MUDABlue

Presented by Peter Rigby

MSR class

November 8, 2006

Introduction

- Goal: Automatically categorize software
- Join communities
  - Leverage each other's work
- Developers can learn “best practices”
- Manual categorization is incomplete and time consuming.

Advantages

- Don't need predefined categories
  - Previous work needed predefined categories
- Multiple membership
  - Not mutually exclusive
- Source code only
  - All projects have source, but not all have documentation

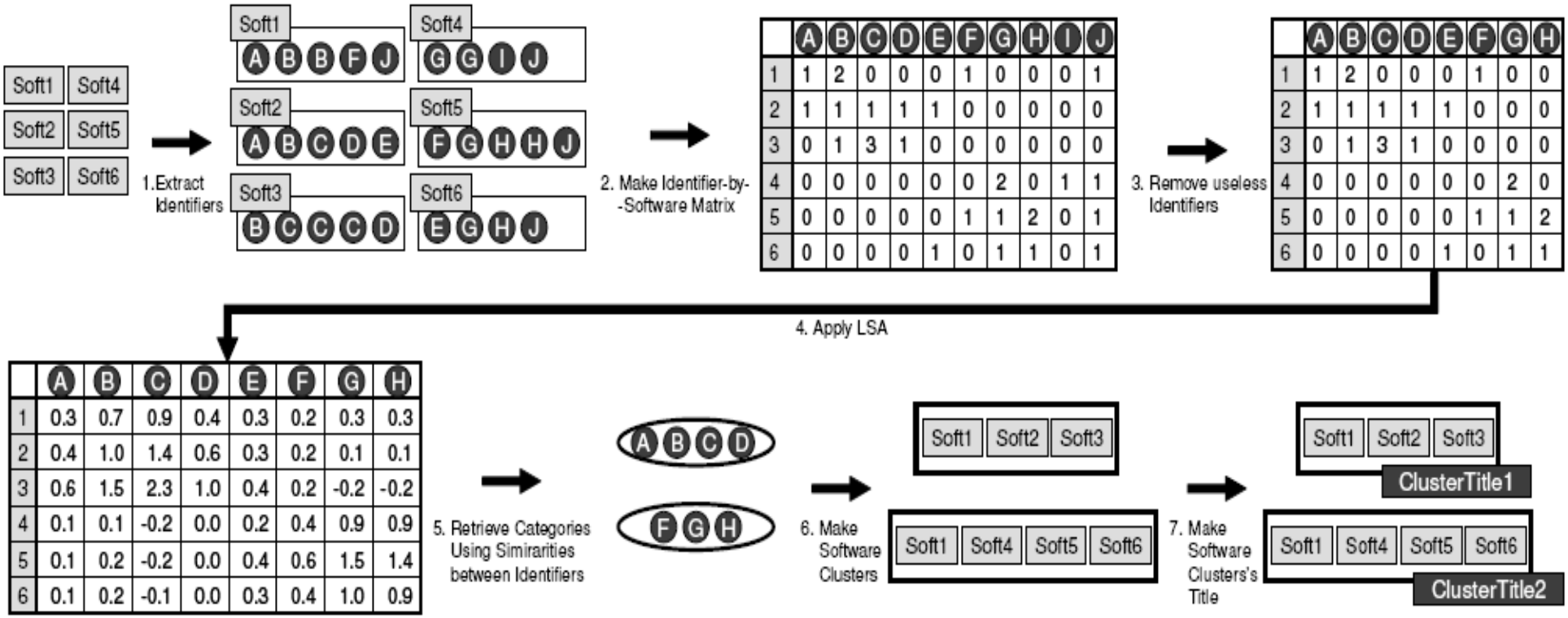
Latent Semantic Analysis

- Statistical technique to extract contextual meaning of words
- Has been used in SE to cluster software components and link documentation and code
- See example later

Technique

- Extract identifiers
  - Exclude comments
- Create identifier-by-software matrix
- Remove meaningless identifiers
  - e.g. Only in one software system
- Apply LSA
- Retrieve categories
  - Cosine criterion
- Create clusters
- Create categories
  - Sum of all identifier vectors for a cluster (10)

Steps



	A	B	C	D	E	F	G	H	I	J
1	1	2	0	0	0	1	0	0	0	1
2	1	1	1	1	1	0	0	0	0	0
3	0	1	3	1	0	0	0	0	0	0
4	0	0	0	0	0	0	2	0	1	1
5	0	0	0	0	0	1	1	2	0	1
6	0	0	0	0	1	0	1	1	0	1

	A	B	C	D	E	F	G	H
1	1	2	0	0	0	1	0	0
2	1	1	1	1	1	0	0	0
3	0	1	3	1	0	0	0	0
4	0	0	0	0	0	0	2	0
5	0	0	0	0	0	1	1	2
6	0	0	0	0	1	0	1	1

	A	B	C	D	E	F	G	H
1	0.3	0.7	0.9	0.4	0.3	0.2	0.3	0.3
2	0.4	1.0	1.4	0.6	0.3	0.2	0.1	0.1
3	0.6	1.5	2.3	1.0	0.4	0.2	-0.2	-0.2
4	0.1	0.1	-0.2	0.0	0.2	0.4	0.9	0.9
5	0.1	0.2	-0.2	0.0	0.4	0.6	1.5	1.4
6	0.1	0.2	-0.1	0.0	0.3	0.4	1.0	0.9

Unifiable Cluster Map

- Allows one to combine clusters visually
- Use a touch graph

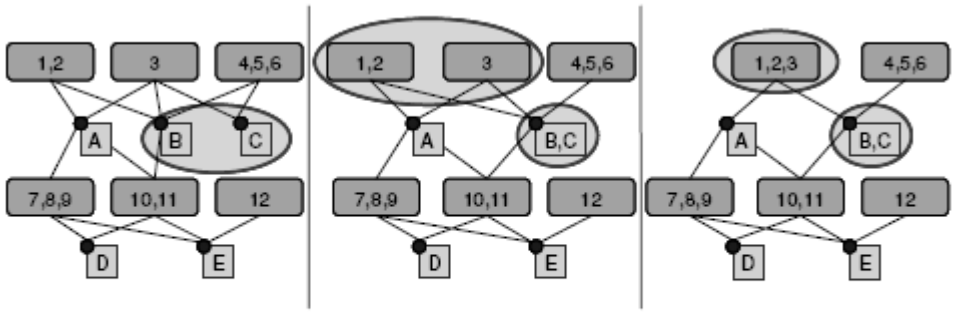


Figure 4. Unification of categories

Category hierarchy View

- Categories grouped by **odds ratio**
- The ratio of an event occurring in one group vs. occurring in another group
- p/(1-p) = p(1-q)
- q/(1-q) q(1-p)
  - > = 1 similar
- Dendrogram
  - A branched diagram representing the apparent similarity or relationship between taxa

- MUDABlue vs Manual Classific.

$$\text{recall} = \frac{\sum_{s \in S} \text{recall}_{\text{soft}}(s)}{|S|}$$

$$\text{recall}_{\text{soft}}(s) = \frac{|C_{\text{MUDABlue}}(s) \cap C_{\text{Ideal}}(s)|}{|C_{\text{Ideal}}(s)|},$$

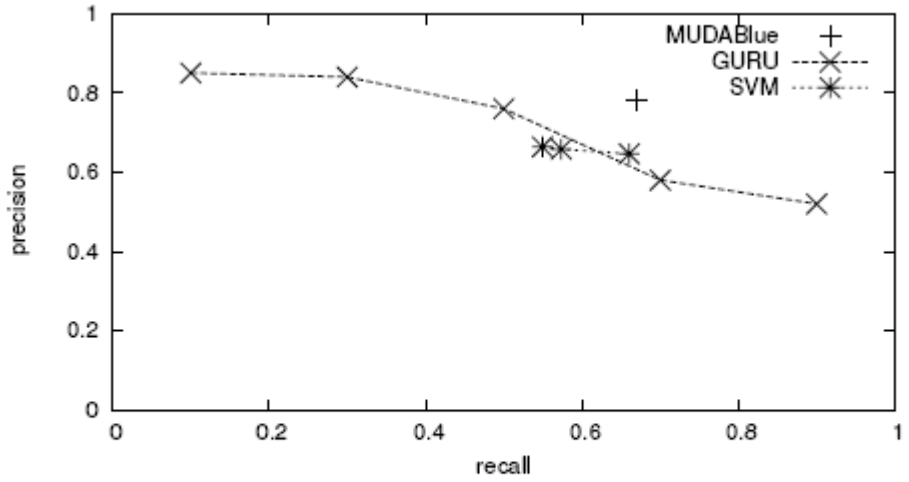


Figure 7. Precision-recall graph

Output

- 41 projects
- 40 categories
- 18 agree with manual
- 8 new (libs etc)
- 14 are?

No.	Title of cluster	Software	# of tokens
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, hdrl_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, GtkWidget, PARAMS, codeset, gpointer, loaded_l10nfile, argz	boardgame/gbatnav-1.0.4, boardgame/gchch-1.2.1, editor/gnotepad+-1.3.3, editor/peacock-0.4	118
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
⋮			
40	clause, cinfo, pred, ci, Group, Np, word, X, A, tmp4	compilers/gprolog-1.2.3, database/postgresql-7.2.1, video-conversion/mjpgTools	6

Table 4. MUDABlue Result (excerpt)

Conclusions

- Strengths
  - Source code only, no predefined categories
  - Validated against manual classification
  - Integrated tool to help with navigation and understanding of categories
- Weaknesses
  - Named outputs are hard to interpret
  - Graphs could be too large
  - Why only 41 projects (it's automated)
  - [The] writing [is] poor