Web Link Analysis: A Personal View

Ricardo Baeza-Yates
Yahoo! Labs
Barcelona, Spain

rbaeza@acm.org

Abstract

In this presentation we give a personal view of link analysis in the Web and the different applications of this technique, in particular ranking in Web search engines. We start with the different levels of link analysis [2], link statistics [4], and the structure of the link graph [8].
Second, we do a historical account with the early work of Bruza & Van der Weide [9], Marchiori [13], Joo & Myaeng [10] and Li [12] to the classic techniques of PageRank [14] and HITS [11], and their derivatives [2].
Third, we cover Web ranking based in link analysis [6]. Although in theory would be enough to apply directly a classical technique, in practice is not that simple. Important issues to consider are the different types of links such as self-links, external/internal links, and navigation links; as well as the different types of pages such as home pages, sink-pages, and page reputation. On the other hand, we also need to consider computational issues regarding off-line and on-line link analysis. But links are not alone, and the anchor text of them plays also an important role in Web search.
With the evolution of the Web, new factors affected link analysis. First, with advertising, the incentives for Web spam increased and link analysis was modified to take in account the length of the support chain of the links leading to a page [2,7]. Second, with the Web 2.0 any person could write links in the Web, decreasing the quality of them and posing new challenges. Third, time became more important, as new and old links are different [1]. More over, the evolution of the Web structure and content was impacted by Web ranking [3,5].
Finally, we try to map all the concepts of link analysis on the field of Scientometrics. Advantages are that the temporal issues are clearer as scientific publications have a fixed and unique date, and the concept of spam does not exist.

References

R. Baeza-Yates, C. Castillo, and F. Saint Jean. Web dynamics, structure and page quality. In M. Levene and A. Poulovassilis, editors, Web Dynamics, pages 93-109. Springer, 2004.

Ricardo Baeza-Yates, Paolo Boldi and Carlos Castillo. Generic Damping Functions for Propagating Importance in Link-Based Ranking, Journal of Internet Mathematics 3(4), 445-478, 2006.

Ricardo Baeza-Yates, Barbara Poblete. Dynamics of the Chilean Web structure. Computer Networks 50:1464-1473, 2006.

Ricardo Baeza-Yates, Carlos Castillo and Efthimis N. Efthimiadis. Characterization of National Web Domains. ACM Transactions on Internet Technology} 7, Issue 2, 2007.

R. Baeza-Yates, A. Pereira, and N. Ziviani. Genealogical trees on the Web: A search engine user perspective. WWW'08: Proceedings of the 17th international conference on the World Wide Web, pages 367--376, Beijing, China, 2008.

R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval, second edition, Addison-Wesley, England, 913 pages, 2010.

Luca Becchetti, Carlos Castillo, Debora Donato, Ricardo Baeza-Yates, and Stefano Leonardi. Link Analysis for Web Spam Detection, ACM Transactions on the Web 2(1), Article 2, 2008.

A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web: Experiments and models. In Proceedings of the Ninth Conference on World Wide Web, pages 309-320, Amsterdam, Netherlands, May 2000. ACM Press.

P.D. Bruza and T.P. Van Der Weide. Stratified hypermedia structures for information disclosure. The Computer Journal, 35(3):208--220, 1992.

W.-K. Joo and S.H. Myaeng. Improving retrieval effectiveness with hyperlink information. Proceedings of International Workshop on Information Retrieval with Asian Languages (IRAL), Singapore, October 1998.

J. Kleinberg. Authoritative sources in a hyperlinked environment. ACM-SIAM Symposium on Discrete Algorithms (SODA), 46(5):604--632, 1998.
Y. Li. Toward a qualitative search engine. IEEE Internet Computing, 2(4):24--29, July 1998.
M. Marchiori. The quest for correct information of the Web: hyper search engines. In Proc. of the sixth international conference on the Web, pages 265--274, Santa Clara, CA, USA, April 1997.
L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: bringing order to the Web. Technical report, Stanford Digital Library Technologies Project, 1998.

Biography:

Ricardo Baeza-Yates is VP of Yahoo! Research for Europe, Middle East and Latin America, leading the labs at Barcelona, Spain and Santiago, Chile, as well as supervising the newer lab in Haifa, Israel. Until 2005 he was the director of the Center for Web Research at the Department of Computer Science of the Engineering School of the University of Chile; and ICREA Professor at the Dept. of Technology of the Univ. Pompeu Fabra in Barcelona, Spain. He is co-author of the best-seller book Modern Information Retrieval, published in 1999 by Addison-Wesley with a second edition in 2010, as well as co-author of the 2nd edition of the Handbook of Algorithms and Data Structures, Addison-Wesley, 1991; and co-editor of Information Retrieval: Algorithms and Data Structures, Prentice-Hall, 1992, among more than 200 other publications. He has received the Organization of American States award for young researchers in exact sciences (1993) and with two Brazilian colleagues obtained the COMPAQ prize for the best CS Brazilian research article (1997). In 2003 he was the first computer scientist to be elected to the Chilean Academy of Sciences. During 2007 he was awarded the Graham Medal for innovation in computing, given by the University of Waterloo to distinguished ex-alumni. In 2009 he was awarded the Latin American distinction for contributions to CS in the region and became an ACM Fellow. In 2011 he became IEEE Fellow.