Using Semantic Web, Ontologies and Blogs for Knowledge Identification, Organisation and Reuse

Azleena Mohd Kassim[1*], Yu-N Cheah[1]

[1] *School of Computer Sciences, Universiti Sains Malaysia, 11800 USM Penang, Malaysia*

In this paper, we propose an approach to make knowledge-based blogging more interesting, reliable and sophisticated by using semantic web with ontology in the blogging environment. The aim of our research is to realise the knowledge management processes of identification, organisation and reuse (adaptation and application) via the blogging environment for them to be available to users in a seamless application. To achieve knowledge identification, we employ natural language processing techniques coupled with ontologies to identify possible categories of new blog posts. For knowledge organisation, we finalise the categorisation of blog posts and store them into a repository. Finally, in knowledge reuse, we create semantic links between blog posts and other related documents. These processes are integrated in a blogging framework which we call SEMblog.

1. Introduction

Knowledge management has always been a key issue in many organisations. As more and more organisations rely on human and intellectual capital to outperform their competitors, the challenge to manage knowledge effectively becomes greater, particularly in the efforts to identify useful knowledge, organise them, and to reuse relevant knowledge.

Current knowledge management methods do not specify how the knowledge is to be handled after it is stored in the database or repository. Knowledge is often dumped in a knowledge base or a repository, and later retrieved upon request. We do not have a set of procedures on how the knowledge is handled within the repository. There are also issues on how to represent knowledge in a machine-understandable form so that it will be easier to be reused over the web or within a centralised system. Semantic web may be an answer to this issue but most semantic web documents that are available are not readable to human users as they are very rigorously formatted, usually in XML. Another problem that arises is that typical knowledge search does not necessarily link to other pages that are relevant.

To address some of these issues, in this paper, we present a blogging framework called SEMblog to identify, organise and reuse knowledge based using semantic web and ontologies. The SEMblog framework brings together technologies such as natural language processing (NLP) and the internet to produce a more intuitive way to manage knowledge, especially in the areas of knowledge identification, organisation and reuse.

2. Related Works

Warren (1) worked on issues pertaining to the use of the semantic web for knowledge management and also the relevance of using ontologies in the semantic web. Warren discussed the building of a knowledge management environment, which starts from a scenario, moving towards utilising technologies for the purpose. Semantic web is seen as a knowledge management scenario that works well as a search mechanism, provided suitable technology is applied to it. He also stressed on the importance of making the semantic web explicitly available and has a business process value.

Stojanovic and Handschuh (2) built a framework using the existing tools in semantic web, which is expected to bring semantic web to a level that can be realised by real-life applications. Both these works (1)(2) have given us motivation on finding a more practical semantic web application whilst still prioritising aspects of knowledge management. We note the suitability and potential of the blogging environment for this purpose.

Avesani et al. (3) worked on a topic-centric view of blogs with the Tagsoratic project, which makes use of the tags and category labels in the project to enable users to find posts that are semantically connected, referring to the same topic being searched. We observed that the use of tags can help to achieve connectivity through all the contents on the semantic web-enabled structure.

Karger and Quan (4) built a tool called Haystack that is reported to be successful and they also created an RDF-encoded blog to check on how RDF representation benefits blogging activities. Haystack provides a robust semantic blogging environment and manages to inter-link blogs. From our perspective, we aim to go beyond creating a semantic blogging environment (where blogs that are relevant to each other are connected), but we are also going into knowledge sharing via semantic blogs.

Moller et al. (5) presented a prototype called semiBlog that focuses on making blog authoring easier to computer users. They used the data that are available on desktops (electronic address books, mp3 metadata, etc.) as semantic metadata that would be imported into the blog. In our SEMblog framwork, we aim to work with simple XML or RSS, but we find that we could add in more meaning to the annotation by having a tag for categories in the RSS module.

3. Our SEMblog Framework

The development of our SEMblog framework consist of three main steps (see Fig. 1): 1. knowledge identification, 2. knowledge organisation, and 3. knowledge reuse.

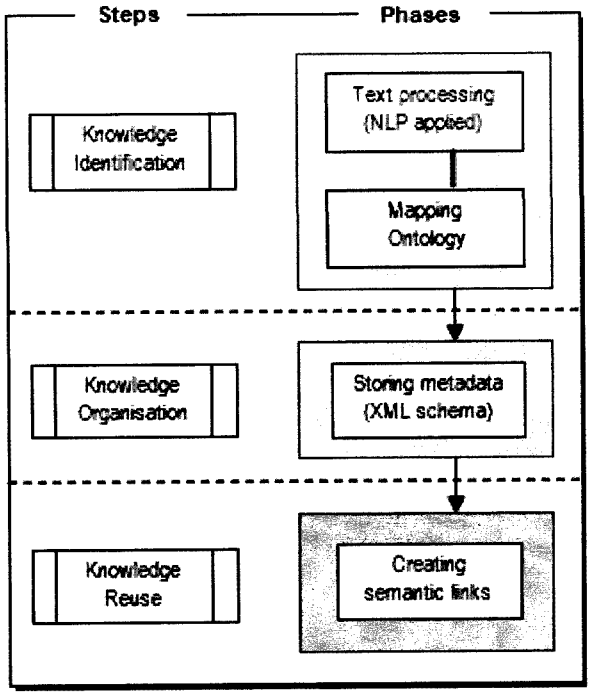

Fig. 1. Our SEMblog framework

3.1 Step 1: Knowledge Identification

In this step, knowledge is acquired from blog posts. These may be in the form of discussions, narrations, thoughts and ideas. The acquired knowledge is then analysed using NLP and intelligent agent techniques, assisted by a domain ontology. Through this analysis, candidate categories for the acquired knowledge are identified (see Fig. 2).

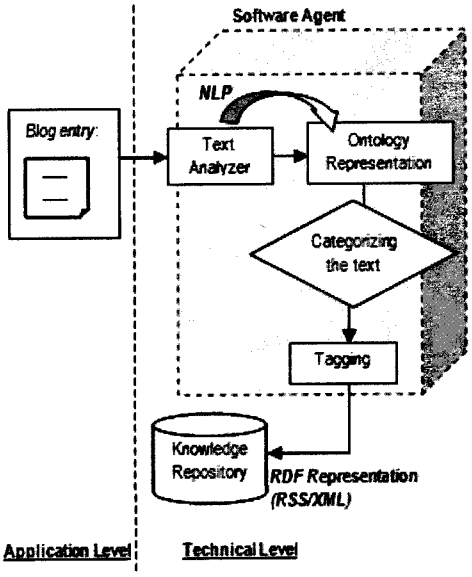

Fig. 2. Knowledge Identification and Organisation

This step is realised in two distinct phases: 1. text processing, and 2. ontology mapping.

3.1.1 Text processing

Text processing is carried out after the SEMblog interface receives a (new) blog post from a user. For this purpose, a software agent is created. The objective of the text processing agent is two-fold.

Firstly, it is to obtain a list of weights (based on word statistical distribution) for nouns and verbs in the blog post. This is done after removing all stop words. The weighted scores are basically the percentage of the word's frequency over the total number of words in the blog post.

Secondly, it processes an OWL ontology (we have adopted the OntoSem (6) ontology) to obtain only the classes and sub-classes. This in effect strips away all RDF tags and leaving behind a list of only the concepts and related concepts.

3.1.2 Ontology mapping

By mapping the weighted words to a particular ontology concept, we hope to identify a knowledge category for the blog post. This in turn achieves a form of knowledge identification.

Ontology mapping is carried out by matching the list of words with their weights with the list of ontology concepts (hierarchy of classes). When a match is found between a word and an ontology class, we calculate the probability (Pw) of that class being a category (see Eq. 1).

$$Pw_C = w \, x \, L \qquad (1)$$

where C is the matched ontology class, w is the weight of the word, and L is the level of exploration. The value of L is initialised to 2 for each matching cycle.

The cycle continues by looking for the superclass (also a class) of the earlier matched ontology class. If a superclass is found, Pw_C is calculated for the matched superclass with the same w of the word but with the value of L decreased by 0.5. The cycle of looking for superclasses stops if a superclass of an earlier class cannot be found, or when $L = 0$, i.e. four levels of ontology classes that is related to the word have been found. The cycle is repeated for other words (and their respective weights).

After all the weighted words have been matched with the list of ontology classes, it is possible for some ontology classes to have more than one Pw values. The final Pw values of these classes are simply the summed value of their respective Pw values.

For our purpose, the higher the Pw value is for a particular ontology class, the high the chance for that class to be chosen as a category for a particular blog post. For this, the ontology classes are listed in descending order of Pw values. These can be considered as candidate categories. We then proceed with knowledge organisation to finalise the categorisation process.

3.2 Step 2: Knowledge Organisation

After the candidate categories for the acquired knowledge have been identified, knowledge organisation involves identifying relevant metadata (most suitable categories) which semantically enriches the knowledge repository (see Figure 2). These lead to the phase of storing of knowledge.

3.2.1 Storing of knowledge

In storing the knowledge (blog post), we have identified two approaches. The first approach is to use the n ontology classes (categories) with the highest Pw values. We then store these categories as annotated metadata to accompany the blog post in the knowledge repository. The advantage of this approach is that we can make semantic linkages to the categories with the highest Pw values (the most viable categories). The disadvantage is that other relevant categories that are not ranked the highest may be missed. Problems may also exist if the n-th and n+1-th categories tie with the same Pw values.

The second approach, which solves some of the issues of the first approach, would be to define a threshold value, Th. The Pw values of the categories are first put through a sigmoid function to obtain a normalised value (S) for each category's Pw value (see Eq. 2).

$$S_C = \frac{1}{1 + e^{-(Pw)}} \qquad (2)$$

For our purposes, we have set a reasonably high threshold value of between 0.95 to 0.98 so that categories with S values $> Th$ will be chosen as metadata for the blog entry. The advantage of this second approach is that categories are chosen based on whether they exceed the threshold value as opposed to just selecting the top n categories. This way, categories do not compete with each other if they have the same Pw value, rather they aim to exceed the threshold instead. The disadvantage is that some irrelevant categories may somehow still be chosen.

Before finalising the metadata, they are checked against a predefined list of categories that are deemed "too general" and are not suitable to be considered as metadata. For example, these may be categories which are too high up the ontology hierarchy. If these are found to be in the list of metadata, they are removed.

The result from this phase is an XML entry in the knowledge repository with the relevant `<title>`, `<link>`, `<description>`, and `<category>` tags. An example blog entry is as follows:

```
<post>
   <post_id>2</post_id>
   <post_date>2006-12-07</post_date>
   <title>Natural disaster in Wikipedia</title>
   <link>http://semblog/user1/post/2.txt</link>
   <description>A natural disaster is the consequence
      of the combination of a natural hazard (a
      physical event e.g. volcanic eruption,
      earthquake, landslide) and human activities.
      Human vulnerability, caused by the
      lack</description>
   <category cat1="human" cat2="primate"
      cat3="disaster-event" cat4="natural-hazard" />
</post>
```

3.3 Step 3: Knowledge Reuse

Knowledge reuse is carried out by forming semantic associations. This ensures that blog posts that are related to a particular blog post are effectively made available to blog readers via these semantic links (see Fig. 3).

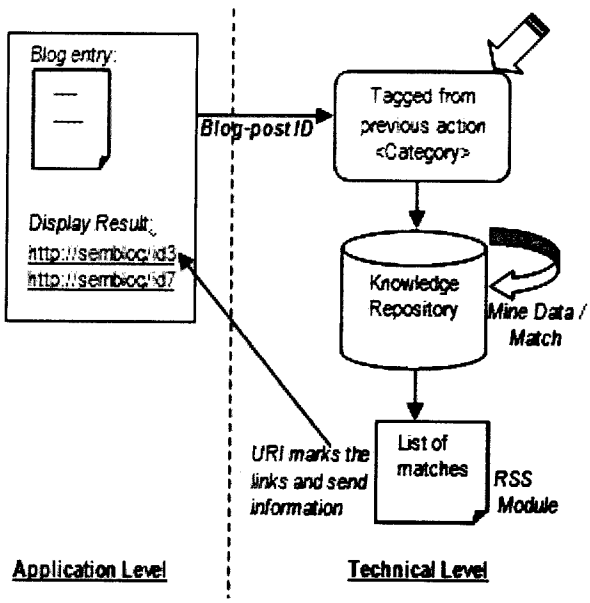

Fig. 3. Knowledge Reuse

3.3.1 Semantic link creation

When the metadata is identified for a new blog post, these are actually the semantic link to other blog posts. So, when a particular blog post is created (or retrieved, e.g. for viewing purposes), the XML knowledge repository is parsed to obtain the URL contained in the `<link>` tags of blog posts that have the same categories as the earlier created blog post. These URLs are also made available with the retrieved blog post. This process is repeated if the user chooses to visit any of these related URLs.

4. Discussion

We have previously identified the use of a threshold value to decide which candidate categories are chosen as metadata for a particular blog post. The choice of the threshold value is not arbitrary but is dependent on the length of the blog post. The realistic assumption here is that the longer the post, the more relevant words that can be found; and the shorter the post, the less relevant words that can be found. So, for our purpose, we propose threshold values as shown in Table 1.

Table 1. Blog post word count and respective threshold values

Blog post word count	Threshold, Th
< 200	0.98
>200 and <500	0.96
> 500	0.95

Table 2. Qualitative evaluation of related projects

Attributes	Our SEMblog	Haystack (4)	semiBlog (5)	Technorati (7)
Knowledge Based	The main objective (multiple flow)	Not defined	Information flow from source to repository	Not defined
Relationship	Semantic links via categorisation	Relationship based exploration - semantic links relationships	Does not specify any relationship	Track links to make out relevance of blogs.
Ontology	Navigate the ontology (OWL) and perform a calculation on metadata up the tree structure	Allow addition bootstrap files and content-loading through the RDF store	Not defined	Not defined
Semantic Associations	XML/RSS format of data repository to enable semantic links	XML/RDF and RSS format storing and access	Uses RDF to represent semantics	Not defined
Categorisation	Automatic categorisation	User-defined categorisation – Users create topic and categorise their content	Not defined	User-defined categorisation using tags
Integrated Environment	Integrate knowledge identification, organisation and reuse	Focuses on the Haystack Client server applications.	Integration between desktop application and the web.	Integrated approach of blogs and forum

From a storage point of view, our semantic knowledge repository uses RSS tags like `<title>`, `<link>` and `<description>`. We would like to highlight that by including SEMblog-related tags like `<post_id>`, `<post_date>` and `<category>`, we have enriched the basic RSS tags and this facilitates the search for related web pages based on the categories. We have also chosen XML for our knowledge repository and metadata as it is portable and platform independent.

Presently, we have not performed any quantitative analysis of the results obtained due to the fact that knowledge captured via blogs is by itself qualitative. However, we were able to make some qualitative comparisons with other related work in this area of semantic blogging. These are summarised in Table 2.

5. Conclusion

The key idea of our research is to find a way to manipulate knowledge, particularly on its identification, organisation and reuse, all in a seamless environment. Thus, we have created an architecture based on blogging to achieve this. We have also adapted web technologies such as semantic web, as well as ontology and NLP techniques. By doing so, we have managed to incorporate semantic web, ontology and NLP into a more user-friendly and practical application.

On the whole, we can see that we have multiple flows of knowledge, triggered by the single process of writing a blog post. When a user blogs, we identify the knowledge, decide on the categorisation in order to organise the knowledge, and make it retrievable or reusable. At the same time, previous knowledge (blog posts) that has been stored is made available to the user by providing links to web pages that are semantically related to the current blog post.

In the future, this system can be improved by focusing on NLP techniques, especially on language, grammar and vocabulary issues. Knowledge involves understanding and in order to make the machine understand the knowledge before conveying it to the users, we can use techniques like NLP or machine translations as an intermediary. Another possible prospect is adapting this research concept into organisational knowledge management. With this, important knowledge flowing in and out of the organisation through transactions or business processes can be identified, organised and reused within the organisation.

In conclusion, we have achieved our objective to provide a framework where users are able to create and access knowledge whilst actively identifying, organising, and reusing knowledge.

6. References

(1) P. Warren: Knowledge Management and the Semantic Web: From Scenario to Technology. *IEEE Intelligent Systems*, Vol. 21, pp. 53-59 (2006)

(2) N. Stojanovic and S. Handschuhs: A Framework for Knowledge Management on the Semantic Web, *The Eleventh International World Wide Web Conference*, Honolulu, Hawaii, USA (2002)

(3) P. Avesani, M. Cova, C. Hayes and P. Massa: Learning Contextualised Weblog Topics. In E. Adar, N. Glance and M. Hurst (eds), *Proceedings of WWW 2005 2nd Annual Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*. (2005)

(4) D. R. Karger and D. Quan: What would it mean to blog on the semantic web? In S.A. McIlraith, D. Plexousakis and F. van Harmelen (eds), *Third International Semantic Web Conference (ISWC-04)*, Hiroshima, Japan, *Vol. 3298 of Lecture Notes in Computer Science*, Springer (2004)

(5) K. Möller, J. Breslin and Stefan Decker: semiBlog – Semantic Publishing of Desktop Data, *Proceedings of the 14th Conference on Information Systems Development (ISD2005)*, Karlstad, Sweden (2005)

(6) Ontosem OWL. Available at: http://morpheus.cs.umbc.edu/aks1/ontosem.owl

(7) Technorati, Inc. Available at: http://technorati.com/about/