

潜在的意味解析法LSAに基づく ソフトウェアシステム分類法の提案

川口 真司

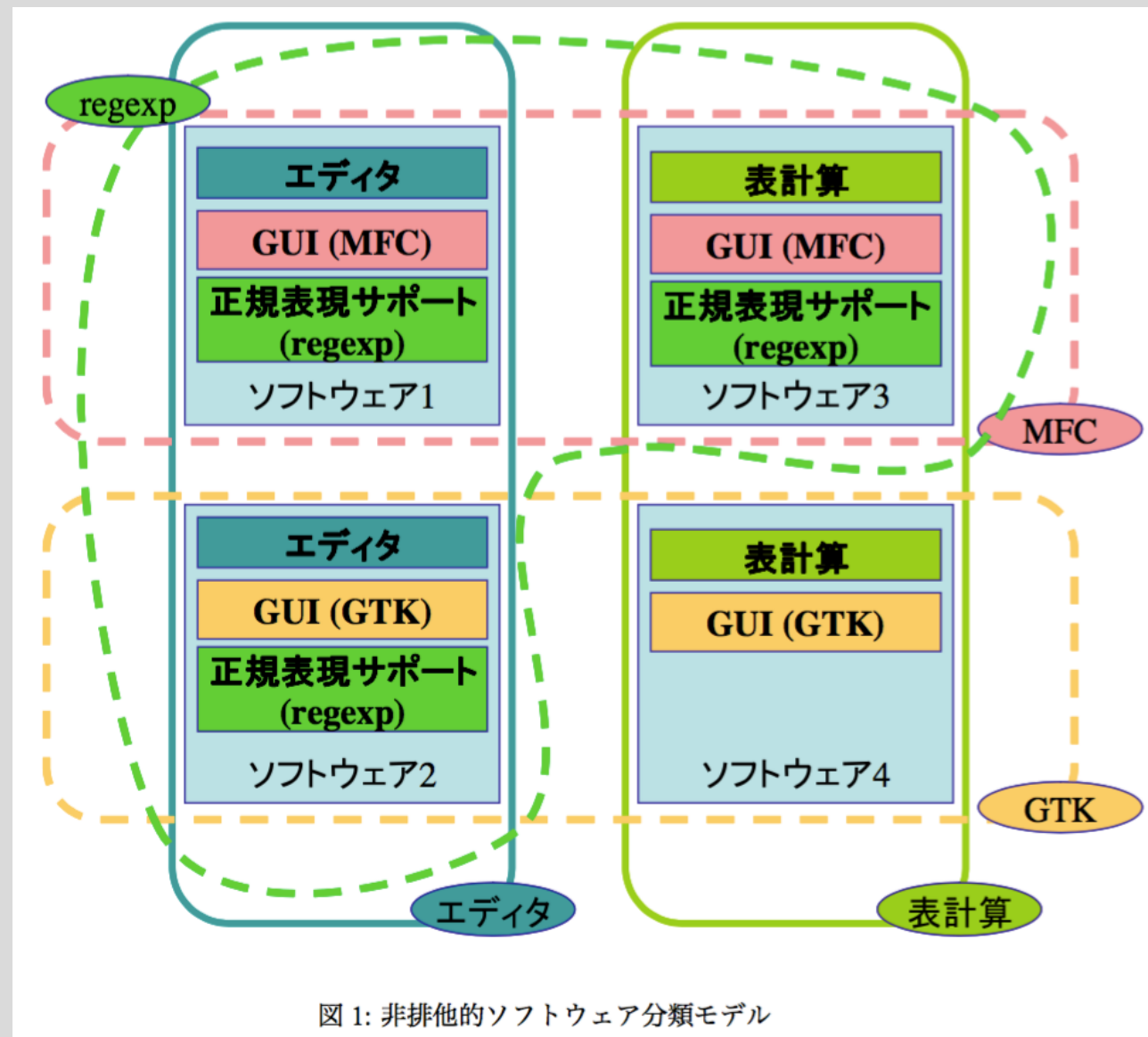
手動ソフトウェア分類の困難化

- SourceForgeに代表されるソフトウェアリポジトリサービスは、未知のソフトウェアプロダクトを検索、発見する場としての価値も高い
- このようなサービスはOSS開発の活発化とともに盛んに利用されるようになった
- 一方、プロジェクト数の増加のため、検索用の分類をサービス管理者が手動で行っていた従来の手法には限界が見えてきている

ソフトウェア分類を自動化したい

LSA (Latent Semantic Analysis、潜在的意味解析法) を応用して、
ソースコードから機械的にソフトウェアを分類する手法を提案する

ソフトウェアの様々な側面と分類方針



ソフトウェアは**複数の側面を内包している**ため、非排他的に分類するほうがより現実に沿った分類であると考えられる(これを実現できる手法が必要である)

既存の関連研究と問題点

ソフトウェアを何らかの機能的単位へ分割する研究

- 各ソフトウェアの関数やファイルがそれぞれ単一の機能を提供していることを前提としていた

ソフトウェアシステム間の関連を対象にした研究

- コードクローンが存在する行の割合を類似度として用いている
- そのため、同じソースから派生したようなソフトウェア間でないと適正な類似度が出力されない(出自が全く異なるようなソフトウェア間では類似度が極端に低くなる)

アプローチ

- プログラムの**識別子** (変数名や関数名など) は、それが使われている文脈での動作を表しているものと考えられる
- 多量の識別子の中から関連の高い識別子群を抜き出せば、それらが一つのカテゴリを表していることが予想される
- このカテゴリを参考にすれば、ソフトウェアを機械的に分類できる
- 関連の高い識別子群を特定するために**LSA**を応用する

手順

1. トークン抽出
2. 共起行列化
3. 出現頻度によるフィルタ
4. LSA
5. トークン間の類似度計測
6. トークンのクラスタ分析
7. クラスタタイトルの作成

1. トークン抽出

各ソフトウェアの識別子を抽出し、トークンとする

※予約語や演算子、コメントをトークンに利用することも可能ではあるが、今回はよりソフトウェアの特徴を表し、各ソフトウェアから同等の頻度、品質で得られる識別子を用いた

2. 共起行列化

いわゆるBag of Words化を行う

3. 出現頻度によるフィルタ

孤立トークン(単一のソフトウェアにのみ存在するトークン)

LSA においては全く無意味なので削除

普遍的トークン(過半数のソフトウェアに存在するトークン)

分析結果にほとんど影響がないと考えられるので削除

4. LSA

1. 共起行列Aの各要素をtf-idf値に変換する
2. 変換後の共起行列Aに対して特異値分解を行い、行列U、S、Vを得る
3. 各行列の次元をrに落とし、再び掛け合わせて行列 $\hat{A}$ を得る

5. トークン間の類似度計測

行列 $\hat{A}$ から単語 (トークン) ベクトルを抽出し、互いの類似度を計算する

6. トークンのクラスタ分析

単語ベクトル間の類似度を用いて、単語の**クラスタ分析**を行う

- クラスタ分析の手法には、全てが独立したクラスタという状態から始めて、もっとも類似度の高いクラスタから順次結合していく方式を採用する
- 用いる距離には、二つのクラスタ間のうち最も距離の長い(類似度の低い)要素間の距離をクラスタの距離とする最長距離法を採用する
- トークン数が膨大なため、以下の二段階に分けてクラスタ分析を行う
 1. 明確なクラスタを形成するトークンを抽出
 2. そのトークンのクラスタを算出

7. クラスタタイトルの作成

- クラスタを構成するトークンから、そのトークンを含むソフトウェアの和集合を一つのソフトウェアクラスタとする
- ソフトウェアクラスタを理解しやすくするために、クラスタタイトル(クラスタに入っているソフトウェア群を代表する文字列)の抽出を行う

手順(再掲)

1. トークン抽出
2. 共起行列化
3. 出現頻度によるフィルタ
4. LSA
5. トークン間の類似度計測
6. トークンのクラスタ分析
7. クラスタタイトルの作成

実験

提案手法を用いて実際にソフトウェアの分類を行い、以下の確認を行った

- 既存の分類と比較して適正な分類が可能かどうか、
- 利用しているライブラリなど既存の分類では考慮されていない分類について分類ができているかどうか

実験対象

実験対象は、無作為に選んだSourceForgeの5つのカテゴリに含まれるC言語のプログラムとした

カテゴリ	ソフトウェア
boardgame	Sjeng-10.0, bingo-cards, btechmux-1.4.3, cinag-1.1.4, faile_1_4_4, gbatnav-1.0.4, gchch-1.2.1, icsDrone, libgmonopd-0.3.0, netships-1.3.1, nettoe-1.1.0, nngs-1.1.14, ttt-0.10.0
compilers	clisp-2.30, csl-4.3.0, freewrapsrc53, gbdk, gprolog-1.2.3, gsoap2, jcom223, nasm-0.98.35, pfe-0.32.56, sdcc
database	centrallix, emdros-1.1.4, firebird-1.0.0.796, gtm_V43001A, leap-1.2.6, mysql-3.23.49, postgresql-7.2.1
editor	gedit-1.120.0, gmas-1.1.0, gnotepad+-1.3.3, molasses-1.1.0, peacock-0.4
videoconversion	dv2jpg-1.1, libcu30-1.0, mjpgTools, mpegsplit-1.1.1
xterm	R6.3, R6.4

表 1: 実験用ソフトウェアの一覧

結果(1)

No.	クラスタイトル	ソフトウェア	トークン数
1	AOP, emitcode, IC_RESULT, IC_LEFT, aop, aopGet, IC_RIGHT, pic14_emitcode, iCode, etype	compilers/gbdk, compilers/sdcc	8597
2	CASE_IGNORE, CASE_GROUND_STATE, screen, CASE_PRINT, CASE_BYP_STATE, Widget, TScreen, CASE_IGNORE_STATE, CASE_PLT_VEC, CASE_PT_POINT	xterm/R6.3, xterm/R6.4	2160
3	YY_BREAK, yyvsp, yyval, DATA, yy_current_buffer, tuple, yy_current_state, yy_c_buf_p, yy_cp, uint32	compilers/gbdk, database/mysql-3.23.49, database/postgresql-7.2.1	223
4	AVI, cinfo, OUTLONG, avi_t, AVI_errno, hdr1_data, OUT4CC, nhb, ERR_EXIT, str2ulong	videoconversion/dv2jpg-1.1, videoconversion/libcu30-1.0, videoconversion/mjpgTools	177
5	domainname, msgid1, binding, msgid2, domainbinding, pexp, _builtin_expect, transmem_list, codeset, codesetp	boardgame/gbatnav-1.0.4, boardgame/gchch-1.2.1	165
6	board, num_moves, ply, pawn_file, npiece, pawns, moves, white_to_move, move_s, promoted	boardgame/Sjeng-10.0, boardgame/cinag-1.1.4, boardgame/faile.1.4.4	154
7	xdrs, blob, DB, UCHAR, XDR, mutex, key_length, logp, page_no, bdb	database/firebird-1.0.0.796, database/mysql-3.23.49	118
8	domainname, N_, binding, gchar, Gtk-Widget, PARAMS, codeset, gpointer, loaded.110nfile, argz	boardgame/gbatnav-1.0.4, boardgame/gchch-1.2.1, editor/gnotepad+-1.3.3, editor/peacock-0.4	118

結果(2)

No.	クラスタイトル	ソフトウェア	トークン数
9	GtkWidget, gchar, gpointer, gint, widget, gtk_widget_show, N_, g_free, dialog, g_return_if_fail	boardgame/gbatnav-1.0.4, editor/gedit-1.120.0, editor/gmas-1.1.0, editor/gnotepad+-1.3.3, editor/peacock-0.4	104
10	AOP, emitcode, esp, IC_RESULT, IC_LEFT, obstack, aop, mov, aopGet, IC_RIGHT	compilers/clisp-2.30, compilers/gbdk, compilers/sdcc	100
11	tuple, uint32, plan, int32, lsn, elm, rec, interp, TCL_ERROR, finfo	database/mysql-3.23.49, database/postgresql-7.2.1	79
12	xdrs, blob, DB, UCHAR, XDR, mutex, key_length, logp, page_no, bdb	database/firebird-1.0.0.796, database/mysql-3.23.49	73
13	UCHAR, relation, stmt, trigger, yyvsp, yyval, t_data, plan, dbname, USHORT	database/firebird-1.0.0.796, database/postgresql-7.2.1	68
14	fout, interp, TCL_ERROR, typ, YY_RULE_SETUP, List, DATA, Tcl_Interp, id, YY_BREAK	compilers/freewrapsrc53, compilers/gbdk, compilers/gsoap2, database/postgresql-7.2.1	50
15	GtkWidget, gchar, gpointer, dlg, gint, g_free, gtk_widget_show, gtk, GList, GTK_BOX	editor/gedit-1.120.0, editor/gmas-1.1.0, editor/gnotepad+-1.3.3	46
16	UCHAR, relation, stmt, trigger, yyvsp, yyval, t_data, plan, dbname, USHORT	database/firebird-1.0.0.796, database/postgresql-7.2.1	43
17	AOP, emitcode, mfp, ic, uchar, IC_RESULT, IC_LEFT, aop, aopGet, IC_RIGHT	compilers/gbdk, compilers/sdcc, database/mysql-3.23.49	36
18	adr, FX, word, stm, ED, xt, REF, prop, term, FP	compilers/gprolog-1.2.3, compilers/pfe-0.32.56	35

結果(3)

No.	クラスタイトル	ソフトウェア	トークン数
19	AOP, emitcode, IC_RESULT, IC_LEFT, aop, aopGet, IC_RIGHT, pic14_emitcode, iCode, etype	compilers/gbdk, compilers/sdcc, database/firebird-1.0.0.796	31
20	dyn, FPRINTF, process_id, p_offset, ctl, rab, que, io_ptr, prior, PRINTF	database/firebird-1.0.0.796, database/gtm.V43001A	29
21	dyn, FPRINTF, process_id, p_offset, ctl, rab, que, io_ptr, prior, PRINTF	database/firebird-1.0.0.796, database/gtm.V43001A	27
22	regparse, dbp, mech, reginput, flagp, NOTHING, tuple, db, _P, regnode	boardgame/btechmux-1.4.3, database/leap-1.2.6, database/mysql-3.23.49	26
23	rectype, argp, rec, fileid, save_errno, data_len, qp, argpp, int4, dbp	database/gtm.V43001A, database/mysql-3.23.49	26
24	AOP, emitcode, IC_RESULT, IC_LEFT, aop, aopGet, IC_RIGHT, pic14_emitcode, iCode, etype	compilers/gbdk, compilers/sdcc, videoconversion/mjpgTools	26
25	jobject, JNIEnv, JNICALL, JNI_EXPORT, jint, jstring, interp, TCL_ERROR, objv, TCL_OK	compilers/freewrapsrc53, compilers/jcom223, compilers/pfe-0.32.56, database/mysql-3.23.49	24
26	entrypoint, USHORT, TEXT, yyvsp, raddr, R, UCHAR, yyval, blob, REQ	compilers/clisp-2.30, database/firebird-1.0.0.796	17
27	int32_t, dbp, cinfo, net, unpack, argp, sinfo, curl, purpose, mysql	database/mysql-3.23.49, videoconversion/mjpgTools	17
28	AOP, emitcode, mfp, ic, uchar, IC_RESULT, IC_LEFT, aop, aopGet, IC_RIGHT	compilers/gbdk, compilers/sdcc, database/mysql-3.23.49	16

結果(4)

No.	クラスタイトル	ソフトウェア	トークン数
29	USHORT, UCHAR, blob, REQ, NULL_PTR, hlcon, SCHAR, interp, wndclass, bdb	compilers/freewrapsrc53, database/firebird-1.0.0.796	16
30	optind, nextchar, _P, optstring, last_nonopt, option_index, uchar, optarg, pfound, dbp	boardgame/ttt-0.10.0, compilers/clisp-2.30, database/mysql-3.23.49	15
31	int4, ctl, tn, rec, semid, blkno, ti, oprtype, save_errno, AH	database/gtm.V43001A, database/postgresql-7.2.1	14
32	notify, mech, PyObject, fargs, Node, Name, pset, zone, tprintf, NOTHING	boardgame/btechmux-1.4.3, database/postgresql-7.2.1	11
33	interp, notify, dbp, tuple, mech, PyObject, uint32, plan, int32, buff	boardgame/btechmux-1.4.3, database/mysql-3.23.49, database/postgresql-7.2.1	10
34	adr, stm, AOP, emitcode, operands, ASSERT, IC_RESULT, pred, lg, REF	compilers/gprolog-1.2.3, compilers/sdcc	9
35	yyvsp, yyn, PARAMS, codeset, domainname, msgid1, binding, msgid2, yylsp, domainbinding	boardgame/gbatnav-1.0.4, boardgame/gchch-1.2.1, compilers/clisp-2.30	9
36	ERREXIT, picture, pool_id, USHORT, get_buffer, output_buf, cinfo, xxx, UCHAR, streams	database/firebird-1.0.0.796, videoconversion/mjpgTools	9
37	REF, dyn, USHORT, vec, path_name, clause, STATUS, E, UCHAR, CSB	compilers/gprolog-1.2.3, database/firebird-1.0.0.796	8
38	AOP, emitcode, pfile, ic, IC_RESULT, IC_LEFT, aop, aopGet, IC_RIGHT, pic14_emitcode	compilers/gbdk, compilers/sdcc, database/postgresql-7.2.1	7
39	ic, ply, npiece, score, AOP, pawn_file, uchar, bking_loc, wking_loc, emitcode	boardgame/Sjeng-10.0, compilers/gbdk	7

結果(5)

No.	クラスタイトル	ソフトウェア	トークン数
40	clause, cinfo, pred, ci, Group, Np, word, X, A, tmp4	compilers/gprolog-1.2.3, database/postgresql- 7.2.1, videoconver- sion/mjpgTools	6

表 2: 41 ソフトウェアの全分類結果

実験結果まとめ

- 同一ジャンル内のクラスタとなっているのが18/40クラスタ
- 既存の分類では同一ではないものの、同一のライブラリを利用している、同一のアーキテクチャを持っているなどの強い関連を持つクラスタが8つ
- 特に結び付きの強い上位20クラスタに限定すると、同一ジャンル、もしくは関連を持つクラスタは17つ
- 41アプリケーションのうち、いずれのカテゴリにも分類されなかったソフトウェアが12個

考察

- 既存の分類と比較した場合、どこにも分類されなかったソフトウェアが多く、既存の分類を十分にカバーできているとは言い難い(→トークン数の少ないものがどこにも分類されずその他となってしまったことが原因)
- 一方、既存の分類では考慮されていない分類については、本手法を用いることで新たな分類を得られることを確認した
- クラスタイトルについては、使用ライブラリが同一のものなどは分かりやすいタイトルがつく傾向にある一方、アプリケーションドメインが同一のものは必ずしもわかりやすいとは言えないタイトルも存在した

所感

- LSAにおける次元圧縮の度合い(論文では次元 r と表現)の具体的な数値と理由が見当たらなかった(非常に残念)
- 実験部分の書き方が雑？