

Motivation

Large amounts of chemical information has been and is being published in a form not directly accessible for machine processing. Chemical names and structure images in patents and journal articles are not immediately usable for virtual screening, docking studies, or QSAR modeling, yet there are vast repositories of such data which can potentially be made available with proper software tools. An example of such a repository is the patent office. Patent documents often contain information not yet published anywhere else and the existing data mining techniques for chemistry-related patents are expensive and laborious.

OSRA aims to facilitate data mining of structure information contained in images in such publications. It is a utility capable of automatically extracting and converting a molecular image into any of the several popular cheminformatics formats.

Current capabilities

OSRA can process multi-page PDF documents as well as many different image formats such as TIFF, GIF, PNG, JPEG, etc. It accepts images scanned at 150 dpi, 300 dpi or higher resolutions, or web-oriented images at 72 dpi. Page segmentation of images from the surrounding text is automatically performed.

OSRA can recognize superatom labels and the distribution contains a user-modifiable dictionary of abbreviations and labels. As it is a command-line utility it can easily be added to a third-party software workflow or used for batch processing. Add-ins exist for Accelrys/Symyx Draw, Chemaxon MarvinSketch, CambridgeSoft ChemOffice, BKChem, etc.

OSRA is free software and can be downloaded as source code, or Windows and Mac installation packages.

It is available at <http://osra.sourceforge.net>

Benchmarks

Set	Size	1.3.6	1.3.7	1.3.8
USPTO	5735	75%	77%	80%
JPO	450	53%	54%	55%

Two sets of clip images – from the United States Patent Office and from the Japanese Patent Office along with the corresponding ground truth molecular files (the first one from the Complex Work Units initiative, the latter courtesy of Chem-Infty project at Kyushu University) are used to represent the progress in recognition rate for OSRA from version 1.3.6 to the most recent – 1.3.8. The recall rates have been calculated by using InChI identifiers.

Differences in scan quality and typographic variability account for the spread in the recognition rates between the two sources.

Challenges

Not all patent documents are created equal and some present more challenges for image recognition software than others. The diversity of rules and regulations between various patent offices in different countries creates large variability in scan and typographic quality. One particular example is WIPO document WO29126624A1(2). Even though it was scanned at a high resolution the submitted document was a fax copy of the original typeset manuscript. This created a highly challenging PDF file as it is much more common to encounter low resolution scans of well-printed papers than a high resolution scans of low quality documents - at least for modern patents and journal articles, not considering historical manuscripts. OSRA contains several specialized image pre-processing routines which are sometimes beneficial for improved recognition. The problem is while some settings allow for better recognition of some images there is no universal “best” set of options for all structure images even within the same document. A proposed solution is to combine the results of several runs with different options and filter the combined set.

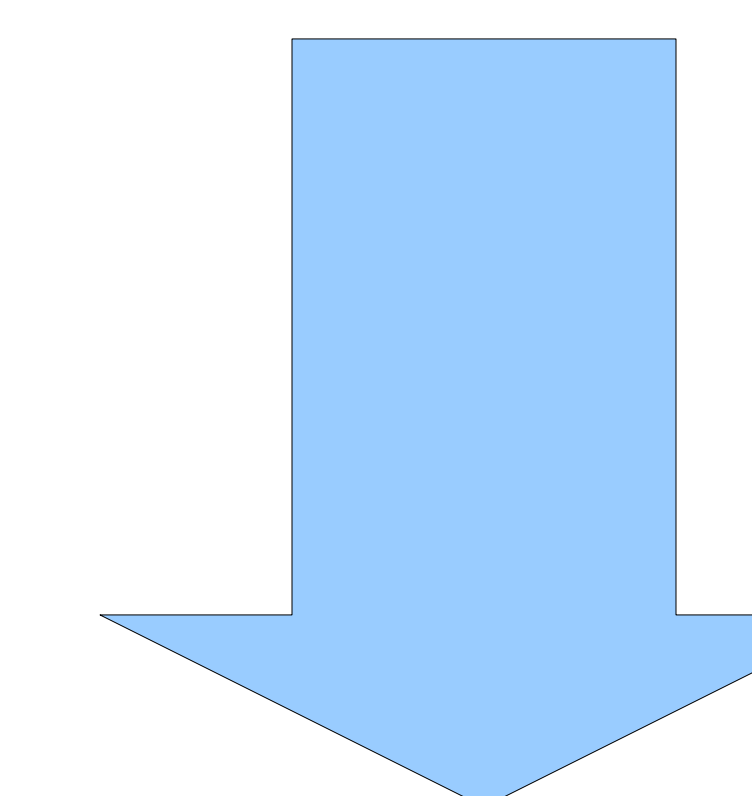
Combination approach

140 unique structures

Options	Recognized	Time
default	24	1m29s
-r 300	9	4m57s
-r 300 -u 1	11	
-r 300 -u 2	13	11m55s
-u 1	16	
-u 2	14	
-j	38	1m29s
-j -r 300	65	4m30s
-j -r 300 -u 1	64	
-j -r 300 -u 2	67	11m4s
-j -u 1	32	
-j -u 2	33	

Default recall: 17% precision: 7%

Best recall: 47% precision: 23%



After combining and filtering:

Recall: 57%

Precision: 51%

Advantages: automatic processing, improved recall and precision rates, no need for manual parameter configuration

Disadvantages: longer processing time

The detailed workflow is available at OSRA SourceForge website.