RECOMMENDING TAGS FOR NEW RESOURCES IN SOCIAL BOOKMARKING SYSTEM

Shweta Yagnik[1], Priyank Thakkar[2], K Kotecha[3]

[1]Assistant Professor, CE Department, L. J. Institute of Engineering & Technology, Ahmedabad - 382 210, Gujarat, India
[2]Assistant Professor, CSE Department, Institute of Technology, Nirma University, Ahmedabad - 382 481, Gujarat, India.
[3]Director, Institute of Technology, Nirma University, Ahmedabad - 382 481, Gujarat, India.

ABSTRACT

Social bookmarking system is a web-based resource sharing system that allows users to upload, share and organize their resources i.e. bookmarks and publications. The system has shifted the paradigm of bookmarking from an individual activity limited to desktop to a collective activity on the web. It also facilitates user to annotate his resource with free form tags that leads to large communities of users to collaboratively create accessible repositories of web resources. Tagging process has its own challenges like ambiguity, redundancy or misspelled tags and sometimes user tends to avoid it as he has to describe tag at his own. The resultant tag space is noisy or very sparse and dilutes the purpose of tagging. The effective solution is Tag Recommendation System that automatically suggests appropriate set of tags to user while annotating resource. In this paper, we propose a framework that does not depend on tagging history of the resource or user and thereby capable of suggesting tags to the resources which are being submitted to the system first time. We model tag recommendation task as multi-label text classification problem and use Naive Bayes classifier as the base learner of the multilabel classifier. We experiment with Boolean, bag-of-words and term frequency-inverse document frequency (TFIDF) representation of the resources and fit appropriate distribution to the data based on the representation used. Impact of feature selection on the effectiveness of the tag recommendation is also studied. Effectiveness of the proposed framework is evaluated through precision, recall and f-measure metrics.

KEYWORDS

Tag recommender, multilabel classification, social bookmaking

1. INTRODUCTION

Social bookmarking system allows user to collect, organize, share and label the resources, here bookmarks or publications, with arbitrary words i.e. Tags. Figure 1 is a snapshot of BibSonomy [1][2], a social bookmark and publication sharing system that supports collaborative tagging where user can post his resources and categorize them from his personal point of view by providing tags. The simplicity of collaborative tagging for user-centric content publishing and

management comes at the cost of challenges [3] like, the freedom of selecting tags compels user to write descriptive tags on his own to define his viewpoint which is burdensome and time consuming task. Hence, user may avoid or assign very small number of tags to resource resulting in very sparse tag space. Further, different users may choose tags based on their knowledge background and preferences i.e. they may describe the same resource based on different granularity level resulting into noisy tag space and creates difficulty to find relevant material based on such tags. Also, synonymous tags increase data redundancy and polysemous tags i.e. a tag that has many contextual meanings, lead to inappropriate connections between resources.

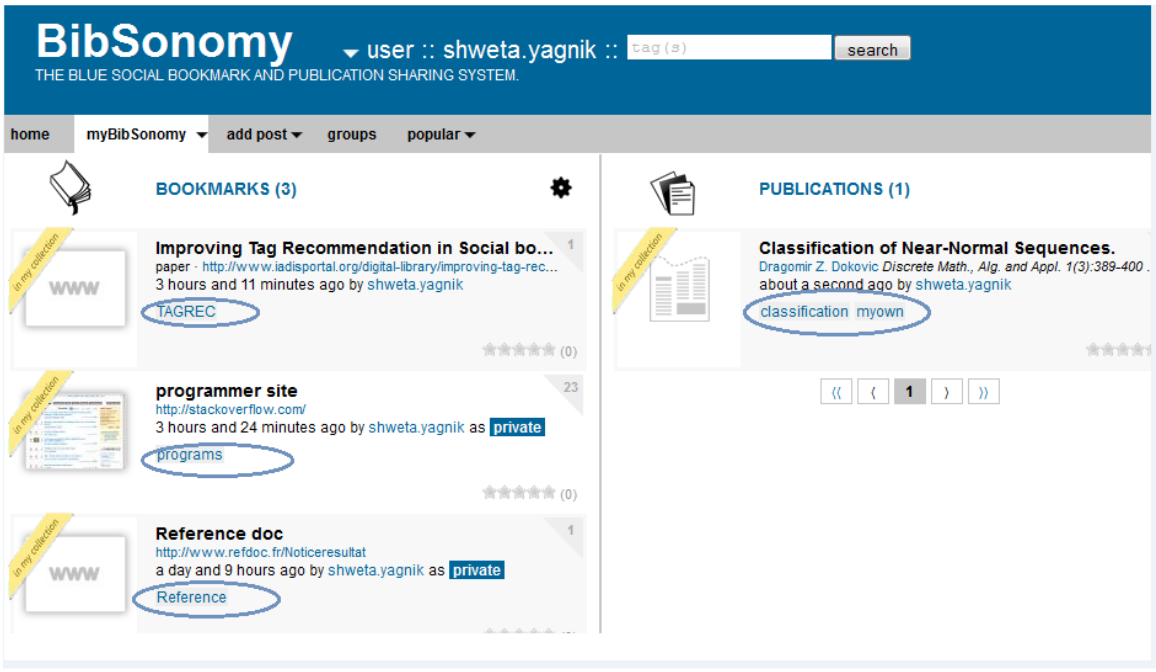

Figure 1. BibSonomy: Social Bookmark and Publication Sharing System

These hurdles in tagging process create very sparse or noisy tag-space that ultimately dilutes the purpose of tagging for information organization. However, it inspires to develop methods that help users while tagging by automatically recommending an appropriate set of tags. The objective of tag recommendation mechanisms is to ease the process of finding useful tags for a resource by reducing his efforts from a manual entry to a mouse click and hence, increasing the chances of getting a resource annotated. It helps in consolidating the vocabulary across users which exposes different aspects of a resource and enriched set of tags help user in reminding what a resource is about. Figure 2 shows tag recommendation in BibSonomy [1][2]. It can be seen that when user posts a bookmark or publication, the system gives suggestion for tags which are appropriate to the resource being submitted.

2. RELATED WORK

Ioannis Katakis et al. [4] also modelled tag recommendation problem as multi-label text classification task. This is different from our work in a way that they used tagging history and represented resources as Boolean feature vectors. Jaschke et al. [5] compared two tag recommendation approaches. First is classic collaborative filtering (CF) and other is graph-based

tag recommendation system based on FolkRank algorithm. To reduce sparsity of folksonomy graph, which is main limitation of graph-based methods, p-core processing i.e. graph pruning was used. The evaluation tests were performed on resultant dense part of dataset, which may not be representative of real life data. K-Nearest Neighbour algorithm was adapted for tag recommendation by Jonathan Gemmell et al. [6]. They also used p-core processing to deal with noisy tag space and effectively worked on dense part of the dataset. Marta Tatu et al.

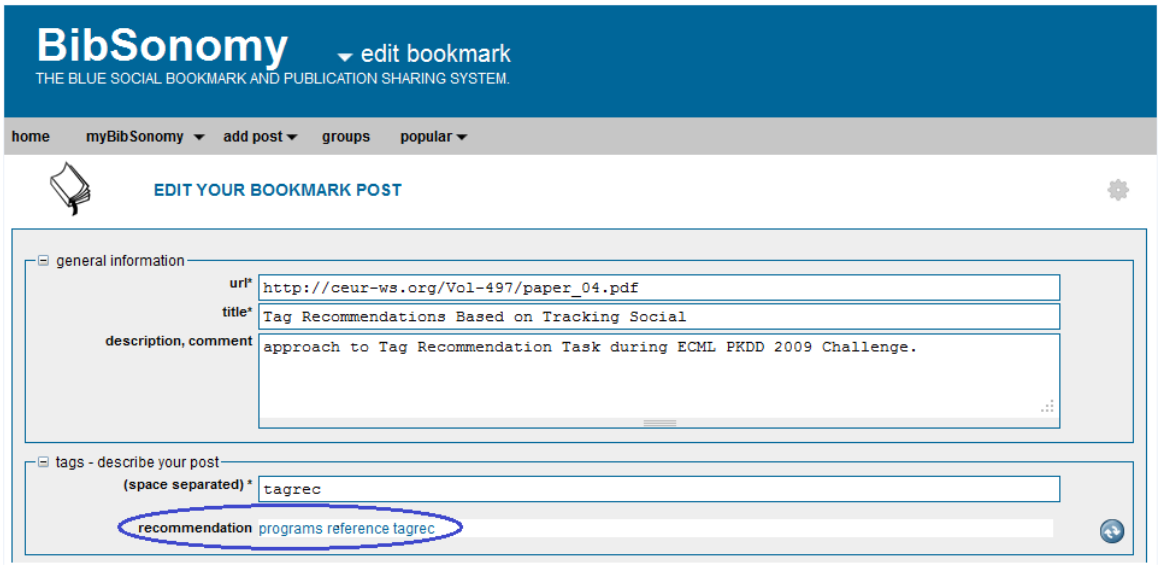

Figure 2. Tag Recommendation in BibSonomy

[7] derived document and user models from the textual content of the post. In this model, tag suggestions were not only from the existing tag space but also from the metadata provided by user to the resource like title, description and the content of the document. User model was derived from user's tagging behaviour. Marek Lipczak et al. [8] showed in his studies that CF based on cosine similarity between users, calculated from resource content is not a good idea for recommending tags. He proved that there is no correlation in cosine similarity between two users calculated based on tags and content of item. He suggested potential sources of tags for recommendation focusing user's personomy. Domenico Gendarmi et al. [9] designed Prompter-A recommender system which suggested tags according to three facets of a social bookmarking system: the personal tagging history, the social tagging behaviour and the textual content of the resource. Evaluation against a snapshot of the BibSonomy dataset [10] revealed that, the combination of these three different tag sources improved the precision of generated tag suggestions in the case where users already had a plentiful tagging history and bookmarks that point to popular resources within the community. Sally Hamouda et al. [11], suggested personalized tag recommendation for social bookmarking system based on finding similar users and similar bookmarks. A generalized tag recommendation framework was proposed by Zinovia Alepidou et al. [12] that conveyed the semantics of resources according to different user profiles. System was built upon resource's title; user's tagging history and other tags. Sanghun Ju et al. [13] have exploited previously annotated tags on the same resource, resource descriptions and previously annotated tags by the same person. They devised and deployed a filtering scheme for removing inappropriate candidates and a weighting scheme for combining information from multiple sources.

In this paper, we model tag recommendation as multilabel text classification problem and experimentally evaluate the impact of different possible representation of the resource and feature selection.

3. PROPOSED APPROACH

We model the social tag recommendation task as multilabel text classification problem. Multi-label classification is a supervised learning problem where an instance may be associated with multiple labels. Tag recommendation task can be modelled as multi-label classification problem as one resource may be annotated with multiple tags based on the relevance with the resource and these different relevant tags facilitate in exposing multiple aspects of a resource. To handle multi-label classification problem there are mainly two approaches [14], first is problem transformation methods that convert the multi-label classification problem into a set of binary classification problems. Binary Relevance (BR), label combination or label power-set method and classifier chains are examples of problem transformation methods. Other approach is algorithm adaptation methods that modify learning algorithm to directly perform multi-label classification.

BR problem transformation method is used in our tag recommendation framework. It is a simple classifier that scales linearly with the number of classes in a multi-label classification dataset [14]. It considers the prediction of each label as an independent binary classification task, thus each binary model is trained to predict the relevance of one of the labels. To accomplish this, the original dataset is transformed into total $|L|$ sets, where L is set of label i.e. set of unique tags in our task. Each dataset D_λ contains all the examples labelled as λ, if in the original dataset they are labelled as λ otherwise as $-\lambda$. It learns a binary classifier $C_\lambda: X \rightarrow \{\lambda, -\lambda\}$or each label.

Naive Bayes is used as a base learner because it is computationally efficient as well as optimal for classification tasks even when the conditional independence between attributes assumption is invalid [15]. Experiments are carried out with Boolean, bag-of-words and TFIDF representation of resources and accordingly multivariate Bernoulli distribution (MVBD), multinomial distribution (MND) or normal distribution (ND) is fitted to the data [16]. We have used Mulan package [17] for our experiments. Performance of multi-label classification is calculated based on standard information retrieval metrics called precision, recall and f-measure [18] as mentioned in Eq. (1), (2) and (3), respectively where m denotes total number of test instances, P_i is a set of predicted labels and Y_i is set of actual labels for instance x_i.

Precision is the number of correct tags retrieved divided by the total number of retrieved tags, thus it gives the percentage of correctly recommended tags among all tags recommended by the tag recommendation algorithm.

$$Precision = \frac{1}{m}\sum_{i=1}^{m}\frac{|P_i \cap Y_i|}{|P_i|} \qquad (1)$$

Recall is the number of correct tags retrieved divided by the total number of correct tags, thus it is the percentage of correctly recommended tags among all tags annotated by the users i.e. actual tags.

$$Recall = \frac{1}{m}\sum_{i=1}^{m}\frac{|P_i \cap Y_i|}{|Y_i|} \qquad (2)$$

It is hard to compare two classifiers using two different evaluation metrics. F-measure is harmonic mean of precision and recall which gives a single metric for comparison. F-measure tends to be closer to smaller of two.

$$F - measure = \frac{1}{m}\sum_{i=1}^{m}\frac{2|P_i \cap Y_i|}{|P_i| + |Y_i|} \qquad (3)$$

4. DATA PREPROCESSING

During the ECML/PKDD Discovery Challenge, Belgium 2008, organizers provided dataset of BibSonomy system [10][1][2]. It contains three training files named tas, bookmark and BibTex. Table 1 reflects the attributes of all three training files.

The original training tas file contains 8,16,197 records, bookmark file contains 1,76,147 and BibTex file contains 92,545 instances. The tas file describes tag assignments made by a user to resource and contains other details like user_id, tag, content id (bookmark.content_id or BibTex.content_id), content type (1 = Bookmark Resource, 2 = BibTeX Resource) and date. For instance, user's tag assignment record is shown in Table 2.

Table 1. Attributes of Three Files

tas	user, tag, content_id, content_type, date
bookmark	content_id, url_hash, url, description, extended description, date
Bibtex	Content_id, journal volume, chapter, edition, month, day, booktitle, howPublished, institution, organization, publisher, address, school, series, bibtexKey, url, type, description, annote, note, pages, bKey, number, crossref, misc, bibtexAbstract, simhash0, simhash1, simhash2, entrytype, title, author, editor, year

Table 2. Tag Assignment to Resource

Example 1	User_id Tag Content_id Content_type Date	27 computer 938977 1 10/10/2005 10:40
Example 2	User_id Tag Content_id Content_type Date	27 quiet 938977 1 10/10/2005 10:40

In our snapshot of BibSonomy dataset, tas file contains total 3,04,118 records, where (user_id, content_id) pair appears multiple times based on number of tag assignments by the user to resource. It reveals total amount of tags assigned by users as each record represents a single tag assigned to the resource. As a part of preprocessing we have converted all tags to lower case and removed punctuation marks and non-English characters from tag string. After this step tas file is left with 3,03,670 records with plain text tag assignments. Thus, each resource is associated with plain text tags that can be accurately processed to generate recommendation. There are total 1,73,568 posts of Bookmark resource and 1,30,102 posts of BibTeX resource. These posts contain 50,000 unique items from each type of resource. For Bookmark total 11,067 unique tags

and for BibTeX 10,878 unique tags are found in the preprocessed dataset. Average tag assignment to Bookmark is 3.4 tags and to BibTeX is 2.6 tags.

The bookmark file contains bookmarked post related information in fields like content id, url hash, url, description, extended description and date. Url hash field uniquely identifies bookmark resource. For instance, one of the web pages bookmarked by user is described by the information shown in Table 3.

Table 3. Bookmarked Web Page in BibSonomy

Content_id	4145011
URL hash	1a4e59c781ba7f9b9dfb63d493738a1a
URL	http://www.epyxmobile.com/
Description	Mobile Internet Telephony :: Skype for the road!
Extended Description	Take Skype with your for the road! Use your mobile phone to call Skype users or receive calls from them, for free! Make phone calls between mobile phones for free, even across country borders!
Date	1/5/1989 10:40

The BibTeX file contains bookmarked publication related information in fields like content id, journal volume, chapter, edition, month, day, booktitle, howPublished, institution, organization, publisher, address, school, series, bibtexKey, url, type, description, annote, note, pages, bKey, number, crossref, misc, bibtexAbstract, simhash0-2, entrytype, title, author, editor and year. Simhash1 uniquely identifies BibTeX entry. Miscellaneous information is collected in the misc field which may include user comments, non-standard BibTeX fields like isbn, bibdate etc. For instance, one of the publications bookmarked by user is described by the information shown in Table 4.

Table 4. Bookmarked Publication in Bibsonomy

Content id	688717
Journal vol	Computer Networks and ISDN Systems
Chapter	30
BibtexKey	brin1998web
URL	http://citeseer.ist.psu.edu/brin98anatomy.html
Pages	107117
Number	17
Misc	keywords = google pagerank searchengine, priority = {3},citeulike-article-id = 922
BibtexAbstract	In this paper, we present Google, a prototype of a large-scale search engine which makes heavy use of the structure present in hypertext. Google is designed to crawl and index the Web efficiently and produce much more satisfying search results than existing systems.
Simhash0	7a736d3fbe3935f4a95181ca5fa0368f
Simhash1	1234ad3633d435ef79d8a7f36dafa0a9
Simhash2	1779c82bd34bbf1ca62956d136a22adf
Entrytype	Article
Title	The anatomy of a large-scale hypertextual Web search engine
Author	Sergey Brin and Lawrence Page
Year	1998

As there exist a huge amount of posted tags, a right number of tags should be selected for reducing the computational cost and avoiding the over fitting problem. We plot histogram for Bookmark and BibTeX dataset to analyse the frequency of occurrence of tag.

Histogram shown in Figure 3 and Figure 4 reflect that low frequency tags i.e. tags which are used single, twice, 5-10 times etc. have dominating count. It reveals the fact that repository has a big number of low-frequency tags which increases sparsity and complicates the process of retrieving good recommendations. High frequency tags should be considered when designing an effective tag recommender. In order to decrease the dimensionality of the problem, we considered high frequency with moderate unique tag count. For bookmark dataset keeping the tag frequency $\geq$ 100 results into 245 unique tags and for BibTeX dataset keeping tag frequency $\geq$ 100 results into 111 unique tags. We have kept 80% of the resultant dataset as training and 20% as testing.

The classifier considers the text representation of the resource for which tags are to be recommended. Experiments are carried out with Boolean, bag-of-words and term frequency-inverse document frequency (TFIDF) representation of the resource. In order to create textual representation for the Bookmark resources, we have used Description and Extended Description fields and for BibTeX resources, we have used the Journal, Booktitle, BitexAbstract and Title fields from the dataset. We have used Weka [19] to convert string attributes into a set of attributes representing presence/absence, count or TFIDF of words [16]. Bookmark and BibTeX resources are represented with 1,439 and 1,173 attributes, respectively. Figure 5 shows the conceptual flow of preprocessing steps we followed for our system.

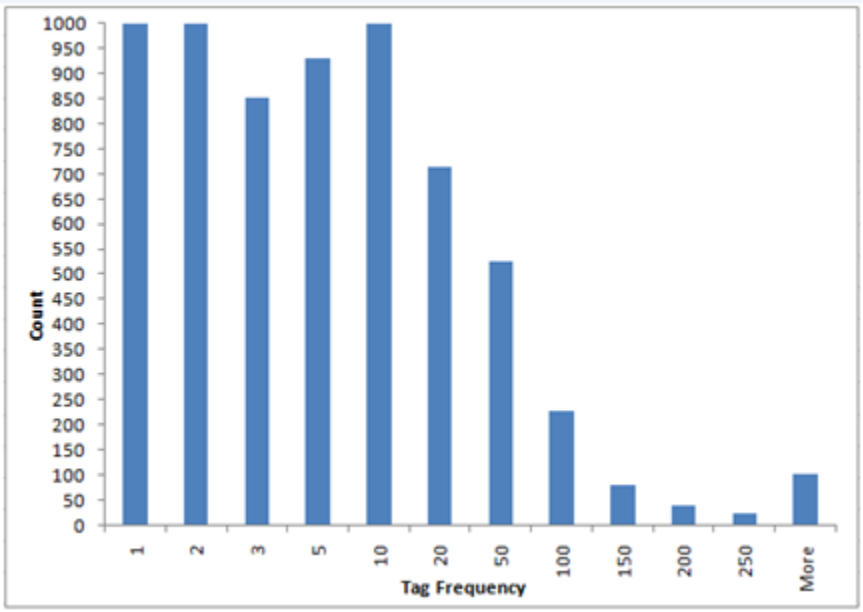

Figure 3. Histogram for Tag Frequency Distribution in Bookmark

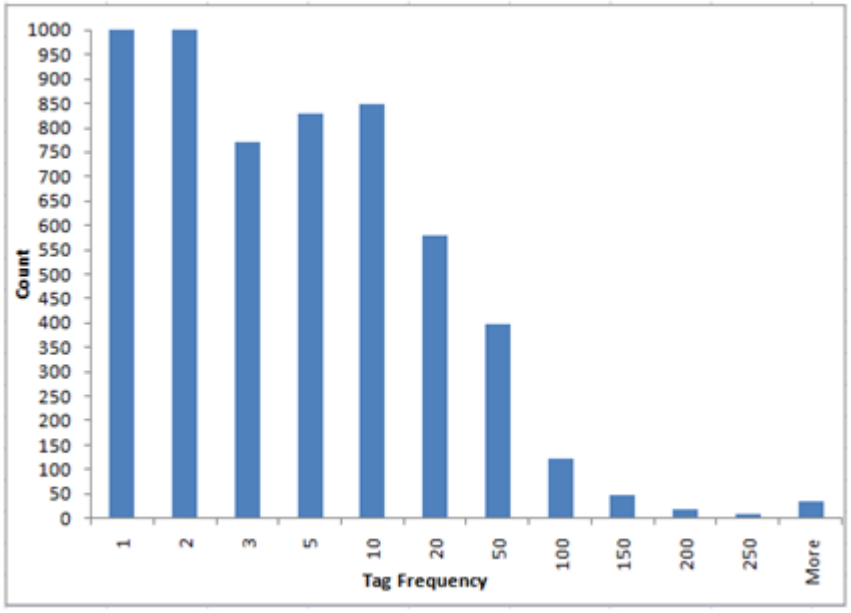

Figure 4. Histogram for Tag Frequency Distribution in BibTeX

5. PROPOSED TAG RECOMMENDATION ALGORITHM

We want to predict tags that a user would assign to a particular resource. The important observation from the given dataset is that particular resource is submitted only once by the user i.e. only once any Bookmark or BibTeX resource is submitted by any user, hence every test is a new unseen resource for which we have to recommend set of tags. In our work, BR classifier from the Mulan package is used. Naive Bayes classifier is used as the base learner of the BR classifier. We propose to use bag-of-words representation rather than Boolean or TFIDF representation of resources which allows us to fit multinomial distribution rather than multivariate Bernoulli or normal distribution to the data. In bag-of-words representation, each attribute is represented as a natural number, indicating the number of occurrences of term in the document. Next we discuss how the multinomial distribution defines the posterior probability [16], $P(c_j|d_i)$ of document d_i belonging to class c_j. Assume that there are m attributes a_1, a_2, …, a_m and n documents d_1, d_2, …, d_m from class T. Number of times attribute a_i occurs in document d_j is denoted as n_{ij}, and the probability with which attribute a_i occurs in all documents from class T as $P(a_i|T)$. It is defined by Eq. (4).

$$P(a_i|T) = \frac{\sum_{j=1}^{n} n_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{n} n_{ij}} \qquad (4)$$

The multinomial distribution defines the probability of document d_j given class T as $P(d_j|T)$ and it is as in Eq. (5).

$$P(d_j|T) = \left(\sum_{i=1}^{m} n_{ij}\right)! \prod_{i=1}^{m} \frac{P(a_i|T)^{n_{ij}}}{n_{ij}!} \qquad (5)$$

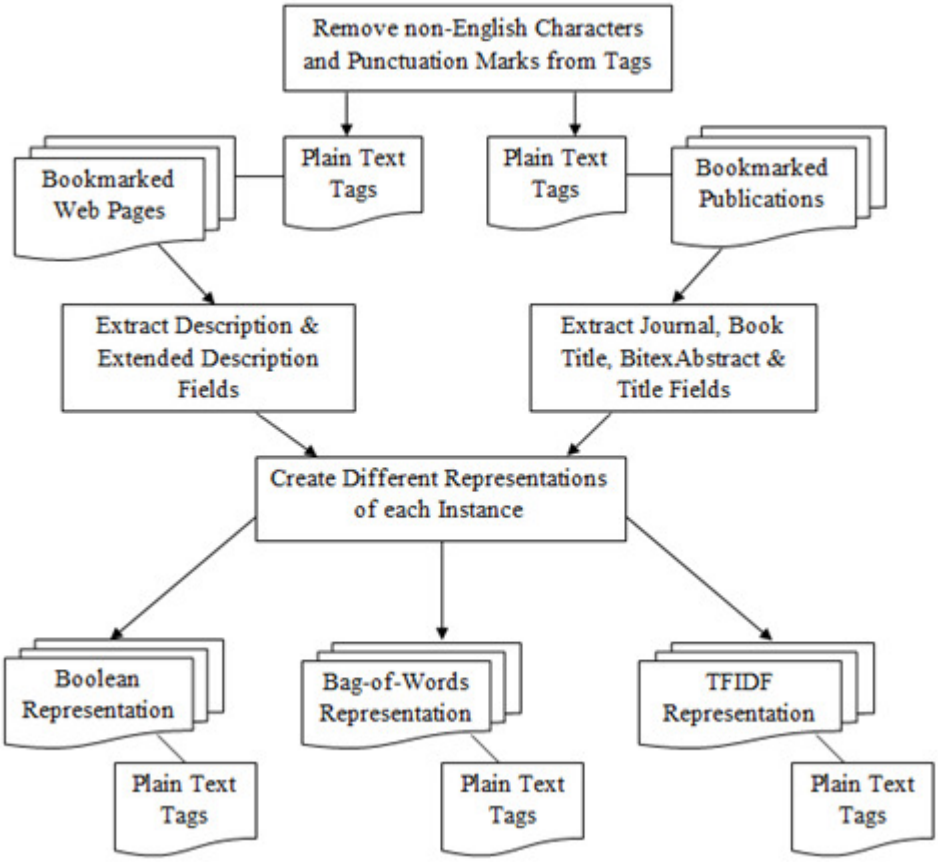

Figure 5. Conceptual Flow of Preprocessing Steps

The ordering of words is ignored in the bag-of-words model. To consider all possible orderings of each word ($n_{ij}!$) and all words in the document $\left(\sum_{j=1}^{m} n_{ij}\right)!$ is added [16].

As stated earlier, we have also experimented with Boolean and TFIDF representation of the resource. In Boolean representation, each attribute is represented by a value 0 or 1 depending on whether or not the corresponding term occurs in the textual representation of the resource. The resources in this representation are binary vectors following the multivariate Bernoulli distribution. In TFIDF representation, each attribute is represented by a value indicating its term frequency-inverse document frequency score. This leads to the continuous vectors following the normal distribution. In case of both the dataset, resources are represented with large number of attributes. It is possible that some of the attributes may not be relevant to the classification task and others may be redundant. Feature selection can help to incorporate only those features which are important for classification task [20]. This may improve performance of classification. To incorporate feature selection, we have used BinaryRelevanceAttributeEvaluator from Mulan package [17]. It evaluates individual attribute based on Weka's GainRatioAttributeEval evaluation metrics [19]. Ranker class is used to give ranking to each attribute. Parameter M decides the number of features to keep in the dataset. We have trained and tested our proposed tag recommender with varying number of features.

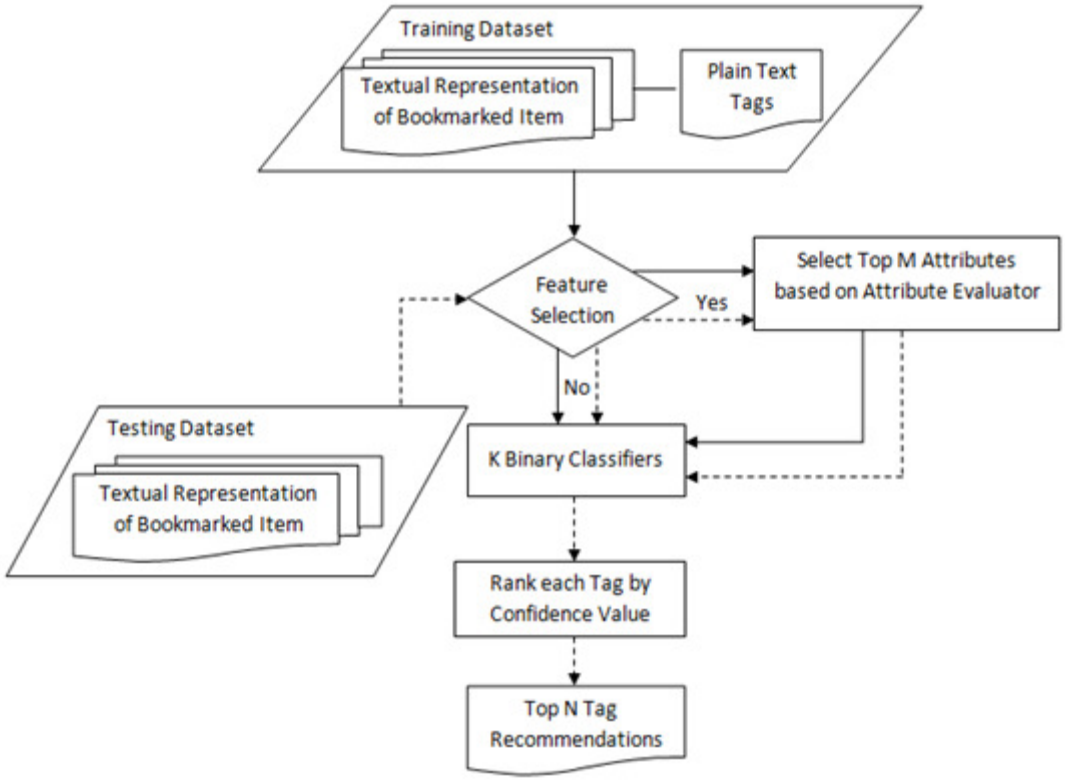

Figure 6. Flowchart of Proposed Tag Recommendation System

We have trained a set of K binary classifiers, each classifier c_k corresponding to tag $t_k \in T$, where T is the set of all available tags. T is decided in the preprocessing step for both the dataset. For any new resource d, each classifier c_k predicts the probability/confidence with which it should be annotated with t_k. The final outcome of the process is the set of TopN tags recommended by the classifiers from the available tags. The set of resources used to train the classifier is 80% of the set of all the resources previously annotated by the user. During training, each resource that is tagged with t_k is considered as a positive example for c_k, while all the other resources which are not tagged with t_k are considered as negative examples for c_k. For each test instance classifier predicts set of tags and gives ranking to each tag based on probability/confidence value. Figure 6 shows the flowchart of the proposed tag recommendation system, where solid lines indicate the learning step, while dotted lines indicate the classification step. To calculate precision, recall and f-measure, we compare tags predicted by the system for each test instance with its true tag assignments.

6. RESULTS AND DISCUSSIONS

Precision(P), recall(R) and f-measure(F) is used to evaluate our Tag Recommender's framework. Using these evaluation measures, performance of Tag Recommender is compared for Boolean, bag-of-words and TFIDF representation of resources. Experiments are carried out in two stages. In the first stage, we have evaluated tag recommender without using feature selection. Results for bookmark and BibTeX dataset are shown in Table 5 and 6 respectively. It is clear from the results that tag recommender performs best when multinomial distribution is fitted to the data.

Table 5. Tag Recommendation in Social Bookmarking System - Results on Bookmark			
P/R/F @TopN	MVBD	MND	ND
P/R/F @Top1	0.218/0.078/0.107	**0.234/0.080/0.112**	0.078/0.030/0.040
P/R/F @Top3	0.136/0.145/0.127	**0.148/0.149/0.135**	0.057/0.064/0.054
P/R/F @Top5	0.103/0.173/0.117	**0.117/0.188/0.131**	0.046/0.082/0.054
P/R/F @Top10	0.071/0.224/0.100	**0.083/0.249/0.115**	0.035/0.114/0.049

Table 6. Tag Recommendation in Social Bookmarking System - Results on BibTeX			
P/R/F @TopN	MVBD	MND	ND
P/R/F @Top1	0.501/0.454/0.465	**0.619/0.569/0.581**	0.180/0.170/0.172
P/R/F @Top3	0.246/0.629/0.339	**0.278/0.712/0.383**	0.090/0.235/0.125
P/R/F @Top5	0.165/0.691/0.256	**0.187/0.784/0.289**	0.069/0.288/0.107
P/R/F @Top10	0.094/0.780/0.163	**0.102/0.835/0.175**	0.052/0.413/0.088

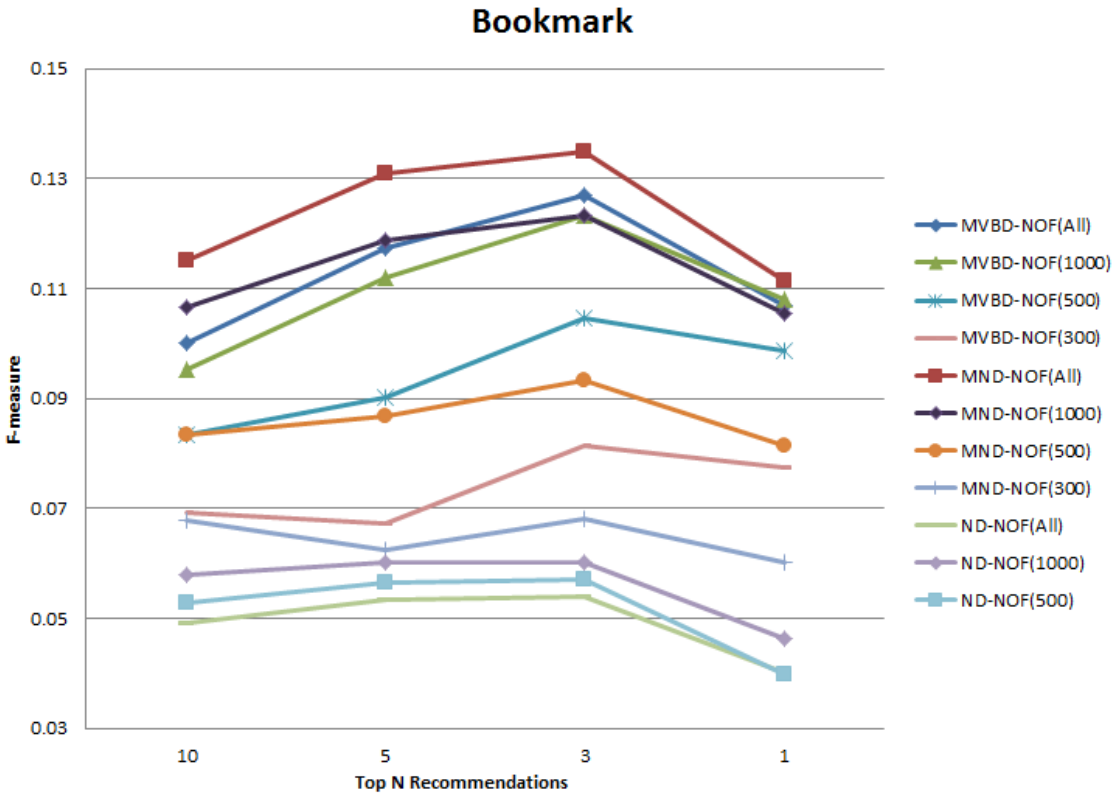

Figure 7. F-measure for Bookmark dataset

In the second stage, we have evaluated tag recommender with feature selection. Results for different number of features (NOF) are summarized in Figure 7 and Figure 8. It can be seen that while using feature selection results are favourable in case of BibTeX dataset but performance of the tag recommender is degraded in case of bookmark dataset.

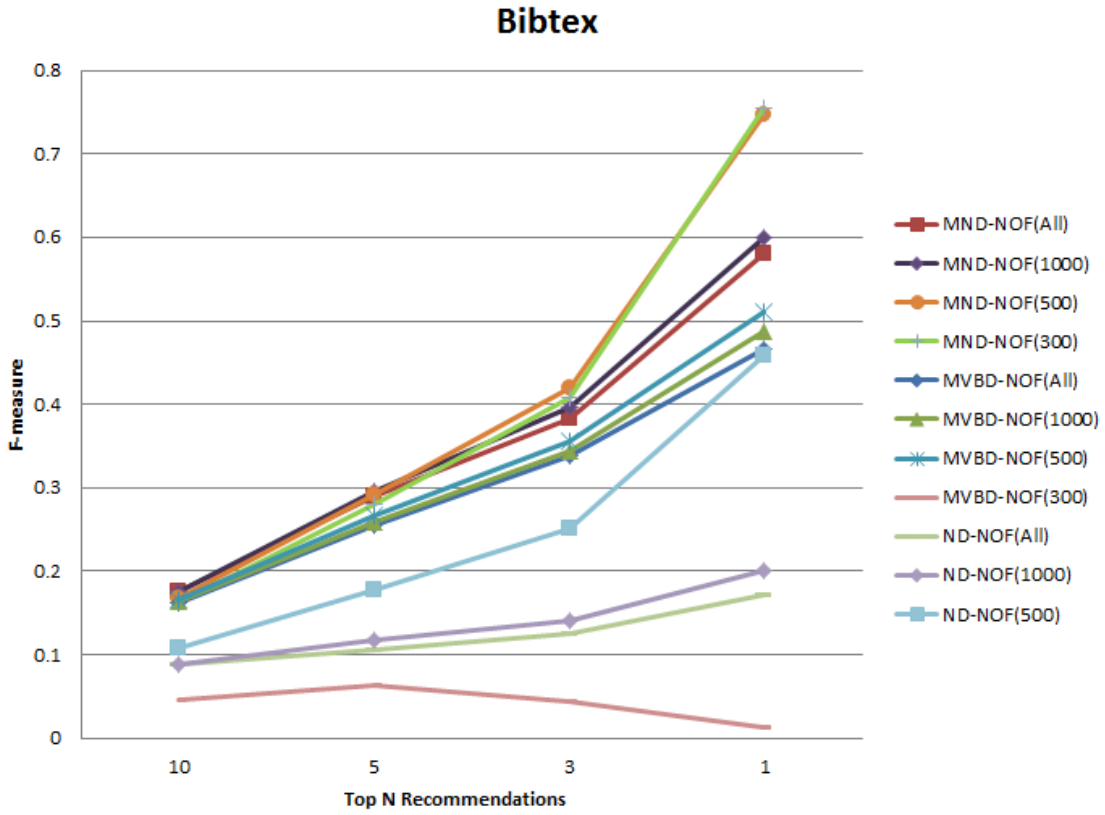

Figure 8. F-measure for BibTeX dataset

6. CONCLUSION & FUTURE WORK

This paper models tag recommendation task as multilabel text classification problem. It is evident from the experimental results that when feature selection is not used, tag recommender performs the best when multinomial distribution is fitted to the data rather than the scenarios when multivariate Bernoulli distribution or normal distribution is fitted. Incorporation of feature selection further improves the performance of tag recommender in case of BibTeX dataset but it affects adversely in case of Bookmark dataset. In future, we plan to investigate reason behind this and experiment with few more datasets. We also plan to tailor tag recommender as personalized tag recommender in future.

REFERENCES

[1] BibSonomy: http://www.bibsonomy.org

[2] Dominik Benz, Andreas Hotho, Robert Jäschke, Beate Krause, Folke Mitzlaff, Christoph Schmitz, and Gerd Stumme, The Social Bookmark and Publication Management System BibSonomy, The VLDB Journal, 19(6):849-875, Dec. 2010.

[3] A. Marchetti, M. Tesconi, F. Ronzano, R. Francesco , M. Rosella and S. Minutoli. Semkey: A semantic collaborative tagging system, Workshop on Tagging and Metadata for Social Information Organization at WWW, vol. 7, p. 8-12, 2007.

[4] I. Katakis, G. Tsoumakas and I. Vlahavas, Multilabel text classification for automated tag suggestion, In Proc. of the ECML/PKDD'08, Discovery Challenge Workshop, Belgium, p.75-83, 2008.

[5] R.Jaschke , L. Marinho, A. Hotho ,L. Schmidt-Thieme and G. Stumme. Tag recommendations in social bookmarking systems, AI Communications, vol. 21, no. 4, p. 231-247, December-2008.
[6] J. Gemmell, T. Schimoler, M. Ramezani and B. Mobasher, Adapting k-nearest neighbor for tag recommendation in folksonomies. Intelligent Techniques for Web Personalization & Recommender Systems, p. 51-62, 2009.
[7] M. Tatu, M. Srikanth and T. DSilva, Tag recommendations using bookmark content, In Proc. of the ECML/PKDD'08, Discovery Challenge Workshop, Belgium, p. 98-107, 2008.
[8] M. Lipczak, Tag Recommendation for Folksonomies Oriented towards Individual Users. In Proc. of the ECML/PKDD'08, Discovery Challenge Workshop, Belgium, p. 84-95, 2008.
[9] D. Gendarmi and F. Lanubile, Improving tag recommendation in social bookmarking systems: A Preliminary Studies, International Conference WWW/Internet, p. 133-140, 2009.
[10] Knowledge and Data Engineering Group, University of Kassel: Benchmark Folksonomy Data from BibSonomy, version of June 30th, 2007. http://www.kde.cs.uni-kassel.de/bibsonomy/dumps
[11] S. Hamouda and N. Wanas, PUT-Tag: Personalized user-centric tag recommendation for social bookmarking systems, in Springer-Verlag, p. 377-385, 2011.
[12] Z. Alepidou, K. Vavliakis and P. Mitkas, A semantic tag recommendation framework for collaborative tagging systems, IEEE International Conference on Social Computing, p. 633-636, 2011.
[13] Sanghun Ju and Kyu-Baek Hwang, A weighting scheme for Tag Recommendation in Social Bookmarking Systems, In Proc. of the ECML/PKDD'09, Discovery Challenge Workshop, Bled, Slovenia, p. 109-118, 2009.
[14] J. Read, B. Pfahringer, G. Holmes, E. Frank, Classifier chains for multi-label classification, Machine Learning Journal, Springer, vol. 85(3), 2011.
[15] H. Zhang, The optimality of naive bayes, In Proc. of the Seventeenth International Florida Artificial Intelligence Research Society Conference, The AAAI Press, p. 562-567, 2004.
[16] Markov, Zdravko, and Daniel T. Larose, Data mining the Web: uncovering patterns in Web content, structure and usage, Wiley-Interscience, 2007.
[17] Multi Label Classification, A Java Library for Multi-Label Learning: http://mulan.sourceforge.net
[18] G. Tsoumakas, I. Katakis, and I. Vlahavas. Mining multi-label data. In Data Mining and Knowledge Discovery Handbook, Springer, 2nd edition, 2010.
[19] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I.H. Witten, The WEKA data mining software: an update, SIGKDD Explorations, Newsl., vol. 11, no. 1, p. 1018, 2009
[20] M.-L. Zhang, J. M. Pena and V. Robles, Feature selection for multilabel naive Bayes classification, Information Sciences, vol. 179, no. 19, p. 3218-3229, 2009.

AUTHORS

Shweta Yagnik has done BE in Computer Engineering in 2004 from C. K. Pithawala College of Engineering and Technology, Surat. She has done M. Tech. in Computer Science & Engineering in 2013 from Nirma University, Ahmedabad and awarded with gold medal for securing 1st rank. She is currently working as an Asst. Professor in Computer Engineering dept. at L. J. Institute of Engineering and Technology, Ahmedabad, Gujarat, India. She has 6 years of teaching experience and 1 and half years of industrial experience. She has carried out her PG dissertation on Tag Recommendation in Social Bookmarking System. Her areas of interest are Data mining, Artificial Intelligence and Information security.

Priyank Thakkar is a Ph.D. Scholar of Nirma University, Gujarat, India. He has done B.E. from Sarvajanik College of Engineering & Technology, Surat in the year 2000. He has done M.E. from BVM Engineering College, V V Nagar in the year 2008. He is presently working as an Assistant Professor, CSE Department, Institute of Technology, Nirma University, Gujarat, India. He has more than 13 years of teaching experience. His area of interest is Data and Web Mining, Machine Learning and Image Processing. He has guided master thesis in the area of web page clustering, web page classification, recommender systems and spam detection.

K Kotecha is Ph.D. from IIT, Bombay. He has more than 18 years of teaching experience and guided several Masters and Doctoral thesis. Currently he is leading Institute of Technology, Nirma University as a Director. He is also a Dean, Faculty of Technology & Engineering and the Director of Academic Development & Research Cell of Nirma University. He has been nominated as a Governing Council Member of the USA based Global Engineering Dean's Council – India Chapter. He has revolutionized the Institute of Technology by implementation of ICT innovations in teaching, learning and administrative work. Institute of Technology has earned several laurels under his able-leadership.