

Problem Statement

NC4 provides their customers with breaking news alerts as incidents occur by monitoring and analyzing over 2,000 news sources.

As NC4's coverage expands, they must view an increasing amount of irrelevant news. Continuing their expansion requires either hiring additional analysts or coming up with a way to remove irrelevant news items before the analysts see them.

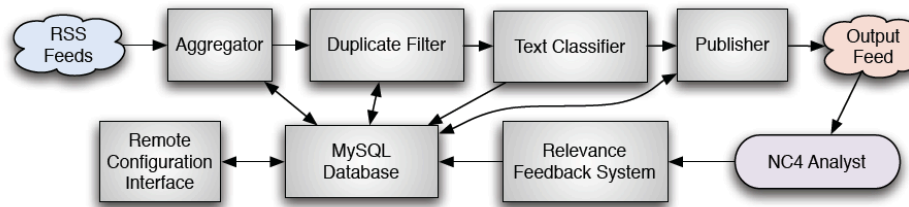
The goal of our project is automatically filter NC4's RSS news.



What is RSS?

RSS, which stands for Really Simple Syndication, is an online publishing format that provides users with an efficient means to see the most recent updates to a website. The website's author publishes a feed containing a list of news items. Each item usually contains a title, a sentence or paragraph describing its content, and a link to the full article.

Solution Overview



Our solution is depicted above. Arrows represent the flow of information.

Main Program

The main program contains an aggregator that monitors a large number of RSS feeds for news updates, a classifier that assigns relevance scores to news items, and a publisher that publishes items to an output feed if their scores are above a threshold. NC4 analysts can view our output feeds with their current feed reader.

Remote Configuration Interface

The remote configuration interface is a web page that allows users to configure the application. It lets one specify source feeds, output feed properties, keywords for use in the keyword filter, and other parameters.

Text Classification

The text classifier decides if a news article is relevant enough to show to NC4's analysts. We explored several different varieties of classifiers to see which was the most effective.

Bayesian Classifiers

Bayesian text classifiers use the probabilities of each word in a piece of text occurring in either a relevant or irrelevant example to generate a composite score. We tested the open-source Bayesian classifier Reverend and implemented our own custom Bayesian classifier.

Relevance Feedback System

We annotate news items with a link that enables NC4 analysts to give feedback on the item's relevance. This data may later be used by NC4 to further train the classifier or to track analyst performance. NC4 is particularly excited about tracking analyst performance, since they currently have no way to do so for RSS.

Duplicate Filter

Analysts commonly receive many copies of the same news item from different sources. Our duplicate filter discards 5-10% of incoming news items. Only about 1/60th of the items it excludes are not actually duplicates. The duplicate filter is one of the most significant improvements we have made to NC4's previous system.

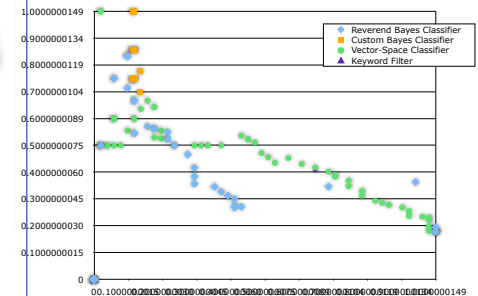
Keyword Filtering

The keyword filter assigns a score to a piece of text by summing scores associated with user specified keywords.

Vector-Space Classifier

Our vector-space classifier works by representing text as a point in a vector-space. We use a neural net to learn a hyper-plane boundary that separates points corresponding to relevant and irrelevant items. In order to classify a new piece of text, we compare the corresponding point to the hyper-plane that separates our training examples.

Classification Results



Completeness is the fraction of the total relevant items that appear in the output feed. Relevance is the fraction of the items in the output feed that are relevant. The above graph shows the relevance vs. completeness curves generated by varying the score threshold for each of our classifiers. We can produce an output feed that has a low completeness but high relevance, or vice-versa.

Conclusion

Classifying news items is extremely difficult; we cannot do it perfectly. However, our text classifiers are competitive with the state of the art. NC4 tested our system and said they noticed an immediate improvement in the usefulness of the RSS items they see.

Acknowledgements

NC4 Liaisons: Karl Kotalik, Juan Matute,

Rajesh Goswami

Faculty Advisor: Prof. Chris Stone

Thanks: Prof. Robert Keller, Prof. Christine Alvarado,

Joyce Greene, Barbara Schade