ソフトウェア成果物の設計根拠の抽出法

山内 崇 (310119)

北陸先端科学技術大学院大学 情報科学研究科

2005 年 2 月 10 日

キーワード: オープンソース・ソフトウェア開発, CVS, 電子メール, ベクトル空間モデル, ヒューリスティック.

1 背景と目的

オープンソース・ソフトウェア開発では、成果物と変更履歴（コミットログ）の管理にバージョン管理システム CVS が、開発者間のコミュニケーションにメーリングリストが多く用いられている。CVS リポジトリには、変更履歴と共に変更した理由が記録される。メーリングリストで行われた変更に関する討議はメーリングリスト・アーカイブに残る。開発者は、CVS の変更履歴やメーリングリスト・アーカイブを閲覧することにより、過去の変更理由やその変更に至るまでの討議を理解できる。

しかし、ソフトウェアの開発期間が長くなると CVS リポジトリやメーリングリスト・アーカイブには大量の情報が蓄積されるため、開発者が必要な情報を探すのが困難になる。FreeBSD の開発者向けメーリングリストを例に挙げると、一ヶ月に約 2500 通のメールが投稿される。そのため、開発者がすべてを読むのは容易ではなく、どのメールに何が書いてあったか覚えておくことも難しい。

そこで本研究では、開発者が変更履歴に対応する討議の検索を容易にするために、CVS リポジトリとメーリングリスト・アーカイブを用い、CVS リポジトリ中のコミットログからメーリングリスト・アーカイブの中の対応するメールスレッドを検索する手法を提案する。

2 提案手法

提案する任意のコミットログに対応するメールスレッドの検索手法を以下に示す。

- **ヒューリスティックによる検索対象の絞り込み**
  検索精度を上げるために、開発者が手作業で探し出すときの手がかりをヒューリス

ティックとして定義し、これを用いて検索対象を絞り込む。
本研究で定義したヒューリスティックを以下に示す。

1. コミットした人が投稿しているメールスレッドのみを残す
2. コミットの日付から前後 n 日以内に投稿されたメールスレッドのみを残す

- **ベクトル空間モデルによる検索処理**
  ベクトル空間モデルは、検索質問と検索対象をベクトルであらわすことにより類似検索を実現する情報検索技術である。本研究では、検索質問にコミットログ、検索対象にメーリングリスト・アーカイブから取り出したメールスレッドを用いる。

ヒューリスティックとベクトル空間モデルを用いた検索処理は以下の手順で行う。1) 検索対象の各メールスレッドの本文から不要語と接尾辞を除去する。2) 検索質問のコミットログから不要語と接尾辞を除去し、残った単語を索引語にする。3) コミットログの索引語ベクトルと各メールスレッドの索引語ベクトルを求める。4) コミットログの索引語ベクトルと各メールスレッドの索引語ベクトルの角度を求めることにより、コミットログとメールスレッドの類似度を求める。この類似度の降順に各メールスレッドを並び替える。

3 評価

“ヒューリスティックによる検索対象の絞り込み”と “ベクトル空間モデルによる検索処理”を、実際の開発プロジェクトに適用する実験を行った。実験対象は、2003 年 12 月から 2004 年 11 月までの 1 年間に FreeBSD CURRENT のメーリングリストへ投稿されたメール（メール総数: 28273, スレッド数: 8480）である。まず、実験対象から、事前に手作業によって “コミットログに対応するメールスレッド”の組み合わせを 12 組探し出した。

以下の 3 通りの場合について各コミットログを基にメールスレッドの検索を行い対応するメールスレッドの順位を調べた。

- すべてのメールスレッド（ベクトル空間モデルのみ）
- コミットした人が投稿しているメールスレッド
- コミットの日付から前後 n 日以内に投稿されたメールスレッド
  (n は 31, 15, 7 に変えて試行する)

ベクトル空間モデルのみを用いた場合は、3 組が 50 位以内に入ったが、500 位以下が 4 組と順位の差が大きく、ベクトル空間モデルのみでは検索精度が十分ではない。

コミットした人が投稿しているメールスレッドを検索対象にした場合は、ベクトル空間モデルのみの場合と比べて 5 組の順位が向上した。中には、1285 位から 11 位に上がったものもある。しかし、7 組は検索対象から外れた。

コミットの日付から前後 n 日以内に投稿されたメールスレッドを対象にした場合は、前後 31 日のとき、6 組が 50 位以内に入り、ベクトル空間モデルのみの場合と比べて 9 組の

順位が向上し、1 組が検索対象から外れた。前後 15 日のとき、ベクトル空間モデルのみの場合と比べて 6 組の順位が向上し、4 組は検索対象にから外れた。前後 7 日のとき、ベクトル空間モデルのみの場合と比べて 4 組の順位が向上し、6 組は検索対象から外れた。このヒューリスティックでは、n が小さくなるにつれて検索対象から外れる割合が大きくなる。

4　まとめと今後の課題

本研究ではヒューリスティックとベクトル空間モデルによる検索手法を提案した。実験の結果、ヒューリスティックを用いた場合に検索対象に入らなかった例が見られたものの、順位が上がっている例が確認できた。これにより、開発者が変更履歴に対応する討議の検索を容易にすることがある程度達成できた。今後の課題として、今回の実験ではベクトル空間モデルで類似度計算をする前にヒューリスティックを適用したが、これ以外の適用方法も検討する必要がある。