Abstract

Spam became a big problem to the society. Some spammers are using templates for sending spam. To send a particular promotion they create some template and merge the details of receivers with the template. Similarities can find among these mails and easily ignore the forthcoming spam. Most high-volume spam is sent using tools those randomizes parts of the message - subject, body, sender address etc. The general form of the template that the spammer is using can often guess by inspecting the features of messages. Most of the spam filters are either rule based models or Bayesian models. The main objective in this paper is to find out semantic distance and evaluate the applicability of the two information retrieval techniques, Simple Vector Space Models (VSM) and VSM using Rocchio Classification in the spam context. Both methods are using cosine similarities to identify the spam.

References

- G. Salton, A. Wong, and C. S. Yang, "A Vector Space Model for Automatic Indexing,"
- J. J. Rocchio. Relevance feedback in information retrieval. In G. Salton, editor, The
- Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze: An Introduction to
- Wilfried N. Gansterer, Andreas G. K. Janecek Robert Neumayer, Spam Filtering Based
on Latent Semantic Indexing
- Tuomo Korenius, Jorma Laurikkala, Martti Juhola, On principal component analysis,
cosine and Euclidean measures in information retrieval, Information Sciences, Volume 177,
Issue 22, 15 November 2007, Pages 4893-4905, ISSN 0020-0255
- Congnan Luo, Yanjun Li, Soon M. Chung, Text document clustering based on neighbors,
Data & Knowledge Engineering, Volume 68, Issue 11, November 2009, Pages 1271-1288,
ISSN 0169-023X, 10.1016/j.datak.2009.06.007.
- Angel R. Martinez, Data Mining of Text Files, In: C.R. Rao, E.J. Wegman and J.L. Solka,
Editor(s), Handbook of Statistics, Elsevier, 2005, Volume 24, Pages 109-131, ISSN 0169-7161,
- Thamarai Subramaniam, Hamid A. Jalab and Alaa Y. Taqa , Overview of textual
1869-1882, 4 October, 2010

Index Terms
Computer Science Communications

Keywords
Spam vector space models Rocchio classification cosine similarity