CoReMo Plagiarism Detector System

CoReMo is a cross-lingual external focusing plagiarism detection system, created to participate in PAN competitions.

CoReMo has participated in PAN since 2010 to 2012 editions, renewed every time, looking for improved performance, high speed responses and low hardware requirements.

The CoReMo effectiveness and its high speed are due to the combination of the speeds of N-gram (xCTNG), a High Accuracy Information Retrieval System (HAIRES), a prune strategy to minimize comparisons (Reference Monotony), an integrated local translation system and, from now on, a precise pairs document comparison module, everything joined to a hard CC++-customized programming job.

CoReMo can be used in either command line or interface ways. At the web interface (http://www.coremedetector.com) it is being tested to offer high quality / low cost analysis services very soon.

Detailed Document Comparison (1)

Trazable Ngram

Every Extended Contextual N-gram obtained is used to generate a Trazable Nagram object, which registers the offset and length time, has a fixed name and is used to register the n-gram matching occurrences for inner and external cases.

F astly Comparable Document

The document is modeled to fit a fast comparison and a matching sections location method. It uses two TrazableNgram vectors: the Natural vector (ordered by the natural position) and the Ordered vector (with former TrazableNgram references in alphabetical order, disregarded by position) got by QuickSort.

After getting the ordered version of the document, filling the inner matching list for every Trazable Ngram is fast and easy, therefore more n-gram comparisons than the N-gram number existing in the largest document.

A Fast Detection/Comparison Software

CoReMo has always been a highlighted fast software in former PAN editions. In PAN2013 it has been highlighted not only for getting the best detections, but as the fastest algorithm for the Text Alignment Task, being at least 4 times faster than any other competitor.

The keys for its remarkable speed are:

- Optimized GCC++ + OpenMP 64 bit programming
- GNU Linux 64bit OS and x64 file system platform
- Internal sort of n-grams by Bubble sort algorithm
- n-grams in a document ordered by Quicksort
- Modified Mergesort algorithm to compare both docs
- Local translations by dictionary mapping
- Taking the advantage of the high-precision automatic modeling when repeated in consecutive comparisons

However, this version takes the advantage of multicores processors technology advantage, and a web interface.

The runtime for the comparison was tested by only a Single Core 2.27 GHz Virtual Machine, needing 75 seconds to analyze 9185 documents pairs.

Using a 6 cores AMD FX810 @ 4.0 GHz, we got the same analysis in less than 5 seconds.

H A I R E S

(Indexed the Extended Contextual Ngrams (90% has or 10% in a few cases))

CoReMo uses this pruny way: “disorder matching if not happening monotonously” in several modules and steps to avoid unneeded deep comparisons. An example is the combined action of HAIRES and another involved sections localization which make any comparisons. RM is also used for optimized web retrieval and discarding noisy matching in the detailed module.

Reference Monotony Pruning Strategy

Gr a d u a l R e d u c t i o n

After the direct detection (leaves), adjacent detections in both source and suspicious documents are joined by a “filtering” phase when the distance is lower than a threshold (4,000 n-grams), getting a pruning execution time when the lowest minimum length is required to consider a detection.

A comprehensive coloured comparison document could be created by combining both direct and filtered detections, detailing the evident direct detection and the jarring effect.

C r o s s l i n g D e t e c t i o n

When a non-English source document is detected, CoReMo can use its word by word stem mapping translation strategy locally for indexing or comparison purposes.

It gets a fastest response and high availability (and economy), as it is not foreign system dependent.

However, it can use improved quality translated versions of the documents of external translation systems.

This simple local translation mapping uses two external dictionaries (Direct2Slem and Slem2Slem) to get the quickest stem translation. The second one gets improved recall when first one fails to find derivated word.

The system is currently using very poor dictionaries, and about 50% of words are not found at any improved dictionary should improve the results.

When getting the n-grams, the original non-English words offset and length are used, to get an exact location of the source section in the detection process.