Dinesh Rathi
School of Library and Information Studies, University of Alberta, Edmonton,
Michael B. Twidale
Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign, IL, USA

Ditch the Smileys: Customizing a Stopword List for Email-based Data

Abstract: The study uses grounded theory approach to develop different categories of stopwords leading to the creation of a stopword list for email-based data. The finding of the study will contribute in better understanding of email as data and developing better algorithms which could automatically remove specific category of stopwords.

Résumé : Cette étude se base sur la théorie à base empirique pour développer différentes catégories de mots vides qui seront utilisés pour créer une liste aux fins d'analyse des données issues de courriels. Les résultats permettront une meilleure compréhension des courriels comme source de données et la création de meilleurs algorithmes de suppression automatique de catégories précises de mots vides.

Text mining typically involves a variety of decisions including algorithm selection, parameter setting (Keogh et al., 2004; Xu and Wunsch, 2005) and the creation of a stopword list. However, the latter is often treated as a tedious but necessary part of the text mining process and is rarely the subject of systematic investigation, unlike, say the validity or utility of the results of the mining process, or of the mining algorithm. As a result, stopword list creation can receive little attention. It is tempting to just re-use a preexisting stopword list. However there are problems with this approach due to the context dependency of language use. Chakrabarti et al. (1997, 1998) argue that “it is tricky to hand-craft the stop words out of domain knowledge of the language; ‘can’ is frequently included in stopword lists, but what about a corpus on waste management?”. Similarly, Silva and Ribeiro (2003) note that “words that are to be included in the stop-list are language and task dependent”.

This work is part of a larger project looking at the text mining of email data, and in particular email relating to issues around technical support. With the widespread adoption of internet technologies in business and everyday life, email has become one of the most important modes of communication for a large number of organizations. For example, computer companies such as Dell are using an email-based system to provide support to their customers. These email-based systems have multiple benefits. They are often adopted as a cost saving measure to manage complex exchanges between people. Technical support conversations can involve multiple exchanges, and even multiple people through escalation before the problem is resolved. Additionally, the nature of email means that there is a trace of the interaction remains, and solutions can potentially be recycled through copy and paste of earlier related emails. Furthermore, the email archive can be analyzed, looking for patterns in the nature of problems that can inform not just future technical support problem solving, but greater efficiency, and indeed identify areas for improvement of the products being supported.

Given the sheer size of typical email archives, these benefits can only be realized by using text mining techniques. The quality of data has implication on the text mining results and hence effort has been made to improve the efficiency of algorithms by removing the non-informative words using the stopword list (Persin, 1994). Several papers have discussed the importance of the quality of data on text mining results (Cooley et al., 1999; Redman, 1998; Jung, 2004; Tayi and Ballou, 1998; Zhang et al., 2004). Data quality is improved by removing so called noise words from the data in the pre-processing stage.

Work on email text mining is relatively focused on classification, filtering, and summarization (Tang et al., 2005) which builds on a much larger body of research on other kinds of text mining (Blake, 2010) Our work focuses on identifying the characteristics of email as data to develop a framework for creating a stopword list for email-based data. We use a grounded theory (Glaser and Strauss, 1967; Corbin and Strauss, 1990) approach to develop a customized stopword list for use in clustering and classification experiments.

A noise words list, also known as a stopword list, has a set of terms (words) that have no usable information. Noise words create poor index terms if not removed from the data (van Rijsbergen, 1979; Sinka and Corne, 2003). Manco et al. (2002) argued that "removal of stopword has the advantage of making the selection of the candidate index terms more efficient and reducing the size of the index structure considerably". The authors defined two categories of stopwords: explicit stopwords and implicit stopwords. Explicit stop words are the standard stopword list such as Smart Experiment Stopword list (Salton, 1971).

Silva and Ribeiro (2003) in their research work created their own stopwords list which was made up of an existing set of general stopwords and the words that appeared in few documents (because those "words are unlikely to represent a category"). Koprinska et al. (2007) in their work on email classification did not use the standard stopword list but removed the words from the data that either appeared only once or were longer than 20 characters.

As argued in the literature, the task should define the customization of the stopword list and this could include both explicit and implicit stopwords so that it is reflective of the data used in this research. The Smart Experiment list (http://www.lextek.com/manuals/onix/stopwords2.html) is an old list and was made from the text which had no or very few spelling errors as compared to text in email. The use of email has led to the advent of new terms. For example there are short forms of regular words such as 'thx' for 'thanks'. Another classic email category is the emoticon, including the smiley: :-). For the purposes of text analysis the smiley is a word. And for certain kinds of analysis it should be treated as a stopword. In addition, email communication and the Internet added new terms which were rare or non-existent before such as "www" or "http". The new stopword list needs to account for such terms.

The creation of a stopword list for conducting experiments on email-based data was a complex decision making process. There were two key reasons for this complexity; first, there is limited literature related to stopword lists for email based data (as the majority of the literature in this domain has focused on other aspects of the text mining process,

relatively ignoring the stopword issue); and second, email as genre and as data has many features which are very interesting, but which make email substantially different from the other text data (e.g., articles, Reuters news, etc.) which are more typically used in text mining. Some of the reasons for the complexity of email-based data include (but are not limited to):

- Email largely consists of unstructured text (content or message) except for headers which is structured and may contain noise, URLs/Hyperlinks (Moreale and Watt, 2002).

- Email is often cursorily written and is different in composition from normal text documents. For example, Reuters' documents are syntactically well formed (Busemann et al., 2000). It may contain badly cased words, intentional/unintentional misspelling (short cuts of terms) or grammatical errors (Tang et al., 2005), classic keyboarding typos, non-stable vocabulary (de Vel et al., 2001), jargon (Busemann et al., 2000), "lexical surrogates for vocalizations" and "visual arrangement of text characters into emoticons" (Corney et al., 2002).

- Email can be in different formats such as HTML, plain and rich text (Tang, et al., 2005).

- Email messages are not very long as compared to research paper / news articles i.e. email generally has few sentences or paragraph (de Vel et al., 2001; Busemann et al., 2000).

Our research was conducted using grounded theory to develop categories of stopwords for email-based data and then to identify the specific terms to be included in each category to create a stopword list. The study was conducted on an email data archive of a computer systems help desk. The data collected for the research was 4 years of email data and had over 15,000 emails. The email data included user problems, spam, reports, etc. Initially, a couple of pilot studies were conducted on a small dataset to gain insight into the terms and their frequency distribution, and to develop initial categories of a stopword list. The categories included an explicit list, names/alias (email-id) list, numeric list, alpha-numeric list, low frequency list, etc. These categories facilitated the inclusion or exclusion of words from the customized stopword list from the large dataset that was used in clustering and classification experiment. In all categories, over 27,000 unique terms were identified for the customized stopword list.

The study provides the framework for developing the implicit stopword list which could be applied not only to email-based data but also to other kinds of data. The creation of categories from this study will not only contribute towards the understanding the characteristics of email as data and but also contribute towards the development of better classification algorithms, such as automatic removal of specific category stopwords as required by the research work, by incorporating the findings in the algorithm's code. The findings will also help other researchers who are researching email-based data for other purposes. For example; researchers, who want to develop a social networking model will consider names/aliases as important data for their research while other categories will be noise for their analysis (Rathi et al., 2007). Hence, the researchers, by using the categories identified in this study, will be able to develop their customized stopword list more efficiently.

The study is relevant to the two areas of the conference theme: knowledge management (KM) aspects of Knowledge and Information Management theme and information retrieval (IR) aspect of Human-Information Interaction theme. For example, text mining is an important KM tools (Marwick, 2001) and the finding of this study will help researchers working in this domain.

References

Busemann, S., Schmeier, S. and Arens, R. G. 2000. Message Classification in the Call Center, Proceedings of the Sixth conference on Applied Natural Language Processing, 158–165.

Blake, C. 2010. Text Mining, ARIST, Vol 45 (To appear)

Chakrabarti, S., Dom, B., Agrawal, R. and Raghavan, P. 1997. Using Taxonomy, Discriminants, and Signatures for Navigating in Text Databases, Proceedings of the 23rd International Conference on Very Large Databases, 446–455.

Chakrabarti, S., Dom, B., Agrawal, R. and Raghavan, P. 1998. Scalabe Feature Selection, Classification and Signature Generation for Organizing Large Text Databases into Hierarchical Topic Taxonomies, The VLDB Journal, Springer-Verlag, 7, 163–178.

Cooley, R., Mobasher, B., and Srivastava, J. 1999. Data Preparation for Mining World Wide Web Browsing Patterns, Knowledge Information System, 1-27.

Corbin, J. and Strauss, A. 1990. Grounded Theory Research: Procedures, Canons, and Evaluative Criteria, Qualitative Sociology, 13(1), 3-21.

Corney, M., de Vel, O., Anderson, A., and Mohay, G. 2002. Gender-preferential Text Mining of E-mail Discourse, The 18th annual Computer Security Applications Conference (ACSAC2002).

de Vel, O., Corney, M. and Mohay, G. 2001. Mining E-Mail Content for Author Identification Forensics, SIGMOD Record, ACM Press, 30(4), 55–64.

Glaser, B., and Strauss, A. 1967. The Discovery of Grounded Theory, Chicago: Aldine Publishing Company.

Jung, W. 2004. An Investigation of the Impact of Data Quality on Decision Performance, Proceedings of the 2004 International Symposium on Information and Communication Technology (ISICT ’04), 166–171.

Keogh, E., Lonardi, S. and Ratanamahatana, C. A. 2004. Towards Parameter-Free Data Mining, Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 206-215.

Koprinska, I, Poon, J., Clark, J. and Chan, J. 2007. Learning to Classify Email, Information Science, 177, 2167–2187.

Manco, G., Masciari, E., Ruffolo, M. and Tagarelli, A. 2002. Towards An Adaptive Mail Classifier, Proceedings of Italian Association for Artificial Intelligence Workshop.

Marwick, A. D. 2001. Knowledge Management Technology, IBM Systems Journal.

Moreale, E. and Watt, S. 2002. Organisational Information Management and Knowledge Discovery in Email within Mailing Lists, In H. Yin et al. (Eds.), Intelligent Data Engineering and Automated Learning, Lecture Notes in Computer Science, 2412/2002, 217-224, Berlin / Heidelberg:Springer-Verlag.

Persin, M. 1994. Document Filtering for Fast Ranking, Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval, Dublin, Ireland, 339-348.

Rathi, D., Twidale, M. B. Singh, V. and Jones, M. C. 2007. Your Mark is My Dirt: Impact of Email Signatures on Decision Making, The 9th International Conference on Decision Support Systems, India.

Redman, T. C. 1998. The Impact of Poor Data Quality on the Typical Enterprise, Communications of the ACM, ACM Press, 41(2), 79–82.

Salton, G. 1971. The SMART Retrieval System—Experiments in Automatic Document Processing, Upper Saddle River, NJ, USA: Prentice-Hall, Inc..

Silva, C and Ribeiro, B. 2003. The Importance of Stop Word Removal on Recall Values in Text Categorization, Proceedings of the International Joint Conference on Neural Networks, 3, 1661-1666.

Sinka, M. P., and Come D. W. 2003. Evolving Better Stoplists for Document Clustering and Web Intelligence, Proceedings of the 3rd Hybrid Intelligent Systems Conference, Australia, IOS Press.

Tang, J., Li, H., Cao, Y. and Tang, Z. 2005. Email Data Cleaning, Proceedings of the eleventh ACM SIGKDD international conference on Knowledge Discovery in Data Mining, Chicago, Illinois, USA, 2005, 489–498.

Tayi, G. K. and Ballou, D. P. 1998. Examining Data Quality, Communications of the ACM, ACM Press, 41(2), 54–57.

Van Rijsbergen, C. J. 1979. Information Retrieval, Newton, MA: Butterworth-Heinemann.

Xu, R. and Wunsch, D. 2005. Survey of Clustering Algorithms, IEEE Transactions on Neural Networks, 16(3), 645-678.

Zhang, S., Zhang, C., and Yang, Q. 2004. Information Enhancement for Data Mining, IEEE Intelligent Systems, March-April, 19(2), 12–13.