Who to follow on Twitter given topics, using TF-IDF and PageRank

GROUP 7

Anton Lindqvist antoli@kth.se
Anna Lindelöf anna@kth.se
Thiago Lobo thiagol@kth.se
Helder Martins helder@kth.se
Casper Renman casperr@kth.se

Abstract

1 Introduction

2 Related work

3 Method

3.1 Crawling Twitter

The Twitter API [5] was used to fetch tweets related to a limited set of hashtags (Appendix A). The crawler collected 10Gb of data which the group decided was sufficient for the use-case.

3.2 The Neo4j database

Neo4j is a graph database that removes the need to explicitly define a schema for the relationships between entities. It has efficient techniques for storing graphs, making it suitable for storing large amounts of data with many relationships. Neo4j allow us to represent the full structure of the Twitter database as a graph, where each different entity is a *node* and the relationship between the nodes an *edge*. Figure 1 details the nodes and the edges of our recommender engine.

The base node of the graph is the *Tweet*, which is a free-text short message sent through Twitter. Every tweet is posted by a *User*, which is represented by the *post* relationship. A tweet may *mention* another user with the @ special character, and it may also *tag* a topic with #, represented by the node *Hash*. To this basic schema derived directly from Twitter, we added the node *Word* which are all the parsed words in the free text of the Tweet. This node is linked to the rest of the graph through the *contains* and *discusses* relationships, with a property specifying the amount of times this relationship happens for every tweet and user. This new entity allows us to easily represent the user as a *bag-of-words* document of every word that the user discusses in all of the user's tweets, thus allowing us to map the recommendation problem to the standard techniques used on information retrieval.

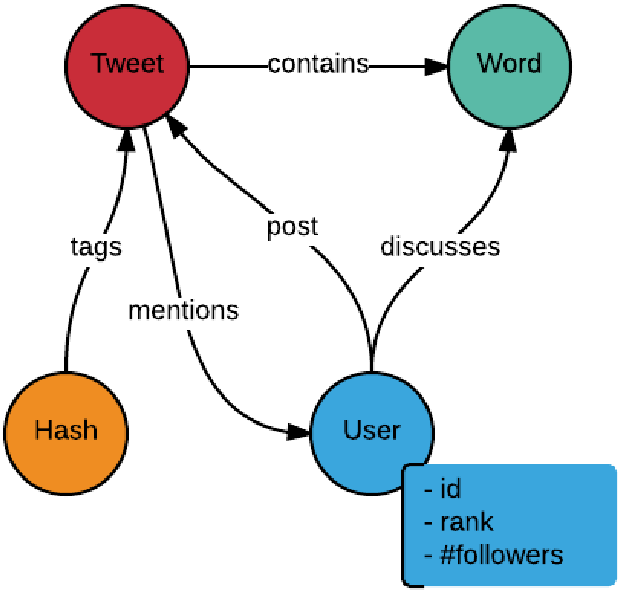

Figure 1: Graph schema used for the Twitter data.

3.3 Parsing tweets and extracting topics

The goal of the project is to recommend users given topics. In order to recommend a user, the user needs to be associated with the topics the user talks about. Therefore the users tweets are parsed and the topics of the tweets are extracted. The topics are extracted by parsing the freetext of the tweets and extracting the nouns and adjectives. The choice of extracting nouns and adjectives was an empiric decision made by the group.

Extracting topics from tweets is done using the Natural Language Toolkit (NLTK) [2] which provides interfaces in Python for things like classification, tokenization and stemming.

3.3.1 Cleaning tweets

A tweet can contain hyperlinks, hashtags, mentions and other symbols. These are removed in order to properly parse the text of the tweet. Specifically, words starting with *#, @, & or http* are ignored. A few other words that commonly occur in a tweet were also ignored as they would not contribute to the cause. These are *don't, i'll, retweet and rt.*

3.3.2 Extracting nouns

The nouns (topics) are extracted by performing the following actions, provided by NLTK:

1. Lowercase all letters and tokenize the text into separate tokens
2. Remove words that are shorter than three characters (This was also a decision made by the group)
3. For each word, remove ignored symbols and words starting with a ignored symbol
4. Part of Speech-tag [4] the words
5. Pick the words that are tagged as `NN` (noun) or `JJ` (adjective)
6. Stem the words and return the result which is a list of words

3.4 PageRank

One of the most well-known ranking and scoring measures is called PageRank [3]. Made famous by Google in late 90's, its main idea is to use the auxiliary information, mainly the *link structure*, present in the World Wide Web as an *authority measure* of the web pages contained within. Representing the web as a graph were each node is a web page and the edges the links between a page and another, it is intuitive to see that nodes with higher number of *inlinks* (that is, the number of links arriving into a node) are of higher importance than the ones with no inlinks at all, just like a scientific article which is cited by several different sources, for example.

3.4.1 PageRank in the Twitter graph

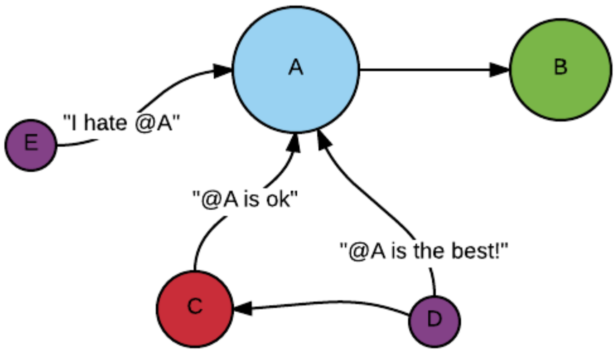

Figure 2: PageRank applied to the Twitter database. Every user is a node, while every mention in tweets is an edge. The size of the node is its relative rank among others.

Although the original PageRank algorithm was modeled with focus on the World Wide Web, its method could be applied to any problem which can be modelled as a graph. Specifically for Twitter, one could see each *user* of the platform as a node and every *mention* in the tweets of a user to another as a link. In the same way that web pages with high number of inlinks have a higher rank, users that are mentioned frequently will be considered more relevant for our recommendation engine, this process can be more clearly seen in Figure 2. Note that we actually do not analyse the content of the tweet, so tweets with positive or negative sentiment will have the same importance for ranking, one could think of it being a ”any publicity is good publicity” kind of model.

The original PageRank algorithm considered following an outlink with equal probability among all the possible links. That is reasonable with the unstructured meta information available in the Web today, but is intuitive to reason that, with more information about these users, different probabilities could be applied to each one of them, depending on the task that we have at hand. For a user recommender engine, our approach used the *number of followers* as a good measure of importance. That is, users with high number of followers will be jumped to with higher probability in the random walk,so their score will be naturally higher. Our engine implemented both methods for evaluation, and the results are reported in the Experiments section.

3.4.2 PageRank Monte Carlo

The standard implementation of the PageRank computation is done via a method called power iteration, which involves finding the largest eigenvector of a transition matrix $\mathcal{P}$ composed of the transition probabilities between every web page of the World Wide Web, a process which is done over several iterations until convergence. This method, although popular and still used today by Google, has its drawbacks mainly regarding the speed of convergence, several passes may be needed until the desired precision is obtained. In our approach we explored a relatively new method, which utilize *Monte Carlo algorithms* to estimate the score of the nodes of the graph. As proposed by Avrachenkov et al. [1], the idea is that, if we sample the web page after a sufficient large amount of random walks, a probability distribution could be calculated with an acceptable degree of precision and with a faster convergence rate. While the power iteration method may require more than 50 iterations for an acceptable ranking to be reached, Avrachenkov et al. proposed method has ranks for the import pages after one iteration only.

Of the several different algorithms proposed, our engine implements the *Monte Carlo complete path*, which is detailed in the Algorithm 1. For every user in the Twitter database, we start a random walk beginning in that user and ending when the user is bored of following mentions. We keep track of the total steps of all random walks and how many times each user was visited. A new user is selected to be followed in the walk from all the users the user mentions, which can be done by applying equal probabilities to each one of them or with increased chance for higher number of followers. If a user does not mention anyone, we consider it a *sink* and jump to any other user in the database with the same method. After every user has been at the beginning of the random walk for a set number of walks, we calculate the user rank by dividing the number of times each user was visited over all random walks with the total steps taken.

3.5 TF-IDF

Ranked user retrieval can be implemented by only using the aforementioned *PageRank* algorithm but by doing so, any query would return the same top listed users. While that might be interesting in some applications, that is not the case in our context. The words in the query should also be used to filter and rank the retrieved users.

TF-IDF is a well known solution to the problem of matching (in a ranked

Algorithm 1 PageRank Monte Carlo, complete path

```
procedure PageRank
    for all walks do
        for all user in users do
            username ← user['username']
            bored ← False
            while ¬bored do
                totalSteps ← totalSteps + 1
                userSteps['username'] ← userSteps['username'] + 1
                mentions ← getUserMentions(username)
                if mentions ∈ ∅ then
                    username ← getRandomUser(users)
                else
                    username ← getRandomUser(mentions)
                bored ← isUserBored()
    for all user in users do
        username ← user['username']
        ranks['username'] ← userSteps['username'] ÷ totalSteps
    return ranks
```

way) documents modelled as *bags-of-words*. Each document (including the input query) is represented by a vector of scores, each of which related to one of the possible terms in our dataset. The scores are calculated as follows: $tf_{w,d} * log_{10}(\frac{N}{df_w})$ where $tf_{w,d}$ is the number of times term w appears in document d, N is the total number of documents and df_w is the number of documents term w appears in. Then, *cosine-similarity* is used to compute how close the query is to each of the documents.

In our implementation, *User* nodes are documents containing each of the *Word* nodes they are linked to. This link contains the number of times this *Word* has been discussed by this *User*, that is, a $tf_{w,d}$ score. The final procedure can be seen in Algorithm 2.

3.6 Final Score

After retrieving the sets of *PageRank* and *TF-IDF* scores for all users that discuss any query term, we decided to first normalize them to zero-average and unitary variance so that they can be mixed together by means of a parameter α. Then, the final score, for each user u, is: $s_u = \alpha * \bar{s}_{p_u} + (1 -$

Algorithm 2 TF-IDF in a Graph Database

```
procedure TF-IDF
    scores ← ∅
    sizes ← ∅
    for token ∈ query do
        users ← query(users that discuss 'token')
        df ← length(users)
        count ← # of occurences of 'token' in 'query'
        wtq ← count * log10(length(documents) / df)^2
        for user ∈ users do
            tf ← query(# of times 'user' discusses 'token')
            scores[user] ← scores[user] + wtq * tf
            sizes[user] ← query(# of words discussed by 'user')
    for user ∈ scores do
        scores[user] ← scores[user] / sizes[user]
    return sort(scores)
```

$\alpha) * \bar{s_{t_u}}$.

4 Experimental results

4.1 Ranking algorithms

4.1.1 Only PageRank

resulting list/table

4.1.2 Only tf-idf

resulting list/table

Small reflection (relevance feedback). What do we think? What should alpha be?

4.1.3 Combination

Try 3 different alphas and show list/table

4.2 Evaluation

Summary of what alpha should be and why.

5 Evaluation of the result

6 Summary and Conclusions

References

[1] K. Avrachenkov, N. Litvak, D. Nemirovsky, and N. Osipova. Monte carlo methods in pagerank computation: When one iteration is sufficient. *SIAM Journal on Numerical Analysis*, 45:890–904, 2007.

[2] Steven Bird. Nltk: the natural language toolkit. pages 69–72, 2006.

[3] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. *World Wide Web Internet And Web Information Systems*, 54:1–17, 1998.

[4] Helmut Schmid. Probabilistic part-of-speech tagging using decision trees. 12:44–49, 1994.

[5] INC Twitter. Twitter api, 2016.

7 Appendix

A Hashtags

#FeelTheBern, #Bernie2016, #BernieSanders, #NotMeUs, #Bernie, #UniteBlue, #StillSanders, #NYPrimary, #WIPrimary, #ImWithHer, #Hillary2016, #HillaryClinton, #Hillary, #Trump2016, #MakeAmericaGreatAgain, #TrumpTrain, #Trump, #tcot, #AlwaysTrump, #TeamTrump, #WakeUpAmerica, #ccot, #TeaParty, #DonaldTrump, #PJNET, #elections2016, #vote, #cir, #USlatino, #AINF, #Latinos, #GOP, #2016Election.