The dataset used in our experiments is WEBSPAM-UK2006 [1], which is obtained from a crawler of the .uk domain. It includes 77.9 million pages belonging to 11402 hosts. There are total 1803 spam and 4411 normal hosts which have both content and transformed link features. Every host has 96 content features and 138 transformed link features. 138 transformed link-based features can be divided into five categories: Degree-related features, PageRank-related features, TrustRank-related features, Truncated PageRank-related features and Supporter-related features.

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