An RSS Feed Analysis Application and Corpus Builder

Shereen Khoja
Pacific University
2043 College Way, Forest Grove, Oregon 97116
USA
shereen@pacificu.edu

Abstract

The RSS Feed Analysis Application and Corpus Builder is a software application that downloads given RSS feeds and compiles them into a corpus. The user simply supplies RSS feed addresses and the application automatically connects to the feeds, downloads them, and strips any formatting tags. The application incorporates the Expat (<eXpat/>) XML parser[1] to identify the tags in the RSS feeds, and the users have the flexibility to define what they would like to keep and what is to be stripped. The application was tested on a project to analyse Middle Eastern Blogs. Thirty-seven blogs were downloaded using the RSS Feed Analyser and compiled into a corpus of 131,836 words. Both the RSS Feed Analyser and corpus are freely available under the GNU General Public Licence.
.

Motivation

The motivation behind building an application to automatically download RSS feeds is the frustration that I feel working in the field of Arabic Computational Linguistics in the lack of corpus creation and analysis tools. I know that I am not alone in feeling this frustration, and researchers use various ad-hoc techniques to get through this basic step. As quoted by Al-Sulatie and Atwell (2004):

> *".. it is still difficult to use corpus analysis tools such as concordancers in handling Arabic text unless they are used in Arabic windows and even so the result is not as tidy as in the case of languages with Roman script. Since our corpus will be available on the internet we hope it would be an interesting challenge for software engineers to develop suitable analysis tools."*

My research in Arabic Computational Linguistics has focused on Modern Standard Arabic, but it is the colloquial form of Arabic and developing linguistic tools for this form of Arabic that is of interest to me now since research on Modern Standard Arabic is quite prolific. My first step in investigating colloquial Arabic is to collect a corpus of Arabic blogs

The growth in blogging has provided language researchers with a new form of writing to investigate various linguistic properties. Examples of this research include: an investigation to determine the mood of a blogger based on the text of a post (Mishne & de Rijke, 2006), an analysis of the genre of blogs (Herring et.al, 2004), and the development of an automated trend discovery for blogs (Glance, Hurst & Tomokiyo, 2004).

There has been little to no research on Arabic blogs and no corpus of Arabic blogs currently exists. Compiling a corpus of Arabic blogs is the first step in analysing the Arab blogosphere and investigating trends within that community.

The RSS Feed Analysis Application

An RSS Feed is an XML document used to publish frequently updated web content such as blogs. The format of these documents is mostly standardised, though there are exceptions. RSS feeds contain metadata such as information about the author and the host, and site data such as a blog item or post and user comments on that post. An example of a blog and its RSS feed are shown in figures 1 and 2 below.

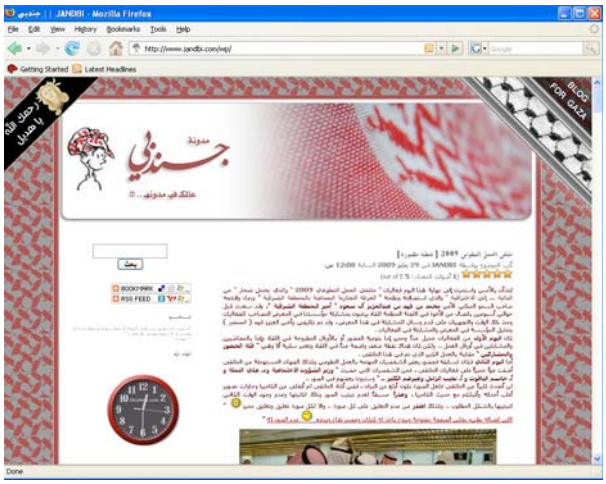

Figure 1: An Arabic Blog Site
http://www.jandbi.com/wp/

[1] http://expat.sourceforge.net/

Figure 2: The Source of an RSS Site

The RSS Feed Analysis Application begins by reading in a file of type *"cbl"*, which is an XML file that contains parameters for the application. *"cbl"* is short for Corpus Builder Language and is an XML format specifically created for the RSS Feed Analysis Application. An example of such a file is:

```
<?xml version="1.0" encoding="UTF-8"?>
<cbl version="1.0">
  <builder>
    <source>
      <outputName>out_jandbi</outputName>
      <dataAdaptor>RSS</dataAdaptor>
      <location>http://www.jandbi.com/wp/?feed=rss2</location>
      <metadata>
        <author>راكد</author>
        <gender>male</gender>
        <country>Saudi Arabia</country>
      </metadata>
    </source>
  </builder>
  <ingester databaseName="db.bin">
    <source name="out_jandbi" type="XML">
    <contentTag sourceTag="content:encoded"/>
  </source>
  </ingester>
</cbl>
```

Figure 3: An Example "cbl" File

The *"cbl"* file consists of two parts: the Builder and the Ingester, which are used by the parts of the application with the same name. The Builder uses all of the sources between the <builder> tags to download the RSS feeds. The example file above contains only one source, but many of these can be described in the *"cbl"* file. No analysis is done at this point.

Source contains a name for the output file, a data adaptor, a location, and metadata. The output file name is used as the name of an XML file produced by the builder that lists the location of the downloaded RSS feed and the source of the feed. The data adaptor is the type of input used to build the corpus. Currently the only acceptable type is RSS but the application is designed in a modular way so that handlers for other adapters such as email or message boards could be added. The location is the URI (Uniform Resource Identifier) of the RSS feed. Finally, the metadata can be anything that the user desires. In the example I have here I am specifying the author, gender, and country of the blogger. This information is not found within the RSS feed and must be added manually.

The Builder

The builder uses the information in the *"cbl"* file to produce three files for each specified source. These files are a "document" file that points to two other files: the data file is a copy of the raw RSS feed, and the info file contains information obtained from the server such as the date downloaded, the date last modified, and the content type.

```
<?xml version="1.0" encoding="UTF-8"?>
<cbl version="1.0">
  <builder>
    <source>
      <outputName>out_jandbi</outputName>
      <dataAdaptor>RSS</dataAdaptor>
      <location>http://www.jandbi.com/wp/?feed=rss2</location>
      <metadata>
        <author>راكد</author>
        <gender>male</gender>
        <country>Saudi Arabia</country>
      </metadata>
    </source>
  </builder>
  <ingester databaseName="db.bin">
    <source name="out_jandbi" type="XML">
    <contentTag sourceTag="content:encoded"/>
  </source>
  </ingester>
</cbl>
```

Figure 4: An Example Document File

The Ingester

The Ingester analyses the downloaded RSS feed to produce the corpus. The second part of the *"cbl"* file contains an ingester tag, which contains the information used to build the corpus. This includes the source that points to the document file produced by the Builder. The ingester tag also contains a content tag that lists the tags that are going to be used to build the corpus. In the *"cbl"* file shown in Figure 1 above, only one tag is used, which is the "content:encoded" tag. Many of these tags could be listed, for example if the user wanted the corpus to contain the title of the article as well as the content. The RSS analysis application is developed in such a way as to give the user flexibility in what they want downloaded and stored in the corpus.

Running the Ingester on the *"cbl"* file produces the raw text that is used to build the corpus. Each word is placed on its own line to aid in the analysis. A sample of the output is shown below.

```
وقد
سعدت
قبل
حوالي
أسبوعين
باتصال
من
الأخوة
في
اللجنة
المنظمة
للقاء
يرغبون
بمشاركة
مؤسستنا
في
المعرض
المصاحب
للفعاليات
ومنذ
ذلك
```

Figure 5: The Arabic Corpus Produced by the RSS Feed Analysis Application

The Corpus

A corpus of Arabic blog was compiled using the RSS Feed Analysis Application.

Before the corpus could be compiled, I needed to determine the blogs that would be included in the corpus. Since I was also interested in the community of Arab bloggers, I decided to compile my corpus around an incident that seemed to unite and shock Arab bloggers from many different countries. This incident was the death of female Saudi Arabian journalist and blogger Hadeel Alhodaif[2], who died at the age of twenty-five after slipping into a coma for a month. Hadeel fought for a freer media in Saudi Arabia and her death made national and international news. After Hadeel slipped into a coma, her father posted a notice on her blog and many Arabic bloggers posted notices and prayers on their blogs.

Blogs were chosen to be included in the blog by searching for the name (Hadeel Alhodaif) in Arabic on Google Blog Search[3].
Thirty seven blogs were collected and the corpus contained 131,836 words. A third of these blogs were written by female bloggers, and the bloggers were from seven countries. The divide is shown below.

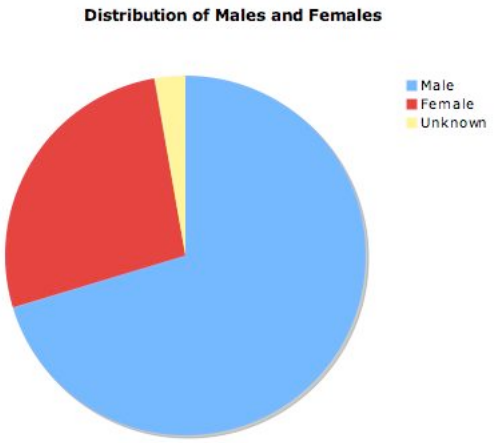

Figure 6: Distribution of Male and Female Bloggers

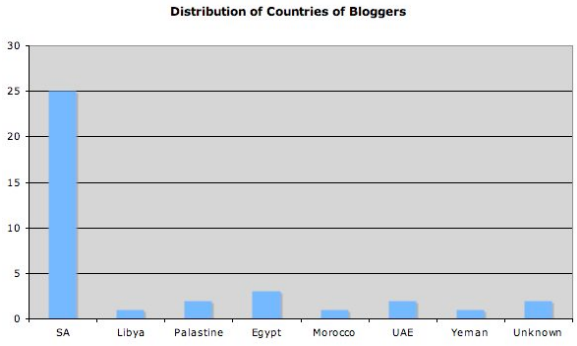

Figure 7: Countries of Origin of the Bloggers

[2] http://www.timesonline.co.uk/tol/news/world/middle_east/article3961731.ece
[3] http://blogsearch.google.com/

The corpus could also be analysed language specific patterns. For example, do the bloggers use Modern Standard Arabic colloquial Arabic or a mixture of both? Sociolinguists may be interested in how bloggers from different countries or genders use the language.

Future Work

The RSS Feed Analysis Application is the first step in developing a multi-purpose corpus creation and analysis application. This application would aid researchers in compiling large corpora from multiple sources and provide basic analysis modules such as stemmers, taggers, and concordancers. Initially these will be targeted for the Arabic language, and support for other languages will be added as required.

The application will be provided to the community for feedback on the various modules that could be incorporated.

Summary

This article described an RSS Feed Analysis Application that is freely available to researchers under the GNU General Public License and can be obtained by contacting the author. The application will run on Windows and can be easily modified to run on Linux and Mac OS X. All the users need to do is use the Corpus Builder Language to specify the RSS feeds they wish to download and the tags they would like included in the corpus.

The application could be used to compile a corpus of RSS of any language simply and quickly for analysis. For example, linguists could investigate multilingual web-based discourse as described in (Baker, 2006).

Furthermore, the 131,836 word corpus is also freely available to the community and can be obtained by contacting the author.

Acknowledgements

I would like to acknowledge the Berglund Center for Internet Studies (http://bcis.pacificu.edu/) for supporting this project.

References

Al-Sulaiti, L. & Atwell, E. (2004). Designing and Developing a Corpus of Contemporary Arabic. In Proceedings of the Sixth TALC Conference (p. 92). Grenada, Spain.

Walters, K. (1996). Diglossia, Linguistic Variation, and Language Change in Arabic. In Perspectives on Arabic Linguistics VIII: Papers from the Eighth Annual Symposium on Arabic Linguistics (pp. 157--197). Amsterdam: Benjamins.

Mishne, G. & de Rijke, M. (2006). Capturing Global Mood Levels using Blog Posts. In Computational Approaches to Analyzing Weblogs (pp. 145--152). Menlo Park, California: AAAI Press.

Glance, N., Hurst, M. & Tomokiyo, T. (2004). BlogPulse: Automated Trend Discovery for Weblogs. In

Proceedings of the WWW Workshop on the Weblogging Ecosystem. New York, New York: ACM.
Herring, S.C., Scheidt, L.A., Bonus, S. & Wright, E. (2004). Bridging the Gap: A Genre Analysis of Weblogs. In Proceedings of the 37th Annual Hawaii International Conference on System Sciences. Los Alamitos, California: IEEE Press.
Baker, P. (2006). Using Corpora in Discourse Analysis. London: Continuum.