Uses of Ontologies in Open-Source Blog Mining

Brian Ulicny, Chris Matheus, Mitch Kokar, Ken Baclawski
VIStology, Inc.
Framingham, MA, USA
{bulicny,cmatheus,mkokar,kbaclawski}@vistology.com

Abstract

The blogosphere provides a novel window into public opinion, but its dynamic nature makes it an elusive medium to analyze and interpret in the aggregate, where it is most informative. We are developing new technology employing ontologies to solve this problem by fusing the signals of the blogosphere and zeroing in on issues that are most likely to migrate offline, enabling analysts to anticipate the threats or opportunities they represent.

There are nearly 16 million active blogs on the Internet with more launched every day. Although much of what's discussed in the blogosphere is of little consequence, increasingly, blogs are emerging as powerful organizing mechanisms, giving momentum to ideas that shape public opinion and influence behavior. For example, Malaysian bloggers have recently become quite effective in confronting perceived corruption in their national government despite governmental control of the major media [4]. The blogosphere is thus a great bellwether of changing attitudes and new schools of thought, but only if analysts know which issues to pay attention to and how to identify those issues early in their lifecycle.

Even where there is freedom of the press, blogs provide a more complete picture of public opinion. For example, the New York Times reports that it receives about 1000 letters daily, but publishes only about 15 [1]. By contrast, Google's blog search engine reveals that 3000 or so blog posts on average cite the New York Times every day, many not in English.

VIStology's IBlogs Project

VIStology's IBlogs (International Blogs) project is a three-year effort funded by AFOSR's Distributed Intelligence program to develop a platform for automatically monitoring foreign blogs. This technology provides blog analysts a tool for monitoring, evaluating, and anticipating the impact of blogs by clustering posts by news event and ranking their significance by relevance, timeliness, specificity and credibility, as measured by novel metrics.

Current blog search engines allow users to discover trends in the blogosphere only by determining the most popular names or news articles (e.g. Blogpulse.com) or by overall popularity of the blog itself (e.g. Technorati.com). These metrics favor attention-grabbing stories that may not have lasting significance.

The IBlogs search engine, in contrast, ranks blog posts by their relevance to a query, their timeliness, specificity and credibility. Briefly, these are computed as follows (see [8] for details). In particular, because of the exophoric and quotational nature of blogs, it is important to identify links to news articles that posts cite and analyze them. Blog posts are not standalone documents; therefore, information retrieval metrics must take into account the articles they cite as well as the commentary they add.

Relevance: What a blog post is about is determined not only by the text of a post, but also by the text of any news article it references. Terms in news articles and blog posts are not ranked by the familiar tf*idf metric standard in information retrieval in light of the clumpiness of the corpus and journalistic conventions.

Timeliness: The timeliness of a blog post is determined by comparing the timestamp of a blog post with the publication date of a news article that it cites. Timeliness, as distinguished from recency, is about proximity to the relevant event.

Specificity: The number of unique individual entities mentioned in a blog post and any news article it cites determines the specificity of a blog post. This is approximated as the number of unique proper nouns and their variants. Attention is also paid to depth in a domain ontology.

Credibility: The credibility of a blog author's posts is determined by the presence of various credibility-enhancing features that we have validated as informing human credibility judgments [7]. These include blogging under one's real name, linking to reputable news outlets, attracting non-spam comments, and so on. This analysis must be computed for each author, since blogs can have multiple authors. The number of inlinks alone does not determine blog credibility.

Ontologies in IBlogs

IBlogs uses ontologies and ontological relations in three ways. First, IBlogs uses an explicit domain ontology in OWL for query expansion. Second, IBlogs uses an ontology of the blogosphere to represent and normalize blog data. Third, IBlogs outputs data expressing explicit ontological relations.

Architecturally, the IBlogs systems includes a document extraction module, a metrics computing module, an indexer (Lucene), a crawler (Nutch), an ontology reasoner (BaseVISor) and a consistency checker (ConsVISor). See Figure 1.

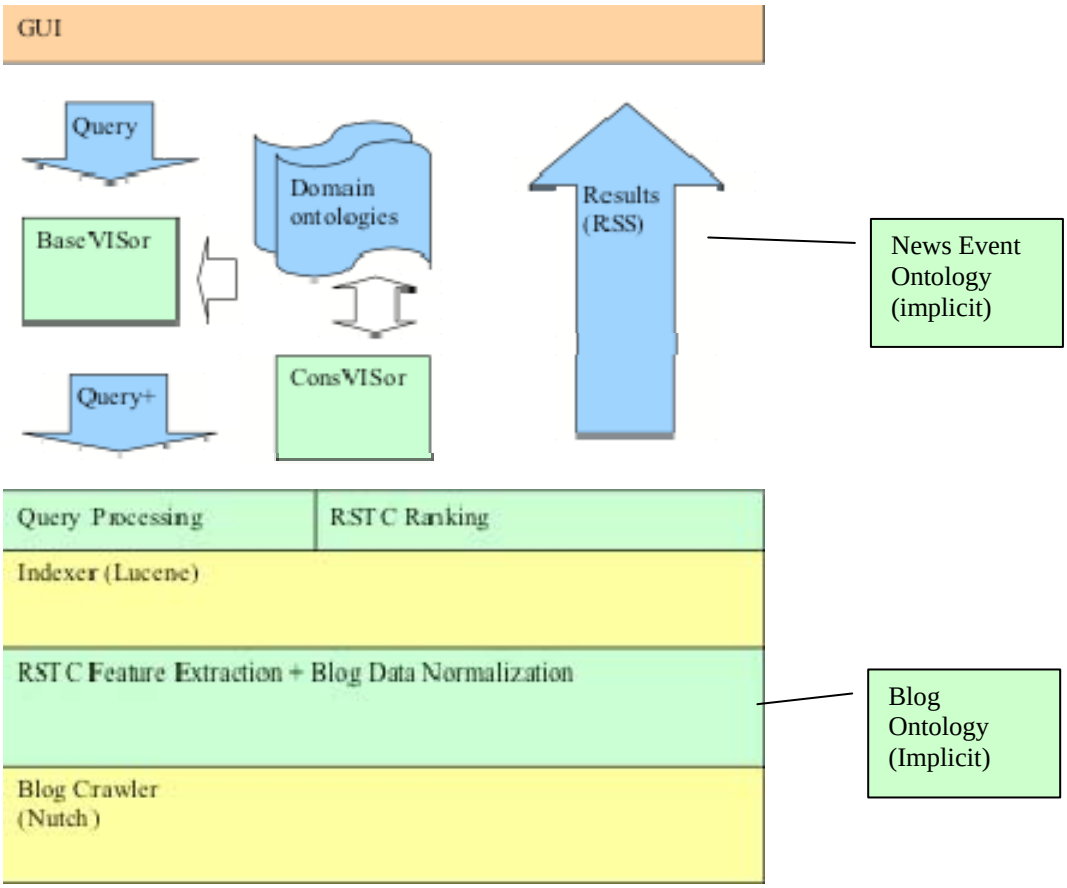

Figure 1: IBlogs Components

VIStology's BaseVISor inference engine is used to query domain ontologies. At present, we are using a terrorism ontology from Teknowledge. BaseVISor [3] is a forward-chaining inference engine that is based on a Rete network optimized for processing triples. It is able to process RuleML rules containing n-ary predicates, and incorporates the axioms and consistency rules for R-Entailment [6]. BaseVISor allows the system to expand queries based on the domain ontology. For example, the query

[class:TerroristFinancier Damascus]

would be expanded to a query that would return blog posts containing any string that has been included in the domain ontology as a member of the class "TerroristFinancier" and the term "Damascus".

ConsVISor is a rule-based system for checking consistency of ontologies represented in RDF, OWL, or DAML. We use ConsVISor to help us mediate conflicts and inconsistencies between multiple domain ontologies. ConsVISor can be used to determine whether two entities (with or without the same name) are coreferential [2].

An ontology of the blogosphere is implicit in the system. The Semantically-Interlinked Online Commnities (SIOC) ontology [5] provided a useful starting place, but we found it necessary to extend it. While the idea of blogging involves certain essential features, platforms for blogging are not standardized. That is, blogs do not specify what links constitute their 'blog roll', or which links are 'trackbacks' to other blogs, and so on. While feeds for blogs may be specified in several syndication standards (RSS 1.0, RSS 2.0, Atom), these feeds require further analysis because the feed itself is not guaranteed to contain the entire blog post, blog comments, images or profile information relevant to determining blog credibility. All this requires parsing and analyzing HTML blog pages that are designed for human consumption.

Finally, IBlogs outputs information annotated according to an ontology of news events and participants. Our goal is to cluster blog posts by the news events that they are about, where any given news event may have more than one news story that reports it, and each of those stories may be published at one or more URLs. A news event is thus typically two levels removed from a blog post that references it. Our system outputs results in the OpenSearch 1.1 RSS standard (opensearch.org), which we have extended with concepts from the Dublin Core metadata standard (dublincore.org) and with our own namespace elements for news event representations.

NewsML (newsml.org), and the associated EventML standard, represent news industry-originated attempts to standardize representations of news articles and the events they report. These standards can be readily converted to OWL ontologies. We will adapt these emerging standards, currently used by Reuters and Agence France Press (AFP) among others, to standardize the representation of news articles and news events in hope that we will be able to directly use output in these formats produced by news providers in the future.

The IBlogs project demonstrates that ontologies are useful for fusing blog information concerning the elements of the blogosphere, topical subject matter and semantic relations between posts.

Acknowledgement

This material is based upon work supported by the United States Air Force under Contract No. FA9550-06-C-0023. The views and findings expressed here do not necessarily reflect the views of the United States Air Force.

References

1. Feyer, T., *Editors' Note; The Letters Editor and the Reader: Our Compact, Updated.* New York Times, May 23, 2004

2. Kokar, K, C. Matheus, J. Letkowsk, K. Baclawski and Paul Kogut, *Association in Level 2 Fusion*. In Proc of SPIE Conference on Multisensor, Multisource Information Fusion, Orlando, FL., April 2004 (vistology.com/consvisor)

3. Matheus, C., K. Baclawski, M. Kokar. *BaseVISor: A Triples-Based Inference Engine Outfitted to Process RuleML and R-Entailment Rules*. In Proceedings of the 2nd International Conference on Rules and Rule Languages for the Semantic Web, Athens, GA, Nov. 2006. BaseVISor is

freely available (vistology.com/basevisor).

4. Open Source Center, *Analysis: Tension Between Malaysian Bloggers, Authorities Appears To Intensify*, September 13, 2007

5. *SIOC Core Ontology Specification* http://rdfs.org/sioc/spec/

6. ter Horst., H, *Combining RDF and Part of OWL with Rules: Semantics, Decidability, Complexity*. In Proc.of the Fourth International Semantic Web Conference. Y. Gil et al. (Eds.): ISWC 2005, LNCS 3729.

7. Ulicny, B., and K. Baclawski, *New Metrics for Newsblog Credibility*, Proceedings of 1st International Conference on Weblogs and Social Media (ICWSM'07). Boulder, CO.

8. Ulicny, B., K. Baclawski, A. Magnus, *New Metrics for Blog Mining*, Proceedings of SPIE -- Volume 6570 Data Mining, Intrusion Detection, Information Assurance, and Data Networks Security 2007, Belur V. Dasarathy, Editor, 65700I (Apr. 9, 2007), Orlando, FL.