(data selec...

File

Control | Flow Variables | Job Manager Selection | Memory Policy

Label: Data Selection

Description: Choose where to get the data from .

Hide in Wizard: □

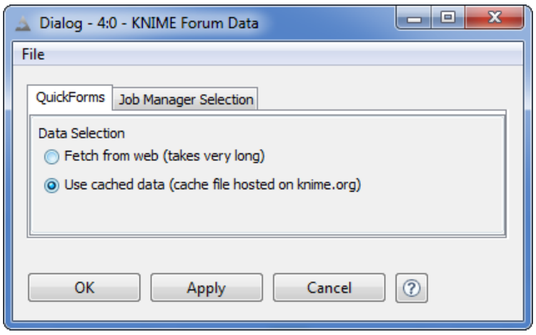
Weight: 1

Variable Name: data_selection

Variable Value: Use cached data (cache file hosted on knime.org)

Possible Choices: es very long), Use cached data (cache file hosted on knii

Figure 43. Configuration Window of the "KNIME Forum Data" metanode containing the "String Radio Buttons" Quickform node



Dialog - 4:0 - KNIME Forum Data

File

QuickForms | Job Manager Selection

Data Selection

○ Fetch from web (takes very long)

● Use cached data (cache file hosted on knime.org)

Figure 44. The "KNIME Forum Data" metanode

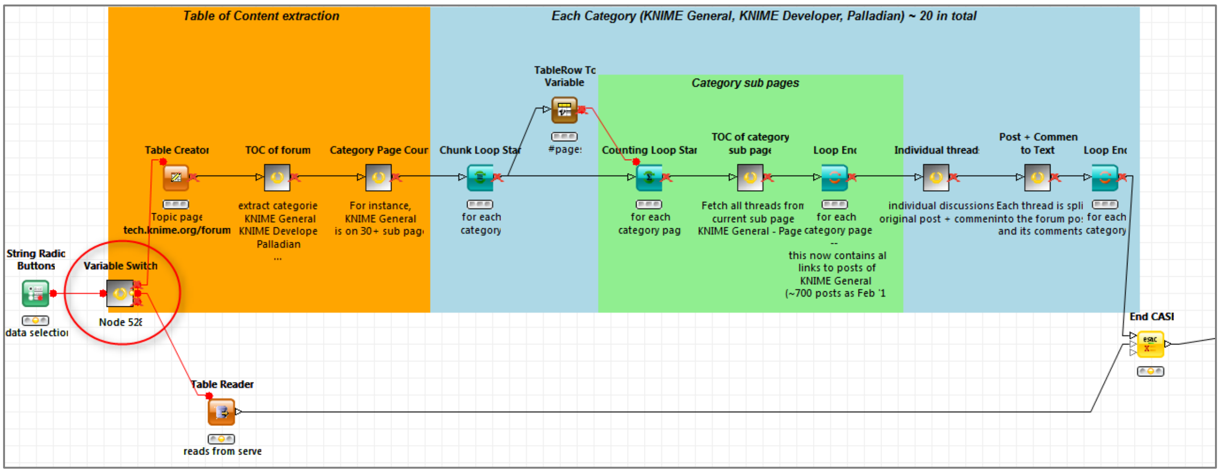


Table of Content extraction
Each Category (KNIME General, KNIME Developer, Palladian) ~ 20 in total
Category sub pages
TableRow Tc Variable
#page:
Table Creator
Topic page tech.knime.org/forum
TOC of forum
extract categorie KNIME General KNIME Develope Palladian ...
Category Page Cour
For instance, KNIME General is on 30+ sub pag
Chunk Loop Star
for each category
Counting Loop Star
for each category pag
TOC of category sub page
Fetch all threads from current sub page KNIME General - Page
Loop End
for each category page -- this now contains al links to posts of KNIME General (~700 posts as Feb '1
Individual thread
individual discussions original post + commen
Post + Commen to Text
Each thread is spli into the forum po: and its comments
Loop End
for each category
String Radio Buttons
data selectio
Variable Switch
Node 528
Table Reader
reads from serve
End CASE

To set a binary value to an integer flow variable, a String Radio Buttons Quickform node has been introduced. The String Radio Buttons node allows for a selection between two strings and outputs two flow variables: a string one containing the selected string value and an integer one containing the index associated with the selected option. The two possible choices to control the CASE Switch node are:

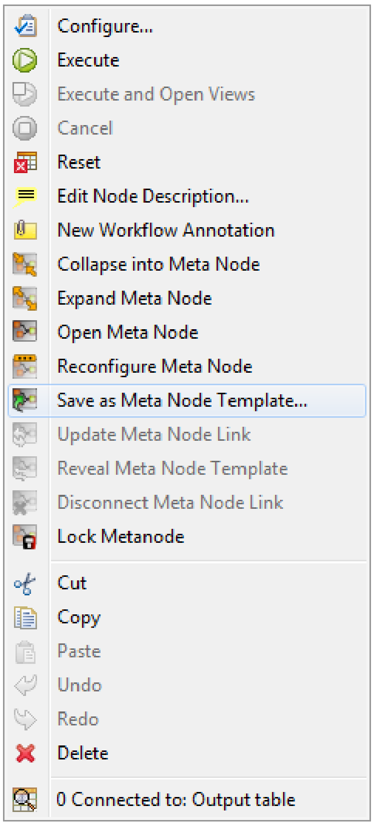
- ○ Fetch from web (takes very long)
- ○ Use cached data (cache file hosted on knime.org)

Figure 42 shows the configuration window of the String Radio Buttons node and figure 43 shows the configuration window of the "KNIME Forum Data" metanode containing the String Radio Buttons node. The final "KNIME Forum Data" metanode is shown in figure 44.

Exploiting the KNIME Server to Share Common Metanodes

All three workflows have the same metanode, described in the subsections above, as starting point. This metanode is supposed to keep performing the same task in all workflows. It would be easier to reduce this metanode to a single metanode template and to use three instantiations of the same metanode template, i.e. three linked metanodes. The maintenance of the metanode sub-workflow would then concentrate on only one spot, i.e. the metanode template, rather than on the three metanode copies.

Figure 45. Context Menu of a Meta Node



The current meta-node can be saved as a meta-node template via its context menu (Fig. 45), by right-clicking the metanode and selecting "Save as Meta Node Template".

After selecting the location on the KNIME server for the metanode template, the current metanode converts into a linked metanode. A linked metanode is just a reference to a metanode template, and, being a reference, it cannot be altered. A linked metanode is recognizable by the little green arrow on the left bottom corner of the metanode icon.

In the context menu of a linked metanode, a number of options become active.

"Update Meta Node Link", for example, looks for updates in the metanode template and refreshes the current instance.

"Reveal Meta Node Template" shows the location of the original metanode template on the KNIME server.

"Disconnect Meta Node Template" breaks the link and produces a physical copy of the metanode template.

The three workflows now share the same metanode to collect or read the data.

Running the Workflows from the Web Portal

Finally, the workflows have been moved onto the production server and ran from the KNIME Web Portal. The KNIME Web Portal compatible with KNIME 2.7 was used for this project.

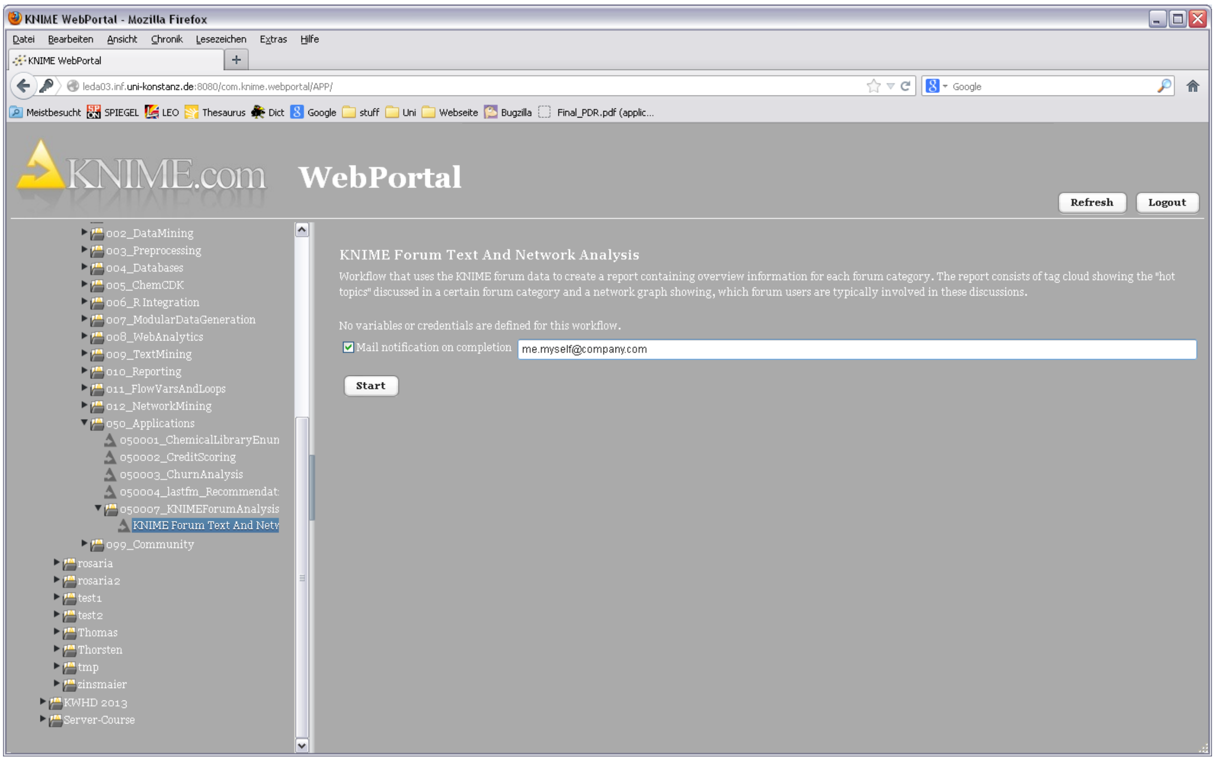
The KNIME Web Portal opens with the list of available workflows on the left, for which the user has execution permission. Selecting a workflow in the list on the left, prepares the execution in the panel on the right. In this section we are running the "KNIME Forum Text and Network Analysis" workflow as an example.

In the first frame of the step-wise execution the user can insert an email address to receive a confirmation email with the results when the execution is finished (Fig. 46). Pressing "Start" starts the step-wise execution of the selected workflow.

The first execution step takes us to the first Quickform node and shows the corresponding GUI. In the case of the "KNIME Forum Text and Network Analysis" workflow, the first Quickform node is the String Radio Buttons node with the choice between web crawling or reading previously stored data. The GUI of this Quickform node, i.e. the two radio buttons, is reported in figure 47.

Pressing "Next" takes us to the next Quickform node, if any, or to the final report (Fig. 48).

Figure 46. First Frame of the Workflow step-wise Execution on the Web Portal


KNIME WebPortal - Mozilla Firefox
Datei Bearbeiten Ansicht Chronik Lesezeichen Extras Hilfe
KNIME WebPortal
leda03.inf.uni-konstanz.de:8080/com.knime.webportal/APP/
Google
Meistbesucht SPIEGEL LEO Thesaurus Dict Google stuff Uni Webseite Bugzilla Final_PDR.pdf (applic...
KNIME.com
WebPortal
Refresh
Logout
002_DataMining
003_Preprocessing
004_Databases
005_ChemCDK
006_R Integration
007_ModularDataGeneration
008_WebAnalytics
009_TextMining
010_Reporting
011_FlowVarsAndLoops
012_NetworkMining
050_Applications
050001_ChemicalLibraryEnun
050002_CreditScoring
050003_ChurnAnalysis
050004_lastfm_Recommendat
050007_KNIMEForumAnalysis
KNIME Forum Text And Netv
099_Community
rosaria
rosaria2
test1
test2
Thomas
Thorsten
tmp
zinsmaier
KWHD 2013
Server-Course
Start

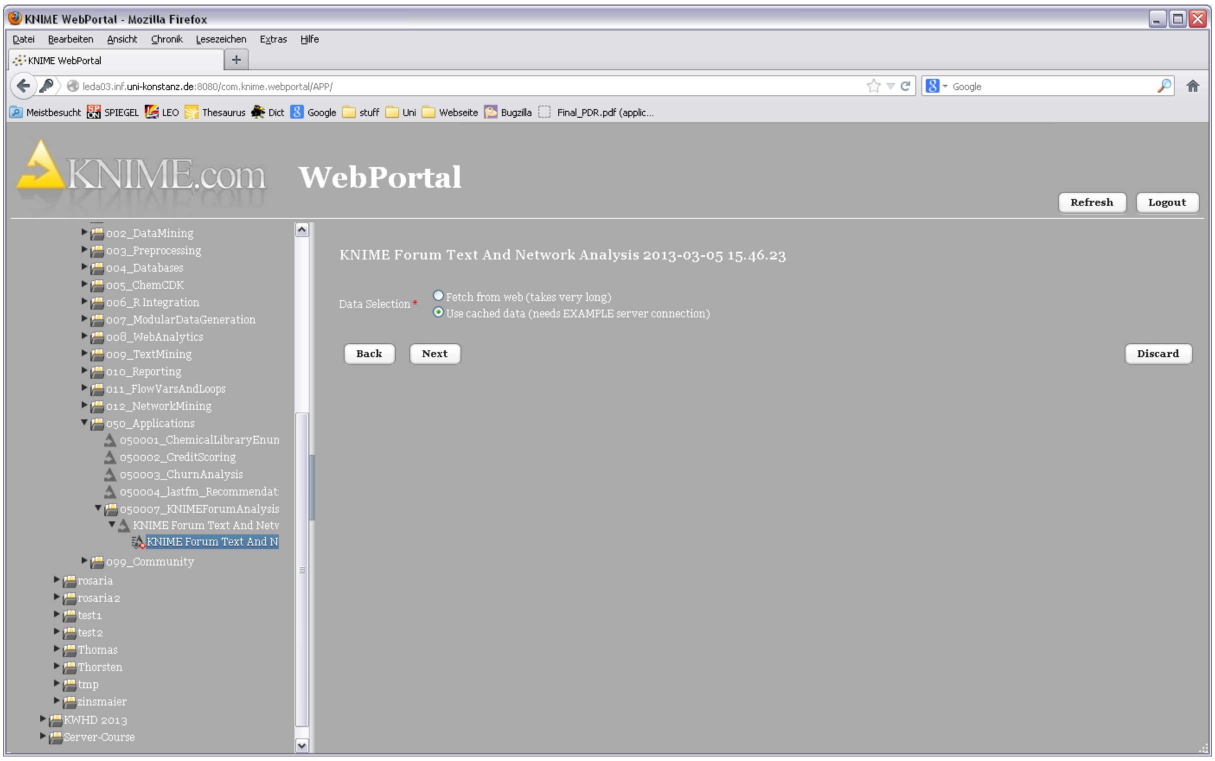
KNIME Forum Text And Network Analysis

Workflow that uses the KNIME forum data to create a report containing overview information for each forum category. The report consists of tag cloud showing the "hot topics" discussed in a certain forum category and a network graph showing, which forum users are typically involved in these discussions.

No variables or credentials are defined for this workflow.

☑ Mail notification on completion me.myself@company.com

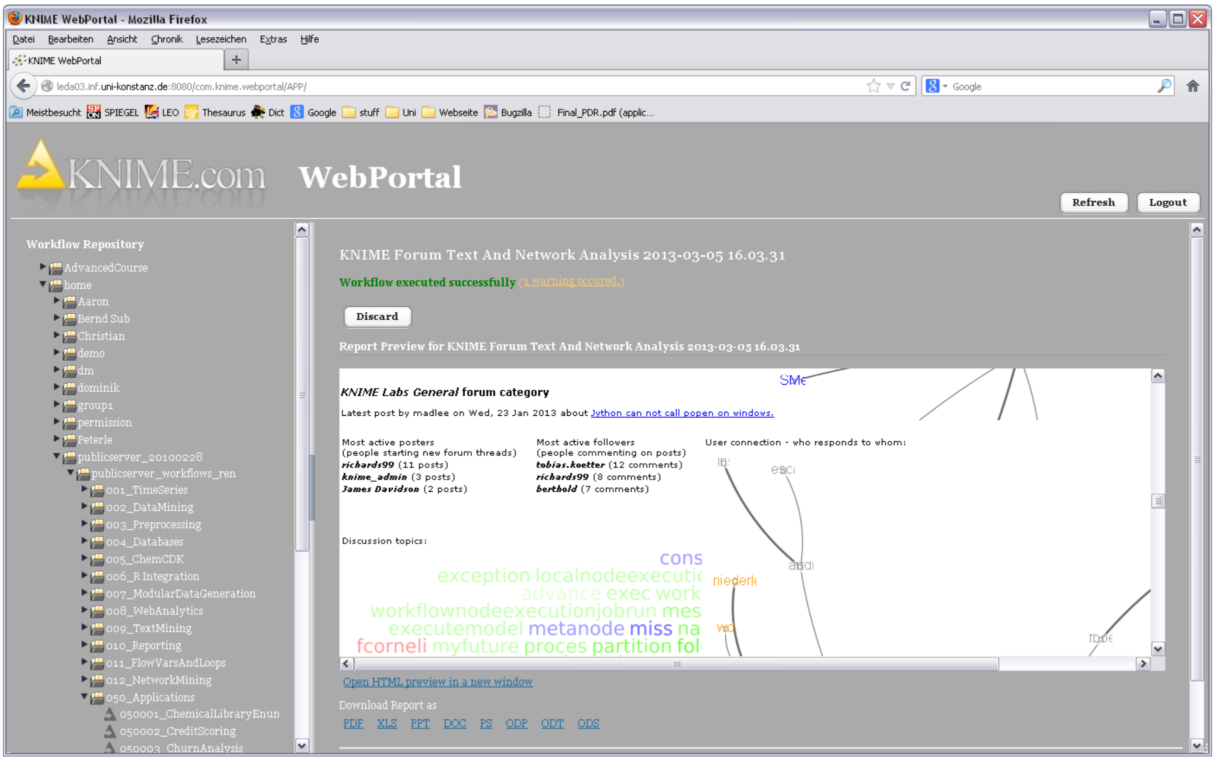
Figure 47. Second Frame for the workflow step-wise execution on the Web Portal, presenting the data input choice via two string radio buttons



KNIME Forum Text And Network Analysis 2013-03-05 15.46.23

Data Selection *
○ Fetch from web (takes very long)
● Use cached data (needs EXAMPLE server connection)

Figure 48. Final Report displayed on the Web Portal



KNIME Forum Text And Network Analysis 2013-03-05 16.03.31

Workflow executed successfully (1 warning occured.)

Report Preview for KNIME Forum Text And Network Analysis 2013-03-05 16.03.31

KNIME Labs General forum category

Latest post by madlee on Wed, 23 Jan 2013 about Jython can not call popen on windows.

Most active posters
(people starting new forum threads)
richards99 (11 posts)
knime_admin (3 posts)
James Davidson (2 posts)

Most active followers
(people commenting on posts)
tobias.koetter (12 comments)
richards99 (8 comments)
berthold (7 comments)

User connection - who responds to whom:

Discussion topics:

Open HTML preview in a new window

Download Report as
PDF XLS PPT DOC PS ODP ODT ODS

Conclusions

This whitepaper has shown how to implement the different steps to collect data from a web forum (or a web site) and extract insights from it. Insights mean the evolution of the basic statistics, the observation of the discussed topics, and the recognition of a central figure for a given topic.

We have shown how to connect to a web site and download its content, how to display the growth of the forum over the years by monitoring its basic statistics, how to use text mining to find out the discussed topics, and how to discover the experts in a given topic. Finally, we have shown how to use the KNIME Server as a productive environment for the final workflows.

All functionalities required for this kind of analysis are available within the KNIME open source data analytics platform. This, combined with the ease of use and the reliability of the platform, has made our work extremely easy, which encouraged us to share the experience with other KNIME users through this whitepaper.

All data and workflows are downloadable from the KNIME public server under 050_Applications/050007_ForumAnalysis, while the KNIME open source platform is available on the KNIME site at www.knime.com.

References

[1] M.F. Porter, 1980, An algorithm for suffix stripping, *Program*, **14**(3) pp 130−137.

[2] Y. Osahwa, N. Benson, M. Yachida, "KeyGraph: Automatic Indexing by Co-occurrence Graph based on Building Connstruction Metaphor", *ADL '98: Proceedings of the Advances in Digital Libraries Conference, page 12. Washington, DC, USA, IEEE Computer Society*, (*1998*)

[3] R. Kuhlen, W. Semar, D. Strauch, Grundlagen der praktischen Information und Dokumentation, *Walter de Gruyter GmbH & Co.KG*

[4] M.A. Beauchamp, 1965, “An improved index of centrality”, Behavioral Science, pp 161-163.

[5] K. Thiel, T. Kötter, M. Berthold, R. Silipo, P. Winters, “Usable Customer Intelligence from Social Media Data: Network Analytics meets Text Mining”, KNIME whitepaper available at: http://www.knime.org/files/knime_social_media_white_paper.pdf

[6] J. Kleinberg, 1999, Authorative sources in a hyperlinked environment, *Journal of the ACM, Col. 36 Nr. 5, pp 604-632*

[7] JUNG the Java Universal Network/Graph Framework available at: http://jung.sourceforge.net/