Abstract

In this paper I discuss a method of content presentation through conversion of text to an image representation. A chosen medium, for example a book or collection of news articles, are converted into a collection of images representing that piece in collage form. Through a combination of Natural Language Processing and ubiquitous tools such as Google and Yahoo! image search, large blocks of text can be artistically represented in a style useful to users.

1 Motivation

An image is worth a thousand words, so the proverb goes. Anyone who has had to sift through headers in a newsgroup, or reads the newspaper in the morning knows that finding even the most salient ideas in text can be a difficult process. Users often have trouble expressing their posts in brief. Newspaper headlines are designed to catch your eye, but even with modern day headlines on the Internet, sifting through text becomes monotonous. Book jackets are designed to sell copies. Text just doesn’t pop out at you the way images do. It would be nice to browse these media by image, but leaving the chosen image of expression may tend to fall into the same dilemmas described above. Perhaps more appropriate images can be chosen by their textual content.

2 Method

The software must be provided with textual data in one of a variety of forms that will be discussed later. In large blocks of text, salient data may be extracted by means of Natural Language Processing
tools. In other media with smaller text samples, it may be more desirable to use less elaborate methods. The data extracted is best limited to at most three words except in certain cases discussed later. This set of words forms the basis for queries on an image search service. A random image will then be chosen from each query result (with a relevance threshold). The collection of images will then be re-sized and resorted among various methods and displayed in collage form.

3 Media-Specific Design Considerations

3.1 Forums, Newsgroups, and Usenet

Newsgroups and forums tend to be different from other forms I consider because most of the content is ad-hoc and user-generated. For this reason, the natural language processing I use in other media tends not to give us anything worthwhile for their trouble. The following are considerations that may be taken with this medium.

Queries can be chosen from several methods:

- a random word chosen from the subject
- a combination of words parsed from the first post based on image availability

It is expected that these options will generate a collage that represents the varied discussions that often occur in groups.

Distinct thread subjects will be sorted in two dimensions, one axis being most recent post date, the other being average message size. The images will then be re-sized based on the thread depth. The result will be similar to a tag cloud, with images instead of key words.

Newsgroups in most forms these days tend to be persistent but dynamic. I offer the following solution. Z-order will be determined by the date of the last message sent in the thread. This will result in older images being covered. An interesting interface option might be to have a time-line slider at the bottom of the visualization that allows users to view the visualization as it would look at different points in time. Even play back a newsgroup like a time-lapse photograph.

3.2 Books and Novels

An image representation of a fiction novel would benefit from Named Entity Recognition. NER is a set of algorithms in Natural Language processing that attempt to identify, classify, and extract parts of speech
from blocks of text. Extracting nouns and verbs from books and weighting them based on number of occurrence is probably a good start to finding useful image queries.

The algorithm would want to avoid personal pronouns since these would most likely be names of fictional characters. Non-fiction books, however, like news articles benefit greatly from detection of pronouns, since a lot of relevant data about the content of non-fiction books tend to be in these words.

### 3.3 News Articles

News articles have the best likelihood of generating queries that result in appropriate and salient images. Also, because they so often contain information about public figures, companies, world locations, and other very specific and relevant data, news articles benefit greatly from Named Entity Recognition. Modern Natural Language Processing have fantastic libraries and tools for performing Named Entity Recognition algorithms on large blocks of text, and can easily extract names of actors, locations of events, and other data by analyzing parts of speech.

The visualization of news articles may also benefit from the z-order and playback features discussed in the above newsgroup considerations.

### 4 Observations

It’s extremely likely that not all of the pictures will match well with the topic of the text. Some will be dead-on, and some will be hopelessly off-topic. This is not all bad. If everyone chose to read exactly what they were looking for, how would they be exposed to new ideas? Sometimes a little serendipity is exactly what we need to discover something extraordinary.