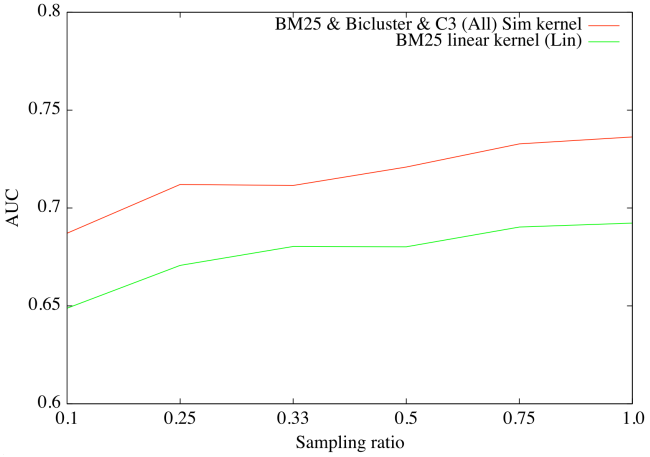

Figure 4: AUC as the function of the size of the training set, given as percent of the full3040, for the baseline BM25 with linear kernel and All with similarity kernel.

0.6657 (3.85%), both with variance 0.004 for ten independent samples. The robustness of the similarity kernel for small training sets is similar to BM25 with linear kernel, as seen in Fig. 4.

5. CONCLUSIONS

Over the C3 data sets, we gave a large variety of methods to predict quality aspects of Web pages, including collaborative filtering and methods that use evaluator and page metadata as well as the content of the page. We achieved best performance by our theoretically justified kernel method over the content of the page and C3 attributes. Our results are promising in that our AUC is stable over 0.7 for all aspects with "presentation" surpassing 0.8. The support vector regression methods also perform with error less than one on the range of 0–4.

6. REFERENCES

[1] J. Abernethy, O. Chapelle, and C. Castillo. WITCH: A New Approach to Web Spam Detection. In *Proc. 4th AIRWeb*, 2008.

[2] R. M. Bell and Y. Koren. Lessons from the netflix prize challenge. *ACM SIGKDD Explorations Newsletter*, 9(2):75–79, 2007.

[3] J. Bennett and S. Lanning. The netflix prize. In *KDD Cup and Workshop in conj. KDD 2007*, 2007.

[4] C. Castillo, K. Chellapilla, and L. Denoyer. Web spam challenge 2008. In *Proc. 4th AIRWeb*, 2008.

[5] C. Castillo, D. Donato, L. Becchetti, P. Boldi, S. Leonardi, M. Santini, and S. Vigna. A reference collection for web spam. *SIGIR Forum*, 40(2):11–24, December 2006.

[6] C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[7] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20, 1995.

[8] B. Z. Daróczy, D. Siklósi, and A. Benczúr. SZTAKI @ ImageCLEF 2012 Photo Annotation. In *Working Notes of the ImageCLEF 2012 Workshop*, 2012.

[9] K. Dave, S. Lawrence, and D. Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proc. 12th WWW*, pages 519–528. ACM, 2003.

[10] I. Dhillon, S. Mallela, and D. Modha. Information-theoretic co-clustering. *Proc. 9th SIGKDD*, pages 89–98, 2003.

[11] M. Erdélyi, A. A. Benczúr, B. Daróczy, A. Garzó, T. Kiss, and D. Siklósi. The classification power of web features. *Internet Mathematics*, 10(3-4):421–457, 2014.

[12] M. Erdélyi, A. Garzó, and A. A. Benczúr. Web spam classification: a few features worth more. In *WebQuality 2011*. ACM Press, 2011.

[13] D. Fetterly and Z. Gyöngyi. Proc. 5th AIRWeb. 2009.

[14] J. Fogarty, R. S. Baker, and S. E. Hudson. Case studies in the use of roc curve analysis for sensor-based estimates in human computer interaction. In *Proc. Graphics Interface*, pp. 129–136, 2005.

[15] Z. Gyöngyi and H. Garcia-Molina. Spam: It's not just for inboxes anymore. *IEEE Computer Magazine*, 38(10):28–34, October 2005.

[16] T. S. Jaakkola and D. Haussler. Exploiting generative models in discriminative classifiers. *Advances in NIPS*, pp. 487–493, 1999.

[17] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[18] Y. Low, D. Bickson, J. Gonzalez, C. Guestrin, A. Kyrola, and J. M. Hellerstein. Distributed graphlab: a framework for machine learning and data mining in the cloud. *Proc VLDB*, 5(8):716–727, 2012.

[19] A. Olteanu, S. Peshterliev, X. Liu, and K. Aberer. Web credibility: Features exploration and credibility prediction. In *Advances in Information Retrieval*, pages 557–568. Springer, 2013.

[20] R. Pálovics, F. Ayala-Gómez, B. Csikota, B. Daróczy, L. Kocsis, D. Spadacene, and A. A. Benczúr. Recsys challenge 2014: an ensemble of binary classifiers and matrix factorization. In *Proc. Recommender Systems Challenge*, page 13. ACM, 2014.

[21] T. G. Papaioannou, J.-E. Ranvier, A. Olteanu, and K. Aberer. A decentralized recommender system for effective web credibility assessment. In *Proc. 21st CIKM*, pp. 704–713. ACM, 2012.

[22] J. C. Platt. Sequential minimal optimization: A fast algorithm for training support vector machines. Technical report, ADVANCES IN KERNEL METHODS - SUPPORT VECTOR LEARNING, 1998.

[23] A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet. simplemkl. *JMLR*, 9:2491–2521, 2008.

[24] S. Rendle, Z. Gantner, C. Freudenthaler, and L. Schmidt-Thieme. Fast context-aware recommendations with factorization machines. In *Proc. 34th SIGIR*, pp. 635–644. ACM, 2011.

[25] B. D. Ripley and F. P. Kelly. Markov point processes. *Journal of the London Mathematical Society*, 2(1):188–192, 1977.

[26] B. Schölkopf, C. J. C. Burges, and A. J. Smola, editors. *Advances in kernel methods: support vector learning*. MIT Press, Cambridge, MA, USA, 1999.

[27] J. Schwarz and M. Morris. Augmenting web pages and search results to support credibility assessment. In *Proc. SIGCHI*, pp. 1245–1254. ACM, 2011.

[28] D. Siklósi, B. Daróczy, and A. Benczúr. Content-based trust and bias classification via biclustering. In *Proc. Webquality*, pp. 41–47. ACM, 2012.

[29] G. Takács, I. Pilászy, B. Németh, and D. Tikk. Investigation of various matrix factorization methods for large recommender systems. In *Proc. 2nd KDD Workshop on Large-Scale Recommender Systems and the Netflix Prize Competition*, pages 1–8. ACM, 2008.

[30] Z. Zheng, H. Zha, T. Zhang, O. Chapelle, K. Chen, and G. Sun. A general boosting method and its application to learning ranking functions for web search. In *Advances in NIPS*, pp. 1697–1704, 2008.