Title	汎用連想計算エンジン GETA を用いたスパム判別に関する研究
Author(s)	石黒，雄輔
Citation	
Issue Date	2008-09
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/4756
Rights	
Description	Supervisor:小川瑞史，情報科学研究科，修士

Finding SPAM Mails
with the associative search engine GETA

Yusuke Ishikuro (0610006)

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

August 8, 2008

Keywords: Statistical method, SPAM filtering, association calculation.

By the spread of Internet, convenience rose with various aspects of the life. E-mails spread now as a communications infrastructure supporting social activity. However, it is said that a SPAM mail holds 90%-95%of the E-mail and becomes a big problem now. As for the existing SPAM mail filter is heuristics filter and Bayesian Filter. heuristics filter is based on the rule by human work, and Bayesian Filter are representative. Enormous labor is necessary by this method to describe all the header analysis rules with hands as problems, and an administrative burden gets heavy. In addition, as for Bayesian Filter, training is necessary for precision improvement, and correspondence takes time for the SPAM mail of a new type. Therefore, by the present, the commercial SPAM mail filter has many cases that compound filters put technology together.

In this research, the possibility of the SPAM distinction by a simple statistical method is pursued from a new angle. The construction of the SPAM filter by the associative calculation only using the text (natural language portions of the text and `Subject`) of e-mail is proposed, and the possibility as a primary filter (or the last stage filter) of the SPAM distinction is explored. In that case, high-speed associative search engine GETAis used on the assumption that practical use of the maximum existing tool. With associative calculation, it is regarded as a multiplex set of a document and a word, and the degree of similar between a word group-document and a

document group-word is calculated only with the frequency of appearance in the document of a word. By this, various associative calculation between a document-document, between word-words, and between document-words is attained.

The association SPAM filter constitutes it after preprocessing of the single language extraction from an email by putting threshold distinction together for an association calculation. The experiment system for evaluating the possibility of an associative SPAM filter consists of two, pretreatment mechanisms, such as the morphological-analysis machine Mecab, and the associative calculation engine GETA. As for the appearance frequency in the document of the word, it is expressed row ingredient in matrix WAM which assumes word frequency vector in the document, column ingredient the appearance frequency vector in the document of the word in GETA. The threshold distinction mechanism is nonimplement under the present conditions. In order to clarify possibility of threshold value distinction, it experimented and the experiment was tried by simple threshold function presumed by only the linear regression about the single degree of similar (real numerical value) and the number of difference words of an e-mail text to a reference SPAM mail group. The data set of an experiment is a reference SPAM mail group alpha is about 80,000 mails, a reference SPAM mail group beta is about 150,000 mails and Required mail group gamma is About 1,000 mails were prepared.

The classification of the mail set, the abstraction of each mail, and the selection of the evaluation function were taken up as a comparison axis in the experiment. Classification of an e-mail set here refers to division of an e-mail set, and the processing to the portion which abstraction of e-mail extracted which portion among the constituent factors of an e-mail simple substance, and was extracted is said. With classification of the mail, it classified in English, and Japanese with abstraction of the mail the filtering due to part of speech (Japanese), demarcation of presence of the repetition and retention, removal and text and Subject of the sign it did. In addition comparison of performance function went with TF and SMARRT measure. When it experiments with these comparison axes, a SPAM mail group and a required e-mail group dissociate notably under suitable abstraction by classifying a reference SPAM mail group to Japanese mail and English mail.

It is necessary to perform an enormous calculation for the setup of the association SPAM filter. There is these two phases of processing and becomes it from making of WAM which is word frequency information making of the email group and an element of GETA. Although the former is heavy processing in the present mounting, once it carries out, additional processing of new reference SPAM mail will be gradual, and it will not become a big problem.In addition, the making of the line is not possible progressively, but is finished in around dozens of seconds if these are SPAM mails of around 100,000. If these setups have been performed, the processing of an associative SPAM filter to input mail is 0.3 or less second, and can be performed efficiently. The present threshold function is restricted to the simple thing of only linear regression with the number of difference words of an e-mail text to the single degree of similar to a spam mail group (real numerical value). The improvement by the SPAM distinction which combined the construction of more suitable threshold function and two or more WAMs is a future work.