[2] N. Ailon, Z. S. Karnin, E. Liberty, and Y. Maarek. Threading machine generated email. In *Proc. of the 6th ACM International Conference on Web Search and Data Mining*, pages 405–414, 2013.

[3] J. Alpert and N. Hajaj. We knew the web was big. *The Official Google Blog*, 21, July 25 2008.

[4] I. Androutsopoulos, J. Koutsias, K. V. Chandrinos, and C. D. Spyropoulos. An experimental comparison of naive bayesian and keyword-based anti-spam filtering with personal e-mail messages. In *Proc. of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 160–167, 2000.

[5] A. Arasu and H. Garcia-Molina. Extracting structured data from web pages. In *Proc. of the ACM SIGMOD International Conference on Management of Data*, pages 337–348, 2003.

[6] R. Bekkerman. Automatic categorization of email into folders: Benchmark experiments on Enron and SRI corpora. *Computer Science Department Faculty Publication Series, University of Massachusetts, Amherst*, (218), 2004.

[7] E. Blanzieri and A. Bryl. A survey of learning-based techniques of email spam filtering. *Artificial Intelligence Review*, 29(1):63–92, 2008.

[8] G. Caruana and M. Li. A survey of emerging approaches to spam filtering. *ACM Computing Surveys*, 44(2):1–27, 2012.

[9] L. A. Dabbish and R. E. Kraut. Email overload at work: An analysis of factors associated with email strain. In *Proc. of the 20th Conference on Computer Supported Cooperative Work*, pages 431–440, 2006.

[10] D. Di Castro, L. Lewin-Eytan, Y. Maarek, R. Wolff, and E. Zohar. Enforcing k-anonymity in web mail auditing. In *Proc. of the 9th International Conference on Web Search and Data Mining*, to appear, 2016.

[11] C. Hachenberg and T. Gottron. Locality sensitive hashing for scalable structural classification and clustering of web documents. In *Proc. of the 22nd ACM International Conference on Information & Knowledge Management*, pages 359–368, 2013.

[12] A. Kannan, K. Kurach, S. Ravi, T. Kaufmann, A. Tomkins, B. Miklos, G. Corrado, L. Lukács, M. Ganea, P. Young, and V. Ramavajjala. Smart reply: Automated response suggestion for email. *CoRR*, abs/1606.04870, 2016.

[13] S. Kiritchenko and S. Matwin. Email classification with co-training. In *Proc. of the Conference of the Center for Advanced Studies on Collaborative Research*, pages 301–312, 2011.

[14] A. Kulkarni and T. Pedersen. Name discrimination and email clustering using unsupervised clustering and labeling of similar contexts. In *Proc. of the 2nd Indian International Conference on Artificial Intelligence*, pages 703–722, 2005.

[15] N. Kushmerick. *Wrapper induction for information extraction*. PhD thesis, University of Washington, 1997.

[16] A. H. Laender, B. A. Ribeiro-Neto, A. S. da Silva, and J. S. Teixeira. A brief survey of web data extraction tools. *ACM Sigmod*, 31(2):84–93, 2002.

[17] J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the dynamics of the news cycle. In *Proc. of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 497–506. ACM, 2009.

[18] D. D. Lewis and K. A. Knowles. Threading electronic mail: A preliminary study. *Information Processing & Management*, 33(2):209–217, 1997.

[19] H. Li, D. Shen, B. Zhang, Z. Chen, and Q. Yang. Adding semantics to email clustering. In *Proc. of the 6th International Conference on Data Mining*, pages 938–942, 2006.

[20] G. A. Miller. WordNet: A lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995.

[21] P. Pantel and D. Lin. Spamcop: A spam classification & organization program. In *Proc. of the AAAI Workshop on Learning for Text Categorization*, pages 95–98, 1998.

[22] S. Sarawagi. Automation in information extraction and integration. In *Tutorial of the 28th Internationl Conference on Very Large Databases*, 2002.

[23] Y.-C. Wang, M. Joshi, W. W. Cohen, and C. P. Rosé. Recovering implicit thread structure in newsgroup style conversations. In *Proc. of the 2nd International Conference on Weblogs and Social Media*, pages 152–160, 2008.

[24] P. Weiner. Linear pattern matching algorithms. In *Proc. of the 14th Annual Symposium on Switching and Automata Theory*, pages 1–11, 1973.

[25] J. B. Wendt, M. Bendersky, L. Garcia-Pueyo, V. Josifovski, B. Miklos, I. Krka, A. Saikia, J. Yang, M.-A. Cartright, and S. Ravi. Hierarchical label propagation and discovery for machine generated email. In *Proc. of the 9th International Conference on Web Search and Data Mining*, 2016.

[26] L. Zhang, J. Zhu, and T. Yao. An evaluation of statistical spam filtering techniques. *ACM Transactions on Asian Language Information Processing*, 3(4):243–269, 2004.

[27] W. Zhang, A. Ahmed, J. Yang, V. Josifovski, and A. J. Smola. Annotating needles in the haystack without looking: Product information extraction from emails. In *Proc. of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2257–2266, 2015.