Report on the Effingham Agreement

Thomas Krichel, Victor M. Lyapunov and Tanya I. Yakovleva

2000–04–04

The lyak1.pl script finds 14,032 unformatted abstract files in the arXiv.org collection. Lyak1.pl converts this data to the following format (“lyak-records”):
paper-code <internal debugging indicator>
SOURCE: untagged Title+Authors+etc. field from abstract file, possible continuation lines starts with a space
Title: The lyak1.pl-proposed Title field, possible continuation lines starts with a space
Authors: The lyak1.pl-proposed Authors field, possible continuation lines starts with a space
<EMPTY LINE as a record separator>

These data are contained in the file "pre-man.txt". A command line running lyak1.pl at openlib.org looks like this:
find ~/data/xxx/ -name *.abs |xargs cat |./lyak1.pl > pre-man.txt
The lyak-records were manually tested and edited if necessary by Tanya. She corrected 3,160 of total 14,032 records, which makes about 22.5%. The result was stored in the file “post-man.txt”. Victor created the file “mod.txt” of lyak1-generated records, which gone under manual editing, and inspected carefully a significant part of it. All the manual corrections of lyak1.pl script’s proposition appeared to be justified.
The expected share, according to the agreement text, of manually corrected records was 10%. The share of manually fixed records is therefore greater than expected. Victor can explain—but not justify—this great difference between the result and the estimation:

- the lyak1.pl considers list of authors follows immediately after the title. A lot of source records, however, contain some publication-relevant data between the title and the author list. Lyak-1 glues this data to the ’title’ field, thus the latter needs cleaning.
- the same situation appears at the end of the author list. The script often glues part of author-irrelevant data to its end.

While doing the estimation in autumn 1999, Victor concentrated on the correct location of the beginning of the author list. This estimation happened very close to the result: the beginning of author list was located incorrectly in 1,426 of 14,032 records, which makes ~10.2% .
Lyak2.pl script processes the post-man.txt file and produces the hash-file “tac.db” with paper-code as key and {Title, Authors, Comments, Journal-ref, Subj-class} lines as value.
Lyak2.pl outputs a diagnostic message for every record it can’t handle:
<Reason> paper-code <tech.info>
S: <source untagged field>
T: <Title>
A: <Authors list>
J: <Journal-ref>
L: <Subj-Class>

Optionally (-v) a text dump of hash-file is done in the same format. The command line to run lyak2.pl looks like this:
./lyak2.pl [-v] < post-man.txt > diag.txt

Most part of rejected records lack either Title or Authors fields, thus being a subject of correcting the source abstract files.
Lyak2.pl could not process some records with disjoint Authors field, even after manual correction, the typical template follows:
SOURCE: title author-name {paper-relevant info} author-workplace
Title: title
Author: author-name author-workplace

There are 110 rejected records in total.
All the files mentioned above are placed at ftp://openlib.org/pub/openlib/effingham