Multi-Relational models of the Blogosphere:
Empirical Characterization and Spam Protection

Pranam Kolari

April 11, 2006
Abstract

Weblogs, or blogs, have become an important new way to publish information, engage in discussions and form communities. Blogs collectively constitute the blogosphere, a highly influential and dynamic subset on the Web. The nature of their content and publishing infrastructure requires that they be modeled, harvested and analyzed differently from the rest of the web.

We first propose a model for the blog graph that extends the more general web graph. The web is viewed as a graph \(G(V, E)\) where \(V\) is the set of pages and \(E\) represents hyperlinks between them. We view the blogosphere at a much lower granularity. In, and around the blogosphere, entities of the set \(V\) can be one of blog post \(P\), blog \(B\), MSM page \(N\), or their complement \((P \cup B \cup N)^c\) that stands for the rest of the Web. The post \(P\) is a tuple consisting of \(Title, Content, Time, Tag, Author\) and \(Comment\). An edge \(E\) represents a link between any of these entities and represents the different kind of relationships that associate them. For instance, such relationships could represent citations between two posts, ownership between a blog and a post, or blogrolls between two blogs. This multi-relational conceptualization, followed by its instantiation is made feasible through the metadata rich publishing on the blogosphere, driven by RSS, OPML (Outline Processor Markup Language), DC (Dublin Core) and FOAF (Friend of a Friend), all of which constitute popular approaches to structuring web content.

To empirically characterize this multi-relational model, we develop a system that discovers and harvests blogs. We focus our characterization towards local content and metadata on blogs, and the link structure involving various entities in our proposed model. Overall, we study the community oriented nature of blogs, identify the boundaries of the blogosphere, and clarify what features renders it different from the rest of Web. We also share the approaches we employed and their effectiveness in harvesting blogs from ping servers, blog/web search engines and the rest of the Web. We recommend and provide new mechanisms that could be useful for timely content harvesting on the blogosphere.

We finally study the nature of spam afflicting the blogosphere by studying which entities in the multi-relational model are prone to spam, and at what infrastructure point. We formally make a distinction between spam blogs with e-mail spam and web spam in general. We then propose algorithms and techniques that employ both local and relational models to detect and eliminate spam posts and the blogs hosting them. We explore how such algorithms can be made to adapt and learn in an adversarial classification setting i.e. as techniques employed by spammers evolve. Based on our own analysis of the algorithms and their cost, we finally recommend their use at various infrastructure points in the blogosphere.