ABSTRACT

Web mining makes use of data mining techniques, to automatically discover and extract the pattern of meaningful information from web documents. Most of the current web documents are the contents of social networking sites, that are highly dominating in the web 3.0. Among the ever-growing trends of social networking sites, blogs have a prominent role in information sharing among the blog users. Blogs are a set of chronological news entries and can be used to publish personal opinions, diary-like articles or news stories relating to a particular interest or product. There are plenty of blogs available for various application domains, such as entertainment, business, politics, social work, natural disaster and education. Many bloggers post blogs about various technologies, which are of current interest to the blog readers. Novices who are interested in learning or gaining knowledge in the specific technology may go through these blogs and infer new ideas or improve their knowledge. Thus, it is evident that social networking sites like blogs play a major role in the web community, and may be viewed as a platform to aid e-learning. In this thesis, the semantically supportive e-learning system has been proposed to enhance the e-learning process.

Blog mining is defined as the process of extracting a useful pattern of blog information from the blogosphere. Blog mining has evolved as an emerging area, and is still in the process of being explored in various contexts and applications of blogs. Though various traditional mining techniques have been applied for mining blog contents, many of these techniques have not
been tuned to provide all user-specific and domain-specific knowledge to the users. Retrieving the required blogs from the blogosphere is a major challenge for the blog users. The blog user searches the blogosphere for a specific topic, using blog search engines. Blog search engines retrieve a huge number of blogs for each topic. Blog users cannot read all the retrieved blogs, since they are voluminous. The retrieved blogs may contain a number of irrelevant, non-English, incomplete and redundant blogs. They may also be displayed to the users in a random order. Therefore, it is necessary to provide appropriate blogs which are relevant to the learner’s expectations.

The problem taken up for study in this research includes a new personalized, trusted blog retrieval system, which summarizes the blog content, and ranks the blog posts, using Content Based Importance (CBI).

In this research work, a Trust based Personalized Blog Ranking and Summarization system (TPBRS) is proposed to aid the e-learning process. An ontology based parallel crawler has been devised to collect the blogs for the subjects defined in the ontology from different blog sources. All the collected blog contents are converted into a Blog Markup Language (BlogML) in order to present the blog information in a generic format. This BlogML includes various tags used for representing the blog content. The BlogML representation of the collected blog information follows the XML data representation. An XML schema is designed to validate the BlogML. In most of the traditional blog search engines, the blogs are retrieved simply based on the blogs existing in the blogosphere. Most of the retrieved contents have different representations; either it may be due to their content or language
representation. In this work, the collected blogs are invoked to refine the meaningful blog content, i.e., maintaining the homogeneous language (English blogs) blogs. For that, a blog filtering system is designed to remove the non-English, redundant, incomplete and irrelevant blogs.

The retrieved blog contents have different types of blog information on a specific topic. The blog readers are not aware of these blogs, and hence, identifying the trustworthiness of blog content becomes a tedious task. Thus, the proposed system uses a trust model concept to compute the trust value of the blogs and to refine the collected blogs based on trust. The blog network is analyzed to find the local trust and global reputation score. To make the blog retrieval task as an easy one, the blogs are clustered using the Hierarchical Agglomerative Blog Clustering Technique. The clustering technique helps to cluster the blogs based on the subject and to organize them in the blog repository. The personalized blog retrieval system is designed and implemented, based on the user profile and user search log. Personalization considerably reduces the retrieval time, as it is not necessary to retrieve all the available blogs related to a specific query. Instead, it is enough only to retrieve the user needed blogs, which can be achieved with the user personalization. A blog ranking system produces better results based on content based importance. Based on the precision and recall values, the retrieved blogs are evaluated. The precision and recall values of the TPBRS system imply that the system performs well in ranking the blogs. Blogs are highly descriptive, and in many cases, the number of blogs appearing for a
common topic is large. It is important to analyze the blog content, and to provide an effective summary by consolidating the blog contents.

The TPBRS system mainly focuses to support e-learning process that helps the blog readers for getting their interested blogs. An e-learning system with various blog mining tasks provides semantically supportive blog reading system. The proposed TPBRS system has high utilization of technical as well as educational blogs for enhanced e-learning.