Web Mining Phase 4

Alimuddin Khan, Gaurav Gawade, Chirag Kular

ABSTRACT

Today finding right content regarding latest topic on the internet is one of the most difficult and time consuming process. We went through different news channels websites, trending topics on different social media like twitter, Facebook etc. and different blogging platforms like Wordpress, Blogger etc. to get the idea of what people are looking for and what these media are providing us. With growing web content, it is tedious to follow precise information and get the content which exactly matters us the most. In this project, using the data mining techniques, we will be filtering out content specific to the user depending on his interests and will give him related content suggestions as well.

1. OVERVIEW

As web mining is a huge topic, we will be focusing only on blogs. For this purpose, we have chosen WordPress as our research blogging platform as it is most widely used blogging platform throughout the world. There are millions of bloggers writing everyday. With the increasing number of blogs, it is very difficult to reach the interesting and famous blogs by just mere Google search. This project not only searches/crawls on blogs hosted on wordpress.org but also on custom domains. We will be using WordPressś RESTful APIs to collect the data.

2. DESIGN

We have designed our application such that while consuming WordPress APIs, we are creating a kind of relationship between posts and hence classifying them as related and non-related posts.

3. IMPLEMENTATION

The project has three stages:

1. *Crawl:*

   This stage collects all the information/data related to blogs like blog id, user id, blog comments, commenterś id, no of comments etc. The ER-diagram of the database we will be generating will look like as follows:

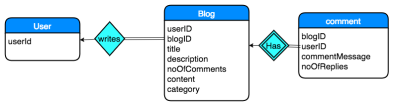

Figure 1: ER-diagram of database

2. *Classify:* In this stage we will classify the blogs into different categories and into different popularity levels depending on user profile and interactions made on that blog.

3. *Presentation:* This stage presents the blog to the end-user depending on his/her profile. This stage is also responsible for suggesting related articles. Most of the UI part will be done in this stage.

3.1 Work Done After Phase 1

We having successfully completed following steps

1. *Model*

   (a) We have successfully created model classes which we will be using for our project. These classes are as follows:

   (i) Post.java:
   - This class is responsible for storing post details like post url, post content, post title, post comments, post likes etc.

   (ii) Reply.java:
   - This class stores the comment details for a particular post.
   - Post class wraps this Reply object to store the comment details.

   (iii) Blogger.java
   - This class store the blogger info such as blogger's site id and list of Posts for that particular blogger.
   - This class wraps Post object to store post details for a particular blogger.

   (b) These classes build the core of our project. Other classes use these core classes to achieve the required objective.

2. *Controller:*
   We have created following controller class:

   (a) BlogMining.java:
   This controller is capable of doing following tasks;

   (i) Requesting WordPress APIs:
   - It can request the WordPress APIs and can collect the response.

   (ii) Parsing the response:
   - It can parse the collected response from the WordPress API into suitable information.

   (iii) Crawl:
   - It can use above two tasks to crawl the WordPress blog.
   - Crawling can be achieved in following ways;
     - Using blog url (Crawls a single post)
     - Using blogger id (Crawls all the posts for that blog id)
     - Using Post Id and Blogger Id (Crawls a single post)

3.2 Work Done After Phase 2

We have created the database to store the required details of the blog.

1. MySQL database:
   - Created a database called web_mining which will store the crawled data in above steps.
   - Our database has following tables:
     - post: To store post details
     - blogger: To store blogger details
     - replies: To store comment details.
     - related_blogs: To store a relation between a blog and related blogs.
2. MySQL controller.
   - Created a controller using jdbc library to insert, delete, update and query the database created in above step.

3.3 Work Done After Phase3

User Interface: We have created a Command Line Interface (CLI). It will be able to do following tasks:

- Fetch a list of famous blogs from the database
- Suggest related blogs depending on user selection.

4. DATA CLASSIFICATION

We are classifying the data i.e. blog posts based on post id, comments, commenter's id. The data classification is performed based on the relationship between the blog posts. We would get an ID of an blog post as input, and then we search for the all the blog posts related to that Post ID. It basically works in two phases

1. Learning Phase - In this phase, it learns the relationship between different posts
2. Classifying Phase - From the learning phase, it classifies the given post ID depending on what it had learned earlier and gives us related posts.

To search for related blogs from a given input URL, if it doesn't find any data from the already learned data, then it crawls the URL to get the related blog.

5. CONCLUSION

We have successfully implemented the program which can do the following:

1. It can find related blogs for any new blog either by crawling it or by data already learned earlier.
2. Present blogs to the end user in decreasing order of their popularity. Hence, blogs which were popular were presented first to the user and user can enjoy the content which is in trend.

6. RELATED WORK

We are referring to a research paper[1] which basically focusses on web intelligence and shows how we can mine the web to get meaningful and useful content.

7. FUTURE WORK

This project has the wide variety of applications. Right now we are just working on Wordpress. We can extend this work to Facebook, twitter, Instagram, Blogger [2] etc. to collect interesting and trending contents from their respective platforms as well. This will put all our things into a single platform and hence we won't have to look all those sites for interesting contens.

8. REFERENCES

[1] P. Mika. Social networks and the semantic web. In *Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 285–291. IEEE Computer Society, 2004.

[2] M. A. Russell. *Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More.* " O'Reilly Media, Inc.", 2013.