CHAPTER 3

ARCHITECTURE OF A TRUST-BASED PERSONALIZED
BLOG RANKING AND SUMMARIZATION SYSTEM

3.1 INTRODUCTION

Blog mining is defined as the process of extracting a useful pattern of blog information from the blogosphere, by incorporating web mining techniques in accordance with the blog data. Blog mining has evolved as a promising field, and it is being explored in various contexts and applications of blogs. Blog mining involves different application areas such as education, corporate businesses and politics. Though the traditional mining techniques have been applied for mining the blog data in different aspects using various methodologies, most of those techniques have not been tuned to deliver all the required and domain-specific knowledge to the users. In order to obtain such user-oriented and domain specific patterns of blog information, a novel architecture of a Trust-based Personalized Blog Ranking and Summarization system (TPBRS) is proposed, as shown in Figure 3.1. The architecture mainly consists of six functional blocks, such as the blog collector, blog filter, Personalized Semantic based Blog Retriever (PSBR), blog organizer, blog summarizer, and blog ranker.
Figure 3.1 Architecture of a Trust-based Personalized Blog Ranking and Summarization (TPBRS) System
The blog collector is used for collecting domain specific blogs with the use of domain ontology. The blog filter performs the process of filtering spam blogs by measuring the trust level of the blog content. The semantic based personalized blog retriever is the typical part of the architecture, that is used to perform personalization of blogs, and thus provide the blogs of interest blogs to the user. The blog organizer arranges the blogs into various clusters with respect to the similarity of their content. The blog summarizer is used to provide a brief summary of the various blog contents that belong to similar blog topics. In response to the user query, the ranker gives the ranked blogs that are based on the degree of blog content relevance.

3.2 BLOG COLLECTOR

The first phase of TPBRS is the blog collector phase. It consists of a domain specific ontology, the blog crawler, blog sites and blogML. The main aim of the blog collector is to collect different blog posts from various blog sites. For that, the blog collector makes use of the input domain ontology to collect domain specific blogs. A vast number of blogs are available on various topics, and also the content richness of the blogs varies depending upon the blog sites. The crawler starts performing the process of crawling blogs using the input seeds given in the input domain ontology. The blog collection involves the process of collecting the blog contents for various subject terms represented in the input domain ontology. The blog crawler is a parallel crawler that has been designed to simultaneously collect blogs from various blog sites. The blog details vary, depending upon the blog sites, from where they are being collected. Thus, it is primarily important to keep a common format for representing the collected blog data. The collected blogs are represented in a general format with the use of the Blog Markup Language (BlogML). BlogML is the incorporation of XML to support the storage and representation of blog information. Blog collection from different sources
enables the user to access the collected blog content from a single blog repository.

### 3.2.1 Ontology

Ontology is one of the ways of representing the domain knowledge in a structured and machine interpretable form. As ontology helps in retrieving the content effectively, the blog collector makes use of the input domain ontology to collect the blogs relevant for a specific domain. The ontology development tool Protégé (http://protege.stanford.edu/) is used for constructing the domain ontology. The ontology is constructed for the education domain with the relevant subjects as most of the blog sites contribute much of their benefits to e-learning based blog contents. The education domain ontology can be updated by adding new ontology classes for each subject, based on the mindmap created for every subject represented in the input ontology. The Blogs are available for all subjects in various blog sources, and have to be collected and related to one another. The ontology relationship between the successor and predecessor of each subject, is used to create a hierarchical relationship between blogs.

Let S be the generalized subject, s(⊆S) be specific subject and, r be the relationship between s and S as the ancestor, descendant and neighbor.

1. S is a superset of s.
2. \( s_1 \cup s_2 \cup \ldots s_n = S \)
3. \((\subseteq S1) \cap (\subseteq S2) = s, (i.e., s \subseteq S1 & s \subseteq S2)\)

Figures 3.2 shows the education domain ontology developed for collecting education oriented blogs from the blogosphere. The ontology starts
with the generalized subject term “Education” with the next sub-level about various types of “Education”, namely, “Pre-school Education”, “Primary Education”, “Secondary Education” and “Higher Education”. The ontology grows with the “Higher Education”. The relationship of the descendant with the ancestor is specified as “is-a” relationship. The subclass of one subject may be the subclass of another subject. One descendant may contain more than one superclass. Figure 3.3. shows the class hierarchy of the education domain ontology on the protégé editor screen.

Figure.3.2 Educational Domain Ontology
In this work, a domain specific ontology is developed with 207 subjects. Each subject has its corresponding classes, sub-classes and properties. The educational domain ontology is shown in Figure.3.2. The developed ontology subjects are available as an input to the blog crawler, to collect the blogs from various blog sites available in the blogosphere. Ontology plays a vital role in the proposed PSBR system. It is used to create the relationship between the collected blogs. The ontological terms act as the seeds for clustering blog posts.

3.2.2 Blog Crawler

In general, a crawler is an automated computer program that visits the given input link, and traverses through the web pages for collecting the
targeted content or links. For blog collection, the crawler makes use of input ontology terms as search keywords, for collecting the blogs from the given blog sites. Crawling starts from the search page, and collects all possible blogs from the blogosphere. The blog posts having common blog information are extracted. The most common information existing in all blogs are, the blog title, author of the blog, blog-posted date, URL, description, tags, and contents of the blog. The crawler digs deep into the blogosphere for collecting all the common blog information from the existing blogs. All the crawlers are independent of each other, as they have to exploit the way the blog template is represented on the blog site, each of which might have a different format. In this work, a parallel crawler is designed and used to collect blogs from different sources in the blogosphere simultaneously.

The pseudo code for the crawler collecting blogs from various blog sources is shown in Figure.3.4. Parallel blog crawlers are used to collect blogs from various blog sources as shown in Figure.3.5. The crawler is designed to extract the required information from the blog sites existing in the blogosphere. The details of blogs such as the title, URL, date, author, description, and content are all crawled, and stored in BlogML. The blogs are converted into a general format using BlogML. Figure.3.6 shows the sample crawled blogs from the blog sites. The sample crawled blogs are for the subject “windows mobile”, given in the input domain ontology. Blog crawlers collect blog posts from various blog sites in the blogosphere and store it in an XML file, that keeps all the blog information in a general format using BlogML. The collected blog posts consist of information such as the keyword, title, description, URL, date, author, and content. The sample blog post in Figure.3.6 shows the complete details of information collected from the blog sites.
Figure 3.4 Pseudo code for crawler

```
Function CrawlBlogsForSiteN(keyword)
{
    Crawl N blog(blogs) results for keyword
    For each blog in blogs
    {
        Extract Title, URL, Description, Date, Author Name, Tags
        Crawl URL
        {
            Retrieve Blog contents
            Retrieve Tags if not available in previous step
        }
        Insert into XML (Title, URL, Description, Date, Author Name, Tags, Content)
    }
}
```

Figure 3.5. Crawling the Blog sites in the blogosphere
3.2.3 BlogML

The crawled blogs are not in a uniform format, as they are collected from different blog sites. To make the blog information in a common format, the Blog Markup Language (BlogML) is used. The collected blogs are represented in BlogML. The blog details are marked with various representations of tags, that include blogpost, title, description, url, date, author, tags, frequency, and contents. Since the blog information is represented in a general format, it becomes a structured blog data. The XML format of blog content representation makes the blog retrieval process easy. Figure 3.7 shows the common BlogML to format the blog in a uniformed structures and the corresponding schema is represented in Figure 3.8. The formatted blogs are inserted in an XML file for further processing. Instead of
storing all the blogs in a single XML file, blog posts are stored in separate XML files using split mechanism that makes the blog processing an easier task. BlogML is the markup language especially used in this work, to generalize the blog posts that have been collected from many blog sites available in the blogosphere. A set of XML tags is defined to collect the common information from the blog sites, as shown in Figure.3.7. The XML schema is designed in this work to map all the necessary blog information in to XML files.

```xml
<?xml version="1.0"?>
<allblogs>
   <blogpost>
      <keyword> </keyword>
      <title> </title>
      <description> </description>
      <url></url>
      <date></date>
      <author></author>
      <tags></tags>
      <freq></freq>
      <contents></contents>
   </blogpost>
</allblogs>
```

Figure 3.7 XML elements in a blog repository
3.2.4 Blog Sites

The blog crawler collects the blogs from various blog sites. The blog sources and their corresponding blog URL are shown in Table 3.1. Blog crawlers are used to collect the blogs simultaneously from different blog sites, such as wordpress.com, blogs.technet.com and blogspot.com. Most of the blogs are collected from wordpress.com, due to the content richness of the blogs. Table 3.2 shows the statistics of blogs collected for the educational domain ontology, shown in Figure 3.2.

```
<?xml version="1.0"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema"

targetNamespace="http://www.w3schools.com"
xmlns="http://www.w3schools.com"

elementFormDefault="qualified">
<xs:element name="blogpost">
  <xs:complexType>
    <xs:sequence>
      <xs:element name="keyword" type="xs:string"/>
      <xs:element name="title" type="xs:string"/>
      <xs:element name="description" type="xs:string"/>
      <xs:element name="url" type="xs:string"/>
      <xs:element name="date" type="xs:date"/>
      <xs:element name="author" type="xs:string"/>
      <xs:element name="tags"="xs:string"/>
      <xs:element name="freq" type="xs:string"/>
      <xs:element name="contents" type="xs:string"/>
    </xs:sequence>
  </xs:complexType>
</xs:element>
</xs:schema>
```

Figure 3.8 XML Schema for Blog Details
Table 3.1 Blog Sources

<table>
<thead>
<tr>
<th>Number</th>
<th>Blog Source</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Wordpress</td>
<td><a href="http://wordpress.com">http://wordpress.com</a></td>
</tr>
<tr>
<td>2.</td>
<td>Blogspot</td>
<td><a href="http://www.blogger.com/home">http://www.blogger.com/home</a></td>
</tr>
<tr>
<td>4.</td>
<td>Technorati</td>
<td><a href="http://technorati.com">http://technorati.com</a></td>
</tr>
<tr>
<td>5.</td>
<td>Blogger</td>
<td><a href="http://www.blogger.com">www.blogger.com</a></td>
</tr>
</tbody>
</table>

Table 3.2 Statistics of Blogs collected for Education Domain Ontology

<table>
<thead>
<tr>
<th>Number</th>
<th>Description</th>
<th>Number of Subject Related Blogs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Number of subjects available in ontology</td>
<td>207</td>
</tr>
<tr>
<td>2.</td>
<td>Number of blogs to be collected per subject by the crawler</td>
<td>50</td>
</tr>
<tr>
<td>3.</td>
<td>Total number of blogs to be collected for all subjects</td>
<td>10350</td>
</tr>
<tr>
<td>4.</td>
<td>Number of blogs collected by the blog crawler</td>
<td>9854</td>
</tr>
</tbody>
</table>

3.3 BLOG FILTER

The blogs are collected from various blog sites. The blogs might have irrelevant information, even though it has been published with the focus
on a specific topic. In many cases, the bloggers’ blog style varies, as they would have different views and thoughts while writing the blogs on a particular topic. When these kinds of blogs are collected from the blogosphere, they would contain different language blogs, irrelevant to the ontology subject and with incomplete content. Thus, it is ultimately important to filter out the blog posts. Therefore, blogs are preprocessed, and also checked for the credibility of trust, based on reputation. The credibility of the blogs is analyzed based on the reputation of the bloggers and trustworthiness of the posted blog content. The filtering process consists of two important tasks, that are, pre-processing and inspecting the trustworthiness of blogs, which leads to the removal of spam blogs.

3.3.1 Blog Pre-Processor

Pre-processing is necessary to obtain the needed information, and helps to achieve further blog related tasks easily, with the use of the preprocessed blog information. Blog preprocessing involves the processes of removing redundant blog details, irrelevant blog content, non-English textual content and incomplete blog posts. The dates on which the blogs were posted, would result in different formats, depending upon the blog sites. For instance, a blog site may use the date format as “dd/mm/yyyy” and the other one may use “dd-mm-yyyy”. Therefore, it is necessary to standardize the posted blog dates with an identical date format. Thus, the blog post dates are converted into a common format. The frequently used words in the blogs are identified, and those words are referred to as tags. The count of each tag is specified as its corresponding frequency.

3.3.2 Trusted Blog System

The trustworthiness of the blogs is computed by analyzing a few trust parameters, such as the reputation of the bloggers, and quality of the
blog content. The reputation of the bloggers is identified by their typical ways of blogging, like creating blogs often, editing the blog contents, and replying to blog comments. The blogs are evaluated based on the trust, and it helps to reveal the reliability of the blog sources. Sometimes, the blog content itself shows the bloggers’ technical and publishing characteristics. The credibility of a blog is found by analyzing the writing style of different bloggers. Identifying a trusted blog provides better usability for the user. Users can read the blogs with trust. Since the blogs are checked for credibility, only highly trusted blogs are processed for further tasks.

3.4 SEMANTIC ANNOTATOR

Blogs of one subject may have some relationship with another subject. Before storing the blogs in the blog repository, the blogs need to be semantically annotated. To annotate, the input domain ontology terms are checked for similar words from the WordNet (http://Wordnet.princeton.edu/). In addition to the ontological terms, the similar words of ontology terms obtained from the WordNet are also added, which helps to annotate the blogs. These annotations help in the retrieval of blogs, as there is a possibility of the existence of blogs with similar words, rather than exact words.

3.5 WORDNET

WordNet is an online lexical reference system, which is designed in such a way that it consists of English nouns, verbs, adjectives and adverbs that are organized into synonym sets (synsets), each representing one underlying lexical concept. The goal of the WordNet is the creation of a dictionary and thesaurus, which could be used automatically. The WordNet is also useful for determining the semantic connections between sets of synonyms, for tracing the morphological connections between words. The ontology is organized not only by the “is-the-synonym-of” relation, but also
by the hierarchically organized verbs and nouns via the hyponym/hypernym relation (Moravec et al 2004).

The term list is verified for having multiple meanings. These terms are stated as valid, if they have only one or don’t have any synonyms. Such terms need not be disambiguated. Terms having more than one meaning indicate, that there is a disambiguation, and it has to be cleared. The (Lesk 1986) algorithm that makes use of WordNet for word sense disambiguation, is used to disambiguate the terms. When the algorithm finds two or more words for a term, then those words have to be added to the existing term; thus, that term is disambiguated. Disambiguation is removed by the inclusion of a term that clearly defines the ambiguous term. The effectiveness of disambiguation is highly dependent on the accuracy of the dictionary.

The input domain ontology subjects are explored with the WordNet to get the relevant synonyms and hypernyms, and the same can be used to create the relation between various blogs.

3.6 BLOG ORGANIZER

The number of blogs is increasing every day. The blog search engines retrieve a high volume of blogs. It is very tedious for a naive user to read every blog. One of the ways of organizing the blogs, is categorizing them, based on their content relevancy. It is ultimately important for the TPBRS system to organize the blogs, to make the retrieval process an easy one. The pre-processed blogs are organized by applying the clustering process. Clustering can be used for organizing the blogs into various clusters, which assist in the retrieval of blogs.
3.6.1 Blog Clustering System

Clustering is the process of grouping the blogs into similar types, based on their relevant content. By clustering those relevant blogs into similar content relevant categories, their retrieval becomes an easy task. It also improves the content organization. Various traditional clustering techniques can be applied for the clustering of blogs. In addition, it is better to use some efficient clustering methods, such as hierarchical agglomerative clustering to effectively cluster the blogs in a tree structure. The clustering method used, is highly efficient to organize the blogs into a proper structure. This improves the retrieval time and efficiency of blog search.

3.6.2 Clustered Blog Repository

Blogs are categorized, by performing hierarchical agglomerative clustering. Since the blogs are clustered hierarchically, retrieval becomes easy. Once the blogs have been clustered, they have to be stored in the repository as multiple segregated clusters for further usage or blog retrieval. Since the clustering of blogs is represented in a tree structure, the process of blog searching becomes easy, and the time complexity is considerably reduced.

3.7 BLOG RANKING SYSTEM

In general, blogs are retrieved in response to the user given keyword, and the results are simply provided without any sort of preference-based ordering in the search results. There may be thousands of blogs on a particular topic. In such cases, the blog readers or content searchers get overwhelmed with the blog contents, and they face difficulty in reading their intended blog content. The readers are benefitted if the blogs are ranked in the order of the most relevant blog content to the user query. For a simple user
query, there may be huge amount of blogs retrieved. The ranking is mainly performed, based on the degree of relevance, and the results of ranking provide better blog contents, i.e., what the users intended to obtain. This helps to rank the blogs and provides the users with an efficient blog retrieval in a sorted order.

3.7.1 Keyword Indexer

The main task in ranking is the process of maintaining the most important as well as the most often accessed terms. By keeping the list of those terms separately, the retrieval process becomes very easy, and also improves the ranking process. Indexing is the process of maintaining the frequently accessed terms for reducing the searching time, and improving the searching process. The frequently used words in the blog are referred to as tags, that are being taken into consideration as indexing terms. Indexing terms contribute greatly in accessing the blogs in less time. Such indexing terms are later used for the efficient retrieval of blog contents.

3.7.2 Similarity Analyzer

The similarity of the blog contents is checked, and the content relevancy is measured by computing the TF-IDF (Term Frequency and Inverse Document Frequency) of the search keyword relevant to the blog content. The given user query is compared with the indexed tags, and the blog similarity analyzer parses the content, to find the TF-IDF of each word in the blog. Once the TF-IDF has been found, the similarity value is calculated to rank the blogs according to the relevance score. The blogs are ranked, based on the relevance score that has been calculated, with respect to the similarity of blog content. The similarity analyzer plays an important role in sorting the blogs in an order of content relevancy. The blogs are evaluated for their content similarity and their content richness is taken into consideration in the process of ranking.
3.8 PERSONALIZED SEMANTIC BASED BLOG RETRIEVAL

The main motive behind personalization is to provide more relevant blog content to the users, pertaining to their interests. The number of blogs is considerably reduced and most of the times those blogs are not relevant to the blog readers. The user personalization mainly involves the processes of maintaining the blog readers’ search interest and their usage logs. The user profile information is also considered as an important criterion, because based on the user profile the blog search results is produced. The blog personalization involves the process of storing the general profile and user profile of the blog readers. The general profile collects the user details, that may include the users’ interest, and information about their educational or working domain. The user profile maintains information regarding the users’ previous search logs. These profiles’ information is the key information, that helps to personalize the blogs effectively.

3.9 BLOG SUMMARIZER

Summarization is the process of giving a brief content on a specific topic, by consolidating the information from various resources. Due to the increased availability of blogs on many domains, users are overwhelmed by the content they are searching for. Though the blog contents are large in number, they are not easily understandable, as they are highly descriptive. It would be better if the blog contents are analyzed for a similar or relevant content, and then combining all the contents that focus on a common topic. The blog summary can be generated to assist the users for getting instant focused content in a very less amount of time. Most of the traditional text based summarizations just concentrate on the available sentences, and simply combine them for producing the summary. Such kind of generated summaries are not always effective. Blog summarization is the process of combining one or more blog contents focusing on a common topic. The blog summarization
can be performed in two ways. First, the blogs are summarized content-wise. Second, the blog repository is summarized to give the overview of subject content in the repository. To summarize the blog content, a new method is used to summarize the minimal number of sentences with meaningful content. Well-known measures like the TF-IDF, the similarity measure, Jaccard coefficient, word matching and stop-words are used for obtaining better summarization results. The blog summarizer yields the summarized version of multiple blogs from the blog repository.

3.10 SUMMARY

- The TPBRS system consists mainly of the functional blocks, viz, the blog collector, blog filtering system, blog organizer, PSBR system, blog summarizer and blog ranker.

- Despite the existing approaches related to various blog analyzing techniques, the TPBRS focuses on personalization, summarization of the blog contents, and ranking of the blogs at a fine-grained level, for effectively processing the blog content.

- In the literature survey, after analyzing the various experimental studies, it has been clearly found that most of the earlier blog content analyzing techniques have not provided any concrete solution to assure the trustworthiness of the blogs at a coarse level, by analyzing blog content oriented trust parameters, that are the content quality and reputation of the bloggers. The approach which we have proposed, ensures the trust level of the blog content by inspecting and measuring various trust oriented parameters, to lead to a better trust-based blog system.
• Personalization of blogs is implemented based on the user profiles. The blog personalization has achieved ultimate importance in TPBRS, when compared to the personalization that had been applied in general web personalization.

• Since the vast amount of blogs have been collected during the blog collection stage, when it comes to produce the blog search results to the users, the ultimate task is to minimize the number of blogs, as well as briefing the blog content, while they share the content on a similar topic. The novel text summarization technique has been incorporated in TPBRS for summarizing the blog contents.

• The organization of the blogs is achieved in TPBRS by implementing the effective clustering technique, known as hierarchical agglomerative clustering, which uses the cosine similarity measure.

• In order to produce efficient results over the clustered blogs, and offer a better improved retrieval process, the blogs are ranked based on content-based importance.