Title	ソフトウェア成果物の設計根拠の抽出法
Author(s)	山内，崇
Citation	
Issue Date	2005-03
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/1859
Rights	
Description	Supervisor:落水　浩一郎，情報科学研究科，修士

Technique of extracting the design basis of software artifact

Takashi Yamauchi (310119)

School of Information Science,
Japan Advanced Institute of Science and Technology

February 10, 2005

Keywords: Open-source-software development, CVS, E-mail, vector space model, heuristics.

1 Background and Purpose

The developer uses version management system CVS to manage the result and the change history (commit log) for open source software development. And, they use the mailing list for the communications between the developers. The CVS repository records the reason changed with the change history.

The developer discusses reason of the change by the mailing list. Then, the discussion remains in the archive. They can understand the discussion until a past reason for the change and the changing from inspecting the change history and the mailing list archive of CVS.

However, the developer's looking for information becomes difficult. The reason is that the barrage of information is accumulated in the CVS repository and the mailing list archive when the development period of software becomes long. For example, about 2500 mails flow to the mailing list for the development of FreeBSD in a month. Therefore, the developer cannot read all mails. And, they are also difficult to remember all mails.

In this paper, I propose the technique for retrieving the mail thread of the commit log in the CVS repository. So that the developer may easily retrieve the discussion corresponding to the change history.

2 Proposal

The following items are the methods of retrieving a mail thread related to an arbitrary commit log.

- **The object of the retrieval is narrowed with heuristic**
  To improve the accuracy of the retrieval, I defined the clue of the retrieval.

  enumerate heuristic defined by this research as follows.

  1. Only the mail thread that the person who had committed it contributed
  2. Only the contributed mail thread : within the day of n from the date of committing.

- **Retrieval processing by vector space model**
  Vector space model is the retrieval question and the retrieval object are described by the vector. It is an information retrieval technology for a similar retrieval.

The following procedures are the retrieval processing that uses heuristic and the vector space model. 1)The system remove an unnecessary word and the suffix from the text of each mail thread. to be retrieved. 2)The system removes an unnecessary word and the suffix from the commit log of the retrieval question, and makes a remaining word the index word. 3)The system calculates the index word vector of the commit log and the index word vector of each mail thread. 4)The system calculates two vectors, and requests a similar level of the commit log and the mail thread. The system permutes each mail thread in the descending order of a similar level.

3 Evaluation

I did the experiment that use “heuristics” and “vector space model”to the software development project. I used E-mail (December,2003 to November,2004; Total of mail: 28273; Total of mail thread: 8480) of the mailing list of FreeBSD CURRENT. First, I searched out 12 combinations of mail

thread "related to " commit log by the hand work from the experiment object.

I retrieved the mail thread based on each commit log and examined the rank of the mail thread that was related to. For about the following three kinds of cases.

- All mail threads(Only the vector space model)
- Mailing list where name of committer exists
- The last date of mail thread before and behind The date of committing is in during n day.
  (n is changed into 31, 15, and 7 and tried)

When I handled only the vector space model, three pairs were within 50th place. And, the difference of those ranks is great because four pairs are in the 500th place following. Therefore, the Case sensitive matching only of the vector space model is insufficient.

When the mail thread which the person who committed has contributed was made applicable to reference, the ranking of 5 pairs improved compared with the case of only a vector space model. There are some which went up from 1285 grades to 11 grades in inside. However, it separated from 7 pairs for reference.

When aimed at the mail thread contributed within the order n day from the date of a commitment, at the time of about 31 day, 6 pairs entered within 50 grades, the ranking of 9 pairs improved compared with the case of only a vector space model, and 1 pair separated for reference. the time of about 15 day the case of only a vector space model comparing the ranking of 6 pairs improving 4 pairs the candidate for reference since it separated At the time of about 7 day, the ranking of 4 pairs improved compared with the case of only a vector space model, and it separated from 6 pairs for reference. In this heuristics, the rate from which it separates for reference becomes large as n becomes small.

4 Consideration

In this research, the reference technique by heuristics and the vector space model was proposed. Although the example which did not go into the

candidate for reference was seen as a result of the experiment when heuristics was used, the example which ranking is going up has been checked. Thereby, it has attained that a developer makes easy reference of the debate corresponding to a change history to some extent. As a future subject, this experiment also needs to consider the application methods other than this, although heuristics was applied before carrying out the degree calculation of similar by the vector space model.