CHAPTER 2

LITERATURE SURVEY

2.1 INTRODUCTION

Web mining aims to discover useful information or knowledge from the web hyperlink structure, web page content and web usage data. Web structure mining and web content mining are the two important areas in which different tasks are performed, such as the link analysis, information retrieval, structured data extraction, and information integration. Mainly, web structure mining aims to generate a structural summary about web sites and web pages. For instance, given a collection of interconnected web documents, interesting and informative facts describing the connectivity of those documents in the web can be discovered (Madria et al 1999). One of the web mining research works carried out by Kosala and Bloomkeel (2000), discussed the categories, of web mining and their corresponding data format, representation, methods and application categories that include categorization, clustering, personalization, summarization and ranking of web documents.

Web content mining refers to the discovery of useful information from the Web contents. It encompasses resource discovery from the Web (Chakrabarti et al 1999; Cho et al 1998), document categorization and clustering (Kohonen et al 2000; Zamir and Etzioni 1999), and information extraction from the Web pages (Chang et al 2003; Tolle and Chen 2000). Many search engines are available on the Internet, each having its own
characteristics, and employing different algorithms to index, rank and present web documents.

Web mining is used in various applications for all processes like classification, document clustering, personalizing the web content and ranking. Krishnapuram et al (2001) presented various clustering algorithms that are used for clustering web documents based on content similarity. Such algorithms are very useful in web mining applications, such as categorization of web documents, snippets, and user sessions.

Most of the web documents available on the Internet are defined through the Hyper Text Markup Language (HTML), that allows an author to organize the presentation of document content by means of special tags, and interpreted by web browsers. Web documents can contain both multimedia information, and connections to other documents through hyperlinks. A hyperlink is often created based on the principle that links are connections among documents that are similar. Hyperlinks are increasingly being used to improve the ability to organize, search and analyze the Web (Brin and Page 1998; Yang et al 2002).

Content management is an important task in organizing the web content. Zettsu et al (2006) discussed knowledge management based on connecting shared intelligence on the Web. The Web can be viewed as a field for individual propagation of personal intelligence with users pursuing their own self interests. A community aggregates personal intelligence, based on its own conceptualization, and forms common understandings as collective intelligence. Furthermore, Zettsu et al (2006) discussed the resulting impacts on the knowledge life cycle, and the challenge to overcome the new paradigm of knowledge management. They also discussed technical issues in implementing knowledge management, based on shared web intelligence. These range from Web information searches, knowledge discovery and data
mining, to a global infrastructure for knowledge sharing, and aggregation of knowledge based on conceptualization in exploiting collective intelligence.

One of the main web mining applications used in content management system is e-learning, that uses web content mining. Chen et al (2009) reported an experience in designing a class project for students in a graduate course, to use open web APIs to develop web mining applications. It is useful to introduce web-based open source software for teaching web mining topics, and study its impact on students’ learning.

2.2 DIRECTIONS OF THE WORK

Information retrieval and knowledge extraction are always areas of research, either in text or web documents. Various web mining techniques provide abundant and essential ways for effectively extracting and, organizing the knowledge contained in many different web pages. Due to the evolution of semantic web standards, there is a shift in the way in which the web is being used. Social networking and the semantic web play an important role in Web 2.0 and Web 3.0 respectively. Blogging is a major area, through which communities in social networking share their knowledge, experience, opinion, and feedback with respect to technical, nontechnical, commercial, personal, political, and educational subjects. These blogs which are posted by different bloggers are stored in different blog sites. These blog contents are viewed similar to web documents stored in web sites. Over a period of time, a large volume of blog contents is made available in the blogosphere. This motivated many researchers to switch their focus from web documents to the blog repository. Many of the web mining techniques which were applied on web documents, were used by the research community (either with necessary modification or with new proposals) to extract useful information and knowledge from the blogosphere.
Our research work also aims to explore information retrieval from the blogosphere. Based on the research articles, it is also understood that the information extracted from various blog sites pertaining to different subjects, may provide value added content in the process of e-learning. This has motivated to do a literature study involving the blog and the role of a blog in e-learning. The rest of the sections discuss the major works carried out relating to blog trust, clustering, personalization, summarization, ranking, and the uses of these of blogs in an e-learning environment.

2.2.1 Trends of Social Networking

The emerging use of the social media, such as social networking sites, wikis and blogs change people’s lives fundamentally, and influence the handling of knowledge and information (Kolbitsch and Maurer 2006). There is a discussion on whether these applications and other technology innovations have produced significant differences, between the earlier and the current generation of people (Oblinger and Oblinger 2005). A social network analysis is useful for the web, because the web is essentially a virtual society or a virtual social network where each page can be regarded as a social actor, and each hyperlink as a relationship. Many of the results from social networks can be adapted and extended for use in the web context. The ideas contributed from the social network analysis are indeed foundational to the success of Web search engines (Liu 2006).

The availability of wealthy social network information, such as online social networking and media sharing sites, have resulted in interesting and innovative research work in recent years. While there is a large body of research on different problems and methods for social network mining, there is a gap between the techniques developed by the research community and their deployment in real-world applications (Bonchi et al 2011). There are many social networking web sites which are popularly used by many people,
and are making a lot of impact on the social network users. Table 2.1 shows the list of popular social networking web sites.

Table 2.1 Statistics of Prominent Social Networking Sites

<table>
<thead>
<tr>
<th>Social Networking sites</th>
<th>Total Active Users</th>
<th>Average Time spent by each User</th>
<th>Companies using Social Media</th>
<th>Top most Countries as Customers</th>
</tr>
</thead>
<tbody>
<tr>
<td>FACEBOOK</td>
<td>850 million</td>
<td>15 hours and 33 minutes/month</td>
<td>71%</td>
<td>USA, Indonesia, Mexico</td>
</tr>
<tr>
<td>TWITTER</td>
<td>465 million</td>
<td>6+ hours per week</td>
<td>59%</td>
<td>USA, Brazil, Japan</td>
</tr>
<tr>
<td>LINKEDIN</td>
<td>150 million</td>
<td>15 minutes 40 seconds/day</td>
<td>44%</td>
<td>USA, Europe</td>
</tr>
<tr>
<td>YOUTUBE</td>
<td>490 million</td>
<td>900 seconds/day</td>
<td>33%</td>
<td>USA, Japan, Germany</td>
</tr>
</tbody>
</table>

Teenagers and young adults use the social media to a very great extent. According to Caruso and Salaway (2009), 88.3% of the undergraduate students had their own laptops in 2009, compared to 65.9% in 2006, and spent an average of 21.3 hours per week online. Moreover, 80.0% of these students rated their information literacy as “very skilled” to “expert”. Some voices in the current debate have even demanded a change in education, in order to respond adequately to the frequent use of the social media by young people. Such a demand should, however, not simply be raised, because the social media is an inherent part of the life of today's teenagers and young adults. It should also be recognized, whether and how these applications are able to promote learning, and knowledge processes. The social media provide
multiple opportunities that may be exploited for learning and knowledge processes in general (Wodzicki et al. 2012).

Social annotations are informative resources for semantic based search and useful information in many applications. Although social annotations are developing fast, they only cover a small portion of the fast growing World Wide Web, and thus suffer from the sparseness problem. Bao et al. (2008) proposed a general propagation model based on a random surfer, that is used to boost the social annotations of web pages. A systematic approach is needed to understand the impacts of the Sense of Community (SOC) on individual’s continued usage behavior in the social networking environment. Built upon theories from the well-established Information Systems (ISs) adoption, usage, and success frameworks, Zhang (2010) developed a theoretical model that integrated SOC and mapped out its structural interactions with other key success factors such as Information Quality (IQ), System Quality (SQ), and user satisfaction, to predict usage. Zhang (2010) modeled the SOC as a formative construct, which consists of multiple independent components.

Social networking is widely used in e-learning. Researchers and academicians publish various articles in social networking sites like Facebook, Twitter and different blogging sites. Conole and Culver (2010) introduced a new social networking site, Cloudworks, which aims to provide a dynamic environment for finding, sharing and discussing learning and teaching ideas and designs. New technologies offer a multitude of opportunities for the creation of innovative, engaging and pedagogically effective learning opportunities. A wiki is a collaborative online space, where knowledge is presented, edited and accepted by those creating the wiki in a style that is open, accessible and responsive to change. Wikis can therefore
create a change in the social values of the learning group, and can be significant to the learning outcomes (Morley 2012).

Most recently, SNA methods have been applied to weblogs. Kumar et al (2003) observed and modeled temporally concentrated bursts of connectivity within blog communities over time, concluding that the blogspace has been expanding rapidly. Adar et al (2004) identified blogs that initiate “information epidemics” and visualize the paths specific infections will take through blogspace. Marlow (2004) used the social network analysis to identify “authoritative” blog authors, and compare them with measures of opinion leadership and authority in the popular press; Delwiche (2005) discussed about the most authoritative blog authors (Xin Li 2009).

One of the most visible trends on the web is the emergence of social web sites which help people to create and gather knowledge, by simplifying user contributions via blogs, tagging and folksonomies, wikis, podcasts, and online social networks. The social web has enabled community-based knowledge acquisition with efforts such as Wikipedia, demonstrating the “wisdom of the crowd” in creating the world’s largest online encyclopedia. Although the process of defining the structures or abstractions of the social web is difficult, it facilitates collaboration and sharing among users, although usually on single sites. Current online-community sites are isolated from one another, like islands in a sea (Uldis Bojārs 2008).

2.2.2 E-Learning

E-learning has the potential to become a cheaper and more efficient educational tool (White 2003). E-learning environment includes humans and learning resources. With continuous interaction between the humans, resources, and the environment, sharable resources are accumulated and better organized. Human knowledge is developed, and therefore, more useful
semantic relations are discovered. Always with the expansion of the linked resources, discovering the hidden semantic community, and the emerging semantic relations in a large network of resources becomes increasingly important.

An e-learning system can select the sequence of web resources and link them into a coherent focused organization (Farrell et al. 2003). E-learning can automatically generate an individual learning path from a repository of XML-based Web resources (Brusilovsky 2004). Many new technologies for e-learning over the Internet have been introduced (Li et al. 2008). The basic issue of realizing effective web-based learning is to discover the interested content, and the basic semantic relation in a large-scale network of contents.

It is important to provide the learning content in accordance with the learners’ context. A number of works have been published, that discuss the process of learning with the assistance of computers, or more specifically, with the assistance of the World Wide Web. For instance, Stojanovic et al. (2001) elaborated ways to search the Semantic Web (Berners-Lee 2000) to enhance e-learning possibilities. The authors noted that it is important to consider the learners’ context in relation to the material they are presented with. Schmidt and Winterhalter (2003) discussed context aware e-learning. Such learners’ context based e-learning systems require semantic, or metadata information about the learning content, in order to form appropriate searches for the learner. In contrast, the model presented in Muhlestein and Lim (2011) infers the context of the learner, by examining the resources commonly accessed with those they are currently requesting.

Once people are familiar with accessing their context oriented e-learning based sites, they prefer better comprehensibility among the e-learning sites. To understand a particular concept in these knowledge bases, a reader needs to learn about related and underlying concepts. For example,
blogs, research repositories, and digital libraries quite often do not link to definitions of the concepts contained in their texts and metadata (Gardner et al 2009).

In order to compromise with the lacking facilities in blogging utilities and blog sites, metadata or tags of blog content have to be taken into consideration for providing an effective e-learning system. To support all these necessities, there is a need for a semantic based e-learning system.

2.2.3 Blogging

Various blog-related topics are described in section 1.4.1 of chapter 1. Blogs also combine personal web pages with tools, to make linking to other pages easier, as well as to post comments and afterthoughts (Blood 2004). Instead of having a few people in control of the threads on traditional Internet forums, blogs basically allow anyone to express his/her ideas and thoughts. Many free blog hosting sites are available to publish blogs. Sites like www.blogger.com, www.xanga.com, and www.livejournal.com, are examples of blog sites, where the blogger can publish blog post (Chau and Xu, 2006).

Weblogs (or blogs) are becoming a new form of mainstream personal communication (Rosenbloom 2004) for millions of people, to publish and exchange knowledge or information, and to establish networks or build relationships in the world of all blogs, called as blogosphere. Weblog-building technologies (or blogging tools) bring new capabilities, such as web publication and communication, to average people, especially nontechnical users. They are designed to facilitate simple and fast creation of web content, without much technical or programming skill. Recent releases of blogging tools have been further improved to provide enhanced features for blog
interactivity; thus promoting the creation of social networks among bloggers (Du and Wagner 2006).

There are two basic blog styles: filters and journals. The filter style focuses on a collection of links to other Web sites. The journal style is an online personal diary with dated entries presented in a “stream of consciousness.” Both styles use headlines and excerpts, putting the most recent entry at the top of the web page to entice readers to investigate further (Lindahl, 2003).

An analysis of BlogPulse.com provides many examples of the types of topics that are found in blogs, such as financial, political, entertainment, and news. As a result, there are millions of blogs. Accordingly, substantial information is put in the blogosphere, of which much may be redundant and un-correlated. Since the information is coming from so many different sources, at different times, there may be false information, not previously realized or recognized, also embedded in the blogs. Blogs are likely to represent a single individual or a group. Blog information may differ from other kinds of text. For example, blogs are not likely to be as well edited, as a newspaper or magazine text. In contrast to other forms of text, blogs may use incomplete sentences and phrases (O’Leary 2011).

Recently, user generated data has been growing rapidly into one of the most important sources of information in the web. This blog content has a lot of information to be processed like opinion, experience, personal, technology and customer reviews, which can be used for analysis or knowledge sharing. Forums, mailing lists, on-line discussions, community question answering sites and social networks like facebook, are some of these information resources that have attracted researchers. The blogosphere (the collection of blogs on the web) is the main source of information for all
domains. Millions of people write about their experience and opinion in their blogs every day, and this provides a huge amount of information to be processed (Keikha and Crestani 2011).

2.2.4 Role of Ontology in Blogs and Social Semantic Web

The semantic web is an evolving extension of the World Wide Web, in which the semantics of information and services are defined, making it possible for web-based tools to understand and satisfy the requests of people and machines to exploit web content (Berners-Lee et al 2001). This model requires a set of knowledge structures to formalize those semantics, and a linkage between the web content and those structures. The semantic web relies on two basic components, such as ontologies and annotations. Ontologies are knowledge structures representing the semantics of domain concepts and their interrelations in a machine-readable way.

The role of ontologies in the architecture of the semantic web and the key elements of the XML and RDF, using the XML as a tool for semantic interoperability, will be ineffective in the long run (Decker et al 2000). Representation and an inference layer is needed on top of the current layers of the web, and to establish such a layer, Decker et al (2000) proposed a general method for encoding ontology representation languages into RDF/RDF schema. There is a lot of family resemblance between blogs and the semantic web model. Blogs enable the average user to talk about, i.e., annotate resources on the web and publish these annotations for others to see. A large portion of these blogs already have machine-readable (table of contents) files encoded in an XML format called RSS. Furthermore, blog entries themselves can be searched over, replied to and referred to in other blogs (Karger and Quan 2005).
Despite the availability of an immense number of blogs, the background knowledge about the blog content of various domains contained in the blogosphere has to be encoded, in accordance with the necessity of the applications. In such cases, ontologies play an ultimate role to encode the background knowledge of specific domains.

A methodology to partially annotate the textual content of web resources in an automatic and unsupervised way is proposed by Sánchez et al (2011). It has used several well-established learning techniques and heuristics, to discover the relevant entities in the text and to associate with classes of an input ontology by means of linguistic patterns. Semantic web mining can improve the results of web mining by exploiting new semantic structures in the web. Stumme et al (2006) focused on an idea to enable search engines and other programs to better understand the content of web pages and sites. The ultimate goal of semantic web mining is “a better web” for all of its users, a “better usable web”. Rossi et al (2006) presented an analysis of different dissimilarity measures, based on the comparison between the semantic structure of the site identified by experts, and the clustering constructed with standard algorithms applied to the dissimilarity matrices generated by the chosen measures. They have shown that the design and the structure of the reference web site can have a strong impact on the outcome of the comparison.

As the semantic web has evolved vastly, various mining techniques and approaches have to be incorporated to enable an easy exchange and understanding of information, based on user-system interaction.
2.2.5 Need for Mining Blogs

Though blogs may share many similarities with web and text documents, existing techniques need to be reevaluated and adapted for mining the multidimensional representations of blog data, which exhibit dimensions not present in traditional documents, such as tags. Blog tags are semantic annotations in blogs, which can be valuable sources of additional labels for myriads of blog documents (Tsai 2011).

A tag is a keyword that can be used to describe a blog. The tag metadata is useful for users to quickly find related blog entries that are tagged to a topic of interest. Tags can be chosen by the blogger, the viewer, or both. If many users tag many items, this tag collection forms a folksonomy. Tagging was popularized by the web 2.0, and is an important feature of many existing services. Many blog systems allow bloggers to add new tags to a post, in addition to placing the post into categories. If a post is incorrectly classified, a blogger can edit the list of tags. The analysis of large data of multiple tags may require the use of dimensionality reduction or projection techniques, to transform the data into smaller sets. Dimensionality reduction finds a smaller set of features that can describe the original set of observed dimensions. Dimensionality reduction can uncover hidden structures which are useful to understand and visualize the data (Tsai 2011). Michael Chau (2009) proposed a general framework for different tasks. It consists of a blog spider, a blog parser, a blog content analyzer, a blog network analyzer, and a blog visualizer.

Though various research works have been done on the blogosphere for digging different blog content, current social networking based applications such as blogs, need more domain or user specific blog mining tasks. Considering these necessities, there needs a system that highly focuses and supports mining or extracting need-based blogs from the blogosphere.
2.2.6 Reputation of Blog Content

Trust is an integral component of many kinds of human interaction, allowing people to act under uncertainty and with the risk of negative consequences. Trust judgments on the web are made by humans based on their prior knowledge about a source’s perceived reputation, or past personal experience about its quality, relative to other alternative sources they may consider. However, weblogs can also be misused to influence and manipulate the readers. Therefore, the credibility of a blog has to be validated before the available information is used for analysis. The credibility of a blog entry is derived from the content, the reliability of the author or blog, and the external references or trackbacks. Trust can be generally categorized as Policy-based, Reputation-based, and Model-based trust, and trust in information resources (Artz and Gil 2007). In particular, trust is a critical factor in Internet applications such as web services, and various online communities or social networks (Ziegler and Lausen 2005; Huang and David 2010).

There are some undesirable contents in the blogosphere to dilute the trustworthiness of blogs; such kinds of blogs are known as spam blogs (splogs). The growth in the numbers of blogs has led to an alarming increase in the number of splogs. Splogs have a detrimental effect on the blogosphere. Splogs not only affect the retrieval of the original blog content, but also mean to attract search engine traffic, used solely for promoting affiliate sites. The splog detection problem is made difficult due to the lack of stable content descriptors. Yu-Ru Lin et al (2008) developed a new technique for detecting splogs, based on the observation that a blog is a dynamic, growing sequence of entries (or posts) rather than a collection of individual pages. In the approach, described by Yu-Ru Lin et al (2008), splogs are recognized by their temporal characteristics and the content of the blog.
In order to preserve the quality of blogs from splogs, various mechanisms have been evolved. A blog becomes a viral marketing site based on peer-production, and it is promoted yet induced by online person to person interactions. Moreover, there exists a large number of information in the blogosphere, including text-based blog entries (articles) and profiles, pictures or figures, and multimedia resources, which serve as an incentive to develop a blog recommender approach, and design an information filtering mechanism. Li and Chen (2009) proposed a blog recommendation mechanism that combines a trust model, social relation and semantic analysis, and illustrates how it can be applied to a prestigious online blogging system.

From the above research studies it is clear there is a need for computing the trust factor in order to handle the collected blogs effectively. The trust factor has to include various trust related parameters that involve all aspects of the blog content.

2.2.7 Blog Categorization

Many researchers have attempted to employ the techniques of traditional text/document clustering, referred to as long text clustering (Bolelli 2007; Zhao 2002; Zhong 2005) to reorganize search results, and then use the cluster labeling method (Treeratpituk and Callan 2006; Cutting 1993) to assign a readable label to each cluster. These clustering methods usually download all the documents of the search results, and perform clustering on them. However, downloading all the documents from the Web is very time-consuming; hence, many researchers try to cluster the title and snippets of the search result (Wang and Zhai 2007; Zeng 2004). Generally, these titles and snippets are referred to as short text, since their size is much smaller than that of the entire document; this kind of clustering process is referred to as short text clustering.
Many social resource-sharing systems have adopted tags to manage their resources. Tags can be helpful for resource sharing and discovery, because they provide meaningful descriptions of resources and help to find what the users want to achieve more easily. Tags in the same clusters are relevant; the resources annotated by these tags also tend to be relevant, as they can help to organize the resources more effectively. Thus, the resource-sharing problem can be alleviated by tag clustering, and users can find the desired resources not only through their own tags, but also through a group of relevant tags (Cui et al. 2011).

In addition to tag based clustering, more methods have to be adopted for analyzing the blogosphere to visualize the clustered blogs. One of the proposed methods is for analyzing a large number of blogs, which introduces a graphical interface to aid in data visualization. Choi and Krishnamoorthy (2007) proposed an algorithm that will return a matrix containing all the similarity values, at which point the K-means algorithm is used to divide the blogs into a number of categories (Choi and Krishnamoorthy 2007).

Apart from these existing clustering approaches, various needs will arise with the increasing growth of the blogosphere. They require new clustering techniques to be developed for functioning beyond their current scope.

2.2.8 Personalized Blogs

Generally, the interface to most of the web search engines is simple. A user enters a few words into a search box and receives a long list of results in return. Despite the simplicity of the interaction, people use Web search engines for many complex tasks, such as research, planning trips, entertaining themselves, purchase items, and find new jobs. The challenge for
a search engine is to translate people’s simple, short queries into lists of documents that satisfy the information needs (Teevan et al 2010). Mostly, search engines presume every user as their typical content searchers.

Most of the web search engines use the content of the web documents and their link structures to assess the relevance of the document to the user’s query. With the growth of the information available on the web, it becomes difficult for such web search engines to satisfy the user information needs expressed in a few keywords. Daoud et al., (2009) presented a personalized search approach that uses a graph-based representation of the user profile. Zeng et al (2008) developed a framework for user activity analysis, based on user interest. The analysis is based on the user interest in the articles that he/she selects to reply, and the articles he/she does not select to reply.

Some personalized techniques make a distinction between long term and short term user profiles and investigate either short term ones (Gauch 2003; Shen 2005) or long term ones (Shen 2005) in a personalized document ranking. A short term user profile refers to the user interests during a short period of time, and inferred from the recent search history. A long term user profile holds persistent user interests generally stable for a long time, and inferred from the whole user search history (Daoud et al 2009).

According to a recent survey made by Technorati (http://technorati.com/), there are about 75,000 new blogs and 1.2 million new posts every day. However, it is difficult and time consuming for a blog reader to find the most interesting posts in the huge and dynamic blog world. An online Personalized Blog Reader (PBR) system allows blog readers to browse the coolest and newest blog posts of their interests, by automatically clustering the most relevant stories. The PBR aims to make a user’s potential
favorite topics always ranked higher than the non favorite ones (Xin Li et al 2009).

Since it is important to distinguish the user needs, the search engines have to be engineered by incorporating user personalization. Hence, there is a need for a personalized information retrieval system that promises to resolve the retrieval problem by modeling the user profile based on the user’s general interests.

2.2.9 Blog Ranking and Summarization

Compared with conventional webs, links in the blogosphere are more abundant, and conversations between bloggers are more frequent. Yuhang et al (2008) proposed a method of ranking bloggers based on the link analysis, which can exemplify the characteristics of blogs, and reduce the influence of link spamming. It has brought convenience to users to read blogs, and it can supply a new methodology for information retrieval in the blogosphere. In order to improve the ranking efficiency, the user’s search information has to be taken into consideration. Zhu et al (2012) demonstrated the importance of mining general web user behavior data, to improve ranking and other web-search experiences, with an emphasis on analyzing individual user sessions for creating aggregate models. They have introduced Click-Rank, an efficient, scalable algorithm for estimating the webpage and website importance, from the general web user-behavior data. They quantitatively evaluated its effectiveness regarding the problem of web-search ranking, showing that it contributes significantly to better performance of retrieval in web-search.

People use weblogs to express thoughts, present ideas and share knowledge. Juffinger et al (2009) introduced an additional dimension to assess the credibility, namely, the quantity structure which uses the content
similarity to rank the blogs. Existing approaches tend to focus on ranking the blog posts according to their recent blog post or popularity only, leaving the problem of retrieving more topic relevant posts to a user's query largely unexplored. Han et al (2009) proposed a novel blog ranking framework, called PTRank that improves the search quality, by taking account of the relevance feedback from users, as well as the various information available from RSS feeds. BinRank is a practical solution for scalable dynamic authority-based ranking. It is based on partitioning and approximation, using a number of materialized sub-graphs (Hwang et al 2010). The effective XML keyword search includes the identification of user search intention, and result ranking in the presence of keyword ambiguities. The statistics are utilized to infer user search intention and rank the query results (Bao et al 2010).

Web spam is a behavior that attempts to deceive search engine ranking algorithms. TrustRank is a recent algorithm that can combat web spam. However, TrustRank is vulnerable in the sense, that the seed set used by TrustRank may not be sufficiently representative to cover well the different topics on the Web. Also, for a given seed set, TrustRank has a bias towards larger communities. Wu et al (2006) proposed the use of topical information to partition the seed set and calculate the trust scores for each topic separately to address the above issues. A combination of these trust scores for a page is used to determine its ranking. Most of the research studies focus mainly to improve ranking based only on the relevant content, and not on the trust versus spam content of blogs. Thus, it requires a method for ranking the retrieved blogs that has to consider both the content richness of the blogs and their trust.

Summarization can be viewed as compressing a given set of transactions into a smaller set of patterns while retaining the maximum possible information. A trivial summary for a set of transactions would be
itself and the information loss here is zero but there is no compaction. Another trivial summary would be the empty set, which represents all the transactions. In this case, the compaction gain is maximal but the summary has no informative content. A good summary is one which is small but still retains enough information about the data as a whole and also for each transaction.

In order to improve the content richness in a summary, some kind of categorization can be applied, for a similar type of content summarization. Such techniques have been proposed, that discuss a direct application of clustering to obtain a summary for a given set of transactions with categorical attributes. The algorithm involves clustering of the data, using any standard clustering algorithm, and then replacing each cluster with a representation, using feature-wise intersection of all transactions in that cluster. Each cluster produces an individual summary, which is essentially the set of feature-value pairs, which are present in all transactions in that cluster. The number of clusters determines the compaction gain for the summary (Chandola and Kumar 2007).

Various summarization techniques have been evolved for providing a better summary. One such technique is the Fuzzy swarm diversity hybrid model for text summarization. The swarm-based text summarization method (Binwahlan et al 2010) generates a summary of the original document by picking up the top sentences which have the highest scores. The top number of sentences is equal to the predefined summary length and the sentences are scored, using the features of the content. A weight is assigned to each feature score. Assigning weights to text features is an attempt to find a way to differentiate between the more and less important features (Binwahlan et al 2010). However, there is a need for efficiently summarizing the blogs based on the available diverse contents into converged and consolidated blog information, to enable the user to get quick knowledge about the retrieved information.
2.2.10 Limitations of Existing Methods

Blogs are unstructured, and a free form of publishing articles by any social actor. Bloggers are allowed to use their own style of writing blog posts, which lead to an irregular corpus that makes the search process complex. Tags play a major role in blog retrieval and personalizing the blog content for the user. There are various works on blog ranking using blogger characteristics and the link analysis. The views of various researchers are used in opinion mining, sentiment analysis and personality identification. Researchers have not yet explored the contribution of blogs in e-learning with various issues. However, these approaches fall short of addressing the following issues in particular.

- Blogs are scattered in various sources and it is very difficult for the user to collect them.
- Blogs are published by multicultural bloggers; so, there is a collection of redundant, non-English and incomplete blogs.
- The credibility of the blog content was not checked using reliable trust parameters.
- The blog search engines retrieve a huge amount of blogs which are of interest to users but some are irrelevant to the user query. Keyword ambiguity is present in blog search engines. For example, the query ‘python’ may refer to python as a snake, as well as the python programming language.
- Blogs are not ordered during retrieval, using blog search engines. The works carried out for ranking the blogs, narrowed down towards the aspect of either blog content or blogger reputation.
2.3 PROBLEM STATEMENT

The problem taken up for study in this research, is the need for a new personalized, trusted blog retrieval systems that summarizes the blog content and ranks the blog posts, using Content Based Importance (CBI). This system uses reputation based trust, to generate trusted blogs and personalize them to the user, in accordance with the user and general profiles. The retrieved blogs are ranked based on the score obtained using CBI. The aim of the study is to first to collect all the technical blogs from the web, and check their credibility. Second, the blogs are personalized for the user query, and finally they are summarized and ranked in the order of their relevancy.

2.4 MOTIVATION OF THIS WORK

Many approaches have been proposed in the literature for the personalization, ranking and summarization of web documents. The literature survey unveils some limitation of the earlier proposed approaches. Though the blogs are structured, they are different from a typical web or text document. Thus, it is very difficult to apply the same web mining techniques for blogs, and for the blog readers to find the relevant blogs from the blogosphere.

Many blogging systems allow bloggers to add new tags to a post, in addition to placing the post into categories. The blogosphere is defined as the collection of all blogs as a community or social network. Because of the large number of existing blog documents (posts), the blogosphere content may be random and chaotic (Tsai, 2011). The blog template varies form one blog site to another. Bloggers can have their own blog template, using blogging software. This becomes a complex task in collecting the blogs from their blog
site. Hence, our research work proposed to address the challenges faced in blog searching with the help of various techniques such as ranking, personalization and summarization.

2.5 THESIS OBJECTIVES

- To create a domain specific (education) ontology and to collect blogs for each subject associated with the ontology, using the parallel crawler from various blog sources.

- To convert the collected blog contents into a common format, using the Blog Markup Language (BlogML), and to design a XML schema to validate the BlogML.

- To design a blog filtering system to remove non-English, redundant, incomplete and irrelevant blogs, and to formulate a trust model to compute the trust value for the blogs and to refine the collected blogs based on trust.

- To devise a clustering technique for clustering the blogs based on the subject, and to organize them in the blog repository for effective retrieval.

- To develop a personalized blog retrieval system for personalizing the blog content to the blog user, based on user profile and user search log.

- To implement a blog ranking system, based on the importance of content for producing the most relevant blog results.

- To create a blog summarizer by grouping the retrieved blog contents based on the ontology.
• To suggest a recommendation system to aid e-learning by collecting the learner’s feedback (The role of blogs for subjects with and without evolving technology) for analysis.

To achieve the above objectives, the architecture of the Trust-based Personalized Blog Ranking and Summarization (TPBRS) system has been proposed in Chapter 3.