Extended Abstract - Making social interactions accessible in online social networks

Fredrik Erlandsson [a], Roozbeh Nia [a], Henric Johnson [b] and Felix S. Wu [a]
[a] *University of California Davis*
[b] *Blekinge Institute of Technology*

1. Introduction

Recently, online social networks, OSNs, have gained significant popularity and are among the most popular ways to use the Internet. Additionally, researchers have become more interested in using the social interaction networks, SINs[1], in order to further enhance and personalize their services[2]. OSNs are also redefining roles within the publishing industry, allowing publishers and authors to reach and engage with readers directly[3]. However, SINs are not very easily available as of today through the current APIs provided by most OSNs. Such applications would therefore spend tremendous amount of time trying to gather the required SINs for their services. Therefore, our research problem is how we can design a system that makes social interactions in OSNs accessible. This also refers to the problem of how to crawl OSNs in a structured way, which is the focus of this short paper.

The nature of OSNs and the amount of information available makes the problem of what to crawl interesting. To narrow down the scope of the proposed research, we are focusing on the interactions in OSNs. By doing this, we noticed that there exist a gap and segregation between content and social graph. To simply provide social informatics for social computing applications, we have developed a crawler that serves as a bridge between the content and social graph in the online world, by not only providing which users have interacted with each other but around exactly which content these interactions have occurred.

Privacy of the user is a major concern when it comes to all online social interactions and crawling as discussed in [4]. We are treating the crawled data with high respect to the integrity of the people behind the users.

2. Related Work

Despite the huge number of social network publications, few have been dedicated to the data collection process. Chau et al. [5] briefly describe using a parallel crawler running breadth-first search, BFS to crawl eBay profiles quickly. The measurement conducted by

Mislove et al.[6] is, to the best of our knowledge the largest OSN crawling study ever published. From four popular OSNs, Flickr, Youtube, LiveJournal, and Orkut, 11.3M users and 328M links are collected. Their analysis confirms known properties of OSNs, such as a power-law degree distribution, a densely connected core, strongly correlated in-degree and out-degree, and small average path length.

Other studies on OSN crawlers include [7,8]. [7] proposed two new unbiased strategies: Metropolis-Hasting random walk (MHRW) and a re-wighted random walk (RWRW). [8] described the detailed implementation of a social network crawler. It used the BFS and uniform sampling as the crawling strategies to run the crawler on Facebook, and then compared the two strategies.

3. A platform to make interactions accessible

Our objective is to design a system that is able to crawl open data. Initially, we will focus on the Facebook Graph API to crawl (gather) all the content that is viewable to the users; such as posts, comments, likes on posts and comments, and shares of posts.

3.1. Design

We have designed our crawler to operate in two stages. *Stage one* uses the Facebook's unique identifier of a public community (page or a group) to find the id of all posts, messages, photos, and links posted on the given community by admins and members. For readability, a post will refer to anything shared on a community on Facebook in this paper. *Stage two* is a bit more complicated; for each post gathered in *stage one* we send at least three to four separate requests (assuming that there are no "likes" on comments), one for the post itself, one for the "likes" on the post (if there exist any), one to get information on who have shared the post and finally one to get all comments (if there exist any). If one of the responses is paginated we have to make consecutive requests to gather the complete view. This also means that for posts with a lot of interactions we have to make multiple requests to the graph. For instance, we have crawled posts with hundreds or thousands of comments each with a few likes, where we have to make a request for each comment to get its likes. To scope the huge number of requests and the requirement to be efficient, our crawler is built as a distributed service much like discussed in [5]. Figure 1 shows a basic sketch how the controller and the crawling agents are connected.

3.2. Statistics

Over the last eight months our tool have gathered a bit over 150GB of structured data, including: 93 million unique Facebook users, 14 million posts, 126 million comments and over 800 million likes.

4. Challenges and requirements

Our crawler highly depends on Facebook's API, and therefore, bugs in Facebook's API will cause problems that we have no control over. Also, resource limitations has forced

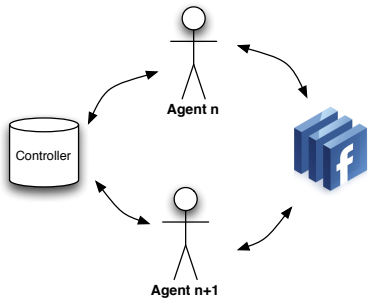

Figure 1. Our distributed crawling mechanism

us to be picky about which communities to crawl. Given enough resources, our crawler can be modified to automatically crawl all public communities on Facebook and other OSNs given an initial set of seeds.

4.1. Requirements

Our crawler tool, from a high level, is simply a black box that takes the identifier of a Facebook community as input and outputs a stream of documents. In addition to capturing the response of API requests, our crawler has to satisfy the following requirements:

Real-time. The information and interactions on Facebook public communities is extremely time-sensitive. In most cases, it is very important to crawl and parse a given post in a community on Facebook online. A few important questions that arise due to the nature of how the interactions around posts evolve are 1) "Which posts do we have to re-crawl to get the most updated information" and 2) "When would be the best time to re-crawl these posts".

Coverage. It is important and desirable to be able to crawl each and every post thoroughly and completely. However, if resources do not allow this, it is more desirable to get all the data from a limited set of posts, rather than less data from a larger set of posts.

Scale. As of today there are over a billion users and millions of public communities on Facebook[9]. There are over 2.7 billion likes and comments posted on Facebook on a daily basis as of February 1st[10].

Data Quality. The crawler should output good quality and uncorrupted data. Therefore, it needs to be able to detect failures in Facebook's current API and be able to restart from exactly where it stops when a failure occurs.

5. Applications of SINs

There are a vast number of applications where SINs can be used, here we give a brief description of two.

Dynamic News Feed: People spend hours on Facebook every day. However, they are only bound to see the posts shared by their immediate friends and pages they have

liked. Using social interactions, it is possible to identify the type of posts the user has been interacting with and find similar posts based on the SIN formed around it that the user has not interacted with. This will create a more dynamic newsfeed rather than the current one where users see the same posts over and over again throughout the day. We can identify which posts the user would be interested in using social interactions but has not interacted with yet. Therefore, the user will only see posts that he/she has not seen before and the content is socially related to what he/she likes.

Social Search: Social Search[11] is one of the hottest areas in the market and companies like Google, Facebook, and Microsoft are spending billions of dollars in the race of building the best social search experience. We believe that the SINs formed around the content shared on these pages and groups give better results when combined with a search engine than the friendship networks currently used. While a group of users have very similar and close interactions around the content shared on Facebook, we can use this information when a person from this group queries something. We know the group's interests and that will help us serve the user with better social search results. Since there is a cap on how many friends users can have on Facebook, the social search will be limited to the number of direct friends. In addition to the limited social network, there are no guarantees that users immediate friends will share the same taste, thought process, or needs. In our approach we can link users with many interactions on related content to provide better search results. Based on the query we can identify the context and use the matching SIN to find related content.

6. Conclusion

We have shown means of building an extensive tool to gather data from public communities on OSNs. Our distributed crawler satisfies all of our requirements in order to retrieve the complete set of non-corrupted data, including all the content shared and all the user interactions around them. We discuss various applications and how they can benefit from leveraging SINs in order to further personalize their services. Finally, we have given a short description of how to design a data-mining tool for OSNs that can be used to gather data.

References

[1] R. Nia, F. Erlandsson, P. Bhattacharyya, R. Rahman, H. Johnson, and F. Wu, "Sin: A platform to make interactions in social networks accessible," in *ASE International Conference on Social Informatics*, dec. 2012.

[2] C. Wilson, B. Boe, A. Sala, K. P. Puttaswamy, and B. Y. Zhao, "User interactions in social networks and their implications," in *Proceedings of the 4th ACM European conference on Computer systems*, EuroSys '09, (New York, NY, USA), pp. 205–218, ACM, 2009.

[3] A. Mrva-Montoya, "Social Media: New Editing Tools or Weapons of Mass Distraction?," *The Journal of Electronic Publishing*, vol. 15, June 2012.

[4] F. Erlandsson, M. Boldt, and H. Johnson, "Privacy threats related to user profiling in online social networks," in *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, pp. 838 –842, sept. 2012.

[5] D. H. Chau, S. Pandit, S. Wang, and C. Faloutsos, "Parallel crawling for online social networks," in *Proceedings of the 16th international conference on World Wide Web*, WWW '07, (New York, NY, USA), pp. 1283–1284, ACM, 2007.

[6] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, "Measurement and analysis of online social networks," in *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, IMC '07, (New York, NY, USA), pp. 29–42, ACM, 2007.

[7] M. Gjoka, M. Kurant, C. T. Butts, and A. Markopoulou, "Walking in facebook: a case study of unbiased sampling of osns," in *Proceedings of the 29th conference on Information communications*, INFOCOM'10, (Piscataway, NJ, USA), pp. 2498–2506, IEEE Press, 2010.

[8] S. Catanese, P. D. Meo, E. Ferrara, G. Fiumara, and A. Provetti, "Crawling facebook for social network analysis purposes," in *WIMS*, p. 52, 2011.

[9] TechChrunch, "Facebook announces monthly active users were at 1.01 billion as of september 30th, an increase of 26% year-over-year," 2012.

[10] T. Cheredar, "Facebook user data," Feb. 2012.

[11] P. Bhattacharyya, J. Rowe, S. F. Wu, K. Haigh, N. Lavesson, and H. Johnson, "Your best might not be good enough: Ranking in collaborative social search engines," in *7th International Conference On Networking, Applications and Worksharing*, 2011.