NLE Application for HPC Challenge:

Indexing of Newswire Content

Baden Hughes {badenh@cs.mu.oz.au}

Within the news media, a common service known as the “newswire” is available. Typically a newswire service is a dedicated feed of stories from a larger news agency, which is provided to smaller content aggregators for syndication through national and international partnerships. Newswire content is primarily received via a dedicated subscription circuit such as a leased line or satellite service.

Newswire content is essentially a continuous stream of text with little internal structure. This format is inherited from earlier implementations where telegraph and serial line printers were used to receive newswires. In electronic form, the majority of newswire services are provided in basic SGML (although this is gradually changing). Newswire content is relatively error free, but may contain occasional transmission or human errors. Newswire services themselves vary between stream-based and chunk-based delivery modes - realtime stream-based services don't have much structural differentiation, whilst chunk- based services have better grouping of content and delineation between content types. Additionally, newswire content may feature "slugs", a short string used by editors to link related stories over a day, which are a kind of internal index.

Within newswire content there is inherent duplication - transmissions are repeated, sometimes with minor alterations as stories change, or sometimes exact copies, depending on the delivery mode and editorial process of the source agency. Furthermore, the style of the newswire may vary based on orientation of the supplier (eg official government information sources vs fully commercial news operations). Another complication is that in some cases newswire services are multilingual, providing multiple translations of each story, thus interleaving linguistic variation along with structural complexity.

Newswire content is divided into 4 main categories. The most typical newswire item is the *story*, a coherent self-contained report on a particular topic or event. Next most frequent is the *multi*, a series of summaries such as news in brief or “today’s news”. Next again is the *other*, which is information intended for general circulation and topically coherent, usually lists of statistics such as sports scores, stock prices, weather forecasts. Finally, the *advis* is a note destined for news editors themselves, and not intended for publication (broadcast embargoes etc).

Whilst newswire is an interesting real time medium which poses processing problems for the receiver, newswire can also be archived for subsequent retrieval. At a basic level this archiving process can be to divide the content by temporal extent (weeks, months, years) and compressing the sources. So, why would we chose to work with such data sources ? Our requirement for search and retrieval assistance may be motivated by a simple

historical enquiry (for example, find all the stories in 1994 about Clinton and Lewinsky). A different requirement may be comparative – evaluating how different agencies reported the same event from different perspectives eg US vs European media, New York vs Los Angeles media, television vs cable vs print vs internet. Even more interesting may be the investigation of newswire stories which were not picked up by syndicators or broadcast media. In all these cases, how do we extract meaningful information from newswire archives ? Our initial requirement is to quantify and extract index information from the original archives themselves, which can then be used as the starting point for subsequent analyses.

Newswire archives are available through self-archiving of newswire subscriptions or through collated corpora. In this experiment we use samples from the Linguistic Data Consoritum's Gigaword Corpus, which is a collection of 4 different newswire sources (Agence France Press English Service, Associated Press Worldstream English Service, New York Times Newswire Service, and Xinhua News Agency English Service) over a period of 7 years. Whilst not being enourmous, newswire achives are reasonable in size. A typical newswire service will generate 15-20Mb per month of raw text. Collecting multiple newswire subscriptions over an extended period (such as the exercise conducted by LDC), will result in a multi-Gb compressed document collection. In this particular demonstration, we are using the 1995 collection from Agence France Press English Service, which contains about 100Mb of newswire text.

In processing newswire archives, there are two different general outputs in which we are interested. The first is basically statistical, determining the number of different types of documents that are in the archive itself. The second is basically indexational, namely to extract all the relevant document ids and headlines for a specific document type to create an index to the archive itself. This allows for greater efficiency in recall since an index is significantly smaller than the archive itself.