An Agent for Categorizing and Geolocating News Articles

Žiga Mahkovec
Faculty of Computer and Information Science
Trzaska cesta 25
SI-1001 Ljubljana
Slovenia
ziga.mahkovec@klika.si

Keywords: Text categorization, support vector machine, RCV1, geolocation, GIS, SVG

Received: June 1, 2004

We present a software agent for categorizing and geolocating news articles. The articles are retrieved from different on-line news sources, such as Google News, Reuters and BBC News. They are parsed, categorized based on crime threat, geographically located and rendered in an SVG widget.
The agent is implemented in Java, using the Scalable Vector Graphics markup language to render the user interface. Text categorization is performed using the Support Vector Machine (SVM) method, with test data from the Reuters RCV1 corpus. The GEONet Names Server and Digital Chart of the World databases are used to geolocate the news articles.

Povzetek: Članek opisuje inteligentnega agenta za lokalizacijo novic.

1 Introduction

The U.S. Department of Homeland Security maintains a special Homeland Security Advisory System, with a 5-stage threat level ranging from green (”Low level of terrorist attacks”) to red (”Severe risk of terrorist attacks”). There are few similar systems available for other countries.

Using on-line news sources, text categorization and geographical localization we were able to develop an agent capable of presenting such information for the entire world. News articles are categorized as ”good” and ”bad” — specifically, to news related to terror, war and violence and other news. They are also parsed to retrieve the exact location of the article’s subject. Using various worldwide news sources and applying categorization and location enables the visualization of the threat level for the entire world.

Text categorization is performed using Support Vector Machines (SVM [3]). This method has proved to be successful in solving several text categorization problems, since it avoids over-fitting when presented with a large number of attributes. For learning and classification we used the SVM-Light implementation [4].

The training set consisted of the Reuters Corpus Volume 1 (RCV1), as modified by David Lewis et al. [5]. The data set consists of over 800,000 news articles, published by Reuters between the years 1996 and 1997.

The *GEOnet Names Server* [6] and *Geographic Names Information System* [7] databases were used to locate the news articles. The articles were then rendered within an SVG widget containing a satellite image of the world and vector political boundaries.

2 News sources

The following news sources were used:

- Google News: `http://news.google.com`
- Reuters: `http://reuters.com`
- BBC News: `http://news.bbc.co.uk`

2.1 Google News

Google News is an aggregator of more than 4,500 news sources from all over the world. The service provides automatic news grouping, pulling together related headlines. The number of related news articles is a good indicator of the importance of a news article.

Google News does not provide RSS feeds [8]. Content retrieval is therefore based on HTML parsing, using the `java.util.regex` package for regular expressions. The parsing engine requires periodical testing, since the format of the Google News articles may change in the future.

Geolocation is performed by parsing the headlines of all related news articles. Most often, the headlines will include the location header, e.g. ”BASRA, Iraq –”. By searching for the most frequent geographical name in the headlines, the article can be accurately located.

Although there are several localized editions available, only the U.S. edition was used for news retrieval.

2.2 Reuters

Reuters, being the largest news agency, publishes some 11.000 stories daily. Its website offers news in 13 differ-

ent categories. RSS feeds are provided, greatly simplifying content retrieval, since the XML format is structured and easy to parse. Figure 1 shows an example of an RSS feed item.

```
<title>
  Blast Near U.S. Compound in Iraq's Basra Kills 2
</title>
<guid isPermaLink="false">6210014</guid>
<link>
  http://www.reuters.com/newsArticle.jhtml?storyID=6210014
</link>
<pubDate>Sat, 11 Sep 2004 13:55:26 GMT</pubDate>
<description>
  BASRA, Iraq (Reuters) — A car bomb exploded near the U.S.
  embassy office in the southern Iraqi city of Basra on
  Saturday, killing two people and wounding three, but no
  Americans were injured, officials and witnesses said.
</description>
</item>
```

Figure 1: An example of a Reuters RSS feed

The Reuters article headers consistently include the location header (e.g. ”BASRA, Iraq”)

2.3 BBC News

BBC News also provides RSS feeds. The articles are already partly geographically located (the geographical categories include Africa, Americas, Asia-Pacific, Europe, Middle East, South Asia and UK).

3 Text categorization

The parsed news articles are represented as a Java class containing the following attributes:

- title
- headline
- full text
- news category
- source URL

The full text is used to categorize the news articles.

3.1 Test collection

The Reuters Corpus Volume 1 (RCV1) is an archive of over 800,000 categorized news articles. Lewis et al. [5] used it to produce the RCV1-v2 corpus, containing some corrections. They also provide vectors for training with SVM classifiers.

The documents are available in XML format. A sample document is shown in figure 2.

The RCV1 documents are coded into three category sets: industry, topics and regions. To separate the ”bad” news from the rest, we used specific topic categories:

- GCRIM: crime, law enforcement
- GVIO: war, civil war

```
<?xml version="1.0" encoding="iso-8859-1" ?>
<newsitem itemid="2440" id="root" date="1996-08-20"
  xml:lang="en">
<DOCNO> RC-2440 </DOCNO>
<title>
  SINGAPORE: Philippines wary of terrorist threats to
  APEC meet.
</title>
<headline>
  Philippines wary of terrorist threats to
  APEC meet.
</headline>
<dateline>SINGAPORE 1996-08-20</dateline>
<text>
<p>The Philippines' top military officer said on Tuesday
Manila was exchanging intelligence information with the
United States and other countries on potential terrorist
threats to the APEC summit later this year.</p>
...
</text>
<copyright>(c) Reuters Limited 1996</copyright>
<metadata>
<codes class="bip:countries:1.0">
  <code code="PHLNS"/>
  <code code="SINGP"/>
  <code code="USA"/>
</codes>
<codes class="bip:topics:1.0">
  <code code="GCAT"/>
  <code code="GVIO"/>
</codes>
<dc element="dc.date.created" value="1996-08-20"/>
<dc element="dc.publisher" value="Reuters Holdings Plc"/>
</metadata>
</newsitem>
```

Figure 2: An example of an RCV1 document

3.2 SVM categorization

The *LYRL2004 split* of the RCV1-v2 corpus contains 23,149 documents. 2,087 of these were labeled as ”bad” using the above criterion. This split was then used to train the SVM-Light classifier.

The news articles were preprocessed using the same technique as the one described in [5]. Stop words were removed; stemming was performed using the Porter stemmer. The $TF \times idf$ weights of the terms in the SVM vectors were computed as follows:

$$w_d(t) = (1 + \log_e n(t,d)) \times \log_e(|\mathcal{D}|/n(t)),$$

where $n(t)$ is the number of documents containing the term t; $n(t,d)$ is the number of occurrences of term t in document d, $|\mathcal{D}|$ is the number of documents used in computing the idf weights.

The feature vectors were also cosine normalized:

$$w_d^{'}(t) = \frac{w_d(t)}{\sqrt{\sum_u w_d(u) \times w_d(u)}}$$

The agent would then run the SVN-Light classifier for each new news article. The articles categorized as ”bad” are specifically marked on the map in the user interface, thus identifying the dangerous regions of the world.

4 Geolocating

The agent tries to accurately locate the venue of each news article. The latitude and longitude of the location are also retrieved, enabling visualization within the map of the world.

Most news sources use a standard header, consisting of the proper location, e.g. "BASRA, Iraq (Reuters)". Reuters is very consistent in producing these headers. Google News aggregates several news sources, thus displaying different header formats. However, by using the most frequent header of a set of related news, the location can be accurately defined.

Two databases were used when finding geographical names and their coordinates: the GEOnet Names Server (GNS) [6] and the Geographic Names Information System (GNIS) [7]. The former contains more than 4 million geographical features for the entire world; the latter contains 2 million features for US only.

The databases consist of several attributes for each of the geographical features:

- full name (including conventional, native and variant names)
- region
- latitude and longitude (in degrees)
- populated place classification (a graduated numerical scale denoting the relative importance of a populated place)
- feature name

The vast amount of geographical data was first filtered; only larger populated areas were retained, reducing the list to 27,290 features for the entire world.

When only partial location information is available for a document (e.g. city only), the most populated feature from the GIS database is used.

5 User interface

The categorized and located news articles are finally rendered in an SVG (Scalable Vector Graphics [2]) widget. The widget can be displayed in a web browser (using Adobe SVG Viewer) or in a standalone application, such as Apache Batik.

The SVG widget consists of:

- A satellite image of the world in 8096×4096 resolution (enabling three levels of zooming).
- Vector political boundaries of 203 countries.
- News pop-ups, shown when hovered; the pop-ups contain an image, title and the headline; by clicking it, a new browser window is opened, following the URL of the article.
- An ECMAScript library enabling client interaction: zoom-in and zoom-out, panning, news pop-ups, article linking, etc.

The result is an SVG widget displaying the map of the world (figure 3). News articles are marked as circles. The circle sizes denote news importance. The "bad" news is marked as red.

6 Conclusion

As expected, the result SVG widget marked the current crisis areas: the war in Iraq, the terrorist attack in Jakarta and violence in the US. Increased news location resolution would enable an even finer outlook of the world threat level. However, many news sources only cite the capitals or even news agency locations instead of the exact venue. This leads to news aggregation and a distorted threat location view.

The SVM classifier in coordination with the RCV1 corpus were successful in categorizing the news articles. They were especially fitting for the Reuters news source.

The SVG markup language proved to be suitable for geographical applications. The mixture of raster and vector graphics provides for a fast and appealing user interface. The `Newsloc` widget is modular and can be used for other GIS-related applications as well.

References

[1] W. Brenner, H. Wittig, and R. Zarnekow. *Intelligent Software Agents: Foundations and Applications*. Springer-Verlag, 1998.

[2] J. Ferraiolo, F. Jun, and D. Jackson. *Scalable Vector Graphics (SVG) 1.1 Specification*. W3C, 2003. `http://www.w3.org/TR/SVG`.

[3] T. Joachims. Text categorization with support vector machines: Learning with many relevant features. *Proceedings of the European Conference on Machine Learning*, 1998.

[4] T. Joachims. Making large-scale SVM learning practical. *Advances in Kernel Methods - Support Vector Learning*, 1999.

[5] D. Lewis, Y. Yang, and F. Li. RCV1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397, 2004.

[6] National Geospatial-Intelligence Agency (NGA). *GEOnet Names Server (GNS)*, 2004. `http://earth-info.nga.mil/gns/html`.

[7] U.S. Geological Survey. *Geographic Names Information System (GNIS)*, 2004. `http://geonames.usgs.gov/gnishome.html`.

[8] D. Winer. *RSS 2.0 Specification*, 2002. `http://blogs.law.harvard.edu/tech/rss`.

Figure 3: The `Newsloc` SVG widget