Identifying Topic-Related Hyperlinks in the Crowd

Patrick Siehndel, Ricardo Kawase, Eelco Herder and Thomas Risse

L3S Research Center, Leibniz University Hannover, Germany
{siehndel, kawase, herder, risse}@L3S.de

Abstract. The microblogging service Twitter has become one of the most popular sources of real time information. Every second, hundreds of URLs are posted on Twitter. Due to the maximum tweet length of 140 characters, these URLs are in most cases a shortened version of the original URLs. In contrast to the original URLS, which usually provide some hints on the destination Web site and the specific page, shortened links do not tell the users what to expect behind them. These links might contain relevant information or news regarding a certain topic of interest, but they might just as well be completely irrelevant, or even lead to a malicious or harmful website. In this paper, we present our work towards identifying credible Twitter users for given topics. We achieve this by characterizing the content of the posted URLs to further relate to the expertise of Twitter users.

1 Introduction

The microblogging service Twitter has become one of the most popular and most dynamic social networks available on the Web, reaching almost 300 million active users [1]. Due to its popularity and dynamics, Twitter has been topic of various areas of research. Twitter clearly trades content size for dynamics, which results in one major challenge for researchers - tweets are too short to put them into context without relating them to other information.

Nevertheless, these short messages can be combined to build a larger picture of a given user (user profiling) or a given topic. Additionally, tweets may contain hyperlinks to external additional Web pages. In this case, these linked Web pages can be used for enriching tweets with plenty of information.

An increasing number of users post URLs on a regular basis. By analyzing the Twitter Stream API, we estimated that, on average, around 800 URLs are posted every second on Twitter. With such a high volume, it is unlikely that all posted URLs link to trustful or relevant sources. Thus, measuring the quality of a link posted on Twitter is an open question.

In many cases, a lot can be deduced just by the URL of a given Web page. For example, if the URL domain is from a news provider, a video hosting website or a social network, the user already knows more or less what to expect after clicking on it. However, regular URLs are, in many cases, too long to fit in a single tweet. Consequently, Twitter automatically reduces the link length using shortening services. This leads to the problem that the user's educated guess of what is coming next is completely gone. Clicking on a link on Twitter is always a surprise and potentially a risk.

In this work, we focus on ameliorating these problems by identifying those tweets containing URLs that might be relevant for the rest of the community. Our method

provides a score for a URL posted by a given user that represents its estimated relevance for a given topic.

The reasonable assumption behind our method is that users who usually talk about a certain topic ('experts') will post interesting links about the same topic. The strong point in our method is that it is independent of the users' social graph. There is no need to verify the user's network or the retweet behavior. Thus, it can be calculated on the fly. To achieve our final goal, we divide our work in two main steps: the generation of user profiles [4] and the generation of URL profiles. In this paper, we focus on the latter step.

2 Methodology

In our scenario, we define a domain expert as a person who has a deeper knowledge regarding a certain domain than the average user. In particular, we analyze if the tweets published by a certain user contain keywords that are very specific and part of a professional vocabulary. The biggest challenge in this task is to find appropriate algorithms and metrics for defining whether a word is part of this professional vocabulary or not. Beside that, we also need to know in which area the user is an expert. The method we developed to solve this task is based on the vast amount of information provided by Wikipedia. We use the link and category information supplied by Wikipedia to define the topic and the expertise level inherent in certain terms. Our method consists of three main steps to create an expert profile for a user.

Extraction: In this step, we annotate the given content (all tweets of a user, or the contents of a Web page) using the Wikipedia Miner Toolkit [3]. The tool provides us with links to Wikipedia articles. The links discovered by Wikipedia Miner have a similar style to the links that can be found inside a Wikipedia article. Not all words that have a related article in Wikipedia are used a links, but only words that are relevant for the whole topic are used as links.

Categorization: In the second stage, Categorization, we extract the categories of each entity that has been mentioned in the users' tweets or in the posted URL. For each category, we follow the path through all parent categories, up to the root category. In most cases, this procedure results in the assignment of several top-level categories to an entity. Since the graph structure of Wikipedia contains also links to less relevant categories, we only follow links to parent categories which distance to the root is shorter or less than the one of the child category. For each category, a weight is calculated by first defining a value for the detected entity. This value is based on the distance of the entity to the root node. Following the parent categories (which are closer to the root), we divide the weight of each node by the number of sibling categories, resulting in each entity receiving 25 category scores. Based on this calculation, we give higher scores to categories that are deeper inside the category graph and more focused on one topic.

Aggregation: In the final stage, Aggregation, we perform a linear aggregation over all of the scores for a document, in order to generate the final profile for the user (or for the website). The generated profile displays the topics a user/website talks about, as well as the expertise in - or focus on - a certain topic.

3 Validation

As mentioned in Section 1, in this paper we focus our attention on the generation of URL profiles and the relation to the corresponding tweets and users. Thus, in order

Table 1. Statistics about the used dataset.

	Items	Annotations	Annotations per Item
Topic Tweets	83,300	88,530	1.06
URLs	40,940	457,164	11.1
All Tweets	11,303,580	30,059,981	3.127

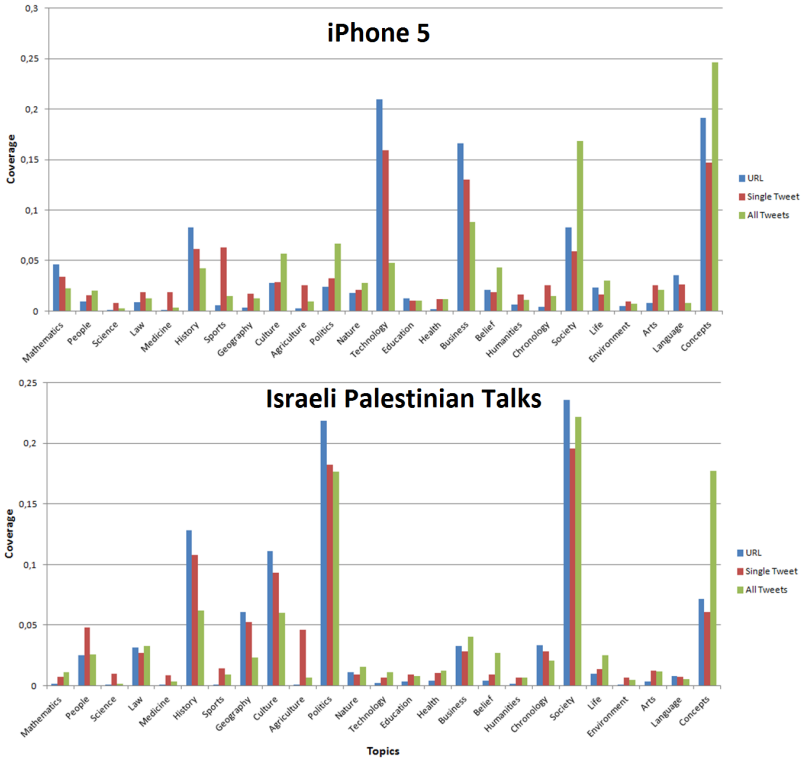

Fig. 1. Coverage of Wikipedia Categories based on the URL Content for each selected topic.

to validate our methodology, we crawled Twitter with a number of predefined queries (keywords) and collected all resulting tweets that additionally contain URLs. We have previously validated our approach by characterizing and connecting heterogeneous resources based on the aggregated topics [2]. Here, the goal is to qualitatively validate if the topic assignment given by our method in fact represents the real topics that are expected to be covered in a given query.

3.1 Dataset

The used dataset consists of around 83,300 tweets related to seven different topics. The idea behind this approach is, to collect a series of tweets that contain links and certain keywords relevant for one particular topic. Within these tweets, we found 40,940 different URLs. For each of these URLs, we tried to download and extract the textual content, which resulted in 26,475 different websites. Additionally we downloaded the last 200 posts for each user. The numbers of the dataset are shown in Table 1.

Table 2. Correlations between created profiles

	URL Content Single Tweet	URL Content User Tweets	Single Tweet User Tweets
Edward Snowden	0.995	0.968	0.961
Higgs Boson	0.812	0.628	0.496
Iphone 5	0.961	0.698	0.664
Israel Palastinian Talks	0.984	0.884	0.867
Nexus 5	0.968.	0.972	0.956
Obamacare	0.983	0.79	0.752
World Music Avards	0.921	0.718	0.614
All topics average	0.946	0.808	0.759

3.2 Topic Comparison

Figure 1 shows the generated profiles for two of the chosen example topics. The shown profiles are averaged over all users and show the profiles based on the content of the crawled web pages, based on the tweets containing the URLs and based on the complete user profile (last 200 Tweets). We can see that for the very specific topic 'Israeli Palestinian Talks' the generated profiles are very similar. For the topic 'iPhone 5' the profiles are less similar, since this topic or keyword is less specific it becomes much harder for a user to find the content he is looking for. A tweet like *'The new iPhone is really cool'* together with a link may be related to many different aspects of the product. Table 2 displays the correlation between the different profiles for the chosen exemplifying topics. While users who write about topic like 'Snowden' or 'Nexus phones' seem to write about related topics in most of their tweets, this is not true for more general topics.

4 Conclusion

In this paper, we presented a work towards the identification of credible topic-related hyperlinks in social networks. Our basic assumption is that users who usually talk about a certain topic ('experts') will post interesting (and safe) links about the same topic. The final goal of our work requires to analyze the quality of the posted URLs. Here, we presented our profiling method with preliminary results of the URL profiles. As future work we plan to analyze the quality of profiles and URLs in order to provide a confidence and quality score for URLs.

5 Acknowledgment

This work has been partially supported by the European Commission under ARCOMEM (ICT 270239) and QualiMaster (ICT 619525)

References

1. Twitter now the fastest growing social platform in the world. http://globalwebindex.net/thinking/twitter-now-the-fastest-growing-social-platform-in-the-world/, Jan. 2013.
2. R. Kawase, P. Siehndel, B. P. Nunes, E. Herder, and W. Nejdl. Exploiting the wisdom of the crowds for characterizing and connecting heterogeneous resources. In *HT*, 2014.
3. D. Milne and I. H. Witten. An open-source toolkit for mining wikipedia. *Artificial Intelligence*, 194:222–239, 2013.
4. P. Siehndel and R. Kawase. Twikime! - user profiles that make sense. In *International Semantic Web Conference (Posters & Demos)*, 2012.