**Identifying influential bloggers in a community**

102502515 韓文彬, 102502551 林佳育

1. **Reference Paper**

Identifying the Influential Bloggers in a Community

Nitin Agarwal, Huan Liu, Lei Tang, Philip S. Yu

ACM, 2008, WSDM '08 Proceedings of the 2008 International Conference on Web Search and Data Mining

<http://www.public.asu.edu/~huanliu/papers/wsdm08.pdf>

1. **Introduction**

The paper is going to find a way to show the influence of bloggers, and show influential bloggers. They get TUAW website’s top 5 active bloggers to compare with their own model, and find some bloggers who are both active and influential.

Based on the model proposed by paper, we improve model by get the mean value because we consider that a blogger’s influence shouldn’t be decided by only one article. Therefore, we take blogger’s average value, deriving from all articles’ scores. We modified the original model to achieve our goal, and observed the result.

1. **Methods**

Formula: Score = numinlink - numoutlink + lengtharticle\*0.001 + numcomment

In order to calculate the score of each article, we use the formula proposed in paper and modified it to correspond to our dataset. Also we run our two model, named Max model and Mean model, on our dataset. After getting all bloggers’ scores, we can rank them via Excel.

For the sake of gaining ground truth, we chose top 10 influential articles and last 10 articles obtaining from Max model, and asked some friends to vote those articles in view of influence. Last, the result of voting is regarded as ground truth.

1. **Dataset**

Our research is consulting on the Pixnet, which is a large blogger in Taiwan. The dataset is getting from Pixnet API. First, we connected to the API to get the weekly hot articles in JSON format. Next, in JSON data, we can know the blogger’s username, which is used in API to get all the blogger’s articles information in JSON format. The last step is visiting the website from the link in JSON data and stored the html page in local-end. After five-day crewing, there is up to 10 GB dataset.

1. **Experiment**

Same as what we have mentioned before, after getting dataset, we just ran our models on them and ranked the scores. And then we can get the result. There is an important thing that the length of article is multiplied by 0.001, which made its portion of the formula not be too high. Finishing the process, we just started to observe the result as compared to the ground truth.

1. **Result**

|  |  |  |  |
| --- | --- | --- | --- |
| Max | Mean | Rank on Pixnet | Ground truth |
| milo0992 | loislinlin | dacota | gogoami |
| stellahyc | yeahyeh70 | ifans | stellahyc |
| gogoami | gogoami | changyang319 | loislinlin |
| loislinlin | stellahyc | isvincent | dacota |
| People2 | People2 | tu0925399900 | ifans |
| mcdull | ofeyhong | ahuiliao | People2 |
| ifans | karellq628 | showmer | mcdull |
| kato3c | felicia8188999 | funiphone | milo0992 |
| yeahyeh70 | dx2a25h | yoyonini362425 | yeahyeh70 |
| dacota | ifans | cpe1208 | kato3c |

1. **Conclusion and Discussion**

There are only two same bloggers in the result of Max model and Rank on Pixnet. However, the Rank on Pixnet is based on the numbers of visit, comment and subscriber. We can infer that influential is not equal to hottest. Also, according to the result, we can learn that the Mean model is not quite good to use. It doesn’t do better.

However, our models can filter advertisement blogger successfully because the last 10 bloggers are all about commerce and their usernames are garbled. They are all fake users! Also, there are many links to unsafe websites in their articles.

To make progress in our model, we should take pictures into consideration because dacota, one blogger, his articles are multiple of pictures but less words, which makes his score lower. This doesn’t reflect the real situation in the world. Therefore, pictures should matter the influential score.

From the model, we also consider the rule of blog in the world. Due to the fast generation coming, there are less and less people using blogs and writing articles on blogs. Influential bloggers can’t be figured out easily because only some people are left using blogs. Maybe one day, blogs will disappear in the world.

Github: <https://github.com/HanVincent/SMM_FinalProject>