An Empirical Investigation of Networks in the Blogosphere

Uldis Bojars Andreas Harth Sheila Kinsella Srinivas Raghavendra
National University of Ireland, Galway
University Road
Galway, Ireland
{uldis.bojars,andreas.harth,sheila.kinsella}@deri.org, s.raghav@nuigalway.ie

Abstract

We seek out to investigate the social networks manifested in weblogs. Our dataset, derived from an initial list of URIs, consists of 3.9M files in XML or RDF format, totalling over 400M statements in RDF. We continue by applying well-known network analysis algorithm to parts of the derived network. Our experiments show that links analysis algorithms benefit from input in a common, structured data model.

1. Introduction

In recent times, weblogs have increasingly become important not only in disseminating information, but influencing people in their decision-making process.

The primary objective of the paper is to detect influential entities in the data graph derived from weblogs. There is a growing body of literature on measuring the influence in the Blogosphere by identifying the most popular blogs, e.g. [3]. Rather than looking only at the link structure between weblogs and posts we are interested in understanding the relations among entities in the Blogosphere, such as people, posts, categories, etc.

We show how to construct a network of relations from a large number of sources, represented in a multitude of feed formats. Then we bring this information to a commond data model and calculate metrics derived from this network of relations.

2. Materials and Methods

In this section we describe a common data model and data sources used in our work, as well as conversion and purification performed on these materials. We use a structured data format in order to more easily extract and process data about different types of entities (blogs, posts, people, topics) and relations between them, and to be able to perform network analysis on multiple dimensions.

Regular weblog pages are meant for human users and lack structural information, however, most weblogs also provide web feeds – a machine readable data format for frequently updated content and a structured information source we can leverage. Most of the feed formats, e.g. RSS 2 and Atom, use XML as the underlying data format, which is based on a tree data model. XML is sufficient to express data in these formats, but is limited in usefulness if the data need to be with new properties or relations, or if different kinds of information need to be integrated.

The Resource Description Framework (RDF) is designed to represent information about resources on the Web, based on a graph model rather than a tree data model. It allows to integrate different types of data, and extend schema information with new relations and properties. We convert all data into RDF according to the data model described below.

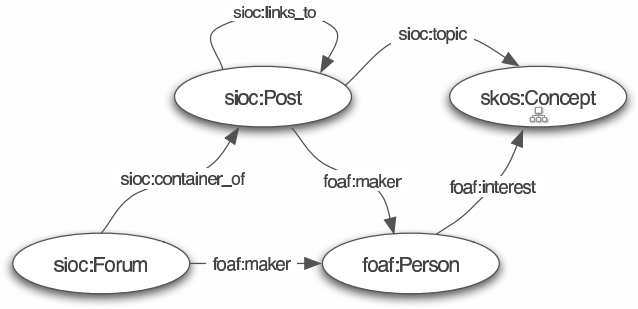

Fig. 1: *Main concepts and connections in the data model.*

Figure 1 shows a high-level view of our data model. Its core is an ontology for Semantically Interlinked Online Communities (SIOC)[1] which allows to describe main concepts of weblogs and other online community sites[1]. SIOC is used in conjunction with Dublin Core (DC) vocabulary for defining additional properties; FOAF vocabulary for describing information about people and their relations; and SKOS[2] to describe categories and tags. A number of sites, such as LiveJournal, use RDF and FOAF to provide information about creators of the content, which can then be integrated with blog content data.

A common problem when working with different feed formats is that some information may be lost when converting data, e.g., from RSS 1.0 to Atom or from Atom to RSS 2.0. The SIOC ontology provides a common framework or data model to which all feed formats can be converted without loosing information.

The initial starting point in our study of the Blogosphere is the ICWSM 2006 conference dataset, consisting of 14M posts from 3M weblogs. The dataset seems to be converted from the HTML and contains little more information beyond the date and time of a post, author name or nickname, title of the post, weblog URL, tags/categories, and outlinks. While exhibiting the desirable property of uniform representation

[1] http://sioc-project.org/
[2] http://www.w3.org/2004/02/skos/

of information across blogging sites and systems, the dataset lacks structural and semantic information that might be of interest in analysis steps. Also, and there are no identifiers for people or categories which could be used to link up these entities across sources. To be able to derive high quality metrics about the Blogosphere, we first require a semantically richer corpus.

We use the ICWSM dataset as a starting reference to extract a large number of URLs to structured content. Even if the raw dataset does not contain links to e.g. RSS or Atom files, we can leverage the fact that most blog sites export structured content from a database and follow a fixed URL scheme to structured content. We manually construct URLs to structured content out of the source URLs from the dataset where possible. Plus, we added URLs to files about information about people (in FOAF) where available.

Table 1 lists the major blogging sites, the data format chosen for conversion, and the number of URLs generated from the initial dataset.

Blogging Site	Data Format	# of URIs
`spaces.live.com`	RSS 2	1,249,104
`livejournal.com`	Atom / FOAF	2 * 808,963
`xanga.com`	RSS 2	363,194
`blogspot.com`	Atom	168,268
`blog.goo.ne.jp`	RSS 1	101,865
`myspace.com`	RSS 2	47,200
`blog.360.yahoo.com`	RSS 2	42,978
`blogfa.com`	RSS 2	40,532
`cocolog-nifty.com`	Atom	31,686
`greatestjournal.com`	Atom / FOAF	2 * 25,656
`wordpress.com`	RSS 2	18,567
`livedoor.jp`	Atom	16,207
`journals.aol.com`	Atom	13,851
`canalblog.com`	RSS 2	13,320
`jugem.jp`	Atom	13,039
Misc	RSS 1/2, Atom	113,497
Total		**3,902,546**

Table 1: *Major blogging sites and associated URIs to structured content pertaining to blogs (Atom, RSS) or to people (FOAF) that can be derived from the ICWSM data set. Individual URIs and smaller small groups are summarised under the Misc category.*

In a next step, we crawled the structured feed files from the respective weblog sites. The crawl resulted in structured content in XML and RDF. During the following purification phase, more low-quality files will be dropped; mainly because XML files were not well-formed or had character encoding issues. In order to use techniques described in this paper the collected data was converted into a unified format (RDF) in SIOC, FOAF and SKOS (for category information) vocabularies.

Transformation of feed syndication formats into RDF is not just syntactic, but also raises the information abstraction level and the data quality. In part it is due to purification that takes place when invalid XML feeds get dropped and data extracted are put into a common model.

We use an XSLT script to convert data from various feed formats into our common data representation. Feed channel and post entry information is converted from RSS and Atom syntax to `sioc:Forum` and `sioc:Post` classes; information about tags and categories is described using `sioc:topic` and `skos:Concept`.

3. Results and Discussion

In the following, we describe initial experiments carried out over a first version of the purified corpus to validate our ideas.

When all the relevant data has been extracted a graph where each instance is modelled as a node, and each link between instances is modelled as a directed edge, is generated and analysed using JUNG[3]. The nodes are labeled with their URI and RDF class, and the edges are labeled with the relation they represent. From this graph, the importance of individual nodes can be calculated using Hubs and Authorities algorithm[2].

For the experiments, we selected a web community according to a topic from the index, namely Web technology. The topic was specified via keyword searches, which resulted in a set of focus nodes. All incoming and outgoing links starting from the focus nodes form the subgraph used in subsequent mining operations. The size of the resulting subgraph is 623,190 nodes and 701,840 edges.

Table 2 lists the ranks calculated for the Web target set.

Rank	Value	URI	Class URI
1	0.999999	sioc:Post	rdfs:Class
2	0.000211	sioc:Forum	rdfs:Class
3	0.000203	blogger:internet	skos:Concept
4	0.000195	blogger:Internet	skos:Concept
5	0.000172	sapart:Web/Tech	skos:Concept
6	0.000168	blogger:web 2.0	skos:Concept
7	0.000152	blogger:web	skos:Concept
8	0.000141	blogger:google	skos:Concept
9	0.000125	blogger:tech	skos:Concept
10	0.000121	blogger:Technology	skos:Concept

Table 2: *Table shows an example of the top 15 nodes (based on authority) in the Web target set.*

4. Conclusion

We have shown how to apply a well-known method for network analysis on information from weblogs weblog, processed using Semantic Web tools and methodologies. By using a common, flexible data model information of different nature can be fused together and analysed on multiple dimensions. This paper describes the first results of such analysis.

We hope that widespread availability of machine-readable social network and online community site data, which can be published directly from databases to e.g. SIOC feeds and FOAF profiles, will foster research on algorithms operating on relations in the Web graph.

References

[1] J. G. Breslin, A. Harth, U. Bojars, and S. Decker. Towards semantically-interlinked online communities. In *2nd European Semantic Web Conference*, pages 500–514, May 2005.

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

[3] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. Structure and evolution of blogspace. *Commun. ACM*, 47(12):35–39, 2004.

[3] `http://jung.sourceforge.net/`