Memeta: A Framework for Multi-Relational Analytics on the Blogosphere

Pranam Kolari and Tim Finin
University of Maryland, Baltimore County
Baltimore, MD 21250
{kolari1, finin}@cs.umbc.edu
http://ebiquity.umbc.edu/project/memeta

Introduction

The "**memeta**" project is developing a framework for studying the structure and content of the blogosphere. We are particularly interested in how metadata about blogs can be discovered, extracted and computed, and how this metadata can be modeled, represented and analyzed to provide new blog related services.

Weblogs, or blogs, are web sites consisting of dated entries (posts) typically organized in reverse chronological order on a single web page. They have become an important new way to publish information, engage in discussions and form web communities. Due to their high influence and specialized publishing infrastructure the subset they constitute is popularly known as the *blogosphere*.

The dialogue oriented nature of the blogosphere has manifested in the use of *Comments* and *Trackbacks* as mechanisms for managing conversations. Blogs are also highly dynamic and embedded in a community of both casual and regular readers, factors that have led to the adoption of new content representation models supported by metadata in several formats. Blogs are considered as "killer applications" for the RSS (RDF Site Summary) and Atom syndication formats which recommend an XML-based representation for machine consumption, representing information like title, date, summary, etc. for content (e.g., a blog post). This metadata is used by aggregators and search engines to provide new blog-related services.

The popularity of blog aggregation has in turn encouraged the publishing of explicit author name and sometimes complete profile information using the Dublin Core and Friend-Of-A-Friend metadata vocabularies. In addition, blogs truly exploit the use of tags and the *folksonomies* they create. Tags are explicit metadata labels associated with a post to provide a highly subjective topic description, and has been found useful for search, browsing and "buzz" analysis. Blogs have also popularized the concept of referrals. Bloggers often publish frequently visited and followed blogs (known as blog subscriptions or blogrolls) in well structured formats like OPML (Outline Processor Markup Language), creating a referral chain that enables interesting topical and community oriented web navigation capabilities.

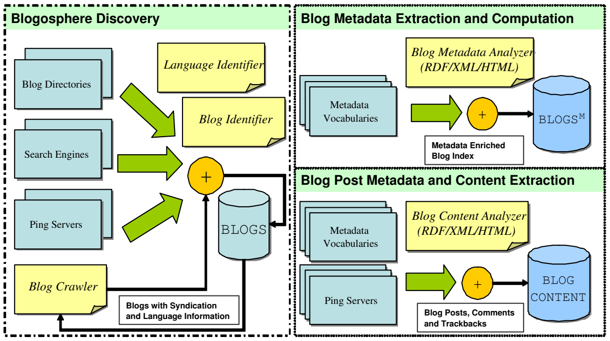
Blogosphere Discovery
Blog Directories
Search Engines
Ping Servers
Language Identifier
Blog Identifier
+
BLOGS
Blog Crawler
Blogs with Syndication and Language Information
Blog Metadata Extraction and Computation
Metadata Vocabularies
Blog Metadata Analyzer (RDF/XML/HTML)
+
BLOGS^M
Metadata Enriched Blog Index
Blog Post Metadata and Content Extraction
Metadata Vocabularies
Blog Content Analyzer (RDF/XML/HTML)
Ping Servers
+
BLOG CONTENT
Blog Posts, Comments and Trackbacks

Figure 1: *Memeta discovers blogs, analyzes them and their posts, extracts and computes metadata, and adds information to its database.*

Motivated by the requirements from the blogosphere, multiple parallel and sometimes conflicting metadata efforts exist and are being used, including those from the *Semantic Web*, *Microformats* and *Structured Blogging* initiatives. Independent of which effort or their combination will finally prevail, the blogosphere has driven, and will continue to drive the publishing of richer metadata on the web. The specific nature of this content demands new techniques for knowledge management and also motivates most of our work.

Discovery Framework

While conventional web search engines use crawlers to discover new content, the blogosphere has evolved a different approach. The community oriented nature of the blogosphere has made the concept of blog directory listings popular, whereas the time sensitive nature of content has created the need for intermediary content brokers known as *Update Ping Servers*. These servers accept explicit pings (i.e., notices of new posts) from newly updated (and created) blogs, and route such notifications to aggregators and blog search engines. We have developed a framework, as shown in Figure 1, that incorporates these additional traits of the blogosphere, combining it with conventional topical crawling mechanisms to create a representative sampling of the blogosphere. The crawlers we employ make use of simple heuristics and Support Vector Machines (Boser, Guyon, & Vapnik 1992) based models for blog identification (Kolari, Finin, & Joshi 2006). Our language detection module employs a cosine similarity based measure to compare content against known language models. Given the computational resource requirements of analyzing the blogosphere as a whole; we fetch metadata and content from blogs on a need-to-use basis. Our system has so far collected and analyzed over ten million blogs.

Modeling and Analysis Framework

Figure 2 shows the multi-relational nature of content extracted through the Memeta framework. This is unlike the conventional web graph model that treats a single entity, a webpage, as a node and models hyperlinks as edges creating a simple relational model. This relational model is then used in many interesting applications, including the now popular social ranking of pages used by search engines. The nature of the blogosphere makes identifying and extracting multiple entities and their inter-relationships possible. This has prompted us to tackle the problem of link-based object classification and link prediction (Getoor & Diehl 2005) in a multi-relational setting. We are developing representation models for the blogosphere to support specialized multi-relational mining algorithms.

Preliminary Work on Splog Detection

Though the blogosphere offers new services providing improved user experience, it is also prone to spam, similar to those on the web in general (Gyöngyi & Garcia-Molina 2005). Referring to the multi-relational model shown in Figure 2, three different entities, namely comments, trackbacks and posts (and the blog as a whole) are usually spammed. We have so far restricted our effort to spam blogs (splogs) which constitute around 20% of the blogosphere according to independent reports and our own analysis. Specific traits of the blogosphere make the splog problem particularly challenging. First, blog search engines do not have knowledge of the entire web graph. Second, faster splog assessment is needed to combat the ease of splog creation in the existing infrastructure. Finally, since blog results are ranked mainly by freshness and relevance, most splogs usually do not have incoming links rendering "link-only" web spam detection ineffective.

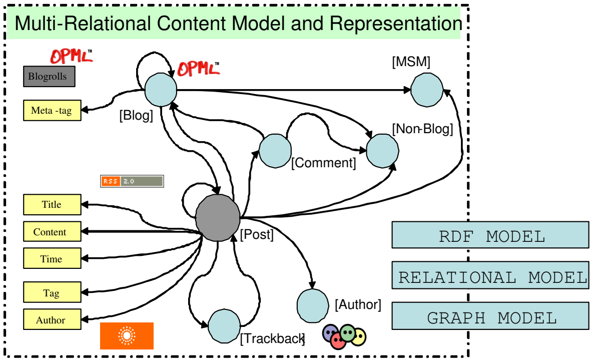

Figure 2: *The blogosphere materializes many relations, making a multi-relational data model appropriate.*

Feature	*Precision*	*Recall*	*F1*
words	.887	.864	.875
urls	.804	.827	.815
anchors	.854	.807	.830

Table 1: *Memeta performs well at splog identification using SVMs with 19000 word features and 10000 each of URL and anchor text features ranked using Mutual Information.*

Results (measured by precision, recall and F1) from our preliminary work on splog detection (Kolari, Finin, & Joshi 2006) using different features are shown in Table 1. The bag-of-words based features slightly outperforms bag-of-outgoingurls (URLs tokenized on '/') and bag-of-outgoinganchors. Demonstrations based on analyzing splogs in ping servers are available online[1].

Continuing Work

The fairly good performance of URL and anchor-based features provide interesting insights into future approaches to the splog detection problem. We are currently experimenting with collective and iterative classification in a multi-relational setting.

Related concrete problems we are addressing include recognizing comment spam and trackbacks; recommending blogs to people; modeling trust relationships in blog communities; and spotting trends. These problems are grounded in the area of content mining and link prediction.

Acknowledgments

We thank Anupam Joshi, Tim Oates and Akshay Java for constructive discussions and James Mayfield for advice and code for language identification.

References

Boser, B. E.; Guyon, I. M.; and Vapnik, V. N. 1992. A training algorithm for optimal margin classifiers. In *COLT '92: Proceedings of the fifth annual workshop on Computational learning theory*, 144–152. New York, NY, USA: ACM Press.

Getoor, L., and Diehl, C. 2005. Link mining: A survey. *SIGKDD Explorations* 7(107).

Gyöngyi, Z., and Garcia-Molina, H. 2005. Web spam taxonomy. In *First International Workshop on Adversarial Information Retrieval on the Web*.

Kolari, P.; Finin, T.; and Joshi, A. 2006. SVMs for the Blogosphere: Blog Identification and Splog Detection. In *AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs, Stanford, March 2006.* To Appear.

[1] See http://ebiquity.umbc.edu/blogger/?p=429.