# 主要参考文献

[1] G. Salton and C. Buckley. Term-weighting approaches in automatic retrieval. Information Processing & Management, 24(5):513-523, 1988

[2] M. P. Jay and W. B. Croft, A language modeling approach to information retrieval, ACM SIGIR 1998

[3] Marc Najork, Allan Heydon, High-Performance Web Crawling, 2001

[4] Boldi, P., Codenotti, B., Santini, M., & Vigna, S. (2004). UbiCrawler: A scalable fully distributed Web crawler. Software - Practice and Experience, 34(8), 711-726.

[5] Lucene in action. Luceneinaction.pdf

[6] J. Wang and Lochovsky, F.H., Data-rich section extraction from HTML pages, Web Information Systems Engineering (WISE) 2002.,Dec. 2002

[7] Gerard Salton and Chris Buckley, Improving retrieval performance by relevance feedback, Journal of the American Society for Information Science, Pages: 288 -297 ,1990

[8] C. Buckley and E. Voorhees, Evaluating evaluation measure stability, SIGIR, 2000

[9] L. Page, S. Brin etc, The PageRank Citation Ranking: Bringing Order to the Web, 1998

[10] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. Journal of the ACM, 46(5–7):604–632, 1999

[11] Yanhong Li, Toward a qualitative search engine, IEEE Internet Computing, 1998

[12] S. Brin, L. Page. The anatomy of a large-scale hypertextual web search engine. In:7th International World Wide Web Conference Proceedings, Brisbane, Australia, 1998:107～117

[13] Barroso, Dean, Hölzle, Web Search for a Planet: The Google Cluster Architecture, IEEE Micro 2003

[14] Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung, The Google File System, 19th ACM Symposium on Operating Systems Principles, 2003

[15] Jeffrey Dean and Sanjay Ghemawat, MapReduce: Simplified Data Processing on Large Clusters, OSDI'04, 2004

[16] Cynthia Dwork, Ravi Kumar, Moni Naor, D.Sivakumar, Rank Aggregation Methods for the Web, WWW10, 2001

[17] Y. Yang and X. Liu, A re-examination of text categorization methods, SIGIR, 1999

[18] B. Florian, E. Martin, and X. Xiaowei, "Frequent term-based text clustering," ACM SIGKDD, 2002