READING:

Retriving nEws Around the worlD for dIscovery and Knowledge Mining

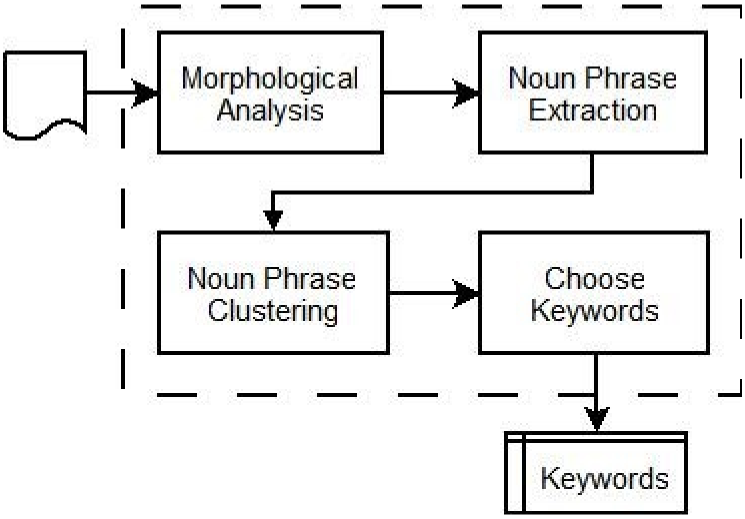
News Classifier Algorithms Overview

1. Introduction

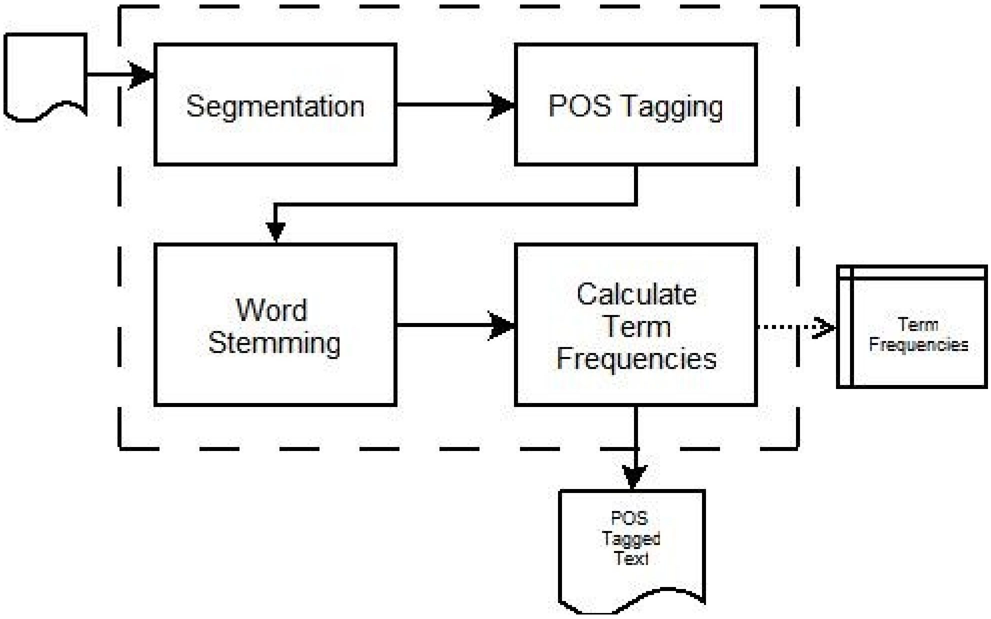
Information about the KANT API can be found in “KantLib.chm.” This paper will give an overview of the algorithms used in the project.

2. Keyword Extraction

The keyword extraction algorithm uses linguistic information to aid in determining the most important phrases in a document. We use noun phrases for keywords as they contain the most information in the document. The algorithm is split into 4 submodules, as shown in the figure below.

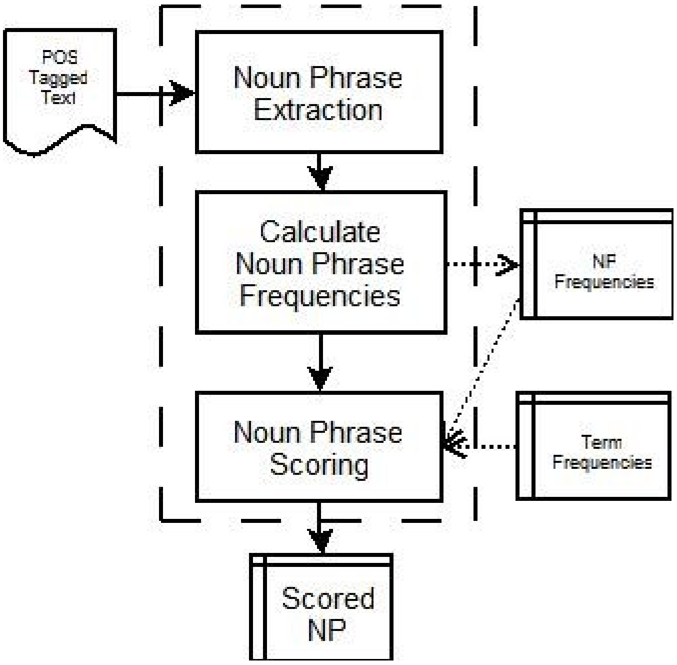


Morphological analysis takes care of word segmentation, part-of-speech tagging and word stemming. An overview can be seen in the figure below.



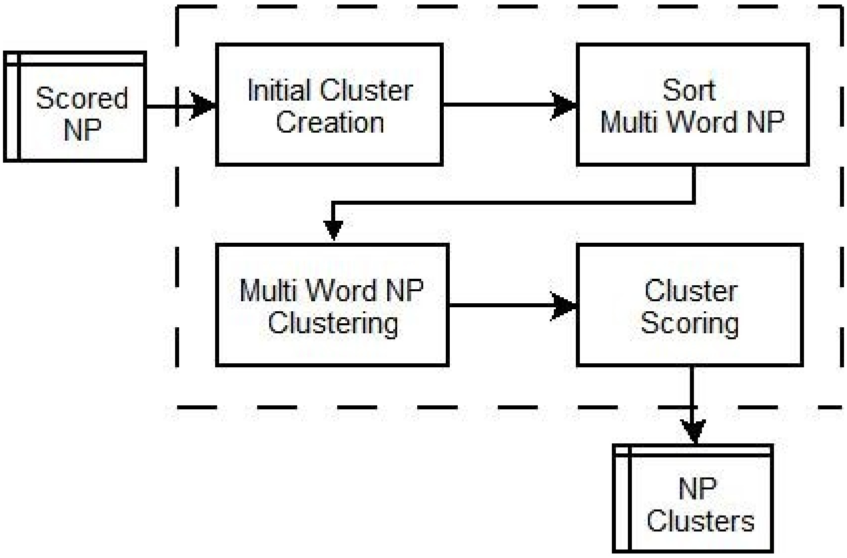
In English, word segmentation is not necessary, but sentence segmentation is. Part-of-speech tagging assigns parts-of-speech tags (noun, verb, adjective, etc.) to words and we are currently using a self-implemented version of Brill's tagger. Word stemming transforms a word into its stem, i.e. running to run, and we are using Porter's stemmer. Calculating term frequencies means to determine the unigram frequency of each word and it is used for scoring keywords.

The next phase is noun phrase extraction and scoring. An overview can be seen in the figure below.



Noun phrase extraction is done using a simple noun phrase grammar. After extracting noun phrases their frequencies are obtained. These frequencies and the frequencies of the individual words are combined to give each noun phrase a score.

Next, the noun phrases are clustered. Clustering attempts to group the noun phrases that have similar semantic meaning together to allow for more diverse keywords to be chosen. An overview can be seen in the figure below.



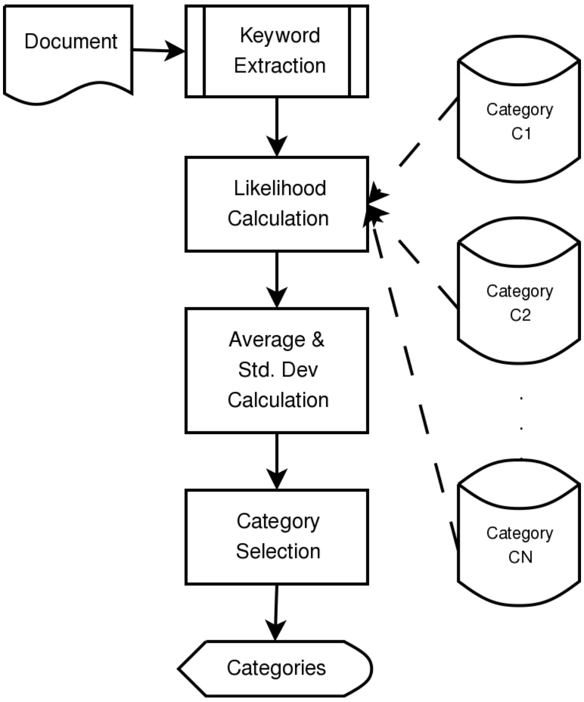
The clustering process first assigns all single word noun phrases to their own cluster. It then sorts the multi word noun phrases. The multi word noun phrases are then added to clusters in which they share an n-gram with. If a multi word noun phrase cannot be assigned to a cluster then it creates its own. After the clustering process is complete each cluster is assigned a score equal the average noun phrase score of the noun phrases in the cluster.

The final step is to choose the keywords. The clusters are sorted by cluster score and the top clusters are used to determine the keywords. From each cluster the keyword with the highest score that has not been assigned as a keyword yet is assigned as a keyword.

3. Category Classification and Topic Discovery and Classification

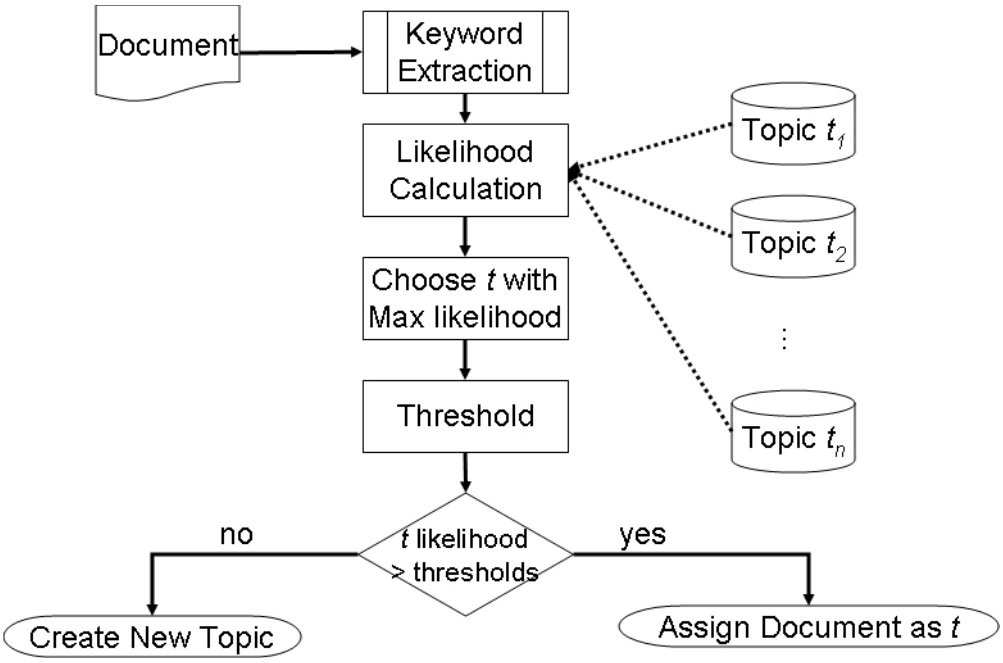
Both category classification and topic discovery and classification are based upon the previously introduced keyword extraction algorithm. They are similar in design in that they represent categories/topics as keyword vectors and use them to determine a similarity or likelihood of a news article to be in the category/topic.

An overview of category classification can seen in the figure below. It takes a document and extracts keywords from it to use a representation of the document. The likelihood that the document is in each of the categories is then calculated using the keywords from the document and the keywords in the category that were gathered through training.



After all the likelihoods are calculated the mean and standard deviation are calculated. Using this information categories are assigned to the document. All categories with a likelihood of more than one standard deviation from the mean are assigned to the document.

Topic discovery and classification works in a similar manner to category classification. An overview can be seen in the figure below. First, keywords are extracted from the given document. Then the likelihood that the document is of each topic is determined using the cosine similarity.

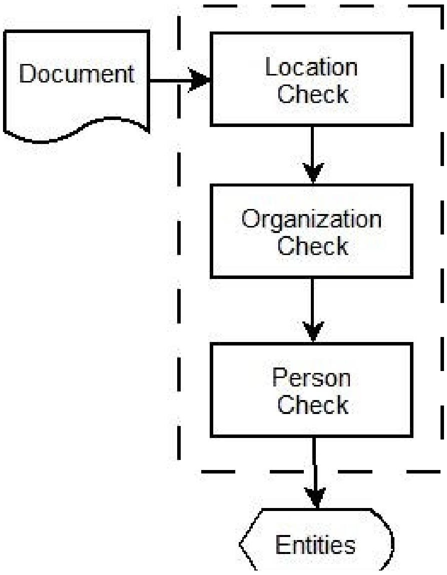


The topic with the highest likelihood is then assigned as the conditional topic for the article. If the likelihood of the conditional topic is greater than some dynamic thresholds then the topic is officially assigned to the document. Otherwise a new topic is created.

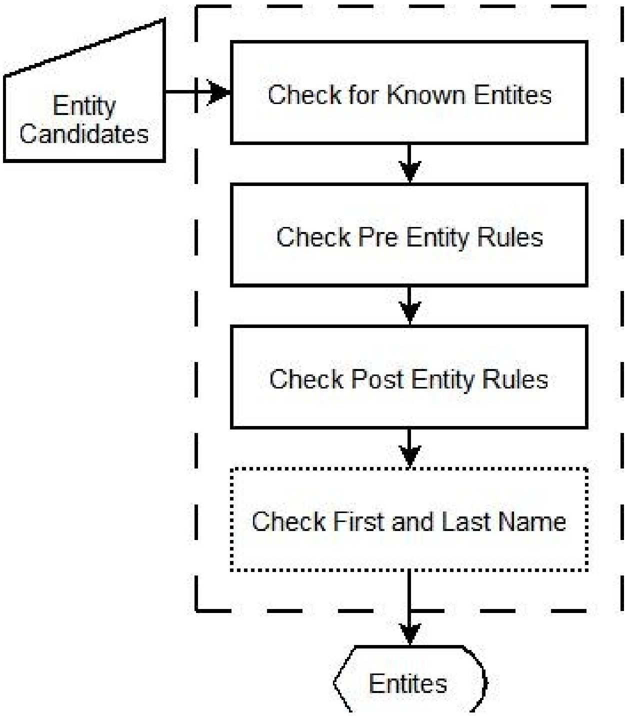
In addition to categories and topics, world regions are also classified. This uses a dictionary containing the names of the countries and their region.

4. Named Entity Recognition

Currently, we limit our named entity recognition to people, locations, and organizations as they are the most abundant in news. We use a standard dictionary and rule-based approach. An overview of the process can be seen in the figure below.

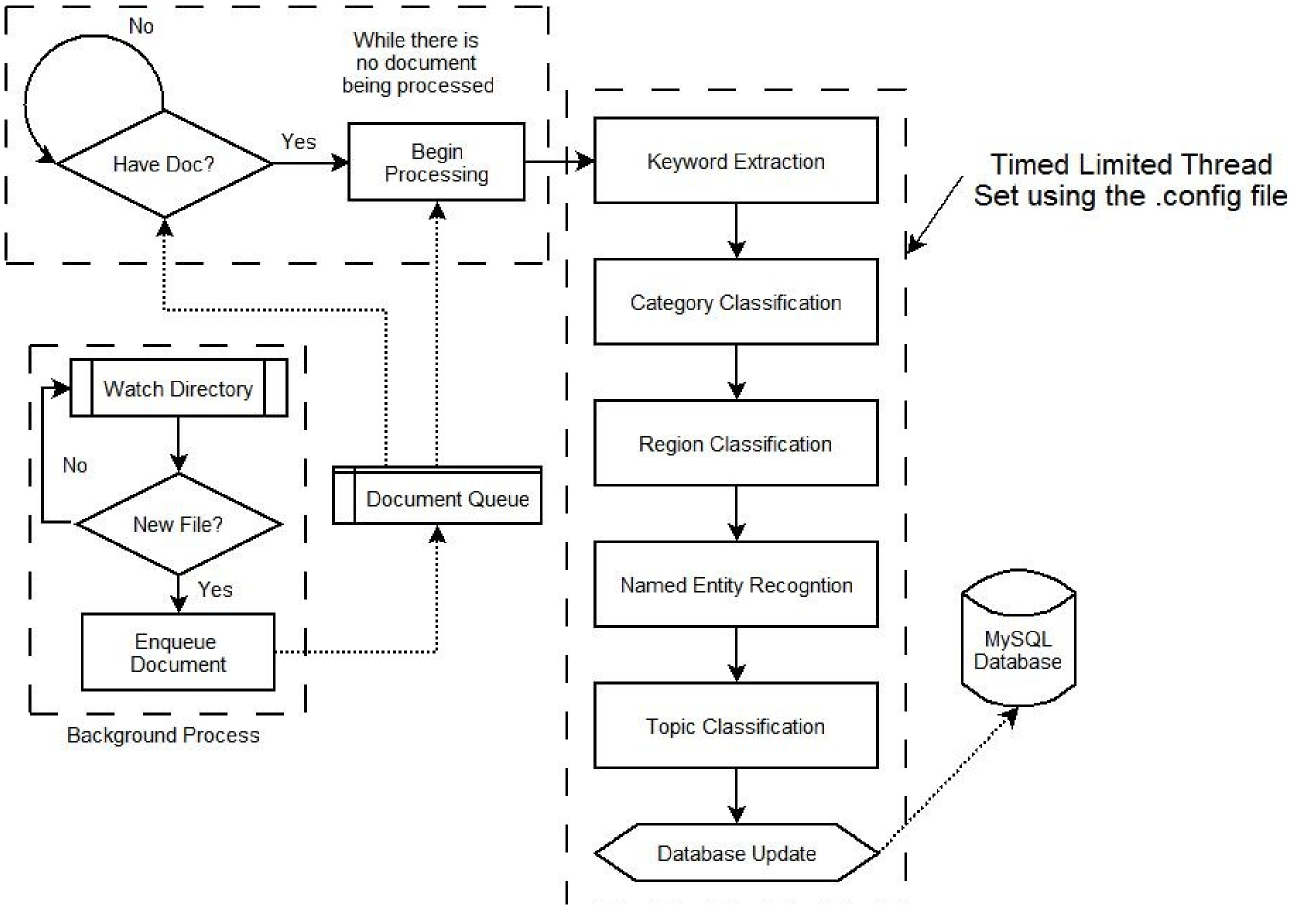


In the figure below, we show how each check is done. Entity candidates are extracted using rules and are specific to each type of entity. This means that when doing a location check the candidate entities will be different than when doing a person check.



Checking for known entities involves looking in a dictionary for the existence of the entity in question. If the entity is not found in the dictionary then the next step is to look at rules for prefixes (Pre Entity Rules). These rules include such things as Mr., Mrs., The nation of, etc. If the entity is still not found then we look for suffixes (Post Entity Rules). These rules include such thins as Co., Inc., corporate, etc. Finally, if we are looking for people and they all fail we look in a dictionary of known first and last names and see there is a match.

5. News Classifier Program Overview



Relevant References

David B. Bracewell, Fuji Ren, and Shingo Kuroiwa. Category classification and topic discovery of news articles. In *Proceedings of Information-MFCSIT 2006*, pages 345-348, 2006.

David B. Bracewell, Fuji Ren, and Shingo Kuroiwa. Multilingual single document keyword extraction for information retrieval. In *Proceedings of the 2005 IEEE International Conference on Natural Language Processing and Knowledge Engineering*, pages 517-522, Wuhan, China, November 2005.

**Document preparer David B. Bracewell