

## ABSTRACT

This book offers a comprehensive overview of the various concepts and research issues about blogs or weblogs. It introduces techniques and approaches, tools and applications, and evaluation methodologies with examples and case studies. Blogs allow people to express their thoughts, voice their opinions, and share their experiences and ideas. Blogs also facilitate interactions among individuals creating a network with unique characteristics. Through the interactions individuals experience a sense of community. We elaborate on approaches that extract communities and cluster blogs based on information of the bloggers. Open standards and low barrier to publication in Blogosphere have transformed information consumers to producers, generating an overwhelming amount of ever-increasing knowledge about the members, their environment and symbiosis. We elaborate on approaches that sift through humongous blog data sources to identify influential and trustworthy bloggers leveraging content and network information. Spam blogs or *splogs* is an increasing concern in Blogosphere, which is discussed in detail with the approaches leveraging supervised machine learning algorithms and interaction patterns. We elaborate on data collection procedures, provide resources for blog data repositories, mention various visualization and analysis tools in Blogosphere, and explain conventional and novel evaluation methodologies, to help perform research in the Blogosphere.

The book is supported by additional material, including lecture slides as well as the complete set of figures used in the book, and the reader is encouraged to visit the book website for the latest information:

<http://tinyurl.com/mcp-agarwal>

## KEYWORDS

blogosphere, weblogs, blogs, blog model, power law distribution, scale free networks, degree distribution, clustering coefficient, centrality measures, clustering, community discovery, influence, diffusion, trust, propagation, spam blogs, splogs, data collection, blog crawling, performance evaluation

*To my parents, Sushma and Umesh Chand Agarwal...-NA*

*To my parents, wife, and sons...-HL*

*...with much love and gratitude for everything.*



# Contents

	Acknowledgments .....	xi
<b>1</b>	<b>Modeling Blogosphere .....</b>	<b>1</b>
	1.1 Modeling Essentials .....	2
	1.2 Preferential Attachment Blog Models .....	8
	1.2.1 Log-normal Distribution Models	12
<b>2</b>	<b>Blog Clustering and Community Discovery .....</b>	<b>15</b>
	2.1 Graph Based Approach .....	17
	2.2 Content Based Approach .....	21
	2.3 Hybrid Approach .....	24
<b>3</b>	<b>Influence and Trust .....</b>	<b>27</b>
	3.1 Influence .....	27
	3.1.1 Graph Based Approach	30
	3.1.2 Content Based Approach	33
	3.1.3 Hybrid Approach	34
	3.1.4 Blog Leaders	40
	3.2 Trust .....	40
	3.2.1 Trust Computation	41
	3.2.2 Trust Propagation	43
<b>4</b>	<b>Spam Filtering in Blogosphere .....</b>	<b>45</b>
	4.1 Graph Based Approach .....	47
	4.2 Content Based Approach .....	49
	4.3 Hybrid Approach .....	51
<b>5</b>	<b>Data Collection and Evaluation .....</b>	<b>53</b>