トップエスイー修了制作

自然言語処理技術の適用による ソフトウェア部品への自動タグ付け

東京大学　馬場 雪乃　ybaba@nii.ac.jp

開発における問題点

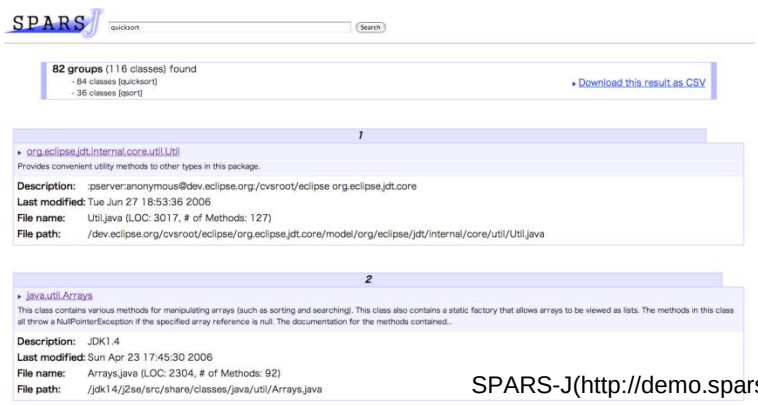
既存のソフトウェア資産からソフトウェア部品を再利用することで，高品質なソフトウェアを効率良く開発することが可能となる．しかし，再利用者が必要とする部品を検索し，検索結果の一覧の中から望む部品を探し出すためには，結果一覧で出力された各部品のソースコードを詳しく見た上で，どの部品が適切なのかを検討する必要がある．

手法・ツールの提案による解決

ソフトウェア部品に対して，その特徴を表すキーワードをタグとして自動的に付与する手法を提案する．タグは，部品の処理内容や対象のオブジェクト，使用しているアルゴリズム等を表す．このタグを部品の検索結果一覧に付与することで，結果一覧のみでも部品の詳細についての情報を得ることが可能となり，求める部品を探し出すことが容易となる．

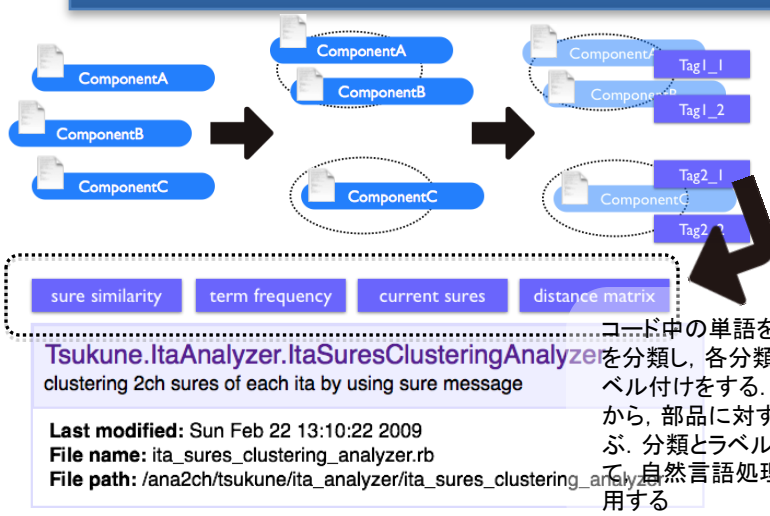
既存の検索システム

検索結果一覧には，クラス名，簡単な説明等しか表示されず，詳細を知るためにはコードの中身を見る必要がある



SPARS-J(http://demo.spars.info/)

タグ付けによる解決



コード中の単語を用いて部品を分類し，各分類に対してラベル付けをする．そのラベルから，部品に対するタグを選ぶ．分類とラベル付けにおいて自然言語処理技術を適用する

処理手順

1. ソースコード中の単語を抽出，出現頻度を数える
   このとき，語の出現位置に応じて重みを与える
2. 部品を，確率的な文書分類手法(LDA)を用いて分類

topic1(CAT)

kitten	0.16
milk	0.08
meow	0.07
...	...

topic2(DOG)

puppy	0.25
bark	0.13
bone	0.12
...	...

➡ d1 TXT

topic1	0.61
topic2	0.10
...	...

[Blei 03] Laten Dirichlet Allocation

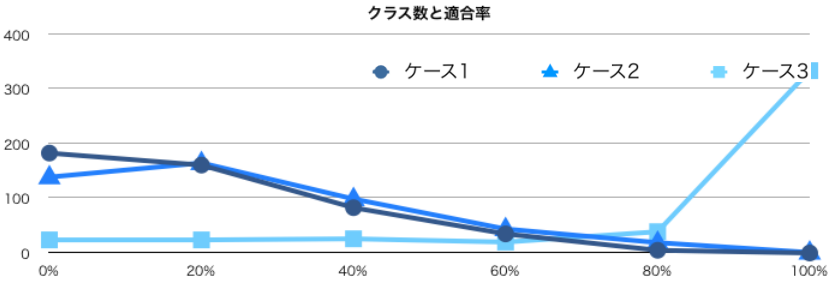
3. 各部品分類に，LDAによる文書分類に対して適切なラベルを付与する手法[Mei 07]を適用
   1. 理解容易性: ~~get~~, ~~posts~~
   2. 関連性: ~~art~~
   3. 網羅性: get posts ~~by time~~
   4. 独立性: ~~id~~, ~~initialize~~

   ➡ ○ get posts
4. 各部品が分類に属する確率と，各分類に対するラベルの適切度を用いて，部品に付与するタグを決定する

[Mei 07] Automatic Labeling of Multinomial Topic Models

評価

Rubyソースコード (クラス数467, LOC 約24,000)に手法を適用．人手で作成した正解データとの適合率で評価



各部品ごとに5つのタグを出力し，部品ごとの適合率を算出．図は，適合率が0%, 20%, …, 100%となったクラスの数を示す

	ケース1	ケース2	ケース3
平均適合率	19.36%	24.67%	84.03%

ケース1: 出力結果と正解のタグが完全に一致している時を正解とした場合
ケース2: タグの順序が逆でも正解とした場合
ケース3: 2語のタグのうち，1語が正解に含まれていれば正解とした場合

NII 国立情報学研究所 National Institute of Informatics
トップエスイー　～サイエンスによる知的のものづくり教育プログラム～
文部科学省 振興調整費産学融合先端ソフトウェア技術者養成拠点の形成