AUTOMATING SEMANTIC BLOGGING

Praphul Chandra
HP Labs,
Bangalore, India

Saravana Mariappan
HP
Bangalore, India

Geetha Manjunath
HP Labs,
Bangalore, India

ABSTRACT
Blogging is an easy and widely usable tool for data exchange and knowledge management. But the data in blogs are limited to the 'followers' (human readers) and not easily available for machine processing to enable information aggregation or further analysis. Thus, there is a need to enrich blog content with additional semantics using metadata. This paper proposes an approach for automating semantic blogging i.e. automatically creating semantic blog entries for certain user web actions. As an example, we show how a review given at a site (e.g. Amazon) can automatically create a semantic blog entry for the user. Our approach is unique in that it focuses on automating the end-to-end process of creating a semantic blog entry. Furthemore, it does not require any change to existing websites The basic approach to automation of semantic blogging is to observe user created data in web actions (e.g. review contents), transform this into XML, parse in to objects, enriched with metadata like microformats (hReview, hCalendar, hProduct and hCard) and publish as blog entries through our automated web browsing technology.

KEYWORDS

Automation, Blogs, Semantic Web, XML, Microformats,

1. INTRODUCTION

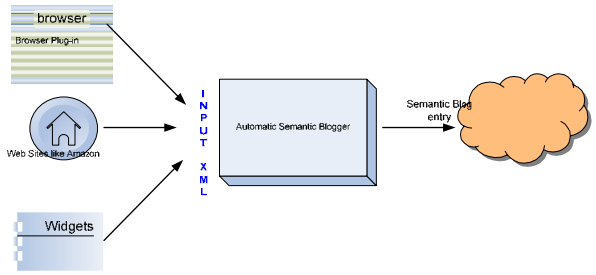
The information available in the web continues to grow at a rapid pace and the web is now the largest human-created information store. This growth in content has gotten a further impetus with Web 2.0 technologies enabling 'user created content' e.g. reviews, tags, favourites, blogs, videos. However, one limitation today is that the user has no ownership or control over the content she creates on the web. One way user can control the content is to publish the content in his blog – however, if the content is published only on a blog, it suffers from limited visibility. Therefore, a reasonable approach is to publish the content both at the destination site (e.g. book review at Amazon) and on the blog. This approach has two limitations - (i) it needs manual effort from user to copy and create a blog entry with the content every time she creates content anywhere on the web and (ii) machine readability of this content poses a big challenge for effectively accessing it and making it available for aggregation and computing.

In current state-of-the-art, OntoBlog [Shakya, 07] attempts linking blogs to existing ontology maintained using available ontology management environment. OntoBlog is a prototype semantic blogging system which employs semi-automatic semantic annotation of blog entries using ontology instances. It provides Semantic navigation to navigate through each blog entry to semantically related blog entries. Semantic aggregation to collect blog entries relevant to the topic of interest is also supported. SemiBlog [Knud, 06 / Knud,Stefan 05] proposes an instantiation of semantic blogging which exhibits tight integration with various desktop applications. SemiBlog extracts data from existing structured desktop data (electronic address books, calendar entries etc) and generates semantic data which is used for semantic blogging. RDF [www –RDF, 09] is used as the representation format. All the above efforts are towards adding more semantics to already blogged entries. They do not automate the end-user creation of semantic blog entries – an important aspect of our work.

We propose an approach where content created by a user (e.g. a review at Amazon) is automatically manifested as a semantic blog entry in the user's blog. This approach enables three functionalities – (a) it allows users to maintain a link (ownership) to the content they create [Chandra, 2009] and (b) it allows computational extraction and aggregation of user created content and (c) it automates posting of entries into a blog and review sites through automated browser functionality. Specifically for the 'review' case, we can implement computing logic to perform queries like *'how many friends of mine have rated Restaurant 'X' more than 4', 'What does 'P' think of 'Y', 'What is the average of ratings given for Movie 'M' by my friends'*

2. AUTOMATING SEMANTIC BLOGGING

Figure 1. Automatic Semantic Blogger

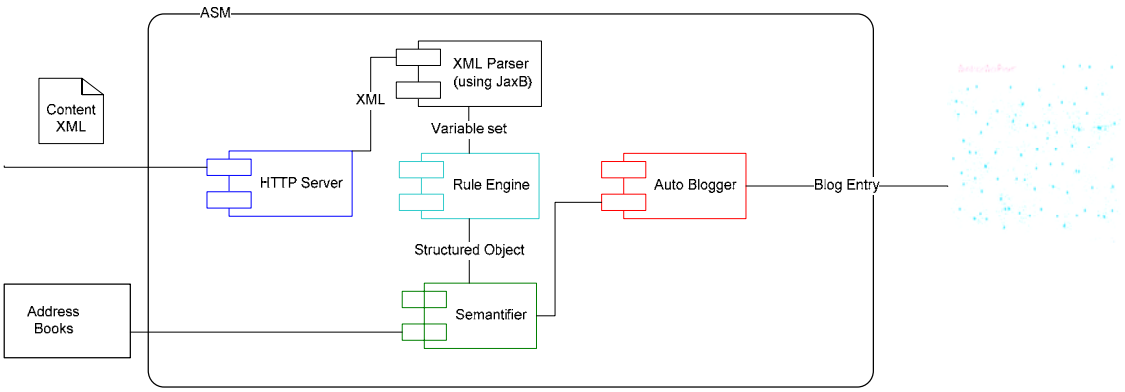


The basic approach to automating semantic blogging is to capture the content of web actions (e.g. the contents of the review that one writes on the web), enrich this content with metadata (Microformats), formulate it to semantic HTML and publish it at a blog-site. This approach is shown in figure 1 where the automatic-semantic-blogger block accepts inputs (as an xml file) from various web-interactions and creates semantic blog entries.

2.1 Capturing content from web actions

To capture the content from a web action, we need a component in the interaction path of the user as shown in Figure 1. One approach is to have a browser plug-in which tracks locally (no privacy concerns) user's browsing and captures user created content whenever pre-specified (and user-configurable) conditions are met (e.g. 'review creation at Amazon'). This browser plug-in then converts the captured content to a prescribed XML format (Listing 1) which is generic enough to capture data that one wants to capture / extract as part of a web action (e.g. a review). This XML is posted to an http port where the automatic semantic blogger is listening. A second approach is to expose an API for the destination web site to invoke this functionality which may support such content capture and transformation to prescribed XML – with automatic blogging as an incentive for users to create content & continue to have control on it. A third approach, and the one that we have adopted is a custom created widget (say, a widget used to write reviews on websites) which captures the content every time it is used, transforms to XML and post that to a configured port. Our widget exploits our previous work which uses a hidden browser technology based on web macro recording, enabling use of existing blog web sites as an interface for pushing the user comments either on their blogs or review site [Manjunath 2009].

Figure 2. ASM - Components



Listing 1. Sample XML for Book Review

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<VariableList>
        <taskletname>BookReview</taskletname>
  <Variable refname="TLREF1">
    <name> ProductType</name>
```

```
    <contentvalue>Product</contentvalue>
    <type>Extract</type>
  </Variable>
  <Variable refname="TLREF6">
    <name>text</name>
    <contentvalue>Very well written about ….. A good book to read if you want to know …</contentvalue>
    <type>Free</type>
  </Variable>
</VariableList>
```

2.2 Adding Semantics

An HTTP Server (A simple Java based program implemented with sockets) listens at the configured port and receives the XML. Then the contents are marshaled to collection of java objects (Objects with Name, Content Value fields) by the XML Parser based on Java architecture for XML Binding. This collection is subjected to a Rule engine which is responsible for imposing pre-described structure on the content (user-created / site-extracted) content where possible.

Rule Engine converts the object collection in to a structured object (Say ReviewVO) based on the 'Rules' created by the user. The 'Rules' are textual mapping between 'Name of the variable' and 'attribute of one of the Micorformats supported by our tool. Microformats [web-microformat, 09] are a set of simple, open data formats built upon existing and widely adopted standards. User can create such rules for each of the Microformat attribute that can be mapped to the extract/input value of user created content (e.g. a review). For each of the object, the value in the name field is checked against the 'Rules' by 'Rules Engine' and based on it, the value in the 'Content value' is assigned to a field of ReviewVO e.g., if the content contains 'ProductType' (e.g. extracted from the website where the user content was posted) and there is a rule that maps 'ProductType' to 'hReview.Category', then the value is correspondingly assigned'. Finally, the ReviewVO is sent to 'Semanticer.

Semanticer forms the XHTML by using Microformats to add annotations/ metadata that describe or define the data content of the review. Semanticer receives the Review Object and using 'Rules' that map fields to the attributes of Microformats, it forms the XHTML by including the right Microformat class tags along with HTML tags. Currently hReview, hProduct, hCard, hGeo and hEvent Microformats are used to enrich review, reviewer, reviewed item and context of the review respectively. This XHTML is posted by the 'AutoBlogger' component using 'iMacro' [web-imacros, 09]. Alternately one can use the API exposed by the respective blog sites.

Listing 2. Snippet of XHTML generated for Sample XML listing in Listing 1, with Microformat Tags(in **Bold**)

```
<div class="hreview">
<span class="type">Product</span><tr><td><p class="summary" align="center">
<p rel="tag">Product</p>:Discovery of India</font></b></u></p>Review (<a
href="http://microformats.org/wiki/hreview"> hReview v<span class="version">0.3</span></a>) by
<span class="reviewer vcard"><span class="fn">Saravana Krishnan</span>
<span class="email">xxxx,</span> at yyyy <abbr class="dtreviewed" title="2009-07-
21T00:00:00">2009-07-21 11:12:88</abbr>)</font></u></p><div class="geo">GEO:<span
class="latitude">123</span><span class="longitude">89</span></div>
<p align="left">Ratings</p></td><td><p align="left" class="rating">4</p></td></tr>
<p align="center" class="description">Very well written ……</p></td></tr></table>…..
```

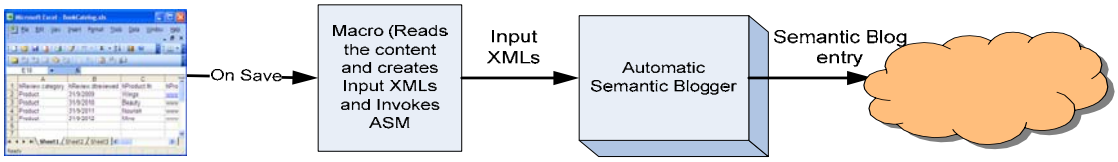
2.3 Publishing on the Blog site using a widget:

In order to publish the semantic XHTML as a blog entry, our automatic blogger employs a hidden browser technology that works on existing websites. As a one-time configuration action, the user uses our browser macro recorder to 'demonstrate' to our tool how and where the blog entries need to made. Our automated widget creation technology [Manjunath, 2009] is then used to create a widget which automates the publishing of the semantic XHTML to the blog site. This enables end-to-end automation to work on existing websites without assuming change in the blog server or review sites. We build on the web recorder from iOpus to enable the above widget.

2.4 Semantic catalogs in Blogs

Semantic blog entries can also be used as online catalog. Such an approach can be useful for e.g. for small micro / casual entrepreneurs who wish to offer an online shopping experience to their customers without incurring the cost of an IT department / vendor for creating and maintaining an updated catalogue of their inventory. We consider a case where a vendor wants to maintain an online catalog for his inventory. In this use case, the vendor would prefer to upload his inventory content as a catalogue in a batch mode instead of invoking a widget for each of the books in his catalog. We have prototyped a simple implementation which uses commonly used spread sheets, macros and some components of our solution.

Figure 5. Online catalog from spreadsheet



A Spreadsheet template with column names mapped to attributes of microformats like hReview, hProduct, hCard is created. A macro is written to read the contents of the spreadsheet, create input XMLs with name-value pairs as listed in Listing 1. The content of the column header which is an attribute of a microformat supported by ASM becomes the name and the corresponding column value becomes the value in the XML.

Similarly for each of the column a name-value pair is formed and thus an XML per row in the spread sheet is generated and moved to a pre-designated folder in the file system. Then a batch program is triggered to initiate the Semantic Blogger. The Blogger component consumes all XMLs in the folder and creates semantic blog entries (one for each XML) as explained in Section 2.2.

2.5 Exploiting the semantics in Blogs

The data available in the blog can be accessed semantically by various techniques. The core intelligence to aggregate / analyze intelligently the microformats can be included inside custom written apps (widgets), RSS aggregators or web crawlers. Many search engines have also started indexing the content in the wild, on Microformats e.g. Google (rel-nofollow) :Google Advanced Search - Usage rights (rel-license),Technorati (rel-nofollow),Technorati ,Tag Search (rel-tag) ,Yahoo (rel-nofollow) ,Yahoo Creative Commons search ((rel-license). ,Yahoo Search Monkey (hCard, hCalendar, hReview, hAtom, others) .

3. CONCLUSION

We believe Automatic semantic blogging is a small step taken in the right direction for achieving the semantic web vision. We have proposed and implemented architecture for automatic capturing of content during user's actions, adding rich metadata to the contents and publishing the semantic blog entry. Since we use the macro-recording technology for automation, our approach does not assume any change on the blogging server or review site.

One sample implementation is completed for writing reviews. This functionality can be easily extended to create other blog entries like articles, stories, recommendations, alerts, transactions etc. Another prototype is completed for creating online product catalogues.

For exploiting the semantic blog entry, web crawlers can be written in the blog to display intelligent aggregation e.g. a list of top rated reviewed items. We can also provide fine resolution filtering of RSS feed. The RSS subscribers can be given contextual Alerts on recommendations for e.g. based on their location and time zone

REFERENCES

Rohit Khare, 2006. Microformats: The Next Small Thing On The Semantic Web. In IEEE Internet Computing, 2006 Vol.10, No.1

Aman Shakya, Vilas Wuwongse, Hideaki Takeda , Ikki Ohmukai, 2007. OntoBlog: Linking Ontology and Blogs.

Steve Cayzer, 2006. What Next For Semantic Blogging. Tech Report

Knud M¨oller and Stefan Decker, 2005. Harvesting Desktop Data For Semantic Blogging.

Knud M¨oller,2006. Using Semantics to enhance blogging experience, Semantic web conference, June, 2006.

Web-SW,2009. Semantic Web, http://semanticweb.org/wiki/Main_Page

Web-MF, 2009. Microformats, http://microformats.org/

Chandra, Gupta , 2009, Retaining personal expression for social search, Internationl World Wide Web Conference, Madrid, Spain, April 2009

Manjunath et al,. 2009. Creating mobile widges without programming, Internationl World Wide Web Conference, Madrid, Spain, April 2009