Personalised Blog Recommendation System (PBRS)

Amit Panjani
Vidyalankar Institute of Technology,
Vidyalankar College Marg, Wadala (E),
Mumbai, India

Bhavik Jain
Vidyalankar Institute of Technology,
Vidyalankar College Marg, Wadala (E),
Mumbai, India

Rahul Bhardwaj
Vidyalankar Institute of Technology,
Vidyalankar College Marg, Wadala (E),
Mumbai, India

Deepali Vora
Vidyalankar Institute of Technology,
Vidyalankar College Marg, Wadala (E),
Mumbai, India

ABSTRACT
Blog provides a simple way for people to share personal experiences and ideas, and has already become an important tool for people to communicate with each other. Due to the vast amount of information on a particular blog, it is often time consuming for reviewing and finding the blog-article to suit the reader's mind. This paper proposes a personalised blog recommendation system that utilises text mining and various recommendation techniques. It aims at providing personalized blog article recommendations with high efficiency and effectiveness. This paper surveys the landscape of actual and possible hybrid recommender systems, and introduces a novel hybrid recommendation method that combines text mining, collaborative filtering, content-based and demographic-based recommendations to recommend blogs.

General Terms
Recommender Systems; blog recommendation; text mining; document classification

Keywords
Blog; recommender system; text mining; hybrid recommendation

1. INTRODUCTION
The amount of data in our world has been exploding, which makes it difficult for the decision-maker to identify useful information. Recent years have witnessed a tremendous growth of the blogosphere. The size of the collection of blogs on the World Wide Web has been lately exhibiting an exponential increase. As blogs become more and more popular, they attract more and more people to get involved and contribute fresh and useful information to blogosphere. Blogs are now one of the main means to spread ideas and information throughout the Web. They discuss different trends, ideas and events.

A Web Log (Blog) is a website maintained by an individual who uses it as a self-publishing media by regularly publishing posts commenting on or describing some event or topic. Blog provides a simple way for people to share their experience. Such a fact attracts many people to participate in blogosphere, and makes it possible for any kind of information to get propagated over the World Wide Web.

In order to get the most value out of their data, the challenge is to ensure the right information is getting to the right employees, because the large amount of information makes it difficult for users to be aware of it or even look through it. The personalized recommender systems are important applications that can address this problem and suggest items that suit the user's needs [1]. Typically, a recommender system compares the user's profile to some reference characteristics, and seeks to predict the rating that a user would give to an item they had not yet considered [2]. The key element of a recommender system is the user model that contains knowledge about the individual preferences, which determines his or her behaviour in a complex environment.

Recommender systems were originally defined as ones in which "people provide recommendations as inputs, which the system then aggregates and directs to appropriate recipients" [3]. The term now has a broader connotation, describing any system that produces individualized recommendations as output or has the effect of guiding the user in a personalized way to interesting or useful objects in a large space of possible options. Recommender systems are software tools and techniques providing suggestions for items to be of use to a user. The suggestions relate to various decision-making processes, such as what items to buy, what music to listen to, or what blog or news to read [4].

This paper proposes a system that provides personalised blog suggestions and makes use of a hybrid approach by combining collaborative, content-based, demographic recommendation techniques with text mining. The approach extends text mining through the Naïve Bayes Text Classification algorithm to classify the blogs to a pre-defined class label or category and thus, providing efficient and effective recommendations by overcoming some of the common problems in recommender systems such as cold start and the sparsity problem.

2. RELATED WORKS
The research closely relates to the works of Recommender systems and blog content mining.

Recommender systems are the software that suggests what we should watch or read or listen to next. Recommender systems are now an integral part of some e-commerce sites such as Amazon.com and CDNow [5]. It is the criteria of "individualized" and "interesting and useful" that separate the recommender system from information retrieval systems or search engines. One common thread in recommender systems research is the need to combine recommendation techniques to achieve peak performance. Recommender systems differ in the way they analyse these data sources to develop notions of affinity between users and items which can be used to identify well-matched pairs. Collaborative Filtering systems analyse historical interactions alone, while Content-based Filtering

systems are based on profile attributes; and Hybrid techniques attempt to combine both of these designs.

Content-based (CB) [7] is a method extended from Information Retrieval technique. CB analyses the attributes and characteristics according to user's historical preference items and, then, matching the suitable ones for user's request. The similarity measure of items or attributes is used for finding the matched items. The measuring technique for CB is not suitable for recommending items such as music, art, movie, audio, photograph, video, etc. However, these items are frequently read in blog sites, hence, these types of article may not easily be analysed for relevant attribute information [8].

Collaborative Filtering (CF) [6][9], on the other hand, is widely applied and used for article, movie, product, etc. CF recommends items based on the similar preference of a group, known as neighbour. The CF technique, therefore, requires methodology for clustering or finding the neighbourhood. Although CF method can handle the wide variety of information, it may still suffer two common problems, i.e., sparsity and cold-start [10]. Sparsity means even if there are many users, it could happen that the user accessing-matrix or rating-matrix is still sparse. This phenomenon will generate the low coefficients of similarity and, therefore, make the recommendation inaccurate. Cold-start is the problem with new users and new items, where not enough information is available for generating recommendations. The ramp-up problem has the side-effect of excluding casual users from receiving the full benefits of collaborative and content-based recommendation. The learning-based technologies work best for dedicated users who are willing to invest some time making their preferences known to the system.

Demographic recommender systems aim to categorize the user based on personal attributes and make recommendations based on demographic classes. The representation of demographic information in a user model can vary greatly. Demographic techniques form "people-to-people" correlations like collaborative ones, but use different data. The benefit of a demographic approach is that it may not require a history of user ratings of the type needed by collaborative and content-based techniques [10].

Hybrid recommender systems combine two or more recommendation techniques to gain better performance with fewer of the drawbacks of any individual one. Most commonly, collaborative filtering is combined with some other technique in an attempt to avoid the ramp-up problem. Some of the combination methods that have been employed are:

- Weighted: The score of different recommendation components are combined numerically.
- Switching: The system switches between recommendation techniques depending on the current situation.
- Mixed: Recommendations from several different recommenders are presented at the same time.
- Feature Combination: Features from different recommendation data sources are thrown together into a single recommendation algorithm.
- Feature Augmentation: Output from one technique is used as an input feature to another.
- Cascade: Recommenders are given strict priority, with the lower priority ones breaking ties in the scoring of the higher ones.
- Meta-level: The model learned by one recommender is used as input to another [10].

Blog content mining finds and extracts information from blogosphere by leveraging natural language processing and data mining technique. Statistics of term frequency and similarity based analysis approaches are usually followed in blog content mining. Information such as post title, tag could be leveraged [11].

Text mining attempts to glean meaningful information from natural language text. It may be loosely characterized as the process of analysing text to extract information that is useful for particular purposes. The field of text mining usually deals with texts whose function is the communication of factual information or opinions, and the motivation for trying to extract information from such text automatically is compelling—even if success is only partial [12]. The phrase "text mining" is generally used to denote any system that analyses large quantities of natural language text and detects lexical or linguistic usage patterns in an attempt to extract probably useful (although only probably correct) information [13].

3. SYSTEM DESIGN

This paper proposes a Personalised Blog Recommendation System (PBRS), which fuses text mining, demographic characteristics and user behaviour information for improving the quality of recommendations. The proposed system consists of the following important elements:

1. Classification of blogs into appropriate categories by using text mining, which classifies any new blog in the system into the proper category by analysing the natural language text.
2. Using the interests of the user, provided during signing up, and user's history of read blogs to determine user's preferences for blog documents.
3. Creating groups of users with similar preferences and demographic profile, and recommending blog content according to those preferences and characteristics.
4. Adjusting the accuracy of recommendations according to user's feedback, provided through semantic rating.

The user interacts with the system through the blogging website which provides, the bloggers, an interface to write blogs on various categories, and the readers, personalised recommendations based on their preferences. The users sign up on the website by providing their personal details and interests, which are stored in a SQL database, and are further used to determine user preferences and creating groups of users. The blogs written by the bloggers are stored in NoSQL document store database. These blog documents are then classified into appropriate categories by using text mining algorithm. After analysing the user preferences and grouping similar users, the blog documents are ranked and personalised recommendations are provided to each user groups. A high level workflow diagram is shown in Fig. 1.

3.1 Document Classification using text mining

Providing suitable recommendations for users requires categorising hundreds of new blog documents into categories that can be matched with user preferences. Blog sites generally contain multiple blog documents on various categories, thus making one blog document the smallest possible unit. The proposed system entails the following steps of text mining to classify blog documents (Fig. 2):

- Text Parsing- It involves extraction of words, parts of speech tagging, word filtering (removing preposition, numbers, and punctuations), synonyms, tokenization, and stemming.
- Text Filtering- Removing irrelevant terms, building stop word dictionary and removing stop words.
- Naïve Bayes Text Classification Algorithm- A probabilistic
- model based on Bayes' theorem, to classify documents in simple and very effective manner.

Text mining is the analysis of data contained in natural language text to deriving high-quality information. Thus, the above steps are applied to scan the blog documents written in natural language for their predictive classification into appropriate category.

3.2 User preference analysis and recommendation generation

The proposed system uses a weighted hybrid approach in which the score of different recommendation components are combined numerically. The various steps which are considered in user preference analysis and generating personalised recommendations (Fig. 3) are as follows:

- The user's interests and user's history of read blogs to determine the user's preferences.
- Creating group of users with similar interests and demographic profile.
- Collaborative filtering algorithm examines large groups of individuals, identify sets of people with similar tastes, and create ranked lists of suggestions.

The final score of a recommended blog is computed from the results of all of the recommendation techniques and a ranked list of suggestions is created to provide personalised recommendations to the user.

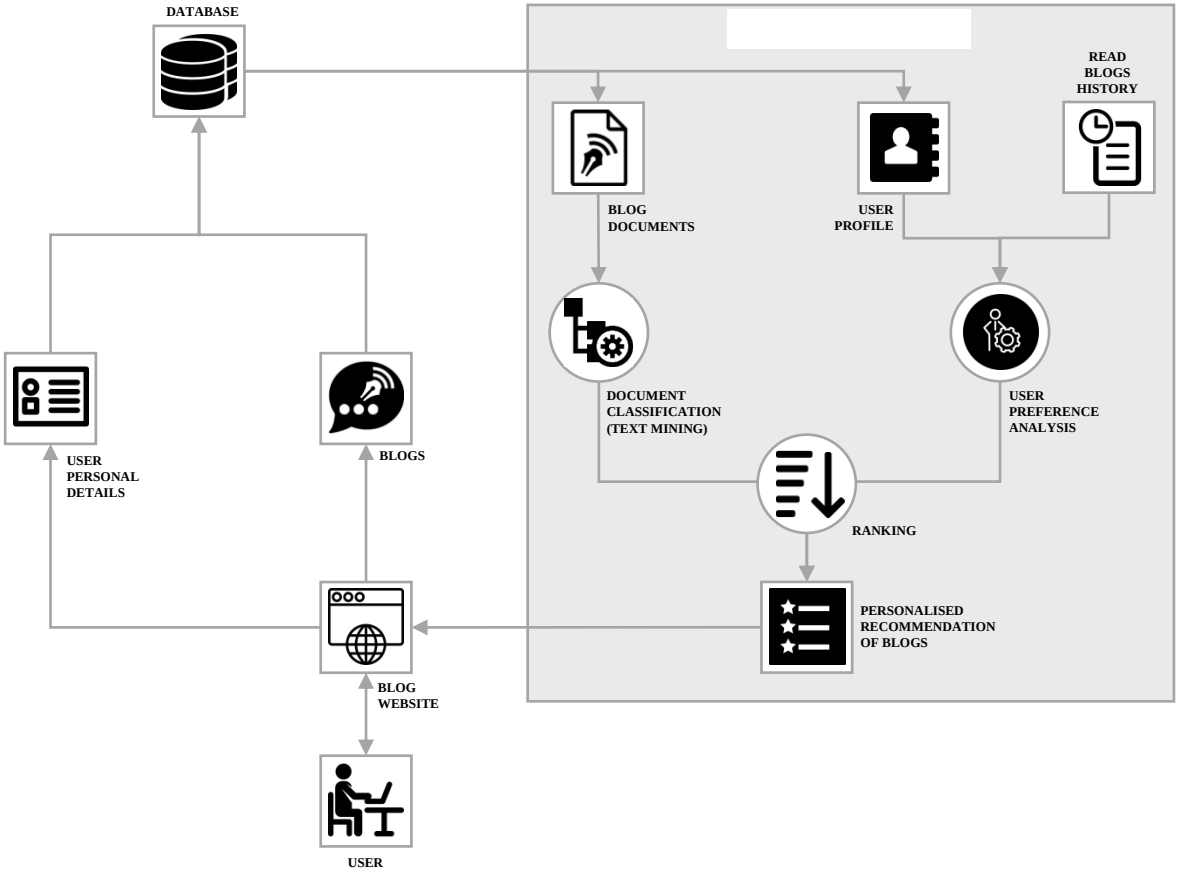

Fig. 1: PBRS high-level workflow

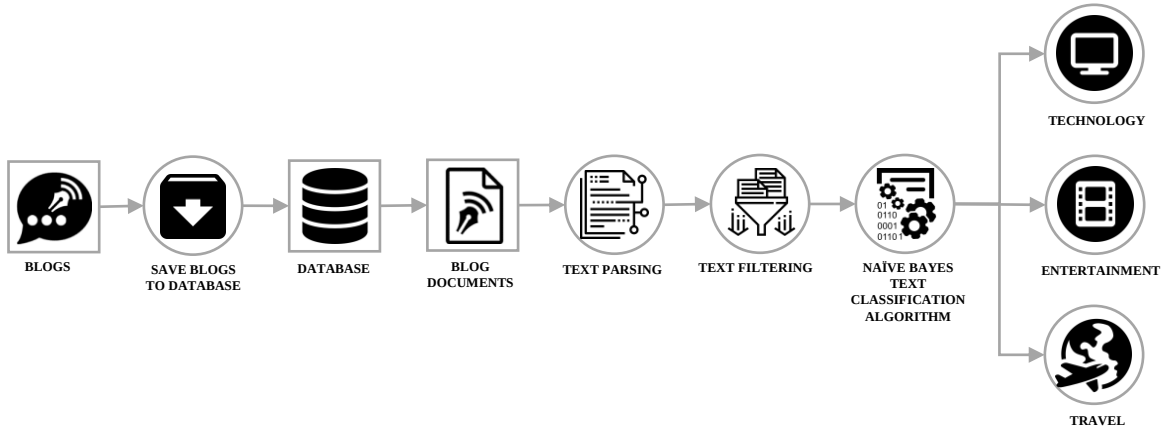

Fig. 2: Document Classification using Text Mining

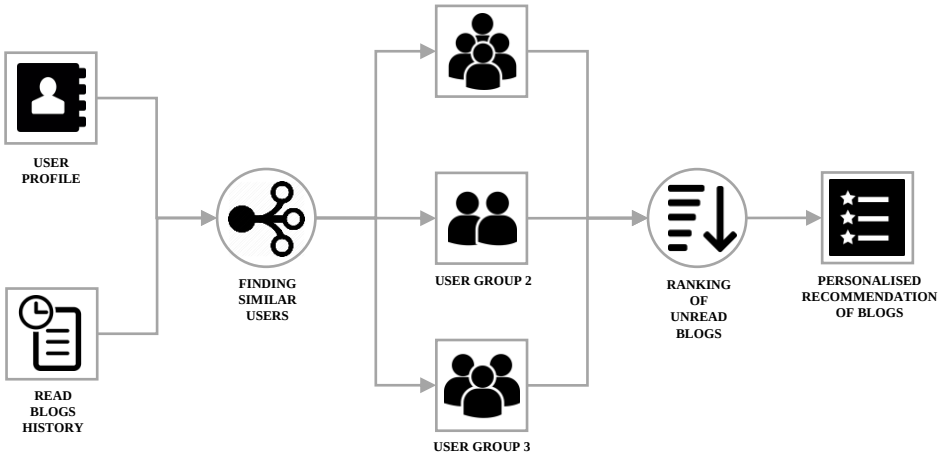

Fig. 3: User preference analysis and recommendation generation

4. CONCLUSION

This paper describes the proposed system, PBRS, which uses a hybrid approach to provide personalised recommendations of blogs by fusing various recommendation techniques and text mining. The system uses text mining to classify blogs to appropriate category and collaborative filtering to group users

with similar preferences and demographic profiles. It creates ranked lists of suggestions by computing final scores from different recommendation techniques and discusses difficulties tied to designing recommender systems. The proposed system tries to reduce the sparsity and cold-start problem, and aims at improving the quality and accuracy of personalised blog recommendations.

5. REFERENCES

[1] Adomavicius G. and Tuzhilin A, 'Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions [J],' Knowledge and Data Engineering, IEEE Transactions on, 2005, 17(6): 734-749.

[2] Shinde S K. and Kulkarni U, 'Hybrid personalized recommender system using centering-bunching based clustering algorithm [J],' Expert Systems with Applications, 2012, 39(1): 1381-1387.

[3] Resnick, P. and Varian, H. R, 'Recommender Systems', Communications of the ACM, 1997, 40 (3): 56-58.

[4] Francesco Ricci, Lior Rokach and Bracha Shapira, 'Introduction to Recommender Systems Handbook,' Recommender Systems Handbook, Springer, 2011: 1-35.

[5] Schafer, J. B., Konstan, J. and Riedl, J., 'Recommender Systems in E-Commerce,' In: EC '99: Proceedings of the First ACM Conference on Electronic Commerce, Denver, CO, 1999: 158-166.

[6] F. Chesani, 'Recommendation Systems,' *Corso di laurea in Ingegneria Informatica*, 2002: 1-32.

[7] Y. S. Kim, B. J. Yum, J. S. Su, and S. M. Kim, 'Development of A Recommender System Based on Navigational and Behavioral Patterns of Customers in E-commerce sites,' *Expert Systems with Applications*, vol. 28, 2005: 381-393.

[8] B. Sarwar, Karypis, G. Konstan, and J. Riedl, 'Analysis of Recommendation Algorithms for E-commerce,' *Proceedings of ACM E-commerce 2000 Conference*, 2000: 158-167.

[9] S. Upendra and M. Pattie, 'Social Information Filtering: Algorithms for Automating 'Work of Mouth',' *Proceedings of the SIGCHI conference on Human factors in computing systems*, 1995: 210-217.

[10] Burke, Robin. 'Hybrid recommender systems: Survey and experiments,' *User modeling and user-adapted interaction* 12.4, 2002: 331-370.

[11] Kening Gao, Yin Zhang, Bin Zhang, Pengwei Guo and Qingpeng Niu, 'Blog Recommendation based on Blog Set similartiy and Mergence,' 2010 Second International Conference on Communication Systems, Networks and Applications, Hong Kong, 2010: 256-259.

[12] Witten, I.H, 'Text mining,' Practical handbook of internet computing, Chapman & Hall/CRC Press, Boca Raton, Florida, 2005: 14-1-14-22.

[13] F. Sebastiani, 'Machine learning in automated text categorization,' ACM Computing Surveys, Vol. 34, No. 1, 2002: 1–47.