Tag Recommendation in Social Bookmarking System

By

Shweta Yagnik

10MCES10

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INSTITUTE OF TECHNOLOGY

NIRMA UNIVERSITY

AHMEDABAD-382481

May 2013

Tag Recommendation in Social Bookmarking System

Major Project

Submitted in partial fulfillment of the requirements

For the degree of

Master of Technology in Computer Science and Engineering

By

Shweta Yagnik

(10MCES10)

Guided By

Prof. Priyank B. Thakkar

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INSTITUTE OF TECHNOLOGY

NIRMA UNIVERSITY

AHMEDABAD-382481

May 2013

Undertaking for Originality of the Work

I, **Shweta Yagnik**, Roll. No. **(10MCES10)**, give undertaking that the major project entitled **"Tag Recommendation in Social Bookmarking System"** submitted by me, towards the partial fulfillment of the requirements for the degree of Master of Technology in **Computer Science and Engineering** of Nirma University, Ahmedabad, is the original work carried out by me and I give assurance that no attempt of plagiarism has been made. I understand that in the event of any similarity found subsequently with any published work or any dissertation work elsewhere; it will result in severe disciplinary action.

Signature of Student

Date:

Place:

Endorsed by

(Signature of Guide)

Certificate

This is to certify that the Major Project entitled "**Tag Recommendation in Social Bookmarking System"** submitted by Shweta Yagnik (10MCES10), towards the partial fulfillment of the requirements for the degree of Master of Technology in Computer Science and Engineering of Nirma University, Ahmedabad, is the record of work carried out by her under my supervision and guidance. In my opinion, the submitted work has reached a level required for being accepted for examination. The results embodied in this major project, to the best of my knowledge, haven't been submitted to any other university or institution for award of any degree or diploma.

Prof. Priyank Thakkar
Guide & Assistant Professor,
CSE Department,
Institute of Technology,
Nirma University, Ahmedabad.

Prof. Vijay Ukani,
Associate Professor & PG-Coordinator(CSE),
CSE Department,
Institute of Technology,
Nirma University, Ahmedabad.

Dr. Sanjay Garg
Professor & Head,
CSE Department,
Institute of Technology,
Nirma University, Ahmedabad.

Dr K Kotecha,
Director,
Institute of Technology,
Nirma University, Ahmedabad.

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 changed the organization of bookmarks from an individual activity limited to a desktop to collective attempt over the web. User can 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. Sometimes user tends to avoid 'Tagging' as they have to describe tags at their own for the resource. These problems result into noisy or very sparse tag space and dilute the purpose of tagging. The effective solution is Tag Recommendation System, that automatically suggest appropriate set of tags while annotating resource. Here, we have studied various approaches and methodologies attempted for tag recommendation system and its performance improvement.

We have modeled tag recommendation task as multi-label classification problem. Naïve Bayes classifier is used on different representations of dataset like boolean, bag-of-words and continuous value using TFIDF representation. The work is extended with feature selection technique to choose good representative attributes from dataset for describing each document for which tags are to be recommended. The approach is evaluated against snapshot of BibSonomy dataset, a social bookmarking system and discussed effectiveness of the framework through precision, recall and f-measure metrics for different representations of dataset. It has been found from the results that the Naïve Bayes classifier for bag-of-words representation outperforms other approaches.

Acknowledgements

My deepest thanks to Prof. Priyank Thakkar, Assistant Professor, CSE Department, Institute of Technology, Nirma University, Ahmedabad, the Guide of the project that I undertook for giving his valuable inputs starting from finding the research problem to the finish of dissertation work and correcting various documents of mine with attention and care. He has taken the pain to go through the project and make necessary amendments as and when needed. I have been amazingly fortunate to have an advisor whose patience and support helped me to overcome many crisis situations and accomplish this project work.

My deep sense of gratitude to Prof. Vijay Ukani, PG-Coordinator, CSE Department, Institute of Technology, Nirma University, Ahmedabad, for an exceptional support and continual encouragement to do this research.

I am obliged to Dr. Sanjay Garg, Head, CSE Department, Institute of Technology, Nirma University, Ahmedabad, for his continuous guidance in my project work.

I would like to thank Dr K Kotecha, Hon'ble Director, Institute of Technology, Nirma University, Ahmedabad, for his unmentionable support, providing basic infrastructure and healthy research environment.

I am also grateful to my Institution, all my faculty members in CSE Department and my colleagues without whom this work would have been a distant reality.

Most importantly, I would like to express my heart-felt gratitude to my family for being constant source of love and providing support and strength throughout this endeavor.

- Shweta Yagnik

Contents

List of Tables

List of Figures

Abbreviations

BR	Binary Relevance
ECML/PKDD	European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases
MN	Multinomial
MVMN	Multivariate Multinomial
TFIDF	Term Frequency - Inverse Document Frequency

Chapter 1

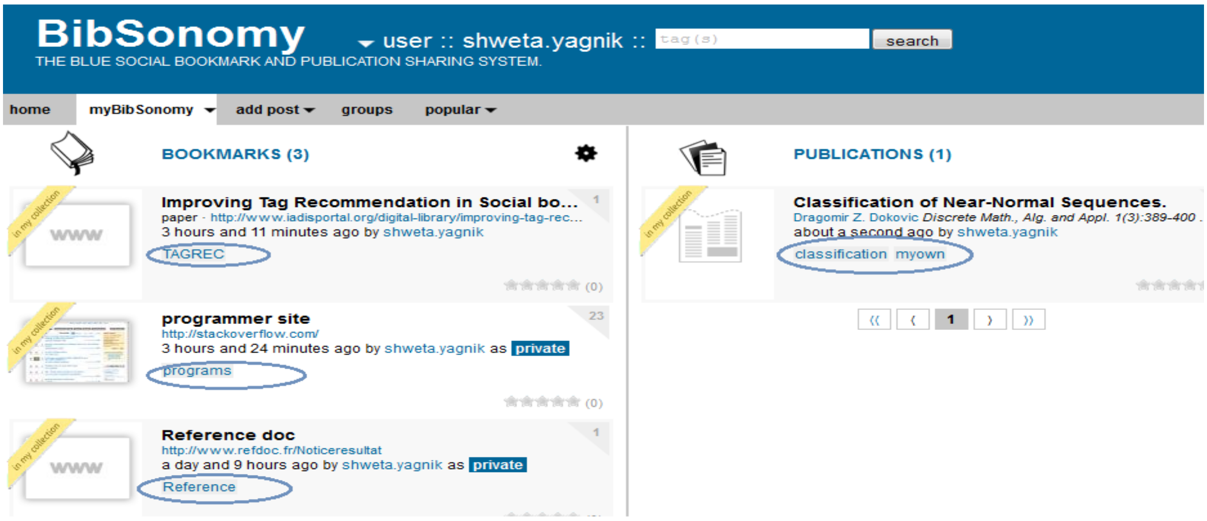
Introduction

1.1 General

Social resource sharing system is web-based system that allows users to upload, share and label the resource with arbitrary words i.e. Tags. The systems can be distinguished according to what kind of resources are supported by the site, several of them solely focus on document sharing. Among the most popular document sharing services are Del.icio.us[22] allows collaborative tagging of website bookmarks, CiteULike[23] supports references to academic publications, Flickr[24] supports sharing of images etc.

1.2 Social Bookmarking System

Social bookmarking system allows user to collect, organize and share bookmarks and publications. They have shifted the paradigm of bookmarks from an individual activity performed on a personal computer to a collective attempt over the Web[15]. All document sharing systems allow users to annotate the shared content using their own keywords i.e.Tags. Figure 1.1 is a snapshot of BibSonomy[25], a social bookmark and publication sharing system that supports collaborative tagging where user can post his bookmarks or publications and categorize them from his personal point of view by providing tags.



Improving Tag Recommendation in Social bo... 1
paper · http://www.iadisportal.org/digital-library/improving-tag-rec...
3 hours and 11 minutes ago by shweta.yagnik
TAGREC
(0)

programmer site 23
http://stackoverflow.com/
3 hours and 24 minutes ago by shweta.yagnik as private
programs
(0)

Reference doc 1
http://www.refdoc.fr/Noticeresultat
a day and 9 hours ago by shweta.yagnik as private
Reference

Classification of Near-Normal Sequences.
Dragomir Z. Dokovic *Discrete Math., Alg. and Appl. 1(3):389-400*
about a second ago by shweta.yagnik
classification myown

Figure 1.1: BibSonomy: Social Bookmark and Publication Sharing System

Annotate bookmark or publication using tags is the process of assigning short textual description to the resource that is used for information organization. Collaborative tagging allows large communities of users to collaboratively create accessible repositories of web resources. It is an additional social dimension that involves multiple users attaching freely selected tags to shared content and later they can be used for searching and categorizing the resource. The traditional hierarchical data structure design based on directories is replaced by flexible tag-based taxonomies defined jointly by users, also called as folksonomy[1].

1.3 Challenges in Tagging Process

The simplicity of collaborative tagging for user-centric content publishing and management comes at the cost of following challenges[17]:

- The freedom of selecting tags compels user to write descriptive tags on their own to define their 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.

- Users may choose tags based on their knowledge background and preferences. They may describe the same object based on different granularity level resulting into noisy tag space. It creates difficulty to find relevant material based on such tags.

- Different tags, which are either synonymous or have closely related meaning increase data redundancy, leading to reduced recall of information [14].

- While tagging users may use polysemous words i.e. a word that has many contextual meanings, this can lead to inappropriate connections between items.

Above mentioned hurdles in tagging process creates very sparse or noisy tag-space that ultimately dilutes the purpose of tagging for information organization.

1.4 Thesis Organization

The rest of the thesis is organized as follows.

Chapter 2, *Introduces tag recommendation in BibSonomy and discusses its objectives. The requirements of effective tag recommendation system are also mentioned.*

Chapter 3, *Literature survey of different approaches for tag recommendation system are mentioned in this chapter.*

Chapter 4, *Problem identified after literature survey is described in this chapter. Multi-label classification approach to solve tag recommendation task is discussed and multi-label evaluation metrics are also described.*

Chapter 5, *Statistics of obtained BibSonomy dataset and its preprocessing task is described in detail. Proposed tag recommendation algorithm is also discussed.*

Chapter 6, *Results obtained after the implementation of proposed algorithm and analysis of the results is described in this chapter.*

Chapter 7, *Concluding remarks of the work done and future scope for further optimization of the proposed algorithm is mentioned.*

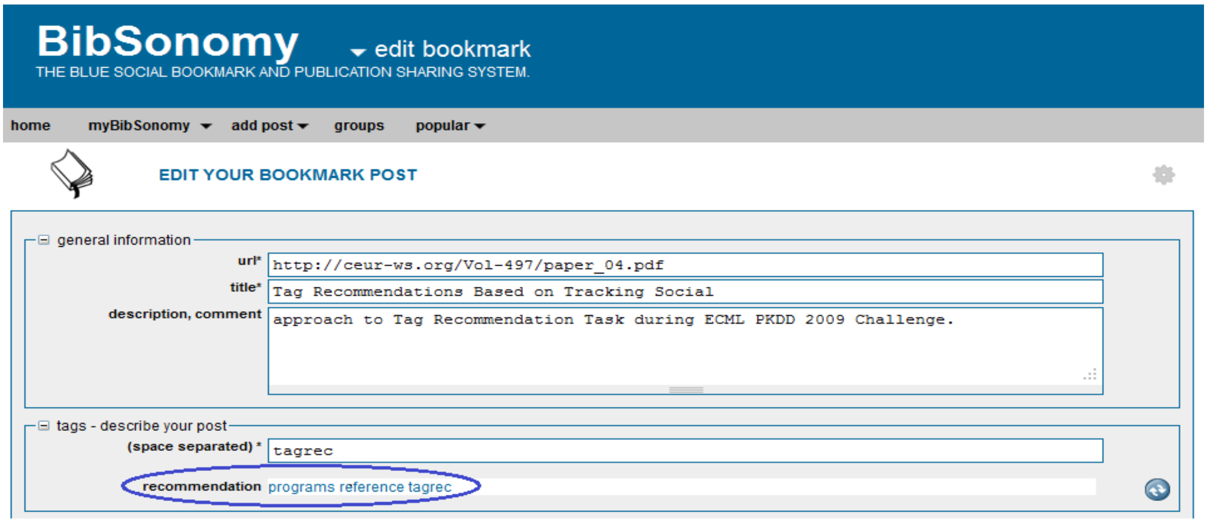
Chapter 2

Tag Recommendation System

2.1 Introduction

All challenges related to annotate a resource with tags, mentioned in previous chapter, inspires to develop methods that assist users while tagging, by automatically suggesting an appropriate set of tags.
Figure 2.1 shows tag recommendation in BibSonomy[25]. When user post a bookmark or publication, system gives suggestion for tags appropriate to the resource, based on user's past tagging history or popularity of tag for that bookmarked URL or publication.

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 increases the chances of getting a resource annotated. It helps in consolidating the vocabulary across users which exposes different facets of a resource and enriched set of tag helps user in reminding what a resource is about. Effective tag recommender avoids ambiguity in tags that helps to offer greater information value. The resultant tag space is cleaner and denser, which is useful not only in resource organization but also provides useful dataset for applying further data mining techniques for knowledge extraction.



url* http://ceur-ws.org/Vol-497/paper_04.pdf

title* Tag Recommendations Based on Tracking Social

description, comment approach to Tag Recommendation Task during ECML PKDD 2009 Challenge.

(space separated) * tagrec

recommendation programs reference tagrec

Figure 2.1: Tag Recommendation in BibSonomy

2.2 Tag Recommender System: Practical Aspects

The requirements of effective tag recommender [1] are as follows:

- **Efficiency**: Tag recommendations are expected to be delivered to the user quickly after the posting process is initialized by the user.

- **Deal with data sparsity**: Very little information about incoming post is available in the system. A popular resource with long posting history of user will naturally have precise recommendation, but such entries are just a small fraction of all encountered posts. The time complexity increases drastically if the system is designed to process all of them exclusively.

- **Deal with open-ended vocabulary**: Tag vocabulary is open-ended and constantly extended by users. Hence newly added tags should be considered in the recommendation process.

- **Generality**: Each collaborative tagging system has its own specific characteristics. Two main types of the system are: broad folksonomy and narrow folksonomy. In broad folksonomy, the same resource e.g. bookmark can be added to the system by many users. These users are not the authors of the resource, but each of them can use a personal set of tags to describe it. In narrow folksonomy a resource e.g. blog post can be added to the system only once by its author. The author is the only person who tags the resource. These differences are likely to have impact on the tagging decisions made by users, hence they must be taken into consideration while designing a tag recommender.

- **Adaptability**: Each recommendation is instantly followed by the real tags entered by the user. This feedback loop constantly brings new valuable information to the system that should be adapted into recommendation model.

Chapter 3

Literature Survey

1. **Ioannis Katakis et. al. [14], have modeled tag recommendation problem as multi-label text classification task**.

   Multi-label classification is a type of classification in which a single instance can be associated with multiple classes. The objective of the model is to predict tags for a particular user and item pair by exploiting prior knowledge about the item and/or the user. It can also make recommendation for unseen users as well as for unseen items. Problem transformation method is used to convert multi-label classification problem into one or more single-label classification task. It uses Binary Relevance classifier for problem transformation and the base learner used along with it is Naïve Bayes classifier.

   Tag Recommendation Task:

   (1) If Item (Bookmark or BibTeX) exists in the training set then suggest top N most popular tags for that item.

   (2) For unseen item, if user has tagging history then suggest top N most popular tags of the user.

   (3) If neither user's nor item's history is found then multi-label classifier is called. The model is built based on the textual content of the items.

(4) The output of the classifier is the predicted set of tags for test-set of item from all available tags in system.

However, they did not conduct an exhaustive study for parameter settings. The results achieved shows slight improvement to previous results, where the best f-measure achieved was 0.0942 and 0.0740 for bookmark & bibtex dataset, respectively.

2. **Jaschke et. al.[13] compared two tag recommendation approaches that are based on the social dimension of a folksonomy.**

(1) The first tag recommendation approach uses classic Collaborative Filtering. Ternary relation of folksonomy is converted to two dimension to apply traditional collaborative filtering. Neighborhood of user is calculated from user-tag OR user-resource projections and from this user similarity, set of Tags are suggested.

(2) Other approach proposed in this paper is a graph-based tag recommendation system based on FolkRank algorithm, an adaptation of PageRank to folksonomy graph. FolkRank algorithm exploits the graph structure of folksonomies. The graph is generated by regarding user, resource and tag i.e. U ∪ R ∪ T as the set of vertices. Edges are defined by two dimensional projections of UT, UR and RT. FolkRank is used to compute the relatedness between tags and the specific user and resource by setting the given user and resource to high preference values in PageRank. Thus, algorithm uses preference based ranking to bias the algorithm towards the query user and resource. Given a resource-user pair the system increases their weights in the folksonomy graph and runs FolkRank to spread the weights in the graph. Tags with the highest weights are returned as recommendations.

The efficiency and scalability of the training process is questionable as the process has to be executed for each incoming post that makes the system inefficient. Apart from the efficiency problem, the main limitation of graph-based methods is the sparsity of the folksonomy graph. To reduce the sparsity problem author used graph pruning up to the point where all nodes have at least p edges i.e. p-cores. The evaluation tests were performed on a dense core of folksonomy, which may not be representative of real life data.

3. **Andrew Byde et. al.[16] described Tagging based and Content based similarity and generate personalized tag recommendation.**

   Their approach is based on recommending tags from URLs that are similar to the one in question according to two variants of cosine similarity metrics, content based and tagging based similarities. Content-based method uses word frequency found in the content of the URL itself. It is capable for recommending tags for URLs that have not been previously tagged by anyone or very sparsely tagged URL. Tagging-based similarity uses common tag frequency of a resource, which it lightweight to implement compared to content based method.
   User's preferred tags are likely to be diluted by other tags in case of URL with a very large number of tags. The method gives personalization by recommending semantic tags on the basis of similarity metrics derived either from tagging data or from content analysis. Similarity between query URL and other URLs that user has already tagged is found and sum up the similarities to give a user-specific weight for each tag i.e give rank to tags that user have already used in past.

4. **Jonathan Gemmell et. al.[10] adapted K-Nearest Neighbor algorithm for tag recommendation.**

   Adapted K-Nearest Neighbor algorithm is query resource oriented while selecting neighbors and tags. It ignores users that have not tagged the query resource while finding the neighborhood of similar users with query user. Thus the number of

similarities to calculate is drastically reduced. After creating N-nearest neighbor, algorithm takes into account the tags which are applied to the query resource.
In collaborative tagging, users often reuse tag, considering this aspect boosting factor is used in addition to weight of the tag in case if the user has a history of that tag for any another resource. This will increase the chance of reusability of the tag.
To deal with noisy tag space and effectively work on dense part of the dataset, a data selection technique called p-core processing is used. Resultant dataset that confirms the occurrence of user, resource and tag in at least p posts.

5. **Marta Tatu et. al.[11] derived document and user models from the textual content of the post provided by the user for URLs and publications.**

   According to this model, tag suggestion are not only from the existing tag space i.e. from the training data, but also from the metadata provided by user to the resource like, title or description well as the content of the document is also considered. Using natural language processing techniques they extract important concepts from the textual metadata and normalize them using WordNet[26]. User model is derived from user's tagging behavior.

6. **Domenico Gendarmi et al. [9] designed Prompter - A recommender system which suggests tags according to three different aspects of a social bookmarking system: the personal tagging history, the social tagging behaviour and the textual content of the resource.**

   Tag Recommendation is retrieved from PersonalTags(u) set that represents the personal tagging history of user u i.e. distinct tags assigned by user u to all the bookmarks he has annotated, SocialTags(b) set that represents the social tagging history of bookmark b and SemanticTags(b) set that is generated by performing a semantic analysis of the text included in the resource.

Prompter is made of three distinct web services:

- Prompter Service invokes the methods provided by the other two web services.
- BibSonomy Service retrieves the personal tagging history of a user and the social tagging history of a bookmark.
- Semantic Service that provide content analysis performed on the resource pointed by the URL.

For a particular(user, resource) it generates four kinds of suggestion. As shown below, each tag-set will be generated based on common tags generated from two or more individual types of tag.

- SharedTags(u,b) = PersonalTags(u) $\bigcap$ SocialTags(b)
- PersonalSemanticTags(u,b) = PersonalTags(u) $\bigcap$ SemanticTags(b)
- SocialSemanticTags(b) = SocialTags(b) $\bigcap$ SemanticTags(b)
- SharedSemanticTags(u,b) = PersonalTags(u) $\bigcap$ SocialTags(b) $\bigcap$ SemanticTags(b)

Evaluation against a snapshot of the BibSonomy dataset revealed that, the combination of above three different aspects of a social bookmarking system improved the precision of generated tag suggestions in the case where users that already have a plentiful tagging history and bookmarks that point to popular resources within the community.

7. **Sally Hamouda et al. [2], suggested personalized tag recommendation for social bookmarking systems based on finding similar users and similar bookmarks.**

   Objective is to address two main limitations of collaborative filtering:

(1) Cold- start user problem: No sufficient user tagging history to use for recommendation.

(2) First-time seen items: Not been tagged before by any user.

Figure 3.1 depicts the flowchart of the Tag Recommendation Task.

(1) If the URL has been previously tagged, the system identifies those users of similar interests and weights their tags using the weighted tag frequency. Similarity of interest between users is fond based on cosine similarity between user's tags assignment to the URL.

(2) For the first time seen bookmark: The content of the bookmarked URL alone is crawled, then top N most frequent words are selected as initial tags for the URL content.

   i. Cold-start user: The tags are suggested using other similar documents from other users.

   ii. Active user: If user has sufficient bookmarking tag history, then a cosine similarity vector is built between the new URL and previously tagged bookmarks.

(3) Personalization is the adjusting recommended tags to reflect the vocabulary of the user. It involves mapping of the suggested tags to the user's personomy i.e. his own vocabulary of tags. This is conducted by building a co-occurrence matrix between the tags placed by user and all other user's tags who have tagged the same bookmarks.

8. **Zinovia Alepidou et. al.[3] proposed a generalized tag recommendation framework that conveys the semantics of resources according to different user profiles.**

They have integrated various models that take into account content, historic

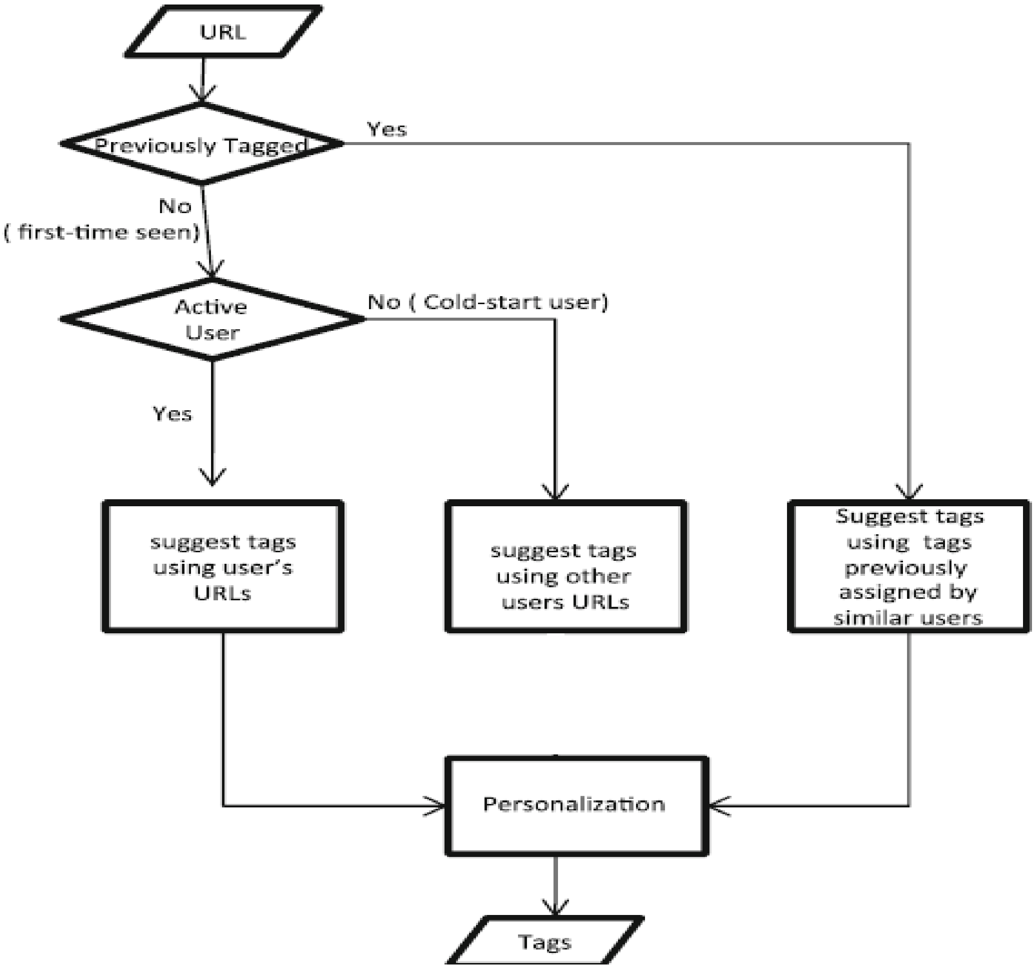

URL
Previously Tagged
Yes
No
(first-time seen)
Active User
No (Cold-start user)
Yes
suggest tags using user's URLs
suggest tags using other users URLs
Suggest tags using tags previously assigned by similar users
Personalization
Tags

Figure 3.1: Tag Recommendation Procedure

values, user preferences and tagging behavior to produce accurate personalized tag recommendations. From this information they build several Bayesian models for tag recommendation.

They created lexicon repository where words extracted from the resource are connected to number of tags. WordNet[26] was used to apply stemming to acquire the root of each term. The final tag suggestion consists of the n tags with the higher co-occurrence probability based on the words that constitute the resource's content. System was built upon resource's title, history,and other tags.

(1) Recommendation based on resource's title- TitleRecommender: Nouns, ad-

jective, verbs are retrieved by Wordnet from resource's title.

(2) Recommendation based on history- UserPersonomy: Assign weight to terms derived in the first step based on the occurrence of it in user's personomy.

(3) Tag to tag recommendation: Given a tag, other tags can be produced from its related terms, like synonyms and hypernyms. This meta-recommenders may support the basic recommenders by improving the quality and quantity of the recommendation, especially when the main recommender fails to provide a sufficient number of tags. A Bayesian 'TagToTagRecommender' which estimates tags occurrence probabilities given other tags occurrence probability (Eq. 3.1). Hence, suggesting one tag t_j leads to suggesting other tags t_i related to the first one.

$$P(t_i|t_j) = \frac{P(t_i|t_j)P(t_i)}{P(t_j)} \tag{3.1}$$

Author also evaluated recommender's performance from etymological point of view by showing results of semantic-based evaluation where content-based method shows good results.

9. **Marek Lipczak et. al. [12] evaluated and suggested potential sources of tags for suggestion focusing user's personomy.**

He has shown in his studies that the collaborative filtering used for recommending tags based on cosine similarity between users that is calculated from resource content is not a good idea for tag suggestion i.e. tags that are associated to a resource by people similar to query user may not be choice of him. He proved there is no correlation in cosine similarity between two users calculated based on tags and content item.

Based on this conclusion the author has used each post's individual content for tag recommendation. He used following three step process for tag recommendations:

(1) **Extraction of title based tags**: Primary set of tags are retrieved from resource title and they are cleaned by removing non-alphanumeric characters.

(2) **Retrieval of tags related to title**: Above set is filtered based on co-occurrence of tags attached to the resource in previous post made by other users. i.e. resource tag history.

(3) **Personomy based filtering**: Then tags are finally filtered by user's own tags he used in history of his personomy. This will reduce the size of the recommendation and improve the precision.

The result is a set of tags related to both the resource and the user.

Chapter 4

Problem Definition and Methodology

Tag recommendation systems have their own practical challenges, mentioned in Chapter 2, which led researchers to develop various methods that assist users in the tagging process, by automatically suggesting an appropriate set of tags. As discussed in previous chapter, various approaches have been made towards suggesting relevant set of tags to user for the item he is posting. But there is still scope of improvement in the performance of tag recommendation system.

4.1 Approach

We have modeled the social tag recommendation task using multi-label classification approach. Multi-label classification is the supervised learning problem where an instance may be associated with multiple labels. Tag recommendation task can be modeled as multi-label classification problem as one resource can be annotated with multiple tags based on the relevance with the resource. Different relevant tags help in exposing multiple facets of a resource.

The conventional single-label classification is learning from a set of examples that are associated with a single label λ from a set of disjoint labels-L, $|L|$ >1. Binary classification is process of classifying the instances into two disjoint sets i.e. positive and negative class. Multi-class classification is the problem of classifying instances into more than two disjoint sets.

To handle multi-label classification problem there are mainly two approaches [4] as mentioned below:

- Problem transformation methods that converts the multi-label 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 method.
- Algorithm adaptation methods that modifies learning algorithm to directly perform multi-label classification.

In our tag recommender, we have used BR problem transformation method. It is a simple classifier that scales linearly with the number of classes in a multi-label classification dataset [4]. BR classifier transforms a multi-label problem into multiple binary classification problems. 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 labeled as λ, if in the original dataset they are labeled as λ otherwise as -λ . It learns a binary classifier C_λ : X → {λ, -λ} for each label.

Along with BR classifier we adopted Naïve Bayes as a base learner because of its computational efficiency as well as its optimality for classification tasks even when the conditional independence between attributes assumption is invalid[7]. We experimented with different representations of dataset like boolean, bag-of-words and continuous value using TFIDF representation.

We have also incorporated feature selection technique to find out subset of attributes that best describe a set of documents, here Bookmark or BibTeX item. With respect to classification task we have selected attributes with which learning algorithm may achieve maximum accuracy. To achieve this task we have used GainRatioAttributeEval, a supervised attribute filter of Weka [18], to select attributes along with BinaryRelevanceAttributeEvaluator class of Mulan package [21].

4.2 Multi-label Evaluation Metrics

Performance of multi-label classification is calculated based on standard information retrieval metrics called precision, recall and f-measure [14] [6].

Notations:

P_i : Set of predicted labels for instance x_i

Y_i : Set of actual labels for instance x_i

m : Total number of test instances

a. **Precision:**

The number of correct tags retrieved divided by the total number of retrieved tags. Tag-Precision is the percentage of correctly recommended tags among all tags recommended by the tag recommendation algorithm.

$$Precision = \frac{1}{m}\sum_{i=1}^{m}\frac{|P_i \bigcap Y_i|}{|P_i|} \tag{4.1}$$

b. **Recall:**

The number of correct tags retrieved divided by the total number of correct tags. Tag-Recall is the percentage of correctly recommended tags among all tags annotated by the users i.e. actual tags.

$$Recall = \frac{1}{m}\sum_{i=1}^{m}\frac{|P_i \bigcap Y_i|}{|Y_i|} \tag{4.2}$$

c. **F-measure:**

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.

$$F - measure = \frac{1}{m}\sum_{i=1}^{m}\frac{2|P_i \bigcap Y_i|}{|P_i| + |Y_i|} \tag{4.3}$$

Chapter 5

Experimental Setup

5.1 Data Preprocessing

During the ECML/PKDD Discovery Challenge,Belgium 2008 [19], organizers provided participants with a dataset containing a portion of the data within the BibSonomy system. It contains three training files named tas, bookmark and bibtex. The original training tas file contains 8,16,197, bookmark file contains 1,76,147 and bibtex file contains 92,545 instances. We have used partial snapshot of the original training files for our experiments whose statistics are mentioned below.

a. tas: File describes tag assignments made by a user to resource and contains following details for each instace:
user_id, tag, content_id (bookmark.content_id/bibtex.content_id), content_type (1 = Bookmark item , 2 = BibTeX item) and date.
For instance, user's tag assignment record is shown in Table 5.1.

Statistics of tas files:

In our snapshot of BibSonomy dataset, tas file contains total 3,04,118 records, where user_id, conten_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 given by a user to the content.

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

Table 5.1: Tag Assignment to Resource

As 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 left with 3,03,670 records with plain text tag assignments. Thus, bookmark/publication is associated with plain text tags that can be accurately processed to generate recommendation.

There are total 1,73,568 posts of Bookmark resource i.e. content_type=1 and 1,30,102 posts of BibTeX resource i.e. conten_type=2. These post contains 50,000 unique items of 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 for BibTeX is 2.6 tags.

b. bookmark: Contains bookmark post related information in bookmark fields like content_id, url_hash, url, description, extended description and date.

Url_hash field uniquely identify bookmark item. For instance, one of the web page bookmarked by user is described by the information shown in Table 5.2.

c. bibtex: Each bookmarked publication is associated with values of bibtex fields like content_id,journal volume, chapter, edition, month, day, booktitle, how-Published, institution, organization, publisher, address, school, series, bib-

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

Table 5.2: Bookmarked Web page in BibSonomy

texKey, 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 5.3.

Because there exist a tremendous amount of posted tags, an appropriate number of tags should be selected for avoiding the over fitting problem and reducing the computational cost. We plot histogram for Bookmark and BibTeX dataset to analyze the frequency of occurrence of tag. The histogram shown in Figure 5.1 and Figure 5.2 reflects that low frequency tags i.e. tags which are used single, twice, 5-10 times etc. have dominating count.

It reveals the fact that the repository has a big number of low-frequency tags that 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 bibtext dataset keeping Tag Frequency $\geq$ 100 results into 111 unique tags. We have kept 80% of the resultant

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. The prototype with a full text and hyperlink database of at least 24 million pages available.
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

Table 5.3: Bookmarked Publication in BibSonomy

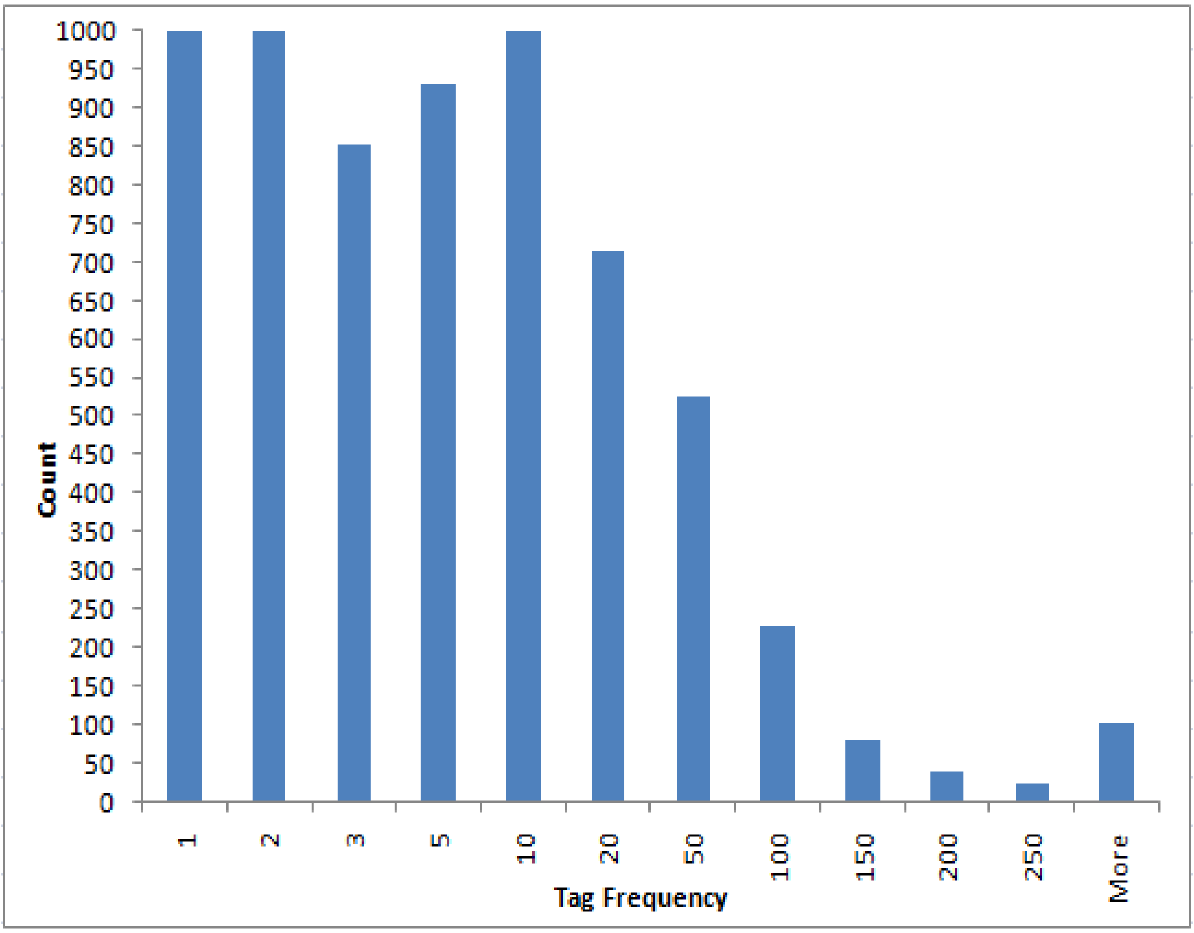


Figure 5.1: Histogram for Tag Frequency Distribution in Bookmark

dataset as training and 20% as testing.

The classifier considers the text representation of the item for which tags are to be recommended. We have experimented with different representations of dataset like boolean, bag-of-words and continuous value using TFIDF representation. In order to create textual representation for the Bookmark item we used Description and Extended Description fields and for BibTeX items, we used the Journal, Booktitle, BitexAbstract and Title fields from the dataset.

Figure 5.3 shows the conceptual flow of preprocessing step we followed for our system. We have used Weka's StringToWordVector tool to convert string attributes

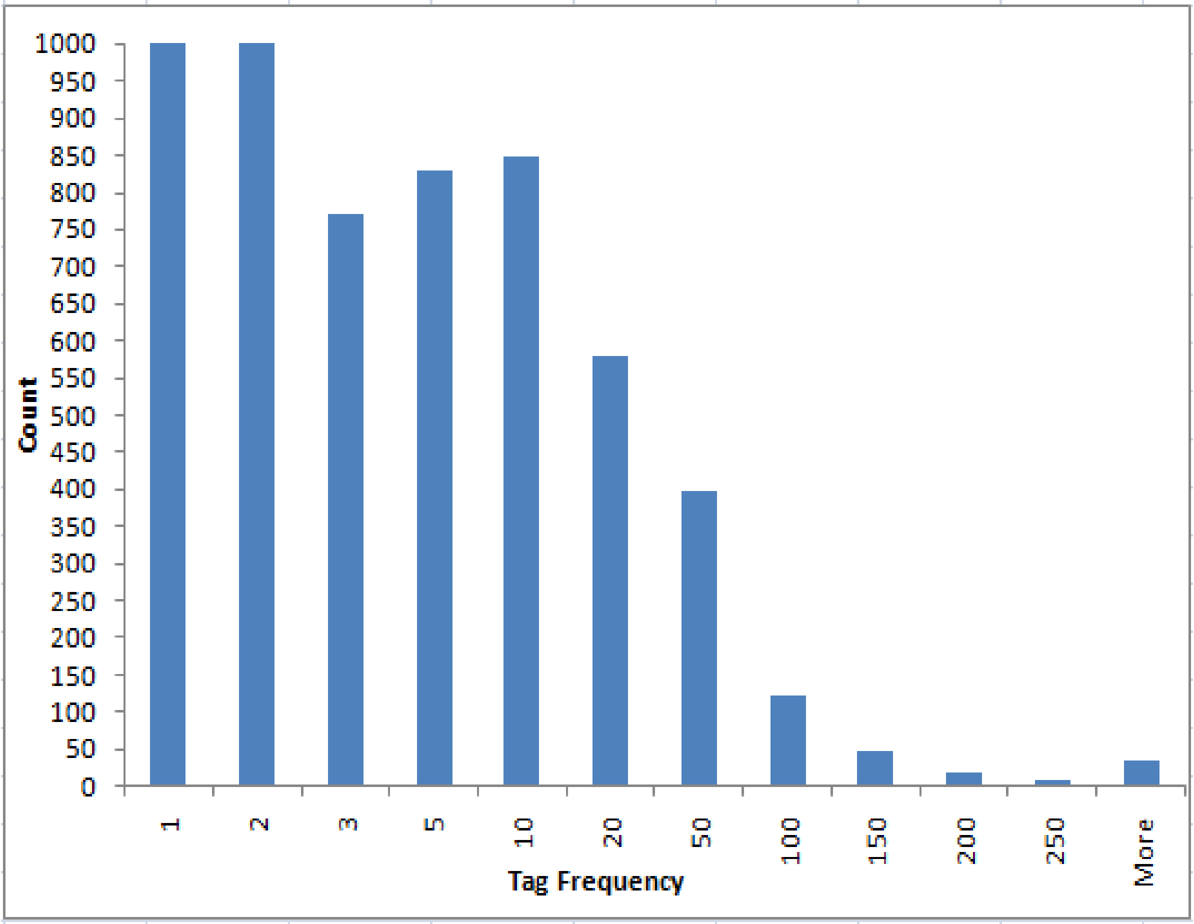


Figure 5.2: Histogram for Tag Frequency Distribution in BibTeX

into a set of attributes representing word occurrence or word frequency [18]. Bookmark and bibtex items are represented with 1,439 and 1,173 attributes, respectively.

5.2 Proposed Tag Recommendation Algorithm

We want to predict what tags a user would assign to a particular item. The important observation from the given dataset is that particular item is submitted only once by the user i.e. only once any Bookmark or BibTeX item is submitted by any user and the same item will not be repeated in the dataset. Hence every test is a

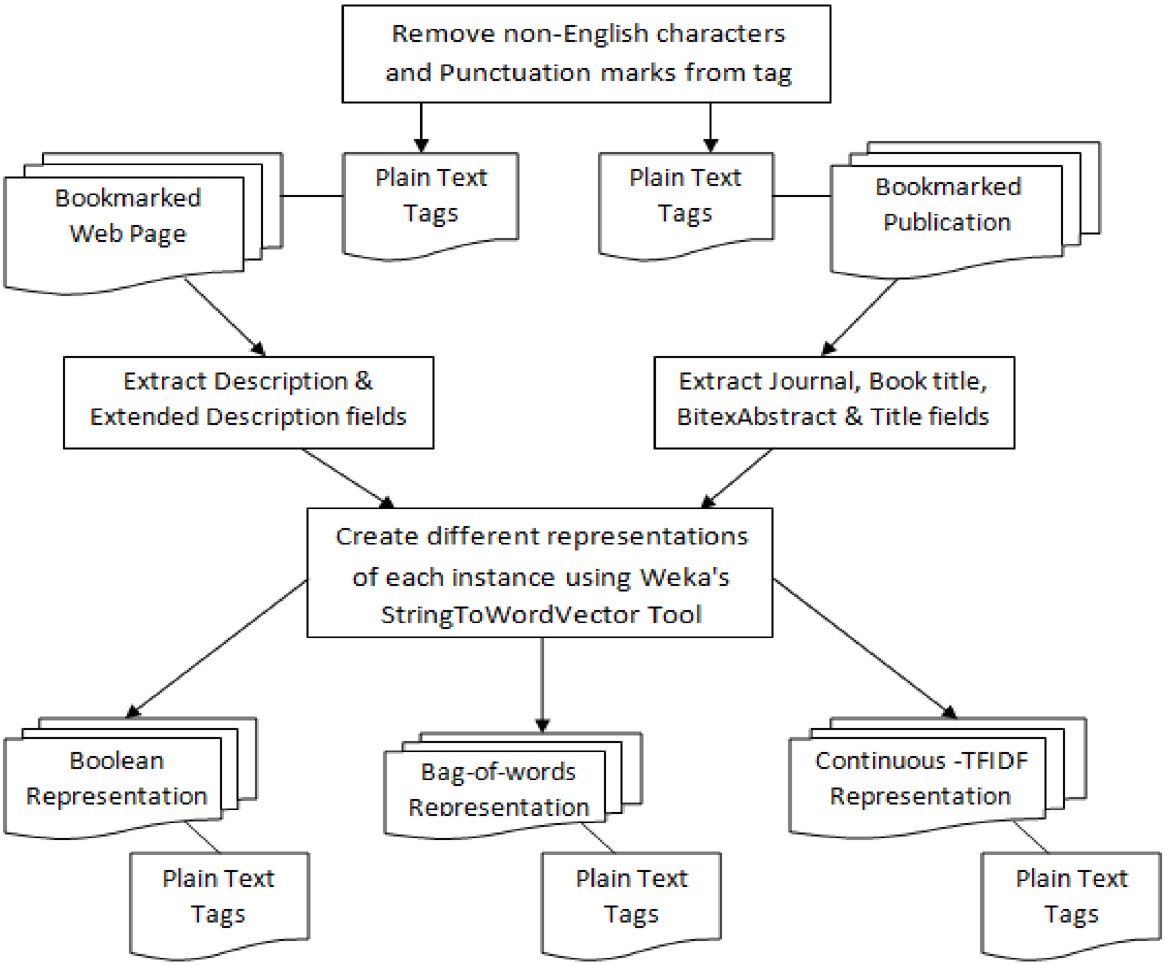

Remove non-English characters and Punctuation marks from tag
Plain Text Tags
Plain Text Tags
Bookmarked Web Page
Bookmarked Publication
Extract Description & Extended Description fields
Extract Journal, Book title, BitexAbstract & Title fields
Create different representations of each instance using Weka's StringToWordVector Tool
Boolean Representation
Bag-of-words Representation
Continuous -TFIDF Representation
Plain Text Tags
Plain Text Tags
Plain Text Tags

Figure 5.3: Conceptual Flow of Preprocessing Step

new unseen item for which we have to recommend set of tags.

We have used Binary Relevance(BR) classifier from the Mulan package[21]. This classifier considers the prediction of each label as an independent binary classification task. The base learner we have used is Naïve Bayes classifier along with BR classifier. Proposed is the Naïve Bayes classifier on document representation by bag-of-words i.e. word count with multinominal distribution. Each attribute is represented as a natural number, indicating the number of occurrences of term in the document. In this representation a document is a vector of natural numbers and its probability is computed according to the multinomial distribution.

According to Bayesian theory, we have

$$P(T|D) = \frac{P(D|T)P(T)}{P(D)} \tag{5.1}$$

In Equation 5.1, P(T|D) means the probability of using tag T given an document D. P(T) is the prior probability of tag T, it will be higher if the tag T appears more frequently. P(D|T) means the frequency of document D in a set of items which are tagged by T. To find this value in case of new document D, we can use the content of the item that can be represented by a language model [5]. Considering boolean document representation, where item D is represented with vector of n attribute values i.e., D = $a_1, a_2, ..., a_n$. Here we make the Naïve Bayes assumption i.e. the statistical attribute-independence, thus the probability of document D given the tag T will be:

$$P(D|T) = P(a_1, a_2, ..., a_n|T) = \prod_{i=1}^{n} P(a_i|t) \tag{5.2}$$

Each P(a_i|T) is calculated as the proportion of documents from class T that include attribute value a_i.

Boolean Naïve Bayes algorithm ignores the term counts. Next we discuss Bayesian model based on term counts. We use the multinomial (MN) model [20] to estimate the a posteriori probability, 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 $d_1, d_2, ..., d_n$ from class T. Let us denote the number of times that attribute a_i occurs in document d_j as n_{ij} , and the probability with which term attribute a_i occurs in all documents from class T as P(a_i|T).

$$P(a_i|T) = \frac{\sum_{j=1}^{n} n_{ij}}{\sum_{j=1}^{m} \sum_{j=1}^{n} n_{ij}} \tag{5.3}$$

The multinominal distribution defines the probability of document d_j given class T as

$$P(d_j|T) = (\sum_{j=1}^{m} n_{ij})! \prod_{i=1}^{m} \frac{P(a_i|T)^{n_{ij}}}{n_{ij}!} \tag{5.4}$$

In the bag-of-words model, the ordering of words is ignored, to consider all possible orderings of each word ($n_{ij}!$) and all words in the document($\sum_{j=1}^{m} n_{ij})!$) is added [18].

We have also experimented Naïve Bayes classifier with other document representations as mentioned below:

- Each attribute is represented by value 0 or 1 depending on whether or not the term occurs in the document. The documents in this representation are binary vectors following a multivariate multinomial distribution (MVMN) i.e. discrete distribution.

- Each attribute is represented as normally distributed continuous variable taking TFIDF values. The documents in this representation are TFIDF vectors following a normal distribution.

Both dataset have great amount of dimensionality of attributes, it is possible that some of the features may not be relevant to the classification task and other may be redundant. Feature selection in classification is necessary to reduce dimensionality and to incorporate only those attributes which are important for classification task[7], this may improve performance of classification task. To incorporate feature selection in both dataset, we have used BinaryRelevanceAttributeEvaluator from Mulan package [21]. It evaluates individual attribute based on GainRatio evaluation metrics and Ranker class is used to give ranking to each attribute. Parameter M will decide number of attributes to keep in the dataset. Then classifier is trained and tested with the reduced dimensionality of dataset.

Here we include a set of n binary classifiers, each classifier c$_k$ corresponding to tag t$_k\epsilon$ T, where T is the set of all available tags, set of tags decided in preprocessing step for both dataset. For any new document d, each classifier c$_k$ predicts whether d should be annotated with t$_k$. The final outcome of the process is the set of N tags recommended by the classifiers from the available tags. The set of documents used

to train the classifier is the set of all the documents previously annotated by the user. Each training document tagged with t_k is considered as a positive example for c_k, while the set of negative examples for c_k is represented by all documents that have not been tagged with t_k. For each test instance classifier make prediction of set of tags and gives ranking to each tag based on confidence value. Figure 5.4 shows the flowchart of the proposed tag recommendation algorithm, where solid line indicate the learning step, while dotted lines indicate the classification step.
The decisions of system for each instance is compared with its true tag assignment and precision, recall and f-measure evaluation metrics are calculated.

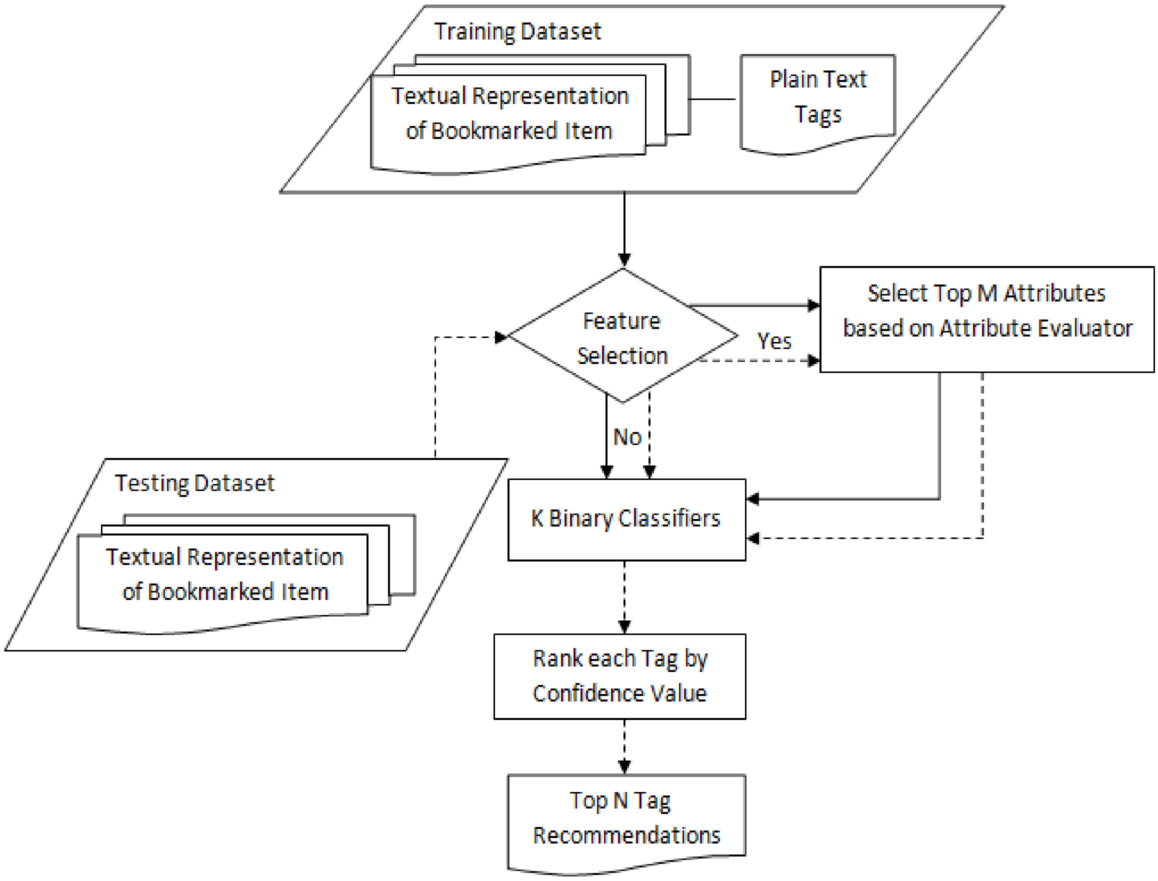

Training Dataset
Textual Representation of Bookmarked Item
Plain Text Tags
Feature Selection
Yes
Select Top M Attributes based on Attribute Evaluator
No
Testing Dataset
Textual Representation of Bookmarked Item
K Binary Classifiers
Rank each Tag by Confidence Value
Top N Tag Recommendations

Figure 5.4: Proposed Tag Recommendation Algorithm Flowchart

Chapter 6

Results and Discussion

We have calculated precision, recall and f-measure as discussed in Chapter 4, to evaluate the framework of our Tag Recommender.

6.1 Multinomial Model (Without Feature Selection)

In this model, multinomial distribution is fitted to the data i.e. documents are represented by bag-of-words and Naïve Bayes classifier is used.

- Table 6.1 shows the results for bookmark dataset.

Top N Predictions	**Precision**	**Recall**	**F-Measure**
10	0.0829	0.2490	0.1151
5	0.1168	0.1876	0.1309
3	0.1481	0.1494	0.1348
1	0.2342	0.0804	0.1115

Table 6.1: Bookmark- Multinomial Model (Without Feature Selection)

- Table 6.2 shows the results for bibtex dataset.

Top N Predictions	Precision	Recall	F-Measure
10	0.1017	0.8348	0.1748
5	0.1866	0.7842	0.2892
3	0.2776	0.7119	0.3831
1	0.6193	0.5689	0.5806

Table 6.2: BibTeX- Multinomial Model (Without Feature Selection)

6.2 Multinomial Model (With Feature Selection)

In this model, multinomial distribution is fitted to the data i.e. documents are represented by bag-of-words with selected set of features and Naïve Bayes classifier is used.

- Table 6.3 shows the results of bookmark dataset with feature selection of 1000 attributes.

Top N Predictions	Precision	Recall	F-Measure
10	0.0767	0.2317	0.1067
5	0.1051	0.1729	0.1189
3	0.1340	0.1382	0.1234
1	0.2202	0.0762	0.1055

Table 6.3: Bookmark- Multinomial Model (Feature Selection: 1000 attributes)

- Table 6.4 shows the results of bookmark dataset with feature selection of 500 attributes.

Top N Predictions	Precision	Recall	F-Measure
10	0.0599	0.1819	0.0833
5	0.0764	0.1284	0.0869
3	0.0996	0.1069	0.0932
1	0.1673	0.0589	0.0815

Table 6.4: Bookmark- Multinomial Model (Feature Selection: 500 attributes)

- Table 6.5 shows the results of bookmark dataset with feature selection of 300 attributes.

Top N Predictions	**Precision**	**Recall**	**F-Measure**
10	0.0485	0.1506	0.0679
5	0.0537	0.0954	0.0625
3	0.0704	0.0806	0.0681
1	0.1200	0.0437	0.0602

Table 6.5: Bookmark- Multinomial Model (Feature Selection: 300 attributes)

- Table 6.6 shows the results of bibtex dataset with feature selection of 1000 attributes.

Top N Predictions	**Precision**	**Recall**	**F-Measure**
10	0.1024	0.8428	0.1763
5	0.1911	0.8049	0.2965
3	0.2868	0.7389	0.3968
1	0.6364	0.5884	0.5996

Table 6.6: BibTeX- Multinomial Model (Feature Selection: 1000 attributes)

- Table 6.7 shows the results of bibtex dataset with feature selection of 500 attributes.

Top N Predictions	**Precision**	**Recall**	**F-Measure**
10	0.0961	0.8301	0.1673
5	0.1840	0.8155	0.2909
3	0.2951	0.8003	0.4188
1	0.7651	0.7410	0.7464

Table 6.7: BibTeX- Multinomial Model (Feature Selection: 500 attributes

- Table 6.8 shows the results of bibtex dataset with feature selection of 300 attributes.

Top N Predictions	Precision	Recall	F-Measure
10	0.0921	0.8151	0.1613
5	0.1755	0.7999	0.2805
3	0.2830	0.7902	0.4076
1	0.7687	0.7515	0.7555

Table 6.8: BibTeX- Multinomial Model (Feature Selection: 300 attributes

6.3 Multivariate Multinomial Model (Without Feature Selection)

In this model, multivariate multinomial distribution is fitted to the data i.e. documents are represented by binary value of attributes and Naïve Bayes classifier is used.

- Table 6.9 shows the results of bookmark dataset without feature selection.

Top N Predictions	Precision	Recall	F-Measure
10	0.0712	0.2239	0.1001
5	0.1027	0.1731	0.1173
3	0.1362	0.1448	0.1270
1	0.2175	0.0782	0.1069

Table 6.9: Bookmark- Multivariate Multinomial Model (Without Feature Selection)

- Table 6.10 shows the results of bibtex dataset without feature selection.

Top N Predictions	Precision	Recall	F-Measure
10	0.0944	0.7802	0.1628
5	0.1654	0.6913	0.2558
3	0.2457	0.6285	0.3389
1	0.5008	0.4541	0.4652

Table 6.10: BibTeX- Multivariate Multinomial Model (Without Feature Selection)

6.4 Multivariate Multinomial Model (With Feature Selection)

In this model, multivariate multinomial distribution is fitted to the data i.e. documents are represented by binary value of selected set of features and Naïve Bayes classifier is used.

- Table 6.11 shows the results of bookmark dataset with feature selection of 1000 attributes.

Top N Predictions	Precision	Recall	F-Measure
10	0.0680	0.2115	0.0954
5	0.0983	0.1645	0.1121
3	0.1320	0.1401	0.1233
1	0.2204	0.0785	0.1080

Table 6.11: Bookmark- Multivariate Multinomial Model (Feature Selection: 1000 attributes)

- Table 6.12 shows the result of bookmark dataset with feature selection of 500 attributes.

Top N Predictions	**Precision**	**Recall**	**F-Measure**
10	0.0590	0.1889	0.0835
5	0.0767	0.1370	0.0901
3	0.1087	0.1219	0.1047
1	0.1941	0.0725	0.0988

Table 6.12: Bookmark- Multivariate Multinomial Model (Feature Selection: 500 attributes)

- Table 6.13 shows the result of bookmark dataset with feature selection of 300 attributes.

Top N Predictions	**Precision**	**Recall**	**F-Measure**
10	0.0488	0.1590	0.0693
5	0.0559	0.1071	0.0674
3	0.0819	0.0982	0.0815
1	0.1458	0.0575	0.0774

Table 6.13: Bookmark- Multivariate Multinomial Model (Feature Selection: 300 attributes)

- Table 6.14 shows the results of bibtex dataset with feature selection of 1000 attributes.

Top N Predictions	**Precision**	**Recall**	**F-Measure**
10	0.0949	0.7832	0.1635
5	0.1671	0.6972	0.2583
3	0.2499	0.6374	0.3442
1	0.5263	0.4762	0.4880

Table 6.14: BibTeX- Multivariate Multinomial Model (Feature Selection: 1000 attributes)

- Table 6.15 shows the results of bibtex dataset with feature selection of 500 attributes.

Top N Predictions	**Precision**	**Recall**	**F-Measure**
10	0.0968	0.8004	0.1667
5	0.1733	0.7223	0.2674
3	0.2596	0.6588	0.3563
1	0.5556	0.4965	0.5106

Table 6.15: BibTeX- Multivariate Multinomial Model (Feature Selection: 500 attributes)

- Table 6.16 shows the results of bibtex dataset with feature selection of 300 attributes.

Top N Predictions	**Precision**	**Recall**	**F-Measure**
10	0.0265	0.2188	0.0455
5	0.0397	0.1785	0.0628
3	0.0311	0.0834	0.0436
1	0.0162	0.0129	0.0135

Table 6.16: BibTeX- Multivariate Multinomial Model (Feature Selection: 300 attributes)

6.5 Normal Model (Without Feature Selection)

In this model, Normal distribution is fitted to the data i.e. documents are represented by TFIDF value of attributes and Naïve Bayes classifier is used.

- Table 6.17 shows the results of bookmark dataset without feature selection.

Top N Predictions	**Precision**	**Recall**	**F-Measure**
10	0.0347	0.1137	0.0492
5	0.0462	0.0818	0.0535
3	0.0566	0.0637	0.0540
1	0.0784	0.0298	0.0401

Table 6.17: Bookmark- Normal Model (Without Feature Selection)

- Table 6.18 shows the results of bibtex dataset without feature selection.

Top N Predictions	Precision	Recall	F-Measure
10	0.0518	0.4131	0.0883
5	0.0690	0.2881	0.1065
3	0.0900	0.2345	0.1251
1	0.1804	0.1697	0.1721

Table 6.18: BibTeX- Normal Model (Without Feature Selection)

6.6 Normal Model (With Feature Selection)

In this model, Normal distribution is fitted to the data i.e. documents are represented by TFIDF value of selected set of features and Naïve Bayes classifier is used.

- Table 6.19 shows the results of bookmark dataset with feature selection of 1000 attributes.

Top N Predictions	Precision	Recall	F-Measure
10	0.0410	0.1327	0.0580
5	0.0516	0.0925	0.0603
3	0.0624	0.0716	0.0603
1	0.0905	0.0344	0.0464

Table 6.19: Bookmark- Normal Model (Feature Selection: 1000 attributes)

- Table 6.20 shows the results of bookmark dataset with feature selection of 500 attributes.

Top N Predictions	Precision	Recall	F-Measure
10	0.0373	0.1211	0.0527
5	0.0484	0.0862	0.0564
3	0.0589	0.0682	0.0572
1	0.0768	0.0292	0.0399

Table 6.20: Bookmark- Normal Model (Feature Selection: 500 attributes)

- Table 6.21 shows the results of bibtex dataset with feature selection of 1000 attributes.

Top N Predictions	Precision	Recall	F-Measure
10	0.0523	0.4183	0.0892
5	0.0759	0.3205	0.1177
3	0.1008	0.2641	0.1405
1	0.2113	0.1983	0.2013

Table 6.21: BibTeX- Normal Model (Feature Selection: 1000 attributes)

- Table 6.22 shows the result of bibtex dataset with feature selection of 500 attributes.

Top N Predictions	Precision	Recall	F-Measure
10	0.0616	0.5361	0.1073
5	0.1110	0.5077	0.1778
3	0.1736	0.4852	0.2504
1	0.4707	0.4540	0.4578

Table 6.22: BibTeX- Normal Model (Feature Selection: 500 attributes)

Figure 6.1 shows graph of f-measures for bookmark dataset obtained in experiments by fitting different distributions of dataset and using Naïve Bayes classifier to generate top N recommendations. It is clear from the figure that when multinomial distribution is fitted to the bookmark dataset, the results are good. Maximum f-measure achieved is 0.1348 when top 3 tag recommendations are generated in this distribution.

Figure 6.2 shows graph of f-measures for bibtex dataset obtained in experiments by fitting different distributions of dataset and using Naïve Bayes classifier to generate top N recommendations. It is clear from the figure that when multinomial distribution with feature selection is fitted to the bibtex dataset, the results are good. Maximum f-measure achieved is 0.7555 when only one tag recommendation is generated in multinomial model with feature selection of 300 attributes. While in case of

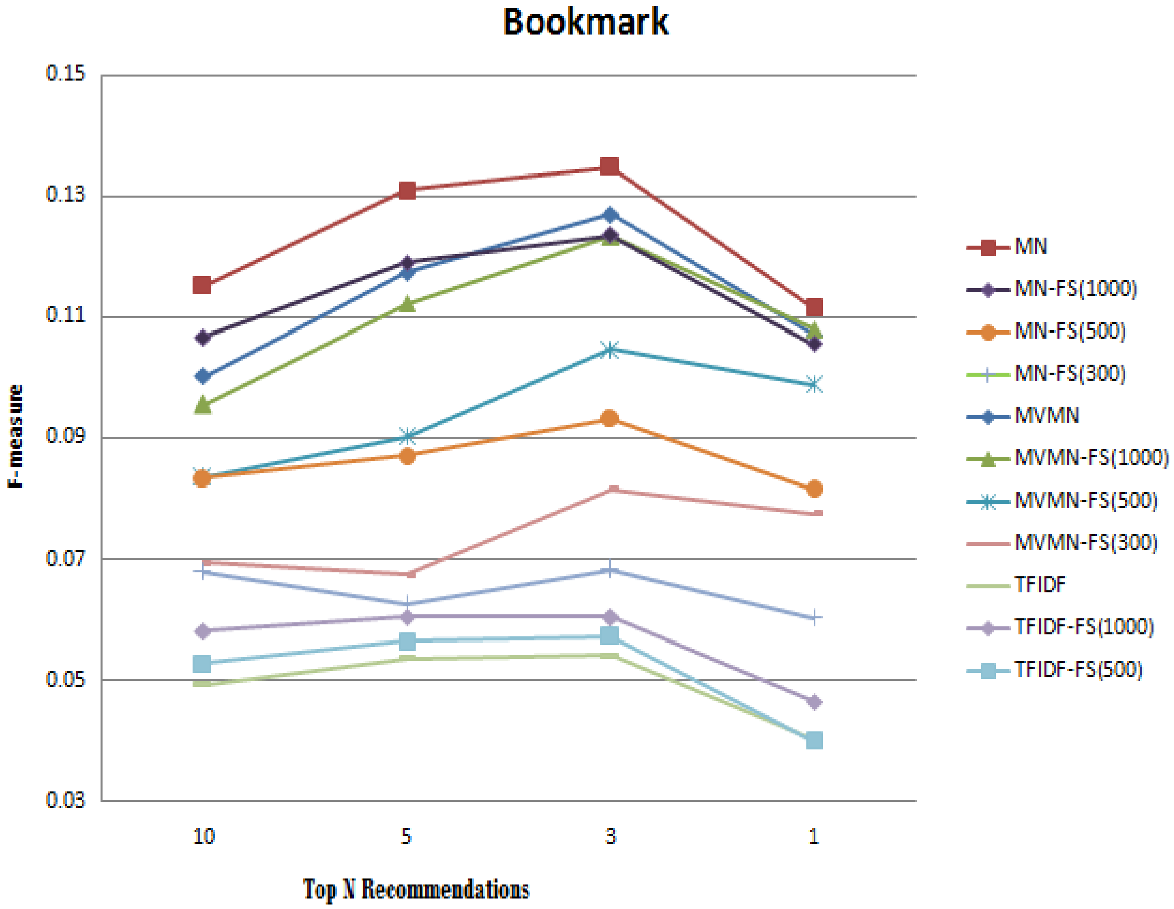


Figure 6.1: F-measure for Bookmark dataset

top 3 tag recommendations, maximum f-measure achieved is 0.4188 in multinomial model with feature selection of 500 attributes.

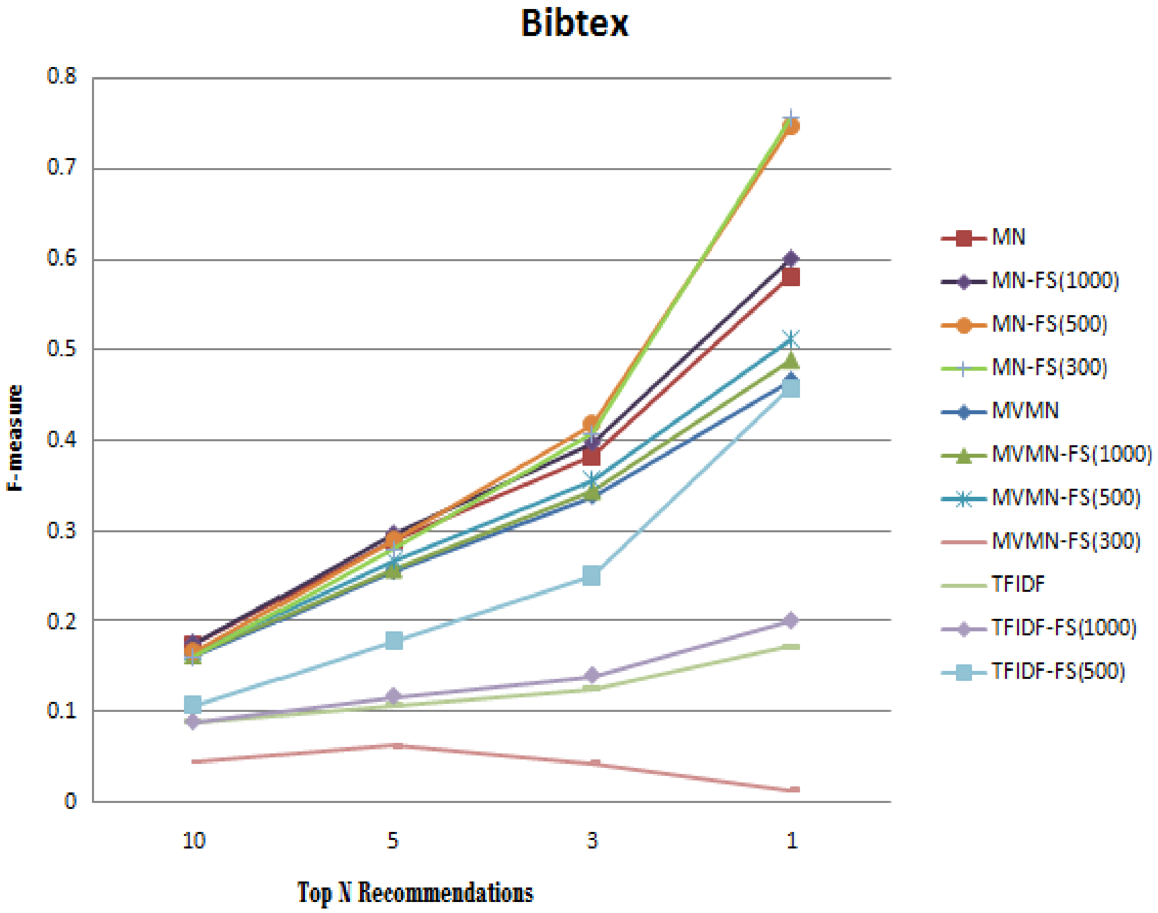


Figure 6.2: F-measure for BibTeX dataset

Chapter 7

Conclusion and Future Work

7.1 Conclusion

Tag Recommendation task is modeled as multi-label classification problem in our implementation. Naïve Bayes classifier is used as a base learner for classification with binary relevance classifier. When multinomial distribution is fitted to the data, results are improved compared to the scenarios when normal or multivariate multinomial distribution is fitted to data and classified with Naïve Bayes classifier. Results show further improvement when feature selection is applied and multinomial distribution is fitted to the bibtex dataset.

7.2 Future Work

In future, current tag recommendation approach can be extended by incorporating history of the user and item. Also, Latent Semantic Indexing can directly be applied to data in order to reduce the dimensionality.

References

[1] M. Lipczak, Hybrid tag recommendation in collaborative tagging systems, Ph.D. Thesis, Dalhousie University, March 2012.

[2] S. Hamouda and N. Wanas, PUT-Tag: personalized user-centric tag recommendation for social bookmarking systems, in Springer-Verlag , p. 377-385, 2011.

[3] 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.

[4] J. Read, B. Pfahringer, G. Holmes, E. Frank, Classifier chains for multi-label classification, Machine Learning Journal, Springer, vol. 85(3),2011.

[5] D. Yin, Z. Xue, L. Hong and B. Davison, A probabilistic model for personalized tag prediction, In Proc. of the 16th ACM SIGKDD international conference on knowledge discovery and data mining, p. 959-968. ACM, July-2010.

[6] G.Tsoumakas M.-L. Zhang, Z.-H. Zhou, Learning from multi-label data tutorial at ECML/PKDD'09, Bled, Slovenia, September-2009.

[7] M.-L. Zhang, J. M. Pena and V. Robles, Feature selection for multi-label naive bayes classification, Information Sciences, vol. 179, no. 19 , p. 3218-3229, 2009.

[8] W. Cheng and E. Hullermeier, Combining instance-based learning and logistic regression for multilabel classification, Machine Learning, vol. 76, no. 2, p. 211-225, 2009.

[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] 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.

[11] 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.

[12] M. Lipczak, Tag Recommendation for Folksonomies Oriented towards Individual Users. In Proc. the ECML/PKDD'08, Discovery Challenge Workshop, Belgium, p. 84-95, 2008.

[13] 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.

[14] 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.

[15] P. Basile, D. Gendarmi, F. Lanubile and G. Semeraro, Recommending smart tags in a social bookmarking system. In Bridging the Gep between Semantic Web and Web 2.0 -SemNet 2007, p. 22-29, 2007.

[16] A. Byde, H. Wan and S. Cayzer, Personalized tag recommendations via tagging and content-based similarity metrics. In Proc. of the International Conference on Weblogs and Social Media, Boulder, Colorado, USA, March-2007.

[17] 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.

[18] Markov, Zdravko, and Daniel T. Larose, Data mining the Web: uncovering patterns in Web content, structure and usage. Wiley-Interscience, 2007.

[19] 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

[20] A. McCallum, K. Nigam. A comparison of event models for naive bayes text classification. In Proc. of the AAAI/ICML-98 workshop on learning for text categorization, vol.752, p. 4148, July-1998.

[21] Multi Label Classification, A Java Library for Multi-Label Learning: http://mulan.sourceforge.net

[22] Del.icio.us: http://www.del.icio.us

[23] CiteULike: http://www.citeulike.org

[24] Flickr: http://www.flickr.com

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

[26] WordNet: http://wordnet.princeton.edu