

TAG RECOMMENDATION USING LATENT DIRICHLET ALLOCATION

by

RAHUL CHOUBEY

B.Tech., Jawaharlal Nehru Technological University, India, 2007

---

A THESIS

submitted in partial fulfillment of the  
requirements for the degree

MASTER OF SCIENCE

Department of Computing and Information Sciences  
College of Engineering

KANSAS STATE UNIVERSITY

Manhattan, Kansas

2011

Approved by:

Major Professor  
Dr. Doina Caragea

# Copyright

Rahul Choubey

2011

# Abstract

The vast amount of data present on the internet calls for ways to label and organize this data according to specific categories, in order to facilitate search and browsing activities. This can be easily accomplished by making use of folksonomies and user provided tags. However, it can be difficult for users to provide meaningful tags. Tag recommendation systems can guide the users towards informative tags for online resources such as websites, pictures, etc. The aim of this thesis is to build a system for recommending tags to URLs available through a bookmark sharing service, called BibSonomy. We assume that the URLs for which we recommend tags do not have any prior tags assigned to them.

Two approaches are proposed to address the tagging problem, both of them based on *Latent Dirichlet Allocation* (LDA) [Blei et al. \[2003\]](#). LDA is a generative and probabilistic topic model which aims to infer the hidden topical structure in a collection of documents. According to LDA, documents can be seen as mixtures of topics, while topics can be seen as mixtures of words (in our case, tags). The first approach that we propose, called topic words based approach, recommends the top words in the top topics representing a resource as tags for that particular resource. The second approach, called topic distance based approach, uses the tags of the most similar training resources (identified using the KL-divergence [Kullback and Liebler \[1951\]](#)) to recommend tags for a test untagged resource.

The dataset used in this work was made available through the ECML/PKDD Discovery Challenge 2009<sup>1</sup>. We construct the documents that are provided as input to LDA in two ways, thus producing two different datasets. In the first dataset, we use only the description and the tags (when available) corresponding to a URL. In the second dataset, we crawl the URL content and use it to construct the document. Experimental results show that the LDA

---

<sup>1</sup><http://www.kde.cs.uni-kassel.de/ws/dc09>

approach is not very effective at recommending tags for new untagged resources. However, using the resource content gives better results than using the description only. Furthermore, the topic distance based approach is better than the topic words based approach, when only the descriptions are used to construct documents, while the topic words based approach works better when the contents are used to construct documents.

# Table of Contents

<b>Table of Contents</b>	<b>v</b>
<b>List of Figures</b>	<b>vi</b>
<b>List of Tables</b>	<b>vii</b>
<b>Acknowledgements</b>	<b>viii</b>
<b>Dedication</b>	<b>ix</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Motivation . . . . .	1
1.2 Problem Definition . . . . .	3
1.3 Overview of the Proposed Approach . . . . .	3
<b>2 Background</b>	<b>6</b>
2.1 Topic Models and LDA . . . . .	6
2.2 Document Generation Using LDA . . . . .	7
<b>3 Data Description and Approaches for Recommending Tags</b>	<b>9</b>
3.1 Dataset Description and Preprocessing . . . . .	9
3.2 Document Construction . . . . .	12
3.3 Topic Words Based Approach . . . . .	13
3.4 Topic Distance Based Approach . . . . .	15
<b>4 Experimental Setup and Results</b>	<b>18</b>
4.1 Experiments Performed . . . . .	18
4.2 Results from Topic Words Based Approach . . . . .	19
4.3 Results from Topic Distance Based Approach . . . . .	22
<b>5 Related Work</b>	<b>26</b>
<b>6 Conclusions and Future Work</b>	<b>30</b>
6.1 Conclusion . . . . .	30
6.2 Future Work . . . . .	31
<b>Bibliography</b>	<b>33</b>

# List of Figures

2.1	Latent Dirichlet Allocation (LDA) plate representation . . . . .	7
3.1	Sample of document constructed based on description and prior tags. . . . .	12
3.2	Sample of document constructed based on the crawled content. . . . .	12
3.3	Topic words based approach . . . . .	14
3.4	Topic distance based approach . . . . .	16

# List of Tables

3.1	Sample data from the <i>bookmark</i> table in the <i>BibSonomy</i> dataset . . . . .	10
3.2	Sample data from the <i>tas</i> table in the <i>BibSonomy</i> dataset . . . . .	11
3.3	Topic words based approach . . . . .	13
3.4	Topic distance based approach . . . . .	17
4.1	Sample URLs used to gain insights into the behavior of the proposed approaches.	21
4.2	Recommended tags for the sample URLs, with 200 and 400 topics, respectively. .	21
4.3	Results on original dataset vs crawled dataset, using the topic words approach . .	22
4.4	Comparison between topic words based approach and topic distance based approach	23
4.5	Results on original dataset vs crawled dataset, using the topic distance approach .	24

# Acknowledgments

This thesis is a result of my individual work coupled with the suggestions, advise, encouragement and support from my major advisor and my committee members. I would like to take the opportunity to thank all of them.

I am grateful to my major advisor, Dr. Doina Caragea, Assistant Professor with the Department of Computing and Information Sciences at Kansas State University. Without her help and support, the work presented in this thesis would not have been possible in the first place. I wish to acknowledge that the ideas presented in this thesis are a result of my many interactions with her. Also, the Information Retrieval and Text Mining course, which I took under her, was essential for providing me with the domain knowledge necessary to perform research work in this area. I am truly blessed to have a major advisor like her.

It was a pleasure to interact with my committee member Dr. Gurdip Singh, Head of the Computing and Information Sciences Department at KSU. The Advanced Operating Systems and the Advanced Computer Networks courses which I took under him helped me understand the multithreading, concurrency and networking concepts in detail. I am grateful for all the help he has provided throughout the duration of my study.

I would also like to thank my committee member Dr. Torben Amtoft, Associate Professor with the Department of Computing and Information Sciences at KSU. He helped me in understanding the algorithm concepts clearly and taught me how to solve hard problems.

I am also thankful to Dr. David Schmidt, Professor with the Department of Computing and Information Sciences and Dr. Daniel Andresen, Associate Professor with the Department of Computing and Information Sciences for their help, support and encouragement.

A special note of thanks to all my friends, especially, Aditya, Sahil, Mallikarjuna, Sruthi, Sanmitra and Singi for helping me in the early days of my masters.



# Dedication

I dedicate this thesis to my parents, Mrs. Anita Choubey and Mr. Pradeep Kumar Choubey. My father has always taught me to stay focused and disciplined while approaching any problem, small or big. My mother has instilled in me the quality of perseverance. I believe, without these qualities and the faith that my parents have in me, I would not have been able to complete the work presented in this thesis.

# Chapter 1

## Introduction

In this chapter, we will first provide motivation for the work in this thesis in Section 1.1, followed by a brief problem definition in Section 1.2 and an overview of the methods used to solve the problem in Section 1.3.

### 1.1 Motivation

With the advent of Web 2.0, we have witnessed a significant increase in activities related to information sharing and collaboration among people on the World Wide Web. Blogs, tagging systems, wikis, social networks, maps, RSS feeds, etc. can be viewed as some of the manifestations of Web 2.0. In Web 2.0, the end user has become both the creator as well as the consumer of information, and as a consequence, we are faced with large amounts of user generated content. The scale of the data available makes it hard for users to quickly find the information that they need. Existing search engines such as [Google](http://www.google.com)<sup>1</sup>, [Bing](http://www.bing.com)<sup>2</sup>, [Yahoo! Search](http://www.search.yahoo.com)<sup>3</sup> simply fetch documents according to the keywords in the query passed by the user. However, taking advantage of the information in the data, beyond keywords, could help retrieve more useful documents. Labeled or tagged data could be useful in this respect. However, most data on the web is unlabeled. Machine learning [[Mitchell, 1997](#)] can be used to label vast amounts of unlabeled data. However, machine learning algorithms themselves

---

<sup>1</sup>[www.google.com](http://www.google.com)

<sup>2</sup>[www.bing.com](http://www.bing.com)

<sup>3</sup>[www.search.yahoo.com](http://www.search.yahoo.com)

need training data in the first place to be able to infer models that can be used to classify new data according to various categories.

Generating training data for a learning algorithm can be very expensive. In order to overcome this problem, it is desirable to have the users themselves provide labels, in the form of tags. In fact, tags can be used not only to categorize data and facilitate information retrieval, but also to assist users in browsing, when they don't have a particular query in mind. In such systems, users can see a list of tags and can choose the ones which are the closest to their needs. A tag can have many sub-tags that categorize data at a finer scale.

The need to support information retrieval from tagged data and assisted browsing, among others, has contributed to a significant growth in folksonomies and social bookmark sharing websites such as [BibSonomy](http://www.bibsonomy.org)<sup>1</sup>, [Delicious](http://www.delicious.com)<sup>2</sup>, [Flickr](http://www.flickr.com)<sup>3</sup> etc. A folksonomy is the result of personal free tagging of information and objects, for one's own retrieval, with the tagging done in a social environment<sup>4</sup>. Social bookmark websites allow users to annotate and share resources on the internet, add tags to resources saved by other users and find content relevant to them that is shared by other users. A resource could be a web page bookmark, a picture, an audio/video file or a scientific publication, to name a few. Using tags, resources can be searched more easily, better categorized and organized.

While folksonomies and social bookmark systems are very useful, sometimes, it can be difficult for a user to come up with relevant tags for resources of interest to him or her. Tag recommendation systems can make this process simpler. Recommender systems can make use of existing tags to recommend new tags for resources relevant to a user. Many recommender systems have become available in the last few years. They can be used to recommend information on a wide range of topics including movies, books, music, images, news stories. Amazon.com recommender system is one of the most popular systems in use at present. However, not many systems for recommending tags are available.

---

<sup>1</sup><http://www.bibsonomy.org>

<sup>2</sup>[www.delicious.com](http://www.delicious.com)

<sup>3</sup>[www.flickr.com](http://www.flickr.com)

<sup>4</sup>Definition provided by T. Vander Wal at <http://www.vanderwal.net/folksonomy.html>

## 1.2 Problem Definition

The aim of this work is to build a tag recommender system that can make use of existing tagged resources (specifically, websites) to produce tags for new resources that have no prior tags assigned to them. This problem is, generally, referred to as the “cold start” problem and represents a challenge for existing recommender systems. We should note that this problem is harder than the related problem of recommending extra tags to a resource that has already been assigned some tags.

Our tag recommender system is specifically designed for the kind of resources provided by the ECML/PKDD Discovery Challenge 2009<sup>1</sup> (DC’09). This challenge consisted of two tasks: content based tag recommendation and graph based tag recommendation for BibSonomy data<sup>2</sup>. BibSonomy is a folksonomy system which allows the sharing of social bookmarks and academic publications. In this challenge, the dataset provided had two parts, one consisting of URLs for which tags have been previously assigned by users (training data) and another one consisting of resources for which no prior tags have been assigned by users (test data). In this work, we focus on content based tag recommendation. We use the training data to build the system and then use the system to recommend tags for resources in the test data.

## 1.3 Overview of the Proposed Approach

To address the tag “cold start” problem defined in Section 1.2, we make use of the Latent Dirichlet Allocation (LDA) method [Blei et al., 2003] (described in detail in Section 2). LDA has been originally designed to model collections of documents, with the goal of finding short descriptions for documents in the collection. Informally, LDA assumes that each document is a mixture of topics and each topic is a mixture of words. Our goal is to use LDA on a collection of URL resources to find topics in the collection. Each resource will be represented

---

<sup>1</sup><http://www.kde.cs.uni-kassel.de/ws/dc09>

<sup>2</sup><http://www.bibsonomy.org>

as a distribution over topics, where each topic consists of tags. Thus, in our work, a resource (i.e., URL) can be seen as the equivalent of a document.

In the dataset provided by DC'09, resources are specified by a unique identifier, URL, title, description, date when the resource was bookmarked, and, possibly, tags assigned by users to that URL (for resources in the training set). Thus, “documents” corresponding to the training data can be represented using title, description and existing tags. “Documents” corresponding to the test data are represented using title and description only, as no tags are available.

We use the LDA implementation in the Machine Learning for Language Toolkit (MALLET) [McCallum, 2002] to infer a topic model for our data. Given the inferred topic model, we represent both training and test resources as topic distributions. There are two approaches that we use to recommend tags to the test resources based on topic distributions, as described below:

- **Topic words base approach:** In the topic distribution associated with a test resource, we identify the most “important” topics (i.e., topics with high probability). We extract the top words in each of these topics and recommend them as tags to the new test resource.
- **Topic distance based approach:** Using the Kullback-Leibler (KD) divergence [Kullback and Liebler, 1951], we calculate the distance between the topic distribution of the test resource and topic distributions of the training resources, and identify the most similar training resources for a given test resource. We use the tags of the most similar training resources to recommend tags for the test resource.

Experimental results on our dataset showed that the URL title and description do not contain enough information for making accurate recommendations using the LDA-based approaches described above. Therefore, to enhance the information related to each resource, we constructed an additional dataset, by crawling the content of each resource in the dataset

(i.e., the corresponding page). Thus, in this case, each resource is represented by the webpage content obtained by crawling that particular resource. As for the dataset, we used LDA to infer a topic model based on content and used the two approaches described above to recommend tags for test resources.

The remainder of the thesis is organized as follows: Chapter 2 provides the background information necessary to understand the work in this thesis. Chapter 3 describes in detail the dataset and the proposed approaches. The experimental setup and the results performed are described in Chapter 4. Chapter 5 discusses the work related to the problem addressed in this thesis. Finally, conclusions and ideas for future work are discussed in Chapter 6.

# Chapter 2

## Background

This chapter contains the background information that is required in order to understand the work presented in this thesis. Section 2.1 describes the topic model and provides an LDA overview, while Section 2.2 describes the document generation using the LDA plate model.

### 2.1 Topic Models and LDA

Probabilistic topic models [Steyvers and Griffiths, 2007] are designed to model large collections of documents, with the goal of identifying the underlying topic structure in the collection and further using this structure to produce short descriptions for the documents in the collection. Topics are defined as cluster of words that appear frequently together. Thus, topics models can be used to identify semantically related words and also to distinguish between different meanings of the same words. The hidden topic structure of a text can be used to enhance browsing, searching or to perform a similarity assessment between documents.

Generally, topic models work under the assumption that documents can be represented as mixture of topics and topics can be represented as mixtures of words. Inferring a model reduces to finding topic distributions given documents and word distributions given topics.

Latent Dirichlet Allocation [Blei et al., 2003] is one of the most popular topic modeling methods. The central theme in LDA is that documents exhibit multiple topics. This idea

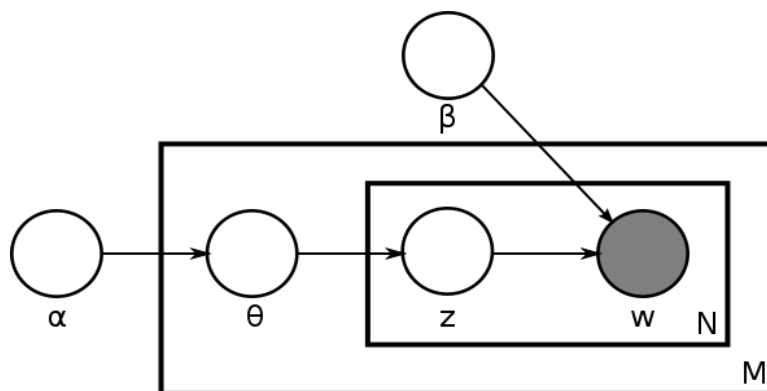
is cast into a generative probabilistic process. In probabilistic modeling, we treat data as observations arising from a generative process that involves some hidden variable, which cannot be observed. When modeling document collections the hidden variables correspond to the underlying topics in the document collection. Using the posterior inference, the hidden topic structure is inferred from the observable data or documents. As LDA is a generative model, it can be used to generate new documents. Informally, to generate a new document, we first choose a topic distribution and select a topic according to that distribution. Then, we select a word according to the word distribution corresponding to that topic. This process is explained more precisely in the following section.

## 2.2 Document Generation Using LDA

We make use of a graphical model to explain the document generation process in LDA. Graphical models are used to represent probabilistic models. In a graphical model, the nodes consist of random variables and edges denote the conditional dependencies between the random variables. The observed variables are shaded and the hidden variables are blank.

Figure 2.1 uses the *plate notation* to depict the LDA model.

Figure 2.1: Latent Dirichlet Allocation Plate Model (Licensed under the Creative Commons Attribute-Share Alike 3.0 Unported license. Accessed from: [http://commons.wikimedia.org/wiki/File:Smoothed\\_LDA.png](http://commons.wikimedia.org/wiki/File:Smoothed_LDA.png), May 2011.)



In this notation, the replicated structure can be duplicated by drawing a box around



the random variable and giving it an index, which denotes the number of repetitions. The LDA employs a probabilistic generative process for every document in a collection. We assume that each document collection has  $M$  documents and that each document  $w$ , has  $N$   $(w_1, \dots, w_n)$  words. The generative process for each document  $w$  in the document collection  $D$  can be described as follows (see [Blei et al., 2003] for details):

1. Choose  $N \sim \text{Poisson}(\xi)$ .
2. Pick  $\theta \sim \text{Dirichlet}(\alpha)$ .
3. For every word in the document:
  - (a) Choose a topic  $z_n \sim \text{Multinomial}(\theta)$ .
  - (b) Choose a word  $w_n$  from  $p(w_n|z_n, \beta)$ .

Several assumptions are made in this model. Firstly, the dimensionality  $k$  of the Dirichlet distribution and henceforth, the dimensionality of the topic variable  $z$  is fixed a priori. Secondly, the word probabilities are parameterized by a  $k \times V$  ( $V$  is the size of the vocabulary) matrix  $\beta$ , where  $\beta_{ij} = p(w^j = 1 | z^i = 1)$ , here  $w^j$  represents the  $j^{\text{th}}$  word in the vocabulary of size  $V$  represented by the  $V$ -vector  $w$ .  $N$  is indicative of the length of the document. It is not associated with any data producing variables such as  $\theta$  and  $z$ . The Poisson assumption is non-consequential and other document length distributions can be used as needed.

From Figure 2.1, we can see that the LDA model can be represented using three levels. Parameters  $\alpha$  and  $\beta$  are parameters at the document-collection level. These are sampled only once while the document collection is being generated. The document-level variables denoted by  $\theta_d$  are sampled only one time per document. Variables  $w_{dn}$  and  $z_{dn}$  are sampled once for every word in the document. In simple language,  $\theta$  indicates the topic histogram, i.e. topics and their probabilities. It is randomly chosen from the distribution over topics. For each word, we choose a topic  $z$  from the topic distribution. To choose a word, we see the topic, find out to what distribution of words for that topic and then choose a word from that distribution. The process is repeated for every word in the document and thus a document is generated. Repeating this process for every document produces a collection.

# Chapter 3

## Data Description and Approaches for Recommending Tags

This chapter describes the dataset used to perform experiments and preprocessing of the data in Section 3.1. Specifically, we used two datasets. The first dataset is the complete data set from ECML/PKDD DC'09, where resources are represented as documents using title, description and (possibly) tags. The second dataset is obtained by crawling the content for a small subset of resources belonging to 10 predefined topics. The procedure used to construct documents is outlined in 3.2. The approaches proposed in the thesis, specifically, the *topic words based approach* and the *topic distance based approach* are discussed in Sections 3.3 and 3.4, respectively.

### 3.1 Dataset Description and Preprocessing

In this thesis we recommend a set of tags for bookmarks (URLs) for which no tags have been assigned previously. The data set was made available by ECML/PKDD DC'09<sup>1</sup>. The dataset was assembled from BibSonomy, a social bookmark and publication sharing system. In the BibSonomy dataset, each post of a user consists of a resource and the set of tags assigned by the user to that resource. In other words, a post is specified by a three key-value structure consisting of the id (used to identify the user), the resource itself (URL) and the

---

<sup>1</sup><http://www.kde.cs.uni-kassel.de/ws/dc09>

set of tags assigned by the user to the resource. An example post from looks like: (342, <http://www.news.com>, (current, latest, news, international, local, weather)). Here 342 is the id used to uniquely identify a user, <http://www.news.com> is the URL corresponding to the resource and (current,latest,news,international,local,weather) denotes the set of tags assigned by the user to the resource.

The BibSonomy dataset is divided into two parts: the training dataset and the test dataset. There are two versions of the training data available. The first version is called the Cleaned Dump training data. This contains all public bookmarks and publication posts of BibSonomy before January 1, 2009. The second version called Post-Core at level 2 is a more refined version of the training dataset, in which processing is done over the training dataset, and all the users, tags and resources which occur less than two times are removed. In this thesis, we used the Cleaned Dump as training data. The test data consists of all the posts between the dates January 1, 2009 and June 30, 2009.

Both training and test datasets are organized into two tables called *tas* and *bookmark*. *Tas* stands for tag assignment. The bookmark table stores the bookmark data and the tag assignment table stores the users and the corresponding resources bookmarked by them. The columns of both tables along with sample rows from the dataset are shown in Tables 3.2 and 3.1, respectively.

Table 3.1: Sample data from the *bookmark* table in the *BibSonomy* dataset

content id	url hash	url	description	extended description	date
8	7edf1e	<a href="http://jo.fr/bstddb">http://jo.fr/bstddb</a>	LaTeX	Search	2008-12-13
9	f363ce	<a href="http://www.netlib.org">www.netlib.org</a>	The Netlib		2006-09-06
10	a1c43	<a href="http://www.gataga.com">www.gataga.com</a>	Gataga	Web Engine	2005-11-11

The *content id* column in the bookmark table matches with the *content id* column of the *tas* table and this column is used to identify a user post in the tables below. The *content id* in the bookmark table uniquely identifies the resource or the URL. The *url hash* column is the md5 value of the url. The *description* column is used to describe the resource and, in

this thesis, we assume it to be the title of the resource. The *extended description* column is used to describe the resource in more detail. The *date* column displays the date on which the resource was added to the system. In the *tas* table, the *content id* column has the same values as those in the *bookmark* table and is used to identify the resources, the *user* column is a number which uniquely identifies the user who has assigned a tag to the resource. The *date* column specifies the date on which the tag was assigned to the content id by the user.

Table 3.2: Sample data from the *tas* table in the *BibSonomy* dataset

user	tag	content id	date
0	bibtex	8	2008-12-13 08:42
1	latex	8	2008-6-14 05:12
2	software	9	2006-09-06 10:25:58
2	math	9	2006-09-06 10:25:58
2	library	9	2006-09-06 10:25:58
3	delicious	10	2005-11-11 01:22:11

To illustrate the column values we consider an example from the Tables 3.2 and 3.1, here the resource or the URL given is `http://jo.fr/bstdb`, the title is identified as `LaTeX` and the set of tags used to identify the resource are `(bibtex, latex)`.

To perform the experiments in this thesis we made use of MALLET [McCallum, 2002], a toolkit containing an implementation of LDA. The tool contains commands for generating the topic distributions for training as well as the test data. We made use of MySQL on the backend. For storing the training and test data we created *tas* and *bookmark* tables, as mentioned in Section 3.1 in the MySQL database. In order to perform experiments, both the training as well as testing data need to be imported into LDA’s internal format.

In order to be able to use the available dataset and implement the two approaches described in this thesis we need to perform some preprocessing. This step involves converting every row in both the test and the training data tables into disparate files. We coded a simple function in Java utilizing the JDBC which reads through every row in the database and performs this action. The actual algorithms begin after performing the preprocessing. Section 3.2 describes the construction of individual train and test files.

## 3.2 Document Construction

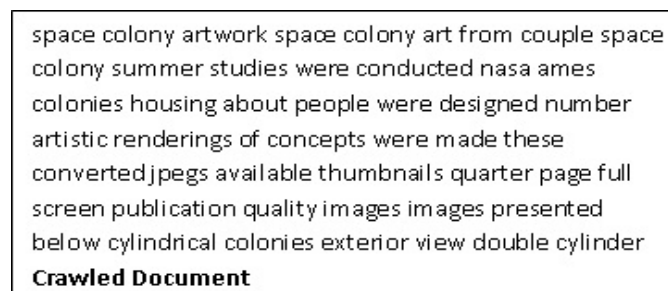
We use two procedures to construct the training documents which need to be given as input to LDA. These two procedures resulted in two different “document” collections that we used to validate our approaches. We know the tag assignments for training data. So, in the first collection, we consider the resources or the URLs along with the set of tags assigned to them by users and their titles/descriptions as documents. A sample document using description (test data) or the tags and description (train data) is shown in Figure 3.1. As



information news blog  
**Tags & Description based Document**

Figure 3.1: Sample of document constructed based on description and prior tags.

can be seen, the titles and descriptions of the resources are very sparse, and may not contain enough information for accurately predicting tags for test resources. To avoid this problem, we crawl the content for a small subset of websites from 10 predefined topics. Thus, in the second document collection, the content of the URL and the tags assigned to it make up the document that is used with LDA. For the test data, we construct documents in a similar manner. However, since we do not have tags assigned to test resources, we construct the test documents by using the description or content only. Figure 3.2 depicts a sample crawled document. As expected, we observed that the documents constructed using the



space colony artwork space colony art from couple space  
colony summer studies were conducted nasa ames  
colonies housing about people were designed number  
artistic renderings of concepts were made these  
convertedjpegs available thumbnails quarter page full  
screen publication quality images images presented  
below cylindrical colonies exterior view double cylinder  
**Crawled Document**

Figure 3.2: Sample of document constructed based on the crawled content.

title, description and tags are sparser than the documents constructed from content.

Table 3.3: Topic words based approach

1. Convert the test and train documents into LDA's internal format.
2. Use the training data to infer an LDA topic model using MALLET. This will produce the model inferencer file and a set of topics. The top ranked words in each topic are also obtained.
3. Generate the topic distribution corresponding to a test resource. Look up the highest probability topics for that resource.
4. Look up the top ranked words in those topics. These words are recommended as tags to the test resource.
5. Repeat the process for all the resources in the test data.
6. Vary the number of topics and repeat steps 1-5.

### 3.3 Topic Words Based Approach

This approach is based on the assumption that the highest probability words of the top ranked topics corresponding to a test resource in a topic model distribution are good choices as tags. Figure 3.3 depicts the whole process in the topic words based approach. The two boxes at the top of the Figure 3.3 represent the two datasets that are given as input to the LDA tool (each dataset is used in a separate experiment). Table 3.3 gives a general overview of the steps of the algorithm. More details for each step are provided below:

1. The goal of the first step is to import the documents into LDA's internal format.
2. Next, we need to infer the topic models from the training data. In order to do this we specify the number of topics as desired by us along with the number of top words in each topic. The topic file contains the topic number, a probability value and a list of top ranked words in that topic. An example from the topic file can be shown as:  
`1,0.0625,(science,biology,weather,bioinformatics,genetics).`

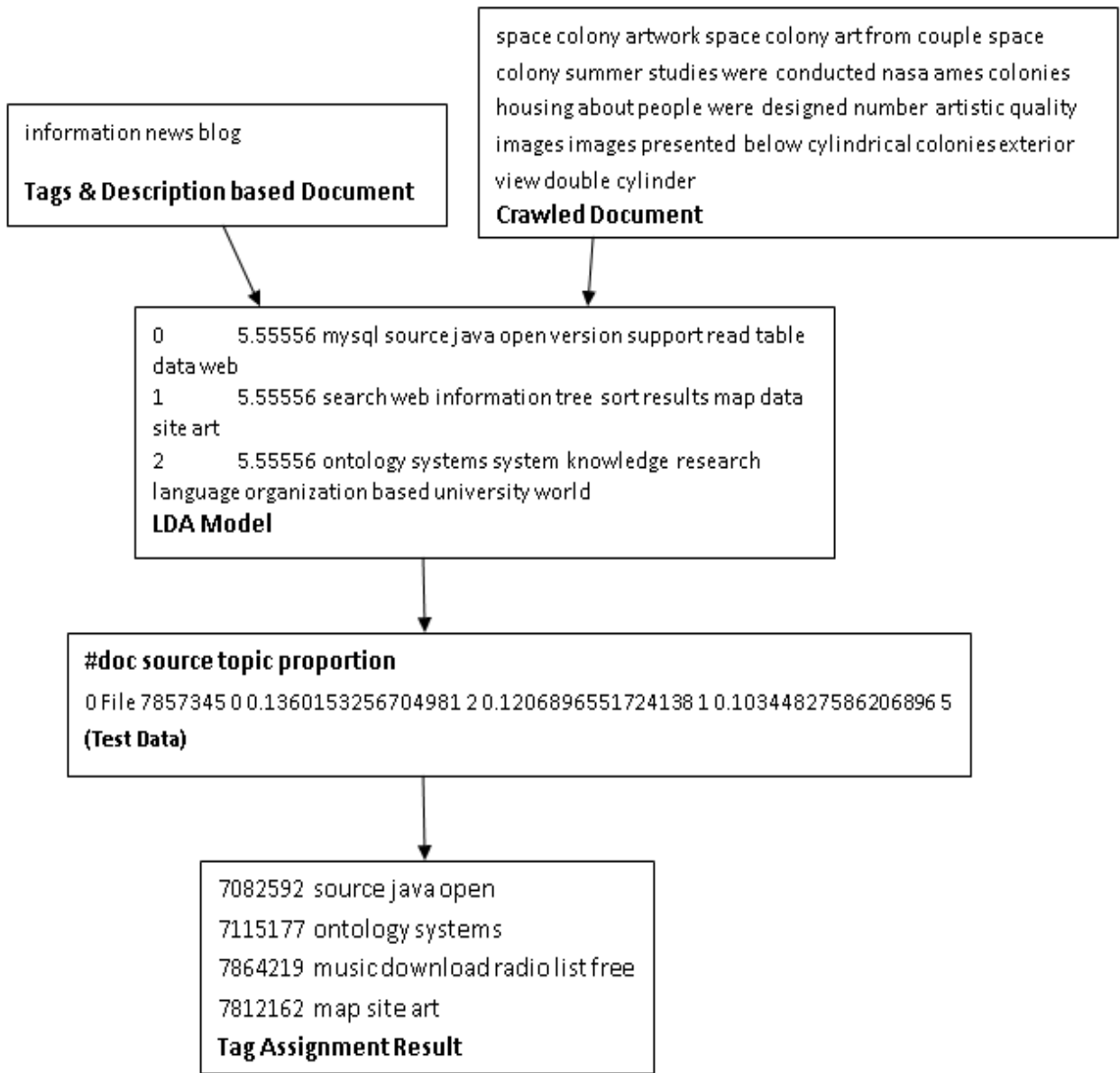


Figure 3.3: Topic words based approach

In this example, 1 indicates the topic number, 0.0625 indicates the probability value and the set of words in braces indicate the set of top ranked words in that topic.

3. After the topic models are obtained, that is, the topics along with the highest probability words in them are obtained, we determine the probability of each topic corresponding to a test resource.
4. In this step, we simply look up the words making up the highest probability topic corresponding to a test resource and suggest all those words as tags to the test resource. A function coded in Java does the work of reading through the output text file and suggesting tags to the test resource.
5. The process described above is used for all the resources in the test dataset.
6. As the last step, we vary the number of topics and observe the results in order to understand what number of topics gives the best results. The results are evaluated using a custom Java program to calculate the precision, recall and F-1 measure and is described in detail in Section 4.1.

### 3.4 Topic Distance Based Approach

In this approach, the basic idea is to calculate the distance between the test and training data topic distributions. The tags from the training document which is nearest to the test document are suggested as tags to the test document. The overview of the topic distance based approach is shown in Figure 3.4. The two boxes at the top of the figure represent the two datasets used for evaluation. As in the case of the topic words based approach, we use only one dataset at a time. Table 3.4 briefly outlines the steps of this approach. The steps are explained in more detail in what follows:

- 1-3. Steps 1. and 2. in this approach are similar to the steps 1. and 2. in the topic words based approach. In step 3, as before we obtain the probability distributions of topics



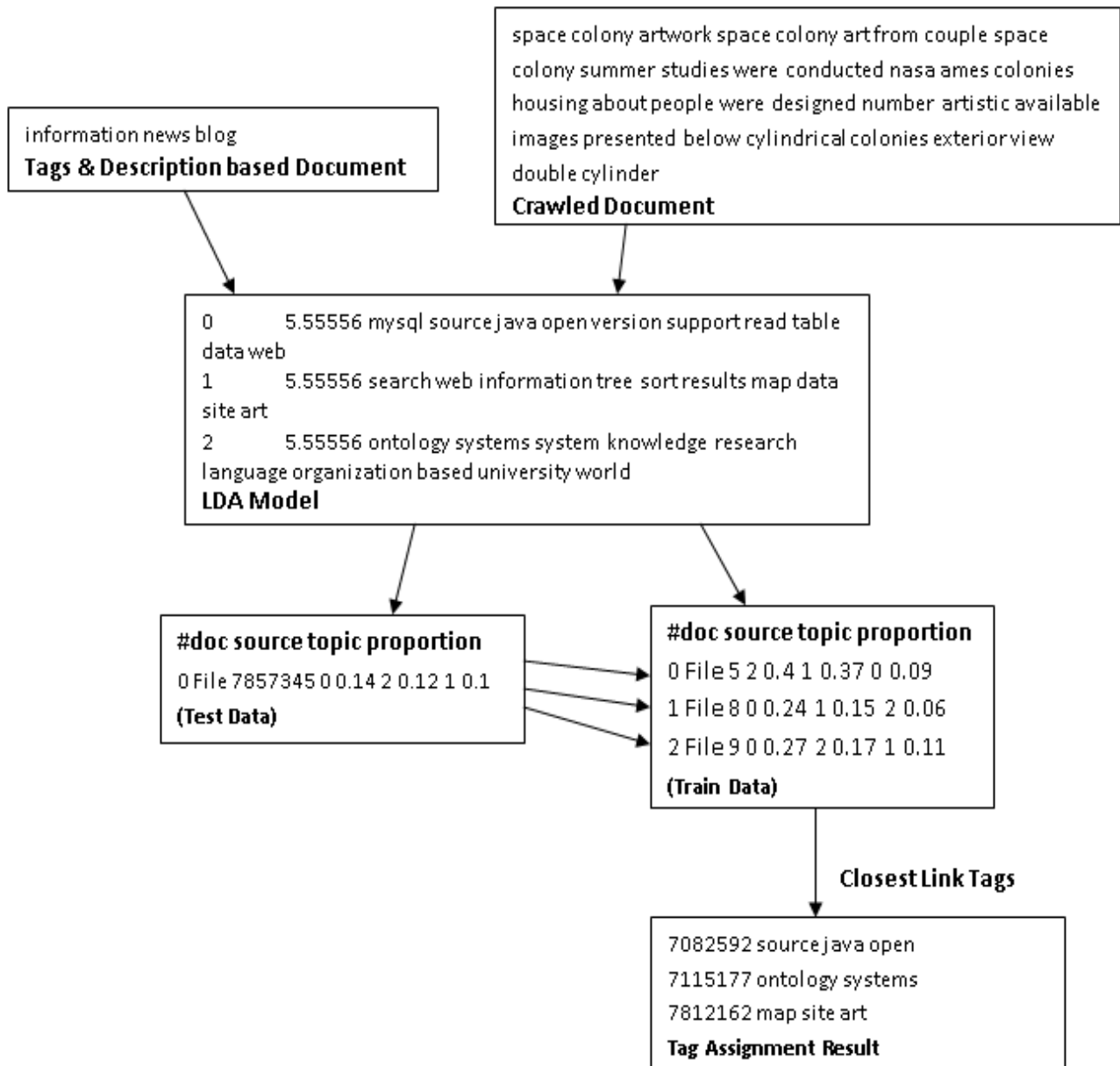


Figure 3.4: Topic distance based approach

Table 3.4: Topic distance based approach

1. Convert the test and train documents into LDA's internal format.
2. Use the training data to infer an LDA topic model using MALLET. This will produce the model inferencer file and a set of topics. The top ranked words in each topic are also obtained.
3. Generate the topic distribution corresponding to a test resource.
4. Make use of a distance metric to calculate the distance from a test resource to every train resource.
5. Take five tags from five training resources having smallest distances to the test resource and recommend them for the test resource.
6. Repeat the process for all the resources in the test data.

for test resources, but also for training resources.

4. After obtaining the topic distribution for both training and test resources, we make use of a distance metric to calculate the distance from a test resource to every train resource. The distance metric used is KL Divergence [Kullback and Liebler, 1951] which calculates the distances between two probability distributions. A function of the method coded in Java is used to calculate the distances. The KL Divergence can be obtained using the formula  $D_{\text{KL}}(P||Q) = \sum_i P(i)\log\frac{P(i)}{Q(i)}$ . Here  $P$  and  $Q$  represent the distributions of the topics corresponding to the train and test resources.
5. In the next step, the five training resources having the smallest distance to the test resource are taken and five tags from them are suggested to the test resource.
6. This process is repeated for all the test data resources. Finally, the results are evaluated using the same technique as the first approach.

# Chapter 4

## Experimental Setup and Results

This chapter starts by describing the experiments conducted to evaluate the two approaches proposed in this thesis and the evaluation criterion in 4.1. The results obtained by performing the experiments are also presented in this chapter. The results are organized into two sections, the Section 4.2 describes the results obtained using the topic words based approach and Section 4.3 describes the results obtained using the topic distance based approach. For both approaches, we first performed experiments on the dataset constructed using the description words and user defined tags, followed by experiments on the dataset constructed by crawling the content of a small number of document.

### 4.1 Experiments Performed

We have performed experiments on the *BibSonomy* data. Our experiments are meant to answer several research questions as stated below:

- How effective are the LDA approaches to recommending tags for untagged test resources? What approach works better, the topic words based approach or the topic distance based approach?
- Can we predict tags for new resources when only the resource title/description is used to represent tags as LDA documents? Do the results improve when the crawled content improves?

- How accurate are the predictions obtained with the two approaches and the two datasets considered?

The experiments performed in this thesis are meant to provide answers to these research questions. We will use both LDA approaches proposed, and both document collections that we constructed in our experiments.

In order to evaluate the tags recommended by performing the above experiments we make use of the F1-measure metric. In order to obtain F1-measure we first calculate the precision and recall values for every resource-tag prediction in the test data. Precision is defined as the fraction of tags retrieved per test document that are relevant to that test document. Recall indicates the fraction of tags that are relevant to a test document that are successfully predicted. The F1-measure can be obtained by using the formula  $(2 * precision * recall) / (precision + recall)$ . The tag table as described above in the dataset description is empty for the test data, this means that there are only null values corresponding to every test resource initially. Now, in order to evaluate the results we write the results in a file in the form of content id and set of tags corresponding to it, the number of tags varying from 5-20 in our case. We use these many tags because in general this is the number of tags recommended originally to a URL. A simple class coded in Java does this for us.

## 4.2 Results from Topic Words Based Approach

We performed experiments using the topic words based approach and the collection of documents constructed based on resource description and prior tags (when available). The F-1 measure value that we obtained was smaller than 0.1. To gain insights into the behavior of the approach and to understand why it does not perform as expected, we consider several sample URLs shown in Table 4.1 as case studies, and analyze the tags recommended by the algorithm by comparison with the actual tags. Table 4.2 shows the top five words in the top three topics, when training the LDA model with 200 and 400 topics, respectively. In this and the following tables, we highlight the related tags in italics and the tags which are

the same as the actual tag set in bold, when they exist. As we can see from the sample of the results in Table 4.2, the recommendations made by the topic words based approach do not result in accurate predictions. In fact, none of the actual tags are recovered in these samples, although some related tags are recommended. In other samples, not shown here, some of the actual tags are recovered. We can also notice that varying the number of topics does not have a significant effect on the predictions made.

Intuitively, the poor performance of the topic words based approach is due to the sparseness of the data used as input for the LDA model. The results suggest that the description words corresponding to a resource are not sufficient to characterize it. We also believe that better results in the work by Lipczak et al. [2009] are attributable to the fact that their approach was used to predict new tags to resources which may or may not have tags assigned to them previously, while our work only focuses on predicting tags to resources to which no tags have been assigned previously.

To investigate if denser documents result in better predictions, we also tested the topic words approach on the content of the crawled documents. This dataset was crawled from 10 topics and contains 500 training data files and 40 test documents. We only crawled a small number of documents, as we had to manually label resources in the 10 topics selected to this part of our study. Table 4.3 shows the comparison between the results on the crawled dataset versus results on the original dataset, using the topic words based approach (with 400 topics). From the table, we can see that the results on the crawled dataset are better than the results on the original dataset. We made use of the evaluation function mentioned in Chapter 4 and found the precision to be 0.18 and recall to be 0.21 for the crawled dataset. The F-1 measure obtained was 0.194 which was slightly better than the F-1 measure obtained (0.187) in the work by Lipczak et al. [2009], even though no resources in our test dataset have prior tags assigned to them. However, we are evaluating on a relatively small dataset and we are using the content of the resources, in this case.

Table 4.1: Sample URLs used to gain insights into the behavior of the proposed approaches.

ID	URL	Title	Actual Tag(s)
1	wrapper.tanukisoftware.org	Java Service Wrapper	java, tomcat
2	radiosure.com	Free internet radio player	radio, record, music, mp3
3	iis.hwanjoyu.org/svm-java	SVM Java	java, svm, tools
4	fomis.org	Ontology and medical science	ontology
5	videogameshop.com	Video games and accessories	games, video

Table 4.2: Recommended tags for the sample URLs, with 200 and 400 topics, respectively.

URL ID	200 Topics			400 Topics		
	Topic 1	Topic 2	Topic 3	Topic 1	Topic 2	Topic 3
1.	press association central international national	die allgemein ist ein wie	labor history global migrate union	ist die nicht ein wie	history labor union migrate chicago	<i>tech</i> commn innovate mobile telephone
2.	lis metadata <i>digital</i> standard meta	ideas commons tour creative sport	bar soft fav apps auto	de le la les en	car compare buy place auto	bar web fav webdev make
3.	ims people fonts peter type	esl education teaching medicine alter	quality university <i>computer</i> safari export	education teaching sort esl english	safari export press <i>computer</i> release	ims peter michael david bad
4.	cms manage books content tagged	food cooking india drink recipies	security privacy hack hacking meta	books tagged xref shelf <i>software</i>	food health recipies cooking india	article toread future interview issue
5.	und crossmedia im medien der	http network server monitor security	health research medical medicine journal	medical research journals medicine journal	network monitor security network protocol	und medien der trust <i>media</i>

Table 4.3: Results on original dataset vs crawled dataset, using the topic words approach

URL ID	Crawled dataset			Original dataset		
	Topic 1	Topic 2	Topic 3	Topic 1	Topic 2	Topic 3
1.	<b>java</b> <i>open</i> <i>source</i> mysql <i>version</i>	xml <i>software</i> org drm www	web info sort search tree	ist die nicht ein wie	history labor union migration chicago	<i>tech</i> co innovate mob telephone
2.	<i>free</i> <b>music</b> <b>radio</b> <i>download</i> <i>list</i>	ago news pm video game	make time work don good	de le la les en	car compare buy place auto	bar web fav webdev make
3.	<b>java</b> <i>open</i> <i>source</i> mysql <i>version</i>	xml <i>software</i> org drm www	sort tree map web data	education teaching sort esl english	<i>computer</i> safari press release export	ims peter michael david bad
4.	<b>ontology</b> research <i>web</i> system language	xml <i>software</i> org drm www	sort tree map <i>web</i> <i>data</i>	books tagged xref shelf <i>software</i>	food health recipies cooking india	article read future interview issue
5.	<b>video</b> <b>game</b> news ago pm	xml http drm <i>software</i> data	list app <b>video</b> live hotel	medical research journals medicine journal	network monitor security network protocol	und medien der trust <i>media</i>

### 4.3 Results from Topic Distance Based Approach

In this approach, we first trained the model on 400 topics and obtained the topic distributions for test resources as well as a training resources. Next, we calculated the distances between test resources and training resources topic distributions and suggested tags from the training resources having the smallest distances to the test resource. Table 4.4 shows a comparison between the topic distance approach and topic words approach, for the original dataset. We

see some missing entries in the table above because for some of the URLs we have less than five tags associated with them.

Table 4.4: Comparison between topic words based approach and topic distance based approach

URL ID	Topic distance based approach			Topic words based approach		
	URL 1	URL 2	URL 3	Topic 1	Topic 2	Topic 3
1.	saar geography met unisbde academia	<i>web</i> <b>java</b> http www <i>server</i>	<i>tech</i> <i>open</i> <i>source</i> <i>gnu</i> linux	ist die nicht ein wie	history labor union migration chicago	<i>tech</i> co innovate mob telephone
2.	geeks foocamp finance reference babsonfip	art site data paint pic	video <i>audio</i> codec <i>tech</i>	de le la les en	car compare buy place auto	bar <i>web</i> fav webdev make
3.	linux debian doc tutorial reference	<b>java</b> <i>tech</i> <i>open</i> <i>source</i>	<i>tech</i> blog web	educate teaching sort esl english	<i>computer</i> safari press release release	ims peter michael david bad
4.	cache bookmarks archive education extension	<i>science</i> <i>tech</i> blog latest	<i>data</i> <i>mining</i> <i>science</i> <i>computer</i> topic	books tagged xref shelf <i>software</i>	food health recipies cooking india	article toread future interview issue
5.	mpr swing perception folder bar	tv show episode free full	online <i>media</i> <b>video</b> tune	medical research journals medicine journal	network monitor security www protocol	und medien der trust <i>media</i>

As can be seen, the results using the topic distance based approach are slightly better than the results using the topic words based approach. This is even more obvious when we perform the evaluation using the F-1 measure, although the F-1 measure result for the topic distance approach is still below 0.1, therefore worst than the best result obtained by Lipczak et al. [2009]. We believe this is again the effect of the sparsity of the data when



only short descriptions of the resources are used to construct documents.

Finally, Table 4.5 compares the results on the crawled data set with the results on the original dataset, when the topic distance based approach is used.

Table 4.5: Results on original dataset vs crawled dataset, using the topic distance approach

URL ID	Crawled content only			Tags and description only		
	URL 1	URL 2	URL 3	URL 1	URL 2	URL 3
1.	article ontology tagging folksonomy	<b>java</b> web <i>service</i> <i>open</i> <i>source</i>	game creative video emulate sort	saar met geography unisbde academia	web http <b>java</b> www <i>server</i>	<i>tech</i> <i>source</i> <i>open</i> gnu linux
2.	article ontology tagging folksonomy	<b>radio</b> <b>music</b> <b>mp3</b> <i>free</i> fnomy	read ontology	geeks foocamp finance reference babsonfip	art site data paint pic	video <i>audio</i> codec <i>tech</i>
3.	<b>java</b> tool <i>tech</i> web service	search web engine semantic ontology	gui apps dev free	linux debian doc tutorial reference	<b>java</b> <i>tech</i> <i>open</i> <i>source</i>	<i>tech</i> blog web
4.	search engine meta small	<i>mining</i> <i>software</i>	games education design free extension	cache bookmarks archive education	<i>science</i> <i>tech</i> blog latest topic	<i>data</i> <i>mining</i> <i>science</i> type
5.	input keyboard <b>game</b> <b>video</b> <i>console</i>	bbc alzheimer brain drug	flickr pics tagging sharing folder	mpr <i>swing</i> <i>perception</i> bar free	tv show episode tune	online <i>media</i> <b>video</b> full

As can be seen from the sample results shown in the table, the results using focused crawling in distance based approach were slightly better than the results on the original dataset. We also observed this when we perform the evaluation using the F-1 measure. However, the results using the topic words based approach are better than the results using

the topic distance based approach on the crawled dataset. We believe this happens because of the fact that we were able to manually identify topics and labeled URLs according to topics, so the URLs represented the topics well.

# Chapter 5

## Related Work

This chapter provides a short overview of the work done previously in the area of Tag Recommendation that is most relevant to our work. We will start with the discussion of a study performed on the same dataset that we have used. Then, we will discuss a few other approaches, including an LDA-based approach that was evaluated on a different dataset.

As mentioned before, the dataset used in this work was obtained from the ECML-PKDD Discovery Challenge 2009<sup>1</sup>. The work by [Lipczak et al. \[2009\]](#) is also based on this dataset. In their work, the authors propose two methods for recommending tags: a content based tag recommendation method and a graph based tag recommendation method. In content based tag recommendation, the URL title and description words in a test resource are scored based on their usage as tags for previously annotated resources, i.e. for resources in the training data. Words with high scores are recommended as tags. Thus, similar to our method, this method uses independent training and test datasets, and can be used to recommend tags for new resources that don't have any prior tags. This approach took the first place in the DC'09 competition, with a reported best F1 measure value of 0.187, when 5 tags were recommended for each test resource. Our work is also content based and as mentioned before we obtained the best F-1 measure of 0.194 when the dataset was constructed using the crawled content. This is better than the best F-1 measure obtained at DC'09 competition.

In graph based tag recommendation, relations among users, resources and tags are rep-

---

<sup>1</sup><http://www.kde.cs.uni-kassel.de/ws/dc09>

resented as a tripartite graph. In this type of recommendation, the test data consists of the resources, users and tags that are present at least two times in the training data. Thus, this method is useful to predict tags to resources which have some prior information about them, but can not be used for new resources. The method took the third place in the DC'09 competition, with a reported best F1 measure of 0.324, when 5 tags were recommended for each test resource.

The paper by [Lu et al. \[2009\]](#) is also very relevant to our work as it proposes a content based approach that can produce tag recommendations for both resources for which a small number of prior tags exist and for new untagged resources. The approach proposed in this paper relies on the observation that similar resources generally have similar tags. It makes use of the cosine similarity metric to propagate tags from richly tagged resources to scarcely tagged or to new untagged resources. The data set used in this study was crawled from the *Delicious*<sup>1</sup> social bookmarking service. The content of each website was used in this approach, as opposed to simply URL titles that are available in the DC'09 dataset. Experimental results showed that the approach can be effectively used to predict tags to new untagged resources. The experiments were performed using 5-fold cross-validation. The training phase requires setting up a similarity threshold  $\epsilon$ : pairs of resources with similarity smaller than  $\epsilon$  are not used in the propagation procedure. Three values were used for this parameter: 0.02, 0.03 and 0.04. The results were reported in terms of precision at 1, 3, 5 and 10. The best precision value of 0.69 is obtained for P@1, when  $\epsilon = 0.02$ .

[Song et al. \[2011\]](#) proposed two document centered approaches (as opposed to user centered approaches) to recommend tags to new resources. The first approach is a graph based, which represents data in the form of two bipartite graphs, one which contains documents and tags, and a second one which contains documents and words. Based on these graphs, topics are identified using graph partitioning methods. The second approach proposed in this work is a prototype-based approach, which finds the most representative documents in a

---

<sup>1</sup>[www.delicious.com](http://www.delicious.com)

collection and uses a Gaussian process classifier to classify documents into multiple classes. For both methods, new documents are classified into one or more topic/classes and tags representative for those tags/classes are recommended to the new untagged document. In this work the authors tested their methods on three datasets crawled from [CiteULike](#)<sup>1</sup>, [Delicious](#)<sup>2</sup> and [BibSonomy](#)<sup>3</sup>. The content of the documents was used in the analysis. The CiteULike dataset has 9,623 distinct papers and 6,527 distinct tags. The Delicious data set was crawled from 20 popular topics and contains 22,656 URLs and 28,457 distinct tags. Finally, the BibSonomy dataset was constructed by first identifying 50 tags and then crawling the content of bookmarks with related tags. This resulted in 14,200 resources with 6,321 tags. Earlier data was used for training and later data was used for testing the proposed approaches. The F1 values reported for the three datasets are in the range of 47% to 54%. The prototype method consistently outperformed the graph-based method. Both approaches outperform an LDA based approach similar to our topic distance based approach.

The LDA technique has been previously applied to the problem of tag recommendation in the work by [Krestel et al. \[2009\]](#). The paper uses the `LingPipe` implementation of LDA<sup>4</sup>. The approach proposed is similar to our topic words based approach, except that the probability distribution of words given documents is obtained by combining the topic distribution given documents with the word distribution given topic. A threshold is used to select tags to be recommended to test resources. The dataset used in this work was made available by [Hotho et al. \[2006\]](#), who crawled the [Delicious](#) website. The original dataset consisted of approximately 75,000 users, 500,000 tags 3,200,000 resources and a total of 17,000,000 tag assignments, but was very sparse. Some preprocessing was applied on the original dataset to generate a denser dataset. Thus, the dataset used in the experiments was relatively small, consisting of 10,000 resources and 3600 tags. Furthermore, the test resources had 1 to 5 tags that were previously assigned and used in the LDA model, in other

---

<sup>1</sup>[www.citeulike.org](http://www.citeulike.org)

<sup>2</sup>[www.delicious.com](http://www.delicious.com)

<sup>3</sup>[www.bibsonomy.org](http://www.bibsonomy.org)

<sup>4</sup>Alias-i. 2008. `LingPipe` 3.7.0. <http://alias-i.com/lingpipe> (accessed October 10, 2008)

words the approach was not used to assigned tags to resources that had no tags assigned to them already. The best F1 value reported was 0.281 and obtained for resources that had 5 prior tags assigned to them. An F1 value of 0.126 was obtained for resources that had only one prior tag.

The work done by [Adrian et al. \[2007\]](#) generates semantic tag recommendations for documents based on a semantic web ontology and Web 2.0 services. Documents are represented in the RDF format and potential tags are extracted using Web 2.0 services. These tags are validated against a domain specific ontology. The dataset used for evaluation consists of 11 documents corresponding to websites about projects or employees. An ontology related to this domain was used for validating the tags. The tags recommended are evaluated and rated by 8 human experts. Precision and recall ratios are calculated based on the ratings provided by the experts. The recall ratio reported was above 60%, with a precision above 70%. However, the downside of this approach is that it uses a small number of documents from a limited domain, as opposed to using a large number of documents from a big variety of domains, as it is the case with the DC'09 dataset.

As opposed to previous approaches, in this work, we use LDA to recommend tags for resources for which no tags have been previously assigned. The dataset that we use consists of 263,004 training resources, 16,898 test resources and 1,401,104 tags (tag vocabulary used to infer the LDA model). Thus, the dataset that we use is much larger than some datasets used in prior studies.

# Chapter 6

## Conclusions and Future Work

In this chapter, we discuss the conclusions of the work presented in this thesis in Section 6.1, followed by some ways by which it can be improved in the future in Section 6.2.

### 6.1 Conclusion

In this section, discuss the research questions raised in Chapter 4 in the light of the results presented in Chapter 4. The first research question which we asked was related to the effectiveness of the proposed LDA approaches for the tag recommendation problem. Experimental results have shown that both the topic words based approach and the topic distance based approach fail to predict accurate tags, especially when only the resource description and available tags are used to construct the documents that are provided as input to MALLET. The recommendations are better when the contents of the resources are used to construct the documents. The topic distance approach works slightly better than the topic words approach when the original dataset is used (resource descriptions and tags only), while the topic words approach works better when the crawled dataset is used (resource contents).

The second questions asked was implying a comparison between the usefulness of the resource descriptions as opposed to resource contents. Based on the experiments performed, it is clear that using the contents of the documents leads to more dense data and therefore better predictions. The descriptions are too brief for providing good recommendations.

At last, the third research question that we raised referred specifically to the accuracy of the results obtained using the LDA approaches. We measured the accuracy using the F-1 measure. The values obtained for both approaches on the original dataset were smaller than 0.1. However, the methods gave an F-1 measure better than the value reported in [Lipczak et al. \[2009\]](#), when the crawled dataset was used.

From the answers above, we can conclude that the performance of the LDA approaches on the tag recommendation problem is not as good as we had expected. The discussion above clearly shows that the method of constructing documents using the resource description doesn't work well. We also see that the crawling approach to constructing documents (using the contents of the resources) works well better. However, the amount of data we used for focused crawling has been very small. Apart from that we had to manually identify the topics and resources belonging to those topics, which was time consuming.

## 6.2 Future Work

As part of the future work, we would like to explore several ideas that could help improve the results in this thesis. These ideas include the following:

- We would like to crawl the entire dataset from ECML/PKDD DC'09 and give it as input to a tool which can apply the LDA technique on whole set of documents. As we have seen from the results shown in Chapter 4, the crawled dataset, where the document contents were used as input to LDA, gives significantly better results.
- We tested our approaches on 200 and 400 inferred topics. We would like to vary the number of topics at a finer scale to see how the performance of the methods described in the thesis varies with the number of topics.
- In this work, we recommended only the top five words corresponding to a topic or a URL. It would be interesting to see if the results improve if we vary the number of top words suggested to a resource, based on a threshold.



# Bibliography

- B. Adrian, L. Sauermann, and T. Roth-Berghofer. ConTag: A semantic tag recommendation system. In *Proceedings of I-Semantics' 07*, pages pp. 297 – 304. JUCS, 2007. URL <http://www.dfki.uni-kl.de/~sauermann/papers/adrian+2007a.pdf>.
- D.M. Blei, A.Y. Ng, and M.I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993 – 1022, 2003. URL <http://www.cs.princeton.edu/~blei/papers/BleiNgJordan2003.pdf>.
- A. Hotho, R. Jaschke, C. Schmitz, and G. Stumme. Information retrieval in folksonomies: Search and ranking. In Y. Sure and J. Domingue, editors, *The Semantic Web: Research and Applications*, volume 4011 of *Lecture Notes in Computer Science*. Springer, Heidelberg, Germany, 2006.
- R. Krestel, P. Fankhauser, and W. Nejdl. Latent Dirichlet allocation for tag recommendation. In *Third ACM Conference on Recommender Systems, RecSys '09*, pages 61 – 68, New York, NY, USA, October 2009. ACM. URL <http://www.13s.de/web/upload/documents/1/recSys09.pdf>.
- S. Kullback and R.A. Liebler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79 – 86, 1951. URL [http://projecteuclid.org/DPubS/Repository/1.0/Disseminate?view=body&id=pdf\\_1&handle=euclid.aoms/1177729694](http://projecteuclid.org/DPubS/Repository/1.0/Disseminate?view=body&id=pdf_1&handle=euclid.aoms/1177729694).
- M. Lipczak, Y. Hu, Y. Kollet, and E. Milios. Tag sources for recommendation in collaborative tagging systems. In *ECML PKDD Discovery Challenge 2009 (DC09)*, volume 497, pages 157–172, September 2009. URL [http://ceur-ws.org/Vol-497/paper\\_19.pdf](http://ceur-ws.org/Vol-497/paper_19.pdf).
- Y.T. Lu, S.I. Yu, T.C. Chang, and J.Y.I. Hsu. A content-based method to enhance tag recommendation. In *International Joint Conference on Artificial Intelligence*, pages 2064

- 2069, Pasadena, California, July 2009. URL <http://ijcai.org/papers09/Papers/IJCAI09-340.pdf>.
- A.K. McCallum. MALLET: A machine learning for language toolkit. 2002. URL <http://mallet.cs.umass.edu>.
- M. T. Mitchell. *Machine learning*. McGraw-Hill Companies Inc., international edition, 1997.
- Y. Song, L. Zhang, and C.L. Giles. Automatic tag recommendation algorithms for social recommender systems. *ACM Transactions on the Web (TWEB)*, 5(1):4, 2011. URL <http://research.microsoft.com/pubs/79896/tagging.pdf>.
- M. Steyvers and T. Griffiths. Probabilistic topic models. In T. Landauer, D McNamara, S. Dennis, and W. Kintsch, editors, *Latent Semantic Analysis: A Road to Meaning*. Laurence Erlbaum, 2007.